SCIENTIFIC AMERICAN

THE SMALL CALCULATOR

$1.25

March 1976

You can own one in the world

Road & Track Magazine tested hundreds

What do you know? The VW Rabbit is the best car in the world for under $3500. Road & Track thinks so, and who are we to argue?

Here's what they have to say: "The Rabbit is something special in a small economy sedan. Its space for passengers and luggage is remarkable for such a small and light car; so is its ride and quietness. It'll be a rare driver who gets less than 25 mpg with it; on a long trip 35 is more like it."

(The 1976 EPA estimates for the standard shift Rabbit shown are even better: 39 mpg highway, 25 mpg city. Actual mileage may vary, depending on type of driving, driving habits, car's condition and optional equipment.)

"It comes as a two-or four-door sedan; either one has a hatchback and a folding rear seat for extra utility.

And—something you don't necessarily expect from a little economy sedan—it is delightful to drive, with peppy performance and first-class handling."

So, friends, if you want one of the 10 best cars in the world, but you don't want to pay more than $3500 for it, you have no choice.

*Suggested 1976 retail price East Coast P.O.E. (4-dr. model higher). Transportation, local taxes, and dealer delivery charges additional. †Suggested

of the 10 best cars for only $3499.*

of 1975 cars. Here are the winners:

currently available retail prices East Coast P.O.E. Transportation, local taxes, and dealer delivery charges additional. ©Volkswagen of America.

Pioneer's new High Polymer Molecular transducer technology will alter the course of high fidelity.

There's a significant new development in high fidelity that is destined to play a vital role in sound reproduction. It is intimately tied in with the piezoelectric principle.

The piezoelectric effect deals with certain crystal devices that flex when voltages are applied to them. Now, Pioneer has discovered a totally new application of the piezoelectric effect by applying the principle to ultra-thin aluminum coated high polymer film.

By employing this film as a low-mass diaphragm and applying audio signal voltages, the material expands and contracts uniformly generating acoustic energy. For the first time it becomes possible to transform electrical energy to an accurate acoustical equivalent. Such thin-film diaphragms properly mounted are capable of reproducing all music frequencies by means of an incredible "breathing" effect. The ramifications of this unique refinement of the piezoelectric principle are far reaching. Consider such immediate applications as microphones, cartridges, speaker systems and headphones — in fact, any type of electromechanical transducer requiring resonance-free performance.

There have been many attempts to create sound using diaphragm motion. For example, electrostatic speakers and headphones. But in contrast to the electrostatic principle, the new application of the High Polymer Molecular principle as discovered and perfected by Pioneer, requires no dangerous, high polarizing voltages.

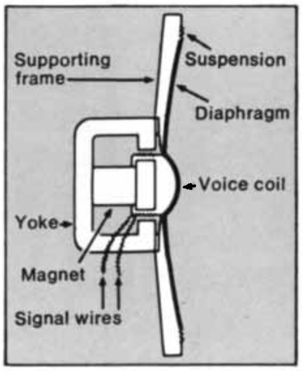


Fig. 1 Principle of conventional dynamic headphones

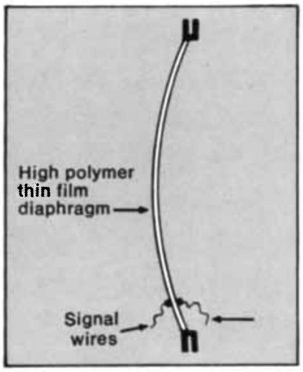


Fig. 2 Principle of new Pioneer High Polymer Molecular headphones

The first totally new concept in headphones in over a decade.

Pioneer has successfully incorporated the High Polymer Molecular transducer principle in two new headphones that are unlike any others. Conventionally designed headphones use moving coils, miniature loudspeaker elements and other mechanical parts — as shown in Figure 1 — all of which come between you and your music. Pioneer's new SE-700 and SE-500 headphones don't. They employ a single thin-film high polymer piezoelectric diaphragm that reproduces sound directly, as shown in Figure 2. Only the diaphragm moves air — and moves it accurately, in exact conformance with the electrical signal applied directly to it. The accurate, low-distortion signals available from any standard headphone jack on your receiver or amplifier are directly translated to equally precise, low-distortion sound by the action of the high polymer film diaphragm. Nothing, absolutely nothing comes between you and the original sound.

Even though you may now own a pair of headphones, you owe it to yourself to hear these new piezoelectric high polymer transducer headphones. In fact, compare them with other types. You'll find a lower level of distortion-free sound than has ever been achieved — even at unprecedented volume levels. The experience of listening with these new Pioneer headphones is a revelation. In addition, the open-back design, light weight and soft, snug fitting earpieces permit hours of comfortable, private listening. You'll come away from your Pioneer dealer thoroughly convinced that Pioneer has altered the course of high fidelity.

SE-700, under $80*; SE-500, under $50*. Both come with a 9¾ foot connecting cable, standard phone plug and storage case.

U.S. Pioneer Electronics Corp., 75 Oxford Drive, Moonachie, New Jersey 07074. West: 13300 S. Estrella, Los Angeles 90248 / Midwest: 1500 Greenleaf, Elk Grove Village, Ill. 60007 / Canada: S.H. Parker Co.

*The value shown is for informational purposes only; the actual resale price will be set by the individual Pioneer dealer at his option.

PIONEER
when you want something better

Established 1845 SCIENTIFIC AMERICAN March 1976 Volume 234 Number 3

ARTICLES

DEPARTMENTS

THE COVER

The painting on the cover shows part of the keyboard of a small electronic calculator, the HP-65, manufactured by the Hewlett-Packard Company. The HP-65 is at the upper end of the range of calculating capability of small electronic calculators (see "The Small Electronic Calculator," page 88). It is programmable by way of both the keyboard and magnetic program cards, and it performs 47 mathematical operations and manipulations of data beyond the four basic functions (addition, subtraction, multiplication and division). The four keys shown at the top of the painting have as their main respective functions, indicated by black letters: entering a number in the machine, changing sign, entering an exponent and clearing the *X* register (an operation that erases the number displayed). Many other options are available through three color-coded shift keys, which are not shown. Two of the shift keys yield the operations indicated in gold letters above the keys and in blue letters on the underside of the keys. The third shift key gives access to associated inverse functions. For the purpose of explaining how a calculator works the article in this issue describes a simpler kind of machine, a hypothetical four-function calculator.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
26–27	Charles S. Lieber, Mount Sinai School of Medicine
28–33	Bunji Tagawa
34	W. Suter, Natural History Museum, Basel
36–40	Allen Beechel
41	R. Sawdo, Woods Hole Oceanographic Institution
42–45	Allen Beechel
47	Gareth P. Williams, Princeton University
48	Ames Research Center, National Aeronautics and Space Administration
49–53	Alan D. Iselin
54	Ames Research Center, National Aeronautics and Space Administration
55	Alan D. Iselin
56	Ames Research Center, National Aeronautics and Space Administration
62–70	Gabor Kiss
72–79	Gabor Kiss
80–86	George V. Kelvin
89	Ben Rose (*top*), Ralph Morse (*bottom*)
90	Rockwell International
92–98	Jerome Kuhl
100	J. Wesley Burgess
101–104	Tom Prentiss
105	J. Wesley Burgess
106	Tom Prentiss
109–110	Jon Brenneis
112–116	Dan Todd
119	"B.C.," by permission of John Hart and Field Enterprises, Inc.
120–122	Ilil Arbel
127–128	Dover Publications, Inc.

LETTERS

Sirs:

In "The Strip-mining of Western Coal" [SCIENTIFIC AMERICAN, December, 1975] Genevieve Atwood apparently has accepted as truth certain fictions concerning coal "development." Contrary to her view that "there are currently no real alternatives to oil, gas and coal as energy sources . . . ," there are, of course, alternatives, among them the development of radically new energy technologies and the implementation of far-reaching measures to encourage energy conservation. However, neither of these alternatives is possible within a society that insists on continuous growth, oil, gas and coal being the darlings produced by the union between the growth economists and the corporate giants.

The view that Western coal "has a low sulfur content" is certainly suspect. Scientists working in southeastern Montana and northeastern Wyoming have recently questioned this view, which was based on the investigation of cores taken from the center of coal seams in the northern Great Plains. Sampling of the edges of the seams as well as more complete sampling of the entire seams have revealed a higher sulfur content than was originally suspected. Further, because of the high water content of much of the coal lying under the northern Great Plains, there are fewer B.t.u.'s produced per ton of Western coal than there are per ton of Eastern coal. Indeed, some authorities are now asserting that the amount of pollution produced per B.t.u. of Western coal will be greater than that produced for Eastern coal.

And reclamation is not working. For example, in Montana, the state noted for having the "toughest" reclamation laws in the U.S., the piles of "topsoil and other unconsolidated material" are allowed to lie untouched and untreated for three years or more—ample time to guarantee the leaching out of toxic materials and the sterilizing of formerly living earth. In fact, Montana reclamation laws are simply not being enforced. It is significant that the Montana Wilderness Association recently (December 7, 1975) went on record advocating the transfer of enforcement functions from the Department of State Lands to the Department of Natural Resources and Conservation in reaction to the failure of State Lands to enforce Montana reclamation laws.

Furthermore, there should be no cause for celebration to be derived from the observation that "the output per man-hour at a strip mine is approximately eight times the output at a fairly modern underground mine." Translated into human terms, this merely means that far too many underground miners will be thrown out of work—sacrificed to the appetite of the dragline. Surely the cost is too high.

Finally, I must take exception to Ms. Atwood's too easy assumption that "in Wyoming, Montana and much of the Great Plains the conditions for surface mining approach the ideal." Conditions are ideal only for the energy-development corporations. Not for the land. *And not for the people.* Destruction of aquifers in the coal beds, deliberate bypassing of reclamation regulations, displacement and eventual phasing out of people, boom, bust: these are the present facts of strip-mining life in the West. In view of these circumstances, would it not be preferable to eliminate strip-mining rather than to cater to the greed of corporate giants beholden only to themselves?

RANDALL GLOEGE

Northern Rockies Representative
Friends of the Earth
Billings, Mont.

Sirs:

I should first like to point out that in amputating the sentence he quotes in his first paragraph Mr. Gloege changes its meaning. The original sentence was clearly referring not to alternative energy sources in general but to short-term alternative sources. As for his assertion that the reclamation of strip-mined lands is not working, I can only say that reclaimed land at several mines I personally have visited is ecologically and hydrologically in balance with adjacent land (rangeland with diverse, self-regenerating vegetation).

Mr. Gloege raises so many questions that it is not possible to respond to all of them in a "letter to the editor." For the rest I shall

Scientific American, March, 1976; Vol. 234, No. 3. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

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

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

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

Name ______

New Address ______

Old Address ______


JAGUAR
Jaguar S
XJ-S

JAGUAR UNLEASHES A NEW BREED OF CAT. THE S-TYPE.

It may well be the best-handling four-passenger car in the world.

The cat is nimble. Jaguars have always been known for their agility rather than brute force. Superb handling and braking were the extra edge that helped carry Jaguar to five victories at Le Mans. With catlike grace they negotiated the twists and turns of the most historic racecourse in the world.

Jaguar's superb handling capabilities have been refined and refined over the years. In doing this, Jaguar has created classic after classic.

This new breed of cat is all that the world has come to expect of Jaguar. It advances the state of the art. It moves and responds even better than previous Jaguars.

In plain fact, the S-type is so thoroughly wedded to the road in cornering and evasive maneuvers, it scarcely seems credible. Until you experience it.

These superb handling characteristics are solidly grounded in the S-type's all-independent suspension, four-wheel power disc brakes and power-assisted rack and pinion steering.

The wind itself makes the S-type handle better. Wind tunnel tests have sculpted the Jaguar S-type into a machine for road-adhesion. A front air dam and an under-engine shield both contribute to increased stability, better handling and quicker acceleration.

The S-type's tires, too, are a handling factor. They have been specially developed by Dunlop to match the unique performance capabilities of the car. And the S-type's wheels are also functional. Their peripheral vents are more than cosmetic: these vents actually deflect significant amounts of heat from the disc brakes and help the brakes perform better.

Yet for all its uncommon handling capabilities, the XJ-S is so very smooth, quiet and thoroughly insulated from the harshness of road surfaces and the wind at high speeds that it rides like a luxury sedan. The smoothness and quiet of the XJ-S in motion would be remarkable in any car. To find these qualities combined with the Jaguar S-type's outstanding level of handling and response is almost miraculous.

Why the cat purrs: the astounding 5.3 litre fuel-injected Jaguar V-12. No conventional powerplant could match the exceptional handling capabilities of the Jaguar S-type.

Therefore its engine is a fuel-injected version of the short-stroke, wide-piston V-12 that holds the 1975 SCCA National Championship in Class B Production. Because of its perfectly balanced configuration, the Jaguar V-12 is exceptionally smooth and quiet. Yet despite its relatively modest displacement (5.3 litres or 326 cubic inches), this engine develops an astonishing SAE net horsepower of 244 at 4,500 rpm, making it one of the most powerful engines offered by any manufacturer, regardless of displacement.

Yet sheer power is not the most beguiling quality of the Jaguar V-12. The magic comes in the engine's unexpected docility at low traffic speeds, and the virtually unequalled smoothness of its acceleration.

And this is an eminently reliable engine. Jaguar test engineers have called it "virtually indestructible." Bosch/Lucas electronic fuel injection and an electronic ignition system that eliminates contact points both contribute to the engine's efficiency and reliability. In environmental tests, the V-12 has proven itself reliable, starting at such extremes of temperature as −40° and +110°.

Astounding? Not for the Jaguar S-type.

The cat is wary. Red alert: the S-type's new instrumentation. Very high performance demands a superb intelligence-gathering system for the driver. To this point, the S-type instrumentation has been completely redesigned to make all vital functions instantly discernible.

All instruments are functionally gathered in one nacelle that is clearly visible through the leather-padded steering wheel. There are large circular dials for the speedometer and tachometer and new quick-reading vertically-calibrated gauges monitoring fuel level, water temperature, oil pressure and battery condition.

A bank of 18 red and amber signal lights above this instrument cluster give the driver an early warning of malfunctions in the S-type's safety, mechanical and electrical systems.

The cat is complete. The S-type is a four place luxury car. The Jaguar S-type comes to you complete: there are no options whatever.

Inside the S-type there is the authentic elegance that only comes when everything is designed to be functional. There is no ostentation about the car: everywhere you look, your eye will be delighted by clean, graceful lines and rich textures.

The seats are covered in topgrain hides. Deep carpeting has been fitted throughout the passenger compartment and inside the spacious trunk.

Every mechanical amenity conceivable has been included: electric windows, electric door-locks, eight-track AM/FM stereophonic sound system and automatic thermostatically-controlled air conditioning system are all standard.

And there is one final luxury that comes as standard equipment on the Jaguar XJ-S. This is the uncommonly thoughtful Jaguar warranty. For 12 months, regardless of mileage, Jaguar will replace or repair any part of the car that is defective or that simply wears out, provided only that the car is properly maintained. The only exceptions are the tires, which are warranted by the tire manufacturer, and spark plugs and filters, which are routine replacement items. Even then, if they are defective, Jaguar will pay to replace them.

Keeping up with the cat. The standards by which personal luxury high-performance cars are measured have just risen by a substantial margin. Drive this remarkable Jaguar S-type soon. You'll discover what a quantum leap forward it marks in the very small world of personal luxury, high-performance motorcars. For the name of the dealer nearest you, call these numbers toll-free: (800) 447-4700, or, in Illinois, (800) 322-4400. British Leyland Motors Inc., Leonia, N.J. 07605.

BRITISH LEYLAND

address myself to two points, because in my view they are widely misunderstood.

First, it is true that the higher moisture content of Western coal increases its sulfur emissions per B.t.u. The emissions per B.t.u. are still, however, lower than those of Eastern coal. The contention that recent sampling has revised the sulfur-content figures is unsubstantiated and is based on tentative, reconnaissance-level information reported in hearings on the Federal Coal Leasing Amendments.

Second, Mr. Gloege implies that strip-mining should be banned in favor of the more labor-intensive and (he says) environmentally sounder underground methods. When surface mining is properly done, it is environmentally, socially and economically more acceptable than underground mining. The reasons are as follows: (1) The environment for the worker is healthier and safer. In 1971 the fatalities per million man-hours in underground mines were .84; in surface mines they were .45. The injuries per million man-hours in underground mines were 52.81; in surface mines they were 25.86. (2) The air in a surface mine may contain some dust, but it is of far better quality than the air in an underground mine. This is reflected in the high incidence of pneumoconiosis among underground miners. (3) Water-quality problems can be rectified at active and abandoned surface mines but not at abandoned underground mines. (4) Underground mining is wasteful of coal resources. Traditional underground methods recover on the average 50 percent of the coal. (Fifty percent remains as supporting pillars or is otherwise unmined.) Surface mining of thick Western seams recovers more than 95 percent of the coal seam. (5) Land-use and hydrologic problems associated with surface mining are severe but so are the land-use and hydrologic problems associated with fires in abandoned underground mines and waste disposal from such mines.

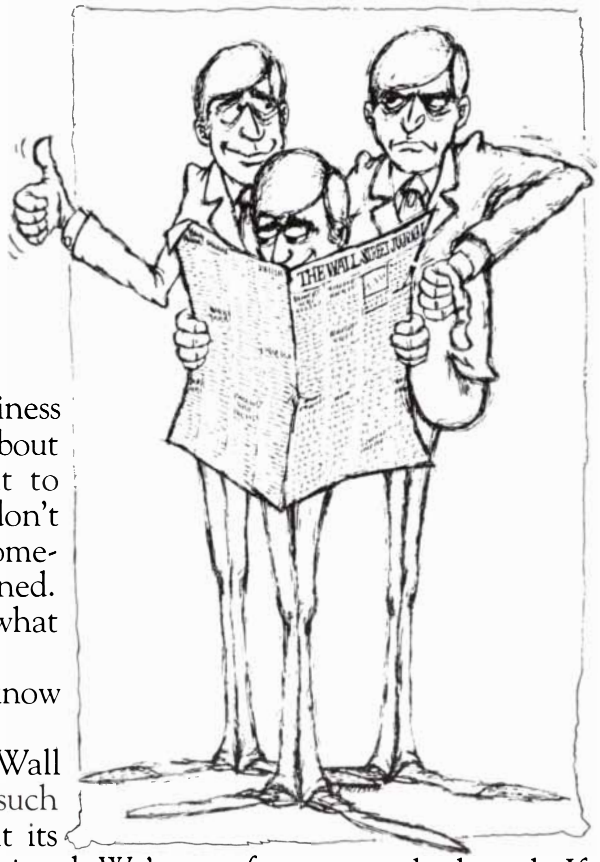
Underground mining is far less economic than surface mining, and it only seems environmentally and socially more acceptable because it is out of sight.

GENEVIEVE ATWOOD

House of Representatives
State of Utah
Salt Lake City, Utah

ERRATUM

In "The Necessity of Fission Power," by H. A. Bethe (SCIENTIFIC AMERICAN, January), it was stated that even in the extremely unlikely event of a severe nuclear accident the chance that a person in the exposed population would ultimately die of cancer from the released radioactivity was 1 percent. The figure should have been .1 percent.



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1926: "The high mark in crude-oil production in the United States was reached in 1925 when 758,500,000 barrels were obtained. Twenty years ago this country's consumption amounted to only about 100,000,000 barrels. Thus has the tremendous growth in the number of automobiles and the general use of oil for power, heat and light given this product front-page importance. Discoveries of new fields and improved knowledge have made oil a respectable industrial commodity. In the United States the most productive field has been California, where in 1925, 231,673,000 barrels were produced, or more than one-third of the country's entire output. Texas, Oklahoma, Wyoming, Kansas, West Virginia, Louisiana, Pennsylvania, Ohio and Kentucky are the other chief producing states."

"Rays that reach the earth from some unknown outside source have proved to be far more penetrating than X rays. The tests made at Muir Lake in California are described by Professor R. A. Millikan: 'At the surface of the lake our electroscopes showed a rate of discharge corresponding to 12.4 ions per centimeter per second. As we sank the electroscope to greater and greater depths this rate of discharge decreased continuously down to a depth of 50 feet, after which it became constant at a value of 3.4 ions per cubic centimeter per second. This meant that we had obtained much more unambiguous and thoroughly quantitative evidence than had before been available for the existence of a radiation so hard that, if it came from outside the atmosphere, it required 50 plus 23, or 73, feet of water to absorb it entirely. The 23 feet is the water equivalent of the atmosphere above Muir Lake. Since at all the altitudes at which we have now experimented we find no measurable variation in the intensity of these rays at any time between midday and midnight, we conclude that they shoot through space equally in all directions.'"

"In a recent address our indefatigable Secretary of Commerce, Herbert Hoover, pointed out that our material civilization rests squarely on the sub-foundation of abstract research and pure science. When you telephone, when you drive your car, when you wash with scientifically made soap or eat scientifically balanced bread, do you think of the ultimate source of the scientific truths that made them possible? Just as engineering, industry, medicine and all the important arts of today rest on fundamental scientific principles, doggedly worked out in the past, so those of the future will depend on the research of today."

"Easily the outstanding feature of the recent annual meeting of the American Association for the Advancement of Science was Professor Dayton C. Miller's announcement that he had carried his previous experiments on ether drift to the extent of making over 100,000 observations, largely during the past summer, and that they apparently prove the existence of an ether drift. The significance of this result, if it is substantiated, is that the Einstein theory must be very seriously affected. In fact, Professor Einstein himself makes the flat statement that 'if Dr. Miller's results should be confirmed, then the special relativity theory, with the general theory in its present form, falls.' Continuing, Professor Einstein says: 'Experiment is the supreme judge.'"

"'When the history of recent stellar astronomy comes to be written, some living names such as H. N. Russell, Shapley, Eddington and Jeans will shine as stars of the first magnitude.' This is what Sir Oliver Lodge, the world's most noted scientist, says in *Nature* about Professor Henry Norris Russell of our editorial staff. Twenty-six years ago an obscure graduate student from Princeton University submitted an article to us about the stars. Ever since then, with almost unbroken regularity, the same author has written a monthly article for this magazine. From obscurity Professor Russell has passed during these years to scientific note, especially because he took the leading part in the development of stellar evolution. We rejoice that we are still able to present to our readers a monthly review of the progress of the sciences of astronomy and astrophysics, written by so high an authority."

SCIENTIFIC AMERICAN

MARCH, 1876: "It has been assumed, too hastily we think, that the growth of the country in population has received a serious check. This conclusion is based on the returns of a part of the census taken last year. So far as is known there were fourteen States that took a census of their population during the year 1875. Of the fourteen censuses, eleven are supposed to be trustworthy and three are not so. The eleven States whose returns are worth any thing show a growth of 1,612,000 in five years. As they represent slightly more than one-fourth of the entire country, it is safe enough to estimate that the increase during the last five years has been five millions, making our total population in the middle of 1875 something more than forty-three millions."

"The Japanese government building at the Centennial grounds in Philadelphia is exciting a great deal of curiosity, not only

Volaré

The accent is on comfort... and space.

The comfort in our new kind of small wagon starts with the doors. There are 4, not 2, like some other small wagons. And every one of those doors is wide for easy entry.

You and your packages. Handled with care.

The comfort and convenience extends to handy storage compartments. And a liftgate with gas pressure props for easy opening (they disappear when closed).

Inside, Volaré has 70% as much cargo space as an average full-size wagon. And yet it is 2 feet shorter, 1,200 lbs. lighter and $1,500 less expensive.

Small wagon efficiency. Big wagon ride.

The Volaré is very easy on gas. According to E.P.A. estimated mileage results, of all Volarés, the wagon did best—30 mpg on the highway, 18 in the city—equipped with 6-cylinder engine and manual transmission. Your actual mileage may differ, depending on how and where you drive, the condition of your car and optional equipment.*

The Volaré is easy on you, too. Because it has an Isolated Transverse Suspension System (pat. pend.) that gives you a smooth, comfortable ride like a big car.

And here's "The Clincher."

For the first 12 months of use any Chrysler Corporation dealer will fix, without charge for parts and labor, any part of our 1976 passenger cars we supply (except tires) which proves defective in normal use. Regardless of mileage. The owner is responsible for maintenance services such as changing filters and wiper blades.

And the part about "regardless of mileage" is especially nice. Because Volaré is one wagon you're going to love to drive. Volaré.

The new small wagon from Plymouth.

*In California, see your dealer for engine availability and mileage data for California equipped vehicles.

on account of its peculiar method of construction but also on account of the peculiar tools and appliances used by the workmen employed in its erection. The workmen, who were sent from Japan with the four car-loads of timber necessary to the construction of the building, appeared to be intelligent and expert artisans, and handled their tools in a manner showing that they feel a complete mastery over them. The members of the building fraternity whom we met on a recent visit, after minutely examining the framework of the structure, pronounced it a 'beautiful piece of workmanship' but 'a complete Chinese puzzle.' Not a nail or a screw is to be found in the whole of it."

"In referring to the progress of steel rail manufacture in this country nearly a year ago, we took occasion to point out the rapid falling off in the importation of steel rails from England and expressed the opinion that still further decrease would follow. We have, moreover, long adhered to the belief that, with our vast sources of mineral wealth supplying the material, all that our people have needed is experience and knowledge in utilizing them, and as that experience and knowledge augmented, so would our reliance on the labor of other countries decline. Steel makers in Sheffield have felt severely the effects of a falling off in their American sales, but none, it appears, were prepared for the alarming announcement that the American market for their rails had practically closed against them. In 1873 Sheffield exported steel goods of all kinds to the United States of an aggregated value of $8,298,865; in 1875 this had fallen to $3,456,160."

"The inhalation of carbon in the form of fine powder is a common evil attached to industrial labor. The coal-miners, the carriers of coal, they who are exposed to dense smoke and they who are exposed to some other occupations, such as walking-stick making, suffer from this irritant. The particles produce comparatively little irritation, but they are carried readily into the bronchial tract through the minute ramifications of the bronchial surface and render the breathing irregular through parts of the lung. Thus an irregular pressure of the inspired air is brought about, and undue pressure is exercised on some portion of the lung structure, there is a rupture or break of the minute vesicular structure, and therewith the development of that disease of the lung which is technically called *emphysema*. The symptoms attendant on this condition are those of suffocated breathing and spasmodic cough. One of these workers, who was greatly reduced by the disease, told us very simply and clearly the facts in a sentence that we repeat. 'It is all right,' he said, referring to charcoal dust, 'it is all right when we can cough it up. If a man get that way, he can stand his work for many years, but eventually the cough always masters us, and we break down.'"

ELECTRONIC BREAKTHROUGH!

New TV Game

Turn your TV into an electronic playground with a new computerized remote control TV game with a psychological twist.

Atari's new remote control, pre-programmed computer ping-pong game can be used with any size TV set—black and white or color.

How many fun things can your entire family enjoy? TV is certainly one of them.

Now through solid-state technology and remote control electronics, you can convert your TV into a ping-pong game so authentic, it even sounds like ping-pong.

This new electronic breakthrough, called Pong, automatically keeps score with its own scoreboard, automatically increases speed, and is psychologically designed to be quick, exciting, and a challenge for all age levels. In fact, with a little practice, it's even possible for an eight year old to beat her father.

FUN AND EASY TO PLAY

A ball is projected from the middle of the screen to one side of the playing field. The paddle is a narrow line on the screen which you move up and down by twisting a knob on the remote control panel. To hit the ball, you position the paddle to collide with the ball.

A speaker in the unit generates a different ping-pong sound when the ball is either served, hits the paddle, or is missed. And each hit of the ball at different angles can propel it in any one of seven different directions, thereby adding the element of "english" to each paddle stroke.

"I beat ya dad!" The screen displays an array of squares signalling the end of the match as one of the players reaches fifteen points.

To start the game, you just turn on your TV to channel three or four, switch on Pong, and press the start button. Your screen is suddenly covered by a playing field (in color on a color TV), and the ball is projected from the middle of the screen. Two can play or you can play by yourself. And, since Pong is on the TV, the whole family can watch and enjoy the action too.

DESIGNED WITH A CHALLENGE

The game is programmed to slowly serve the ball at a variety of different angles to the player who lost the last point and then speed up after four strokes. The ball will then speed up again after four more strokes. The double speed-up feature keeps Pong a constant challenge as you improve your skill.

A typical game goes quickly—normally between 2½ to 4 minutes. After each point is scored, the new score is flashed on the screen and then disappears as the ball is served again. The first player to reach 15 points wins, and a pattern of squares fills the screen to signal the end of the game. To play again, just press the start button.

NO SPECIAL INSTALLATION REQUIRED

Is Pong hard to install? Not at all. A special receiver/switch clips onto the antenna connection of any TV and you're ready to play. It's just that simple. You can also leave the switch permanently clipped onto your TV while you watch your favorite TV shows.

PONG WAS NO ACCIDENT

Pong was invented and built in America by Atari, an American company (with a Japanese name) that manufactures 65% of the nation's coin-operated electronic games—games that sell for thousands of dollars and must produce a profit for their owners.

Atari knew from experience what the public enjoyed in a game. It was their business to know. So Pong was designed with all the proven fun, psychological twists and appeal of the expensive coin-operated models.

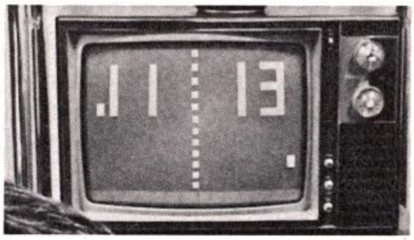
DON'T CONFUSE PONG

There are other TV games on the market that appear to be the same as Pong. Don't be confused. Pong is the only game that uses a single, large-scale integrated and pre-programmed circuit. There's nothing to put over your TV screen to make it work, and it's completely portable. You can play it from as far away as fifteen feet and use it on other TV sets.

AND TALK ABOUT ABUSE

You can really abuse Pong. Its solid-state circuitry and high impact case mean that you can drop it, kick it, and abuse it, without fear. Pong should never require service, but if it does during or after the unit's 90 day warranty, mail it in its handy mailer carton to Atari's prompt national service-by-mail center.

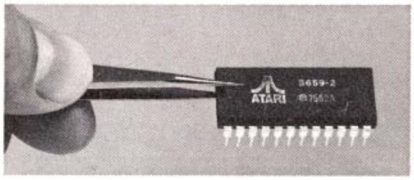
Pong comes complete with four "D" cell flashlight batteries that give over 50 hours of fun. With its low voltage and the unit's fully-sealed, tamper-proof enclosure, there is absolutely no danger to you or your children. Even the FCC (Federal Communications Commission) has personally inspected and approved the game's amazing electronics. Pong will not cause interference to other TV sets on your block or in your house.



The automatic scoreboard displays the score in large numerals after a point is won and then disappears when the next ball is served.

AN INVESTMENT IN YOUR FAMILY

Pong will bring your entire family closer together with a game all of you can play whether you're five or ninety-five. Your children will spend less time watching TV shows and more time playing with their friends. Pong will teach coordination, speed and dexterity. The young, the elderly, the sick or the bed-ridden—all will find Pong fun. Pong is also the perfect gift, an unexpected executive toy, or an exciting way to entertain at parties.



The large-scale integrated circuit with 2,487 active transistors is the new computer component that makes Pong the nation's most advanced and easiest-to-operate TV game.

We feel so strongly about Pong, that we'll make this offer. Order Pong. When it arrives, give it a real workout. Play it with your family and friends. If you do not thoroughly enjoy it or if you do not think that it is a great investment in your family's entertainment, return it within two weeks after receipt for a prompt and courteous refund. There is no risk. No other company offers you the opportunity to personally try this product in your own home with no obligation. JS&A is America's largest single source of space-age products for business and pleasure.

To order Pong, credit card buyers may call our toll-free number below or send a check for $102.95 ($99.95 plus $3.00 for postage, insurance and handling) to the address shown below, (Illinois residents add 5% sales tax). You'll receive by return mail, one complete remote control unit, clip-on receiver/switch, batteries and instructions. An optional AC adapter for $9.95 is available. Additional receiver/switches for other TVs in your house are also available for $9.95. But don't delay. Start enjoying the fun of computer electronics at home on your own TV screen, today.

JS&A NATIONAL SALES GROUP

DEPT. SA JS&A Plaza
Northbrook, Illinois 60062
CALL TOLL-FREE .. 800 323-6400
In Illinois call 800 648-4990

THE AUTHORS

CHARLES S. LIEBER ("The Metabolism of Alcohol") is professor of medicine at the Mount Sinai School of Medicine of the City University of New York and chief of the research program on liver disease and nutrition at the Bronx Veterans Administration Hospital. A Belgian by birth, Lieber was educated at the University of Brussels, where he received his M.D. in 1955 as the top-ranking student in his class. After clinical training in internal medicine he turned to full-time research, working first at the Queen Elisabeth Medical Foundation in Brussels. He came to the U.S. in 1958 to continue his studies at the Harvard Medical School and the Boston City Hospital and was appointed to the Harvard faculty two years later. Before taking up his present posts in 1968, he was associate professor of medicine at the Cornell University Medical College and director of the liver-disease and nutrition unit at Bellevue Hospital in New York. The author or coauthor of some 300 scientific publications, he has in recent years served as president of both the American Medical Society on Alcoholism and the American Society for Clinical Nutrition. Lieber wishes to acknowledge the fact that the original studies reviewed in his article were supported by the Veterans Administration, the National Institute of Alcoholism and Alcohol Abuse, the National Institute of Arthritis, Metabolism, and Digestive Diseases and the National Council on Alcoholism.

MAX BLUMER ("Polycyclic Aromatic Compounds in Nature") is a senior scientist at the Woods Hole Oceanographic Institution. Born and raised in Basel in Switzerland, he obtained his Ph.D. in chemistry from the University of Basel in 1949. He worked in a number of industrial and academic laboratories in Europe and the U.S. until 1959, when he joined the staff at Woods Hole. He writes: "I have been fortunate in finding an environment where I could remain intimately involved with the experiment. Nature is too complex and my anticipation of the unexpected is too limited to allow me to delegate most observations. This attitude has been rewarded many times when I have suddenly found a consistent explanation for seemingly independent observations, often separated by many years."

ANDREW P. INGERSOLL ("The Meteorology of Jupiter") is associate professor of planetary science at the California Institute of Technology. A graduate of Amherst College, he went on to Harvard University, receiving his Ph.D. in 1965. A year later he moved to California, where, he writes, "I live with my wife Sarah, our three kids and six student types in what might be called a commune for hardworking scientists. We tried being a nuclear family for a few years, but we got lonely. Now we have a built-in social life in our own home." Commenting on his work, Ingersoll adds: "My professional dream is to put together a general theory of atmospheres that works for all the planets, including the earth."

J. RICHARD GOTT III, JAMES E. GUNN, DAVID N. SCHRAMM and BEATRICE M. TINSLEY ("Will the Universe Expand Forever?") began their collaboration in 1974, while Gott and Gunn were at the California Institute of Technology and Schramm and Tinsley were at the University of Texas at Austin. The key connection came when Gott traveled to Austin to give a talk and the four realized, in Tinsley's words, "that pieces of evidence from our various research specialties could be fitted together into a surprisingly clear case for an open universe. There were several subsequent visits between Texas and California and many long telephone conversations while the work on our joint paper for *The Astrophysical Journal* was in progress, but it was not until the summer of 1975 that all of us actually got together. Then, at the Institute of Astronomy of the University of Cambridge, we used the material from our published paper to write (with the literary assistance of Rosemary W. Gunn) this article for *Scientific American*." Gott, who was graduated *summa cum laude* from Harvard University in 1969 and obtained his Ph.D. in astrophysics from Princeton University in 1972, was recently appointed to the Princeton faculty. Gunn, a graduate of Rice University, received his Ph.D. from Cal Tech in 1965. He is now professor of astronomy at Cal Tech and a staff member of the Hale Observatories. Schramm acquired his undergraduate degree from the Massachusetts Institute of Technology in 1963 and his doctorate from Cal Tech in 1971. (That same year he won the national Greco-Roman wrestling championship.) He is currently associate professor of astronomy and astrophysics at the Enrico Fermi Institute of the University of Chicago. Tinsley has an M.S. in physics from the University of Canterbury in New Zealand (1963) and a Ph.D. in astronomy from the University of Texas (1967). She is now associate professor of astronomy at Yale University.

JOHN ROSS ("The Resources of Binocular Perception") is professor of psychology at the University of Western Australia. He is a graduate of the University of Sydney and has a Ph.D. in psychology from Princeton University. His current professional interest "is with applications of discoveries about human vision. Together with my associate Monte Sala I have patented two inventions. One is a picture-transmission system based on the response of vision to random scans. The other is a new type of display based on phase sensitivities." He adds: "I do not have any hobbies in the usual sense, although I enjoy horse racing, which poses a whole set of interesting problems in applied probability theory."

EUGENE W. McWHORTER ("The Small Electronic Calculator") works as an engineer for the Texas Eastman Company, a chemical-manufacturing facility of the Eastman Kodak Company. From 1966 to last year he was employed by Texas Instruments Inc., where he put his knowledge of semiconductor electronics to use writing and supervising production of various educational materials for the Texas Instruments Learning Center. He was the author, for example, of *Understanding Solid-State Electronics,* a basic textbook on semiconductor principles and applications. His degrees include an M.S. in chemical engineering from the University of Texas at Austin and an M.B.A. from Southern Methodist University.

J. WESLEY BURGESS ("Social Spiders") is a graduate student at North Carolina State University, where he is conducting an investigation of spider behavior as a member of the Department of Mental Health Research. Burgess is a graduate of Purdue University, where he was a member of the Undergraduate Entomology Research Committee. Concerning the subject of his present article he writes: "Finding animals living together in cooperation is exciting whether they are spiders or any other animal group, and I am sure it is a phenomenon that will turn up more often, as people's interest in and awareness of simple social life is stimulated."

LEONARD A. HERZENBERG, RICHARD G. SWEET and LEONORE A. HERZENBERG ("Fluorescence-activated Cell Sorting") have worked together for several years at Stanford University on the development of the novel technique described in their article. Leonard Herzenberg is professor of genetics at the Stanford School of Medicine. A graduate of Brooklyn College, he acquired his Ph.D. from the California Institute of Technology in 1955. He then spent two years as an American Cancer Society Fellow working in the laboratory of Jacques Monod at the Pasteur Institute in Paris before coming back to the U.S. to continue his research at the National Institutes of Health. He moved to Stanford in 1959. Sweet, who serves as chief engineer on the cell-sorter project, majored in electrical engineering as an undergraduate at Cal Tech. He began his association with Stanford in 1956 at its Applied Electronics Laboratory, where he worked on the development of high-speed ink-jet recording and printing devices. After four years at Varian Associates he went back to Stanford in 1971 to join the cell-sorter project. Leonore Herzenberg has collaborated on a number of research projects with her husband, although, she adds, "each of us maintains a separate research identity." Her chief current interest involves the regulation of the immune response by lymphocytes that originate in the thymus gland.

The New Standard.

Unheard-of power and untold luxury.

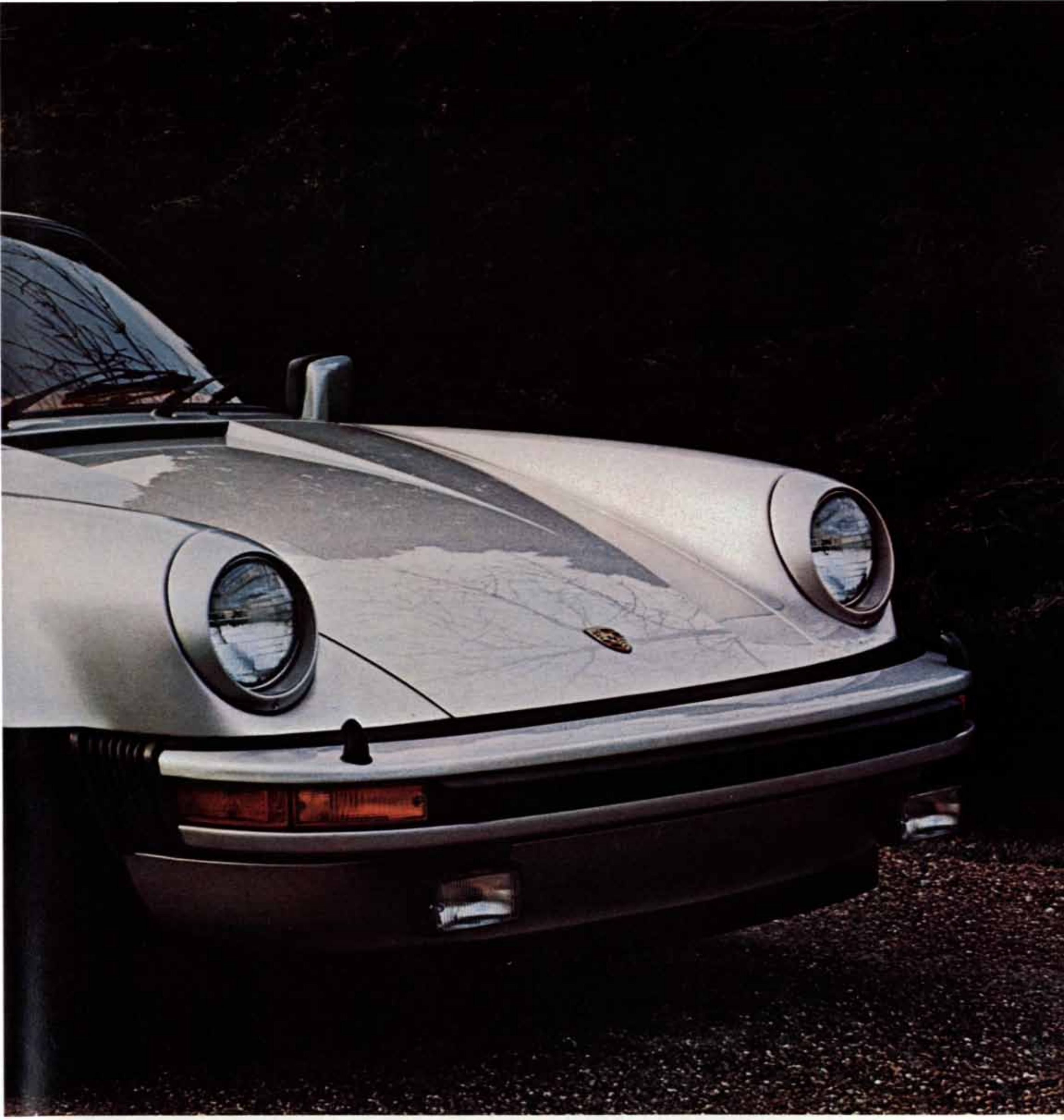
The fastest production Porsche ever made. Zero to sixty in under 6 seconds.

Built to such exacting standards that fewer than 600 people can own one this year.

The Porsche Turbo Carrera.

At its heart is the only turbocharged powerplant available to the American driver as standard equipment. Simply put, the turbocharger (driven by the engine's exhaust gases) forces more air into the engine than it could normally breathe. Which yields dramatically increased horsepower without attendant increases in fuel consumption or pollutants.

Yet, for all its astounding power, the Turbo Carrera is exceedingly well-mannered in traffic. It may be handled quietly and unobtrusively in any situation. Still, even at rest,



Porsche Turbo Carrera.

you can sense its immense potential. Such performance is enjoyed from a splendid vantage point. The inside of the Turbo Carrera.

There is, in a word, leather. It is supple and rich and it surrounds you. There is sumptuous cut-pile carpeting, air conditioning (and an automatic temperature control) and a four-speaker stereo radio with tape deck. The windows, radio antenna and headlight washers are electrically operated, as is the side-view mirror which de-ices and adjusts electrically.

This, then, is touring in the grand manner.

This is power hitherto unknown and a new standard of elegance.

So, for those who have always considered Porsche the ultimate sports car, we give you the new standard.

Turbo Carrera.

The ultimate Porsche.

Established 1845 SCIENTIFIC AMERICAN March 1976 Volume 234 Number 3

The Metabolism of Alcohol

Overconsumption of alcohol can cause cirrhosis and death not only because alcoholism promotes malnutrition but also because alcohol and its products disturb liver metabolism and damage the liver cells

by Charles S. Lieber

For all the attention being directed toward heroin, cocaine and marijuana, the favorite mood-altering drug in the U.S., as it is in almost every human society, is alcohol. Its psychic effects, both pleasant and unpleasant, are well enough known. What is less well known is that alcohol, in different quantities for different people, is a toxic drug: its overconsumption taxes the body's economy, produces pathological changes in liver tissue and function and can cause disability and death. As the incidence of alcoholism has risen in the U.S. population, so has the incidence of cirrhosis of the liver, which in 1974 climbed past both arteriosclerosis and influenza and pneumonia to rank seventh among the leading causes of death; in some urban areas (including New York City) it is actually the third most frequent cause of death between the ages of 25 and 65. Such a disease deserves intensive study. It is only in recent years that the changes in liver tissue and liver function have been directly associated with specific steps in the metabolism of alcohol, for the first time giving some hope that rational methods can be developed for preventing and treating alcoholic liver disease.

What we call alcohol is of course ethyl alcohol, or ethanol (CH_3CH_2OH). On the earth ethanol is probably almost as old as life itself. Whereas for man it is something to be ingested, it is a waste product for yeasts, the microorganisms that manufacture it. Ethanol is what is left over after fermentation, the process by which yeasts, through the agency of their particular enzymes, derive energy from various vegetable sugars. Man probably encountered ethanol during prehistoric times in the form of naturally fermented fruit juice (wine), honey (mead) and malted grain (beer).

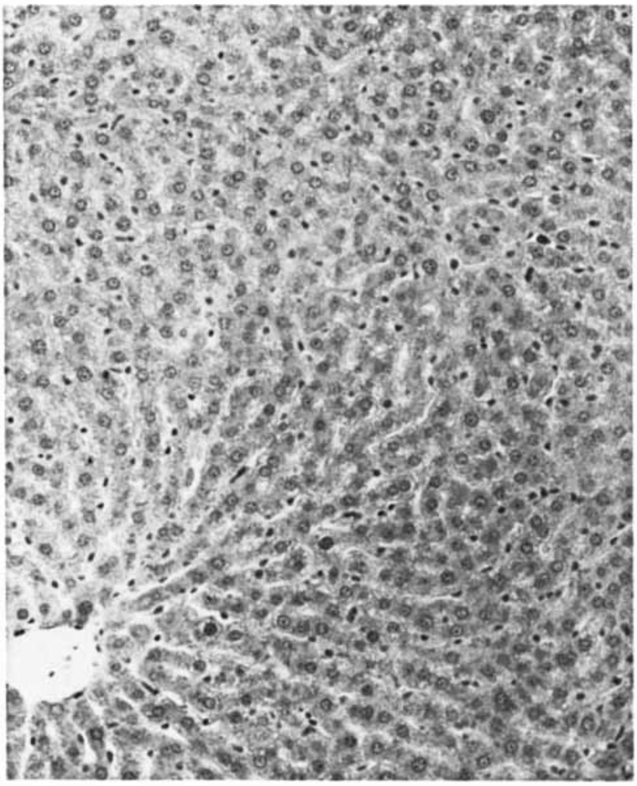
At least since the anatomical studies of Vesalius in the 16th century it has been recognized that heavy ingestion of alcohol is associated with diseases of several tissues, notably those of the liver. Until recently, however, such disease was attributed not to alcohol itself but to the malnutrition that is so often an accompaniment of excessive drinking. Alcohol was considered not a drug but a special kind of food with interesting psychotropic effects, and one that was metabolized by the body like any other energy-rich food. As recently as 1949 the distinguished physiologist Charles H. Best and his colleagues wrote that alcohol's metabolic contribution was simply to supply calories and that "there is no more evidence of a specific toxic effect of pure ethyl alcohol upon liver cells than there is for one due to sugar." Perhaps it was wishful thinking on the part of people in general and physicians in particular that installed as accepted fact the concept that alcohol lacked intrinsic toxicity.

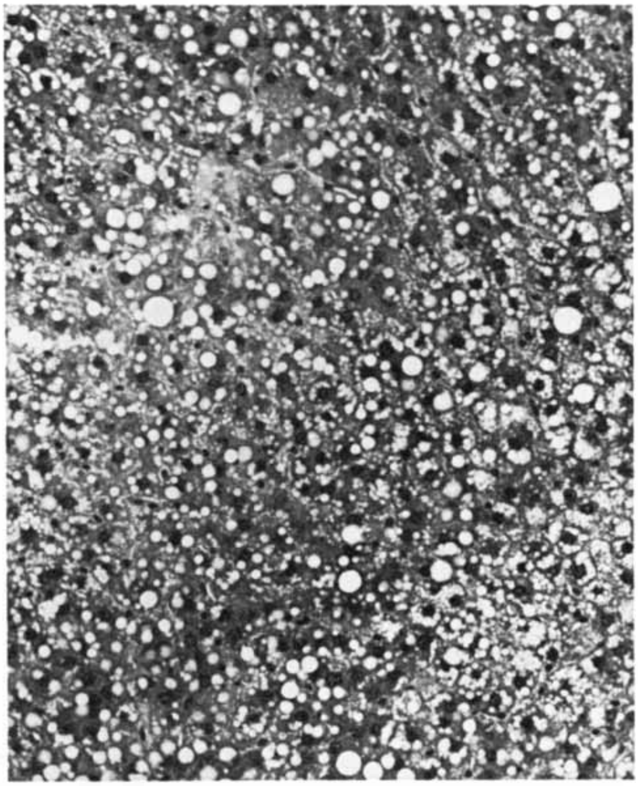
It was a rather naïve belief. Ethanol has little in common with other energy-rich compounds. Carbohydrates and fats can be synthesized in the body as well as ingested in the diet, but alcohol is essentially foreign to the body. As do carbohydrates and fats, alcohol has a high caloric value and is readily absorbed from the gastrointestinal tract; unlike them, however, it is not effectively stored in the tissues. Moreover, very little alcohol can be disposed of through the lungs or the kidneys, so that the body can get rid of alcohol only by oxidizing it. Again unlike fat and most carbohydrates, which can be oxidized by almost all tissues, alcohol must be oxidized in the liver, the organ that contains the bulk of the enzymes necessary for initiating the process. The organ-specificity of alcohol explains to a large extent the concentration of so many of alcohol's deleterious effects in the liver, which is the body's main chemical plant, the primary site of metabolic processes ranging from the synthesis of proteins to the detoxification of drugs and other foreign substances.

The purpose of this article is to relate how our studies at the Bronx Veterans Administration Hospital and the Mount Sinai School of Medicine of the City University of New York demonstrated that alcohol's toxicity for the liver is independent of poor diet and have shown how specific events in the metabolism of alcohol are linked to specific biochemical and structural changes in the liver and other tissues.

There are reasons enough, both socio-economic and physiological, for an alcoholic to suffer from malnutrition. The alcoholic spends his time, money and energy drinking and may tend to disregard such routine tasks as the preparation of meals and even eating. Unlike other drugs, alcohol has a high caloric value: 7.1 calories per gram, so that a pint of 86-proof spirits (not an unusual daily intake for an alcoholic) represents about half of the daily caloric requirement. As a result the appetite for food may be diminished. Ethanol calories, however, are "empty calories," quite unassociated with nutritive proteins, minerals and vitamins. Moreover, alcohol directly enhances malnutrition. By causing inflammation of the stomach, pancreas and intestine it can impair the digestion of food and absorption of nutrients into the blood; malnutrition can in turn cause the intestine to function still less effectively. Finally, ethanol and its primary conversion product, acetaldehyde, can interfere with the activation of vitamins by liver cells. As a result of all these effects malnutrition is common in alcoholics, and malnutrition alone does unquestionably impair liver function, as is observed in experimental animals that are fed severely deficient diets.

It was therefore quite natural to speculate





FATTY LIVER was produced in laboratory rats by alcohol in spite of an adequate diet. Liver-tissue sections from two rats are enlarged 240 diameters in these photomicrographs. One rat (*left*) was fed a liquid diet without alcohol. The other (*right*) had a liquid diet containing ethanol as 36 percent of total calories. After 24 days the micrograph of the ethanol-fed rat's liver shows numerous globular fat droplets.

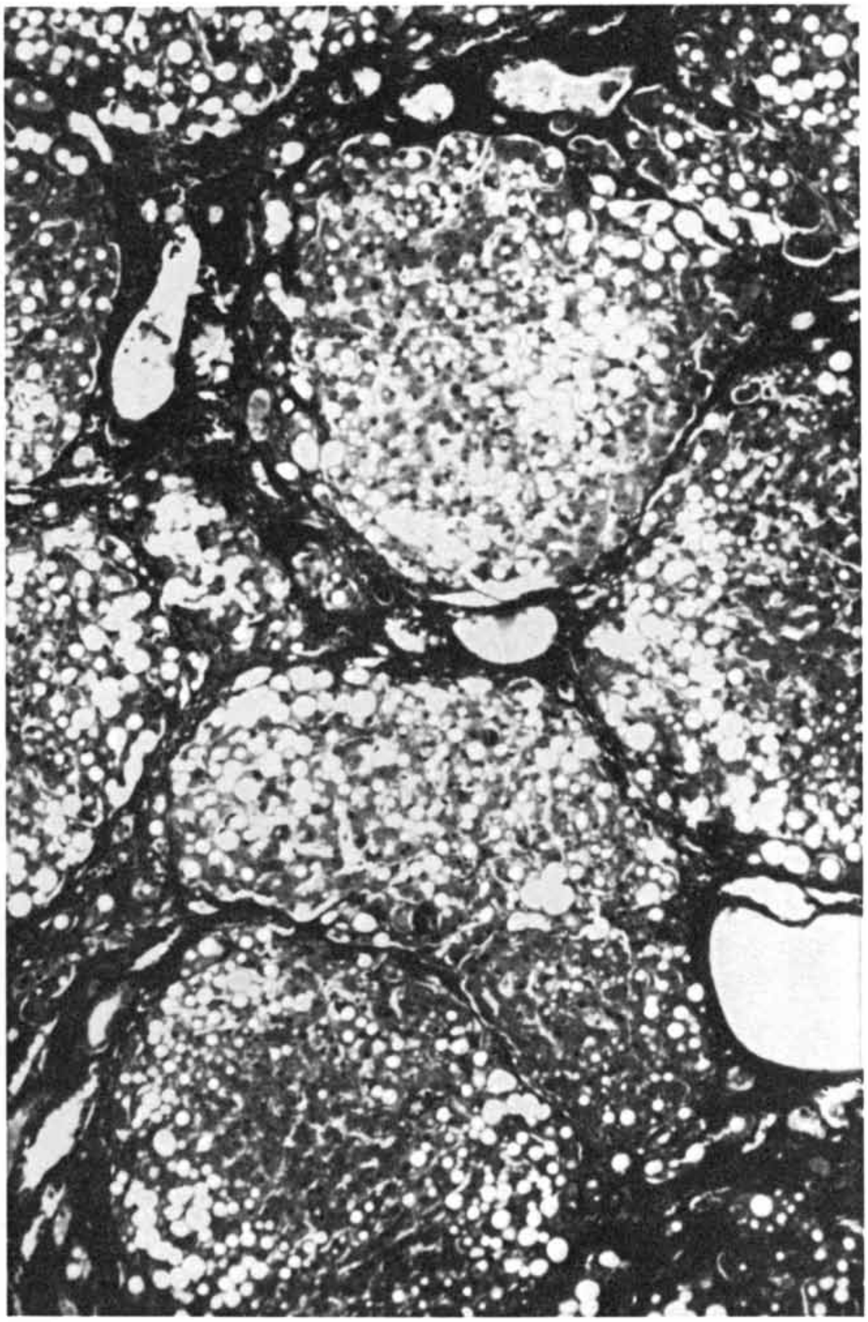
MICROSTRUCTURAL CHANGES are revealed in electron micrographs made by Oscar A. Iseri and the author, in which part of the liver cell is enlarged 16,000 diameters. In the control liver (*left*) the gray organelles with infolding membranes (cristae) are normal mitochondria; between them the parallel arrays of ribosome-studded rough endoplasmic reticulum are seen, along with some vesicular (sac-shaped) smooth endoplasmic reticulum. In the liver of the ethanol-fed rat (*right*) the mitochondria are swollen and distorted; in some mitochondria the external membrane and cristae are disrupted. There is marked proliferation of the smooth endoplasmic reticulum.

that malnutrition might be the cause of alcoholic liver disease even if one recognized that alcohol was not a food. The fact remains that in medical practice one encounters alcoholics who develop liver disease in spite of what is apparently an adequate diet. In the late 1950's I therefore began to wonder whether alcohol might exert direct toxic effects on the liver. Such a finding would clearly be of therapeutic as well as theoretical interest. Belief in the exclusively nutritional theory was leading many physicians to advise alcoholic patients that normal liver function could be maintained in spite of heavy alcohol consumption as long as the diet was adequate. If that was not true, it would be good to know it.

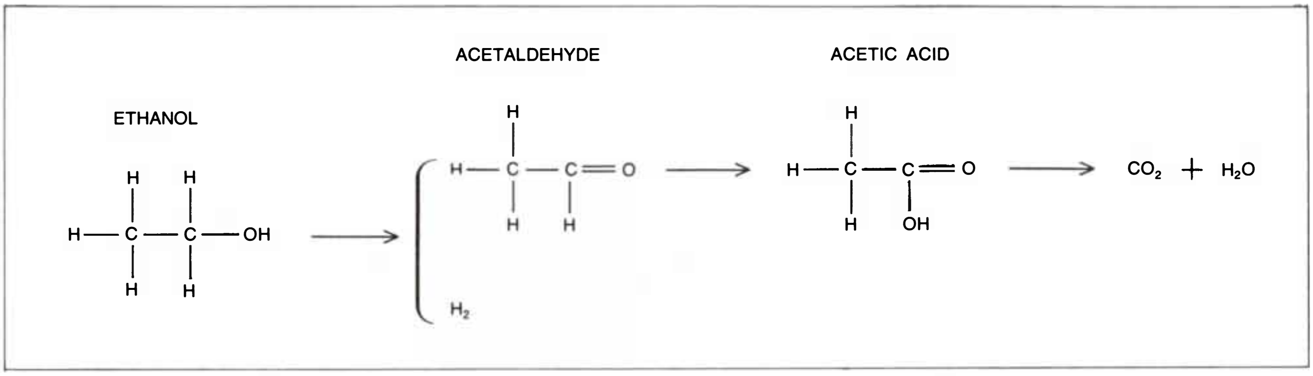
The importance of the issue, along with our confidence that even heavy consumption of alcohol carries a negligible risk if it is only for a brief period and under supervised conditions, led to a decision to study the effect of ethanol directly in volunteers. We fed the volunteers a nutritionally optimal low-fat diet, in which protein accounted for 25 percent of the calories, or two and a half times the recommended percentage. That meant beef for breakfast, cheese and fish for lunch and meat or fowl for dinner, with supplementary vitamin and mineral capsules. The volunteers also had six drinks a day, a total of about 10 ounces of 86-proof alcohol. On that regimen there was a progressive rise in liver fat; routine thin-needle biopsy revealed significant fat accumulation after a few days, and there was an average eightfold increase over the 18-day course of the experiment. Here was the characteristic "fatty liver," the first (and typically reversible) stage of liver disease. There were striking changes in the ultrastructure of the liver cell: the mitochondria, the energy-converting organelles, were enlarged and distorted, and the smooth membranes of the endoplasmic reticulum, the site of enzymes associated with the metabolism of alcohol and other substances, proliferated. These liver changes were brought about by a rather moderate ingestion of ethanol, which did not result in any clinical signs of intoxication; the blood levels peaked at from 80 to 90 milligrams of alcohol per 100 milliliters of blood, which is below the levels (100 or 150 milligrams) that constitute evidence of excessive drinking in most states. Drunkenness is no prerequisite for the development of liver damage.

In order to verify these effects and study their mechanisms in detail we needed to reproduce them in an experimental animal. The laboratory rat is a common subject for such experiments. Ethanol is usually fed to rats in their drinking water, however, and under these circumstances the animals generally refuse to take enough ethanol to develop liver injury. Leonore M. DeCarli and I countered the rat's aversion to alcohol by developing a new technique: we incorporated the ethanol in a nutritionally adequate but totally liquid diet, so that in order to eat or drink the animals also had to take the alcohol. With this method we produced fatty liver in rats in spite of an adequate and fully controlled diet. The changes in the liver were comparable to those produced in man, both on gross examination and when tissue was viewed with the light microscope [*see top illustration on opposite page*]. The ultrastructural changes seen with the electron microscope were also similar, including enlargement and disruption of the mitochondria and proliferation of the smooth endoplasmic reticulum, part of the system of involuted membranes within the cell [*see bottom illustration on opposite page*].

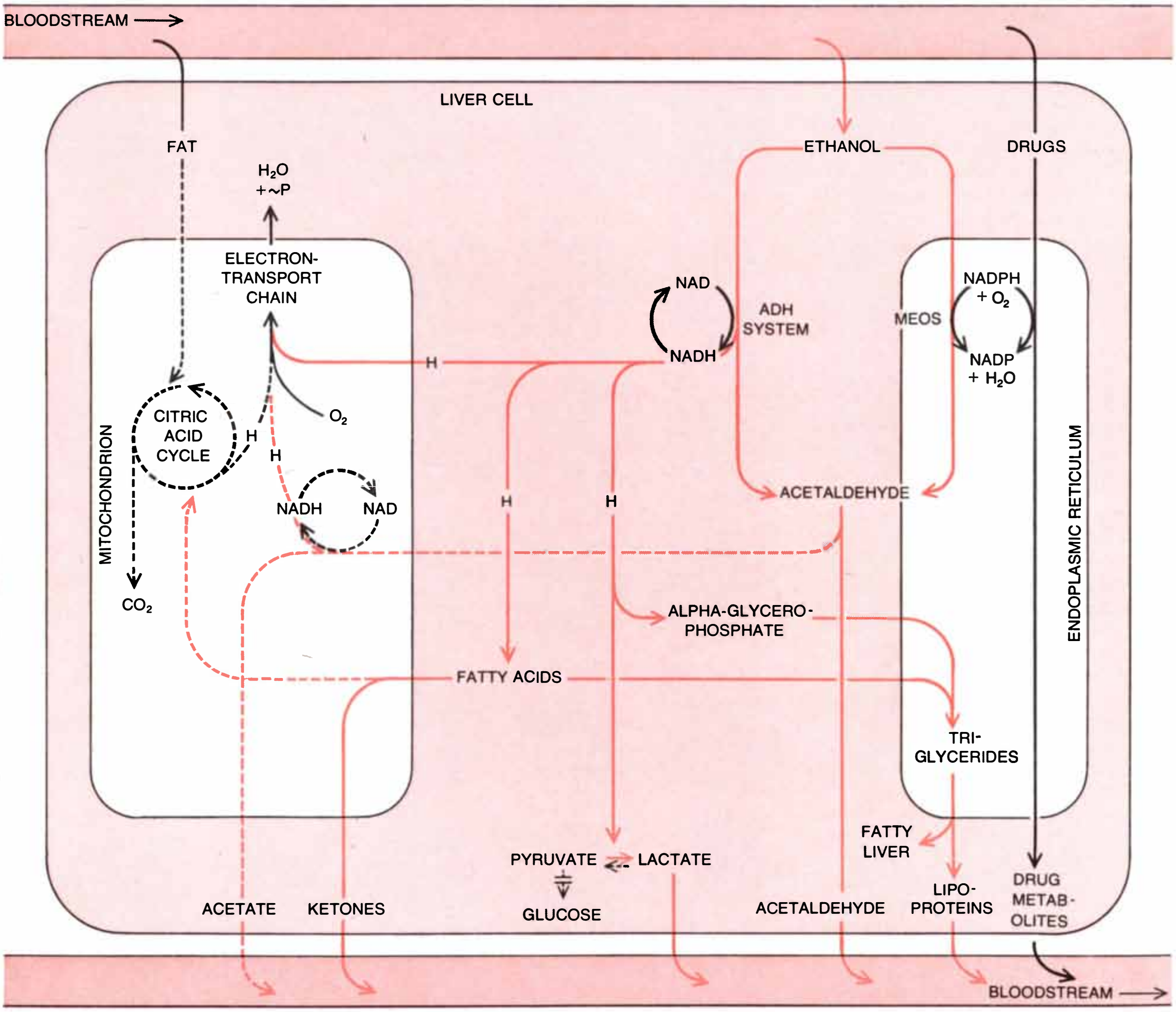
Although our alcohol-fed rats developed a fatty liver, they did not go on to incur the more severe forms of alcoholic liver injury seen in human beings. The first of these stages is alcoholic hepatitis, in which decreased liver-cell function leads to the death of cells, inflammation and a mortality rate of from 10 to 30 percent. The final stage is cirrhosis, when fibrous scars disrupt the normal architecture of the liver and give rise to a number of potentially fatal compli-



CIRRHOSIS OF LIVER was produced in a baboon after four years of heavy alcohol consumption. This section of the baboon's liver, enlarged 70 diameters, shows the broad strands of fibrous connective tissue, like scar tissue, that have disrupted the once orderly arrays of liver cells, segregating nodules of irregularly arranged cells and heavily concentrated globules of fat.



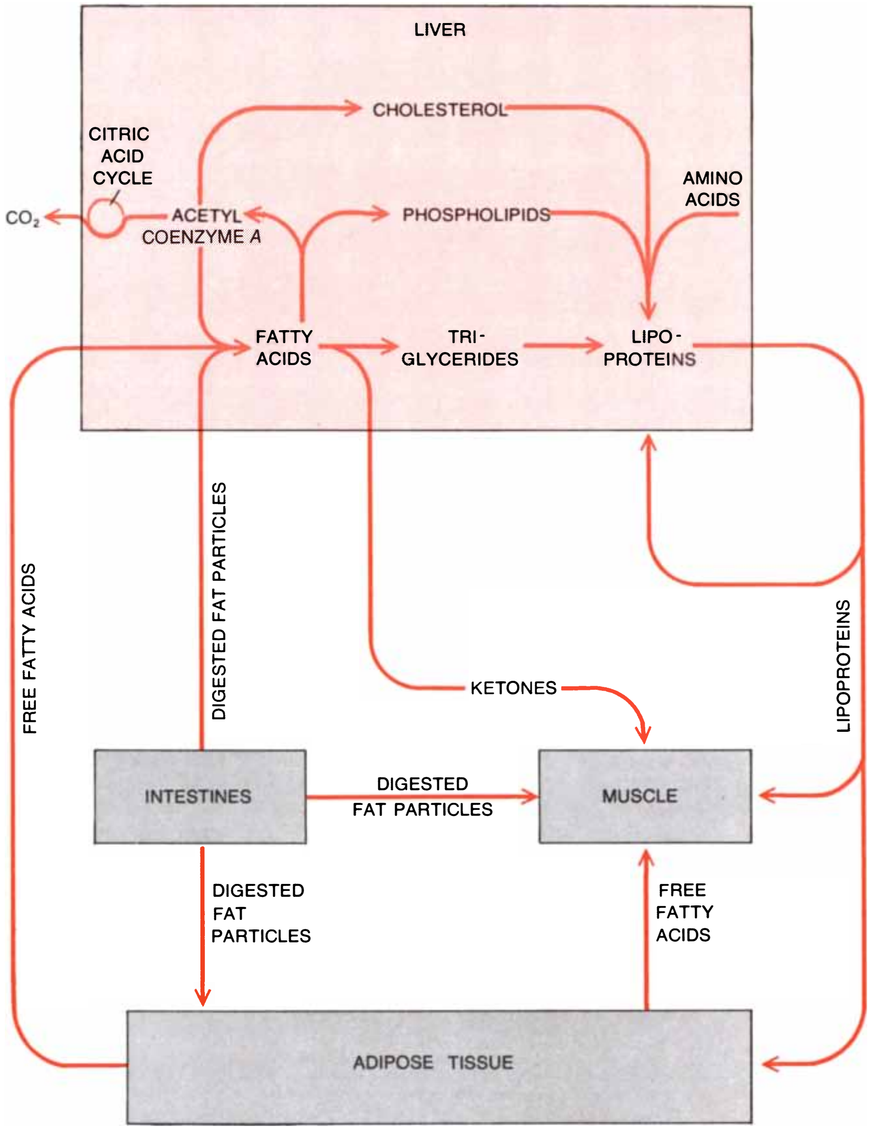
OXIDATION OF ETHYL ALCOHOL (ethanol) is accomplished in several stages. In the liver cells two hydrogen atoms are removed from each molecule of ethanol to form acetaldehyde. The acetaldehyde is normally oxidized, primarily in the liver, to form acetic acid (as acetates), which is eventually broken down into carbon dioxide and water. The illustration omits the various enzymes and cofactors.



METABOLISM OF ETHANOL in the liver cell by the alcohol-dehydrogenase (*ADH*) system and the microsomal ethanol-oxidizing system (*MEOS*) is represented schematically, with the two ethanol pathways and the movements of their products shown in color. Pathways that operate in the absence of ethanol are shown in black and those whose activity is decreased by ethanol are shown by broken lines. In the primary ethanol pathway alcohol dehydrogenase catalyzes the removal of hydrogen atoms (*H*) through reduction of a cofactor, nicotinamide adenine dinucleotide (*NAD*). The excess hydrogen is shunted into various processes. For example, it supplies the energy-producing electron-transport system, supplanting fats, the normal fuel. Hydrogen also contributes to the synthesis of excess triglycerides, or fats. A secondary pathway that comes into play at high alcohol levels utilizes the microsomal system, which also metabolizes certain drugs and other foreign compounds. Here the cofactor is reduced nicotinamide adenine dinucleotide phosphate (*NADPH*).

cations. We could imagine two reasons for the rats' failure to develop hepatitis and cirrhosis. Even with our liquid diet their alcohol intake did not exceed 36 percent of the total calories, which corresponds to moderate consumption in man. Moreover, it takes from five to 25 years of steady drinking for a human being to develop cirrhosis, and a rat lives for only about two years.

We turned to the baboon, which is long-lived and more closely related to man, and which can tolerate liquid diets containing alcohol as 50 percent of the total calories, about the same as an alcoholic human being. At that level the baboons were visibly intoxicated, and when alcohol was temporarily withdrawn, they gave signs of physical dependence, such as tremors and seizures. At the Laboratory for Experimental Medicine and Surgery in Primates we put 16 baboons on the ethanol-rich diet and 16 others on an alcohol-free control diet with the same caloric content. Whereas the liver remained normal in the control animals, all 16 animals given ethanol soon showed excessive fat accumulation in the organ; five of them showed typical alcoholic hepatitis, and six of the animals kept on the ethanol diet developed cirrhosis after between two and four years. The rat and baboon experiments demonstrated clearly that severe liver injury can be produced by prolonged heavy ingestion of alcohol even when the diet is good. They also gave us for the first time experimental animal models that reproduce all the liver lesions seen in the alcoholic human being. The models have enabled us to find biochemical explanations for the early stages of alcoholic liver disease (fatty liver and the metabolic disturbances associated with it) and for some of the persistent lesions that seem to lead to the more severe stages: hepatitis and cirrhosis.



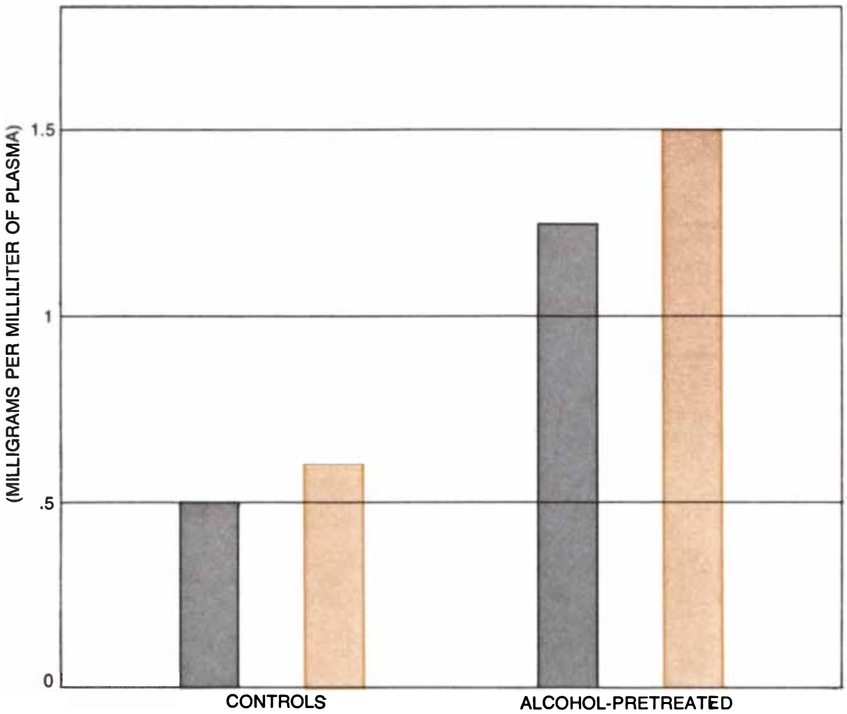
LIVER FAT comes from three sources: from the diet, by way of the intestines; in the form of free fatty acids mobilized from storage in adipose tissue, and through synthesis within the liver itself. A fatty liver can develop if there is an excessive supply of fat or if disposal of fat from the liver (by oxidation to carbon dioxide or by secretion into the blood as lipoproteins) is impaired.

The first step in alcohol's primary metabolic pathway is catalyzed by the enzyme alcohol dehydrogenase. (In spite of its name, alcohol dehydrogenase serves as a generalized remover of hydrogen atoms from various compounds, including some steroids; it was therefore probably present in the liver of the prehistoric men who first sampled alcohol, available to take on what has since become a major function.) Alcohol dehydrogenase catalyzes the transfer of hydrogen atoms from ethanol to a cofactor, nicotinamide adenine dinucleotide (NAD), converting the ethanol into acetaldehyde. The acetaldehyde is then oxidized, primarily in the liver, to form acetate, which is eventually converted into carbon dioxide and water. A number of the metabolic effects of alcohol are directly linked to the two first products of its oxidation: hydrogen and acetaldehyde.

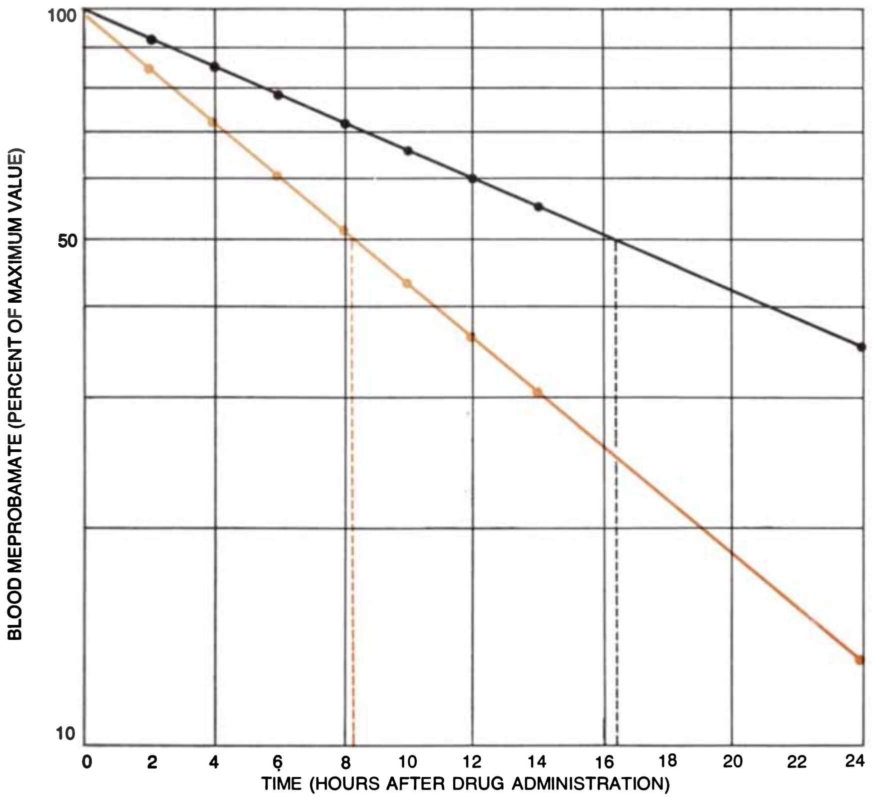
The excess hydrogen from alcohol unbalances the liver cell's chemistry. In order to live the cell must get rid of the hydrogen, and it does so by shunting hydrogen ions into one or more of several pathways dependent on them, sometimes with deleterious effects. One such pathway is the process whereby amino acids (derived from the breakdown in the liver of proteins) are converted, with pyruvate as an intermediate, into glucose. In the presence of excess hydrogen ions the process is turned in a different direction: the pyruvate is reduced to lactate instead of being converted into glucose. Blood sugar is derived primarily from three sources: gluconeogenesis, or synthesis from amino acids in the liver, breakdown of glycogen stored in the tissues and conversion of carbohydrates in the diet. If the alcoholic person has been drinking and not eating, there are no dietary carbohydrates and glycogen may be used up; if gluconeogenesis is then blocked by the diversion of pyruvate to lactate, the level of sugar in the blood will be lowered. Low blood sugar, or hypoglycemia, is known to be a complication of acute alcoholism, but it is often overlooked. When an intoxicated person is brought into a hospital emergency room, it is important to test for hypoglycemia, since crucial organs, including the brain, can be critically affected by a lack of sugar; some deaths of "unknown origin" in alcoholics may be attributable to the condition.

The lactate buildup from excess hydrogen has other effects. The lactate moves into the blood, bringing on lactic acidosis. In the kidney it interferes with the excretion of uric acid. A high uric acid level in the blood (hyperuricemia) exacerbates gout, so that the process I have described may explain the ancient clinical observation that excessive drinking can trigger or aggravate gout attacks.

There are other ways in which the liver cell can rid itself of alcohol's excess hydro-



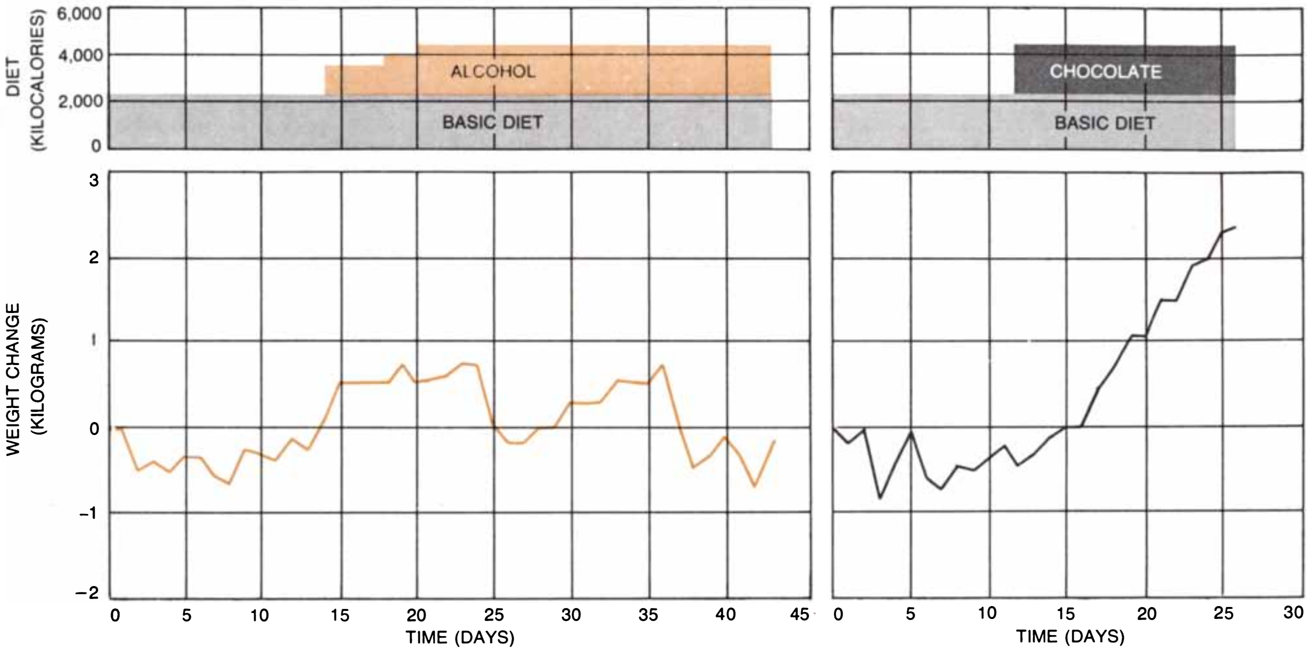
ALCOHOL-FED RATS develop a tendency toward hyperlipemia, or high blood fat. Animals were fed either the control diet or the alcohol-containing diet of the same caloric value for about a month. Then, after fasting, the animals were given a "load" of one or the other diet. The rats' fasting blood-lipid levels had been about the same. In response to the load the animals that had been chronic alcohol consumers (*right*) developed markedly higher blood-lipid levels, whether the load contained alcohol (*colored bars*) or did not contain alcohol (*gray bars*).



