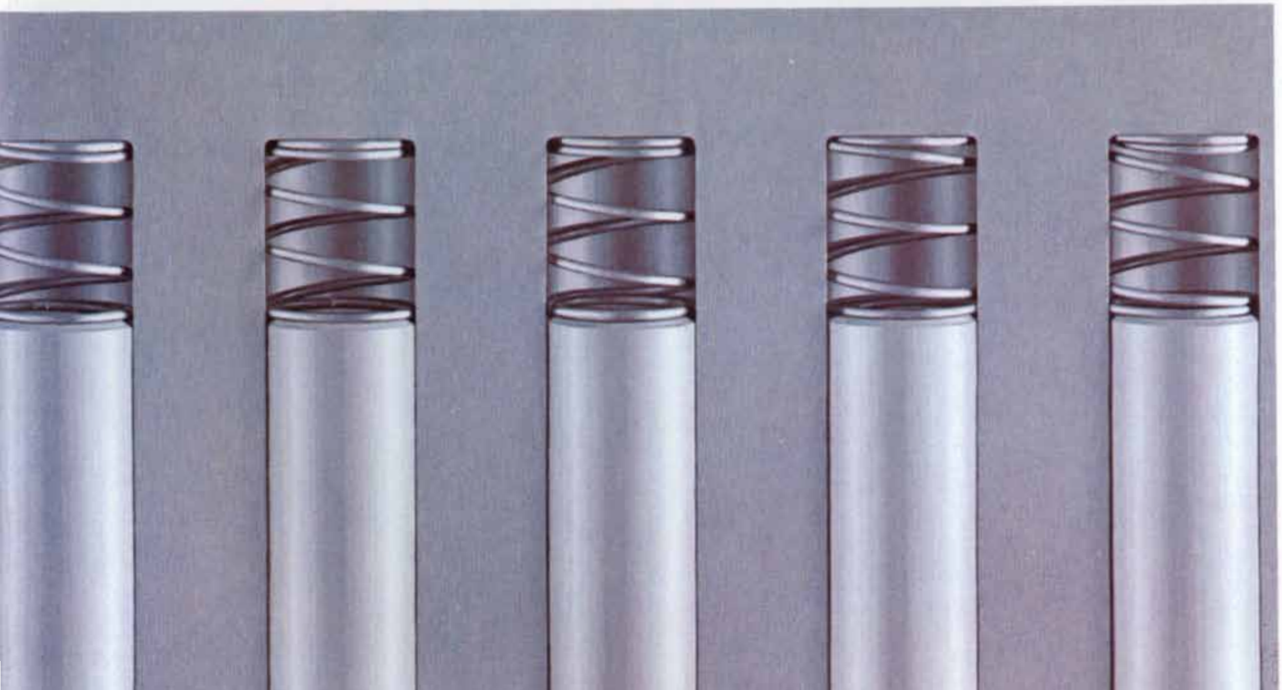
SCIENTIFIC AMERICAN



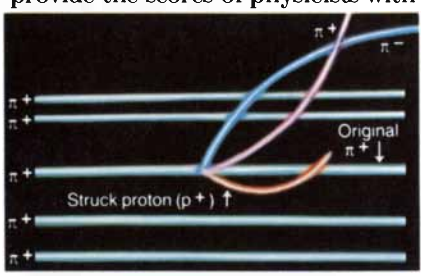
MICROELECTRONIC PACKAGING

$2.50

July 1983

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

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



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

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

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

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

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

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

IBM®


"When your traveling companion is worth $2.5 million you better be sure about your transportation."
BOB NORRIS, AMERICAN QUARTER HORSE ASSOCIATION

Established 1845 **SCIENTIFIC AMERICAN** July 1983 Volume 249 Number 1

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/83. $2.00 + 0.00.

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

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

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

Name

New Address

Street

City

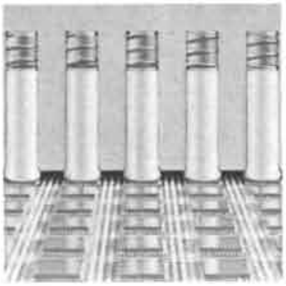
State and ZIP

Old Address

Street

City

State and ZIP



THE COVER

The columns rising above a mosaic floor in the painting on the cover represent part of a cooling module that houses more than 100 silicon chips in the central processing unit of a large digital computer. The chips are mounted face down on a ceramic substrate that supplies some 12,000 terminals for power and signals. The columns, or pistons, are pressed against the back of the chips to carry off heat. In many computers the technology of interconnecting and cooling microelectronic devices is the major factor limiting the performance of the central processor (see "Microelectronic Packaging," by Albert J. Blodgett, Jr., page 86). The module is made by the International Business Machines Corporation.

THE ILLUSTRATIONS

Cover painting by Marvin Mattelson

Page	Source
14–20	William S. Huff, State University of New York at Buffalo
34	National Aeronautics and Space Administration, Toby Carlson, Pennsylvania State University, and Annabelle Carlson
35–41	Albert E. Miller
42	U.S. Geological Survey, EROS Data Center
44	Andrew Tomko
45	Andrew Tomko (*top*); Vincent Courtillot, University of Paris (*bottom*)
46–49	Andrew Tomko
50	Weststock
52	Enid Kotschnig
54–56	Sally Black
57	University of Washington
58	Lauren R. Donaldson, University of Washington College of Ocean and Fishery Sciences
71	Pierre A. Henkart, National Cancer Institute (*top*); Alan M. Kuzirian, Marine Biological Laboratory, Woods Hole, Mass. (*bottom*)
72	Jerome Kuhl
73–74	Daniel L. Alkon, Marine Biological Laboratory, Woods Hole, Mass. (*top*); Jerome Kuhl (*bottom*)
76–84	Jerome Kuhl
87	International Business Machines Corporation
88–90	Gabor Kiss
91	International Business Machines Corporation
92–94	Gabor Kiss
95	International Business Machines Corporation (*top*), Gabor Kiss (*bottom*)
96	International Business Machines Corporation
99	© Walter Chandoha
100	Ilil Arbel
101–102	Donald R. Kaplan, University of California at Berkeley
104	Ilil Arbel
105	Donald R. Kaplan, University of California at Berkeley
107	Allen Beechel
108	Edward Bell
109–115	Allen Beechel
117	Collection of Ray Jerome Baker and Robert E. Van Dyke
118	Ilil Arbel
119–122	Edward Bell
124–128	Michael Goodman

LETTERS

Sirs:

"Hidden Visual Processes," by Jeremy M. Wolfe [SCIENTIFIC AMERICAN, February], is a provocative challenge to traditional theories of visual perception and a refreshing alternative to commonplace conceptions of human perceptual abilities. The article also raises epistemological issues that deserve further attention. In his effort to show that the human visual system is composed of a number of specialized subsystems acting coordinately to effect our awareness of the world, he argues that "no amount of introspection can make us aware of the subsystems themselves."

Wolfe's experiments are ingenious. His conclusions regarding their cognitive significance are questionable. In his discussion of one typical procedure, that involving "accommodation" (the ability to control the muscles focusing the eye on objects at various distances by changing the shape of the lens) he concludes that since this ability is notably independent of color perception, its discovery is "a finding that no amount of introspection could have suggested about vision." His detailed account of the experimental procedures, however, contains numerous (and unavoidable) references to what the subject saw or thought he saw. For example: "Thus our subjects could readily tell that the isoluminant stimulus was an edge between red and green, but it was impossible for them to bring the edge into focus." How are we to suppose this information was obtained if not by asking the subjects to report their occurrent visual experiences, that is, to "introspect"? These references do not in any way undermine the empirical significance of Wolfe's research, but they do underscore the inconsistency of his methodical repudiation of introspection and his repeated appeal to what are essentially introspective reports of visual experiences *under experimentally directed conditions.* The fact that some perceptual processes may be undetectable under normal or unrestricted conditions is neither evidence nor argument for the assumption that their detection requires a bypass of introspective awareness. On the contrary, the detection of such processes is both psychologically and logically dependent on the ability to produce just such *anomalous* experiences as Wolfe's experiments elicit. Introspection has not been eliminated; it has merely been redirected.

Methodologically, Wolfe's research exemplifies commitments that are common enough, even orthodox in experimental psychology. It is all the more ironic, then, that this discipline, which has been systematically intent on excluding appeal to anything as "subjective" and unquantifiable as introspection, should rely so consistently, if covertly, on the introspective judgments of subjects whose reports are regarded as neutrally objective data when they are produced under the controlled conditions prescribed by the investigators.

ALAN TORMEY

University of Maryland
Baltimore County
Catonsville

Sirs:

I was impressed by the experiments Jeremy M. Wolfe described. I do not, however, agree with his conclusions. At the beginning of his article he is referring to our work on residual vision in brain-injured patients (Ernst C. Pöppel, Richard Held and Douglas Frost in *Nature,* Vol. 243, No. 5405, pages 295–296; June 1, 1973). He writes that we "have found that when people who are perceptually blind because of an injury to the cerebral cortex are asked to direct their eyes toward a spot of light, they do surprisingly well. The subjects report that they cannot see the spot, and so they think they are guessing, but they look in roughly the right direction more often than chance would allow. "Brain-injured people thus show evidence of multiple visual processes: some that are damaged, and hidden ones that remain functional." Although the description of our experimental evidence is correct. I cannot accept Wolfe's conclusion on "multiple visual processes," in the light of his final remark that the visual sense is actually many senses and "that the full range of human visual senses remains to be discovered."

I still believe that there is only *one* visual sense and that what Wolfe is talking about—also in the case where he is referring to our results—are not different visual senses but operations within the one sense that enable the visual system to do the one thing it has been developed for: to recognize objects in the visual world.... A theory of object recognition putting service operations on a lower level is quite different from a theory that defines vision as activity of many visual senses.

ERNST PÖPPEL

Ludwig-Maximilians-Universität
Munich

Sirs:

Alan Tormey raises a common but, I think, incorrect objection to the idea of "hidden" visual processes. He notes that the data are obtained by asking the subject what he or she sees and concludes that since the subject can answer, the processes under study are accessible to conscious visual perception. What the subject "sees" in these experiments and what the experimenter "infers" are two very different things, however. Consider the accommodation experiment. The subject reports on the alignment of two line segments. We, of course, are not interested in the ability to perceive alignment. We use the information to measure the response of the accommodative system, how well the subject was focusing. If we asked directly, "At what distance are you focused?" the subject could not give us a meaningful answer in this experiment.

Our situation is similar to that of the physician. If you have an abdominal pain, the doctor does not ask you which of your various internal organs is causing the problem. Instead you are asked where it hurts, and from this information the doctor can try to diagnose the underlying cause. The apparent locus of the pain may be some distance from the site of the problem. No matter; the doctor is not interested in your experience of pain but only in what it can reveal about the state of your health.

In the experiments described in my article we used two different line-alignment tasks and a measure of apparent self-motion. From this we inferred that the accommodative system is colorblind and that there are at least three different binocular processes in human vision. The specific tasks we used were irrelevant to our conclusions. The conclusions cannot be derived from conscious experience either in or out of the experiment. This is so even though the conclusions are based on interpretations of what our subjects report seeing.

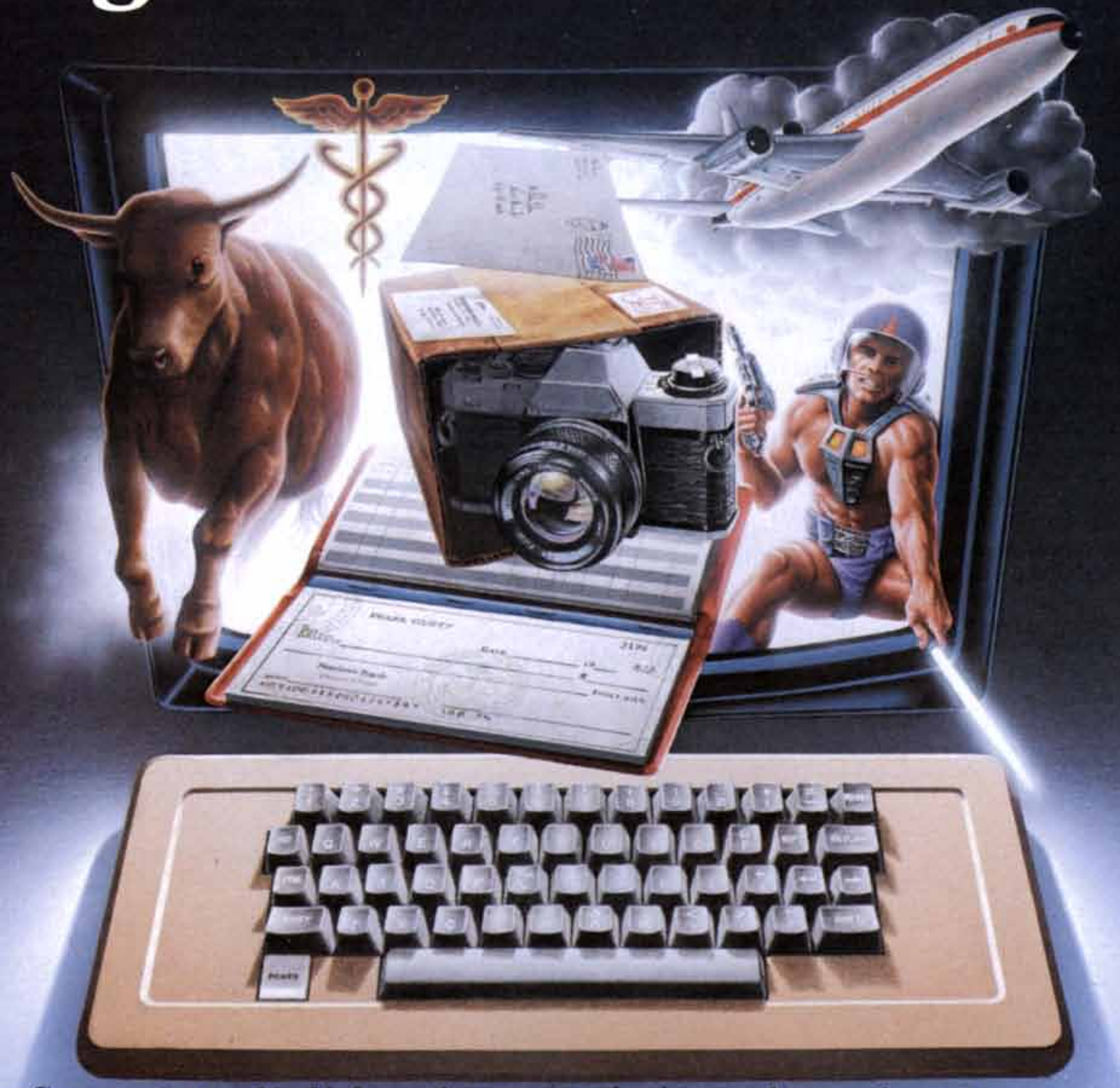
Ernst Pöppel is, I think, correct when he makes object recognition the "primary task" of the visual system. As he says, under normal circumstances functions such as accommodation serve to make the task easier. Nevertheless, the subservient process may still be treated as separate, self-contained and "hidden" when it is studied in the laboratory. The accommodative process is responsive to certain stimuli and gives rise to certain responses. In the laboratory it can be studied in the way we would study the vision of an animal. We present stimuli and make inferences on the basis of a motor output. I would not question Dr. Pöppel's main point, that there is a unified visual sense. I would suggest that the unity of vision may be like the unity of an army. At one level it moves with a single purpose. On more detailed examination it is composed of a number of autonomous parts acting together.

JEREMY M. WOLFE

Department of Psychology
Massachusetts Institute of Technology
Cambridge


PAN AM
VICTOR

We don't care which computer you buy. We'll help you get the most out of it.



CompuServe puts a world of information, communications, and entertainment at your fingertips.

CompuServe is the easy to use videotex service designed for the personal computer user and managed by the communications professionals who provide business information services to over one fourth of the FORTUNE 500 companies.

Subscribers get a wealth of useful, profitable, or just plain interesting information like national news wires, electronic banking and shop at home services, and sophisticated financial data. Plus, a communications network for electronic mail, a bulletin board for selling, swapping, and personal notices and a multi-channel CB simulator.

You get games on CompuServe, too. Classic puzzlers, educational, sports and adventure games and fantastic space games featuring MegaWars, the "ultimate computer conflict."

To learn more about CompuServe, call toll-free, 800-848-8199, for an illustrated guide to the CompuServe Information Service. The videotex service for you, no matter which computer you buy.

CompuServe

Consumer Information Service
2180 Wilson Rd., Columbus, Ohio 43228
800-848-8199 In Ohio Call 614-457-8650
An H&R Block Company

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1933: "Professor R. A. Millikan's suggestive theory of the origin of cosmic rays is based on the assumption that the rays are electromagnetic radiations, or photons, similar in type to X rays and gamma rays but of shorter wavelength. About five years ago two German physicists, Walter Bothe and Werner Kolhörster, did an experiment with counting tubes that convinced them the cosmic rays are electrically charged particles. If this conclusion is correct, it means there should be a difference in the intensity of the rays over different parts of the globe. The earth acts as a huge magnet, which should deflect the electrified particles as they shoot toward it. The effect should be least near the magnetic poles and greatest near the Equator, resulting in an increasing intensity from the Equator toward the poles. A group led by Arthur H. Compton of the University of Chicago has organized nine different expeditions over the past 18 months, going to different parts of the globe to measure cosmic rays. When the results of these expeditions are brought together, it is found that the cosmic-ray intensity near the poles is about 15 per cent greater than that near the Equator. These results show that a considerable part, at least, of the cosmic rays consists of electrified particles."

"Invisible 'black light' was described by Samuel G. Hibben of the Westinghouse Electric and Manufacturing Company at a recent meeting of the Illuminating Engineering Society in New York City. 'The "black light,"' said Mr. Hibben, 'is ultra-violet radiation, and it is 99 per cent free of visible light. It is produced by two new black-bulb lamps, one consuming two amperes, the other five. They are made of a special cobalt glass. The bulbs of these lamps absorb 99 per cent of the visible light and transmit 80 to 85 per cent of the ultra-violet wanted. The radiations are relatively long in wavelength, in the range of 3,200 to 4,000 angstrom units. This long-wave ultra-violet lends itself to many photographic and fluorescent effects.'"

"From the time the *Graf Zeppelin* was first placed in service in 1928 until the end of last year the airship made 290 voyages, including 33 ocean crossings, traveling 330,000 miles in 5,370 hours, at an average speed of 61 miles per hour, and carrying almost 17,000 passengers, 35,000 pounds of mail and 80,000 pounds of freight, without accident or mishap, a remarkable demonstration of safety and dependability. A number of improvements are incorporated in a new leviathan of the air that is now nearing completion in Germany and that is to supplement the South American service of the *Graf Zeppelin*. The airship and the airplane differ in that the airplane is primarily a fast, comparatively short-radius craft, whereas the airship is slower and reaches full efficiency only on long voyages, particularly across oceans. In the aviation of the future there is every reason to believe that the airplane will find its place as the carrier for relatively short distances—even making contact with the dirigible, as has already been successfully accomplished during the regular Germany-Brazil service of the *Graf Zeppelin*—and that the dirigible will be the vehicle of transcontinental and intercontinental travel."

"Former industrial workers, storekeepers, barbers, clerks, electricians, miners—these and other unemployed men from many other walks of life are the direct beneficiaries of the new Federal reforestation program. A quarter of a million men will be paid for their reforestation work, will be assured their subsistence for the next few months and will get the healthy, outdoor work that will help to make them physically fit and restore their mental equilibrium weakened by privation and despair. No less direct will be the benefit to the entire American people."

SCIENTIFIC AMERICAN

JULY, 1883: "Nearly all the petroleum that goes into the world's commerce is produced in a district about 150 miles long, with a varying breadth of from one mile to 20 miles, lying mainly in the state of Pennsylvania but lapping over a little on its northern edge into the state of New York. This region yielded, in 1882, 31,398,750 barrels. A little petroleum is obtained in West Virginia, a little at isolated points in Ohio and a little in the Canadian province of Ontario. There is also a small field in Germany, a larger one, scantily developed, in southern Russia and one still larger, perhaps, in India. The total production of all the fields, outside of the region here described, is but a small fraction in the general account. Furthermore, the oil of these minor fields, whether in America or the Old World, is of inferior quality."

"The Westinghouse brake is fitted upon the engine, tender and each vehicle of the train. Air compressed by a pump on the locomotive to, say, 70 lb. or 80 lb. to the square inch fills the main reservoir on the engine, and flowing through the driver's brake valve and main pipe it also charges the supplementary reservoirs throughout the train. When the train is running, uniform air pressure exists throughout its length, that is to say, the main reservoir on the engine, the pipe from one end of the train to the other, the valves and supplementary reservoirs on each vehicle are all charged ready for work, the brake cylinders being empty and the brakes off. The essential principle of the system is that maintaining the pressure keeps the brakes off, but letting the air escape from the brake pipe, on purpose or by accident, instantly applies them. It follows, therefore, that the brake may be applied by the driver or any of the guards, or if necessary by a passenger, by the separation of a coupling, or the failure of or injury to a vital part of the apparatus; since the brake on each vehicle is complete in itself and independent, should the apparatus on any one car be torn off, the brakes will nevertheless remain applied for almost any length of time in the rest of the train."

"It is an old and probably a true saying that every element could be detected everywhere had we sufficiently delicate tests for it. Early observations had prepared William Crookes, F.R.S., for the wide distribution of yttrium, and no sooner had Mr. Crookes developed his method of radiant-matter spectroscopy than the exquisite sensitiveness of this spectrum test forced itself on his notice when he sought for yttrium in various minerals. Pink coral contains one part of yttrium in 200 parts; strontianite, one part of yttrium in 500 parts; chondrodite from Mount Somma, one part in 4,000; calcite, one part in 10,000; ox bone, one part in 10,000; an earthy meteorite (Alfianello), one part in 100,000, and tobacco ash, one part in 1,000,000."

"Should women ride like men? Although it may not appear to be the case, the seat a woman enjoys on a side-saddle is fully as secure, and not nearly as irksome, as that which a man has to maintain, unless he simply balances himself and does not gripe the sides of his horse with either the knee or the side of the leg. It is curious to note the different ways in which the legs of men who pass much time in the saddle are affected. Riding with a straight leg and a long stirrup almost invariably produces what are popularly called knocked knees. On the other hand, riding with a short stirrup produces bowed legs. No deformity necessarily follows the use of the side-saddle if the precaution is taken with growing girls to change sides on alternate days, riding on the left side one day and on the right the next."

THE AUTHORS

LARRY LONG and DIANA DEARE ("The Slowing of Urbanization in the U.S.") are members of the staff of the U.S. Bureau of the Census. They are utilizing the results of the census of 1980, among other data, to study the distribution of population and economic activity in the U.S. Long is at the bureau's Center for Demographic Studies. He received his undergraduate education at the University of Texas at Austin, going on to obtain his Ph.D. in sociology there in 1969. He did postdoctoral work at the University of Pennsylvania before moving to the Census Bureau while the results of the census of 1970 were being tabulated. DeAre is with the bureau's Population Division. She got her B.A. in 1971 at DePaul University. Her M.A. (1973) and Ph.D. in geography (1976) are from the University of Texas at Austin. She joined the bureau in 1975.

VINCENT COURTILLOT and GREGORY E. VINK ("How Continents Break Up") are geophysicists who did independent work on the subject of their article before collaborating on writing it. Courtillot is professor of geophysics and chairman of the department of earth sciences at the University of Paris. He attended the Paris School of Mines as an undergraduate. He came to the U.S. to pursue his studies, earning his M.S. in geophysics at Stanford University in 1972 before returning to France; his Ph.D. in geophysics was awarded in 1977 by the University of Paris. Courtillot was a member of the recent French-Chinese geological expedition to Tibet. Vink is a graduate student at Princeton University. He received his B.A. in 1979 from Colgate University. He is writing a dissertation on the plate-tectonic evolution of the Norwegian-Greenland Sea and the Arctic Ocean.

LAUREN R. DONALDSON and TIMOTHY JOYNER ("The Salmonid Fishes as a Natural Livestock") are experts in fisheries. Donaldson is professor emeritus of fisheries at the University of Washington. He got his B.S. from Intermountain Union College. He obtained his M.S. and Ph.D. from the University of Washington before joining the faculty of the College of Ocean and Fishery Sciences there in 1932. He has made a significant contribution to the large fisheries program at Washington. He developed a brood stock of rainbow trout with rates of growth, survival and egg production more than 20 times those of previous trout stock. He directed the work that established runs of salmon and steelhead trout to the university's holding ponds. Donaldson served for many years as director of the Laboratory of Radiation Biology at Washington, which evaluates the biological effects of radioactivity. Joyner earned his Ph.D. in fisheries in 1962, partly for work done at the Laboratory of Radiation Biology. He went to Brown University as an undergraduate before serving in the Marine Corps in the Pacific in World War II. After getting his doctorate he went to work for the National Marine Fisheries Service (then the Bureau of Commercial Fisheries). He retired in 1975 and is now a writer and consultant on fisheries. Joyner writes that he is "following with great interest a project I helped to start while I was with the National Marine Fisheries Service: rearing selected stocks of Atlantic salmon in saltwater pens in Puget Sound to help provide eggs for the restocking of rivers in New England."

DANIEL L. ALKON ("Learning in a Marine Snail") is research medical officer and chief of the section on neural systems at the National Institute of Neurologic and Communicative Disorders and Stroke. He is also on the staff of the Marine Biological Laboratory in Woods Hole, Mass.; his laboratory is at Woods Hole. His B.S. (1965) is from the University of Pennsylvania. His M.D. (1969) is from the Cornell University Medical College. After completing his internship he went to the National Institutes of Health. Alkon's research interests include neuropharmacology, sensory physiology and animal behavior.

ALBERT J. BLODGETT, JR. ("Microelectronic Packaging"), is manager of the manufacturing research laboratory at the Thomas J. Watson Research Center of the International Business Machines Corporation. He was graduated from Yale University with a B.E. in electrical engineering in 1960. In the same year he moved to the product-development laboratory at IBM. Two years later he left for Stanford University on an IBM resident fellowship; at Stanford he earned his M.S. and Ph.D. degrees in electrical engineering before returning to IBM in 1966. Among other projects, Blodgett was responsible for the design of the solid-logic-technology modules that are the basic logic devices in the S/360 data processor and the development of the multilayered ceramic modules utilized in the 43XX and 308X data-processor units.

DONALD R. KAPLAN ("The Development of Palm Leaves") is professor of botany at the University of California at Berkeley. He got his A.B. in

1960 at Northwestern University. His Ph.D. in botany, granted in 1965, is from Berkeley. In 1965 he served as a National Science Foundation fellow at the Royal Botanical Garden at Kew in England. He returned to the U.S. to take up a job in the department of organismic biology at the University of California at Irvine. He went to Berkeley in 1968. In addition to Kaplan's botanical interests he is a free-lance writer on (and photographer of) railroad trains.

NARIMAN B. MISTRY, RONALD A. POLING and EDWARD H. THORNDIKE ("Particles with Naked Beauty") are members of a group of 75 physicists from seven universities who are investigating the physics of beauty-flavored mesons with the detector named CLEO at the Cornell Electron Storage Ring (CESR). Mistry is senior physicist at the Newman Laboratory of Nuclear Studies at Cornell University. A native of India, he received his undergraduate education at St. Xavier's College in Bombay. He came to the U.S. in 1958 to continue his training in physics; Columbia University awarded him a Ph.D. in 1963. He has been at Cornell since then. Poling is research associate in physics and astronomy at the University of Rochester. His B.S. (1976) is from the State University of New York at Buffalo and his Ph.D. (1981) is from Rochester. Thorndike is professor of physics at Rochester. He was graduated from Wesleyan University in 1956 with an A.B. and went on to earn his M.S. at Stanford University in 1957. His Ph.D. in physics (1960) is from Harvard University. He has served as spokesman of the CLEO collaboration since 1981.

DEREK BICKERTON ("Creole Languages") writes: "I acquired my B.A. from the University of Cambridge in 1949 but did not start my academic career until 1964, when I taught English literature at the University College of Cape Coast in Ghana. In 1966 and 1967 I studied linguistics at the University of Leeds before becoming senior lecturer in the English language at the University of Guyana, where I remained for four years. I taught briefly at the University of Lancaster in England and joined the University of Hawaii at Manoa in the fall of 1972. In the British system it isn't necessary to have a doctorate but in the American it is, and so I took advantage of a regulation that enables graduates of Cambridge to submit published work instead of a regular dissertation. The work I submitted was *Dynamics of a Creole System* (Cambridge University Press, 1975). I obtained my doctorate in 1976. I have held the rank of full professor at Hawaii since the same year. I have also written several novels, the most recent of which was *King of the Sea* (Random House, 1980)."

METAMAGICAL THEMAS

Parquet deformations: patterns of tiles that shift gradually in one dimension

by Douglas R. Hofstadter

What is the difference between music and visual art? If someone asked me this question, I would have no hesitation in responding. To me the major difference is temporality. Works of music intrinsically involve time; works of visual art do not. More precisely, pieces of music consist of sounds intended to be played and heard in a specific order and at a specific speed. Music is therefore fundamentally one-dimensional; it is tied to the rhythms of our existence. Works of visual art, in contrast, are generally two- or three-dimensional. Paintings and sculpture seldom have any intrinsic "scanning order" built into them that the eye must follow. Mobiles and other pieces of kinetic art may change over time, but often without any specific initial state or final state or intermediate states. You are free to come and go as you please.

There are, of course, exceptions to this generalization. European art has grand friezes and historic cycloramas, and Oriental art has intricate pastoral scrolls up to hundreds of feet long. These types of visual art impose a temporal order and speed on the scanning eye. There is a starting point and a final point. Usually, as in stories, these points represent states of relative calm, particularly at the end. In between, various types of tension are built up and resolved in an idiosyncratic but pleasing visual rhythm. The calmer end states are usually orderly and visually simple, whereas the tenser intermediate states are usually more chaotic and visually confusing. If you replace "visual" by "aural," virtually the same can be said of music.

I have been fascinated for many years by the idea of trying to capture the essence of the musical experience in visual form. I have my own ideas about how this can be done; in fact, I spent several years working out a form of visual music. By no means, however, do I think there is a unique or best way to carry out this task of "translation," and indeed I have often wondered how others might attempt to do it. I have seen a few such attempts, but most of them struck me as being unsuccessful. One striking counterexample is the set of "parquet deformations" meta-composed by William S. Huff, professor of architectural design at the State University of New York at Buffalo.

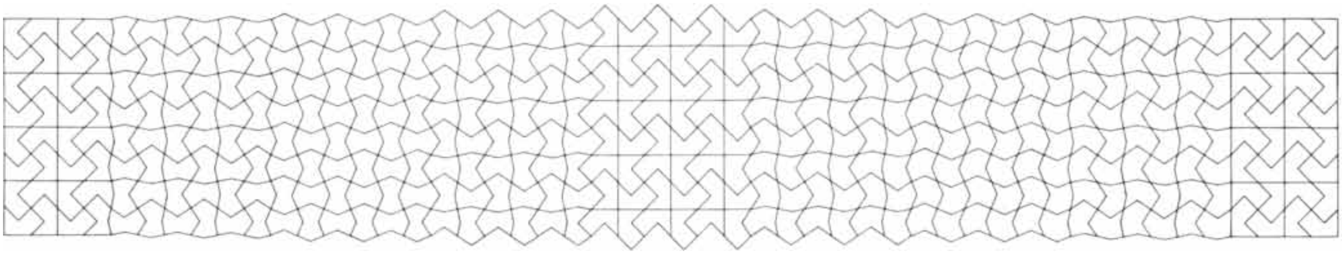
I say "meta-composed" for good reason. Huff himself has never executed a single parquet deformation. He has elicited hundreds of them, however, from his students, and in so doing he has brought this form of art to a high degree of refinement. He might be likened to the conductor of a fine orchestra. Although the conductor makes no sound in the course of a performance, we give much credit to the person doing the job for the quality of the sound. We can only guess at how much preparation and coaching went into the performance.

So it is with William Huff. For 20 years his students in Buffalo and also at Carnegie-Mellon University have been prodded into flights of artistic inspiration, and it is thanks to Huff's vision of what constitutes quality that some beautiful results have emerged. Not only has he elicited outstanding work from students; he has also carefully selected what he thinks are the best pieces. These he is keeping in archives. For these reasons I shall be referring to "Huff's creations" but always in this more indirect sense of meta-creations.

Not to take credit from the students who executed the individual pieces, there is a larger sense of the term "credit" that goes exclusively to Huff, the person who has shaped this entire art form himself. Let me use an analogy. Gazelles are marvelous beasts, yet it is not they themselves but the selective pressures of evolution that are responsible for their species' unique and wondrous qualities. Huff's judgments and comments have here played the role of those impersonal evolutionary selective pressures, and out of them has been molded a living and dynamic tradition, a "species" of art exemplified and extended by each new instance.

All that remains to be said by way of introduction is the meaning of "parquet deformation." Actually it is nearly self-explanatory. Traditionally a parquet is a regular mosaic made out of inlaid wood on the floor of an elegant room, and a deformation is a deformation. Huff's parquets are more abstract: they are regular tessellations (or tilings) of the plane, ideally drawn with zero-thickness line segments and curves. The deformations are not arbitrary but must satisfy two basic requirements: (1) there must be change only in one dimension, so that it is possible to see a temporal progression in which one tessellation gradually becomes another, and (2) at each stage the pattern must constitute a regular tessellation of the plane, that is, there must be a unit cell that could combine with itself so that it could cover an infinite plane exactly.

From this very simple idea emerge some stunningly beautiful creations. Huff explains that he was originally inspired, back in 1960, by the M. C. Escher woodcut "Day and Night." In that work forms of birds tiling the plane are gradually distorted (as the eye scans downward) until they become diamond-shaped, looking like the checkerboard pattern of cultivated fields seen from the air. Escher of course became famous for



A parquet deformation titled "Fylfot Flipflop" his tessellations, both pure and distorted, as well as for the other haunting games he played with art and reality.

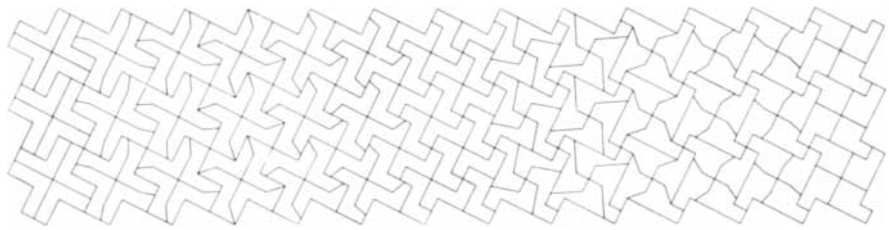
Whereas Escher's tessellations are almost always based on animal forms, Huff decided to limit his scope to purely geometric forms. In a way that is like a decision by a composer to follow austere musical patterns and to totally eschew anything that might conjure up a "program" (that is, some kind of image or story behind the sounds). An effect of this decision is that the beauty and visual interest must come entirely from the complexity and subtlety of the interplay of abstract forms. There is nothing to "charm" the eye, as there is with pictures of animals. There is only the unembellished perceptual experience.

Because of the linearity of this form of art, Huff has likened it to visual music. He writes: "Although I am spectacularly ignorant of music, tone-deaf and hated those piano lessons (yet can be enthralled by Bach, Vivaldi or Debussy), I have the students 'read' their designs as I suppose a musician might scan a work: the themes, the events, the intervals, the number of steps from one event to another, the rhythms, the repetitions (which can be destructive, if not totally controlled, as well as reinforcing). These are principally temporal, not spatial, compositions (although all predominantly temporal compositions have, of necessity, an element of the spatial and vice versa—e.g., the single-frame picture is the basic element of the moving picture)."

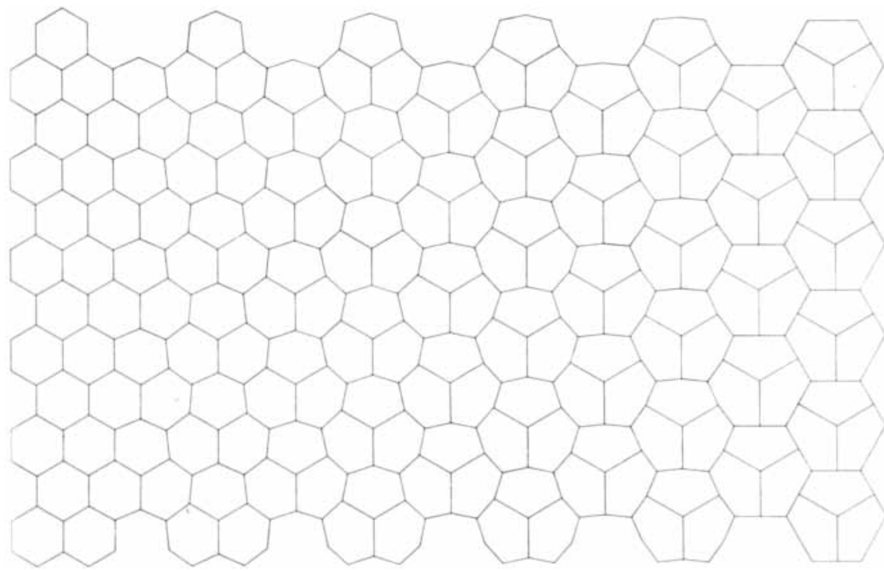
What are the basic elements of a parquet deformation? First there is the class of allowed parquets. On this Huff writes: "We play a different (or rather, tighter) game from Escher's. We work with only *A* tiles (i.e., congruent tiles of the same handedness). We do not use, as he does, *A* and *A'* tiles (i.e., congruent tiles of both handednesses), although an exception to this rule is the example called Dual. Finally, we don't use *A* and *B* tiles (i.e., two different interlocking tiles), since two such tiles can always be seen as subdivisions of a single larger tile."

The other basic element is the repertory of standard deforming devices. Typical devices include lengthening or shortening a line; rotating a line; introducing a "hinge" somewhere inside a line segment so that it can "flex"; introducing a "bump" or "pimple" or "tooth" (a small intrusion or extrusion having a simple shape) in the middle of a line or at a vertex; shifting, rotating, expanding or contracting a group of lines that form a natural subunit, and variations on these themes. To understand these descriptions you must realize that a reference to "a line" or "a vertex" is actually a reference to a line or a vertex inside a unit cell, and therefore when *one* such line or vertex is altered, all the corresponding lines or vertexes that play the same role in the copies of that cell undergo the same change. Since some of those copies may be at 90 degrees (or other angles) with respect to the master cell, one locally innocent-looking change may induce changes at corresponding spots, resulting in unexpected interactions whose visual consequences can be quite exciting.

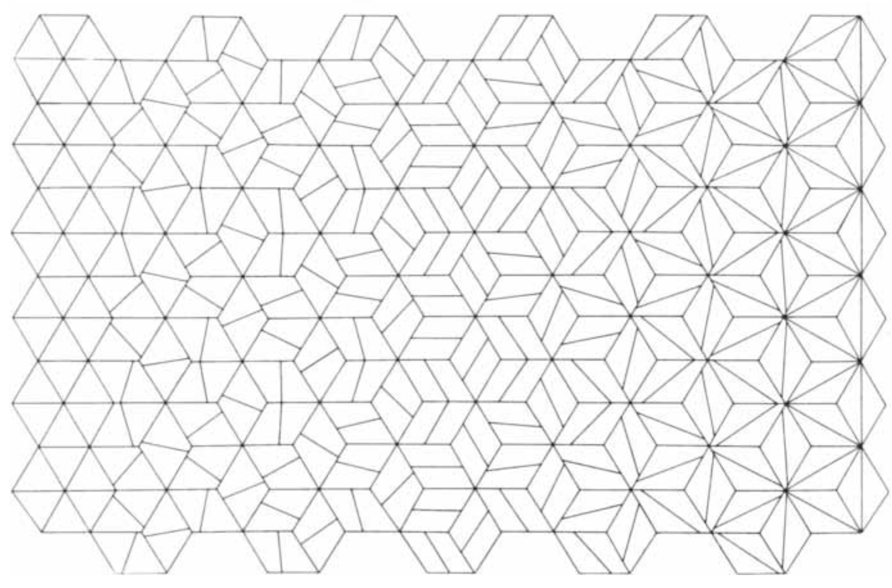
Without further ado, let us proceed to examine some specific pieces. Look at the one titled "Fylfot Flipflop" [*opposite page*]. It is an early one, exe-



"Crossover"

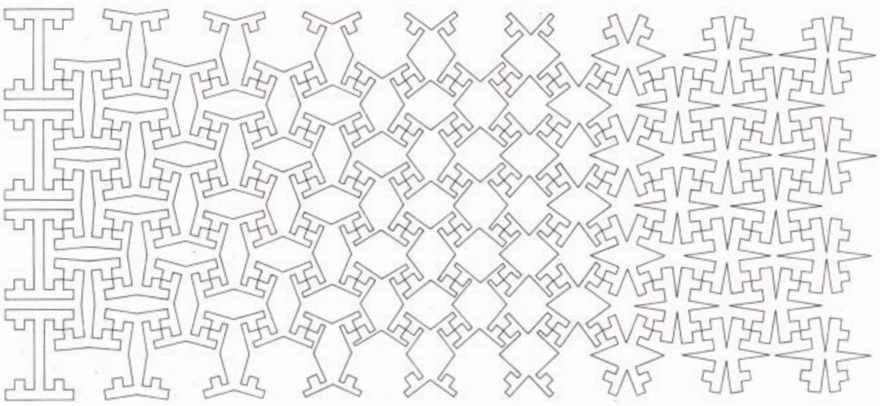


"Dizzy Bee"

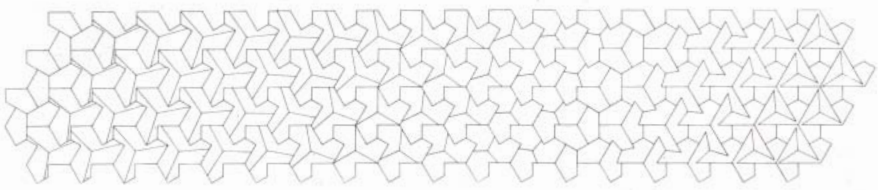


"Consternation"

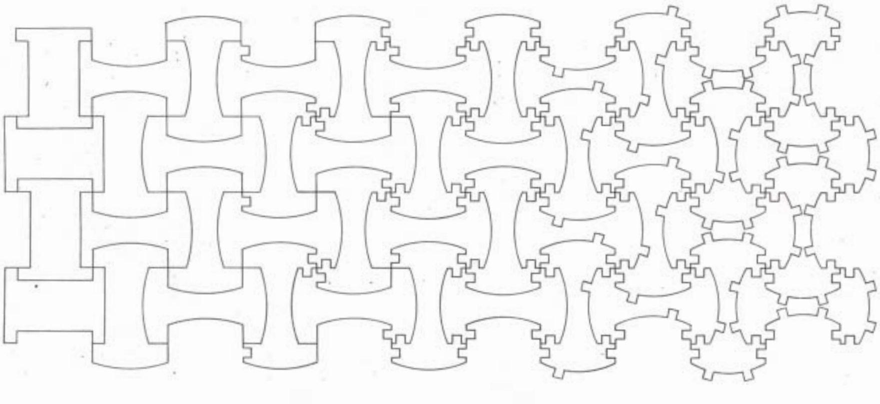
cuted by Fred Watts at Carnegie-Mellon in 1963. If you simply let your eye skim across the top line, you will get the distinct sensation of scanning a tiny mountain range. At each edge you begin with a perfectly flat plain and then move into gently rolling hills, which become taller and steeper, eventually turning into jagged peaks; past the center point these start to soften into lower foothills, which gradually tail off into the plain again. This much is obvious even at a casual glance. Subtler to see is the line just below, whose zigging and zagging is 180 degrees out of phase with the top line. Notice that in the very center that line is completely at rest: a perfectly horizontal stretch flanked on each side by increasingly toothy regions. Below it there are seven more horizontal lines. Thus if one completely filtered out the vertical lines, one would see nine horizontal lines stacked above one another, the odd-numbered ones jagged in the center, the even-numbered ones smooth in the center.



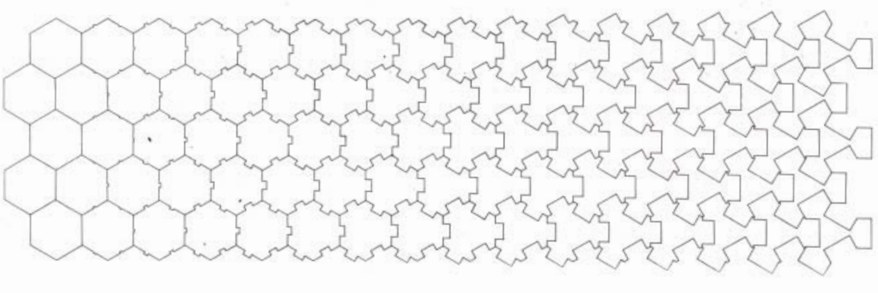
"Oddity out of Old Oriental Ornament"



"Y Knot"



"Crazy Cogs"



"Trifoliolate"

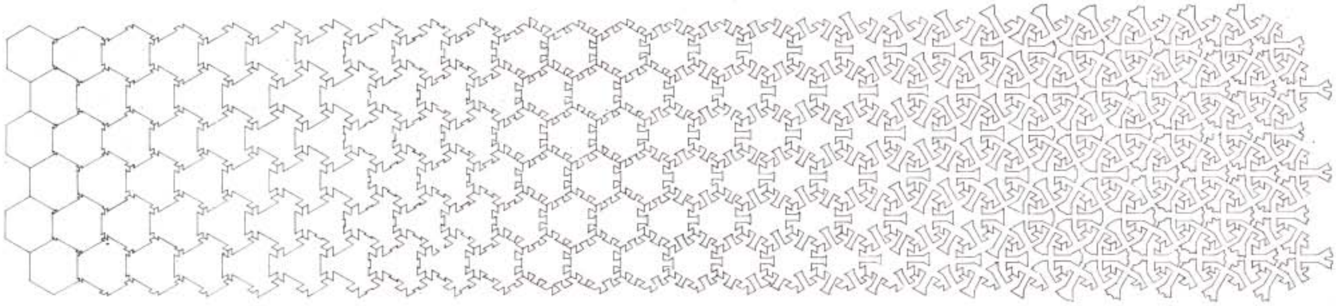
What about the vertical lines? Both the left-hand and the right-hand borderlines are perfectly straight vertical lines. Their immediate neighbors, however, are as jagged as possible, consisting of repeated 90-degree bends, back and forth. The next vertical line nearer the center is practically straight up and down again. Then there is a wavy line again, and so on. As you move across the picture you see that the jagged lines gradually get less jagged and the straight ones get increasingly jagged, until in the middle the roles are completely reversed. The process then continues, so that by the time you have reached the other side the lines are back to normal. If you could filter out the horizontal lines, you would see a simple pattern of quite jagged lines alternating with less jagged ones.

When these two extremely simple, independent patterns—the horizontal and the vertical—are superposed, what emerges is an unexpectedly rich perceptual feast. At the far left and right the eye picks out fylfots—that is, swastikas—of either handedness contained inside perfect squares. In the center the eye immediately sees that the central fylfots are all gone, replaced by perfect crosses inside pinwheels.

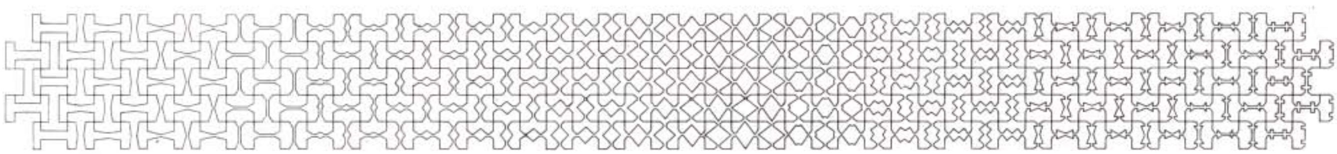
And then a queer perceptual reversal takes place. If you just shift your focus of attention diagonally by half a pinwheel, you will notice that there is a fylfot right there before your eyes! In fact, suddenly fylfots appear all over the central section where before you had been seeing only crosses inside pinwheels. And conversely, of course, now when you look at either end, you will see pinwheels everywhere with crosses inside them. No fylfots! It is an astonishingly simple design, yet the effect catches nearly everyone off guard.

This is a simple example of the ubiquitous visual phenomenon called "regrouping," in which the boundary line of the unit cell shifts so that structures jump out at the eye that before were completely submerged and invisible, whereas conversely, of course, structures that a moment ago were completely obvious are now invisible, having been split into separate conceptual pieces by the act of regrouping, or shift of perceptual boundaries. It is both a perceptual and a conceptual phenomenon, a delight to the subtle combination of eye and mind that is most sensitive to pattern.

For another example of regrouping take a look at "Crossover," executed



"Arabesque"



"Razor Blades"

by Richard Lane, also at Carnegie-Mellon in 1963 [*page 15*]. Something really amazing happens in the middle, but I won't tell you what. Just find it yourself by careful looking.

By the way, there are still features left to be explained in "Fylfot Flipflop." At first it appears to be mirror-symmetrical. For instance, all the fylfots at the left end are spinning counterclockwise, while all the ones at the right end are spinning clockwise. So far, so symmetrical. In the middle, however, all the fylfots go counterclockwise. This surely violates the symmetry. Furthermore, the one-quarter-way and three-quarter-way stages of the deformation, which ought to be mirror images of each other, bear no resemblance at all to each other. Can you figure out the logic behind this subtle asymmetry between the left and right sides?

This piece also illustrates yet another way parquet deformations resemble music. A unit cell—or rather, a vertical cross section consisting of a stack of unit cells—is analogous to a measure in music. The regular pulse of a piece of music is given by the repetition of unit cells across the page. And the flow of a melodic line across measure boundaries is modeled by the flow of a visual line—such as the mountain-range lines—across many unit cells.

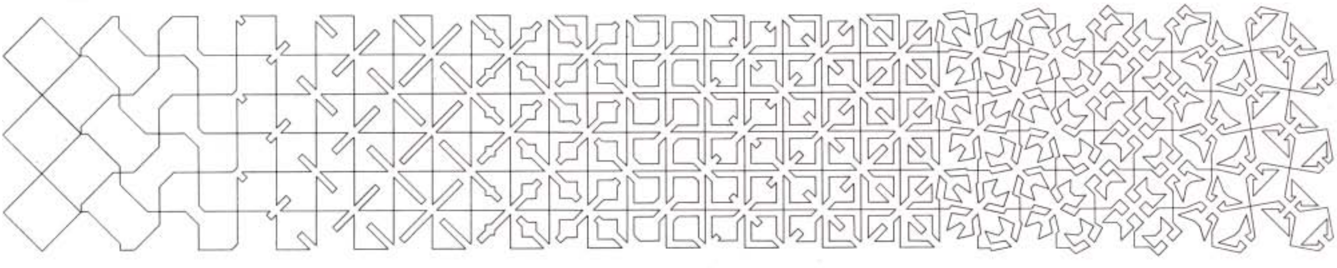
Bach's music is always called up in discussions of the relation between mathematical patterns and music, and this occasion is no exception. I am reminded particularly of some of Bach's texturally more uniform pieces, such as certain preludes from "The Well-tempered Clavier," where in each measure there is a certain pattern executed once or twice and possibly more times. From measure to measure this pattern undergoes a slow metamorphosis, meandering in the course of many measures from one region of harmonic space to far-distant regions and then slowly returning by some circuitous route. For specific examples you might listen to (or look at the scores of) Book I, No. 1 and No. 2, and Book II, No. 3 and No. 15. Many of the other preludes have this feature in places, although not for their entirety.

Bach seldom deliberately set out to play with the perceptual systems of his listeners. Artists of his century, although they occasionally played perceptual games, were considerably less sophisticated about, and less fascinated by, issues we now deem part of perceptual psychology. Such phenomena as regrouping would have intrigued Bach, and I sometimes wish he had known of certain effects and had been able to try them out, but then I remind myself that whatever time Bach might have spent playing with newfangled ideas would have had to be subtracted from his time for producing the masterpieces we know and love, and so why tamper with something that precious?

On the other hand, I do not find this argument 100 percent compelling. Who says that if you are going to imagine playing with the past, you have to hold the lifetimes of famous people constant in length? If we can imagine telling Bach about perceptual psychology, why can't we also imagine adding a few extra years to his lifetime to let him explore it? After all, the only *divinely* imposed (that is, absolutely unslippable) constraint on Bach's years is that they and Mozart's years add up to 100, no? Hence if we give Bach five extra years, then we merely take five away from Mozart. It is painful, to be sure, but not all *that* bad. We could even let Bach live to 100! (Mozart would never have existed.)

Although it is difficult to imagine and impossible to know what Bach's music would have been like if he had lived in the 20th century, it is certainly not impossible to know what Steve Reich's music would have been like if *he* had lived in this century. In fact, I am listening to a record of it right now. Now, Reich's music really *is* conscious of perceptual psychology. All the way through he plays with perceptual shifts and ambiguities, pivoting from one rhythm to another, from one harmonic origin to another, constantly keeping the listener on edge and tingling with nervous energy. Imagine a piece resembling Ravel's "Boléro," only with a much finer grain size, so that instead of its having roughly a one-minute unit cell it has a three-second unit cell. Its changes are so tiny that sometimes you can barely tell it is changing at all, whereas at other times the changes jump out at you. What Reich piece am I listening to? Well, it hardly matters, since most of his music satisfies this characterization, but for the sake of specificity you might try "Music for a Large Ensemble," "Octet" or "Violin Phase."

Let us now return to parquet deformations. "Dizzy Bee," executed by Richard Mesnik at Carnegie-Mellon in 1964 [*page 15*], involves perceptual tricks of another kind. The left side looks like a perfect honeycomb or, somewhat less poetically, a perfect bathroom floor. When we move to the right, its perfection seems in doubt as the rigidity of the lattice gives way to shapes that seem rounder. Then we notice that three of them have combined to form one larger shape: a superhexagon made up of three somewhat squashed pentagons. The curious thing is that if we now sweep our eyes from right to left back to the beginning, we can no longer see the left side in quite the way we



"Cucaracha"

saw it before. The small hexagons now are constantly grouping themselves into threes, although the grouping changes quickly. In our mind we experience "flickering clusters" where groups form for an instant and then disband, their components immediately regrouping in new combinations. The poetic term "flickering clusters" comes from a famous theory of how water molecules behave, in which the bonds are hydrogen bonds rather than mental ones.

Even more dizzying, perhaps, is "Consternation," executed by Scott Grady at SUNY at Buffalo in 1977 [*page 15*]. This is another parquet deformation in which hexagons and cubes vie for perceptual supremacy. It is so complex and agitated in appearance that I scarcely dare to attempt an analysis of it. In its intermediate regions I find the same exciting kind of visual pseudochaos as exists in Escher's best deformations.

Perhaps irrelevantly, but I suspect not, the names of many of these studies remind me of pieces by Zez Confrey, a composer best known in the 1920's for his novelty piano pieces such as "Dizzy Fingers" and "Kitten on the Keys" and—my favorite—"Flutter By, Butterfly." Confrey specialized in pushing rag music to its limits without losing musical charm, and some of the results seem to me to have a saucy, dazzling appeal not unlike the jazzy appearance of this parquet deformation.

The next parquet deformation, "Oddity out of Old Oriental Ornament," executed by Francis O'Donnell at Carnegie-Mellon in 1966 [*page 16*], is based on an extremely simple principle: the insertion of a "hinge" in one single line segment and the subsequent flexing of the segment at that hinge. The reason for the stunningly rich results is that the unit cell giving rise to the tessellation occurs both vertically and horizontally, so that flexing it one way induces a crosswise flexing as well, and the two flexings combine to yield this curious and unexpected pattern.

Another deformation that shows the amazing results of an extremely simple but carefully chosen transformation principle is "Y Knot," executed by Leland Chen at SUNY at Buffalo in 1977 [*page 16*]. If you look at it with full attention, you will see that its unit cell is in the shape of a three-bladed propeller and that the unit cell never changes in shape. All that does change is the *Y* lodged tightly inside the unit cell. And the only way the *Y* changes is by very slowly rotating clockwise. Admittedly in the final stages of rotation this forces some previously constant line segments to extend themselves a bit, but that does not change the outline of the unit cell in any way. It is remarkable what well-chosen simplicity can do.

Three of my favorites are "Crazy Cogs" (Arne Larson, Carnegie-Mellon, 1963), "Trifoliolate" (Glen Paris, Carnegie-Mellon, 1966) and "Arabesque" (Joel Napach, SUNY at Buffalo, 1979) [*pages 16 and 17*]. They all share the feature of getting increasingly intricate as you move to the right. Most of the preceding deformations do not have this extreme quality of irreversibility, that is, the ratcheted quality signaling that an evolutionary process is taking place. I can't help wondering if the designers did not think they had painted themselves into a corner, particularly in the case of "Arabesque." Is there any way you can back out of that supertangle except by retrograde motion, namely by retracing your steps? I suspect there is, but I wouldn't care to try to find it.

As a contrast consider "Razor Blades," an extended study in relative calmness [*preceding page*]. It is unsigned but was executed at Carnegie-Mellon in 1966. Like "Fylfot Flipflop," the first piece I described, this one can be broken up into long, wavy horizontal lines and vertical structures crossing them. It is a little easier to see them if you start at the right side. For instance, you can see that just below the top there is a long snaky line with numerous little nicks in it, undulating its way to the left and in so doing shedding some of those nicks, so that at the very edge it has degenerated into a perfect square wave, as such a periodic waveform is called in Fourier analysis. Complementing this horizontal structure is a similar vertical structure that is harder to describe. The thought that comes to my mind is that of two ornate, rather rectangular hourglasses with ringed necks, one on top of the other. You can see for yourself.

As with "Fylfot Flipflop," each of these patterns by itself is intriguing, but of course the real excitement comes from the daring act of superposing them. Incidentally, I know of no piece of visual art that better captures the feeling of beauty and intricacy in a Steve Reich piece, created by slow "adiabatic" changes floating on top of the chaos and dynamism of the lower-level frenzy. Looking back, I see I began by describing this parquet deformation as "calm." Well, what do you know? Perhaps I would be a good candidate for one of *The New Yorker's* occasional notes titled "Our Forgetful Authors."

More seriously, there is a reason for this inconsistency. One's emotional response to a given work of art, whether the work is visual or musical, is not static and unchanging. There is no way of knowing how you will respond the next time you hear or see one of your favorite pieces. It may leave you unmoved or it may thrill you to the bone. It depends on your mood, on what has recently happened, on what happens to strike you and on many other subtle intangibles. One's reaction can even change in the course of a few minutes. And so I won't apologize for this seeming lapse.

Let us now look at "Cucaracha," executed by Jorge Gutierrez at SUNY at Buffalo in 1977 [*above*]. It moves from the utmost geometricity—a lattice of perfect diamonds—through a sequence of gradually more arbitrary modifications until it reaches some kind of near-freedom, a dance of strange, angular, quasi-organic forms. This fascinates me. Is entropy increasing or decreasing in this rightward flow toward freedom?

A gracefully spiky deformation is the one wittily titled "Beecombing Blossoms," executed by Laird Pylkas at SUNY at Buffalo this year [*opposite page*]. Huff told me Pylkas struggled for weeks with it and at the end, when she had resolved her difficulties, she mused: "Why is it that the obvious ideas always take so long to discover?"

As our last study let us take "Clearing the Thicket," executed by Vincent Marlowe at SUNY at Buffalo in 1979, which mixes straight lines and curves, right angles and cusps, explicit squarish swastikoids and implicit circular holes [*page 20*]. Rather than demonstrating my inability to analyze the ferocious complexity of the design, I should like to use it as the jumping-off point for a brief discussion of computers and creativity.

Some totally new things are going on in this particular parquet deformation, things that have not appeared in any previous one. Notice the hollow circles at the left that shrink as you move to the right; notice also that at the right there are hollow "anticircles" (concave shapes made from four circular arcs turned inside out) that shrink as you move to the left. Now, according to Huff, such an idea had not appeared in any previously created deformations. This means that something unusual happened here: something genuinely creative, something unexpected, unpredictable, surprising, intriguing and, not least, inspiring to future creators.

Thus the question naturally arises: Would a computer have been able to invent this parquet deformation? Well, put this way it is a naive and ill-posed question, but we can try to make some sense of it. The first thing to remind ourselves of is that the term "computer" refers to nothing more than an inert hunk of metal and semiconductors. To go along with this bare computer, this hardware, we need software and energy. The former is a specific pattern inserted into the hardware, binding it with constraints and yet imbuing it with goals; the latter is what breathes "life" into it, making it act according to those constraints and goals.

The next point is that the software is what really controls what the machine does; the hardware simply obeys the software's dictates, step by step. And yet the software could exist in a number of different "instantiations": realizations in different computer languages. What really counts about the software is not its literal aspect but a more abstract, general, overall "architecture," which is best described in a nonformal language such as English. We might say that the plan, the sketch, the central idea of a program is what we are talking about here, not its final realization in some specific formal language or dialect. That is something we can leave to apprentices to carry out after we have presented them with our informal sketch.

Hence the question actually becomes less mundane and more theoretical and philosophical: Is there an architecture to creativity? Is there a plan, a scheme, a set of principles that, if it were elucidated clearly, could account for all the creativity embodied in the collection of all parquet deformations, past, present and future?

Note that we are asking about the *collection* of parquet deformations, not about some *specific* work. It is a truism that any specific work of art can be re-created, even re-created in various slightly novel ways, by a programmed computer.

For example, Piet Mondrian evolved a highly idiosyncratic, somewhat cryptic style of painting over a period of many years. You can see, if you trace his development over time, exactly where he came from and where he was headed. If you focus on a single Mondrian work, however, you cannot sense this stylistic momentum, this quality of dynamic, evolving style that any great artist has. Looking at one work in isolation is like taking a snapshot of something in motion: you capture its instantaneous position but not its momentum. Of course, the snapshot might be blurred somehow, in which case you would get a sense of the momentum but would lose information about the position. When you are looking at just a single work of art, however, there is no mental blurring of its style with that of recent works or soon-to-come works; you have exact position information ("What is the style *now?*") but no momentum information ("Where was the style and where is it going?").

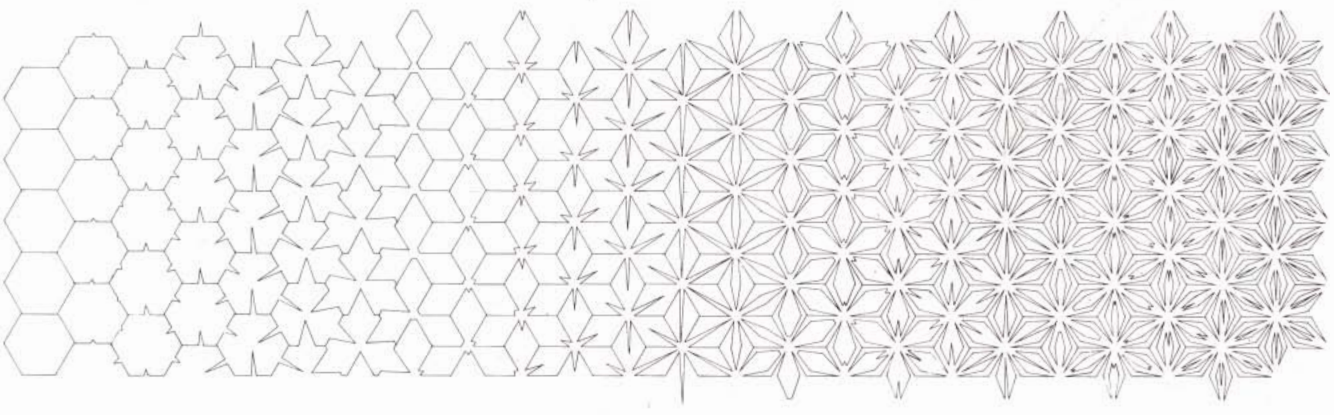
Some years ago A. Michael Noll, a mathematician and computer artist, took a single Mondrian painting—an abstract, geometric study with seemingly random elements—and from it extracted some statistics on the patterns. Given these statistics, he programmed a computer to generate numerous pseudo-Mondrian paintings having the same or different values of these randomness-governing parameters. Then he showed the results to viewers who had no foreknowledge of what he was up to. The reactions were interesting: more people preferred one of the pseudo-Mondrians than preferred the genuine Mondrian!

This is quite amusing, even provocative, but it also is a warning. It proves that a computer can certainly be programmed, after the fact, to imitate—and imitate well—mathematically capturable stylistic aspects of a given work. It also warns us, however: Beware of cheap imitations!

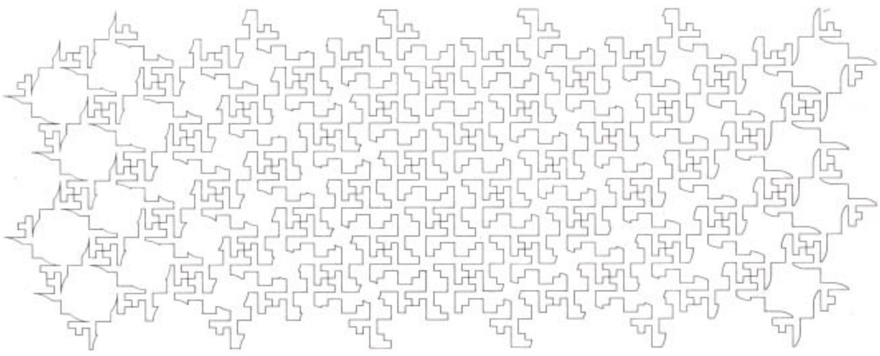
Consider parquet deformations. Undoubtedly a computer could be programmed to do any *specific* parquet deformation or minor variations on it without too much trouble. There simply are not that many parameters to any given deformation. The essence of any artistic act lies not, however, in selecting particular values for certain parameters but far deeper: in the balancing of myriad intangible and mostly unconscious mental forces, a judgmental act that results in many conceptual choices eventually adding up to a tangible, perceptible, measurable work of art.

Once the finished work exists scholars looking at it may seize on certain qualities of it that lend themselves to being easily parametrized. *Anyone* can do statistics on a work of art once the work is there for the scrutiny, but the ease of doing so can obscure the fact that no one could have said a priori what kinds of mathematical observables would turn out to be relevant to the capturing of stylistic aspects of the as-yet-unseen work of art.

Huff's own view on this question of mechanizing the art of parquet de-



"Beecombing Blossoms"



"Clearing the Thicket"

formations closely parallels mine. He believes some basic principles could be formulated at the present time that would enable a computer to generate relatively stereotyped yet novel creations of its own. He stresses, however, that his students occasionally come up with rule-breaking ideas that enchant the eye for reasons deeper than any he has been able to verbalize. In this way the set of explicit rules gets gradually enlarged.

Comparing the creativity that goes into parquet deformations with the creativity of a great musician, Huff writes: "I don't know about the consistency of the genius of Bach, but I did work with the great American architect Louis Kahn (1901–1974) and suppose it must have been somewhat the same with Bach. That is, Kahn, out of moral, spiritual and philosophical considerations, formulated ways he would and ways he would not do a thing in architecture. Students came to know many of his ways, and some of the best could imitate him rather well (although not perfectly). But as Kahn himself developed he constantly brought in new principles that brought new transformations to his work, and he even occasionally discarded an old rule. Consequently he was always several steps ahead of his imitators who knew what *was* but couldn't imagine what *will be.* And so it is that computer-generated 'original' Bach is an interesting exercise. But it isn't Bach—that unwritten work that Bach never got to, the day after he died."

The real question is: What kind of architecture is responsible for *all* these ideas? Or is there any *one* architecture that could come up with them all? I would say that the ability to design good parquet deformations is probably deceptive in the same way as the ability to play good chess is: it seems more mathematical than it actually is.

A brilliant chess move, once the game is over and can be viewed in retrospect, can be seen as logical, as "the correct thing to do in that situation." But brilliant moves do not originate from the kind of logical analysis that occurs after the game; there is no time during the game to check out all the logical consequences of a move. Good chess moves spring from the organization of a good chess mind: a set of perceptions arranged in such a way that certain kinds of ideas leap to mind when subtle patterns or cues are present. The way perceptions have of triggering old and buried memories underlies skill in any type of human activity, not only chess. It is just that in chess the skill is particularly deceptive, because after the fact it can all be justified by a logical analysis, a fact that seems to hint the original idea *came from* logic.

Writing lovely melodies is another one of those deceptive arts. To the mathematically inclined, notes seem like numbers and melodies like number patterns. Therefore all the beauty of a melody *seems* as if it ought to be describable in some simple mathematical way. So far, however, no formula has produced even a single good melody. Of course, you can look back at any melody and write a formula that will produce it and variations on it. But this is retrospective, not prospective. Lovely chess moves and lovely melodies (and lovely theorems in mathematics) have this in common: every one has idiosyncratic nuances that seem logical a posteriori but are not easy to anticipate. To the mathematical mind chess-playing skill and melody-writing skill and theorem-writing skill seem obviously formalizable, but the truth turns out to be more tantalizingly complex than that. Too many subtle balances are involved.

So it is with parquet deformation, I reckon. Each one taken alone is in some sense mathematical. Taken as a class, however, they are not mathematical. This is what is tricky about them. Don't let the apparently mathematical nature of an *individual* deformation fool you; the architecture of a program that could create all these parquet deformations and more good ones would have to incorporate computerized versions of *concepts* and *judgments*, and those are much more elusive and complex than mere numbers.

At this point many critics of computers and artificial intelligence, eager to find something "computers can't do" (and never will be able to do), often go too far: they jump to the conclusion that art and, more generally, creativity are fundamentally uncomputerizable. This is hardly the implied conclusion! The implied conclusion is that if computers are to be enabled to act human, we shall have to wait until we have good computer models of such human properties as perception, memory, mental categories, learning and so on. We are a long way from that. There is no reason to assume, however, that those goals are in principle unattainable, even if they remain far off for a long time.

In this column I have been playing with the double meaning of the term "architecture": it means both the design of a habitat and the abstract essence of a grand structure of any kind. The former has to do with hardware and the latter with software. In a certain sense, William Huff is a professor of both brands of architecture. Obviously his professional training is in the design of "hardware," namely genuine habitats for human beings, and he works in a school where that is what they do. He is also, however, in the business of forming in the minds of his students a softer kind of architecture: the mental architecture that underlies the skill to create beauty. Fortunately for him he can take for granted the complexity of a human brain as the starting point on which to build this architecture.

When I first met Huff and saw how abstract and seemingly impractical were the marvelous works produced in his design studio—ranging from parquet deformations to strange ways of slicing a cube to gestalt studies using thousands of dots to eye-boggling color patterns—I at first wondered why this man was a professor of *architecture.* But after talking with him and his colleagues my horizons were extended about the nature of their discipline.

The architect Louis Kahn had great respect for the work of William Huff, and it is with his words that I should like to conclude:

"What Huff teaches is not merely what he has learned from someone else but what is drawn from his natural gifts and belief in their truth and value. In my belief what he teaches is the introduction to discipline underlying shapes and rhythms, which touches the arts of sight, the arts of sound and the arts of structure. It teaches students of drawing to search for the abstract and not the representational. This is good as a reminder of order for the instructors/architectural sketchers (like me), and good especially for the student sketchers without background. It is the introduction to exactitudes of the kind that instill the religion of the ordered path."

BOOKS

Cholera, whales, statistical snapshots of the U.S., rare wonders of the Incas

by Philip Morrison

CHOLERA: THE AMERICAN SCIENTIFIC EXPERIENCE, 1947–1980, by W. E. van Heyningen and John R. Seal. Westview Press, Boulder, Colo. ($27.50). A thousand miles of the leeward southern slopes of the Himalayas receive the fierce monsoon rains, to drain into the sea through the many mouths of the broad Ganges. That intricate delta is the heart of Bengal, populous and fertile, 200 miles across between its two capitals east and west, Dacca and Calcutta. For a long time the dwellers in those wetlands have endured an endemic bacterial disease, cholera, that nearly every year takes a high toll in human life. It is an abrupt and terrifying affliction, "the bowels turning to water"; the patient—chilled, thirsty, vomiting as he passes a clear, watery stool by the pint every hour or two, yet alert—soon develops an "intense fear of impending death." That fear is justified within a few days, for as many as two cases out of three.

Two centuries ago, perhaps following some genetic change in the cell strain, cholera learned to travel. Six pandemics followed; the calamity became a catchword and a curse throughout the West. Americans counted their dead from cholera by the hundred thousand as recently as the years after the Civil War. During the worst visitation of Britain, in the 1850's, the queen's "incomparable" physician, John Snow, demonstrated beyond doubt that the infection was carried in water contaminated by sewage. He did not even need to know that its cause was a comma-shaped bacillus, found swarming in the intestinal contents of victims in the same years by Filippo Pacini. In capital-rich countries the spread of *Vibrio cholerae* has for most of a century been avoided by the scrupulous separation of piped drinking water from sewage, although it took a generation before Snow's proof led to public action. But the world's resources are so ill-shared that most tropical countries cannot yet afford such prevention for the million villages at risk.

This engrossing and varied volume, as much at home in the laboratory politics of medical expatriates and their capricious institutional patrons as it is in cyclase biochemistry, brings the story up to date. Cholera is expensive to prevent, although the developed world has managed it. Over the past decade, however, it has become cheaply curable, through insight. Two authors, the first an Oxford biochemist, the second an American medical-research administrator, are veterans of the campaign. A 30-year war against the vibrio has been waged across the world by a shifting international team of men and women, largely Americans who knew little of cholera but much of physiology and biochemistry, allied with the experienced doctors of the cholera countries.

Deadly cholera is a superficial disease, only skin-deep. The thin layer of cells affected is the epithelial lining of the small intestine, an area measuring 200 square meters, a broad living carpet "fit for the Palace of Versailles," all cunningly folded into the 10-foot tube of the small intestine packed within each human abdomen. Only a single layer of cells is affected by the microbes, which secrete a specific protein toxin. The toxin quickly enters and raises the level of the messenger substance (cyclic adenosine monophosphate) in the cells, in analogy to cell response to hormones. This intricate piece of biochemistry is quite well understood, although just how the message changes the cell is not yet known.

The result is a shift in the balance between the absorption and the secretion of water and its dissolved ions by that cell layer. Each hour the normal intestinal wall passes four or five liters of water; ordinarily the balance inward and outward is elegantly struck. Only a few minutes' flow is needed to replace the normal net loss over a day. A photograph shows a slight, cheerful man, a newly recovered cholera patient, "surrounded by 96 litre bottles" that had held the intravenous fluid that pulled him through a few days' illness. A loss of 30 quarts of water a day is not untypical. The disease affects only that thin cell layer; the toxin never enters the bloodstream.

Those cells are in any case sloughed off steadily from the gut; in normal circumstances they live and die in less than a week. The toxin, however, poisons them at once; they simply remain without adequate function for a few days, until they are duly replaced by unaffected new cells in the usual sequence. The disease is therefore quickly self-limiting, less systemic than a common cold. The patient dies, but not from the toxin, which has only induced an invisible biochemical lesion in that single layer of short-lived cells. Death is from intense dehydration and demineralization, the lifeblood turned dark and thick. Early workers tried to relieve the symptoms by giving the patient a lot of water to drink, but their efforts worked only transiently; an hour or two later the shadow of death was back. Enough water just would not go in.

