"A Marvelous Proof"

Fernando Q. Gouvêa

No one really knows when it was that the story of what came to be known as "Fermat's Last Theorem" really started. Presumably it was sometime in the late 1630s that Pierre de Fermat made that famous inscription in the margin of Diophantus' *Arithmetic* claiming to have found "a marvelous proof". It seems now, however, that the story may be coming close to an end. In June, 1993, Andrew Wiles announced that he could prove Fermat's assertion. Since then, difficulities seem to have arisen, but Wiles' strategy is fundamentally sound and may yet succeed.

The argument sketched by Wiles is an artful blend of various topics that have been, for years now, the focus of intensive research in number theory: elliptic curves, modular forms, and Galois representations. The goal of this article is to give mathematicians who are not specialists in the subject access to a general outline of the strategy proposed by Wiles. Of necessity, we concentrate largely on background material giving first a brief description of the relevant topics, and only afterwards describe how they come together and relate to Fermat's assertion. Readers who are mainly interested in the structure of the argument and who do not need or want too many details about the background concepts may want to skim through Section 2, then concentrate on Section 3. Our discussion includes a few historical remarks, but history is not our main intention, and therefore we only touch on a few highlights that are relevant to our goal of describing the main ideas in Wiles' attack on the problem.

Thanks are due to Barry Mazur, Kenneth Ribet, Serge Lang, Noriko Yui, George Elliot, Keith Devlin, and Lynette Millett for their help and comments.

1 PRELIMINARIES. We all know the basic statement that Fermat wrote in his margin. The claim is that for any exponent $n \geq 3$ there are no non-trivial integer solutions of the equation $x^n + y^n = z^n$. (Here, "non-trivial" will just mean that none of the integers x, y, and z is to be equal to zero.) Fermat claims, in his marginal note, to have found "a marvelous proof" of this fact, which unfortunately would not fit in the margin.

This statement became known as "Fermat's Last Theorem," not, apparently, due to any belief that the "theorem" was the last one found by Fermat, but rather due to the fact that by the 1800s all of the other assertions made by Fermat had been either proved or refuted. This one was the last one left open, whence the name. In what follows, we will adopt the abbreviation FLT for Fermat's statement, and we will refer to $x^n + y^n = z^n$ as the "Fermat equation."

The first important results relating to FLT were theorems that showed that Fermat's claim was true for specific values of n. The first of these is due to Fermat himself: very few of his proofs were ever made public, but in one that was he shows that the equation

$$x^4 + y^4 = z^2$$

has no non-trivial integer solutions. Since any solution of the Fermat equation with exponent 4 gives a solution of the equation also, it follows that Fermat's claim is true for $n = 4$.

Once that is done, it is easy to see that we can restrict our attention to the case in which n is a prime number. To see this, notice that any number greater than 2 is either divisible by 4 or by an odd prime, and then notice that we can rewrite an equation

$$x^{mk} + y^{mk} = z^{mk}$$

as

$$(x^m)^k + (y^m)^k = (z^m)^k,$$

so that any solution for $n = mk$ yields at once a solution for $n = k$. If n is not prime, we can always choose k to be either 4 or an odd prime, so that the problem reduces to these two cases.

In the 1750s, Euler became interested in Fermat's work on number theory, and began a systematic investigation of the subject. In particular, he considered the Fermat equation for $n = 3$ and $n = 4$, and once again proved that there were no solutions. (Euler's proof for $n = 3$ depends on studying the "numbers" one gets by adjoining $\sqrt{-3}$ to the rationals, one of the first instances where one meets "algebraic numbers.") A good historical account of Euler's work is to be found in **[Wei83]**. In the following years, several other mathematicians extended this step by step to $n = 5, 7, \ldots$. A general account of the fortunes of FLT during this time can be found in **[Rib79]**.

Since then, ways for testing Fermat's assertion for any specific value of n have been developed, and the range of exponents for which the result was known to be true kept getting pushed up. As of 1992, one knew that FLT was true for exponents up to 4 000 000 (by work of J. Buhler).

It is clear, however, that to get general results one needs a general method, i.e., a way to connect the Fermat equation (for any n) with some mathematical context which would allow for its analysis. Over the centuries, there have been many attempts at doing this; we mention only the two biggest successes (omitting quite a lot of very good work, for which see, for example, **[Rib79]**.

The first of these is the work of E. Kummer, who, in the mid-nineteenth century, established a link between FLT and the theory of cyclotomic fields. This link allowed Kummer to prove Fermat's assertion when the exponent was a prime that had a particularly nice property (Kummer named such primes "regular"). The proof is an impressive bit of work, and was the first general result about the Fermat equation. Unfortunately, while in numerical tests a good percentage of primes seem to turn out to be regular, no one has yet managed to prove even that there are infinitely many regular primes. (And, ironically, we do have a proof that there are infinitely many primes that are *not* regular.) A discussion of Kummer's approach can be found in **[Rib79]**; for more detailed information on the cyclotomic theory, one could start with **[Was82]**.

The second accomplishment we should mention is that of G. Faltings, who, in the early 1980s, proved Mordell's conjecture about rational solutions to certain kinds of polynomial equations. Applying this to the Fermat equations, one sees that for any $n \geq 4$ one can have only a *finite* number of non-trivial solutions. Once again, this is an impressive result, but its impact on FLT itself turns out to be minor because we have not yet found a way to actually determine how many solutions should exist. For an introduction to Faltings' work, check **[CS86]**, which contains an English translation of the original paper.

Wiles' attack on the problem turns on another such linkage, also developed in the early 1980s by G. Frey, J.-P. Serre, and K. A. Ribet. This one connects FLT with the theory of elliptic curves, which has been much studied during all of this century, and thereby to all the machinery of modular forms and Galois representations that is the central theme of Wiles' work. The main goal of this paper is to describe this connection and then to explain how Wiles attempts to use it to prove FLT.

Notation. We will use the usual symbols $\mathbb{Q}$ for the rational numbers and $\mathbb{Z}$ for the integers. The integers modulo m will be written[1] as $\mathbb{Z}/m\mathbb{Z}$; we will most often need them when m is a power of a prime number p. If p is prime, then $\mathbb{Z}/p\mathbb{Z}$ is a field, and we commemorate that fact by using an alternative notation: $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$.

2 THE ACTORS.

We begin by introducing the main actors in the drama. First, we briefly (and very informally) introduce the p-adic numbers. These are not so much actors in the play as they are part of the stage set: tools to allow the actors to do their job. Then we give brief and impressionistic outlines of the theories of Elliptic Curves, Modular Forms, and Galois Representations.

2.1 p-adic Numbers. The p-adic numbers are an extension of the field of rational numbers which are, in many ways, analogous to the real numbers. Like the real numbers, they can be obtained by defining a notion of distance between rational numbers, and then passing to the completion with respect to that distance. For our purposes, we do not really need to know much about them. The crucial facts are:

1. For each prime number p there exists a field $\mathbb{Q}_p$ which is complete with respect to a certain notion of distance and contains the rational numbers as a dense subfield.
2. Proximity in the p-adic metric is closely related to congruence properties modulo powers of p. For example, two integers whose difference is divisible by p^n are "close" in the p-adic world (the bigger the n, the closer they are).
3. As a consequence, one can think of the p-adics as encoding congruence information: whenever one knows something modulo p^n for every n, one can translate this into p-adic information, and vice-versa.
4. The field $\mathbb{Q}_p$ contains a subring $\mathbb{Z}_p$, which is called the *ring of p-adic integers*. (In fact, $\mathbb{Z}_p$ is the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$.)