ALCOHOL ACTIVATES the enzyme systems of the smooth endoplasmic reticulum, stepping up the metabolism of certain drugs. The curves show the disappearance from the blood of a dose of the tranquilizer meprobamate given to volunteers before (*black*) and after (*color*) a month of alcohol ingestion. Alcohol activation cuts the drug's half-life (*broken lines*) in half.

gen. Several of them involve the formation of lipids, or fat. The hydrogen can be shunted directly into the synthesis of alpha-glycerophosphate and fatty acids. Those are the two precursors of the triglycerides, and triglycerides are the lipids that accumulate in the alcoholic fatty liver. The main mechanism for disposing of hydrogen is more indirect, but it has a similar result. The hydrogen is transferred to the mitochondria, the cell organelles that produce the energy for liver functions. Normally it is fat that is oxidized—in effect burned—in the mitochondrial citric acid cycle to produce usable energy in the form of phosphate ions in an energized state. The plentiful hydrogen from alcohol, however, provides an alternate fuel that is oxidized instead of the hydrogen from fat. In promoting the oxidation of alcohol the substitution exacts a price: the lipids accumulate, again leading to fatty liver. If alcohol is ingested along with a diet containing fat, the fat of dietary origin accumulates in the liver; even when alcohol is taken with a low-fat diet, fat made in the liver itself is deposited there. In addition, when alcohol is ingested in very large quantities, it can trigger hormonal discharges that mobilize fat from stores of adipose tissue and move it toward the liver [*see illustration on preceding page*].

What can the liver do with the accumulating fat? One possibility is for it to be secreted into the bloodstream, which delivers blood lipids to provide fuel for peripheral tissues such as muscle and deposits extra supplies for storage in adipose tissue. The secretion is complicated by the fact that lipids are not soluble in water; they have to be made soluble by being wrapped in a thin coat of protein to form lipoproteins. The assembly of lipoproteins is carried out in the liver in the smooth membranes of the endoplasmic reticulum. I have alluded to the proliferation of the smooth endoplasmic reticulum that comes with heavy alcohol consumption in both rats and human beings, which I observed with Oscar A. Iseri, Bernard P. Lane and Emanuel Rubin. The proliferation is reflected, Jean-Gil Joly, Lawrence Feinman and I found, in the increased activity of certain enzymes in the smooth reticulum, which enlarges the liver's capacity for secreting lipoproteins. A liver that has thus adapted to alcohol after being conditioned by a period of heavy alcohol intake will respond with exaggerated secretion of lipoproteins even after a normal meal is eaten, producing hyperlipemia, or an abnormally high level of fat in the blood. Enrique Baraona and I observed that effect in rats, and William H. Perlow, Stephen A. Borowsky and I have now verified it in man. The effect is of particular significance in people who have underlying abnormalities of either lipid or carbohydrate metabolism and therefore a propensity for elevated blood-lipid levels. Hyperlipemia is a major predisposing factor for heart attacks. What is usually overlooked is that in individuals with preexisting hyperlipemia alcohol is proba-



ALCOHOL CALORIES are apparently not fully equivalent to other calories. When a basic 2,200-calorie diet was supplemented by 2,000 calories of alcohol (*left*) and of chocolate (*right*), the gain in weight caused by the alcohol was much less marked and more erratic.

bly the single most important aggravating factor that is amenable to correction.

Still another way for the liver to dispose of excess fat is to convert some of it into the water-soluble breakdown products called ketone bodies and secrete them into the bloodstream. In some susceptible people this response may be exaggerated, resulting in an elevation of ketone bodies in the blood that mimics the condition known as ketoacidosis in diabetic patients.

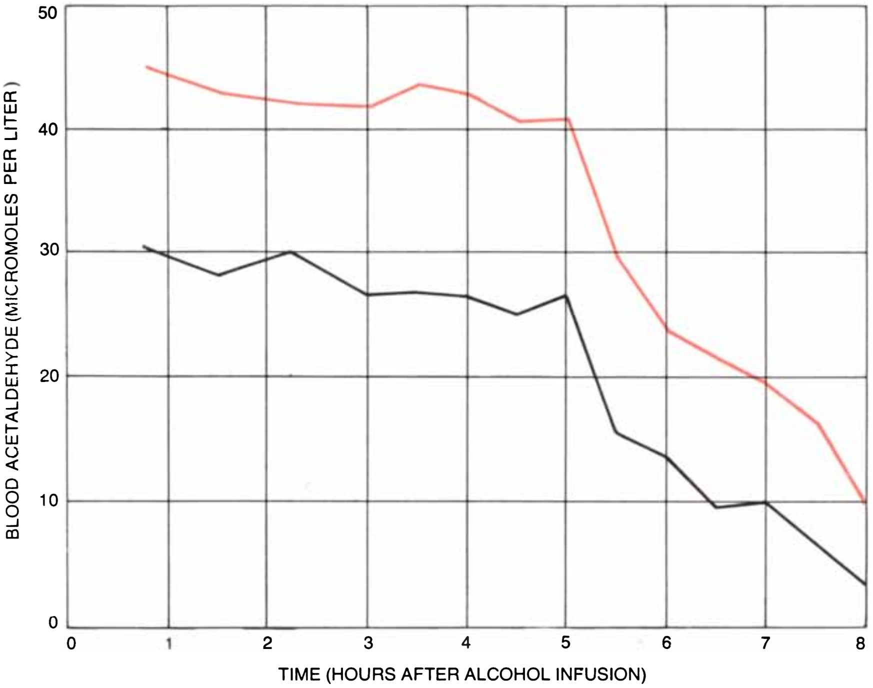
The conversion of lipids into lipoproteins is just one of a number of functions of the endoplasmic reticulum of liver cells, which also inactivates a wide variety of drugs and other foreign substances and converts them into water-soluble products that can be excreted [see "How the Liver Metabolizes Foreign Substances," by Attallah Kappas and Alvito P. Alvares; SCIENTIFIC AMERICAN, June, 1975]. I therefore wondered whether the proliferation of smooth reticulum after chronic alcohol consumption would be reflected in an increased capacity of the alcoholic liver to metabolize various drugs. That turned out to be the case. After repeated administration of alcohol (but before the evolution of severe liver injury) the enzymes of the smooth reticulum that inactivate tranquilizers, anticoagulants and other drugs and that detoxify certain food additives, cancer-causing substances and insecticides do increase their activity, enhancing the body's capacity to rid itself of these compounds. For example, Prem S. Misra and I gave the tranquilizer meprobamate to volunteers and measured its rate of disappearance from the bloodstream. The time for the blood level of the drug to fall to half its original value decreased from 16 hours during a control period to eight hours after a month of alcohol consumption.

Anesthesiologists had known for many years that larger doses of sedatives are required to achieve a given effect in alcoholics than in other people. That drug tolerance was attributed to adaptation in the central nervous system: increased resistance to sedatives by the brain of the alcoholic. Our finding pointed to a metabolic adaptation as well: increased capacity of the alcoholic liver to inactivate and excrete sedatives and other compounds detoxified by the endoplasmic reticulum. At the adaptive stage of their disease (which, as we shall see, does not continue indefinitely) alcoholics therefore require larger doses of many drugs. That is true, however, only when the individual is sober. When the alcoholic has been drinking, the effect is quite the opposite. The reason, we found, is that one of the drugs the smooth reticulum metabolizes is alcohol itself, by way of an accessory pathway that supplements the basic alcohol-dehydrogenase system.

DeCarli and I demonstrated the accessory pathway by spinning liver tissue in the ultracentrifuge and isolating the endoplasmic reticulum as what is called the microsomal fraction. We then found that a preparation of the microsomal fraction would oxidize ethanol. We called the accessory pathway the microsomal ethanol-oxidizing system, and Rolf Teschke, Kunihiko Ohnishi and I were able to obtain it in a semipurified form. We have found that this microsomal ethanol pathway comes into operation after the blood alcohol reaches a certain level. The alcohol thus enters into competition with other drugs whose metabolism shares some elements of the microsomal system, thereby slowing down the metabolism of those drugs and enhancing their effect. That is why simultaneous drinking and taking of tranquilizers is particularly dangerous: the alcohol can accentuate the action of the drug not only because the effect of the two drugs on the brain may be additive but also because the presence of alcohol can interfere with the liver's capacity to inactivate the drug, so that at a given dosage more of the drug remains active for a longer time.

Along with the rest of the microsomal enzyme systems, the ethanol-oxidizing system adapts to heavy alcohol consumption by increasing its activity, thus contributing to alcohol tolerance. The alcoholic's ability to drink more than most nonalcoholics is primarily due to brain tolerance, a progressive decrease in the effectiveness of alcohol's action on the brain. The alcoholic also develops an increased capacity to metabolize alcohol, not only through the microsomal system but also through the alcohol-dehydrogenase pathway and perhaps a third pathway that depends on the enzyme catalase. After heavy or prolonged drinking, however, this adaptation can be offset by progressive liver injury, so that the liver's overall ability to handle alcohol remains about the same or even decreases.

The microsomal changes I have described are beneficial in that they help to rid the liver of fat and speed the detoxification of many drugs, food additives and other foreign compounds. Hyperlipemia is one undesirable concomitant of the adaptive process, and there are others. Some foreign substances are activated rather than inactivated by the conversions they undergo in



ACETALDEHYDE LEVEL IN BLOOD is higher in alcoholics than in nonalcoholic individuals. Identical quantities of pure ethanol were infused into the blood of alcoholic and nonalcoholic volunteers, achieving similar blood-alcohol levels. The acetaldehyde plateau was higher in the alcoholics (*colored curve*) than in the nonalcoholics (*black*). Apparently the alcoholics metabolized acetaldehyde less effectively, perhaps because of ethanol-induced liver damage.

the endoplasmic reticulum. Certain potentially cancer-causing substances become carcinogenic only after activation by the microsomes, and other substances become toxic to the liver itself only after such activation. For example, exposure to carbon tetrachloride is known to cause liver damage, but the familiar dry-cleaning agent is harmless to the liver until it is activated by the liver-cell endoplasmic reticulum. Yasushi Hasumura and I found that the increase in microsomal activity induced by alcohol enhances the toxicity of carbon tetrachloride: rats chronically treated with alcohol were much more susceptible to the toxic effects of carbon tetrachloride than matched control animals. This effect presumably explains the clinical observation that alcoholics are particularly susceptible to carbon tetrachloride poisoning in dry-cleaning plants that still use the compound. The enhanced susceptibility of alcoholics very probably extends to a number of other foreign compounds that are harmless to most people but that may become toxic in the alcoholic as an undesirable side effect of increased microsomal activity.

Another side effect is energy waste. Microsomal activity requires energy. Moreover, it is a peculiarity of the various microsomal oxidations that they produce heat without conserving chemical energy. This could conceivably contribute to the impaired growth we observe in animals that are fed ethanol, since the production of heat beyond what is required to maintain body temperature represents a waste of energy. Such waste may explain at least in part my observation with Romano C. Pirola that the addition of ethanol to the diet results in less weight gain than the addition of the same number of calories from other sources [*see illustration on preceding page*]. That is, with respect to body weight ethanol calories apparently do not fully "count," at least in alcoholics and when the alcohol intake is large.

Apart from the metabolic complications of excess hydrogen and of the adaptive changes in microsomal activity, heavy alcohol consumption has direct toxic effects on liver tissue. It appears that a central role in the toxicity of alcohol may be played by acetaldehyde, a product of alcohol's several metabolic pathways that is extremely reactive and that affects most tissues in the body. Most of the acetaldehyde is converted into acetate by the liver mitochondria, but some of it escapes into the bloodstream. Both the alcohol dehydrogenase and the microsomal ethanol-oxidizing pathways become saturated when the liver is loaded with a large amount of alcohol, so that the acetaldehyde level in the blood reaches a plateau and stays there until the alcohol level drops, at which point the microsomal pathway apparently cuts out. Mark A. Korsten, Shohei Matsuzaki, Feinman and I found that the acetaldehyde plateau is significantly higher in alcoholics than it is in nonalcoholics, even when the same amount of alcohol has been given to both groups and the same blood level of alcohol has been attained [*see illustration at left*]. This abnormally high acetaldehyde level in alcoholics could result from faster metabolism of alcohol to form acetaldehyde (because of the adaptive increase in microsomal activity) or from slower disposition of the acetaldehyde (because of impaired acetaldehyde metabolism).

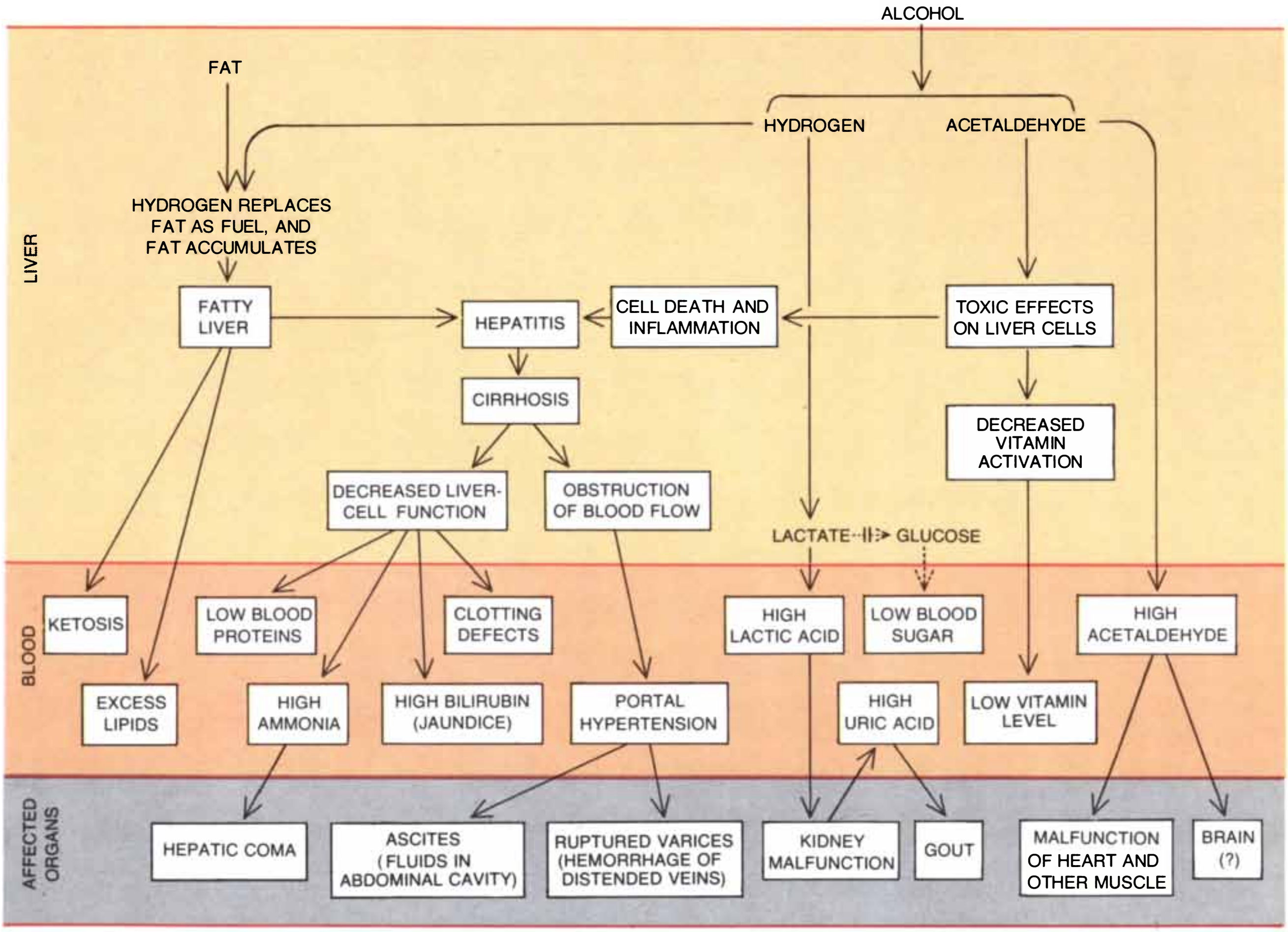
Probably both factors are involved, at least in the beginning. I have mentioned the striking alterations in the mitochondria revealed by the electron microscope even in the early stages of heavy alcohol consumption. Hasamura and I isolated the damaged mitochondria and found that their capacity for metabolizing acetaldehyde to acetate is reduced. Acetaldehyde itself may be responsible for part of the decrease in mitochondrial function: Arthur I. Cederbaum, Rubin and I showed that it has a toxic effect on the organelles. The alcoholic may therefore be the victim of a vicious circle: a high acetaldehyde level impairs mitochondrial function in the liver, acetaldehyde metabolism is decreased, more acetaldehyde accumulates and causes further liver damage. The acetaldehyde affects not only the liver but also other tissues. For example, Marcus A. Rothschild and his colleagues at the Manhattan Veterans Administration Hospital have reported that acetaldehyde levels no higher than we have measured in the blood of alcoholics can inhibit the synthesis of proteins in heart muscle. Such an effect could in part explain the impaired cardiac function that is common in alcoholics. Acetaldehyde may affect the functioning of other muscles too.

Acetaldehyde has striking effects on the brain. Several investigators have suggested that acetaldehyde rather than alcohol itself may be responsible for the development of the dependence that, along with tolerance, characterizes alcohol addiction. Dependence manifests itself by a state of extreme discomfort, often accompanied by physiological disturbances such as tremors and seizures, produced by withdrawal of a drug. Several mechanisms mediated by acetaldehyde have been proposed to explain dependence but none has yet been confirmed. One proposal begins with the fact that certain amine neurotransmitters, which transmit nerve impulses from one cell to another in the brain, are inactivated by the enzyme monoamine oxidase to form an aldehyde that is then converted into an acid. The conversion to acid requires an enzyme that is also active in the metabolism of acetaldehyde. Virginia E. Davis and Michael J. Walsh of the Houston Veterans Administration Hospital have suggested that if acetaldehyde is present in the brain, it may compete for the enzyme; unmetabolized neurotransmitter aldehydes may therefore accumulate; the aldehydes may then combine with the neurotransmitter, forming compounds that are startlingly similar to certain morphine derivatives known for their ability to promote dependence.

Another possibility, pointed out by Gerald Cohen of Mount Sinai, is that the acetaldehyde combines directly with amines to form isoquinoline derivatives, potent psychoactive compounds that could play a role in developing dependence. Alcohol dependence is probably determined by a number of factors acting in concert. It is at least possible, however, that acetaldehyde is involved in the predisposition to alcoholism: a predisposed person may have deficiencies of acetaldehyde metabolism that make for a higher blood acetaldehyde level and a propensity to dependence because of the effect of the higher acetaldehyde level on the brain.

If heavy drinking continues, reversible fatty liver evolves, in most people, toward more severe and irreversible liver disease: hepatitis and then cirrhosis. Just why and how the liver may lose its ability to adapt to the alcohol load is not clearly established. Even in the fatty-liver stage there are some indications that more severe lesions are developing. There may be a ballooning of the liver cells, which is usually attributed to fat accumulation. Baraona, Maria-Anna Leo, Borowsky and I have noted that proteins, of which the liver is the major synthesizer, also pile up in the cells, whose ability to export proteins is somehow depressed by large amounts of alcohol. The engorgement of liver cells with fat and protein interferes with their normal functioning. So does the reduced energy production to which I have alluded. For all these reasons some liver cells may die, and the necrosis can trigger an inflammatory process that is characteristic of the stage called alcoholic hepatitis. The acute reduction in liver-cell function at this stage is enough to cause death in some patients.

Cell necrosis and inflammation in turn promote the next stage: fibrosis, or the development of scar tissue, the hallmark of cirrhosis. The fibrous connective-tissue barriers between groups of liver cells interfere with the flow of blood to and from the cells, further decreasing liver function. Partially obstructed in the liver, blood backs up, increasing the pressure in the portal system, which brings blood to the liver from the intestines. Abnormal channels may therefore develop in the venous system so that some of the blood can bypass the circulatory block established by the cirrhotic liver. Some newly overloaded veins (notably in the esophagus) may become distended, and such veins, called varices, can rupture and hemorrhage. Bleeding varices are a major cause of death in cirrhosis. Moreover, under high pressure plasma leaks out of the portal-system blood vessels. Extra lymph is also formed, and it leaks out of the lymphatic vessels. Both sources contribute to ascites, the accumulation of fluid in the abdominal cavity. A final complication stems from the liver's inability to clear from the blood ammonia and other nitrogenous compounds produced by bacteria in the intestines. As such compounds accumulate they act on the brain and may cause functional disturbances, hepatic coma and death.



COMPLICATIONS of excessive alcohol consumption stem largely from excess hydrogen and from acetaldehyde. Hydrogen produces fatty liver and hyperlipemia, high blood lactic acid and low blood sugar. The accumulation of fat, the effect of acetaldehyde on liver cells and other factors as yet unknown lead to alcoholic hepatitis. The next step is cirrhosis. The consequent impairment of liver function disturbs blood chemistry, notably causing a high ammonia level, which can lead to coma and death. Cirrhosis also distorts liver structure, inhibiting blood flow. High pressure in vessels supplying the liver may cause ruptured varices and accumulation of fluid in abdominal cavity. There are individual differences in response to alcohol; in particular, not all heavy drinkers develop hepatitis and cirrhosis.

Polycyclic Aromatic Compounds in Nature

These multiple-ring hydrocarbon molecules have been found in soils and sediments around the world. They are unusually stable, and their origins have presented an intriguing puzzle

by Max Blumer

Polycyclic aromatic compounds are found in great variety in the natural world; among them are plant and animal pigments of notable beauty and unusual chemical stability. Alizarin, the red of military uniforms in Napoleon's day, belongs to this class of compounds; it exhibits its original brilliance in museum collections, where other plant pigments have long since faded. Deep-sea animals, particularly sea urchins and sea lilies, owe their vivid coloration to such pigments. Polycyclic aromatic compounds are also noted for their biological effects. Some of them can induce cancer or cause mutations, even in very low concentrations.

Just what are polycyclic aromatic compounds? They are one of the principal classes of substances where the central molecular structure is held together by stable carbon-carbon bonds. The great variety of carbon compounds—and of life itself, which is based on such compounds—reflects the almost limitless possibilities in the spatial arrangement of carbon atoms in groups, clusters, chains and rings. Long carbon chains are common in natural products; they have been recognized ever since organic chemistry became a separate discipline. Branched carbon chains are equally common and also have a long scientific history. The recognition in 1865 that carbon atoms can be linked in closed rings was one of the great conceptual advances of chemistry. The discovery was made by Friedrich Kekulé, who solved the riddle of the structure of the benzene molecule after having a dream in which dancing snakes bit their own tails.

From Kekulé's insight it was only a short step to the recognition that the naphthalene molecule has two fused rings and the anthracene and phenanthrene molecules have three. Such multiple-ring, or polycyclic, compounds are said to be saturated if all the bonds of the carbon atoms, beyond the minimum needed for carbon-carbon bonding, are linked to hydrogen atoms. They are called aromatic if some of the carbon atoms are doubly bonded to other carbon atoms. In Kekulé's day the rings of aromatic compounds were thought to have single and double bonds in alternation. That traditional but mistaken view survives in the common graphical representation of the hexagonal benzene ring. According to current views, the bonds that link the carbon atoms in the benzene molecule and its polycyclic analogues all play an equal role and have an equal value. The theory of chemical bonding helps to explain the exceptional stability of six-carbon-ring systems.

Polycyclic aromatic hydrocarbons consist of three or more fused benzene rings in linear, angular or cluster arrangements. By definition they contain only carbon and hydrogen. Nitrogen, sulfur and oxygen atoms can, however, be readily substituted for carbon atoms in the rings. The resulting "heteroaromatic" compounds are commonly grouped with the hydrocarbons, which they resemble in their properties and their behavior in analysis.

The identification of polycyclic aromatic hydrocarbons and the study of their natural pathways are pursued by geochemists, environmental chemists and toxicologists. Since the quantity of such hydrocarbons in most samples is low, concentration and preliminary separation are usually necessary. In my laboratory at the Woods Hole Oceanographic Institution we use a three-step procedure consisting of gel filtration, column chromatography and precipitation. These methods interact in different ways with the hydrocarbon mixture; in combination they isolate the constituents of the mixture rapidly. In the first step the polycyclic hydrocarbons are separated from larger molecules. In the second they are separated from substances that are held either less or more strongly in a chromatographic column of alumina. In the third step a solid precipitate is formed with a reagent that combines specifically with aromatic compounds. Foreign materials are excluded from the precipitate and can be removed by washing with a solvent. This three-step separation will work with samples weighed in micrograms. For very complex mixtures an additional chromatographic column serves to resolve the aromatic-hydrocarbon concentrate into fractions of molecules that have equal numbers of rings.

The polycyclic aromatic fractions from such a separation can be analyzed by various techniques; as a rule no single existing method will provide a complete analysis. We have found most informative a combination of three techniques: measurement of ultraviolet absorption, separation by distillation in a high vacuum and mass spectrometry. The ultraviolet spectra of the hydrocarbon samples tell us much about the spatial arrangement of the benzene rings in them, but they yield little information on the presence and nature of substituents: atoms that are extraneous to the rings. Distillation further resolves complicated mixtures; it simplifies mass-spectrometric analysis and provides structural information, based on the relations between structure and volatility. The mass spectra complement the other data; they measure the overall size of the molecules, detect the presence of substituents and often indicate their character.

Until recently the detailed composition of many mixtures of polycyclic aromatic hydrocarbons found in nature was unknown, and the sources of the compounds were mysterious. Can chemical analyses

FOSSILIZED SEA LILY of the genus *Millericrinus* was found to contain a series of pigments. The photograph on the opposite page shows a section through a rootstock of the fossil animal, which lived about 150 million years ago on the muddy bottom of the Jurassic Sea, southwest of what is now Basel in Switzerland. The middle of the stalk contains crystallized pigments, given the name fringelites after Fringeli, the mountain where the fossils were found.

provide clues to the processes of formation, transformation and transportation of organic compounds in nature? To answer the question I shall first take up some of the principles governing the formation of polycyclic aromatic hydrocarbons, after which I shall discuss some case histories that reveal correlations between structure and origin.

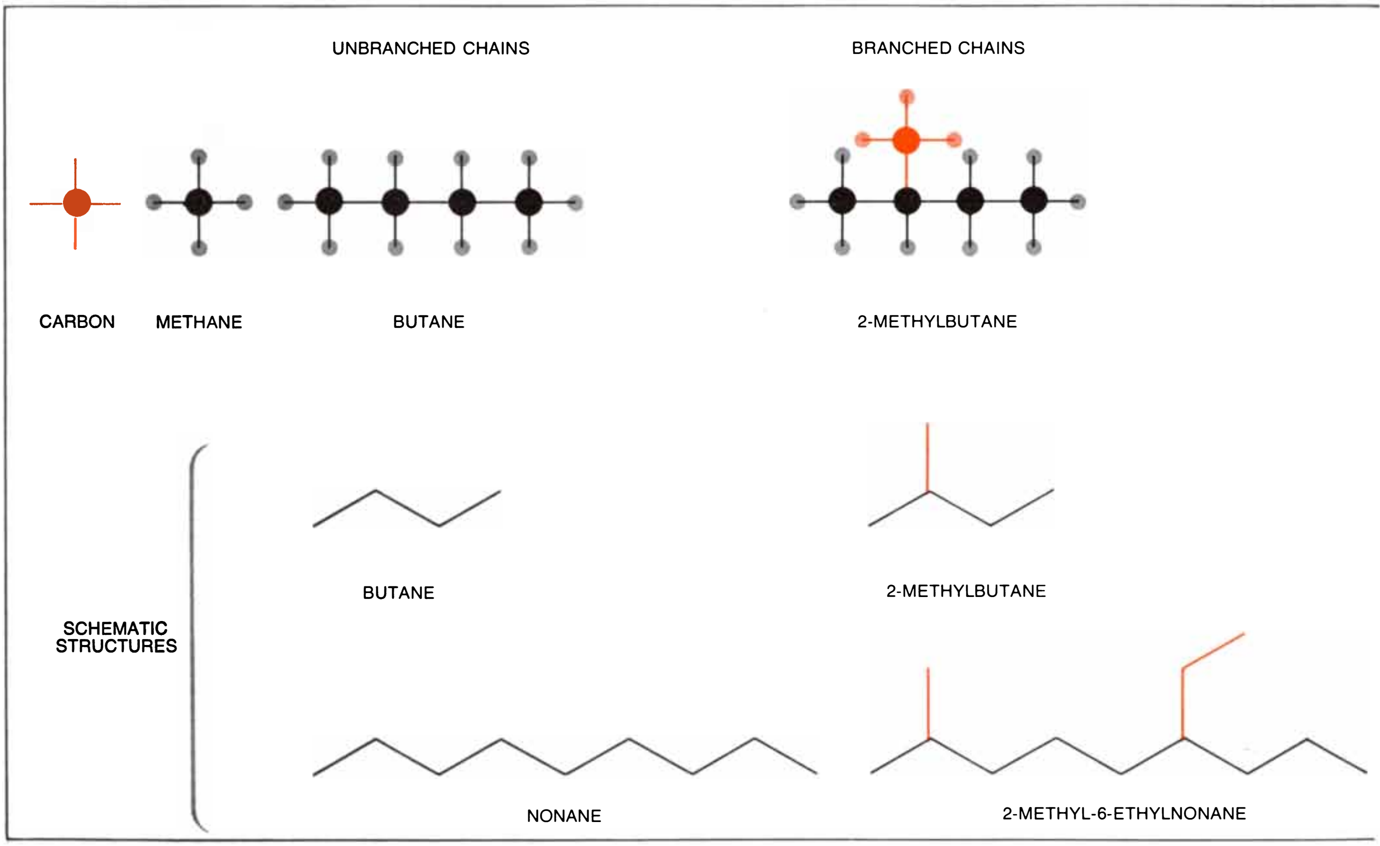
Polycyclic aromatic molecular structures are formed whenever organic substances are exposed to high temperatures. In this process, called pyrolysis, energy is released, and the aromatic products that are formed are stabler than their precursors. For example, pyrolysis occurs when a match is being charred by the advancing flame. The charcoal formed has the structure of graphite: extended giant molecules consisting of tightly linked benzene rings. High temperatures and open flames are not required, however, for the aromatization and graphitization of organic matter; even the heat from an electric iron is sufficient to cause incipient graphitization and scorch fabric. Given enough time, aromatization proceeds at even lower temperatures. The aromatic hydrocarbons of crude oil are formed over millions of years in sediments that are at temperatures between 100 and 150 degrees Celsius.

The composition of the products of thermal aromatization depends on the nature of the starting material and on the transformation temperatures. Burning a log in a fireplace or scorching a fabric with an iron leads to products that are quite unlike petroleum, although like petroleum they contain polycyclic aromatic hydrocarbons. One feature is particularly dependent on the formation temperature of the hydrocarbon mixtures: the abundance and relative distribution of aromatic hydrocarbons that carry substituent side chains of various lengths, known collectively as alkyl groups. At very high temperatures, as in the coking of coal, the products consist of a relatively simple mixture of unsubstituted hydrocarbons, presumably because of the rapid cleavage of the less stable alkyl bonds. At intermediate temperatures, as in the smoldering of wood in a smokehouse, complex mixtures of alkylated rings survive. Yet even there long alkyl chains are not favored. Unsubstituted hydrocarbons outnumber the substituted ones, and the percentage of alkylated hydrocarbons decreases rapidly as the length and number of the alkyl chains increase.

A quite different pattern is observed when the temperature of formation is lower. That pattern is best illustrated by analyses of crude oils; there alkylated polycyclic aromatic hydrocarbons far exceed the unsubstituted ones, and the average degree of alkylation and the maximum number of carbons on the aromatic rings are far higher than they are in samples produced by high-temperature pyrolysis. This fact reflects the conditions of petroleum formation; the time is adequate to accomplish the energetically favored aromatization but the temperature is not high enough to fragment even the weaker carbon-carbon bonds in the alkyl chains.

It has been conjectured that polycyclic aromatic hydrocarbons are being produced today by thermal processes in soils at environmental temperatures. Such mixtures should be even more heavily alkylated than petroleum and should therefore possess characteristics that are readily detectable, even when they are mixed with hydrocarbons from other sources. As far as I know, no such mixtures have been found.

Are polycyclic aromatic hydrocarbons also formed by living organisms? If so, can we distinguish between them and those from pyrolytic sources? Many beautiful and stable plant and animal pigments with structures based on several aromatic rings are indeed formed by living organisms. By definition, however, such pigments are not



VARIETY OF CARBON COMPOUNDS is virtually without limit. With its four valences, or bonds, the carbon atom readily forms chains, branched and unbranched, and ring structures in enormous variety. In the "ball and stick" models in the top row the large balls represent carbon atoms and the small balls hydrogen atoms. The most ubiquitous carbon-containing ring structures in nature are those based on the benzene ring: C_6H_6. Benzene is described as being unsaturated because not all the available carbon bonds are taken up by

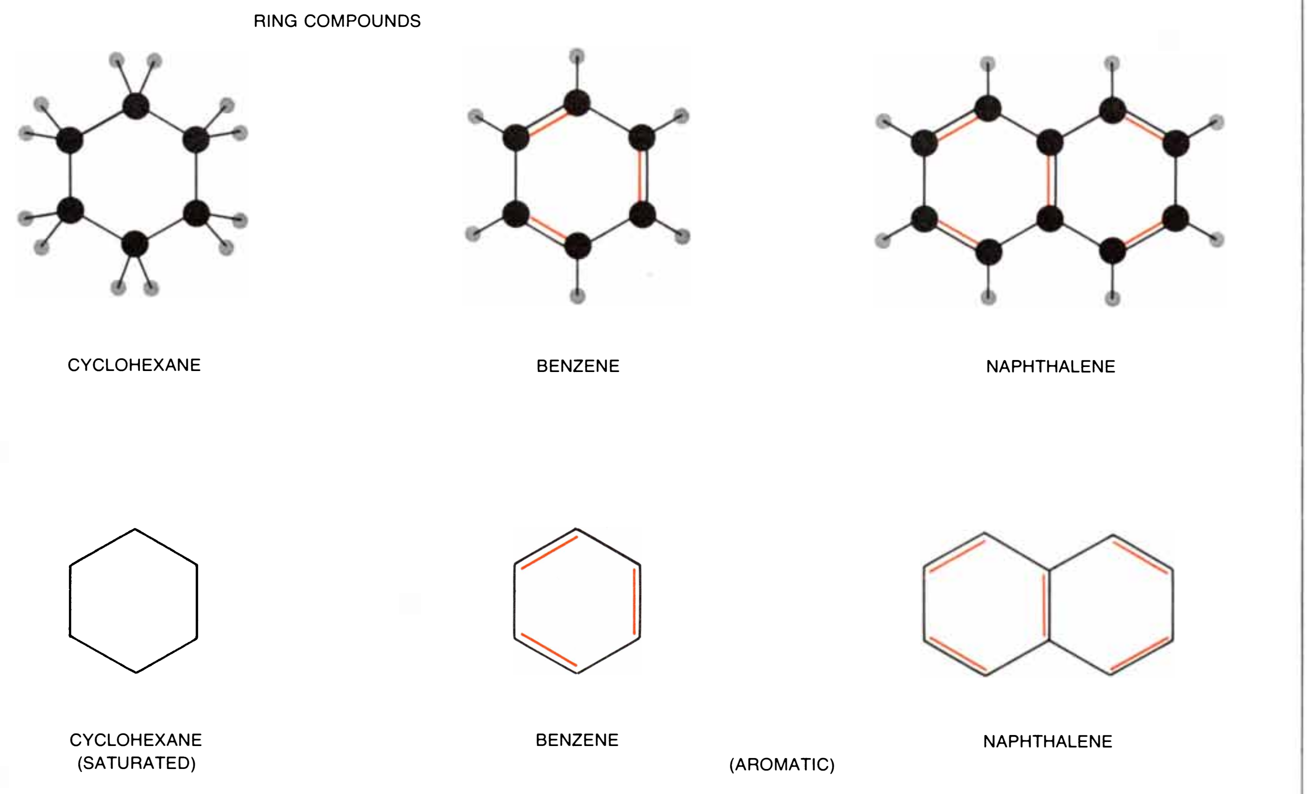
hydrocarbons; they incorporate other elements, usually oxygen and nitrogen. A vigorous discussion continues about suggestions that polycyclic aromatic hydrocarbons are also synthesized by living organisms, directly or by the transformation of precursors that might resemble the surviving structures. It is particularly difficult to resolve the controversy experimentally because of the ubiquity of aromatic hydrocarbons in nature and the ease with which they are transported and contaminate experiments.

In order to exclude such contamination Gernot Grimmer of the Biochemical Institute for Environmental Carcinogens in Hamburg has grown plants under carefully controlled conditions in greenhouses monitored for the appearance of outside contamination. In spite of extensive air-filtering the most volatile aromatic hydrocarbons entered the greenhouse from the outside. Less volatile aromatic-ring systems with four or more benzene rings, however, were not found in the plants, which appears to exclude their biosynthesis. The investigations were necessarily limited to a small number of plant species, and synthesis may still occur in other organisms.

There may be an indirect way to resolve the controversy, or at least to suggest samples that could be analyzed for biosynthetic mixtures of polycyclic aromatic hydrocarbons. Biosynthesis differs from pyrolytic and geochemical synthesis in its great selectivity. The living cell synthesizes only a very limited number of compounds out of an immense range of theoretically possible ones; those are the compounds whose particular properties are required by the organism. When several members of one chemical family (for example several straight-chain hydrocarbons) are formed by the cell, they are often formed in markedly different concentrations. In pyrolysis and geochemistry few selection rules are at work; the resulting mixtures are exceedingly complex, and adjacent members of hydrocarbon families are found in similar concentrations. The characteristic biosynthetic selection pattern is observed not only among the straight- and branched-chain hydrocarbons in organisms but also among the fatty acids, amino acids and carbohydrates; indeed, it is seen among most if not all constituents of the cell. This appears to be such a fundamental biochemical principle that one expects to observe it also in any mixture of polycyclic aromatic hydrocarbons that might be formed by organisms.

We now have the means of recognizing a thermal or biochemical contribution to an environmental assemblage of polycyclic aromatic hydrocarbons. For pyrolysates we can even estimate the formation temperature. I shall describe some natural polycyclic mixtures, give their hydrocarbon composition and attempt to deduce how they were formed. As I proceed I shall note other correlations between the chemical structures and the processes of formation of such hydrocarbon mixtures.

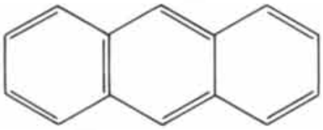
Compounds with an intense fluorescence in ultraviolet radiation are found in extracts from soils. Some 30 years ago W. Kern, then an assistant at the laboratories of Hoffmann–La Roche in Basel, became curious about such extracts and studied them in his spare time. His discovery in 1947 of the four-ring aromatic hydrocarbon chrysene in a garden soil made him the originator of a new branch of environmental chemistry. Related compounds were soon found. The isolation of the strongly carcinogenic hydrocarbon benz[*a*]pyrene in rural soils stimulated worldwide analyses. By the end of the 1960's it was generally accepted that from 10 to 15 unsubstituted hydrocarbons are present in most if not all soils in similar proportions everywhere. Various speculations attributed their formation to soil bacteria, to the decay of plant material

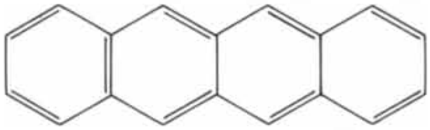


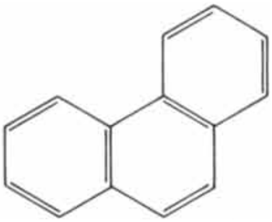
other kinds of atoms. Chemists use the word aromatic to describe unsaturated ring compounds. Naphthalene is the simplest polycyclic, or multiple-ring, aromatic hydrocarbon. Cyclohexane, C_6H_{12}, is the saturated analogue of benzene. In the familiar diagram of the benzene ring depicted here single and double bonds are shown as alternating around the ring. Actually all carbon-carbon linkages are equivalent, as if each pair of carbon atoms had one and a half bonds. In the illustrations that follow, simple schematic structures are used.

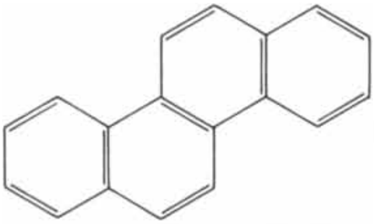
or to fallout from polluted air. The limitations of that picture rapidly became evident when newer analytical methods with higher resolving power were applied to the study of polycyclic-hydrocarbon mixtures in soils and recently deposited sediments. Within a short time the number of known or suspected members of this family had increased by at least two orders of magnitude.

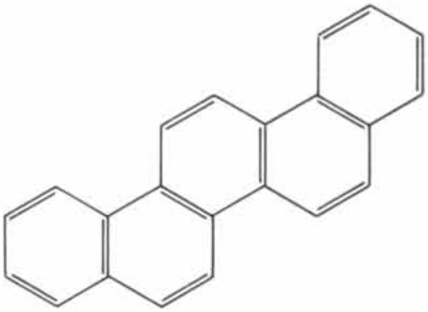
The following picture of the hydrocarbon compounds present in soils and recent sediments has emerged. Unsubstituted aromatic hydrocarbons are the most abundant; they are accompanied by extended series of substituted members containing methyl groups, alkyl chains and saturated five-member rings. Sulfur-containing analogues of the aromatic hydrocarbons are also present. These principal structural elements are encountered in mixed systems in an almost limitless number of permutations. The

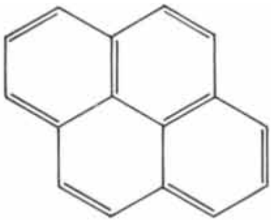

ANTHRACENE

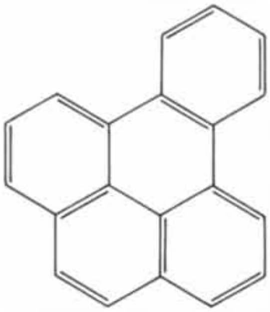
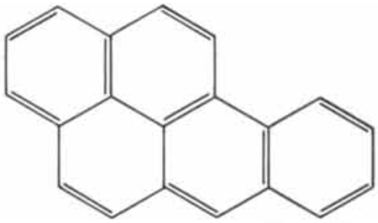

TETRACENE

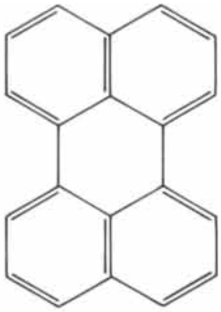

PHENANTHRENE

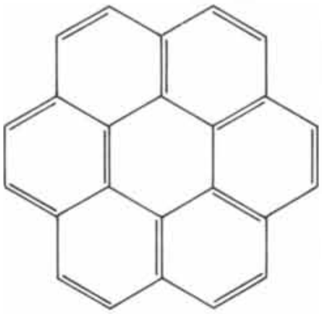

CHRYSENE

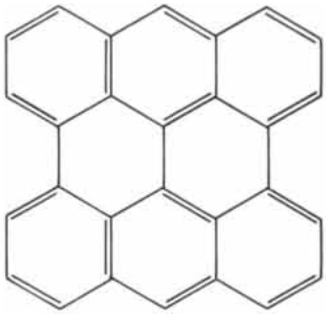

PICENE


PYRENE



ISOMERIC BENZOPYRENES


PERYLENE


CORONENE


MESO-NAPHTHODIANTHRENE

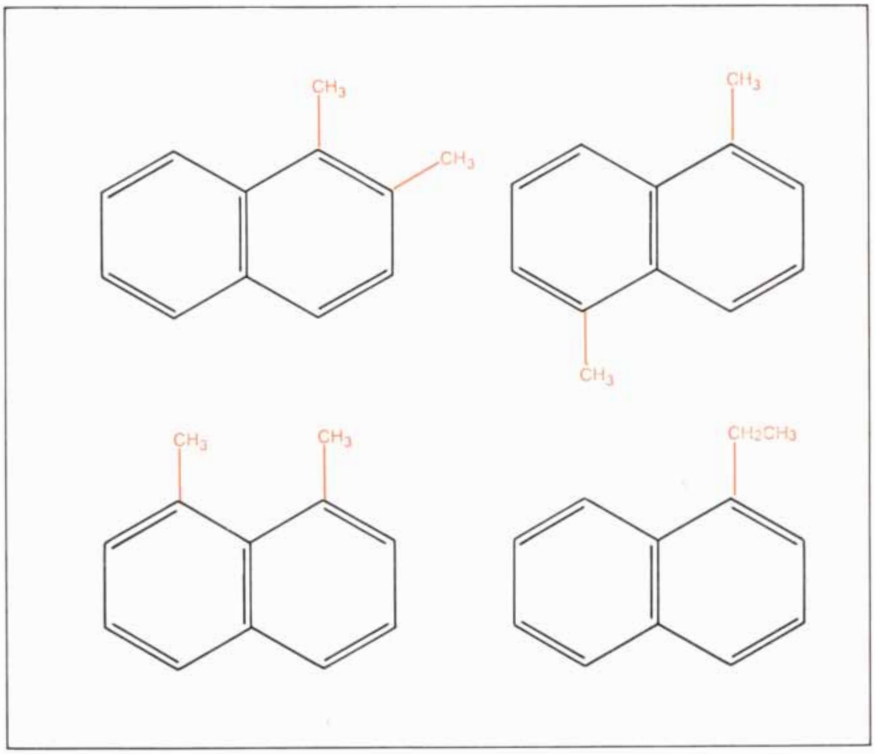
POLYCYCLIC AROMATIC HYDROCARBONS consist of two or more rings fused together in various ways. Rings of six carbon atoms have the highest stability because they require the least distortion of the natural bond angles of the carbon atoms. The rings may form linear arrangements (*top row*), angular arrangements (*second row*) or clusters, as is illustrated by the remaining configurations. The two benzopyrenes are known as isomers because they differ in geometry. Isomeric form at right, benz[*a*]pyrene, is a cancer-causing agent.

hydrocarbon composition remains remarkably constant over a wide geographic range, from continental soils to sea-bottom sediments and from deposits that vary from oxidizing to strongly reducing. It is most unlikely that the great diversity of organisms associated with such a wide variety of sites would contribute the same hydrocarbon series in the same proportions. The complexity of composition and the similarity in the concentrations of related compounds also argue against a biochemical source.

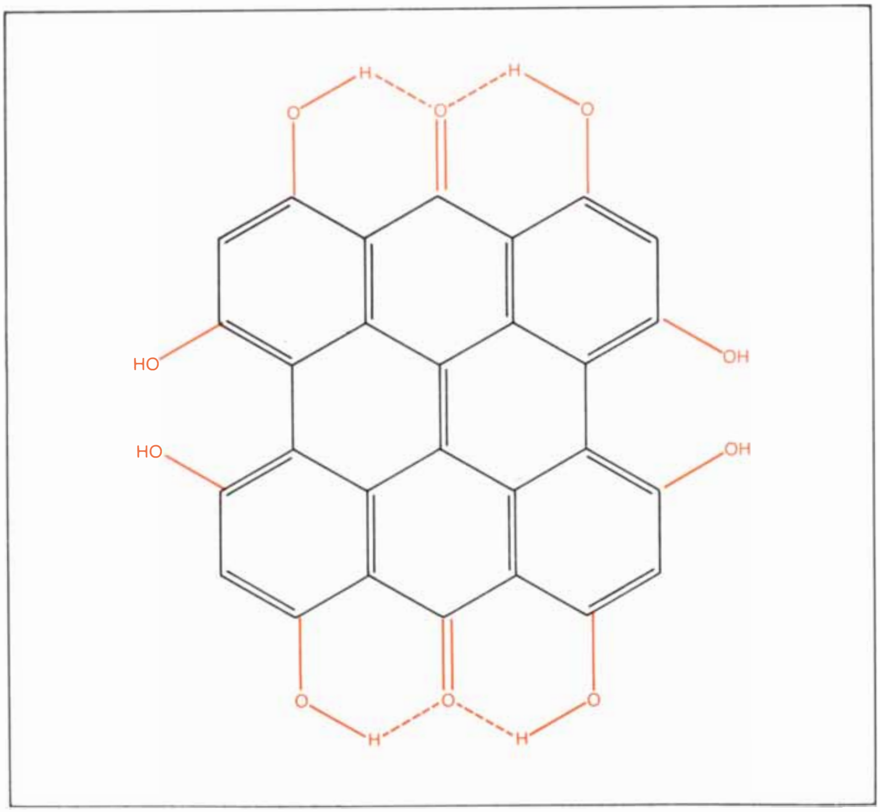
How good, then, is the case for a thermal origin of these sedimentary hydrocarbon mixtures? I believe it is excellent. In each series the unsubstituted hydrocarbon predominates. Alkyl derivatives with up to 10 substituted carbon atoms are present, but the concentrations decrease rapidly with increasing alkylation [*see bottom illustration on next page*]. That is exactly the pattern expected for medium-temperature pyrolysis. Are there any pyrolysates of natural products that exhibit such a pattern? My colleague William W. Youngblood of Florida Technological University suggested that we analyze "hickory-smoke flavor," a commercial flavoring obtained by the distillation of wood at moderately high temperatures.

Remarkably, the alkyl distribution pattern of that pyrolytic material matched our observations in soils and young sediments. This evidence excluded several thermal sources as the principal contributors to the sedimentary hydrocarbons. For example, the polycyclic aromatic hydrocarbons in polluted air, derived from incomplete combustion in furnaces and engines at higher temperatures, contain fewer alkyl derivatives than are found in the sediment samples. Crude oil, at the other end of the spectrum of formation temperatures, is much more heavily alkylated. If the polycyclic hydrocarbons were formed in decaying plant material, they should exhibit even more alkylation. In any case decaying plant material must be considered an unlikely contributor to the polycyclic aromatic hydrocarbon fraction found worldwide in sediments for the reason I have already noted: the huge diversity of organisms involved. Thus all the evidence suggests that the polycyclic aromatic hydrocarbon fraction is generated by natural pyrolytic processes. But how is it possible for the composition of the fraction to be so constant over such a wide sampling range?

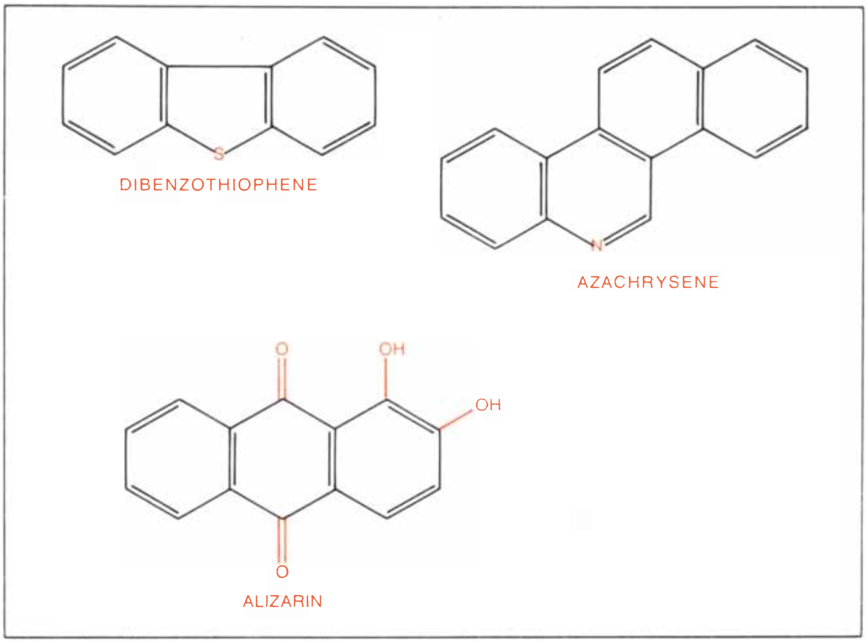
I have a provocative suggestion that would explain the relative uniformity of the polycyclic aromatic hydrocarbon fraction in soils and young marine sediments. Large quantities of pyrolytic products are formed in forest and prairie fires and are widely dispersed by prevailing winds. The haze over the North Atlantic is attributed in part to such fires. Indeed, Dwight M. Smith, John J. Griffin and Edward D. Goldberg of the Scripps Institution of Oceanography have found carbon particles with a recog-



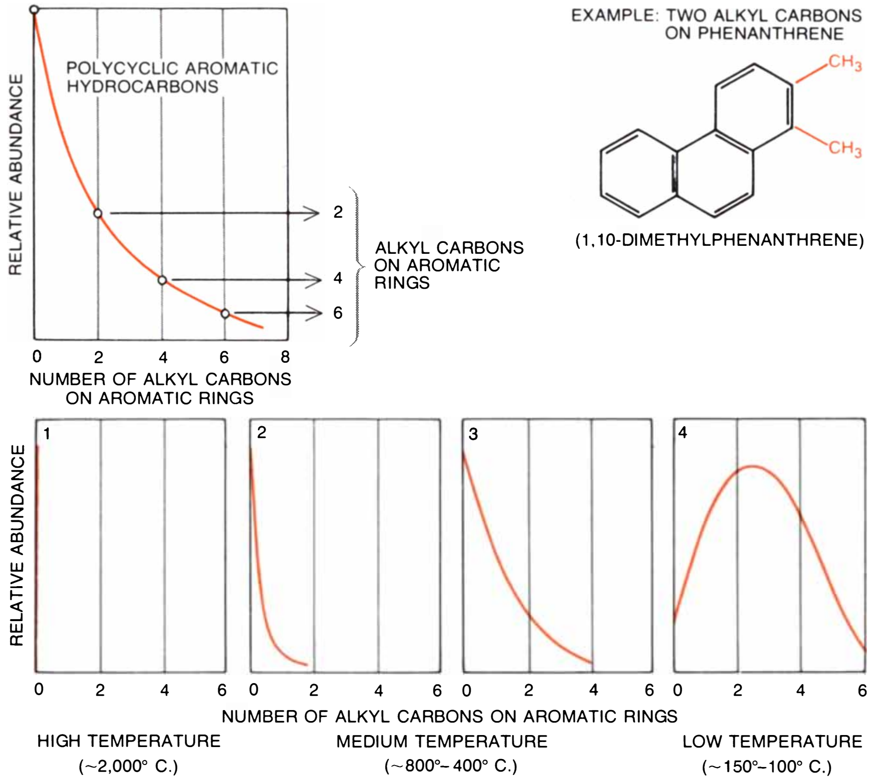
ISOMERS OF ALKYLNAPHTHALENES contain by definition the same number of carbon and hydrogen atoms. These are only four of many possible isomers. The term alkyl refers to the side chains on the rings. Although isomers are chemically identical, their structural differences give rise to differences in properties. It is nonetheless difficult to isolate isomers in pure form.



PRIMITIVE RED PIGMENT was extracted by the author from the fossilized rootstalk of the sea lily *Millericrinus* (*see illustration on page 34*). The pigment is one of a family of the fringelites that differ in the number and position of the hydroxyl (OH) groups on their molecules.



ADDITION OF FOREIGN ATOMS to polycyclic aromatic hydrocarbons greatly increases the number of structural permutations possible. Alizarin, the brilliant red pigment of Napoleon's time, has oxygen atoms and hydroxyl groups as peripheral substituents on its molecule. It bears a family resemblance to the fringelite structure shown at the bottom of preceding page.



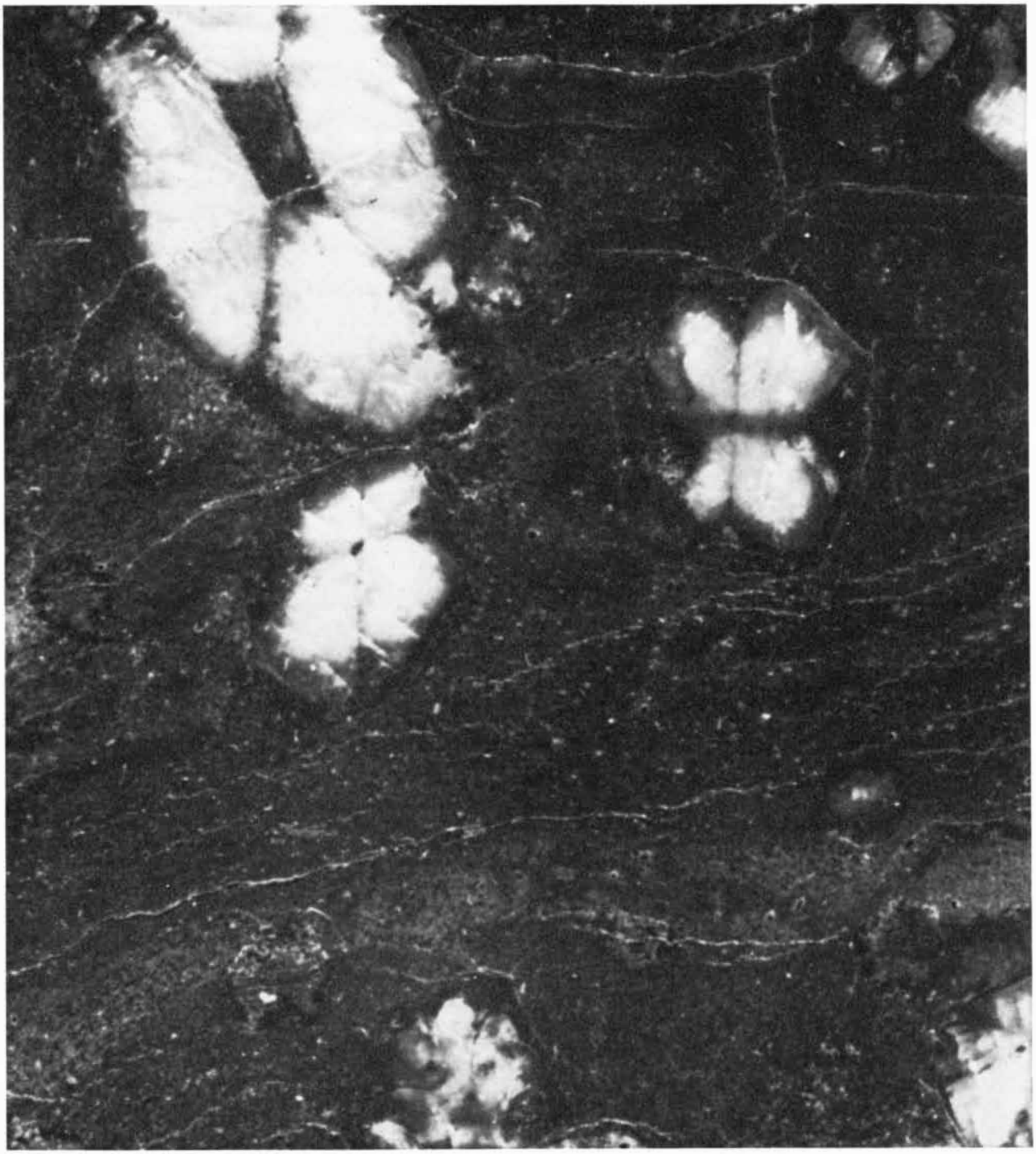
NUMBER OF ALKYL CARBONS present as side chains on polycyclic aromatic hydrocarbons correlates closely with the temperature at which the compounds were formed. In a typical mixture (*curve at top left*) rings with no side chains are the most common; in general the more side-chain carbons there are, the lower the abundance is. An example of a polycyclic compound with two alkyl carbons is 1,10-dimethylphenanthrene. When the aromatic compounds are formed at the temperature typical of gas flames (*1*), side chains are virtually absent. At temperatures typical of wood fires (*2, 3*) abundance of side chains increases with decreasing temperature. At the low temperatures that are associated with the formation of petroleum, however, the polycyclic compounds with two or three alkyl carbons outnumber the other configurations (*4*).

nizable wood texture in deep marine sediments. Other investigators have described elemental carbon in deep-sea manganese deposits where we had already discovered aromatic hydrocarbons.

In addition to such carbon particles forest and prairie fires produce aromatic hydrocarbons in great abundance. As the cooling mass of gas ascends from those fires the freshly formed active carbon particles can readily pick up the aromatic hydrocarbons and protect them from light-induced oxidation during their movement through the atmosphere. Mixing during transport would explain the uniform composition over a wide area and explain why the organisms and the chemical conditions at the site of deposition are not reflected in the chemical makeup of the hydrocarbon mixtures. Finally, the wide range of combustion temperatures in natural fires would explain the observed range of alkylated derivatives. That assumption is further supported by other structural features of the sedimentary hydrocarbon fraction. I shall return to these features when I take up some other assemblages of aromatic hydrocarbons in nature.

Some rare organic minerals with a beautiful fluorescence have been found in California and in eastern Europe; they occur together with mercury ores and sometimes with mineral waters and at vents of flammable gases. The study of these minerals goes back to the French chemist Jean-Baptiste-André Dumas, who worked on them in the first half of the 19th century. In the 1880's German chemists isolated two four-ring aromatic hydrocarbons in samples from mercury mines at Idrija in Yugoslavia. The idrialite from that location, curtisite from Skaggs Springs, Calif., and related minerals from other locations have since been the subject of several investigations. The simple analytical methods used until recently suggested a simple composition with a few components, perhaps only one. As with the sedimentary hydrocarbons, our newer methods have revealed an entirely unanticipated compositional complexity. Although idrialite and curtisite are distinctly different chemically, the hydrocarbon series in them overlap to some degree. They contain at least several hundred polycyclic aromatic hydrocarbons, together with their sulfur and nitrogen analogues and alkyl and cycloalkyl derivatives in many combinations of substitution.

Can chemical analysis tell us something about the origin of these minerals? The alkyl distribution of the polycyclic aromatic hydrocarbons in idrialite and curtisite resembles the distribution in soils and young sediments. Thus the unsubstituted hydrocarbons are the most abundant. The alkyl series does not extend as far, however, and the drop in concentration from one member of a series to the next is steeper than it is in the hickory-wood distillate. That suggests a pyrolytic origin at temperatures higher



HALO OF GRAPHITE forms around growing crystals of chiastolite (aluminum silicate) in certain minerals associated with deeply buried sediments. The organic matter in such sediments is converted into mixtures of organic compounds that vary in volatility. The most refractory fraction is graphite, which remains in specimens such as this one from R. Sawdo of the Woods Hole Oceanographic Institution.

than those in forest and prairie fires but lower than those in furnaces and engines.

Curiously, the polycyclic assemblages in idrialite and curtisite differ from those discussed so far in the way the benzene rings are arranged. The ring arrangements can be divided into three broad categories: linear (with all the rings in a line), angular (with the rings in steps) and clustered (with at least one ring surrounded on three sides). The various arrangements differ in their stability. The linear arrangement in anthracene and tetracene is the least stable; tetracene and more complex equivalents to it can be prepared in the laboratory, but they do not survive in nature. Clusters of benzene rings, as in pyrene, benzopyrene and perylene, are stabler and are commonly found in pyrolysates. The stablest configuration is the angular arrangement of benzene rings in phenanthrene, chrysene and picene. Such stable series abound in idrialite and curtisite, which contain few of the hydrocarbons of the cluster type that are common in soils, sediments, wood tar, tobacco smoke, petroleum and automobile-exhaust gases.

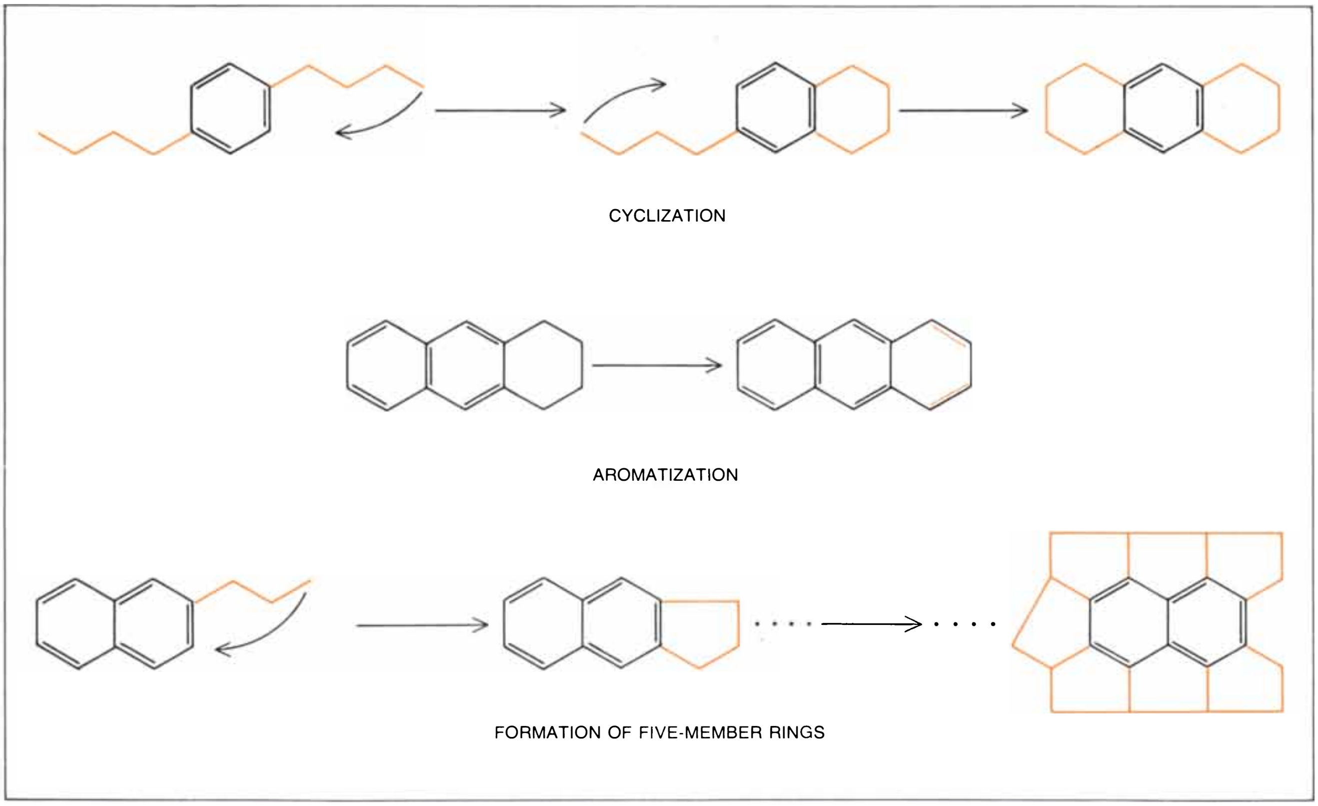
Evidently series of the cluster type are formed in pyrolysis, regardless of the temperature. They survive if the reaction products are rapidly cooled to temperatures where the decomposition of the less stable products is arrested, as it is in the case of wood tar; they also persist in petroleum because it has never reached temperatures capable of eliminating or of rearranging the less stable ring systems. This suggests that idrialite and curtisite lack the unstable cluster configurations because those configurations have been eliminated by prolonged exposure to elevated temperatures.

With such chemical clues and the geological background, one can interpret the formation and evolution of these hydrocarbon minerals as follows. Sediments containing organic compounds are carried by movements of the earth's crust to regions much deeper than those where petroleum is typically formed. In those deeper regions pyrolytic temperatures probably go as high as 400 to 500 degrees C. The original organic material is destroyed and its constituents are rearranged and thermodynamically stabilized. One of the products is graphite, which remains at the original depth. The stable hydrocarbon gases (in particular methane) and the polycyclic aromatic hydrocarbons (along with their sulfur and nitrogen analogues) remain near the original depth long enough for the least stable hydrocarbons to be destroyed. Eventually the rest of the hydrocarbons move toward the surface together with the gases, the mineral waters and the mercury ores, which are recognized for their geochemical mobility. Along the migration path the hydrocarbons are separated by fractional crystallization. Fractions of higher molecular weight and higher melting point form idrialite and pendletonite; fractions of lower molecular weight and lower melting point end up as curtisite.

Chemical analysis has again given us clues to the formation of a natural hydrocarbon assemblage; it suggests the gross mode of formation and provides evidence on the temperature of formation and the duration of thermal exposure. As our investigations proceeded we visualized an explanation for the formation of hydrocarbons in such diverse materials as soils, young sediments, wood tar, tobacco smoke, engine exhaust and minerals such as idrialite and curtisite. We obtained further correlations between chemical structure and the processes of hydrocarbon formation from studies of the polycyclic aromatic hydrocarbons in petroleum.

Petroleum may well be the most complex organic mixture on the earth. It is formed from the residues of earlier life in buried sediments by chemical reactions that require millions of years for completion. The transformation resembles pyrolysis, but the reactions are exceedingly slow because of the modest temperatures involved: probably less than 150 degrees C. Products that would not be observed in high-temperature pyrolysates are formed and retained.

The chemical analysis of the polycyclic aromatic fraction in petroleum is a challenging task. Adequate analytical resolution requires the combination of many different techniques; even then pure single compounds are rarely isolated. Harold J. Coleman and his co-workers at the Bartlesville Energy Research Center in Oklahoma have made extensive analyses of the polycyclic aromatic fractions in crude oils, particularly those from the region of Prudhoe Bay in Alaska. Polycyclic aromatic hydrocarbons are abundant, amounting to about a sixth of the oil fraction distilling between 370 and 535 degrees C. A similar propor-

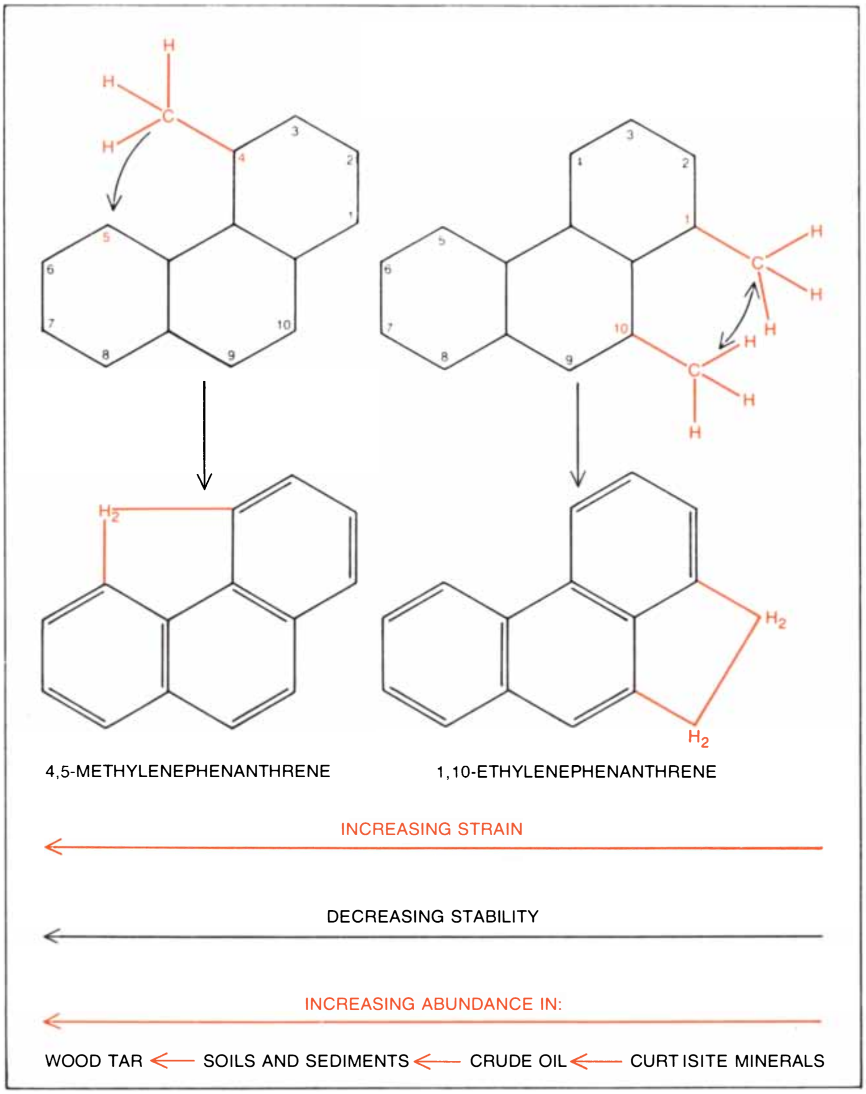


CYCLIZATION AND AROMATIZATION of saturated hydrocarbon chains produce the complex mixture of polycyclic aromatic compounds found in petroleum. The source material of petroleum, largely derived from plants, is rich in long carbon chains. At elevated temperatures in buried sediments some of the chains lose a few hydrogen atoms and are converted into six-member rings (*top*). With further loss of hydrogen the saturated rings are aromatized (*middle*). Additional substitution by chains, followed by cyclization and aromatization, gradually creates compounds with many rings. Five-member rings, which also form readily (*bottom*), cannot be converted into aromatic rings. If several accumulate around an aromatic nucleus, they block its further growth to a still larger aromatic system. Aromatic hydrocarbons surrounded by several five-member rings are abundant in petroleum. Their production is favored by a reaction time of millions of years and by the absence of temperatures able to break the somewhat strained carbon-carbon bonds in the five-member rings.

tion is present in some fractions that have a lower boiling point, and in their higher distillates and distillation residue. Three compositional features are particularly pertinent here: the extreme complexity and the uniformity in the concentration of adjacent compounds within the many hydrocarbon series, the frequent occurrence of compounds with saturated five-member rings and the presence of compounds in which the substituent groups show considerable steric strain, that is, show bond angles distorted beyond their normal range.

The complexity and the uniformity in the concentration of adjacent compounds characterize the origin of this crude-oil fraction; it is not a biochemical pattern but the result of extensive geochemical scrambling of organic compounds by nonselective reactions. In the formation of petroleum some linear carbon structures are transformed into saturated five- and six-member rings. Such reactions are possible because they stabilize the source material thermodynamically. The six-member saturated rings thus created are readily transformed by aromatization—the loss of hydrogen—into the still stabler aromatic systems. Those systems can grow through the accretion and aromatization of new saturated six-member rings.

The five-member rings, on the other hand, are not readily converted into aromatic structures, so that with the passage of time two, three or more such rings may accrete around an aromatic nucleus, thereby blocking its further growth into a large aromatic system. The average number of five-member rings around an aromatic nucleus is far greater in petroleum than it is in hydrocarbon series formed at higher temperatures. That fact reflects the much greater time available for petroleum formation; the marshaling of rings can continue, although at low intensity, for millions of years. In pyrolysis at higher temperatures five-member rings in the aromatic structures are less abundant either because the reaction time is short, because the source material is rapidly converted into stable six-member and aromatic systems or because the temperature is high enough to break the carbon-carbon bonds in saturated rings.



STRAINED FIVE-MEMBER RINGS are formed by pyrolysis (exposure to high temperatures) from alkyl side chains. Only one alkyl carbon is needed (*left*) if the parent structure supplies three of the five sides of the ultimate ring. In other cases (*right*) two alkyl chains may be needed. Because the carbon-carbon bonds are under greater strain in five-member rings than in six-member ones they have a greater energy content and hence a lower stability. Five-member rings survive most readily if the pyrolytic mixture is quickly quenched, as it is in the case of smoke from a wood fire, or if the pyrolysis continues for a long time at low temperature.

Highly strained molecules in which a single CH_2 group forms a bridge between two aromatic carbon atoms, as in 4,5-methylenephenanthrene [*see illustration on this page*], form with different relative abundances in various pyrolysates. They are most abundant in wood tar, lower by one order of magnitude in soils and sediments and by two orders of magnitude in crude oils. In idrialite and curtisite such compounds are nearly absent. The distorted bond angle at the CH_2 bridge implies that such molecules have a higher energy content than molecules with less strain and therefore less stability. We attribute the formation of strained CH_2 bridges in pyrolysis to energetic reactions and their survival either to rapid quenching (as in tar and smoke) or to low temperatures (as in petroleum formation) that depress the rate at which structural rearrangement can proceed. CH_2 bridges have disappeared from idrialite and similar minerals because they were exposed to high temperatures for long periods after the initial pyrolysis.

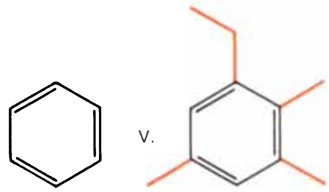
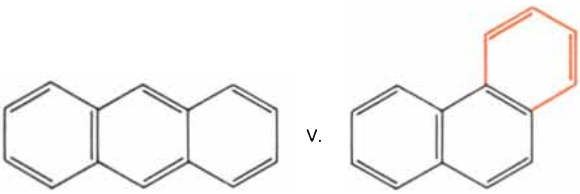
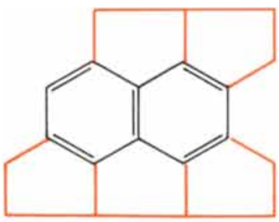
These examples will serve to show the richness of information that is encoded in the structure of organic compounds in nature. The parameters I have discussed here reflect the character of the source materials, the processes of formation, the temperatures at which the compounds form and the reaction time. The relations among these parameters suggest that the compounds could be used as geological "thermometers" and "clocks." Their calibration, however, presents a difficult problem. There must be a complex interdependence among the parameters, since time and temperature both influence the survival of unstable structures. Yet even at the present level of our understanding, detailed chemical analysis has helped us to recognize a wide range of environmental processes and has guided us to a unified view of materials that are remote in terms of sources and formation processes.

Not all natural mixtures of polycyclic aromatic hydrocarbons have a single predominant origin. Often there is a more complex situation where more than one hydrocarbon source contributes to the sample. For instance, a recent marine sediment may contain fossil fuels from an oil spill in addition to the hydrocarbons we attribute to the fallout of soot particles from forest and prairie fires. The different patterns of composition should make it possible to recognize these multiple sources.

With my co-worker Jeremy Sass I have studied one such instance. An oil spill in Buzzards Bay off Massachusetts has contaminated the near-shore bottom sediment with a fuel oil that has a comparatively low boiling point. Extraction and isolation of the higher-boiling-point polycyclic aromatic fraction and subsequent analyses reveal the normal background pattern we observe not only throughout the bay but also in the soil of the adjacent land. The unsubstituted aromatics predominate and the alkyl distribution pattern is the one characteristic of woodsmoke, that is, of the hydrocarbons we attribute to the fallout of soot. Within the range of boiling points of the compounds in the spilled fuel, however, alkyl derivatives are more abundant than the unsubstituted hydrocarbons. That is the petroleum pattern, and it is remarkable that it can still be observed now, almost six years after the spill.

A particularly interesting example of a dual source comes from our earlier investigation of hydrocarbons in a fossil sea lily from the Jura Mountains of Switzerland. This fossilized deep-sea animal is vividly colored by pigments that are dispersed throughout the rock matrix of the fossil. Present as tiny crystals, the pigment is a true organic mineral; we named it fringelite after Fringeli, the mountain where we discovered it. Since similar compounds are still found in living relatives of this animal species, we assume that the pigment, or at least its close precursors, were part of the animal that lived in the Jurassic Sea 150 million years ago.

Fringelite and its close relatives are aromatic compounds. When we attempted to resolve the fossil pigments by chromatography, we noted in the fraction that washed out of the column ahead of the pigments some intensely fluorescing but nearly colorless materials. Further study revealed a continuous polycyclic series with from three to seven aromatic rings. The lower members of the series are not particularly abundant and closely resemble aromatic mixtures that are found in many other geological specimens. They may be the survivors of compounds that were synthesized in the geologic past or, what is more likely, they may be geological transformation products, originating in deep sediments that were not sufficiently rich in organic material to yield petroleum. At higher molecular weights we find unusual hydrocarbons, never before isolated from sediments, that exhibit precisely those features we had pre-

STRUCTURAL FEATURES	STRUCTURAL PREFERENCE	CONDITION OF FORMATION
 DEGREE OF ALKYLATION	LOW DEGREE OF ALKYLATION	HIGH FORMATION TEMPERATURE HIGH CARBON CONTENT IN SOURCE MATERIAL
 RING ARRANGEMENT (LINEAR, ANGULAR, CLUSTER)	ANGULAR RING ARRANGEMENT	LONG EQUILIBRATION TIME HIGH EQUILIBRATION TEMPERATURE
 STRAIN	HIGH DEGREE OF RING STRAIN	SHORT EQUILIBRATION TIME LOW FORMATION TEMPERATURE
 NUMBER OF FIVE-MEMBER RINGS	MULTIPLE FIVE-MEMBER RINGS	LONG REACTION TIME LOW FORMATION TEMPERATURE

CONFIGURATION OF HYDROCARBON STRUCTURES found in geochemical mixtures provides rich information about the conditions under which the structures formed. The type and spatial arrangement of the alkyl chains and of the saturated and aromatic rings can vary over a wide range. The presence of preferred structures in natural mixtures of polycyclic aromatic hydrocarbons reflects the composition of the source material, the temperatures of formation and transformation and the duration of chemical processes involved.

