



Bronze, Granite, Clay. *“Figure Eight Knot Complement vii/CMI”* is mathematician–sculptor Helaman Ferguson’s icon for the Clay Mathematics Institute. The large piece (bottom), done in Inner Mongolian Black Granite, is located at the CMI in Cambridge, Massachusetts. Smaller bronzer versions (top) are given as the annual Clay Research Award. (Photos courtesy of Helaman Ferguson.)

Think and Grow Rich

It's a common question encountered by students of mathematics: If you're so smart, why aren't you rich? In fact, math majors usually *do* place near the top of the list of entry-level salaries for college graduates. But, as a quick route to easy street, immersing oneself in abstract algebra, complex functions, and topology is not the way to go.

Or is it?

The Clay Mathematics Institute, a non-profit foundation based in Cambridge, Massachusetts, has offered a million dollars each for solutions to seven open problems in mathematics. Announced in May 2000, the seven “Millennium Prize Problems,” as the CMI calls them, are widely considered among the most important—and most difficult—unsolved problems today. Two problems come from number theory, two from topology, two from mathematical physics, and one from theoretical computer science.

The oldest Millennium Prize problem is a number-theoretic challenge known as the Riemann Hypothesis, which dates to 1859. It concerns a property of what's called the Riemann zeta function, customarily defined by the formula

$$\zeta(s) = 1 + 1/2^s = 1/3^s + 1/4^s + \dots ,$$

where s is a complex number. In a famous paper published in 1859, the German mathematician Bernhard Riemann used the zeta function to study the distribution of prime numbers.

Riemann's starting point was the observation, first made in 1737 by Leonhard Euler, that the zeta function can also be written as a *product* involving all primes:

$$\zeta(s) = 1/(1 - 2^{-s})(1 - 3^{-s})(1 - 5^{-s})\dots .$$

By analyzing the zeta function, Riemann obtained amazingly precise formulas for the distribution of primes. His analysis, which was based on the then-developing theory of complex variables, was later made rigorous, and led to a complete proof of the Prime



Landon T. Clay. (Photo courtesy of the Clay Mathematics Institute, Inc.)

The Riemann Hypothesis has baffled mathematicians for the last century and a half.

Number Theorem: The number of primes less than x is approximately $x/\ln x$ (where $\ln x$ is the natural logarithm of x).

The formulas relating the zeta function to the distribution of primes are stated in terms of the “zeroes” of the zeta function: the values s for which $\zeta(s) = 0$. The zeroes are of two types: the “trivial” zeroes at the negative even integers ($s = -2, -4, \text{etc.}$), and the “non-trivial” zeroes, all of which have real parts between 0 and 1. In essence, the non-trivial zeroes of the zeta function contain detailed information about the error term in the approximation $x/\ln x$. (The approximation that’s actually used is a function known as the logarithmic integral.) Riemann thought it “very likely” that the real parts of these zeroes are all equal to $1/2$ —that is, that the non-trivial zeroes of the zeta function all lie on a line. “It would, of course, be desirable to have a rigorous proof,” he added.

Indeed it would. The Riemann Hypothesis, as the conjecture came to be called, has baffled mathematicians for the last century and a half. If true—and no one seriously doubts that it is—the Riemann Hypothesis has profound implications for number theory. And beyond mathematics as well: In the last decade, mathematical physicists have found echoes of the Riemann Hypothesis in the theory of quantum chaos (see “A Prime Case of Chaos,” *What’s Happening in the Mathematical Sciences*, Volume 4).

The other number-theoretic prize problem is known as the Birch–Swinnerton-Dyer conjecture. In the 1960s, British mathematicians Brian Birch and Peter Swinnerton-Dyer made an extensive computational study of the L-functions of elliptic curves (see “New Heights for Number Theory,” page 2). These functions are closely related to the Riemann zeta function. In every example they studied, Birch and Swinnerton-Dyer found that the value of $L_E(s)$ at $s = 1$ predicts whether the elliptic curve E has finitely many or infinitely many rational solutions. Their conjecture is that E has infinitely many rational solutions if and only if $L_E(1) = 0$.

