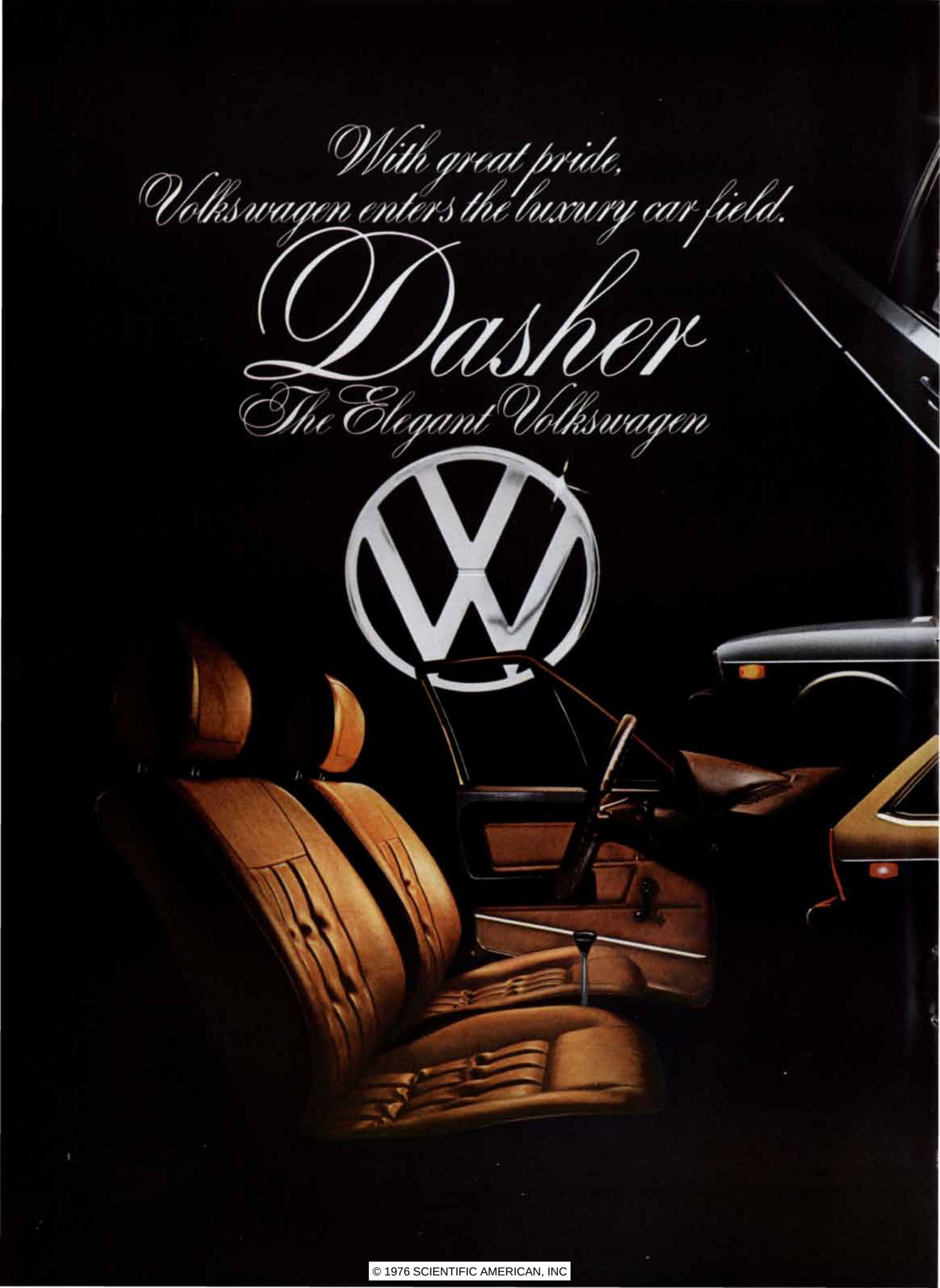
SCIENTIFIC AMERICAN

THE PHOTOGRAPHIC LENS

$1.25

August 1976



The elegant Volkswagen. New for us, but not strange for us. All Volkswagens have been elegant in their simplicity. Elegant in design. In concept. In function. Dasher is all of these. With the added elegance of timeless styling. A striking interior. Rich appointments. Dasher is a cultivated car. In size. (Extravagant inside, conservative outside.) In performance. (0-50 mph in 8.0 seconds.) In economy: 37 mpg highway, 24 mpg city. (1976 EPA estimates with standard transmission. Actual mileage may vary with your type of driving, driving habits, car's condition and optional equipment.) In serviceability. In features like steel-belted radial tires, fuel injection and front-wheel drive. You may drive the graceful sedan or the gracious station wagon. We offer these cars with great pride, to be owned with great pride. Volkswagens before Dasher have been elegantly simple. Dasher is simply elegant.

Some speakers sound fantastic on part of the music.

One of today's best speakers is famous for its highs. Another has been said to deliver the best bass, and a third is loved for its mid-range performance.

Each "best" speaker is terrific in its own particular area because it's engineered that way. To please people who are hung up on a particular kind of sound.

The new HPM-100 speaker system is different. It produces superlative sound across the whole spectrum.

That is not an accident.

Originally it was a challenge to our engineers: specialize in perfecting everything. Their response was to create a whole new universe of loudspeaker technology.

For the superhigh frequency range, to begin with, they gave the HPM-100 a radically new kind of supertweeter, made with a newly developed substance called High Polymer Molecular film. This HPM film is used to convert electrical energy into sound. It does so with a degree of accuracy that was previously only a theory.

This one sounds fantastic on all of it.

That's a lot to claim, but these speakers really are that different from everything that has gone before. They work without any magnets, voice coils or domes. In fact, without any moving parts at all.

The tweeter, mid-range driver and woofer all depart just as radically from conventional speaker technology. The key innovation in all of them is a process that turns out speaker cones heavily impregnated with carbon fibers. They are rigid and tough, but still thin and light. As a result, each of them can move easily, like a well-oiled piston, to produce high and middle and low frequencies that are clearer, more natural and far more transparent than anything you'd expect to hear from a four-way speaker system.

Talk to your high-fidelity dealer about the HPM-100.

Take along a favorite recording and listen to it alternately through HPM-100's and some comparably priced speakers. Especially speakers that sound fantastic on the high frequencies, or the low ones, or the middle range.

If there's something you like about each of today's best speakers, this is the one that can give it all to you.

PIONEER

U.S. Pioneer Electronics Corp., 75 Oxford Drive, Moonachie, New Jersey 07074.

HPM-100

The all-around great speaker.

Established 1845 **SCIENTIFIC AMERICAN** August 1976 Volume 235 Number 2

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $15 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $18 PER YEAR, ALL OTHER COUNTRIES.

THE COVER

The painting on the cover shows a summer landscape reflected in an ultra-wide-angle lens for a 35-millimeter camera. The antireflection coatings on the lens alter the normal colors in the outdoor scene. Thus the sky has a purplish cast and the reflection of the sky in the water is greenish. The sun is reflected several times by the lens's interior elements, of which there are 11 in all. If the lens were used to make a picture of the reflected scene, the image on the film would cover a 180-degree field of view, measured diagonally on the standard 24-by-36-mm. frame of a 35-mm. camera. Because of the lens's extremely wide angle of view and also because it makes the lines of rectilinear objects appear to be curved, it is known to photographers as a fish-eye lens. Made by the Minolta Corporation, it has a focal length of 16 mm. and a maximum aperture of *f*/2.8. The technology of creating such high-performance lenses is described by William H. Price in his article "The Photographic Lens" (page 72).

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
19–23	Total Communications Industries, Ltd.
25–29	Alan D. Iselin
30–31	Philip E. L. Smith, University of Montreal
32	EROS Data Center, U.S. Geological Survey
33–37	Adolph E. Brotman
47	EROS Data Center, U.S. Geological Survey
48–56	George V. Kelvin
57	EROS Data Center, U.S. Geological Survey
60	Charles Lane, Medical Research Council Laboratories, London
62–70	Bunji Tagawa
73	Norman Goldberg, courtesy of *Popular Photography*
74–82	Dan Todd
85–87	Tom Prentiss
88	Tom Prentiss (*top*); Rolf Schumacher, University of Bonn (*bottom*)
91	The Bettmann Archive, Inc.
92–98	Andrew Christie
99	Maurits C. Escher
100–105	Andrew Christie
106	Andrew Christie (*top*), Fritz Goro (*bottom*)
107–108	Andrew Christie
111	Morton Salt Company
112	Humberto Fernandez-Moran, University of Chicago
113	*Black Holes, Quasars, and the Universe.* © 1976, Houghton Mifflin Company
114	Crown Copyright. The Science Museum, London

What's new in physics...chemistry... applied analytical math... computer data base organizations... and the entire universe around us!

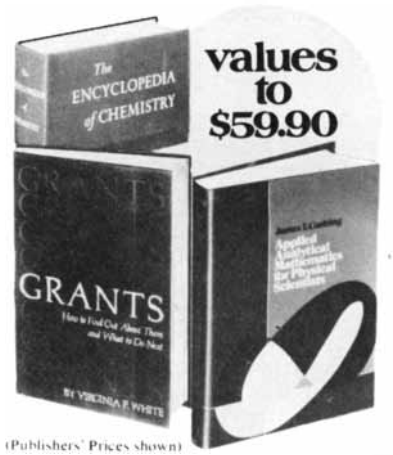
To keep in touch with everything that's happening in every scientific area that interests you, get set now for a great reading spree –

take any 3 books all for only $3.95

as a member of The

Library of Science

if you will join now for a trial period and accept only 3 more books – at member discount prices – within the next 12 months.



(Publishers' Prices shown)

50551. GRANTS: How to Find Out About Them and What to do Next. *Virginia P. White.* A "must" for anyone who wants to get, keep, or increase grant funding. **$19.50**

55375. INFRARED: The New Astronomy. *David R. Allen.* The first comprehensive account of the last decade's remarkable breakthroughs in infrared astronomy. **$12.00**

68750. PHYSICAL THOUGHT FROM THE PRESOCRATICS TO THE QUANTUM PHYSICISTS. *Samuel Sambursky.* Impressive history of the most important events in the development of the physical sciences. Includes excerpts of great scientists from Aristotle to Pauli. **$20.00**

73191. QUEUEING SYSTEMS VOLUME II: Computer Applications. *Leonard Kleinrock.* Takes informational processing theories and demonstrates how they can be used to solve real-world problems, particularly those relating to computer systems. *Counts as 2 of your 3 books.* **$25.00**

39865. COMPUTER DATA-BASE ORGANIZATION. *James Martin.* How to use the latest design and implementation techniques to improve real-time accessibility and improve flexibility – and do it at the lowest possible cost. *Counts as 2 of your 3 books.* **$26.50**

52190. HANDBOOK OF APPLIED MATHEMATICS. *Edited by Carl Pearson.* Over 1000 pages of vital and practical techniques for scientists and engineers who use mathematics as a tool. *Counts as 2 of your 3 books.* **$37.50**

73190. QUEUEING SYSTEMS VOLUME I: Theory. *Leonard Kleinrock.* "...a beautifully written innovative book. If you study queues, own this book." – Ivan T. Frisch, Editor-in-Chief of *Networks.* **$19.95**

70190. PRINCIPLES OF OPERATIONS RESEARCH. Second Edition. *Harvey M. Wagner.* This 1039-page book offers a complete overview of all the fundamental concepts. **$19.95**

44941. THE ENCYCLOPEDIA OF PHYSICS. Second Edition. *Edited by Robert M. Besancon.* Over 300 distinguished authorities in 344 articles on every major topic in physics today. More than 1000 pages. *Counts as 2 of your 3 books.* **$37.50**

34140. APPLIED ANALYTICAL MATHEMATICS FOR PHYSICAL SCIENTISTS. *James T. Cushing.* Deals with topics ranging from spectral analysis of linear operators to complex variables, to group theory. **$19.95**

36395. BLACK HOLES, QUASARS AND THE UNIVERSE. *Henry L. Shipman.* Black holes, quasars, white dwarfs, neutron stars, supernovae, Seyfert galaxies – explores the frontiers of astronomy. **$12.95**

70170. PRINCIPLES OF HOLOGRAPHY. Second Edition. *Howard M. Smith.* Covers every aspect of holography from fundamentals to applications. **$16.50**

61610. MAYA: The Riddle and Rediscovery of a Lost Civilization. *Charles Gallenkamp.* Compelling view of the life, arts, religion and science of this brilliant civilization that emerged, rose above, and then returned to the jungle. **$12.95**

57670. LANDMARK EXPERIMENTS IN TWENTIETH CENTURY PHYSICS. *George L. Trigg.* Two dozen trail-blazing experiments by some of this century's greatest physicists that signaled a marked change in man's concept of the physical world. **$18.50**

49340. FUNDAMENTAL ASTRONOMY: Solar System and Beyond. *Franklyn W. Cole.* The mysteries of the universe from earliest times to the frontiers of modern astronomical thought. **$16.50**

56380. INTRODUCTION TO PROBABILITY. *Richard L. Scheaffer and William Mendenhall.* Covers everything from probability distribution to the central limit theorem. **$13.50**

How the Club Operates • The Book Club **News**, describing the coming Main Selection and Alternate Selections, will be sent to you 15 times a year at 3-4 week intervals. • If you wish to purchase the Main Selection, do nothing and it will be shipped to you automatically. • If you prefer one of the Alternates, or no book at all simply indicate your decision on the reply form always enclosed with the **News** and mail it so that we receive it by the date specified. • The **News** is mailed in time to allow you at least 10 days to decide if you want the coming Main Selection. If, because of late mail delivery of the **News**, you should ever receive a Main Selection without having had the 10 day consideration period, that Selection may be returned at Club expense. • Continued membership past the trial period brings you the added benefits of our Bonus Book Plan – an important way to save even at least 70% off publishers' prices.

44861. THE ENCYCLOPEDIA OF CHEMISTRY. Third Edition. *Edited by Clifford A. Hampel and Gessner G. Hawley.* Over 1200 pages, with charts, diagrams, and illustrations. *Counts as 2 of your 3 books.* **$39.50**

81150. SUBATOMIC PHYSICS. *Hans Frauenfelder and Ernest M. Henley.* Covers everything from how accelerators and detectors work to the structure of subatomic particles. **$17.50**

61350. MATHEMATICAL TECHNIQUES IN CHEMISTRY. *Joseph B. Dence.* From elementary algebra through calculus to tensors, matrices, Bessell functions, and much more. **$14.50**

The Library of Science 2-4AE
Riverside, New Jersey 08075

Please accept my application for membership and send me the three volumes indicated, billing me only $3.95 for all three. I agree to purchase at least three additional Selections or Alternates during the first 12 months I am a member, under the club plan described in this ad. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments. **Send no money. Members are billed when books arrive.**

3 books for $3.95
Indicate by number the 3 books you want.

[ ] [ ] [ ]

A few expensive books (noted in book descriptions) count as 2 choices.

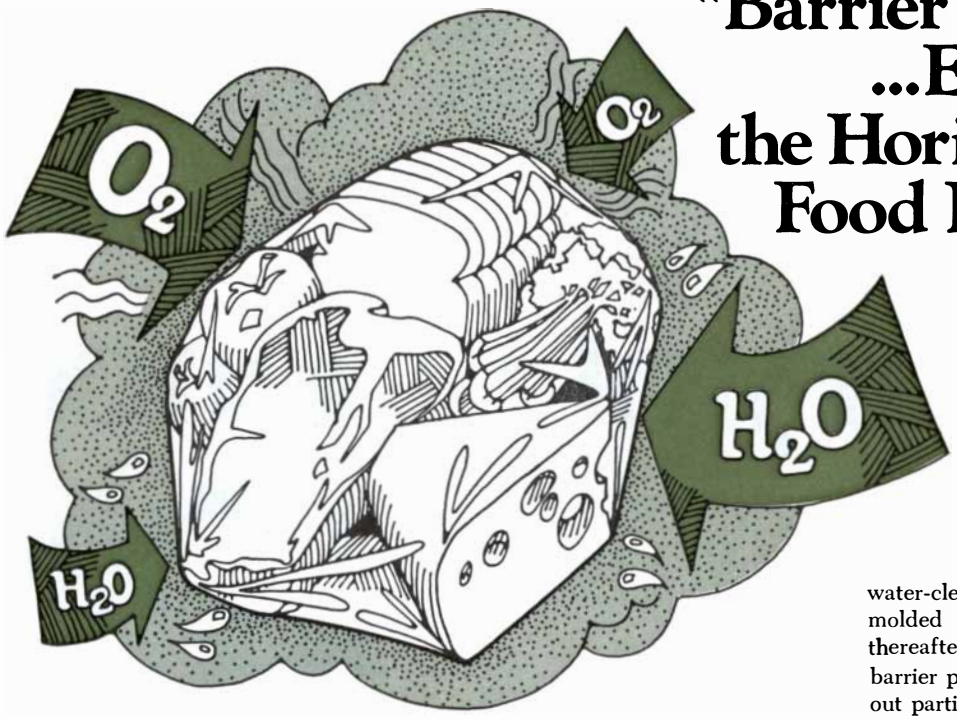
Name_____

Address_____

City_____ State_____ Zip_____

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada.) Book selections purchased for professional purposes may be a tax-deductible expense.

"Barrier Plastics" ...Expanding the Horizons for Food Packaging



John Wesley Hyatt, acclaimed the father of plastics, would be stunned by the sophistication of polymer architecture today. He would be equally confounded by the enormous variety of end-use requirements – performance demands that polymer chemists now intrepidly set about trying to meet. And none of these is more stringent than plastic packaging for food products.

As a start – food producers understandably demand virtually perfect preservation of taste, physical texture, appearance, freshness along with acceptable costs. The FDA severely scans potential contamination, often in terms of parts per billion. Distributors and retailers look for good mechanical protection, minimum handling hazards, high buyer-appeal, and long shelf-life. When the cumulative list of musts is drawn up, the expert resin maker of more than a decade ago might well have thrown in the towel.

Monsanto researchers, however, merely wiped their collective brows. After all, they had succeeded in architecting resins with high modulus, optical clarity, food grade purity, and good heat resistance. Each of these key properties was *de rigueur* for particular uses. And when you got down to the nitty-gritty of food packaging – the biggest hurdle was impermeability. That meant barrier properties – a resin that held out (or in) air oxygen, water vapor and selected gases. (Because even now some dried meats, cheese, and ground coffee are packed blanketed with nitrogen).

Assess The Building Blocks

In examining the functional building blocks for making big resin molecules, each moiety offers tantalizing virtues and frustrating vices. Take styrene. Long chains make beautiful, water-clear polystyrene resins – easily molded or extruded as sheets and thereafter vacuum formed. But the barrier properties (except for keeping out particulate dirt) are poor. In fact, Monsanto resin makers actually produce a crystal-clear polystyrene film that breathes, making it tough by molecular orientation, and supplies it for fresh-flower wrap. It's so permeable that the CO_2 consumed by foliage inside the wrap comes in from outside the wrap to be constantly renewed.

Barrier anomalies abound in commercial plastics. Take the Permeability Factor as a basis for comparison. This is a measure of the amount of gas, or water vapor, that passes through the barrier the particular polymer presents. Gas permeability is measured in terms of c.c. per mil thickness/100 sq. ins./day per atmosphere. Water vapor permeability is expressed in *grams* per mil at 100°F.

	Permeability Factor (Median Values)		
Commercial Polymer	Oxygen	Carbon Dioxide	Water Vapor
Polystyrene	416	1,250	13
Styrene-Acrylonitrile (Copolymer (25% AN)	70	280	4.15
Polyethylene (High Density)	110	300	0.5
Polypropylene	150	450	0.5
(PET) Polyethylene terephthalate	7.00	30.5	0.77
Polyvinyl alcohol	0.002	0.005	1,200

These figures show that normally a polymer capable of passing the two gases like a sieve can still be a good water vapor barrier. Conversely, polymers that let water vapor come and go freely set up a firm no-no to the gases. The few exceptions, such as polyvinylidene chloride, form the select group of high barrier polymers.

Nature decreed the barrier performance of homopolymers. But Monsanto researchers reasoned that certain moieties that make up big homopolymer molecules when married to other moieties in varying proportions could have a profound effect on barrier properties. Particularly when laced-in to make a new resin molecule in specific molecular positions. What especially caught their fancy was the enormously better gas barrier properties that high AN acrylonitrile-styrene polymers possessed. With 55-90% AN content, the gas stoppage was far higher than with polystyrene. Moreover, the AN-loaded resin had unexpectedly good water vapor barrier properties. Doubtless the acrylonitrile moiety was responsible, yet pure polyacrylonitrile did not measure up as a barrier resin on several counts, processability being one.

Built-in Gas Stoppers

An ingenious research method later confirmed the early discovery that high nitrile polymers would indeed be the best candidates as barrier resins for critical packaging applications. This method involved determining the permachor values for a spectrum of resin making moieties. The screening method Monsanto researchers developed was based upon determining a numerical rating for a variety of specific moiety substitutions in a polymer, working with the simplest polyethylene for the backbone. This rating correlated the effect that the particular substitution has upon permeability – since each moiety will exert its effect regardless of the polymer system into which it is introduced. Thus a phenyl group has a particular value, a methyl or nitrile group another, etc. In this way, the permachor value of a particular polymer could be computed by adding up all the individual permachor contributions in the polymer chain. The higher the total, the lower the permeability. Adjustments are made for positioning and chain order, to reflect the decrease in diffusion created by crystallites and chain packing. A given moiety in the backbone of an amorphous polymer would have a lower permachor value than when locked in a crystalline network.

Early on, an astute member of the Monsanto research team had discovered the optimum range of nitrile content and the most promising copolymer structure. Subsequently, a mass of laboratory data and permachor value computer read-outs confirmed the initial pencil chemistry. Bench scale resins proved the discovery was right! The end result was an M-1500 series of particularly-structured high-nitrile content styrene copolymers that had these virtues:

They turned out to be top drawer gas barriers.

They proved to be almost the best water vapor barriers to boot.

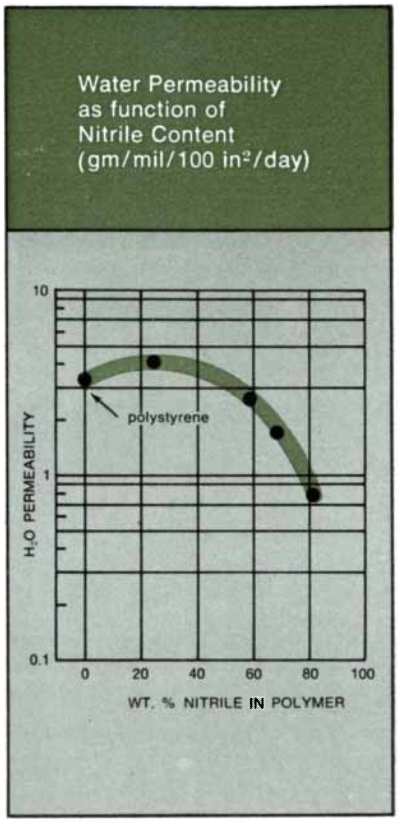
Equally important for food packaging, they are transparent, extrudable as film or sheet, blow-moldable and Monsanto developed a process for making them in food-grade purity. (FDA clearance is now in the works.)

Copolymers With New Properties

Most intriguing, the oxygen permeability of these acrylonitrile-styrene copolymers is *not* adversely affected by contact with water. Other polymers, such as cellophane and polyvinyl alcohol, allowed enormously greater penetration of oxygen when wet than when dry. Consequently, many packaging uses require "coated" or laminated materials before the barriers are capable of stopping oxygen ingress. M-1500 series resins can do it on their own.

Another anomaly – the water vapor barrier properties did not materially drop off as impermeability to gas increased. This phenomena was contrary to what might have been expected from low acrylonitrile/styrene polymers compared with polystyrene. Actually, it was found that increasing amounts of acrylonitrile did increase water vapor permeability – up to a point. Then above 40% AN, a drastic inversion occurs due to polymer chain packing and resistance of the chain to open into diffusion paths.

These new oxygen-barrier polymers can now widen the horizon for packaging – from clear film wrapping, to blister packs, to molded containers. They are the most effective plastics to keep atmosphere oxygen out and the products' freshness-moisture inside and intact. With a variety of such gas barrier resins to choose from, it remains to be seen which does the best job with meats, cheese, coffee or the rest of the gamut of flavor-sensitive foods.



For specific data on M-1500 series copolymers, request a copy of:
"Barrier Resins – Specialty Copolymers."

Address inquiries to:
Monsanto Company
Department E3SB
800 N. Lindbergh Blvd.
St. Louis, Mo. 63166

Monsanto
the science company.

We asked Control Data

to help us consolidate our data processing into a single, more efficient facility.*

They did it.

(and reduced our computer rental costs by $200,000 per year)

*Details on request. Call collect (612) 853-7600 for **Ramsey County** System Summary. Or write Control Data Corporation, Box 1980, Airport Station, Minneapolis, MN 55111.

CONTROL DATA CORPORATION

LETTERS

Sirs:

Scientific American's scholarly standing is so high that it is no pleasure to call into question the title of Stillman Drake's article in your April issue, "Galileo and the First Mechanical Computing Device." Professor Drake describes a species of mechanical instrument introduced by Galileo for doing certain geometry problems. In point of historical fact the use of mechanical computing devices in geometry was well known to mathematicians of the ancient world, so that the term "first mechanical computing device" is highly misleading.

Plutarch, in his life of Marcellus, when referring to machines designed and contrived by Archimedes "not as matters of any importance but as mere amusements in geometry," continues:

"Eudoxus and Archytas had been the first originators of this far-famed and highly prized art of mechanics, which they employed as an elegant illustration of geometric truths, and as means of sustaining experimentally, to the satisfaction of the senses, conclusions too intricate for proof by words and diagrams. As for example, to solve the problem, so often required in constructing geometric figures, given the two extremes, to find the two mean lines of a proportion, both these mathematicians had recourse to the aid of instruments, adapting to their purpose certain curves and sections of lines."

A topical journal of science can hardly avoid giving prominence, once in a while, to erroneous statements. But it is surely stretching editorial license to misassign a priority by a margin of more than 1,500 years, particularly since Professor Drake in his excellent article makes no such sweeping claim.

DONALD MICHIE

Professor of Machine
Intelligence
University of Edinburgh
Visiting Professor
of Computer Science
University of Illinois
Urbana

Sirs:

Mine is the full responsibility for the title of the recent article that Professor Michie has characterized as highly misleading. My intention was to establish priority of the sector not over everything before it but over the slide rule and the calculating machine familiar to everyone today. I mentioned the abacus at the outset, to indicate the antiquity of calculating devices as such. I did not mention instruments limited to special purposes, such as those for drawing mechanical curves, the mesolabe and the astrolabe. In my article I distinguished elevation gauges, triangulation instruments and drafting devices from computing instruments, although they certainly could be used for certain calculations. My classification was arbitrary, as I believe any classification is, and I apologize to any readers who were misled by it.

I should like to add that in describing Galileo's sector as having been introduced "for doing certain geometry problems," Professor Michie himself may mislead some readers. What was important about it was that it enabled users to solve all practical geometric problems of Galileo's day and in addition some physical and numerical problems of proportionality, square- and cube-root extraction, and computation of compound interest, for all of which operations Galileo gave instructions. His "adjoined lines" (my Scale VII) made it possible to obtain in square measure any area bounded by straight lines and circular arcs of any radius in any combination. Whatever we now think of mere Euclidean mathematics, Galileo's contemporaries had no practical geometric problems of greater generality than that. As Felix Klein emphasized, all practical mathematics is mathematics of approximation. Hence the importance of Galileo's sector to the practical men of his time is hard to exaggerate.

STILLMAN DRAKE

Institute for the History
and Philosophy of Science
and Technology
University of Toronto
Toronto

Sirs:

In your April issue there is a letter from George D. Cody commenting on H. A. Bethe's article, "The Necessity of Fission Power," in the January issue that contains references to one of my papers dealing with "net energy." Cody's calculations are at variance with the facts, and although Professor Bethe has dealt with this in his reply appearing in the same issue, I should like to discuss the matter further because of the naïve and unrealistic types of calculation along the lines suggested by Cody that are appearing in various journals.

Cody's calculations seem to be based on starting to build six generating units of 1,175 megawatts (MW) each in 1976 (a total of 7,050 megawatts) and increasing the capacity 18 percent each year to 72 units of 1,175 megawatts each in 1990 (84,400 megawatts), then stopping. That would lead to an increase of 430,000 megawatts, or a total of 520,000 megawatts, of nuclear-plant generating capacity in the year 2000. This is based on a capacity of 31,000 megawatts in operation now, 59,000 megawatts under construction and a 10-year lead time between the initiation of a plant and its going into operation.

These numbers are incorrect. The mathematical approach disregards the way in which capacity would be added, and is being added, on a realistic basis. At the end of 1975 nuclear plants with a capacity of 39,000 megawatts were licensed for operation and plants with a capacity of another 71,000 megawatts had received construction permits. The time required for design, manufacturing and construction, the time during which there would be significant dollar inputs and energy, is seven years or less. Three years or more are now required for environmental and safety reports, hearings and so on. A reasonable, practical program can be outlined for achieving the much too modest goal of 520,000 megawatts at the end of the year 2000 that would lead to the following nuclear capacities and amounts of power needed for the design, manufacturing and construction of the nuclear plants.

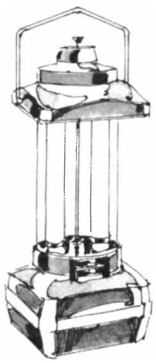
End of year	Nuclear capacity (MW)	Annual rate of additions (MW)
1976	39,000	
1980	82,000	12,500
1982	110,000	15,200
1985	161,000	18,700
1990	266,000	23,100
1995	388,000	25,700
2000	520,000	26,600

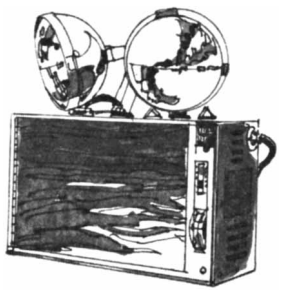
This program would match the present plant-building capacity and predicted expansion to provide about 15,800 megawatts for design, manufacturing and construction in 1976. That figure would rise to 26,600 megawatts in 1993, the year in which actual work would start on the plants for the year 2000. These figures are average annual ones, since historically the number of plants ordered and put into production each year has varied with circumstances.

The program outlined in my table would represent an annual growth rate in "starts" of a little more than 3 percent, which is quite different from the 18 percent cited by Cody. The actual plant-building capacity is substantially higher, and growth rates could also be higher. If the capacity were fully utilized, it would give more than the total of 520,000 megawatts assumed for the year 2000.

It is easy to demonstrate (as I did in the paper mentioned by Cody) that the current nuclear-energy output is greatly in excess of the energy required to build

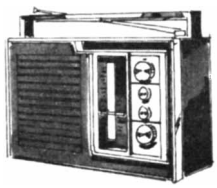
We light-up





, start-up

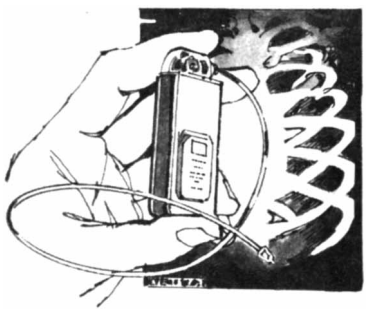
tune-in



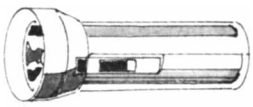
, simplify

protect

, and

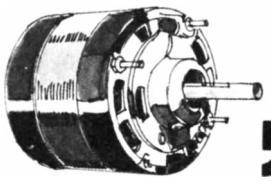


your life.



, liven-up

, turn-over



,

, amplify

,

maybe even prolong

As you can see, we do a lot. And we're into a lot of different things.

Most of what we do, however, involves the storage of energy.

In batteries.

Batteries with names that are household words: Exide, Ray-O-Vac and Willard.

But what you may not be aware of is the exhaustive amount of technology and research we put into batteries and related devices.

Today, we're working on solutions to tomorrow's energy needs. In areas as diverse as solar energy and medical electronics. Miniaturized energy units for the electronics industry and five-ton mining batteries. Molten salt batteries for large-scale energy storage. To name a few.

Bringing products like these to the marketplace involves a constant investment in energy research. Last year we spent millions of dollars on it. This year we'll spend even more.

If you'd like to learn more about our ideas in energy, write: Communications Department, ESB Incorporated, 5 Penn Center Plaza, Philadelphia, PA 19103. Or call: (215) 564-4030.

ESB INCORPORATED
An INCO Company

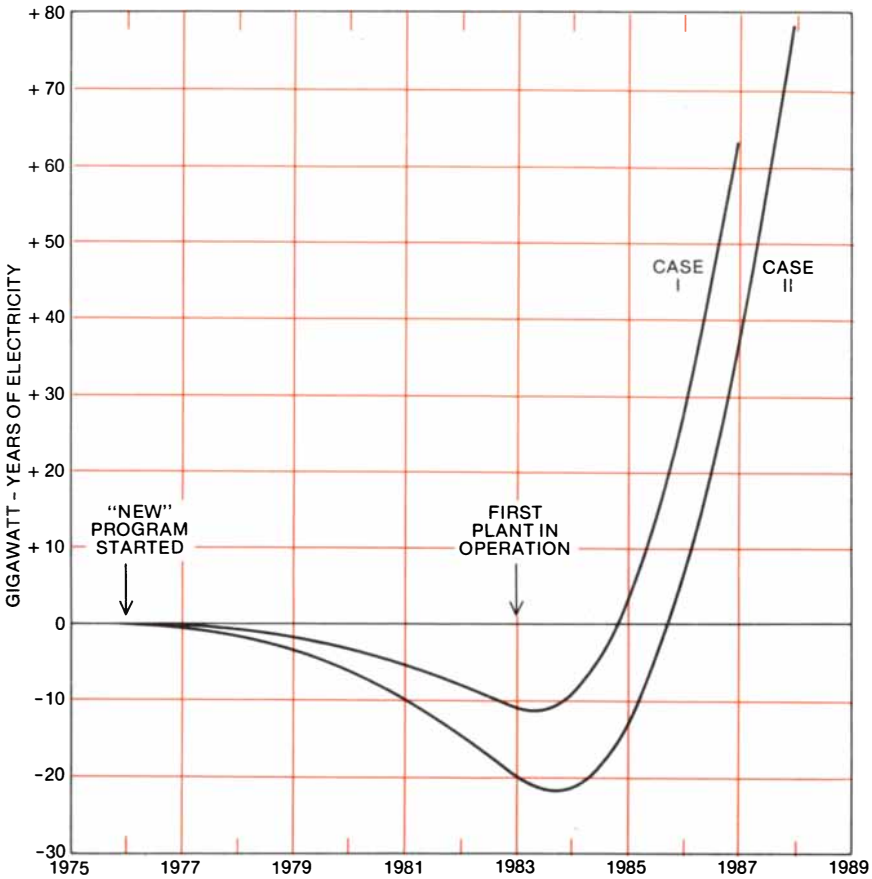
ESB

Ideas in energy.

new nuclear plants. If, however, one asks the question in the form he did, unrealistic as it is, it can still be shown that the energy required for a "new" nuclear program is paid back rapidly once the plants start producing power. (It is difficult to repay it sooner, although that seems to be one of the principal complaints!)

Assume that the "new" program is that nuclear plants will go into the process of design, manufacturing and construction starting in 1976 (and into operation starting in 1983) on a scale that will match the capacity figures cited in my table. These are the plants on which we might still have a reasonable basis for a "go" or "no-go" decision. Assume that the plants operate at an average capacity of 70 percent and that 5 percent of their output goes to the nuclear-fuel cycle (mostly for uranium enrichment). The net production factor would then be 66.5 percent for base-load operation, which should hold until the year 2000. My construction energy estimates, which are based on actual plant figures and energy ratios rather than on dollars (as the references cited by Cody are), lead to a requirement of 1,580 kilowatt-hours per kilowatt of capacity to build the plant and provide the first load of fuel.

How long does it take to return all the energy invested under these assumptions? From the curve for Case I in the accompanying chart it can be seen that the period is about 1.8 years from the time (the beginning of 1983) that the reactors from the "new" program begin operation. Thus by the end of 1984 there would be a net energy profit from this program.



The net energy balance of a nuclear-power program

If we were to reduce the assumed load factor to 65 percent and increase the fuel-cycle losses to 6 percent, and if we were also to assume that the Chapman-Price estimates for energy used in plant construction are correct (which I dispute), then we would get the energy balance shown by the curve for Case II in the chart. In this most unfavorable case it takes 2.7 years for the net balance to become positive, that is, it takes until the fall of 1985. In both cases the reactors scheduled for completion after 1990 are included in the construction energy requirement, which Cody does not do. These payout periods would look most inviting in any financial analysis. Indeed, they would be as good as one could reasonably expect from any program of energy production.

Cody also makes the statement that "the net thermal-equivalent benefit to the society would be only about 70×10^{15} B.t.u.'s" up to the year 2000. This insignificant amount of energy is equal to the entire amount consumed in the U.S. in one year today. Here again, however, the figure is in error. The figures in my calculations for the period from 1983 to the year 2000 are 189 quads (10^{15} B.t.u.'s) for Case I and 167 quads for Case II. Since the maximum debit shown in the chart is from 11 to 22 gigawatt-years, or from one to two quads, a net positive return of this scale by the year 2000 seems to be a good reason for pursuing nuclear power rather than one for opposing it.

Nuclear power cannot provide all the solutions to our growing energy problems. Along with coal and our best efforts at conservation, however, it will have to play a major role for the next generation or more. It is irresponsible to suggest that other energy sources will be developed and put into large-scale use before that time. It is equally irresponsible to claim too much for conservation, which so far has accomplished very little, even though a great deal more can and should be done. Professor Bethe has in my opinion done a public service with his factual and well-reasoned article in *Scientific American,* and I commend it to the attention of all thoughtful people.

W. KENNETH DAVIS

Vice-President
Bechtel Power Corporation
San Francisco

Scientific American, August, 1976; Vol. 235, No. 2. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

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

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 30 cents.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name

New Address

Old Address

8 sound reasons to buy our new receiver. Plus its sound.

Sony's new, more powerful STR-6800SD receiver should get a warm reception. Because it not only looks different from other receivers, it *is* different.

It has some features found in more expensive separate components—and other features found nowhere else at all.

1. The most-used controls all in one place. The level control, muting switch, tuning knob and input and tape selectors are all in the upper right-hand corner.

2. A dial pointer that doubles in length when it's close to a station. Together with the signal strength meter and the center channel meter, this Sony exclusive helps you tune more accurately.

3. A stepped level control to keep both channels equal. It guarantees unprecedented accuracy—to within ½ db over the whole volume range.

4. MOS FET front end electronics unitized tuning. Because it's unitized, the receiver tunes the same whether it's cold or warmed up. And MOS FET gives it a very wide dynamic range.

5. Dolby noise reduction system. So you can benefit from Dolby broadcasting. Instead of being an extra, it's built in—operated from the front panel.

6. Phase locked loop. It gives you better stereo separation and less distortion.

7. LEC (low emitter concentration) transistor. This Sony exclusive in the preamp phono stage yields tight RIAA equalization, low noise, low distortion and a wide dynamic range.

8. Sony's most powerful receiver. It delivers 80 watts minimum RMS continuous power per channel at 8 ohms from 20 Hz to 20,000 Hz with no more than 0.15% total harmonic distortion. It has a direct-coupled power amplifier with true complementary symmetry output stages.

And more. To these specifications (remember, we state them conservatively), add Sony's proven reliability. And you get a receiver that produces a sound that'll make you understand why you have ears.

That's the STR-6800SD at $600. Or, for less power and a few less features—but no loss of fidelity—the STR-5800SD at $500 and the STR-4800 at $400 (all suggested retail prices).

A sound investment.

SONY

How to cut startup time, downtime and overtime.
These three men count on Digital around the clock.

"I've bought hundreds of computers from a dozen manufacturers. Digital's the best by far."

LFE is a high-technology company breaking new ground across a range of critical automatic control frontiers—industrial processes, traffic control, and supervisory control including

"We wanted a company that would take the time to work with us. That's why we called on Digital."

American Systems, Inc. of Watertown, Mass. offers companies a complete teleprocessing system, with special emphasis on large-vocabulary voice-response capability.

Mr. Marvin Heit, Vice President of ASI, explains the logic behind going with Digital. "Our company is young, yet doubling in size every year, and that means we can use all the help we can get. Digital gave it to us—the equipment, the technical expertise, to get it started and the field engineers to keep it going."

Currently ASI specializes in

"We teamed up with Digital, and saved $60,000 a year in overtime alone."

Since 1968, Town Talk Baking Company of San Diego, California, has tripled their product line and increased their sales staff by 40%.

Today, they offer customers in and around San Diego over 300 different products and annual sales are $11 million.

Mr. Frank Fornaca, Vice President, explains how Digital became involved. "We were faced with a situation of increased production demands and complicated route scheduling. And it became clear to us that we needed

Saving managers millions.

Low cost computers that pay for themselves over and over again is what has made Digital one of the two computer companies in the world to install over 50,000 computers.

And what this means to you is that right at the time you need it most you have access to an enormous reservoir of experience and ideas for cutting costs while getting more done.

Digital is the pioneer in minicomputers, the leader in the development of small- to large-scale timesharing systems, and the company that has made many of the major breakthroughs in computerizing techniques.

Because of Digital, interactive computing, computer hierarchies and networks, data communications systems and new uses of computer systems are growing at an unprecedented rate.

And every year Digital places a remarkably comprehensive selection of computer tools within the financial reach of thousands and thousands of new users.

Today anyone who has a good applications idea for a computer can afford one. And just as important, they can get the help they

gigantic hydro-electric systems.

You'll find their equipment in heating and air conditioning systems, maximizing yield and profits in a range of chemical processes, controlling pipelines, indicating radiation levels, flood control—anywhere a high level of control is vital.

Their systems obviously have to be as fail-safe as man can make them and that includes the computers that run them.

"The essential features we look for are reliability and long product life," says Herb Nyser, Corporate Vice President for Engineering and Development. LFE now has over 100 Digital computers at various sites. It has 20 powerful PDP-11/35's on the TVA project alone, monitoring and controlling substations and hydro-electric facilities, reporting status and alarm conditions, closing circuit breakers and performing many other functions.

"Our trademark is excellence but to maintain that reputation we can't cut corners. We go with the best computers—the best companies. That means Digital." ■

---

setting up a complete teleprocessing system using CRT's, teleprinters, and voice-response devices for financial institutions. And they've just installed the first 3 sections of a statewide teleprocessing network for a large New York bank. "When the system is complete, 150 bank branches will have a complete teleprocessing communications capability. For example, a teller from one bank will be able to dial a computer to check the credit on an out-of-town customer and receive the answer in English."

What does the future hold for teleprocessing and ASI? "You'll be able to call up your bank right from your own home, and tell the computer what you want done. We also see a need for voice-response capabilities in insurance, manufacturing, retailing and the distribution industries. As we continue to expand, we're going to keep on working with Digital." ■

---

an automated data processing system."

Town Talk chose a DEC DATASYSTEM 535 computer. It handles all their order processing, route accounting, accounts receivable, production planning, accounts payable and general ledger.

"Thanks to this system, we've been able to handle a tremendous increase in workload, and also cut our bad debt losses by 50% and improve our cash flow situation. In fact, I would say we have from $100,000 to $300,000 more investment and operating capital than we would have had without our computer system. Our DATASYSTEM also reduced the time our route salesmen spent doing their paperwork, saving $60,000 a year in overtime costs."

Mr. Fornaca sums up Digital's role this way. "Digital has given us the basis for more sound business management, so that now we can operate more profitably even in a highly competitive market." ■

---

need from a company with over 4000 field representatives in 25 countries around the world.

To find out how Digital can help you save money, send for your free copy of "50,000 Computers Saving Managers Millions." Fill out the coupon below, or call or write Digital Equipment Corporation, Maynard, MA (617) 897-5111, Ext. 4089. European headquarters: 81 route de l'Aire, 1211 Geneva 26. Tel: 42 79 50. In Canada: Digital Equipment of Canada, Ltd.

Please send me your free book "50,000 Computers Saving Managers Millions" to show me how I can get more work done at lower cost.
My computer application is _____

Name _____ Title _____

Company _____ Address _____

City _____ State _____ Zip _____

SA 86

digital

50,000 computers saving managers millions.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1926: "When President Coolidge appointed the Federal Oil Conservation Board, he wrote a letter that showed what a thorough grasp he had obtained of the fundamental facts governing the demand and supply of oil. The letter opens by stating, 'It is evident that the present methods of capturing oil deposits is wasteful to an alarming degree.' In another paragraph he says that he is advised that our current oil supply is kept up only by drilling many thousands of new wells each year, and that 'the failure to bring in producing wells for a period of two years would slow down the wheels of industry and bring about a serious industrial depression.' Julian D. Sears of the United States Geological Survey reminds us that it required 41 years and four months to produce the first billion barrels of oil, that it took only eight years and one month to bring in another billion barrels and that only one year and seven months was required for the seventh billion to be brought to the surface. In the presence of this rate of increase anyone who tells the American people that the cistern has no bottom is doing the country a great disservice."

"To make a trip to the West Coast 'by the Canal' is one of the most popular excursions of the day, and justly so. Not only is the Panama Canal one of the most important links in the great trade routes of the world but also it is one of the most majestic and inspiring works of construction ever achieved by the hand of man. The vast sums of money invested on the Canal were spent partly for national defense and partly for commercial purposes. At the end of 10 years of operation, on July 1, 1924, the excess of total earnings over expenses stood at $33,241,425. The rapid increase of traffic passing through the Canal during the few years preceding and including 1924 was due largely to the great discoveries of fruitful oil fields in California and the large number of tankers that conveyed the oil from the Pacific to the Atlantic."

"The chemist is fast becoming master in his own house. Everywhere about us are manifestations of the part he has played in making this world a safer, saner and better place in which to live. Let us take, for example, but one achievement of the chemist—the manufacture of artificial silk, now called rayon, in which he has triumphed over nature. It was not until 1891 that artificial silk was produced on a commercial scale. The industry languished for a while but not so the chemist. He was ever on the alert, always hunting and searching for some way, either chemical or mechanical, to improve his product, and now, after 35 years of painstaking research, his efforts have been crowned with success and the world's output of rayon is greater than that of natural silk."

"A six-tube neutrodyne circuit designed to operate directly on the alternating current of house lighting mains, dispensing with all batteries, has been introduced by the Garod Radio Corporation. No hum from the alternating current could be noticed in the initial demonstration, and when asked regarding the life of the tubes, B. F. Miessner, the engineer who designed the receiver, said that less current is consumed by the tubes from the lighting circuit than is ordinarily consumed by tubes from batteries, and for that reason longer life could be expected."

SCIENTIFIC AMERICAN

AUGUST, 1876: "The first century of the United States closed last month. It has been a century of development without parallel in history. The population has increased from 2,750,000 to 44,675,000. The area has been extended from 800,000 to 3,603,844 square miles. The development of agriculture under the pressure of immigration and the stimulus of mechanical invention has been utterly without precedent. The value of manufactures has advanced from $20,000,000 to $4,200,000,000."

"The reasons for and against making the metric weights and measures the only legal standards in this country are thoroughly canvassed in the majority and minority reports of the committee of the Franklin Institute appointed to consider the question at the request of the Boston Society of Civil Engineers. The majority report, submitted by Messrs. Coleman Sellers and W. P. Tatham, urgently opposes the change, believing that the possible benefit to be reaped from it would not make up for the damages done during the transition. The report of the minority of the committee, Mr. Robert Briggs, is less an argument than a vigorous protest against the positions taken by the majority as being untrue, irrational or absurd. Much better than any protest against the statements of the majority of the committee would have been an array of facts showing that the metric system had been adopted by countries other than France without the evil results predicted."

"The Pennsylvania Transportation Company has been chartered by the Commonwealth of Pennsylvania for the purpose of transporting oil from the oil regions to the principal Atlantic seaport cities. The plan proposed is to run the oil through a four-inch pipe laid on the surface. The first objective point or terminus will be Baltimore, as being the most feasible and direct route for the pipes from western Pennsylvania. The oil will be distributed from the pipes into immense reservoirs, with refining establishments adjacent."

"Is the universe composed entirely of hydrogen? There are many eminent chemists, Professor Cooke among them, who believe that instead of there being 64 elements there is but one. What force we shall employ to dissociate the elements and convert them into that primitive form we are at a loss to say as yet, but the spectroscope leads us to think that heat, if sufficiently intense, may accomplish it."

"It is now 20 years since the French chemist Deville demonstrated the possibility of producing the metal aluminium upon a large scale, but the extent to which this praiseworthy invention has been utilized has only to a slight degree realized the sanguine expectations that intelligent minds of all professions have associated with it. In the manufacture of jewelry and other articles of luxury aluminium may in time play an important part. For the manufacture of philosophical and engineering instruments, especially the latter, the indifference of aluminium to atmospheric influences and its extreme lightness have met with general recognition. For household utensils aluminium has been highly recommended, but for this purpose it has been opposed by the despotic rule of custom or fashion. At present there are four aluminium works in existence, of which three are in France and one is in England. Their total production amounts to about 3,500 lbs. yearly. The market value of aluminium has averaged about $15 per lb., and has for some years been stationary at that figure."

"The record of boiler explosions kept in this office since our last annual report shows 139, by which 191 persons were killed and 267 injured. The principal kinds of boilers involved in the explosions are as follows: railroad locomotives (23 explosions, 40 killed, 51 injured), steamboats and tugs (18 explosions, 30 killed, 31 injured), saw mills (17 explosions, 34 killed, 52 injured) and threshing machines (nine explosions, 28 killed, 10 injured)."

SHOULD A CAR WITH A REPUTATION FOR BEING SO SAFE GO SO FAST?

Over the years, Volvo has become the very symbol of the safe, sane automobile, designed for people with a rational view of life.

But anyone who slides behind the wheel of a 1976 Volvo 240 may discover it's something more.

As *Road Test* magazine has put it: "This is one fun car to drive."

This year, Volvo has introduced a new fuel-injected, overhead cam 4-cylinder engine. It has extremely fast pickup in the 20-55 m.p.h. range where most serious driving is done.

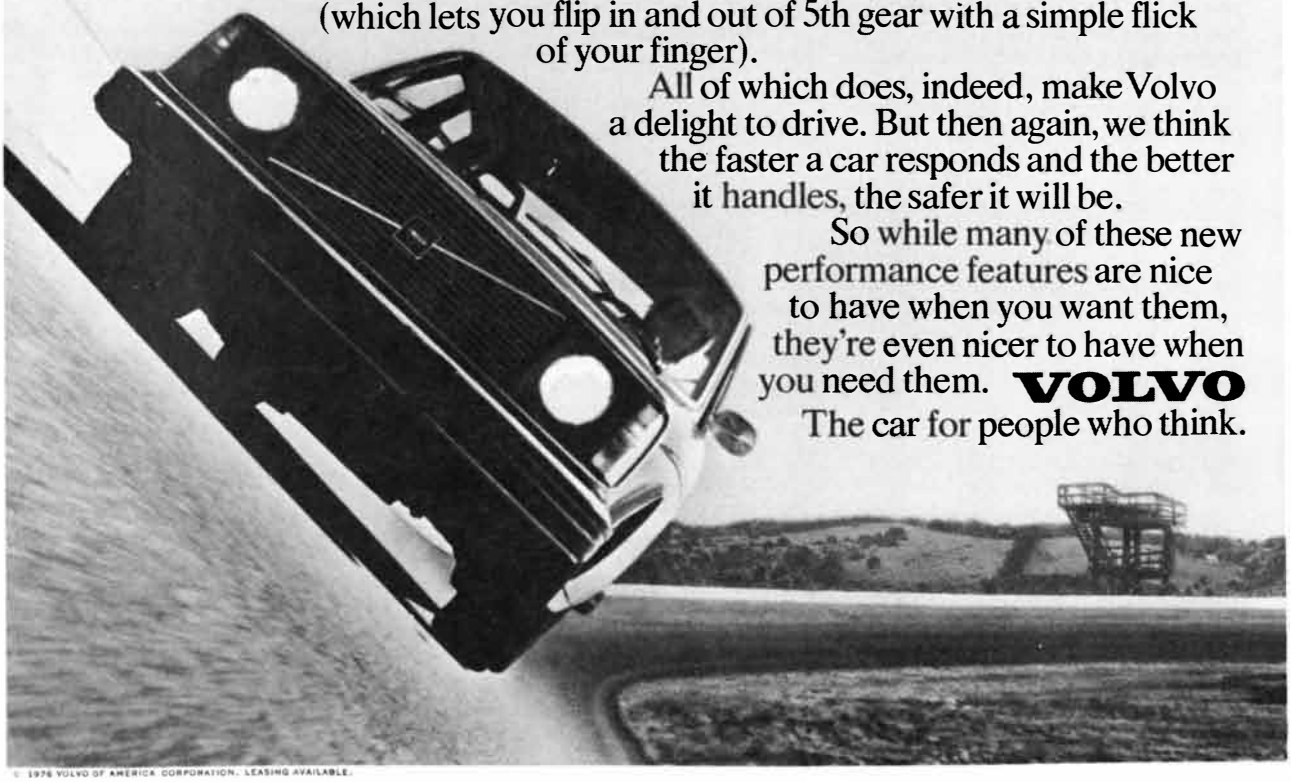
In a comparison of passing times, a Volvo 242 with a 4-cylinder engine was faster than a Mercedes 280 with a six.

Volvo also gives you rack and pinion steering to help you take life's curves. And a spring-strut front suspension designed to keep the car steady and level even if you take them fast. You get 4-wheel power disc brakes. And you can order a 4-speed manual transmission with electrically-operated overdrive (which lets you flip in and out of 5th gear with a simple flick of your finger).

All of which does, indeed, make Volvo a delight to drive. But then again, we think the faster a car responds and the better it handles, the safer it will be.

So while many of these new performance features are nice to have when you want them, they're even nicer to have when you need them. **VOLVO**

The car for people who think.



THE AUTHORS

DAVID S. RUBSAMEN ("Medical Malpractice") is a physician and attorney who specializes in the medical aspects of litigation. He is also editor and publisher of *Professional Liability Newsletter,* a monthly periodical of topics related to medical malpractice, and a consultant to the Doctors' Company, a new doctor-owned malpractice-insurance carrier in California. Rubsamen received his B.A., M.D. and LL.B. degrees from Stanford University in 1945, 1948 and 1959 respectively. He practiced internal medicine until 1962, when he set out on his present career. A past chairman of the California Medical Association's Legal Liaison Committee to the California Bar Association and a current member of the board of governors of the American College of Legal Medicine, he has served as a medical-legal consultant to both the executive and the legislative branches of the Federal Government.

JOHN R. PAPPENHEIMER ("The Sleep Factor") is George Higginson Professor of Physiology at the Harvard Medical School. He is a graduate of Harvard University and obtained his Ph.D. from the University of Cambridge. A member of the Harvard Medical School faculty since 1946, he was selected as a career investigator of the American Heart Association in 1953. His interest in the physiology of sleep, he writes, "started by chance in 1965 when I was browsing through some old books and came across Piéron's monograph of 1913. I realized at once that our group at Harvard had the techniques to confirm or disprove the Piéron phenomenon and to extend his observations if they were correct." Pappenheimer spent the academic year 1975–1976 on sabbatical as Eastman Professor at the University of Oxford; in his absence the work at Harvard on the purification of the sleep factor was carried on principally by Manfred L. Karnovsky, Greg E. Koski and James M. Krueger.

PHILIP E. L. SMITH ("Stone-Age Man on the Nile") is professor of anthropology at the University of Montreal. A native of Newfoundland, he studied history and economics at Acadia University in Nova Scotia. Later, "after several misguided years in the business world," he entered Harvard University to study prehistoric archaeology, receiving his Ph.D. in 1962. He also did graduate work at the University of Bordeaux in France. He taught anthropology at the University of Toronto for several years before moving in 1966 to Montreal, where, he writes, "I find that lecturing in French helps to filter out some of the jargon endemic in English-language archaeology today." An active and wide-ranging fieldworker, he has done archaeological excavating in the U.S., Mexico, Iraq, France, Egypt and Iran. He is married to Fumiko Ikawa-Smith, whose specialty is Japanese archaeology and who currently heads the anthropology department of McGill University.

KEVIN C. BURKE and J. TUZO WILSON ("Hot Spots on the Earth's Surface") are geophysicists with a common interest in plate tectonics. Burke, who is chairman of the department of geological sciences at the State University of New York at Albany, obtained his Ph.D. from University College London in 1953. He has done geological fieldwork in 27 countries. In the 1950's he taught at the University of Ghana and later did research for the British Geological Survey on nuclear raw materials. From 1961 to 1965 he was involved in establishing a geology department at the University of the West Indies in Jamaica, and from 1965 to 1971 he taught at the University of Ibadan in Nigeria. Wilson, who played a key role in the revival of the theory of continental drift first put forward by Alfred L. Wegener in 1912, is currently director general of the Ontario Science Centre. His degrees include a B.A. from the University of Toronto, a B.A. and an M.A. from the University of Cambridge and a Ph.D. from Princeton University. Before taking up his present post he was for many years professor of geophysics at Toronto, where he also served as director of the Institute of Earth Sciences.

CHARLES LANE ("Rabbit Hemoglobin from Frog Eggs") is a biologist on the staff of the Medical Research Council Laboratories at Mill Hill in London. He has also done research at the University of Oxford (where he received his Ph.D. in zoology in 1971), the University of Cambridge and the Massachusetts Institute of Technology. His training regimen for his current series of microinjection experiments, he reports, includes "slalom racing, tennis, squash and swimming." Lane is a member of the scientific and banking Rothschild family, which was most recently represented in the pages of SCIENTIFIC AMERICAN by a contribution from his mother, Miriam Rothschild, who wrote (with Y. Schlein, K. Parker, C. Neville and S. Sternberg) "The Flying Leap of the Flea" in the issue of November, 1973.

WILLIAM H. PRICE ("The Photographic Lens") is manager of optical engineering for the apparatus division of the Eastman Kodak Company, where he began his professional career in 1950, shortly after his graduation from Middlebury College with a B.A. in physics. He later acquired an M.S. in optics from the University of Rochester. Before taking on his present assignment he worked for Kodak in the positions of quality control engineer, optical designer, optical design supervisor and head of optical technology. He holds 10 patents in the field of optics and has written several technical papers on subjects related to the design, specification and manufacture of lens systems.

LORUS J. MILNE and MARGERY MILNE ("The Social Behavior of Burying Beetles") are behavioral ecologists at the University of New Hampshire. Both members of this husband-and-wife research team hold Ph.D.'s in biology, from Harvard University and Radcliffe College respectively. They began their investigation of the behavior of burying beetles a few years ago when Lorus Milne, a Canadian by birth, encountered some of the insects outside a summer cottage in Haliburton County in Ontario. There mice trapped in the cottage vanished mysteriously overnight when they were left outside on the woodland floor. Tying a string to the dead mouse did no good; the string was cut regularly. A fine wire tether kept the burying beetles from concealing their prize so quickly, however, and their secret was discovered. Later the Milnes recorded their activities on 16-millimeter color motion-picture film and began to accumulate a growing file on these curious insects.

J. J. CALLAHAN ("The Curvature of Space in a Finite Universe") is associate professor of mathematics at Smith College. A 1962 graduate of Marist College, he went on to obtain his Ph.D. in mathematics from New York University. He was a Benjamin Peirce lecturer at Harvard University for three years before joining the faculty at Smith in 1970. His primary mathematical interests, he notes, are "differential analysis and catastrophe theory and its applications. The themes of this article—geometry and the history of mathematics—are things I am interested in teaching, but I am not a specialist in them." The article developed, he adds, "out of an attempt to explain Einstein's concept of a finite but unbounded space to my nonscientific colleagues at Smith. They found it tough going, and some simply dismissed a finite universe as impossible, because Kant had done so when he studied the question 300 years ago. A sabbatical this past year at the University of Warwick gave me a chance to read what Kant said about space. He does indeed epitomize the commonsense view, which Einstein (and Riemann before him) shattered. I was unable, however, to find any satisfactory explanation of just how the old and the new ideas fit together, so I attempted one myself."

"The most complete and most scholarly dictionary of the English language"

—*The Christian Science Monitor*

As an introduction to membership in the

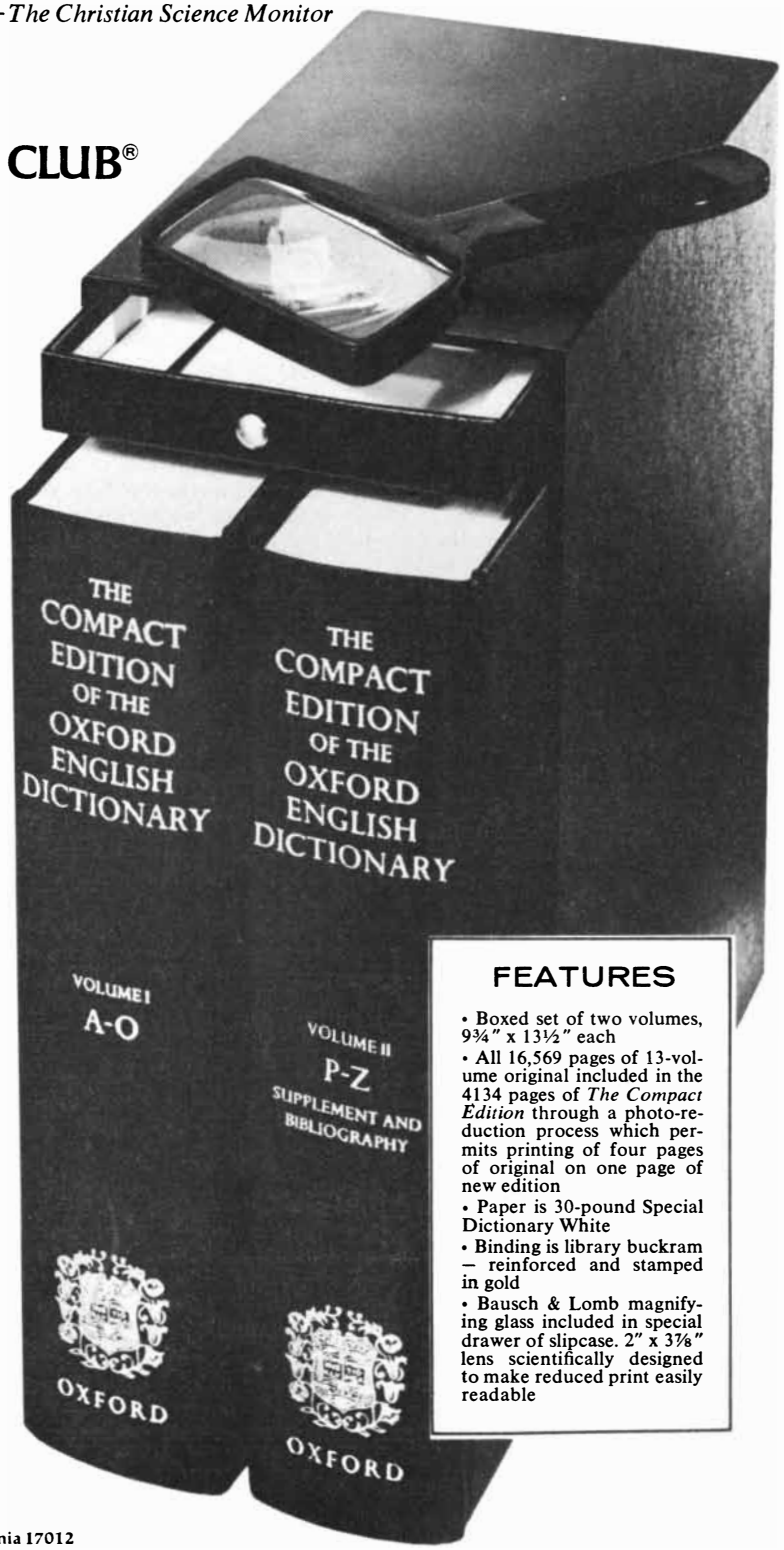
BOOK-OF-THE-MONTH CLUB®

yours for only

$17.50

[Publisher's list price $95]

THE SUGGESTED TRIAL: You simply agree to buy four Club choices within a year at substantial savings on most books you choose.



THE OXFORD ENGLISH DICTIONARY is generally regarded as the final arbiter of the English language. Until recently it had been available only as a thirteen-volume set, currently priced at $395. Now, through the combination of an ingenious method of micrographic reproduction and a fine Bausch & Lomb optical lens, every single one of its 16,569 pages, fifty million words and close to two million illustrative quotations appears, in easily readable form, in *The Compact Edition.*

> The *New York Times* book critic Christopher Lehmann-Haupt has said of this edition: "It is something of a miracle.... *The Compact Edition* is easier to work with than the original with its 13 separate volumes."

Even more extraordinary, as a trial member of the Club you may obtain the two-volume set for only $17.50. And as long as you remain a member, you will receive the *Book-of-the-Month Club News*,® a literary magazine announcing the coming Selection and describing other important books, known as Alternates, most of which are available at substantial discounts — up to **40%** on more expensive volumes.

If you continue after your trial membership, you will earn at least one Book-Dividend® Credit for every Selection or Alternate you buy. These Credits enable you to acquire for your library a wide variety of books, called Book-Dividends, at astonishing savings — *at least 70% of publishers' list prices.*

All Book-of-the-Month Club books, including Book-Dividends, are identical to the publishers' editions in content, format, size and quality.

FEATURES

- Boxed set of two volumes, 9¾" x 13½" each
- All 16,569 pages of 13-volume original included in the 4134 pages of *The Compact Edition* through a photo-reduction process which permits printing of four pages of original on one page of new edition
- Paper is 30-pound Special Dictionary White
- Binding is library buckram — reinforced and stamped in gold
- Bausch & Lomb magnifying glass included in special drawer of slipcase. 2" x 3⅜" lens scientifically designed to make reduced print easily readable

BOOK-OF-THE-MONTH CLUB, INC., Camp Hill, Pennsylvania 17012

Established 1845 SCIENTIFIC AMERICAN August 1976 Volume 235 Number 2

Medical Malpractice

Medical care has improved so much that patients expect more of it and are quicker to sue when the outcome is unsatisfactory. Awards have risen sharply, precipitating a crisis that remains unresolved

by David S. Rubsamen

In 1972, according to a Federal commission studying medical malpractice, the cost to the average patient of malpractice-insurance charges passed on to him by physicians and hospitals was from 20 to 50 cents of every $10 he paid to a physician and about 52 cents for every day he spent in a hospital. It has been estimated that by 1980, if the present trend continues, the charges will have increased tenfold. The figures dramatize what is widely known as "the medical malpractice crisis," which paradoxically arises from the great progress of medical science over the past three decades and the resulting public expectation that medicine should have a cure for every ill. Now errors of omission are brought into court as well as errors of commission, and juries tend to make big awards. My experience as a consultant on the medical aspects of litigation has led me to the conclusion that the solution of the malpractice crisis requires two things: (1) better control by the medical profession of the quality of medical practice and (2) a rational system of compensating negligently injured patients.

Until quite recently a complete recovery from an illness was rarely a certainty, and the physician's management was seldom questioned after an unsatisfactory outcome. For example, a failure to diagnose appendicitis, followed by perforation and death from peritonitis, was regarded as one of life's hazards. The timely detection of cancer was the exception rather than the rule for many years after the development of surgical techniques that made it possible to cure a malignancy discovered early.

Even the best care in those times was frequently ineffective. Parents who raised children before the discovery of antibiotics can recall the feeling of helplessness evoked by a diagnosis of pneumonia or infection of the middle ear. It was not until about 1945 that medicine recognized the role of an imbalance of electrolytes in the fatal outcome of a variety of acute disease states. Such techniques as electronic monitoring, easily available studies of blood gases and the rapid determination of serum electrolytes—all indispensable in the effective treatment of many severe illnesses today—are developments of the period since World War II. It is only a slight exaggeration to say that the era of scientific medicine began about 30 years ago.

Improved medical technology disclosed previously unsuspected pathophysiological changes in serious illness and trauma. Therapies were developed to deal with many such changes. As a result the prognosis for a large number of sick people improved sharply. The proportion of patients of whom physicians had to conclude that "there's nothing we can do here" decreased dramatically; most of them entered an arena of challenging medical problems where good results could reasonably be expected. (This prospect was much less likely for the chronically ill patient, for whom prevention was and still is the only real solution.)

Naturally the expectations of patients have matched the progress of medical science. More important, in terms of malpractice litigation, is a similar shift in attitude among physicians. This fact leads, subtly but inevitably, to the crucial issue of how medical negligence is defined.

The simplest statement of the issue is that the physician owes it to his patient to be careful. If, as a result of the physician's carelessness, the patient suffers injury, then liability exists. Like most legal formulations, however, this one is simple in theory but complex in practice. It requires closer examination.

Both carelessness and injury must be established; neither is sufficient alone. An example is provided by the case of a 55-year-old housewife in California who consulted her family physician in July, 1970, for treatment of cough and chronic fatigue. She saw him 23 times over the next six months, but her complaints continued and she lost 60 pounds. Finally, in January, 1971, the physician ordered a chest X ray, which revealed a bilateral pulmonary malignancy. The patient died a few weeks later.

Her husband sued. Three medical witnesses testified that, considering the patient's symptoms, a chest X ray was required by at least the fall of 1970. On the other hand, the attorney for the defendant physician presented two medical witnesses who testified that the cancer was multifocal in origin and therefore inoperable from the beginning. The defense made no attempt to justify the doctor's delay in accomplishing the diagnosis. The jury accepted the testimony of the defense on the origin of the illness and returned a verdict in favor of the doctor.

Sometimes a physician's carelessness may be egregious, and yet no injury results. Years ago a well-trained obstetrician was called to the hospital late at night to perform an emergency Caesarean section. He was moderately intoxicated, and his initial incision was accomplished with a sweep of the scalpel that cut down through the wall of the uterus and even through the amniotic membrane. The infant lay exposed for

extraction, and there was not a scratch on him. The physician's reckless action, let alone the fact that he was operating under the influence of alcohol, clearly identified his carelessness. There was no injury, however, and so, although the physician might have been disciplined by his county medical society or expelled from the staff of the hospital, neither the mother nor the child had a basis for a negligence suit.

Although these questions of causation (that is, the connection between carelessness and injury) are not infrequent in malpractice litigation, most of the time the central issue concerns the definition of carelessness. Many physicians believe only the most obvious errors of commission should be called to account, and many contend too that errors of omission—the "honest mistakes of judgment"—should not provide grounds for a lawsuit. There is, however, no logical reason why an error of judgment should be any less subject to a malpractice action than an overt blunder. The test remains the same: Did the physician meet his legal duty to the patient to be careful? The discussion can become abstruse at this point, turning on such matters as what the standard of care in medical practice should be and how it would be elucidated in an enlightened system. In the real world of a jury trial, however, the issue is simple: the physician has violated the standard of due care when the jury, having considered all the evidence and the judge's instructions, votes against him.

What, then, is the basis for a jury's judgment? In most cases testimony from at least one physician appearing as a witness for the patient provides the critical evidence. Sometimes the patient's attorney will obtain the necessary testimony by skillful questioning of the defending physician, and in some states the law allows the jury to infer negligence from certain types of medical accident. Ordinarily, however, the jury hears one physician or more testify for the patient and one physician or more testify for the doctor defendant and then decides what standard of care should be applied in the case and whether the doctor violated it. Thus medical testimony is usually vital to the patient's cause.

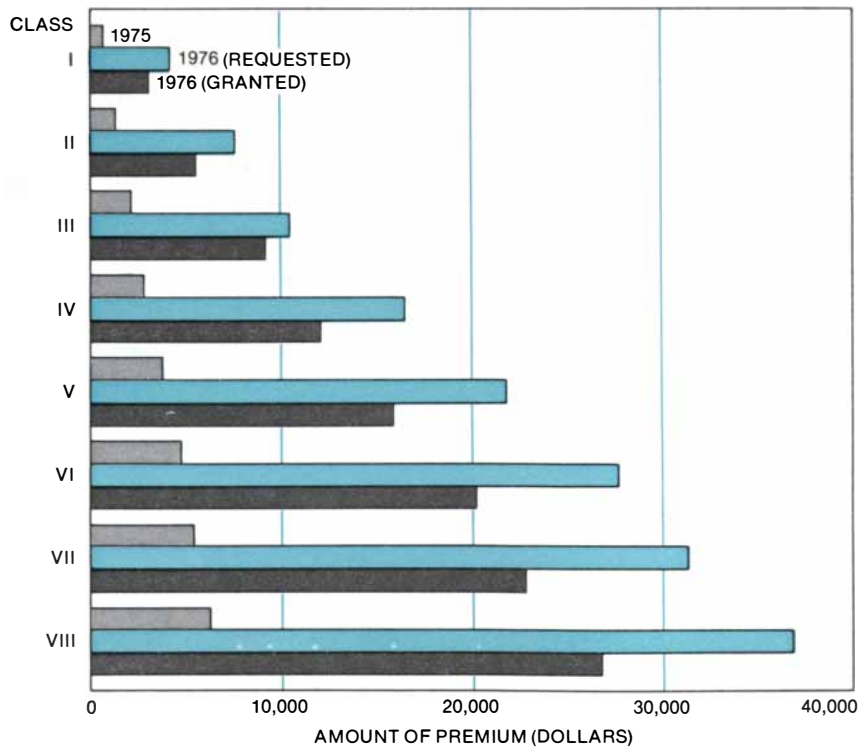
What legal principle makes the medical witness necessary? Medical-malpractice litigation involves scientific facts. The common law has long recognized that juries cannot resolve conflicting points of view on scientific matters without hearing expert witnesses.

The development of medical-malpractice litigation in the U.S. has paralleled the increasing availability of expert medical witnesses. In other words, it has depended on the willingness of at least some physicians to testify against other physicians. Thirty years ago physicians were so reluctant to testify against one another that the phrase "conspiracy of silence" had real meaning. In some states the same situation prevails today, but a majority of the states have experienced a growth in malpractice litigation as expert witnesses have become available to plaintiffs. In the states leading this trend (notably California, New York, Florida, Illinois, Pennsylvania and New Jersey) there has also been an extraordinary escalation in the size of awards.

California's experience, involving both physicians and hospitals, is illustrative. In 1969 the state had only three cases in which a verdict or an out-of-court settlement of a malpractice action amounted to $300,000 or more. The number rose steadily to about 34 in 1974. By 1975 the attitude of jurors had apparently been influenced by the publicity about California's malpractice crisis, and the figure dropped to 24. The state's first case resulting in an award of $1 million or more was in 1967; since then there have been 22 more, 14 of them after January, 1974.

What determines the accessibility of medical witnesses for patients who become plaintiffs? The absolute requirement is an expansion, usually by judicial fiat, of the traditional "locality rule." Under this rule the only expert witnesses who can testify in a malpractice action are physicians practicing in the defendant physician's community. The rule reflects a situation of more than 50 years ago, when schools of thought on medical practice were often geographically insular. Given the professional and social contacts among physicians in a community, the rule effectively suppresses malpractice litigation. Most states have abandoned the rule as the courts have come to the view that medical schools and the professional literature are national in scope. Discarding the locality rule, however, does not automatically open the door to malpractice litigation. Something else must happen.

Again California is illustrative. Although the state's Supreme Court eliminated the locality rule in 1952, it was another eight years or so before medical witnesses for plaintiffs began appearing more than rarely, and their appearance was not common in malpractice actions throughout the state un-



INCREASE IN PREMIUMS for physicians buying malpractice insurance reflects the rise in the number of malpractice lawsuits and in the size of the awards made to successful plaintiffs. The chart shows the premiums paid by various classes of physicians in southern California in 1975, the level of premiums sought for 1976 by the insurance company and the amounts granted by the state's insurance commissioner. The premiums are for coverage of $1 million in a single case and an aggregate of $3 million for two or more cases in one year. The class numbers refer to different types of practice; the highest premiums (Class VIII) apply to physicians who do specialized surgery or specialize in obstetrics and gynecology. Although the data are limited to southern California, they are indicative of what has been happening elsewhere.

til about 1969. The availability of these witnesses has brought about a striking change in the type of malpractice case litigated in California. The shift occurred between 1968 and 1972. The pattern of change is bound to appear in any state where medical witnesses for plaintiffs in malpractice actions become freely available.

I have alluded to the fact that negligent medical injury can be divided into two categories: errors of commission and errors of omission. The first category includes the slip of the surgeon's scalpel, the inadvertent administration of an injurious drug, the application of a cast so tightly that circulatory impairment results and so on. The amount of injury resulting from errors of commission is quite limited; taking all negligent medical injuries together (not just the cases that are litigated), only a small percentage are caused by errors of commission.

Errors of omission can be characterized as mistakes in judgment. If the physician had given this drug instead of that one, if his therapeutic intervention had been better timed, if his diagnostic acumen had been a bit sharper, then the unsatisfactory outcome could have been avoided.

In the majority of states malpractice litigation involves mainly errors of commission. Unless a judgmental error is gross it will not be the subject of a lawsuit. The reason is that physicians in those states still share the belief that "honest mistakes" should not be litigated. In the half-dozen or so states where this attitude no longer exists, however, the incidence of successful malpractice actions has risen sharply. The reason is that the number of unsatisfactory outcomes that might be regarded as the consequence of judgmental errors is large, provided that a high enough standard of care is applied.

Two cases in California are illustrative. Both of them entailed complex medical facts of the kind that are common when alleged errors of judgment are litigated. The first case arose in December, 1968, when a college student suffered a dislocation of his right knee while playing touch football. The popliteal artery (the major artery, lying just behind the knee, that carries blood to the lower leg and the foot) was severed by the injury. The broken ends of the artery constricted, thus preventing bleeding and the formation of a local hematoma, which would have signaled the catastrophic event.

The patient was seen soon after his injury at the emergency room of the college hospital. The physician on duty could not identify the pulsations normally found in the two branches of the popliteal artery that supply blood to the foot, and he called for an orthopedic consultation. The orthopedist saw the patient shortly thereafter and noted on the patient's chart a "probable" injury to the popliteal artery, but he concluded that there was most likely just a spasm of the artery, since he thought he could detect faint pulsations in the arteries of the foot. A general surgeon was also consulted. He wrote that the arterial pulsations were "extremely weak on the right," but he agreed with the orthopedist that the popliteal artery was probably only in spasm. "I think we can safely observe without arteriogram or exploration," he wrote.