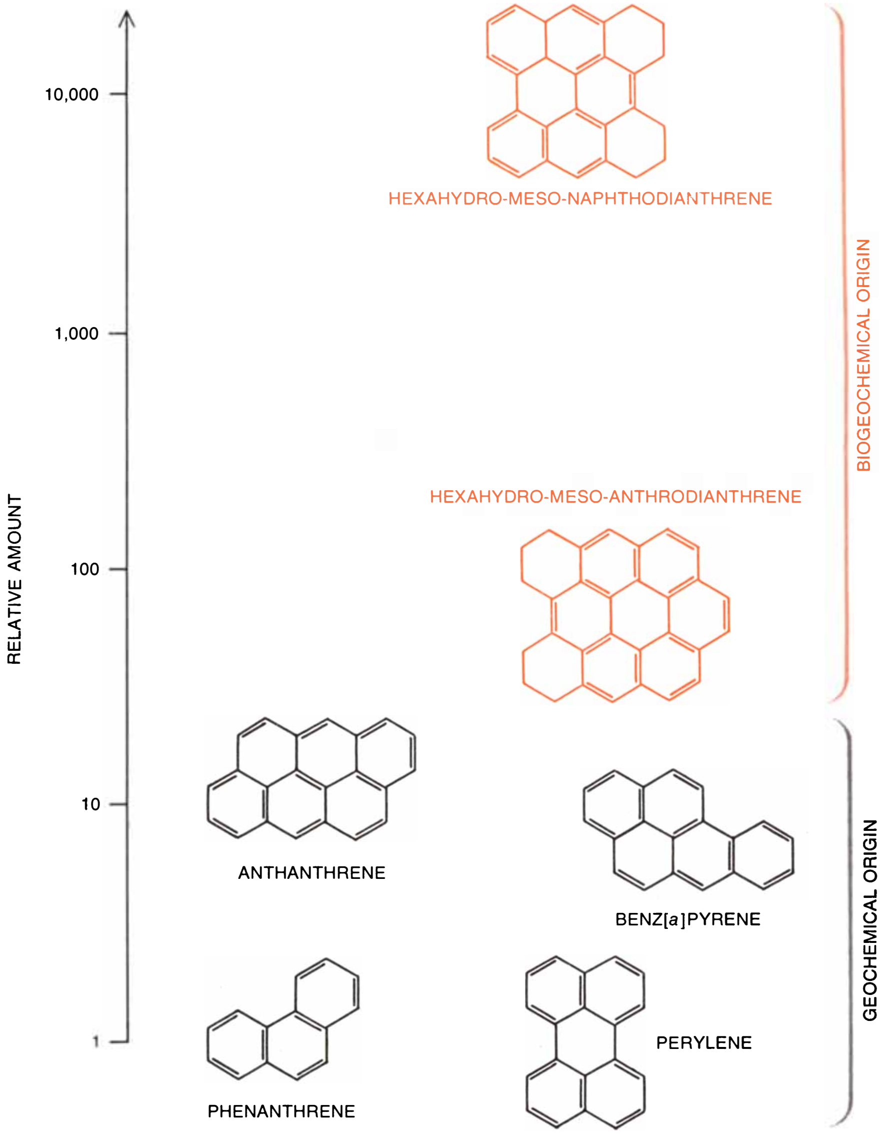
dicted for substances of biochemical rather than geochemical origin. The number of such compounds is limited, the chemical structures are select and unusual, and the concentrations are far higher than the normal background [*see illustration at right*].

More detailed structural analysis supports this interpretation. The exotic hydrocarbons are the bare aromatic ring systems on which the fringelite pigments are based. We are not certain whether those hydrocarbons were already present in the living animals and therefore represent genuinely biochemical products or whether they are geochemical transformation products of biochemical precursors closely related to the fringelite pigments. Either would support the predictive value of the relations I have suggested between chemical structure and biological origin.

Our original view of aromatic hydrocarbons in nature has now changed dramatically. The picture of a dozen or so simple compounds distributed over the earth in similar ratios has given way to the realization that the mixture is exceedingly complex and variable and that its origins are diverse. The mixture is not yet fully analyzed; indeed, it may not be fully resolvable at the present state of the analytical art. Yet even our incomplete analyses reveal the operation of a number of fundamental chemical processes, and we are learning to read in the hydrocarbon structures a code that can be deciphered in terms of sources and processes of formation, transformation and mass movement. We have spanned a wide range of samples, and the evidence for a thermal origin of polycyclic aromatic hydrocarbons in nature is overwhelming. The biosynthesis of selected aromatic assemblages under special circumstances has not been ruled out, but neither has it been demonstrated convincingly.

Polycyclic aromatic hydrocarbons from pyrolysis have been present on the earth for a long time. Man has been in contact with combustion products throughout his history, and natural fires and reactions in sediments formed polycyclic aromatic hydrocarbons long before the advent of man. Moreover, we now know that the hydrocarbons in smoke, in fallout from the air, in sediments and in fossil fuels include, in addition to already recognized compounds that can give rise to cancer and mutations, new carcinogens and mutagens. That raises a new question: Have such materials contributed significantly to the role of mutation in the evolution of species? They might then rank among other natural mutagens such as ultraviolet radiation and background nuclear radiation.

Polycyclic aromatic hydrocarbons now enter the environment in larger amounts than they did in the human and geologic past, and from some new sources. Among the dominant sources are the incomplete combustion of wood, coal and petroleum and the spillage of raw or refined petroleum.



TWO GROUPS OF HYDROCARBONS are found in the fossil sea lily *Millericrinus*. The less abundant group consists of fairly small structures of three to six fused benzene rings also widely found in ancient sediments. The more abundant group consists of much larger polycyclic structures (*color*) that have never been found in other deposits. Their structures are so similar to the fringelite pigments that a common origin seems very likely. This is one of the few cases where at least part of a natural mixture of polycyclic aromatic hydrocarbons does not originate with pyrolysis but can be traced back to a substance that was synthesized by a living organism.

Attempts have been made to assess the effects of this new influx of aromatic hydrocarbons into the environment, yet these attempts predate the recent realization that the environmental mixture of aromatic hydrocarbons is exceedingly complex, that it has many sources, that it is difficult to analyze and that many components have never been tested for their biological activity. It seems important to reexamine the question of the environmental effects of the new influx of aromatic substances in the light of our present knowledge and of our remaining ignorance.

Since Dumas's first study of idrialite and Kern's discovery of chrysene in soil, our knowledge of polycyclic aromatic hydrocarbons in the environment has developed in step with the analytical art: slowly during periods of analytical stagnation, rapidly with the confluence of many modern analytical methods. I see a similar development in other areas of environmental organic chemistry. In each our best techniques fail to approach the complete resolution of chemical compounds. The question "How complex is nature?" is an important one and has many implications. As a geochemist I see an almost limitless opportunity to interpret a complex world in terms of the processes that have shaped it. As a part-time environmental biologist I am frustrated by the difficulty, if not impossibility, of predicting the effect of organic chemicals in nature without a fuller understanding of their structure.

The Meteorology of Jupiter

The visible features of the giant planet reflect the circulation of its atmosphere. A model reproducing those features should apply to other planetary atmospheres, including the earth's

by Andrew P. Ingersoll

Every feature that is visible in a picture of the planet Jupiter is a cloud: the dark belts, the light-colored zones and the Great Red Spot. The solid surface, if indeed there is one, lies many thousands of kilometers below the visible surface. Yet most of the atmospheric features of Jupiter have an extremely long lifetime and an organized structure that is unknown in atmospheric features of the earth. Those differences, and the fact that atmospheric features are easy to observe on Jupiter, make the giant planet a laboratory where terrestrial meteorologists can test theories about atmospheric dynamics in ways not possible on the earth.

Jupiter's diameter is 11 times larger than the earth's; its surface gravity is 2.4 times stronger; it rotates on its axis 2.4 times faster (once every 10 hours). There is little or no change of season on Jupiter because the axis of the planet's rotation is nearly parallel to the axis of its orbit around the sun.

Jupiter's atmosphere and interior are mostly hydrogen, with other elements such as helium, carbon, oxygen and nitrogen mixed with the hydrogen in the same proportions as they are in the sun. Because that mixture does not solidify at the temperatures and pressures that have been calculated to exist on Jupiter, the planet is probably gaseous or liquid throughout its interior. The mixture is gently stirred at all depths by convection currents that carry heat from the interior to the surface. Theoretical calculations indicate that the internal heat is most likely left over from Jupiter's initial gravitational contraction, when it condensed out of the nebula that also gave rise to the sun and the other planets. The amount of internal energy Jupiter releases at present is approximately equal to the amount of energy it absorbs from the sun.

From the point of view of meteorology the most important differences between Jupiter and the earth lie in the fact that Jupiter has an appreciable internal energy source and probably lacks a solid surface. The other differences—in radius, gravity, rate of rotation and so on—are mainly differences of degree. Even the chemical composition of Jupiter's atmosphere is similar to the chemical composition of the earth's as far as its effect on meteorology is concerned. The atmospheres of the two planets consist chiefly of noncondensable gases: hydrogen and helium on Jupiter, nitrogen and oxygen on the earth; mixed in are small amounts of water vapor and other gases that do condense, forming clouds. In terms of the temperature changes that would occur on the two planets if the condensable vapors were entirely converted into liquid or solid form, thus releasing all their latent heat, Jupiter's atmosphere would be somewhat less affected by condensation than the earth's. Clouds, condensation and precipitation are nonetheless important in the dynamics of both atmospheres.

Our picture of the average composition and vertical structure of the Jovian atmosphere is based partly on observation and partly on theory. Spectroscopic studies from the earth have established that the atmosphere is mostly molecular hydrogen (H_2), with smaller amounts of methane (CH_4), ammonia (NH_3) and a growing list of other gases. In those studies absorption features in the infrared spectrum of Jupiter are compared with absorption features in the spectrum of gases in the laboratory. Since each gas has characteristic wavelengths at which it absorbs radiation, absorption spectra of the Jovian atmosphere make it possible to positively identify most of the gases in it even in very small concentrations. The exception is helium, which absorbs radiation only in the ultraviolet region of the spectrum at wavelengths that cannot be observed through the earth's atmosphere. Helium was recently detected, however, by an ultraviolet spectrometer on the spacecraft *Pioneer 10,* and the effect of the helium on the infrared spectra of other gases in the Jovian atmosphere has also been observed.

From the relative strengths of the absorption spectra of hydrogen, methane and ammonia the relative abundances of those gases can be determined. On Jupiter the ratio of the number of carbon atoms to hydrogen atoms is about 1 : 3,000 and the ratio of nitrogen atoms to hydrogen atoms is 1 : 10,000. Those abundances are close to the abundances of the same elements in the sun. Within wide limits the ratio of helium to hydrogen in Jupiter is consistent with the inferred ratio of helium to hydrogen in the sun (1 : 15). It is the abundance ratios, together with the low density of Jupiter as a whole, that suggest that the planet is very much like the sun in its composition.

The amount of heat Jupiter radiates implies that the interior of the planet is hot. If it were cold, there would not be enough heat in the interior to have lasted until the present time. A consequence of the hot-interior model of Jupiter is that solids cannot form in it. According to current thinking, the planet is mostly liquid, with a gradual transition to a gaseous atmosphere in the outermost few thousand kilometers.

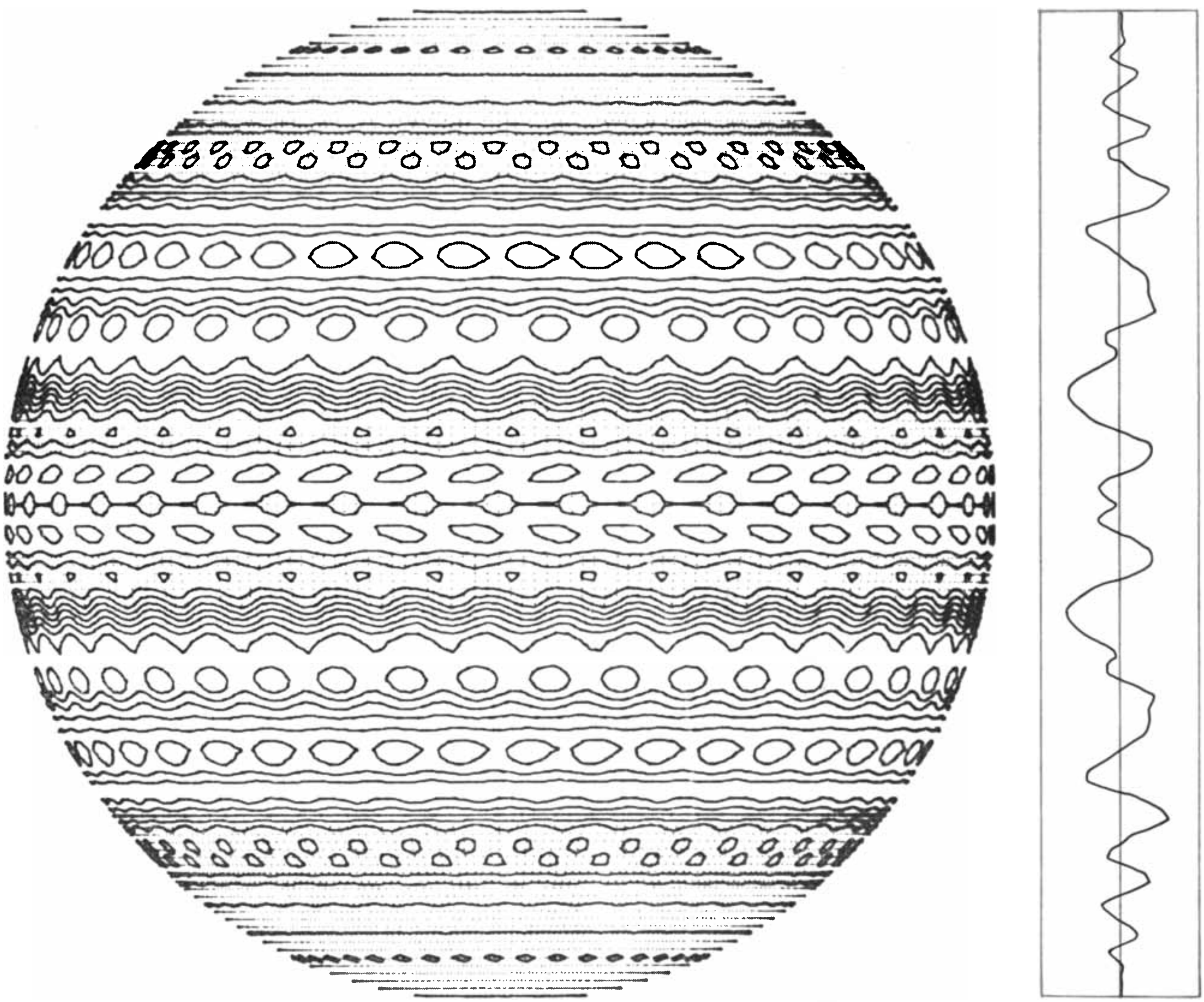
Spectroscopic data at infrared and radio wavelengths also yield information about the temperature and pressure in the Jovian atmosphere. At the deepest levels the temperature decreases with altitude at the rate of some two degrees Celsius per kilometer. That rate is close to the adiabatic lapse rate, which is the value of the temperature gradient in a well-mixed atmosphere. In such an adiabatic atmosphere parcels of gas move vertically without exchanging heat with neighboring parcels. They get cooler or warmer, but they do so only because their pressure changes as they ascend or descend. Neighboring parcels at the same level in the atmosphere are indistinguishable from one another. An adiabatic temperature gradient also usually implies that the atmosphere is stirred by convection.

The temperature in Jupiter's atmosphere is 165 degrees Kelvin (degrees C. above absolute zero) at the level where the pressure is one atmosphere (the pressure of the earth's atmosphere at sea level). The temperature continues to decrease with height until it reaches a minimum value of 105 degrees K. at the level where the pressure is .1 atmosphere. At that point it begins to rise slightly once again. By analogy with the earth's atmosphere this minimum marks the beginning of the Jovian stratosphere. In that layer of the Jovian atmosphere the temperature is controlled largely by radiation rather than by convection.

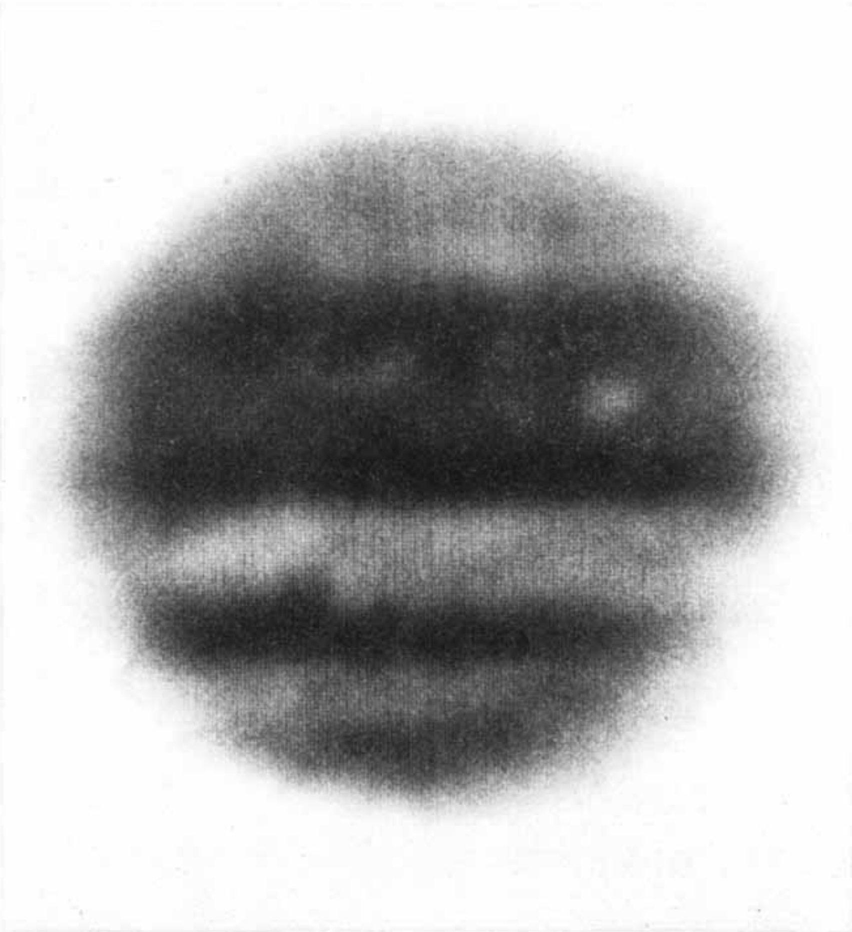
There are thin clouds on Jupiter as high as the base of the stratosphere. Broken thick clouds, with holes opening into deeper levels, begin to be present where the pressure is between .6 and one atmosphere. The pressure at the cloud tops on Jupiter is thus similar to the pressure at the cloud tops on the earth. The temperatures at those levels on Jupiter are much lower, however, since Jupiter is five times farther from the sun than the earth is.

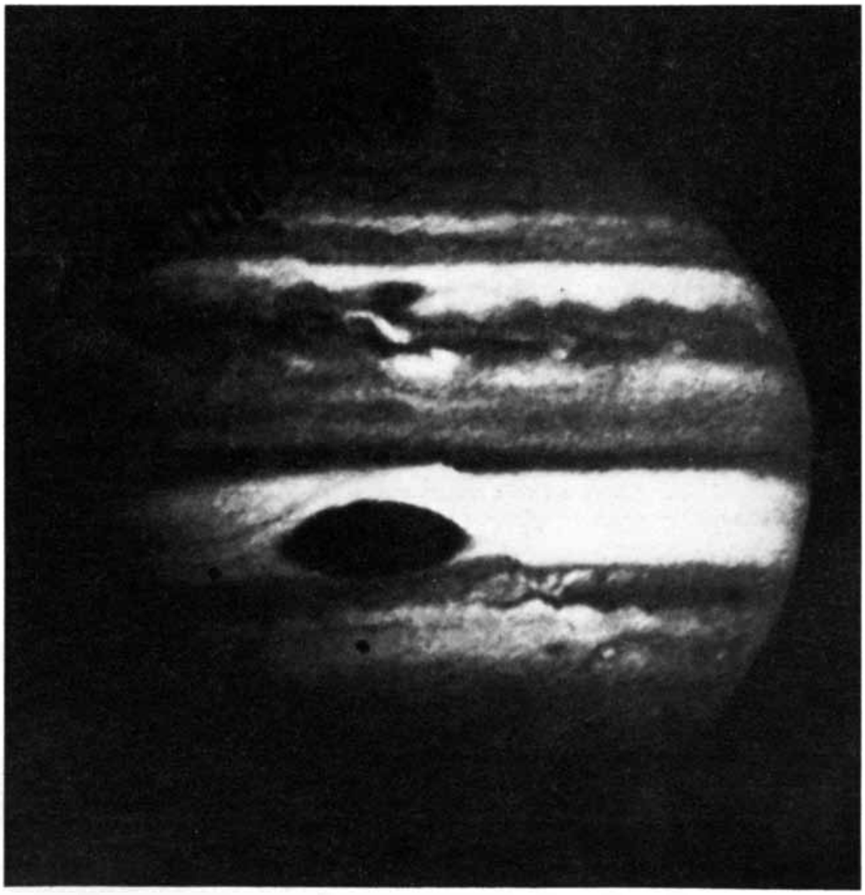
The composition of the Jovian cloud particles and the nature of the material that endows them with their remarkable colors cannot be determined from spectroscopy alone. Here theoretical calculations by John S. Lewis of the Massachusetts Institute of Technology have been useful. Lewis began by assuming that Jupiter's atmosphere has the same composition as the sun's at the ranges of temperature and pressure observed on the planet, and that all its constituents are at chemical equilibrium. He then calculated the amount of solid and liquid matter in the atmosphere as a function of height above an arbitrary base level. His calculations show that the deepest and thickest clouds are condensed water, since both oxygen and hydrogen are abundant in the atmosphere. Above those clouds are clouds of ammonium hydrosulfide (NH_4SH) and above those in turn are clouds of pure ammonia (NH_3). The level of each cloud depends on the vapor pressure of the particular condensate, which is a strong function of temperature. The less volatile substances such as water condense at higher temperatures (deeper in the atmosphere) than the more volatile substances such as ammonia. The relatively low vapor pressure of water at the temperatures characteristic of the Jovian cloud tops also explains why water vapor was found there only recently: water cannot be detected spectroscopically unless it is in the form of a vapor. In 1974 Harold P. Larson and his co-workers at the University of Arizona detected water on Jupiter for the first time by observing radiation that emerged through holes in the upper cloud layers from water molecules at deeper levels.

Lewis' model, which assumes that all the substances are at chemical equilibrium at each level, cannot account for the colors of the clouds. The cloud particles in that model are white, whereas the Jovian clouds are subtle shades of red, brown, white and blue. With convection bringing up exotic material from deeper levels, however, and with ultraviolet radiation from the sun energizing chemical reactions at the top of the atmosphere, it is not surprising that portions of Jupiter's atmosphere do depart from local chemical equilibrium. Only small amounts of coloring material are needed to explain the observations, and an atmosphere with the same composition as the



TURBULENCE OF JOVIAN ATMOSPHERE is depicted in a computer simulation devised by Gareth P. Williams of Princeton University, who used one model that attempts to account for Jupiter's east-west bands. The lines plotted are streamlines along which the atmosphere flows. Flow is fastest where the lines are closest together. The relative direction of the flow is indicated by the graph at the far right. Where the curve travels to the right of the vertical line the net flow is eastward, or faster than the planet's mean rotational velocity. Where the curve travels to the left of the vertical line the net flow is westward, or slower than the mean rotational velocity. The relative velocity of the flow, which is up to 50 meters per second, is indicated by the magnitude of the deviation of curve from vertical line.





sun's would contain all the necessary elements in a wide variety of compounds. It is currently believed that the clouds are basically as Lewis has predicted, with the colors explained by the addition of small amounts of sulfur, red phosphorus and complex organic molecules.

Data on the horizontal structure and motions of the Jovian atmosphere are obtained mainly from photographs at visible wavelengths. The data include ground-based observations beginning at the end of the 19th century, together with the high-resolution images made from the spacecraft *Pioneer 10* and *Pioneer 11*. The ground-based observations before 1955 are summarized by B. M. Peek in his classic work *The Planet Jupiter*. It is evident from surveying these observations that most of the principal features of Jupiter's atmosphere have existed for decades or longer. Such longevity of cloud patterns is remarkable by terrestrial standards. On the earth cloud patterns rarely last longer than one or two weeks, unless they are directly connected with underlying topography such as mountain ranges. On Jupiter there is no underlying topography as far as anyone knows. Hence it would be of great interest to know why the Jovian cloud patterns are so long-lived.

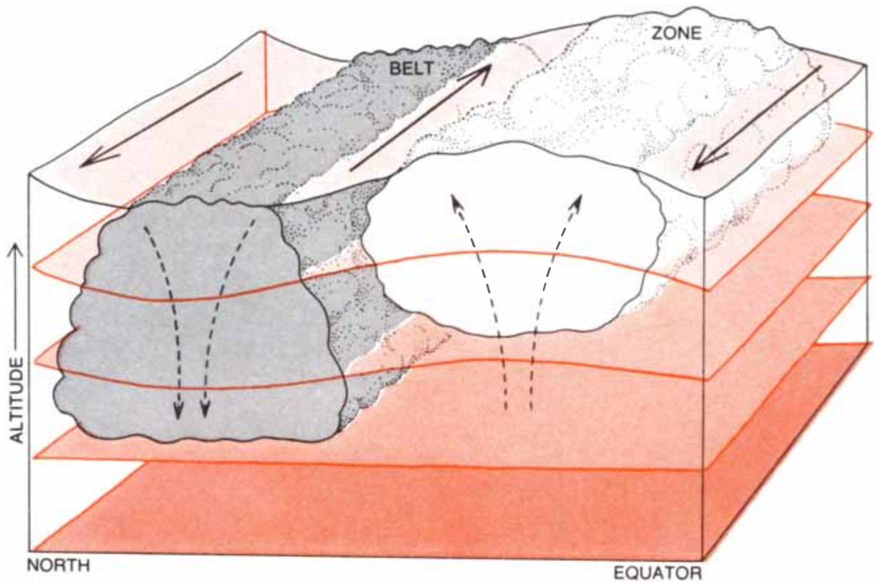
Perhaps the simplest and most direct explanation of the longevity of Jupiter's clouds is one provided by Peter J. Gierasch, who is now at Cornell University, and Richard M. Goody of Harvard University. Gierasch and Goody point out that on Jupiter the radiative time constant is extremely long. The radiative time constant is the time it takes for a mass of air to warm up or cool off by radiating in the infrared portion of the spectrum. On the earth the radiative time constant is a few weeks, which is comparable to the lifetime of flow features in the atmosphere. On Jupiter the radiative time constant is longer than a year. The differ-

INFRARED FEATURES of the atmosphere of Jupiter (*top*) photographed from the spacecraft *Pioneer 10* reveal the distribution of the temperature at the cloud tops. Light areas are regions of low infrared emission and are evidence of thick, high clouds. Dark areas are regions of high infrared emission and are evidence for holes in upper cloud layer exposing warmer layers below. When infrared image is compared with an image made in visible light (*bottom*), it is evident that the regions of high clouds correspond to the light-colored "zones" and that the regions of few clouds correspond to the dark-colored "belts." Since clouds usually form on warm, rising currents in an atmosphere and disperse on cooler, sinking currents, the zones must correspond to regions of warm rising gas and the belts to cooler sinking gas. Great Red Spot can be seen at the lower left in both images; it is an area of particularly low infrared emission and resembles the zones more closely than the belts. On image made in visible light the comblike feature at top left, two square black dots near Great Red Spot and apparent discontinuity running across photograph are artifacts of way in which picture was made.

ence is the result of the lower temperatures in Jupiter's atmosphere, the smaller amount of heat the planet receives from the sun (only 4 percent of the amount the earth receives) and the fact that the gases of Jupiter's atmosphere emit less infrared radiation than the gases of the earth's atmosphere. Furthermore, if the masses of gas in the Jovian atmosphere store their heat at altitudes as low as the base of the clouds, then the radiative time constant may be even longer than a year because of the large volume of gas that must be heated or cooled. Thus the long lifetime of atmospheric phenomena on Jupiter could be largely due to the low rate at which temperature differences are radiated away.

An equally remarkable aspect of Jovian cloud features is their organization and spatial regularity. At any one time there are usually at least 10 bands, the belts and the zones, each circling the planet on a line of constant latitude. Such a high degree of symmetry around the axis of rotation does not appear in satellite photographs of terrestrial clouds.

The belts are either brown or brown tinged with blue, and the zones are either white or white tinged with red. The reddest features, such as the Great Red Spot and smaller spots, are usually located in the zones. Infrared observations that measure the temperature of the cloud tops suggest that there is a basic difference between the zones and the red spots on the one hand and the belts on the other.



REGIONS OF HIGH AND LOW PRESSURE on Jupiter are compared in a cross section of the atmosphere made along a meridian of longitude. Atmospheric pressure increases toward the center of the planet. At great depths, where the atmosphere rotates uniformly with planet's mean rotation period, "surfaces" of constant pressure (*darker colors*) in the atmosphere are horizontal. At intermediate depths the higher temperatures of the zones cause the surfaces of constant pressure (*lighter colors*) to bulge upward, creating higher pressures in the zones (*right*) than in the belts (*left*). Coriolis forces generated by the rotation of the planet then cause the atmosphere to flow in a direction out of the page (westward) along the equatorward edges of the zones, and into the page (eastward) along poleward edges of zones. Slower, secondary circulation is produced in atmosphere as warmer gas rises in zones and cooler gas sinks in belts.

On Jupiter, as on the earth, the most dramatic differences in temperature within a limited area are found between one level of the atmosphere and another. Since in general temperature decreases with altitude, a high infrared temperature indicates that at high altitudes there are only thin, transparent clouds or no clouds at all, and a low infrared temperature indicates that at high altitudes there are relatively thick, opaque clouds. On the earth weather satellites utilize that principle for photographing clouds both in the daytime and at night at infrared wavelengths: a convective storm such as a hurricane always appears as a region of low infrared temperature because of its extensive cover of high clouds.

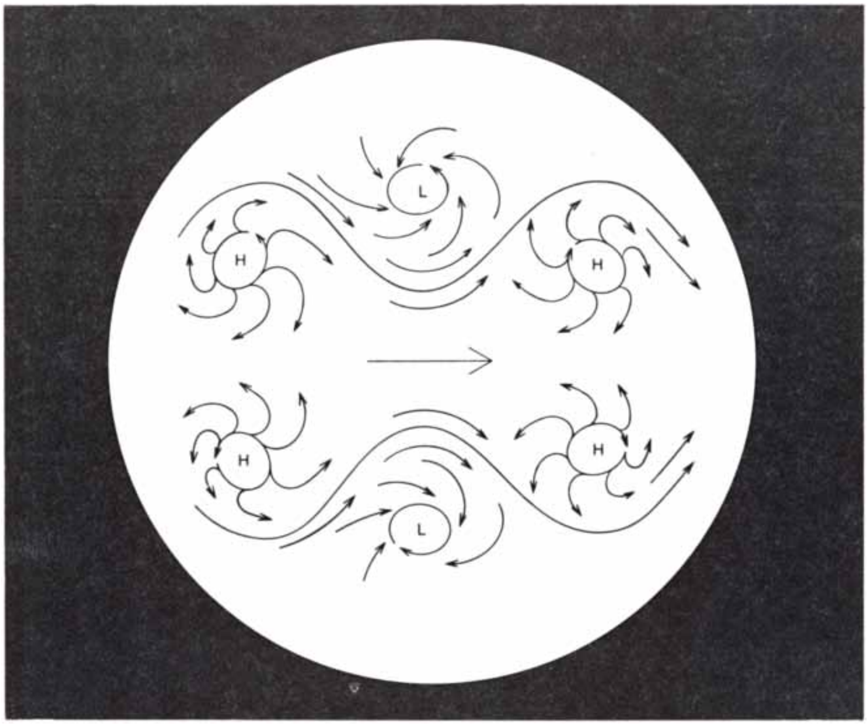
Pioneer 10 made two infrared images as it sped past Jupiter on December 3, 1973. *Pioneer 11* made four infrared images a year later. A comparison of one of the infrared images with a similar image made in the visible region of the spectrum reveals that the zones are regions of low infrared temperature and the belts are regions of high infrared temperature. The Great Red Spot has a particularly low infrared temperature, and therefore it is similar to the zones rather than to the belts. Hence by analogy with the earth one can infer that the red spots and white zones are regions of active convection and rising motion in the atmosphere, whereas the brown belts are regions of cloud dispersal and sinking motion.

Although the infrared observations help to classify Jupiter's atmospheric features, they do not answer the more basic question of why such features exist. Some clues to the answer to that question are provided by observing the motions of the clouds. From the earth observers can chart the position of small spots with respect to the belts and zones and other large features. Photographic records of the features go back about 100 years. The smallest spots that can be seen from the earth are about 3,000 kilometers in diameter, which is about a fiftieth the diameter of Jupiter or approximately the diameter of a large storm system on the earth. Such storm systems on the earth generally move with the mean wind direction; for example, in North America they move from west to east. As the storm systems evolve and decay, however, their apparent motion with respect to the mean motion of the wind sometimes changes. Hence there are problems in using cloud features to track winds on a large scale. Such features are nonetheless the best wind data we have for Jupiter at present.

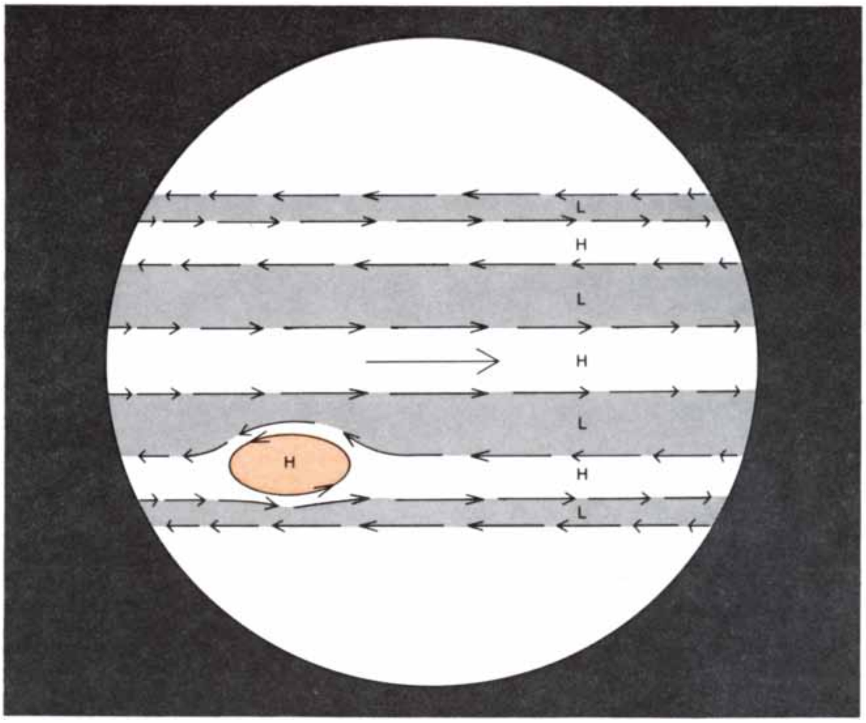
Observed over a period of a few weeks the smaller spots around the Great Red Spot always move counterclockwise; therefore the winds around the Great Red Spot probably blow in that direction. Moreover, the direction and magnitude of the winds are typical of those observed in the same region decades earlier. Elsewhere on Jupiter the wind blows along lines of constant latitude; the largest relative velocities are at the boundaries between belts and zones.

As on the earth, the relative velocities are small compared with the mean velocity of the atmosphere associated with the planet's rotation. The rotation of Jupiter's interior can be determined from the rotation of the planet's magnetic field, which has a period of nine hours 55 minutes 29.7 seconds. The same rotation period is typical of the clouds in Jupiter's middle latitudes, although the winds at the edges of belts and zones cause small regions to depart from the mean rotation period by as much as five minutes. Within a band of latitudes extending some eight degrees on each side of the Jovian equator the rotation period of the atmosphere is about five minutes shorter than the rotation period of the magnetic field. The shortest rotation period of all is found away from the equatorial region. It is associated with the boundary between a prominent belt and a zone in the northern hemisphere. There the winds complete a circuit of the planet once every nine hours 49 minutes. At that boundary the velocity of the clouds relative to the velocity of Jupiter's rotating interior is about 120 meters per second, or more than 270 miles per hour.

Whenever atmospheric motions persist for intervals that are long compared with a planet's period of rotation, the Coriolis force must play an important role in the dynamics of the atmosphere. In the northern hemisphere of Jupiter or the earth the Coriolis force acts to the right of the direction of motion. In order to balance the force there must be a pressure force to the left, that is, there must be a high-pressure area to the right of the flow direction. Thus on the earth the wind blows counterclockwise around a low-pressure center (forming a cyclone) in the Northern Hemisphere and clockwise around a high-pressure center



EARTH'S ATMOSPHERE IS BAROCLINIC on a global scale, that is, there are large horizontal temperature differences between the Equator and the poles. As a result of the differences a longitudinal wave forms in the atmosphere in both hemispheres, traveling from west (*left*) to east (*right*) and transporting heat from the Equator to the poles. The troughs (equatorward excursions) of the wave are associated with low-pressure regions (*L*) and the crests (poleward excursions) of the wave are associated with high-pressure regions (*H*). The flow of air around low-pressure regions is cyclonic and flow of air around high-pressure regions is anticyclonic.



JUPITER'S ATMOSPHERE IS BAROTROPIC on a global scale, that is, there is very little horizontal temperature difference between the equator and the poles. Regions of high pressure (*H*) and low pressure (*L*) are linear features along parallels of latitude: the zones (*white*) and the belts (*gray*). High-speed winds flow at the boundaries between the bands; where relative wind motion is westward (*toward the left*) the winds are known as retrograde jets. Above a latitude of 45 degrees in both hemispheres banded pattern breaks down (*see illustration on page 54*). Color oval in southern hemisphere is Great Red Spot. Winds around it are anticyclonic.

(forming an anticyclone). The pattern is reversed in the Southern Hemisphere. For winds moving in latitudinal zones (as at the boundaries between Jupiter's belts and zones) the highs and lows are found in bands between the regions of maximum and minimum zonal velocity.

On Jupiter the zones and the Great Red Spot are high-pressure regions (anticyclonic) and the belts are low-pressure regions (cyclonic). That was first pointed out in 1951 by Seymour L. Hess of Florida State University and Hans A. Panofsky of New York University. Hence the zones and the Great Red Spot seem to be fundamentally different from terrestrial storms, which are usually cyclonic at sea level. The difference is not, however, as fundamental as it seems. Since clouds tend to form in rising air, and since rising air tends to be warm, it is reasonable to assume that the zones and the Great Red Spot are warmer than their surroundings at any particular level within the clouds. In that respect they resemble tropical cyclones (hurricanes) and mature extratropical cyclones on the earth, most of which are also warm. The resemblance is significant because terrestrial air masses that are warm tend to have high-pressure centers and anticyclonic circulation at high altitudes, the altitudes to which the Jovian observations refer.

The reason that such storms are anticyclonic at great heights is that the pressure drop with altitude in a warm air mass is less than the pressure drop with altitude in a cold one. This fact is a consequence of the hydrostatic relation between pressure change and density: when the density is low, as it is when the air is warm, the pressure drop with altitude is also low. Thus a warm air mass tends to become a high-pressure (anticyclonic) region with increasing altitude. If the pressure is very much lower than the surroundings at low altitudes, as it is with a terrestrial cyclonic storm, the transition to anticyclonic circulation may not arise. In general, terrestrial hurricanes are strongly cyclonic at sea level and weakly anticylonic at high altitudes. In other words, the observed anticyclonic circulation of the Jovian zones and the Great Red Spot is consistent with the evidence from infrared observations that they are warm centers of rising motion. Hence in one way they are similar to warm convective storms on the earth.

As we have seen, however, there is a fundamental difference between the earth and Jupiter in that the earth's atmosphere is bounded at the bottom by a relatively undeformable surface of water and land. The surface can support large differences in air pressure at sea level from one place to another, and so there can be strong winds close to the ground. We have little information about the atmosphere below the visible clouds on Jupiter, but it is reasonable to assume that below some level it rotates uniformly with a period equal to the rotation period of the magnetic field. The situation is equivalent to saying that the atmosphere

FEATURE	INFRARED MEASUREMENTS	CLOUD HEIGHT	VORTICITY	PRESSURE	TEMPER-ATURE	VERTICAL VELOCITY	EXPECTED CLOUDS	COLOR
BELT	HOT	LOW	CYCLONIC	LOW	COLD	DOWN	LOW, THIN	DARK
ZONE	COLD	HIGH	ANTICYCLONIC	HIGH	HOT	UP	HIGH, THICK	LIGHT
GREAT RED SPOT	COLD	HIGH	ANTICYCLONIC	HIGH	HOT	UP	HIGH, THICK	ORANGE

CHARACTERISTICS OF FEATURES in the Jovian atmosphere are summarized. The zones and the Great Red Spot are similar in all important respects except shape. The belts are the reverse of the zones in all respects. Taken together the data suggest that all the features are not different and isolated phenomena, and that they all must be linked together as part of one global weather pattern.

has a fluid lower boundary that deforms easily when pressure is applied from above. The deformation prevents the buildup of pressure differences at deep levels and therefore keeps the circulation at those levels weak. Thus whereas a warm air mass on the earth can have either cyclonic or anticyclonic circulation at low altitudes, on Jupiter the low-altitude circulation must be zero. In that respect the circulation on Jupiter may resemble currents in the earth's oceans more closely than currents in its atmosphere. Ocean currents tend to be weak at great depths and are always anticyclonic near the surface if the water between the two levels is warm.

There have been several attempts to put these arguments on a quantitative basis. In 1969 Jeffrey Cuzzi, who was then a graduate student at the California Institute of Technology, and I decided to study the observed zonal wind patterns on Jupiter. We first rediscovered the qualitative relation discussed by Hess and Panofsky in 1951. We then made a rough estimate of how great a difference in temperature between zones and belts would be needed to account for the observed winds. Actually the quantity one wants to determine is not the temperature difference alone but rather the product of the average temperature difference and the depth to which the difference extends. If the observed zonal motions refer to the tops of the ammonia clouds in Lewis' model of the Jovian atmosphere, then the depth must be measured downward from that level.

On the basis of the observed Jovian winds the value of the product of the temperature difference and the depth is about 150 kilometer-degrees. In other words, the actual depth over which there are appreciable temperature differences between the zones and the belts is unknown, but if the depth were 15, 150 or 1,500 kilometers, the temperature difference at those depths would have to be respectively 10 degrees, 1 or .1 degree C. in order to account for the observed winds. The magnitude of the temperature difference is related to the heat sources and heat sinks that are maintaining the atmospheric circulation. The critical depth is the thickness of the source-sink region. To discuss the situation further we need to consider the source-sink mechanism.

In 1971 Gierasch, who was then at Florida State University, and Albert I. Barcilon suggested that Jupiter's atmosphere is similar to the earth's atmosphere over the Tropics. There the sun heats the surface of the ocean, causing the water to evaporate. The moist air close to the surface becomes unstable, leading to convection and the formation of cumulus clouds. Large-scale, organized motions are present over the Tropics because the small-scale cumulus convection, which is the principal means by which heat is transferred from the ocean to the atmosphere, varies in response to the large-scale motions. The maximum horizontal temperature difference from one place to another in the Tropics is found between air that has been heated by condensation and air that has not.

According to Gierasch and Barcilon, the temperature difference between the zones and the belts on Jupiter is due to the release of latent heat, as are the temperature differences over the earth's Tropics. Lewis' model of the clouds is assumed to apply to the Jovian zones; the belts are assumed to be dry, without any condensable vapors. The zones are then about two degrees hotter than the belts at all levels above the level where water vapor condenses. (Compared with the clouds composed of water, the clouds of ammonium hydrosulfide and of ammonia have only a small effect on the temperature.) Thus a temperature difference of about two degrees extends over a depth of about 75 kilometers, which is the total thickness of the water-ammonia cloud system. Moreover, the product of the temperature difference and the depth agrees with the 150 kilometer-degrees deduced from the wind observations.

Gierasch and Barcilon did not offer a mechanism by which condensable vapors remain concentrated in the zones. In a tropical hurricane on the earth the cyclonic winds exert a cyclonic stress (force per unit area) on the ocean surface. That stress, interacting with the Coriolis force, causes air to flow inward toward the low-pressure center of the hurricane just above the surface and causes water to flow outward just below the surface. On the earth the flow in the ocean can be ignored, but the convergence of moist air toward the center of the hurricane renews the hurricane's supply of latent heat and maintains its circulation.

On Jupiter the cloudy regions are anticyclonic and rotate in a direction opposite to that of cloudy regions on the earth. There is no ocean; the lower fluid is simply the atmosphere below the clouds. An anticyclonic stress from above on the lower fluid causes the lower fluid to flow inward and the upper atmosphere to flow outward. Here, however, it is the convergence of the lower fluid toward the center that matters. The convergence occurs below the base of the clouds, at levels where precipitation is evaporating. It is this convergence that continuously resupplies the zones with condensable vapor. Such a process is one possible mechanism by which the temperature difference between zones and belts is maintained.

So far I have said nothing about the circulation of the atmosphere in middle latitudes on Jupiter and on the earth. On the earth the basic energy source of the atmosphere at middle latitudes is the horizontal temperature gradient. The gradient arises because the sun heats the Equator far more than it heats the poles. An atmosphere with such horizontal temperature differences at any given altitude is said to be baroclinic. It is usually unstable. The instability manifests itself as a wavelike pattern in the basic eastward flow of the atmosphere. Each trough of the wave is the point where the air makes its maximum excursion toward the Equator and each crest is where it makes its maximum excursion toward the poles. The wave transports heat both upward in the atmosphere and toward the poles. The upward heat transport, in which hot air rises and cold air sinks, tends to lower the center of mass of the atmosphere, since a given volume of cold air is denser and therefore heavier than the same volume of hot air. The heat transport thereby releases gravitational potential energy. The troughs and crests of the baroclinic wave are respectively the cyclones and anticyclones that dominate the earth's weather at the middle latitudes.

An important question is whether or not such baroclinic waves are present on Jupiter. If the heat flux from Jupiter's interior were uniform from the equator to the poles, the total heat supplied to the atmosphere from both the sun and the interior would still be twice as great at the equator as at the poles, because of the difference in the degree to which those regions are heated by the sun. Such a difference in heating could cause a considerable amount of baroclinic instability if Jupiter were like the earth at middle latitudes.

The axially symmetrical, banded appearance of Jupiter suggests, however, that Jupiter must be quite different from the earth at those latitudes. Mixing across circles of latitude is an essential aspect of baroclinic instability. And although there are some disturbances resembling baroclinic waves on Jupiter, they are confined to narrow bands of latitude and do not seem to give rise to large-scale mixing between one band and another. In addition, if heat transport toward the poles by baroclinic instability were the only means of balancing the extra amount of solar heating at Jupiter's equator, an appreciable temperature difference between the equator and the poles would develop.

Peter H. Stone of M.I.T. has a general theory of baroclinic instability in any planetary atmosphere. For Jupiter he estimates that the equator would be 30 degrees hotter than the poles if the solar heat were transported poleward by baroclinic instability alone. His estimate was recently tested when the infrared radiometers on *Pioneer 10* and *Pioneer 11* obtained the first good heat measurements of the Jovian poles.

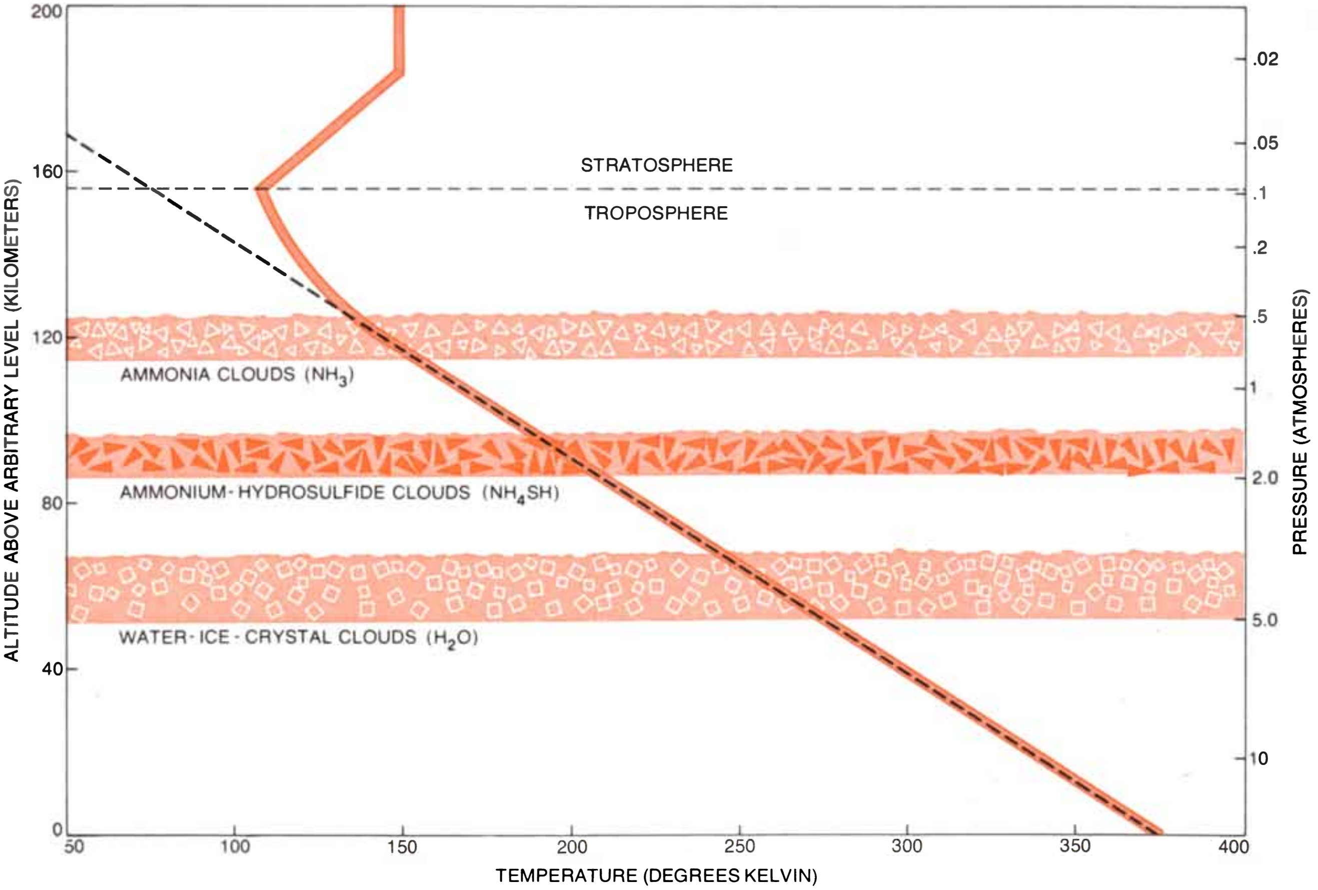
The measurements showed that the heat emitted by Jupiter is about the same over both the poles and the equator. This implies that the poles are at about the same temperature as the equator at the same levels of pressure in the atmosphere. In fact, the observed temperature difference is no greater than three degrees, as compared with the value of 30 degrees predicted by Stone's baroclinic-instability model. The implication is that Jupiter is not like the earth at middle latitudes, and that the transport of heat in the Jovian atmosphere by baroclinic instabilities is insignificant.

If that is the case, the difference in solar heating between the equator and the poles of Jupiter must be balanced by heat from the planet's interior. Recently I suggested a mechanism whereby the difference could be equalized by the convective flow of the internal heat. The mechanism works like a thermostat. Where the actual vertical temperature gradient in the atmosphere is equal to the adiabatic value the convective heat flux is zero. Where the temperature falls with altitude at a rate slightly higher than the adiabatic rate, however, the convective heat flux is large. Thus a slight cooling of the poles relative to the equator because of the difference in amount of solar energy absorbed soon leads to a large increase in the convective heat flux to the poles relative to the flux to the equator. The convective thermostat ensures that Jupiter is very nearly adiabatic throughout. Hence the relation between temperature and pressure at both the equator and the poles must be very nearly equal. The possibility that there is greater convective flow of internal heat at the Jovian poles seems to be confirmed by the appearance of convectionlike patterns north of a latitude of about 45 degrees.

The internal heat source might also play a role in supplying energy for the circulation observed within the Jovian belts and zones. It has been suggested that the belts and zones are simply surface manifestations of giant convection cells. The giant cells might be as deep as they are wide, and therefore they might extend an appreciable fraction of Jupiter's radius downward toward the center of the planet.

There are several arguments against the convection-cell hypothesis, although none of them is conclusive. First, if Jupiter's atmosphere is indeed analogous to the earth's tropical atmosphere, it does not need a deep, large-scale circulation below the clouds. Condensable vapor could be concentrated in the zones and dispersed in the belts in a shallow layer just below the base of the clouds. The temperature differences arising from the condensation within the clouds seem fully capable of accounting for the observed winds. Therefore the belts and zones could well be a superficial phenomenon, extending only slightly deeper than the base of the deepest clouds.

Second, if the belts and zones were associated with the large-scale convection of internal heat, one would expect that the infrared emission from Jupiter would be strongest in the zones, since the zones are the site



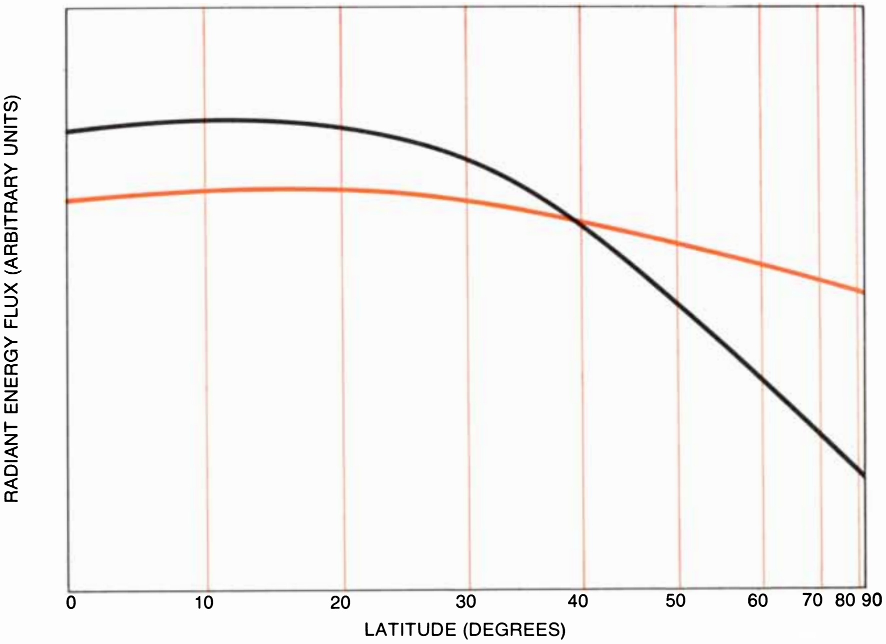
VERTICAL CLOUD STRUCTURE of the atmosphere of Jupiter has been computed from a theoretical model devised by John S. Lewis of the Massachusetts Institute of Technology. In an atmosphere composed primarily of the noncondensable gases hydrogen and helium there are distinct layers of clouds. The color line indicates the temperature of the atmosphere at various depths and pressures; it is based on an analysis of infrared data made by Glenn S. Orton of the Jet Propulsion Laboratory of the California Institute of Technology. The black broken line indicates the theoretical temperature of the atmosphere at the same depths and pressures if the atmosphere were perfectly adiabatic (completely mixed by convection). The planet's atmosphere is nearly adiabatic except at uppermost levels.

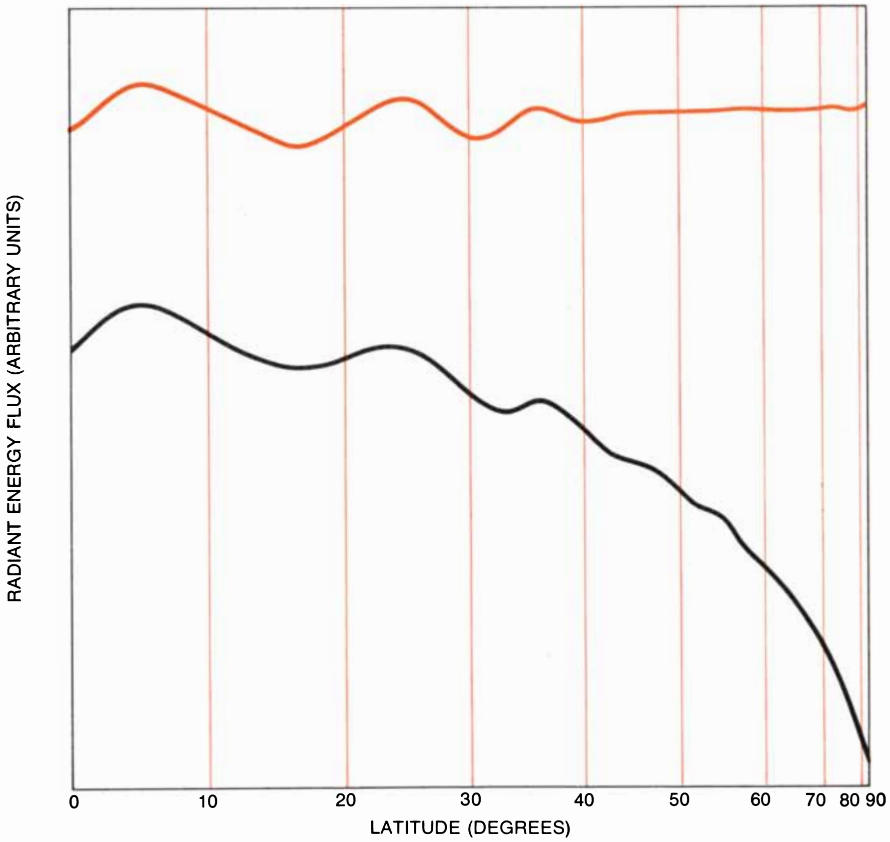
of rising motion. It is true that we can observe the heat flux only at the top of the atmosphere; nevertheless, the zones do not seem to be the principal regions of infrared emission. In fact, it is the belts that both emit the greater amount of infrared energy and absorb the greater amount of solar energy. On the other hand, the net flux, that is, the difference between the amount of energy emitted and the amount absorbed, is about equal in the zones and the belts. This equality implies that the internal heat flux at the base of the clouds is the same in the belts and the zones, provided that there is no latitudinal heat transport within the clouds. The point is that the radiative-flux data provide no evidence to support the convection-cell hypothesis, even though they do not rule it out.

Third, the convection-cell hypothesis has more than its share of theoretical difficulties. Convection is typically a small-scale phenomenon. In the sun or in the earth's atmosphere the size of the dominant energy-carrying cells is on the order of one scale height, that is, the vertical distance in which the density and the pressure increase by a factor of *e*, or 2.718. On Jupiter the scale height is a few tens of kilometers, and the belts and zones are some 10,000 kilometers wide. Moreover, on a rotating body convection patterns tend to develop a longitudinal structure, just the opposite of what is observed on Jupiter. Such arguments suggest that small-scale convection is the dominant mode of heat transfer, at least below the clouds, where radiation and condensation are unimportant.

There are other ways, however, in which Jupiter's belts and zones might be part of a deep pattern of convection extending down a significant fraction of the planet's radius. Several theoretical and laboratory studies have shown that convection in a rotating sphere usually takes the form of long, thin columns, the axes of which are parallel to the axis of rotation. The two ends of each column intersect the visible surface of the sphere at opposite latitudes in the northern and southern hemispheres. Friedrich H. Busse of the University of California at Los Angeles has suggested that the belts and zones are the surface manifestation of such long, thin columns. So far the columns have been shown to exist only in laboratory experiments with comparatively viscous liquids. Extending the hypothesis to cover the compressible gases on a rotating sphere the size of Jupiter is therefore somewhat speculative.

Whether the belts and zones are a shallow phenomenon confined to the cloud layers or are part of a deep convection pattern, there is still no satisfactory explanation of why they are oriented in the east-west direction. The earth's tropical atmosphere also is banded, with a cloudy zone around the Equator and clear belts to the north and south. The equatorial intertropical convergence zone is a band of cumulus clouds and thunderstorms that generally occupies a strip some five to 10 degrees north of the





RADIATION BALANCE in the atmosphere of the earth (*top*) and of Jupiter (*bottom*) shows dramatic differences between the planets. For both planets the amount of thermal radiation emitted (*color*) is nearly independent of latitude, whereas the amount of sunlight absorbed (*black*) is strongly affected by the oblique angle of the sun at the poles. The earth has a negligible source of internal heat, and so the areas under the two curves for the earth are equal. The excess amount of solar heating at the earth's Equator is carried poleward by currents in the atmosphere and ocean. Jupiter has an appreciable source of internal heat: it emits 1.9 times as much energy as it receives from the sun. Thus the area under the color curve for Jupiter is 1.9 times larger than the area under the black curve. In addition very little heat is carried poleward by currents in Jupiter's atmosphere. Hence Jupiter's internal heat must reach the surface preferentially at the poles. The smaller bumps in the curves for Jupiter are created by the belts and zones; more thermal radiation is emitted and more sunlight is absorbed in the belts than in the zones. Curves for both planets are averaged relative to longitude, time of day and season.

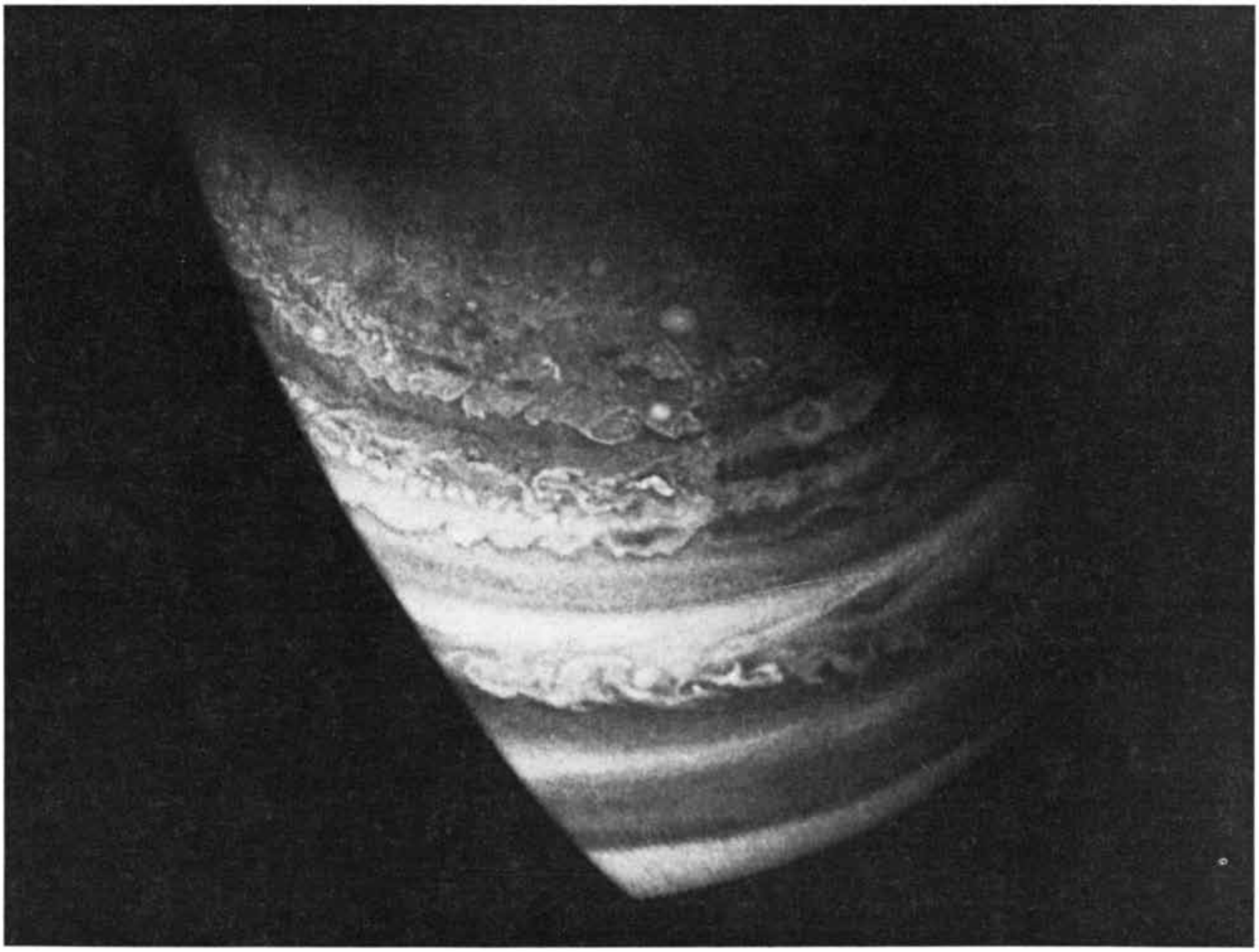
Equator. It is a region of net upwelling in the atmosphere, even though heat is carried upward to the base of the stratosphere by a system of both updrafts and downdrafts. On each side of the intertropical convergence zone there is a dry belt extending to 30 degrees north and south of the Equator. There the motion is almost everywhere downward; updrafts are confined to a thin layer close to the ground. The skies are generally clear and heat is transferred vertically mainly by infrared radiation. The northward offset of the intertropical convergence zone is apparently due to the fact that the continents and oceans are distributed differently in the Northern and Southern hemispheres.

The difference between Jupiter and the earth seems to come at higher latitudes. On Jupiter the banded pattern repeats itself several more times; on the earth the pattern breaks up into baroclinic instabilities. The difference may be due to the thermostatic effect of the internal heat source on Jupiter. On the earth baroclinic instabilities provide the only way for the atmosphere to balance the solar heating: the air over the Equator heats up relative to the air over the poles, instabilities develop and soon the middle-latitude circulation is dominated by cyclones and anticyclones. On Jupiter the equatorial regions heat up only slightly relative to the polar regions, and the internal heat flux readjusts itself by an amount that is sufficient to balance the difference in the solar heating.

The instability of the belts and zones is an interesting question in its own right. Any equilibrium state of an atmosphere, except the state of uniform rotation, has kinetic energy and gravitational potential energy. A small-amplitude disturbance may be able to extract some of that energy and grow at the expense of the basic state. The basic state is stable only if all the disturbances remain small. It is unstable if there is at least one kind of disturbance that continues to grow. In any real situation all kinds of disturbances are always present to some extent, so that an unstable basic state always breaks up or evolves in a finite time.

The simplest basic state is one in which an atmosphere flows strictly from west to east and the wind velocity varies only with latitude. Superficially that is the case on Jupiter, although it may not be valid to neglect the dependence of velocity on altitude. The only available energy in that basic state is kinetic, and it is associated with the different velocities at the different latitudes. There are no horizontal temperature gradients from one place to another at the same level. The atmosphere is said to be barotropic, as opposed to baroclinic, where the available energy is mainly gravitational and is associated with horizontal temperature gradients.

The different velocities of the eastward winds can be plotted with respect to latitude to yield a velocity profile. For a barotropic atmosphere an instability is possible only when the curvature of the velocity profile exceeds a certain critical value. That value is equal to twice the cosine of the latitude times the planet's rotation rate divided by the radius. In 1969 Cuzzi and I applied the stability criterion to Peek's long-term data on Jupiter's rotation period with respect to latitude. We found that according to this criterion there are only a few latitudes on Jupiter where an instability is even marginally possible. The instabilities are found at the center of retrograde jets in



PATTERN OF BANDS BREAKS DOWN at high latitudes on Jupiter, as is shown by this high-resolution image of the planet's northern hemisphere made from *Pioneer 11*. The equator is the second light zone from the bottom. The disturbed region to the north is at a latitude where the basic flow between a belt and a zone is theoretically expected to be unstable. Farther north such an instability dominates at all latitudes. Borders of picture are distorted because of rotation of planet and motion of spacecraft during time picture was made.

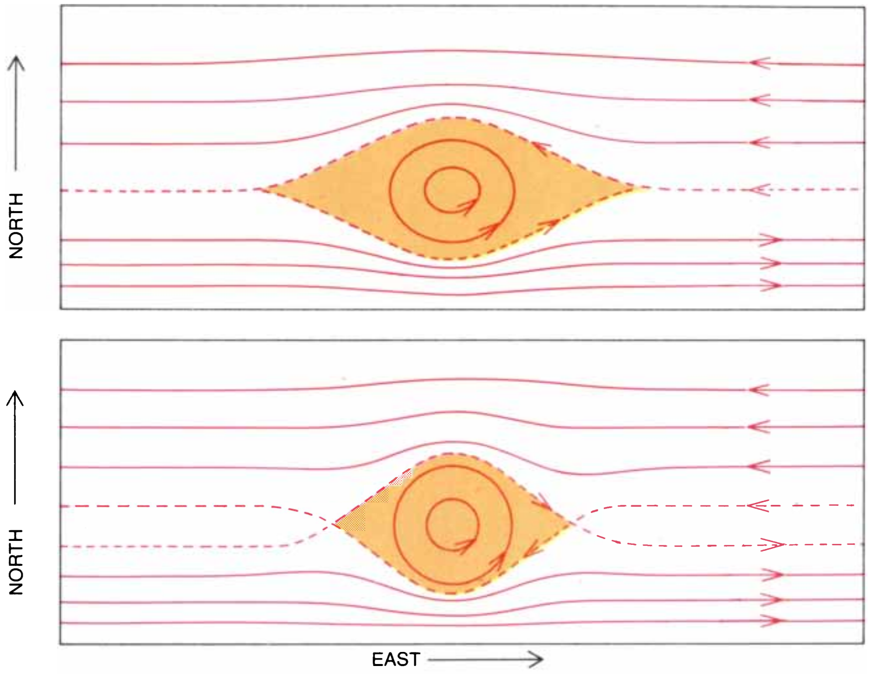
Jupiter's atmosphere: latitudes where the rotation period is the longest. The disturbance that grows the fastest is a longitudinally oscillating wave. Thus we expect the basic east-west flow to break up into a wavy pattern in the vicinity of the retrograde jets whenever the stability criterion is exceeded. When the stability criterion is not exceeded, the east-west flow pattern remains constant in time.

A confirmation of sorts was provided by *Pioneer 11* in its closeup pictures of the Jovian northern hemisphere. The pictures show several bands where a wavelike pattern suggests the location of an instability. The most prominent patterns are seen on the equatorward edges of zones (the poleward edges of belts). According to Peek's data, those same positions are the location of the retrograde jets. In other words, the instability seems to be developing where the barotropic-stability criterion says it should.

The *Pioneer 11* pictures also suggest that the parallel flow of the belts and the zones breaks down entirely at latitudes higher than about 45 degrees. That is at least qualitatively consistent with the barotropic-stability criterion, because the critical value of the curvature is proportional to the cosine of the latitude, and the cosine of the latitude drops to zero at the poles. Hence according to the stability criterion, a pattern of east-west bands at the poles would be unstable, and therefore such a pattern could not be a permanent atmospheric feature.

A north-south pattern is different. The most general north-south pattern is a mixture of waves that also propagate westward at various speeds. The difference between the north-south flow patterns and the east-west ones seems to explain the recent computer results of Gareth P. Williams of the Geophysical Fluid Dynamics Laboratory at Princeton University. He began with the barotropic model of Jupiter's atmosphere and introduced into the equations a term that described a randomly varying force and a term that described the dissipation of energy attributable to the viscosity of the atmosphere. He found that initially both the east-west and the north-south motions appeared in his model. Later the north-south motions continued to form and decay, whereas the east-west pattern grew slowly to the limiting amplitude imposed by the stability criterion. By adjusting the balance between the randomly-varying-force term and the dissipation term, Williams was able to produce a quite realistic-looking computer model of Jupiter [*see illustration on page 47*].

It is still hard to understand why the barotropic model should be so successful in explaining Jupiter's atmospheric features. The deep atmosphere must rotate with the interior, so that the speed of the currents at the boundaries of the belts and the zones must vary with depth. Associated with such a variation with depth there must be temperature differences between the belts and the zones on horizontal surfaces. In other words, the Jovian atmosphere should be baroclinic, not barotropic. The barotropic



COMPUTER MODEL OF GREAT RED SPOT (*top*) in Jupiter's atmosphere duplicates several of the atmospheric flow features that are observed. It shows the counterclockwise winds around the spot and the proper directions of the east-west flow of the gas north and south of the spot. Moreover, it shows pointed tips at the east and west edges of the spot. A variation of the model (*bottom*) shows how smaller dark spots that occasionally approach the Great Red Spot can sometimes suddenly change latitude and begin to recede, carried by a current of gas back in the direction from which they came. In both versions of the model the horizontal scale of the diagram has been compressed by a factor of three with respect to the vertical scale.

model would still be a valid approximation if the actual lapse rate were very much less than the adiabatic lapse rate, or if temperature increased with altitude. Those, however, are unlikely possibilities; both theory and observation suggest that the actual lapse rate is close to the adiabatic lapse rate except in the stratosphere, which is above the region of cloud motions. Perhaps the success of the barotropic model is due to our overzealous efforts to match the observational data with a theory that is based on terrestrial experience. The real mechanisms at work on Jupiter might still be eluding us.

Let us now turn to the Great Red Spot and similar spots in Jupiter's atmosphere. For a number of years the only theory that seemed capable of explaining the Great Red Spot was the Taylor-column hypothesis put forward by Raymond Hide, who was then at M.I.T. [see "Jupiter's Great Red Spot," by Raymond Hide; SCIENTIFIC AMERICAN, February, 1968]. Because of the Great Red Spot's long lifetime, its constancy in latitude and its uniqueness it seemed that it must be connected with an underlying solid object or topographic feature that was giving rise directly to the flow patterns at the visible surface. A Taylor column is the cylinder of stagnant fluid that was believed to join the solid object to the red cloud we see at the top of the Jovian atmosphere.

Since it now seems likely that Jupiter has no solid surface at any depth, the Taylor-column hypothesis is not as compelling. Moreover, the recent observations have brought out the similarity between the Great Red Spot and the zones. In fact, the spot's long lifetime and its constancy in latitude are no more unique than the long lifetime and the constancy of latitude of the zone in which it is embedded. The Great Red Spot also drifts irregularly westward relative to Jupiter's magnetic field, which would presumably be rooted in the solid surface of the planet if there were one. Finally, other zones seem to have their own red spots, suggesting that the Great Red Spot is not unique.

The prevailing view now is that all the red spots, like the zones, are basically a meteorological phenomenon. Since the Great Red Spot has higher clouds and a more strongly anticyclonic circulation than any of the zones, it could be called a superzone. If the zones and red spots are driven by the latent heat of condensation, then it could also be called a Jovian hurricane.

As with a terrestrial hurricane, there is probably a net rising motion within the Great Red Spot and a relatively uniform sinking motion in the regions immediately outside it. By far the highest velocities, however, are associated with the horizontal flow around the spot; the vertical motion of the atmosphere and its associated spreading motion at the cloud tops are too small to be measured. Perhaps the vertical heat transfer and energy release, and the energy dissipation as well, are small. Therefore to a first approximation one might neglect all the dissipative processes and ask whether there is an adequate model of the horizontal flow

SMALLER RED SPOT resembling the Great Red Spot is shown in this image made from *Pioneer 10*. Such spots, lasting for about two years, are not uncommon and are usually found in the zones. Their similarity to the Great Red Spot suggests that all such features are purely long-lived meteorological phenomena and are not associated with any solid feature within Jupiter.

GREAT RED SPOT, photographed from *Pioneer 11*, is now believed to resemble a terrestrial hurricane. The observed flow of the atmosphere in the zone and around the spot is anticyclonic, implying that the pressure within the spot is high. The flow near the top of a terrestrial hurricane is also anticyclonic, implying that it too is a high-pressure region at high altitudes. The Great Red Spot has been a feature on Jupiter for some three centuries. Hydrodynamic models of the spot show that in the absence of the dissipation of energy such a flow feature could in principle last forever. Even a small amount of dissipation will not affect it appreciably, and on Jupiter the energy that would need to be dissipated could be completely provided by the release of latent heat from the rising gas within the zone condensing to form clouds.

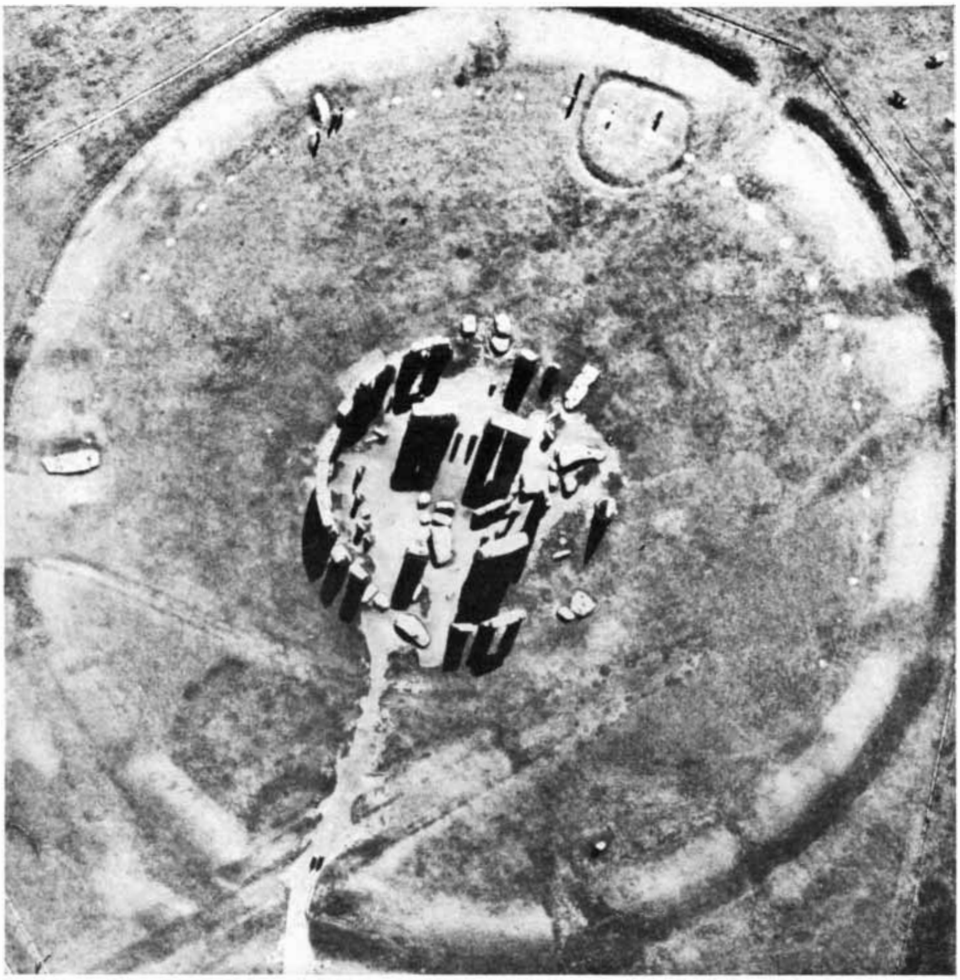
around the spot. Once again the simplest case is for a barotropic atmosphere. The aim of such a study would be to see if there are any special conditions under which red spots can or cannot exist, and to see if the model can predict any details of the flow pattern of such spots that can be compared with observation.

The Great Red Spot is embedded in a zone in the Jovian southern hemisphere, and it rotates counterclockwise like a wheel between two oppositely moving surfaces. Northward (equatorward) of the spot, along the northern edge of the zone, the flow is to the west. Southward (poleward) of it the flow is to the east. In 1970, working with a computer model, I found that the rolling-wheel configuration was the only one that gave a steady (time-independent) configuration of flow. The configuration shows that anticyclonic spots such as the Great Red Spot can exist only in an anticyclonic linear feature such as a zone. That aspect of the computer model is confirmed in the association between red spots and zones in both the northern and the southern hemisphere. The computer model also predicts that the Great Red Spot should have pointed tips at its east and west ends where the opposing wind currents of the zone divide. That aspect is confirmed in the close-up pictures of the spot, where the pointed tips are clearly visible. Recently Tony Maxworthy and Larry G. Redekopp of the University of Southern California have duplicated additional features of the flow around the Great Red Spot, including the interaction of other large spots with it over a long period of time. The basic flow is assumed to be barotropic, so that this model suffers from the weakness common to barotropic models: the actual lapse rate must differ from the adiabatic lapse rate by an amount that is inconsistent with other data on Jupiter's atmosphere.