In fact, the conjecture says something more precise. Even when an elliptic curve has infinitely many rational solutions, they are all generated from a finite set. That is, the group of rational solutions has an abstract structure of the form $T + Z^r$, where T denotes a finite “torsion” group, Z denotes the integers (the German word for number is *Zahl*), and r is a non-negative integer called the “rank” of the curve. (This in itself is a deep theorem. It was conjectured by Henri Poincaré in 1901, and proved by L.J. Mordell in 1922.)

The Birch–Swinnerton-Dyer conjecture says that E has rank r if and only if $L_E(s)/(s-1)^r$ has a non-zero limit as s tends to 1.

The conjecture goes on to relate the value of this limit to various properties of the elliptic curve, but that’s not required for the Millennium Prize. Some progress has been made in the last quarter century. In 1977, John Coates and Andrew Wiles at the University of Cambridge proved that, for a special class of elliptic curves with a property known as complex multiplication, if $L_E(s)$ has a nonzero limit at $s = 1$, then the curve’s rank is 0, as the conjecture predicts. In 1983, Benedict Gross at Brown University and Don Zagier at the University of Maryland showed that, for “modular” elliptic curves—that is, for all elliptic curves with an associated modular form—if $L_E(s)/(s-1)$ has a nonzero limit at $s = 1$, then the rank is at least 1. In 1990, Victor Kolyvagin at the Steklov Institute in Moscow strengthened these results. He showed that the Coates–Wiles theorem holds for modular elliptic curves (a class that was already known to include curves with complex multiplication). Kolyvagin also showed that in the Gross–Zagier theorem, the rank is *exactly* 1.

Now that all elliptic curves are known to be modular—thanks to the Taniyama–Shimura conjecture having been proved—Kolyvagin’s result holds across the board. But it falls short of the Birch–Swinnerton-Dyer conjecture, because it’s silent about curves for which $L_E(s)/(s-1)^r$ is nonzero at $s = 1$ for r greater than 1. Conceivably, some elliptic curve lurks in the back alleys of number theory with $L_E(s)/(s-1)^2$ nonzero at $s = 1$, but with rank something other than 2—perhaps even 1 or 0.

The more famous of the two topology problems is known as the Poincaré conjecture. Loosely speaking, it asserts that a topologically complicated three-dimensional space (or “manifold”) cannot masquerade as something simple. Stated technically, the conjecture asserts that a compact 3-d manifold has a trivial fundamental group if and only if the manifold is homeomorphic to the 3-sphere. A similar result is known to hold for 2-d manifolds.

“Homeomorphic” is a fancy way of saying that two things are essentially the same. In the rubber-sheet metaphor of topology, it means that a balloon is a balloon is a balloon—until you puncture, rip, or tear it. A manifold

Even when an elliptic curve has infinitely many rational solutions, they are all generated from a finite set.



Arthur Jaffe. *CMI President.* (Photo courtesy of Bachrach Photographers.)

is something that, locally at least, looks like ordinary euclidean space, be it the 2-d plane, 3-d space, or something of higher dimension. The easiest manifolds to picture are two-dimensional; these are simply curved surfaces. The “compact” ones, which include the sphere, the torus, and a host of more complicated, pretzel-like surfaces (see Figure 1), are all bounded.

The fundamental group is an algebraic structure associated with closed curves in a manifold. In the 2-d case, picture a loop of wire that is confined to the surface, but may otherwise be moved around, stretched or shrunk. For the 2-d sphere, it’s easy to see—though not quite so easy to prove—that any such loop, no matter how complicated, can be shrunk down to a single point. For the torus, by contrast, there are loops that cannot be shrunk to a point, namely the ones that go all the way around, or through, the “hole” in the torus. The upshot is that the fundamental group of the 2-sphere is “trivial,” whereas the fundamental group for the torus (as well as all the other, more complicated surfaces) is non-trivial.