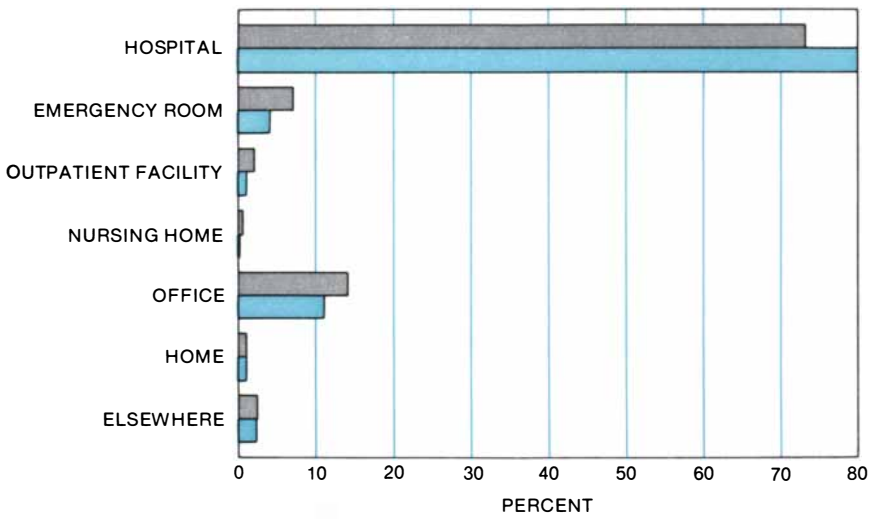
These examinations were made during the first two hours after the patient was injured. There was then a delay of about 18 hours before an arteriogram was performed. It demonstrated the obstruction of flow through the popliteal artery. The patient's leg had to be removed.

When the case was tried in October, 1971, the attorney for the plaintiff concentrated on the early medical decisions and called them substandard. He presented the testimony of a well-qualified vascular surgeon who practiced in a community 400 miles away. This expert witness concentrated on two themes in telling the jury what standard of care should be applied. First, since the popliteal artery is the main source of blood to the lower leg and the foot, an obstruction of its flow for more than about six hours creates a substantial risk of gangrene, and the risk increases with time. Second, an arteriogram and/or a surgical exploration of the area behind the knee are fairly simple procedures, and they make possible a definite diagnosis. Therefore since the defendants suspected an arterial injury but could not be certain about an obstruction, their failure to perform an arteriogram or a surgical exploration in the period shortly after the injury violated the standard of due care.

The expert medical testimony for the defense stressed the judgmental aspect of the case and the fact that the physicians who saw the patient were well-qualified specialists. It was noted that the patient did not present the typical signs of arterial interruption: severe pain in the foot accompanied by a pronounced drop in the temperature of the skin and by blanching of the skin. The judge gave the jury the usual instructions, pointing out that the mere fact of an error is not a basis for liability on the part of the physician and that the plaintiff must show that the error was a product of carelessness, that is, of substandard care. The jury nonetheless returned a verdict for the plaintiff in the amount of $475,000.

Did the jury apply too high a standard of care? Many surgeons with whom I have discussed the case feel that it did not. Yet five years earlier the case almost certainly would not have been brought to trial because the plaintiff's attorney would not have been able to find an expert witness to comment adversely on the errors of omission.

What is the limit on errors in judgment? Conceivably it could be as large as the number of untoward medical occurrences. Such a conclusion assumes, however, that juries might eventually adopt a rule of prescience for the medical profession, which would be absurd. Nevertheless, the complexity of the medical facts in most cases of judgmental error means that juries will often not be able to deliberate intelligently



PLACE OF INJURY is indicated for malpractice-insurance claims closed between July, 1975, and February, 1976, covering a total of 9,471 claim reports received by the National Association of Insurance Commissioners. The gray bars refer to the number of incidents occurring in each place and the colored bars indicate the amount of indemnity paid on the basis of those incidents. Hospitals accounted for 80 percent of incidents and 84 percent of indemnity.

about them. The second case from California illuminates the point.

In the fall of 1973 a 33-year-old woman consulted a general surgeon because she had recently felt a lump in her right breast. The lesion was somewhat tender and the skin over it showed redness. The doctor concluded that the patient had a localized infection. He put her on antibiotic therapy and told her to return in two weeks.

According to the doctor, on the second visit the inflammation appeared to have subsided. He told the patient to discontinue the antibiotic, observe the effect and return in a month. Two weeks later she telephoned to say that the lesion had flared up. The doctor told her to go back on the antibiotic and to keep the appointment she had been given.

On the third visit she reported that the inflammation had subsided until the day before and then had flared up again. The doctor told her to continue the antibiotic for two weeks and said that if it had not subsided by then, he would perform a biopsy. He was concerned about the possibility that the lesion might be an inflammatory carcinoma, an extremely rare but highly malignant breast cancer.

This time the antibiotic was ineffective, and the patient returned in 11 days. A biopsy the following day revealed an inflammatory carcinoma. The lesion measured 7.5 centimeters in diameter; on the initial examination it had been about a third that size. The pathologist who studied the tissue saw evidence of an associated bacterial infection, thus explaining the patient's initial response to antibiotic therapy.

Now the woman was treated by the removal of her adrenal glands, followed by radiotherapy and chemotherapy. (Mastectomy was not done, since it does not improve the prognosis in inflammatory carcinoma.) When the case was tried, early in 1974, there was no evidence of cancer, but all the treating physicians agreed that this response to the intensive therapy must be viewed as temporary and that the patient's prognosis for an extended remission was poor.

At the trial the patient contended that the antibiotic therapy had never been significantly effective. The medical witness who testified on the patient's behalf said the lesion should have been biopsied within two weeks after the first examination. The defending physician testified that he had told the patient from the beginning that the diagnosis was either an infection or a malignancy. He said the question of cancer was of concern to him, but he was reluctant to perform a biopsy since cutting into infected tissue virtually guaranteed a bad scar.

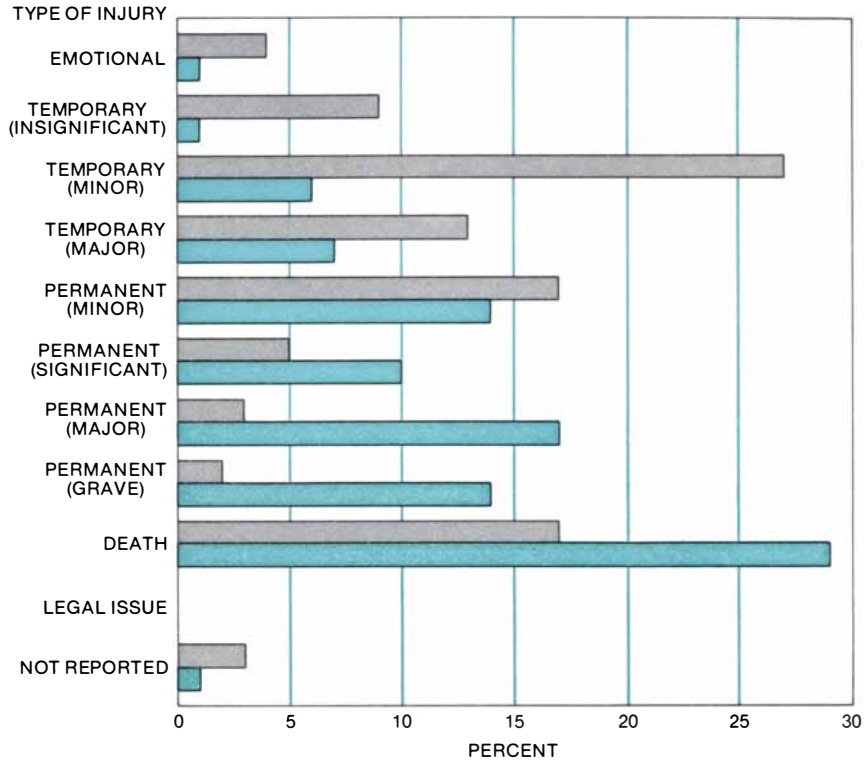
Three expert witnesses for the defendant testified that the delay of some six weeks in doing a biopsy was well within the normal standard of care. They emphasized that the treatment would have been identical even if the cancer had been identified on the patient's first visit to the physician. The plaintiff's expert witness agreed with this point but asserted that a six-week delay in treatment could make a difference in the patient's response to treatment, even though success in this context was regarded as affecting the length of her life rather than the possibility of a cure.

The jury returned a verdict of $200,000 in favor of the patient. Most observers of the trial thought the verdict was a compromise, since an award two or three times larger would have been expected for a woman of the patient's age if the jury had been convinced that the physician had been negligent. The defense filed a motion for a new trial, but the day before the judge was to rule on the motion the case was settled for $100,000.

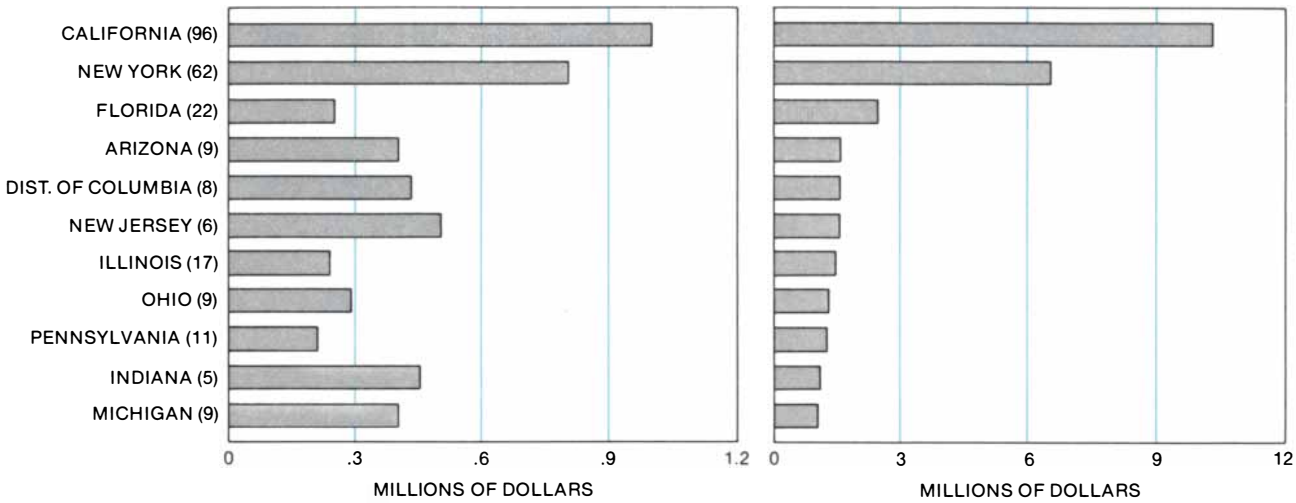
These two cases show that medical negligence is not to be equated with what is normally regarded as bad practice. From a strictly legal point of view the young man who lost a leg was a victim of negligent treatment because the plaintiff's expert said so and the jury accepted his testimony. From a medical point of view it is not unreasonable to say that the jury was right, but it is also apparent that the jurors were applying a relatively high standard of care to the conduct of the defendants.

As for the second case, it is difficult to believe the judge would not have ordered a new trial on the ground that the jury's verdict was unreasonable. The verdict emphasizes the capriciousness of juries in cases of medical malpractice and demonstrates why many authorities in medical law argue for specialty courts, where experienced judges would sit without juries. An alternative is a private system of arbitration for malpractice cases. Nevertheless, most defense attorneys prefer the jury system. The reason is that in the rural areas of all states and even in a number of metropolitan areas juries generally favor physicians. Therefore it is the quite obvious error of commission that characterizes the great majority of adverse malpractice verdicts.

Perhaps the most interesting aspect of medical-malpractice litigation is the manner in which standards of medical care can be set by juries. For example, in 1973 a jury in California returned a verdict of $1.5 million against an anesthesiologist. The plaintiff was a woman who had suffered a cardiac arrest at the conclusion of a brief elective operation. She was left with moderately severe brain damage. During the trial her attorney presented medical testimony in support



TYPE OF INJURY is indicated for the malpractice-insurance claims represented by the illustration on the opposite page. The gray bars show the percent of malpractice incidents represented by each type of injury; colored bars, the percent of indemnity paid in each category. "Legal issue" ordinarily is whether the physician obtained the informed consent of the patient.



AMOUNT OF PAYMENT for medical malpractice is listed for the 11 jurisdictions in the U.S. where the largest amounts were paid between July, 1975, and February, 1976, according to data provided to the National Association of Insurance Commissioners. Associated with the name of each area is a number in parentheses giving the number of claims of $50,000 or more paid by insurance companies on behalf of defendants in malpractice actions. The bars at the left represent the largest payment made in each jurisdiction and the bars at the right show the full amounts paid. On a national basis 311 claims involved payments of $50,000 or more; they represented only 3 percent of the total number of claims but accounted for 63 percent of the total indemnity paid for all the claims included in this survey.

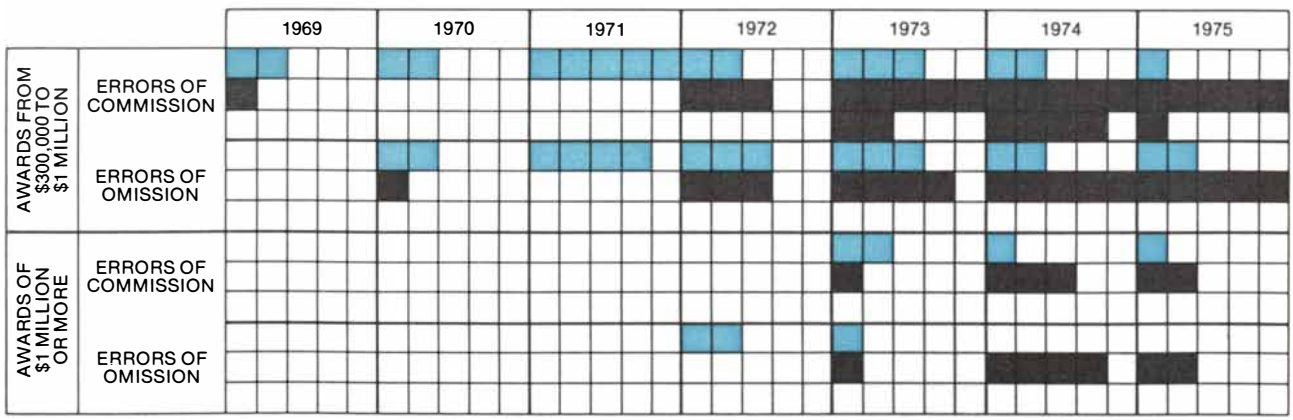
of two arguments. First, there were certain oversights in the management of the anesthesia that led to the cardiac arrest. Second, even with the arrest the patient would not have suffered significant brain damage if the anesthesiologist had been monitoring her pulse constantly, so that the arrest would have been detected the moment it occurred and resuscitative measures could have been started immediately. Interviews with jurors after the trial revealed that this second argument was central to their decision.

Does due care require an anesthesiologist to monitor the pulse continuously as a matter of routine? The defendant and three other witnesses said that this was not the standard practice in the local community in a simple case such as this one. Their standard was to take the pulse for 15 seconds every two minutes. The anesthesiologist who testified for the plaintiff asserted that the standard of due care requires constant monitoring of the pulse in every case.

Shortly after the trial I asked several anesthesiologists in teaching centers in California if constant monitoring of the pulse during all anesthesia really was the due-care standard in California. Each replied that he adhered to this standard himself and regarded it as being the preferred practice, but most of them doubted that it was widely enough observed to constitute the standard of practice among anesthesiologists generally. It does constitute the standard now, however, particularly since verdicts in three of four subsequent cases in California involving the same issue have gone against the anesthesiologists.

Against this background the origins of the malpractice crisis, which began about two years ago, are readily seen. The immediate cause was the sudden and drastic increase in the rates for malpractice insurance in most states. Malpractice insurance has become much more difficult to get in recent years; some 22 commercial carriers offered it five years ago, whereas perhaps seven offer it today. Resistance among physicians to higher premiums can lead the only company left in a state to withdraw, thereby compounding the crisis. This has already happened in Maryland and is a possibility in several other states.

One would suppose that the high level of premiums (as much as five times higher than they were in 1974) would cause the commercial companies to hurry back into the market. The main reason they do not is that they cannot be sure of making a profit. More than half of the states have statutes of limitation



RISE IN AWARDS in malpractice actions in California is charted. The chart covers awards of $300,000 or more made as a result of jury verdicts (*color*) or settlements reached out of court (*black*) and also shows whether the injury on which the malpractice award was based resulted from an error of commission by the defendant physician or hospital or was an error of omission, that is, of judgment.

that allow suits for negligent medical injury to be brought for as long as three years after the discovery of the injury (not the date when the injury occurred). The malpractice actuary will therefore not have a good idea of his losses within 12 months after a policy is issued. Indeed, it will be at least three years before he will know of even half of the suits attributable to a given policy year.

This extended "tail" of liability has two important consequences. First, the adequacy of the malpractice premiums charged for a given year cannot be judged for three or four years. Second, the length of the tail is uncertain. How many of the negligent medical injuries that occurred in 1975 will have been reported by 1978 or 1982? What will the effect of currency inflation have been by then? Will juries be more willing to decide complex cases against physicians and to award large amounts of money to injured patients?

In some areas physicians have responded to the problems of high premiums and limited availability of malpractice insurance by forming their own insurance companies. Such carriers have two advantages: as new companies they do not suffer from an overhang of losses from previous years, and they can take full advantage of cost-containment legislation that may have been adopted recently in their state. Existing commercial companies are likely to find that this legislation will not be retroactive.

A number of states have recently enacted legislation directed at the malpractice crisis. The new laws accomplish such things as clarifying judicial rules, providing for stricter statutes of limitation and creating mandatory review boards to examine malpractice cases before they go to trial in order to discourage frivolous suits. It is likely that most of these reforms will be largely ineffective in reducing malpractice awards.

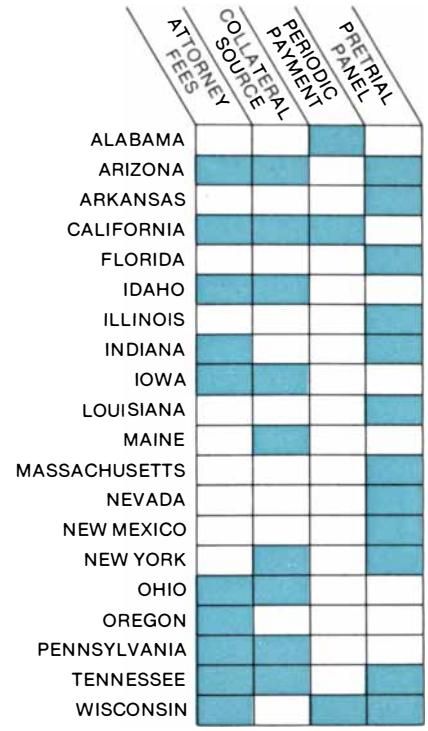
A few effective cost-containment laws have been enacted. They reflect two general approaches, exemplified by the statutes of Indiana and California. Indiana's law limits the liability of the "health-care provider" to $100,000. In addition a compensation fund was created by a charge of up to 10 percent of the malpractice premiums paid by physicians and hospitals; the patient can receive compensation from the fund for as much as $400,000. Florida, Louisiana, Oregon, Wisconsin, Idaho, Pennsylvania and Illinois have followed this general approach. Unless a fund has no ceiling or a high one, however, the statute is vulnerable to the challenge that it is unconstitutional; a challenge of this kind has already been successful in the Supreme Court of Illinois.

California's legislation contains four important reforms. When the patient receives a verdict or a settlement that exceeds $50,000 for the payment of future damages, the money will be paid periodically in amounts ordered by the judge, thus avoiding prolonged payments for damages that may become moot, as when a patient dies sooner than had been expected. Collateral benefits arising out of the patient's injury are disclosed to the jury under California's new law, the idea being that juries will make lower awards when it is known that a patient is receiving money from health or disability insurance, Social Security payments and the like. Where suits are settled out of court, as more than 80 percent are, the benefits will reduce settlements dollar for dollar. California also has limited to $250,000 the amount the plaintiff can recover for noneconomic losses such as pain and suffering and loss of bodily function. Finally, the state has adopted a sliding scale of contingent fees for the plaintiff's attorney.

Among physicians in the states where the malpractice crisis has been the most severe there is substantial pressure for taking malpractice claims out of the jury system. I have already mentioned the speciality court. Other possibilities are arbitration, which retains the negligence rule, and a "no fault" arrangement under which the concept of malpractice would be abolished and a patient who had been injured would be compensated regardless of the culpability of the physician.

The cost of the no-fault approach could be enormous. It should also be kept in mind that a more objective method of determining negligent medical injury (such as arbitration, where textbook evidence can be substituted for the testimony of a medical expert) would greatly increase the number of awards granted to patients in the states that now have a low incidence of malpractice claims. Hence it would be inadvisable to introduce such a reform without effective cost-containment legislation.

Ultimately the losses on malpractice insurance should stabilize at an affordable level, given fair determination of issues of liability, reasonable apportionment of damages, effective cost-containment legislation and the establishment of insurance companies owned by physicians. A vital proviso would be for such companies to apply carefully considered underwriting restrictions, aimed at making certain that an insured physician assumes only those medical tasks he can do well. It can be argued that such a system would amount to a form of licensing through insurance rather than by state law. Such limitations are rational, however, in the light of the growing complexity of modern medicine. Moreover, if physicians own the insurance companies, the problem of fairness in applying the limitations should be minimal.



RECENT LEGISLATION directed at the medical-malpractice crisis is summarized. The chart reflects laws that were passed or became effective between January 1, 1975, and March 15, 1976, but does not include all the laws. Laws affecting attorneys' fees limit the amount a lawyer can charge as a fee in a malpractice case. Collateral-source laws require disclosure to the jury of payments the defendant will receive from sources other than his lawsuit; the aim is to reduce correspondingly the size of the award made by the jury. Periodic-payment laws apply to future damages and seek to eliminate the windfall that results when a plaintiff receives a lump-sum award but then dies shortly afterward. Pretrial panels review malpractice claims with the aim of discouraging frivolous malpractice lawsuits.

A number of physicians believe restrictions of this kind will have little value, since it is generally the "good" doctor who is sued; the less adequate practitioner, who is likely to have a stable practice in a small community, will escape lawsuits regardless of his mistakes. Although this generalization overlooks the numerous jury verdicts and out-of-court settlements against substandard physicians, it contains an element of truth. The best physicians are likely to be in charge of the most difficult cases and so may be the target of a number of claims. Most such claims will not result in the payment of an indemnity, although even the best physician can be guilty of a "slip" that is equated with negligence by expert testimony. It will be the job of a rational system for determining facts in malpractice cases—a system that will do away with juries—to achieve better justice for patients and physicians alike.

The Sleep Factor

During prolonged wakefulness a sleep-promoting substance collects in the brain. Extracted from goats, it can induce excess sleep in rats and rabbits. Efforts are under way to identify its chemical structure

by John R. Pappenheimer

The proposition that natural sleep is brought on by a substance secreted by the brain during wakefulness is in accord with everyday experience. No one who has observed a very sleepy child will find it hard to believe that the intense desire to sleep involves some potent chemical mechanism in the central nervous system. The comment "I slept as though I were drugged" expresses the intuitive feeling that some natural "drug" takes part in the induction and maintenance of a deep and satisfying sleep. Iago taunts Othello:

> ...Not poppy, nor mandragora,
> Nor all the drowsy syrups of the world
> Shall ever medicine thee to that
> sweet sleep
> which thou ow'dst yesterday.

Does the brain produce each waking day some natural "drowsy syrup" that induces sleep each night?

Experimental evidence suggesting that a sleep-promoting factor may be manufactured by the brain during prolonged sleep-deprivation was obtained more than 60 years ago by the French physiological psychologist Henri Piéron. Piéron and his colleagues transfused cerebrospinal fluid from sleep-deprived dogs to the cerebrospinal system of normal dogs and noted that the recipients slept for several hours following the transfusion.

Although simple in principle, the Piéron experiment was fraught with technical difficulties and confusing interpretations. The transfusions involved great stress for both the donor and the recipient animals. Donor dogs were kept awake for 10 days or longer by tying them in a sitting position during the day and walking them constantly at night. The withdrawal and transfusion of cerebrospinal fluid involved puncturing the atlanto-occipital membrane at the base of the brain, an operation that was carried out without anesthetics; the animals struggled violently and often collapsed or were partially paralyzed after the puncture. Piéron's experiments were conducted more than 25 years before the introduction of electroencephalography as a means of evaluating sleep, and his results were necessarily based on simple observation of the animals following the transfusions. Nevertheless, after many experiments, Piéron came to the conclusion that dogs receiving fluid from sleep-deprived donors were likely to sleep for several hours, whereas recipients of fluid from normal donors usually remained alert. In 1913 he published a monograph on sleep that included an extensive account of his transfusion experiments.

The Piéron phenomenon was reinvestigated in 1939 by Jerome G. Schnedorf and A. C. Ivy of Northwestern University. They reported positive results in nine out of 20 transfusions of cerebrospinal fluid from sleep-deprived dogs to normal recipients. The onset of sleep in the recipients came about 45 minutes after the transfusion, lasted between two and four hours and was usually associated with a slight increase in body temperature. Schnedorf and Ivy were convinced of the reality of the Piéron phenomenon, but they questioned its relevance to normal sleep because of the variability of the response, the associated increase in temperature and the extreme degree of sleep deprivation required to obtain a positive response.

No further attempts were made to investigate the Piéron phenomenon until my colleagues and I in the department of physiology at the Harvard Medical School took up the problem in 1965. At the time we were studying the ionic composition of cerebrospinal fluid in relation to the problem of how an animal maintains control of breathing. For this purpose we had devised techniques for perfusing the cavities of the brain in specially prepared unanesthetized goats. The shape and dimensions of the skull in these animals enable one to implant a Teflon guide tube through the occipital bone over the cisternal cavity without interfering with the freedom of movement of the head and the neck [*see illustration on opposite page*]. The membranes over the cisternal cavity can then be punctured at will through the guide tube and cerebrospinal fluid can be withdrawn continuously at the rate of six milliliters per hour without disturbing the animal. With the addition of a second guide tube the entire ventricular-cisternal system of the unanesthetized animal can be perfused with artificial cerebrospinal fluid. After the needles are withdrawn the membranes seal themselves, so that there is little danger of infection, provided that strict asepsis is maintained during the punctures. Between experiments the guide tubes are protected from mechanical damage by a metal cap attached to the animal's horns.

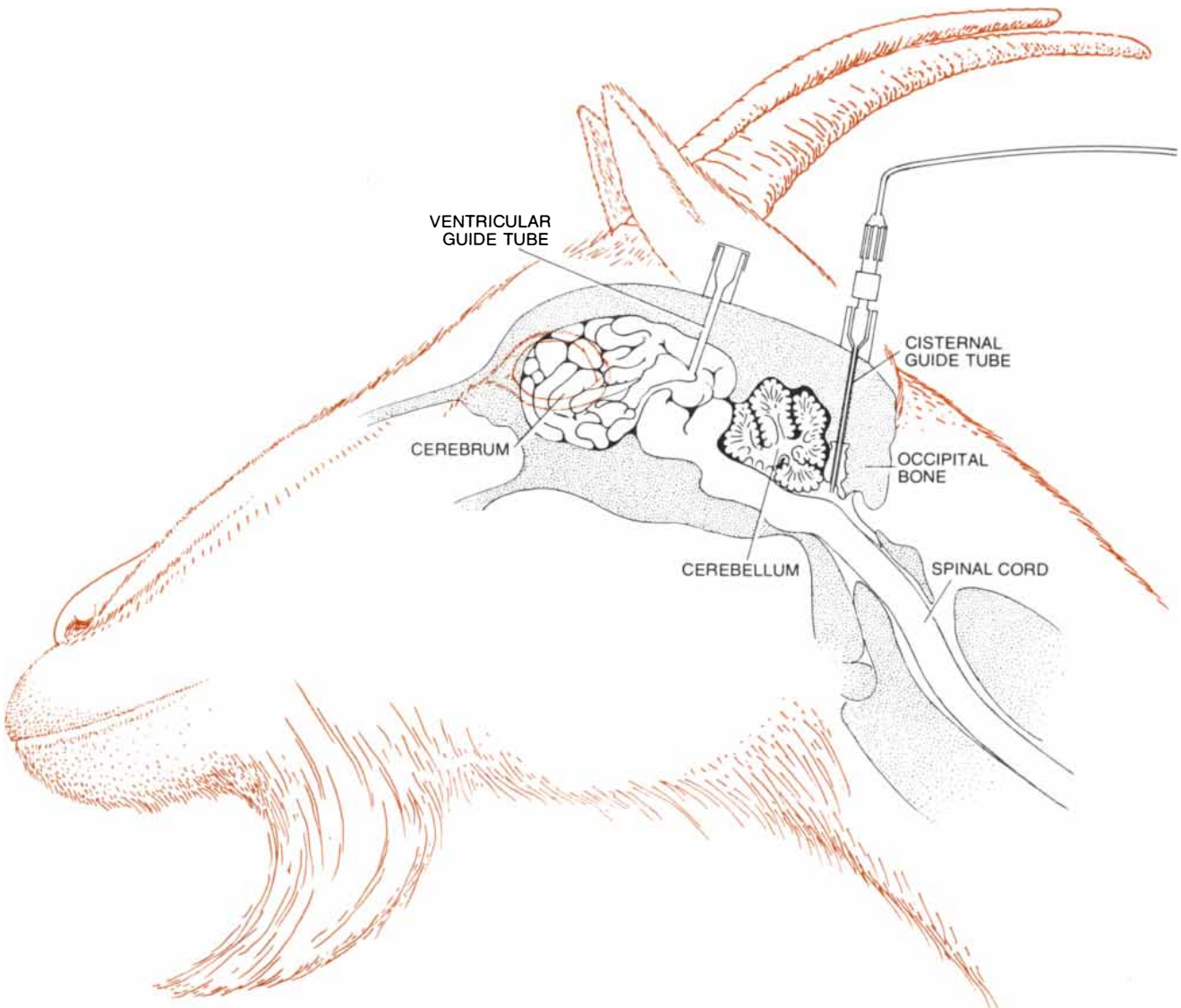
Given goats prepared in this way, it seemed to us that we were in an advantageous position to examine the Piéron phenomenon under conditions involving little discomfort to the animals. We assumed at the outset that the phenomenon, if it were of fundamental importance to the sleep mechanism, would be demonstrable after moderate degrees of sleep deprivation and that it would not be species-specific. (We considered it highly unlikely that each species would synthesize a unique factor to serve a function as generalized as sleep.) With these thoughts in mind we decided to try some pilot experiments on the transfusion of cisternal fluid from our goats to the ventricular system of cats. To this end we prepared a few cats with implanted ventricular guide tubes, so that we could infuse fluids at a low rate into the ventricular system repeatedly and without disturbing the normal conscious animal.

In our first experiments we judged the results by simple observation, just as Piéron had 55 years earlier. We did not really expect positive results. Yet all of us who participated in these pilot experiments (Tracy B. Miller, Cecilie A. Goodrich, James Nicholl and I) had the

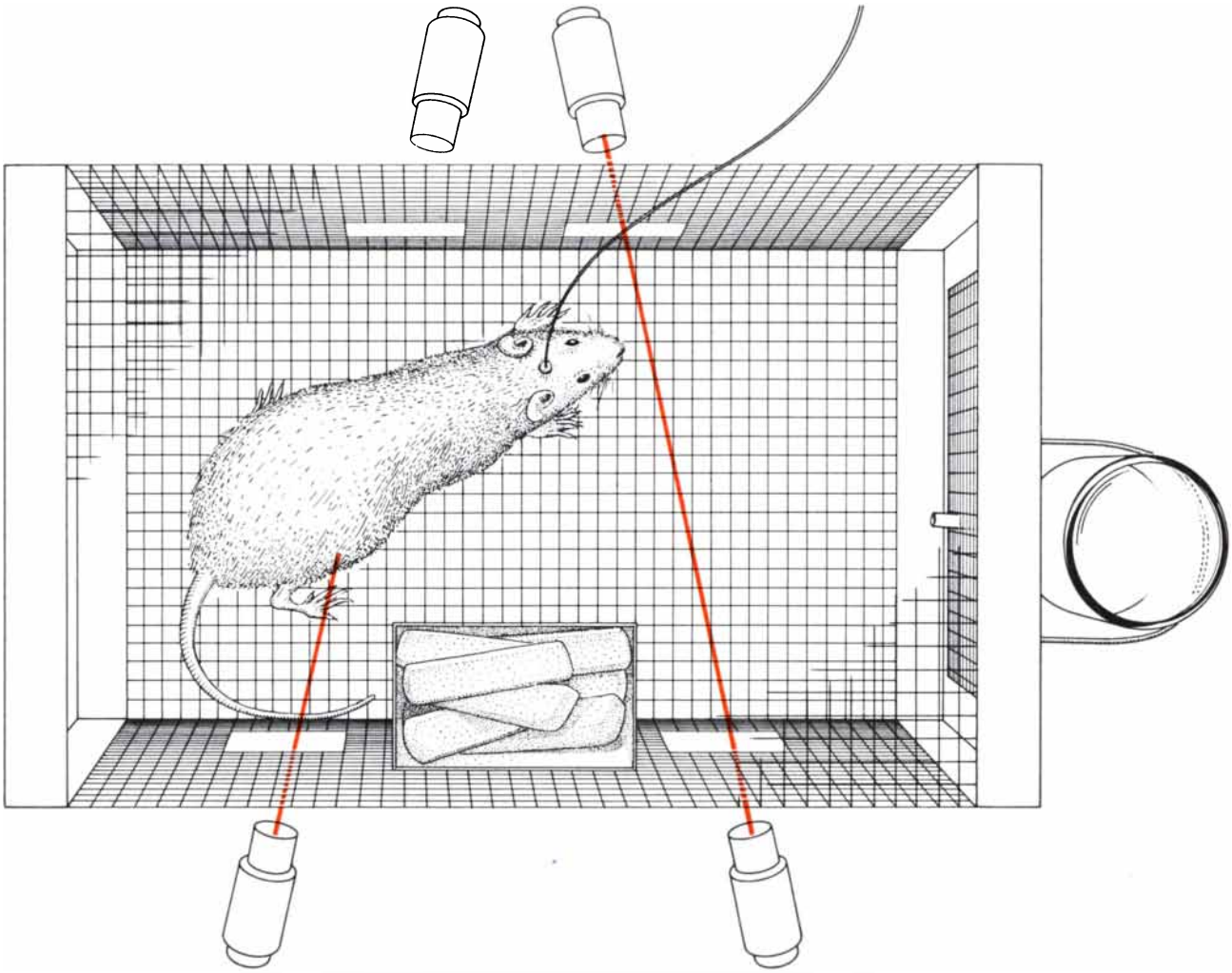
strong impression that cats infused with fluid from goats that had been deprived of sleep for 48 hours were abnormally sleepy in comparison with the same cats infused with fluid taken from the same goats when the goats had not been deprived of sleep. Indeed, the results were so striking and of such potential physiological significance that we put other plans aside in order to devote full time to the systematic exploration of the Piéron phenomenon. Our interest was further stimulated by reports from the laboratory of Marcel Monnier in Basel. He and his co-workers had just found that a neurohumoral factor mediating sleep is released into the venous blood of the brain during electrical stimulation of certain areas of the part of the brain known as the ventromedial thalamus.

Our first concern was to devise a suitable quantitative assay for the sleep-promoting effect. We had it in mind that isolation and purification of the active factor might require hundreds of assays. The prospect of estimating the duration of sleep from an analysis of thousands of hours of conventional electroencephalographic recordings was not an appealing one; we therefore developed a fairly simple indirect assay based on the nocturnal activity of rats. Rats have a well-defined diurnal cycle of sleep and activity: they sleep about 65 percent of the time during the daylight hours and less than 30 percent of the time during the night. They eat and drink mostly at night. Our plan was to infuse the rats with the suspected sleep factor some two hours before the beginning of the dark phase; we had the thought that the exogenous, or non-native, sleep-inducing factor might overcome the animals' natural wakefulness, causing them to sleep more and hence be less active at night.

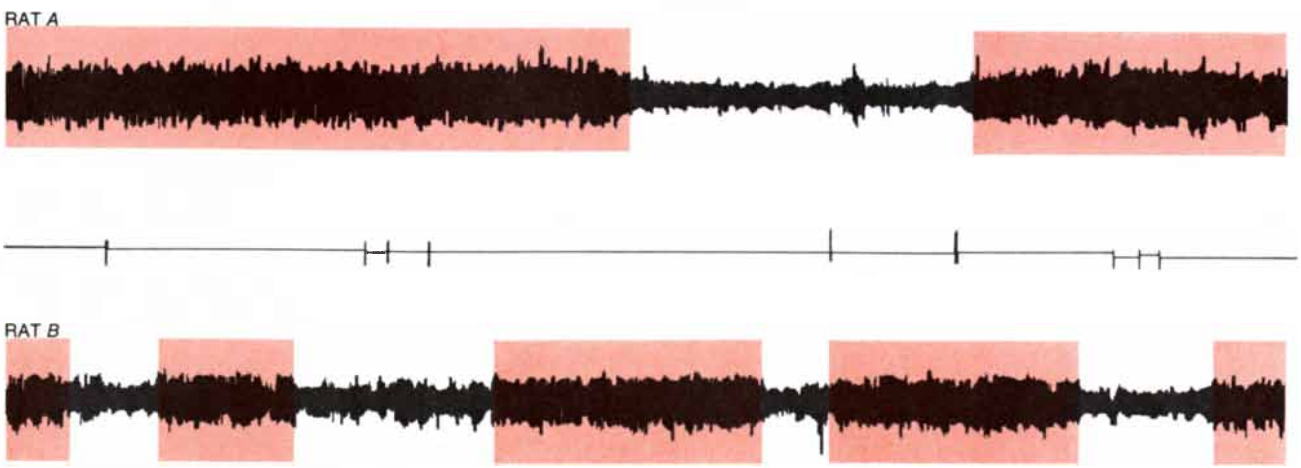
We constructed individual activity cages for 16 rats. Each cage was crisscrossed by dim red light beams focused



SLEEP FACTOR IS OBTAINED from the cerebrospinal fluid of unanesthetized goats that have first been equipped with a pair of Teflon guide tubes implanted through the bone between the horns. The membranes over the fluid-filled cavities of the brain can be punctured at will through the guide tubes; after the needles are removed the membranes seal themselves. Using the arrangement shown in this illustration, for example, normal cerebrospinal fluid can be withdrawn continuously from the cisternal cavity of the goat's brain through the posterior guide tube at a rate of six milliliters per hour without disturbing the animal. With the aid of the second guide tube the entire ventricular-cisternal system can be perfused with artificial cerebrospinal fluid. Between experiments tubes are shielded by a metal cap.



SLEEP FACTOR IS INFUSED slowly into the ventricular system of an experimental animal such as the rat through an implanted guide tube without disturbing the normal conscious animal. The sleep-promoting effect of such an infusion is measured by monitoring the nocturnal activity of the test rat in a specially constructed activity cage, viewed here from above. The cage, one of a bank of 16 such cages in the author's laboratory at the Harvard Medical School, is criss-crossed by dim red light beams focused on photoconductive cells. The interruption of a light beam for two seconds actuates a counter in the next room, providing a 24-hour record of activity of each rat.



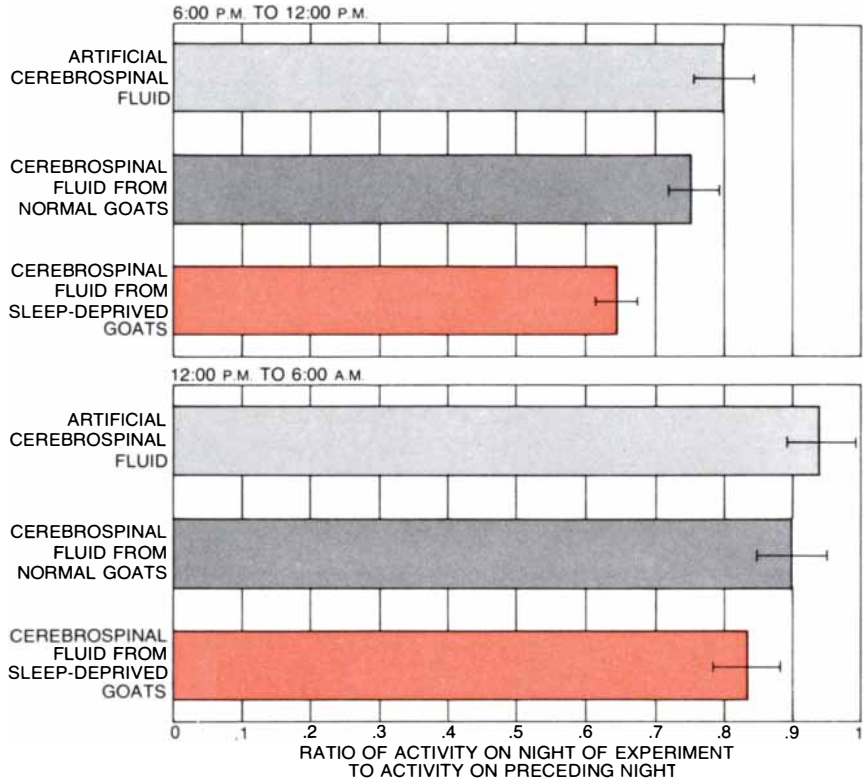
ELECTROENCEPHALOGRAPHIC RECORDINGS of the rats' brain waves are made continuously during the nocturnal-activity measurements in order to obtain more specific evidence of a sleep-inducing effect following an infusion of cerebrospinal fluid from sleep-deprived goats. The electroencephalograms shown here, obtained simultaneously for two different rats, are characterized by episodic cycles of slow-wave sleep (*colored bands*). The episodes, each lasting for a few minutes, are marked by fairly large voltage changes at low frequencies. The nocturnal-activity record for both rats appears in the center. In this case an upward deflection of the pen marker (signaling the interruption of a light beam) refers to rat *A* and a downward deflection refers to rat *B*. As this short section of the combined record shows, brief periods of wakefulness are often accompanied by signals indicating that the rats had crossed the light beams.

on photoconductive cells. The interruption of a light beam for two seconds actuated a counter in the next room. The bank of 16 counters was photographed automatically every six hours to obtain a 24-hour record of the activity of each rat. The rat room was air-conditioned and temperature-controlled and was maintained on a 12-hour light-dark cycle. We developed simple techniques for implanting ventricular guide tubes in the rats, and we found that the animals tolerated repeated intraventricular infusions at a rate that enabled us to introduce .1 milliliter of fluid over a period of 30 minutes.

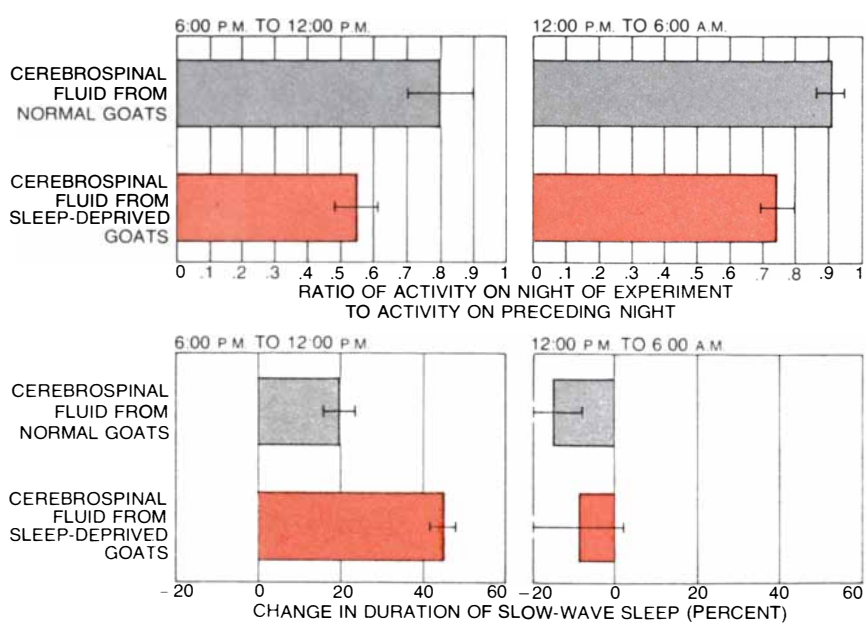
We soon learned that it was necessary to work only with male rats in order to avoid the large variations in nocturnal activity associated with the four-day sexual cycle in females. There were also large variations of activity between individual male rats, but the nocturnal activity of any given rat was fairly constant from one night to the next. Each rat could therefore serve as its own experimental control, with the average of its preinfusion nocturnal-activity counts providing a baseline for comparison with the night of infusion.

The results of our first series of experiments on whole, sterile cerebrospinal fluid taken from normal goats and from goats subjected to 48 hours of sleep deprivation showed that the nocturnal activity of the test rats was reduced to about 63 percent of normal during the first six hours after the infusion of .1 milliliter of cerebrospinal fluid from sleep-deprived goats. Furthermore, we found that this reduction was significantly greater than the reduction following the infusion of sterile saline solution or fluid from normal goats. The reduction in nocturnal activity following the infusion of normal fluids has since been attributed to the fact that the fluids were infused during the two hours before the dark cycle, at a time when the animals are normally asleep. They are therefore deprived of almost two hours of sleep and tend to make it up later on. Sham infusions have the same effect. Nevertheless, the difference between the effects of infusing normal cerebrospinal fluid and cerebrospinal fluid from sleep-deprived goats was statistically quite significant.

The assay based on nocturnal activity is useful for screening large numbers of samples, but it is not specific for sleep. More rigorous evidence of a sleep-inducing effect requires continuous recording or averaging of electroencephalographic recordings. It is well known that sleep, as revealed by electroencephalograms, is characterized by episodic cycles that vary in duration from about 90 minutes in man to a few minutes in small laboratory animals. In adult rats and rabbits each episode of



NOCTURNAL ACTIVITY of experimental rats was reduced to approximately 63 percent of the normal level during the first half of the night following the intraventricular infusion of a tenth of a milliliter of cerebrospinal fluid from sleep-deprived goats (*upper set of bars*). The reduction of activity that follows the infusion of the two control fluids (artificial cerebrospinal fluid and cerebrospinal fluid from normal goats) is related to the fact that all the infusions took place two hours before the dark cycle at a time when the rats normally sleep; they are therefore partially sleep-deprived and tend to make up the deficit later on, mostly in the first half of the night. During the second half of the night the activity levels of the experimental rats showed an almost complete recovery, approaching the normal activity level (*lower set of bars*).



SIMULTANEOUS MEASUREMENTS of the rats' nocturnal activity and the duration of their slow-wave sleep episodes were obtained following intraventricular infusions of cerebrospinal fluid from sleep-deprived goats and normal control goats. The evident correlation between reduction of nocturnal activity and increased duration of slow-wave sleep associated with the infusion of the fluid from the sleep-deprived goats provided strong evidence that a sleep-promoting material accumulates in cerebrospinal system during prolonged wakefulness.

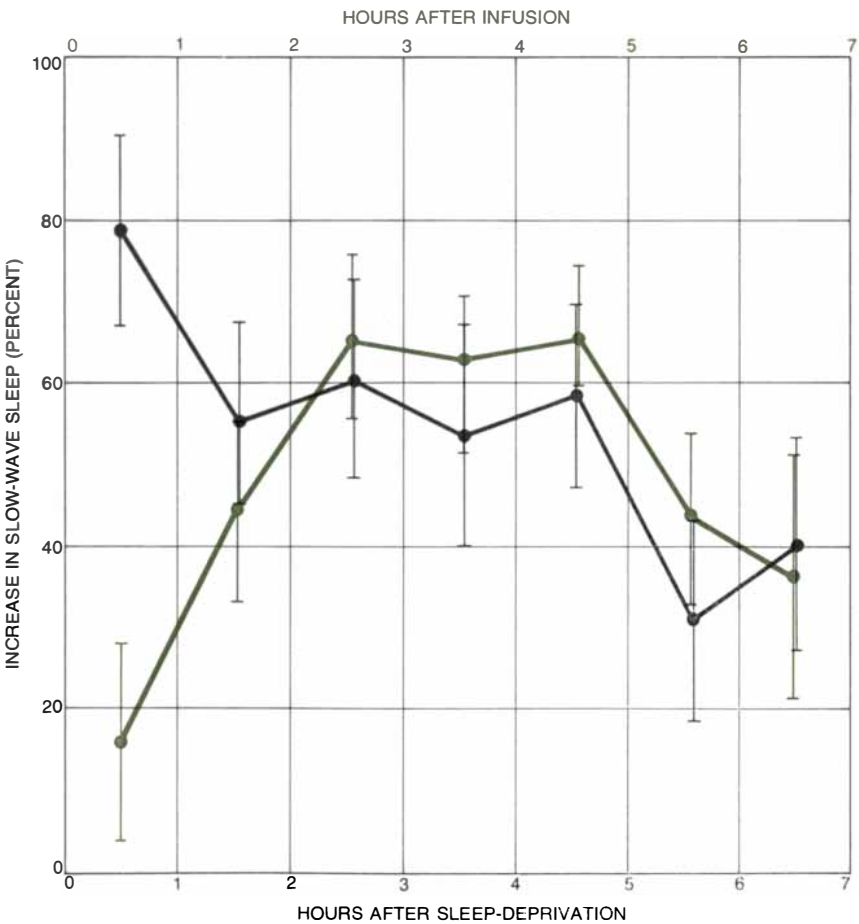
sleep consists mainly of what is called slow-wave sleep, that is, sleep that is accompanied by fairly large electroencephalographic voltage changes at low frequencies.

Working with rats provided with implanted electrodes (in addition to ventricular guide tubes), we were able to show that the reduction of nocturnal activity following an infusion of cerebrospinal fluid from sleep-deprived goats is indeed associated with the increased duration of slow-wave sleep as measured on the basis of continuous electroencephalographic recordings [*see bottom illustration on page 26*]. These results provided strong evidence of the reality of the Piéron phenomenon. In addition we had demonstrated that the sleep-promoting material was present in the cerebrospinal fluid of goats after only a day or two of sleep deprivation and that it was not species-specific. These observations lent support to the view that we were dealing with a phenomenon of broad physiological significance.

With this information in hand, we joined forces with Manfred L. Karnovsky of the department of biological chemistry at the Harvard Medical School in a program designed to identify the sleep-promoting factor. That program has lasted five years and is still incomplete. Nevertheless, considerable progress has been made toward the purification of the sleep factor from cerebrospinal fluid, most recently from the whole brain of sleep-deprived animals. In order to obtain enough starting material we maintained a colony of 20 goats with permanently implanted cisternal guide tubes. We could not puncture the brain of the goats more than once every three weeks, however, without running the risk that adhesions of the cisternal membranes would form. On this schedule, allowing for some failures, it took almost three years to collect the six liters of fluid needed to develop the purification procedures. The work was hampered throughout by the fact that standard fractionation procedures introduced contaminants that interfered with the assay for natural sleep when they were introduced directly into the cerebral ventricles. Moreover, some of the normal constituents of the cerebrospinal fluid can cause excitatory reactions when they are infused into the ventricular cavity in high concentrations. In the end, however, we succeeded in achieving a considerable degree of purification and concentration of the sleep factor from large volumes of cerebrospinal fluid. Solutes with a molecular weight greater than about 500 were removed by filtering the fluid through a series of ultrafine molecular sieves. Solutes with a molecular weight less than 350 were mostly eliminated by passing the fluid through a series of gel filters.

A concentrated solution of cerebrospinal fluid, containing only solutes with a molecular weight in the range between 350 and 500, was found to reduce the nocturnal activity of rats to 50 percent of the normal level for 12 hours following infusion into the ventricles of the rats' brain; in rabbits the normal sleep time was increased by about 50 percent for six hours following intraventricular infusion. The sleep-promoting action of the concentrates could be destroyed by incubating them with a proteolytic (protein-cleaving) enzyme, indicating that the molecule of the sleep factor might have at least one amino acid bond of the peptide type. The active fraction contained primary amino groups that reacted to standard fluorescence tests in a manner characteristic of peptides. It is possible, however, that peptides were present as impurities in the active fraction, and it would be premature to conclude that the sleep factor is a simple peptide.

The amount of the sleep factor in the cerebrospinal fluid is evidently very small. Indeed, most of the material we collected so laboriously from liters of cerebrospinal fluid was used up in the process of learning how to purify and concentrate the factor. Since many substances found in cerebrospinal fluid are present in much higher concentrations in the brain, we sought to extract the sleep factor from the brain tissues of sleep-deprived goats, working with some of the fractionation procedures developed for cerebrospinal fluid. The initial extractions of whole brain, brain stem and cerebral cortex were carried out by procedures known to be effective for the extraction of small-peptide hormones. The crude extracts from the brain were then subjected to the same purification procedures employed with cerebrospinal fluid. Sleep-promoting ac-

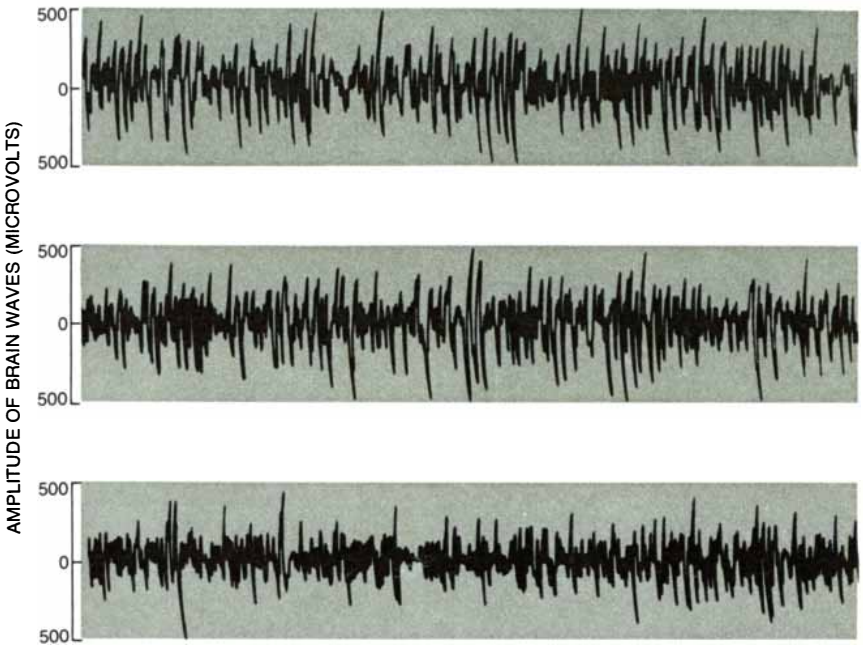


EXCESS SLOW-WAVE SLEEP was also observed in rabbits following intraventricular infusions of concentrated sleep factor from the brain of sleep-deprived goats or sheep (*colored curve*). The increase in slow-wave sleep associated with such infusions is compared here with the increase in slow-wave sleep observed for the same rabbits without infusions but following 24 hours of sleep deprivation (*black curve*). The delay of two or three hours in the onset of sleep following the infusions suggests that the exogenous, or non-native, sleep factor must diffuse slowly to its site of action in the brain, whereas the endogenous, or native, sleep factor formed during sleep deprivation is immediately available at its site of action in the brain. The error bars give the statistical error probability for 13 separate tests on five different rabbits.

tivity was found in the same fractions as those that contained the sleep factor from cerebrospinal fluid. When we tested the sleep-promoting effect in rabbits infused with purified extracts from the brains of sleep-deprived goats, we found that sleep-promoting activity, expressed per gram of brain tissue, appeared to be somewhat greater in the extracts prepared from the brain stem than in those prepared from cerebral cortex, but the assays were not sufficiently quantitative for us to be certain on this point.

The excess sleep induced by an infusion of the sleep factor is not continuous; rather it represents an increase in the duration and the frequency of sleep episodes. The experimental animals are easily aroused by noise, and they wake up at intervals to eat or drink. We noticed, however, that the amplitude of the slow-wave voltage during episodes of sleep was considerably greater than the amplitude observed normally. This was easily seen on standard electroencephalographic recordings, but in order to investigate the effect quantitatively we led the output of the electroencephalograph through electrical filters, rectifiers and integrators to get the mean integrated voltage. The integrations were printed on tape every two minutes. We found that the amplitude of the voltage in the low-frequency range was about 50 percent greater in rabbits during slow-wave sleep induced by the sleep factor than it was during normal sleep. The amplitudes in other frequency bands were unaffected. Evidently the sleep substance has a specific effect on the events that lead to slow-wave voltage.

If exogenous, or non-native, sleep factor prepared from the brain tissues of sleep-deprived animals could induce excess sleep with abnormally large slow-wave voltages, it seemed reasonable to predict that depriving the test animals of sleep would lead to similar electrical events through the accumulation of endogenous, or native, sleep factor. That proved to be the case. The amplitude of low-frequency electroencephalograms during slow-wave sleep following 24 hours of sleep deprivation was indistinguishable from the electroencephalogram recorded from the same rabbit after infusion of the exogenous sleep factor. The course of development of high-amplitude, slow-wave sleep following sleep-deprivation, however, differs from that following infusion of the sleep factor. In the former case high-amplitude, slow-wave sleep starts almost immediately, whereas excess sleep induced by exogenous sleep factor does not reach a maximum until two or three hours after the infusion. The delay in the onset of sleep following infusion into the ventricles of the brain can be attributed to the time required for the exogenous sleep factor to diffuse from the ventricular system to its site of action within the brain tissue. Endogenous sleep substance, on the other hand, may accumulate within the brain close to its site of action.



ELECTROENCEPHALOGRAMS OF RABBIT obtained after the infusion of concentrated sleep factor (*top trace*) and after 24 hours of sleep deprivation (*middle trace*) are presented here along with the normal electroencephalogram of the same rabbit (*bottom trace*). Each of these one-minute recordings is typical of the average of 20 similar recordings made in the hours before and after the traces shown. The first two traces show the high-amplitude slow waves that are characteristic of both the sleep induced by the sleep factor and the deep sleep that follows sleep deprivation. Presumably the latter is induced by an accumulation of native sleep factor during the period of sleep deprivation. The mean rectified voltages in the three examples shown are 140 microvolts (*top*), 138 microvolts (*middle*) and 91 microvolts (*bottom*).

The successful extraction and purification of the sleep factor from the brain tissues of sleep-deprived animals is an important advance because it opens the way to the acquisition of large quantities of starting material for systematic studies of chemical structure and physiological function. Our exploration of brain tissues as a source of starting material coincided with independent studies by Koji Uchizono and his associates at the University of Tokyo. They have extracted and purified a sleep-promoting material from the brain stem of 1,000 sleep-deprived rats. Their material appears to have some of the same properties that ours does. It appears in the same purification fractions and has similar effects on the nocturnal activity and sleep time of rats. So far neither group has obtained material of sufficient purity to allow definitive structural studies, but work toward this goal is being pursued vigorously in both Tokyo and Boston.

The amount of active material extracted from the brain of sleep-deprived animals is extremely small: about a millionth of a gram per 100 grams of brain. It is therefore evident that natural sources are unlikely to provide enough pure material for systematic physiological and biochemical studies of the factor's mode of action. Hence our energies are currently focused on solving the chemical structure with a view toward synthesizing the material.

Many important and interesting questions could be investigated experimentally if the sleep factor could be synthesized. What, for example, is the relation of the sleep factor to Michel Jouvet's "neurotransmitter" theory of sleep [see "The States of Sleep," by Michel Jouvet; SCIENTIFIC AMERICAN, February, 1967]? How is the sleep factor related to biochemical changes known to take place during sleep? Can the sleep factor be labeled with a radioactive tracer so that its origin and site of action can be determined at the cellular level? Can the molecule be altered in such a way as to facilitate its entry into the brain after oral or intravenous administration, thus paving the way for clinical applications? Given adequate supplies of the pure material, there is every reason to expect that progress will be made toward solving one of the most challenging problems of contemporary physiology, namely the function and mechanism of sleep.

Stone-Age Man on the Nile

Thousands of years before the first pharaohs, hunters and gatherers lived along the great river. Their adaptations to their environment underlay the later development of agriculture and high civilization

by Philip E. L. Smith

In Egypt today the annual rainfall south of Cairo, if any, is measured in millimeters. This has been true since the days of the pharaohs. In earlier times, however, the level of the Nile was much higher. The great river and its seasonal tributaries watered a countryside that was fertile and temperate in climate even though it was surrounded by inhospitable desert. This combination of circumstances makes prehistoric Egypt virtually a laboratory microcosm for the study of human adaptations.

When did men first inhabit the Nile Valley? Only in recent years has much been learned about Nilotic prehistory; even today knowledge is spotty. We do have proof that people representative of two general stages of human techno-cultural development, the Lower and Middle Paleolithic periods, were present in Lower (or northern) Egypt from about a million to 30,000 years ago. Apart from the evidence to be gleaned from stone tools, however, little is known about those early inhabitants of Egypt. For example, not a single human fossil from these periods has been found.

As we come to within some 22,000 years of the present, in the latter part of the Upper Paleolithic, a more detailed picture of Egyptian prehistory comes into view. We can begin to speak with some conviction of the way people lived, of their settlement patterns, economic activities and technological proficiencies, and of the environmental and climatic conditions that prevailed. It is of more than passing interest because the Old World during the Upper Paleolithic has traditionally been seen against the background of the harsh glacial environment of Europe at that time, when small groups of hunters pursued such cold-climate animals as the woolly mammoth and the reindeer. A greater contrast in settings can scarcely be imagined than that between the chill, art-rich caves of the Pyrenees and the Dordogne on the one hand and the lush green sloughs and side channels of the Nile on the other, where the hippopotamus, the hartebeest and the gazelle took the place of the reindeer, the bison and the wild horse.

Most of our new knowledge about prehistoric Egypt is a by-product of the dam construction at Aswan in Upper (or southern) Egypt in the 1960's. The Egyptian and Sudanese governments, in association with the United Nations Educational, Scientific and Cultural Organization, invited many foreign archaeologists to salvage monuments and sample as many as possible of the sites that would eventually be submerged by the enlarged Nile reservoir. Some of those who responded, mainly prehistorians from Canada, the U.S. and the U.S.S.R., concentrated their efforts on sites that had been inhabited by early man. Their work, undertaken in collaboration with colleagues in such related fields as geology and paleontology, has greatly enhanced our understanding of the preagricultural populations of the Nile Valley at the time when the last glaciers were beginning to retreat in Europe.

Not all the early-man sites investigated during the 1960's were in danger of flooding. The area where my group from the National Museum of Canada conducted excavations in 1962 and 1963 was threatened in a different way. The Kom Ombo Plain ("kom" is the Egyptian equivalent of the Near Eastern "tell" or "tepe," mound), about 50 kilometers north of Aswan, is an extensive area of ancient alluvial silts, which, although they are now desiccated, need only water pumped from the Nile to transform them into fertile farmland. Reclamation of this kind had begun at the turn of the century, when European promoters established pumping stations and developed sugar plantations on the plain. Now the Egyptian government decided to follow suit and resettle most of the population that would be flooded out of Egyptian Nubia in the remainder of the plain.

Many valuable Paleolithic sites at Kom Ombo had already been planted with sugarcane. My group, sponsored by the Canadian government, and a group from Yale University undertook to salvage or sample a fair number of those that were still undisturbed. In so doing we were following in the footsteps of a French engineer and amateur archaeologist, Edmond Vignard, who worked in a Kom Ombo sugar refinery in the early 1920's and made many useful observations at a time when the early-man sites of the plain were still relatively intact.

Other advantages in addition to Vignard's reconnaissance were available to us at Kom Ombo. One was that the

KOM OMBO PLAIN, on the east bank of the Nile some 500 kilometers south of Cairo, is where the author and his colleagues inves-

stratigraphy, and thus the chronology, of the geological formations there has been worked out in considerable detail by Karl W. Butzer of the University of Chicago (together with his student Carl L. Hansen) and also by R. J. Fulton of the Geological Survey of Canada. These studies made it possible to match the various prehistoric sites with the geological record of fluctuations in the level and course of the Nile and also with changes in the local tributary streams that reflect past changes in local rainfall. The second major advantage was that whereas animal remains from the Paleolithic are both scarce and poorly preserved in most of Egypt, large quantities of bones are present at the Kom Ombo sites. The bones, of course, provide the investigator with invaluable clues to the early inhabitants' subsistence activities.

The environment in late Paleolithic times at Kom Ombo was the product of a complex interplay of factors. The behavior of the Nile itself—its long-term cyclical fluctuations, annual inundations, volume and velocity—was determined by climatic events far off in East Africa. Moreover, the local rainfall, the temperature and the behavior of the tributary streams flowing across the plain from their headwaters in the Red Sea Hills some 150 kilometers away were products of patterns of atmospheric circulation in the Northern Hemisphere, where the European glaciers were still influential. Both factors interacted further with the geomorphology of the plain and of adjoining areas. The result was a mosaic of microenvironments and habitats in a restricted geographical zone. In such an unstable ecosystem rapid shifts in the inventory of plants and animals, and in the exploitative methods of the human cultures subsisting on them, might be expected to occur as one or another variable was altered.

In the Nile Valley immediately north and south of Kom Ombo the river and its floodplain have long been confined to a narrow corridor that runs between sandstone cliffs and high terraces. At Kom Ombo, however, a series of geological faults has caused the cliffs to retreat eastward, so that a wide depression, extending over 500 square kilometers, lies along the east bank. From the prehistorian's viewpoint the most interesting features of the local geology are the sediments—silts, sands and pebbles—that were deposited between 15,000 and 10,000 B.C. by the Nile and its tributary streams. The Nilotic silts, the products of soil erosion far to the south in Ethiopia, were laid down when the river was considerably higher and more vigorous than it is today. They stand some 15 meters above the modern floodplain. It is in and on these silts, known as the Gebel Silsila Formation, that the late Paleolithic sites are found, in some places buried deep and revealed only by erosion or artificial cuts and in others lying exposed on the surface where the desert wind has blown away the concealing silt.

The plain today, with miles of green sugarcane plantations, vegetable fields and irrigation canals surrounding the new town of Kom Ombo, is very different from the arid, dusty wasteland of less than a century ago. It is also very unlike the plain of the late Paleolithic period. Archaeologists no longer accept the notion that before agriculturists transformed the Nile Valley in Neolithic times it was hostile jungle and swamp, difficult to reach, inhabited by dangerous animals and holding little attraction for man. We now know that on the contrary the valley was a zone of fairly open terrain where hunters, gatherers and fishermen had access to a biomass of aquatic and terrestrial resources that could support a considerable population. Moreover, the Kom Ombo Plain was a better than average segment of the Nile Valley, and its inhabitants must have been among the best-nourished people in the late Paleolithic world.

The climate in Upper Egypt at that time was only slightly less arid than it is today. We may guess at an annual

tigated a series of Stone Age living sites dating back to between 15,000 and 10,000 B.C. The site being excavated here belongs to cultures designated Sebekian and Silsilian, dating back to between 13,000 and 12,000 B.C. In it the investigators found stone tools characteristic of the cultures and bones of animals hunted by inhabitants. Cliff in the background is part of the northern boundary of the plain.

LANDSAT SATELLITE PHOTOGRAPH shows the valley of the Nile as the colorful streak running at the left through the desert. The now intensively cultivated Kom Ombo Plain is the bulge to the right of the river. In this false-color photograph the green of the cultivated areas is translated into red. The widening of the river at the lower left is Lake Nasser, which is backed up behind the Aswan High Dam. At the right are the uplands leading to the Red Sea Hills. The valleys leading down to the Nile are wadis that were once seasonal tributaries of the river. They are now dry except on the rare occasions when there is a cloudburst in the Red Sea Hills.

precipitation of 10 or 20 millimeters, falling mainly in the Red Sea Hills during the winter months. The permanent river and the seasonal tributaries flowing westward from the Red Sea uplands, however, largely canceled out the effects of inadequate rainfall at Kom Ombo. The range of temperatures was lower than it is today by perhaps 10 degrees Celsius, and in winter there could be frost.

By that time the Nile had long established its modern regime of summer flooding, induced by the monsoon rains in East Africa. From August to October the river rose to inundate areas that today lie far beyond its shrunken floodplain. At its height the swollen Nile cut long, meandering side channels across the Kom Ombo Plain, creating what were in effect islands until the waters ebbed late in the fall. The floodplain was probably about five kilometers wide. The human settlements were concentrated along the levees of the seasonal overflow channels when the Nile was high and were shifted to the lower floodplain as the river receded. The banks of the tributaries, now dry wadis, have yielded few campsites; perhaps they were less attractive for settlement.

Little concrete paleobotanical evidence has been recovered at Kom Ombo, but it is possible to gain an overall impression of the vegetation, generally subtropical, that grew on the plain. In the low floodplain, in addition to a grassy mat that covered much of the area, a gallery forest of acacia, tamarisk, sycamore and Egyptian willow probably stood beside the main stream. Less dense growths of the same trees would have occupied the channel levees. Thorn trees probably grew in the larger tributary valleys, and the higher water table outside the zone of annual flooding probably supported a semidesert vegetation of low scrub or brush with dry grassland on the hills and scattered desert shrubs farther east. We can also assume a rich growth of aquatic flora—reeds, sedges, lotuses and papyrus plants—along the river, the side channels and the sloughs.

C. S. Churcher of the University of Toronto has analyzed the animal remains from our excavations. His work has revealed a surprisingly wide range of vertebrates: at least a dozen taxa of mammals, 22 of birds, three of fishes and one taxon of reptiles. Prominent among the mammals are a now extinct large wild ox (*Bos primigenius*), the bubal hartebeest (*Alcelaphus buselaphus,* a species still living in the Sudan) and several species of gazelle. These and the hippopotamus were the most important game animals. There were in addition hares, hyenas, a species of canine, bandicoot rats and possibly the so-called Barbary sheep (the aoudad, *Ammotragus lervia,* still found in North Africa).

In the streams and pools lived the

large Nile catfish, the Nile perch and the African barbel, as well as clams, the Nile oyster and a species of soft-shelled turtle. Many of the bird bones are representative of migratory species; the Nile Valley was probably then, as it is now, an important flyway between Europe and Africa. Wading and diving birds included numerous goose and duck species, the cormorant, the heron, the flamingo, the spoonbill, the crane and the curlew. Apparently the elephant, the giraffe, the rhinoceros and large carnivores such as the lion and the leopard were not present. The ostrich, the wild pig, the zebra and the crocodile may well have existed there, but we found no trace of them among the animal bones.

Such an abundant concentration of plant and animal resources must have made Kom Ombo one of the most attractive human habitats available anywhere in late Paleolithic times. An economy at once river-oriented and diversified was to emerge at Kom Ombo and flourish for at least 5,000 years. Even though much archaeological evidence has been destroyed and precise information on the vegetation is meager, we can to some extent plot the seasonal flow of food energy through the plain and attempt to show how human activities were accommodated to long- and short-term fluctuations in the energy flow.

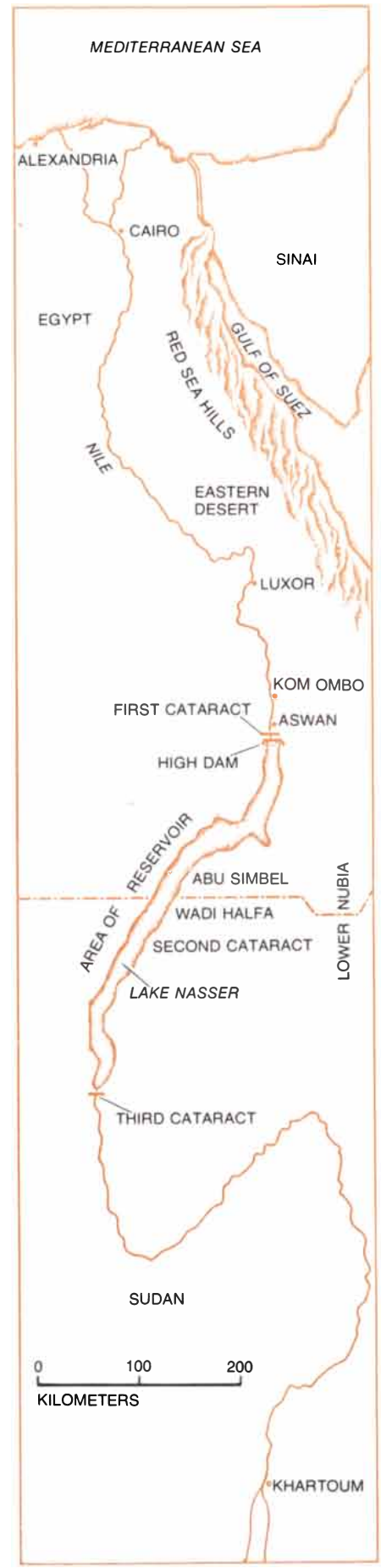
The earliest of the Upper Paleolithic sites at Kom Ombo are about 17,000 years old. There were probably people on the plain before that time, but either their sites have not been preserved or they remain undiscovered. In any event in the centuries immediately preceding 15,000 B.C. rainfall at Kom Ombo was minimal, there was little seasonal runoff in the tributaries and vegetation and game were probably sparse away from the Nile. From about 15,000 to 10,000 B.C., however, rain was generally more plentiful and tributary runoff was greater. The climatic change evidently contributed to rich and varied cultural developments. The same general phenomenon, although differing in detail, has been reported by other excavators who have worked recently in Lower Nubia and Upper Egypt. Evidence of a cultural flowering up and down the Nile Valley has caused prehistorians to revise the traditional view that the later Paleolithic of the area was impoverished. At Kom Ombo alone during this 5,000-year interval we find emerging a series of styles in the manufacture of stone tools (which prehistorians call industries or sometimes, as a convenient fiction, cultures); they vary considerably in the form of the tools, the methods of manufacture and the kinds of stone the toolmakers preferred. We are still not entirely certain how this unexpected and seemingly anarchic diversity in tool production should be interpreted, but it is surely one of the most intriguing new aspects of the prehistory of Egypt.

The majority of the stone tools are small and light. Small flakes and blades were struck from a stone "core" and then chipped into tools. We find no implements that can be interpreted as axes or adzes, and only a small number of heavy tools (usually roughly split or chipped pebbles) appear to have been used for smashing or chopping. Some of the tools are only a few centimeters in length, small enough to be characterized as microliths. One can only assume that most of the stone artifacts were associated in one way or another with the subsistence activities of their makers; unfortunately, as is usually the case in Paleolithic studies, it is hard to ascertain the precise function or functions of an artifact with any degree of certainty.

The first late Paleolithic stone-tool industry recognized on the Kom Ombo Plain is called the Halfan. Carbon-14 determinations at several small campsites place the Halfan industry around 15,000 B.C. It is a curious industry combining relatively archaic and relatively advanced technological features. The archaic feature is what prehistorians call the Levallois method of removing large flakes from a specially prepared core. The method was known in Egypt, in Europe and elsewhere for hundreds of thousands of years before the Upper Paleolithic period. The advanced feature is the fabrication of small flakes with lightly retouched edges.

Only a few Halfan sites have been found. Perhaps the Kom Ombo Plain was not densely occupied at the time. Indeed, it is not until about 13,000 B.C., following a phase when the volume of the Nile had decreased somewhat and its annual inundations were lower, that sites on the plain become fairly abundant.

A second industry, which we have called the Silsilian, and a third, known as the Sebekian, appear roughly between 13,000 and 12,000 B.C. The Silsilian industry specialized in microlithic tools: many small "backed" blades (that is, blades blunted on one edge) and even tiny triangles and trapezoids made on blades of such exotic multicolored stone as agate, jasper and quartz. The Sebekian industry featured longer, narrower blades with the edges lightly retouched by "nibbling," usually near the base; the makers showed a preference for gray or buff-colored flint. Beginning about 11,000 B.C. and continuing for several



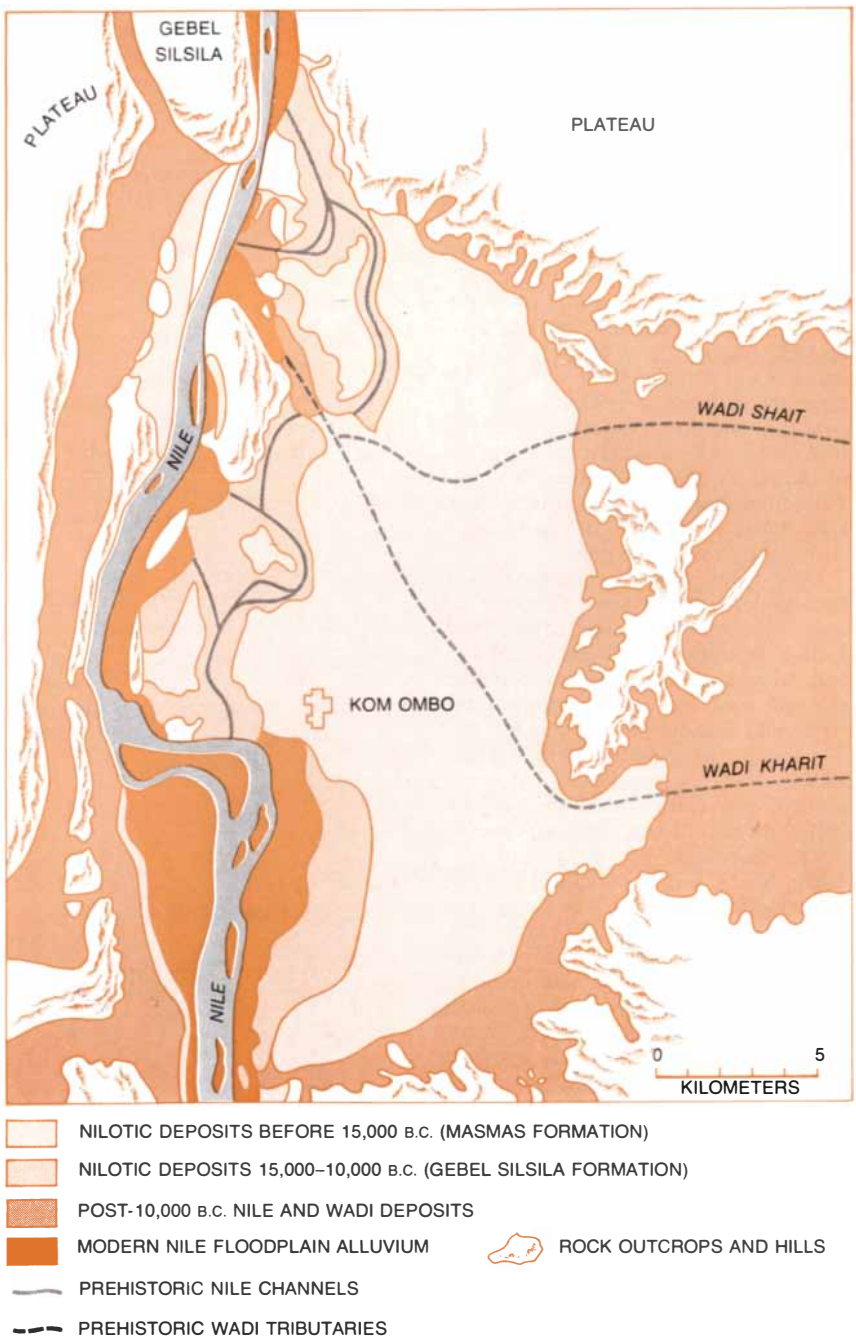
MAP OF THE NILE shows the course of the river from the Sudan (*bottom*) to the Mediterranean (*top*). Some 15 years ago the Egyptian and the Sudanese governments invited many foreign scholars to investigate archaeological sites, actual or potential, along the river that later were flooded by the new reservoir or were destroyed by land reclamation.

millenniums thereafter, a fourth industry, the Sebilian (identified and named by Vignard half a century ago), is found at Kom Ombo. Here the old Levallois technique of core preparation reappears: many of the flakes struck from the core were broad and thin. They were then chipped into geometric shapes, including large triangles and trapezoids as well as microliths. More or less contemporaneous with the Sebilian, and sharing certain of its traits, is a fifth group of artifacts we have named the Menchian industry. Many of the Menchian tools are made on rather thick, heavy flakes and blades; they may have been used for scraping. Both the Sebilian and the Menchian artifacts seem to be associated with sandstone slabs and handstones that were evidently used for grinding or pulverizing.

