Report on Fermat's Last Theorem

V. Kumar Murty

Fermat's Last Theorem is the assertion that for any integer $n > 2$ there are no non-zero integers x, y, z such that $x^n + y^n = z^n$. Attempts to prove it over the last three and a half centuries have resulted in the formulation of many deep concepts, tools and techniques. Recent developments on Fermat's Last Theorem have brought these concepts together in a spectacular fashion. In this article, we attempt to describe some of these ideas in general terms, so as to convey a flavour of the subject.

A question studied by the ancients is to find all solutions in integers of the equation

$$X^2 + Y^2 = Z^2. \tag{1}$$

For example (3, 4, 5), (5, 12, 13) are solutions. One way of producing infinitely many solutions is to take $y = 0$ and $x = z$. But we shall consider these to be degenerate and the problem is to find all *non-degenerate* solutions (that is, triples of integers (x, y, z) satisfying (1) and with $xyz \neq 0$). At first sight, this seems to be a problem in number theory, but it is most easily solved from a geometric point of view. In the first place, it is enough to find all solutions of

$$u^2 + v^2 = 1 \tag{2}$$

with u, v rational numbers, for then we could clear the denominator to get an integral solution of (1). Equation (2) defines the unit circle. Given a point (u_0, v_0) on it, consider the line through (u_0, v_0) and (0, 1) and let t_0 be the point where it intersects the u-axis.

By writing the equation of the line, it is easy to check that u_0 and v_0 are rational numbers if and only if t_0 is rational. Thus, the rational solutions of (2) are in one-to-one correspondence with the rational values of t and in fact are given by

$$u = \frac{2t}{t^2+1} \quad v = \frac{t^2-1}{t^2+1},$$

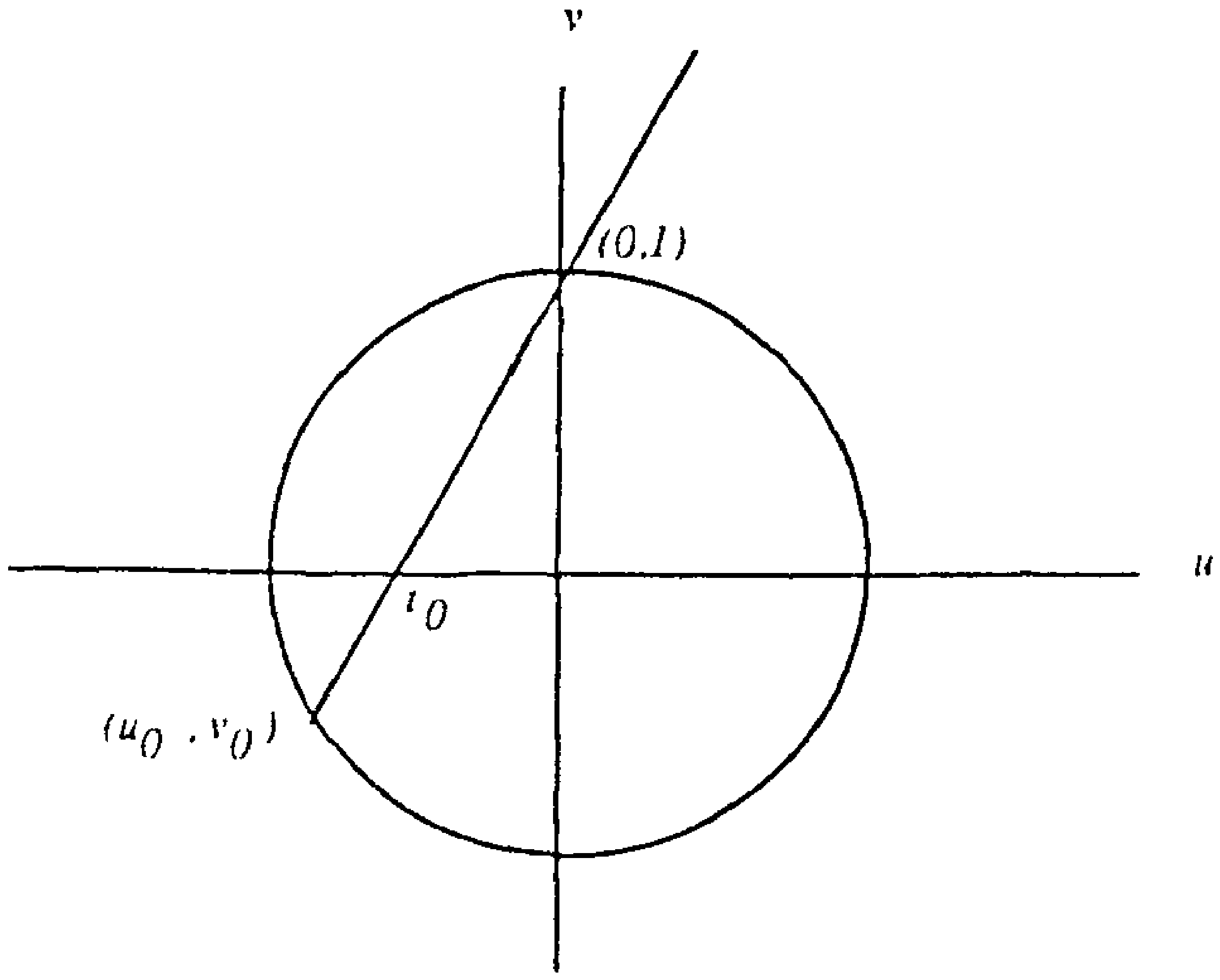
where t is rational.