Much the same happens with 3-d manifolds. The fundamental group of the 3-sphere is trivial, but the fundamental group of the 3-d torus is not. In 1904, Henri Poincaré conjectured that all other

3-d manifolds also have non-trivial fundamental groups. This is certainly true for all *known* 3-d manifolds. Unfortunately, the classification of manifolds, which is extremely simple in two dimensions, is itself still conjectural in 3-d. Indeed, the best way to prove the Poincaré conjecture—“best” because it would prove so much more—would be to prove the Thurston Geometrization Conjecture (see “Topology Makes the World Go Round,” page 38).

Curiously, higher-dimensional analogues of the Poincaré conjecture have already been proved, even though the classification theory in higher dimensions is extremely primitive. The analogues say that an n -dimensional manifold has trivial

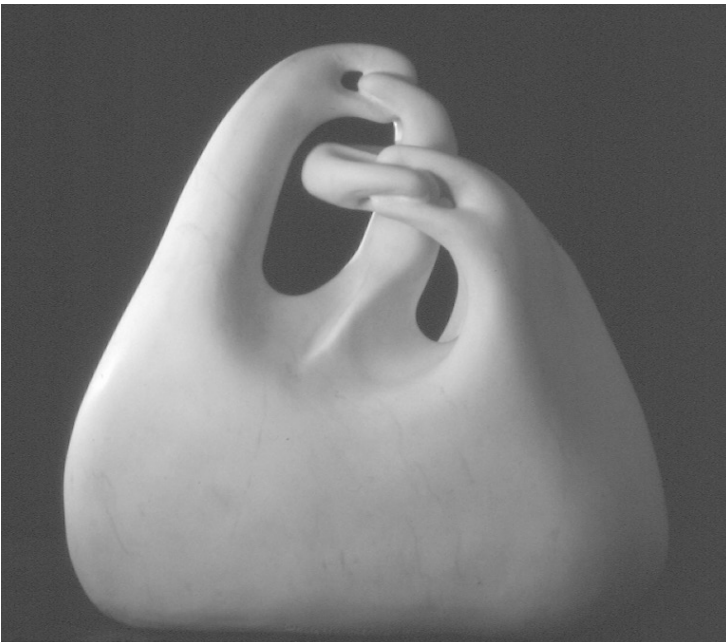


Figure 1. In *Whaledream II*, mathematician–sculptor Helaman Ferguson portrays the quiet complexity of topological space. (Photo courtesy of Helaman Ferguson.)

Sphere Basics

What is the 3-sphere? It's best approached through lower dimensions. The 1-sphere is simply the ordinary circle. The theory of compact 1-d manifolds is extremely simple: The 1-sphere is the only one. (Incidentally, the fundamental group of the 1-sphere is not trivial: The circle *itself* is a loop, and there's no way you can contract it to a point.) The 2-sphere is the sphere of ordinary speech. Like the circle, it can be defined algebraically, by the equation $x^2 + y^2 + z^2 = 1$. Topologists call this an "embedding" of the 2-sphere in R^3 . A nice property of 2-d manifolds is that every (orientable) 2-d manifold can be embedded R^3 .

The 3-sphere is most easily defined algebraically, as the solution set in R^4 of the equation $x^2 + y^2 + z^2 + w^2 = 1$. Unfortunately, objects in 4-dimensional space are hard to picture. However, there's another way to understand the 3-sphere that leads to a fairly good grasp of it. We start again with the 1- and 2-spheres.

Another view of the circle is as the entire real line with one extra point added to close the loop. The correspondence comes courtesy of a simple map called the stereographic projection (see Figure 2). Similarly, the 2-sphere is the entire infinite plane with one extra point (sometimes called the north pole). In other words, the 1-sphere is much like a line, and the 2-sphere is much like a plane—with the one difference that they've been "compactified" by the addition of one extra "point at infinity." Again, the 3-sphere is similar: It's basically just euclidean 3-d space with one slightly mysterious extra point glued on.