There is, of course, a lot more to say, and the reader will find it said in many references, such as **[Kob84]**, **[Cas86]**, **[Ami75]**, and even **[Gou93]**. The p-adic numbers were introduced by K. Hensel (a student of Kummer), and many of the basic ideas seem to appear, in veiled form, in Kummer's work; since then, they have become a fundamental tool in number theory.

[1] Many elementary texts like to use $\mathbb{Z}_m$ as the notation for the integers modulo m; for us (and for serious number theory in general), this notation is inconvenient because it collides with the notation for the p-adic integers described below.

2.2 Elliptic Curves. Elliptic curves are a special kind[2] of algebraic curves which have a very rich arithmetical structure. There are several fancy ways of defining them, but for our purposes we can just define them as the set of points satisfying a polynomial equation of a certain form.

To be specific, consider an equation of the form

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

where the a_i are integers (there is a reason for the strange choice of indices on the a_i, but we won't go into it here). We want to consider the set of points (x, y) which satisfy this equation. Since we are doing number theory, we don't want to tie ourselves down too seriously as to what sort of numbers x and y are: it makes sense to take them in the real numbers, in the complex numbers, in the rational numbers, and even, for any prime number p, in $\mathbb{F}_p$ (in which case we think of the equation as a congruence modulo p). We will describe the situation by saying that there is an underlying object which we call *the curve* E and, for each one of the possible fields of definition for points (x, y), we call the set of possible solutions the "points of E" over that field. So, if we consider all possible complex solutions, we get the set $E(\mathbb{C})$ of the complex points of E. Similarly, we can consider the real points $E(\mathbb{R})$, the rational points $E(\mathbb{Q})$, and even the $\mathbb{F}_p$-points $E(\mathbb{F}_p)$.

We haven't yet said when it is that such equations define elliptic curves. The condition is simply that the curve be *smooth*. If we consider the real or complex points, this means exactly what one would expect: the set of points contains no "singular" points, that is, at every point there is a well-defined tangent line. We know, from elementary analysis, that an equation $f(x, y) = 0$ defines a smooth curve exactly when there are no points on the curve at which both partial derivatives of f vanish. In other words, the curve will be smooth if there are no common solutions of the equations

$$f(x,y) = 0 \qquad \frac{\partial f}{\partial x}(x,y) = 0 \qquad \frac{\partial f}{\partial y}(x,y) = 0.$$

Notice, though, that this condition is really algebraic (the derivatives are derivatives of polynomials, and hence can be taken formally). In fact, we can boil it down to a (complicated) polynomial condition in the a_i. There is a polynomial $\Delta(E) = \Delta(a_1, a_2, a_3, a_4, a_6)$ in the a_i such that E is smooth if and only if $\Delta(E) \neq 0$. This gives us the means to give a completely formal definition (which makes sense even over $\mathbb{F}_p$). The number $\Delta(E)$ is called the *discriminant* of the curve E.

Definition 1. *Let K be a field. An elliptic curve over K is an algebraic curve determined by an equation of the form*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

where each of the a_i belongs to K and such that $\Delta(a_1, a_2, a_3, a_4, a_6) \neq 0$.

Specialists would want to rephrase that definition to allow other equations, provided that a well-chosen change of variables could transform them into equations of this form.

[2] Perhaps it's best to dispel the obvious confusion right up front: ellipses are not elliptic curves. In fact, the connection between elliptic curves and ellipses is a rather subtle one. What happens is that elliptic curves (over the complex numbers) are the "natural habitat" of the elliptic integrals which arise, among other places, when one attempts to compute the arc length of an ellipse. For us, this connection will be of very little importance.

It's about time to give some examples. To make things easier, let us focus on the special case in which the equation is of the form $y^2 = g(x)$, with $g(x)$ a cubic polynomial (in other words, we're assuming $a_1 = a_3 = 0$). In this case, it's very easy to determine when there can be singular points, and even what sort of singular points they will be. If we put $f(x, y) = y^2 - g(x)$, then we have

$$\frac{\partial f}{\partial x}(x, y) = -g'(x) \quad \text{and} \quad \frac{\partial f}{\partial y}(x, y) = 2y,$$

and the condition for a point to be "bad" becomes

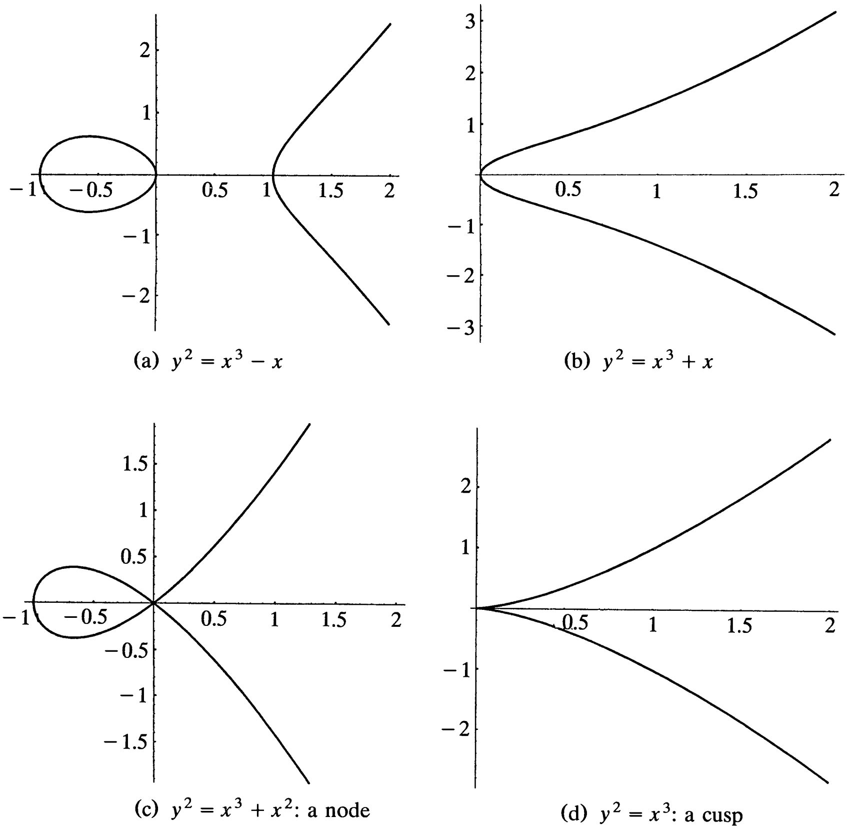
$$y^2 = g(x) \qquad -g'(x) = 0 \qquad 2y = 0,$$

which boils down to $y = g(x) = g'(x) = 0$. In other words, a point will be bad exactly when its y-coordinate is zero and its x-coordinate is a *double root* of the polynomial $g(x)$. Since $g(x)$ is of degree 3, this gives us only three possibilities:

- $g(x)$ has no multiple roots, and the equation defines an elliptic curve;
- $g(x)$ has a double root;
- $g(x)$ has a triple root.

Let's look at one example of each case, and graph the real points of the corresponding curve.

