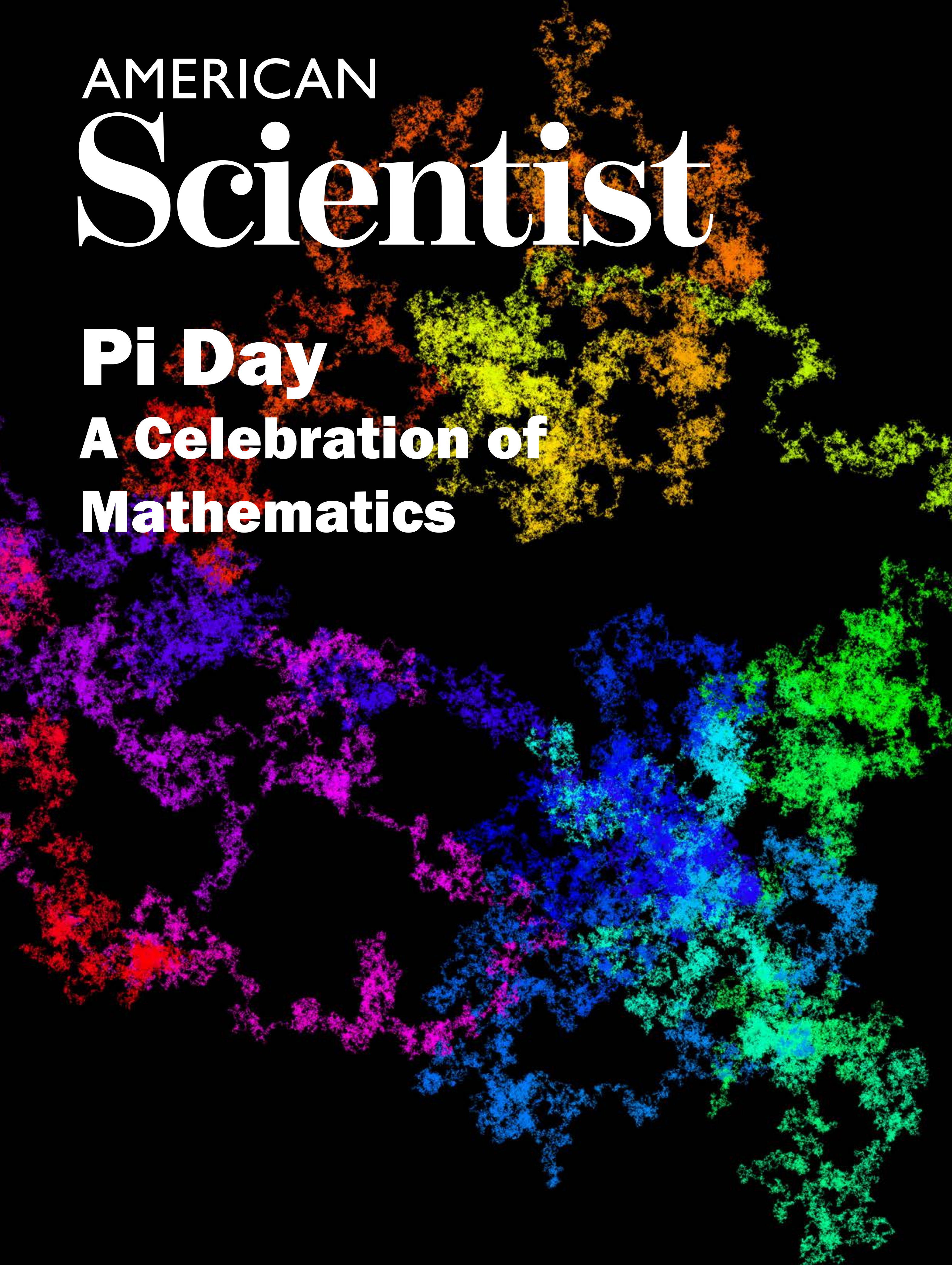
AMERICAN
Scientist

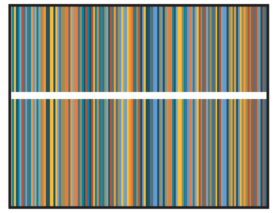


Pi Day

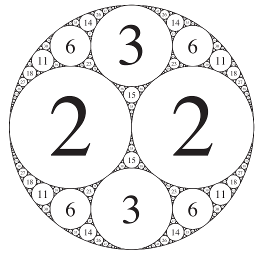
A Celebration of Mathematics

AMERICAN Scientist

Special Collection

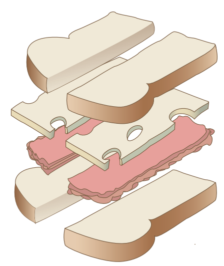


page 4



page 8

page 22



page 54

Cover Illustration Credit

Cover image created by Francisco J. Aragón Artacho under the supervision of Jonathan M. Borwein at CARMA (Centre for Computer-Assisted Research Mathematics and its Applications) of the University of Newcastle, Australia

Welcome to *American Scientist*!

Oh my, here come pi!
three-point-one-four-one-five
The constant we all know, the famous ratio.
Oh my, here come pi!
How many can you memorize?
It goes on and on, a never-ending song.
Hey, hey, it's pi day!
Make a circle, celebrate!
Irrational action, that can't be expressed as a simple fraction!
Pi!

This catchy ditty, sung by kids taking a bus ride to math camp in the 2019 movie *Wonder Park*, shows how pi has become the most famous star of the math world, and also captures some of the fun associated with Pi Day. For those who are not math enthusiasts, the subject might not bring up party imagery, but as we showcase in this collection, pi and its mathematical brethren have long been used in games and even in art.

Pi Day, recognized by the U.S. House of Representatives in 2009, is celebrated on March 14, because in the month-day format 3–14, that date calls to mind the usual abbreviated form of pi, 3.14. The idea for Pi Day is attributed to the late physicist Larry Shaw of the Exploratorium in San Francisco, who organized the first celebration in 1988; it involved prodigious pie-eating and large parades of people marching in circles while holding up signs depicting the digits of pi.

Pi is a ubiquitous constant: For any circle, its circumference divided by its diameter will be pi. This property has been known for millennia, and later the use of the lowercase Greek letter π came into play as an abbreviation of the Greek word *periféreia*, meaning circumference. However, the use of the Greek letter to designate the constant itself only became popular with the work of Swiss mathematician Leonhard Euler in the mid-1700s.

Pi remains something of a mystery—is there ever any pattern to its endless digits? To better visualize the huge data set created by the digits of pi, mathematician Francisco J. Aragón Artacho of the University of Newcastle in Australia and his colleagues created a computerized image to measure its randomness (*shown on the cover*). Aragón Artacho and his colleagues converted the first 100 billion digits of pi into base 4, so all digits were represented as 0, 1, 2, or 3, and each digit became a step in a random walk pattern. Each of the four numbers dictated the direction of the "walk" in the image (0 right, 1 up, 2 left, 3 down), and the colors indicate the path followed by the walk, red being first. After 100 billion steps, the walk seemed to come back close to where it started. But the randomness of the picture of the walk provides visual support for the conjecture that the digits of pi may be random.

Pi is famous for its geometrical roots, but its use extends into many branches of math and physics that have nothing to do with circles. This collection begins with a history of the calculation of pi, but then we broaden out to include discussions of such topics as the beautiful aspects of geometrical forms, ways that statistics has benefited science, and the reasons we should defend so-called pure mathematics, among many other topics. And there's a lot more math content that we could not fit into this collection, so check out the *American Scientist* website for a special listing of other math articles of interest.

We hope that reading this collection will inspire you to delve further into pi and other math topics. Maybe you'll be inspired to have a Pi Day celebration of your own!

Fenella Saunders
Editor-in-Chief

Pencil, Paper, and Pi

A gargantuan calculation of π in the 1850s ran up against the limits of manual arithmetic; figuring out where it went wrong calls for forensic mathematics.

Brian Hayes

William Shanks was one of the finest computers of the Victorian era—when the term *computer* denoted not a machine but a person skilled in arithmetic. His specialty was mathematical constants, and his most ambitious project was a record-setting computation of π. Starting in 1850 and returning to the task at intervals over more than 20 years, he eventually published a value of π that began with the familiar digits 3.14159 and went on for 707 decimal places.

Seen from a 21st-century perspective, Shanks is a poignant figure. All his patient toil has been reduced to triviality. Anyone with a laptop can compute hundreds of digits of π in microseconds. Moreover, the laptop will give the *correct* digits. Shanks made a series of mistakes beginning around decimal place 530 that spoiled the rest of his work.

I have long been curious about Shanks and his 707 digits. Who was this prodigious human computer? What led him to undertake his quixotic adventures in arithmetic? How did he deal with the logistical challenges of the π computation: the teetering columns of figures, the grueling bouts of multiplication and division? And what went wrong in the late stages of the work?

One way to answer these questions would be to buy several reams of paper, sharpen a dozen pencils, and try to retrace Shanks's steps. I haven't the stamina for that—or even the life expectancy. But by adapting some pencil-driven algorithms to run on silicon computers, I have gotten a glimpse of what the process might have been like for Shanks. I think I also know where a couple of his errors crept in, but there are more that remain unexplained.

Brian Hayes is senior writer for American Scientist. *Additional material related to the* Computing Science *column can be found online at http://bit-player.org. E-mail: brian@bit-player.org*

Scanty Intervals of Leisure

Biographical details about William Shanks are hard to come by. It's known that he was born in 1812, married in 1846, and died in 1882. He came from Corsenside, a village in the northeast of England, near the Scottish border. After his marriage he lived in Houghton-le-Spring, another small northeastern town, where he ran a boarding school.

Some sources identify Shanks as a student of William Rutherford, a mathematician who taught at the Royal Military Academy and also dabbled in π calculations. It's true that Shanks studied with Rutherford, but this was not the relationship of a graduate student with a thesis advisor. When Shanks published a small book on π in 1853, he dedicated it to Rutherford, "from whom I received my earliest lessons in numbers." It turns out that Rutherford taught at a school not far from Corsenside in the 1820s. Shanks was then a boy of 10 or 12, and he must have been one of Rutherford's pupils.

I have not been able to learn anything about Shanks's further education; there is no mention of a university degree. Rutherford remained a mentor and became a collaborator. The two men cross-checked their calculations of π and published some of the results jointly.

The available evidence suggests that Shanks was an amateur and a marginal figure in the mathematical community, but not a crank. He published 15 papers in the *Proceedings of the Royal Society*. Although he was never a member, he apparently had no trouble persuading Fellows to submit manuscripts on his behalf. These sponsors—some of whom were also listed as subscribers to his 1853 book—included prominent figures in British science and mathematics: George Stokes, George B. Airy, William Whewell, Augustus De Morgan.

Pencil-and-paper computation was a skill more highly prized in the 19th century than it is today. Even then, however, grinding out 707 decimal places of π was more of a stunt than a contribution to mathematical research. Shanks seems to have understood the borderline status of his project. The book he wrote about his calculations begins:

> Towards the close of the year 1850, the Author first formed the design of rectifying the Circle to upwards of 300 places of decimals. He was fully aware, at that time, that the accomplishment of his purpose would add little or nothing to his fame as a Mathematician, though it might as a Computer; nor would it be productive of anything in the shape of pecuniary recompense at all adequate to the labour of such lengthy computations. He was anxious to fill up scanty intervals of leisure with the achievement of something original, and which, at the same time, should not subject him either to great tension of thought, or to consult books.

He was surely right about the limited payoff in fame and funds. I hope he managed to avoid tension of thought.

The Recipe for Pi

There are countless ways of computing π, but almost all 19th-century calculators chose arctangent formulas. These methods begin with a geometric observation about a circle with radius 1 and circumference 2π. As shown in the diagram below, an angle drawn at the center of the circle defines both an arc along the circumference and a right triangle with sides a, b, and c. The arctangent function relates the length of side b (the "side opposite" the angle) to the length of the arc. In particular, when b has length 1, the arc is one-eighth of the circumference, which is equal to $\pi/4$. The equation arctan $1 = \pi/4$ is the key to computing π. If you can assign a numerical value to arctan 1, you get an approximation to $\pi/4$; multiply this number by 4 to get a value for π itself.

The next question is how to compute an arctangent. The pioneers of calculus devised an infinite series that gives the value of arctan x for any value of x between –1 and +1:

$$\arctan x = \frac{x^1}{1} - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

For the case of $x = 1$, the series assumes a particularly simple form:

$$\arctan 1 = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

Hence to calculate π one can just add up the terms of this series—the reciprocals of successive odd numbers, with alternating plus and minus signs—until the sum attains the desired accuracy.

Lamentably, this plan won't work. At $x=1$ the arctan series converges at an agonizingly slow pace. To get n digits of π, you need to sum roughly 10^n terms of the series. Shanks would have had to evaluate more than 10^{700} terms, which is beyond the means of even the most intrepid Victorian scribbler.

All is not lost. For values of x closer to zero, the arctan series converges more quickly. The trick, then, is to combine multiple arctan calculations that sum up to the same value as arctan 1. Shanks worked with the following formula, discovered in 1706 by the English mathematician John Machin:

$$\frac{\pi}{4} = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239}$$

He had to evaluate two arctan series rather than just one, but both of these series converge much faster.

The upper illustration on page 344 traces the evaluation of the first three terms of the series for arctan 1/5 and arctan 1/239, retaining five decimal places of precision. The error in the computed value of π is 0.00007. No extraordinary skill in arithmetic is needed to carry out this computation by hand. But now imagine scaling it up to several hundred terms and several hundred decimal places. The basic operations remain the same, but keeping all the figures straight becomes a clerical nightmare.

In computing arctan 1/5, Shanks evaluated 506 terms, each carried to 709 decimal places. Most likely he performed separate summations of the positive and negative terms. If he tried to write down such an addition problem all in one piece—253 rows of 709-digit numbers, or almost 180,000 digits in all—it would fill a sheet of paper two meters wide by a meter high. Breaking the task down into smaller pieces makes it less awkward physically but entails other costs: extra copying of intermediate results, transferring carry digits, the risk of misaligning columns or rows.

Erwin Engert, a Shanks enthusiast, has tested the travails of pencil-and-paper calculation by doing 20-digit and 40-digit evaluations of Machin's arctan formula. The results are on his website at http://engert.us/erwin/Miscellaneous.html. The challenge of keeping digits aligned became severe enough that Engert printed ruled forms for the larger computation. Shanks may well have done the same, although we have no direct evidence.

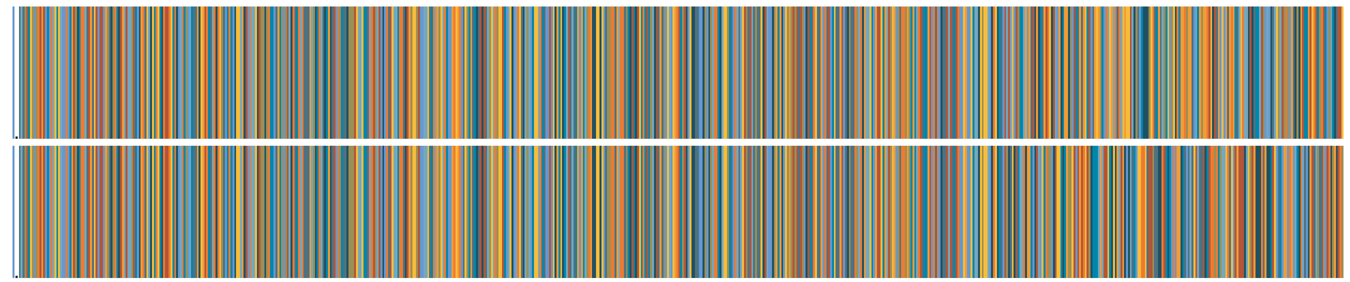
Pencil-Friendly Algorithms

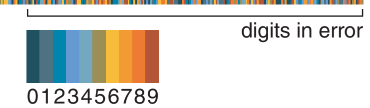
In silico, summing n terms of the series for arctan x takes just a few lines of code:

```
function arctan(x, n)
  sum = 0
  for k from 0 to n−1
    sign = (−1)^k
    m = 2 × k + 1
    term = sign × x^m / m
    sum = sum + term
  return sum
```

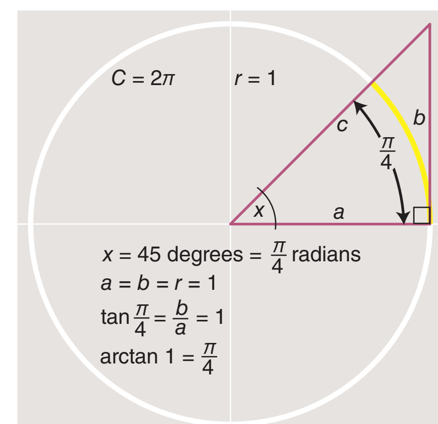
For each integer k from 0 to $n-1$, the program generates an odd integer m and the corresponding term of the arctan series, x^m/m. The expression $(-1)^k$ sets the sign of the term—plus for even k, minus for odd. When the loop completes, the function returns the accumulated sum of the n terms. The only hidden subtlety here is that the numeric variables must be able to accommodate numbers of arbitrary size and precision.

No one doing arithmetic with a pencil would adopt an algorithm anything like this one. After every pass through the loop, the program throws away





The digits of π are encoded in ribbons of color. (The mapping of digits to colors is given in the key at right.) The upper band shows 707 correct decimal places of π; below are the digits computed between 1850 and 1873 by William Shanks. Errors begin in the 528th decimal place.



The pie slice that helps determine the value of π is an eighth of a circle, with an angle of 45 degrees, or π/4 radians. The tangent of this angle, defined as the ratio *b/a* in the red triangle, is equal to 1. Hence, computing the arctangent of 1 yields a numerical value for π/4.

All illustrations by Brian Hayes

$$
\begin{array}{lcccccccr}
\arctan x & = & \frac{x^1}{1} & - & \frac{x^3}{3} & + & \frac{x^5}{5} & & \\
\arctan \frac{1}{5} & = & 0.20000 & - & 0.00267 & + & 0.00006 & = & 0.19739 \\
 & & & & & & & \times & 4 \\ \cline{9-9}
 & & & & & & & & 0.78956 \\
\arctan \frac{1}{239} & = & 0.00418 & - & 0.00000 & + & 0.00000 & = & -\ 0.00418 \\ \cline{9-9}
 & & & & & & & & 0.78538 \\
 & & & & & & & \times & 4 \\ \cline{9-9}
 & & \multicolumn{5}{c}{\pi = 4\left(4 \arctan \frac{1}{5} - \arctan \frac{1}{239}\right)} & \approx & 3.14152
\end{array}
$$

A crude computation of π proceeds by summing the first three terms in an infinite series for arctan 1/5 and arctan 1/239. Each term is evaluated to five decimal places. Plugging these values into John Machin's formula *(bottom)* yields four correct decimal places of π.

all its work except the variables k and *sum*, then starts from scratch to build the next term of the series. A manual worker would surely save the value of x^m as a starting point for calculating the next power, x^{m+2}. And exponentiating –1 is not how a human computer would keep track of alternating signs.

It's not hard to transform the program into a more pencil-friendly procedure, avoiding needless recomputation and saving intermediate results for future use. Moreover, the computer can be programmed to use digit-by-digit algorithms—the ones we all learned in elementary school, and forgot soon after—for multiplication and long division. But these alterations still fail to capture some important practices of a shrewd human reckoner.

Most of the terms in the series for arctan 1/5 are repeating decimals with a short period. For example, the term $(1/5)^9/9$ works out to 0.000000056888…. A naive computer program would go on dividing digit after digit out to the limit of precision, but Shanks surely just filled in a string of 8s.

There are also peculiarities of base 10 to be taken into account. For generating the sequence of odd powers of 1/5, the basic step is dividing by 25. Engert suggests dividing by 100 (a shift of the decimal point) and multiplying by 4. Another option is to calculate $(1/5)^m$ as $2^m/10^m$ (where again division by a power of 10 is just a decimal-point shift). I mention this latter possibility because Shanks's book on the π computation includes a table of the powers of 2 up to 2^{721}. Did he use those numbers to compute his powers of 1/5, or were they just for checking values computed in some other way?

Shanks doesn't reveal much about his computational methods, and I remain unsure about several aspects of his strategy. For example, a term in the series for arctan 1/5 can be written either as $(1/5)^m/m$ or as $1/(m5^m)$. Mathematically these expressions are identical, but they imply different computations. In the first case you multiply and divide long decimal fractions; in the second you build a large integer and then take its reciprocal. Which way did Shanks do it? He doesn't say. If I were to attempt to replicate his work, I might stick with decimal fractions for arctan 1/5, because of the many short-period repetitions, but I might choose the reciprocal method for arctan 1/239, because taking a reciprocal is a little easier than other forms of division.

Where He Went Wrong

As Tolstoy might have said, all correct computations are alike, but every erroneous one errs in its own way. In that spirit, the incorrect digits in Shanks's result are much more informative than the correct ones. If nothing else, they might reveal just where and how his computation went off the rails.

Shanks published his value of π in three stages. A January 1853 article (under Rutherford's byline) includes 530 decimal places; 440 of those figures were confirmed by Rutherford, and the rest were also correct apart from a few typographical errors and a discrepancy in the last two digits that could be attributed to round-off.

In the spring of 1853, Shanks extended his calculation from 530 to 607 decimal places, publishing these results in a privately printed book, *Contributions to Mathematics, Comprising Chiefly the Rectification of the Circle to 607 Places of Decimals*. This is where the errors creep in. His value of arctan 1/5 goes awry in the 530th decimal place, right on the boundary between the old and the new computations. Because arctan 1/5 is multiplied by 16 in the Machin formula, the error propagates back to the 528th decimal place in the value of π. Shanks's sum for arctan 1/239 is also incorrect, starting at the 592nd decimal place.

After bringing out his book, Shanks put π aside for 20 years. When he took up the task again in 1873, he extended the two arctan series to 709 decimal places and π to 707. Because these computations were built atop the flawed earlier work, they were doomed from the start. The errors weren't noticed until 75 years later, when D. F. Ferguson, working with a mechanical desk calculator, extended a new calculation of π beyond 700 digits.

Trying to discover where Shanks went wrong is an interesting exercise in forensic mathematics. Usually, one strives to find the correct answer to a problem; here the aim is to get the wrong answer—but the *right* wrong answer. We want to take a correct value and find some way of modifying it that will yield the specific erroneous output reported by Shanks. It's like searching for a suspicious transaction when your checkbook disagrees with the bank statement, except that we have no access to the individual checkbook entries, only the final balance.

To search for an error in the arctan 1/5 series, I took the difference between

arctan 1/5	24383 02697 56051 83775 74220 87783 58531 52464 74933 09145 87633 82311 24903 32030 12680 51006 70223 31257 50509 42448
term 248	24383 02697 56051 83776 17781 64242 33783 03370 18192 64880 28277 68629 15647 78710 20728 79980 54529 14758 51113 04621
term 72	24383 02697 56051 83776 17781 64242 33783 03370 18192 64880 28277 68611 91509 85606 75901 21359 85563 63034 37319 94276
Shanks 53	24383 02697 56051 83776 17781 64242 33783 03370 18192 64880 28277 68611 91509 85606 75901 21359 85563 63034 34783 9926
Shanks 73	24383 02697 56051 83776 17781 64242 33783 03370 18192 64880 28277 68611 91509 85606 75901 21359 85563 63034 32100 56649
	520 530 540 550 560 570 580 590 600 610

Forensic analysis tries to identify simple errors that transform the correct value of arctan 1/5 *(top)* into the erroneous values published by Shanks *(bottom)*. Omitting a 0 at position 530 in term 248 "uncorrects" 39 digits *(yellow band)*. A five-digit omission in term 72 leads to a match with another 33 digits of the Shanks value *(orange band)*. Further errors remain, but the situation is confusing; the final eight digits of a 609-place computation from 1853 were changed without explanation when Shanks returned to the task in 1873.

the true value and Shanks's value, then subtracted this discrepancy from each of the 506 terms of the series. In most cases the result was uninformative, but my eye was drawn to this pattern, in the 248th term:

T: 7444668008048289738430583501
S: 7444668008483897384305835010

Sequence *T* comes from the true arctan sum, starting at decimal place 520; sequence *S* is the same region after subtracting the discrepancy. In the first 10 positions the two numbers agree, but thereafter *S* is a shifted version of *T*, created by omitting the 0 marked in red and letting the rest of the digits slide left one place. (There's also a substitution a few digits later, where a 2 becomes a 3.)

Without further documentary evidence, it's not possible to prove that this spot marks the site of Shanks's first error, but it's certainly a plausible hypothesis. When Shanks extended this term from 530 digits to 609, he didn't need to do any actual arithmetic. The term is a repeating decimal with a period of 210 digits, so he merely needed to copy a segment from earlier in the sequence. It seems likely that he missed that 0 digit while copying. I was not the first to discover this error; Erwin Engert identified it before I did.

If you inject this one-digit shift error into the arctan calculation, the output matches the Shanks value in the region following decimal place 530, but the agreement does not continue all the way to the end. At decimal place 569 the two sequences part ways again. Evidently there's another mistake.

I wasn't the first to notice this problem, either. In 1946 Ferguson called attention to an anomaly in term 72 and suggested that Shanks had omitted all the digits of this term from position 569 on. I believe that Ferguson correctly identified the trouble spot, but his diagnosis is not quite right. Truncating term 72 in this way does not transform the correct sum into the Shanks value. But another simple change does work: omitting five digits at position 569 and shifting the rest of the term to the left.

With these two "*un*corrections," we can transform the true value of arctan 1/5 into the Shanks value through decimal place 601. At that point there must be yet another error, but the situation is confusing. The last eight digits of the 609-place value published in 1853 differ from the corresponding digits listed in 1873. I have not found a simple error that yields either version. The error in arctan 1/239 also remains unexplained.

It's curious that Shanks produced almost 530 flawless digits of π, then made at least four mistakes in the next 80 digits. All four errors date from March or April of 1853, and they seem to be clerical rather than mathematical. I can only speculate on the cause of this sudden spate of carelessness. Perhaps Shanks was hurrying to get his book into the hands of the subscribers. Or maybe, at age 41, he was experiencing the early symptoms of presbyopia.

Stories about Shanks tend to focus on the mistakes. We look back with pity and horror on all those pages of meticulous arithmetic rendered worthless by a slip of the pencil. But I would argue that even with the errors, Shanks's computation of π was an impressive endeavor. His 527 correct digits were not bettered for almost a century. Augustus De Morgan, one of the leading mathematicians of the era, had his doubts about Shanks's work, but he also spoke admiringly of "the power to calculate, and… the courage to face the labour."

For further material on Shanks, including references and programs for exploring his computation, see http://bit-player.org/shanks.

A Tisket, a Tasket, an Apollonian Gasket

Dana Mackenzie

In the spring of 2007 I had the good fortune to spend a semester at the Mathematical Sciences Research Institute in Berkeley, an institution of higher learning that takes "higher" to a whole new extreme. Perched precariously on a ridge far above the University of California at Berkeley campus, the building offers postcard-perfect vistas of the San Francisco Bay, 1,200 feet below. That's on the west side. Rather sensibly, the institute assigned me an office on the east side, with a view of nothing much but my computer screen. Otherwise I might not have gotten any work done.

However, there was one flaw in the plan: Someone installed a screen-saver program on the computer. Of course, it had to be mathematical. The program drew an endless assortment of fractals of varying shapes and ingenuity. Every couple minutes the screen would go blank and refresh itself with a completely different fractal. I have to confess that I spent a few idle minutes watching the fractals instead of writing.

One day, a new design popped up on the screen *(see the first figure)*. It was different from all the other fractals. It was made up of simple shapes—circles, in fact—and unlike all the other screen-savers, it had numbers! My attention was immediately drawn to the sequence of numbers running along the bottom edge: 1, 4, 9, 16 … They were the perfect squares! The sequence was 1-squared, 2-squared, 3-squared, and so on.

> *Fractals made of circles do funny things to mathematicians*

Before I became a full-time writer, I used to be a mathematician. Seeing those numbers awakened the math geek in me. What did they mean? And what did they have to do with the fractal on the screen? Quickly, before the screen-saver image vanished into the ether, I sketched it on my notepad, making a resolution to find out someday.

As it turned out, the picture on the screen was a special case of a more general construction. Start with three circles of any size, with each one touching the other two. Draw a new circle that fits snugly into the space between them, and another around the outside enclosing all the circles. Now you have four roughly triangular spaces between the circles. In each of those spaces, draw a new circle that just touches each side. This creates 12 triangular pores; insert a new circle into each one of them, just touching each side. Keep on going forever, or at least until the circles become too small to see. The resulting foam-like structure is called an Apollonian gasket *(see the second figure)*.

Something about the Apollonian gasket makes ordinary, sensible mathematicians get a little bit giddy. It inspired a Nobel laureate to write a poem and publish it in the journal *Nature*. An 18th-century Japanese samurai painted a similar picture on a tablet and hung it in front of a Buddhist temple. Researchers at AT&T Labs printed it onto T-shirts. And in a book about fractals with the lovely title *Indra's Pearls*, mathematician David Wright compared the gasket to Dr. Seuss's *The Cat in the Hat*:

> The cat takes off his hat to reveal Little Cat A, who then removes his hat and releases Little Cat B, who then uncovers Little Cat C, and so on. Now imagine there are not one but three cats inside each cat's hat. That gives a good impression of the explosive proliferation of these tiny ideal triangles.

Getting the Bends

Even the first step of drawing an Apollonian gasket is far from straightforward. Given three circles, how do you draw a fourth circle that is exactly tangent to all three?

Apparently the first mathematician to seriously consider this question was Apollonius of Perga, a Greek geometer who lived in the third century B.C. He has been somewhat overshadowed by his predecessor Euclid, in part because most of his books have been lost. However, Apollonius's surviving book *Conic Sections* was the first to systematically study ellipses, hyperbolas and parabolas—curves that have remained central to mathematics ever since.

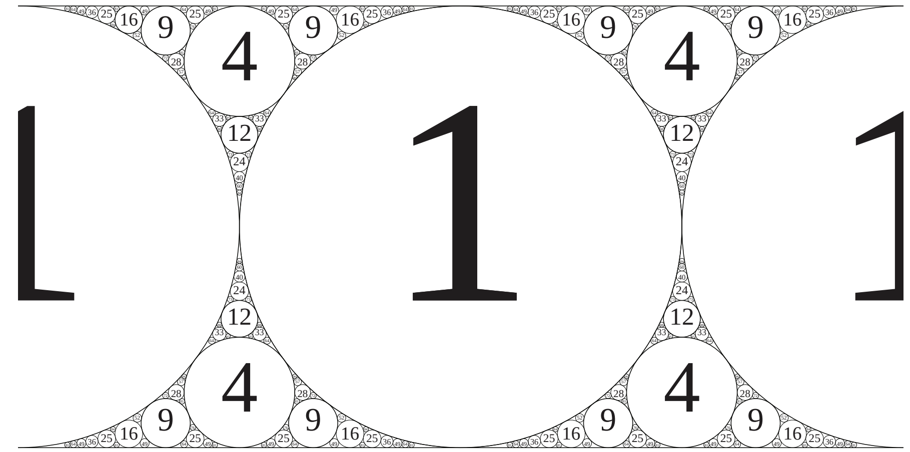
One of Apollonius's lost manuscripts was called *Tangencies*. According to later commentators, Apollonius apparently solved the problem of drawing circles that are simultaneously tangent to three lines, or two lines and a circle, or two circles and a line, or three circles. The hardest case of all was the case where the three circles are tangent.

No one knows, of course, what Apollonius' solution was, or whether it was correct. After many of the writings of the ancient Greeks became available again to European scholars of the Renaissance, the unsolved "problem of Apollonius" became a great challenge. In 1643, in a letter to Princess Elizabeth of Bohemia, the French philosopher and mathematician René Descartes correctly stated (but incorrectly proved) a beautiful formula concerning the radii of four mutually touching circles. If the radii are r, s, t and u, then Descartes's formula looks like this:

$$\frac{1}{r^2}+\frac{1}{s^2}+\frac{1}{t^2}+\frac{1}{u^2}=\frac{1}{2}\left(\frac{1}{r}+\frac{1}{s}+\frac{1}{t}+\frac{1}{u}\right)^2.$$

All of these reciprocals look a little bit extravagant, so the formula is usually simplified by writing it in terms of the *curvatures* or the *bends* of the circles. The curvature is simply defined as the reciprocal of the radius. Thus, if the curvatures are denoted by a, b, c and d, then Descartes's formula reads as follows:

$$a^2+b^2+c^2+d^2=(a+b+c+d)^2/2.$$

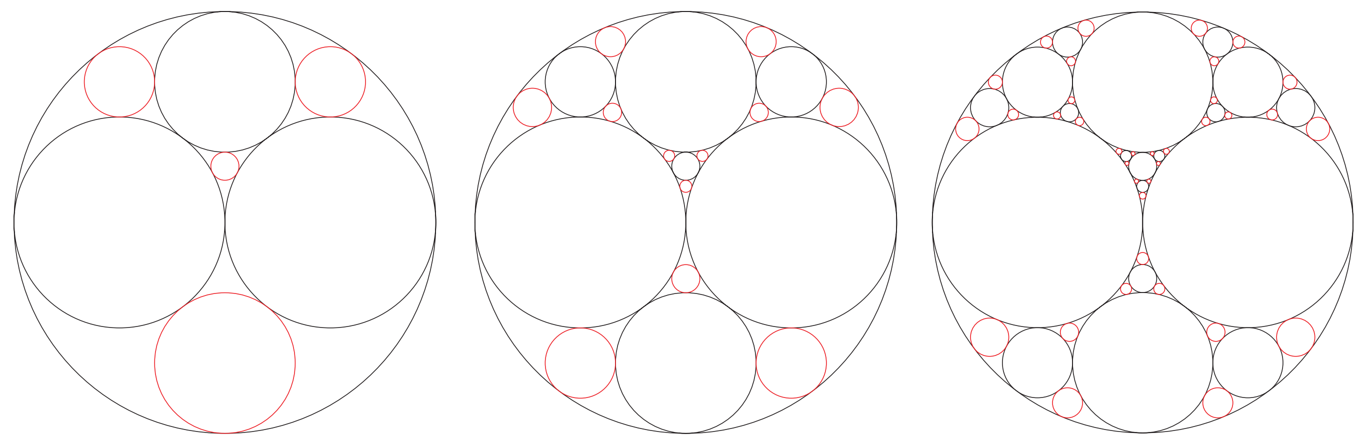


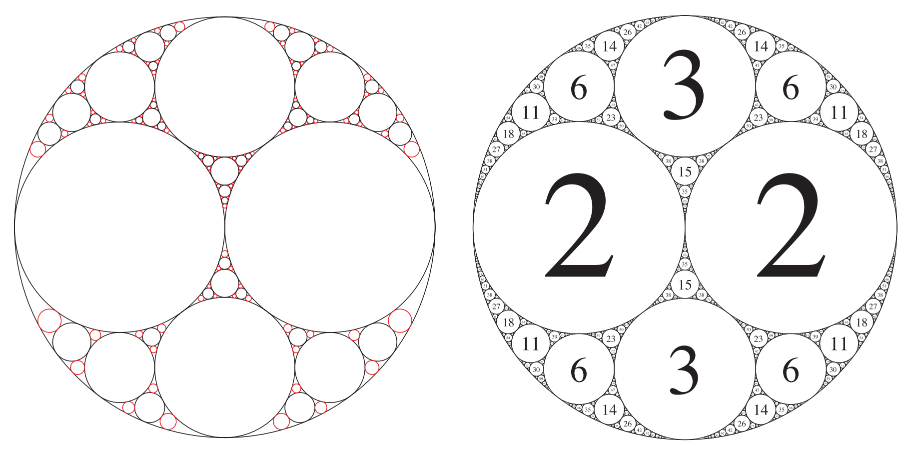
Numbers in an Apollonian gasket correspond to the curvatures or "bends" of the circles, with larger bends corresponding to smaller circles. The entire gasket is determined by the first four mutually tangent circles; in this case, two circles with bend 1 and two "circles" with bend 0 (and therefore infinite radius). The circles with a bend of zero look, of course, like straight lines. (Image courtesy of Alex Kontorovich.)

As the third figure shows, Descartes's formula greatly simplifies the task of finding the *size* of the fourth circle, assuming the sizes of the first three are known. It is much less obvious that the very same equation can be used to compute the *location* of the fourth circle as well, and thus completely solve the drawing problem. This fact was discovered in the late 1990s by Allan Wilks and Colin Mallows of AT&T Labs, and Wilks used it to write a very efficient computer program for plotting Apollonian gaskets. One such plot went on his office door and eventually got made into the aforementioned T-shirt.

Descartes himself could not have discovered this procedure, because it involves treating the coordinates of the circle centers as complex numbers. Imaginary and complex numbers were not widely accepted by mathematicians until a century and a half after Descartes died.

In spite of its relative simplicity, Descartes's formula has never become widely known, even among mathematicians. Thus, it has been rediscovered over and over through the years. In Ja-





An Apollonian gasket is built up through successive "generations." For instance, in generation 1 *(top left)*, each of the red circles is inscribed in one of the four triangular pores formed by the black circles. The final gasket shown here, whimsically named "bugeye" by Katherine Sanden, an undergraduate student of Peter Sarnak at Princeton University, has circles with bends –1 (for the largest circle that encloses the rest), 2, 2 and 3. The list of bends that appears in a given gasket (here, 2, 3, 6, 11, etc.) form a number sequence whose properties Sarnak would like to explain—but, he says, "the necessary mathematics has not been invented yet." (Image courtesy of Katherine Sanden.)

Dana Mackenzie received his doctorate in mathematics from Princeton University in 1983 and taught at Duke University and Kenyon College. Since 1996 he has been a freelance writer specializing in math and science, and he has frequently edited articles on mathematical topics for American Scientist. His published books include The Big Splat, or How Our Moon Came to Be *(Wiley, 2003), and volumes 6 and 7 of* What's Happening in the Mathematical Sciences *(American Mathematical Society, 2007 and 2009). Email: scribe@danamackenzie.com*

pan, during the Edo period, a delightful tradition arose of posting beautiful mathematics problems on tablets that were hung in Buddhist or Shinto temples, perhaps as an offering to the gods. One of these "Japanese temple problems," or *sangaku*, is to find the radius of a circle that just touches two circles and a line, which are themselves mutually tangent. This is a restricted version of the Apollonian problem, where one circle has infinite radius (or zero bend). The anonymous author shows that, in this case, $\sqrt{a}+\sqrt{b}=\sqrt{c}$, a sort of demented version of the Pythagorean theorem. This formula, by the way, explains the pattern I saw in the screensaver. If the first two circles have bends 1 and 1, then the circle between them will have bend 4, because $\sqrt{1}+\sqrt{1}=\sqrt{4}$. The next circle will have bend 9, because $\sqrt{1}+\sqrt{4}=\sqrt{9}$. Needless to say, the pattern continues forever. (This also explains what the numbers in the first figure mean. Each circle is labeled with its own bend.)

Apollonian circles experienced perhaps their most glorious rediscovery in 1936, when the Nobel laureate (in chemistry, not mathematics) Frederick Soddy became mesmerized by their charm. He published in *Nature* a poetic version of Descartes' theorem, which he called "The Kiss Precise":

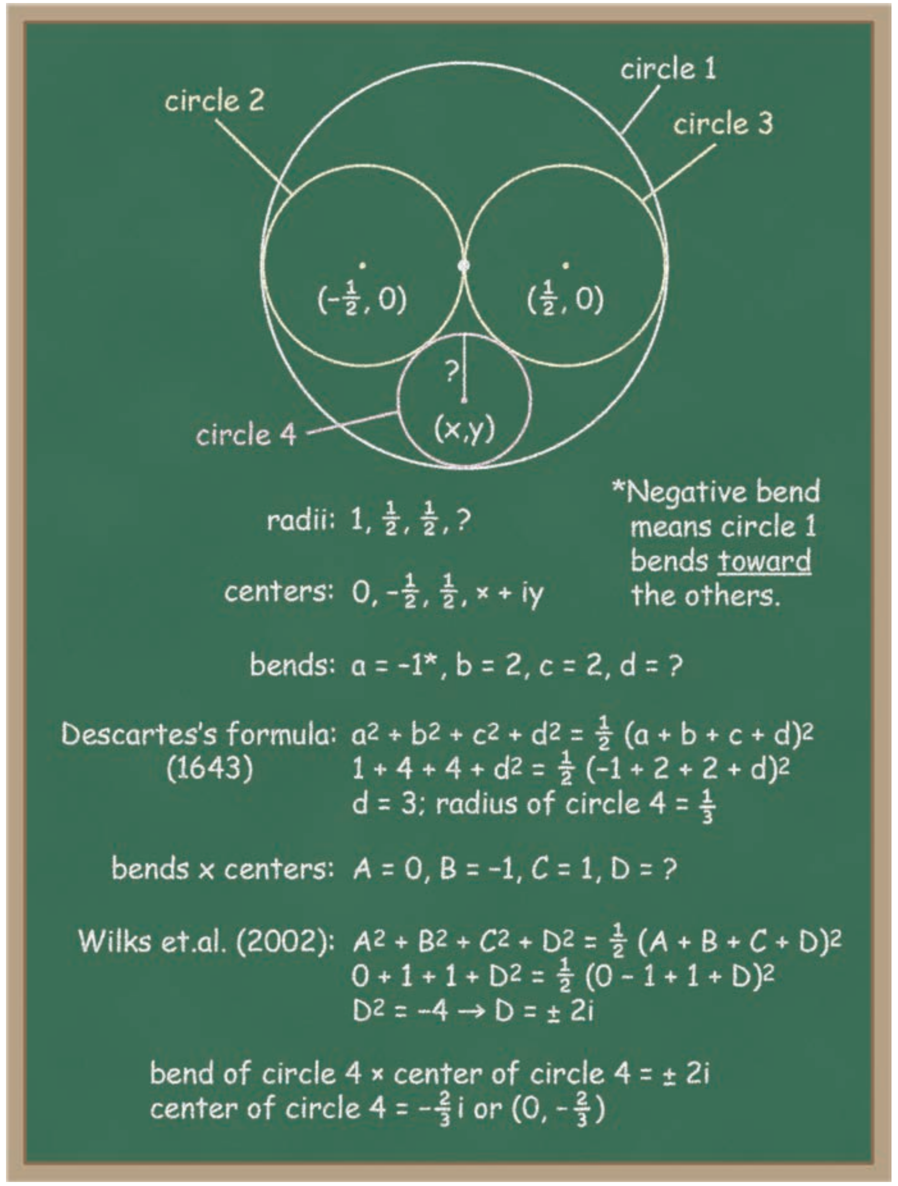


illustration by Barbara Aulicino

In 1643 René Descartes gave a simple formula relating the radii of any four mutually tangent circles. More than 350 years later, Allan Wilks and Colin Mallows noticed that the same formula relates the coordinates of the centers of the circles (expressed as complex numbers). Here Descartes's formula is used to find the radius and center of the fourth circle in the "bugeye" packing.

Four circles to the kissing come
The smaller are the benter.
The bend is just the inverse of
The distance from the center.
Though their intrigue left
Euclid dumb,
There's now no need for rule
of thumb.
Since zero bend's a dead
straight line,
And concave bends have
minus sign,
The sum of the squares of all
four bends
Is half the square of their sum.

Soddy went on to state a version for three-dimensional spheres (which he was also not the first to discover) in the final stanza of his poem.

Ever since Soddy's prosodic effort, it has become something of a tradition to publish any extension of his theorem in poetic form as well. The following year, Thorold Gosset published an *n*-dimensional version, also in *Nature*. In 2002, when Wilks, Mallows and Jeff Lagarias published a long article in the *American Mathematical Monthly*, they ended it with a continuation of Soddy's poem entitled "The Complex Kiss Precise":

Yet more is true: if all four discs
Are sited in the complex plane,
Then centers over radii
Obey the self-same rule again.

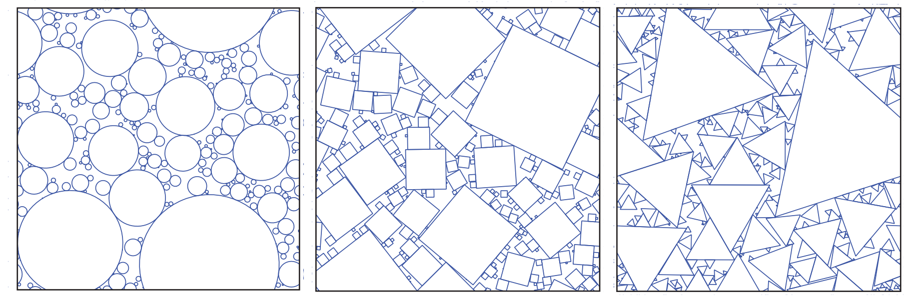
(The authors note that the poem is to be pronounced in the Queen's English.)

A Little Bit of Gasketry

To this point I have only written about the very beginning of the gasket-making process—how to inscribe one circle among three given circles. However, the most interesting phenomena show up when you look at the gasket as a whole.

The first thing to notice is the foam-like structure that remains after you cut out all of the discs in the gasket. Clearly the disks themselves take up an area that approaches 100 percent of the area within the outer disk, and so the area of the foam (known as the "residual set") must be zero. On the other hand, the foam also has infinite length. Thus, in fact, it was one of the first known examples of a *fractal*—a curve of dimension between 1 and 2. Even today its dimension (denoted δ) is not known exactly; the best-proven estimate is 1.30568.

