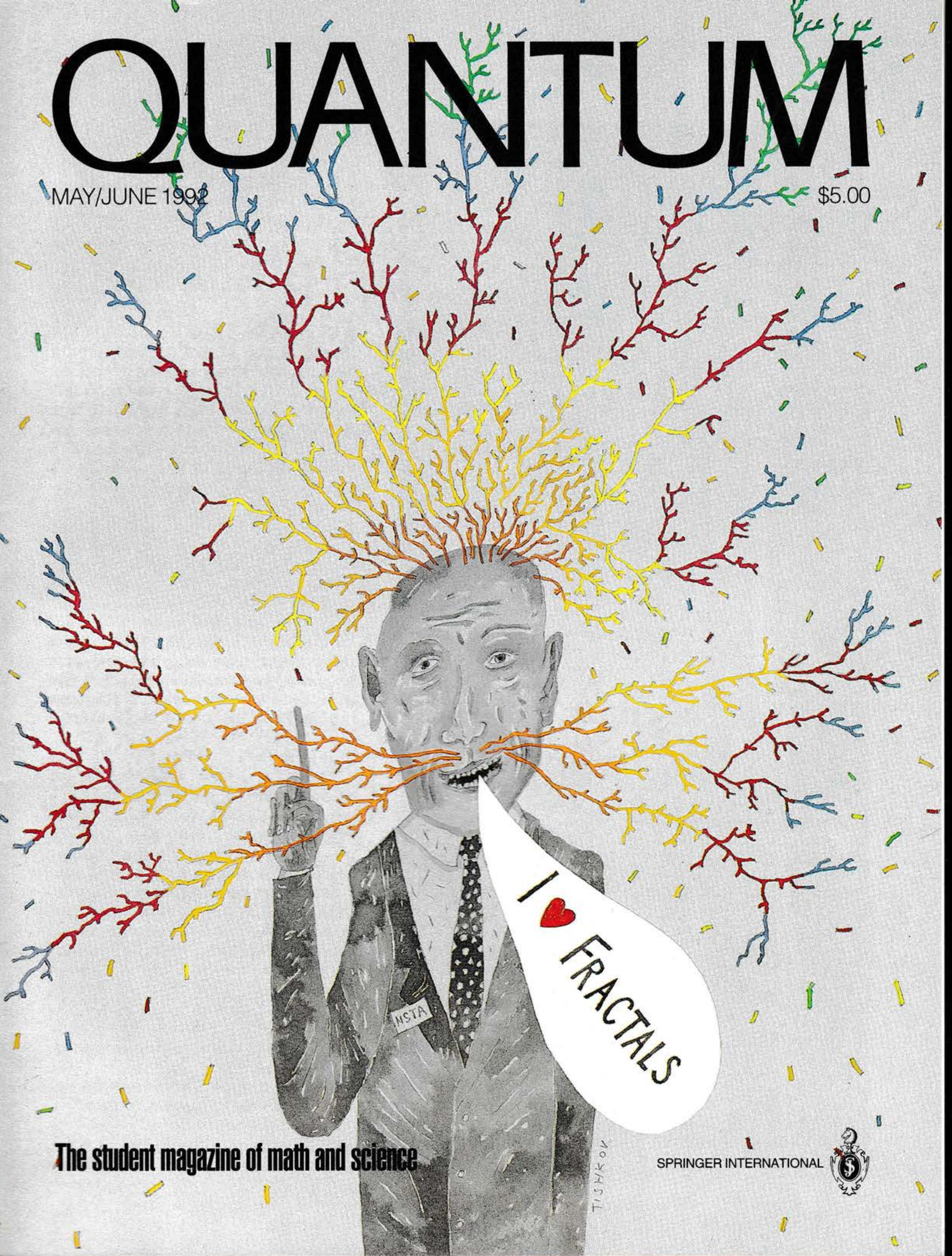
QUANTUM

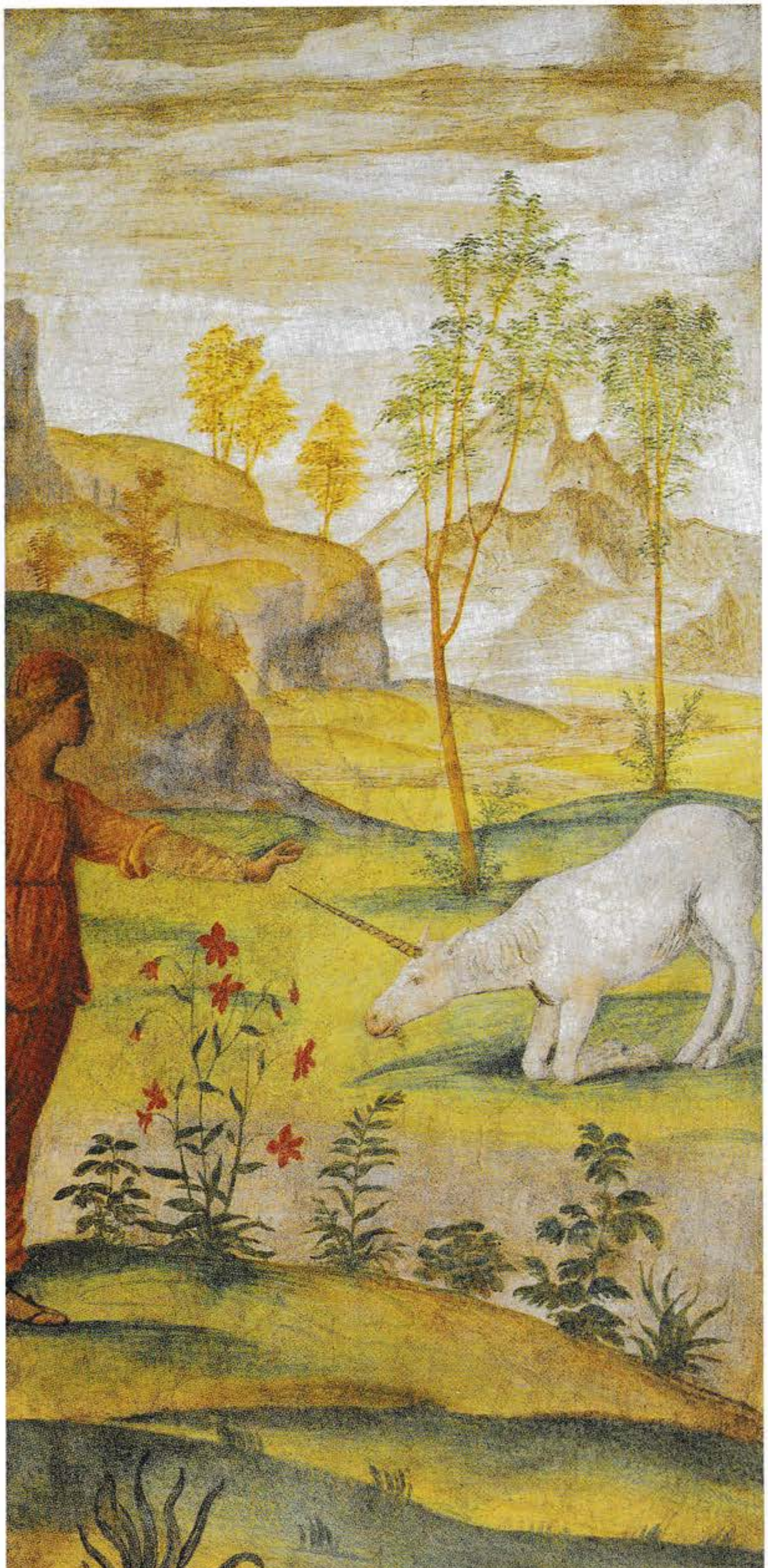
MAY/JUNE 1992

$5.00



The student magazine of math and science

SPRINGER INTERNATIONAL



National Gallery of Art, Washington (Collection of Samuel H. Kress) © NGA

Procris and the Unicorn (c. 1520) by Bernardino Luini

ACCORDING TO ONE LEGEND, the unicorn runs so swiftly no hunter can harm it, but it bows submissively when approached by a virgin. The maiden in this fresco by Bernardino Luini (d. 1532) is clearly unaware of the ways of fantastic beasts—she raises her arm to protect herself. Through a trick of perspective, the point of the beast's horn seems perilously close to her outstretched hand.

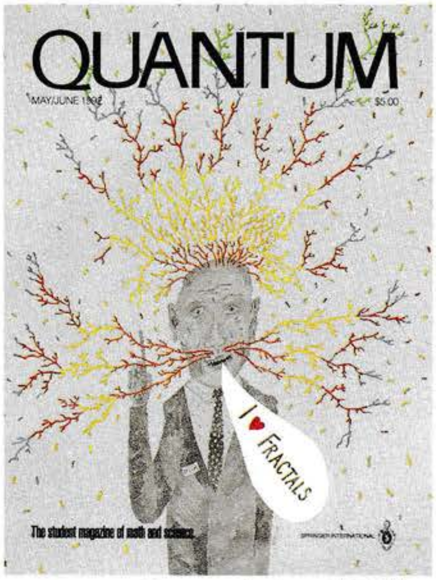
Procris and the Unicorn is one of a series of frescoes about the ill-fated lovers Cephalus and Procris. Cephalus took great delight in hunting, and Procris gave him a javelin that never failed to hit its mark (originally a gift from the goddess Diana). But Procris came to suspect that Cephalus had another lover in the woods where he spent so much of his time. She went to the forest and hid in a bush to watch him. Hearing a rustling, Cephalus mistook her for his prey and killed her with the javelin she had lovingly given him.

The unicorn lives on as an evocative and enigmatic symbol. It was first depicted on Assyrian reliefs and has appeared in Chinese as well as European art through the ages. The first description in Greek literature was penned in the fifth century B.C. by Ctesias (who may actually have been describing the Indian rhinoceros). Ctesias says the *monokeros* (which became *unicornis* in Latin) has a white body, a purple head, and blue eyes; from its forehead protrudes a cubit-long horn, red at the tip, black in the middle, and white at the base. It was thought that those who drink from its horn are protected from stomach trouble, epilepsy, and poison.

Believe it or not, all this has something to do with the Kaleidoscope in this issue.

QUANTUM

MAY/JUNE 1992 VOLUME 2, NUMBER 5



Cover art by Leonid Tishkov

Who *doesn't* like fractals? In addition to making abstract mathematics stunningly visual, they have opened new paths to an understanding of the physical world. Fractals are intimately related to chaos theory, which explores the interconnected order and disorder in phenomena as diverse as plant growth and the behavior of weather systems.

One young woman who likes fractals is Ashley Reiter of North Carolina, who wrote the winning paper in the 1991 Westinghouse Science Talent Search. Ashley used a certain definition of dimension (the "Hausdorff dimension") and found a research article in which this dimension was determined empirically. She obtained contradictory theoretical result. In the course of investigating this discrepancy, Ashley determined the dimensions of fractals generated by Pascal's triangle and its higher-dimension analogues.

Turn to page 6 for a look at this hot topic. (The wild hair on the cover is a fanciful rendering of diffusion-controlled aggregation, which is discussed in the article.)

FEATURES

6 Chaos & Order
A ride on Sierpinski's carpet
by I. M. Sokolov

12 Contemplation & Calculation
The universe discovered
by Yury Solovyov

24 Remainders & Congruences
Go "mod" with your equations
by Andrey Yegorov

34 Ways & Means
The power of dimensional thinking
by Yuli Bruk and Albert Stasenko

DEPARTMENTS

2 Publisher's Page

19 How Do You Figure?

20 In the Lab I
Neutrons seek the murderer!

29 Brainteasers

30 Math Investigations
Triangles of differences

32 Kaleidoscope
One, two, many

40 Physics Contest
Shake, rattle, and roll

44 Looking Back
Calculating π

52 In the Lab II
Double, double toil and trouble

54 Happenings
Bulletin Board

56 Answers, Hints & Solutions

64 Toy Store
Further adventures in Flexland

PUBLISHER'S PAGE

Russian bazaar

Further notes of an American traveler

IN MARCH, DURING ONE OF my periodic visits to Moscow, I had a chance to visit Izmaylovo Park, located northeast of the Kremlin on the outskirts of the city. In the old days Izmaylovo was a baronial estate; in Soviet times it became a public recreational facility; now it's a huge open-air bazaar and flea market. Everything is for sale: icons, quilts, coins, stamp collections, furniture, clothing, Communist Party memorabilia . . .

Спасение утопающих—
дело рук самих утопающих—
Ильф и Петров

A drowning man must use
his own arms to save himself—
Ilf and Petrov

The Izmaylovo phenomenon isn't unique. Muscovites are gathering elsewhere to buy and sell in this rudimentary market economy. Entire city blocks in Moscow are set aside for such trade in private—one might say personal—property. It's an astonishing sight.

I had mixed feelings as I worked my way through the dense, shuffling mass of bundled-up humanity. I admired the Russian people's will to survive, their patience, their solidarity in bad times, their peculiar blend of optimism and fatalism. But I couldn't help feeling sad as I saw them selling their personal belongings in the street. It seemed as if they were selling their past—their family histories as well as their shared national experience—for their daily bread.

There is talk of redistributing land to those who can farm it properly. But my friends in Moscow told me any action would be too late for the spring planting. They said the foodstuffs people had put away were depleted this past winter. Amid the political talk that the Russian economy will likely turn the corner in six months to a year, Muscovites are already thinking ahead to the winter of '92–'93 and wondering how they will get by.

В чужой монастырь
со своим уставом не ходят—
русская пословица

Don't bring your rules
to our monastery—
Russian saying

Many economic "monks" from the West have been eager to impose their rules on the shambling Russian economy. They insist on a transition to a kind of "free market" that doesn't even exist in the West. They would in effect make

A small portion of the crowd at Izmaylovo Park. I would estimate that there were at least 30,000 people there.

aid and loans contingent on a high unemployment rate and the removal of social programs, most of which constitute the kind of "safety net" that has become (after years of struggle) an integral part of the Western democracies.

Other economists caution the Russians against a headlong dive into market capitalism without adequately protecting the weakest members of their society. But their voices seem lost in the chorus cheering Russia on to a harsh new economic plan, one that gives free rein to well-positioned Russians and non-Russians to exploit this potentially rich but unstable land.

Соловья баснями не кормят—
русская пословица

Nightingales don't live on fairy tales—
Russian saying

While in Moscow I attended several wonderful performances of ballet and folk dance. The theaters were packed with ordinary people, not VIPs. They had paid 6–8 rubles to get in—less than 10 cents at the current exchange rate. I feared I was seeing the last instances of open access to the arts in Russia. When monetary reforms take hold, tickets will cost hundreds of rubles, just as theater and symphony tickets in the US cost $25 (or more), putting them out of the reach of many working people.

True, Russia and the other republics in the Commonwealth of Independent States need a new class of businessmen and economists if they are to join the world economy and materially improve the lives of their citizens. Individual initiative and self-interest are important values in a market economy. But I hope the people of the CIS don't lose sight of other values that have produced great scientists, artists, and thinkers. One such value is concern for the common good and for those at the bottom of society. You can't pull yourself up by your own bootstraps if you have no boots.

—Bill G. Aldridge

QUANTUM

THE STUDENT MAGAZINE OF MATH AND SCIENCE

A publication of the National Science Teachers Association (NSTA)
& Quantum Bureau of the Russian Academy of Sciences
in conjunction with
the American Association of Physics Teachers (AAPT)
& the National Council of Teachers of Mathematics (NCTM)

Quantum (ISSN 1048-8820) is published bimonthly by the National Science Teachers Association in cooperation with Springer-Verlag New York Inc. Volume 2 (6 issues) will be published in 1991–1992. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by the Russian Academy of Sciences and the Russian Academy of Pedagogical Sciences, as well as original material in English. **Editorial offices:** NSTA, 3140 North Washington Boulevard, 2nd Floor, Arlington, VA 22201. **Production offices:** Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010. Production Editor: Madeline R. Kraner

Advertising:

Advertising Representatives: (Washington) Paul Kuntzler (202) 328-5800; (New York) Thomas Heitzman (212) 460-1675; and G. Probst, Springer-Verlag, Heidelberger Platz 3, 1000 Berlin 33, FRG, telephone (0) 30-82 07-1, fax 30-82 07-1.

Second class postage pending at New York, NY, and additional mailing ofices. **Postmaster:** send address changes to: *Quantum* Springer-Verlag New York, Inc., Journal Fulfillment Services Department, 44 Hartz Way, Secaucus, NJ 07096-2491.

Subscription Information: North America

Student rate: $14; Personal rate (nonstudent): $18; Institutional rate: $28. Rates include postage and handling. (Canadian customers please add 7% GST to subscription price. Springer-Verlag GST registration number is 123394918.) Subscriptions begin with next published issue (backstarts may be requested). Bulk rates for students are available. Send orders to *Quantum*, Springer-Verlag New York, Inc., P.O. Box 2485, Secaucus, NJ 07096; or call 1-800-SPRINGER (in New York, call (201) 348-4033).

All Countries Outside North America

Subscription rates in U.S. currency as above (all rates calculated in DM at the exchange rate current at the time of purchase) plus postage and handling. SAL (Surface Airmail Listed) is mandatory for Japan, India, Australia, and New Zealand. Customers should ask for the appropriate price list. Air mail delivery to all other countries is available upon request. Orders may be placed through your bookseller or directly through Springer-Verlag, Heidelberger Platz 3, 1000 Berlin 33, FRG.

This project was supported, in part, by the
National Science Foundation
Opinions expressed are those of the authors and not necessarily those of the Foundation


They were talking about Fractals TISHKOV

A ride on Sierpinski's carpet

And a cruise along an infinite coastline

by I. M. Sokolov

THE OBJECTS CALLED "fractals" were initially created in the imagination of mathematicians at the beginning of this century. Hardly anyone would have thought that there might be anything like these far-fetched and elegant curves in nature. And though this article will mostly touch on physical systems, it will have to begin with a short, nonrigorous mathematical introduction.

Self-similarity

A self-similar geometrical figure (or solid) is a figure that can be cut into a finite number of identical figures similar to it.

Let me remind you of the general definition of similarity: two figures are called similar if they have the same shape, even though their sizes may be different; so one of them is an enlarged or diminished copy of the other. More exactly, one of two similar figures can be mapped onto the other so that the distance between any two of its points is increased or decreased in the same ratio, called the *ratio of similarity.* Examples of self-similar figures are given in figure 1: a straight line segment, an equilateral triangle, a square, and a cube.

Figure 1

Art by Leonid Tishkov

The object presented in figure 2 looks a bit more complicated, but it's constructed quite simply. We start with an equilateral triangle with side length l_0 and repeat (to infinity) the following procedure: every straight segment of the curve obtained on the previous step is divided into three parts and the middle part is replaced by two segments of length $l/3$, where l is the length of the initial segment. The first stages of this procedure are seen in figure 2. At the nth stage the curve becomes a polygonal line consisting of $3 \cdot 4^n$ line segments, each $l_0/3^n$ units long, and its total length is

$$L = 3l_0(4/3)^n.$$

This polygonal line is called the *Koch triadic curve* or *Koch snowflake* (after the Swedish mathematician who invented it).

Strictly speaking, the Koch snowflake isn't a self-similar object according to our definition above. But it consists of three self-similar curves "grown" from the three sides of the initial triangle: each of the four segments that replaced a side of the initial triangle at the first step of construction eventually yields a curve similar to the one produced by the whole side (with a ratio of 1/3). Moreover, any segment of the polygonal curve obtained at the nth step of the construction also produced a similar curve, the ratio of similarity being 3^{-n}.

The objects in figure 3 are also self-similar. They're called the *Sierpinski triangle* and the *universal Sierpinski curve,* or "Sierpinski carpet," after the Polish mathematician Waclaw Sierpinski (1882–1969). You can see how they're constructed: the first is obtained by repeatedly joining the midpoints of the sides of corresponding equilateral triangles; the second is obtained by infinitely repeating the procedure of cutting out the center portion of a square divided into nine square parts.

Now let's get back to the Koch curve and try to determine its length with a compass. We can do this, for example, by opening the feet of the compass to some length λ and marking off steps of length λ along the segments of the curve. The length L of the curve is then approximately λn,

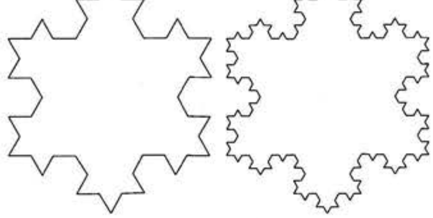


Figure 2

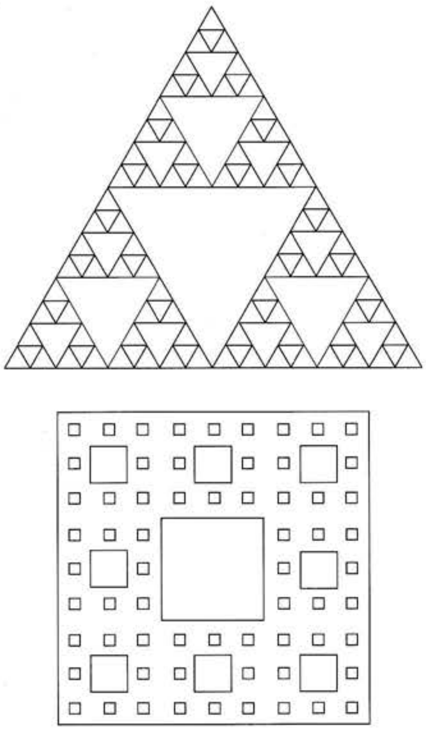


Figure 3

where n is the number of steps we've taken. The value of λ is called the scale of measurement.

Let's see how this process looks when used on a circle of radius $R = 1$. For $\lambda = 1.0$ m, we find $L = \lambda n = 6$ m. For $\lambda = 0.1$ m, we get $L = 6.2$ m; and for $\lambda = 0.001$ m, we get $L = 6.28$ m. As $\lambda \to 0$, L tends to the limit of $2\pi R = 6.28318\ldots$ m.

But if we try to repeat the same procedure with the Koch curve, we'll be convinced that there is no limit that could be considered the length of the curve. Choosing the scale $\lambda = l_0/3^n$, we find that the measured length of the curve is equal to the length of a polygonal line corresponding to the nth stage of construction—$L = 3l_0(4/3)^n$; so it grows indefinitely as $n \to \infty$.

Attempts to measure the lengths of other self-similar curves would give analogous results: as the scale of measurement decreases, the length of the curve increases without limit.

Here I should point out one very important factor that distinguishes a real self-similar object from an ideal mathematical one: real objects have a minimum scale of measurement $\lambda_{\min}$.

For example, let's take the actual process of constructing the Koch curve with paper and pencil. Suppose we construct a curve starting with a triangle whose side is 1.0 m long and use a pencil that produces a line a_0 = 0.1 mm = 10^{-4} m wide. From the mathematical point of view, the procedure for constructing the curve can last for ever. But the real process will stop as soon as the length of a straight segment between two neighboring corners becomes comparable to the width of the line. It's easy to calculate that this will occur at step $n = \ln(l_0/a_0)/\ln 3 \cong 9$. The length of our line will be $L \cong 40$ m. So the real self-similar curve has a finite length.

Now let's return to ideal mathematical objects. The formula for the length of the Koch curve can be expressed as

$$L = A\lambda^{-\alpha}, \qquad (1)$$

where $A = 3l_0^{\ln 4/\ln 3}$, $\alpha = \ln 4/\ln 3 - 1$. (You can prove to yourself that this expression is equivalent to the formula $L = 3l_0(4/3)^n$.) The exponent α has to do with the dimension of the curve.

What is dimension?

There are several definitions of dimension, based on completely different ideas. Let's take a look at a few of them.

The first definition has to do with the number of coordinates needed to unambiguously locate a point. In our space this number is three; on the plane, two coordinates are enough; on a straight line, only one coordinate is needed. In this sense space is three-dimensional, a plane is two-dimensional, and a line is one-dimensional. Naturally, according to this definition, dimension is always an integer number.

A second way of defining dimension is based on the observation that in order to cut a figure into disconnected parts, one merely has to remove a set whose dimension is 1 less than that of the figure. For instance, to dissect a line one merely removes one of its points; to dissect a plane figure, we can cut it along some curve; and to dissect a solid, we can cut it along some surface. So dimension can be defined inductively: the dimension 0 is assigned to a single point, or more generally, to any finite or infinite but countable set (that is, a set whose points can be enumerated by the numbers 1, 2, 3, ...); and the dimension of any other set is assumed to be 1 more than the dimension of the section dividing it into disconnected parts. Such a dimension, called *inductive*, again is always an integer.

Now let's move on to a third, and for us the most interesting, definition of dimension—or rather, to the definition of a whole class of like notions of dimension. The simplest of them is the *dimension of self-similarity*.

The dimension of self-similarity D can be defined by the formula

$$D = \frac{\ln N}{\ln n},$$

where N is the number of identical parts into which the given self-similar object can be partitioned and n is the ratio of similarity of the object to its parts. Look at figure 1. Cutting a square as depicted there, we divide it into $N = 4$ squares with sides half as long as those of the original square ($n = 2$). The cube with side length 1 consists of $N = 8$ cubes with side length 1/2 ($n = 2$). So the dimension of self-similarity for a square is $\ln 4/\ln 2 = 2$; for a cube it's $\ln 8/\ln 2 = 3$; and, obviously, the dimension of a straight line segment is 1.

If we calculate the dimension of the objects shown in figures 2 and 3 in the same way, we'll see that the dimension of any segment of the Koch curve (and the dimension of the whole curve) is $D = \ln 4/\ln 3 \cong 1.2618$; for the Sierpinski triangle it's $\ln 3/\ln 2 \cong 1.5849$; for the Sierpinski carpet it's $\ln 8/\ln 3 \cong 1.8727$. These strange curves have a fractional dimension.

Now let's go back to formula (1) for the length of the Koch curve. Using the above definition of dimension D we can rewrite the formula as

$$L = 3l_0^D\lambda^{1-D}.$$

We see that the rate of growth of the measured length of a self-similar curve as a function of the decreasing scale of measurement depends on the dimension D of the curve. More exactly, L/λ—which is approximately the number of steps we made with our compass in measuring the curve—is proportional to λ^{-D}. And this prompts a new definition of dimension.

How do we measure dimension?

The dimension of self-similarity can be determined only for very regular objects that are constructed according to definite rules. If the deviations from regularity are small, the object can be considered approximately self-similar. But what happens if they're large?

Let's use another definition of dimension, one that is often used to experimentally measure the dimension of various physical systems.

The space in which the investigated object is situated is divided into boxes with side length λ (for instance, a square grid with side λ is drawn on the plane of a photo of the object). Boxes that contain points of the object are tallied. The partition is repeated at a smaller scale $\lambda' < \lambda$ (fig. 4). The dependence of the number of boxes containing points of the object on the size of the box is expressed by the law $N = A\lambda^{-D}$, where A is a constant and D is the unknown dimension. Investigating a flat region with area S (such as the triangle in figure 4), we can easily prove that $N \cong S/\lambda^2$, so $D = 2$. For a line segment, $N = BL/\lambda$, where L is the length of the line segment and B is a coefficient that depends on its orientation. A line segment's dimension D is 1. If we repeat this procedure with the objects in figures 2 and 3, we'll obtain values of D coinciding with their dimension of self-similarity. To determine the dimensions of real objects, the graph of $\ln N$ as a function of $-\ln \lambda$ is drawn. It is a straight line whose slope give us the value of D.

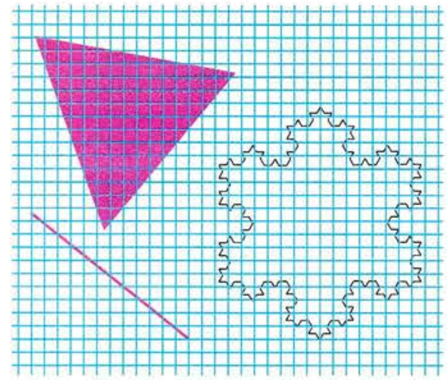


Figure 4

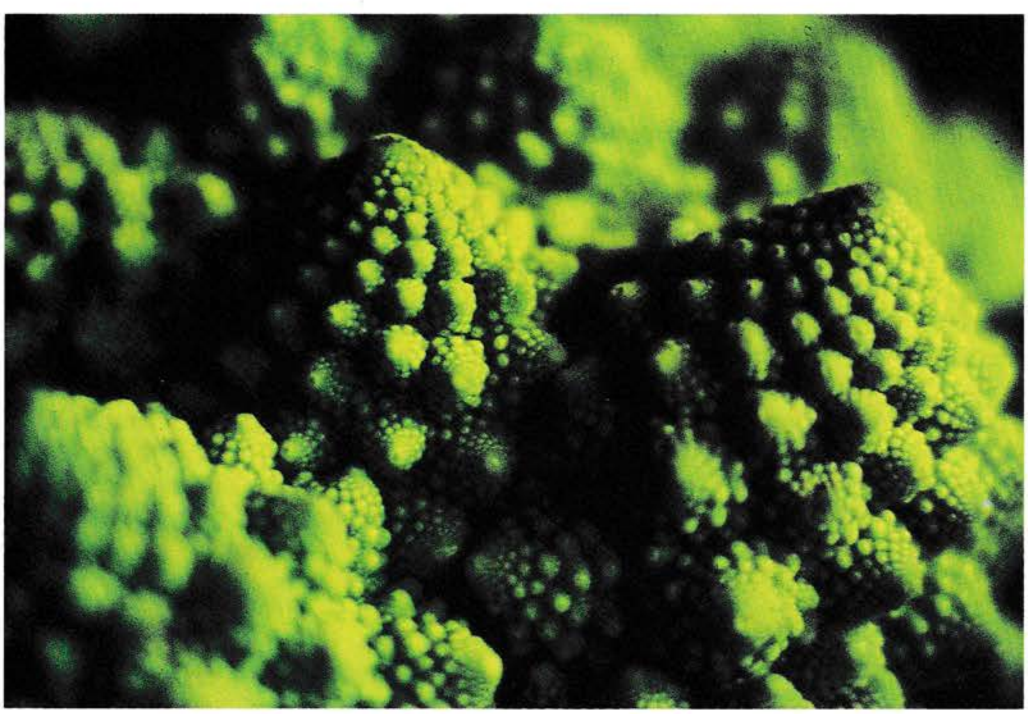
Natural fractals

In 1961 an article by the English scientist L. Richardson (1881–1953) appeared that was devoted to the measurement of the length of coastlines. The author proved that the measured length of a seacoast increases as the scale decreases according to the law $L = A\lambda^{-\alpha}$ (Richardson's law), where the exponent α for the British coast, for example, equals 0.24; for the Australian coast it's 0.13. And though this law resembles the formulas for the length of self-similar curves, this work of Richardson existed independently. In physics there were some other examples related to self-similar objects. But it was all so piecemeal . . .

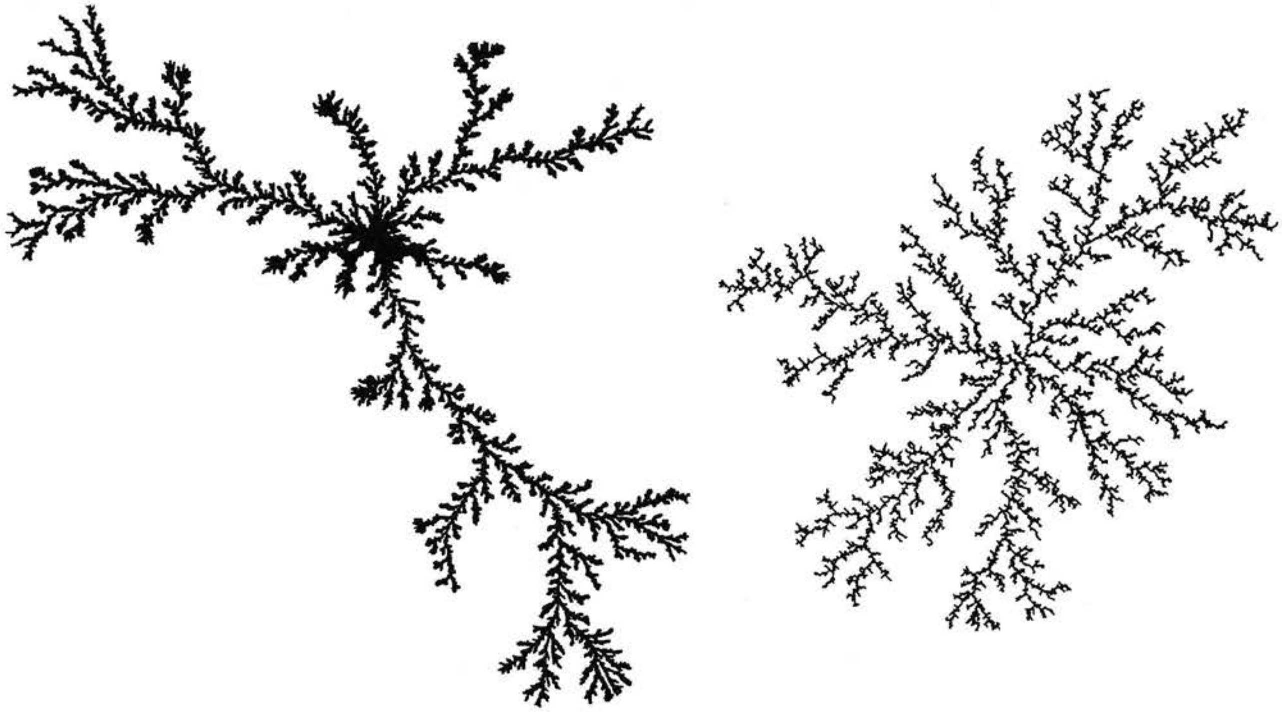
Everything changed drastically with the publication of a book by Benoit Mandelbrot (a French mathematician now working in the United States). It was published in 1975 in French and in 1977 in English. This book brought together many mathematical and physical examples and made them the common property of scientists everywhere. But Mandelbrot's greatest service was thinking up a name for all this.