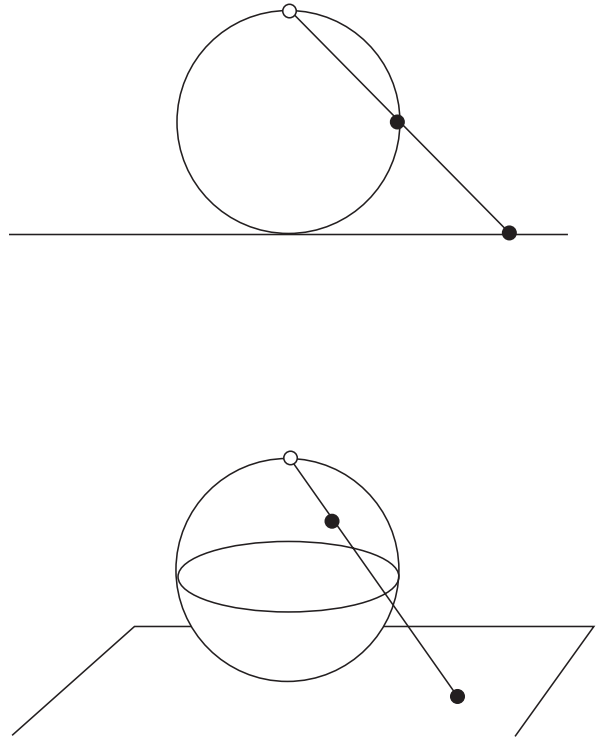


Figure 2. Stereographic projections equate the circle and the sphere with the line and the plane, respectively, each with an extra "point at infinity."

fundamental group if and only if it's homeomorphic to the n -sphere. For dimensions 5 and higher, this was first proved in 1960 by Stephen Smale at the University of California at Berkeley. The 4-d case resisted until 1982, when Michael Freedman at the University of California at San Diego knocked it off.

One reason the Poincaré conjecture is so hard in 3-d has to do

The Hodge conjecture concerns the analysis of high-dimensional manifolds defined by algebraic equations.

with the nature of the surface that's swept out when a loop is contracted to a point. The surface is topologically a disk—that is, it looks like the interior of a circle—but it can wind up embedded with all kinds of kinks and self-intersections. This can't happen for loops on the 2-sphere. But loops in 3-d are often knotted, and that makes for a mess. Even in 4-d it may be impossible to avoid self-intersections when contracting to a point. But in 5-d and higher, there's enough "room" to maneuver, so that the disk can always be swept out in a nice way.

The other million-dollar topology problem is the Hodge conjecture, named for William Hodge, who first proposed it in 1950. The Hodge conjecture concerns the analysis of high-dimensional manifolds defined by systems of algebraic equations. It says, very roughly, that everything you always wanted to know about algebraically defined manifolds (but were afraid to ask) is to be found in the theory of calculus.

An algebraically defined manifold is a space that's defined by a system of polynomial equations. As a rule, one equation in n variables defines a "hypersurface" of dimension $n - 1$. For example, the equation $x^2 + y^2 + z^2 = 1$ in 3 variables defines the 2-dimensional sphere. For a system of, say, k equations, the intersection of the corresponding hypersurfaces is, in general, of dimension $n - k$.

The high-dimensional analogues are impossible to picture, especially if you let each variable be a complex number, but the basic mathematical principles carry over from lower-dimensional settings. In particular, it's possible to do calculus on these manifolds. The formalism goes by the name "cohomology" theory. At the heart of this theory is a vector space of differential forms known as the Hodge space. (In calculus, a differential $f dx$ is "what gets integrated" in the integral $\int_a^b f(x)dx$. Differential forms are higher-dimensional generalizations.) The Hodge conjecture asserts that, for each algebraically defined manifold, this vector space is spanned by differential forms associated with algebraically defined submanifolds.