For the first case, consider the curve given by $y^2 = x^3 - x$. Its graph is in figure 1 (a) (to be precise, this is the graph of its real points). A different example of the same case is given by $y^2 = x^3 + x$; see figure 1(b). (The reason these look so



(a) $y^2 = x^3 - x$

(b) $y^2 = x^3 + x$

(c) $y^2 = x^3 + x^2$: a node

(d) $y^2 = x^3$: a cusp

different is that we are only looking at the real points of the curve; in fact, over the complex numbers these two curves are isomorphic.)

When there are "bad" points, what has happened is that either two roots of $g(x)$ have "come together" or all three roots have done so. In the first case, we get a loop. At the crossing point, which is usually called a "node," the curve has two different tangent lines. See Figure 1(c), where we have the graph of the equation $y^2 = x^3 + x^2$ (double root at zero).

In the final case, not only have all three roots of $g(x)$ come together, but also the two tangents in the node have come together to form a sort of "double tangent" (this can be made precise with some easy algebra of polynomials, but it's more fun to think of it geometrically). The graph now looks like Figure 1(d), and we call this kind of singular point a "cusp".

How does all this relate to the discriminant Δ we mentioned above? Well, if r_1, r_2 and r_3 are the roots of the polynomial $g(x)$, the discriminant for the equation $y^2 = g(x)$ turns out to be

$$\Delta = K(r_1 - r_2)^2(r_1 - r_3)^2(r_2 - r_3)^2,$$

where K is a constant. This does just what we want: if two of the roots are equal, it is zero, and if not, not. Furthermore, it is not too hard to see that Δ is actually a polynomial in the coefficients of $g(x)$, which is what we claimed. In other words, all that the discriminant is doing for us is giving a direct algebraic procedure for determining whether there are singular points.

While this analysis applies specifically to curves of the form $y^2 = g(x)$, it actually extends to all equations of the sort we are considering: there is at most one singular point, and it is either a node or a cusp.

One final geometric point: as one can see from the graphs, these curves are not closed. It is often convenient to "close them up." This is done by adding one more point to the curve, usually referred to as "the point at infinity." This can be done in a precise way by embedding the curve into the projective plane, and then taking the closure. For us, however, the only important thing is to remember that we actually have one extra point on our curves. (One should imagine it to be "infinitely far up the y-axis," but keep in mind that there is only one "point at infinity" on the y-axis, so that it is *also* "infinitely far down.")

With some examples in hand, we can proceed to deeper waters. In order to understand the connection we are going to establish between elliptic curves and FLT, we need to review quite a large portion of what is known about the rich arithmetic structure of these curves.

The first thing to note is that one can define an operation on the set of points of an elliptic curve that makes it, in a natural way, an abelian group. The operation is usually referred to as "addition." The identity element of this group turns out to be the point at infinity (it would be more honest to say that we *choose* the point at infinity for this role).

We won't enter into the details of how one adds points on an elliptic curve. In fact, there are several equivalent definitions, each of which has its advantages! The reader should see the references for more details of how it is done (and the proof that one does get a group). The main thing to know about the definition, for now, is that it preserves the field of definition of the points: adding two rational points gives a rational point, and so on.

What this means is that for every choice of a base field, we can get a group of points on the curve with coordinates in that field, so that in fact an elliptic curve gives us a whole bunch of groups, which are, of course, all related (though sometimes related in a mysterious way). So, given an E, we can look at the complex points $E(\mathbb{C})$, which form a complex Lie group which is topologically a torus, or we can look at the real Lie group $E(\mathbb{R})$, which turns out to be either isomorphic to the circle S^1 or to the direct product $\mathbb{Z}/2\mathbb{Z} \times S^1$. (Look back at the examples above; can you see which is which?)

From an arithmetical point of view, however, the most interesting of these groups is the group of rational points, $E(\mathbb{Q})$. A point $P \in E(\mathbb{Q})$ gives a solution in rational numbers of our cubic equation, and looking for such solutions is, of course, an example of solving a diophantine equation, a sort of problem that is quite important in number theory. What is especially nice about $E(\mathbb{Q})$ is the fact, proved by L. Mordell (and extended by A. Weil) in the 1920s, that it is a *finitely generated* abelian group. What this means is just the following: there is a finite list of rational points on the curve (or, if one prefers, of rational solutions to the equation) such that every other rational solution is obtained by combining (using the addition law) these points with one another. These points are called the *generators* of the group $E(\mathbb{Q})$, which is usually called the *Mordell-Weil group* of E.

The curves we considered earlier have very simple Mordell-Weil groups. For the curve given by $y^2 = x^3 - x$ (figure 1a), it has four points; and for $y^2 = x^3 + x$ (figure 1b) it has two. It is easy, though, to give more interesting examples. Here is one, chosen at random from **[Cre92]**: if E is the curve defined by $y^2 + y = x^3 - x^2 - 2x + 2$, the Mordell-Weil group $E(\mathbb{Q})$ is an infinite cyclic group, generated by the point $(2, 1)$.

Of course, knowing that we have a finitely generated group raises the obvious question of estimating or computing the number of generators needed and of how one might go about actually finding these generating points. Both of these questions are still open, even though there are rather precise conjectures about what their answers should be. For many specific curves, both the number and the generators themselves have been completely worked out (see, for example, the tables in **[Cre92]**), but the general problem still seems quite difficult.

A fundamental component of the conjectural plan for determining the generators is considering, for each prime number p, the reduction of our curve modulo p. The basic idea is quite simple: since our equation has integer coefficients, we can reduce it modulo p" and look for solutions in the field $\mathbb{F}_p$ of integers modulo p. This should give a finite[3] group $E(\mathbb{F}_p)$, whose structure should be easier to analyse than that of the big group $E(\mathbb{Q})$. It's a rather simple idea, but several complications[4] arise.

The main thing that can go wrong is that the reduction modulo p may fail to be an elliptic curve. That is actually very easy to see. To tell whether the curve is elliptic (that is, if it has no singular points), one needs to look at Δ. It is perfectly possible for Δ to be nonzero (so the curve over $\mathbb{Q}$ is elliptic) while being at the same time congruent to zero modulo p (so that the curve over $\mathbb{F}_p$ is singular). This phenomenon is called *bad reduction*, and it is easy to come up with examples. One might take $p = 5$, and look at the curve $y^2 = x^3 - 5$. This turns out to be an elliptic curve over $\mathbb{Q}$, but its reduction modulo 5 is going to have a cusp. One says, then, that the curve has bad reduction at 5. In fact, the discriminant turns out to be

[3] It is finite because, apart from the point at infinity, there are only p^2 possible points. In fact, the maximum possible number of points is smaller than that, but that fact takes some proving.

[4] It may seem a bit perverse to dwell on the nature of these complications, but it will turn out that we need to have at least some understanding of how this goes later on.

$\Delta = -10800$, which is clearly divisible by 2, 3, and 5, so that the curve has bad reduction at each of these. (In each case, it's easy to verify that the reduced curve has a cusp.)

We want to classify the possible types of reduction, but there is one further glitch that we have to deal with before we can do so. To see what it is, consider the curve $y^2 = x^3 - 625x$. At first glance, it seems even worse than the first, and the discriminant, which turns out to be $\Delta = -15625000000$, looks *very* divisible by 5. But look what we can do: let's change variables by setting $x = 25u = 5^2u$ and $y = 125v = 5^3v$. Then our equation becomes

$$(5^3v)^2 = (5^2u)^3 - 625(5^2u),$$

which simplifies to

$$5^6v^2 = 5^6u^3 - 5^6u,$$

and hence to

$$v^2 = u^3 - u,$$

which is not only a nice elliptic curve, but has good reduction at 5. In other words, this example shows that *curves which are isomorphic over* $\mathbb{Q}$ *can have very different reductions modulo* p.