Maybe you remember the primary contribution of the character Athos in *Twenty Years Later* by Dumas—he came up with a name for the operation: "The Family Affair." This stroke was considered equal to the sword of d'Artagnan and the money of Porthos. To coin a good name is a great achievement.

For objects of nonintegral dimension—or



A close-up view of broccoli Romanesco, a cross between cauliflower and broccoli. This hybrid displays a striking degree of self-similarity: the individual nodules are smaller versions of the entire cluster; the nodules are in turn composed of smaller nodules shaped like the larger ones; and so on. (From Fractals for the Classroom *by Peitgen, Jürgens, and Saupe, New York: Springer-Verlag, 1992)*



The object at left is an instance of diffusion-controlled aggregation of zinc at the interface between a solution of zinc sulfate and n-butyl-acetate. The figure at right is a computer simulation of the same phenomenon based on the Brownian motion of single particles. (From Fractals for the Classroom*)*

rather, for objects whose dimension measured by the last of the above methods is greater than their topological dimension—Mandelbrot invented the word "fractal." This word comes from the Latin word *fractus*—fractional, broken.

Mandelbrot's first book was called *Fractals: Form, Chance, Dimension.* His second, published in 1982, was entitled *The Fractal Geometry of Nature*—and the title couldn't have been more apt.

Many geographical objects have fractal properties: shorelines, rivers, mountains, canyons. The borders of countries, if they correspond to natural landmarks and aren't drawn on the map with a ruler and then determined on location (like the border between Egypt and Sudan), are also fractals. The length of the Portuguese–Spanish border (given in Portuguese reference books) and the length of the Spanish–Portuguese border (according to official Spanish data) differ by 20% because of the different scales used. This proves once again that the notion of the length of fractal curves doesn't make too much sense.

It turned out that curves like the Koch curve are the rule rather than the exception in nature. It's clear that the self-similarity of real natural objects is violated by accidental deviations from strict regularity. For example, different parts of a coast aren't identical—they just resemble each other. And all real systems have a minimum scale of measurement. These circumstances should be taken into account when you analyze any physical situation.

In order to discuss the fractal properties of a system, the difference between the minimum and maximum scales must be large enough. If we take a shoreline, the maximum scale will be about 1,000 km = 10^6 m, and the minimum scale, determined by the instability of the shore because of waves, tides, and so on, is of the order of 1^{-10} m. These scales differ by a factor of a million!

Another example of fractal curves is the visible edge of a cloud. Here the difference between the minimum and maximum scales is even larger: there are data on clouds from several hundred meters across, which have visible details of about 1 m, to the size of the Earth (cyclonic regions). The dimension of a cloud edge is $D = 1.35$.

So far we've limited our discussion to fractal curves, by and large—that is, to extremely convoluted lines like the Koch curve; our geographic examples were mainly curiosities. But there are many physical processes that create more complicated and more important fractal structures.

No doubt many of you have grown crystals from a saturated solution. If the solution isn't oversaturated and it's well mixed, then a beautiful, regular crystal will grow from a "seed" hanging on a thread immersed in the solution. The crystal grows because some molecules in the course of their thermal motion approach places on the surface where they can "stick," occupying the most advantageous position with regard to their energy state. Of course, most of the molecules land in less favorable places, but sooner or later they move back into the solution because their bonds with the crystal aren't strong enough. Due to this equilibratory growth we obtain a crystal without cavities and with perfectly smooth, flat facets.

If crystallization and dissolution aren't in equilibrium (which can occur with rapid crystallization from an oversaturated solution or crystallization from the gas phase), crystals of another type appear. You've seen the frosty overcoat your freezer acquires from time to time, and the icy patterns on your windows in the winter. These rather porous formations arise from the condensation of water in the air. At first, separate clusters of molecules are formed; then they multiply and unite, and the clusters create patterns. The conditions for the growth of clusters resemble the conditions for the formation of snowflakes in clouds.

This process of growth, called diffusion-controlled aggregation, causes the formation of small fractal-shaped crystals called dendrites. The fractal dimension of dendritic crystals is determined by the specific mechanisms of their growth. Depending on the interaction of the molecules forming the crystal and on the size of the crystal, the dendrite may have a random, irregular form or, on the other hand, it may seem to be a perfectly regular figure—like a snowflake, for example. But in reality we can speak of the regular form of a snowflake only if the scale is large enough (the size of the snowflake itself); on a smaller scale there is no regularity—this is a reflection of the random processes that led to its formation.

The existence of a minimum scale (which in this case may be of the same order as, or much larger than, a molecule) means that total number of molecules in a crystal (or its mass) depends on its size according to the law $N_{\text{mol}} \sim M \sim l^D$. So we can determine the dimension of dendritic crystals by using the dependence of their mass on their size.

Shapes that look very much like dendritic crystals can arise in dielectrics. If a strong spark strikes a dielectric plate, it leaves a distinct pattern on the surface—so-called Lichtenberg figures, named after the German physicist and experimenter who discovered them in the 18th century. The resemblance of Lichtenberg figures and dendritic crystals is no accident—their formation is theoretically described by similar equations.

The fractal dimension is a very important and measurable characteristic of a physical system. It can also be calculated by means of various theoretical models. By comparing the measured and calculated values, one can decide which model is better. In addition, when we calculate the physical properties of fractal systems (for example, the resilience of snow and other porous materials), we can use a mathematical method developed specially for this instance.

Many systems that have long been used for practical purposes have fractal properties. For example, the surface of activated charcoal, used as an absorbent in protective masks, is fractal. Its dimension is greater than 2; it has an extremely large area (formally infinite, in the sense that the Koch curve is infinite); and it has holes of all sizes that can catch and firmly hold particles of any size, from a speck of dust to a large molecule. The surfaces of many solid catalysts used in chemistry also are fractal. Their catalytic activity depends on the fractal properties of their surfaces, which are determined by the method used to prepare and process them.

We've gotten to know many objects of noninteger dimension. So the question arises: is the space we live in three-dimensional? We can give a definite answer to this question. The fractal dimension of space determines the expression of many familiar physical laws. For example, the exponent 2 in the denominator of Coulomb's law $F = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}$ is in fact $D - 1$, where D is the fractal dimension of space. Analysis of data collected to verify physical laws whose formulas depend on the dimension of space have shown that its dimension differs from 3 by not more than 10^{-10}. Our space is indeed "very three-dimensional."

Instead of a conclusion

The history of the study of fractal systems is rather instructive. At first fractals seemed like a mental game of those engaged in pure mathematics, and researchers in the natural sciences showed no interest in these objects. At the same time, there were some poorly understood facts (like the immeasurability of coastlines) that weren't important enough to attract general attention and not interesting enough to investigate for their own sake. The number of such facts kept growing, but they were still isolated and of little interest. Then they were given an all-encompassing name and soon (after only 10 years!) the "fractal boom" began in physics. One scientist even called fractals the infection of the late 20th century.

Why did fractals catch on? First of all, it turned out that we're surrounded by such systems and that we encounter them practically every day. Second, such objects have many unusual properties. Without understanding these properties, we can't understand even such simple things as the shapes of clouds or snowflakes. Third, everything turned out to be more complicated than it seemed at first: a fractal must be described not by a single fractal dimension but by a set, a spectrum of different dimensions, each of which becomes equal to the dimension of Euclidean space as soon as we pass from fractals to ordinary bodies. The different properties of fractal systems depend on the different dimensions. Fourth . . . Fifth . . . Tenth . . . —new questions arise more quickly than the old ones are answered.

Many theories have passed through the stage of accumulating questions before achieving harmony and completion. So the best time for fractals is still to come.

The universe discovered

How misconceptions about the heavens were stripped away, one by one

by Yury Solovyov

OF ALL THE IMAGES of nature that unfold before the human eye, the most magnificent is the view of the starry sky. From ancient times this sight has stirred the human imagination, evoking that powerful current of thought we now call science. How did mankind penetrate the secret of the universe, the secret of the motion of the heavenly bodies?

The first notions of the ancients about the universe were based on what was directly visible. The ancient Egyptians and Babylonians identified the universe with the Earth, which they took to be an enormous disk floating in a boundless ocean. They envisioned the sky as an overturned cup resting on a plane, the inner surface of the cup sprinkled with stars. The stars form ordered configurations, called constellations. The constellations remain unchanged with each passing day, year, or even century. During the night the stars rotate about a stationary point, located nowadays near the North Star, as if the cup as a whole were rotating about an axis that passes through this point and the observer's eye. Careful observations show that the cup makes a complete rotation in 23 hours, 56 minutes.

Because of the cup's rotation, some stars set in the west while others are rising in the east. These observations make one think that the cup is part of a complete sphere and that the stars are distributed on its surface. This sphere is called the *celestial sphere* or the sphere of fixed stars. The two stationary points at which the celestial sphere intersects the rotation axis are called the *celestial poles*. An imaginary circle on the celestial sphere whose points are equidistant from both poles is called the *celestial equator*.

Spherical Earth

So, observations of the diurnal (daily) rotations of the stars led ancient astronomers to the concept of the celestial sphere. But it was harder for them to draw the conclusion that the Earth is shaped like a ball. Ancient Greek philosophers came up with this idea as early as the beginning of the fifth century B.C. They relied on accounts of travelers who noticed that the height of the north celestial pole increased as they moved to the north and decreased as they moved southward (fig. 1).

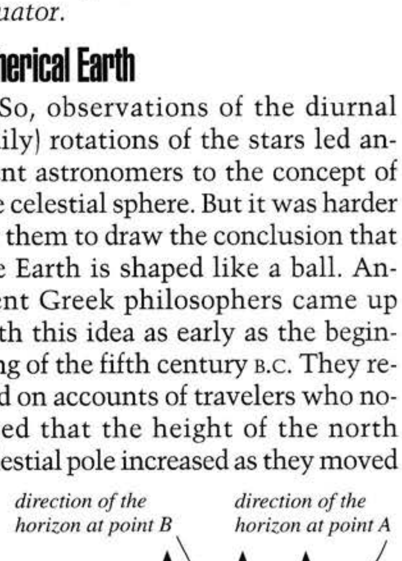


Figure 1

The first true measurement of the Earth's radius was carried out by Eratosthenes (ca. 276–ca. 194 B.C.), a Greek who was a native of Syene (now Aswan) in southern Egypt. While still a youth, he noticed that in Syene every year at noon on June 21 the Sun stays exactly overhead and that tree trunks cast no shadows. Later, in Alexandria, which is in northern Egypt, he discovered that shadows don't disappear there on the same date, and it dawned on him (a stroke of genius!) that this is due to the curvature of the Earth's surface. Alexandria is 770 km north of Syene, so when the Sun is at its zenith over Syene, it must be some angle away from the zenith over Alexandria. This angle α can be measured by observing the shadow cast by a tree trunk or column at Alexandria (fig. 2) at the moment when no shadow is cast at Syene (at noon on June 21). Angles α and β are equal as alternate interior angles formed by a line intersecting two parallel lines. The measured angle was $\alpha = 7°$, so an angle of 7° whose vertex lay at the center of the Earth swept a circular arc 770 km long on the Earth's surface. Since a full circle comprises 360°, the Earth's circumference must be 39,600 km, and its radius must be about 6,400 km (precise modern values are 40,200 km for the circumference and 6,378 km for the radius).

Art by Elena Trofimova

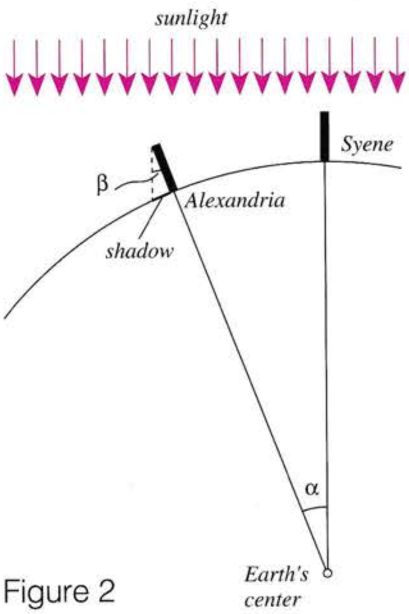


Figure 2

The idea of a spherical Earth made it possible to simplify the geometric model of the universe. It was only natural to take the terrestrial and celestial spheres to be concentric and the rotation axis of the celestial sphere to be an extension of the Earth's polar axis.

Wandering stars

Besides fixed stars, we can also see heavenly bodies on the celestial sphere that change their positions from day to day. These bodies are called planets because *planetes* in ancient Greek means "wanderer." Seven such "wandering stars" were known from ancient times: the Moon, Mercury, Venus, the Sun, Mars, Jupiter, and Saturn.

To understand how the Sun moves along the celestial sphere, let's recall that a day (divided into 24 hours) is the period between two successive sunrises. Since the celestial sphere makes a full rotation about its axis in slightly less than a day (23 hours and 56 minutes), the Sun must move along the celestial sphere in the direction opposite to the sphere's rotation. That's why every day the Sun rises four minutes later than the stars that rose with the Sun the day before. By marking the position of the Sun with respect to the stars every day at sunrise, we can trace its trajectory along the celestial sphere. The trajectory turns out to be another circle whose center coincides with the Earth's center, while its plane is tilted at an angle of 23°30′ to the celestial equator. Along this circle, called the *ecliptic*, the Sun moves from west to east with an almost constant angular velocity, approximately equal to 1° per day, and makes a full revolution in about 365 days, 6 hours.

The Moon also continually moves with respect to the stars. Its orbit is likewise a circle with the Earth at its center. The plane of the circle is tilted at an angle of 5° to that of the ecliptic (fig. 3). The Moon moves almost uniformly along its trajectory in the same direction as the Sun (that is, opposite to the diurnal rotation of the celestial sphere), making a full turn in slightly more than 27 days. Because of this motion, the Moon, like the Sun, falls behind the stars in their diurnal rotation, though the time lag for the Moon is greater: it's not 4 minutes, as for the Sun, but almost a full hour.

The other five wandering stars also move along the celestial sphere, but their motions are much more intricate than those of the Sun and Moon (fig. 4a). Ancient astronomers divided these five planets into two groups—the inferior planets (Mercury and Venus) and the superior planets (Mars, Jupiter, and Saturn) according to their

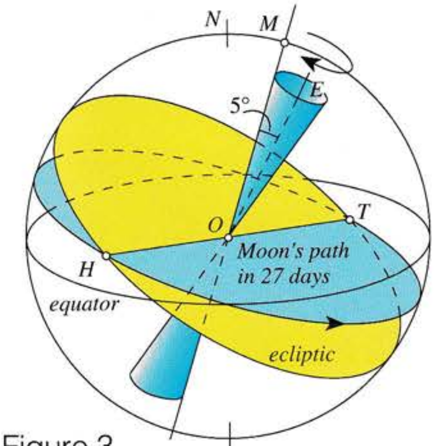
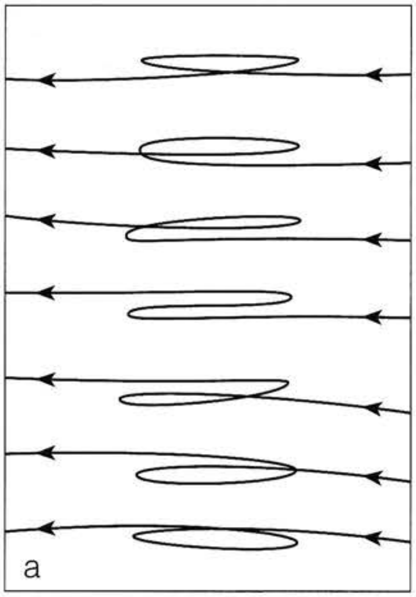


Figure 3
The trajectory of the Moon is a great circle on the celestial sphere tilted at 5° to the plane of the ecliptic. The axis OM *perpendicular to the plane of the Moon's orbit rotates about the axis* OE *perpendicular to the plane of the ecliptic, describing a cone with a period of 18.6 years.*

apparent motions, which are considerably different. (As a matter of fact, this difference is explained by the different positions of the two kinds of planets with respect to the Sun and the Earth: the inferior planets are those nearest the Sun, while the superior planets are further from the Sun than the Earth is.)

Figures 4b and 4c show the two kinds of planetary trajectories plotted against the background of the fixed stars. As with the Sun and Moon, it's important to keep in mind that all the planets, together with the fixed stars, participate in the diurnal rotation. So, when we talk about the planets' motion along the celestial sphere, we in effect subtract this rotation from the motions actually observed.

The inferior planets, Mercury and Venus, don't deviate far from the Sun. The angular distance between a planet and the Sun is called the planet's *elongation*—eastern or western, depending on whether the planet is located to the east or to the west of the Sun. The maximum elongation for Mercury is 28°; for Venus—47°. When an inferior planet's eastern elongation is greatest, it can be observed soon after sunset in the western sky, and it sets a little after the Sun does. Day by day it moves, at first slowly, then more quickly, westward—that is, against the Sun's motion. This type of planetary motion is called *retrograde*. As the days pass, it gradually approaches the Sun, hides in its rays, and can no longer be seen. At this moment the *inferior conjunction* of the planet and the Sun occurs. Some time after the inferior conjunction the planet becomes visible again—this time in the east, shortly before sunrise. Meanwhile the planet continues its retrograde motion, gradually getting farther away from the Sun. After its retrogression slows and it reaches its greatest western elongation, the planet stops and switches to *direct* motion (eastward). At first it moves slowly, then its motion gradually gets faster. Its distance from the Sun decreases, and finally it hides in the morning rays of the Sun—that is, its *superior con-*



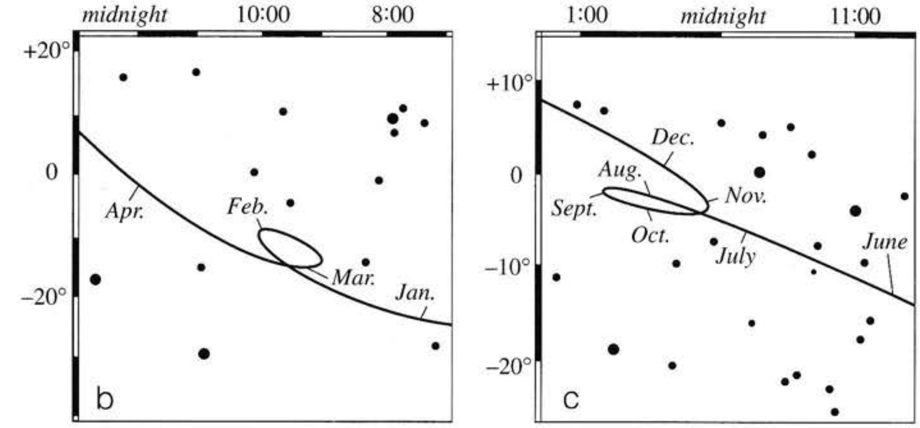


Figure 4
Different kinds of loops in the apparent paths of the planets are shown in figure 4a; portions of the trajectories of Mercury and Mars in 1988 are depicted in figures 4b and 4c.

junction with the Sun occurs. Some time later, it can be seen once again in the evening twilight. As it continues to move directly, the planet loses speed. After reaching its maximum distance to the east (that is, its greatest eastern elongation), the planet stops, changes direction (against the Sun's motion), and the process starts all over again. The period of one such "oscillation" is 88 days for Mercury and 225 days for Venus.

The apparent motion of the superior planets is rather different. When a superior planet is seen after sunset in the western sky, its motion among the stars is direct—that is, from west to east, just like the Sun. But it moves more slowly than the Sun, so the Sun catches up and the planet is lost to our sight for some time, since it rises and sets at almost the same time as the Sun. After the Sun has passed the planet, the planet can be seen in the east before sunrise. The speed of its direct motion gradually decreases, the planet stops, and then it starts its retrograde motion among the stars from east to west. Some time later the planet stops again and switches to direct motion; the sun overtakes it again from the west, and the planet once again ceases to be seen. These phenomena are repeated, over and over, in exactly the same order.

Armed with our modern understanding of the solar system, one could explain all the peculiarities of the apparent motion of the planets without too much difficulty. (Try to do it!) Imagine, though, the feat of these ancient astronomers. Their work was based on wrong, misleading ideas, yet they not only contrived to explain their observations coherently, they managed to predict planetary motion with remarkable precision (given their starting point).

First models of the universe: Eudoxus

The first model of the universe was extremely simple. Extended observations convinced the ancient Greeks that the other planets were spherical like the Earth. In addition, over time so much data was accumulated about the two seemingly largest heavenly bodies—the Sun and the Moon—that they began to be considered "relatives" of the Earth. There was no reason to consider the other wandering stars dissimilar to the Sun and Moon. So, it was thought, all of them are more or less like the Earth, and the differences in their apparent sizes can be explained by their different distances from the Earth.

But these huge bodies, eternally hurtling above our heads as they circle the Earth, must be attached pretty firmly to something. The celestial sphere didn't seem suited for this purpose, since the planets move independently of it. This is why the Greeks imagined seven new spheres—one for each planet—concentric with the sphere of fixed stars but smaller. All seven planetary spheres take part in the motion of the celestial sphere, which makes one rotation a day, and they also go through their own independent motions.

This model later gave rise to the idea of the "music of the spheres." The Greeks related each of the seven planetary spheres to one of the tones of the octave, and the sphere of fixed stars represented the eighth tone. The Greek philosophers thought that the huge hollow spheres to which bodies as large as the Sun and Moon (and all the other planets) are attached should give out sounds as they rotate, just as the spinning wheels of a mechanical device do. The different tones produced in this way combine to create a splendid melody whose powerful sounds fill the whole universe. And only we—imperfect Earth-dwellers—are unable to hear the sounds of this heavenly music, the eternal delight of the Olympian gods.

As more knowledge about the motion of heavenly bodies was acquired, the idea of spheres constituting the order in the universe had to be elaborated further. An unshakable underpinning of the ancient Greek worldview was the conviction that the Earth is the center of everything, the principal body in the universe. So the only way to explain all the numerous complications discovered by ancient astronomers—the irregular motions of the planets, their retrogressions, and so on—was to insert new spheres that would influence one and the same heavenly body by their combined motion. Eudoxus (408?–355? B.C.) constructed a new system of the universe, consisting of 27 spheres: three spheres for the Sun, three for the Moon, four spheres for each of the remaining five planets, and one large sphere for the fixed stars. It soon became clear, however, that 27 spheres are insufficient to describe the apparent motion of the planets. So Callipus soon added another 22 spheres to those proposed by Eudoxus. As time passed, the heavenly machine became more and more complicated. Eventually the need for a simpler, clearer description made itself acutely felt.

The eccentricities of Hipparchus

All the ideas of the ancients about the universe were based on the principle of uniform circular motion. This principle was first shaken by the Alexandrian astronomer and mathematician Hipparchus (2nd century B.C.), who discovered that the seasons of the year vary in duration. Hipparchus was the first to find the Sun's perigee and apogee and to establish that the Sun moves faster near the perigee than near the apogee. But the axiom of uniform motion was too deeply embedded in the flesh and blood of ancient science for Hipparchus to venture to destroy it.

To explain his discoveries, Hipparchus resorted to another method. He assumed that the Sun moves uniformly in a circle whose center does not coincide with the Earth's center but lies somewhere in free space outside it. Then, indeed, the Sun would seem to move irregularly—faster in the portion of the circle nearer the Earth and slower in the opposite portion. Figure 5 illustrates this mechanism: the center of the Sun's motion is at the intersection of the broken lines, while the solid lines coincide at the Earth's center. By trial-and-error we can find the location of the point at which an observer would see the aforementioned peculiarities of the Sun's motion, even though in reality it's still uniform circular motion. Hipparchus called the line connecting the perigee and apogee the *line of apsides* (pronounced AP-sih-deez). He called the ratio of the distance between the center of the solar orbit and the Earth's center, measured along this line to the orbit's radius, the *eccentricity* of the orbit. These terms have been retained in astronomical discourse down to the present.

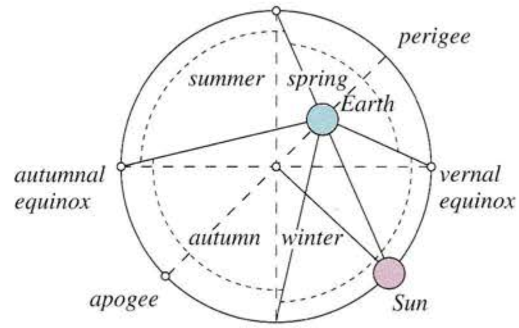


Figure 5

As with the Sun, Hipparchus placed the center of the Moon's orbit outside the Earth. He calculated the eccentricity, perigee, and apogee of its orbit and the direction of the line of apsides. Hipparchus determined the motions of the Sun and Moon with an accuracy surprising for that time. For instance, with the data Hipparchus obtained one could calculate to within a day the dates of the full moon for the present time—some 2,000 years later. The great Alexandrian astronomer also began to study the motions of other planets, which present much greater difficulties. But it was Claudius Ptolemaeus, or Ptolemy (2nd century A.D.), following in Hipparchus's footsteps, who succeeded in making significant progress in this direction.

The Ptolemaic system

Ptolemy's system of the universe, which reigned for 1,500 years without being doubted by anybody, was based on Hipparchus's observations and calculations. Ptolemy set forth his system in *He mathematike syntaxis* ("the mathematical collection"), which eventually became known as *Ho megas astronomos* ("the great astronomer"). In the ninth century, Arab astronomers used the Greek superlative *Megiste* to refer to the book. When the Arabic definite article *al* was added, the title became *Almagest*, the name still used today. Up to the end of the Middle Ages this work was honored almost on a level with divine revelation. It was considered a crime to doubt the words of the *Almagest*.

Four postulates underlie the Ptolemaic system:

1. The Earth is at the center of the universe.

2. The Earth is at rest.

3. All the heavenly bodies move around the Earth.

4. The motion of heavenly bodies proceeds in circles with constant velocities—that is, uniformly.

Ptolemy based his system on Hipparchus's eccentric circles. However, according to Ptolemy, none of the luminaries (except the Sun) revolves directly along such a circle; this is the path of the center of another circle along which the planet moves (fig. 6). This "other circle"—the planet's orbit—is

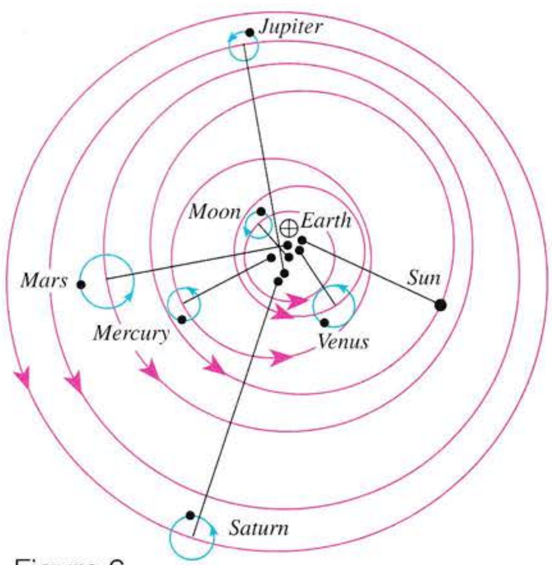


Figure 6
The Ptolemaic system.

called the *epicycle*; the circle traced by the center of an epicycle is called the *deferent*. The Sun's deferent and the deferents and epicycles of the other planets lie inside the sphere of fixed stars.

The diurnal motion of all the heavenly bodies was explained by the rotation of the universe as a whole about the stationary Earth. The direct and retrograde motions of planets were explained as follows.

Let a planet at some moment be at point P_1 of its epicycle (fig. 7), and let the center of the epicycle be at point N_1 of the deferent. The planet revolves about point N_1 with angular velocity α, and point N_1, as the center of the epicycle, revolves about the Earth with angular velocity ω. In the course of the uniform circular motion of both P_1 and N_1, the planet describes a loop that will be seen by an observer in a projection onto the celestial sphere. Why the loop forms is obvious: at point P_1 the motions along the epicycle and the deferent are directed similarly—from right to left. After describing the 180° arc, the planet moves along its epicycle from left to right. If the angular velocity α is greater than ω, the direction of the apparent motion near P_2 changes—the planet's motion is retrograde.

For each planet Ptolemy worked out the relative sizes of the radii of its epicycle and deferent, the position of the deferent's center, and the speeds of the planet's motion along the epicycle and deferent so that the resulting motion, when observed from the Earth, would match the actual apparent motion. This turned out to be possible under certain conditions, which Ptolemy took as postulates. These postulates are as follows:

1. The centers of the epicycles of the inferior planets lie on the line directed from the Earth to the Sun.

2. The radii of the epicycles of all the superior planets, drawn to their positions, are parallel to the same direction.

So the direction to the Sun turned out to be preeminent in the Ptolemaic system.

The Ptolemaic system not only qualitatively explained the apparent motions of the planets, it also made it possible to calculate their positions in the future with a rather high accuracy. Discrepancies between theory and observation that arose as the precision of observations improved were eliminated at the expense of making the system more complex. For example, certain irregularities in the apparent motions of the planets were attributed to the existence of secondary epicycles: the original epicycle of a given planet was thus considered the path of the center of a second epicycle, along which the planet actually moves. When even such a construct was insufficient for a planet, a third epicycle was introduced, and a fourth, and so on, until the position of the planet in the last epicycle produced more or less acceptable agreement with observations. By the beginning of the 16th century the Ptolemaic system comprised 40 circles in all.

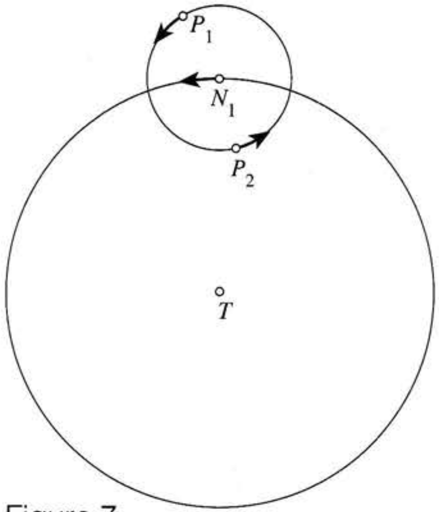


Figure 7

Let's return to the *Almagest* and look at a table (below) with the numbers Ptolemy gave for planetary motions along their epicycles and the motions of the epicycles themselves along their deferents.

A surprising fact leaps out at us: for the inferior planets (Mercury and Venus) the motion of the epicycle's center proceeds at the same rate as the motion of the Sun around the Earth. For the superior planets (Mars, Jupiter and Saturn) these numbers differ, but the sums of both motions give the very same value—that of the Sun's motion. Consequently, the motion of the Sun is contained in all the planetary motions. Such a phenomenon naturally seemed very strange. The obvious question arises: isn't there a common cause for all these equal values?

Doubtless many ancient and medieval thinkers posed just such a question. For instance, the ancient astronomer Aristarchus proposed that the Sun occupies the central position in the universe. However, the first person who dared to develop, in a mathematically rigorous way, the idea that all the planets revolve around the Sun was the Polish astronomical genius Nicolaus Copernicus (1473–1543).

Planets	**Diurnal motion along epicycles**	**Diurnal motion of epicycle's center along deferent**	**Sum**
Sun	0°00′00.0″	0°59′08.3″	0°59′08.3″
Mercury	3°06′24.1″	0°59′08.3″	4°05′32.4″
Venus	0°36′59.4″	0°59′08.3″	1°36′07.7″
Mars	0°27′41.7″	0°31′26.6″	0°59′08.3″
Jupiter	0°54′09.0″	0°04′59.2″	0°59′08.3″
Saturn	0°57′07.7″	0°02′00.6″	0°59′08.3″

The Copernican system

Copernicus's lifework, *On Revolutions of Celestial Spheres*, was published in 1543, not long before his death. In it Copernicus elaborated his understanding of the Earth's motion and laid the foundation for a new astronomy. The system of the universe he constructed is called heliocentric and is based on the following statements:

1. The Sun rather than the Earth is at the center of the universe.

2. The spherical Earth rotates about its axis, and this rotation accounts for the seeming daily motion of the heavenly bodies.

3. The Earth and all the other planets revolve about the Sun, and this revolution accounts for the apparent motion of the Sun among the stars.

4. All the motions are represented as combinations of uniform circular motions.

5. The seeming direct and retrograde planetary motions belong not to the planets but to the Earth.

In addition, Copernicus believed that the Moon revolves around the Earth and that they both revolve around the Sun.

The postulate of uniform circular motion forced Copernicus, like Ptolemy, to resort to epicycles and to shift the centers of the deferent circles with respect to the Sun's center. As a result, the Copernican model was no simpler than Ptolemy's old model—suffice it to say that it contained 48 circles instead of the 40 circles in the geocentric system. Nor was it any more accurate. But it contained what the geocentric system lacked—the grain of scientific truth that grew into the tree of a new astronomy.

By a twist of fate, the task of confirming Copernicus's conclusion fell to the Danish scientist Tycho Brahe (1546–1601), one of the greatest astronomers of all time, who had very solid grounds for not accepting the heliocentric system. His main argument against Copernicus amounted to this: if the Earth were revolving around the Sun, then Venus and Mercury would have phases like the Moon's, which no serious astronomer had ever observed. These arguments sounded convincing, and although the predicted phases actually do exist (as we know now), the lack of optical instruments prevented their detection. Nonetheless, it was the precise observations of Tycho Brahe that eventually justified Copernicus's point of view. The data gathered by Brahe allowed his student Johannes Kepler to announce, after eight years of work, that each planetary orbit is an ellipse with the Sun at one focus, and that the line joining the Sun and a planet sweeps out equal areas in equal times. And this was how the Pythagorean harmony of perfect circular orbits centered at the God-given special location of our planet fell by the wayside. In turn, Kepler's laws (much more than the falling apple of lore) formed the foundation of Newton's law of gravity, which for almost three centuries has been the basis of physics and cosmology.