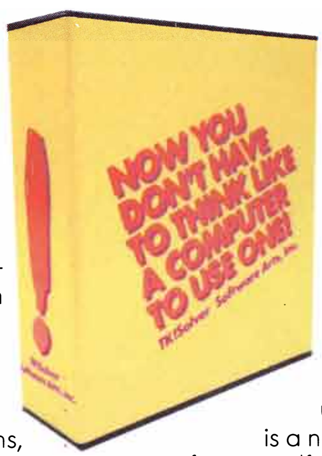
It was in the fall of 1947, during a rogue outbreak of cholera along the Nile, that the nature of death from cholera was grasped. An epidemiological team concluded from its surprise encounter with the acute disease that "the diarrhoea... constitutes the entire pathologic aspect of the disease." That team was a new U.S. Navy group, its head a Rockefeller Institute physiologist, Comdr. Robert Allan Phillips, M.C., U.S.N. Phillips had come to wartime Cairo as a naval medical officer to join the U.S. Typhus Commission, set up there in 1942 to unite American, British and Egyptian efforts to deal with typhus, feared in North Africa both during and after the war.

Phillips had a taste for the role, and so once the war ended and his Cairo laboratory was about to be disbanded he "rushed back to Washington and made his pitch for support to every likely agency." Only the Navy, with its long tradition of concern for tropical disease and a flexible and enthusiastic postwar view of research, acted. Cholera attacked almost at once; the unit's improvised but brilliant defense marked the beginning of the modern American campaign against cholera, led over the years by a shifting set of delicate alliances among "the Navy, the Army, the Department of State, the Department of Health, Education, and Welfare, and the academic world."

This chronicle is as diverse as the story of any modern war. The politics of the committee room late at night in some Atlantic City hotel can fix or move forces halfway across the world; foreign policy opens and closes research. Field leaders are removed or promoted as at home their friends and critics clash or concur. In the field changing ideas and rapid actions count, yet chance shapes results. Directors matter, but so does the talent of Mr. Razzack, the mechanic. He kept at work the 37 motor launches whose river ambulance patrols in the

CRUNCH!

The inventor of VisiCalc,® the world's leading software product, introduces TK!Solver™—the revolutionary way to crunch problems on a personal computer.



With VisiCalc, Software Arts™ created an entirely new application for personal computers—the electronic spreadsheet.

Now Software Arts has developed TK!Solver, another entirely new and different application for personal computers—the problem solver. Business and financial problems. Scientific and engineering problems. Architectural and building design problems. Any kind of problem you can dream up, in any area.

What's so revolutionary about that? Simply this: the power of the TK!Solver program comes from the ease with which you can set up problems, vary assumptions, find solutions, and display results. All the facilities needed to solve the problems—simple or complex—are built-in, and need not be developed. Once you have stated your problem, simply enter the known values, then solve the problem with a single keystroke.

For even easier problem solving use a TK!SolverPack™ in which models for solving common problems in such areas as finance, engineering and education have already been formulated by professionals in these fields.

There is no software product in existence that solves problems as quickly, flexibly, and easily as TK!Solver—The Problem Cruncher™. In fact, no other software product even works the way TK!Solver does. It is a necessity for personal computer owners. See for yourself. Visit your nearest computer store for a demonstration and find out how much you can do with a personal computer when it thinks as flexibly as you do.

Software Arts™

27 Mica Lane, Wellesley, MA 02181

cholera season along the waterways a couple of hours out of Dacca by boat were the chief means of the field trial of cholera vaccine there in 1964. So does the courage of the Baltimore volunteers who earlier tested the vaccine.

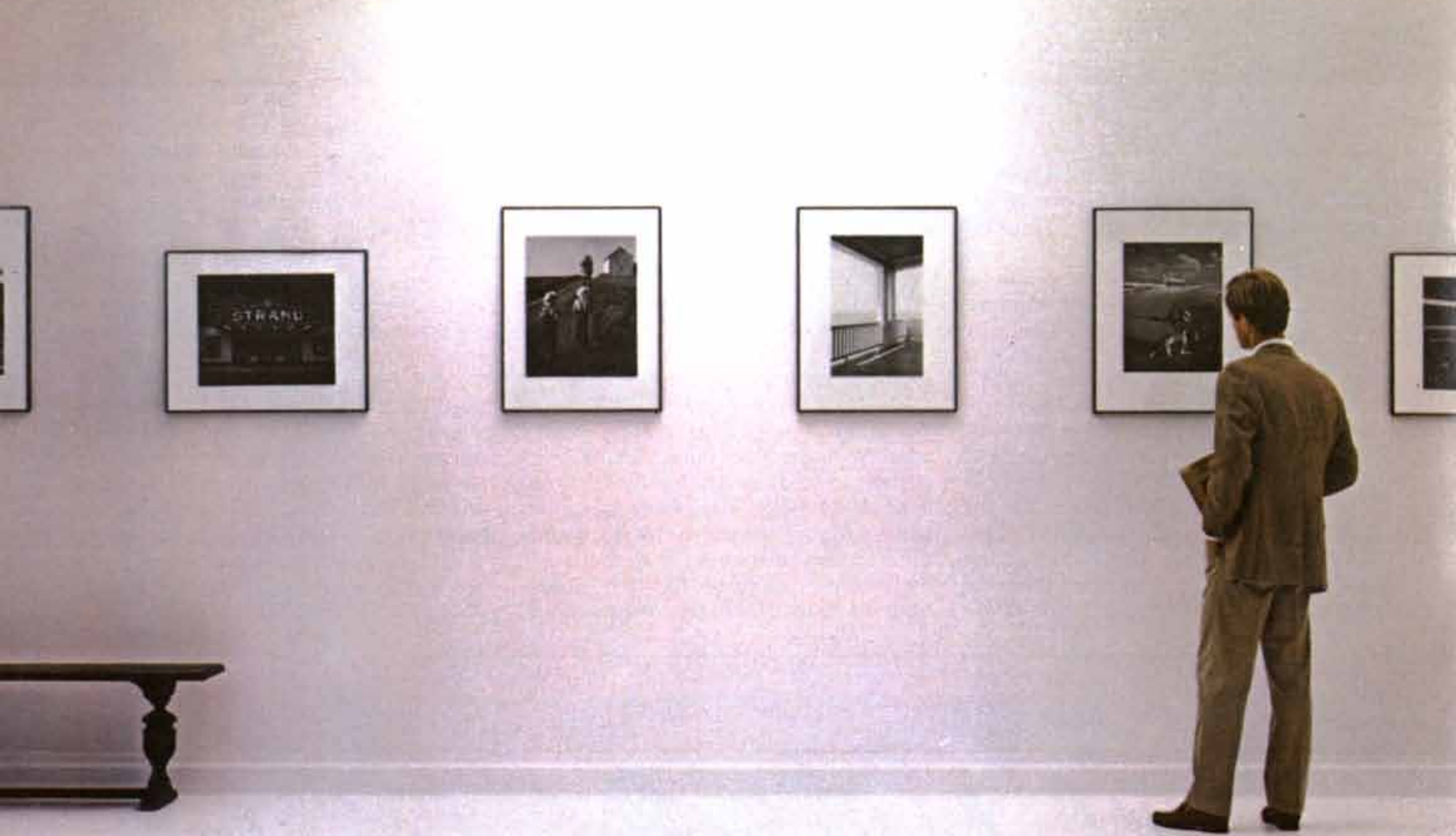
The battles and the medical units are too many to list here. The first decade and more were spent in learning how to save cholera patients by the prodigal intravenous injection of sterile, isotonic, pyrogen-free saline water under knowing control. In Thailand, in the Philippines, in South Vietnam, wherever the teams were at work, cholera took less than 1 percent of those who came under the painstaking care required for injecting floods of intravenous saline. But few cholera patients could come under such care, which demanded at least a large corps of trained medical aides, if not doctors and hospitals. The defeat of vaccines—both those using cells and those using the newly crystallized toxin—has been disappointing, although not unexpected in a disease that injures so quickly and so superficially. The cheap antibiotic tetracycline kills the vibrios, but only after their thin-layer damage is already done.

The simple cure, almost if not quite a final victory, came in about 1968, presaged by an unhappy but revealing trial under Phillips in Manila in 1962 and by various laboratory results even earlier. The Dacca and Calcutta centers both confirmed about then that adding glucose to drinking water enabled the patient to absorb enough water; the rationale appears to be that the passage of a triple complex of sodium ion, glucose molecule and water molecule through the decisive membrane is not at all inhibited by the cholera toxin. Field trials were forced by history on the Calcutta laboratory run by Johns Hopkins in the monsoon months of 1971, when six million refugees fled westward from the civil war in East Pakistan. Triumph! Nearly 4,000 patients were treated, with an overall fatality rate under 4 percent. The treatment was oral fluids made up of local materials packaged on the spot in the laboratory's library, enough for all the patients at a total cost of $750.

The victory is not quite total; it may be that some people will always need intravenous rehydration and not just those cheap cupfuls of sugared and salted drinking water. But it looks as though the toxin of cholera, as deadly as that of tetanus or botulinus but specific to epithelial cells instead of to the deep-lying nerve endings of the central or motor system, has been routed. The defeat is not by medical prevention—unless the piped water and adequate nutrition of the well-to-do can be called medical—but by cure, a simple cure by drinking water in the right way.

There is a new center now in Dacca, internationally supported and administered but at home there, dedicated by the president of Bangladesh in 1979. It has high hopes and a wide brief: no less than global attack on all the diarrheal diseases, together accounting for 500 million cases each year, a chief cause of death in children particularly. In 1978 *The Lancet* wrote: "The discovery that sodium transport and glucose transport are coupled in the small intestine, so that glucose accelerates absorption of solute and water, was potentially the most important medical advance this century." The total dollar cost to U.S. taxpayers of their generation-long contribution to the war against cholera was about the cost of one of today's air-superiority fighter planes.

MOBY-DICK; OR, THE WHALE, by Herman Melville. Illustrated with 100 woodcuts by Barry Moser. University of California Press ($25). THE ECOLOGY OF WHALES AND DOLPHINS, by D. E. Gaskin. Heinemann Educational Books, Inc. Exeter, N.H. ($45). Does the whale's magnitude diminish? Will he perish? In Chapter 105 of his ocean of a book Melville asks those questions and then answers them: "And as upon the invasion of their valleys, the frosty Swiss have retreated to their mountains; so, hunted from the savannahs and glades of the middle seas, the whalebone whales can at last resort to their polar citadels...and bid defiance to all pursuit from man.... Wherefore we account the whale immortal in his species." Melville wrote of three-masted ships, before Svend Foyn and his cannon with its grenade-tipped harpoon, before the powered catcher boat and the big floating factories. This is no place to appraise Melville's own prose leviathan; we only note here a typographically beautiful edition of the work, decorated with striking woodcuts that avoid visual interpretation of character or event, to present only the tools, the prey, the ships and the ports. The handsome folio was hand-set in Goudy Modern and hand-printed by Andrew Hoyem in 1979 in a few hundred copies; this edition is a reduced facsimile of that collector's plea-



It's now possible for people who've only admired great pictures to take them.

The Programmed FG.

Inside everyone who's ever gazed appreciatively at a photograph is the urge to take great pictures themselves.

And now, no matter how much or how little picture taking experience you've had, there's a camera worthy of an admiring look.

The remarkable programmed FG.

The FG is a camera that ambitious beginners can still use after they've become accomplished veterans.

In the programmed mode, just focus and shoot. A microchip computer automatically selects the proper lens opening and shutter speed.

And through-the-lens flash metering with the optional SB-15 speedlight makes flash pictures virtually fool-proof.

Automatic and manual, too.

But what happens when you want a camera that allows you more flexibility?

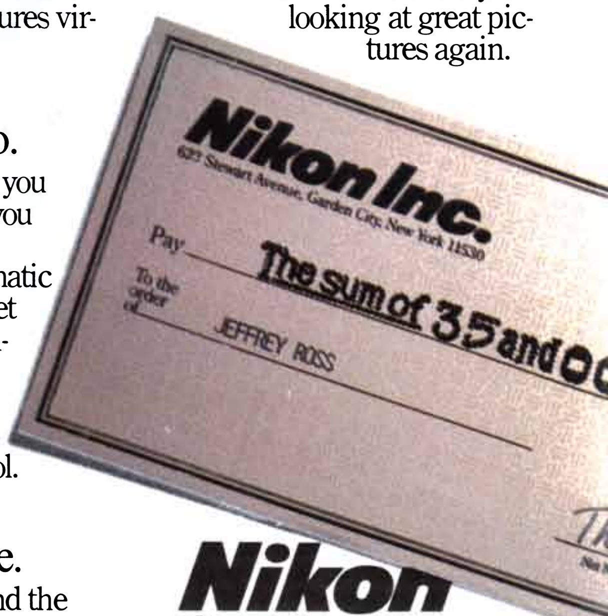
Then the FG is an automatic camera that allows you to set the aperture yourself to control depth-of-field. And, in the manual mode, you can make all settings yourself for complete creative control.

$35 rebate for a limited time.

But the real genius behind the FG is that it costs so much less than you'd think. And until August 15, 1983, there is a $35 cash rebate.*

So visit your Nikon camera dealer soon and check out the ingenious programmed FG.

You'll never be content just looking at great pictures again.



Nikon

We take the world's greatest pictures.™

*Offer applies only to products which include the Nikon, Inc. limited U.S.A. warranty form. Photographs courtesy of the Witkin Gallery, Inc., N.Y.C. © Nikon Inc. 1983. For further information write Dept. 47, Garden City, N.Y. 11530.

sure, down in page area by a factor of about two and in price by one of 40.

D. E. Gaskin addresses the very question Melville set. He is a marine biologist at the University of Guelph in Ontario. He counts 20 years in the study of whales and their kin, with field experience from the circumpolar Southern Ocean to the Newfoundland banks. His volume is a critical technical review of what is known—and not known—about the main species of cetaceans, looked at from the viewpoint of populations over time and space, with only the indispensable minimum of material on the individual animals and their adaptations. As Melville's whale was metaphysical, so is our whale today political, "a semimythological beast."

A summary of such a thorough review of an exponentially growing scholarly literature is perforce schematic. In the first third of the book whales are presented as seasonably migratory animals (several species are discussed in detail) that find distinct requirements for breeding and for the taking of food; their distribution is a long-evolved compromise between these two needs. Their travels are empathetic: winter in the sun, summer in the cool. It is the prey that leads the whale; sperm whales take mostly squid in the open sea, mostly fish closer to shore. Sperm migrations—recall the white whale—are diffuse, ill-defined. (Although the big males are locally overexploited, this wide-ranging species numbers some 600,000 worldwide: "a satisfactory condition.") But the decimated plankton-feeding giants, the blues, move and group regularly from Tropic to pole with the sun because their polar food, the krill, are particularly seasonal, patchy and discontinuous.

How whales navigate and even how they hunt we do not really know; echolocation, star sights, even shore landmarks may play a part for one species or another. A recent experiment with a single humpback, temporarily captive, found that the blindfolded animal could not run a simple maze of obstacles. It is such doubt, raised by experience, of the easy and attractive assumptions that follow from our modest insight into the animals that sets the tone of Professor Gaskin's cool book.

The metabolism of whales is of course a key for any fundamental ecological study. In one very interesting chapter it is made clear that whales and porpoises probably do not much differ from other mammals in their energy needs, the weight taken into account. An estimate for a 50-ton male cachalot, such as Moby-Dick, suggests an intake of squid at 3 percent of the whale's body weight per day; the female sperm whale, smaller by a factor of four or so, has a narrow margin of nutrition, particularly in view of the demands of lactation. This species lives usually segregated by sex; perhaps—it is a conjecture—that is because the energy demands of the two sexes are so different.

The big blues are biggest in the Antarctic; there the single-species patches of krill abound for only four months. In the Northern Hemisphere the fewer, smaller blues take more species of the zooplankton over a six-month season. The large size of the blue may be an adaptation to the short feeding season; the energetic issue is between staying on in a marginal feeding ground or cruising off to warmer waters. The matter is more complex still, because it appears (there is some doubt) that the big rorquals are too dense to float and must spend power just to stay near the surface to breathe. The management question hidden in all the careful budgets for propulsion, heat loss by conduction and radiation, and the like, is the retrospective calculation of how much krill the blues ate when they were numerous. The guesses suggest that the standing crop of krill was 10 or 20 times the total annual catch of all the world's fisheries. If the little krill live for a few years, as studies show, the future commercial krill take might be indefinitely sustainable at a level comparable to all the fishing the world now does for all species. Maybe.

Whales are gregarious to a degree. This text is rather skeptical about social implications, finding that simple groups and mutual finds of food are as likely as more integrated and complex behavior. It seems clear that groups of sperm whales keep in close acoustic-ranging contact as they forage in the depths. The animals disperse on diving but signal their identity frequently; each knows where the others are. They can return to the surface in a tight school; the calves, unable to dive as deep, may be protected by this stratagem from sharks.

A quite technical and uncertain third of the book seeks to understand the evolution and speciation of the order of whales and porpoises. Recent results on chromosome structure support the idea that all the cetaceans, small and large alike, came from a single line in the Paleocene, a view not popular until now.

Finally the author takes up the analysis and management of whale populations, the core of Melville's concern and our own. It is not easy to count whales: they are not randomly scattered, markers are hard to plant and to recover, age determination and recognition of individuals by their natural patterns are less than reliable. In 1980 a finback whale was for the first time fitted with a radio transmitter that survived; the animal was followed for 1,000 miles. These years we are passing a phase in animal ecology "largely as a result of intense love affairs with high-speed computing" that has plenty of models and rather less input data. Paper whales are simulated well with silicon chips, real ones and their hunters much less so.

Models of single species, like the lumped quotas and regulations dear to international agreements, are abstractions that may prove dangerously misleading. The chapter on the sorry history of postwar pelagic whaling is a gloomy one; Gaskin inclines to the view that mere short-run economics has been used to justify the near-extinction of species. The profitable idea is to match the lifetime of the gear to that of the fishery! The fishery declines sadly; years later there is a new big crop and new fishing investment: call it pulse fishing. Twenty years of pure research would not be a bad international plan for the present. Even though our analysis is poor, it remains true that the major failures have been ethical ones. The regime of local 200-mile limits, the lure of the krill, climatic effects on grain crops—all suggest that the diminished whales may not rest safe even in their lonely citadel.

U/S: A STATISTICAL PORTRAIT OF THE AMERICAN PEOPLE, edited by Andrew Hacker, with the assistance of Lorrie Millman. The Viking Press and Penguin Books ($25). STATISTICAL ABSTRACT OF THE UNITED STATES: NATIONAL DATA BOOK AND GUIDE TO SOURCES, 103RD EDITION, 1982–1983, U.S. Department of Commerce, U.S. Bureau of the Census. Superintendent of Documents, U.S. Government Printing Office ($15). The map is clear: since 1790 the point that marks the decennially reckoned centroid of the population of the U.S. has moved westward from the shores of Chesapeake Bay to vault the broad Mississippi for the first time in 1980. Ever since 1910 it has drifted in a straight line with a southerly trend, pulled by that distant southwestern desert sun.

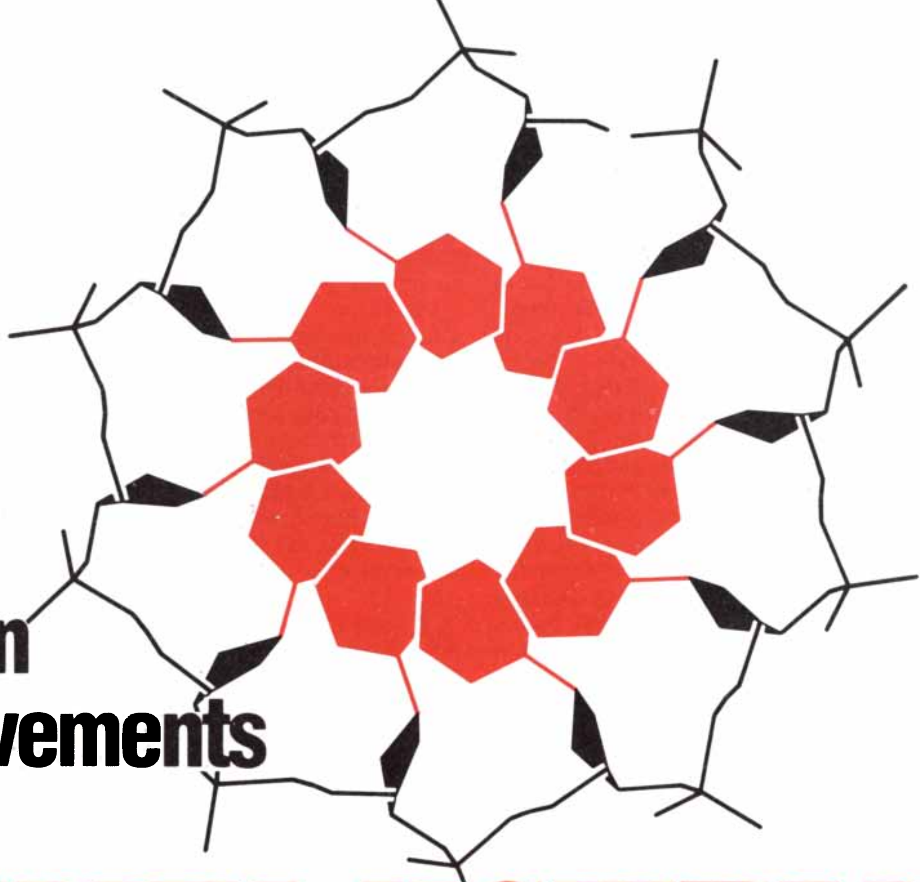
The constitutional head count fixed the number of men, women and children resident in these United States on April 1, 1980, at just 226,545,805. This angelic accuracy is a fiction; the Bureau of the Census reckoned before the fact that it would be about 222 million people. The bureau received in the end about 2.2 percent more official name forms than that, a result seeming to mean that the undercounting by 2.5 percent the bureau itself had corrected for in 1970 had ceased. By 1982, however, the bureau, under suit in the courts, reversed its views. Black Americans were still undercounted, the bureau conceded, by nearly 5 percent. The report remains officially unchanged, although a figure near 228 million might be held a better approximation.

The decennial count is not the only source of these data today. Each month, mainly seeking the unemployment rate, the bureau questions a careful sample of about 71,000 households in 1,100 counties to get a Current Population Report, which it can use for projections. On that basis the population of the U.S., augmented by the Armed Forces overseas, will be some 234 million, plus or minus about .8 million, on July 1, 1983. This figure is up about 1 percent per year from 1980, the net result of a birth rate of about 1.6 percent, a death rate of .85 percent and a net immigration rate of .2 percent. The projection for the year 2000 is 268 million, give or take 12 or 15 million.

To count is human, and by no means easy. The small book by the sociologist Andrew Hacker and his colleague is a critical presentation of the results of the census and of many other continuing statistical series, offered both as a book to read through and as a useful reference. Much of the material is tabular, but the tables are set into a running comment and comparison that is quite readable. The focus is demographic: how many of us there are, where we live, the vital events of birth and death, health, households, marriage and divorce. Like the census itself, however, *U/S* treats jobs and incomes, taxes and government, crime, education, housing and travel. Major questions are included in full, such as income distribution by family, by sex and race, by occupation and source, and much more.

A little cheerful numerical gossip also enters, for example the compensation of top corporate executives in 1981. Half a dozen tycoons collected more than $1 million, like top professional outfielders, even though one or two of the companies they headed were far from Easy Street. (Mobil and Union Oil could afford it.) In sum, the book offers a skeptical yet smoothed entry into what remains a rather austere assessment of American life, and it deserves both readers and browsers. It has an excellent index and a careful list of sources.

The second volume reviewed here is another matter: the annual official compendium of some 1,600 tables and quite a few graphs in 1,000 close-set pages. It was commented on in this column in the edition of 1972, after the preceding census. This time around it is still a big bargain, and its print and paper are decidedly better. It presents the unadorned tables of its sources without comment, although it carefully describes, perhaps too formally, the samples and their pitfalls. The Census Bureau undertakes to compile statistics from a wide variety of Government and non-Government sources, from the United Nations to the Federal Bureau of Investigation and from the Corporation for Public Broadcasting to *Jane's Fighting Ships.* It is plain that data are viewed as power; every bureau and industrial association tends to count the items in its own



A celebration of the achievements of a science

BIOCHEMISTRY

SECOND EDITION

LUBERT STRYER

Stanford University

949 pages
983 illustrations
$32.95

Lubert Stryer's *Biochemistry* has become the standard against which all other biochemistry books should be measured. It has shaped and continues to shape the subject it teaches.

First published eight years ago, *Biochemistry* immediately set new standards in the teaching and presentation of biochemistry. Emphasizing the process of discovery, Stryer aimed "to provide an introduction to the principles of biochemistry that gives the reader a command of its concepts and language." Reviewers commented that it was "doubtful whether there will be another text which will appeal to the student as strongly as this one does" (*Nature*).

The Second Edition, "bigger and even more splendid," is "a most handsome and thought-provoking new edition of a fine textbook" (*TIBS*). "It continues the tradition of excellence established in his first edition" (*American Scientist*), brilliantly achieving the author's aim as stated in the Preface: "to weave new knowledge into the fabric of the text...to enhance the book's teaching effectiveness by centering the exposition of new material on common themes wherever feasible and by citing recurrent motifs....and to convey a sense of the intellectual power and beauty of the discipline of biochemistry."

With a clarity and conciseness achieved through the unified narrative of a single author, "....the brevity of the text often belies its depth: the author has a unique ability to expound difficult concepts in a few words." (*Nature*)

Stryer's *Biochemistry*, Second Edition—giving up-to-date coverage of new developments while communicating the excitement and progress of biochemical frontiers.

If you are not using Stryer, compare *Biochemistry* with your present text. After the comparison, even *more* the choice.

W. H. Freeman and Company
41 Madison Avenue
New York, N.Y. 10010

universe, from mobile homes to violent assaults.

They are all here. It is the catholicity of the tables that wins the count-intoxicated reader. The catch of menhaden remains the biggest fishery in point of weight landed, as it has been for decades, although its value as a source of fish meal is only a few cents per pound. The sunflower has risen to the second most valued crop in the state of North Dakota; it is a monopoly of that state, at least as a major crop. The Federal agency with the largest budget for space was the National Aeronautics and Space Administration only from 1961 to 1981; before and since that 20-year period the Department of Defense led. In 1981 more musical works were registered for copyright than any other category of work, including serials and monographs, if the Library of Congress is right. Nearly 60 million people either hunted or fished legally for sport in 1980. Total food consumption per head remained the same (to an accuracy of 1 percent!) from 1970 to 1980. The most conspicuous changes in the diet appear to be the rise of chicken, the decline of eggs, fluid milk and cream, an increase in cheese and in fresh fruit and a surprising fall in refined sugar (more than compensated, however, by a rise in corn sweetener).

Crimes went steadily up from 1970 to a peak in 1980, with a small decrease since then. The gross death rate has been flat for 20 years and more, the rising life expectancy about keeping pace with the aging population. Newspaper circulation has been flat since 1970, and so has television viewing (except that the television has gone from black and white to color). The number of overseas telephone calls has risen by a factor of about eight. Per capita mail volume has gone up by about 15 percent over the decade; most of that rise (as the reader may have noted) is in third-class items. Union membership has grown very slowly in the past 10 years, much more slowly than the labor force. Women have found work as barbers, lawyers and computer specialists, to cite the few occupations in which there have been sizable increases in the employment of women. For minority employees office management and aircraft maintenance show a similar record. Serious and trivial together, these pages exemplify what it is fashionable to call a data base. Some base!

There is one disappointment. Last time one could pore over the economic and demographic parameters of a couple of hundred urban centers, to compare them easily. Alas, although the new list of urban centers is complete with populations, all the other comparative data are displaced into a supplementary volume: *State and Metropolitan Area Data Book*. It is not easy to characterize urban centers; they are complex. If one examines the pages carefully, it appears that great New York, with 16 million souls from New Brunswick to Montauk, is at the top of a list of 11 leading centers and Cleveland-Akron-Lorain, with just under three million, is at the bottom.

As before, the volume is not parochial; it includes a fine set of comparative international statistics, most of the economic data from UN sources. On a long list of countries Saudi Arabia leads in the relative economic importance of construction and is last in that of agriculture. Such a fact is not really surprising, but it expresses what one might surmise with an economy that is all but poetic.

MONUMENTS OF THE INCAS, text by John Hemming, photographs by Edward Ranney. A New York Graphic Society Book. Little, Brown and Company ($45). The joint effort of an American photographer and an English author, this extremely attractive volume transcends the merely showy to offer an honestly comprehensive if elegant photographic documentation of Tawantinsuyu, the empire built by the Incas within less than a century. What is left of its monuments lies beautiful and enigmatic before the lens, against the snowy peaks of the Andes along a sweeping arc from broad Lake Titicaca north to equatorial Quito. There are plenty of books about that empire conquered near its zenith, and images of its remains are commonplace, but Ranney has done more than any other photographer to show the works of the Incas as they are set into the numinous mountain landscapes that somehow called forth those superb forms in stone.

The first chapter of 15 sketches Inca history and presents an overview of Inca architecture, drawing explicitly on the admirable recent monograph of Graziano Gasparini and Luise Margolies, reviewed here (in translation) in March, 1981. Then the chapters treat, one after another, the famous localities more or less as they might be visited by a lucky traveler heading north from the Island of the Sun all the way to Ingapirca, the only surviving Inca ruin in Ecuador. About two-thirds of the book is devoted to the central region around Cuzco, that mountain capital planned in the form of a puma, here studied from Raqchi north to Vilcabamba. The illustrations are evocative and knowing, no less for a key carved detail than for a vista of meadow and peak. The text is sensitive and well documented; telling citations from the old chroniclers jostle the archaeological reports and the travelers' traditions. Hemming makes good use of the indispensable work of contemporary Peruvian scholarship; the footnotes are full and the bibliography is a find for serious beginners.

An armchair reader, envious of the fortunate who will visit the sites with this book behind them as a kind of hyperguide, found on page after page remarks and particularly scenes that escape notice in many a worthwhile book on the Inca world. The snow-covered peaks grandly tower over the rolling meadows of Soccllacasa pass, "near the rock shrines of Saihuite." There two rocks lie alone in the high open rangeland, about 100 miles toward the sea from Cuzco. The principal stone of the pair "looks from a distance like a broken flint." It lies where it has lain for a very long time, a boulder of 200 or 300 tons, never much moved by human hands. But that broken surface 15 feet across has been fashioned into "an intricate mass of carving." Small figures, pumas and hunters, lizards, frogs, snakes, vicuñas and more, have been released there, all set into a labyrinth of tiny stairs, altars and platforms, through which a network of grooves once led liquid down from the highest point in some unknown "ceremony of divination." Iconoclasts have defaced the figures, save for one brooding puma. The two photographs of this monument are extraordinary; it is a statement as timeless as it is contemporary. A microcosm was somehow carved out by the Inca's craftsmen; the 19th-century travelers who saw an Andean topographic map in that carving were not far wrong.

The vista of the real city of Machu Picchu from the southern rise of its saddle in the mountains, a familiar and beautiful shot, given us by Ranney also, seems now at a glance only one more Saihuite rock. Notches in the stone rim offer evidence that the face of the Saihuite boulder once had a copper or gilded-copper cladding, probably worked at a level of detail the relief carving in stone can never transmit to us. Perhaps the golden cities of the tall tales were real after all, even if they were only models. The strange gnomonlike platforms called inti-huatanas are called "hitching posts of the sun." Several are shown as they stand now in the Cuzco region. They are not, however, working gnomons or zenith markers as far as anyone can now make out; it seems a reasonable conjecture that here too what we now see is only a carefully carved stone base, its surfaces rounded and bland, never itself an astronomical instrument but only a kind of armature for a more intricate and controlled metal cladding. Together were they not working solar instruments?

The photographs and their excellent reproduction in duotone are so beautiful that they satisfy entirely; the work is a black-and-white prize and an archaeological introduction of real merit.

Introducing the new Osborne™ Executive™ Series.

We listened when you told us you wanted a personal business computer.

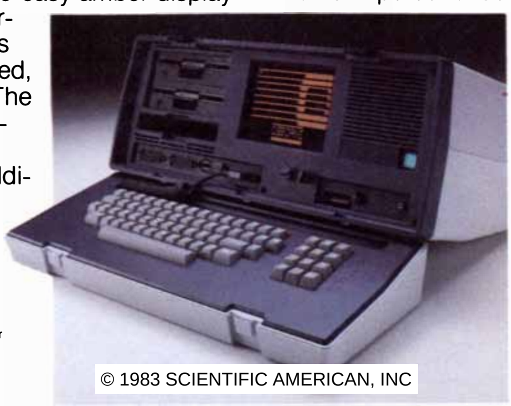
We delivered the Osborne 1.

And in the two years since its introduction, the Osborne 1 has become the largest selling portable business computer on earth.

We listened harder.

The result is the new Osborne Executive Series. Consider these new advantages:

☐ A big, 80-column, eye-easy amber display lets you see more of the information you're using, with less effort. ☐ Memory is expanded, for larger spreadsheets. ☐ The new Executive Series now includes a database system—Personal Pearl.™ (That's in addition to an already formidable list of standard Osborne software, including word processing and electronic spreadsheets.) ☐ Terminal Emulation provides access to information stored on other computers. ☐ IBM Personal Computer™ compatibility (on the Executive 2) lets you use software packages designed for the IBM system.



And there's much more.

Add up the value of all the hardware features and software systems included with either the Executive or Executive 2 and you will see that the sticker price is one of the incredible bargains available in personal computing.

Visit your local authorized Osborne dealer. Once again, Osborne has set the standard for portable business computers.

For the name of your nearest Osborne dealer, dial (800) 772-3545 ext. 905 (in California); (800) 227-1617 ext. 905 (outside California).

OSBORNE™
COMPUTER CORPORATION

Trademarks: Osborne 1, Executive, Executive 2: Osborne Computer Corporation. Personal Pearl: Relational Software. IBM Personal Computer: International Business Machines.

THE WORLD'S FASTEST BAGGAGE HANDLER?

If there's a station wagon in the world that's faster than the Turbo Wagon from Volvo, it's a very well-kept secret.

The Turbo Wagon can take you from a standing start to the legal speed limit in 9.2 seconds. As *Motor Trend* puts it: "You keep hoping for a stray Saab 900 Turbo or even a BMW you can sniff out and then send scrambling to the roadside while you blow its doors in."

We assure you, a test drive in the Turbo Wagon will be very revealing.

Especially if you thought the only time a station wagon could haul was when it had something in the back.

THE TURBO WAGON

By Volvo.

Established 1845 SCIENTIFIC AMERICAN July 1983 Volume 249 Number 1

The Slowing of Urbanization in the U.S.

The rural population is growing as fast as the urban population for the first time. The geographic expansion and coalescence of urban areas, however, has led to the appearance of massive "supercities"

by Larry Long and Diana DeAre

One of the most fundamental features of a modern society is the demographic predominance of the cities over the countryside. Urbanization, along with the industrial growth that has generally accompanied it, has historically distinguished the developed countries from the developing ones. In a country that is undergoing economic development two tendencies usually contribute to the increasing dominance of the urban areas. First, the population of established cities grows faster than the population of the nation as a whole. Second, in some rural areas population growth is rapid enough for the areas to become urban. In the U.S. both of these basic forms of urbanization have been operating almost continuously at least since the decennial census was initiated in 1790.

The detailed results of the census of 1980, however, show that urbanization in this traditional twofold form was strongly modified in the U.S. between 1970 and 1980. The first form of urbanization ceased during the decade: the population of rural areas grew as fast as that of urban areas. Only the second kind of urbanization is still in progress: between 1970 and 1980 the population of many rural areas grew enough for them to be reclassified as urban. If it had not been for growth in areas that were considered rural in 1970, the fraction of the U.S. population living in urban areas would have decreased for the first time. As the growth of cities was largely attributable to economic factors, so too the decline appears to have economic causes. The available data suggest that the rural areas' attractiveness for industry is responsible for much of their population growth.

Although the deceleration of urban growth is widespread, it is by no means geographically uniform. In the Northeast both kinds of urbanization have ended. In the South and West both kinds are still in progress, but the overall rate of urbanization is lower than it was in the past. Moreover, the current trends have quite different effects on the various geographic components of a city. In the 1950's and 1960's a common pattern was for the city proper to lose population while the suburbs gained. In many instances the city itself continued to lose people in the 1970's and the area subject to population loss expanded to include the inner suburbs. The outer suburbs, however, continued to grow both in population and in area. As a result the urban areas of the U.S. have become larger and less densely populated.

Urban areas that are expanding in area can ultimately run into one another. The coalescence of neighboring cities in the 1970's led to an increase in the number of "supercities": continuously built-up urban areas with a population of a million or more. The growth of the supercities shows that the current trends do not necessarily mean the end of large urban centers. Recent developments do suggest, however, that the economic advantages of locating industry in cities may have been reduced. As a result new and less concentrated forms of human settlement may be in prospect.

Measuring the number of people who live in urban areas requires a definition of what is urban, which would appear to be a straightforward matter. One could, for example, simply count the people who live within the boundaries of cities of a certain size. Such a definition, however, would omit many of the residents of the suburbs, who are certainly part of the urban population. Moreover, the question would naturally arise of how many residents a city must have before it is included in the count.

Because of such ambiguities there are many possible methods of distinguishing the urban population from the rural. The Bureau of the Census has several ways of identifying the urban part of the U.S. The oldest definition includes all incorporated places with a population of 2,500 or more. Such a broad category almost inevitably includes some areas that do not correspond to the ordinary notion of what is urban.

The Census Bureau also employs a narrower category that more closely approximates the usual notion. The second category defines an urbanized area as a city with a population of 50,000 or more together with all the contiguous built-up land that has a population density of at least 1,000 per square mile. The large city included in the urbanized area is termed the central city. In this article we refer to the suburbs that share a boundary line with the city as the inner suburbs; all other parts of the urbanized area are referred to as the outer suburbs.

The concept of the urbanized area was introduced in analyzing the results of the census of 1950. The motivation for its introduction was the rapid growth of the suburbs in the years after World War II. The urbanized area provides a good measure of the demographic limits of a city as opposed to its administrative boundaries, which often do not coincide with the extent of urbanization.

If the urbanized areas are removed from the broad category of all urban places, what remains is the group of nonsurburban cities and towns with a population of more than 2,500 but less than 50,000. In the terminology of the Census Bureau such areas are referred to as "other urban" places; here we shall call them the small cities and towns. Territory that is outside both the urbanized areas and the small cities and towns is considered rural.

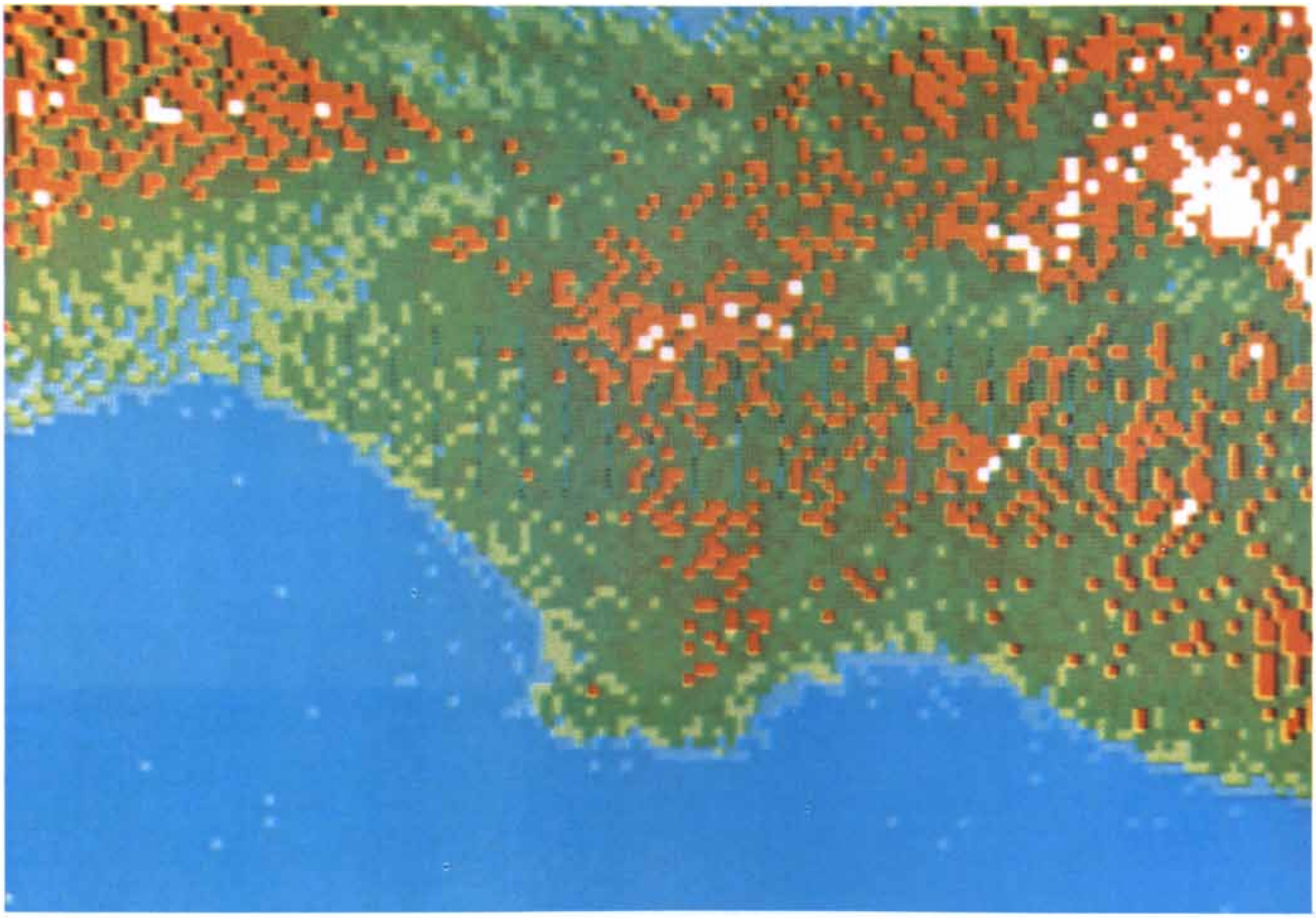
We have employed the returns from the census of 1980 to analyze recent trends in the growth of the urban and rural populations. We have examined developments in both the urbanized areas and the small cities and towns, and we shall describe some of the trends in the territory that is urban in the broad sense. We shall concentrate, however, on the urbanized areas. The reason is that included among the small cities and towns are places that in an important sense are more rural than urban. Furthermore, many small cities and towns outgrow that classification to become urbanized areas and others are absorbed in the expansion of the large cities.

The definitions of what is urban employed by the Census Bureau have undergone technical changes since 1970. We have adjusted the data from the census of 1980 to make them comparable with the data from the three preceding censuses. The adjustment has enabled us to make consistent comparisons of trends in population, geographic area and population density in the U.S. from 1950 to 1980.

The most general and significant conclusion that emerges from our work is that the level of urbanization stopped increasing for the first time in the 1970's. The population of the urbanized areas is still growing, but at a rate lower than that of the rural areas. As a result the fraction of the nation's population in urbanized areas is no longer increasing. In 1950, 46 percent of the U.S. population lived in an urbanized area; in 1960, 53 percent did so; in 1970, 58 percent. In 1980 the fraction was still 58 percent. In some parts of the country the fraction is decreasing.

If the urbanized areas are classified according to when they became large enough to be designated as such, the deceleration of the urbanizing process appears even more dramatic. The oldest urbanized areas are the ones designated in 1950. In the 1950's and 1960's these older cities grew substantially faster than the population of the nation. By the 1970's, however, the older urbanized areas, with an average annual population growth rate of .8 percent, were growing slower than the national population, which had a rate of 1.1 percent. If it had not been for the faster-than-average growth of the newer urbanized areas, the fraction of the population in the large cities would have declined.

These data imply that one of the two fundamental forms of urbanization—the growth of preexisting urban cen-



THERMAL ISLAND is shown on a map of the area around Los Angeles and Long Beach in California. The map indicates the daytime surface temperatures. The data were collected by a satellite operated by the National Aeronautics and Space Administration and were plotted by a computer. White corresponds to the highest temperature; dark green, light green and red correspond to intermediate temperatures; blue corresponds to the lowest temperature. A thermal island is an area of elevated temperature resulting from the high density of manmade structures. The hottest part of the region is an industrial park (*white area at upper right*). Los Angeles and Long Beach make up an urbanized area: a city with 50,000 people or more and all the contiguous territory with a density of at least 1,000 per square mile.

ters—stopped in the U.S. in the 1970's. The slowing of the growth of the older cities was very rapid. In the 1950's the rate of population growth in the older urbanized areas and those added in 1960 was 23 times the rate of areas outside them. In the 1960's the rate of growth in the older urbanized areas and those that had been added by 1970 was 16 times the rate of the rest of the country. In the 1970's the rate of growth in the urbanized areas was only marginally greater than the rate in other places, and it was higher only because of rapid growth in areas that had been considered rural in 1970. This development is unprecedented in U.S. demographic history.

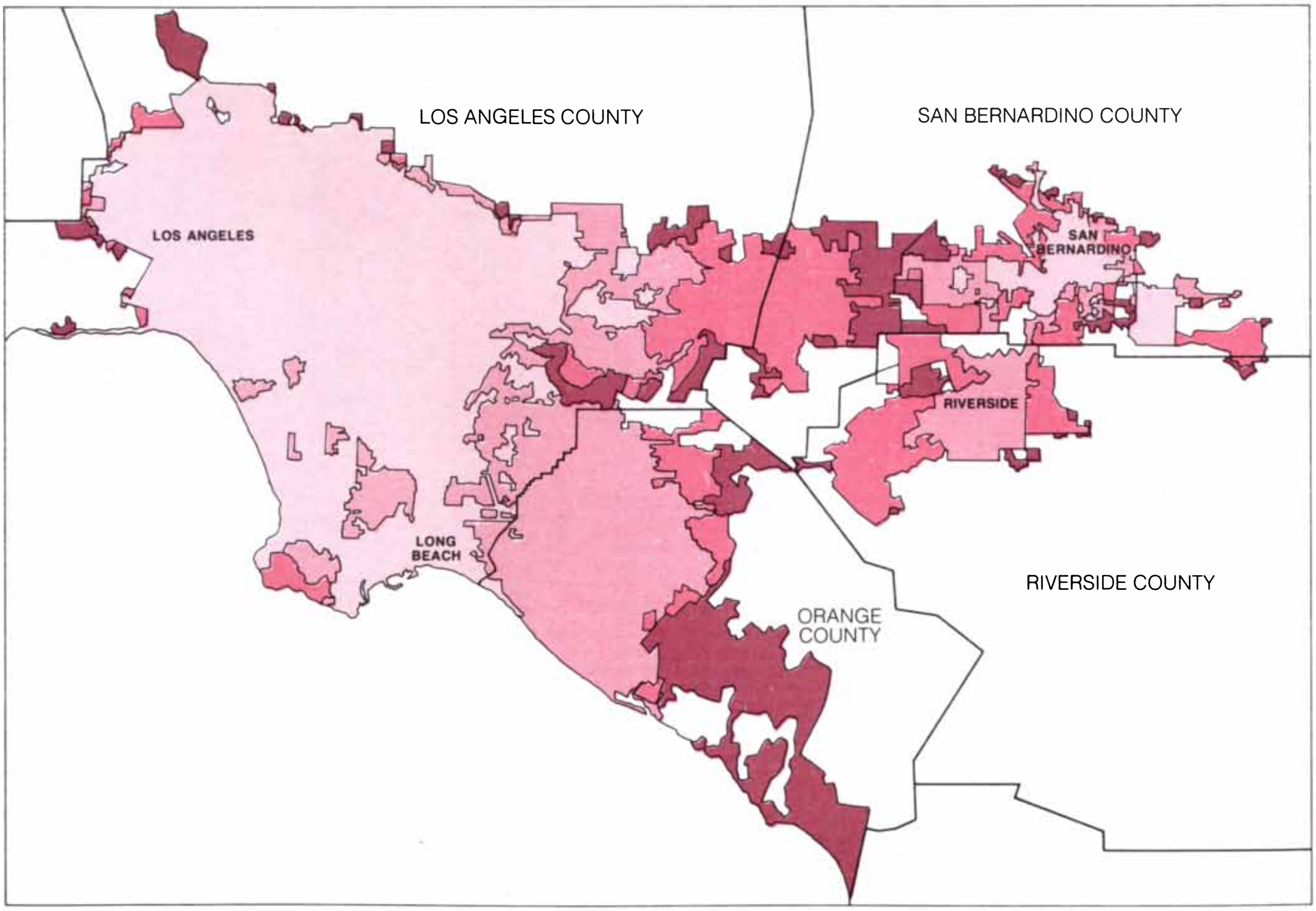
If the broad concept of what is urban is employed, the level of urbanization did not merely stop increasing in the 1970's but actually declined. If those who live in small cities and towns are included, the fraction of the U.S. population in urban areas was 64 percent in 1950. It increased to 70 percent in 1960 and to 73.5 percent in 1970 before decreasing to 73.3 percent in 1980.

A decline of less than one percentage point is not substantial; its significance is not in its size but in the fact that it comes after a long period of uninterrupted growth. An advantage of the more inclusive definition of urban territory is that it can be applied to census records extending far into the past. The level of urbanization in the broad sense declined only once before: between 1810 and 1820. That decline was the result of population growth in several rural states of the Middle West that had been added by the Louisiana Purchase.

The slowing of the growth of the urban population may be a temporary phenomenon that could be reversed by changes in economic factors such as the price of fuel. It is more probable, however, that we are witnessing a basic demographic reversal. The immediate cause of the reversal is not a decrease in the absolute size of the urban population but a rapid increase in the rate of population growth in rural areas. The rural population declined slightly in the 1950's and 1960's. In the 1970's, however, the rural population increased by 12 percent, a rate of growth about equal to that of the urbanized areas and greater than that of the small cities and towns, which increased by 9 percent. Thus the demographic growth outside the large cities took place in the open countryside rather than in the small towns.

The rate of urbanization decreased in the U.S. as a whole in the 1970's. The national trends, however, conceal significant differences among the regions of the country. For the purposes of the census the U.S. is divided into four major regions. The Northeast region includes the New England states as well as New Jersey, New York and Pennsylvania. The North Central region includes the Great Lakes states and Iowa, Kansas, Missouri, Nebraska, North Dakota and South Dakota. The South and West regions consist of the remaining states, with Texas being the westernmost state in the South region.

In the Northeast in the 1970's there



SOUTHERN CALIFORNIA SUPERCITY is made up of the Los Angeles, Long Beach, San Bernardino and Riverside urbanized areas. As the urbanized areas expanded geographically in the 1970's they ultimately encountered the boundaries of other urban places. When this happened, they merged into supercities: continuously built-up urban masses with a population of a million or more. The map shows the boundaries of the urbanized areas of southern California as determined by the censuses of 1950 (*lightest color*), 1960 (*light color*), 1970 (*dark color*) and 1980 (*darkest color*). By 1980 the Los Angeles–Long Beach area and the San Bernardino–Riverside area had merged into a massive supercity. Such expansion has generally been accompanied by a decrease in the population density of the urbanized areas.

was a decrease in the population of the urbanized areas from 34.8 million to 34.4 million. The fraction of the total population of the region in urbanized areas declined from 71 to 70 percent. Furthermore, no new urbanized areas came into being in the region during the decade. The Northeast is therefore a highly urbanized, demographically mature region that has become less urbanized because both basic forms of urbanization have ended.

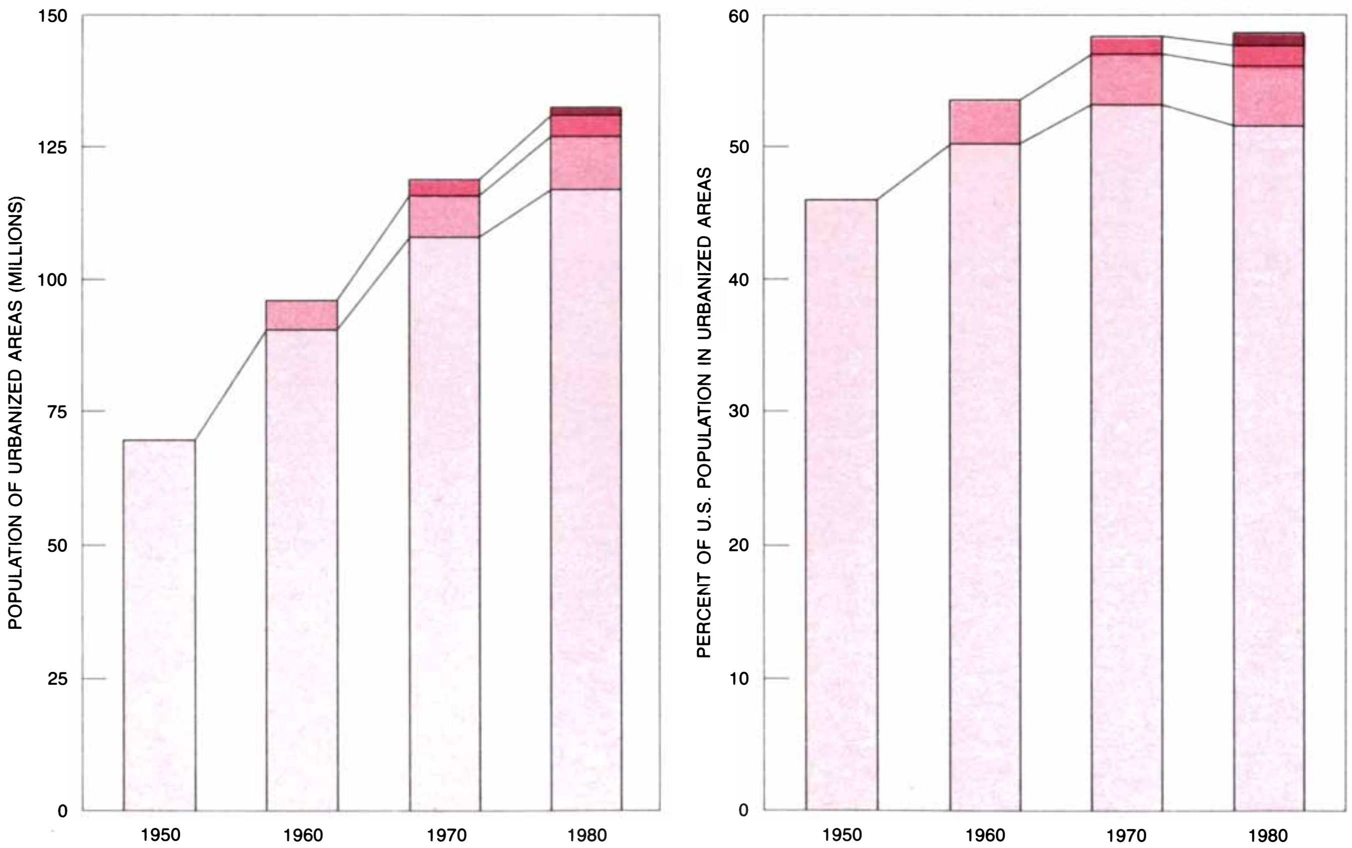
In the North Central region there was a slight increase in the population of the urbanized areas: from 31.5 to 31.9 million. The population outside the urbanized areas, however, grew at a much higher rate and as a result the fraction of the population in the urbanized areas declined from 56 to 54 percent. The newer urban centers in the North Central region (those designated in 1960 or 1970), on the other hand, grew faster than the regional average. In addition a number of places grew rapidly enough to be newly classified as urbanized areas. Hence in the North Central region the decrease in the level of urbanization was solely the result of population decreases in the older urban centers.

In the South and West the degree of urbanization increased according to the twofold traditional pattern. Intriguingly, although the West is often thought of as a region of "wide open spaces," it is in reality almost as urban as the Northeast, which is much more readily imagined as a highly urban region. Between 1970 and 1980 the fraction of the population of the West in an urbanized area increased to 69 percent; as we have indicated, the fraction in the Northeast decreased to 70 percent. This startling comparison emphasizes the difference between the demographic and the geographic measures of urbanization, a distinction that is quite important in analyzing current trends in urbanization. Although most of the land in the West is not devoted to urban uses, most of the region's inhabitants live in urban places.

It is probable that the level of urbanization in the West will for some time remain higher than the level in the North Central and South regions. In the South the population of urbanized areas increased from 29 to 36 million; in 1980 this represented 48 percent of the total population. In both the West and the South the relative increase in the population of the urbanized areas was smaller in the 1970's than it had been in the two preceding decades.

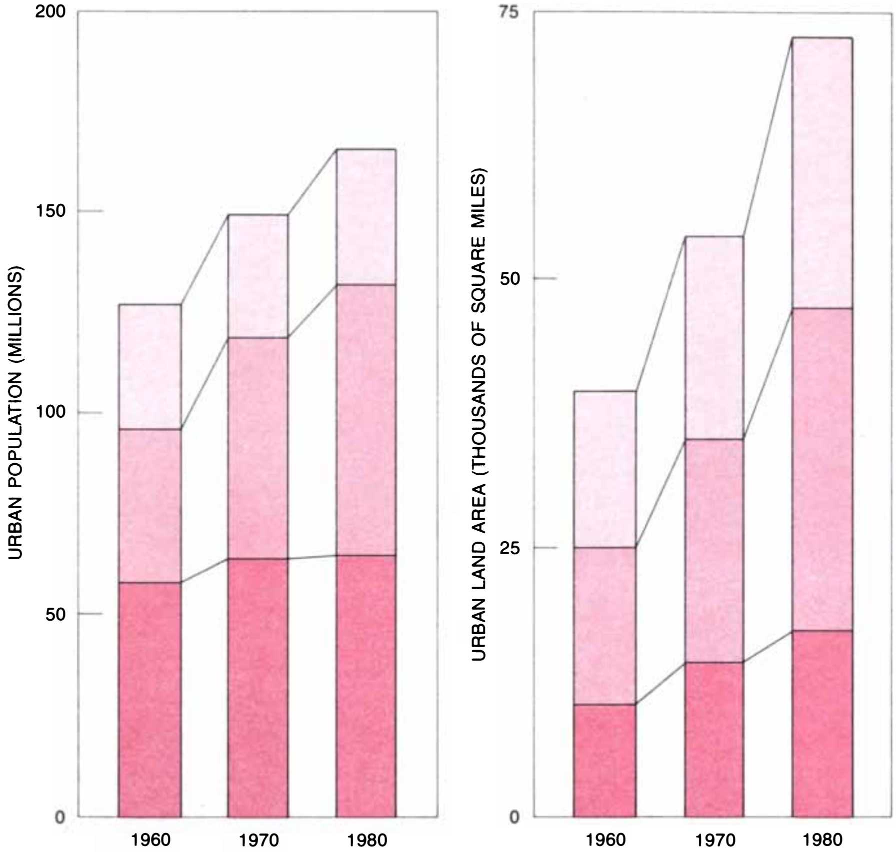
When we examined census results for individual states, we found that "ruralization," the increase in the fraction of the population in rural areas, was widespread and that it occurred in states with quite different levels of urban concentration. Ruralization occurred in 18 states in the 1970's, a substantial increase over the 1960's, when it occurred in only four. The 18 states include some where the population is growing rapidly, such as Texas and Oklahoma, and some where the total population is decreasing, such as New York. The process could be detected in both highly urban and highly rural states. Indeed, the reversal was so complete that there were only three states where the absolute size of the rural population decreased. Twenty-nine states had such losses in the 1960's, and most of the 29 had also lost rural population in the 1950's.

A vivid example of the pervasiveness



LEVEL OF URBANIZATION, or the fraction of the population in the urbanized areas, has almost stopped increasing. If the population of a small town grows enough, the town becomes an urbanized area. The subdivisions of the bars correspond to urbanized areas added after the censuses of 1950, 1960, 1970 and 1980. Between 1970 and 1980 the total population of the urbanized areas continued to grow (*left*). Because of renewed population growth in the rural areas, however, the fraction of the population in the urbanized areas increased little (*right*). The fraction would have decreased except for rapid growth in the urbanized areas added in 1980.

of rural demographic growth in the 1970's is provided by the section of the Eastern Seaboard often referred to as Megalopolis. The term Megalopolis was coined by Jean Gottmann of the University of Oxford in a book of the same name published in 1961. Gottmann defined the region as extending from southern New Hampshire to northern Virginia and including Boston, New York, Philadelphia, Baltimore, Washington and the smaller metropolitan areas in between. The area Gottmann specified also includes the rural areas on the fringes of the urban centers.



GEOGRAPHIC COMPONENTS of urban territory have been affected quite differently by recent population trends. Central cities (*dark color*) are the major cities included in the urbanized areas. The suburban fringe (*intermediate color*) makes up the remainder of an urbanized area. Small cities and towns (*light color*) are those with a population of from 2,500 to 50,000. From 1960 to 1980 the total urban population increased by 33 percent (*left*). The total urban land area increased by 84 percent in the 20-year period (*right*). The result was a sharp decrease in the population density of urban territory. The population and land area of the suburbs increased as a proportion of the total. The population and land area of the other types of urban territory (the central cities and the small cities and towns) decreased as a proportion of the total.

In 1960, 37 million people, a fifth of the nation's population, lived within the boundaries of Megalopolis. Surprisingly, between 1970 and 1980 Megalopolis stopped becoming more urban. Its rural population grew 10 times as fast as its urban population and the urban population decreased from 87 to 86 percent of the total. Demographic growth on the fringe of the urban settlements continued, but the central cities lost population. Population growth in the rural areas accelerated moderately, enough for it to surpass the overall urban rate.

It should not be assumed that the renewed demographic vitality of rural areas is due to an increase in the number of people who live and work on farms. On the contrary, the number of people who live on farms continued its longstanding decline in the 1970's. Unlike the count of population, data for employment are not generally tabulated according to urban and rural categories. The Census Bureau does, however, compile annual data on the number of jobs in each county. The information is collected from the reports employers file with the Social Security Administration and the Internal Revenue Service.

We classified each county according to whether it had an urban settlement in it. A county was considered rural if it included no settlement with a population of 2,500 or more. A county was considered metropolitan if it included a city with a population of 50,000 or part of the suburbs of such a city. Between the two extreme categories of metropolitan and rural counties are many counties whose largest settlement is one of the "other urban" areas. When the counties had been classified, we counted the new jobs brought into being in each one by the renewed economic activity that followed the recession of 1974–75.

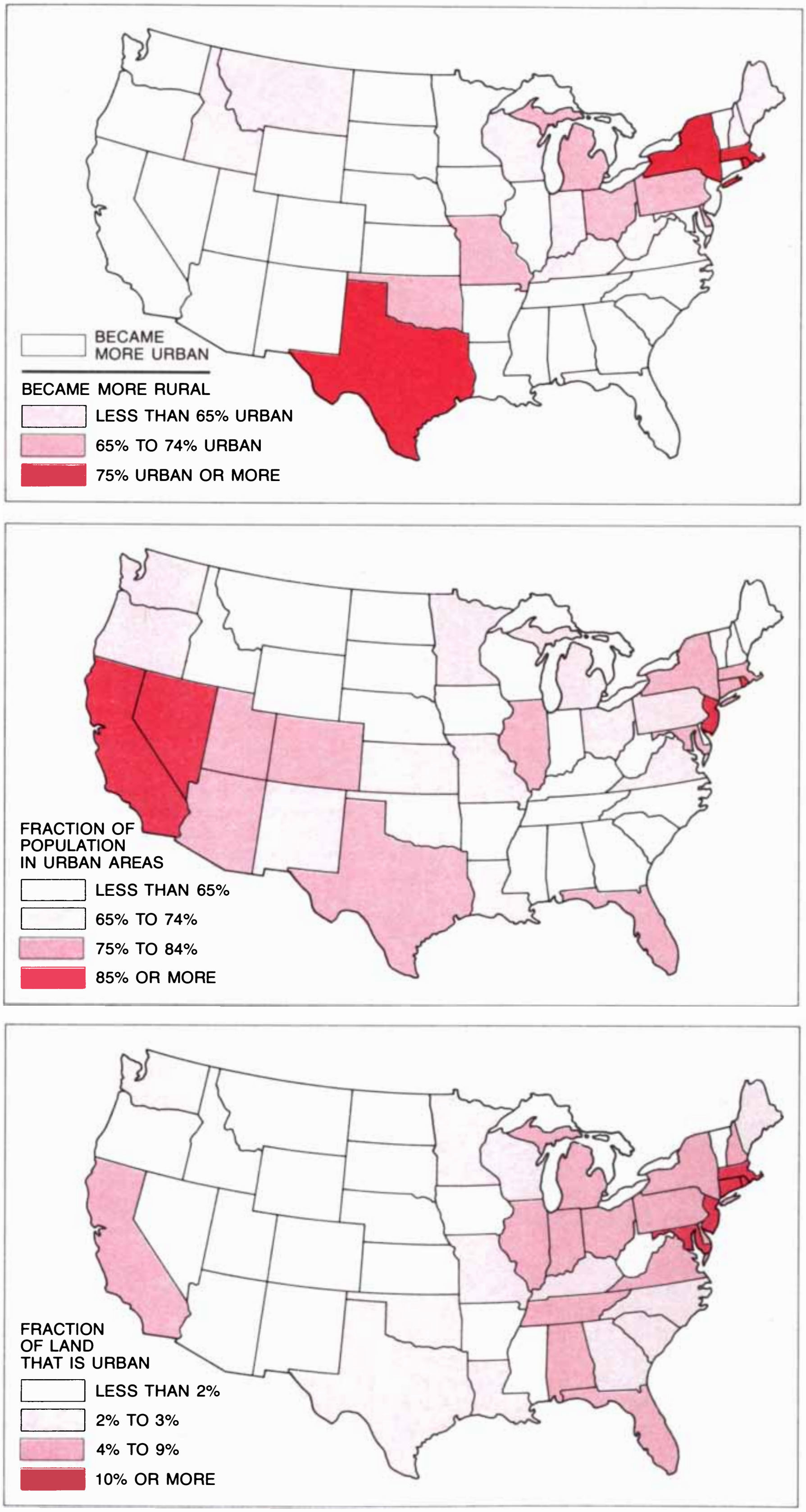
Our analysis suggests that one of the major factors in the deconcentration of the U.S. population is the fact that new jobs are being added as rapidly in rural areas as they are in urban areas. Overall the rate of increase of employment was greater in the rural counties. Furthermore, the rural rate was greater in many kinds of businesses, including manufacturing companies. In some regions and in some industries the rate of growth in jobs was far greater in the rural counties than it was in the metropolitan counties.

Thus economic forces could be responsible for much of the deceleration of the growth of the urban population. Although population growth in the urbanized areas has slowed, the areas have continued to expand rapidly in a geographic sense. The disjunction between demographic and geographic growth has led to significant changes in patterns of land use and population density.

Employing the broad definition of what is urban, about 14,400 square miles were converted from rural to urban uses in the 1960's; in the 1970's about 18,700 square miles were converted. Some 74,000 square miles, or 2 percent of all land in the U.S., is now covered by cities, suburbs and small towns. It is readily apparent that the increase in urban land area has been much faster than the increase in urban population. From 1960 to 1980 the area of urban settlement increased by 84 percent while the urban population increased by only 33 percent. What is more, the disparity between the demographic and the geographic rates is increasing. In the 1960's the rate of spatial expansion was not quite twice the rate of population growth; in the 1970's the spatial rate was three times the demographic one.

If the area of urban settlement is increasing faster than the urban population, it follows that the density of the urban population is decreasing. Before turning to the effect of such a trend on the components of the urbanized areas, we shall briefly consider the relation between population and geographic area at the state level. Analyzing urban growth in terms of both land use and population can provide a necessary corrective to an exclusively demographic approach. For example, the state with the greatest fraction of its population in urban areas is California. The West is highly urbanized, but California is exceptional even in the region: 91 percent of its population is in urban areas, broadly defined. Geographically, however, California is rural: 96 percent of its land is outside the urban areas.

New Jersey presents a contrasting picture. The fraction of the state's population that is urban is about the same as it



RURALIZATION, or an increase in the fraction of the population in rural areas, occurred in 18 states in the 1970's (*top*). This was a substantial increase from the 1960's, when only four states became more rural. The 18 were in all four census regions; they included both highly urban and highly rural states. Geographic and demographic measures of urbanization yield quite different pictures of the states. The demographic measure is the fraction of the state population in cities and towns (*middle*). By this measure California is the most urban state. The geographic measure is the fraction of the state's land area covered by cities or towns (*bottom*). By this measure California is much less urban than New Jersey, the most urban of the states.

is in California (89 percent) but a smaller fraction of land is rural (71 percent). Hence New Jersey is the most densely populated state. Its residential density of 986 people per square mile is greater than that of Belgium, one of the most densely populated nations in Europe.

When one considers a typical New Jersey landscape, it might seem that much more than 29 percent of the state's land is urban. The definition of what is urban, however, refers to residential density and therefore excludes certain areas that might be counted as urban by other standards, such as land given over to factories, oil refineries, storage facilities, industrial parks and highways that lie outside urbanized areas and the corporate boundaries of towns or cities with more than 2,500 inhabitants. The land under such facilities is rarely considered rural, but it is not urban in the demographer's sense because it provides homes for very few people.

As we observed at the beginning, urban growth and industrial development have historically been closely connected. In the traditional pattern of industrial development factories were built in or at least near urban centers. The factors that affect where industrial plants are constructed have apparently been changing in recent years. The result has been the rapid growth of industry in rural settings. The people who moved to the rural areas to take up the new jobs there, however, do not conform to the ordinary image of the rural population. Rather than being farm workers, they are in many instances either workers with the kind of technical training required in modern industry or professionals. The new rural demographic concentrations appear to represent small centers of urban culture transplanted to the countryside and enabled to survive by recent advances in communications, transportation and methods of industrial production.

There has been a substantial decline in the average population density of the urban areas. The density of urban places in the broad sense decreased from 3,200 people per square mile in 1960 to 2,300 in 1980. The decline was faster in the 1970's than it was in the 1960's. Since 1950 the average density of urbanized areas has been reduced by almost half.

The reduction in density has two causes. First, existing urban places have become less densely populated. In addition new urban areas generally have a lower density than older ones. The reduction in urban population density together with the growth of the rural population has led to a slight narrowing of the differential in density between the urban and the rural populations. The rural population, however, with a density of 17 people per square mile, is still much sparser than the urban population.

Declines in urban population density

were widespread in both the 1960's and the 1970's. In the 1960's 81 percent of all urbanized areas became less densely populated; in the 1970's the fraction was 87 percent. The exceptions were for the most part cities where geographic circumstances prevented further expansion. For example, the population density of Honolulu increased because territorial growth is made difficult by mountains and the seacoast. The density of the urbanized area that includes Miami, Fort Lauderdale and West Palm Beach increased because the three communities have merged and population increases must be accommodated by the filling in of vacant suburban land.

The coalescence of the three Florida cities has resulted in the formation of an urban mass with a population of more than three million covering more than 800 square miles. The formation of this mass exemplifies a process that continued in the 1970's. As the urbanized areas have expanded geographically they have in many instances reached the boundaries of adjacent cities. The contiguous urban areas then join to form amorphous urban regions covering large amounts of land.

Unlike Megalopolis, which includes a substantial rural periphery, these supercities represent continuously built-up urban territory. There are now 29 urban centers in the U.S. with a population of a million or more. About half of them were created by the fusion of two or more preexisting cities. The supercities are not recognized as legal or political entities, although they are interdependent in planning for the provision of such services as public transportation, sewage and water.

The supercities represent a demographically vital part of urban territory. Even as the rate of urbanization decreased, the proportion of the population living in the supercities increased. In 1970 there were 24 supercities with a population of 76 million, or 37 percent of the U.S. population. In 1980 the population of the 29 supercities was 89 million, or 39 percent of the total. The supercities have expanded by incremental growth at their suburban fringes, by

SUPERCITY	1980 POPULATION	LAND AREA (SQUARE MILES)	POPULATION DENSITY (PEOPLE PER SQUARE MILE)
NEW YORK CITY/ NORTHEASTERN NEW JERSEY/ SOUTHERN CONNECTICUT	17,606,680	3,656	4,816
LOS ANGELES/LONG BEACH/ SAN BERNARDINO/RIVERSIDE	10,184,611	2,187	4,657
CHICAGO/NORTHWEST INDIANA/ AURORA/ELGIN/JOLIET	7,212,778	1,675	4,306
PHILADELPHIA/ WILMINGTON/TRENTON	4,779,796	1,270	3,764
SAN FRANCISCO/SAN JOSE	4,434,650	1,122	3,952
DETROIT	3,809,327	1,044	3,649
BOSTON/BROCKTON/LOWELL/ LAWRENCE/HAVERHILL	3,225,386	1,084	2,975
MIAMI/FORT LAUDERDALE/ HOLLYWOOD/WEST PALM BEACH	3,103,729	816	3,804
WASHINGTON, D.C.	2,763,105	807	3,424
HOUSTON/ TEXAS CITY/LAMARQUE	2,521,857	1,169	2,157
CLEVELAND/AKRON/ LORAIN/ELYRIA	2,493,475	989	2,521
DALLAS/FORT WORTH	2,451,390	1,280	1,915
ST. LOUIS	1,848,590	597	3,096
PITTSBURGH	1,810,038	713	2,539
SEATTLE/ EVERETT/TACOMA	1,793,612	672	2,669
MINNEAPOLIS/ST. PAUL	1,787,564	980	1,824
BALTIMORE	1,755,477	523	3,357
SAN DIEGO	1,704,352	611	2,789
ATLANTA	1,613,357	905	1,783
PHOENIX	1,409,279	641	2,199
TAMPA/ST. PETERSBURG	1,354,249	527	2,570
DENVER	1,352,070	439	3,080
CINCINNATI/HAMILTON	1,228,438	465	2,642
MILWAUKEE	1,207,008	496	2,433
NEWPORT NEWS/HAMPTON/ NORFOLK/PORTSMOUTH	1,099,360	614	1,790
KANSAS CITY	1,097,793	589	1,864
NEW ORLEANS	1,078,299	230	4,688
PORTLAND	1,026,144	349	2,940
BUFFALO	1,002,285	266	3,768

SUPERCITIES, the urban masses with a population of a million or more, include both incorporated areas and built-up suburban territory. In many cases they have been brought into being by the merging of two or more large cities. In 1970 there were 24 supercities; by 1980 there were the 29 shown here. The development of the supercities shows that the peaking of urbanization need not halt the development of very large urban centers. All the supercities increased in area during the 1970's, but several lost population (*bold type*). Although some cities within the boundaries of large urban complexes have been losing population for decades, not until the 1970's did any of the large merged urban masses lose residents.

absorbing smaller cities and by fusion with cities of equal size. The process by which such places have developed indicates that the slowing of urbanization need not entail the disappearance of large urban centers.

Examining the continuously built-up urbanized areas helps to overcome the difficulties inherent in taking as the unit of analysis the city proper, cities whose boundaries can be changed only with difficulty. The very large urbanized area in New York, northern New Jersey and southern Connecticut that includes New York City is still the nation's largest supercity. Its population in 1980 was 17.6 million, and although it lost people during the decade, it continued to expand geographically. The importance of the distinction between a city and an urbanized area can be shown by comparing Chicago with Los Angeles. Chicago itself has more residents than the city of Los Angeles, but the urbanized area around Los Angeles surpassed the one around Chicago in population during the 1950's. By 1980 urban development had joined Los Angeles, Long Beach, San Bernardino, Riverside and many smaller places. The result is an area that is geographically as well as demographically larger than the urban nexus that includes Chicago, Aurora, Elgin and Joliet in Illinois and Gary in Indiana.

One tends to think of southern California as a more or less suburban area with a low population density. The Los Angeles supercity, however, has a population of about 10 million on 2,000 square miles of land. Its density of 4,700 people per square mile is greater than the densities of the supercities around Chicago, Philadelphia, Boston and several other older urban areas.