It turns out that among all the possible equations for our curve, one can choose an equation that is *minimal*, in the sense that its discriminant will be divisible by fewer primes than the discriminant for other equations. Since the primes that divide the discriminant are the primes of bad reduction, a minimal equation will have reduction properties that are as good as possible. When studying the reduction properties of the curve, then, one must also pass to such a minimal equation (and there are algorithms to do this).

Well, then, suppose we have done so, and have an elliptic curve E given by a minimal equation. Then we can classify all prime numbers into three groups:

- *Primes of good reduction*: those which do not divide the discriminant of the minimal equation. The curve modulo p is an elliptic curve, and we have a group $E(\mathbb{F}_p)$.
- *Primes of multiplicative reduction*: those for which the curve modulo p has a node. If the singular point is (x_0, y_0), it turns out that the set $E(\mathbb{F}_p) - \{(x_0, y_0)\}$ has a group structure, and is isomorphic to the multiplicative group $\mathbb{F}_p - \{0\}$.
- *Primes of additive reduction*: those for which the curve modulo p has a cusp. If the singular point is (x_0, y_0), the set $E(\mathbb{F}_p) - \{(x_0, y_0)\}$ once again has a group structure, and is isomorphic to the additive group $\mathbb{F}_p$.

No curve can have good reduction everywhere, so there will always be some bad primes, but the feeling one should get is that multiplicative reduction is somehow not as bad as additive reduction. There are various technical reasons for this, which we don't really need to go into. Instead, we codify the information about the reduction types of the curve into a number, called the *conductor* of the curve. We define the conductor to be a product $N = \prod p^{n(p)}$, where

$$n(p) = \begin{cases} 0 & \text{if } E \text{ has good reduction at } p \\ 1 & \text{if } E \text{ has multiplicative reduction at } p \\ \geq 2 & \text{if } E \text{ has additive reduction at } p \end{cases}$$

(The exact value of $n(p)$ for the case of additive reduction depends on some rather subtle properties of the reduction modulo such primes; most of the time, the exponent is 2.) The result is that one can tell, by looking at the conductor, exactly what the reduction type of E at each prime is.

The elliptic curves we will want to consider are those whose reduction properties are as good as possible. Since good reduction at all primes is not possible, we opt for the next best thing: good reduction at almost all primes, multiplicative reduction at the others. Such curves are called *semistable*:

Definition 2. *An elliptic curve is called semistable if all of its reductions are either good or multiplicative. Equivalently, a curve is semistable if its conductor is square-free.*

A crucial step in the application of Wiles' theorem to FLT will be verifying that a certain curve is semistable. Just to give us some reference points, let's look at a few examples.

1. Let E_1 be the curve $y^2 = x^3 - 5$, which we considered above. One checks that this equation is minimal, and that the curve has additive reduction at 2, 3, and 5, so that it is not semistable. The conductor turns out to be equal to 10800 (essentially, the same as the discriminant!).
2. Let E_2 be the curve $y^2 + y = x^3 + x$. This has multiplicative reduction at 7 and 13 (checking this makes a nice exercise) and good reduction at all other primes. Hence, E_2 is semistable and its conductor is 91.
3. Let E_3 be the curve $y^2 = x^3 + x^2 + 2x + 2$ (which is minimal). This has discriminant $\Delta = -1152 = -2^7 \cdot 3^2$, so that the bad primes are 2 and 3. It turns out that the reduction is multiplicative at 3 and additive at 2, and the conductor is 384; the curve is not semistable.
4. *The main example for the purpose at hand*: Let $a, b,$ and c be relatively prime integers such that $a + b + c = 0$. Consider the curve E_{abc} whose equation[5] is $y^2 = x(x - a)(x + b)$. Depending on what $a, b,$ and c are, this equation may or may not be minimal, so let's make the additional assumptions that $a \equiv -1 \pmod{4}$ and that $b \equiv 0 \pmod{32}$. In this case, the equation is *not* minimal. A minimal equation for this curve turns out to be
$$y^2 + xy = x^3 + \frac{b - a - 1}{4}x^2 - \frac{ab}{16},$$
which we get by the change in variables $x \to 4x, y \to 8y + 4x$. One can then compute that the discriminant is $\Delta = a^2b^2c^2/256$ (not surprising: a constant times the product of the squares of the differences of the roots of the original cubic), and that the curve is semistable. The primes of bad reduction are those that divide abc (this would be easy to see directly from the equation, by checking when there is a multiple root modulo p), and therefore the conductor is equal to the product of the primes that divide abc:
$$N = \prod_{p|abc} p$$

[5] It may strike the reader as funny that c is absent from the equation. Keep in mind, however, that c is completely determined by a and b, so that it is really not as absent as all that. The crucial point is that the roots of the cubic on the right hand side are 0, a, and $-b$, so that the differences of the roots are (up to sign) exactly $a, b,$ and c.

(this number is sometimes called the *radical* of abc). We will be using curves of the form E_{abc} (for very special a, b, and c) when we make the link with FLT.

We need a final bit of elliptic curve theory. It is interesting to look at the number of points in the groups $E(\mathbb{F}_p)$ as p ranges through the primes of good reduction for E. Part of the motivation for this is the reasoning that if the group $E(\mathbb{Q})$ is large (i.e. there are many rational solutions), one would expect that for many choices of the prime p many of the points in $E(\mathbb{Q})$ would survive reduction modulo p, so that the group $E(\mathbb{F}_p)$ would be large. Therefore, one would like to make some sort of conjecture that said that if the $E(\mathbb{F}_p)$ are very large for many primes p, then the group $E(\mathbb{Q})$ will be large.

Elaborating and refining this idea leads to the conjecture of Birch and Swinnerton-Dyer, which we won't get into here. But even this coarse version suggests that the variation of the size of $E(\mathbb{F}_p)$ as p runs through the primes should tell us something about the arithmetic on the curve. To "encode" this variation, we start by observing that the (projective) line over $\mathbb{F}_p$ has exactly $p + 1$ points (the p elements of $\mathbb{F}_p$, plus the point at infinity). We take this as the "standard" number of points for a curve over $\mathbb{F}_p$, and, when we look at $E(\mathbb{F}_p)$, record how far from the standard we are. To be precise, given an elliptic curve E and a prime number p at which E has good reduction, we define a number a_p by the equation

$$\#E(\mathbb{F}_p) = p + 1 - a_p.$$

For primes of bad reduction, we extend the definition in a convenient way; it turns out that we get $a_p = \pm 1$ when the reduction is multiplicative (with a precise rule to decide which) and $a_p = 0$ when it is additive.

The usual way to "record" the sequence of the a_p is to use them to build a complex analytic function called the *L-function* of the curve E. It then is natural to conjecture that this L-function has properties similar to those of other L-functions that arise in number theory, and that one can read off properties of E from properties of its L-function. This is a huge story which we cannot tell in this article, but which is really very close to some of the issues which we do discuss later on. Suffice it to say, for now, that we get a function

$$L(E,s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s},$$

where the a_p are exactly the same as the ones we just introduced, the a_n are determined from the a_p by "Euler product" expansion for the L-function, and the series can be shown to converge when $\mathrm{Re}(s) > 3/2$. The L-function is conjectured to have an analytic continuation to the whole complex plane and to satisfy a certain functional equation.