All the studies of the Great Red Spot demonstrate that such flow features are possible in a steady state without a solid surface to define the shape of the pattern. In the absence of energy dissipation such a flow feature will last forever. Introducing a small amount of dissipation, as is appropriate for Jupiter, will not change the flow patterns appreciably. The energy that would need to be dissipated might be completely provided by the convergence between moist and dry gas and the release of latent heat.

Someday I should like to develop a computer model that works for both Jupiter and the earth. With the appropriate energy sources and sinks, with more stringent boundary conditions, with allowances made for Jupiter's internal heat source and so on the model should realistically predict the behavior of the atmosphere of either planet. Having such a model would mean that we understood terrestrial meteorology at a much deeper level, since the universality of the principles involved would have been tested by their applicability in two independent atmospheric systems.

We want to be useful...and even interesting

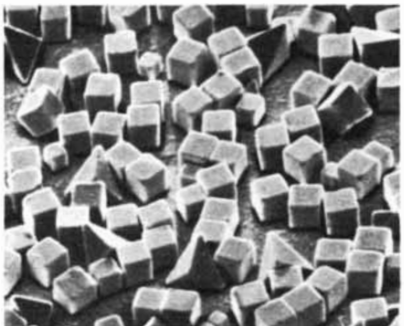


Kenting Earth Sciences Limited, Ottawa, Canada

Silver halide imagery, among other uses, illuminates mankind's path from its past towards its future. Few traces of our forebears are as obvious as Stonehenge. Nor does aerial photography merely lead the ethnographers and archaeologists to interesting sites. They need it as much to see why the sites were chosen. Their disciplines are useful, with pertinence to any regional planning commission caught between the environmentalists and the booster types.

Homo sapiens uses grass fires or nuclear fission, a lichen uses rock-dissolving acids, and both organisms change the surface of Earth in pursuing their respective needs. Seeking a longer perspective and better guide to understanding human needs than the bottom line on an accountant's balance sheet, scholars study photographs of our old planet's surface for subtle clues left in geologically recent times by our kind. The aerial camera is the tool for surface study, just as the shovel and brush are for vertical exploration, but the price of air-survey contracts sometimes frustrates scholars when dealing with people interested in balance sheets.

Let them take heart. All existing aerial photography has not yet been drained of ethnographic excitement. For the sparks to fly, the photography and a specific conceptual framework must strike together. If you've got the concepts, we may have hints on where to find the photography. Just please don't explain the concepts. Our mind is on a balance sheet. State question to E. G. Tibbils, Aerial Mapping Markets, Kodak, Rochester, N.Y. 14650.



Silver halide itself is to colloid science and surface chemistry almost what the colon bacillus and the fruit fly are to genetics. For doing science, a dispersion like this of near-perfect AgBr cubes—ranging narrowly in edge length size around 4 micrometers—has come a long way from the wild type of AgX crystal on which practical photography based itself until rather recently. This is quite aside from the uniquely retentive memory that such a crystal has for a few photons of light. The practical consequences thereof have kept numerous scientists the world around in meat and potatoes. So intensely has interest been focused on this stuff that scientists pursuing pure, fundamental studies in adsorption phenomena, dielectric loss, electrophoresis, light scatter, etc. like to work with silver halide and its extensive literature. If you want the name of someone whose interests run in that direction, try Dr. A. H. Herz, Kodak Research Laboratories, Rochester, N.Y. 14650.

Kodak

Before you buy any computer or peripheral, read about these men.

They've bought hundreds from Digital.

"We depend on over 100 computers so New Yorkers can depend on us."

That's Al Feyler speaking for the Planning and Mechanization Department of the New York Telephone Company, New York City region.

"Our region handles 7 million main stations and some 30 million

"After 4 solid months of competitive testing, we found nothing could compete with Digital."

The business is leasing data communications equipment. The company is Leasco Data Communications Corp., a subsidiary of Reliance Group.

Leasco has put thousands of data communications systems in hundreds of companies all across the country. As part of its agreement, Leasco provides fast maintenance service, no matter where

"We didn't get a computer, we got a company."

Tano Corporation is one of the hundreds of OEM customers that put Digital computers in their products. James Reiss, President of the company, explains the reasons for going with Digital.

"Our company designs complex electronic control systems for oil and gas pipelines, chemical plants, utilities and municipal services. We needed a company that could provide sophisticated hardware and software to match."

At their present rate of growth, Tano expects to double their already significant PDP-11 pur-

Saving managers millions.

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

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

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

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

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

Today anyone who has a good applications idea for a computer can afford one. And just as important, they can get the help they phone calls daily. Computer power helps us provide our customers with the best possible service."

In only 4 years, Mr. Feyler's department has installed over 104 computers. 76 from Digital Equipment Corporation.

"Our computers perform a wide number of administrative and maintenance operations. Among other things, they report network and switching problems, monitor traffic flow, analyze trunk call data, make equipment assignments, switch messages, and report coin box troubles.

"In short, they help keep one of the world's most complicated communications networks running smoothly."

As far as the future for New York Telephone and New York Telephone customers is concerned, Mr. Feyler explains: "Right now, we're planning a far more complex network by 1985. In fact, we'll be using about 600 minicomputers in all phases of our operation."

Why so many Digital computers? "Cost, architecture and overall reliability were the big factors in our decision." ■

---

the unit is. Since every service call results in a direct loss to the company, the obvious goal is to minimize service calls. According to Executive Vice President, Carl English, one of the best ways to handle that is to select the most reliable equipment.

"Before we chose a peripheral vendor, we put four competitive teleprinter terminals through a grueling four-month test. Then we had a team of technical experts evaluate the results and the equipment to determine which teleprinter would ultimately be the most reliable. Digital was the unanimous choice."

Since June, Leasco has bought hundreds of DECwriter Teleprinters from Digital. With hundreds more on order.

"Our customers want continuous use of their machines," English says, "and Digital gives it to them." ■

---

chases in the coming year.

"We depend on timely delivery to keep our production going. That's why we picked Digital. They've given us the true production line, high volume operation we need.

"What's more, they back their hardware and software with good service.

"When we're trying to sell a system to somebody for a remote area, the first question asked is, what if it breaks down. We just give him the telephone number and address of the local Digital field service office. That's a powerful selling point. And one that no other minicomputer company can match.

"After buying 150 Digital computers, I can't say enough about them. Hardware, software, support – they have everything." ■

---

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

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

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

My computer application is ______________________

Name ______________________ Title ______________________

Company ______________________ Address ______________________

City ______________ State ______________ Zip ______________

digital

50,000 computers saving managers millions.

450 SL

SCIENCE AND THE CITIZEN

Meter by Meter

The pace of conversion to the metric system in the U.S. can be expected to quicken under the impetus of the Metric Conversion Act of 1975, which was signed by President Ford in December. The act provides for the creation of a United States Metric Board of 17 members (to be nominated by the President and approved by the Senate), who will have the responsibility for coordinating a "voluntary conversion" by means of "a broad program of planning, coordination and public education." The legislation puts into effect, with modifications, the recommendation of the U.S. Metric Study (made between 1968 and 1971 under legislation passed by Congress in 1968) that the U.S. adopt the metric system "through a coordinated national program over a period of 10 years." The 1975 act provides for the coordinated national program but does not specify that the conversion should be completed within a certain period of time.

The system of measurement that the Metric Board will be fostering is formally known as the International System of Units (often abbreviated SI, for Système International d'Unités). The SI has seven base units (the meter for length, the kilogram for mass, the second for time, the ampere for electric current, the kelvin for temperature, the mole for the amount of a substance and the candela for luminous intensity) and two supplementary units (the radian for a plane angle and the steradian for a solid angle). Units for all other quantities are derived from these nine.

In deriving units the SI emphasizes that they should be formed from the base and supplementary units in a coherent manner. One result of this principle is a series of 16 prefixes that should be applied to the basic units to indicate large or small quantities. For large quantities they are deka (10^1), hecto (10^2), kilo (10^3), mega (10^6), giga (10^9), tera (10^{12}), peta (10^{15}) and exa (10^{18}). For small ones they are deci (10^{-1}), centi (10^{-2}), milli (10^{-3}), micro (10^{-6}), nano (10^{-9}), pico (10^{-12}), femto (10^{-15}) and atto (10^{-18}). Another result is that certain units widely regarded as being part of the metric system are not part of the SI. They include the liter (the SI's derived unit for volume is the cubic meter, and in strict terms one should order fuel and beer in fractions or multiples of that unit) and the metric ton (1,000 kilograms, properly expressed as one megagram). The National Bureau of Standards takes the position, however, that units such as the liter, and also the minute, the hour and the day, "are used so widely that it is impractical to abandon them."

Among the SI's derived units with special names are those for frequency (the hertz, or cycles per second), force (the newton, or meter-kilogram per second per second), pressure or stress (the pascal, or newton per square meter), energy, work or quantity of heat (the joule, or newton-meter), power or radiant flux (the watt, or joule per second), electric charge or quantity of electricity (the coulomb, or ampere-second), electric potential, potential difference or electromotive force (the volt, or watts per ampere), capacitance (the farad, or coulombs per volt), electric resistance (the ohm, or volts per ampere), conductance (the siemens, or amperes per volt), magnetic flux (the weber, or volt-second), density of magnetic flux (the tesla, or webers per square meter), inductance (the henry, or webers per ampere), luminous flux (the lumen, or candela-steradian), illuminance (the lux, or lumens per square meter), radioactivity (the becquerel, or spontaneous nuclear transitions per second) and absorbed dose of radiation (the gray, or joules per kilogram).

It seems likely that a number of units that are not part of the SI will continue in use no matter how extensively the metric system is adopted in the U.S. For some units the reason will be familiarity, and for others it will be the possibility that an SI unit would impede communication. In addition to the units that the Bureau of Standards says it would be impractical to abandon because they are so familiar, certain other units will be acceptable "for a limited time" because "their usage is already well established." They include the nautical mile, the knot, the barn (a unit employed in nuclear physics to express the effective cross section of an atomic nucleus) and the gal (a unit employed in geodesy and geophysics to express the acceleration due to gravity). According to the bureau, "it is believed by many persons familiar with metrication efforts on the part of business, industry and the nation's school systems that metric will be the predominant, although not exclusive, measurement system within 10 years."

Slowdown

In the view of the U.S. nuclear-power industry 1975 was "a year of consolidation—a corporate euphemism for rotten business." That assessment sums up both the content and the tone of the annual review of the U.S. nuclear-power program issued by the Atomic Industrial Forum, an official organ of the industry. "While other sectors of the economy began to shake off recessionary symptoms," the Forum observes, the nuclear program "barely held its own in the face of sluggish overall demand for electric energy and unabated inflation." Moreover, "the continuing struggle to forge an integrated, commercial fuel cycle... was complicated by a cautious Federal bureaucracy, a skeptical Congress and a clamorous band of critics."

The dimensions of what is described as "the statistical decimation of nuclear power in 1975" are summarized in the review. Up to 1974 the consumption of electric power in the U.S. had grown for some time at an average rate of 7 percent per year. In 1974 that annual growth rate was cut to zero. Nonetheless, economists for the electric-utility companies projected a renewed growth rate of 5 percent for 1975. The actual growth rate for the year turned out to be less than 2 percent. Although residential and business consumption of electricity increased substantially, industrial consumption fell below even the 1974 level.

As a result the utilities chose to continue the "recession-inspired wave" of cutbacks in the construction of nuclear and non-nuclear power plants, a trend that seemed to the nuclear industry "to have spent itself in early 1975." Approximately 180,000 megawatts of projected new electric-generating capacity had been deferred in 1974, some two-thirds of it nuclear. In 1975 the utilities deferred an additional 40,000 megawatts of capacity, this time half of it nuclear. Eight power-reactor projects were canceled altogether.

According to the Atomic Industrial Forum, "the disproportionate impact of utility retrenchment on nuclear power plants is not difficult to understand," since nuclear plants tend to be the largest and most costly additions to utility systems. It has been estimated that a 1,000-megawatt power reactor ordered in 1975 will cost $868 million ($868 per kilowatt) to build, exclusive of fuel, and will be ready for operation in 1984. The same plant, if it had been ordered in 1969, would have cost $226 million and would probably have been in operation today.

Turning to the problems of the nuclear fuel cycle, the Atomic Industrial Forum notes that at "the front end, it was suddenly a seller's market for uranium." Although this is expected to "rekindle" the pace of exploration in 1976, the result in terms of new uranium discoveries is "less easily discernible." In any case "the ultimate uranium supply may be determined as much by the tail end of the cycle as by the availability of the raw material." Here, the Atomic Industrial Forum observes, "there was great commotion but little forward movement in 1975." Plans to recycle plutonium in the spent reactor fuel, a stratagem that could reduce uranium requirements by as much as 40 percent, have been held up pending approval of the concept by the Nuclear Regulatory Commission. Meanwhile the question of whether or not private industry is to enter the uranium-enrichment business is the subject of Congressional debate. The Energy Research and Development Administration (ERDA) has recently announced that it must know by the end of March whether private uranium enrichment is to be expected or whether the Federal Government will need to expand its own enrichment capacity in order to prevent a possible shortage of reactor fuel by about 1984.

In spite of the generally "dismal" performance of the industry in 1975, the Atomic Industrial Forum states that "the year had its compensations," among them the fact that the year ended with 58 power reactors licensed to operate in the U.S., 69 under construction and 101 more "in the pipeline." Nuclear power plants had about 40,000 megawatts of generating capacity in 1975, accounting for roughly 8 percent of the total U.S. capacity. That percentage is still projected by ERDA to rise to more than 23 percent by 1985, 40 percent by 1995 and 45 percent by 2000.

Anatomy of an Antigen

The immune response and all other phenomena of immunochemistry depend on the highly specific manner in which an antibody "recognizes" and binds to its antigen. It has been clear for some time that what a combining site on an antibody molecule recognizes is the pattern of an antigenic determinant, or reactive region: a rather small patch on the outside of the protein, polysaccharide, nucleic acid or other large molecule that acts as an antigen. In the case of protein antigens those regions must be short stretches of amino acid units on the complexly folded chains of amino acids that constitute a protein molecule.

Now for the first time the complete antigenic structure of a protein has been worked out. In an immunological tour de force M. Z. Atassi has precisely mapped the five antigenic reactive regions on the molecule of myoglobin, the oxygen-transporting protein of muscle. The task, which took 11 years to accomplish at the State University of New York at Buffalo School of Medicine and at the Wayne State University School of Medicine, was described in more than 50 step-by-step reports by Atassi and a succession of co-workers and has been reviewed in the journal *Immunochemistry* by Atassi, who is now at the Mayo Medical School. "In retrospect," he writes, "protein chemistry was not... sufficiently developed" to do the job early in the 1960's; he and his colleagues developed new reactions and other techniques as they went along.

Atassi chose myoglobin because it was then the only protein whose three-dimensional structure was known; the sequence of most of its amino acids was also known. The 153 amino acids of myoglobin form a single chain that is folded into a compact molecule in which straight helical segments alternate with sharp bends, and within which is nested a single iron-containing heme group: the oxygen carrier. One could not simply take the chain apart and check each peptide, or segment of two or more amino acids, for antigenicity. That had been attempted with some proteins, but the resulting information revealed little about the active sites of a native, or folded and intact, protein molecule. Atassi relied instead on five distinct approaches, the results of each serving to correct or confirm the results of the others.

First he studied the effect of changing the conformation of the native protein by substituting variously modified heme groups for the normal one (which had been shown not to be in an active region) and testing the antigenicity of the resulting molecules in the presence of a number of different antibodies. (The antibodies had been prepared by injecting myoglobin into goats and rabbits.) The results showed that the overall shape of the molecule affects antigenicity, at least in the case of myoglobin. Then Atassi examined the effect of modifying specific amino acids in the molecule, one by one and in combinations. Some modifications affected the antigenicity of the molecule and others did not; the location of those that did (without causing any conformational change in the protein) provided the first evidence for the approximate locations of reactive regions.

The next stage was to dissect the molecule into overlapping peptide fragments, taking care (on the basis of the preceding step's results) not to cleave the chain within an indicated reactive region. An isolated, unfolded peptide does not always display antigenicity even if it comprises a reactive region, but under certain conditions it may. Positive results with two or more overlapping segments, for example, indicated that the reactive region was within the overlapping one. Then the immunochemically active peptides were modified at individual amino acids; again if a modification affected reactivity, the modified amino acid must be within a reactive region.

By combining the results of these four procedures Atassi was able to define five segments of the chain within which there appeared to be reactive regions. Some of those segments could not, for technical reasons, be precisely dissected out of the myoglobin molecule for final testing and narrowing down. The final step was therefore to prepare synthetic peptides that mimicked amino acid sequences within the suspected segments and test each of those peptides for antigenicity. And so the reactive regions were eventually mapped: five sequences, each of them six or seven amino acids in length, situated either at or near bends in the chain (but not at all bends). Essentially the same regions (with a shift of no more than one amino acid unit in a few cases) were recognized by antibodies from several different goats and rabbits.

Energy Plantations

Only 100 years ago wood supplied more than half of the energy used in the U.S. The material is now being reconsidered as perhaps the cheapest way to capture solar energy. The Energy Research and Development Administration (ERDA) has recently written several large contracts for the study of biomass energy systems, which would have trees, grasses or other crops grown on "energy plantations" as a source of either electric power or synthetic fuels. Plausible estimates suggest that B.t.u.'s from plant cellulose would be competitive with low-sulfur coal in parts of the U.S. and substantially cheaper than oil everywhere.

For example, the Green Mountain Power Company of Burlington, Vt., which now pays more than $50 per ton for low-sulfur coal, is seriously considering fueling its next generation of power stations with wood chips, provided by systematic pruning and clearing of existing forests.

Proponents of energy plantations are confident that unless there are unforeseen reductions in the cost of solar-collection systems, which depend on solid-state solar cells, thermal panels or sun-tracking mirrors, the natural process of photosynthesis offers by far the simplest and cheapest way of trapping solar energy on a large scale. A major advantage of energy plantations over artificial systems is that the solar energy is captured and stored simultaneously. It is estimated that a movable-mirror-and-boiler collection system, perhaps the cheapest of the proposed technical methods, would require an investment of roughly $2,500 per installed kilowatt. The capital investment in equipment and starting material for an energy plantation, according to one estimate, should be no more than $150 per kilowatt, assuming that the land was rented at a rate that would yield the owner a return of 10 percent. To this must be added perhaps another $750 per kilowatt for the power plant itself. (A nuclear power plant would cost from 15 to 20 percent more.)

Two general kinds of crops are considered suitable for energy plantations: warm-season grasses (such as Bermuda grass and Sudan grass) and certain deciduous trees (such as poplar, eucalyptus, alder, cottonwood and sycamore). The ideal crop would be a perennial that reproduces from cuttings (since seed requires costly collection and sowing), that grows rapidly when young and that resprouts vigorously after harvesting. Warm-season grasses, including sugarcane and sorghum, may be preferable to trees in the Gulf Coast states, where rainfall is plentiful. Farther north trees may provide higher yields.

Grasses would be harvested every few weeks. Trees would be planted several inches apart (up to 10,000 per acre) and would be harvested every three or four years on a continuous rotating schedule. The stumps should sprout many times before replanting is necessary.

Yields considered realistic for either warm-season grasses or trees are seven to 11 tons of dry matter per acre per year, which corresponds to the capture and net conversion of .6 to .8 percent of the incident sunlight. (More optimistic estimates project yields two or three times larger.) The typical conversion efficiency for wild forests is .1 to .2 percent and for trees grown on pulpwood farms about .3 percent. Each acre of an energy plantation should supply enough fuel for two kilowatts of installed generating capacity. Thus a 500-megawatt power plant would consume the output from 250,000 acres, or about 400 square miles. It is estimated that there are between 70 and 100 million acres in the U.S., now little used and unsuitable for food or fiber production, that could be converted into energy plantations. The potential biomass harvest from 100 million acres would provide enough fuel for 200,000 megawatts of electric-generating capacity, equivalent to about 40 percent of the nation's present installed capacity. Alternatively, depending on the economics, the biomass from energy plantations could be converted into synthetic gaseous and liquid fuels.

Clocks

The first three-quarters of this century have witnessed a revolution in prehistory: the progressive substitution of absolute chronologies for relative ones. Early in that period the archaeologist's only precise chronological technique was one borrowed from the earth sciences: relative stratigraphic position. Admirably accurate at any one stratified site, particularly if the uppermost strata included remains from historical periods, relative stratigraphy was almost useless when it came to comparing data from widely separated places. Soon after World War I a powerful new dating technique—tree-ring analysis—enabled archaeologists in the southwestern U.S. to construct the first extensive regional absolute chronology, reaching back from the present to the centuries just before European contact with the New World. Soon after World War II the introduction of carbon-14 dating gave archaeologists a tool for the establishment of absolute chronologies over a range of thousands of years around the world. Since then new absolute methods have appeared, not least among them the combining of tree-ring and carbon-14 analyses to detect irregularities in the carbon-14 clock and to recalibrate it. The journal *World Archaeology* recently invited contributors to summarize developments in absolute dating techniques. The highlights are as follows.

Tree-ring chronologies in the Old World now include a sequence based on samples of juniper wood at Gordion in Turkey covering the years from 1500 to 700 B.C., a European sequence based on subfossil oaks from streambed gravel deposits that may extend to about 7000 B.C. and shorter sequences in England and Ireland that overlap with the Roman, Viking and medieval periods. Tree-ring chronologies in the U.S. now include sites in Alaska and the Mississippi valley; the longest southwestern sequence now goes back to 332 B.C. Of major importance is the chronology based on living and dead bristlecone pines from the White Mountains of California. It provides an unbroken record going back to 6000 B.C. and may soon be extended by another 2,000 years. It is the bristlecone chronology that made possible recalibration of the absolute dates derived from carbon-14 analysis; these proved to be as much as 1,000 years in error for the period from 4000 to 5000 B.C.

Carbon-14 dates based on wood and charcoal samples begin to accumulate excessive probability errors when the samples are more than 35,000 years old. A new carbon-14 technique, based on samples of bone collagen, retains tolerable margins of error up to 50,000 years. Sample enrichment may be able to extend the bone-collagen range back to 75,000 years.

Another new absolute dating technique is based on the fact that after an animal has died some of the amino acids in its bones gradually shift from their natural "left-handed" molecular structure into a "right-handed" structure that is unknown in living organisms. Age determinations based on this amino acid racemization have proved to correlate well with carbon-14 findings from samples of the same specimens. Future prospects for the technique seem promising on two grounds. First, the bone sample required is much smaller than that needed for bone-collagen extraction. Second, the probability errors remain within acceptable limits until the specimens are more than 100,000 years old.

An established technique that determines the age of mineral specimens on the basis of their proportional content of potassium and argon has been greatly refined in recent years. The method is based on the fact that a natural isotope of potassium decays radioactively into argon. Thus over a period of time the proportion of argon to potassium increases. Absolute dates can be determined in a range from less than 50,000 to more than 50 million years.

Absolute dating by means of fission-track analysis has come to be recognized as perhaps the most versatile of the new methods. In crystals and even in amorphous materials such as glass, obsidian and amber microscopic tracks are produced by the spontaneous fission of trace amounts of uranium. The abundance of the tracks, which are revealed by etching, is proportional to the time elapsed since the material was first formed or, in the case of certain artifacts, was last exposed to high temperatures.

The method has now been used to date materials ranging from glass only a century old to 2,000-year-old pottery from Japan, zircon inclusions in volcanic ash from Valsequilla in Mexico (where certain ambiguous human artifacts have been unearthed) that are as much as 500,000 years old, and pumice from Olduvai Gorge in Tanzania as much as 2.3 million years old. A potassium/argon finding on samples from the same pumice formation, indicating that its age was 1.75 (plus or minus .05) million years old, is in good agreement with the fission-track finding.

Catastrophe Theory

Any object or concept can be represented as a form, a topological surface, and consequently any process can be regarded as a transition from one form to another. If the transition is smooth and continuous, there are well-established mathematical methods for describing it. In nature, however, the evolution of forms is rarely smooth; it usually involves abrupt changes and perplexing divergences.

Such discontinuous and divergent phenomena have long resisted mathematical analysis, and in most cases quantitative description is still unattainable. In the past few years, however, a method for the construction of qualitative, topological models has been developed. The method was devised by René Thom of the Institut des Hautes Études Scientifiques at Bures-sur-Yvette in France, and numerous applications of it have been proposed by other investigators. Thom's treatise, *Structural Stability and Morphogenesis,* has recently been published in an English translation by D. H. Fowler.

Thom proved that in a space having no more than four dimensions (such as the four-dimensional space-time of the real world) there are just seven kinds of transformation. Because these transformations represent sudden disruptions of otherwise continuous processes he termed them the seven elementary catastrophes. They are represented by graphs that have folds and pleats and other more complicated surfaces; discontinuities are represented in the topological models by sudden jumps from one surface to another.

One of the early applications of the theory, which Thom discussed extensively, was to the growth and differentiation of the embryo. Since all the cells in an embryo contain the same genetic information, it is not obvious how the cells differentiate into ectoderm, endoderm and mesoderm, or how some of them later specialize to form particular organs. Without proposing explicit biochemical mechanisms, Thom suggested that the process of differentiation is a series of catastrophes. In principle the model could describe each point of bifurcation, where the development of a cell diverges from that of its immediate neighbors.

E. C. Zeeman of the University of Warwick in England has employed the theory to describe a number of other phenomena, including some that are quite commonplace. The course of an argument between two people, for example, often involves an escalation of emotion, making rational discussion impossible. In the model this divergence becomes a catastrophe: the most probable behavior is either aggression or submission, and sudden, catastrophic shifts from one attitude to the other are not uncommon. A similar model extends the principle to international politics: a threatened nation may offer concessions or may declare war. Both models explicitly account for the observation that a man or a nation "pushed too far" may suddenly lash out. The adage about the straw that broke the camel's back might also be cited; indeed, the sudden buckling of a flexible beam, whether it is the spine of an animal or the girder supporting a bridge, is readily explained as one of the elementary catastrophes.

Catastrophe theory has been particularly interesting in its applications to the biological and social sciences, perhaps because discontinuous and divergent phenomena abound there. For example, Thom suggests applications not only in embryology but also in the theory of evolution, in reproduction, in the process of thought and in the generation of speech. For the living cell and for the organism as a whole life is one catastrophe after another.

Booming Dunes

A sand dune would not seem to be a very likely candidate as a natural sound generator. The fact is that dunes in many parts of the world squeak, roar or boom. Acoustic sands have been described in desert legend for at least 1,500 years, but they have received little scientific attention. Recently, however, David R. Criswell of the Lunar Science Institute in Houston, Tex., John F. Lindsay of the University of Texas at Galveston and David L. Reasoner of the Marshall Space Flight Center in Huntsville, Ala., conducted the first quantitative analysis of the properties of an acoustic dune.

The dune they investigated was Sand Mountain near Fallon, Nev., which has often been heard to boom. In order to obtain simultaneous recordings of both seismic and acoustic emissions from the sand, the experimenters and two assistants set up a geophone for monitoring vibrations transmitted through the sand and a microphone for receiving sounds transmitted through the air. After trying several different methods they found that the sand boomed loudest when a trench was rapidly dug in it with a flat-bladed shovel. The sound was like a short, low note on a cello; it lasted for less than two seconds and was readily audible at a distance of 30 meters. The booming could also be produced by pulling the sand downhill with the hand; in that case strong vibrations reminiscent of a mild electric shock could be felt in the fingertips.

When the sand was disturbed, the sound was detected by the air microphone about a quarter of a second before seismic vibrations were detected by the geophone. When the disturbance was stopped, both types of vibration ceased simultaneously. The experimenters interpret this to mean that the disturbance induced multiple modes of vibration in the ground, the earliest of which was not detected by the geophone.

Although the appearance of booming sand dunes is indistinguishable from that of ordinary silent dunes, examination of sands in the scanning electron microscope revealed that the individual grains of booming sand are more highly polished than the grains of silent sand. Twenty-nine of the 31 known booming dunes are composed primarily of quartz sand; the exceptions are two dunes in Hawaii that are principally calcite sand.

The tonal quality of the booming is remarkably pure, comparable to an organ tone. "One must be amazed that as unlikely a medium as sand can produce such pure oscillations," the experimenters comment in *Journal of Geophysical Research*. "A quantitative theory of the booming process is not presently available. . . . The problem is intriguing and very likely far more complicated physically and mathematically than the simple processes controlling oscillations in an organ pipe."

Will the Universe Expand Forever?

The recession of distant galaxies, the average density of matter, the age of the chemical elements and the abundance of deuterium together suggest that the expansion cannot be halted or reversed

by J. Richard Gott III, James E. Gunn, David N. Schramm and Beatrice M. Tinsley

Cosmological inquiry is ancient, but only in the past 50 years or so have we begun to understand how the universe began and what its ultimate fate may be. The crucial perception came in the 1920's, when Edwin P. Hubble demonstrated that the spiral nebulas are not local objects but independent systems of stars much like our own, and thereby showed that the universe is a much larger place than had been imagined. Hubble showed further that the entire observable system of galaxies is in orderly motion. As is now well known, the nature of that motion is expansion: all distant galaxies are receding from us.

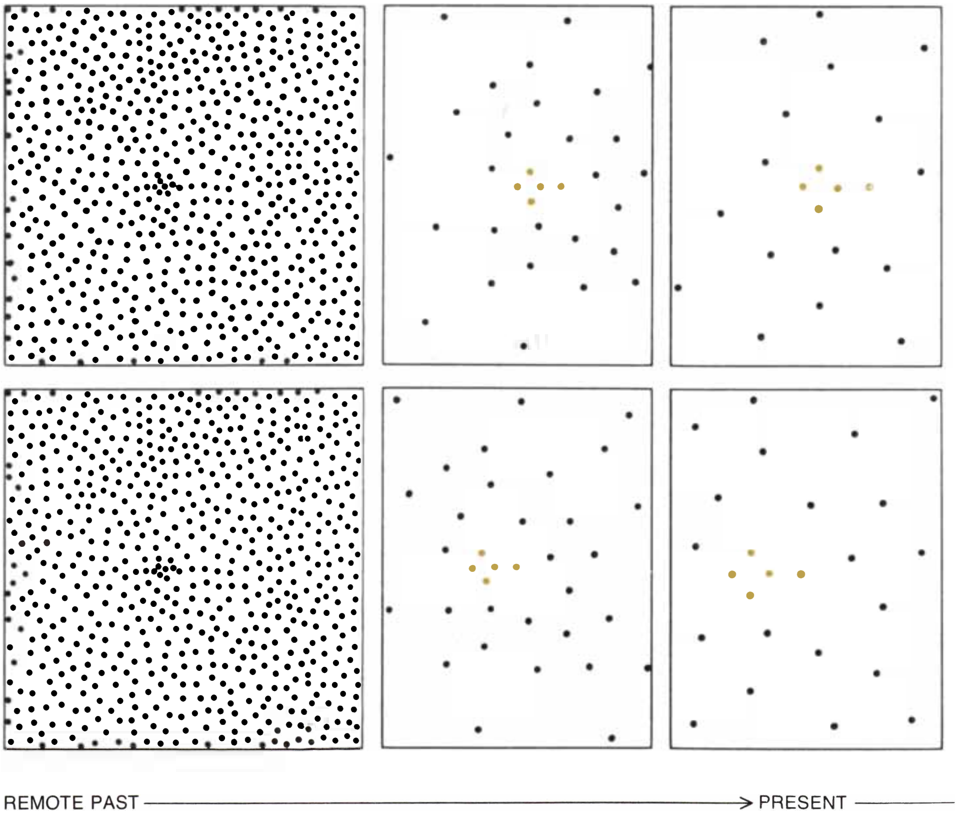
That the universe is expanding is today considered established. A question that remains unsettled is whether the expansion will continue forever or whether the receding galaxies will someday stop and then reverse their motion, eventually falling together in a great collapse. The answer to this question determines the geometrical character of the universe, that is, it determines the nature of space and time. If the expansion continues perpetually, the universe is "open" and infinite; if it will someday stop and reverse direction, the universe is "closed" and of finite extent.

In order to choose between those possibilities, astronomers construct mathematical models of the universe and then attempt to find observable features of the real universe that would confirm one of the models and exclude all others. So far no single measurement has been made with enough precision to settle the question unambiguously. Several independent tests are possible, however, and pieces of the puzzle have been supplied by many workers employing quite different techniques. It now seems feasible to assemble the pieces. Taken together, the available evidence suggests that the universe is open and that its expansion will never cease.

Isotropic Expansion

Hubble detected the recessional motion of the distant galaxies through measurements of their optical spectra. The spectra of most stars (and hence of galaxies) are interrupted by dark lines representing the absorption of particular wavelengths by atoms in the cooler, outer layers of the stellar atmosphere; each chemical element generates a characteristic pattern of lines whose wavelengths are precisely known from laboratory measurements. When the galaxy is moving away from the observer, the wavelength of each spectral line is increased as a result of the Doppler effect, so that all the lines appear to be displaced toward longer wavelengths and in particular toward the red end of the visible portion of the spectrum. The displacement is called a red shift, and by measuring its magnitude the velocity of recession can be calculated. When an object is moving toward the observer, the wavelengths of the spectral lines are decreased by the Doppler effect and the lines appear to be displaced toward the blue end of the spectrum, an effect called a blue shift. All the distant galaxies whose spectra were measured by Hubble and by later observers show red shifts; they are therefore assumed to be receding from us.

The recessional motion has several remarkable properties. Hubble showed that the velocity with which a galaxy recedes is proportional to its distance from us, so that a constant ratio of distance to velocity can be calculated. The ratio is such that a galaxy 10 million light-years from us recedes with a velocity of 170 kilometers per second; another galaxy twice as far away recedes twice



FATE OF THE UNIVERSE is described through mathematical models of its behavior. Two classes of models are generally considered plausible; in both the universe begins in a compact state of infinite density (the big bang). In one class of models the universe expands continuously and indefinitely, albeit at an ever lower rate (*upper series of drawings*). In the other class as fast, or 340 kilometers per second [*see illustrations on next two pages*]. Small departures from this rule are commonly observed because most galaxies are members of groups or clusters and have orbital motions with a component along the line of sight connecting the earth with the galaxy. Those motions are essentially random, however, so that in any large sample of galaxies they cancel one another. Nonrandom, systematic variations from the ratio have been found only for galaxies at the most extreme distances; as we shall see, these variations do not invalidate the rule but provide important information about the history of the universe.

A second characteristic of the cosmic expansion is its isotropy: it is the same in all directions. No matter where in the sky a galaxy is found, its recessional velocity is related to its distance by the same proportionality. This observation seems to suggest that the universe is remarkably symmetrical and, what is even more extraordinary, that we happen to be at its very center. The crystal spheres of medieval cosmologies were no more geocentric.

There is, of course, another explanation, which can be understood most readily by considering a simple two-dimensional model of an expanding universe. Imagine a spherical balloon with small dots painted on its surface, each dot representing a galaxy. As the balloon is inflated the distance between any two dots (always measured on the surface of the sphere) increases with a speed proportional to the distance between them. No matter which dot is designated the center, all the other dots recede from it uniformly in all directions. Thus each dot observes the same expansion and no one of them has a privileged position. Such an expansion has no center; more precisely, every point is its center.

It follows from this analysis of the expansion that the geometrical configuration of the dots cannot change. A balloon bearing a picture of Mickey Mouse continues to bear the same picture as it is inflated. All distances between points on the balloon are multiplied by the same factor. Similarly, in the real universe eight galaxies that happen to lie at the corners of a cube in one epoch will remain at the corners of a cube, albeit a larger one, as the universe expands.

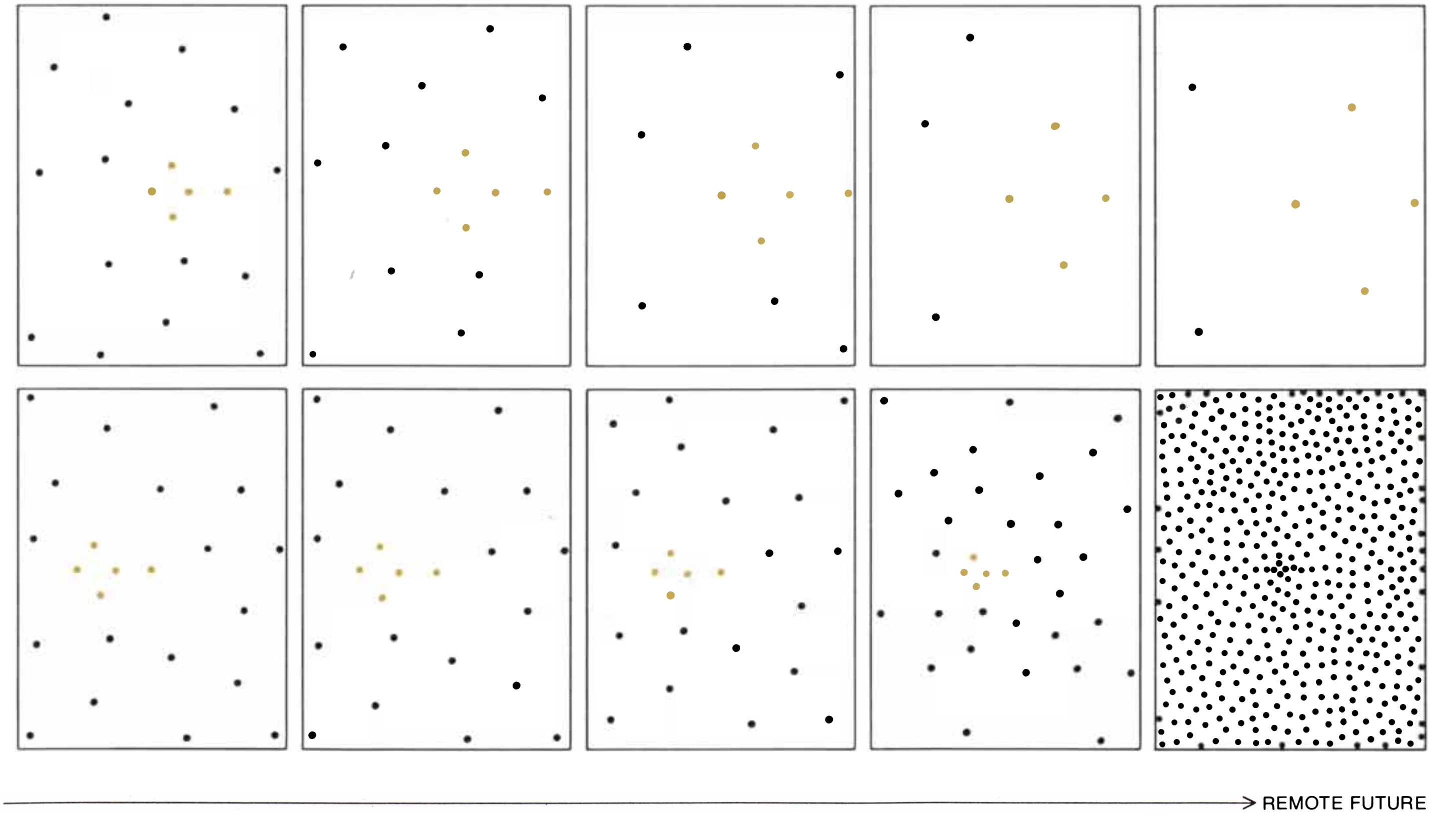
The Big Bang

Since Hubble's original discovery increasingly precise observations have shown that it is not only the cosmic expansion that is isotropic; all the large-scale features of the universe are indifferent to direction. For example, the distribution of galaxies on the celestial sphere and the distribution of extragalactic radio sources appear to be quite uniform. The most compelling evidence of isotropy was discovered in 1965 by Arno A. Penzias and Robert W. Wilson of the Bell Laboratories; it is the microwave background radiation that seems to bathe the entire universe. The microwave radiation has since been shown to be highly isotropic; it varies by less than one part in 1,000 over the entire area of the sky.

The observation of such remarkable isotropy has led to the adoption of a powerful generalization called the cosmological principle. It states that the universe appears isotropic around all observers participating in the expansion everywhere and at all times. In other words, our galaxy is indeed at the center of the universe, but so is every other galaxy.

The cosmological principle also governs the behavior of the two-dimensional model universe represented by a spherical balloon. If the painted dots are distributed with uniform density over the surface of the balloon, the neighborhood of any chosen point is statistically the same as that of any other point and no direction has any special significance. Indeed, it is not necessary to postulate independently that the dots (or, in the three-dimensional universe, the galaxies) are uniformly distributed. If the universe is isotropic for all observers, then the distribution must be homogeneous; if it were not, an observer at the edge of a density fluctuation would not see a uniform distribution independent of direction.

For the purposes of this discussion we shall adopt the cosmological principle, but it must be remembered that its appeal is mostly philosophical. It has not been ade-

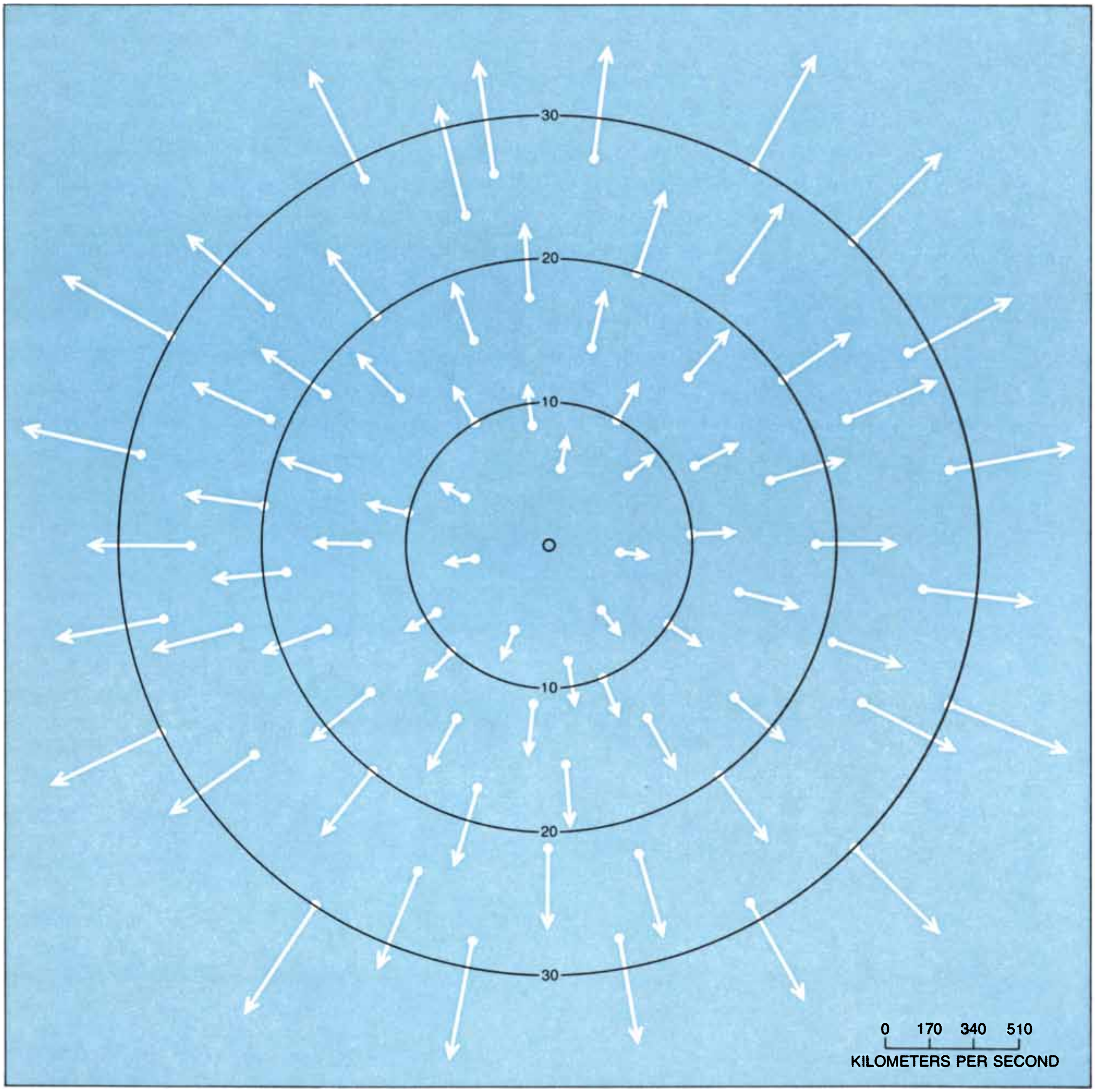


the universe expands to a maximum size, then begins to contract, eventually reaching infinite density again (*lower series of drawings*). The alternatives are illustrated here in an arbitrary region of space in which the cosmic expansion is represented by a decrease in average density. The expansion is isotropic, that is, the same in all directions. As a result an observer at any point perceives himself as being at the center of the expansion, and the shape of any pattern (such as the arbitrary pattern of colored dots) will be preserved in all epochs.

quately tested, and indeed adequate tests may never be possible.

Given our knowledge of the universe as we observe it today, what can we deduce about its history? A simple, hypothetical model with which to begin is one where the recessional velocity of every galaxy has remained unchanged through all time. It is then apparent that any galaxy now receding from us was once arbitrarily close, and that the time that has elapsed since then is equal to the ratio between the galaxy's distance and its velocity. Since the ratio is the same for all galaxies, all of them must have been nearby at the same time; in other words, at some unique time in the past all the matter in the universe was compressed to an arbitrarily great density. The time calculated to have elapsed since that compact state existed, assuming the rate of expansion has not changed, is called the Hubble time. Its reciprocal, by which one multiplies the distance of a galaxy to obtain its recessional velocity, is called the Hubble constant. Measurements of the Hubble time are complicated by uncertainties as to the distance to galaxies, and the measurements have been repeatedly revised since Hubble's first estimate of about two billion years. The Hubble time is now thought to be between 12 and 25 billion years, and the most likely value is about 19 billion years.

If the motions of the galaxies are extrapolated into the past as far as possible, a state is eventually reached in which all the galaxies were crushed together at infinite density. That state represents the big bang, and it marks the origin of the universe and everything in it. By the simple mathematics of



COSMIC EXPANSION seems to place the observer at the center of the universe, from which all distant galaxies are fleeing. The velocity with which a galaxy recedes is proportional to its distance from the observer, a relation first established in the 1920's by Edwin P. Hubble from observations made with the 100-inch telescope at the Mount Wilson Observatory. The principle that the ratio of velocity to distance is constant has since become known as Hubble's law. It can be interpreted most simply as evidence that the expansion of the universe began with the big bang, since the relation implies that in the past all the galaxies were packed together with infinite density. Distances are given here in millions of light-years; velocities are represented by the lengths of arrows, measured on the scale at lower right.

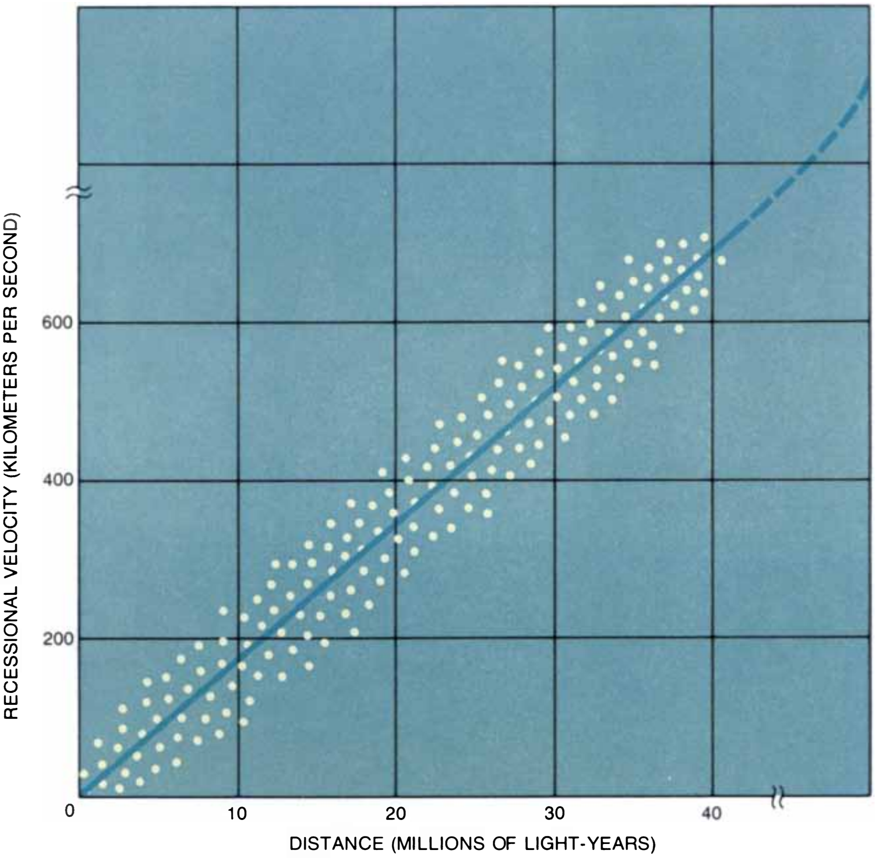
proportionality, if the recessional velocities have not changed, the date of the big bang must be exactly one Hubble time ago. Actually the rate of expansion has almost certainly not been constant, but that does not alter the fact of the big bang; it merely changes the date.

That the universe began with a big bang is an inevitable conclusion if the known laws of physics are assumed to be correct and in some sense complete. It is conceivable, however, that there are laws of nature whose effects are negligible on the scale of the physics laboratory, or even on the scale of the solar system, but that might predominate in determining the behavior of the universe as a whole. One theory requiring new laws of nature was the steady-state cosmology, in which the universe is unchanging and infinitely old. In order to explain the cosmic expansion the steady-state theory postulated the continual creation of matter from the void.

In the steady-state model of the universe it was particularly difficult to account for the microwave background radiation. This radiation field has the spectral characteristics of the thermal radiation emitted by a body at a temperature of 2.7 degrees Kelvin. It seems to be satisfactorily explained only as a relic of an epoch when the universe was very hot and very dense. A steady-state universe cannot have had such conditions, since in that model all conditions, by definition, have not changed.

In big-bang models the background radiation is a natural consequence of conditions in the early universe. The initial state in these models is one of high temperature and density, a state sometimes called the cosmic fireball. At this stage the matter and the electromagnetic energy composing the universe are thought to have been in thermodynamic equilibrium, and as a result the radiation spectrum was that of a very hot body. As the universe has expanded the radiation has cooled, reaching the low-temperature spectrum observed today. The cooling can be understood as an enormous red shift; since all galaxies are constantly receding from the radiation, its spectrum is constantly displaced toward the longer wavelengths associated with lower energies and lower temperatures. In 1946 George Gamow predicted the existence of a thermal background radiation entirely from the theoretical framework of the big-bang model. He estimated its present temperature as being about five degrees K. The general agreement between Gamow's prediction and the observations of Penzias and Wilson is the most compelling evidence for the big bang.

Thus it appears that the universe began from a state of infinite density about one Hubble time ago. Space and time were created in that event and so was all the matter in the universe. It is not meaningful to ask what happened before the big bang; it is somewhat like asking what is north of the North Pole. Similarly, it is not sensible to ask where the big bang took place. The point-universe was not an object isolated in space; it was the entire universe, and so the only answer can be that the big bang happened everywhere.



HUBBLE'S LAW is established by measuring the ratio of velocity to distance for many galaxies. The best estimate of the ratio (*solid colored line*) is about 17 kilometers per second per million light-years. Individual galaxies (*white dots*) depart from that value because most are members of clusters and have orbital velocities. The inverse of the ratio is the Hubble time: the time it would have taken for any given galaxy at its present velocity to reach its present position, or in other words the time since the big bang if the velocities have not changed. Actually it is thought that the recessional velocities have declined under the influence of gravitation; as a result the ratio is thought to increase at extreme distances (*broken colored line*).

In most models of the evolving universe the receding galaxies are assumed to follow ballistic trajectories, roughly analogous to those of a thrown ball or an artillery shell. The galaxies were propelled apart by forces acting at the moment of the big bang, but since then they have moved in free flight, without further propulsion. They should therefore continue in uniform motion if no other forces acted on them. Actually the galaxies continue to interact as they fly apart. If only the familiar forces that express the known laws of physics are to be allowed in our models, then only one force can have a significant effect on the expansion: gravitation. We can therefore hope to understand the dynamics of an expanding universe if we can describe the gravitational interactions of all its components.

Gravitational Deceleration

The gravitational force affects all matter, it is always attractive and its range is infinite. Moreover, gravitation has a peculiar geometrical property that significantly aids analysis: A hollow sphere exerts no net gravitational forces on masses in its interior. (Actually, of course, the mass of the hollow shell attracts the masses in the interior, but all the forces exactly cancel, so that at every point in the interior the resultant force is zero.) This proposition was first proved by Newton, but it applies equally well to more recent theories of gravitation, such as the general theory of relativity.

If a spherical region of the universe is selected for examination, the rest of the universe surrounding it can be regarded as a hollow spherical shell, since the cosmological principle requires that the surrounding matter be uniformly distributed in all directions. The selected sphere can then be studied as if it were isolated and not subject to forces from the outside. The cosmological principle also ensures that any selected sphere of galaxies will expand or contract by the same factor as the universe as a whole, regardless of the sphere's location or size. In order to characterize the dynamics of the universe, therefore, we need only examine the dynamics of a representative sphere within it. If the sphere chosen is a small one, the velocities of the galaxies will be much smaller than the speed of light, and their motions can be described in terms of Newtonian mechanics.

A galaxy at the edge of such a small

Fallingwater at Mill Run, Pennsylvania. Designed in 1936 by Frank Lloyd Wright. Photographic sequence ranging from 10.4 inches (upper left) to one-te *right) records varied aspects of the house from its smallest detail to its overall integration with the building site. 'ographs above are reproduced actual size.*

ONE CAMERA: 10.4 INCHES TO INFINITY.

Slip a deluxe SX-70 Land camera in one jacket pocket, and some film packs in another, and you're ready to document your field trip.

Your field may be architecture, archeology, or criminology. It doesn't matter. What matters is the SX-70's portability and versatility.

Consider these unique virtues:

With a touch of the finger and a glance through the split-image rangefinder, you can focus to get a one-half life-size image of a subject 10.4 inches from your lens. Or a panoramic view at infinity. Or anything in between. (You can even get a *life*-size image, if you need it, by attaching our optional close-up lens.)

When you're working in the field, you have to concentrate on your surroundings. So it's a blessing to have a camera that doesn't compete for your attention. With the SX-70 you just focus on your subject, frame and press the electronic button. Period.

There's no worry that part of the subject will be cut out of your print. (Our single-lens reflex viewing system delivers what you see in the viewfinder.)

No preoccupation with your light conditions. (Our automatic electronic exposure control system takes care of that.)

No need to time the development of your prints as they emerge from the camera into broad daylight. And nothing to peel or throw away. (Our film takes care of all that.)

With the SX-70, you can take a sequence of up to ten photographs at intervals as short as 1.5 seconds.

The color values of those photographs are incredibly true-to-life. And our revolutionary metallized dyes will keep those colors from fading for an incredibly long time.

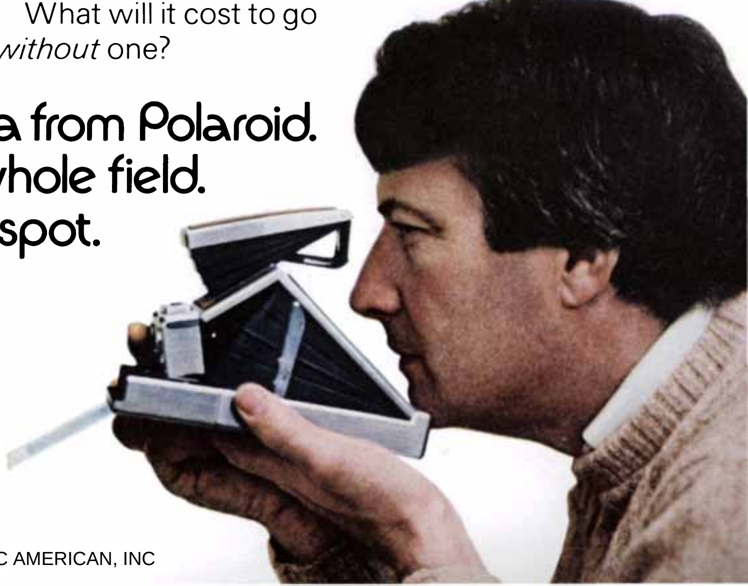
Furthermore, since SX-70 prints are virtually grainless, you could photograph a saguaro cactus (for example) at infinity, and identify its resident owl under a magnifying lens back at your desk.

But before you leave to return to your desk, you'll know that each of your photographs contains exactly what it's supposed to. Because the print develops itself completely in minutes. So, if the camera moved, or you want to change the angle of a particular shot, you can reshoot then and there.

What will it cost to have an SX-70 on your next trip? Wrong question.

What will it cost to go *without* one?

**The SX-70 camera from Polaroid.
Covers the whole field.
On the spot.**



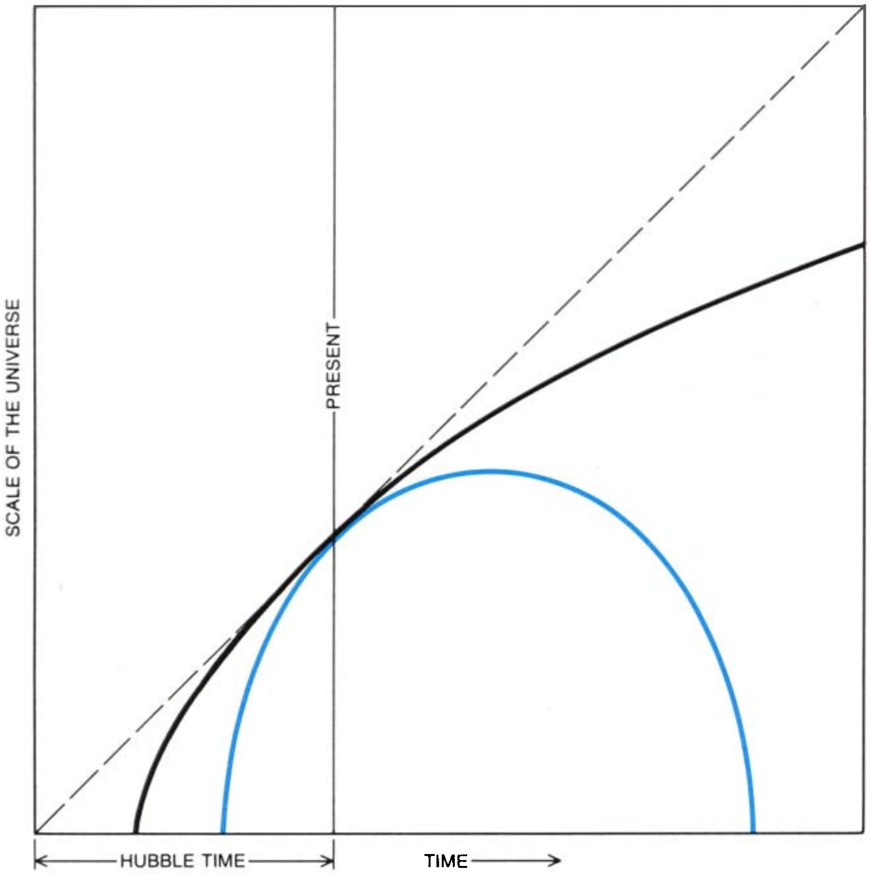
spherical region feels only the gravitational forces generated by the matter inside the sphere. If that matter is distributed homogeneously, then the resultant force acting on the galaxy attracts it to the center of the sphere. As a consequence the test galaxy does not recede with constant velocity; instead its recessional motion is at all times decelerated. It is thus obvious that in the past the test galaxy and all other galaxies must have been moving faster than they are today. Ignoring the deceleration leads to an overestimate of the age of the universe. That age is one Hubble time only if the rate of expansion has not changed; since the rate has slowed under the influence of gravitation, the big bang must have taken place more recently than one Hubble time ago.

The magnitude of the gravitational deceleration clearly depends on the amount of mass inside the selected sphere. If the sphere contains a great deal of matter, the test galaxy must eventually stop and fall toward the center; the representative spherical region begins to collapse and, on the cosmological principle, so does the entire universe. If there is little matter, the test galaxy will decelerate continuously but never stop. Both the spherical region and the universe as a whole will expand indefinitely. The situation is analogous to that of a projectile shot upward from the surface of the earth: the projectile decelerates but nevertheless will not return to the surface if its velocity exceeds a certain critical value, the escape velocity.

The escape velocity for objects leaving the earth is determined by the mass of the earth; for a test galaxy at the edge of an arbitrary sphere in space the escape velocity is determined by the total mass within the sphere. From the ratio of velocity to distance the actual velocity of the test galaxy with respect to the center of the sphere is known. Its ultimate fate therefore depends on the value of the escape velocity and hence on the mass within the sphere.

Since the universe is assumed to be homogeneous, the determining quantity is the average density of matter in the universe. If the density is smaller than some critical value, the effect of gravitation is too small to halt the cosmic expansion, and all galaxies will recede forever (although ever more slowly). If the density is greater than the critical density, gravitation will prevail, and the expansion will slow to a stop and begin an accelerating contraction ending in a final cataclysm: what might be called the big crunch. The actual value of the critical density depends on the Hubble time, which is not precisely known. If the Hubble time is 19 billion years, the critical density is 5×10^{-30} gram per cubic centimeter, the equivalent of about three hydrogen atoms per cubic meter. That seems to be an exceedingly small density, but it should be remembered that on the average the universe is quite empty.

The effect of gravitation on the cosmic expansion can be incorporated into mathematical models most conveniently by introducing a dimensionless number called the density parameter and denoted by the Greek letter omega (Ω). The density parameter is defined as the ratio of the actual density of the universe to the critical density. If the universe is to expand forever, that ratio must be less than or equal to 1; if Ω equals exactly 1, the universe is expanding everywhere at just the escape velocity, and if Ω is greater than 1, the universe must eventually collapse.



MODELS OF COSMIC EVOLUTION describe changes in the scale of the universe with the passage of time. All models must be consistent with the scale and the rate of expansion observed today, so that all their graphs must be tangent at the present moment. If the rate of expansion is unchanging (*broken black line*), the age of the universe is the Hubble time. Decelerating universes are younger, and both their history and their destiny depend on the magnitude of the deceleration. With modest deceleration the expansion can continue indefinitely, albeit at an ever lower rate (*solid black line*). Greater deceleration implies that the cosmic expansion must stop and then reverse, leading to an eventual collapse (*colored line*). The infinitely expanding universe is said to be "open" and the collapsing universe, which is also the youngest, "closed."

The Geometry of Space

The foregoing discussion could have been derived entirely from the Newtonian theory of gravitation, although it is also valid in the general theory of relativity. In the general theory, however, the value of the density parameter has further consequences; in particular it determines the geometry of space. In the high-density universe fated to collapse, gravitation is sufficiently strong to "close" space. The total volume of the universe is finite at all times, although there is nevertheless no boundary or edge to the universe. A two-dimensional analogue of such a three-dimensional space is the surface of a sphere, which similarly is finite in area although it has no boundary.

If Ω equals 1, so that the universe expands with exactly the escape velocity, the geometry of space is "flat"; it is the familiar Euclidean geometry, and it is represented in two dimensions by an infinite plane.

The geometry of a perpetually expanding universe in which Ω is less than 1 is more difficult to illustrate. The two-dimensional analogue is the surface of a figure called a pseudosphere, and a complete pseudosphere cannot be constructed in three-dimensional space. A saddle-shaped surface has some of the properties of such a space, but it is a defective model in the important respect that it has a center, whereas the real space defines no preferred position [*see illustration on page 70*]. Perhaps the best two-dimensional representation of such a space is a projection of a pseudosphere onto a plane, a device that is employed in several of the works of the Dutch artist M. C. Escher [*see illustration on page 71*].

The three possible kinds of three-dimensional space are distinguished by several ge-

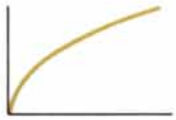
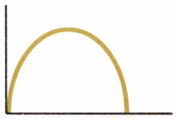
We asked Control Data

to integrate our branches and data processing operations into the most advanced bank network in the U.K.*

They did it.

*Details on request. Call collect (612) 853-7600 for **Barclays Bank** CDC Communications System Summary. Or write Control Data Corporation, Box 1980, Twin Cities Airport, St. Paul, MN 55111

CONTROL DATA CORPORATION

	OPEN	CRITICAL	CLOSED
DENSITY PARAMETER Ω $\frac{\text{ACTUAL DENSITY}}{\text{CRITICAL DENSITY}}$	$\Omega < 1$	$\Omega = 1$	$\Omega > 1$
DECELERATION PARAMETER q_0 DECELERATION $\frac{\text{DISTANCE}}{(\text{VELOCITY})^2}$	$q_0 < \frac{1}{2}$	$q_0 = \frac{1}{2}$	$q_0 > \frac{1}{2}$
GEOMETRY OF SPACE	HYPERBOLIC (NEGATIVE CURVATURE)	FLAT (ZERO CURVATURE)	SPHERICAL (POSITIVE CURVATURE)
FUTURE OF THE UNIVERSE	PERPETUAL EXPANSION	PERPETUAL EXPANSION	EVENTUAL COLLAPSE
			

OPEN AND CLOSED MODELS of the universe are distinguished mainly by the average density of matter and by the value of the cosmic deceleration. Density is a crucial factor because in models described by the general theory of relativity it is the sole determinant of the gravitational forces that slow the cosmic expansion. Density is most easily treated as a dimensionless parameter, the ratio of the actual density to the critical density just needed to halt the expansion. Deceleration can also be expressed as a dimensionless number, the deceleration parameter, which in the models considered here is always equal to half the density parameter. These two parameters determine not only the future of the universe but also the geometry of space. The open universe is of infinite size at all times, and in it space has hyperbolic, or negative, curvature. In the universe with critical density, in which the density parameter is exactly 1, space has zero curvature, it is the flat space of Euclidean geometry. The closed universe is of finite size; in it space has spherical, or positive, curvature.

ometric properties, some of which can be represented in the two-dimensional models. A flat plane, of course, is the basis of Euclid's geometry, and on it all the Euclidean axioms and the theorems derived from them are obeyed. On a plane exactly one line can be drawn through a given point parallel to another line; the sum of the included angles in a triangle is always 180 degrees; the circumference of a circle increases in proportion to the radius, and the area of a circle increases in proportion to the square of the radius.

On the surface of a sphere no two lines are parallel, provided that a straight line is defined as one taking the shortest path between two points. Such lines are called geodesics, and on the sphere they are the great circles, any two of which always intersect. Similarly, on a sphere the sum of the included angles in a triangle is always greater than 180 degrees; the circumference of a circle increases more slowly than in proportion to the radius, and the area of a circle increases more slowly than in proportion to the square of the radius.

The surface of a pseudosphere possesses properties opposite to those of a sphere. Through a given point infinitely many lines can be drawn that are parallel to another line, or geodesic. The sum of the angles of a triangle is less than 180 degrees. The circumference of a circle increases faster than in proportion to the radius, and the area of a circle increases faster than in proportion to the square of the radius. The geometry of the three-dimensional space represented by a pseudosphere was first studied in 1826 by Nikolai Lobachevski.

In the simple cosmological models discussed here the geometry of space is uniquely related to the future behavior of the universe. It is notable that in models with Ω greater than 1 the universe is closed in both space and time. The volume of space is finite, and there are definite temporal limits, beginning with the big bang and ending with the big crunch. Models in which Ω is less than or equal to 1 are open in both space and time. Such models have a definite starting point (the big bang), but they are always infinite in extent and they expand indefinitely into the future.

Measurements of Deceleration

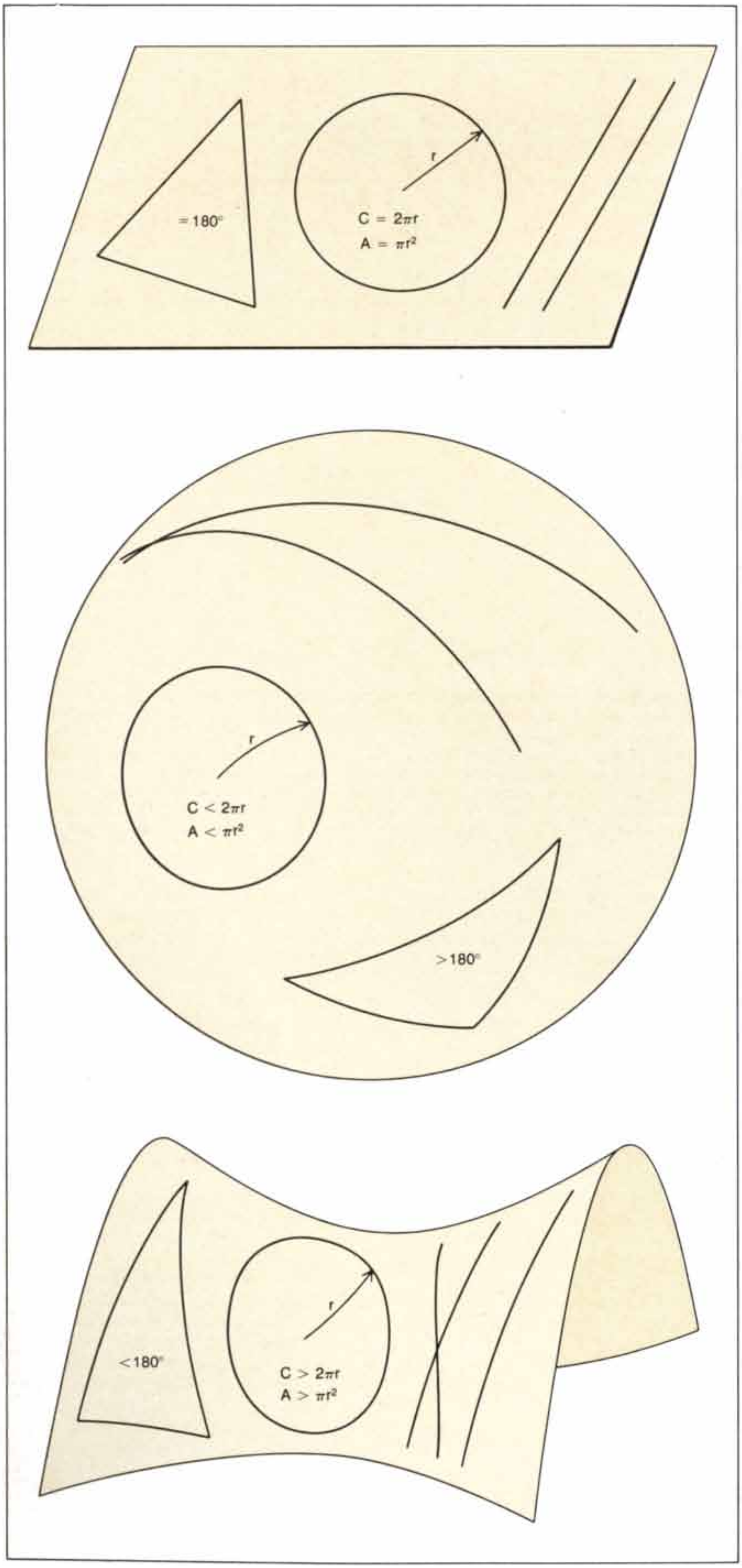
There are several possible ways of determining whether the actual universe is open or closed. All of them lead ultimately to an estimate of the rate at which the cosmic expansion is decelerating. One method is simply to measure the deceleration directly, by observing distant galaxies. It is also possible to measure the age of the universe, and by comparing it with the Hubble time (the age if there were no deceleration) to derive an estimate of how much the velocity of expansion has changed. Since the deceleration is a gravitational phenomenon, an equivalent measure is the average density of matter; comparing the actual density with the critical density gives the ratio Ω. Finally, the present abundance of certain chemical elements represents a kind of fossil record of conditions in the very early universe, including the density, and from that information too the value of Ω can be calculated. Evidence from each of these methods has contributed to our present knowledge of the state of the universe.

The deceleration of the cosmic expansion is usually expressed in terms of a dimensionless number called the deceleration parameter and symbolized q_0. Since the deceleration is a gravitational effect, the deceleration parameter is closely related to the average density of matter. In the cosmological models considered here, which are constructed according to the general theory of relativity, q_0 is always equal to exactly half the density parameter Ω. Thus if q_0 is greater than 1/2, the universe is decelerating rapidly enough, because of its high density, to stop expanding and subsequently collapse. If q_0 is less than 1/2, the expansion cannot stop because the density is too low to halt it.

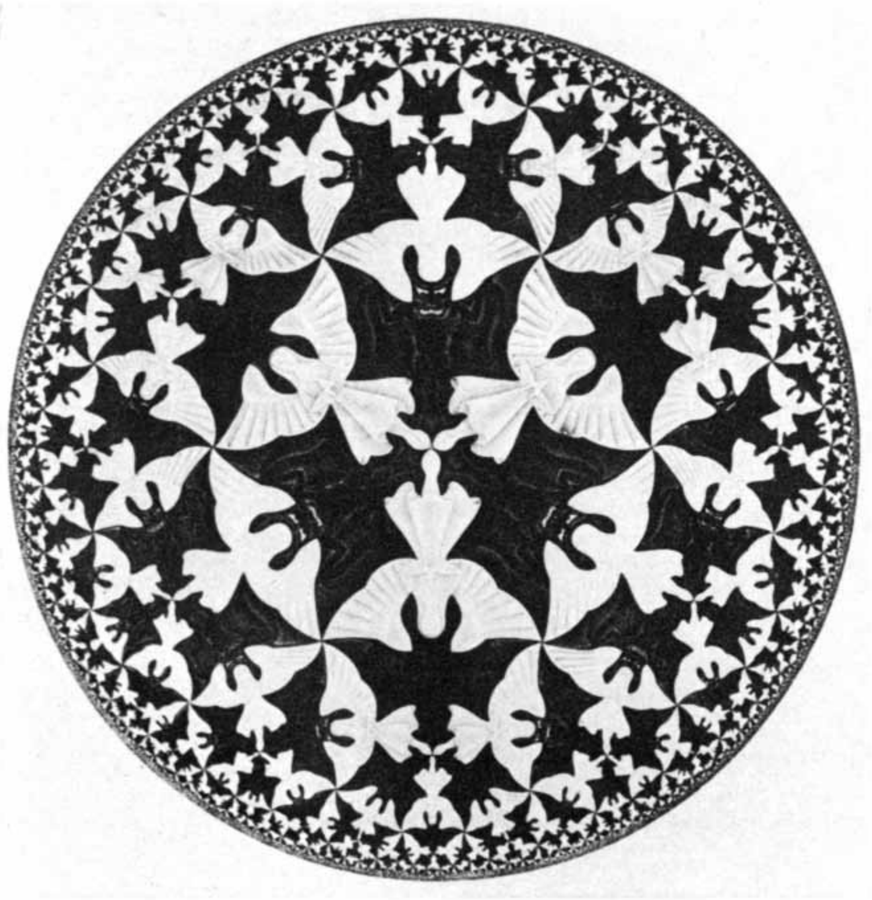
An obvious approach to determining the rate of deceleration would be to measure the radial velocity of a single galaxy at two times in order to see how much it has slowed down in the interval. Unfortunately the change in velocity expected in a human lifetime is far too small to be detected; indeed, the experimental errors involved in the determination are many orders of magnitude larger. Because of the finite speed of light, however, it is possible to measure the velocities of galaxies in the remote past and to compare them with velocities representing more recent eras. The comparison is possible because as we look at increasingly distant objects in the sky we also see farther and farther into the past. The relation is obvious when distances are measured in light-years: if a galaxy is a billion light-years away, the light we receive from it today was emitted a billion years ago, and the Doppler shift in its spectrum must reflect the distant galaxy's velocity then with respect to our own velocity now. Thus if the cosmic expansion is decelerating, the constant ratio of velocity to distance discovered by Hubble is not expected to hold for the most distant galaxies. At extreme distances the ratio should increase, or in other words the velocities should be greater than those predicted by Hubble's law.

In order to measure the deceleration by this method it is necessary to have an independent measure of the distances of the galaxies. For all but the nearest galaxies the only practical method of estimating the distance to a galaxy is from its apparent luminosity. If all galaxies at all times had the same intrinsic luminosity, then their apparent brightness would vary simply as the inverse of the square of their distance, and the calculation of distance would be a straightforward procedure. Of course, they do not all have the same intrinsic luminosity.

Random variations in brightness (caused, for example, by differences in size) may produce errors in any individual measurement. Because of such variations it is necessary to acquire a large volume of data and to submit it to statistical analysis, but in principle random variations are not a serious con-



GEOMETRY OF SPACE characteristic of each model universe has an analogous surface. The properties of the surfaces are defined by the Euclidean axioms and theorems on parallel lines, on the included angle of a triangle and on the circumference and area of a circle. The flat space of a critical universe is represented by a plane, and the positively curved space of a closed universe corresponds to the surface of a sphere. Some of the properties of the negatively curved space of an open universe can be demonstrated on a saddle-shaped surface, but the saddle is an imperfect analogue because it has a center. The best representation of an open universe is an infinite surface called a pseudosphere, which cannot be constructed in a three-dimensional space.



SURFACE OF A PSEUDOSPHERE is represented in an etching, *Circle Limit IV,* by M. C. Escher. In the etching the surface is projected onto a plane. As in any map projection, the scale is not constant; on the pseudosphere itself the figures of angels and demons would all be the same size. If a single figure is regarded as a unit of measure, it is apparent that the circumference of a circle increases much faster than in proportion to the radius. Similarly, each figure defines a triangle (with the vertexes at the feet and the wing tips); from the number of triangles that meet at each vertex it can be shown that on the pseudosphere the sum of the angles of a triangle is less than 180 degrees. The pseudosphere is an infinite surface of negative curvature, analogous to space in a universe that expands forever. It has no privileged position that could be considered a center, and projection would be unchanged if it were centered on any other point.

cern, since in any large sample they can be expected to cancel out. Systematic variations, however, require explicit correction.

Theories of stellar evolution suggest that the combined light from all the stars in an isolated galaxy probably declines at a rate of a few percent per billion years. Galaxies were therefore probably brighter in the remote past. If the change in brightness were neglected in making measurements of the deceleration, the calculated distances would be too small and as a result the rate of deceleration would be overestimated. The decline in brightness would seem to be quite modest, but it changes the calculated value of the deceleration parameter q_0 by about 1, which is more than enough to decide between an open universe and a closed one. The best current observations, which take into consideration the changes in luminosity resulting from stellar evolution, suggest that q_0 is closer to zero than to 1/2 and therefore that the universe is open and perpetually expanding.

There is a further large uncertainty in the determination of the deceleration. Most of the observed galaxies are situated in relatively dense clusters, and possible interactions between galaxies ought to be taken into account. For example, it has recently been shown that in clusters large galaxies swallow smaller ones, with a consequent change in luminosity and size. It is not yet possible to predict the magnitude of the change, or even to be sure whether it makes the measured luminosity increase or decrease. Adding stars to a galaxy should make it brighter, but in cosmological observations only the luminosity of the central part of the galaxy is measured. If the cannibal galaxy swells significantly, the number of stars in the central region might be reduced and the galaxy would appear fainter.

The Age of the Universe

As a result of statistical uncertainties and our imperfect knowledge of galactic evolution the value of q_0 derived from measurements of recessional velocity is very uncertain. From this test alone one cannot conclude that q_0 is less than 1/2 and that the universe is open; on the other hand, very large values of q_0, such as q_0 equals 2, do seem to be excluded.

The second approach to determining the fate of the universe is to measure its age. If the expansion were not decelerating at all, the age would be equal to the Hubble time. Since it is decelerating it must be somewhat younger than the Hubble time. By finding the difference between the actual age and the Hubble time it is possible in principle to calculate the deceleration parameter q_0.

The age of the universe can be estimated by two methods; both actually yield only lower limits, since they measure the ages of objects in the universe, but those objects were probably formed within the first billion years or so after the big bang. The first method consists in determining the age of the oldest stars that can still be observed today. The oldest stars close enough for detailed observation are thought to be those in the globular clusters associated with our own galaxy. Models of stellar evolution indicate that they are between eight and 16 billion years old.

The age can also be estimated from measurements of the relative abundance of certain heavy elements. All the elements heavier than iron, including several radioactive ones, are thought to have been formed in supernovas, which have probably been exploding in the galaxy since its formation. Because each radioactive element decays at a constant rate, the ratio of the abundance of each radioactive element to the abundance of its decay products can reveal the average age of the heavy elements. The ratios indicate that the age of the galaxy is between six and 20 billion years. The two calculated ages are thus consistent, and they suggest that the big bang took place between eight and 18 billion years ago.