The precise statement of the conjecture is highly technical and the concepts involved are extremely subtle. In fact, Hodge himself first stated the conjecture incorrectly! (He also proposed a generalization that turned out to be incorrect. The French mathematician Alexandre Grothendieck corrected the latter in a 1969 paper with an eye-catching title: "Hodge's general conjecture is false for

trivial reasons.”) Such missteps are common when fields are first developing. It takes time for researchers to determine which analogies are the most appropriate.

By now mathematicians understand the theory well enough to be confident the Hodge conjecture is properly stated, and they expect that it’s correct. Still, the conjecture could be false for “trivial” reasons that have escaped notice for the last 50 years. The CMI is willing to award a prize for a counterexample to any of its problems if the counterexample radically alters the nature of the theory, but reserves the right to reformulate the problem if, as happened with Grothendieck, the counterexample reveals only a minor flaw. (Due to the substantial sums involved, the CMI has written its rules very carefully.)

The two Millennium problems in mathematical physics are concerned with our mathematical understanding of two familiar features of the real world: the fact that water (or any other fluid, including the air we breathe) flows in a smooth (if sometimes turbulent) manner, and the fact that matter has mass—i.e., the fact that there’s something for force to accelerate!

The first problem refers to the existence and smoothness of solutions to the Navier–Stokes equation (which actually consists of two equations). The Navier–Stokes equation essentially applies Newton’s famous formula (force equals mass times acceleration) to fluids: It describes the way that fluid flow changes in space and time.

The equation is extremely difficult to solve in general. Exact solutions are possible in only a few settings. (For example, if the fluid starts out perfectly still, then, according to the Navier–Stokes equation, it remains so for all time. But as solutions go, that one’s about as dull as ditchwater!) For the most part, fluid-flow scientists, which include a range of researchers from pure mathematics to aerodynamics, are content to use numerical algorithms that produce approximate solutions. But they would love to know more about the exact solutions. As a first step, they would simply like to know that exact solutions really do exist, no matter what the initial conditions.

And there’s the hang-up: It’s *not* known that the Navier–Stokes equation always has a solution. It’s conceivable that a fluid could begin flowing in such a way that it eventually develops a “singularity”—a point or points where the flow is no longer continuous,

It’s not known that the Navier–Stokes equation always has a solution.

The 3-d obstacle is turbulence.

such as a vortex around which the fluid spins with greater and greater velocity the closer in you look. The Millennium problem is to prove that this can't happen: If the initial conditions are smooth, then the flow remains smooth for all time.

In fact, the Millennium problem could go either way: You can win the prize by finding a smooth set of initial conditions for which the flow *doesn't* stay smooth. In two dimensions, the analogous problem was solved in the 1960's, by Olga Ladyzhenskaya at the Steklov Institute in St. Petersburg, Russia, who showed that smooth conditions necessarily spawn smooth flows. And partial results for the 3-d case are known: All smooth flows stay smooth for at least a short amount of time, and if a flow is slow enough to start with, then it stays smooth forever.

The 3-d obstacle is turbulence. To keep track of a turbulent fluid, you need to make finer and finer observations of it: The fluid continually generates features at smaller and smaller scales. Turbulence does not exist in two dimensions, but it plagues the 3-d theory. Very roughly speaking, the question of existence and smoothness for the Navier–Stokes equation amounts to asking how quickly turbulent flow reaches these smaller and smaller scales. If, say, you have to double the magnification once per second, then the flow will always be seen as smooth—after n seconds, you just have to look 2^n times closer. But if you have double the magnification first after one second, again just a half second later, yet again just a quarter second later, and so forth, then you're going to have a mess in two seconds.