It is time to introduce the other actors in the play and to explain how they relate to elliptic curves. The reader who would like to delve further into this theory has a lot to choose from. As an informal introduction, one could look at J. Silverman's article **[Sil93]**, which relates elliptic curves to "sums of two cubes" and Ramanujan's taxicab number. Various introductory texts are available, including **[Cas91]**, **[Hus87]**, **[Kna92]**, **[Sil86]**, and **[ST92]**. Each of these has particular strengths; the last is intended as an undergraduate text. In addition to these and other texts, the interested reader might enjoy looking at symbolic manipulation software that will handle elliptic curves well. Such capabilities are built into GP-PARI and SIMATH, and can be added to *Mathematica* by using Silverman's *EllipticCurveCalc* package (which is what we used for most of the computations in this paper), and to *Maple* by using Connell's *Apecs* package. See[**C** +], [**Z** +], [**SvM**], [**Con**].

2.3 Modular Forms. Modular forms start their lives as analytic objects (or, to be more honest, as objects of group representation theory), but end up playing a very intriguing role in number theory. In this section, we will *very* briefly sketch out their definition and explain their relation to elliptic curves.

Let $\mathfrak{h} = \{x + iy | y > 0\}$ be the complex upper half-plane. As is well known (and, in any case, easy to check), matrices in $\mathrm{SL}_2(\mathbb{Z})$ act on $\mathfrak{h}$ in the following way. If $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ is the matrix

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

(so that $a, b, c,$ and d are integers and $ad - bc = 1$), and $z \in \mathfrak{h}$, we define

$$\gamma \cdot z = \frac{az + b}{cz + d}.$$

It is easy to check that if $z \in \mathfrak{h}$ then $\gamma \cdot z \in \mathfrak{h}$, and that $\gamma_1 \cdot (\gamma_2 \cdot z) = (\gamma_1 \gamma_2) \cdot z$.

We want to consider functions on the upper half-plane which are "as invariant as possible" under this action, perhaps when restricted to a smaller group. The subgroups we will need to consider are the "congruence subgroups" which we get by adding a congruence condition to the entries of the matrix. Thus, for any positive integer N, we want to look at the group

$$\Gamma_0(N) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) | c \equiv 0 (mod N) \right\}.$$

We are now ready to begin defining modular forms. They will be functions $f : \mathfrak{h} \to \mathbb{C}$, holomorphic, which "transform well" under one of the subgroups $\Gamma_0(N)$. To be specific, we require that there exist an integer k such that

$$f\left(\frac{az + b}{cz + d}\right) = (cz + d)^k f(z).$$

Applying this formula to the special case in which the matrix is

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

shows that any such function must satisfy $f(z + 1) = f(z)$, and hence must have a Fourier expansion

$$f(z) = \sum_{n=-\infty}^{\infty} a_n q^n \quad \text{where } q = e^{2\pi i z}.$$

We require that this expression in fact only involve non-negative powers of q (and in fact we extend that requirement to a finite number of other, similar, expansions, which the experts call the "Fourier expansions at the other cusps"). A function satisfying all of these constraints is called *a modular form of weight k on $\Gamma_0(N)$*. The number N is usually called the *level* of the modular form f.

We will need to consider one special subspace of the space of modular forms of a given weight and level. Rather than having a Fourier expansion with non-negative powers only, we might require *positive* powers only (in the main expansion and in the ones "at the other cusps"). We call such modular forms *cusp forms*; they turn out to be the more interesting part of the space of modular forms.

Finally, one must make a remark on the relation between the theory at various levels: if N divides M, then every form of level N (and weight k) gives rise to (a number of) forms of level M (and the same weight). The subspace generated by all forms of level M and weight k which arise in this manner (from the various divisors of M) is called the space of *old forms* of level M. With respect to a natural inner product structure on the space of modular forms, one can then take the orthogonal complement of the space of old forms. This complement is called the space of *new forms*, which are the ones we will be most interested in.

What really makes the theory of modular forms interesting for arithmetic is the existence of a family of commuting operators on each space of modular forms, called the Hecke operators. We will not go into the definition of these operators (they are quite natural from the point of view of representation theory); for us the crucial things will be:

- For each positive integer n relatively prime to the level N, there is a Hecke operator T_n acting on the space of modular forms of fixed weight and level N.
- The Hecke operators commute with each other.
- If m and n are relatively prime, then $T_{nm} = T_n T_m$.

We will be especially interested in modular forms which are eigenvectors for the action of all the Hecke operators, i.e., forms for which there exist numbers λ_n such that $T_n(f) = \lambda_n f$ for each n which is relatively prime to the level. We will call such forms *eigenforms*.

This is all quite strange and complicated, so let's immediately point out one connection between modular forms and elliptic curves. Suppose one has a modular form which is

- of weight 2 and level N,
- a cusp form,
- new,
- an eigenform.

If that is the case, one can normalize the form so that its Fourier expansion looks like

$$f(z) = \sum_{n=1}^{\infty} a_n q^n \quad \text{with } a_1 = 1.$$

Suppose that, once we have done the normalization,

- all of the Fourier coefficients a_n are integers.

Then there exists an elliptic curve whose equation has integer coefficients, whose conductor is N, and whose a_n are exactly the ones that appear in the Fourier expansion of f. In particular, the L-function of E can be expressed in terms of f (as a Mellin transform), and the nice analytic properties of f then allow us to prove that the L-function does have an analytic continuation and does satisfy a functional equation.

This connection between forms and elliptic curves is so powerful that it led people to investigate the matter further. The first one to suggest that *every* elliptic curve should come about in this manner was Y. Taniyama, in the mid-fifties. The suggestion only penetrated the mathematical culture much later, largely due to the work of G. Shimura, and it was made more precise by A. Weil's work pinning down the role of the conductor. We now call this the "Shimura-Taniyama-Weil

Conjecture." Here it is:

Conjecture 1 (Shimura-Taniyama-Weil). *Let E be an elliptic curve whose equation has integer coefficients. Let N be the conductor of E, and for each n let a_n be the number appearing in the L-function of E. Then there exists a modular form of weight 2, new of level N, an eigenform under the Hecke operators, and (when normalized) with Fourier expansion equal to $\Sigma a_n q^n$.*

For any specific curve, it is not too hard to check that this is true. One takes E, determines the conductor and the a_n for a range of n. Since the space of modular forms of weight 2 and level N is finite-dimensional, knowing enough of the a_n must determine the form, and we can go and look if it is there. (In general, given a list of a_n, it is not at all easy to determine whether $\Sigma a_n q^n$ is the Fourier expansion of a modular form, so we need to do it the other way: we generate a basis of the space of modular forms, then try to find our putative form as a linear combination of the basis.) If we find a form with the right (initial chunk of) Fourier expansion, this gives prima facie evidence that the curve satisfies the STW conjecture. To clinch the matter, one can use a form of the Čebotarev density theorem to show that if *enough* (in an explicit sense) of the a_n are right, then they all are.

This method has been used to verify the STW conjecture for any number of specific curves (see, for example, **[Cre92]**). The conjecture has a really crucial role in the theory of elliptic curves; in fact, curves that satisfy the conjecture are known as "modular elliptic curves," and many of the fundamental new results in the theory have only been proved for curves that have this property.