HOW DO YOU FIGURE?

Challenges in physics and math

Math

M51

Penultimately even. Prove that for any integer $n > 2$, the next-to-last digit of 3^n is even. (V. Plachko)

M52

Factors, sums, equal powers. The positive integers a, b, c, and d satisfy the equality $ab = cd$. Prove that the number $a^{1992} + b^{1992} + c^{1992} + d^{1992}$ is composite. (A. Anjans)

M53

Two circles inside a third. Inside a circle there are two intersecting circles. One of them touches the big circle in point A, the other in point B. Prove that if segment AB meets the smaller circles at one of their common points (fig. 1), then the sum of their radii equals the radius of the big circle. Is the converse true? (A. Vesyolov)

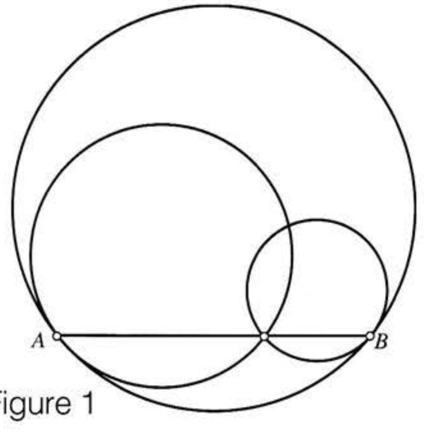


Figure 1

M54

Skinny rectangles. A continuous and monotonic function is defined on the closed interval [0, 1] and takes values in the same interval. Prove that, for any n, its graph can be covered by n rectangles of area $1/n^2$ whose sides are parallel to the coordinate axes. (A. Anjans)

M55

Out of sight. All planets of some planetary system are spheres of unit radius. Mark on each of the planets the set of points that are invisible from any point on any of the other planets. Prove that the sum of the areas of all these sets is equal to the surface area of one planet.

Physics

P51

Bouncing dumbbell. A dumbbell consisting of a weightless bar with identical small balls at both ends stands vertically on a smooth level table. A horizontal velocity v is imparted to the upper ball by hitting it. What is the minimum bar length for which the lower ball of the dumbbell loses contact with the table immediately after the upper ball is struck? (A. Zilberman)

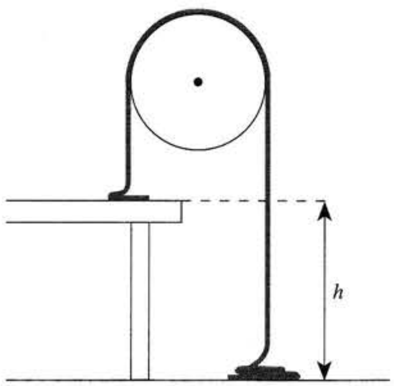


Figure 2

P52

Cable on a table. A cable is thrown over a cylinder so that some of the cable is on the table and some is on the floor. After the cable is released it begins to move without friction. Find the speed of the cable after uniform motion has been established. The table height is equal to h (fig. 2) (A. Zilberman)

P53

Vapor over salted water. It is well known that if ordinary water is salted, its boiling point rises. Will the density of saturated water vapor be higher or lower at the boiling point? (A. Buzdin, S. Krotov)

P54

Thermal oscillations. A set of experiments was performed to examine the properties of a nonlinear resistor. First, the dependence of its resistance on temperature was studied. It was found that as the temperature increased, the resistance instantaneously jumped from $R_1 = 50\ \Omega$ to $R_2 = 100\ \Omega$ at $t_1 = 100°C$; as the resistor was cooled, a corresponding drop was found to occur at $t_2 = 99°C$ (see figure 3). In another experiment a constant voltage $V_1 = 60$ V was applied to the resistor, and its temperature turned

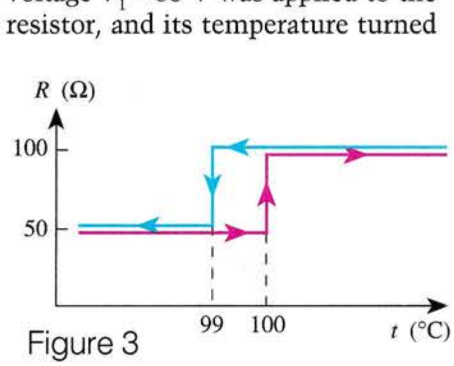


Figure 3

CONTINUED ON PAGE 23

IN THE LAB I

Neutrons seek the murderer!

Records of neutron activation analysis

by A. S. Shteinberg

THE TITLE OF THIS ARTICLE was prompted by a detective story—if you're patient, you'll read all about it before too long. But for now we'll deal with things rather distinct from criminology and look at how the chemical structure of a substance can be determined—that is, how one can recognize what elements are present in a sample and in what proportions.

You know from studying chemistry that this can be done by means of chemical reactions that are characteristic of certain substances and elements—methods of so-called "wet chemistry." In some cases, though, it's difficult to apply these techniques—for example, when one has to chemically analyze a very small sample, or when a high degree of precision is needed in simultaneously analyzing several elements, or if one has to detect one "alien" atom among a million or 100 million atoms and determine what kind it is.

Such situations are by no means speculative. In particular, they arise in modern semiconductor technology. More and more often it's said that after the stone, bronze, and iron ages, the silicon age has arrived. Today's electronics would be impossible without miniature integrated circuits, in which silicon plays an essential role. But the properties of these circuits are acutely dependent on incredibly small amounts of various admixtures. Without the ability to determine very precisely the composition of small silicon samples, progress in developing new generations of computer would be impossible.

So how is it done?

It has been known for about 100 years that physical methods often turn out to be the most effective in chemical analysis. And among these, a true champion in sensitivity is neutron activation analysis (NAA). Before I describe how it works, let's recall how the atomic nucleus is structured.

The nucleus of an atom consists of positively charged protons and electrically neutral neutrons. The number of protons determines the atom's nuclear charge and is its basic characteristic. One element differs from another mainly because of the number of protons in its nucleus. Hydrogen (H), the first element in Mendeleyev's table, has one proton in its nucleus; helium (He), the second, has two; and so on. Well, how many neutrons does the nucleus contain? It turns out that different numbers of neutrons can be contained in nuclei of the same element. Nuclei with the same number of protons and different numbers of neutrons are called isotopes.

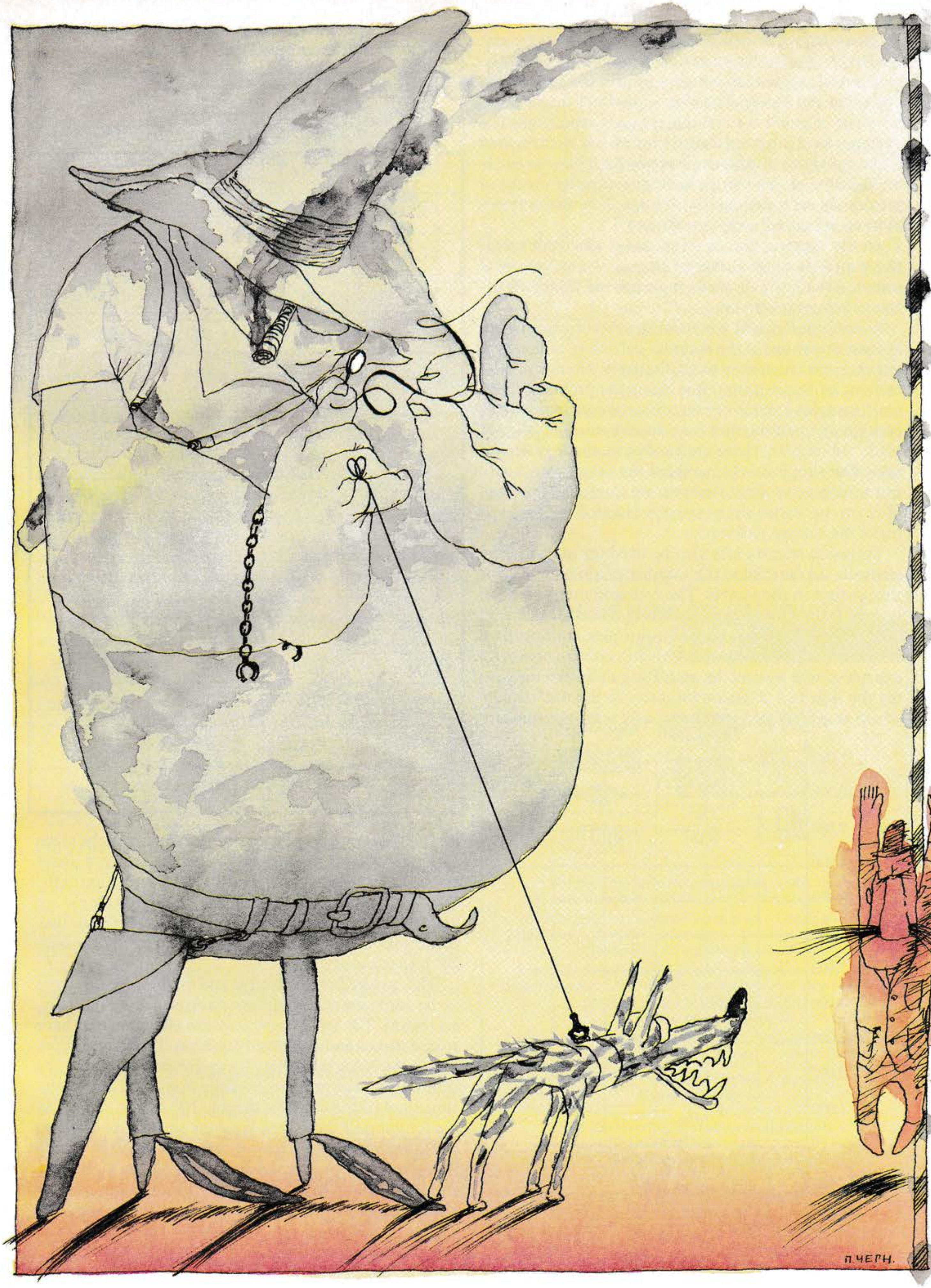
Let me give you an example. There are always two protons in the helium nucleus. The number of neutrons can be 1, 2, 4, or 6; that is, there are four isotopes of helium: ^{3}He, ^{4}He, ^{6}He, and ^{8}He (the superscripts show the total number of protons and neutrons). Tin has the record number of isotopes at 10, xenon has 9, both cadmium and tellurium have 8, and so forth. Under natural conditions, different isotopes aren't distributed uniformly. As a rule, one single isotope has the lion's share (90% or more) of all natural deposits of an element.

The chemical properties of an element's isotopes are identical. As for their physical properties, the differences can be very noticeable. For instance, the atomic weights of various isotopes are different because of the "extra" neutrons. But for us, this isn't the main issue. A great many isotopes have an amazing property—radioactivity. They are a source of radiation no one can see or feel, but it can be detected with special instruments. The simplest of these detectors are the photographic plate and the Geiger counter. I won't get into the question of the nature of radiation. The topic is complex and lies outside the framework of this article. But to keep things clear, you can imagine radiation as a flow of invisible particles.

The emission of any radioactive isotope is specific to the highest degree. That is its individual portrait—as unique as a human fingerprint. The most characteristic features of this portrait are the emission energy (for us, the energy of the particles) and the half-life.

It's rather easy to explain what an element's half-life is. A detector can not only register the emission but also measure its intensity—that is, determine how many particles have fallen on a unit of area during a unit of time. The greater the intensity, the more often the Geiger counter clicks. The time it takes for an isotope's emission intensity to decrease by a half is called its half-life (denoted by $T_{1/2}$). During this time, half of the radioactive nuclei have decayed. Half-lives differ from one isotope to

Art by Pavel Chernusky


П.ЧЕРН.

another—from fractions of a second to millions of years. There are special detectors that not only measure the intensity of an emission but also construct its spectrum—a graph of the dependence of an emission's intensity on its energy. Figure 1 is an example of such a spectrum. The sharp peaks of intensity (called spectral lines) correspond to the emissions of different isotopes. Each line can easily be linked with the corresponding isotope by means of previously compiled tables (this was how the isotopes marked in figure 1 were "recognized").

So the construction of an emission spectrum solves the problem of qualitative chemical analysis of a sample—that is, it allows us to determine what radioactive elements it contains.

Everything I've told you so far should convince you of at least one thing: if the samples consist of radioactive isotopes, it's rather easy to qualitatively determine their content by their spectra. But here's the problem: "normal" samples consist of nonradioactive isotopes. So to perform an analysis, they have to be activated—that is, made radioactive. There are various methods of activation. The simplest is to bombard the sample with neutrons in a reactor. Since neutrons are electrically neutral, they can penetrate the positively charged nucleus and make the isotope radioactive.

Emission spectra can also be used for quantitative analysis—to determine the number of atoms of this or that isotope in the sample. The emission intensity of an isotope is directly proportional to the number of nuclei in it. Using standard samples (samples containing a known amount of the given element), one can construct a graph of this relation by activating them. By measuring the isotope's emission intensity in the test sample, which was activated simultaneously with the standard

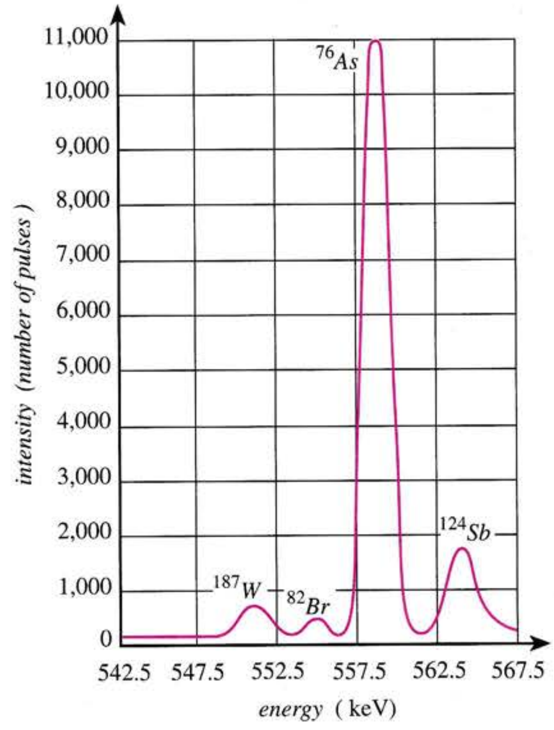


Figure 1

Neutron activation analysis of silicon impurities

Basic natural isotope	Detected isotope	Sensitivity (atoms/cm³)	Half-life
^{23}Na	^{24}Na	1×10^{12}	15 hours
^{39}K	^{42}K	1×10^{13}	12.4 hours
^{52}Cr	^{51}Cr	5×10^{12}	27.7 days
^{56}Fe	^{59}Fe	1×10^{14}	44.6 days
^{59}Co	^{60}Co	2×10^{11}	5.27 years
^{63}Cu and ^{65}Cu	^{64}Cu	2×10^{12}	12.7 hours
^{64}Zn	^{65}Zn	1×10^{13}	244 days
^{75}As	^{76}As	1×10^{10}	26.3 hours
^{80}Br	^{82}Br	5×10^{10}	35.3 hours
^{98}Mo	^{99}Mo	1×10^{11}	66.2 hours
^{108}Ag	^{110}Ag	2×10^{11}	252 days
^{122}Sb	^{124}Sb	5×10^{10}	60.4 days
^{181}Ta	^{182}Ta	1×10^{10}	115 days
^{185}W	^{187}W	2×10^{11}	23.9 hours
^{197}Au	^{198}Au	2×10^{9}	2.7 days
^{202}Hg	^{203}Hg	1×10^{11}	47 days

sample, one can easily determine the amount of the given isotope from the graph. (After what I've told you about the half-life, you should understand why activation of the test sample and activation of the standard sample are performed simultaneously. If this condition isn't fulfilled, one would have to introduce a correction to take the half-life into account.)

This method of qualitative and quantitative analysis is the one I mentioned earlier—neutron activation analysis (NAA). The table above lists the elements that have to be detected and quantified in silicon. The third column shows the NAA sensitivity for each element (that is, the minimum concentration for detection). One cubic centimeter of silicon contains roughly 10^{23} atoms, so you can see for yourself the record-breaking sensitivity of NAA.

Figure 2 gives NAA data on impurity levels in silicon from three American firms that supply silicon to the electronics industry. Relatively high concentrations of some elements—iron and chromium in particular—require

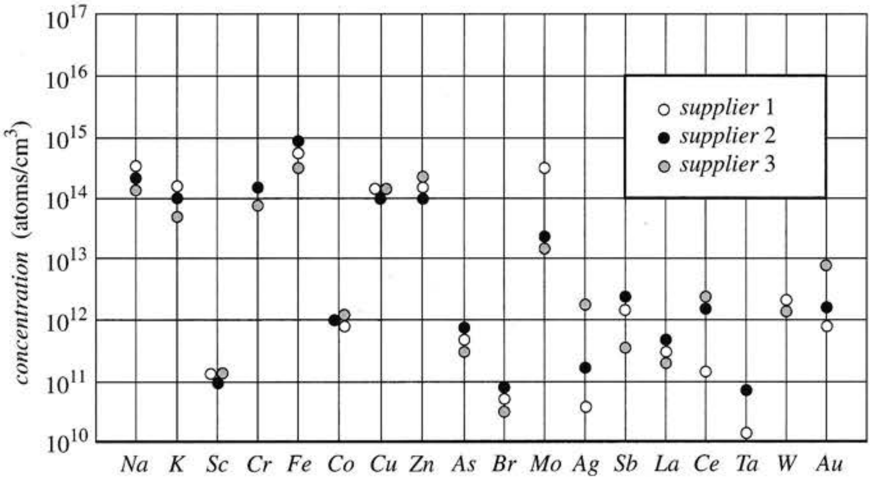


Figure 2

that the silicon be purified further before it can be used in integrated circuits. Otherwise, the computer won't be worthy of your trust.

And finally—the promised detective story.

On May 13, 1958, in the Canadian city of Edmundston, near the Canadian–American border, the corpse of a 16-year-old girl, Gaetane Bouchard, was found. An American by the name of John Follman, who traveled frequently to Edmundston on business, was suspected on the basis of circumstantial evidence. Follman categorically denied having anything to do with the crime. The investigation was badly in need of physical evidence.

Upon careful examination of Bouchard's corpse, a single hair found in the girl's tightly clenched fist. It could belong only to her or to the murderer.

Is it possible to unmistakably identify a person by a single hair? The police put this question to NAA specialist Robert Jervie. The answer required some special research. The basic idea was that every person's hair has its own characteristic set of microelements in unique concentrations. To prove it, Jervie had to investigate the hair of hundreds of people. The concentrations of the characteristic microelements—arsenic, sodium, copper, zinc, bromine—in human hair is less than one atom in a hundred million. No other method but NAA could solve the problem of identifying a man by his hair.

As a result of painstaking work, it was proved that the hair in the victim's hand belonged to Follman, and this became the decisive evidence of his guilt.

"CHALLENGES IN PHYSICS AND MATH" CONTINUED FROM PAGE 19

out to be t_3 = 80°C. Finally, when voltage V_2 = 80 V was applied to the resistor, spontaneous current oscillations were detected in the circuit. Determine the period of these oscillations as well as the maximum value of the current. The air temperature in the laboratory is constant at t_0 = 20°C. The rate of heat loss from the resistor is proportional to the difference between its temperature and the temperature of the surrounding air. The resistor's heat capacity is C = 3 J/K. (A. Buzdin)

P55

Say "cheese"? It's dangerous to photograph a tiger at a distance of less than 20 m. If you were using a classic camera obscura with an aperture that is 1 mm in diameter, how large must the camera be so that the tiger comes out striped in the photograph? The distance between the stripes on the tiger is equal to 20 cm. (A. Stasenko)

ANSWERS, HINTS & SOLUTIONS ON PAGE 56


Д. Крылов

Go "mod" with your equations

The modus operandi of moduli

by Andrey Yegorov

ON MONDAY, September 2, 1991, I was making some long-range plans, and I had to figure out what day of week the 20th of December, 1992, will be. I had no calendar at hand and was forced to get busy computing. I knew that the interval from September 2, 1991, to December 20, 1992, is $28 + 31 + 30 + 20 + 366 = 475$ days, or 67 full weeks and 6 days more ($475 = 67 \cdot 7 + 6$). So I concluded that December 20, 1992, will be a Sunday.

Squaring a multidigit number, a student obtained 46,991,075. The teacher, glancing into the student's notebook, said, "Your answer's wrong!" How did the teacher know?

Exercise 1. Think about whether the square of an integer can end in 75.

The solutions to both of these unsophisticated problems, as well as many others, are based on considerations of divisibility. Not long ago we touched on this subject (see "Divisive Devices" in the September/October 1991 issue of *Quantum*). This time we'll examine it from another angle. But the starting point is the same: we must begin with a short reminder about division with and without remainders.

Division with a remainder

If $a = qb + r$, where a, b, q, r are integers, and $0 \le r < |b|$, we say that a has been divided by b with a remainder of r. For instance, dividing 5 by 7 we get $5 = 0 \cdot 7 + 5$; dividing -224 by -9 we get $-224 = 15 \cdot (-9) + 1$; and so forth.

If the remainder is zero—$a = qb$—we say that *a is divisible by b*. It's obvious but important that *for any integers k and l the number ka + lb is divisible by c whenever a and b are divisible by c*. Here's how this fact can be used.

Problem 1. For certain values of n, the numbers $7n + 1$ and $8n + 3$ have a common divisor d, $d \ne 1$. Find d.

Solution. Since $7(8n + 3) - 8(7n + 1) = 13$ is divisible by $d \ne 1$, and 13 is a prime number, $d = 13$.

As a matter of course, n and d here were assumed to be integers. I'll make this assumption throughout the article: all the numbers considered will be integers, although at times it will be convenient to mention it again.

Exercise 2. Divide with a remainder: (a) 1931 by 17, (b) -295 by 31, (c) $-1{,}005$ by -98.

Exercise 3. The number $17x + 3y$ is divisible by 61. Prove that the same is true for $8x + 5y$.

Exercise 4. Find the remainders in dividing (a) n by $n - 1$ and by $n - 2$; (b) $n^2 + n + 1$ by $n + 1$ and $n + 2$; (c) $n^4 + 1$ by $n + 3$ ($n \ge 80$).

Exercise 5. Find all integers n such that the numbers (a) $(n^2 + 1)(n - 1)$, (b) $(n^5 + 3)(n^2 + 1)$ are integers.

Congruences

Consider the following problem.

Problem 2. What is the last digit of 2^{999}?

Solution. Write out consecutive powers of two:

$$2, 4, 8, 16, 32, 64, \ldots.$$

We see that the last digits of these numbers are repeated after every four numbers, so the last digit of 2^n depends only on the remainder of the exponent n when divided by 4. Since $999 = 4 \cdot 249 + 3$, the answer to this problem is $2^3 = 8$ (that is, the third number in the string above).

In this example the set of all exponents was partitioned into four classes consisting of numbers of the form

$$4k,\ 4k + 1,\ 4k + 2,\ 4k + 3.$$

Generally, for any positive integer m all integer numbers (not necessarily positive!) fall into m classes: each class comprises all the numbers having the same remainder when divided by m.

Here are these classes:

(0) The numbers a of the form $a = km$,

(1) The numbers a of the form $a = km + 1$,

.
.
.

$(m - 1)$ The numbers a of the form $a = km + (m - 1)$.

Clearly, any number belongs to one of these classes. The difference

Art by Dmitry Krymov

between two arbitrary numbers from *one and the same* class is divisible by m, while the difference between numbers from different classes is not divisible by m.

DEFINITION. *If the difference between integers a and b is divisible by m, then the numbers a and b are said to be congruent modulo m.*

In Latin, *modulo* is the ablative case of *modulus*; in mathematics, it means "with respect to a modulus of . . ." This is the notation for congruence modulo m:

$$a \equiv b \pmod{m}.$$

The numbers a and b are congruent modulo m if and only if they belong to the same class—that is, have the same remainders when divided by m. In other words, $a \equiv b \pmod{m}$ *means that* $a = b + km$ *for some integer* k.

For instance, $27 \equiv 7 \pmod{10}$, $78 \equiv 6 \pmod{24}$, $6 \equiv 0 \pmod{3}$, $25 \equiv -4 \pmod{29}$.

Exercise 6. Prove that (a) $a^3 \equiv a \pmod{6}$, (b) $a^5 \equiv a \pmod{5}$ for all natural numbers a.

Exercise 7. Prove that $2^{100} \equiv 3^{100}$ modulo 5, 13, 211.

Exercise 8. Prove that $11^{10} - 1$ is divisible by 100.

Exercise 9. Let $S(N)$ be the sum of the digits in the number N. Prove that $N \equiv S(N)$ modulo 3 and 9.

Exercise 10. Let $S(A) = S(5A)$. Prove that $A \equiv 0 \pmod{9}$.

Exercise 11. The decimal notation of a certain number consists of (a) 1991 or (b) 1992 ones and some number of zeros. Can this number be the square of an integer?

Exercise 12. Prove the following test for divisibility by 11: $a = \overline{a_n a_{n-1} \ldots a_0} \equiv 0 \pmod{11}$ if and only if $(-1)^n a_n + (-1)^{n-1} a_{n-1} + \ldots + a_0$ is divisible by 11. (The bar over the expression means that the symbols below stand for decimal digits, which stand next to each other to indicate place value as if they were numerals.)

Properties of congruences

Many of the properties of congruences are quite similar to those of regular equations:

1. If $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$.

Further, if $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then

2. $a + c \equiv b + d \pmod{m}$,
3. $a - c \equiv b - d \pmod{m}$,
4. $ac \equiv bd \pmod{m}$.

That is, congruences can be added, subtracted, and multiplied just like regular equations.

For example, let's prove property 4. Since $a \equiv b$, $c \equiv d$, the differences $a - b$ and $c - d$ are divisible by m. From $ac - bd = a(c - d) + d(a - b)$, it follows that $ac - bd$ is divisible by m, or

$$ac \equiv bd \pmod{m}.$$

Exercise 13. Prove properties 1, 2, and 3.

Let $a \equiv b \pmod{m}$. The above properties imply

5. $a^k \equiv b^k \pmod{m}$.

Sometimes common factors of both parts of a congruence can be canceled out:

6. If $ac \equiv bc \pmod{m}$, and the numbers c and m are coprime,[1] then $a \equiv b \pmod{m}$;

7. If $a \equiv b \pmod{m}$ and $a = ka_1$, $b = kb_1$, $m = km_1$ for some positive integer k, then $a_1 \equiv b_1 \pmod{m_1}$.

In other words, both parts of a congruence and its modulus can be reduced by their common factor.

Let's prove property 6. The number $c(a - b)$ is divisible by m; c and m are coprime. Therefore, $a - b$ is divisible by m: $a \equiv b \pmod{m}$.

Exercise 14. Prove property 7.

Summing up, we can say that *for any algebraic expression obtained from integers a, b, c by means of addition, subtraction, and multiplication, it is possible to replace these numbers with their remainders when they are divided by m without changing the remainder obtained when the entire expression is divided by m.*

Consider two applications of this statement.

Problem 3. Find the remainder when

$$N = (1^2 + 1)(2^2 + 1)(3^2 + 1) \ldots (1{,}000^2 + 1)$$

is divided by 3.

Solution. The above statement implies that, modulo 3,

$$\begin{aligned} N &\equiv (1^2 + 1)^{334} \cdot (2^2 + 1)^{333} \cdot (3^2 + 1)^{333} \\ &\equiv 2^{334} \cdot 2^{333} \cdot 1^{333} \\ &\equiv 2^{667} \equiv (2^2)^{333} \cdot 2 \\ &\equiv 1^{333} \cdot 2 \equiv 2. \end{aligned}$$

Problem 4. For what positive integers n is the number $8n + 3$ divisible by 13? (Compare this with problem 1.)

Solution. The above properties give the following chain of equivalent congruences:

$8n + 3 \equiv 0 \pmod{13}$,
$8n \equiv -3 \pmod{13}$,
$64n \equiv -24 \pmod{13}$ (we've multiplied by 8, which is coprime with 13),
$-n \equiv -11 \pmod{13}$ (since $64 \equiv -1 \pmod{13}$ and $-24 \equiv -11 \pmod{13}$),

and, finally,

$$n \equiv 11 \pmod{13}.$$

Indeed, if $n = 13k + 11$, then $8n + 3 = 8(13k + 11) + 3 = 104k + 91 = 13(8k + 11)$. So $8n + 3$ is divisible by 13 if and only if $n = 13k + 11$.

Exercise 15. Find the remainders of (a) $2^{1992} - 1$ modulo 17, (b) $(3^{20} + 11)^{55}$ modulo 13.

Exercise 16. Prove that (a) $2^{50} + 1$ is divisible by 125, (b) $2^{48} - 1$ is divisible by 105, (c) $2^{3^n} + 1$ is divisible by 3^{n+1} but is not divisible by 3^{n+2}.

Exercise 17. Find all prime numbers p such that $20p^2 + 1$ is also a prime.

[1]Two natural numbers are called "coprime" or "relatively prime" if they have no common factor other than 1.

Exercise 18. Prove that (a) $1^{1991} + 2^{1991} + \ldots + 30^{1991}$ is divisible by 31; (b) $1^m + 2^m + \ldots + (n-1)^m$ is divisible by n for any odd m and n.

Exercise 19. For what positive integers n is the number $20^n + 16^n - 3^n - 1$ divisible by 323?

Exercise 20. Prove that $5^{2n+1} + 3^{n+1} \cdot 2^{n-1}$ is divisible by 19 for any positive integer n.

Exercise 21. For what values of n can the fraction $(15n + 2)/(14n + 3)$ be simplified?

The Chinese Remainder Theorem

Consider m terms of an arithmetic sequence

$$a, a + d, \ldots, a + (m - 1)d,$$

where a and d are integers. The following theorem is often very helpful:

THEOREM 1. *If d is coprime with m, then the sequence a, $a + d$, ..., $a + (m - 1)d$ contains exactly one number divisible by m.*

PROOF. The difference between the kth and lth terms of the sequence is equal to $d(k - l)$ and is not divisible by m; otherwise m would divide $k - l$, which is impossible, since $|k - l| < m$. Consequently, no two numbers in the sequence are congruent with each other and, therefore, all these numbers have different remainders modulo m. So they represent *all* the classes modulo m—that is, each of the remainders $0, 1, 2, \ldots, m - 1$ is congruent with exactly one of the numbers in the sequence.

Thus, we've proved even a somewhat stronger statement than theorem 1.

Exercise 22. Find all triples of primes of the form p, $p + 2$, $p + 4$.

Exercise 23. Find a finite arithmetic sequence of maximal length consisting of prime numbers and having a difference of 6.

Exercise 24. Fifteen prime numbers make an arithmetic sequence with a difference of d. Prove that $d > 30{,}000$.

Now let's apply theorem 1 to prove the so-called *Chinese Remainder Theorem*. This theorem was already known more than 2,000 years ago in China.

THEOREM 2. *Given n numbers m_1, m_2, ..., m_n coprime with each other and n numbers $r_1, r_2, \ldots, r_n$ such that $0 \le r_i \le m_i - 1$ ($i = 1, 2, \ldots, n$). Then there exists a number N whose remainder when divided by m_i is r_i for all $i = 1, 2, \ldots, n$.*

In other words, $N \equiv r_i \pmod{m_i}$ for all $i = 1, 2, \ldots, n$.

PROOF. Let's use induction over n. For $n = 1$ the statement of the theorem is trivial. Suppose it's true for $n = k - 1$ and consider $n = k$. By the inductive assumption, there exists a number M such that

$$M \equiv r_i \pmod{m_i} \text{ for } i = 1, 2, \ldots, k - 1.$$

Let $d = m_1 m_2 \ldots m_{k-1}$. Consider the numbers

$$M, M + d, M + 2d, \ldots, M + (m_k - 1)d.$$

Since d is coprime with m_k, it follows from the proof of theorem 1 that one of these numbers, N, has the remainder r_k when divided by m_k. At the same time, $N \equiv M \pmod{m_i}$ for any $i = 1, \ldots, k - 1$, so the remainder of N divided by m_i, $i < k$, is r_i, completing the proof.

Just one more theorem.

THEOREM 3. *For any numbers m_1, m_2, ..., m_n coprime with each other and any remainders r_1, r_2, ..., r_n modulo m_1, m_2, ..., m_n, one can find n successive numbers a, $a + 1$, ..., $a + n - 1$ such that $a \equiv r_1 \pmod{m_1}$, $a + 1 \equiv r_2 \pmod{m_2}$, ..., $a + n - 1 \equiv r_n \pmod{m_n}$.*

In other words, for any set of pairwise coprime moduli m_1, m_2, ..., m_n, one can find n consecutive positive integers that would yield any desired remainders when divided by m_1, m_2, ..., m_n, respectively.

PROOF. By the Chinese Remainder Theorem there is a number a such that

$$\begin{aligned} a &\equiv r_1 \pmod{m_1}, \\ a &\equiv r_2 - 1 \pmod{m_2}, \\ &\vdots \\ a &\equiv r_n - n + 1 \pmod{m_n}. \end{aligned}$$

Then the numbers a, $a + 1$, ..., $a + n - 1$ satisfy the requirement of our theorem.

Exercise 25. Prove that (a) among any 10, (b) among any 16 consecutive positive integers there is a number that is coprime with all the other numbers. (c) Is this statement true for any 17 consecutive positive integers?