The concept of fractional dimension was popularized by Benoît Mandelbrot in his enormously influential book *The Fractal Geometry of Nature*. Although the meaning of dimension 1.30568 is somewhat opaque, this number is related to other properties of the foam that have direct physical meaning. For instance, if you pick any cutoff radius r, how many bubbles in the foam have radius larger than r? The answer, denoted $N(r)$, is roughly proportional to r^{δ}. Or if you pick the n largest bubbles, what is the remaining pore space between those bubbles? The answer is roughly proportional to $n^{1-2/\delta}$.



Physicists study random Apollonian packings as a model for foams or powders. In these simulations, new bubbles or grains nucleate in a random place and grow, either with rotation or without, until they encounter another bubble or grain. Different geometries for the bubbles or grains, and different growth rules, lead to different values for the dimension of the residual set—a way of measuring the efficiency of the packing. (Image courtesy of Stefan Hutzler and Gary Delaney.)

Physicists are very familiar with this sort of rule, which is called a *power law*. As I read the literature on Apollonian packings, an interesting cultural difference emerged between physicists and mathematicians. In the physics literature, a fractional dimension δ is *de facto* equivalent to a power law r^{δ}. However, mathematicians look at things through a sharper lens, and they realize that there can be additional, slowly increasing or slowly decreasing terms. For instance, $N(r)$ could be proportional to $r^{\delta}\log(r)$ or $r^{\delta}/\log(r)$. For physicists, who study foams empirically (or semi-empirically, via computer simulation), the logarithm terms are absolutely undetectable. The discrepancy they introduce will always be swamped by the noise in any simulation. But for mathematicians, who deal in logical rigor, the logarithm terms are where most of the action is. In 2008, mathematicians Alex Kontorovich and Hee Oh of Brown University showed that there are in fact no logarithm terms in $N(r)$. The number of circles of radius greater than r obeys a strict power law, $N(r) \sim Cr^{\delta}$, where C is a constant that depends on the first three circles of the packing. For the "bugeye" packing illustrated in the second figure, C is about 0.201. (The tilde (~) means that this is not an *equation* but an *estimate* that becomes more and more accurate as the radius r decreases to 0.) For mathematicians, this was a major advance. For physicists, the likely reaction would be, "Didn't we know that already?"

Random Packing

For many physical problems, the classical definition of the Apollonian gasket is too restrictive, and a random model may be more appropriate. A bubble may start growing in a randomly chosen location and expand until it hits an existing bubble, and then stop. Or a tree in a forest may grow until its canopy touches another tree, and then stop. In this case, the new circles do not touch three circles at a time, but only one. Computer simulations show that these "random Apollonian packings" still behave like a fractal, but with a different dimension. The empirically observed dimension is 1.56. (This means the residual set is larger, and the packing is less efficient, than in a deterministic Apollonian gasket.) More recently, Stefan Hutzler of Trinity College Dublin, along with Gary Delaney and Tomaso Aste of the University of Canberra, studied the effect of bubbles with different shapes in a random Apollonian packing. They found, for example, that squares become much more efficient packers than circles if they are allowed to rotate as they grow, but surprisingly, triangles become only slightly more efficient. As far as I know, all of these results are begging for a theoretical explanation.

For mathematicians, however, the classical, deterministic Apollonian gasket still offers more than enough challenging problems. Perhaps the most astounding fact about the Apollonian gasket is that if the first four circles have integer bends, then *every other circle* in the packing does too. If you are given the first three circles of an Apollonian gasket, the bend of the fourth is found (as explained above) by solving a quadratic equation. However, every *subsequent* bend can be found by solving a *linear* equation:

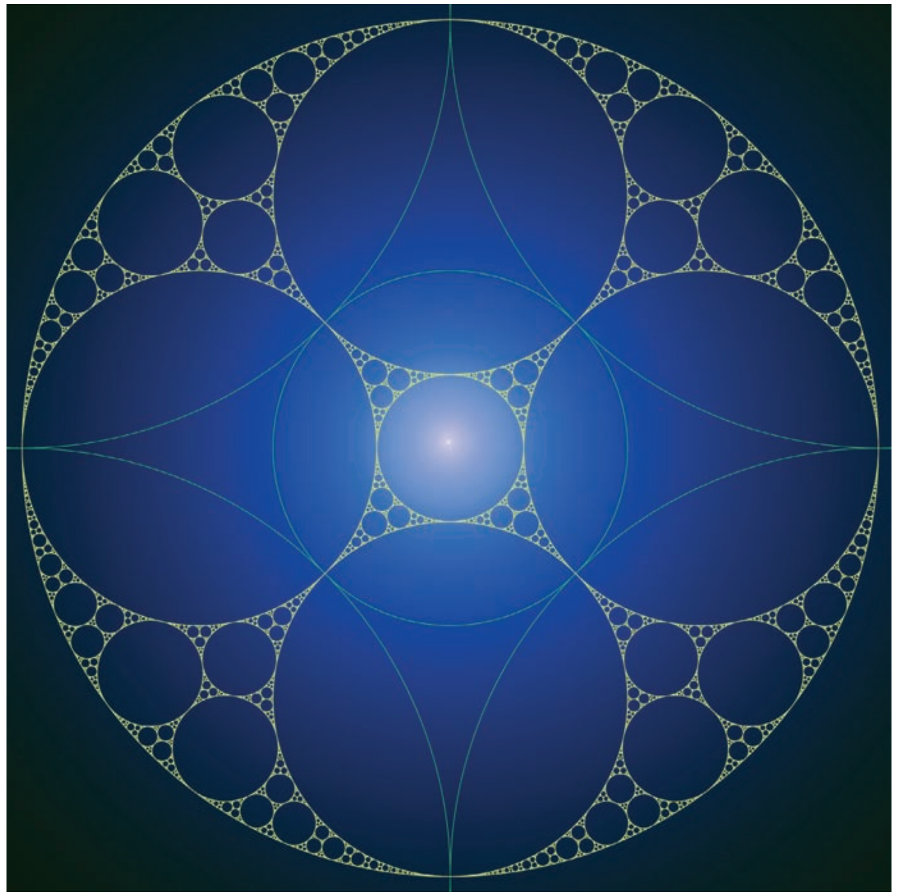
$$d+d'=2(a+b+c)$$

For instance, in the "bugeye" gasket, the three circles with bends $a=2$, $b=3$, and $c=15$ are mutually tangent to two other circles. One of them, with bend $d=2$, is already given in the first generation. The other has bend $d'=38$, as predicted by the formula, $2+38=2(2+3+15)$. More importantly, even if we did not know d', we would still be guaranteed that it was an integer, because a, b, c and d are.

Hidden behind this "baby Descartes equation" is an important fact about Apollonian gaskets: They have a very high degree of symmetry. Circles a, b and c actually form a sort of curved mirror that reflects circle d to circle d' and vice versa. Thus the whole gasket is like a kaleidoscopic image of the first four circles, reflected again and again through an infinite collection of curved mirrors.

Kontorovich and Oh exploited this symmetry in an extraordinary and amusing way to prove their estimate of the function $N(r)$. Remember that $N(r)$ simply counts how many circles in the gasket have radius larger than r. Kontorovich and Oh modified the function $N(r)$ by introducing an extra variable of position—roughly equivalent to put-

A favorite example of Sarnak's is the "coins" gasket, so called because three of the four generating circles are in proportion to the sizes of a quarter, nickel and dime, respectively. (Image courtesy of Alex Kontorovich.)



Many variations on the Apollonian gasket construction are possible. In this beautiful example, each pore is occupied by three inscribed circles rather than by one. Light blue arcs represent five "curved mirrors." Reflections in these curved mirrors—known technically as circle inversions—create a kaleidoscopic effect. Every circle in the gasket is generated by repeated inversions of the first six circles through these curved mirrors. (Image courtesy of Jos Leys.)

ting a lightbulb at a point x and asking how many circles illuminated by that lightbulb have radius larger than r. The count will fluctuate, depending on exactly where the bulb is placed. But it fluctuates in a very predictable way. For instance, the count is unchanged if you move the bulb to the location of any of its kaleidoscopic reflections.

This property makes the "lightbulb counting function" a very special kind of function, one which is invariant under the same symmetries as the Apollonian gasket itself. It can be broken down into a spectrum of similarly symmetric functions, just as a sound wave can be decomposed into a fundamental frequency and a series of overtones. From this spectrum, you can in theory find out everything you want to know about the lightbulb counting function, including its value at any particular location of the lightbulb.

For a musical instrument, the fundamental frequency or lowest overtone is the most important one. Similarly, it turned out that the first symmetric function was all that Kontorovich and Oh needed to figure out what happens to $N(r)$ as r approaches 0.

In this way, a simple problem in geometry connects up with some of the most fundamental concepts of modern mathematics. Functions that have a kaleidoscopic set of symmetries are rare and wonderful. Kontorovich calls them "the Holy Grail of number theory." Such functions were, for instance, used by Andrew Wiles in his proof of Fermat's Last Theorem. An interesting new kaleidoscope is enough to keep mathematicians happy for years.

Gaskets Galore

Kontorovich learned about the Apollonian kaleidoscope from his mentor, Peter Sarnak of Princeton University, who learned about it from Lagarias, who learned about it from Wilks and Mallows. For Sarnak, the Apollonian gasket is wonderful because it has neither too few nor too many mirrors. If there were too few, you would not get enough information from the spectral decomposition. If there were too many, then previously known methods, such as the ones Wiles used, would already answer all your questions.

Because Apollonian gaskets fall right in the middle, they generate a host of unsolved number-theoretic problems. For example, which numbers actually appear as bends in a given gasket? These numbers must satisfy certain "congruence restrictions." For example, in the bugeye gasket, the only legal bends have a remainder of 2, 3, 6 or 11 when divided by 12. So far, it seems that every number that satisfies this congruence restriction does indeed appear in the figure somewhere. (The reader may find it amusing to hunt for 2, 3, 6, 11, 14, 15, 18, 23, etc.) "Computation indicates that every number occurs, but we can't prove that even 1 percent of them actually occur!" says Ron Graham of the University of California at San Diego. For other Apollonian gaskets, such as the "coins" gasket in the fifth figure, there are some absentees—numbers that obey the congruence restrictions but don't appear in the gasket. Sarnak believes, however, that the number of absentees is always finite, and beyond a certain point any number that obeys the congruence restrictions does appear somewhere in the gasket. At this point, though, he is far from proving this conjecture—the necessary math just doesn't exist yet.

And even if all the problems concerning the classic Apollonian gaskets are solved, there are still gaskets galore for mathematicians to work on. As mentioned before, they could study random Apollonian gaskets. Another modification is the gasket shown in the last figure, where each pore is filled by three circles instead of one. Mallows and Gerhard Guettler have shown that such gaskets behave similarly to the original Apollonian gaskets—if the first six bends are integers, then all the rest of the bends are as well. Ambitious readers might want to work out the "Descartes formula" and the "baby Descartes formula" for these configurations, and investigate whether there are congruence restrictions on the bends.

Perhaps you, too, will be inspired to write a poem or paint a tablet in honor of Apollonius' ingenious legacy. "For me, what's attractive about Apollonian gaskets is that even my 14-year-old daughter finds them interesting," says Sarnak. "It's truly a god-given problem—or perhaps a Greek-given problem."

Science Observer

A Helix with a Handle

Mathematicians prove the existence of a new class of minimal surfaces

Dip a loop of wire into a soapy solution, and the film that covers the loop will be what mathematicians call a minimal surface. The soap forms such a shape because it minimizes surface tension. At any point, a minimal surface is maximally curved in one direction and minimally curved in the opposite direction, but the amount of curvature in each direction is exactly the same. As a result, each point on the surface is either a flat plane or a saddle shape, never a sharp peak or valley. But a minimal surface doesn't have to be flat or simple overall: A plane can be twisted into a parking-ramp shape called a helicoid, which mathematicians proved over two centuries ago is also a minimal surface.

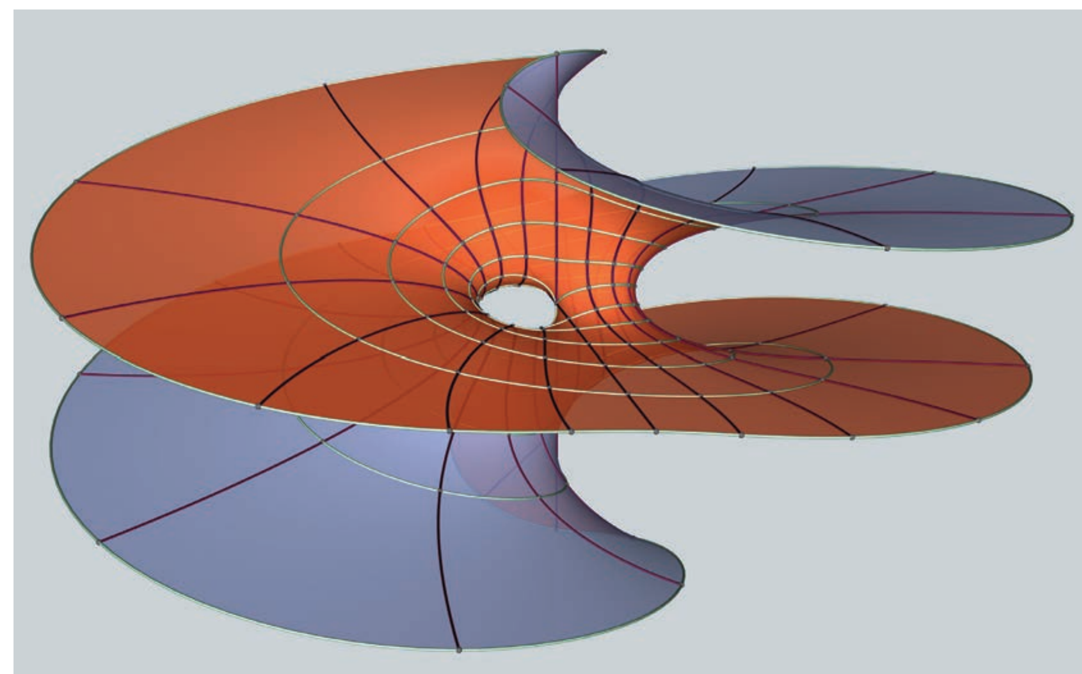
Mathematicians have proved the existence of a class of minimal surfaces that cannot be embodied by soap bubbles but can be visualized by computer simulation. This surface, called a genus-one helicoid, is a variation on a standard helicoid, but there is a tunnel through the deck of the parking-ramp spiral. When untwisted, this surface looks like a flat sheet with a coffee-mug-handle shape grafted onto it. "Think of a torus, like an inner tube," says Matthias Weber of Indiana University. "Now imagine that you puncture the torus. This results in a surface that can be stretched and deformed into the genus-one helicoid. I think that's a real mind bender."

As they reported in the November 15, 2005, issue of the *Proceedings of the National Academy of Sciences*, Weber, David Hoffman of Stanford University and Michael Wolf of Rice University have proven that such shapes, whether they have one or an infinite number of handles, are indeed minimal surfaces that can go on forever in all directions and never fold back to intersect themselves.

Over a decade ago, Hoffman, with Fusheng Wei, then of the University of Massachusetts at Amherst, and Hermann Karcher of the University of Bonn in Germany, had created computer simulations of such handled helicoids, but an airtight demonstration of minimal surfacehood eluded them. "Computer graphics programs enabled us to visualize these surfaces, but we couldn't bring them back into the mathematical fold," says Hoffman. "I think the information about how to solve this problem was lurking in the pictures all the time, but we just had to think about it for a long time and have the theory catch up with the evidence we had." Catching up can be hard to do: The mathematical proof takes up more than 100 pages.

An advanced understanding of minimal surfaces could be relevant to materials science; for instance, some compound polymers, such as Kevlar, have interfaces between molecules that are approximately minimal surfaces, the shape of which can influence the chemical properties of the material.

As mathematicians, Weber and his colleagues are most excited about a potentially large, new class of minimal surfaces that have not been found in nature and which no investigators had imagined could exist until recently. "It's easy to come up with one new example of a minimal surface, but this one is of a very different nature than others that have been found before," Weber said. "So it's opened a new field within the theory of minimal surfaces." —*Fenella Saunders*



Matthias Weber

A new type of minimal surface, found through computer simulation, will extend infinitely in all directions without crossing back to intersect itself. The distinct orange and blue colors indicate that the shape is a double spiral.

The Bootstrap

Cosma Shalizi

Statistics is the branch of applied mathematics that studies ways of drawing inferences from limited and imperfect data. We may want to know how a neuron in a rat's brain responds when one of its whiskers gets tweaked, or how many rats live in Manhattan, or how high the water will get under the Brooklyn Bridge, or the typical course of daily temperatures in the city over the year. We have some data on all of these things, but we know that our data are incomplete, and experience tells us that repeating our experiments or observations, even taking great care to replicate the conditions, gives more or less different answers every time. It is foolish to treat any inference from only the data in hand as certain.

If all data sources were totally capricious, there'd be nothing to do beyond piously qualifying every conclusion with "but we could be wrong about this." A mathematical science of statistics is possible because, although repeating an experiment gives different results, some types of results are more common than others; their relative frequencies are reasonably stable. We can thus model the data-generating mechanism through probability distributions and stochastic processes—random series with some indeterminacy about how the events might evolve over time, although some paths may be more likely than others. When and why we can use stochastic models are very deep questions, but ones for another time. But if we *can* use them in a problem, quantities such as these are represented as "parameters" of the stochastic models. In other words, they are functions of the underlying probability

Statisticians can reuse their data to quantify the uncertainty of complex models

distribution. Parameters can be single numbers, such as the total rat population; vectors; or even whole curves, such as the expected time-course of temperature over the year. Statistical inference comes down to estimating those parameters, or testing hypotheses about them.

These estimates and other inferences are functions of the data values, which means that they inherit variability from the underlying stochastic process. If we "reran the tape" (as Stephen Jay Gould used to say) of an event that happened, we would get different data with a certain characteristic distribution, and applying a fixed procedure would yield different inferences, again with a certain distribution. Statisticians want to use this distribution to quantify the uncertainty of the inferences. For instance, by how much would our estimate of a parameter vary, typically, from one replication of the experiment to another—say, to be precise, what is the root-mean-square (the square root of the mean average of the squares) deviation of the estimate from its average value, or the *standard error?* Or we could ask, "What are all the parameter values that *could* have produced this data with at least some specified probability?" In other words, what are all the parameter values under which our data are not low-probability outliers? This gives us the *confidence region* for the parameter—rather than a *point estimate,* a promise that either the true parameter point lies in that region, or something very unlikely under any circumstances happened—or that our stochastic model is wrong.

To get standard errors or confidence intervals, we need to know the distribution of our estimates around the true parameters. These *sampling distributions* follow from the distribution of the data, because our estimates are functions of the data. Mathematically the problem is well defined, but actually *computing* anything is another story. Estimates are typically complicated functions of the data, and mathematically convenient distributions all may be poor approximations of the data source. Saying anything in closed form about the distribution of estimates can be simply hopeless. The two classical responses of statisticians have been to focus on tractable special cases, and to appeal to asymptotic analysis, a method that approximates the limits of functions.

Origin Myths

If you've taken an elementary statistics course, you were probably drilled in the special cases. From one end of the possible set of solutions, we can limit the kinds of estimator we use to those with a simple mathematical form—say, mean averages and other linear functions of the data. From the other, we can assume that the probability distributions featured in the stochastic model take one of a few forms for which exact calculation *is* possible, either analytically or via tables of special functions. Most such distributions have origin myths: The Gaussian bell curve arises from averaging many independent variables of equal size (say, the many genes that contribute to height in humans); the Poisson distribution comes from counting how many of a large number of independent and individually improbable events have occurred (say, radium nuclei decaying in a given second), and so on. Squeezed from both ends, the sampling distribution of estimators and other functions of the data becomes exactly calculable in terms of the aforementioned special functions.

That these origin myths invoke various limits is no accident. The great re-

Cosma Shalizi received his Ph.D. in physics from the University of Wisconsin–Madison in 2001. He is an assistant professor of statistics at Carnegie Mellon University and an external professor at the Santa Fe Institute. Address: 132 Baker Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. Internet: http://www.bactra.org

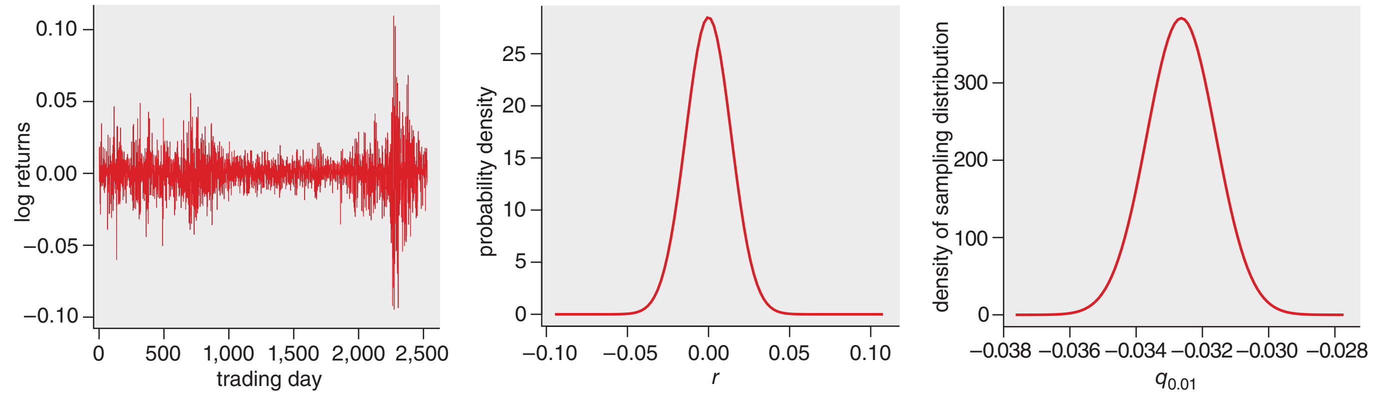


Figure 1. A series of log returns from the Standard and Poor's 500 stock index from October 1, 1999, to October 20, 2009 *(left)*, can be used to illustrate a classical approach to probability. A financial model that assumes the series are sequences of independent, identically distributed Gaussian random variables yields the distribution function shown at center. A theoretical sampling distribution that models the smallest 1 percent of daily returns (denoted as $q_{0.01}$) shows a value of -0.0326 ±0.00104 *(right)*, but we need a way to determine the uncertainty of this estimate.

sults of probability theory—the laws of large numbers, the ergodic theorem, the central limit theorem and so on—describe limits in which *all* stochastic processes in broad classes of models display the same asymptotic behavior. The central limit theorem (CLT), for instance, says that if we average more and more independent random quantities with a common distribution, and if that common distribution is not too pathological, then the distribution of their means approaches a Gaussian. (The non-Gaussian parts of the distribution wash away under averaging, but the average of two Gaussians is another Gaussian.) Typically, as in the CLT, the limits involve taking more and more data from the source, so statisticians use the theorems to find the asymptotic, large-sample distributions of their estimates. We have been especially devoted to rewriting our estimates as averages of independent quantities, so that we can use the CLT to get Gaussian asymptotics. Refinements to such results would consider, say, the rate at which the error of the asymptotic Gaussian approximation shrinks as the sample sizes grow.

To illustrate the classical approach and the modern alternatives, I'll introduce some data: The daily closing prices of the Standard and Poor's 500 stock index from October 1, 1999, to October 20, 2009. (I use these data because they happen to be publicly available and familiar to many readers, not to impart any kind of financial advice.) Professional investors care more about changes in prices than their level, specifically the *log returns*, the log of the price today divided by the price yesterday. For this time period of 2,529 trading days, there are 2,528 such values *(see Figure 1)*. The "efficient market hypothesis" from financial theory says the returns can't be predicted from any public information, including their own past values. In fact, many financial models assume such series are sequences of independent, identically distributed (IID) Gaussian random variables. Fitting such a model yields the distribution function in the center graph of Figure 1.

An investor might want to know, for instance, how bad the returns could be. The lowest conceivable log return is negative infinity (with all the stocks in the index losing all value), but most investors worry less about an

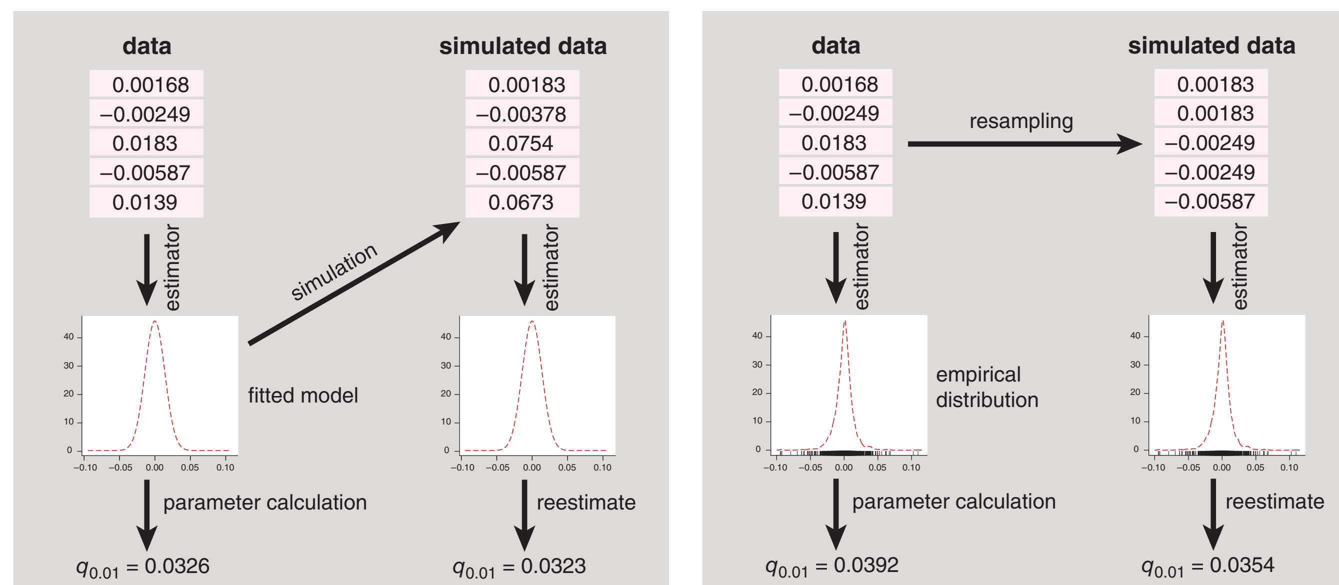


Figure 2. A schematic for model-based bootstrapping *(left)* shows that simulated values are generated from the fitted model, and then they are treated like the original data, yielding a new parameter estimate. Alternately, in nonparametric bootstrapping, a schematic *(right)* shows that new data are simulated by resampling from the original data (allowing repeated values), then parameters are calculated directly from the empirical distribution.

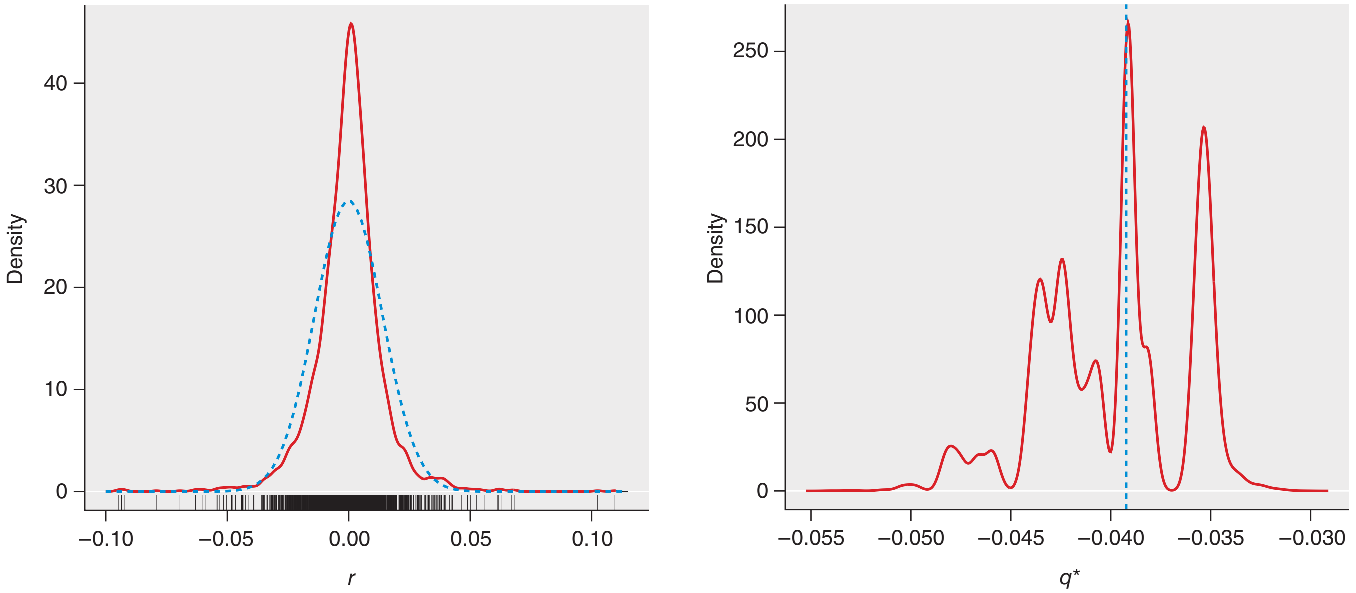


Figure 3. An empirical distribution ***(left, in red, smoothed for visual clarity)*** **of the log returns from a stock-market index is more peaked and has substantially more large-magnitude returns than a Gaussian fit** ***(blue).*** **The black marks on the horizontal axis show all the observed values. The distribution of $q_{0.01}$ based on 100,000 nonparametric replications is very non-Gaussian** ***(right, in red).*** **The empirical estimate is marked by the blue dashed line.**

apocalyptic end of American capitalism than about large-but-still-typical losses—say, how bad are the smallest 1 percent of daily returns? Call this number $q_{0.01}$; if we know it, we know that we will do better about 99 percent of the time, and we can see whether we can handle occasional losses of that magnitude. (There are about 250 trading days in a year, so we should expect two or three days at least that bad in a year.) From the fitted distribution, we can calculate that $q_{0.01} = -0.0326$, or, undoing the logarithm, a 3.21 percent loss. How uncertain is this point estimate? The Gaussian assumption lets us calculate the asymptotic sampling distribution of $q_{0.01}$, which turns out to be another Gaussian *(see the right graph in Figure 1),* implying a standard error of ± 0.00104. The 95 percent confidence interval is (–0.0347, –0.0306): Either the real $q_{0.01}$ is in that range, or our data set is one big fluke (at 1-in-20 odds), or the IID-Gaussian model is wrong.

Fitting Models

From its origins in the 19th century through about the 1960s, statistics was split between developing general ideas about how to draw and evaluate statistical inferences, and working out the properties of inferential procedures in tractable special cases (like the one we just went through) or under asymptotic approximations. This yoked a very broad and abstract theory of inference to very narrow and concrete practical formulas, an uneasy combination often preserved in basic statistics classes.

The arrival of (comparatively) cheap and fast computers made it feasible for scientists and statisticians to record lots of data and to fit models to them. Sometimes the models were conventional ones, including the special-case assumptions, which often enough turned out to be detectably, and consequentially, wrong. At other times, scientists wanted more complicated or flexible models, some of which had been proposed long before but now moved from being theoretical curiosities to stuff that could run overnight. In principle, asymptotics might handle either kind of problem, but convergence to the limit could be unacceptably slow, especially for more complex models.

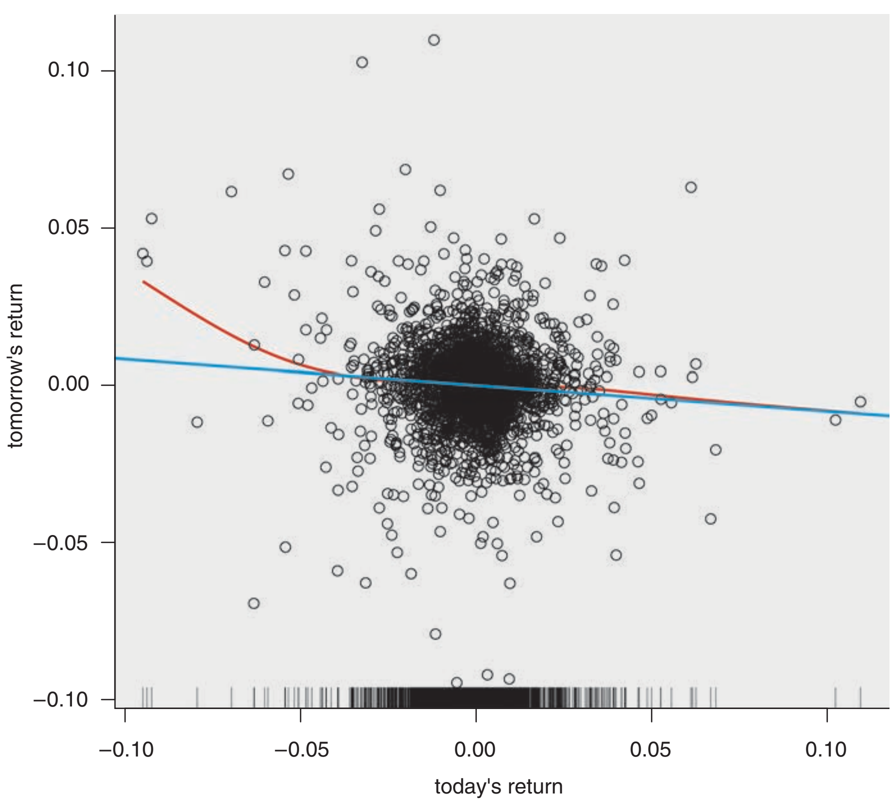


Figure 4. A scatter plot of black circles shows log returns from a stock-market index on successive days. The best-fit line ***(blue)*** **is a linear function that minimizes the mean-squared prediction error. Its negative slope indicates that days with below-average returns tend to be followed by days with above-average returns, and vice versa. The red line shows an optimization procedure, called** ***spline smoothing,*** **that will become more or less curved depending on looser or tighter constraints.**

By the 1970s statistics faced the problem of quantifying the uncertainty of inferences without using either implausibly helpful assumptions or asymptotics; all of the solutions turned out to demand *even more* computation. Perhaps the most successful was a proposal by Stanford University statistician Bradley Efron, in a now-famous 1977 paper, to combine estimation with simulation. Over the last three decades, Efron's "bootstrap" has spread into all areas of statistics, sprouting endless elaborations; here I'll stick to its most basic forms.

Remember that the key to dealing with uncertainty in parameters is the sampling distribution of estimators. Knowing what distribution we'd get for our estimates on repeating the experiment would give us quantities, such as standard errors. Efron's insight was that we can *simulate* replication. After all, we have already fitted a model to the data, which is a guess at the mechanism that generated the data. Running that mechanism generates simulated data that, by hypothesis, have nearly the same distribution as the real data. Feeding the simulated data through our estimator gives us one draw from the sampling distribution; repeating this many times yields the sampling distribution as a whole. Because the method gives itself its own uncertainty, Efron called this "bootstrapping"; unlike Baron von Münchhausen's plan for getting himself out of a swamp by pulling himself out by his bootstraps, it works.

Let's see how this works with the stock-index returns. Figure 2 shows the overall process: Fit a model to data, use the model to calculate the parameter, then get the sampling distribution by generating new, synthetic data from the model and repeating the estimation on the simulation output. The first time I recalculate $q_{0.01}$ from a simulation, I get -0.0323. Replicated 100,000 times, I get a standard error of 0.00104, and a 95 percent confidence interval of (–0.0347, –0.0306), matching the theoretical calculations to three significant digits. This close agreement shows that I simulated properly! But the point of the bootstrap is that it doesn't rely on the Gaussian assumption, just on our ability to simulate.

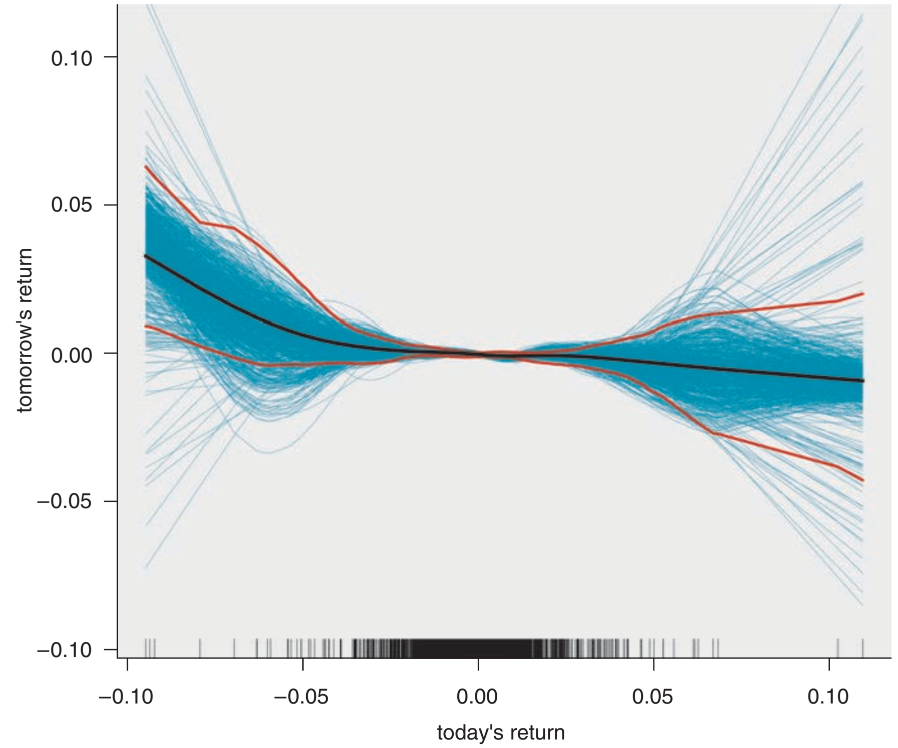


Figure 5. The same spline fit from the previous figure ***(black line)*** **is combined with 800 splines fit to bootstrapped resamples of the data** ***(blue curves)*** **and the resulting 95 percent confidence limits for the true regression curve** ***(red lines).***

Bootstrapping

The bootstrap approximates the sampling distribution, with three sources of approximation error. First there's *simulation error,* using finitely many replications to stand for the full sampling distribution. Clever simulation design can shrink this, but brute force—just using enough replications—can also make it arbitrarily small. Second, there's *statistical error:* The sampling distribution of the bootstrap reestimates under our fitted model is not exactly the same as the sampling distribution of estimates under the true data-generating process. The sampling distribution changes with the parameters, and our initial fit is not completely accurate. But it often turns out that distribution of estimates *around* the truth is more nearly invariant than the distribution of estimates themselves, so subtracting the initial estimate from the bootstrapped values helps reduce the statistical error; there are many subtler tricks to the same end. The final source of error in bootstrapping is *specification error:* The data source doesn't exactly follow our model at all. Simulating the model then never quite matches the actual sampling distribution.

Here Efron had a second brilliant idea, which is to address specification error by replacing simulation from the model with resampling from the data. After all, our initial collection of data gives us a lot of information about the relative probabilities of different values, and in certain senses this "empirical distribution" is actually the least prejudiced estimate possible of the underlying distribution—anything else imposes biases or preconceptions, which are possibly accurate but also potentially misleading. We could estimate $q_{0.01}$ directly from the empirical distribution, without the mediation of the Gaussian model. Efron's "nonparametric bootstrap" treats the original data set as a complete population and draws a new, simulated sample from it, picking each observation with equal probability (allowing repeated values) and then re-running the estimation (as shown in Figure 2).

This new method matters here because the Gaussian model is inaccurate; the true distribution is more sharply peaked around zero and has substantially more large-magnitude returns, in both directions, than the Gaussian *(see the left graph in Figure 3).* For the empirical distribution, $q_{0.01} = -0.0392$. This may seem close to our previous point estimate of –0.0326, but it's well beyond the confidence interval, and under the Gaussian model we should see values that negative only 0.25 percent of the time, not 1 percent of the time. Doing 100,000 non-parametric replicates—that is, resampling from the data and reestimating $q_{0.01}$ that many times—gives a very non-Gaussian sampling distribution (as shown in the right graph of Figure 3), yielding a standard error of 0.00364 and a 95 percent confidence interval of (–0.0477, –0.0346).

Although this is more accurate than the Gaussian model, it's still a really simple problem. Conceivably, some other nice distribution fits the returns better than the Gaussian, and it might even have analytical sampling formulas. The real strength of the bootstrap is that it lets us handle complicated models, and complicated questions, in exactly the same way as this simple case.

To continue with the financial example, a question of perennial interest is predicting the stock market. Figure 4 is a scatter plot of the log returns on successive days, the return for today being on the horizontal axis and that of tomorrow on the vertical. It's mostly just a big blob, because the market is hard to predict, but I have drawn two lines through it: a straight one in blue, and a curved one in black. These lines try to predict the average return tomorrow as functions of today's return; they're called *regression lines* or *regression curves*. The straight line is the linear function that minimizes the mean-squared prediction error, or the sum of the squares of the errors made in solving every single equation (called the *least squares* method). Its slope is negative (–0.0822), indicating that days with below-average returns tend to be followed by ones with above-average returns and vice versa, perhaps because people try to buy cheap after the market falls (pushing it up) and sell dear when it rises (pulling it down). Linear regressions with Gaussian fluctuations around the prediction function are probably the best-understood of all statistical models—their oldest forms go back two centuries now—but they're more venerable than accurate.

The black curve is a nonlinear estimate of the regression function, coming from a constrained optimization procedure called *spline smoothing:* Find the function that minimizes the prediction error, while capping the value of the average squared second derivative. As the constraint tightens, the optimal curve, the spline, straightens out, approaching the linear regression; as the constraint loosens, the spline wiggles to try to pass through each data point. (A spline was originally a flexible length of wood craftsmen used to draw smooth curves, fixing it to the points the curve had to go through and letting it flex to minimize elastic energy; stiffer splines yielded flatter curves, corresponding mathematically to tighter constraints.)

To actually get the spline, I need to pick the level of the constraint. Too small, and I get an erratic curve that memorizes the sample but won't generalize to new data; but too much smoothing erases real and useful patterns. I set the constraint through *cross-validation:* Remove one point from the data, fit multiple curves with multiple values of the constraint to the other points, and then see which curve best predicts the left-out point. Repeating this for each point in turn shows how much curvature the spline needs in order to generalize properly. In this case, we can see that we end up selecting a moderate amount of wiggliness; like the linear model, the spline predicts reversion in the returns but suggests that it's asymmetric—days of large negative returns being followed, on average, by bigger positive returns than the other way around. This might be because people are more apt to buy low than to sell high, but we should check that this is a real phenomenon before reading much into it.

There are three things we should note about spline smoothing. First, it's much more flexible than just fitting a straight line to the data; splines can approximate a huge range of functions to an arbitrary tolerance, so they can *discover* complicated nonlinear relationships, such as asymmetry, without guessing in advance what to look for. Second, there was no hope of using a smoothing spline on substantial data sets before fast computers, although now the estimation, including cross-validation, takes less than a second on a laptop. Third, the estimated spline depends on the data in two ways: Once we decide how much smoothing to do, it tries to match the data within the constraint; but we also use the data to decide how much smoothing to do. Any quantification of uncertainty here should reckon with both effects.

There are multiple ways to use bootstrapping to get uncertainty estimates for the spline, depending on what we're willing to assume about the system. Here I will be cautious and fall back on the safest and most straightforward procedure: Resample the points of the scatter plot (possibly getting multiple copies of the same point), and rerun the spline smoother on this new data set. Each replication will give a different amount of smoothing and ultimately a different curve. Figure 5 shows the individual curves from 800 bootstrap replicates, indicating the sampling distribution, together with 95 percent confidence limits for the curve as a whole. The overall negative slope and the asymmetry between positive and negative returns are still there, but we can also see that our estimated curve is much better pinned down for small-magnitude returns, where there are lots of data, than for large-magnitude returns, where there's little information and small perturbations can have more effect.

Smoothing Things Out

Bootstrapping has been ramified tremendously since Efron's original paper, and I have sketched only the crudest features. Nothing I've done here actually proves that it works, although I hope I've made that conclusion plausible. And indeed sometimes the bootstrap fails; it gives very poor answers, for instance, to questions about estimating the maximum (or minimum) of a distribution. Understanding the difference between that case and that of $q_{0.01}$, for example, turns out to involve rather subtle math. Parameters are functions of the distribution generating the data, and estimates are functions of the data or of the empirical distribution. For the bootstrap to work, the empirical distribution has to converge rapidly on the true distribution, and the parameter must smoothly depend on the distribution, so that no outlier ends up unduly influencing the estimates. Making "influence" precise here turns out to mean taking derivatives in infinite-dimensional spaces of probability distribution functions, and the theory of the bootstrap is a delicate combination of functional analysis with probability theory. This sort of theory is essential to developing new bootstrap methods for new problems, such as ongoing work on resampling spatial data, or model-based bootstraps where the model grows in complexity with the data.