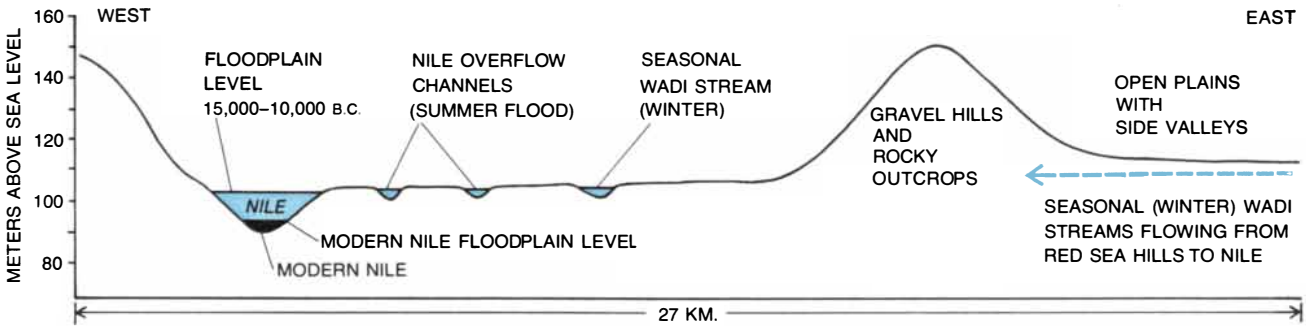
To what extent these variations in stone-tool industries reflect distinct sociocultural groupings, or specialized subsistence activities, or the evolution of one or more traditions over a period of time, it is still difficult to say. The long-term cyclical oscillations of the Nile may well have had some impact on the cultural situation. When the level of the river dropped, as it periodically did for centuries or millenniums, many of the valley zones outside the Kom Ombo Plain where the floodplains were narrower would have been adversely affected as the annual inundations were more restricted. The total biomass of the plants and animals in such areas would have been reduced for long periods, and under such conditions there might have been a tendency for the human groups living in them to move into larger and stabler zones such as the Kom Ombo Plain. The migrations, whether temporary or permanent, of outside groups to the Kom Ombo Plain may help to explain some of the rapid cultural changes and replacements the stone-tool industries seem to reflect between 15,000 and 10,000 B.C.

The exploitable part of the Kom Ombo Plain, which includes the former Nile floodplain, the wadis and the groundwater zones but excludes the modern floodplain (then largely submerged) and isolated rocky outcrops, was probably about 400 square kilometers, or about 150 square miles. It is not easy to calculate the density of Paleolithic populations, but a very approximate figure, based on recorded populations of recent and still living nonagricultural peoples with diversified patterns of subsistence, is about one person per square mile. Thus it is likely that the Kom Ombo Plain could have supported at least 150 people and perhaps as many as 300 under optimum conditions. It is of course unlikely that the population density was constant over the 5,000-year period.

Analyses of the animal remains, together with what we can infer of the river regime and the vegetation patterns, strongly suggest that the plain was capable of supporting human life not just seasonally but all year round. It is nonetheless highly unlikely that the population of the plain could have remained together as a single group throughout the year, or could have remained permanently in a single locality. Probably at any one time the population was split into a number of small bands, perhaps composed of related families who tended to hang together in a loose kind of organization. Each band probably moved in an annual cycle related to the seasonal availability of different food resources. Each of the bands may even have been identified with a certain territory on the plain, although these territories were probably not exclusive. Whether the entire population, or only those who recognized themselves as being culturally related, periodically came together for economic or social purposes we do not know, but judging by



MAP OF THE KOM OMBO PLAIN shows the prehistoric Nile channels, wadi streams and geological deposits. In the period between 15,000 and 10,000 B.C. the river was considerably higher and more vigorous and the climate somewhat less arid than today. Map is based on studies undertaken by Karl W. Butzer of the University of Chicago and his student Carl L. Hansen.



SCHEMATIC CROSS SECTION OF THE PLAIN on a line running east and west in the period between 15,000 and 10,000 B.C. shows height of the Nile, overflow channels and wadi streams. Annual flood of river covered a much greater area than it does today.

the behavior of living hunting peoples it is not unlikely.

The settlement system and subsistence strategy that prevailed were undoubtedly fluid, since they would have had to be correlated with seasonal variations in the abundance of food, just as biomass output itself was linked to short- and long-term pulsations in rainfall and river height. The output was probably spread fairly evenly over most or all of the year, although winter and spring (approximately from November through April) would have been the seasons of abundance. Much of the aquatic biomass, including fish, clams, oysters, waterfowl, turtles and hippopotamus, and such edible plants as water lily, water chestnut, water lettuce, water plantain, papyrus and other reeds, was probably available in all seasons. These foods would have been exploited along the side channels during the flood period from August through October and along the main stream at other seasons. Dry-land provender, including fruits, berries, nuts and edible gums from the acacia, the palm, the sycamore and other trees, and perhaps melons, cucumbers, the "Abyssinian banana" (*Ensete edula*) and various wild grass seeds, would have been most abundant in winter and spring, during and just after the rains. These plants would have been most common along the wadis but would also have grown on the desert steppe beyond them. Roots and bulbs should have been available throughout the year.

Most of the large mammals, particularly the wild ox and the hartebeest (the principal sources of meat), would also have been hunted in winter and spring. One hunting area was the marshy floodplain with its natural pastures beside the lowered river. The wadis and the grassy hills and plains to the east of Kom Ombo were a second area. The herds of wild cattle, unable to go for long periods without water, were almost certainly never far from the channels, pools and pastures of the floodplain and might even have been systematically culled by the hunters all year long. The hartebeests would have ranged more widely, and in the winter, when the tributary streams were flowing, they were probably found along the wadis. Remains of gazelles and asses and perhaps Barbary sheep suggest hunting forays into the drier open grasslands, hills and the fringes of the desert. Fowling was probably in the main a winter activity; many of the birds identified by the bones unearthed at Kom Ombo (for example the crane, smew, goosander and several other species of duck and goose) are migratory.

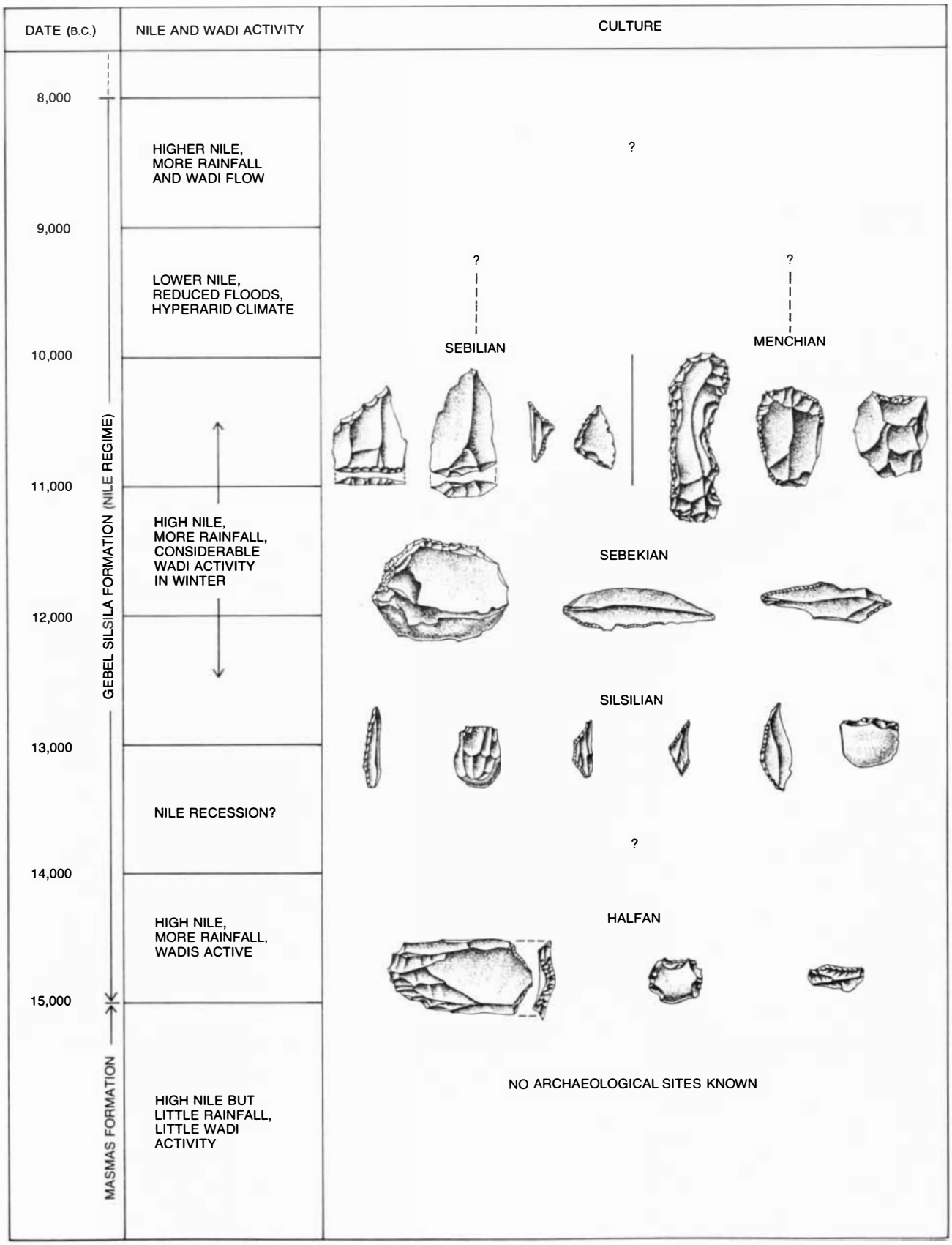
Late spring and early summer (from April through June) was probably a more difficult period of subsistence. As the heat and aridity increased, the Nile shrank to its lowest level, the grassy vegetation diminished and some of the game dispersed. At this time of year the hunting bands may have been forced to split up into smaller units, each perhaps consisting of a few nuclear families, and to spread out more widely in order to exploit the less abundant food resources of the plain until the summer floods resumed.

In terms of reconstructing seasonal activities on the Kom Ombo Plain, one of the most informative sites we excavated was Gebel Silsila III, near the northern end of the plain about four kilometers east of the present Nile. The site stands near a side channel of the Nile that 15,000 years ago ran a few hundred meters from a range of sandstone cliffs. When we found the site in 1962, it was almost entirely buried under an accumulation of hard, carbonate-rich Nilotic silt as much as a meter thick, deposited by river floods soon after the site was abandoned. Only a scatter of flints and bones exposed on the wind-eroded surface betrayed the site's existence. With the aid of Robert Fulton, my assistant Morgan Tamplin and I spent several months excavating here in the winter of 1962–1963. The site proved to be stratified in two main levels: a rich Sebekian occupation, containing many stone tools and animal remains, overlay a somewhat poorer Silsilian one.

Between 13,000 and 12,000 B.C. the prehistoric inhabitants of the plain had settled near a small depression, some 10 meters in diameter, that held water during at least part of the year. Here they lived and ate, built their fires and made their stone artifacts. The large quantity of flint tools manufactured and discarded on the spot, the great number of broken animal bones and the overall size of the area occupied indicate that, at least for the Sebekian people who lived there last, this may have been an important base camp. They hunted cattle, hartebeest and gazelle (a variety of game that reflects the exploitation of several distinct microenvironments), ate an occasional hippopotamus, caught turtles, catfish and perch and killed at least 14 different kinds of birds. It is possible that the Sebekian group occupied the site during more than one season of the year, although not necessarily the year round or even for very long at any one time. For example, the bones of migratory birds suggest that the Sebekian group was at the site at some time between September and April. The bones of immature gazelles, which presumably, like their modern counterparts, were born either in January or in late July and August, show that the group was present at one or both of these times. The bones of immature hartebeests, born at any time between May and December, also point to summer and/or fall occupancy.

In addition to stone tools and animal bones two more intimate relics of human presence have come from the Gebel Silsila III site. One, found in the lower level, is a milk tooth from a Silsilian child less than seven years old. This evidence that children were among the early occupants of the site lends weight to the view that we are dealing here with a family group rather than, say, a temporary encampment of adult hunters. The other relic is a lump of hardened mud from the upper, Sebekian level. It clearly shows impressions from the palm of a human hand, probably an adult's, and is one of the rare examples of a skin impression from the Paleolithic.

Putting together the data from all the sites at Kom Ombo, we are able to reach certain tentative conclusions about other aspects of the inhabitants' lives. First, there are no caves or sheltering rock overhangs in the surrounding cliffs, and



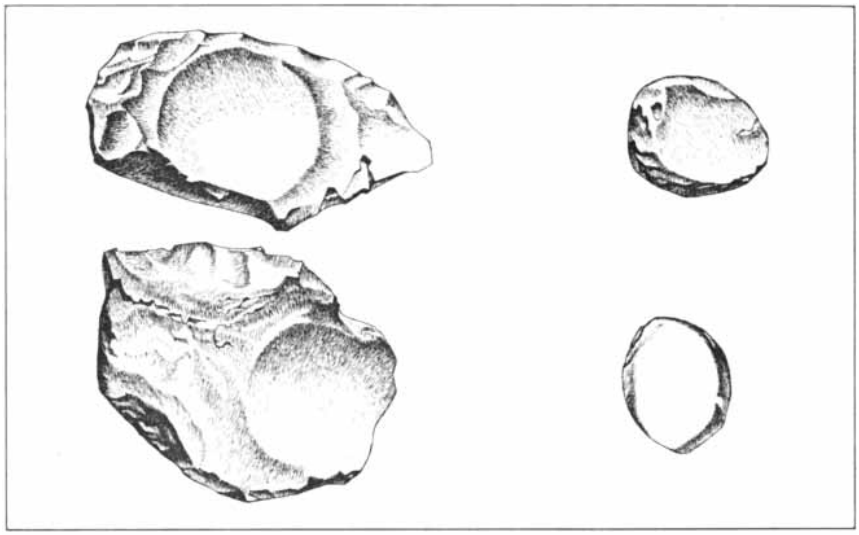
CHRONOLOGY OF CLIMATES AND CULTURES on the Kom Ombo Plain shows the relations between the two. The stone tools of the Halfan culture are characterized by flakes struck from a specially prepared core and by small blades with lightly retouched edges. The Silsilian culture specialized in microlithic tools, many of which are "backed," or blunted on one edge. The Sebekian culture featured long, narrow blades with edges lightly retouched by "nibbling." The Sebilian flakes, like the Halfan, were struck from a specially prepared core and were then often chipped into geometric shapes. The Menchian tools are heavy; many may have been used for scraping.

so all the living sites on the plain were in the open. At none of the sites have we found traces of permanent dwellings: stone or mud construction or postholes that would indicate substantial wood shelters. We can assume that the inhabitants built brush huts, windbreaks or light tents, shelters that would have left few traces, much as the modern Bushmen and Australian aborigines do.

Second, tools made of bone or horn are rare. Wood was probably used more generally, but no trace of wood implements has survived. The inhabitants of the plain did not have pottery, so that it also seems likely that they used containers made of skin, basketry, bark or wood. They probably made nets and lines for fishing and perhaps fowling. Again, however, none of these artifacts have survived. Whether the bow and arrow were present is not known, but the small stone points of several industries may well have served as arrowheads. Small flint blades may have been set into wood knife handles or fish spears. Dugout or reed canoes and rafts would have been useful during the flood season, and perhaps for crossing to the west bank of the Nile, but again no evidence of such craft has survived.

Like most of the other Paleolithic peoples of the Nile Valley (but unlike many contemporaneous groups elsewhere), the inhabitants of Kom Ombo seem to have shown no interest in beads, pendants, bracelets or other personal adornments. Perhaps the red and yellow ocher found in some of the sites was used for body decoration. It is also possible that the spectacular plumage of such birds as the golden eagle, the osprey and the black kite, whose bones are found at the sites, were used for decorative purposes. We have no good evidence that any of the inhabitants engraved or sculptured bone and stone, as their Magdalenian contemporaries in Europe did. Their art, if they had any, may have been expressed in more perishable materials. Interesting scenes of wild animals, including cattle and hippopotamus, are engraved on the cliffs near our Gebel Silsila sites, but no one can prove they were the work of a late Paleolithic group. Physically the Paleolithic inhabitants of the Kom Ombo Plain were fully modern representatives of *Homo sapiens* and were apparently rather robust in build.

With the benefit of hindsight we can now see that many late Paleolithic peoples in the Old World were poised on the brink of plant cultivation and animal husbandry as an alternative to the hunter-gatherer's way of life. The new livelihood had its formal beginnings around the start of postglacial times in southwestern Asia and perhaps elsewhere as well. One current hypothesis about the origins of agriculture is that it was related to late Pleistocene population growth and increased pressure on food resources. This, the hypothesis contends, led in some cases to the greater exploitation of foods that up to that time had been comparatively neglected, particularly plants, smaller animals, birds, fish and mollusks.



GRINDING STONES are associated with Sebilian and Menchian artifacts. Some were possibly used to process wild plant foods. Stones at left are grinding slabs; those at right are rubbers.

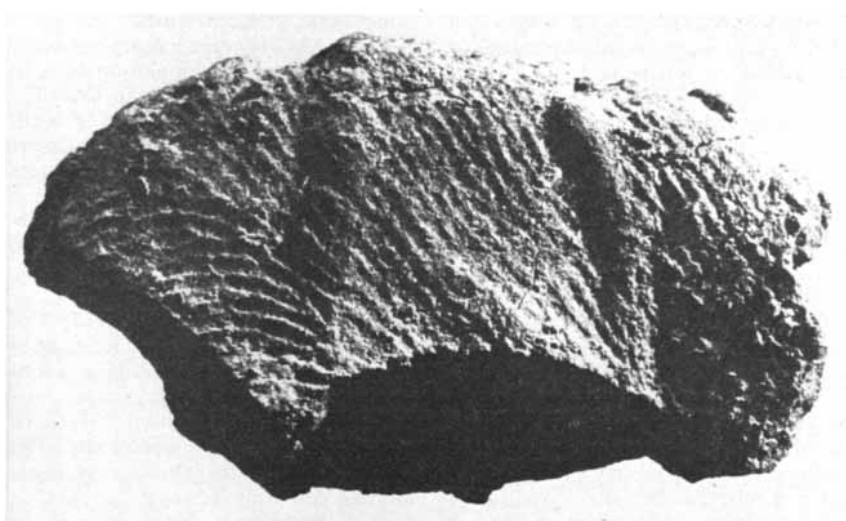
J. Desmond Clark of the University of California at Berkeley and others have recently argued that some evidence for this trend may be seen at Kom Ombo and elsewhere in the lower Nile Valley. Between 13,000 and 10,000 B.C. there appears to have been a general tendency toward population increase and more numerous and larger settlements. At some sites, particularly those of the Menchian and Sebilian cultures on the Kom Ombo Plain, we find many grinding stones: artifacts that suggest the processing of plant foods, perhaps even wild seed-bearing grasses. None of these grasses has yet been identified with certainty, although millets, sorghum and even barley have been postulated. At similar sites elsewhere in Egypt flint blades with a gloss or polish resembling that found on much used stone sickle blades have been reported. It has also been suggested that there may have been some tentative efforts at controlling or taming wild cattle, gazelles and other animals at about this time.

If there was such a trend toward the manipulation and domestication of plants and animals in Egypt at the end of the Ice Age, it would seem to have been a false dawn. We do not know much about human activities at Kom Ombo during the interval after 10,000 B.C. that is sometimes called the Epipaleolithic or Mesolithic. We do know, however, that a complex series of small-scale climatic and environmental changes took place. Contrary to earlier hypotheses, although there was a warming trend in Egypt at the end of the Pleistocene, it was not accompanied by either a sudden or a continuous period of desiccation. From about 10,000 B.C. until 3000 B.C. conditions fluctuated between the semiarid and the hyperarid. But no catastrophic drought, as some earlier archaeologists had believed, forced the hunter-gatherers out of the hills and plains into the narrow confines of the Nile Valley, where all at once, in propinquity with the appropriate plants and animals, they "invented" agriculture. Indeed, local rainfall seems actually to have increased for a time after 9000 B.C. and again about 5000 B.C.

Nevertheless, even though the tributaries continued their seasonal flow, the hydrological budget of the Nile itself now tended to be reduced as a result of climatic changes in East Africa. The river's floods were more restricted, the water table was lowered and the wetlands shrank in extent. Although there were still periodic fluctuations in the size of the Nile, conditions were never quite the same again at Kom Ombo. The plain seems to have become a marginal zone, unable to support a population as large as that of the late Paleolithic. Over the next few thousand years most of Kom Ombo slowly reverted to near-desert conditions; the ephemeral wadi waters gradually faded away, and sometime late in the third millennium B.C. the present hyperarid climate of Egypt became established.

We know that in other parts of Upper Egypt after 10,000 B.C., and in Lower Nubia as well, a hunting and fishing way of life continued for at least 4,000 years more, although its practitioners appear to have been reduced in numbers. At Kom Ombo too some hunting-gathering-fishing groups probably still lived on the western edge of the plain near the river. The general reaction of the inhabitants of the Nile Valley

HIPPOPOTAMUS JAW in a late Pleistocene deposit is cleared by author and a local worker. Presence of the animal four kilometers from today's Nile is evidence of swampy conditions.



IMPRESSION OF THE PALM OF A HUMAN HAND appears on a lump of hardened mud found in a Sebekian deposit. It is a rare example of a skin impression from the Paleolithic.

to the changing environmental conditions seems to have been to place a greater emphasis on fishing and the procurement of other riverine foods in order to supplement the increasingly scarce supply of game animals and land plants. Possibly some of the brief hyperarid periods after 10,000 B.C. led to a rapid reduction in the populations of these plants and animals. Unlike certain groups in southwestern Asia, however, the early Egyptians had not developed a sedentary way of life based on villages and on the collection of wild plants, which under conditions of demographic or ecological stress would have promoted the plants' domestication. If an indigenous trend toward plant and animal domestication developed in the Nile Valley at all, it apparently never passed the incipient level. Such a trend may, however, have helped to preadapt the Egyptians to a ready acceptance of food production later.

When diversified food production finally arose in Egypt, perhaps around 5000 or 4000 B.C., or at least 2,000 years after its development in western Asia, it was evidently introduced from outside and utilized the familiar animals (goats, sheep and pigs) and cereals (wheat and barley) domesticated long before in Asia. The imported plants quickly took the place of those indigenous to Egypt because they were more productive, particularly after the advent of irrigation. They and the imported animals provided the economic base for Pharaonic civilization, which emerged about 3000 B.C. Nevertheless, it is interesting to note that in the earliest Pharaonic era, that of the Old Kingdom, the Egyptians showed a lively interest in domesticating local animals—wild cattle, gazelles, antelopes and even hyenas—and continued to make use of cranes, geese and ducks in sacrifices and for food. The roots of this practice may lie in the preagricultural traditions of the Nile Valley.

In recent years prehistorians working in many areas of both the Old World and the New have uncovered a great diversity of human specializations and adaptations that developed in the closing phases of the Pleistocene. Our knowledge of the riparian hunters, fishermen and gatherers of the Nile Valley provides a valuable addition to the data on these processes of local adaptation. Although we should not exaggerate the role of environmental change in our attempts to explain cultural change, no one denies that there are close, if still poorly understood, articulations between the two, particularly at the hunting-gathering level of cultural evolution. What is being learned from the work on the Kom Ombo Plain and at other localities along the Nile should be of value to all prehistorians in search of the principles underlying the development of human behavior in the distant past.

What constitutes "publication"?

Microfiche. Is this enough?

Science without communication to colleagues might as well be sorcery or fantasy. "Publish or perish" says it another way. But publishing has its economics, and they can be oppressive if the communication needs color but the world needs fewer than 10,000 copies of it. Which is where microfiche makes its contribution. Professional associations have caught on to microfiche to the point that we now publish (on paper) a pamphlet entitled "Micropublishing for Professional Associations." Microfiche is not comfortable to read lying in bed; it requires a reading device. A Kodak Ektalite reader weighs about 6 pounds. A professional association that buys at least 100 of them from us can resell them to its members for about $100 apiece if it needn't turn a profit on them. We can make a master color fiche at a price of $1.50 for each color slide or page sent us to put on it, and the fiche can take up to 120 such images. We strike off copies for distribution at $1.50 apiece for 50, and we drop down by steps all the way to 66¢ apiece for 1700 or more copies. Black-and-white microfiche costs less and has its place in reducing bulk for conveying voluminous data or graphics that do not require color. (Prices, of course, are subject to change without notice.)

Or does it have to be this?

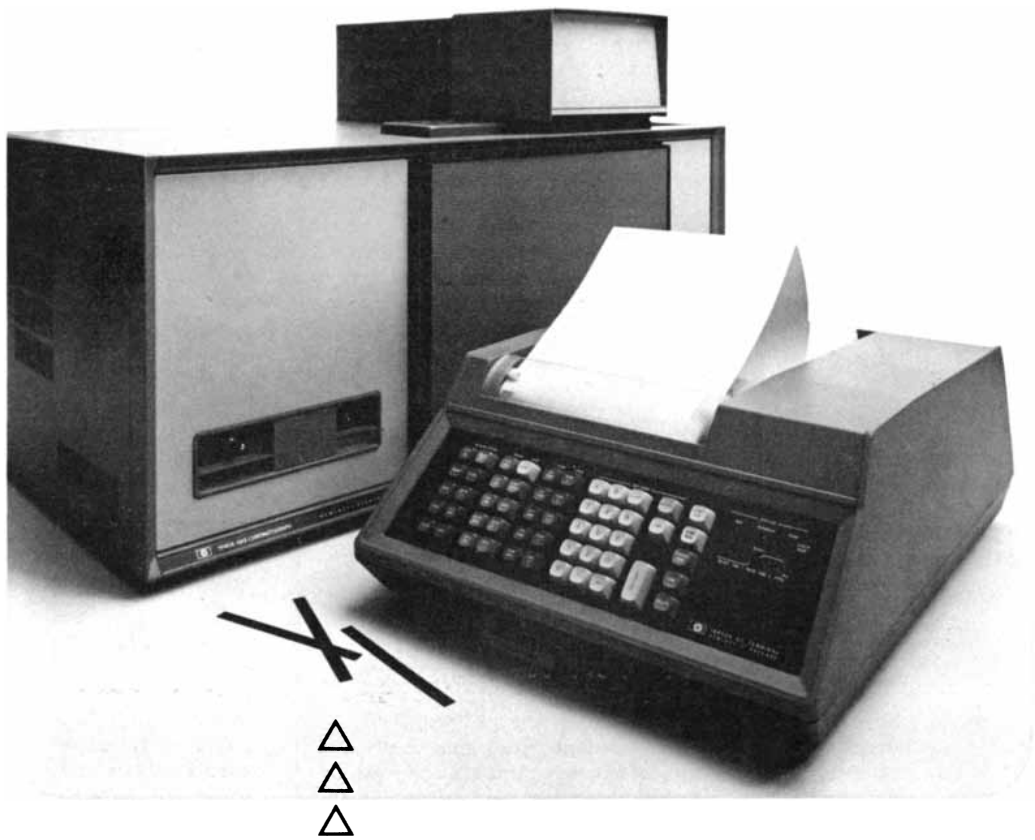
Jaded is the mind that finds no satisfaction in seeing its product committed to printer's type. There is a bit more to it than just communication. Gutenberg has had great impact on our sense of worth, despite abundant proof that a press can just as easily print lies and trash. Within these last few years (and with some help from Kodak technology), Gutenberg's method of raised letters to apply ink is rapidly giving way to photo offset. This has had the effect of broadening the range of price and quality that the industry can offer. The industry itself ranges from the operator of an antiquated press tossed out by a plant that needed more speed and efficiency all the way to the fully integrated house that not only sets type (which the small operator may need to have done elsewhere) but also offers design service that sets the tone of the printed product in the reader's mind more than the reader realizes. An editor facing decisions on which way to go will find upon investigation that some printers are too small and some are too big and one is just right.

And now there is this wonderful machine.

Publication by means of the newly introduced Kodak Ektaprint 150AF copier-duplicator is so quick, so easy, so inexpensive as to be almost too tempting! Place as many as 50 pages of typescript or graphics in the feeder box on top, press the buttons, and before you return with your cup of coffee, stapled duplicates of the entire multipage work are filling the removable tray at the rate of 4,200 pages an hour and in consistently high quality, crisper perhaps than the typing. Editorial deadline could be the same afternoon as the mailing, but the minicomputer inside that takes the instructions from the buttons—the number wanted, one-sided or two-sided, reduced in size or not, one staple or two, darker or lighter—lacks editorial judgment. In-house duplicating departments and commercial duplicating services are now acquiring these Ektaprint machines.

To find out about micropublishing or Ektaprint duplicators, write to Dept. 55W, Kodak, Rochester, N.Y. 14650. To find the right commercial printing house, you don't need us.

Kodak



These are the "keys" to repeating a program: You can record a complete keyboard-entered program on one of these magnetic cards. Inserting it in the 5840A's built-in card reader reprograms the GC quickly and without error.

Programmable GC analyzes a series of samples and gives you the answer to each analysis—automatically.

Two years ago, HP introduced the first "smart" gas chromatograph—the 5830. It quickly gained wide acceptance for its ease of use, and because it automatically presented the user with a complete report of the analysis. Now a second-generation instrument, the HP 5840A, retains all its predecessor's advantages and introduces a new level of automation.

Like the original HP Analytical Answer-Machine, the new HP 5840A provides an answer to an analytical problem completely unattended between sample injection and final report. Its built-in digital processor continuously monitors detector output; reduces it to peak areas and times; identifies individual sample components and computes their concentrations; generates the chromatogram; and presents a complete analytical report on its own printer/plotter—all without operator attention.

In addition, through a powerful set of programming capabilities, the new 5840A analyzes a whole series of samples automatically, each according to the instructions that you give it before the analysis is begun:

- *Time programming* makes changes during an analysis in any integration or GC parameter or in any external device (e.g., sample injector) at precisely the retention time you designate.
- *Run programming* lets you designate changes to be made between analyses, for an entire series of samples, and carries them out automatically at the right time, (e.g., recalibrate the instrument periodically, using any standard sample you designate).
- *Magnetic card programming* lets you record on a small magnetic card a complete keyboard-entered program for single or multiple analyses—and repeat the program simply by inserting the card in a built-in card reader. Magnetic card programming saves the time that it takes to enter complex programs manually and, especially, prevents programming errors.

Prices for the 5840A Analytical Answer-Machine start at $11,525* for a complete system including digital GC, digital processor and magnetic card reader, and printer/plotter with keyboard control.

Keystroke programmability plus continuous memory: a major advance in professional hand-held calculators.

The new HP-25C combines 49 steps of keystroke programmability with a continuous memory. Whenever you switch it off, it retains all the data you've stored in its program memory and addressable registers. The result is a new level of capability you won't easily outgrow.

The basic usefulness of a programmable calculator with continuous memory is immediately apparent. Once you've entered and debugged a program, you have it available for as long as you need it.

We've made this possible by incorporating CMOS technology in the new HP-25C: all data in the program memory and the eight addressable registers are stored in two CMOS chips. Because they consume only 1/80,000 of the calculator's normal operating power, these chips are powered continuously, even when the calculator is turned off, without substantially decreasing the operating life of the battery. Useful life between battery charges is in fact considerably *increased*, because the calculator need be turned on only when it is actually being programmed or used for calculations.

Continuous memory is also useful in some not-so-obvious ways:

- You can add your own special functions to the calculator's preprogrammed functions—as many as will fit into 49 steps. You can store them as permanently as you wish, and recall any of them for immediate use with a few strokes of the programming keys.
- You can add conversion constants in the storage registers—any constant you frequently use in your work—and recall them at a single keystroke no matter how often you switch the calculator off.
- You can use the storage registers as a notebook to save data from previous problems for use in calculations at another time and place, or to accumulate the sum of a number of data entries made during the course of a long-term experiment.

- And you can manage your time more efficiently by programming your HP-25C *in advance* of meetings, seminars, lab sessions, or classes.

Except for the addition of continuous memory, the HP-25C is identical to the HP-25 Scientific Programmable: 72 scientific, engineering, and mathematical functions and operations; 49-step programming memory with decision branching capability; a choice of display formats in fixed point, scientific, and engineering notations—power and flexibility enough to cope with the growing complexity of professional calculating requirements.

The HP-25C costs $200* including comprehensive owner's handbook, applications programs book, a booklet on "Getting the most from continuous memory," battery pack, recharger/AC adapter, and soft carrying case. Price of the HP-25 *without* continuous memory is $145*. You can get them at quality department stores and campus bookstores, or from HP. Call 800-538-7922 in the U.S.A. (in California, 800-662-9862) for the name of the HP retailer nearest you.

For more information on these products, write to us. Hewlett-Packard, 1503 Page Mill Road, Palo Alto, California 94304. *Domestic USA prices only.

HEWLETT hp PACKARD

Sales and service from 172 offices in 65 countries.

For assistance call: Washington (301) 948-6370, Chicago (312) 677-0400, Atlanta (404) 434-4000, Los Angeles (213) 877-1282, Toronto (416) 678-9430

Mail to: Hewlett-Packard, 1503 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP-25C Scientific Programmable Calculator
( ) HP 5840A Programmable GC

Name _____
Company _____
Address _____
City _____ State _____ Zip _____

00646

SCIENCE AND THE CITIZEN

Rules of the Game

Molecular biologists now have a detailed list of do's and don'ts covering their work with recombinant DNA, a line of research with the potential for much benefit and a still uncertain degree of risk. In June the National Institutes of Health issued guidelines for experiments in which bits of the genetic material DNA taken from different sources are combined into biologically viable molecules that can be introduced into a host and there reproduce themselves. The NIH action was the culmination of a series of events new to scientific history: investigators themselves had become concerned about the possible consequences of novel procedures, had made their concern public, had voluntarily restricted their experimentation and had asked an agency that can effectively regulate them to formalize the restrictions and thus ensure their general application.

The experiments in question grew out of the discovery in the late 1960's and early 1970's of various enzymes that could cut DNA sequences in specified ways and of other enzymes that could link DNA segments together. It became possible to select a carrier molecule (typically a virus DNA or one of the nonchromosomal bits of bacterial DNA called plasmids), to splice into it a bit of foreign DNA (from another virus or plasmid, a bacterium or a higher organism) and to introduce the resulting chimera, or combined molecule, into a host cell (typically a bacterium). The chimera replicates itself dependably in the host, giving it new genetic characteristics that allow it to be isolated and cloned into a line of specifically endowed cells (see "The Manipulation of Genes," by Stanley N. Cohen; SCIENTIFIC AMERICAN, July, 1975). The procedure breaches the natural barriers among species, which means that it brings the biologist into new and untested territory.

In 1973 investigators gathered at a biology conference suggested that the National Academy of Sciences consider whether the evolving recombinant-DNA research required some form of control. A National Academy committee was formed, and in 1974 it proposed that certain potentially hazardous recombination experiments be deferred, that the NIH set up a committee to evaluate recombinant research and establish guidelines and that an international meeting of investigators be convened to review progress in the field and consider the technique's potentialities. Such a meeting was held at Asilomar in California early in 1975. The participants agreed that recombinant research should go forward but that specific kinds of precaution should be taken, matched to the potential hazard of particular procedures, and that some recombinations should not be carried out at all. An NIH Recombinant DNA Molecule Program Advisory Committee then drafted a set of guidelines, which went through several changes before being approved and published.

The guidelines are intended to protect people engaged in the experiments, the general public and the environment. Five kinds of experiment are "not to be initiated": cloning of DNA's from disease-causing organisms or from tumor viruses; recombinations involving genes for the synthesis of potent toxins; recombinations that might increase the virulence or range of plant-disease genes; deliberate release of hosts containing recombinants; transfer of a drug-resistant trait to an organism that does not acquire it naturally, if the transfer could compromise control of disease. As for permissible experiments, four levels of physical containment are described, ranging from ordinary laboratory conditions to the sterile, sealed-off conditions hitherto required only in facilities handling the most dangerous microorganisms. Various host categories are also established, graded according to the degree of "biological containment" they provide by reason of mutations that make them unlikely or unable to survive except in the laboratory. Finally, combinations of physical-containment levels and permissible hosts are prescribed for various categories of experiments.

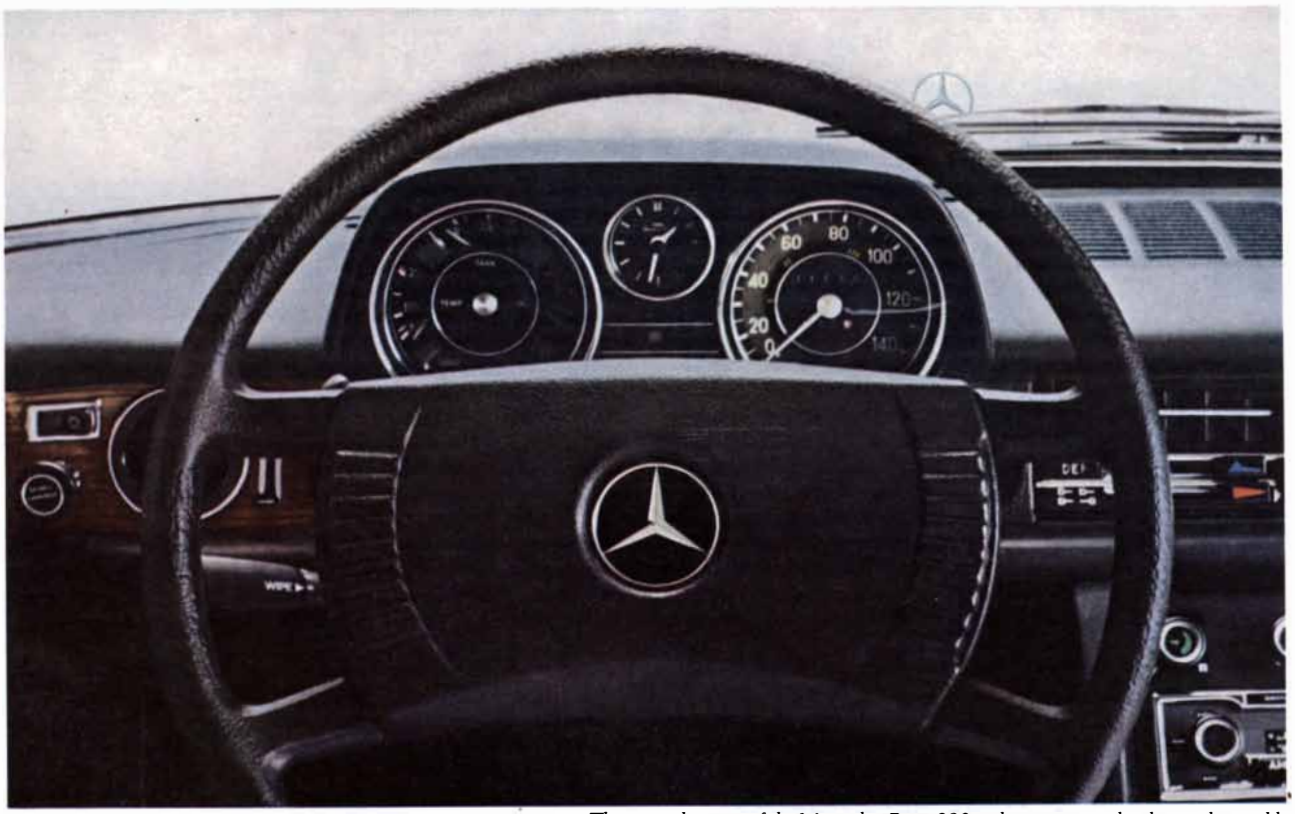
The NIH rules follow the principles outlined by the Asilomar report but are far more specific, and they are more rigorous in that they require particular procedures and documentation. A few investigators in the field apparently feel that the regulations are not sufficiently restrictive. Others believe that, by calling public attention to DNA recombination and cloning as a more hazardous procedure than other procedures in biology, investigators may have occasioned more public concern than is warranted. In any case the guidelines are likely to be reviewed and revised frequently as new experimental results reveal special hazards or (more likely) demonstrate that certain procedures are safer than had been expected.

Teravolt Territory

Enrico Fermi once drew a sketch of the ultimate particle accelerator: a machine that circled the globe at the Equator. At a recent meeting in Serpukhov, near Moscow, physicists from the U.S.S.R., Western Europe, the U.S. and Japan discussed the desirability of building a "world machine," somewhat smaller than Fermi's but nonetheless its symbolic equivalent. The world machine would be roughly 10 miles in diameter and would be capable of accelerating protons to 10,000 gigavolts (GeV), or 10 teravolts (TeV). A few days before the Serpukhov meeting the world's largest accelerator at the Fermi National Accelerator Laboratory ("Fermilab") in Illinois, with a ring 1.25 miles in diameter, for the first time accelerated protons to 500 GeV. A similar machine just completed near Geneva for the European Organization for Nuclear Research (CERN) has reached 400 GeV. The physicists at Serpukhov were of the opinion that the world machine, also dubbed VBA (very big accelerator), is beyond the resources of a single country or even a small group of countries. The VBA would probably cost between $1 billion and $2 billion, from three to six times more than the large machines at Fermilab and CERN.

Like the Fermilab and CERN accelerators, the VBA will be a proton synchrotron with a fixed target. Because essentially all the accelerated particles collide with other particles in the target, fixed-target machines have a high "luminosity" and thus are capable of generating intense beams of secondary particles, such as mesons and neutrinos. Much higher effective energies can be achieved by machines in which beams of particles, countercirculating in storage rings, are made to collide, but the luminosity, or number of collisions per cycle, is far lower than it is in fixed-target machines. Three large electron-positron colliding-beam facilities are under construction: PEP at the Stanford Linear Accelerator Center in the U.S., PETRA in West Germany and VEPP-4, a somewhat smaller facility, in the U.S.S.R. In PEP and PETRA electrons accelerated to 18 GeV will collide with positrons of the same energy. The collision energy, measured in the center-of-mass system, is the sum of the energies of the colliding particles, or 36 GeV. (To achieve the same energy release in a fixed-target machine the impinging particles would have to be accelerated to an energy of about 600 GeV.)

Plans for large proton-antiproton colliding-beam facilities have been laid at the Brookhaven National Laboratory and at CERN. In the Brookhaven unit, ISABELLE, the colliding beams would each have an energy of 200 GeV; in the CERN unit, LSR, each beam would have an energy of 400 GeV. The respective center-of-mass energies of the two facilities would therefore be 400 and 800 GeV. (To achieve a center-of-mass energy of 800 GeV a fixed-target ma-



The control center of the Mercedes-Benz 280 — the most copied sedan in the world.

What does it feel like to drive the most copied sedan in the world?

Eight of the world's major automobile makers have either compared their cars to the Mercedes-Benz 280—or have actually tried to copy it. Your first drive will show you that no one has *copied* the 280 at all.

When you take the wheel of a 280 Sedan, the automobile states its own case. Listen to the engine. It's a sophisticated overhead camshaft six.

Press the 280 into a tight turn. Fully independent suspension gives you uncanny road adhesion. And, because *every* wheel has its own separate suspension system, bumps or potholes can affect only one wheel.

It's a different story with most of the imitators. Their rear axles are single rigid units. So a jounce on one wheel produces a bounce on its mate.

You get what you pay for

Others have copied the 280's lines and many of its dimensions. But few have attempted to copy the completeness of its standard features. And none has matched all of its engineering features. Features that give you greater safety, comfort and performance. Features that account for the 280 Sedan's price.

Finally, consider this financial fact. Based on the average official used car prices over the past five years, a Mercedes-Benz holds its value better than any make of luxury car sold in America. And even among the Mercedes-Benz models listed, the 280 Sedan's figures are outstanding.

You get what you pay for in the 280: a unique driving experience that is the sum of *all* the reasons why the 280 is the most copied sedan in the world.

Mercedes-Benz

Engineered like no other car in the world.

The 280 Sedan: Rewards you may never have experienced in an automobile.

IF YOU THINK YOUR PTA CAN'T DO ANYTHING ABOUT JUVENILE DELINQUENCY, READ SOME ARRESTING FACTS FROM UNIVERSITY CITY, MISSOURI.

Before he was grown up, Billy was almost sent up. Billy was caught selling dope. He was one of ten boys — all in the ninth grade — who had just been arrested for their first offense.

The St. Louis County PTA Council decided to do something before the second. Working with a representative from the juvenile court, they created a program to tutor these boys. To help with their studies. And teach them a job skill.

Twice a week the boys met after school to learn tailoring. They made their own shirts. They made their own trousers. They made friends. None of the boys have been in trouble with the police or courts since.

What made the St. Louis County PTA Council so effective? For one thing, they're part of the National PTA. An organization that offers information, aid, a voice in government, even sources for funding. But for the most part, it was the people in University City themselves who made the difference. They simply got involved. They tried to do something. And they found out they could.

You can do something too. About juvenile delinquency, school policy, health, whatever. Call your school office. And join The Today PTA. We are what you make us.

THE TODAY PTA. JOIN.

chine would need a beam of 320,000 GeV.) If the U.S. goes ahead with ISABELLE, as seems likely, CERN will probably decide to cancel LSR.

Meanwhile Fermilab is well along with the "doubler," the installation of a second ring of magnets (superconducting) in the tunnel of the 500-GeV synchrotron. The second ring will raise the 500-GeV beam to 1,000 GeV. Fermilab has also proposed the construction of a storage-ring facility, POPAE (protons on protons and electrons), in which a 1,000-GeV beam of protons would collide either with a proton beam of equal energy or with an electron beam of 20 GeV. The Russians, whose largest accelerator is the 70-GeV proton synchrotron at Serpukhov, are working on plans for UNK, a fixed-target machine of 2,000 GeV, to which storage rings could be added to generate center-of-mass collision energies of 4,000 GeV, or 4 TeV.

Nuclear Power Abroad

In the midst of the current debate over the appropriate role of nuclear power in the future U.S. energy budget it is easy to lose sight of the fact that elsewhere in the world the commitment of nations to the generation of electricity from atomic fission is proceeding apace. According to the latest available statistics, compiled by the Atomic Industrial Forum, there are at present 18 countries other than the U.S. where nuclear power plants are generating electricity. The combined capacity of the 112 operating power reactors in those 18 countries is rated at 35,882 megawatts of electricity. The U.S., in comparison, has 60 nuclear power plants licensed to operate, with a capacity of 41,954 megawatts.

In terms of the nuclear share of each country's total electrical supply the U.S., which has the capacity to produce more than 8 percent of its electricity from nuclear power, ranks behind several western European countries, including Switzerland (18 percent), West Germany (15 percent), Sweden (13 percent), the United Kingdom (10 percent) and France (10 percent). In addition Japan can now produce about 8 percent of its electricity from nuclear energy.

More to the point, however, is the fact that the number of foreign countries that have nuclear power plants under construction, on order or planned has risen sharply in recent years and now stands at 41. The nuclear-plant commitments made by those 41 countries as of mid-1976 total 454 reactors with a combined capacity of 343,355 megawatts. This last figure represents an increase of more than 17 percent from 1975, when commitments totaled 294,278 megawatts, and more than 56 percent from 1974, when they totaled 220,575 megawatts.

After the U.S., France leads in total nuclear-plant commitments: 47 reactors with a projected total capacity of 39,345 megawatts. Other countries with large projected nuclear-generating capacities include Spain (35,845 megawatts), West Germany (28,683 megawatts), Iran (27,200 megawatts), Italy (21,386 megawatts) and Japan (20,002 megawatts). At the last count the U.S. was committed to a total of 228 reactors with a combined capacity of 226,189 megawatts.

Naked Charm

The quark model of elementary particles has often seemed an exercise in pure reason: the quarks were invented to explain the internal structure of particles that were not known to have an internal structure. A celebrated addendum to the theory, the "charm" hypothesis, went on to explain the structure of other particles whose very existence was in doubt. These audacious conjectures are now being substantiated. The quark model as a whole is deemed to be a success, and a few particles discovered in recent months give every indication of possessing the new property of matter called charm.

The new particles were observed with the SPEAR particle-storage ring, the machine at the Stanford Linear Accelerator Center (SLAC) in which electrons and positrons collide head on and annihilate each other. SPEAR is one of the laboratories where the psi, or *J*, particle was first seen two years ago. It was immediately proposed that the psi consists of a charmed quark bound to a charmed antiquark, and that explanation of it has been supported by the subsequent discovery of several related states. The psi, however, is a particle that hides its charms: the property cannot be observed because the charm of the quark and that of the antiquark cancel out. The newly discovered particles are not so modest.

Like the psi, the new particles are mesons, members of the class of particles that consist of a quark and an antiquark. The prevailing interpretation is that only one of these quarks has charm, so that the property can be displayed overtly. The mesons were discovered by a group of 41 physicists from the Lawrence Berkeley Laboratory of the University of California and from SLAC, who report their results in *Physical Review Letters*. They have postponed naming the new states until all their important properties have been determined. Many of the same investigators participated in the discovery of the psi.

The new mesons decay too quickly for them to be observed directly; they were found as enhancements at particular energies in the production of *K* mesons and pi mesons, which are interpreted as being the decay products of some new state of matter. The first particle detected was an electrically neutral one that decays into one *K* meson and one pi

meson or into one *K* meson and three pi mesons. It has a mass of 1.865 GeV. (The mass of a particle is usually given in terms of its equivalent energy in electron volts; 1 GeV is a billion electron volts.) It is always created in pairs consisting of a particle and an antiparticle, and one of these is apparently formed in an excited state that has a mass of 2.01 GeV.

More recently a second particle has been found that decays into one *K* meson and two pi mesons. This particle is electrically charged, and its mass is slightly greater than that of the neutral particle. If the new states are charmed mesons, they consist of charmed quarks in combination with other quarks designated *u* and *d;* these are the commonest quarks, the constituents of the protons and neutrons in ordinary matter.

Evidence of suspected charmed particles was detected earlier in the debris produced by the bombardment of protons and neutrons with high-energy neutrinos. Nevertheless, the observation of the particles at SPEAR had been awaited with some anxiety. If charmed mesons exist, the theory holds that they must be produced in electron-positron annihilations. Moreover, some properties of the particles can be determined more conveniently by one technique than by the other. The neutrino experiments give an excellent measure of lifetime, which is difficult to determine at SPEAR; on the other hand, electron-positron annihilation gives a much more precise measurement of mass.

The search for charmed particles was begun at SPEAR soon after the psi was found. It is now apparent that they were not seen sooner only because the signal of their presence is exceedingly faint. In an analysis of 29,000 electron-positron collisions only about 200 could be attributed to the new particles. (A few hundred more have since been recorded.) In order to help isolate these few events the detector at SPEAR was modified so that pi mesons and *K* mesons could more readily be distinguished. In this way collisions that gave rise only to pi mesons or to *K* mesons could be excluded from consideration. Still, an elaborate statistical analysis was needed before the signal could be seen clearly.

All the properties of the new particles that have been determined so far are those expected of charmed mesons. Their decay into *K* and pi mesons had been predicted; that is why enhancements in the production of those particles were searched for at SPEAR. The particles are made only in pairs, and they are not made at all below a threshold energy. That is the behavior expected of a particle bearing a new property of matter such as charm. Finally, the mass of the least massive charmed meson was calculated theoretically more than a year ago by Sheldon L. Glashow and his colleagues at Harvard University. They predicted a value of 1.83 GeV, which is in good agreement with the observed mass of 1.865 GeV. The authors of the SPEAR report stop short of declaring their discovery to be a charmed meson, but any other explanation would now seem farfetched.

Hot and Heavy

The celestial X-ray sources that were studied in the five years after the first source was discovered in 1962 either radiated steadily, like the Crab Nebula and other supernova remnants, or varied regularly, like Her X-1 in the constellation Hercules. In 1967 the first transient X-ray source to be detected flared up in Centaurus; since then scores of X-ray novas have been identified from such X-ray satellites as *Uhuru, Copernicus, Ariel-5,* ANS (the Astronomical Netherlands Satellite) and SAS-3 (the third Small Astronomy Satellite). Recently five members of still another class of celestial X-ray sources have been discovered: highly variable and extremely massive objects at the center of globular clusters of stars.

Some of the regularly varying X-ray sources appear to be members of eclipsing binary systems; others vary so rapidly (from less than a second up to 10 seconds) that they appear to be pulsars. The transient X-ray sources, like visible novas, tend to flare up and remain extremely bright in the X-ray sky for days and weeks before they fade back to normal. Some of them may emit their radiation when gas they eject is heated by shock waves traveling through it. Others may be energized when matter falls onto a rapidly rotating, magnetized, collapsed object similar to the object that may account for some of the nontransient sources such as Her X-1.

The new type of object has a pattern of variability that is completely unlike the pattern of any of the previously known variable X-ray sources. The brightest of the new sources, 3U 1820–30 in the globular cluster NGC 6624, is described in *The Astrophysical Journal* by a group of American and Dutch workers (J. Grindlay, H. Gursky, H. Schnopper, D. R. Parsignault, J. Heise, A. C. Brinkman and J. Schrijver). The source emits powerful bursts lasting less than a second and then returns to its normal energy output some eight seconds later. There appears to be no regular period between bursts: the object is highly variable on a time scale ranging from minutes to months.

The pattern of variability and the nature of the spectrum of 3U 1820–30 could be explained if in the center of NGC 6624 there were a black hole about the size of Saturn, 1,000 times more massive than the sun and surrounded by a cloud of hot gas some three million kilometers deep. The gas would normally be in equilibrium with the black hole, either surrounding it as an atmosphere or rotating around it as a disk of matter accreting to it. Occasionally some kind of instability might arise that would temporarily alter that equilibrium, generating a quick burst of X rays. The X rays would propagate outward through the hot gas. Some of them would get through the gas unhindered; others would first be absorbed by atoms or ions of gas and then be reradiated a short time later. An observer would thus see an initial sudden burst of X rays that would gradually trail off as the delayed radiation was reemitted by the hot gas.

The new class of X-ray sources could be important to astrophysics in two major ways. First, the previously known X-ray sources seem to be largely confined to the plane of our galaxy, indicating that they may be very young objects embedded in the galaxy's spiral arms. The fact that objects of the new class have been found within globular clusters, which are among the oldest objects in the galaxy, indicates that at least one kind of X-ray source is also very old.

Second, the lack of similarity between the new class of objects and the previously known variable X-ray sources implies that the new objects are themselves very different from the common binary X-ray stars. In fact, the way in which the new objects emit short, sharp bursts of X rays is reminiscent of the mysterious objects that emit short, sharp pulses of gamma rays. Grindlay, Gursky and their colleagues remark: "We cannot exclude a possible relation between these events and the γ-ray bursts... which have yet to be identified with known cosmic objects."

Errant Eddies

The Gulf Stream is the swiftest and most energetic current in the North Atlantic. About 100 kilometers wide, it meanders up the coast of North America from Cape Hatteras to the Grand Banks like a long, twisting ribbon. This is only its average path, however; recent research indicates that the structure of the Gulf Stream is considerably less stable than had been supposed.

The vagaries of the current were first quantitatively measured in the mid-1950's, when the pioneering studies of Frederick C. Fuglister showed that the Gulf Stream is not homogeneous but consists of several narrow bands of rapid flow separated by countercurrents. Occasionally an individual band wanders off its usual course in a broad loop some 500 kilometers long. The loop then pinches off from the main current, forming a ring of swiftly moving Gulf Stream water around a core of seawater of different origin.

Although the existence of Gulf Stream rings has been known for 25 years, the details of their formation, movement and decay are only now be-

The Black Watch Kit $29.95

The Black Watch by Sinclair is unique. Controlled by a quartz crystal... powered by two hearing aid batteries. Styled in understated elegance the Sinclair way. No knobs or buttons. To see the exact time or date just touch the face of the case. A re-set control is on the back.

Correctly assembled watches are guaranteed for a full year...Accuracy within a second per day.

The kit contains:

1. Printed Circuit Board.
2. Unique Sinclair Designed IC.
3. Encapsulated Quartz Crystal.
4. Trimmer.
5. Capacitor.
6. LED Display.
7. 2-Part Case with Window in Position.
8. Batteries and Clip.
9. Black Strap.
10. Full Instructions.

All you provide is a fine soldering iron and wire snips.

Build Your Own Scientific Calculator in 3 Hours for $14.95

Designing the Sinclair Scientific was no small feat of engineering. You don't have to be an engineer to assemble it yourself with our kit.

Put together this thinnest, lightest scientific calculator from eight groups of components, using only a soldering iron and wire snips. Complete instructions are included.

The assembled calculator can do:

- Five-function memory
- Algebraic logic
- Degrees/radians switch
- 8 digit mantissa, 2 digit exponent
- Ln and e^x
- Sine, cosine, tangent, arcsine, arccosine, arctangent
- Square root, pi, reciprocal
- Automatic constant
- Floating decimal point

sinclair™

Sinclair Radionics Inc.
375 Park Avenue, New York, N.Y. 10022

Please send me:

_____ Sinclair Black Watch Kit(s) with date @ $29.95 ea. $_____

_____ Sinclair Scientific Calculator Kit(s) @ $14.95 ea. $_____

Sales Tax (N.Y. Residents) $_____

Shipping & Handling $2.50 per unit $_____

Enclosed is my check/m.o. for total $_____

Name _____

Address _____

City _____

State _____ Zip _____

ing revealed. In *Oceanus,* the quarterly review of the Woods Hole Oceanographic Institution, Philip Richardson and Peter Wiebe discuss the importance of ring eddies in redistributing water and marine life in the western North Atlantic.

Perhaps half a dozen rings, analogous to the cutoff highs and lows created by jet streams in the atmosphere, break off on each side of the Gulf Stream every year. Those forming to the south or east of the main current trap cold water from the continental slope and are called cold-core rings. They are huge structures, extending horizontally from 150 to 300 kilometers and vertically from 2,500 to 3,500 meters. Carrying some 10 trillion tons of subarctic water in their core, they move slowly south or southwest, decaying as they go. They have been known to persist as physically identifiable structures, however, for two years or longer. Data recently gathered suggest that at any given time there may be as many as 15 of them wandering across the Sargasso Sea west of Bermuda. The warm-core rings that form north or west of the Gulf Stream are shallower structures (about 1,000 meters deep) and generally last only about six months before they coalesce with the main current near Cape Hatteras.

For the marine biologist cold-core rings are of great interest because their formation isolates zooplankton and other organisms of subarctic waters within the core structure. Since many of the subarctic species are distinct from those indigenous to the Sargasso Sea, the southward movement of the rings results in the large-scale invasion of one community by another. On the basis of his studies of the species distribution of the shrimplike euphausiids in migrating Gulf Stream rings, Wiebe observes that the formation and decay of a cold-core ring is a kind of natural ecological experiment in which an oceanic community is transplanted from one environment to another and also exposed to biological interactions with the community adapted to the second environment.

Cold-core rings may also have evolutionary significance. Some of them are known to coalesce with the Gulf Stream after a period of weeks or months, reuniting expatriated organisms with similar forms in their home range. If, as seems to be the case, populations of subarctic species living in the rings are exposed to increasing environmental stress as they are carried southward, that stress may provide a progressive selection mechanism. Cold-core rings may thus be a means by which genetically altered populations are introduced into the parent population, thereby increasing the overall adaptability of the species and extending its range. Since the Gulf Stream has been flowing in its present course for millions of years, ring formation and decay over a similar time period may be at least partly responsible for the well-defined biogeographic distribution of oceanic populations in the western North Atlantic.

Salt-Water Cereal

The development of crops that can tolerate a high level of salt would open two large new opportunities for food production: irrigation with seawater and the growing of crops on land where salinity has prevented or driven out farming (see "Salt-Water Agriculture," by Hugo Boyko; SCIENTIFIC AMERICAN, March, 1967). By screening barley for salt tolerance and selecting about a dozen strains that appeared likely to be successful in a salty environment, Emanuel Epstein of the University of California at Davis and his associate Jack Norlyn have succeeded in growing to maturity barley that was planted in a sandy plot and irrigated with seawater. They are now developing a salt-tolerant tomato and planning a similar venture with wheat.

Epstein and Norlyn began with barley because it was known to be more resistant to salt damage than most other crops. Having ascertained that tolerance to salt is inherited and that barley plants could be selected for that trait, they obtained seed of a composite cross representing a reservoir of genetic variability derived from 6,200 strains from all over the world. Using this seed, they embarked on their program of screening and breeding. The most promising strains were planted in an experimental plot near the Bodega Marine Laboratory in California. Water for irrigation was drawn from Bodega Bay. Since seawater contains most of the nutrients a plant needs for growth (potassium, calcium, magnesium, sulfur and a number of trace elements), the experimenters supplied only nitrogen, phosphorus and potassium. (The potassium was not needed but was contained in the commercial fertilizer the experimenters employed.) Most of the plants matured and developed grain.

The results, Epstein said, represent "the beginning of a new system of crop production based on seawater, which provides both irrigation and nutrients."

Vaults of the Toreadors

The skill and nerve of the modern bullfighter are impressive, but they seem commonplace compared with the performance of his Bronze Age predecessors. In the latter half of the second millennium B.C. unarmed and unprotected acrobats in Crete and on the Greek mainland ritually baited bulls into charging; then they leaped over the bull's horns, landed on its back and vaulted to the ground. The ritual captured the imagination of the Bronze Age society; artists depicted the bull-baiters

in seal engravings, sculpture and frescoes. Heinrich Schliemann uncovered one such fresco at Tiryns in the 1880's, and Sir Arthur Evans found others at Knossos in the 1920's.

As with so many other Bronze Age discoveries in Crete, the first definitive description of the ritual to be published in English was written by Evans. Depending for his interpretation primarily on one miniature bronze sculpture of a bull and an acrobat, Evans recreated the action as follows. First, as the baited bull charged with its head down the acrobat seized the animal's horns. Aided by the upward thrust of the horns as the bull raised its head, the acrobat did a handstand on the horns and a back flip, so as to land upright on the bull's back at or near the rump. The maneuver was completed by jumping lightly to the ground behind the moving bull.

John G. Younger of Duke University recently reviewed the 60-odd most legible authentic depictions of Greek bull-leaping now available. Writing in *American Journal of Archaeology,* he notes first that unless the Greeks had carefully trained the bulls that participated in the ritual, the style of leaping proposed by Evans would have been virtually impossible. When a charging bull encounters its target, it does not lift its head straight up; it rotates its head so as to hook one horn upward. Such a motion not only would have made the acrobat's postulated handstand precarious but also would have hindered rather than helped his subsequent back flip.

Younger found that 19 of the acrobatic depictions, most of them on seals, indicate a different approach to bull-leaping. The leaper, Younger suggests, never grasped the bull's horns but sprang, as if diving, over the head of the advancing animal with his arms extended. The leaper's forearms absorbed the force of his headfirst landing on the charging bull's back. Then, as the leaper's momentum carried his legs up and over his head, he pushed downward with his arms, thereby completing the back flip that carried him safely over the bull's tail to the ground.

Significantly, Younger also finds evidence that the ritual had a relatively short vogue. Seventeen of the later depictions show the leaper in a "static . . . pose above the bull. The leaper seems to float there horizontally." These lifeless treatments of the subject suggest to Younger that although the artists may have seen depictions of bull-leaping or may have had the practice described to them, they had never witnessed an actual performance. The interval when representations were made that imply eyewitness familiarity with bull-leaping can be narrowed down to the four centuries between 1700 and 1300 B.C. Thereafter there may well have been no more bull-leaping (or bull-leapers).

Hot Spots on the Earth's Surface

These regions of unusual volcanic activity record the passage of plates over the face of the earth. They may also contribute to the fracturing of continents and the opening of new oceans

by Kevin C. Burke and J. Tuzo Wilson

Scattered around the globe are more than 100 small regions of isolated volcanic activity known to geologists as hot spots. Unlike most of the world's volcanoes, they are not always found at the boundaries of the great drifting plates that make up the earth's surface; on the contrary, many of them lie deep in the interior of a plate. Most of the hot spots move only slowly, and in some cases the movement of the plates past them has left trails of extinct volcanoes. The hot spots and their volcanic trails are milestones that mark the passage of the plates.

That the plates are moving is now beyond dispute. Africa and South America, for example, are receding from each other as new material is injected into the sea floor between them. The complementary coastlines and certain geological features that seem to span the ocean are reminders of where the two continents were once joined. The relative motion of the plates carrying these continents has been reconstructed in detail, but the motion of one plate with respect to another cannot readily be translated into motion with respect to the earth's interior. It is not possible to determine whether both continents are moving (in opposite directions) or whether one continent is stationary and the other is drifting away from it. Hot spots, anchored in the deeper layers of the earth, provide the measuring instruments needed to resolve the question. From an analysis of the hot-spot population it appears that the African plate is stationary and that it has not moved during the past 30 million years.

The significance of hot spots is not confined to their role as a frame of reference. It now appears that they also have an important influence on the geophysical processes that propel the plates across the globe. When a continental plate comes to rest over a hot spot, the material welling up from deeper layers creates a broad dome. As the dome grows it develops deep fissures; in at least a few cases the continent may rupture entirely along some of these fissures, so that the hot spot initiates the formation of a new ocean. Thus just as earlier theories have explained the mobility of the continents, so hot spots may explain their mutability.

Plate Tectonics

The modern theory of plate tectonics divides the superficial regions of the earth into two layers. The lithosphere, the outermost layer and the only one directly accessible to us, is cold and rigid. Below it is the asthenosphere, which is white hot and capable of being slowly deformed. The asthenosphere is not liquid, although there is a small amount of melted rock in the earth's interior. The asthenosphere is a solid, but one that flows under stress. It is not unlike ice, which seems brittle in the form of an ice cube but is quite plastic in a glacier flowing down a mountain valley.

The distinction between lithosphere and asthenosphere is based on rigidity and to a large extent reflects differences in temperature. An older distinction, based on chemical composition, divides the upper earth into the crust and the mantle. The boundary between these layers does not correspond to that between the lithosphere and the asthenosphere. The crust is the upper portion of the lithosphere, and the lithosphere also contains the topmost part of the mantle. The asthenosphere usually lies entirely within the mantle.

Under the oceans the crust is composed primarily of basalt; the continents, on the other hand, are made largely of granitic rock. Granite is lighter than basalt, and the continents are considerably thicker than the oceanic crust, with the result that the continents float well above the ocean floor. It was once proposed that the continents move through the ocean floor like ships, but that hypothesis had to be abandoned. Actually the continents are carried by the lithosphere like rafts locked in the ice of a frozen river.

The lithosphere is broken into about a dozen plates, in which the continents are firmly anchored. The plates separate from one another at the crests of the mid-ocean ridges, where new lithosphere is created. The ridges wind through all the world's oceans and constitute the largest mountain system on the earth. At the crests of the ridges undersea volcanism adds new material to the plates, pushing them apart. The opposite process—the consumption of lithospheric plates—is observed where the plates converge and overlap. In those regions, called subduction zones, one plate plunges under another and is reabsorbed into the mantle.

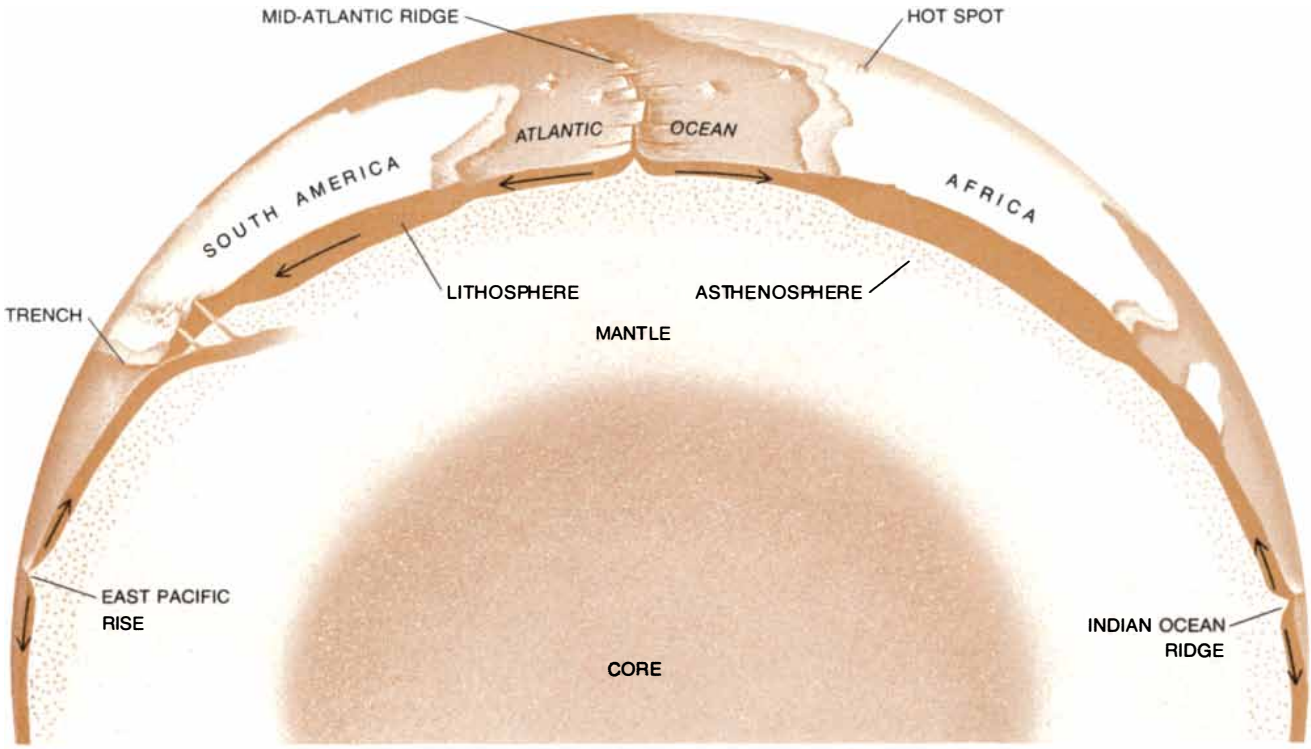
The movement of the lithospheric plates is thought to be associated with large-scale convection currents in the mantle. The currents may actually drive the plate movements, but too little is known about convection in the mantle to warrant firm conclusions.

Hot Spots and Plumes

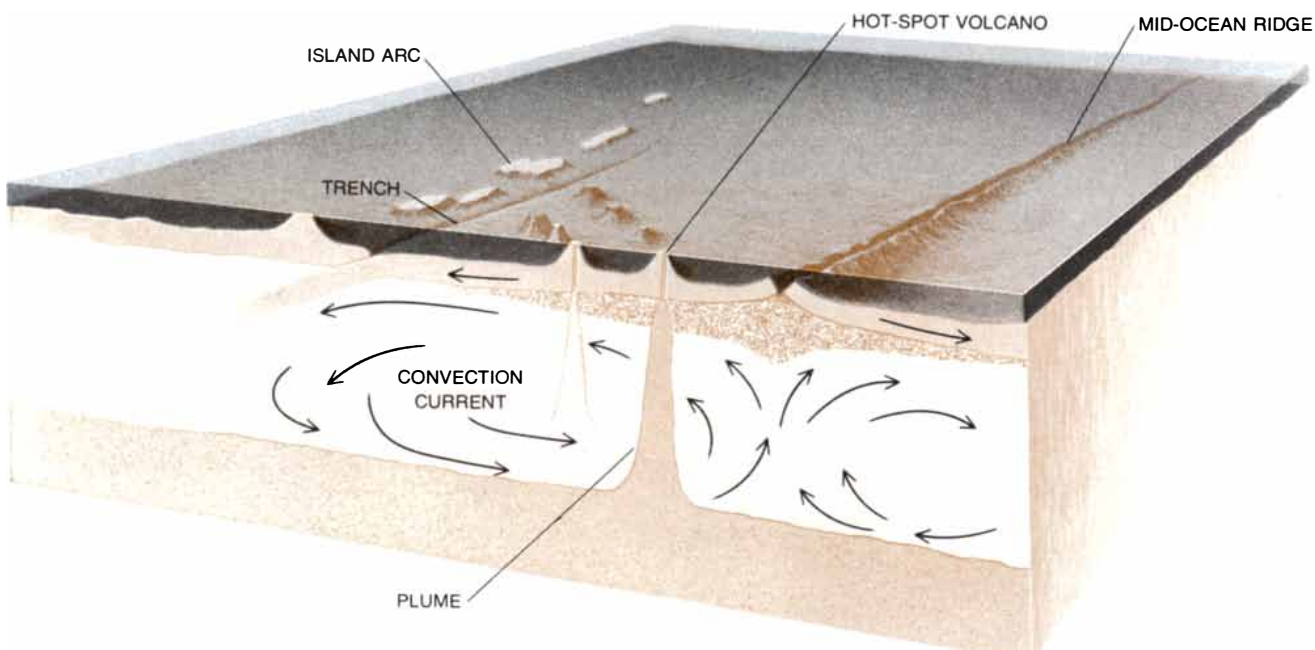
Almost all volcanic activity is confined to the margins of the plates. Along the full length of the mid-ocean ridges there is undersea volcanism in which the lava erupted is predominantly basalt. At convergent plate boundaries lavas are formed by the melting of lighter constituents of the subducted plate. The upwelling lava can create an island arc, such as the arcs of the Philippines, Japan and the Aleutians, or a volcanic mountain system, such as the Andes and the Cascade Range of the Americas. The lavas associated with convergent plates differ from the basalts of the mid-ocean ridges. They are called andesite lavas and they contain more silicon, calcium, sodium and potassium than basalt and less iron and magnesium.

Volcanism that is not associated with plate margins accounts for a small proportion of the world's volcanic activity, probably much less than 1 percent. It is these few isolated volcanoes that have been named hot spots. They are distin-

HOT SPOT in North Africa is an isolated group of volcanic mountains surrounded by the Sahara. The group is called Tibesti, and it lies in northeast Chad, near the Libyan border. The photograph was made from an altitude of about 920 kilometers by the LANDSAT earth-resources satellite. Dark blotches on the landscape are relatively recent lava flows. Two large, recent volcanic craters are visible; the one at the lower right is Emi Koussi, which has an elevation of 3,415 meters and is the highest peak in the region. Some older craters are recognizable, but they have been severely eroded. At Tibesti and at other African hot spots lavas of different ages are piled on top of one another, suggesting that the continent is stationary with respect to the hot spots and probably has been for 30 million years.



MOVEMENT IN THE EARTH is described by the theory of plate tectonics. The lithosphere, the cool and rigid layer that includes the crust of the earth, is broken into about a dozen large plates. These plates move over the asthenosphere, a layer that is hotter and capable of slow deformation. The crust is the top of the lithosphere; the rest of the lithosphere and all of the asthenosphere are parts of the mantle. Lithospheric plates move apart as new material is added to them along mid-ocean ridges. Where two plates come together one dives under the other and is reabsorbed; this process, called subduction, results in extensive volcanic activity. Hot spots are small volcanic regions typical of neither mid-ocean ridges nor subduction zones. Unlike most other volcanoes, they are often found far from plate margins, and even when they are near a plate margin, they can be distinguished by the volume of lava they eject and by its composition.



SOURCE OF A HOT-SPOT VOLCANO is thought to be a "plume" rising from deep within the mantle. Differences in composition between the lavas ejected at hot spots and those characteristic of plate-margin volcanism suggest that the two kinds of lava come from different parts of the mantle; indeed, the source of the hot-spot lavas may have been isolated for as long as two billion years. Much of the mantle is probably stirred by convection currents, so that the plumes must originate in some region isolated from this circulation. For example, they might come from a stagnant zone in the middle of a convection cell, or from a layer below the reach of the mantle currents. A lithospheric plate moving over a plume leaves a trail of volcanoes that grow older with distance from present site of volcanic activity.

guished by their very isolation: in the middle of a rigid lithospheric plate, far from centers of seismic activity, a hot spot may be the only distinctive feature in an otherwise monotonous landscape. Almost all hot spots are regions of broad crustal uplift, and this swelling is distinct from the smaller-scale mountain-forming or island-forming activity characteristic of all volcanoes. Finally, the lavas associated with hot spots differ from those found both at the mid-ocean ridges and at subduction zones. The hot-spot lavas are basalts, like those of the ocean ridges, but they contain larger amounts of the alkali metals (lithium, sodium, potassium and so on). Alkali-rich lavas are rare at plate margins.