Exercise 26. Prove that for any n there are n consecutive positive integers each of which is divisible by the square of some integer other than 1.

Exercise 27. Is there a moment in the day when the hour, minute, and second hands of a correctly running watch make angles of 120° with one other?

Exercise 28. Find the least positive integer yielding the remainders 1, 2, 4, 6 when divided by 2, 3, 5, 7, respectively.

Exercise 29. Find the least positive even number a such that the numbers $a + 1$, $a + 2$, $a + 3$, $a + 4$, $a + 5$ are divisible by 3, 5, 7, 11, 13, respectively.

How to solve congruences

In problem 4 we found all integers n such that $8n + 3$ is divisible by 13. That is, we solved for n the congruence

$$8n + 3 \equiv 0 \pmod{13}.$$

Now we can generalize this problem.

Let a and m be coprime numbers. We want to solve for n the congruence

$$an \equiv b \pmod{m}$$

for arbitrary b.

By theorem 1 there exists some k such that $ak \equiv 1 \pmod{m}$. Multiplying the two sides of the given congruence by k, we get

$$(ak)n \equiv n \equiv bk \pmod{m},$$

which means

$$n = bk + ml,$$

where l is an arbitrary integer.

Of course, the question arises: How can we find k for a particular congruence?

If m is not too large, this can be done simply by trial-and-error. As to the general solution, you can find it, albeit in implicit form, in the article "Divisive Devices" mentioned above. There it was shown that for any two coprime numbers a and b there exist integers x and y such that $ax + yb = 1$; also, it was explained how the numbers x and y can be found by means of Euclid's algorithm. Switching to our notation, we replace b with m (coprime with a) and x with k. Then the equation $ak + ym = 1$ can be rewritten in the form of a congruence: $ak \equiv 1 \pmod{m}$—the very one we wanted to solve.

Problem 5. Solve the congruence

$$32n \equiv 7 \pmod{37}.$$

Solution. Since $32 \equiv -5 \pmod{37}$, we can rewrite the congruence in the form

$$5n \equiv -7 \equiv 30 \pmod{37},$$

or

$$n \equiv 6 \pmod{37}$$

(we've canceled out the factor 5 legitimately because 5 and 37 are coprime).

Linear equations with integer coefficients, also discussed in "Divisive Devices," can be solved by reducing them to congruences.

Problem 6. Find all pairs of integers x, y satisfying the equation $7x - 23y = 131$.

Solution. Since $23 \equiv 2 \pmod{7}$, $131 \equiv -2 \pmod{7}$, the given equation can be rewritten as $-2y \equiv -2 \pmod{7}$, or $y \equiv 1 \pmod{7}$.

Thus, $y = 1 + 7k$ for any integer k. Now we can easily find x:

$$7x - 23(1 + 7k) = 131,$$
$$7x = 154 + 23 \cdot 7k,$$

and finally,

$$x = 22 + 23k, \quad y = 1 + 7k.$$

Exercise 30. Solve the congruences (a) $17x \equiv 19 \pmod{37}$, (b) $147x \equiv 63 \pmod{29}$.

Exercise 31. Solve for integers x and y (a) $7x + 8y = 1$, (b) $13x - 15y = 16$, (c) $257x + 18y = 175$.

Exercise 32. Solve for integers x, y, and z the system of equations

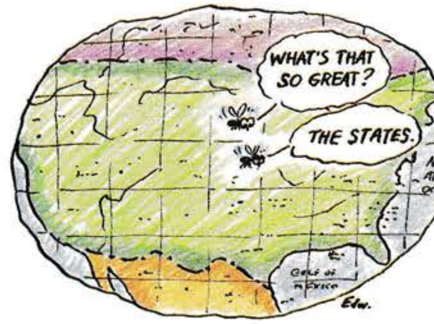
$$\begin{cases} 3x + 5y - 7z = 1, \\ 4x + 9y + 11z = 2. \end{cases}$$

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

BRAINTEASERS

Just for the fun of it!

B51



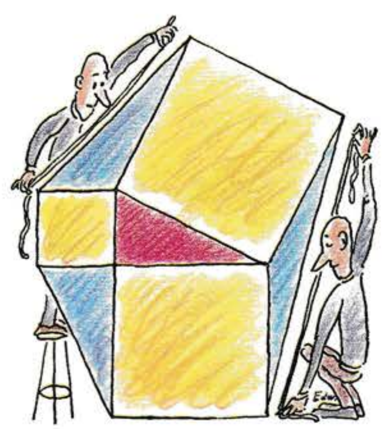
Crush on a map. The population of the United States is more than 200 million. It would seem that on a map with a scale of 1 : 5,000,000 (1 centimeter to 50 kilometers) there should be enough room for 1/5,000,000 as many people—that is, for more than 40. But you can check experimentally that five persons would have some difficulty squeezing onto such a map, and ten would find it impossible. Why? (G. Galperin)

B52

Divisibility test. Prove that a number is divisible by 13 if and only if, after deleting its last digit and adding 4 times this last digit to the remainder, we get a number divisible by 13. (B. Goncharenko)

B53

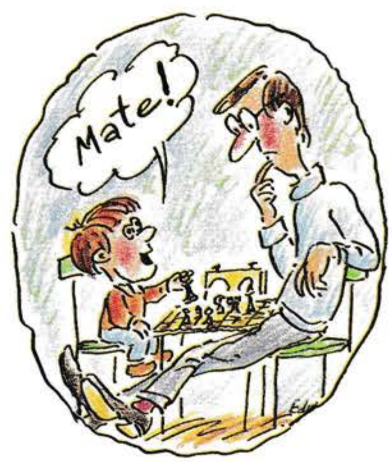
Twilight in the mountains. Mountaineers say that high up in the mountains, twilight is noticeably shorter than down on the plains. What, in your opinion, is the reason for this? (A. Buzdin)



B54

Pythagoras revisited. The figure at left shows a right triangle with three squares on its sides; the vertices of the squares are joined to form three triangles. Prove that these triangles have equal areas. (N. Avilov)

B55



From points to players. Two precocious students from an elementary school took part in a chess championship at a nearby university. Each participant plays each of the others once; a win is worth one point, a draw is worth half a point, and players receive no points for a loss. The combined score of the elementary school students was 6.5; the scores of the university students all happened to be the same. How many university students participated in the championship? (A. Markosian)

ANSWERS, HINTS & SOLUTIONS ON PAGE 59

Art by Edward Nazarov

MATH
INVESTIGATIONS

Triangles of differences

And witnesses for the minimum

by George Berzsenyi

IT ISN'T DIFFICULT TO SHOW that if $n \geq 3$ and $k \geq 1$ are integers, and if one calculates the forward differences of the members of the sequence

$$1, n, n^2, n^3, \ldots, n^{k-1}$$

and successive forward differences thereof, then all the entries of the resulting triangle of numbers will be distinct positive integers. I posed this as a problem in a recent round of the USA Mathematical Talent Search, and I wish to pose it now as the first challenge in this column. The triangle of differences corresponding to $(n, k) = (3, 6)$ is shown below:

1 3 9 27 81 243
2 6 18 54 162
4 12 36 108
8 24 72
16 48
32

As a second challenge, I wish to pose a more general, though a bit more vague, question: **For what other "naturally arising" sequences of positive integers is it true that all successive forward differences are distinct positive integers?"**

Some years ago, when I first thought of this problem, I imposed yet another condition on the members of such sequences: I wanted their largest members to be as small as possible. In particular, I found that for $k = 1, 2, 3, 4, 5$, and 6, this minimal value is $N_k = 1, 3, 8, 20, 43$, and 98, respectively. As one can see from its complete table of differences, shown below, the sequence $\langle 10, 12, 17, 26, 46, 98 \rangle$ is such a "witness" for $N_6 = 98$, whose minimality was established by a computer search:

10 12 17 26 46 98
2 5 9 20 52
3 4 11 32
1 7 21
6 14
8

Can you show that for each value of N_k, the witnesses come in pairs?

Subsequent to my findings, one of my mathematical friends, Dr. Stanley Rabinowitz, was able to determine the next three values of N_k—they are $N_7 = 212$, $N_8 = 465$, and $N_9 = 1{,}000$. The value of $N_{10} = 2{,}144$ was conjectured by a colleague of his; and a search by Dr. Basil Rennie, editor of the (new) *James Cook Mathematical Notes* (*JCMN*), suggests that $N_{11} = 4{,}562$. Can you verify the values of N_{10} and N_{11} and extend these findings?

I reported the first ten values of N_k in the June 1984 issue of *JCMN* (pp. 4054–4055) and gave there a listing of some of the witnesses. Apart from the trivial $\langle 1 \rangle$ for N_1, and $\langle 1, 3 \rangle$ and $\langle 2, 3 \rangle$ for N_2, I conclude now with a complete listing of exactly half of them for $3 \leq k \leq 9$; upon constructing their complete tables of differences, and recalling that witnesses come in pairs, you should have no difficulty in listing the remaining ones.

$k = 3$: $\langle 2, 3, 8 \rangle$
$k = 4$: $\langle 2, 5, 9, 20 \rangle$, $\langle 4, 5, 8, 20 \rangle$, $\langle 4, 6, 9, 20 \rangle$
$k = 5$: $\langle 6, 10, 15, 23, 43 \rangle$
$k = 6$: $\langle 8, 14, 21, 32, 52, 98 \rangle$
$k = 7$: $\langle 11, 18, 27, 39, 59, 102, 212 \rangle$
$k = 8$: $\langle 15, 25, 37, 54, 80, 126, 224, 465 \rangle$
$k = 9$: $\langle 17, 28, 46, 73, 112, 171, 273, 485, 1{,}000 \rangle$,
$\langle 17, 30, 49, 78, 122, 189, 299, 511, 1{,}000 \rangle$

There are many interesting relationships among these witnesses and the resulting complete tables of differences. I encourage you to investigate them.

The purpose of this column is to direct the attention of *Quantum*'s readers to interesting problems in the literature that deserve to be generalized and could lead to independent research and/or science projects in mathematics. Students who succeed in unraveling the phenomena presented are encouraged to communicate their results to the author either directly or through *Quantum*, which will distribute among them valuable book prizes and/or free subscriptions.

KALEIDOSCOPE

One, two, many

Nonmathematicians reckon that mathematicians reckon. Do they reckon rightly?

EACH ONE OF YOU HAS PROBABLY HEARD of people in primitive tribes in Africa and South America that count like this: "one, two, many." But there's another tribe, scattered all over the world, whose representatives count in just the same way. It's the "tribe" of scientists—in particular, mathematicians. You don't believe it? Open an encyclopedia and look at all the words beginning with mono-, di-, poly- and uni-, bi-, multi- (the prefixes meaning "one," "two," "many" in Greek and Latin). Here are some of them.

Monotonicity is a useful and important property of a function or number sequence: a monotonic function (or sequence) either increases or decreases everywhere. By the fundamental theorem of Weierstrass, a bounded monotonic sequence always has a limit. Applied to the monotonic sequence of perimeters of regular n-gons inscribed in a given circle, this theorem makes it possible to define the length of the circumference (see the insert "What is π?" in the article "Calculating π"). Monotonic functions occur elsewhere in this issue—see problem M54 on page 19.

A **monograph** is a book devoted to one particular subject. If a book is called, say, *Monotonic Functions*, be sure it's a monograph. But this doesn't necessarily mean the book has only one author!

A **monomial** (an economical version of mononominal) is the simplest kind of algebraic expression—a product of numbers, variables, and their powers, like x^3, 17, $-zy^2$.

A **dichotomy** is a division or the process of dividing into two, usually repeated many times. It's such a common tool in calculus (for approximating roots of equations and proving some basic theorems to boot) that it has entered the mathematical folklore. Here's one of the numerous "scientific" ways of catching a lion in a desert: build a fence dividing the desert in two and choose the half with the lion; bisect this half with a new fence; choose the part with the lion again; and so on. By the so-called "principle of nested intervals," the successively chosen halves have a unique common point. So the lion will be found at this point surrounded by the fences. Another practical example of dichotomy—a very vivid, though rather cruel one—is given by Harry Graham's short poem:

A **unicursal curve** can be traced without lifting the pen off the curve or tracing any portion of it twice. To tell whether a given curve is unicursal, one must count its odd nodes—that is, the points from which one can move along the curve in an odd number of directions. For a curve to be unicursal it is necessary and sufficient that it have not more than two odd nodes. This was one of the first theorems of topology, proved by Leonhard Euler in 1736.

Uniform means the same throughout.

Univalent mapping never maps two different points onto one and the same point.

Bifurcation means "forking," "dividing into two branches." This term is used in a general situation when something depending on a parameter undergoes a qualitative change at some threshold value of the parameter—the bifurcation point. A fixed point of a function forks, say, into two points mapped onto each other. In the article "Halving Some More" (see the last issue of *Quantum*) you met an example of bifurcation: the simultaneous birth of two cusps on an originally smooth envelope. Some problems connected with bifurcations lead to fractals—those shaggy structures have a lot of branchings in them (see the illustrations to the article on fractals in this issue).

Binary means "in two variables." In the binary number system, all numbers are represented in terms of two symbols, 0 and 1; a binary operation works on two arguments, like the subtraction of numbers or dot product of vectors.

The **binomial theorem** is a famous formula for the expansion of $(a + b)^n$.

A **bisector** is a line dividing an angle, segment, and suchlike into two equal halves.

vivid, though rather cruel one—is given by Harry Graham's short poem:

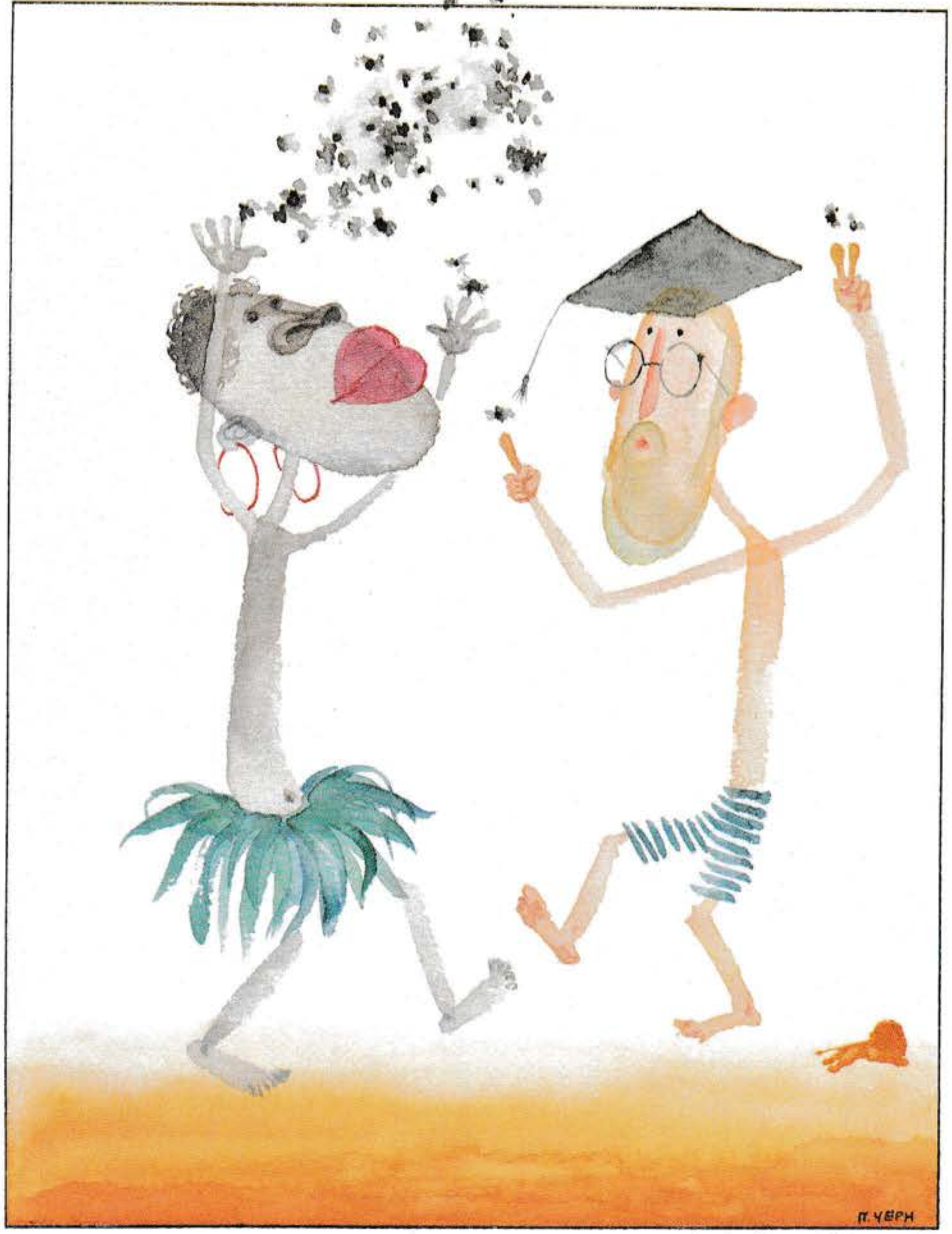
"There's been an accident," they said,
"Your servant's cut in half; he's dead!"
"Indeed!" said Mr. Jones, "and please,
Send me the half that's got my keys."

A **dihedral group** is the symmetry group of a regular polygon. You may have read about it in the articles "Marching Orders" and "Getting It Together with Polyominoes" (see the November/December 1991 issue of *Quantum*).

A **dilemma** is a choice between two possibilities. Perhaps the most famous dilemma is that of the horse that starved to death between two bales of hay.

A **polynomial** is a sum of monomials, like $x^3 - x + 17$ (a polynomial in one variable) or $x^3 - 3xy^2 + xy + 2y$ (a polynomial in two variables). This is one of the most important notions in algebra. And some polynomials are so important they've acquired their own names—Legendre polynomials, Chebyshev polynomials, and so on.

Polygon, polyhedron, polytope—these words mean many-angled, many-faced, and many-placed (so to speak) in Greek. And everybody understands, more or less, what the first two words mean in geometry (the third term is used sometimes for polyhedrons of four or more dimensions). But to give an accurate mathematical formulation of what everybody understands—for instance, what is meant by "polyhedron"—is a far from easy task.



Art by Pavel Chernusky

halves.

Multiplicative means related to multiplication. Multiplicative number theory explores the properties of integers connected with their factors—for example, of prime numbers.

Multinomial—although the literal meaning of this word is the same as that of "polynomial," it's used in a more narrow sense: in connection with the formula for $(a_1 + a_2 + \ldots + a_k)^n$ generalizing the binomial theorem, and mostly in combinatorics and probability.

Multidimensional—how big is "multi" here? In other words, when should one say "many" counting dimensions? The natural boundary is three, because it's the dimensionality of the world we inhabit. But in studying regular polyhedrons, for instance, one finds out that there are five different types in three-dimensional space, six different types in four-dimensional space (and infinitely many different regular polygons on the plane), whereas in any space of dimension $n \geq 5$ there are only three types, and they're the same for all dimensions—one such polygon is the n-dimensional cube. So in this case "many" starts from 5 rather than 4.

Sometimes two numerals are found in one term, as in *biunique* (one-to-one). You may have noticed that there's no law as to which prefix is used, Greek or Latin: *binomial* has a Latin prefix, while *polynomial* has a Greek one (even though the root is Latin); Greek *dichotomy* and Latin *bisection* are absolute equivalents but are used in different contexts.

Similar words occur in science—*multiplet, dioxide, polypod, biceps*, as well as in everyday life—*university, bicycle, monologue* (but not *dialogue*: here the prefix is dia-, "through, across"; compare *diagonal* and *diameter* in geometry). Words beginning with tri-, tetra-, penta-, … and ter-, quadr-, quint-, …—that is, three, four, five, … in Greek and Latin—are also used in math and science (*triangle, triode, quadric*), but they're less frequent. So it's not far from the truth to say that scientists count things around them according to the principle "one, two, many."

The power of dimensional thinking

"In physics . . . there is no place for muddled thinking. . . . Those who really understand the nature of a given phenomenon must obtain the basic laws from dimensional considerations."

—Enrico Fermi

by Yuli Bruk and Albert Stasenko

AT SCIENTIFIC SEMINARS and conferences, where new theoretical or experimental findings are discussed, just about every paper begins with a qualitative description and estimate of the effect the speaker wants to talk about. In even the most thorough paper, lecture, or article, it's impossible to report all the experimental details or all the theoretical "tricks" that were essential in conducting the research or solving various problems. Yet there are certain points one must always address without waiting for queries from the listeners or readers.

First and foremost, one must offer an evaluation of the order of magnitude of the anticipated effect; simple extreme cases; and the nature of the functional connections among the quantities determining the phenomenon. Essentially, analysis of these closely interlinked questions is what we call the qualitative description of the physical situation.

One of the most effective methods for conducting this analysis is the dimensional method. We'll look at its underpinnings in this article. It isn't an exaggeration to say that dimensional analysis is extremely efficient, sparing mountains of paper for the theorist and saving money and time for the experimenter. Rapid estimation of the scale of a phenomenon, development of the principal framework of the experiment, discovery of the qualitative and functional relations, restoration of forgotten formulas during examinations—these are only a few of the merits and applications of dimensional analysis.

Analysis of dimensions has been used in physics since Newton's time. In fact, it was Newton who formulated the principle of similarity, which is closely linked with dimensional analysis. We'll illustrate Newton's principle with a very simple and well-understood example.

Imagine that a body of mass m moves in a straight line under the action of a constant force F. If the body's initial velocity is zero and the velocity after it travels a distance s is equal to v, we can write the equation for the law of conservation of energy as $mv^2/2 = Fs$. So we see there's a functional relation among the quantities of v, F, m, and s.

Now let's suppose that we don't yet know the energy conservation law (or don't want to use it), but we do know that there exists a functional dependence among v, F, m, and s. Very often (but not always, of course!) the functional dependence of the physical quantities is a power law. Let's assume that this is the case here. We can put it another way: we consider that the formula determining the velocity v as a function of F, m, and s takes the form

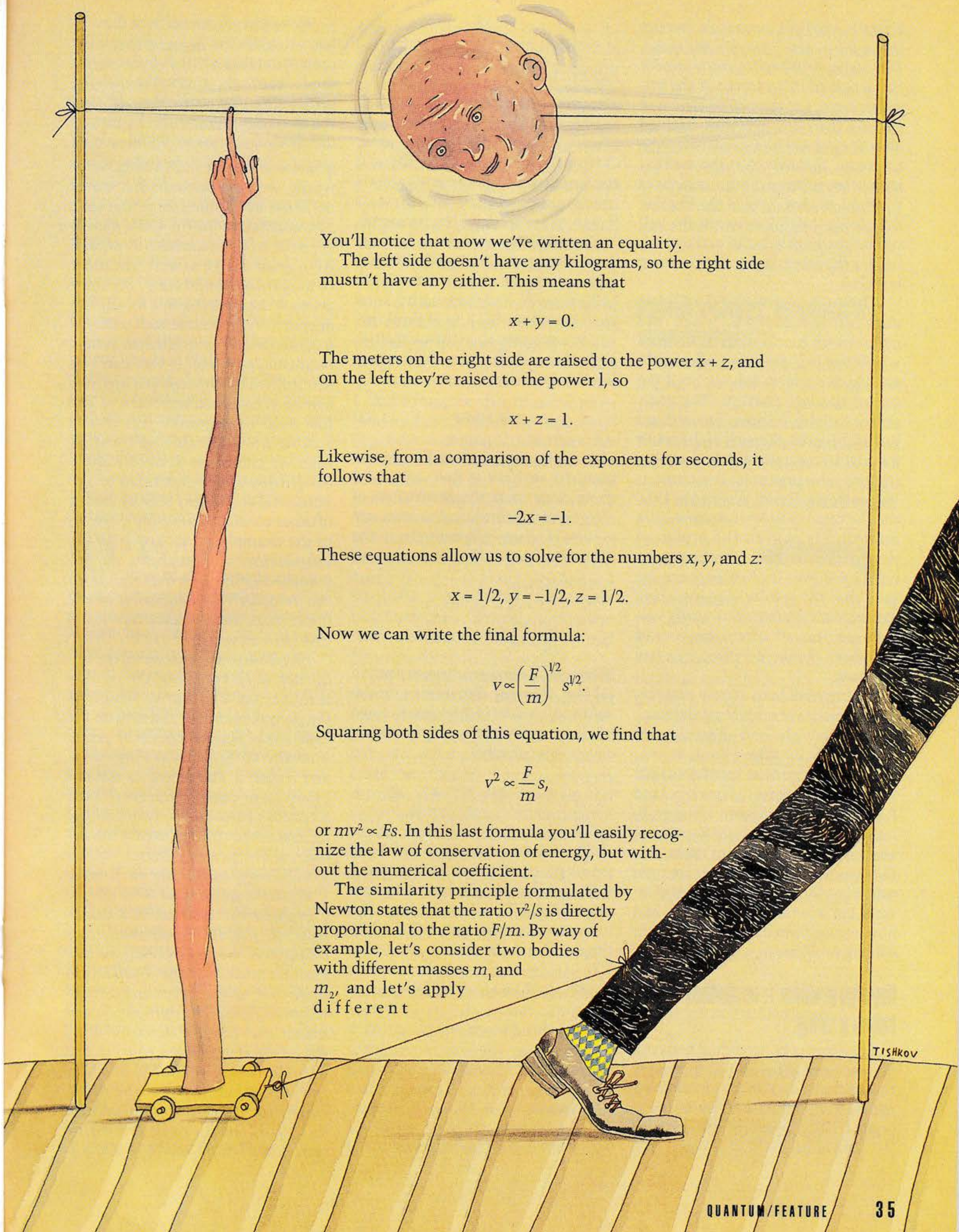
$$v \propto F^x m^y s^z. \qquad (1)$$

Here x, y, and z are numbers we have yet to determine. The "$\propto$" sign means that the left side of the formula is proportional to the right side—that is, $v = kF^x m^y s^z$, where k is a numerical coefficient. Since k is a dimensionless quantity, naturally it's impossible to determine this coefficient by means of dimensional analysis.

Of course, the left and right sides of equation (1) must be measured in the same units—that is, they must have the same dimension. We'll measure v in meters/second, F in newtons, m in kilograms, and s in meters. In other words, we'll choose the dimensions for the quantities v, F, m, and s as follows: $[v] = \mathrm{m/s} = \mathrm{m} \times \mathrm{s}^{-1}$, $[F] = \mathrm{N} = \mathrm{kg} \times \mathrm{m} \times \mathrm{s}^{-2}$, $[m] = \mathrm{kg}$, $[s] = \mathrm{m}$. (The symbol $[A]$ means the dimension of the quantity A.) Then we can write the condition that the dimensions of the left and right sides of formula (1) are the same:

$$\mathrm{m} \cdot \mathrm{s}^{-1} = \mathrm{kg}^x \cdot \mathrm{m}^x \cdot \mathrm{s}^{-2x} \cdot \mathrm{kg}^y \cdot \mathrm{m}^z = \mathrm{kg}^{x+y} \cdot \mathrm{m}^{x+z} \cdot \mathrm{s}^{-2x}.$$

Art by Leonid Tishkov



You'll notice that now we've written an equality.

The left side doesn't have any kilograms, so the right side mustn't have any either. This means that

$$x + y = 0.$$

The meters on the right side are raised to the power $x + z$, and on the left they're raised to the power 1, so

$$x + z = 1.$$

Likewise, from a comparison of the exponents for seconds, it follows that

$$-2x = -1.$$

These equations allow us to solve for the numbers x, y, and z:

$$x = 1/2,\ y = -1/2,\ z = 1/2.$$

Now we can write the final formula:

$$v \propto \left(\frac{F}{m}\right)^{1/2} s^{1/2}.$$

Squaring both sides of this equation, we find that

$$v^2 \propto \frac{F}{m} s,$$

or $mv^2 \propto Fs$. In this last formula you'll easily recognize the law of conservation of energy, but without the numerical coefficient.

The similarity principle formulated by Newton states that the ratio v^2/s is directly proportional to the ratio F/m. By way of example, let's consider two bodies with different masses m_1 and m_2, and let's apply different forces F_1 and F_2 to them such that the ratios F_1/m_1 and F_2/m_2 are the same. The bodies will begin to move under the action of these forces. If the initial velocities of the bodies are equal to zero, then the velocities attained after they travel a distance s will be the same. We arrived at the similarity law by equating the dimensions of the right and left sides of the formula describing the link between the values for the final velocity and the values for the force, mass, and distance traveled.

Although dimensional analysis was introduced back when the groundwork for classical mechanics was being laid, it wasn't effectively applied in problem solving until the end of the last century. The great physicist John William Strutt (Lord Rayleigh) deserves much of the credit for popularizing this method and using it to solve important and interesting problems. Rayleigh wrote in 1915 that he often wondered at how little attention is paid to the profound similarity principle, even by prominent scientists. It often happens, he said, that the results of painstaking research are presented as newly discovered "laws" when they could have been obtained *a priori* in a few minutes.

It's impossible to accuse present-day physicists of neglecting the similarity principle and dimensional analysis. Let's take a look at two classic problems that are often called "Rayleigh problems." Of course, Lord Rayleigh examined many more problems, and he used dimensional analysis to solve many of them. But the ones we'll discuss below are typical. From these and other examples, we'll get a better feel for when and how to use dimensional analysis in solving problems.

Rayleigh problem 1: oscillations of a ball on a string

Let a string be stretched between points A and B (fig. 1). The tension of the string is F. A heavy ball is attached in the middle of this string at point C. The length of the segment AC (and similarly CB) is equal to l.

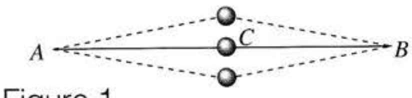


Figure 1

The mass M of the ball is much greater than the mass of the string. The string is pulled and released. Obviously the ball will oscillate. If the amplitude of these oscillations is much smaller than the string's length, the process will be harmonic.

Rayleigh showed how to find the dependence of the oscillation frequency ω on the string tension F, the ball's mass M, and the length l. Now we'll reproduce his line of thinking.

Let's suppose that the quantities ω, F, M, and l are linked by a power law dependence:

$$\omega \propto F^x M^y l^z \qquad (2)$$

The exponents x, y, and z are the numbers we have to find. As we did above, let's write the dimensions of the quantities of interest to us in any system of units—for example, in the SI system:

$$[\omega] = \text{s}^{-1},$$
$$[F] = \text{N} = \text{kg} \cdot \text{m} \cdot \text{s}^{-2},$$
$$[M] = \text{kg},$$
$$[l] = \text{m}.$$

If formula (2) expresses the actual physical law, then the dimensions on the right and left sides of this formula must be the same—that is, the following equation is correct:

$$\text{s}^{-1} = \text{kg}^x \cdot \text{m}^x \cdot \text{s}^{-2x} \cdot \text{kg}^y \cdot \text{m}^z = \text{kg}^{x+y} \cdot \text{m}^{x+z} \cdot \text{s}^{-2x}.$$

There are no meters and kilograms on the left side of this equation, and the seconds have an exponent of 1. This means that the following equations are satisfied by the numbers x, y, and z:

$$x + y = 0, \ x + z = 0, \ -2x = -1.$$

Solving this system, we get

$$x = 1/2, \ y = -1/2, \ z = -1/2.$$

Therefore,

$$\omega \propto F^{1/2} M^{-1/2} l^{-1/2}.$$

We've already mentioned that we had to replace the equal sign with the proportionality symbol because we don't know the numerical coefficient. It's interesting, though, that the exact formula for the frequency differs from the one we found only by a factor of $\sqrt{2}$ ($\omega^2 = 2F/Ml$). In other words, we can conclude that we've obtained an estimate for ω that isn't just qualitative (in the sense that it depends on the quantities F, M, and l) but quantitative as well. The order of magnitude found for F, M, and l gives us the correct value for the frequency. (We'll do more with order-of-magnitude estimates later.) In simple problems, coefficients that can't be determined by dimensional analysis can be considered numbers of the first order of magnitude. We should bear in mind that this isn't a strict rule. We can come to our ultimate conclusion about the value of the numerical coefficient only by means of other considerations. (By the way, in the example we looked at in the introductory section above, the numerical coefficient in the formula for the velocity v as a function of the force, mass, and distance traveled is close to 1 as well.)

Needless to say, instead of the frequency ω we could use the oscillation period $T = 2\pi/\omega$, which is uniquely related to it, and look for the exponential dependence of the period T on the string tension F, mass M, and length l. The factor 2π doesn't "spoil" (nor does it "improve"!) the formula obtained by dimensional analysis—we still can't write the numerical coefficient without strictly solving the equation for the oscillations.

Another simple and well-known example—oscillations of a mathematical pendulum—offers an interesting case. We could obtain the precise formula for the oscillation frequency $\omega^2 = g/l$ by using dimensional analysis. The numerical coefficient that can't be determined by dimensional analysis is equal to 1. If from the outset we tried to find how the oscillation period of the mathematical pendulum was linked with its length l and the free-fall acceleration g, we would arrive at the formula $T \propto \sqrt{l/g}$, which differs from the exact solution by a factor of 2π. Yet it doesn't follow from this example that it's more advantageous to use the oscillation frequency than the period when dimensional analysis is used; the appearance of the factor 2π is linked only with the definition $\omega = 2\pi/T$.

Let's return to the Rayleigh problem and state once more the assumptions allowing us to solve it by dimensional analysis. First, we assumed that a link in fact exists between the quantities ω, F, M, and l. Second, we considered the formula expressing this connection to be a power law: $\omega \propto F^x M^y l^z$.

Dimensional analysis helps us find the functional relations among different physical quantities, but only when these relations are power laws. Fortunately, there are many such relations in nature, and dimensional analysis should become our reliable assistant.