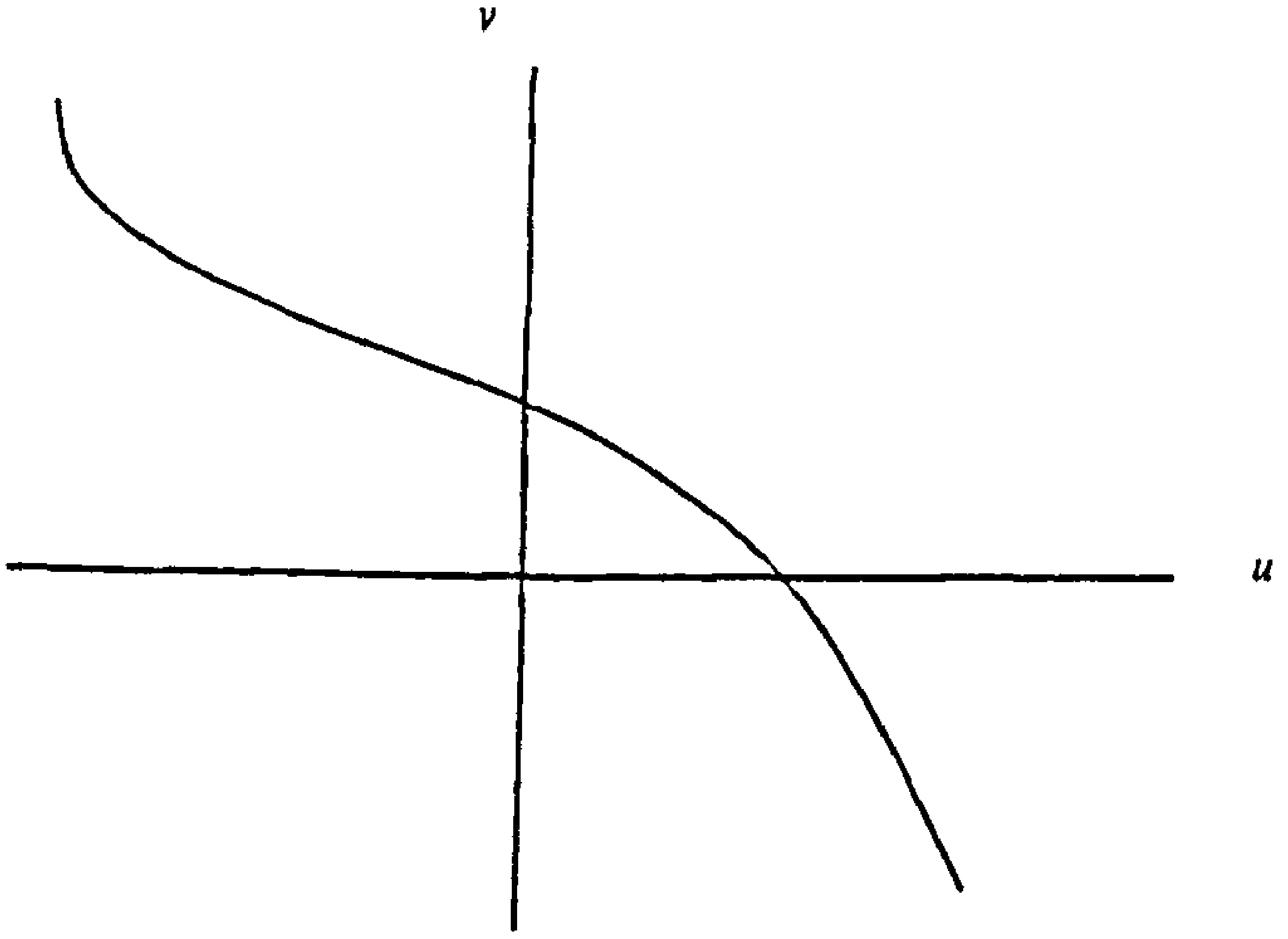
The saga of Fermat's last theorem begins with the consideration of the cubic analogue of (1). The problem is to find all solutions in integers of

$$X^3 + Y^3 = Z^3. \tag{3}$$

Once again, we are looking for non-degenerate solutions. If we attempt to mimic the approach used above, we are led to consider rational points on the cubic curve

$$u^3 + v = 1. \tag{4}$$





Since a line through two points on the curve will intersect the curve at a third point, we cannot expect to

V Kumar Murty is in the Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A1 parametrize the points on the curve using points on a fixed line. To solve (4) (or equivalently (3)) requires more arithmetic insight.