The mechanism that generates hot spots must be sought in the mantle. They may be surface manifestations of "plumes": rising, columnar currents of hot but solid material. The plumes might well up from below the asthenosphere, at a phase-change boundary a few hundred miles inside the mantle. The distinctive composition of the hot-spot lavas argues that their source is isolated from the general circulation pattern of the mantle. For example, plumes might be generated in stagnant regions in the center of a circular convection current, or they might come from a very deep layer of the mantle, below the region that is effectively stirred by convection. The circulation of the mantle is still poorly understood, however, and for the moment any attempt to explain the origin of hot spots must remain speculative. Here we are concerned mainly with surface manifestations, which are not strongly dependent on the exact source of the magma. It is possible to formulate a consistent interpretation of hot spots even without a detailed model of the earth's interior.

Island Chains

Perhaps the most prominent and most easily recognized hot spot is the one that has formed the Hawaiian Islands. On an expedition to the South Seas in 1838, James Dwight Dana, an American geologist, noted that these islands become progressively older as one proceeds northwest from Kilauea and Mauna Loa, the active volcanoes on Hawaii itself. (Dana estimated the ages of the islands from the extent to which they had been eroded.)

It is now apparent that all the islands in the Hawaiian chain were created by a single source of lava, over which the Pacific plate has passed on a course proceeding roughly toward the northwest. The plate has carried off a trail of volcanoes of increasing age, in much the same way that wind passing over a chimney carries off puffs of smoke.

Dana also called attention to two other chains of Pacific islands whose trend is parallel to that of the Hawaiian chain. These are the islands of the Austral Ridge and the Tuamotu Ridge; the latter group includes Pitcairn Island. Like the Hawaiian Islands, these chains become older toward the northwest, and in each of them the most recent volcanic activity is near the eastern terminus. It would be difficult to ignore the inference that all three chains were generated as a result of the same plate motion. Indeed, from the configuration of the islands the apparent course of the plate can be mapped.

Leonhard Euler, the 18th-century Swiss mathematician, proved that on the surface of a sphere the only possible motions are rotations. It is therefore always possible to describe the movement of a lithospheric plate as a rotation around a pole. (The pole does not have to pass through the plate itself.) W. Jason Morgan of Princeton University has been able to show that the Hawaiian, Austral and Tuamotu chains could all have been generated by the rotation of the Pacific plate around the same pole. Employing a somewhat different approach, Jean-Bernard Minster and his colleagues at the California Institute of Technology have deduced from observed rates of sea-floor spreading that if the African plate has been stationary, then the Pacific plate must have moved along the trajectory defined by the Hawaiian chain.

At the western end of the Hawaiian Islands a string of submerged mountains, the Emperor Seamounts, strikes to the north. It is appealing to consider the entire system of islands and seamounts as a single chain that has changed direction, and age determinations support that interpretation. The oldest of the Hawaiian Islands, near the bend, are about 40 million years old. The Emperor Seamounts continue the age sequence without interruption, beginning near the bend with an age of 40 million years and continuing to an age of about 80 million years where the chain ends off the Kamchatka Peninsula. Morgan has found he can account for the formation of the seamounts by the rotation of the Pacific plate around a different pole, suggesting a remarkably simple sequence of events: about 40 million years ago the motion of the Pacific plate shifted to a new pole of rotation and thereby changed its direction of migration, causing an abrupt kink in the Hawaiian chain.

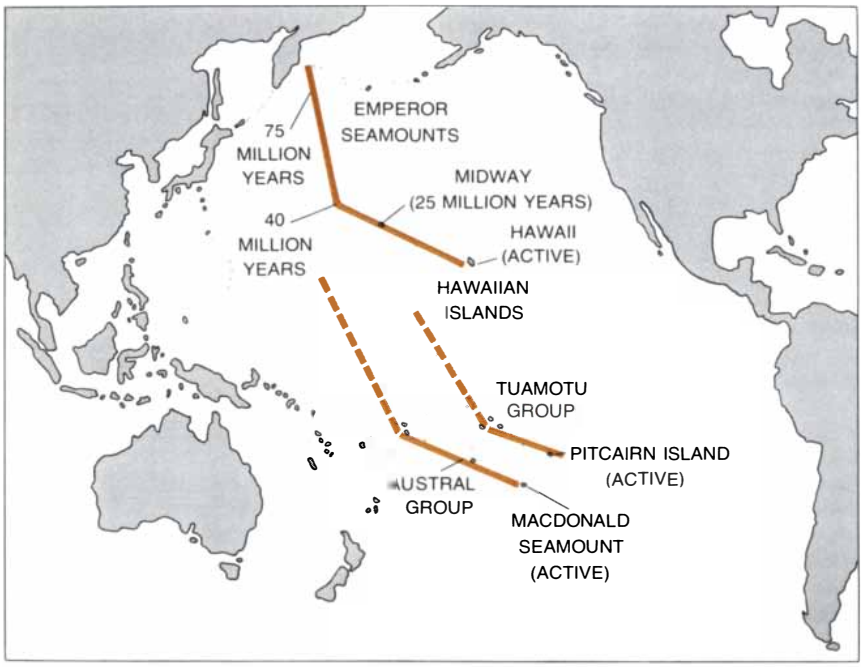
Another tantalizing inference from geography suggests a possible confirmation of this theory. The Austral and the Tuamotu island chains also seem to bend sharply at an age of about 40 million years, and each is continued in a line of seamounts. These seamounts are parallel to the Emperor system, and they could have been formed by the rotation of the plate around the same pole. For this conclusion to be accepted, however, it must be shown that the seamounts become progressively older to the north. As yet there are few dates established for the seamount series; those that have been obtained suggest a more complex interpretation.

The reconstruction of plate motions from the tracks of hot-spot volcanoes depends ultimately on the belief that the hot spots themselves are immobile or nearly so. This assumption appears to be justified. Minster and his colleagues have made accurate maps of relative plate motions by methods that do not rely on the hot-spot positions. Their work shows that prominent hot spots throughout the world have not moved in relation to one another during the past 10 million years. Other investigators have compared the positions of hot spots over much longer periods. Those determinations seem to show that groups of hot spots in one ocean have moved with respect to groups in other oceans over the past 120 million years—since the supercontinent Pangaea broke apart. This wandering of groups of hot spots, however, is slow compared with the shifting of the lithospheric plates.

The Population of Hot Spots

A census of the world's hot spots suggests that at least 122 have been active in the past 10 million years. Most of them meet all the particulars of the definition and can be classified without ambiguity. They are centers of volcanism that are not associated with plate boundaries and that form elevated domes with a diameter of up to about 200 kilometers. Also counted in the census, however, are several regions that lie on mid-ocean ridges or close to them; prominent among these are Iceland, the Azores and Tristan da Cunha, a small group of islands in the South Atlantic. The reason for the inclusion of these areas is that they seem more characteristic of hot spots than of normal mid-ocean ridges. The volume of material they have ejected greatly exceeds the norm for mid-ocean ridges; that is why they have been built up into islands while the rest of the ridge crest has remained submerged. More important, the lavas of these regions are the alkali-rich basalts that are rare at plate margins but that are typical of hot spots.

Our census probably underestimates the number of hot spots. There are domes or rises on some plates that are not capped by volcanoes; in spite of similarities in shape and geophysical properties we have not included them. There are probably also small, active volcanoes on the ocean floor that remain to be discovered. Finally, we have not attempted to include hot spots on con-



PACIFIC ISLAND CHAINS can be interpreted as tracks formed by the movement of the sea floor over stationary hot spots. The Hawaiian Islands grow older toward the northwest, beginning with Hawaii itself. Two other chains parallel to the Hawaiian Islands display a similar pattern of ages. They are the Austral group, which begins with the MacDonald Seamount, and the Tuamotu group, which begins with Pitcairn Island. All three chains could be generated by the same clockwise rotation of the Pacific plate. The age sequence of the Hawaiian Islands is continued in the Emperor Seamounts, which strike northward from a bend formed 40 million years ago. The change in direction implies that until then the rotation of the plate was centered on a different pole. Seamounts also extend to the north from the Austral and Tuamotu groups (*broken lines*), but there is no convincing evidence their ages form a linear sequence.

verging plate margins. In these areas volcanic activity is both abundant and complex, and it would be difficult to isolate the contribution of hot spots from other sources of volcanism. It should be noted, however, that basalts rich in alkali metals are found in some converging-plate zones.

Of the 122 hot spots we have identified, 53 are in ocean basins and 69 are on continents. Among the oceanic hot spots there is a tendency to congregate on mid-ocean ridges: 15 lie on the crests of ridges and nine others are near the crests. The greatest concentration, however, is in Africa. The African plate has 25 hot spots on land, eight at sea and 10 more on or near the surrounding ocean ridges, for a total of 43.

Even allowing for possible errors in our census, the inhomogeneity is striking. The African plate constitutes 12 percent of the world's surface area, but it has 35 percent of the hot spots. The large-scale topography of the African continent is also unusual. It is characterized by basins and swells, and in recent epochs South and East Africa have been greatly uplifted to produce highlands and the Great Escarpment. The topography and the abundance of hot spots are almost certainly related. Both can be explained by the hypothesis that Africa has come to rest over a population of hot spots.

The most compelling evidence that Africa is stationary is that at some hot spots lavas of several ages are superposed. If the continent were moving, of course, these lavas would be spread out in a chronological sequence. A few hot spots in the vicinity of Cameroon seem at first to be aligned like the island chains of the Pacific. It has been found, however, that these volcanoes are not arranged in chronological sequence. Their alignment is presumably a coincidence; it cannot have been caused by the motion of the plate.

Africa's basin-and-swell topography and the uplifting of large regions could be a direct result of the continent's immobility. Seismic studies have shown that the mantle is not homogeneous, and if there are variations in composition, there may be local concentrations of radioactive elements. The decay of these elements, which contributes a major part of the heat generated in the interior of the earth, would heat and expand some parts of the mantle more than others. The effects of the expansion might uplift regions of a stationary continent, but on a moving continent the uplift would be smeared out and would not be detected.

It is tempting to generalize from these observations, and it does seem there is a relation between the number of hot spots on a continent and the speed with which the continent is moving over the mantle. Antarctica, China and Southeast Asia, like Africa, have relatively large numbers of hot spots on land. Rates of sea-floor spreading imply that if Africa is stationary, these other regions are moving only slowly. In contrast, on rapidly moving plates, such as the North and South American ones, hot-spot volcanism is uncommon.

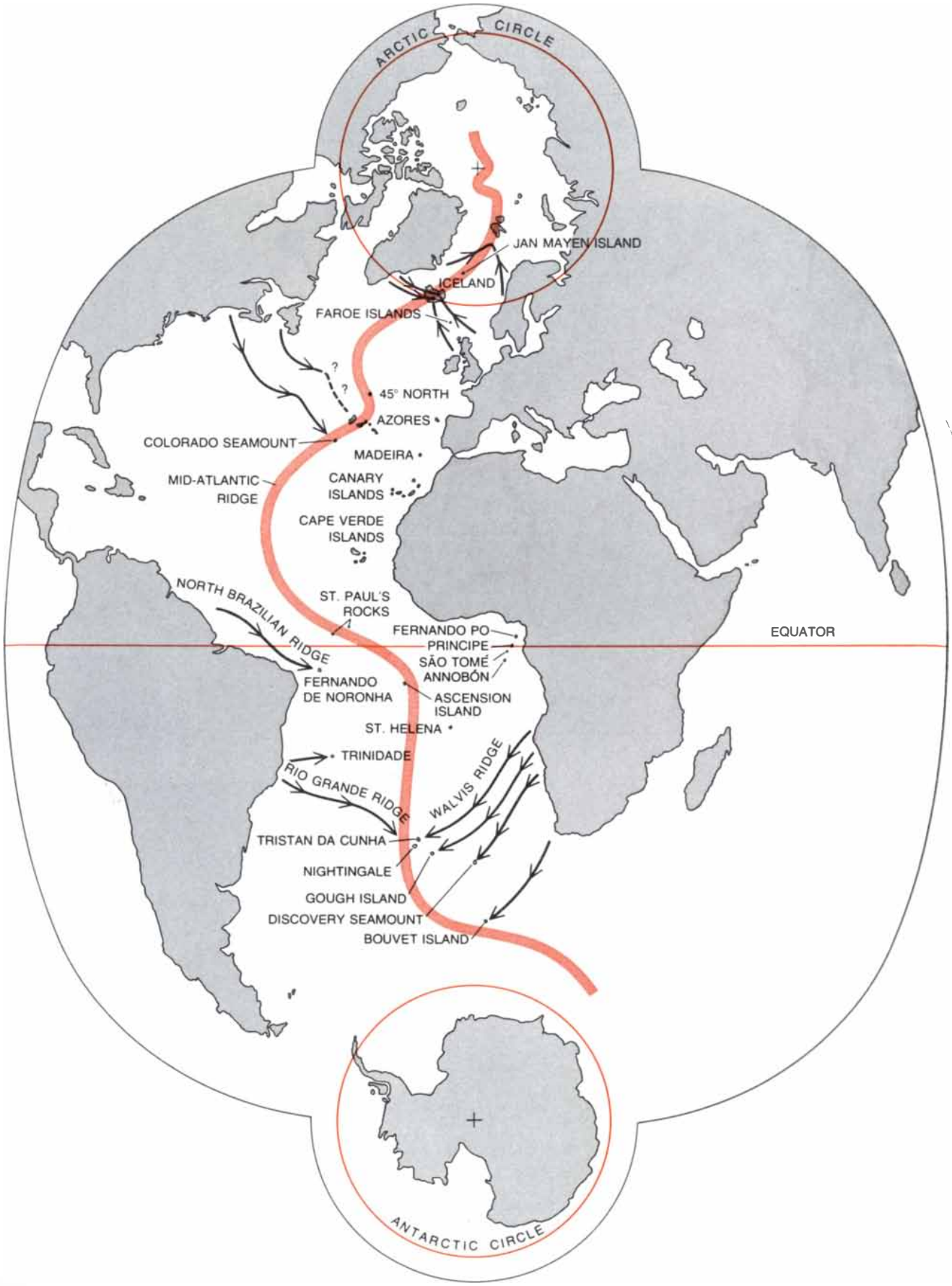
Opening of the Atlantic

Dated sequences of rock imply that Africa had numerous active volcanoes until the breakup of Gondwanaland 120 million years ago. The volcanic activity then ceased, and it did not resume until 30 million years ago. The two periods of activity and the long intermission between them can be read as signposts indicating the stages in the formation of the Atlantic Ocean.

The earlier episode of volcanic activity suggests that when Africa was a component of Gondwanaland, it was stationary over the mantle. When the supercontinent fractured along the present line of the Mid-Atlantic Ridge, Africa moved east. The motion over the mantle extinguished the volcanism for the next 90 million years. It is convenient to assume that the developing mid-ocean ridge was then stationary and that the two continents spread symmetrically away from it. They rotated in opposite directions around a pole near Cape Farewell on the coast of Greenland.

About 30 million years ago the African plate came to rest; volcanic activity on the continent resumed, and it has continued to the present. Although the African plate had stopped, sea-floor spreading had not. As a result the Mid-Atlantic Ridge was forced to begin drifting west. The relative motion of Africa and South America was unchanged, but the speed of the South American plate with respect to the mantle was doubled. When the Mid-Atlantic Ridge began its migration, the hot spots on the crest were left behind. Today a row of hot spots, which includes Tristan da Cunha and Ascension Island, is found a few

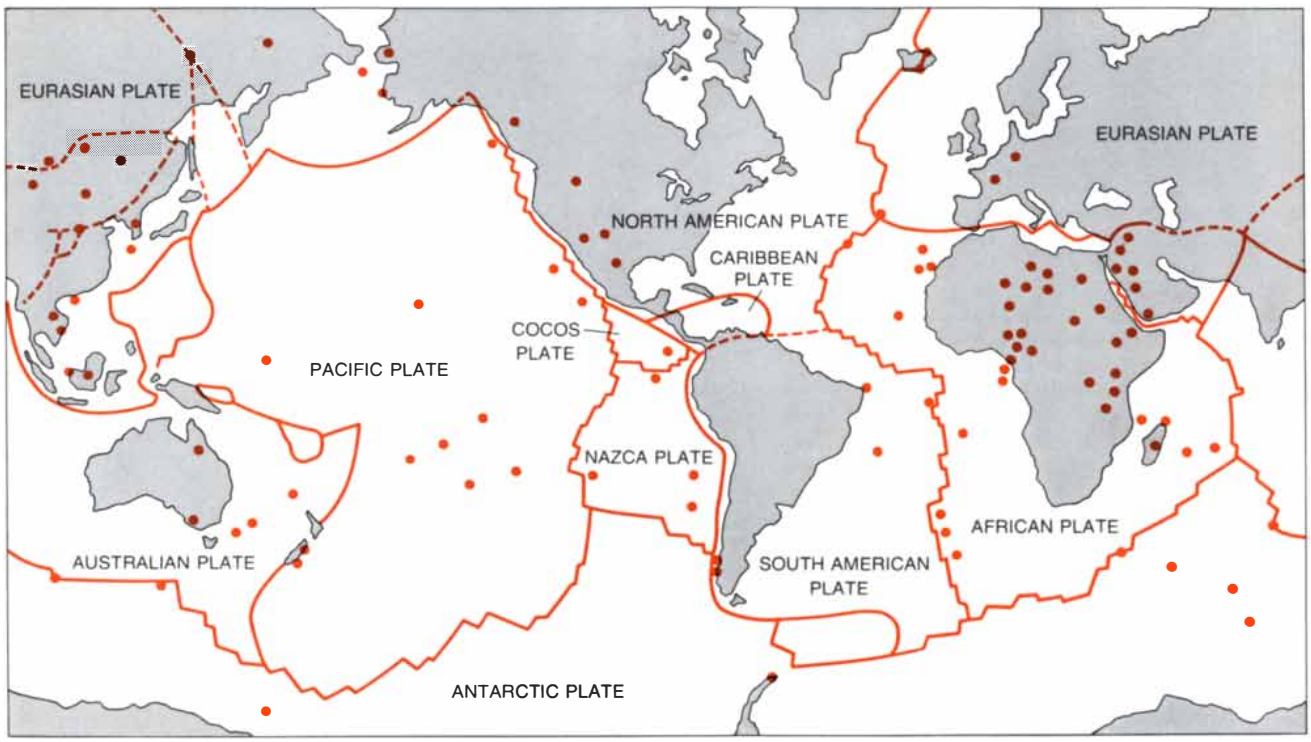
ATLANTIC HOT SPOTS also record the passage of the lithospheric plates. Several of these hot spots are on or near the Mid-Atlantic Ridge; a notable example is Iceland, which has been built up from the massive eruptions of volcanoes on the ridge crest. From some hot spots transverse ridges of volcanic rock extend back to the continental margins, indicating that the mid-ocean ridge developed over these volcanoes and that they were already active when the continents separated.

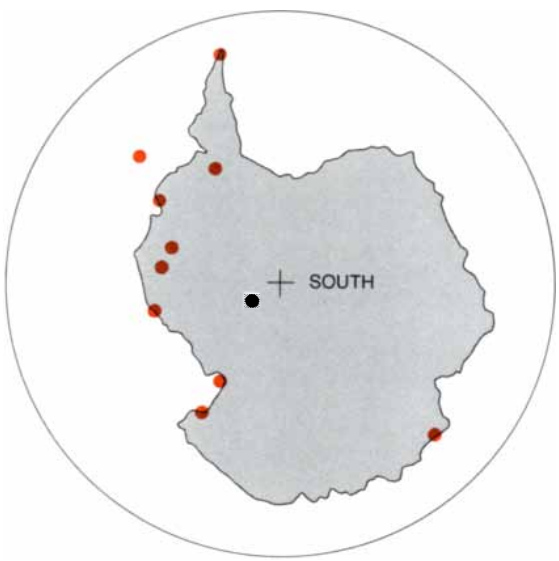


hundred kilometers east of the crest on lithosphere 30 million years old.

The evidence for the mobility of the Mid-Atlantic Ridge lies on the sea floor. From Tristan da Cunha a range of volcanic debris called the Walvis Ridge extends to the northeast. It is believed to be the track of the hot spot during the earlier part of the expansion (when the crest was fixed and Africa was moving), since it extends to lavas on the African coast that date from the disintegration of Gondwanaland. On the other side of the Mid-Atlantic Ridge another line of volcanic debris, the Rio Grande Ridge, extends to the Brazilian coast. There is no hot spot at its seaward terminus, which is separated from the mid-ocean ridge by a gap equivalent to 30 million years.

The disposition of these surface features can be explained by assuming that when the Atlantic was born, Tristan da Cunha was already an active volcano lying directly on the rift that opened to form the ocean. Lava from the hot spot overflowed onto both sides of the ridge and was rafted away by the spreading plates; continued eruption formed a *V*-shaped pair of tracks. When the mid-ocean ridge began to move west, the hot





POPULATION OF HOT SPOTS includes at least 122 that have been active in the past 10 million years. They are found on all the major plates and on both oceanic and continental crust, but their distribution is decidedly nonuniform. There is a concentration along mid-ocean ridges, and in particular along the Mid-Atlantic Ridge; what is even more conspicuous, of the 122 hot spots 43 are on the African plate. Together with other evidence, this abundance of hot spots suggests that the African plate is stationary over the mantle. If the African plate is adopted as a frame of reference, other areas that have many hot-spot volcanoes, such as Antarctica and Southeast Asia, are found to be moving only slowly; on fast-moving plates hot-spot volcanism is rare. The map is based on one prepared by W. S. F. Kidd.

spot was left behind on the stationary African plate. It could no longer produce a lateral ridge; instead its successive lava flows simply piled up, one on top of the other. Today at Tristan da Cunha young volcanic rocks are found along with lavas at least 18 million years old. Since lava was no longer deposited on the American plate, the Rio Grande Ridge was also terminated.

Because hot spots are particularly common along ridges and seem to exercise some control over their location, it is reasonable to suppose that the crest of the Mid-Atlantic Ridge will someday jump back to the hot spots it has abandoned. If it does, the 30-million-year gaps in the Walvis Ridge and the Rio Grande Ridge will remain as a record of the interlude.

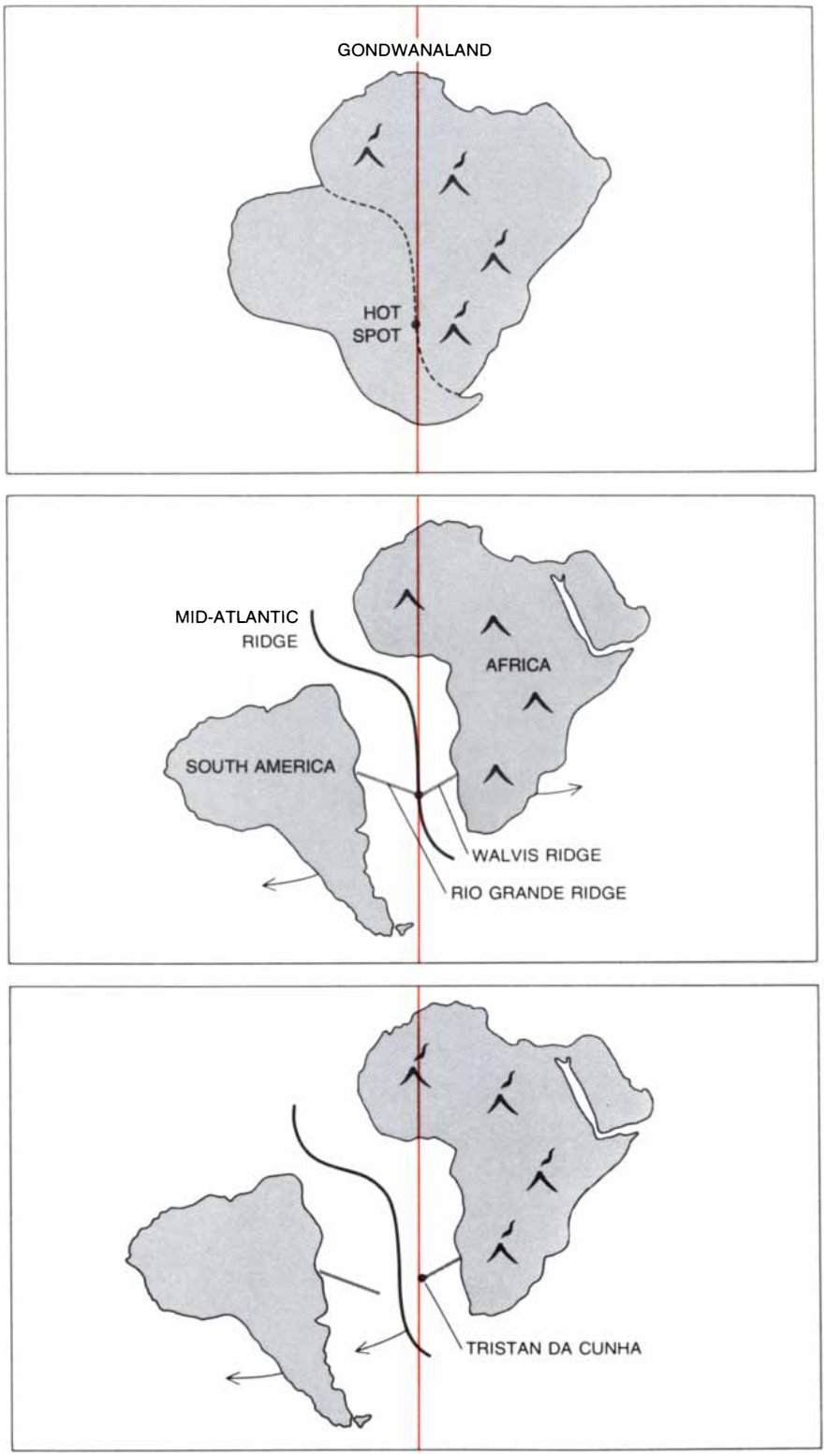
In the North Atlantic a somewhat different history can be read from the sea floor. The North Atlantic formed by the rotation of the Eurasian and North American plates around a pole in the Arctic Ocean. As we noted above, however, the American plate was already rotating around a pole in Greenland, near Cape Farewell, as a result of its separation from Africa. A single plate cannot rotate around two poles that are both fixed, and in this case the Arctic pole was itself in motion. The result was a shift in the position of the northern Mid-Atlantic Ridge.

When the North Atlantic began to open 80 million years ago, the locus of sea-floor spreading was west of Greenland. Spreading continued there until 50 million years ago and created Baffin Bay. An extinct hot spot left a pair of lateral ridges that trace this movement, extending to Disko Island in Greenland and to Cape Dyer on Baffin Island. Meanwhile 60 million years ago a new mid-ocean ridge developed east of Greenland. The continents have continued to diverge along that line since then.

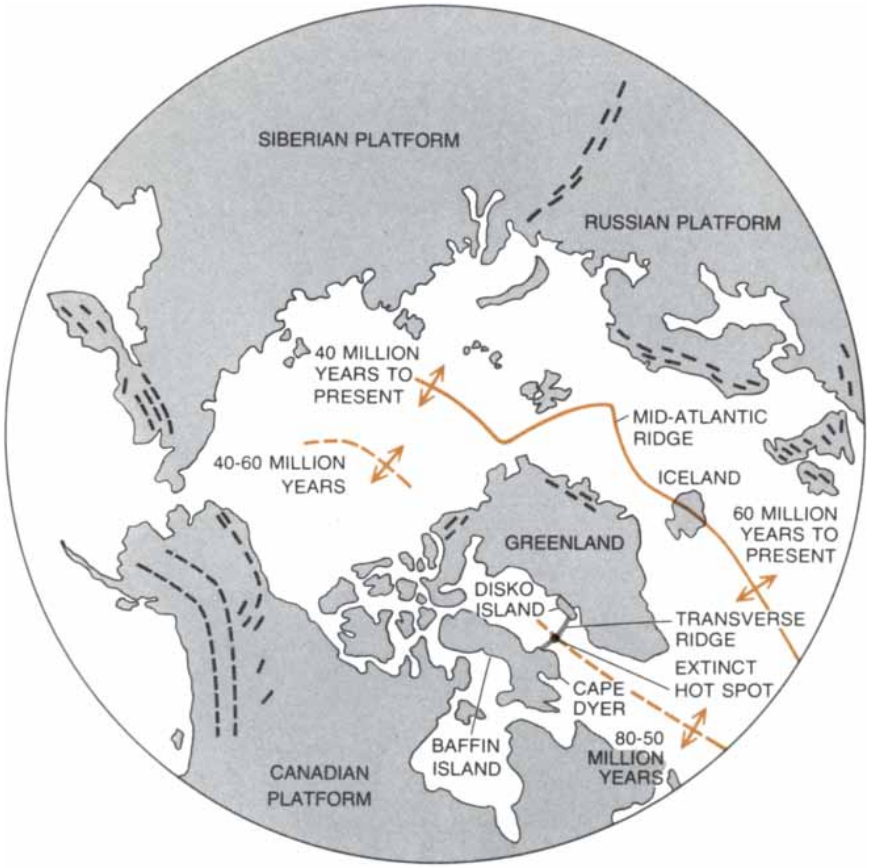
Motion at Subduction Zones

We have seen that hot spots provide a method for translating the relative motion of lithospheric plates into motion with respect to the mantle. This frame of reference has been employed in clarifying an important aspect of the behavior of the plates that had been imperfectly understood.

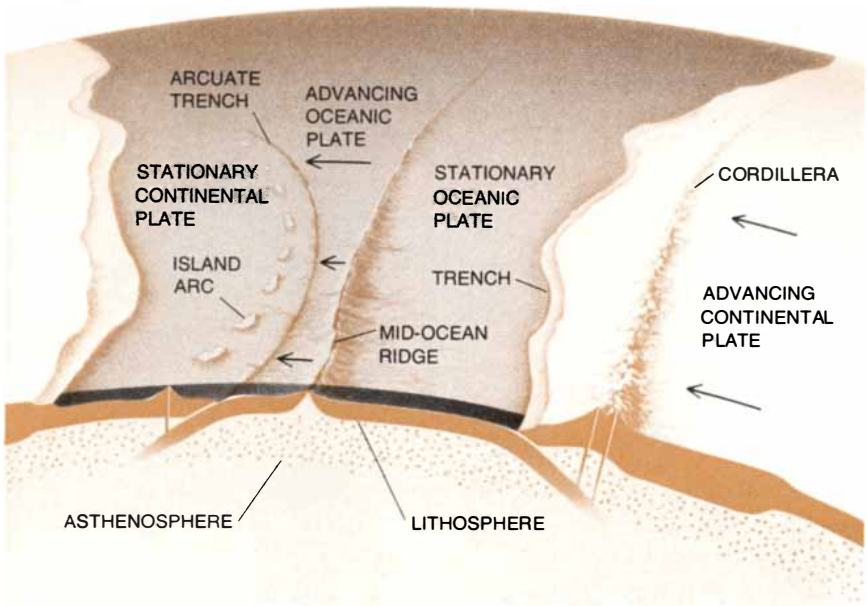
When an oceanic plate collides with a continental one, the oceanic plate usually dives toward the mantle and is subducted. That is because the continental plates are thicker and more buoyant. The partial melting of the sinking plate leads to volcanic activity above the subduction zone, but this activity can have two quite different surface expressions. In some cases an island arc forms offshore. The most prominent examples of this process are in South Asia, where



OPENING OF THE SOUTH ATLANTIC began 120 million years ago, when the great southern continent Gondwanaland broke apart. Until then hot-spot volcanoes had been abundant in Africa, suggesting that the continent was stalled over the mantle. When a fissure separated the continents, they receded symmetrically from the developing mid-ocean ridge and the motion of the African plate extinguished the hot spots. About 30 million years ago Africa came to rest and the present era of volcanism on the plate began. Because sea-floor spreading continued at the Mid-Atlantic Ridge, the ridge itself was forced to move west and the speed of South America was doubled. The ridge formed along a line that included several hot spots (*only one is shown*). As long as the ridge was stationary these hot spots generated trails of volcanic rock that extended back to the continental shores. When the mid-ocean ridge began its migration, the hot spots "fell off" the ridge crest and are now isolated on the African plate.



RIDGES IN THE NORTH ATLANTIC and Arctic oceans suggest that the locus of sea-floor spreading in this region shifted between 50 and 60 million years ago. Initially the continents separated along a ridge west of Greenland, opening Baffin Bay. An extinct hot spot has left a record of this movement in trails of volcanic rock that extend from the dormant ridge to Cape Dyer on Baffin Island and to Disko Island in Greenland. About 60 million years ago sea-floor spreading began at the present site of the Mid-Atlantic Ridge, which passes east of Greenland.



CONVERGENT PLATE MARGINS can assume two different forms. Where an oceanic plate is advancing on a stationary continent, the thin and flexible sea floor buckles offshore in a characteristic arcuate pattern; the volcanoes rising above the subduction zone create an island arc like the arcs of Japan and Indonesia. When a moving continent overrides a stationary oceanic plate, the descending slab of lithosphere is forced to bend at the coastline; as a result the volcanoes rise through the continent, forming a mountain system such as the Andes.

the subduction of the Indian-Australian plate has generated the Indonesian archipelago, and in East Asia, where the sinking Pacific and Philippine plates have produced the islands of Japan and the Philippines. In other cases the volcanic activity appears on the continental landmass. The Andes, for example, were thrown up by the subduction of the Nazca plate, and the Sierra Nevada of California and the Coast range of British Columbia derive from the subduction of the Pacific plate.

It has not been clear why the same process should have two dissimilar manifestations. By referring the plate motions to the hot spots we have attempted to resolve the question. Island arcs form when the continent is stationary over the mantle and the ocean floor moves under it; coastal mountain ranges are raised up when the continent overrides a stationary oceanic plate.

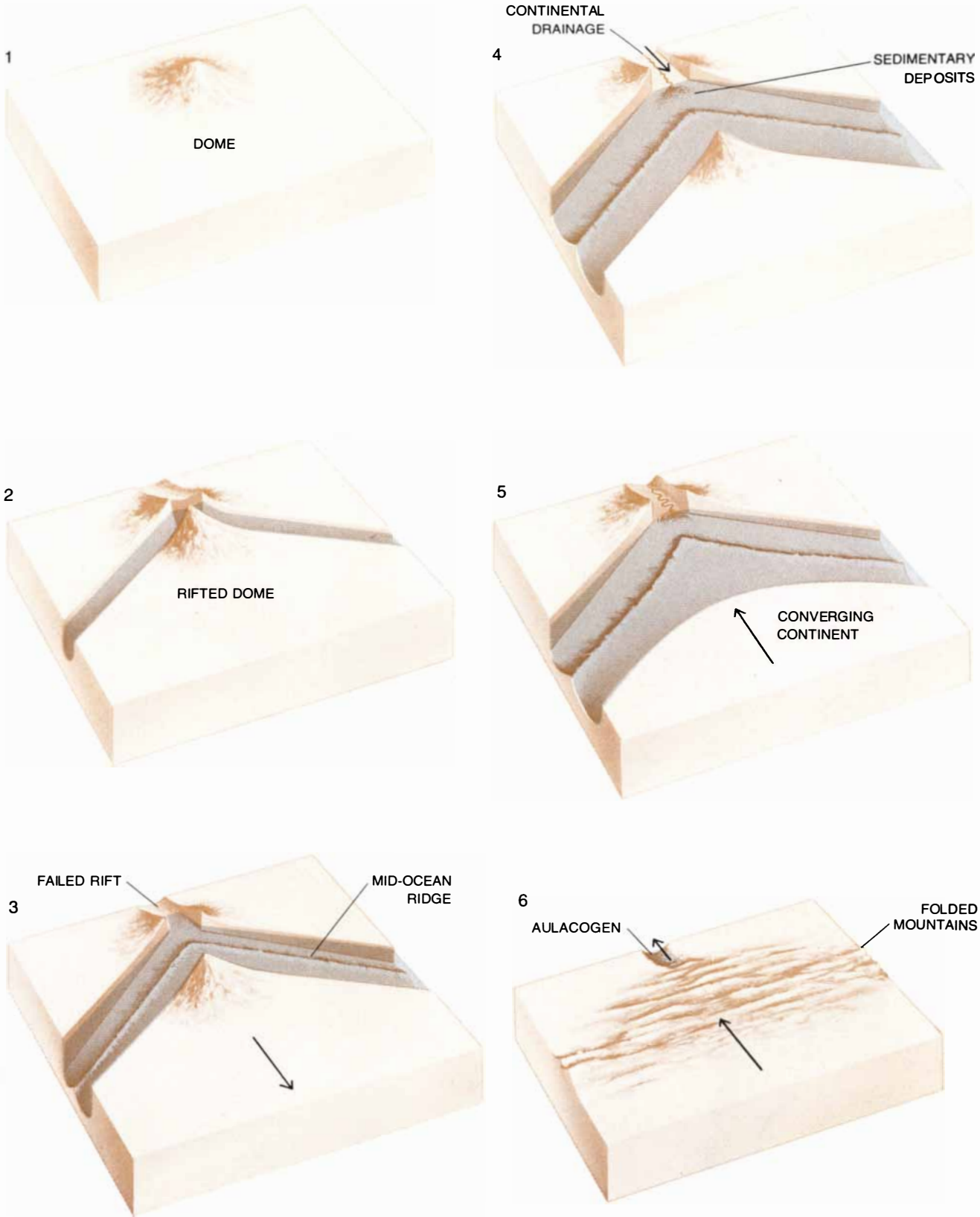
The only plausible explanation for the regular shape of island arcs has been proposed by F. C. Frank of the University of Bristol. He has pointed out that a flexible but inextensible thin spherical shell can bend in on itself only along a circular bend or fracture. This can readily be demonstrated by denting a ping-pong ball. It is suggested that where the ocean floor is moving and free to adopt the preferred shape offshore islands are formed in the characteristic arcuate pattern. When the continent is advancing, on the other hand, the oceanic plate is submerged before it can develop an offshore arc. The known motions of the plates in the Pacific region support this conjecture. The oceanic plates of the Pacific are advancing on Eurasia and underthrusting it, but they are being overtaken by the Americas.

Doming and Rifting

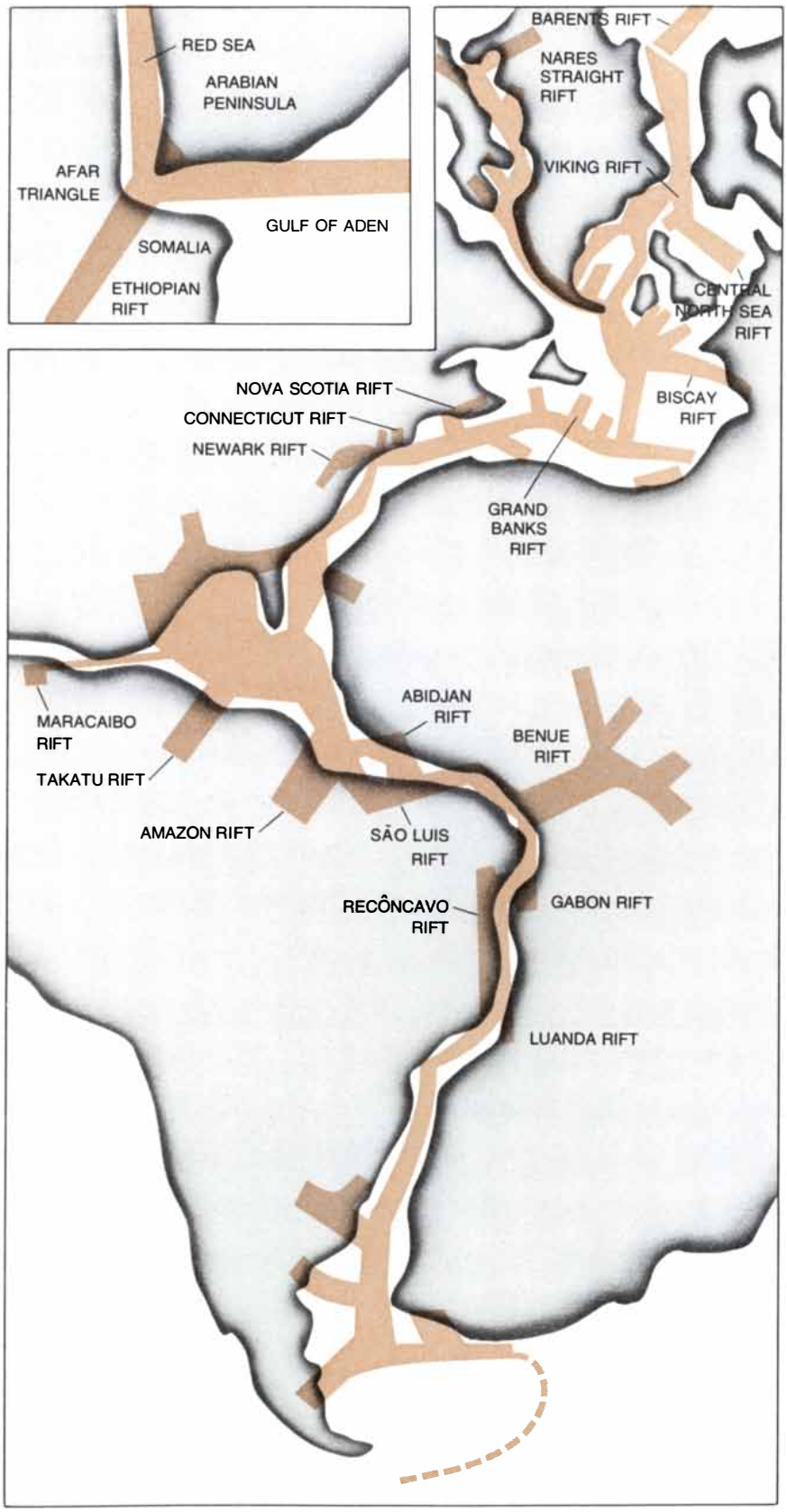
So far we have considered hot spots mainly as indicators of plate motion. They may also act to initiate cycles of tectonic activity.

When a continent comes to rest, the dome that swells up over a hot spot is subject to fracturing. When a rift appears, it very often has a characteristic three-armed pattern. Forty years ago Hans Cloos, a German geologist, recognized the prevalence of such three-armed rifts and showed that they are often related to doming of the continental crust. We would suggest that these rifts are often the seed from which an ocean grows. It follows that the ultimate cause of the rupturing of a continent may be the continent's coming to rest over the mantle. The hot spots appear to guide the fracturing, although they are not necessarily its only cause.

The observed concentration of hot spots on mid-ocean ridges would be accounted for if this mechanism is com-



DOMES AND RIFTS associated with hot spots may be involved in the fracturing of continents and the opening of oceans. A dome, often capped with volcanoes (*1*), forms on a continent that is at rest over the mantle. Rifts develop in the dome (*2*), frequently in a three-armed pattern. Two of the arms widen and eventually become the basin of an ocean (*3*), but the third arm fails to develop further. This failed rift may become a major river valley draining the continent and transporting sediments to the new sea (*4*). Later another continent approaches the site of the original rift and closes the ocean (*5*). The collision pushes up a belt of folded mountains, reversing the drainage pattern and carrying sediments back into the failed arm of the rift. Eventually the rift is filled; what remains is a trough of deep sediments roughly perpendicular to the mountain belt (*6*). Nicholas Shatsky of the U.S.S.R. has named such features aulacogens.



PREVALENCE OF THREE-ARMED RIFTS is revealed by reassembling the continents surrounding the Atlantic. In most cases two of the arms were incorporated into the Atlantic while the third remained a blind rift. A similar process can be observed today where the Arabian Peninsula is splitting away from Africa (*upper left*). The Gulf of Aden and the Red Sea form two branches of a rift; the third extends from the Afar Triangle into Ethiopia and Somalia.

monly involved in the formation of oceans. The breakup of Gondwanaland accords well with this interpretation; it will be remembered that Africa was stationary until the disintegration began.

Typically two arms of the rift open to form an ocean basin, but the third arm fails and remains as a fissure in the continental landmass. By restoring the margins of the Atlantic Ocean to their positions before Pangaea split apart, an abundance of three-armed rifts is revealed. The successful arms merged to create the ocean, whereas the unsuccessful ones remained as rifts extending into the continents. The best example of such a failed rift on the Atlantic coasts is the Benue Rift, which strikes away from the Gulf of Guinea into equatorial Africa.

A much more recent and more conspicuous example can be observed today where the Arabian Peninsula is splitting away from Africa. The Red Sea and the Gulf of Aden both represent arms of a three-armed rift. The third, dry arm strikes into Ethiopia from the Afar Triangle. The symmetry of the pattern is remarkable. The fact that Africa has been stationary over the mantle for 30 million years and that it bears extensive evidence of doming and rifting suggests that we could be witnessing the early stages in the disintegration of the African continent.

Aulacogens

The present cycle of tectonic activity, which dates from the breakup of Pangaea, is not the only one in the earth's history. The recognition that hot spots, domes and rifts form a sequence in the fragmentation of continental landmasses has led to the discovery of a clue that could prove valuable in attempts to reconstruct the earlier wanderings of the lithospheric plates.

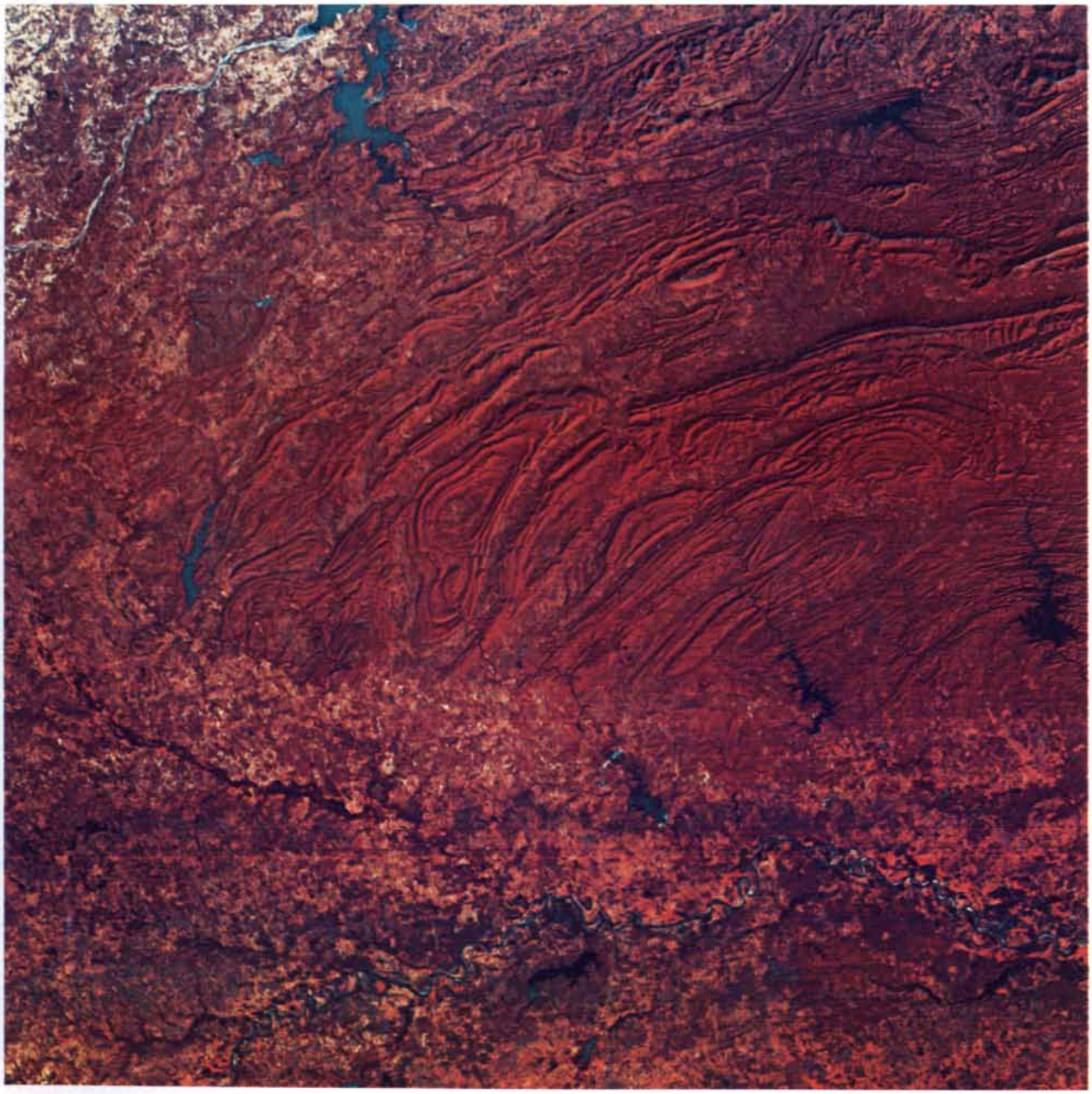
In 1941, as German forces threatened the main oil-producing area of the U.S.S.R., Nicholas Shatsky, a Russian stratigrapher, began a search for sedimentary basins that might contain new oil reserves. From the stratigraphic sequences compiled by Shatsky and his colleagues a previously unrecognized pattern emerged. Over much of the Russian and Siberian platforms the sedimentary layer is about a kilometer thick, but they found several narrow troughs up to 800 kilometers long where the sequence is three times the normal thickness. They named these formations aulacogens, from Greek words meaning "born of furrows." Aulacogens are rifts that extend from belts of folded mountains into continental platforms.

Aulacogens can now be recognized as failed arms of three-armed rift systems. When the two successful arms opened to form an ocean, the failed arm remained as a rift valley running inland from the

new seacoast. The rift became a feature of the drainage pattern of the continent, accumulating a thick deposit of sediments. Later another continent approached the coast, closing the ocean and blocking the rift. Compressional forces generated by the collision pushed up a chain of folded mountains. The remnant of the rift was a deep bed of sediments striking almost perpendicularly to the mountain chain.

The aulacogens that Shatsky recognized in the U.S.S.R. were of Paleozoic age (between 225 and 600 million years old). Paul Hoffman of the Geological Survey of Canada has since described a formation called the Athapuscow aulacogen that is two billion years old; it underlies the eastern arm of Great Slave Lake in northern Canada. Shatsky himself recognized what is probably the best-developed aulacogen in North America. It is a bed of sediments 15 kilometers deep in southern Oklahoma, parallel to the Texas border. It formed as a rift 600 million years ago when an ocean opened up roughly where the North Atlantic is today. The closing of that ocean was responsible for the building of the Caledonian, Appalachian and Ouachita mountains.

These ancient aulacogens are evidence that the cycle of continental disintegration and reassembly has been going on for at least two billion years. The development of domes and rifts in continents that come to rest over hot spots may have been a part of the process throughout this period.



AULACOGEN in southern Oklahoma is a remnant of an earlier cycle of continental drift. The photograph is a false-color image made in December, 1972, by the LANDSAT satellite. The aulacogen starts in the belt of flatland in the lower half of the photograph and extends another 400 kilometers to the west. To the north are the Ouachita Mountains. When a sea opened up to the south and east 600 million years ago, the aulacogen was the failed arm of a three-armed system of rifts. The closing of that sea raised up the Ouachita range as well as the Appalachians. Erosion of the Ouachitas has added to sediments already in the rift to form a layer of sediments 15 kilometers deep.

Polaroid Type 105 Positive/Negative Film is now a preservative agent: The Royal Photographic Society of Great Britain chose it to copy over 12,000 rare masterpieces (like the one below) from its Archives. Type 105 provides instant 3¼ x 4¼ inch positives for reference purposes. And negatives of superb quality for high fidelity reproductions. The negatives' high resolution permits great enlargement without loss of detail. And their wide tonal range reproduces every shading nuance of the orig-

FROM THE ARCHIVES OF THE ROYAL PHOTOGRAPHIC SOCIETY OF GREAT BRITAIN.

THE SHE

(1887, 18½" × 11½" P

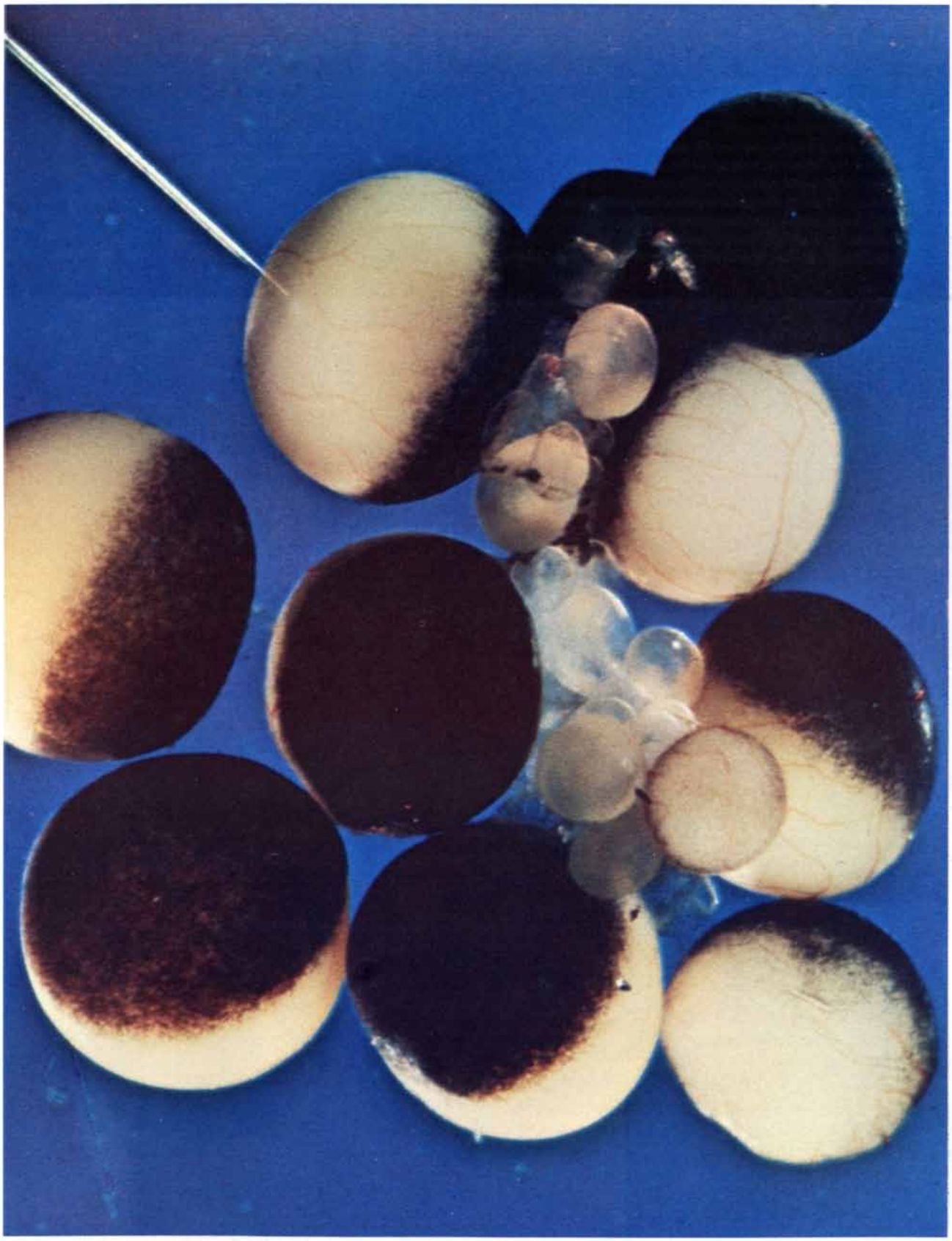
Photographed by

inal prints. Furthermore, since the Polaroid prints and negatives are produced in only 30 seconds, the original photographs can lie, untouched, under the camera until the archivist knows he has satisfactory copies. The irreplaceable originals can then be removed to safe storage under optimum conditions with no further handling risks. Thus, in a manner of speaking, some of the great pioneers in the history of British photography will owe their lasting reputations to a film made by Polaroid Corporation.

REPRODUCED FROM A POLAROID TYPE 105 NEGATIVE (3¼" X 4¼")

IMPERS.
ATINUM PRINT)

ry Peach Robinson



LIVING OÖCYTE of the South African clawed frog *Xenopus laevis* is microinjected with a foreign messenger RNA. The messenger codes for rabbit globin, and the unspecialized oöcyte, a precursor of the frog egg, will synthesize rabbit hemoglobin, which is ordinarily produced only in specialized red blood cells of a very different animal. The large, bicolored oöcytes are about a millimeter in diameter; the smaller pale cells are immature oöcytes. The sharpened glass needle (*top left*) with a tip about 20 micrometers (thousandths of a millimeter) in diameter, connected by oil-filled tubing to a syringe, can accurately deliver volumes of fluid as small as 10 nanoliters (billionths of a liter), and it is possible to inject more than 200 oöcytes an hour. The photomicrograph was made by the author and Neil Papworth.

Rabbit Hemoglobin from Frog Eggs

Messenger RNA from a specialized cell of one species is translated by the egg of another species. Such molecular biology experiments, done in living cells, give information on control of gene expression

by Charles Lane

What happens if a frog egg is given instructions, in the form of a particular molecule of the nucleic acid RNA, to make rabbit hemoglobin? The egg makes rabbit hemoglobin. What if the frog egg is supplied with the RNA molecule that specifies honeybee venom? It makes honeybee venom. Such results are not merely bizarre biochemical anomalies. They provide significant information about the expression of particular genes in particular cells. The control of gene expression is central to normal animal development and to such abnormal processes as cancer, birth defects and metabolic disorders of genetic origin. The experiments I shall describe are attempts to gain an understanding of that control by doing molecular biology in living cells. We inject various components of differentiated cells back into embryonic cells in order to study the molecular mechanisms involved in development.

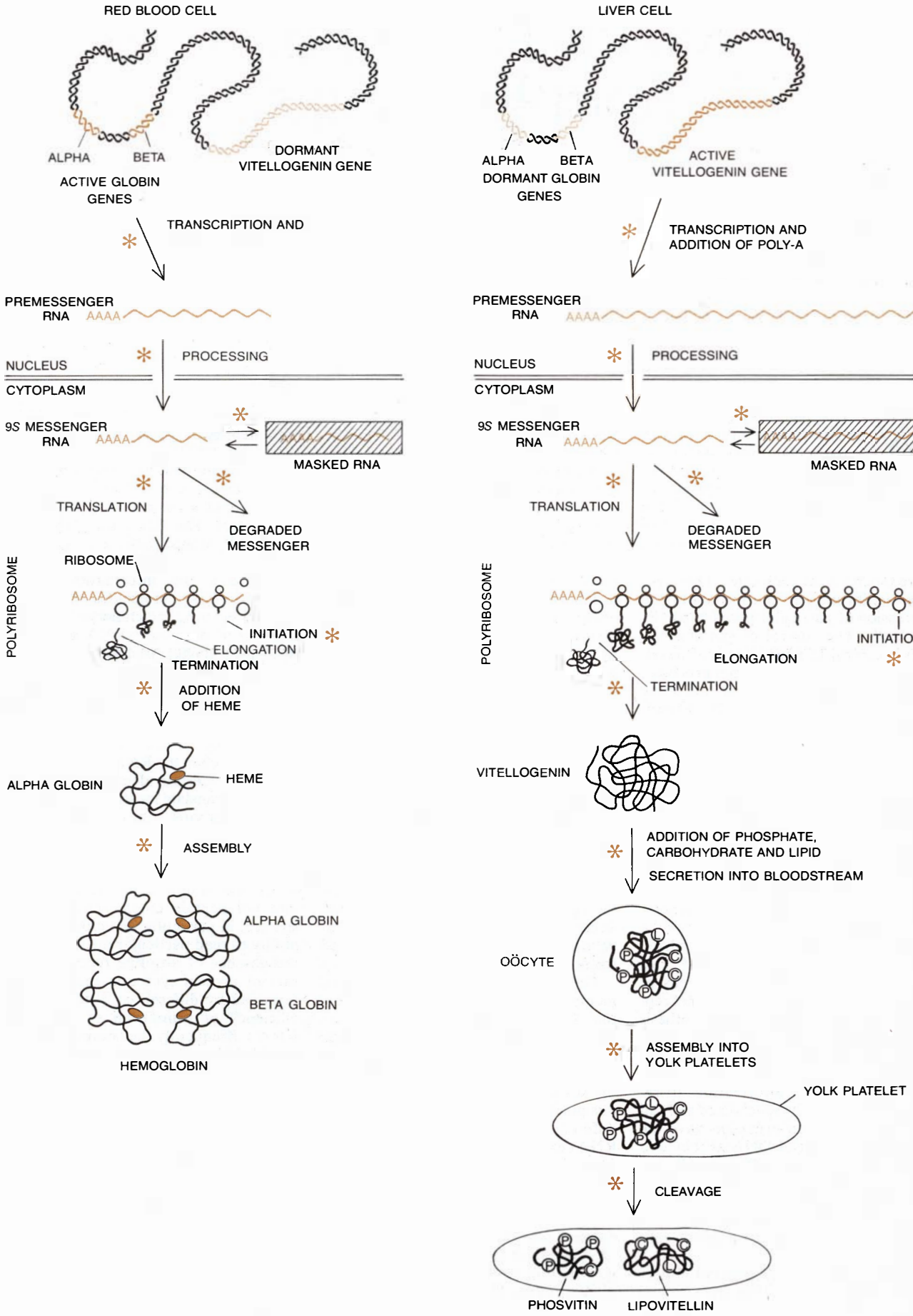
The fertilized egg contains, encoded in the nucleotide sequences of DNA molecules, all the genetic information that has been passed from one generation to the next and that will specify the nature of the many different specialized cells of the adult organism: blood cells, muscle cells, skin cells and dozens of other cell types, each with a distinct repertory and schedule of protein synthesis and therefore of functions. How do all those specialized cells arise? It is not that each cell type is allocated only the particular genetic information it requires to become its specialized self; all the cells of a given organism have the same complement of DNA as they embark on differentiation. If each cell contains a complete set of genes, it must be that different genes are expressed in different cells and at different times. How, then, is the expression of genes controlled?

Many of the early triumphs of molecular biology stemmed from the purification and structural analysis of such large molecules as those of DNA, which carries the genetic information, the various RNA's, which mediate the translation of the information into proteins, and some of the proteins thus produced. Next the function of such molecules was studied in cell-free systems: laboratory solutions whose chemistry is nicely adjusted to support many of the reactions that take place in living cells. Knowledge of the structure and behavior of molecules in laboratory glassware does not yield unequivocal information, however, on their nature and function in the normal living cell. It is better to work, if possible, in the cell itself. For example, the most dramatic confirmation that specialized cells contain a complete set of genes, and hence a full set of DNA molecules, was produced in 1962 by John B. Gurdon of the University of Oxford, who extracted nuclei taken from the intestinal epithelial cells of tadpoles and transplanted them into frog eggs from which the nuclei had been removed [see "Transplanted Nuclei and Cell Differentiation," by J. B. Gurdon; SCIENTIFIC AMERICAN, December, 1968]. Some of the eggs developed into normal, fertile adult frogs, which meant that epithelial cells must contain all the genes required for normal development.

Those nuclear-transplantation experiments were the technical and intellectual starting point for the microinjection experiments with which this article deals. Whereas transplanting nuclei into eggs establishes variable gene activity as the key to differentiation, injecting various purified macromolecules (primarily different kinds of information-bearing RNA) explores the means by which the normal living cell brings about this selective expression of genetic information.

There are many stages at which control might be exerted. The flow of information from the genes in the cell nucleus to the protein-synthesizing machinery in the cytoplasm is a complex process involving a multitude of molecular components, and the flow could be regulated, both qualitatively and quantitatively, at many points [*see illustration on next page*]. The transcription of the gene's DNA coding sequences into RNA provides one obvious opportunity for control. The RNA molecules that are initially transcribed (heterogeneous, or premessenger, RNA) are probably larger than the information-bearing messenger-RNA molecules that reach the synthesizing machinery; both the cleaving of messenger-RNA sequences from their longer precursors and their passage out of the nucleus might be subject to control. Thereafter specific components could combine with particular messenger RNA's and "mask" them; the genes coding for such RNA's would be expressed when the RNA was somehow unmasked. Certain elements of the systems that translate messenger RNA into protein—perhaps the transfer-RNA molecules that deliver the appropriate amino acid units, or the ribosomes (with their associated protein factors) on which the amino acids are assembled into polypeptide chains—could be restricted, in particular cell types, in their ability to read particular messengers. In the absence of absolute restriction the rate of protein synthesis could be determined by differences in the relative efficiency of translation of particular RNA's. Frequently the completed polypeptide chain does not become a functional gene product until after secondary modification of the amino acid sequence or assembly into a multiple-chain structure or even incorporation in a subcellular organelle; such modification and assembly could also be controlled to modulate the expression of genes.

We have been testing many of these possible control points by injecting material into the eggs or oöcytes (egg precursors) of the South African frog *Xenopus laevis*. They are easy to obtain in large numbers. They are enormous; a fully grown oöcyte, just before it is



transformed by a hormonal stimulus into an egg, is more than a millimeter in diameter, some 10,000 times larger than a frog liver cell. And it is resilient enough to withstand the injection of an amount of fluid equal to 5 percent of its volume.

In 1963 Gerard Marbaix, A. Burny and H. Chantrenne of the Free University of Brussels purified an RNA species they believed to be the messenger specifying globin, the protein component of hemoglobin. It was the expected size (nine Svedberg units) and was found linking polyribosomes, or groups of ribosomes, in rabbit reticulocytes: immature red blood cells that make hemoglobin and little else. Six years later the Belgian workers had still not been able to demonstrate biological activity in their 9*S* fraction, but that year Raymond E. Lockard and Jerry B. Lingrel of the University of Cincinnati managed to translate a similar 9*S* fraction in a cell-free system, a lysate of red blood cells. At that time Gurdon and I were experimenting with the injection into oöcytes of reticulocyte polyribosomes. Marbaix joined our laboratory at Oxford, and we undertook to inject the purified 9*S* fraction.

We dissolved the 9*S* RNA in a simple buffer solution and injected about 50 nanoliters (billionths of a liter) into each of 20 oöcytes. The oöcytes were incubated overnight with a radioactively labeled amino acid. Then an extract of the oöcytes, to which unlabeled rabbit hemoglobin had been added as a marker, was passed through a column of Sephadex, which separates molecules according to size. The same things were done with control oöcytes, which had been injected with buffer lacking globin RNA. The extract from RNA-treated oöcytes turned out to contain radioactive molecules that were in the same Sephadex fraction—were the same size—as the unlabeled rabbit hemoglobin molecules (which were identified by their color); the extract of the control oöcytes did not contain those radioactive molecules [*see illustrations on next two pages*]. In other words, the RNA-treated oöcytes appeared to be synthesizing spectacular amounts of hemoglobin.

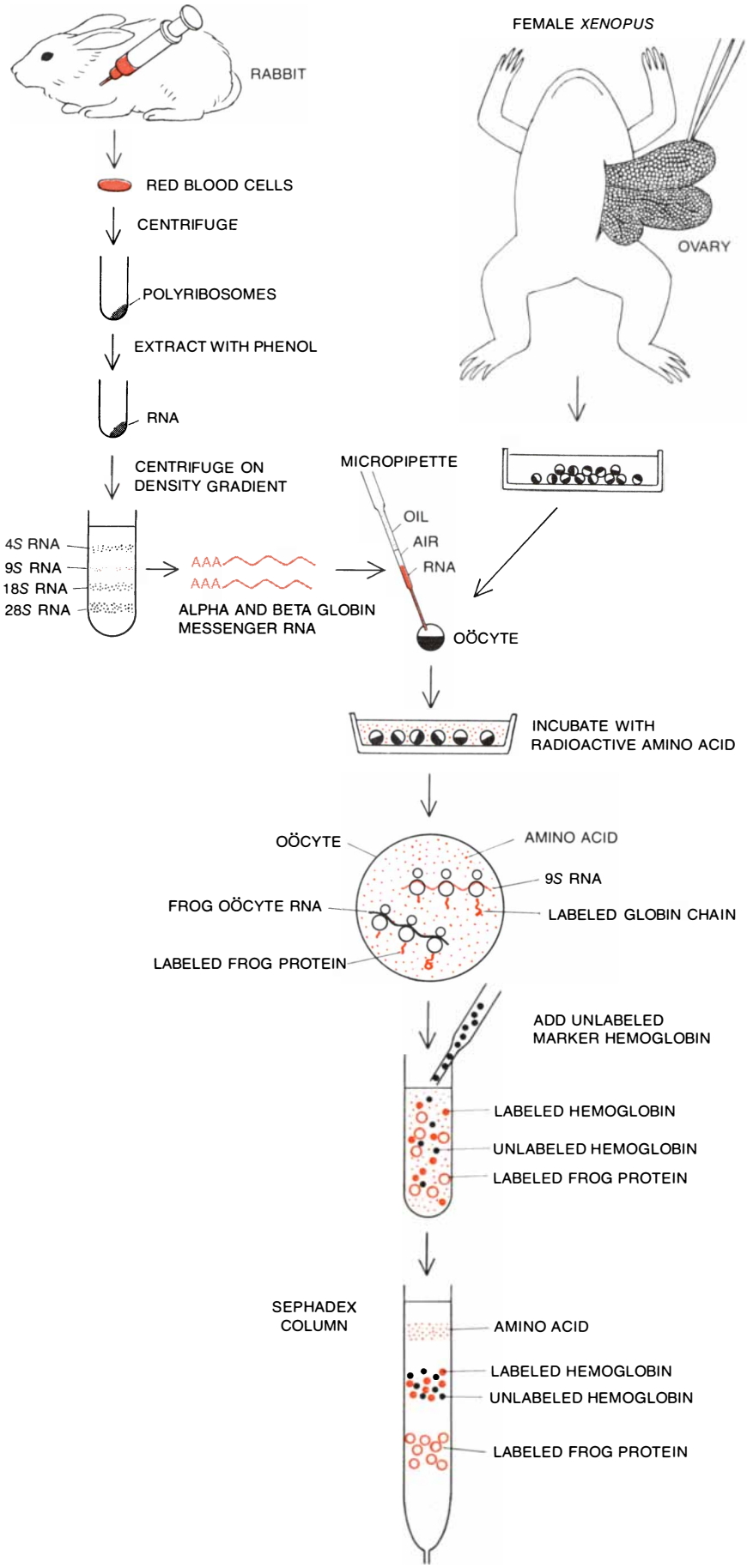
What kind of hemoglobin was it? Was it really the injected rabbit messenger RNA that was directing the synthesis or could the dormant hemoglobin genes of the oöcyte itself have been awakened? A hemoglobin molecule is a complex of two pairs of globin chains, usually called the alpha and the beta chains, and four heme groups. The two kinds of globin chains can be separated from each other by ion-exchange chromatography; moreover, globin chains of different animal species can be resolved into distinct, identifiable peaks on the chromatogram. We isolated globin messenger RNA from duck and mouse red blood cells as well as from rabbit cells and repeated the first experiment with each kind of messenger—injected the RNA into oöcytes, cultured the oöcytes with radioactive amino acids and added marker globin from the appropriate species—and put the extracts through a Sephadex column to partially purify the hemoglobin and then through an ion-exchange column to separate the globin chains. In each case the radioactive oöcyte-derived chains chromatographed with the appropriate marker globin chains; in the case of control oöcytes into which messenger RNA had not been injected there was no synthesis of radioactive globin chains of any kind [*see illustration on page 66*]. In other words, the response to an injected globin messenger depends on the species of the messenger: rabbit messenger directs the synthesis by frog oöcytes of rabbit alpha and beta globins, duck RNA specifies the production of duck globins and so on. The machinery of the unspecialized frog oöcyte can translate the biological information contained in messenger RNA from the highly differentiated, specialized rabbit, duck or mouse red blood cell; therefore at least some of the translation systems of the oöcyte can handle messenger RNA from both another cell type and another species; they are neither "cell type-specific" nor "species-specific."

It was just conceivable that the presence of a messenger from an alien species had broken down normal restrictions on the translation of a messenger from another cell type. We needed to rule out the possibility that there are cell type-specific translation restrictions that are themselves species-specific. Michael Berridge, Hugh R. Woodland and I prepared messenger RNA's for frog globin, frog egg-yolk protein and frog albumin, extracting and partially purifying the last two from the liver, where yolk protein and albumin are synthesized. When we injected them into frog oöcytes, the liver messenger directed the synthesis of either yolk protein or albumin and the red-cell messenger brought about hemoglobin synthesis. It was now strictly logical to conclude that some of the translation systems of the oöcyte are not cell type-specific.

If messenger RNA's from vertebrates as different as the mouse and the duck could function in the amphibian oöcyte, was it possible that RNA from even a different phylum could function there? Together with Ingela Kindas-Mugge and Gunther Kreil of the Austrian Academy of Sciences, we decided to test an insect RNA: the messenger for honeybee venom. The venom gland of the honeybee, *Apis mellifera,* synthesizes promelittin, a polypeptide of 34 amino acids that is subsequently cleaved to form the toxin melittin. Since the amino acid sequence of promelittin was known, the faithfulness of translation of an injected messenger could be rigorously tested. We isolated messenger RNA from venom glands and injected it into frog oöcytes. After incubation overnight we made an oöcyte extract from which we were able to purify a protein with the characteristics of promelittin. Apparently the translation machinery of the oöcyte is—again at least in part—not even phylum-specific. Moreover, the vertebrate system translates the arthropod RNA faithfully: most of the amino acid sequence of the oöcyte product was determined, and it was found to perfectly match the known sequence of the venom-gland product [*see illustration on page 67*]. The oöcyte does, however, appear to lack posttranslation modification systems that are present in the venom gland. An amino group (NH_2) is added to the amino acid (glutamine) at one end of the promelittin from honeybees, but the oöcyte product does not undergo this amidation. And the oöcyte promelittin, unlike the bee protein, is not cleaved to form melittin.

Workers in several laboratories have now tested a diverse spectrum of messengers, and the results indicate that any messenger RNA from a eukaryotic cell (a cell with a membrane-bounded nucleus) can be translated by the oöcyte. It is difficult to prove that a particular RNA cannot be translated, since negative results could be the consequence of messenger instability before injection, of degradation, modification or loss of the translation product or of other factors; it does seem, however, that only eukaryotic RNA's function in the oöcyte.

FLOW OF INFORMATION from nucleus to cytoplasm, the pathway leading to gene expression, might be subject to control at a number of points (*colored stars*). All genes are present in almost all cells, but different genes are expressed in different specialized cells: the genes for alpha globins and beta globins in red blood cells (*left*), for instance, and the gene for vitellogenin in liver cells of the female frog (*right*). The nucleotide sequences of the DNA (the genes) are transcribed into similar sequences of RNA. The initial transcript is probably into a precursor to which a string of adenylic acid nucleotides (*A*) is added; this premessenger RNA is processed to form messenger RNA, which leaves the nucleus and—if it is not masked or degraded—becomes attached to ribosomes and is translated into a polypeptide chain. The completed chain may be modified in various ways, as in the case of the proteins whose formation is illustrated. Two alpha and two beta globin chains are combined with one another and with four heme groups to form hemoglobin; the large vitellogenin chain is modified chemically and assembled into the yolk platelets in the oöcyte, where it is cleaved to form two yolk proteins.

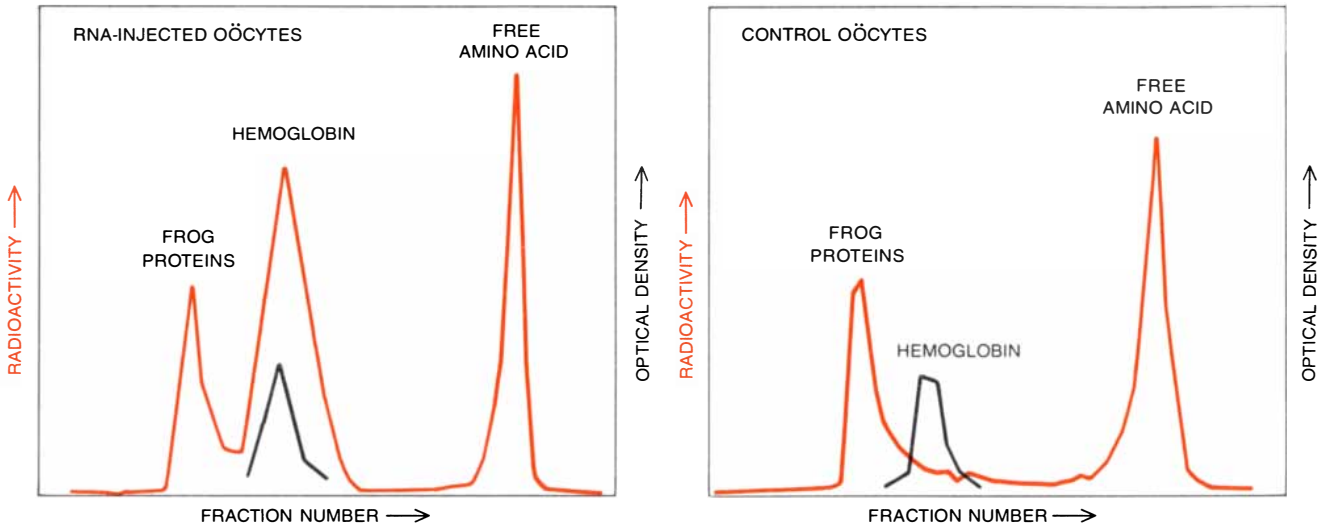


For example, synthetic polymers consisting of repeated single nucleotides or repeating polynucleotide sequences do not seem to be translated. Neither do viral RNA's if they are from viruses that infect bacteria (the latter are prokaryotes and lack nuclei) rather than plants or animals.

The very general nature of this translation system suggests, if it does not prove, that any of the oöcyte's protein-manufacturing assemblies can be programmed by any eukaryotic messenger RNA. One can, of course, argue that cell type-specific messengers require specific translation factors, and that the frog cell contains a complete set of such factors. Since promelittin has never been found in any vertebrate tissue, however, the idea that a factor specific for promelittin messenger RNA preexists in the oöcyte seems farfetched. The oöcyte might possibly contain factors that are specific for classes of messenger RNA's, and promelittin messenger might happen to fall in one such class. The generality of translation by the oöcyte is most simply explained, however, by the lack of any requirement for translation factors specific to particular messenger RNA's. This is not to say that message-specific factors might not exist in more specialized cell types.