The bootstrap has earned its place in the statistician's toolkit because, of all the ways of handling uncertainty in complex models, it is at once the most straightforward and the most flexible. It will not lose that place so long as the era of big data and fast calculation endures.

Bibliography

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7:1–26.

First Person: Tim Davis

Recognized internationally for his innovative algorithms and software, Tim Davis's research has been applied in a vast range of real-world technological and social applications, from MATLAB to Google Street View, from aiding the FBI to creating original art. He is the recipient of the 2018 Sigma Xi Walston Chubb Award for Innovation, an annual award established in 2006. Previous awardees include engineer Akhlesh Lakhtakia, computer scientist Rosalind W. Picard, and materials scientist Stan Ovshinsky, among others. American Scientist*'s digital managing editor, Robert Frederick, spoke with Davis about the software tools he creates, which he describes as "mathematical monkey wrenches," some of which he has used to create artworks based on music.*

What prompted you to make these mathematical tools in the first place? Was it a problem you were trying to solve, or something else?

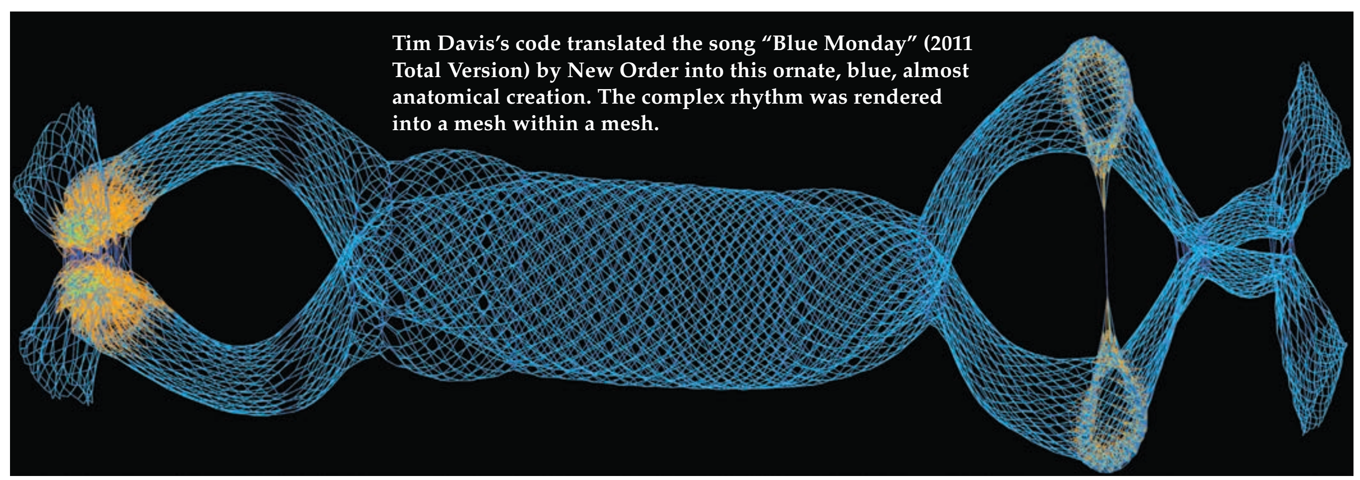
I did all my degrees in electrical engineering and was also interested in computer architecture. But all along I was interested in software—just writing the code—all the way back to being a high school student. In 1978, maybe, I was a sophomore or a junior in high school. My brother was an undergraduate in mechanical engineering at Purdue University and said, "Oh, Tim, here, I'll get you an account on the Purdue mainframe. Here's a book on FORTRAN. Have at it." And so I did. I went through all the problems in the book and typed up my punch cards and went to the mechanical engineering lab and stuck in my cards and computed pi or whatever from the projects in the book. I had fun. By the time I was done, I'd worked through the whole book, and that next summer I worked as a consultant in the computer room with all the other students. I was helping students with their Runge–Kutta differential equation solvers. I had no idea what Runge–Kutta was, but I could help them with their FORTRAN. And being really focused on finding programming bugs in other people's code—it's a fun puzzle—is a great way to learn how to find bugs in your own code.

So what was it that led to your work with matrix computations, now powering the "backslash" command in MATLAB, for example?

Going through my undergraduate at Purdue, I wasn't doing linear algebra. I wasn't doing a lot of matrix computation. But I was still writing some software. Then when I was in grad school at the University of Illinois, with an interest in computer architecture, I was looking at how to get the data to the processor faster and I came up with some ideas. Then I thought, "Well, it's hard to think about that in isolation, so let's look at an algorithm—why not?" I haven't stopped looking, basically. I came up with some ideas of how to move the data around a little better for that algorithm. Eventually, my thesis turned into one minor section on architecture and just about all of it on a graph and a matrix algorithm. Then, for a postdoc, I had the opportunity to work with Iain Duff, who's one of the top experts in the field of sparse matrix computations. He's in southern France, so I spent a year in Toulouse, which was a lot of fun. That's where I started working on the algorithm that, 10 years later, became "backslash" in MATLAB.

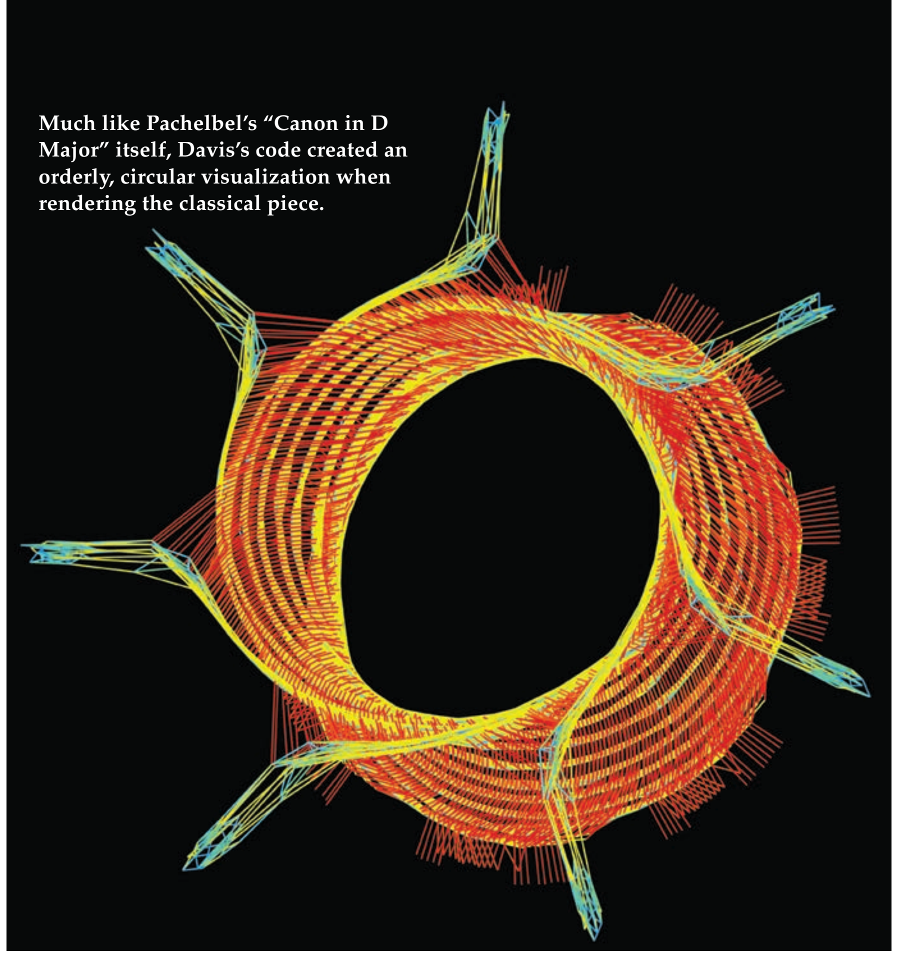
Was that when you first started explicitly making algorithms for other people, for problems that weren't your own?

Well, already as a PhD student, I wrote an algorithm that then didn't solve anybody else's problems. It was a matrix solver. But it was as a postdoc that I started writing another solver when I didn't have a matrix I wanted to solve. So yes, at that point, I was creating algorithms for other people, like Iain Duff does. He creates algorithms to solve matrix problems, not because he has matrices to solve, but because he's



Tim Davis's code translated the song "Blue Monday" (2011 Total Version) by New Order into this ornate, blue, almost anatomical creation. The complex rhythm was rendered into a mesh within a mesh.

Robert Frederick; Tim Davis



Much like Pachelbel's "Canon in D Major" itself, Davis's code created an orderly, circular visualization when rendering the classical piece.

a computational mathematician. So I'm doing like him, basically. I'm creating solvers that other people will be able to use. It's viewed in my domain as an academic contribution just like writing a paper is, or discovering a new galaxy would be to an astrophysicist.

And all of it is about solving the matrix equation, AX=B?

Yes! Isn't that crazy? It's just all AX=B. I don't care what X is, really. I don't. You give me A and B, I don't care where your A and B come from. But I want to give you the right X. I don't have a personal attachment to X [laughs]. I just know I want to give you the right one. And that's why we should care that code is written well, because if it's written properly and elegantly and is easy to read and understand, then other people can understand and rely on these solvers to do things that are important to them.

People out there are using my work to build power networks and circuits, and fly drones, and even rescue girls from the sex-slave trade. Seeing the tool out there and getting used is really heartwarming. Google used my code for a year to place all the photos in Street View without even asking me how my code works and without even needing to tell me. They download it and use it and there it is. All these crazy things that I have no idea how it is that they're doing that, except that it is AX=B somewhere.

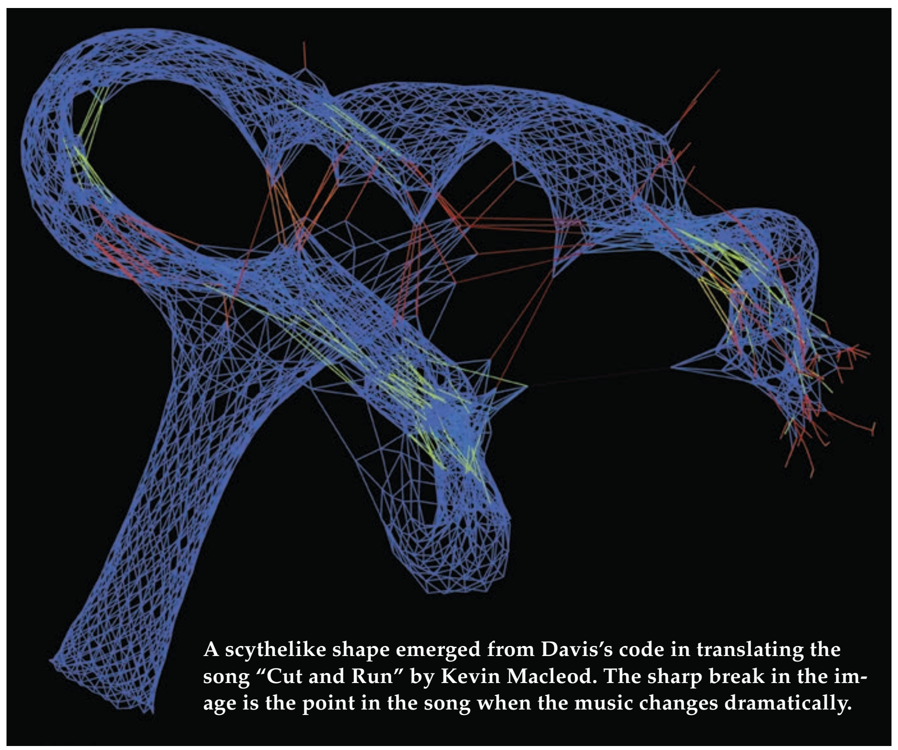
Now, the more recent stuff I'm doing—that's coming out of this group consortium, GraphBLAS.org—we're not solving AX=B, surprisingly. It's things such as matrix multiply. So it's still linear algebra, and yet it's beautiful, and it's solving important problems, and, if you do it well and do it right and do it fast—asymptotically fast and fast in practice—then the world will use it, and then the world will solve problems with it. That's fun.

Why do you say "It's beautiful"?

For me, I see code—and people like me see code—as kind of like the proof of a mathematical theorem. And just like mathematicians might publish many versions of their proofs, there's value in refinement. Turning a complicated, 10-page proof into an elegant, more understandable, more powerful 1-page proof that also proves the same theorem might open doors to other methods that could be used to solve other problems. Just like what mathematicians do for mathematical proofs, we do this for software.

Once you build one of these tools, do you later see further elegance—further improvements—and so go back to work on that tool? Or do you say, "That tool is good enough—done—I'm moving on to something else"?

There's really multiple aspects to that question. When I write a piece of software and publish it, I get a paper out of it. I get people using it. It builds impact, which is good. But then it can't stop there, can it? My code—all code—always has to be updated and maintained at some level. And then someone sends me an email and says, "Hey, I'm using your code to do this, and what about that?" "Oh, yeah," I'll say, "that would be easy, I'll add that!" So I'll add a new feature or do something different with it. I'll make it easier to compile or run, and update the user guide. Yes, I could say, "Okay, I'm done. I'll stop." But there's almost always more things to do. So I maintain the code, even though it's published and done. I need to periodically bring the car into the shop and tinker with it, right?



A scythelike shape emerged from Davis's code in translating the song "Cut and Run" by Kevin Macleod. The sharp break in the image is the point in the song when the music changes dramatically.

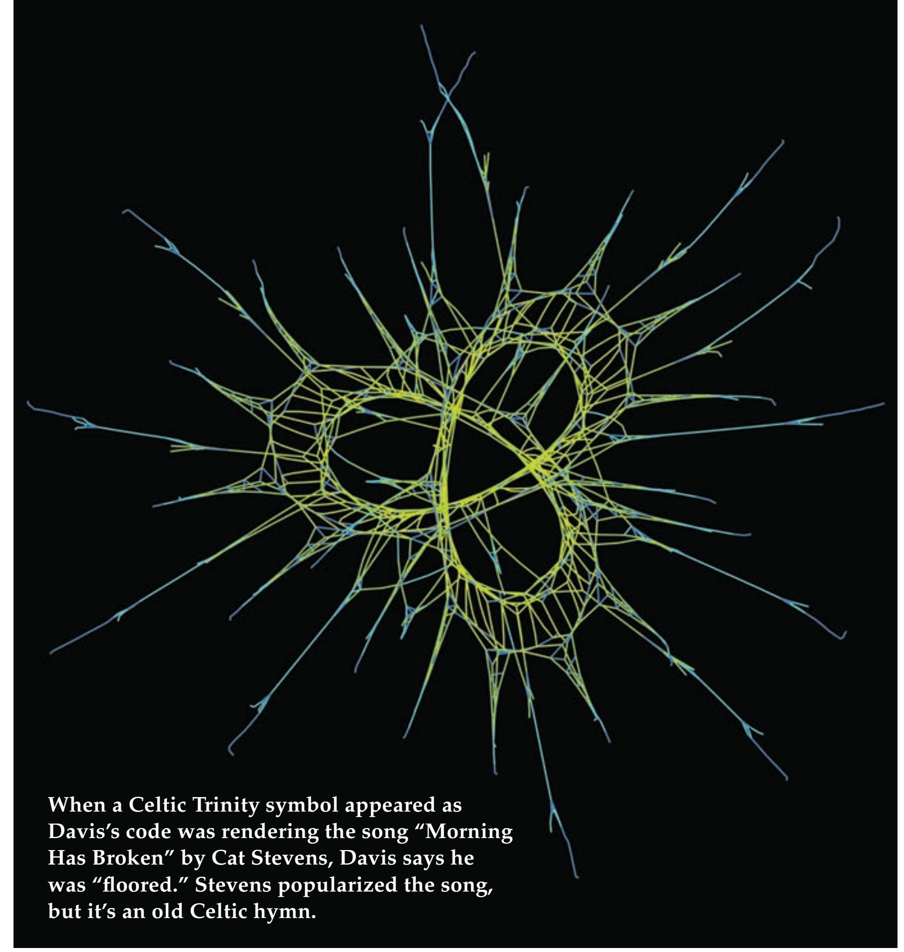
Tim Davis

Tim Davis

What kind of support, if any, do you get as an academic to do this kind of software-maintenance work?

Even if I don't get a paper out of it, it's still worth my time doing the software maintenance, if you will, because I've got users to keep happy, basically. I don't want to let them down. But, in general, building tools as an academic like this—and there are other computer scientists who are tool builders like me—we face a risk. That's because writing really good software, software that is better than commercial or government quality, is crazy as an academic, because the payoff is only the impact: It's a valuable contribution, from creating new algorithms and new methods, to making these methods fast—10,000, 5,000, 100 times faster than prior methods for some of these problems, although sometimes a factor of two is all I get, or sometimes it ties.

Of course, if you have that impact, that shows academic credibility to your algorithms and your methods and your work. But what if you build this great mathematical wrench, but it's just not the very best wrench that's out there? Or what if the world doesn't actually need that kind of wrench and so people don't use it? I think I have around 40 journal publications, which in some domains is minuscule. But one paper has 300 pages of bulletproof code behind it. That's a hefty contribution. But if that code is not used, then it won't have the impact and I'd only gotten one paper out of it, if that. So it's risky, and I kind of barely squeaked through my earlier promotions. But now at Texas A&M University, my department head, who appreciates my work, she says, "Tim, you know you took a crazy risk to do all this?" Yeah, but I loved it, and I still do.



When a Celtic Trinity symbol appeared as Davis's code was rendering the song "Morning Has Broken" by Cat Stevens, Davis says he was "floored." Stevens popularized the song, but it's an old Celtic hymn.

What about your artwork and using your code to turn music into visual art? Do you polish that artwork too?

The code itself to create that artwork is actually fairly rough. It's what I would call my own "internal prototype." I'm not distributing the code, though. I'm distributing the art. The software—

People out there are using my work to build power networks and circuits, and fly drones, and even rescue girls from the sex-slave trade. Seeing the tool out there and getting used is really heartwarming.

that tool—is my paintbrush. So when I use it to render a visualization based on a new piece of music, sometimes I'll decide, "Nah, I need something different. Let me think about these rules I've got and let me add to them." I'll modify my software to create different kinds of visualizations.

So there's an aesthetic, meaning that you're looking at these images that are generated across various parameter spaces and saying, "Okay, that one appeals to me aesthetically, I'll chose that one." Or you'll tweak the code if you're not happy with any of them?

Yes, exactly. It's almost like a photographer. What'll happen is I'll do this image sweep and look across it and say, "Oh my goodness, I've never seen this before." I just rendered Cat Stevens's rendition of "Morning Has Broken," which is a Celtic hymn, and I came up with a symbol, a graph, that looks like Celtic Trinity symbol. I was floored. It's usually like seeing images in clouds. They're not there, but your mind puts them there, and then that's the image. Now, I have to create 1,000 clouds, and some of them are beautiful. Many of them are beautiful. But then I'll say, "That's a particularly beautiful cloud for that piece of music."

Online: Hear the music that was rendered by Davis's code into these visualizations, and see a video of a visualization being made in real time.

Recreational Computing

Erik D. Demaine

MARTIN GARDNER was a great man of many talents. He was an amateur mathematician, a puzzler, a professional magician, a debunker of pseudoscience, and a popular writer about all of these topics. He wrote more than 65 books and published a column, "Mathematical Games," in *Scientific American* for 25 years, from 1957 to 1982. Because of his influence on countless readers, Gardner became known as the father of "recreational mathematics"—playful mathematical problems designed and solved purely for fun. Gardner's accessible, inviting prose and his ability to correspond with impressive numbers of readers gave the general public the opportunity to enjoy mathematics and to participate in mathematical research. Many of today's mathematicians, including myself, entered the field at least in part due to Gardner's influence.

Sadly, Gardner died on May 22, 2010, at the age of 95. His death has been sorely felt by mathematicians around the world. But rather than dwell on our loss, I feel compelled to celebrate the tradition that Gardner started. Roughly every two years since 1993, Tom Rodgers has organized a conference in Atlanta called the Gathering for Gardner. It brings together mathematicians, puzzlers, magicians and debunkers who love the work of Martin Gardner and the spirit he embodied—playful intellectual curiosity. Gardner's own absence from the Gathering since 1996 has not stopped it from continually growing in participation and intensity. The ninth Gathering, held last March, was the most prodigious yet, with 300 participants, a half-day sculpture-building party and two evening magic shows.

Erik Demaine is an associate professor in computer science at the Massachusetts Institute of Technology. Address: MIT CSAIL, 32 Vassar St., Cambridge, MA 02139. Internet: http://erikdemaine.org/

All illustrations by Tom Dunne

Puzzles and tricks from Martin Gardner inspire math and science

I am a theoretical computer scientist, which puts me at the boundary of computer science and mathematics. The goal of the field is to use mathematics to understand computation—what it is and what it can do. Readers of this column already know that computation is extremely powerful, offering new perspectives, approaches and solutions in perhaps every discipline. Computer science is highly unusual in this universality of influence—the only other example I know of is mathematics—and it's what excites me about the field. The interdisciplinary field of "computational x" is central to most fields where it has been considered (the "x" could be biology, chemistry, neuroscience, geometry, linguistics, finance, and so on)—and for other fields, I believe it is simply yet to be discovered.

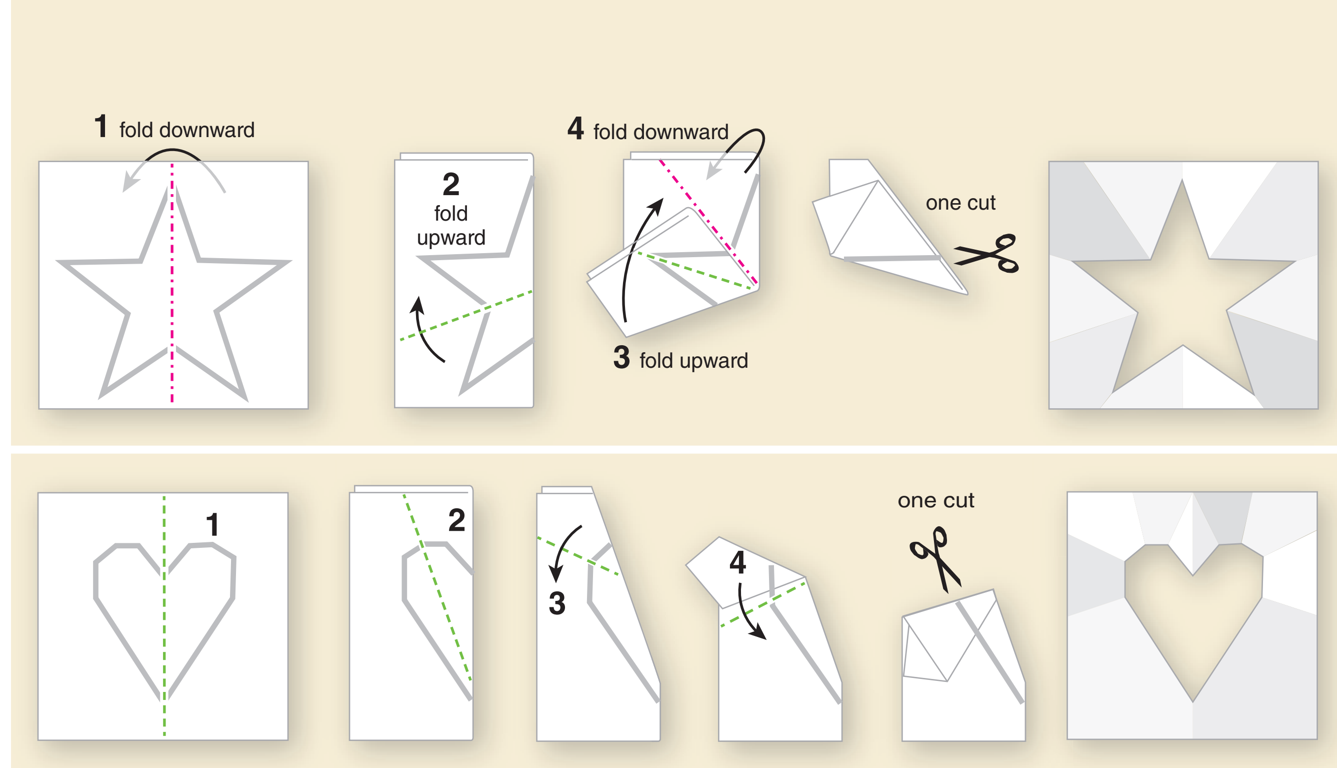
What I'd like to show here is that computation is a useful way to think about more recreational pursuits, too—specifically puzzles and magic. Martin Gardner is my inspiration: He did not consider puzzles, magic and mathematics as separate pursuits, but blurred the traditional boundaries between them. He routinely illustrated mathematics using puzzles and magic, and he studied puzzles and magic using mathematics. I like to apply the same spirit to theoretical computer science, where the computational perspective offers new ways to think about puzzles and magic—specifically, how to design challenges and tricks automatically. Voilà, recreational computer science!

Gardner's work continues to influence researchers such as myself. The three examples I'll describe are solutions to problems that Gardner posed—ones he stated explicitly or ones that have been inferred from his work. Throughout Gardner's writings are countless mathematical questions, puzzles and magic tricks that deserve further research and extension. I encourage everyone to read through his collected works, for the fun this always brings, as well as to help find these seeds for future research. I will collect your suggestions, which you can send to martingardner@csail.mit.edu. Long live the spirit of Martin Gardner!

One-Cut Magic

Our first example of recreational computer science is inspired by a magic trick performed by Harry Houdini before he was an escape artist. In his 1922 book *Paper Magic*, he describes how to take a standard sheet of paper, make a sequence of folds, cut along a straight line, unfold, and obtain a perfect five-pointed star *(see the top of the first figure, next page)*. In 1960, Gardner wrote a *Scientific American* column that described a few such magic tricks, producing "simple geometrical figures" by a sequence of folds and one complete straight cut.

Gardner included the tantalizing statement: "More complicated designs … present formidable problems." To a theoretical computer scientist such as myself, this screams "Unsolved problem!" Whenever we have several examples of a particular style of magic trick, it is natural to wonder whether it represents a general principle. In this case, we know how to make several simple figures by folding and making one complete straight cut. But what is the complete range of figures that are possible to make in this way, and what sequence of folds will make them? To put it more computationally, can a computer algorithm tell us whether a particular shape can be made, and if so, tell us where to fold and cut to make it?



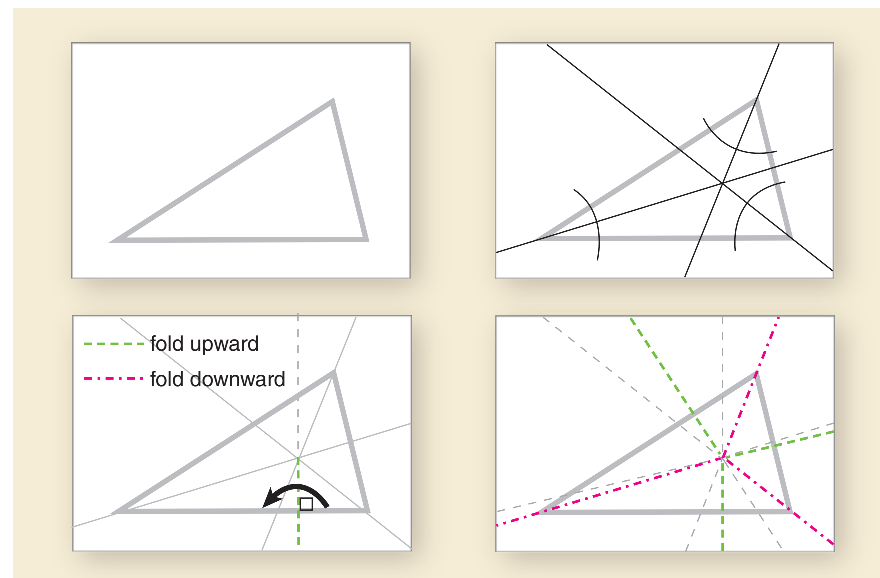
Before he was an escape artist, Harry Houdini described how to fold a piece of paper so that a five-pointed star could be made from it with one cut *(top)*. Martin Gardner wrote about such one-cut magic tricks in 1960. The author and his colleagues developed a similar series of folds to produce a heart shape with one cut *(bottom)*. Red lines indicate where the paper is folded downward so that the creased fold points towards you (called a mountain fold); green lines show an upward fold of the paper, resulting in a creased fold pointing away from you (a valley fold).

I started thinking about this problem in 1996 when I was a starting Ph.D. student at the University of Waterloo in Canada. My father, Martin Demaine, then an artist and an avid puzzler and now also a mathematician, had read Martin Gardner's article back in 1960, remembered the problem and suggested we work on it. So we tackled it, together with my Ph.D. advisor Anna Lubiw, and essentially solved it two years later.

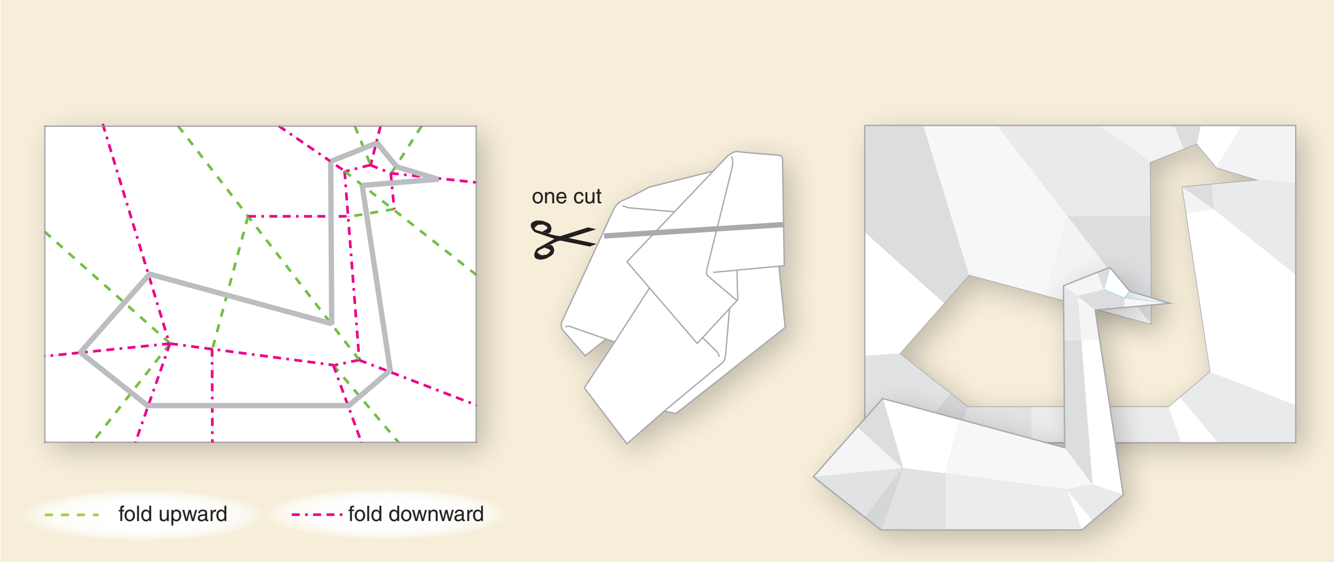
Initially we experimented, folding and cutting lots of pieces of paper. One of our first recognizable creations was a heart shape *(see the bottom of the first figure, above)*. This shape, like the five-pointed star, has a line of reflectional symmetry—the center line along which the pieces on either side are mirror images of each other—which is the first fold to make. Why should we fold along symmetry lines? Our effective goal is to fold the paper to *align* the sides of the shape we want to cut out; once the sides lie along a common line, we simply cut along that line. So if the shape has a line of symmetry, we should fold along it, aligning the two halves of the shape, leaving just one half to align.

What if the shape we want doesn't have a line of symmetry? This is where the problem gets really interesting. Initially we thought that such shapes were impossible: What else could the first fold be? But as we learned more about origami mathematics, which at the time was just making its debut in the scientific literature, we realized that there were many more types of folds than just "fold along a line." Armed with the idea of such complex flat folds, we quickly made progress on solving the problem, building general techniques for producing more and more shapes.

One early step in this progression was making any triangle *(see the second figure, below)*. To try this out, draw the three sides of a triangle on a piece



Shapes do not need to have a line of symmetry in order to be folded and produced with one cut. The author and his colleagues were able to demonstrate a method by which any triangle can be produced. Gray and black lines are for visual reference; only colored lines are folded.



The fold-and-one-cut method can be used to produce shapes of theoretically limitless complexity. One such shape is a swan, with fold lines as indicated at left, leading to the collapsed paper with one cut line at center, and resulting in the final figure at right. To download and print this example, go to http://amsci.org/swan.pdf. For other examples, see the website http://erikdemaine.org/foldcut/.

of paper. Now find the lines that bisect each of the angles of the triangle by starting at one of its corners, folding one edge on top of another, unfolding, and repeating for the two other corners. A classic theorem from high-school geometry is that the three angular bisectors meet at a common point. Now fold along a line through this point so as to bring one of the three triangle sides onto itself, and unfold. The fold will be perpendicular to the triangle side. If you like, you can repeat with the two other sides, though only one perpendicular fold is necessary. The final step, which is the hardest if you've never made a "rabbit ear" in origami, is to fold along all the creases at once, with the angular bisectors folding one way (called "mountain" folds) and the perpendicular ones folding the other way (called "valley" folds). Once you do this, all three triangle sides lie along a line, with the inside and outside of the triangle lying on opposite sides of that line. What's cool is that this works for any triangle—no line of symmetry required.

To our surprise, we also found a way to make any polygon with any number of sides, not just triangles. So you can make the silhouette of your favorite shape, such as the swan in the third figure *(above)*, by folding and then making one complete straight cut. The fold is substantially more complicated, but not too hard with practice. I recommend precreasing—folding and unfolding each crease line—before attempting to fold all the creases simultaneously and collapse the swan down to a line.

And there's more: You can make several polygons at once with a single cut. This result is especially useful for spelling initials, such as the ones that I prepared for my first Gathering for Gardner (the fifth one, abbreviated G4G5) in 2002. It's rather difficult to fold anything beyond a few letters, but in principle you could fold a piece of paper, make one complete straight cut, and produce the entire Gettysburg Address.

Although this research was motivated by magic tricks, the theoretical computer science that resulted turns out to have more serious applications as well. A closely related problem is that of compactly folding an airbag flat so that it fits into a small compartment such as your steering wheel. The solution to the fold-and-cut problem leads to a natural way to collapse three-dimensional surfaces such as airbags, and this approach has been applied in some simulations of airbag deployment. So recreational computer science may even save lives.

Coin-Sliding Puzzles

Our next example of recreational computer science is inspired by a collection of classic "penny puzzles," which Gardner wrote about in *Scientific American* in 1966. Helena Verrill, a mathematician now at Louisiana State University, introduced my father and me to puzzles of this type. You are given a pyramid of six coins, shown on the left in part *a* of the fourth figure *(next page)*, and your goal is to make a sequence of moves to arrange the coins into a line. But each move you make must place an existing coin into a position that touches at least two other coins. This "two-adjacency" constraint is what makes the puzzle challenging. It might initially seem impossible to get rid of all the three-coin triangles, but with a little insight, it is possible.

There are several coin-sliding puzzles of this type, with the two-adjacency constraint on moves. So again we wondered: How does it generalize? The natural computational question here is whether computers can solve this type of puzzle: Given a starting configuration and a goal configuration of coins, can a computer algorithm determine whether the puzzle is solvable under the two-adjacency constraint, and if so, find a sequence of moves to solve it? Even better, can it find the shortest sequence of moves to solve the puzzle?

Both of these questions are actually still unanswered. I suspect that the second question has a negative answer, and it is computationally intractable to determine the fewest moves needed to solve a puzzle, but no one has proved that yet. However, the first question might have a positive answer, which would be much more interesting—it would essentially provide a general theory for this type of puzzle.

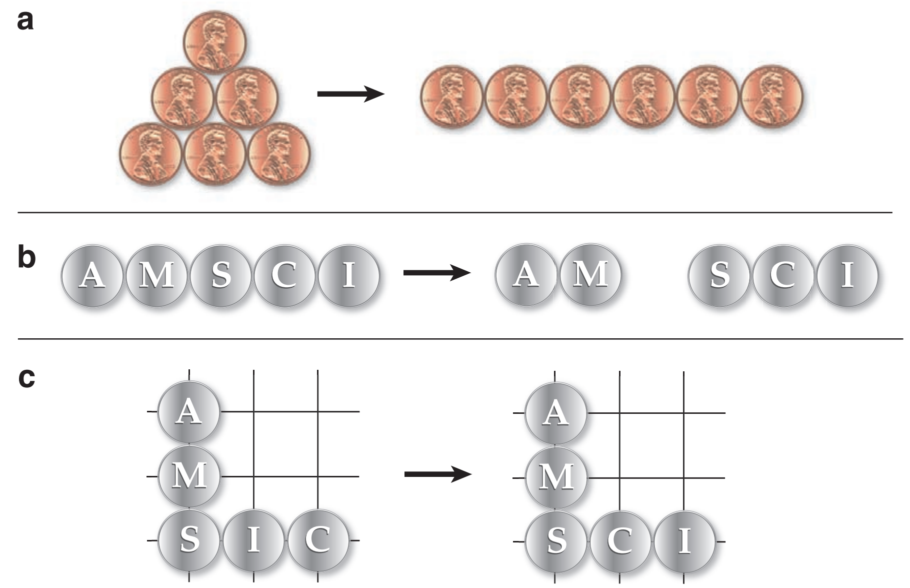
In 1998 Verrill visited my father and me for a couple of days of problem solving, during which we came up with and tackled the first problem. We observed that the puzzle in the fourth figure adheres to a triangular grid: If you draw an equilateral triangular grid where the triangle side length equals the diameter of the coin, then the centers of the coins always remain at grid intersections. This property is a neat consequence of the two-adjacency rule because the coins start on this grid.

For puzzles on the triangular grid, we were able to develop a general computational theory. We found some very simple conditions for when it is possible to transform a starting configuration into a goal configuration via two-adjacency moves. First, the number of coins must be the same in the two configurations, and it must be possible to make at least one two-adjacency move from the starting configuration. Second, and more interesting, the goal configuration must have at least one of three patterns that make it possible to have a last move that solves the puzzle. The patterns are a triangle of three coins, a connected group of four or more coins, and a connected group of three coins along with another connected group of two coins.

As long as the goal configuration has at least one of these patterns, the puzzle is guaranteed to be solvable, and a computer algorithm can tell you how. This result holds even if the coins have different labels (for example, some are heads and some are tails), provided there are more than three coins total.

This kind of simple characterization of solvable puzzles not only helps to find outcomes, but also makes it easy to design new puzzles. Armed with a guarantee about when a puzzle will be solvable, we can design a new one within those constraints that will also be visually interesting. For example, can you solve the puzzle in part *b* of the fourth figure?

Another special type of coin-sliding puzzle arises if we replace the triangular grid with a square grid, requiring every move to place a coin on this grid. These puzzles are substantially harder to solve, both in practice and in theory. We came up with a nearly complete characterization of solvable puzzles, showing how to complete challenges with at least two "extra coins" to help navigate the other pieces. Puzzles without any extra coins are impossible to solve, and puzzles with just one extra are typically either unsolvable or not too engaging—but we still don't have a computer algorithm to tell us exactly which puzzles can be solved. The situation with two extra coins is quite interesting, however, and has again allowed us to design several puzzles. Try, for instance, the puzzle in part *c* of the fourth figure. To see more coin-sliding puzzles, visit http://erikdemaine.org/slidingcoins/.



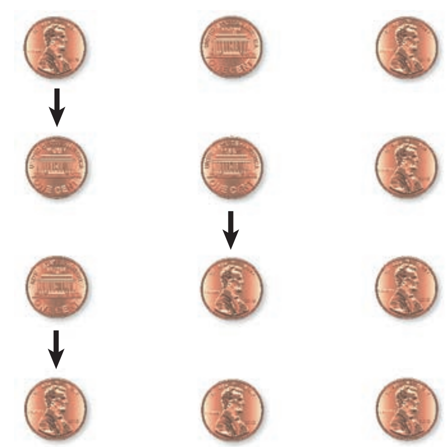
Reconfiguring coins becomes an interesting challenge when there are constraints. For instance, can you transform a pyramid of coins *(a)* into a line using only moves that place coins in contact with at least two other coins? The minimum possible number of moves is seven; a solution is shown on the next page. Similarly, can you split the abbreviation of this magazine's title *(b)* into two words using only two-adjacency moves? Again, the minimum number of moves is seven. Finally, in a new coin-sliding puzzle, see if you can correct the spelling of the abbreviation *(c)* using only two-adjacency moves on a square grid. This time the minimum number of moves is eight. To see more puzzles of this type, go to http://erikdemaine.org/slidingcoins/.

Coin-Flipping Magic

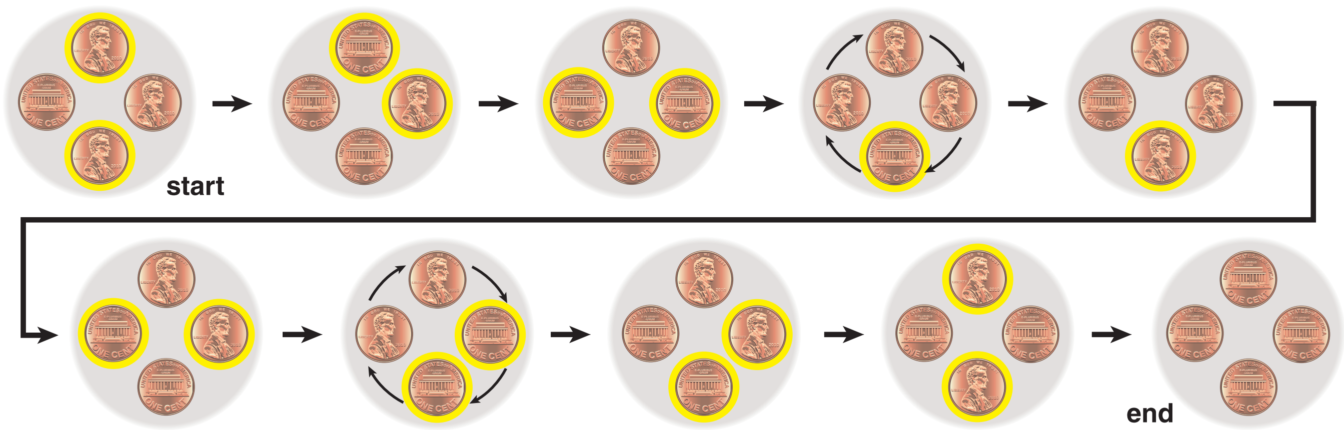
Our last example of recreational computer science is inspired by two tricks with a blindfolded magician, which involve flipping coins into a desired configuration. Because a volunteer manipulates the coins, these tricks have the distinction of being performable over the telephone, on radio or on television, thus being performed "personally" for many people at once.

The first trick, independently invented by Martin Gardner and Karl Fulves before 1980, involves three coins arranged in a line. The spectator arranges the coins as heads or tails, in any combination they like. The magician's goal—without seeing the coins—is to make them all the same—all heads or all tails. Naturally, the spectator should not choose this outcome as the starting configuration, or else the trick will be over rather quickly. The magician now gives a sequence of instructions: Flip the left coin, flip the middle coin, and flip the left coin again. In between, the magician asks whether the coins are yet all the same, and continues to the next instruction only if the trick is not yet over. It's no surprise that the magician can eventually equalize all the coins, but it's impressive that it always takes at most three moves *(see the fifth figure, at right)*.

Why these three moves work is closely linked to a set of codes in computer science that are widely used today to reduce errors when representing digital data with analog signals. These so-called Gray codes, named after Frank Gray from Bell Labs, who patented the system in 1947, have the feature that every two successive binary values differ by only one bit. A con-

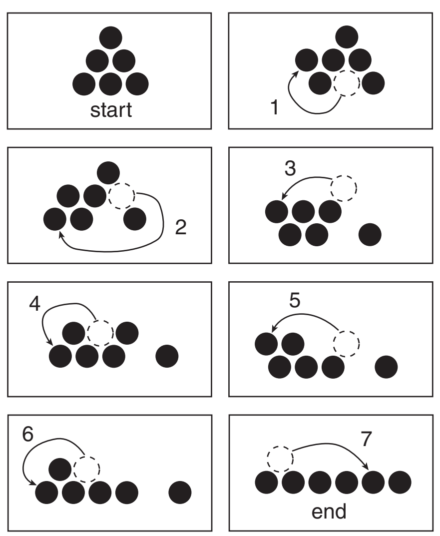


Three coins are arranged in a line *(top row)* in any starting sequence. A blindfolded magician can make them all heads or tails in at most three moves *(black arrows)*. Here the instructions are to flip the left coin *(second row)*, then the middle coin *(third row)*, then flip the left coin again *(fourth row)*, resulting in all heads.