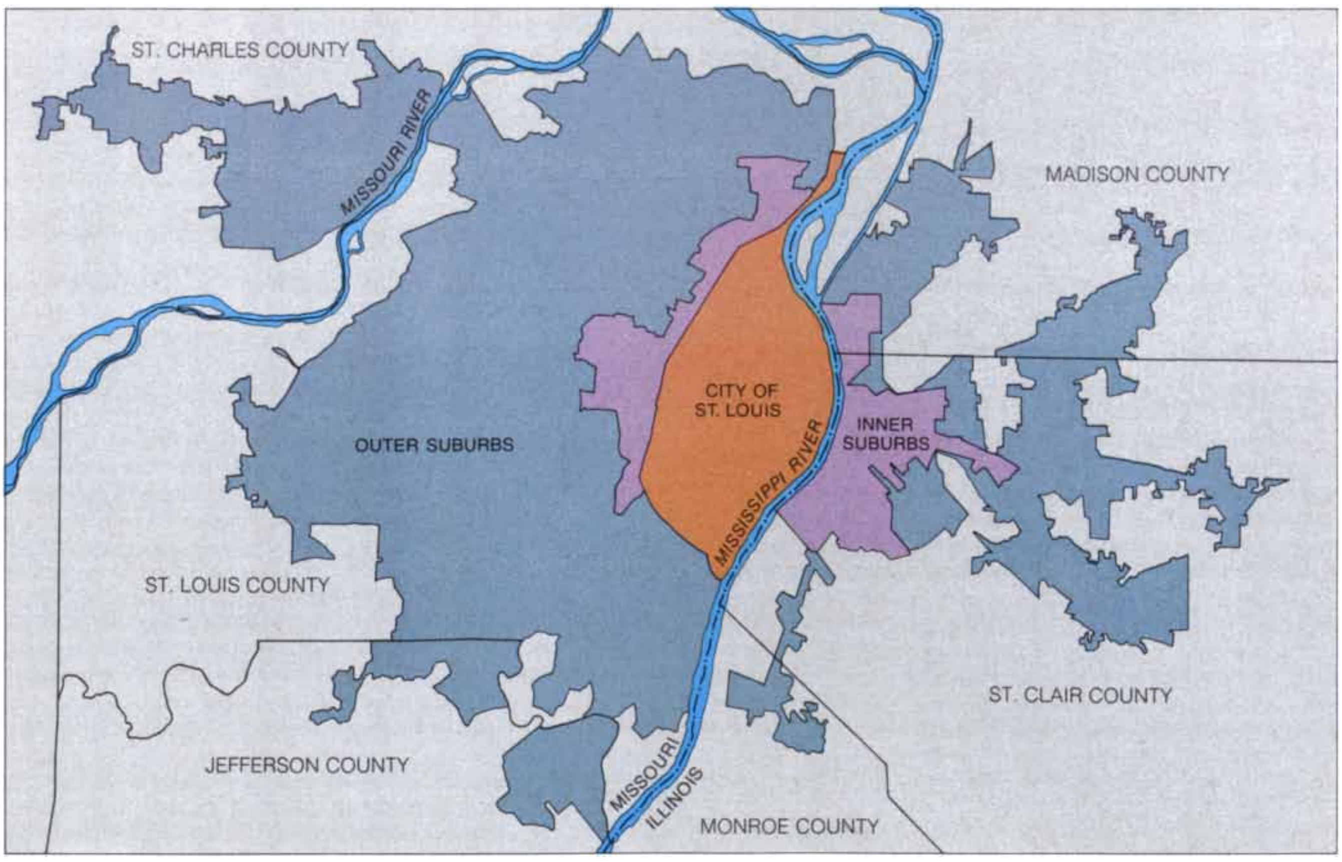
Density is high in California because the population is growing and spatial expansion is limited by the coastline, the mountains and the large areas that are already taken up by dense settlement and highways. An observer may be less conscious of the density of population in southern California than he is of that in the other regions because the steep gradients of population density that distinguish cities in the Northeast or the Middle West from their suburbs do not exist on the West Coast. Nevertheless, the urbanized part of southern California is more densely populated on the average than many major urban centers in the Northeast and the Middle West are.

The changes in density and geographic area in the 1970's have affected all parts of the urbanized areas from the central cities to the outer suburbs. The changes, however, have not affected the components in the same way: a disproportionate amount of growth in both population and area has taken place in the suburbs. In 1960, 30 percent of the total urban population lived in the suburbs, which occupied 36 percent of urban land. By 1980, 41 percent of urban residents lived in the suburbs, which occupied 42 percent of urban land. In the same period the fraction of the urban population living in the central cities decreased, and so did the fraction living in small cities and towns.

The increase in the suburban share of the urban population has been even more rapid in the older cities than the national data suggest. In 1950 about 30 percent of the population of the urbanized areas lived in the suburbs; in 1980 more than 50 percent did. The increase has been slower in the newer urbanized areas, partly because in many instances a new city can annex surrounding territory. Many older cities are surrounded by incorporated areas that can resist annexation; hence geographic expansion of the central city is limited.

Because the suburbs were areas of relatively rapid demographic growth in the 1970's, their population density decreased less than that of other places. The average density of the suburbs decreased by 15 percent, that of the central cities by 17 percent and that of the small cities and towns by 18 percent. The changes were too small to affect the overall ratio of population density between the cities and the suburbs, which was 1.7 in both 1970 and 1980.

The fact that the ratio of population density in the cities and suburbs did not change conceals some significant demographic trends. In many urbanized areas the central city has been losing popula-



tion for decades. In general the urbanized area as a whole has gained population because the growth in the suburbs has compensated for the losses in the central city. The 1980 census has shown, however, that in many cases an entire urbanized area is now losing population. One reason is that the rate of population loss in the central city has increased. A second reason is that the suburbs closest to the central city have begun losing population.

The urbanized area that includes St. Louis provides a good example of the process. The city of St. Louis lost population in both the 1960's and the 1970's; the rate of loss was higher in the 1970's. The losses have been so great and so sustained that the city now has roughly the same number of residents it had in 1890. Like most other cities in the Middle West and Northeast, St. Louis has long been surrounded by incorporated areas. These inner suburbs began to lose people in the 1960's, and their population declined by about 15 percent in the 1970's. Beyond the ring of inner suburbs population is still growing, but at a sharply diminishing rate. St. Louis' outer suburbs grew by 103 percent in the 1950's, by 55 percent in the 1960's and by 17 percent in the 1970's. Other Middle Western and Northeastern cities show a similar pattern.

The image of a doughnut is often employed to describe an urbanized area with an older central city whose population is decreasing. The image is not new; such population losses have been going on for decades. What is new is the expansion of the region of population loss (the "hole" in the doughnut) to include the inner suburbs. The hole has grown largest in the urbanized areas where the doughnut pattern has prevailed longest.

Moreover, although the doughnut has long been a familiar image in the industrial centers of the Northeast, in the 1970's it spread for the first time to much of the rest of the U.S. About half of all urbanized areas show the pattern, including a fourth of the cities in the South. The city of Atlanta, after growing in the 1950's and 1960's, lost population in the 1970's at a rate greater than that for Newark, N.J.

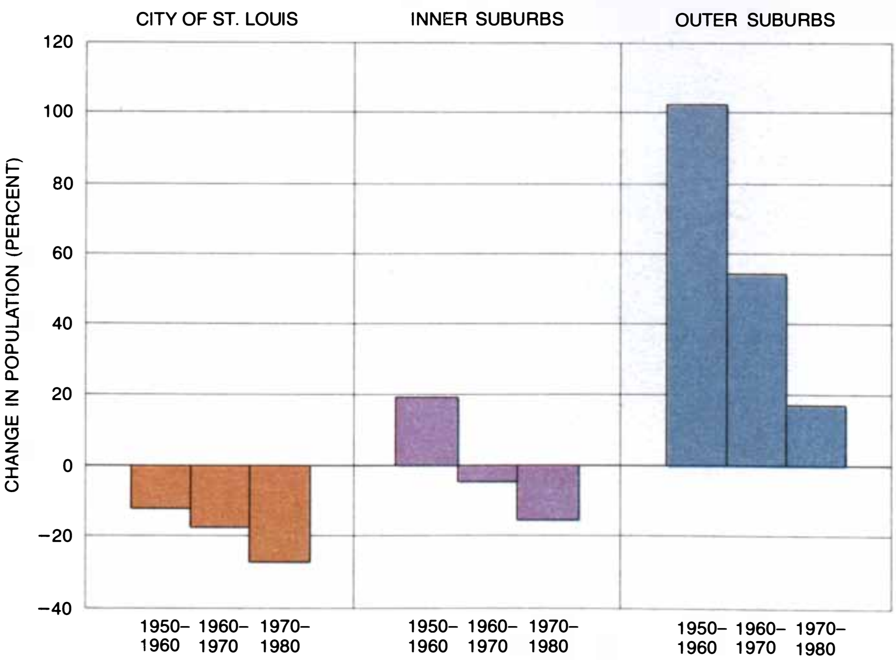
Population losses in the central cities and inner suburbs have become so great that an increasing number of urbanized areas are losing population. In the 1950's only one area in 20 lost residents; in the 1970's one in six did. Most such urban population declines are invisible to the casual observer. Indeed, in the light of "gentrification" (the movement of young professionals into inner-city neighborhoods) and new housing construction in the outer suburbs many people will find it hard to believe the population of the entire urbanized area around St. Louis declined by 2 percent in the 1970's.

In spite of the fact that they are losing population, many older cities and urbanized areas are gaining households. The number of households in the St. Louis urbanized area increased by 12 percent in the 1970's. What makes such apparently contradictory developments possible is the fact that the average number of members of a household is declining. The fertility of married women has decreased, and changes in living arrangements have been brought about by divorce, separation and the virtual disappearance of the extended family living in one dwelling. All these factors have contributed to a decrease in the size of the average household.

For this reason some of the most significant demographic trends of the past three decades in urban areas have been largely invisible. A count on a large scale, such as the census, is needed to measure them precisely. Our purpose in this article has been to analyze such trends by employing definitions that can be applied consistently over the entire period. The definitions themselves, however, do not remain static. Based on the results of the 1980 census, demographers at the Census Bureau have already begun to redefine the urbanized area. Urban areas are no longer limited to the territory around cities with a population of 50,000 or more. As of 1980 an urbanized area is any group of 50,000 people with an average density of 1,000 per square mile, regardless of the size of the largest city. Thus many people who live in or near a small town have become part of the urbanized-area population.

The change is not particularly surprising. Almost every redefinition of what is urban has expanded the number of people in the category. Indeed, as urbanization reaches its historical peak some uncertainty about who is an urban dweller is almost inevitable. The inability of demographers to agree on what is urban territory may be taken as a sign that urbanization has peaked at a high level.

The uncertainty undoubtedly extends even further than we have proposed here. For some time almost the entire population of the U.S. has been urban in the sense of having access to the amenities of urban life whether or not they live in cities. Perhaps the best model of the population is a continuum in which one end represents the most nearly urban part and the other represents the most nearly rural part. In the 1970's there was an unexpected increase in the demographic growth of the most nearly rural parts of the nation; at the same time the rate of loss increased in the most nearly urban parts. The fundamental cause of the reversal appears to be economic: population has followed jobs. The future distribution of population in the U.S. could well be determined by which regions are able to compete for new jobs both within the U.S. and in international markets.



ST. LOUIS URBANIZED AREA shows a pattern that prevails in many older cities. The boundaries of the city proper (*orange*), the inner suburbs (*purple*) and the outer suburbs (*blue*) are shown on the map at the left. The rate of population growth in each area is shown in the chart above. The city of St. Louis has been losing residents for decades and now has the same population it had in 1890. Many people who left the city moved to the surrounding suburbs. Such a pattern is often described as a doughnut. In the 1970's the area of population loss (the "hole" in the doughnut) spread outward: the inner suburbs lost a substantial fraction of their residents for the first time. The outer suburbs are still gaining population, but at a rapidly decreasing rate.

How Continents Break Up

They rift apart over millions of years, and in the process they deform. A study of the rifting is beginning to reveal the properties of the plates that compose the earth's crust

by Vincent Courtillot and Gregory E. Vink

The theory of plate tectonics holds that the crust of the earth consists of plates some 100 kilometers thick that are moving with respect to one another. It accounts for a wealth of facts. The distribution of earthquakes and volcanoes around the world is largely explained by the motions of the plates: their creation at midocean ridges; their destruction by collision or at subduction zones, where they plunge downward into the earth's mantle; their friction as they slide past each other along what are called transform faults. Moreover, the similarity in shape between the edges of two continents that today are separated by thousands of kilometers (for example the eastern edge of South America and the western edge of Africa) shows that they rifted apart. In other words, they were once on a single plate. Corroborating evidence is supplied by the distribution of plant and animal fossils on the continents.

If it is assumed that plates do not deform as they move, their relative motions can be calculated by means of mathematical theorems describing the motion of a rigid body on a sphere. Indeed, the motions of most of the earth's large plates over the past 150 to 200 million years have been worked out with considerable precision. On the other hand, the motions of plates that are assumed to be rigid cannot yield more than an understanding of their kinematics. The assumption that plates are rigid means it must also be assumed that their boundaries have a fixed geometry, and this latter assumption makes it impossible to propose a realistic model of how continental crust breaks up. For example, the breakup that created South America and Africa must be taken to have happened instantaneously along a line, a most unlikely event.

Actually there is ample evidence that continental crust does deform. Mountain ranges, which often result from the collision of plates, are the most obvious evidence. In addition the distribution of earthquakes at plate boundaries suggests that the boundaries are not narrow, and the occurrence of earthquakes well away from plate boundaries suggests the release of stress, and thus the occurrence of deformation, in the middle of a plate. The work we shall describe in this article suggests that the breakup of continental crust must be seen as a process taking millions of years along a zone several hundreds of kilometers wide. It suggests, moreover, that the crust at such a zone is unlikely to behave like anything rigid.

The key evidence supporting the theory of plate tectonics is found on the ocean floor in the form of magnetic anomalies. Fundamentally the volcanic material that rises into place to become new oceanic crust at the axis of a midocean ridge is hot enough for the minute magnetic domains in it to become aligned with the magnetic field of the earth. As the material cools, locking the alignment into place, it spreads away from the ridge axis. Behind it along the axis newer crust is rising into place. Now, the magnetic field of the earth varies in an irregular pattern, and its polarity reverses approximately five times every million years. Therefore as the new oceanic crust spreads away from the ridge axis two divergent magnetic "tapes" record the pattern of reversals. This makes it possible to measure the age of the ocean floor and also determine the velocity at which plates have diverged. That velocity proves to vary from less than a centimeter per year to more than 15 centimeters. Moreover, when the ridge axis is offset by a series of transform faults, the magnetic pattern in the oceanic crust should reproduce the offset.

There are instances, however, where the magnetic pattern fails to match the offset. These mismatches were initially explained in terms of sudden jumps of the ridge to a new location. The problem is that some of the jumps are so numerous and so closely spaced in time that the process is best seen as the continuous evolution of the boundary of a plate. Consider the peculiar *V*-shaped magnetic patterns found in the Pacific, first near the Galápagos Islands and then on the Juan de Fuca Ridge off the Pacific Coast of North America. A study of those patterns led Richard N. Hey, who was then working at the Hawaii Institute of Geophysics, and his colleagues (particularly Frederick K. Duennebier of the Hawaii Institute, W. Jason Morgan of Princeton University and Peter R. Vogt of the U.S. Naval Research Laboratory) to develop a model in which a rift propagates continuously through the oceanic crust at a rate comparable to the rate at which plates spread apart at the rift.

As Hey and Tanya M. Atwater of the University of California at Santa Barbara explain it, the process occurs where two rifts are offset by a transform fault. If the process is modeled in steps, each step is a sudden jump in which the tip of the rift propagates forward, establishing a new fault and moving oceanic crust from one side of the rift (that is, from one plate) to the other [*see illustrations on next page*]. The result is a *V*-shaped

AFAR DEPRESSION in northeastern Africa is one of two places on the earth (the other is Iceland) where a midocean ridge comes up onto dry land. The axis of each such ridge is a rift where two great plates diverge and new oceanic crust rises into place. The positions of the rifts at Afar are controversial. According to the authors, one rift courses westward from the Gulf of Aden into the Gulf of Tadjurah, the beak-shaped body of water extending into the image from the right. Its tip, currently at Lake Asal just west of the Gulf of Tadjurah, is propagating into the depression and will ultimately join with a rift whose tip is moving southward in the Red Sea. (The mouth of the Red Sea is at the upper right.) The rifting will then have given rise to a new ocean basin. The image is a mosaic of six frames made in false colors by the Landsat satellites.

wake consisting in part of small pieces of crust bounded by abandoned faults. If the process is taken to be continuous, as it in fact is in the crust of the earth, the wake is a *V*-shaped discontinuity in the pattern of magnetic anomalies. Hey pointed out that the discontinuity can mistakenly be attributed to a fracture. (The attribution is mistaken because no slippage between plates ever occurred along the discontinuity.) He called each limb of the *V* a pseudofault.

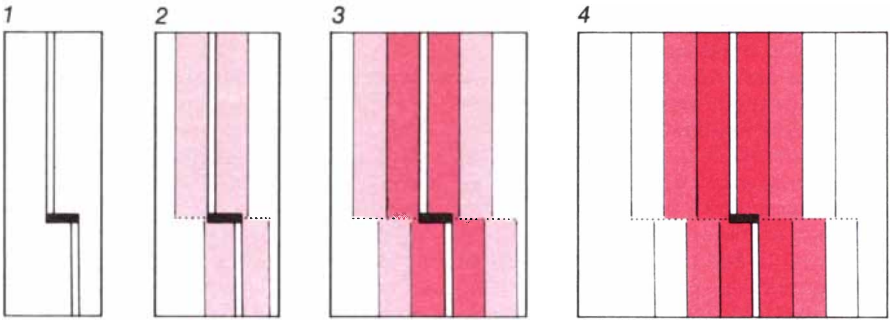
The breakup of continental crust must also involve the propagation of rifts, yet the details of the process have been elusive. In the early 1970's J. Tuzo Wilson of the University of Toronto and Kevin C. Burke and John F. Dewey of the State University of New York at Albany suggested that there might be a link between "hot spots," triple junctions and continental breakup. They suggested that if a continent becomes stationary with respect to the underlying mantle, thermal anomalies (hot spots) in the mantle might make the continent dome upward and rupture in a three-branched pattern of rifts. (Doming and rupture often seem to give rise to a three-branched pattern. They do it, for example, in the crust of a pie baking in the oven.) The three rifts might then develop into two active rifts and an aborted one. (Burke and Dewey thought they could see examples along the rifted margins of the Atlantic.) The active rifts propagating away from a number of junctions could eventually join up and thereby fragment the continent.

More recently two other proposals have been made. First, Morgan has proposed that rifting often occurs along a hot-spot track. Here the motion of a continent over a thermal anomaly in the underlying mantle gives rise to a line of volcanoes at the surface. Under the surface, meanwhile, the thermal anomaly thins and weakens the crust, making it prone to rift. Second, Hey and Vogt have suggested that a process resembling the model of a rift propagating through oceanic crust might account for some of the rifting on the continents. On the ocean floor, however, the model requires that one rift recede while another one advances. Thus the model can only modify the geometry of preexisting plate boundaries. It cannot apply to the breakup of a continent, a process in which a new plate boundary arises.

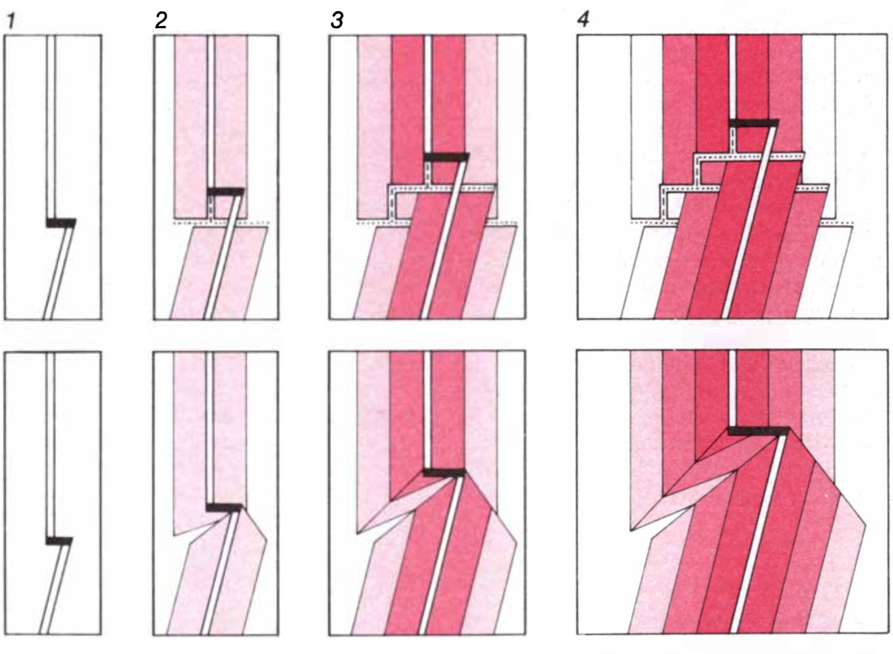
At this point magnetic anomalies are again important evidence. In 1977 one of us (Courtillot), working with Jean-Louis Le Mouël and Armand Galdeano of the University of Paris, made a magnetic map of part of the western Gulf of Aden and part of the adjacent land mass, the Afar depression in Djibouti and Ethiopia. Afar is a natural laboratory where sea-floor spreading can be studied without the hindrance of thousands of meters of overlying water. Like Iceland, it is an exposed part of the worldwide midocean-ridge system. Many investigators believe much can be learned about the fundamental process of rifting in these dry-land settings. As early as 1938 the French paleontologist and philosopher Pierre Teilhard de Chardin was convinced that Afar was the place where Alfred Wegener's theory of continental drift should be tested. More recently the investigators who dived in research submarines to the mid-Atlantic ridge at 37 degrees north latitude as part of the French-American Mid-Ocean Underwater Survey (FAMOUS) were struck by the resemblance of the rift at Afar to what they had seen under almost three kilometers of water.

The clearest clues to what has happened at Afar are the magnetic maps [*see top illustration on opposite page*]. The maps reveal that two types of crust are juxtaposed. In one type the magnetic anomalies are intense and form parallel bands. The pattern is typical of the one that results from sea-floor spreading. The bands can be correlated with the periods between reversals of the earth's magnetic field, and so they can be dated. In the other type of crust the magnetic anomalies are wider, rounder and less intense. They describe a magnetic quiet zone, or MQZ.

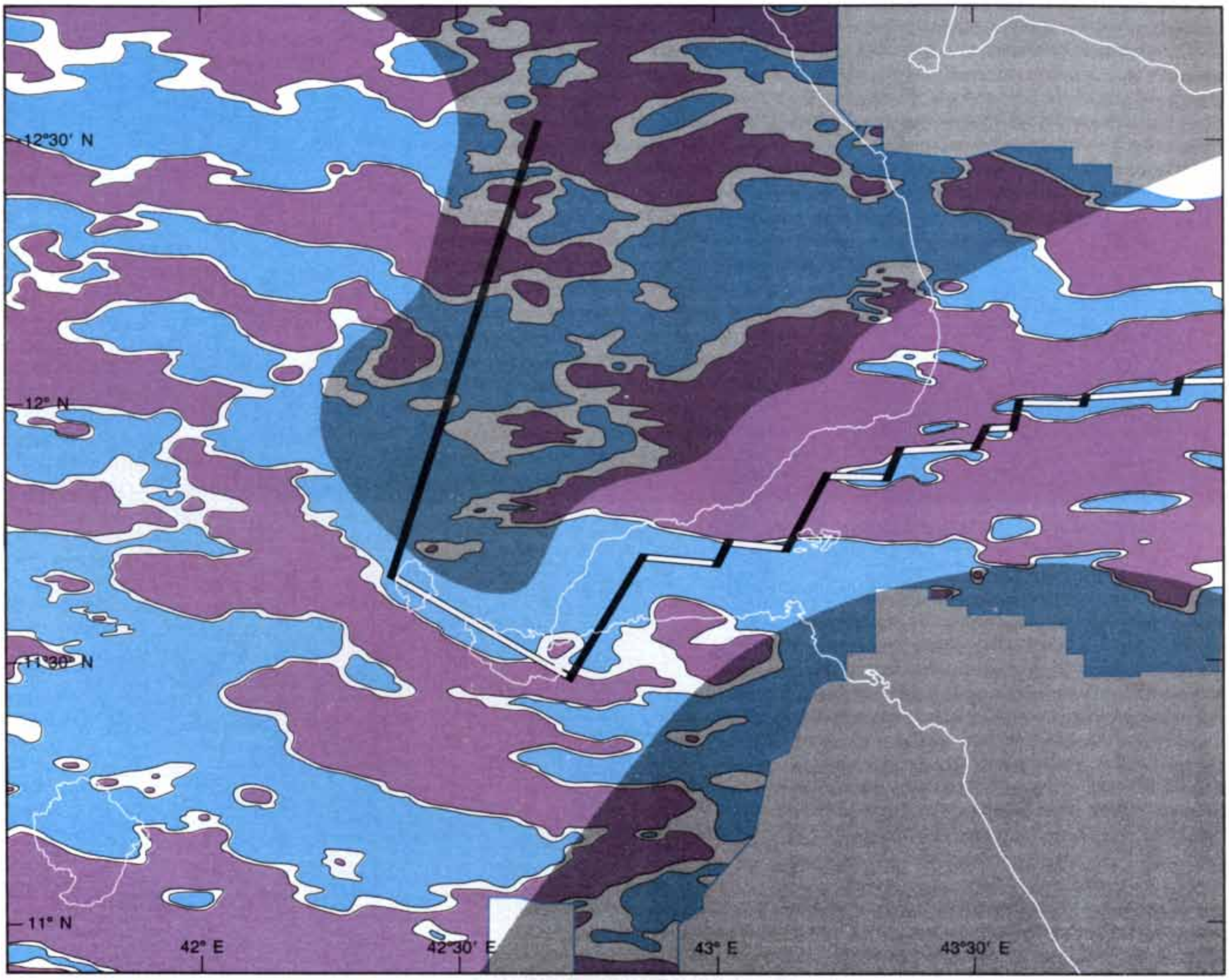
The MQZ forms a huge reclining *V*



MAGNETIC ANOMALIES on the ocean floor are the strongest evidence that new oceanic crust spreads away from each midocean rift. The rift itself (*1*) is offset by a transform fault (*black*). New crust arises continuously. At first it is hot; thus its intrinsic magnetization becomes aligned with that of the earth. As it cools and solidifies, the magnetization is locked into place; hence at each side of the rift a band of magnetized crust diverges (*2*). The earth's field occasionally reverses its polarity. Each such reversal entails two new magnetic anomalies (*3, 4*).

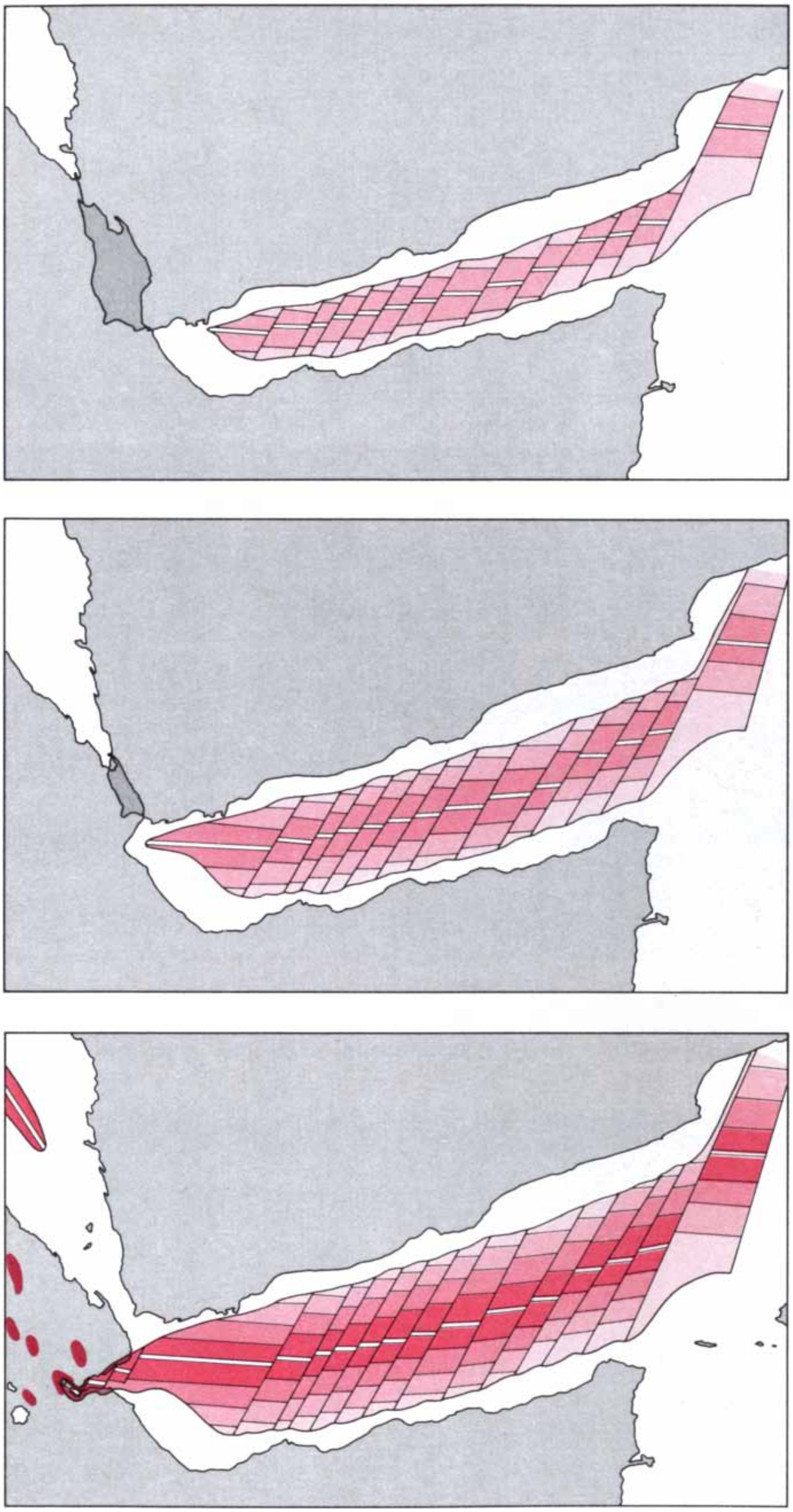


DISRUPTED PATTERN of magnetic anomalies is evidence that rifts propagate. In the top row a model proposed by Richard N. Hey of the University of California at San Diego and his colleagues is shown in steps. Two rifts at an angle to each other are offset by a transform fault (*1*). New oceanic crust rises into place and spreads; then the tip of the southern rift propagates northward, transferring a block of the new oceanic crust from east to west of the rift, that is, from one plate to another (*2*). Successive episodes of sea-floor spreading and rift propagation (*3, 4*) produce a *V*-shaped zone of magnetic discontinuity. In the bottom row of the illustration the process is continuous; thus the sea floor spreads while the rift is propagating.



MAP OF MAGNETIC ANOMALIES at Afar and on the floor of the Gulf of Aden reveals that different types of crust are juxtaposed. In the gulf the anomalies (shown in alternating blue and red) are intense and elongated, in a pattern typical of oceanic crust. Elsewhere, and particularly toward the northeast, the anomalies are subdued. They mark out a funnel-shaped region called the magnetic quiet zone, or MQZ (*gray*). The MQZ may represent continental crust deformed by the propagation of the rift along the floor of the gulf.

VIEW DOWN THE AXIS of the rift at the Gulf of Aden is afforded to the east from a point on the western shore of Lake Asal, 155 meters below sea level. The rift is at the center. It is flanked on each side by faults running parallel to the rift that produce a set of escarpments. The pattern is much like the one observed at midocean ridges from research submarines under almost three kilometers of water.



RECONSTRUCTIONS OF THE GULF OF ADEN as it looked 10 million years ago (*top*) and seven million years ago (*middle*) are made by taking the present-day map of magnetic anomalies on the floor of the gulf (*bottom*) and removing the anomalies that are younger than those ages. The direction of rift propagation (east to west) is plain. Overlaps between Arabia and Africa (*dark gray*) show that the plates stretched as they rifted. In the present-day map a number of active volcanic ranges (*solid color*) indicate deformation ahead of the tip of the rift. The Red Sea rift is at the upper left of the present-day map. It too is flanked by magnetic anomalies.

that abruptly interrupts the oceanic pattern. It therefore interrupts the bands of magnetic anomalies that mark sea-floor spreading; hence the edges of the MQZ can be dated rather precisely. Between 45 and 46 degrees east longitude in the Gulf of Aden the edges are about 10 million years old; they intersect the magnetic anomaly representing ocean floor of that age. Toward the west the edges become progressively younger. Near Lake Asal in Afar they are very young indeed. A number of earthquakes and the birth of a volcano near Lake Asal in November, 1978, confirm that the vicinity of the lake is a site of tectonic activity today.

The simplest interpretation of these various observations is that a rift is propagating westward into preexisting oceanic and continental crust. The edges of the MQZ were created by past propagation, which formed new oceanic crust in its wake. The Gulf of Aden is the result. The opening angle of the *V* formed by the edges is about 30 degrees. It arises by simple trigonometry from the rate of the sea-floor spreading that has opened the Gulf of Aden (1.5 centimeters per year) and the rate at which the tip of the rift has moved westward (three centimeters per year).

Support for this interpretation has come from a study of seismic activity in the region undertaken by Jean-Claude Ruegg and Jean-Claude Lépine of the University of Paris. They observe that the foci of most earthquakes under the Gulf of Aden cluster in a band approximately 10 kilometers wide that closely follows the rift. (The position of the rift is deduced from a mapping of the bottom of the gulf and from the position of the youngest magnetic anomaly.) The band of earthquake foci extends from the Gulf of Aden into the Gulf of Tadjurah, a beak-shaped westward extension of the Gulf of Aden. After that the band coincides with the Ghoubbet-Asal Rift, which runs from the Gulf of Tadjurah to Lake Asal. Apparently it ends near Lake Asal, where, we suggest, the tip of the propagating rift is currently situated. West of Lake Asal the seismicity becomes diffuse, suggesting that there the crust is deforming over a wider area.

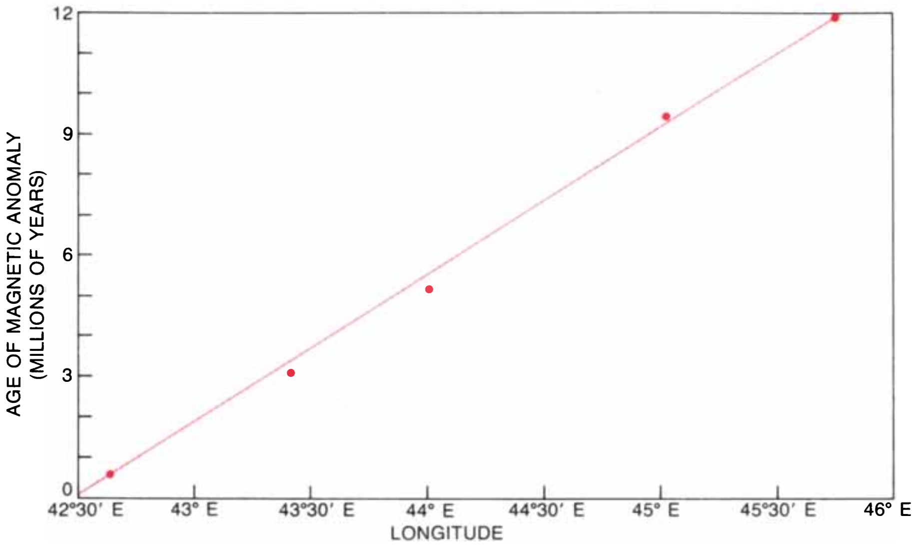
Further evidence comes from the age and the chemical composition of the basalts at Afar, as they have been determined by Michel Treuil of the University of Paris, Jacques Varet of the Bureau de Recherches Géologiques et Minières and Olivier Richard of the University of Paris at Orsay. Basalts are the volcanic rock that composes oceanic crust. The basalts found on the edges of the Gulf of Tadjurah have a tholeiitic composition: they are poor in the mineral olivine but contain the mineral orthopyroxene. Such basalts are formed by the large-scale partial melting of the mantle under a midocean ridge. Closer to Lake Asal and the tip of the rift the basalts are more alkaline, that is, they are richer in sodium and potassium. This indicates a lesser degree of partial melting than tholeiites undergo. The age of the basalts at each location closely matches the age deduced for the oceanic crust at that location from the measured magnetic anomalies.

In sum, the principal sign of a propagating rift at Afar is a distinctly oceanic pattern of magnetic anomalies that is disrupted by the intersection of the pattern with the rifted edge of a continent. The magnetic quiet zone at Afar is therefore best taken to be continental crust that has deformed in the course of the propagation.

Signs that rifts have propagated elsewhere in the world, so that continents deformed, emerge from reconstructions: the efforts made by earth scientists to determine how continental fragments fitted together before they rifted apart. The usual method is to try to match up their edges. Thus each fragment is rotated about an axis passing through the center of the earth until magnetic anomalies of the same age are made to coincide. If the edge of each fragment arose at a single instant in time, so that the edge is precisely parallel to the oldest magnetic anomalies on the ocean floor that has arisen between the fragments, the reconstruction will succeed: it will match the fragments precisely and at the same time yield the age of the rifting. If, however, the rift has propagated, the reconstruction will not quite succeed: the edge of each fragment will not have arisen all at once, and so the edge will cut across magnetic anomalies.

In such instances bringing together the edges of two continents will not reproduce the configuration they had before they rifted apart. A more accurate reconstruction will match the points where rifting began. The continents will then overlap in places, but that is quite appropriate. The overlaps represent the stretching of the continental crust that took place as the crust was broken apart by propagating rifts.

This appears to be what happened in the Gulf of Aden. If, for example, an attempt is made to reconstruct the relative positions of the African plate and the Arabian plate 10 million years ago (the age of what is called magnetic anomaly 5), it is found that the gulf closes up. Moreover, it is found that the plates begin to overlap at 43 degree 30 minutes east longitude. Ten million years ago the tip of the rift was there. This explains the observation, made by A. S. Laughton and his colleagues at the National Institute of Oceanography in England in 1970 and confirmed recently by James R. Cochran of the Lamont-Doherty Geological Observatory, that anomaly 5 and anomalies older than anomaly 5 cannot be found west of 43 degrees 30 minutes east longitude.



SPEED OF PROPAGATION of the rift in the floor of the Gulf of Aden is deduced from the reconstructions. Each point plots the westernmost limit of a magnetic anomaly (*horizontal axis*) and the age of the anomaly (*vertical axis*). The slope of the line connecting the points implies a velocity of three centimeters per year. The end of the line at an age of zero and a longitude of about 42 degrees 30 minutes demonstrates that the tip of the rift is now at Lake Asal.

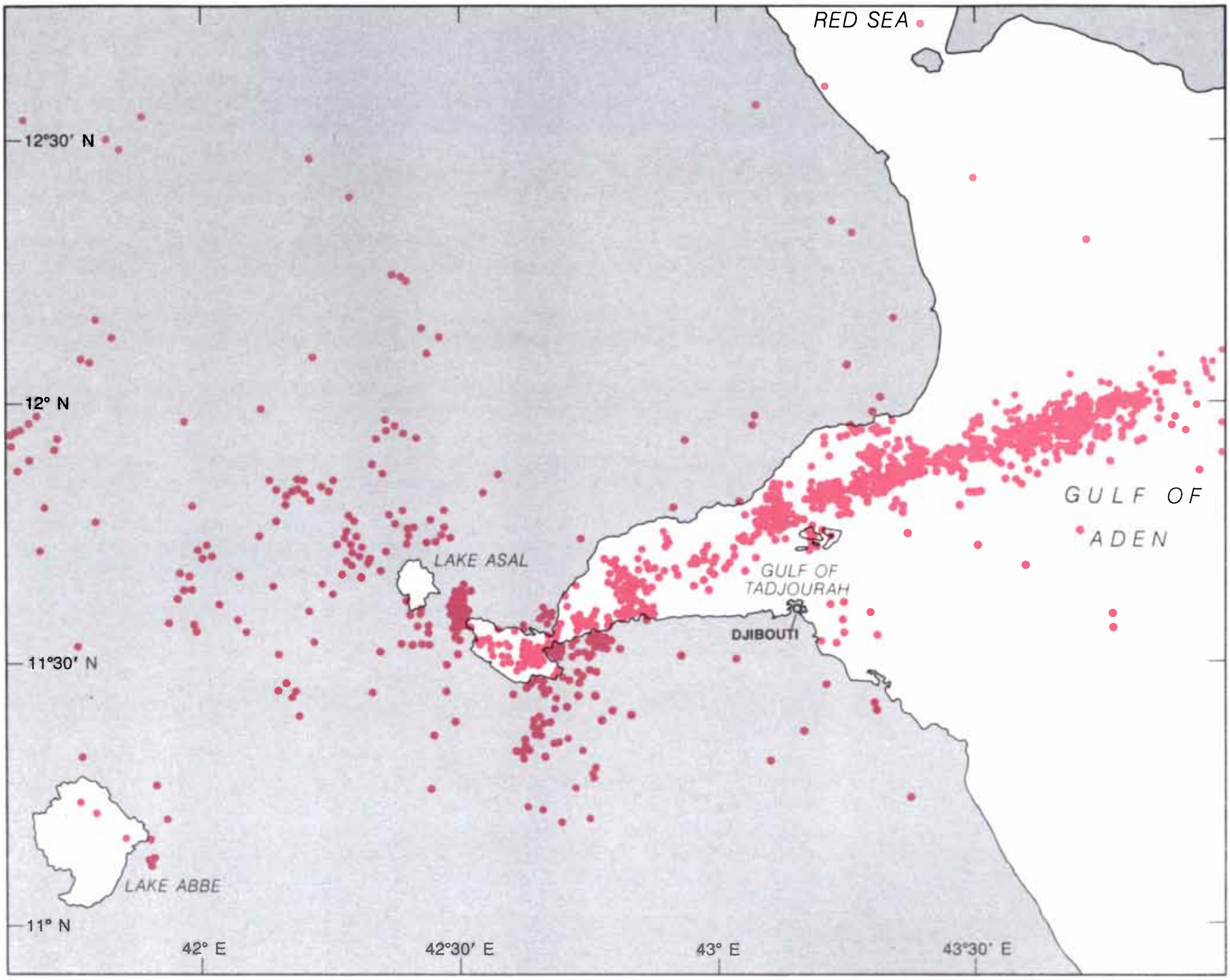
Overlaps also emerge in reconstructions of the North Atlantic made by one of us (Vink). Specifically overlaps between northern Greenland and Spitsbergen suggest that the edge of each of these bodies of land stretched by as much as 100 kilometers. Still another example emerges from the fit of South America and Africa. The classic reconstruction was done by Sir Edward Bullard of the University of Cambridge. Bullard used the 1,000-meter-depth contour lines in the Atlantic to represent the edge of each continent. He rotated the lines together. A computer program found the fit that minimized gaps and overlaps. Even so, there remained a gap of 250 kilometers between the southern tip of Africa and the edge of an undersea plateau that includes the Falkland Islands. The Deep-Sea Drilling Project of the Scripps Institution of Oceanography has since shown that the Falkland plateau is continental crust. It is a submerged part of South America. The gap cannot have existed. When it is closed in a reconstruction, there are overlaps between South America and Africa.

In any case magnetic surveys of the South Atlantic show that the oldest magnetic anomalies on the ocean floor do not parallel the edges of South America and Africa. The surveys confirm that propagating rifts created the South Atlantic. The oldest magnetic anomalies (one with an age of 135 million years near the edge of Africa and one with an age of 127 million years near the edge of South America) are found only in the southernmost part of the Atlantic. The floor of the Atlantic was therefore spreading in the south at a time when Africa and South America were still joined in the north. The rift had not yet propagated that far.

The available evidence suggests the following picture of how a rift propagates to break up a continent and form a new ocean basin. First a system of fractures develops in a plate, possibly along a preexisting weak zone such as an old suture zone (the locus of an earlier collision between continents), a large strike-slip fault system (a fault system where the slippage is horizontal) or, as Morgan has suggested, a hot-spot track. Let us assume for the sake of simplicity that the fractures lie in a linear zone, and let the average width of the zone be 200 kilometers, as it is in the Rift Valley of East Africa. As the fractures develop, the continental crust is stretched and thinned. Nevertheless, the relative motion of the plates that will emerge from the process is negligible: the total stretching is limited to a few tens of kilometers and occurs over periods of a few tens of millions of years.

As the stretching progresses, increasing lengths of the future plate boundary have decreasing strength. The mechanics of the process are unlikely, however, to be uniform. Thus the plates remain attached at a number of "locked zones." These zones cannot prevent the failure of the crust that lies between them. Hence basaltic oceanic crust begins to rise into place. In short, rifting begins. Each rift elongates rapidly through the thinned crust. Then the tips of the rifts enter the remaining locked zones and begin to break them apart.

Let us now consider a single locked zone. As in Hey's model of sea-floor spreading, a sequence of discrete steps is easier to explain than the actual contin-



FOCI OF EARTHQUAKES at Afar were recorded between 1974 and 1980 by investigators from the Institut de Physique du Globe in Paris. They confirm the position of the rift. Scattered foci west of the tip of the rift suggest that the crust there is now deforming.

uous process. In the first step the tip of a rift has reached a locked zone from each side [*see illustration on opposite page*]. Along the length of each rift new oceanic crust is being created. It fills the space freed as the two plates move apart. Inside the locked zone, however, no new crust develops. Therefore the locked zone must stretch.

In the next step the tips of the rifts penetrate an incremental distance into the locked zone. In this increment the crust of the locked zone stops stretching; the creation of oceanic crust takes over. Hence an increment of the locked zone is carried away by each plate. The part of the locked zone not yet invaded by the rifts continues to stretch. The process goes on until the rifts meet and the two plates separate completely. What remains of the locked zone is a deformation in the edge of each newly formed continent. The deformation is maximal at the point where the continents last touched.

The model can be refined so that the process is continuous. It then emerges, naturally enough, that if parameters such as the rate at which the rifts propagate change with time, the deformation of the edge of the continent produces various shapes. For instance, the shape of certain overlaps that arise in attempts to match the edges of continents that broke apart suggests that the propagating rifts may slow down as they first cut through a locked zone and then may accelerate in the final phase of the rupture. (The result in such a case is a deformed zone with a flattened tip.) More generally the model predicts that the strain in a continental plate that breaks up occurs in a band some 100 kilometers wide at the margin of each newly formed plate and also in the locked zones, which stretch dramatically. The disruption of the crust in a locked zone may be responsible for the lack of intense magnetic anomalies in the magnetic quiet zone at Afar.

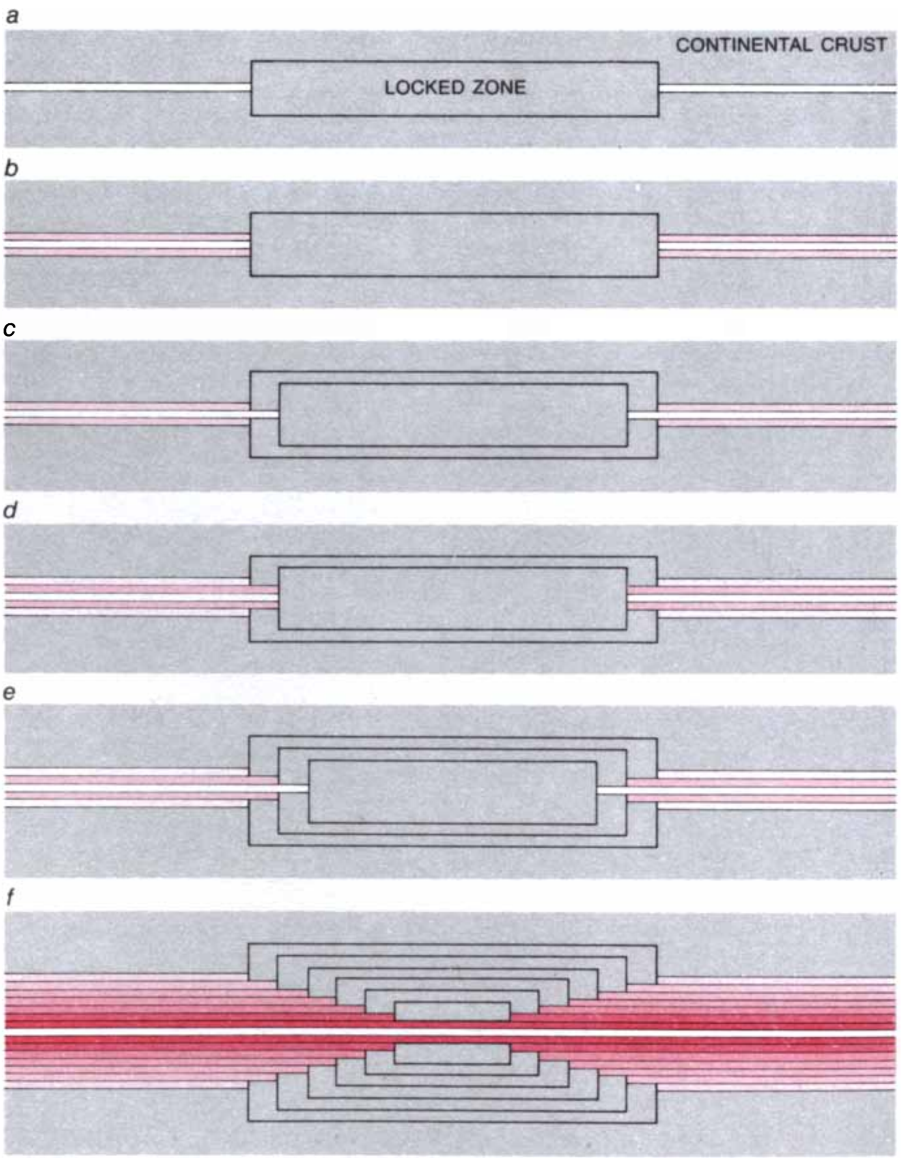
Yet if Afar is to be viewed as a locked zone in the act of deforming, a second propagating rift in addition to the one that opened the Gulf of Aden must be identified. The East African Rift Valley, which has a long history of very small motion, may be an aborting rift and is not a good candidate. The younger and more active southern Red Sea is more promising. Ron W. Girdler and Peter Styles, who were then working at the University of Newcastle upon Tyne, have attempted to identify magnetic anomalies in the Red Sea so that the seafloor spreading there can be related to the spreading in the Gulf of Aden. The effort has been complicated by the intricate faulted structure of the floor of the Red Sea and by the slow rate at which it is opening. So far Girdler and Styles have been able to recognize two probable phases of spreading, one between 30 and 15 million years ago and then a recent one beginning five million years ago that is continuing today. Magnetic mapping by Hans A. Roeser of the Federal Institute for Geosciences and Natural Resources in Hanover reveals at the Red Sea a *V*-shaped zone of oceanic anomalies quite like the one in the Gulf of Aden. The magnetic quiet zone at the Red Sea appears to correspond to the older phase of spreading.

The tip of the putative rift that is opening the Red Sea is now between 15 and 16 degrees north latitude. Its distance from the tip of the Asal-Aden rift is 300 kilometers. Presumably that is the length of the Afar locked zone remaining to be broken apart. A comparison of Afar with the model described above suggests that the locked zone was originally 750 kilometers long and 150 kilometers wide. Each of the propagating rifts has therefore torn through 200 kilometers of the zone. An extrapolation of the current rates of propagation suggests that the locked zone will break up completely in less than 10 million years. The zone will by then have undergone a mean deformation that will have stretched it by 80 percent of its original width. The last part of the zone to rupture will have stretched twice that amount.

How do the findings at Afar bear on the suggestion by Wilson, Burke and Dewey that rifting is related to the doming and rupturing of the continental crust above a hot spot in the mantle? Afar has repeatedly been considered one of the clearest cases of a triple junction caused by a hot spot. The junction is that of the Red Sea rift, the Gulf of Aden ridge and the Ethiopian rift (the northernmost part of the East African rift). Our evidence, however, indicates that the rifts at Afar are propagating toward, not away from, an eventual junction. One of us (Courtillot) has suggested that a first phase of tectonic activity at Afar may have generated a knee-shaped rift that opened the Red Sea and left the East African Rift Valley. Preexisting zones of weakness and heterogeneity in the composition of the crust might be responsible for the "knee." A second phase of activity, still in progress today, may then have opened the Gulf of Aden and produced the Red Sea's second episode of spreading. Again a knee-shaped rift appeared, but since the Afar locked zone is not yet cut through, the rifting is not complete. The Afar junction would thus be the superposition of two knee-shaped rifts.

Rifts that are propagating through a continent and are breaking it apart will doubtless be found in places other than Afar. The Gulf of California is a likely candidate, even though its relatively small scale and its great disruption by transform faults make the magnetic data difficult to interpret. The process of rift propagation is well worth understanding. For one thing, the rifting of continental margins and the subsequent deposition of sediments above them govern the evolution of oil-bearing formations. In addition the study of rifting may yield insight into the mechanical properties of continental crust. The pattern of rifting that separated South America and Africa, for example, suggests the locked zones typically have a length of 400 kilometers, a width of 150 kilometers and are spaced at a mean interval of 700 kilometers. The pattern further suggests that sea-floor spreading begins (that is, oceanic crust begins to rise into place at a rift) when about 65 percent of the length of the future plate boundary has lost its strength to resist the breakup.

In attempting to understand how a plate breaks up over a length of thousands of kilometers, we have dealt with temporal scales of no more than a million years and spatial scales of hundreds of kilometers. On such scales we find the continental crust deforming. Among the tools that facilitate the study of how the deformation proceeds two are crucial: the detailed examination of magnetic anomalies and the detailed examination of the edges of the continents. To the extent that the work goes beyond the assumption that plates are rigid it should provide a better view of how a plate evolves.



RIFTING OF A CONTINENT into two separate plates is shown in steps according to a model developed by one of the authors (Courtillot). The process begins along a preexisting zone of weakness. The crust, however, is heterogeneous and so a number of areas resist. One such "locked zone" is shown; it prevents two rifts from meeting (*a*). Along the length of each rift new oceanic crust rises into place; the new crust fills the space freed as the incipient plates move apart. The locked zone, in contrast, must stretch (*b*). The tips of the rifts then begin to invade the zone (*c*). Still more oceanic crust is created. Meanwhile what remains of the locked zone stretches again (*d*). The tips of the rifts move closer together (*e*). Ultimately the rifts meet and the plates separate (*f*). The locked zone has become a deformation in the edge of each new continent. The model can be made into a continuous one in which the size of the locked zone and the velocity of tips of the rifts change with time. Deformations with various shapes result.

The Salmonid Fishes as a Natural Livestock

The remarkable genetic adaptability of this family, which includes salmon and trout, is making it possible to replenish depleted stocks, to establish new ones and to adapt the fishes for "ranching" operations

by Lauren R. Donaldson and Timothy Joyner

Thousands of years ago the salmonid fishes, a group including salmon and trout, flourished throughout the Northern Hemisphere in cool lakes, streams and rivers and in the open sea. In historic times the hand of man has changed this picture through habitat-destroying activities and heavy fishing, with the result that the salmonids have disappeared from many bodies of water and in others their numbers have greatly diminished. Several recent developments, however, offer excellent opportunities for stemming or even reversing the downward trend. The developments rest on the remarkable genetic adaptability of the salmonids and the ability of fish breeders to turn that adaptability to advantage.

The 60-odd species of the salmonid family include thousands of races. Each race is adapted to a unique environment. This diversity can provide fish breeders with an enormous pool of genetic material from which desirable traits can be selected. They include size and shape; rate of growth; the color, taste and texture of the flesh; tolerance of fresh or salt water, and the timing and patterns of migration.

Drawing on these properties, breeders so far have achieved several significant results. Among them are: (1) Salmonids can be induced to "home," or return to spawn after a period at sea, in any cool body of water where they were released as fingerlings, which means that they can be established or reestablished in bodies of water that do not now have them; (2) some salmonid species will adapt to year-round life in fresh water even though they would normally spend part of their lives at sea, and (3) some species can be successfully "farmed" or "ranched" by techniques of aquaculture.

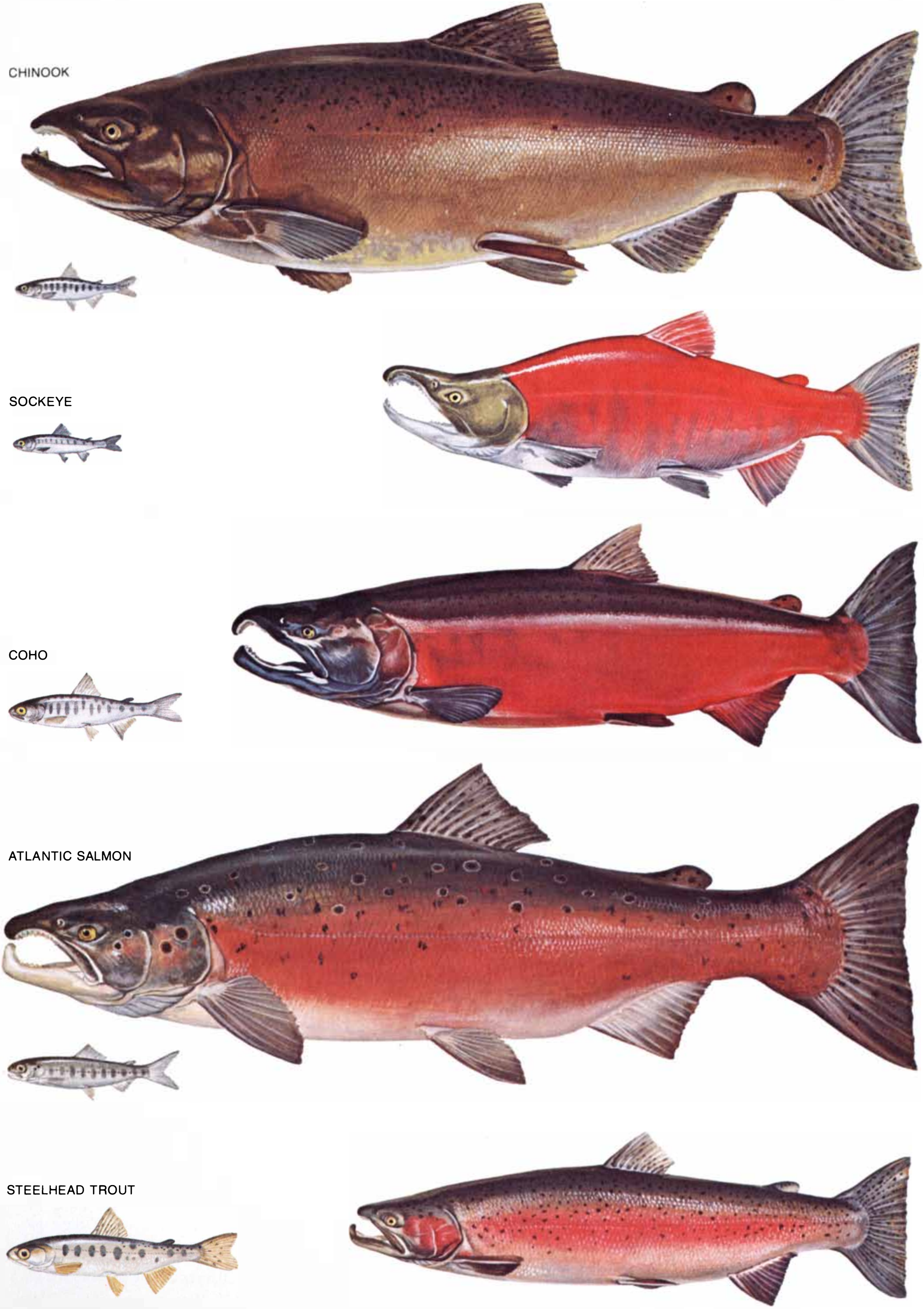
Ranched salmon are homing in large numbers to release sites in North America, Japan and Siberia. Salmon have been successfully transplanted to Chile. The fjords and sounds of the Chilean archipelago have abundant stocks of plankton on which the newly transplanted salmon can feed. In the future carefully selected stocks of salmon planted at the southern tip of South America may be able to migrate to the edge of the Antarctic to feed on the summertime concentrations of krill and deliver that resource back to the release sites splendidly packaged as bright red salmon flesh. Salmon introduced into the Great Lakes have adjusted to that freshwater habitat in a way many people thought would be impossible. Along the rocky coasts of Scotland and Norway fish farmers are raising salmon and rainbow trout to large size in blocked-off sea lochs, fjords and floating pens.

The natural distribution of the salmonids before human beings began to affect it has been fairly well established. In the genus *Salmo* Atlantic salmon (*S. salar*) and European trout species were found around the margins of the North Atlantic and in adjacent parts of the Arctic Ocean. The rainbow-steelhead and cutthroat species of trout were found around the margins of the North Pacific. The chars, members of the genus *Salvelinus,* lived around the margins of both the North Atlantic and the North Pacific basins and in the waters draining into the Arctic Ocean. The six species of Pacific salmon (genus *Oncorhynchus*) were originally limited to the margins of the North Pacific from Taiwan to California, including the Bering and Okhotsk seas and the adjacent parts of the Arctic Ocean.

All salmonids hatch in fresh water and spend at least their first weeks there. Some members of all three genera are anadromous: they migrate to the sea, often for great distances, before returning to their native streams to spawn. Atlantic salmon return to spawn year after year; Pacific salmon return once and die after spawning. Some species of salmonids remain in fresh water all their lives. Several species of Pacific salmon can adjust readily to living permanently in fresh water. Relict populations of cherry and sockeye salmon survive in lakes that have been landlocked for thousands of years.

The homing ability of migratory salmonids is legendary. They swim upstream for many miles, leaping in the air to climb waterfalls and fish ladders, and unerringly make every turn needed to get back to their native stream. Some years ago Arthur D. Hasler and James A. Larsen of the University of Wisconsin at Madison established that the fish home by smell, distinguishing the unique chemical signature of their native stream in concentrations as dilute as parts per billion. The signature is derived from the rocks, the soil and the or-

HOMING SALMON clogging the Tazimina River in Alaska appear in the photograph on the opposite page. They are sockeye, or red, salmon (*Oncorhynchus nerka*), returning to the place where they hatched. They left the river at from two to four years of age, lived for from one to three years at sea and are returning to spawn, finding their way back by the unique chemical "smell" of their native stream. The fish in ranks at the edges of the stream are resting before continuing upstream. The rank formation arises because they orient themselves upstream; they are side by side owing to their numbers. The other fish are making nests in the gravel. The photograph was made by Thomas C. Kline, Jr., of the University of Washington as part of an annual survey designed to forecast for commercial fishermen and conservation authorities how many sockeyes can be expected when the fish hatched from the eggs eventually return.


CHINOOK
SOCKEYE
COHO
ATLANTIC SALMON
STEELHEAD TROUT

ganic matter associated with the water.

Attempts at artificially rearing the salmonids started in Germany before anything was known about homing except that the fish did it with great success. In 1763 it was observed that the eggs could be squeezed out of the females, fertilized with sperm from the males and incubated in cool, flowing water until they hatched. The early investigators worked with the brown trout (*Salmo trutta*) and the Atlantic salmon (*S. salar*). In the U.S. the effort to culture the salmonids began in 1804 with the artificial spawning and incubation of the eastern char (*Salvelinus fontinalis*).

The first hatchery for Pacific salmon was built on the McCloud River in California in 1870. As in all previous efforts to propagate these fish, the river was robbed of its wild stock by seining the spawning fish from their ancestral spawning grounds to provide the "seed" for the hatchery. The same practice was followed in all the hatcheries established later along the entire Pacific coast from California to Japan. The hatcheries concentrated solely on the incubation of eggs, giving no thought to the problems of feeding and rearing the fry, which were released unready for the challenges of a hostile environment. As a result the early efforts contributed little to the salmon runs and in fact often impaired them.

In 1895 the state of Washington built a hatchery on the Kalama River, a tributary of the Columbia River. This station concentrated on the chinook (*Oncorhynchus tshawytscha*), the largest of the Pacific salmon. At the station procedures were developed for spawning, hatching, rearing, releasing and recapturing the fish to be spawned to provide eggs for the next generation: a spectrum of actions that made it possible to recharge the salmon runs. As the system developed it began to yield surplus spawn that could be transferred to other stations to start new runs.

In 1901 a salmon hatchery was built on the Green River, a small river that enters Puget Sound in the Seattle harbor. Chinook eggs from the Kalama hatchery were brought in to start a new run. After a period of adjustment these chinooks began homing on the Green River hatchery in such numbers that a surplus was soon available. The Green River hatchery became a source of chinook eggs for many other hatcheries in the Puget Sound area, including (in 1949) the experimental hatchery on the campus of the University of Washington. These chinooks, along with the stocks of rainbow and cutthroat trout and coho salmon already established there, provided the university with its own "herds" of salmonids and also provided the basis for a number of experimental projects.

Experimenters elsewhere had for some time been concocting specific recipes of chemicals to be put in the water where the young salmonids were released in order to ensure that they would return there. At the University of Washington we soon learned that such a step is unnecessary; the natural chemical signature of the water is sufficient. A major project entailed selective breeding to increase the proportion of chinook salmon that return to the university pond after two or three years at sea in relation to the number returning after four or five years. This project has been highly successful. So has a project for accelerating the growth of juvenile coho salmon by giving them special diets and controlling the water temperature. Under this regime our coho are able to smolt and go to sea after only six or seven months, compared with the normal period of from 18 to 30 months. These stocks and the techniques employed to rear them are now becoming the basis for commercial salmon ranching and farming in the U.S., Canada, Chile and France.

The chinook salmon brought to the university had to adjust to an environment far different from the environments commonly frequented by this species. Instead of coming back to big, deep rivers such as the Sacramento, the Columbia, the Fraser and the Yukon, to which many of their relatives are accustomed, chinooks returning from the sea to their "home" at the university must enter Puget Sound, turn left, enter the Lake Washington ship canal and pass through the locks either with the ship traffic or by way of the fish ladder along the south bank. Then, after a three-and-a-half-mile trip through the congested industrial area along the shores of Lake Union, they must turn left again, climb a small ladder and enter a collecting pond on the campus.

The attraction water that guides the fish into this last step of their journey is pumped from the ship canal into the collecting pond, from which it flows by gravity back into the canal. This tiny flow has served as the home stream for many generations of salmon and sea-run trout bred at the university. The university's chinook salmon have adjusted to this unusual environment so well that each fall the return to the campus, even after the chinooks have run the gauntlet of heavy sport and commercial fishing in the ocean, the Juan de Fuca Strait and Puget Sound, yields from 10 to 20 times the number of eggs needed to perpetuate the run.

The surplus eggs have served to build many other salmon runs. From 1967 to 1970 several hundred thousand eyed eggs (eggs in which the developing eyes can be seen, the earliest stage at which eggs can be handled carefully without damage) were sent to Michigan as seed for stocks that have become the basis for the lively new sport fishery for chinooks that has developed in Lake Michigan. In 1970 eggs were shipped to the New York State Department of Environmental Conservation for seeding Lake Erie and Lake Ontario with chinooks.

In 1980 eyed eggs from the university's stock of chinooks were sent to Chile. The young fish reared from those eggs were released into waters leading to the Southern Channels zone of that elongated country; there an abundance of crablike and shrimplike plankton is an ideal food source for foraging salmon. Two years later the Chilean and U.S. sponsors of the project were delighted by a return of adult chinooks that far exceeded their modest expectations. These fish are now providing a source of eggs for an expanded planting program that will soon include releases into the Strait of Magellan. Salmon venturing out to sea from those waters should have access to the enormous stocks of krill that congregate in the summer in the region of the Antarctic Convergence, where waters flowing outward from the Antarctic meet warmer waters from the ocean basins to the north. The fish should also be able to forage in the plankton-rich waters of the Patagonian Shelf, which might be more hospitable for them during the winter.

Chinooks are not the only salmonids whose numbers have been increased by these techniques. Programs also have been applied with success to the coho salmon (*Oncorhynchus kisutch*), the pink salmon (*O. gorbuscha*), the chum salmon (*O. keta*), the Atlantic salmon (*Salmo salar*) and the rainbow trout (*S. gairdnerii*).

Coho salmon were historically distributed from the Arctic to central California. Since they tend to spawn and rear their young in small streams, their homing journeys scattered them, and unique stocks or races of cohos evolved in each of thousands of spawning areas

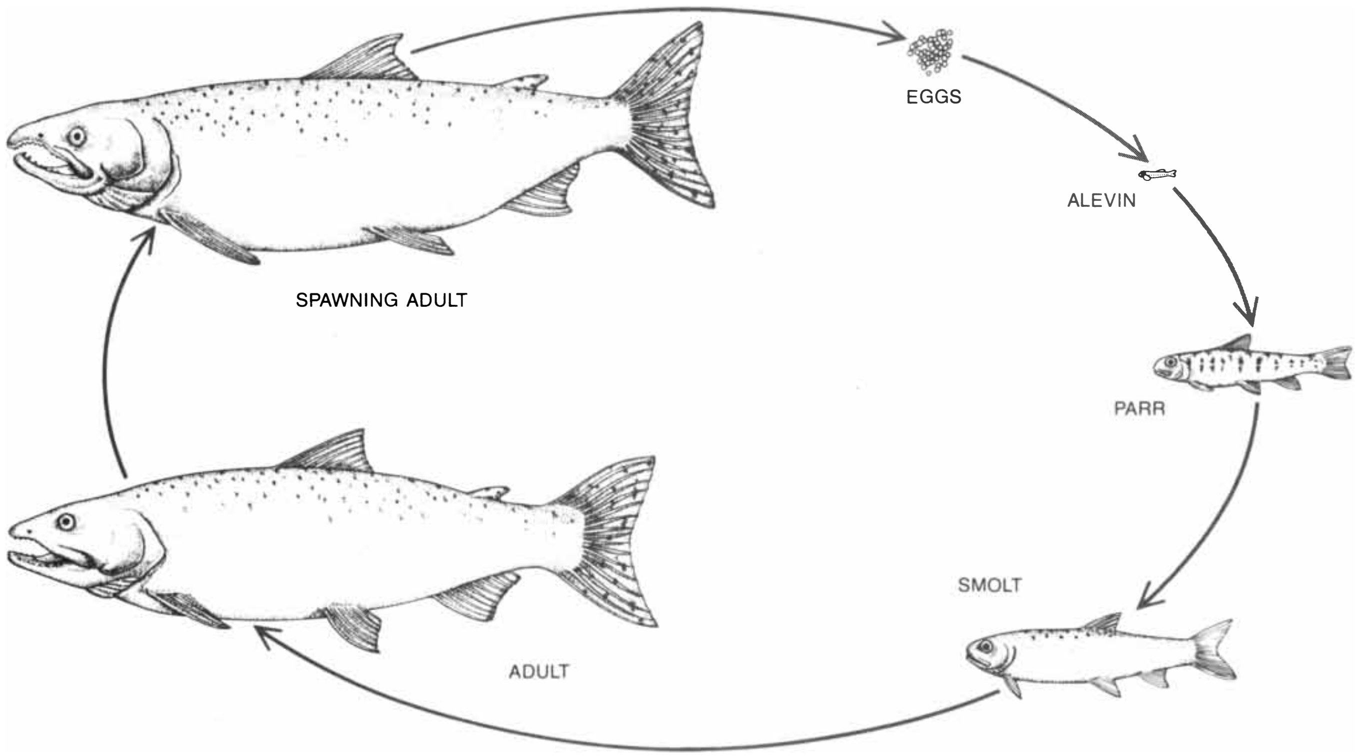
FIVE SPECIES of salmonid fishes, a family that includes the salmon and the trout, are portrayed on the opposite page at a common scale. All are males; at the right they are shown at spawning age and at the left as young parr, still in the freshwater stream where they hatched. Not long after the parr stage the fishes become smolts, that is, they go through the physiological changes that make it possible for them to migrate to sea and live in salt water. The top three species are Pacific salmon, which return to their native stream to spawn only once and then die. The Atlantic salmon and the steelhead trout return to spawn every year for many years.

in streams and tributaries with access to the North Pacific.

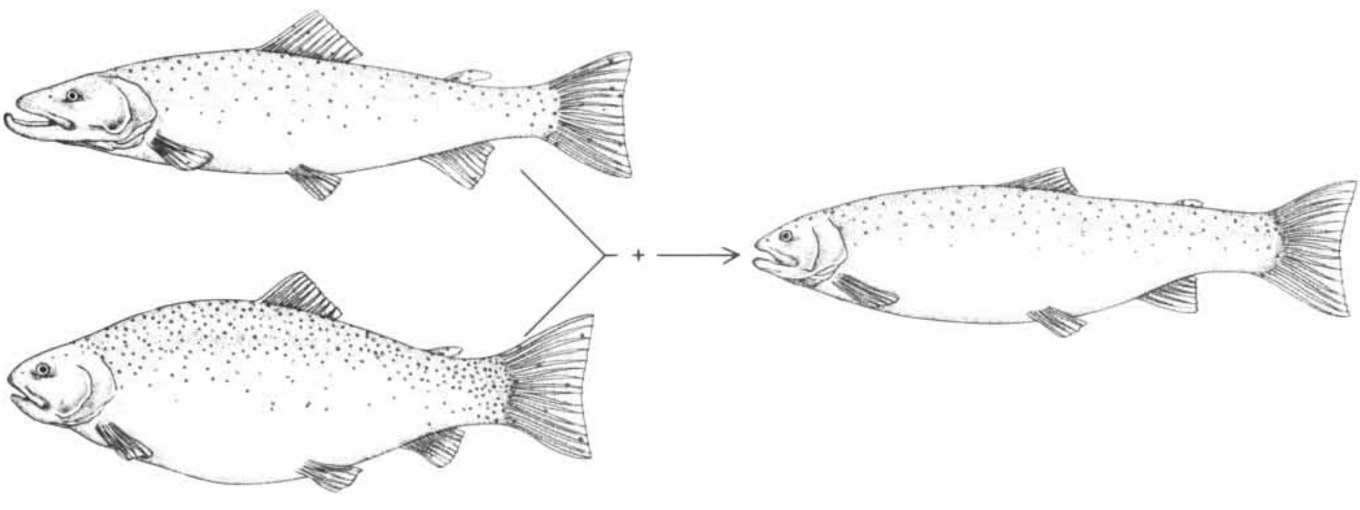
The adult cohos return to their home streams from their feeding areas in the ocean late in the fall, when the seasonal rains increase the flow and reduce the temperature of the water. In the long winter months the eggs lie covered with protective layers of gravel. They hatch between March and June. During the summer the young cohos feed on aquatic insects in the small streams where they were hatched. In the fall, when the adults return to spawn, the young from the previous spring feed on the food organisms dislodged by the spawning fish as they dig nests. The young also eat any eggs that are not immediately covered with gravel.

During the second winter the young cohos remain in the stream, virtually dormant. In their second spring the young in the southern part of the coho range become smolts: they go through the physiological changes that enable them to migrate to sea. (Without these complex changes the salmon, which have a salt content of about 12 parts per 1,000 in their blood, would dehydrate on entering seawater, which has a salt content of from 30 to 32 parts per 1,000.) Cohos stay at sea for a year and a half, returning at the age of three to start the cycle anew. In the colder parts of the range cohos usually need two and a half years to reach the smolt stage; after a year and a half in the ocean they return as four-year-old adults.