The oöcyte provides more than a set of nonspecific translation systems. The translation machinery is not isolated in a test tube but is part of the living cytoplasm and is influenced by it; the polypeptide chains formed under the direction of a foreign messenger are exposed to the enzyme systems present in the cytoplasm of the injected cell. When we programmed oöcytes with messenger RNA from lens cells of the calf eye, we were surprised to see not only that the lens protein crystallin was synthesized in the oöcytes but also that an acetyl group (CH_3CO) was added at the amino (NH_2) end of the protein chain. This is the same modification that crystallin undergoes when it is made in the calf lens cell itself. The secondary-modification reaction that results in the

HEMOGLOBIN is synthesized in *Xenopus* oöcytes injected with messenger RNA from rabbit red blood cells. The messenger is extracted from reticulocytes, immature red cells, and is injected into oöcytes that have been removed from the ovary. The injected cells are incubated with a radioactive amino acid (*colored dots*), thus labeling the proteins that are synthesized. An extract of the cells, to which rabbit hemoglobin has been added as a marker, is passed through a Sephadex column. The column separates the various proteins according to size: radioactive frog-oöcyte proteins (*colored circles*), rabbit hemoglobin (*black disks*) and also a radioactive protein (*colored disks*) that subsequent analysis shows is hemoglobin (*see illustration on opposite page*).



OÖCYTE-SYNTHESIZED HEMOGLOBIN is identified by Sephadex chromatography. The various fractions, separated by size, are examined for radioactivity and for optical density, or color intensity. The charts show the results for oöcytes microinjected with rabbit messenger (*left*) and for control oöcytes that were injected with buffer solution only (*right*). In both cases a major peak in the optical-density curve (*black*) identifies the red rabbit hemoglobin added as a marker. Both radioactivity curves (*color*) have peaks for newly synthesized frog proteins and for unincorporated amino acid. The radioactivity curve for RNA-injected oöcytes, but not the one for control oöcytes, has another peak that coincides with the hemoglobin peak: the RNA-injected oöcytes appear to have synthesized hemoglobin.

acetylation of crystallin is not, in other words, confined to the highly differentiated lens cell. That some proteins are acetylated and others are not must stem from the specificity of the acetylating enzyme for particular amino acid sequences; provided only that those sequences appear (in this case on a polypeptide chain specified by a foreign RNA), the enzyme does its work. Secondary modifications of this type appear to be a consequence rather than a cause of cell differentiation. So far there has been no exception to the general finding that the secondary-modification reactions occurring in the specialized cells of vertebrates also occur in oöcytes, for example acetylation, hydroxylation, phosphorylation and polypeptide-chain cleavage. We expect to find certain exceptions, however. The inability of oöcytes to cleave insect promelittin into the secreted product melittin probably stems not from the invertebrate origin of the molecules but from a general lack in the oöcyte of the enzyme systems that are associated with external secretion.

The cytoplasmic milieu supports assembly processes as well as enzymic reactions. For example, the globin chains made under the direction of injected messenger RNA go on to combine with newly synthesized oöcyte heme; there is indirect evidence that some of the chains assemble, within the oöcyte, to form four-chain hemoglobin molecules. Injected messenger RNA for frog egg-yolk protein programs the synthesis of the giant polypeptide vitellogenin. The polypeptide is incorporated into yolk platelets and is cleaved by an enzyme into the main constituents of egg yolk, the proteins phosvitin and lipovitellin. Vitellogenin is normally made in the liver of the adult female frog and is then transported to the oöcyte; the experiment shows that the subcellular destination and processing of yolk proteins do not depend on the oöcyte's having absorbed vitellogenin from the bloodstream.

All this information narrows down the range of levels within which control must be mainly exerted. We have found that in the oöcyte the interaction between an available "free" messenger and the living cytoplasm leads automatically to translation, to secondary modification and even to assembly into multiple-chain proteins and into subcellular organelles. Other living cells may differ from the oöcyte in the precise nature of the controlling step, although many appear to be similar in that the flow of genetic information seems to be controlled at some level prior to the appearance of free messenger RNA; in the oöcyte, once the messenger is available the expression of genetic information follows as an automatic consequence. Qualitative regulation, then, implies the modulation of such mechanisms as transcription, the selection of particular molecules of premessenger RNA, the movement of messenger RNA out of the nucleus and the unmasking of stored messengers. The oöcyte and egg systems can be exploited for identifying and analyzing the critically important control elements that must operate at one of these pretranslation levels or more.

We found that the injection of either reticulocyte polyribosomes or complexes of messenger RNA with its attendant proteins leads, as the introduction of purified messenger RNA does, to the synthesis of globin chains. Tracing the possible control stages back toward the hemoglobin gene itself, the next logical question is whether or not the large messenger-RNA precursor has a globin-synthesizing capability. When Sau-Ping Kwan and I tested the globin-forming activity of precursor preparations derived from mouse and duck cells, the oöcytes did synthesize mouse and duck globins. The difficulty is that the oöcyte system detects minute amounts (less than a picogram, or a trillionth of a gram) of globin messenger, and so the precursor preparation must not be contaminated with even a trace of mature messenger. Contamination is hard to discern, because small messenger RNA's might aggregate to form what seems to be a large precursor molecule. We therefore mix in some rabbit-reticulocyte polyribosomes containing mature globin messenger; if what is intended to be a mouse or a duck messenger-precursor preparation makes any rabbit globin, it must be contaminated with mature messenger. Unfortunately that has been the case in all the experiments to date, and so we can only say it is probable that the information in messenger-RNA precursors can be translated by the oöcyte. The point remains to be proved.

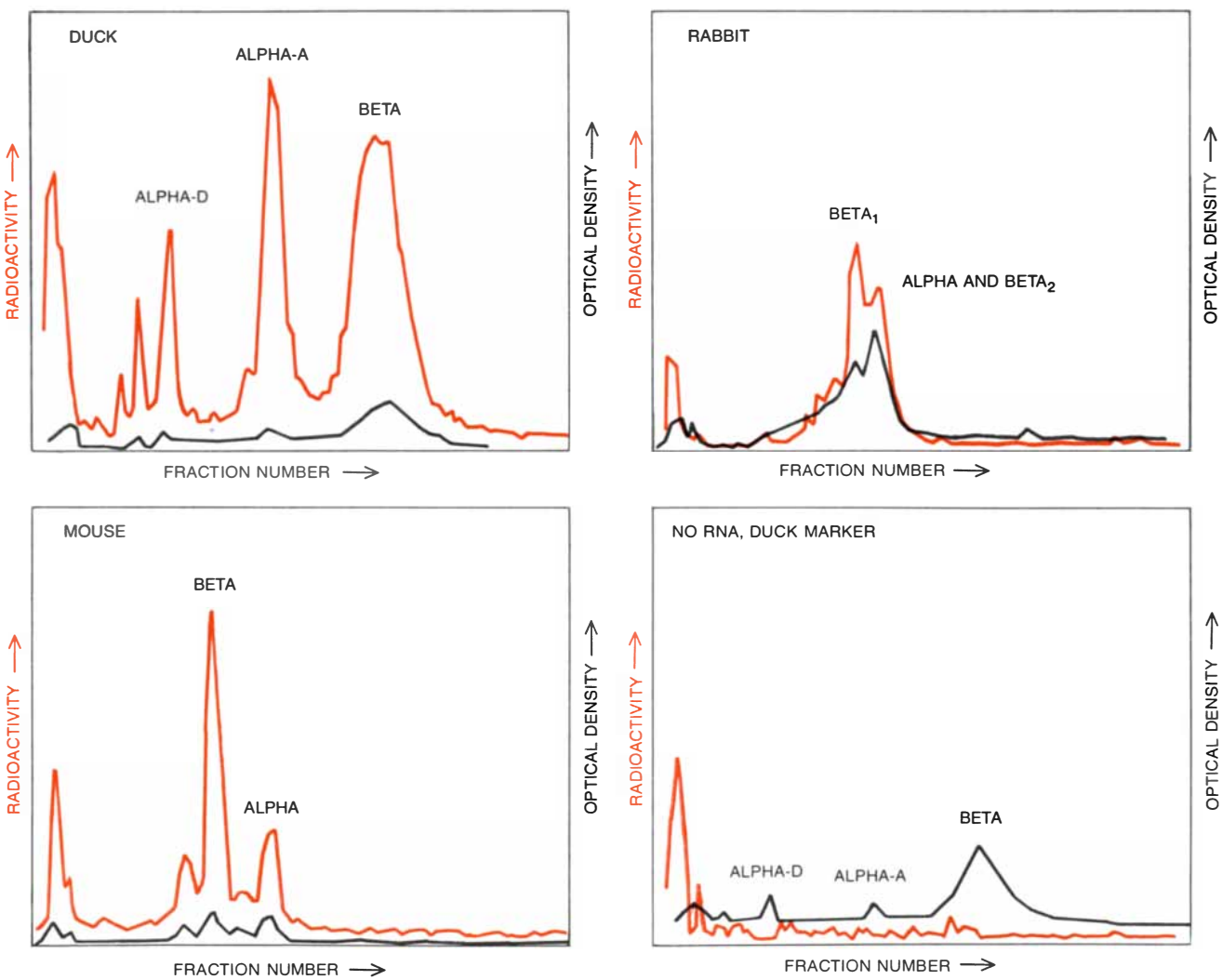
The mass of evidence accumulated from other developing systems suggests that control of gene expression at the level of messenger-RNA synthesis is of primary importance. Microinjection experiments, by showing the lack of qualitative control once messenger is available, also point to events occurring within the nucleus. The unmasking of preformed messengers, however, may also play an important regulatory role,

particularly during early development. Oöcytes probably contain a stockpile of inactive messengers in addition to the ones that are actively engaged in the synthesis of proteins.

The fact that there are nonspecific translation systems within eggs and oöcytes does not in itself prove that the machinery of more specialized cell types is equally general; the differentiated state could be stabilized, or even created, by selectivity at the level of the translation machinery. However, in an elegant experiment Gurdon, Woodland and Lingrel have produced muscle cells that make hemoglobin and have thus gone some way toward showing that even highly specialized cells can translate any messenger. In their experiment fertilized eggs are injected with mouse or rabbit globin messenger. Many of the eggs develop normally, right up to the swimming-tadpole stage. If radioactive amino acids are supplied at any stage of development, one can detect the synthesis of globin chains. Moreover, the response is species-specific: tadpoles raised from eggs injected with mouse messenger make mouse globin and tadpoles raised from eggs containing rabbit messenger make rabbit chains. At least some of the injected messengers must be stable and functional in developing cells.

How are globin messengers segregated, or apportioned as the cells divide, in the growing embryo? This question was answered by dissecting tissues out of a tadpole that was raised from an RNA-injected fertilized egg and then culturing the tissue fragments in a medium containing radioactive amino acids. Woodland, Gurdon and Lingrel found that several different tissues and regions of the tadpole made globin chains at similar rates. That is to say, the messenger injected into the frog egg seems to be divided roughly equally during the course of development, a finding that is not easily accommodated by theories of cell differentiation that rely on the unequal distribution of an original stockpile of egg RNA's.

More important, even muscle tissue isolated from tadpoles that were raised from eggs injected with messenger RNA from rabbit or mouse cells respectively made rabbit or mouse globin chains. More than 85 percent of the cells in these bits of tissue were either specialized muscle, nerve or backbone cells, and yet their rate of globin synthesis was comparable to that of less specialized regions of the tadpole. One might argue (but not without invoking complex selection mechanisms) that in the egg or during the formation of the embryo the globin RNA's had picked up specific



ABILITY OF OÖCYTE to translate the information encoded in a foreign RNA was established by repeating the experiment with RNA's from three different species and by breaking down the hemoglobin into globin chains, which can be identified by species. After injection with either duck, rabbit or mouse messenger RNA and incubation with radioactive amino acid, the oöcytes were mixed with the appropriate hemoglobin and an extract was made and passed down a Sephadex column. The partially purified hemoglobin thus obtained was subjected to ion-exchange chromatography. Radioactivity (*color*) and optical-density (*black*) curves showed that injected oöcytes produced globin chains that chromatographed with the appropriate marker chains; when no foreign RNA had been injected, there was no synthesis of globin chains (*chart at bottom right*). Note that ducks have two kinds of alpha chain, certain rabbits two kinds of beta chain.

One of a series of reports on the first hundred years of the telephone.

Why the cost of telephone service has gone up less than the cost of almost anything else.

In the late 1920's, in a Chicago factory, the history of industrial relations reached a turning point. The plant, the Hawthorne Works of the Western Electric Company, made telephones and telephone equipment for the Bell System. And in 1927 its managers had a puzzle.

The Hawthorne Works of the Western Electric Company, scene of what one economist called "the most exciting and important study of factory workers ever made."

For more than two years the company had been studying plant lighting and its relation to efficiency. (It was the era of the "efficiency expert" and "scientific management.") Increases in illumination were followed by increases in production, as expected. But decreases in light levels were also followed by increases in production. Two young women even maintained good production under light no brighter than moonlight.

It became clear that light had only a minor effect, and that there were many other variables to be identified. To solve the puzzle, the company undertook a further study, carried on jointly with the Harvard Graduate School of Business Administration. The researchers selected a group of six competent, experienced women, explained what they were trying to do, and requested their cooperation. Over a period of twenty-six months, the researchers added rest periods and snacks to the group's work schedule, shortened the work day, and then returned to the original schedule. The group showed an almost unbroken rise in average hourly production and also in total weekly production, even when the week was five hours shorter than at the start. At the end, their production was 30 percent above the beginning level.

The Hawthorne Experiments made it clear that the "scientific management" theory of the day relied too heavily on methods borrowed from the physical sciences. Two major conclusions are widely accepted now, but then their application to factory work was new:

People work better when they feel they are part of a team.

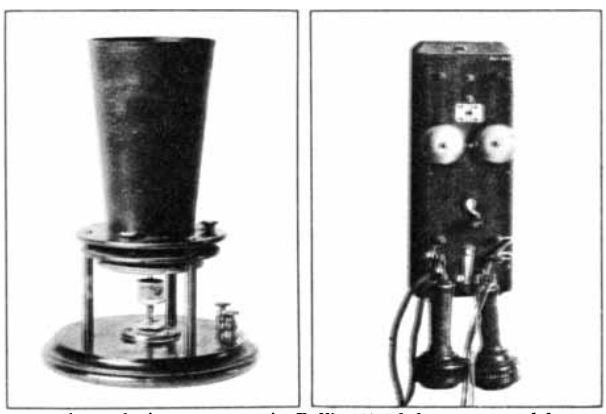
People work better when they feel what they are doing is important.

Today most businesses are aware of "group dynamics" and "job enrichment." But the lessons of Hawthorne have shaped

policy for many years, not just in Western Electric factories but in all parts of the Bell System. And the Bell System is still a leader in the exploration of factors affecting industrial productivity.

The prices of most telephone equipment made by Western Electric currently average 20 percent below the prices of other suppliers. Why?

New products for the Bell System are usually designed at Bell Laboratories, the research and development arm of the System. Bell Labs also sets quality standards. But at an early stage manufacturing engineers from Western Electric sit down with the designers and look for ways to save. All companies know that's the best time to cut costs; in the Bell System, that knowledge guides practice. The search for reduced costs continues after production begins. Every aspect of manufacturing is under constant reexamination. Western Electric's engineering cost reductions alone totaled $198 million in first-year savings in 1975.



An early improvement in Bell's 1876 laboratory model phone was the addition of a bell.

Yet none of these achievements would take place unless the people involved were convinced that it is important to produce good telephones at low cost. Phillip S. Babb of McKinsey and Co., management consultants, made this analysis in an interview published in the journal *International Management:*

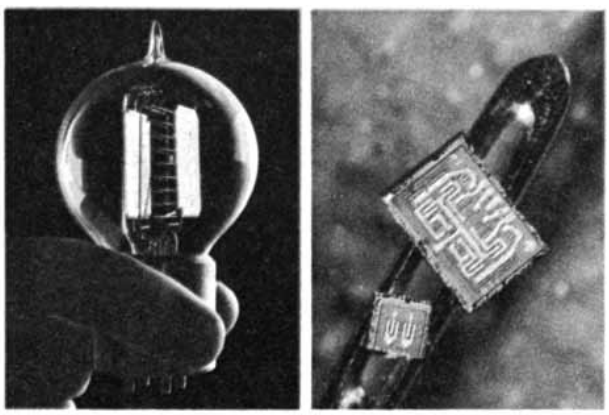
> Western Electric has succeeded in making cost-cutting a central part of the ethos, the value system, by which their people live. Driving costs down—with retained high quality—is what they spend their working lives at. It is what they take their pride in. It is their way to the corporate top.

To put it another way, the business of the Bell System is providing good telephone service at reasonable cost; Western Electric's activities are directed toward that service goal, rather than toward simply making products.

That service goal characterizes all parts of the Bell System, including the twenty-three regional operating companies and the Long Lines Department. All the parts work closely together to achieve that goal, and all benefit as a result. The operating companies provide telephone service and report, through AT&T, to Bell Labs and Western Electric their needs and the needs of telephone users. Bell Labs and Western Electric design and manufacture equipment to meet known needs as well as the best estimates of expected needs. And the local companies are assured of having the products customers want. To use the vocabulary of the economist, vertical integration with organizational feedback enhances productivity throughout the Bell System.

Touche, Ross & Co., acting as consultants for the staff of the Federal Communications Commission, made a study of how this corporate structure affects costs for telephone service. According to their report, written in 1974:

Western Electric's efficient performance has resulted in lower costs than otherwise would have been incurred. Because of Western's pricing policies and practices, these lower costs have not increased profits, but have been passed on to operating companies in the form of lower



Western Electric's first commercially successful vacuum tube (left) was used in 1915 in the amplifiers that made possible the first transcontinental telephone call. Solid-state electronics, begun at Bell Labs, makes it possible for tiny integrated circuits (right) to do the work of many vacuum tubes.

prices....The effect of the interrelationship between Bell and Western Electric is to operate Western, not as a manufacturing concern, but as an integral part of a vertically integrated communications firm. These interrelationships result in a favorable impact upon Western's costs, prices and service to operating companies.

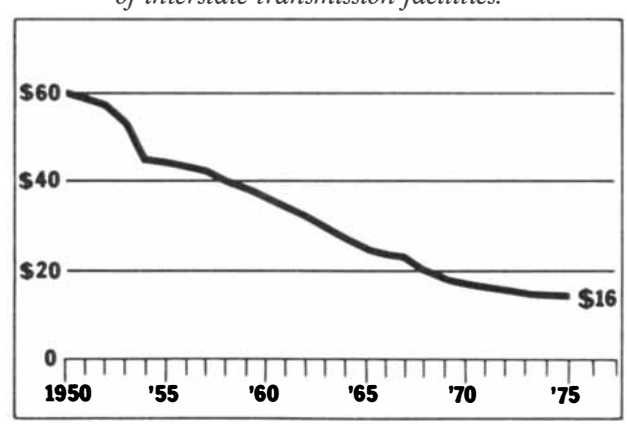
Another major factor affecting productivity is investment in new technology—in research and development. Bell Laboratories is recognized worldwide as one of the leading development and research institutions anywhere. The achievements of Bell Labs people have won two Nobel prizes, one for the demonstration of the wave nature of matter, and the other for the invention of the transistor.

The search for new and better technology has always been a part of the telephone industry. On March 10, 1876, Alexander Graham Bell and Thomas A. Watson achieved that famous first telephone message, "Mr. Watson, come here. I want to see you." At once they began to improve the instrument, and make it more usable. The goal, then as now, was to provide good telephone service at a price almost every American could afford.

The effect of research and engineering on costs can be seen most readily by considering how it has changed methods of transmitting calls. Bell Labs scientists found ways to send many conversations simultaneously through a pair of wires, and later through coaxial cables. They incorporated microwave radio into transmission systems for long distance calls.

In just the last quarter century, such improvements have reduced the average cost per circuit mile of the Bell System nationwide long distance network from $60 to $16. (See graph.) The cost of the newest coaxial cable system is less than $2 per circuit mile.

Average cost per circuit mile of interstate transmission facilities.



But Bell engineers are not satisfied. Already they are preparing the technology for even greater economies and capacities when call volumes reach a level to justify using it. The new Comstar domestic satellite—being used jointly by the Bell System and GTE Satellite Corporation—in addition to standard communications traffic will

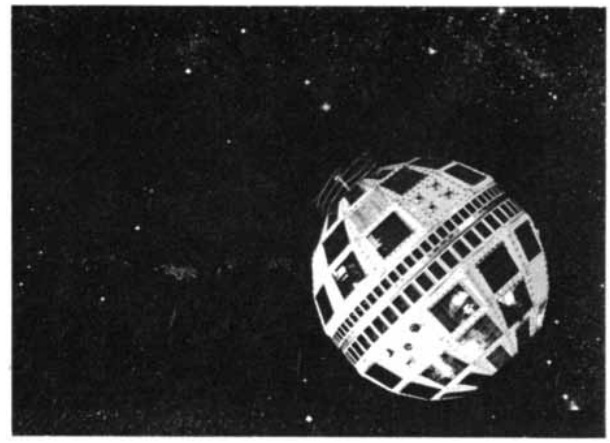
beam experimental signals to an extraordinarily precise antenna so that Bell Laboratories scientists can investigate super-high frequencies that could provide increased satellite capacity in the future. And new systems, using millimeter waveguides or laser light and glass fibers, are expected to reduce transmission costs and add new capacity also.

Another simple way to measure how technology improves productivity is to look at the number of Bell System people required to serve each 10,000 telephones. In 1925 it took 246. In 1958 it took 148. Today it takes 65.

Finally, the Bell System seeks to improve productivity by improving the methods used to manage the telephone business. For instance, the teamwork of Western Electric and Bell Labs people was cited earlier. To facilitate their interaction, some Bell Labs people work adjacent to Western Electric plants. A significant reduction has resulted in the time required to get a new design from drawing board to actual production.

The Bell System is placing greater emphasis on computerized information retrieval for the mountain of data connected with serving 118.5 million telephones. It is moving toward a standardized format for recording and storing data, to make more efficient use of computers.

The Bell System is placing greater emphasis on new methods of employee training, on the restructuring of jobs, and on



The Bell Systems's Telstar® satellite demonstrated the feasibility of using space satellites for communications.

efforts to build more responsibility, challenge and satisfaction into jobs at all levels. Experience to date indicates that these changes help people do a better job, reduce employee turnover, and consequently improve productivity as well.

That is exactly the result one would expect, on the basis of the 1927 Hawthorne findings. The Bell System has changed a lot since 1927. But it still emphasizes people, research, manufacturing efficiency and an organizational structure that fosters teamwork.

Data issued by the federal Bureau of Labor Statistics show that overall the productivity of the telephone industry has increased 50% since 1965. That is two-and-a-half times the productivity increase of the United States economy as a whole.

In that same decade, the cost of living rose 75%. Telephone rates for local service went up only 40%. And interstate long distance rates went up about 4%. Now 95% of all American homes have telephones.

One Bell System. It works.

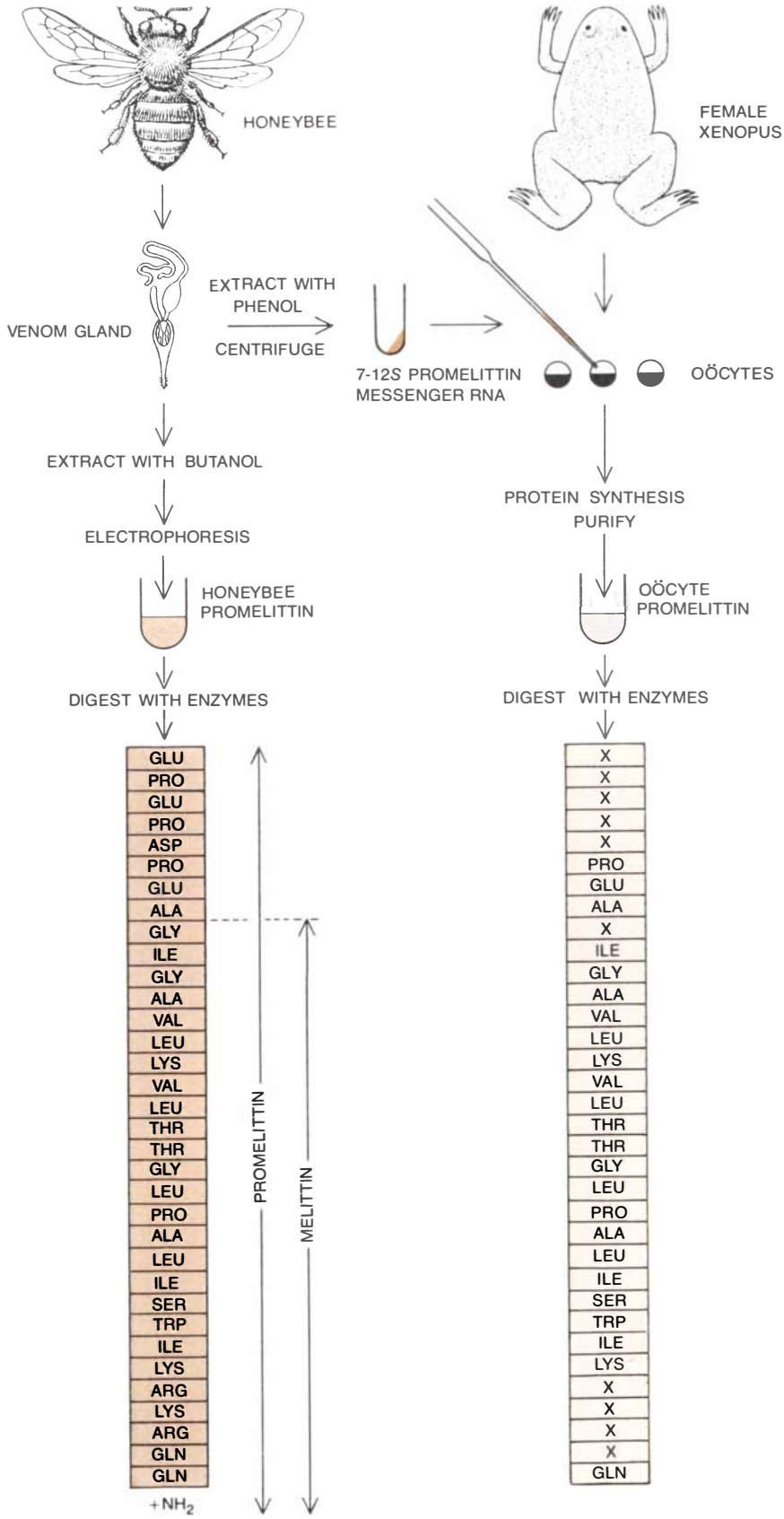
Bell System

factors that allowed them to be translated in the muscle cell. The simplest explanation, however, is that normal muscle cells are able to translate globin messenger if the globin RNA is available. It therefore seems likely that even specialized cells can translate any messenger. Taken in combination with the oöcyte and egg data, this result suggests that any cell can be programmed by any messenger.

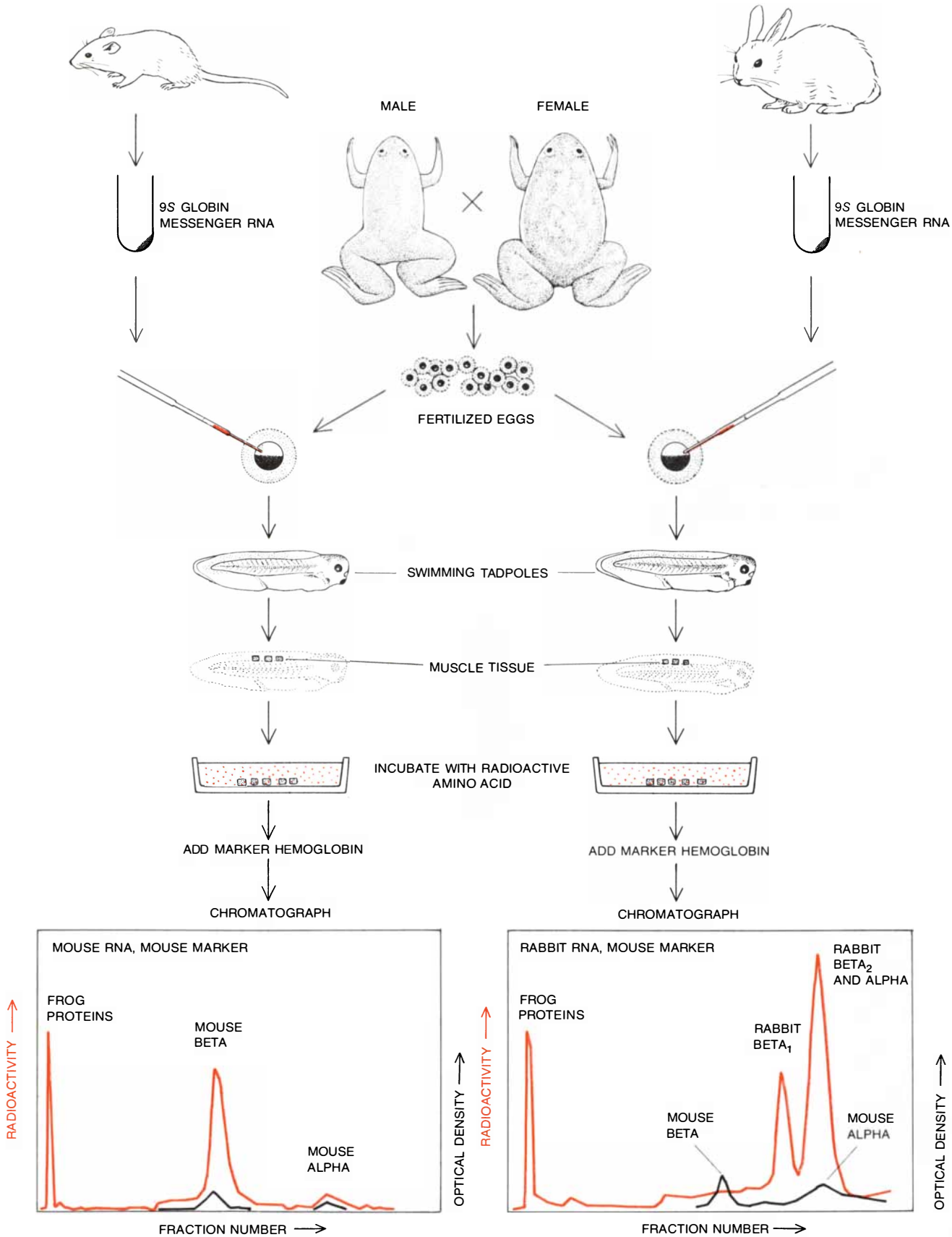
The qualitative selection of genetic information thus seems to involve controlling the supply of available messenger-RNA species rather than selecting, at the translation level, among a population of freely available messengers. The possibility should not be dismissed, however, that the efficiency of RNA translation can be varied to provide quantitative control.

The oöcyte and egg are excellent systems for investigating control at the level of translation and for assaying either minute quantities of messenger or, more important, putative translational control elements. When small amounts (less than five nanograms) of globin messenger are injected into oöcytes, there is an approximately linear response, as measured by the incorporation of a radioactive amino acid, histidine, into hemoglobin: the more RNA, the more radioactive hemoglobin. This means that the oöcyte can be used as an assay system. It also implies that the supply of messenger is one of the factors limiting the rate of synthesis of a particular product. Nevertheless, the overall rate of protein synthesis in the oöcyte cannot be severely limited by the supply of messenger. Injected messengers seem to compete with the cell's own messengers: the injection of messengers that were purified from the oöcyte back into the oöcyte raises the overall rate of protein synthesis by less than 20 percent.

The sensitivity of the oöcyte system is quite remarkable. One can calculate that each injected beta-globin messenger is translated about once every two or three minutes at 19 degrees Celsius, or almost as fast as it would be translated in a red blood cell at the same temperature. Synthesis proceeds at about that rate as long as the oöcytes stay alive in culture. (Some oöcytes injected with globin messenger have lived for more than two weeks, and one might predict survival times of months when better conditions for long-term oöcyte culture are defined.) The efficient synthesis in oöcytes, over a long period of time, of a stable, highly radioactive product in effect enormously magnifies the activity of the injected material and hence enables one to assay minute quantities of messenger—less than a trillionth of a gram of beta-globin messenger RNA. If, on the other hand, extreme sensitivity is not required, it can be more convenient to assay messenger RNA in a cell-



HONEYBEE VENOM is synthesized by frog cells. Venom glands were removed from honeybees and a messenger-RNA fraction was isolated. The messenger was injected into frog oöcytes, which synthesized promelittin, the precursor of the venom component melittin. The amino acid sequence of the oöcyte-synthesized promelittin (*right*) was compared with the sequence of the honeybee protein itself, removed directly from the venom gland (*left*). To the extent that it was determined, the sequence of the oöcyte promelittin was the same as the honeybee version. Secondary modifications such as addition of NH_2 were not carried out in oöcyte.



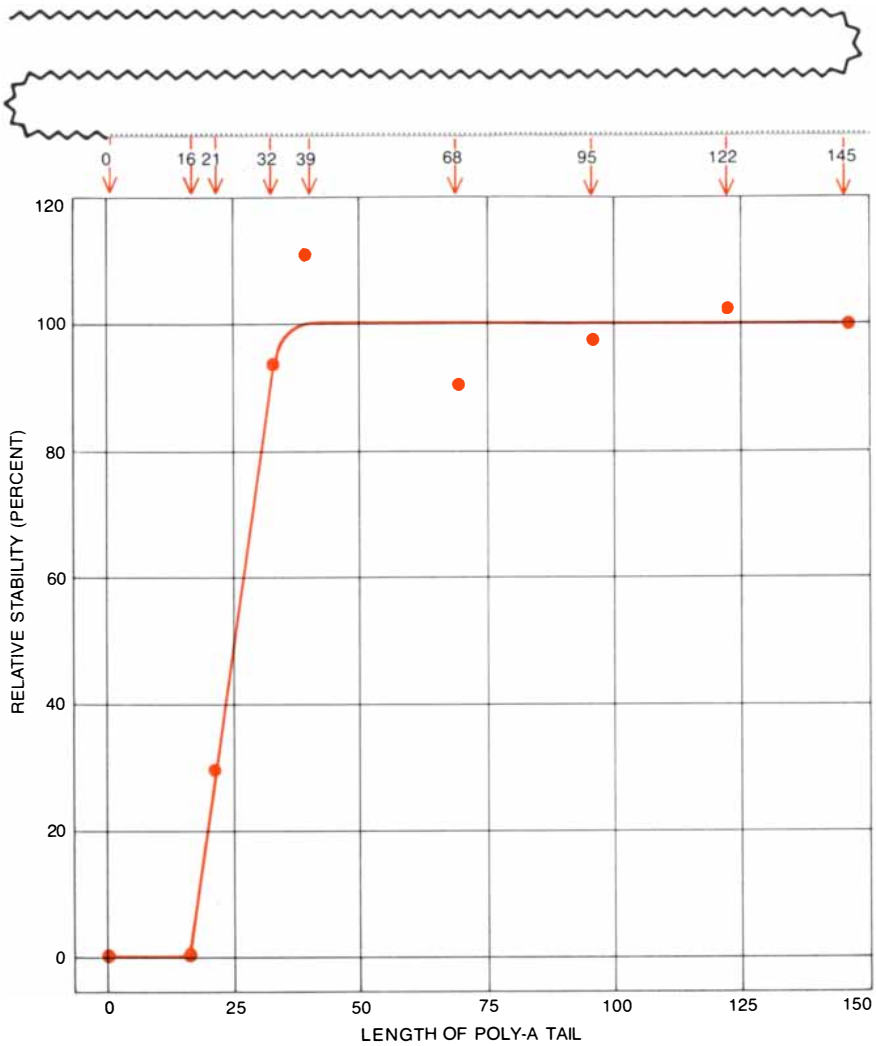
SPECIALIZED CELLS can also translate foreign messenger. At the Laboratory of Molecular Biology in Cambridge, Hugh R. Woodland, John B. Gurdon and Jerry B. Lingrel raised tadpoles from normally fertilized eggs into which messenger RNA for mouse or rabbit globin had been injected. Muscle tissue was dissected from tadpoles and incubated with a radioactive amino acid; mouse hemoglobin was added as a marker. Ion-exchange chromatography shows that chains synthesized in the tissue correspond to the species of the injected messenger: chains made under the direction of mouse RNA form radioactive peaks that coincide with mouse marker peaks (*left*); chains programmed by rabbit messenger yield peaks characteristic of rabbit globin and hence not coincident with the mouse marker peaks (*right*).

free system, where the confusing background of frog-protein products is not present. For assaying putative translational control elements, however, the oöcyte is best because what happens in it is likely to be relevant to what happens in normal living cells. For such purposes one can mix possible regulatory molecules with a messenger RNA and inject the combination or else inject the messenger and, once it is established, introduce the regulator.

The approach is illustrated by an experiment to determine the role of heme in controlling the translation of globin messenger. Crude globin messenger contains just a little more alpha messenger RNA than beta messenger RNA and programs the synthesis of roughly equal amounts of the two chains in a cell-free system. In contrast, oöcytes injected with mouse or rabbit messenger make at least five times as much beta globin as alpha globin. If heme is mixed with the messenger or is injected after the messenger is established, the translation of alpha messenger is enhanced so that the alpha/beta ratio approaches unity. This effect does not necessarily imply that there are elements of the translation system that interact with only one messenger. Rather, the heme probably increases the supply of factors that initiate polypeptide synthesis and thus preferentially affects the alpha messenger, which has a lower affinity for these initiation factors than the beta messenger.

The fertilized egg is also a good medium for studying translational control elements. Components injected into a fertilized egg end up in the rapidly dividing and differentiating—and hence biologically interesting—cells of the growing embryo. In one experiment of this kind the injected messenger served as a probe for the appearance of particular factors during the course of early development. We introduced mouse or rabbit globin messenger and measured the alpha/beta ratio at various developmental stages. Mouse messenger yields the characteristic excess of beta chains. So does rabbit messenger, in the egg and during the initial cleavage stages, but just before the stage called gastrulation the synthesis of alpha chains increases, and it remains at an elevated level at least until the swimming-tadpole stage [*see illustration on page 71*]. Some factor, similar to heme in effect, that specifically stimulates the translation of rabbit alpha-chain messenger appears to be formed at a particular developmental stage.

Whole-cell assay systems present an excellent opportunity for studying messenger stability. The lifetime of a messenger in a cell-free system sheds little light on what its life expectancy would be in a normal living cell; yet to study a cell in which the messenger is being synthesized is difficult, since at any given time the messenger population reflects a balance between synthesis and decay. By injecting a known messenger into an egg or an oöcyte one can study the decay process alone. Messenger stability is relevant to the control of gene expression because instability could be a means of changing the population of proteins as cells differentiate. In this regard results obtained with fertilized eggs are likely to be of particular interest, since here one can look at messenger stability in a developing embryo. So far only rabbit and mouse globin messengers have been tested in this way, and they both appear to be stable at least until the swimming-tadpole stage, after more than 20 cell divisions. Rabbit globin is also stable in the oöcyte, and injected oöcytes will synthesize globin for several weeks. Two frog-liver messengers, those coding for albumin and for the enormous vitellogenin molecule, are also stable, showing that stability is not an artifact arising from the introduction of foreign messenger. (All the messengers so far tested code for differentiated products and are stable in their normal cellular milieu; messengers with a short half-life have not been injected into eggs or oöcytes, and it is possible that they would retain the characteristics expressed in the cell type from which they were derived.)



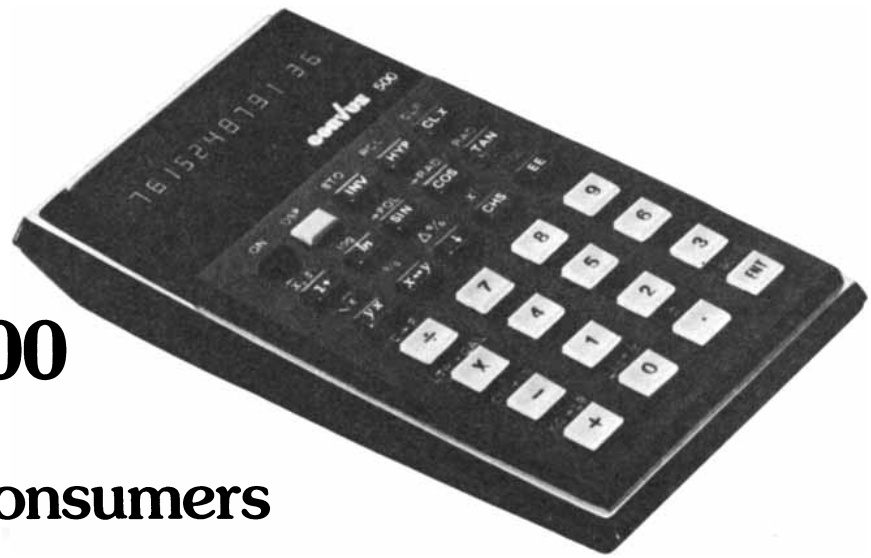
FUNCTIONAL STABILITY in oöcytes of globin messenger molecules with poly-A tails of different lengths was tested by a Belgian-Israeli group. By treating globin messenger RNA with the enzyme polynucleotide phosphorylase for varying periods they produced messenger species with tails of different lengths; the species were purified by chromatography and then injected into oöcytes. After 44 hours of incubation a radioactive amino acid was added and the rate of hemoglobin synthesis was compared with the rate in oöcytes programmed with native, or unshortened, globin messenger. Some 30 adenylic acids are required to stabilize messenger.

Perhaps the most elegant exploitation of the oöcyte system for investigating messenger-RNA stability has been the work of a group of investigators from the Free University of Brussels (G. Huez, Marbaix, E. Hubert and M. Leclercq) and the Weizmann Institute of Science in Israel (Uri Nudel, Hermona Soreq, R. Salomon, B. Lebleu, Michel Revel and Uriel Z. Littauer), who studied the part played by the "poly-A tail" of the messenger RNA in determining

Number One Rated!

corvus 500

TK Enterprises Reporting to Consumers



With MOSTEK single chip technology, the new Corvus 500 is the first non-Hewlett-Packard calculator with Reverse Polish Notation, 10 addressable memories, 4 level roll down stack to be introduced. If you compare the Corvus 500 feature by feature with the HP-45, you will find striking similarities. There are also some important differences:

	Corvus 500	HP-45
RPN (Reverse Polish Notation)	Yes	Yes
Memory Store and Recall 10 Registers	Yes	Yes
4 Level Stack, Rotate Stack	Yes	Yes
10 MEMORY EXCHANGE WITH X	Yes	No
Log, LN	Yes	Yes
Trig (Sine, Cosine, Tangent, INV)	Yes	Yes
HYPERBOLIC (SINH, COSINH, TANH, INV)	Yes	No
HYPERBOLIC RECTANGULAR	Yes	No
y^x, e^x, 10^x, $\sqrt{}$ ×, 1/×, ×!, ×↔y, π,CHS	Yes	Yes
$\sqrt[x]{y}$ through INVERSE	Yes	No
GRADIANS	No	Yes
DEGREE-RADIAN CONVERSION	Yes	No
Degree-Radian Mode Selection	Yes	Yes
DEC-DEG-MIN-SEC	No	Yes
Polar to Rectangular Conversion	Yes	Yes
Recall Last ×	Yes	Yes
Scientific Notation, Fixed and Floating	Yes	Yes
Fixed Decimal Point Option (0-9)	Yes	Yes
DIGIT ACCURACY	12	10
DISPLAY OF DIGITS	12	10
%, Δ %	Yes	Yes
GROSS PROFIT MARGIN %	Yes	No
Mean and Standard Deviation	Yes	Yes
Σ+, Σ−	Yes	Yes
Product—Memories	Yes	Yes
C.F. DIRECT CONVERSION	Yes	No
F.C. DIRECT CONVERSION	Yes	No
LIT-GAL, DIRECT CONVERSION	Yes	No
KIL-LBS, DIRECT CONVERSION	Yes	No
GAL-LIT, DIRECT CONVERSION	Yes	No
LBS-KIL, DIRECT CONVERSION	Yes	No
CM-INCH DIRECT CONVERSION	Yes	No
INCH-CM DIRECT CONVERSION	Yes	No

As you can see, the Corvus 500 IS a lot more calculator for $69.95.

We have listed some of the many features, but let's amplify on some highlights:

1. RPN (Reverse Polish Notation) "COMPUTER LOGIC" and 4 LEVEL STACK. Your problem is solved the way it is written, left to right sequence, eliminating restructuring, unnecessary keystrokes, and the handicap of having to write down intermediate solutions. And all information is at your disposal—just roll the stack (R↓) to any intermediate information desired. You arrive at your solution faster, more simply and therefore more accurately.

Perhaps at this point we should address ourselves to the controversy between algebraic entry and RPN. One question we must ask is why proponents of algebraic entry *always* use an example of sum of products and *never* an example of product of sums:

(2+3) × (4+5)=

Algebraic: 2+3=MS 5+4=×MR=
TOTAL 12 keystrokes (SR51, add 2 more keystrokes)

RPN: 2 Enter 3+4 Enter 5+×
TOTAL 9 keystrokes

2. THE CORVUS 500 and HP-45 HAVE 10 ADDRESSABLE MEMORY REGISTERS, 4 LEVEL OPERATIONAL STACK, and a "LAST X" REGISTER (10th Mem. Reg.) With 10 addressable memories, you have access to more entries, or intermediate solutions; less remembering, or writing down, YOU have to do. And less chance for error.

The stack design also permits X and Y register exchange, and roll-down to any entry to the display for review or other operation.

The "last x" register permits error correction or multiple operations when a function is performed, the last input argument of the calculation is automatically stored in the "last x" register, which can be quickly recalled to correct an error, or to perform another operation using the same number.

3. DIRECT HYPERBOLIC and HYPERBOLIC RECTANGULAR to POLAR, and INVERSE. For those of you electronic and computer science engineers who require access to this specialized application, the Corvus 500 solves "your" problems.

4. A WORD ABOUT CORVUS 500 12 DIGIT DISPLAY AND ACCURACY. Finally you have displayed 12 digit accuracy in business format and 10 + 12 in scientific notation. LED is manufactured by Hewlett-Packard.

FOR THE FIRST TIME you can raise the number 10 to 199th power or calculate Factorial (x!) of up to 120. Unbelievable!

5. The CORVUS 500 CONVERTS METRIC/U.S. and INVERSE UNITS DIRECTLY. These preprogrammed functions act not just as constants, eliminating unnecessary keystrokes.

WHAT ABOUT CONSTRUCTION? With so many features, the next most obvious question must be in regard to the quality of the unit itself. We are proud to report the Corvus 500 to be double injected molded, with "tactile" feedback keyboard. The compact, contoured case is 5½" long by 3" wide by 1¼" high, and weighs just 8 oz.

The COMPLETE CORVUS 500 for $69.95 includes:

- **Rechargeable and replaceable Nickel Cadmium batteries. Optional 3 AA batteries.**
- **Adaptor/Charger**
- **Owner's handbook**
- **Soft carrying case**

The Corvus 500 is warranted by the manufacturer against defects in materials and workmanship for one year from date of delivery.

For those of you who have the HP-21 or 45 or any other advanced calculator on order, aren't you glad you still have the opportunity to take advantage of the release of the Corvus 500 for $69.95? Hurry! Order yours today.

***AN INVITATION:** TK Enterprises is proud that we are able to introduce the MOSTEK chip Corvus-500 to you, our discerning readers. (Incidentally, we have received orders for Corvus 500 from almost every major company in the USA, in every state in the Union (including HP & TI people). Many are already taking advantage of this incomparable price performance ratio instrument. We invite you to experience the new Corvus 500 by mailing in the coupon today, or call our order desk **(800) 421-0367; in CA. (213) 370-5795.** From the tremendous initial response, please allow 3 days for delivery. We are sure you will agree that the CORVUS 500 is the calculator of the year, and the one you have been waiting for. However, you may return the remarkable Corvus 500 within 7 days without obligation, for any reason whatsoever. So, please mail or phone today . . . before the news of Corvus 500's demand exceeds supply!*

To order your Corvus 500 Scientific Calculator, phone: (800) 421-0367; (213) 370-5795 or mail coupon today!

tk enterprises
16611 Hawthorne Blvd.
Lawndale, Ca. 90260
(213) 370-5795
(800) 421-0367

Yes . . . I'd like to try the Corvus 500 for 7 days. Please ship me the Corvus 500 Advanced Scientific Calculator, complete with accessories, at $69.95, plus $2.50 for shipping and insurance. I understand that if I'm not completely satisfied, I may return the unit and accessories within 7 days for full refund (or, at my option, receive full credit toward the purchase of any other calculator carried by T.K. Enterprises).

NAME _____

ADDRESS _____

COMPANY _____

PHONE _____

☐ **Cash Payment:** Check or money order enclosed. (Please be sure to add any applicable state and local taxes.) To speed-up delivery send money order.

☐ **Bill My Credit Card:** ☐ BankAmericard ☐ Master Charge Bank No. ☐☐☐☐

Account No. ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

Expiration Date _____

Signature: _____

SA #8-76

Other Values from TK:

Texas Instruments

PC-100	$229.95
SR50A	$ 47.95
SR51A	$ 62.95
SR52	$229.95
SR56	$ 89.95
5050M	$109.95

Libraries, cards available.

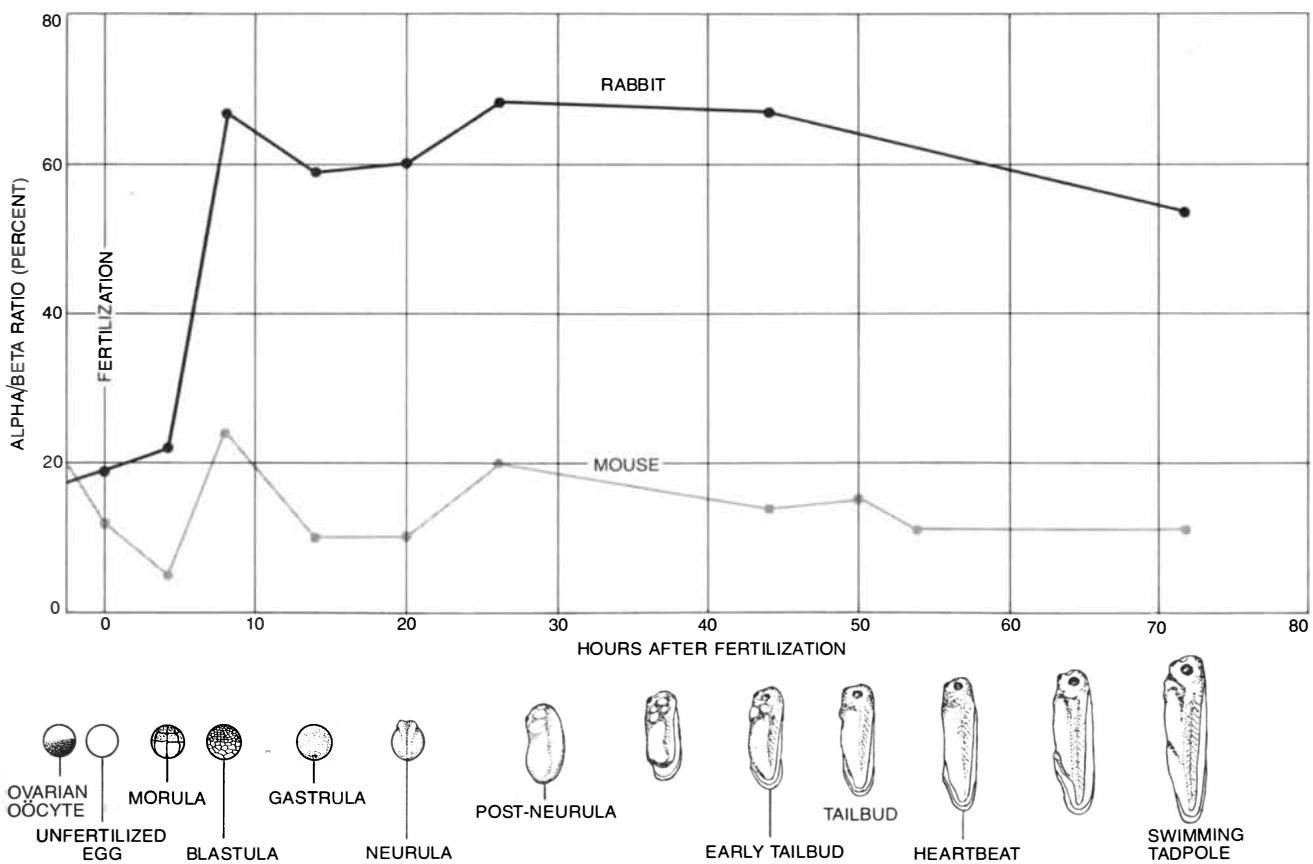
NOVUS Consumer Products from National Semiconductor

Mathematician #4510	$17.88
Scientist #4520	$29.95
Scientist PR #4525	$54.95
NS #4615	$48.88
NS #4640	$54.95
Financier	$28.45
Statistician	$28.45

the life expectancy of the whole molecule. One of the most distinctive features of most kinds of messenger RNA is the presence at one end of the molecule of a long string of a single kind of nucleotide, adenylic acid. Yet the function of these rather striking poly-A tails remained a mystery, not least because it was known that messengers lacking tails could be translated normally in certain cell-free systems. The Belgian-Israeli group injected into oöcytes rabbit globin messenger RNA with and without poly-A tails. They found that messengers without tails are translated well enough at first but do not function efficiently for very long: after 48 hours the RNA with poly-A was making about 25 times as much globin as the tailless RNA. The poly-A is thus implicated in preserving the functional stability of messenger RNA. The tailless messengers are actually destroyed by the oöcyte; they can hardly be detected physically 48 hours after injection. Since the tails were removed by stepwise degradation with the enzyme polynucleotide phosphorylase, one might argue that the enzyme also affected some other part of the globin messenger, for example the region adjacent to the tail. The investigators went on, however, to restore poly-A tails to the RNA with the enzyme adenyl transferase, and they found that stability was restored. In other words, it really is the presence of the tail that confers stability on the messenger molecule.

When polynucleotide phosphorylase is used to shorten the tail rather than to remove it completely, a critical tail length of about 30 adenylic acid units turns out to be required before a messenger molecule is stabilized [*see illustration on page 69*]. One can speculate that in cells that, unlike the oöcyte, require a rapidly changing population of messengers the life expectancy of a "stabilized" molecule may depend on how much longer the tail is than the critical length, given that many such cells contain enzymes that shorten poly-A tails. Although many of the above results could be peculiar to the oöcyte system, careful reexamination of data from cell-free systems has confirmed the tendency of messenger without poly-A to be less stable. With their combination of sophisticated chemistry and simple microinjection experiments the Belgian-Israeli group may well have solved the mystery of the poly-A tail.

Experiments in the microinjection of RNA demonstrate the advantages of combining the unrelated disciplines of biochemical analysis and micromanipulation. The approach is not limited to amphibian eggs and oöcytes. The egg of the fruit fly *Drosophila* is several hundred times smaller than the *Xenopus* oöcyte and the mouse egg is several thousand times smaller, yet both of these small eggs can be microinjected. Ordinary cells growing in culture have also been injected, and the technique is not limited to the injection of RNA. Gurdon, Alan Colman, Christopher Ford and Woodland have already injected eggs and oöcytes with eukaryotic and viral DNA and with histones, antibodies and other proteins. Some synthetic and naturally occurring DNA's are transcribed, and the DNA replicates in the egg cytoplasm. Introducing specific genes into living frog cells could improve our understanding of development. The raw materials for such experiments might be provided by recently developed techniques for cloning recombinant DNA molecules.



CONTROL OF TRANSLATION at various developmental stages is studied by injecting mouse or rabbit globin messenger into fertilized eggs and analyzing the globin synthesized by the embryos for its content of alpha and beta chains. Eggs injected with mouse RNA produce about a fivefold excess of beta chains (*gray curve*), that is, the alpha/beta ratio is about 20 percent (as it is also in injected oöcytes). In eggs programmed with rabbit globin messenger the alpha/beta ratio is about the same at first (*black curve*), but production of alpha chains increases suddenly just before gastrulation; some factor that stimulates alpha-chain synthesis is formed at that stage of development.

The Photographic Lens

New optical materials (both glasses and plastics), antireflection coatings, laborsaving computer programs and new production methods yield high-performance lenses at increasingly low cost

by William H. Price

Designing a lens can be compared to playing chess. In chess a player tries to trap his opponent's king in a series of moves. In creating a lens a lens designer attempts to "trap" light by forcing all the rays arising from a single point in the subject to converge on a single point in the image, as a consequence of their passing through a series of transparent elements with precisely curved surfaces. Since in both cases the ultimate goal and the means by which it can be attained are known, one is tempted to think there will be a single best decision at any point along the way. The number of possible consequences flowing from any one decision is so large, however, as to be virtually, if not actually, infinite. Therefore in lens design as in chess perfect solutions to a problem are beyond reach. Although this article will be concerned only with the design of photographic lenses, the same principles apply to all lenses.

The lens designer has one enormous advantage over the chess player: the designer is free to call on any available source of help to guide him through the staggering number of possibilities. Most of that help once came from mathematics and physics, but recently computer technology, information theory, chemistry, industrial engineering and psychophysics have all contributed to making the lens designer's job immeasurably more productive. Some of the lenses on the market today were inconceivable a decade ago. Others whose design is as much as a century old can now be mass-produced at low cost. With the development of automatic production methods lenses are made by the millions, both out of glass and out of plastics. Today's lenses are better than the best lenses used by the great photographers of the past. Moreover, their price may be lower, in spite of the fact that 19th-century craftsmen worked for only a few dollars a week and today's lenses are more complex. The lens designer cannot fail to be grateful for the science and technology that have made his work easier and his creations more widely available, but he is also humbled: it is no longer practical for a fine photographic lens to be designed from beginning to end by a single human mind.

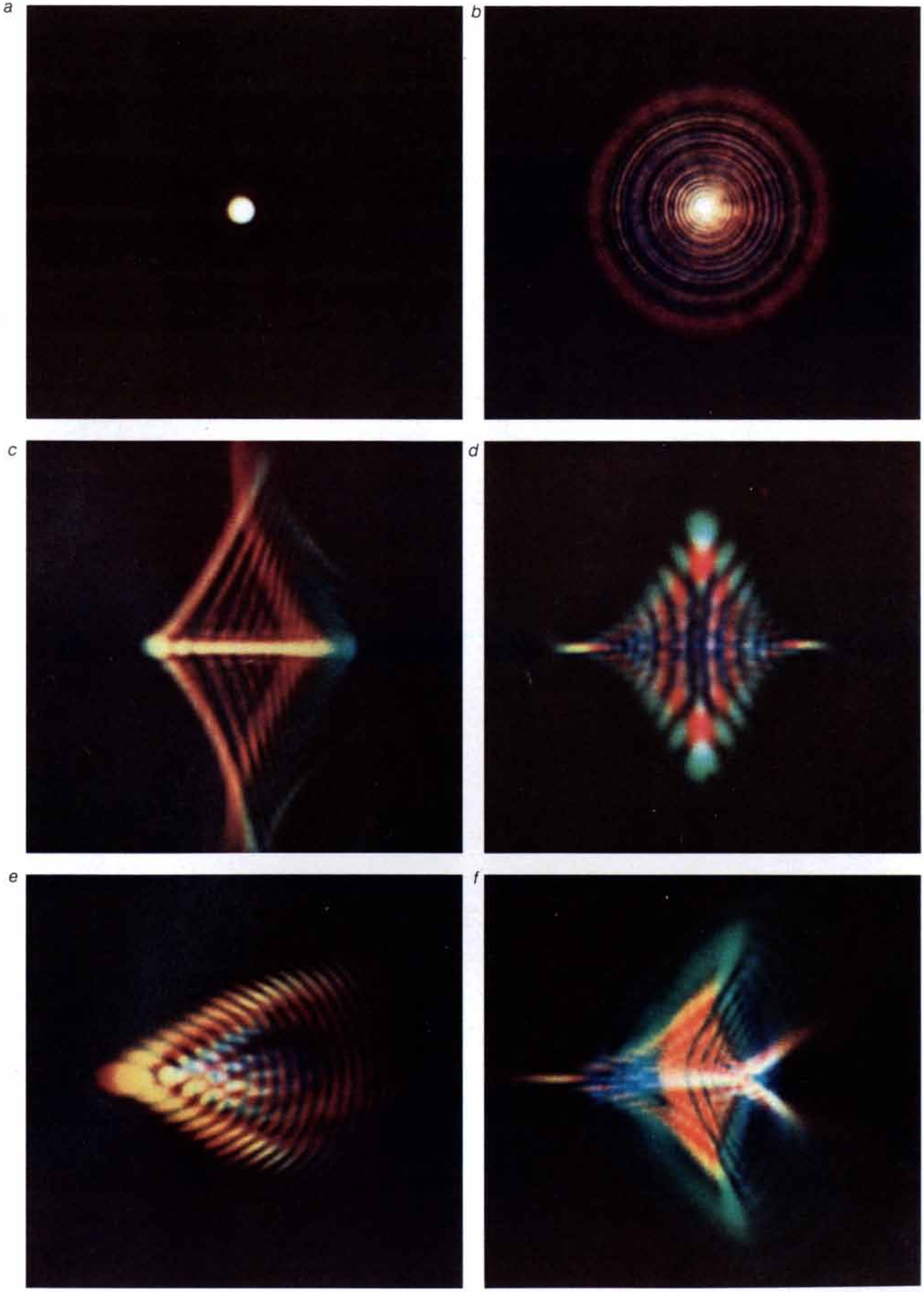
What kind of lens was the first to be used in photography is not known because the inventor of photography, Joseph Nicéphore Niepce, left no written record of his experiments. It is believed, however, that Niepce's first picture (which is lost to history) was made in 1822 with a meniscus lens in a camera obscura. "Meniscus" (from *meniskos*, the Greek diminutive for moon) describes the cross-sectional shape of the simplest practical photographic lens: a crescent moon, formed by two arcs of different radii. Simple eyeglasses for reading are meniscus lenses.

The camera obscura (the Latin for dark room) had been known from antiquity. Leonardo da Vinci described a simple form of camera obscura in which light entered the room through a small hole and formed a faint image on the opposite wall. In the 16th century the hole was replaced by a meniscus lens, which made the image many times brighter. The camera obscura was a popular tool of artists, who used it to trace the outlines of their subjects.

If Niepce made his first photographs with a camera obscura and a meniscus lens, he soon sought something better. It is reported that Charles Louis Chevalier, of a firm of engineers and instrument makers in Paris, provided him with a two-element achromatic lens. Such lenses, designed to minimize the chromatic aberration, or color fringes, produced by simple meniscus lenses, had been introduced to astronomy in 1758 by the English optician John Dollond, but they were still novelties early in the 19th century. At about the same time that Chevalier was adding an achromat to Niepce's camera, Joseph Jackson Lister and Giovanni Battista Amici introduced achromatic lenses to microscopy, thereby removing chromatic aberration from microscopes and for the first time making bacteria visible.

Chromatic aberration results from the phenomenon in which a prism disperses white light into a spectrum of colors. In a vacuum all colors, or wavelengths, of light travel at the same velocity. In a material medium the velocity of light is always reduced, and the velocity of the shorter wavelengths is reduced more than that of the longer ones. Thus when light leaves one medium and enters another at an angle, its path is sharply refracted, or bent, either toward or away from a line perpendicular to the interface between the two mediums, depending on whether its velocity in the second medium is lower or higher.

LENS ABERRATIONS can be studied by magnifying the images formed at the focal plane when a point light source is beamed at the lens. In the micrographs on the opposite page, made by Norman Goldberg, technical director of *Popular Photography*, the images are magnified 600 diameters. Ideally the image should itself be a point (*a*), but the ideal is usually achieved, as in this case, only when the light source is in line with the axis of the lens. Image *b* exhibits one of the most common lens defects, spherical aberration. (The sources of the various aberrations are illustrated schematically on page 75.) Another common defect, astigmatism, accounts for the strong horizontal line in image *c*, which also exhibits coma and chromatic aberration. Astigmatism shows up in purer form in image *d*. If the focus of the lens were moved slightly either forward or back, the astigmatism would produce either a sharp horizontal line or a sharp vertical line. Coma, a familiar type of aberration that arises when the light source is off axis, is depicted in image *e*. The final image (*f*), which combines a complex mixture of coma, astigmatism and chromatic aberration, is typical of the off-axis images produced when fast, modern lenses are used at full aperture. After allowance has been made for the great magnification of the image, however, it is evident that the lens focuses most of the light energy within a very small "blur disk," which in this particular case is a circle with a diameter of .03 millimeter.



The susceptibility of light to refraction is the cornerstone of refractive optics, or the theory of lenses. Some 60 years before Isaac Newton took up the study of optics the phenomenon of refraction was described and graphically depicted by the Dutch mathematician Willebrord Snell van Royen. Thereafter René Descartes formulated the precise law of refraction, which became known as Snell's law: the sine of the angle of incidence times the index of refraction of the first medium is equal to the sine of the angle of the refraction times the index of refraction of the second medium. Snell's law is still the lens designer's single most useful formula. It tells him how to bend light to his will.

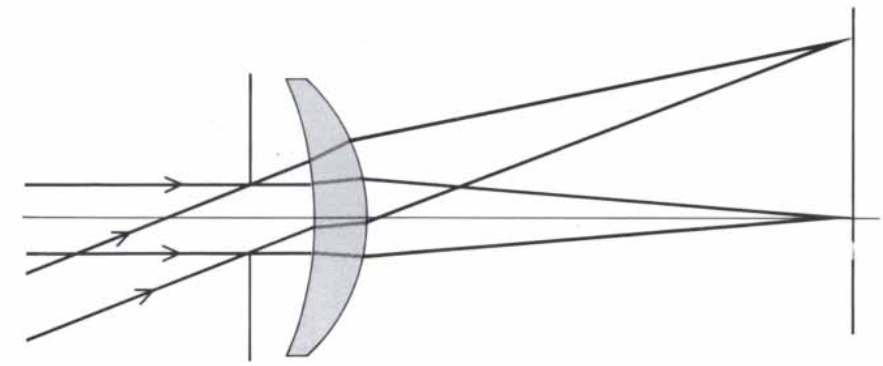
When a beam of white light strikes glass at an angle, the blue-violet wavelengths are bent the most and the red wavelengths the least. It is for this reason that the edges of the emerging beam appear to be fringed with color. From his work with prisms Newton was the first to perceive that white light is a mixture of colors. His first proposition in *Opticks* states that "Lights which differ in Colour differ also in Refrangibility." He was swept too far, however, by the force of his perception. He incorrectly jumped to the conclusion that since chromatic aberration is inherent in lenses, nothing can be done about it. (Whereupon he invented the reflecting telescope, which does away with lenses and the troublesome problem of refrangibility.)

What Newton failed to observe is that in passing through glasses of different compositions colors exhibit different degrees of refrangibility, or, as we would say today, glasses can have different dispersions. This is the weapon with which the lens designer can combat chromatic aberration. The trick is to make a lens with at least two elements. The first element is a convex lens made of a glass that disperses the colors to a minimum degree. The second is a concave lens made of a glass that disperses the colors to a maximum degree. In optical terminology a positive lens of lower dispersion is combined with a negative lens of higher dispersion. If everything is planned just right, the dispersion almost cancels out but the compound lens can still refract substantially [*see bottom illustration at left*].

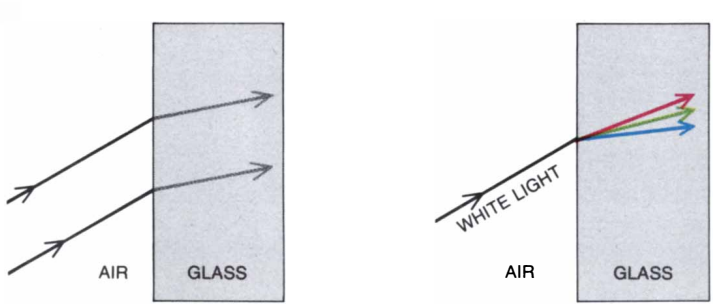
That roughly describes the standard landscape photographic lens of the 19th century. In fact, the two-element photographic lens is still in production today, both as a low-cost lens for simple cameras and as a fairly expensive telescopic lens for nature and sports photography. In such lenses a long focal length is needed to bring objects closer, and a small field of view is acceptable. Good as the original two-element recipe was, it was still only a recipe. The early achromats were empirical lenses. There was no exact theory of how to match up the variables.

Part of the solution came in 1841, when Carl Friedrich Gauss published his theory of lenses. The theory conceptually simplified a lens to ignore all rays except those that lie in a plane either through the lens axis or close to it. Such rays are called paraxial. Even though the Gauss model is simplified, it includes focal length (the distance from the optical center of the lens to the plane in which the rays are brought to a focus), magnification, the location of the principal points of the lens and the location of the image. Gauss's theory was to lens design what trigonometry was to navigation. Paraxial ray tracing served for several decades as the principal tool of photographic lens designers. The simplification implicit in Gauss's procedures does, however, exact a penalty: the plane of the sharpest image usually does not coincide precisely with the plane defined by the Gaussian rules. Nevertheless, as a first cut the analysis comes very close to the bone.

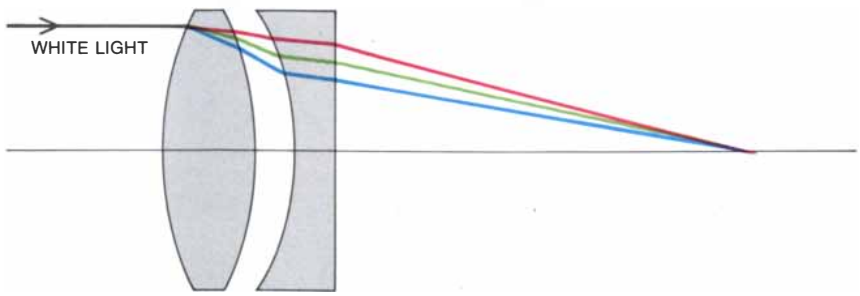
The Gaussian image plane also serves as a benchmark for measuring the entire range of aberrations with which the lens designer must contend, even if



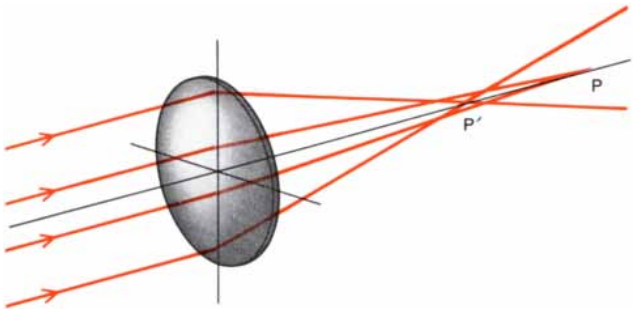
SIMPLE MENISCUS LENS, a crescent-sectioned element with two spherical surfaces, was probably in camera with which Joseph Nicéphore Niepce took the first photograph in 1822.



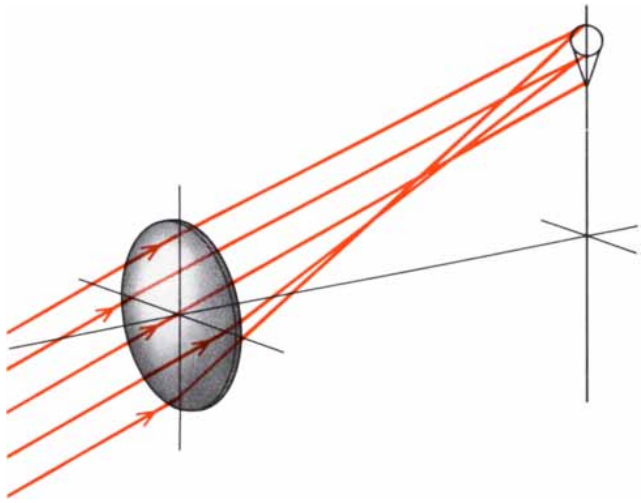
LIGHT IS REFRACTED, or bent, when its velocity is changed as a result of passing from one medium to another. Because short wavelengths of light travel more slowly in glass than long wavelengths, white light is dispersed into its spectral colors. The degree of dispersion differs according to the composition of the glass (or composition of some other transparent medium).



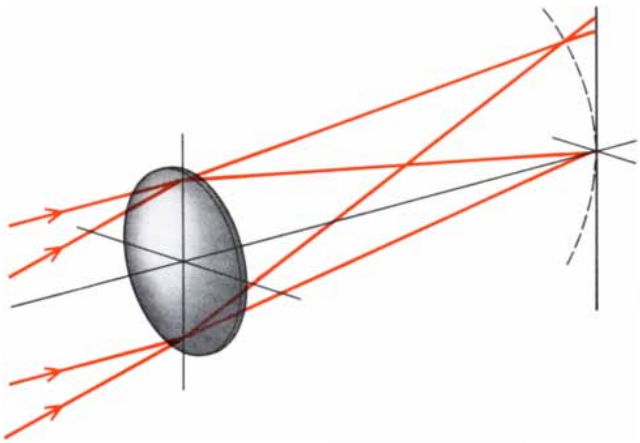
CHROMATIC ABERRATION can be corrected by combining glasses with different dispersions. In this example of a simple two-element lens first element is made from low-dispersion glass. The second element, of opposing power but weaker, is made from high-dispersion glass. Dispersion is canceled out but focusing power remains. Dispersion of colors is exaggerated.



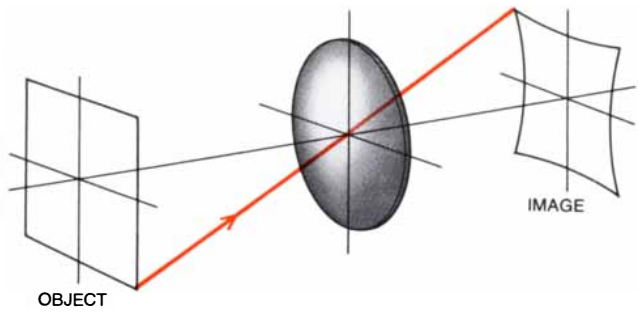
SPHERICAL ABERRATION is an inherent characteristic of any lens whose surface is a section of a sphere. Light originating from the same object point comes to a focus at slightly different points (*P* and *P′*), depending on whether the rays pass through the center of lens or the periphery. Distance separating *P* from *P′* varies with aperture.



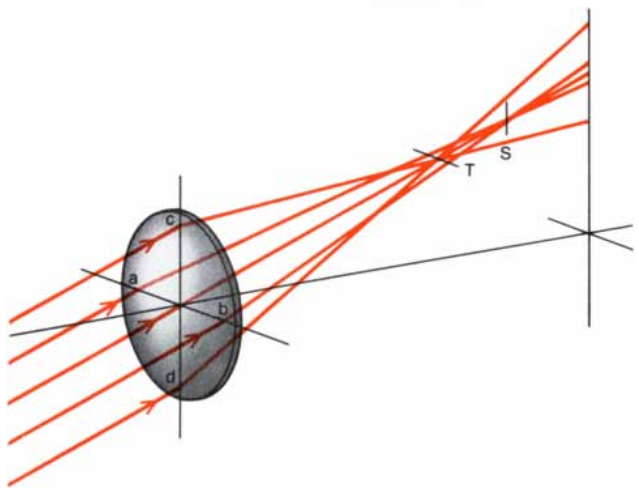
COMA is produced when light from a point off the axis of the lens passes through the perimeter of the lens and comes to a focus in a ring displaced radially from the focus of light that has passed through the lens center. Coma appears as a bright core of light with a spreading tail. Word is derived from the Greek for flowing hair.



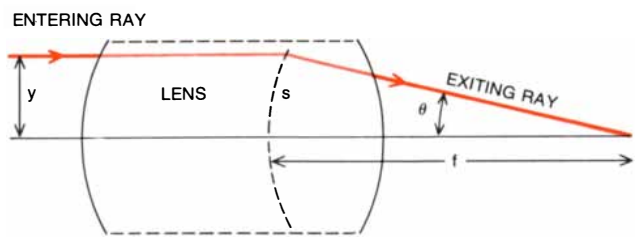
CURVATURE OF FIELD results when an oblique beam of light is brought to a focus closer to the lens than an axial beam of light is. This type of aberration gives rise to an image surface that is curved.