Average Density

Whether a given age within the allowed range corresponds to an open universe or a closed one depends on the value of the Hubble time, and as we have seen that value is not easily determined. Moreover, even if the Hubble time is assumed to equal the recent best estimate of 19 billion years, neither the age limits nor the exclusion of q_0 values greater than 2 is sufficient to decide whether the universe is open or closed [*see illustration on page 72B*]. The issue can be decided only by imposing further constraints.

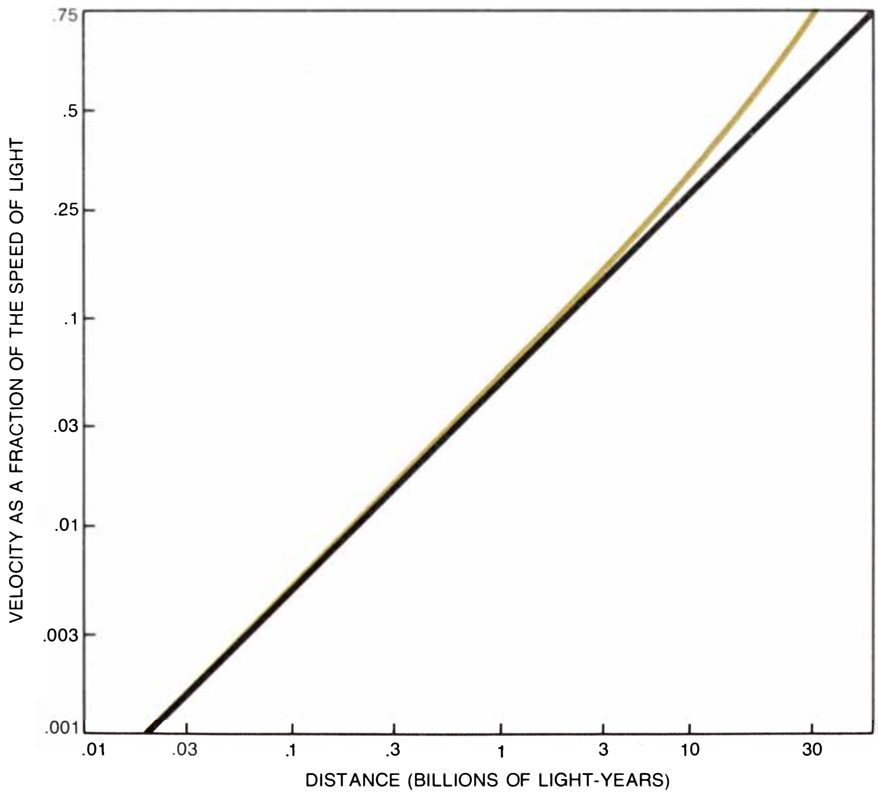
The third test consists in measuring the average density of matter in the universe and thereby deriving the density parameter Ω. A lower limit to the density can be obtained by considering only the mass associated with visible galaxies. The density is found by counting the galaxies in a given volume of space, multiplying by the masses of the galaxies and dividing by the volume.

Weighing a galaxy is not as difficult as it might at first seem. Few galaxies are completely isolated; most are found in small groups or in large clusters, and their mass can be deduced from observations of the gravitational effects they exert on one another. Two galaxies in orbit around each other, for example, must have a gravitational attraction just sufficient to balance the centrifugal force. If their separation and their velocities with respect to each other are known, the determination of their combined mass becomes a simple exercise in Newtonian mechanics. The procedure for clusters of many galaxies is only slightly more complicated. Significantly, the mass calculated in this way includes not only the mass of the galaxies but also the mass of any other matter in the cluster. Constituents that would not be visible, such as black holes or extragalactic dust and gas, are automatically taken into account.

Estimates of the mass of a great many galaxies, combined with counts of the galaxies in large volumes of space, give an indication of the value of the density parameter Ω. If the mass associated with galaxies represents all the mass in the universe, then Ω is only about .04 and the universe must definitely be open and infinitely expanding. This value could be uncertain by a factor of 3, so that a value of Ω as great as .12 would still be consistent with observations, but that is still well below the value of 1 required to close the universe.

The density of the universe can also be estimated by comparing the behavior of distant galaxies with the behavior of those in the local supercluster, the aggregate of galaxies that includes our own local group in addition to many other small groups and the somewhat larger Virgo cluster. Within the local supercluster the mean density of galaxies is two and a half times greater than that in the universe as a whole. If all mass is associated with galaxies, then the average density of matter must also be two and a half times greater in the supercluster than outside it. The difference in density should give rise to a difference in the rate of expansion; because the local density is greater, nearby galaxies should be more strongly decelerated. The magnitude of the difference depends on the value of Ω; if Ω is large, there should be a considerable difference. If Ω is small, then the deceleration everywhere is small, and even a local enhancement in density by a factor of two and a half would cause little change. In fact, the difference is undetectable, being smaller than the probable observational errors. The most straightforward conclusion is that Ω is very small, probably no larger than .1.

Both methods of estimating density are explicitly confined to the matter associated with galaxies, and an obvious objection to them is that there might be substantial amounts of matter elsewhere in the universe. That possibility cannot be excluded,



DECELERATION of the cosmic expansion can be detected in the recessional velocities of galaxies in the remote past. It is possible to look into the past by observing the most distant galaxies, since light reaching us now was emitted a length of time ago given by the galaxy's distance in light-years. The deceleration is thus perceived as a departure from Hubble's law; if there were no deceleration, the ratio of velocity to distance would be constant (*black line*); with deceleration the ratio increases at extreme distances (*colored line*). Because of the difficulty of estimating the distance to galaxies it has not been possible to measure the rate of deceleration precisely, but values of the deceleration parameter greater than about 2 have been excluded.

A MITS Altair Computer Report

Technology of Three Altair Computers

MITS Altair Computers are built around recently developed "microprocessor" integrated circuits. These compact, wafer shaped "chips" are about 2 inches long, 1/2 inch wide, and 1/16 inch thick. They represent over 10,000 electronic components, and they contain all the logic circuitry of a full-blown computer.

1. The Altair 8800 Computer is an "open-ended" general purpose computer built around the 8080 microprocessor chip. Its basic configuration includes a CPU (Central Processing Unit) circuit board, front-panel control board, power supply, and case. Up to 16 circuit boards can be added inside the computer simply by plugging them in. These boards could include a wide variety of memory boards, interface boards, and processor option boards.

The Altair 8800 Computer can be programmed from the front panel indicator lights (LED's) and switches, or it can be interfaced to any number of computer peripherals. These peripherals include teletypewriters, line printers, floppy disks, paper tape reader/punch, CRT terminals, and more.

Since the Altair 8800 can be configured to meet the needs of the user, its applications are virtually unlimited.

2. The Altair 680 Computer is built around the 6800 microprocessor chip. It is smaller and more compact than the Altair 8800, measuring just 11 inches wide by 11 inches deep by 4-11/16 inches high.

While the Altair 680 was designed primarily for dedicated programming—such as industrial process control, several hundred Altair 680's have been sold to hobbyists for experimentation. One reason for this is that the Altair 680 is a complete computer in itself. Its main component board contains the CPU, 1,024 words of memory (RAM), a PROM monitor for loading paper tapes and an I/O port that can be wired for one of four different types of peripherals. Like the Altair 8800, it too can be programmed from the front panel.

3. The Altair 8800B Computer, MITS' newest computer, is basically a second generation design of the Altair 8800. This machine incorporates some of the most recent advances in computer technology. More information can be obtained from the factory.

The Advent of the Computer Club

Since the introduction of the **Altair 8800 Computer** in January of 1975, computer clubs have been springing up across the country. The largest of these, the Southern California Computer Society, now has a membership of over 2000.

Computer clubs are groups of individual computer owners who meet regularly to discuss mutual problems and carry out joint projects. In addition to using computers for traditional applications such as computer games, computer art, and educational programming, many computer hobbyists are experimenting with more bizarre applications. These applications include voice input/output and biofeedback controlled peripherals.

The Computer as a Household Pet

One computer hobbyist has an Altair based computer, named Ralph, which he regards as a household pet. Besides being inexpensive to feed and care for, Ralph can perform a number of entertaining and practical tricks. These include playing blackjack, balancing a checkbook, teaching basic mathematics, turning on the coffee pot in the morning, controlling the temperature and humidity of the house, flipping on the yardlights at dusk, and acting as a burglar alarm if need be.

Computer Costs

Altair computers, marketed in both kit and assembled units, have helped to bring about drastic cuts in the price of computing. The Altair 680, for instance, is currently selling for $345 in kit form. A complete Altair 8800 system with 16K of memory, a floppy disk, Teletype, and Extended BASIC language software sells for under $4,000.

These low costs have opened the doors to thousands of individuals and small businesses. And they have made it practical to use the computer for a wide range of new applications.

Altair Customers

While the majority of Altair owners have some sort of technical background, they include a broad range of people from engineers to retail managers to artists, teachers, doctors, editors, housewives, musicians, lab technicians, businessmen, attorneys, and factory workers. In addition to some of the above mentioned applications, they are using their Altairs for such applications as medical electronics, instrument control, model train and airplane control, text editing, mailing list maintenance, software development, music synthesis, interface to larger computers, graphics display, OSCAR tracking, bookkeeping, and timeshare services.

More Information

Space does not permit us to present a complete discussion of low-cost computing here, but we have prepared a complete Altair documentation notebook for those of you who wish to investigate the matter further. This notebook includes a catalog of all Altair products, technical literature, a more complete discussion of the home computer, a list of computer clubs, a list of authorized Altair dealers, a sample Altair Computer Notes newsletter, and much more in a sturdy 3-ring binder. Until April 30, 1976, it will sell for $5 plus $1 for postage and handling.

mits

Altair Coupon

☐ Please send me a copy of your Altair Documentation Notebook. Enclosed is $5 plus $1 to cover postage and handling.
☐ Please send me a free catalog
☐ Enclosed is a check for $______
☐ BankAmericard #______ ☐ or Master Charge #______
NAME______
ADDRESS______
CITY______ STATE & ZIP______

MITS/2450 Alamo SE/Albuquerque, NM 87106/505-265-1951

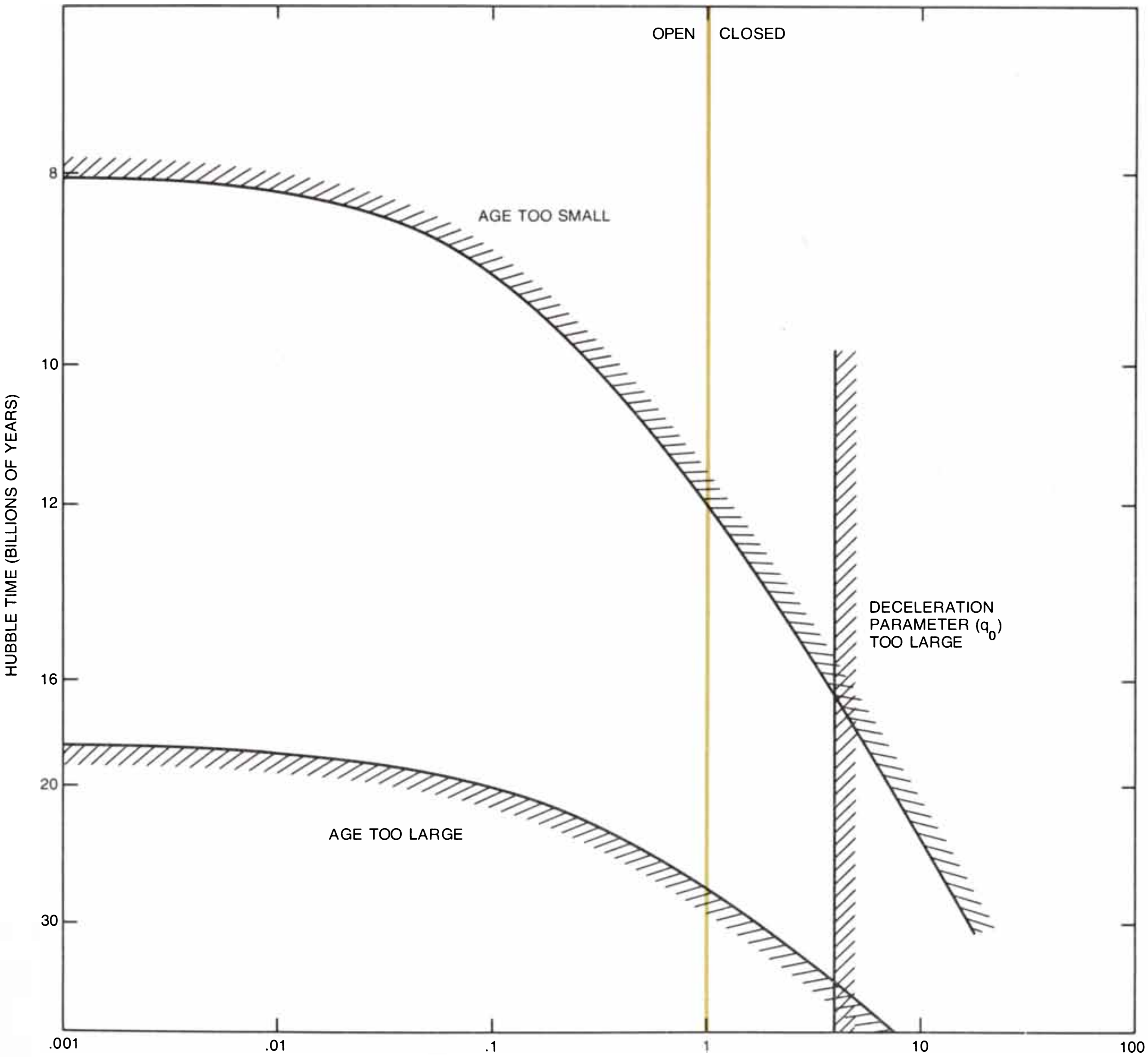
MITS/2450 Alamo SE/Albuquerque, NM 87106/505-265-1951

but there is no evidence to substantiate it.

Current theories show that clusters of galaxies could have formed from a universe in which matter was distributed much more smoothly than it is now. Debris left over from the formation of galaxies would also be collected by the clusters. Thus any particles that are not now in the clusters must have preferentially avoided them, that is, the particles must have had the special and unusual initial positions and velocities that would enable them to escape capture. Even if a large amount of matter were distributed uniformly outside the clusters now, it would fall into them in a few billion years.

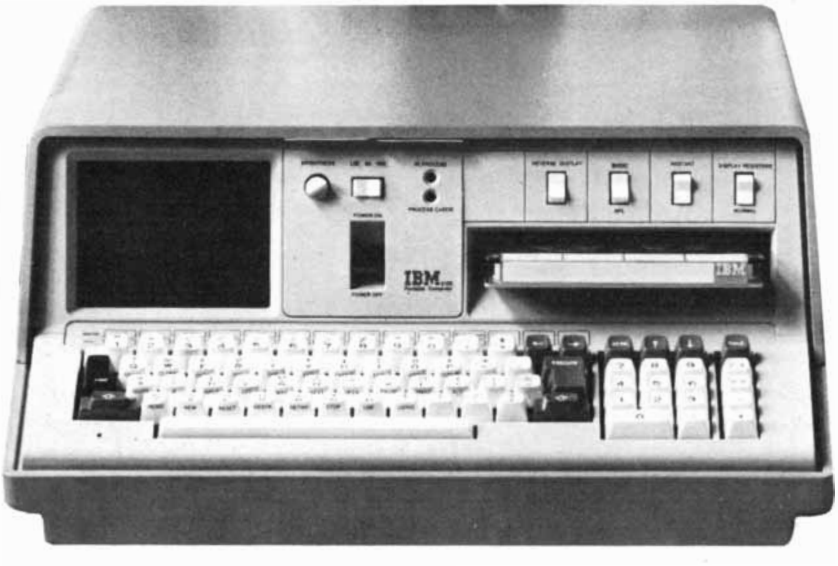
Alternatively, the necessary mass could consist of some uniformly distributed medium with enough internal pressure to be unaffected by the gravitation of galaxies. It might, for example, be made up of large numbers of neutrinos or of gravitational waves. There is, however, a strong argument against such a pervasive "radiation-like" medium: it would almost certainly have prevented galaxies and clusters of galaxies from ever forming.

The density of all matter in the universe, whether or not it is associated with galaxies, can in principle be determined, but only by extrapolating from conditions in the present universe to those a few minutes after the big bang. The simplest assumptions about that early period suggest that the temperature and density were high enough for some subatomic particles to interact and form sizable amounts of some of the lighter nuclei. In particular a proton and a neutron could fuse to make a nucleus of deuterium, and most of the deuterium nuclei would quickly combine to form helium nuclei, composed of two protons and two neutrons. The proportion of deuterium and helium



INITIAL CONSTRAINTS on the state of the universe are provided by determinations of its age and of the deceleration parameter. Estimates of the age of the oldest stars and of the average age of heavy elements suggest that the universe is between eight and 18 billion years old; the corresponding limits to the Hubble time depend on the density. Observations of distant galaxies provide an upper limit for the deceleration parameter: it cannot exceed 2, and the density parameter therefore cannot be greater than 4. The constraints derived from these measurements alone do not determine whether the universe is open or closed, since they encompass both kinds of model.

The new IBM 5100. An ideal personal computer for scientists and engineers.



Available with three language options: BASIC, APL or both.

The new IBM 5100 is a sophisticated problem-solving aid incorporating the very latest in semi-conductor technology.

It features an alpha-numeric keyboard, a 15 key calculator pad, 14 interactive command keys and a video display screen that shows 1024 characters in sixteen, 64 character lines. In addition, the 5100 display screen has a scrolling capability that allows you to move information up or down for easy editing.

The 5100 goes to work wherever you need it to help solve the most complicated problems.

The 5100 uses a tape data cartridge which holds more information than the cassettes used in other small computers.

Still another advantage of the 5100 is its large main storage capacity, up to 64K bytes.

The 5100 also features an optional 80 character per second bi-directional printer, an auxiliary tape unit that reads up to 2850 characters per second and an optional communications adapter that turns the 5100 into a terminal capable of transmitting information to a larger processor.

The most powerful language in small computers

The 5100 is available with three language options: BASIC, APL or both. BASIC is a generally well known, easy-to-use language using English-like commands to solve engineering and scientific problems.

But it is APL that really sets the IBM 5100 Portable Computer apart. APL is the most powerful language in small computers today. It uses concise symbols to solve complex problems, including matrix commands that use a single symbol to solve multiple equations. APL routines are as close to the original mathematics as possible and take up just a fraction of the space required for the same routines written in other high level languages.

Just flip a switch to go from BASIC to APL.

The IBM 5100 Portable Computer is also available with extensive problem-solver libraries. These IBM developed programs are highly interrogative to help you focus on getting answers instead of how to run the program.

Statistical library

The statistical library provides a comprehensive capability covering the most commonly used statistical techniques. It covers such areas as:

Elementary Statistics
- Histogram
- Cross Tabulation
- Transformation, alters columns in matrices
- Tally
- Moment
- Correlation
- t-Test
- Chi-square Test
- Survival Rate
- Probit Analysis

Time Series Analysis
- Moving Average
- Seasonal Analysis
- Cyclical Analysis
- Autocovariance and Autocorrelation
- Triple Exponential Smoothing
- Crosscovariance and Crosscorrelation

Regression and Design Analysis
- Simple Regression
- Multiple Regression
- Polynomial Regression
- Stepwise Regression
- One-way analysis of variance
- Factorial Design

Multivariate Analysis
- Discriminant Analysis
- Canonical Correlation
- Factor Analysis

Non-Parametric Statistics
- Kendall Rank Correlation
- Kendall Coefficient of Concordance
- Sign Test
- Wilcoxon Matched Pairs Signed Rank Test
- Mann-Whitney U-Test
- Friedman Two-way Analysis of Variance
- Cochran Q-Test
- Biserial Correlation
- Point Biserial Correlation
- Tetrachoric Correlation
- Phi Coefficient

Plus there are four data generation routines that provide facilities to enter and display or print data. Or correct, modify, generate or smooth it.

The IBM data cartridge holds up to 204,000 characters on highly-reliable quarter-inch magnetic tape.

Mathematical library

The mathematical library provides pre-programmed routines featuring a conversational interactive capability that allows you to define a problem through a series of questions and answers. The areas covered include the following:

- Solutions to Simultaneous Linear Equations
- Matrix Eigenproblem
- Eigenvalues and Eigenvectors
- Integration
- Differentiation
- Interpolation, Approximation and Smoothing
- Zeroes and Extrema of Functions
- Ordinary Differential Equations
- Finite Fourier Transform
- Special Functions
- Linear Programming

Business Analysis Library

There are 30 interactive routines in the business analysis library. They all utilize the BASIC language and can be used to solve problems without the help of specialized data processing personnel. Included in the library are easy-to-use procedures for:

- Spread Sheet Analysis
- Investment Analysis
- Break Even Analysis
- Depreciation Analysis
- Time Series Analysis
- Graphic Presentation

Or, you can program the 5100 yourself. It comes with a self-study training package that makes it easy to learn and easy to use without taking any classes or relying on specially trained experts.

The IBM 5100 Portable Computer. It can help solve complicated problems wherever they happen to be. Because unlike most other so-called small computers the 5100 is truly portable, weighing about 50 lbs.

To arrange for a demonstration, call your local IBM General Systems Division office or fill out this coupon.

IBM®

IBM General Systems Division
P. O. Box 2068, Atlanta, Georgia 30302

☐ I would like more information about IBM's new 5100.
☐ I would like a demonstration of IBM's new 5100.
My major area of interest is:
☐ Engineering/Scientific ☐ Statistical Analysis
☐ Business/Financial Analysis

Name

Title

Company

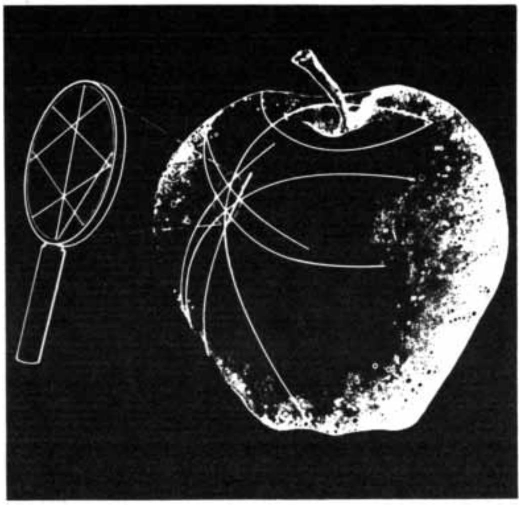
Address

City State Zip

Phone

Some of the best science books come from W. H. Freeman and Company

Gravitation



by CHARLES W. MISNER, University of Maryland,
KIP S. THORNE, California Institute of Technology, and
JOHN ARCHIBALD WHEELER, Princeton University

"A definitive, stunning introduction to geometrodynamics (general relativity) in the crisp modern symbolism of differential geometry. . . . An unrivalled exposition of one of the supreme theories of mathematical physics." —*American Mathematical Monthly*, April 1974

"*Gravitation* is one of the great science texts of our day. . . . "
—*Journal of the British Astronomical Association*, October 1974

"A pedagogic masterpiece. . . . In this era of intense interest in gravitation we may be grateful to Misner, Thorne, and Wheeler for guiding us through its intricacies with such insight."
—D. W. Sciama, *Science*, March 22, 1974

"Required reading for all those working in, or wondering about, the fascinating new field of gravitational physics . . . rarely has a scientific field been so well served."—*American Scientist*, January/February 1974

1973, 1279 pp., 137 illus., 163 boxes, cloth: ISBN 0-7167-0334-3, $49.50; paper: ISBN 0-7167-0344-0, $22.50

DNA Synthesis

by ARTHUR KORNBERG, Stanford University

Written by a Nobel Laureate in Medicine, this is the first comprehensive treatment of DNA synthesis emphasizing its biochemical aspects and recent developments. *DNA Synthesis* contains discussions of the structure and function of DNA, its precursors, enzymology, repair, recombination, restriction, transcription, and the synthesis of genes. The 194 illustrations include electron micrographs and many colored figures. There is an extensive and up-to-date bibliography.

"An excellent book about a complex area of investigation, written with great clarity by a master of the subject. . . . Will long remain as a stimulating and useful work. . . . "
—*Times Higher Education Supplement*, December 20, 1974

1974, 399 pp., 194 illus., 69 tables, ISBN 0-7167-0586-9, $19.00

Astronomy and Cosmology *A MODERN COURSE*

by FRED HOYLE

Fred Hoyle has created a unique and exciting introduction to astronomy. In it he abandons the traditional encyclopedic approach to the subject in favor of a fuller treatment of such topics as the large-scale structure of the universe, the history of matter, the origin of the planets, and the origin of life. "I have tried to emphasize," he writes, "the uncertainties and tensions to which the astronomer and physicist are exposed when they work at the frontiers of their sciences."

Astronomy and Cosmology contains over 600 illustrations, including a special section of photographs in full color. The reader needs no more than a knowledge of elementary algebra to follow the arguments of this provocative and profound book.

1975, 711 pp., 617 illus., 45 tables, ISBN 0-7167-0351-3, $15.95

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, California 94104, U. S. A.

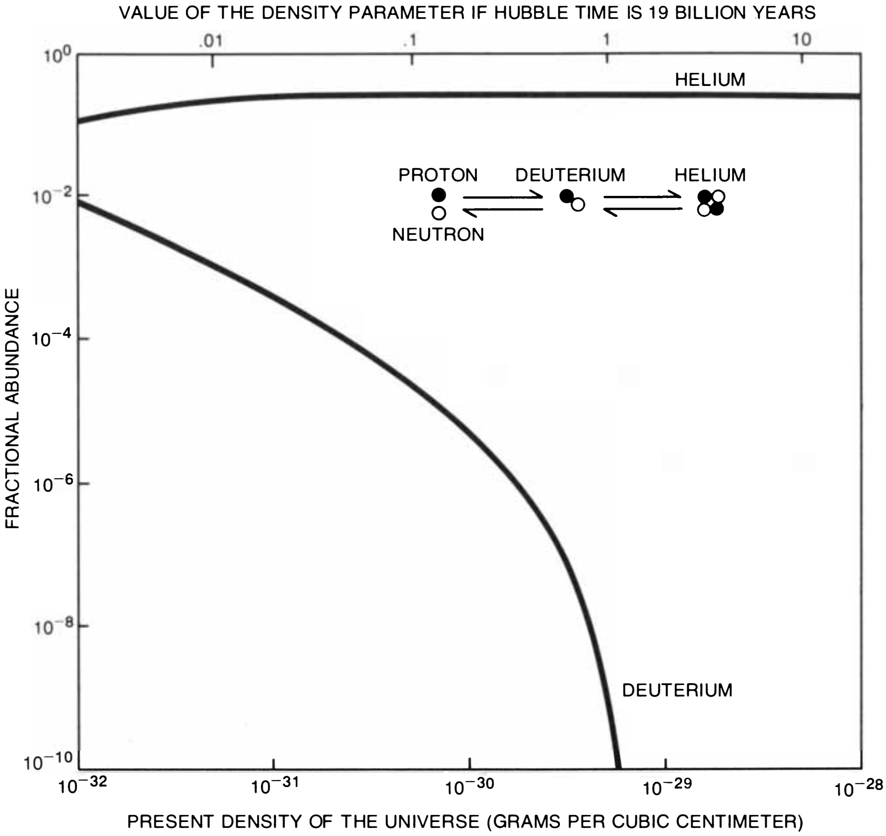
formed in this way depends on the density of the universe at the time when it was hot enough for the reactions to take place. From the early density and from the present temperature of the microwave background radiation it is possible to deduce the density today.

Primeval Density

Mathematical models indicate that for the entire possible range of densities in the early universe between 20 and 30 percent of the matter is converted into helium. The helium abundance measured in a variety of astronomical objects is in this narrow range, which strongly supports the fundamental assumption that the universe went through a period of extreme temperature and density shortly after the big bang. The present abundance of deuterium depends strongly on the early density [*see illustration at right*]. The relative abundance of deuterium in nearby interstellar space has been measured by the third Orbiting Astronomical Observatory satellite, named *Copernicus*. After taking account of the deuterium depleted by nuclear reactions in stars, the measured abundance yields an average present density of about 4×10^{-31} gram per cubic centimeter. The measurement is a sensitive indicator of density: if the universe were 10 times denser, the big bang would have made less than a thousandth the observed abundance of deuterium. For this reason uncertainties in the measurement do not result in large uncertainties in the estimated density.

Whether the density determined by the deuterium abundance represents an open universe or a closed one depends on the Hubble time. As we have seen, if the Hubble time is 19 billion years, the critical density is 5×10^{-30} gram per cubic centimeter, so that Ω, the ratio of actual density to critical density, is about .08. For any value of the Hubble time between 13 and 19 billion years, the value of Ω derived from the deuterium abundance is consistent with that derived from the density of galaxies. Conversely, for any plausible value of the Hubble time, a value of Ω as great as 1 is inconsistent with the density required for the manufacture of deuterium.

The abundance of deuterium would seem to provide powerful evidence that the universe is open; unfortunately the arguments supporting that conclusion are somewhat insecure. In extrapolating from the present state of the universe to conditions soon after the big bang the simplest possible model has been employed; other models might allow the observed amounts of helium and deuterium to be made in a much denser, closed universe. Those models are more complicated, even somewhat contrived, but they cannot be excluded. Moreover, the significance of the deuterium abundance depends entirely on the assumption that all the deuterium in the universe was made shortly after the big bang. Other sources, such as supernovas, have been suggested, but so far no mechanism has been found that would create a significant amount of deuterium without violating other constraints.



DENSITY OF THE EARLY UNIVERSE influenced the synthesis of deuterium and helium, and from the relative abundances of those elements the present density can be inferred. Deuterium is thought to have been formed by the fusion of protons and neutrons in the first few minutes after the big bang, but if the density then was too great, most or all of the deuterium would have been converted into helium. The abundances of both elements are shown as fractions (by mass) of all the matter in the universe. If the simplest models of the early universe are correct, and if deuterium has not been formed in more recent events, the observed abundance suggests that the density of the universe cannot be greater than about 4×10^{-31} gram per cubic centimeter.

Plausible Models

The measurements of the deceleration parameter, of the age of the universe, of the density of galaxies and of the abundance of deuterium all impose independent constraints on the state of the universe. If the measurements are consistent, there must be some class of models of the universe that is allowed by all the constraints. Indeed there is, and moreover it is a relatively small class, so that interesting predictions about the future of the universe are possible [*see illustration on page 79*]. If the universe is not too old, and if its density is at least equal to that observed in galaxies but not too great to make deuterium, the value of Ω must be between .04 and .09. That is far below the value required for a closed universe.

Two additional observations are consistent with the allowed values of Ω and the Hubble time. The calculated age of the stars in globular clusters is sensitive to the abundance of helium, and as we have seen that in turn is determined by the density of the universe. It is therefore encouraging to find that the age and the helium abundance allowed by the combined constraints are compatible with what is known of the globular-cluster stars.

The constraints also require that the Hubble time itself be between 13 and 20 billion years. The direct determination of the Hubble time is difficult, but in recent years much effort has been expended on the problem by Allan R. Sandage and Gustav A. Tammann of the Hale Observatories. Their best value is 18 ± 2 billion years. Robert P. Kirshner and John Kwan of the California Institute of Technology have employed a different technique, relying on the properties of exploding stars in distant galaxies, to make an independent measurement of the Hubble time. They place the value between 13 and 22 billion years.

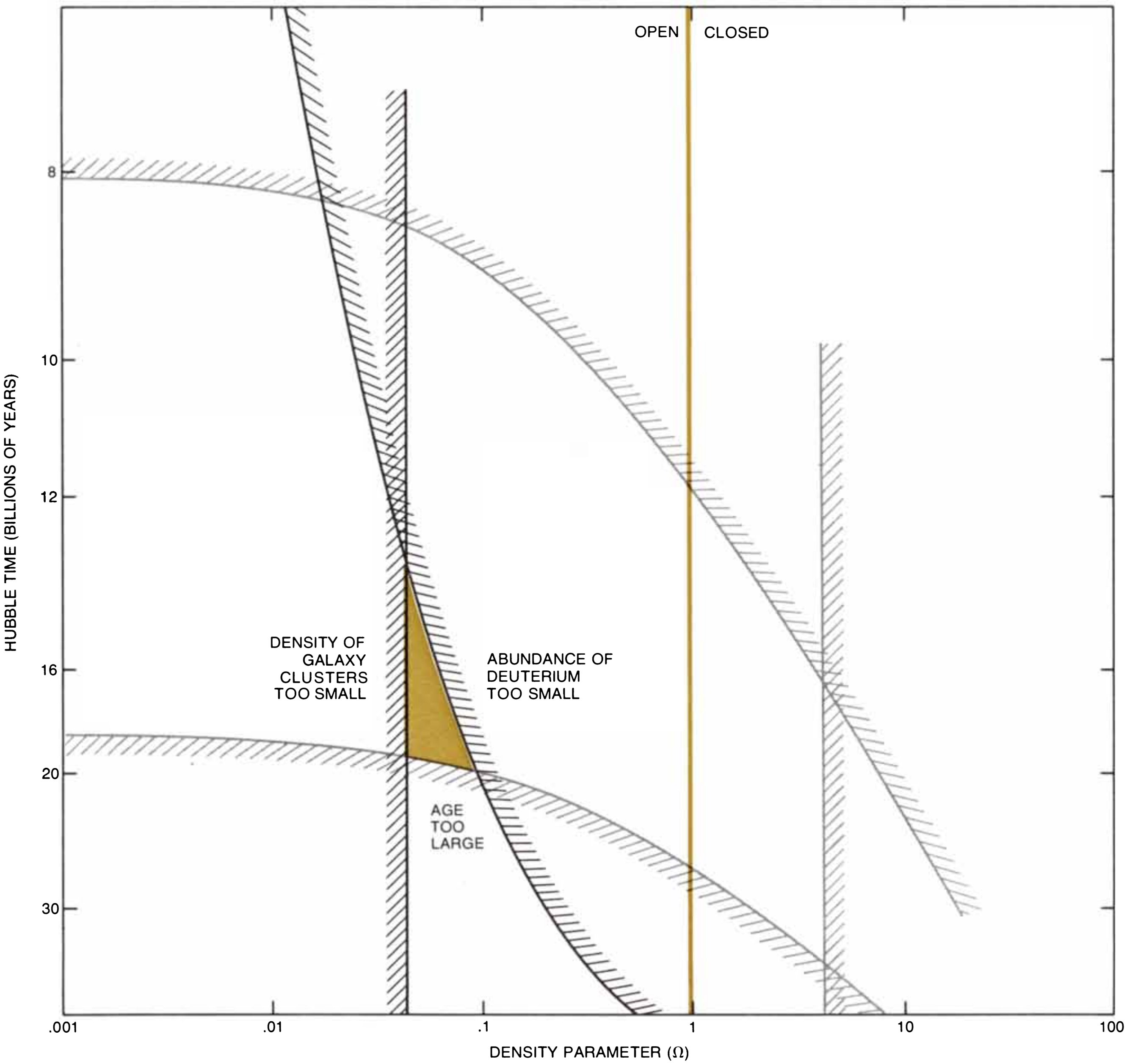
The consistency of the results obtained by such diverse methods is gratifying, and it encourages confidence that the cosmological model is well determined and the fate of the universe is known. Because of uncertainties in the data and in the theory em-


WHEN YOU SPEND $10,000 FOR A CAR, YOU SHOULDN'T BE AFRAID TO DRIVE IT.
Any man who has traveled the highway to success shouldn't feel he has to detour around potholes.
Yet it seems many big, expensive cars today are better prepared for country club driveways than city streets and back roads.
The elegant new Volvo 264 is not your commonplace rich man's car. It offers more than luxury. It's engineered to afford you the privilege of abusing it.
A new front suspension combining springs *and* struts absorbs jolts and increases stability by reducing roll. Thousands upon thousands of spot-welds (each one strong enough to support the entire weight of the car) fuse body and frame into one solid, silent unit.
The Volvo 264 is extremely agile. A new light alloy, fuel-injected overhead cam V-6 cuts weight. (The 264 is 1,100 pounds lighter and almost a foot shorter than the new "small" Cadillac Seville. Not to mention almost $4,000 smaller in price.)
The 264 GL is also the most lavishly equipped Volvo we make. Leather everywhere you sit. A heated driver's seat. Power front windows. Sunroof. And air conditioning.
So if you're thinking about buying a luxury car, give some thought to the Volvo 264.
You've worked hard to afford the best. You deserve a car that can take the worst.
304 CWK
VOLVO 264
The car for people who think.

ployed to interpret them, however, the apparent agreement may yet prove to be fortuitous.

A firm prediction of the models considered here is that the deceleration parameter equals half the density parameter, and as we have seen this prediction cannot yet be tested. If in the future it is found to be wrong, more complicated cosmological models will be required. For example, one class of models employs a modification of general relativity once suggested by Einstein, in which a parameter called the cosmological constant is introduced. In these models space itself generates an attractive or repulsive gravitational force, and as a result the deceleration is no longer related in a simple way to the density.

Taken one at a time, each of the constraints we have discussed has possible loopholes. In particular, some of our colleagues would disagree with the small density derived from estimates of the mass associated with galaxies, and with the inclusion of a constraint on density derived from the production of deuterium. Our arguments and conclusions, however, derive their credibility from the fact that a consistent cosmological model can be constructed by the most straightforward interpretation of each piece of evidence. It is remarkable that such diverse factors as the age of stars, the mass of galaxies, the abundance of chemical elements and the observed rate of expansion of the universe can all be interpreted naturally in terms of one of the simplest cosmological models. This model describes a universe that is infinite in extent and that will expand forever. The case for an open universe is by no means closed, but it is strongly supported by the weight of the evidence.



ADDITIONAL CONSTRAINTS combine to suggest that the universe will expand forever. The abundance of deuterium implies an upper limit to the density of all matter in the universe, and therefore also limits the density parameter, although the numerical value of that limit depends on the Hubble time. A maximum value of the Hubble time itself is defined by the estimates of the ages of stars and the heavy elements. Finally, calculations of the mass associated with clusters of galaxies supply a lower limit to the density parameter. Barring seemingly improbable complications, these constraints confine all allowed models to a small range of values of the density parameter and the Hubble time (*colored area*); all models in that range describe a universe that is open, infinite and perpetually expanding.

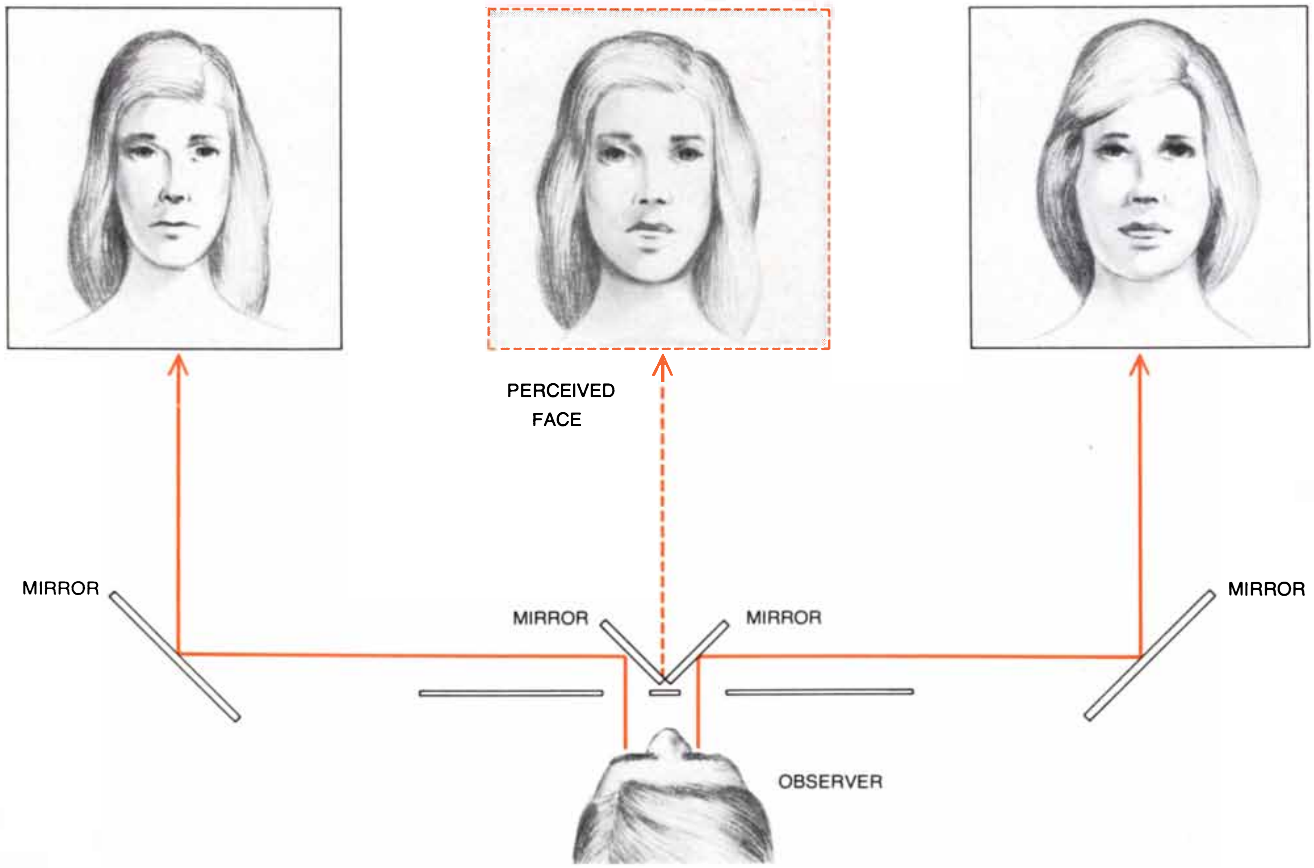
The Resources of Binocular Perception

Studies with stereograms that change at random reveal that the perceptual system extracts depth and motion information from the visual input even before we are conscious of what we see

by John Ross

When we view objects in three-dimensional space with two eyes, our visual system does far more than simply combine the images projected on the two retinas. It is well known that binocular depth perception depends on the disparities between the images formed by the two eyes. In addition to binocular perception's exquisite sensitivity to disparities in relative position it displays a remarkable capacity for decision making. The visual system in effect constructs three-dimensional scenes from the two-dimensional images formed on the retinas by fitting the visual information into a conceptual framework. Among the major questions that arise are the following: What are the information-processing capacities involved in binocular perception? And what are the resources at the disposal of those capacities?

With the invention of the stereoscope by Charles Wheatstone in 1838, it became possible to create three-dimensional effects with two photographs of a scene taken from slightly different viewpoints. The stereoscope presents one of the photographs to the right eye and the other photograph to the left. When the two pictures are viewed simultaneously, they combine to form a three-dimensional scene. Since each eye



BINOCULAR COMBINATION of two different faces results in a curious blending of the two, often with a distinct improvement in the facial features. The mirrors are arranged so that the observer's left eye sees the face on the left and his right eye sees the face on the right. Unusual effects can be obtained when one face is a man's and the other is a woman's. The selective blending of features that occurs in binocular combination indicates that the visual system apparently has the ability to accept or to reject information on aesthetic grounds.

sees a complete picture, early investigators of binocular depth perception thought that such perception resulted from an optical combination of the monocular images, or at least depended on recognition of monocularly perceptible features.

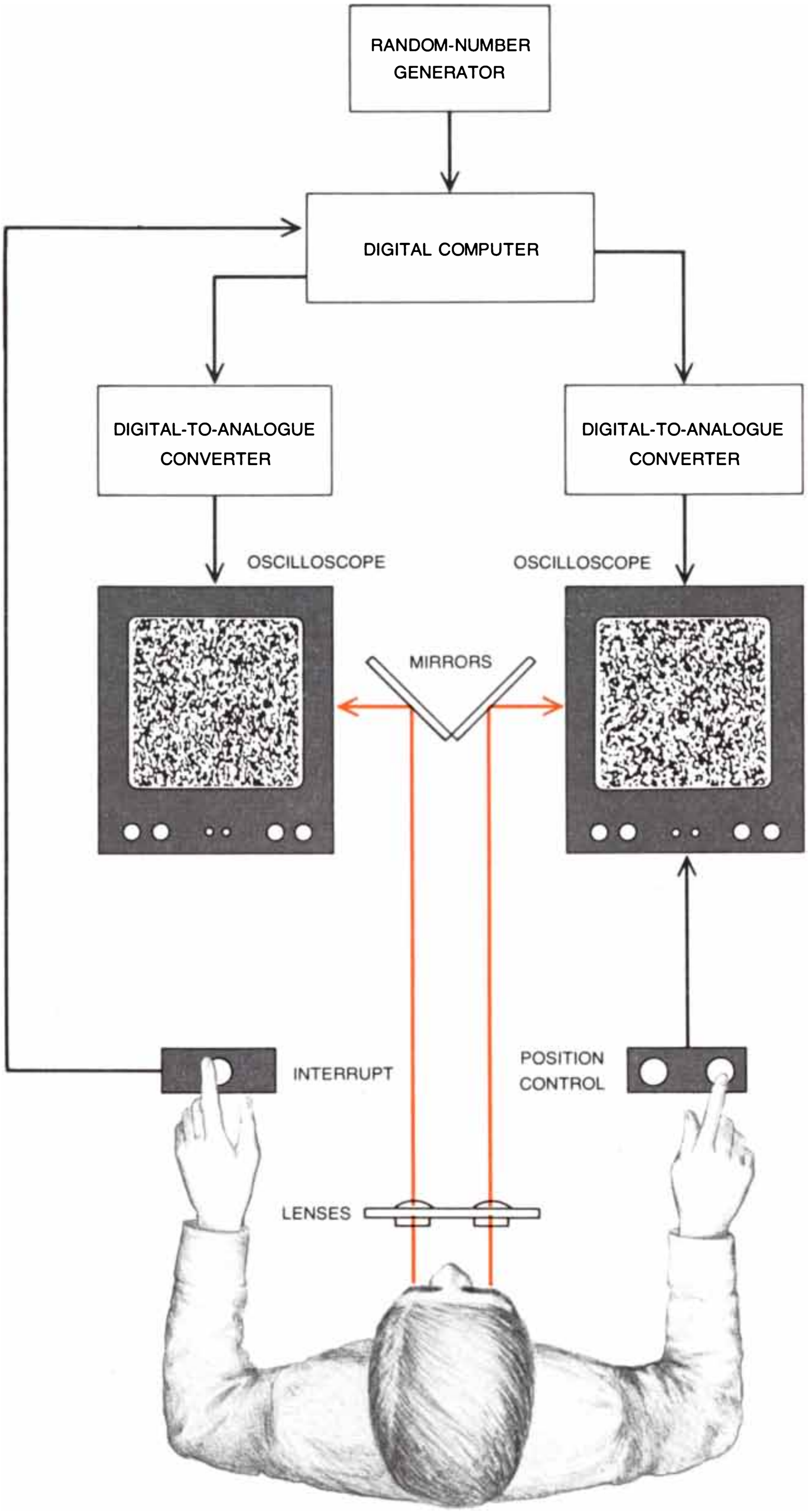
Other investigators, however, did not accept the monocular-combination hypothesis. In 1877 A. L. Austin of Invercargill in New Zealand wrote to Charles Darwin about a curious discovery he had made: "Although a perfect stranger to you, and living on the reverse side of the globe, I have taken the liberty of writing to you on a small discovery I have made in binocular vision in the stereoscope. I find by taking two ordinary carte-de-visite photos of two different persons' faces, the portraits being about the same sizes, and looking about the same direction, and placing them in a stereoscope, the faces blend into one in a most remarkable manner, producing in the case of some ladies' portraits, in every instance, a *decided improvement* in beauty."

Darwin passed the letter on to Francis Galton, who confirmed Austin's observations. Galton also knew of a rather egocentric mathematician who had combined two photographs of himself in a stereoscope, "the one," Galton reported, "taken with an assumed stern expression, the other with a smile, and this combination produced a curious and effective blending of the two." Galton disagreed with the monocular-combination explanation because the binocular perception of two different faces was so unlike a true optical combination of the faces.

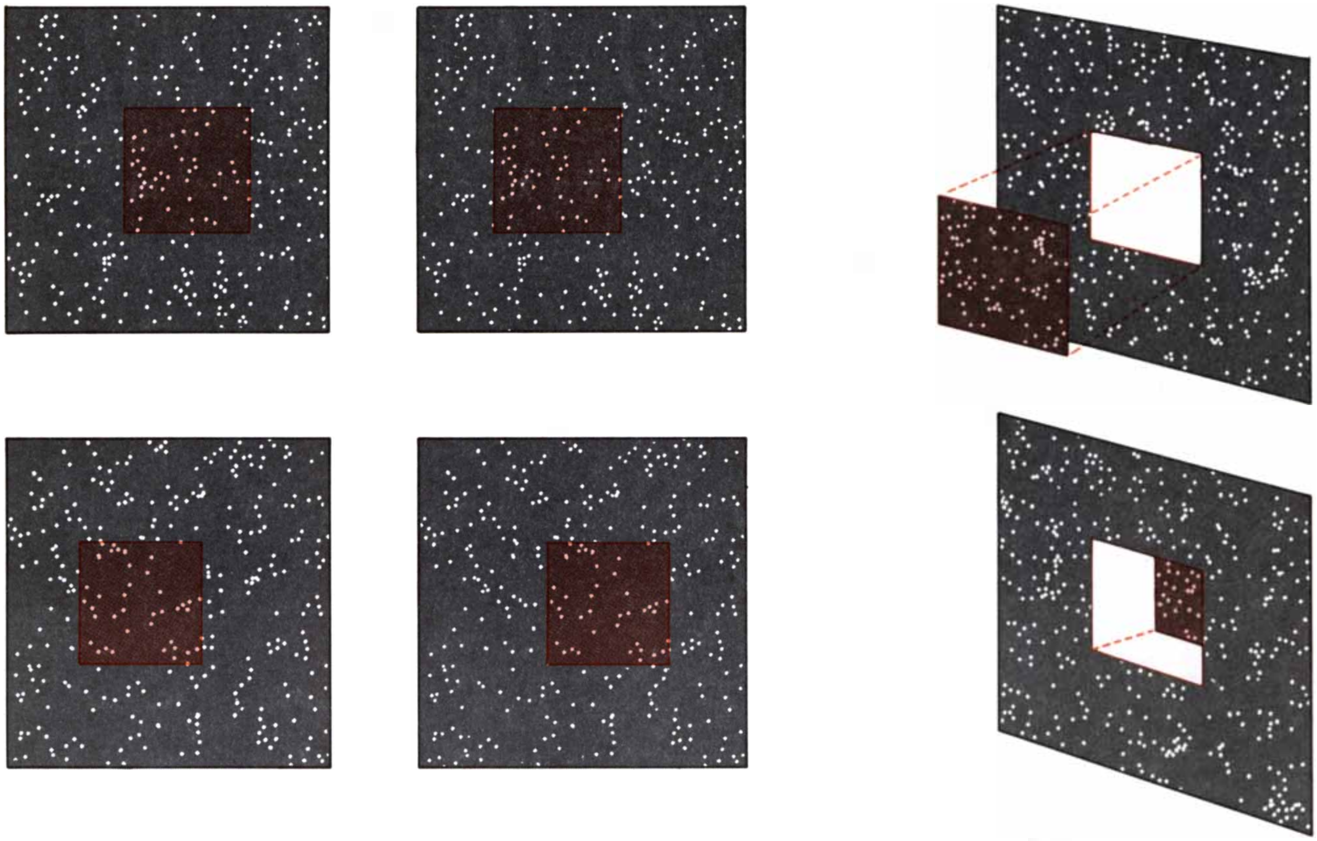
A similar effect is produced when two real faces are viewed by means of mirrors so that one face is seen by the left eye and the other face is seen by the right eye [*see illustration on opposite page*]. The point I want to make here is that binocular combination reveals some critical faculty in the visual system that is capable of making decisions and of rejecting information, apparently on aesthetic grounds.

Just how resourceful binocular perception can be was demonstrated by Bela Julesz of Bell Laboratories with his random-dot stereograms [see "Texture and Visual Perception," by Bela Julesz; SCIENTIFIC AMERICAN, February, 1965]. With the aid of a computer he created patterns of random dots that, when they are viewed with a stereoscope, combine to form three-dimensional scenes. When the same stereograms are viewed monocularly, they appear to be of completely random texture and show no recognizable forms or shapes. Julesz had demonstrated conclusively that it is possible for binocular depth perception to occur in the absence of recognizable monocular cues.

In our laboratory at the University of Western Australia, John H. Hogben and I, working in close collaboration with Monte Sala, an electronics engineer, have developed a computer system that extends Julesz' random-dot-stereogram method. The delivery end of our system consists of a pair



COMPUTER SYSTEM for generating completely random stereograms is shown in schematic form. A random-number generator delivers the coordinates for the points of light that are to be displayed, and a computer adds depth information to the stream of random points. The points of light appear briefly in pairs on the oscilloscopes. For each point on the left screen there is a corresponding point on the right screen. The observer sees thousands of points of light appearing and disappearing at random. In spite of the fact that the scene is continually changing, the visual system is capable of matching the pairs of points on the left and right screens. The observer perceives the displays as a single scene in vivid depth, with objects standing free in space.



DEPTH INFORMATION IS ADDED to random-point patterns by a horizontal shift of the points in the regions to be seen in depth. To create a raised central square, for example, points are shifted to the right in the left image and to the left in the right image (*top*). Where there is an overlap of points the background points are eliminated by the computer. Random points are added to any blank areas created by the shifts. To create a depressed surface, points are shifted to the left in the left image and to the right in the right image (*bottom*).

of point-plotting oscilloscopes that are optically separated so that one is visible to the left eye of an observer and the other is visible to the right. Points of light are displayed briefly in pairs on the oscilloscopes, one point on the left oscilloscope and one on the right. Each eye of the observer views a rapid succession of points of light, each of which is independent of those that went before or those that follow. When the observer shuts one eye, all he sees is a swarm of dots of light apparently in motion over the entire square display area, rather like the "snow" seen on a television receiver.

The sequence of points of light is not made up in advance, nor does it repeat itself. A specially designed random-number generator delivers pairs of coordinates for points on the 256-by-256 grid of the oscilloscope screens. Each point of light is thus generated at random.

Depth information is added to the random stream by a computer that controls the display oscilloscopes. The computer holds a blueprint of the scene to be portrayed. As each random point is intercepted the computer determines the shift, if any, in the location of the point that is to be applied. The principles governing the shifts are those of classical stereoscopy. Points of a near surface are shifted to the right in the left eye's field and to the left in the right eye's field. The reverse is done for points of a far surface. The greater the shift, the greater the perceived depth. To ensure a uniform distribution of points in regions where the shift creates an overlap of points, the background points are eliminated by the computer. In regions where the shift creates a blank the computer adds points of light to fill it.

Consider what the visual system must do in order to create a binocular percept under these circumstances. Thousands of points appear and disappear at random. For each point seen by one eye the other eye must single out the point with which it is paired. In addition a pair of points that have been shifted with respect to each other must be interpreted as a single point, not as different points seen by different eyes. Finally, if a surface is to be seen in depth as a form with definite edges, the visual system must be able to recognize all the points within that form as having a common disparity.

An observer with normal stereoscopic vision, after a brief period of adjustment, sees the output on the oscilloscopes as a scene in depth. With practice the time required to achieve depth perception decreases until it is virtually immediate. To see a scene produced by random-dot stereograms for the first time is an uncanny experience. Objects stand free in space, vivid and sharp-edged. Let me describe a simple case. A square region at the center of the oscilloscope display is shifted by the computer so that on the right oscilloscope points in the central square are shifted to the left and on the left oscilloscope they are shifted to the right. When either display is viewed with only one eye, the observer sees points of light in motion on a dark background. When he views the displays with both eyes, the scene changes dramatically. At the center of the display a square appears to float in front of the background. The square is like a sheet of dark plastic on which points of light move about. Although the pattern of points of light on the surface changes continuously, the square itself presents a solid and unchanging appearance. Binocular perception has constructed the solid square (which is actually not there) in order to account for the information it is receiving from the random-dot stereograms.

Something else about the binocular scene strikes the observer after a moment's reflection: the points of light within the small square are trapped on its surface. Remember that when the display is viewed with one eye, the points of light move all over the display area quite uniformly. Now in the binocular view some points are lifted out of the background and never cross the border of their new territory. Instead they seem to bounce back from the edges of the floating square. What this means is that the visual system attaches functional significance to the edges constructed by the binocular process, and that motion across such edges from one depth level to another is prohibited. Furthermore, when monocular perception conflicts with binocular constructs, the monocular percepts are suppressed.

In our experiments we have not found an upper limit to the rate at which binocular perception can handle the input of random pairs of points. Our computer system can display up to 30,000 pairs of points per second, and that rate is handled easily by the visual system. We have an optical system that can deliver points at rates of up to 250,000 pairs per second, and even at that rate binocular perception shows no sign of being deluged with more information than it can handle. There must of course be some upper limit, since when the rate of delivery is sufficiently high, the displays will be flooded with light. Even then the limitation may be at the level of the receptors in the retina rather than at the level of the information-processing capacities of the mechanisms that serve to compare the input from the two eyes.

There is, however, a lower limit below which information is received at too low a rate to sustain the perception of forms in depth. That limit varies with the complexity of the scene presented and with the size of the detail in the scene. For single shapes of a reasonable size the lower limit is about 2,000 pairs of points per second. Below that rate points can appear to be at different depths, but form is lost and there are no definite edges.

Binocular perception is capable of constructing very complex scenes from the random-dot stereograms, provided that the rate of delivery of information is high enough to maintain the scene. A sequence of 10 descending steps (in effect 10 different depth planes) can easily be perceived. It is even possible for the computer to present targets that are seen as slanted surfaces.

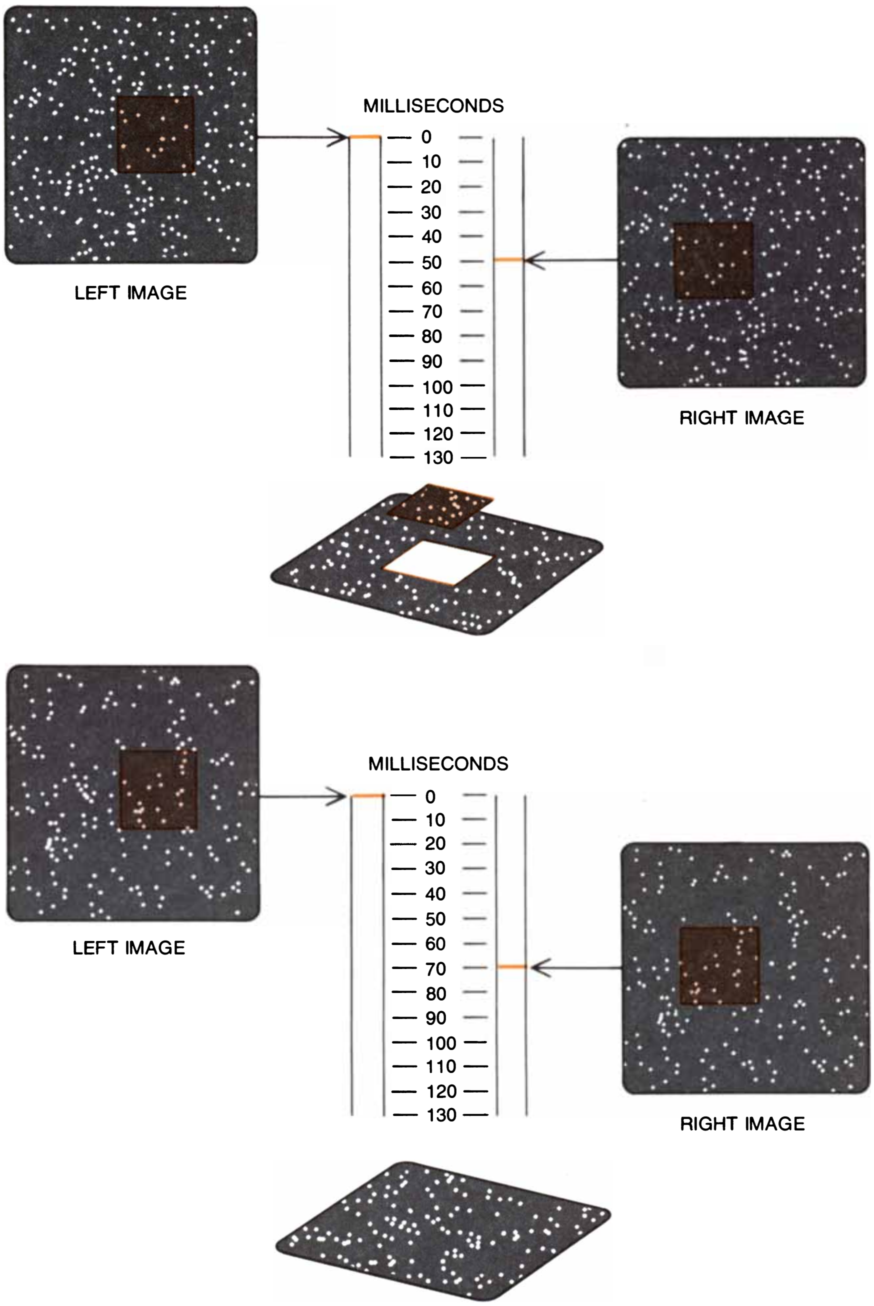
When the computer's blueprint specifies two surfaces in one region of the display, the result might be expected to be chaos, but it is not. Half of the points of light are assigned to the upper surface and half to the lower. The problem of making both surfaces visible is elegantly solved by the visual system: the surface in front appears to be transparent and the one behind opaque. This shows that there is an intuitive grasp of spatial relations built into binocular perception that transcends mere geometry.

As I have mentioned, the constructs of binocular perception control the apparent motion that is perceived by prohibiting movement across an edge between regions of different depths. Is it possible, we wondered, to cause a target to move toward or away from an observer? We modified the computer blueprint so that the disparity of the points on the surface of the target changed continuously. Provided that the change is neither too fast nor too slow, the observer does perceive the target moving without any loss in its structure. The target holds its form and its edges remain solid, retaining their functional control over the apparent motion of the points within the target. There is no disruption of depth perception as the target moves back and forth.

There is a limit to the amount of spatial disparity that can be introduced into stereograms. Too large a disparity gives rise to double images, although not necessarily to a complete loss of the sense of depth. Hogben and I set out to determine whether there might be a limitation in time as well as in space. In other words, could we present a train of points to one eye and the paired train of points to the other eye a short time later and still obtain binocular fusion?

To do this we imposed a delay on the train of random points destined for one of the eyes. The points destined for the left eye were presented immediately but the points for the right eye were put into a delay loop inside the computer so that they were presented at a later time. Both oscilloscopes received points at the same rate and both displays looked identical when they were viewed monocularly, but the stream of points on one display were behind the other, as if they had started later.

Consider the problem with which binocular perception is now confronted by the random-dot stereograms. Two points arrive at the same time, one to each eye, but they are completely independent of each other and therefore carry no disparity information. New points arrive at the rate of 10



TIME LIMITATIONS OF BINOCULAR PERCEPTION were studied by immediately presenting the stream of points to the left eye and delaying the stream to the right eye. When the delay between left image and right image is less than 50 milliseconds (*top*), the targets are seen in depth. When the delay exceeds 50 milliseconds, the impression of depth collapses (*bottom*).

per millisecond. It would be impossible to match pairs of points unless a record of the thousands of points seen is stored in some way. What is more, the record must be extremely accurate if it is to be utilized to find the exact disparity between pairs of points.

We found that binocular perception can tolerate a delay of about 50 milliseconds (a twentieth of a second) but not a longer one. After some practice observers can see targets in depth and identify them as long as the delay between points is less than 50 milliseconds. If it is longer by even a few milliseconds, the impression of depth collapses and the targets can no longer be seen. The tolerable delay varies a little above or below 50 milliseconds for different observers, but it is quite constant for a given observer.

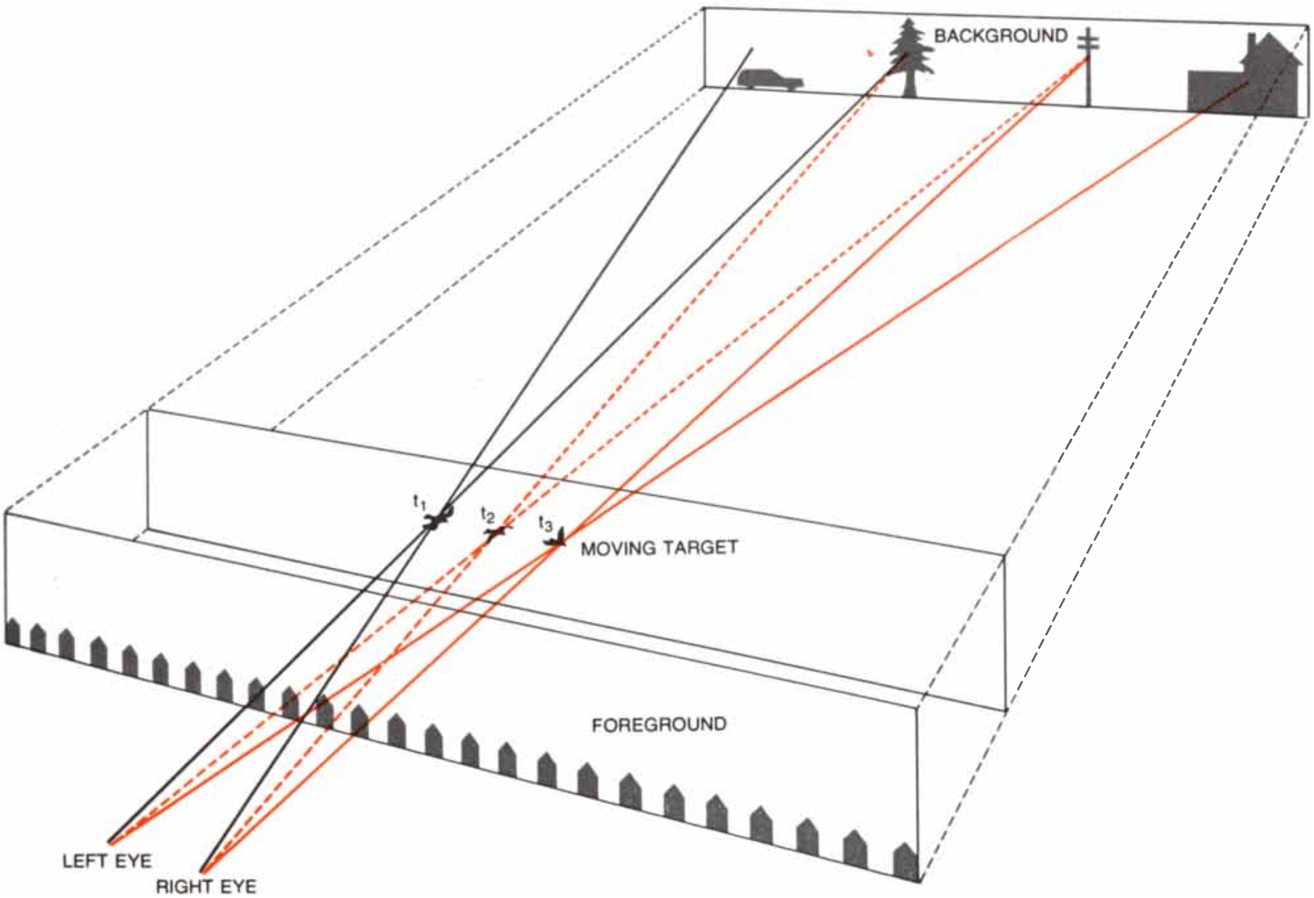
Two important conclusions can be drawn from the results of this experiment. The first is that binocular perception is subject to a measurable limitation in time as well as in space. Disparity information must arrive within an interval of 50 milliseconds, just as the spatial disparity must not exceed a certain distance. The second point is that binocular perception must incorporate some form of visual memory capable of maintaining a fine positional record of thousands of points for at least 50 milliseconds.

Under the conditions of our experiment, however, if a point of light is seen at all, it remains visible for about 130 milliseconds. That visibility time can be established in various ways, for example by simply counting all the points visible at any instant on a small portion of the screen. The problem, then, is to explain why the time limit for binocular comparison is shorter than the time the point of light remains visible. If a point of light can still be seen by one eye, why can it not be paired with another point when the other point becomes visible to the other eye? Before we consider this paradox let us consider the role of binocular perception in tracking laterally moving objects.

When we fix our eyes on a target, the target forms images in identical positions on the retinas of both eyes. Objects farther away than or closer than the target form images in different positions on each retina, giving rise to the disparity that is the basis for stereoscopic depth perception. Therefore at any time each eye has a different image of the background and the foreground but an almost identical image of the target. If the target moves and the eyes follow it, each eye maintains an identical image of the target, but now each eye encounters the same elements of the background and foreground at different times [*see illustration below*]. If the eyes track a target moving from left to right, the left eye will encounter each part of the background before the right eye encounters it. The situation is the reverse for the foreground: the right eye will encounter each part of the foreground before the left eye encounters it.

It occurred to me that the visual input to the eyes from a moving target that is being tracked could be mimicked with the random-dot stereograms. In addition the stereograms would eliminate any monocular cues about the target, the foreground and the background. In order to determine whether or not time differences encountered in tracking a moving target play a role in binocular perception, our computer was programmed to display the target at exactly the same time on both screens. Points outside the target were displayed immediately to one eye but were delayed for a brief time before being displayed to the other eye [*see illustration on opposite page*]. There is no spatial disparity in the stereograms. All points are plotted in exactly the same position on both display screens.

The observer sees the target in depth, which is not surprising since the target area



TRACKING A MOVING OBJECT creates time differences in the visual input to the two eyes. When the eyes converge on a target, say a bird, moving from left to right, the left eye will see the elements in the background before the right eye does. For example, at t_1 the left eye will see the tree in the background but the right eye will not encounter the tree until t_2. The left eye meanwhile has moved ahead and is aimed toward the telephone pole. At t_3 the right eye encounters the telephone pole and the left eye sees the house. The situation is the reverse for the foreground. As the eyes follow the moving bird the right eye leads and the left eye encounters the same elements later.

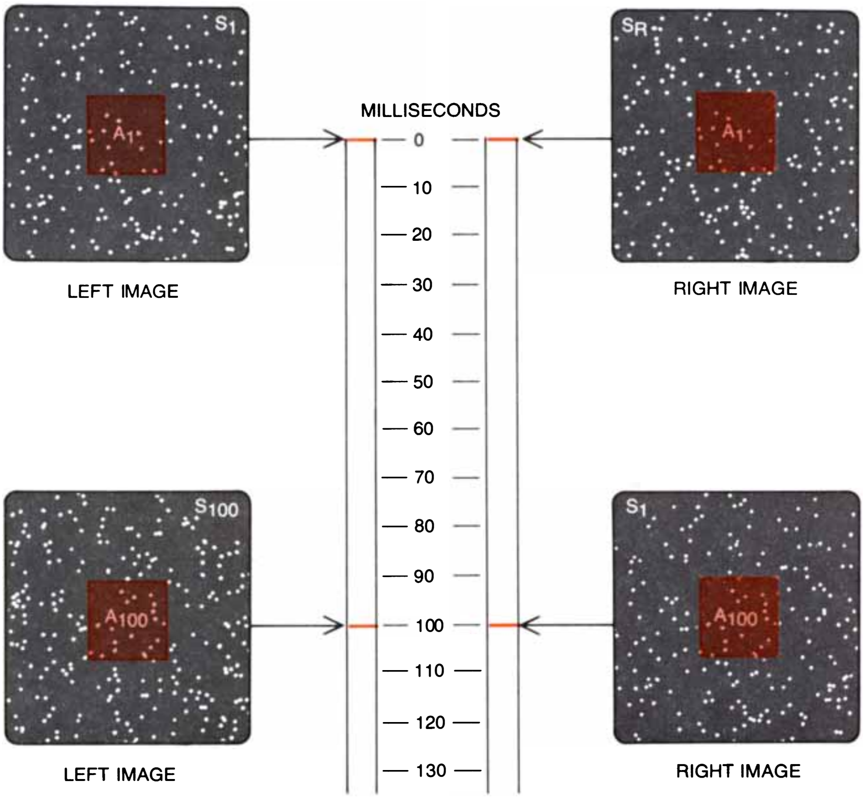
is shown simultaneously to both eyes and the background is not. What is surprising is that the area surrounding the target is perceived as both foreground and background, the foreground moving in one direction and the background moving in the opposite direction.

If the points of light from the surrounding area reach the left eye first, the direction of the background motion is from right to left, as if the target were moving from left to right across the background. If the points reach the right eye first, the background motion is from left to right. In both cases the foreground motion is the reverse of the background motion. In fact, foreground and background may combine to give a strong impression of an upright cylinder that is rotating around its vertical axis, with the target in the middle of the cylinder.

It is possible to get a fair impression of the effect at home by tuning a television receiver to an empty channel and adjusting the contrast controls so that the "snow," or contrast noise, is readily visible. Place a dark filter or one lens of a pair of sunglasses in front of one eye and look at the screen with both eyes. The filter delays the information received by that eye, and you will be able to see the rotating-cylinder effect. The snow will stream in one direction in front of the screen and in the opposite direction behind it. When the filter is moved to the other eye, the direction of the streams will be reversed.

In our experiment there were two conflicting indications of depth: disparity consistent with a static depth scene, and delay consistent with streaming. At delays below 50 milliseconds the static scene was observed, but at delays above 70 milliseconds it yielded to the streaming effect. Delays between 50 and 70 milliseconds caused the scene to shift back and forth from static to streaming.

The results show that binocular perception can attune itself to different sources of information in the one situation. For delays below 50 milliseconds spatial disparity holds sway. For delays of 70 milliseconds and, we think, up to two seconds, time differences result in the perception of streaming in depth.



TIME DIFFERENCES in visual input that are encountered when the eyes track a moving object can be mimicked with dynamic random-dot stereograms. Points of light in the central target area are displayed simultaneously to both eyes, but the right eye sees the points in the surrounding area after a brief delay. For example, the left eye sees A_1 and S_1 while the right eye sees A_1 and a different surrounding texture (S_R). One hundred milliseconds later the left eye sees A_{100} and S_{100} while the right eye sees A_{100} and S_1. There is no spatial disparity: both eyes see all points in exactly the same position. The time delay between left and right eyes, however, creates an effect of depth and motion, which is illustrated on the next page.

Under other circumstances delays of less than a millisecond, as little as 160 microseconds, are enough to provide a clear indication of depth. David Burr and I have shown this to be true with stroboscopic sequences that give the illusion of smooth motion. Each eye views the same sequence, but there is a phase difference between the sequences as observed by the two eyes. The phase difference is sensed and interpreted as an indication of depth.

We now believe that vision has two equally powerful ways of dealing with objects in motion. We can sense the instantaneous spatial disparity to determine depth, or we can sense the phase difference with which objects pass common reference points in the two eyes. That is to say, vision can adopt one of two alternative perceptual attitudes toward the information being received. In the real world the percepts resulting from the different attitudes are usually at least consistent with each other, but under experimental conditions such as ours the scene may shift with the attitude.

Let us return to the paradox of why a point of light that is visible for 130 milliseconds cannot be paired with its partner when the latter is delayed for more than 50 milliseconds. Apparently in that case we can see something but cannot utilize information from it. On the other hand, when longer time delays cause streaming to be perceived, we are obviously making use of visual information about a collection of points after they have disappeared from view. This means that some kind of visual record must be maintained for more than 130 milliseconds. In other words, when lateral motion is being tracked, records of visual input are maintained beyond the time limit of visibility.

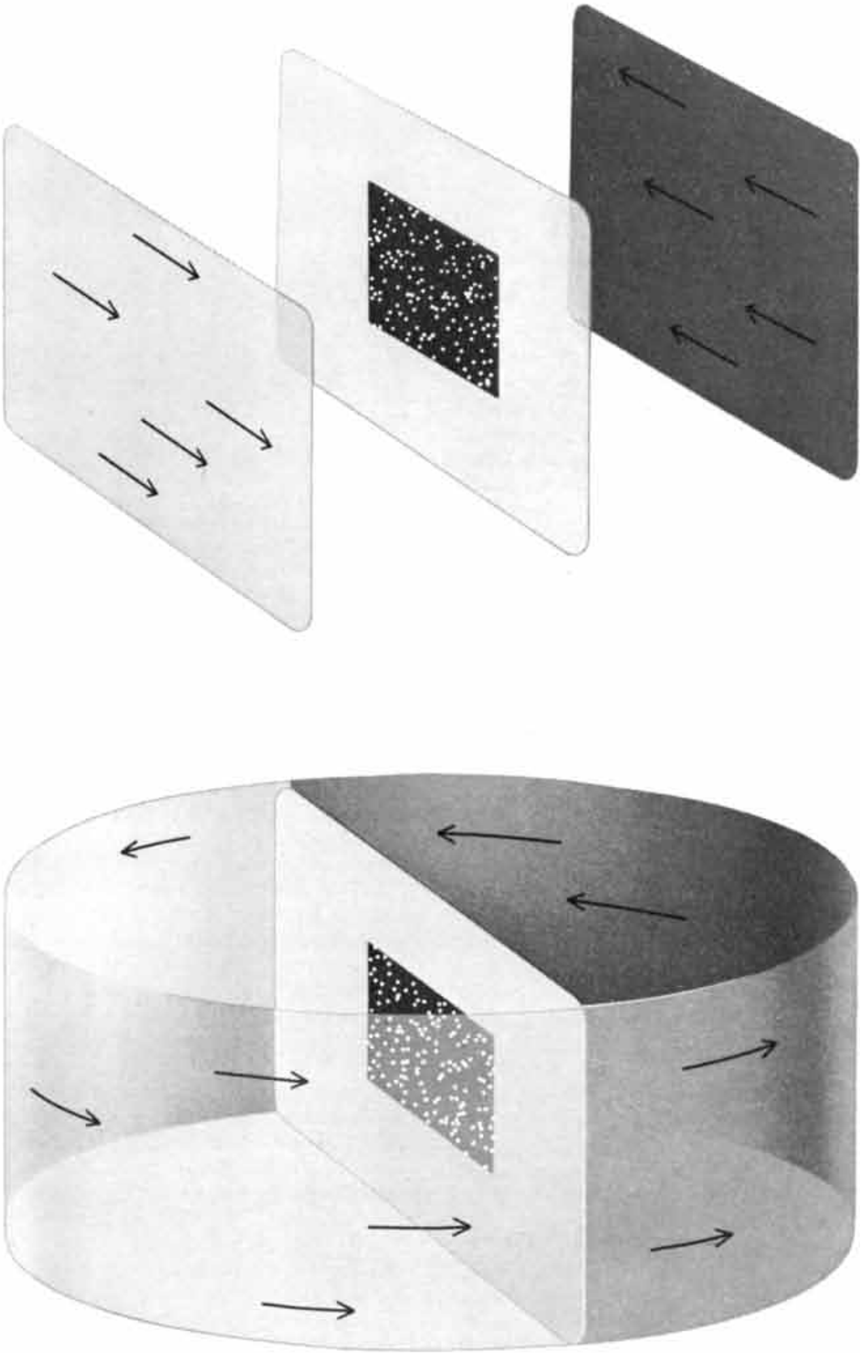
I believe we must conclude that binocular perception has access to records of visual input that are independent of what we see. This represents a radical break with the commonsense view that what we see constitutes all the sense data from which higher perceptual processes develop a conception of the scene before our eyes. It appears there are records of visual input that can be consulted before anything at all is seen in order to determine the proper framework for perception.

When we look at a random stereogram, each eye views a separate stereogram but we do not "see" two stereograms. Our binocular percept is the result of the space and time differences in the stereograms. The subtle discrepancies between the two displays must be recorded separately and very accurately. What these records are, how many separate repositories there are and what their physical basis is we do not know. A number of puzzling visual effects such as apparent motion, visual masking and the perception of symmetry can be brought together by the assumption that perception must wait on the analysis of independent visual records before we are able to perceive.

Some of the skills of binocular perception, for example aligning two images that are different in size or recognizing an entire region as having a common disparity value, are elegantly explained by the spring-coupled dipole model of stereoscopic depth perception proposed by Julesz. But other skills of selection, analysis and synthesis, such as the perception of an idealized face, and of sensing time differences still remain unexplained.

There may be a clue in the fact that

shapes made visible by the binocular combination of random-dot patterns are more idealized than real shapes. Square targets are more perfectly square, with more perfect edges than any real square. They are like the Platonic form of a square. What we observe in random-dot stereograms may well be idealized conceptions, imposed on the external flow of information by something within the visual system. What we observe may be a structure built by vision when it adjusts itself as it tunes in to the sources of the information it can sense. Just as a computer has a program, so may the visual system have a program of arrangements for shapes in space and time. What we see is an interpretation of the external world, ordered within a framework the visual system imposes because of the attitude it adopts. In other words, we adopt a perceptual attitude in order to comprehend the world.