Coho salmon have been the "bread



SIX STAGES in the life of the coho salmon (*Oncorhynchus kisutch*) are depicted at a common scale. They are: the eggs; the alevin, to which the yolk sac is still attached; the parr, a spotted juvenile still living in fresh water; the smolt, a silvery fingerling ready to migrate to sea; the mature adult ready to return from the sea, and the mature adult coho after it has returned to its native stream to spawn.



HYBRIDIZATION OF TROUT was carried out at the University of Washington. Sperm from steelheads, which are anadromous (migratory) rainbow trout, served to fertilize eggs from a select stock of nonmigratory rainbows that had been developed at the university. Here the male steelhead is at the top left. The resulting hybrid grows faster than the steelheads but slower than the nonmigratory rainbows. It also migrates, and on its return it displays the rainbow's trait of biting at a lure, which makes it a good fish for sport fishermen.

and butter" fish for salmon hatcheries on the West Coast of the U.S. The millions of coho smolts turned out each year by Federal, state, Indian and private hatcheries now constitute the largest single source of salmon for the commercial and sport fisheries in California, Oregon and Washington. In 1967 one of us (Donaldson) organized a small research project at the University of Washington aimed at determining whether the growth rate of coho salmon could be accelerated with improved diets and warmed water so that fully smolted fingerlings would be ready to migrate in six or seven months instead of 18. We found that the objective was indeed attainable. Most of the cohos released as accelerated smolts returned to the university pond after 18 months at sea, that is, when they were two years old. By selecting brood stock from this group we were able to increase the rate of returning adults from 1.25 percent in 1969 to 6.42 percent six years later.

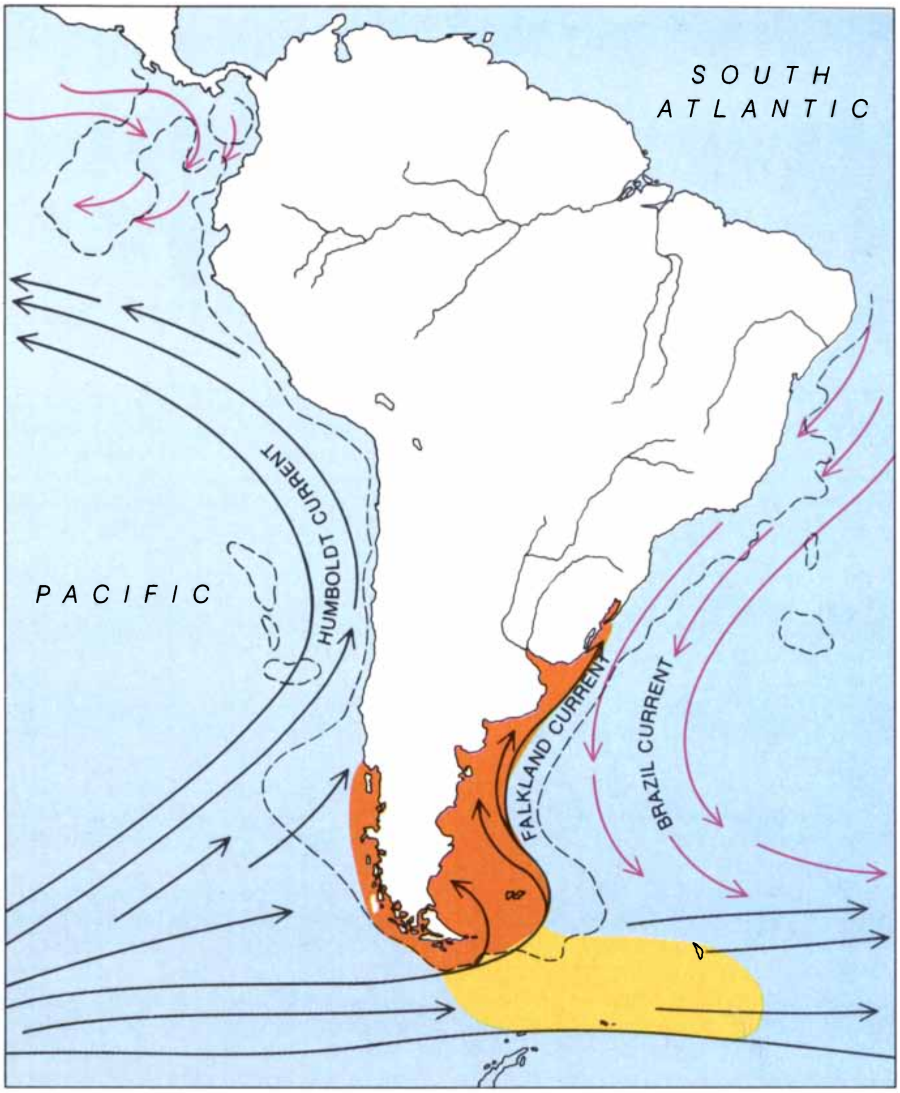
That improvement resulted from our application of recent research findings on the nutrition of the salmonids by developing diets with appropriate amounts of digestible protein, fats, minerals and vitamins. Fish meal has been the major source of the protein, but the quality of the product from Peru has been poor and supplies from the Northern Hemisphere are becoming both expensive and scarce. Yeast, vegetable protein and bacterial protein have been substituted in various amounts with various degrees of success. Fish-processing wastes preserved by acidification hold much promise.

Heat as a by-product of the generation of electricity or other industrial processes can be employed to hold the temperature of hatchery water in the optimal range for the growth of salmonids (usually from 8 to 12 degrees Celsius, or from 46 to 54 degrees Fahrenheit). Several advantages are gained by controlling the water temperature to accelerate the rearing of hatchery smolts: harmful extremes of temperature in the summer and winter can be avoided; smolts can be released before water levels get dangerously low in the summer; exposure to disease is reduced by the shorter residence time of the fish in the crowded conditions of hatchery ponds; less food is needed with shorter rearing times, and the time to harvest is reduced.

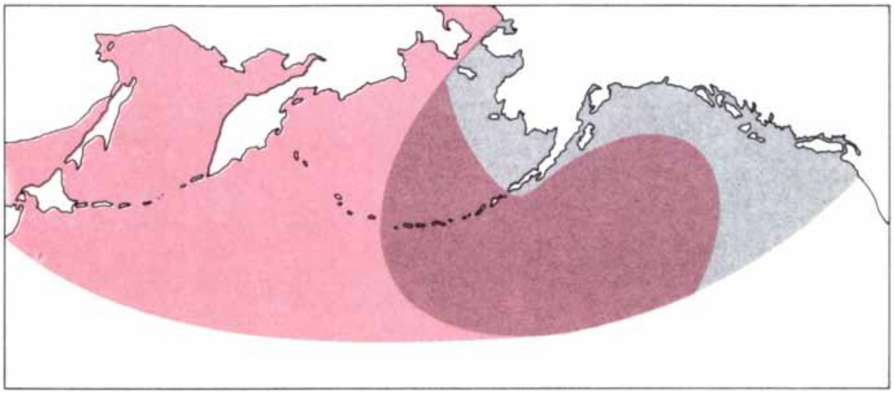
Many of the natural spawning and nursery areas for salmon have been destroyed in the developed nations of the Northern Hemisphere by the cutting down of forests, by farming and by the spread of urban-industrial complexes. Several countries have turned to artificial production to replace the salmon lost by the destruction of habitats. Over the past 20 years the Japanese have had remarkable success in producing chum salmon at hatcheries. Nearly all the chums now harvested in Japanese territorial waters are of hatchery origin.

Although a much greater variety of methods for artificially rearing many species of salmon and trout in hatcheries have been developed in North America and Europe, none of the methods has been applied with the single-minded intensity that has characterized the Japanese application of a system imported from New England in the 18th century, nor have the results been as spectacular. Nevertheless, out of all these efforts a new concept of ocean ranching of salmonids has begun to emerge. It differs from earlier hatchery techniques in that the young fish are carefully nurtured until their prospects for survival in the ocean are optimized. Only then are they released to migrate to the sea, where they can feed and mature on the open range until they return to their native coasts and rivers to be harvested. Ocean ranching will have a profound effect on the evolution and survival of salmon and sea-run trout. If it is managed well, the effect will be beneficial; if it is managed badly (so that, say, inappropriately selected stocks are put into a coastal ecosystem where they usurp habitats of other valuable stocks), the effect could be detrimental to both the ecosystem and the rational management of the salmonid fisheries.

In Oregon, where the ocean ranching of coho, chinook and chum salmon is



NEW OCEAN RANGE has been opened to Pacific salmon as a result of the transplantation of fertilized eggs to Chile from North America and Japan. The most successful transplant was made in 1980 from the University of Washington's stock of chinooks. Returns from the stock are providing a source of eggs for an expanded planting program that will release salmon into the Strait of Magellan. The map shows prospective ranges for salmon in the sub-Antarctic.



LONG MIGRATION of chum salmon (*O. keta*) reared artificially in Japan and the U.S.S.R. takes them to the Gulf of Alaska. The Asian chums are represented by the colored area, the North American chums by the gray area, and the region where they mingle by a mixture of color and gray. On their annual return to Asia the chums yield a harvest of 200 million pounds.

legal under close supervision by the state, a subsidiary of the Weyerhaeuser forest-products company has built a large hatchery near Springfield. There the heated discharge from a paper mill warms the icy water of the McKenzie River in the winter to a temperature of 11 degrees C. (52 degrees F.) for incubating coho eggs and rearing the fry. When the fry reach the smolt stage in the late spring, they are trucked to the coast at Newport and put in ponds that receive salt water pumped from the ocean. After from 10 to 14 days in the ponds the fish become adapted to full-strength ocean water. After a final feeding they are released into the ocean. In about 18 months they home on the artificial saltwater river that flows out of the ponds from which they were released, delivering themselves in prime condition for the harvest. Boats and nets are not needed. It is not even necessary for the ranchers to round up their animals and herd them into a corral; the fish go into the ponds on their own.

In the wild the pink salmon and the chum are usually in the more northerly parts of the salmon range. Ordinarily they spawn close to the sea; some races even spawn on tide flats, where fresh water seeps into the sea. After incubating in the gravel during the cold winter months the young fry pop out of the nest in the spring and go almost immediately into the sea. Unlike the other salmonids, the pink and the chum salmon can smolt as soon as the yolk sac is absorbed.

Most students of salmon biology have believed the chum and particularly the pink salmon could not live in fresh water but had to go to sea to survive. It was therefore surprising to find that a small release of pinks into Lake Superior showed up as adults three years later. This odd-year return is unknown in seagoing pinks, which normally return after two years. Undoubtedly a fishery will develop in the Great Lakes to exploit this new stock.

In the years after World War II huge hatchery programs were built up in northern Japan and on the island of Sakhalin in the U.S.S.R. to enhance the production of pink and chum salmon. Nearly all the returning salmon are caught live in traps or set nets rather than from boats as they are in most of the North American salmon fisheries. They are therefore unmarked and in prime condition when they are delivered to market.

The well-coordinated Japanese program has expanded so that more than a billion chum fry are being released each spring. The young smolts leave the coasts of Japan for ocean feeding grounds south of the Aleutian Islands and in the Gulf of Alaska. After from three to five years in the ocean they return to the Japanese rivers from which they migrated to sea. In the fall of 1982 the harvested return was well over 20 million salmon, compared with annual catches of from two to three million soon after the turn of the century, when the fishery was based on naturally spawning populations.

The Russians are releasing millions of pink and chum fry from hatcheries into the small streams of Sakhalin. They plan to increase the number by 100 million per year until the total reaches three billion in the year 2000. Like the Japanese, the Russians favor terminal harvesting. The spawners swim up collecting troughs along the riverbanks. The fishermen standing along the banks can pick them up and drop them directly into the processing lines. Terminal harvesting is by far the most economical and efficient way to harvest such a resource. It also makes possible accurate counts and thereby encourages precise and effective management.

The Atlantic salmon has been considered by some experts to be an endangered species, but a brighter future is beginning to emerge. National policy in Iceland has succeeded in maintaining and even increasing the number of salmon returning to Icelandic streams. The policy aims at protecting the resource rather than accommodating the commercial fishing industry as unfortunately many governments in Europe and North America do.

The salmon returning to Iceland belong to the farmers who own the land along the rivers. Only in the lower reaches of four of the 250 rivers and at three locations along the coast, however, are the farmers allowed to follow the ancient practice of netting salmon. This net fishery brings in about 40 percent of the annual catch. The rest is harvested by sport fishermen using rod and line. Both commercial and sport fishing rights must be leased from the farmers who own the land. The management of the salmon fisheries is strictly controlled by the government, which sees to it that an ample number of fish are allowed to proceed to the spawning grounds.

That program has been supplemented by artificial rearing and by the opening of rivers formerly impassable to salmon. At the government's experimental fish farm in Köllafjördur much progress has been made in the use of geothermally heated water for the accelerated rearing of Atlantic-salmon smolts, in controlling the length of time the fry are exposed to light each day to promote feeding and growth, in developing brood stocks for seeding in streams newly made accessible to salmon and in mixing salt water in rearing ponds to trigger the smolting process. Streams formerly impassable to salmon have been opened by building fish-passage facilities. The government also promotes and aids the development of commercial salmon ranching.

In western Europe the runs of salmon have been greatly reduced from the level of preindustrial times, not only by the unintentional impairment of habitats but also by interceptions at sea in which the drift nets of one country take salmon that originated in another country. In Scotland, Ireland, Norway and Canada entrepreneurs are taking steps to ensure the supply of Atlantic salmon for their markets by rearing the fish to maturity in sea ponds and floating cages. Norway, with its long coastline of protected fjords, is the leader in this new type of sea farming. In 1981 Norway marketed 8,000 tons of farmed Atlantic salmon; the output is expected to triple over the next three years.

Of all the salmonids the rainbow trout exhibit the greatest genetic diversity. From their original range along the margin of the eastern Pacific they have spread or been spread to almost every part of the globe. They thrive even in the Tropics at higher elevations.

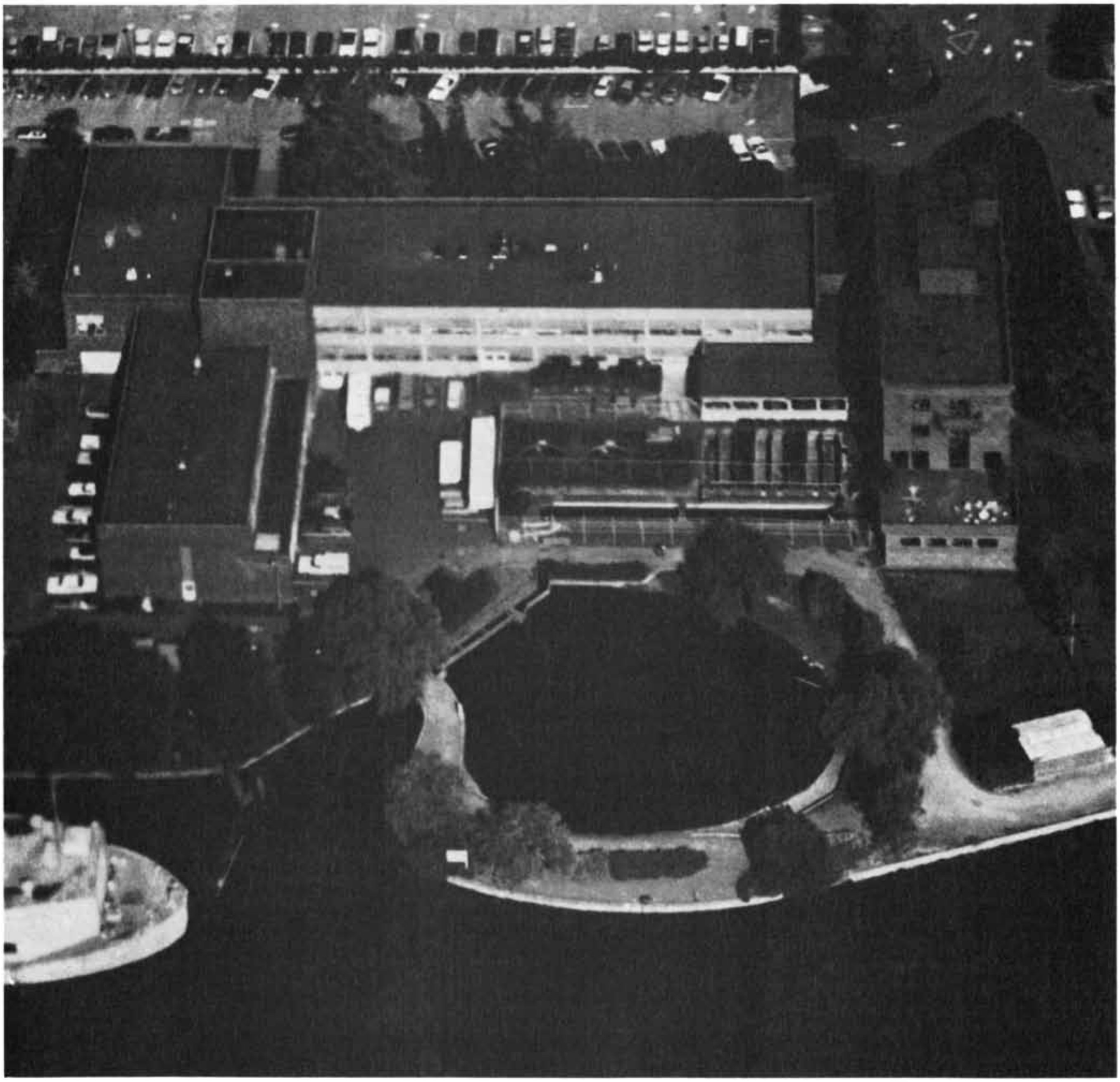
In addition to providing valuable

sport fisheries the rainbow trout are the basis of big commercial enterprises. Italy alone produces 21,000 tons of rainbow trout per year for marketing; Denmark, Finland, France, Norway and the U.K. all have major commercial operations. In southern Idaho, in the valley of the Snake River, trout farms yield about 13,000 tons of rainbow trout per year.

In the wild state rainbow trout spawn only in the spring. Now selectively bred stocks are available that spawn in nearly every month, providing a sound basis for commercial trout farming.

A program of selective breeding of rainbow trout has been in operation at the College of Ocean and Fishery Sciences of the University of Washington since 1932. The objective has been to turn out large numbers of healthy, acclimated fish for lake- and stream-management programs and for commercial farms. In the early years the trout reached maturity in their fourth year, weighing on the average 680 grams (1.5 pounds). Females yielded from 400 to 500 eggs in the first spawning. Some males of our select stock now reach 680 grams in their first year. The females, which mature in their second year at an average weight of 4,500 grams (10 pounds), yield about 10,000 eggs each. Although the rate of increase in the average length of the fish selected as brood stock has slowed, the average weight of the fish continues to increase. The result is a deep-bodied fish with better flesh. We like to think our results compare with the poultrymen's development of the broad-breasted turkey.

The university's strain of rainbow trout has been cut off from the sea for many generations. Nevertheless, and somewhat to our surprise, smolts from eggs of this strain adapted well when they were introduced to floating sea cages in Norway. They grew to a large size (from five to 15 pounds) in what



"HOME STREAM" for the chinook and coho salmon reared artificially at the University of Washington is a manmade pond at the edge of the campus. To reach it the 5,000 to 6,000 salmon that return to the pond from the sea each fall have to make several turns and pass through a ship canal. The water that attracts them is pumped from the canal to the pond and flows back to the canal by gravity.

we thought would be a quite stressful environment.

The rearing of rainbow trout in sea cages has become a profitable business in Norway. Some 4,000 tons were marketed in 1981, mostly in Europe. The market is expanding to the U.S., where large, fresh, pen-reared rainbows from Norway now appear in Seattle, the salmon capital of the world.

The crossbreeding of closely related stocks or races of plants and animals often results in "hybrid vigor." This tendency can be observed among the salmonids. Interracial crosses between steelheads (anadromous rainbow trout) and the select nonmigratory rainbows developed at the University of Washington result in hybrids that grow faster than the steelheads but slower than the nonmigratory rainbows. It appears that the dominant growth factor in the university stock has been transmitted, at least in part, to the steelheads. Moreover, the hybrids seem to retain the steelhead characteristic of migrating to the sea, where they grow rapidly. When they return at maturity, they provide exciting sport for recreational fishermen because they have the rainbow's trait of biting at a lure.

The selective breeding of salmonids can in a relatively short time affect their form, size, behavior, taste and food value as profoundly as 1,500 years of stock breeding have affected dairy and beef cattle. The future evolution of salmonid species now lies chiefly in the hands of the people who are beginning to breed them on a large scale. The task embodies challenges. One challenge is to maintain genetic diversity by providing appropriate sanctuaries to ensure the survival of wild stocks. Another is to utilize the enormous adaptability of these remarkable animals by selectively breeding stocks that will take hold both in new environments and in places where native stocks have been displaced by human activities.

FISH FARMING of Atlantic salmon and rainbow trout is carried out by Norwegian entrepreneurs in floating cages of the kind shown here. In 1981 Norway marketed more than 8,000 tons of farmed Atlantic salmon (*Salmo salar*). The output is expected to triple by 1986.

1700 pounds of optical perfection.

Perfection is not a term we use lightly. Or frequently. We've been taught for generations that perfection, in lens-optics terms, means $\frac{1}{4}$ wavelength. We've bettered this tolerance in the manufacture of an "egg crate" mirror, one-fifth the weight of a solid mirror. Manufactured for the Space Telescope Program, every point on its nearly 8-foot-diameter aspheric surface is within $\frac{1}{10}$ wavelength of perfection. So while we're not 100% perfect, we're within several millionths of an inch of that goal.

We put all our experience into play in the manufacture of this massive optical element $97\frac{1}{2}$ inches in diameter and 1700 pounds in weight. And our challenges were many. Since this lens is intended to perform in orbit, where the pull of gravity has been neutralized, we had to solve a gravity-distortion problem during the manufacturing cycle. We solved the problem with an air cushion, a membrane that inflates for testing to provide a uniform support and thus eliminate gravity-induced strain. The support deflates for polishing operations.

An aspheric mirror to gather information from the universe is an example of our large-optics capability. (Optics of this kind are ground, polished, and tested in our own lab with its vibration-isolated, six-story-high interferometric test towers.) But we're big in small things, too. Like lenses for cameras and copiers. We meet customers' needs with pride, precision, and stringent quality control.

In addition to massive, sophisticated equipment, a century of optical progress, and systems utilizing our ability to the fullest, we offer customers a tradition of realism when predicting costs and delivery.

That's important. Whether you seek to get a clearer picture of your neighbor or of celestial objects.

For more information about Kodak's contribution to the Space Telescope Program, write for the booklet *Mirror for the World's Most Powerful Telescope.* Send your request to: Richard M. Price, Research and Engineering, Kodak Apparatus Division, 901 Elmgrove Road, Rochester, NY 14650.

Kodak

Kodak. Where technology anticipates need.

Out to Launch

For more than two decades it has been the official policy of the U.S. to structure the nation's strategic nuclear forces in such a way as to ensure their ability to "ride out" a full-scale attack by the U.S.S.R. before the order is given to launch a retaliatory attack. The delayed-response strategy is intended to minimize the chance of starting a nuclear war by accident, on the basis of an electronic false alarm, say, or perhaps a human error. At the same time U.S. military and political leaders have been careful to avoid ruling out in public the possibility that a retaliatory second strike might be launched sooner, either on receiving early warning that a Russian first strike was on the way or just after it was learned that the incoming warheads had begun to reach their targets. The rationale offered for leaving open such a "launch on warning" or "launch under attack" option is that it adds to the uncertainty of military planners in the U.S.S.R. and hence reinforces the credibility of the U.S. deterrent.

Recent statements by members of the Reagan Administration have tended to draw increased attention to the early-launching option. The issue has been raised in the context of efforts by the Administration to get congressional approval of its plan to deploy 100 MX missiles in existing Minuteman missile silos. Advocates of the MX program had previously argued that the same silos were in danger of being destroyed by increasingly accurate Russian missiles; indeed, the vulnerability of the Minuteman system was advanced as one of the main justifications for proceeding with the MX program. Opponents of the Administration's newest basing scheme for the MX have pointed out that in spite of plans to "superharden" the existing silos, MX missiles based in them would be no more likely to survive a Russian "counterforce" attack than the Minuteman missiles they were designed to replace. Indeed, because the huge, 10-warhead MX missiles would constitute a more attractive target than the smaller three-warhead Minuteman missiles, the end result of this part of the Administration's "strategic modernization" program could well be a less survivable land-based missile force.

In testifying before the Senate Appropriations Committee in April, Secretary of Defense Caspar W. Weinberger and Chairman of the Joint Chiefs of Staff General John W. Vessey, Jr., acknowledged the theoretical vulnerability of MX missiles deployed in Minuteman silos. According to General Vessey, no more than 30 percent of the missiles could be expected to survive a Russian counterforce attack "if we ride out the attack." He added, however, that the Russians "have no assurance that we will ride out the attack." When the question was raised of whether the Administration was therefore moving in the direction of a strategy of launching under attack, Secretary Weinberger refused to discuss the matter further in public.

Although the statements of officials such as Secretary Weinberger and General Vessey on the release of nuclear weapons do not differ fundamentally from those of their counterparts in previous administrations, they do seem to reflect growing pressures on the leaders of both the U.S. and the U.S.S.R. to resort to the early-launching option in formulating their nuclear-war plans. Another indication of the concerns engendered by the advent of a new generation of "counterforce-capable" weapons in the arsenals of the two superpowers is the renewed attention being paid to command, control and communications (C^3) systems, which are generally thought to be the most vulnerable part of any deterrent force. The comparative "softness" of C^3 targets means they are vulnerable in theory not only to highly accurate long-range intercontinental ballistic missiles (ICBM's), which would take about 30 minutes to reach their targets, but also to shorter-range submarine-based and "theater" nuclear weapons, which have flight times of 15 minutes or less. Future discussions of arms-control issues are likely to focus increasingly on the dangers inherent in the seemingly inexorable technological trend toward reduced decision-making time and on the prospects for improving the reliability of systems designed to prevent the outbreak of an accidental nuclear war in the emerging "hair trigger" environment.

Reactor Power

Six countries now generate more than 25 percent of their electricity with nuclear power. (The U.S. has 72 commercial reactors in operation, far more than any other nation, but because of the country's size they account for only 13 percent of the power.) In France 32 operating reactors produce more than 40 percent of the power; France intends to have 85 percent of its installed capacity nuclear by the end of the century, with essentially only peak loads supplied by other sources. Nuclear reactors are generating power for the electric grid in 23 nations besides the U.S., and 17 more nations have firm programs for nuclear power. By the end of last year 9 percent of the world's electricity was being generated by nuclear reactors, up from 8 percent in 1981.

The data on the 40 nations other than the U.S. with nuclear-power programs were assembled by the Atomic Industrial Forum, Inc. In terms of reliance on nuclear energy the survey found the leaders other than France to be Sweden (36 percent), Finland (35.8), Switzerland (28.1) and Belgium (25.4). In terms of number of operating reactors the leaders were the U.S.S.R. (37, with no figure available for how much of the country's electricity they generate), France and the U.K. (32 each), Japan (25), West Germany (15) and Canada (12).

The survey tallied 207 operable reactors outside the U.S., eight more than there were at the end of 1981. In other categories there were 163 reactors under construction, 13 on order and 172 planned. The total capacity of the operating plants is 105,823 net megawatts, up 9 percent over the figure for 1981.

The survey took note of four "striking developments overseas" in 1982. One was the first order for commercial nuclear reactors from the People's Republic of China; it calls for two 300-megawatt pressurized-water reactors to be built near Shanghai by 1988. Another was the entry of Brazil into the ranks of the countries with operating commercial reactors. The third was "the start of full commercial operation by both India and Japan of spent-fuel reprocessing plants, signaling a major move toward nuclear self-sufficiency for both nations." The fourth was that South Korea became the eighth country to have a program for fast-breeder reactors.

Four in Every Ten

Forty percent of the world's people live in India and China. Therefore government policies bearing on population growth in the two countries have a significant relevance to the size and well-being of the human population as a whole. The recent demographic histories of India and China offer an instructive comparison, in particular because of the notable similarities between the two countries. Both are relatively poor, heavily agricultural nations. Both had a largely colonial status until after World War II, when they gained economic independence through powerful revolutionary movements. Reducing the birth rate has been a goal held strongly by government officials in both countries. There is, of course, a striking political and economic difference between India and China. Whereas the economy of India is largely a market one, the economy of China is almost entirely state-controlled. In a recent issue of *Proceedings of the National Academy of Sciences* Ansley J. Coale of Princeton University summarizes the significant demographic disparities that have resulted partly from the political differences between the two countries.

The current population of China is one billion. The population of India is 700 million. Between 1951 and 1981 the Chinese population increased by about 80 percent and the Indian population by about 90 percent. In China, however, there was a substantial reduction in the overall rate of population growth in the 1970's. There was no corresponding reduction in India. The overall rate of population growth is determined by the birth rate and the death rate. In both countries the death rate after World War II was high. In China in the 1950's there were 18 deaths per 1,000 people per year. By the 1970's the rate had been reduced to seven per 1,000, roughly equal to the rate in many developed countries. In India, on the other hand, the death rate in the 1950's was 26 per 1,000 and by the 1970's it had decreased only to 17 per 1,000.

According to Coale, the greater improvement in mortality in China is due largely to the fact that the system of medical care there is better and more egalitarian than the one in India. China has a higher ratio of physicians to population than India. The Chinese physicians are distributed more equitably than they are in India, where the doctor-patient ratio is eight times higher in the cities than it is in the countryside. Moreover, China has an extensive government-sponsored system of paramedical personnel, many of whom practice in rural areas. In both India and China more than 80 percent of the people live in the countryside. The better health care received by the rural majority in China appears to account for much of the difference in mortality rates.

In China fertility has decreased along with mortality. The birth rate in the 1960's was 36 per 1,000 of population. In the 1970's the rate was 23. The decline was due mainly to a vigorous official family-planning program. The intense coordination of all social and political activity in China contributed much to the effectiveness of the program. Such all-embracing social coordination raises the possibility that coercion is practiced in family planning. Chinese couples must obtain a "planned-birth certificate" before the wife becomes pregnant. Coale notes that there have been reports of forced abortions in instances where a woman was pregnant without a permit.

Coercion has also been a significant issue in India, where official family-planning efforts have been much less successful. The birth rate in India in the 1970's was 39 per 1,000 people per year, only very slightly less than the rate in the 1960's. According to Coale, "despite official support (for 30 years) of family planning, the government of India has not been able to organize a birth-control program that regularly provides adequately staffed services for most of the population." Partly out of frustration at past failures, the government of Prime Minister Indira Gandhi instituted a draconian sterilization program under the Emergency Act of 1975. Charges that Indian villagers had been sterilized against their will were soon widespread. Coale notes that "coercion was denied by the government but was almost inevitable, whether or not intended, because, for example, local officials set quotas for villages and for the number of motivated couples a schoolteacher was to round up." Resentment of the sterilization program contributed to the defeat of the Gandhi government in the 1977 elections. Thereafter family planning disappeared from the national agenda, to return only with Mrs. Gandhi's reelection.

Clearly both the Chinese and the Indian government are committed to reducing the rate of population growth. In explaining the much greater success of the Chinese in actually doing so, Coale cites their form of government and also the Chinese culture, which could be more conducive to the use of contraceptives and to the idea of limiting family size than Indian culture is. In spite of the past successes of the state birth-control program, however, the Chinese population will continue to increase for several decades because of the large number of young people who are now reaching childbearing age. Coale estimates that the population could cease to grow in 2020, when there will be some 1.2 billion Chinese. The smaller Indian population has a much higher rate of growth. Coale estimates that if fertility in India begins to fall soon, the ultimate stable size of the Indian population could also be about 1.2 billion. He concludes that in both countries "a lower birth rate now is desirable, but the ideal rate is not zero. There are social and political costs of excessive emphasis on the immediate achievement of very small families; the rights and sensibilities of the current population and the disequilibrating effects of drastic changes in age composition must enter the calculation of desirable population policies."

Still at Large

Recent experimental successes in elementary-particle physics have been confined largely to detecting particles that can exist only in the extreme environment of the high-energy accelerator. Two of the three intermediate vector bosons, the W^+ and the W^-, are inferred from the products of five energetic collisions seen earlier this year by two groups of investigators at the European Organization for Nuclear Research (CERN). As of this writing four more events have been recorded, bringing the total to nine. The third intermediate vector boson, the Z^0, may have been detected as well; a single pair of electrons, each with an energy of about 50 GeV (50 billion electron volts), were observed moving in opposite directions from the site of a collision between a proton and an antiproton. Such an event is widely regarded as a good signature for the presence of the Z^0.

A preliminary account of an analysis of the intermediate-vector-boson data, done by theorists at CERN, also suggested that the long-sought *t*, or top, quark has been found; more careful analysis must be carried out, however, before the existence of either the Z^0 boson or the *t* quark can be confirmed. Mesons that show naked beauty, or have a quark composition for which the property *b*, or beauty, does not cancel, have been observed following high-energy collisions between electrons and positrons (see "Particles with Naked Beauty," by Nariman B. Mistry, Ronald A. Poling and Edward H. Thorndike, page 106). The detection of all these exotic particles is important for the confirmation of theory, but in other respects the particles are ephemeral: they may not exist anywhere in the universe outside the instruments of physicists.

There are several phenomena, however, that remain undetected but should theoretically be found in ordinary matter. They are among the most sought-after and most elusive prizes in physics. Three fundamental questions have been raised: Is there a magnetic monopole? Does the neutrino have mass? Does the proton decay? Not only would affirmative answers to such questions be spectacular confirmation of theoretical predictions but also any definitive answer would have profound implications for efforts to forecast the long-range future of the universe. Nevertheless, in spite of years of work, no definitive answers have emerged; experimental physicists have only continued to narrow the limits within which the phenomena can be detected.

The monopole has attracted attention because its existence is predicted by the grand-unification theories, which link the strong, the weak and the electromagnetic force, three of the four fundamental forces of nature. (The fourth is gravity.) According to such theories, isolated magnetic poles having a mass of about 10^{16} times the mass of the proton should have been created within the first 10^{-35} second of the big bang. Such particles are many orders of magnitude more massive than the most massive particles that can be generated in accelerators; nevertheless, if they were created at the beginning of the big bang, most of them should still exist, perhaps as dense matter collected at the core of the earth or as micrometeoroids gravitationally bound to our galaxy or to the solar system [see "Superheavy Magnetic Monopoles," by Richard A. Carrigan, Jr., and W. Peter Trower; SCIENTIFIC AMERICAN, April, 1982].

The monopoles of cosmic origin should be the easiest to detect. Blas Cabrera of Stanford University recently devised an instrument in which three coils of superconducting niobium wire are mounted along three mutually perpendicular axes. If a monopole were to pass through the coils or through a shield that is mounted outside the coils, a small but measurable current would be induced in at least one of them. Cabrera found evidence for a single monopole last year by detecting the change in current in a smaller, less elaborate coil, but it now appears likely that the signal was spurious. With the new device Cabrera has seen nothing; if the mass of the monopole is indeed 10^{16} times that of the proton, his data now indicate there can be no more than one monopole for every 10^{17} protons in the universe.

Cabrera's superconducting coils can detect monopoles moving at any velocity, but the sampling rate is low because the coils must be relatively small. A scintillation counter can sample matter more rapidly, but slow-moving monopoles may evade its detection. Investigators from Stanford and from the University of Utah have attempted to find monopoles with a scintillation counter. They report that no monopoles are observed having velocities between 1.4×10^{-4} and 3×10^{-2} times the speed of light. The intensity of monopole radiation in the vicinity of the earth is probably less than 5×10^{-12} per square centimeter per steradian per second.

If supermassive magnetic monopoles are found, they could make up a significant part of the total mass of the universe. The attempt to detect a nonzero mass for the neutrino is also partly motivated by cosmology: if neutrinos have mass, the sum of it may be sufficient for gravity to arrest the expansion of the universe and drag all matter and energy back into the primordial singular point from which it began.

One mass-detection strategy takes advantage of the fact that there are three kinds of neutrino; if neutrinos have mass, quantum mechanics implies the three kinds of neutrino should oscillate, or freely exchange their identities. Investigators from the California Institute of Technology, the Technical University of Munich and the Swiss Institute for Nuclear Research are measuring the flux of neutrinos outside a nuclear power reactor at Gosgen in Switzerland. The apparatus is sensitive to the presence of the electron neutrino; if electron neutrinos change into muon neutrinos or tau neutrinos as they leave the reactor, the population of electron neutrinos should decrease with distance from the reactor. Felix H. Boehm of Cal Tech reports that no such variations are observed. Therefore neutrino oscillations, if they exist, are quite small.

The oscillation experiments do not give direct upper limits on the mass of the neutrino; they only make it possible to estimate the square of the difference between the masses of two kinds of neutrino. An experiment done three years ago by Russian investigators estimated the neutrino mass directly by measuring the decay rate of tritium, the radioactive isotope of hydrogen. They concluded from their results that the neutrino has a small but finite mass, between 14 and 46 electron volts. Reanalysis of the Russian data by J. J. Simpson of the University of Guelph, however, suggests that the data are consistent with a zero mass for the neutrino. Boehm's group has recently set an upper limit for the neutrino mass of 10 to 20 electron volts; the group measured the rate of the decay process called double beta decay in the isotope germanium 76. A larger-scale experiment is now planned at Cal Tech, where double-beta-decay rates will be measured in 3,000 liters of xenon 136.

Perhaps the most sobering experimental result for grand-unification theories is a new lower bound on the lifetime of the proton. In the simplest form of grand-unification theory the proton should decay in less than 10^{29} years, with the large possible error of plus or minus 2 in the exponent. Such infrequent events can be detected by searching for the proton decay in a large mass of ordinary matter. Investigators from the University of California at Irvine, the University of Michigan and the Brookhaven National Laboratory are monitoring 8,000 tons of water, or about 10^{33} protons, in a salt mine near Cleveland, but to date no event that resembles a proton decay has been observed. The findings imply the lifetime of the proton is greater than 6.5×10^{31} years, which is too long for the prediction derived from the simple form of grand-unification theory even if the potential error in the prediction is taken into account.

There still remains some comfort for the theorists. Two other experiments have reported evidence that the proton decays. In the mines at Kolar Gold Fields in southern India three candidate decay events have been reported. Investigators at CERN, monitoring plates of iron suspended in the Mont Blanc Tunnel between France and Italy, have announced one candidate event. The jury remains out on whether or not these results are real proton decays, and if they are, how such results can be reconciled with the null result in Cleveland. In any case, according to Maurice Goldhaber of Brookhaven, "the idea of grand unification will live forever." There are many more complex versions of the attempt to unify the forces of nature that may also get their day in court.

Hot Potato

The cultivated potato plant, *Solanum tuberosum,* is susceptible to a variety of insect pests. The wild potato *Solanum berthaultii* is not. Its resistance seems to be associated with two kinds of tiny glandular hairs that cover its stems and leaves; aphids in particular have been thought to be trapped physically by the hairs and their sticky exudates. The antiaphid strategy of *S. berthaultii* turns out to be more sophisticated than that. Investigators at the Rothamsted Experimental Station in England have found that one type of hair releases a hydrocarbon that is the main component of an aphid pheromone. A pheromone is a hormonelike substance that is secreted by an animal and influences the behavior of other individuals of the same species, and this particular pheromone serves as an alarm signal among aphids. In effect the plant masquerades chemically as an aphid warning its fellow aphids to keep away.

R. W. Gibson and J. A. Pickett report in *Nature* how they found the bogus pheromone by extracting *S. berthaultii* leaves with alcohol and analyzing the extract by gas chromatography and mass spectrometry. Of several hydrocarbons present in the extract but not in a similar extract of cultivated-potato leaves, one was shown to be (*E*)-beta-farnesene, the aphid pheromone. Its source was found to be a sticky droplet at the tip of the type-*B* hairs. To demonstrate that the substance is volatile enough to affect approaching aphids, Gibson and Pickett collected air from above intact leaves, passed the air through a solvent, pentane, and measured the amount of farnesene that had been dissolved in the pentane. Some 50 billionths of a gram of the pheromone was thus extracted from 20 milliliters of air, a concentration high enough to generate an alarm response.

The Rothamsted workers went on to demonstrate such a response directly. *S. berthaultii* leaves were put in a syringe and air was expelled from the syringe toward a colony of wingless aphids (*Myzus persicae*) about a centimeter away. The aphids dispersed, showing the same behavior as aphids exposed to authenticated samples of the pheromone; air from an empty syringe or one loaded with leaves of the cultivated potato had no such effect. Then aphids were placed a centimeter away from a potato-plant leaf, facing toward the leaf. Most of the aphids walked right onto the leaf of a cultivated potato plant. Only six of 48 aphids tested did so in the case of a wild-potato leaf; the others either turned away at a distance of from one millimeter to three millimeters or walked parallel to the edge of the leaf, maintaining a similar distance. Again the behavior is like that of aphids faced with a plant treated with (*E*)-beta-farnesene.

It is not hard to cross *S. berthaultii* with the cultivated potato, Gibson and Pickett point out. Some hybrid plants have the type-*B* hairs in quantity. If the ability to secrete the pheromone could be widely introduced into *S. tuberosum,* it should protect potato crops against the sapsucking damage done by aphids and also against viruses spread by the insects, a still more important threat.

Ice Hole

Few places on the surface of the earth embody the subtle, complex interactions that command the attention of earth scientists better than the ones where ice, air and water meet, and at these places of meeting few things exemplify the complexity better than polynyas: regions in frozen seas that stay anomalously free of ice. The word is the Russian for open place in the ice.

Two groups of investigators have contributed papers on polynyas to an issue of *Journal of Geophysical Research* devoted entirely to "marginal ice zones." J. D. Schumacher and C. H. Pease of the National Oceanic and Atmospheric Administration in Seattle and K. Aagaard and R. B. Tripp of the University of Washington investigated the polynya in the Bering Sea off the southern coast of St. Lawrence Island. There currents flowing north toward the Arctic Ocean meet winds blowing south. Ice forms continuously but the winds drive it off. At locations near the polynya the workers measured ocean currents, temperature and salinity and also wind velocity from November, 1980, through June, 1981. About two-thirds of the significant fluctuations in the currents detected that winter could be accounted for by variability in the large-scale pattern of oceanic circulation or by "direct wind forcing." Eleven times, however, something more complex happened. The ocean current reversed, and the water temperature decreased, so that it remained near the freezing point of the salt-laden sea. Evidently "brine rejection" (the expulsion of salt from freezing ocean water) increased the density of

The world's densest computer is now the heart of the world's most 32-bit computer.

Our 32-bit CPU.

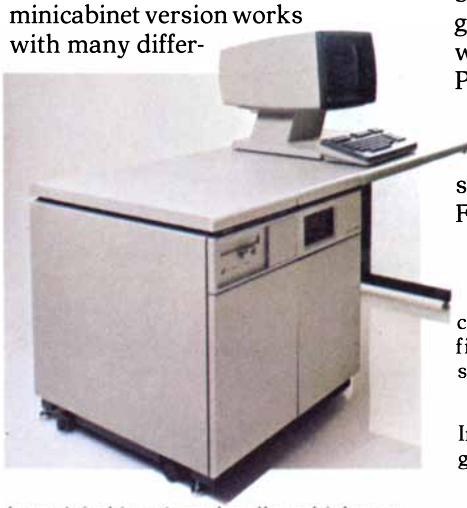
From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

32-bit computers for 32-bit applications.

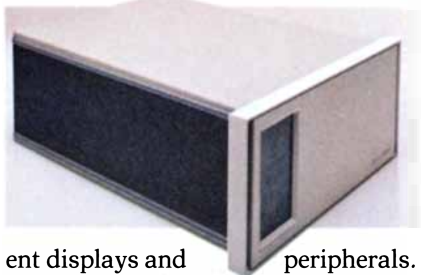
The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits—these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.



As a minicabinet, it can handle multiple users.

A rack-mountable version is available, too.



All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX.® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating

The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

chip
affordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals. Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.

HEWLETT PACKARD

the underlying sea, giving rise to "an offshore pressure gradient which, in concert with offshore wind stress, reversed the mean flow."

The other group, consisting of D. R. Topham and R. G. Perkin of the Institute of Ocean Sciences in British Columbia, S. D. Smith and R. J. Anderson of the Bedford Institute of Oceanography in Nova Scotia and G. den Hartog of the Atmospheric Environment Service in Ontario, investigated the polynya in Baffin Bay between Devon Island and Dundas Island. The area is notable for strong tides and currents; in addition the narrowest part of Pioneer Channel, which runs between the islands, proves to be bridged "by a shallow sill at a depth of approximately 40 meters." The workers measured ocean currents, temperature and salinity in March and April, 1980. The temperature varied with the tides by as much as .2 degree Celsius, resulting in "pulses" of ocean water. The colder pulses were "close to the local freezing point." They "probably originate in shallow nearshore regions." The warmer pulses rose to the surface at the south side of the channel. They appear to have derived from "strong tidal flows across [the] shallow sill." Their slightly higher temperature is sufficient to establish the polynya.

Help!

As many a tale of adventure fiction has it, the best way for a hunter to attract a tiger to his blind is to tether a young goat nearby. The bleating kid, frightened by restraint and darkness, is irresistible tiger bait. Goats are by no means the only mammals that scream with fright; so do, among others, rabbits, lambs, horses and human beings. The fear scream, however, is particularly common among birds; in many species cries of fright by young birds quickly attract a parent.

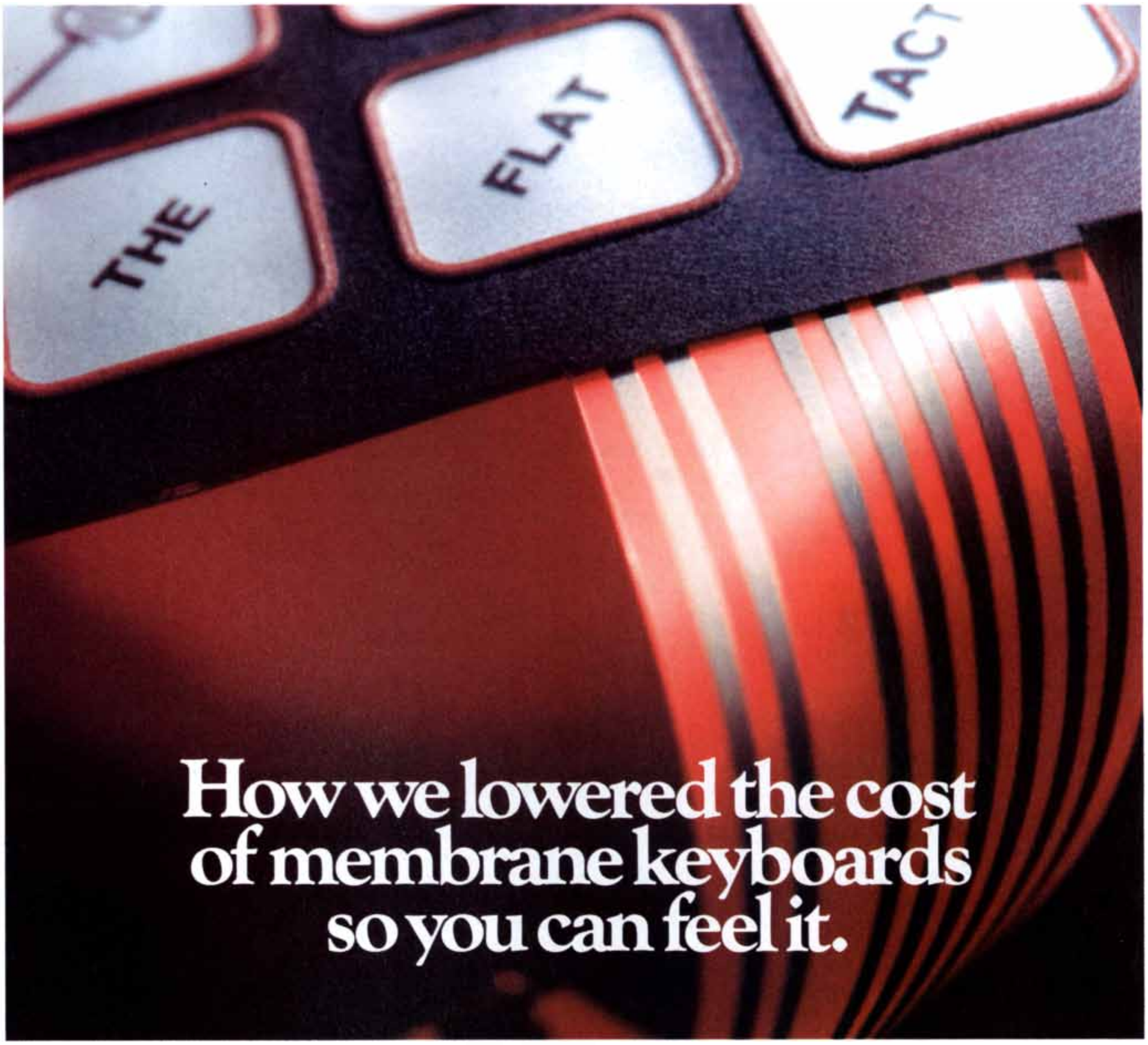
What is the adaptive value of such screams? The question was framed and answered recently by Göran Högstedt of the University of Lund. Writing in *The American Naturalist*, Högstedt dismisses one interpretation: that the victim's shrieks serve to warn others of its kind that a predator is present, so that its behavior can be classed as a form of altruism. Observations show that most birds actually ignore the fear screams of their own kind. Is screaming juvenile behavior, as with a threatened nestling, that has somehow persisted in the adult? Probably not. Frogs and toads that grow up without interacting with their parents show the same fear-scream behavior. Does screaming startle the predator, giving the prey a chance to escape? This is a speculation along the right line, Högstedt notes, but one that does not go far enough. What the scream actually achieves is the attraction of other predator species to the site of an attack in the hope of an easy meal: the prey, the original predator or both.

To test his proposal Högstedt recorded fear screams of the European starling and played them in the field when potential second predators happened to approach within 100 to 250 meters of his concealed loudspeaker. The most numerous predators he saw were common buzzards. Of the 32 individuals of the species that came into range, 14 quickly approached the source of the recorded screams. So did one out of three rough-legged buzzards that came within range, two out of four sparrowhawks, four out of 10 hen harriers and a single goshawk. The only predatory birds that failed to respond to the screams at all were seven kestrels. Adult kestrels weigh only 200 grams, compared with a goshawk's 1,300 grams; they may have been avoiding approaching larger predatory birds. The starling screams also attracted two predatory mammalian species: all of the four foxes that came within range and two of the three feral cats.

The field test left no doubt that fear screams attracted diverse predators. Predatory birds, however, hunt more by eye than by ear. In open terrain they would probably spot a predator attacking prey well before the prey's screams could be heard. This being so, the adaptive advantage of the prey's summoning a second predator should be manifested to a greater extent among birds that habitually live out of sight in dense cover and to a lesser extent among birds that live primarily on the wing or on open ground.

Högstedt tested this prediction by determining the proportion of fear-screamers among six species of aerial and open-ground birds and among six species of dense-cover birds. Of the six species in the first category (swifts, sand martins, swallows, house martins, tawny pipits and white wagtails) two did not scream at all, and on the average only 5 percent of the individuals of the remaining four species proved to be fear-screamers. Of the six species in the second category (wrens, thrush nightingales, marsh warblers, reed warblers, icterine warblers and garden warblers) more than 80 percent of the individuals of two species were fear-screamers, and none of the other four species showed less than 32 percent fear-screamers.

Högstedt concludes that fear-screaming is not an altruistic adaptation but a selfish one, a "cry for help" directed at any animal capable of interfering with the primary predator. The value of the adaptation is the possibility—slight, but better than none—that a clash between a primary predator and a "pirate" one will facilitate the escape of the screaming prey.



Companies everywhere are looking for the technological leadership that will improve their quality and productivity.

Those looking to the future with electronic keyboards are finding it with AMP technology.

The way we design membrane keyboards to work with either flat or keybutton designs, or our use of the latest conductive ink technology to stand up under millions of actuations—these are examples of that leadership.

Add the recognized advantage of AMP interconnection systems, and our customers get the mass production and reliability needed to accommodate today's electronic technology. And at a cost that's lower than conventional systems.

How can we help you?

Whether you need keyboards, connectors or the technology for either, come to us and you'll get what the world has come to expect. Industry-leading products that are designed to reduce the cost of making yours.

AMP has a better way.

For product innovations, for new applications, for ways to save you money, AMP has the better ways to make your company more productive.

Write for our brochure, "Why AMP means productivity." AMP Incorporated, Harrisburg, PA 17105.

Auckland • Barcelona • Brussels • Buenos Aires • Dublin • Frankfurt • Harrisburg • Helsinki • s-Hertogenbosch • Kowloon • London • Luzern • Mexico City
Montreal • Oslo • Paris • San Juan • Sao Paulo • Singapore • Stockholm • Sydney • Taipei • Turin • Toronto • Tokyo • Valley Forge • Viby • Vienna

AMP means productivity.

Learning in a Marine Snail

Hermissenda can be conditioned, or trained to associate two stimuli. Neural mechanisms underlying associative learning have been defined in the snail, and they may not be very different in the human brain

by Daniel L. Alkon

Can sophisticated mental functions in higher animals ever be accounted for in terms of the electrical activity of particular nerve cells and the chemical events that generate the activity? Neurobiologists can indeed describe some simple, stereotyped behaviors at that rigorously physical level. For example, the reflexive withdrawal of a limb in response to a harmful stimulus can be accounted for in terms of the electric impulses of a few identifiable cell types; many of the electrophysiological and biochemical mechanisms underlying the firing of those impulses are known. There is an immense gulf between a reflex action and, say, human memory and learning. Yet there is at least a chance that one can begin to bridge the gulf through the analysis of simple forms of learning in animals with a simple nervous system.

Pigeons, dogs and human beings can be conditioned, that is, they can be trained to associate one stimulus with another temporally related stimulus. The dog learns—and remembers—that the sound of a bell precedes and presages the smell of meat, and so it salivates at the sound of the bell. The vertebrate brain is far too complex to allow an analysis of such associative learning at the level of individual cells with present-day technology. For several decades, however, neurobiologists have been studying, in invertebrate animals, small networks of nerve cells that are responsible for integrating sensory information; in the past 20 years some have gone on to study the neural circuitry underlying simple reflexive behaviors. Some 12 years ago my colleagues and I at the Marine Biological Laboratory in Woods Hole, Mass., wondered whether such animals, with many orders of magnitude fewer neurons (nerve cells) than the trillion or so cells in the human brain, might be susceptible to conditioning. If we could train such an animal to associate two stimuli, we would try to identify the neural pathways activated by those stimuli and find out how the pathways converge, providing the physical basis for the learned association. By studying the activity of neurons in the pathways we would then try to identify specific changes effected by the conditioning and so begin to define "learning" in electrical and molecular terms.

Training

We chose to work with a gastropod mollusk, the shell-less marine snail *Hermissenda crassicornis*. In its natural environment *Hermissenda* exhibits positive phototaxis during the daylight portion of its light-dark cycle, that is, it moves toward light. The adaptive value of this behavior is clear: hydroids, the microorganisms on which the sea snail feeds, are concentrated in well-lighted water near the surface. When the ocean becomes turbulent, as in the course of a storm, *Hermissenda* responds by decreasing its overall velocity of movement and increasing the surface area of its muscular "foot." These responses enable it to cling to hard surfaces deeper in the water. Again the survival value is clear. An animal approaching the surface during a storm is shaken vigorously. It can survive for many weeks without feeding, but it cannot survive structural damage wrought by a turbulent sea.

Perhaps the experience of encountering turbulence in association with light at the surface causes the snail to "learn" to associate the light with the turbulence. The light would come to presage the onset of turbulence, and the learned association would tend to retard the animal's movement toward maximal light and thus into an area of maximal turbulence. We set out to reproduce the effects of light and turbulence in the laboratory.

We trained the snails in glass tubes filled with artificial seawater and arranged radially on a turntable. First we measured the average time untrained animals took to move to an illuminated area at the center of the turntable. Then we rotated the turntable at various speeds, subjecting the animals to a centrifugal force simulating the effects of turbulence. The conditions varied. For some animals light and rotation were precisely and repeatedly paired; five control groups were subjected to neither light nor rotation, to one or the other or to different combinations and temporal associations of the two stimuli.

After training we retimed the animals' movement toward light and calculated the effect of the training. For animals subjected to paired light and rotation (and only for those animals) the velocity of movement toward light was less than a third of the pretraining velocity. They had learned to associate light (which in this case served as the conditioned stimulus) with rotation (the unconditioned stimulus). The effect of the conditioning was to make the response to light more like the response to rotation.

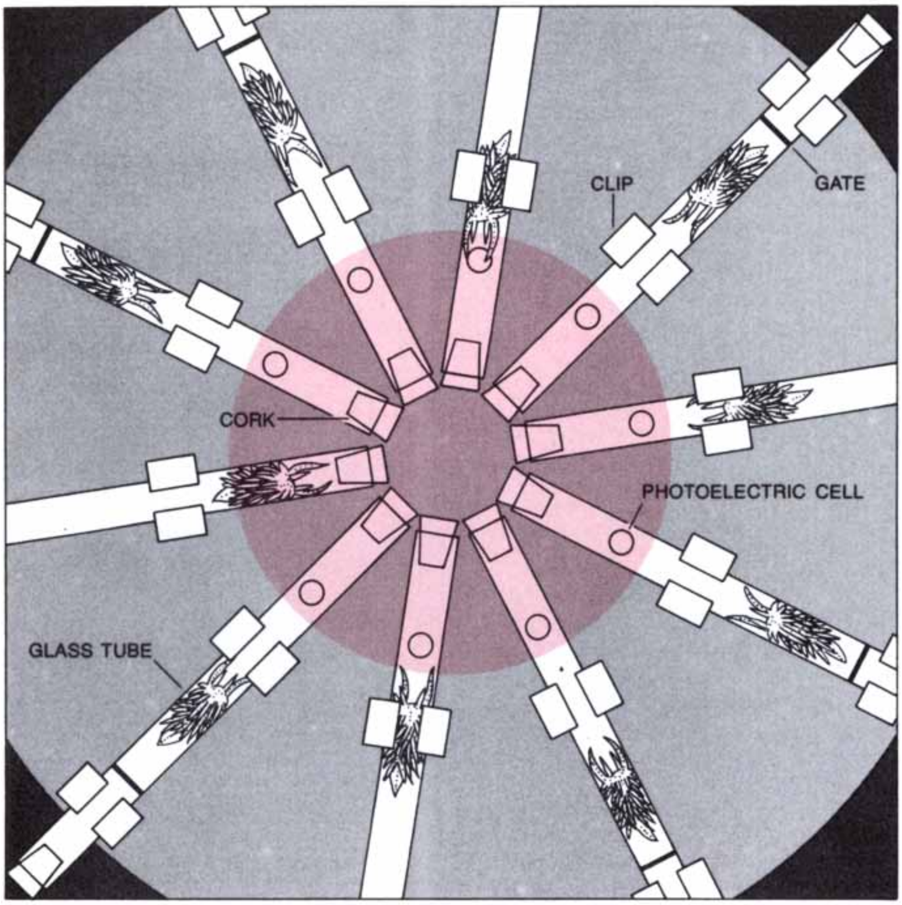
Conditioning

Our various control procedures indicated that the change of light-elicited behavior after training truly reflected associative learning. Exposure to light alone did not change the subsequent response to light, ruling out a nonassociative behavioral change known as habituation: a progressive decrease in responsiveness resulting from the repeated presentation of a stimulus. Exposure to rotation alone did not change the subsequent response to rotation, ruling out the possibility of sensitization: an increase in responsiveness resulting from the repeated presentation of a strong motivational stimulus. (Both habituation and sensitization can confuse a finding that associative behavorial change has taken place; such effects have not been completely ruled out in the case of certain behavioral changes reported to follow associative training in other snail species.) Moreover, *Hermissenda* was not conditioned when light and rotation were alternated during training or when the stimuli were presented at random. Terry J. Crow and I found the two stimuli had to be presented together. (More specifically, the onset of light preceded

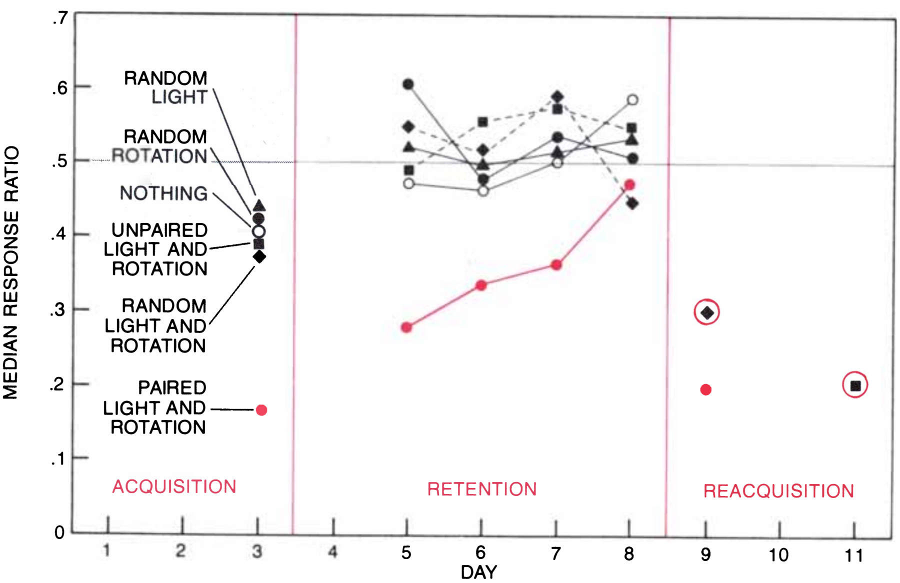
MARINE SNAIL *Hermissenda crassicornis,* pictured in a seawater aquarium in this photograph made by Pierre A. Henkart of the National Cancer Institute, is about an inch and a half long. The two large oral tentacles are tactile and chemosensory organs; the smaller dorsal tentacles called rhinophores are thought to sense water movement. The animal's central nervous system is on the buccal crest just behind the rhinophores. The plumelike appendages covering the snail's back, the cerata, may serve, like gills, as gas-exchange organs.

CENTRAL NERVOUS SYSTEM of *Hermissenda* is enlarged some 50 diameters in a photomicrograph made by Alan M. Kuzirian of the author's laboratory. The methylene-blue stain picks out the globular ganglia, or clusters of related nerve cells, and a number of nerves radiating over the buccal crest. The eyes (*black*) and the primitive vestibular organs called statocysts (*white dots*) are between the two pedal ganglia (*lower left and right*) and the more medial pleural ganglia.

EYE AND STATOCYST are enlarged about 300 diameters in a micrograph made by Kuzirian. The light-sensitive membranes of five photoreceptor cells extend up through the black pigment to abut on the underside of the lens. The roughly spherical statocyst wall is made up of 13 hair cells. Their motile cilia (not visible here) keep a cluster of crystals, called statoconia, centered; changes in the direction of gravitational force move the crystals, exciting particular hair cells.



SNAILS ARE TRAINED AND TESTED in glass tubes on a turntable. First each snail's phototactic response to light before training is measured: the time the snail takes to reach a spot of light (*color*) is recorded automatically when the animal reaches a photoelectric cell. Then the snails are trained by being rotated while confined to the outer end of the tube and thus subjected to a centrifugal force that is sensed by the statocysts. For one group the rotation is precisely paired with a 30-second period of light; various control groups either are not trained or are subjected to light or rotation alone or to the two stimuli alternately or at random. Finally the snails' velocity of movement toward light is timed again to determine the effect of training.



CONDITIONING IS DEMONSTRATED by these data. The "response ratio" measures the suppression of the phototactic response. A ratio of .5 means in effect that there has been no change in the velocity of movement toward light; a lower ratio signifies a reduction in the velocity. Animals subjected to paired light and rotation (*colored dots*) are conditioned after three days of training: they move more slowly toward light. They "forget" that behavior slowly but reacquire it after one day of retraining. Control animals show no significant change in behavior (*black*). Yet they too can be conditioned (*colored circles*) when subjected to paired stimuli.

the onset of maximum rotation by about a second.) It was the fixed temporal relation between stimuli that the animal learned and remembered.

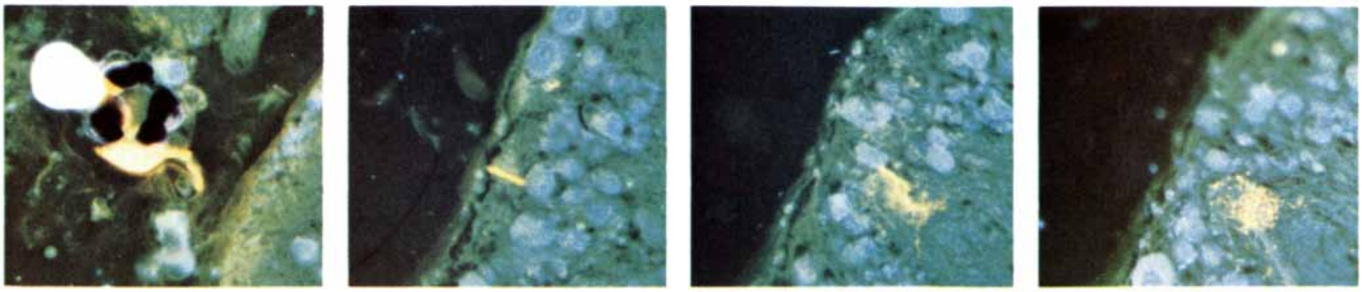
The behavioral change showed other characteristics of classical conditioning. It showed "acquisition" (the amount of change increased with practice), it could last for many weeks (close to a generation for this snail) and it did not affect the animal's response to stimuli other than the two being manipulated. William Richards and Joseph Farley of Princeton University showed that the learning was subject to extinction: after training was complete the animals "forgot" the learned association more quickly the more often light was presented without rotation. Crow showed, moreover, that after extinction the animals relearned the association with less training than they had the first time; the accelerated relearning is known as savings. Finally, Farley showed that *Hermissenda* learned more slowly when extra, unpaired stimuli (light or rotation) were interposed during conditioning. This degradation of learning by unassociated stimuli demonstrated contingency: the animals were learning that rotation is contingent on light, and when it was not, the learning process was impaired.

The behavioral changes I have described were observed in snails taken from their natural environment on the coast of California. Such animals displayed a wide range of responses to light before training, presumably reflecting individual differences in experience as well as in genetic background. June F. Harrigan therefore developed a laboratory strain of *Hermissenda* whose pretraining responses were much less variable, as were the behavioral changes observed in the course of training. We have also been able to raise animals with a defect in the organ that senses rotation and so obtain further evidence for associative learning. Such snails respond to light normally but cannot associate light with rotation.

All in all, these results suggested that *Hermissenda*'s ability to learn associatively could not be readily distinguished from the ability of a dog (or a whale or a human being, for that matter). This being the case, it seemed to us the underlying processes responsible for such learning might transcend differences among species. The first step toward understanding such processes was to trace the pathways mediating the association between light and rotation in the snail.

The Neural Pathways

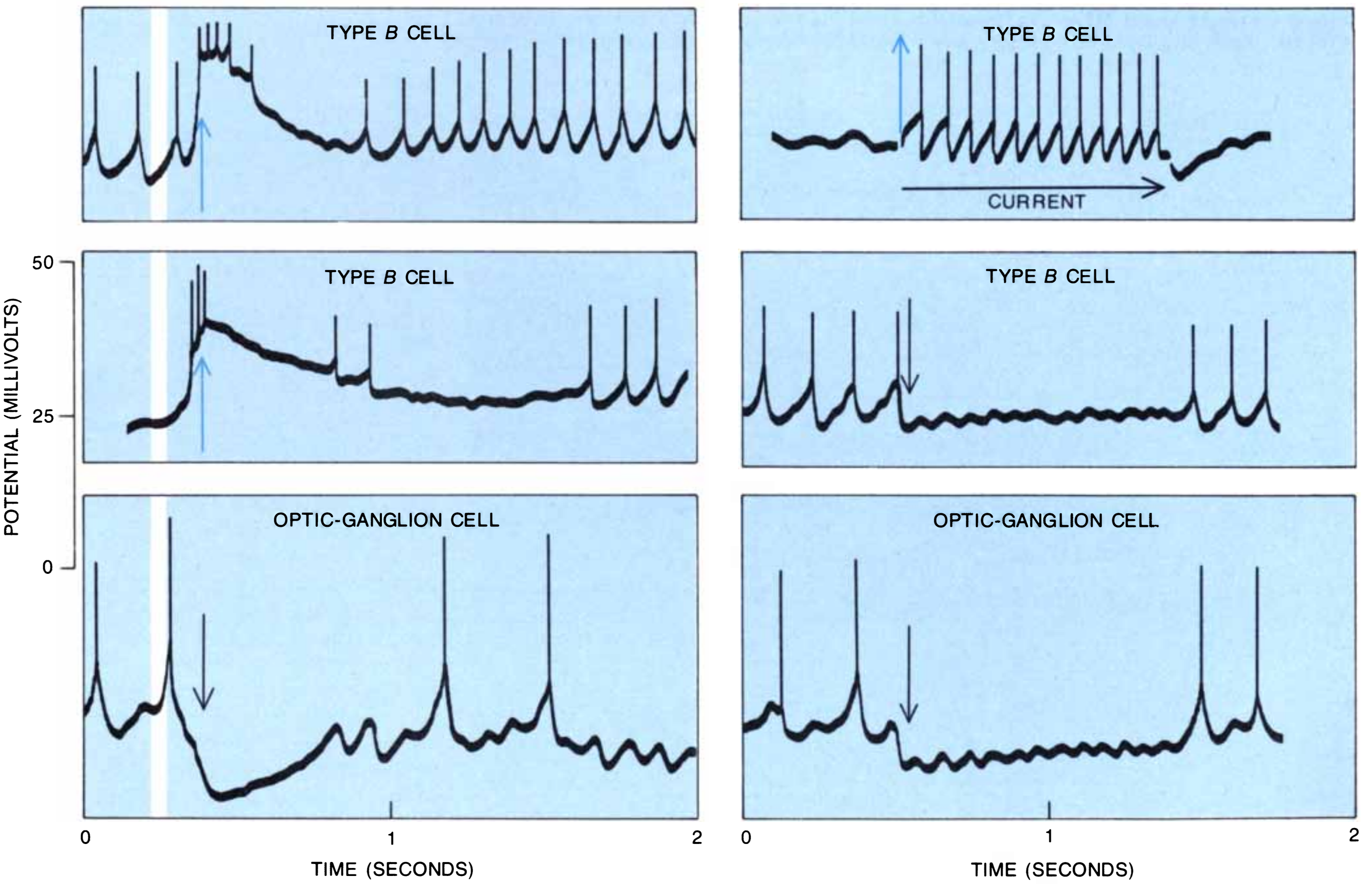
We began by mapping a wiring diagram of *Hermissenda*'s nervous system, with particular attention to two sensory pathways, one of them responsive to the conditioned stimulus (light) and the other to the unconditioned stimulus (rota-



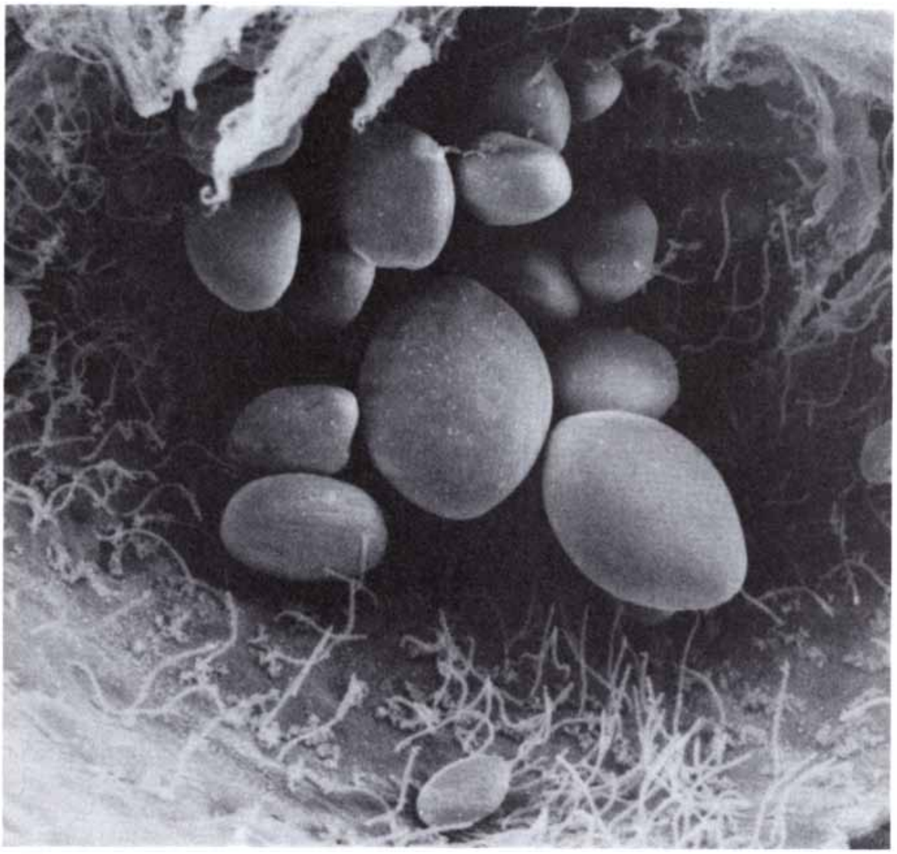
TYPE *B* PHOTORECEPTOR is mapped in micrographs made by Leona Masukawa. The dye Procion yellow was injected into a Type *B* cell and was transported along its axon to the axon terminals. In the first micrograph the dye fills the cell body (near the white lens) and the beginning of the axon. In successive thin sections of tissue the axon is seen (sectioned obliquely) entering the pleural ganglion and, in the last micrograph, ending in a spray of fine branches. The terminals of the branches form synapses (junctions) with other cells.

tion). Light generates electrical signals in two Type *A* photoreceptor cells and in three Type *B* photoreceptor cells in each of the animal's eyes; these signals are transmitted sequentially to interneurons, motor neurons and groups of muscles. Rotation is transduced into electrical signals by 13 hair cells (named for the hairlike cilia that line their inner surface) in each of two primitive vestibular organs analogous to the mammalian inner ear: the statocysts. These signals too are transmitted successively to interneurons, motor neurons and muscle groups. Some terminals of hair-cell axons (an axon is the long main process that carries the nerve impulse from one neuron to others) have synapses (junctions) with the axons of visual receptor cells. These and other, indirect interconnections between the visual and the vestibular pathways are the loci of convergence between the two sensory modalities.