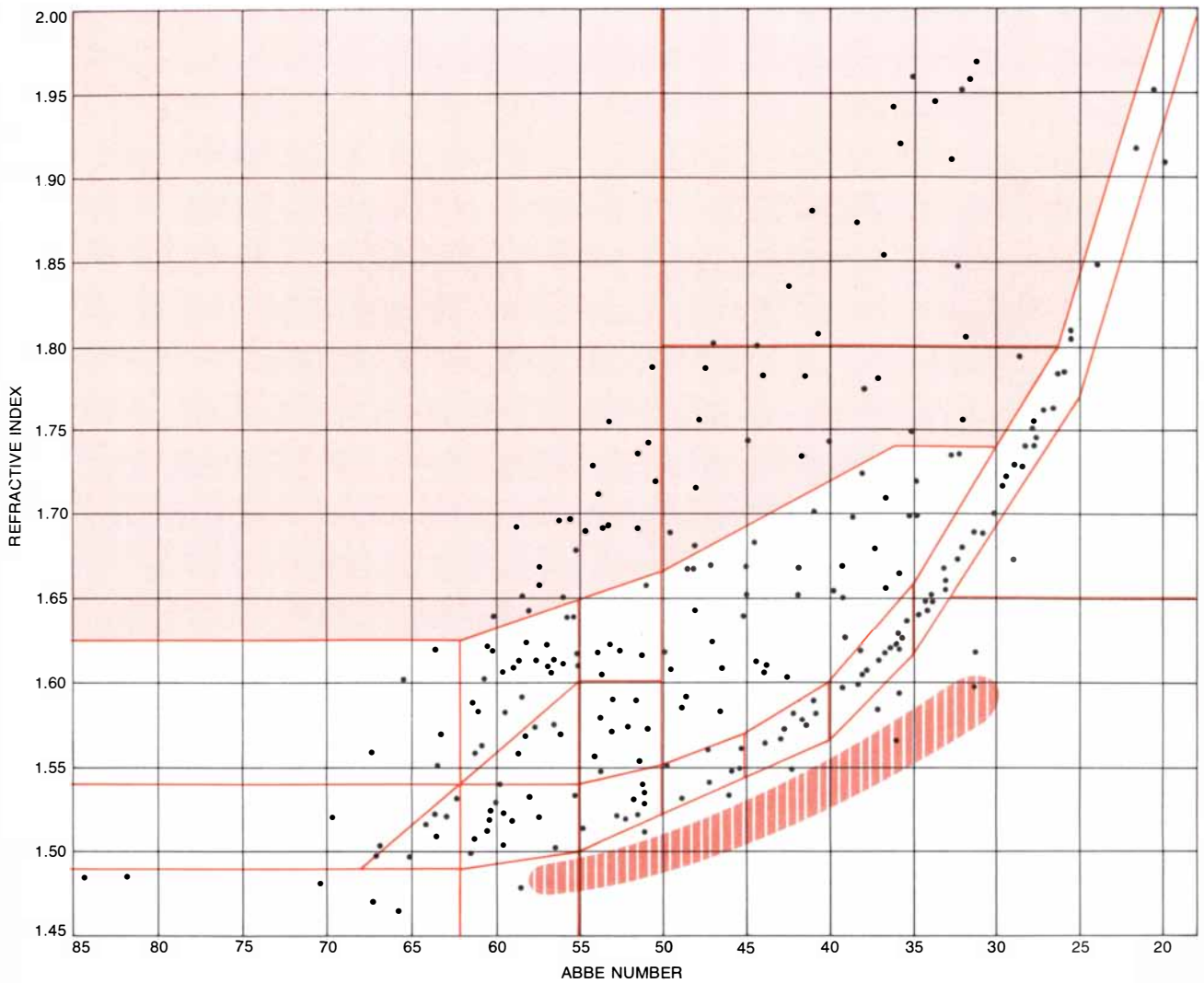
DISTORTION is a lens aberration in which the magnification of the image varies with the obliquity of the entering rays. The consequence of this aberration is that straight lines appear curved in image.



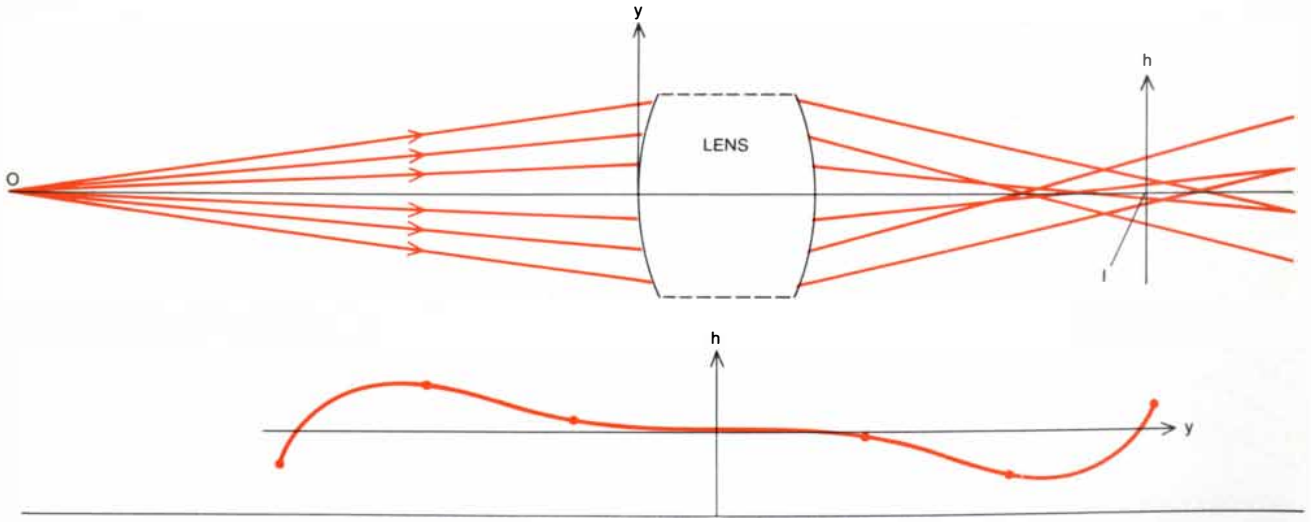
ASTIGMATISM is another common defect of off-axis images created by lenses. Light from an off-axis point that passes through the lens along the axis *a,b* is focused at *S*, whereas light from the same object point that passes through the lens along axis *c,d* is focused at *T*. *S*, which is known as the sagittal, or radial, focus, is a line image that is perpendicular to the optical axis; *T*, the tangential focus, is a line image that is tangent to a circle centered on the optical axis.



ABBE SINE CONDITION, formulated by Ernst Abbe, specifies the condition for simultaneous correction of spherical aberration and coma. When the rays entering the lens and the rays leaving the lens are extended, they intersect at a surface, *S*, defined as the locus of the points at which all rays from an infinitely distant point source on the axis of the lens appear to be refracted. The dual correction is achieved when *y* is equal to $f \cdot \sin \theta$; therefore surface *S* is spherical.



PROPERTIES OF OPTICAL GLASSES are commonly characterized by their refractive index, n, which measures the light-bending power of the glass, and by their Abbe number, ν, which measures the extent to which the glass disperses white light into its spectral colors. The lower the Abbe number, the greater the dispersion. Each point on the chart represents a different optical glass. The rare-earth glasses (*area tinted in color*) combine high refractive index with low dispersion. Optical plastics lie in hatched area below optical glasses.



TRANSVERSE SPHERICAL ABERRATION, h, of a given lens, L, is a measure of the failure of light rays originating at O to converge at the Gaussian image point, I. The aberration varies with lens aperture y in a manner described by a series expansion, $h = ay^3 + by^5 + cy^7 + \ldots$. Plotted curve of h against y shows such an equation fitted to a few traced rays. The aberrations of other rays passing through lens may be interpolated from curve or calculated from equation, after solving for a, b, $c \ldots$, without further ray tracing.

his lens is to be used only for photography in light of one color. The primary monochromatic aberrations, which were defined mathematically by Ludwig Seidel in 1856, are spherical aberration, coma, astigmatism, curvature of field and distortion.

Spherical aberration arises because lens surfaces are sections of spheres, and light passing through the edge of a lens comes to a focus at a point different from that for light passing through the center of the lens. Coma also involves the spherical nature of lens surfaces; images formed off the central axis of the system tend to be asymmetrical. "Coma" has the same root as "comet": the Greek for flowing hair. Because of coma the point of focus seems to have a cometlike tail.

Astigmatism means not coming to a point; it is from the Greek *a-*, not, and *stigma,* mark or spot. It is also due to the asymmetry of off-axis images, but its effect is to cause light to be spread along a line either in a plane through the image point and the lens axis or at right angles to that plane. It has the curious effect of blurring horizontal lines and sharpening vertical lines, or vice versa.

Curvature of field arises if the locus of the sharpest points in the image lies on a curved surface instead of on a plane. It is this aberration that accounts for photographs that are sharp in the center and blurred around the edges. In certain types of large telescopes the aberration is left uncorrected and the photographic plate is curved to compensate for it. The last of the principal aberrations, distortion, gives the appearance of a curved object projected onto a flat surface, or vice versa. It is analogous to the effect that makes Greenland so large in a map made on Mercator's projection.

The early 1840's saw not only the genesis of scientific lens design based on Gaussian principles but also the birth of two men who were to make the next great contributions to the lens-design problem: Ernst Abbe and John William Strutt, later Lord Rayleigh. Born in 1840, Abbe became the chief physicist and lens designer for the famous optical firm of Carl Zeiss. Among his many contributions were the Abbe number, which is used in the classification of optical glasses, and the Abbe sine condition, which defines a lens that is free of coma. The Abbe number is the reciprocal of the degree of dispersion; it incorporates the difference in refraction of two wavelengths that are widely separated in the spectrum. Guided by the Abbe number, the lens designer can cancel out chromatic aberrations for any two wavelengths of light. The Abbe number was the lens designer's answer to Newton.

In selecting glasses for a new lens the designer consults a chart of glass properties in which the horizontal axis is marked off in Abbe numbers [*see top illustration on opposite page*]. The scale on the vertical axis shows the index of refraction for light of a specified color in the middle region of the spectrum. With the chart the designer can see at a glance the interaction between the available glasses and light of the primary colors.

The Abbe sine condition states that coma is eliminated when the distance from the axis of the lens to the point where a ray enters the lens, traveling parallel to the axis, is equal to the focal length of the lens times the sine of the angle the ray makes with the axis at the point of focus [*see illustration at bottom right on page 75*]. Today nearly every fine camera lens meets the Abbe sine condition.

Lord Rayleigh, two years younger than Abbe, presented lens designers with the ultimate challenge by stating the conditions that must be met by a perfect lens. He showed mathematically that an image formed by an optical system will not differ sensibly from a perfect image only if all the rays travel over optical paths of equal length. Rayleigh found that in practice the perfect image is attained if the difference between the shortest path and the longest does not exceed a quarter of a wavelength of light. Such a lens is said to be diffraction-limited. Ideally a lens should be at the Rayleigh limit for light of all wavelengths. If this goal is to be closely approached, the glasses in the lens must all have the same partial dispersion even though their individual dispersions may vary. (Partial dispersion is the rate of change of dispersion with wavelength. Dispersion proper is the rate of change of the index of refraction with wavelength.)

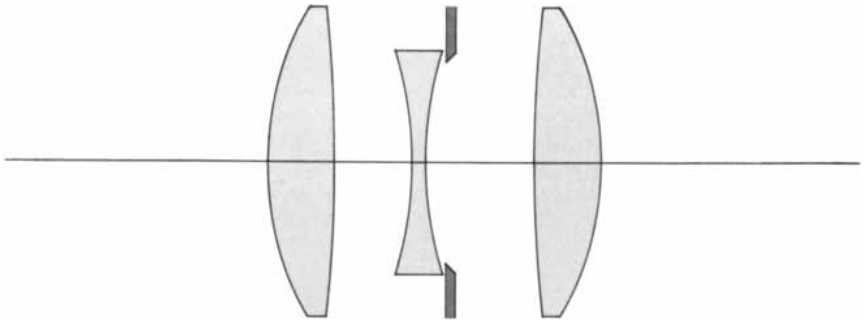
During the 19th century and early in the 20th Józef Miksa Petzval, Henry Coddington and A. E. Conrady, among others, derived mathematical relations and developed techniques for evaluating the magnitude of specific aberrations on the basis of the smallest possible amount of data, in order to reduce the crushing burden of calculation. Petzval, the inventor of the Petzval portrait lens, discovered that the curvature of field of a lens, in the absence of astigmatism, is a relatively simple function of the index of refraction of the lens elements and their radii of curvature. Coddington is credited with the derivation of simple formulas for the calculation of astigmatism for small lens apertures. Conrady is considered by many to be the father of modern optical design, through his application of the concept of optical-path difference to the primary aberrations, both monochromatic and chromatic.

At the end of a century of photographic-lens development, say by the mid-1920's, the creation of lenses was no longer a cookbook operation, but lens design still remained more of an art than a science. Confronted with a lens-design problem, a few men with a genius or a knack could sense the direction to take. Then it was a matter of the laborious application of Snell's law over and over again as designs were checked out by tracing the paths that rays of light might take from the object to the image. Few outside the profession could comprehend the magnitude of the task. The only tools available were a six-place table of common logarithms and by the 1930's a desk-top mechanical calculator. A difficult lens could absorb several man-years of computation. The key to a successful outcome was persistence.

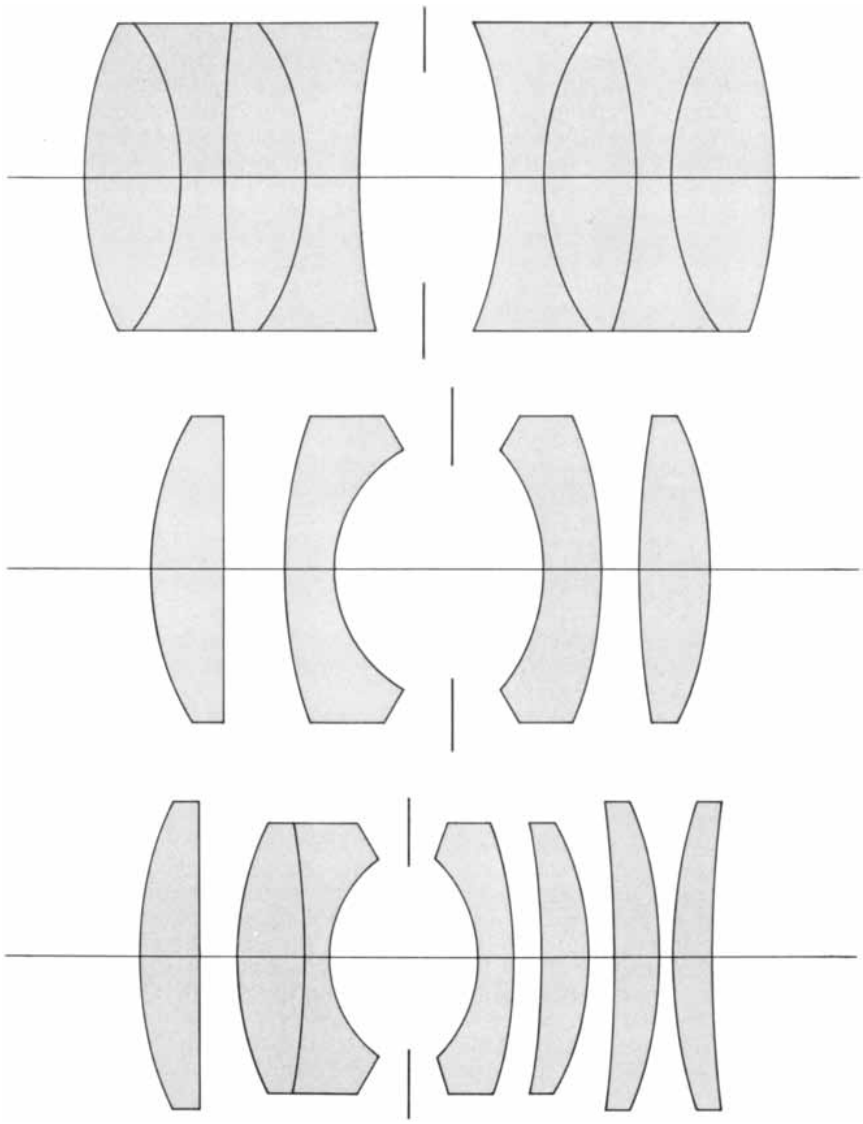
For all the inelegance of such brute-force methods, nearly all the fundamental types of lenses on the market today were pioneered during that period. Some of them, such as the Zeiss Sonnar *f*/1.5, were impressive for their speed. Others, such as the Cooke triplets, were remarkable for their simplicity. (The designation *f*/1.5 means that the focal length of the lens is 1.5 times the maximum aperture of the lens. The smaller the *f* ratio is, the more light a lens can gather in a given time. Light-gathering power, or speed, is inversely proportional to the *f* number squared.)

A brief excursion into geometric optics will help to explain the unique position of the Cooke design, first described in 1893 by H. Dennis Taylor of the British optical house of Cooke and Sons. Geometric optics, unlike the more general field of physical optics, ignores all the known facts about light except those affecting the path of its propagation. Consider the spherical aberration of the hypothetical lens depicted at the bottom of the opposite page. The designer would like the lens (*L*) to bring all the rays from a point of origin (*O*) to a point of focus (*I*). One can see that lens *L* fails badly in this respect. Although some of the rays do converge at *I,* most of them do not.

Any failure of a ray to strike point *I* can be measured in the same way a miss in target practice can be measured, by finding the distance from the bull's-eye to the striking point. In optics such a deviation is called a transverse aberration, a distance measured perpendicular to the path of the light ray. In the diagram for lens *L* the transverse aberration from the aiming point *I* is designated as the magnitude *h* (for height above or below the axis). Inspection of the diagram tells us that *h* is a function solely of the point where the ray enters the lens; that point will be designated as a magnitude along dimension *y*. In other words, the transverse spherical aberration of rays from any point is a function of the aperture of the lens. This is a general characteristic of lenses, which is why most of them achieve their maximum sharpness when they are used at less



COOKE TRIPLET LENS, originally conceived in 1893 by H. Dennis Taylor, is probably the most studied and refined type of photographic lens in service today. Its three elements provide the simplest arrangement that enables the lens designer to eliminate all seven basic aberrations out to third order. Terms of higher order tend to be small. The seven are spherical aberration, coma, astigmatism, distortion, curvature of field and chromatic aberration along two axes.



REFLECTIONS FROM LENS SURFACES hampered lens design before the development of antireflection coatings in the late 1930's. When many elements were desirable for high performance, designers were limited to systems in which several of the elements were cemented together (*top*) to eliminate the reflections that normally occur at air-to-glass surfaces. Because the cemented surfaces had to have the same curvature, the designer had fewer degrees of freedom with which to reduce aberrations. The lenses were also costly to manufacture. For many purposes Gaussian designs with fewer elements (*middle*) were nearly as satisfactory. Today antireflection coatings make air-spaced systems of many elements feasible (*bottom*), so that highly corrected, large-aperture (fast) photographic lenses are available at moderate cost.

than their full aperture. Spherical aberration applies only to the images of object points lying on the optical axis of a lens. It may vary with the distance from the lens of the object point O.

The exact position of any of the infinite number of possible aberrant rays on dimension h is found by solving the equation $h = ay^3 + by^5 + cy^7 \ldots$. The coefficients $a, b, c \ldots$ are found by tracing several rays at various apertures, y, to calculate the specific values of h, then solving the resulting simultaneous equations in $a, b, c \ldots$. The values differ with each lens design. Once the values are derived for a few magnitudes of y, they hold for all intermediate magnitudes of y, and therefore they comprehensively describe all the aberrant light in the image.

The reader may wonder why only odd-numbered exponents appear in the equation. The first-order exponent, y^1, is absent because it represents nonaberrant light that converges exactly at I. All even-numbered terms are absent because regardless of whether y is positive or negative, y^2, y^4 and so on would all be positive. Image formation, however, is symmetrical. Even-numbered terms drop out because they implicitly contradict symmetry.

For lenses of small aperture and small field the higher-order terms tend to be small. If the aberrations represented by the third power of the aperture and the field are corrected, most of the light energy is concentrated in the image point. The first practical formulas for calculating the third-order values of the primary aberrations were published by Seidel 120 years ago. Canceling out lower-order aberrations does not necessarily reduce aberrations of higher order, but they nonetheless tend to be reduced. And if they are reduced, a very good lens results.

The great virtue of the Cooke triplet is that it contains the smallest number of elements by means of which all seven of the third-order aberrations can be eliminated. The seven are spherical aberration, coma, astigmatism, distortion, curvature of field and two chromatic aberrations (along two axes, longitudinal and lateral). Besides controlling these aberrations the lens designer must deal with one more variable, the focal length of the lens, which determines the magnification.

Faced with eight dependent variables (seven aberrations and the focal length), the designer must have at least that many independent variables under his control or he is helpless to effect a solution. For a given selection of glasses the independent variables, or degrees of freedom, available to the designer of a three-element lens are as follows. There are two separations: the distance from the first element to the second and from

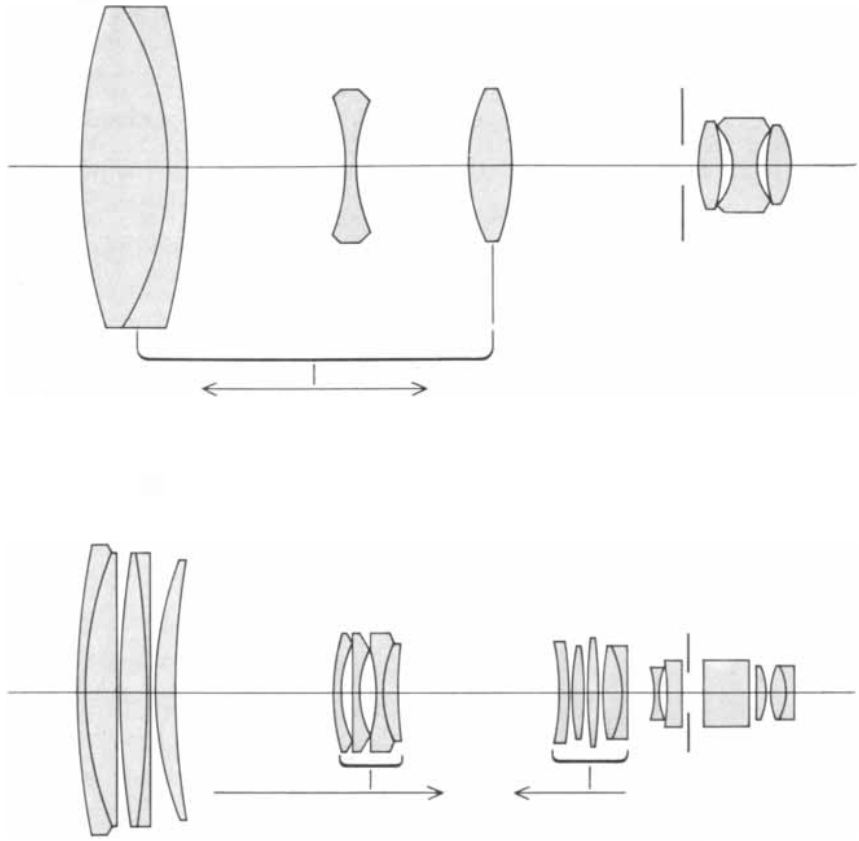
the second to the third. For each of the three elements the designer can choose the power, or magnification. Finally, each lens element has one independent surface of curvature. (The other is fixed by the choice of the first curvature and the power.) Thus, given enough experience and time, the designer can in principle find some combination of the eight variables in a Cooke triplet that will eliminate all third-order aberrations.

The second century of the photographic lens has seen a different line of progress. In 1927 George W. Morey of the Geophysical Laboratory of the Carnegie Institution of Washington realized that the formulation of optical glass was a field dominated by tradition. He was convinced that many potentially interesting formulations remained to be explored, although he was not sure what new properties might be most useful. He approached Charles W. Frederick, the chief lens designer of the Eastman Kodak Company. Frederick was interested. To answer Morey's question Frederick's department designed a number of lenses incorporating hypothetical (and unobtainable) glasses. Morey and Frederick came to the conclusion that what was really needed was a glass with both low dispersion and a refractive index much higher than that of any glass then available.

Toward the end of 1932 Kodak signed a contract with Morey. He was to experiment in the basement of his home to try to make glass of the required type. Morey's samples showed that he was making progress in the right direction, but his glass was much too dark for lenses. Although he was unable to reduce the coloration, he did arrive at the sought-after values of refraction and dispersion. To make his all but opaque glass Morey used boric oxides and the rare-earth element lanthanum.

The Kodak Research Laboratories set up a small pilot plant to determine the cause of the coloration and to eliminate it. Analysis indicated that it was due to impurities, mostly metal oxides, introduced in the making of the glass. Using a platinum crucible to make the glass eventually reduced the coloration to yellow, which was bad for many lenses but was quite acceptable for the lenses used by the U.S. in aerial-reconnaissance cameras during World War II. The yellow eliminates some of the effects of atmospheric haze.

Continued research on chemical purification removed the last vestige of color by reducing the impurities in the glass to the level of less than one part per billion. For some glasses this required a gold crucible rather than a platinum one. These glasses came to be known as EK glasses after their prefixes in the Kodak catalogue. The hypothetical lens designs had been prophetic. Today all manufacturers of optical glass make rare-earth glasses. At least one lens element of such glass is employed in practically every high-performance photographic objective made in the world.



"ZOOM" LENSES, which can vary the magnification of the image over a considerable range, also became practical with the advent of antireflection coatings. The seven-element zoom lens at the top, designed in the early 1960's for eight-millimeter motion-picture cameras, has a relative aperture of *f*/1.9 and provides a continuous range in focal lengths from 10 to 30 mm. The first and third components move together to achieve the threefold change in focal length while simultaneously keeping the image focused on the film. The zoom lens at the bottom provides a 20-to-one range in focal length, a range widely used in television. The second and third groups of elements are moved in opposite directions nonlinearly to produce the zoom and maintain focus. Large number of elements are required to control lens aberrations over long zoom range.

Almost predictably, the new rare-earth glass created a new problem even as it solved many old ones. Since internal reflection in a lens increases with the lens's refractive index, the new lenses were more susceptible to flare, or non-image light. Flare has more effect on photography than one might think. It destroys information and is therefore analogous to noise in a communications system.

It had been known since 1936 that a thin coating of transparent material on the surface of a lens could counteract flare, and that for any one wavelength such a coating could in fact eliminate flare entirely. The coating materials had to have a refractive index equal to the square root of the refractive index of the lens and had to be applied with a thickness of a quarter of the given wavelength. Such a coating not only cancels out reflection of that wavelength but also increases the transmission of light through the lens.

Repeated attempts to coat the surface of glass chemically with a thin film were inconsistent and unsatisfactory. The key was turned in 1936 when John D. Strong of the California Institute of Technology reported success in evaporating a film of fluorite (calcium fluoride) on glass in a vacuum. The first fluorite coatings did not adhere well, however. They were soft and rubbed off easily. The problem was solved by heating the lens during the deposition of the coating to drive off impurities. For virtually all applications fluorite has now been supplanted by more durable coatings of magnesium fluoride.

Among the lenses that lens coating and rare-earth glass made practical was the "zoom" lens, which had first appeared in the early 1930's. (The Busch Vario-Glaukar of 1931 was the first.) Zoom lenses have grown from the early seven-element configurations with a

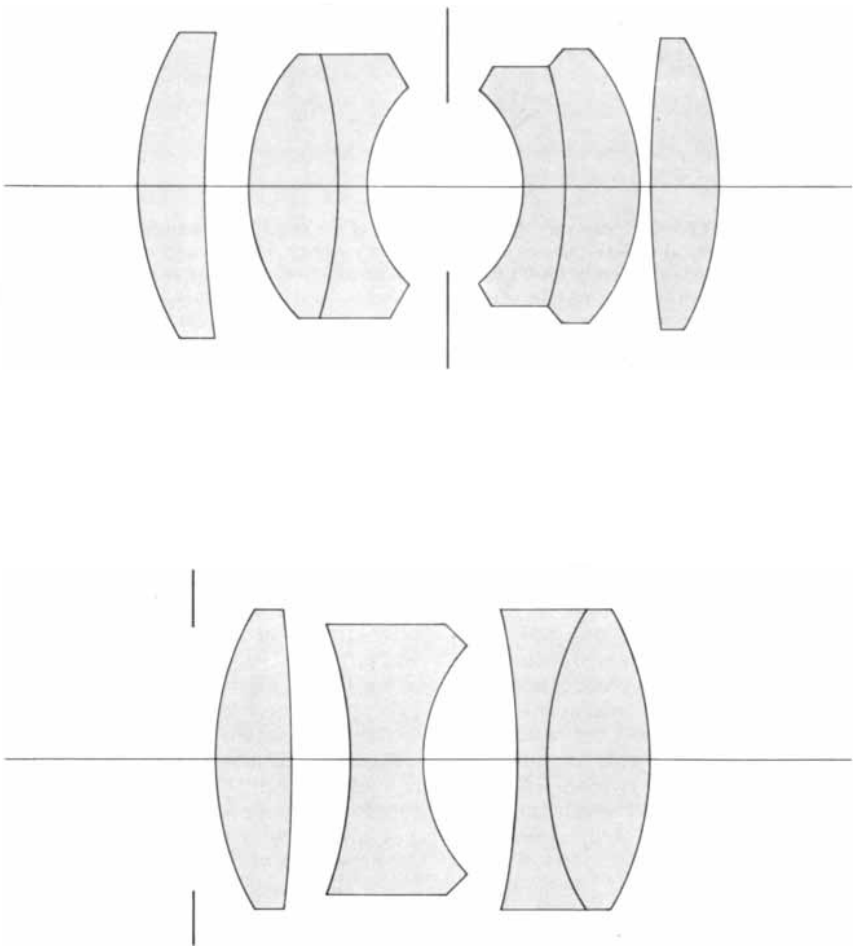
zoom ratio of three to one to today's 20-to-one lenses used in telecasts of sports events to carry the viewer from the tee to the green or from a close-up of the quarterback to a view of the entire play. These lenses contain 20 or more elements and are equipped with motor-driven zoom, focus and aperture controls.

Given relative freedom from flare, designers of high-performance lenses could abandon an expensive path they had long trod. To avoid flare they had been creating lenses with many elements and perfectly mated surfaces that were cemented together to minimize interfaces between glass and air. They could now design lenses of the Gauss type, which exploit glass-air interfaces. A Gauss lens may have as few as four elements but it has at least eight such interfaces. No mating is required because air conforms perfectly to any curved surface. Moreover, for every two elements separated the lens designer acquires another independent variable to work with. Today practically all the fastest camera lenses (of *f*/2 or lower) use air-spaced elements.

New process-control techniques have made it possible to deposit high-efficiency antireflection coatings consisting of multiple layers to minimize reflection over the entire visible spectrum. For several years devices and techniques developed by the Optical Coating Laboratory, a California company, have been widely used in Japan to make multiply coated camera lenses.

Once a water white rare-earth glass was achieved, attempts to make low-dispersion glasses with a high index of refraction were not forgotten. Whereas in the 1930's the aim was to get a low-dispersion glass with a refractive index of 1.75 (as compared with a value of 1.62 for high-index crown glass), in the 1970's optical-glass manufacturers succeeded in making experimental glass with a refractive index of 2.01 and relatively low dispersion.

The Kodak glass plant is now able to make economical production melts of water white, relatively low-dispersion glass with an index of 1.95 to 2.0. This unique material has given lens designers the freedom to create more effective lenses without incurring prohibitively high manufacturing costs. One application lies in increasing the aperture of a camera lens to make it possible to take color pictures at light levels that are lower than those currently practical without long exposures or flash. A lens aperture of *f*/1.9 is desirable. Heretofore *f*/1.9 lenses of suitable quality required at least six elements. The new glass makes it possible to produce an *f*/1.9 lens of the same quality with four elements [*see illustration on this page*]. Many more people should be able to afford the four-element lens when it reaches the market in the near future than can now afford a six- or seven-element lens.



NEW HIGH-INDEX GLASSES enable designers to reduce the number of elements needed to achieve a given level of freedom from aberration. The two lens configurations shown here are both *f*/1.9 and yield images of equivalent quality. In the six-element Gaussian lens the refractive index of the elements ranges from 1.6 to 1.75. In the four-element Tessar lens the refractive index of elements ranges from 1.9 to 1.95. An added benefit of the compact Tessar design is that the aperture stop, or the diaphragm, of the camera can be located in front of the lens, thus making it possible to align the four elements of the lens precisely in a single mount.

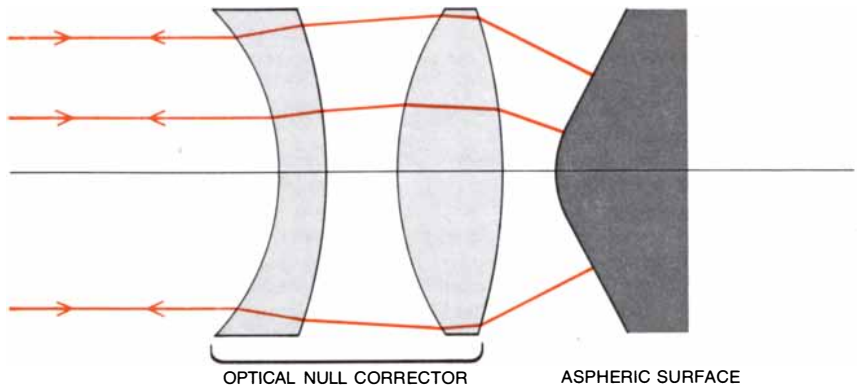
In an early effort to reduce costs camera-lens manufacturers experimented in the 1930's with lenses made of plastics. With the compression-molding techniques of that time it was not possible, however, to achieve the mirror-smooth surfaces required. Although satisfactory lenses could be made by casting, the process was too slow and expensive. Ultimately injection molding proved to be more promising than either of the other methods. In 1952 viewfinder lenses made of molded transparent plastic were introduced in box cameras. This success led by 1957 to the use of plastic for some of the simplest camera lenses, and by 1959 triplet lenses were being made by injection molding. That was no mean achievement, since several difficult problems had to be overcome.

One problem with plastic lenses is thermal change. The lenses are less dense and their index of refraction is lower in warm weather than it is in cold. At Kodak we puzzled over this for some time. We found we could design the lenses so that as one element changes with the heat to shift the focus to the rear, a compensating element changes with the heat to shift the focus forward by an exactly corresponding amount.

A more insidious problem was that of achieving a strain-free lens element. Strain in a lens has unwelcome optical properties. What is more, the strain in plastic is eventually relieved by thermal cycling (as happens, for example, in cold weather when a camera is moved from indoors to outdoors and back again). As the strain diminishes with many such cycles, the dimensions of the lens change enough to ultimately degrade its image.

We tried every kind of steel mold. None would yield a strain-free product. Finally we found a special ceramic that can be polished to the required high luster and has heat-transfer characteristics that make it possible to fabricate a lens element that is free of strain. This technique has also made it practical to coat plastic lenses to reduce reflections, just as is done with glass lenses.

Meanwhile we have continued to study the molding process and have modeled it mathematically with the intention of getting closer tolerances. We are now able to hold tolerances through the center of a lens of 1/1,000 inch and across the diameter of a lens of 1/10,000 inch. We can also hold surface-contour tolerances of 1/200,000 inch during 30,000 molding cycles. Since as many as 16 lenses are molded in each cycle, we can produce 500,000 lenses before a mold has to be reconditioned.



TESTING OF ASPHERIC SURFACES, recently introduced in some lenses, creates special problems. One testing scheme uses an optical null corrector, which converges an entering parallel beam of light (one with a plane wave front) into a beam with an aberrated wave front that conforms to the aspheric surface. Hence light reflected from the aspheric surface retraces the incoming path and emerges from the null corrector as a plane wave again. The unit is placed in one arm of an interferometer. Any departure of the aspheric surface from the desired shape shows up in the interferometer image as a deviation of the emerging wave front from a plane.

The spherical surface of a lens is only an approximation of the ideal surface. In a reflecting telescope the ideal mirror has a parabolic section. Perfect lenses would have a surface of rotation slightly more complex than that. Aspheric, or nonspherical, refraction optics have been in use since Bernhard Schmidt somewhat accidentally discovered in 1930 a manual technique for making an aspheric correcting lens to work in conjunction with a spherical telescope mirror. The process, still in service today, uses the surface tension of glass to retain a smooth polish when a plate is heated and allowed to sag into a mold. The surface that goes against the mold will be poor and is subsequently ground and polished flat, but the free surface retains its original high polish in its "sagged" contour. The result was the wide-field Schmidt telescope. Apart from this application, however, aspheric lens elements are available commercially only in certain professional motion-picture camera lenses and in some half a dozen lenses for 35-millimeter cameras. The 35-mm. camera lenses, one made in West Germany and the others in Japan, have speeds of up to *f*/1.2 and retail for about $1,000. High cost has put aspheric lenses out of reach for most nonprofessional photographic purposes.

Aspheric molds for plastic lenses promise to change that. We have now developed practical techniques of making aspheric molds for plastic lenses with the precision and repeatability required for photographic lenses. So far lenses with one aspheric surface on a plastic element are able to gather twice as much light (that is, they are one full *f* stop larger) as an all-spherical lens comparable in quality and number of elements. The testing of aspheric surfaces has required the development of new techniques. In one such technique a lens called an optical null corrector converts a parallel beam of light into an intentionally distorted, or aberrated, wave front that is directed at the aspheric lens surface to be tested. If the test surface has the proper aspheric curvature, it reflects the aberrated incident light in such a way that when the light passes through the optical null corrector, the parallel plane wave front is restored. Placed in an interferometer, the aspheric surface can now be tested in the same way that a spherical or plane surface is.

With injection molding we have found it possible, and sometimes preferable, to mold a plastic mount directly onto a ground, polished and coated glass lens element. Mounting problems present some of the severest trials of glass lens manufacture. Injection molding provides a degree of exactness and repeatability in lens assembly that is not otherwise economic.

As matters now stand, glass is superior to plastic in dimensional stability, elasticity, hardness and refractive index. Optical plastics are still limited to a few types of polymer, mainly acrylic, styrene and styrene-acrylonitrile. Plastic camera lenses, which we now confidently predict can be mass-produced with aspheric surfaces, will require fewer elements for the same results and will offer the possibility of higher-order correction.

An alternative to aspheric surfaces for correction of the limitations of spherical surfaces is gradient-index glass. As we have seen, spherical aberration is typical of these limitations. The spherical surface is too strong at the edge. This may be overcome by aspherizing one surface to make it weaker toward the edge, or alternatively lowering the refractive index toward the edge. Research on gradient-index glass is being actively pursued at the Institute of Optics of the University of Rochester in cooperation with Bausch & Lomb Inc. At the Kodak Research Laboratories gradient-index plastics are under development, and research on the mathematics required for the design of gradient-index lenses is in progress. Since light in a gradient-index medium does not travel in a straight line, the mathematical calculation of the light path and the wave-front contour becomes significantly more complex. Attending these developments is the need for specifications, tolerances, testing techniques and the manufacturing control of the gradients. Although this field is now in its infancy with respect to photographic lenses there seems to be room in the future for both aspheric surfaces and gradient refractive index.

My predecessor as the manager of the Optical Design Department at Kodak, Rudolf Kingslake, has often remarked that lens designers may have benefited more than anyone else from the introduction of computing machinery. We certainly would be hard put to calculate our aspheric surfaces and zoom lenses without them.

By 1950 ray-tracing programs had been written for a number of computers, including the IBM card-programmed calculator, the National Bureau of Standards Eastern Automatic Computer and the Harvard Mark I. By 1954 work was under way on the automatic design of lenses at Harvard, at the University of Manchester and at the Bureau of Standards. In 1956 Kingslake hired Donald P. Feder to develop a practical automatic-design program for Kodak. In the beginning we shared the company's business computer, but in time the demand for scientific computing warranted separate facilities. These have been brought

up to date several times through the years.

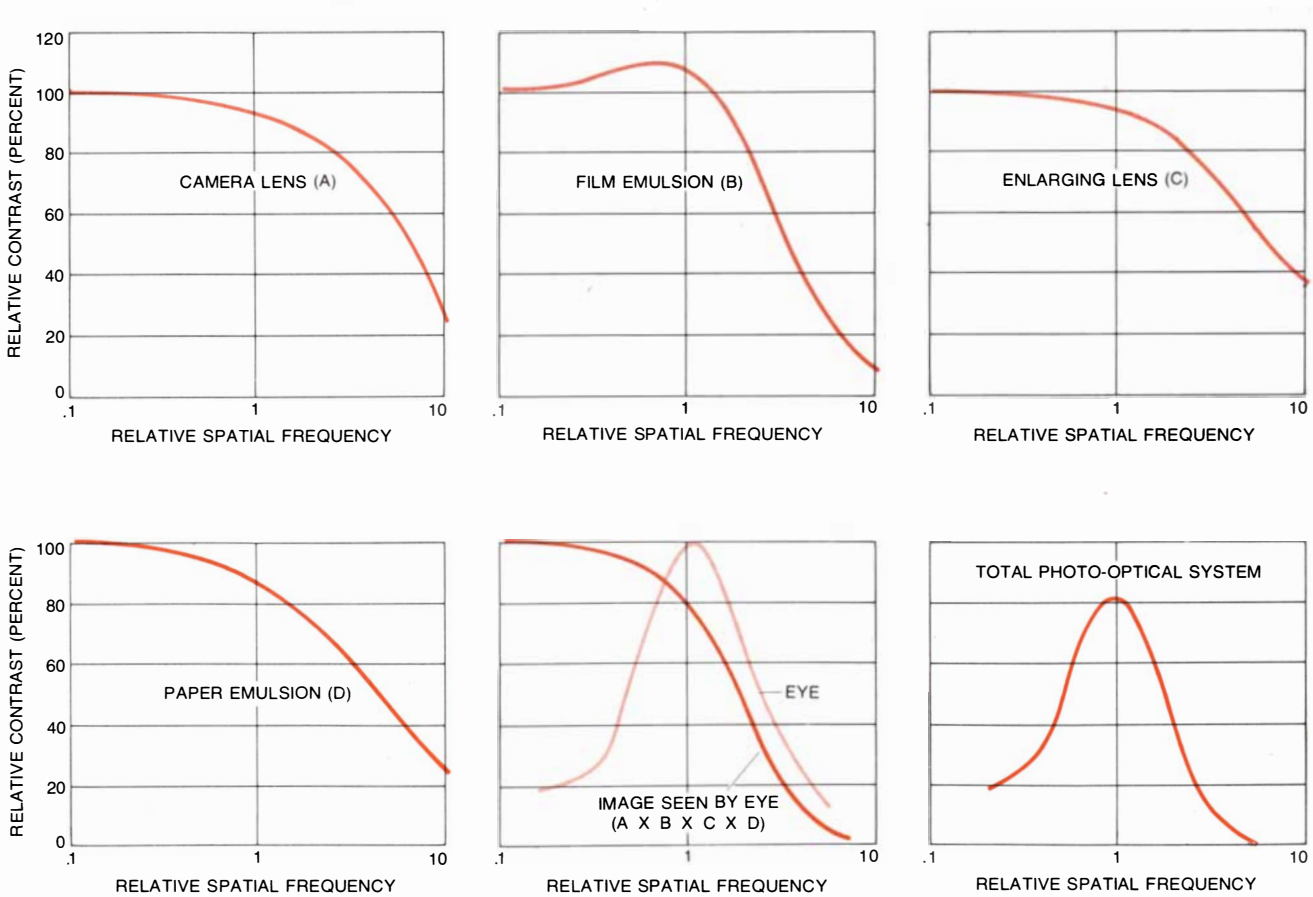
Feder had developed a program at the Bureau of Standards to analyze lens designs for the Air Force. It cost about $2,000 to analyze a lens by this method in 1956. In 1957 Feder wrote a more ambitious program for Kodak that performed the analysis for $100. A 1971 program written by Philip E. Creighton at Kodak was able to produce a complete set of analyses for eight focal planes, five wavelengths and five field angles of a lens with up to 12 surfaces. The analyses cost less than $5.

The real strength of the computer, however, lies not in analysis but in its ability to improve a lens design. The aim is to reduce lens errors to an acceptably small amount. We should like in fact to reduce the errors to their mathematical minimum, but that is not possible because to do so would require solving a large number of simultaneous nonlinear equations with a large number of unknowns. The task is beyond the reach of modern mathematics. What we can do with the powerful assistance of a computer is to arrive at a series of successively closer approximations to a flawless lens.

That this could be done in short order was publicly demonstrated during a symposium on optics at the University of Rochester in 1962, when Feder and his colleagues designed a four-element lens from beginning to end during an evening session. The job took two and a half hours of machine time, and the design became known as the "symposium lens." Feder estimates that present-day computers working with the same program would take a couple of minutes. When it was applied later that year to the improvement of a high-quality microfilm lens that had been designed by hand, the program provided a more precise optical system that was cheaper to manufacture. The day of automatic optical design had arrived.

Since the early 1960's lens-design computer programs have been developed at the University of Rochester and other universities, at lens-manufacturing companies throughout the world and by independent consulting firms such as David Grey Associates, Inc., of Waltham, Mass. There is no longer any doubt that the best of the computer programs will produce results superior to those that can be achieved by precomputer methods. It is estimated that the use of computers has increased the productivity of lens designers tenfold. Moreover, designers are now much more confident than they were in the reliability of their predictions regarding lens performance and manufacture. Part of this confidence stems from a Monte Carlo technique for analyzing the sensitivity of a lens design to cumulative manufacturing variations within the limits of normal process control. The technique makes it possible to decide whether or not a lens can be effectively manufactured.

Professional and advanced amateur



MODULATION TRANSFER FUNCTION of a complete photographic system involves the modulations, or losses, introduced at each step. In this analysis fine details in an image are regarded as variations in light intensity in space in the same way that variations in signal strength with time are regarded by a radio engineer when he is evaluating the performance of radio equipment. At each step in a photographic system fine details (high spatial frequencies) are reproduced with loss of contrast (that is, modulated). These curves are calibrated with the performance of the eye normalized to show peak response (100 percent relative contrast) at a relative spatial frequency of 1. Unlike a lens, the eye degrades the contrast of images that are both lower and higher than 1 in relative frequency. The image seen by the eye is the product of the first four modulation curves (lens times film emulsion times enlarging lens times paper emulsion). The quality of image transmitted to the brain is proportional to area under last curve, which is the product of modulation curves of all stages.

photographers used to argue vigorously over the merits of various lenses, particularly the high-performance lenses made for expensive 35-mm. cameras. Although lenses could be tested objectively with resolution charts, it was generally recognized that resolving power alone (defined as the number of high-contrast lines per millimeter that a lens can resolve in various parts of the image field at various apertures) is a surprisingly unreliable guide to the quality of pictures a lens might produce. Much of the mystery surrounding lens "quality" was cleared up in 1951 when Otto H. Schade, Sr., of the Radio Corporation of America described his investigations of the lenses used in the entire chain of information transmission represented by a television system. Schade was able to show that the recording of fine detail was not necessarily related to the general efficiency of information transfer. His most surprising result was that some highly rated lenses were not as good for television purposes as lenses that were thought to be inferior.

Schade's investigations added a dimension to the Rayleigh definition of image quality. The Rayleigh criterion is now seen as being a limiting case: it locates one end of the quality continuum. It tells us when a lens is approaching perfection but does not tell us which of two imperfect lenses is the better.

By regarding light-intensity variations in space across the image formed by a lens as the radio engineer regards signal-strength variations in time when he measures the performance of transmitters, receivers and amplifiers, Schade was able to apply information theory to lenses and to provide for them an "optical transfer function." The fact that the transfer function coincides rather well with criteria that lens designers have been using for generations is an indication of its validity. Even more important, the transfer function of a lens can be combined with the transfer function of film, of photographic printing devices, of projection lenses and so on. The transfer function can be computed for the lens design and also measured on the manufactured lens. Hence with the computer we can mathematically model the entire photographic system, beginning with the subject and ending with the transfer function of the viewer's eye.

The comparison of such objective calculations and measurements with people's subjective reactions to the corresponding photographic results tells us, in terms of design requirements, just what constitutes a better picture. Such conceptual models have greatly helped the photographic industry in deciding just where to concentrate its research-and-development efforts, thus improving the ratio of quality to cost for people who make pictures.

The Social Behavior of Burying Beetles

These insects, usually working as a mated pair, can quickly inter a carcass many times their size. The carcass serves them and their larvae as food, and both parents participate in caring for the young

by Lorus J. Milne and Margery Milne

An observant person who sees the carcass of a small animal lying on the ground is likely to find, if he returns to the spot the next day, that the object has disappeared. The chances are that it has been buried, either there or nearby, by a pair of *Nicrophorus* (or *Necrophorus*) beetles. They will use it as food for their young during the larval stages. The feat of these small insects in rapidly interring a carcass that is many times their size is remarkable enough, but it is only a step toward the most advanced form of parental cooperativeness known among the Coleoptera (the beetles). We have spent much time watching these burying beetles (also called carrion beetles and sexton beetles) and putting them to various tests, which demonstrate an impressive plasticity in the behavior of the insects.

The patient French naturalist Jean-Henri Fabre set out fleshy bait of several kinds to lure burying beetles to where he could watch them. He admired these little gravediggers of the animal world, describing them as being "elegantly attired" in black, with a "double, scalloped scarf of vermilion" across their shining wing covers. The observer cannot watch for long. Unlike the scarab beetle of Mediterranean countries, which walks in plain view while rolling a ball of dung to some still undiscovered place of burial, a burying beetle quickly slides out of sight below the carcass of a mouse or a bird it has found. There, lying on its back, the insect uses all six of its powerful legs as levers to shift its prize. From time to time it rights itself and bulldozes headfirst into the earth to loosen the soil and push it away. Inconspicuously, a fraction of an inch at a time, the carcass moves horizontally or disappears into the ground.

Nicrophorus beetles are by no means the only insects that sequester food for their larvae before they lay the eggs that will give rise to the larvae, but they work as a team, whereas the others (the scarab and its relatives and various solitary bees and wasps) work alone. Either a male or a female *Nicrophorus* will initiate the flexible behavior that gets the larval food into a safe place. At any time during the operation a mate is likely to arrive. The partner is accepted with no time off for courtship. The two labor together at intervals and also separately in a loose cooperation that advances the common effort. Yet either member of the pair may also creep into a more or less concealed place and appear to sleep for as much as half an hour or depart on feet or wings to some unknown destination for a comparable period, thereafter returning and resuming the work. Ordinarily copulation is deferred until the beetles are securely in possession of their carrion in a chamber of their own making, an inch or more below the surface of the ground.

At this juncture the male might be expected to perform his brief sexual duty and depart. The inseminated female would then carry on alone to the end of the sequence of behavior specified by her inheritance. Occasionally this pattern is followed, but usually both parents remain. Together they work the mass of food into a compact ball. They free it of fur or feathers, perhaps adding secretions that modify the course of decomposition.

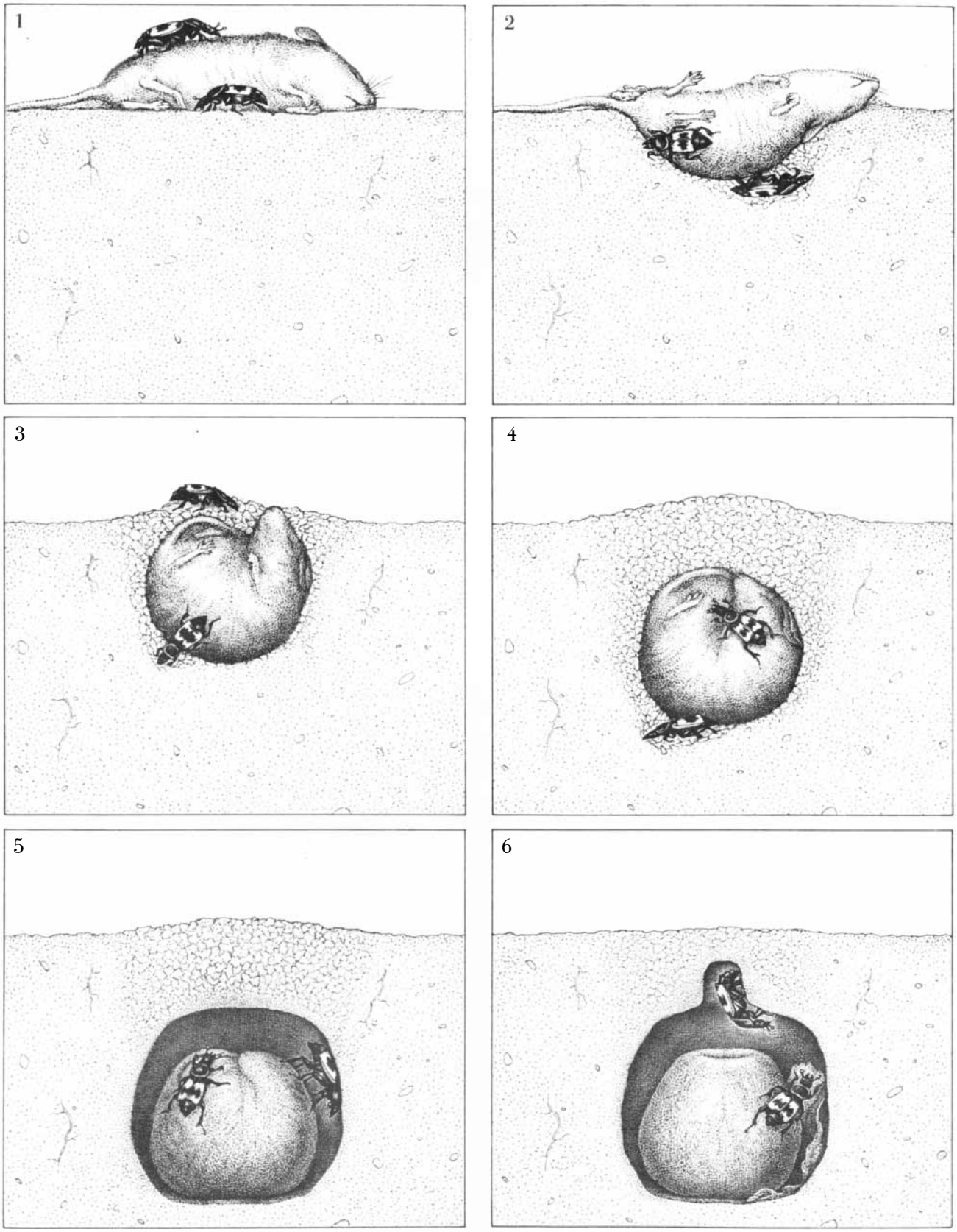
As the insects clamber around the carcass, which will provide food for them as well as for their young, the floor, the walls and the roof of the earthen chamber become firmly packed. The female constructs a short vertical extension of the chamber above the carrion and lays her eggs in the side walls of the passageway. She returns to the carcass, and by a combination of selective feeding and clawing at the upper surface prepares a conical depression. Both beetles regurgitate into the depression droplets of partly digested tissue. The fluid accumulates as a pabulum for the larvae that will soon hatch.

This much Fabre or any other persistent observer could discover by exhuming the beetles and their food supply at the proper time, just before the young hatch. Erna Pukowski, studying species native to her Polish countryside, managed to learn more. She made captive burying beetles so much at home, notwithstanding the unnatural condition she created by illuminating their burial chamber, that she could follow the next steps.

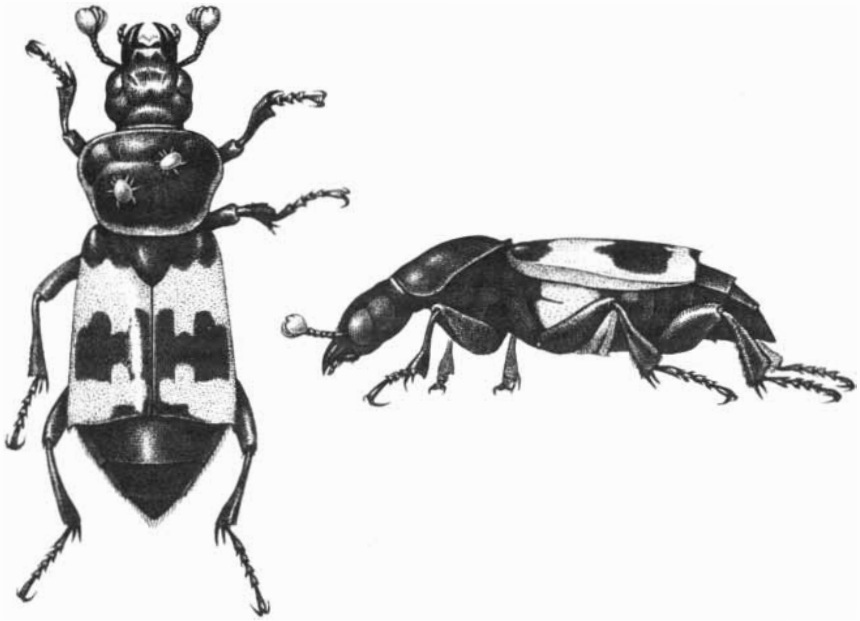
One beetle (perhaps the female, although the members of a pair are too much alike externally for an observer to distinguish sex) stood beside the pool of liquid nourishment and began to stridulate. The sound brought hatchling larvae (some two or three millimeters long and almost like maggots in appearance) to the parent's side. The parent sipped from the pool and then transferred the fluid food to one larva after another. The larvae lifted their mouth ends, the better to receive the food. Sometimes both parents shared in the feeding operation.

The British entomologist R. L. Morley discovered in 1902 that the sound of stridulation arises when twin plectrums on the inner surface of the cover of a beetle's wing fret against crosswise ridges on the fifth segment of the abdomen. The sound is clearly audible to human ears. Pukowski noticed it also during the three or four seconds when copulation is in progress. We have heard it when burying beetles are under stress, as they are in repelling an insect of another species or a smaller member of the same sex and in confronting an obstruction that impedes the movement of a carcass.

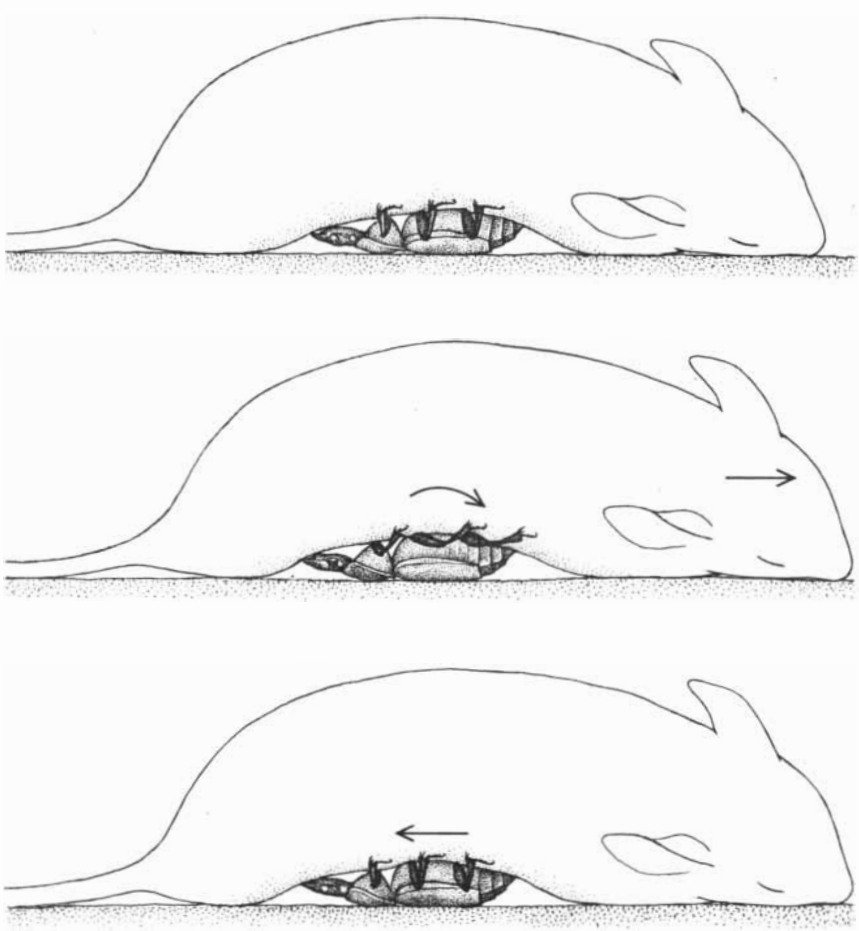
In 1972 Carsten Niemitz of Justus



BURIAL OF MOUSE by a pair of *Nicrophorus* beetles is depicted. Usually a carcass is discovered by one beetle, which is soon joined by a mate. Here a male and a female, indistinguishable externally, are shown as they inter the carcass of a mouse in a burial chamber they prepare by moving earth and packing it radially and upward. The scale is about two-thirds life size. At the end of the process (*6*) the carcass is about an inch below the surface of the ground. The beetles have shaped it into a ball and have removed the skin and tail. They create a pool of liquid food for their larvae in the top of the ball, and the female then lays her eggs in a small chamber above the pool.



BURYING BEETLE of the species *N. marginatus* is portrayed in top and side views. The beetle employs all six of its powerful legs in moving and burying a carcass. A beetle is likely to carry at all times a cargo of mites of the genus *Poecilochirus,* which evidently live in some kind of symbiotic arrangement with the beetles. Two such mites appear in the top view of the beetle.



TRANSPORT OF CARCASS is a technique that *Nicrophorus* beetles employ when the ground where they find the body is too hard. The beetle lies on its back and uses its legs as levers to shift the carcass. If a pair of beetles are present, they work together at times and also separately in a loose cooperation. They will transport a carcass several meters if necessary.

Liebig University in West Germany discovered that very young larvae will orient themselves to the sound of an adult's stridulation recorded on tape. This response disappears, however, after the larvae have molted for the first time. Even so, older larvae renew their solicitation of regurgitated food for a few hours after each molt by approaching any adult that is close to the pool of food and pressing their mouthparts against its jaws or palps. This action stimulates regurgitation as before. Otherwise the growing larvae feed directly from the pool or pull fragments from the surface of the carrion.

The larvae receive parental care all through their period of feeding growth. The parents may even prepare a horizontal passageway into which the fully grown larvae can crawl to pupate. Only then, when the adults can contribute nothing more to their brood, do they force their way upward through the soil and fly away.

We have not yet marked and followed the departing parents to see whether they do the same thing all over again. They probably do, since adult beetles live from three to 15 months, depending on the species. They search widely for the odor of recent death and are remarkably efficient at finding carrion. Frantisek Petruska, a Czechoslovak ethologist, has found, by capturing beetles with carrion bait, marking them and releasing them at various distances, that they will return to the carrion within 24 hours from as much as four kilometers away.

During one period of four hours, beginning just 35 minutes after we had laid out a newly dead mouse on birch-leaf litter, nine burying beetles arrived. Each beetle followed the guidance of the olfactory organs in its antennae. It dropped to the ground within three meters of the mouse, quickly folded its flying wings under its wing covers and came crashing through the litter to the carcass. There, after only a moment's hesitation, the beetle turned over onto its back, slid under the body and lifted the mouse slightly from the ground, apparently to test whether the body was movable. Emerging on the other side of the mouse and righting itself, the beetle began testing the soil.

We had placed our bait on hard ground. Each beetle rejected the site for burial and began to explore, seemingly at random, for softer earth. This was our cue to remove the active beetle and wait for the next one. Each one of the nine beetles followed essentially the same routine, even though four of them were members of one species and five were members of another.

Competition for small carcasses is frequently intense. Ants and flies (par-

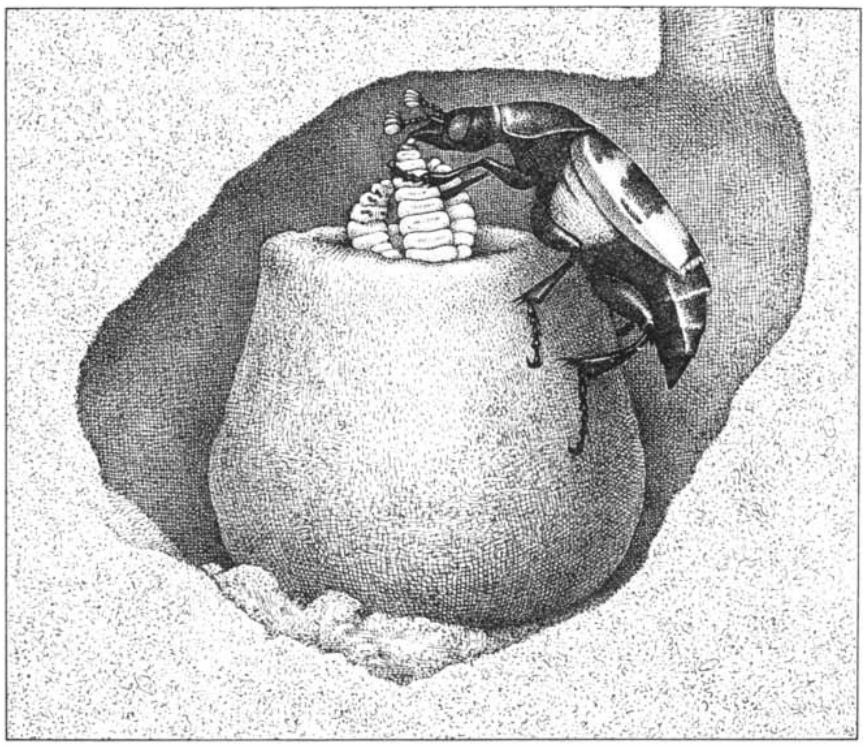
ticularly blowflies, which deposit active maggots) tend to take over during the day. Burying beetles of the species that are most active by day succeed only if they can inter a body quickly. For species that are active at night the competition is mainly from other species. The largest beetle generally repels all the others except a mate. That is probably why one more often finds a large male cooperating with a small female (or a large female cooperating with a small male) than one finds two beetles of the same size cooperating.

Burying beetles have other ways of reducing competition. Each species has a preferred combination of temperature range and relative humidity. This pattern, as Jean Théodoridès of the University of Paris showed in his laboratory, keeps certain beetles in woodland and others in open fields. Burying beetles that are active in the spring belong to species that go through the winter as adults, whereas the beetles found competing in the summer are likely to represent species that spend the cold months dormant as pupae or full-grown larvae.

Animals that eat insects are likely to constitute a hazard for burying beetles that are active by day. At least one diurnal species of burying beetle in Europe and one in North America may escape being eaten as a result of their resemblance to a small bumblebee. Unlike most burying beetles, these species have golden hair over some of their hard black surfaces. Color, sound and style of flight combine so convincingly that the British biologists Charles Lane and Miriam Rothschild have suggested that this is an example of mimicry, at least with respect to sound. Even a superficial resemblance might have survival value for the beetles.

The most spectacular feature of the activity of burying beetles is the way they transport a carcass from hard ground to soft, in one steady direction, a fraction of an inch at a time. A beetle that has yet to acquire a mate may identify a suitable burial site several meters away from the carrion. The beetle will alternate between loosening the earth at the burial site and rushing back to the carcass. There it performs its lifting feat, starting under the body at the end closest to the burial site. The dead weight is progressively shifted until the beetle emerges from under the opposite end. The insect may run around the carcass and repeat the process time after time. If a mate arrives, the progress is more nearly continuous.

A measure of the success of this way of life can be seen in the fact that the genus *Nicrophorus* includes almost 100 species, with some overlap in distribution. About half of the species are Asiatic. Almost the only areas where burying beetles have not been found are the West Indies, Africa south of the great deserts, Australia and New Zealand.



FEEDING OF LARVAE by a burying beetle of the species *N. vespillo* is shown on the basis of photographs made by the Polish naturalist Erna Pukowski. The parent beetle sips from the pool of fluid food in the top of the buried carcass and then transfers the material to one larva after another. The larvae rear up instinctively, much the way nestling birds do when feeding.

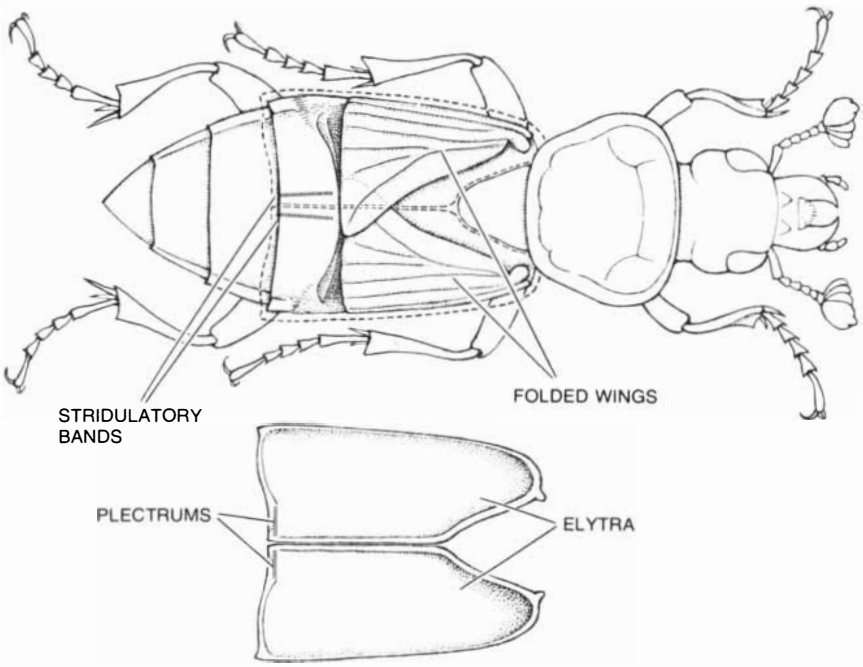
Adult burying beetles range in length from 10 to 35 millimeters, with considerable variation within a species. All of them appear able to transport the compact body of a bird or a mammal weighing up to 100 grams—up to the size of a rat or a big robin. Anything heavier is usually abandoned unless it is only slightly overweight and can be interred where it is found. A dead snake, however, can weigh more and still be buried expeditiously. Its carcass is subdivided into two or more zones of operation. One pair of beetles attends to each zone.

As Fabre observed, burying beetles show considerable plasticity in behavior. Noting the number of obstacles a typical environment is likely to present to beetles trying to move or bury a body, he wrote that the insect therefore "cannot employ fixed methods in performing its task. Exposed to fortuitous hazards, it must be able to modify its tactics within the limits of its modest discernment. To saw, to break, to disentangle, to lift, to shake, to displace— these are so many means that are indispensable to the gravedigger in a predicament. Deprived of these resources, reduced to uniformity of procedure, the insect would be incapable of pursuing its calling."

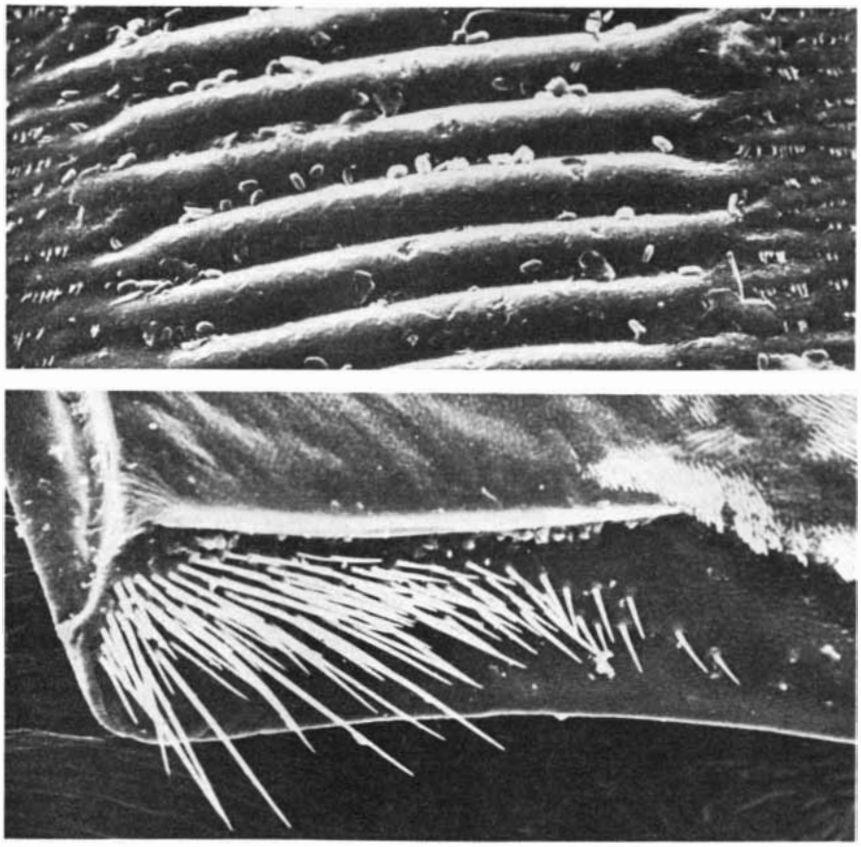
Fabre's evaluation rested on watching the beetles at work rather than on experiments. He was reluctant to disturb the carrying and burying operations because so few *Nicrophorus* beetles came to his bait. Our studies have centered in countryside (New Hampshire and Ontario) where richer woodlands and more varied fields support a larger population of beetles. Simple tests confirm the versatility encompassed within the insects' programmed patterns of behavior.

To create a reasonable facsimile of the type of obstacle a burying beetle might encounter naturally, we place a dead mouse close to a clover plant and then tie the carcass down by arching the stalk of a leaf over the torso and fastening the stalk to the ground with a hairpin. The first burying beetle to arrive discovers that the front and rear ends of the carcass can be raised but that the middle cannot. The beetle promptly climbs over the mouse, discovers the tight leaf stalk, forces its head under the stalk and pushes forward. The stalk does not break, but it stretches enough to release the carcass for transport and burial. Repetitions of the experiment with other beetles all have the same result.

Once we drove a good-sized stake into the ground at a 45-degree angle and tied a strong cotton string around its upper end. We tied the dangling end of the string around a hind leg of a dead mouse lying on soft ground. A pair of *Nicropho-*



STRIDULATING MECHANISM is employed by a *Nicrophorus* beetle to call larvae to food and also in times of stress. Here the elytra, or wing covers, are shown (*below*) removed from the beetle's back and turned over, so that the plectrum at the bottom of each elytron is visible. The sound is made when the plectrums are rubbed against ridges on a segment of the abdomen.



STRIDULATORY APPARATUS of the *Nicrophorus* beetle consists of the *pars stridens* (*top*), which has the crosswise abdominal ridges the plectrums rub against, and of the plectrums, one of which is shown at bottom. The plectrum is the ridged white structure that is in a form approximating a right angle; the bottom part does the striking and the part at the left serves as a guide. In these scanning electron micrographs, made by Rolf Schumacher of the University of Bonn, the *pars stridens* is enlarged about 445 diameters and the plectrum 120.

rus beetles pushed away the soil below the body until the mouse hung from the tethered leg over a cup-shaped depression. The insects cleared a space the thickness of their bodies between the mouse and the soil and then kept swiveling the carcass in wide arcs. The tail of the mouse dragged on the rim of the depression until one of the beetles chewed it off.

That did not solve the problem, and so both beetles explored the surface of the carcass. Only about six hours after they had begun to work did one of them discover the tether. In less than a minute the insect settled down to gnaw through the cotton fibers. By dawn the carcass had been liberated and buried.

To test the strength of *Nicrophorus* beetles we rested one end of a flat rock on the body of a 50-gram mole. The rock applied about half a kilogram of unyielding weight to the body. Two beetles were nonetheless able to work the body free. First they took up positions side by side with their back against the rock and their legs against the body. They shifted the body about a centimeter in relation to the rock and then repeated the performance with respect to the hard soil below the body. Alternating between these two areas of contact, they freed the carcass in less than half an hour, whereupon they transported it to soft ground and quickly buried it.

In tests of the memory of burying beetles we have found that if a beetle has had 15 or 20 minutes of experience with a suitable carcass, it can be removed and held captive for at least 16 hours without losing its readiness to return to the body within minutes of being released. After 24 hours of separation from its trophy the beetle is more likely to fly off. Two beetles of the species *Nicrophorus orbicollis* that had shifted a mouse about six inches were picked up and put in separate boxes with moist earth. Two hours later, while two smaller beetles of the species *N. tomentosus* were working on the mouse, we released the male *orbicollis* six inches to the east of the body and the female six inches to the north. They both feigned death for a few seconds and then set out almost directly for the body of the mouse. They repelled the *tomentosus* beetles and resumed their normal activities.

In this test the female was the larger *orbicollis*. If the disparity had been the other way, the behavior probably would have been different. We find that females are much more aggressive than males in ousting rivals from carrion. A male is more likely to allow other *Nicrophorus* beetles, particularly those of other species, to work for a while before repelling them. In the end each manageable carcass serves as food for the adults and larvae of only one pair of *Nicrophorus* beetles—the ones that bury it.

Beetles need both memory and some special sense (probably olfaction) to recognize a particular trophy. If we move a carcass a meter or less to one side while the beetles of a pair are momentarily away from it, they immediately begin exploring on their return to the vacant site. In a few minutes, aided no doubt by scent, they find the carcass and resume work as though nothing had happened.

Possibly burying beetles mark a carcass with a chemical secretion, which would explain what happens when the beetles return to a site where they have been working on a carcass only to find a different carcass there. They examine the substitute and then go off exploring. If they find the original carcass within a meter or less, they resume work on it. If they fail to locate their prize, they are as likely to fly away as they are to accept the substitute. A volatile substance that conferred a distinctive odor on a carcass, as a message to be read later by the same insect or its mate, might serve also as a pheromone. Pukowski noticed that a lone *Nicrophorus* beetle, after laboring for a long time without being joined by a mate, would climb on top of a plant or a stone, elevate its abdomen obliquely and extend it as though emitting a secretion.

The social behavior of burying beetles fits between extremes in the behavior of other insects. In the most primitive insect social behavior the parent or parents attend only to their own offspring. The most advanced social insects have a female at least providing care for the offspring of other females, often as a sterile surrogate parent. Burying beetles often show some altruistic behavior in that small members of the same species or a different one may contribute significantly to the rapid burial of a carcass and then leave, taking no part in reproduction. The dominant, mated pair take over the food supply. The female, at least, remains to care for the larvae, but she will not tend the larvae of any other female.

Parental interactions that promote the survival of further generations have evolved independently in more than two dozen families of insects. Among the beetles, which are the most varied order of animals, *Nicrophorus* is unique in extending maternal care so far and in having the aid of the male so often until the larvae are ready to pupate. No other members of the same superfamily show social behavior of any kind. Indeed, entomologists regard the superfamily (Staphylinoidea) and the family (Silphidae) to which the *Nicrophorus* beetle belongs as being made up of rather unspecialized beetles. It is odd that behavior of such plasticity should have arisen at all and then should have succeeded so widely.

The Curvature of Space in a Finite Universe

Curvature of a surface is an intrinsic property that gives rise to distortion of distances on a map. The same is true for curvature of space, where the map is Einstein's general theory of relativity

by J. J. Callahan

Is the universe finite or infinite? According to numerous ancient mythologies, it has a complex structure but is nonetheless finite. That viewpoint developed in Greek philosophy and culminated in the cosmology of Eudoxus and Aristotle: the earth is a ball surrounded by a series of concentric crystalline spheres, the outermost sphere carrying the fixed stars and containing within it the entire material universe. The primary purpose of this cosmology was to explain the motions of the planets and other celestial bodies. Each body was carried around the earth by the rotation of the sphere in which it was embedded. Nevertheless, an integral part of the theory was that the universe was finite.