The "N − K = 1" rule

The idea of the dimension of a physical quantity is introduced when certain basic physical quantities have already been chosen and their units of measurement have been determined. In mechanics, we consider mass, length, and time the traditional basic quantities. These quantities are measured in the SI system in kilograms, meters, and seconds; in the CGS system—in grams, centimeters, and seconds.

We'll remind you that the basic units of measurement in each system (the basic dimensions) are called the units of measurement (dimensions) for the basic quantities. The units of measurement for all other nonbasic quantities are expressed in the basic units of measurement. For example, in the SI system the unit of measurement for force, the newton, is $\text{kg} \cdot \text{m/s}^2$, and the unit of measurement for energy, the joule, is $\text{kg} \cdot \text{m}^2/\text{s}^2$; while in the CGS system these are the dyne, which is $\text{g} \cdot \text{cm/s}^2$, and the erg, which is $\text{g} \cdot \text{cm}^2/\text{s}^2$. Units such as the dyne and the erg are called derived units.

If we're dealing with problems in which nonmechanical quantities appear (electrical charge, temperature, and so on), we can increase the number of basic quantities. Current is included in the basic quantities in the SI system (it's measured in amperes), as is temperature (measured in kelvins), and so on.

Generally, the basic quantities and their units of measurement can be chosen in a variety of ways. Much depends on convenience, tradition, and current standards and conventions. We'd like to emphasize that one can use dimensional analysis in any system of units. One must, of course, write the expressions for the dimensions of different quantities in the same system every time.

Let's imagine that in some problem we have to find the functional dependence among N quantities. Assuming that this dependence is a power law, we can try to solve the problem by dimensional analysis. *If the dimensions of the all N quantities are expressed through the dimension of K of the basic quantities, and if $N - K = 1$, then there exists a single formula defining the exponential dependence among the N quantities, and this formula can be found by using dimensional analysis.*

It's not hard to convince ourselves that this is true. We write the general form of the desired formula as follows: one of the quantities is on the left side, and the product of the powers of the other $N - 1$ quantities is on the right side. The exponents are still unknown. The number of unknown exponents is $N - 1$. To determine these exponents we need $N - 1$ equations. We obtain each equation by comparing the exponents on the left and right sides of the equation for one of the basic dimensions. If we have exactly $N - 1$ dimensions in our problem, then we obtain exactly the number of equations we need. These equations are linear, but the existence and uniqueness of the solution of the system of such equations guarantee the existence and uniqueness of the power law in question. The examples given above illustrate this rule quite well. When we write the formulas $v \propto F^x m^y s^z$ or $\omega \propto F^x M^y l^z$, we had $N = 4$ quantities each time, while the number of unknown exponents $N - 1 = 3$ coincided with the number of basic dimensions $K = 3$ (kg, m, s). The systems of three linear equations for the three variables each had a single solution, and the final formulas for v and ω also were the only ones possible. So we'll always be able to build only one formula for the four functionally connected quantities if the number of basic dimensions encountered in the problem equals 3.

Rayleigh problem 2: oscillations of a spherical droplet

Let a drop flow from a round opening (fig. 2). It's natural to assume that the drop must be spherical in the equilibrium state because this makes the surface energy a minimum and any system tends to the state with minimum energy. Even very small deformations of the drop lead to pulsations due to surface tension, so the shape of the drop changes periodically. We assume that the oscillations continue long enough and their damping is small. What interests us is the frequency (or period) of the process. This frequency can depend, obviously, on the surface tension of the liquid σ, the density of the liquid ρ, and the radius of the drop r.[1] We'll look for this dependence in the form

Figure 2

$$\omega \propto \sigma^x \rho^y r^z.$$

Let's write the dimensions of all the quantities in the SI system:

[1]Possibly a question has occurred to you: why not assume that the frequency can depend also on the force of gravity acting on the drop? It's an appropriate question, and we promise to discuss it below.

$[\omega] = s^{-1}$,
$[\sigma] = N \cdot m^{-1} = kg \cdot m \cdot s^{-2} \cdot m^{-1} = kg \cdot s^{-2}$,
$[\rho] = kg \cdot m^{-3}$,
$[r] = m$.

The number of linked quantities we're looking for is again one more than the number of basic dimensions. From the $N - K = 1$ rule it follows that the formula obtained for the frequency must be unique. The equations for determining x, y, and z are obtained from the condition

$$s^{-1} = kg^x \cdot s^{-2x} \cdot kg^y \cdot m^{-3y} \cdot m^z = kg^{x+y} \cdot s^{-2x} \cdot m^{-3y+z}.$$

There are three equations for the three variables:

$$-2x = -1, \; x + y = 0, \; -3y + z = 0.$$

This system has the unique solution

$$x = 1/2, \; y = -1/2, \; z = -3/2.$$

Finally, we get the formula for the oscillation frequency:

$$\omega \propto \sqrt{\frac{\sigma}{\rho r^3}}.$$

This formula also suggests a possible method of experimentally determining σ. We need to know the density of the liquid ρ and the radius of the drop r, and we need to determine the frequency ω experimentally. We don't know the numerical coefficient in this formula, but this isn't a serious obstacle. We can determine it, for instance, from an experiment with a liquid whose surface tension is known.

Essentially we're faced now with a simple case of modeling—we can simulate the oscillations of the drop of liquid being studied by means of the oscillations of a drop of liquid with known σ and ρ. We can also compare the oscillations of the drop's shape for the two different liquids.

We can interpret the formula for the oscillation frequency ω in another way. Let's write it as follows:

$$\frac{\sigma}{\rho} \propto r^3\omega^2.$$

Since σ and ρ are parameters that characterize the liquid and so are the same for drops of this liquid of different sizes, we come to the conclusion that the periods $T_1 = 2\pi/\omega_1$ and $T_2 = 2\pi/\omega_2$ of the oscillations of two drops of the same liquid and their radii r_1 and r_2 are linked by the correlation

$$\frac{T_1^2}{T_2^2} = \frac{r_1^3}{r_2^3}$$

—the squares of the oscillation periods of the two drops are proportional to the cubes of their dimensions! Does this remind you of one of Kepler's laws? It's a strained analogy perhaps, but we're talking about periodic processes in both cases!

We'll make one more observation. Even as we're writing out the set of quantities whose interrelation we're trying to find, we have to realize what's essential to the concrete physical phenomenon and what's not. If we're talking about dynamics (for example, oscillations), then the characteristics of force and mass must appear in the problem. In the problem about the oscillations of a drop, σ played the role of the force characteristic; ρ, the liquid's density, played the role of the mass characteristic. Essentially, we considered the oscillations to be governed by the surface tension only. The solution obtained undoubtedly applies to a drop oscillating in a spaceship. Does it work for a drop near the Earth? Shouldn't we also take gravity into account?

Let's think it through. The force of gravity is $F_g \propto \rho r^3 g$, while the surface tension is $F_t \propto \sigma r$. Clearly, for small enough r the surface tension is greater than the force of gravity. Omitting the numerical coefficient, let's state the inequality expressing the condition for which we can neglect gravity: $\sigma r >> \rho r^3 g$. This inequality is equivalent to $r << (\sigma/\rho g)^{1/2}$. We can state that, for small enough drops, gravity should not influence the oscillations. How big can a drop of water get before we can no longer use the formula obtained above for the oscillation frequency? Try to estimate this maximum size on your own.

The oscillation frequency of atomic nuclei

It turns out we can use the formula for the oscillation frequency of a drop to determine the oscillation frequency of atomic nuclei.

In the droplet model of the atom's nucleus, the nucleus is considered a drop of nuclear material—a "liquid" consisting of protons and neutrons. Surface tension keeps the nuclear droplet from decaying.

Nucleons (protons and neutrons) are inside the nucleus in a bound state. This means that a certain energy must be expended to pull them away from each other. This energy is equal to $\varepsilon \cong 13 \cdot 10^{-13}$ joules per nucleon. The radius of the nucleus is $r \cong 10^{-14}$ m; the proton mass is $m_p \cong 1.7 \cdot 10^{-27}$ kg. Let's try to use all this information to calculate typical oscillation frequencies of atomic nuclei—the droplets of nuclear material.[2]

We could use the same formula for the oscillation frequency of a droplet of nuclear material as for the oscillations of drops of an ordinary liquid if only we could learn to calculate the surface tension of the "nuclear liquid." The total surface energy in the drop model must be of the same order of magnitude as the binding energy of all particles inside the drop. If the number of nucleons in the nucleus is A (the mass number), then the total binding energy of all the nucleons is $A\varepsilon$, and the surface energy of the nuclear droplet is of the same order of magnitude as $A\varepsilon$. Dividing it by the surface area of the drop $S = 4\pi r^2$, we obtain an estimate of the surface tension $\sigma \propto A\varepsilon/4\pi r^2$. The mass of a nucleus consisting of

[2]At this point we're interested only in the qualitative dependence of the oscillation frequency on the parameters of the nucleus and the quantitative order-of-magnitude estimate; therefore, we can neglect the difference between the masses of the proton and neutron and consider the mass of the nucleon equal to that of the proton, for instance.

A nucleons is close to Am_p, and the volume of the nucleus is $(4/3)\pi r^3$; so the density of the "nuclear liquid" is of the order of $\rho \propto 3Am_p/4\pi r^3$. Substituting the expressions obtained for σ and ρ in the formula, we get the result we were looking for:

$$\omega \propto \frac{1}{\sqrt{3}r}\left(\frac{\varepsilon}{m_p}\right)^{1/2}$$
$$\propto \frac{1}{r}\left(\frac{\varepsilon}{m_p}\right)^{1/2}.$$

The typical "nuclear" frequencies are close to 10^{22} s^{-1}. Check that the formula written here leads to a similar result—assuming a numerical coefficient (which we didn't determine) of the order of 1.

Meters "along" and "across"

The problems we considered up to now were in essence solved identically and uniquely. Don't think it's always that way. In some situations the $N - K = 1$ rule doesn't hold, and so we must have recourse to new ideas. One idea is to try to increase the number of basic quantities—that is, to study the problem in a system with a larger number of basic dimensions.

To illustrate this idea, let's look at two simple problems.

No doubt you're acquainted with the first problem. Let's imagine that a ball falls from a table of height H onto the floor. The velocity of the ball is horizontal and equal to v_0 at the moment it falls off the table. Obviously we can correlate the distance it flies with H and v_0. Do these simple calculations before reading any further.

Finished? Now think—is it possible to find the connection between H and v_0 using dimensional analysis? Let's try. Let the distance that it flies be equal to x_0. Undoubtedly the free-fall acceleration g is an essential quantity in this problem along with H and v_0. The answer shouldn't depend on the ball's mass, since this is a purely kinematic problem. So we've got four quantities: x_0, v_0, H, and g. The equations for the dimensions of all these quantities consist only of meters and seconds—that is, $N = 4$, $K = 2$, and $N - K = 2 > 1$! If we write $x_0 \propto v_0^{\alpha}H^{\beta}g^{\gamma}$, then we can get only two equations for the three variables α, β, γ. What do we do now?

Let's introduce separate quantities for measuring the vertical and horizontal distances: we'll measure distances along the horizontal axis x in "horizontal" meters m_x and distances along the vertical axis y in "vertical" meters m_y (fig. 3). Then the dimensions are

$$[x_0] = m_x,$$
$$[v_0] = m_x s^{-1},$$
$$[H] = m_y,$$
$$[g] = m_y s^{-2}.$$

Now the basic dimensions are m_x, m_y, and s—for $N = 4$ we now have $K = 3$. The formula $x_0 \propto v_0^{\alpha}H^{\beta}g^{\gamma}$ leads to the condition

$$m_x = m_x^{\alpha} \cdot s^{-\alpha} \cdot m_y^{\beta} \cdot m_y^{\gamma} \cdot s^{-2\gamma}$$
$$= m_x^{\alpha} \cdot s^{-\alpha - 2\gamma} \cdot m_y^{\beta + \gamma}.$$

The system of equations

$$\alpha = 1, \ -\alpha - 2\gamma = 0, \ \beta + \gamma = 0$$

has the single solution

$$\alpha = 1, \ \beta = \frac{1}{2}, \ \gamma = -\frac{1}{2},$$

and we get the following answer:

$$x_0 \propto v_0\sqrt{\frac{H}{g}}.$$

Compare this solution with the answer you got by exact calculation.

The second problem that illustrates the same idea has to do with the kinetic theory of gases. Molecules of gas are finite in size and can collide with each other even in a rarefied gas. The average distance the molecules travel between two consecutive collisions is called the mean free path. We want to know how the mean free path l depends on the size of molecules r and their concentration n.

Let's write the dimensions: $[n] = m^{-3}$, $[l] = m$, $[r] = m$. When we try to link l, r, and n, we again find that the $N - K = 1$ rule doesn't hold: $N = 3$ and $K = 1$—there is only one basic dimension (meters) in the problem.

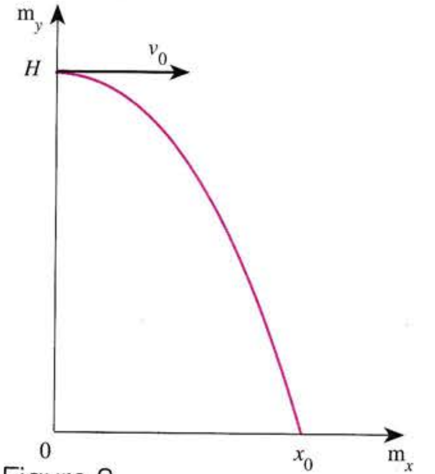


Figure 3

It's convenient to introduce again the "longitudinal" and "transverse" lengths. Let's take the molecules to be little spheres and let's follow one of them. We'll measure the distance along the molecule-ball's trajectory in "longitudinal meters"—$m_{\parallel}$. Obviously the only molecules that can "block" the movement of our chosen molecule are those in the cylinder whose axis is parallel to the trajectory and whose base is the cross section of the molecule-ball perpendicular to the trajectory. The area of this cross section is proportional to r^2. In this situation it's natural to suppose that r is measured in "transverse meters"—$m_{\perp}$. Then the volume is measured in units of $m_{\parallel} \cdot m_{\perp}^2$ and the dimension n is $m_{\parallel}^{-1} \cdot m_{\perp}^{-2}$. After these considerations we've got two basic dimensions—$m_{\parallel}$ and $m_{\perp}$—for three quantities l, r, and n. That's enough for us to obtain the simple formula linking them. It's not difficult to verify that this formula is

$$l \propto \frac{1}{nr^2}.$$

Problems and recommendations

Now you're acquainted with the basic elements of dimensional analysis and how to solve problems with it. We'll emphasize again that the formulas obtained by such considerations allow one to make quantitative estimates as well. You have to be careful, of course, but it would be a

CONTINUED ON PAGE 43

PHYSICS CONTEST

Shake, rattle, and roll

"She stood in silence, listning to the voices of the ground . . ."
—William Blake, "The Book of Thel"

by Arthur Eisenkraft and Larry D. Kirkpatrick

ONE PERSON DESCRIBED how the bedroom wall moved across the room. Another watched as a huge wave of concrete traveled along the highway. We all saw the massive destruction when one bridge roadway collapsed on top of another. The earthquake in the San Francisco area that coincided with the 1989 World Series gave us a glimpse of the power and energy in our planet.

In the fury of the destruction, the Earth is whispering secrets about its composition. The Earth is not solid rock. The Earth is not of uniform density. Longitudinal and transverse waves, called *P* and *S* waves, travel through the Earth as a result of the quake. The differences in *P* and *S* wave behaviors can give us clues about the structure of the Earth while also allowing us to locate the epicenter of the quake.

Although the speeds of the *P* and *S* waves vary within the Earth, the *P* waves always travel faster than the *S* waves. This fact gives us the ability to locate the epicenter of the quake. By knowing the relative speeds of the *P* and *S* waves and measuring the delay in the arrival of the *S* waves, we can determine the distance from the epicenter. Here's an analogy. If you run at 3 m/s and a friend walks at 1 m/s, you will always arrive at a given location before your friend. If you arrive 10 seconds earlier, the distance traveled was 15 meters. If you arrive 20 seconds earlier, the distance traveled was 30 meters.

Let's assume that the epicenter is near the Earth's surface and that the *P* and *S* waves have constant but unequal velocities. If at one location on the Earth the waves arrive with a time difference of 2 minutes, we know that the epicenter of the quake must be situated a specified distance from this location. But in which direction? We don't know. We therefore trace the circumference of a circle on the surface with a radius specified by this time delay. The epicenter can be located on any part of this circumference. If we have a second location with a (different) time delay, this will provide us with a second circle. A third location and a third circle will uniquely determine the actual point location of the epicenter.

The *P* waves are able to travel through solids and liquids, while the *S* waves travel only through solids. The *P* waves arrive at locations on the opposite side of the Earth; the *S* waves do not. This information leads us to conclude that a portion of the interior of the Earth is liquid. By carefully observing where the *P* waves travel and where the *S* waves do not, we can infer more about the size of this liquid core of the Earth.

More curious is the observation that there are positions on the Earth where neither the *P* nor the *S* waves arrive. These shadow zones are somehow protected from disturbances at some locations. What could cause such a shadow region? One explanation is that the *P* waves travel at a different speed within the liquid core. A *P* wave traveling from a solid mantle into a liquid core will change speeds and change direction (that is, they will refract). The result of this refraction is the creation of the shadow region.

Professor Cyril Isenberg, academic leader of the British Physics Olympiad Team, challenged students worldwide in the 1986 International Physics Olympiad with a problem about the *P* and *S* waves of an earthquake. We present parts of that problem as a challenge to our readers.

Let's assume that the Earth is composed of a central liquid spheri-

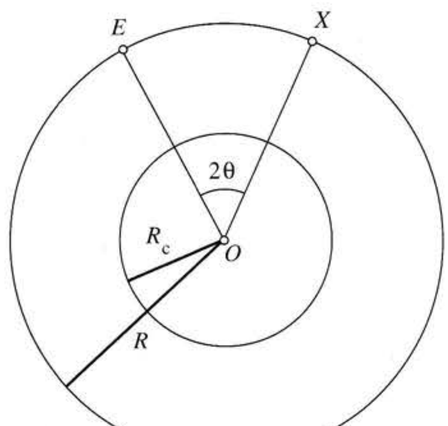


Figure 1

Art by Tomas Bunk


THE WORLD
THE EARTH
YOU ARE HERE!
THE EARTH
CRUST
LITHOSPHERE
MANTLE
ASTHENOSPHERE
OUTER CORE
TOM BUNK

cal core of radius R_c that is surrounded by a solid, homogeneous mantle of radius R. The velocities of the S and P waves through the mantle are v_S and v_P, respectively. An earthquake occurs at point E on the surface of the Earth and produces P and S seismic waves. A seismologist observes the waves at location X. The angular separation between E and X measured from the center of the Earth O is 2θ, as shown in figure 1.

(A) Our beginning physics students should try to show that the seismic waves that travel through the mantle in a straight line arrive at X at a time t (the travel time after the earthquake) given by $t = 2R \sin\theta/v$ for $\theta \le \arccos(R_c/R)$, where $v = v_P$ for the P waves and v_S for the S waves.

(B) After an earthquake an observer measures the time delay between the arrival of the S wave and the P wave as 2 minutes, 11 seconds. Deduce the angular separation of the earthquake from the observer using these data:

R = 6,370 km
R_c = 3,470 km
v_P = 10.85 km/s
v_S = 6.31 km/s

(C) The observer in part B notices that at some time after the arrival of the P and S waves, there are two further recordings on the seismometer separated by a time interval of 6 minutes, 37 seconds. Explain this result and verify that it is indeed associated with the angular separation determined in part B.

(D) For those of you who wish to plunge deeper, draw the path of a seismic P wave that arrives at an observer, where $\theta \le \arccos(R_c/R)$, after two refractions at the mantle–core interface. Obtain a relation for P waves between θ and i, the angle of incidence of the seismic P wave at the mantle–core interface.

(E) For our advanced problem solvers, using the data above and the additional fact that the speed of the P waves in the liquid core is 9.02 km/s, draw a graph of θ versus i. Comment on the physical consequences of the form of this graph for observers stationed at different points on the Earth's surface.

(F) Sketch the variation of the travel time taken by the P and S waves as a function of θ for $0 \le \theta \le 90$ degrees.

Send your answers to *Quantum*, 3140 North Washington Boulevard, Arlington, VA 22201. If you're a student, please indicate your grade.

The leaky pendulum

Very good solutions were submitted to our contest problem in the November/December issue by Ben Davenport (Charlotte, North Carolina), Samuel Dorsett (Mitchell, Indiana), and Jesse Tseng (Little Rock, Arkansas). We also received a packet of solutions from Campbell High School in Campbell, Missouri.

We begin by calculating the length of time t_{max} it takes for the fluid to run out of the container. This is given by the initial mass M_0 of the fluid divided by the rate r at which the fluid runs out. Therefore,

$$t_{\max} = \frac{M_0}{r} = \frac{8a^3\rho}{r},$$

where ρ is the density of the fluid. In order to avoid writing this set of constants repeatedly, let's define a new dimensionless, timelike variable τ by the relationship

$$\tau = \frac{t}{t_{\max}}, \quad 0 \le \tau \le 1.$$

In terms of τ, the mass M remaining as a function of time is given by

$$M(t) = M_0(1 - \tau)$$

and the height l of the fluid as a function of time is given by

$$l(t) = 2a(1 - \tau).$$

Since the center of mass of the fluid is located at its geometrical center, the length of the pendulum is given by

$$L(t) = L_0 + a - \frac{l(t)}{2} = L_0 + a\tau.$$

This expression could have been written down directly by realizing that the center of mass moves from L_0 to $L_0 + a$ at a uniform rate during the "time" τ.

Under the simplifying assumptions of this problem, the period of the pendulum is

$$T = 2\pi\sqrt{\frac{L_0 + a\tau}{g}} = T_0\sqrt{1 + \frac{a\tau}{L_0}},$$

where $T_0 = 2\pi\sqrt{L_0/g}$ is the initial period of the pendulum. The graph of the period versus time is shown as the upper curve in figure 2 for the case $L_0 = 2a$. Although the curve appears to be straight, it actually has a slight curve due to the square root. Note that the period is not defined after the fluid has all run out as the pendulum no longer has any mass.

When the container has a mass

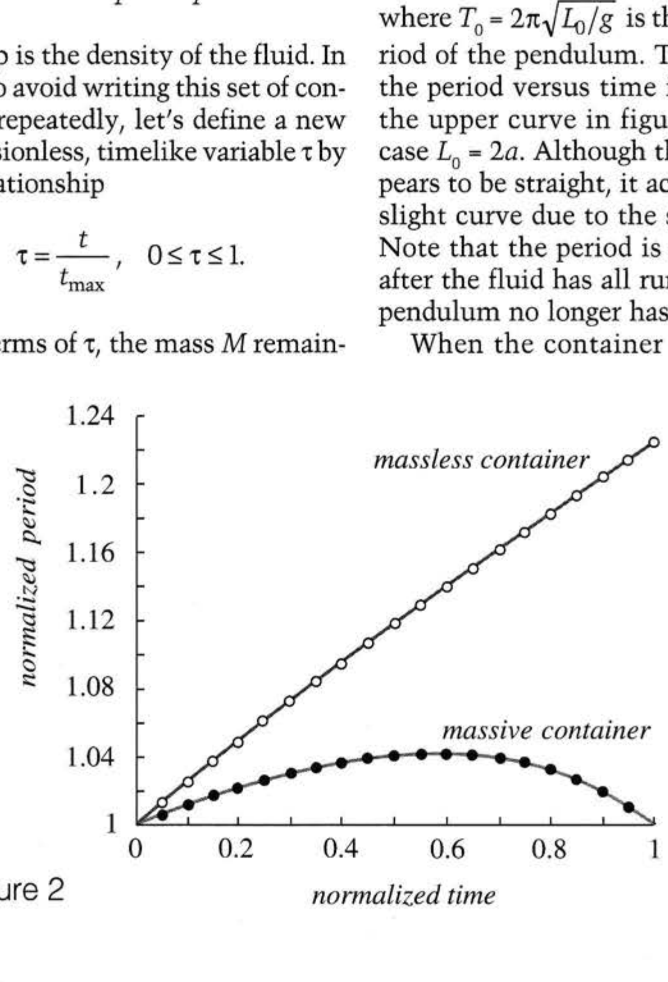


Figure 2

M_0, we must calculate the combined center of mass x_{cm} of the remaining fluid and the mass of the container

$$x_{\mathrm{cm}} = \frac{m_c x_c + m_f x_f}{m_c + m_f},$$

where the subscripts "c" and "f" refer to the container and fluid, respectively. If we choose to calculate the center of mass x_{cm} relative to the center of the container, we have $x_c = 0$ and

$$x_{\mathrm{cm}} = \frac{M_0(1-\tau)a\tau}{M_0 + M_0(1-\tau)}$$
$$= a\frac{\tau(1-\tau)}{2-\tau}.$$

Therefore, the period of the pendulum as a function of time is

$$T = T_0\sqrt{1+\frac{a\tau(1-\tau)}{L_0(2-\tau)}}.$$

This function is shown by the lower curve in figure 2. Note that the maximum period occurs about 60% of the way through the time period and that the maximum period is less than for the case with the massless container. Note also that the period returns to its initial value when the fluid has completely run out. You probably anticipated this because the center of mass must return to the center of the container.

Those of you who know some calculus can differentiate the expression for the center of mass to find that it reaches its maximum value when

$$\tau = 2-\sqrt{2} = 0.586.$$

Notice that this maximum occurs 58.6% of the way through the time period independent of the rate of flow. This value can now be substituted back into the expression for the period to find its maximum value

$$T_{\max} = T_0\sqrt{1+\frac{\left(3-2\sqrt{2}\right)a}{L_0}}.$$

We hope you enjoyed thinking about the leaky pendulum.

"DIMENSIONAL THINKING"
CONTINUED FROM PAGE 39

mistake not to use the fact that the numeric coefficients in formulas often turn out to be of order unity.

It goes without saying that estimation, the construction of simple models, and the use of analogies are only the first steps in studying physical processes. A more precise and rigorous examination of the phenomena must follow. We don't want you to get the impression that dimensional analysis is omnipotent. Before using it in any of its forms, you should try to imagine the physical process more clearly and think hard about the characteristics that are of interest. Only on this condition can one hope to succeed.

We leave you with some problems to think about on your own.

Problems

1. Find the formula describing the relation between a planet's mass M, its radius R, and the minimum period of rotation about its axis. Take into consideration that the existence of such huge "balls" as the planets depends on the gravitational interaction of the particles from which these "balls" are made. What is the minimum rotation period for a planet whose mass and radius are equal to the Earth's? Work out an order-of-magnitude estimate.

2. The linear dimensions of two geometrically similar steel tuning forks differ by a factor of 3. How do the frequencies of these tuning forks differ?

3. Find the dependence of the pulsation period of a gas bubble formed at a point explosion deep underwater if energy E was released and the water pressure equals p. Also estimate the maximum size of the gas bubble. How does the pulsation period depend on the depth H?

4. Estimate the pressure in the center of a star of mass M and radius R. Calculate the same thing for the Sun ($M_s = 2\cdot10^{30}$ kg, $R_s = 7\cdot10^5$ km), a white dwarf ($R_d = 10^3$ km), and a neutron star ($R_n = 20$ km). The masses of the white dwarf and neutron star are close to that of the Sun.

5. Compare the surface tension of the "nuclear liquid" discussed above and that of water.

ANSWERS IN THE
JULY/AUGUST ISSUE


π=3
Д. Крылов

Calculating π

The contribution of Christiaan Huygens

by Valery Vavilov

IN HIS FAMOUS TREATISE *Measurement of the Circle*, Archimedes (ca. 287–212 B.C.), the greatest mathematician of antiquity, gave the following approximation for the value of π:

$$3\frac{10}{71} < \pi < 3\frac{10}{70},$$

or, in decimals,

$$3.14084507 < \pi < 3.14285714.$$

(The correct value of π to eight decimals is 3.14159265.) He created the classic and, perhaps, the most natural method for computing π: since it is the length of a circle of diameter 1, this number can be approximated by the perimeters p_n and q_n of regular n-sided polygons inscribed in and circumscribed about this circumference (fig. 1), so for $n \geq 3$,

$$p_n < \pi < q_n.$$

As the number n of sides increases, the polygons approach a circle, and their perimeters monotonically approach π. So, in principle, we can attain as small an error as we want by choosing a sufficiently large n. [As obvious as they seem, these considerations turn out to be rather difficult to corroborate thoroughly when treated more rigorously—you'll hardly find a geometry textbook in which it's proven that, say, the sequence p_n has a limit. For this reason, we've inserted a more detailed discussion of the definition of π (beginning on the next page), which can be regarded as a sort of introduction and postscript to this article.—*Ed.*]

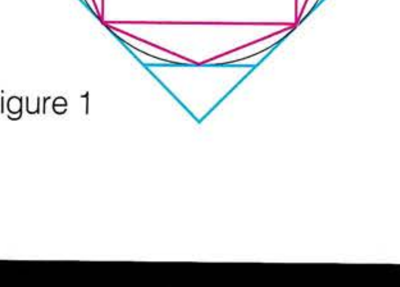


Figure 1

Art by Dmitry Krymov

To obtain his approximation, which remained unsurpassed for many centuries, Archimedes had to compute the perimeters of inscribed and circumscribed 96-gons; in fact,

$$3\frac{10}{71} < p_{96} < q_{96} < 3\frac{1}{7}.$$

His followers, in their struggle for an ever more accurate approximation, managed to improve it by increasing the number of sides. (A short history of these efforts can be found in the Kaleidoscope of the January/February 1991 issue of *Quantum*.) But in 1654 a totally unexpected result was discovered by the great Dutch scientist and mathematician Christiaan Huygens (who was only 25 years old at the time): to get the precision Archimedes had attained in computing π, regular 12-gons will suffice. This result was published in his treatise *On the Discovered Size of the Circle* (*De Circuli Magnitudine Inventa*). Huygens achieved his result only by improving the technique for calculating the perimeters of regular polygons. His work was based on new ideas, which were further developed in our time. His basic idea can be formulated as follows.

Since $p_n < \pi < q_n$ for all $n \geq 3$, π lies in the interval (p_n, q_n). Let's divide this interval into three equal parts (fig. 2). Can we say in advance which of these parts contains the number π?

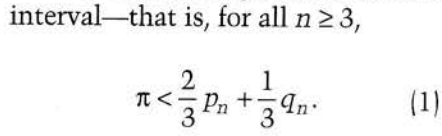


Figure 2

One of the main results of Huygens's work is that π always lies in the left interval—that is, for all $n \geq 3$,

$$\pi < \frac{2}{3}p_n + \frac{1}{3}q_n. \qquad (1)$$

It is this fundamental inequality that allowed Huygens to surpass Archimedes. Let's prove Huygens's result and then consider some related problems.

The proof of the Huygens formula

I'll break the proof of inequality (1) down into several steps. Throughout the article, parentheses will denote the areas of polygons; so (XYZ) is the area of triangle XYZ, and so on. The term "segment" will be used for a portion of a circle cut off by a chord.

CONTINUED ON PAGE 48

What is π?

The first section of this note is a sort of preamble to "Calculating π"; the second part is a kind of afterword.

The definition

Some students, when they are asked to prove that $3 < \pi < 4$, say something like "There's nothing to prove, because everyone knows that $\pi = 3.14$." This answer reminds me of my friend's favorite story about a handbook for foresters, which gave, along with lots of other useful information, an approximate formula for estimating the volume of a tree: $V = \pi D^2 h/4$, where D is the diameter of the tree, h is its height, and π is "a mathematical constant that for pine trees is assumed to be equal to 3."

To tackle the problem of estimating the number π correctly, one must start with a definition of π. The most common one (though not the only one) defines π as the ratio of the circumference of a circle to its diameter. (Of course, it must be accompanied by a justification—that is, a proof that this ratio is the same for all circles.) Now one may, and even should, ask: what's the length of a circumference? This question also has a standard answer: it is the value approached by the sequence of perimeters p_n of regular n-sided polygons inscribed in the circle as n approaches infinity; in other words, the limit of this sequence. In particular, for a circle of diameter 1, this limit simply equals π, and this will be the case we'll examine further.

So now we have the accurate definition we wanted. But it doesn't help us much, because we don't know how the sequence p_n approaches π: do the inequalities $3 < p_5 < 4$ imply a similar estimation of π or don't they? Also, now that we've stepped onto the path of mathematical rigor, we should go all the way to the end and justify the definition of the circumference, too—that is, prove that the sequence p_n does approach some value after all. To this end it suffices to show that this sequence is monotonic (in the case of p_n, increasing) and bounded; then the existence of the limit will follow from the well-known Weierstrass theorem—one of the first theorems on limits[1]. The boundedness of the perimeters p_n is the easier part of the matter: p_n for any n is less than the perimeter of an arbitrary circumscribed polygon because, in general, the perimeter of any convex polygon is less than the perimeter of any polygon containing it. (To prove the last statement, one can turn the outer polygon into the inner one by successively cutting off pieces along the sides of the inner polygon, as in figure 1, where the numbers indicate the order of the cuts—each cut diminishes the perimeter of the outer polygon, since a straight line is the shortest distance between two points.)