As our final remark on modular forms, we point out that it is possible, for any given N, to determine (essentially using the Riemann-Roch theorem) the exact dimension of the space of cusp forms of weight 2 and level N. This gives us a very good handle on what curves of that conductor should exist (if the STW conjecture is true).

For more information on modular forms, one might look at **[Lan76]**, **[Miy89]**, or **[Shi71]**. There is an intriguing account of the Shimura-Taniyama-Weil conjecture, in a very different spirit, in Mazur's article **[Maz91]**, and a useful survey in **[Lan91]**.

2.4 Galois Representations. The final actors in our play are Galois representations. One starts with the Galois group of an extension of the field of rational numbers. To understand this Galois group, one can try to "represent" the elements of the group as matrices. In other words, one can try to find a vector space on which our Galois group acts, which gives a way to associate a matrix to each element of the group. This in fact gives a group homomorphism from the Galois group to a group of matrices (this need not be injective; when it is, one calls the representation "faithful").

Rather than work with specific finite extensions of $\mathbb{Q}$, we work with the Galois group $G = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ of the algebraic closure of $\mathbb{Q}$. This is a huge group (which one makes more manageable by giving it a topology) that hides within itself an enormous amount of arithmetic information. The representations we will be considering will be into 2×2 matrices over various fields and rings, and they will (for the most part) be obtained from elliptic curves and from modular forms.

To see how to get Galois representations from an elliptic curve, let's start with an elliptic curve E, whose equation has coefficients in $\mathbb{Z}$. Choose a prime p. Since the (complex, say) points of E form a group, one can look in this group for points which are of order p (that is, for points (x, y) such that adding them to themselves

p times gives the identity). It turns out that (over $\mathbb{C}$) there are p^2 such points, and they form a subgroup which we donate by $E[p]$. In fact, this group is isomorphic to the product of two copies of $\mathbb{F}_p$:

$$E[p] \cong \mathbb{F}_p \times \mathbb{F}_p.$$

Now, the points in $E[p]$ are a priori complex, but on closer look one sees that in fact they are all defined over some extension of $\mathbb{Q}$, and in particular that transforming the coefficients of a point of order p by the Galois group G yields another point of order p. In fact, it's even better than that: since the rule for adding points is defined in rational terms, the whole group structure is preserved. Since $E[p]$ looks like a vector space of dimension 2 over $\mathbb{F}_p$, this means that each element of G acts as a linear transformation on this space, and hence that we get a representation

$$\bar{\rho}_{E,p}: G \to \mathrm{GL}_2(\mathbb{F}_p).$$

(We use a bar to remind ourselves that this is a representation "modulo p.")

Now, $\mathrm{GL}_2(\mathbb{F}_p)$ is a finite group, and G is very infinite, so this representation, while it tells us a lot, can't be the whole story. It turns out, however, that we can use p-adic numbers to get a whole lot more. Instead of considering only the points of order p, we can consider points of order p^n for each n. This gives a whole bunch of subgroups

$$E[p] \subset E[p^2] \subset E[p^3] \subset \dots$$

and a whole bunch of representations, into $\mathrm{GL}_2(\mathbb{F}_p)$, then into $\mathrm{GL}_2(\mathbb{Z}/p^2\mathbb{Z})$, then into $\mathrm{GL}_2(\mathbb{Z}/p^3\mathbb{Z})\dots$. Putting all of these together ends up by giving us a p-adic representation

$$\rho_{E,p}: G \to \mathrm{GL}_2(\mathbb{Q}_p)$$

which hides within itself all of the others. The representations $\rho_{E,p}$ contain a lot of arithmetic information about the curve E.

And how does it look on the modular forms side? Well, it follows from the work of several mathematicians (M. Eichler, G. Shimura, P. Deligne, and J.-P. Serre) that, whenever we have a modular form f (of any weight) which is an eigenform for the action of the Hecke operators and whose Fourier coefficients (after normalization) are integers, we can construct a representation

$$\rho_{f,p}: G \to \mathrm{GL}_2(\mathbb{Q}_p)$$

which is attached to f in a precise sense which is too technical to explain here. (The construction of the representation is quite difficult, and in fact no satisfactory expository account is yet available.)

The crucial thing to know, for our purposes, is that *when an elliptic curve E arises from a modular form f, then the representations $\rho_{E,p}$ and $\rho_{f,p}$ are the same.* In fact, a converse is also true: given a curve E, if one can find a modular form f such that $\rho_{E,p}$ is the same as $\rho_{f,p}$ then E will be modular.

3 THE PLAY. We are now ready to take the plunge and try to see how all of this theory relates to Fermat's Last Theorem. The idea is to assume that FLT is false, and then, using this assumption, to construct an elliptic curve that contradicts just about every conjecture under the sun.

3.1 Linking FLT to Elliptic Curves. So let's start by assuming FLT is false, i.e., that there exist three non-zero integers $u, v,$ and w such that $u^p + v^p + w^p = 0$ (as we know, we only need to consider the case of prime exponent p, which is therefore odd, so that we can recast a solution in Fermat's form to be in the form above). Since we know that the theorem is true for $p = 3$, we might as well assume that $p \geq 5$. We may clearly assume that $u, v,$ and w are relatively prime, which means that precisely one of them must be even. Let's say v is even. Since p is bigger than two, we can see, by looking at the equation modulo 4, that one of u and w must be congruent to -1 modulo 4, and the other must be congruent to 1. Let's say $u \equiv -1 \pmod 4$.

Let's use this data to build an elliptic curve, following an idea due to G. Frey (see **[Fre86]**, **[Fre87a]**, **[Fre87b]**). We consider the curve

$$y^2 = x(x - u^p)(x + v^p).$$

This is usually known as the Frey curve. Following our discussion, above, of the curve $E_{abc,}$ we already know quite a bit about the Frey curve. Here's a summary:

1. Since v is even and $p \geq 5$, we know that we have $v^p \equiv 0 \pmod{32}$. We also know that $u^p \equiv -1 \pmod 4$. This puts us in the right position to use what we know about curves E_{abc}.
2. The minimal discriminant of the Frey curve is
$$\Delta = \frac{(uvw)^{2p}}{256}.$$
3. The conductor of the Frey curve is the product of all the primes dividing $u^p v^p w^p$, which is, of course, the same as the product of all the primes dividing uvw.
4. The Frey curve is semistable.

Now, as Frey observed in the mid-1980s, this curve seems much too strange to exist. For one thing, its conductor is extremely small when compared to its discriminant (because of that exponent of $2p$). For another, its Galois representations are pretty weird. Very soon, people were pointing out that there were several conjectures that would rule out the existence of Frey's curve, and therefore would prove that Fermat was correct in saying that his equation had no solutions.

3.2 FLT follows from the Shimura-Taniyama-Weil Conjecture. It was already clear to Frey that it was likely that the existence of his curve would contradict the Shimura-Taniyama-Weil conjecture, but he was unable to give a solid proof of this. A few months after Frey's work, Serre pinpointed, in a letter to J.-F. Mestre, exactly what one would need to prove to establish the link. In this letter (published as **[Ser87a]**), Serre describes the situation with the phrase "STW + ε implies Fermat." Because of this, the missing theorem became known, for a while, as "conjecture epsilon." This conjecture was proved by K. A. Ribet in **[Rib90]** about a year later, and this established the link. A survey of these results can be found in **[Lan91]**.