Aristotle's picture of the world was widely accepted in medieval Europe; it appears, for example, in scholastic philosophy and in Dante Alighieri's *Divine Comedy*. In fact, Dante actually extended Aristotle's picture in a radical and thoroughly modern way. I shall take up Dante's interpretation in my conclusion. In spite of the popularity of the finite-world picture, however, it is open to a devastating objection. In being finite the world must have a limiting boundary, such as Aristotle's outermost sphere. That is impossible, because a boundary can only separate one part of space from another. This objection was put forward by the Greeks, reappeared in the scientific skepticism of the early Renaissance and probably occurs to any schoolchild who thinks about it today. If one accepts the objection, one must conclude that the universe is infinite.

The notion of infinity has always been wrapped in mystery, and historically it triggered apprehensions that have only gradually been overcome. During the scientific Renaissance, Euclidean geometry became the main instrument for comprehending infinite physical space. Euclidean geometry contends that a straight line is the shortest distance between two points, and that the sum of the angles in a triangle will always be 180 degrees. The Renaissance scientists saw that Euclidean geometry treated ideal objects in a mathematical context that was infinite, but that its axioms and propositions exactly described the spatial relations of the real world. Leibniz and Newton shared the view that physical space was infinite and Euclidean. They disagreed, however, on how matter was situated in space. For Leibniz a finite group of stars was unthinkable: such a group would have to be in some specific location in space and God would have had no sufficient reason to put it in one place rather than in some other. Leibniz thus concluded that the universe must be infinite. Newton rejected that possibility, however, on the grounds that God is the only possible actual infinity. Although today these arguments may not seem persuasive, at the time they were considered sufficient.

Who was right? Both arguments were essentially negative. Leibniz denied that the universe was finite, Newton denied that it was infinite. Neither was enthusiastic, however, about the alternative with which he was left. In 1781 Immanuel Kant offered in his *Critique of Pure Reason* a thorough analysis of the entire problem of space, including a bold and novel resolution of the dispute between Newton and Leibniz. Kant said that they were both right and that we must admit paradoxically that the universe is neither finite nor infinite! This basic contradiction between principles that seem equally necessary and reasonable is known as Kant's antinomy of space. The antinomy of space is one of several antinomies that in Kant's view pointed to "a hereditary fault in metaphysics that cannot be explained, much less removed, except by ascending to its birthplace, pure reason itself." A major aim of the *Critique of Pure Reason* was to remove such hereditary faults from metaphysics. Kant's method was drastic. He argued that since we cannot conceive of the universe as being either finite or infinite, we shall never be able to discover empirically whether it is either finite or infinite. In other words, it is not an objective property of the universe to be either finite or infinite. Furthermore, space is not a thing but is a form through which we perceive things, and we make a fundamental error when we treat space as a thing. The antinomy reflected a basic limitation in the mental processes we use to describe the world. Kant would insist that we discard our question as being meaningless.

Today Kant's metaphysical analysis of space is disregarded by modern science because its foundation—notably Euclidean geometry—has been broadened by revolutionary developments in mathematics and physics. Einstein's general theory of relativity provides a new geometry of space, and it opens another approach, unforeseen by Kant, to the question about the finiteness of the world. For Kant the question had simply been invalid. Einstein restored its validity by arguing that Kant's antinomy of space is only apparent and that it can be understood without resorting to metaphysics. In short, Einstein shows that a finite universe is a real possibility.

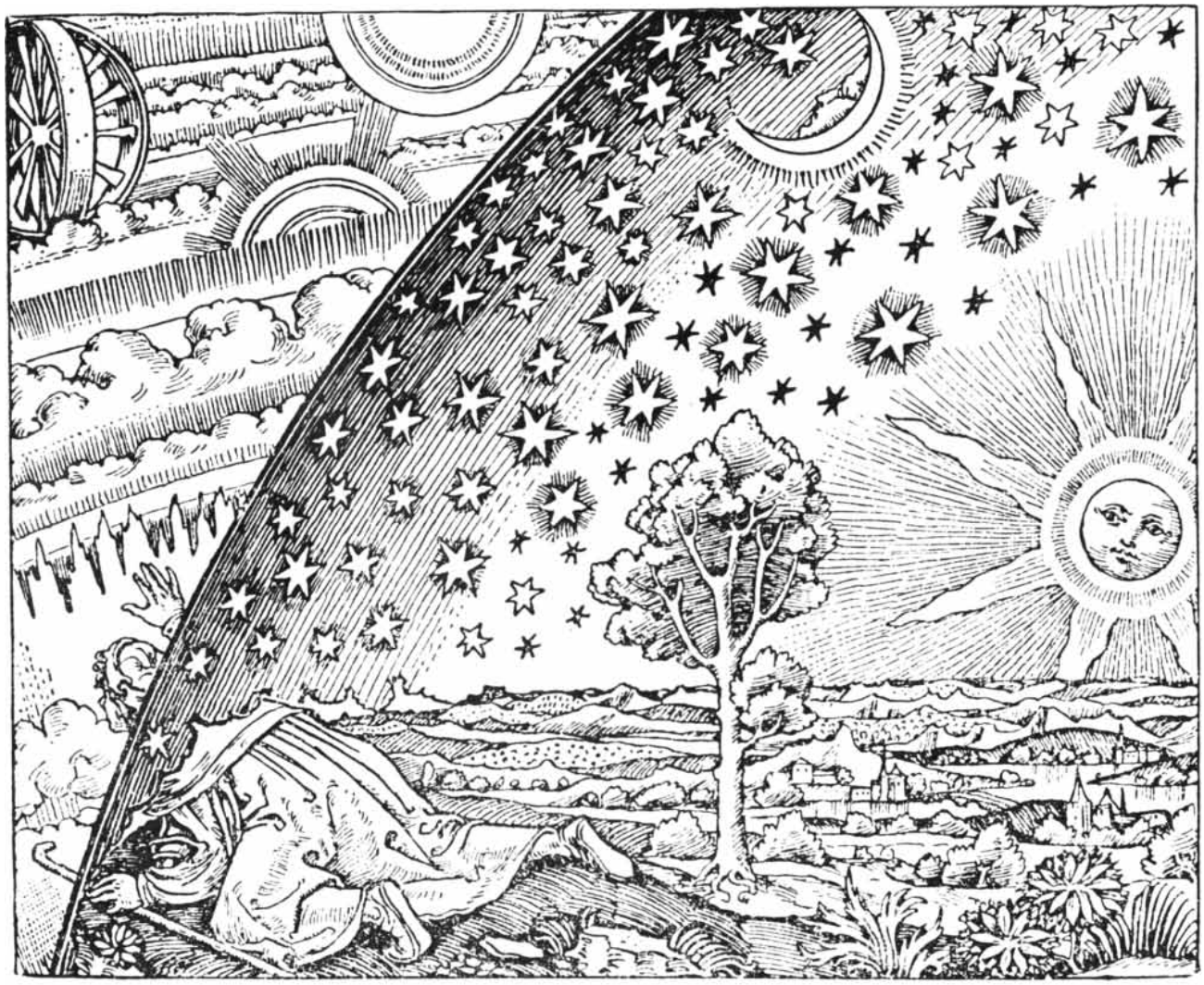
Like any other physical theory, the general theory of relativity deals with matter and its properties. It regards a galaxy as being perhaps the most natural unit of matter on the cosmic scale. Thus at this level the problem of space is the problem of understanding how the galaxies fit together. A convenient way of visualizing Einstein's solution to the problem is to construct a laboratory model of the entire galactic system. One could construct the model out of balls and sticks, like a model of a large molecule, except that each ball would repre-

sent a galaxy and each stick the distance between two galaxies. Before examining Einstein's model, however, let me go back and translate the views of Newton, Leibniz and Kant into the language of models. That may demonstrate more clearly both what they said and how the general theory of relativity went beyond them.

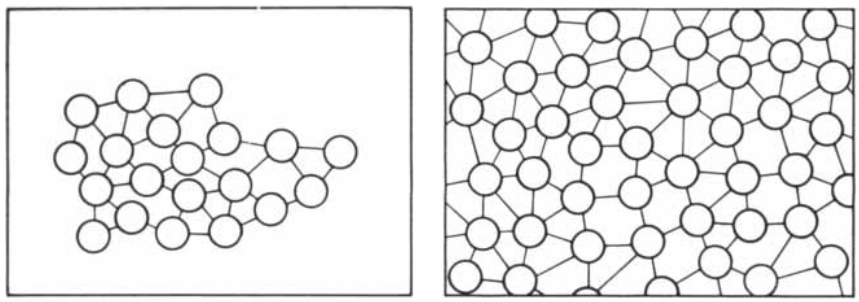
Let us turn first to Newton and his assertion that the universe is finite. If Newton is right, then there can be only a finite number of galaxies. Furthermore, it is reasonable to expect that in time one might devise instruments for locating all of them, and then a complete model of the galactic system could be built. In any case such a model is possible as a mental construct, and that is sufficient for our purposes. If Newton had thought in terms of such a model, he certainly would have had in mind an exact scale model of the universe: one in which the distances between balls were exactly proportional to the distances between the galaxies they represented.

The main consequence of exact scaling is that any metric feature of the model (that is, any feature that depends only on distance) will be shared by the galactic system. In other words, the model and the galactic system should have the same metric features. The laws of Euclidean geometry are known through direct observation to hold in the terrestrial laboratory, and they dictate all the metric properties of the model. Hence those same laws must dictate the metric properties of the galactic system. That conclusion is very important; in fact, it is the key to everything that follows. It states that the laws of Euclidean geometry are valid in the galactic system not because they are directly verifiable through observation and measurement in the galactic system but because the system can be reproduced in a scale model. The converse will also be true: if it is impossible to reproduce the galactic system in an exact scale model, then we must abandon the conviction that the geometry of intergalactic space is Euclidean.

Now, a geometric figure and any scale model of it are similar, meaning that corresponding angles in the figure and in the model are identical and corresponding sides are directly proportional to each other in length. Thus we are essentially saying that space can contain simi-



SPACE WAS FINITE and had a definite edge, according to the Aristotelian cosmology accepted during medieval times. Here a man is shown looking beyond the edge of space to the Empyrean abode of God beyond. The illustration is often said to be a 16th-century German woodcut; according to Owen Gingerich of Harvard University, it is more likely a piece of art nouveau that was apparently published for the first time in 1907 in *Weltall und Menschheit,* edited by Hans Kraemer. In either case the picture clearly demonstrates a dilemma posed by Immanuel Kant known as Kant's antinomy of space. Kant believed that the universe had to be finite in extent and homogeneous in composition, and that space had to obey the laws of Euclidean geometry. Actually, however, all those assumptions cannot be true at once. Newton, Leibniz and Einstein had different ways of resolving the dilemma, shown in illustrations on next two pages.



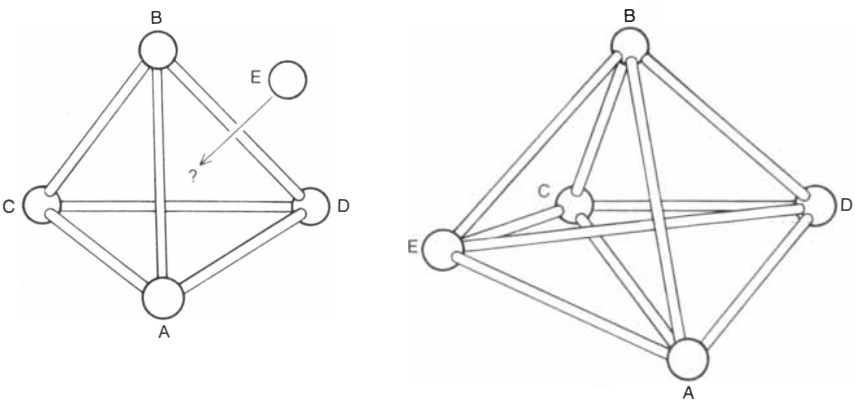
TWO "BALL AND STICK" COSMOLOGICAL MODELS demonstrate the philosophical views of Newton and Leibniz; each ball represents a galaxy and each stick represents the distance between galaxies. Although both men accepted Kant's assumption that space obeyed Euclidean geometry, Newton believed the galactic system was finite and inhomogeneous (*left*); thus his model of the galactic system had both a center and a boundary. Leibniz, however, believed the galactic system was infinite and homogeneous (*right*), with no center or boundary.

lar figures of arbitrary sizes if and only if its structure is given by Euclidean geometry. That result is actually very old: it was first obtained by the English mathematician John Wallis in 1663. In either case, using the language of models or of similar figures, we end up with a criterion for determining the geometric properties of space.

Any model of a finite universe has two metric features of special interest. First, the model has a "geographic" center. Second, it has a boundary, consisting of those balls with neighbors on only one side. Therefore if Newton is right, the galactic system must also have a center and a boundary, because it possesses all the metric properties of its scale model.

It is the inhomogeneity of Newton's universe (the fact that not all galaxies have neighbors on all sides) and not its finiteness as such that Leibniz could not accept. Any finite model has a boundary and a center; in order to eliminate those features one would have to add an infinite number of new balls. Then the model becomes impossible to build. Nevertheless, it is possible to imagine an arrangement of balls reproducing exactly the arrangement of galaxies and fading into the distance in all directions. Let us take that mental construct as Leibniz' scale model. It resembles a model one can build of a crystal, which is also infinite in a theoretical sense. Hence Leibniz' model is actually no stranger than Newton's. If the galaxies are more or less uniformly distributed, then the model will have neither a center nor a boundary.

Like Leibniz, Newton and everyone else in the 18th century, Kant believed in the validity of Euclidean geometry. Unlike many, he knew that Euclidean geometry could not be justified by experience alone. In fact, our key argument, that we judge space to be Euclidean not through observation but by constructing a model of the galactic system, is due to Kant. In the *Critique of Pure Reason* he actually did not address either models or galaxies. He declared that a direct intuition of space—what I am calling a model—is given to each of us, and through it we discover the properties of space. Because the intuition is universally shared among human beings, it is one of the dictates of pure reason and must be put on an equal footing with sensory experience in investigations of the world. What is more, since the intuition does not depend on experience, which can be faulty or incomplete, the knowledge intuition gives must necessarily be true. As Kant put it, Euclidean geometry is synthetic a priori, by which he meant that it is a special kind of knowledge that is truly descriptive of the world of experience but is not itself derived from that experience. Just as today every science strives to become exact, in the 18th century classical physics saw geometry as its ideal. Kant made his beliefs explicit because he wanted to exploit the special status of geometry to refute the claim of empiricists, most successfully advanced by David Hume, that *all* knowledge of the world is sensory. But is such a direct intuition absolutely necessary to understanding the universe? Must the world admit a scale model? Something fundamental in Kant's philosophy would be undermined if a different situation were to be perceived.

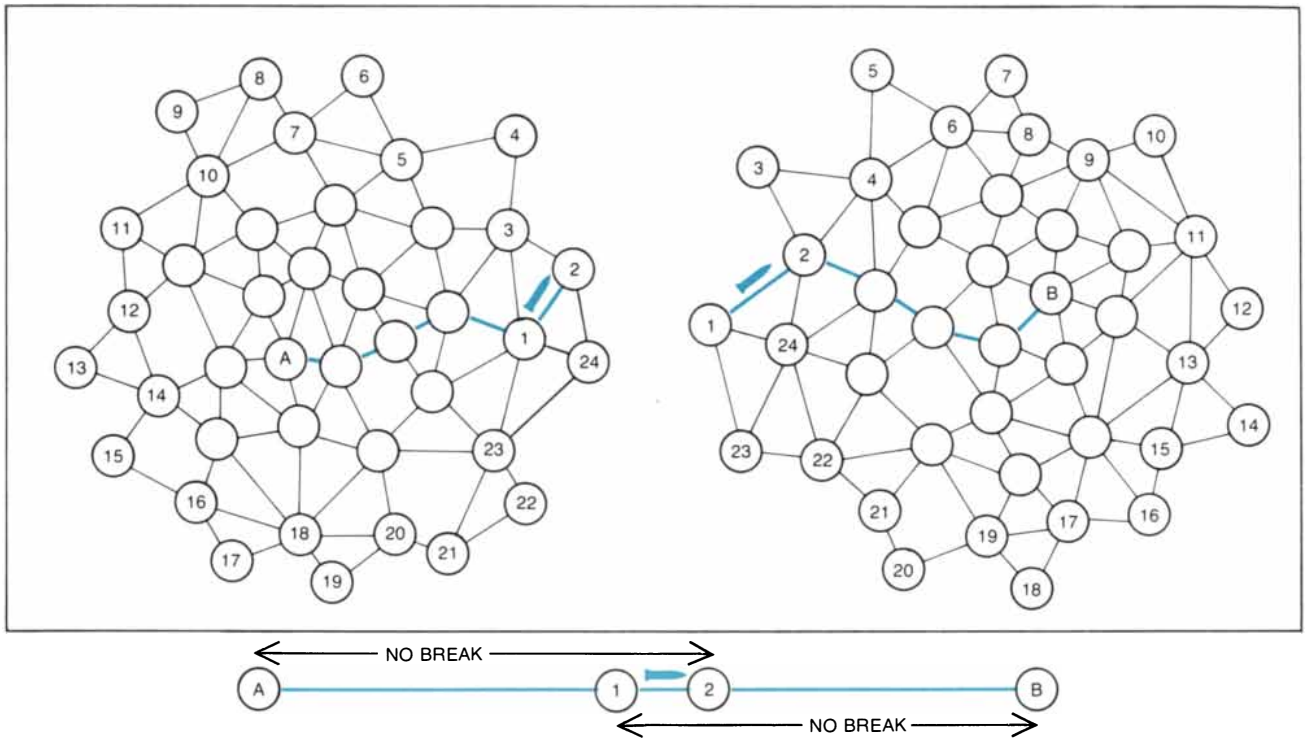


EINSTEIN'S COSMOLOGICAL MODEL achieves Kant's desideratum of a finite and homogeneous universe, but only by rejecting Kant's assumption that space was Euclidean. To say that space at large is curved means that it may not be possible to build an exact scale model of the galactic system in the laboratory, where the laws of Euclidean geometry rule. For example, five equidistant galaxies may exist in space, but any attempt to construct a scale model of such a system will fail (*left*). Einstein resolves Kant's antinomy of space by suggesting that when one builds the ball-and-stick model of the system, one should just say that one long stick joining two of the galaxies represents same distance in space as the nine shorter sticks (*right*).

Newton's and Leibniz' models are two clearly distinct models of the galactic system. Kant rejected them both. He had to, because each lacked what he felt was an essential property. For his part he maintained that any study of the material universe must begin by acknowledging three facts: first, the galactic system is finite; second, it is homogeneous and unbounded; third, it can be reproduced in an exact scale model. Those three facts, however, cannot be simultaneously true. In other words, no exact scale model can be both finite and homogeneous. Once again we have arrived at Kant's antinomy of space, this time through the language of models.

Why does the antinomy arise? Kant blamed it on inherent limitations in the mental processes we use to describe the world. In this case the mental process is model building. Kant is saying that we are unable to build a model of the galactic system because our minds cannot tell us how.

There is a way to get out of the antinomy: simply refuse to accept one of Kant's "facts." Remember, not one of them is a physical fact established by direct observation. They are all intuitive assumptions. Even Kant admitted that.



EINSTEIN'S MODEL for a finite galactic system is shown here in somewhat more detail. The model is entirely contained within the limits of the illustration. The model appears to be highly inhomogeneous: it is shown in two disconnected pieces each with a center and a boundary, and not all galaxies appear to have neighbors on all sides. Furthermore, two numbered balls, one in each piece, represent the same galaxy. Unnumbered balls represent distinct galaxies that are not duplicated. The model is not built to scale, however; therefore its peculiar properties are not necessarily reflected in the galactic system itself. In fact, the galactic system is actually quite homogeneous. The color line between ball *A* and ball *B* represents a continuous path through the galactic system being taken by a rocket (also shown twice). The continuous nature of the rocket's path is illustrated at bottom. One can think of the model as a pair of three-dimensional viewing screens in which every galaxy appears on at least one screen, and any galaxy at the edge of one screen also appears on the other.

One possibility is to follow Newton and reject Kant's second assumption that the galactic system is homogeneous and unbounded. Then there is no problem in accepting the remaining two assumptions and building a suitable model. That may in fact be the correct choice, since a finite but inhomogeneous galactic system has never been ruled out experimentally. In any case the antinomy is gone. Another possibility is to follow Leibniz and reject the first assumption, that the galactic system is finite. This choice also removes the antinomy, since a homogeneous but infinite system has not been ruled out experimentally either. Or, finally, we can follow Einstein.

Einstein actually achieved Kant's ambition by constructing a down-to-earth model of a finite homogeneous galactic system. He did it by rejecting Kant's third assumption, that the model must be exactly scaled. For example, in Einstein's model a three-inch stick in one location may represent an intergalactic distance of, say, 50 million light-years, whereas in another location it may represent 60 million light-years. To see the impact of Einstein's innovation consider the following imaginary scene.

One day at some time in the distant future the members of an intergalactic surveying team return to their home base, having measured all the distances between five galaxies in which they are particularly interested. They exchange information and discover that each of the five galaxies is equidistant from the other four. Immediately they can stick four balls together at the vertexes of a regular tetrahedron to represent four of the equidistant galaxies. Where should the fifth ball go? They can attach it to any three of the other four balls and form a second tetrahedron. It is then equidistant from the three balls but not from the fourth ball. Actually there is no way to build a laboratory model consisting of five equidistant balls. Forget about a grand design for the entire galactic system; here is a mere handful of galaxies presenting a crisis in intuition. As a spokesman for the Euclidean position, Kant would have argued that the surveyors are mistaken, because space "is not like that." But the surveyors have already rechecked their work and there is no mistake. Space *is* like that. The assumption that we can build a scale model of any physical system, which is equivalent to the assumption that the geometry of space is Euclidean, is thus revealed to be an attempt to make reality conform to our preconceptions. Einstein turns that around and makes the model conform to reality: he takes the model already built and declares that one long stick should represent the same intergalactic distance as the other nine joining all the balls.

It is only a short step from five galaxies to Einstein's model of the entire galactic system. The illustration above presents a simplified version of Einstein's model using several dozen balls instead of the millions that might actually be required. The number of balls does not really matter, since a larger model will exhibit the same essential features. Two features of the model are particularly noticeable. First, in this illustration there are two disconnected pieces, each with a boundary; second, sometimes two different balls, one in each piece, represent the same galaxy. Because the model is not built exactly to scale, however, those peculiar properties need not carry over to the galactic system itself. In fact, the system is actually quite homogeneous, as one can see by thinking of the model as a pair of three-dimensional viewing screens. In that case every galaxy appears in at least one of the screens, and to guarantee that none is overlooked, any galaxy appearing at the

edge of one screen also appears at the edge of the other screen. That explains why two balls sometimes represent the same galaxy.

Now look at one of the galaxies that is visible in both screens. Half of its neighbors appear in one screen and half in the other. Thus although each ball representing that one galaxy is on the boundary of its own screen, the galaxy itself must be completely surrounded by neighboring galaxies. Since the remaining galaxies appear in the middle of either one screen or the other, they also have neighbors on all sides. Hence the galactic system has no boundary. Furthermore, the galactic system is connected, because an object can move continuously from any one galaxy to any other, although it may be necessary to switch viewing screens at some point in order to follow the motion.

Distance distortion is a third essential feature of the model, but it is not particularly noticeable in such an incomplete illustration. Exactly how much distances are distorted is a technical matter, fully treated in Einstein's detailed model (the general theory of relativity). The simple presence of the distortion is what concerns us, because it reveals a fundamental property of space, that is, of the metric relations of the galactic system. Whether or not a model must distort distances is determined solely by the nature of space. As we have already seen, if the galactic system admits an exact scale model, space is Euclidean. If no scale model is possible, that fact must likewise be due to some contrary property of space. This property has been given the name curvature.

The word curvature may seem an odd choice for a property of space, and so I shall discuss the reason for it below. In any event curvature of space refers simply to the need for distorted models and has nothing to do with space being somehow mysteriously bent. Taking Einstein's model into consideration, one can now draw one of the fundamental conclusions of the general theory of relativity: Kant's antinomy of space arises from an unwarranted assumption that space is Euclidean. A finite and homogeneous galactic system is conceivable, and it is the curvature of space that makes it so.

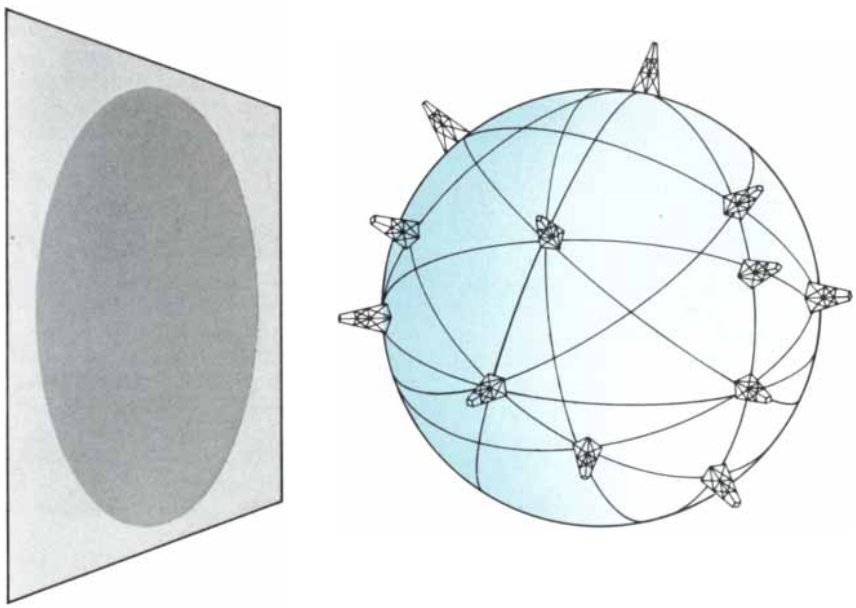
The connection between curvature and a finite world is a common subject in popular accounts of relativity, but it is usually explained by drawing an analogy with what can happen on surfaces. On a flat plane any finite set of points has a boundary. On the spherical surface of the earth, however, one can imagine a more or less uniformly distributed network of, say, weather stations, and although there are a finite number of stations, every station has neighbors on all sides, so that the network has no boundary. The analogy suggests that the space of a finite homogeneous galactic system must therefore be like the surface of the earth, only one dimension larger. Now, the earth's surface is only two-dimensional, which means that the position of a point on it is determined by just two numbers, latitude and longitude. The reason the earth's surface has no boundary, we feel, is that it curves around into a third dimension. We are led to infer that a three-dimensional "spherical" space must somehow curve around into a fourth dimension. The analogy collapses because it is hopeless to imagine what the extra spatial dimension looks like; no one has ever seen it.

The analogy is actually a good one, but oddly enough it suffers from too detailed a picture of the earth and not too scant a picture of space. A surface has two kinds of geometric properties, intrinsic and extrinsic. An intrinsic property is one that refers to measurements that can be carried out entirely on the surface itself; any other property is extrinsic. For example, in the third century B.C. Eratosthenes deduced the radius of the earth by combining an intrinsic fact about the earth's surface (the fact that the distance from Syene to Alexandria is 5,000 stadia) with an extrinsic fact (the fact that when the sun is directly overhead at Syene, it is 7.2 degrees from the zenith at Alexandria).

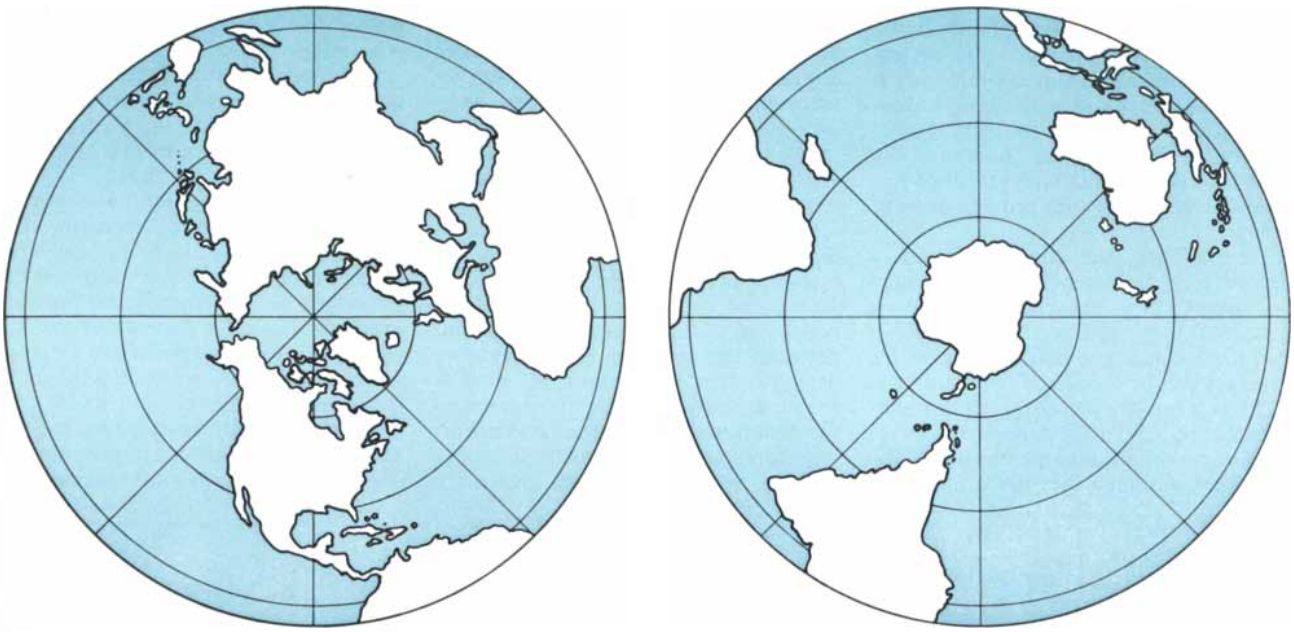
Physical space has no extrinsic geometry that we know of, however, because every spatial property we know relates to figures and measurements made in space itself. Thus space cannot be analogous to the surface of a sphere, because it has nothing comparable to the sphere's extrinsic geometry. An analogy between different objects can exist in our minds only where we find it possible to overlook their differences.

Still, there is a way to conceive of the curvature of space. Our actual goal is to understand what a finite homogeneous galactic system must be like. Here it is useful to consider the example of a network of weather stations on the surface of the earth. Notice that the relevant geometric properties of the network are all intrinsic: each weather station is surrounded by neighbors, and the network of stations, although finite, has no boundary. Since those are intrinsic properties, we lose nothing by ignoring the earth's extrinsic geometry; moreover, since the extrinsic part was the stumbling block all along, we are even better off ignoring it. Irrelevant information can be just as distracting to a mathematician as it is to a reader of mystery stories. In mathematics the job of isolating what is genuinely relevant is called abstraction.

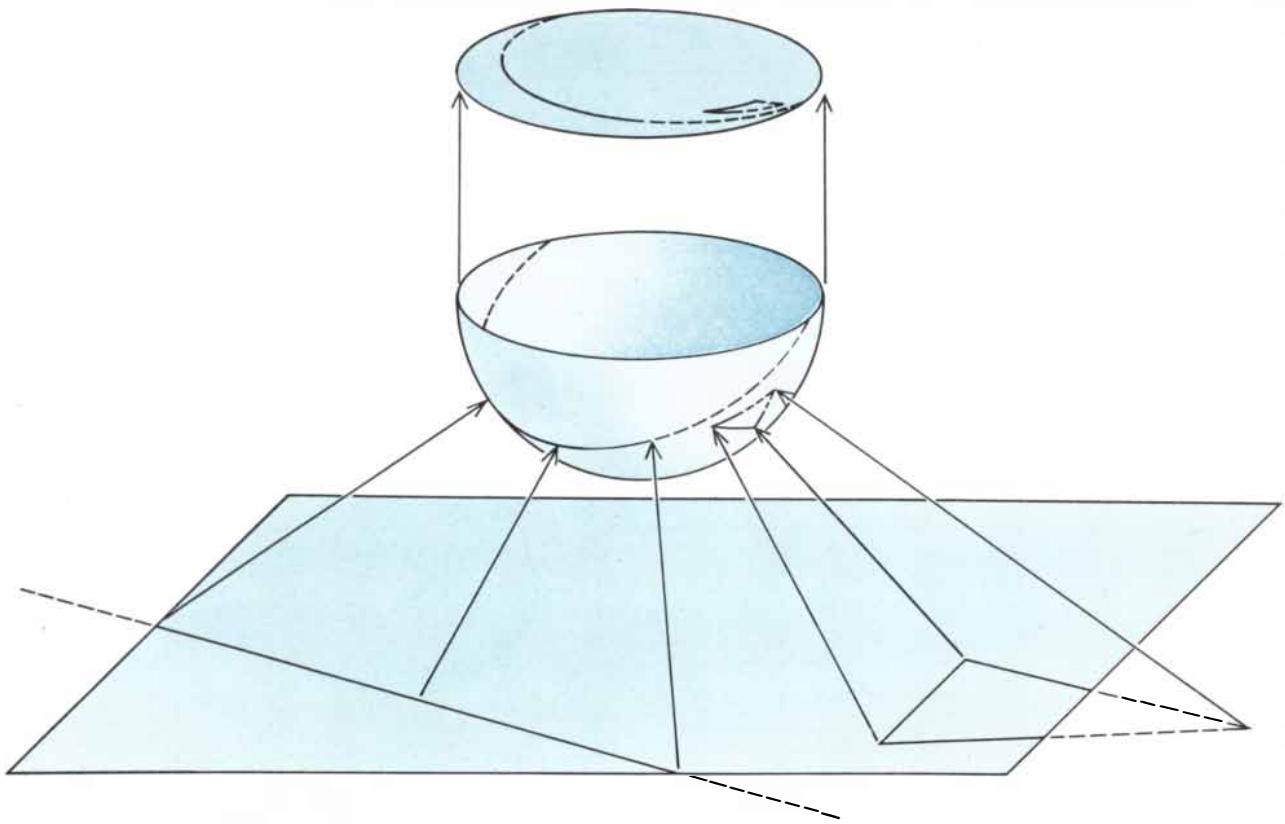
The language of models can help once again. A terrestrial globe is an exact



INTRINSIC GEOMETRY AND EXTRINSIC GEOMETRY of the surface of an object such as a sphere are quite different. Any intrinsic property of a surface refers to measurements that can be carried out on the surface itself; any other property is extrinsic. For example, the fact that on a sphere a network of weather stations can be both finite and unbounded is an intrinsic property. The fact that the sphere casts a circular shadow from all viewpoints is an extrinsic geometric feature. Physical space has no known extrinsic geometry because every property we know of relates to figures and measurements made in space itself. Thus it is hopeless to try to imagine curved space as being mysteriously bent through a fourth dimension, since we cannot take an extrinsic view of space by getting outside it and looking back at it.



ATLAS OF CHARTS of the spherical earth captures all the intrinsic properties of the earth while filtering out the extrinsic geometry. Since the curved surface of the earth has been flattened out on the charts, distances are distorted. The amount and kind of distortion contain all the information needed to reconstruct the full geometry of the earth. Hence we no longer need a third dimension in order to understand the curvature of the surface of a sphere; similarly, we do not need a fourth dimension to understand the curvature of space.

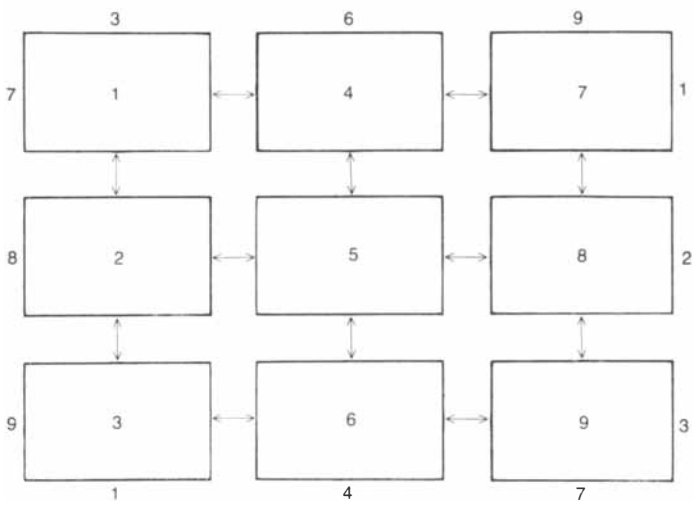


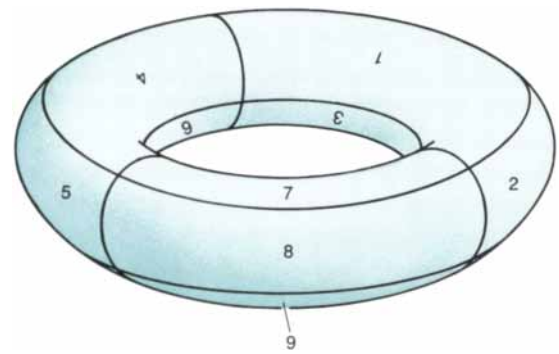
INFINITE PLANE CAN BE MAPPED onto an atlas of charts as well as a sphere can. In this particular case the atlas consists of one chart, the disk, which is made by first projecting the plane onto the hemisphere and then up onto the disk. (This is only one of many ways of mapping the plane.) The geometric shapes again reveal the distance distortions created by the mapping. Any unbounded two-dimensional surface can be mapped by an atlas of charts; such a surface, including the plane and sphere, is a two-dimensional manifold.

scale model of the earth's surface. It is too good, however, because it reproduces all the earth's geometric features, extrinsic as well as intrinsic, with perfect clarity. Ironically, using it denies us the understanding we seek. We need a model that captures all the intrinsic geometry but at the same time filters out the extrinsic. Let us turn now to the other familiar model of the earth: an atlas of maps.

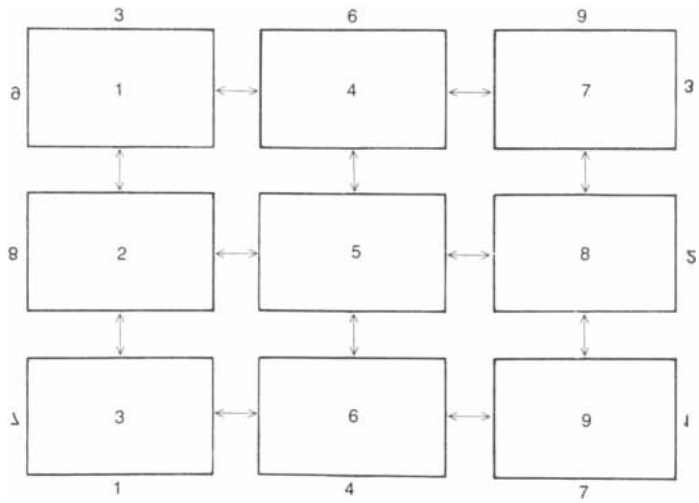
In many ways any atlas is inferior to a globe. It represents the earth as a collage of overlapping flat charts, and each chart distorts distances. Nevertheless, all the intrinsic geometry of the earth's surface can be recovered from an atlas. This is actually a surprising result mathematically, and it is difficult to prove. The fact remains that the inevitable distortions, although they are a nuisance, are manageable. How else could worldwide air and sea navigation be based on charts? One might object that an atlas does not look like the earth, that it does not help one to grasp the earth's extrinsic geometry. For our purposes, however, the flatness of the charts, far from being a disadvantage, is their greatest virtue. It implies that one can understand all the intrinsic geometry of a sphere without ever leaving a flat two-dimensional plane. That fact is a tremendous economy, and it shows what abstraction can achieve: one does not need a third dimension in order to understand the structure of the weather network; hence one does not need a fourth dimension in order to understand the structure of the galactic system. Furthermore, what seems to be a creaky analogy between the earth and space is in fact a real analogy between their abstract intrinsic structures: Einstein's model (the general theory of relativity) is an atlas of space.

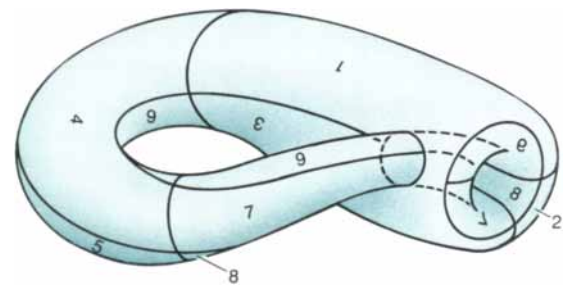
How does curvature fit in? The way it is used in the general theory of relativity can be traced back to the work of Carl Friedrich Gauss, in a theory of curved surfaces he published in 1827. Gauss was the first to recognize that a surface has a separate intrinsic geometry. His most remarkable discovery, however (and he even called it that), was that the curvature of a surface is an intrinsic property. Basically Gauss said the following: Take a small portion of a curved surface and flatten it out on a plane. In other words, make a chart. This can generally be done only by stretching the surface, that is, by distorting distances. It should be evident that the original cur-





AN ATLAS FOR A TORUS (*left*) might consist of nine charts; the numbers around the edges of each chart indicate which of the other charts it overlaps. The numbered regions on the torus corresponding to the nine different charts in the atlas are shown at the right.





AN ATLAS FOR A KLEIN BOTTLE (*left*) might also consist of nine charts. It is very similar to the atlas for the torus except in the way in which the charts overlap. That change, however, makes it impossible to paste the charts together without making the surface intersect itself in places where it should not. The Klein bottle is a form that has one surface. It has no "inside," as a torus and a sphere do.

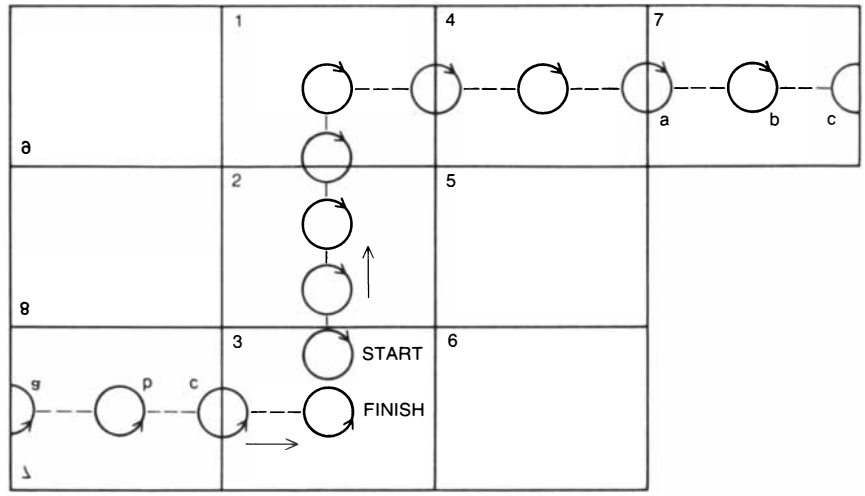
vature of the surface determines the precise kind and amount of distortion. Gauss's remarkable discovery is the converse of that fact: the curvature of a surface can be completely determined solely from the distance distortions present in a chart. In other words, distortion of a chart and curvature of a surface are different aspects of the same thing: when the intrinsic geometry of a portion of a surface is abstracted to a chart, its curvature is given by distance distortion. That is why the completely analogous distortions in Einstein's model of the galactic system are attributed to the curvature of space.

Gauss studied the intrinsic geometry of a small portion of a surface by making a chart. In order to map an entire surface it is generally necessary to resort to several charts, or an atlas. A surface that can be mapped by an atlas of charts is called a two-dimensional manifold. The term is meant to emphasize that the surface is usually made up of many two-dimensional patches instead of a single chart.

Every smooth surface without a boundary is a two-dimensional manifold. For example, an infinite plane is a two-dimensional manifold, and so is a torus. One can also start with any arbitrary collection of charts that overlap in a coherent way; they constitute an atlas for some two-dimensional manifold.

The notion of a manifold, which grew out of the attempt to understand the intrinsic geometry of a surface, actually yields something more general. One can make charts out of solid, three-dimensional "blobs" instead of flat, two-dimensional "patches" and have something that makes sense in three dimensions. The resulting object is called, naturally enough, a three-dimensional manifold, and it is a volume rather than a surface. Einstein's model of the universe is an atlas for one particular three-dimensional manifold: it is called the three-dimensional sphere, or three-sphere. (The surface of an ordinary sphere is a two-sphere.) There are countless other examples. The ordinary Euclidean space of mathematics is given by a single chart, analogous to the chart for the infinite plane. Other spaces quickly become too messy to describe in detail. Mathematicians have even defined manifolds for arbitrary numbers of dimensions, by finding a way of getting around the need to visualize the fundamental building blocks: the dimension of a manifold is simply interpreted as the number of variables needed to locate a point on it. Manifolds have thus become a natural setting for problems requiring many variables, and their use now extends beyond mathematics to science in general.

One of the most basic perceptions we have of our environment is that three



ORIENTABILITY is a global property of a surface. A circle with an arrowhead on its circumference is said to define a local orientation (either clockwise or counterclockwise) when it is drawn on a chart for a manifold. That orientation can be extended to other charts by moving the circle from one chart to another in the atlas. Moving along two different routes, however, may produce conflicting orientations of the circular arrowhead in a distant chart. If that happens, the manifold is said to be globally nonorientable. A Klein bottle is globally nonorientable; shown is its atlas with one chart (*chart 7*) and three of the positions of the circular arrowhead (*a–c*) included twice. When the circular arrowhead returns to its starting position in chart 3, it is an upside-down mirror image of its former self. In some manifolds, for example the torus, such a conflict can never arise; therefore such a manifold is said to be globally orientable.

variables (height, width and breadth, or *x, y* and *z*) are needed to label completely all positions within it. Physical space is a three-dimensional manifold. The mistake of common sense was to assume that it must be the one given by a single chart: Euclidean space. We can now see why this was such a natural error to fall into. Locally (that is, in the immediate neighborhood of any point) all manifolds of the same dimension look alike. In principle they are distinguishable because of the presence of curvature; in practice, however, curvature may not be detectable with sufficient experimental accuracy when it is measured over a small region. (The surveyors in my little scene had to travel to other galaxies to get their results; our own measuring instruments have not yet left the solar system.) In other words, for all practical purposes space *is* Euclidean if it is taken in small enough pieces. The entire cosmos is another matter, and the notion of the manifold provides a wealth of new possibilities for understanding the structure of the universe in the large. Far from being bankrupted by the apparent nonsense of questions about the structure of space, as Kant suggested, mathematics has been substantially enriched.

The most striking differences between one manifold and another are global differences, differences that can be discerned only by studying entire atlases, not single charts. One global property of a surface is orientability. For example, the atlases for a torus and for the surface known as a Klein bottle are quite similar, but the Klein bottle has two peculiar features that set it quite apart from the torus. First, the Klein bottle cannot be constructed in space without intersecting itself in places where it should not intersect itself. Second, it cannot be oriented in space. Self-intersection cannot be inferred directly from the atlas. It is an extrinsic feature. Nonorientability is an intrinsic feature, however, and it can be discovered by following a moving clock face from one chart in the atlas to the next. Nonorientability is a global property. A second global property is connectivity, arising from the question of whether every closed loop sliced on a surface separates the surface into two pieces. A sphere and a torus differ in their connectivity. There are analogous notions of connectivity for higher-dimensional manifolds. The study of global properties is a part of topology. It thus becomes evident that in a manifold—including Einstein's three-sphere model of space—a number of small localities having a quite ordinary and familiar geometry can be combined to produce a global effect that is both novel and surprising. This aspect of manifolds is also a part of the fascination of the drawings of Maurits C. Escher.

In the search for the origins of Einstein's ideas much attention has been given to the non-Euclidean geometry that was developed in the early 1800's. That emphasis is somewhat misleading. Several mathematicians, including Johann Heinrich Lambert, who lived between 1728 and 1777 and who was a friend of Kant's, had come to the con-

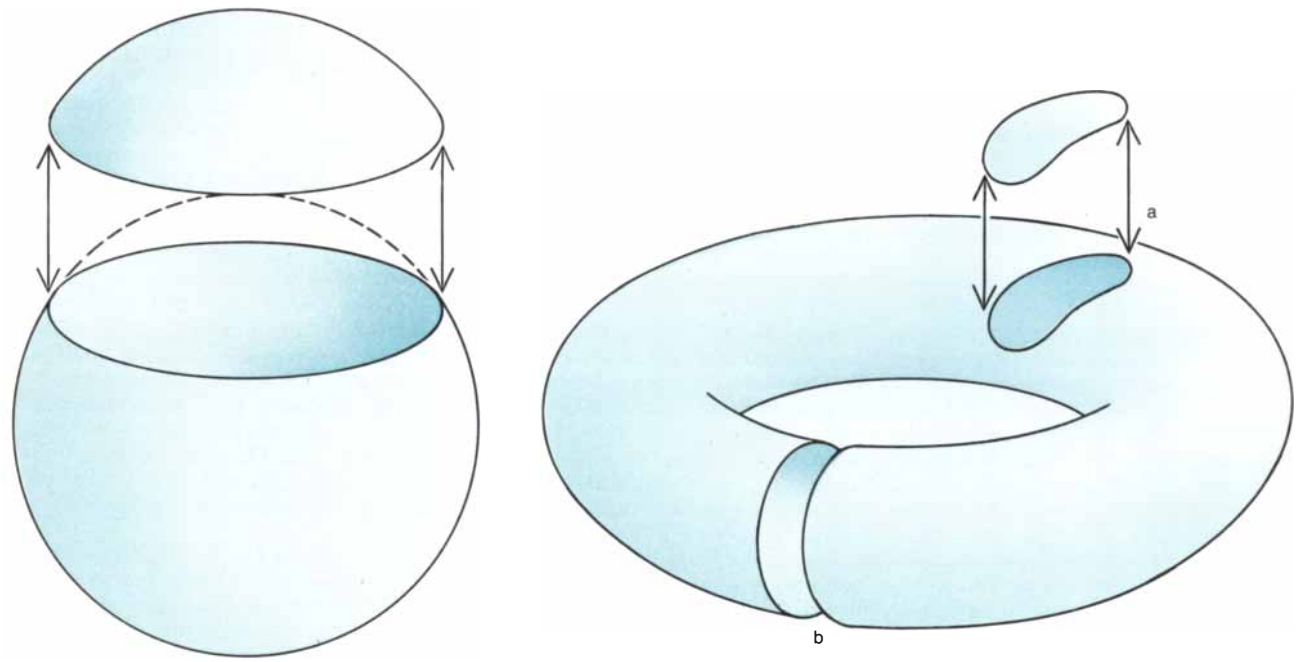
clusion that a different collection of theorems, logically as sound as those of Euclidean geometry, could be derived from a new system of axioms that differed slightly from Euclid's. A smaller number of 19th-century mathematicians, notably Gauss, János Bolyai and Nikolai Lobachevski, saw further that such a new system could plausibly describe the structure of physical space just as well as Euclid's. It was no longer possible to maintain Kant's position that Euclidean geometry was synthetic a priori. The antinomy of space persisted, however, because the new Lobachevskian space (as it is often called) had the same overall topological structure as Euclidean space. In particular any finite collection of galaxies in it would still have to have a boundary.

The 19th century was a time of tremendous developments in geometry. By the 1870's comprehensive systems embracing all Euclidean, non-Euclidean and projective geometry had evolved. Curiously, the question of whether the geometric systems were physically relevant drifted from the center of attention and even became rather confused. Henri Poincaré, one of the most eminent mathematicians of the time, even maintained that there was no strictly correct geometry for the description of space. The choice of one geometry instead of another was purely a matter of convention, he said, or was no more consequential than the choice between Fahrenheit and centigrade thermometers to measure temperature.

The mathematics of Einstein's general theory of relativity is quite distinct in style from that of the prevailing systematic geometries at the beginning of the 20th century. It is a part of differential geometry, which, as the name suggests, exploits the power of calculus. With it Einstein showed how to interpret gravitation as being a curvature of space. In other words, distance distortions will be present in a model of even a small portion of space if the space contains an appreciable amount of matter. That distortion, and not gravitational "force," then dictates the paths of moving bodies. Curvature is therefore a local phenomenon as well as a global one. In fact, the general theory of relativity is mainly occupied with the local consequences of curvature. It is difficult to imagine how the geometry of the general theory of relativity, inextricably bound up with matter as it is, could have evolved from the tidy axiomatic systems of Euclid and Lobachevski. Borrowing a locution from the world of the New York theater, where plays can be presented off-off-Broadway, we could say that the geometry Einstein used is non-non-Euclidean. Einstein's ideas were cast in a language very different from even non-Euclidean geometry, called the absolute differential calculus. Until Einstein used it and changed its name to tensor analysis, it had the reputation of being the kind of pure mathematics that had no connection with the real world.

Ironically tensor analysis did not start that way. Its origins can be found in the work of Bernhard Riemann (1826–1866). On the occasion of Riemann's appointment to the faculty of the University of Göttingen in 1854 at the age of 27 he opened an entirely new line of thinking in his probationary lecture, titled *On the Hypotheses That Lie at the Foundations of Geometry*. The topic had been selected for him by his teacher and colleague, Gauss. The entire modern viewpoint can be found in Riemann's lecture: the concept of an n-dimensional manifold; the study of a manifold's intrinsic geometry—particularly curvature—by extending Gauss's work on surfaces; even the radical notion that geometry and physics are inseparable, that is, that the presence of matter determines the curvature of space. Riemann's central concern, however, was with the implications his ideas had for the structure of physical space. His own words (translated) explain it well:

"That space is an unbounded three-dimensional manifold is an assumption that is employed for every apprehension of the external world, by which at every moment the domain of actual perceptions is filled in and the possible locations of a sought-for object are constructed; in these applications it is continually confirmed. For that reason the unboundedness of space has a greater



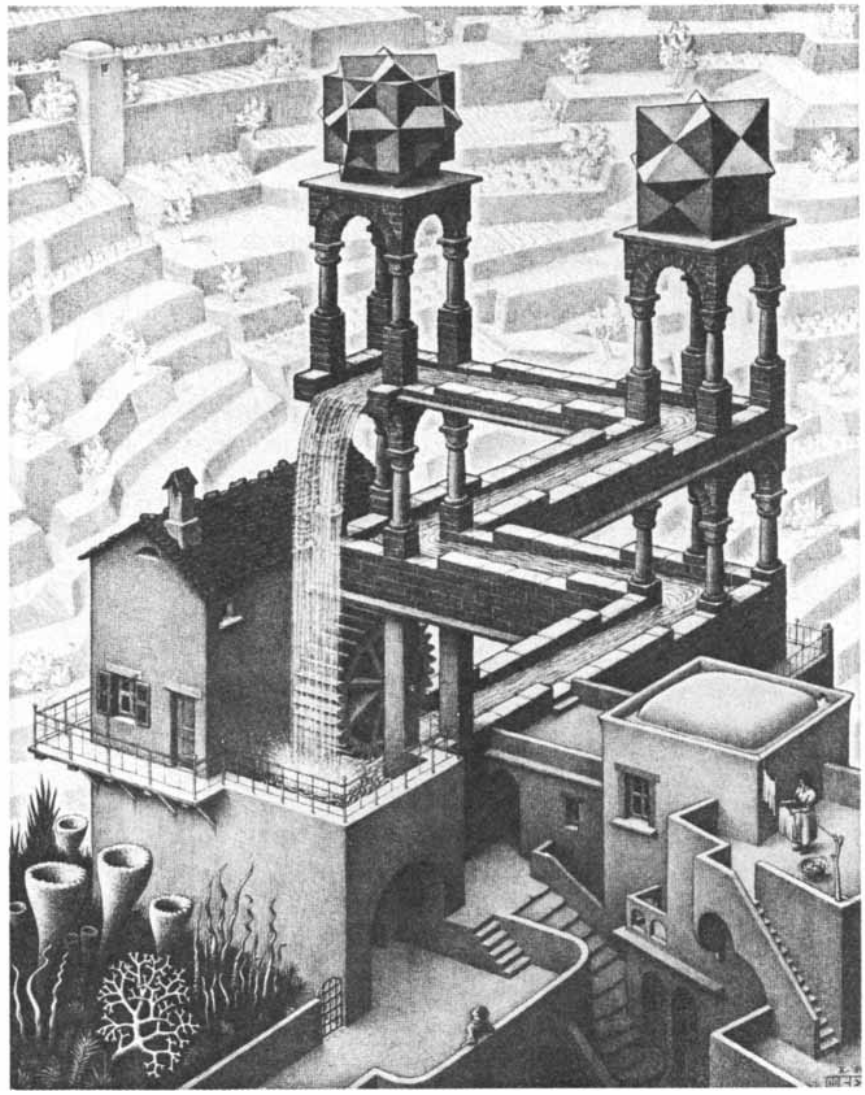
CONNECTIVITY is another global property of a surface, and it refers to the question of whether or not every closed loop on a surface separates the surface into two pieces. For example, the connectivity of a sphere and that of a torus are different. Any loop on a sphere separates it into two pieces (*left*). Such loops (*a*) also exist on the torus (*right*), but there are certain loops (*b*) that do not separate it. The fact that nonseparating loops can exist for some surfaces and not for others reflects a basic topological difference between those surfaces.

certainty than any external experience. But its infinitude in no way follows from this. On the contrary, space would necessarily be finite if we assumed that bodies exist independently of position—so that we could ascribe constant curvature to space—as long as this curvature had a positive value, however small."

Riemann treated space as a manifold, and he recognized that one of its possible structures is given by what we have been calling Einstein's three-sphere. Thus it was Riemann who first saw that a finite unbounded universe was conceptually possible. He, not Einstein, actually provided the way around Kant's antinomy of space.

Although Gauss, for one, was astonished by what he heard Riemann say, the wider scientific community did not become generally aware of Riemann's lecture until it was published posthumously in 1868. (Riemann died of tuberculosis at 39.) During the next half-century Riemann's mathematical ideas were developed extensively, but their applications to space were virtually forgotten. A generation later Einstein apparently rediscovered the full wealth of Riemann's ideas about the physical world. Einstein's work, however, is no mere copy of what Riemann had already done. The general theory of relativity is above all a physical theory, a coherent and detailed account of the underlying geometric character of gravitation. Riemann provided the geometric language, but Einstein's physics was radically new. Riemann had only hinted at it.

The positions of Einstein and Kant are by no means antithetical. On the contrary, as we have seen, one of the fundamental tenets of Kant's metaphysical idealism is that space is not a thing but one of the forms through which we organize our perceptions of things. Moreover, the structure of our perceptual organization is given a priori. Those thoughts of Kant's are implicitly accepted by relativity. The basic quarrel between Einstein and Kant is over the structure of space. Kant assumed, partly because he saw no more general possibility, that space must be Euclidean; Einstein maintained that it is Riemannian, and he includes Kant's position as a special case. That explains how the antinomy of space could arise in the first place, and also how Riemann and Einstein could resolve it without completely undermining Kant.



"WATERFALL," a black-and-white lithograph by Maurits C. Escher, is an example from the world of art of how space might be locally quite ordinary and yet globally quite surprising.

A final word about Dante, following some observations made by Andreas Speiser in his book *Klassische Stücke der Mathematik*. In the first two books of *The Divine Comedy* Dante traverses the material world from the icy core of the earth, the abode of Lucifer, to the Mount of Purgatory. In the last book, *Paradise,* Dante's beloved Beatrice guides him up through the nine heavenly spheres, each sphere larger and more rapidly turning than the last, until he reaches the Primum Mobile, the ninth and largest sphere and the boundary of space. His goal is to see the Empyrean, the abode of God. It finally appears to him, as a blinding point of light surrounded by nine concentric spheres that represent the angelic orders responsible for the motions of the material spheres. Dante is puzzled, however, because the smaller the radius of each Empyrean sphere is, the faster the sphere turns. Beatrice explains that there is no paradox between the material and the spiritual spheres; every sphere, whether material or spiritual, turns faster the more perfect or divine it is. The spiritual world completes the material world exactly as one viewing screen of Einstein's model of the galactic system completes the other. The overlap between the two is revealed by the correspondence between the celestial spheres and their angelic counterparts, and again as in Einstein's model the farther a sphere is from the center of one chart, the nearer its counterpart is to the center of the other. And the speeds of the material spheres and of the spiritual spheres are in harmony.

Speiser suggests that Dante was able to come to this remarkable vision because his geometric knowledge was derived from astronomy and not from Euclid, which he scarcely knew. Here is a translation by Barbara Reynolds (Penguin, 1962) of the conversation between Dante and Beatrice in Canto XXVIII:

About this Point a fiery circle
whirled,

With such rapidity it had outraced
The swiftest sphere revolving round the world.

This by another circle was embraced,
This by a third, which yet a fourth enclosed;
Round this a fifth, round that a sixth I traced.

Beyond, the seventh was so wide disclosed
That Iris, to enfold it, were too small,
Her rainbow a full circle being supposed.

So too the eighth and ninth; and each and all
More slowly turned as they were more removed
Numerically from the integral.

Purest in flame the inmost circle proved.
Being nearest the Pure Spark, or so I venture,
Most clearly with Its truth it is engrooved.

Observing wonder in my every feature,
My Lady told me what I set below:
"From this Point hang the heavens and all nature.

Behold the circle nearest it and know
It owes its rapid movement to the spur
Of burning love which keeps it whirling so."

"If manifested in these circles were
The cosmic order of the universe,
I should be well content," I answered her;

"But in the world below it's the reverse,
Each sphere with God's own love being more instilled
The further from its centre it appears;

Whence, if my longing is to be fulfilled,
Here in this wondrous and angelic fane,
Where love and light alone the confines build,

I must entreat thee further to explain
Why copy from its pattern goes awry,
For on my own I ponder it in vain."

"There's naught to marvel at, if to untie
This tangled knot thy fingers are unfit,
So tight 'tis grown for lack of will to try."

Then she went on: "This is no meagre bit
I'll give to thee. Wouldst thou be filled? Then take,
And round its content ply thy subtle wit.

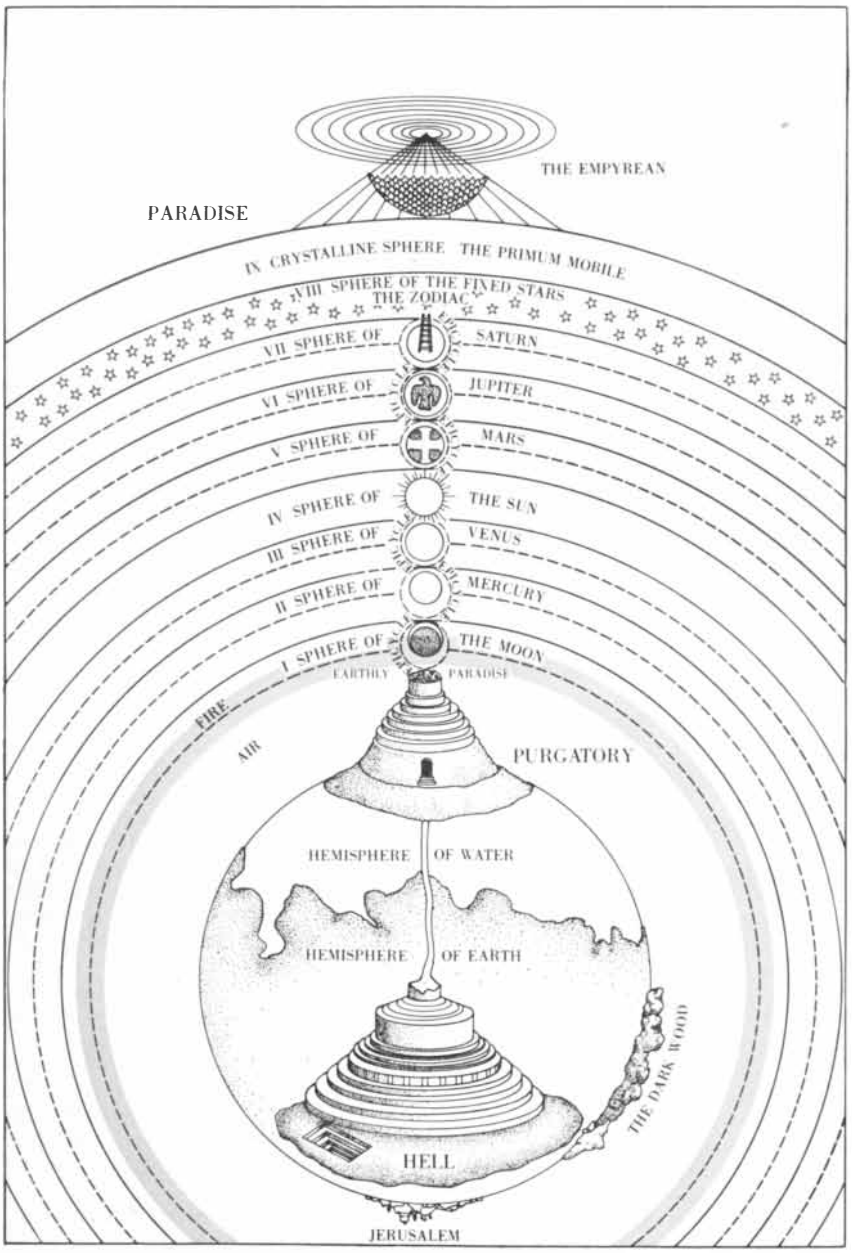
Material circles in the heavens make
Their courses, wide or small, as more or less,
Through all their parts, of virtue they partake.

The greater good makes greater blessedness;
More blessedness more matter must enclose,
If all its parts have equal perfectness.

It follows that the sphere, which as it goes
Turns all the world along, must correspond
To this, the inmost, which most loves and knows.

Hence, if thou wilt but cast thy measure round
The angels' *power,* not their circumference
As it appears to thee, it will be found

That wondrous is the perfect congruence
Which every heaven with every mover shows
Between their corresponding measurements."



DANTE ALIGHIERI'S SCHEME OF THE UNIVERSE in illustration from "Paradise" in *The Divine Comedy* extends Aristotelian cosmology in a modern way. It is discussed in text.

Freeman Books on Land and Landscape

Examine them free for 15 days.

GEOLOGY ILLUSTRATED

John S. Shelton, Drawings by Hal Shelton

Geologist/pilot Dr. John Shelton has been taking aerial photographs since 1939, and this is his master work. *Geology Illustrated* is a breathtaking view of our home planet: its rocks, minerals, volcanoes, glaciers, lakes, rivers, earthquake faults, mountain ranges, and other features. The nearly 400 photographs and crystal-clear text provide an enjoyable way to understand geology. Shelton addresses the book to "thoughtful people who enjoy the outdoors"—to hikers, rockhounds, pilots (and passengers), environmentalists, and many others. "An artist of aerial photography, Shelton uses some 400 of his finest photos to illuminate a discussion of the whole-earth system. Not a traditional textbook, but a fascinating exploration of problems posed by asking, 'How did that come about?' Worth buying for the photos and book design alone, but you'll probably find yourself becoming interested in geology regardless of your original intentions."
—*Whole Earth Catalog,* Summer 1971

1966, 434 pages, 382 illustrations, $14.50

PLANTS IN THE LANDSCAPE

by Philip L. Carpenter, Theodore D. Walker, and Frederick O. Lanphear, all of Purdue University

Plants in the Landscape is a unique introduction to the principles and practices of ornamental horticulture in landscape architecture. No other single volume covers such a broad range of topics. Here is a wealth of information on subjects ranging from the landscape practices of ancient civilizations to no-nonsense, dollars-and-cents descriptions of the present-day landscaping industry. The focus throughout is on landscape plants: how to select and maintain them for healthy growth in a wide range of settings.

Plants in the Landscape is a beautiful book. The text is closely integrated with more than 400 illustrations (about 300 photographs). Whether you are building a new home or remodeling an old one, whether you are a professional landscaper or a weekend gardener, you'll find *Plants in the Landscape* a valuable addition to your library.

1975, 481 pages, 534 illustrations, 32 tables, $16.00

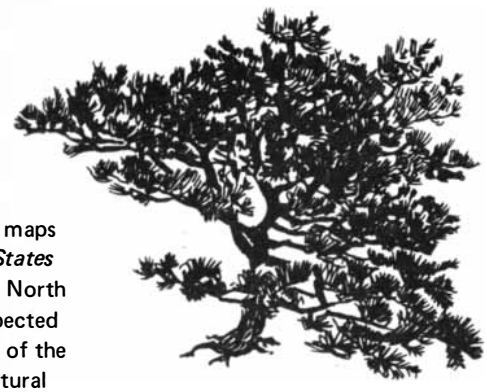
NATURAL REGIONS OF THE UNITED STATES AND CANADA

Charles B. Hunt, The Johns Hopkins University

With its generous use of photographs, line drawings, and maps (558 illustrations in all), *Natural Regions of the United States and Canada* provides a fascinating environmental tour of North America. Charles B. Hunt, a widely experienced and respected scholar, describes and explains the physical environment of the North American continent, Puerto Rico, and Hawaii. Natural regions are discussed in terms of their topography, geology, climate, soil, water, plants, and the interrelationship among all these factors.

"Just open this book to an area you already know. Within pages, the landforms, soils, colors, plants, and importance of your watershed will intensify. An encyclopedia of physiography written in academic style. Lots of maps. Lots of geologic history. The best book to practice mapping North America into your brain."—*Whole Earth Epilog,* 1974

1974, 725 pages, 558 illustrations, 22 tables, $15.95



TO ORDER, use the adjacent card. If card is missing, send order with payment to:

W. H. FREEMAN AND COMPANY
Dept. SA, 660 Market Street, San Francisco, CA 94104

MATHEMATICAL GAMES

The symmetrical arrangement of the stars on the American flag and related matters

by Martin Gardner

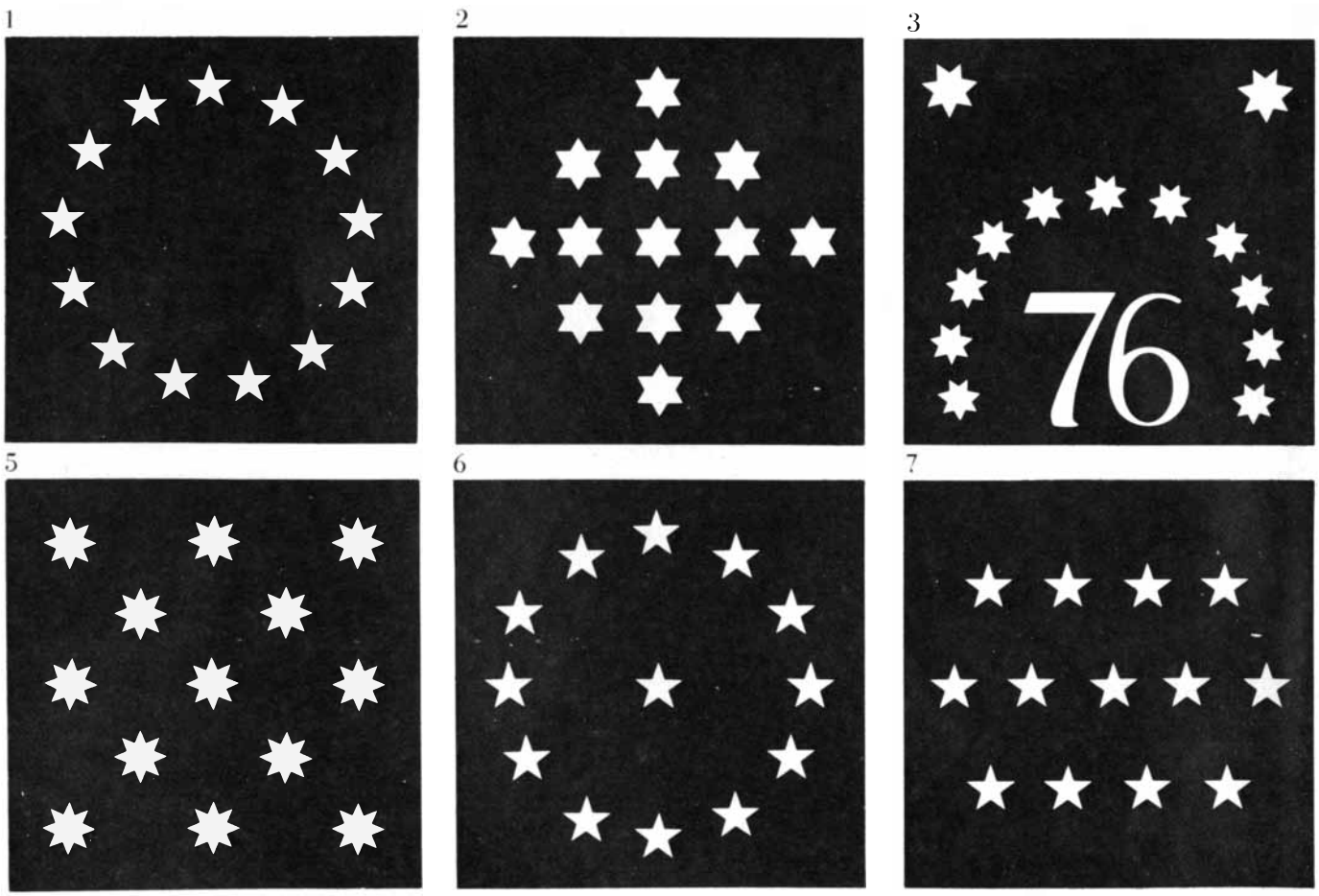
Whose broad stripes and bright stars. . . .
—"The Star-spangled Banner"

Look at the back of a dollar bill and you will see above the eagle on the Great Seal of the United States 13 five-pointed stars—symbols of the 13 original states—arranged as a six-pointed star. The pattern shows that 13 is the first nontrivial (larger than 1) figurate number of a type called "star numbers." (See this department for July, 1974.)

There are, of course, a multitude of other ways to arrange 13 points on the plane to meet the demands of aesthetic symmetry or the provisos of recreational mathematicians. In a moment we shall consider two unsolved problems, far from trivial, that concern ways of arranging 13 points. But first, in honor of the Bicentennial, let us take a look at how the Colonists arranged the 13 bright stars on their earliest flags.

According to popular legend, the first stars-and-stripes flag was sewn by Betsy Ross, who based it on a rough drawing supplied by George Washington. She is said to have displayed her handiwork to Washington and others in May or June of 1776 at a house somewhere on Arch Street in Philadelphia. To show how she made the pattern for her stars she is said to have folded a sheet of paper and then, with one snip of her scissors, to have cut out a perfect pentagram. The 13 stars, so the story goes, were arranged on a blue field in a circle to imitate King Arthur's Round Table. It is the first design in the illustration below. Replicas of this "Betsy Ross flag" are flapping this year all over the U.S. The flag is on a 13-cent stamp. It appears in such famous paintings as "Washington Crossing the Delaware," "The Spirit of '76" and the picture that used to be in many schoolbooks showing Betsy's nimble fingers at work on Old Glory.

Alas, the story has been totally discredited. Its sole source was Betsy's grandson, who said he heard it when he was 11. Not a single flag with 13 stars in a circle has survived, and there is no evidence that such a flag existed in Revolutionary times. Many historians doubt that a stars-and-stripes flag of any



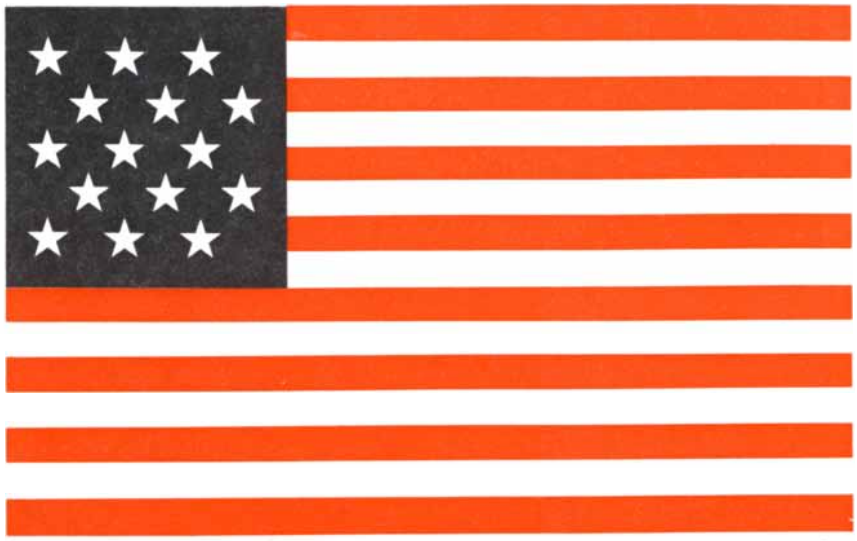
Alleged arrangements of the 13 stars on early flags of the U.S.

design flew during a single sea or land battle of the Revolution.

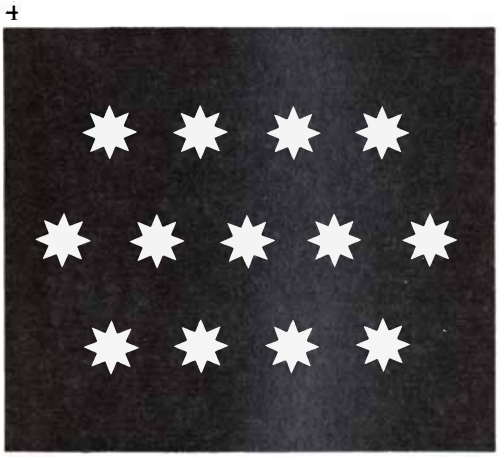
The second design in the illustration shows how the stars are arranged on a flag that John Hulbert, captain of a company of minutemen from Long Island, is said to have flown in 1775. This arrangement too is not supported by any evidence. The truth is that no one knows who designed the first stars-and-stripes flag and that almost nothing is reliably known about the flag's earliest history.

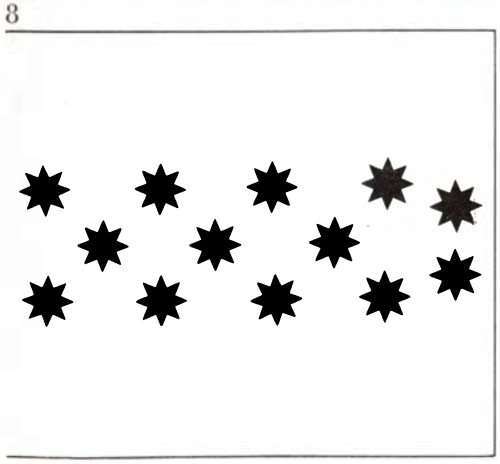
We do know that on June 14, 1777, the Second Continental Congress resolved "that the flag of the United States be thirteen stripes, alternate red and white; that the union be thirteen stars, white in a blue field, representing a new constellation." There is not a word about the number of star points, how the stars are to be arranged or which color of stripe should outnumber the other. As a result, from 1777 until 1795 there were wild variations in flag designs. Some flags even violated the Congressional order by having red and blue stripes or red, white and blue stripes; some had blue stars on a white background.