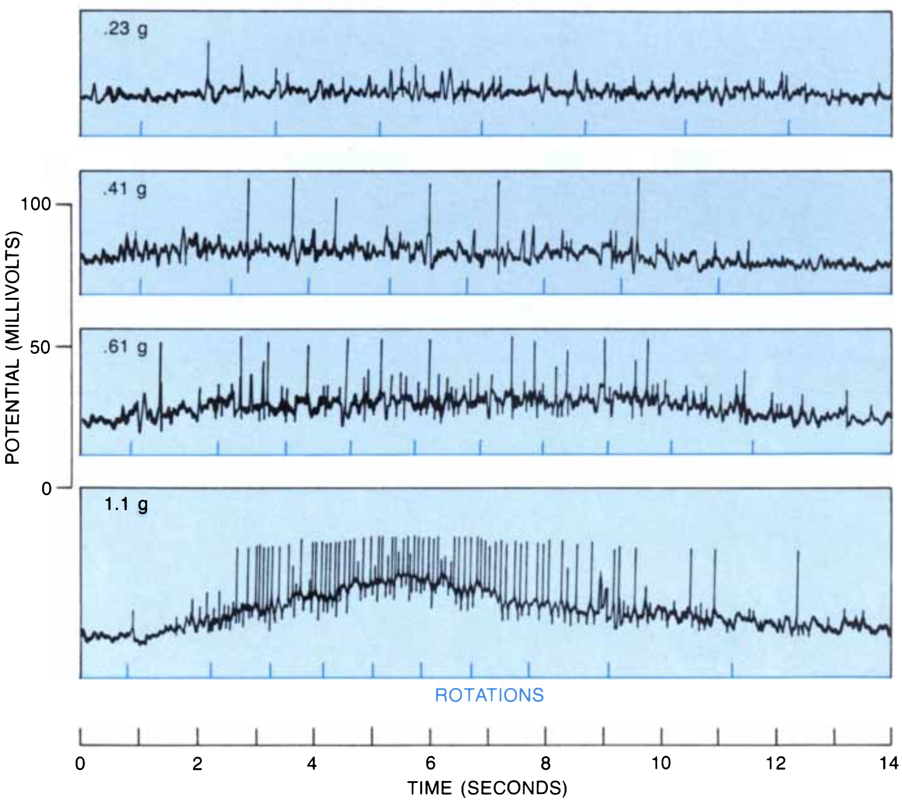
We mapped the two pathways both by electrophysiological methods and by making micrographs of stained preparations. The electrical responses of cells to light or rotation (or to the injection of microcurrents simulating the natural stimuli) were recorded with microelectrodes. By stimulating and recording from pairs of cells in thousands of animals we learned when and how specific receptor cells respond to particular stimuli by firing nerve impulses along the length of their axon; we learned to which postsynaptic neurons the signal is transmitted and whether it excites or inhibits such neurons. Repeating the process for successively connected cells enabled us to work out the pathways from the receptors that sense light and rotation to the muscles that impel *Hermissenda* through the water. The electrophysiological data were supplemented



INTRACELLULAR RECORDINGS show the effect of light (*left*) and of a current injection (*right*) on Type *B* photoreceptors and hence on an optic-ganglion cell they inhibit. In response to a brief flash of light (*white strip*) two *B* cells become depolarized, or more positively charged (*colored arrows*). Nerve impulses are triggered. These signals generate inhibitory postsynaptic potentials in the optic-ganglion cell, making it more negatively charged (*black arrow*); fewer impulses are triggered. Injection of a positive current (*horizontal arrow*) into either of two *B* cells simulates the effect of light, eliciting impulses that inhibit both the other *B* cell and the optic-ganglion cell.



HEMISPHERE OF STATOCYST is enlarged some 2,000 diameters in a scanning electron micrograph made by Eliahu Heldman. One sees the inner surface of the hair cells and a number of the statoconia (dispersed here because they are not kept clustered by the motion of the cilia).



RESPONSE OF A HAIR CELL was recorded as an isolated nervous system was being rotated at various speeds. Centrifugal force (indicated as a gravitational force, *g*) drove the statoconia against the cilia, making the hair cell's membrane potential more positive and giving rise to impulses. The cell's axon had been cut in order to eliminate any synaptic effects on the cell.

by information from photomicrographs and electron micrographs. A dye injected into a cell body moves along the axon to its farthest terminals, and the stained fiber can be traced in successive thin tissue sections; when two communicating neurons are stained, the synaptic junction between them can be visualized.

By these means we built up a wiring diagram: an interconnected network of chains of linked neurons [*see illustration on page 76*]. The diagram was essentially the same in every adult *Hermissenda* we studied. Clearly it is specified genetically; it is the invariant substrate for whatever changes are brought about by a learning process. How does its activity change with conditioning? Our investigation of the changes was made at three successively more localized and specific levels.

The first level was neuroanatomical. We measured the changes, during and after conditioning, in the signals that arise within individual neurons and are transmitted to other cells. Then we pinpointed the neurons in which the signals change first and those in which the signals remain changed, thereby identifying the events that actually cause the acquisition and the retention of the conditioned behavior. At a second, biophysical level we dealt with the cell membrane of the neurons in which the critical changes take place. We found differences in the currents passing across those membranes—in the flux of specific ions, or electrically charged atoms—that account for the changed signals. Finally, at the biochemical level, we identified chemical reactions that control the altered flow of ions.

The Excitable *B* Cells

In essence our neuroanatomical study revealed a circuit of neurons whose activity is modified depending on the pattern of stimuli to which it is exposed. The potential for this redesigning is implicit in the wiring of the visual and vestibular pathways and their interconnections. When light is paired with rotation during training, the net response of the system is to excite the three *B* photoreceptor cells in each eye more than usual and to render them more easily excited. That is, the excitation of the Type *B* cells that normally results from stimulation by light is enhanced and prolonged after training.

This is not a direct effect but a result of the integrated response of the system to the pairing of light and rotation. Normally the hair cells that respond most strongly to rotation fire even in the absence of excitation. This "tonic" activity sends inhibitory signals to the Type *B* cells. When the hair cells are stimulated by rotation alone, their activity is increased and they inhibit the *B* cells more strongly. When rotation stops, however,

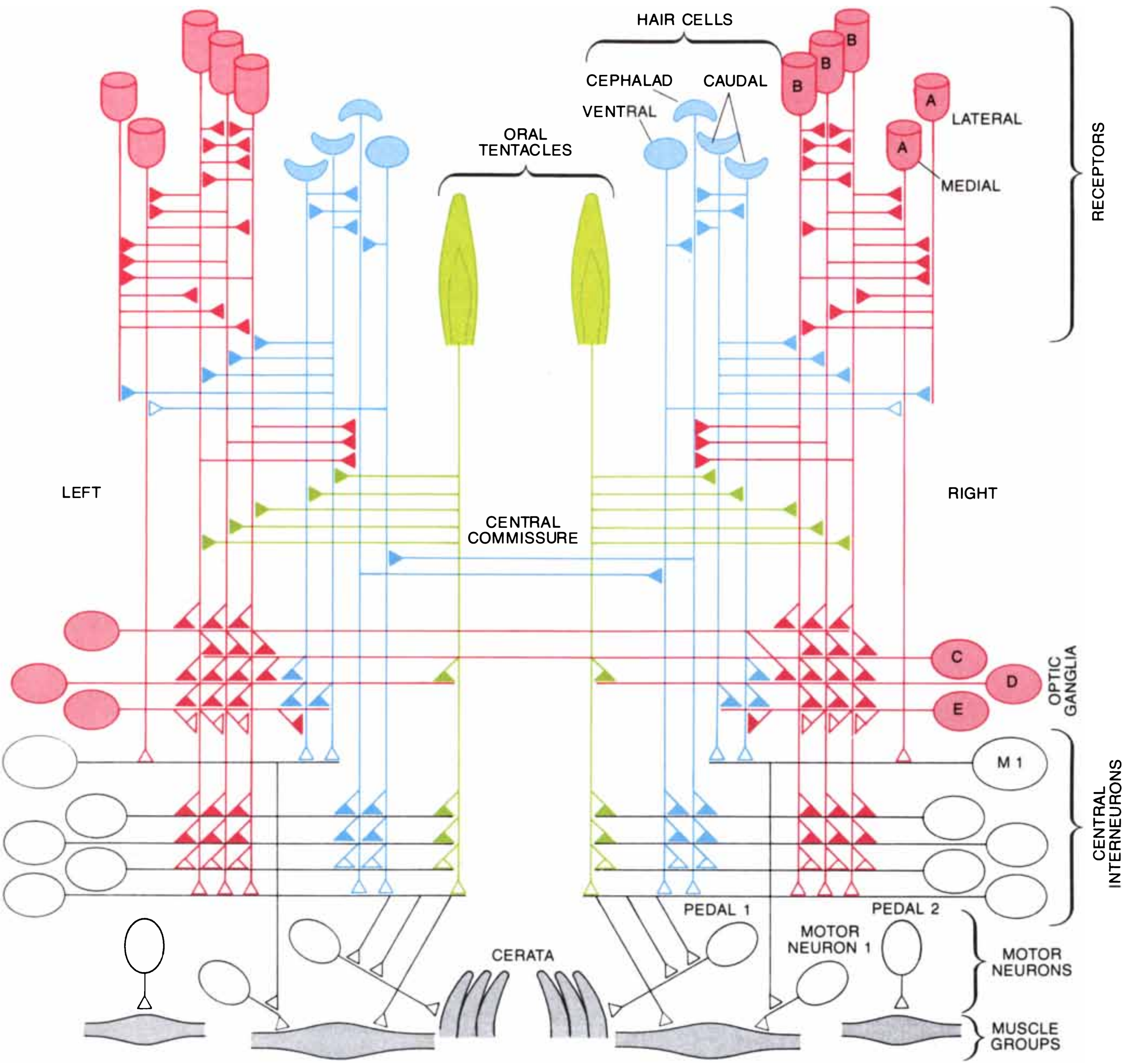
hair-cell activity is reduced even below the unstimulated, tonic level and the *B* cells are released from inhibition. When rotation is paired with light, there are additional effects. Currents that flow when the *B* cells transduce light reduce the electrical resistance of the photoreceptors' membrane, "shunting" the inhibitory signals from the hair cells and thereby lessening the effect of rotation-induced inhibition. Moreover, excitatory signals from second-order visual cells (the *E* cells of the optic ganglion) to the Type *B* cells increase in frequency immediately after rotation and light are paired. The effects of enhanced excitation accumulate in the course of training and the *B* cells become tonically more excitable: their response to light is progressively enhanced and prolonged.

Now, a salient effect of *B*-cell activity, as Yasumasa Goh showed in our laboratory, is to inhibit a chain of neurons that begins with the medial Type *A* cell and is responsible for motor-neuron impulses driving the muscle contractions that move the animal toward light. An enhanced and prolonged *B*-cell response has the effect of increasing the inhibition and thereby reducing the movement toward light.

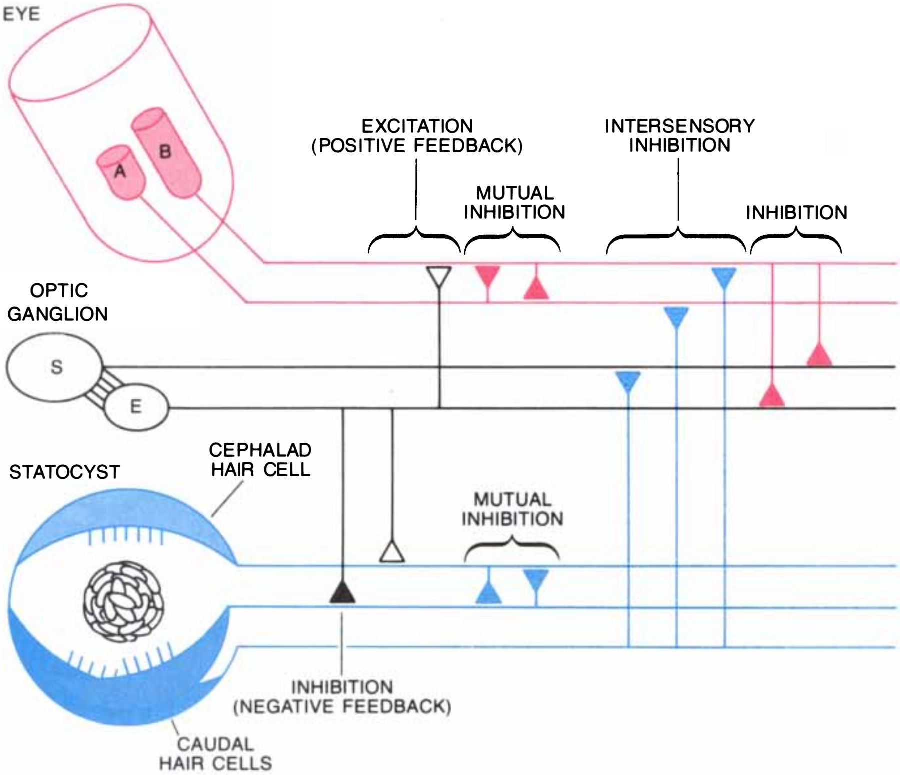
Causality

A number of experimental findings demonstrated that the modification of the Type *B* cell's response is not merely correlated with the learning behavior but is at least in part responsible for causing the behavior. These confirming experiments were "blind," that is, the investigators did not know which training procedure particular animals had undergone. In one study we measured the activity of motor neurons mediating the turn toward light. As predicted, the motor neurons of conditioned animals (in whom the inhibitory input originating with the Type *B* cells would be expected to be enhanced) indeed showed reduced activity when the animals were exposed to light.

In another experiment on conditioned and control-group animals during the



WIRING DIAGRAM of *Hermissenda*'s nervous system is summarized schematically. Representative cells (and their axons) of the visual (*red*) and vestibular (*blue*) pathways are shown, as are some axons of the tactile-chemosensory pathways (*green*) and some interneurons, motor neurons and muscles (*black*). Excitatory synaptic interactions are indicated by open triangles, inhibitory ones by solid triangles. Each interaction has been shown to be reliably present by repeated simultaneous recordings of presynaptic and postsynaptic potentials.



SITES OF CONVERGENCE between the visual and the vestibular pathways are indicated in this diagram of representative cells responsive to light and rotation. They include a "silent" optic-ganglion cell (*S*) whose effects are transmitted by way of a cell (*E*) to which it is coupled electrically. The caudal hair cells exhibit steady ("tonic") activity tending to inhibit Type *B* photoreceptors. Paired presentation of light and rotation is followed by reduced hair-cell inhibition of the Type *B* cell and by increased synaptic excitation of it by the *E* cell (positive feedback).

period of retention, the axon of Type *B* cells was cut close to the cell body, which was thus isolated physically and electrically from the rest of the nervous system. The excitability of the isolated cells in response to light or to the injection of a positive current was assessed. Several different indexes of excitability were increased in the Type *B* cells of the animals that had been exposed to paired light and rotation. The conditioning-induced increase in *B*-cell excitability, in other words, is intrinsic to (and is stored in) the cell body; it is not a passive reflection of changes elsewhere in the system.

Farley and I also demonstrated causality by manipulating the snail's neural circuits artificially. By injecting microcurrents into individual cells of the two converging pathways we caused the cells to send the same signals they send in response to the natural stimuli. In one experiment we worked with the nervous system isolated from the snail's body, injecting currents into a hair cell and a second-order visual cell of the optic ganglion. Repeated current injections, paired with light, gave rise to increased *B*-cell excitability; unpaired with light, the same injections (which now simulated one of our control procedures in training) had no such effect. In another experiment we achieved the behavioral effect of conditioning simply by pairing light with repeated injections, into a single Type *B* cell exposed in living animals, of a current calculated to simulate the synaptic effects on a *B* cell when hair cells are stimulated by rotation. We found that *B*-cell excitability was duly increased. On days following the injection, after the animals had recovered from the procedure, the velocity of their movement toward a light source was reduced, just as it was in the case of conditioned animals. Control animals that received the current injections unpaired with light displayed no such change.

A final demonstration of the causal role of the increased excitability of Type *B* cells came from confirmation of a behavioral prediction. The connections among visual and vestibular cells are such that rotating the animal with its head toward the center of rotation stimulates what are called the caudal hair cells, with the result that the Type *B* cells are eventually made more excitable, as described above. Training the animal with its head in the opposite direction instead stimulates what are called the cephalad hair cells and so results in increased inhibition of the Type *B* cells; the phototactic response, we reasoned, should be enhanced. We tried it and the prediction was borne out: the Type *B* cells became less excitable and the animals moved more quickly toward light. (Shaking without restriction in orientation, as in a turbulent sea, has the net effect of exciting the Type *B* cells because the caudal hair cells have more direct and more effective synaptic connections with the photoreceptors than the cephalad cells do.)

Membrane Changes

Our neuroanatomic results constituted the first evidence for any animal, vertebrate or invertebrate, that lasting associative learning is a direct result of a change in the excitability of particular, identified neurons. Excitability is a property of the nerve-cell membrane, and so the obvious question became: What happens to the Type *B* cell's membrane in the course of conditioning? Before discussing our findings I need to explain briefly how changes in the neuronal membrane generate electrical signals. At rest a neuron—the cell body and its processes—is polarized: the inside of the cell is about 70 millivolts negative with respect to the outside. The potential difference is maintained by the concentration of ions within and outside the cell; the concentration is regulated by the selective permeability of the cell's membrane, which is studded with gated protein channels that are specific for each of several different ions. The major factors in establishing the membrane potential are positive potassium and sodium ions. There is a much higher concentration of potassium inside the cell and a much higher concentration of sodium outside. The membrane is normally much more permeable to potassium than it is to sodium, and so the potassium ions tend to leak out, leaving the interior of the resting cell negative.

Ionic channels in the cell membrane can open in response to a sensory stimulus (in the case of a receptor cell), to the effect of a chemical transmitter on the postsynaptic membrane or to a shift of membrane potential. When sodium channels open, sodium ions (and in the Type *B* cell also positive calcium ions) flow in, depolarizing the cell (making the inside more positive). Almost immediately this depolarization causes potassium channels to open, and the outflow of potassium acts to restore a negative resting potential. These changes of membrane potential can spread locally along the membrane and in some cases give rise to a much larger depolarization that can propagate regeneratively along the entire length of the axon; this is the action potential, or nerve impulse.

We found that the outflow of potassium ions (through what are called *A* channels and probably through other potassium channels too) during a depolarization induced by a current injection in the absence of light is reduced on days after conditioning; the inflow of calcium ions (through other channels) follow-

You can count on 3M diskettes. Day after day.

Just like the sun, you can rely on 3M diskettes every day. At 3M, reliability is built into every diskette. We've been in the computer media business for over 30 years. And we've never settled in. We're constantly improving and perfecting our product line, from computer tape and data cartridges to floppy disks.

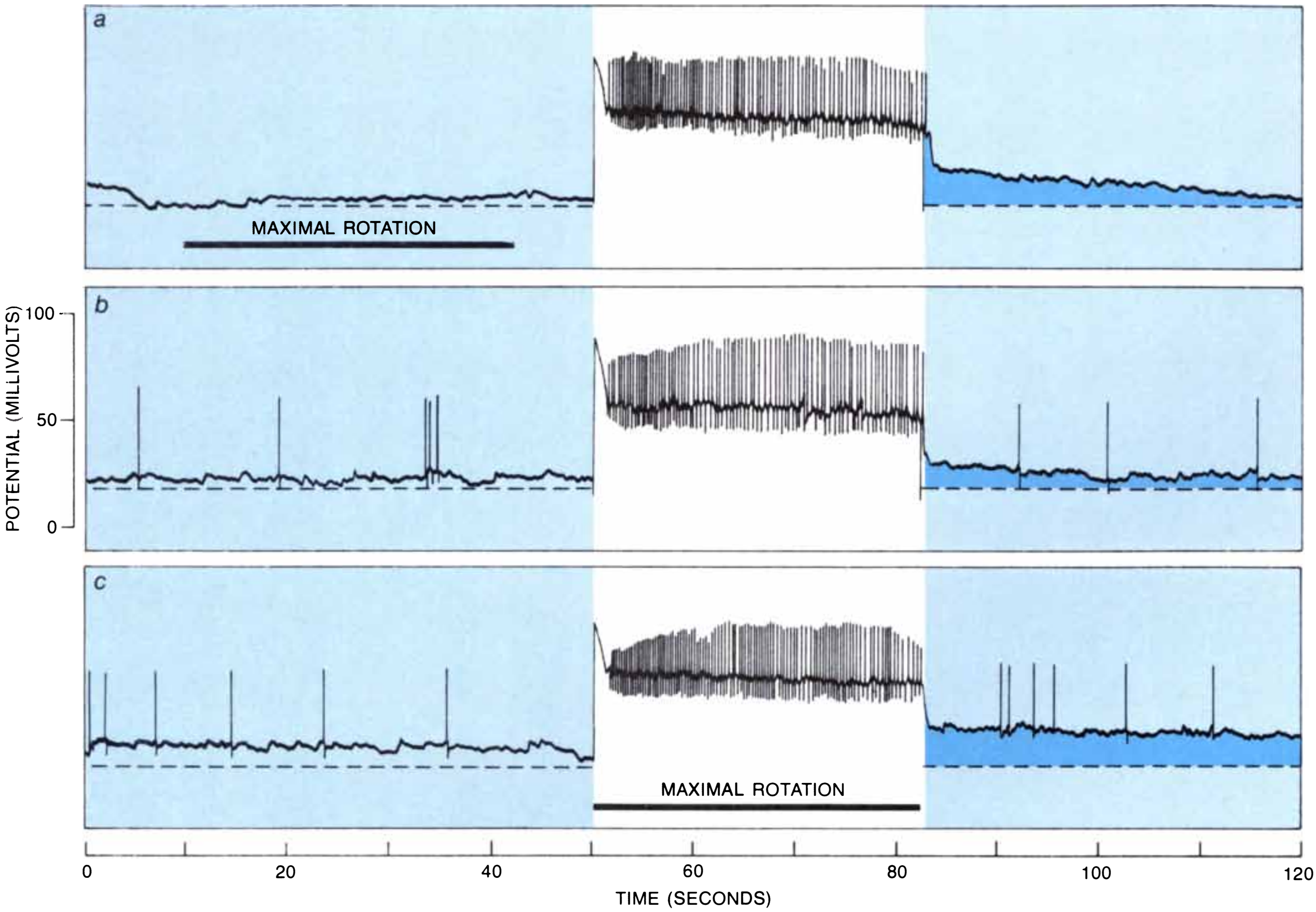
3M diskettes are made at 3M. That way, we have complete control over the entire manufacturing process. And you can have complete confidence in the reliability of every 3M diskette you buy.

Look in the Yellow Pages under Computer Supplies and Parts for the 3M distributor nearest you. In Canada, write 3M Canada, Inc., London, Ontario. If it's worth remembering, it's worth 3M diskettes.

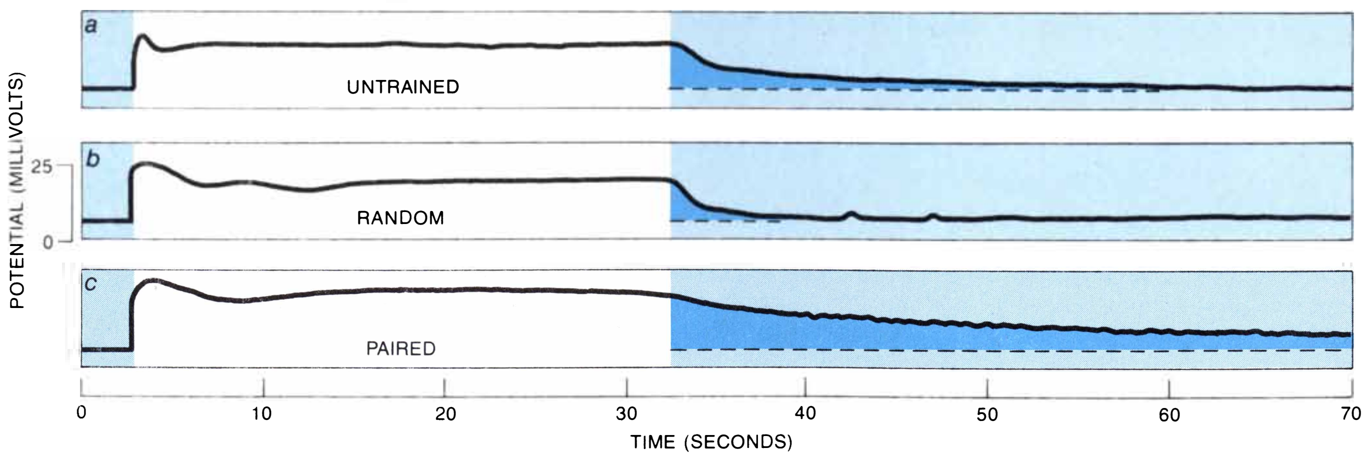
3M hears you...

3M

ing light-induced excitation is increased. Both of these changes cause the Type *B* cell to depolarize more in response to excitation by light or a current: it is more excitable. (It may be, although it is not yet established, that the increased calcium flux is simply the result of the increased excitability arising from the reduced flow of potassium.) John A. Connor of Bell Laboratories and I found, moreover, that calcium ions tend to accumulate within the Type *B* cell during the acquisition of the conditioned behavior. Prolonged elevation of intracellular calcium reduces the number of open *A* potassium channels. The reduced outflow of potassium makes the membrane more excitable. In the presence of light it depolarizes more readily; more calcium channels open, enhancing the flow of calcium and thereby further



TYPE *B* CELL'S RESPONSE to light alternating with rotation (*a*), light alone (*b*) or light paired with rotation (*c*) is shown. In each case the second of two successive periods of light is shown (*white area*); the first had ended 40 seconds before these traces begin. The cell's initial resting voltage is indicated by the broken line; the amplitude of the long-lasting depolarization that follows the second exposure to light is emphasized by the dark color. The paired stimuli can be seen to cause the most long-lasting depolarization of the Type *B* cells (*c*).



CHANGE IN EXCITABILITY is shown by these recordings to be a property of the Type *B* cell itself. The voltage of three *B* cells (whose axons had been cut to isolate them from all other neurons) was measured. The cells came from an untrained animal (*a*) and, on days after training, from an animal that had been subjected to light and rotation at random (*b*) and from one that had been conditioned by paired light and rotation (*c*). The long-lasting depolarization (*dark color*) following exposure to light is largest in the conditioned animal (*c*).

SCIENCE/SCOPE

Vacuum-tube computers spanning half an acre of floor space will be replaced by modern computers the size of two vending machines when North America's new air defense system goes into operation late this year. Hughes Aircraft Company's Joint Surveillance System will replace aging SAGE (Semi-Automatic Ground Environment) and BUIC (Back-Up Interceptor Control) systems. It will link U.S. Air Force surveillance radars, civil air traffic control radars, and Canadian radars into a shared system. Seven regional control centers -- each equipped with the smaller computers -- will monitor skies 200 miles beyond North American borders. An eighth center will monitor skies surrounding Hawaii.

The U.S. Army will save over $200 million by using simulators to train troops to use and repair Firefinder weapon-locating radars. The Firefinder detects and tracks enemy artillery and mortar fire with a pencil-thin electronic beam. It instantly backplots their trajectories so counterfire can be directed with pinpoint accuracy. The Firefinder trainer simulates battlefield conditions so troops can learn to operate the radar without using live artillery fire and without taking a radar out of deployment for instruction. Also, where only one student could operate an actual radar, six students can train at once under the control of one instructor. Hughes builds the Firefinder radars and trainers.

Mexico will inaugurate a national communications satellite system in 1985 with the launch of two Hughes spacecraft from NASA's space shuttle. The satellites will carry advanced telecommunications services for the entire country. Plans include educational and commercial TV programs, telephone and facsimile services, and data and business transmissions. Mexico now leases communications capacity on two other Hughes spacecraft -- an Intelsat IV from the International Telecommunications Satellite Organization and a Westar from Western Union. There are 157 satellite receiving stations operating throughout Mexico.

Paperless planning is making its debut to guide assemblers with step-by-step instructions for manufacturing electro-optical hardware. Color video monitors having twice the quality of home TV sets are replacing thick planning books at Hughes. The monitors are used in conjunction with video discs that hold up to 50,000 full-size color pictures, each equivalent to one sheet of planning. The discs store three-dimensional computer graphics or standard video. Assemblers can review still images or sequences showing how a product is to be built. Paperless planning is faster, more accurate, and less costly than conventional methods. It will also reduce manufacturing errors.

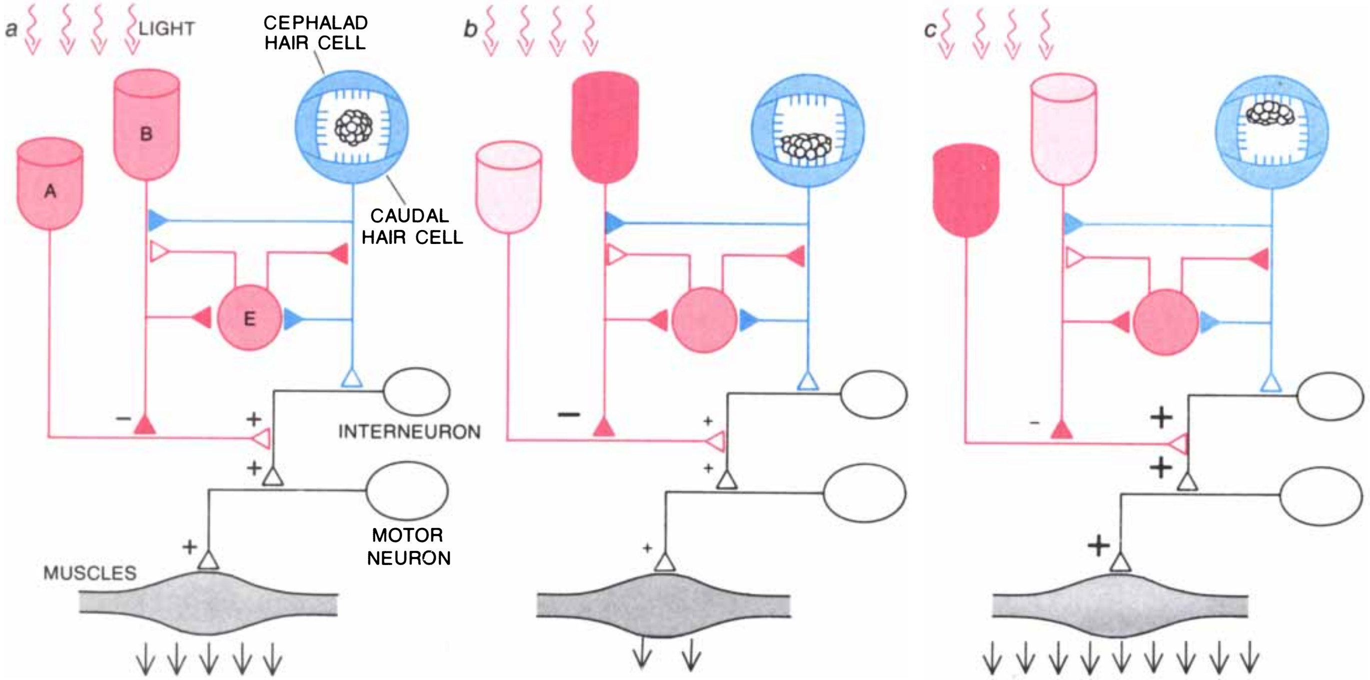
Hughes Missile Systems Group in Tucson, Arizona, has built a reputation for the development and manufacture of sophisticated electronics. We have enhanced our leadership in this field with these missiles: the electro-optical/laser/infrared-guided Maverick, the radar-guided Phoenix, the wire-guided TOW, and now the radar-guided AMRAAM. We have openings for microwave specialists, circuit designers, hybrid microelectronic engineers, and production support test engineers. Send your resume to Professional Employment, Dept. SE, Hughes Missile Systems Group, P.O. Box 11337, Tucson, AZ 85734. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For more information please write:
P.O. Box 11803, Los Angeles, CA 90291



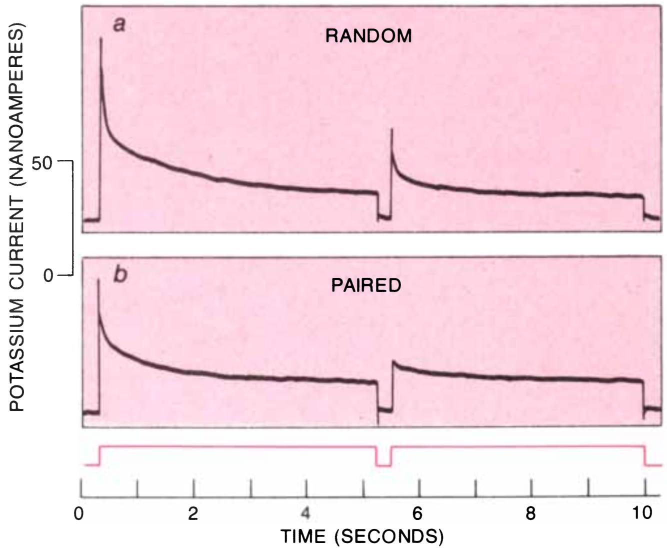
ANIMAL'S RESPONSE TO LIGHT is changed by the effects of conditioning on the neural system. The basic wiring diagram is invariant: the Type *B* photoreceptor inhibits the medial Type *A* photoreceptor, whose stimulation by light causes excitation of the interneuron, motor neuron and muscles in turn. During a burst of hair-cell impulses the *B* cell is inhibited (directly and by way of the second-order *E* cell). Following a burst the hair cell becomes quiescent and the *E* cell more active; the *B* cell is consequently excited. This is the situation in the untrained animal (*a*); the synaptic effects are enhanced when light and rotation are paired even once. Repeated pairing of light with rotation, with the snail oriented to excite caudal hair cells (negative conditioning), results in a long-lasting increase in the excitability of the *B* cell, increasing its inhibition of the Type *A* cell during and after exposure to light and thus slowing movement toward light (*b*). Positive conditioning (with the animal in the opposite orientation) excites cephalad hair cells and has the opposite effect (*c*).

increasing intracellular calcium, which further reduces the number of open potassium channels, and so on in a positive-feedback cycle.

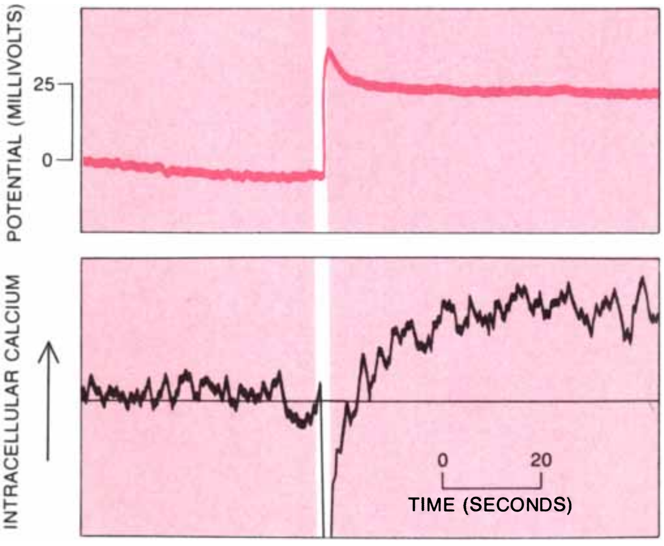
The cycle is reinforced during learning. With successive repetitions of stimulus pairings at sufficiently brief intervals, there is a progressively larger residual increase of Type *B* depolarization, which in turn seems to amplify the interaction of the calcium and potassium channels. It is possible too that the channels undergo complementary changes in synaptically linked neurons. We have preliminary evidence that the membrane of the medial Type *A* photoreceptor, the one inhibited by Type *B* cells, is less excitable after conditioning—the inverse of the situation in the Type *B* cells.

Biochemical Mechanisms

If the calcium level in Type *B* cells rises during conditioning and the rise



POTASSIUM OUTFLOW is reduced in the Type *B* cells of a conditioned animal on days after training. A change of potential (a 50-millivolt increase) was imposed on a *B* cell (*colored trace*) and the resulting current equivalent to the outflow of potassium ions was measured (*black traces*). The flow is larger in a cell from an animal exposed to light and rotation at random (*a*) than it is in a cell from an animal conditioned by paired light and rotation (*b*); the difference is greater for the second of the two successive changes in voltage.



INTRACELLULAR CALCIUM LEVEL is increased following a light stimulus. A Type *B* cell was injected with a dye that combines with calcium to form a complex absorbing light at a specific wavelength; the absorbance is recorded to serve as a measure of the calcium concentration in the cell. In response to a flash of light (*white strip*) the interior of the cell becomes more positive (*colored trace*). The calcium level (*black trace*) increases, and it remains elevated as long as the photoreceptor's membrane potential remains elevated.

is accompanied by changes in the flow of ions through membrane channels, we hypothesized, perhaps the calcium affects the channels by initiating or stimulating certain biochemical reactions. Calcium is implicated in many regulatory mechanisms, notably the activation of enzymes responsible for attaching phosphate groups to proteins; the phosphorylation in turn often changes the character of the protein. It seemed possible that elevated intracellular calcium activates enzymes that attach phosphate groups to membrane proteins controlling the opening and closing of *A* potassium channels and calcium channels.

A number of observations support the hypothesis. First Joseph T. Neary measured the incorporation of phosphate groups (labeled with a radioactive isotope) into certain proteins that appear to constitute or to be associated with potassium and calcium channels in the membrane. The net number of phosphate groups incorporated in one of those proteins is different in conditioned animals from what it is in control animals. Paul Greengard of the Yale University School of Medicine had found earlier that the phosphorylation of certain membrane proteins in several vertebrate species is mediated by calcium. Recently Leonard K. Kaczmareck and Susan A. DeRiemer of Yale, collaborating with Greengard and us, were able to link Neary's finding with Greengard's. They showed that in *Hermissenda* the phosphorylation of several nervous-system proteins, including the one Neary had found to change with conditioning, depends on the calcium concentration.

We tested calcium's primary role in increasing excitability by injecting the ion into Type *B* cells. The injection caused a prolonged reduction of potassium flow through the *A* channel, just as conditioning does. Finally, following a suggestion made by Howard Rasmussen of Yale, we injected into *B* cells a calcium-dependent protein kinase, an enzyme that catalyzes calcium-mediated phosphorylation. The injection enhanced the effect of calcium on the *A* channels. It simulated the effect of conditioning, giving rise to the same increased excitability due to a reduced flow of potassium ions and an increased flow of calcium ions.

In summary, our biochemical studies suggest that during acquisition the intracellular calcium level is progressively elevated. The ion activates calcium-dependent enzymes that facilitate the phosphorylation of membrane proteins regulating the flow of calcium and potassium ions. After training ends, intracellular calcium subsides to a normal level, but throughout the period of retention of learned behavior the phosphorylation of proteins persists. (As Rasmussen has pointed out, calcium-mediated phosphorylation can be a fairly irreversible process.) The high level of intracellular calcium during training could cause a long-lasting activation, through phosphorylation, of protein kinases. The activated enzymes could then be responsible for the persistent membrane changes that maintain increased excitability.

Our findings demonstrate, for the first time for any nervous system, that associative learning lasting for several days involves changes in the long-term biochemical regulation of ion flow across the cell membrane. The changes we find are at cell bodies and along axons, not at synapses. It may turn out that synaptic changes are indeed implicated in associative learning, but we have not found evidence for such changes. Alterations of nonsynaptic membrane permeability can account for the learned behavior we have studied in *Hermissenda.*

Learning in Vertebrates

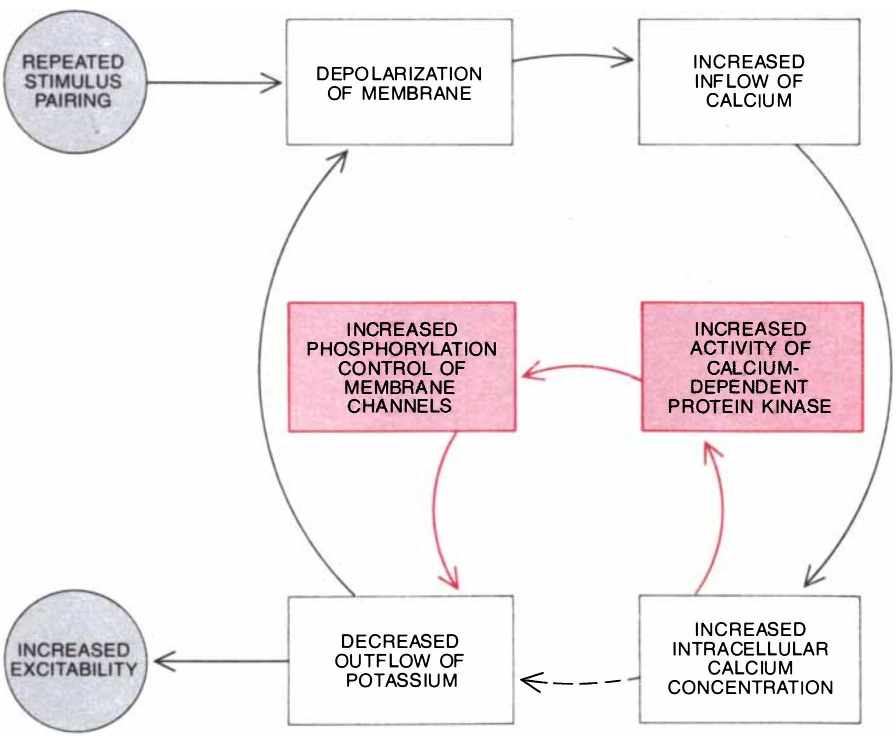
Marine organisms serve as marvelous models, revealing basic facts about many aspects of nervous function that have consistently proved to be generally applicable in higher animals. We have begun to test the applicability to the vertebrate brain of what we have learned about the biochemical basis of associative learning in *Hermissenda*. Recently Charles D. Woody of the School of Medicine of the University of California at Los Angeles, Bruce A. Haye and I injected the protein kinase I mentioned above (the enzyme whose injection into Type *B* cells simulated the effects of conditioning) into certain neurons in the brain of the cat. The activity of these neurons has been correlated with learning. The enzyme injection gave rise to changes in the excitability of the cat-neuron membrane that are similar to those generated in the Type *B* cell.

Such persistence through evolution of a common biochemical mechanism does not, to be sure, imply identity of function. Even if one considers only associative learning, the gulf between the nervous function of a snail and that of a human being is wide, as I said at the outset. To begin with, there is a vast difference in capacity to resolve sensory information and to associate stimuli. Lower species can learn to associate only a narrow range of stimuli; human beings can associate whatever they perceive. Human consciousness makes possible willful forgetting or repression; it places perceptions in a rich emotional context, with gradations of positive and negative feelings that invest associations with value and meaning.

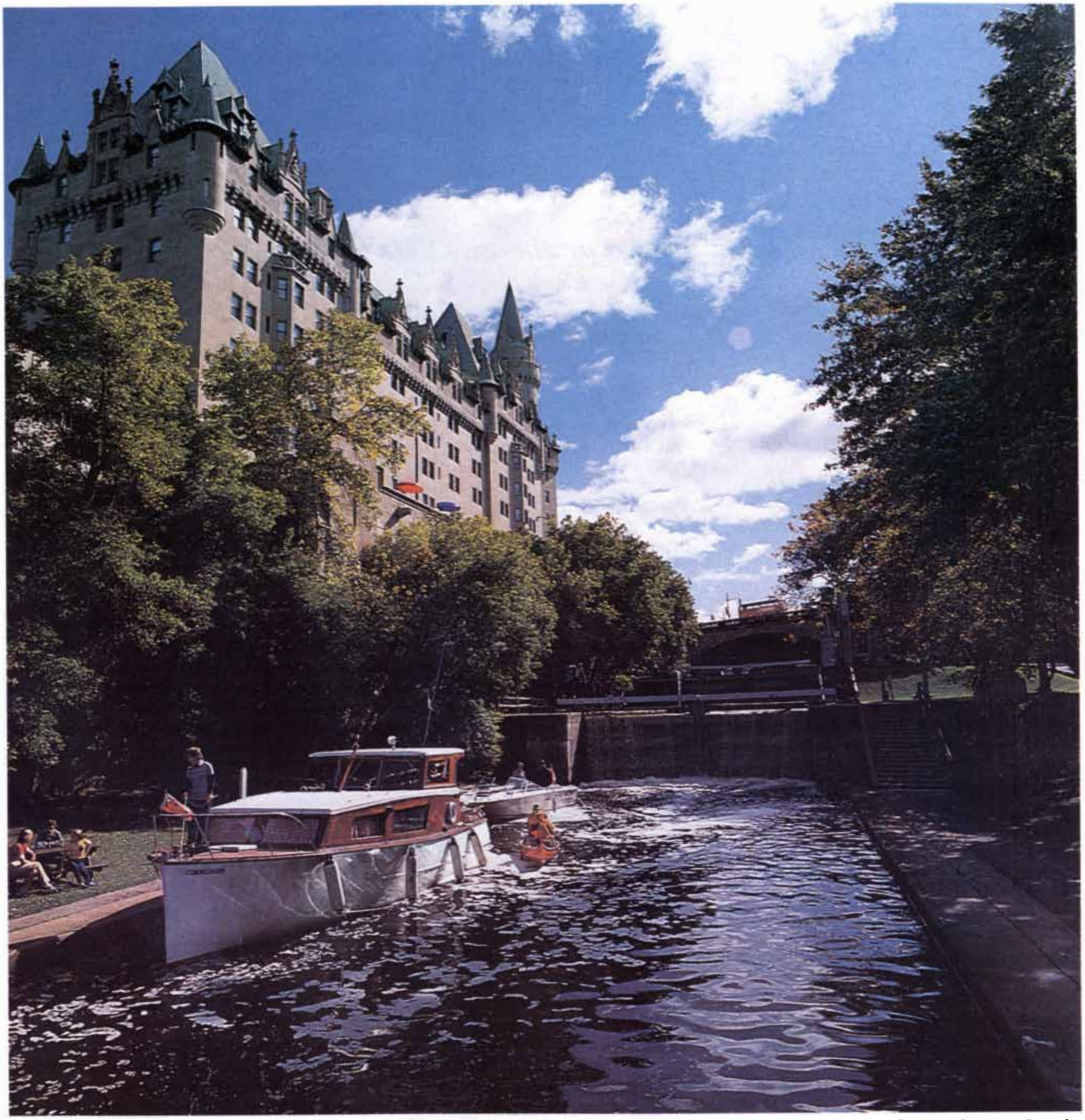
For all of that, much of the biophysics and biochemistry of learning may be quite similar in snails and in human beings. It may be the neural circuitry that makes the difference. Recall that in *Hermissenda* the Type *B* photoreceptors become more excitable in the course of conditioning because fewer potassium channels and more calcium channels open in response to a stimulus—because, in turn, of the increased activation of an enzyme by calcium ions. The Type *A* photoreceptors have the same kind of channels, however, and the same potential for undergoing changes in excitability. It is the wiring diagram of the visual and vestibular pathways that makes the Type *B* cells receive more excitation and so become more excitable when the snail is trained with its head toward the center of rotation. The same wiring diagram determines that the Type *A* receptors, not the Type *B* cells, become more excitable when the animal is trained in the opposite orientation. It is the impact of a particular combination of stimuli, with a particular temporal relation, on a particular genetically programmed wiring diagram that excites one set of cells and inhibits another set.

Consider that billions of neurons in the human brain may have the same potential as the *Hermissenda* cells to become more excitable or more inhibited. It is not an unreasonable assumption; similar potassium and calcium channels have been characterized in invertebrates and in higher vertebrates. Consider also that the relative amount of excitation received by each neuron in a subpopulation of human brain cells must depend on a preexisting wiring diagram and on the nature and timing of sensory stimuli. Just as the relative excitability of photoreceptors is altered when *Hermissenda* learns, so the relative excitability of neurons in the human brain may undergo long-lasting shifts determined by conscious experiences. The fundamental biophysical and biochemical changes responsible for storing learned information need not be very different in human beings from the mechanisms now being revealed for *Hermissenda.*

On the other hand, just what can be learned by an animal, and how much, is structured and limited by the wiring diagram of its neural system. The enormous number of neurons in the human brain and the complexity of its circuitry may provide the potential for recording subtle differences among stimuli, for associating stimuli and for associating the associations, that is, for learning by progressive abstraction. The difference between what people can learn and what a snail can learn may result from the difference in wiring diagrams rather than from any special properties of the cell membrane or any special biochemical control mechanisms that uniquely distinguish human neurons from those of a snail.



POSITIVE-FEEDBACK LOOP accounts for the increased excitability of the Type *B* cell's membrane. At the biophysical level (*black boxes*) increased depolarization during conditioning increases the calcium flux; the increase in intracellular calcium reduces the flux of potassium, further depolarizing the membrane. Two biochemical regulatory steps are shown (*colored cycle*). The high calcium concentration activates a calcium-dependent protein kinase; the enzyme phosphorylates proteins that regulate the membrane channels, reducing potassium flux.



Ottawa, Ontario/Canada

Microelectronic Packaging

Methods of housing, cooling and interconnecting the silicon chips in a digital computer have an important bearing on the machine's performance. The aim is to fit the most chips in the least volume

by Albert J. Blodgett, Jr.

The packaging of the microelectronic silicon chips in a high-speed digital computer might seem only incidental to the machine's design. One could argue that the actual processing of information is done entirely by the circuitry on the chips themselves; the functions of the packaging are simply to interconnect the chips and other devices, to distribute electric power and to provide a suitable operating environment. Clearly, every computer needs packaging of some kind, but the nature of the packaging would not appear to have much influence on the functioning of the machine. The facts are otherwise. In many high-speed data-processing units packaging technology is the factor that determines or limits performance, cost and reliability.

One reason packaging has become so important is the imperative to make the central elements of a computing system exceedingly compact. Improvements in the design and fabrication of microelectronic devices have greatly increased the number of logic functions that can be put on a chip as well as the speed at which arithmetic operations are performed. As a result a major source of delay in the central processing unit of many computers now is the time needed for a signal to pass from one chip to another. In order to reduce the delay the chips must be put closer together. Fitting many chips, each with many terminals, into a small volume challenges packaging technology in several ways. First, there is little space available for the thousands of conductors needed to distribute electric power and information-bearing signals among the chips. Furthermore, the electrical properties of this network of conductors must be designed to minimize the distortion of signals, which becomes more difficult as switching speeds are increased and dimensions are reduced. Finally, a dense array of chips gives off a fair amount of heat, which must somehow be removed if the circuits are to operate properly. In many instances the thermal problem is the most challenging one.

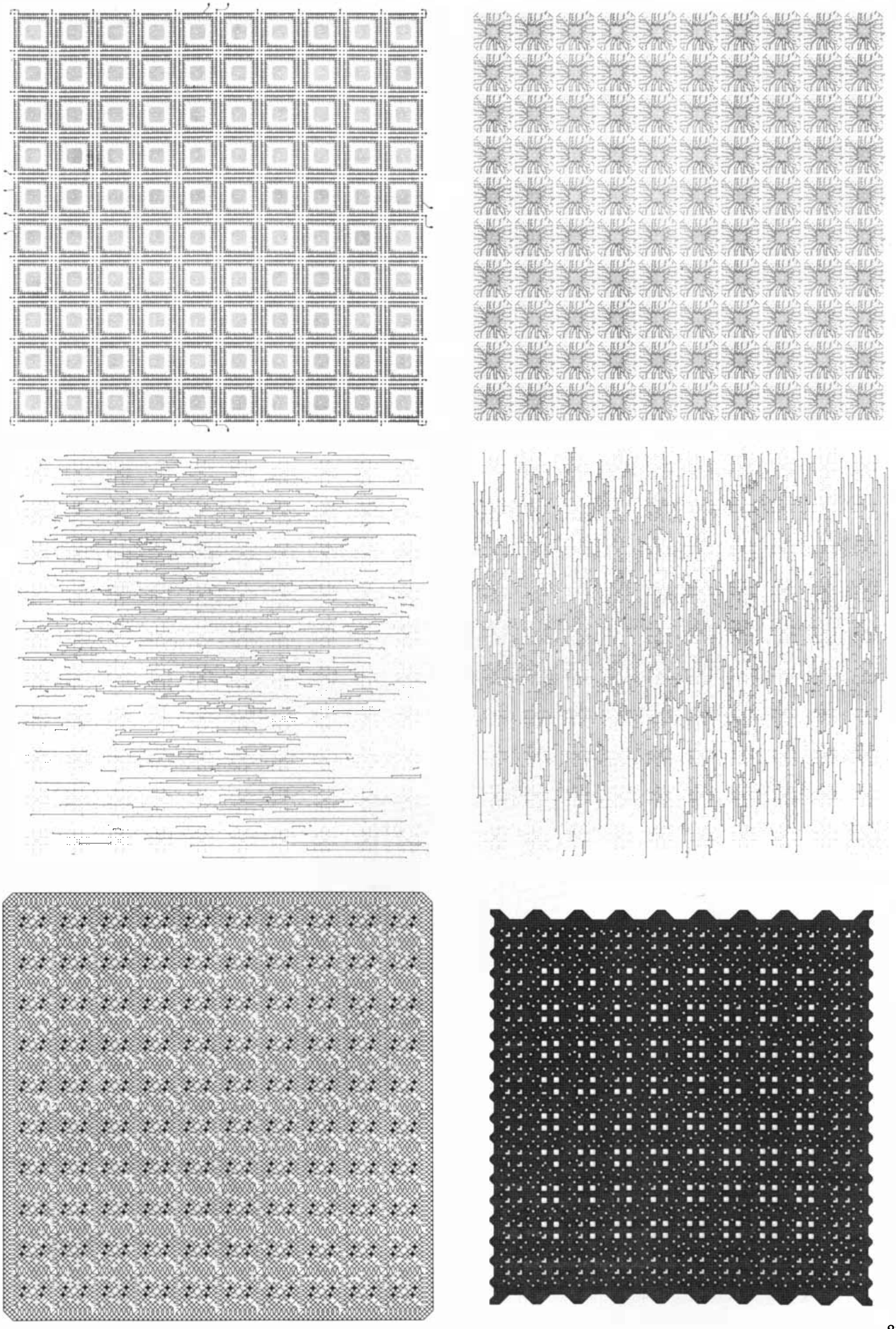
The designers of high-performance digital computers have devised a number of ingenious and quite varied solutions to these problems. Here I shall discuss some general considerations that go into the design of any packaging technology. I shall also describe in some detail the system developed for one high-performance system, the model 3081 processor of the International Business Machines Corporation. It is large "mainframe" computers of this kind that make the severest demands on packaging methods and materials. The technology I shall describe is capable of interconnecting more than 100 chips—which might incorporate circuitry equivalent to the entire central processing unit of an earlier computer—in a module that can be held in one hand.

For more than a decade the hierarchy of packages in a typical mainframe computer has had three levels. Each chip is permanently mounted on an individual chip carrier, a plastic or ceramic housing with a dozen or more metal leads. A number of chip carriers, along with a few discrete components such as resistors and capacitors, are mounted on a card made of fiberglass impregnated with an epoxy resin. Metal conduction paths are printed on the surface of the card, and in some cases the card is a laminate with multiple layers of printed wiring. Several cards are mounted through connectors along one edge to a larger board whose printed circuitry establishes connections between the cards. The entire assembly of chip carriers, cards and main board is tied to the rest of the system (including other assemblies of the same kind) by multiconductor cables.

The three-level hierarchy of packages is still the standard in many computers, most notably the ones called minicomputers. In both the lowest- and the highest-speed systems, however, technological advances have made it possible to dispense with one level in the hierarchy. In the low-performance realm of the microcomputer each chip is still housed in a separate carrier; owing to the dramatic increase in the number of circuits per chip, however, the number of chips has been reduced to the point where they all fit on one board. In some of the latest large mainframe computers another approach has been taken: many chips are bonded to a single substrate, forming a multichip module that plugs directly into a large board. Thus in both cases the card level is eliminated.

The evolution of microelectronic devices and that of packaging technology are strongly interdependent. For example, putting more circuits on a single chip generally requires adding more electrical contacts, but that is feasible only if the package can accommodate additional connections and signal paths. Similarly, chips with more circuits as a rule consume more power; they can be incorporated into a system only if the package can safely remove the addition-

LACEWORK OF CONDUCTORS interconnects 100 silicon chips in the central processing unit of a "mainframe" computer. Each of the patterns on the opposite page forms one layer of a ceramic substrate that can have as many as 33 layers. At the upper left is the top layer, to which the chips are bonded, establishing electrical connections through a dense array of contact pads in the middle of each chip site. At the upper right is one of five redistribution layers, which carry signals from the chip contact pads to engineering-change pads that surround the chip site. The two layers in the middle are signal planes that provide for communication between chips; on a given plane all the conductors are oriented parallel to either the *x* or the *y* axis. A complete substrate has 16 signal planes. Between each pair of signal planes is a voltage-reference plane, such as the one at the lower left, that helps to define the electrical properties of the signal lines. The layer at the lower right is one of three that distribute power to the chips. In the finished substrate the layers are stacked, laminated and fired, so that they fuse into a tilelike ceramic. The patterns are shown at their approximate actual size after firing.



al heat. Because of these relations the constraints that apply to packaging design today can be fully understood only by first considering the nature of the chips that go into the packages.

The crucial event in the evolution of semiconductor technology was the introduction of the integrated circuit, which combines a number of transistors and other basic circuit elements on a single piece of silicon. The first integrated circuits, in the early 1960's, included only from six to eight transistors, diodes and resistors, enough to implement a "gate," or simple logic function. A typical gate receives two or three input signals and issues a single output; the pattern of the input signals determines whether the output is on or off. Several hundred gates might be needed to build a functional unit of a computer such as an adder.

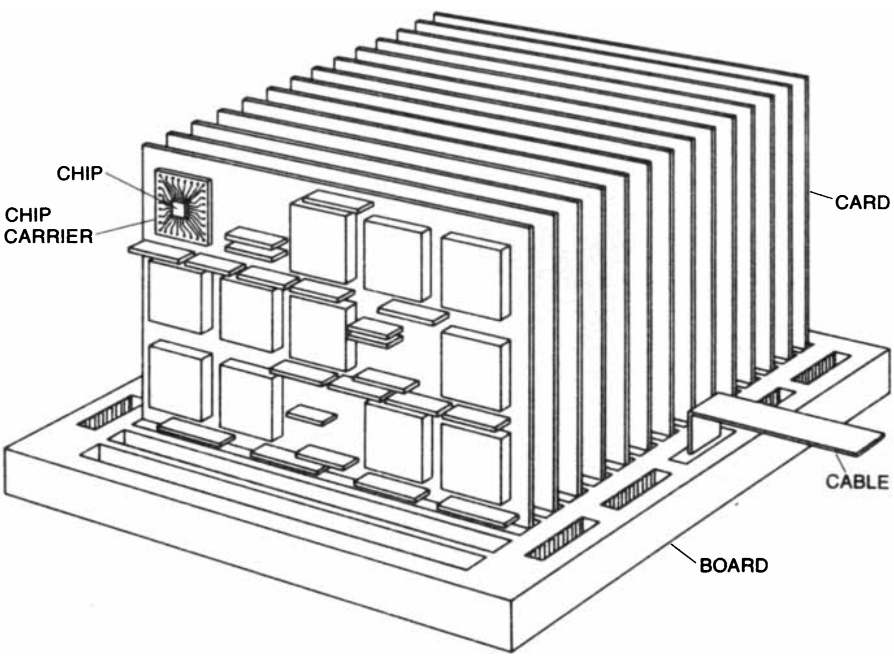
The early integrated circuits were built on a silicon chip with a surface area of a few square millimeters, and the smallest features in the pattern that defined the circuitry were from 10 to 20 micrometers across. The chip communicated with the rest of the system through eight or 10 terminals. The switching delay, which represents the time from the arrival of an input signal to the production of an output signal, was generally between 20 and 40 nanoseconds, or billionths of a second.

Advances in the art of semiconductor manufacture over the past 20 years have been dramatic. The maximum available surface area on a chip has increased by more than 10 times (to about 50 square millimeters), and the minimum feature size has decreased by a similar factor (to less than 1.5 micrometers). High-speed logic chips have been built with more than 40,000 transistors and other circuit elements, organized into 5,000 gates that switch in about two nanoseconds. A chip of this kind can have 200 electrical contacts.

The integrated circuits employed in a high-performance mainframe computer are built out of basic circuit elements called bipolar transistors. Chips based on another kind of transistor, the field-effect transistor, can fit many more circuit elements in a given area. For example, microprocessors and related devices have been built on a single chip with more than 100,000 field-effect transistors. Why, then, do mainframe computers employ bipolar transistors? The reason is that they are faster: as a rule they have a shorter switching delay.

The chips with the highest transistor count are random-access memory arrays, whose function is to store information. Early semiconductor memory devices, introduced in 1971, had 128 cells; each cell was capable of holding one bit (binary digit) of information. Today memory chips with a capacity of 256,000 bits are in production.

For the packaging designer the logic chips of the central processor present a greater challenge than memory chips. In general logic chips dissipate more power and hence give off more heat; they also require more electrical contacts for signal communication. Similarly, a chip based on bipolar transistors consumes more power than a chip with equivalent functions made with field-effect transistors. It follows that packaging design is most critical in the central processor of a high-speed mainframe computer.



HIERARCHY OF PACKAGES in a large digital computer has traditionally had three levels. Each chip is housed in an individual carrier; a number of carriers and other components are attached to a printed-circuit card; several cards are in turn mounted on a larger printed-circuit board. The assembly is connected to the rest of the system by cables. With the recent introduction of a module capable of holding more than 100 chips it has been possible to eliminate the card level and thereby to reduce the number of interconnections and the total wiring length.

Packaging begins where the chip ends: at the metallic pads on the surface of the chip where contact is made with the external wiring. The length of that wiring is among the packaging designer's primary concerns.

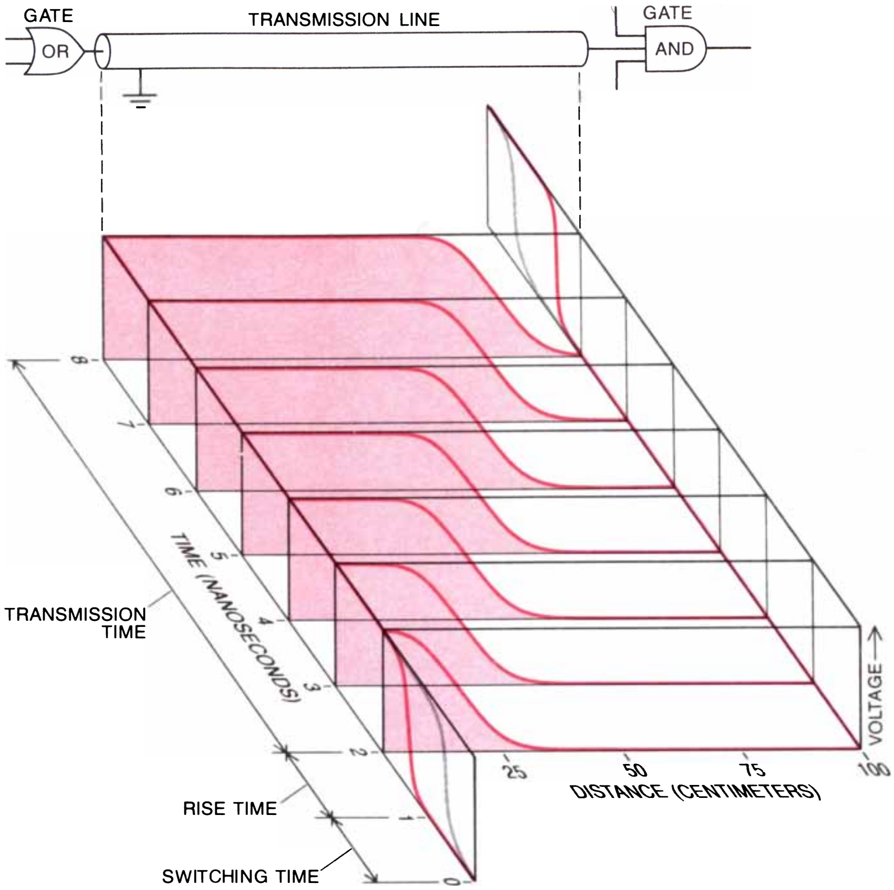
In the analysis of many electrical circuits it is reasonable to assume that a voltage applied at one end of a wire appears simultaneously at all points along the length of the wire. Actually the speed of a voltage signal along a wire is finite; it is determined by a property of the insulator surrounding the wire, namely its relative dielectric constant. If the insulator is air (or a vacuum), the relative dielectric constant is 1 and the signal travels with the speed of light in free space, which in units appropriate to this discussion is about 30 centimeters per nanosecond. In other insulators the dielectric constant is larger and the speed is reduced by a factor proportional to the square root of the dielectric constant.

For a fiberglass printed-circuit board the dielectric constant is approximately 4, and so the propagation speed is reduced by a factor of two; in other words, signals travel through the conductors in the board at about 15 centimeters per nanosecond. Because a signal may have to go appreciably farther than 15 centimeters to get from one chip to another, propagation delays can exceed a nanosecond. In slower digital devices a delay of this magnitude is insignificant because the switching delays of the logic gates are tens or hundreds of nanoseconds. In a computer built out of devices that switch in a nanosecond, however, propagation delays clearly have a major influence on the overall speed of operations. It is for this reason that minimizing wire lengths and maximizing circuit density are of critical importance in the design of packaging.

When propagation delays are important, the wire that carries a signal cannot be considered a simple conductor but instead must be treated as a transmission line. The signal is represented as a wave propagating along the transmission line, and the voltage at any point along the conductor depends on both distance from the source and time since the signal was emitted.

In a transmission line the electrical resistance of the conductor is not the only property that affects the propagation of a signal. It is also important to know the inductance, which determines the amount of energy stored in the magnetic field set up by a passing current, and the capacitance, which determines the energy stored in the corresponding electric field. The inductance and the capacitance depend on the geometry of the transmission line and on electrical and magnetic properties of the materials it is made from; together they define the impedance of the line. For a low-resistance transmission line the impedance is equal to the square root of the ratio of the inductance per unit length to the capacitance per unit length. It is measured in ohms, the same unit employed for resistance, but its effects on a propagating signal are more complicated than the effect of resistance on a steady current.

One characteristic of all waves is that they can be reflected. Similarly, a digital signal can be partially reflected from a discontinuity in the transmission line or from the end of the line. The reflection coefficient, which gives the fraction of the signal reflected, is determined by the impedance and by the load resistance that terminates the line. Suppose a given transmission line has an impedance of 100 ohms. If the load resistance is also 100 ohms, the signal is totally absorbed by the load and none of it is reflected back into the line; this is the ideal situation. If the load resistance is 200 ohms, however, a third of the signal is reflected and adds to the initial signal on the line. A load resistance of 50 ohms also yields a reflection coefficient of one-third, but the reflected signal is subtracted from the initial one. It is clear that such reflections must be controlled to prevent switching errors.



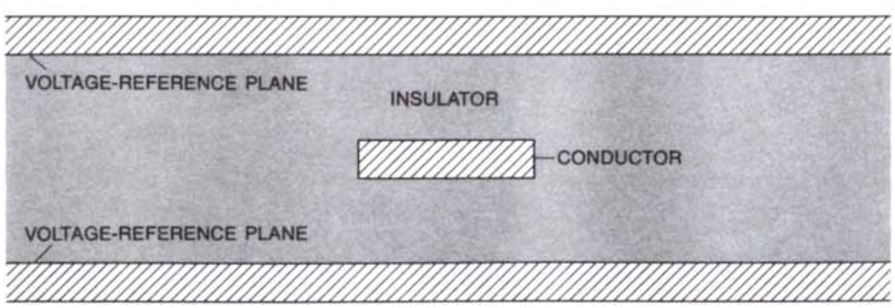
SIGNAL DELAY in a digital computer limits the maximum speed of operation. The delay has three components: the rise time of a voltage applied to a logic gate, the switching time of the gate itself and the transmission time needed for the signal to reach the next gate. When the signal must pass from one chip to another, the transmission delay, which is determined by the packaging technology, is often the longest of the three. Here the rise time and the switching time are each about a nanosecond, but transmitting the signal over a distance of roughly a meter takes six nanoseconds. The signal propagates as an electromagnetic wave in a transmission line; the wave is represented by a graph of voltage as a function of time and distance.

Reflections are only one of several ways the electrical design of a package can modify a signal or inject "noise" into a circuit. For example, two adjacent conductors can be coupled through their mutual inductance and capacitance, so that a signal sent down one line may also appear on the other. Such "crosstalk" must be avoided if the behavior of the system is to be predictable.

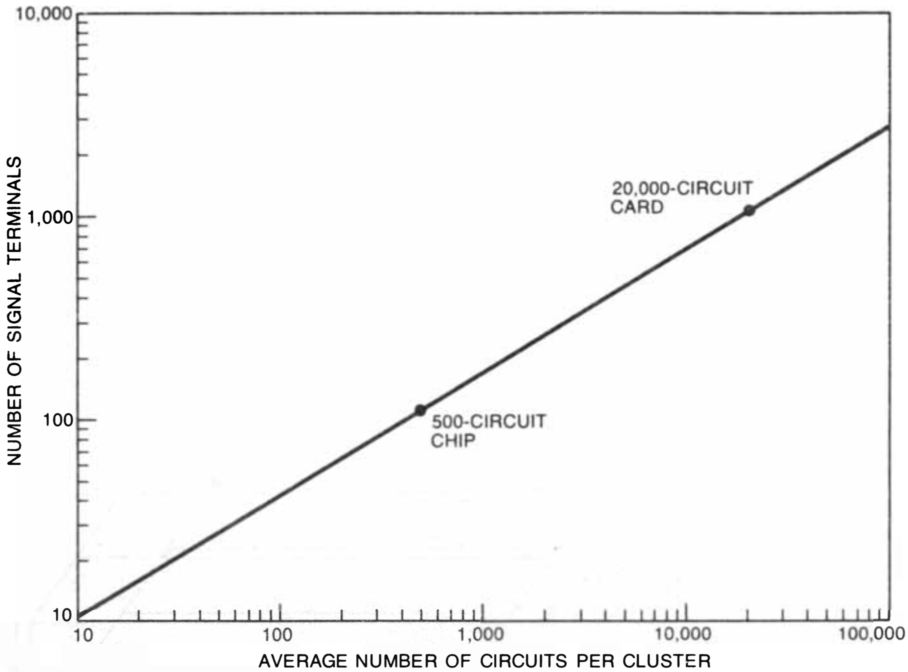
In a high-performance package the basic method of controlling the characteristics of the transmission lines is to separate layers of signal wires with conductive sheets called voltage-reference planes. The reference planes can also provide a path for return currents. Each plane is at a uniform electric potential, either zero volts (ground voltage) or one of the supply voltages needed by the chips and other components. (Hence the planes can also be employed to distribute power.) A signal line encased in an insulating medium and sandwiched between two such planes makes a transmission line whose properties can be calculated. The planes give the line a uniform and well-defined impedance and also inhibit crosstalk between lines in adjacent layers.

The design of a transmission line begins with the specification of its direct-current resistance. The resistance must be small compared with the load resistance or the input voltage will be seriously attenuated when it reaches the output of the line. The resistance per unit length is determined by the resistivity of the material and the cross-sectional area of the conductor; once the material is chosen only the latter property can be altered by the designer. For printed circuits and conductive traces fabricated by similar techniques the cross section is a flattened rectangle.

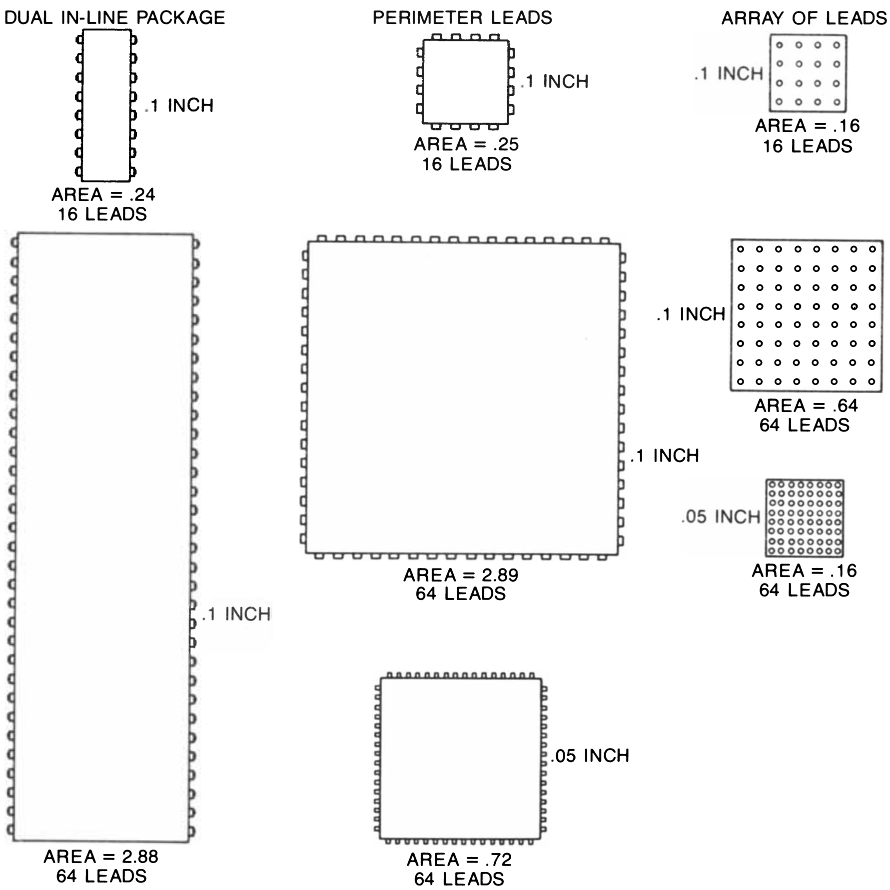
Given the dimensions of the conductor, the line impedance is determined by two additional factors: the dielectric constant of the insulating medium and the distance between the voltage-reference planes. For a particular insulating material the distance between reference planes is adjusted to achieve the desired impedance. The design value depends



TRANSMISSION LINE consists of a conductor embedded in an insulator and sandwiched between two conductive planes. The electrical characteristics of the line depend on the dimensions and the physical properties of the component parts. For example, the speed of a wave propagating through the line is determined by the dielectric constant of the insulator. The parallel conductive planes reduce coupling, or crosstalk, between lines in adjacent layers.



NUMBER OF TERMINALS needed for communication with a cluster of logic circuits can be estimated from an empirical relation called Rent's rule. In a mainframe computer a cluster of C logic circuits needs $2.5C^{.61}$ signal terminals. Thus a chip with 500 circuits would need about 110 terminals, and a card with 20,000 circuits would need about 1,000 signal terminals.



"FOOTPRINT" OF A CHIP CARRIER, or the area it occupies on a card or board, depends on the spacing and arrangement of the terminals. The dual in-line package, with two rows of pins .1 inch apart, is efficient for a chip with 16 leads but not for one with 64. A square package with leads at the perimeter saves no space if the distance between terminals remains .1 inch, but it covers only one-fourth the area when the leads are on .05-inch centers. The most efficient practical configuration for a chip carrier is a full grid of terminals. Indeed, a square carrier with 64 leads on a .05-inch grid takes up less space than a 16-pin dual in-line package does.

on many factors, including the electrical properties of the chips in the system, the dimensions and other specifications of the package and the amount of power available to drive the transmission lines. Typically the impedance is in the range from 50 to 100 ohms.

A conductor between two reference planes can only approximate an actual transmission line. In practice a signal line connecting two chips may follow a tortuous route, threading from one layer of wiring to another. At transitions between packaging levels—such as those from the chip to the carrier or from the card to the board—the electrical properties of the wiring depart significantly from the ideal. As noted above, such discontinuities can cause reflections. They also introduce additional delays, proportional to their capacitance and inductance. The extra delays must be added to the basic propagation delay of the circuit to determine the total delay introduced by the packaging system.

A single transmission line extending between two isolated terminals could be given precisely defined characteristics. The actual task of the packaging designer, however, is to create a network of intermeshed pathways to interconnect thousands of device terminals. The topological complexity of the network is formidable.