While the Navier–Stokes equation concerns classical mechanics, the other physics problem on the CMI's hit list comes from quantum theory. The problem is to show that the framework mathematical physicists use to study quantum fields—an approach known as Yang–Mills theory—really does describe the subatomic world of quarks, gluons, and the rest of the particle zoo. In particular, the theory should predict a “mass gap” in the energy spectrum. Loosely speaking, this means that in a theory of, say, quarks, the energy of empty space is 0, but as soon as even one particle appears, the energy is at least some minimum value, E . Assuming this, you can assign the particle the mass m via Einstein's formula $E = mc^2$.

The current version of Yang–Mills theory, known as quantum chromodynamics (QCD), almost certainly predicts a mass gap,

along with other desirable properties such as asymptotic freedom and quark confinement. (The former means that at shorter and shorter distances, the quantum behavior of the field is better and better approximated by the classical theory; the latter explains why we live in a sea of protons and pions and such, rather than an ocean of quarks.) Computational studies of QCD indicate the theory works, and experiments to date confirm its predictions. But rigorous proof is lacking.

The official problem goes beyond chromodynamics. QCD is only one Yang–Mills theory, based on a “gauge” group of symmetries known as $SU(3)$. (Roughly speaking, $SU(3)$ is the group of rotations in 3-d complex space.) The Millennium problem is to establish a Yang–Mills theory for *all* gauge groups, and to show that each one has a mass gap. A tall order indeed, but theoretical physicists consider it essential to a proper understanding of quantum fields.

The final Millennium problem, called the “P versus NP” problem, goes to the heart of computer science. At issue is whether many problems that are now out of reach computationally will always remain so. In other words, are there computational problems that computers will *never* handle?

Technically, both P and NP are classes of “decision” problems, which seek a simple Yes/No answer. For example: Does 1,002,011 have any divisors less than 1000? Many optimization problems, such as finding the shortest route in the travelling salesman problem, can be reduced to a set of such Yes/No questions.

For problems in class P, the work required to solve a representative problem grows no faster than a *polynomial* in the “size” of the problem. To decide, for example, whether two N -digit numbers have a common factor requires a computation—the Euclidean algorithm—that grows like N^2 . Such problems are relatively “easy,” because the extra time needed to solve a slightly bigger instance is a tiny fraction of the time spent on the smaller version. If it takes, say, 1 second to find the greatest common factor of two 1000-digit numbers, then it should take only 2 more milli-seconds to handle two 1001-digit numbers.

The class NP—the moniker means *non-deterministic polynomial time*—is trickier to describe. Loosely speaking, NP consists of problems for which proposed solutions are easy to *check*. Factorability is a good example: The answer to the question about

The “P versus NP” problem goes to the heart of computer science. At issue is whether many problems that are now out of reach computationally will always remain so.

1,002,011 is Yes, as you can easily check on a pocket calculator by dividing it by 733. “Easy” means that the proposed solution can be verified with a computation that grows polynomially in the size of the problem. But the proposed solution can come from anywhere: an exhaustive search, a lucky guess, or some sort of insider knowledge. That’s what makes them non-deterministic.

More precisely, NP consists of Yes/No problems that are easy to check *when the answer is Yes*. The flip side is called co-NP. Asking whether a number is prime, for example, is co-NP, since it’s easy to demonstrate a factor if you happen to know one.

(Actually primality/factorability is in *both*

NP and co-NP. Number theorists

have developed a theory of

“prime certificates,”

which enables one

to verify that a

large number

is prime with

a polynomial

amount of

computation.)

P is a sub-

set of both NP

and co-NP. All

three classes belong

to a larger class called

PSPACE, which belongs to a

yet larger class of problems called

EXP (see Figure 3). EXP consists of Yes/No

problems that can take an exponential amount of time to solve,

possibly because they require intermediate results that have exponentially many digits (for example, does the decimal expansion of 2^{2^n} contain an even number of 1’s?). PSPACE problems demand less extensive bookkeeping. You can think of them as problems that can be worked out in an exponential amount of time on a polynomial-sized—but *erasable*—blackboard.