DEPTH AND MOTION are perceived when time delays are introduced into random-dot stereograms in the manner depicted in the illustration on the preceding page. The target is seen as a solid object. The surrounding area, however, is perceived as both foreground and background (*top*). Points of light in the foreground move in one direction and points in the background move in the opposite direction, making it appear that target is moving from left to right. Foreground and background may combine to give impression of a cylinder rotating around target (*bottom*).

It comes as close to speaking for itself as any car ever built.

Aggressive, unmistakable profile. Strong, mellow, no-nonsense sound. 280-Z, by Datsun.

Its particulars bear out its promise. Under the sculpted hood, an efficient 2800cc overhead cam engine with electronic fuel injection. The 6-cylinder plant has an 8.3:1 compression ratio for optimum performance. The transmission is an all-synchro manual 4-speed; an automatic is available.

Perfection. Fuel injection.

Additional GT requisites include rack and pinion steering. Power assisted front disc brakes. 195/70 HR 14 steel belted radials. Reclining bucket seats. AM/FM radio with power antenna and electric rear window defogger standard, of course.

The fuel injected Datsun 280-Z. 2-seater or 4-seater. Commune with one.

The fuel injected
Datsun 280-Z

The Small Electronic Calculator

Pocket-sized calculators are based on a single microelectronic "chip." The chip circuits and their associated components form an information-processing system of considerable sophistication

by Eugene W. McWhorter

In the five years since pocket-sized electronic calculators came on the mass market millions of people have had the delightful experience of seeing the answer to a calculation such as 953.22 times 14.331 (and much more complicated calculations) light up in the display the instant they pushed the "equals" key. It seems safe to say, however, that few of those who have enjoyed this substantial acceleration and expansion of their calculating abilities have had more than a general idea of what was going on inside the calculator. It is well known that the works of a small electronic calculator are based on a tiny microelectronic "chip." But how are the logical organization and the numerical routines of a small electronic calculator built into those works?

In terms of both history and operating principles the small electronic calculator is a direct outgrowth of electronic "office" and "professional desk-top" calculators. Some 10 years ago a typical four-function (addition, subtraction, multiplication and division) calculator of that type incorporated hundreds of separate microelectronic integrated circuits and cost many hundreds of dollars. Visionary workers in the electronics industry, however, were looking to the day when the same concepts could be incorporated in machines that would be small, simple and inexpensive enough to gain a mass market. As early as 1965 work was in progress at Texas Instruments Incorporated on an experimental four-function pocket calculator based on a single integrated circuit. A patent on that device dates from 1967.

The history of semiconductor technology has been marked by regular and even predictable increases in the complexity of integrated circuits, coupled with reductions in cost for virtually all semiconductor products, including the light-emitting diodes that in most calculators display the numbers. This evolutionary pattern is often interpreted in terms of a "learning curve" based on the experience gained from a steadily rising volume of production. In any event it became possible by about 1970 to put all of a basic calculator's logic on a single metal-oxide-semiconductor integrated-circuit chip priced at less than $100. Within a year that technology gave rise to a new generation of small four-function calculators. A typical machine had a single integrated circuit for all calculating functions; it also had certain supporting circuits.

Some of the new calculators were pocket-sized, battery-powered models; others were compact desk-top machines that operated on alternating current. The important thing was that the prices (well under $200) gave the devices a handhold at the upper end of the mass market. Now that the industry is about 100 million units farther along the learning curve the low end of the price range is well below $20. The result is that for millions of people everyday arithmetic will never be the same again.

Small calculators are evolving so rapidly and are so varied in sophistication that no one explanation of how they work will fit all calculators. It is nonetheless possible to explain the principles of how they work on the basis of a hypothetical four-function calculator built around a typical real chip. The chip is not the last word in microelectronic devices, but it is sufficiently representative of integrated circuits that are still in service throughout the industry. It is a metal-oxide-semiconductor chip of the large-scale-integrated type, meaning that it carries on a surface that is about five millimeters, or .2 inch, on a side thousands of components such as transistors, resistors and diodes. It has 28 pins, or terminals, which is a standard in the industry. Power is supplied to it at plus and minus seven volts. Its operations are synchronized by a 250-kilohertz clock signal, that is, a regular time pulse running at a rate of 250,000 cycles per second. Since similar chips have been described in detail in this magazine [see "Metal-Oxide-Semiconductor Technology," by William C. Hittinger; SCIENTIFIC AMERICAN, August, 1973], I shall not dwell on the physical structure of the chip.

In addition to the chip the calculator has several other components, some visible to the user and some not. The most conspicuous visible component is the keyboard for entering numbers and instructions. Among the other components are the display system, the oscillator circuit that generates the clock signal, a voltage regulator for the power supply, a pack of rechargeable batteries and a plastic case.

The 28 pins of the chip may seem like too few to receive and transmit all the information the chip can handle. Why they are enough can best be explained by beginning with the display system of the calculator. If one looks closely at the numeral 8 when it is illuminated in the display, one sees that it is made up of seven segments: three forming the upper part of the 8, three the lower part and one the cross-link. Any other numeral from 0 through 9 can be formed with fewer than seven such segments, as can a minus sign and any of several symbols needed to indicate error or overflow [*see illustration on pages 92 and 93*]. Each segment is an elongated light-emitting diode, a tiny piece of semiconductor material with two terminals. An eighth diode forms the decimal point that can be illuminated to the right of each digit.

Our hypothetical calculator has a row of nine such arrays of eight light-emitting diodes, 72 diodes in all. The eight arrays beginning at the right end of the row accommodate the eight-digit display of which the calculator is capable; the ninth array, at the left end of the row, is for the minus sign, the decimal point farthest to the left and the symbols for error and overflow. A diode lights up when both of its terminals are connected to the power supply, with the cathode being more positive than the anode.

The 28 pins of the chip must also serve as inputs to the chip from the keyboard. The calculator has 18 push-button switches and two slide switches (not counting the switch that turns the calculator on and off). Each has two terminals. Pressing a key on the keyboard holds the corresponding switch closed.

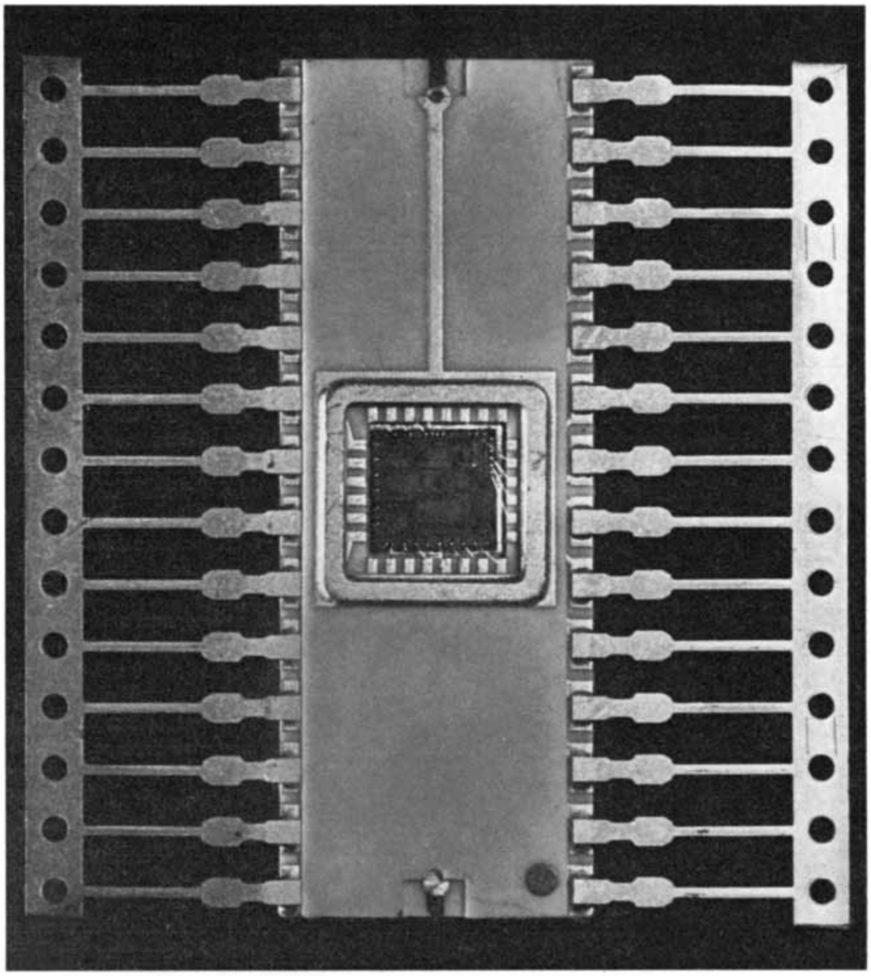
The problem of connecting 72 light-emitting diodes and 20 switches to the chip through 28 pins is solved as follows. The keyboard is actually connected to the chip through only four input terminals, which can be designated *N, O, P* and *Q*. The light-emitting diode segments are connected through eight output terminals, which can be designated by the letters *a* through *h*. At any given instant the chip interacts with no more than one switch for each keyboard-input line and illuminates no more than one of the nine diode arrays. This is made possible by the provision of 11 scan lines, which can be designated by the numerals *1* through *11*. They originate at the chip and lead to both the keyboard and the digit display.

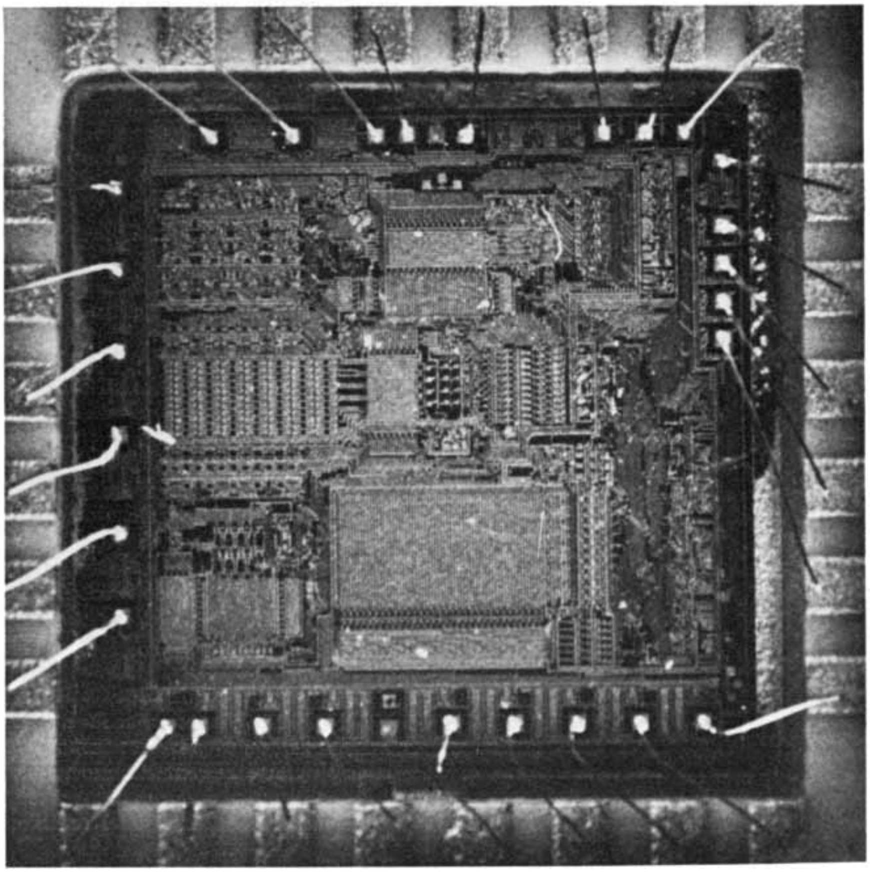
A timing mechanism in the chip, synchronized by the signal from the main clock, switches one scan line at a time from a normal voltage of zero to a higher voltage and back again. The voltage pulse lasts for only 132 microseconds (33 cycles of the main clock). There is a six-cycle delay between the end of one scan pulse and the beginning of the next, so that a complete scanning cycle (during which each scan line is pulsed once) takes 429 clock cycles, or about 1.7 milliseconds.

How do the scan lines make it possible to control 72 light-emitting diodes with eight segment outputs? Each of nine scan lines (*1* through *8* plus *11*) provides the input to an inverting current-booster circuit that is external to the main chip. The circuit is termed a digit-driver. Each such driver, while it is activated by a scan pulse, provides a high-capacity current path to ground from all eight of the anodes in one array of light-emitting diodes. Thus a digit position can be illuminated only during the time in each scanning cycle when the corresponding scan line is activated.

The particular digit that is displayed at a given time in one display position or another is determined by which of the eight segment-output terminals are activated during a scan pulse. Each segment output controls an external current-booster circuit called a segment-driver. Each segment-driver is connnected to the cathode of the light-emitting diode in the same position in all nine digit displays. When a segment-driver is activated, it connects its corresponding cathodes to the positive power supply. Current from this particular segment-driver then flows through the only light-emitting diode whose anode is connected to ground through a digit-driver.

An example of what results from this arrangement can be followed by supposing that scan line *2* and outputs *abcdgh* are active. The display (for 132 microseconds) is "3." in the second position from the right. Then this scan line and digit display go off for 24 microseconds while the segment outputs change to the pattern required for the next digit to the right. Suppose the pattern





MICROELECTRONIC CHIP performs the calculating functions of a pocket calculator. At top the square chip, measuring five millimeters (.2 inch) on a side, is shown exposed in its ceramic housing, which has been sawed away to reveal the chip. Associated with the chip are 28 pins, or terminals, which connect it to other parts of the calculator. At bottom is a micrograph in which the chip is enlarged 17 diameters. The chip and its housing, made by Texas Instruments Incorporated, constitute an integrated circuit with thousands of electronic functions. Among the structural features in the micrograph is the read-only memory, rectangular area at bottom center.

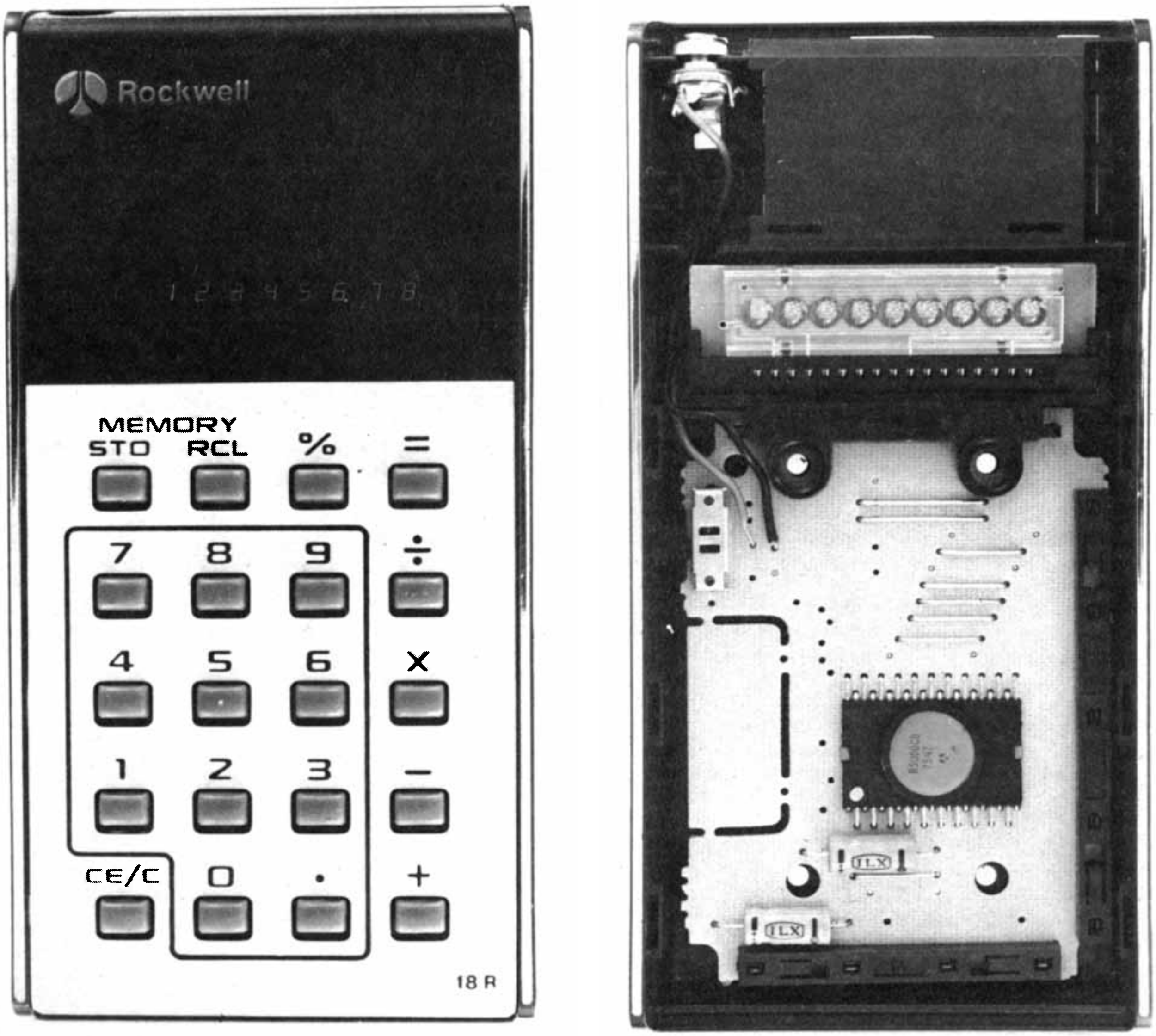
is *abcdef*. As a result "0" appears in the first position on the right for 132 microseconds (while scan line *1* is active). Although each of these digits is on for only about 8 percent of the time, the flicker is so rapid that the eye sees an unwavering "3.0" in the last two positions on the right.

So much for the display. How do the scan lines make it possible for signals from 20 keyboard switches to go into only four keyboard-input terminals on the chip? Each switch provides the only possible connection between a given scan line and a given input line. At any given instant input signals can be received only from the switches connected to the scan line that happens to be activated at that instant. (For scan line *1* they are the chain/constant switch, the decimal-selector switch, the "multiply" key and the "1" key.) Any switch (or more than one) in this group that happens to be closed while the particular scan line is being pulsed transmits the pulse to the corresponding input line (or lines).

On receiving a pulse in a keyboard-input line the main chip notes which scan line is currently active and thereby deciphers which switches of the several that could possibly be closed along this input line are closed. Thus during a full scanning cycle the calculator "looks at" each switch once. In this way signals from 20 keyboard switches are time-multiplexed into only four inputs to the main chip.

Even when the calculator is merely displaying a number and waiting for the next keystroke, the main chip is active, exercising its "idle routine" in time with the pulses of the clock signal. In this routine the chip continuously scans the keyboard and the display while its segment outputs twinkle on and off in a precisely changing pattern to create the apparently steady display.

In addition to maintaining the display during the idle routine the chip must deal with inputs from the keyboard. During the idle routine it watches for a signal on either the *N* input (which connects to all 10 number keys) or the *O* input (connected to all the operation keys). On detecting a pulse in one of these lines the chip must first verify that the signal is not merely random "noise." The calculator takes care of such noise by not acting on an input in the *N* or *O* line until it has verified that the signal is still there at the end of the next scanning cycle. In addition, if a scan line shows that both a number key and an operation key are closed, the chip always deals only with the operation signal.



ELECTRONIC CALCULATOR made by Rockwell International is shown at left as it appears to the user and at right with its top removed. Among the objects visible in the internal view are the adapter jack at upper left, below it a horizontal row of light-emitting diodes that illuminate the calculator's display, the on-off switch (*left of wires*) and (*lower right*) the rectangular housing containing the chip.

While the chip is acting on a number signal or an operation signal, it ignores further inputs in the N and O lines. It consults the P and Q lines as necessary, however, to ascertain the existence of instructions relating to the position of the decimal point and to the type of calculation (chain or constant).

It is possible for the chip to complete an operation even before the operator releases the key that triggered the operation. This means that the same signal must not be acted on more than once. Hence before the chip resumes the idle routine it verifies that all keys have been released.

It will be evident from the foregoing that the calculator does a good deal more than just add, subtract, multiply and divide. It is in fact a multipurpose information-processing system of considerable sophistication. Keying and display are important, but what really makes the hypothetical calculator work are the architecture, or logical organization, and the algorithms, or numerical routines. The principal electronic subsystems built into the main chip are depicted schematically in the illustration on page 94. In what follows I shall describe what most of the subsystems do and then trace a simple calculation as they carry it out.

The functional heart of the system is the adder-subtracter, which is usually called simply the adder. It operates on numbers that have been put into a binary code. In the pure binary number code each bit, or binary digit, represents a power of 2. A binary number such as 10010 is most easily read from the right, where the first bit represents the number of 1's, the second the number of 2's, the third the number of 4's and so on. The binary notation 10010 therefore encodes a decimal-system number with no 1's, one 2, no 4's, no 8's and one 16. To put it another way, the encoded number is $0 + 2 + 0 + 0 + 16$, or 18.

The hypothetical calculator employs a variant of this system known as binary-coded-decimal notation. In the BCD system each digit of a decimal number is represented by a four-bit pure binary code. Thus the decimal number 97 would be 1001 0111. A 1 is represented by a higher voltage and a 0 by a lower voltage; there are no intermediate voltages in the system.

In a typical adding or subtracting operation two such binary-coded-decimal integers are presented to the adder, one pair of four-bit digits at a time, beginning with the least significant digit. For example, if one were adding 56 and 43, which would be encoded as 0101 0110 and 0100 0011, the right-hand digits (0110 and 0011) would be presented to the adder first.

The basic building blocks of digital electronic systems are the circuits called logic gates, with such designations as AND, OR, NAND, NOR and inverter gates. The output of an AND gate is 1 if, and only if, all inputs to it are 1. The output of an OR gate is 1 if at least one input is 1. The output of an inverter gate is the opposite of its input. A NAND gate acts as an AND gate followed by an inverter gate. Similarly, a NOR gate acts as an OR gate followed by an inverter gate.

Virtually any arithmetical or logical operation can be performed by an appropriately interconnected combination of AND, OR and invert functions. In the adder subsystem various combinations of gates add two bits of corresponding significance and a third bit carried over from adding the next least significant pair of bits. The result is one "sum" bit and one "carry" bit. Several such circuits in what is called a ripple-carry configuration add the multiple-bit groups that constitute binary-coded-decimal digits.

Another basic building block in calculator circuits is the flip-flop, so named because its output can be flipped or flopped to one logic state or the other by various input signals. Flip-flops of various kinds provide one form of cell for the temporary storage of information. The "latch" shown with the output decoder in the illustration of the chip's logical design on page 94 consists of four flip-flops of a certain type. Another form of cell for the temporary storage of information is the dynamic memory cell, which holds a bit by briefly storing an electric charge.

A register consists of several memory cells that always function together in a fixed close association. In a shift register bits are synchronously shifted in bucket-brigade fashion through a series of memory cells under the command of a common signal. Several shift registers can be operated synchronously in parallel.

The main clock signal is a square-wave signal of constant frequency common to all parts of the calculator. Its pulses serve both to synchronize the activity of the various subsystems and to ensure that each new set of output signals from a given subsystem has settled down to a valid state before the signals are accepted by other subsystems. Like many calculator systems, this one has in addition to the main clock network three others working in association. The signals in the additional networks are called phases. An internal mechanism pulses the phases one at a time in a repetitive cycle (in step with the main clock), providing a convenient source of stepped sequencing control for certain subsystems that must perform their functions in sequential steps.

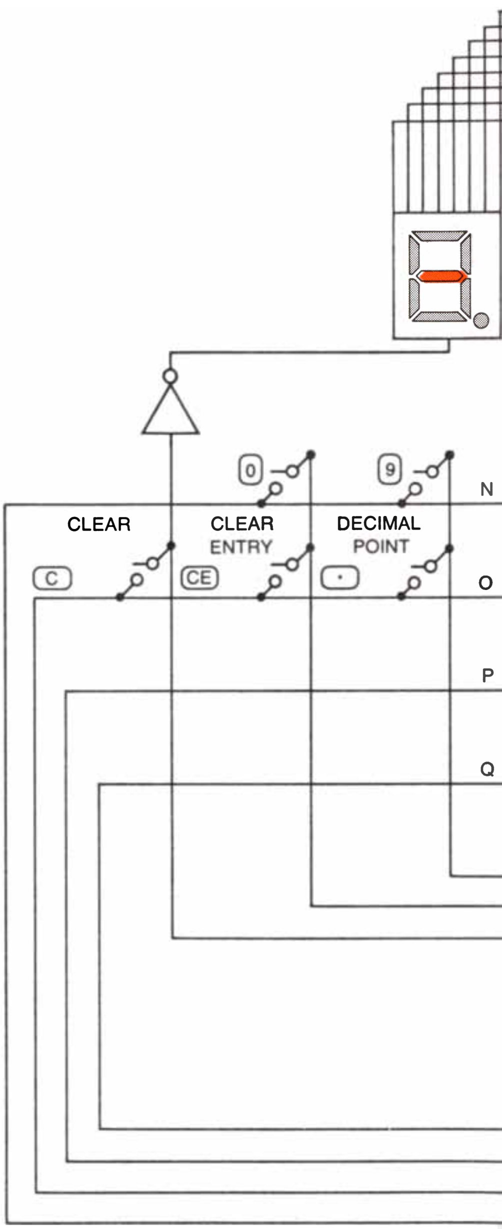
As the calculator adds or subtracts, working on one pair of four-bit digits at a time, the adder also accepts another bit as a carry (or a borrow) from previously added (or subtracted) digits. It produces a binary-coded-decimal digit that represents a sum or a difference, meanwhile storing a carry bit or a borrow bit in a dynamic storage cell within the circuitry of the adder. In adding, say, 835 and 974 [*see illustration on page 98*] the calculator works on each pair of decimal digits (the 5 and 4 first in this example) with two stages of standard ripple-carry addition. Subtraction is performed by first converting the subtracted digit into "1's-complement" form (inverting all four digits, so that 1101 becomes 0010) and then adding. Multiplication and division are performed by programmed sequences of addition or subtraction.

The two numbers put into the adder for a given operation are typically taken from the display register and the operand register, and the result is put into the accumulator register. All three registers are 4-by-13-bit shift registers. The routing of the numbers is accomplished by circuitry within the "control" block. Under most conditions the routing circuitry maintains the data in the three registers by mutually synchronous recirculation. The three registers are in fact integral parts (along with what is called the flag register) of a single 14-by-13-bit dynamic shift register that continuously makes one shift every three pulses of the main clock. The interval is counted out by the three-phase clock generator and is described as one state time [*see top illustration on page 96*]. Thirteen consecutive state times (one for each digit of storage in a 4-by-13-bit register) define a longer period called an instruction cycle. During one instruction cycle, therefore, the contents of a register normally complete precisely one recirculation.

Routing paths are also provided within the control block for exchanging numbers between registers as they are recirculated and for replacing one of the recirculating numbers with the output from the adder. Furthermore, certain of the routing paths include an extension or a delay of either one or two state times (three or six clock pulses). By this means the routing and delay circuits can impose a longer or a shorter delay on the recirculation of one register than is imposed on the others, so that the number contained in this register can be shifted in one direction or the other with respect to the other registers. Finally, one input to the routing circuits is for a digit synthesized by other circuits in the control block. This is the path where a digit decoded from a keyboard input can be started on its way through the system of registers.

The decoding, timing and control circuitry, which can be called the controller, regulates all the other subsystems in the chip. What it does is principally governed by one or another of 320 instruction words (each consisting of 11 bits) delivered to the instruction register (and thence to the controller) from the read-only memory, so named because it contains a programmed set of operational instructions that cannot be changed after the manufacture of the calculator. Each instruction word, obtained from the read-only memory by nine-bit address words, establishes the operating rules that apply during one instruction cycle of 13 state times (39 clock cycles).

During each instruction cycle the shift registers may perform not only one complete circulation but also an addition or a subtraction if such an operation is called for. Register shifts and exchanges may also take place during the cycle. Further details of actions taken during an instruction cycle are conditioned by signals arriving at the chip through the four lines from the keyboard, by the counting of clock pulses, by the carry or borrow bit from the last adder operation and by 13 successive pairs of bits in a circulating 2-by-13-bit shift register

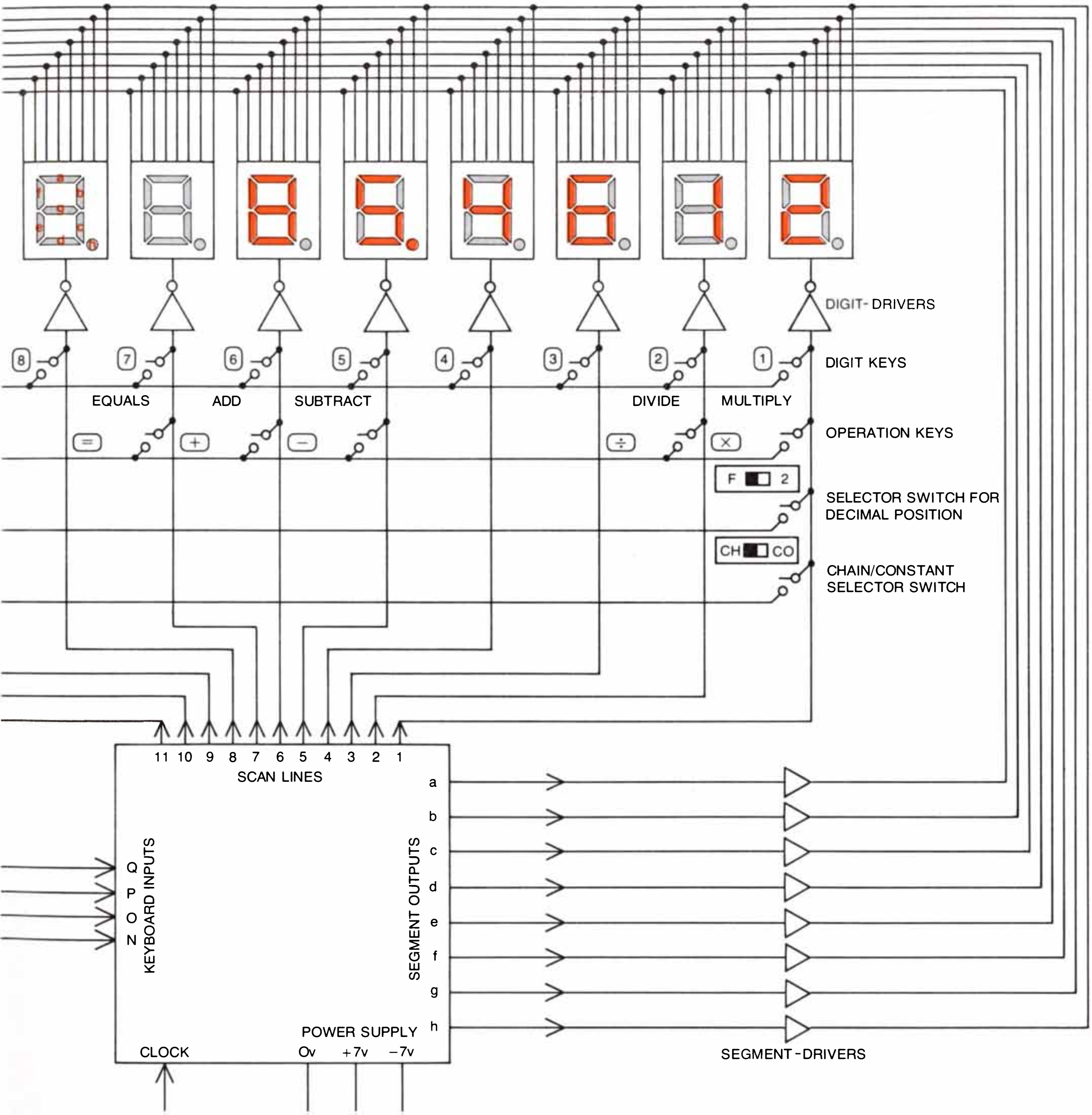


KEYBOARD AND DISPLAY circuits external to the main chip are diagrammed. The chip is represented by the square at bottom center. By means of the circuitry the chip communicates with 20 keyboard switches and

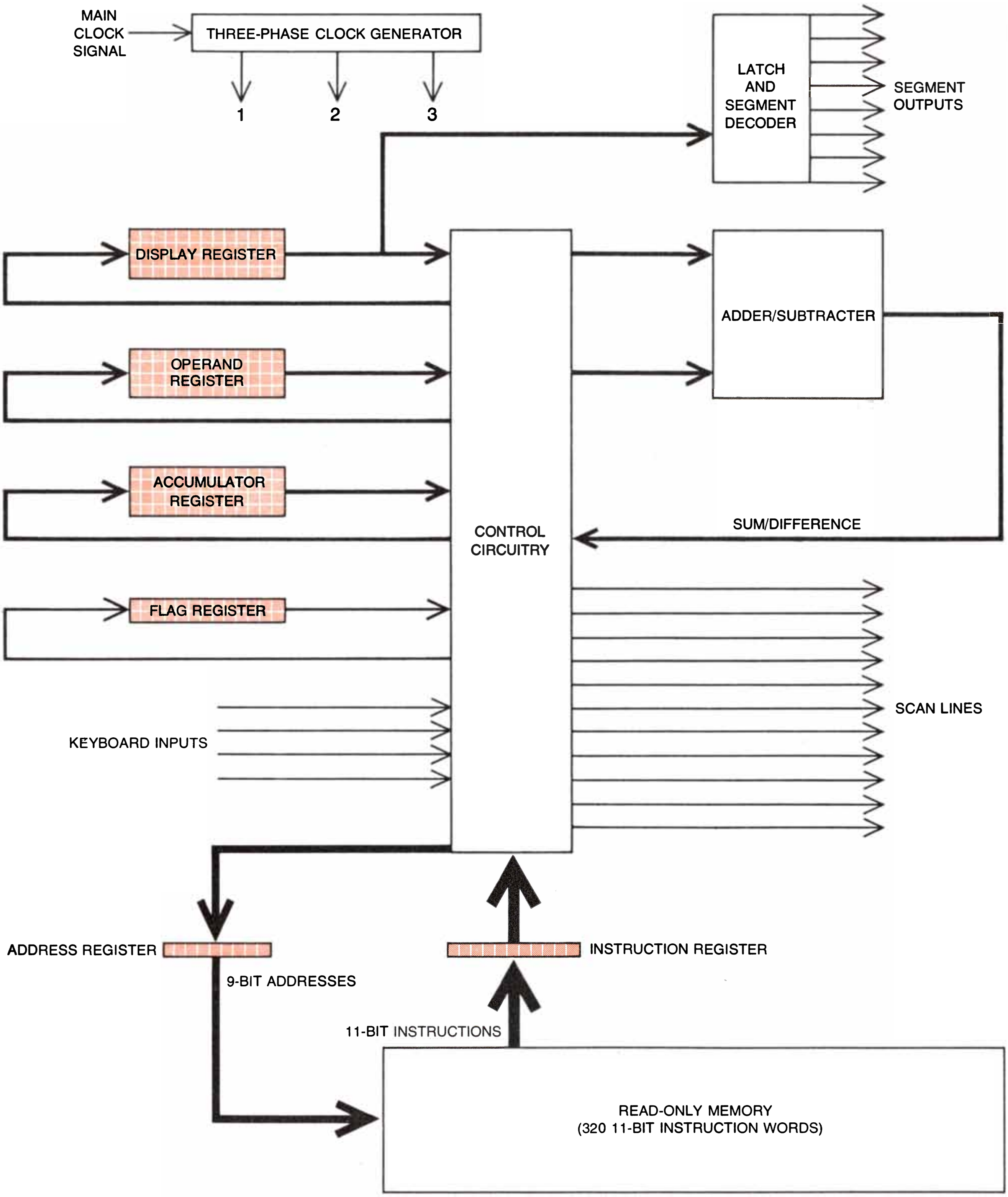
that serves as a kind of "scratch pad" memory at the disposal of the controller. It is this component that is known as the flag register.

Suppose the user wants the calculator to perform the calculation 2 + 3, which for the circuits of the chip entails the same type of procedure that would be followed in a much more complicated calculation. He turns the machine on and pushes the "2" key. The signal thus activated goes through one of the keyboard-input lines to the controller, where it causes the insertion of a 2 into the display register. From the display register numbers (including the 2) are transmitted to the lighted display panel of the calculator.

Next the user pushes the "plus" key. The signal thus activated is stored as a code in the flag register, available for consultation by the controller. Now the user pushes the



72 light-emitting diodes through only 23 of its 28 pins. In the matrix formed by the points where the 11 scan lines cross over the four keyboard-input lines each of the 44 crossovers is a possible location for a switch. Only 20 switches are needed for the hypothetical but representative four-function calculator shown in this diagram. When the calculator is operating, the scan lines are energized one at a time for 132 microseconds in a repetitive cycle, so that numerals illuminated in the display flicker imperceptibly. A displayed numeral is made up of a maximum of seven light-emitting diodes, all of which appear in color in the illuminated 8. An eighth diode forms the decimal point.



CONCEPTUAL ORGANIZATION OF CHIP is portrayed schematically. Each box represents a major electronic subsystem. The size of the arrows indicates the amount of information moving simultaneously from one section to another; the thinnest arrows represent one bit, or binary digit, and the thickest arrow, which emerges upward from the read-only memory, represents 11 parallel bits. In adding, say, 5 and 4 the user would begin by pushing the "5" key on the keyboard. The signal would go through one of the keyboard-input lines to the control circuitry, where (according to appropriate instructions from the read-only memory) it would be coded into binary form and transmitted to the display register. The user would then push the "plus" key, and the information would be stored in the flag register. When the "4" key is pushed, the signal would be handled as the 5 was handled earlier, and the 5 stored in the display register would be shifted to the operand register. Finally, the user would push the "equals" key. The control circuitry, still drawing instructions at each step from the read-only memory, would ascertain from the flag register what was supposed to be done with the two stored numbers. It would present them to the adder, which would add them and put the result in the accumulator register. From there the result would be immediately shifted to the display register, whence it would be transmitted to the latch and segment decoder for display on display panel.

"3" key. The 2 already stored in the display register is shifted to the operand register and the 3 goes through the same steps that the 2 went through earlier, ending up in the display register and in the lighted display.

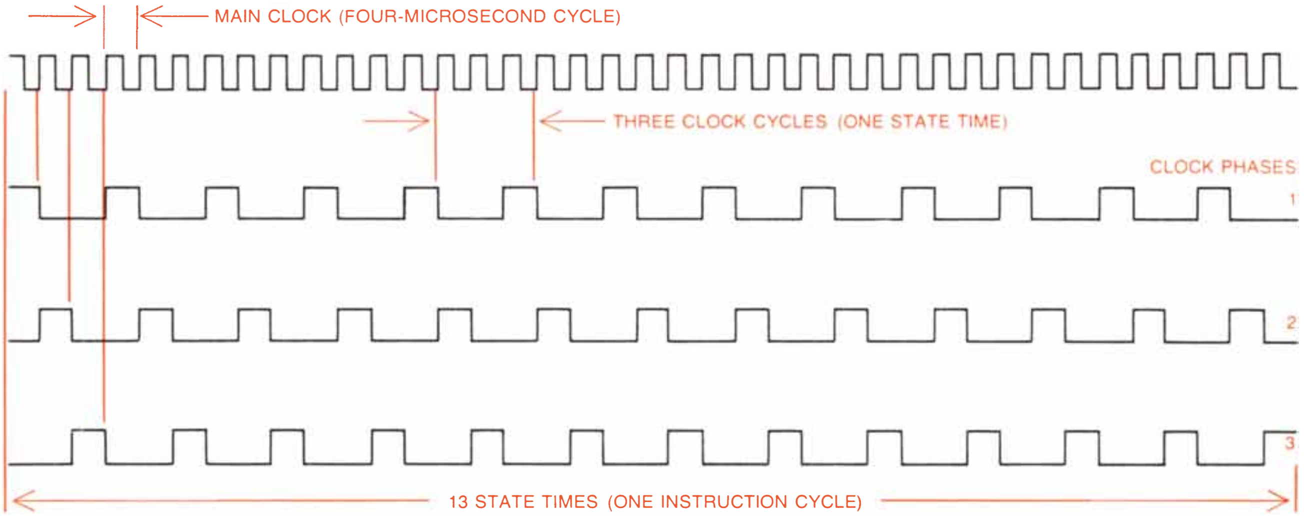
Finally the user pushes the "equals" key. The controller consults the flag register to find out what was supposed to be done with the two numbers (2 and 3) now circulating in the display register and the operand register. It ascertains that they are to be added. The controller takes the numbers from the registers where they are stored and presents them to the adder, the 2 as 0010 and the 3 as 0011. There they are added, and the result is put into the accumulator register. From there the result is shifted into the display register and displayed.

Even calculations scarcely more complex than 2 + 3 may require several passes through the routing and delay circuits and possibly through the adder. Suppose the first number in the calculation is 25.6. The 2 is entered first and appears in the display with a decimal point to its right. When the 5 is entered, the 2 must be shifted one place to the left in the display, but the decimal point must remain to the far right. Next the decimal-point key is pushed; the display does not change, but the signal from the key is stored as a notation in the flag register. Now when the "6" key is pushed, the decimal point must be moved one place to the left, along with the 2 and the 5. Shifting the decimal point requires the calculator to go through an adding cycle that is applied to the decimal-point digit, which is one of the 13 digits in the display register and is employed to keep track of the decimal point.

If the next command from the keyboard is to add and the number to be added (to 25.6) is 33.14, the first task is to line up the decimal points in the two registers involved, just as one would do in a manual addition. In other words, the decimal-point digits in the two registers whose contents are being added must be made equal by shifting the 25.6 one place to the left and adding a place to the decimal-point digit (yielding 25.60). After that has been done the two numbers can be added in one pass through the adder.

Those examples should suffice to indicate the kind of sequential operations, tests and branchings that must be executed in response to an input from the keyboard. A complete addition or subtraction involving a floating decimal point may call for perhaps 300 instruction cycles (some 12,000 clock cycles). The precise number depends on the programming and the size of the numbers entered in the calculation. Multiplication and division, which call for repetitive additions, subtractions and shifts, require correspondingly longer routines.

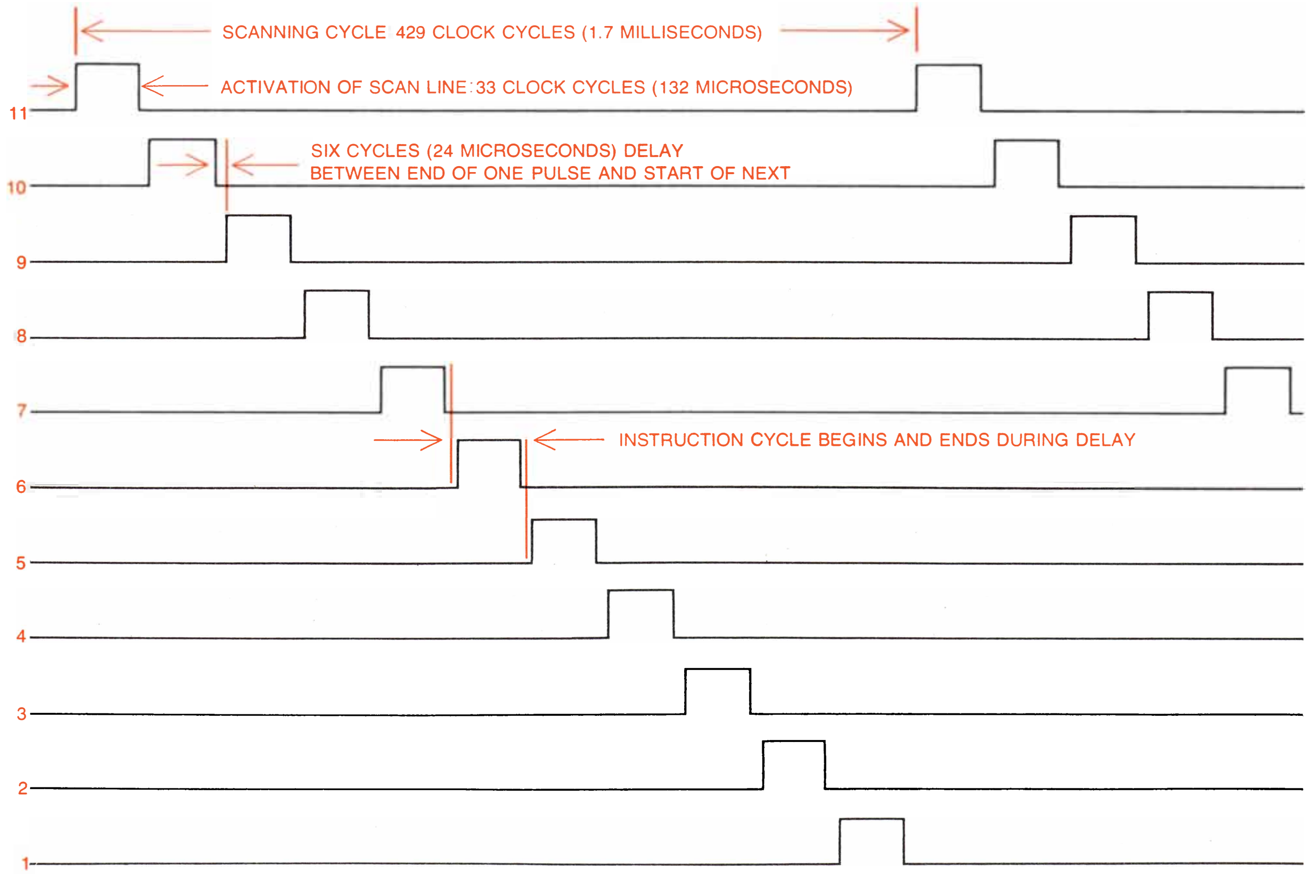
Although 12,000 clock cycles may sound like a long time for just an addition or a subtraction, one should remember that the actual time involved is less than .05 second. It would indeed seem an interminably long time for a large electronic computer, which would have a clock frequency perhaps 100



MAIN CLOCK of a typical four-function calculator operates at a rate of 250 kilohertz, or 250,000 cycles per second. Among other things, the clock establishes the timing for the circulation of data through the display, operand, accumulator and flag registers, which constitute a single 14-by-13-bit register (the random-access memory) that continuously makes one shift every three main-clock pulses. Three such pulses constitute one "state time," and 13 consecutive state times (one for each digit of storage in a 13-bit register) make up what is called an instruction cycle. During one instruction cycle the contents of a shift register usually complete exactly one recirculation.

times faster than the calculator's and would also have more efficient programming, mainly by virtue of its employing longer instruction words. To the human operator, however, the calculator appears to perform its operations almost instantaneously.

Our hypothetical calculator, like most other calculators, has many features of design that are reminiscent of a large computer. For example, the adder, together with the parts of the controller that route and delay the passing digits, corresponds to the central processing unit of a computer. The three data registers correspond to the computer's main memory. The remaining parts of the controller are equivalent to the timing and control circuitry that governs the operation of a computer at the lowest level, the level at which the modification of operations calls for changes in wiring configuration. The programmed routines stored in



SCANNING CYCLE employed for successively activating the 11 scan lines takes up 429 of a typical calculator's four-microsecond clock cycles. Each trace denotes the voltage in one scan line as time passes; vertical displacements show the timing of activation pulses.

the read-only memory are roughly equivalent to the microprograms in certain computers. (A microprogram in a computer is a more or less fixed routine, directly governing the control circuits, that is called into play by an instruction in a machine-language program.)

The operation of a microprogram in a computer or a calculator resembles that of a true program in that it generally involves fetching one instruction after another from a memory of some kind in accordance with addresses determined partly by testing the results of earlier operations. The concept of a program, however, ordinarily connotes a set of instructions that can be more readily composed and altered apart from the machine itself than microprograms can. The distinction is plain in "programmable" calculators, where each instruction of a program (possibly one composed by the operator) typically gives rise to a microprogrammed routine of the kind I have described. The distinction becomes somewhat blurred when one considers how electronic calculators perform functions other than the basic four, such as obtaining a square root or a sine or calculating y^x. Such functions are usually executed by a programmed sequence (achieved through microprograms) of additions, subtractions, multiplications and divisions.

One result of the computerlike structure of most small electronic calculators is that, as with a computer, a variety of differently programmed routines can be executed by the same framework of electronic circuitry. Different programs are ordinarily put in a computer by reading instructions into the memory, however, whereas a small calculator's microprograms are loaded into its read-only memory during manufacture.

Several themes can be discerned in the evolution of the small electronic calculator. They converge on more functions and more convenience for less cost. The sources of these advances include economies of scale, experience gained through high-volume production and advances in the technology of semiconductors.

On the one hand, significant cost reductions have been achieved by measures so seemingly trivial as making the printed-wiring conductors of the keyboard matrix perfectly orthogonal, so that no two scan lines and no two keyboard-input lines have to cross each other. In other words, a diagram of the keyboard matrix corresponds exactly to the actual layout of the keyboard. The step adds somewhat to the complexity of programming the routines set in motion by keyboard entries, but it reduces the number of costly interconnected layers required on the printed-wiring board under the keys. On the other hand, far-reaching technological innovations may soon make it economically practical to integrate all of a calculator's electronic circuitry on one chip, with an attendant reduction in the cost of assembly and improvement in reliability.

Although it is hazardous to forecast what might happen in such a rapidly changing industry, one can predict that the price of the cheapest basic calculators (already under $10) will go lower still. At the other extreme one can predict the development of fully programmable pocket-sized microcomputers costing less than $100. A certain amount of work has been done on calculators small enough to be worn like a wristwatch. Displays are expected to become larger and easier to read, perhaps through improvements in the technology of liquid crystals, which would also pay dividends in terms of reduced power consumption and longer service from batteries. Even though the small electronic calculator has revolutionized the way people deal with numbers, its full potential has not yet been realized.

MEANING OF NUMBERS		PLACES IN REGISTER				STEPS DURING ADDITION IN ONE PASS THROUGH ADDER
		THOUSANDS	HUNDREDS	TENS	ONES	
AUGEND	0835	0000	1000	0011	0101	NUMBERS IN OPERAND AND DISPLAY REGISTERS TO BE ADDED IN ONE PASS THROUGH ADDER.
ADDEND	0974	0000	1001	0111	0100	
	5				0101	ADD LEAST SIGNIFICANT DIGITS.
	4				0100	
SUM 9, CARRY 0				0	1001	BINARY SUM IS NOT GREATER THAN 9, SO PASS IT TO ACCUMULATOR. STORE THE CARRY IN AN INTERNAL LATCH.
	3			0011		ADD NEXT DIGITS, PLUS CARRY.
	7			0111		
PURE BINARY 10			0	1010		BINARY SUM IS GREATER THAN 9, SO ADD BCD CORRECTION OF 6.
CORRECTION 6			0	0110		
SUM 0, CARRY 1			1	0000		CARRY GOES TO LATCH, SUM TO ACCUMULATOR REGISTER.
	8		1000			ADD MOST SIGNIFICANT DIGITS, PLUS CARRY.
	9		1001			
PURE BINARY 18		1	0010			BINARY SUM IS GREATER THAN 9, SO ADD 6.
CORRECTION 6		0	0110			
SUM 8, CARRY 1		1	1000			PASS FINAL SUM AND CARRY TO ACCUMULATOR REGISTER.
SUM	1,809	0001	1000	0000	1001	FINAL SUM IN ACCUMULATOR REGISTER.

EXAMPLE OF ADDITION shows how the decimal numbers 835 and 974 are added by a calculator. The calculator operates on binary-coded-decimal numbers (BCD), in which each digit of a decimal number is encoded as four binary digits. The 5 of 835, for example, is encoded as 0101; reading from the right the code means one 1, no 2's and one 4. Each four-bit group, however, is necessarily added in pure binary code. For example, 3 (0011) plus 7 (0111) yields 1010, which is the correct 10 in pure binary code but is meaningless in BCD. Hence a second binary adder is provided in the BCD adder network to add 6 (0110) to any pure-binary sum greater than 9 to convert it to BCD form. Here pure-binary 10 (1010) plus 6 (0110) yields pure-binary 16 (10000), which is correctly interpreted in BCD form as 1 0000, or 10. Two such BCD corrections are necessary here as the pocket calculator carries out the operation of adding 835 and 974.


How did Dodge Colt put so much in such a little car?
Revolutionary Silent Shaft Engine
5-speed manual transmission
Flow-through ventilation
Tinted glass
Reclining bucket seats
Front disc brakes
Bumper guards—front and rear
Carpeting
Power brakes
Locking gas cap
33 MPG highway, 20 MPG city EPA estimate*
Adjustable steering column
Camping equipment not included
(But there's 58.3 cubic feet of space for it.)
Introducing the '76 Dodge Colt Estate Wagon. With all the standard equipment listed here, this Colt got 33 MPG on the highway, 20 MPG in the city. Colt comes in four other models, too: Coupe, Carousel, 4-door Sedan and GT. **Prices start at $3,175.** (Base sticker price for a 1976 Colt Coupe. Not including taxes, destination charges, license and title fees and optional equipment. California prices slightly higher.)
Dodge
*EPA estimate for Colt Estate Wagon. Your actual mileage may differ, depending on how and where you drive, the condition of your car and its optional equipment. In Californi ata.
CHRYSLER CORPORATION

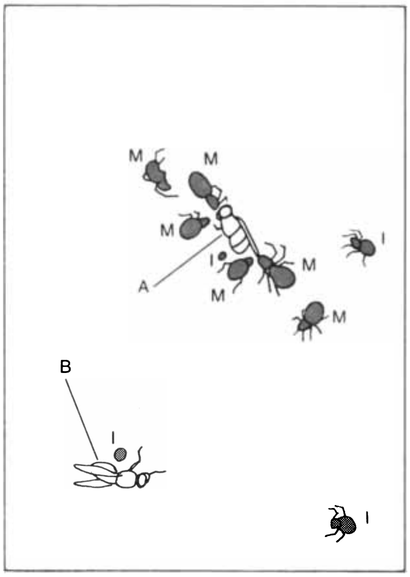
Social Spiders

Most adult spiders lead solitary lives. A few species, however, are gregarious and others even build large communal webs. Both degrees of spider sociality can be observed among species native to Mexico

by J. Wesley Burgess

Among insects—notably bees, ants and termites—social behavior is common. Among spiders it is rare. All spiders are predatory carnivores; among many of them even the male of the species cannot approach the female without risk of being attacked and killed. It is therefore paradoxical that there are any social spiders at all. How, then, can such spiders exist?

The number of social spiders is small; only in 12 genera distributed among nine families of spiders is any kind of sociality known. The 12 genera are, however, widely distributed, with representatives in both the Old World and the New. Two of the New World species are found in Mexico. I recently visited areas near Guadalajara where both species are present, observed the social spiders in their natural habitat and brought home to North Carolina a number of specimens for rearing and further observation in the laboratory.



COOPERATIVE CAPTURE of a fly (*A*) by several spiders is seen in the photograph on the opposite page. The diagram above identifies prey and predators. Spiders labeled *M* are mature; those labeled *I* are immature. Only two of the many flies on the web (*A, and B at bottom left*) are new catches. The spiders are of the social species *Mallos gregalis*. The cluster of mature spiders is feeding or preparing to feed. One immature spider has been drawn to the scene; another is approaching. The photograph was made in the author's laboratory; spiders were collected in Mexico.

The two Mexican species lead distinctively different lives. *Mallos* (formerly *Coenothele*) *gregalis* is a small spider, with a body that rarely exceeds five millimeters in length. It builds a large colonial web, surrounding the branches of a tree with a continuous sheet of silk. Its aggregations may be socially the most complex spider colonies in North America. *Oecobius civitas* is an even smaller spider; few have bodies more than two and a half millimeters long. It lives gregariously, spinning its silk shelter and alarm-system web in a dark and narrow microhabitat: the underside of a rock.

Spider societies are different from the societies found among the higher social insects both in kind and in degree. One reason is that a spider's web extends its range of sensory perception in a way that has no analogy among insects. Another is that the structure of a spider's mouthparts is such that it can feed only on other animal life. Any animal of appropriate size that a spider encounters, including a spider of another species or even the same species, is potential prey. It will nonetheless be useful in describing the sociality of the social spiders to sketch the probable evolution of different degrees of sociality among insects.

As Edward O. Wilson of Harvard University has pointed out, the eusocial insects, or higher social insects, have three traits in common: cooperative care of the young, a division of labor whereby more or less sterile individuals attend to the needs of fertile individuals, and a life cycle long enough for the offspring at some point to share the activities of the parental generation. The evolutionary routes that may have led from nonsocial to eusocial behavior appear to be traceable in terms of the less than eusocial behavior found among various insect relatives of the eusocial species. Charles D. Michener of the University of Kansas has outlined two such possible routes.

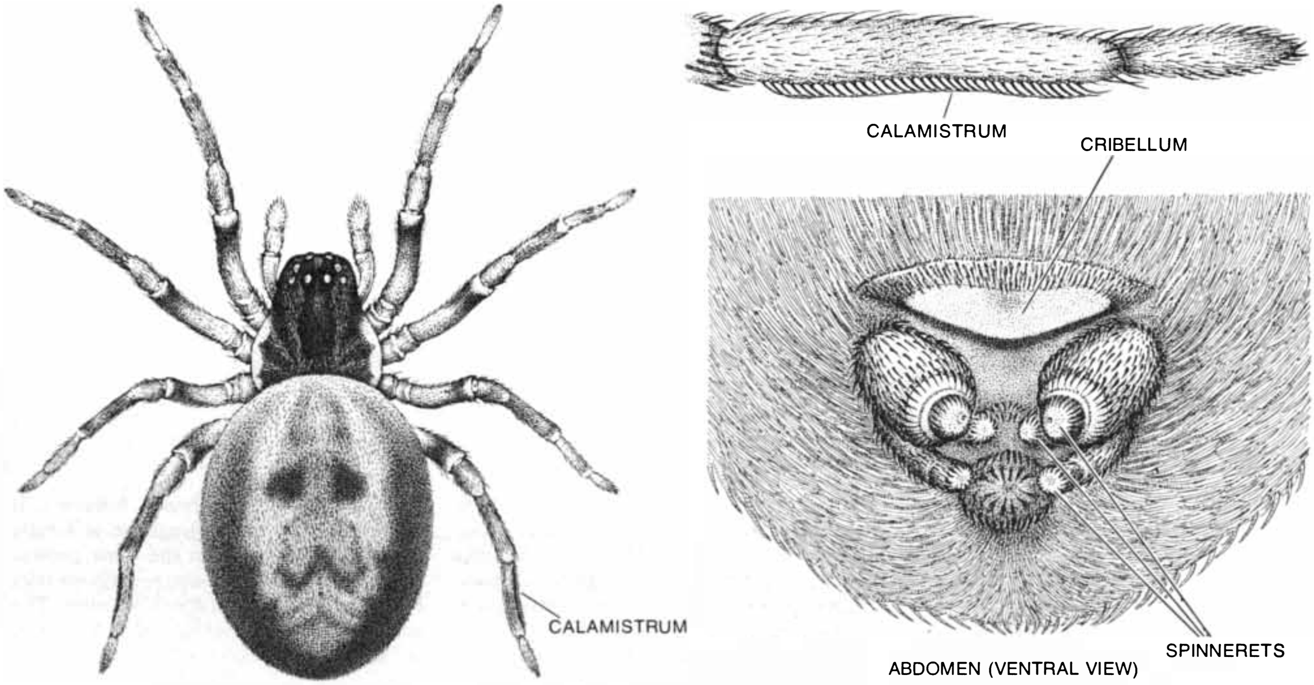
The first route Michener calls parasocial; on it there are three levels of increasingly complex behavior on the way to eusociality. The lowest level, communal behavior, is characterized by an aggregation of female insects, all belonging to the same generation; once the females have aggregated they build a communal nest for their young. The next level, quasi-social behavior, is characterized by cooperative care of the young. The third level, semisocial behavior, is characterized by the appearance of different castes that serve different roles. Thereafter eusociality is achieved when the life cycle is extended so that parents and mature offspring coexist in the same colony.

Michener's second evolutionary route he calls subsocial. On this route only one level of behavior precedes eusociality; it is characterized by solitary rather than communal nest building. The solitary female remains at the nest, however, and cares for her young. Eusociality is achieved in one step when the nest builder lives long enough to have the assistance of its first daughter generation in caring for subsequent, caste-differentiated daughter generations.

Looked at in these terms no social spider is eusocial. We must define the common base of spider sociality in much more restricted terms: the existence of various degrees of communality and of characteristic interactions among the members of communal aggregations.

Here it should be noted that with few exceptions even spiders that are solitary in habit go through a semicommunal stage early in their life cycle. Unlike insects, spiders do not have a larval stage. Each emerges from the egg as a functioning miniature adult, although it retains a yolk sac that supplies it with nutrients for several days. It grows in size and develops its sexual characteristics through a series of successive molts, the earliest of which takes place within the shelter of the parental egg sac. It leaves the egg sac fully prepared to spin silk and disable prey.

One might therefore expect that the spiderlings of the solitary species would scatter as soon as they leave the egg sac. Instead for the duration of a period known as the tolerant phase the spiderlings aggregate, and many of them join in the labor of building a small sheet web. They may even attack any small prey animal that blunders into the



COMMUNAL SPIDER *Mallos gregalis* has an average body length of five millimeters. Its complex flytrap web incorporates many sticky bands of silk that entangle intruders. The sticky silk is drawn from hundreds of microscopic pores in an abdominal plate (*right*), the cribellum. The spider combs out the silk with its calamistrum, an array of bristles that grows on the metatarsal of each of its hind legs.

web and wrap the intruder in silk. At this early stage, however, they never feed on the prey. After several days of the tolerant phase have passed the spiderlings disperse, build individual webs and feed on the prey they capture. All the spiderlings appear to adopt the solitary behavioral pattern simultaneously.

It is also noteworthy that in certain solitary-spider species (including representatives of the families Eresidae, Theridiidae and Agelenidae) the adult female does not abandon the egg sac after constructing it but remains with it, or carries it with her, until the spiderlings emerge. The female may then allow them to share her captured prey or may nourish them with regurgitated food or special secretions. Such parental care of the offspring bears a certain resemblance to the lower level of Michener's subsocial route to higher sociality. Thus even among the spider species that are recognized as being solitary, transient episodes of sociality may be observed.

When spiders live in groups, a number of additional interaction patterns are evident. Spider groups form in a variety of ways. For example, adult spiders of some species in the families Uloboridae and Araneidae will aggregate without regard to whether they are the offspring of the same parents or different ones. Each individual in these aggregations spins its own web. Among some species the individual may also contribute silk to a communal web area. Some of these groups may be made up of as many as 1,000 adults. In general each individual lives independently. All, however, share the benefits of a large aggregate web surface and of monopolizing a habitat that might otherwise have been shared by competitive species.

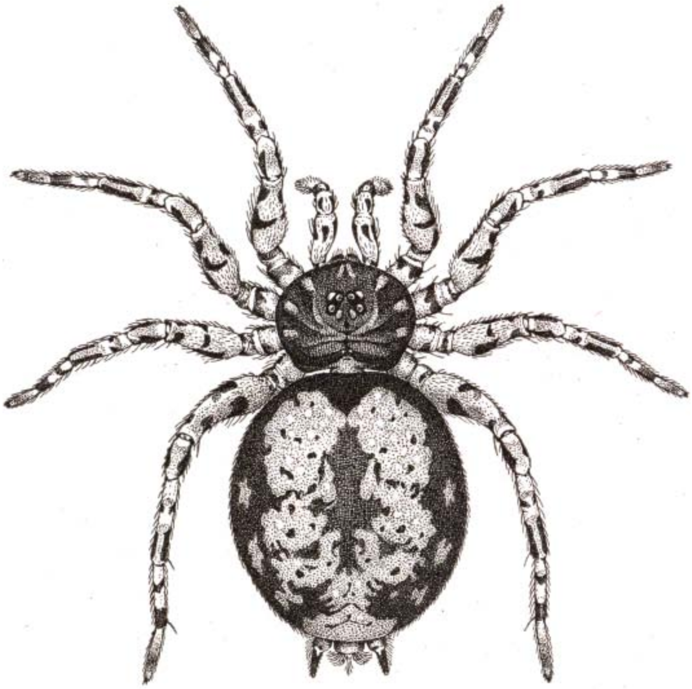
The viability of simple aggregations such as these demonstrates the existence of a tolerance mechanism in the individual adult spiders. At the very least the mechanism must be strong enough to keep the spiders from eating one another when prey are scarce. Evidently the mechanism is also species-specific; it is not limited to simply ensuring that the spiders are tolerant of all the other spiders in the aggregation. They are also tolerant of any spider of their own species. This has been demonstrated as follows. Individuals of the species *Metepeira spinipes,* a member of the family Araneidae, have been taken from populations living hundreds of miles away and introduced into local aggregations of *M. spinipes.* The presence of strangers did not disrupt the tolerance mechanism within the local aggregation, nor was any difference noted in the behavior of the two groups.

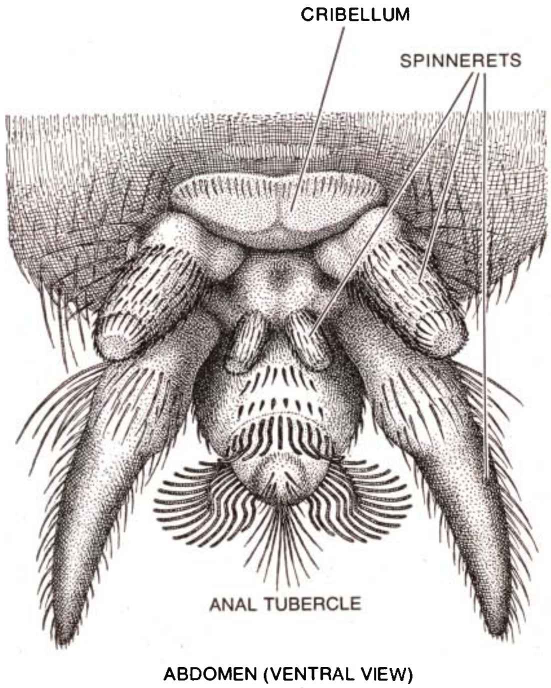
The most dramatic examples of spider sociality involve interactions substantially more complex than those I have been describing so far. These interactions are known only for four (possibly five) spider species. Two of the species are African: *Agelena consociata* and *Stegodyphus sarsinorum.* The others are New World spiders: *Anelosimus eximus* (and possibly a second species of the genus, *A. studiosus*) in South America and one of the species I have collected in Mexico, *Mallos gregalis.* All have in common the habit of constructing a large central web that is occupied by all the spiders in the aggregation. By combining their labors the spiders are able to construct a web that is much larger and far more elaborate in architecture than the web of any single spider; the structure is occupied by successive generations.

These spiders also collaborate in capturing prey much larger than prey any one of them could capture alone. Moreover, after the prey has been captured the spiders feed on it communally. Interactions as complex as these imply that these species have in addition to a tolerance mechanism a capacity for the coordination of individual responses to stimuli and an ability to recognize intraspecies sensory cues or to respond to some other kind of information. As an example, each spider seems to be able to distinguish between the web vibrations caused by a fellow member of the community and the vibrations caused by potential prey.

Bertrand Krafft of the University of Nancy has observed *Agelena consociata* in Gabon. He found that close-quarters tolerance in the species is mediated at least in part by chemotactic cues. Uninjured members of the community tolerate one another. An injured spider, or one whose normal superficial odor has been artificially altered by a washing in alcohol and ether, is attacked immediately. Neither the chemotactic cues nor other possible but still unidentified components of the spiders' tolerance mechanism are confined to local populations of the species. As with *Metepeira spinipes,* individual spiders of the same species can be moved from one colony to another without disrupting the communal activity pattern.

There is no evidence that any of these spider species has evolved a caste system such that the adults differ in form in accordance with any division of labor. Some difference in behavioral roles may exist as a result





GREGARIOUS SPIDER *Oecobius civitas* has a body averaging two and a half millimeters in length. Like *Mallos gregalis* it has a cribellum, but it uses its sticky silk actively, wrapping its prey rather than waiting for it to become entangled. The spider combs out a thread of the silk with its anal tubercle (*right*) and winds silk around the prey by circling it, abdomen foremost (*see middle illustration on page 106*).

of age or variations in biological rhythms, but just how cooperation is cued remains unknown. The pattern of behavior is nonetheless an example of sociality that is not easily equated with any pattern of sociality among insects. These spiders' behavior may well deserve a category of its own: communal-cooperative.

The Mexican social spider *Mallos gregalis* traps mostly flies on the sticky surface of the communal sheet web it spins around the branch of a tree. The spider has long been known to Mexicans as *el mosquero,* the fly-killer, and in the rainy season, when houseflies are particularly oppressive, those who live in the Guadalajara countryside will bring a web-covered branch into their house in much the same way that other people might string up flypaper. A member of the family Dictynidae, *M. gregalis* is a cribellate spider. Such spiders have a sieve-like plate, the cribellum, on the underside of their abdomen [*see illustration on opposite page*]. Sticky silk emerges from fine holes in the cribellum and is combed away with the two hind legs that bear a special row of bristles known as the calamistrum. This is the silk that forms the sticky prey-trapping areas on the outside of the spider's web. The web as a whole is an elaborate structure that includes supporting lines running between the surface sheet and the twigs and leaves of the branch, sheltered retreats for the spiders and special chambers where the female spiders live with their egg sacs. The sacs, thin wrappers of silk, contain from 10 to 20 eggs. The surface sheet is perforated in places with holes that provide access to the interior of the web.

The communal web of *M. gregalis* can be very large. One I saw near Guadalajara covered the limbs and branches in the upper three-quarters of a 60-foot tree of the mimosa family. Where the limbs met the trunk the silk of the sheet web was gray, but near the tips of the branches the silk was new and white. Evidently construction was continuing outward along the limbs. The spiders were not confined to the newer portions but were active in all parts of the web.

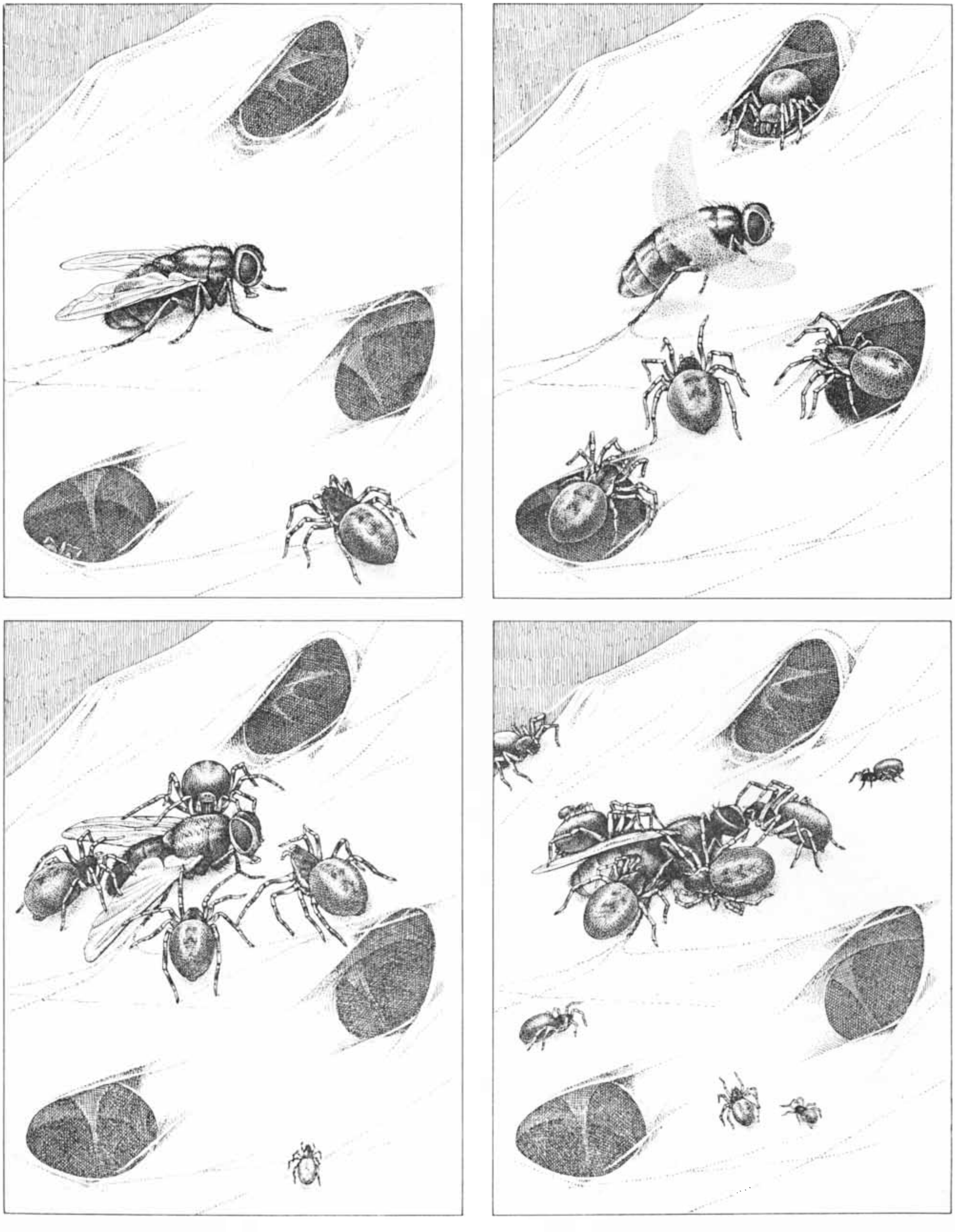
Both field and laboratory observations confirm that the construction of the *M. gregalis* web is a mutual effort. If a laboratory colony of the spiders has some treelike support available, such as an upright stick, the spiders will build their characteristic enveloping sheet web. In the absence of such a support they will build the kind of three-dimensional web that is typical of other dictynid species. Although this web looks different from the natural one, it too includes retreats and egg-sac chambers. In the laboratory web a task begun by one spider may be finished by another. I have also seen one spider of the colony lay down strands of ordinary silk, after which other spiders added bands of the sticky cribellate silk.

Observed in nature, the spiders seem to move around at random and without haste, emerging from and disappearing into the holes in the surface of the web. Their fly prey are trapped by the sticky web when they alight on it. When a fly gets stuck, two or three spiders approach the buzzing insect, immobilize it with their venomous bites and then feed on it. On occasion the spiders can be seen carrying flies down the holes into the interior of the web.

The spiders' predatory behavior can be observed in detail in the laboratory. We feed our colonies once every five days, which increases the probability that a majority of the spiders will be ready to feed at the same time. At any given moment one or two individuals in a colony of about 100 spiders are usually on the surface of the web; the other spiders will be in the web's interior. When a housefly is put into the spider cage and flies about, it makes a humming noise that is audible to the experimenter but causes no apparent change in the random activity of the spiders.

A fly that lands on a nonsticky part of the web and walks around stimulates a localized response; some of the spiders will turn to face in the direction of the fly, but that is all. If the fly gets entangled in a sticky part of the web and begins to buzz loudly, the behavior of the spiders changes abruptly. Throughout the web spiders that have been at rest turn toward the trapped fly and begin to approach it in short jumps. The fly continues to buzz even after the first spiders to reach it start their attack, usually by biting a leg or a wing. The buzzing draws more attackers; they move directly toward the fly over the web surface until eventually the prey almost disappears under the feeding spiders. Both male and female spiders attack. Even immature spiders take part, swarming over the adults in search of a place to feed.

Even though the attacking spiders in the caged colony are quite aggressive, we have never observed one spider attacking another. As we canvass the behavioral repertory that differentiates social spiders from solitary ones, this aspect of feeding in aggregations is significant. For example, young soli-



RESPONSE TO AN INTRUDER by a colony of communal spiders is reconstructed in these drawings on the basis of laboratory observation. Unlike the members of a wild colony all the spiders in the laboratory colony have fasted the same length of time. The sound of a passing fly is audible to a human observer but attracts no attention from the spiders. Even when a fly lands on the surface of the web (*top left*), only the nearby spiders reorient themselves. The buzzing of a fly entangled in sticky silk (*top right*) stimulates a response throughout the colony, and the spiders advance on the prey in quick jumps. The bites of the first spiders to reach the fly (*bottom left*) give rise to a louder buzzing that further stimulates spiders to approach the prey. As feeding begins (*bottom right*) immature spiders join the adults.

tary spiders such as those of the species *Araneus diadematus,* when they are artificially confined in close quarters, will also feed communally. Among the artificially confined solitary spiderlings, however, a tolerance mechanism, if it exists at all, operates only imperfectly; they will feed communally both on captured flies and on one another. This suggests that it is a strong tolerance mechanism that accounts for communal feeding in *Mallos gregalis* just as coordination mechanisms account for communal capture of prey.

The tolerance mechanism at work in *M. gregalis* colonies is being studied in our laboratory. It is evident from our observations that the mechanism is strong and that it operates both at close quarters and over considerable distances. Indeed, several separate mechanisms may be at work, perhaps mediated by cuing systems that allow discrimination between, say, the web vibrations caused by trapped prey and those caused by members of the colony. To test this hypothesis we are subjecting the colonies to the stimuli of various web vibrations in the hope of isolating such cues.

WEB-COVERED BRANCHES of a species of mimosa near Guadalajara support part of the communal web of a *Mallos gregalis* colony. Scattered holes allow the spiders to move freely from areas inside the web to the sticky outer surface where intruders become trapped. In the fly season local people often bring such branches indoors to serve as a kind of natural flypaper.

The social behavior of the second Mexican spider, *Oecobius civitas,* at first seems to be principally aggregative, like the behavior of other spiders that build their nests in close proximity. The darkness of this spider's microhabitat makes observation of its behavior difficult, but its unusual method of prey capture has been recorded. *O. civitas* has a finger-shaped organ, the anal tubercle, on the abdomen near its silk-extruding spinnerets. With this appendage it can comb sticky silk out of its cribellum in a rope that it winds around its prey [*see illustration on page 103*].

A closer study of the sociality of *O. civitas* proves that it is more than merely aggregative. The spiders' behavior features a curious combination of tolerance and avoidance. On the underside of the rock that shelters the spiders each individual weaves a small open-ended tube of silk that is its hiding place; around this retreat the spider constructs a thin, encircling alarm-system net close to the surface of the rock. The pair of structures makes up the spider's web, which is generally found in a hollow or a crevice of the rock. If a spider is disturbed and driven out of its retreat, it darts across the rock and, in the absence of a vacant crevice to hide in, may seek refuge in the hiding place of another spider of the same species. If the other spider is in residence when the intruder enters, it does not attack but darts out and seeks a new refuge of its own. Thus once the first spider is disturbed the process of sequential displacement from web to web may continue for several seconds, often causing a majority of the spiders in the aggregation to shift from their home refuge to an alien one.

Field observations and experiments indicate that, as with *Metepeira* and *Mallos,* the mechanisms responsible for the combination of tolerance and avoidance extend beyond the local population to include other spiders of the same species. Moreover, within the local population the shift to another spider's shelter may be a semipermanent move. The reason is that when the spiders are undisturbed, they occupy a fixed web position for long periods. In any event the behavioral pattern of the species benefits the individual spider by providing more than one available retreat in an emergency.

The group behavior of *Oecobius civitas* is far simpler than that of *Mallos gregalis.* It is nonetheless effective in enabling the spiders to live together under crowded conditions. No doubt the avoidance mechanism makes a major contribution toward the spiders' ability to maintain a high population density in their restricted microhabitat. Other contributing factors probably include the spider's unusual predatory technique and the spacing of individual webs. In any case, although we remain largely ignorant of the mechanisms underlying avoidance and tolerance, they appear to be the basic building blocks that provide a foundation for more complex group behavior.

It has been suggested that *Oecobius civitas* displays an even more remarkable kind of sociality: construction of a communal egg sac by the females in the aggregation. The possibility of such a behavioral advance, unknown among spiders, came to light recently when William A. Shear of Hampton-Sydney College undertook a taxonomic review of the oecobiid spiders. He was assisted by a number of colleagues who donated specimens to the project. Among the donors was Willis J. Gertsch, curator emeritus of spiders at the American Museum of Natural History, who had collected specimens of *O. civitas,* its web and its egg sacs in the Guadalajara area.

The usual oecobiid egg sac contains from five to 10 eggs. In the preserved material donated by Gertsch, however, Shear found two groups of more than 200 immature spiders. Each group was contained in what gave every appearance of being a single egg sac. Shear published his observation in 1970, suggesting that *O. civitas* might be a communal egg layer.