The stars had five, six, seven or eight points and were arranged in all kinds of ways. No. 3 in the illustration is on a flag said to have been used by the Bennington militia in 1777. Nos. 4 and 5 appear in paintings by a Dutch artist on flags alleged to have been flown by John Paul Jones on ships in 1779. No. 5, with its field stretched horizontally, was perhaps the most common pattern on flags from 1777 to 1795. (Note that it is No. 2 rotated 45 degrees.) No. 6 is supposed to have been on the flag of a Maryland regiment and No. 7 on a flag flown in Boston, both around 1781. No. 8 is credited to the North Carolina militia of 1781.



The flag of 1794, with 15 stripes and 15 stars





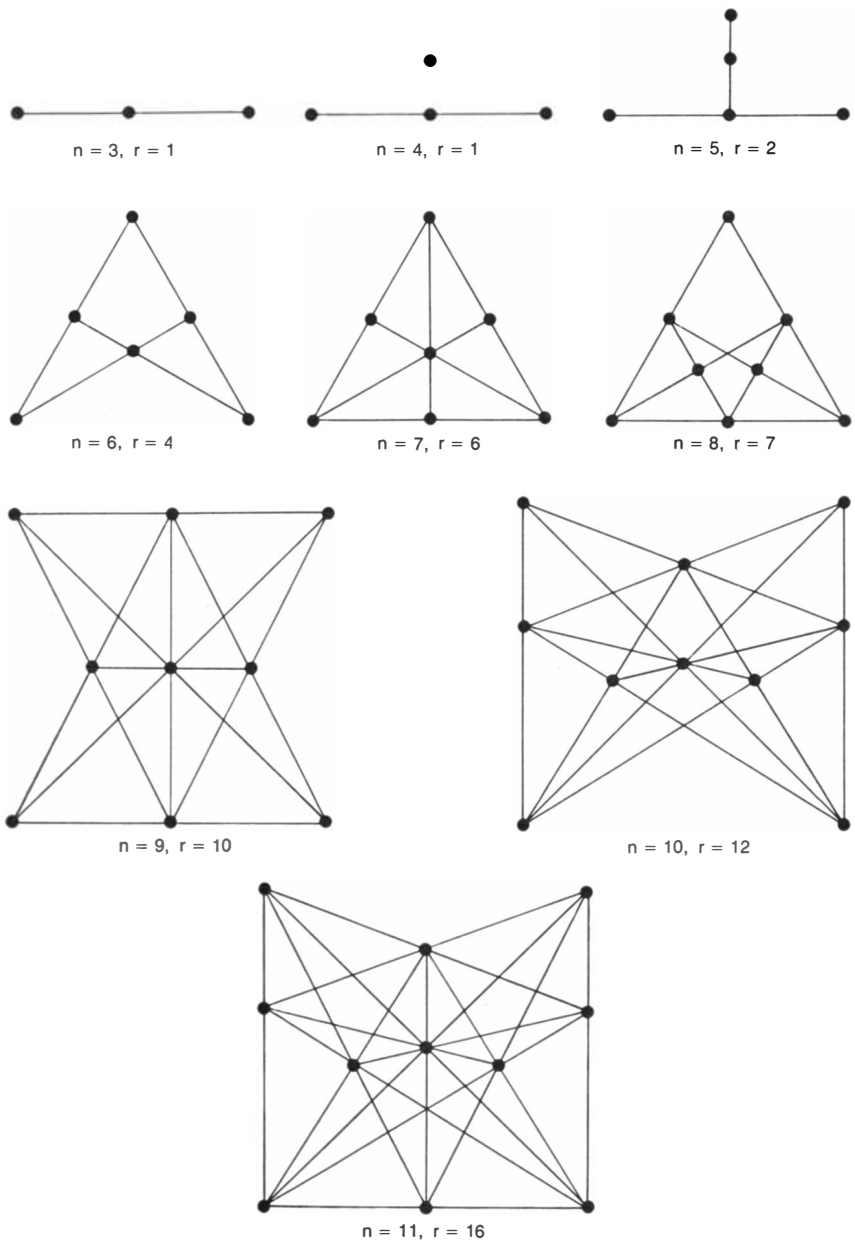
In 1794, after Vermont and Kentucky had become states, Congress decided to add two more stars and two more stripes, arranging the stars as is shown in the illustration above. One congressman thought this was "a trifling business which ought not to engross the attention of the House." Another called it "a consummate specimen of frivolity." But the bill was passed, and the flag became official the following year. This was the flag that Francis Scott Key saw flying over Fort McHenry during the War of 1812 and that inspired him to write "The Star-spangled Banner."

In 1817, after five more states had joined the union, Congress decided to go back to 13 stripes and to have a new star added each time a state was admitted. A flag with 20 stars in a four-by-five rectangle became official in 1818. From then until 1913 there were 24 changes of the flag as stars were added. The stars were usually five-pointed and in staggered rows, although many other arrangements were popular, including star-shaped patterns. After the admission of New Mexico and Arizona in 1912 the flag was stable for 46 years, its 48 stars arranged in a six-by-eight rectangle. In 1959, to accommodate a new star for Alaska, President Eisenhower ordered a seven-by-seven field, the rows staggered with odd rows aligned at the left. When Hawaii became a state later in 1959, he ordered the field changed to its present form of 50 stars in staggered rows of six and five.

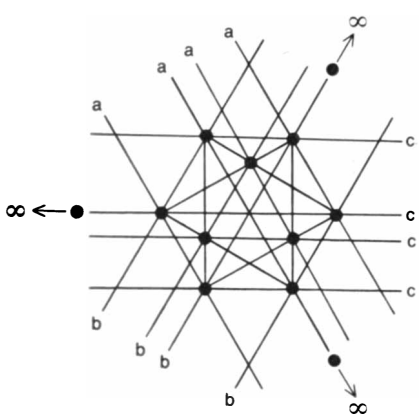
There are many puzzles based on the arrangement of n spots on a field, but the oldest and most popular are known as "tree plant" problems. They have that name because in early puzzle books they were usually presented with a story about a farmer who wishes to plant a certain number of trees in an orchard so that the pattern of trees will have r straight rows of exactly k trees in each row. The puzzles were made difficult by maximizing the number of rows. Surprisingly, the general problem of determining the largest number of rows, given n and k, is nowhere near solved even when k is 3 or 4.

"These tree-planting puzzles," wrote Henry Ernest Dudeney, England's greatest puzzle expert (in *Amusements in Mathematics*), "have always been a matter of great perplexity. They are real 'puzzles' in the truest sense of the word, because nobody has yet succeeded in finding a direct and certain way of solving them. They demand the exercise of sagacity, ingenuity and patience, and what we call 'luck' is also sometimes of service. Perhaps some day a genius will discover the key to the whole mystery."

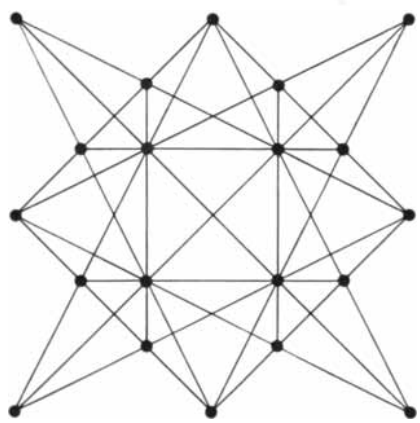
When k is 2, the problem is trivial. If n points are arranged so that no three are in line, every pair forms a row of two. When k is 3, the problem not only becomes interesting but also is related to such mathematical topics as balanced-block designs, Kirkman-Steiner triples, finite geometries, Weierstrass elliptic functions, cubic curves, projective planes, error-correcting codes and many other aspects of significant mathematics. The latest and most definitive paper



Some solutions to the three-in-a-row problem



Twelve points in 19 three-point rows



Sam Loyd's solution for rows of four

on the problem is by Stefan A. Burr, Branko Grünbaum and N. J. A. Sloane, all leading mathematicians. Called "The Orchard Problem," it appeared in *Geometriae Dedicata,* Vol. 2 (1974), pages 397–424. What follows is taken mainly from that paper.

The first nontrivial reference on tree-plant problems is a book called *Rational Amusement for Winter Evenings,* by John Jackson, published in London in 1821. According to Dudeney, who owned a copy, it contains 10 such puzzles. The mathematician J. J. Sylvester worked continually on the general problem from the late 1860's until his death in 1897. (The temperamental Sylvester had a stormy career. He was denied a degree at Cambridge because of his Jewish faith but obtained one at Trinity College of the University of Dublin. He was a professor at the University of Virginia for three months until an altercation with a student led to his resignation. At Johns Hopkins he founded *American Journal of Mathematics.* One of his books is *The Laws of Verse* [he was fond of writing poetry], and for many years in England he was a barrister.)

Maximum solutions for three-in-a-row tree plants, from 3 through 11 points, are shown in the top illustration at the left. Note that not until n equals 9 does the maximum number of rows exceed n. It is easy to get eight rows with nine points (a three-by-three square array does it), but adding two more rows is a bit tricky. In his book Jackson introduced the problem as follows:

Your aid I want, nine trees to plant
In rows just half a score;
And let there be in each row three.
Solve this: I ask no more.

The solution, derived from a famous theorem in projective geometry called the Pappus theorem, has been credited to Isaac Newton.

The 11-point pattern is given by Dudeney (Problem 213 of his *Amusements*) as a military puzzle. In lecturing on tree-plant problems Sloane has simplified Dudeney's narrative line by describing a World War I battlefield on which 11 Turks were surrounded by 16 Russians. Each Russian fired once, and each bullet passed through exactly three Turkish heads. How were the Turks standing? The remarkable solution—11 points in 16 rows of three each—is said by Dudeney to have been constructed about 1897 by the Reverend Dr. Wilkerson. (Does any reader know who he was?) Sloane tells me that this is the only practical application of the orchard problem he knows, although he once invented a fictional "Haltwhistle triode" of n pins, which, because of unexplained capacitance effects, have to be arranged in rows of three.

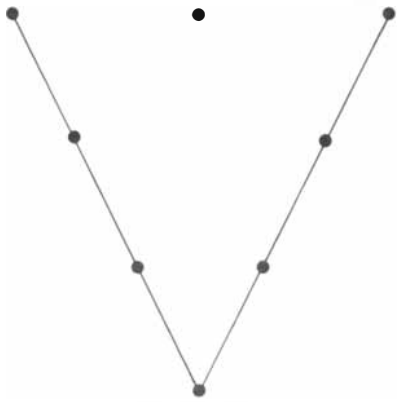
Twelve points will make 19 rows.

n = 4, r = 1

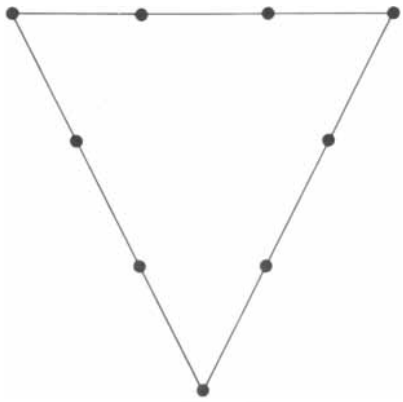
n = 5, r = 1

n = 6, r = 1

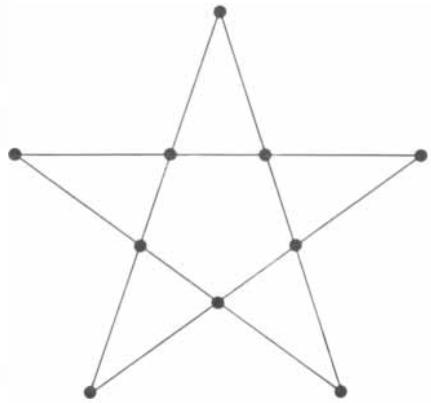
n = 7, r = 2



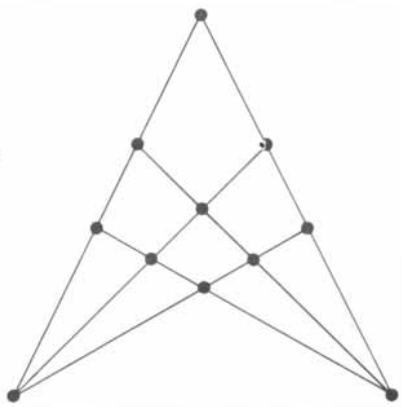
n = 8, r = 2



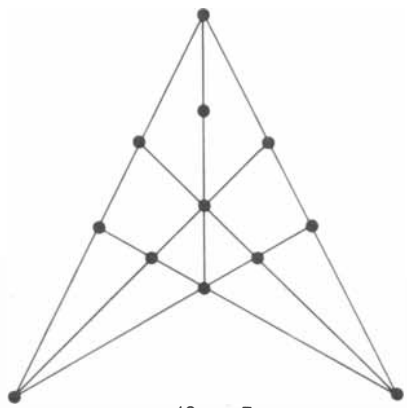
n = 9, r = 3



n = 10, r = 5

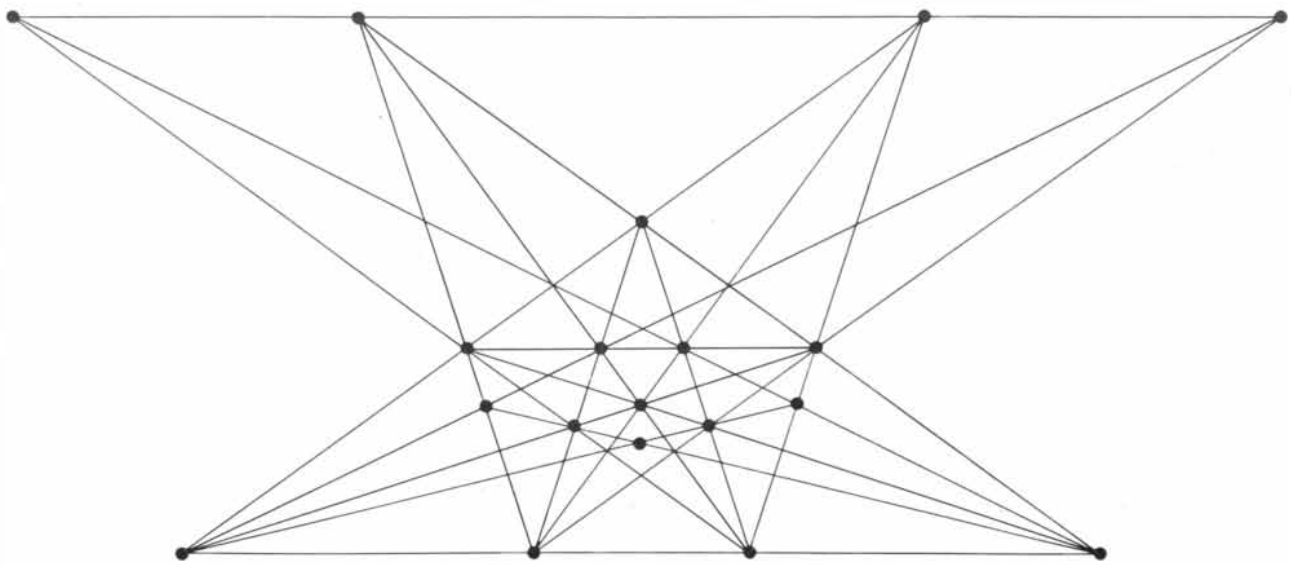


n = 11, r = 6

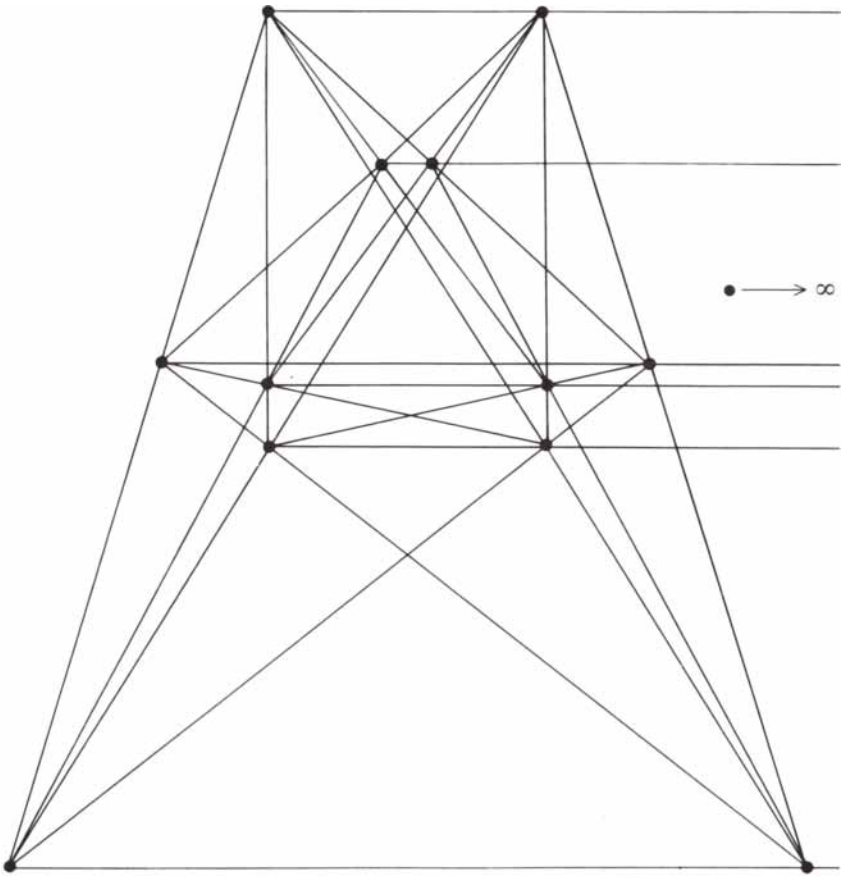


n = 12, r = 7

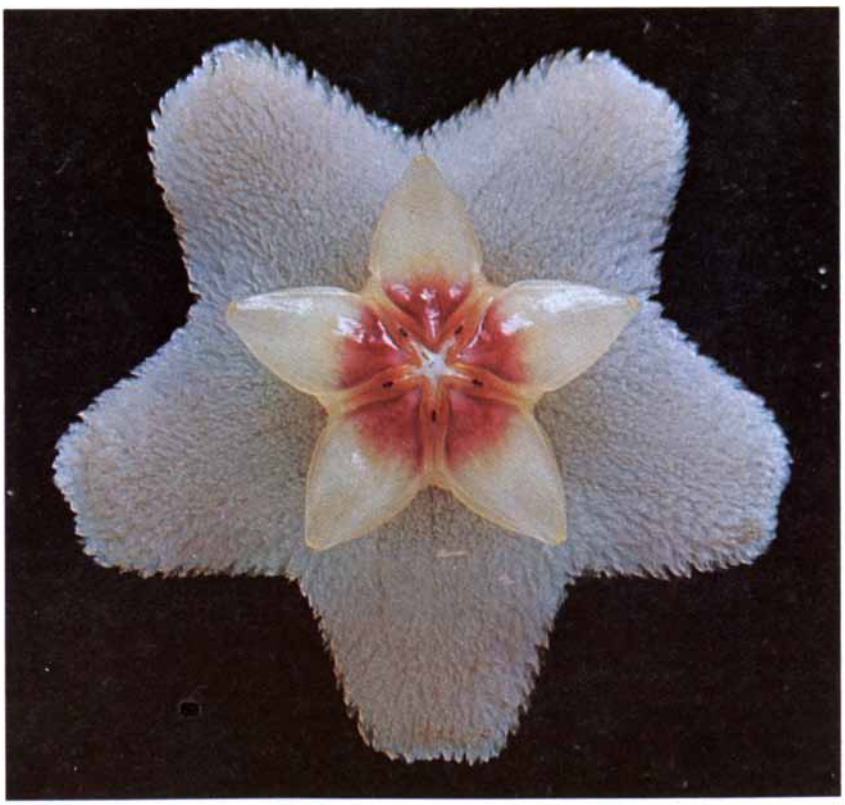
More solutions for rows of four points



Nineteen points in 19 rows of four



Thirteen points (one at infinity) in three-point rows



A blossom of **Hoya carnosa,** *a milkweed, in a pattern of points*

This result was announced, apparently for the first time, by R. H. Macmillan in a note to *The Mathematical Gazette,* Vol. 30 (1946), page 109, and is proved maximal in the Burr-Grünbaum-Sloane paper. The illustration at the bottom left on page 104 shows a way of drawing the pattern symmetrically by placing three points and one line at infinity. This pattern can be projected to give a standard solution, but it is difficult to show on a small sheet of paper. Imagine the pattern viewed in perspective with the eye below and to the right. Each of the three sets of four parallel lines (labeled with *a*'s, *b*'s and *c*'s) will converge, the three meeting points lying on the horizon to form the 19th row.

Only one other three-in-a-row case has been solved, that of $n = 16$. Burr, Grünbaum and Sloane prove the maximum number of rows to be 37. Thus the lowest unsolved case is $n = 13$. The best-known result, 22 lines, is shown in the top illustration at the left. One point is at infinity. If the pattern is viewed in perspective from the left, the six parallel horizontal lines, of two points each, will converge on the 13th point. In other words, the pattern can be projected to give a standard solution, but it is difficult to show it except on a large sheet. The best-known results for n equals 14 through 20 are 26, 31, 37, 40, 46, 52 and 57.

When the number of points demanded for each row rises to four, the problem becomes more difficult. As in the case of $k = 3$, maximums have been established through 12 points, with 13 as the lowest unsolved case. Examples of best patterns from four through 12 points are shown in the top illustration on the preceding page. They are taken from Grünbaum's "New Views on Some Old Questions of Combinatorial Geometry," a paper he gave at the International Colloquium on Combinatorial Theories in Rome in 1973. It is scheduled for publication (any year now) in the colloquium's *Proceedings*. The best-known results for n equals 13 through 20 are 9, 10, 12, 15, 15, 18, 19 and 20.

The case of $n = 10$ has many topologically distinct solutions (see Chapter 2 of my *Mathematical Carnival*), providing Dudeney and his American rival, Sam Loyd, with more than a dozen puzzles. When n equals 16, the best-known result (15 rows) is an elegant pattern of three nested pentagons surrounding a central point [*see illustration on opposite page*]. In *The Canterbury Puzzles and Other Curious Problems,* where this arrangement solves Problem 21, Dudeney admits he cannot prove it but says he has a "strong pious opinion" that 15 rows is maximal. It is surprising and infuriating that no one has done better with 17 points than this same pentagonal pattern with the 17th point added as a total irrelevancy.

The pentagonal pattern appears in the

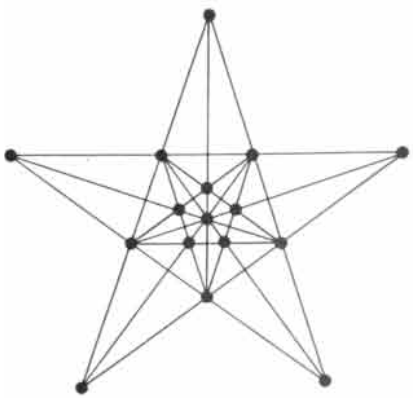
photograph on the opposite page of a blossom of *Hoya carnosa,* a member of the milkweed family. Imagine the sides of the outer petals extended to points as shown in the diagram below. The three largest sets of pentagonally placed vertexes, together with the flower's center, beautifully give the 15 rows of four each in the best-known solution for 20 points.

The number of rows of four begin to exceed the number of points when n equals 21 (for which 23 is believed maximal), but when n equals 18, 19 or 20, rows equal to n are possible. The bottom illustration on page 105 shows how Grünbaum gets 19 rows with 19 points. The case of $n = 20$ is answered in two of Loyd's puzzle books with 18 rows [*see illustration at bottom right on page 104*]. In about 1945 Macmillan found a simple, beautifully symmetrical way to make 20 rows with 20 points. It was later rediscovered by Grünbaum, who was unaware at the time of Macmillan's unpublished results. Can readers construct it before I reveal it next month?

When n equals 13, no one has done better than the nine rows shown in the illustration on page 109. Dudeney gives this as the solution to Problem 149 of his *Modern Puzzles and How to Solve Them,* reprinted as Problem 435 in *536 Puzzles & Curious Problems* (Scribner's, 1967), a collection of his puzzles. If any reader can improve on this solution, I shall report it in a later column.

Very little work has been published on rows of five or more points. According to Grünbaum, the best results known for $k = 5$, points five through 20, are 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 6, 6, 7, 9, 10 and 11. I do not know how many of them have been proved maximal. A set is known of 35 symmetrically arranged points that determine 36 rows of five. Grünbaum conjectures that no smaller example exists in which the number of rows of five exceeds the number of points. I know of no work that has been done on the extension of tree-plant problems to spaces of three dimensions or more.

Burr, Grünbaum and Sloane conjec-



The flower pattern: 16 points in 15 rows

SCIENCE/SCOPE

Now, with the launch of the second Marisat, the new maritime communications satellite system blankets the sea lanes in a 120-million-square-mile area of the Atlantic and Pacific. For the first time, ships can communicate clearly and without interruption with other ships and with shore bases. The two Marisat spacecraft -- built by Hughes for COMSAT General Corporation -- are relaying high-quality voice, telex, facsimile, and data transmission. Vessels flying the flags of 10 nations already are equipped for satellite communications. The U.S. Navy also is leasing a substantial portion of Marisat's capacity for its fleet communications needs.

Men at sea have depended in the past primarily on international Morse code and radio telephony that often was disrupted and delayed by adverse weather conditions, signal fading, or crowded frequency channels. The Marisat system has eliminated these delays.

Marisat earth stations in Connecticut and California, interconnected with terrestrial networks and linked by 24-hour voice and data lines to COMSAT General's Washington, D.C., control center, complete the Marisat satellite system.

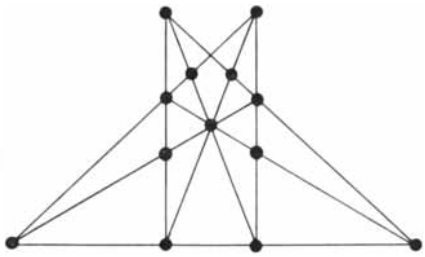
Digital watch modules with three new features have been introduced by Hughes. One is both a regular watch and a stopwatch (or chronograph) which measures to 1/100th second. Another module, in addition to the usual time and date readout, displays a pre-programmed message of up to four five-letter words. Custom modules with special commercial messages or slogans are available.

A wrist watch that's also a nine-function calculator is the third new type in the Hughes line. The solid-state electronic module, which measures only 1.4 by 1.25 inches, features an eight-digit LED display for both calculator and watch functions. The keys of its standard calculator keyboard can be pressed with a pencil point. Hughes does not market a watch to consumers, but is the largest producer of modules for watch manufacturers.

Hughes needs systems-level engineers: Sonar Systems -- design and develop surface and subsurface ASW tactical and surveillance sonar systems for Undersea Systems Laboratory....Special Projects -- engineering analysis of foreign air defense communications, command and control systems. Evaluation of C^3 networks and system performance evaluation. Requirements: BS or higher degree, U.S. citizenship. Please send resume to: Engineering Employment, Hughes Aircraft Company, P.O. Box 3310, Fullerton, CA 92634. An equal opportunity employer.

The location of enemy artillery can be pinpointed in seconds with the U.S. Army's new AN/TPQ-37 artillery locating radar. Its three-dimensional antenna scans the horizons with a pencil-shaped beam which moves so fast it forms an electronic barrier that can detect incoming projectiles as they rise above the horizon. These are tracked and their trajectories are back-plotted to locate the firing weapons, often before the first shell hits the ground. A contract for further development and for limited production of 10 ALRs has been awarded to Hughes.

The ALR can be deployed quickly and set up in 30 minutes. It has two main units: an antenna trailer towed by a 5-ton truck which carries the transmitter, receiver, and generator, and an operations unit housed in a standard S-280 shelter mounted on a 2½-ton truck. The shelter has room for two operators and a supervisor, although one man can do the job if necessary.



Thirteen points in nine rows of four

ture that, with the four exceptions of n equals 7, 11, 16 and 19, the formula for the maximum number of rows of three, given n points, is

$$1 + \left[\frac{n(n-3)}{6}\right].$$

The brackets indicate rounding down to the nearest integer. If the conjecture is correct, no one can do better with 13 stars than 22 rows. For rows of more than three points there are not even good conjectures.

The solution to last month's cryptarithm is

$$\begin{array}{r} 4379 \\ \underline{\quad 79} \\ 39411 \\ \underline{30653\ } \\ 345941. \end{array}$$

The problem appears in Joseph S. Madachy's *Mathematics on Vacation* (Scribner's, 1966). Other questions in July's column are answered as follows.

To prove that the series

$$1/1 + 3/2 + 5/4 + 7/8 + \ldots$$

converges on 6, first halve each term to obtain

$$1/2 + 3/4 + 5/8 + 7/16 + \ldots.$$

Subtract this from the original series:

$$\begin{array}{l} 1/1 + 3/2 + 5/4 + 7/8 + 9/16 + \ldots \\ \quad - 1/2 + 3/4 + 5/8 + 7/16 + \ldots \\ \hline 1 + 1 + 1/2 + 1/4 + 1/8 + \ldots. \end{array}$$

The sequence at the right of the first term has a sum of 2, so that the entire series must have a sum of 3. Since this series is half of the original series, the original has a sum of 6.

If we begin with any natural number and then repeatedly take the nth root and multiply it by m, the limit, L, is

$$m^{n/(n-1)}.$$

Don Morran's proof is as follows. Assume that $mL^{1/n}$ approaches a nonzero limit L. At the limit $mL^{1/n} = L$; $m^nL = L^n$; $L^n - m^nL = 0$; $L(L^{n-1} - m^n) = 0$; $L = m^{n/(n-1)}$.

BOOKS

Snack-food factories, biological energy, the new universe and Newton's alchemy

by Philip Morrison

SNACK FOOD TECHNOLOGY, by Samuel A. Matz, Ph.D. The Avi Publishing Company, Inc. Westport, Conn. ($29). Popcorn has been an American foodstuff at least since the ancient deposits in Bat Cave, which were in place 4,000 years ago. The attractive, convenient, crisp white puffs expand from the unique kernel, which bursts sharply when it is heated at atmospheric pressure. Hundreds of square miles of good Corn Belt farmland are planted to this crop and about as much again grows in home gardens. It is no traditional area of husbandry: some 30 years ago the volume expansion of a good grade of popcorn was about 26 to one, whereas the expansion of the hybrids patiently bred at Purdue and Iowa State is now as much as 40 to one. (This seems a benign enough change: the producer cuts his costs and the consumer of a fixed volume ingests fewer calories!)

Popcorn and potato chips, baked wheaten pretzels and crackers and bits of cornmeal dough extruded at high temperature and pressure so that they puff into a variety of forms for subsequent drying—those are the chief American snacks. Peanuts and other nut meats are popular (P. T. Barnum brought roasted peanuts in the shell to New York in 1870); even pickled pigs' feet and beef jerky are well known, if rather restricted in appeal. The big volume, however, is in starchy foodstuffs that are crisped, coated with oil, flavored primarily with salt and marketed to be eaten out of the hand. Insofar as French-fried potatoes and a dozen other savory temptations eaten avidly in the marketplace and the home the world over fit this description, it must be a mark of some quite general human indulgence.

This coolly technical book is devoted to the snack-food industry in the U.S., and in particular to packaged snacks to be eaten cold, between regular meals. It treats systematically the ingredients, equipment and processes required to produce the entire wide range of snacks, together with packaging and even the development of novelties. The aim is to introduce the reader to the quite substantial periodical and patent literature. The general reader will not find it all of equal interest, but the picture the book conveys is engrossing: from the *p*H of vanilla wafers to the tradition of secrecy in the pretzel trade. ("The production of pretzels is more of an art than a science.")

For example, a special pretzel salt from a particular salt dome of the Louisiana coast has the peculiar characteristic that "it breaks up into uniformly flat, rectangular-shaped particles upon crushing." A photograph is presented in evidence. Such granules adhere particularly well to the smooth, hard pretzel surface, and the natural salt impurities add a note of "huskiness." Pretzel-twisting machines are not described in detail; stick pretzels can be made with almost any flour, but "the flour used in twisted pretzels is very critical."

The slice of peeled potato from one millimeter to two millimeters thick that is fried to become the natural potato chip is generally acceptable only if its color after frying is golden rather than dark brown. The slice is darkened by the presence of reducing sugars that react with cellular amino acids. Potatoes ought therefore to be stored near room temperature so as to slow the accumulation of the sugars; the lower the temperature is, the more rapidly the unwanted sugars form. The apparent packed bulk density of the thin, bent plates of fried potato we call chips is less than a twentieth of their true density, and so the packing and shipping are substantially controlled by that loose, bulky fit. In 1971 the ingenious A. L. Liepa patented a scheme for simulating potato chips by rolling out thin sheets of quite solid dough in which reconstituted dehydrated potatoes play a major role, with support from cereals and other additives. Carefully cut pieces are held in a saddle shape by molds during the frying process. Such mashed-potato chips stack neatly in a shipping cylinder as Pringles (the registered trademark of the makers, Procter & Gamble) at an apparent density three times greater than conventional chips, and they now enjoy commercial success. It is a clear victory of geometry over substance.

The oil layer of the fried snack is a source of serious problems because the limit to the shelf life of such foods is set by rancid taste, which develops from the slow oxidation of double carbon bonds in the oils. The reaction is catalyzed by light and by trace amounts of metal ions, particularly copper. No copper, bronze, brass or Monel metal is allowed to touch the fat; equipment surfaces are preferably of welded stainless steel; the empty space at the top of tanks of oil is filled with nitrogen. Expensive products such as nut meats are packed in a vacuum or are held under refrigeration. Opaque pigmented plastic films are almost universal in food pouches; moisture loss and light leak are fought with composite bagging materials and sometimes with a layer of aluminum foil.

The oils favored in order to avoid rancidity are the saturated ones, which are perhaps the least healthful. The fried snack foods are prepared with hydrogenated soybean or cottonseed oils. Coconut oil is widely used for spray fat, as on the familiar unsweetened but rich round crackers. That oil is based mostly on lauric acid, which is highly saturated but has a rather low melting point; the main carbon chain is short, only 12 atoms long instead of the 18 or 20 of the commonest fatty acids. Antioxidant additives are incorporated to fight rancidity; they include such unexpected snack ingredients as tertiary butylhydroquinone (TBHQ). The Food and Drug Administration allows only 200 parts per million of fat and oil for such constituents, whose side effects the book does not mention.

The mechanical engineering of all this slicing and extruding and frying and packing is described here quite clearly, with a good many references to specific manufacturers. If you want to start up a corn-puff product line, expect to spend some $60,000 before installation in order to turn out 300 pounds of cheese curls per hour. You will own an auger feeder for the cornmeal mix, a pressurized collet machine with a water-cooled die head to turn out the puffs, a conveyor that feeds the puffs steadily to the drying oven, plenty of hot air and another transfer belt that drops the dried curls into the tumbler where they are coated with oil and cheese powder. Pumps and blenders are required too. The packaging machinery is extra. A cheese curl of quality will run about 60 percent cornmeal collets, 25 percent oil and from 10 to 12 percent cheese powder (mostly a spray-dried emulsion of cheddar, well salted and with added color).

Flavors are the product of quite a distinct industry, since synthetic flavors call for a heavy investment in sophisticated analytical techniques before the flavor can be caught. Snack foods generally depend on rather few and simple flavorings. As a response to attacks on snack foods as "empty calories," the

present trend is to enrich them with supplemental nutrients. Vitamins are useful in such small amounts that it is not difficult to add them, but effective amounts of minerals are hard to include without "textural and visual defects," which are apparent with, say, half a gram of tricalcium phosphate per portion. Protein can be added—with care—in the form of defatted soy flour, which may constitute up to 15 or 20 percent of the protein of the cornmeal puff. Pure amino acid supplements are possible, but they are tricky and not cheap. The only fat required in the human diet, arachidonic acid or its linoleic acid precursor, is needed mainly for infants; there is little indication for adding it to the fatty snack foods.

The technology laid out here is rather disarming. The snack-food technologist is trying to please, and by and large he takes no part in fraud or deception. That the result is not very praiseworthy seems to reflect on grander problems, from the motives and scale of U.S. production to the uses of leisure, to television commercials and on to the social role of youth and who knows what profundities of depth psychology. Popcorn will be munched, one expects, as long as maize grows. Close-packed potato chips and TBHQ antioxidant appear more transient; there will be ebb and flow in the salience of the junk-food junkies. Sugar and cigarettes belong to other and gloomier chapters.

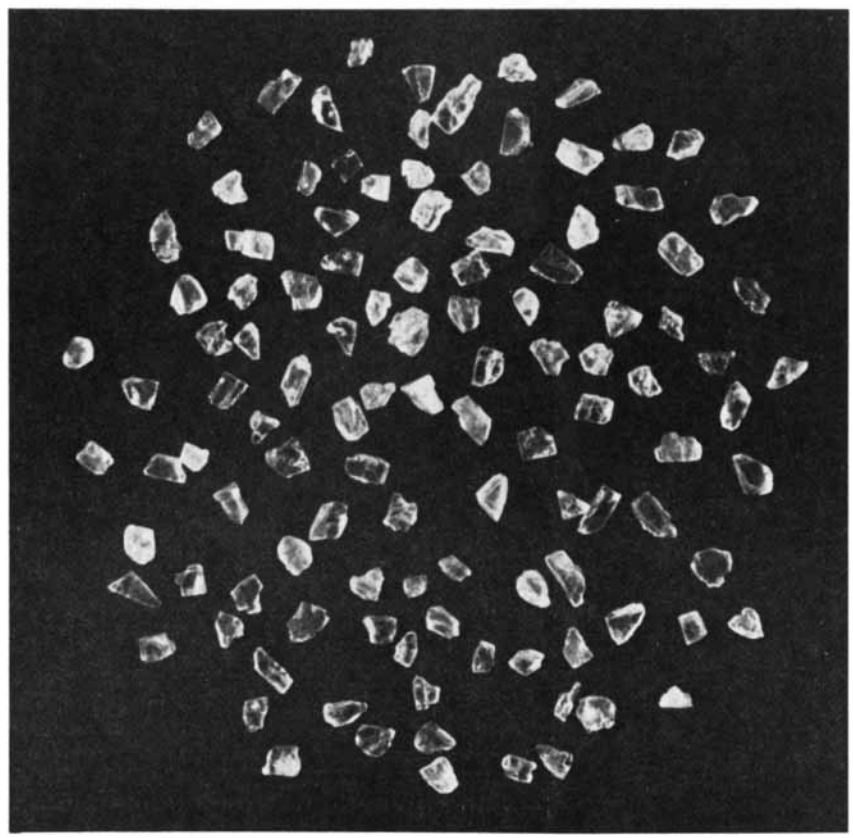
ENERGY TRANSFORMATION IN BIOLOGICAL SYSTEMS. CIBA FOUNDATION SYMPOSIUM 31. IN TRIBUTE TO FRITZ LIPMANN ON HIS 75TH BIRTHDAY. Elsevier/Excerpta Medica/North Holland ($29.25). The genetic specificity of evolving life turns out to be a matter of a marvelous system of information storage and transfer, a codicil of the second law of thermodynamics. On the other hand, the activity of living forms not as pedigrees and species but as individuals functions of course under the sway of the first law of thermodynamics, which governs energy transfer. It is no less important and no less molecular, for all that it has some of the flavor of classical physiology. Behind the information lie the long helical molecules of the nucleic acids, and behind the transfer of energy lies the relatively small molecule of adenosine triphosphate (ATP). The energy of ATP is released when the molecule is cleaved by the enzyme adenosine triphosphatase (ATPase). In the opening paper of this broad and lively symposium, written by Fritz Lipmann, there is a remarkable electron micrograph, made by Humberto Fernandez-Moran, that shows the ATPase molecules lined up along the surface of two mitochondria like so many tadpoles.

It was Lipmann who, in the years before World War II, saw the key importance of the "high-energy phosphate bond" and gave it its wiggle symbol (~P). He reviews the "roots of bioenergetics" in the symposium, whose participants include the household names of modern bioenergetics. Besides Lipmann himself there were Sir Andrew Huxley, Sir Alan Hodgkin, Sir Hans Krebs and a cluster of other British, U.S., French and German research workers, young and old. Their papers are complemented by a report of the discussion that followed each of them, from which it is no surprise to glean a lot of new and fascinating ideas and puzzles.

After the opening paper two reviews discuss the function and origin of two widespread cellular organelles: chloroplasts and mitochondria. In the chloroplasts light energy is converted into ATP and coenzyme and is stored in carbon-carbon bonds. Those in turn are burned in the mitochondria with oxygen to form ATP again for powering metabolism throughout the cell. The flow of carbohydrates and gases can be macroscopic; the entire atmosphere is part of the process. The elaborate circuit is still far from being known in detail, in particular the structures of the enclosing membranes that arrange for some remarkable chemical gradients. That both of these organelles arose from free-living prokaryotic cells is widely believed, although the descent of green chloroplasts from blue-green algae is the more easily accepted. Why are these organelles, like all prokaryotic cells, so slow to evolve, small and single-celled? Are they biochemical inventors rather than elaborating genetic systems? Lipmann cites François Jacob's answer: Prokaryotes do not die. A cycle of life and death is essential for evolution.

One participant, F. R. Whatley of Oxford, tells of a form of amoeba that slowly poisons itself by secreting lactic acid. It has no mitochondria, and so it is in fact an anaerobic organism, but it needs oxygen in order to dispose of the lactic acid. Encapsulated within itself in small cavities, it holds a supply of aerobic bacteria; the metabolism of those aerobic residents is able to "mop up the unwanted lactic acid" even though they make ATP not for the host but only for themselves. There is an instance of symbiosis prior to the postulated origin of the mitochondrion ("an evolutionary fairy story").

Then there is the matter of light and life. An admirable modern review of bioluminescence is full of novelty. Certain colorless proteins extracted from coelenterates emit light in a flash when calcium ion is added, with a good yield of 10 or 20 percent. The stored molecule need not be enzymatically modified to yield light; a protein-bound excited state radiates when it is triggered by the ion.



Special salt crystals used to coat pretzels, from **Snack Food Technology**

executive health

the report that briefs you on what to watch

Volume XII, Number 9 • June, 1976 • Pickfair Bldg., Rancho Santa Fe, California 92067 • Area 714:756-2600

ON "THE UNSUSPECTED ILLNESS" – HYPOTHYROIDISM

How it can subtly sap health - physical or mental, or both - in a remarkable variety of ways . . . and a simple test you can use to help uncover it.

Of all the problems that can affect health, none, it now appears, may be more common, more easily corrected, yet more often untreated and unsuspected than hypothyroidism, or low thyroid gland functioning.

It's hardly a new story that severe thyroid deficiency can have its obvious and devastating effects. But evidence has been mustered now that hypothyroidism in mild or moderate form can be responsible for any or many of an astonishingly large number of diverse problems, ranging from low energy and excessive fatigue to repetitive infections and chronic headaches, and from circulatory disturbances and stubborn skin disorders to difficulties with memory and concentration, and even mental depression, to name just a few.

Moreover, thyroid deficiency—with startlingly high incidence, affecting, according to one estimate, 40% of the population to some degree—may have much to do with blood cholesterol elevation, atherosclerotic disease of the arteries, and heart attacks.

Yet it commonly escapes diagnosis—for one reason, because of the failure of many physicians as well as lay people to recognize that low thyroid function can have effects that vary considerably from one victim to another; and, for another reason, because commonly used tests are not always reliable.

The controller and the confusion

It's almost impossible to exaggerate the importance of the thyroid, a small butterfly-shaped gland in the neck that weighs less than an ounce.

It is the thyroid which controls metabolism—the process by which nutrients are transformed into energy and many essential chemical reactions in the body are carried out.

Minute secretions of the gland—less than a spoonful a year—are responsible for much of the body's heat production; they help maintain blood volume and the vital flow of the circulatory system; are essential for muscle health; and heighten the sensitivity of nerves. Every organ, tissue and cell in the body is affected by the hormone secretions of the thyroid.

Both the cretin child and the myxedematous adult demonstrate the all-pervading influence of the thyroid.

YOU HAVE BUT ONE LIFE

— doesn't it make sense to find out how to take the best care of it you can?

What keeps well people well? Medical men have long concentrated upon sick people and how to get them well . . . not upon well people and how to keep them well. Now many top research scientists are concentrating their efforts on *preventive medicine . . . how to keep well people well.*

May we suggest that you get the benefits of this new research for yourself? Do as so many thousands of executives do. Subscribe to Executive Health Report. The members of our *Editorial Board* are among the world's most distinguished authorities on *preventive medicine.* Their wise advice can help you not only live longer but enjoy those extra years! (Note their high qualifications as shown on opposite page.)

Executive Health Report is not sold on newsstands but only by private subscription at $18 a year in the U.S.A. and its possessions. $19 a year in Canada and Mexico. All other countries $24 by surface mail, $28 by air mail. Individual reports (back issues) $1.50 per copy.

Subscribe now under this unusual introductory offer:

(1) Your choice of any three of the reports listed below ($1.50 each) FREE!

(2) Your money back at anytime during the entire year if you do not find our reports live up to your expectations.

Never forget: *"Men's lives are chains of chances"* but as Euripides saw clearly so long ago: *"Chance fights ever on the side of the prudent."* Your only insurance against "tomorrow" is what you do today. You have but one life . . . doesn't it make sense to find out how to take the best care of it you can?

Please study the reports listed here and **circle your three choices:**

Sir Hans Krebs, M.D.: On the overuse and misuse of medication.

Dr. Alton Ochsner, "On the Role of Vitamins C and E in Medicine."

James Greenwood, Jr., M.D. On Vitamin C in the Treatment of Back Pain.

Mark D. Altschule, M.D.: On the Much Maligned Egg. Are we taking a lot of pleasure out of eating unnecessarily?

Mark D. Altschule, M.D.: What Causes Your Arteries to Harden.

On Fatigue, the Great Deceiver: Why most men never get their "second wind."

If You Sit At a Desk All Day, how to avoid a pot-belly and double-chin.

Never Take a Chance. Never Ignore Any Sign of a TIA (Transient Ischemic Attack) such as a sudden temporary weakness of an arm or leg.

The Anatomy of Courage. Some truths a man must learn to accept about life.

120/80 or what? High blood pressure: the hidden hazard of your executive job.

The executive's dilemma: "Now I lay me down to sleep" (I hope!).

On "The Horizontal Exercise." New research findings about sex and how to keep your middle age young!

Professor John Yudkin: On "This Slimming Business" . . . The truth about the prevention and cure of overweight!

Dr. Linus Pauling: What About Vitamin E? Eminent investigators now suspect it may be one of the key factors to help resist disease and slow the aging process.

Go easy gentlemen, too much social drinking damages your liver (even though you may feel well, eat well, and never get tight!)

Nothing to fool with: It's that dangerous old devil, your prostate.

"The non-dieting diet." Walking is magic to keep or win back a strong, lean body (and that is just the first of its great health benefits.)

On how to live 90 to 100 healthy years! (The syndrome of longevity . . . its 7 great "constants").

On The Arthritis Mystery. Can what you eat—or don't eat—make you arthritis-prone or arthritis-resistant?

Alton Ochsner, M.D.: On "The Chair Disease" . . . Why blood clots in your veins are a little-realized occupational hazard of desk-bound executives.

Trace Minerals . . . Part I. On chromium deficiency and atherosclerosis.

Trace Minerals . . . Part II: On your danger from cadmium in the water you drink and the food you eat—and how to protect yourself!

Dr. Hans Selye: On Stress Without Distress. Your mind can make or break you!

The B Vitamins . . . Part I. On B_{12} . . . When to suspect a deficiency of this remarkable vitamin in your diet.

Stephen R. Elek, M.D. On "The Hurry-up Disease." Why it may be a key factor that triggers heart attacks at 40.

The B Vitamins . . . Part II. On Vitamin B_6 (Pyridoxine) "The Sleeping Giant of Nutrition."

Dr. Linus Pauling: For The Best of Health, How Much Vitamin C Do You Need?

The B Vitamins, Part III . . . On Biotin and Pantothenic Acid.

The B Vitamins, Part IV . . . On Folic Acid: The most commonly deficient B vitamin.

George C. Griffith, M.D.: On those irregular heart beats (arrhythmias). Some mean little or nothing, but others warn your heart is in trouble. All call for a doctor's immediate examination.

Ross Hume Hall, Ph.D.: Beware of those fabricated foods. There is still too much we do not know about all the chemical additives that are hidden away in the fine print on the labels of processed foods.

Roger J. Williams, Ph.D., D.Sc.: On your startling biochemical individuality. Some amazing facts about your body you need to know if you want to understand yourself (and other people) better.

On "The Unsuspected Illness"—Hypothyroidism. How it can subtly sap health in a variety of ways . . . and a simple test to uncover it.

Please use the coupon below, under our money-back at any time guarantee.

EXECUTIVE HEALTH, Pickfair Bldg., Rancho Santa Fe, CA 92067

Gentlemen: Enclosed is my check for $_____ for a year's subscription to Executive Health to start with this month's issue. I have circled the three $1.50 reports I am to receive free. It is understood that I am to get my money back if at any time during the entire year I become dissatisfied with your reports. In addition, I would appreciate your sending me a complete list of your 70 other reports because among them may be some from which I might greatly benefit and would otherwise miss.

NAME (please print)_____

ADDRESS_____

STATE/COUNTRY_____ ZIP_____ **SA-D**

111B

In life the package is held within some compartment and triggered, most likely, by electrical action potentials. The tiny sparks of light that glow with every wave and ripple in tropical seas display the emission by dinoflagellates. Particles in the cell flash when the *p*H is lowered in the presence of oxygen. Here the "precharged" carrier protein releases its excitation whenever the molecule is freed to reach the enzyme. The tiny cell moves strictly with the sea, but the inertial difference between the cell and its flagella flexes that delicate bond and evokes some action potential that triggers the flash.

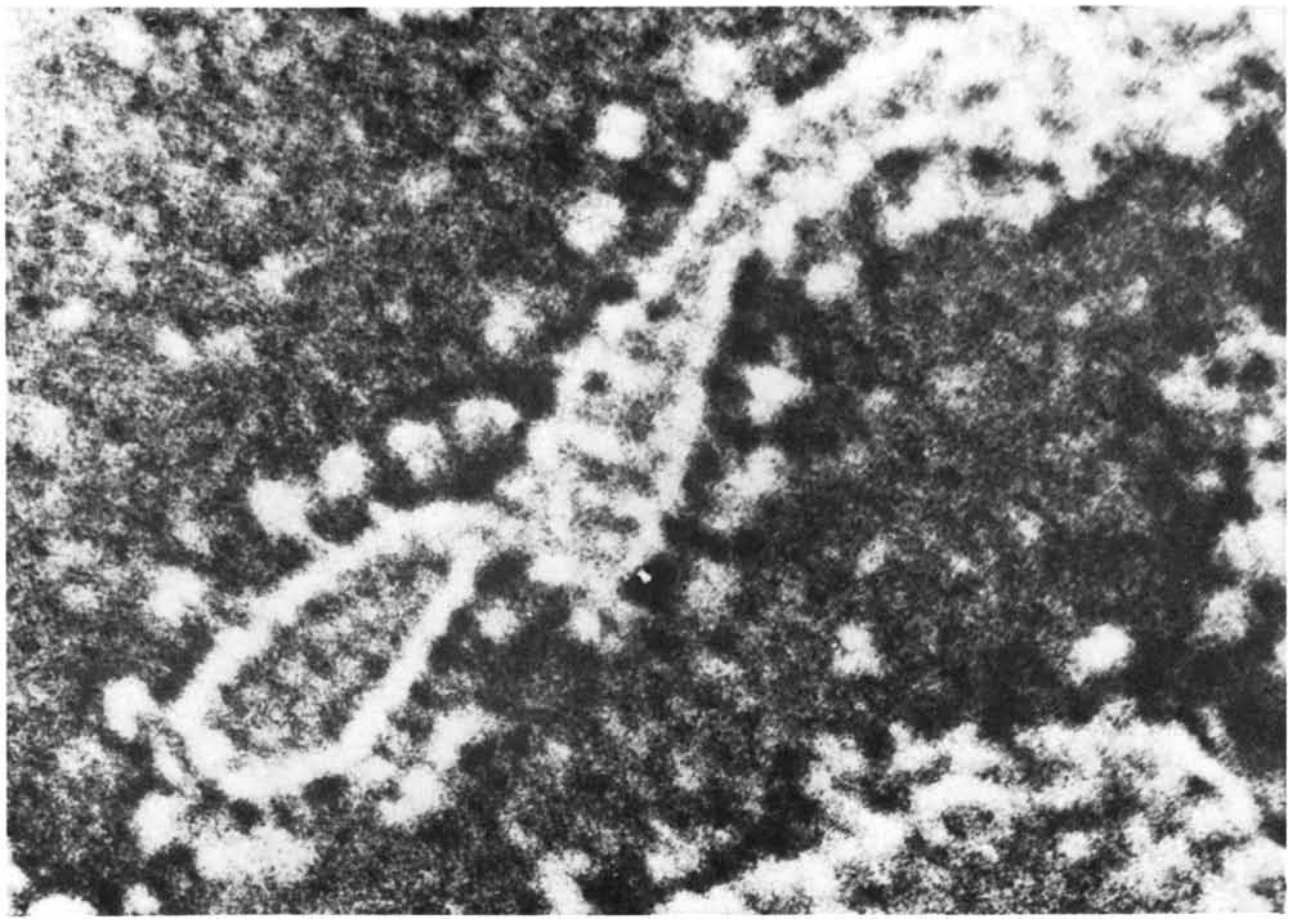
Why should a microorganism glow? It has no eyes, no signaling partners. To be sure, the total energy evolved in the flash is small; glowing cells expend perhaps 1 percent of their total energy in light, but our eyes are so sensitive that we perceive that emission. Perhaps the flashes frighten the grazing predators of the algae or even expose the grazers in turn to their own predators. One form of reef fish has luminous bacterial symbionts that it keeps in a lantern pouch, complete with shutter.

Visual purple, or rhodopsin, is a familiar specialized protein in the retina, through which the energy of photons is converted into the firing of neurons. A brief but comprehensive piece by the Tübingen biologist Dieter Oesterhelt reviews the remarkable work he and his California colleagues have done in the past few years. They have found that same purple substance in the membrane of a bacterial cell living in extreme salt concentrations. The patches of retinal purple serve, in these halobacteria, to produce ATP with photon energy in place of chlorophyll, which these bacteria lack. The informational photosensor here becomes an energetic phototransducer. There is evidence that the fundamental mechanism is a light-driven proton pump, in which ATP is made at once when the light is turned on. The pump—within the structure of the purple membrane—produces gradients in potential and *p*H that in some way convert adenosine diphosphate (ADP) to ATP without a chain of intermediate compounds.

Finally, energy must lead to life in motion. ATP somehow energizes the long, interleaved, parallel fibers of striated muscle, sliding them past one another without any change in fiber length. This model is discussed in depth by Huxley and Torkel Weis-Fogh in particular. Their discussion brings out the existence of three distinct biological motors. One, on the smallest scale, is the rotation of bacterial flagella at their roots. It is not immediately fueled by ATP. The second is the world of muscular motion we mammals share with all other advanced forms, the "active sliding" fueled by the hydrolysis of ATP. The third is new: a genuinely contractile system of rods, more or less rubberlike structures whose length is directly affected by calcium-ion concentration. These are seen clearly in some small colonial marine ciliates resembling tiny sea anemones.

Not all the papers here are sufficiently general for the inexpert reader; one of them dwells on a heterodox view of thermodynamics, not at all persuasively. Overall, however, the symposium is a feast that allows any reader with an interest in biology at the molecular or biochemical level to come in direct contact with a few outstanding research workers who span a full generation.

BLACK HOLES, QUASARS, AND THE UNIVERSE, by Harry L. Shipman. Houghton Mifflin Company ($12.95). "While I try to avoid speculation, I have

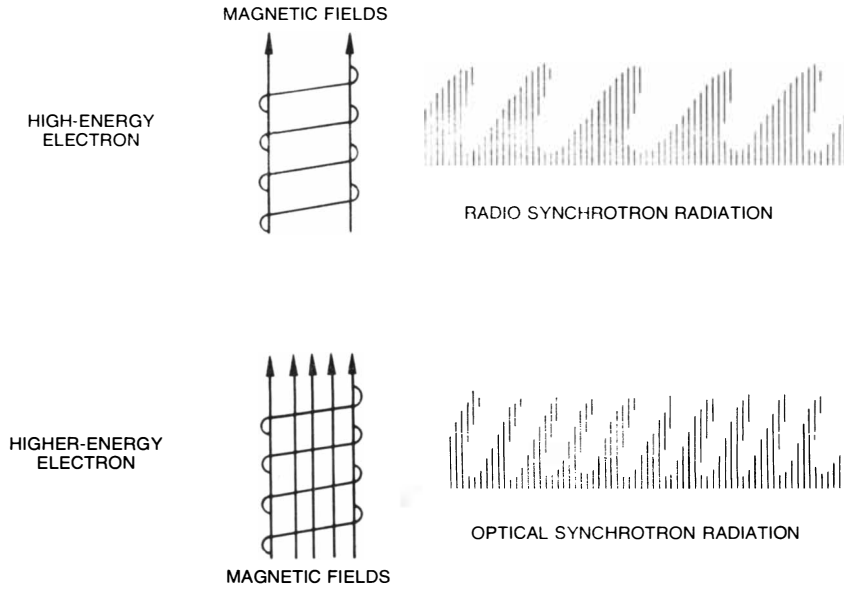


Globular ATPase molecules are lined up along two mitochondria in electron micrograph by Humberto Fernandez-Moran

found that nonastronomers are generally interested in the most speculative parts of the field." The young theorist who has written this elementary introduction to the trendiest topics in physical science today has faced that problem with discrimination, honesty and good humor, attributes that are lacking in the other books aimed at the same readers. There is not one equation, although the book does not avoid the indispensable numerical results that alone can make such a strange domain accessible. Professor Shipman's style is informal and ingratiating, with little of the detached passivity of the journal page; at the same time his aim is clarity, not exhortation. From the beginning he makes it plain that theories are "castles in the air," describing a model world and not a real one. The observations too are mere images: voltages and developed silver grains are not themselves quasars. We try to match the airy world and the imaged one; in astronomy today, unlike in the days of Newton and Halley, we claim prediction rather less than we seek understanding. The Pygmalion syndrome—infatuation with one's own model—is a classical danger and is "moderately prevalent among black hole theorists, so watch out for it." A successful revolution in science is not the usual outcome of each powerful challenge to the established view. The odds are with the conservatives. It seems now that the ingenious and attractive steady-state cosmology was one more attempt at a revolution that failed. Like history, science does not end; the few loyal partisans of such a cosmology who survive are off in some mathematical fastness of infinities and antimatter, laying their plans to capture the microwave radiation.

This approach, a candid and self-conscious discussion of method and change in science, takes little space, but it marks this book apart from the thrillmongers. The substance comes out, for example, in several summary lists that present in clear tabular form the various successes and failures—the evidence for and against certain novel ideas.

The first of the three parts concerns black holes. It begins quite helpfully with an overview of the larger ideas of stellar evolution, and of white dwarfs and neutron stars in particular. Then it treats carefully the strange model of the event horizon, aided by a long numerical table of what a distant observer and an infalling probe (or unhappy astronaut) might experience. The topic turns next to the search for real black holes, the ambiguous case of the peculiar binary Epsilon Aurigae and the more persuasive, if still open, example of Cygnus X-1. A brief account of "frontiers and fringes" follows, in an effort to part the wilder proposals of wormholes and white holes from better-founded black-hole structure theory. A skeptic might complain a little that even the Einstein field equations, well tested and elegant, are not able to yield black holes without certain quite far-reaching extrapolations.



Synchrotron radiation is explained in **Black Holes, Quasars, and the Universe**

The second portion of the text treats galaxies (a little sketchily) as a prelude to active galaxies. Much space is given to a very simple account of radio interferometry, the wonderful means of finding the radio fine structure of quasars and their kin. Quasar theories are treated with as much care as they deserve, and the rebellious view that the red shift is misleading us about quasar distances is fairly, if skeptically, presented.

The third and shortest portion of the text surveys cosmology. There is no more up-to-date popular treatment of the troubles of the old Hubble program and its present state. The microwave background radiation is given the top billing it requires, with a helpful set of graphs reviewing measurements. The book ends with a neat two pages summarizing cosmological facts, theory, informed opinion, unanswered questions and speculation, "leaving the problems of creation to the philosopher and the theologian."

Handsome diagrams adorn many pages of this well-presented book. The level of exposition, although logically high, is surprisingly elementary in terms of assumed preparation in science. Simplified sketches show a rocket that reaches—and one that does not reach—the velocity of escape, synchrotron radiation drawn as waves of differing length out of a coil-wrapped sheaf of lines of force, the blue shifts and red shifts of a diagrammatic spectrum from a single-line spectroscopic binary and so on. On the other hand, there are diagrams of the innermost galaxy, of a quasar model and of the stages of the big bang that are much more demanding.

There are a few oversights. The editors must have cut out the comparison photographs that would have made the supernova of 1939 meaningful. The text certainly plays down the enormous energy in the extended radio lobes of such quasars as 3C 273; Professor Shipman sets quasars off from radio galaxies a good deal more strongly than seems justified. The volume is a real success on its own honest terms, and it can be recommended for the general scientific reader who wants a treatment that will supplement articles in magazines or enrich and update a more formal course in astronomy at the introductory level. There is a good, if brief, annotated bibliography that includes original papers.

Are we at the edge of new physical laws? Sir Fred Hoyle asked three years back: "Do we cross a bridge into wholly unfamiliar territory?" Shipman comments: "As I have worked on this chapter, I have seen this fascinating bridge in front of me. But does it lead to the Promised Land or off a cliff?" Perhaps it leads, as it has so often before, to the qualitative modification of a splendid theory such as general relativity—but only across the stream, beyond the domain in which it has so far withstood quantitative test.

THE FOUNDATIONS OF NEWTON'S ALCHEMY, OR "THE HUNTING OF THE GREENE LYON," by Betty Jo Teeter Dobbs. Cambridge University Press ($22.50). "For no man lives that ever hath seene / Upon foure feet a *Lyon*

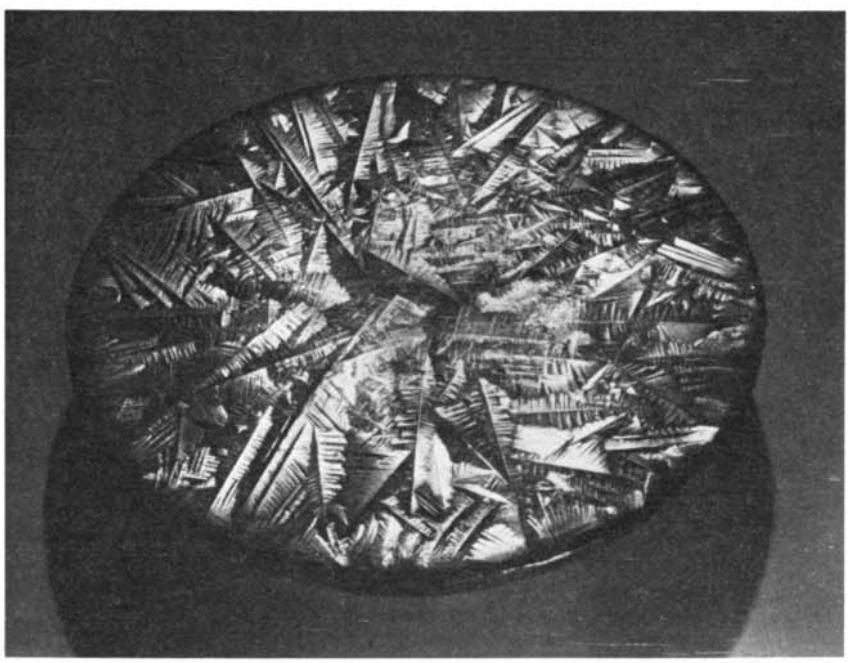
colloured *greene*." So runs a verse published by an English alchemist, a contemporary of Isaac Newton. The verse was abstracted at Trinity College by Newton sometime between his early optical publications and his great book of mathematical-mechanical principles. Newton added a couple of pages of his own notes to this rather undistinguished work of "spiritual alchemy." "The script contains ye Regimen of ye work in common gold after ye Philosophers' mercury is made." Newton himself aimed at the great work: transmutation.

This study is the most explicit examination of the bulky alchemical papers of Newton (a couple of thousand pages, almost all in his own hand) that has been reported in a long time. We know well that Newton worked seriously at transformations of matter. "About 6 weeks at spring, and 6 at the fall, the fire in the elaboratory scarcely went out.... His brick furnaces ... he made and altered himself." Professor Dobbs has made an admirable start at the refining of Newton's mind, but that great work has only just begun.

It is made pretty clear that Newton was primarily devoted not to the spiritual alchemy, which was "a way of life" for the quasi-religious adept, but rather to laboratory alchemy, which was at heart a true study of matter, if a more philosophical study than most metallurgy today. Another historian has made it plain that all Newton's studies were animated by "one overwhelming desire, to know God's will through His works in the world." Newton had good reason for seeing in himself a scholar set apart, destined to sift the knowledge of past and present. All one to him were cryptic allusions by alchemical Rosicrucians or by the Fathers of the Church, the work of a prism on sunlight, theorems on conic sections, the orbital measures of the Cassinis or the look of the crucible full of the reduced "Greene Lyon," which was the "crude and immature" ore of antimony.

The most remarkable yield of Professor Dobbs's learned and painstaking study is the evidence of Newton's successes in the toxic small-scale heavy-metal metallurgy he undertook so laboriously (and perhaps dangerously) at those brick furnaces. Where the adepts wrote of a "Philosophers' sea" with fat and scaly little fishes, and of the skillful fisherman who might take them "with a finely woven net," Newton saw a literal network formed over the surface of the frozen impure alloy of copper and antimony: "noe pit but a net work forme spread all over ye top." He made many trials and summarizes the results over a range of proportions. In another and fascinating paper (Keynes manuscript 18), which the author argues strongly is Newton's own, the alchemist went on to describe "a mercury as living and mobile as any mercury found in the world. For it makes gold begin to... spring forth into sprouts and branches, changing colors daily, the appearances of which fascinate me everyday." He had reached, he may well have thought, the philosophers' mercury, the essence of fluidity in metals, with the tail of the peacock. "I reckon this a great secret in Alchemy."

We no longer look for a seamless positivism in the old masters. Indeed, it little



"Star regulus" pattern in antimony metal, from **The Foundations of Newton's Alchemy**

becomes us who live in a world where plutonium is a potent commodity to regard transmutation as irrational or inconceivable; it is now done wholesale. All simple particulate theories of matter surely predict it. Newton, like Boyle, tried to find it. They failed only quantitatively: they did not have enough energy. Neither, of course, was free of myth, but they sought the key to the inner nature of matter eventually in the furnace, guided but never ruled by the strange allusions and metaphors of their mystical predecessors.

Here the historian has examined mainly the writings and intercourse of the 17th-century alchemists, among whose work Newton was convincingly at home. Her arguments flow from weighing the experiences of Newton as he read, copied, argued, pondered. The case seems strong: he did not flee the mystery men, as he did not flinch from the Book of Revelation. He looked to both for clues and guides. One piece of the furnace world enters this history in evidence. It is the star regulus of antimony, shown in a splendid photograph from the Science Museum in London. The remarkable dendritic network seen in this material can either suggest a radiation outward or, seen as "lines radiating in to a central point," connect directly with the ideas of particulate attraction that underlie Newton's most evident success. That success, however, lay not in the idea itself, widespread and at least as old as the works of William Gilbert and Kepler, but in Newton's magisterial brand-new quantitative exploitation of it. "Halley asked him how he knew it? Why, replied he, I have calculated it." On the other hand, he published no alchemical results, perhaps because he could find there no single proposition as clean as the dispersion of a prism or the beat of a pendulum.

It is a real debt we owe historians such as Dobbs. The word is a mirror of thought, and they carefully assay the words. But Newton is not all in words, or in metaphysics, or even in the ambient world of thought. He is revealed in other tongues, surely in mathematics but also in the daily renewed fascination of the furnace. It is not enough to mention vaguely "some sort of stable intermetallic compounds" or conclude that "now no one considers doing such an experiment." Paradigms and preconceptions are certainly important, but the actual journey to a world as pungent as the "elaboratory" cannot be slighted. Of course, we do not expect to find in the fire a mystical peacock or even the true transmuting mercury, but we may find among Newton's "sprouts and branches" more of the mind of the 17th-century founders of science than we yet know. Enough of bookstacks; Clio, to your furnaces!

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

MEDICAL MALPRACTICE

THE CITATION. Vols. 1–32. American Medical Association, 1960–1975.

PROFESSIONAL LIABILITY NEWSLETTER. Vols. 1–7. Edited by David S. Rubsamen. Berkeley, Calif., 1968–1976.

MEDICAL MALPRACTICE: REPORT OF THE SECRETARY'S COMMISSION ON MEDICAL MALPRACTICE. Department of Health, Education and Welfare, January 16, 1973.

MEDICAL MALPRACTICE LAW. Angela Roddey Holder. John Wiley & Sons, 1975.

THE SLEEP FACTOR

PERFUSION OF THE CEREBRAL VENTRICULAR SYSTEM IN UNANESTHETIZED GOATS. J. R. Pappenheimer, S. R. Heisey, E. F. Jordan and J. deC. Downer in *The American Journal of Physiology,* Vol. 203, No. 5, pages 763–774; November, 1962.

EXCHANGE OF MATERIAL BETWEEN CEREBROSPINAL FLUID AND BLOOD. J. R. Pappenheimer and S. R. Heisey in *Proceedings of the First International Pharmacological Meeting: Volume IV.* Pergamon Press, 1963.

THE IONIC COMPOSITION OF CEREBRAL EXTRACELLULAR FLUID AND ITS RELATION TO CONTROL OF BREATHING. John R. Pappenheimer in *The Harvard Lectures: Series 61.* Academic Press Inc., 1967.

CEREBRAL VENTRICULAR INFUSIONS IN UNRESTRAINED RATS. C. A. Goodrich, B. Greehey, T. B. Miller and J. R. Pappenheimer in *Journal of Applied Physiology,* Vol. 26, No. 1, pages 137–140; January, 1969.

SEPHADEX: GEL FILTRATION IN THEORY AND PRACTICE. Pharmacia Fine Chemicals, 1970.

STONE-AGE MAN ON THE NILE

NEW INVESTIGATIONS IN THE LATE PLEISTOCENE ARCHAEOLOGY OF THE KOM OMBO PLAIN (UPPER EGYPT). Philip E. L. Smith in *Quaternia,* Vol. 9, pages 141–152; 1967.

DESERT AND RIVER IN NUBIA: GEOMORPHOLOGY AND PREHISTORIC ENVIRONMENTS AT THE ASWAN RESERVOIR. Karl W. Butzer and Carl L. Hansen. The University of Wisconsin Press, 1968.

EGYPTIAN PREHISTORY: SOME NEW CONCEPTS. Fred Wendorf, Rushdi Said and Romuald Schild in *Science,* Vol. 169, No. 3951, pages 1161–1171; September 18, 1970.

HOT SPOTS ON THE EARTH'S SURFACE

PHYSICS AND GEOLOGY. J. A. Jacobs, R. D. Russell and J. Tuzo Wilson. McGraw-Hill Book Company, Inc., 1959.

IS THE AFRICAN PLATE STATIONARY? K. Burke and J. T. Wilson in *Nature,* Vol. 239, No. 5372, pages 387–390; October 13, 1972.

TWO TYPES OF MOUNTAIN. J. Tuzo Wilson and Kevin Burke in *Nature,* Vol. 239, No. 5373, pages 448–449; October 20, 1972.

PLUME GENERATED TRIPLE JUNCTIONS: KEY INDICATORS IN APPLYING PLATE TECTONICS TO OLD ROCKS. Kevin Burke and J. F. Dewey in *The Journal of Geology,* Vol. 81, No. 4, pages 406–433; July, 1973.

RELATIVE AND LATITUDINAL MOTION OF ATLANTIC HOT SPOTS. Kevin Burke, W. S. F. Kidd and J. Tuzo Wilson in *Nature,* Vol. 245, No. 5421, pages 133–137; September 21, 1973.

RABBIT HEMOGLOBIN FROM FROG EGGS

RABBIT HAEMOGLOBIN SYNTHESIS IN FROG CELLS: THE TRANSLATION OF RETICULOCYTE 9 S RNA IN FROG OOCYTES. C. D. Lane, G. Marbaix and J. B. Gurdon in *Journal of Molecular Biology,* Vol. 61, No. 1, pages 73–91; October 14, 1971.

THE CONTROL OF GENE EXPRESSION IN ANIMAL DEVELOPMENT. J. B. Gurdon. The Clarendon Press, Oxford University Press, 1974.

MOLECULAR BIOLOGY IN A LIVING CELL. J. B. Gurdon in *Nature,* Vol. 248, No. 5451, pages 772–776; April 26, 1974.

THE INJECTION OF RNA INTO LIVING CELLS: THE USE OF FROG OOCYTES FOR THE ASSAY OF MRNA AND THE STUDY OF THE CONTROL OF GENE EXPRESSION. Charles D. Lane and John Knowland in *The Biochemistry of Animal Development: Vol. III,* edited by Rudolf Weber. Academic Press, 1975.

THE PHOTOGRAPHIC LENS

APPLIED OPTICS AND OPTICAL DESIGN: PART TWO. A. E. Conrady. Edited by Rudolf Kingslake. Dover Publications, Inc., 1960.

PHOTOGRAPHY: ITS MATERIALS AND PROCESSES. Edited by C. B. Neblette. D. Van Nostrand Company, Inc., 1962.

OPTICAL COMPONENTS. Edited by Rudolf Kingslake in *Applied Optics and Optical Engineering: Vol. III.* Academic Press, 1965.

MODERN OPTICAL ENGINEERING: THE DESIGN OF OPTICAL SYSTEMS. Warren J. Smith. McGraw-Hill Book Company, 1966.

SPSE HANDBOOK OF PHOTOGRAPHIC SCIENCE AND ENGINEERING. Edited by Woodlief Thomas, Jr. John Wiley & Sons, 1973.

THE SOCIAL BEHAVIOR OF BURYING BEETLES

LES GRANDS NÉCROPHAGES DU GLOBE: SILPHINI-NECRODINI-NECROPHORINI. G. Portevin in *Encyclopedie Entomologique VI.* Paul Lechevalier, 1926.

ÖKOLOGISCHE UNTERSUCHEN AN NECROPHORUS F. Erna Pukowski in *Zeitschrift für Morphologie und Ökologie der Tiere,* Vol. 27, pages 518–586; 1933.

THE INSECT SOCIETIES. Edward O. Wilson. Harvard University Press, 1971.

DIEL PERIODICITIES OF CERTAIN CARRION BEETLES (COLEOPTERA: SILPHIDAE). Paul P. Shubeck in *The Coleopterists Bulletin,* Vol. 25, No. 2, pages 41–46; June, 1971.

BEITRAG ZUR KENNTNIS DER STRIDULATIONSAPPARATE EINHEIMISCHER NECROPHORUS-ARTEN (*NECROPHORUS HUMATOR* Ol., *Necrophorus investigator* Zetterst., *NECROPHORUS VESPILLOIDES* Herbst) (Insecta, Coleoptera). Rolf Schumacher in *Zeitschrift für Morphologie der Tiere,* Vol. 75, pages 65–75; 1973.

THE CURVATURE OF SPACE IN A FINITE UNIVERSE

EINSTEIN'S GENERAL THEORY OF RELATIVITY. Max Born in *Einstein's Theory of Relativity.* Dover Publications, Inc., 1962.

PARADISE. Dante Alighieri, translated by Dorothy L. Sayers and Barbara Reynolds in *The Comedy of Dante Alighieri.* Penguin Books, 1962.

MATHEMATICAL THOUGHT FROM ANCIENT TO MODERN TIMES. Morris Kline. Oxford University Press, 1972.

GRAVITATION. Charles W. Misner, Kip S. Thorne and John Archibald Wheeler. W. H. Freeman and Company, 1973.

MATHEMATICAL GAMES

THE HISTORY OF THE UNITED STATES FLAG: FROM THE REVOLUTION TO THE PRESENT, INCLUDING A GUIDE TO ITS USE AND DISPLAY. Milo M. Quaife, Melvin J. Weig and Roy E. Appleman. Harper and Brothers, Publishers, 1961.

THE ORCHARD PROBLEM. Stefan A. Burr, Branko Grünbaum and N. J. A. Sloane in *Geometriae Dedicata,* Vol. 2, No. 4, pages 397–424; February, 1974.

You always knew there could be cars like this.

You always knew that there could be cars that deliver luxury, engineering, styling and performance. Now your long wait is over. Lancia Beta. Seventy years of Lancia racing and car building experience culminating in a superb driving experience. It's your intelligent alternative to the average and the over-priced automobile.

LANCIA

The intelligent alternative.

For the name of your nearest Lancia dealer call toll free (800) 447-4700 or in Illinois (800) 322-4400.

Overseas delivery, leasing and daily rental arranged through your participating dealer.
Lancia of America. Division of Fiat Motors of North America, Inc., Montvale, New Jersey 07645

Never buy a white Bordeaux by the bottle. Buy a crisp, white Bordeaux by the label.

B&G.
The B&G label.
250 years of wine making history from all the great wine regions of France: Bordeaux, The Loire Valley, Beaujolais, Côtes du Rhône and Burgundy.

Now, B&G brings you Pontet-Latour, the crisp white wine from Bordeaux.

Elegant. Affordable. Buy it by the label.

B&G.

Letters of recommendation from France