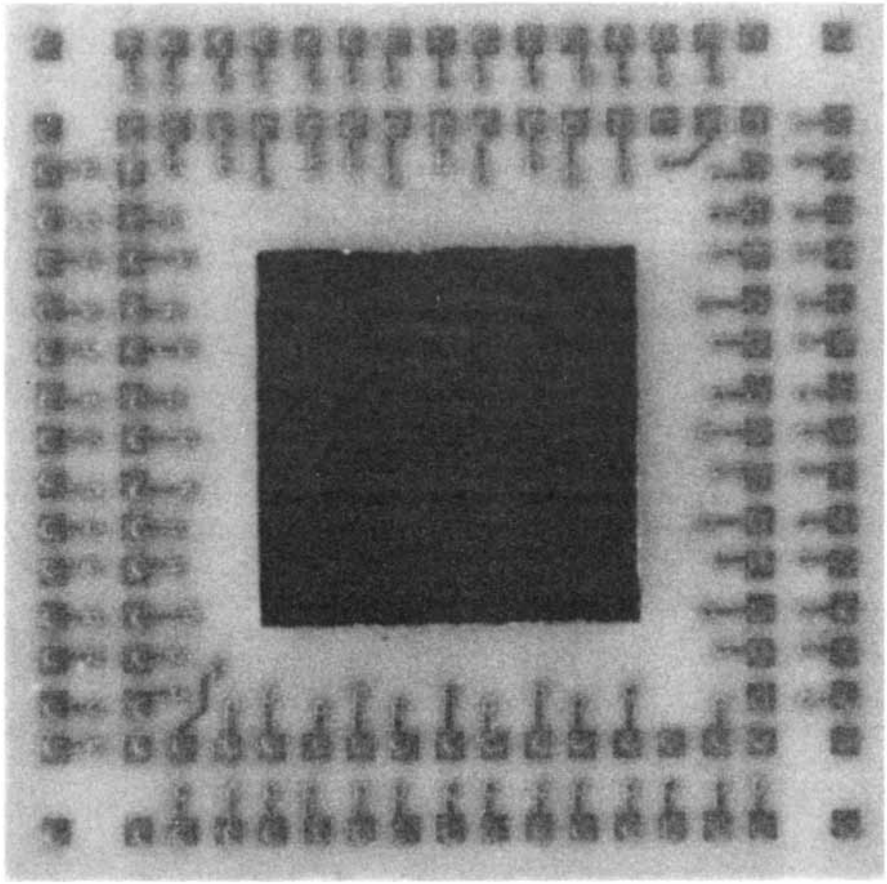
The number of signal terminals needed to establish connections at any given level in the packaging hierarchy can be estimated from an empirical relation called Rent's rule, developed in 1960 by E. F. Rent of the International Business Machines Corporation. The rule, which can be applied to chips, multichip modules, cards or boards, gives the approximate number of terminals as a function of the number of logic circuits, C. A form of the rule based on data obtained by analyzing large data-processing systems indicates that the average number of terminals is equal to $2.5C^{.61}$. Thus the number of terminals needed to support 100 circuits is about 40 and the number needed for 1,000 circuits is almost 170.

There are certain limitations to the scope of Rent's rule. First, it can be applied only to the logic elements of a processing system; arrays of memory cells need far fewer terminals. Second, each cluster of circuits must be a small, "random" subset of the entire logic complex. If a cluster constitutes a complete functional unit of the computer, then again fewer terminals suffice. Third, it is assumed that information passed between packages is not specially encoded for serial transfer, a technique that can reduce the number of terminals, although only at the cost of slower operation. The clusters of logic circuits for the mainframe computers being considered here meet these conditions.

Consider the packaging of a functional unit with 10,000 logic circuits, built up out of chips that have a maximum of 25 circuits each. The chips are housed in carriers with 14 signal pins and mounted on boards with 100 signal terminals. At first it might seem that a 10,000-circuit unit could be assembled out of 400 of the 25-circuit chips. Rent's rule shows, however, that a package with 14 signal pins can support an average of only 17 circuits; hence the actual number of chips needed is 10,000 divided by 17, or 588. (Almost a third of the capacity of the chips remains idle, suggesting that a chip carrier with more than 14 signal pins would be more efficient.)

A second application of Rent's rule indicates that the 100 signal terminals on each board provide enough communications capacity for 424 circuits. The number of boards needed is therefore 10,000 divided by 424, or 24, and each board can hold an average of about 25 chip carriers.

Rent's rule estimates only the number of packages needed at a given level in the hierarchy; the actual size of a card or board must be determined from the details of the physical design. Much depends on the lowest-level package, which is generally the single-chip carrier. The factors that need to be considered in card design include the configuration of the terminals on the carrier, the maximum density of the signal paths, the electrical performance required, the power requirements of the chips and the cooling capacity of the overall packaging design.



MULTICHIP CARRIER can greatly reduce the average distance between chips and hence the average transmission delay. Here a single chip is shown mounted face down on a multilayer ceramic substrate that can accommodate about 100 chips. Both power and signals pass through conductors buried within the substrate. The double row of pads surrounding the chip gives access to the signals, so that the assembly can be tested and connections can be altered.

A chip carrier in its simplest form is a space transformer. It serves as a bridge between the small and closely spaced contact pads on the surface of the chip and the larger wiring network on the card or board. One of the commonest chip carriers is the dual in-line package, a rectangular enclosure of plastic or ceramic with a row of leads along each of its two longer sides. The leads are arranged on .1-inch centers and are inserted into holes in the printed-circuit board. The dual in-line package makes fairly efficient use of space on the card for chips that have comparatively few terminals (say from 10 to 20), but it becomes unwieldy as the number of leads increases. For a 64-pin chip not only is the package greatly lengthened but also its width must be increased to allow space inside the package for connections between the leads and the chip contact pads.

The industry standard is now shifting from the dual in-line package to a square chip carrier with leads, or contacts, on all four sides. The leads are on .05-inch centers and are soldered to corresponding pads on the surface of the printed-circuit card. The design significantly reduces space requirements. A square chip carrier with 64 perimeter leads on .05-inch centers occupies only a fourth the area taken up by a 64-pin dual in-line package.

The densest terminal configuration for a chip carrier is a full array of pins or contact pads. Sixty-four terminals in an eight-by-eight array on .05-inch centers take up less than a fourth the space needed for the same number of leads in a square perimeter arrangement at the same spacing. The advantage of the array format becomes even greater as the number of terminals increases. Making connections to the denser array of pins or contacts, however, requires a printed-circuit board that is more complex and hence more expensive.

As the number of terminals per chip and the number of chips per unit area increase, the major limiting factor becomes the capability of the board or multichip module to provide interconnections. The simplest printed-circuit boards have only one layer of signal lines, and so the wiring must be laid out with careful attention to an obvious constraint: no two conductors can be allowed to cross. Multiple layers of wiring can eliminate this problem, particularly if all the conductors in a given layer are oriented along one major axis, either x or y. On the other hand, with a multilayer board space is needed for vias, or vertical pathways, that connect the layers. A signal crossing the board diagonally would be directed through a via to an x plane, where a signal line would take it along one edge of the board, through another via to a y plane, with a signal line perpendicular to the first one, and finally through a third via to the surface of the board at the destination.

The layout of a multilayer board is governed by design rules that specify the dimensions of the conductors and where they can be placed. For example, the vias are generally arranged in a grid; not every intersection of the grid lines necessarily has a via, but vias can be installed only at the grid points. The signal lines must run between the rows of vias. Inserting multiple conductors between adjacent vias can increase the overall wiring density, but it also complicates the design of the board since a conductor may be blocked from a particular via by another conductor.

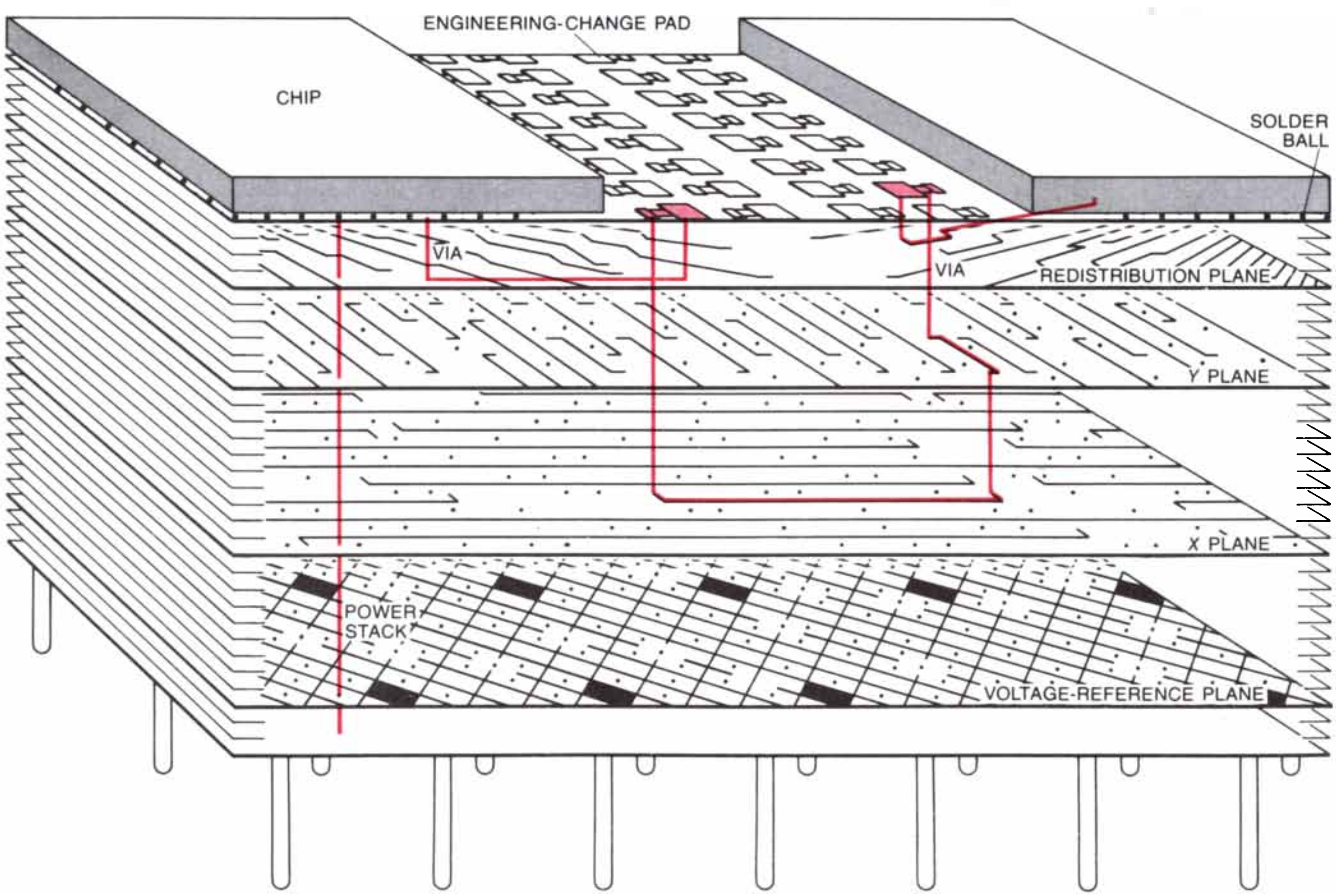
The design rules establish the maximum wiring capacity of a packaging technology. The maximum length of wiring per unit area of board surface is equal to the number of signal layers multiplied by the number of signal lines between adjacent vias and divided by the grid spacing of the vias. For example, a board with two signal planes, two wires between vias and vias on .1-inch centers has a wiring capacity of 40 inches per square inch. Both the wiring capacity and the efficiency with which it can be put to use are strongly influenced by the via size and the availability of vias for communication between signal planes.

A technology that packs many chips into a small volume is futile if the heat generated by the chips cannot be safely carried off. The property that determines the cooling capacity of a chip carrier or module is its thermal resistance, which is closely analogous to electrical resistance. According to Ohm's law, the resistance between two points in an electrical circuit is equal to the voltage difference between the points divided by the current flowing between them. Similarly, the thermal resistance of a package is the temperature difference between the heat source (the chip) and the ultimate heat sink (the ambient air) divided by the heat flux passing through the package. When the system is in a steady state, the heat flux must be equal to the power being dissipated by the chip. The thermal resistance can be expressed in units of degrees Celsius per watt.

The thermal path from the chip to the air can be considered in two parts. First, heat is conveyed, usually by conduction, from the semiconductor junctions—the areas within the structure of a chip where most of the heat is generated—to the surface of the package. The thermal resistance of this part of the path is called the internal resistance and depends mainly on the geometry of the chip carrier and the thermal conductivity of the materials. Second, heat is removed from the package itself, in most cases by moving air; the mechanism of heat transfer is either forced or natural convection. The external thermal resistance is a complicated function of many factors, including the area and thermal emissivity of the module and the velocity and turbulence of the airstream.

The overall thermal resistance of a typical plastic dual in-line package in moving air is about 50 degrees C. per watt. Thus a chip dissipating .5 watt would undergo a 25-degree rise in temperature. The internal resistance can be reduced by replacing the plastic package with a ceramic one, which has a higher thermal conductivity, or by enlarging the area over which the chip is bonded to the carrier. Such measures can reduce the thermal resistance by half and thereby double the power capacity. The external resistance can also be reduced, for example by adding a heat sink to increase the effective area of the package or by increasing the velocity of the air, but it is very difficult to reduce the overall resistance by more than another factor of two.

For a package with a given thermal resistance the maximum power dissipation depends on the temperature difference available for cooling. Although some silicon devices can operate at a junction temperature exceeding 150 degrees C., the maximum operating temperature of digital circuits is usually



MULTILAYER CERAMIC SUBSTRATE can accommodate 320 centimeters of wiring per square centimeter of surface area. Here a small region of the substrate is shown in a schematic and greatly magnified cross section. A typical signal path proceeds from a chip down a "via" that passes between the layers of the structure. A conductor on one of five redistribution layers carries the signal to another via that returns to the surface at an engineering-change pad. The signal then passes through a metallic link along the surface and into a third via, which takes it to a signal plane deep within the substrate. A conductor oriented along the *x* axis leads to still another via and a *y*-oriented conductor; a final via returns to the surface at another chip site to complete the signal path. Some signal paths go to pins on the bottom of the substrate, where connections are made to other multichip modules. Power is distributed by three planes at the bottom of the substrate. The pattern of connections can be modified by cutting the link at an engineering-change pad and bonding a new wire to the pad.

limited to between 75 and 85 degrees to ensure reliability and uniformity of electrical properties from chip to chip. The inlet temperature of the cooling airstream can be as high as 30 degrees, and the air heats up (by as much as 10 to 15 degrees) as it passes through the package. Hence the maximum available temperature difference is less than 50 degrees. These calculations imply that the power-handling capacity of a ceramic dual in-line package is less than two watts. In general the maximum heat flux with air cooling is about two watts per square centimeter at the chip level and about .5 watt per square centimeter at the module level. (The values can be increased by using special heat sinks and high-velocity chilled air.) These limits represent important constraints on circuit power and circuit density and therefore on performance.

Some designers of high-speed computing systems have explored a number of alternatives to packaging technology based on the air-cooled chip carrier. Here I shall describe one high-performance technology, which I believe has extended the state of the art. It was developed for a new series of mainframe computers by scientists and engineers at the IBM facilities in East Fishkill, Endicott and Poughkeepsie, N.Y., and in Sindelfingen, West Germany. Two fundamental design objectives were to reduce the number of interconnections between packaging levels and to reduce the total wiring length. These goals offered three potential benefits: higher speed, lower cost and improved reliability. The package that resulted from these efforts has two main components: a multilayer ceramic substrate on which the chips are mounted and through which all interconnections are made, and a module assembly that provides a direct thermal path from the back of the chips to a water-cooled heat sink.

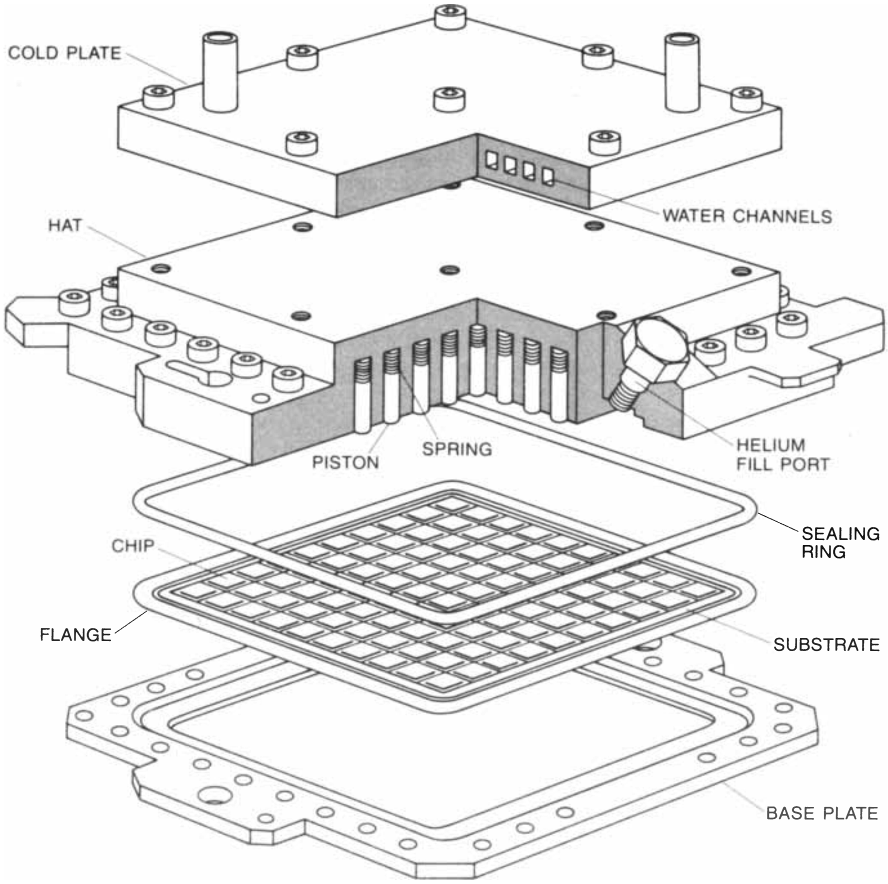
The ceramic substrate compresses a wiring network of extraordinary complexity into a square, tilelike object 90 millimeters on a side and roughly five millimeters thick. On the top surface are sites for between 100 and 133 high-speed chips, with a total of more than 12,000 chip contact pads. On the bottom surface of the substrate are 1,800 pins that supply power to the chips and route signals to and from the next level in the packaging hierarchy. Both the chip terminals and the module terminals are organized as two-dimensional arrays to minimize the space required. The substrate itself has 33 layers of conductors, which are interconnected by more than 350,000 vias.

Sixteen of the layers in the substrate are *x* or *y* planes of signal wiring. The design rules for these layers allow vias to be placed on grid points with a spacing of .5 millimeter. Just one signal line is allowed between adjacent vias, eliminating the possibility of conflict for access to a via. Thus the 16 signal layers have a maximum wiring capacity of 320 centimeters per square centimeter of substrate. A typical substrate has 130 meters of wires on these planes. A voltage-reference plane is inserted between each pair of *x* and *y* layers to control the impedance of the signal lines. The characteristic impedance is 55 ohms.

The top five layers of the substrate have the most vias. Here the vias are on a .25-millimeter grid to match the area array of 120 contact pads on each chip. These layers are employed to redistribute the signal lines (96 per chip site) from the chip pads to a set of contact pads that surround each site. From there the signal lines return to the interior of the substrate. The surface pads allow the module to be tested with the chips in place. Furthermore, if changes in wiring are needed, a small link on the surface can be severed to isolate any signal line from the internal wiring of the substrate. A new connection can then be made by bonding a fine wire to the metal pad and laying it in the channels between the chips. The ability to make such engineering changes is particularly important during the development of a new product. There are also methods for replacing an individual chip because of either a design change or the failure of a component.

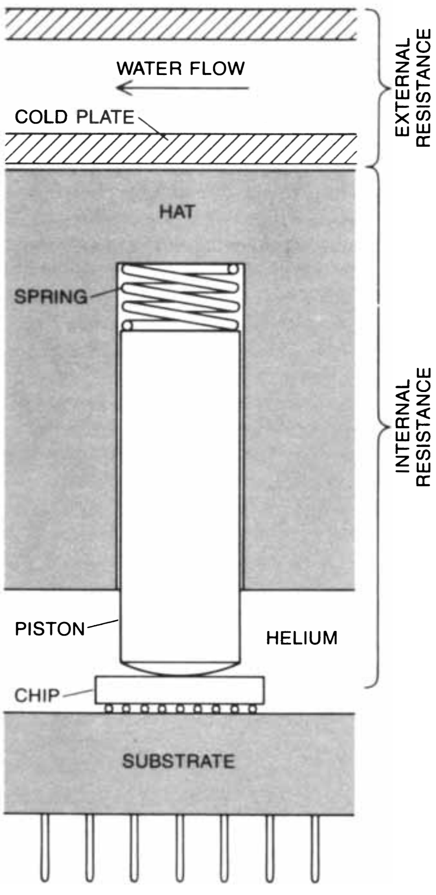
Power is distributed by three planes at the bottom of the substrate; two of the planes carry the voltages required by the chips and the third is at ground potential. The planes themselves are supplied in parallel by an array of pins distributed uniformly across the bottom surface; 500 of the 1,800 available pins are used for this purpose. From the power planes the current flows directly to the chip power pads through parallel vertical stacks of vias. The substrate is designed to supply up to four watts to each chip site, although not all the chips draw the maximum current. The module as a whole is limited to a total of 300 watts, or an average of about three watts per chip. The power-supply voltage loss due to resistance in the package is less than 15 millivolts.

As might be expected, some compromises had to be made in matching the



THERMAL-CONDUCTION MODULE houses the multilayer substrate and cools the 100 or more chips mounted on its surface. The substrate is clamped between a baseplate and a "hat" equipped with spring-loaded metal pistons. Each piston presses against the back of a chip, conducting heat to a cold plate bolted on top of the hat. The cold plate in turn gives up the heat to chilled water that is pumped through channels in the plate. The module's ability to dissipate heat is enhanced by filling the internal volume with helium, which has a higher heat conductivity than air. The rated power-handling capacity of the thermal-conduction module is 300 watts.

design to the materials for such a complex packaging technology. The principal component of the substrate is alumina, or aluminum oxide, a ceramic that was chosen because of its superior mechanical properties and its ability to withstand the various chemical and thermal processes employed in manufacturing the substrate and module. The principle drawback of alumina is its relatively high dielectric constant of 9.4. As a result the propagation speed of signals in the module is lower than it would be in a fiberglass printed-circuit board; since chips can be mounted directly on the ceramic substrate, however, the average distance between chips is much smaller than it would be with card-on-board technology, and so the total propagation delay is significantly less. Another compromise was made in choosing the material of the conducting paths inside the substrate. Because the ceramic must be fired at 1,500 degrees C., the conductors must be made out of a refractory metal; other metals have higher electrical conductivity, but their melting point is too low. (Copper, for example, melts at 1,083 degrees C.) The metal chosen is molybdenum, whose bulk resistivity is approximately three times that of copper. This constraint is compensated for by the high density of vias in the design; large currents are supplied directly to the chips by stacks of vias.



THERMAL RESISTANCE of the module determines the maximum temperature rise of a chip with a given power dissipation. The resistance of the heat path from the chip through the piston (and the helium atmosphere) to the hat is defined as internal; the resistance from the cold plate to the flowing water is external. The total resistance of the package is some 11 degrees Celsius per watt per chip site. Therefore a chip that dissipates four watts undergoes a rise in temperature of 44 degrees above the water temperature.

The fabrication of the multilayer ceramic substrate begins with the casting of the individual layers. Ceramic and glass powders are mixed with an organic binder and solvent to form a slurry, which has the consistency of paint. The slurry is deposited on a moving plastic belt and passes under a blade that sets the thickness of the layer. A long drying oven drives off the solvent, leaving a cohesive but still flexible material that looks rather like thick paper. Square blanks called green sheets are then cut from the web. ("Green" is the ceramist's term for unfired material; the sheets are actually white.) Alignment holes are punched in the corners of each sheet to aid in subsequent operations.

The next step is the punching of via holes. It is done under computer control by a high-speed multiple-punch-and-die machine. In the top layer of a 100-chip module, for example, every chip site has an identical array of holes. One hundred punches are therefore mounted in a grid with a spacing equal to the spacing of the chip sites. Each operation of the tool punches a single hole at the same relative position in each chip site; the entire sheet is then moved slightly and the next set of 100 holes is punched.

The metal patterns are laid down on the green sheets by a process similar to stenciling or silk-screen printing. A paste of molybdenum in a binder and solvent is extruded through a metal mask that has the pattern of the wiring cut into it. (The pattern itself is generated automatically by data from the computer-aided-design system.) The paste is applied under pressure so that the punched via holes are also filled. The metallized sheets are then dried and inspected. The detection of flaws in individual sheets before they are combined with other sheets to form a complete substrate is important to maintaining the overall yield of the manufacturing process.

The sheets that pass the inspection are stacked in the appropriate sequence and laminated under high pressure at 75 degrees C. Because vias only 120 micrometers in diameter must be continuous from one layer to the next, control of dimensions and alignment is critical. The green laminate is trimmed to size and subjected to a long firing cycle in which the peak temperature of more than 1,500 degrees is reached in a hydrogen atmosphere. At lower temperatures the organic material decomposes and becomes volatile, and at the higher temperatures the ceramic and the metal sinter into a monolithic structure.

The rate of heating must be carefully controlled; if the temperature were to rise too quickly, the organic binder would volatilize faster than it could diffuse to the surface, causing the substrate to delaminate. During the sintering process the substrate shrinks by approximately 17 percent in each dimension, for a total volume reduction of about 40 percent. Given the tight dimensional tolerances of the final assembly, it is clear that the amount of shrinkage must be known precisely when the patterns are first inscribed on the green sheets. After firing, the substrate has the size, shape and characteristic hardness of a ceramic tile; if it is struck, it rings.

The exposed metal areas on both surfaces of the completed substrate are plated with nickel and then with gold. An extensive electrical test is done by an automated tester that again utilizes data from the computer-aided-design system to verify the correct pattern of interconnections. The machine must confirm that each pad is connected to other pads as specified by the design; in addition the machine must make certain there are no extra, improper connections. When the testing is completed, the 1,800 pins are brazed to the bottom surface; a metal flange is attached, also by brazing, in the same operation.

The integrated circuits are attached to the substrate by a method developed at IBM for an earlier generation of computers. First a tin-lead solder is evaporated through a metal mask onto the contact pads on the surface of the chip. The chip is then heated in an inert atmosphere to melt the solder, which, under the influence of surface tension, forms a spherical droplet on each terminal. The solder is allowed to harden again, and the chip is inverted over the substrate with the contact pads aligned. When all the chips are in place, the assembly is once more heated to the melting point of the solder; each solder pad assumes the shape of a truncated sphere, connecting the chip and the substrate terminals electrically but still holding the chip above the surface. The mounted chips can be tested through the surrounding engineering-change pads. Any surface wires needed are then attached by ultrasonic bonding.

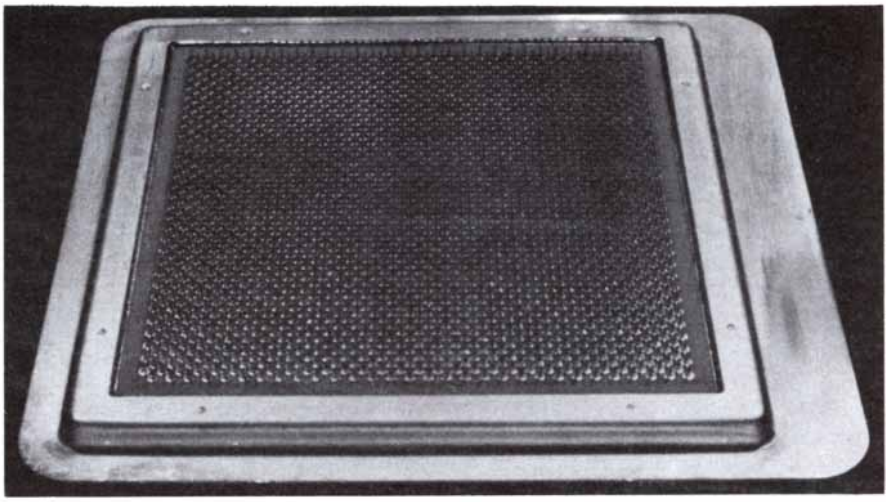
The multilayer ceramic substrate, with its flange and pins, forms the base of the thermal-conduction module, which has two other major components, called the hat assembly and the cold plate. The hat is clamped to the substrate flange, compressing a pliable ring with a *C*-shaped cross section and thereby sealing the internal volume. The cold plate is bolted on top of the hat. Within the hat is an array of spring-loaded aluminum pistons with the same pattern as that of the chips on the substrate. Each piston presses against the back of a chip, conducting the heat upward to the cold plate. The cold plate in turn is cooled by water chilled to 24 degrees C. flowing through internal channels at a rate of 40 cubic centimeters per second. The thermal properties of the module are enhanced further by filling the sealed volume with helium, which at room temperature has a much higher thermal conductivity than air. The helium reduces the internal thermal resistance by more than half.

The assembled module has an internal thermal resistance (measuring from the chip up to the cold plate) of nine degrees per watt per chip site and an external resistance of two degrees per watt per chip site. The module is conservatively specified for use at a maximum of four watts per chip and a total of 300 watts. From the thermal resistance it can be calculated that a chip dissipating four watts should reach a temperature of 68 degrees, which is well below the maximum operating temperature of the circuits. The heat flux is about 20 watts per square centimeter at the chip level and four watts per square centimeter at the module level, an order of magnitude greater than the heat flux in a typical air-cooled package.

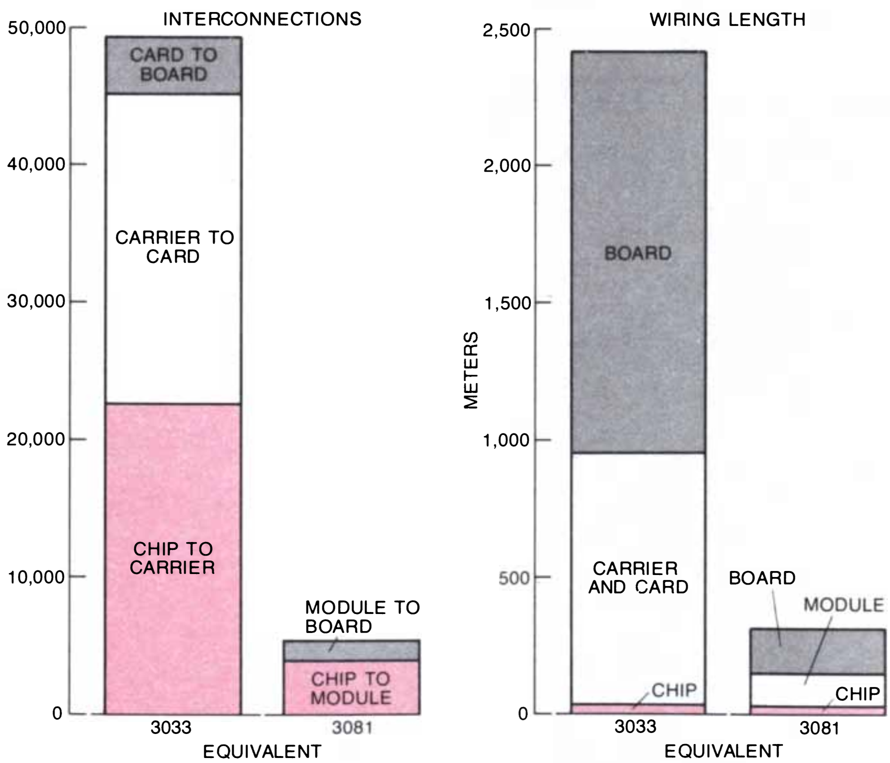
The thermal-conduction module represents only one level out of three levels in an advanced-technology central processing unit. The first level is that of the chips themselves, which were developed at the IBM laboratory in East Fishkill. The logic chips all have the same underlying structure of 704 basic logic cells; over this structure, however, there are three levels of metallic conductors that are customized to create various configurations of gates and other devices. The metal layers also supply power to the circuits and distribute signals between the cells and the contact pads. The bipolar logic circuits on the chips have a switching delay of 1.1 nanoseconds.

The third component of the system, which represents another major advance in packaging technology, is the large printed-circuit board that interconnects up to nine thermal-conduction modules. The board, developed at IBM's Endicott facility, has 20 layers, including six impedance-controlled signal planes. A nine-module board measures 60 by 70 centimeters and provides connections between the 16,200 pins of the modules and more than 2,000 additional terminals for cables leading to other subassemblies. The board also supplies up to 600 amperes of current to the modules.

The thermal-conduction module and the associated technologies were introduced in 1981 in the IBM 3081 high-performance data processor, developed at the Poughkeepsie laboratory. In a typical module in the 3081, 52 of the chip sites are occupied by logic chips. There are also 34 array chips, a form of high-speed semiconductor memory employed for data and instructions that must be immediately available to the central processing unit, and five terminator chips, which are arrays of resistors used to match the impedance of long, off-chip transmission lines. A module with this complement of chips holds more than 25,000 logic circuits,



BED OF NAILS on the underside of the thermal-conduction module is made up of 1,800 connecting pins. Five hundred of them supply power to the module; the rest are available for communication with other modules and with other components of the computing system.



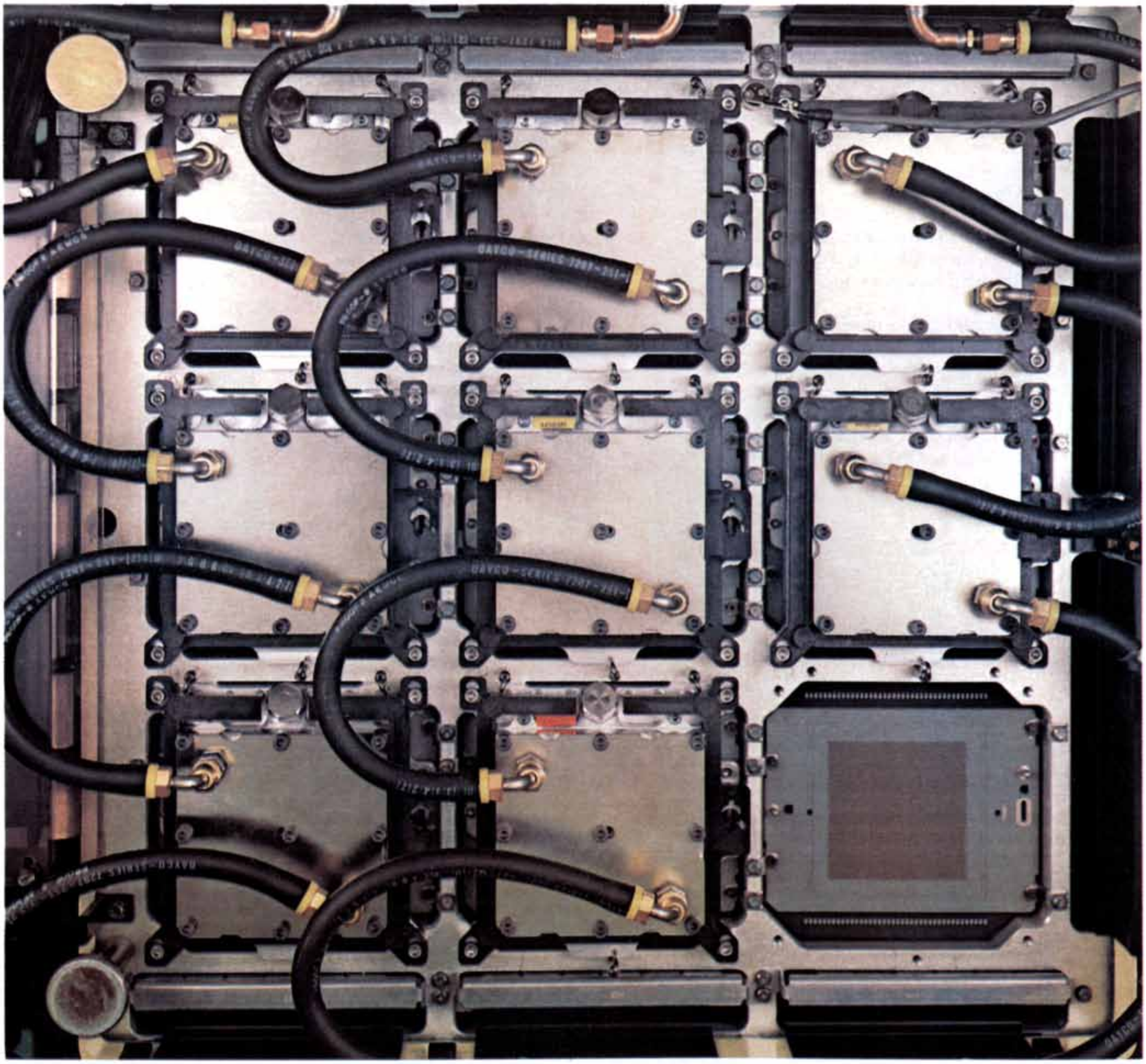
EFFECT OF PACKAGING on performance and reliability is suggested by a comparison of the packaging technologies employed in two mainframe computers made by the International Business Machines Corporation. The IBM 3033 is built with individual chip carriers, cards and boards; the 3081 employs the thermal-conduction module. Here the comparison is made between a single thermal-conduction module and a set of 3033 components with the same number of logic circuits. The number of connections between packaging levels in the thermal-conduction module is smaller by a factor of 10; furthermore, most of the remaining connections are solder joints between the chip and the substrate, which are more reliable than mechanical connections. The signal wiring length in the thermal-conduction module is one-eighth that in the 3033 equivalent, leading to a commensurate reduction in signal-transmission time.

65,000 array bits and almost 500 terminating resistors.

It is instructive to compare the packaging technology of the 3081 with that of another high-performance IBM computer, the 3033, which is built with a more conventional packaging technology. In the 3033 each chip has an average of 12 logic circuits and is housed in an individual chip carrier. The carriers in turn are mounted with terminating resistors and other components on printed circuit cards, which connect along one edge to larger circuit boards. If the functions of the typical thermal-conduction module described above were implemented in the 3033 technology, 1,880 chip carriers would be needed for the logic circuits, another 80 chip carriers for the memory arrays and additional components for the line terminators. To mount and interconnect all these components would take 52 printed-circuit cards, four larger boards and the cabling required for connections between the boards. Note that these components would be needed to reproduce the functions of a single module; a full-featured 3081 system has 26 modules.

Three effects of the packaging technology based on the thermal-conduction module are particularly noteworthy. First, the large reduction in packaging hardware results in a significantly lower cost. Second, the system has improved reliability. The main sites of failure in electronic assemblies are the connections between package levels. The thermal-conduction module has eliminated one level of packaging entirely and has reduced the number of logic signal connections between levels by a factor of almost 10. Furthermore, most of the remaining connections are the chip-to-module solder joints, which are inherently more reliable than mechanical connections at higher packaging levels.

The third effect is improved performance. The total length of logic wiring in the 3081 processor is roughly one-eighth what it would be in equivalent 3033 technology. The result is a fourfold reduction in the total packaging delay of the central processing unit, including the modules, the boards and the cables; this in turn allows a twofold decrease in processor cycle time.



PRINTED-CIRCUIT BOARD mounted on a massive steel frame represents the next level (after the thermal-conduction module) in the packaging hierarchy of the 3081 computer. The board accepts nine modules. Here eight modules have been installed and connected to the water-cooling system; the ninth position is left vacant, exposing the connectors that receive the 1,800 pins on each module.

The Development of Palm Leaves

The compound leaves of plants usually arise either from differential growth or from selective cell death. Palm leaves, however, follow a developmental pathway that combines both of these processes

by Donald R. Kaplan

One of the central questions of modern biology is: How does a mature plant or animal develop from the fertilized egg and then from the embryo? Pioneer biologists described the stages of development in many organisms and tissues in great detail, but the investigation of the underlying processes that control development had to await the rise of molecular biology. Moreover, even describing the subtle stages of development in many tissues had to await the rise of methods such as electron microscopy. Clearly the underlying processes in the stages of development of a tissue would be difficult to identify without identifying the stages themselves.

A prime example of where a complex developmental pattern has had to be clarified before its control could be investigated is the leaf of one family of flowering plants: the palms (Palmae). All flowering plants have three main structures: roots, stems and leaves. Of these structures leaves are the most diverse. For example, a leaf, in addition to its principal role as an organ of photosynthesis, may be modified into a protective seal over a new bud, into a climbing organ, into the reproductive organs of a flower or even into an insect trap. An unmodified leaf is characterized by a broad, flat green expanse, the part of the leaf that has the greatest concentration of chlorophyll, intercepts sunlight and acts as an organ of gas exchange. This, however, is only one of a leaf's three component parts. It is known formally as the lamina (from the Latin for "plate") and less formally as the blade.

What supports the blade is the second component of the leaf: the petiole (another borrowing from the Latin, this time by the great Linnaeus, that means "little foot"). The petiole serves largely as a conduit, carrying the nutrient products of photosynthesis from the leaf blade to the third leaf component. That component is the leaf base, which joins the other parts of the leaf to the plant stem. As well as providing mechanical support for those parts and conveying nutrients from the petiole to the stem (and thus to the plant as a whole) the base surrounds and protects the rest of the leaf in its younger and more embryonic form, when a leaf-to-be exists only as part of a terminal bud on a shoot.

Exactly how does a leaf develop from the bud? Although I shall be discussing the process primarily in terms of representatives of the palms, it will be useful to include representatives of a not very distantly related family: the arums (Araceae). The palms and the arums grow leaves from the embryonic state to maturity by means of distinctly different pathways that were not fully understood until recently.

Plant development is quite unlike animal development. With most animals the formation of new organs is confined to the earliest phases of embryonic growth. With plants the process is a continuous one. New organs arise from perpetually embryonic growth centers: undifferentiated tissues consisting of cells capable of transformation into a variety of plant organs. The growth centers, at the extremities of the plant, are embryonic tissue at the apex of the root and shoot. Such tissue is termed the apical meristem.

In a typical shoot system the meristem initiates new increments of stem growth, together with leaves along the stem, in a precise geometric pattern. Since new stem increments and leaves arise over a prolonged period, it might seem that a plant shoot would be a set of structurally identical units, like the segments of an earthworm. For example, those who study leaf development have taken for granted that if the leaves below the terminal bud of a shoot were arranged in order of increasing age, the sequence would illustrate the stages a leaf at a given position on the stem would go through if its development were followed over some length of time.

Actually such an assumption is valid only if the plant exhibits growth of the "steady state" kind, where successive leaf and stem units can be shown to be identical. Numerous examples are known, however, where leaf and stem structures change markedly as the shoot grows. To compile a record of leaf growth under such circumstances it is necessary first to specify a certain leaf position along the shoot and then to concentrate on the development of that particular leaf.

When a leaf originates, in the form of an outgrowth from the periphery of the shoot apex, it typically takes the form of a flattened pad that shows no signs of differentiation. As its growth continues, the first two parts to be defined, in outward order, are the leaf base and the blade. The petiole, if one develops at all, arises later as an insert between the base and the blade.

The leaves of flowering plants of course have a large variety of shapes and sizes. One of the commoner variants in shape is the dissected, or compound, leaf. The blades of these leaves are cut into segments or leaflets, known to botanists as pinnae (from the Latin for "feathers"). In terms of comparative development dissected leaves are of particular interest because they so clearly illustrate how different paths of development can lead to leaves that are closely similar in appearance. The giant fronds of palm trees are the largest and most complex of all dissected leaves. At the same time the genera of the tropical arums, a family famous for its diversity of leaf forms, provide a classical example of two sharply contrasting modes of leaf dissection. Let us take up the arums first.

The mature leaves of the arum species *Zamioculcas zamiifolia* have a pinnately compound blade: four to five pairs of blade segments appear as appendages on each side of an elongated stalk, or rachis (a Latinized borrowing from the Greek for "rib"). When the terminal bud of a *Z. zamiifolia* shoot is dissected by removing successively younger leaves, a tiny dome-shaped structure, 100 micrometers in diameter, is at last revealed. This is the apical meristem of the shoot; from it arise the youngest primordia, or embryonic leaves, each successively forming a small hood that arches over the shoot apex. As each new leaf primordium grows upward it completely enshrouds the apical dome and the two free edges of its hood are pressed closely together.

The growth of different tissues at different rates soon gives the leaf's future leaflets the appearance of bumplike protuberances along the leaf's two free edges. The older and slightly larger bumps appear toward the tip of the leaf; the younger and smaller ones arise progressively nearer the leaf base. Once a complement of four or five leaflet pairs has appeared, enlargement continues and the bumps soon assume the shape of mature leaflets.

This mode of leaflet initiation from the free margins of a leaf is the commonest mechanism of development among higher plants with dissected leaves, whether they are simple vascular plants such as ferns, gymnosperms such as cycads or the higher flowering plants. Through progressive degrees of elaboration it is possible for compound leaves to be twice, three times or even many times dissected. Indeed, in principle there is no limit to the degree of elaboration that can be found in a dissected leaf.

The arum that shows a sharply different mode of leaf dissection is popularly known as the split-leaf philodendron, one of the climbing genus *Monstera.* These plants are noted for the distinctive holes in the blades of their leaves and are widely grown as ornamental plants for that reason. The size and shape of the holes vary in different parts of the blade because different parts of the blade have different rates of growth. For example, the first holes to appear lie close to the edge of the blade; at maturity they are large and elliptical in outline

COCONUT PALM, probably the most numerous of all the cultivated members of the palm family, displays an abundance of compound leaves and clusters of commercially valuable nuts. The individual frond leaflets take form in a complex process of development.

because that part of the blade undergoes the greatest amount of lateral expansion. The holes that develop later are found closer to the leaf midrib. They tend to be smaller and to retain their original rounder outline because the tissue surrounding them does not undergo as much lateral growth.

It has been known for more than a century that the holes in philodendron leaves originate as localized areas of cell death. The scanning electron microscope now makes it possible to observe the process in considerable detail. The earliest sign of the impending event is the appearance of a slightly indented circular area on the surface of the leaf blade. The indentation reflects the diminished turgor (fluid content) of the affected cells. When these cells finally die, the part of the leaf composed of them simply dries up and falls away, leaving behind a perforation.

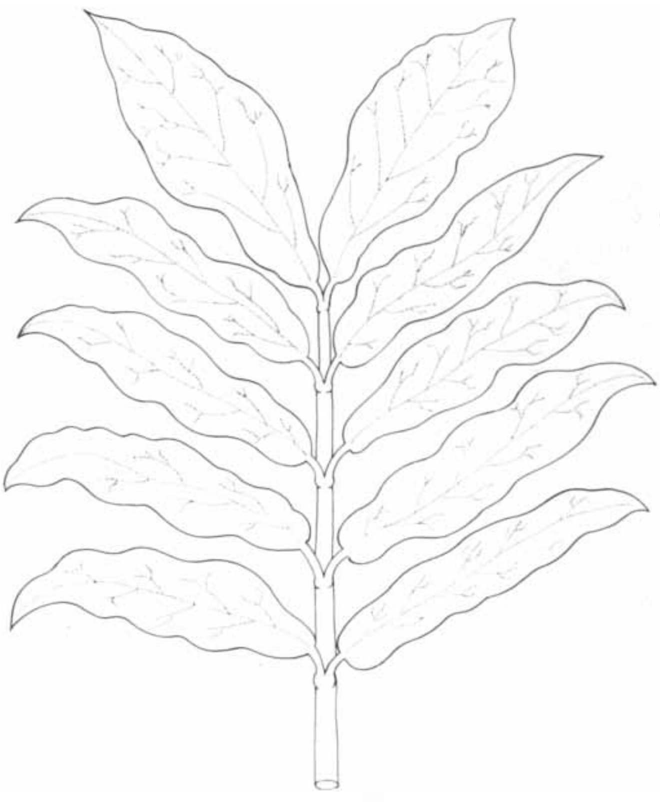
In many species of the genus *Monstera* the marginal strip of blade tissue adjacent to the most peripheral hole will not grow in width at a rate that matches the growth of the rest of the blade. As a result this thin bridge of marginal tissue usually breaks, converting what had been a perforated blade surface into a marginally lobed one. Indeed, in some species (for example *M. subpinnata, M. tenuis* and *M. dilacerata*) the leaf blade forms only a single set of broadly elliptical perforations near the blade edge. When these perforations lose their bridges, the result is a pinnately lobed leaf blade that bears a striking resemblance to the pinnate leaves of *Zamioculcas zamiifolia*. If one did not know that the dissected *Monstera* leaf had been formed by the process of cell death and the dissected *Zamioculcas* leaf by the process of lobing, the fact that each was the result of a markedly different developmental program would be impossible to guess.

Like the arums, the palms make up a large family of tropical plants that is worldwide in distribution. Not only are their fronds, or leaves, large and complex but also one species of the palm genus *Raffia* grows the largest leaves in the entire plant kingdom. Its leaflets can be 60 feet long. The palms produce a dissected leaf-blade surface in a way that is distinctly different from either of the arums' modes of leaf production. In some respects, however, the process represents a combination of the one arum's lobing and the other's cell death. Because of the greater complexity of the palm's mode of leaf production and many unresolved aspects of the process, my colleagues and I at the University of California at Berkeley have recently reinvestigated palm-leaf development in order to define the fundamentals of the process.

All palm leaves have three standard components: an elongated blade and petiole, and a leaf base that is characteristically tubular and completely encircles the stem. Among palms the leaf base plays a particularly significant role. Not only does it support the enormous and heavy photosynthetic organs of the blade and petiole; it is also a source of support for the shoot in the younger regions of the palm's stem (or trunk), where the internodes (or individual stem units) are still elongating. That is why all the bases of all palm leaves incorporate an elaborate network of fibrous tubular bundles; these provide the leaf base both with mechanical reinforcement and with flexibility.

Palm-leaf blades typically show one or the other of two major configurations that reflect differences in the distribution of growth during the course of dissection. One of the two is characterized as pinnate because of its feathery appearance. The other is characterized as palmate because it is fanlike, rather like a hand with its fingers spread. Pinnate fronds tend to have short petioles; palmate fronds have long ones.

What makes the developmental mode of palm leaves distinctive is that at first their blade surfaces are thrown into a series of pleats known as plications. A process of tissue separation along cer-



DISSECTED LEAVES of two tropical arums look much the same at maturity although they followed entirely different developmental pathways. At the left is the leaf of *Zamioculcas zamiifolia;* its five paired leaflets began as bumplike protuberances along the edge of the embryonic leaf blade (*see micrographs on opposite page*). At the right is the leaf of *Monstera subpinnata,* popularly known as the split-leaf philodendron. Its 12 leaflets originally formed a single leaf blade before the development of areas of localized cell death dissected the blade into separate leaflets (*see micrographs on page 102*). Evidence of the separate developmental pathways is not apparent at maturity.

tain of these pleats then cleaves the pleated surface into a series of leaflets. It is easy to see, in looking at a fully developed leaf of the palmate kind, that the segmented nature of the leaf originates from the pleating of the leaf blade. Indeed, as with a Japanese fan, the leaf can be closed by pressing the pleats together. Since each leaflet is *V*-shaped in cross section, it appears to have been cut out of a pleated surface.

The development of a palm leaf of the pinnate kind in general involves the same process. It is, however, subtler and more complicated, and it can best be understood by tracing the development of the leaf of the "feather" palm from inception to maturity. The lateral leaflets of the feather palm are dispersed along the leaf's rib axis; they are formed while the rib axis is elongating.

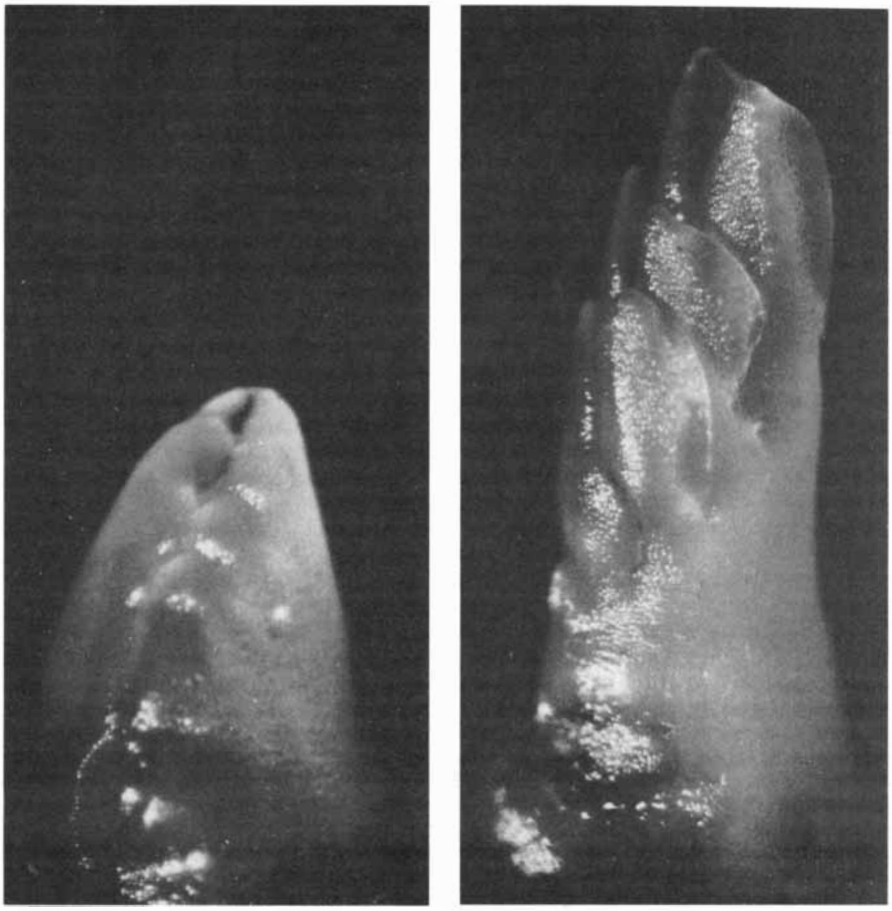
The primordium of a young feather-palm leaf arises as an outgrowth from the shoot apex. It resembles a helmet fitted over the apical dome. The part of the primordium that projects above and away from the shoot apex is the future leaf blade. The future leaf base in turn appears as a collarlike growth initiated around the periphery of the shoot apex; the shape of this early base structure foreshadows the subsequent complete encirclement of the stem by the leaf base.

The young primordium soon looks more like a hood than a helmet. Its more readily observable outer surface corresponds to the lower surface of the future leaf blade; its narrower inner surface will become the future blade's upper surface. The margins, or free edges, of the primordium represent the boundary between the two surfaces of the future blade.

Where in the lobe-growing arum *Zamioculcas* the initiation of leaflets involves lobing along this leaf-blade edge, the feather palm's future leaflets first appear as a series of ripples or ridges within the surface of the blade some distance from the edge of the blade. The folds are not only particularly evident on the readily observable lower blade surface-to-be but also have their counterparts on the upper surface. A section made at a right angle to one of the blade margins cuts across these ridges and reveals that the lower and upper ridges actually make up a single set of pleats running the entire length of the blade in a compressed zigzag.

What makes these pleats unique in leaf development is their spatial restriction. For example, they never extend to the blade margins, with the result that the growing leaf carries a narrow unfolded strip of tissue around the entire periphery of the blade. Similarly, on the back side of the leaf the pleats extend only a short distance toward the axis of the thickened leaf rib.

As the leaf continues to grow, new pleats appear. Most of them are added toward the base of the leaf blade but a few may appear toward the tip of the blade as well. The total number of pleats typically corresponds to the number of leaflets found in the mature leaf, and so the development of additional pleats is terminated when that number has been attained. At this stage the pleats remain tightly pressed together, like the folds of a camera bellows.



EMBRYONIC LEAF of *Z. zamiifolia* is seen at an early stage in the photomicrograph at the left and at a slightly later stage in the photomicrograph at the right. The specimens are 1.7 and three millimeters long. Bumps on the free margins of the younger leaf are more recognizable as emerging leaflets on the older one; they are starting to approximate their mature shape.

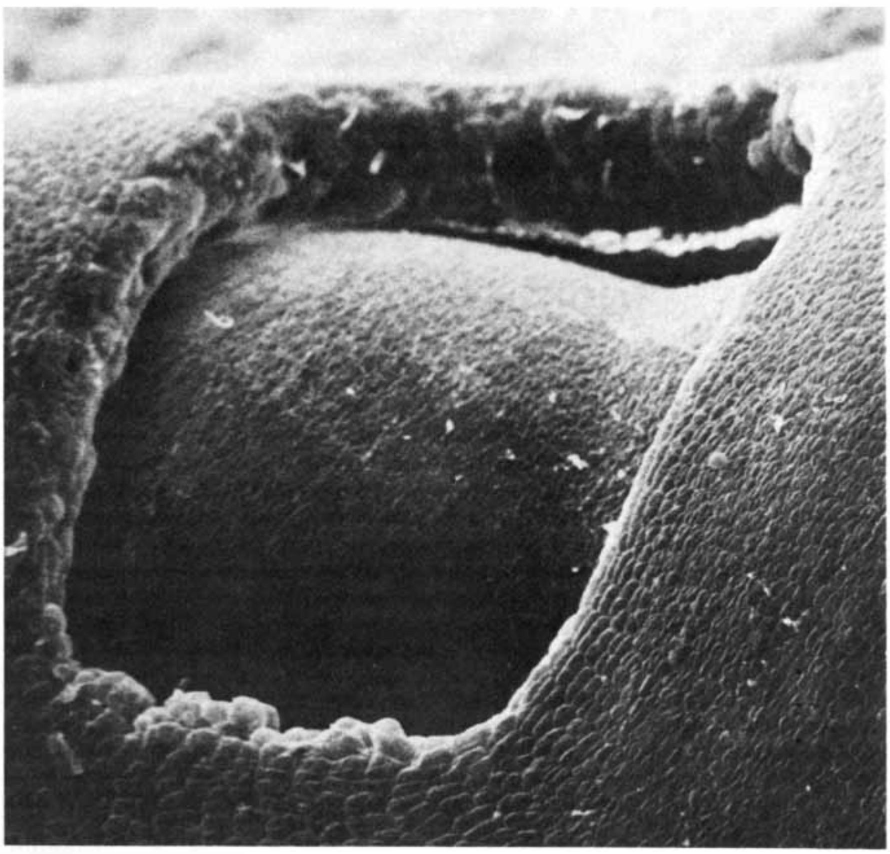
The next and most intricate steps in the development of the feather-palm leaf take place after the pleats have undergone further growth in depth. The steps are two: separation of the pleats into individual leaflets and disposal of the peripheral leaf-blade tissue that still binds the leaflets together at their tips. In the first step the adjacent pleats separate from each other. Although the cellular features of the process are not yet known in detail, it is assumed that separation is effected by a breakdown of the intercellular cementing substance along the line of the separation. It is certain that selective cell death is not involved: there is no evidence of cell damage along the separation.

Before proceeding to the second step in leaflet formation some general features of palm-leaf development in the first step should be mentioned. One of them is that in the majority of pinnate-leaved palm species the separation process takes place along the ridges of the outer, or lower, leaf-surface-to-be. The process gives rise to individual *V*-shaped leaflets. Observation of the process by scanning electron microscopy shows that an indentation develops along the top of the ridge. As tissue separation takes place the ridge indentation is transformed into a distinct slit.

Returning to the second step, even though the leaflets have been separated from one another over most of their length, their tips remain attached to the strip of leaf-edge tissue. The condition does not prevail for long, although even after the breakup of the tissue begins, strips of the nearly separated margin may still remain attached to the tips of individual leaflets. Because of their resemblance to parts of a horse's bridle, these dangling strips are called the "reins" of the leaf.

The mode of origin of palm-leaf pleating has been a subject of contro-

LEAF PERFORATIONS that will eventually dissect the blade of a *Monstera* leaf into individual leaflets can be seen in this scanning electron micrograph, which magnifies a developing leaf 35 times. In the younger part of the leaf, at the right, the perforation-to-be appears as an indentation on the surface of the leaf blade. In the older part, at the left, some of the dying tissue has broken away from the surrounding edge of still healthily turgid leaf-blade cells.



LATER STAGE OF LEAF PERFORATION is shown in a scanning electron micrograph of a *Monstera* leaf blade. The dead tissue has fallen away. As the leaf blade continues to grow, the tissue at the bottom, between the lower edge of the perforation and the blade margin, will tear apart, turning the two sides of the perforation into the lower edge of one leaflet and the upper edge of the adjacent one. This *Monstera* specimen, 17 millimeters long, is magnified 35 times.

versy in botanical circles for a century and a half. Each of two separate developmental hypotheses, representing radically different morphogenetic alternatives, has had until recently its own rank of supporters. The simplest and most direct of the two hypotheses proposes that pleating results from an initial folding produced by differential growth within the young leaf blade.

The second hypothesis proposes that pleating begins with an initial process of tissue separation, which is then followed by differential growth. According to this hypothesis, alternating clefts develop along both the upper and the lower surfaces of the blade because progressive cell separation begins on the blade surface and forms a furrow that extends to the interior of the leaf. As the cells separate from one another this interior penetration deepens until the resulting zigzag outline of the blade is difficult to distinguish from an outline produced by folding, particularly after the folds have become compressed.

What makes the second hypothesis so much more complex than the first is that it calls for some remarkable feats of cell differentiation. For example, consider the layer of cells on the surface of the leaf blade: the leaf epidermis. According to the hypothesis, the development of furrows would divide the leaf epidermis into isolated surface patches. If, as is observed, the sides of the furrows are also covered by a layer of epidermal cells, then some of the cells must have regenerated from cells deep in the interior of the leaf. These interior cells, however, would not normally have been destined to become epidermal cells. Is such a regeneration possible? If it is, it would represent a mechanism unique among higher plants.

The question was sufficiently intriguing to enlist Nancy and Ronald Dengler of the University of Toronto as additional allies in my investigation. Although what I shall hereafter refer to as the "tissue splitting" hypothesis lacked unequivocal evidence in its favor, there remained several reasons such a mechanism might in fact be operating in palms. The first was that some kind of tissue separation was clearly involved in the cleavage of palm-leaf pleats into individual leaflets. This being so, could not the same process be responsible for the origin of the pleats in the first place? William of Occam would surely have approved of a single developmental program rather than two different ones.

A second datum in favor of tissue splitting was an observation made as young palm leaves were being examined by scanning electron microscopy: the initial pleat furrows looked very much like narrow slits cut into the leaf surface. This fact, together with the observation that in some palm species the pleats appear to be block-shaped and are separated by narrow slits, further biased us in favor of the tissue-splitting hypothesis.

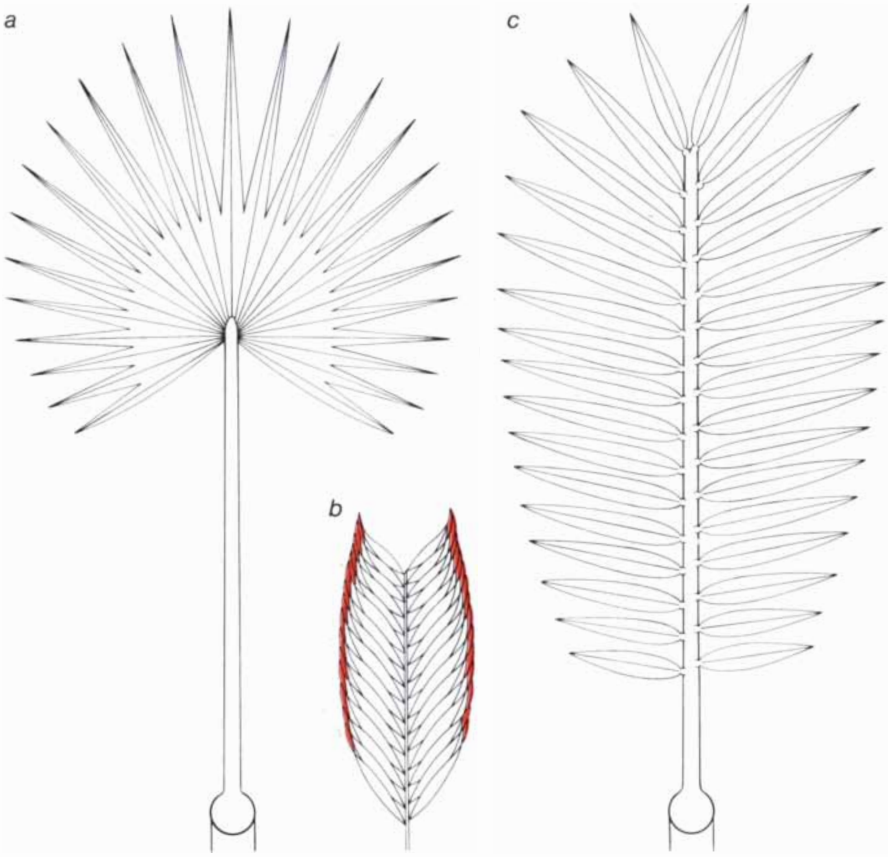
A survey of the scientific literature showed that one reason the two hypotheses had remained unresolved for so long was that earlier investigators, lacking more powerful instrumentation, had necessarily accepted as definitive data the superficial appearances of pleating shapes. They took these indications as proof of how the process occurred rather than relying on more fundamental measurements that were independent of surface appearances.

We set as our goal the study of the cellular and anatomical bases of pleating origins in the leaves of several different palm species, utilizing scanning electron microscopy in addition to light microscopy in order to pay particular attention to the three-dimensional aspects of leaf-blade development. We also studied thin sections of plant tissue (1.3 micrometers thick), prepared by the same techniques applied in transmission electron microscopy. (In transmission electron microscopy, unlike scanning electron microscopy, the electrons must pass through the specimen.) The use of these thin sections gave us light-microscope images of superior resolution, enabling us to eliminate many of the light-microscope artifacts, or misleading false elements, that had plagued earlier investigators.

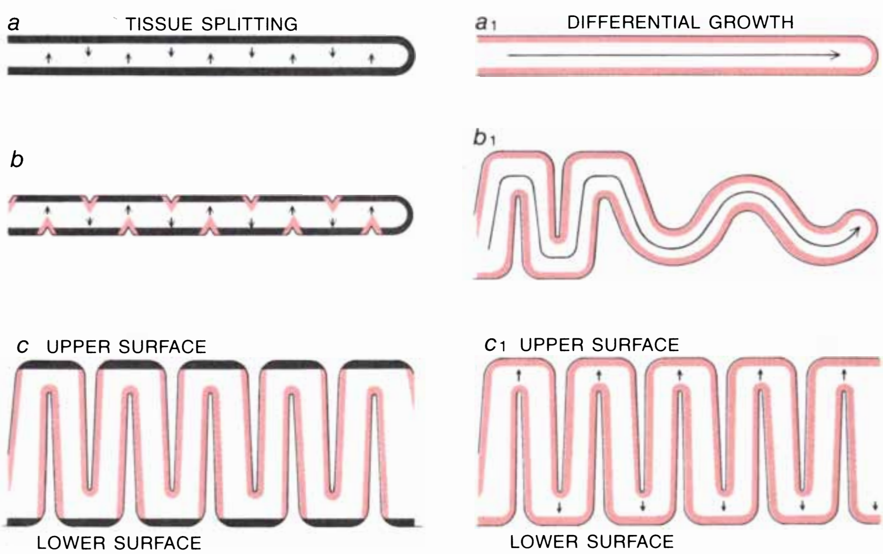
Now, according to both the tissue-splitting and the differential-growth hypotheses on the origin of leaf pleats, the furrows that separate the pleat ridges must deepen as growth continues. In the tissue-splitting model the furrows themselves would deepen. In the differential-growth model the increase in depth would be caused by growth of the ridges upward from the base of the furrow. Thus it should be possible to determine unequivocally which of the two models was operative by simply counting the number of cells that remained in contact in the lower and upper pleat ridges, either below or above the respective furrow lines.

For example, we predicted that if tissue splitting was the process in action, then the number of cells remaining in contact opposite the base of a furrow should decrease over time because cells were being pulled apart by the same forces of cell separation that were deepening the furrow. We also predicted that if the increased depth of the furrow were the result of ridge upgrowth, in accord with the differential-growth model, then the number of cells in contact along the furrow line should either increase (if the growth of the adjacent ridges involved both cell division and cell enlargement) or remain the same (if the growth involved cell enlargement alone).

Our counts of cells proved that the



MATURE PALM LEAVES generally show either a fanlike form (*a*), with a long petiole separating the leaf blade from the leaf base, or a featherlike form (*c*), with a short petiole and leaflets on each side of a long central rib. Before either array of leaflets can unfold (*b*) it must shed its "reins" (*color*): the thin strips of leaf-blade margin connecting the tips of the leaflets.



PALM-LEAF PLEATING had been thought to be due to one of two mutually exclusive patterns of development. The first pattern, shown in three stages at the left in this diagram, would initiate (*a*) as "tissue splitting" (cell separation) on the upper and lower surfaces of the leaf blade, a process that would divide the epidermal layer of the blade (*gray*) into unconnected units (*b*). As leaf-blade growth continued (*c*) the zones of mesophyll tissue connecting the upper and lower surfaces would expand substantially (*color*) and somehow regenerate as epidermal cells. The alternative pattern of development, at the right, visualizes pleating as the result of "differential growth" (a_1, b_1), whereby the pleats would first arise from crowding during growth and thereafter be extended by upward and downward processes of growth within the leaf blade, maintaining an unbroken layer of epidermal tissue (*color*) throughout the process (c_1). The author and his colleagues, studying scanning electron micrographs and thin sections of different palm leaves, concluded that the tissue-splitting hypothesis could not be supported because no cell separation was visible, whereas cell growth and enlargement could be seen.

deepening of the furrow lines was the result not of cell separation but of ridge upgrowth. Although the process was not identical for the lower and upper pleat ridges, no decrease in cell numbers (indicative of splitting) was evident in either. In the upper ridges cell numbers initially increase; in the lower ridges the numbers remain about the same. Further observations gave additional support to the differential-growth hypothesis. For example, we found that the pleats had a continuous epidermal layer over their folded surfaces at all stages of development. By the same token there was no evidence at any stage of a splitting and redifferentiation of the epidermal layer.

Our studies included both palmate and pinnate palm species and gave virtually identical results regardless of the species' blade morphology. Therefore we suggest that our findings are probably valid for members of the palm family in general. One other issue, however, still remained unresolved. Why did the primordial folds in the leaves of some palm species assume a shape superficially suggestive of tissue splitting? Having been misled ourselves, we were anxious to find out, once the pleats were initiated, what forces might be influencing their shape.

It had been suggested in the 19th century that space limitations within the terminal bud were responsible for the initial pleating of palm-leaf primordia. The proposal was almost certainly inspired by the sight of tightly folded young leaves in the crown of palms. Could this proposed relation between pleating and space restrictions, obvious in the later stages of leaf development, apply to the stage where the leaf primordia were virtually microscopic in size and crowding seemed unlikely? Circumstantial evidence suggests not. For example, if pleating were simply a response to space limitations, one might logically expect the entire leaf-blade surface to be folded. Yet the margins of the leaf blade are never folded. What is more, when one analyzes the relation between the space within the sheath of the next-older leaf and the shape and distribution of pleats on the blade of the next-younger leaf, one finds no correlation between the space available for expansion and the orientation and extent of the pleats.

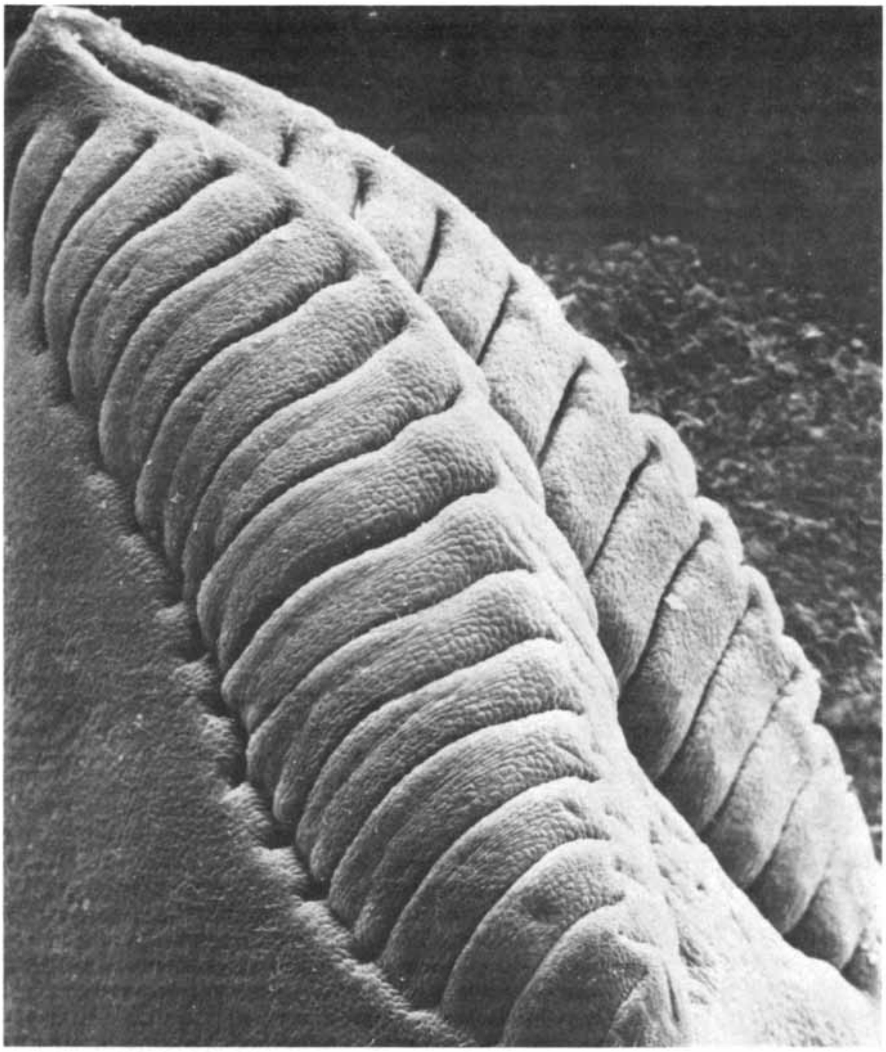
This is not to say that space limitations in the bud do not have an effect on the shape of the pleats. For example, in palms that have the bud leaves tightly packed together the outer leaf surface appears flat and the pleat furrows resemble slits cut into the leaf blade. In palms that have less tightly compacted bud leaves, however, the pleats bulge quite conspicuously, giving the impression that they arise as a result of blade folding rather than of tissue splitting.

An acid test of the possible role of crowding in leaf-blade pleating would be the growth of isolated leaf primordia in a nutrient medium to allow observation of their development in an uncrowded environment. Pilot experiments have established the possibility of rearing such primordia in vitro from an early stage, when they are less than a millimeter in length. We have not yet, however, undertaken a rigorous analysis of their development to see how it may compare with the situation in vivo.