What Serre noticed was that the representation modulo p

$$\bar{\rho}_{E,p}: G \to \mathrm{GL}_2(\mathbb{F}_p)$$

obtained from the Frey curve was rather strange. It looked like the sort of representation one would get from a modular form of weight 2, but if one applied the "usual recipe" for guessing the level of that modular form, the answer came out to be $N = 2$. He also showed that the modular form must be a cusp form. The problem is that *there are no cusp forms of weight* 2 *and level* 2!

So suppose there is a solution of the Fermat equation for some prime p, and use this solution to build a Frey curve E. Let N be the conductor of E (which we determined above). Suppose, also, that STW holds for E, so that there exists a modular form of weight 2 and level N whose Galois representation is the same as the one for E. Then we have the following curious situation: we have a representation $\bar{\rho}$ which we know comes from a modular form of weight 2 and level N, but which *looks* as if it should come from a modular form of smaller level.

Here is where Ribet's theorem comes in: he proves that (under certain hypotheses which will hold in our case) whenever this happens the modular form of smaller level must actually exist! Notice that this doesn't mean that the original modular form came from lower level; what it means is that there is a form of lower level whose representation reduces modulo p to the same representation.

The upshot of Ribet's theorem is the following:

Theorem 1 (Ribet). *Suppose STW holds for all semistable elliptic curves. Then FLT is true.*

This is true because if FLT were false, one could choose a solution of the Fermat equation and use it to construct a Frey curve, which would be a semistable elliptic curve. By STW, this curve would be attached to a modular form, so that its Galois representation is attached to a modular form. By Ribet's theorem, there must exist a modular form of weight 2 and level 2 which gives the same representation modulo p. Just a little more work allows one to check that this modular form must be a cusp form. But this is a contradiction, because there are *no* cusp forms of weight 2 and level 2.

3.3 Deforming Galois Representations. It is now that we come to Wiles' work. His idea was that one can attack the problem of proving STW by using the Galois representations, and in particular by thinking of "deformations" of Galois representations. The idea is to consider not only a representation modulo p, but also *all* the possible p-adic representations attached to it (one speaks of "all the possible lifts" of the representation modulo p). These can be thought of as "deformations" because, from the p-adic point of view, they are "close" to the original representation.

This sort of idea had been introduced by B. Mazur in **[Maz89]**. Mazur showed that one could often obtain a "universal lift," i.e., a representation into GL_2 of a big ring such that all possible lifts were "hidden" in this representation. If one knew that the representation modulo p were modular, then one could make another big ring "containing" all the lifts which are attached to modular forms. The abstract deformation theory then provides us with a homomorphism between these two rings, and one can try to prove that this is an isomorphism. If so, it follows that all lifts are modular.

What Wiles proposes to do is very much in this spirit, except that he restricts himself to lifts that have especially nice properties. He starts with a representation modulo p, and supposes that it is modular and that it satisfies certain technical assumptions. Then he considers all possible deformations which "look like they could be attached to forms of weight 2," and gets a deformation ring. Considering all deformations which are attached to modular forms of weight 2 gives a second ring (which is closely related to the algebra generated by the Hecke operators, in fact). Wiles then attempts to prove, using a vast array of recent results, including ideas of Mazur, Ribet, Faltings, V. Kolyvagin, and M. Flach, that these two rings are the same.

It is not hard to see that the homomorphism between the two rings we want to consider is surjective. The difficulty is to prove it is also injective. Wiles reduces this question to bounding the size of a certain cohomology group. It is here that the brilliant ideas of Kolyvagin and of Flach come in. About five years ago, Kolyvagin came up with a very powerful method for controlling the size of certain cohomology groups, using what he calls "Euler systems" (see **[Kol91]** and the survey of the method in **[Maz93]**). This method seems to be adaptable to any number of situations, and has been used to prove several important recent results. The initial breakthrough showing how one could begin to use Kolyvagin's method in our context is due to Flach (see **[Fla92]**), who found a way to construct something that can be thought of as the beginning of an Euler system applicable to our situation. Wiles called on all these ideas to construct a "geometric Euler system" which plays a central role in the argument. (*It is at this point that the current difficulty lies.*)

From the bound on the cohomology group one will get a proof that the two rings are in fact isomorphic. Translated back to the language of representations, this means that if one starts with a representation modulo p which satisfies Wiles' technical assumptions (and is modular), then any lift of the kind Wiles considers is also modular.

3.4 Put it all together.... Assume, then, that one can prove that all lifts of a modular representation are still modular. Now suppose we have an elliptic curve E whose representation modulo p we can prove (by some means) to be modular. Suppose also that this representation satisfies Wiles' technical assumptions. Then any lift of this representation is modular. But the p-adic representation $\rho_{\mathrm{E,p}}$ attached to E is one such lift! It follows that this representation is modular, and hence that E is modular.

All we need, now, is to prime the pump: we must find a way to decide that the representation modulo p is modular, and then use that to clinch the issue. What Wiles does is quite beautiful.

First of all, he takes a semistable elliptic curve, and looks at the Galois representation modulo 3 attached to this curve. At this point, there are two possibilities. The representation, as we pointed out above, amounts to an action of the Galois group on the vector space $\mathbb{F}_3 \times \mathbb{F}_3$. Now, it may happen that there is a subspace of that vector space which is invariant under every element of the Galois group. In that case, one says that the representation is *reducible*. If not, it is *irreducible*.

One has to be just a little more careful, Just as it sometimes happens that a real matrix has complex eigenvalues, it can happen that the invariant subspace only exists after we enlarge the base field. We will say a representation is *absolutely irreducible* when this does not happen: even over bigger fields, there is no invariant subspace.

Well, look at $\bar{\rho}_{E,3}$. It may or may not be absolutely irreducible. If it is, Wiles calls upon a famous theorem of J. Tunnell, based on work of R.P. Langlands (see

[Tun81], **[Lan80]**) to show that it is modular, and hence, using the deformation theory, that the curve is modular.

If $\bar{\rho}_{E,3}$ is not absolutely irreducible, Wiles shows that there is another elliptic curve which has the same representation modulo 5 as our initial curve, but whose representation modulo 3 *is* absolutely irreducible. By the first case, it is modular. Hence, its representation modulo 5 is modular. But since this is the same as the representation modulo 5 attached to our original curve, we can apply the deformation theory for $p = 5$ to conclude that our original curve is modular.

If Wiles' strategy is successful, we get:

Theorem 2. *The Shimura-Taniyama-Weil conjecture holds for any semistable elliptic curve.*

And, since the Frey curve is semistable,

Corollary 1. *For any $n \geq 3$, there are no non-zero integer solutions to the equation $x^n + y^n = z^n$.*

Of course, this is just *one* corollary of the proof of the STW conjecture for semistable curves, and it is certain that there will be many others still. For example, as Serre pointed out in **[Ser87b]**, one can apply Frey's ideas to many other diophantine equations that are just as hard to handle as Fermat's. These are equations that are closely related to the Fermat equation, of the form

$$x^p + y^p = Mz^p,$$

where p is a prime number and M is some integer. From Serre's argument and Wiles' result, one gets something like this:

Corollary 2. *Let p be a prime number, and let M be a power of one of the following primes:*

$$3, 5, 7, 11, 13, 17, 19, 23, 29, 53, 59.$$

Suppose that $p \geq 11$ and that p does not divide M. Then there are no nonzero integer solutions of the equation $x^p + y^p = Mz^p$.