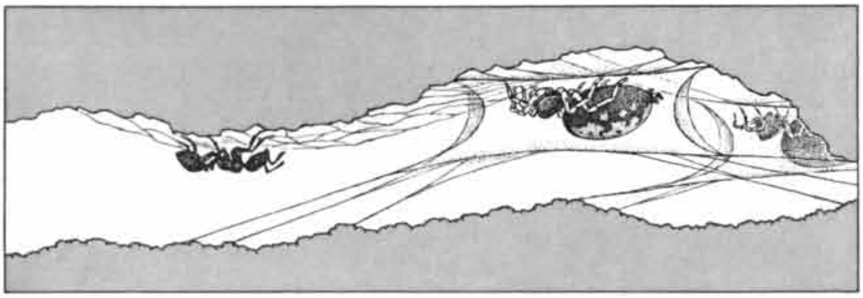
When I collected specimens of *O. civitas* and its egg sacs in the area near the shores of Lake Sayula, where Gertsch had done his collecting, I found that several other species of spiders shared the rocky habitat with the oecobiids. As a result a variety of egg sacs could be collected. This I did, sealing individual egg sacs in individual tubes. I was disappointed to find that only the small sacs, averaging seven eggs to a sac and mainly collected in or near *O. civitas* web retreats, hatched oecobiids.

After rearing this spider in the laboratory for three generations and observing only individual egg sacs containing from five to 10 eggs, I consider that to be the normal pattern of reproductive behavior in *O. ci-*

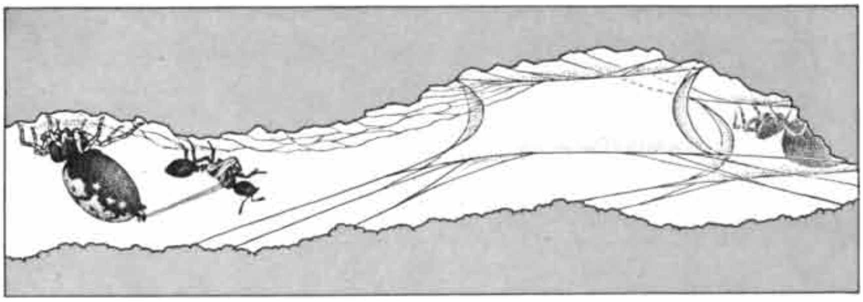
vitas. To resolve the question beyond all doubt other single *O. civitas* egg sacs containing eggs or immature spiders in large numbers will have to be collected in the field.

Mating behavior has not yet been observed in our laboratory populations of either *Mallos gregalis* or *Oecobius civitas.* Solitary male spiders go through elaborate pre-mating maneuvers, so-called courtship patterns that supposedly inhibit predation in the female at the time of copulation. Among social spiders, which live in tolerant aggregations, such maneuvers would not seem necessary. Indeed, if differences in copulatory patterns between solitary and social spiders do exist, they may even provide clues to the evolutionary background of spider sociality. In this connection we have made one possibly significant observation concerning fertility. Solitary spiders raised in the laboratory retain the cyclical breeding rhythms characteristic of their wild state, but when our *M. gregalis* colonies are provided with a uniform environment and controlled periods of darkness and light, they produce fertile eggs throughout the year.

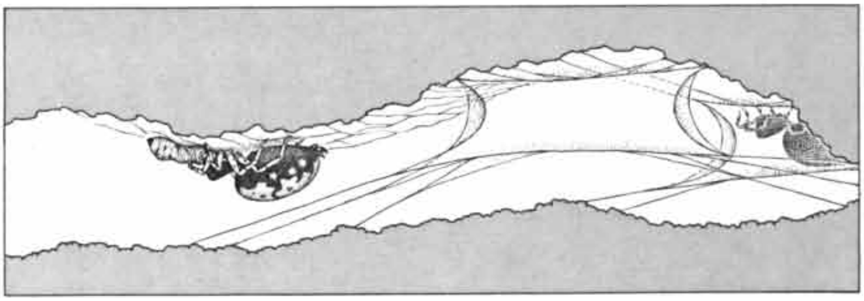
Observation of the two Mexican spiders has uncovered a substantial amount of information about their sociality, but that information more often than not merely defines the extent of our ignorance. For example, we do not know what conditions favor the development of spider sociality or even what mechanisms are involved in tolerance, avoidance, the formation of groups or the coordination of activity. Moreover, it is not known how different forms of spider sociality are related to one another or how, in complex interactions, intragroup information is transferred. The search for answers nonetheless seems to offer one certainty: The more we learn about the sociality of comparatively simple animals, the better we shall be able to understand the sociality of more complex species, including our own.



CAPTURE OF PREY, usually a foraging ant, by a spider of the gregarious species *Oecobius civitas* follows a complex pattern that begins when the intruder disturbs the spider's alarm web.



ALARMED SPIDER leaves its shelter and moves in circles around its prey, its abdomen foremost and raised clear of its legs, while it combs out a strand of sticky silk with anal tubercle.



WRAPPED IN SILK, the ant is immobilized. The spider may rest for a time or may turn (*left*) to bite and disable its prey. Only the captor feeds on the prey; nearby spiders do not approach.

Western Electric Reports:

Ion implantation with a new twist.

Western Electric produces millions of semiconductor components a year for use in Bell System telecommunications equipment. An essential step in the process is "doping"—introducing a precisely controlled impurity into the semiconductor material to alter its electrical characteristics. Until now, that's generally been done through diffusion techniques which entail masking a semiconductor wafer and "baking in" impurities—basically a broad brush process.

Bell Labs engineers felt that using ion beams to implant the dopants would be better than diffusion techniques. Ion implantation would improve the performance of some existing devices and would permit the design of new devices that require very precise control during manufacture.

But until recently, conventional ion implantation systems had serious drawbacks for heavy doping of wafers at fast production rates.

Electrical systems, in which a moving beam scans a stationary wafer, tolerate only relatively low currents before the beam starts to spread apart. And low currents mean low production rates.

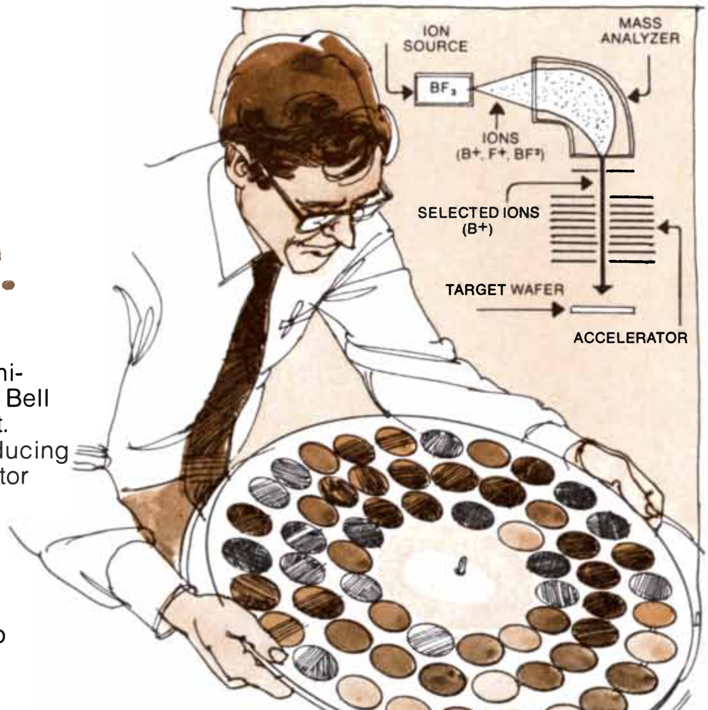
Mechanical (x-y) systems in which wafers are moved back and forth across a stationary beam, can use high currents but are unacceptably slow.

Now, engineers at Western Electric's Engineering Research Center in Princeton, N.J. have developed a new rotating mechanical scanner that overcomes the low speed limitation while still accommodating a high beam current.

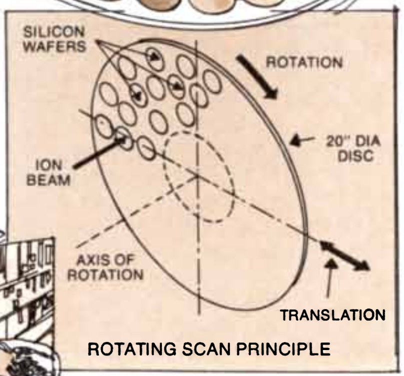
The heart of the rotating scanner is a 20-inch aluminum disc which holds 60 silicon wafers in concentric rings. The disc rotates at about 1000 rpm as it moves from side to side through the path of a stationary ion beam. The beam traces a long overlap spiral on the disc.

To ensure uniform dosage, a digital control system adjusts the speed of the disc's side-to-side motion as the spiral moves toward the center of the disc.

Benefit: Rotating ion implantation techniques are one more manufacturing innovation that allows the Bell System to meet your communications needs reliably and economically.



Ion implantation: a gas is ionized. Ions are drawn into a beam, mass selected and implanted into target wafers. With the rotating mechanical scan method, a 3-inch silicon wafer can be implanted in a few seconds. Conventional x-y mechanical scans require about a minute.



The scanning device is a disc which spins on an axis parallel to the ion beam. The second direction of scan is provided by translating the entire rotating disc in the plane of the disc and perpendicular to the beam.

Ion implantation techniques are used in the manufacture of semiconductor devices used throughout the Bell System from switching and transmission gear to terminal equipment.

Western Electric

Fluorescence-activated Cell Sorting

This new technique, which is capable of identifying and isolating closely related types of animal cells at rates of up to 5,000 cells per second, holds much promise for investigations of cell biology

by Leonard A. Herzenberg, Richard G. Sweet and Leonore A. Herzenberg

Perhaps the chief obstacle to the study of cell biology today is the problem of obtaining pure populations of different kinds of living cells from a given organ. Imagine molecular biology without tools such as ultracentrifugation, column chromatography and electrophoresis for the isolation of DNA, RNA, enzymes and other constituents from the molecular "soup" obtained by the disruption of cells. Detailed descriptions of the mechanisms involved in the metabolism of sugar and the synthesis of protein—two outstanding achievements of recent years—would still be a gleam in the biochemist's eye. An important prerequisite for studying the components of any biological system, be they the molecules in a cell or the cells in an organ, is the ability to isolate those components from one another so that they can be characterized and recombined under controlled conditions.

Of course, cell biology is not entirely lacking in techniques for isolating and studying living cells. Several methods, including centrifugation and electrophoresis, have played a key part in fundamental discoveries. As knowledge in the field expands, however, so does the quest for a more refined method to separate closely related yet functionally distinct types of cells. Here we shall describe a promising new approach to the problem of cell separation, pursued mainly in our laboratory at Stanford University.

The common starting point for all cell-separation techniques is the breakdown of the organ (or organism) to obtain a suspension of viable dissociated cells floating free in a suitable medium. That is fairly easy to accomplish with many tissues, where it is generally adequate to crush the excised tissue against a fine wire mesh. Other tissues may call for enzymes or other chemicals to sever the intercellular connections that bind the cells together. From some organs only one or a few of the cell types present can be recovered, whereas for other organs no techniques have yet been found for obtaining viable cell suspensions in good yields.

The development of techniques for preparing suspensions of dissociated cells was in itself a major advance in cell biology. Much of what we know today about cell functions comes from studies in which suspensions of cells are either injected back into animals or cultured for longer or shorter periods in the laboratory. As the cells obtained from a given organ were found to be responsible for a given function, however, questions inevitably arose: Which cells in the suspension were actually doing the work? Was an interaction between two or more different cell types required?

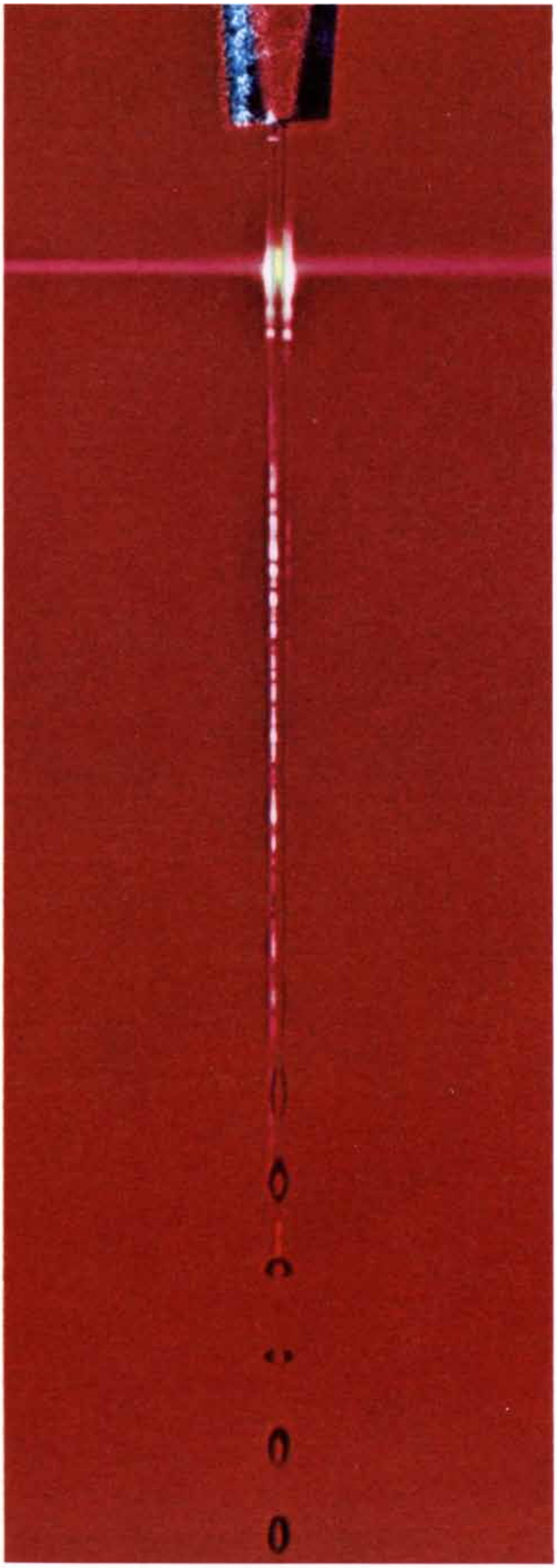
The existing methods for separating cells grew out of the need to answer such questions for cell suspensions from a variety of organs being studied for a variety of functions. Some of those methods depend on the selective killing of unwanted cells in order to enrich the suspension in the surviving type (or types) of cells. Other techniques aim for selective enrichment based on physical or chemical differences between the cells in the suspension. In either case the method of choice is dictated by the cells under study, their susceptibility to damage by various treatments, the methods used to assay cell function and, most important, by the nature of the differences between the cells that must be separated from one another.

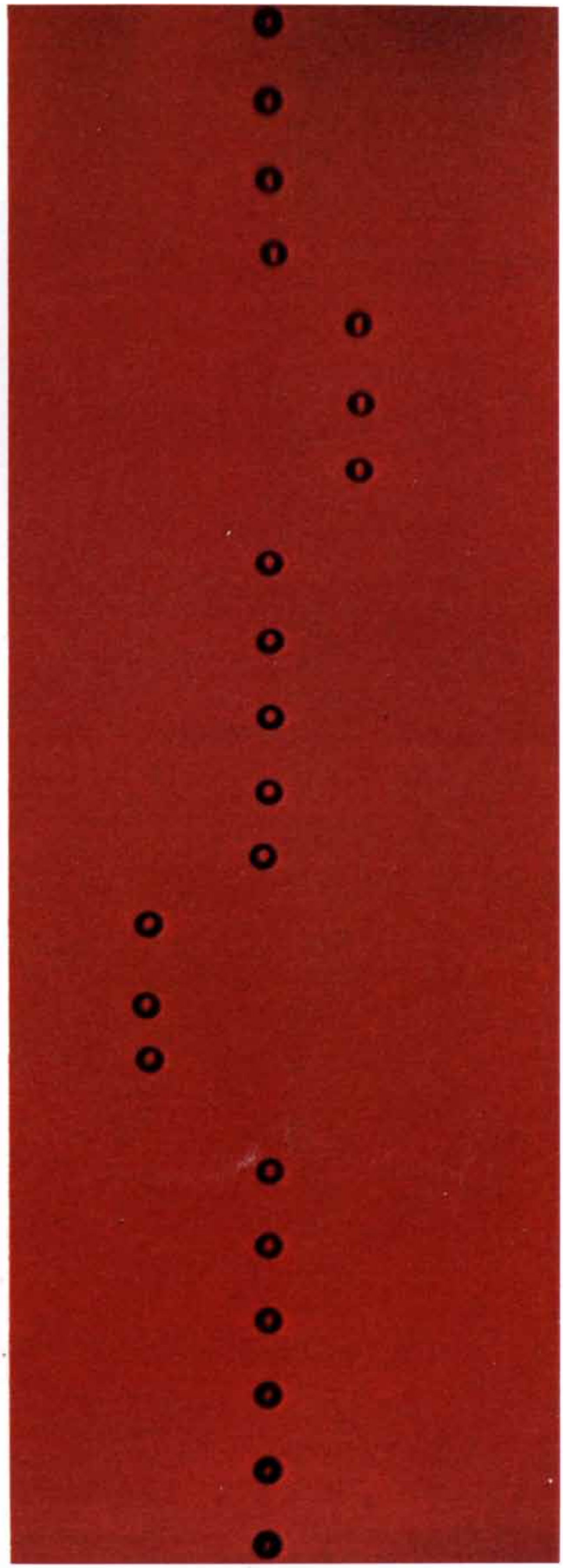
Cell-separation methods can take advantage of differences in the physical properties of the cells (such as size or density) or they can exploit differences in the properties of the surface membranes of the cells (such as their electric charge or their ability to adhere to glass or plastics). Because functionally different cells may differ from one another only in the amount or type of a single surface membrane constituent, however, many laboratories, including our own, have sought techniques with more specificity than is afforded by physical criteria alone.

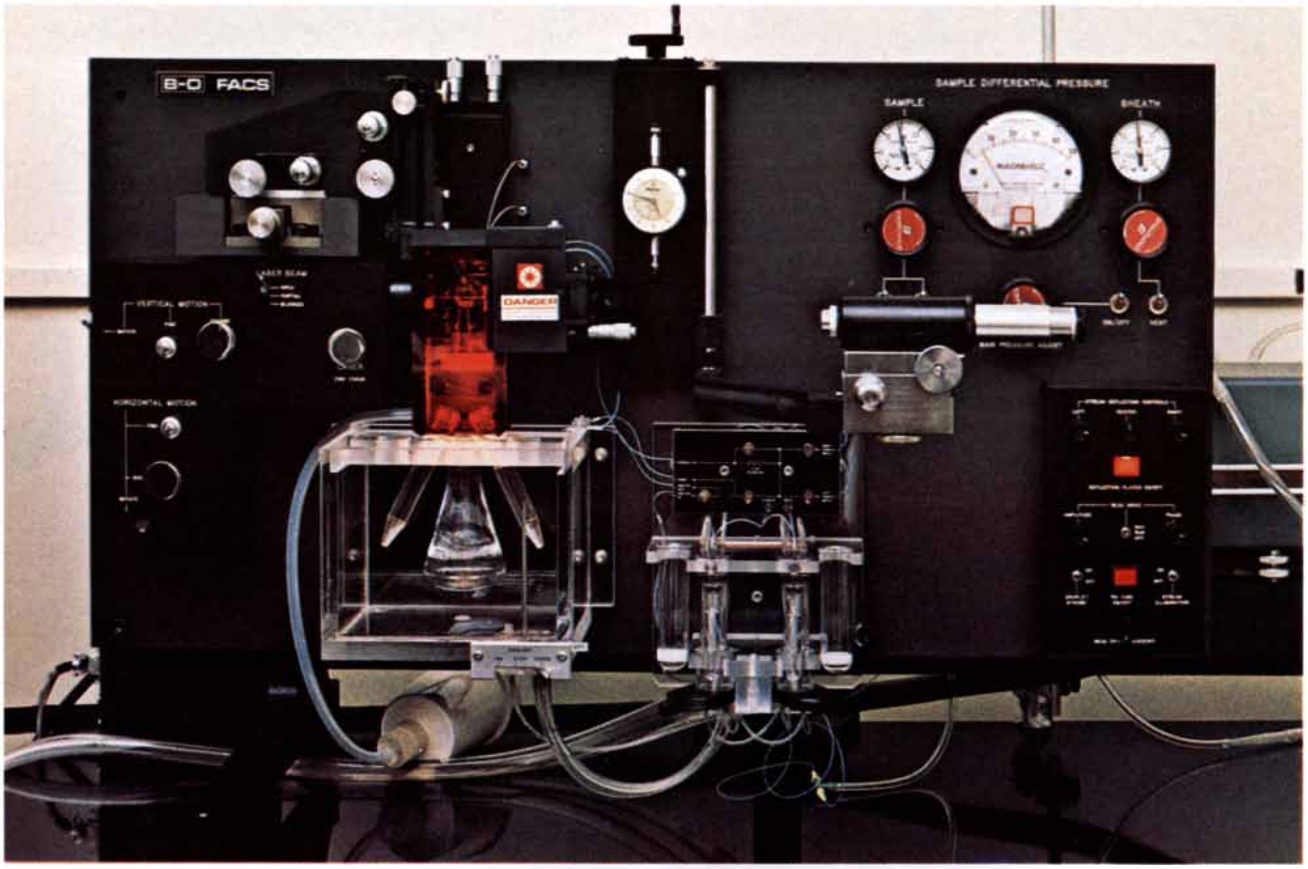
Several years ago we were searching for methods for isolating variants that appeared infrequently in cultured animal-cell lines. We wanted variants that differed only slightly from the parent cell line, perhaps only by the gain or loss of a single structure on the surface of the cell. We were struck by the apparent ease with which certain structures on the cell surface could be coated with a specific fluorescent substance and by the almost infinite variety of surface structures that could be labeled in this way to distinguish the cells from one another. It was clear that we could identify the variants we wanted with the aid of the fluorescent-marker method and the fluorescence microscope. Other investigators had already done so in similar experiments, but what we needed for our experiments was a method not only for identifying fluorescence-tagged cells but also for isolating them.

We were at that time also becoming interested in the interactions of lymphocyte cells in the immune response and in the genetics and biology of lymphatic tumors. Again we recognized the need for a method for isolating various kinds of lymphocytes, many of which deviate from other lymphocytes only by small surface differences. And again be-

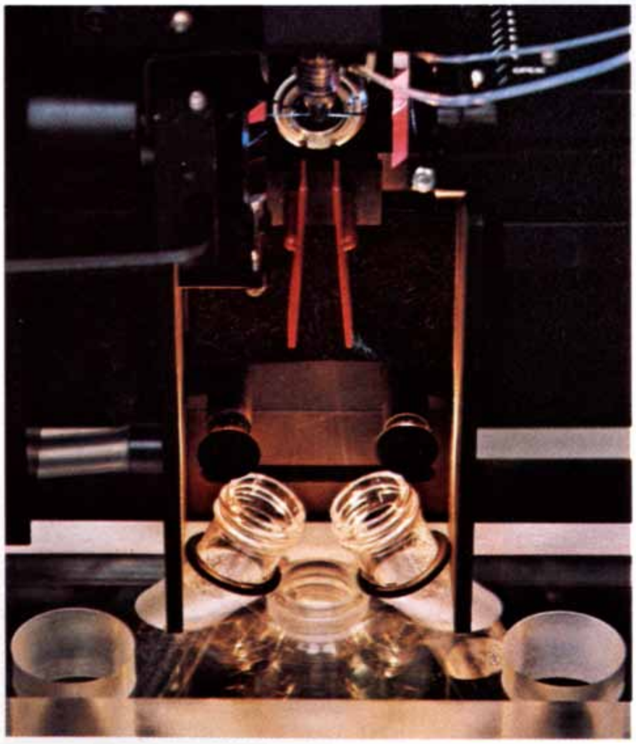
FLUID JET containing a mixture of different cells, some "tagged" with a fluorescent dye, is shown emerging from a rapidly vibrating nozzle and breaking into a uniform stream of tiny droplets in the left-hand photograph on the opposite page. Just below the nozzle, before the droplets form, the jet is illuminated with the focused blue light of an argon-ion laser, which excites a yellow-green fluorescence in the tagged cells. The fluorescent light is detected, along with light scattered forward out of the illuminating beam by the passing cell, and the resulting electrical signals are used to charge the liquid stream exactly when the droplet containing a desired cell is forming. Farther downstream the droplets, which retain their charge after they separate from the stream, pass through a constant electric field across their path. As the right-hand photograph shows, charged droplets are deflected toward appropriate collecting reservoirs, whereas uncharged droplets continue on their original course. Both of these demonstration photographs are multiple exposures made with the aid of a microscope and a stroboscopic lamp timed to flash synchronously with the procession of falling droplets; thus the image of each droplet seen in the photographs is actually made up of thousands of identical droplet images. Here the fluorescence was enhanced by adding extra fluorescent dye to cell suspension.



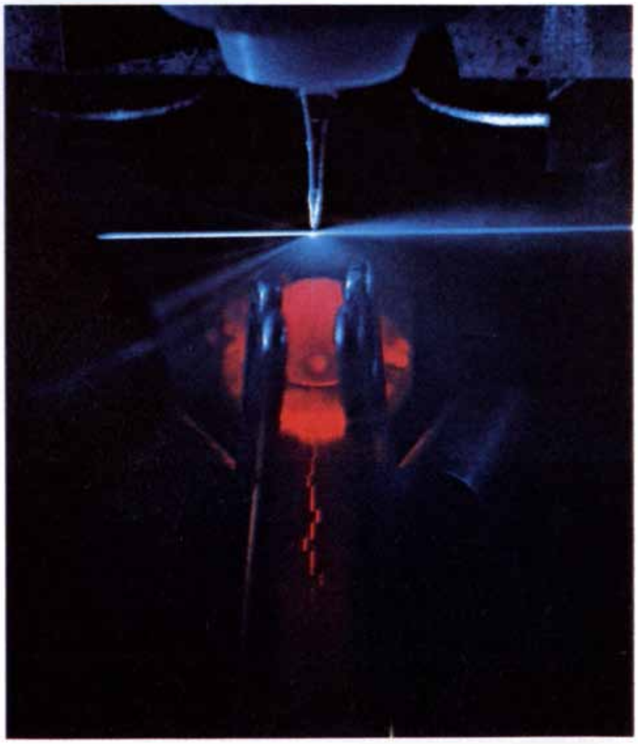




CURRENT MODEL of the fluorescence-activated cell sorter, the FACS-II, was photographed in the authors' laboratory at Stanford University, where it was originally developed. The machine, which is now manufactured by the Becton Dickinson Electronics Laboratory, is capable of separating as many as 90 million cells, differing by as little as 5 percent in their fluorescence, in a typical run of five hours.



CLOSEUP PHOTOGRAPH shows the portion of the Stanford cell sorter where the fluid droplets are actually formed and separated. The deflected streams of droplets are barely visible in this normal exposure as faint lines curving toward the collecting vessels at bottom.



STROBOSCOPIC-FLASH PHOTOGRAPH similar to photographs on the preceding page but at a macroscopic scale manages to capture the deflected droplets in mid-flight near the electrically charged plates. The laser beam was made visible by blowing smoke in its path.

cause these differences could be detected by the fluorescence microscope, we wanted a method of isolating cells tagged with a fluorescent marker.

Those needs led us to become intrigued with the idea of building a fluorescence-activated cell-sorting machine that would enable us to do the desired experiments. We contacted Louis A. Kamentsky of the International Business Machines Corporation, who had already built a machine that could separate fluorescent cells. The goals set for his instrument, however, were different from our own. Kamentsky was concerned with isolating small numbers of potential cancer cells—not necessarily viable after isolation—and depositing them on a slide for microscopic study by a pathologist, whereas what we wanted was to obtain large numbers of isolated cells, both viable and suitable for a variety of functional biological studies.

We would probably have gone no further than dreaming about a cell sorter that met our needs if it had not been for the fortunate proximity in another laboratory at Stanford of a team of investigators engaged in the National Aeronautics and Space Administration's Exobiology Program. Over coffee and lunch we talked with our engineer neighbors about the feasibility of jointly building a cell sorter. Eventually we got some initial funding from various sources and decided to attempt to build such a machine. Several years later a prototype model of a fluorescence-activated cell sorter, designed principally by William A. Bonner and H. Russel Hulett, went into operation in our laboratory.

The fluorescence-activated cell sorter, or FACS, separates cells according to how much fluorescent dye is bound to each cell. The current model, FACS-II, detects cells with as few as 3,000 molecules of fluorescein (a common fluorescent dye) on each cell. In addition, by means of correlated light-scattering, the device simultaneously measures the sizes of all cells; hence it can be set to separate those cells that fall within desired ranges of both size and fluorescence. The sorter can operate efficiently at rates of up to 5,000 cells per second, which means that in a typical five-hour run 90 million cells can be processed for various experimental purposes.

Briefly FACS-II works as follows. As cells emerge from the starting reservoir they are rapidly passed single file across a small area intensely illuminated by a laser beam; they are then isolated in tiny droplets that are electrostatically charged and deflected to different collecting reservoirs [*see illustration on next page*]. As the cells flash across the laser beam, the optical signals generate electric pulses, which are rapidly compared with preset criteria so that the droplet containing the observed cell can be deflected into the appropriate reservoir.

The techniques we rely on for generating and controlling fluid droplets have their origin in the work of the 19th-century physicist Félix Savart, who showed that a small fluid jet would break up into a procession of droplets having remarkable uniformity and regularity if the nozzle or orifice that formed the jet was vibrated at the proper frequency. Savart's work, published in 1833, stimulated research by other investigators, including Lord Rayleigh, Chichester Bell (Alexander Graham Bell's cousin) and C. V. Boys. In Boys's well-known book on soap bubbles and other surface-tension phenomena there appears a photograph of a water jet made with the light of an electric spark. A musical note, acoustically coupled to the jet nozzle, produced minute bulges in the cylindrical column of fluid as it emerged from the nozzle, and surface-tension forces amplified those regular disturbances until they severed the jet. By deflecting droplets generated in this way with the field produced by a piece of electrified sealing wax Boys showed that processions of identical droplets could be made to follow precise curved trajectories.

Those early studies were limited to jets with a diameter on the order of a millimeter and a repetition rate of less than 1,000 droplets per second. In 1961, after preliminary experiments had verified that much higher drop frequencies could be obtained with smaller jets, that the drops could be independently and precisely charged and that useful deflections could be produced with available electrostatic technology, a project was initiated by one of us (Sweet) at Stanford's Applied Electronics Laboratory to utilize those techniques as the basis of a high-speed recorder that would write with electrostatically deflected droplets of ink. The Stanford project developed techniques for accurately and independently deflecting ink droplets at rates of up to 200,000 per second, a capacity that has since led to the production of high-speed ink-jet computer printers, mail addressers and other printing equipment.

In 1964 the Stanford work on ink-jet recording was noticed by Mack J. Fulwyler and Marvin A. Van Dilla of the Los Alamos Scientific Laboratory. Fulwyler had been studying individual cells with a Coulter counter, an instrument that determines the volume of a cell by flowing an electrically conducting saline solution containing suspended cells through a tiny aperture and measuring the change in the electrical resistance of the fluid in the aperture as each cell or other insulating particle passes through it. Fulwyler successfully combined this counting principle with the Stanford droplet generator in an instrument that initially was capable of sorting cells at rates of up to about 1,000 per second. In describing their apparatus Fulwyler and Van Dilla noted that measurements of the optical properties of cells (such as their fluorescence) might also serve as the basis of a sorting technique.

The present Stanford fluorescence-activated cell sorter is an outgrowth of that work. Cells in liquid suspension are forced under pressure through a micronozzle into the center of a stream of cell-free fluid and then out through an effluent nozzle 50 micrometers in diameter. This design creates a coaxial flow that keeps the cells near the axis of the effluent jet. The nozzle assembly is vibrated axially at an ultrasonic frequency of 40,000 cycles per second, breaking the jet into 40,000 uniform droplets per second.

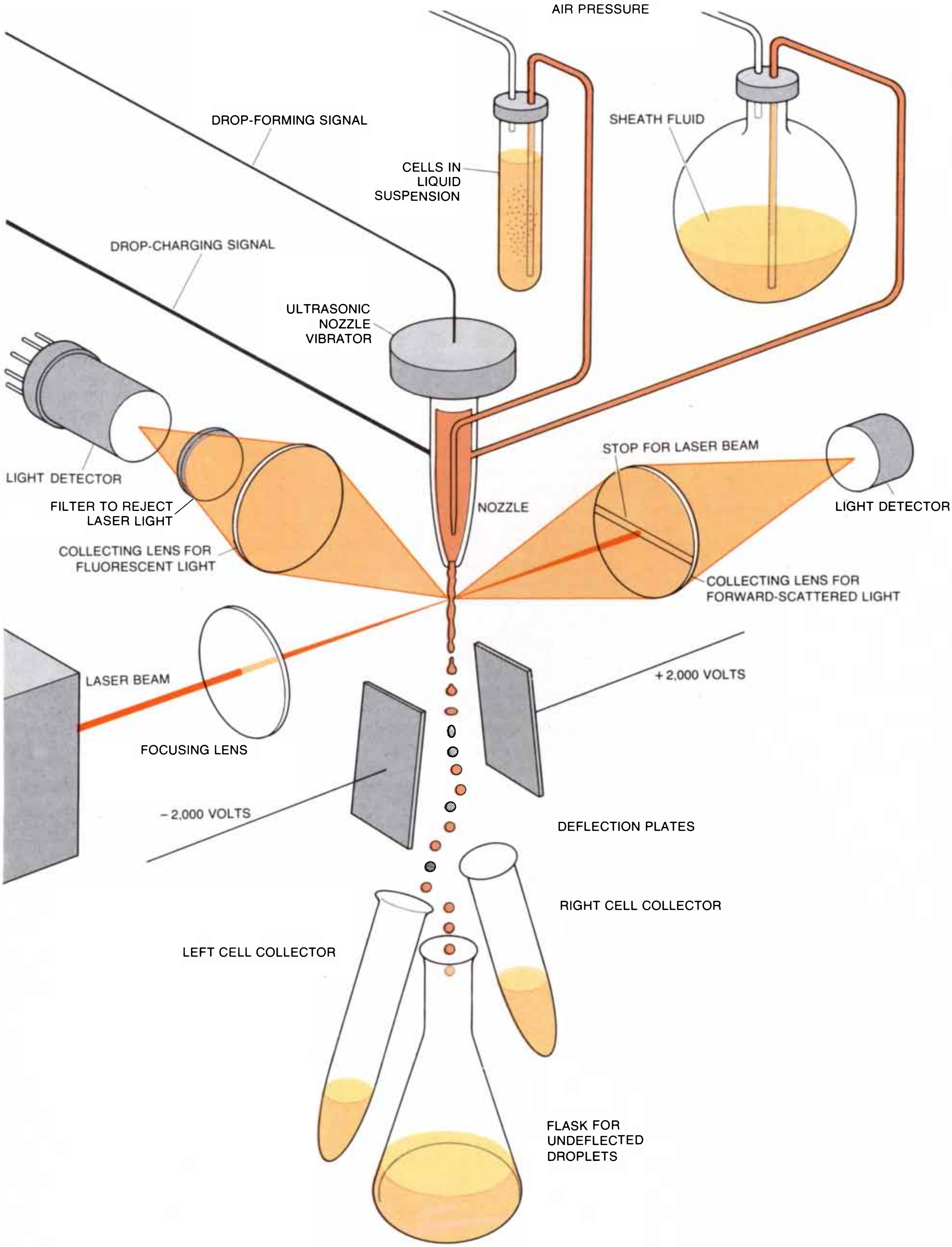
Immediately below the nozzle, before the droplets form, the jet is illuminated by the focused blue or green light of an argon-ion laser, operating at a wavelength selected to excite fluorescence in cells tagged with the appropriate fluorescent material. Some of the fluorescent light, filtered to remove the exciting wavelength, is focused onto a photomultiplier tube, which generates an electrical signal proportional to the number of fluorescent molecules on each cell. A second signal, related to the volume of the cell, is generated by detecting the light scattered forward out of the illuminating beam by the passing cell. The signals are processed, delayed and combined to produce other electric pulses, which serve to charge the liquid stream exactly when the droplet containing a desired cell is forming [*see illustration on page 113*].

Droplets that separate from the stream while it is charged retain their charge. The droplets, still traveling at the velocity of the jet from which they formed (about 10 meters per second), are directed between two charged plates that establish a constant electrostatic field across their path. Charged droplets are deflected appropriately, whereas uncharged droplets continue on their original course.

The deflection principle we employ is similar to that of the familiar cathode-ray tube but with an important difference. In the cathode-ray tube the electrons all have identical charges and are deflected by varying the field through which they pass. The deflecting force on all the electrons in the field at a given instant is in the same direction, and neighboring electrons cannot be directed along different paths. That factor limits the minimum response time of the cathode-ray tube to the comparatively short time required for electrons to pass through the deflection system.

In the cell sorter, on the other hand, the droplets have different charges. Moreover, they also have a much lower ratio of charge to mass than electrons, and hence if they are to be sufficiently deflected, they must spend a longer time (about three milliseconds) in the deflecting field. In a varying-field system the field could not be changed more rapidly than about every three milliseconds, or 333 times per second. By controlling the droplet's charge instead, and by specifying two droplets per sorted cell, the deflection direction can be changed as many as 20,000 times per second, an improvement in speed of about sixtyfold over a high-speed varying-field system.

The electrical signals generated by the two light detectors are processed by comparing the signals with preset amplitude limits that define cells by class. Upper and lower limits for the amplitudes of both signals can be set independently in order to



MAIN COMPONENTS of the fluorescence-activated cell-sorting machine are depicted in this schematic diagram. The cells are kept near the axis of the effluent jet by forcing them under pressure through a nozzle into the center of a stream of cell-free sheath fluid. The nozzle assembly is vibrated axially at 40,000 cycles per second, breaking the jet into 40,000 droplets per second. The detected fluorescent light generates an electrical signal that is proportional to the number of fluorescent molecules on each cell. The signal that is generated by detecting the forward-scattered light is related to the volume of the passing cell. The drop-charging signal is triggered by means of an electrical system that processes the signals that are received from two light detectors (*see illustration on opposite page*).

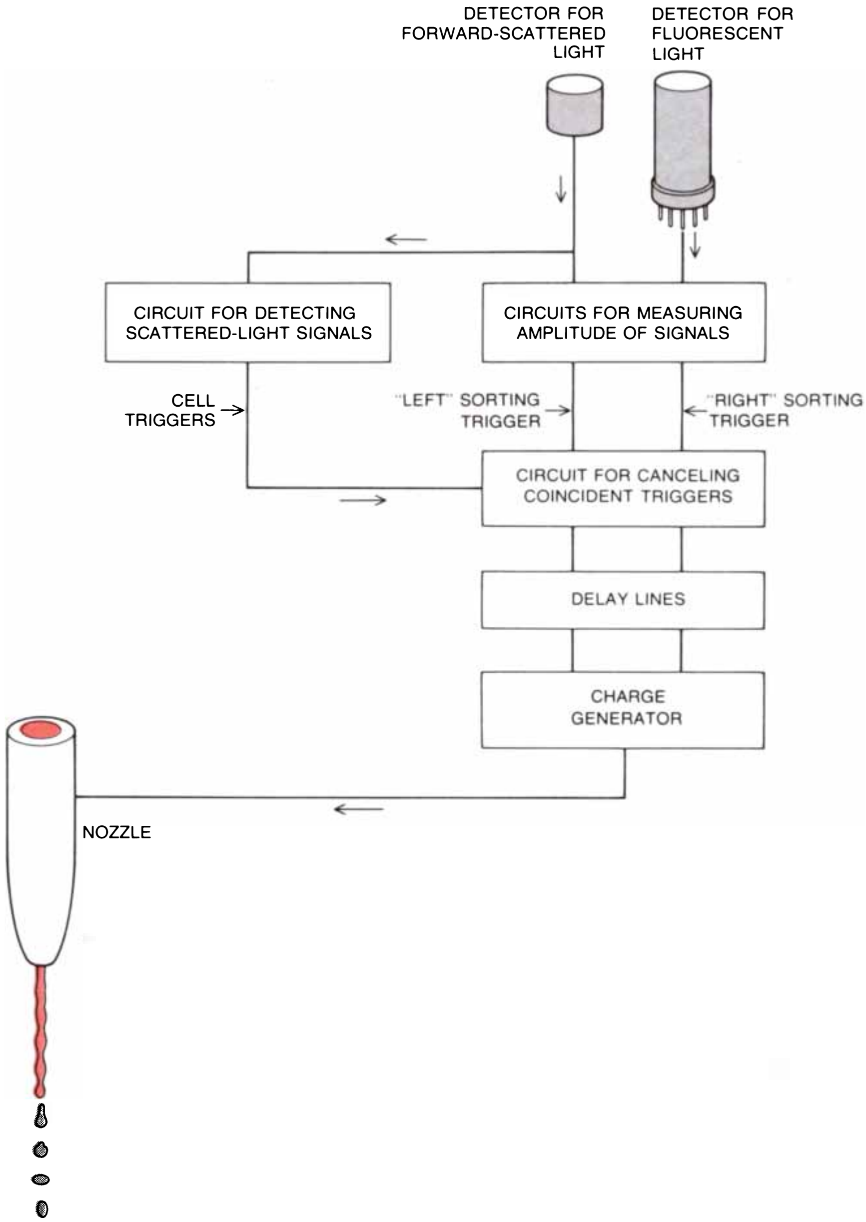
define two cell fractions to be sorted from the input sample.

Each cell that satisfies the criteria for sorting actuates a "left" or "right" sorting trigger. To ensure that only the desired cells are deflected a control device checks for adequate time separation between the cells that actuate sorting triggers and the unwanted cells, and cancels those triggering events that correspond to cells too closely spaced to separate. The triggering events are then delayed until the cell arrives at the point where droplets form. The charge generator is switched on just before the desired cell enters a droplet, charging the entire stream, and is switched off after the droplet separates, marking the droplet with a trapped positive or negative charge, which remains unchanged during the droplet's journey to the collecting tube. To be certain of charging the right droplet, at least two droplets are charged for each cell to be sorted. A third reservoir, the contents of which are usually discarded, collects undeflected droplets (which are empty or contain unwanted cells, debris and cells too closely spaced to separate).

At the typical sorting rate of 5,000 per second (approximately 18 million cells per hour) the fluorescence-activated cell sorter yields separated samples with a purity of between 90 and 99 percent. The viability of the cells is not affected by their passage through the instrument; in fact, the percent of living cells in the input sample can actually be increased by setting the scattered-light threshold to exclude dead cells. The temperature of the input and output reservoirs is controlled, and all components of the instrument that are in contact with cells can be sterilized. The duration of a run, which is determined by the number of separated cells required and the stability of the biological material, often lasts for up to eight hours.

The analytical capabilities of the fluorescence-activated cell sorter make it a powerful tool for measuring the distribution of cells in a given sample according to either size or fluorescence. The range for fluorescence detection extends well below the usual detection level in the fluorescence microscope. Cells that differ by as little as 5 percent in fluorescence can be differentiated. Furthermore, by appropriate manipulation of the thresholds it is possible to determine the size of cells characterized by a certain level of fluorescence or the level of the fluorescence of cells within a particular size range. These analyses have been greatly facilitated by coupling the cell sorter to a computer that stores, manipulates, displays and plots the accumulated data.

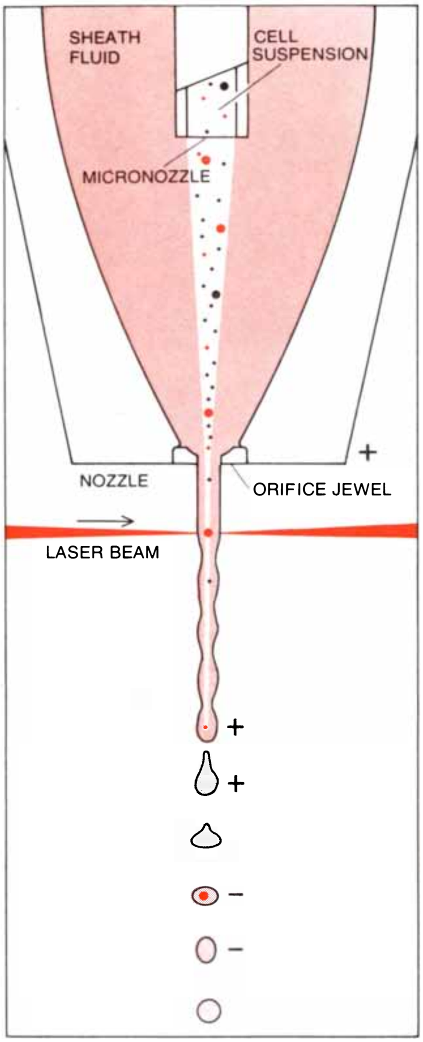
Of course, analysis and separation go hand in hand. Without the analytical ability of the cell sorter the choice of separation thresholds would be an excursion into the unknown; conversely, without the ability to separate subgroups of cells identified by analysis the correlation of function with the identified characteristics would be extremely difficult if not impossible.

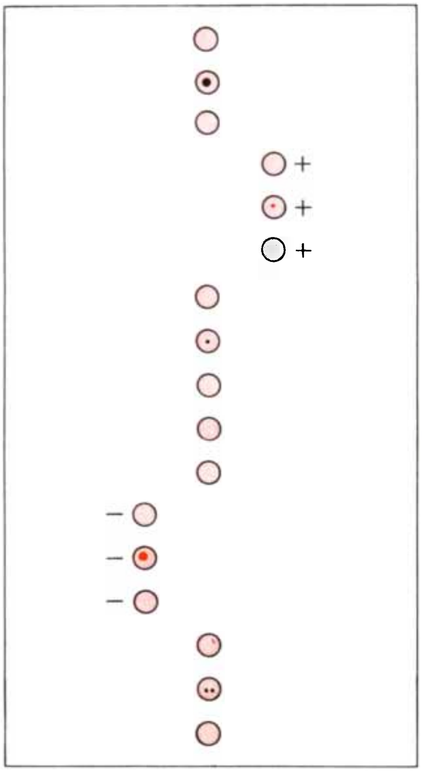


SIGNALS FROM LIGHT DETECTORS are processed and combined to produce electric pulses called sorting triggers, which are used to charge the fluid jet exactly when the droplet containing a desired cell is forming. Preset amplitude limits can be established for both signals in order to define two cell fractions to be sorted from the input sample. Triggering events that correspond to cells too closely spaced to separate are canceled. To make certain of charging the droplet containing the cell, at least two droplets are charged for each cell to be sorted.

By examining the computer-generated record of the distribution of cells as a function of the amount of fluorescence per cell or the amount of light scattered per cell the investigator can often tell whether the cell suspension under study consists of more than one resolvable population of cells. In addition, because the fluorescence measurements and the forward-scattered light measurements are made simultaneously for each cell, the investigator can selectively determine the fluorescence distribution of those cells that fall above or below a given size threshold. Alternatively he can determine the size distribution of those cells that generate fluorescence signals above or below a given fluorescence threshold. Thus, for example, the existence of a population of "medium-sized bright cells" can be recognized by appropriate manipulations of upper and lower thresholds. Once such a population is recognized it can be separated and studied functionally.

This "gating" technique also makes it possible to identify and remove fluorescent artifacts caused by the presence of odd-sized bits of fluorescent "junk" and dead cells in the suspension. Dead cells, it turns out, scatter less light than live cells, and hence the dead cells can be resolved into a separate peak in the curve representing the forward-scattered light distribution [*see bottom illustration on page 115*]. This finding might be just a curiosity if it were





not for its great practical significance in actual sorting experiments. Dead cells often bind the fluorescein-tagged protein reagents we use to stain the surface structures of cells. Therefore the almost unavoidable presence of a small percentage of dead cells in a suspension increases "noise" in the experiment, and such cells must be eliminated, particularly when the number of viable fluorescent-tagged cells is small. If the forward-scattered light threshold is set above the range associated with dead cells, the cell sorter will ignore all fluorescence signals associated with dead cells and direct them to another collection vessel, thus enabling the experimenter to separate and analyze only live cells.

FACS-I and FACS-II machines, produced commercially by the Becton Dickinson Electronics Laboratory, are now installed in laboratories in Britain, Germany and the U.S. Since those machines have only recently gone into service, however, the work we shall describe below was done mostly in our laboratory with the prototype model of the FACS-I.

We have used our machine mainly to study the roles that different types of lymphocytes play in the immune response. These cells, found in the spleen, the lymph nodes, the bone marrow, the thymus gland and the blood, constitute the body's chief defense against infection and are responsible for, among other things, the rejection of organ transplants. Although lymphocytes are structurally quite similar, they are functionally diverse and have been shown to fall into several distinct subpopulations. With the fluorescence-activated cell sorter we have been able to confirm some fundamental hypotheses about the functions of lymphocytes. Moreover, we have been able to add new information obtainable only with a method that enables one to separate and directly measure certain surface characteristics of these cells.

One major group of lymphocytes, called *B* cells (because of their origin in the bone marrow), includes precursor cells that may be triggered by an invading pathogen (or some other foreign substance) to divide and differentiate, forming cells that release large numbers of antibody molecules, which in turn react with the triggering antigen. The triggering process usually also requires the participation of lymphocytes belonging to another group, called *T* cells (because they originate in the thymus gland).

ENLARGED VIEWS of tip of nozzle (*top*) and region downstream near deflection plates (*bottom*) reveal how cells can be sorted into two classes depending on their size and their fluorescence. At instant shown in top drawing nozzle and liquid stream have positive charge, since "small bright cell" in droplet that is about to break off is to be sorted into "right" fraction. The undeflected droplets are either empty or contain unwanted cells, debris or cells too closely spaced to separate. Relative differences in cell sizes have been exaggerated in these drawings. The diameter of a cell is about a tenth of the diameter of a droplet.

In a typical immune response to an antigen an animal will produce a large number of different kinds of structurally similar antibody molecules, which differ only in the amino acid sequence of a small part of the antibody molecule. These structurally unique regions constitute the combining sites that make antibodies capable of reacting specifically with the triggering antigen. In general the antibody is highly specific for the antigen and will combine poorly if at all with other molecular structures. Similar antigenic structures do "cross-react" with antibodies elicited by one or the other structure, but even a very small modification of an antigen, such as the substitution of a single amino acid or a single sugar in a large protein or carbohydrate molecule, can be distinguished by antibody populations.

The potential for responding to antigens by producing antibody molecules with a wide range of combining-site structures has been traced to the precursor *B*-cell population. Each mature antibody-producing cell gives rise to only a single species of antibody molecule with a unique combining site that is reactive with the stimulating antigen. Therefore the interaction of the antigen and the precursor cells must be responsible for determining what antibody molecules will be produced.

According to current theory, that decision results from the selective triggering of precursor *B* cells, each of which is already committed to the production of a unique combining site. The combining-site structure, incorporated in an antibody molecule, is manufactured in small amounts by the committed precursor and is exposed on the surface membrane of the cell. When the appropriate antigen comes in contact with the surface antibody, it is firmly bound, and the precursor cell is triggered (with the help of a cooperating *T* cell) to divide and differentiate.

The combining site is at one end of the antibody molecule; the other end defines a group of physiologically important characteristics of the antibody. During differentiation the progeny of a given *B* cell remain committed to the production of the combining-site end of the molecule found on the precursor cell, but they may switch the other end to one of several different structures.

The concept that antibody production results from the combination of an antigen with a specific combining site on a precursor cell was first postulated by Paul Ehrlich around the turn of the century. Succeeding generations of immunologists have refined and modified the concept to include the commitment of the precursor cell to a unique specificity that reflects the structure of the combining-site end of the molecule and the selective triggering by the antigen. By isolating specific precursor cells with the fluorescence-activated cell sorter we have now been able to obtain direct evidence supporting the predictions of this commitment theory.

Before we did this work a number of laboratories had established that precursor cells with a given kind of surface-antibody specificity could be depleted from lymphoid-cell populations by allowing the cells to react with the appropriate target antigen. Since the precursor cells could not be recovered and tested directly in these studies, however, we decided to attack the problem with the FACS-I.

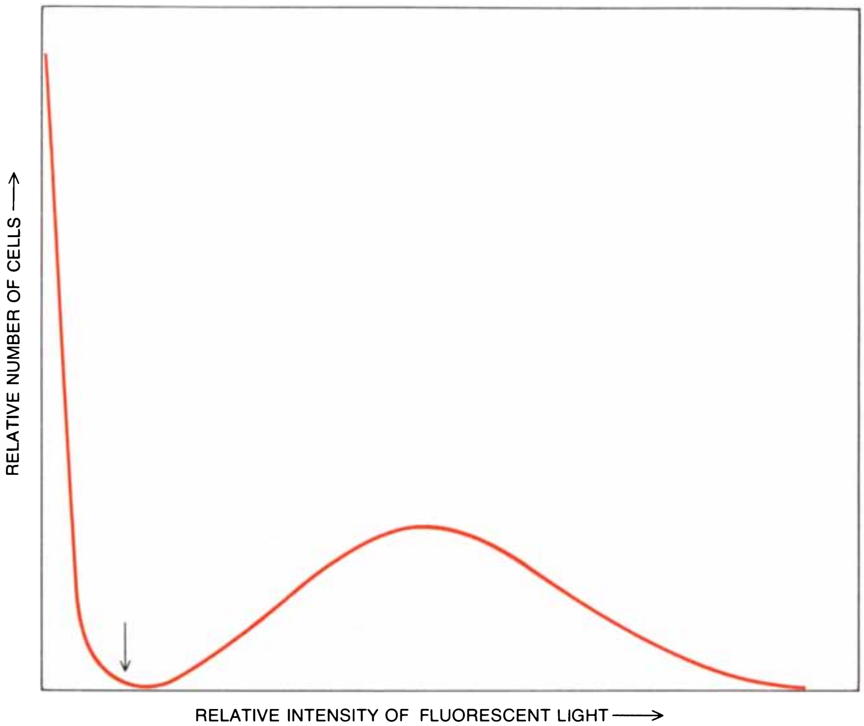
Michael Julius and Tohru Masuda, then working in our laboratory, isolated a subpopulation of mouse-spleen cells that bound a given antigen. They showed that this population of cells was highly enriched for the precursors of cells that produced antibody reactive with the antigen.

For the experiment the particular antigen chosen (a hemocyanin obtained from the keyhole limpet, a marine gastropod) was tagged with fluorescein and incubated with the cells. The cells were then washed free of excess antigen and passed through the cell sorter to separate those cells that bound the hemocyanin from those that did not. The two resulting cell fractions, labeled fluorescence-positive and fluorescence-negative, were injected into mice along with the appropriate cooperative *T* cells and fresh unlabeled antigen to allow the precursor cells to mature to form antibody-producing cells. An unrelated antigen was also injected into some of the mice to provide a control for testing the specificity of the separated precursor cells.

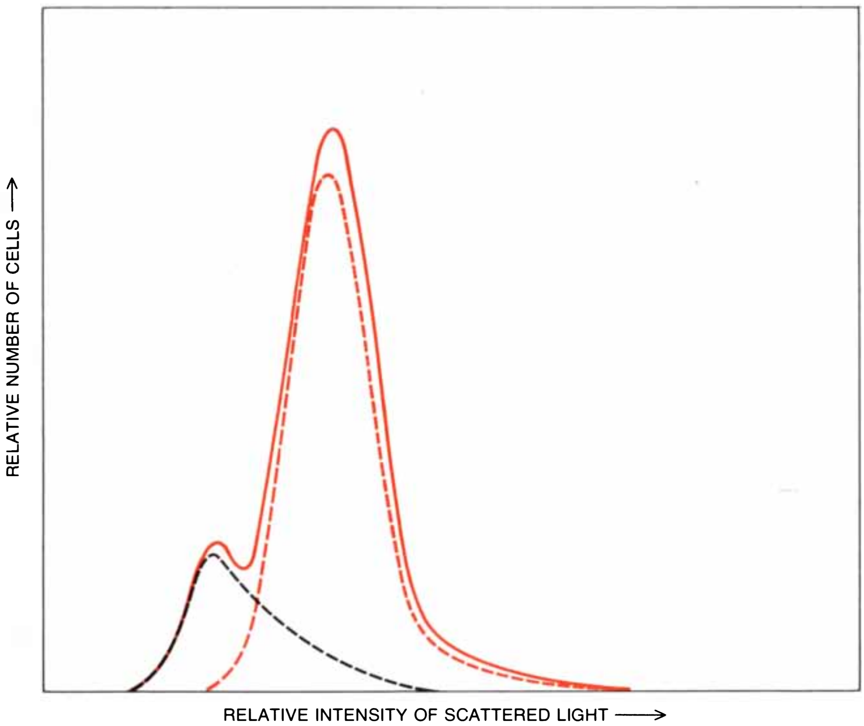
The results showed that the mouse recipients of the hemocyanin-binding, or fluorescent, cells manufactured antibody reactive with hemocyanin, whereas the recipients of the nonfluorescent cells did not. The latter mice, however, did produce antibody to the unrelated antigen, whereas the fluorescent cells could not. Thus the specificity of the antigen-combining site produced by the precursor cell determined the specificity of the progeny's antibody-producing cells.

Another study supporting the commitment hypothesis was done as a joint project involving our laboratory and workers in the laboratory of John J. Cebra at Johns Hopkins University. Working with genetically determined structural variations found in rabbit antibodies near the part of the molecule that incorporates the combining site, Patricia Jones of Johns Hopkins discovered a strict correlation between the antibody type of the precursor cell and that of the producer cell. The structural differences she worked with, called allotypes, are easily detectable by antibodies that react specifically with the allotypic structure of the antibody molecule.

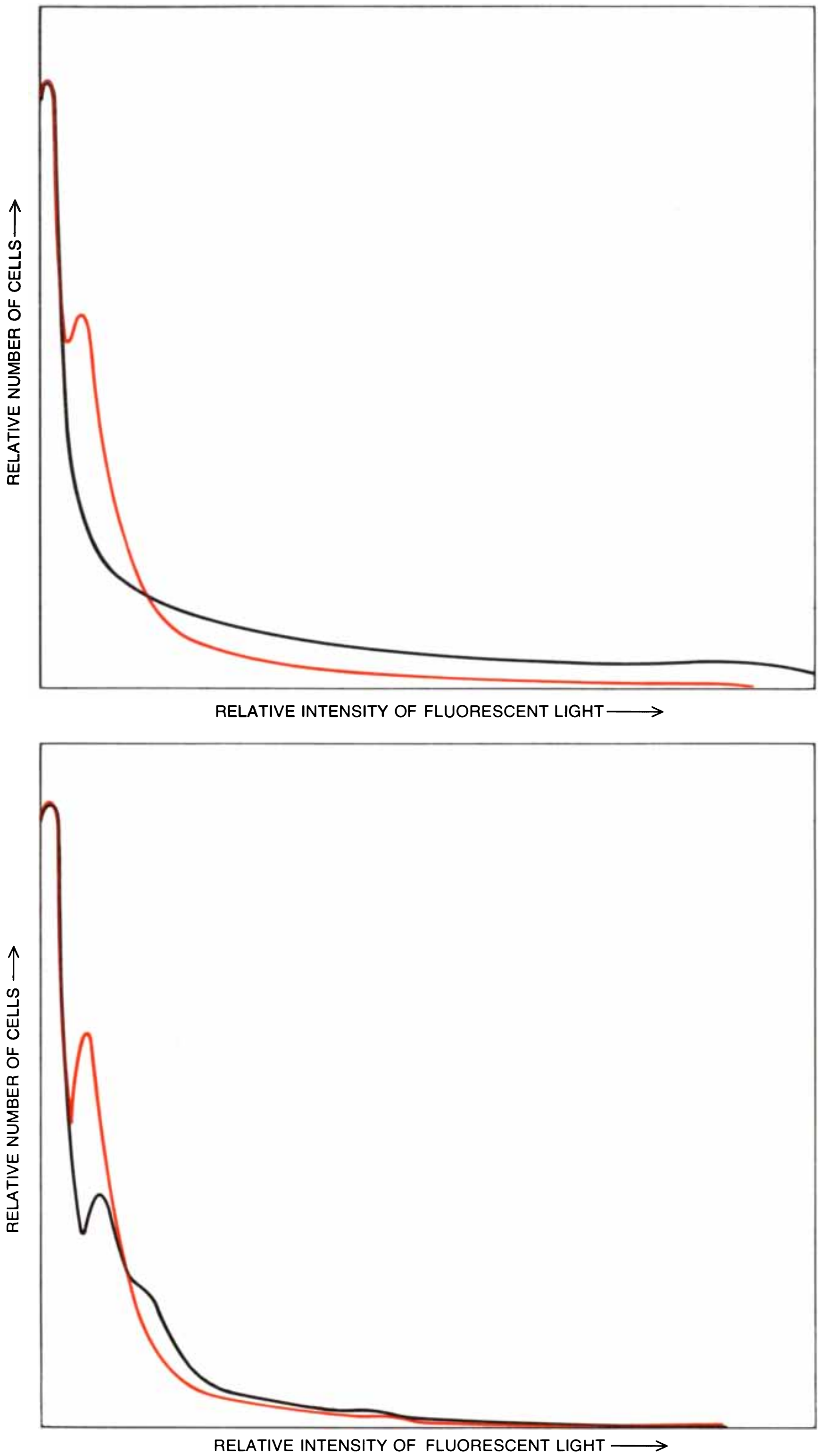
In studies of this kind the separation of lymphocytes was based on minute qualitative differences in the structure of surface-membrane components. The fluorescence-activated cell sorter has also been brought into play to separate cells where quantitative differences in the amount of surface antigens distinguish the subpopulations. A good example of that ability is the definition of thymus-derived-lymphocyte (*T*-cell) sub-



ONE WAY to distinguish live cells from dead cells in a given cell population by means of the fluorescence-activated cell sorter is represented by this graph. The data were obtained from a study in which a sample of human spleen cells was treated with fluorescein diacetate, a nonfluorescent substance that is converted to fluorescent fluorescein after entering a cell. Since live cells retain fluorescein but dead cells do not, the two fractions can be separated by adjusting a threshold setting (*arrow*) that divides the "bright" cells (*right*) from the "dark" ones (*left*).



ANOTHER WAY to distinguish live cells from dead cells is illustrated here for the same fluorescein-diacetate-stained spleen-cell suspension. The solid colored curve shows the distribution of all the cells in the sample according to how much light they scatter forward out of the laser beam. The broken colored curve shows the "fluorescence-gated" scattered-light distribution, which is composed only of fluorescent, or live, cells. The broken black curve shows the scattered-light distribution for nonfluorescent, or dead, cells. Obviously dead cells scatter less light and can be resolved into a separate peak in the overall scattered-light distribution curve.



NEW ANTISERUM found by workers at the Roswell Park Memorial Institute in Buffalo to contain antibodies that are reactive with antigens on the surface of certain cells in the thymus gland and spleen of mice was analyzed by Robert Stout at Stanford, using the FACS-I. He was able to show that the antiserum actually contained antibodies reactive with two independent cell-surface antigens, one of which (called *ThB*) was present on two cell types (spleen cells and thymus cells), whereas the other (called *ML*2) was present on only one of the two types (spleen cells). The FACS-I compared the amount of fluorescence on spleen cells incubated either with the intact serum, containing both anti-*ThB* and anti-*ML*2 antibody (*black curves*), or incubated with a specific anti-*ML*2 reagent prepared by absorbing the intact serum with thymus cells that remove the anti-*ThB* antibody (*colored curves*). The findings, which are charted here for three-week-old animals (*top graph*) and three-month-old animals (*bottom graph*), showed that the amount of *ThB* found on the spleen cells decreased progressively as the animal aged.

populations. In experiments exploiting the fluorescence-activated cell sorter workers in our laboratory have developed major evidence for the sequence of *T*-cell development in the thymus gland and for the identification of *T*-cell subgroups in the spleen and lymph nodes.

So far we have discussed the fluorescent-antibody preparations used to detect the various cell-surface antigens as if immunized animals normally make antibody only against the desired determinant structure. Actually antiserums prepared against such antigens frequently contain several populations of antibodies, each of which reacts with an independent antigenic structure. Therefore each antiserum employed was carefully examined for irrelevant antibodies and if necessary was purified so that the final reagent was specific for the antigen under study.

Occasionally, in spite of the multitude of tricks developed by immunologists over the years, important antiserums defy resolution into specific antibody populations and must be regarded as mixed reagents. We shall discuss one such case, where the FACS-I, because of its ability to detect the amount of fluorescent antibody bound to cells, was able to measure quantitative and qualitative differences in surface antigens of lymphocyte populations during development even though the antiserum detecting the antigens was only partially resolved.

Robert Stout, working in our laboratory, used the FACS-I to study an interesting antiserum obtained from M. Yutoku, Allan L. Grossberg and David Pressman of the Roswell Park Memorial Institute in Buffalo. The Roswell Park workers had shown that the antiserum in question contained antibodies reactive with a new type of antigen present on the surface of about half of the lymphocytes in the thymus gland (thymocytes) and also on some spleen cells in mice. Stout confirmed the original observation and went on to show that the antiserum contained antibodies reactive with two independent surface antigens, one of which (called *ThB*) was present on thymocytes and *B* cells, whereas the other (called *ML*2) was present only on *B* cells. (Neither antigen was present on peripheral *T* cells.) Furthermore, he showed that the amount of *ThB* found on *B* cells decreased progressively as the animal aged.

In order to demonstrate the developmental changes, Stout compared the amount of fluorescence on spleen cells incubated either with intact serum, containing anti-*ThB* and anti-*ML*2, or incubated with a specific anti-*ML*2 reagent prepared by absorbing the intact serum with *T* cells that remove the anti-*ThB* antibody. The findings showed that spleen cells from three-week-old animals and three-month-old animals bind about the same amount of fluorescent antibody when they are incubated with the specific anti-*ML*2 serum. Incubation of cells with the intact serum, containing antibody to both *ThB* and *ML*2, was found to increase the amount of fluorescence per cell, indicating the presence of *ThB* and *ML2* on the same cells. The increase, however, is much greater with cells taken from the three-week-old animals. The broad, flat distribution of fluorescence seen with these cells suggests that they carry variable amounts of *ThB*, sometimes up to five times the amount present on cells from older animals [*see illustration on opposite page*]. With this method of analysis we are now examining the developmental patterns for cells carrying these antigens.

The applications of our technique that we have cited have been chosen to give some idea of the kinds of problem a fluorescence-activated cell sorter can be expected to solve. Similar investigations are in progress in a number of laboratories on a wide range of cell types, including normal and leukemic blood cells, cells in the nervous system, cultured normal cells and cultured tumor cells.

Recent work on human leukemias done with the FACS shows promise for substantially improving the diagnosis and treatment of this group of diseases. Investigators at University College London and the Harvard Medical School report that with the aid of a newly developed battery of fluorescent reagents they can now quickly identify many of the different types of leukemia from small samples of blood taken from patients. Because FACS analysis requires only a blood sample rather than the costlier and more painful bone-marrow sample currently used to identify leukemia, the new approach may enable the physician both to diagnose the disease more rapidly and to monitor therapy at frequent intervals, watching for the disappearance or reappearance of leukemic cells. It may also lead to the development of new methods for treating specific types of leukemia and even "tailor-made" treatments designed to suit each individual patient.

Fluorescence sorting and analysis is also being used to attack other problems of immediate clinical relevance, such as measuring the DNA content of individual chromosomes in suspension with the aim of developing rapid ways to define the chromosome constitution (karyotype) of cells for genetic analysis and attempting to automate the counting of different types of white blood cells, a very common test in the clinical laboratory. At Stanford we are now working on a safe screening method for prenatal diagnosis of genetic disease based on isolating and testing the small number of fetal cells reportedly present in maternal blood quite early in pregnancy.

As these few examples show, the potential applications of a technique for separating fluorescence-labeled cells range over virtually all biology and medicine. Future directions will be determined by the ingenuity of investigators not only in finding ways to put fluorescent labels on cells but also in finding ways to make use of the separated cells.

Introducing the kind of car you thought you could never own.

Most men dream of owning a high-performance, classic grand touring car. But that's all they do. They dream. And for good reason.

Cars of this type are generally expensive to buy and even more expensive to keep up. And what they don't take out of you in dollars, they take out in time and aggravation.

Why do we bring up the problems of these cars? Because we are happy to offer one with none of these problems. Possibly the first car of its kind within your reach. The Lancia Beta.

Lancia? What's a Lancia?

You may not know the name Lancia. Not too many have ever been to America. But to the car aficionado the name is legend.

Lancias have been built since 1906. Over the years, they've become synonymous with technical innovation, to say nothing of a long history of Grand Prix racing, including some historic triumphs over Ferrari. In fact, even today, Lancia is the Manufacturers' World Rally Champion.

The new Lancia Beta.

The car we're introducing to America is called the Lancia Beta. It is a highly refined touring car in the original tradition.

In performance, the Beta is nothing short of extraordinary. It is remarkably responsive through each of its five forward gears. And it corners like its racing forebears.

The car is meticulously finished. The doors, the trunk, all the seams fit. The paint is burnished to a level you simply don't see on mass-produced vehicles. The upholstery on the Coupe is pure leather.

The pleasure without the pain.

For all its exceptional performance and handling, the Beta is not the least bit nervous or temperamental. It doesn't require a mechanic as a permanent houseguest. In fact, it doesn't demand any more from you than the car you now own.

The good news.

For all its workmanship, its performance, and its prestige, you might expect the Beta Coupe to cost $10,000, $12,000, maybe $15,000.

Happily, you'd be very wrong.

The base price of the Coupe is $7,510.* Even more surprising, our 4-door Saloon is about $1,000 less.

All of which means that maybe, just maybe, you can own the kind of car you thought you could never own.

LANCIA
A Division of Fiat

For the name of your nearest Lancia dealer, call toll-free: 800-447-4700. In Illinois, call: 800-322-4400.

*1976 Manufacturer's suggested retail price POE. Inland transportation, dealer preparation and local taxes additional. Car rental, leasing, and overseas delivery arranged through your participating dealer. Lancia of America, 155 Chestnut Ridge Road, Montvale, N.J. 07645.

MATHEMATICAL GAMES

On the fabric of inductive logic, and some probability paradoxes

by Martin Gardner

"The universe, so far as known to us, is so constituted, that whatever is true in any one case, is true in all cases of a certain description; the only difficulty is, to find what description."

—JOHN STUART MILL, *A System of Logic*

Imagine that we are living on an intricately patterned carpet. It may or may not extend to infinity in all directions. Some parts of the pattern appear to be random, like an abstract expressionist painting; other parts are rigidly geometrical. A portion of the carpet may seem totally irregular, but when the same portion is viewed in a larger context, it becomes part of a subtle symmetry.

The task of describing the pattern is made difficult by the fact that the carpet is protected by a thick plastic sheet with a translucence that varies from place to place. In certain places we can see through the sheet and perceive the pattern; in others the sheet is opaque. The plastic sheet also varies in hardness. Here and there we can scrape it down so that the pattern is more clearly visible. In other places the sheet resists all efforts to make it less opaque. Light passing through the sheet is often refracted in bizarre ways, so that as more of the sheet is removed the pattern is radically transformed. Everywhere there is a mysterious mixing of order and disorder. Faint lattices with beautiful symmetries appear to cover the entire rug, but how far they extend is anyone's guess. No one knows how thick the plastic sheet is. At no place has anyone scraped deep enough to reach the carpet's surface, if there is one.

Already the metaphor has been pushed too far. For one thing, the patterns of the real world, as distinct from this imaginary one, are constantly changing, like a carpet that is rolling up at one end while it is unrolling at the other end. Nevertheless, in a crude way the carpet can introduce some of the difficulties philosophers of science encounter in trying to understand why science works.

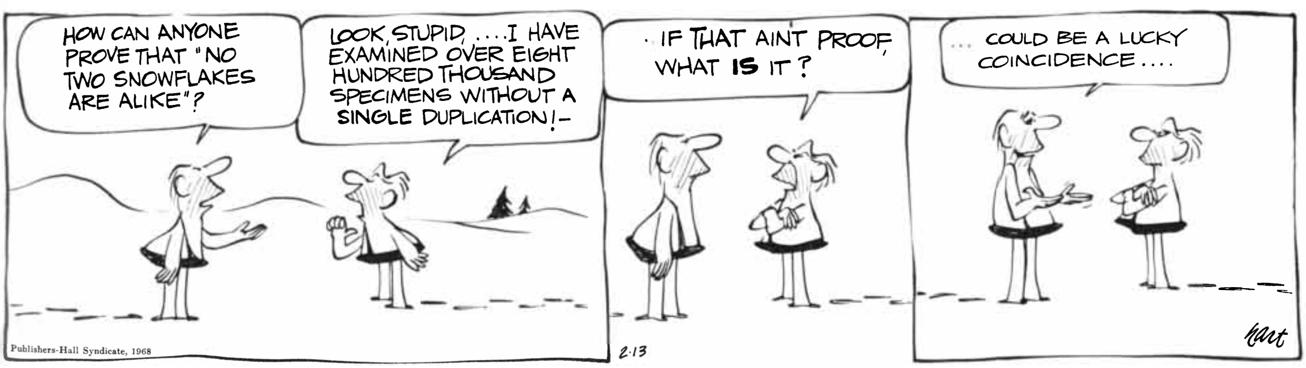
Induction is the procedure by which carpetologists, after examining parts of the carpet, try to guess what the unexamined parts look like. Suppose the carpet is covered with billions of tiny triangles. Whenever a blue triangle is found, it has a small red dot in one corner. After finding thousands of blue triangles, all with red dots, the carpetologists conjecture that all blue triangles have red dots. Each new blue triangle with a red dot is a confirming instance of the law. Provided that no counterexample is found, the more confirming instances there are, the stronger is the carpetologists' belief that the law is true.

The leap from "some" blue triangles to "all" is, of course, a logical fallacy. There is no way to be absolutely certain, as one can be in working inside a deductive system, what any unexamined portion of the carpet looks like. On the other hand, induction obviously works, and philosophers justify it in other ways. John Stuart Mill did so by positing in effect that the carpet's pattern has regularities. He knew this reasoning was circular, since it is only by induction that carpetologists have learned that the carpet is patterned. Mill did not regard the circle as vicious, however; and many contemporary philosophers (R. B. Braithwaite and Max Black, to name two) agree. Bertrand Russell, in his last major work, tried to replace Mill's vague "nature is uniform" with something more precise. He proposed a set of five posits about the structure of the world that he believed were sufficient to justify induction.

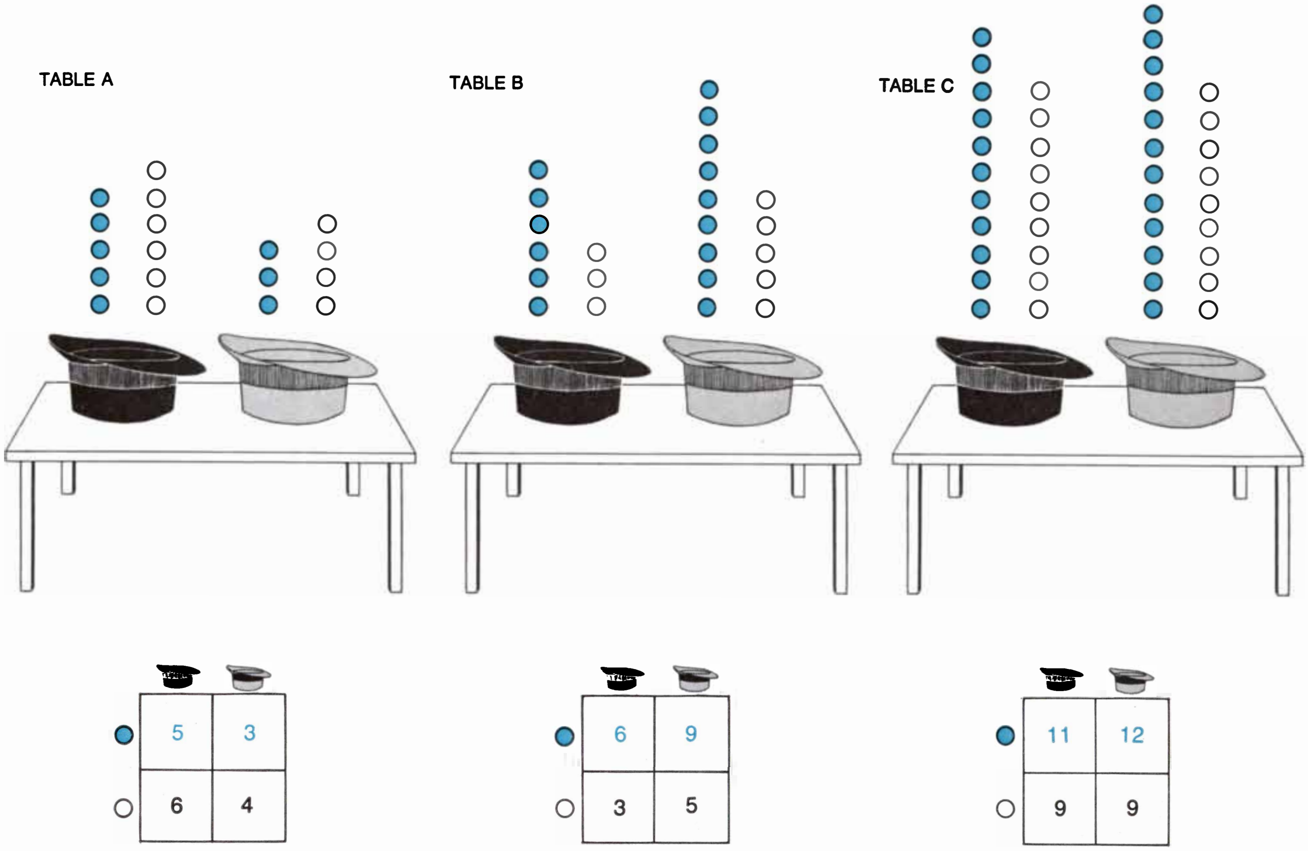
Hans Reichenbach advanced the most familiar of several pragmatic justifications. If there is any way to guess what unexamined parts of the carpet look like, Reichenbach argued, it has to be by induction. If induction does not work, nothing else will, and so science might as well use the only tool it has. "This answer is not fallacious," wrote Russell, "but I cannot say that I find it very satisfying."

Rudolf Carnap agreed. His opinion was that all these ways of justifying induction are correct but trivial. If "justify" is meant in the sense that a mathematical theorem is justified, then David Hume was right: there is no justification. But if "justify" is taken in any of several weaker senses, then, of course, induction can be defended. A more interesting task, Carnap insisted, is to see whether it is possible to construct an inductive logic.

It was Carnap's great hope that such a logic could be constructed. He foresaw a future in which a scientist could express in a formalized language a certain hypothesis together with all the relevant evidence. Then by applying inductive logic he could assign a probability value, called the degree of confirmation, to the hypothesis. There would be nothing final about that value. It would go up or down or stay the same as new evidence accumulated. Scientists already think in terms of such a logic, Carnap maintained, but only in a vague, informal



A cartoon comment on inductive reasoning



E. H. Simpson's reversal paradox

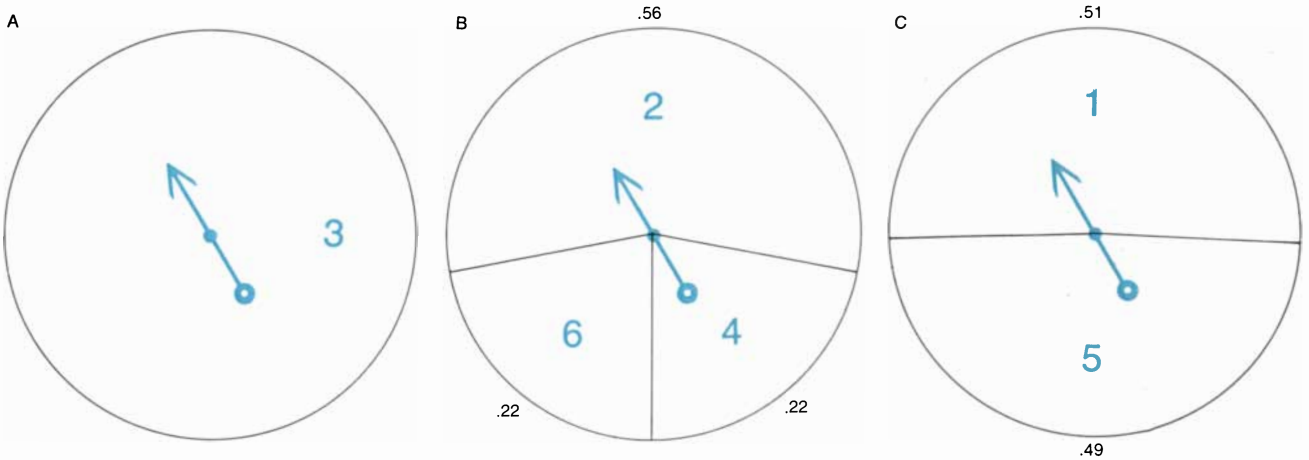
way. As the tools of science become more powerful, however, and as our knowledge of probability becomes more precise, perhaps eventually we can create a calculus of induction that will be of practical value in the endless search for scientific laws.