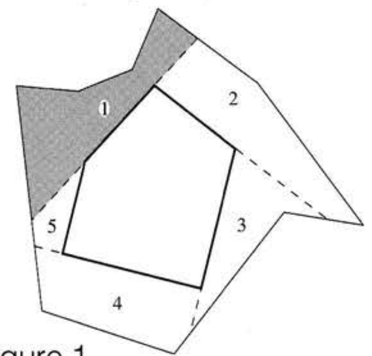


Figure 1

But what about the monotonic growth? Clearly, $p_n < p_{2n}$ for any $n \geq 3$, because the inscribed $2n$-gon can be always constructed so as to contain the n-gon (see figure 1 in "Calculating π") by adding n midpoints of its circumcircle arcs subtended by its sides to the n vertices of the n-gon. Thus, any *subsequence* of the sequence p_n obtained by consecutive duplication of the number of sides—that is, any subsequence of the form $p_k, p_{2k}, p_{4k}, \ldots, p_{2^n k}, \ldots$—increases monotonically. So we can simply confine ourselves to one of these bounded and increasing subsequences—say, the subsequence $p_3, p_6, p_{12}, \ldots, p_{3\cdot 2^n}, \ldots$, and *define* π as the limit of this subsequence.

[1]This theorem can be found in any beginning book on real analysis—for example, *Real Variables* by John M. H. Olmsted (New York: Appleton-Century-Crofts, 1956).—*Ed.*

Such a definition is absolutely correct, though somewhat clumsy: it leaves us in the dark as to whether the limits of other such sequences (like p_{2^n} or $p_{5\cdot 2^n}$) are the same; and we still haven't proved that the sequence p_n has any limit at all. However, this definition allows us to determine π to any desired accuracy (for instance, to prove the inequalities with which we started: $3 < \pi < 4$). Indeed, since it's the limit of the strictly increasing sequence $p_{3\cdot 2^n}$, the number π is greater than any of its terms; in particular, $\pi > p_6 = 3$ (the side length of a regular inscribed hexagon is equal to the radius, which is 1/2 in our case). On the other hand, as we've mentioned, the perimeter q_k of any circumscribed regular polygon is greater than any of the perimeters p_m, so it's not less than their limit $\lim_{n\to\infty} p_{3\cdot 2^n} = \pi$. In particular, for a circumscribed square we have $q_4 = 4 \geq \pi$. To prove the strict inequality $\pi < 4$, one can insert the perimeter q_8 of the circumscribed octagon between π and q_4: $\pi \leq q_8$, and $q_8 < q_4$, and also $q_{2n} < q_n$ for any $n \geq 3$ (the $2n$-gon can be obtained by cutting off all the corners of the n-gon, thus diminishing its perimeter—see figure 1 in "Calculating π").

By the way, we've come across an important inequality here: $q_n > q_{2n}$, which means that $q_{k\cdot 2^n}$ is a decreasing sequence and, therefore, has a limit. This fact can be used to prove (at last!) that the sequence p_k (and q_n) converges to some limit.

First, we notice that

$$0 < q_n - p_n = p_n(1 - q_n/p_n) \to 0$$

as $n \to \infty$, because the p_n's are bounded, and q_n/p_n is the ratio of similarity of the circumscribed regular n-gon to the inscribed regular n-gon and, obviously, tends to 1 as $n \to \infty$. In particular, as $k \to \infty$, $q_{3\cdot 2^k} - p_{3\cdot 2^k} \to 0$, so $\lim_{k\to\infty} q_{3\cdot 2^k} = \lim_{k\to\infty} p_{3\cdot 2^k} = \pi$. Now, $p_n < q_{3\cdot 2^k}$ for any $n \geq 3$, $k \geq 0$, so $p_n \leq \lim_{k\to\infty} q_{3\cdot 2^k} = \pi$; in the same way, the inequality $q_n > p_{3\cdot 2^k}$ implies $q_n \geq \pi$ for any $n \geq 3$. Finally, $0 < \pi - p_n < q_n - p_n \to 0$ as $n \to \infty$, which means that $\lim_{n\to\infty} p_n$ does exist and equals π. Similarly, $\lim_{n\to\infty} q_n = \pi$.

So, with some effort we've managed to prove that the equality $\lim_{n\to\infty} p_n$ can be correctly adopted as a definition of the number π. The proof that p_n increases is even more sophisticated and lies outside the scope of this note.

You may want to return to the article at this point and meet me here again when you're through.

Hands-on computing

Now we'll follow in the ancient footsteps of Archimedes (and follow up on the definition of π presented in the first section of this note). To find an approximate value of π, one can calculate the perimeters p_n of regular n-gons inscribed in a circle.

Figure 2 suggests an efficient way to perform these calculations. Denote by a_n the side length of such an n-gon for a circle of radius 1. Then in figure 2 we can label $AB = a_{2n}$, $AC = a_n/2$. Applying the Pythagorean theorem to the right triangles ABC and OAC (where O is the circle's center), we get

$$\begin{aligned} a_{2n}^2 &= AB^2 \\ &= AC^2 + BC^2 \\ &= \frac{a_n^2}{4} + (OB - OC)^2 \\ &= \frac{a_n^2}{4} + \left(1 - \sqrt{1 - \frac{a_n^2}{4}}\right)^2 \\ &= 2 - 2\sqrt{1 - \frac{a_n^2}{4}} \\ &= 2 - \sqrt{4 - a_n^2}. \end{aligned}$$

So we can start with some k-gon and, using this formula over and over, find the side length and the perimeter for the $k \cdot 2^n$-gon for an arbitrarily large n. Then the perimeter will be an approximation of 2π. In particular, for $k = 4$—that is, starting with a square whose side length is $a_4 = \sqrt{2}$—we get

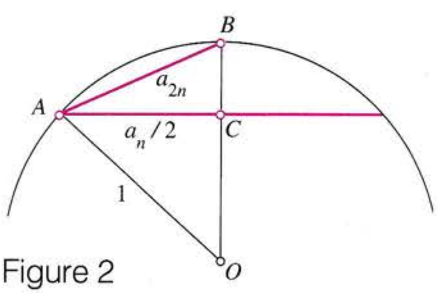


Figure 2

$$\begin{aligned} a_{2^n}^2 &= 2 - \sqrt{4 - a_{2^{n-1}}^2} \\ &= 2 - \sqrt{4 - \left(2 - \sqrt{4 - a_{2^{n-2}}^2}\right)} \\ &= 2 - \sqrt{2 + \sqrt{4 - a_{2^{n-2}}^2}} \\ &= \cdots \\ &= 2 - \sqrt{2 + \sqrt{2 + \cdots + \sqrt{2 + \sqrt{4 - a_4^2}}}} \\ &= 2 - \underbrace{\sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}_{n-2 \text{ roots}}. \end{aligned}$$

Thus, the number π is the limit of the following expression as $m \to \infty$:

$$\frac{1}{2} p_{2^{m+1}} = 2^m \underbrace{\sqrt{2 - \sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}}_{m \text{ roots}}.$$

This formula is beautiful but not too convenient for calculations. We'll see below that it's much better to compute both perimeters p_n and q_n simultaneously: the calculations will be simpler; both the lower and upper bounds will be obtained; and we'll be able to apply the Huygens formula, which is more exact (see "Calculating π").

Look at figure 5 in "Calculating π." We can assume that the chord AB in this figure is a side of a regular inscribed $2n$-gon, so $AB = a_{2n}$. Then segments BC and CA are halves of two sides of the regular circumscribed $2n$-gon, so $BD = 2BC = b_{2n}$—the side length of the circumscribed $2n$-gon. Also, it's clear that segments AH and BE are halves of sides of the inscribed and circumscribed n-gon, respectively: $AH = a_n/2$, $BE = b_n/2$. By the formula obtained in step 4 of the proof of the Huygens formula in "Calculating π," BD is the harmonic mean[2] of AH and BE, so

$$b_{2n} = \frac{2 \cdot a_n/2 \cdot b_n/2}{a_n/2 + b_n/2} = \frac{1}{2} h(a_n, b_n),$$

where h denotes the harmonic mean. Multiplying both parts by $2n$, we get

$$q_{2n} = h(p_n, q_n) = \left(\frac{p_n^{-1} + q_n^{-1}}{2}\right)^{-1}.$$

(Check this formula yourself.) Further, the right triangles AHB and BAD are similar ($\angle BAH = \angle ABD$). Therefore, $AB : AH = BD : AB$, so $AB^2 = BD : AH$, or

$$a_{2n} = \sqrt{b_{2n} \cdot \frac{a_n}{2}} = g\left(b_{2n}, \frac{a_n}{2}\right),$$

where g denotes the geometric mean. Multiply the last formula by $2n$ again:

$$p_{2n} = g(p_n, q_{2n}).$$

So we start with some small k, find q_k and p_k, and then successively compute the harmonic and geometric means of the last two terms of the emerging sequence $q_k, p_k, q_{2k}, p_{2k}, q_{4k}, p_{4k}, \ldots$. The results for $k = 3$ are presented in table 1 in "Calculating π." The limiting value of this sequence, as we know very well, is $2\pi r$, where r is, of course, the radius of the circle in question. I advise you to check table 1 on your own—the most primitive calculator will suffice; and if you choose the right sequence of calculations, you won't even have to write down any intermediate results—just the values of the perimeters. As a theoretical exercise, try to prove that when we start with two arbitrary numbers q and p, $q > p > 0$, the limit of our sequence is $\left(qp/\sqrt{q^2 - p^2}\right) \cdot \arccos(p/q)$. What will the limit be if we replace the harmonic mean in our procedure with the arithmetic mean? (In this case, the limit is called the *Schwab mean* of the initial numbers q and p.)

—Vladimir Dubrovsky

[2] If a and b are positive real numbers, then their geometric mean is defined as $\sqrt{ab}$, while their harmonic mean is $2ab/(a + b)$. It's not hard to show that the harmonic mean of two numbers is never greater than their geometric mean. For a fuller discussion see, for example, *An Introduction to Inequalities* by Edwin Beckenbach and Richard Bellman (MAA New Mathematical Library, 1961).—*Ed.*

CONTINUED FROM PAGE 45

Step 1. Let MAB be an isosceles triangle inscribed in a segment cut off from a circle by the triangle's base AB, and let CKL be the triangle formed by the tangents to the circle at points A, B, and M (fig. 3). Then

$$(MAB) < 2(CKL). \qquad (2)$$

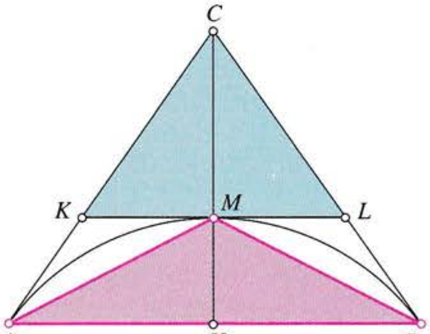


Figure 3

Proof. Draw the altitude CH. By the similarity of triangles CKL and CAB,

$$\begin{aligned}\frac{(MAB)}{(CKL)} &= \frac{AB \cdot HM}{KL \cdot MC} \\ &= \frac{AC}{KC} \cdot \frac{AK}{KC} \\ &= \left(1 + \frac{AK}{KC}\right) \cdot \frac{AK}{KC}.\end{aligned}$$

Since the tangents KA and KM are of equal length, $AK/KC = KM/KC < 1$. It follows that the last expression in the equalities above is less than 2.

Step 2. The area s of the circular segment AMB in figure 3 satisfies the inequality

$$s < \frac{2}{3}(ABC). \qquad (3)$$

Proof. Let's approximate the segment from the inside by the sequence of inscribed polygons obtained by adding isosceles triangles (colored red in figure 4) constructed on the sides of the previous polygon, starting with the triangle ABM. First, we add triangle AMM_1, where M_1 is the midpoint of arc AM, and do likewise for the side MB; then we add

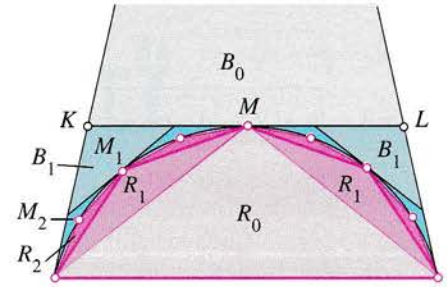


Figure 4

four more triangles, all congruent to AM_1M_2, where M_2 is the midpoint of arc AM_1; and so on. If R_0 is the area of triangle ABM, R_1 is the area of AMM_1, R_2 the area of AM_1M_2, and R_n the area of each of two triangles added at the nth step, then

$$R_0 + 2R_1 + 2^2R_2 + \ldots + 2^nR_n + \ldots = s.$$

Similarly, let's approximate the segment from the outside by cutting pieces from triangle ABC (the blue triangles in figure 4), making the cuts along the tangents at points M, M_1, and all the other vertices of the red triangles as they appear during the process of approximation. If B_0 is the area of triangle KLC, and $B_1, B_2, \ldots$ are the areas of the successive blue triangles, then

$$\begin{aligned}B_0 + 2B_1 + 2^2B_2 + \ldots \\ + 2^nB_n + \ldots = (ABC) - s.\end{aligned}$$

By inequality (2) in the first step of the proof, $R_n < 2B_n$; multiplying by 2^n and summing over all n, we get $s < 2[(ABC) - s]$, so $s < (2/3)(ABC)$.

Step 3. Now, at last, we can get the length l_{AB} of arc AB. Let's prove the following lemma of Huygens, which is interesting in its own right (refer to figure 5):

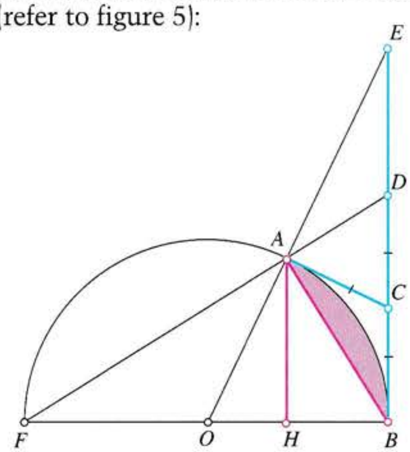


Figure 5

Lemma. *If AH is the perpendicular dropped from A onto the diameter BF of the circle and D is the point where the extended chord FA meets the tangent at B, then*

$$l_{AB} < \frac{1}{3}AH + \frac{2}{3}BD. \qquad (4)$$

Proof. The arc length l_{AB} can be represented as the area S of the sector OAB of the given circle (O is its center) divided by $r/2$—half the circle's radius. Indeed, if $\angle ABC = \theta$ (in radian measure), then arc $AB = (\theta/2\pi)(2\pi r) = r\theta$, while the area of $AOB = (\theta/2\pi)(\pi r^2) = \theta r^2/2$, so arc AB = (area of sector)/$(r/2)$. So let's calculate the area S. Since tangents CA and CB to circle O are equal, the midpoint C of the hypotenuse BD of right triangle ABD is equidistant from B and A. Therefore, ABC is just the same isosceles triangle with respect to the circular segment AB as the one we considered in the previous steps, and the area of the segment $s < (2/3)(ABC)$ by inequality (3). It follows that the area of the sector is

$$\begin{aligned}S &= (OAB) + s < (OAB) + \frac{2}{3}(ABC) \\ &= \frac{1}{3}(OAB) + \frac{2}{3}(OACB),\end{aligned}$$

where $(OACB)$ is the area of quadrilateral $OACB$. Now

$$(OAB) = \frac{1}{2}OB \cdot AH = AH \cdot \frac{r}{2},$$

$$(OACB) = 2(OBC) = BC \cdot r = BD \cdot \frac{r}{2},$$

and therefore

$$l_{AB} = \frac{S}{r/2} < \frac{1}{3}AH + \frac{2}{3}BD.$$

Step 4. In the situation established in the previous step (fig. 5), let the extensions of the radius OA and the tangent BD meet at E. Then

$$l_{AB} < \frac{2}{3}AH + \frac{1}{3}BE.$$

Proof. All we need to do is show that BD is less than the arithmetic mean $(AH + BE)/2$ of AH and BE, since in this case, according to inequality (4),

$$l_{AB} < \frac{1}{3}AH + \frac{2}{3}\frac{AH+BE}{2}$$
$$= \frac{2}{3}AH + \frac{1}{3}BE.$$

Some fancy algebra will allow us to compute BD. By the similarity of triangles FBD and FHA and of OBE and OHA, we have

$$BD = AH\cdot\frac{FB}{FH} = AH\frac{2FO}{FO+OH}$$
$$= \frac{2AH}{1+OH/OB} = \frac{2AH}{1+AH/BE}$$
$$= \frac{2AH\cdot BE}{AH+BE}.$$

The last expression is called the *harmonic mean* of $a = AH$ and $b = BE$ (see footnote 2 on page 47); it's quite easy to prove that the harmonic mean of positive numbers a and b never exceeds their arithmetic mean:

$$\frac{2ab}{a+b} \le \frac{(a-b)^2/2 + 2ab}{a+b}$$
$$= \frac{(a+b)^2}{2(a+b)} = \frac{a+b}{2}.$$

Now for the last step, in which we derive inequality (1).

Step 5. Let AA_1 be a side of a regular inscribed n-gon, and let it be parallel to side EE_1 of the regular n-gon circumscribed about the same circle so that its sides are parallel to those of the inscribed n-gon (fig. 6). Figure 6 can be obtained from figure 5 by adding its reflection through the diameter FB. So the length of arc AA_1 is

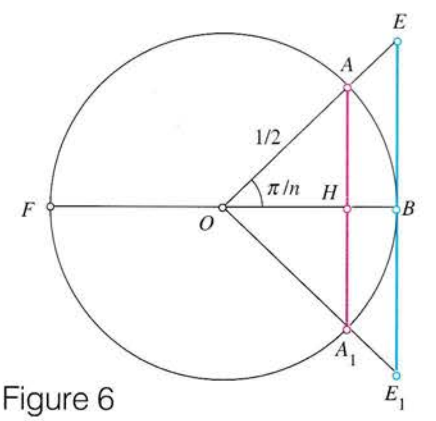


Figure 6

$$l_{AA_1} = 2l_{AB} < 2\left(\frac{2}{3}AH + \frac{1}{3}BE\right)$$
$$= \frac{2}{3}AA_1 + \frac{1}{3}EE_1.$$

Multiplying this by n, we get the basic Huygens formula in inequality (1).

In pursuit of accuracy

We've established that the number π lies in the first third of interval (p_n, q_n) for all $n \ge 3$. To determine its location more exactly, consider the ratio of the lengths of intervals (π, q_n) and (p_n, π). Calculations show that this ratio, $(q_n - \pi)/(\pi - p_n)$, is quite close to 2 for large enough n (see tables 1 and 2). Based on these calculations we can suppose with certainty that this ratio approaches 2 as n increases indefinitely—that is,

$$\lim_{n\to\infty}\frac{q_n - \pi}{\pi - p_n} = 2. \qquad (5)$$

I'll prove this fact using some elementary calculus.

Inspecting figure 6, in which segments AA_1 and EE_1 can be considered sides of an inscribed and a circumscribed n-gon, respectively, we find $p_n = n \sin \pi/n$, $q_n = n \tan \pi/n$. So

$$\frac{q_n - \pi}{\pi - p_n} =$$
$$\frac{1}{\cos(\pi/n)}\cdot\frac{\sin(\pi/n) - (\pi/n)\cos(\pi/n)}{(\pi/n) - \sin(\pi/n)}.$$

As $n \to \infty$, the first factor on the right side of this equation approaches 1 ($\cos x \to 1$ as $x \to 0$). So it suffices to show that

$$f(x) = \frac{\sin x - x\cos x}{x - \sin x} \to 2$$

as $x \to 0$ (I replaced π/n with x). A little later, I'll prove the following estimates: for any $x > 0$,

$$x - \frac{x^3}{6} < \sin x < x - \frac{x^3}{6} + \frac{x^5}{120}, \qquad (6)$$

$$1 - \frac{x^2}{2} < \cos x < 1 - \frac{x^2}{2} + \frac{x^4}{24}. \qquad (7)$$

These inequalities imply $\sin x \cong x - x^3/6$, $\cos s \cong 1 - x^2/2$, and so $\sin x - x\cos x \cong x^3/3$, $x - \sin x \cong x^3/6$ (the

n	p_n	q_n
3	2.59807621	5.19615242
6	**3.**00000000	**3.**46410161
12	**3.1**0582854	**3.**21539030
24	**3.1**3262861	**3.1**5965994
48	**3.1**3935020	**3.14**608621
96	**3.14**103195	**3.14**271459
192	**3.14**145247	**3.14**187304
384	**3.141**55760	**3.141**66274
768	**3.141**58389	**3.141**61017
1,536	**3.14159**046	**3.14159**703
3,072	**3.14159**210	**3.14159**374

Table 1

n	$(q_n - \pi)/(\pi - p_n)$
3	3.78012440
6	**2.**27772383
12	**2.0**6345553
24	**2.0**1552959
48	**2.00**386204
96	**2.000**96424
192	**2.000**24098
384	**2.000**6024
768	**2.0000**150
1,536	**2.00000**376
3,072	**2.000000**94

Table 2

n	$\frac{2}{3}p_n + \frac{1}{3}q_n$
3	**3.**464101615137
6	**3.1**54700538379
12	**3.142**349130544
24	**3.141**639056219
48	**3.14159**5540408
96	**3.141592**83380
192	**3.1415926**64850
384	**3.1415926**54293
768	**3.14159265**3633
1,536	**3.1415926535**92
3,072	**3.141592653589**

Table 3

error of approximation is negligibly small with respect to x^3). Thus,

$$f(x) \cong \frac{x^3/3}{x^3/6} = 2.$$

To prove inequalities (6) and (7), consider the function

$$f_0(x) = -\sin x + x - \frac{x^3}{6} + \frac{x^5}{120}$$

and its iterated derivatives

$$f_1(x) = f_0'(x) = -\cos x + 1 - \frac{x^2}{2} + \frac{x^4}{24},$$
$$f_2(x) = f_1'(x) = \sin x - x + \frac{x^3}{6},$$
$$f_3(x) = f_2'(x) = \cos x - 1 + \frac{x^2}{2},$$
$$f_4(x) = f_3'(x) = -\sin x + x.$$

By the well-known inequality, $\sin x < x$ for $x > 0$. So the derivative $f_4(x)$ of the function $f_3(x)$ is positive for $x > 0$, which means that $f_3(x)$ increases for $x > 0$. But $f_3(0) = 0$, so $f_3(x) > 0$ for $x > 0$. This just happens to be the left inequality of inequality (7). Applying exactly the same argument to function $f_2(x)$ and its derivative $f_3(x)$, we prove $f_2(x) > 0$ for $x > 0$, which is the left inequality of inequality (6). Similarly, $f_1(x) > 0$ and $f_0(x) > 0$, which are the right inequalities of inequalities (7) and (6), respectively, which completes the proof of equation (5).

Equation (5) brings us to the following qualitative conclusion: *the number π, lying in the interval (p_n, $(2/3)p_n + (1/3)q_n$), for all sufficiently large values of n is located much closer to the right end of the interval than to the left end.*

The efficiency of the Huygens formula

It's interesting to compare the two approximate formulas for π:

$$\pi \cong p_n,$$

which might be called the Archimedes formula, and

$$\pi \cong \frac{2}{3} p_n + \frac{1}{3} q_n,$$

which Huygens obtained in the work mentioned above.[1]

The greater efficiency of the second formula can be shown by direct computation. It's really exciting to do this work yourself with your calculator, and if you follow the guidelines on pages 46–47 ("What Is π?"), you'll complete your own tables 1 and 3 in no time at all (in these two tables the correct decimal digits of π are printed in bold).

Another way to compare the efficiency of the Archimedes and Huygens formulas doesn't require any numeric calculations of p_n and q_n at all. One can derive and use so-called *a priori* estimates of the accuracy of the formulas—that is, inequalities that allow one to estimate the differences $\pi - p_n$ and $(2/3)p_n + (1/3)q_n - \pi$ in terms of n in advance, without computing them. Such estimates allow us to plan calculations, and this is their special advantage: from them we can find how big n must be for the desired accuracy to be achieved. Using inequalities (6) and (7) and carrying out simple arithmetic calculations, it's easy to obtain the following *a priori* estimates (you could do it as an exercise):

$$\frac{5}{n^2} < \pi - p_n < \frac{6}{n^2}, \qquad (8)$$
$$\frac{2}{3} p_n + \frac{1}{3} q_n - \pi < \frac{64}{n^4}. \qquad (9)$$

It follows immediately from these inequalities that the rate of convergence of the difference $(2/3)p_n + (1/3)q_n - \pi$ to zero as $n \to \infty$ is greater than that of the difference $\pi - p_n$.

[1]Actually, the work of Archimedes has reached us as a fragment containing three statements. Only one of them can be regarded as an indication of which of the two values—p_{96} or q_{96}—he considered an approximation of π. And this value is q_{96} rather than p_{96}. But historians think that this statement was added to the text later. Besides, it's pretty likely that q_{96} was preferable to p_{96} simply because the Archimedean approximation for the former (22/7) is simpler than for the latter (223/71).—*Ed.*

To compare these rates, let's find the values of n that will ensure an accuracy of π to, say, the nearest hundredth. Since $\pi - p_n < 6/n^2$, to calculate π using the Archimedes formula to the desired accuracy it's enough to take n such that $6/n^2 < 10^{-2}$—that is, $n \geq 25$. Using the Huygens formula, it's enough to choose n such that $64/n^4 < 10^{-2}$—that is, $n \geq 9$.

You can see from this example, as well as from inequalities (8) and (9), that the Huygens formula gives us the desired accuracy more quickly.

You can try to obtain your own formulas for π using polynomial approximations for $\sin x$ and $\cos x$. For example, from inequality (6) one can derive

$$\left| \pi - \frac{4p_{2n} - p_n}{3} \right| < \frac{c}{n^4}$$

with some constant c. It's interesting that this estimate involves only inscribed regular polygons.

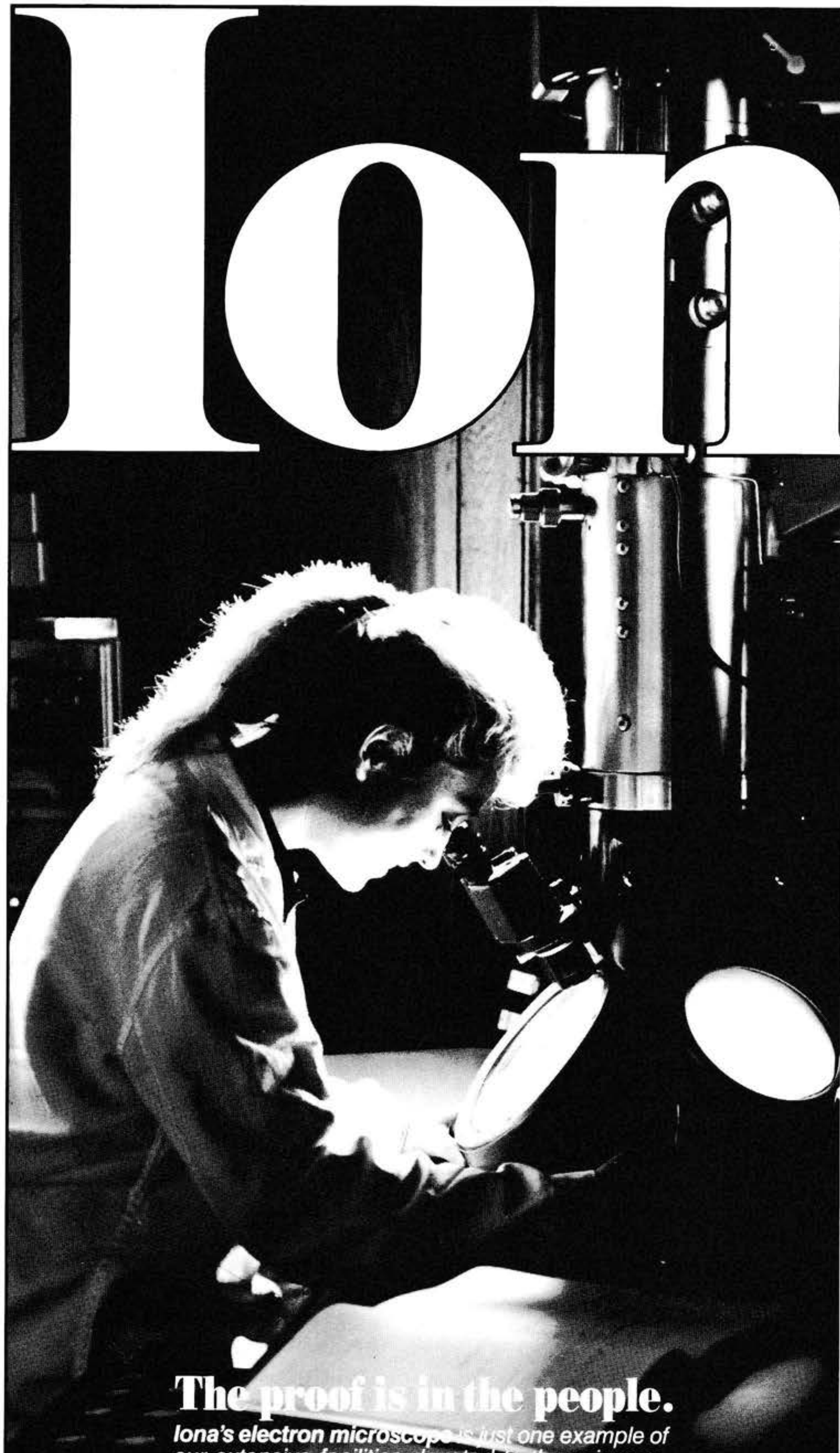
The keen observation made by Huygens—that the circumference is half as far from the perimeter of the inscribed regular n-gon as from that of the circumscribed regular polygon—was generalized not so long ago. It was proven that a similar statement is true for any smooth curve (without corners) bounding a convex figure. More exactly, it reads like this.

Let L be the length of a convex closed curve without corners, P_n the maximum of the perimeters of n-sided polygons inscribed in the curve, and Q_n the minimum of the perimeters of circumscribed n-gons. (Polygons with such extremal perimeters do exist!) Then

$$\lim_{n\to\infty} \frac{Q_n - L}{L - P_n} = 2.$$

The proof of this and many other interesting theorems can be found in the book *Lagerungen in der Ebene, auf der Kugel und im Raum* by L. Fejes Toth (Springer-Verlag, in German).

Iona



Be your best at Iona.

Our "One-on-One Science Program" pairs selected science majors with faculty mentors in important, original research in such studies as water quality, lead poisoning and computer modeling.

Our science and math programs offer 10 majors —from biochemistry to ecology to physics. Iona has also made a major commitment to math and science studies.

Scholarship and financial aid for qualified students are part of Iona's tradition: making quality education affordable.

National honor societies in the sciences and math are active at Iona. And our honors programs recognize and reward outstanding students, encouraging original research and independent study.

For an application call 1-800-231-IONA.

Iona College

715 North Avenue
New Rochelle, NY 10801-1890

Circle No. 9 on Reader Service Card

Double, double toil and trouble

Boiling a witch's brew of toluene and water

by A. Buzdin and V. Sorokin

THE PROCESS OF BOILING is familiar to everyone. The water in a kettle is said to be boiling when an intensive formation of vapor occurs throughout the liquid. Here lies the principal difference between boiling and evaporation: with evaporation the processes of vaporization take place on the open surface only, whereas boiling begins when the saturated vapor pressure in the bubbles becomes equal to the external (atmospheric) pressure.[1] If the saturated vapor pressure is lower than the atmospheric pressure, the bubbles that randomly appear in the liquid collapse and disappear.

Boundary boiling

It's well known that every liquid is characterized by a definite boiling point at a specific atmospheric pressure. For example, water boils at 100°C at normal atmospheric pressure (p_0 = 760 mm-Hg), while the boiling point of toluene (C_7H_8) is 111°C.

What's the boiling point of a "mixture" of toluene and water (two liquids that don't actually mix). You might suppose the boiling point is somewhere between 100°C and 111°C. But if you pour toluene into a test tube containing water (toluene's density is lower, so it will form the upper layer of the "mixture") and heat the tube in a water bath, you'll find that boiling begins at approximately 95°C! (You'll have to use a sufficiently precise thermometer.)

What's going on here? The experiment itself suggests the answer. If you watch carefully, you'll notice that boiling begins at the boundary dividing the two liquids. Here we're dealing with so-called "boundary boiling." In this case the gas bubble appearing at the boundary forms from both water and toluene, and the bubble contains saturated vapor of each liquid. At the boiling point, the vapor pressure in the bubble equals the sum of the partial saturated vapor pressures of toluene $p_t(T_b)$ and water $p_w(T_b)$:

$$p_0 = p_t(T_b) + p_w(T_b),$$

which is Dalton's law. So the saturated vapor pressure of water and toluene (each taken separately) must be lower than the atmospheric pressure, which means that the boundary boiling point T_b is lower than the boiling point of toluene or water.

The following experiment is particularly impressive. Put some carbon tetrachloride (CCl_4), whose boiling point is 76.7°C, in a test tube, and then pour some water on top of it. To make the boundary more distinct, you can color the carbon tetrachloride with an iodine solution beforehand (fig. 1). When the test tube is heated in a water bath, boundary boiling starts at only 65°C. When you do this experiment, make sure you heat the water gradually so you don't cause "bump" boiling of the liquids, which could result in splattering.

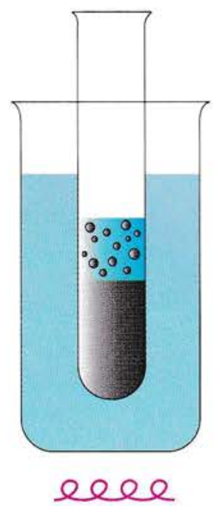


Figure 1

When heating the water bath you should always use a hot plate; never use an open flame (Bunsen burner or alcohol lamp).

Carbon tetrachloride and toluene are among the reagents that can be found in any high school chemistry lab. But you should keep in mind that they're hazardous and must be handled with care. Use small amounts and conduct the experiments in a hood.

It's interesting that the bubbles rise to the surface and burst. The carbon tetrachloride vapor then condenses, and the drops sink.