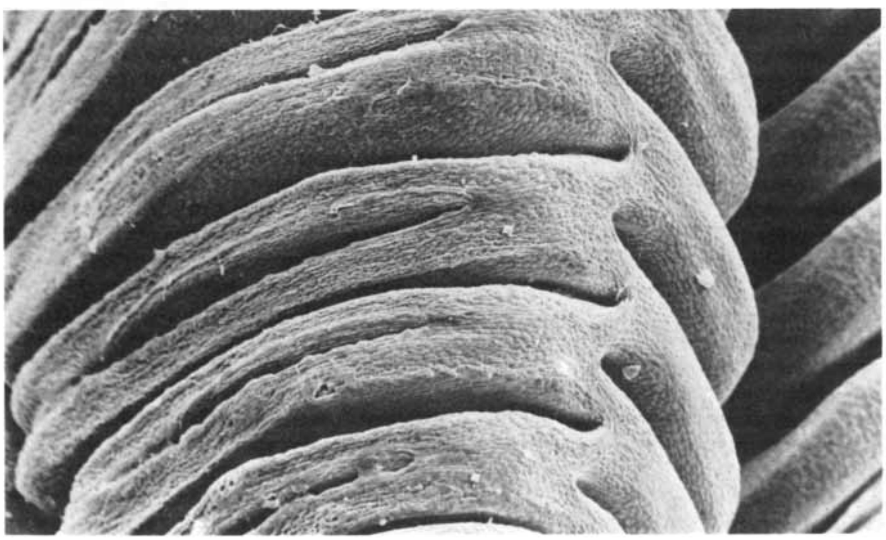
Our studies of palm-leaf development were inspired by the suggestion that the leaf blade's dissection into leaflets involved a process unique in the plant kingdom: tissue splitting. The investigation established, however, that whereas a process of tissue separation is indeed sometimes responsible for cleavage of the leaf blade's initial folds into distinct segments, the development of a pleated leaf blade involves a process no more complicated than differential growth.

In essence, then, the morphogenetic process responsible for palm-leaf dissection is really much the same process responsible for the more typical development of leaflets in the arum *Zamioculcas*. The two processes differ only in the site where the differential growth takes place. In the arum this site is at the free edge of the leaf blade, whereas in the palm the growth takes place within the surface of the blade at some distance from the blade edge. Indeed, one can argue that all the complexity characteristic of palm-leaf development, from pleating and tissue separation along selected pleats down to the final abscission of the reins, is a consequence of leaflet initiation within the blade rather than at its edge. Viewed in this light the many idiosyncrasies of palm-leaf development seem far less aberrant.

In the long run, of course, the interests of botany will be poorly served by any mere cataloguing of developmental pathways in plants. Instead the botanist's ultimate goal must be the coupling of such developmental information with the rich data from the molecular levels of biological organization. For example, most molecular biologists assume that the systems of control they have diligently traced in plants are to be found in all the higher plants. This may not be true. Perhaps the different kinds of development described here are indicative of greater differences among molecular control systems than seem apparent now. It may well be that future studies of morphogenesis will play a significant role in investigators' thinking about the mechanisms of plant growth and development only if they effectively integrate information available from the other fields of biological research.



PLEATING FOLDS in the maturing leaf blade of a pinnate palm of the genus *Chamaedorea* are seen in this scanning electron micrograph. The specimen, three millimeters long, is magnified 45 times. The inner ridges at the right, like those partly visible just to the left, will become the midribs of each of the leaflets-to-be. Some outer leaflet ridges, visible only to the left, show developing central axial depressions. Tissue cleavage will continue along these depressions until each of the leaflets is separated from the next (*see the further advanced specimen below*).



LEFT OUTER RIDGES of the same pinnate leaf blade are seen in a scanning electron micrograph at the same magnification. The tissue splitting of the outer ridges has progressed and the relation between the leaflet halves-to-be and their respective midribs is more readily evident.

Particles with Naked Beauty

The fifth quark, embodying the "flavor" known as beauty, has now been seen in combination with an antiquark not of its own flavor. The beauty of the new composite particles is accordingly exposed

by Nariman B. Mistry, Ronald A. Poling and Edward H. Thorndike

The ultimate building blocks of matter are currently thought to be the small set of indivisible particles called leptons and quarks. The world we normally experience is composed almost entirely of one kind of lepton, the electron, and two kinds of quark, arbitrarily labeled "up" and "down" (or simply *u* and *d*). The fractionally charged quarks, which are distinguished from one another by the quantum properties known as flavors, are bound together in different combinations to constitute the particles of the atomic nucleus: the proton (which has the composite structure represented by the letters *uud*) and the neutron (*udd*).

A complete description of all the known subatomic particles, including those observed only fleetingly in high-energy particle accelerators, requires more of these fundamental constituents. Six kinds of lepton are recognized, falling naturally into three doublets, or pairs. Until recently four kinds of quark were also well established, forming two similar doublets. In addition to the *u* and *d* quarks of ordinary matter a second pair of quarks is needed to embody the special flavors termed strangeness (*s*) and charm (*c*). This article summarizes the experimental evidence for a fifth kind of quark: the bottom, or *b*, quark, which embodies the flavor referred to as beauty.

The first indication of the existence of quarks endowed with beauty came six years ago as the result of experiments in which a beam of high-energy protons was directed against a stationary target. Among the products of the bombardment were a number of massive two-quark particles that appeared to consist of a new quark, the *b* quark, bound to its own antiparticle, the $\bar{b}$ quark. In such a particle the opposite flavors cancel and the particle is said to be flavorless; in other words, the observed particles were judged to have hidden, or concealed, beauty.

For the past few years our group at Cornell University has taken another approach to the search for particles with beauty. In our experiments oppositely directed beams of high-energy electrons and positrons are brought into collision; the colliding particles of matter and antimatter annihilate each other in a flash of electromagnetic radiation from which new particles can materialize. With this method we have succeeded in creating particles in which the *b* quark is apparently bound to an antiquark of a different flavor. In these particles the beauty is no longer hidden; indeed, it is said to be naked. The observation of particles with naked beauty is expected to yield new information on the relations among the fundamental particles. It is also expected to focus renewed attention on the search for yet another quark to complete the third quark doublet. The predicted but so far unobserved sixth quark is labeled the top, or *t*, quark, and its flavor is referred to as truth. Before telling the story of the discovery of the beauty quark we shall review in somewhat greater detail the evolution of the lepton/quark description of matter.

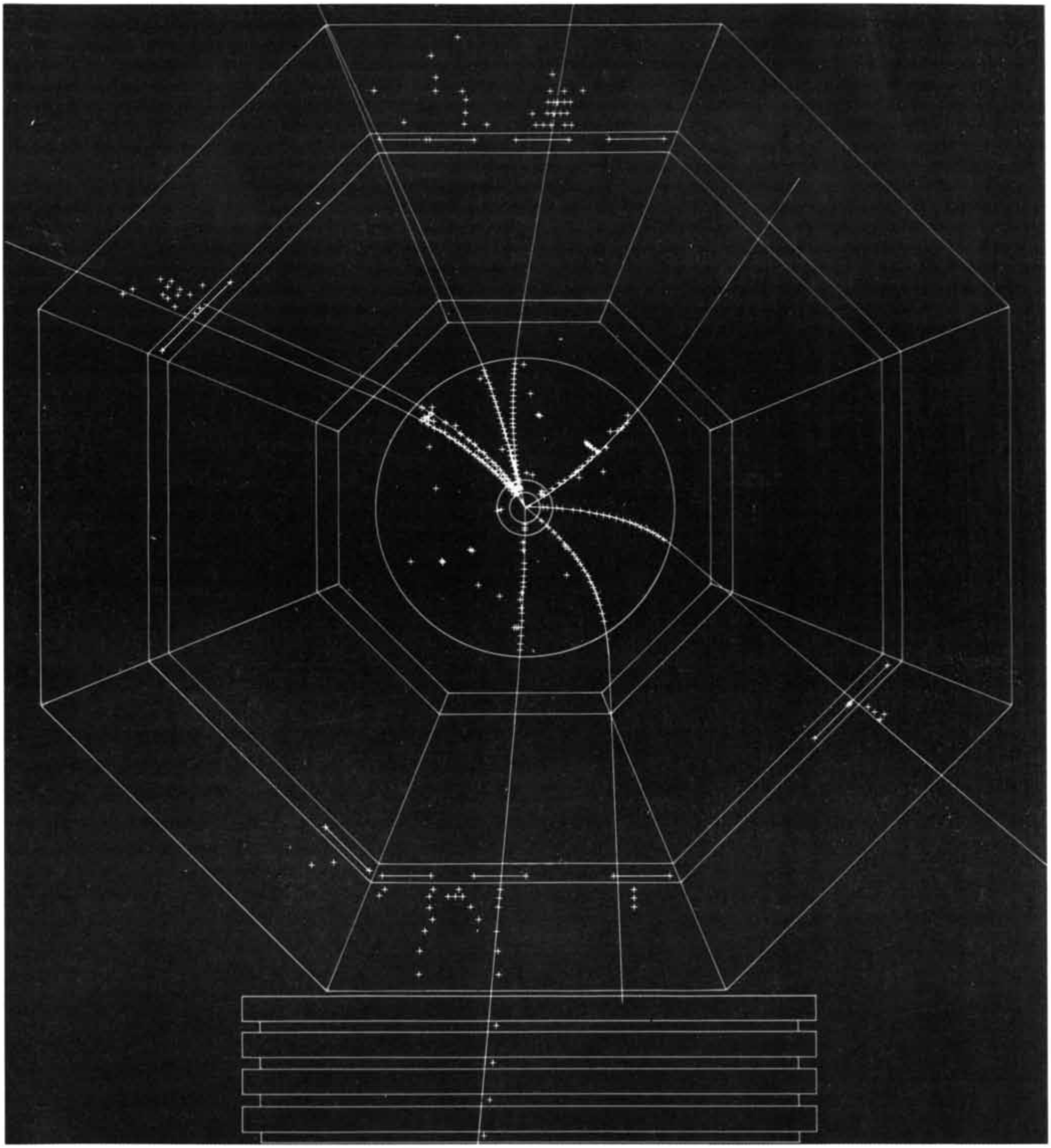
The three lepton doublets consist of the electron (e^-) and its uncharged partner, the electron neutrino (ν_e), the muon (μ^-) and its neutrino (ν_μ) and the tau (τ^-) and its neutrino (ν_τ). For each lepton there is an antilepton, and so the full membership of the class also includes the positron (e^+) and its antineutrino ($\bar{\nu}_e$), the positive muon (μ^+) and its antineutrino ($\bar{\nu}_\mu$) and the positive tau (τ^+) and its antineutrino ($\bar{\nu}_\tau$). The leptons are very small, no more than 10^{-15} centimeter in diameter, and are probably pointlike. Although the charged leptons behave alike, their masses are quite different. The muon is 206 times as massive as the electron, and the tau is almost 3,500 times as massive as the electron. The neutrinos are very light and may actually be massless.

Each of the three quark doublets (*u* and *d*, *s* and *c*, *b* and *t*) is made up of one quark with charge 2/3 and another with charge $-1/3$. Like the leptons, the quarks seem very small and are usually thought of as point objects. Similarly, for each quark there is an antiquark, with opposite charge and opposite flavor. For example, since the *b* quark has charge $-1/3$ and -1 unit of beauty, its antiquark, the $\bar{b}$, has charge $+2/3$ and $+1$ unit of beauty.

With 24 building blocks (leptons, antileptons, quarks and antiquarks) it is quite remarkable that under most circumstances matter can be accounted for with only one lepton doublet (e^-, ν_e) and one quark doublet (*u*, *d*). In addition to the three fundamental particles already mentioned as the main constituents of ordinary matter the fourth member of these first two doublets, the electron neutrino, is emitted by matter as a by-product of natural radioactivity.

Quarks are bound in other combinations to make up many particles besides protons and neutrons. Two possible combinations exist: groups of three quarks, called baryons, and quark-antiquark pairs, called mesons. Together the two classes of particles built from quarks are called hadrons. Hadrons, although they are small, are not pointlike. Their diameter of about 10^{-13} centimeter reflects their composite structure. Because of the strength of the force that binds quarks together, quarks are not detected in isolation. Physicists learn about quarks by studying hadrons. More than 100 different hadrons are known; most of them, however, are seen only in high-energy collisions produced by particle accelerators.

Of the four forces of nature three—the strong force, the electromagnetic force and the weak force—play important roles in the interactions of leptons and quarks. The fourth force, gravity, which is important in the macroscopic world, is negligible on the scale of quarks and leptons. In the modern view of the forces two particles interact by exchanging entities called gauge bosons. The exchanged objects behave somewhat like particles, but they do not have well-de-



BEAUTY-FLAVORED MESONS (that is, particles made of two quarks, only one of which is endowed with the property called beauty) are responsible for the pattern of particle tracks represented in white in this computer-generated display. The origin of the tracks was a head-on collision between an electron and a positron. The colliding particles of matter and antimatter had been accelerated in opposite directions within the circular vacuum chamber of the Cornell Electron Storage Ring (CESR). The collision took place inside the CLEO detector, one of two large particle-identification systems set up at the interaction points where the counterrotating beams cross. (A plan view of the accelerator and a three-dimensional view of the CLEO detector appear on page 111.) The mutual annihilation of the electron and the positron resulted in a flash of electromagnetic radiation, from which there promptly materialized a massive unstable meson called the upsilon (Υ'''), consisting of a beauty quark (b) loosely bound to its own antiquark ($\bar{b}$). In the subsequent decay of the Υ''' two beauty-flavored mesons, designated B and $\bar{B}$, were formed, and they in turn decayed into an assortment of other particles. The passage of the electrically charged decay products through various parts of the detector was recorded as a pattern of "hits" (*white crosses*). By fitting tracks to the hits an automatic pattern-recognizing computer program determined the direction and momentum of each particle and in some cases identified the type of particle. For example, the track at the six-o'clock position was attributed to an extremely penetrating muon because of the string of hits registered in a special set of detection chambers arrayed around the outside of the detector. The track at 12 o'clock was identified as a probable electron on the basis of information recorded about the rate at which it deposited energy in an inner set of chambers, together with the telltale "shower" pattern it left in the electromagnetic-shower chamber at the top of the detector. A complete analysis of such collision events relies on data from all the detector's components, including some not shown here.

fined masses and they exist very briefly.

The strong force acts only among quarks; it is transmitted by the gauge bosons called gluons, which are so named because they are the "glue" that binds quarks together into hadrons. The electromagnetic force acts between any pair of charged particles; it is transmitted by the photon, which is the massless quantum of electromagnetic radiation. The weak force affects both quarks and leptons; it is transmitted by the gauge bosons designated W^+, W^- and Z^0 (also known as intermediate vector bosons). Since the strong force is the strongest of the three forces, it plays the dominant role in quark and hadron processes, unless it is inhibited for some reason. In contrast, the weakness of the weak force, in the range of energies currently being examined, ensures that it is significant only when the strong and electromagnetic forces are suppressed. As the available energy increases, the weak force becomes stronger, until it eventually becomes comparable in strength to the other forces.

The gauge bosons associated with the respective forces act as the agents in all processes in which those forces are involved. The strong bonding of a d and a $\bar{u}$ quark to form a negative pion, or π^- meson, for example, is attributable to the exchange of gluons between the two quarks. Similarly, the electromagnetic force between two electrons is mediated by the exchange of photons. Another kind of electromagnetic interaction is observed when an electron and a positron collide and annihilate each other to form a photon, which promptly materializes into a positive muon and a negative muon. The weak interaction is responsible for processes such as the decay of a tau lepton, in which the tau is transformed into a tau neutrino by emitting a W^- boson. The W^- in turn decays into a negative muon and a muon antineutrino.

The heavier quarks and leptons are unstable and decay into lighter quarks and leptons. The quark decays entail a change in flavor. Gluons and photons are insensitive to flavor and cannot cause such changes; hence the quark and lepton decays are classified as weak interactions. In the decay of the s quark, for example, the s quark becomes a u quark when it emits a W^-. The W^- in turn can materialize into several possible pairs of particles, such as an electron and an electron antineutrino. One might also expect to observe the decay of an s quark into a d quark with the s quark emitting a Z^0 (the neutral counterpart of the W^+ and W^-), which could then decay into an electron and a positron. Careful searches have shown that such processes involving the Z^0 do not occur.

The need to incorporate this experimental fact into the theory led to the hypothesis that the s quark is the charge $-1/3$ member of a doublet with a charge $2/3$ quark of a new flavor, labeled charm. With two quark doublets the interplay of the quarks is such that the probability of the decay of an s quark into a nonstrange quark by the emission of a Z^0 vanishes. The existence of the c quark was predicted on this basis almost 20 years ago, 10 years before the dramatic discovery of psi (ψ) particles, which are mesons composed of a c quark and a $\bar{c}$ antiquark [see "Electron-Positron Annihilation and the New Particles," by Sidney D. Drell; SCIENTIFIC AMERICAN, June, 1975]. This mechanism for suppressing certain strangeness-changing processes can be generalized to other flavors. In particular, if the b quark shares a doublet with a t quark, beauty-changing processes involving a Z^0 are forbidden. If there is no t quark, however, such processes must happen.

After the charmed quark was discovered in 1974 the lepton/quark description of matter seemed to be on a sound footing. The two lepton doublets and the two quark doublets recognized by that time accounted nicely for all the known particles. There was no compelling reason to expect any other leptons or quarks. The discovery of the tau lepton in 1975, however, suggested that there might be more quarks too [see "Heavy Leptons," by Martin L. Perl and William T. Kirk; SCIENTIFIC AMERICAN, March, 1978]. Evidence for a fifth quark was not long in coming. Indeed, there are many parallels between the discovery of the c quark and that of the b.

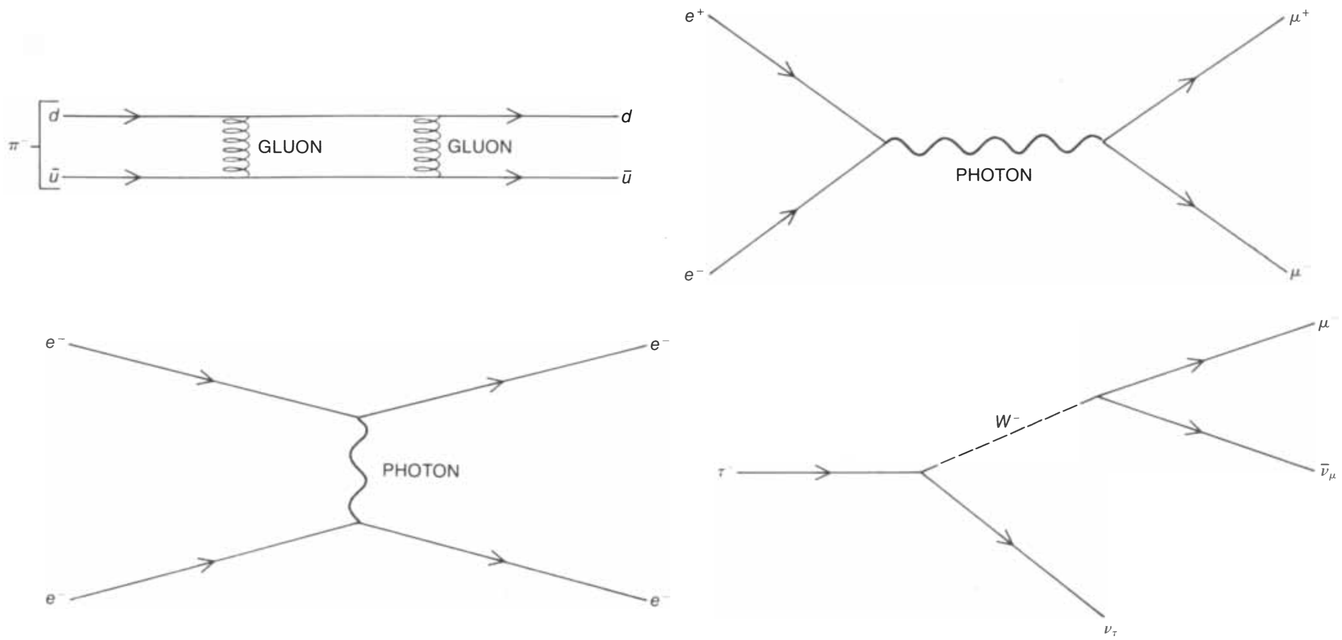
The first piece of evidence resulted from a 1977 experiment by a group of physicists at the Fermi National Accelerator Laboratory. The experimenters, led by Leon M. Lederman, then of Columbia University, bombarded a copper target with a beam of protons at an energy of 400 billion electron volts (GeV) and searched for muons produced in the collisions. When they observed a pair of oppositely charged muons, they considered the possibility that a new particle had been created and had subsequently decayed into the $\mu^+\mu^-$ pair. By measuring the momentum and the direction of each of the muons the experimenters were able to determine the mass of the new particle. The mass distribution that was obtained showed a broad peak at 10 GeV, indicating the presence of the new particle, which was named the upsilon (Υ). Lederman and his colleagues interpreted the shape of their peak as showing that they were in fact seeing two or more narrow peaks (probably three), smeared together by the limited experimental resolution [see "The Upsilon Particle," by Leon M. Lederman; SCIENTIFIC AMERICAN, October, 1978].

The decay of the new particle into a $\mu^+\mu^-$ pair is an electromagnetic process. If the upsilon were to decay by the strong force acting at full strength, the electromagnetic decays would be insignificant in comparison and the $\mu^+\mu^-$ decay would not have been detected. Its

CHARGE	LEPTON DOUBLETS		ANTILEPTON DOUBLETS		CHARGE
0	ELECTRON NEUTRINO	ν_e	$\bar{\nu}_e$	ELECTRON ANTINEUTRINO	0
−1	ELECTRON	e^-	e^+	POSITRON	+1
0	MUON NEUTRINO	ν_μ	$\bar{\nu}_\mu$	MUON ANTINEUTRINO	0
−1	NEGATIVE MUON	μ^-	μ^+	POSITIVE MUON	+1
0	TAU NEUTRINO	ν_τ	$\bar{\nu}_\tau$	TAU ANTINEUTRINO	0
−1	NEGATIVE TAU	τ^-	τ^+	POSITIVE TAU	+1

CHARGE	QUARK DOUBLETS		ANTIQUARK DOUBLETS		CHARGE
+2/3	UP	u	$\bar{u}$	ANTIUP	−2/3
−1/3	DOWN	d	$\bar{d}$	ANTIDOWN	+1/3
+2/3	CHARMED	c	$\bar{c}$	ANTICHARMED	−2/3
−1/3	STRANGE	s	$\bar{s}$	ANTISTRANGE	+1/3
+2/3	TOP (TRUTH)	t	$\bar{t}$	ANTITOP (ANTITRUTH)	−2/3
−1/3	BOTTOM (BEAUTY)	b	$\bar{b}$	ANTIBOTTOM (ANTIBEAUTY)	+1/3

LEPTONS AND QUARKS, the two classes of indivisible, pointlike particles currently assumed to be the ultimate building blocks of matter, fall naturally into pairs called doublets. The leptons, which have integer charges (−1, +1 or 0), are observed routinely in the laboratory. The fractionally charged quarks are not observed in isolation; they apparently exist only as the constituents of the composite particles known as hadrons. For each lepton or quark there is a corresponding antilepton or antiquark with the same mass but with opposite charge and flavor.



THREE FORCES important in the interaction of subatomic particles operate by the exchange of intermediary particles called gauge bosons. In these schematic diagrams showing how such forces work time goes from left to right. The gluon, the gauge boson of the strong force, is responsible for binding the quarks together into hadrons; in this case the d quark and the $\bar{u}$ quark that constitute a negative pion, or π^- meson, are tightly bound by the periodic exchange of a gluon (*diagram at upper left*). The photon, the massless quantum of electromagnetic radiation, serves as the carrier of the electromagnetic force; for example, when two electrons approach each other, they interact electromagnetically by exchanging a photon (*diagram at lower left*). In another kind of electromagnetic interaction an electron and a positron collide and annihilate each other to form a photon, which can then materialize into new particles, such as a muon-antimuon pair (*diagram at upper right*). Leptons and quarks decay through the weak interaction. Thus when a negative tau lepton decays, it is transformed into a tau neutrino by the emission of the intermediate vector boson labeled W^-; the W^-, one of three such carriers of the weak force, decays into one of several possible combinations of particles, such as a negative muon and a muon antineutrino (*diagram at lower right*).

observation suggested that the strong decay of the upsilon is somewhat inhibited. It was recognized that if the upsilon were a meson composed of a new quark and its antiquark ($b\bar{b}$), such a suppression could be readily explained. With enough energy the b and $\bar{b}$ quarks in the meson could separate, creating a new quark-antiquark pair and yielding "beauty-flavored" mesons, for example $b\bar{u}$ and $\bar{b}u$.

The latter decay is a strong-interaction process that would not be suppressed if it were energetically allowed. The observed suppression suggests that there is too little energy available for this process; in other words, the mass of the upsilon is less than the sum of the masses of two beauty-flavored mesons. The only alternative strong-decay mechanism is for the b and $\bar{b}$ quarks to annihilate each other and produce hadrons. This type of annihilation, although it is mediated by the strong interaction, is known to be suppressed. Indeed, it is because of the suppression that the electromagnetic decay is important. In the electromagnetic process the b and b quarks annihilate each other to form a photon, which then materializes into a $\mu^+\mu^-$ pair.

The suppression of the strong decay is expected to result in a lifetime for the upsilon that is considerably longer than the 10^{-23}-second lifetime of a typical strongly decaying hadron. Time intervals on this order are still too brief to be measured directly, but they can be determined from a relation between the lifetime of a particle and the precision to which its mass can be measured. If a particle lives forever, its mass can be determined with perfect accuracy. If its lifetime is very brief, measurements of its mass will give a distribution with a width related to the value of the lifetime; for a normal strongly decaying hadron the width of the mass distribution would be equivalent to an energy range of about .15 GeV. Lederman's experiment did not have sufficient resolution to measure the width of the upsilon's mass-distribution peaks; this determination awaited the arrival of a different technique.

The collision technique by which the upsilon family was discovered was not well suited to the detailed study of these particles. Upsilons are rarely created in proton-nucleus collisions, and when they are, they are usually accompanied by many other particles, making it difficult to follow the process. A far better technique for studying the upsilon family, and the b quark in general, is one in which beams of electrons and positrons are made to collide with each other. Sometimes an electron and a positron will annihilate each other to form a photon, which can materialize into a quark-antiquark pair, such as $b\bar{b}$. When the energy of the electron and that of the positron add up to the mass of an upsilon particle, there is an enhanced probability of the formation of a $b\bar{b}$ pair, which can bind to form an upsilon. Since upsilons decay primarily into hadrons, the experimental technique is to vary the energy of the colliding electron-positron beams and to measure the rate of formation of hadrons. This technique was previously exploited at the electron-positron storage ring SPEAR, a facility of the Stanford Linear Accelerator Center (SLAC), in the discovery of the psi meson.

An experiment to search for the upsilons was carried out in 1978 with the electron-positron storage ring DORIS at the Deutsches Elektronen-Synchrotron (DESY) in Hamburg. Plots of the rate of hadron formation as a function of energy showed two sharp spikes at about 10 GeV, confirming the earlier indications that there were two or more distinct upsilon states. The DORIS machine had been strained to its limit in order to reach the energy of the second upsilon state, however, and it could go no higher. For this reason the third possible upsilon state was out of reach and, of greater importance, so were all $b\bar{b}$ systems in which the b and $\bar{b}$ were not tightly bound.

It was now crucial to see whether

there is indeed a third bound state of $b\bar{b}$ and, at a slightly higher energy, another "resonance" where the b and $\bar{b}$ are only loosely bound and are produced with enough energy to separate and appear as beauty-flavored particles. Theorists had already predicted the energy levels of the bound $b\bar{b}$ system based on evidence obtained from the study of the $c\bar{c}$ system of the psi mesons. It was now up to the experimentalists to confirm or refute that picture and to study the behavior of the b quark through its decays and interactions.

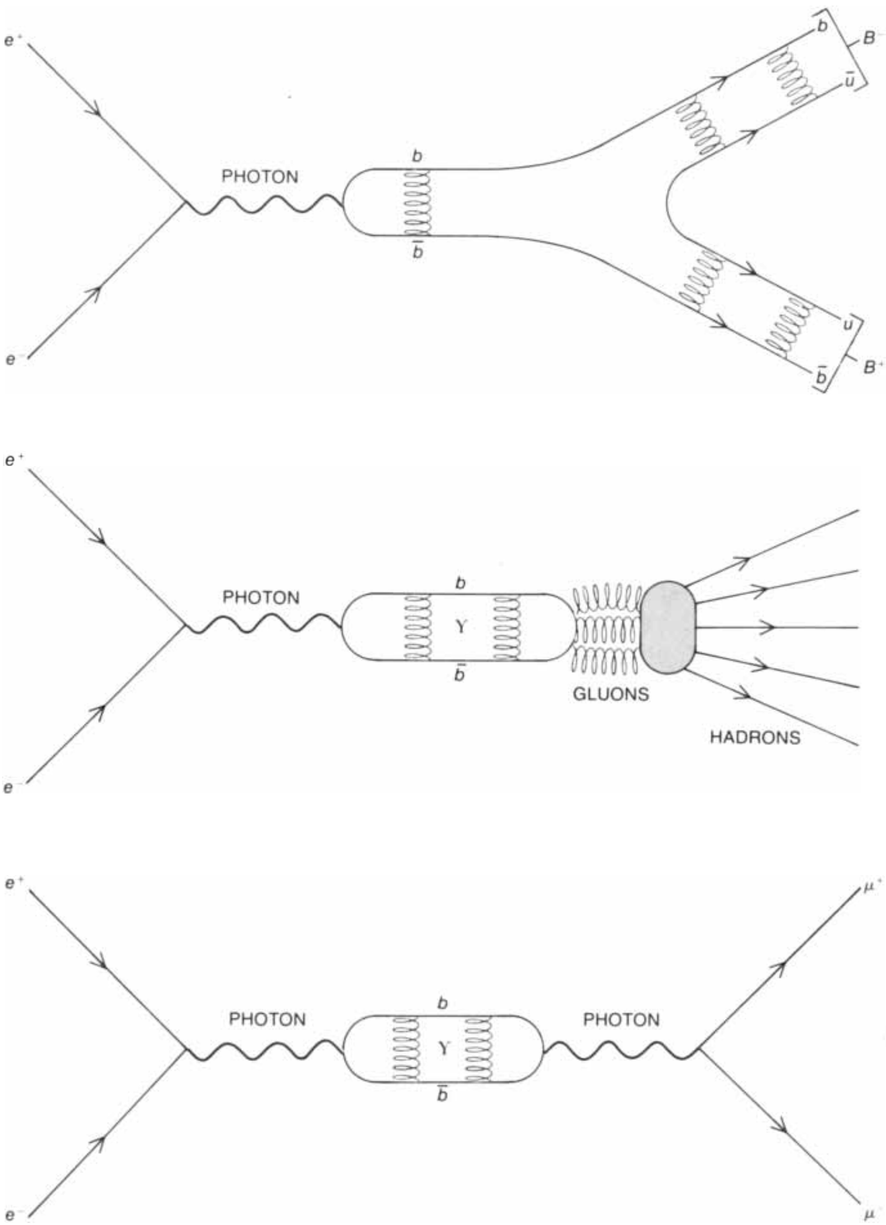
At the time there was a growing sense of excitement among the physicists working on the preparation of the Cornell Electron Storage Ring (CESR). This machine, built under the direction of Boyce D. McDaniel and Maury Tigner, was designed in 1975 before the existence of the upsilon was known. It was a stroke of good fortune that the upsilon system lay right in the middle of the energy range planned for the CESR. Since the CESR was first turned on in April, 1979, it has been the site of most of the detailed work done on the b quark.

At the CESR a "bunch" of more than 10^{11} electrons and another bunch of roughly the same number of positrons travel in opposite directions in a ring with a circumference of 768 meters. The bunches circulate within a toroidal vacuum chamber maintained at a pressure of 10^{-8} torr (10^{11} times less than atmospheric pressure). Bending magnets and focusing magnets confine the particles to a circular path within the vacuum chamber and maintain the size and shape of the bunches. A constant supply of power to the beams compensates for the energy that is radiated away. Because of the comparatively low density of the electrons and the positrons in the bunches, there are very few head-on collisions at each crossing of the bunches. The high rate of crossings (400,000 per second) ensures, however, that there is an adequate rate of electron-positron annihilations. In the argot of particle physicists these collisions are "events." Hadronic events, those in which the collision products include hadrons, are observed a few times per minute.

The detailed study of the production and decay of b quarks requires the observation and study of hundreds of thousands of hadronic events. The rate at which these events happen is low, and so it is up to the experimenters to make sure their detectors are capable of recording every interesting collision. Furthermore, the detector must not miss many particles created in the events, because such particles may be crucial to understanding what happened in the production and decay of new particles.

A great deal of effort has been invested in the design and construction of systems for detecting and studying the products of such interactions. The challenges are formidable. Events may produce many particles, both charged and neutral, that can project in all directions. Furthermore, the rarity of interesting events makes high efficiency essential. The detectors must measure and record particles that pass through the detection zone in a few billionths of a second. At the CESR the electron and positron beams cross at two diametrically opposite intersection regions. The detectors surrounding the respective crossing points reflect two different approaches to the design of detection systems for such colliding-beam experiments.

The CLEO detector, which occupies the south interaction region at the CESR, is a general-purpose detector built and operated by a group of 75 physicists (including the three of us) from seven institutions: Cornell University, Harvard University, Ohio State University, the University of Rochester, Rutgers Uni-

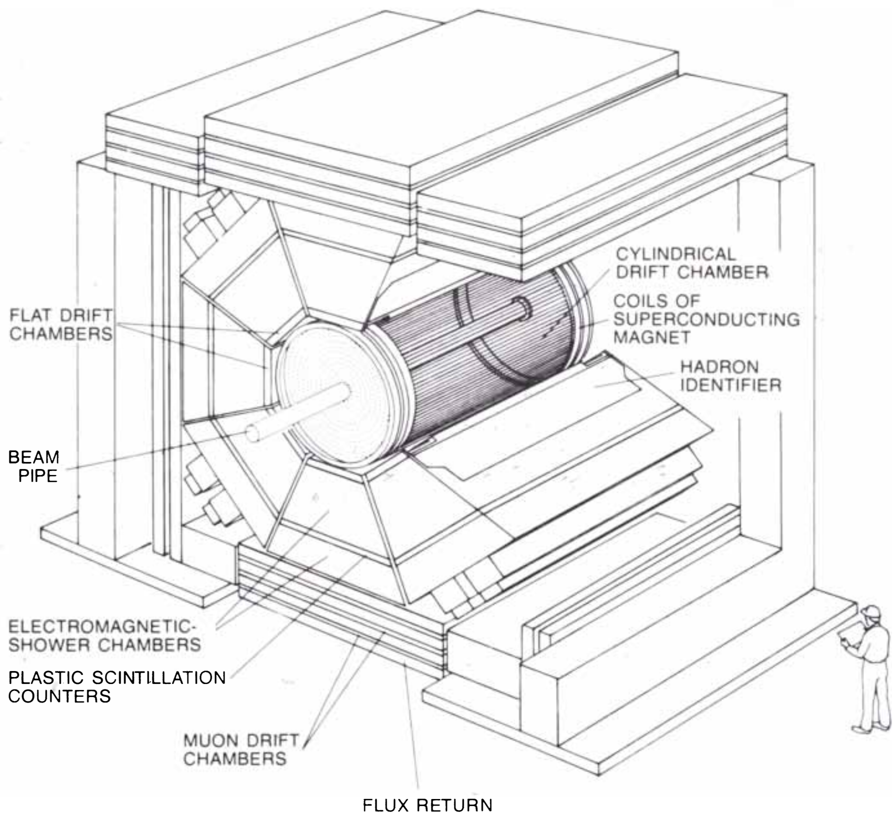


BEAUTY QUARKS ARE CREATED in electron-positron collisions when the photon resulting from the annihilation of the two colliding particles materializes into a $b\bar{b}$ pair. If there is enough energy, an additional quark-antiquark pair can be created and the b and $\bar{b}$ quarks can appear in combination with the other quarks as beauty-flavored mesons (*top diagram*). At certain lower energies, where it is impossible to produce the two mesons, the $b\bar{b}$ pair is created as a bound system: the upsilon meson. Since the quarks cannot separate, they annihilate each other in one of two ways. The first type of annihilation is by the strong interaction with the emission of gluons; the gluons in turn give rise, by a poorly understood process, to hadrons, which are observed in the aftermath of the decay (*middle diagram*). This type of decay, although it is a strong-interaction process, is inhibited, with the result that electromagnetic decays of the upsilon become significant in comparison. In the latter type of decay the $b\bar{b}$ pair annihilates into a photon, which can then form, among other things, a muon-antimuon pair (*bottom diagram*).

versity, Syracuse University and Vanderbilt University. It was designed to provide as much information as possible about each event. That includes detecting charged particles and measuring their momentum, measuring the energy and direction of photons and identifying specific kinds of particles, such as pions, kaons (*s*-flavored mesons), protons, electrons and muons. Because particles can emerge in any direction, it was necessary to surround the interaction point as completely as possible with detector components. As a result CLEO is a very large device, occupying a space six meters long, eight meters wide and nine meters high, and weighing almost 1.5 million kilograms (1,500 metric tons). Careful computer analysis of the thousands of measurements done by individual components makes it possible to develop a clear understanding of what happens in specific annihilation events.

At the heart of CLEO is a superconducting solenoid magnet one meter in radius and three meters long. Inside the magnet is a cylindrical detector, called a drift chamber, that tracks charged particles as they emerge from the collision. As many as 17 separate measurements of the position of a charged particle are recorded, enabling the experimeters to trace the particle's trajectory. By measuring the curvature of the path in the magnetic field the momentum of the particle can be determined. The process of track detection and reconstruction is a crucial element for almost all work with CLEO.

Outside the coil of the solenoid magnet are components that serve to identify different kinds of charged particles. There is a set of plastic scintillation counters, which emit a brief flash of light when a charged particle passes through them. A precise measurement



SITE OF THE DISCOVERY of beauty-flavored mesons is depicted at two scales in this pair of drawings. The plan view at the top shows the overall layout of the CESR facility. Electrons and positrons are accelerated first in a linear accelerator and then in a synchrotron ring before being injected into the storage ring. The particles are confined to separate "bunches," which circulate in opposite directions and cross at two diametrically opposed interaction points, where detectors have been installed to study the products of the electron-positron annihilations. The three-dimensional view at the bottom shows the detailed structure of the CLEO detector. A superconducting solenoid magnet sets up a strong magnetic field in the central cylindrical chamber, where the momentum of each charged particle is determined from the curvature of its trajectory. Other sensing devices are mounted in eight identical modules surrounding the central magnetic field region and in several outlying pieces of equipment.

of the time of the flash, combined with accurate knowledge of the time at which the beams cross, makes it possible to calculate the time needed for a particle to travel from the collision to the counter. From this interval the speed of the particle can be determined. The measurement of the speed combined with the momentum measured with the central drift chamber leads to the determination of the particle's mass. On this basis it is possible to differentiate pions, kaons and protons.

In the original configuration of the CLEO detector the speed of a particle was also measured by a set of Cerenkov counters, which detected the light emitted when a particle passing through them exceeded the velocity of light in the gas with which the counters are filled. (The particle in this case acts somewhat like an airplane breaking the sound barrier.) The Cerenkov counters were useful for identifying electrons. They have now been replaced by gas-filled detectors that measure the rate of energy loss of each charged particle as it travels through the gas. The information can be combined with the measured momentum of the particle to yield the particle's velocity and hence its mass. In this way protons, pions, kaons and electrons can be independently identified.

Electrons have their identity confirmed in yet another part of the detector. When an electron passes through matter, it loses energy by producing what is called an electromagnetic shower. The process begins when a fast electron is deflected by a nucleus and emits an energetic photon. When the photon passes close to another nucleus, it gives rise to an electron-positron pair. The electron and the positron in turn emit photons, and the process continues with the production of a shower of secondary electrons, positrons and photons until all the energy is absorbed. The CLEO electromagnetic-shower detector is constructed of alternate layers of lead (to provide a medium for the production of showers) and charged-particle detectors (to sample the energy and extent of the shower as it develops). The shape and the energy of the showers associated with tracks coming from the collision point are analyzed to identify electrons. Showers that are not associated with any track are recognized as photons.

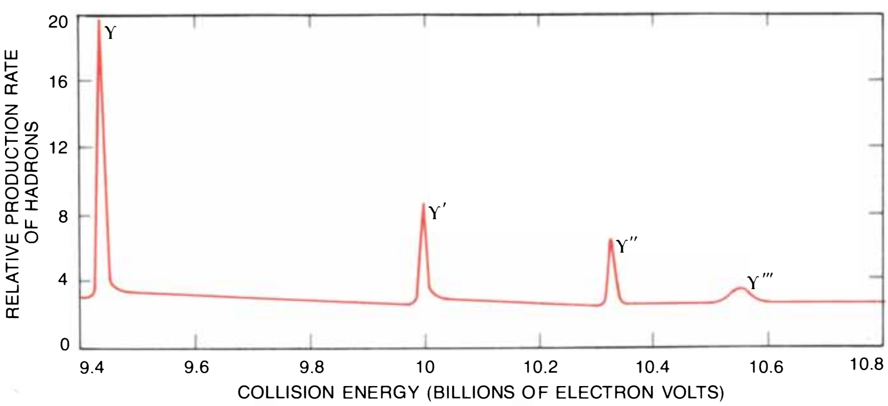
The detection of muons is based on their ability to penetrate large thicknesses of material. A hadron passing through matter is likely to interact strongly, and an electron will generate a shower; a muon, however, loses energy very slowly and can travel through much material without stopping. To exploit this property CLEO is surrounded by a steel absorber about a meter thick. Covering the outside of the steel are flat drift chambers that detect all charged particles emerging from the absorber. A charged particle that is observed to penetrate the steel is identified as a muon.

The CUSB detector in the north intersection region of the CESR was designed with a specific purpose in mind. It was built primarily as a high-resolution photon detector by a group of physicists from Columbia University, Louisiana State University, the Max Planck Institute for Physics and Astrophysics at Munich and the State University of New York at Stony Brook. Its main component is a set of shower detectors consisting of sodium iodide crystals. With these detectors it is possible to make precise measurements of the energy of photons. CUSB is particularly well suited to the measurement of photon energies in the range that is expected to be characteristic of transitions among the upsilon states. The CUSB photon detectors also serve to identify electron showers. Close to the collision point CUSB has tracking chambers to detect charged particles. Parts of the detector are surrounded by steel absorbers and also by counters for detecting muons.

The study of *b*-quark physics at the CESR began in late 1979. The upsilon-particle resonances Υ, Υ′ and Υ″ were quickly found at both CLEO and CUSB by measuring the rate of hadron production as a function of energy. The three resonances appeared as narrow peaks at the respective energies of 9.46, 10.02 and 10.35 GeV. This was the first time the Υ″ was clearly observed as a separate resonance, an event that was announced to the world of high-energy physics in a Christmas card. Since then the resonances have been carefully studied through the observation of several hundred thousand events. Some transitions among the energy states have been observed, and as rarer transitions are identified it should be possible to map out the complete spectroscopy of the upsilon system.

Since the upsilon resonances Υ, Υ′ and Υ″ are hadrons with both beauty and antibeauty, the flavors cancel and the hadrons are flavorless. Accordingly these hadrons are said to have hidden beauty. When such a hadron decays, it does so by the mutual annihilation of the b and $\bar{b}$ quarks, and in that case the decays of the b quarks themselves cannot be examined. At electron-positron energies greater than the equivalent mass of the upsilon states $b\bar{b}$ pairs can be produced with enough energy for the two quarks not to be bound, allowing them to move away from each other. As they separate, the strong force between the quarks causes a light quark-antiquark pair ($u\bar{u}$ or $d\bar{d}$) to be created. The quarks group together to form a B meson and a $\bar{B}$ meson ($b\bar{u}$ and $\bar{b}u$, or $b\bar{d}$ and $\bar{b}d$). In the B meson the beauty is no longer hidden. It is this particle that is said to display naked beauty (or bare bottom). In contrast to hidden beauty, naked beauty must decay by the weak interaction. Therefore by studying the beauty-flavored B mesons one can study the weak interactions of the b quark.

The upsilon resonances were originally interpreted as being three different energy states of the bound $b\bar{b}$ system. It seemed reasonable to expect that other resonances would exist above the threshold for decay into beauty-flavored hadrons. A bump-hunting expedition in the energy region above the Υ″ early in 1980 yielded the expected resonance, at 10.58 GeV. Unlike the first three upsilon resonances, which were all extremely narrow, the fourth resonance, the Υ‴, was quite wide, indicating that it was decaying by the unsuppressed route into B and $\bar{B}$; in other words, it was a source of beauty-flavored particles.



UPSILON FAMILY OF RESONANCES, representing several discrete energy states of the $b\bar{b}$ meson, are responsible for the peaks in this plot of the overall rate of hadron production as a function of the energy of the colliding electron-positron beams recorded in the CLEO detector. The Υ, Υ′ and Υ″ peaks are quite narrow, indicating that they correspond to tightly bound $b\bar{b}$ states created at energies below the threshold for the production of beauty-flavored mesons. The Υ‴ peak is wider than the other three peaks, indicating that the $b\bar{b}$ system at the higher energy state must be decaying to yield a $B\bar{B}$ pair (*see illustration on opposite page*).

To establish clearly that this interpretation was the correct one it was necessary to find evidence for particles carrying the beauty flavor in Υ''' decays. As we have seen, quarks of one flavor can decay into those of another flavor only through the weak interaction. One class of unmistakable weak decays consists of "semileptonic" decays in which a meson decays into a muon or an electron, together with a neutrino and one or more hadrons. CLEO was well equipped for identifying both muons and electrons, and the photon detectors of CUSB were also suited to the identification of electrons; accordingly both groups of investigators set about measuring the rate of single-lepton production in Υ''' decays in order to confirm the presence of the weak decays of beauty-flavored particles.

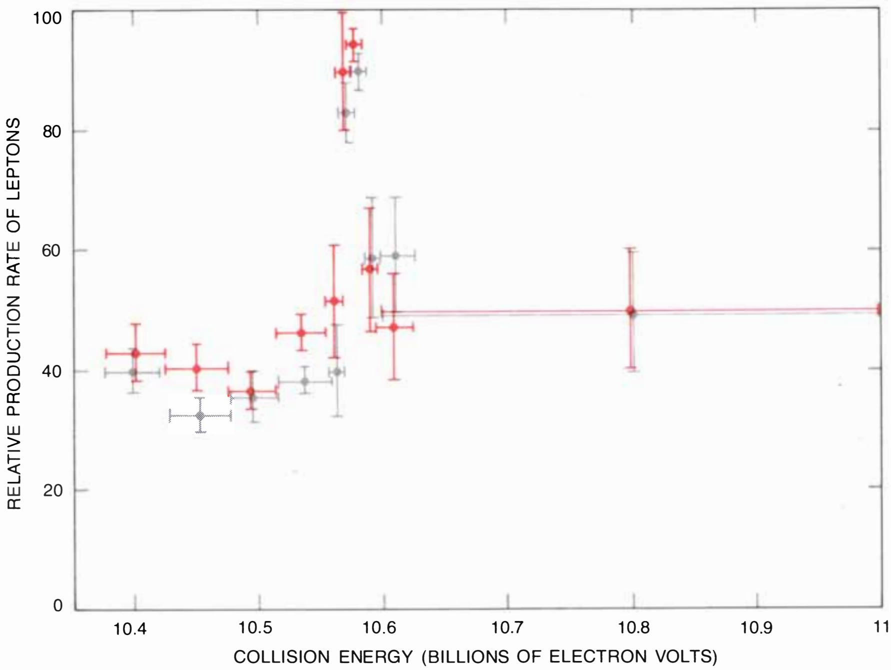
Because of the difficulty of detecting leptons and the low rate at which Υ''' events were accumulated, it was some time before a definitive conclusion could be drawn from the lepton yields. Through the first half of 1980, however, it gradually became clear that the rate of lepton production for events at the Υ''' energy is significantly greater than it is for events at neighboring energies. Before accepting the conclusion that we were seeing beauty-flavored hadrons, however, we had to consider the possibility that the leptons are produced by some mechanism other than the weak decay of *B* mesons. The observed distribution of lepton momenta is qualitatively what one would expect from the decay of a heavy meson such as the *B* and not from the decay of a lighter charmed meson. Since the leptons are produced singly and not in oppositely charged pairs, they cannot be the product of an electromagnetic process. When all alternative explanations were ruled out, the conclusion became clear: we were seeing naked beauty.

Although the lepton yields demonstrated the presence of naked beauty in Υ''' decays, we knew little about the *B* meson. Since the Υ'' is below the threshold for $B\bar{B}$ decay and the Υ''' is above it, the mass of the *B* meson must lie between half of the mass of the Υ'' and half of the mass of the Υ''', or in the range from 5.18 to 5.29 GeV. To demonstrate that the *B* meson is a real object and to measure its mass, it is necessary to reconstruct one from the debris of an Υ''' event. To do so calls for taking the tracks of an event in various combinations and determining whether any particular combination can be demonstrated to be the decay products of a *B* meson. The procedure is difficult because a *B* meson often decays into a large number of particles, some of which are neutral and can escape detection. Others emerge along the beam pipe and are not seen for that reason. What is more, each event has both a *B* and a $\bar{B}$ meson, and so the number of incorrect combinations can be extremely large.

By concentrating on decay modes of the *B* that involve only a few particles the CLEO group has recently succeeded in reconstructing *B* mesons. To do so a sample of 150,000 events collected over a nine-month period was searched for the charmed particles known as neutral *D* mesons that decayed to form a kaon-pion pair. In some cases the D^0 was produced directly in the decay of a *B* meson. In others the *B* decayed into a charged excited state of the charmed meson, which in turn decayed into a D^0 and a charged pion. The reconstructed *B*'s had all decayed into a D^0 or a charged excited state and one or two charged pions, giving a total of from three to five charged particles and no neutral ones. This is in marked contrast to the typical *B*, which decays into nine particles, three of which are neutral. The masses of the 18 *B* candidates cluster around 5.274 GeV.

Our mass measurement for the *B* demonstrates that the Υ''', at 10.58 GeV, is only .03 GeV above the threshold for $B\bar{B}$ production (twice the *B* mass, or 10.55 GeV). When the Υ''' resonance decays into $B\bar{B}$, there is just enough energy to make up the mass of the two mesons; as a result they are produced virtually at rest and with no accompanying particles to confuse the experimenter. Thus the Υ''' is an ideal "factory" for making beauty-flavored particles: it is both an abundant source of *B* mesons and an uncluttered place to observe and study the decays of *b* quarks. There are many features of these decays we should like to explore. For instance, it is important to learn how quickly the *b* quark decays and what it decays into. If history is any guide, these experiments will lead to an improved understanding of the relations among the different flavors of quarks.

Although the *b* quark is expected to decay slowly on the time scale of elementary-particle processes, the decay seems very fast on a macroscopic time scale. The lifetime of the *b* is predicted to be between 10^{-13} and 10^{-14} second. In that brief interval a *B* meson traveling at a speed of 10 percent of the speed of light will move only .0002 centimeter. This distance is too small to be measured with the CLEO detector. So far the best that can be said from the CLEO data is that the lifetime is less than 10^{-10} second. A similar study done at the electron-positron storage ring named PETRA at DESY had the advantage of higher-energy (and therefore faster-moving) *B*'s. In spite of an extremely low rate of *B* production, that experiment gave an up-



CLEARER INDICATION that beauty-flavored particles are produced in Υ''' decays is given by these plots of the rate of lepton production in electron-positron collisions as a function of the collision energy. The sharp increases in the production rate of electrons (*gray*) and muons (*color*) appear at an energy equivalent to the mass of the Υ''' resonance. The Υ''' evidently decays into some combination of products that in turn decay into a lepton and other particles. Such "semileptonic" decays are a telltale sign of a weak-interaction process and confirm that the Υ''' resonance decays to yield a $B\bar{B}$ pair. Symbols give range of experimental error.

per limit on the B lifetime of 5×10^{-12} second.

What are the products of the decay of the b quark? The possibilities can be grouped into three classes. The standard view is that the b emits a W^- and is transformed into a c or a u. An alternative is for the b to emit a Z^0, transforming the b into an s or a d. This is a beauty-changing process involving a Z^0 and is forbidden if the b is part of a doublet with t. Finally, it is possible that the b decays in some entirely new and unexpected way, not involving the emission of either a W^- or a Z^0. Let us consider each of these possibilities in turn.

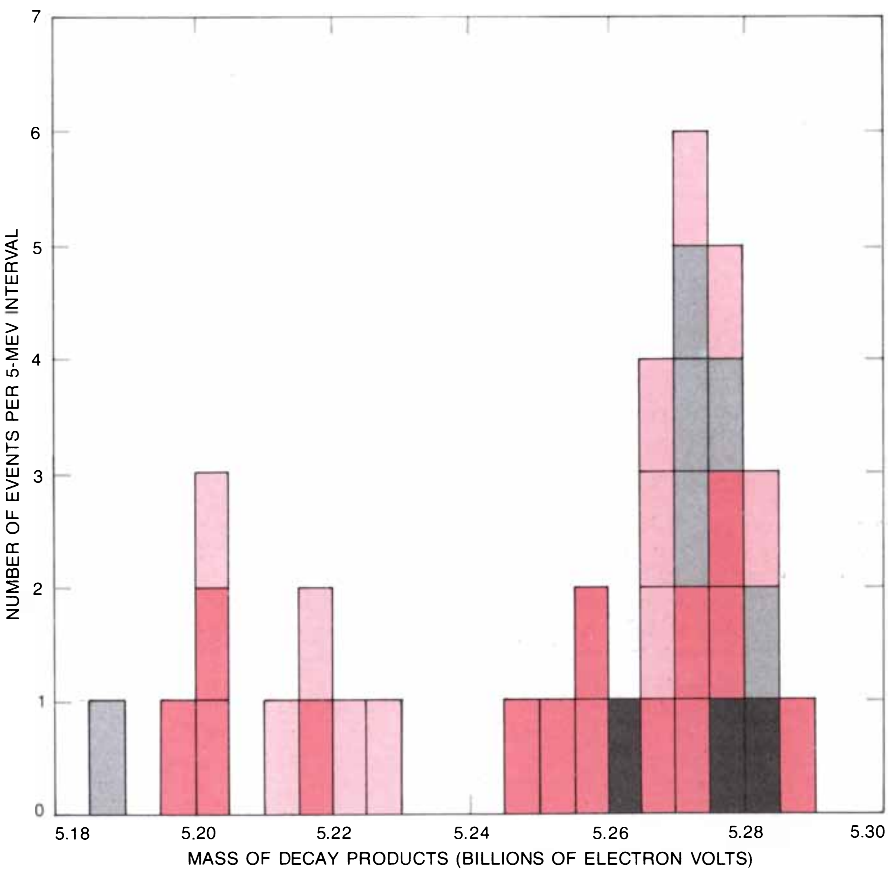
In the standard view, with the b decaying into a W^- and a u or a c, the end products of b decay must include the particles into which the W^- itself decays. The options are $e^-\bar{\nu}_e$, $\mu^-\bar{\nu}_\mu$, $\tau^-\bar{\nu}_\tau$, $\bar{u}d$ and $\bar{c}s$. In general the quark options are three times likelier than the lepton options. By counting the possible outcomes and allowing for some necessary theoretical adjustments one can predict that the end products of b decay will include an electron about 13 percent of the time and a muon the same percentage of the time. The prediction agrees quite well with experimental measurements from CLEO and CUSB, lending support to the standard view of b decay.

This theoretical picture makes no prediction about whether the b quark decays more often to c or to u. Accordingly that is one of the most important things one can hope to learn about the b quark. The recent direct observations of D^0 mesons in B-meson decay demonstrates that the b quark does in fact decay to c at least some of the time. There are two procedures by which the decay preference of the b quark has been measured. The well-known dominance of the decay of c into s suggests the most obvious approach: one would expect B decays to produce more strange particles (kaons) if the b quark decays predominantly into c than if b decays into u. CLEO can detect both charged and neutral kaons. A careful study of kaon production shows there is a definite enhancement in the rate of kaon production in Υ''' events compared with nonresonance events, which translates to a combined rate for charged and neutral kaon production of about 1.4 particles per B decay. Theoretically one would expect about .8 kaon for each B decay if b always decays into u and 1.6 if b always decays into c. The measurement suggests that the decay of b into c is dominant.

A better way to examine the question is to analyze the momentum distribution of the leptons from semileptonic B decays. Among the products of one of these decays there must be a lepton, a neutrino and one or more hadrons. If b decays into c, a charmed particle must be created. The lightest charmed particle is the D meson, with a mass of 1.87 GeV. Since this mass is quite large, it limits the amount of energy that can be carried away by the lepton as momentum. In contrast, if b decays into u, it is not necessary to create such a massive hadron, and it is possible for the lepton to have more momentum. By examining the production of high-momentum leptons and comparing these observations with what would be expected if b decays into c (or if b decays into u) it has been possible to demonstrate again that b decays primarily into c. In fact, it appears from this evidence that b decays into u no more than 5 percent of the time.

The standard view of the decay of the b quark, which has been the framework for the foregoing discussion, demands that the b quark share a doublet with the t quark. The discovery of the t quark is crucial to the confirmation of this model. The t quark has been sought in experiments at the electron-positron storage rings PEP (at SLAC) and PETRA (at DESY). The failure to find the t may simply mean that the maximum energy of these searches (37 GeV) is too low to allow the production of bound $t\bar{t}$ states. The alternative is that the standard theory is incorrect and another theoretical description must be found.

If the standard theory is wrong and the b quark does not have a partner, one must consider the possibility that it decays by emitting a Z^0. (Remember, this process is not allowed in the standard theory.) Independent of whether the b turns into a d or an s, the decay products of the Z^0 must be present among the end products of b decay. The Z^0 can produce e^+e^-, $\mu^+\mu^-$, $\tau^+\tau^-$, $\nu\bar{\nu}$, $u\bar{u}$, $d\bar{d}$, $s\bar{s}$ or $c\bar{c}$. Here the e^+e^- and $\mu^+\mu^-$ options are very attractive because they yield clear signals. A search for indications of two routes of b decay, $b \rightarrow \mu^+ + \mu^- +$ hadrons and $b \rightarrow e^+ + e^- +$ hadrons, has yielded no evidence of such



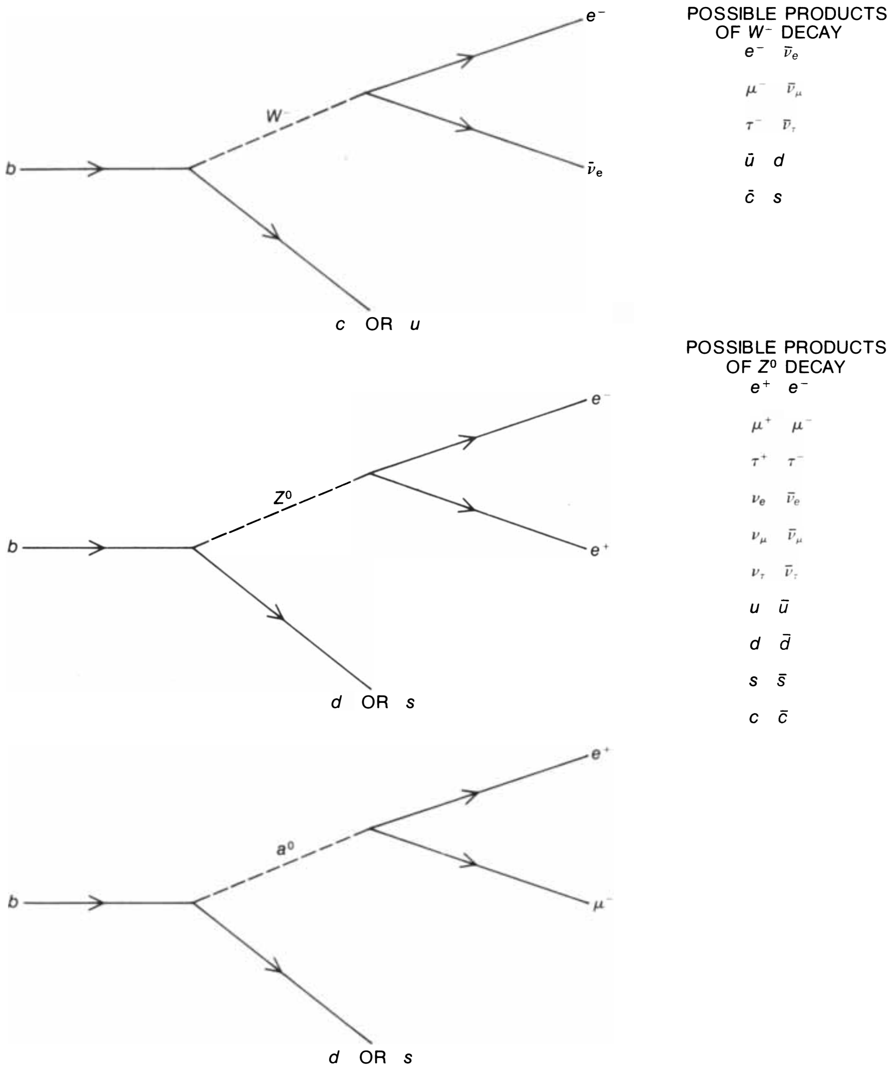
RECONSTRUCTION OF BEAUTY-FLAVORED MESONS was accomplished by the CLEO group by examining the charged-particle tracks from the decay of the Υ''' resonance and tabulating them in various combinations according to four different models of the subsequent decay of a beauty-flavored meson (*key at left*). The resulting bar chart shows a peak in the vicinity of 5.274 billion electron volts. The specific decay modes considered are unusual in that the B meson decays into a small number of charged particles and no uncharged particles (other than the extremely short-lived D^0, or neutral charmed meson). The rarity of these decays and the difficulty of filtering them from other Υ''' decay processes is reflected in the fact that the small number of entries shown is the result of nine months of data collection, totaling 150,000 collision events. The abbreviation MeV in the vertical scale stands for millions of electron volts.

decays, leading to the conclusion that the Z^0 is not a major contributor to b decay.

Finally, there is the possibility that the b decays neither by W^- emission nor by Z^0 emission but in some entirely new way. Within the limits of rather general theoretical considerations only two additional kinds of decay are possible. The first is for the b to decay into a quark and two leptons from different doublets. The second is for the b to decay into a lepton and two antiquarks. If the b quark decays by the first of these exotic processes, with two leptons per decay, one would expect events with many muons, electrons, taus and neutrinos in various combinations. The taus in turn would decay, releasing part of their energy to neutrinos, which would be undetectable. If this kind of decay were significant, one would expect to observe frequent electron and muon production and considerable missing energy in B-decay events. By simultaneously measuring the yield of muons and electrons and the detectable energy in each Υ''' event CLEO has shown that this kind of decay is not dominant.

In the second class of exotic b decays three more antiquarks than quarks are created (one that was originally bound to the b quark and two that were created in its decay). The presence of the three antiquarks results in the creation of one antibaryon per B-meson decay. Similarly, the excess of quarks in $\bar{b}$ decay results in one baryon per $\bar{B}$ decay. Although there are many kinds of baryons (and antibaryons) that can be produced, all eventually decay into protons or neutrons (or into antiprotons or antineutrons). In CLEO protons and antiprotons can be identified, but neutrons and antineutrons escape undetected. If the b decays into a lepton and two antiquarks, one would expect B-decay events to yield, say, a muon, an electron, an antiproton or a large amount of missing energy. By simultaneously measuring the yields of muons, electrons and antiprotons, along with the detected energy per event, the CLEO experimenters have shown that this second kind of exotic decay is not dominant either.

In sum, all the evidence supports the view that the b quark decays by emitting a W^-. Furthermore, there are indications that most of the time it turns into a c quark rather than a u quark.



SEVERAL POSSIBILITIES for the decay of the *b* quark were investigated by the CLEO group. In the standard theoretical model, based on the assumption of three quark doublets, the *b* quark decays into either a *c* quark or a *u* quark by emitting a W^- boson (*top diagram*). The W^- can then decay into five possible final states, two of which ($e^-\bar{\nu}_e$ and $\mu^-\bar{\nu}_\mu$) are the semileptonic modes that were observed in the discovery of naked beauty. If the *t* quark, the *b* quark's hypothetical partner, does not exist, there must be a second mechanism for *b* decay, namely one in which the *b* quark decays into a *d* quark or an *s* quark by emitting a Z^0 (*middle diagram*). Again there is a range of possible final states, two of which (e^+e^- and $\mu^+\mu^-$) have been shown by the CLEO group to occur rarely, if at all. Exotic decay processes, not involving the usual gauge bosons, have been suggested by some theorists engaged in exploring alternatives to the standard model. One example is the decay of the *b* quark into a *d* quark or an *s* quark by the emission of a^0, an exotic neutral boson (*bottom diagram*). The a^0 (for which there is no experimental evidence) need not obey the rules that apply to conventional gauge bosons and so can decay into two leptons of different types, such as a negative muon and a positron. Experiments with CLEO have shown that such exotic decay processes also are uncommon.

Although an understanding of the b quark and its interactions is developing, a number of open questions remain. What is the lifetime of the b quark? At present only upper limits have been established; they are far above the expected lifetime, however, and therefore provide no new information for evaluating the theory. What fraction of the time does b decay into u rather than into c? Data from CLEO and CUSB show that the fraction of b decays into u is small, but how small is it? Does the t quark exist? By showing that beauty-changing processes involving a Z^0 occur infrequently (if at all) and by showing that the unexpected processes that involve neither W^- nor Z^0 are also rare (if not absent) the CLEO experiments have ruled out almost all theories that can avoid having a t quark. Perhaps the t quark has not been found yet because it is too massive, and searches with higher-energy colliding-beam machines will find it. If the t does not exist, the theory of quark decays is in deep trouble.

Even if it is assumed that the t quark will ultimately be found, there remain some profound questions about the quark and lepton doublets. A perfectly satisfactory world can evidently be built of one lepton doublet (e^-, ν_e) and one quark doublet (u, d). Yet at least three doublets of each type are thought to exist, and more may still be discovered. Why does nature behave in such an uneconomical way? How are the various doublets related and how do they differ? What does "flavor" really mean? Further study of the b quark may help to answer some of these questions.

Creole Languages

These widely scattered languages show striking similarities. The development of Creole in Hawaii suggests children learn a language by first constructing an abstract form of a creole

by Derek Bickerton

The ancient Greek historian Herodotus records the story of Psamtik I, pharaoh of Egypt in the seventh century B.C., who set out to discover the original language of humanity. On royal decree two infants were taken away from their parents and put in the care of a mute shepherd, who was instructed to raise the children in isolation from other people. The shepherd was to take note of the first word uttered by the children; "uncorrupted" by the language of their forefathers, Psamtik reasoned, they would begin to speak in the pure tongue from which all other languages were derived. The first intelligible sound the children made was "bekos," which meant bread in the ancient language Phrygian. Therefore, Psamtik maintained, the original language of humanity is Phrygian.

The story has amused generations of linguistics students. Most linguists, who have taken it for granted that no such experiment should ever be carried out, have dismissed the Psamtik experiment as being defective in design and unlikely to yield any useful result. Indeed, the assumption that an "original" vocabulary can be recovered is overoptimistic, and linguistic isolation of the individual, which has been documented in a few cases of severe child abuse, usually results in the absence of language. Nevertheless, a modified form of the experiment has been repeated many times over the past 500 years among the children of slaves and laborers who were pressed into service by the European colonial powers.

These laborers, who were shipped from many parts of the world to tend and harvest crops in Africa, the Indian Ocean region, the Orient, the Caribbean and Hawaii, were obliged to communicate within their polyglot community by means of the rudimentary speech system called pidgin. Pidgin speech is extremely impoverished in syntax and vocabulary, but for the children born into the colonial community it was the only common language available. From these modest beginnings new native languages evolved among the children, which are generically called creole languages. It can be shown that they exhibit the complexity, nuance and expressive power universally found in the more established languages of the world.

Taken at face value, the development of many different creole languages suggests that the search for a single, original language is misguided. For many years, however, scholars have noted a remarkable similarity of structure among all the creole languages. It can now be demonstrated, by considering the origin of creole language in Hawaii, that similarities among creoles cannot be accounted for by contact with other languages, either indigenous or imported. The finding suggests that what is common to creole languages may indeed form the basis of the acquisition of language by children everywhere. There is now an impressive body of evidence to support this hypothesis: between the ages of two and four the child born into a community of linguistically competent adults speaks a variety of language whose structure bears a deep resemblance to the structure of creole languages. Hence, by an ironic stroke of justice, the surviving linguistic remnants of colonialism may offer indispensable keys to the study of our own linguistic heritage.

The historical conditions that favored the development of creole languages are well known. Between 1500 and 1900 England, France, the Netherlands, Portugal and Spain established numerous labor-intensive, agricultural economies on isolated littorals and underpopulated tropical islands throughout the world. The colonies were engaged primarily in monoculture, usually sugar, and their economic viability depended on an abundance of cheap labor imported from distant regions under conditions of chattel slavery. Workers were drawn first from West Africa and later from East Africa, India and the Orient, and they spoke a variety of mutually incomprehensible languages.

Under more salutary conditions of immigration the workers or their children would eventually have learned the language of the local colonial power, but two factors combined to keep them from doing so. First, the number of speakers of the colonial languages rarely exceeded 20 percent of the total population, and it was often less than 10 percent. In other words, there were relatively few people from whom the dominant language could have been learned. Second, the colonial societies were small, autocratic and rigidly stratified. There were few chances for prolonged linguistic contact between field laborers and speakers of the dominant language.

Except in Hawaii, there is little reliable documentary evidence concerning the early linguistic history of the colonial societies. It has generally been assumed that pidgin developed as a contact language solely to allow communication between masters and workers and among workers from various immigrant groups. Creole languages then arose among the children of the workers through the "expansion" of pidgin; there was little occasion for the children to use the ancestral languages of their parents, and they still lacked access to the language of the dominant culture. What is meant by the term "expansion" has remained obscure until my colleagues and I began our studies in Hawaii.

The unique advantage for the study of creole language in Hawaii is that the details of its formation can be reconstructed at least in part from the speech of people still living. Although Hawaiian contact with Europeans goes back to 1778, it was not until 1876 that a revision in the U.S. tariff laws, allowing the free importation of Hawaiian sugar, enabled Hawaiian sugar plantations to increase their output by several hundred percent. A polyglot force of indentured laborers, made up of Chinese, Filipinos, Japanese, Koreans, Portuguese, Puerto Ricans and others, began to be assembled, and by 1900 it outnumbered the other groups in Hawaii, both native and European, by a ratio of two to one.

A pidgin based on the Polynesian language Hawaiian initially served as a means of communication between immigrants and the locally born, but the annexation of Hawaii by the U.S. in 1898 eventually led to the replacement of Hawaiian by English. After 1900 the Hawaiian language declined, and pidgin Hawaiian was replaced as a lingua franca by a pidgin based on English. By the time we began our intensive study of language variation in Hawaii in the early 1970's there were still many survivors, both immigrant and locally born, from the years 1900 until 1920.

Our recordings of locally born people make it clear that the process of creolization was under way by 1900 and was certainly complete by 1920. Most of the linguistic features that characterize Hawaiian Creole English are present in the speech of working-class people born in Hawaii since 1905; before that date the proportion of Creole speakers to the rest of the population falls off rapidly. On the other hand, the speech of immigrants is always some form of pidgin, although just what form it takes depends on the date of the immigrant's arrival in Hawaii as well as the immigrant's language background. The pidgin spoken by the earliest immigrants among our subjects is much more rudimentary than that spoken by the later ones, probably because the latter were exposed to Creole as well as pidgin. Nevertheless, the distinction between pidgin and Creole remains fundamental: anyone familiar with Hawaii can quickly identify the ethnic origins of any immigrant on the basis of speech patterns alone. Without a conversational topic or a person's physical appearance as a guide, however, no one can reliably identify the ethnic origins of any locally born speaker solely on the basis of the speaker's pronunciation or the grammatical structure of the utterances.

One of the main characteristics of pidgin, therefore, is its variability from speaker to speaker. Each immigrant seems to have gone about the task of inventing a makeshift language in some individual way. For example, pidgin speakers of Japanese ancestry generally place the verb at the end of a sentence, as in "The poor people all potato eat" ("All that the poor people ate were potatoes"). Filipino pidgin, however, places the verb before the subject: "Work hard these people" ("These people work hard"). More often word order follows no fixed principle except the pragmatic one that old, shared information is stated near the beginning of a sentence and new information near the end.

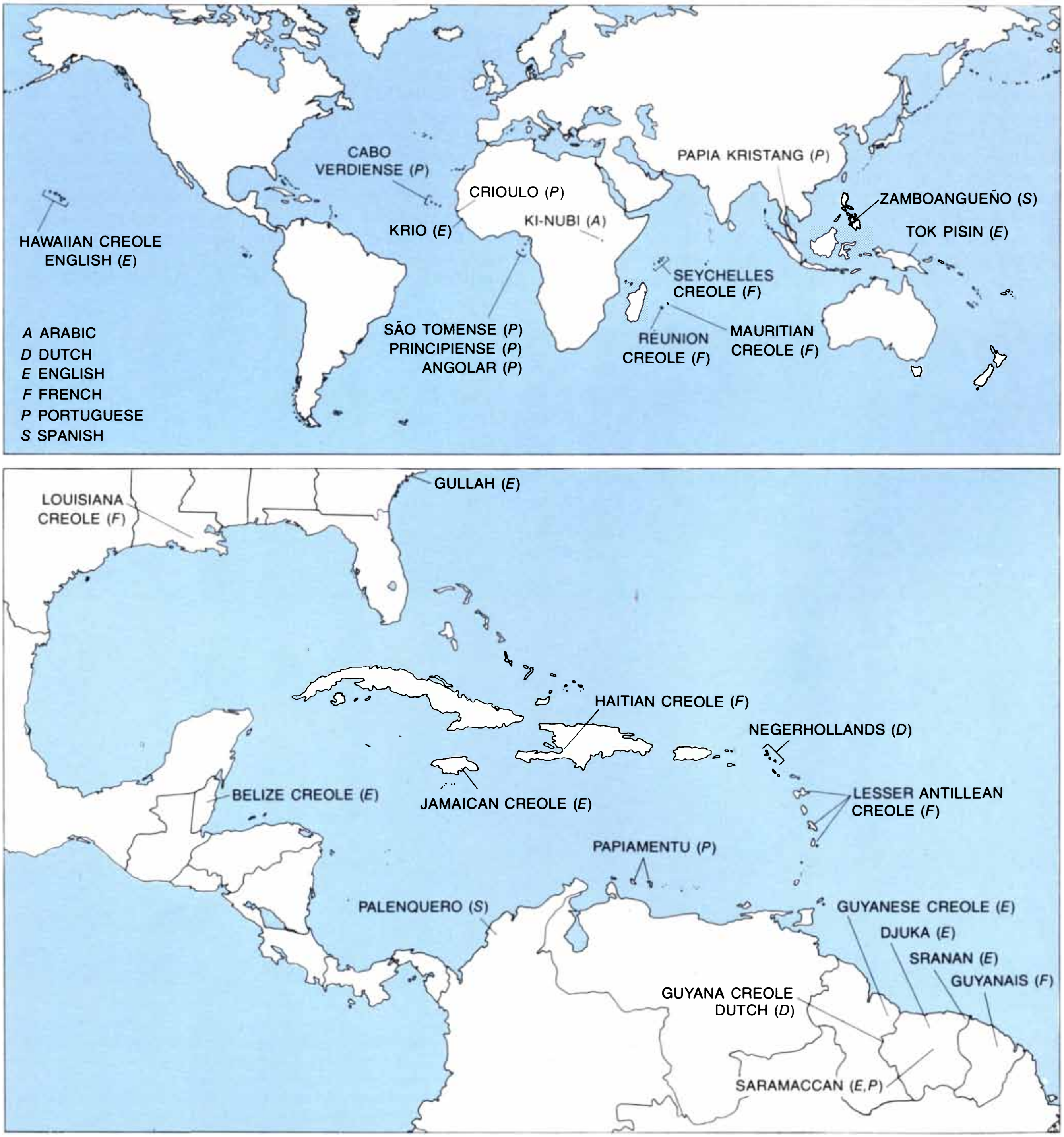
It is probably the case that anything expressible in Creole, or in English for that matter, can also be expressed in pidgin. Nevertheless, the pidgin speaker is at a great disadvantage, because pidgin lacks many of the building blocks possessed by all native languages. Such everyday necessities of language as articles, prepositions and auxiliary verbs are either absent or appear sporadically in a quite unpredictable fashion. Pidgin sentences have no subordinate clauses, and single-clause utterances frequently lack verbs.