Four coins arranged in a circle are a variation on the previous coin-flipping trick. The magician tells a volunteer which coins to flip, but this time, before each move, the volunteer can rotate the circle of coins however he or she likes. The puzzle starts at the top left then continues on the second line. Yellow outlines indicate the coins to be flipped in the next move. In this case the volunteer rotates the coins only between steps two and three, and between steps four and five. The magician still accomplishes the trick in seven moves.

figuration of three coins can be seen as a corner of a three-dimensional cube: There are three coins, and each can be either heads or tails, making $2^3=8$ corners. But the cube is effectively folded in half because the all-heads configuration is just as good as the all-tails configuration, leaving just $2^2=4$ “double configurations” arranged in a two-dimensional square. The Gray code tells us how to traverse all these nodes by changing one coin at a time, without ever repeating a configuration. Because we are counting moves instead of configurations visited, we get to subtract 1, for a total of 3 moves. More generally, if we had n coins, we'd have 2^n configurations, 2^{n-1} double configurations, and $2^{n-1}-1$ moves in the worst case. This study makes it clear why the trick is for three coins: Four coins would already require $2^3-1=7$ moves.



A solution for the first coin-sliding puzzle, shown on the previous page, transforms a pyramid of coins into a line. With the rule that each move must place a coin so it touches at least two others, seven moves are required.

To make the trick more impressive, we can give the spectator more freedom. A 1979 letter from Miner Keeler to Gardner, which Gardner wrote about in *Scientific American* in the same year, describes a trick involving four coins arranged in a circle. The setup and goal of the trick are the same as before, but this time the spectator can rotate the circle of coins however he or she likes before following each of the magician's instructions. The magician follows these seven steps: Flip the top and bottom coins (after rotation), flip the top and right coins (after rotation), flip the left and right coins (after rotation), flip the bottom coin (after rotation), flip the left and right coins (after rotation), flip the right and bottom coins (after rotation) and flip the top and bottom coins (after rotation). Despite the spectator's apparent flexibility, the magician equalizes all the coins in no more moves than the original four-coin trick (*see the sixth figure, above*).

What makes this trick possible? A recent paper by two MIT students, Nadia Benbernou and Benjamin Rossman, along with my father and myself, analyzes what type of spectator moves such as “rotate the table” still let the magician equalize the coins with a clever choice of moves. The solution is closely tied to group theory, a field crucial to modern cryptography. The key requirement turns out to be that the number of different moves that a spectator can make is a power of 2. In the case of a rotating table, the number of coins must be a power of 2 (as in the four-coin trick above). But it is also possible, for example, to allow the table to be flipped over, because this precisely doubles the number of possible spectator moves.

Like many people, I love puzzles and magic. I also love theoretical computer science. It is wonderful to combine these loves, following in the footsteps of Martin Gardner, and I hope that more computer scientists will consider the recreational side as a good source of fun problems to solve, which may also lead to practical research. Let's keep carrying the torch that Gardner left burning.

Bibliography

Benbernou, Nadia, Erik D. Demaine, Martin L. Demaine and Benjamin Rossman. 2008. Coin-flipping magic. Presented at Gathering for Gardner 8, March. http://erikdemaine.org/papers/CoinFlipping.

Cipra, Barry, Erik D. Demaine, Martin L. Demaine and Tom Rodgers (eds). 2004. *Tribute to a Mathemagician*. Natick, Mass: A. K. Peters, pp 23–30 and 63–72.

Demaine, Erik D., Martin L. Demaine and Helena A. Verrill. 2002. Coin-moving puzzles. In *More Games of No Chance*. Cambridge, U.K.: Cambridge University Press, pp 405–431.

Demaine, Erik D., and Joseph O'Rourke. 2007. *Geometric Folding Algorithms: Linkages, Origami, Polyhedra*. Cambridge, U.K.: Cambridge University Press.

Gardner, Martin. 1989. Penny puzzles. In *Mathematical Carnival*, chapter 2. Washington, D.C.: Mathematical Association of America. Originally published in *Scientific American*, February 1966.

Gardner, Martin. 1995. Paper cutting. In *New Mathematical Diversions*, chapter 5. Washington, D.C.: Mathematical Association of America. Originally published in *Scientific American*, June 1960.

Science Needs More *Moneyball*

Frederick M. Cohan

The *Moneyball* story, in book and film, champions a data-mining revolution that changed professional baseball. On the surface, *Moneyball* is about Billy Beane, the general manager of the Oakland A's, who found a way to lead his cash-strapped club to success against teams with much bigger payrolls. Beane used data to challenge what everyone else managing baseball "knew" to be true from intuition, experience and training. He pioneered methods to identify outstanding players he could afford because they were undervalued by the traditional statistics used by the baseball elite.

This film was marketed as a sports movie. When I saw it, I knew right away what *Moneyball* is really about: the thrill and triumph of data mining. It's an instructive tale of how existing data can be examined for meaning in ways that were never intended or imagined when they were originally collected. Beane and his colleagues challenged the time-honored trinity of batting average, home runs and runs batted in (RBIs) as the essence of the offensive value of a player, replacing these statistics with newer measures based on the same data. They worked off theories developed by baseball writer and historian Bill James, who posited in the 1970s that the traditional stats were really imperfect measurements. James's approach didn't just replace one intuition with another. He let the game decide which stats did the best job of predicting offensive output.

Baseball's data-mining methods are starting a similar revolution in research

This approach is not easy. Trying to directly predict the number of games won would confound the skill of a team's offense with its pitching and fielding. James figured that one could test each offensive stat by trying to predict the total number of runs produced by each team over the course of a season, thus eliminating any effects of defense. It turned out that on-base percentage and slugging percentage were far superior to any individual offensive statistics used up to that point. James and others similarly devised statistics for pitching and fielding that were more independent of context.

Beane's use of the new statistics is appealing because it defeated the wisdom and training of other industry experts. His approach is summed up in one of the best scenes from the *Moneyball* film. Armed with his new data-mining methods, Beane challenges other talent evaluators about a player they all deem "good." A scout counters him, praising the player's swing. Beane's reply: "If he's such a good hitter, why doesn't he hit good?"

In other words, expert intuition aside, the data don't lie.

Frederick M. Cohan is a professor of biology and a professor of environmental studies at Wesleyan University. He earned his B.S. in biology at Stanford University and was awarded the first Ph.D. from Harvard University's then-new Department of Organismic and Evolutionary Biology. He studies the origins of ecological diversity in bacteria and is working to develop an "idiot's guide" to bacterial systematics—a universal system to identify the most newly divergent bacterial species before we know anything about the ecological or physiological differences that distinguish them. He teaches various courses in evolutionary biology, bioinformatics and the effects of global change on infectious disease. Address: Department of Biology, Wesleyan University, Middletown, CT 06459-0170. E-mail: fcohan@wesleyan.edu

Bacteria Stats

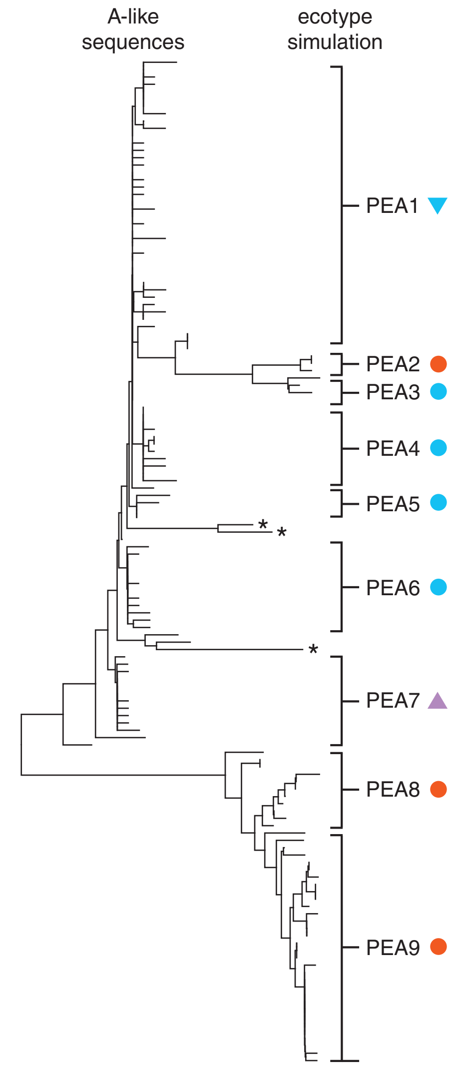
As I see it, the baseball revolution produced an "idiot's guide" to creating a team roster—a handbook based on things one can learn not through decades of experience and intuition but by applying general quantitative methods. It's the same kind of approach we should employ more in the sciences. Mountains of data and a capacity for analyzing them have also become available to science in the past few years. Data are now poised to trump the intuition of experts and the "facts" that scientists have championed over the years.

For instance, consider my own field, biology. Every biologist "knows" what a species is—a group of organisms that can successfully produce viable and fertile offspring. Biologists have long believed that species defined this way represent the fundamental units of ecology and evolution.

In the case of evolutionary microbiology (my specialty), it is particularly important to be able to recognize all the fundamental units of ecology among closely related bacteria. We especially need to distinguish those that are dangerous from those that are not and those that are helpful from those that are not. Indeed, we would like to identify all the bacterial populations that play distinct ecological roles in their communities.

As in baseball, the discovery of bacterial diversity has experienced a transition from relying on the subjective judgment of experts to objective and universal statistical methods. Originally, discovery and demarcation of bacterial species required a lot of expertise with a particular group of organisms, involving difficult measures of metabolic and chemical differences. To make the taxonomy more accessible, decades ago the field complemented this arduous approach with a kind of idiot's guide, where anyone could use widely available molecular techniques to identify species—for example, a certain level of overall DNA sequence similarity.

One popular universal criterion (among others) is to identify species as



The author and David Ward of Montana State University study a genus of bacteria called *Synechococcus*, found in hot springs such as this one in Yellowstone National Park *(above)*. The photosynthetic bacteria form mats *(inset)*, which have horizontal and vertical gradients. Temperature and nutrients diminish horizontally with distance from the spring source, and light brightness and solar spectrum increase vertically within the mat. The team used DNA sequences to place extremely close relatives of *Synechococcus*, classifiable within one species according to the traditions of systematics, on an evolutionary tree *(right)*. Their algorithm for ecotype simulation found nine groups most likely to be ecologically distinct. These groups differ from one another in their associations with temperature and depth. Blue is 60 degrees Celsius, purple is 63 degrees and red is 65 degrees. Circles indicate temperature independent of position, but colored up and down arrows correspond to upper and lower depths of the mat. (Photographs courtesy of David Ward.)

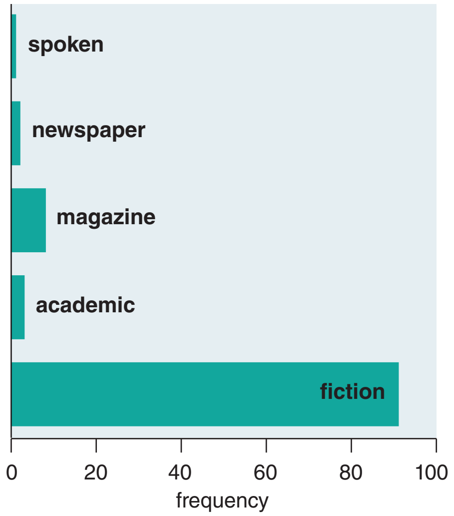
groups of organisms that are at least 99 percent similar in a particular universal gene. The problem is that—like the case of baseball where batting average, RBIs and home runs were used to supplement expert knowledge—nobody in microbiology tested whether the new molecular techniques actually came closer to solving the problem of recognizing the most closely related species.

Unfortunately, microbiology's current DNA-based idiot's guide, as well as the expert-driven metabolic criteria that preceded it, has yielded species with unhelpfully broad dimensions. For example, *Escherichia coli* contains strains that live in our guts peaceably, as well as various pathogens that attack the gut lining and others that attack the urinary tract. Moreover, established fecal-contamination detection kits that are designed to identify *E. coli* in the environment are now known to register a positive result with *E. coli* relatives that normally spend their lives in freshwater ponds, with little capacity for harming humans. And *E. coli* is not alone—there is a Yugoslavia of diversity within the typical recognized species: Much like the veneer of a unified country that hid a great diversity of ethnicities and religions, *E. coli* (and most recognized species) contains an enormous level of ecological and genomic diversity obscured under the banner of a single species name.

We can fix this confusion the same way that baseball improved its data analysis: by letting the game—or in our case, nature—decide which stats best predict what we most want to know. In microbiology the trick is to let the bacteria tell us what DNA sequence approach most accurately identifies the bacteria that are significantly different in their habitats and ways of making a living. Two teams, including Martin Polz's group at the Massachusetts Institute of Technology and my group at Wesleyan and Montana State Universities, have developed computer algorithms for identifying groups of bacteria specialized to different habitat types within an officially recognized species. These algorithms reject the expert-based criteria for how much diversity should be placed within a species. Instead, they analyze the dynamics of bacterial evolution to let the organisms themselves tell us the DNA sequence criterion that best demarcates ecologically distinct populations for a particular group of bacteria.

Another opportunity for discovery in biology through data mining stems from the new Human Microbiome Project. Here, DNA sequences are collected from various bacteria-laden human habitats, such as the gut, mouth, skin and genitals, with samples taken from individuals of different age, sex, health, weight and diet.

For example, Dusko Ehrlich of the French National Institute for Agricultural Research and his colleagues recently analyzed the bacterial genes purified from the feces of 39 humans from six European countries, amounting to about 100 million bases of bacterial DNA per person. They attempted to identify bac-



A search of the Corpus of Contemporary American English reveals that the phrase "she drew her breath" and its variants (he drew his breath, she draws her breath, draws breath, draw breath, drew breath, etcetera) is far more common in literature and other written sources than in spoken language. Such examples demonstrate that before audio recording became available, dialog in novels and stories may give poor examples in the analysis of how spoken English language usage has changed over time.

terial biochemical functions associated with age and body mass. Their intuition suggested various guesses for the identity of these genes, which were largely supported, but data-driven methods identified genes that gave much stronger relationships. One important data-driven discovery indicated a negative relation between obesity and the microbes' capacity for harvesting energy.

Ongoing massive sequencing projects in human, marine and soil environments allow us to characterize the diversification of bacteria: to discover the most newly divergent bacterial species, to characterize them as specialized to different habitats and to identify the biochemical functions most important in each habitat. However, the approach depends critically on how well we describe the habitats we sample.

Word Mining

Beyond the field of microbiology, data-mining revolutions are extending across the natural and social sciences (although meteorology and economics, with decades-long access to mountains of data, are still the granddaddies of this approach). In the social sciences, it is particularly interesting to see how data mining has recently helped linguists analyze how words are actually used in writing and speech—for example, as seen in the challenge of producing a dictionary. Traditionally, analysis of language use has involved assessment of written texts, usually from a canon of books accepted by experts as exemplars of "proper" usage, a step that required an army of volunteers who sent in quotations to the dictionary editors. Then the appointed set of language experts made subjective decisions about new usage—what is acceptable, what is vulgar and what is vile. A data revolution in linguistics is freeing us from needing the army of volunteers, as well as from the opinions of the learned experts. Language analysis is heading toward a data-driven idiot's guide that can decide on acceptable usage based on what is actually accepted in writing and in speech.

Various corpora of written and spoken language have emerged online, and these allow extensive analysis of how and where words are used. Entire uploaded texts can be searched and analyzed. The largest is the Oxford Corpus, launched in 2006 and covering texts from the entire Anglosphere. The U.S.-centered Corpus of Contemporary American English (COCA) features a user-friendly website (http://corpus.byu.edu/coca/). These corpora, when searched, give a 10-word neighborhood around each use of the word, which yields much information. For instance, a searcher can see whether the word is used in the singular or plural form, as well as words that are frequently co-located with it and so on. In *Damp Squid*, Jeremy Butterfield describes how these corpora can yield a picture of English (or potentially any language) as it is actually used, as validated by the entire community of writers and speakers.

One way that corpus-based analysis bucks expert opinion is in deciding when an evolutionary change in usage has become acceptable simply by the criterion of being frequently accepted. For example, the word "criteria," on entering the English language from Greek, maintained its original meaning as the plural of "criterion." Cringe though we may, our own experiences plus analysis of the Oxford Corpus show that use of "criteria" as the singular is catching up on its use as the plural. The corpus also allows us to note changes in old expressions that still hold meaning for us, but only if we change the words a little. Shakespeare's "in one fell swoop" is still a popular phrase four centuries later, but only through changing the obsolete adjective "fell" to one that sounds similar and holds a similar meaning, which is "foul." Despite the resistance of experts, the language is de facto evolving, and the corpus allows us to validate these changes.

Lost to the Past

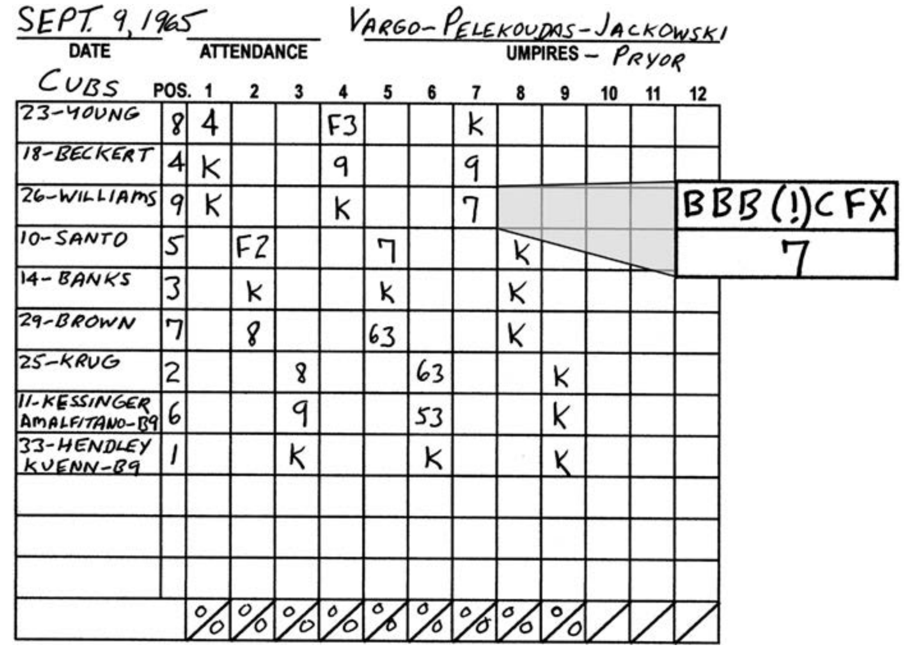
As useful as the idiot's guide approach has been across fields, gleaning meaning from old data serves up severe challenges. Difficulties can arise because at the time events happened, the data recorders did not anticipate that the information would be analyzed in ways not yet imagined. In cases stretching across baseball, biology and language, important items were not reported or, in some cases, observed at all. There is a twin problem to using past data, which is a communitarian challenge—appreciating that data are often used in ways unimagined at the time of collection, how can we make the data we record today more usable and valuable in the future?

As baseball and the sciences have taken an interest in mining old data for new insights, it has turned out that the old data sets are often sufficiently complete for us to discover new "laws" of baseball or science. Yet in far too many cases, fresh scrutiny of old data reveals painful omissions proving that science has missed an opportunity.

In retrospect, I am amazed at how little interest baseball and biology have shown for the future use of data. In baseball, the traditional play-by-play record of games was all that was reliably available until 1988, when the pitch-by-pitch record became the standard. The new record turned out to be important in many ways—for example, in managing a pitcher's productivity, health and longevity.

Until recently, biology was equally shortsighted in its data collection; this has created a problem for biologists who would like to analyze other scientists' published data. For example, Cathy Lozupone and Rob Knight at the University of Colorado figured out from analyses of others' data that the most difficult evolutionary transition in the history of bacteria has been from saline to nonsaline environments and vice versa. However, because the original researchers did not record the actual salinity levels, Lozupone and Knight could not pinpoint the precise concentration of salinity that has been most difficult to cross.

Previous standards of data collection in biology were typically limited to what might be interesting for the experiment



In 1965, baseball pitcher Sandy Koufax had a perfect game—allowing no runners from the Chicago Cubs—which the author saw with his Little League team. A scorecard shows Koufax struck out batters ("K") and others had fly balls to right ("9") or left ("7") field. But the play-by-play scoring format missed the game's pin-drop moment during Billy Williams's seventh inning at bat. Although Williams eventually flied out ("7"), a pitch-by-pitch record *(inset)* shows that before he had a strike ("C"), a foul ("F") and a hit ball ("X"), Koufax initially pitched three balls ("B"), one errant pitch away from walking the hitter to base and messing up perfection. (Image courtesy of the author.)

at hand or perhaps for some future experiment in the same lab. Today, biologists are increasingly expected to anticipate likely uses by others of the data we gather and are taking pains to do so, but this forethought is not easy.

I recently met with Hilmar Lapp, a database expert at the National Evolutionary Synthesis Center (NESCent), and discussed how researchers could avoid omitting important elements of data. He said that it is too much to expect, in the case of biology, for one researcher to think to include all the observations worthy of recording for posterity; he suggests what is needed is a "crowd intelligence." Accordingly, NESCent and other organizations have sponsored working groups to pool ideas and propose standards and directions of biological data collection in novel areas of inquiry—that is, to foster crowd intelligence. For example, the Genome Sequencing Consortium recently established standards for recording environmental data when genes and genomes are sampled; earlier action might have avoided the debacle of the missing salinity data Lozupone and Knight encountered.

In some cases, we do not have data on old events, not because of a lack of imagination but because the appropriate technology was not available at the time. In the case of baseball, the new, high-tech Advanced Value Metrics (AVM) system automatically describes each hit ball by its trajectory, velocity and point of hitting the ground. The AVM description of a hit allows analysis of how frequently a fielder can catch a ball that usually ends up being a double. But no one could analyze the skill of fielders at this level prior to the advent of this technology.

Until recently in biology, a lack of microbiological technology limited plant ecologists' understanding of the factors allowing a particular plant species to grow. Plant ecologists discovered only recently that the success of many plant species in nature is determined by helpful and harmful microbes that live in the soil. Therefore, decades of studies trying to understand the successes and failures of plants came up short because they failed to collect data on soil microbes.

In linguistics, the lack of technology for audio recording has hindered an analysis of spoken English usage over time. You might think that dialog written in novels and stories would be a good substitute for actual sound recordings; these pages are frequently as good a record as we will get. However, it is discouraging that a corpus-based analysis of word usage in speech versus fiction by lexicographer and author Orin Hargraves has shown that certain clichéd phrases, which appear to mimic spoken language, are actually used far more frequently in literature than in real life. For example, hardly anyone really says "he bolted upright" or "she drew her breath," but these forms are found with surprisingly high frequency in literature. Consequently, an unbiased, corpus-based account of spoken English usage begins with abundant voice recording in the 20th century.

Analyses of huge data sets allow us to move beyond our previous understanding, which was based on much less data than we have available to us today. There is so much possibility for a data-driven explosion of understanding of games, creatures and words by explorers today and in the future. We owe these future explorers the best and most complete record of life today that we can offer.

The *Moneyball* film opens with wisdom from Mickey Mantle: "It's unbelievable how much you don't know about the game you've been playing all your life." Surely the same is true for many in the natural and social sciences, pondering the areas they have been studying all their careers.

Bibliography

Arumugam, M., et al. 2011. Enterotypes of the human gut microbiome. *Nature* 473:174–180.

Becraft, E., F. M. Cohan, M. Kühl, S. Jensen and D. M. Ward. 2011. Fine-scale distribution patterns of *Synechococcus* ecological diversity in the microbial mat of Mushroom Spring, Yellowstone National Park. *Applied and Environmental Microbiology* 77:7689–7697.

Butterfield, J. 2008. *Damp Squid: The English Language Laid Bare*. Oxford: Oxford University Press.

Cohan, F. M., and S. M. Kopac. 2011. Microbial genomics: *E. coli* relatives out of doors and out of body. *Current Biology* 21:R587–R589.

James, B. 2011. *Solid Fool's Gold: Detours on the Way to Conventional Wisdom*. Chicago: ACTA Sports.

Kopac, S., and F. M. Cohan. 2011. A theory-based pragmatism for discovering and classifying newly divergent bacterial species. In *Genetics and Evolution of Infectious Diseases*, ed. M. Tibayrenc. London: Elsevier.

Lewis, M. 2003. *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton.

Lozupone, C. A., and R. Knight. 2007. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the U.S.A.* 104:11436–11440.

Wiedenbeck, J., and F. M. Cohan. 2011. Origins of bacterial diversity through horizontal gene transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews* 35:957–976.

Zimmer, B. 2011. The jargon of the novel, computed. *New York Times*, July 29.

Ode to Prime Numbers

Primes offer poetry both subject matter and structure

Sarah Glaz

"No branch of number theory is more saturated with mystery and elegance than the study of prime numbers," wrote Martin Gardner in his essay, "Patterns and Primes." It is therefore no wonder that prime numbers show up in another human endeavor that delves into mysteries in search of patterns and elegance—poetry. As a mathematician and poet, I have long been interested in this confluence.

Some poems, echoing the purpose of early poetic treatises on scientific principles, attempt to elucidate the mathematical concepts that underlie prime numbers. Others play with primes' cultural associations. Still others derive their structure from mathematical patterns involving primes. Whatever the mode of introduction, the meeting of poetry and primes—"those exasperating, unruly integers that refuse to be divided evenly by any integer except themselves and 1," as Gardner described them—is often an eventful one.

Poetic Mathematics

Gardner often quoted poems in his *Mathematical Games* column for *Scientific American*, and he wrote several essays on prime numbers. He could hardly have found a better poem for the subject than British poet Helen Spalding's "Let Us Now Praise Prime Numbers," which he reprinted in the essay "Strong Laws of Small Primes." The poem captures elements that have made primes an object of fascination since the time of Euclid. Spalding (1920–1991) is herself a mysterious figure whose life is difficult to trace after her last publication in *The London Magazine* in 1961.

Sarah Glaz is a professor of mathematics at the University of Connecticut. She is the author of Commutative Coherent Rings *(1989) and coeditor of two volumes of articles, all published by Springer. A third edited volume,* Commutative Rings, Integer-valued Polynomials and Polynomial Functions, *is forthcoming from Springer in 2014. She is coeditor of the poetry anthology* Strange Attractors: Poems of Love and Mathematics *(CRC Press/A K Peters, 2008), and her new edited collection,* Bridges 2013 Poetry Anthology, *will appear this summer from Tessellations Publishing.*

Let Us Now Praise Prime Numbers

Let us now praise prime numbers
With our fathers who begat us:
The power, the peculiar glory of prime numbers
Is that nothing begat them,
No ancestors, no factors,
Adams among the multiplied generations.

None can foretell their coming.
Among the ordinal numbers
They do not reserve their seats, arrive unexpected.
Along the lines of cardinals
They rise like surprising pontiffs,
Each absolute, inscrutable, self-elected.

In the beginning where chaos
Ends and zero resolves,
They crowd the foreground prodigal as forest,
But middle distance thins them,
Far distance to infinity
Yields them rare as unreturning comets.

O prime improbable numbers,
Long may formula-hunters
Steam in abstraction, waste to skeleton patience:
Stay non-conformist, nuisance,
Phenomena irreducible
To system, sequence, pattern or explanation.

—*Helen Spalding*

The poem's first stanza alludes to the Fundamental Theorem of Arithmetic. This theorem states that every positive integer greater than 1 is either a prime number or can be expressed as a unique product of prime numbers. Thus the primes are the building blocks of the integers and, consequently, of the entire real number system. In the second and third stanzas, Spalding suggests how prime numbers appear among the other numbers: Scattered without a discernible pattern, they fan out and occur less frequently as the numbers grow larger. However, despite this reduction in frequency, an infinite number of primes exists. Euclid's proof of the infinitude of prime numbers, circa 300 BCE, is considered to be one of the most elegant proofs in mathematics—a poem in its own right. Michael Szpakowski's *Proof, a Short Opera* offers a poetic and musical rendition of this proof. The piece can be viewed at www.somedancersandmusicians.com/proof/.

Courtesy of Paul Ashwell

Prime numbers capture the attention of visual artists and poets alike. *Prime Mark*, a 2010 work by Paul Ashwell, consists of 72 small canvases, each of which displays symbols that represent a number. Nonprime numbers are shown by combinations of symbols that indicate their prime factors. See more at http://paulashwell.co.uk/.

In the poem's final stanza, Spalding touches on one of the deep mysteries associated with prime numbers: our inability to pin them down with a formula. Prime numbers smaller than a given number N can be found through a technique called the Sieve of Eratosthenes—named for Eratosthenes (ca. 276–195 BCE), the Greek mathematician who discovered it. The "sifting" consists of a simple divisibility test and the systematic deletion of all the proper multiples of the prime numbers up to the largest prime smaller than the square root of N. The method works best when N itself is small. For $N = 100$, for example, the deletion leaves in the sieve the first 25 primes:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97

Since the time of Eratosthenes, many techniques have been invented to "catch" prime numbers, but as yet no formula has been found that covers them all. In particular, it is notoriously difficult to produce very large primes. Neither has a pattern been found to predict their distribution within a given interval of numbers. In 2000, the Clay Mathematics Institute listed seven of the most important open problems in mathematics. The institute offers an award of $1 million to anyone who publishes a solution to one of these Millennium Prize Problems. One problem, the Riemann Hypothesis, formulated by Bernhard Riemann (1826–1866), celebrated its 150th anniversary in 2010. It is a conjecture about the zeros of the Riemann zeta function. The function, ζ, is defined for complex variables, s, and a value of s for which $\zeta(s) = 0$ is called a zero of zeta. The zeta function was introduced by Leonhard Euler in the early 1800s as a function of a real variable. Riemann extended the function to complex numbers and established a connection between its set of zeros and properties of prime numbers. The Riemann Hypothesis is considered to be the most important open problem in pure mathematics, and its solution would advance our knowledge of the distribution of prime numbers. Tom Apostol's poem, "Where Are the Zeros of Zeta of *s*?," playfully imparts the excitement generated by the chase after its solution. It begins:

Where are the zeros of zeta of *s?*
G. F. B. Riemann has made a good guess;
They're all on the critical line, saith he,
And their density's one over 2pi log *t.*

This statement of Riemann's has been like a trigger
And many good men, with vim and with vigor,
Have attempted to find, with mathematical rigor,
What happens to zeta as mod *t* gets bigger.

—Tom Apostol, from "Where Are the Zeros of Zeta of s?"

Many other questions about prime numbers remain unanswered. Some of these problems and their partial solutions, as well as the spell cast by primes on the mathematicians who study them, have also made their way into poetry.

Prime Culture

Prime numbers have applications in computational fields, including cryptography and statistics, as well as in many scientific domains, such as engineering and physics. They also offer what Richard Crandall and Carl B. Pomerance call, in their 2005 book *Prime Numbers: A Computational Perspective,* "cultural connections." These cultural connections manifest themselves in poetry in a variety of ways.

The concept of primality is employed in poems as a metaphor for the intoxicating mysteries of life and human behavior. An example of this phenomenon is found in "Prime Numbers," by Jim Mele.

Prime Numbers

Prime numbers,
I remember them
like drinks
following complicated folk laws.
Out in California
a friend visits a pebble
beach, indivisible
in this uncertain life.

—Jim Mele

The depth of the cultural connection between primes and poetry becomes more apparent when we examine the inclusion of specific prime numbers in poems. The affinity between numbers and words has roots in the invention of alphabetic writing by the Phoenicians in the 2nd millennium BCE, when numbers came to be denoted by letters of the alphabet. In ancient poetry, especially in the domain of magic, mysticism and divination, every word acquired the number value of the sum of its letters and every number attained the symbolic values of one or more words in whose spelling it appeared. Historian of mathematics David Eugene Smith notes that 3 and 7 "were chief among mystic numbers in all times and among all people." This, he proposes, is because "3 and 7 are the first prime numbers—odd, unfactorable, unconnected with any common radix, possessed of various peculiar properties." In other words, 3 and 7 acquired a special importance precisely because of their primality. Vestiges of such significance, combined with layers of cultural, sociological and historical meaning, allow prime numbers to evoke powerful images and emotions, both personal and collective. Poems featuring the prime number 7 exemplify this effect. Perhaps most notably, 7 appears in key religious texts. It shows up in the first poem of *Genesis,* the first book of the Bible, as well as in the New Testament, the Koran, and others. Seven also appears in the *Epic of Gilgamesh*—one of the earliest known works of literature, dated around 2,000 BCE. The contemporary poems "Reasons for Numbers," by Lisel Mueller, and "How I Won the Raffle," by Dannie Abse, reflect the layers of history and mystery that the number 7 carried with it into the present; both are excerpted below:

7

Because luck
is always odd
and the division
of history
into lean and fat
years
mysterious

—Liesel Mueller, from "Reasons for Numbers"

I chose 7 because those ten men used to dance
around the new grave seven times.

Also because of the pyramids of Egypt;
the hanging gardens of Babylon;
Diana's Temple at Ephesus;
the great statue of Zeus at Athens;
the Mausoleum at Halicarnassus;
the Colossus of Rhodes;
and the lighthouse of Alexandria.

—Dannie Abse, from "How I Won the Raffle"

An even earlier poem features 7 as a lucky number. Langston Hughes's "Addition [1]" employs the form of a math problem to comment on the addition of "love" to "luck."

Addition [1]

7 x 7 + love =
An amount
Infinitely above:
7 x 7 – love.

—Langston Hughes

Lewis Carroll's classic poem, *The Hunting of the Snark,* mentions 7 in company of other numbers for an amusing mathematical effect. Do the math!

"Taking Three as the subject to reason about—
A convenient number to state—
We add Seven, and Ten, and then multiply out
By One Thousand diminished by Eight.

"The result we proceed to divide, as you see,
By Nine Hundred and Ninety and Two:
Then subtract Seventeen, and the answer must be
Exactly and perfectly true."

—Lewis Carroll, from The Hunting of the Snark

Aesthetics and Structure

Poems rarely call on prime numbers for their visual appeal. A notable exception is William Carlos Williams's imagist poem, "The Great Figure."

The Great Figure

Among the rain
and lights
I saw the figure 5
in gold
on a red
firetruck
moving
tense
unheeded
to gong clangs
siren howls
and wheels rumbling
through the dark city.

—Williams Carlos Williams

Williams's poem makes clear the aesthetic quality of the figure 5 he describes. American artist Charles Demuth's painting *I Saw the Figure 5 in Gold* was inspired by it. A series of multimedia works based on the poem are available at the website Poems that Go (poemsthatgo.com).

More often, numbers contribute to the structure of a poem. Poetry's musicality depends not only on words but also on quantifiable structural elements, and formal poetry relies on counting: metrical feet, rhyme words, line length, number of lines in a stanza, number of stanzas in the poem and more. A certain amount of mathematical calculation, either formal or intuitive, is involved in free verse as well. And some nontraditional poetic structures and procedures rely explicitly on the mathematical properties of prime numbers.

One such technique employs the Fundamental Theorem of Arithmetic. To construct a poem using this theorem, you decide on the length of the poem and then number the poem's lines consecutively from bottom to top, starting at 2. Then choose a word that stands for multiplication and a word that stands for exponentiation. The next step is to write the lines marked by prime numbers. Each line numbered with

Wikimedia Commons

Charles Demuth (1883–1935) painted *I Saw the Figure 5 in Gold* in response to a William Carlos Williams poem.

a prime is a building block of the other lines, much like the prime numbers build the positive integers. The first poem written with this structure was Carl Andre's poem "On the Sadness." My poem, "13 January 2009," was also made using this approach. The form does not require the writer to note the mathematics that undergirds it, but in this instance the notation is part of the poem.

13	**January 2009**
$12=2^2$x3	Anuk is dying for Anuk is dying in the white of winter
11	The coldest month
10=2x5	Anuk is dying in the falling snow
$9=3^2$	The white of winter for Anuk is dying
$8=2^3$	Anuk is dying for the white of winter
7	The drift of time
6=2x3	Anuk is dying in the white of winter
5	The falling snow
$4=2^2$	Anuk is dying for Anuk is dying
3	The white of winter
2	Anuk is dying
1	.

—Sarah Glaz

Here the word *in* stands for multiplication, and the word *for* stands for exponentiation. The poem is generated from the prime numbered lines—2, 3, 5, 7, and 11, which are written first—as follows: Factor each nonprime line number into a product of powers of distinct primes. For example, 12 = 2^2 x 3. The primes appearing in the number 12, arranged in increasing order, are 2 and 3. Line 2 is: *Anuk is dying,* and line 3 is: *The white of winter.* To construct line 12, replace the number 2 with line 2, the number 3 with line 3, multiplication with *in* and exponentiation with *for.* This makes line 12:

Lighght Prime

```
11111111111111111111111111111111111111111111111111111111111111111111111111
10000000000000000000000000000000000000000000000000000000000000000000000001
10000000000000000000000000000000000000000000000000000000000000000000000001
10008800000000888888000888888000880000880008888880008800008800888888880001
10008800000000008800008800008800880000880088000088008800008800000880000001
10008800000000008800008800000000880000880088000000008800008800000880000001
10008800000000008800008800000000888888880088000000008888888800000880000001
10008800000000008800008800888000880000880088008880008800008800000880000001
10008800000000008800008800008800880000880088000088008800008800000880000001
10008800000000008800008800008800880000880088000088008800008800000880000001
10008888888800888888000888888000880000880008888880008800008800000880000001
10000000000000000000000000000000000000000000000000000000000000000000000001
10000000000000000000000000000000000000000000000000000000000000000000000001
11111111111111111111111111111111111111111111111111111111111111111111111111
                                                                   x 10^1280 – 1
```

—*Jason Earls*

Anuk is dying for Anuk is dying in the white of winter. The same procedure is used to generate each line of the poem. When the poem is read aloud, the echo created by the repetition of prime-numbered lines evokes an elegiac mood.

Another method involves the aesthetic manipulation of very large primes. Jason Earls's concrete prime poem, "Lighght Prime" *(shown above)* is based on Aram Saroyan's poem, "Lighght." (The history of this poem, which engendered considerable controversy when it was first published, is worth looking up.)

Earls used zeroes and ones to create a visual representation of the poem. The word "lighght" appears in the interior of a rectangular array of digits, all of which are 0s and 1s. Taking the digits of this rectangular array and placing them in the same order on a straight line creates a long number. Multiplying this number by 10^{1280}, and then subtracting 1, yields a very large prime number. Verifying that this number is indeed prime involves the use of a computer program. Earls's book, *The Lowbrow Experimental Mathematician*, includes additional information on this poetic form and more concrete prime poems.

Yet another technique for constructing poems involves the prime number 7. This method, called the *n + 7 algorithm*, was invented by the Oulipian poet Jean Lescure. The literary movement known as Oulipo—Ouvroir de Litterature Potentielle (Workshop of Potential Literature)—was founded by Raymond Queneau in 1960. Its members invented constraints that generate literature; many of these constraints are mathematical. The $n + 7$ algorithm replaces each noun in a given poem with the seventh noun that follows in a specified dictionary. Mathematically, the procedure is a function on the set of nouns—one that "translates" each noun by 7 units. The results are often amusing. Computer programs make it easy to run this algorithm on longer texts, and to do so using numbers other than 7. You can try out the procedure using a dictionary or at www.spoonbill.org/n+7/.

Whether they are invoked as lucky numbers, employed as generative constraints, or just lauded in all their unruliness, primes in poetry lend both elegance and unpredictability. This dual nature—both exemplar and irritant—is familiar to poetry lovers. "Stay non-conformist, nuisance," Spalding urges the primes. It's a directive that the best poems often follow as well.

Acknowledgments

An earlier version of this essay appeared as "The Poetry of Prime Numbers" in the *Proceedings of Bridges Coimbra*, 2011, pp. 17–24. For permission to reprint their poems, we are thankful to Jim Mele and Jason Earls. "Addition [1]" is from *The Collected Poems of Langston Hughes*, by Langston Hughes, edited by Arnold Rampersad with David Roessel, associate editor, copyright 1994 by the estate of Langston Hughes. Used by permission of Alfred A. Knopf, a division of Random House, Inc. "The Great Figure," from *The Collected Poems of William Carlos Williams*, vol. 1, 1909–1939, by William Carlos Williams, is copyright 1938 by New Directions Publishing and is reprinted by permission.

Bibliography

Abse, D. 2000. *Be Seated Thou*. Rhinebeck, New York: The Sheep Meadow Press.

Crandall, R. E., and C. Pomerance. 2005. *Prime Numbers: a Computational Perspective*. New York: Springer.

Earls, J. 2010. *The Lowbrow Experimental Mathematician*. Raleigh, NC: Pleroma Publications.

Gardner, M. 2005. *Martin Gardner's Mathematical Games CD: The 6th Book of Mathematical Diversions* (1971). Washington, DC: Mathematical Association of America.

Glaz, S. 2010. "Poetry inspired by mathematics." *Proceedings of Bridges Pécs*, 35–42.

Glaz, S. 2011. "13 January 2009." *Recursive Angel*, May–June.

Glaz, S., and J. Growney, eds. 2008. *Strange Attractors: Poems of Love and Mathematics*. London: CRC Press/AK Peters.

Hughes, L. 1994. *The Collected Poems of Langston Hughes*. New York: Vintage Books.

Jennings, E., ed. 1961. *An Anthology of Modern Verse 1940–1960*. London: Methuen.

Mueller, L. 1986. *Second Language*. Baton Rouge: Louisiana State University Press.

Saroyan, A. 2007. *Complete Minimal Poems of Aram Saroyan*. New York: Ugly Duckling Presse.

Smith, D. E. 1958. *History of Mathematics*, volume II. Mineola, NY: Dover Publications.

Williams, W. C. 1985. *Selected Poems of Williams Carlos Williams*. New York: New Directions.

The Music of Math Games

Keith Devlin

Search online for video games and apps that claim to help your children (or yourself) learn mathematics, and you will be presented with an impressively large inventory of hundreds of titles. Yet hardly any survive an initial filtering based on seven, very basic pedagogic "no-nos" that any game developer should follow if the goal is to use what is potentially an extremely powerful educational medium to help people learn math. A good math learning game or app should avoid:

- Confusing mathematics itself (which is really a way of thinking) with its representation (usually in symbols) on a flat, static surface.
- Presenting the mathematical activities as separate from the game action and game mechanics.
- Relegating the mathematics to a secondary activity, when it should be the main focus.
- Adding to the common perception that math is an obstacle that gets in the way of doing more enjoyable activities.
- Reinforcing the perception that math is built on arbitrary facts, rules and tricks that have no unified, underlying logic that makes sense.
- Encouraging students to try to answer quickly, without reflection.
- Contributing to the misunderstanding that math is so intrinsically uninteresting, it has to be sugar-coated.

Of the relatively few products that pass through this seven-grained filter—

Video games that provide good mathematics learning should look to the piano as a model

which means they probably at least don't do too much harm—the majority focus not on learning and understanding but on mastering basic skills, such as the multiplicative number bonds (or "multiplication tables"). Such games don't actually provide learning at all, but they do make good use of video game technology to take out of the classroom the acquisition of rote knowledge. This leaves the teacher more time and freedom to focus on the main goal of mathematics teaching, namely, the development of what I prefer to call "mathematical thinking."

Many people have come to believe mathematics is the memorization of, and mastery at using, various formulas and symbolic procedures to solve encapsulated and essentially artificial problems. Such people typically have that impression of math because they have never been shown anything else. If mention of the word *algebra* automatically conjures up memorizing the use of the formula for solving a quadratic equation, chances are you had this kind of deficient school math education. For one thing, that's not algebra but arithmetic; for another, it's not at all representative of what algebra is, namely, thinking and reasoning about entire classes of numbers, using logic rather than arithmetic.

What's in a Game?

So how to go about designing a good video game to help students learn mathematics? The first step should be to read—several times, from cover to cover—the current "bible" on K–12 mathematics education. It is called *Adding it Up: Helping Children Learn Mathematics*, and it was published by the National Academies Press in 2001. The result of several years' work by the National Research Council's Mathematics Learning Study Committee, a blue-ribbon panel of experts assembled to carry out that crucial millennial task, this invaluable volume sets out to codify the mathematical knowledge and skills that are thought to be important in today's society. As such, it provides the best single source currently available for guidelines on good mathematics instruction.

The report's authors use the phrase *mathematical proficiency* to refer to the aggregate of mathematical knowledge, skills, developed abilities, habits of mind and attitudes that are essential ingredients for life in the 21st century. They break this aggregate down to what they describe as "five tightly interwoven" threads. The first is *conceptual understanding*, the comprehension of mathematical concepts, operations and relations. The second is *procedural fluency*, defined as skill in carrying out arithmetical procedures accurately, efficiently, flexibly and appropriately. Third is *strategic competence*, or the ability to formulate, represent and solve mathematical problems arising in real-world situations. Fourth is *adaptive reasoning*—the capacity for logical thought, reflection, explanation and justification. Finally there's *productive disposition*, a habitual inclination to see mathematics as sensible, useful and worthwhile, combined with a confidence in one's own ability to master the material.

The authors stress that it is important not to view these five goals as a checklist

Keith Devlin is a Stanford University mathematician and recently a founder of a small educational video-game studio. His company, InnerTube Games (innertubegames.net) will release its first game, Wuzzit Trouble, *in early March, and it will be available initially for iPhone and iPad, with other platforms to follow. He blogs at profkeithdevlin.org, and this article has been adapted from one of his blog series.*

to be dealt with one by one. Rather, they are different aspects of what should be an integrated whole, with all stages of teaching focused on all five goals.