The proof is precisely parallel to what we have done before: given a solution, construct a Frey curve, and consider the resulting modular form. Apply Ribet's theorem to lower its level, and then study the space of modular forms of that level to see if the form predicted by Ribet is there. If there is no such form, there can be no solution.

In fact, one can even get a general result, as Mazur pointed out:

Corollary 3. *Let M be a power of a prime number ℓ, and assume that ℓ is not of the form $2^n \pm 1$. Then there exists a constant C_ℓ such that the equation $x^p + y^p = Mz^p$ has no nonzero solutions for any $p \geq C_\ell$.*

However successful they may be in the end at proving the Shimura-Taniyama-Weil conjecture, Wiles' new ideas are certain to have enormous impact.

REFERENCES

[Ami75] Y. Amice, *Les nombres p-adiques*, Presses Universitaires de France, Paris, 1975.

[C^{+}] Henri Cohen et al., *GP-PARI*, a number-theoretic "calculator" and C library. Available by anonymous ftp from math.ucla.edu.

[Cas86] J. W. S. Cassels, *Local fields*, Cambridge University Press, Cambridge 1986.

[Cas91] J. W. S. Cassels, *Lectures on elliptic curves*, Cambridge University Press, Cambridge, 1991.

[Con] Ian Connell, *APECS*: *arithmetic of plane elliptic curves*, an add-on to *Maple*. Available by anonymous ftp from math.mcgill.edu.

[Cre92] J. E. Cremona, *Algorithms for modular elliptic curves*, Cambridge University Press, Cambridge, 1992.

[CS86] Gary Cornell and Joseph H. Silverman (eds.) *Arithmetic geometry*, Springer-Verlag, Berlin, Heidelberg, New York, 1986.

[Fla92] M. Flach, *A finiteness theorem for the symmetric square of an elliptic curve*, Invent. Math. 109 (1992), 307–327.

[Fre86] G. Frey, *Links between stable elliptic curves and certain diophantine equations*, Annales Univesitatis Saraviensis, Series math. 1 (1986), 1–40.

[Fre87a] G. Frey, *Links between elliptic curves and solutions of* $A - B = C$, J. Indian Math. Soc. 51 (1987), 117–145.

[Fre87b] G. Frey, *Links between solutions of* $A - B = C$ *and elliptic curves*, Number Theory, Ulm 1987 (H.P. Schlickewei and E. Wirsing, eds.) Lecture Notes in Mathematics, vol. 1380, Springer-Verlag, 1987.

[Gou93] Fernando Q. Gouvêa, *p-adic numbers*: *an introduction*, Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[Hus87] Dale Husemöller, *Elliptic curves*, Springer-Verlag, Berlin, Heidelbert, New York, 1987.

[Kna92] Anthony W. Knapp, *Elliptic curves*, Princeton University Press, Princeton, 1992.

[Kob84] N. Koblitz, *p-adic numbers*, *p-adic analysis, and zeta-functions*, second ed., Springer-Verlag, Berlin, Heidelberg, New York, 1984.

[Kol91] V. Kolyvagin, *Euler systems*, The Grothendieck Festschrift, vol. 2, Birkhäuser, 1991, pp. 435–483.

[Lan76] Serge Lang, *Introduction to modular forms*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[Lan80] R.P. Langlands, *Base change for* GL(2), Ann. of Math. Stud., vol. 96, Princeton University Press, Princeton, NJ, 1980.

[Lan91] Serge Lang, *Number theory III*, Encyclopedia of Mathematical Sciences, vol. 60, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[Maz89] Barry Mazur, *Deforming Galois representations*, Galois Groups Over $\mathbb{Q}$ (Y. Ihara, K. A. Ribet, and J.-P.Serre, eds.) Springer-Verlag, 1989.

[Maz91] Barry Mazur, *Number theory as gadfly*, American Mathematical Monthly 98 (1991), 593–610.

[Maz93] Barry Mazur, *On the passage from local to global in number theory*, Bull. Amer. Math. Soc 29 (1993), 14–50.

[Miy89] Toshitsune Miyake, *Modular forms*, Springer-Verlag, 1989.

[Rib79] Paulo Ribenboim, 13 *lectures on Fermat's Theorem*, Springer-Verlag, Berlin, Heidelberg, New York, 1979.

[Rib90] Kenneth A, Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ *arising from modular forms*, Invent. Math. 100 (1990), 431–476.

[Ser87a] Jean-Pierre Serre, *Lettre à J-F Mestre*, Current Trends in Arithmetical Algebraic Geometry (Kenneth A. Ribet, ed.), Contemporary Mathematics, vol. 67, American Mathematical Society, 1987.

[Ser87b] Jean-Pierre Serre, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Duke Math. J. 54 (1987), 179–230.

[Shi71] G. Shimura, *Introduction to the arithmetic theory of automorphic forms*, Princeton University Press, 1971.

[Sil86] Joseph H. Silverman, *The arithmetic of elliptic curves*, Springer-Verlag, Berlin, Heidelberg, New York, 1986.

[Sil93] Joseph H. Silverman, *Taxicabs and sums of two cubes*, American Mathematical Monthly 100 (1993), no. 4, 331–340.

[ST92] Joseph H. Silverman and John Tate, *Rational points on elliptic curves*, Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[SvM] J. H. Silverman and P. van Mulbregt, *EllipticCurveCalc, a Mathematica* package. Available by anonymous ftp from gauss.math.brown.edu; contact jhs@gauss.math.brown.edu for information.

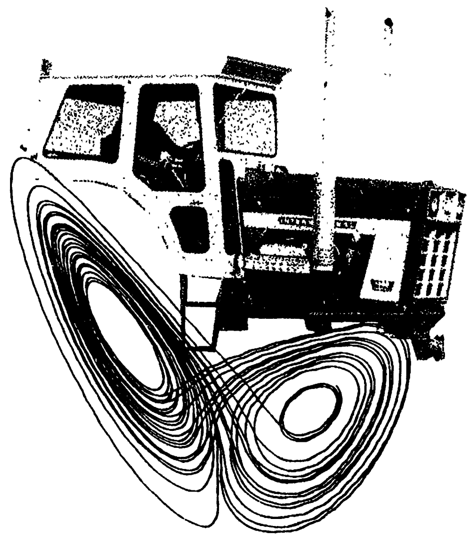
[Tun81] J. Tunnell, *Artin's conjecture for representations of octahedral type*, Bull. Amer. Math. Soc. (N. S.) **5** (1981), 173–175.

[Was82] Larry C. Washington, *Introduction to cyclotomic fields*, Springer-Verlag, Berlin, Heidelberg, New York, 1982.

[Wei83] André Weil, *Number theory: an approach through history, from Hammurapi to Legendre*, Birkhäuser, 1983.

[Z$^+$] H. G. Zimmer et al., *SIMATH*, a computer algebra system with main focus on algebraic number theory. Contact simath@math.uni-sb.de for more information.

Colby College
Department of Mathematics and Computer Science
Waterville, ME 04901
fqgouvea@colby.edu



A Strange Tractor

Submitted by Alberto Guzman
Department of Mathematics
The City College of CUNY
New York, NY 10031