The first of the following examples

INDENTURED SUGARCANE WORKERS, who spoke a rudimentary language called pidgin, are shown in a photograph made in Hawaii by the late Ray Jerome Baker in 1924. Thousands of such workers from many countries were brought to Hawaii in the late 19th and early 20th centuries to meet the labor demands of large sugarcane and pineapple plantations. Pidgin language developed out of the need for communication among the various language groups within this polyglot labor force; the workers in the photograph, for example, although primarily from the Philippines, originally spoke a variety of mutually incomprehensible languages such as Visayan, Ilocano and Tagalog. The children of pidgin-speaking parents remained largely isolated from speakers of Hawaiian or English, and they inherited no consistent linguistic model from the previous generation that could meet their needs. There arose among these children a language much more sophisticated than pidgin called Hawaiian Creole English, whose vocabulary is primarily English but whose structure, or grammar, is quite distinct from that of English or any other noncreole language. Socioeconomic circumstances similar to those in Hawaii frequently gave rise to pidgin languages throughout the world; these languages were later developed into creole languages by the children of immigrant laborers. The original photograph is in the collection of Ray Jerome Baker and Robert E. Van Dyke in Honolulu.

was recorded from a pidgin-speaking Korean; omitted words are bracketed in the translation: "And a too much children, small children, house money pay" ("And [I had] too many children, small children, [I had] to pay the rent"). The second example was recorded from a Japanese speaker: "Before mill no more Filipino no nothing" ("Before the mill [was built, there were] no Filipinos here at all"). The third example, recorded from the speech of a retired bus driver, illustrates the heroic measures needed to say anything out of the ordinary in pidgin: "Sometime good road get, sometime, all same bend get, enguru [angle] get, no? Any kind same. All same human life, all same" ("Sometimes there's a good road, sometimes there's, like, bends, corners, right? Everything's like that. Human life's just like that").

The language-learning task confronted by the child born into a community of such speakers is far different from the task imposed on the child who is surrounded by linguistically competent adults. The children of English or Chinese parents, for example, are presented with accurate models to follow. Although their mistakes are seldom overtly corrected, they can almost constantly check their own utterances against those of older speakers and adapt them



WORLDWIDE DISTRIBUTION of creole languages reflects the historical circumstances of their development. Almost all creoles arose on isolated tropical littorals or islands, where colonial powers had established agricultural economies based on cheap immigrant labor. The geographic dispersion of the colonies suggests that creole languages developed independently of one another. The letters in parentheses after the name of each language indicate the colonial language from which most of the vocabulary of the creole is borrowed.

where necessary. When they have mastered the simpler structures of their language, more complex structures are readily available.

For the Hawaiian-born child of immigrant parents, however, there was no consistent linguistic model for the basic word order of simple sentences and often no model at all for the more complicated structures of language. Many such children were born of interethnic or interracial marriages, and so even at home there was little occasion to speak the native language of either parent. Moreover, even among the children not born of linguistically mixed parents there was considerable incentive to abandon the parents' native language and adopt some version of pidgin in the company of peers and neighboring adults. Like first-generation immigrant children elsewhere, the children of Hawaiian immigrants often became bilingual or even trilingual, and they adopted the common language of their peers as a native language in spite of considerable efforts by their parents to maintain the ancestral tongue.

The historical evidence is consistent with the view that the structure of Creole arose without significant borrowing from other languages. Bilingual or trilingual children of school age need not (and usually do not) mix up the structural features of the languages they speak, and there is no reason to suppose such crossovers were common in Hawaii. The most compelling argument for the autonomous emergence of Creole, however, is its observed uniformity. How, within a single generation, did such a consistent and uniform language develop out of the linguistic free-for-all that was pidgin in Hawaii? Even if all the children of various immigrant groups had begun by learning the languages of their parents, and even if the differences among the various pidgins had been smoothed by interaction and contact among the children, the homogeneity of the language that developed remains in need of explanation. Fifty years of contact among pidgin-speaking adults were not enough to erase the differences among the national language groups; the homogeneity must have resulted from the differences between children and adults.

One might still suppose the structural uniformity of Creole is derived from certain structures of one of the ancestral languages or perhaps from certain structures of English, the language of the plantation owners. There are numerous differences, however, between the structure of Creole and the structure of any of the languages with which Creole speakers might have been in contact. In English, for example, it is possible to refer to an object or a group of objects in a nonspecific way, but English grammar forces the speaker to state in advance whether the number of unspecified objects is one or many, singular or plural. One must say either "I am going to the store to buy a shirt" or "I am going to the store to buy shirts," even though one may not want to commit oneself in advance to buying any particular number of shirts.

In Creole a grammatically neutral marker for number can be employed on the noun "shirt" in order to avoid specifying number: "I stay go da store for buy shirt" ("I am going to the store to buy shirt"). Moreover, in Creole the addition of a definite or an indefinite article to "shirt" changes the meaning of the sentence. In saying "I stay go da store for buy one shirt" the Creole speaker asserts the shirt is a specific one; in the sentence "I stay go da store for buy da shirt" the speaker further presupposes

PIDGIN	HAWAIIAN CREOLE ENGLISH
Building—high place—wall part—time —now-time —and then —now temperature every time give you.	Get one [There is an] electric sign high up on da wall of da building show you what time an' temperature get [it is] right now.
Now days, ah, house, ah, inside, washi clothes machine get, no? Before time, ah, no more, see? And then pipe no more, water pipe no more.	Those days bin get [there were] no more washing machine, no more pipe water like get [there is] inside house nowadays, ah?
No, the men, ah—pau [finished] work—they go, make garden. Plant this, ah, cabbage, like that. Plant potato, like that. And then—all that one—all right, sit down. Make lilly bit story.	When work pau [is finished] da guys they stay go make [are going to make] garden for plant potato an' cabbage an' after little while they go sit down talk story ["shoot the breeze"].
Good, this one. Kaukau [food] any kind this one. Pilipin island no good. No more money.	Hawaii more better than Philippines, over here get [there is] plenty kaukau [food], over there no can, bra [brother], you no more money for buy kaukau [food], 'a'swhy [that's why].

PIDGIN AND CREOLE versions of identical sentences illustrate the structural differences between pidgin and Creole in Hawaii. Pidgin, which is spoken only by immigrants, varies widely from speaker to speaker. Although one can probably say anything in pidgin that can be said in English or Creole, the structure of pidgin is extremely rudimentary. Pidgin sentences are little more than strings of nouns, verbs and adjectives, often arranged to place old, shared information first and new information later in the sentence. Creole arose in Hawaii only among the children of immigrants, and it is much richer in grammatical structure than pidgin. Moreover, the rules of Creole grammar are uniform from speaker to speaker, and they resemble the structural rules of other creoles. English versions of words and phrases are given in brackets.

ENGLISH	HAWAIIAN CREOLE ENGLISH
The two of us had a hard time raising dogs.	Us two bin get hard time raising dog.
John and his friends are stealing the food.	John-them stay cockroach the kaukau.
He doesn't want to play because he's lazy.	He lazy, 'a'swhy he no like play.
How do you expect to finish your house?	How you expect for make pau you house?
It would have been better if I'd gone to Honolulu to buy it.	More better I bin go Honolulu for buy om.
The one who falls first is the loser.	Who go down first is loser.
The man who was going to lay the vinyl had quoted me a price.	The guy gon' lay the vinyl bin quote me price.
There was a woman who had three daughters.	Bin get one wahine she get three daughter.
She can't go because she hasn't any money.	She no can go, she no more money, 'a'swhy.

STRUCTURAL DIFFERENCES between sentences in Hawaiian Creole and their English equivalents show that the grammar of Creole did not originate as a grammar borrowed from English. For example, the past perfect tense of a verb in Creole is expressed by the particles "bin" or "wen," which precede the main verb, instead of by the suffix "-ed." Nonpunctual, or progressive, aspect is expressed by the word "stay" instead of by the suffix "-ing." In the English sentence "The two of us had a hard time raising dogs" the rules of grammar oblige the speaker to indicate that the noun "dog" is either singular or plural. In the Creole version of the sentence, however, neither singular nor plural is implied. There are also relatively insignificant lexical differences between the two languages: "cockroach" is picturesquely employed as a verb, and "kaukau," which may be derived from the Chinese pidgin term "chowchow," is a common word for "food." Equally striking structural differences are found between Hawaiian Creole and other languages, such as Chinese, Hawaiian, Japanese, Korean, Portuguese, Spanish or the Philippine languages, with which speakers of Hawaiian Creole might have been in contact.

that the listener is already familiar with the shirt the speaker is going to buy.

There are many other features of Creole that distinguish it from English. Whereas in English there is a past tense, which is usually marked with the suffix "-ed," in Creole there is a tense called the anterior tense, which is marked with "bin" for older speakers and with "wen" for younger speakers. The anterior tense is somewhat like the English past perfect: "had walked" in English is "bin walk" in Creole, and "walked" in English is simply "walk" in Creole. In order to distinguish irreal, or possible, actions or processes from actual ones, English employs the conditional or the future tense. In Creole all such irreal circumstances are expressed by the particle "go," which is placed before the main verb and marks what linguists call modality. For example, the sentence "If I had a car, I would drive home" is rendered in Creole as "If I bin get car, I go drive home."

There is also a Creole auxiliary verb that marks what linguists call aspect; it too is placed before the main verb and indicates that the action expressed by the verb is nonpunctual, or in other words repeated, habitual, continuing or incomplete. In order to say "I run in Kapiolani Park every evening" in Creole one must say "I stay run in Kapiolani Park every evening." If the particle "stay" is omitted by the Creole speaker, the action is understood to be completed or nonrepetitive.

In English there is no straightforward way to distinguish purposes that have been accomplished from those that have not. The sentence "John went to Honolulu to see Mary" does not specify whether or not John actually saw Mary. In Creole grammar the ambiguity must be resolved. If John saw Mary and the Creole speaker knows that John saw Mary, the speaker must say, "John bin go Honolulu go see Mary." If John did not see Mary or if the speaker does not know whether or not John saw Mary, the speaker must say, "John bin go Honolulu for see Mary."

Similar distinctions could be drawn between the grammatical structure of Creole and the structure of other contact languages, such as Hawaiian, Ilocano (the language spoken in the north of the Philippine island of Luzon) and Japanese. There are also resemblances, but most of them are confined to idiomatic expressions. For example, the Creole expression "O the pretty," which means "How pretty he [she/it] is," is a literal translation of the Hawaiian-language idiom "O ka nani." In the main, however, our investigations strongly suggest that the basic structures of Creole differ from those of other languages. Although it might seem that some children of immigrants could transfer the structures of their parents' native languages onto the evolving Creole language, they did not do so. The structural linguistic input that was available to the children was apparently not used in the development of Creole.

Even if it could be demonstrated that all the grammatical structures of Creole were borrowed, cafeteria-style, from one contact language or another, the uniformity of Creole would present a difficult question: How did the speakers who invented Creole come to agree on which structure to borrow from which language? Without such agreement Creole could not be as uniform as it is. Yet it seems highly implausible that the agreement could have been reached so quickly. If there had been massive borrowing from ancestral languages, differences in the version of Creole spoken by various groups would have persisted at least one generation beyond the first generation of speakers.

There is another dimension to the problem of the uniformity of Hawaiian Creole. It turns out that creole languages throughout the world exhibit the same uniformity and even the same grammatical structures that are observed in Hawaii. The finding is all the more remarkable when it is compared with the rather poor correspondence in structure I have noted between Hawaiian Creole and other contact languages in Hawaii. For example, the distinction made in Hawaiian Creole between singular, plural and neutral number is also made in all other creole languages. Similarly, in all other creole languages there are three invariant particles that act as auxiliary verbs and play the roles that "bin," "go" and "stay" play in Hawaiian Creole.

In Haitian Creole, for example, the word "té" marks the anterior tense of the verb, the word "av(a)" marks irreal modality and the word "ap" marks the aspect of the verb as nonpunctual. Thus in Haitian Creole the phrase "I have been walking" is rendered "m [I] t'ap [té + ap] maché." Similarly, in Sranan, an English-based creole found in Surinam (formerly Netherlands Guiana), the anterior tense marker is "ben," the irreal modality marker is "sa" and the nonpunctual aspect marker is "e." The phrase "He would have been walking" is rendered "A [he] ben sa e waka." Most important, there is a strict order that must be followed in all creole languages when more than one of these markers is present in a sentence. The particle for tense precedes the particle for modality, and the particle for modality precedes the particle for aspect.

Finally, consider the grammatical distinction I have noted between purposes accomplished and unaccomplished. The same distinction, absent in English, is found in all creoles. In Mauritian Creole, a creole based on the French vocabulary that is used on the island of Mauritius, a sentence such as "He decided to eat meat" can be expressed in two ways. If the subject of the sentence carried out his decision, the sentence is rendered "Li ti desid al mâz lavian," which literally means "He decided go eat meat." If the decision was not carried out, the sentence is rendered as "Li ti desid pu mâz lavian," or literally "He decided for eat meat." In Jamaican Creole the sentence "He went to wash" must be rendered either as "Im gaan fi bied" ("He went with the intention of washing") or as "Im gaan go bied" ("He went to wash and completed the task").

These examples only suggest the extent of the structural similarities among creole languages. The similarities seem unaffected by the wide geographic dispersion of the creoles and the variation among the languages such as Dutch, English and French from which they draw the greatest part of their vocabulary. Scholars such as Hugo Schuchardt began to point out the resemblance in the 19th century, and in the 1960's many examples were explored in detail by Douglas Taylor, by Robert Wallace Thompson of the University of the West Indies and by Keith Whinnom of the University of Exeter. Thus even before the development of Hawaiian Creole was reasonably well understood the grammatical similarities among the creole languages of the world were recognized as an important finding that required explanation.

The linguist's first reaction to such a finding is to look for a common ancestor of the similar languages. For ex-

VERB FORM
BASE FORM ("HE WALKED"; "HE LOVES")
ANTERIOR ("HE HAD WALKED"; "HE LOVED")
IRREAL ("HE WILL/WOULD WALK"; "HE WILL/WOULD LO
NONPUNCTUAL ("HE IS/WAS WALKING")
ANTERIOR + IRREAL ("HE WOULD HAVE WALKE "HE WOULD HAVE LOVED")
ANTERIOR + NONPUNCTUAL ("HE WAS/HAD BEEN WALKING")
IRREAL + NONPUNCTUAL ("HE WILL/WOULD BE WALKING")
ANTERIOR + IRREAL + NONPUNCTUAL ("HE WOULD HAVE BEEN WALKING")

CONJUGATION OF THE VERB is similar in all creole languages, in spite of superficial lexical differences. Moreover, the creole system is quite distinct from the one encountered in English and in most other languages. The table gives conjugations in Hawaiian Creole, Haitian Creole and Sranan (an English-based creole spoken in Surinam, the former Nether-

ample, it has been conjectured that the linguistic ancestor was a contact language that grew out of Portuguese and certain West African languages in the course of the first Portuguese explorations of Africa in the 15th and 16th centuries. According to the hypothesis, this contact language was subsequently spread around the world by Portuguese sailors, changing its vocabulary but not its syntax or semantics as it entered the sphere of influence of another colonial power. Superficially such an explanation might seem to be consistent with the development of Creole in Hawaii, because Portuguese laborers were brought to the islands in large numbers during the late 19th and early 20th centuries.

There are several serious flaws in the account. First, Hawaiian Creole bears scant resemblance to any of the contact languages, including Portuguese. Second, the claims of linguistic similarity between creoles and Portuguese or between creoles and West African languages are grossly exaggerated. Most important, our study of hundreds of Hawaiian speakers has made it clear that Hawaiian Creole almost certainly originated in Hawaii. We found no surviving immigrant who speaks anything approximating a creole language; instead every immigrant we surveyed speaks some variety of pidgin. If Hawaiian Creole was primarily an imported language, it would have been carried by immigrants, and presumably it would have been learned by others among the immigrant population. One must therefore conclude that Hawaiian Creole arose among the children of immigrants, where it is now found. Moreover, if a creole language could develop in Hawaii without ancestry, it can arise anywhere else in a similar way.

The implications of these findings are far-reaching. Because the grammatical structures of creole languages are more similar to one another than they are to the structures of any other language, it is reasonable to suppose most if not all creoles were invented by the children of pidgin-speaking immigrants. Moreover, since creoles must have been invented in isolation, it is likely that some general ability, common to all people, is responsible for the linguistic similarities.

The suggestion that people are biologically predisposed to use language is not a new one: for more than two decades Noam Chomsky of the Massachusetts Institute of Technology has argued that there is an innate universal grammar underlying all human languages. The universal grammar is postulated largely on the grounds that only by its means could children acquire a system as enormously complex as a human language in the short time they do. Studies by the late Eric H. Lenneberg tend to confirm Chomsky's hypothesis. The acquisition of language resembles the acquisition of other complex and flexible aspects of the child's behavior, such as walking, which are undoubtedly controlled to some degree by neurophysiological development. The universal grammar conjectured by Chomsky is a computing device, somehow realized neurologically, that makes a wide range of grammatical models available to the child. According to Chomsky, the child must then "select" which of the available grammatical models matches the grammar of the language into which the child is born.

The evidence from creole languages suggests that first-language acquisition is mediated by an innate device of a rather different kind. Instead of making a range of grammatical models available, the device provides the child with a single and fairly specific grammatical model. It was only in pidgin-speaking communities, where there was no grammatical model that could compete with the child's innate grammar, that the innate grammatical model was not eventually suppressed. The innate grammar was then clothed in whatever vocabulary was locally available and gave rise to the creole languages heard today.

The implications of this hypothesis call into question an idea that most linguists, including Chomsky, have tacitly accepted for many years, namely that no one of the world's languages is easier or harder for the child to acquire than any other. If there is a creole grammar somehow imprinted in the mind, creole languages should be easier to acquire than other languages. How is it, then, that not all children grow up speaking a creole language? The answer is they do their best to do just that. People around them, however, persist in speaking English or French or some other language, and so the child must modify the grammar of the native creole until it conforms to that of the local language.

Two kinds of linguistic evidence are relevant for testing the hypothesis. First, if some grammatical structure of creole is at variance with the corresponding grammatical structure of the local language, one should find that children

NONSTATIVE VERBS			STATIVE VERBS		
HAWAIIAN CREOLE	HAITIAN CREOLE	SRANAN	HAWAIIAN CREOLE	HAITIAN CREOLE	SRANAN
E WALK	LI MACHÉ	A WAKA	HE LOVE	LI RÊMÊ	A LOBI
E BIN WALK	LI TÉ MACHÉ	A BEN WAKA	HE BIN LOVE	LI TÉ RÊMÊ	A BEN LOBI
E GO WALK	L'AV(A) MACHÉ	A SA WAKA	HE GO LOVE	L'AV(A) RÊMÊ	A SA LOBI
E STAY WALK	L'AP MACHÉ	A E WAKA	—	—	—
E BIN GO WALK	LI T'AV(A) MACHÉ	A BEN SA WAKA	HE BIN GO LOVE	LI T'AV(A) RÊMÊ	A BEN SA LOBI
E BIN STAY WALK	LI T'AP MACHÉ	A BEN E WAKA	—	—	—
E GO STAY WALK	L'AV AP MACHÉ	A SA E WAKA	—	—	—
E BIN GO STAY WALK	LI T'AV AP MACHÉ	A BEN SA E WAKA	—	—	—

lands Guiana) for stative and nonstative verbs. Stative verbs are verbs such as "like," "want" and "love," which cannot form the nonpunctual aspect; in English, for example, one cannot add "-ing" to a finite stative verb. The base form of the verb refers to the present for stative verbs and to the past for nonstative verbs. The anterior tense is roughly equivalent to the English past tense for stative verbs and to the English past perfect tense for nonstative verbs. The irreal mode includes the English future, conditional and subjunctive. In all the creole languages the anterior particle precedes the irreal particle, and the irreal particle precedes the nonpunctual particle. In Hawaiian Creole, however, "He bin go walk" has come to mean "He walked" instead of "He would have walked," and the forms "He bin stay walk," "He go stay walk" and "He bin go stay walk," although they were widespread before World War II, are now almost extinct because of the growing influence of English in Hawaii. The bracketed English translations are provided only as a rough guide to the meaning.

CHILD LANGUAGE	ENGLISH CREOLES
Where I can put it?	Where I can put om? (Hawaii)
Daddy throw the nother rock.	Daddy t'row one neda rock'tone. (Jamaica)
I go full Angela bucket.	I go full Angela bucket. (Guyana)
Lookit a boy play ball.	Luku one boy a play ball. (Jamaica)
Nobody don't like me.	Nobody no like me. (Guyana)
I no like do that.	I no like do that. (Hawaii)
Johnny big more than me.	Johnny big more than me. (Jamaica)
Let Daddy get pen write it.	Make Daddy get pen write am. (Guyana)
I more better than Johnny.	I more better than Johnny. (Hawaii)

SENTENCES SPOKEN BY CHILDREN between two and four years old, all born of English-speaking parents, are strikingly similar to sentences in English-based creole languages. The similarities among creole languages and the likelihood that the languages arose independently of one another suggest that creoles develop among children whenever there is no adequate native language to serve as a model. The author conjectures that if children were removed from their native English-language community at the age of about two, they would grow up speaking a language whose vocabulary would be primarily English but whose grammar would be a creole.

make systematic errors with respect to the structure of the local language. On the other hand, if the two grammatical structures tend to agree, one should find extremely early, rapid and errorless acquisition of the local-language structure.

Consider the systematic error observed by David McNeill of the University of Michigan in the speech of a four-year-old boy. In one of McNeill's observing sessions the boy complained, "Nobody don't like me," and the boy's mother responded by correcting the sentence: "Nobody likes me." The boy then repeated his sentence and the mother repeated her correction no fewer than eight times. Finally, the child altered his sentence and shouted in exasperation, "Nobody don't likes me."

The error is found in many English-speaking children between three and a half and four years old, including children who are not exposed to dialects of English that employ double negatives. There are many languages, such as French and Spanish, that also employ double negatives, but the only languages that allow negative subjects with negative verbs are creoles. For example, in Papia Kristang, the Portuguese-based creole language of the Malay Peninsula, one can say, "Angkosa nte mersimentu," or literally, "Nothing not-have value." In Guyanese Creole, which is based on English and found in Guyana (formerly British Guiana), one can say, "Non dag na bait non kyat," or literally, "No dog did not bite no cat."

A second instance of systematic error is found in the formation of children's questions. Children learning English often indicate questions only by their intonation; the subject and the auxiliary verb are almost never reversed. For example, children repeatedly say things such as "You can fix this?" even though they have heard countless questions such as "Can you fix this?" Similarly, no creole language distinguishes questions and statements on the basis of word order; the difference is marked by intonation alone.

Consider the sentence "A gon' full Angela bucket." Although such a sentence is unacceptable in English, it is perfectly acceptable in Hawaiian Creole, Guyanese Creole or any of several other creoles related to English. It is synonymous with the sentence "I'm going to fill Angela's bucket," but it differs from the structure of the English sentence in the following ways. First, the first-person pronoun "I" is reduced to "A"; second, the auxiliary verb "am" is omitted; third, the forms "go" or "gon" are used to mark the future tense; fourth, the word "to" in the infinitive is omitted; fifth, the adjective "full" is employed as if it were a transitive verb, and sixth, the possessive marker "-'s" is omitted. All these features are characteristics of creoles, but this sentence was not uttered by a creole speaker. It was spoken by the three-year-old daughter of an English-speaking linguist.

When a feature of the local language matches the structure of creole, children avoid making errors that would otherwise seem quite natural. For example, children learning English acquire the suffix "-ing," which expresses duration, at a very early age. Even before the age of two many children say things such as "I sitting high chair," where the verb expresses a continuing action. One would expect that as soon as the suffix was acquired it would be applied to every possible verb, just as the suffix "-s" that marks the English plural is frequently overgeneralized to nouns such as "foot" and "sheep."

One would therefore expect children to utter ungrammatical sentences such as "I liking Mommy" and "I wanting candy." Remarkably, such errors are almost never heard. Children seem to know implicitly that English verbs such as "like" and "want," which are called stative verbs, cannot be marked by the suffix "-ing" to indicate duration. The distinction between stative and nonstative verbs is fundamental to creole languages, however, and no marker of continuing action can be employed with a stative verb in creoles either.

The distinction between specific and nonspecific reference, which I have already discussed, is an important feature of creole languages. In English the distinction can be subtle, but young children nonetheless acquire it with ease. Michael P. Maratsos of the University of Minnesota constructed a series of sentences for children to complete, for which the completions depended on the distinction between specific and nonspecific reference. For example, the sentence "John has never read a book," which makes nonspecific reference to the noun "book," can be completed by the phrase "and he never will read a book"; it cannot be completed by the phrase "and he never will read the book." Similarly, the sentence "John read a book yesterday," in which a specific book is presupposed, can be completed by the phrase "and he enjoyed the book"; it cannot be completed by the phrase "and he enjoyed a book." Children as young as three years were able to make such distinctions correctly about 90 percent of the time.

Many more studies of language acquisition will have to be carried out before the structure of creole languages can be firmly accepted as the basis of first-language acquisition. Daniel Isaac Slobin of the University of California at Berkeley has suggested that there is a set of processes children apply to any language they hear, which he calls basic child grammar. Slobin's most recent work, which is not yet published, cites evidence from several languages for the hypothesis, and it now appears that basic child grammar and creole languages may have much in common.

If creole languages represent the manifestation of a neurologically determined program of child development, then Psamtik was by no means the fool he has been taken for. It may be possible to discover, at least in general outline, the structure of human language in the early stages of its development. Moreover, in attempting to reconstruct such a language linguists may be able to answer questions the pharaoh did not even ask: How did human language originate? What are the minimum prerequisites for such a thing as language to arise in a species? If such questions can be answered or even formulated in a precise and coherent way, we shall be much closer to understanding what makes the human species different from others.

The IBM Personal Computer. More bang for your buck.

Now you can have the best prize of all: IBM quality at a new price* that's better than ever. The system includes 64KB of user memory, one 160KB diskette drive plus a color/graphics adapter.

However you plan to use your IBM Personal Computer, you'll win big—because it can help suit your needs in the office, home or school.

Growth options? Certainly. Add a color monitor if you'd like. Or a printer. Or more memory. And you can choose from IBM's growing library of personal computer software to make your system a truly useful, valuable and versatile tool.

To get all the news about this important IBM price/performance story, ring the bell at your authorized retail dealer.

IBM

*This price applies to IBM Product Centers. Prices may vary at other stores.
For more information on where to buy the IBM Personal Computer, call 800-447-4700. In Alaska or Hawaii, 800-447-0890.

THE AMATEUR SCIENTIST

The physics of the follow, the draw and the massé (in billiards and pool)

by Jearl Walker

Billiards and pool have the feel of physics. Balls collide with each other and the rails of the table like gas molecules in some two-dimensional tank. Actually the physics of billiards and pool is subtler. For example, a skilled player can impart spin to the ball, achieving such effects as the follow, the draw and the massé. Indeed, the interaction of the cue and the ball may be the most challenging application of classical mechanics. To master the forces and trajectories of both billiards and pool one must play the game often and analytically. A helpful step toward that objective is an understanding of the physics of the games.

Recently Todd King of Temple City, Calif., sent me his analysis of some of the classic shots in pool. Until a few years ago almost the only study concerning the dynamics of billiards was in lecture notes by Arnold Sommerfeld, who is better known for his work on early quantum mechanics. Last year David F. Griffing of Miami University devoted a chapter to pool and billiards as part of his book *The Dynamics of Sports: Why That's the Way the Ball Bounces.* These three sources inform the following discussion of the physics of billiards and pool. After describing some simple relations I shall take up a few of the famous trick shots outlined in *Byrne's Treasury of Trick Shots in Pool and Billiards,* by Robert Byrne.

When the cue strikes the cue ball, both horizontal and rotational motions are imparted. For simplicity assume that the cue stick is horizontal and delivers only a horizontal force. Although the force can be applied anywhere on the ball's surface facing the player, the shock sets the ball in horizontal motion just as if the force were applied at the center of mass.

Now assume that the stroke by the cue is in the vertical plane passing through the ball's center of mass, namely on an imaginary vertical line running through the center of the face toward the player. The location of the blow along this line has no direct bearing on either the ball's initial velocity or its momentum (the product of mass and velocity). They are set by two other factors in the collision. One factor, over which the player has virtually no control, is the duration of the collision. The second factor, easily controlled by the player, is the force on the ball. A "hard" shot generates more velocity and momentum than a "soft" shot because the force in the collision is greater.

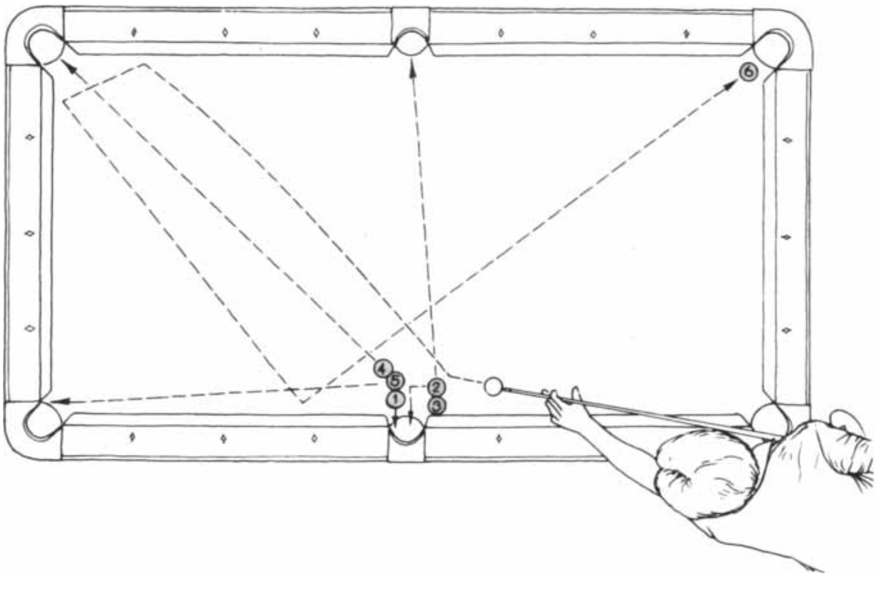
In addition to horizontal motion the cue also generates a torque that makes a ball rotate about its center of mass. The magnitude of the torque is equal to the product of the force and a lever arm that represents the vertical distance between the middle of the ball and the point where the cue strikes. The torque increases as the distance of the blow from the middle of the ball increases.

The torque determines the initial rate at which the ball spins about the center of mass. The spin is proportional to the torque divided by the ball's moment of inertia (a number that takes into account not only the mass of an object but also the distribution of the mass around the axis of rotation). For a cue ball rotating about an axis through its center of mass the moment of inertia is two-fifths of the mass multiplied by the square of the radius. The factor of 2/5, which arises from the shape of the ball, plays a role in a player's decision about where to stroke the ball in certain shots.

If the player wants an initially nonspinning ball, he should strike it at the height of its center of mass. The lever arm for such a stroke is zero, and so the torque and spin are zero. With a higher blow the collision has a lever arm and hence a measurable torque. The ball moves forward because of the force of the collision, and it spins about its center of mass because of the torque. The ball has topspin: the top of the ball moves away from the player faster than it otherwise would. Striking the ball below the center results in backspin.

The player's stroke therefore controls three features of the motion. The force determines the velocity of the ball over the table. The lever arm of the force determines the direction of spin. The product of the force and the lever arm determines the rate of spin.

Without friction from the surface of the table the cue ball would continue moving until it hit a rail or another ball. Even a surface worn smooth from play can provide significant friction, however, if the ball slips on the cloth. The friction can be high enough to alter both the horizontal and the rotational motion of the ball and thereby significantly change the shot. If the ball rolls over the table without slipping, the friction is low and affects little more than the maximum distance of roll.



The "just showin' off" shot

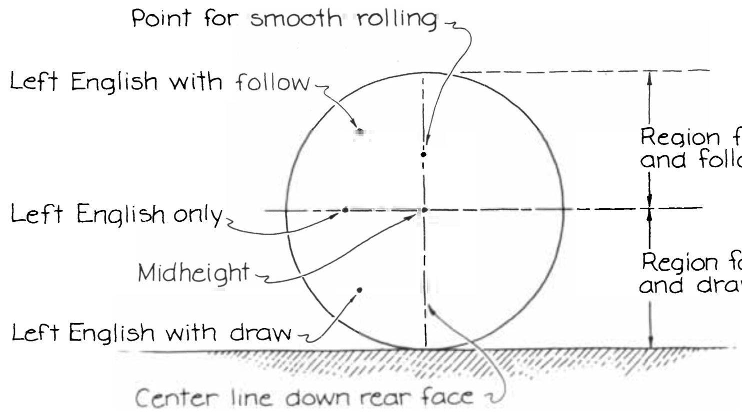
The friction on a slipping ball depends on the weight of the ball and on the surface texture of the cloth and the ball, but it is independent of the rate at which the ball slips. The direction of the force depends entirely on the direction of slip. Suppose the player delivers a large topspin to the cue ball; the bottom surface then slides toward him and the center of mass moves away. At the point of contact the friction force is opposite to the sliding. (The force is away from the player.) Since the friction opposes the sliding, it begins to decrease the spin of the ball about its center of mass. And since the friction force is away from the player, it continues to propel the ball forward and away from the player. A ball given a large topspin can run for a long time because of this additional propulsion.

Suppose the player imparts backspin. The force sends the ball away and the torque sets it into a spin that makes the bottom slide over the cloth in the same direction. The friction force on the bottom on the ball is thus toward the player. Again the friction tends to slow the spin, but this time its effect on the center of mass is rearward. As the ball slides, its forward motion and spin are slowed. Eventually the backspin is eliminated and the ball begins to roll without sliding. A ball hit with backspin runs only a short distance because the friction opposes the motion of the center of mass.

A ball with topspin will slip unless the speed of its center of mass is equal to the spin multiplied by the radius. Then the ball's forward motion is exactly matched by the motion of the bottom surface through the point of contact with the table. To achieve this match instantly the player must stroke the ball at a point that is exactly above the middle by a distance equal to two-fifths of the radius. This relation is established by the fact that the 2/5 in the formula for the moment of inertia must be canceled by the 2/5 in the formula for the lever arm of the torque.

If the cue ball is struck higher than at this special point, the spin rate is at first too large for a match to be achieved, but friction tends to force the ball into the matched state. The friction (directed forward) reduces the spin and increases the speed of the center of mass. If the ball does not run into anything, the match of speeds is eventually made and the ball begins to roll smoothly.

If the ball is struck between the middle and the special point, its spin is in the right direction for smooth rolling but is too small. At the point of contact with the cloth the ball's surface has a small rearward motion owing to the spin and a larger forward motion owing to the speed of the center of mass. The net slipping motion there is forward, creating a rearward friction force that tends to increase the spin and decrease the motion of the center of mass until the match that causes a smooth roll is achieved.



Strategic places to strike the cue ball

If the ball is struck below the middle, the spin is in the wrong direction for smooth rolling. This time the friction induced by slipping reduces both the spin and the speed of the center of mass. Eventually the spin stops and the ball begins to roll smoothly.

A skilled player can impart a long or short run to a ball by striking it at a point relative to the special point for smooth rolling. If he wants the ball to reach the far side of the table quickly, he must strike it above the special point so that the friction propels the ball.

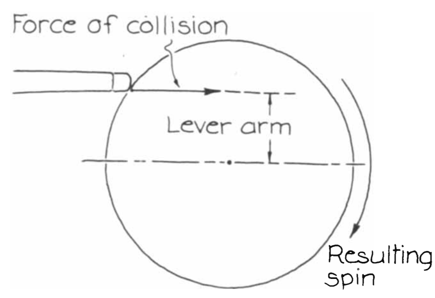
The player is more likely to be concerned with how the ball is rotating when it strikes another ball. (The other balls are called the object balls.) A collision between a cue ball and an object ball transfers momentum from the cue ball. In a head-on collision the transfer is complete, leaving the cue ball with a motionless center of mass. In a glancing collision the cue ball loses only part of its momentum and continues to travel. In any collision virtually none of the cue ball's rotation is transferred because the friction between the surfaces of the colliding balls is minute and the collision is brief. Only with significant friction could the cue ball transfer spin to an object ball.

The absence of transferred spin leads to two interesting shots. Suppose the cue ball is hit with topspin and collides head on with an object ball while it is still sliding. Just after the collision the cue ball's center of mass is motionless but the ball continues to spin. The forward-directed friction generated by the spin slows the spin and begins to propel the center of mass. Soon the cue ball begins to roll again, following the object ball. This is a follow shot. A cue ball with topspin is often said to have "follow" or "follow English."

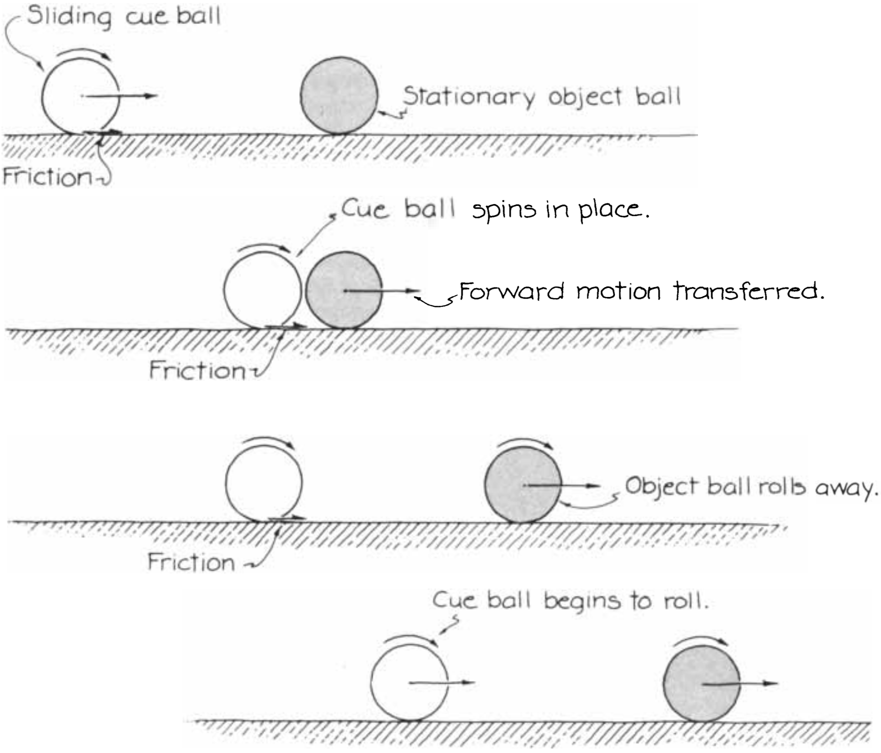
If the cue ball is given backspin, it will return to the player after hitting an object ball head on. The collision leaves the cue ball with a motionless center of mass but with the same amount of spin. The friction generated by the sliding surface is toward the player. As the friction slows the spin and propels the center of mass, the ball begins to roll smoothly toward the player. This is a draw shot. A ball with backspin is often said to have "draw" or "draw English."

A follow shot is depicted in the middle illustration on the next page. A player who wants to pocket the four ball and the seven ball with a single shot strikes the cue ball with follow, thus causing the four ball to ricochet off the seven ball and into the pocket. The collision of the cue ball with the four ball leaves the cue ball momentarily spinning in place, but sliding friction soon propels it toward the pocket again. In the meantime the seven ball has bounced off the rail near the pocket and come into line between the cue ball and the pocket. The cue ball then pockets the seven ball and comes to a stop.

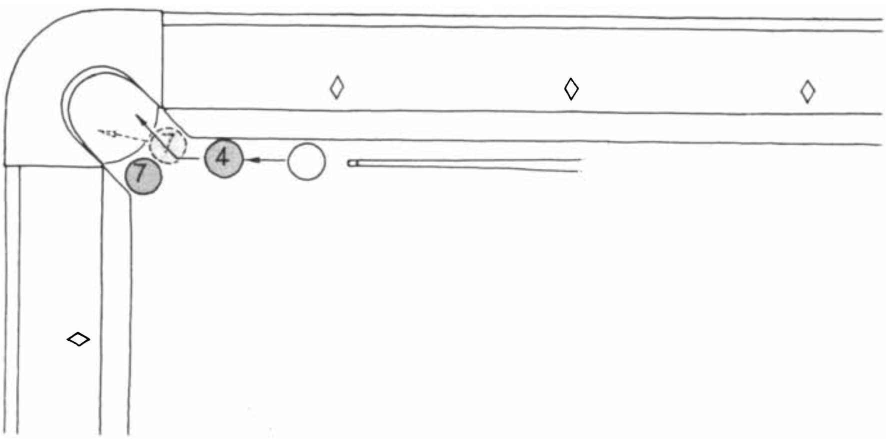
So far I have written only of stroking the cue ball along a vertical line through the center. The result is topspin or backspin about a horizontal axis. If the ball is stroked elsewhere, the axis of spin still passes through the center of mass but is no longer horizontal. A stroke on the left side of the ball is said to give left English, on the right side right English. From overhead left English is a clockwise spin about the vertical. As before, the rate of spin depends on the lever arm associated with the force. The farther



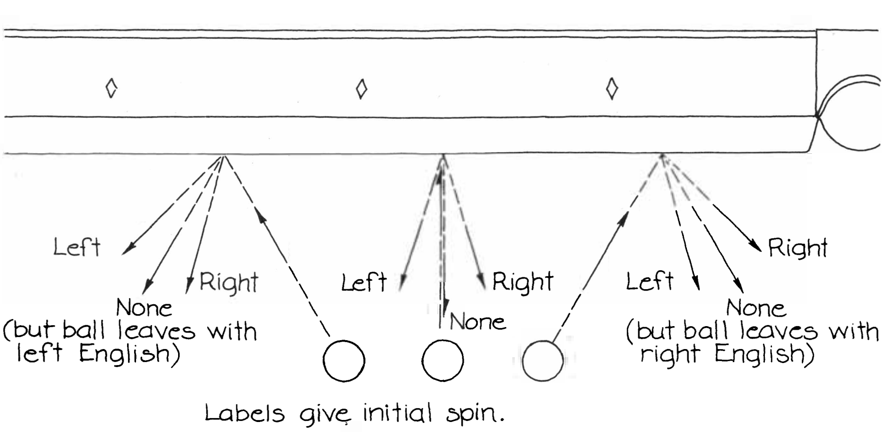
Factors in a topspin shot



Four stages in a follow shot



A follow shot demonstrated



Effects of putting English on the cue ball

off center the ball is struck, the larger the lever arm is and the faster the spin around the vertical is. The friction of the ball with the table serves only to diminish the spin.

If the cue strikes the ball high or low on the side, the ball spins about an axis that is between the horizontal and the vertical. A stroke below the middle and to the left of center results in a draw with left English. The center of mass is given momentum in the usual way and the ball spins about an axis tilted out of the vertical toward the player's left. One can view this rotation as being two simultaneous spins, one that is clockwise about the vertical and another that is backspin about the horizontal. The primary friction force on the ball is imparted by the backspin.

A cue ball with side English travels in a straight line like a ball with no English, but the angle at which it rebounds from a rail is remarkably different. When the ball has no initial English, its angle of rebound is the same as its angle of approach to the rail. If the approach to the rail is perpendicular, as is shown in the bottom illustration at the left, the ball must retrace its path after rebounding. If it has left English, however, it returns along a path on the player's left because of friction during contact with the rail. From an overhead view the ball with left English turns clockwise. When it slides along the rail, the friction force is to the left. When the ball rebounds, it has not only its velocity perpendicular to the rail but also this leftward component. The ball returns along a straight path resulting from the combination of the two motions.

A left-English ball approaching the rail at any other angle is similarly redirected. The opposite effects arise with a right-English ball. A simple way to remember the difference is to associate the direction of English with the rotation of the rebound path: left English results in a clockwise rotation of the path and right English results in a counterclockwise rotation.

If a ball approaches the rail with no English and at some angle other than 90 degrees, friction from the rail gives it English. Consider the ball's approach to be a combination of one motion perpendicular to the rail and another parallel to it. Once contact is made the parallel component generates friction on the ball. The resulting torque spins the ball about the vertical, imparting English. Suppose a ball is sent into the rail from the player's right. If it initially had no spin, it leaves the rail with left English.

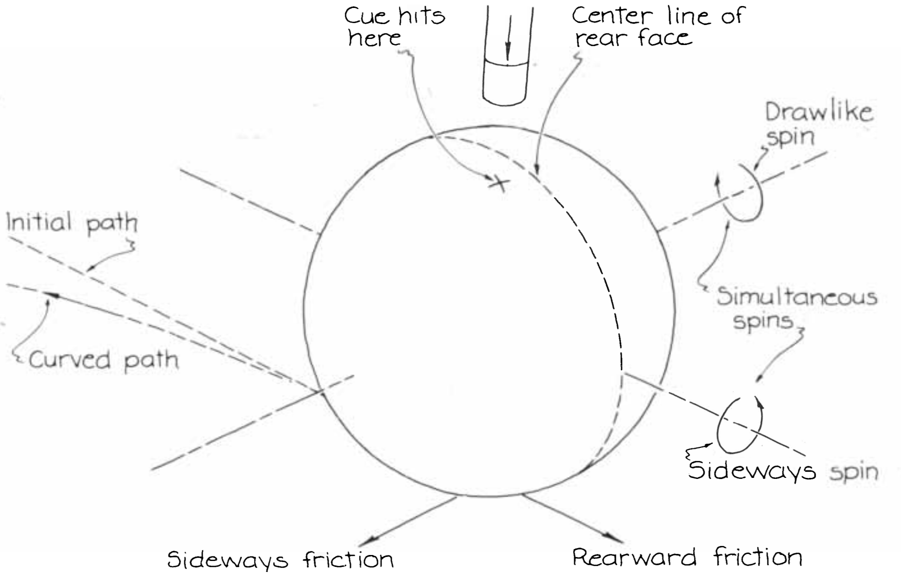
All these rotations of spin are about an axis in a plane perpendicular to the direction of travel of the ball. The massé shot supplies spin about an axis out of that plane. With the cue almost vertical the player strikes downward on the side of the cue ball. The horizontal part of the stroke determines the initial path of the ball, but the spin given to the ball generates a friction from the table that ends up curving the path.

Suppose the player strikes sharply the left side of the ball. Since the blow is hard and the lever arm is large, the spin imparted to the ball is large. The spin is about an axis that is approximately in a horizontal plane but is not perpendicular to the initial path of the ball. To simplify the spin one can regard it as consisting of two simultaneous spins about different axes. One is parallel to the initial path and the other is perpendicular to it. The spin about the axis perpendicular to the initial path is similar to that of a simple draw shot. The spin about the other axis forces the ball to slip perpendicular to the initial path, thereby generating a friction force that is also perpendicular to the path. This sideways force curves the path of the ball.

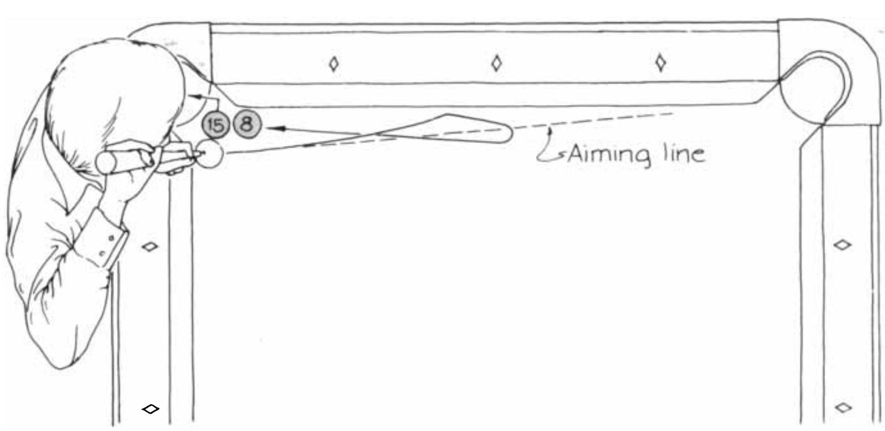
The massé is commonly employed to send the cue ball around an obstacle to reach a hidden ball. A more complex massé shot is shown in the lower illustration at the right. The idea is to sink the 15 ball and the eight ball with a single shot and to have the eight ball enter the pocket last. The cue ball is given a massé shot, knocks in the 15 ball, misses the eight ball and heads toward the rail in a curved path because of sideways friction. After rebounding from the rail the ball stops and then reverses its horizontal motion, heading back to pocket the eight ball.

The initial stroking of the cue ball gives it both backspin for a friction force like the one in the standard draw shot and sideways spin for a sideways friction to curve the path into the rail. The rebound off the rail is little affected by the backspin, but the sideways friction keeps the ball near the rail. After the rebound the backspin finally stops the horizontal motion of the ball's center of mass. Since the ball still spins, that friction force then brings it back toward the player. The remaining sideways friction continues to drive the ball toward the rail. Hence after the reversal of path resulting from the drawlike component of the shot the ball comes back near the rail to knock the eight ball into the corner pocket.

When a cue ball collides with an object ball, part of the momentum and kinetic energy of the center of mass of the cue ball is transferred. The transfers are almost total if the collision is head on. In a glancing collision the transfers cause the two balls to separate along approximately perpendicular paths. (In practice a small amount of energy is lost by the balls in the collision, and the angle between their paths is a bit less than 90 degrees. I shall disregard this complication.)



Factors in a massé shot

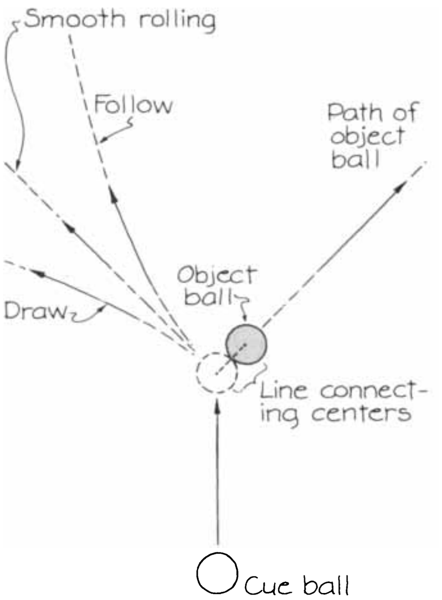


A massé shot demonstrated

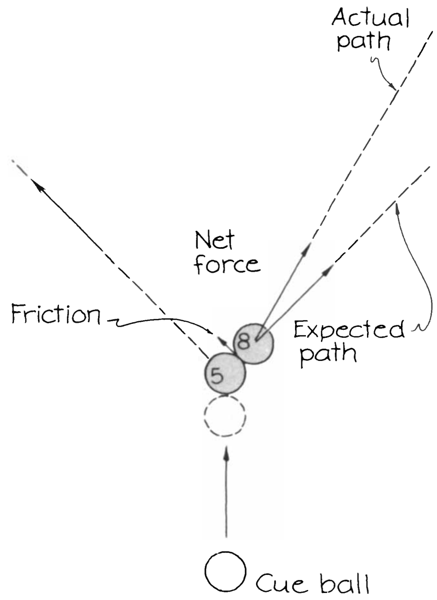
You can easily predict where the cue ball and the object ball will travel after a collision. Imagine the instant the two balls touch and mentally draw a line between their centers. At that instant the object ball acquires two forces from the cue ball. At the contact point and perpendicular to the line between the centers is a small friction force. It is almost always small enough to be ignored. Parallel to the line is a larger force that pushes the object ball off along a path that is also parallel to the line. The direction given to the object ball depends almost entirely on the orientation of the line between the centers of the balls at the instant of contact. Through experience the skilled player can direct the cue ball so that it makes contact in just the way necessary to send an object ball into a pocket. The player can be certain the cue ball will travel perpendicularly to that path.

If the cue ball is given follow or draw and still has the associated sliding when it reaches the object ball, the collision changes somewhat in that the cue ball leaves the collision site on a curved path. Suppose the cue ball has been given a large follow. The collision transfers part of the kinetic energy and momentum of the center of mass. If one can disregard entirely the friction between the colliding balls, none of the spin of the cue ball is transferred to the object ball. The cue ball begins to move away from the collision site along a path perpendicular to the path taken by the object ball. The cue ball still has topspin. The curious feature is that the ball no longer slides parallel to its path. The component of the sliding perpendicular to the path provides a sideways friction force that pushes the ball into a curved path. Therefore when the cue ball is given follow, it tends to curve back toward its original direction after a collision. A cue ball with draw tends to curve away from the original direction.

Normally the friction between two colliding balls is negligible. It can be greatly increased if the surfaces are cov-



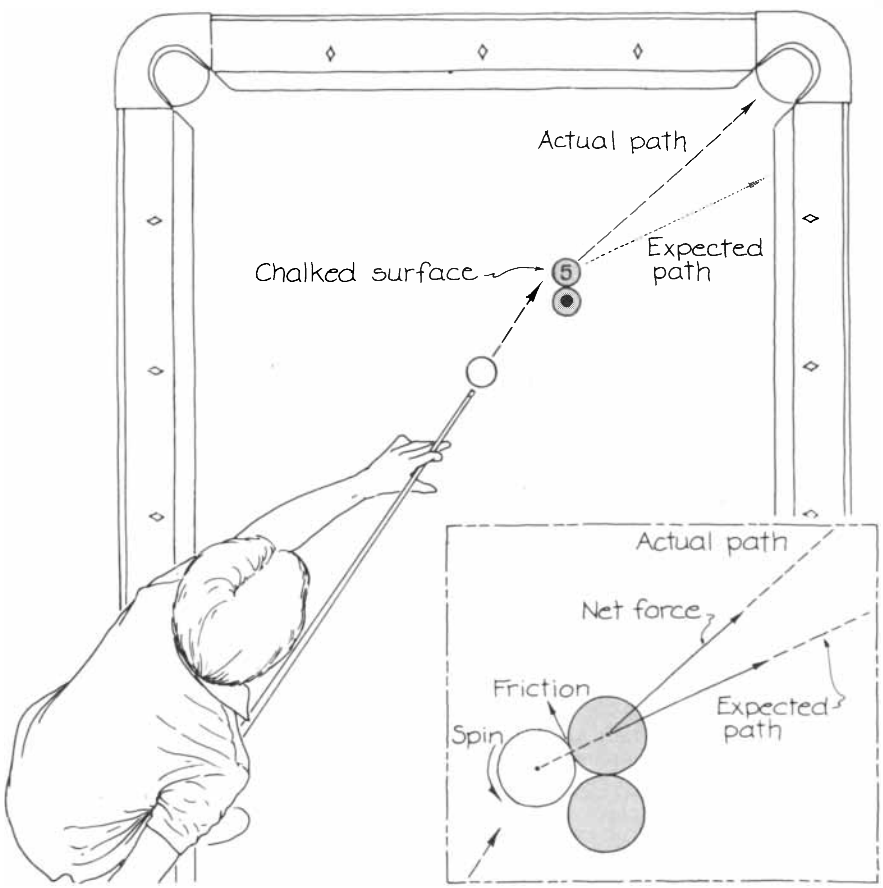
A cue ball hits an object ball



A cue ball hits frozen object balls

ered with chalk. My favorite example comes from Byrne's book. The illustration below depicts the setup: the player must get the five ball into the pocket at the right. Can the shot be made without contact between the cue ball and the spotted ball? Normally the shot is impossible. The five ball can travel to the pocket only if the player has aligned the collision so that the line joining the centers of the cue ball and the five ball points to the pocket. The spotted ball is clearly in the way.

The shot can be made if the player chalks the left side of the five ball and then sends the cue ball into the side with a little right English. With chalk on the collision area the friction between the balls is no longer negligible. The five ball is subjected to two forces during the collision, one force parallel to the line connecting the centers and the other force (friction) perpendicular to that line. The five ball heads off in the direction of the net force, which by design is toward the corner pocket.



Creating friction between two object balls

Chalking a ball is certain to get you thrown out of a pool game, but a similar application of friction between balls can be achieved more acceptably. When a cue ball strikes an object ball that is already touching another object ball (the two object balls are said to be frozen), the collision creates a friction between the object balls that can significantly alter the path of one of them. Consider the situation in the illustration at the top right on this page. The cue ball is sent directly into the five ball, which is frozen to the eight ball.

It is best to analyze the collision in two steps. First the cue ball transfers energy and momentum to the five ball, which then collides with the eight ball. The five and the eight should separate along perpendicular paths, but the eight ends up traveling more in the forward direction because of friction between the object balls.

During the collision between the five ball and the eight ball the five begins to move perpendicular to the line between the centers. The eight ball should move parallel to that line because of the force of collision from the five. Since the balls are initially frozen, however, the motion of the five ball rubs the surface of the eight ball, generating a friction force that briefly pushes the eight ball perpendicular to the anticipated path. The actual path is then set by the combination of these two forces on the eight ball during the collision; the path is more in the forward direction than it would be if the balls were not initially frozen.

My last example, the "just showin' off" shot, has become famous because Steve Mizerak, a master of pool, performed it in a television commercial. Five balls are clustered around a side pocket. The six ball lies at the mouth of a corner pocket. Can all six of the balls be pocketed with a single shot? I certainly cannot make the shot, but Mizerak is said to be successful three times out of four.

The cue ball is sent into the two ball with follow and left English. Imagine the position of the balls and the forces between them at the instant the cue ball reaches the two ball. The two ball has three forces on it. One force is parallel to the line connecting its center with the center of the cue ball. Another is along the line connecting the centers of the two ball and the three ball. Because those balls were initially frozen, the two ball also has a rightward friction force on it at the point of contact when it slides off the three ball. The net force sends the two ball over to the five ball, where it ricochets into the side pocket.

Meanwhile the three ball has been moving. When the cue ball hit the two ball, the three ball received two forces from the two ball. A force parallel to the line connecting their centers knocked the three ball hard into the rail. The second force was leftward friction generated by the rubbing of the balls when the two ball departed toward the left. (This friction arises because the two ball and the three ball were initially frozen.) The rail pushes back on the three ball, directing it straight across the table. The ball also travels somewhat to the left because of its brief friction with the two ball. The three ball ends up in the pocket on the far side.

The five ball was initially frozen to the one ball and the four ball. When the two ball hits it, the five ball is subjected to several forces. One force is along the line connecting its center and the center of the two ball. Two more forces lie along lines connecting the center of the five ball with the centers of the one and the four. In addition the five ball has friction forces from being frozen to the other balls. The net force on the five ball is neatly toward the pocket in the left corner. The net force on the one ball is toward the nearby side pocket. The net force on the four ball is toward the far left pocket. Five balls are down.

After this cluster of balls has spread the cue ball returns to the area. It had been launched with follow (for a long run) and with left English. Its collision with the two ball left it traveling toward the far rail with most of its initial spin. The spin, however, is now somewhat sideways to the path. The cue ball curves to the left, rebounding closer to the corner pocket than it would without the sideways force. Its bounce from the rail removes most of the spin. Thereafter it travels in straight lines, bouncing twice more from rails. It finally reaches the six ball at the other end of the table and pockets it.

I have of course been describing only a limited selection of shots. Thousands of interesting shots remain to be analyzed. You might be particularly interested in figuring out the physics of shots into large clusters of frozen balls. Byrne has several curious examples devised by 19th-century masters of pool and billiards. You might also be interested in jump shots, in which the cue ball is sent hopping over the table or even between two tables. Be careful. Proprietors of pool halls rarely tolerate such shenanigans, even in the interest of scientific investigation.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

THE SLOWING OF URBANIZATION IN THE U.S.

NEW DIRECTIONS IN URBAN-RURAL MIGRATION: THE POPULATION TURNAROUND IN RURAL AMERICA. Edited by David L. Brown and John M. Wardell. Academic Press, 1980.

NONMETROPOLITAN AMERICA IN TRANSITION. Edited by Amos H. Hawley and Sara Mills Mazie. University of North Carolina Press, 1981.

POPULATION DECONCENTRATION IN THE UNITED STATES. John F. Long. U.S. Department of Commerce, U.S. Bureau of the Census, U.S. Government Printing Office, 1981.

HOW CONTINENTS BREAK UP

PROPAGATION OF AN ACCRETING PLATE BOUNDARY: A DISCUSSION OF NEW AEROMAGNETIC DATA IN THE GULF OF TADJURAH AND SOUTHERN AFAR. V. Courtillot, A. Galdeano and J. L. Le Mouël in *Earth and Planetary Science Letters,* Vol. 47, No. 1, pages 144–160; March, 1980.

PROPAGATING RIFTS ON MIDOCEAN RIDGES. Richard Hey, Frederick K. Duennebier and W. Jason Morgan in *Journal of Geophysical Research,* Vol. 85, No. 37, pages 3647–3658; July 10, 1980.

PROPAGATING RIFTS IN SEAFLOOR SPREADING PATTERNS. Tanya Atwater in *Nature,* Vol. 290, No. 5803, pages 185–186; March 19, 1981.

PROPAGATING RIFT EXPLANATION FOR THE TECTONIC EVOLUTION OF THE NORTHEAST PACIFIC: THE PSEUDO-MOVIE. R. N. Hey and D. S. Wilson in *Earth and Planetary Science Letters,* Vol. 58, No. 2, pages 167–188; April, 1982.

PROPAGATING RIFTS AND CONTINENTAL BREAKUP. V. Courtillot in *Tectonics,* Vol. 1, No. 3, pages 239–250; June, 1982.

THE SALMONID FISHES AS A NATURAL LIVESTOCK

THE ORIGIN AND SPECIATION OF ONCORHYNCHUS. F. Neave in *Transactions of the Royal Society of Canada,* Vol. 3, Series 3, Section 5, pages 25–39; June, 1958.

FISH OF RARE BREEDING: SALMON AND TROUT OF THE DONALDSON STRAINS. Neal O. Hines. Smithsonian Institution Press, 1976.

SALMON RANCHING. Edited by J. E. Thorpe. Academic Press, 1980.

PROCEEDINGS OF THE SALMON AND TROUT MIGRATORY BEHAVIOR SYMPOSIUM. Edited by E. L. Brannon and E. O. Salo. School of Fisheries, University of Washington, 1981.

LEARNING IN A MARINE SNAIL

CELLULAR ANALYSIS OF A GASTROPOD (*HERMISSENDA CRASSICORNIS*) MODEL OF ASSOCIATIVE LEARNING. Daniel L. Alkon in *The Biological Bulletin,* Vol. 159, No. 3, pages 505–560; December, 1980.

CALCIUM AND CYCLIC-AMP AS SYNARCHIC MESSENGERS. Howard Rasmussen. John Wiley & Sons, Inc., 1981.

CONDITIONING: REPRESENTATION OF INVOLVED NEURAL FUNCTIONS. Edited by Charles D. Woody. Plenum Press, 1982.

PRIMARY CHANGES OF MEMBRANE CURRENTS DURING RETENTION OF ASSOCIATIVE LEARNING. Daniel L. Alkon, Izja Lederhendler and Jonathan J. Shoukimas in *Science,* Vol. 215, No. 4533, pages 693–695; February 5, 1982.

MICROELECTRONIC PACKAGING

PHYSICAL LIMITS IN DIGITAL ELECTRONICS. R. W. Keyes in *Proceedings of the IEEE,* Vol. 63, No. 5, pages 740–767; May, 1975.

MICROELECTRONICS. *Scientific American,* Vol. 237, No. 3; September, 1977.

VLSI SUPPORT TECHNOLOGIES: COMPUTER-AIDED DESIGN, TESTING, AND PACKAGING. Edited by Rex Rice. IEEE Catalog No. EHO 191–7, IEEE Computer Society Press, 1982.

IBM 3081 SYSTEM DEVELOPMENT TECHNOLOGY. *IBM Journal of Research and Development,* Vol. 26, No. 1; January, 1982.

IBM PACKAGING TECHNOLOGY. *IBM Journal of Research and Development,* Vol. 26, No. 3; May, 1982.

PAPERS ON MANUFACTURING TECHNOLOGY—PACKAGING. *IBM Journal of Research and Development,* Vol. 27, No. 1; January, 1983.

THE DEVELOPMENT OF PALM LEAVES

ANATOMY OF THE MONOCOTYLEDONS, VOL. 2: PALMAE. P. B. Tomlinson. Oxford University Press, 1961.

THE NATURAL HISTORY OF PALMS. E. J. H. Corner. University of California Press, 1966.

ANATOMY OF SEED PLANTS. Katherine Esau. John Wiley & Sons, Inc., 1977.

THE MECHANISM OF PLICATION INCEPTION IN PALM LEAVES: PROBLEM AND DEVELOPMENTAL MORPHOLOGY. Donald R. Kaplan, Nancy G. Dengler and Ronald E. Dengler in *The Canadian Journal of Botany,* Vol. 60, pages 2939–2975; 1982.

ALTERNATIVE MODES OF ORGANOGENESIS IN HIGHER PLANTS. Donald R. Kaplan in *Contemporary Problems in Plant Anatomy,* edited by William C. Dickinson and Richard A. White. Academic Press, in press.

PARTICLES WITH NAKED BEAUTY

THE UPSILON PARTICLE. Leon M. Lederman in *Scientific American,* Vol. 239, No. 4, pages 72–80; October, 1978.

OBSERVATION OF A FOURTH UPSILON STATE IN e^+e^- ANNIHILATIONS. D. Andrews et al. in *Physical Review Letters,* Vol. 45, No. 4, pages 219–221; July 28, 1980.

DECAY OF *b*-FLAVORED HADRONS TO SINGLE-MUON AND DIMUON FINAL STATES. K. Chadwick et al. in *Physical Review Letters,* Vol. 46, No. 2, pages 88–91; January 12, 1981.

EVIDENCE FOR NEW-FLAVOR PRODUCTION AT THE Υ(4*S*). C. Bebek et al. in *Physical Review Letters,* Vol. 46, No. 2, pages 84–87; January 12, 1981.

OBSERVATION OF EXCLUSIVE DECAY MODES OF *b*-FLAVORED MESONS. S. Behrends et al. in *Physical Review Letters,* Vol. 50, No. 12, pages 881–884; March 21, 1983.

CREOLE LANGUAGES

BIOLOGICAL FOUNDATIONS OF LANGUAGE. Eric H. Lenneberg. John Wiley & Sons, Inc., 1967.

RULES AND REPRESENTATIONS. Noam Chomsky. Columbia University Press, 1980.

ROOTS OF LANGUAGE. Derek Bickerton. Karoma Publishers, Inc., Ann Arbor, 1981.

ISLE DE FRANCE CREOLE. Philip Baker and Chris Corne. Karoma Publishers, Inc., Ann Arbor, 1982.

THE AMATEUR SCIENTIST

MECHANICS. Arnold Sommerfeld. Academic Press, 1964.

BYRNE'S TREASURY OF TRICK SHOTS IN POOL AND BILLIARDS. Robert Byrne. Harcourt Brace Jovanovich, Publishers, 1982.

THE DYNAMICS OF SPORTS: WHY THAT'S THE WAY THE BALL BOUNCES. David F. Griffing. Mohican Publishing Company, 1982.

W I N E & F O O D

Chardonnay

The golden age of California Chardonnay is upon us.

Though this grape is a native of the Burgundy region of France, Chardonnay is now firmly rooted in the red soil of the Sonoma Valley. Here it produces a white wine of world-class quality and reputation.

WINE

After we have picked the grapes at optimum maturity and cold-fermented the wine to dryness, Chardonnay is then elevated to greatness by the judicious use of oak barrel aging.

Following an extensive two year barrel experiment, we selected French Nevers oak barrels to lend the desired character and complexity to our 1981 Proprietor's Reserve Chardonnay. The wine is made in a style that complements

food. Ripe apple and citrus aromas are followed by rich varietal flavors carefully balanced to the oak. The finish is classic Chardonnay, full and spicy.

FOOD

The full flavor of Chardonnay makes it an ideal complement to seafood, veal, turkey and chicken. Also, my wife Vicki and I enjoy our Chardonnay as an aperitif wine, served with sliced cheeses and fruit.

Because the marriage of wine and food is the most graceful expression of our art form, Vicki creates many recipes to enhance the unique character of our wines. One of my favorites is *Grilled Fish Steaks, Sonoma Style*. Seasoned with an unusual blend of pimento, parsley, watercress and garlic, the light and spicy flavors of this dish are a perfect match for Chardonnay.

If you would like to sample some of our Chardonnay recipes, and learn more about wine, please write to the address below for your free brochure, "Wine and Food: Chardonnay." We would enjoy sharing them with you.

Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.

Sam J. Sebastiani

Sebastiani Vineyards, Dept. G, P.O. Box AA, Sonoma, CA 95476

What if you chose as a technical

"At Union Carbide, an HP computer network automates production management and has paid for itself in less than two years."

Union Carbide's Linde Division in Tonawanda, N.Y., manufactures custom production equipment. Factory systems manager Doyce Coffman says, "When production demand fluctuates, it complicates our scheduling, product costing, and long range planning.

"To solve these problems we automated production and materials management with a network of HP 1000 and 3000 computers and 43 data capture terminals.

"Because these systems continuously monitor operations and respond in real time, we can identify and eliminate production bottlenecks on line. The payback? Improved efficiency factory-wide and a dollar savings that covered the cost of our entire HP network in less than two years."

What should you expect from a computer partner?

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

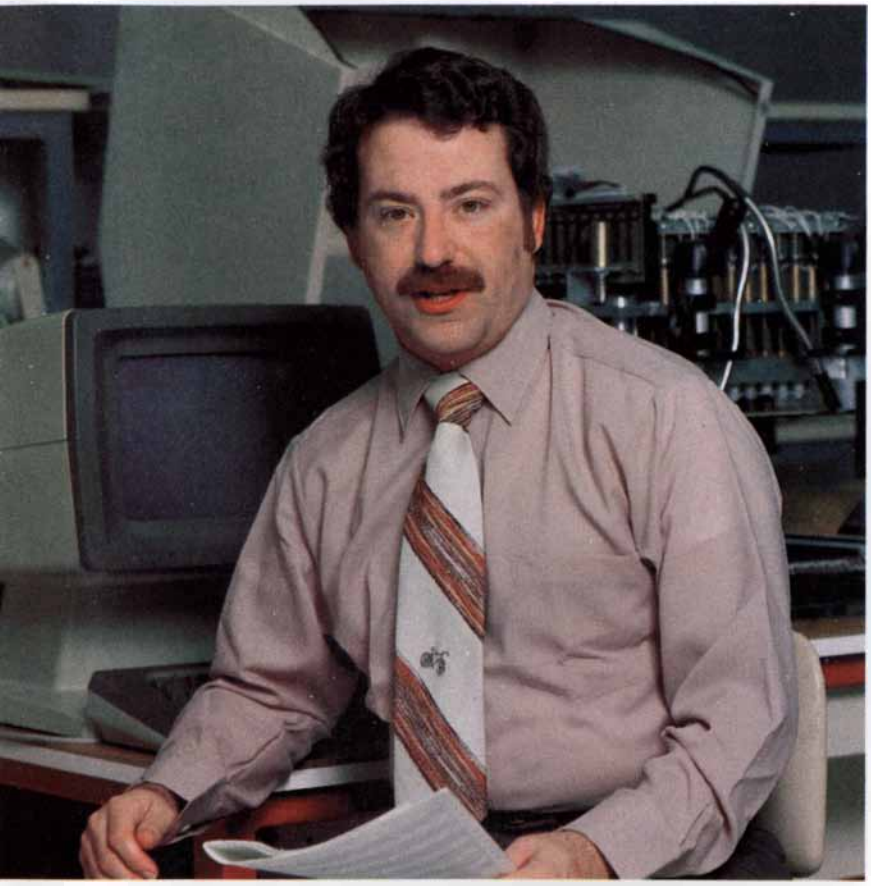
Quality, by HP's definition, encompasses more than the product.

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

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

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

Hewlett-Packard computer partner?



"At Amdahl, an HP computer network helps control testing. And it's paid for itself in two years."

Amdahl manufactures large-scale general purpose computers. In Sunnyvale, Calif., Bob Henderson, Manager of Test Systems Software, says: "We use a network of HP 1000 computers to control various tests on our Model 580 computers. With the HP systems, we can achieve the advanced testing and stringent quality control our high-speed, high-density products demand.

"Replacing batch systems with an HP computer network saved us nearly $100,000 in equipment costs. And by using HP 1000's we're obtaining test data ten times faster than before.

"Our test network helps streamline manufacturing processes and has paid for itself in just 2 years. In addition, the network saves us up to $34,000 a year by reducing paperwork and test cycle time."

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

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

You should consider a working partnership with HP. Now.

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

When performance must be measured by results

hp HEWLETT PACKARD

0002342

INTRODUCING THE PERFORMANCE SEDAN WITH A GENETIC ADVANTAGE.

Fifteen years ago, when big was beautiful and the auto industry was churning out ungainly V-8's, BMW introduced the 2002. A performance car.

Nine years later, big was out and the future was conceded to the tiny economy car. BMW introduced the 320i. A performance car.

Today the auto industry has embraced performance as the wave of the present. And BMW has introduced the 318i. A performance car—but one with an important difference.

It arrives with a heritage, as opposed to a history of recurring identity crises.

WHY A CAR'S PAST IS ITS FUTURE.

All cars, especially those aspiring to high performance, are products of engineering, technology—and genealogy.

The BMW 318i enjoys a substantial head start on all counts.

Its engine anticipated by years the current preference for fuel injection. The BMW 3-Series power plant serves as the basis for the BMW Formula One race engine—modified on the 318i by a new L-Jetronic system that generates even greater quickness and torque.

Its gearbox has been refined to match the greater torque, while retaining the characteristic precision that has inspired comparison to the linkage of fine timepieces.

Its new independent suspension is a virtual syllabus of lessons learned on racecourses, delivering handling and road holding Car and Driver judged "magnificent."

It offers a new graduated power-assist steering system, larger disc brakes, and such technological innovations as an electronic system that actually calculates when routine servicing is needed.

It is also among the automotive leaders in incongruity, having managed to boost fuel efficiency* at the same time it was improving its performance.

"THE FINEST QUALITY CAR AVAILABLE."

Inside, the 318i rebuts the myth that a noisy, bone-rattling ride under cramped conditions somehow certifies a vehicle as a true performance car.

It is roomier than its predecessors and well equipped with noise- and vibration-dampening measures. It also offers orthopedically-designed bucket seats and other functional amenities bearing the imprint of its racing lineage.

Its "overall ergonomics," summed up AutoWeek, "demonstrate why many consider the BMW the finest quality car available on the market today."

In short, the 318i continues and advances a 55-year commitment to performance. As today's fledgling performance cars will discover, there is no available shortcut to that experience.

If you see no need to discover the point along with them, we invite you to visit your nearest BMW dealer and test drive the 318i.

THE ULTIMATE DRIVING MACHINE.

*EPA-estimated 27 mpg, 38 highway. Figures are for comparison purposes only. Your actual mileage may vary, depending on speed, weather and trip length; actual highway mileage will most likely be lower. European Tourist Delivery can be arranged through your authorized U.S. BMW dealer.