So it's not that the crucial information about mathematics learning required to design good learning video games is not available—in a single, eminently readable source—it's that few people outside the math education community have read it.

Combining Skills

The majority of video games designed to provide mathematics learning fail educationally for one of two reasons: Either their designers know how to design and create video games but know little about mathematics education (in particular, how people learn mathematics) and in many cases don't seem to know what math really is, or they have a reasonable sense of mathematics and have some familiarity with the basic principles of mathematics education, but do not have sufficient experience in video game design. (Actually, the majority of math education games seem to have been created by individuals who know little more than how to code, so those games fail both educationally and as games.)

To build a successful video game requires an understanding, at a deep level, of what constitutes a game, how and why people play games, what keeps them engaged, and how they interact with the different platforms on which the game will be played. That is a lot of deep knowledge.

To build an engaging game that also supports good mathematics learning requires a whole lot more: understanding, at a deep level, what mathematics is, how and why people learn and do mathematics, how to get and keep them engaged in their learning, and how to represent the mathematics on the platform on which the game will be played. That too is a lot of deep knowledge.

In other words, designing and building a good mathematics educational video game—be it a massively multiplayer online game (MMO) or a single smartphone app—requires a team of experts from several different disciplines. That means it takes a lot of time and a substantial budget. How much? For a simple-looking, casual game that runs on an iPad, reckon nine months from start to finish and a budget of $300,000.

Following the tradition of textbook publishing, that budget figure does not include any payment to the authors who essentially create the entire pedagogic framework and content, nor the project's academic advisory board (which it should definitely have).

In the author's game, *Wuzzit Trouble*, the cute and fuzzy creatures must be freed from traps controlled by gearlike combination locks. Players collect keys to open the locks by solving puzzles of varying difficulty. (Image courtesy of InnerTube Games.)

The Symbol Barrier

Given the effort and the expense to make a math game work, is it worth the effort? From an educational perspective, you bet it is. Though the vast majority of math video games on the market essentially capitalize on just one educationally important aspect of video games—their power to fully engage players in a single activity for long periods of time—all but a tiny number of games (fewer than 10 by my count) take advantage of another educationally powerful feature of the medium: video games' ability to overcome the *symbol barrier*.

Though the name is mine, the symbol barrier has been well known in math education circles for over 20 years and is recognized as the biggest obstacle to practical mastery of middle school math. To understand the symbol barrier and appreciate how pervasive it is, you have to question the role symbolic expressions play in mathematics.

By and large, the public identifies doing math with writing symbols, often obscure symbols. Why do they make that automatic identification? A large part of the explanation is that much of the time they spent in the school mathematics classroom was devoted to the development of correct symbolic manipulation skills, and symbol-filled books are the standard way to store and distribute mathematical knowledge. So we have gotten used to the fact that mathematics is presented to us by way of symbolic expressions.

But just how essential are those symbols? After all, until the invention of various kinds of recording devices, symbolic musical notation was the only way to store and distribute music, yet no one ever confuses music with a musical score.

Just as music is created and enjoyed within the mind, so too is mathematics created and carried out (and by many of us enjoyed) in the mind. At its heart, mathematics is a mental activity—a way of thinking—one that over several millennia of human history has proved to be highly beneficial to life and society.

In both music and mathematics, the symbols are merely static representations on a flat surface of dynamic mental processes. Just as the trained musician can look at a musical score and hear the music come alive in her or his head, so too the trained mathematician can look at a page of symbolic mathematics and have that mathematics come alive in the mind.

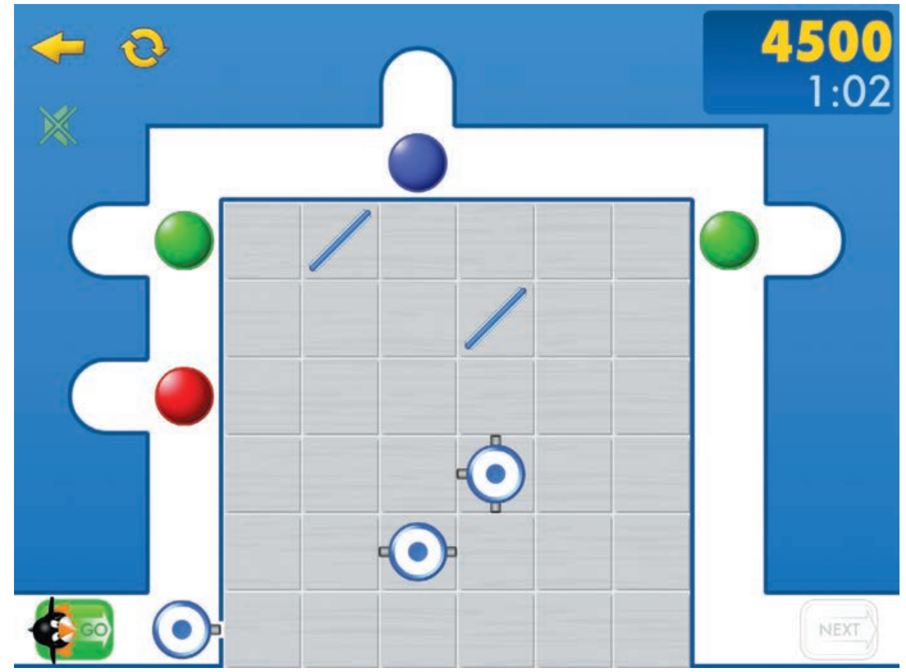
So why is it that many people believe mathematics itself is symbolic manipulation? And if the answer is that it results from our classroom experiences, why is mathematics taught that way? I can answer that second question. We teach mathematics symbolically because, for many centuries, symbolic representation has been the most effective way to record mathematics and pass on mathematical knowledge to others.

Still, given the comparison with music, can't we somehow manage to break free of that historical legacy?

Though the advanced mathematics used by scientists and engineers is intrinsically symbolic, the kind of math important to ordinary people in their lives—which I call everyday mathematics—is not, and it can be done in your head. Roughly speaking, everyday mathematics comprises counting, arithmetic, proportional reasoning, numerical estimation, elementary geometry and trigonometry, elementary algebra, basic probability and statistics, logical thinking, algorithm use, problem formation (modeling), problem solving, and sound calculator use. (Yes, even elementary algebra belongs in that list. The symbols are not essential.)

True, people sometimes scribble symbols when they do everyday math in a real-life context. But for the most part, what they write down are the facts needed to start with, perhaps the intermediate results along the way and, if they get far enough, the final answer at the end. But the doing-math part is primarily a thinking process—something that takes place mostly in your head. Even when people are asked to "show all their work," the collection of symbolic expressions that they write down is not necessarily the same as the process that goes on in their minds when they do math correctly. In fact, people can become highly skilled at doing mental math and yet be hopeless at its symbolic representations.

With everyday mathematics, the symbol barrier emerges. In their 1993 book *Street Mathematics and School Mathematics*, Terezinha Nunes, David William Carraher and Analucia Dias Schliemann describe research carried out in the street markets of Recife, Brazil, in the early 1990s. This and other studies have shown that when people are regularly faced with everyday mathematics in their daily lives, they rapidly master it to an astonishing 98 percent accuracy. Yet when faced with what are (from a mathematical perspective) the very same problems, but presented in the traditional symbols, their performance drops to a mere 35 to 40 percent accuracy.



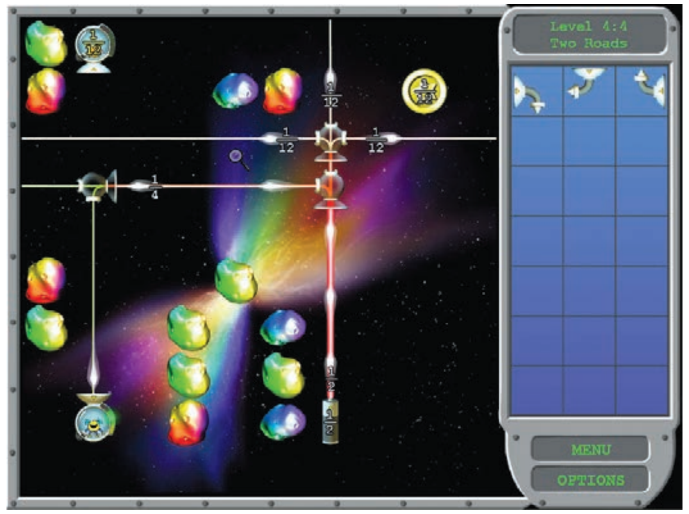
***KickBox* uses a penguin character called JiJi that players must help get from one end of the corridor to the other. Players position beam-splitters and reflectors to direct lasers that knock out obstacles in JiJi's path. Solving such a puzzle provides excellent practice in mathematical thinking, completely separate from the more familiar formulas, equations and dreaded "word problems." (Image courtesy of the MIND Research Institute.)**

It simply is not the case that ordinary people cannot do everyday math. Rather, they cannot do symbolic everyday math. In fact, for most people, it's not accurate to say that the problems they are presented in paper-and-pencil format are "the same as" the ones they solve fluently in a real life setting. When you read the transcripts of the ways they solve the problems in the two settings, you realize that they are doing completely different things. Only someone who has mastery of symbolic mathematics can recognize the problems encountered in the two contexts as being "the same."

The symbol barrier is huge and pervasive. For the entire history of organized mathematics instruction, where we had no alternative to using static, symbolic expressions on flat surfaces to store and distribute mathematical knowledge, that barrier has prevented millions of people from becoming pro-

***MotionMath* is a *Tetris*-inspired game that uses the motion sensors in a smartphone or tablet to allow players to tilt the screen to direct descending fractions to land on the right location on the number line. This game is an excellent introduction to fractions for younger children, as it connects the abstract concept to tactile, bodily activity. (Image courtesy of MotionMath Games.)**

ficient in a cognitive skill set of evident major importance in today's world, on a par with the ability to read and write.



In the math puzzle game *Refraction*, players learn about fractions and algebra. In this puzzle, the player has to split a laser beam a sufficient number of times to power all of the alien spaceships on the screen. The game is also designed to be modified on the fly, in an effort to capture data about what teaching methods and reward systems work best for students. (Image courtesy of the University of Washington.)

Going Beyond

With video games, we can circumvent the barrier. Because video games are dynamic, interactive and controlled by the user yet designed by the developer, they are the perfect medium for representing everyday mathematics, allowing direct access to the mathematics (bypassing the symbols) in the same direct way that a piano provides direct access to the music.

It's essentially an interface issue. Music notation provides a useful interface to music, but it takes a lot of learning to be able to use it. It's the same for mathematics notation.

The piano provides an interface to music that is native to the music, and hence far more easy and natural to use. When properly designed, video games can provide interfaces to mathematical concepts that are native to those concepts, and thus far more easy and natural to use.

Consider some of the reasons so many people are able to master the piano. You learn by doing the real thing (initially poorly, on simple tunes, but getting better over time). You use the very same instrument on Day 1 that the professionals use. You get a sense of direct involvement with the music. You get instant feedback on your performance—the piano tells you if you are wrong and how you are wrong, so you can gauge your own progress. The instructor is your guide, not an arbitrator of right or wrong. And the piano provides true *adaptive learning.*

We read a lot today about adaptive learning, as if it were some new invention made possible by digital technologies. In fact it is a proven method that goes back to the beginning of human learning.

What's more, the proponents of today's digital version have gotten it all wrong, and as a result produce grossly inferior products. They try to use artificial intelligence so an "educational delivery system" can modify the delivery based on the student's performance.

Yet tens of thousands of years of evolution have produced the most adaptive device on the planet: the human brain. Trying to design a computer system to adapt to a human's cognitive activity is like trying to build a cart that will draw a horse. Yes, it can be done, but it won't work nearly as well as building a cart that a horse can pull.

The piano metaphor can be pursued further. There's a widespread belief that you first have to master the basic skills to progress in mathematics. That's total nonsense. It's like saying you have to master musical notation and the performance of musical scales before you can start to try to play an instrument—a surefire way to put someone off music if ever there was one. Learning to play a musical instrument is much more enjoyable, and progress is much faster, if you pick up—and practice—the basic skills as you go along, as and when they become relevant and important to you. Likewise, for learning mathematics, it's not that basic skills do not have to be mastered, but rather it's how the student acquires that mastery that makes the difference.

When a student learning to play the piano is faced with a piece she or he cannot handle, the student (usually of his or her own volition) goes back and practices some more easier pieces before coming back to the harder one. Or perhaps the learner breaks the harder piece into bits, and works on each part, at first more slowly, then working up to the correct tempo. What the player does not do is go back to a simpler piano (one with fewer keys, perhaps?), nor do we design pianos that somehow become easier to play. The piano remains the same; the player adjusts (or adapts) what they do at each stage. The instrument's design allows use by anyone, from a rank beginner to a concert virtuoso.

This lesson is the one we need to learn in order to design video games to facilitate good mathematics learning. For over 2,000 years, commentators have observed connections between mathematics and music. We should extend the link to music when it comes to designing video games to help students learn math, thinking of a video game as an instrument on which a person can "play" mathematics.

A Mathematical Orchestra

The one difference between music and math is that whereas a single piano can be used to play almost any tune, a video game designed to play, say, addition of fractions, probably won't be able to play multiplication of fractions. This means that the task facing the game designer is not to design one instrument but an entire orchestra.

Can this be done? Yes. I know this fact to be true because I spent almost five years working with talented and experienced game developers on a stealth project at a large video game company, trying to build such an orchestra. That particular project was eventually canceled, but not because we had not made progress—we had developed over 20 such "instruments"—but because the pace and cost of development did not fit the company's entertainment-based financial model. A small number of us from that project took all that we had learned and formed our own company, starting from scratch to build our own orchestra.

In the meantime, a few other companies have produced games that follow the same general design principles we do. Some examples include the games *MotionMath* and *MotionMath Zoom,* which use the motion sensors in a smartphone or tablet to allow players to interact directly with numbers. The puzzle game *Refraction* was produced by a group of professors and students in the Center for Game Science at the University of Washington, and was designed as a test platform that could be altered on the fly to see what teaching methods and reward systems work best for students learning topics such as fractions and algebra. *DragonBox* focuses on learning algebra in a puzzle where a dragon in a box has to be isolated on one side of the screen. *KickBox* uses physical concepts—such as positioning lasers to get rid of obstacles for the game's penguin mascot—to learn math concepts. The same producer, the MIND Research Institute, also developed *Big Seed,* a game where players have to unfold colored tiles to completely fill a space. These games all combine the elements of math learning with game play in an effective, productive fashion.

DragonBox challenges players to isolate the glittering box (containing a growling dragon) on one side of the screen. What they are doing is solving for the x in an algebraic equation. But there isn't an x to be seen in the early stages of the game. As the player progresses through the game, mathematical symbols start to appear, first as names for the objects, later replacing the object altogether. This game demonstrates very clearly that solving an algebraic equation is not fundamentally about manipulating abstract symbols, but is reasoning about things in the world, for which the symbols are just names. *DragonBox* provides a more user-friendly interface to algebraic equations—but it's still algebra, and even young children can do it. (Image courtesy of We Want To Know Games.)

The game produced by my colleagues and me, because we were working in our spare time and were entirely self-funded until early last year, has taken us three years to get to the point of releasing. Available in early March, *Wuzzit Trouble* is a game where players must free the Wuzzits from the traps they've inadvertently wandered into inside a castle. Players must use puzzle-solving skills to gather keys that open the gear-like combination locks on the cages, while avoiding hazards. As additional rewards, players can give the Wuzzits treats and collect special items to show in a "trophy room."

We worked with experienced game developers to design *Wuzzit Trouble* as a game that people will want to play purely for fun, though admittedly mentally challenging, puzzle entertainment. So it looks and plays like any other good video game you can play on a smartphone or tablet. But unlike the majority of other casual games, it is built on top of sound mathematical principles, which means that anyone who plays it will be learning and practicing good mathematical thinking—much like a person playing a musical instrument for pleasure will at the same time learn about music. Our intention is to provide, separately and at a later date, suggestions to teachers and parents for how to use the game as a basis for more formal learning. *Wuzzit Trouble* might look and play like a simple arithmetic game, and indeed that is the point. But looks can be deceiving. The puzzles carry star ratings, and I have yet to achieve the maximum number of stars on some of the puzzles! (I never mastered Rachmaninov on the piano either.) The game is not designed to teach. The intention is to provide an "instrument" that, in addition to being fun to play, not only provides implicit learning but may also be used as a basis for formal learning in a scholastic setting. We learned all of these design lessons from the piano.

Slide Rules: Gone But Not Forgotten

Many of these well-made mechanical calculating aids have outlasted the engineers who knew how to use them, but they remain culturally pervasive.

Henry Petroski

In my recent column on paperweights (July–August 2016), I described how the electrical engineer Charles Steinmetz (1865–1923) often worked in a canoe afloat near his camp on a tributary of the Mohawk River, just outside Schenectady, New York, not far from the General Electric Research Laboratory where he was employed. Often among the papers, pencils, and other objects on the board he set across the gunwales to serve as a desk was a set of tables of logarithms that he used in making calculations. I also described another photograph, showing Steinmetz at work at a table inside the rustic camp, noting the presence of a slide rule that I suggested he used when he did not require the highest precision in his calculations.

Steinmetz worked at his Camp Mohawk through the first two decades of the 20th century. The slide rule and logarithm tables he used eventually were superseded by desktop electromechanical calculators that were large, heavy, and expensive, making them practical investments only for large engineering firms such as General Electric. Tables of logarithms and the slide rules based on them remained in common use by independent engineers and students into the 1960s, and first-year engineering students of the time were introduced to their theory, use, and application. The introduction in the 1970s of the handheld scientific calculator, at the time often referred to as an electronic slide rule, pretty much ended the era of both logarithms and the slide rule in calculations of all kinds, and by the 1980s the only engineering students who even knew what a slide rule was tended to be the young relations of engineers and scientists, who had inherited a family "slip stick" and rudimentary instruction in its use. By early in the current millennium the slide rule had been long forgotten in the back of bottom drawers, kept there largely for sentimental reasons—or for when batteries ran out or the power failed.

Some especially nostalgic engineers who actually did use a slide rule in the early days of their career mounted it on the wall as if it were the first dollar they had earned in a small business enterprise that had succeeded. Not a few engineers who had advanced into senior management positions were known to keep a small slide rule in a top desk drawer, ready to be used to check the results of computer calculations that were brought to them by their younger colleagues. Retired engineers with hairy ears often wore a working replica of a slide rule as a tie clip. Museum curators turned down donations of familiar slide rules because discarded models of the instruments had become so common and numerous in their collections.

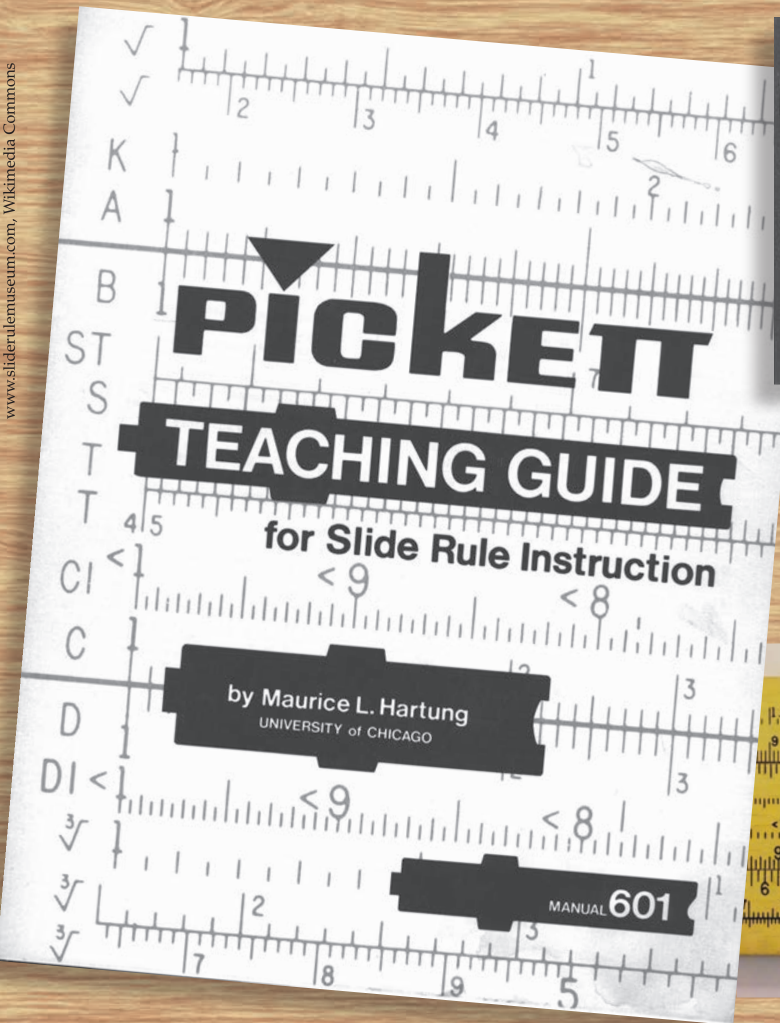
Most slide rules outlasted the engineers that used them, in large part because they were so well made, carefully used, and well maintained. Into the 1960s, first-year engineering students were advised to invest in a good slide rule, because it was something they would need and use for the rest of their engineering careers, and it was indeed expected to last that long. An investment of the order of $20, a not insignificant sum at the time, bought a top-of-the-line model, such as a Keuffel and Esser Log Log Duplex Decitrig, whose mahogany body was faced with white celluloid on which as many as two dozen scales were inscribed. A similar sum could also buy a Versalog with a bamboo structure, manufactured by the Frederick Post Company, or an aluminum-bodied rule made by Pickett with its scales incised on a yellow background.

Henry Petroski is the Aleksandar S. Vesic Professor of Civil Engineering and a professor of history at Duke University. His most recent book is The Road Taken: The History and Future of America's Infrastructure. *Address: Box 90287, Durham, NC 27708.*

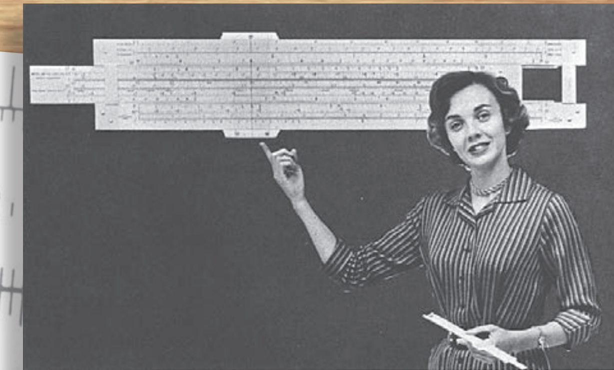
Instruments of Distinction

These precision mechanical analog computers were meant to be coddled, not thrown around or dropped, lest they get out of alignment. Each company's slide rule came with its own distinctive protective case, which was essentially a hard, boxlike sheath lined inside with soft material. The Keuffel and Esser was distinguished by its tan-orange case, the Post by its dark brown one, and the Pickett by its lighter brown one. For the standard 10-inch student model rule—the size refers to the length of the scales; the physical length of the rule was an inch or so longer—all cases were fitted with a cover flap and a means of attaching the case to its owner's belt. The sight of a student wearing a slide rule scabbard was an almost sure sign that he was an engineering student (and most were male in the slide rule era).

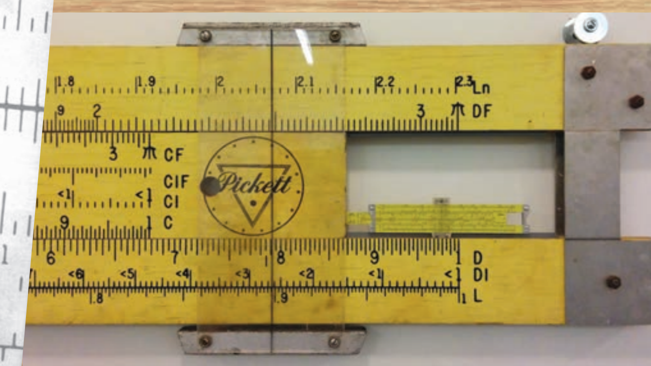
As can be imagined from the variety of materials of which slide rules were made, each had a different look, feel, and action. But they all were of pretty much the same basic construction: The moveable slide, from which the instrument took its name, was framed between a pair of fixed pieces known as *stators,* which were held together at their ends by metal fixtures or, on smaller versions, within the track formed in a base part. Those models consisting of stators and slide were marked on both sides with as many as a couple dozen scales, which were the key elements by which numbers were manipulated for calculations. The main scales were conventionally labeled A, B, C, and D, from top to bottom. Calculations involving nonadjacent scales were aided by the presence of a *cursor,* which consisted of a fine line inscribed on a glass or plastic windowlike device that could be slid from one end of the rule to the other.



www.sliderulemuseum.com, Wikimedia Commons



Pickett and other manufacturers produced an oversized slide rule designed to be hung on the wall of classrooms and used for teaching demonstrations (*above*); companies also usually supplied a teacher guide with the device (the one at left from 1960 was intended for junior and senior high school classes). The teaching slide rule below is about 8 feet long and is shown with a standard sized model in the slide gap for size comparison.



On a typical slide rule, the scales that are used most often are not arranged linearly, with the numbers equally spaced as they are on a measuring tape, but logarithmically, with the distance between, say, 1 and 2 being much greater than that between 8 and 9. This layout reflects the nature of a logarithm, which is the power to which a base number must be raised to give a target number. One familiar base number, which arises ubiquitously in mathematical physics problems and calculations, is known as Euler's number in honor of the 18th-century physicist and engineer Leonhard Euler (1707–1783), who designated it as e. The base of natural logarithms, e, is also known as Napier's constant, after the Scottish mathematician John Napier (1550–1617), who invented logarithms and also the calculating device known as Napier's bones, a precursor to the slide rule.

So-called common logarithms are derived from the base 10, and the "logarithm base 10" of a number y is written $\log_{10} y$, or simply $\log y$, when the base 10 is understood. Thus $\log 100 = 2$ because $10^2 = 100$. The familiar slide rules used by engineers were based on base 10 logarithms, as can be seen on a rule containing C, D, and L scales. The L scale gives the logarithm of the numbers aligned with it on the C or D scale. Thus at the extreme right, both C and D scales read 10, whereas the L scale reads 1.0, which is the base-10 logarithm of 10.

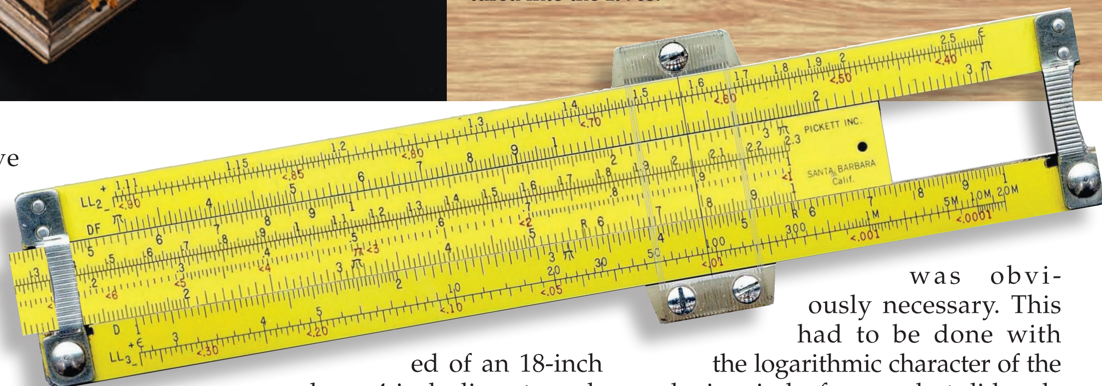
Because the rules for multiplying and dividing logarithms are $\log(x \cdot y) = \log x + \log y$, and $\log(x/y) = \log x - \log y$, to multiply two numbers on a slide rule, the numbers should be added; in order to divide the first by the second, the numbers should be subtracted. This means that adding the distance x to the distance y on the C and D scales, respectively, produces the product $x \cdot y$. Subtracting the distance y from the distance x gives the result of dividing x by y. These additions and subtractions could be done easily by sliding the C scale along the D scale. With practice, the operation was quick and accurate—but only to three or four significant digits.

The same computation can be alternately performed—not so quickly or easily, but much more accurately—by using a table of logarithms, which typically would record the logarithms of a number to many more significant digits. Because each number has a unique logarithm, it could be determined once and for all to a large number of decimal places and entered into a table of logarithms. It was a book of such tables that Steinmetz took out with him in his canoe when he wanted to make calculations to a very high degree of accuracy. When the simpler computation of addition or subtraction of logarithms had been performed, the result's inverse logarithm—known as the antilogarithm or antilog—could be determined. Contrary to popular opinion, slide rules could not be used for the addition or subtraction of the numbers themselves.

Slide Rule Predecessors

The true sliding rule was invented in about 1622 by the English mathematician William Oughtred (1574–1660),

Scottish mathematician John Napier (1550–1617) invented logarithms and a precursor to the slide rule called a calculating table (a version from 1680 is shown at left). A cylindrical calculating instrument was patented by Edwin Thacher in 1881 (*above*). The modern duplex slide rule (*below*), with scales on both front and back, was first patented in 1890 and manufactured into the 1970s.



© National Museums Scotland, U.S. Food and Drug Administration, www.sliderulemuseum.com

who is also said to have been the first to use the × symbol to represent multiplication and the abbreviations *sin* and *cos* for *sine* and *cosine*, respectively. Before Oughtred, multiplication and division by logarithms were performed using a single logarithmically scaled stick supplemented with a pair of dividers. To multiply two numbers, the first was located on the stick and then the second was stepped off from it by means of the dividers. Oughtred's inventive leap was to employ two sliding scales, first in a concentric circular arrangement and later in the linear form that became the iconic slide rule design.

What came to be the standard layout of scales is attributed to a French student named Victor Mayer Amédée Mannheim (1831–1906), whose name became associated with the familiar arrangement. However, Mannheim slide rules had scales on only one face of the rule, leaving the back not fully utilized. Slide rules incised with scales on front and back, effectively doubling the number of them that could be fitted onto the rule, became known as duplex models. A further improvement was made by the American civil engineer Edwin Thacher (1839–1922). His "Calculating Instrument" was patented in 1881 (U.S. Patent 249,117). It consisted of an 18-inch long, 4-inch diameter cylinder whose scales were made of 20 segments that added up to a 30-foot-long scale, off of which exceptionally accurate numbers could be read.

By incising a pair of sliding sticks or rods of a more conventional arrangement with the numbers from 1 and 10, the marks representing them spaced not uniformly but in proportion to their base-10 logarithms, the multiplication and division of two numbers could be easily done by a simple manipulation of the scales. To multiply 2.75 and 4.8, for example, the index—or 1 mark—on the C scale (which is located on the sliding center piece) would be aligned with the 2.75 mark on the D scale on the fixed body of the device. The cursor could then be slid along the rule to the 4.8 mark on the C scale, effectively adding the length representing 4.8 to that representing 2.75 and getting 7.55. Reading with the aid of the cursor the answer on the D scale completed the multiplication.

When the input or output number did not coincide exactly with a marked division of the scales, interpolation was obviously necessary. This had to be done with the logarithmic character of the scales in mind, of course, but slide rule neophytes were instructed how to do so. They soon became adept at making very good estimates of the third and fourth significant digits, the former at the left end of the scales, where the distance between numbers was more spread out than at the right end, where the distance was compressed because of the nature of logarithms. Engineers and scientists thus became quite confident in their answers to three digits.

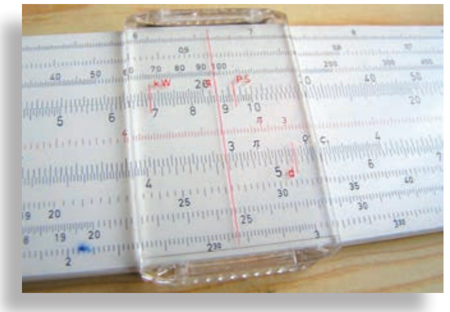
Enter the Electronic Age

One thing the slide rule did not do was locate the decimal place in an answer, and this fact was among the objections to allowing students to use the new electronic slide rules when they were first introduced. But another major reason for opposition was that the earliest pocket calculators were priced in the hundreds of dollars—compared with $20 for a high-quality slide rule—and faculty members worried about the unfair advantage that wealthy students might have over the less well-to-do. Not surprisingly, as academics are wont to do, committees were formed to study the issue and make recommendations as to how all students might be put on a level footing when taking exams. As with committees generally, those debating issues related to the new technology took their time to reach a conclusion. In the meantime, the price of handheld electronic calculators dropped precipitously, and before the study committees could issue their recommendations, the matter of manual versus electronic slide rules based on price had become moot.

In any case, whether by using a traditional slip stick or an electronic version, numbers of any accuracy could be multiplied and divided much more quickly by punching keys on a keypad than by looking up their multidigit logarithms in a book of tables, adding or subtracting the numbers manually, and then looking up the antilog of the answer. In the former case, the precision of the result would generally be limited to three or four significant digits and, depending upon how carefully the markings on the sliding rules were made and the ability of the user to interpolate between the nonlinear divisions, the result could be good enough—especially for initial or tentative engineering work, where the numbers are subject to revision anyway.

When more accuracy was desired of a slide rule, one with longer scales could be used. Twenty-inch rules, designed to be used at a desk or drafting table, served this purpose. So did circular slide rules, because the length of their scales could be as much as three times (π times to be exact) the diameter of the device. Sometimes greater compactness was more important than greater accuracy, and five-inch slide rules were commonly carried around among the pens and pencils packed into the shirt pockets or pocket protectors of busy engineers. A Picket five-inch pocket slide rule was taken to the Moon by Apollo astronauts.

For many decades students were instructed in classrooms outfitted with a six-foot or longer working slide rule hung front and center over the blackboard. The rudiments of slide-rule principles and use were explained on this large model. Instruction typically began with basic multiplication and division as done on the C and D scales, but instruction was also given in the use of several of the other straightforward scales, such as those used for calculating squares and cubes and taking square roots. Students were often left on their own, with their slide rule's owner's manual, to master others of the two dozen scales relevant to their specific needs. The Log Log Duplex Decitrig, for example, had trigonometric scales with angles measured in tenths of a degree (hence the name decitrig). There were also specialized slide rules more suited to, say, electrical than mechanical engineering problems. Slide rules suited to technicians were even more specialized.



Wikimedia Commons

A slide rule's cursor consists of a transparent plate inscribed with a thin line; it is helpful in aligning figures for calculations using scales that are not adjacent.

The fact that, on a slide rule, adding two marked lengths did not give the sum of the numbers, but the logarithms of the numbers, formed the basis for a common shibboleth to separate initiated from uninitiated users of the slip stick. Furthermore, by writing numbers differing by orders of magnitude in scientific notation, say 3.784532×10^6 and 0.097354×10^6, the numbers can be multiplied or divided using logarithms (or a slide rule), but where the decimal point in the result is located is left to the judgment of the person doing the calculation. As long as the slide rule was in common use, the skill to do this was a learned ability. The engineer thus had to have at all times a sense of the expected magnitude of the answer to a calculation.

Without developing the skill of estimating orders of magnitude, engineers were not prepared to recognize when a computer's output was absurd.

Among the objections to the introduction of the electronic slide rule was that it automatically located the decimal point in the answer, thereby not forcing engineering students—and perhaps even mature engineers—ever to develop the skill of estimating orders of magnitude. This was not a moot point, for without developing such a sense, engineers were not prepared to recognize when a calculator's or computer's answer was unreasonable, if not absurd. Whatever the electronic display or printout provided was taken as gospel and this total reliance on the machine could on occasion lead to over- or underdesigned structures and systems, or to their failure. Not having a sense of magnitude could allow, for example, an engineer to be oblivious to the fact that units of feet were being used in place of meters—or vice versa—in design calculations.

Even though they are no longer manufactured by the likes of Keuffel and Esser, in the culture of engineering the slide rule remains pervasive. A slide rule appears on the logos of long-established engineering societies, ranging from the British Institution of Mechanical Engineers to the Institution of Engineers, Malaysia. Joe Miner, the mascot of the Missouri School of Mines (recently renamed Missouri University of Science and Technology) carries a pickax in one hand and an outsize slide rule over his shoulder. The engineer-writer Nevil Shute Norway, who wrote under his first two names only, titled his autobiography, *Slide Rule: The Autobiography of an Engineer.* And at the University of Maryland, there is a long, low structure known as the "slide rule building," which was architecturally designed to resemble the calculating device in accordance with the wishes of its benefactor, Glenn L. Martin (1886–1955), the aviation pioneer and founder of the aircraft company headquartered in nearby Bethesda.

Although the slide rule may no longer be on every engineer's desk or drafting board, it remains embedded in the culture of the profession and serves as a symbol of what is at the root of engineering practice: calculation. Whether of stresses, strains, voltages, currents, flow rates, concentrations, lift, drag, or whatever is relevant to the field being practiced, even as it has been superseded by the handheld calculator and digital computer, the slide rule continues to symbolize the engineer at work. It was certainly seldom far from Charles Steinmetz's reach, whether he was working at a desk at the General Electric Research Laboratory or in his canoe at his Camp Mohawk.

In Defense of Pure Mathematics

After 75 years, Godfrey Harold Hardy's A Mathematician's Apology *still fuels debate over pure versus applied mathematics.*

Daniel S. Silver

Godfrey Harold Hardy was one of the greatest number theorists of the 20th century. Mathematics dominated his life, and only the game of cricket could compete for his attention. When advancing age diminished his creative power, and a heart attack at 62 robbed his physical strength, Hardy composed *A Mathematician's Apology*. It was an *apologia* as Aristotle or Plato would have understood it, a self-defense of his life's work.

"A mathematician," Hardy contended, "like a painter or poet, is a maker of patterns.... The mathematician's patterns, like the painter's or the poet's, must be beautiful; the ideas, like the colours or the words, must fit together in a harmonious way." It was a personal and profound view of mathematics for the layman, unlike anything that had appeared before. The book, which this year reaches the 75th anniversary of its original publication, is a fine and most accessible description of the world of pure mathematics.

Ever since its first appearance, *A Mathematician's Apology* has been a lightning rod, attracting angry bolts for its dismissal of applied mathematics as being dull and trivial. The shaft that lit up the beginning of a review in the journal *Nature* by Nobel laureate and chemist Frederick Soddy was particularly piercing: "This is a slight book. From such cloistral clowning the world sickens."

Hardy's opinions about the worth of unfettered thought were strong, but stated with "careful wit and controlled passion," to borrow words of the acclaimed author Graham Greene. They continue to find sympathetic readers in many creative fields. They were prescient at the time, and remain highly relevant today.

Daniel S. Silver is a research mathematician at the University of South Alabama. His research explores the relation between knots and dynamical systems, as well as the history of science and the psychology of invention. E-mail: silver@southalabama.edu. This is an extended version of the article that appeared in print and digital editions.

The Art of Argument

Hardy was born on February 7, 1877, in Cranleigh, Surrey. His parents valued education, but neither had been able to afford university.

Hardy grew up to be a scholar, a sportsman, an atheist, and a pacifist, but above all, he was an individualist. In an obituary of him, the mathematician E. C. Titchmarsh recalled: "If he dined at high table in tennis clothes it was because he liked to do so, not because he had forgotten what he was wearing."

Portents of Hardy's interest in mathematics as well as his lack of interest in religion were apparent at an early age. In church his energies were usually directed toward factoring numbers on hymn boards rather than toward worship. But his attitude about religion went deeper than mere disinterest. According to Titchmarsch:

> Hardy always referred to God as his personal enemy. This was, of course, a joke, but there was something real behind it. He took his disbelief in the doctrines of religion more seriously than most people seem to do.

Hardy exhibited his disbelief in odd ways. For example, he refused to enter any religious building. Mathematician George Pólya remembered that whenever he and Hardy walked past a church, Hardy would be sure that Pólya was between him and the building. Pólya never knew the reason.

Hardy began a famous collaboration with analyst J. E. Littlewood in 1911. Two years later they would pore over strange handwritten mathematical manuscripts that had been sent unsolicited by a young Indian civil servant, Srinivasa Ramanujan. Together they would decide that it was the work of a true genius. After considerable effort, Hardy succeeded in bringing Ramanujan to the University of Cambridge, where Hardy was a professor. The "one romantic incident in my life" is how Hardy described his discovery and collaboration with his young protégé, who tragically died of illness seven years later.

The atmosphere at Cambridge was nearly unbearable for Hardy during the World War I years of 1914 to 1918. Many of his friends and colleagues, including Littlewood, had gone off to fight. Hardy was a pacifist but not a conscientious objector: He volunteered for service only to be rejected on medical grounds. His deep regard for German culture and equally deep distrust of politicians compounded his emotions. In 1916 his pacifist friend Bertrand Russell was dismissed from his lectureship at Trinity College of Cambridge for printing "statements likely to prejudice the recruiting and discipline of His Majesty's forces." Hardy felt quite alone.

In 1919, Hardy moved to Oxford University, where his eccentricities thrived, and he was again happy and productive. In his rooms he kept a picture of Vladimir Lenin. He shunned mechanical devices such as telephones, would not look into a mirror, rarely allowed his photograph to be taken, and was very shy about meeting people. Nevertheless, he was a superb conversationalist, able to carry on talk about many subjects (including, of course, cricket). Titchmarsh recalled: "Conversation was one of the games which he loved to play, and it was not always easy to make out what his real opinions were."

Pólya had a similar recollection: "Hardy liked to shock people mildly by stating unconventional views, and he liked to defend such views just for the sake of a good argument, because he liked arguing."

It is clear that Hardy enjoyed teasing his audience, something one should keep in mind when reading *A Mathematician's Apology.*

The offer of the Sadleirian chair of pure mathematics at Cambridge was too great a temptation for Hardy. In 1931 he returned to the university by the River Cam, once home to Isaac Newton and James Clerk Maxwell. As his illness and debilitation progressed, academic honors accumulated: the Chauvenet Prize from the Mathematical Association of America in 1932, the Sylvester Medal from the Royal Society in 1940. The Copley Medal from the Royal Society was to be presented to him on December 1, 1942, the day he died.

Work for Second-Rate Minds

Many reviews of *A Mathematician's Apology* appeared during the first few years of its life. Most were favorable. The author of one such review, published in *The Spectator* in 1940, was Graham Greene. Hardy must have been flattered to read: "I know no writing—except perhaps Henry James's introductory essays—which conveys so clearly and with such an absence of fuss the excitement of the creative artist."

Other reviews were less enthusiastic. Today, as then, there are several reasons to be offended by *A Mathematician's Apology,* especially if you are a scientist.

If you are the author of expository articles (such as this one), then you don't have to wait long for an insult to be hurled your way. Hardy's book began:

> It is a melancholy experience for a professional mathematician to find himself writing about math-

The second edition of *A Mathematician's Apology* featured Hardy's now-iconic photograph on the dust jacket. For the first edition, Hardy sent postcards requesting that presentation copies be sent to colleagues including C. D. Broad and J. E. Littlewood, the physicist Sir Arthur Eddington, chemist and novelist C. P. Snow, cricketer John Lomas (to whom he dedicated the book), and his sister Gertie. He also requested copies be sent to colleagues in the United States. (Postcards photograph courtesy of Cambridge University Library.)

Top: Wikimedia Commons. Background: Internet Archive, archive.org ematics. The function of a mathematician is to do something, to prove new theorems, to add to mathematics, and not to talk about what he or other mathematicians have done. ...Exposition, criticism, appreciation, is work for second-rate minds.

Most reviewers pardoned Hardy for this assertion. Félix de Grand'Combe, professor of French at Bristol University, did not. Writing in the *Journal of Education* in August 1943, the former French army officer exploded:

> It really is a touching—albeit ostentatious—confession of a local intellectual debility... It is clear that Prof. Hardy is a great mathematician. It is no less clear, from his own showing, that one can be a great mathematician and yet fail to understand things that are readily comprehensible to an ordinary, well-educated mind.

Artists resent art critics, musicians scorn music critics. It is an ancient story, as Grand'Combe reminded readers in his lengthy review. However, he argued that observing and reformulating can be creative and illuminating acts:

> When Linnaeus devised his wonderful classification of plants he didn't "make" anything, he merely discovered a pre-existing treasure, explaining and rendering perceptible to all eyes a series of coherent relationships actually present in Nature, but his work altered and clarified our whole conception of the vegetable world; it gave informing reason to apparent chaos, life to what the ancients had seen as a dark welter of "non-being."

"A mathematician, like a painter or poet, is a maker of patterns.... The mathematician's patterns... must be beautiful; the ideas, like the colours or the words, must fit together in a harmonious way."