Computer scientists believe that P is *strictly* smaller than NP, that NP is strictly smaller than PSPACE, and that PSPACE is strictly smaller than EXP. All that’s now known for sure is that P is strictly smaller than EXP—there are indeed problems that cannot

EXP

PSPACE

NP

co-NP

P

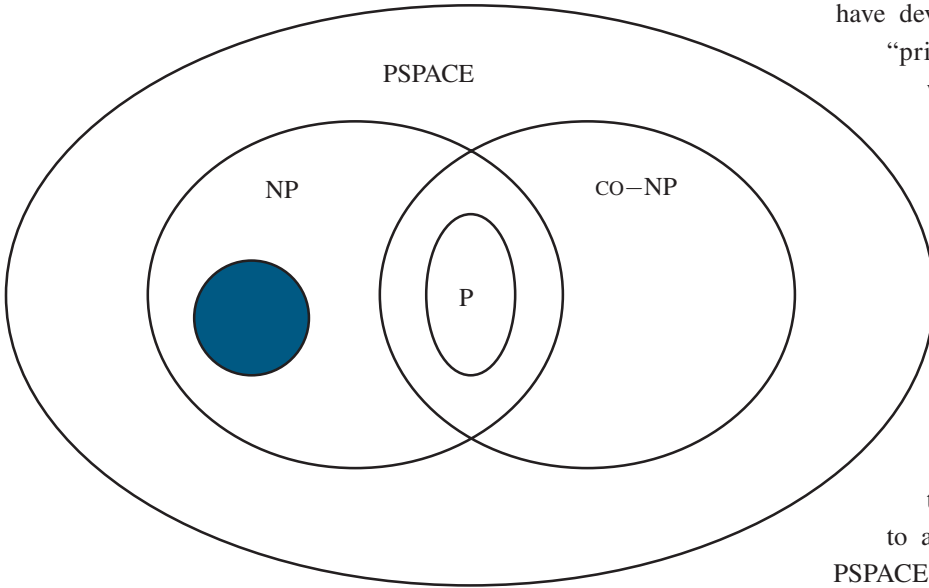


Figure 3. *The hierarchy of complexity theory. The colored circle indicates the class of NP-complete problems.*

be solved in polynomial time. But it may still be the case that $P=NP=PSPACE$, $NP=PSPACE=EXP$, or any other combination that doesn't involve $P=EXP$.

The Millennium problem focuses on P and NP because they contain most of the problems that computers face in practice. To put it jokingly, P includes the problems scientists *can* solve, while NP contains the ones they *want* to solve.

Occasionally a problem formerly only known to be in NP is shown to be in P . The problem of linear programming enjoyed such a demotion in 1979, when Leonid Khachian found a polynomial-time algorithm for it. (Khachian's method was dramatically improved by Narendra Karmarkar at Bell Labs in 1984.) But some NP problems belong to P only if *all* NP problems belong to P . These problems are called NP -complete. The con-

cept was introduced in 1971 by Stephen Cook

at the University of

Toronto, who gave

the first such exam-

ple. Thousands of

problems are now

known to be NP -com-

plete (see "Ising on the

Cake," page 88). In

essence, each one contains all

the difficulties of anything in NP . So if

you want to find a problem in NP can't be solved

in polynomial time, you may as well focus on the ones that are NP -complete.

Alternatively, you might *find* a polynomial-time algorithm for an NP -complete problem. If you do, then you have, in effect, found polynomial-time algorithms for *all* NP problems, and the Clay Mathematics Institute will make you a millionaire. But that million will pale beside what you'll gain by cracking all the cryptographic systems based on the difficulty of NP problems.

The CMI, whose institutional mission is to "further the beauty, power, and universality of mathematical thought," hopes that the prize offer will not only attract new efforts to solve these seven important problems, but recruit more newcomers to mathematics. The intellectual rewards are enormous. The pay's not bad, either.

Computer scientists believe that P is *strictly* smaller than NP , that NP is strictly smaller than $PSPACE$, and $PSPACE$ is strictly smaller than EXP .