In Carnap's *Logical Foundations of Probability* and also in his later writings he tried to establish a base for such a logic. How successful he was is a matter of dispute. Some philosophers share his vision (John G. Kemeny for one) and have taken up the task where Carnap left off. Others, notably Karl Popper and Thomas S. Kuhn, regard the entire project as having been misconceived.

Carl G. Hempel, one of Carnap's admirers, has argued sensibly that before we try to assign quantitative values to confirmations we should first make sure we know in a qualitative way what is meant by "confirming instance." It is here that we run into the worst kinds of difficulty [see "Confirmation," by Wesley C. Salmon; SCIENTIFIC AMERICAN, May, 1973].

Consider Hempel's notorious paradox of the raven. Let us approach it by way of 100 playing cards. Some of them have a picture of a raven drawn on the back. The hypothesis is "All raven cards are black." You shuffle the deck and deal the cards face up. After turning 50 cards without finding a counterinstance the hypothesis certainly becomes plausible. As more and more raven cards prove to be black, the degree of confirmation approaches certainty and may finally reach it.

Now consider another way of stating the same hypothesis: "All nonblack cards are not ravens." This statement is logically equivalent to the original one. If you test the new statement on another shuffled deck of 100 cards, holding them face up and turning them as you deal, clearly each time you deal a nonblack card and it proves to have no raven on the back, you confirm the guess that all nonblack cards are not ravens. Since this is logically equivalent to "All raven



The three spinners for Colin R. Blyth's paradox cards are black," you confirm that also. Indeed, if you deal all the cards without finding a red card with a raven, you will have completely confirmed the hypothesis that all raven cards are black.

Unfortunately when this procedure is applied to the real world, it seems not to work. "All ravens are black" is logically the same as "All nonblack objects are not ravens." We look around and see a yellow object. Is it a raven? No, it is a buttercup. The flower surely confirms (albeit weakly) that all nonblack objects are not ravens, but it is hard to see how it has any relevance at all to "All ravens are black." If it does, it equally confirms that all ravens are white or any color except yellow. To make things worse, "All ravens are black" is logically equivalent to "Any object is either black or not a raven." And that is confirmed by any black object whatever (raven or not) as well as by any nonraven (black or not). All of which seems absurd.

Nelson Goodman's "grue" paradox is equally notorious. An object is "grue" if it is green until, say, January 1, 2000, and blue thereafter. Is the law "All emeralds are grue" confirmed by observations of green emeralds? A prophet announces that the world will exist until January 1, 2000, when it will disappear with a bang. Every day the world lasts seems to confirm the prediction, yet no one supposes that it becomes more probable.

To make matters still worse, there are situations in which confirmations make a hypothesis less likely. Suppose you turn the cards of a shuffled deck looking for confirmations of the guess that no card has green pips. The first 10 cards are ordinary playing cards, then suddenly you find a card with blue pips. It is the 11th confirming instance, but now your confidence in the guess is severely shaken. Paul Berent has pointed out a similar example. A man 99 feet tall is discovered. He is a confirming instance of "All men are less than 100 feet tall," yet his discovery greatly weakens the hypothesis.

Confirmations may even falsify a hypothesis. Ten cards with all values from the ace through the 10 are shuffled and dealt face down in a row. The guess is that no card with value n is in the nth position from the left. You turn the first nine cards. Each card confirms the hypothesis. But if none of the turned cards is the 10, the nine cards taken together refute the hypothesis.

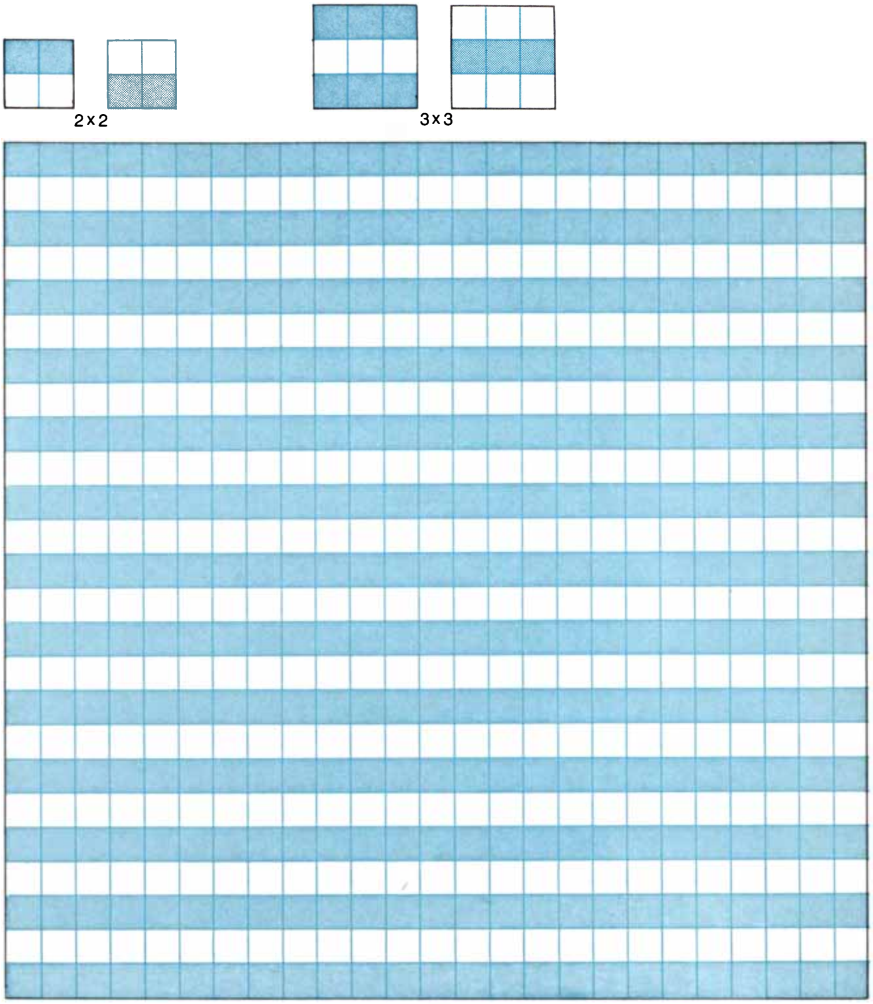
Carnap was aware of such difficulties. He distinguished sharply between "degree of confirmation," a probability value based on the total relevant evidence, and what he called "relevance confirmation," which has to do with how new observations alter a confirmation estimate. Relevance confirmation cannot be given simple probability values. It is enormously complex, swarming with counterintuitive arguments. In Chapter 6 of Carnap's *Logical Foundations* he analyzes a group of closely related paradoxes of confirmation relevance that are easily modeled with cards.

For example, it is possible that data will confirm each of two hypotheses but disconfirm the two taken together. Consider a set of 10 cards, half with blue backs and half with green ones. The green-backed cards (with the hearts and spades designated *H* and *S*) are Q*H*, 10*H*, 9*H*, K*S*, Q*S*. The blue-backed cards are K*H*, J*H*, 10*S*, 9*S*, 8*S*. The 10 cards are shuffled and dealt face down in a row.

Hypothesis *A* is that the property of being a face card (a king, a queen or a jack) is more strongly associated with green backs than with blue. An investigation shows that this is true. Of the five cards with green backs, three are face cards as against only two face cards with blue backs. Hypothesis *B* is that the property of being a red card (hearts or diamonds) is also more strongly associated with green backs than with blue. A second investigation confirms this. Three green-backed cards are red but there are only two red cards with blue backs. Intuitively one assumes that the property of being both red and a face card is more strongly associated with green backs than blue, but that is not the case. Only one red face card has a green back, whereas two red face cards have blue backs!

It is easy to think of ways, fanciful or realistic, in which similar situations can arise. A woman wants to marry a man who is both rich and kind. Some of the bachelors she knows have hair and some are bald. Being a statistician, she does some sampling. Project *A* establishes that 3/5 of the men with hair are rich but only 2/5 of the bald men are rich. Project *B* discloses that 3/5 of the men with hair are kind but only 2/5 of the bald men are kind. The woman might hastily conclude that she should marry a man with hair, but if the distribution of the attributes corresponds to that of the face cards and red cards mentioned in the preceding example, her chances of getting a rich, kind man are twice as great if she sets her cap for a bald man.

Another research project shows that 3/5 of a group of patients taking a certain pill are immune to colds for five years, compared with only 2/5 in the control group who were given a placebo. A second project shows that 3/5 of a group receiving the pill were immune to tooth cavities for five years, compared with 2/5 who got the placebo. The combined statistics could show that twice as many among those who got the placebo are free for five years from both



David A. Klarner's impossibility proof colds and cavities compared with those who got the pill.

A striking instance of how a hypothesis can be confirmed by two independent studies, yet disconfirmed by the total results, is provided by the following game. It can be modeled with cards, but to vary the equipment let us do it with 41 poker chips and four hats [*see top illustration on page 120*]. On table *A* is a black hat containing five colored chips and six white chips. Beside it is a gray hat containing three colored chips and four white chips. On table *B* is another pair of black and gray hats. In the black hat there are six colored chips and three white chips. In the gray hat there are nine colored chips and five white chips. The contents of the four hats are shown by the charts in the illustration.

You approach table *A* with the desire to draw a colored chip. Should you take a chip from the black hat or from the gray one? In the black hat five of the 11 chips are colored, so that the probability of getting a colored chip is 5/11. This is greater than 3/7, which is the probability of getting a colored chip if you take a chip from the gray hat. Clearly your best bet is to take a chip from the black hat.

The black hat is also your best choice on table *B*. Six of its nine chips are colored, giving a probability of 6/9, or 2/3, that you will get a colored chip. This exceeds the probability of 9/14 that you will get a colored chip if you choose to take a chip from the gray hat.

Now suppose that the chips from both black hats are combined in one black hat and that the same is done for the chips in the two gray hats [*see table C in top illustration on page 120*]. If you want to get a colored chip, surely you should take a chip from the black hat. The astonishing fact is that this is not true! Of the 20 chips now in the black hat, 11 are colored, giving a probability of 11/20 that you will get a colored chip. This is exceeded by a probability of 12/21 that you will get a colored chip if you take a chip from the gray hat.

The situation has been called Simpson's paradox by Colin R. Blyth, who found it in a 1951 paper by E. H. Simpson. The paradox has turned out to be older, but the name has persisted. Again, it is easy to see how the paradox could arise in actual research. Two independent investigations of a drug, for example, might suggest that it is more effective on men than it is on women, whereas the combined data would indicate the reverse.

One might imagine that such situations are too artificial to arise in statistical research. In a recent investigation to see if there was sex bias in the admissions of men and women to graduate studies at the University of California at Berkeley, however, Simpson's paradox actually turned up. Independent studies of admissions of men and of women in the fall of 1973 showed a positive sex bias against female applicants. Then when the data for men and women were combined, there was a small but statistically significant bias *in favor* of women (see "Sex Bias in Graduate Admissions: Data from Berkeley," by P. J. Bickel, E. A. Hammel and J. W. O'Connell in *Science*, Vol. 187, February 7, 1975, pages 398–404).

Blyth has invented another paradox that is even harder to believe than Simpson's. It can be modeled with three sets of cards or three unfair dice that are weighted to give the required probability distributions to their faces. We shall model it with the three spinners shown in the bottom illustration on page 120, because they are easy to construct by anyone who wants to verify the paradox empirically.

Spinner *A*, with an undivided dial, is the simplest. No matter where the arrow stops, it gives a value of 3. Spinner *B* gives values of 2, 4 or 6 with the respective probability distributions of .56, .22 and .22. Spinner *C* gives values of 1 or 5 with the probabilities of .51 and .49.

You pick a spinner; a friend picks another. Each of you flicks his arrow, and the highest number wins. If you can later change spinners on the basis of experience, which spinner should you choose? When the spins are compared in pairs, we find that *A* beats *B* with a probability of $1 \times .56 = .56$. *A* beats *C* with a probability of $1 \times .51 = .51$. *B* beats *C* with a probability of $(1 \times .22) + (.22 \times .51) + (.56 \times .51) = .6178$. Clearly *A*, which beats both of the others with a probability of more than 1/2, is the best choice. *C* is the worst because it is beaten with a probability of more than 1/2 by both of the others.

Now for the crunch. Suppose you play the game with two others and you have the first choice. The three spinners are flicked and the high number wins. Calculating the probabilities reveals an extraordinary fact. *A* is the worst choice; *C* is the best! *A* wins with a probability of $.56 \times .51 = .2856$, or less than 1/3. *B* wins with a probability of $(.44 \times .51) + (.22 \times .49) = .3322$, or almost 1/3. *C* wins with a probability of $.49 \times .78 = .3822$, or more than 1/3.

Consider the havoc this can wreak in statistical testing. Assume that drugs for a certain illness are rated in effectiveness with numbers 1 through 6. Drug *A* is uniformly effective at a value of 3 (spinner *A*). Studies show that drug *C* varies in effectiveness. Fifty-one percent of the time it has value 1 and 49 percent of the time it has value 5 (spinner *C*). If drugs *A* and *C* are the only two on the market and a doctor wants to maximize a patient's chance of recovery, he clearly chooses drug *A*.

What happens when drug *B*, with values and a probability distribution corresponding to spinner *B*, becomes available? The bewildered doctor, if he considers all three drugs, finds *C* preferable to *A*.

Blyth has an even more mind-blowing way of dramatizing the paradox. Every night a statistician eats at a restaurant that offers apple pie and cherry pie. He rates his satisfaction with each kind of pie in values 1 through 6. The apple pie is uniformly 3 (spinner *A*); the cherry varies in the manner of spinner *C*. Naturally the statistician always takes apple.

Occasionally the restaurant has blueberry pie. Its satisfaction varies in the manner of spinner *B*.

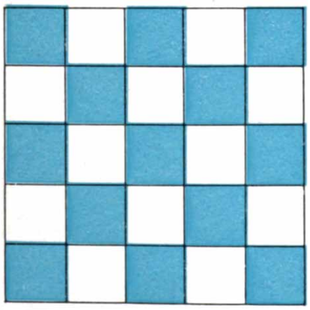
Waitress: "Shall I bring your apple pie?"

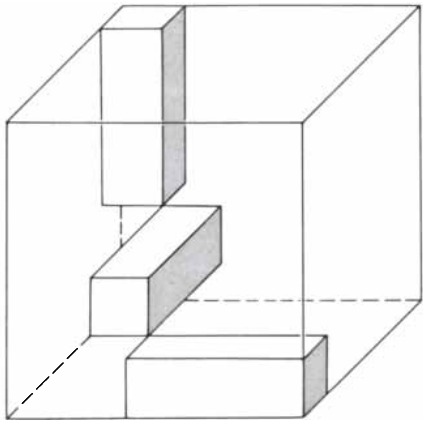
Statistician: "No. Seeing that today you also have blueberry, I'll take the cherry."

The waitress would consider that a joke. Actually the statistician is rationally maximizing his expectation of satisfaction. Is there any paradox that points up more spectacularly the kinds of difficulty Carnap's followers must overcome in their efforts to advance his program?

Readers were asked last month to prove that it is impossible to tile a square of side 25 with a mixture of 2×2 and 3×3 squares. David A. Klarner's proof begins by two-coloring the square with stripes [*see illustration on preceding page*]. No matter how a 2×2 square is placed, it will cover two colored cells and two white ones. Since the large square contains an excess of 25 colored cells, the excess will remain no matter how many 2×2 squares are placed or where they are placed. If the square can be tiled, we must therefore find a way to place a set of 3×3 squares so that they cover 25 more colored cells than white ones.

It is apparent that a 3×3 square, howev-





Key to John Horton Conway's cube puzzle

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

—*The Christian Science Monitor*

As an introduction to membership in the

BOOK-OF-THE-MONTH CLUB®

yours for only

$17.50 [Publisher's list price $90]

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

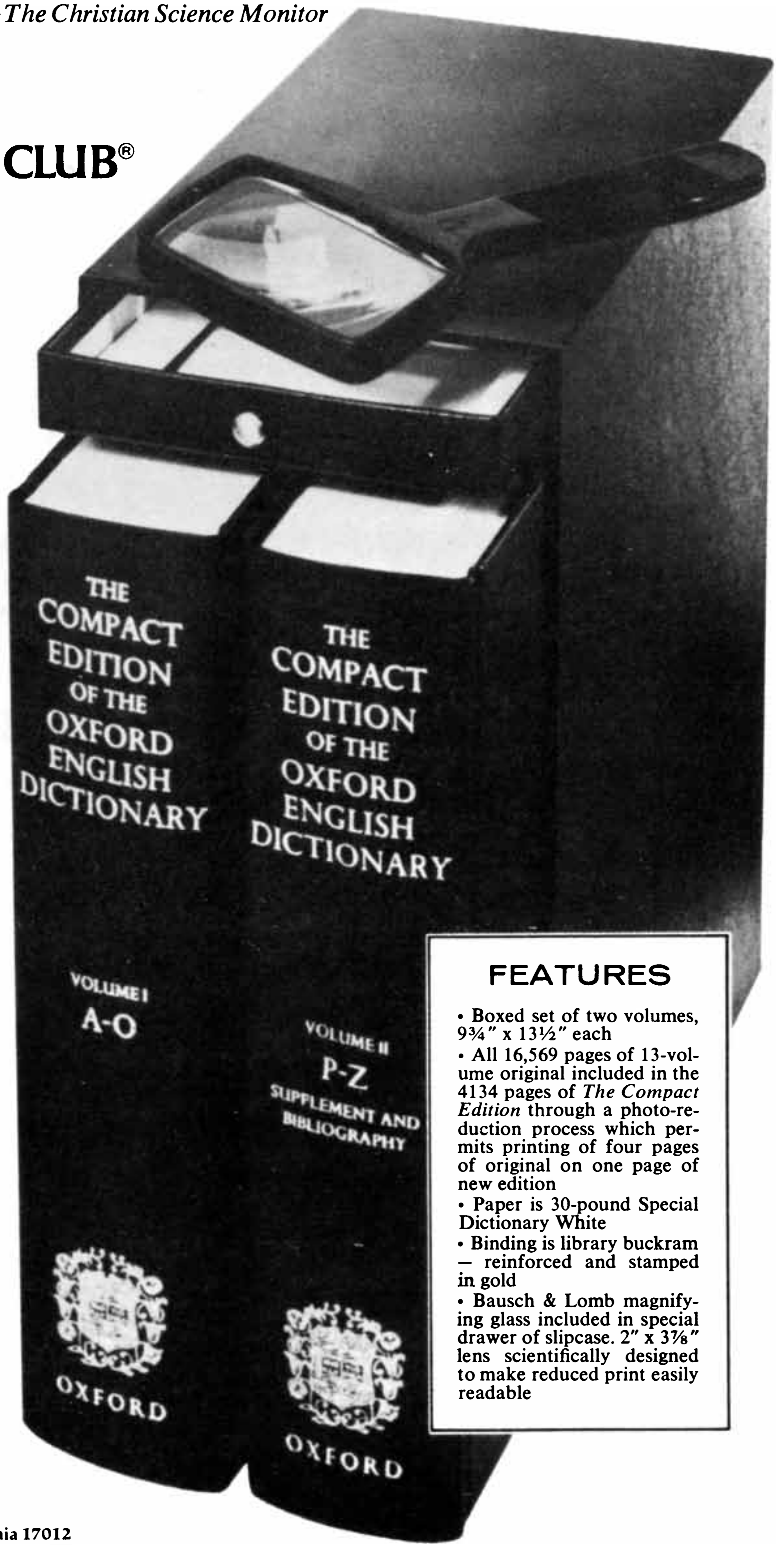
THE OXFORD ENGLISH DICTIONARY is generally regarded as the final arbiter of the English language. Until recently it had been available only as a thirteen-volume set, currently priced at $350. Now, through the combination of an ingenious method of micrographic reproduction and a fine Bausch & Lomb optical lens, every single one of its 16,569 pages, fifty million words and close to two milion illustrative quotations appears, in easily readable form, in *The Compact Edition.*

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

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

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

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



FEATURES

- Boxed set of two volumes, 9¾" x 13½" each
- All 16,569 pages of 13-volume original included in the 4134 pages of *The Compact Edition* through a photo-reduction process which permits printing of four pages of original on one page of new edition
- Paper is 30-pound Special Dictionary White
- Binding is library buckram – reinforced and stamped in gold
- Bausch & Lomb magnifying glass included in special drawer of slipcase. 2" x 3⅞" lens scientifically designed to make reduced print easily readable

BOOK-OF-THE-MONTH CLUB, INC., Camp Hill, Pennsylvania 17012 er placed, covers either six colored cells and three white ones or six white cells and three colored ones. In each case the difference is 3. But 25 is not a multiple of 3, and so there is no way the 3×3 squares can cover an excess of 25 colored cells. Therefore the tiling is impossible.

Klarner has shown in general that if a rectangle can be tiled at all by squares of $a \times a$ and $b \times b$, it can always be split into two rectangles (one possibly empty), one of which is tilable by $a \times a$ squares alone and the other tilable by $b \times b$ squares alone.

Last month's second task was to build a $5 \times 5 \times 5$ cube with 18 specified bricks. In working on puzzles of that type one tends to place large pieces first, then try to fit smaller ones into the gaps. In this case it is a fatal strategy. If you worked on the puzzle by trial and error, you probably found that if you left the three small bricks to the last, they never fitted. The puzzle is difficult precisely because those three pieces must be in a unique configuration, and it is unlikely that you would hit on it by chance.

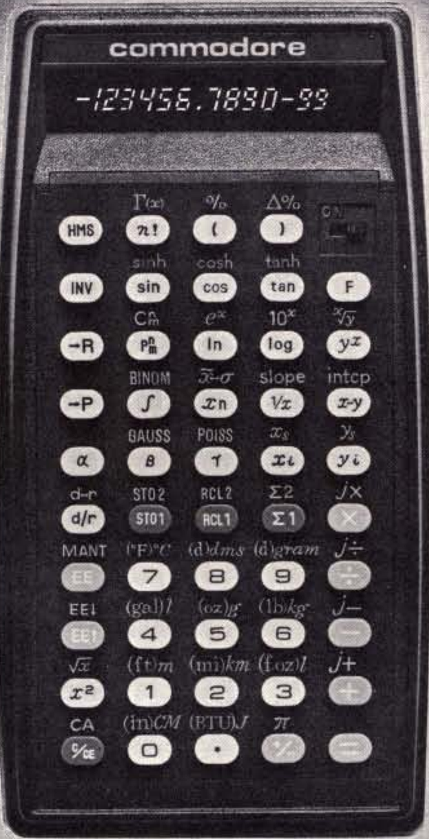
Let us approach the problem by way of combinatorial geometry. If we checkerboard-color a 5×5 cross section, we find that it contains 13 cells of one color and 12 of the other [*see illustration on page 122*]. Consider all the bricks except the three small ones. No matter how one brick is placed, it must fill an even number of cells (0, 2, 4 or 8) in every section. Half of the cells it fills will be one color and half will be the other.

Since there are an odd number of cells in each section, one or three of its cells must be occupied by a $1 \times 1 \times 3$ brick. In addition the brick must be placed so that if it fills one cell, it will be a cell of the same color as the central cell. If it lies entirely within the section, two of its three cells must match the color of the central cell.

No matter how a small piece is placed, it will occupy five sections. It follows that the three pieces, if they contribute to all 15 sections (as they must), will have to be placed so that no two contribute to the same section. There is only one way to do that and at the same time meet the coloring requirements. Once the small pieces are correctly positioned it is not hard to find one of the more than 500 ways of packing the larger bricks around them.

John Horton Conway, who invented this cube puzzle, could have substituted another $1 \times 2 \times 4$ for the $2 \times 2 \times 2$, thus providing the maximum number of canonical bricks that will go into a $5 \times 5 \times 5$ box, as was shown last month. Adding the order-2 cube, however, provides amusing misdirection. The puzzle can also be given with 29 bricks of $1 \times 2 \times 2$, together with the three $1 \times 1 \times 3$ bricks. Whether most people would find that easier or more difficult is hard to say. The principle behind Conway's cube generalizes in the sense that three $1 \times 1 \times (2n - 1)$ bricks have a unique configuration in a cubical box of side $2n + 1$ that makes it possible to pack the rest of the box with $1 \times 2 \times 2$ bricks.

The most diversified scientific calculator you can own...and you can, for just $79.95



The Commodore rechargeable SR4190R

A powerful preprogrammed multifunction calculator with 106 directly accessible keyboard functions.

The complete mini-computer for
Every Engineer
Physicist
Chemist
Geologist
Demographer
Ecologist
Mathematician
Statistician
Quality Control Analyst
Time and Motion Expert
Advanced Student
. . . and so many others

There has never been an electronic calculator with so extensive a performance range as the brand new Commodore SR4190R. Its operating capacity, combined with the speed and accuracy with which computations are performed is unequalled by any other calculator.

We are proud to offer it to you on a, "no-questions asked", full refund guarantee. More than its diversified applications, we want you to appreciate its many exclusive advantages for yourself.

Discover its unique **Hours-Minutes-Seconds** in digital clock format. Ambiguity is a thing of the past. And you can perform arithmetic operations, such as time study and motion analysis in this mode as easily as in the decimal mode.

Complex numbers, difficult on some calculators, impossible on others, are directly accessible from the SR4190R keyboard. Thanks to its extraordinary accuracy limit, **Combination and Permutation** operations are not hindered by the overflow which occurs when the factorial is larger than 10^{100}.

Among the 106 direct entry functions are **Poisson and Binomial Probability Densities; Gaussian Distribution; Linear Regression Analysis; Mean and Standard Deviation**; and so many more. Most important, all are on the keyboard. Several thousand internal preprogrammed steps put all of these functions at your fingertips . . . instantly. You needn't bother with preparing formulas, creating programs or maintaining an elaborate library. It's all right there. The broadest range of diversified applications is at your command.

commodore
A Name You Can Trust

Commodore is perhaps the oldest brand in consumer electronic calculators. The SR4190R was designed, developed and produced in Palo Alto, California. Other Commodore plants are located in Osaka, Japan, Toronto, Canada and Eaglescliff, England.

SATISFACTION OR YOUR MONEY BACK IN FULL

Order your SR4190R Rechargeable Scientific on a "No-Questions-Asked" trial basis. If not satisfied with the performance of your SR4190R, return it for a full refund.

A Brief Summary of the SR4190R

10-digit Mantissa, 2-digit exponent with Variable Exponent Integer Increase and Decrease
MANT, EE, EE↑, EE↓

Register Keys
STO 1, RCL 1, Σ1, STO 2, RCL 2, Σ2, x↔y, Xn, Xi, Yi, α, β, γ, ( ).

One Real Variable Function Keys
ln, log, e^x, 10^x, 1/x, $\sqrt{x}$, x^2
SIN, COS, TAN, INV SIN, INV COS, INV TAN
sinh, cosh, tanh, INV sinh, INV cosh, INV tanh

Two Real Variable Arithmetic Function keys:
+ − × ÷

Two Complex Variable Arithmetic Function Keys:
j +, j −, j ×, j ÷

Two Real Variable Analytical Function Keys:
→P, →R, P^n_m, y^x, $\sqrt[x]{y}$, %, Δ%, C^n_m

Statistical Function Keys:
x↔y, SLOPE, INTCP, GAUSS, BINOM, POISS, x_s, y_s

Hours-Minutes-Seconds Mode:
HMS

Unit Conversions:
(°F) C, (d) dms, (d) gra, (gal) l, (oz) g, (lb) kg, (ft) m, (mi) km, (f oz) l, (in) cm, (BTU) J

Degree/Radian Conversion and Mode Keys:
d/r, d↔r

Numerical Entry Keys:
0-9, π

3 Angular Units
radians, degrees, grads

Poisson and Binomial probability densities. Gaussian distribution.

Mean and Standard Deviation PLUS Linear Regression Analysis

Power supply: Built-in rechargeable Ni-Cad batteries. AC adapter/recharger included.

One Year Warranty

Order Direct

MR. CALCULATOR

39 Town and Country Village, Palo Alto, California 94301 Phone (415) 328-0740

Complete this coupon and enclose with payment or credit card information.

Please send me__________ Commodore
(Quantity)
SR4190R(s). I understand I may return them within 10 days after receiving them and receive a full refund if I am not satisfied.

Name____________________
(Please print)
Street____________________
City____________________
State__________ Zip__________
☐ Check or money order enclosed.
☐ Please charge my order to my credit card.
Exp. Date____________________
BankAmericard #____________________
Master Charge #____________________
Your Signature____________________
Cost of Calculators Ordered @ $79.95 ea. $______
California residents add 6% sales tax $______
Plus Postage and Handling ($2.50 per Calculator) $______
Total Amount Enclosed (or Charged) $______

Mr. Calculator retail stores are located in California: San [illegible] Palo Alto, San Jose, San Diego. Texas: Houston, Austin.

BOOKS

The lost lore of pre-Columbian America, and an artist's vision of Galileo's period

by Philip Morrison

ARCHAEOASTRONOMY IN PRE-COLUMBIAN AMERICA, edited by Anthony F. Aveni. University of Texas Press ($16.50). THE CODEX NUTTALL: A PICTURE MANUSCRIPT FROM ANCIENT MEXICO, edited by Zelia Nuttall, with an introduction by Arthur G. Miller. Dover Publications, Inc. ($7.50). At the summer solstice of 1973 an international assembly was convened in the Valley of Mexico to discuss a wide range of disciplines of science and scholarship. An illustrated volume (set in left-justified typewriter script and hence bulky for its cargo of words) presents almost a score of papers from one session there, by archaeologists, astronomers, anthropologists, historians, an architect and even an amateur or two. The subject is evidently a burgeoning one, and this is a diverse offering. About a third of the book is devoted to astronomy in the southwestern U.S. among Pueblo and Navajo peoples and their ancestors; there are studies of archaeological sites (most of them north of Panama) whose form or orientation may be astronomical, a number of papers on Maya calendrics in manuscript and stone, a paper treating of an Olmec mosaic pendant and a couple of excellent reviews.

It is abundantly clear that we have only a little understanding of the rich astronomy of pre-Columbian America. The 16th-century Spanish scholars, our best witnesses to the high culture the conquest forever ended, do not tell us much about American concepts of the heavens. Like most ethnographers after them, they were men whose knowledge of the sky was scanty. The great Indian astronomical specialists and their schools were eliminated unheard. Even the folk cultures that survive today have been little understood, since "there is scarcely an ethnologist or social anthropologist who can identify anything other than the moon and the Big Dipper in the night sky." The indigenous people are a good deal wiser, although even among them there are only a few experts, who must be sought out and listened to.

Some seeking out of this type is exhibited here. One paper is devoted to elaborating the well-known assertion that certain pictographs of a crescent with some second luminary may record the sudden appearance of the Crab supernova in the morning sky, together with the crescent moon, in July of A.D. 1054. The fact is that it was not the custom of the priestly sky-watchers—who still serve the northern Tewa pueblos—to record unexpected phenomena. In 1067 there was a great volcanic eruption and ash fall in northern Arizona; it led to no known petroglyph, even though fields and houses had to be abandoned for years. On the other hand, associations of sun, moon and star petroglyphs are found even today at sites that were used for calendric divination and dating. Florence H. Ellis and her informants have made a strong case against the supernova identification, a case that is assisted by the large differences among the many diagrams interpreted as marking the moon-supernova association. In a similar vein Navajo experts have made it clear to Claude Britt, Jr., that the "planetaria" (rock-shelter ceilings marked with diagrams resembling star maps, which are found in the Canyon de Chelly in particular) are not ancient Anasazi maps of the sky, made perhaps to celebrate the time of the supernova, but are much later Navajo star paintings that had ceremonial and training functions for their medicine men.

Other papers offer original suggestions, although the research has in general yet to mature to conviction. Marion P. Hatch finds a remarkable clue in the Madrid Codex, a Maya hieroglyphic almanac: a table accompanied by successive drawings of a serpent, which she identifies with the constellation Draco turning around the third-magnitude star Eta Draconis. That star is remarkable because of its position close to the pole of the ecliptic, which has allowed the data of its meridian transit to remain nearly unchanged over millenniums during which stars that are not far away shifted in meridian transit by a couple of months. If this were to be confirmed, it would raise many important issues, not least among them a search for similar concern with the conspicuous star Canopus, far to the south but equally close to the other precessional pole. Canopus was to the Greeks the rudder of the ship Argo (a possible allusion to its axial steadfastness?).

The old issue of the off-meridian alignment of the great precinct of Teotihuacán, as well as a score of other ancient cities of Mexico and Guatemala, is carefully examined by the editor of the volume. The shift some 15 degrees east of north was no error; in Teotihuacán, at least, evidence is preserved of remarkable surveying accuracy, to a tenth of a degree or better. Perhaps the orientation was to mark the heliacal rising of the Pleiades, which took place on the very day that the sun, in its annual passage north, reached the zenith of the tropical site. That motion has long been connected with the curious 260-day cycle of the Maya world, since the sun returns southbound through the zenith again just 105 days later. Another paper suggests that some effigy mounds (perhaps even those strange curly figures written large on the Peruvian desert at Nazca) represent not alignments but templatelike mappings of the starry world above, just as the familiar swastika might stand for the form of Ursa Minor turning around Polaris nightly.

This collection of papers offers a good deal to ponder and to hope for. The subject is not new, but it is newly reborn. The volume should please many readers even though false starts and wishful thinking abound; it is less a demonstration of arrival at the truth than a guide to long and hopeful journeys that lie ahead and are sure to surprise us all.

The Codex Nuttall, which offers the eye a strange and polychrome journey hard to match in books of moderate price, gives direct access to the symbolic life of the second growth toward civilization that was so forcibly grafted back to the Old World by Cortes and Pizarro. About 15 painted "books" from the Mixtec culture are still preserved. This is one of them; it surfaced in a Florentine monastery, and in the 19th century it went as a gift "to the Hon. Robert Curzon, 14th Baron Zouche." Zelia Nuttall, an energetic and original Americanist, first published it from Harvard's Peabody Museum for scholars everywhere. She arranged for it to be rendered by an unnamed artist in careful facsimile on foldouts. It is a photographic reproduction of that 1902 publication we have here, now printed on cut book pages rather than on foldouts, with the reverse and obverse pages running consecutively. Although the white book paper is not as clean a white as the lime-stuccoed bark paper, and although the bright colors are not as intense as the natural mineral pigments of the Mixtec artists, this reproduction is "more faithful to the original than many photographic facsimiles," says Dr. Miller, an archaeologist at the Center for Pre-Columbian Studies at Dumbarton Oaks.

It is a striking book, visually much more appealing than the great Maya almanacs one can see in facsimile. Their texts are strings of complex but nearly monochromatic Maya glyphs; they nourish the cryptographic taste, but this book is only a half-rebus, more resembling the Bayeux chronicle. Some 1,200 to 1,400 little painted figures crowd its pages, among the temples, palaces and fishy blue pools of their world. Here march grave masked princes in profile, their richly cloaked backs heavily laden with symbolic accouterments, hands bearing a white dog or a great plumed shield, or even locked in dread combat. Here a naked priest snatches the heart out of a prisoner, there a noble wedding pair pledge their troth with a beaker of frothing chocolate. Regard a curious struggle, jaguar and turkey, and, almost at the end, a brave king arrowed to death as he stands tied to an executioner's scaffold.

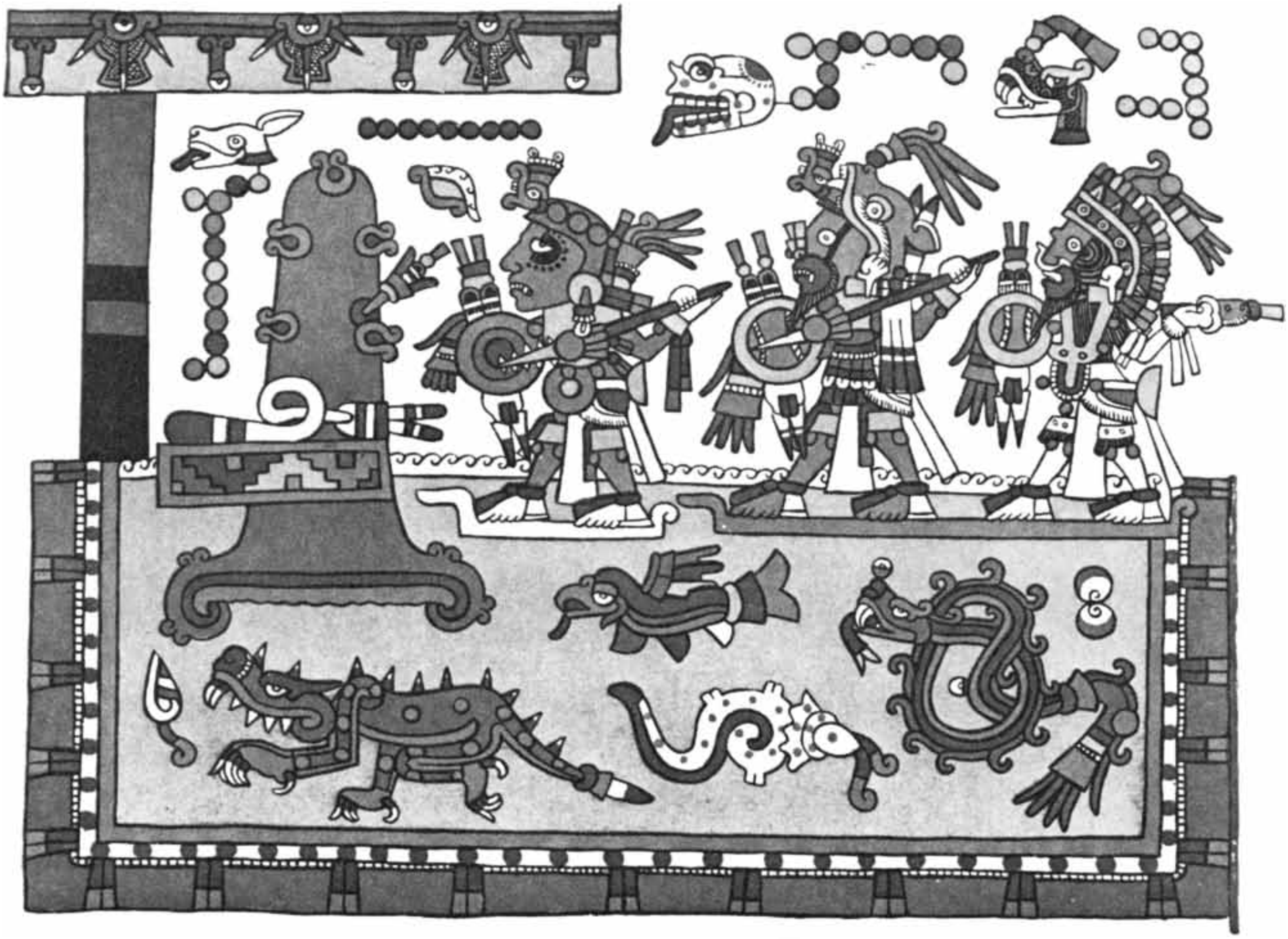
Everywhere on the dense yet composed pages float the linked dots and little objects that spell Mexican names and dates. We know, mainly from the work of the Mexican archaeologist Alfonso Caso, that the book narrates first the lineage and then the exploits of a hero-king of a 12th-century Oaxacan state: Eight-Deer Tiger Claw of Tilantongo. Caso's translation of the complex picture book with its mix of lively hieratic representation, ideograph and phonetic symbol was published 10 years ago; he died in 1970. Two pages of his translation are cited here; with the guide thus offered, the reader's fancy is given buoyant aid. There is much that not even Caso could read; perhaps the many who will see the published pages can add to our understanding.

Nuttall herself thought this book might have been among the treasures Cortes sent off to his emperor from Vera Cruz that first year. Modern scholars exclude the possibility because the text contains a number of pen-and-ink notations in Spanish script with Nahuatl words; no one could write that way in 1520, and such notes seem highly improbable in Europe even during a long unknown history. (One hopes that the large crimson *H* unmistakably blazoned at the base of a temple pyramid on page 45 and elsewhere is not the copyist's waggish joke on Harvard!)

MODELS OF MADNESS, MODELS OF MEDICINE, by Miriam Siegler and Humphry Osmond. Macmillan Publishing Co., Inc. ($8.95). The febrile, failing and unresponsive child seemed unlikely to recover. Then a hasty visit by the great clinician, by chance costumed in his glorious academic robes, entranced the child into eating a sugared peach fed him bit by bit by the splendid visitor, who assured the boy that he would not be sickened by this very special fruit. It worked, and for the next 40 days Sir William Osler "put on his doctor's robes in the hall" outside the sickroom. Medicine is a social relationship, not a mere science, and it follows patterns of symbolic behavior evolved in the widest spectrum of societies.

Clinical medicine is built on the relation of two roles: the "Aesculapian authority" of the doctor (seen in its full graciousness with Osler) and the submissive but yet hopeful role of the sick person, exonerated from blame and exempt from normal tasks but constrained to seek the authoritative help of the physician. "Doctor's orders" excuse almost any behavior, yet they are mere advice. The physician's authority has three sources: knowledge and wisdom, the morality and goodness of his task ("socially right as well as individually good") and a charismatic element, flowing from a special grace, that transcends reason because it is plain that life and death are arbitrary and not fully knowable. The role of the sick person is in no way deviant; it is normal, assigned to almost everyone at some time in life. The role of the sick person seems primary; to confer that status is perhaps the key function of Aesculapian authority. One cockney patient saw this truth: "Of course Guy's Orspital is a very fine Orspital, I know it well.... And there are some very good doctors—I knows 'em.... But what I want to know is this—what would Guy's Orspital be without the patients?"

Along such sociological lines, deriving from Talcott Parsons and T. T. Paterson, these authors have given a detailed typology of "models of madness." They spend most of their effort quite disarmingly analyzing the standard medical world view. They come not to bury it but to praise it. Defective, transparent as its anthropological nature may be, it is a practical institution, the consequence of long social evolution. Those who would deal with disorders as serious as the common forms of schizophrenia abandon medicine at the peril of placing heavy responsibility on those who are too dam-



Three Mixtec figures in **The Codex Nuttall: A Picture Manuscript from Ancient Mexico** aged to bear it. The authors contrast in detail a long series of alternative models: the moral one, in which the behavior of the client must be changed from its bad state; the "impaired" model, where the patient is recognized as being "normal" for his handicapped state, and the aim becomes to let him live as benign a second-class life as he can manage; the psychoanalytic model, based on the particularity of the individual case and seeking a transference of what is mainly symbolic behavior, and four still more radical views belonging mainly to the 1960's, in which the disease is seen variously as a sickness of society, an attempt to break the bonds of a narrow conformity, an escape from a conspiracy of mislabeling or a mere index of a pathologically maneuvering family.

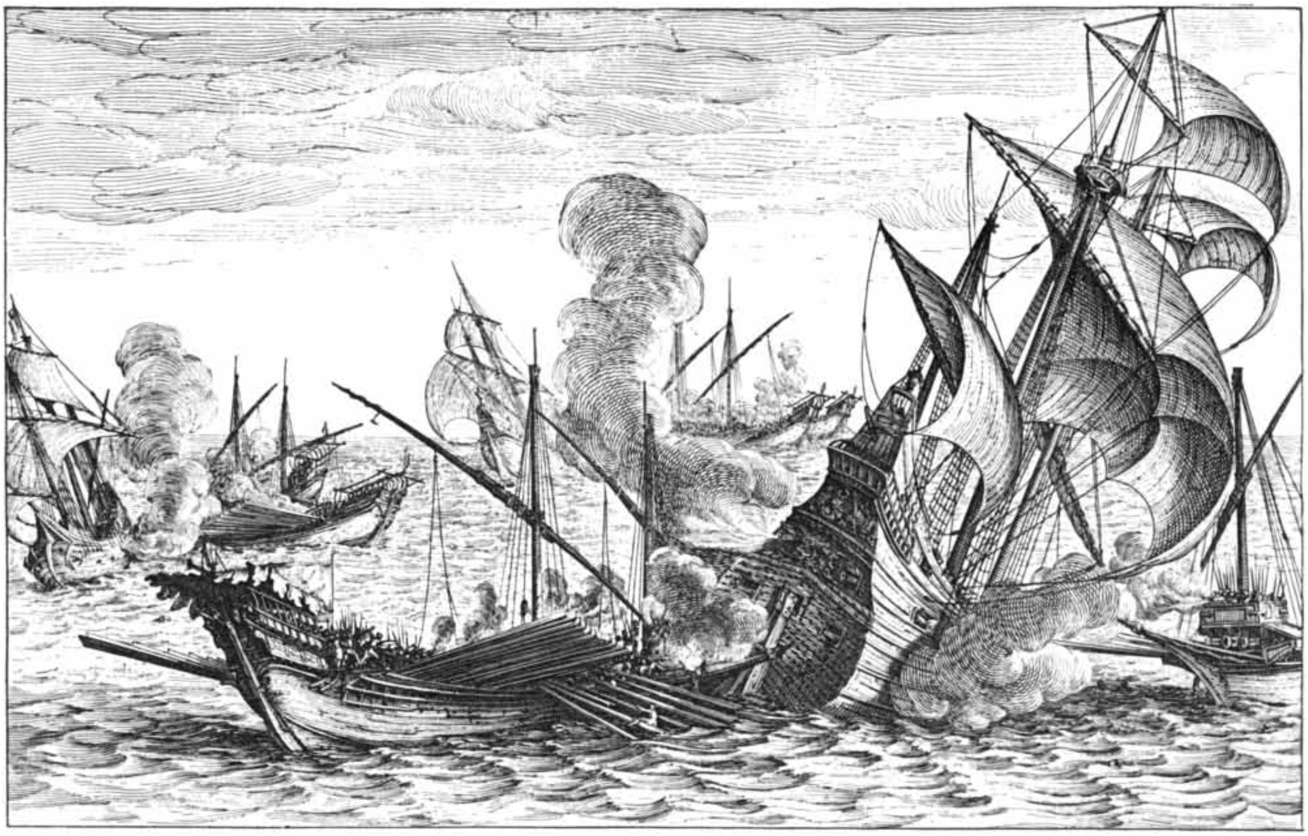
For Siegler and Osmond the grave illness is centered in perceptual and sensory disorders, and it is medical because it probably arises from some dimly perceived genetic-biochemical fault. ("Madness" is to be read narrowly as schizophrenia.) Their discussion is richly made, with material from the clinic, the hospital, the novelist and various expositions of the alternatives over the years. No doubt they scamp some of the models proposed by others, but their broad position seems fairly and honestly supported. If one looks away from the frankly psychotic disorders to the subtler neurotic ones, the psychoanalysts, whose depth is not much lessened by the probability that their method has gone beyond science to the particular and the unverifiable, still have a strong inning to come. The more fragile other views seem transient as models, however fashionable the book of the moment and its useful part truth.

Public-health doctors and "science doctors" are included as submodels of normal medicine. They share with it the etiological assumption that disease arises neither from the supernatural nor from malice and misbehavior. They do not, however, play opposite the sick role. The public-health doctor has no patients; he treats unseen citizens, with an authority that is structural, even legal. The investigator treats not patients but experimental subjects who are mainly standing in for patients of the future. His authority is entirely sapiential; his subject deserves to be compensated, whether in money, in time or in a sense of shared contribution.

A final compact account of models of drug addiction and of alcoholism completes this easily read, provocative book, everywhere written with a winning goodwill and candor, a book of clarity and energy that even its opponents will find useful. A frequent triteness of expression is its chief weakness.

CALLOT'S ETCHINGS. Edited by Howard Daniel. Dover Publications, Inc. ($4.50). If you would see the Florence that Galileo Galilei looked on—the *commedia dell'arte,* fireworks over the Arno, a country fair, a peasant defecating, a pageant on the Piazza della Signoria—then view it through the faithful eye and precise hand of Jacques Callot. This rich paperbound selection is a real bargain; most of the photo-offset reproductions from excellent impressions are at approximately the original scale. After the death of Cosimo II, Callot left Florence, which would no longer spend much money on art. Back in his native Nancy he was drawn into the Thirty Years' War, which he depicts with anger and compassion: camp followers, formal drill, wild pillage and grandiose siege. The expert editor endorses the plausible guess that Callot used both optical novelties of his day: the telescope to view entire armies in the field and the hand magnifier to control his minutely worked copper surfaces. Those masterful, intricate lines and textures represent to us across three centuries a clotted Florentine crowd or serried squadrons of cavalry, each vivid in an area smaller than the palm of your hand.

ATLAS OF OPTICAL TRANSFORMS, by G. Harburn, C. A. Taylor and T. R. Welberry. Cornell University Press ($15). The delicate current in a radio receiver antenna is extraordinarily intricate. Consider that the electrons respond all at once to the signals from dozens, even hundreds, of broadcast programs. Their resulting motion, more complicated than the stormy sea, would seem beyond analysis. Yet a simple tuned circuit that allows the adjustment of one quantity only, the frequency, makes possible the neat separation of that cacophony into individual signals, one after another. The entire response is the carefully



A Florentine naval battle depicted by Jacques Callot in **Callot's Etchings** phased simultaneous sum of the responses to each individual frequency. (Set aside for a moment the necessarily finite bandwidth of the tuner.) Indeed, the ear works in an analogous way.

That scheme, which represents a general function of time as the sum of very simple oscillatory functions of time, each with its own frequency to separate it from the rest, is known—after its renowned developer—as Fourier analysis. It is with such equivalence techniques that the power of mathematical physics first departs from the intuitive. The scheme can be much generalized, to dimensions other than time and to many dimensions at once. Such a mapping in two-space, in the plane of an object or its image, is called an optical transform. If I illuminate a mask with coherent parallel light of a single wavelength, the mask will diffract that light to form a pattern of a new kind. The pattern appears with the geometry of the object not simply magnified or diminished, as by a lens, but transformed by the interference of the coherent waves along the several paths they travel to the recording surface. That diffraction pattern is a two-dimensional Fourier transform. Just as a signal with rapid variation in time requires a wide frequency band in the case of the radio receiver, so an object with fine detail in space will affect a broad range of positions in the final pattern. The two spaces, the domains of frequency and of time, or the positions in object and in diffraction pattern, are said to lie in reciprocal spaces.

The matter is not easy. Formally it is clear and powerful, but it is far from our conscious sensory responses and difficult to welcome to intuition. Yet it lies deep beneath much of science, from oceanography and radio astronomy to X-ray crystallography. It is perhaps the crystallographers who have made the most of it. Their aim is to interpret the X-ray-diffraction patterns of a three-dimensional lattice of atoms: to perceive the lattice's true shape in space from the arrangement of spots and lines in the cross sections of their photographs.

This unusual atlas is the crystallographers' effort—fascinating to any thoughtful examiner—to build intuition in reciprocal space. Three Cardiff physicists have done the work, under the sponsorship of a UNESCO pilot project on the teaching of crystallography. What they have done, following an idea of Sir Lawrence Bragg's of some 40 years ago, is to make analogues for the main features of X-ray crystal analysis. The model lattices, called masks, are more or less complex arrays of black marks. The diffraction patterns are produced photographically. The X-ray beam is modeled, on a scale enlarged to suit the coarseness of human vision, by optical light.

The authors present a long series of elegant full-page plates. On a left-hand page you see a set of masks, ranging from one as simple as a single dot up to regiments of dots parading in lattices and then, with all manner of imperfections, up to helixes, double helixes and finally to Mickey Mouse, displayed as a very spotty pattern indeed. The facing page shows the reciprocal-space diffraction pattern photographed from each mask. It is instructive to see the wide bright spot made by a tiny obstacle, and to follow the way the fine detail grows as the mask is enlarged with more and more dots. Disorder and its effects become plainly visible, and one finds true lattice patterns made by masks with many thousands of dots. It is a real wonder to note how, whereas crispness waits on strongly repetitive order, the symmetry is built up almost immediately. More than 30 plates are here, each showing both the actual mask and the diffraction pattern of a dozen logically related examples, with descriptions in French and English.

The simple but painstakingly controlled setup of three lenses (one to magnify), laser source and film is sketched and described. Some masks are made by punching holes in exposed film and making contact prints of the result. Most are produced, however, by an automatic device that, as instructed by a magnetic tape, exposes a photographic film to precise dots on any nodes of a square lattice: up to 80 million dots on a full sheet of film! The light of a helium-neon laser is used to make the diffraction photographs. Care is taken to control phase, even to the point of deliberate modification of optical phase by the use of polarized light and mica half-wave plates in order to model certain advanced tricks of the X-ray crystallographer. It is sure that students and practitioners of the interferometric sciences will benefit from long study of these pictures. Any general scientific reader owes it to himself to look for 15 minutes at the beginning examples, which build up the fundamentals, and then perhaps at the diffraction diamond, which is characteristic of a helix. One complaint: we are given no hint anywhere as to the full size of masks that are so complex as to be shown only in sample.

CONTROLLED ENVIRONMENTS FOR PLANT RESEARCH, by Robert Jack Downs. Columbia University Press ($12). Three decades back campus after campus saw the construction of a curious building, without classrooms but full of humming machinery, where the physicists worked happily by day and night. The places shared a name: the cyclotron.

There is contagion in the name. For 20 years now, since a prototype was built in the 1950's at the California Institute of Technology, another similarly intricate and imposing style of laboratory has spread across the campuses. This time it was the plant scientists who were cheerful; the places were called phytotrons. A phytotron is a combination of temperature-controlled greenhouses and artificially lighted rooms; here temperature, light and humidity can be controlled at will for the purpose of studying the response of plants to their environment.

The author is the director of the phytotron at North Carolina State University; his brief and informed book is an effort to show biologists some of the practical problems their demands set before engineers and to let the engineers in more plainly on the requirements of plants. The task seems simple; after all, room air conditioning for people works reasonably well, and greenhouses have been used for centuries. Yet there is trouble in these little green Edens. The rules of human comfort are plainly not good enough. The light brings a heavy load of heat, and it cycles on and off in unfamiliar ways.

There have been plenty of failures in expensive, specially designed growth chambers; the engineers too often act as though the job were just another school building. The mistakes of others are ignored. Even the prefabricated products offered by a dozen U.S. companies and another dozen in other countries are often a disappointment. There are no tight standards for comparison. Many rooms work well—when they are empty. Plants interfere with the airflow and liberate much latent heat with their water vapor, and such factors are hard to estimate. The lesson is to buy only after careful study and detailed specification.

A growth chamber (which may be a spacious walk-in room or a little cabinet) consists of a refrigeration system for controlling temperature and relative humidity within a tight structure that has a natural or artificial source of light. Temperature control is the key. Relative humidity, regulated by a humidifier to bring in mist or steam and a cold plate to dehumidify, is generally less important and is more difficult: "Where precise control of humidity is required, costs begin to rise dramatically." A single human hair is still the heart of the most common hygrometric control, although thermal dew-point detectors and resistance sensors are now available. Light supply is no simple matter; many biologists believe the maximum maintainable intensity from reasonably priced lamps is only half of what such crops as corn, milo and cotton can use.

The biological side, of course, is still less routine. Data on the right temperature for plants are sparse and often conflicting. Temperature optimums are by no means independent of light intensity, carbon dioxide concentration or daily and seasonal temperature progression. "Fieldlike tobacco plants are difficult to obtain in controlled-environment facilities" without attention to a seasonal temperature shift. In most growth rooms carbon dioxide is short during lighted periods; the plants appear to be healthy but are always a little hungry. Supplementary carbon dioxide yields tobacco more like "normal" plants. Light quality too is tricky. Warm white light seems in practice the best, watt for watt, even though red and blue lamps match the photosynthetic action spectrum better. The red to far-red ratio, particularly at the time just before darkness, is of real importance.

This small book is candid testimony to how far we still are from detailed understanding of the plants by whose growth we live. It is a cautionary tale, but a living argument for the importance and interest of fundamental studies on the whole plant.

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

THE METABOLISM OF ALCOHOL

THE BIOLOGY OF ALCOHOLISM, VOL. I, VOL. II, VOL. III. Edited by Benjamin Kissin and Henri Begleiter. Plenum Press, 1971, 1972, 1973.

INCREASE OF ETHANOL, MEPROBAMATE AND PENTOBARBITAL METABOLISM AFTER CHRONIC ETHANOL ADMINISTRATION IN MAN AND IN RATS. P. S. Misra, A. Lefèvre, H. Ishii, E. Rubin and C. S. Lieber in *The American Journal of Medicine,* Vol. 51, No. 3, pages 346–351; September, 1971.

SEQUENTIAL PRODUCTION OF FATTY LIVER, HEPATITIS, AND CIRRHOSIS IN SUB-HUMAN PRIMATES FED WITH ADEQUATE DIETS. Charles S. Lieber, Leonore M. DeCarli and Emanuel Rubin in *Proceedings of the National Academy of Sciences,* Vol. 72, No. 2, pages 437–441; February, 1975.

HIGH BLOOD ACETALDEHYDE LEVELS AFTER ETHANOL ADMINISTRATION: DIFFERENCE BETWEEN ALCOHOLIC AND NONALCOHOLIC SUBJECTS. M. A. Korsten, S. Matsuzaki, L. Feinman and C. S. Lieber in *The New England Journal of Medicine,* Vol. 292, No. 8, pages 386–389; February 20, 1975.

MEDICAL CONSEQUENCES OF ALCOHOLISM. *Annals of the New York Academy of Sciences,* Vol. 252; April 25, 1975.

POLYCYCLIC AROMATIC COMPOUNDS IN NATURE

ÜBER DAS VORKOMMEN VON CHRYSEN IN ERDE. W. Kern in *Helvetica Chimica Acta,* Vol. 30, No. 6, pages 1595–1599; October 15, 1947.

BENZPYRENES IN SOIL. M. Blumer in *Science,* Vol. 134, No. 3477, pages 474–475; August 18, 1961.

INVESTIGATION OF BIOSYNTHETIC FORMATION OF POLYCYCLIC HYDROCARBONS IN HIGHER PLANTS. Gernot Grimmer and Dietrich Düvel in *Zeitschrift für Naturforschung,* Vol. 25b, No. 10, pages 1171–1175; October, 1970.

HYDROCARBONS IN HUMAN ENVIRONMENT: PART IX. CONTENT OF POLYCYCLIC HYDROCARBONS IN ICELAND SOIL. G. Grimmer, J. Jacob and A. Hildebrandt in *Zeitschrift für Krebsforschung und Klinische Onkologie,* Vol. 78, No. 1, pages 65–72; 1972.

ELEMENTAL CARBON IN MARINE SEDIMENTS: BASELINE FOR BURNING. Dwight M. Smith, John J. Griffin and Edward D. Goldberg in *Nature,* Vol. 241, No. 5387, pages 268–270; January 26, 1973.

POLYCYCLIC AROMATIC HYDROCARBONS IN THE ENVIRONMENT: ISOLATION AND CHARACTERIZATION BY CHROMATOGRAPHY, VISIBLE, ULTRAVIOLET, AND MASS SPECTROMETRY. Walter Giger and Max Blumer in *Analytical Chemistry,* Vol. 46, No. 12, pages 1663–1671; October, 1974.

THE METEOROLOGY OF JUPITER

THE PLANET JUPITER. Bertrand M. Peek. Faber and Faber, 1958.

THE ATMOSPHERE OF THE JOVIAN PLANETS: PAPERS FROM THE THIRD ARIZONA CONFERENCE ON PLANETARY ATMOSPHERES. *Journal of the Atmospheric Sciences,* Vol. 26, No. 5, Part 1, pages 795–1001; September, 1969.

A SURVEY OF THE OUTER PLANETS: JUPITER, SATURN, URANUS, NEPTUNE, PLUTO, AND THEIR SATELLITES. R. L. Newburn, Jr., and S. Gulkis in *Space Science Reviews,* Vol. 14, No. 2, pages 179–271; 1973.

WILL THE UNIVERSE EXPAND FOREVER?

MODERN COSMOLOGY. D. W. Sciama. Cambridge University Press, 1971.

THE AGE OF THE ELEMENTS. David N. Schramm in *Scientific American,* Vol. 230, No. 1, pages 69–77; January, 1974.

WHAT CAN DEUTERIUM TELL US? David N. Schramm and Robert V. Wagoner in *Physics Today,* Vol. 27, No. 12, pages 41–47; December, 1974.

AN UNBOUND UNIVERSE? J. Richard Gott III, James E. Gunn, David N. Schramm and Beatrice M. Tinsley in *The Astrophysical Journal,* Vol. 194, pages 543–553; December 15, 1974.

THE RESOURCES OF BINOCULAR PERCEPTION

TEXTURE AND VISUAL PERCEPTION. Bela Julesz in *Scientific American,* Vol. 212, No. 2, pages 38–48; February, 1965.

THE INTELLIGENT EYE. Richard L. Gregory. McGraw-Hill Book Company, 1970.

FOUNDATIONS OF CYCLOPEAN PERCEPTION. Bela Julesz. University of Chicago Press, 1971.

STEREOPSIS BY BINOCULAR DELAY. John Ross in *Nature,* Vol. 248, No. 5446, pages 363–364; March 22, 1974.

THE SMALL ELECTRONIC CALCULATOR

MOS/LSI DESIGN AND APPLICATION. William N. Carr and Jack P. Mize, edited by Robert E. Sawyer and John R. Miller. McGraw-Hill Book Company, 1972.

THE 'POWERFUL POCKETFUL': AN ELEC-

TRONIC CALCULATOR CHALLENGES THE SLIDE RULE. Thomas M. Whitney, France Rodé and Chung C. Tung in *Hewlett-Packard Journal,* Vol. 23, No. 10, pages 2–9; June, 1972.

A POCKET-SIZED ANSWER MACHINE FOR BUSINESS AND FINANCE. William L. Crowley and France Rodé in *Hewlett-Packard Journal,* Vol. 24, No. 9, pages 2–9; May, 1973.

FLOATING POINT ARITHMETIC. S. Kelly and S. MacFarlane in *Electronic Engineering,* Vol. 46, No. 554, pages 53–56, April, 1974; No. 555, pages 77–80, May, 1974; No. 557, pages 43–47, July, 1974.

SOCIAL SPIDERS

THE INSECT SOCIETIES. Edward O. Wilson. The Belknap Press of Harvard University Press, 1971.

EVOLUTION OF SOCIAL BEHAVIOR IN SPIDERS (ARANEAE; ERESITAE AND THERIDIIDAE). Ernst J. Kullmann in *American Zoologist,* Vol. 12, No. 3, pages 419–426; August, 1972.

THE SHEET WEB AS A TRANSDUCER, MODIFYING VIBRATION SIGNALS IN SOCIAL SPIDER COLONIES OF *MALLOS GREGALIS.* J. Wesley Burgess in *Neuroscience Abstracts, Society for Neuroscience Fifth Annual Meeting;* 1975.

FLUORESCENCE-ACTIVATED CELL SORTING

FLUORESCENCE ACTIVATED CELL SORTING. W. A. Bonner, H. R. Hulett, R. G. Sweet and L. A. Herzenberg in *The Review of Scientific Instruments,* Vol. 43, No. 3, pages 404–409; March, 1972.

DEVELOPMENT AND APPLICATION OF A RAPID CELL SORTER. H. R. Hulett, W. A. Bonner, R. G. Sweet and L. A. Herzenberg in *Clinical Chemistry,* Vol. 19, No. 8, pages 813–816; August, 1973.

THE FC RECEPTOR ON THYMUS-DERIVED LYMPHOCYTES: I. DETECTION OF A SUBPOPULATION OF MURINE T LYMPHOCYTES BEARING THE FC RECEPTOR. Robert D. Stout and Leonard A. Herzenberg in *The Journal of Experimental Medicine,* Vol. 142, No. 3, pages 611–621; September 1, 1975.

THE FC RECEPTOR ON THYMUS-DERIVED LYMPHOCYTES: II. MITOGEN RESPONSIVENESS OF T LYMPHOCYTES BEARING THE FC RECEPTOR. Robert D. Stout and Leonard A. Herzenberg in *The Journal of Experimental Medicine,* Vol. 142, No. 5, pages 1041–1051; November 1, 1975.

MATHEMATICAL GAMES

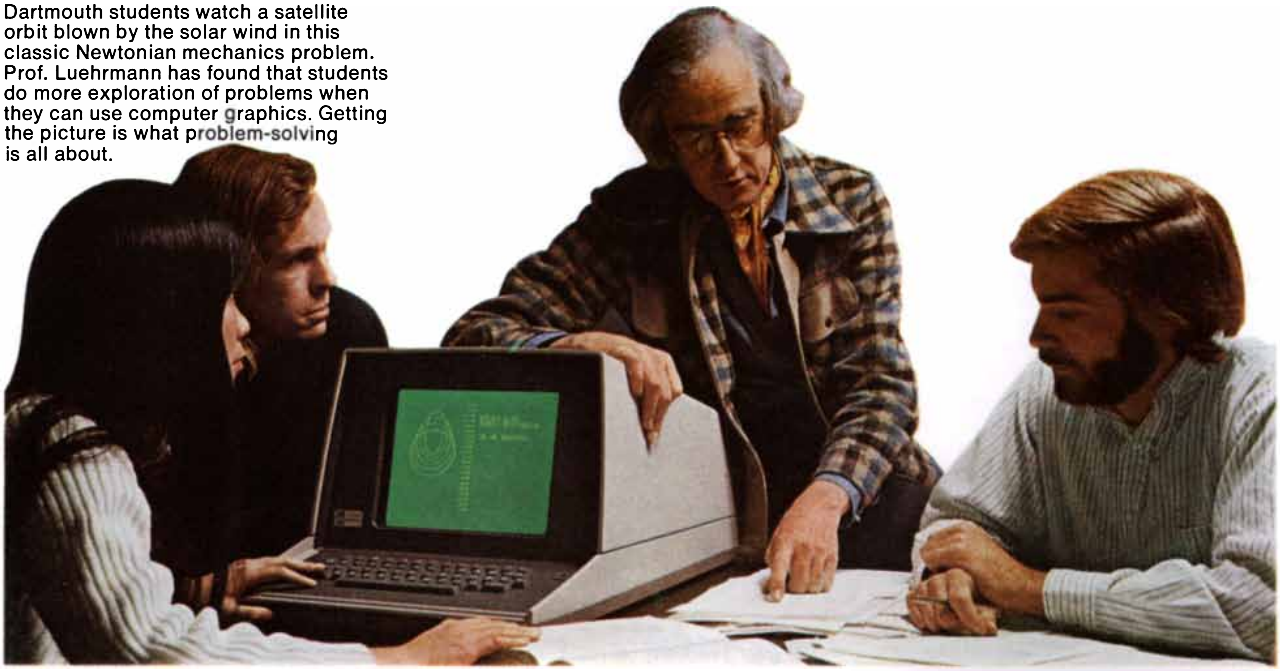
ON SIMPSON'S PARADOX AND THE SURE-THING PRINCIPLE. Colin R. Blyth in *Journal of the American Statistical Association,* Vol. 67, No. 338, pages 364–381; June, 1972.

THE JUSTIFICATION OF INDUCTION. Edited by Richard Swinburne. Oxford University Press, 1974.

"With computers and graphics at Dartmouth we're taking students out of their passive receptor situation and putting them into the active inquiry role of a researcher."

Prof. Arthur Luehrmann
Office of Academic Computing, Dartmouth College, Hanover, N.H.

Dartmouth students watch a satellite orbit blown by the solar wind in this classic Newtonian mechanics problem. Prof. Luehrmann has found that students do more exploration of problems when they can use computer graphics. Getting the picture is what problem-solving is all about.



Arthur Luehrmann was one of the first professors in the country to use time-shared computer graphics extensively in teaching. In his field — physics — a primary stumbling block for beginning students was the lack of an intuitive "feel" that made sense of all the calculus and diagrams. Words and equations simply failed to convey the idea of a system evolving, step-by-step, from an initial condition through a series of neighboring states. Computer algorithms and graphic output did the job.

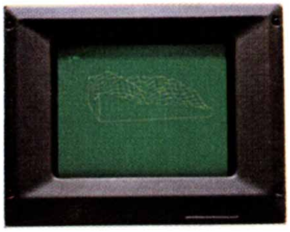
At Dartmouth, in 1968, Prof. Luehrmann began using x-y plotters to convert tables of numbers into pictures. In 1969, Tektronix supplied Dartmouth with an early Model 4002, the first graphic terminal priced under $10,000. Today, Dartmouth has dozens of graphic terminals and the majority bear our name. We are participating in Prof. Luehrmann's development of new graphic syntax in BASIC, the computer language.

Prof. Luehrmann says that the exciting part of using computer graphics in education is that it causes students to formulate problems in a new language. "Not just programming language," he says, "but conceptual language where the student is thinking in terms of procedures and sub-procedures and loops and branches and all of these complex ideas that are very powerful. Students develop new ways of thinking about problems."

Prof. Luehrmann has been something of an evangelist about computer graphics, writing enhancements of BASIC and showing his colleagues, both at Dartmouth and beyond, how this emerging tool can help both their teaching and research.

Students watch business-cycle and growth-pattern data, plotted in a time series, come to life as moving lines on the projection screen. "For teaching, there's nothing like graphics," Prof. McGee says. Working before the class at a graphic terminal instead of a blackboard, while the terminal's picture is projected via a scan converter, Prof. McGee is here *generating* data according to one of the Box-Jenkins forecasting models. While the numbers are being generated they can also be stored while a second program analyzes the data empirically. Exactly the same boxes appear in the *analyzed* data, but this time the autocorrelation and the partials are derived from the data, not from theory. "I use these programs in front of the class to show how the theoretical properties of a time series model can be obscured by 'data noise' when we do the analysis."

Computer graphics helps Prof. Victor McGee teach statistics to students in Dartmouth's Amos Tuck School of Business Administration.



Optimization Prof. Alvin Converse teaches engineering courses involving computer simulation, and regularly uses the lecture hall graphic terminal and TV projector for demonstrating concepts like optimization. "Students are highly motivated by graphic output, as most everyone is," he says. "You can actually watch the progress of the algorithm trying to find the top of the surface. For instance, if the test function is a curved valley, you can watch certain methods get hung up and just not move. And you can watch other methods come in very quickly, or rapidly diverge if they're not very stable. Graphic output has turned out to be an interesting pedagogical tool."

Mapping Prof. Robert Huke, of the Geography Department, finds that using a Tektronix graphic tablet — an electronic grid of thousands of points teamed up with the terminal — saves him and his students days of laborious tracing and drafting. With planimeter mapping, the scientist uses an electronic stylus to trace the area of interest on a map, which has been overlaid on the graphic tablet. The terminal simultaneously shows the shape being traced, and then displays the computer's calculations of the area. Not ready for hard copy? Fine. Store the data, recall it later. For choropleth mapping, graphically expressed values for anything—rainfall, for example—appear exactly where they pertain—measuring stations within a state, for example. The uses go on and on.

From the beginning, students are taught to write programs that make use of the graphic software already in the computer. Program writing is not considered a rote exercise. This is the way Arthur Luehrmann puts it: "When a person is writing a computer program he is cast in the role of a teacher explaining to a perfectly logical infinitely patient, and totally unimaginative student how to carry out a sequence of tasks. It is axiomatic that one never learns a subject so well as when he teaches it to someone else."

These are just a few of the many ways in which computer graphics is helping to improve the learning process at Dartmouth. It is growing and changing constantly, as more people discover that computer graphics is easier to use than they think.

The graphic tools available to Dartmouth students and faculty stem from a large library of graphic software developed by Arthur Luehrmann and his associates. The package of graphic sub-programs is written in Prof. Luehrmann's enhanced BASIC, so that no user need be locked into using any particular manufacturer's hardware.

The central computer is a large Honeywell 66-40, connected to a network of more than 300 terminals in 25 buildings on the campus. In addition, an educational network reaches dozens of other schools and colleges from New York City to northern Maine, and west as far as Chicago.

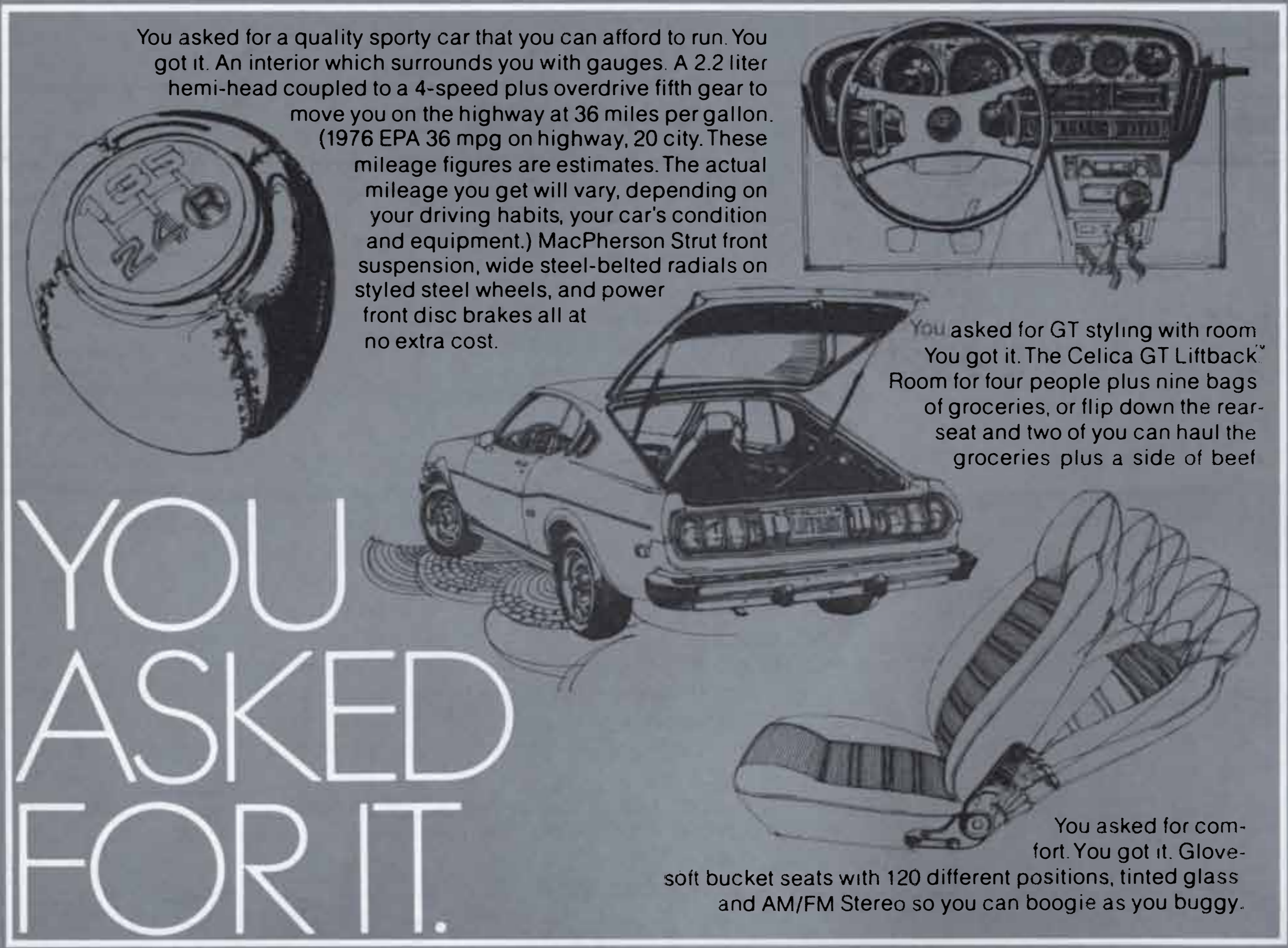
Tektronix has been working hard to remove the cost barriers to even wider use of computer graphics. The new Tektronix 4006 terminal, shown middle left on the opposite page, is priced at only $2995, a new standard of value. The new Tektronix 4051 desktop computer provides graphics power, either standing alone or on-line with a larger computer, for only $7,000.

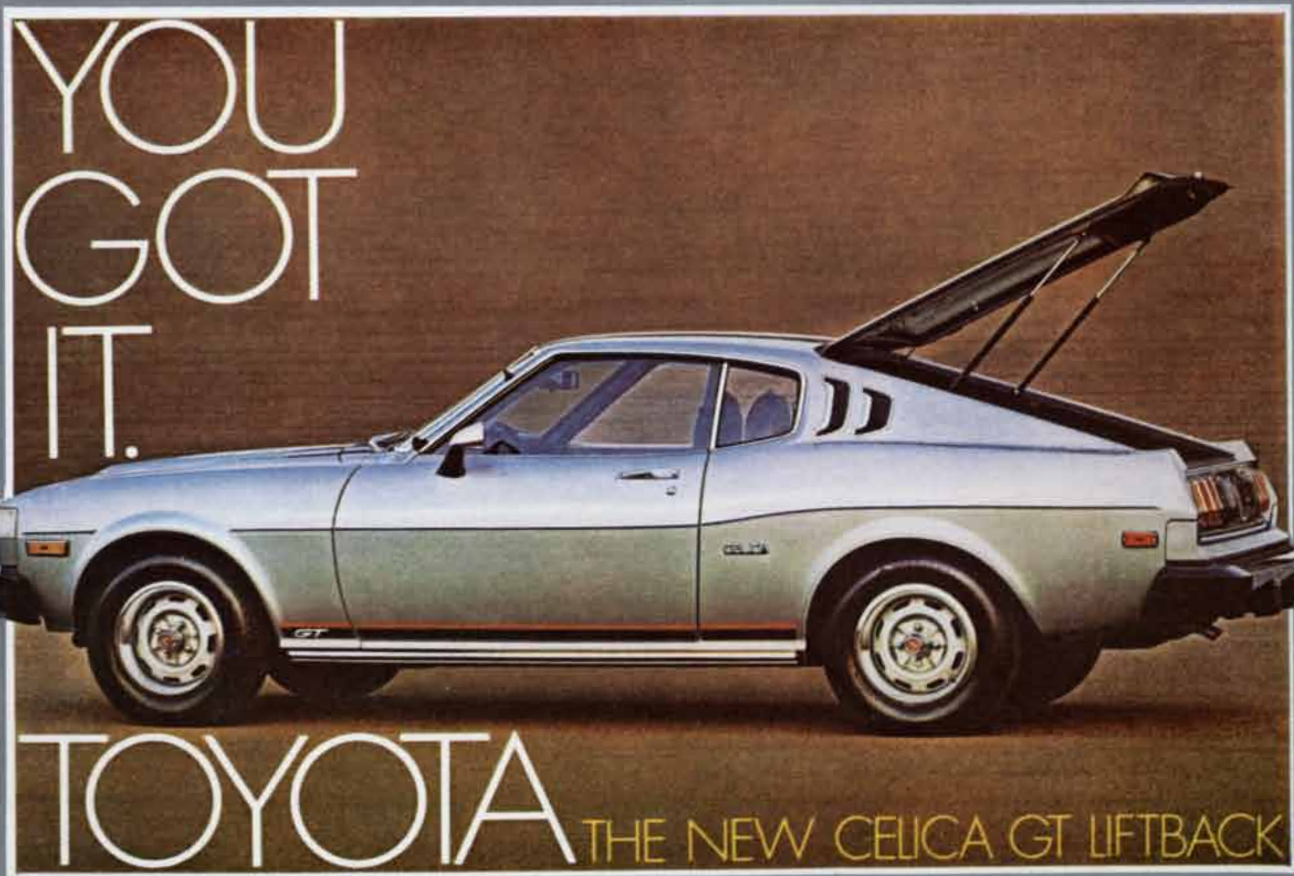
U.S. sales prices FOB, Beaverton, Oregon

These are just two of the 835 Tektronix products and instruments serving the test, measurement and display needs of thousands of customers worldwide. Not only in education but also in science, health, communications and industry. For additional information contact Bob Worsley, Education Marketing Manager. Tektronix, Inc., P.O. Box 500, Beaverton, Oregon 97077

TEKTRONIX®
committed to excellence

Tektronix graphic terminals are now located in many parts of the Dartmouth campus. Here in the Mathematics Department students are visualizing functions in three dimension, working on their own. A nearby hard-copy printer captures permanent records, whenever needed, of any picture that shows on the screen.


You asked for a quality sporty car that you can afford to run. You got it. An interior which surrounds you with gauges. A 2.2 liter hemi-head coupled to a 4-speed plus overdrive fifth gear to move you on the highway at 36 miles per gallon. (1976 EPA 36 mpg on highway, 20 city. These mileage figures are estimates. The actual mileage you get will vary, depending on your driving habits, your car's condition and equipment.) MacPherson Strut front suspension, wide steel-belted radials on styled steel wheels, and power front disc brakes all at no extra cost.
You asked for GT styling with room. You got it. The Celica GT Liftback™ Room for four people plus nine bags of groceries, or flip down the rear-seat and two of you can haul the groceries plus a side of beef.
YOU ASKED FOR IT.
You asked for comfort. You got it. Glove-soft bucket seats with 120 different positions, tinted glass and AM/FM Stereo so you can boogie as you buggy.


YOU GOT IT.
TOYOTA
THE NEW CELICA GT LIFTBACK