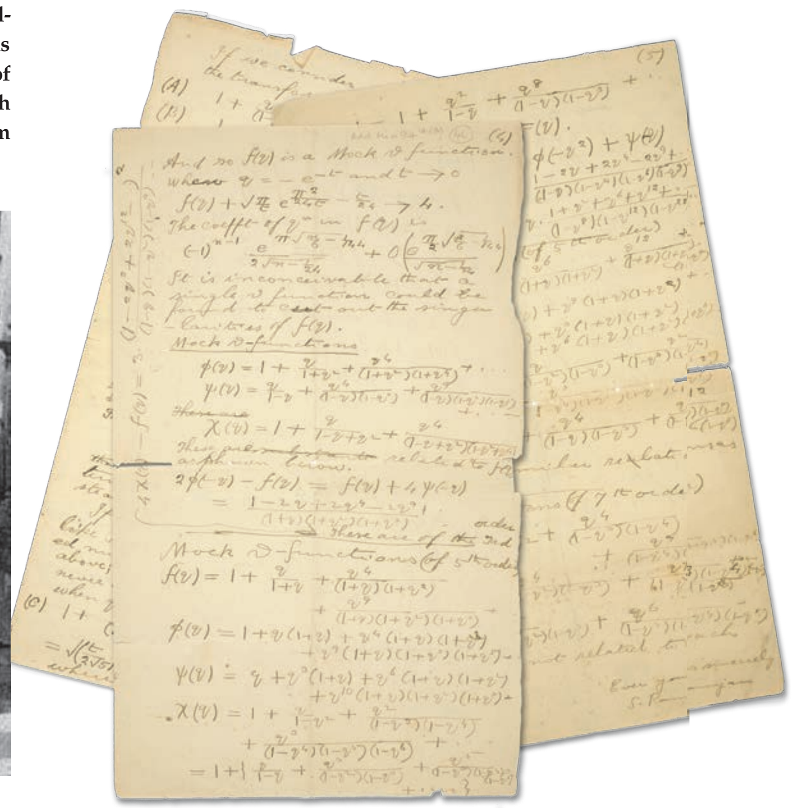
If Hardy thought that exposition, criticism, and appreciation is work for second-rate minds, then he must have come to that conviction late in life. During his prime years, he wrote book reviews for *The Cambridge Review, The Times Literary Supplement, Nature*, and *The Mathematical Gazette*. He was an effective and enthusiastic lecturer, constantly in demand. His textbook *A Course in Pure Mathematics*, published in 1908 and still in print, is entirely expository.

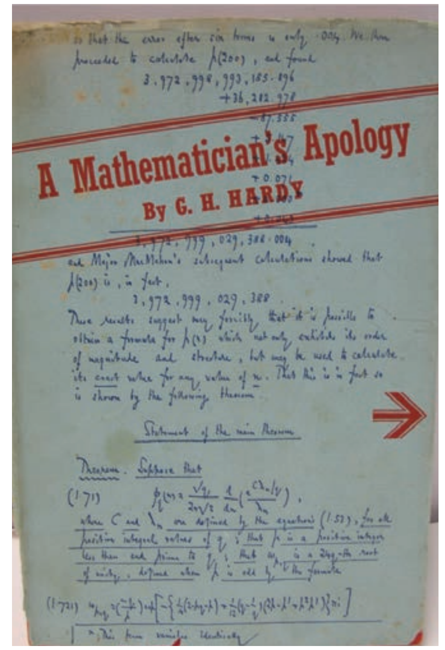
Greene and Grand'Combe had curiously different reactions to Hardy's mathematical calligraphy sampled on the dust jacket of *A Mathematician's Apology*. While Greene found mystical allure in "the author's tiny algebraic handwriting, as beautiful as Greek," Grand'Combe experienced nothing but incomprehension:

> In a book whose jacket is illustrated by what, I presume, is a sample of creative mathematical calculation, culminating in a formula, wherein an array of more than twenty-five figures, interlaced with abundant pluses and minuses is locked in the cold embrace of at least three different kinds of brackets, Professor G. H. Hardy of Cambridge purports to address the layman.

Hardy (*far right*) and his protégé Srinivasa Ramanujan (*center*) are shown with colleagues at Cambridge University (*below*). Ramanujan send Hardy many theorems in letters (*right*), and worked closely with Hardy for five years on various aspects of number theory, including highly composite numbers, which are positive integers with more divisors than any smaller positive integer. Ramanujan received a doctorate from Cambridge for this work in 1916. (Letters image courtesy of Ken Ono.)


Wikimedia Commons





The original printing of Hardy's book included mathematical calculations in the author's handwriting on the cover. Some reviewers found the imagery charming; others thought it unapproachable for the layman. (Photograph courtesy of the author.)

If you are a biochemist in search of a cure for a dreadful disease, then you might be insulted by Hardy's summary of your true motives. There are three: intellectual curiosity, professional pride (including anxiety to be satisfied with one's performance), and ambition. According to Hardy:

> It may be fine to feel, when you have done your work, that you have added to the happiness or alleviated the sufferings of others, but that is not why you did it.

Writing in *The News Letter* in 1941, the English physicist and philosopher of science Norman Campbell took Hardy's assertions at face value. However, if a mathematician's principal motivation is to benefit himself rather than society, he asked "why should we provide ...so many more comfortable jobs for mathematicians than for, say, poets or stamp-collectors?"

Should you be an older mathematician, you might be vexed by Hardy's reminder: "No mathematician should ever allow himself to forget that mathematics, more than any other art or science, is a young man's game." After observing that French mathematical prodigy Évariste Galois died at 20, Ramanujan at 33, and Bernhard Riemann at 40, Hardy added: "I do not know an instance of a major mathematical advance initiated by a man past fifty."

Cambridge University philosopher C. D. Broad responded:

> To produce, as [Hardy] does, a list of persons who did supreme creative work in mathematics and then died young... is surely irrelevant. I suppose that the suppressed premise is that the work which they did before their early deaths was so stupendously great that it is incredible that they should have equaled it if they had lived.

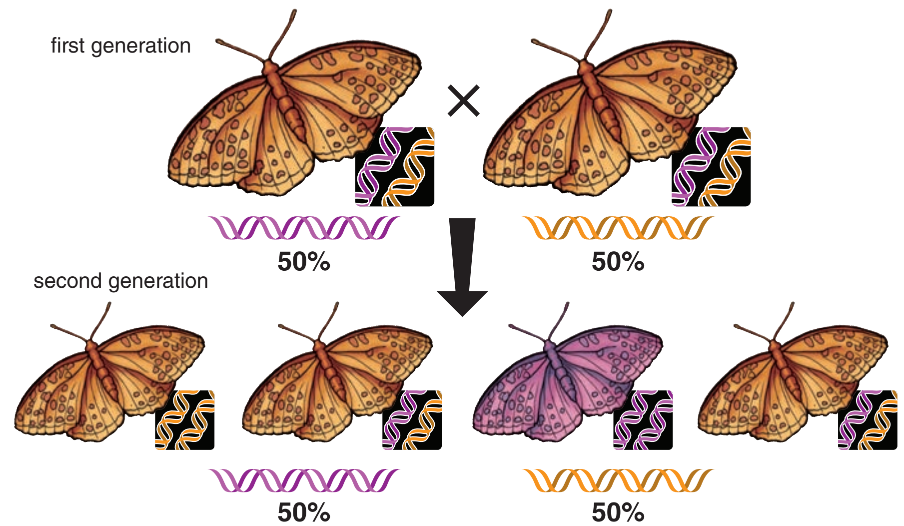
Examples of mathematicians who have made significant discoveries after the age of 50 can be given easily. Littlewood, who remained productive well after the age of 90, is one counterexample. Nevertheless, Hardy's belief is common today among mathematicians. It is encouraged by the fact that the Fields Medal, the highest award in mathematics, is awarded only for work done before the age of 40.

If you are a scientist whose feathers are not yet ruffled, Hardy's main contention will surely disturb your plumage. "Real" mathematics, he argued, is almost wholly "useless" whereas useful mathematics is "intolerably dull." By "real" mathematics, Hardy meant pure mathematics that tends to be abstract and general and, in Hardy's opinion, has the most aesthetic value. Opposed to it is the bulk of mathematics seen in school: arithmetic, elementary algebra, elementary geometry, differential and integral calculus, mathematics designed for computation and having the least aesthetic appeal.

Hardy was both prosecutor and defender in an imaginary trial to determine whether his life had been worthwhile:

> I have never done anything "useful." No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world. ...I have just one chance of escaping a verdict of complete triviality, that I may be judged to have created something worth creating.

As more than one observer has noted, it is ironic that Hardy is perhaps most widely known for a discovery



It is ironic that Hardy, a believer in pure and "useless" mathematics, is best remembered by many for the Hardy-Weinberg Principle, which quantifies how the frequencies of genetic traits remain constant across generations, in the absence of mutations or other evolutionary influences. Hardy was introduced to the problem by geneticist Reginald Punnett, with whom he played cricket. In a 1908 letter to *Science*, Hardy called his mathematical derivation of this problem "very simple." (Illustration by Barbara Aulicino.)

Hardy and analyst J. E. Littlewood, with whom he collaborated for many years starting in 1911. (Photograph courtesy of the University of Cambridge.)

about genetics. A theorem that he and Wilhelm Weinberg independently proved is well known today as the Hardy-Weinberg principle.

But for Hardy, who lived through two world wars, number theory provided a retreat that was, thankfully, useless to military planners. Hardy had opposed England's entry into the first World War, a deadly conflict made more so by science and technology. *A Mathematician's Apology* was first published in 1940, when England was again at war. Borrowing from his own article, "Mathematics in Wartime," published in the journal *Eureka* in 1940, Hardy wrote in *Apology* the same year that "When the world is mad, a mathematician may find in mathematics an incomparable anodyne. For mathematics is, of all the arts and sciences, the most austere and the most remote."

According to Hardy, the mathematician's world is directly linked to reality. Theorems are non-negotiable. In contrast, he says, the scientist's reality is merely a model. "A chair may be a collection of whirling electrons, or an idea in the mind of God," he declared. "Each of these accounts of it may have merits, but neither conforms at all closely to the suggestions of common sense." The pure mathematician need not be tethered to physical facts. In Hardy's words: "'Imaginary' universes are so much more beautiful than this stupidly constructed 'real' one."

Frederick Soddy, who had helped the world understand radioactivity, was disgusted by such sentiments. In his review in *Nature,* he said that if Hardy were taken seriously, then the "real mathematician" would be a "religious maniac."

Hardy enjoyed teasing his audience, something one should keep in mind when reading *A Mathematician's Apology.*

Soddy added a rebuke:

> Surely in these times a little appreciation of the military virtues, rather than the conventional vilification of the profession of arms, is called for by religious people of all sorts, especially since their sort of "education" is really at the bottom of the whole tragedy, and the chemist's bombs and poison-gas are such a heaven-sent whipping boy for their own.

Hardy was aware of Soddy's review. He might have been amused by it. A letter to him from R. J. L. Kingsford at Cambridge University Press, dated January 1941, concluded: "I quite agree that Soddy's amazing review in *Nature* is a most valuable advertisement. I enclose a copy of the review, herewith."

The notion that mathematicians form a self-protected priesthood with their own religion was popular. It runs throughout Lancelot Hogben's *Mathematics for the Million,* a highly successful book about mathematics that has sold well since 1937. After condemning Pythagoras and Plato for their excessive fondness for abstraction, Hogben wrote:

> The fact that mathematicians are often like this may be why they are so inclined to keep the high mysteries of their Pythagorean brotherhood to themselves.

In their reviews both Soddy and Broad suggested that Hardy's book might have been in part a response to Hogben. Hogben criticizes Hardy in *Mathematics for the Million,* which adds credence to the theory.

Another condemnation of *A Mathematician's Apology* came from E. T. Bell, a mathematician and science fiction writer who is best remembered for his 1937 book *Men of Mathematics.* In his review, published in *The Scientific Monthly* in 1942, Bell recommended Hardy's book to "solemn young men who believe they have a call to preach the higher arithmetic to mathematical infidels." He concluded: "Congenital believers will embrace [Hardy's book] with joy, possibly as a compensation for the loss of their religious beliefs of their childhood."

It Won't Make a Nickel for Anyone

Hardy had intended to publish *A Mathematician's Apology* with Cambridge University Press at his own expense. However, Press Secretary S. C. Roberts recognized the value of the 90-page essay, and endorsed it to the Syndics, the governing body of the Press. At meetings in July 1940, it was decided that 4,000 copies of *A Mathematician's Apology* would be printed. The sale price would be 3 shillings, 6 pence each, roughly 8 pounds in today's British currency.

Postcards from Hardy requested that presentation copies be sent to colleagues, including Broad and Littlewood, the physicist Sir Arthur Eddington, chemist and novelist C. P. Snow, cricketer John Lomas (to whom he dedicated the book), and his sister Gertrude, with whom he was very close. He also requested copies be sent to colleagues in the United States.

A letter to him, dated May 29, 1941, is a reminder of the devastation being caused by the war: "...copies which were being bound for a second shipment have unfortunately been destroyed by enemy action at one of our binders."

A Mathematician's Apology sold well. Additional printings of 2,500 copies were made in 1941. Another 2,000 copies were printed in 1948, the year following Hardy's death.

In June 1952, Hardy's sister wrote to the Press:

> As *A Mathematician's Apology* is now impossible to get, both first hand and second hand, I expect that you will in time be reprinting it; I think that it would be a good idea to have a photograph of my brother in it granted that it did not make it too expensive. I have the negative of which the enclosed photograph is an enlargement; it is an amateur snap and extremely characteristic.

The photograph that she sent would eventually appear on the dust jacket of the second edition, and has become a well-known image of her brother wearing a white suit, seated in a wicker chair. Hardy, peering over his glasses, seems to be examining the photographer. Secretary R. W. David, who was now handling the project, replied to Gertrude enthusiastically. She wrote back a few days later: "I think that it is far the best ever taken of my brother."

Reissuing *A Mathematician's Apology* would be difficult. Inflation in Britain had made it impossible to reprint so small a book at a reasonable sales price. Some sort of material would be needed to extend it. The Rouse Ball Lecture, "Mathematical Proof," which Hardy had delivered in 1928 was considered. So was "What is Geometry?," Hardy's Presidential Address to the Mathematical Association of America in 1925. Nothing seemed appropriate.

In 1959, 11 years later, chemist and writer C. P. Snow suggested that he might write an introduction. It seemed a superb idea. Snow is best remembered for his 1959 essay *The Two Cultures,* describing the purported conflict between scientists and humanists. He had advised Cambridge University Press during the war years. He also had been a close friend of Hardy, and had offered him advice about the book. However, Snow would not commit to a deadline. An internal memorandum from David in September 1966 complained that "we have been chasing Snow for copy at roughly yearly intervals."

Then in 1967 word came that Snow (now Lord Snow) had finished his piece. Unfortunately, he had written it for another book, *Variety of Men,* which Macmillan in London and Scribner in New York would soon publish. In addition to Hardy's profile, it would include biographies of H. G. Wells, Einstein, Churchill, Stalin, and other notables.

Could the Syndics get permission to reprint Snow's article? *Variety of Men* would also be serialized in magazines, adding a second layer of legal complication. The idea of finding another writer to introduce Hardy was considered but rejected. A privately printed pamphlet by Hardy on Bertrand Russell's dismissal from Trinity was also considered. David reported that "it would not be at all suitable for marrying to the *Apology.*"

David's determination to pursue Snow was undiminished. On October 17, 1966, he wrote:

> We are agreed that our first choice is to persevere with the original plan. Snow's introduction is fluent, anecdotal, a "lively sketch" such as might be given in a radio talk in commemoration of a great man. Some may consider it shallow and may regret that the great man needs to be introduced by the lesser. But... those who wish Hardy to be more widely appreciated must not scorn the honest populariser.

A few days later, all the problems seemed to have been resolved. David wrote: "You will see from the Syndics'

Hardy leads a team of mathematicians in a cricket match, during a British Association meeting at Oxford in August 1926. His team included E. C. Titchmarsh, Bernard Bosanquet, Edward Linfoot and William Ferrar. (Photograph courtesy of the University of Oxford Mathematical Institute.)

minutes of 21 October that the way is now clear to proceed."

And then there was more trouble. Hardy's sister died, leaving all rights to Hardy's work to the London Mathematical Society. The Society objected to "filling out" *A Mathematician's Apology* with writing by anyone other than Hardy.

In a memorandum dated January 3, 1967, and marked "Urgent," David noted that Cambridge University Press owned the copyright of the book together with control of the project. "Yet I am sure," he added, "that the Syndics would not wish to ride rough-shod over the wishes of the [London Mathematical Society]. ... The project must for the moment be put absolutely on ice until I get a further reply from the London Mathematical Society."

Apparently an agreement was reached, and the second edition of *A Mathematician's Apology* was in bookstores by the end of 1967. Snow's contribution added literary charm. It began:

> It was a perfectly ordinary night at Christ's high table, except that Hardy was dining as a guest.... This was 1931, and the phrase was not yet in English use, but in later days they would have said that in some indefinable way he had star quality.

As Cambridge University Press anticipated, the new edition of *A Mathematician's Apology* was received well in the United States. Byron Dobell, an author and editor in New York who helped many young writers, including Tom Wolfe and Mario Puzo, seasoned his praise with a sprinkle of caution:

> It is the kind of book you wish was being read by all your friends at the very moment when you are reading it yourself. It is one of those secret, perfect works that makes most writing seem like a mixture of lead and mush. It's the under-the-counter book we're touting this month. It has nothing to do with anything but the joy of life and mind. The price is $2.95 and, with a title like that, it won't make a nickel for anyone.

Physical Connections

The second edition of *A Mathematician's Apology* appeared as mathematics was becoming increasingly abstract. Many mathematicians rejoiced at this change of direction in their field. Others lamented. If the trend continued, some believed, mathematics would become irrelevant.

One mathematician who celebrated was University of Chicago professor Marshall Stone. His article "The Revolution in Mathematics," which first appeared in the journal *Liberal Education* in 1961 and was reprinted the same year in *American Mathematical Monthly,* saw abstraction bridging areas of mathematics that had previously been isolated islands of thought. The identification of mathematics and logic, he argued, was greatly responsible:

> Mathematics is now seen to have no necessary connections with the physical world beyond the vague and mystifying one implicit in the statement that thinking takes place in the brain. The discovery that this is so may be said without exaggeration to mark one of the most significant intellectual advances in the history of mankind.

Stone noted a paradox: Increasing abstraction was spawning new applications. He listed the mathematical theory of genetics and game theory, as well as the mathematical theory of communications with contributions to linguistics.

A line between pure and applied mathematics exists at most universities today. Too often it is a battle line, witnessing skirmishes over scant resources and bruised egos.

A very different opinion was expressed the following year in "Applied mathematics: What is needed in research and education," published in *SIAM Review.* It was the transcript of a symposium chaired by mathematician H. J. Greenberg. Its panel consisted of mathematicians George Carrier, Richard Courant, Paul Rosenbloom, and physicist C. N. Yang. Stone's article with its embrace of abstraction was discussed with alarm. The panel members urged a more traditional vision of mathematics, one that draws its inspiration from science. Courant's warning sounded like a review of *A Mathematician's Apology:*

> We must not accept the old blasphemous nonsense that the ultimate justification of mathematical science is "the glory of the human mind." Mathematics must not be allowed to split and diverge towards a "pure" and an "applied" variety.

Despite Courant's warning, a line between pure and applied mathematics exists at most universities today. Too often it is a battle line, witnessing skirmishes over scant resources and bruised egos. It is a line that has perhaps been blurred a bit by pure mathematicians' widespread use of computers and technology's urgent need for sophisticated algorithms. Mathematicians who share Hardy's sentiments might feel reluctant to express them in the face of soaring costs of higher education. Students with mounting debts have become increasingly impatient with teachers who digress from material directly needed for their exams. Administrators drool over research grants in medicine and cyber-security while finding less filling the meager grants awarded in pure mathematics.

The line between pure and applied mathematics might be blurred, but it will not soon be erased. As long as it exists, G. H. Hardy's *A Mathematician's Apology* will be read and—usually—enjoyed. No finer summary can be offered than that written by J. F. Randolph in his 1942 review:

> This book is not only about mathematics, it is about ideals, art, beauty, importance, significance, seriousness, generality, depth, young men, old men, and G. H. Hardy. It is a book to be read, thought about, talked about, criticized, and read again.

Acknowledgments

The author is indebted to Dr. Samuel Harrison, Editor (Mathematical Sciences), Cambridge University Press, and Dr. Rosalind Grooms, Archivist, Cambridge University Press, for arranging access to files related to Hardy and *A Mathematician's Apology.* Travel support was provided by a collaborative research grant by the Simon's Foundation. Thanks also to Dr. Susan G. Williams for her advice, technical help, and encouragement.

Bibliography

Bell, E. T. 1937. *Men of Mathematics.* New York: Simon and Schuster.

Bell, E. T. 1942. Confessions of a mathematician (Review of G. H. Hardy, *A Mathematician's Apology*). *The Scientific Monthly* 54:81.

Broad, C. D. 1941. Review of G. H. Hardy, *A Mathematician's Apology*. *Philosophy* 16: 323–326.

Cambridge University Press. *Press Syndicate Minute Books, 1935–1947.* Cambridge, UK: Cambridge University Library.

Cambridge University Press. *Author file: G. H. Hardy.* Cambridge, UK: Cambridge University Library.

Carrier, G. F., R. Courant, P. Rosenbloom, C. N. Yang, and H. J. Greenberg. 1962. What is needed in research and education: A symposium. *SIAM Review* 4:297–320.

Dobell, B. 1967. Review of G. H. Hardy, *A Mathematician's Apology. Book World* (a supplement to the *Chicago Tribune*) December 15.

Grand'Combe, F. 1943. Review of G. H. Hardy, *A Mathematician's Apology. The Journal of Education* 175:369–370.

Greene, G. 1940. The austere art (Review of G. H. Hardy, *A Mathematician's Apology.*) *The Spectator* December 20:18.

Hardy, G. H. 1925. What is geometry? (Presidential address to the Mathematical Association). *The Mathematical Gazette* 12(175):309–316.

Hardy, G. H. 1929. Mathematical proof. *Mind, A Quarterly Review* 38(149)1–25.

Hardy, G. H. 1940. Mathematics in wartime. *Eureka* 3:12–15.

Hardy, G. H. 1940. *A Mathematician's Apology.* Cambridge, UK: Cambridge University Press.

Hardy, G. H. 1942. *Bertrand Russell and Trinity.* Cambridge, UK: Cambridge University Press.

Hogben, L. 1937. *Mathematics for the Million.* NewYork: W. W. Norton and Company.

Pólya, G. 1969. Mathematicians I have known. *The American Mathematical Monthly* 76:746–753.

Randolph, J. F. 1942. Review of G. H. Hardy, *A Mathematician's Apology. The American Mathematical Monthly* 49:396–397.

Snow, C. P. 1966. *Variety of Men.* New York: Charles Scribner's and Sons.

Soddy, F. 1941. Qui S'Accuse S'Acquitte (Review of G. H. Hardy, *A Mathematician's Apology*). *Nature* 147:3–5.

Stone, M. 1961. The revolution in mathematics. *The American Mathematical Monthly* 68:715–734.

Titchmarsh, E. C. 1949. Godfrey Harold Hardy (1877–1947). *Obituary Notices of Fellows of the Royal Society* 6:446–461.

For relevant web links, consult this issue of *American Scientist Online:*

https://www.americanscientist.org/magazine/issues/2015/november-december

What Our Readers Are Saying

I read most of the articles in each issue. I find them informative and/or entertaining.

I appreciate the diversity of articles and their accessibility to non-experts. I learn a lot about science going on outside my discipline (chemistry) and also about the history and social issues associated with doing science (recent article about women's contributions to Nobel Prize).

I look forward to every issue. Like the diversity of topics very much. Always enjoy Henry Petroski's articles. Please keep the book reviews.

Slicing Sandwiches, States, and Solar Systems

Can mathematical tools help determine what divisions are provably fair?

Theodore P. Hill

Gerrymandering is making headlines once again, with a case already before the Supreme Court regarding partisan redistricting in Wisconsin, and another from Pennsylvania waiting in the wings. At the core of the problem of redrawing congressional districts is the issue of "fairness," and that is tricky business indeed. The general subject of fair division has been studied extensively using mathematical tools, and some of that study has proved very useful in practice for problems such as dividing estates or fishing grounds. For gerrymandering, however, there is still no widely accepted fair solution. On the contrary, this past October Pablo Soberón of Northeastern University showed that a biased cartographer could apply mathematics to gerrymander on purpose, without even using strange shapes for the districts. The underlying idea traces back to one of mathematicians' favorite theorems, which dates back to World War II.

The late 1930's were devastating years for the Polish people, but they were years of astonishing discovery for Polish mathematicians. Between the rock of the Great Depression and the hard place of impending invasion and occupation by both Nazi and Soviet armies, a small group of mathematicians from the university in Lwów (today Lviv) met regularly in a coffee shop called the Scottish Café to exchange mathematical ideas. These ideas were not the mathematics of complicated calculations (which were then done with the aid of slide rulers), but rather were very general and esthetically beautiful abstract concepts, soon to prove extremely powerful in a wide variety of mathematical and scientific fields.

The café tables had marble tops, and could easily be written on in pencil and then later erased like a slate blackboard. Since the group often returned to ideas from previous meetings, they soon realized the need for a written record of their results, and purchased a large notebook for documenting the problems and answers. The book, kept in a safe place by the café headwaiter and produced by him upon the group's next visit, was a collection of these mathematical questions, both solved and unsolved, that decades later became known in international mathematical circles as the *Scottish Book*.

The general subject of fair division has been studied extensively using mathematical tools, and some of that study has proven useful for problems such as dividing estates or fishing grounds.

The Ham Sandwich Problem

Problem No.123 in the book, posted by Hugo Steinhaus, a senior member of the café mathematics group and a professor of mathematics at the University of Lemberg (now the University of Lviv), was stated as follows:

"Given are three sets A_1, A_2, A_3 located in the three-dimensional Euclidean space and with finite Lebesgue measure. Does there exist a plane cutting each of the three sets A_1, A_2, A_3 into two parts of equal measure?"

To bring this question to life for his companions, Steinhaus illustrated it with one of his trademark vivid examples, one that reflected the venue of their meetings, and also perhaps their imminent preoccupation with daily essentials: Can every ordinary ham sandwich consisting of three ingredients, say bread, ham, and cheese, be cut by a planar slice of a knife so that each of the three is cut exactly in half? (*See the figure at the bottom of page 44.*)

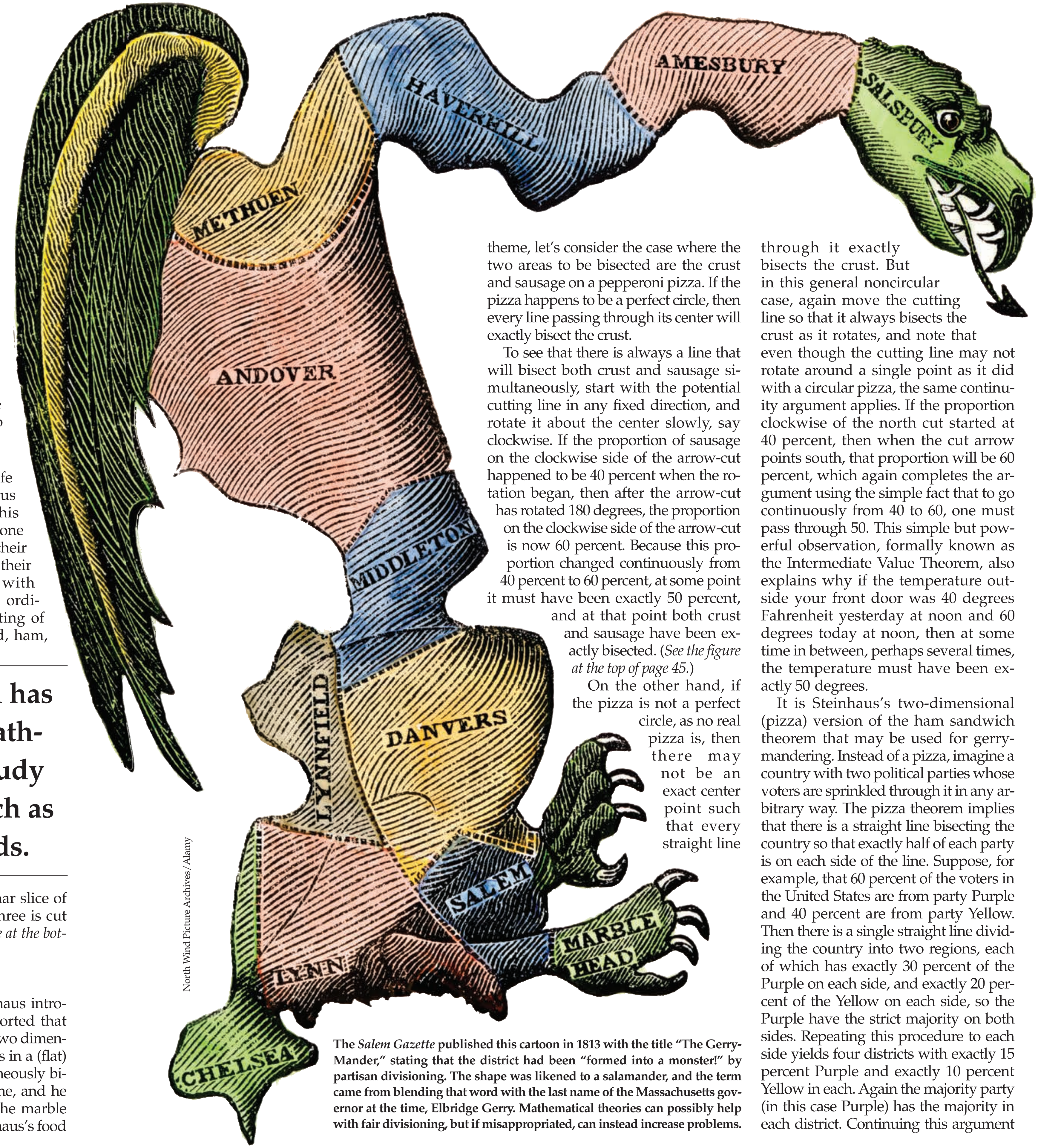
A Simpler Problem

At the meeting where Steinhaus introduced this question, he reported that the analogous conclusion in two dimensions was true: Any two areas in a (flat) plane can always be simultaneously bisected by a single straight line, and he sketched out a solution on the marble tabletop. In the spirit of Steinhaus's food theme, let's consider the case where the two areas to be bisected are the crust and sausage on a pepperoni pizza. If the pizza happens to be a perfect circle, then every line passing through its center will exactly bisect the crust.

To see that there is always a line that will bisect both crust and sausage simultaneously, start with the potential cutting line in any fixed direction, and rotate it about the center slowly, say clockwise. If the proportion of sausage on the clockwise side of the arrow-cut happened to be 40 percent when the rotation began, then after the arrow-cut has rotated 180 degrees, the proportion on the clockwise side of the arrow-cut is now 60 percent. Because this proportion changed continuously from 40 percent to 60 percent, at some point it must have been exactly 50 percent, and at that point both crust and sausage have been exactly bisected. (*See the figure at the top of page 45.*)

On the other hand, if the pizza is not a perfect circle, as no real pizza is, then there may not be an exact center point such that every straight line through it exactly bisects the crust. But in this general noncircular case, again move the cutting line so that it always bisects the crust as it rotates, and note that even though the cutting line may not rotate around a single point as it did with a circular pizza, the same continuity argument applies. If the proportion clockwise of the north cut started at 40 percent, then when the cut arrow points south, that proportion will be 60 percent, which again completes the argument using the simple fact that to go continuously from 40 to 60, one must pass through 50. This simple but powerful observation, formally known as the Intermediate Value Theorem, also explains why if the temperature outside your front door was 40 degrees Fahrenheit yesterday at noon and 60 degrees today at noon, then at some time in between, perhaps several times, the temperature must have been exactly 50 degrees.

It is Steinhaus's two-dimensional (pizza) version of the ham sandwich theorem that may be used for gerrymandering. Instead of a pizza, imagine a country with two political parties whose voters are sprinkled through it in any arbitrary way. The pizza theorem implies that there is a straight line bisecting the country so that exactly half of each party is on each side of the line. Suppose, for example, that 60 percent of the voters in the United States are from party Purple and 40 percent are from party Yellow. Then there is a single straight line dividing the country into two regions, each of which has exactly 30 percent of the Purple on each side, and exactly 20 percent of the Yellow on each side, so the Purple have the strict majority on both sides. Repeating this procedure to each side yields four districts with exactly 15 percent Purple and exactly 10 percent Yellow in each. Again the majority party (in this case Purple) has the majority in each district. Continuing this argument



North Wind Picture Archives/Alamy

The *Salem Gazette* published this cartoon in 1813 with the title "The Gerry-Mander," stating that the district had been "formed into a monster!" by partisan divisioning. The shape was likened to a salamander, and the term came from blending that word with the last name of the Massachusetts governor at the time, Elbridge Gerry. Mathematical theories can possibly help with fair divisioning, but if misappropriated, can instead increase problems.

Theodore P. Hill is a professor emeritus of mathematics at Georgia Institute of Technology, and currently research scholar in residence at California Polytechnic State University in San Luis Obispo. He received his PhD in mathematics from the University of California, Berkeley. One of his hobbies is tracking down early American mathematics books, and the resulting collection now resides at the Bancroft Library at UC Berkeley. Website: http://www.math.gatech.edu/~hill

ZUMA Press, Inc./Alamy

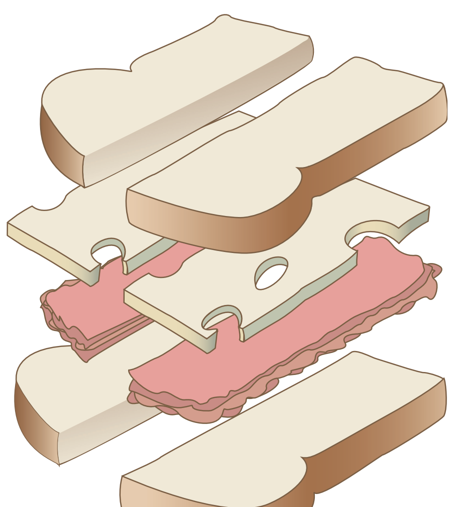
Protesters rally outside the United States Supreme Court in October 2017 while the court hears arguments in a case about gerrymandering in Wisconsin. The court has not considered the constitutionality of gerrymandering in more than a decade, and the court has not previously been able to agree on a standard for when a redistricting map goes too far for the sake of partisanship. A 2013 poll found that across party lines, seven in ten Americans agreed that those who stand to benefit from drawing electoral lines should not have a say in the way those lines are drawn.

shows that whenever the number of desired districts is a power of two, there is always a straight-line partition of the country into that number of districts so that the majority party also has the majority of votes in every single district. (*See the map on page 46.*)

This repeated-bisection argument may fail, however, for odd numbers of desired districts. Sergei Bespamyatnikh, David Kirkpatrick, and Jack Snoeyink of the University of British Columbia, however, found a generalization of the ham sandwich theorem that does the trick for any number of districts, power of two or not. They showed that for a given number of Yellow and Purple points in the plane (no three of which are on a line), there is always a subdivision of the plane into any given number of convex polygons (districts) each containing exactly the same numbers of Yellow points in each district, and the same number of Purple. (*See the map on page 47.*)

Direct application of the ham sandwich theorem would not fix the gerrymandering problem, but would make it worse.

In his application of this theorem to gerrymandering, Soberón observed that for any desired number of districts, this theorem implies that there is always a subdivision into that number of polygonal districts so that each district has exactly the same number of Purple, and exactly the same number of Yellow. Whichever party has the overall majority in the country will also have the majority in every district. Thus, as he found, a direct application of ham sandwich theory would not help fix the problem, but would actually make it worse, and the electorate should be wary if the person drawing congressional maps knows anything about that theory. No wonder the Supreme Court balked on all three of the most recent cases it has heard on partisan gerrymandering.



Illustrations by Theodore P. Hill and Barbara Aulicino

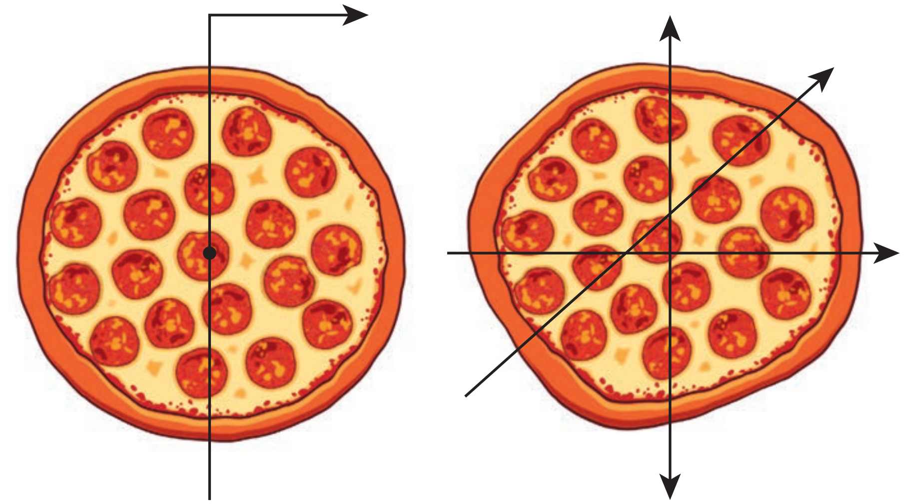
The ham sandwich problem asks whether any three objects, such as the ham, cheese, and bread in a sandwich, can be bisected simultaneously by a single straight cut of a knife. The objects need not be connected to one another, or even to themselves—the bread, for example, might already be in two slices, or even broken into crumbs and scattered about.

The Scottish Café

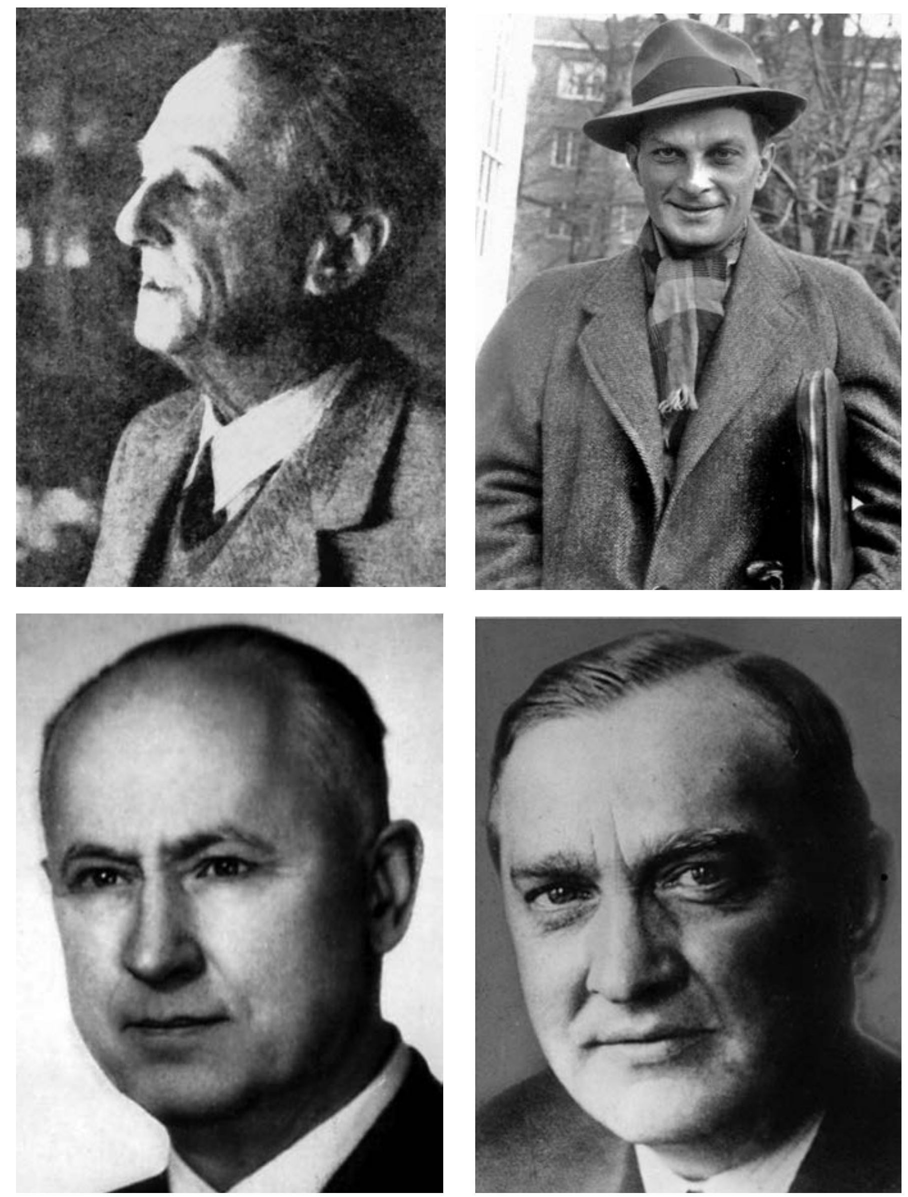
After giving his argument for the two-dimensional case of the ham sandwich theorem, Steinhaus then challenged his companions to prove the 3-dimensional version. The same basic Intermediate Value Theorem argument of continuity that worked for the pizza theorem will not settle the "ham sandwich" Problem 123 question, simply because there is no single "direction" to move a given starting plane passing through the sandwich, guaranteeing a return to the same spot having bisected both of two other objects somewhere along the way.

Two gifted students and protégés of Steinhaus, Stefan Banach and Stanisław Ulam, were also members of the Scottish Café group. Using a discovery Ulam had made around the same time with Karol Borsuk, another Scottish Café comrade, Banach was able to prove the sandwich conjecture of Steinhaus. The key to Banach's proof, called the Ulam-Borsuk Theorem, was another general continuity theorem similar in spirit to the Intermediate Value Theorem, but much more sophisticated. Steinhaus also brought that abstract theorem to life with another of his colorful real-life examples: the Ulam-Borsuk Theorem, he said, implies that at any given moment in time there are two antipodal points on the Earth's surface that have the same temperature and the same atmospheric pressure.

If there are more than three solid objects, or more than two regions in the plane, then it may not be possible to bisect all of them simultaneously with a single plane (or line), as can easily be seen in the case where four small balls are located at the vertices of a pyramid. Also the conclusion of bisection cannot generally be relaxed. For example, if your goal is to split a pizza (or political territory) into two pieces so that one side contains exactly 60 percent of each, that may not always be possible. (*See the figure at the bottom of page 47.*)

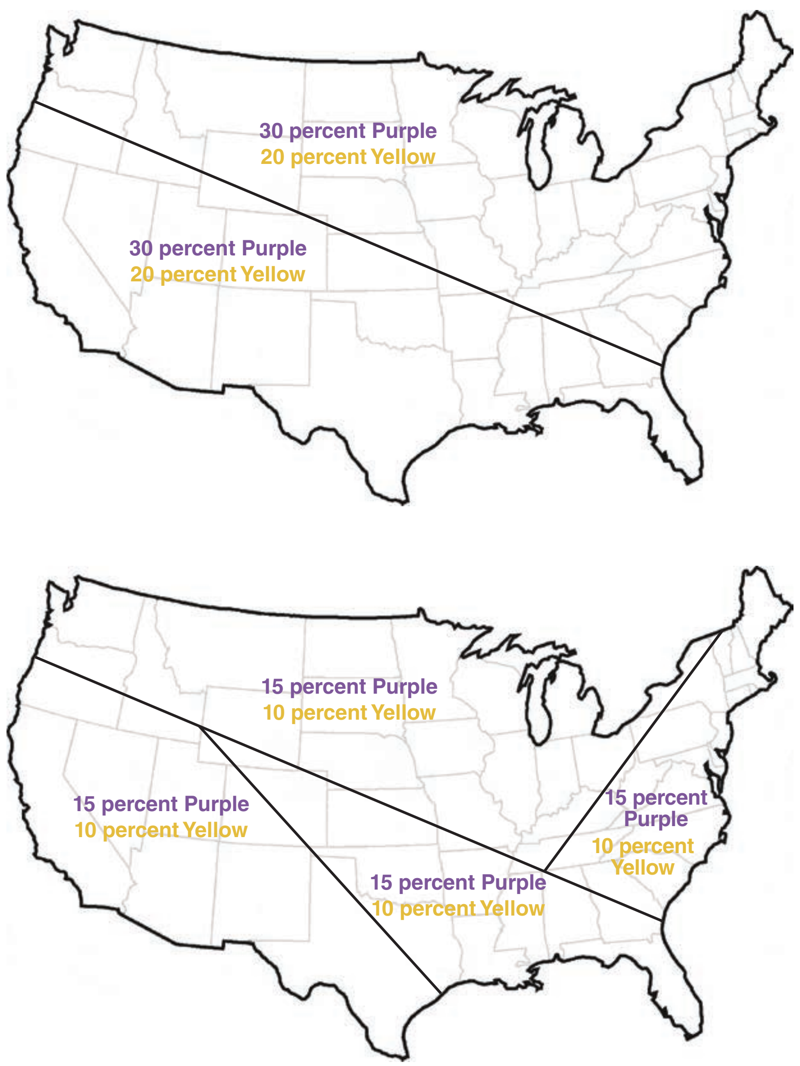


If a pizza is a perfect circle, then every line through the center will bisect the crust. If the cut started with 40 percent of the sausage in the direction of the arrow, after rotating 180 degrees, 60 percent of the sausage will be in the direction of the arrow. So somewhere in between, the line will hit 50 percent and the same cutting line will bisect both crust and sausage. If the pizza is not a perfect circle, the crust-bisecting lines may not all pass through the same point, but the same argument applies.