An experiment for observing boundary boiling that's easy to do involves kerosene poured on top of water in a test tube (the kerosene can be colored with a drop of iodine solution). You can see quite clearly that

[1]Actually, for the bubbles to grow the vapor pressure in them must exceed the sum of the atmospheric pressure, the pressure of the liquid lying above, and the pressure caused by the surface tension of the liquid. In most cases, though, the second and third pressures are much less than the atmospheric pressure and so we can neglect them.

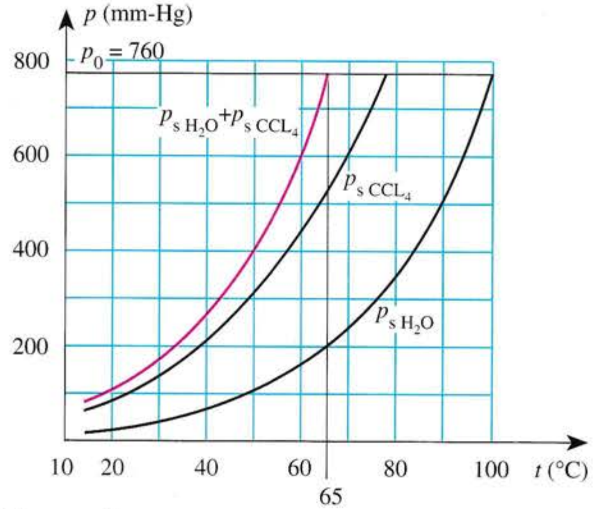


Figure 2

boiling begins at the boundary, but the boundary boiling point is so close to the boiling point of water that you must heat it slowly.

Is it possible to predict the boundary boiling point? It's easy for carbon tetrachloride and water if you happen to have data on the dependence of the saturated vapor pressure of these liquids on temperature.

Figure 2 shows a graph of the temperature dependence of saturated vapor pressure for carbon tetrachloride and water (black lines); the red line shows the sum of the two:

$$p_s(T) = p_{s\,H_2O}(T) + p_{s\,CCl_4}(T).$$

The point where the sum intersects the line $p = p_0 = 760$ mm-Hg gives the boiling point at the boundary of the water and carbon tetrachloride: T_b = 65°C. The vapor pressure of H_2O is 190 mm-Hg, while the vapor pressure of CCl_4 is 570 mm-Hg; the sum equals the normal atmospheric pressure.

Unfortunately a simple calculation like this doesn't always lead to the actual result. Quite often you have to consider the mutual solubility of the components.

It's curious that if you continue the boundary boiling for some time, you'll notice that carbon tetrachloride boils away more quickly than water does. Why? Try to estimate relative rate at which the two liquids boil away (the density ρ of CCl_4 is 1,600 kg/m^3).

Saturated vapor pressure

We're used to determining the saturated vapor pressure of water by just looking it up in a book. But can we obtain the temperature dependence of the saturated vapor pressure ourselves? That is, can we do an experiment and compile our own reference table?

Let's look at one of the simplest methods of obtaining such a relationship. We'll need a beaker, a graduated test tube, a thermometer, a ring stand and clamp, and a hot plate.

Fill the beaker with cold water and place it on the cold hot plate. Partially fill the graduated test tube with water and, turning it upside-down, place it in the beaker. Use the clamp to keep it in place. Put the thermometer in the beaker and clamp it in place as well. Begin heating the beaker. As the temperature rises, record the change in volume of the vapor in the tube.

The vapor in the tube is a mixture of air and saturated water vapor. The total pressure, naturally, equals the atmospheric pressure (we can neglect the pressure of the water column—it amounts to only a few millimeters): $p_0 = p_a(T) + p_s(T)$, where $p_a(T)$ is the air pressure in the tube and $p_s(T)$ is the saturated vapor pressure at temperature T. At the start of the experiment you can neglect the water vapor pressure (at 20°C it equals only 17 mm-Hg). As the temperature increases, the contribution of the water vapor increases.

Using the Clausius–Clapeyron–Mendeleyev equation, we can write out the following statement for the air in the test tube:

$$\frac{p_0 V_0}{T_0} = \frac{p_a V}{T}$$

and

$$p_a = p_0 \frac{V_0}{V}\frac{T}{T_0},$$

where p_0 is the initial pressure (equal to the atmospheric pressure), V_0 and T_0 are the initial volume and initial air temperature, respectively. Now that we know how the volume and temperature change, we can find the saturated vapor pressure at the given temperature:

$$p_s = p_0\left(1 - \frac{T}{T_0}\frac{V_0}{V}\right).$$

This relationship between saturated vapor pressure and temperature that we found experimentally agrees with the reference table. If the temperature is higher than 80°C, the error is less than 5%. Low temperatures cause a deterioration in accuracy.

Why does this happen? Think about what causes the error and try to come up with ways of improving the accuracy of our method.

Art by Sergey Ivanov

Bulletin Board

Results of Twenty-Four Challenge

In our September/October issue, we gave you the opportunity to show off your math skill and ability by playing Twenty-Four®, a game where teams of players must add, subtract, multiply, and/or divide to arrive at the solution of 24. We received more than 30 entries, from classes at all grade levels, and we judged them in two categories: 8th grade and below, and 9th grade and above. Due to space limitations, we were unable to report the outcome in the March/April issue. Here are the results.

8th grade and below. *Grand Prize:* Bernadette Vachetto's 8th grade at Churchville-Chili Middle School in Churchville, New York. *First Runner-Up:* Mr. Kuster's 8th grade at Churchville-Chili Middle School in Churchville, New York. *Second Runner-Up:* Sherry Welch's 8th grade at Gates-Chili Middle School in Rochester, New York. *Third Runner-Up:* Phyllis Perkins's 5th–6th grade at University Elementary School in Bloomington, Indiana. *Fourth Runner-Up:* Paul Larson's 8th grade/periods 2 and 3 at Loyal Public School in Loyal, Wisconsin.

9th grade and above. *Grand Prize:* Jean Kahn's 9th–12th grade at Shoreham-Wading River High School in Shoreham, New York. *First Runner-Up:* Thomas Morrow's 10th–12th grade at East High School in Rochester, New York. *Second Runner-Up:* Mr. Detzel's 10th grade/period 4 at Shaler Area Senior High School in Pittsburgh, Pennsylvania. *Third Runner-Up:* Mrs. Schilstra's 9th grade/period 1 at Penfield High School in Penfield, New York. *Fourth Runner-Up:* Mrs. Schilstra's 9th grade/period 4 at Penfield High School.

Prizes were generously donated by the Eastman Kodak Company–21st Century Learning Challenge. Congratulations and thanks to all who participated.

The Rain Forest Imperative

For years scientists and conservationists have been spreading the word about the plight of tropical rain forests. Now there is an educational study unit and video to help students understand the complex and urgent issues surrounding the rain forest crisis.

"The Rain Forest Imperative," developed by the nonprofit Conservation International and sponsored by McDonald's Corporation, challenges students to learn more about tropical rain forests, their modern day economic and cultural value, and the threat of deforestation. The 25-minute video highlights the Amazonian rain forest in Brazil and is accompanied by a study unit providing classroom activities in life science, geography, and social science curricula. The imperative encourages students to define the many issues contributing to deforestation and identify its real-world solutions.

To order "The Rain Forest Imperative," send $9.95 to McDonald's Education Resource Center, PO Box 8002, St. Charles, IL 60174-8002, or call 800 627-7646.

Quest for the Thinking Computer

Humans will be pitted against computers once again in the second annual Quest for the Thinking Computer. The contest, administered by the Cambridge Center for Behavioral Studies, will be held in November at the Computer Museum in Boston, which features a collection of vintage computers and robots with over 100 hands-on exhibits illustrating the evolution and impact of computers.

Last year's contest drew an audience of more than 200 people. Ten judges conversed at each of eight computer terminals in an attempt to determine which terminals were controlled by people and which by computers. Then they ranked the terminals according to how human-like the conversations were. As expected, the two hidden humans had the highest overall rankings. But, surprisingly, one of the six computers fooled half the judges into thinking it was a person. And three of the judges thought one of the humans was a computer!

The contest was inspired by a paper published in 1950 by the English mathematician Alan Turing, one of the creators of the modern computer. The 1992 contest will be a restricted one, requiring computers to be conversant on only one topic in order to give them a fighting chance of deceiving the judges. The 1992 winner will receive a bronze medal and $2,000. Periodically, an open-ended test will be held. If a computer can fool judges into thinking it's a person in an open-ended test, at least $100,000 will be awarded to the designers of the system, and the contest will be abolished.

Lively debate surrounds the contest: will we have the right to turn off a winning entry? Who should get the prize, the designer or the computer? Could the winning computer be considered "intelligent" or "self-aware" or "conscious"?

Applications for the 1992 competition must be postmarked by July 31. Submissions will be evaluated and screened, and no more than ten finalists will be selected by September 30. A printout of the conversations generated during the 1991 competition is available for a nominal fee; a diskette is also available. For an application and additional information, contact Kathleen Towne at the Cambridge Center for Behavioral Studies, 11 Waterhouse Street, Cambridge, MA 02138, or call 617 491-9020.

Physical simulations on the Mac

Students can now formulate complex motion experiments on the Macintosh and see the results of their experiments in full animation. Interactive Physics™ II, a new software package from Knowledge Revolution, is a complete motion laboratory that simulates and measures objects in motion, driven by physical laws. The software includes a new simulation engine, research-level modeling capabilities, and custom experiment authoring tools.

Interactive Physics lets users draw and build any number of objects on a Macintosh screen (squares, rockets, cars), define motion parameters for each object (mass, elasticity, charge, velocity), set the environment of an experiment (gravity, air resistance, electrostatics), and then immediately "run" the experiment, simulating how the objects would interact in reality. The dynamic simulation engine mathematically creates smooth animationlike simulations, whose measurement data can be displayed simultaneously in graphical, meter, or table format.

Interactive Physics is available from Knowledge Revolution and Macintosh resellers for a suggested retail price of $399. For additional information contact Knowledge Revolution, 15 Brush Place, San Francisco, CA 94103, or call 800 766-6615.

Duracell recognizes young inventors

Six talented high school juniors and seniors with interests in mechanics and electronics are winners in the tenth annual Duracell NSTA Scholarship Competition. To enter, a 9–12th grade student had to design and build a device that is energy efficient, practical, and powered by one or more Duracell batteries. Over 500 inventions were submitted. The devices illustrate the inventiveness and creativity of American high school students today.

The first-place scholarship of $10,000 was awarded to Curt Klaustermeier, a senior at Triad High School in Illinois, for his Battery-Powered Adjustable Wrench. Aided by a gear run from a small battery-powered motor, the wrench opens and closes with ease. Five second-place winners each received a $3,000 scholarship: Sean Burrows, a junior at Shoreham Wading-River High School in New York, invented a visual detection and warning device; Richard Peirce, also a junior and Shoreham Wading-River High School, created radio-controlled life preserver; Glenn Scott Simmonds, a senior at the North Carolina School of Science and Mathematics, designed an emergency system that detects light and noise levels and responds by turning on its own light; William Thomas Chi, a senior at Mission San Jose High School in California, developed a small portable device that neatly administers eye drops; and Daniel Jacob Shapiro, a junior at Beaverton High School in Oregon, devised a portable alarm that safeguards books by sounding an 85-decibel buzzer when tilted. Ten students were given $500 scholarships, and 25 students received $100 cash awards.

To find out how to enter the eleventh annual Duracell NSTA Scholarship Competition, write to Katie Rapp, National Science Teachers Association, 1742 Connecticut Avenue NW, Washington, DC 20009, or call 202 328-5800.

—Compiled by Elisabeth Tobia

Be a factor in the
QUANTUM
equation!

ANSWERS, HINTS & SOLUTIONS

Math

M51

Let's use induction over n. For $n = 3$ the statement to be proved is true: $3^3 = 27$, and 2 is an even number. Now assume that the next-to-last digit of the number $N = 3^n$ is even, and prove that the same is true for $3N = 3^{n+1}$. Let ab be the last two digits of N. Then the two last digits of $3N$ coincide with those of the number $3(10a + b) = 30a + 3b$, and since a is even, the next-to-last digit of $3N$ is of the same parity as the next-to-last digit of $3b$. Write out the first few powers of 3:

$$3^1 = 3,\ 3^2 = 9,\ 3^3 = 27,\ 3^4 = 81.$$

Their last digits are 3, 9, 7, 1, and they'll keep repeating periodically: 3^5 ends in 3, and so on. These are in fact the values taken by b, so $3b$ equals 9, 27, 21, or 3. The next-to-last digits of these numbers are all even (for 9 and 3 they're simply 0), which completes the proof by induction.

M52

Since the number c divides ab, it can be represented as a product uv, where u divides a and w divides b (we can take for u the greatest common divisor of a and c). Let $w = a/u$, $t = b/v$; then $d = ab/c = wt$. It follows that

$$\begin{aligned}&a^{1992} + b^{1992} + c^{1992} + d^{1992}\\ &= (uw)^{1992} + (vt)^{1992} + (uv)^{1992} + (wt)^{1992}\\ &= (u^{1992} + t^{1992})(v^{1992} + w^{1992}),\end{aligned}$$

so the sum in question can be factored.

M53

Denote by O, O_1, O_3 the centers of the given circles and by r, r_1, r_2 their respective radii (fig. 1). If segment AB meets circles O_1 and O_2 at their common point C, then isosceles triangles

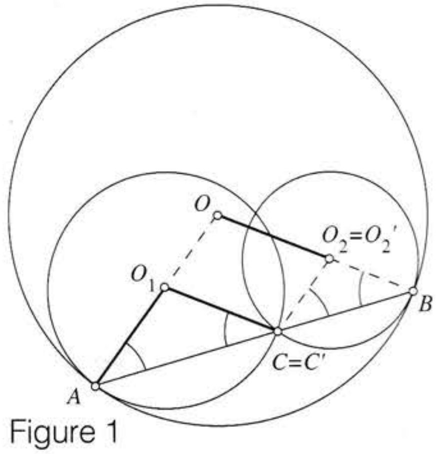


Figure 1

OAB, O_1AC, O_2BC are similar to one another (triangles OAB and O_1AC, say, have a common angle at vertex A). So the opposite sides of quadrilateral OO_1CO_2 are parallel and, therefore, congruent (by a property of the parallelogram). It follows that

$$\begin{aligned}r &= OA = OO_1 + O_1A = O_2C + O_1A\\ &= r_2 + r_1.\end{aligned}$$

The converse is also true: if $r = r_1 + r_2$, then segment AB passes through one of the common points of the smaller circles. To prove it, construct the parallelogram $OO_1C'O_2'$, whose vertices C' and O_2' lie on segments AB and OB, respectively (clearly this can always be done in a unique way). Triangles O_1AC' and $O_2'C'B$ are similar to isosceles triangle OAB, so

$$O_1C' = O_1A = r_1,$$

which means that C' lies on circle O_2, and

$$\begin{aligned}O_2'B &= O_2'C' = OO_1 = OA - O_1A\\ &= r - r_1 = r_2,\end{aligned}$$

which means that $O_2' = O_2$ and that C' lies on circle O_2, which is what we had to prove. (V. Dubrovsky)

M54

The proof is based on the following idea: for any two points $A(a, f(a))$ and $B(b, f(b))$ on the graph of the given function f (fig. 2), the rectangle $r(A, B)$ with diagonal AB covers the segment AB of the graph, its area S satisfying

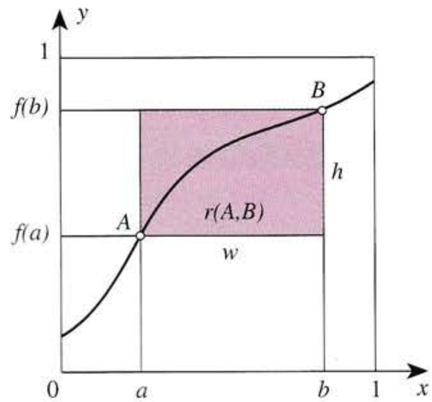


Figure 2

inequality $2\sqrt{S} \le w + h$, where $w = b - a$ is its "width" and $h = f(b) - f(a)$ is its height; this inequality is a direct consequence of the well-known inequality of arithmetic and geometric means: $\sqrt{wh} \le (w + h)/2$. [This inequality follows from the fact that squares are never negative: $(\sqrt{w} + \sqrt{h})^2 = w + h + 2\sqrt{wh} \ge 0$, so $(w + h)/2 \ge \sqrt{wh}$. This holds so long as w and $h \ge 0$.—*Ed.*]

We'll demonstrate two ways of using this idea. Assume for the time being that $f(0) = 0$, $f(1) = 1$.

(1) Extend function f onto the whole positive half-axis so it remains continuous and monotonic (for instance, we can define $f(x) = x$ for $x > 1$). Assign the label A_0 to the initial point (0, 0) of its graph. Take points $A_1, A_2, \ldots, A_n$ on the graph such that the area of each rectangle $r(A_{i-1}, A_i)$, $i = 1, 2, \ldots, n$, equals $1/n^2$. Such points exist because the function is continuous and monotonic and grows indefinitely. (When point P moves along the graph, starting from A_{i-1}, the area of $r(A_{i-1}, P)$ grows indefinitely, taking on all intermediate values, $1/n^2$ among them.) For each of these rectangles, the sum of the width and height is not

less than $2\sqrt{1/n^2} = 2/n$, so the sum of the coordinates of point A_n, which are equal to the sums of all widths and all heights, is not less than 2. Therefore, at least one of these coordinates is not less than 1, which means that A_n lies on the extension of the graph and the rectangles cover the graph.

(2) Let's choose the points B_0, B_1, ..., B_n on the graph, starting with $B_0(0, 0)$, such that the sum of the width and height of each rectangle $r(B_{i-1}, B_i)$, $i = 1, 2, \ldots, n$ is $2/n$. (Point B_i is simply the point where the graph intersects the line $y = 2i/n - x$; in particular, B_n is the endpoint (1, 1) of the graph—see figure 3.) The area of each of the rectangles is not greater than $[(w + h)/2]^2 = 1/n^2$. Dilating the rectangles, if necessary, to make their areas exactly $1/n^2$, we'll get the required covering.

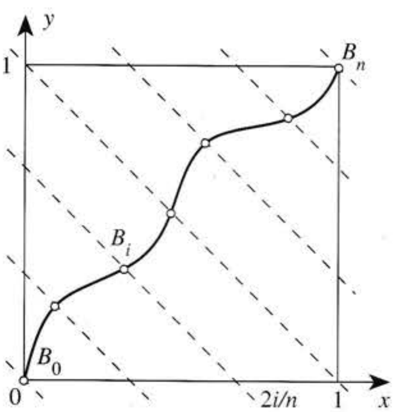


Figure 3

If $f(0) > 0$ or $f(1) < 1$, the graph must be completed by vertical segments to form a continuous curve joining points (0, 0) and (1, 1) (fig. 4). Similarly, a curve can be made out of the graph of any monotonic function (not necessarily continuous—see figure 5).

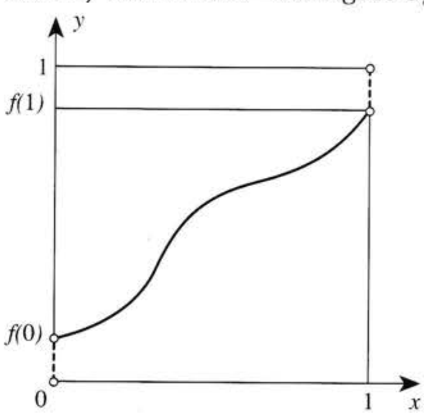


Figure 4

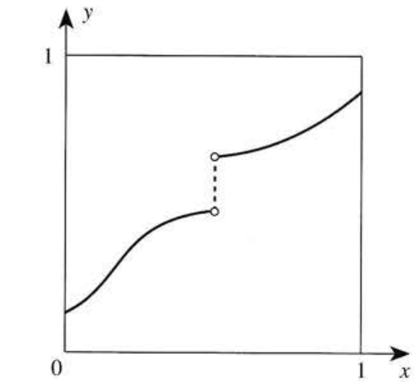


Figure 5

Both arguments work with these curves as well. (N. Vasilyev, V. Dubrovsky)

M55

Consider two arbitrary planets. Let P and Q be their centers. Point X on the planet centered at P and point Y on the other will be called *corresponding* if the rays PX and QY have the same direction (fig. 6). The regions on the two planets that are not visible from the other planet are obviously two "exterior" hemispheres cut on the planets by the planes through P and Q perpendicular to line PQ; they're shaded in figure 6. Of any two

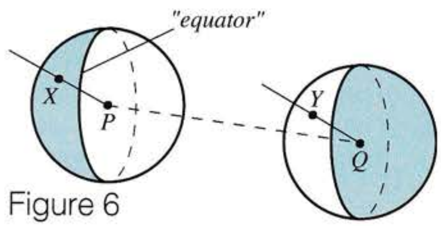


Figure 6

corresponding points on our two planets (except points on the "equators" of the hemispheres), only one can belong to the respective nonvisible hemisphere. As to the points of the "equators," it's a matter of convention whether such a point is visible from the other planet (that is, from the corresponding point on the other planet's equator); but since there are only a finite number of such equators for all pairs of planets, their total area is zero, so we can and will simply disregard them.

Now, take an auxiliary unit sphere S and denote by I_P the set of its points corresponding to the points on planet P that are not visible from any other

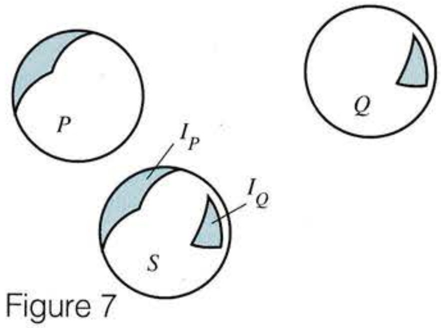


Figure 7

planet (fig. 7). It suffices to show that the sets I_P for all planets cover sphere S without gaps and overlaps; or that, except for the boundary points of regions I_P (which are disregarded), every point X of sphere S belongs to one and only one set I_P.

If X belongs simultaneously to two sets I_P and I_Q, then the points on planets P and Q corresponding to X are each out of sight from the other planet, which is impossible, as we've seen above. So there are no overlaps.

Finally, take an arbitrary point X of S. Suppose for convenience that the radius OX is pointing upward (fig. 8).

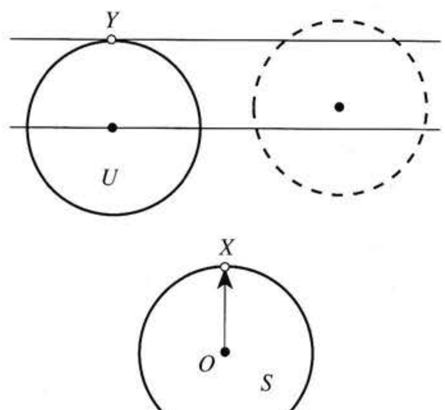


Figure 8

Let U be the uppermost planet; then point Y on this planet corresponding to X is completely out of sight—otherwise, the planet from which it is seen would be above U. So X belongs to I_U, and the covering of sphere S has no gaps. (V. Dubrovsky)

Physics

P51

After the horizontal velocity $\mathbf{v}$ is imparted to the upper ball, the balls will revolve about the center of mass of the system, which in turn moves forward with speed $\mathbf{v}/2$. If the upper ball immediately loses contact with the table, the only force acting on the system is the force of gravity. It imparts acceleration $\mathbf{g}$ to the system and therefore to the lower ball. For the lower ball to lose contact with the table, its centripetal acceleration must be greater than $\mathbf{g}$.

In the coordinate system moving with the velocity of the center of mass, both balls have velocities $v/2$ that are equal in magnitude, and the centripetal acceleration equals

$$a_c = \frac{(v/2)^2}{l/2} = \frac{v^2}{2l}.$$

The lower ball will lose contact with the table if

$$\frac{v^2}{2l} \geq g$$

—that is,

$$l \leq \frac{v^2}{2g}.$$

P52

The simplest way to solve this problem is to use Newton's second law. In a time Δt a length of cable $\Delta l = v\Delta t$ is put into motion. If the mass per unit length of cable is denoted by μ, the mass of the length of cable Δl is equal to $\Delta m = \mu\Delta l = \mu v\Delta t$. This means that a change in momentum $\Delta mv = \mu v^2\Delta t$ is given to the mass Δm in time Δt. The impulse that causes this change in momentum is due to the difference between the force of gravity acting on the left and right portions of the cable. This difference is equal to mgh. According to Newton's second law we can write

$$\mu v^2\Delta t = \mu gh\Delta t.$$

From this we get

$$v = \sqrt{gh}.$$

P53

Considering that the gas density is $\rho = M/V$, we can write the equation of state for water vapor in the form $p = (\rho/\mu)RT$, where ρ and μ are the density and molar mass of water vapor. Boiling takes place when the saturated vapor pressure becomes equal to the atmospheric pressure. If the boiling point of the salted water has been raised and the atmospheric pressure is constant, the density of the saturated water vapor must decrease.

P54

The thermal power generated when the current passes through the resistor is partially spent on heating the resistor and partially released to the surroundings. In the state of thermal equilibrium (when the resistor's temperature remains constant) all the power is released to the surroundings.

Let's denote the proportionality factor between the power generated in the resistor and the difference between the temperatures of the resistor and the surrounding air by α. At a temperature $t_3 = 80°C$ the resistance equals $R_1 = 50\ \Omega$ (see figure 9). Then from the equality

$$\frac{V_1^2}{R_1} = \alpha(t_3 - t_0)$$

we get

$$\alpha = \frac{V_1^2}{R_1(t_3 - t_0)} = 1.2\frac{\text{V}}{\Omega \cdot \text{K}}.$$

The spontaneous current oscillations with the voltage $V_2 = 80$ V across the resistor are caused by the oscillations of its resistance. When the temperature of the resistor reaches $t_1 = 100°C$, its resistance jumps abruptly from $R_1 = 50\ \Omega$ to $R_2 = 100\ \Omega$. As a result, the thermal power generated in the resistor decreases, and the resistor begins to cool because the rate at which energy is being generated in the resistor is less than the rate at which energy is released to the surroundings. When the resistor's temperature falls to $t_2 = 99°C$, its resistance changes abruptly from $100\ \Omega$ back to $50\ \Omega$. The thermal power dissipated in the resistor increases, and the resistor heats up again because the rate of energy generation is now greater than the rate of energy release. At a temperature $t_1 = 100°C$ the resistance jumps again, and the entire process is repeated.

The period of the oscillations in the circuit equals

$$T = \tau_1 + \tau_2,$$

where τ_1 is the time it takes the resistor's temperature to increase from t_2 to t_1, and τ_2 is the time it takes the resistor to cool from t_1 to t_2. Let's write the corresponding equations for the heat balance:

$$\frac{V_2^2}{R_1}\tau_1 = \tau_1\alpha(t_1 - t_0) + C(t_1 - t_2),$$

$$\frac{V_2^2}{R_2}\tau_2 = \tau_2\alpha(t_1 - t_0) - C(t_1 - t_2)$$

(since the relative change in the resistor's temperature is small, we can assume that the thermal power released to the surroundings is constant and equals $\alpha(t_1 - t_0)$). Substituting the given numerical values, we find

$$\tau_1 = \tau_2 \cong 0.1 \text{ s} \Rightarrow T \cong 0.2 \text{ s}.$$

The maximum value of the current is obviously equal to

$$I_{max} = \frac{V_2}{R_1} = 1.6 \text{ A},$$

and the minimum value is

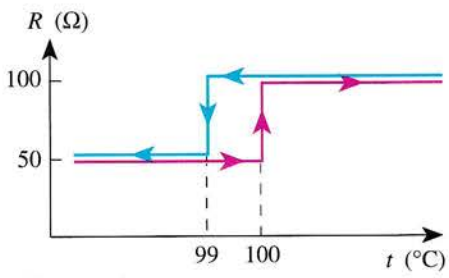


Figure 9

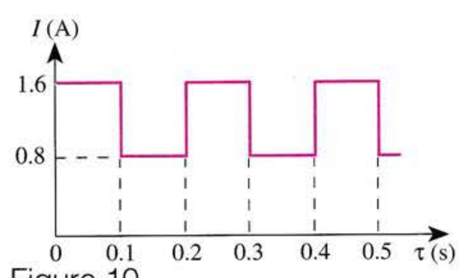


Figure 10

$$I_{\min} = \frac{V_2}{R_2} = 0.8 \text{ A}.$$

Figure 10 shows the graph of the time dependence of current in the circuit.

P55

A camera obscura is simply a box with a pinhole aperture. A broad beam of light rays reflect from every point of the object and strike the photographic film inside the camera (fig. 11). For the tiger to appear striped in

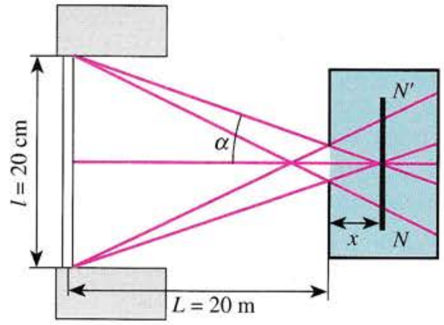


Figure 11

the photograph, the beams coming from the nearest points of adjacent stripes on the tiger must not overlap on the film. Therefore, the film must be a distance x from the aperture that is greater than $d/\tan\alpha$, where d is the diameter of the aperture.

Since

$$\tan\alpha = \frac{l/2}{L+x},$$

then

$$x > \frac{2d(L+x)}{l}.$$

From this we get

$$x > \frac{2dL}{l\left(1-\frac{2d}{l}\right)}.$$

Since $2d/l << 1$, we can neglect $2d/l$ as compared to 1 in the denominator of the formula. We then get

$$x > \frac{2dL}{l} = 20 \text{ cm}.$$

So the camera must measure at least 20 cm from front to back.

Brainteasers

B51

The area of a map with the scale $1 : k$ is $1/k^2$ that of the territory it represents. The number of people that can fit in a certain region is proportional to its area. So if this region is as big as the map in the problem, its "population" will be about $(2 \cdot 10^8)/(25 \cdot 10^{12}) = 0.000008$ person.

B52

Let the given number be $n = 10a + b$ (b is the last digit). Then the described operation yields $m = a + 4b$. Since $4n - m = 39a$ is a multiple of 13, the divisibility by 13 of the number m is equivalent to that of the number $4n$ and, therefore, of n.

B53

Twilight is the period of semidarkness between sunset and nightfall. The daylight at this time of day is the result of sunlight being dispersed by the atmosphere at high altitudes, where the Sun has not yet set behind the horizon. High up in the mountains the atmospheric layer over the Earth is thinner and the period of its illumination after sunset is shorter. So twilight in the mountains is shorter too. (Planets devoid of atmosphere have no twilight at all, nor does the Moon.)

B54

In the notation of figure 12 the area of triangle 1 is equal to $(1/2)ac \sin(180° - \alpha) = (1/2)ac \sin\alpha$, which is the area of the shaded triangle. Similarly, one can show that the other two triangles are also equal in area to the shaded one. (Actually, triangle 3 is even congruent to it.) (V. Dubrovsky)

B55

If x is the unknown number of university students, then each of $x + 2$ participants in the championship played $x + 1$ games, so their total score, equal to the total number of games, is $(x + 2)(x + 1)/2$. The total score of the x university students equals $(x + 2)(x + 1)/2 - 6.5 = (1/2)(x^2 + 3x - 11)$ and is equally distributed among them: each of them got $(1/2)(x + 3 - 11/x)$ points. This must be an integer number of half points, so 11 is divisible by x. If follows that $x = 11$ (in the case of $x = 1$, the score of each university student would be negative).

Go "mod"

1. No, it can't. If the square of an integer ends in 5, then this integer has the form $10n + 5$, so the square equals $(10n + 5)^2 = 100(n^2 + n) + 25$, and its second last digit is 2, not 7.

2. (a) The quotient $q = 113$, the remainder $r = 10$; (b) $q = -10$, $r = 15$; (c) $q = 11$, $r = 73$.

3. $(17x + 3y) \cdot 22 = 8x + 6 \cdot 61x + 5y + 61y = (8x + 5y) + 61(6x + y)$, so $17x + 3y$ and $8x + 5y$ are both divisible by 61.

4. (a) 1 and 2; (b) $n^2 + n + 1 = n(n + 1) + 1$, remainder = 1; $n^2 + n + 1 = (n - 1)(n + 2) + 3$, remainder = 3; (c) $n^4 + 1 = (n^3 - 3n^2 + 9n - 27)(n + 3) + 82$, $n + 3 > 82$, remainder = 82.

5. (a) $(n^2 + 1)/(n - 1) = n + 1 + 2/(n - 1)$ is a whole number when $n = 3, 2, 0$, or -1; (b) $n = 1, 0, -1$, or -3.