Wikimedia Commons

Members of the Scottish Café mathematicians who worked on the ham sandwich theorem in two and three dimensional cases included (*clockwise from top left*) Hugo Steinhaus, Stanisław Ulam, Stefan Banach, and Karol Borsuk.



According to the two-dimensional (pizza) version of the ham sandwich theorem, there is a straight line across the United States so that exactly half of the Purple and half of the Yellow party voters are on either side (*top*). Bisecting each of those (*bottom*), the same argument shows that there are four regions with equal numbers of Purple and equal numbers of Yellow in each of them. Thus the party with the overall majority also has the majority in each of the districts.

Generalizations

During World War II, the statement of this colorful and elegant new mathematical result—that any three fixed objects simultaneously can be bisected by a single plane—somehow made it through enemy territory and across the Atlantic, long before email or smartphones or Skype. Mathematicians Arthur Stone and John Tukey at Princeton University learned about this new gem of a theorem via the international mathematics grapevine, and improved the result to include nonuniform distributions, higher dimensions, and a variety of other cutting surfaces and objects. The new Stone and Tukey extensions also showed, for example, that a single circle simultaneously can bisect any three shapes in the plane. For example, there is a location for a telecommunications satellite and a power level so that its broadcasts will reach exactly half the Yellow, half the Purple, and half the Teal (Independents). (*See the map on page 48.*)

Formally speaking, of course, drawing a line to bisect two discrete mass distributions such as Yellow and Purple voters may require splitting one of the voter-points, which may not always be possible (or desirable). If a distribution has an odd number of indivisible points of one type, for example, then clearly no line can have exactly half those points on each side of the line. Inspired by the success of my PhD advisor Lester Dubins in addressing a different fair division problem involving indivisible points (professors, in that case), I wondered whether the conclusion of the ham sandwich theorem might be extended to also include mass distributions with indivisible points—such as grains of salt and pepper sprinkled on a table—by replacing the notion of exact bisection of distributions by a natural generalization of the statistical notion of a median.

Recall that a median of a distribution, say of house prices in a neighborhood, is a price such that no more than half of all the house values are below and no more than half are above that price. Extending this notion to higher dimensions yields the concept of median lines, median planes, and median hyperplanes in higher dimensions. Using the Ulam-Borsuk Theorem again, but this time applied to a different "midpoint median" function, it was straightforward to show that for any two arbitrary random distributions in the plane, or any three in space, there is always a line median or plane median, respectively, that has no more than half of each distribution on each side.

Some 20 years later, Columbia University economist Macartan Humphreys used this result to solve a problem in cooperative game theory. In a setting where several groups must agree on allocations of a fixed resource (say, how much of a given disaster fund should be allocated to medical, power, housing, and food), the objective is to find an allocation that no winning coalition could override in favor of another allocation. He showed that such equilibrium allocations exist precisely when they lie on "ham sandwich cuts."

Touching Planes

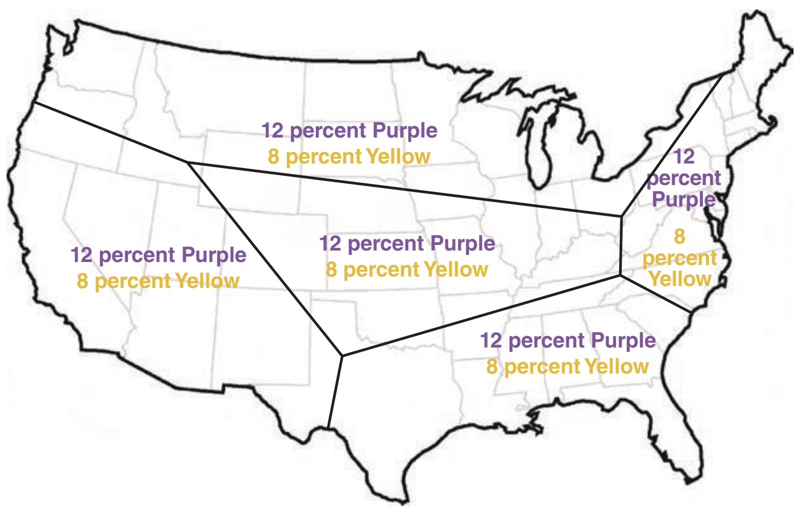
In explaining the beauties of the ham sandwich theorem to nonmathematician friends over beer and pizza, one of my companions noticed that often there is more than one bisecting line (or plane), and we saw that some bisecting lines might touch each of the objects, whereas others may not. I started looking at this observation more closely and discovered that in every case, I could always find a bisecting line or plane that touched all the objects. When I could not find a reference or proof of this concept, I posed the question to my Georgia Tech friend and colleague John Elton, who had helped me crack a handful of other mathematical problems: Is there always a bisecting plane (or hyperplane, in dimensions greater than 3) that also touches each of the objects?

Together, he and I were able to show that the answer is yes, which strengthens the conclusion of the classical ham sandwich theorem. For example, this improved version implies that at any instant in time, in our Solar System there is always a single plane passing through three bodies—one planet, one moon, and one asteroid—that simultaneously bisects the planetary, the lunar, and the asteroidal masses in the Solar System. (*See the figure at the bottom of page 48.*)

Diverse Divisions

The ideas underlying the ham sandwich theorem have also been used in diverse fields, including computer science, economics, political science, and game theory. When I asked my friend Francis Su, Harvey Mudd College mathematician and fair-division expert, about his own applications of the ham sandwich theorem, he explained how he and Forest Simmons of Portland Community College had used ham sandwich results to solve problems in *consensus-halving*. In particular, they used it to show that given a territory and $2n$ explorers, two each of n different specialties (two zoologists, two botanists, two archeologists, etc.), there always exists a way to divide the territory into two regions and the people into two teams of n explorers (one of each type) such that each explorer is satisfied with their half of the territory.

As a more light-hearted application during a keynote lecture at Georgia Tech, Tel Aviv University mathematician Noga Alon described a discrete analog of the ham sandwich theorem for splitting a necklace containing various types of jewels, as might be done, he said, by mathematically oriented thieves who steal a necklace and wish to divide it fairly between them. Even though it had been offered as an amusement, his result had applications, including to VSLI (Very Large Scale Integrated) circuit designs where an integrated chip composed of two different types of nodes is manufactured in the shape of a closed circuit (much like a necklace), and may be restructured after fabrication by cutting and regrouping the pieces. Alon's theorem answers this question: How many cuts need to be made of the original circuit in order to bisect it into two parts, each containing exactly half of each type of node?

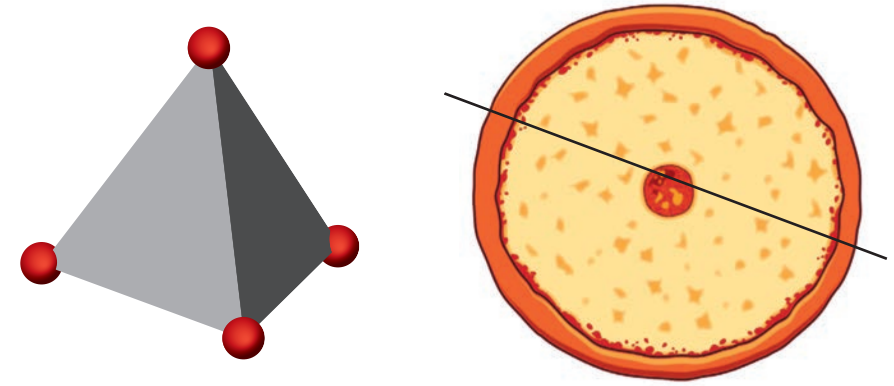


For odd numbers of desired districts, the repeated-bisection argument of the two-dimensional version of the ham sandwich theorem may fail. However, a generalization of the theorem works for any number of districts, by showing that for a given number of Purple or Yellow points in a plane (no three of which are on a line), there is always a subdivision of the plane into any given number of convex polygons, each of which contains exactly the same number of Yellow, and the same number of Purple, points.

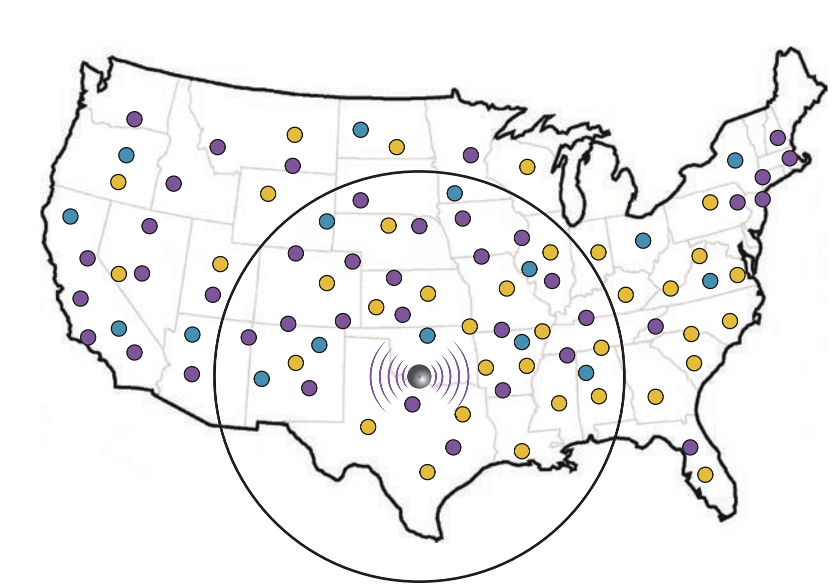
Revisiting the Café

Steinhaus published the proof of the ham sandwich theorem in the local Polish mathematical journal *Mathesis Polska* in 1938, the year of the infamously violent *Kristallnacht*. The Scottish Café mathematics gatherings continued for a few more years, despite the invasion of western Poland by the German army and the Soviet occupation of Lwów from the east, but the difficult times would soon disperse both scholars and their works. Ulam, a young man in his 20s and, like Steinhaus, also of Jewish roots, had left with his brother on a ship for America just two weeks before the German invasion.

Banach, nearing 50 and already widely known for his discoveries in mathematics, was appointed dean of the University of Lwów's department of mathematics and physics by the Soviets after they occupied that city, under the condition that he promised to learn Ukrainian. When the Nazis in turn occupied Lwów, they closed the universities, and Banach was forced to work feeding lice at a typhus research center, which at least protected him from being sent into slave labor. (Banach, like many others, was made to wear cages of lice on his body, so they could feed on his blood. The lice, which are carriers of typhus, were



It is not always possible to bisect simultaneously more than three objects with a single plane (such as points at the corners of a pyramid, shown at left), nor to separate simultaneously three objects by the same unequal ratios. The analog in two dimensions (*right*) shows that the pizza cannot be cut by a straight line so that exactly 60 percent of the crust and 60 percent of the sausage are on the same side of the line.



Mathematicians Arthur Stone and John Tukey of Princeton University extended the ham sandwich theorem to nonuniform distributions, higher dimensions, and a variety of other cutting surfaces and objects. One of their examples showed that a single circle simultaneously can bisect any three shapes in the plane. For instance, it is always possible to design the power and location of a telecommunications satellite so that its broadcasts will reach exactly half the Yellow, half the Purple, and half the Teal (Independents).

used in research efforts to create a vaccine against the disease.) Banach was able to help reestablish the university after Lwów was recaptured by the Soviets in 1944, but died of lung cancer in 1945.

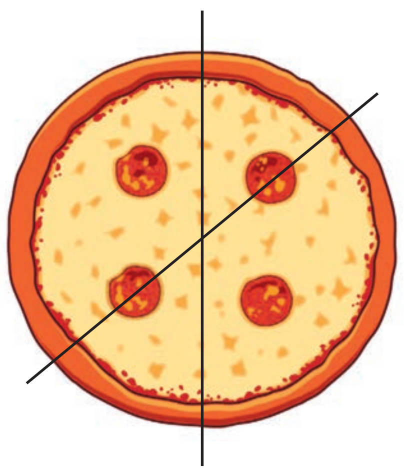
Although the correct statement of the crisp ham sandwich theorem had made it through the World War II mathematical grapevine perfectly, the proper credit for its discoverers was garbled en route, and Stone and Tukey mistakenly attributed the first proof to Ulam. Sixty years later the record was set straight when a copy of Steinhaus's article in the *Mathesis Polska* was finally tracked down, and we now know that Steinhaus posed the problem and published the first paper on it, but it was Banach who actually solved it first, using a theorem of Ulam's.

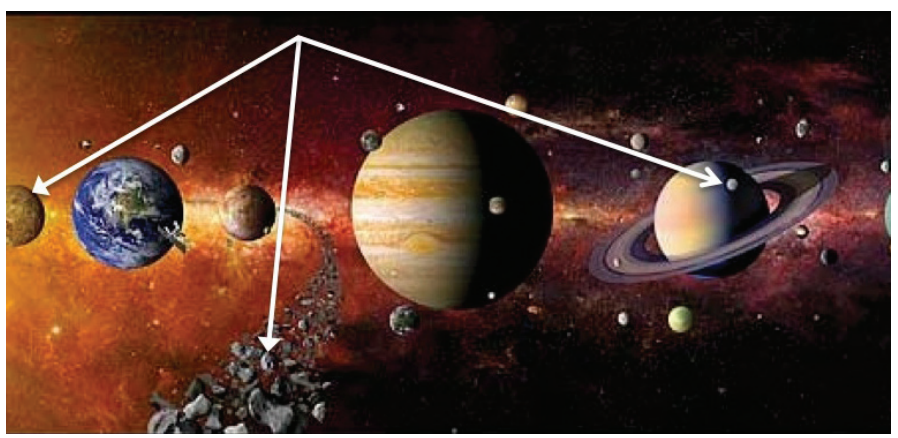
Today Banach is widely recognized as one of the most important and influential mathematicians of the 20th century, and many fundamental theorems, as well as entire basic fields of mathematics, that are based on his work are now among the most extensively used tools in physics and mathematics.

Ulam went on to work as one of the key scientists on the Manhattan Project in Los Alamos, achieving fame in particular for the Teller-Ulam thermonuclear bomb design, and for his invention of Monte Carlo simulation, a ubiquitous tool in economics, physics, mathematics, and many other areas of science, which is used to estimate intractable probabilities by averaging the results of huge numbers of computer simulations of an experiment.

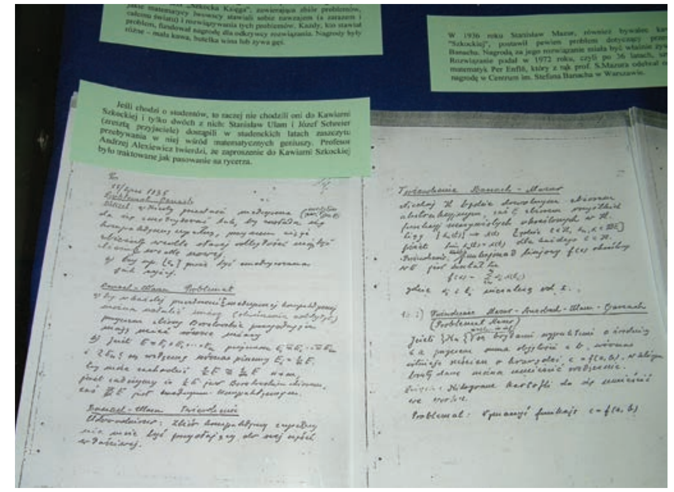
After the war, Steinhaus would have been welcomed with a professorship at almost any university in the world, but he chose to stay in Poland to help rebuild Polish mathematics, especially at the university in Wrocław, which had been destroyed during the war. During those years in hiding, Steinhaus had also been breaking ground on the mathematics of fair division—the study of how to partition and allocate portions of a single heterogeneous commodity, such as a cake or piece of land, among several people with possibly different values. One of Steinhaus's key legacies was his insight to take the common vague concept of "fairness" and put it in a natural and concrete mathematical framework. From there it could be analyzed logically, and has now evolved into common and powerful tools. For example, both the website Spliddit, which provides free mathematical solutions to complicated everyday fair division problems from sharing rent to dividing estates, and the eBay auction system, which determines how much you pay—often below your maximum bid—are direct descendants of Steinhaus's insights on how to cut a cake fairly.

Some bisecting lines or planes may touch each of the objects, whereas others may not, as shown on the pizza below. Nevertheless, there is always a single bisecting line or plane (or hyperplane, in higher dimensions) that touches all of the objects. For example, at any instant in time in our Solar System there is always a single plane passing through three bodies—one planet, one moon, and one asteroid—that simultaneously bisects the planetary, the lunar, and the asteroidal masses in the Solar System.





Stanisław Kosiedowski/Wikimedia Commons



Stako/Wikimedia Commons

The building that housed the Scottish Café where the group of mathematicians met in the late 1930s is still standing in Lwów (*above left*). Copies of the Scottish Book with the original entries by Banach and Ulam are on display at the Library of the Mathematical Institute of the Polish Academy of Sciences in Warsaw (*above right*). The original book remains in the custody of the Banach family, who took it with them after Banach's death and the war ended, when they were required to resettle in Warsaw. Steinhaus kept in touch with the family, and after the war, copied the book by hand to send to Ulam at Los Alamos in 1956. Ulam translated the book into English and had 300 copies made at his own expense. Requests for the book became so numerous that another edition was printed in 1977. After a "Scottish Book Conference" in 1979, in which Ulam participated, the book was again reissued with updated material and additional papers.

These ideas, born of a mathematician living and working clandestinely with little contact with the outside world for long periods of time, and undoubtedly facing fair-allocation challenges almost daily, have inspired hundreds of research articles in fields from computer and political science to theoretical mathematics and physics, including many of my own. Steinhaus eventually became the first dean of the department of mathematics in the Technical University of Wrocław. Although I never met him in person, I had the good fortune to be invited to visit that university in December 2000, and it was my privilege to lodge in a special tower suite right above the mathematics department and to give a lecture in the Hugo Steinhaus Center.

Steinhaus had made the last entry in the original *Scottish Book* in 1941 just before he went into hiding with a Polish farm family, using the assumed name and papers of a deceased forest ranger. The *Scottish Book* itself also disappeared then, and when he came out of hiding and was able to rediscover the book, Steinhaus sent a typed version of it in Polish to Ulam at Los Alamos, who translated it into English. Mathematician R. Daniel Mauldin at the University of North Texas, a friend of Ulam, published a more complete version of the *Scottish Book* including comments and notes by many of the problems' original authors. Their Problem 123, which evolved into the ham sandwich theorem, continues to fascinate and inspire researchers, and Google Scholar shows that eight decades later, several dozen new entries on the topic still appear every few months.

But what about that pesky gerrymandering problem? Negative results in science can also be very valuable; they can illuminate how a certain line of reasoning is doomed to failure, and inspire searches in other directions. That outcome is exactly what happened when the negative ham sandwich gerrymandering result showed that a redistricting attempt might still be radically biased even if the shapes of the districts are quite regular. That insight led researchers to drop the notion of shape as the key criterion, and to look for another approach. The result was a new "efficiency gap" formula that quantifies how much a map is gerrymandered based on vote shares, not on shape. This formula, too, has problems, and in turn it inspired me and my colleague Christian Houdré at Georgia Tech to look for a better measure of "gerrymandered-ness" using combinatorial models involving balls and urns. And so the exciting cycle of scientific discovery that started with the ham sandwich theorem continues.

A great many mathematicians today owe a huge debt to those intrepid Polish academics, and we raise our cups of java to those original Scottish Café mathematicians!

Bibliography

Bellos, A. 2014. *The Grapes of Math*. New York: Simon and Schuster.

Bespamyatnikh, S., D. Kirkpatrick, and J. Snoeyink, J. Generalizing ham sandwich cuts to equitable subdivisions. *Discrete and Computational Geometry* 24:605–622.

Elton, J., and T. Hill. 2011. A stronger conclusion to the classical ham sandwich theorem. *European Journal of Combinatorics* 32:657–661.

Hill, T. 2000. Mathematical devices for getting a fair share. *American Scientist* 88:325–331.

Humphreys, M. 2008. Existence of a multicameral core. *Social Choice and Welfare* 31:503–520.

Mauldin, R. D. (ed.) 2015. *The Scottish Book: Mathematics from the Scottish Cafe*, 2nd Edition. Basel:Birkhäuser.

Simmons, F., and F. Su. 2003. Consensus-halving via theorems of Borsuk-Ulam and Tucker. *Mathematical Social Sciences* 45:15–25.

Soberón, P. 2017. Gerrymandering, Sandwiches, and Topology. *Notices of the American Mathematical Society* 64:1010–1013.

Steinhaus, H. 1938. A note on the ham sandwich theorem. *Mathesis Polska* XI:26–28.

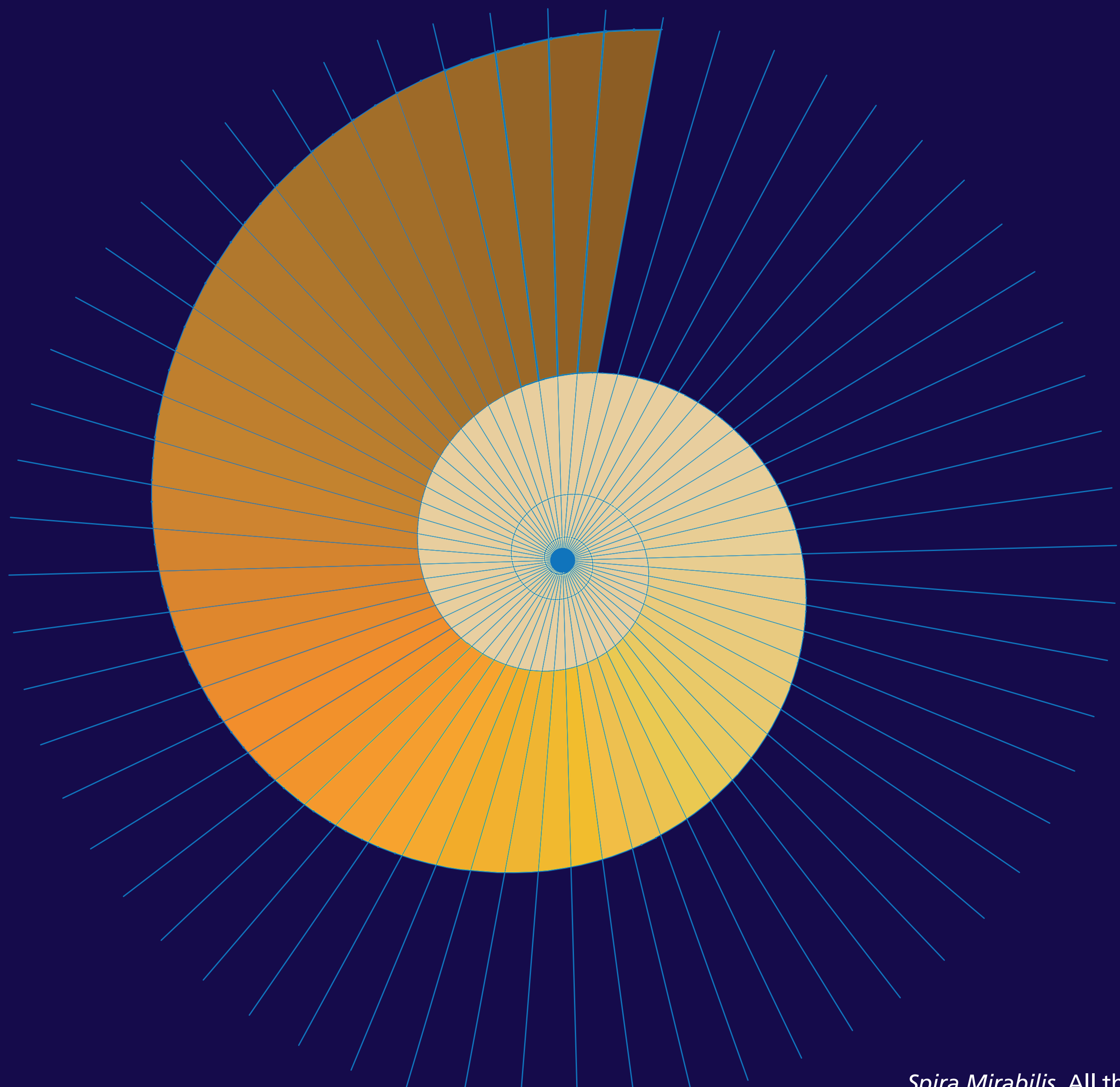
For relevant Web links, consult this issue of *American Scientist Online:*

www.amsci.org/magazine/issues/2018/january-february

Twisted Math and Beautiful Geometry

Four families of equations expose the hidden aesthetic of bicycle wheels, falling bodies, rhythmic planets, and mathematics itself.

Eli Maor and Eugen Jost



Spira Mirabilis. All the geometric artwork in this article is produced by Eugen Jost.

Spira Mirabilis

Of the numerous mathematical curves we encounter in art, geometry, and nature, perhaps none can match the exquisite elegance of the logarithmic spiral. This famous curve appears, with remarkable precision, in the shape of a nautilus shell, in the horns of an antelope, and in the seed arrangements of a sunflower. It is also the ornamental motif of countless artistic designs, from antiquity to modern times. It was a favorite curve of the Dutch artist M. C. Escher (1898–1972), who used it in some of his most beautiful works, such as *Path of Life II*.

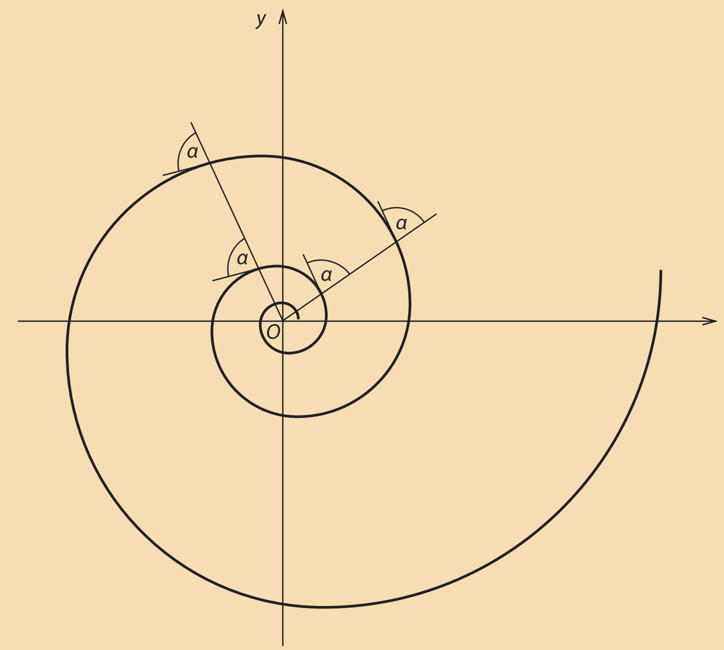
The logarithmic spiral is best described by its polar equation, written in the form $r = e^{a\theta}$, where r is the distance from the spiral's center O (the "pole") to any point P on the curve, θ is the angle between line OP and the x-axis, a is a constant that determines the spiral's rate of growth, and e is the base of natural logarithms. Put differently, if we increase θ arithmetically (that is, in equal amounts), r will increase geometrically (in a constant ratio).

The many intriguing aspects of the logarithmic spiral all derive from this single feature. For example, a straight line from the pole O to any point on the spiral intercepts it at a constant angle α. For this reason, the curve is also known as an equiangular spiral. As a consequence, any sector with given angular width $\Delta\theta$ is similar to any other sector with the same angular width, regardless of how large or small it is. This property is manifested beautifully in the nautilus shell (*right*). The snail residing inside the shell gradually relocates from one chamber to the next, slightly larger chamber, yet all chambers are exactly similar to one another: A single blueprint serves them all.

The logarithmic spiral has been known since ancient times, but it was the Swiss mathematician Jakob Bernoulli (1654–1705) who discovered most of its properties. Bernoulli was the senior member of an eminent dynasty of mathematicians hailing from the town of Basel. He was so enamored with the logarithmic spiral that he dubbed it "spira mirabilis" and ordered it to be engraved on his tombstone after his death. His wish was fulfilled, though not quite as he had intended: For some reason, the mason engraved a linear spiral instead of a logarithmic one. (In a linear spiral the distance from the center increases arithmetically—that is, in equal amounts—as in the grooves of a vinyl record.)The linear spiral on Bernoulli's headstone can still be seen at the cloisters of the Basel Münster, perched high on a steep hill overlooking the Rhine River.

But if a wrong spiral was engraved on Bernoulli's tombstone, at least the inscription around it holds true: *Eadem mutata resurgo*—"Though changed, I shall arise the same." The verse summarizes the many features of this unique curve. Stretch it, rotate it, or invert it, it always stays the same.

A nautilus shell, cut in half to reveal its chambers.



In a logarithmic spiral, a straight line drawn outward from the center always intercepts the spiral at a constant angle.

Notes

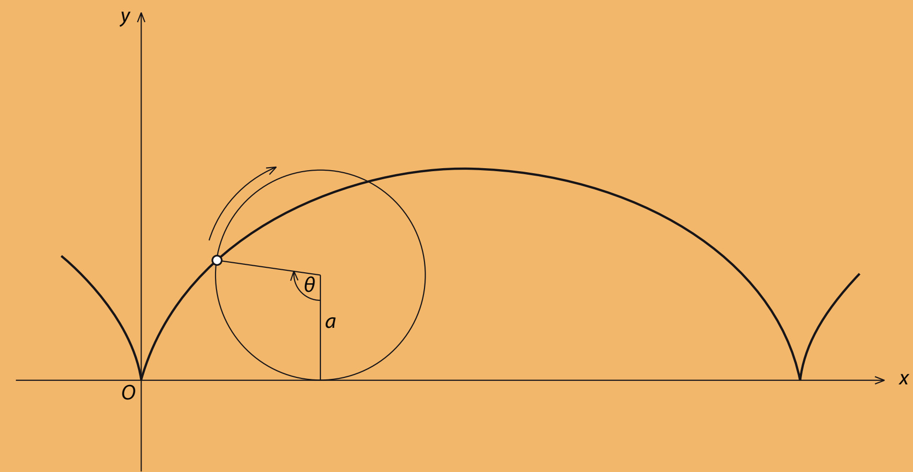
This angle is determined by the constant a; in fact, $\alpha = \cot^{-1}a$. In the special case when $a = 0$, we have $\alpha = 90°$ and the spiral becomes the unit circle $r = e^0 = 1$. For negative values of a, the spiral changes its orientation from counterclockwise to clockwise as θ increases.

For more on the logarithmic spiral, see Maor, *e: The Story of a Number*, chapter 11.

The Cycloid

Rivaling the logarithmic spiral in elegance is the cycloid, the curve traced by a point on the rim of a circle that rolls along a straight line without slipping (*right*). The cycloid is characterized by its arcs and cusps, with each cusp marking the instant when the point on the wheel's rim reaches its lowest position and stays momentarily at rest.

The cycloid has a rich history. In 1673, the Dutch physicist Christiaan Huygens (1629–1695) solved one of the outstanding problems that had intrigued 17th-century scientists: to find the curve down which a particle, moving only under the force of gravity, will take the same amount of time to reach a given final point, regardless of the initial position of the particle. This problem is known as the tautochrone (from the Greek words meaning "the same time"). To his surprise, Huygens found that the curve is an arc of an inverted cycloid. He tried to capitalize on his discovery by constructing a clock whose pendulum was constrained to swing between two adjacent arcs of a cycloid, so that the period of oscillations would be independent of the amplitude. (In an ordinary pendulum this condition holds only approximately.) Unfortunately, although the theory behind it was sound, the performance of Huygens's clock fell short of his expectations.



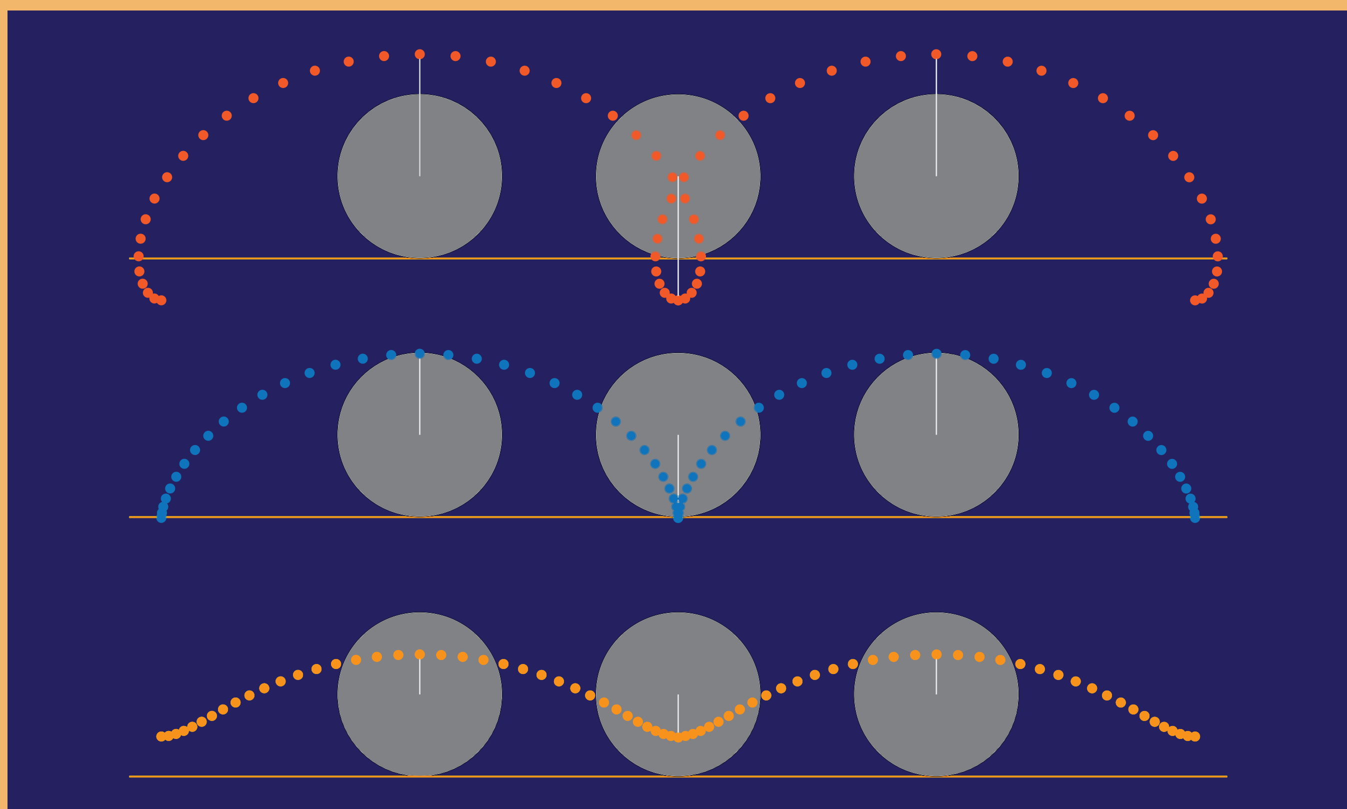
When a circle rolls along a straight line, the path traced by any given point on the outer edge of the circle takes the form of a cycloid.

Shortly thereafter, the cycloid made history again. In 1696 Johann Bernoulli (1667–1748), the younger brother of Jakob (of logarithmic spiral fame), posed this problem: to find the curve along which a particle, again subject only to the force of gravity, will slide down in the least amount of time. You might think the answer should be the straight line connecting the initial and final positions of the particle, but this is not so: Depending on the path's curvature, the particle may accelerate faster at one point and slower at another, showing that the path of shortest *distance* between two points is not necessarily the path of shortest time.

Known as the brachistochrone ("shortest time"), this problem was attempted by some of the greatest minds of the 17th century, including Galileo, who incorrectly thought the required path is an arc of a circle. In the end, five correct solutions were submitted—by Isaac Newton, Gottfried Wilhelm Leibniz, Guillaume de L'Hospital (famous for a rule in calculus named after him), and the Bernoulli brothers, who worked on the problem independently and used different methods. To their surprise, the curve turned out to be an inverted cycloid—the same curve that solved the tautochrone problem. But instead of rejoicing in their success, the two brothers became embroiled in a bitter priority dispute, resulting in a permanent rift between them.

The cycloid had some more surprises in store. Evangelista Torricelli (1608–1647), inventor of the mercury barometer, is credited with finding the area under one arc of the cycloid: The area turned out to be $3\pi a^2$, where a is the radius of the generating circle. A few decades later Christopher Wren (1632–1723), London's venerable architect who rebuilt the city after the Great Fire of 1666, determined that the length of each arc is $8a$; surprisingly, the constant π is not involved. This was one of the first successful rectifications of a curve—finding the arc length between two points on the curve. With the invention of calculus in the decade 1666–1676, such problems could be solved routinely, but in the early 17th century they presented a challenging task.

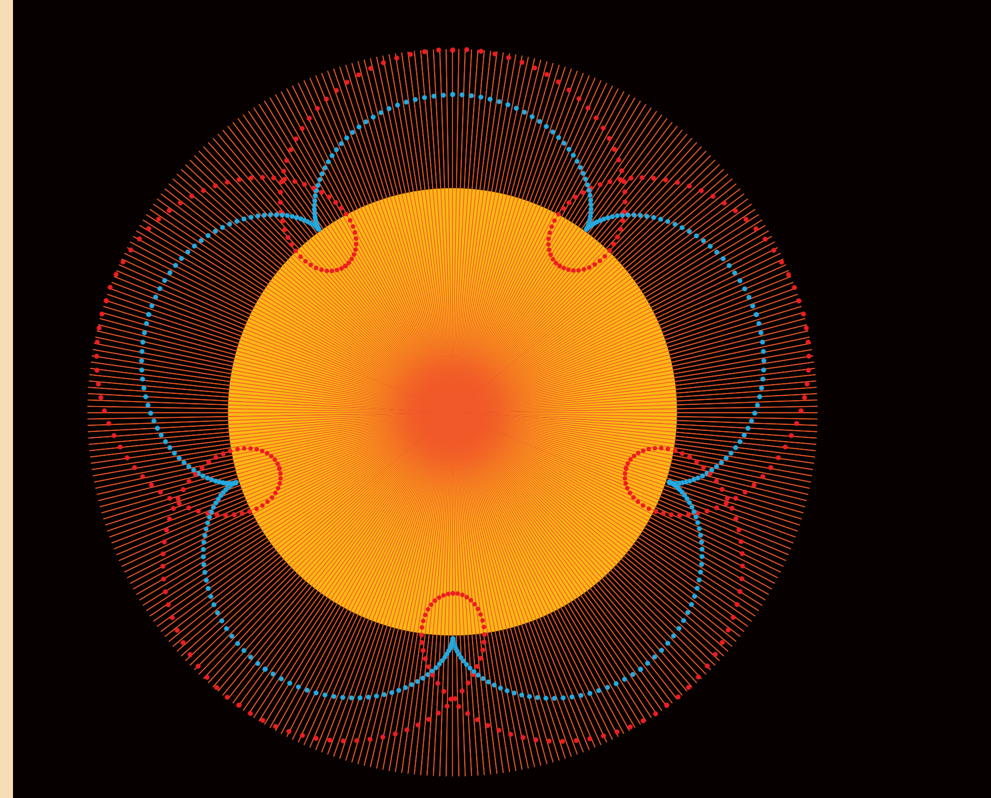
Reflections on a Rolling Wheel (*left*), shows the path of a luminous point attached to a rolling wheel at three different distances from the center. At top, the point is outside the wheel's rim (as on the flank of a railroad car wheel); at the middle, it is exactly on the rim; and at the bottom, inside of it. The top and bottom curves are called prolate and curtate cycloids, respectively, while the middle curve is the ordinary cycloid. You can see the curtate variant at night as the path traced by the reflector on a bicycle wheel while the cyclist moves forward.



Reflections on a Rolling Wheel. Several different types of cycloid *(dotted red, blue, and orange lines)* take shape as the circle rolls along.

Note

For a full history of the cycloid, see the article "The Helen of Geometry" by John Martin, *The College Mathematics Journal* (September 2009, pp. 17–27).



An epicycloid (*blue dotted lines*) is the path traced by a circle as it rolls along the outer edge of another circle. The path of Venus, seen from Earth, appears as a prolate epicycloid (*red dotted lines*).

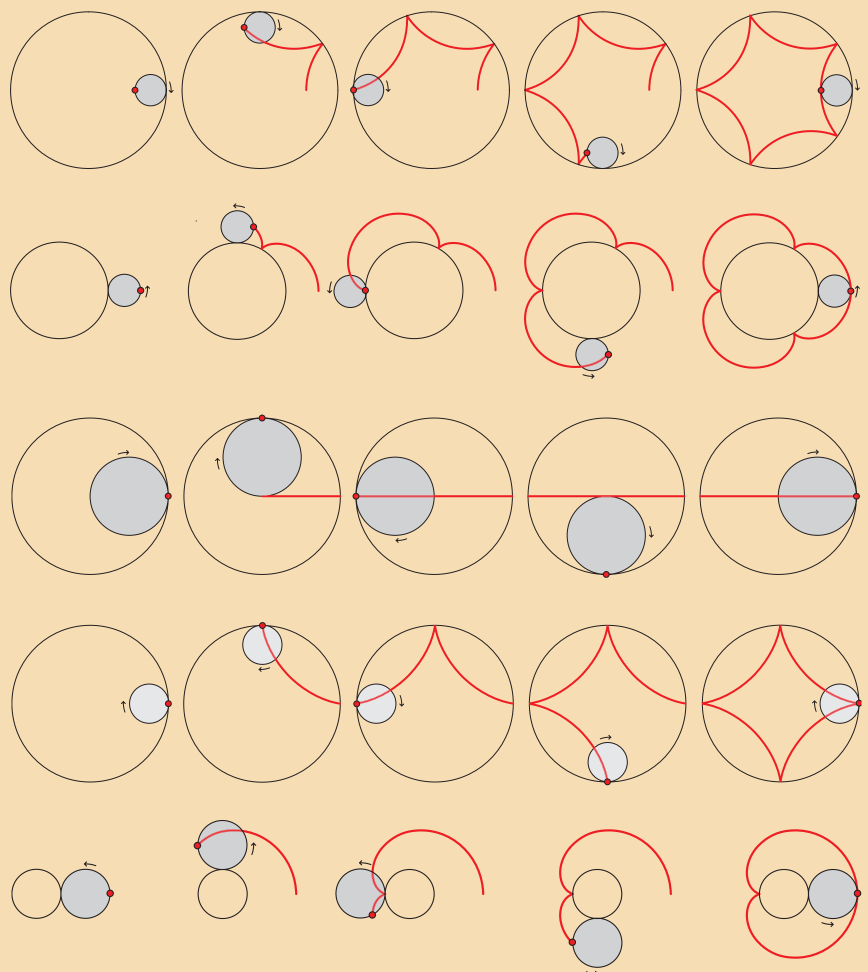
Epicycloids and Hypocycloids

Whereas the cycloid is generated by a point on the rim of a wheel rolling along a straight line, a related type of curve arises from a wheel rolling on the outside of a second, fixed wheel. The resulting curve is an epicycloid (from the Greek epi, meaning "over" or "above"). Alternatively, we can let the wheel roll along the inside of a fixed wheel, generating a hypocycloid (hypo = "under"). The epicycloid and hypocycloid come in a great variety of shapes, depending on the ratio of the radii of the two wheels. Let the radii of the fixed and moving wheels be R and r, respectively. If R/r is a fraction in lowest terms, say m/n, the curve will have m cusps (corners), and it will be completely traced after n full rotations around the fixed wheel. If R/r is not a fraction—if it is irrational—the curve will never close completely, although it will nearly close after many rotations.

For some special values of R/r the ensuing curves can be something of a surprise. For example, when $R/r = 2$, the hypocycloid becomes a straight-line segment: Each point on the rim of the rolling wheel will move back and forth along the diameter of the fixed wheel (*see next page*). Thus, two circles with radii in the ratio 2:1 can be used to draw a straight-line segment! In the 19th century this type of curve provided a potential solution to a vexing problem: how to convert the to-and-fro motion of the piston of a steam engine into a rotational motion of the wheels. It was one of many solutions proposed, but in the end it turned out to be impractical.

When $R/r = 4$, the hypocycloid becomes the star-shaped astroid (from the Greek astron, a star). This curve has some interesting properties of its own. Its perimeter is $6R$ (as with the cycloid, this value is independent of π), and the area enclosed by it is $3\pi R^2/8$, that is, three-eighths the area of the fixed circle. Imagine a line segment of fixed length with its endpoints resting on the x- and y-axes, like a ladder leaning against a wall. When the ladder is allowed to assume all possible positions, it describes a region bound by one-quarter of an astroid. This shows that a curve can be formed not only by a set of points lying on it, but also by a set of lines tangent to it.

Turning now to the epicycloid, the case in which the fixed and the moving wheels have the same radius ($R/r = 1$) is of particular interest: It results in a



Epicycloids and hypocycloids (*red lines*) can take a variety of forms, depending on the size and position of the rotating circle relative to the fixed circle.

cardioid, so called because of its heart-shaped form. This romantic curve has a perimeter of $16R$ and its area is $6\pi R^2$.

The Greek astronomer Claudius Ptolemaeus, or Ptolemy (ca. 85–165 C.E.), invoked epicyclic motion in an attempt to explain the occasional retrograde motion of the planets—a movement from east to west in the sky, instead of the usual west to east. He ascribed to them a complex path in which each planet moved along a small circle whose center moved around Earth in a much larger circle. The resulting epicycle has the shape of a coil wound around a circle. When this model still failed to account for the positions of the planets accurately, more epicycles were added on top of the existing ones, making the system increasingly cumbersome. Finally, in 1609, Johannes Kepler discovered that planets move around the Sun in ellipses, and the epicycles were laid to rest.

The illustration on page 143 shows a five-looped epicycloid (blue) and a prolate epicycloid (*red*) similar to Ptolemy's planetary epicycles. This latter curve closely resembles the apparent path of Venus against the backdrop of the fixed stars. Earth and Venus follow an eight-year cycle during which the two planets and the Sun will be aligned almost perfectly five times. Surprisingly, eight Earth years also coincide with 13 Venusian years, locking the two planets in an 8:13 celestial resonance and giving Fibonacci aficionados one more reason to celebrate!

Notes

We might mention in passing that the astroid has the unusual rectangular equation $x^{2/3} + y^{2/3} = R^{2/3}$.

For nice simulations of how these curves are generated, go to http://mathworld.wolfram.com/Hypocycloid.html; see also http://mathworld.wolfram.com/Epicycloid.html.

For more on the properties of epicycloids and hypocycloids, see Maor, *Trigonometric Delights*, chapter 7.

Steiner's Porism

The first half of the 19th century saw a revival of interest in classical Euclidean geometry, in which figures are constructed with straight-edge and compass and theorems are proved from a given set of axioms. This "synthetic," or "pure," geometry had by and large been thrown by the wayside with the invention of analytic geometry by Pierre de Fermat and René Descartes in the first half of the 17th century.

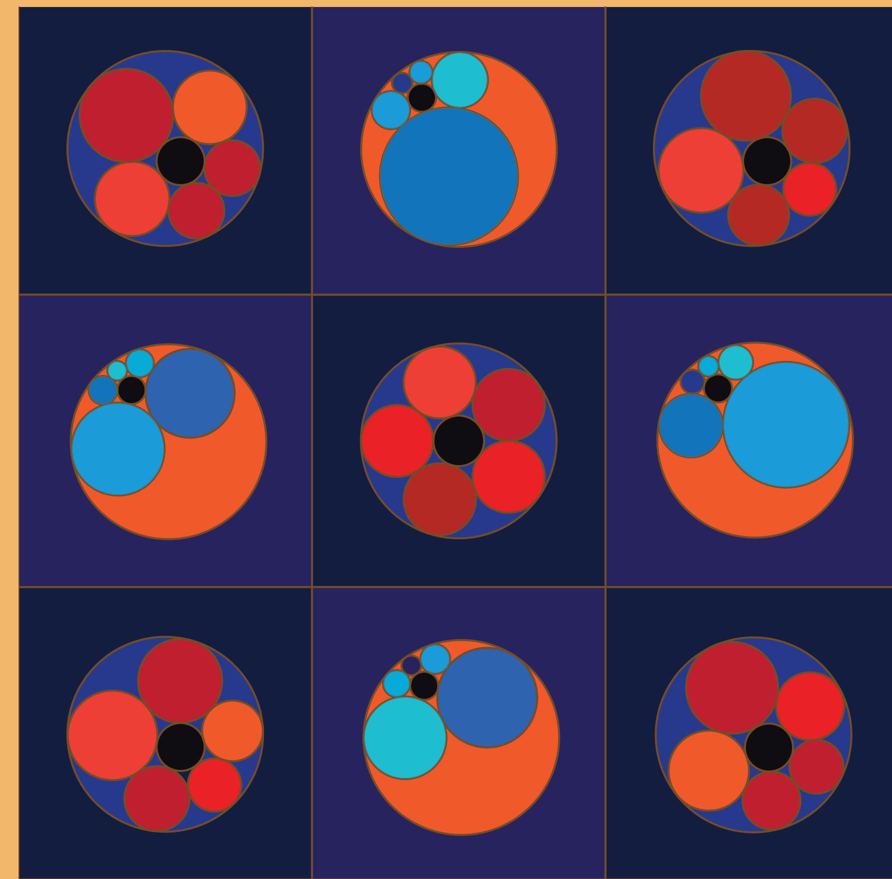
Analytic geometry is based on the idea that every geometric problem could, at least in principle, be translated into the language of algebra as a set of equations, whose solution (or solutions) could then be translated back into geometry. This unification of algebra and geometry reached its high point with the invention of the differential and integral calculus by Newton and Leibniz between 1666 and 1676; it has remained one of the chief tools of mathematicians ever since. The renewed interest in synthetic geometry came, therefore, as a fresh breath of air to a subject that had by that time been considered out of fashion.

One of the chief protagonists in this revival was the Swiss geometer Jacob Steiner (1796–1863). Steiner did not learn how to read and write until he was 14, but after studying under the famous Swiss educator Heinrich Pestalozzi, he became completely dedicated to mathematics. Among his many beautiful theorems we bring here one that became known as Steiner's porism (more on that odd name in a moment).

Steiner considered the following problem: Given two nonconcentric circles, one lying entirely inside the other, construct a series of secondary circles, each touching the circle preceding it in the sequence as well as the two original circles (*see figure at lower right*). Will this chain close upon itself, so that the last circle in the chain coincides with the first? Steiner, in 1826, proved that if this happens for any particular choice of the initial circle of the chain, it will happen for *every* choice.

In view of the seeming absence of symmetry in the configuration, this result is rather unexpected. Steiner devised a clever way of exposing hidden symmetry by inverting the two original circles into a pair of *concentric* circles. As a result, the chain of secondary circles (now inverted) will occupy the space between the (inverted) given circles evenly, like the metal balls between the inner and outer rings of a ball-bearing wheel. These can be moved around in a cyclic manner without affecting the chain.

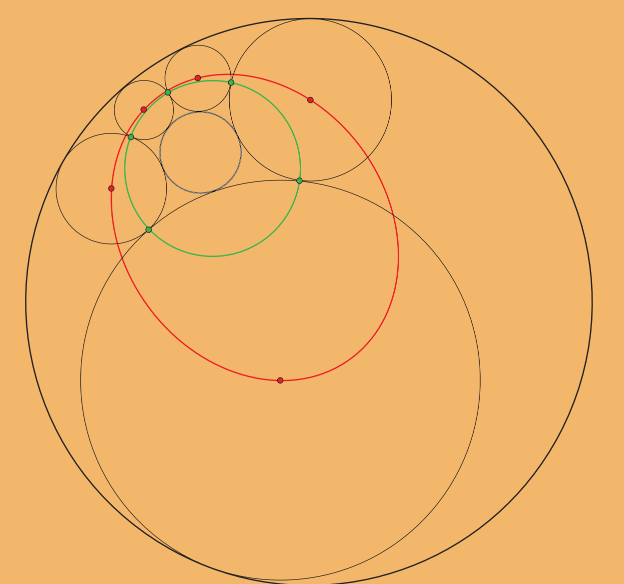
But that's not all: It turns out that the centers of the circles of the Steiner chain always lie on an ellipse (*marked in red*), and the points of contact of adjacent circles lie on yet another circle (*marked in green*).



These nine examples of Steiner chains each consist of one large circle containing six other circles, some overlapping, of various sizes.

The images at left illustrate nine Steiner chains, each comprising five circles that touch an outer circle (*alternately colored in blue and orange*) and an inner black circle. The central panel shows this chain in its inverted, symmetric "ball-bearing" configuration. As happens occasionally, a theorem that has been known in the West for many years turned out to have already been discovered earlier in the East. In this case, a Japanese mathematician, Ajima Chokuyen (1732–1798), discovered Steiner's porism in 1784, almost half a century before Steiner. An old Japanese tradition, going back to the 17th century, was to write a geometric problem on a wooden tablet, called *sangaku*, and hang it in a Buddhist temple or Shinto shrine for visitors to see. A fine example of Steiner's—or Chokuyen's—chain appeared on a sangaku at the Ushijima Chomeiji temple in Tokyo. The tablet no longer exists, unfortunately, but an image of it appeared in a book published about the same time as Steiner's discovery.

It is somewhat of a mystery why this theorem became known as Steiner's porism. You will not find the word *porism* in your usual college dictionary, but the online Oxford English Dictionary defines it as follows: "In Euclidean geometry: a proposition arising during the investigation of some other propositions by immediate deduction from it." Be that as it may, the theorem again reminds us that even good old Euclidean geometry can still hold some surprises within it.



Jacob Steiner's study of nonconcentric circles gave rise to the mathematical proposition named after him.

Notes

Steiner chains enjoy many additional properties. See http://en.wikipedia.org/wiki/Steiner_chain. For a proof of Steiner's porism, see Coxeter, *Introduction to Geometry*, p. 87.

See Hidetoshi and Rothman, *Sacred Mathematics: Japanese Temple Geometry*, p. 292.

 Eli Maor is the author of To Infinity and Beyond *and* The Pythagorean Theorem: A 4,000-Year History, *among other books (all Princeton University Press) and has taught the history of mathematics at Loyola University Chicago. Eugen Jost is a Swiss artist whose work is strongly influenced by mathematics.*

For relevant web links, consult this issue of *American Scientist Online:*

https://www.americanscientist.org/magazine/issues/2014/march-april

Scientists' Nightstand

Stats and Fiction

AN ADVENTURE IN STATISTICS: The Reality Enigma. Andy Field. Illustrated by James Iles. 746 pp. SAGE Publications, 2016. $56.

In graduate school, I searched and searched for a good applied statistics textbook—one that not only explained analyses and how they work but also covered how to prepare and check one's data, write the programming code, and read the output. Like most ecologists, I needed to learn a vast array of analytical techniques. I ended up cobbling together what I needed using several books. *A Primer for Ecological Statistics,* by Nicholas J. Gotelli and Aaron M. Ellison, was fine for checking the basics. For multivariate analyses, I referred to *Analysis of Ecological Communities,* by Bruce McCune and James B. Grace, and *Using Multivariate Statistics,* by Barbara G. Tabachnick and Linda S. Fidell. Over the course of those doctoral research years, as well as when I began teaching undergraduate biology, I picked up a variety of statistics textbooks and put most of them right back down.

Further complicating matters, researchers who rely on statistical analysis of their data must typically be familiar with some sort of programming code to run the numbers. Mastering how to write the code and interpret the output can be big hurdles for early-career scientists—especially as the number of analyses they may need to have in their toolboxes has proliferated. During my last year of dissertation research, I read about the code language for the free program R using Michael J. Crawley's *The R Book.* This freeware had vast online help networks that enabled me to find what I needed fairly quickly and cheaply; it made much nicer visual graphics than SAS software did; and it offered me the ability to do analyses that other statistical software packages couldn't easily perform.

Now, many years later, I have at last encountered a book that provides solid, innovative statistics instruction alongside lessons in coding. And it's fair to say that it does so like no other. Andy Field's *An Adventure in Statistics: The Reality Enigma*—an introductory statistics educational text embedded in a science fiction story with graphic-novel artwork—has caught my attention and kept it. If only I'd had this book back in grad school.

Field, a professor of child psychopathology at the University of Sussex, is the author of the popular textbook *Discovering Statistics Using SPSS* [Statistical Package for the Social Sciences], which has gone through three editions, selling hundreds of thousands of copies. When Field asked his publisher, SAGE Publications, for permission to write a "statistics for dummies" book as part of a series put out by a rival publisher, he was told that if he would write the book for SAGE instead, he would be given complete authorial control—freedom to do whatever he wanted. *An Adventure in Statistics* was the result. Field has created (if you'll forgive the pun) a truly novel textbook: one driven by a fictional plot, full of quirky science fiction tropes, in which readers accompany the protagonist on a quest to learn statistics. Like a standard textbook, it is organized into a logical sequence of instructional chapters, with review questions and activities at the end of each. But unlike most textbooks, the fictional plot guides the reader throughout and is accompanied by comic-book–style illustrations. Field also freely blends elements from the thriller and horror genres into the tale as his protagonist races to locate a missing person and faces a zombie apocalypse. The book is unlike anything else out there, but it works despite—or maybe because of—its peculiarity.

Field uses the book's prologue to set the scene, introducing readers to a dystopian future in which the invention of a "reality prism" has made it possible for anyone wearing the device to see truth objectively and to separate out subjective experience. This invention, developed a few decades before the story's action begins, has brought about a revolution through the demise of not only propaganda and media spin but also religion, art, music, creativity, and people's sense of purpose. When a new World Governance Agency embeds in its citizens Wi-Fi–enabled microchips that record what a person sees, thinks, and hears in real time, a schism emerges: On one side are those who accept the chips in order to join a virtual hive mind; on the other are those who refuse them, preferring instead a steampunk-like love of anachronism. In Field's hands, the reality prism serves as more than an interesting premise. He uses the invention to cheekily make points about the difficulty of defining objectivity, adding depth and dimension to a question at the root of the practice of statistics.

Illustration by Andy Field, © 2016. Courtesy of SAGE Publishing.

Having caught a fleeting glimpse of his missing lover, Zach (*left*) fears she has abandoned him for a life immersed in science. In the second panel, Milton (the cat) makes a call on his Proteus (a device that has replaced smart phones) to a shadowy figure he can see through an attached monocle. The reader is left to wonder who is on the other end of that call. From *An Adventure in Statistics.*

Taking this destabilized world as its backdrop, Field's tale centers on two characters who have been romantic partners for 10 years and share an apartment: Zach, the lead singer in a metal band called The Reality Enigma, who follows his gut feelings, and Alice, a scientist who bases her decisions on evidence. Zach is in awe of Alice's scientific prowess, although he doesn't always understand her work. When Alice disappears, with all records of her existence having been erased, Zach decides that in order to understand her research and why she might have disappeared, he has to learn science and statistics—even though he hates math and admits that it made him feel "inferior and frustrated" in school. His quest brings him into contact with a passel of wacky characters, including Milton—a talking ginger cat that keeps texting him statistics hints and is, incidentally, a scientist trapped in a cat's body—and Celia, a beautiful fan of his music who has a big crush on him and who also happens to work at a mysterious scientific research institution, JIG:SAW, which was mentioned multiple times in the data files that Alice left behind on the day she disappeared.

As Zach progresses through his quest, he receives a comprehensive introduction to statistics. Like many introductory statistics texts, this one starts with basic ideas about sampling designs and the distribution of data, and it ends with a common method for comparing two or more means—analysis of variance, including factorial and repeated-measures designs. The text does not give a comprehensive overview of nonlinear or multivariate models. It covers the basics, however, and provides guidelines for avoiding pitfalls commonly encountered by novice researchers, both of which it does considerably better than many other textbooks I've examined. Each chapter ends with a set of activities and questions (labeled

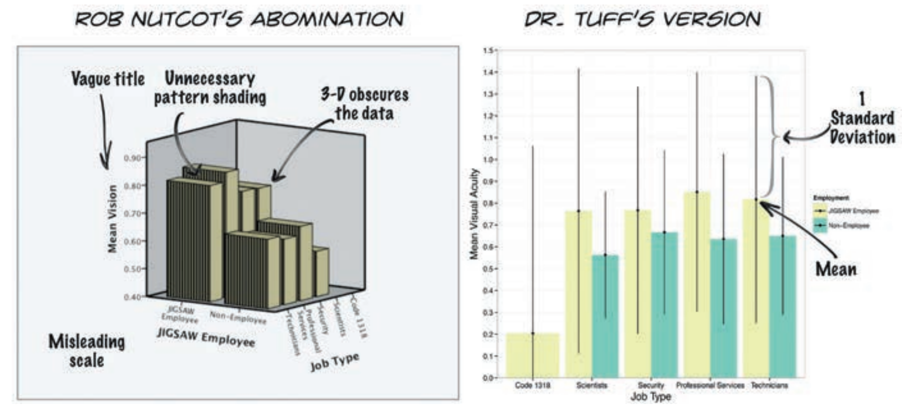


Illustration by Andy Field, © 2016. Courtesy of SAGE.

In a section on the art of presenting data, Field provides several illustrations that contrast a terrible graph or chart (*left*) with a more elegant version (*right*). The flawed examples are attributed to Rob Nutcot, ostensible head of the research institution JIG:SAW. Most of the elegant versions are attributed to a Dr. Sisyphus Tuff, described as "the world expert on displaying data"; his last name is clearly meant to evoke pioneer of data visualization Edward Tufte. From *An Adventure in Statistics.*

"puzzles") that help the reader review the concepts covered. Unlike the standard textbook examples and exercises, however, these consider topics such as zombie rehabilitation, the psychology of cheating on one's partner, and the business of successfully promoting a metal band with merchandise.

In addition, data files and R scripts for some of the problems are available. I like that Field offers these, as well as an ample number of images that show effective data visualizations. The examples of code and output in R for particular analyses are an essential part of an applied statistics textbook if one is using it to teach oneself and is applying the lessons to one's own data. Readers can also find videos of lectures by Field on his YouTube page (http://bit.ly/2kWEhfv), along with tutorials for both his earlier statistics textbook and this one.

Field's clear and fun explanations demonstrate that he is an experienced and conscientious teacher. Through Zach's first-person narration, Field shows that the protagonist's biggest hindrance is his own insecurity about math, not any inability to do statistics and understand it. And Field gives Zach—and readers—reassurance when the topic is especially difficult. For example, when Milton explains degrees of freedom, Zach responds, "That made no sense whatsoever." Milton answers, "Worry not: Nobody understands degrees of freedom." Presenting statistics instruction in a narrative format enables Field to create an emotional connection with readers that typical textbooks, and many teachers, do not.

As an experienced educator, Field has a good sense of where a student might get held up, and he makes sure to cover such topics repeatedly to emphasize certain points. But he maintains a teacher's sense of humor about students and their tendency not to listen well to their instructor. At one point, Zach gets confused about why a technique for repelling zombies doesn't work, even though data supporting the technique are available. He asks Milton, "Why would you have a model that fits well but doesn't turn out to be much use in the real world?" Field's description of Milton's reaction to this remark depicts teacherly exasperation: "Milton's face contorted into a strange mix of admiration and suicidal ideation. 'I spent a great deal of time telling you about sources of bias that can influence the linear model. Must you subject me to the utter tedium of explaining all of that again?'" Then Milton proceeds to give Zach a quick overview of the main points already made about bias. Field clearly wants to emphasize the importance of understanding bias in linear-model statistics, but he also seizes the opportunity to playfully tease those readers in need of a recap.

This failure to repel zombies is not the only occasion on which statistics obscure the truth. All the characters struggle to trust one another, and many discover others to be "lying" with statistics—through poor choice of analysis, failure of the data to conform to assumptions, misapprehension of the data's structure or outliers, or the creation of misleading data visualizations. In this way, Field teaches that statistics is a tool that can be used not just to solve problems and comprehend complex patterns, but also to deceive—or to confirm biases. Often this subject is not addressed so overtly in statistics classes, especially in cases in which it might court controversy or complicate homework assignments. The fictionalized data avoid these downsides while communicating important cautionary notes.

Milton ends up being Zach's de facto statistics teacher for most of the book. He is incredibly hard on Zach in an ironic, catty way, but when Zach loses confidence or when others attack his knowledge of statistics, Milton has his back. Most of the time, Milton displays a quirky and brusque sense of humor. For example, when a chimera threatens Zach as he fumbles trying to interpret some data, Milton bristles, "Look, lizard . . . . Three weeks ago this ape thought that kurtosis was a dental hygiene problem; all things considered, we are moving swiftly." Later, Milton even congratulates Zach for sticking with it, giving him one of the only straightforward compliments in the whole book: "You are the best student I've ever had. I have taught many brilliant scientists, but they are naturals . . . . You are different: You find this hard, people have told you that you can't do it, but . . . you've never given up."

Field's world-building and character development in the story animate the often contentious matter of attempting to separate objectivity from subjectivity, science from art, realism from relativism, logic from intuition, and rational thought from emotion. After all, Milton advises against dichotomizing continuous variables, saying it is "rarely sensible." Through depicting his characters' struggles, Field shows that both sides of each of these dichotomies are necessary for solving problems well—and that when the opposing sides are at odds, problems may not be solved well and can become more polarizing. Field makes this point most vividly when he has Sister Price, a druidic figure who represents a group called the Doctrine of Chance, explain the drawbacks of null-hypothesis significance testing (NHST): "The recipe-book nature of NHST encourages people to think in this all-or-nothing way. The dogmatic application of the 0.05 rule [for *p*-values] can mislead scientists." Indeed, this pitfall has led to the current debate among scientists over reproducibility and fishing for *p*-values below the threshold for significance.

Field drives this point home later in the explanation, making it clear that despite claims that this type of analysis is more objective and less biased, it is not necessarily so. He has Zach realize that "the scientist's intentions before data collection affect the actual value of *p*." From here, Field guides the narrative into a lesson on effect sizes and Bayesian statistics, a realm of analysis that is not explored in detail in all statistics textbooks. In the story, a cult has emerged around NHST, traditionally considered a gold-standard falsifiable test impervious to the effects of bias. The Doctrine of Chance—Sister Price's cult, which advocates for Bayesian statistics—arose in response. Critical of the traditional method's shortcomings, they argue that it indeed allows bias to enter by several possible avenues, including flawed experiment design, overestimating the importance of small effects deemed statistically significant, or by outright fishing for significance. The fictional narrative isn't too far off from the truth, given that these two camps in statistics have been at odds in the past (and occasionally still are). Field avoids the controversy by putting a humorous fictional spin on it, but he also makes it clear that any statistical technique is suspect when it is applied blindly or dogmatically.

The mythology that Field builds shows that he values the importance of art and emotion as a driver for one's use of statistics and desire to learn it. Indeed, Zach's tendency to follow his gut feelings comes in handy throughout the story. By showing how all the characters use statistics along with their other skill sets, Field humanizes statistics, depicting it as a tool wielded by people who may be good or bad, are certainly complex, and are not always in agreement about how they see the world.

The mythology that Field builds shows that he values the importance of art and emotion as a driver for one's use of statistics.

The fictional story exists in service of the statistics instruction, as the narrative flow is driven by wherever the statistics lessons need to go next. Although on its own the tale would not garner praise from literary critics, it succeeds in making a normally dry read into one that is fun, emotive, and even suspenseful. Field uses fiction to talk about contentious topics in science and statistics in entertaining and indirect ways, and he also uses the story to show that behind every statistical analysis is a plot with characters, each of whom has his or her own worldview, ethics, desires, and emotions. In this way, the book stands out as being especially instructive about the application and interpretation of statistics in the messy real world, in contrast to the many textbooks that show only the application of statistics in an idealized world. Sometimes fiction is the best vehicle for showing us our own reality, even in a field developed to separate facts from fictions.

Katie L. Burke is digital features editor of American Scientist. *She received her PhD in biology from the University of Virginia in 2011. She blogs about ecology at the* Understory.

[*This review was originally published in the March–April 2017 issue.*]

Information, Reimagined

A MIND AT PLAY: How Claude Shannon Invented the Information Age. Jimmy Soni and Rob Goodman. 366 pp. Simon and Schuster, 2017. $27.

A Mind at Play, Jimmy Soni and Rob Goodman's new biography of Claude Shannon, the mathematician considered to be the father of information theory, introduces us to its subject with an anecdote: After falling out of sight during the 1960s, Shannon made an unannounced appearance in 1985 at the International Information Theory Symposium in Brighton, England. The shy, white-haired celebrity was eventually spotted and soon afterward mobbed by autograph-seeking fans. Persuaded by the symposium chairman to come to the podium at the evening banquet, the reluctant Shannon had to endure hearing himself introduced as "one of the greatest scientific minds of our time." When the cheering and applause finally subsided, Shannon could only say, "This is—ridiculous!" He reached into his pocket, produced three balls, and began to juggle.

Photo courtesy of the Shannon family

Claude Shannon was an avid unicyclist who enjoyed coming up with eccentric designs to build, including one with an off-center hub that caused the rider to bob up and down while pedaling. Whether Shannon was redesigning data transmission or unicycles, the authors note that his work displayed a "mastery of model making: the reduction of big problems to their essential core." From *A Mind at Play*.

The chairman later described the bizarre scene: "It was as if Newton had showed up at a physics conference." Although hyperbolic (Newton, as far as we know, could not juggle), his comparison expresses an admiration for Shannon that has only grown stronger through the years. It reached dramatic height last year, the centennial of Shannon's birth. Celebratory conferences were held around the world. A Google doodle marked the day, April 30, 1916, when Claude Elwood Shannon was born in Petoskey, Michigan. One wonders what Shannon would have thought of all the fuss.

The fuss, however, is understandable: Shannon's landmark innovations—especially in laying theoretical groundwork for encoding messages for transmission and by determining how digital circuits could be designed—link him inextricably to today's information age. And in the wake of the centennial, Soni, a journalist, and Goodman, a writer and political scientist, have handily supplied curious readers with more of the modest mathematician's story.

Shannon's most productive years, those between 1940 and the mid-1950s, were spent in Manhattan at Bell Telephone Laboratories (which later moved to Murray Hill, New Jersey). During World War II he worked on a variety of projects involving electronics and cryptography. However, Shannon's enduring fame rests mainly on his landmark paper, "A Mathematical Theory of Communication," published in 1948 in the *Bell System Technical Journal* and republished by University of Illinois Press in 1963.

In the short paper Shannon considered the problem of transmitting digital data (that is, sequences of zeroes and ones) along a noisy channel. Many had believed that in order to increase the rate at which information can be transmitted, one should simply increase the power of the signal source. Building on earlier work by Harry Nyquist and Ralph Hartley, two colleagues at Bell Labs, Shannon showed that in fact there is a maximum rate of transmission over any channel. Assuming that the channel interference is caused by white noise, Shannon gave an easily computable formula for the maximum rate in terms of bandwidth and signal-to-noise ratio. Its calculated rate is a sharp maximum, meaning that it can be approached as closely as we desire, but it can never be exceeded.

Any transmission is vulnerable to error—random zeroes received as ones or vice versa. Shannon showed that if the transmission rate is less than the maximum, then there exist ways to send the data (by "coding" the transmission) so that the probability of error can be made arbitrarily small. The work of finding such codes, however, was left to others who took up the challenge. Today, data compression algorithms that rely on Shannon's theorems are used for an array of digital tasks, from recording music to sending pictures from Mars.

Shannon's appropriation of the term "entropy" inspired a productive debate about deep connections between information and thermodynamics. Mathematicians in probability and dynamical systems found that it could be extended and used effectively in their own work.

An abstract interpretation of the word *information* lies at the heart of Shannon's theory. Gone are semantic meanings. Any string of zeroes and ones satisfying a particular list of rules (for example, "zero cannot be followed immediately by zero") could be acceptable. English words can be communicated in this way by assigning different strings of zeroes and ones to individual letters (including an additional "letter" for a space).

As Shannon observed, our language has a certain amount of redundancy built in. For example, you can read this sentence, but, as Soni and Goodman relay, Shannon observed that "MST PPL HV LTTL DFFCLTY N RDNG THS SNTNC" as well—a condition familiar to anyone who sends text messages. Shannon gave a definition for the amount of information transmitted in a message. He then defined the rate of information transmitted, which he called *entropy*. For example, if we restrict ourselves to messages of zeroes and ones, then a source that can produce only ones would have zero entropy, whereas a source that produces zeroes and ones with the flip of a coin would have the largest possible entropy.

Soni and Goodman relate a famous story about Shannon's choice of the word "entropy." The mathematician John von Neumann noted the uncanny similarity between Shannon's notion and one that had been used in thermodynamics for decades. "You should call it entropy, for two reasons," von Neumann advised. "In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage."

"Almost certainly, this conversation never happened," insist the authors, echoing doubts raised elsewhere. However, Shannon himself related the story, exactly as above, in a 1971 interview with the engineer Myron Tribus. Regardless of whether the story is true, Shannon's appropriation of the term "entropy" inspired a productive debate about deep connections between information and thermodynamics. Mathematicians who work in the areas of probability and dynamical systems then heard about Shannon's definition and found that it could be extended and used effectively in their own work. It is not difficult to imagine that von Neumann, one of the greatest mathematical minds of the 20th century, anticipated some of these later developments.

It is a temptation to look back at the early life of a genius and search for signs that promise future greatness. In the case of Claude Shannon, however, we find few indicators. We read that Shannon won a third-grade Thanksgiving story-writing contest and that he played alto horn in school musicals. He loved to build and fix things, especially radios, as did many youngsters in the 1920s. He found mathematics easy and enjoyed its competitive aspects, but no evidence is offered of exceptional mathematical ability.

What little is revealed about Shannon's college career also fails to predict eminence. He attended the University of Michigan, where he earned dual bachelor degrees in mathematics and electrical engineering. He was elected to the Phi Kappa Phi and Sigma Xi honor societies. He published two solutions to questions proposed in the *American Mathematical Monthly*, an expository journal intended for both students and faculty. These accomplishments are laudable but certainly not rare. Unfortunately, we don't learn who Shannon's teachers were or what mathematics and science courses he took at the University of Michigan. Such information might have helped the reader anticipate the first blaze of Shannon's genius, his master's thesis, completed in 1937.

Serendipity is a standard ingredient of notable careers, and for Shannon it was added during his master's program, in the spring of 1936, when he noticed a typed card stuck to a bulletin board. It advertised a graduate assistantship at the Massachusetts Institute of Technology with the duty of running a *differential analyzer*, a mechanical computer designed to solve differential and integral equations. Such analog computers had been around since 1876, but this one also had some digital components and was the first capable of general applications. Eventually it would solve differential equations with 18 independent variables. Its inventors were Harold Hazen and Vannevar Bush. "I pushed hard for that job and got it," Shannon recalled. "That was one of the luckiest things of my life."

Vannevar Bush was a tall figure in American science. He had joined MIT's electrical engineering department in 1919, and three years later, he founded a military supplier now called the Raytheon Company. In 1941, Bush would help convince President Roosevelt to begin building an atomic bomb, and he would take a leading role in its development. At MIT Bush recognized Shannon's brilliance and took a serious interest in his career, guiding him through graduate school and on to Bell Labs.

Shannon's thesis, *A Symbolic Analysis of Relay and Switching Circuits*, is regarded as one of the most important master's theses ever written. Completed in 1937, it used century-old ideas of the British logician George Boole to simplify the arrangement of relays comprising electrical networks. Elegant and practical, Shannon's system provided a basis for modern digital circuit design. Most mathematicians who teach applications of Boolean algebra to electrical circuits in courses of discrete mathematics do not realize they are presenting the ideas in Shannon's thesis. More than 50 years later, Shannon downplayed the significance of his discovery. "It just happened no one else was familiar with both fields at the same time," he told an interviewer, adding, "I've always loved that word. ‘Boolean.’"

Bush was not only a good judge of intellect, he was also a shrewd observer of temperament. He might have been worried about his new protégé. Shannon had lost his father in his sophomore year, and for some reason stopped speaking to his mother shortly afterward. Bush encouraged Shannon to spend time at Cold Springs Harbor Laboratory and apply Boolean algebra to Mendelian genetics. He would be supervised by Barbara Stoddard Burks, a sympathetic psychologist interested in the genetics of genius and keenly interested in questions of nature versus nurture. The Genetics Records Office at Cold Springs Harbor had more than 25 years of data for Shannon to contemplate. In less than one year, Shannon had learned enough of genetics to complete his Ph.D. dissertation, *An Algebra for Theoretical Genetics*, a masterful but overly theoretical work that would have little to offer geneticists. The experience confirmed the opinion that Bush and Burks shared: Shannon was a genius who could acquire knowledge of a new subject quickly and from it create significant mathematics. However, Shannon had little regard for the work. He fled the field and never bothered to publish his dissertation. Some years later he remarked, "I had a good time acting as a geneticist for a couple of years."

> A prolific tinkerer with a singular sense of humor, Shannon invented bizarre devices, including a calculator that operates with Roman numerals.

After receiving his doctorate Shannon spent a summer at Bell Labs and a year at the Institute for Advanced Study in Princeton, and he finally found full-time employment back at Bell Labs.

Shannon was fortunate to work at Bell Labs during a period when research and development in the United States was generously funded. His brilliance entitled him to a freedom that seems impossible today, in a time of international competition and demands by shareholders for fast profits. With characteristic modesty, Shannon once admitted to a supervisor, "It always seemed to me that the freedom I took [at the Labs] was something of a special favor."

Lured by a change of scene and the relative security of academia, Shannon accepted a position at MIT in 1958. He retired in 1978. The good luck that had followed him for so long finally departed in the early 1980s as Shannon began displaying signs of Alzheimer's disease. He died from the illness in 2001.

A Mind at Play is a loving biography recounted by two admirers of Claude Shannon. It is especially good at relating the many stories that have contributed to the growing fascination with its hero. A prolific tinkerer with a singular sense of humor, Shannon invented bizarre and amusing devices, many of which are described. They included a motorized pogo stick, a rocket-powered frisbee disk, a juggling machine, a calculator that operates with Roman numerals, and a relay-controlled robotic mouse that could solve a maze and keep track of its solution. An invention of Shannon's that became known as the "Ultimate Machine" fascinated science-fiction writer Arthur C. Clarke during a visit to Bell Labs. In his 1958 book *Voice across the Sea*, Clarke offered a description of the machine's workings: "When you throw the switch, there is an angry, purposeful buzzing. The lid slowly rises, and from beneath it emerges a hand. The hand reaches down, turns the switch off, and retreats into the box. With the finality of a closing coffin, the lid snaps shut, the buzzing ceases, and peace reigns once more."

A Mind at Play is somewhat less successful when mathematics appears. For example, both the conclusion of Shannon's "Theorem on Color Coding" and Hartley's formula for information are misstated. The authors do an admirable job of describing Shannon's entropy for a coin toss, but they stop short of explaining it for a more general information source. Readers wishing to learn details of Shannon's work would do better to go to Shannon's papers, which are well written and freely available online.

More distressing than minor technical slips is the authors' discussion of the criticism that followed publication of *The Mathematical Theory of Communication*. After citing a sharp comment by probabilist Joseph Doob in a review, the authors imagine that pure mathematicians formed a cabal to condemn Shannon's applied work. Certainly Shannon's definitions and proofs were not always complete and correct. (For example, Shannon's theorem about the optimum use of noisy channels by coding, discussed previously, was finally proved by Amiel Feinstein in 1954, and today it is known as the Shannon-Feinstein theorem.) Nevertheless, Shannon's work was and continues to be used and admired by the mathematical community. *Mathematical Reviews*, in which Doob published his odd remark, contains nearly 2,000 reviews that refer to Shannon's entropy.

Shannon did more than open up the new field of information theory. He also demonstrated what can be accomplished by combining passionate inquiry with a fondness for levity. *A Mind at Play* is an enjoyable biography that unites us with the singular spirit of Claude Shannon.

Daniel S. Silver is an emeritus professor of mathematics at the University of South Alabama. His research explores the relation between knots and dynamical systems, as well as the history of science and the psychology of invention.

[*This review was originally published in the January–February 2018 issue.*]

Math with Attitude

MATH WITH BAD DRAWINGS: Illuminating the Ideas That Shape Our Reality. Ben Orlin. 367 pp. Black Dog and Leventhal, 2018. $27.99.

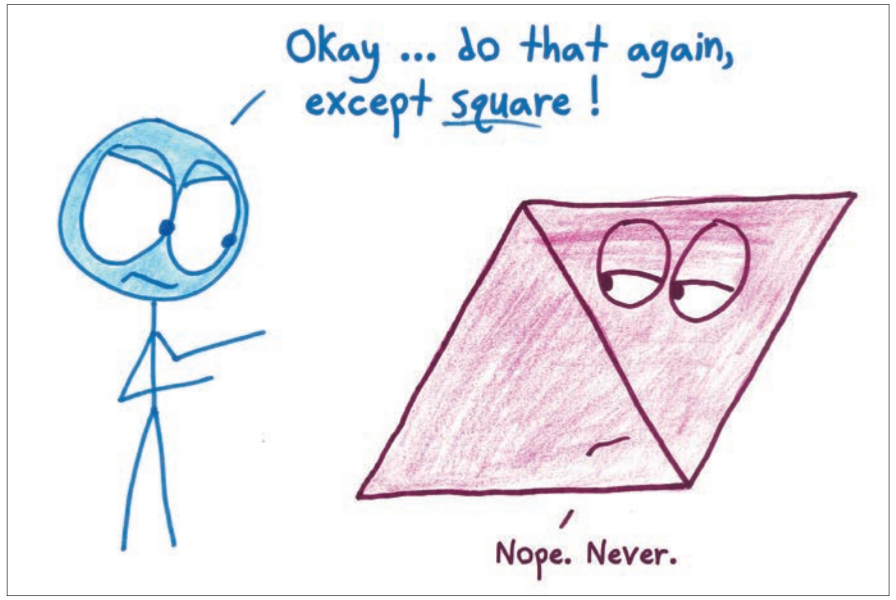
Math books meant for a broad audience are often tinged with evangelical fervor. They yearn to reinspire all those millions who lost their faith in numbers somewhere between flash-card drills and the quadratic formula. Real mathematics isn't like that, the books assure us. Real mathematics is filled with exciting adventures: turning a sphere inside out without piercing the surface, tiling an infinite bathroom with a pattern that never repeats, drawing curves so squiggly they fill all of space, strolling around a Möbius band and returning as your own mirror image.

I have read and thoroughly enjoyed many books in this genre, and I've even written a couple of them myself. However, I've never really believed they are likely to convert anyone who's not already singing in the mathematical choir. The sad fact is, outside the circle of math enthusiasts, people aren't all that interested in sphere eversion and aperiodic tiling.

Ben Orlin's *Math with Bad Drawings* may have a better chance of reaching lost souls. Orlin has an advantage over ivory tower types like me. As a K–12 classroom teacher, he comes face-to-face with skeptical youth every day. When he asked a group of ninth graders why they study math, they settled on the answer, "to prove to colleges and employers that we are smart and hardworking." Orlin comments:

> The students weren't wrong. Education has a competitive zero-sum aspect, in which math functions as a sorting mechanism. What they were missing—what I was failing to show them—was math's deeper function.

"Deeper function" is a revealing phrase. If I were writing that sentence, I might have said "math's deeper meaning" or perhaps "math's inner beauty." But Orlin is listening to his students, and they are telling him, "Keep your feet on the ground." In these pages there are no mind-boggling excursions into *N*-dimensional



"Say you're creating a square, and you want its diagonal to be the same length as its sides." Ben Orlin proposes this geometric impossibility, only to have it vetoed by the personification of a rhombus in one of his "bad drawings." From *Math with Bad Drawings*.

geometry or puzzles about self-referential sentences that are true only if they're false. In this book, mathematics is a down-to-earth tool for describing and understanding the world, not an art form or a quest for esoteric truths. Orlin applies this tool to the activities of everyday life: rolling the dice, paying your taxes, rescuing the global economy from daredevil bankers, fixing the Electoral College, designing a Death Star for Galactic Emperor Palpatine. (I'll concede that Death Star engineering is not an everyday task for most of us, but even a math teacher deserves a little fun every now and then.)

The volume is organized in five parts. Part I is a brief introduction exploring what mathematics looks like to students and teachers as well as to mathematicians. Part II takes up geometry and design, praising the virtues of the triangle as a structural element and bravely taking on the contentious issue of A4 versus U.S. letter-size stationery. Part II is also where the Death Star turns up. Parts III and IV, comprising almost half the book, deal with probability and statistics: lotteries, baseball box scores, *p*-hacking in the sciences, and the curious practice of putting world literature through a statistical meat grinder. Part V turns to some economic and political themes.

As presented in *Math with Bad Drawings*, these topics require no mathematical knowledge or skills beyond the ken of a ninth grader—elementary arithmetic, some basic concepts in probability, enough geometry to recognize a right triangle. It's ordinary schoolroom math—just the sort of thing that has bored and alienated generations of students. And yet Orlin spins it into a charming book you'll want to take to the beach, or at least keep handy by the commode.

What's his secret? Well, first of all, there are the bad drawings—although in truth they're not half bad. Not even quarter bad. Or maybe I'm just unusually susceptible to stick figures with oversize bubbleheads, whose eyes communicate a surprising gamut of human emotions. The expressive eyes sometimes migrate to other objects—polygons, coffee cups, gemstones, maps of Minnesota—where they are just as endearing. I want them in all my math books from now on, please.

The prose is also chipper and cheerful. I'll content myself with a single example, which happens to address one of the main messages of both the text and the bad drawings:

> Fables and math have a lot in common. Both come from dusty, moth-eaten books. Both are inflicted upon children. And both seek to explain the world through radical acts of simplification.
>
> If you want to reckon with the full idiosyncrasy and complexity of life, look elsewhere. Ask a biologist, or a painter of photorealistic landscapes, or someone who files their own taxes. Fable tellers and math makers are more like cartoonists. By exaggerating a few features and neglecting all the rest, they help explain why our world is the way it is.

Orlin has a third secret ingredient, but it's invisible; it's something that's been deliberately left out of the recipe. "Do the math" and "show your work" are phrases that never turn up in these pages. There are no homework problems, no exercises for the reader, not even worked examples. The focus is on concepts, not algorithms or formulas or equations. Orlin occasionally gives the result of a numerical calculation, but he doesn't dwell on where the answer came from or explain how one might tackle similar problems. This mode of discourse would not be at all unusual in a work of history or literary criticism, but it's a radical departure in mathematics, where learning by doing is a way of life, and problem-solving is both a pastime and a rite of passage.

I was a few chapters into the book before I became fully conscious of this curious absence. My first reaction was: "No! Wait! You can't do that. You can't write a math book with no math in it." But why not? Authors in other disciplines are under no such compulsion. A book on music doesn't have to teach you how to play the guitar or compose a string quartet. Likewise not all books about food are full of recipes. Why should reading mathematics always be a roll-up-your-sleeves participatory process? As Orlin demonstrates, it's entirely possible to say interesting things about mathematics without showing people how to do mathematics. And this more discursive approach may help bring the gospel to an audience that would be turned away by scary notation.

If I haven't quite convinced you of the wisdom of mathless math writing, that's because I haven't quite convinced myself either. After all, mathematical notation has a purpose: It clearly expresses ideas that would be hard to communicate without it. Consider a passage in the introduction to the section on probability. After noting that the outcome of a single coin toss is 50/50, Orlin writes:

> But if you could flip a trillion coins, you'd find yourself approaching a different world altogether: a well-groomed land of long-term averages. Here, half of all coins land on heads . . . and one-in-a-million events happen a millionth of the time, give or take.

These statements convey a deep truth: that random events in large enough numbers converge toward their average or expected behavior. Nevertheless, I worry that some readers will come away with a mistaken intuition about that experiment with a trillion coins. In particular, what is the probability of seeing exactly equal numbers of heads and tails? The phrase "half of all coins land on heads" might be taken to imply that the probability of this outcome grows larger as the number of coins increases, and that heads = tails would become a certainty with infinitely many coins. In fact, the probability of observing equally many heads and tails in a trillion coin flips is less than one in a million, and as the number of flips goes to infinity, the probability of an equal division wilts away to zero.

My point is not that Orlin's statement about long-term averages is incorrect; at worst it's slightly imprecise or incomplete. My point is that full understanding of a mathematical fact is hard to attain without doing some mathematics. Stands to reason, no? But then I see one of Orlin's sleepy-eyed stick figures demanding, "Okay. Show me the math." I'll give it a try.

Allow me to start with an easier problem: the probability of getting equal numbers of heads and tails when flipping 100 coins rather than a trillion. The number of possible head-tail sequences in 100 coin flips is 2^{100}. How many of those sequences have exactly 50 heads and 50 tails? The answer is $100!/(50!\times 50!)$, where the exclamation point denotes the factorial function: $100! = 100\times 99\times 98\times \ldots \times 3\times 2\times 1$. Stacking up all those multiplications produces some very large numbers, but with computer assistance it's not hard to calculate them. The probability we're seeking is the number of 50-head sequences divided by the total number of sequences; it's about 0.08.

To be thorough I would have to explain where those formulas came from and why you should believe they give the right answer, but I'm not going to bother, because the formulas are useless for the full-scale computation anyway. The number of possible outcomes when you flip a trillion coins is 2 raised to the trillionth power, which is a number with too many bits to fit in my computer's memory. To complete the computation I must resort to shortcuts or stratagems, such as working with logarithms of factorials. With some algebraic hocus-pocus, the formula for the probability of equal heads and tails can be reduced to a remarkably simple approximation: $1/\sqrt{\pi n/2}$, where n is the number of coins being flipped. For $n = 1$ trillion, this works out to 8×10^{-7}. The challenge, of course, is explaining the hocus-pocus. Perhaps I could do so in terms the stick figure would understand, but it would take at least a few paragraphs, and I'm sure those droopy eyes would close before I could finish.

Euclid supposedly declared, "There is no royal road to geometry." He was scolding an overprivileged pupil who was tired of ruts and potholes and wanted a well-paved route to the summit of knowledge. Orlin hasn't built the royal road, but he's offering aerial tours of the mountainside that are well worth taking. The details may be hard to discern from this altitude, but the scenery is great! I look forward to the sequel, although I am disappointed to learn it will not be titled *More Math with Worse Drawings.*

Brian Hayes is a former editor and columnist for American Scientist. *His most recent book is* Foolproof, and Other Mathematical Meditations.

[*This review was originally published in the July–August 2019 issue.*]