6. (a) $a^3 - a = (a - 1)a(a + 1)$ is divisible by 2 and by 3 (therefore, by 6) as a product of three consecutive numbers, so a^3 and a have equal remainders modulo 6. (b) $a^5 - a =$

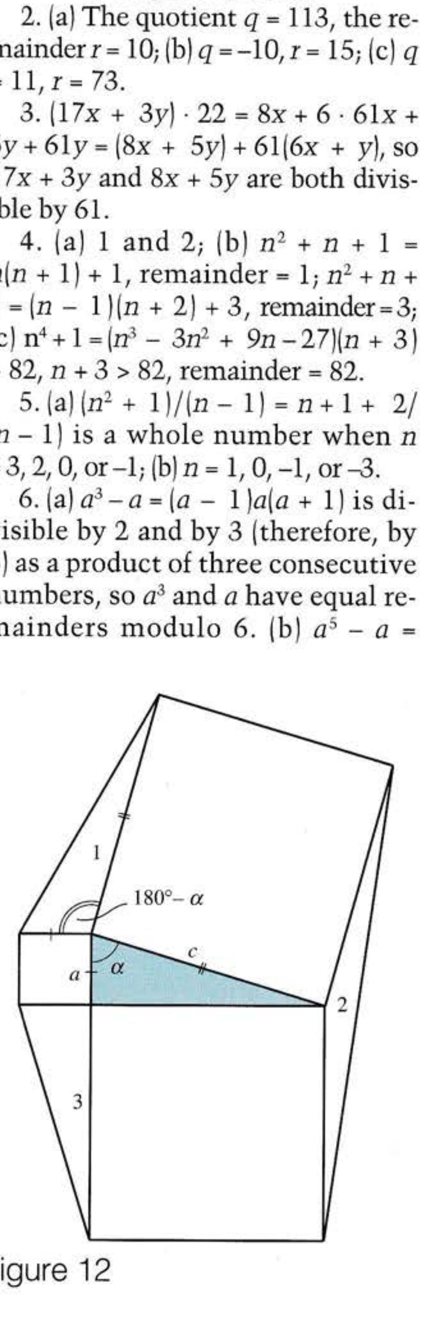


Figure 12

$a(a^4-1) = a(a^2-1)(a^2+1) = (a-1)a(a+1)(a^2+1)$, and $a^2+1 \equiv a^2-4 = (a-2)(a+2) \pmod 5$; one of the 5 consecutive numbers $a-2$, $a-1$, ..., $a+2$ is always divisible by 5.

7. $2^{100} = 4^{50}$, $3^{100} = 9^{50} = (5+4)^{50}$, and the last number is obviously congruent to 4^{50} modulo 5; $3^{100} = (5-2)^{100} = 5N + 2^{100}$ for some integer N; $3^5 = 243 = 211 + 32 = 211 + 2^5$, so $3^{100} = (211 + 2^5)^{20} \equiv 2^{100} \pmod{211}$.

8. $11^{10} - 1 = (11-1) \cdot (11^9 + 11^8 + \ldots + 11 + 1)$. The first factor is 10; the second factor is the sum of 10 numbers each ending in 1, so it ends in 0 and is divisible by 10 too, which means that the product is divisible by 100.

9. For $N = \lim_{n\to\infty} p_{3\cdot 2^n}$, we have $N - S(N) = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \ldots + a_0 - a_n + a_{n-1} + \ldots + a_0) = a_n(10^n - 1) + a_{n-1}(10^{n-1}-1) + \ldots + a_1 \cdot (10-1)$, but $10^k - 1 = 99\ldots9$ (k nines) and is divisible by 9 (and by 3).

10. By the previous exercise, $S(5A) - S(A) \equiv 5A - A = 4A \pmod 9$. So if $S(5A)$ and $S(A)$ yield the same remainders when divided by 9, then 9 divides $4A$ and, therefore, A.

11. (a) By exercise 10, if N is any number in question, then $N \equiv S(N) = 20 \equiv 2 \pmod 3$. But the square of any number either is divisible by 3 or has a remainder of 1 when divided by 3, because $(3k \pm 1)^2 = 3(3k^2 \pm 2k) + 1$. So N is not a square.

(b) The answer is again no. In this case the sum of digits $1 + 9 + 9 + 2$ is divisible by 3 but not divisible by 9, which is impossible for the square of an integer.

12. Follow the solution to exercise 9 and use the divisibility of $10^{2k} - 1 = 11 \cdot 9090\ldots9$ (k nines) and $10^{2k+1} + 1 = 10 \cdot (10^{2k} - 1) + 11$ by 11.

15. (a) $2^{1992} - 1 = 16^{498} - 1 \equiv (-1)^{498} - 1 \equiv 0 \pmod{17}$. (b) Since $3^3 = 27 \equiv 1 \pmod{13}$, $(3^{20} + 11)^{55} \equiv [(3^3)^6 \cdot 3^2 - 2]^{55} = 7^{55} = 49^{27} \cdot 7 \equiv 10^{27} \cdot 7 \equiv (-3)^{27} \cdot 7 \equiv -7 \equiv 6 \pmod{13}$.

16. (a) $2^{50} + 1 = (2^{10})^5 + 1 = 1{,}024^5 + 1 \equiv 24^5 + 1 = (24+1)(24^4 - 24^3 + 24^2 - 24 + 1) \pmod{125}$. The first factor is 25; the second factor divisible by 5, since $24 \equiv -1 \pmod 5$; so the product is divisible by 125.

(b) Note that $105 = 3 \cdot 5 \cdot 7$. Now, $2^{48} - 1$ $(-1)^{48} - 1 \equiv 0 \pmod 3$, $2^{48} - 1 = 4^{24} - 1 \equiv (-1)^{24} - 1 \equiv 0 \pmod 5$, and $2^{48} - 1 = 8^{16} - 1 \equiv 1^{16} - 1 \equiv 0 \pmod 7$, so $2^{48} - 1$ is divisible by 3, 5, and 7.

(c) Use induction over $n \geq 0$. As the equation

$$2^{3^n} + 1 = \left(2^{3^{n-1}} + 1\right)\left(2^{3^{n-1}\cdot 2} - 2^{3^{n-1}} + 1\right)$$

shows, the crucial point is to prove that the second factor on the right side is divisible by 3 but not by 9. Find its remainders:

$$\begin{aligned} 2^{3^{n-1}\cdot 2} - 2^{3^{n-1}} + 1 &\equiv (-1)^{3^{n-1}\cdot 2} - (-1)^{3^{n-1}} + 1 \\ &\equiv 3 \equiv 0 \pmod 3; \end{aligned}$$

$$\begin{aligned} 2^{3^{n-1}\cdot 2} - 2^{3^{n-1}} + 1 &\equiv 8^{3^{n-2}\cdot 2} - 8^{3^{n-2}} + 1 \\ &\equiv 1 - (-1) + 1 \equiv 3 \pmod 9. \end{aligned}$$

17. $p = 3$, since for any integer n not divisible by 3, $20n^2 + 1$ is divisible by 3.

18. For $n = 2k+1$, $1^m + 2^m + \ldots + (n-1)^m = [1^m + (n-1)^m] + [2^m + (n-2)^m] + \ldots + [k^m + (n-k)^m] \equiv [1^m + (-1)^m] + [2^m + (-2)^m] + \ldots + [k^m + (-k)^m] \equiv 0 \pmod m$.

19. The number n must be even, which follows from the following relations: $323 = 17 \cdot 19$, $20^n + 16^n - 3^n - 1 \equiv 3^n + (-1)^n - 3^n - 1^n \pmod{17}$; $20^n + 16^n - 3^n - 1 \equiv 1^n + (-3)^n - 3^n - 1 \pmod{19}$.

20. $5^{2n+1} + 3^{n+2} \cdot 2^{n-1} = 25^n \cdot 5 + 6^{n-1} \cdot 3^3 \equiv 6^{n-1}(6 \cdot 5 + 27) \equiv 0 \pmod{19}$.

21. Any common divisor of $15n + 2$ and $14n + 3$ also divides $15n + 2 - (14n + 3) = n - 1$ and $14n + 3 - 14(n-1) = 17$; so $n - 1$ must be divisible by 17, or $n = 1 + 17k$. The fraction in question can be simplified for any such n.

22. $p = 3$ (consider remainders modulo 3).

23. 5, 11, 17, 23, 29. Any 5 successive terms of the sequence in question have different remainders when divided by 5; therefore, one of them must be divisible by 5. Since it is prime, it must be equal to 5.

24. By theorem 1, the difference d can't be coprime with $m = 15$ or with any smaller number. So $d \geq 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 = 30{,}030$.

25. (a) Any number dividing a and b must also divide $a - b$, so a prime common divisor of any two of our 10 consecutive integers divides a number not greater than 9 and therefore can be equal only to 2, 3, 5, or 7. Five of the given integers are odd; only two of them can be divisible by 3; one can be divisible by 5, and one by 7. This leaves at least one odd number not divisible by any of the numbers 3, 5, 74,34—and this number is coprime with all the rest.

(b) The proof is similar to that in part (a) above but needs more subtle reasoning. Let the first of the numbers in question be even, and let $a_1, a_2, \ldots, a_8$ be the odd numbers in increasing order. Possible common divisors are 2, 3, 5, 7, 11, 13. The number 3 can divide at most three of the numbers $a_1, a_2, \ldots, a_8$; 5 and 7 can each divide two of them at most; 11 and 13 can each divide one at most (because, for instance, if 11 divides two numbers, a_m and a_n, then $|a_m - a_n| \geq 22$).

But none of the numbers a_3, a_4, a_5 can have 11 or 13 as a common divisor with any other of our 16 numbers, since $a_5 - 11 < a_1 - 1$, $a_3 + 11 > a_8$. Suppose none of the 16 numbers is coprime with the rest. Then 3 must divide one of the numbers a_3, a_4, a_5; 5 must divide another one; and 7 must divide the third. Therefore, 7 divides only one of the numbers $a_1, a_2, \ldots, a_8$. Now, the best we can do is to assume that 5 divides two numbers—a_3 and a_8; and 3 divides three other numbers—a_1, a_4, a_7; then 7 divides a_5. But 13 can divide neither a_2 ($a_2 + 13 > a_8$) nor a_6 ($a_6 - 13 < a_1 - 1$), so one of these two numbers is coprime with all the rest. The argument is almost the same if the first 16 numbers are odd.

(c) Each of 17 numbers 2,184, 2,185, ..., 2,200 has a common divisor with any other.

26. Use theorem 3.

27. No.

28. $-1 + 2 \cdot 3 \cdot 5 \cdot 7 = 209$.

29. 788.

30. (a) $x = 12 + 37k$, (b) $x = 17 + 29k$ for any integer k.

31. (a) $x = -1 + 8k$, $y = 1 - 7k$; (b) $x = 7 + 15k$, $y = 5 + 13k$; (c) $x = -1 + 18k$, $y = 24 - 257k$.

32. $x = -17 + 118k$, $y = 9 - 61k$, $z = -1 + 7k$.

Tartu '91

Here are the answers to the problems posed in the article "Tartu in the Summer of '91" in the last issue. The answers (as well as the questions) were prepared by Ilya Itenberg and Dmitry Fomin, who were kind enough to share them with *Quantum*'s readers.

1. Refer to figure 13. If $\angle BAL = \angle CDK$, then quadrilateral $AKLD$ can be inscribed in a circle and $\angle ADL + \angle AKL = 180°$. Since $\angle ADL + \angle BCL = 180°$ (lines AD and BC are parallel), we must have $\angle AKL = \angle BCL$. Similarly, $\angle BKL = \angle ADL$, and $\angle BKL + \angle BCL = 180°$. This means that quadrangle $BCLK$ is also cyclic. Thus we have $\angle CKD = 180° - \angle AKD - \angle BKC = 180° - \angle ALD - \angle BLC = \angle BLA$.

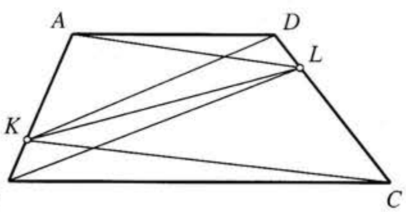


Figure 13

2. The sum of these numbers is zero, so they cannot be positive simultaneously.

3. The given algebraic condition on F implies that if $A = F(x)$ for some real x (that is, if A is in the range of F), then $F(A) = 1/A$. We'll prove now that there exists real x such that $F(x) = 500$. It then follows that $F(500) = 1/500$.

We know that $F(1{,}000) = 999$. Therefore $F(999) = 1/999$. The function F is continuous, and 999 and 1/999 are in its range. So (by the Intermediate Value Theorem of analysis) the number 500 is in its range as well, and $F(500) = 1/500$.

4. The beginning of the game is determined: $2 \to 3 \to 4$. Let's suppose that the second player has a winning strategy. If the first player replaces 4 with 5, the second player must replace 5 with 6. A winning strategy must therefore involve leaving the number 6 on the blackboard. But the first player can arrange to achieve this position by replacing the 4 with 6 instead of 5. Thus the second player cannot have a winning strategy. Since the numbers on the blackboard keep increasing, someone must win. So there must exist a winning strategy for the first player.

5. If x and y are positive real numbers, then it isn't difficult to show that $1/x + 1/y \ge 4/(x + y)$. Indeed, $(x - y)^2 \ge 0$ (the square of a real number is never negative), so

$$\begin{aligned} x^2 - 2xy + y^2 &\ge 0, \\ x^2 + y^2 &\ge 2xy, \\ x^2 + 2xy + y^2 &\ge 4xy, \\ (x+y)^2 &\ge 4xy, \\ \frac{(x+y)^2}{xy(x+y)} &\ge \frac{4xy}{xy(x+y)}, \\ \frac{x+y}{xy} = \frac{1}{x} + \frac{1}{y} &\ge \frac{4}{x+y}. \end{aligned}$$

So we have

$$\begin{aligned} \frac{1}{A} + \frac{1}{B} + \frac{4}{C} + \frac{16}{D} &\ge \frac{4}{A+B} + \frac{4}{C} + \frac{16}{D} \\ &\ge \frac{16}{A+B+C} + \frac{16}{D} \\ &\ge \frac{64}{A+B+C+D}. \end{aligned}$$

6. We can reformulate the question as follows: There exist nonnegative integers x, y, z, u, v, w, t such that

$$\begin{aligned} A &= x + 2y + 5z + 10u + 20v \\ &\quad + 50w + 100t, \\ B &= x + y + z + u + v + w + t. \end{aligned}$$

Then

$$\begin{aligned} 100B &= 100t + 100w + 100v \\ &\quad + 100u + 100z + 100y + 100x, \end{aligned}$$

or

$$\begin{aligned} 100B &= 1(100t) + 2(50w) + 5(20v) \\ &\quad + 10(10u) + 20(5z) \\ &\quad + 50(2y) + 100x, \end{aligned}$$

so it's possible to take $100t$ one-cent coins, $50w$ two-cent coins, $20v$ five-cent coins, ..., and x dollars, which is a set of A coins worth B dollars.

7. Let's denote the given 70,000,000-digit number by A and the sum of all natural numbers having at most 7 digits by B. Direct division shows that the number 9,999,999 is divisible by 239. We use this fact to show that A is congruent to B modulo 239. Let's mentally dissect the decimal representation of A into ten million septuples, representing all 7-digit numbers: $a_1, a_2, a_3, \ldots, a_{10{,}000{,}000}$. Then we have

$$\begin{aligned} A &= a_1 + (10^7)a_2 + (10^{14})a_3 + \ldots \\ &\quad + (10^{7\cdot 9999999})a_{10{,}000{,}000}. \end{aligned}$$

Since 10^7 is congruent to 1 modulo 239, we know that 10^{7k} is also congruent to 1 for every natural k. So A is congruent to $a_1 + a_2 + a_3 + \ldots + a_{10{,}000{,}000} = B$.

Now, the number B is divisible by 239, because all natural numbers less than 9,999,999 can be divided into pairs with the sum of each pair equal to 9,999,999. So A is also divisible by 239.

8. We fix an initial city A and consider all cities $B_1, B_2, \ldots, B_k$ where the roads starting from A end. These cities, together with the roads from A to each B_i, form a system that we'll call S.

Now consider the cities $C_1, C_2, \ldots, C_m$ (each of which differ from the cities already in S), which are endpoints of the roads starting from $B_1, B_2, \ldots, B_k$. Let's add these to the system S, together with the roads from each B_i to each C_j: one new road for each city added. Continuing in this way, we'll end by including all N cities in system S, together with $N - 1$ roads: one road ending at each city except for A.

Similarly, we can construct another system T of cities and roads by traveling "backwards." Starting at the same initial city A, we consider all cities $B'_1, B'_2, \ldots, B'_p$ from which a road leads to city A. These cities, together with the roads from each B'_i to A, form a system that we'll call T.

Now consider the cities $C'_1, C'_2, \ldots, C'_m$, (each of which differs from the cities already in T) from which a road leads to one of the cities B'_p. We add these to the system T, together with the roads from each C_i to each B_j: one new road for each city added. Continuing in this way, we'll end by including all N cities in system T, together with $N - 1$ roads: one road ending at each city except for A.

Now we take the union of S and T. This union contains all N cities, and we can easily get from one city to another, staying on roads in S and T—for example, by passing through city A. The union of S and T contains at most $2N - 2$ roads (some of the roads in S may also be in T). So there is at least one road not included in the union of S and T, and this road can be closed without affecting the given property.

9. Let's consider first $2^n + 1$ natural numbers. We assign to each of them—say, P—a certain string of length n consisting of ones and zeroes defined as follows: if P belongs to A_k, then the kth symbol in the string is equal to 1; otherwise, the kth symbol will be 0.

Now there are 2^n possible strings of length n, each of whose symbols is either 0 or 1, and we are looking at a collection of $2^n + 1$ of these. Therefore, the pigeonhole principle assures us that there exist two natural numbers X and Y such that corresponding strings are equal. These numbers are the desired ones. This simple argument completes the proof.

10. The basic idea of the solution is the common but important idea of an *invariant*, which appears in both mathematics and science.

Let's define a quantity S as the sum of the number of piles and the number of stones. At each move we decrease the number of stones by one, but we increase the number of piles by one. So S remains invariant, no matter what moves we may make.

The initial value of S is 1,002. Suppose it were possible, after a number of moves, to obtain exactly n piles, each containing exactly 3 stones. Then there would be $3n$ stones altogether, and S would be equal to $n + 3n$, or $4n$. But the number 1,002 cannot be represented as $4n$ (for a natural number n), since it is not a multiple of 4.

This contradiction shows it is impossible to end up with piles consisting of exactly three stones each.

11. You can trust him. In fact, we'll construct such an orchard. In this solution, we'll refer to a distance of ten meters as a "unit."

We first pick a point for the position of an apple tree, then plant ten pear trees around a unit circle centered at the apple tree and located at the vertices of a regular decagon. We now construct the required orchard in eleven steps. At each step, we'll create several copies of the orchard at the previous step, then "erase" the previous step.

As a second step, we draw ten unit vectors, each directed along one of the sides of the decagon in the first orchard. We then translate the ten pear trees, and the apple tree, ten times: once along each unit vector. Finally, we erase the original decagon and the apple tree at its center.

What is the result of this operation? The decagon has ten images, making 100 pear trees. The original apple tree also has ten images, making 10 apple trees. But there are in fact more apple trees to be planted. To see this, we follow the ten images of one of the pear trees from step 1. Each image is a pear tree that is one unit from the original tree, in a direction parallel to the decagon of step 1. Thus the images all lie on a circle of unit radius, and we can plant an apple tree at the center of this circle—this (step-2) apple tree is planted at the spot where a (step-1) pear tree originally stood. In this way, we have in fact 20 apple trees, rather than 10, in the step-2 orchard.

We now take step 3 in exactly the same way we took step 2. The step-2 orchard can be viewed as ten decagons of pear trees, the images of the original (step-1) decagon—plus some all-important apple trees, which we'll ignore for the time being. We select one decagon, draw ten unit vectors along its sides, and translate the step-2 orchard ten times. We then erase the step-2 configuration. What remains is ten copies of the 100 step-2 pear trees, making 1,000 step-3 pear trees. Let's count the apple trees in step 3. Certainly there are 10 copies of each of the 20 step-2 apple trees, making 200 apples trees for step 3. But we can look at the set of images of each step-2 pear tree. These form a decagon of ten pear trees, each one unit from where the step-2 pear tree stood. We can therefore replant this spot with a (step-3) apple tree. Since there were 100 step-2 pear trees, we must plant 100 new (step-3) apple trees, making 300 apple trees in all.

We continue deriving step 4 from step 3, step 5 from step 4, and so on, keeping track of the number of apple trees and the number of pear trees at each step:

Step	Apple trees	Pear trees
1	1	10
2	20	100
3	300	1,000
4	$4 \cdot 10^3$	10^4
.	.	.
.	.	.
.	.	.
10	$10 \cdot 10^9$	10^{10}
11	$11 \cdot 10^{10}$	10^{11}

At step 10 we have equal numbers of apple and pear trees, while at step 11 the number of apple trees overtakes the number of pear trees.

There's one case that's an exception to this process. It may happen that at a given step, two of the new pear trees accidentally coincide. It would be very difficult to predict this collision geometrically, but it doesn't matter. The construction doesn't depend on the exact direction of the translation vectors, only on the fact that they are of unit length and that there are ten images for each pear tree. Thus, if any two vectors take two pear trees from step n onto the same spot in step $n + 1$, simply adjust the direction of one of the vectors by a degree or two so that the new pear trees don't coincide. This is always possible, since there are only a finite number of ways two pear trees can coincide. If we make one translated copy of the step-n orchard through this new direction, we can continue the construction as described.

12. (a) Let's label all the centers of these squares and look at the vertical lines on the chessboard. The centers of the squares, being lattice points, lie on these vertical lines. Since there are 9 such lines and 55 centers, the pigeonhole principle assures us that we can find 7 centers on the same vertical line (since $55 > 54 = 9 \times 6$).

So on this vertical line, 7 of the 9 lattice points are labeled. The same pigeonhole principle shows us that there must be three consecutive labeled points on the line.

This means that we can delete one of the three squares of which these are the centers: the one in the middle.

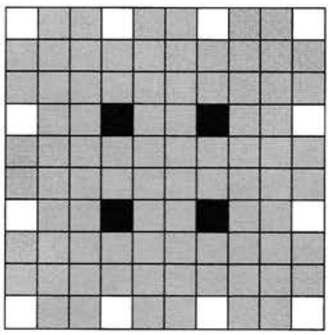


Figure 14

(b) We mark 16 squares on the board (12 white and 4 black) as in figure 14. Because the white squares are too far apart, no 2×2 square can cover more than one of them, so there are 12 squares covering the 12 white boxes, and it can be easily checked that there are exactly 8 boxes on the border that are not covered with these 12 squares. So there must be 8 other squares containing these remaining boxes on the border. We now have 20 squares covering the "border region," forming a frame with the width 2. In the central 6×6 square we must similarly have four different squares containing the four black boxes. These squares cover 16 of 36 central boxes, and it's now possible to pick 20 squares—one for each remaining central box—to obtain 24 squares covering the entire central 6×6 square. Because $20 + 24 = 44$, we can delete one of 45 given squares so that board remains covered.

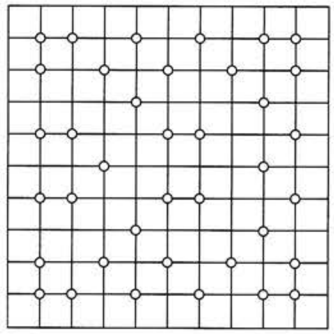


Figure 15

(c) The exact answer is 38. We can choose 38 lattice points for the centers of the 2×2 squares (fig. 15) so that if we delete any 2×2 square, the board does not remain covered. For more than 38 squares, one can always be deleted, no matter how they are arranged. The case-by-case proof is too long to give here.

Further adventures in Flexland

With our accommodating and tireless guide, Mr. Flexman

by Alexey Panov

LAST TIME MR. FLEXMAN acquainted us with the hexaflexagon. Now we'll meet other kinds of flexagon. Mr. Flexman assures me he has lots of interesting things in store for us. He begins with the simplest.

A two-way hinge

Take a piece of stiff paper and cut out two rectangles measuring 2.5 cm × 5 cm and two strips measuring 1 cm × 7 cm. Fold each of the narrow strips at both ends to form square flaps. Color each rectangle blue on one side, red on the other; color the flaps of the strips blue and the rest of the strips red on both sides. Glue the flaps at opposite corners of the bottom side of one rectangle as in figure 1. Then wrap the strips around the rectangle (fig. 2), put the second rectangle, blue surface up, on top of the first one, and glue the free flaps to this surface (fig. 3). And there you have a "two-way hinge." This might be called the simplest flexagon, because it has only two stable states. To change the state of the open flexagon, you have to draw the far edges of the rectangles together (fig. 4) and open it from the other side. Mr. Flexman has noticed that by assigning the value 0 to one state and the value 1 to the other, you turn the flexagon into a "bit" of computer memory. So, he says, it's only natural to join together a number of such bits to form a byte. I think that's pretty reasonable—let's follow his advice.

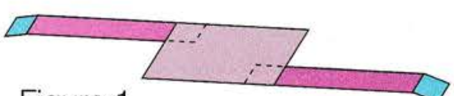


Figure 1


Figure 2


Figure 3

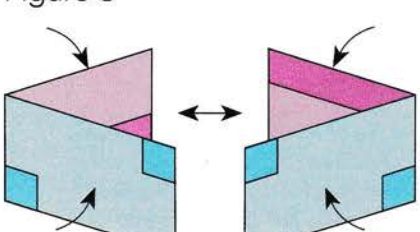


Figure 4

Flexchain

Prepare some more two-color rectangles and strips—half of the strips red with blue flaps, the other half blue with red flaps. Take the flexagon you made (I'm assuming you followed along with us and actually made a two-way hinge), put it on a table as shown in figure 3, and glue two blue strips to the two free corners of the bottom surface of the upper rectangle. Wrap the strips around this rectangle, put the next one (red side up) on top, bend the strips, and paste the flaps to the upper surface. You get a chain of three rectangles. Then attach another one to the upper rectangle of the chain with the red strips such that the new upper surface is blue, and continue in this way, alternating the colors of the rectangles and strips, until a long enough chain appears (say, eight or nine "bits"). When set on a table edgewise, the chain whimsically curves and at the slightest touch will change its shape (fig. 5). If it's made of $n + 1$ rectangles, it has n joints and can exist in 2^n states. For, say, $n = 21$, it has $2^{20} = 1{,}048{,}576$ different states. (You can do the computation for $n = 8$.) Mr. Flexman has informed me that he's now working on a flexcomputer with a flexchain memory; he's going to supply all of Flexland with them soon. But what's more exciting for me is another property of flexchains.

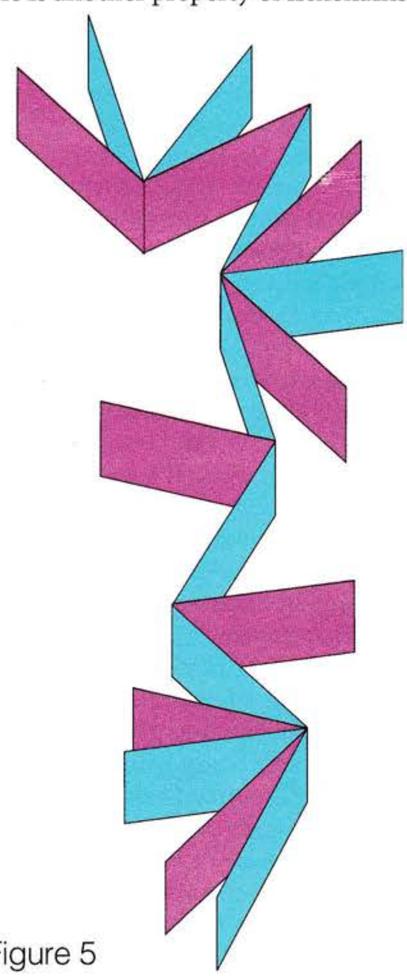


Figure 5

Flexons

Take a flexchain in any state and stretch it by pulling it apart at its ends. It will run through a series of states, ending up in a certain stable state—for instance, like the one in figure 6. A stable state is a rectilinear chain at some of whose joints single rectangles stick out. The rectangles that stick out like that will be called "flexons," and we'll call the whole stable configuration a "flexon state." These flexons have two remarkable features:

1. They can move along a flexchain.
2. When the flexons collide, they're annihilated.

To make a flexon move, push its outer edge toward the flexchain (fig. 7), activating the hinge and causing the flexon to join the chain while the neighboring rectangle springs out. This creates the impression that the "disturbed" link shifts along the flexchain. As the flexon runs along, it changes the flexchain's color. When two flexons collide, they vanish, but the flexchain becomes two units (rectangles) longer.

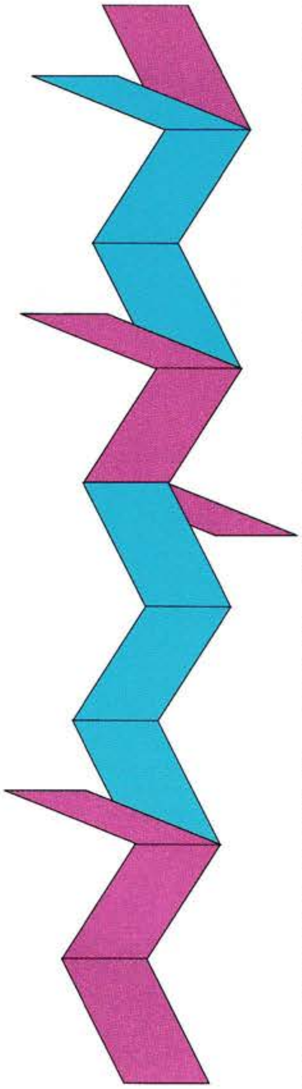


Figure 6

Mr. Flexman is also a specialist in flexphysics. He says that the flexon's length is the fundamental length in the flexuniverse, and at present he is exploring the behavior of closed flexchains with flexon perturbations.

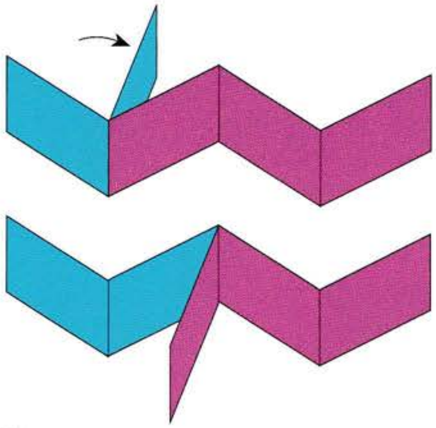


Figure 7

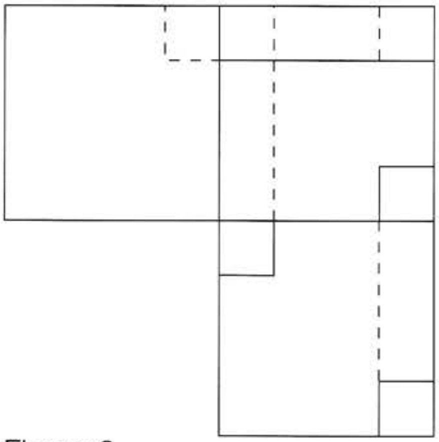


Figure 8

By the way, one kind of closed flexchain can be made out of squares (instead of rectangles). These chains are much more diverse because, in constructing such a chain, we can join the next square to any of the three free edges of the previous one. The simplest closed flexchain is made of four squares and eight strips. Instead of detailed instructions on how to assemble it, Mr. Flexman supplied us only with figures 8 and 9. He says it's a special challenge for Flexland visitors to restore the construction and color scheme of this flexagon using these figures. His only hint is that figure 9 shows all four possible colorings of the surfaces of the flexagon, and that the colorings of the two surfaces are (A, B) in the first state, (B, C) in the second, (C, D) in the third, and (D, A) in the fourth.

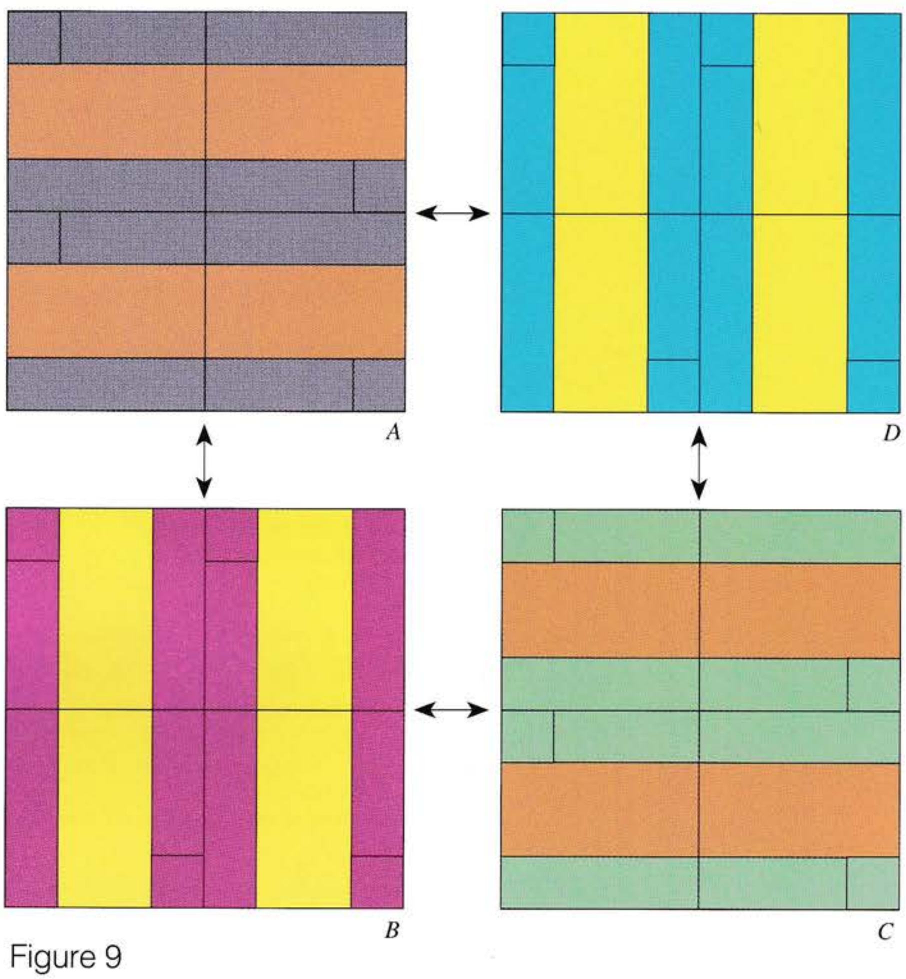


Figure 9