It was Gauss who first showed that (3) has no non-degenerate solutions. In essence, the idea is to factorize the left-hand side. Let $\zeta = \exp(2\pi i/3)$ be a third root of unity. Then (3) is the same as

$$(X+Y)(X+\zeta Y)(X+\zeta^2 Y) = Z^3. \qquad (5)$$

If ζ had been an ordinary integer, we might have proceeded as follows: The three factors on the left are essentially relatively prime (that is, have no common divisors) and so if their product is a cube, then each factor itself must be a cube. This follows from the unique factorization property of the integers. Gauss developed the 'arithmetic' of the field $\mathbb{Q}(\zeta)$ to an extent where this calculation still makes sense and thus, he could conclude that a solution (x, y, z) of (5) satisfies

$$x+y=u^3$$

$$x+\zeta y=v^3$$

$$x+\zeta^2 y=w^3.$$

Now rearranging these equations and using the fact that $1+\zeta+\zeta^2=0$, one deduces that there is a new solution

$$a^3+b^3=c^3$$

with $\max(|a|, |b|, |c|) < \max(|x|, |y|, |z|)$. Moreover, if we begin with a non-degenerate solution (x, y, z), we again get a non-degenerate solution (a, b, c). Thus, repeating this procedure eventually produces a contradiction and shows that (3) has no non-degenerate solution.

It is interesting to note that if we alter the equation slightly to

$$X^3+Y^3+Z^3+W^3=0$$

then we do find non-trivial solutions. Indeed $3^3+4^3+5^3+(-6)^3=0$. Another solution was discovered by Ramanujan, namely (12, 1, –10, –9). Ramanujan observed that this solution gives the smallest number which can be expressed as a sum of two cubes in two different ways, namely $1729 = 12^3 + 1^3 = 10^3 + 9^3$. Later he found infinitely many solutions. In 1988, Elkies and Zagier independently discovered a solution of $X^4+Y^4+Z^4-W^4=0$, namely (2682440, 15365639, 18796760, 20615673).

In 1637, the lawyer and amateur mathematician Pierre de Fermat obtained a copy of a translation of the Greek work by Diophantus. He read the description of the solutions of (1). Fermat asked himself the question whether

$$X^n+Y^n=Z^n \qquad (6)$$

has any non-degenerate solutions. Apparently, he convinced himself that for $n > 2$ it did not and he made a marginal note to this effect. This part of the story is too well-documented to warrant elaboration here.

After Fermat's death, his marginal note was discovered and when attempts at reconstructing his marvelous proof failed, the problem acquired some notoriety. Fermat made many assertions without proof. Some of them have subsequently been proved correct and others have been shown to be false. This marginal note is the only (known) remaining assertion that has yet to be proved or disproved, hence the epithet Fermat's last theorem.

It is, at least at first glance, rather surprising that the problem should have attracted as much attention and interest as it has from the public, both mathematical and non-mathematical. It is important to consider this phenomenon as it bears on the very foundation of scientific discovery and on the public perception of science.

The word 'science' comes from the Latin verb 'scindere' which means to dissect or to take apart for the purpose of analysis. It refers, therefore, not so much to a body of knowledge as to an approach to knowledge itself. By observation of phenomena, science seeks to discover the principle which underlies them.

The use of the word discovery is deliberate: it is indicative of a feeling shared by many thinkers that there is a hidden structure or underlying harmony in nature. We do not invent it; we can only try to reveal it and describe it. Philosophy goes further and speaks of a 'harmony of harmonies' that is so all-encompassing that it is beyond description. But science deals with the describable, and mathematics is the language of science.

As we make language more precise, the field of ideas which can be described in that language becomes narrower. Science has chosen precision and accepted this limitation of field, and the most precise of languages is mathematics. However, even here, we should not forget that it is not a machine that does science or mathematics, but the human being. The infinite vista of the human mind peers through from behind the most immaculate and imposing mathematical formula. The finite and known always points to the infinite and the unknown and in this sense, knowledge is infinite. A formula may be the end of one discovery, but it is also the beginning of a new investigation. It is this that makes mathematical discovery possible.

Mathematical discoveries are the discoveries of new concepts and their relationships. How does one go about making such discoveries? 'There is no royal road to geometry.' The only way is by patient and careful study. Even Ramanujan worked through 5000 problems to awaken his skill and insight.

A mathematical mind is awake to problems which might suggest themselves. Thus it is not unreasonable or very surprising that Fermat should have asked himself about (6) after reading a discussion of (1).

In June 1993, newspapers around the world flashed the headline that the 350-year old Fermat problem had been solved. Professionals and amateurs alike were glued to their computers to read the latest postings on the electronic mail networks. And perhaps somewhat surprisingly, the 'man on the street' wanted to know the latest on Fermat. Why?

Partly, it is an awe of technological and scientific achievements, an awe that has been instilled in us by the astounding and apparently miraculous successes in these fields. But the interest shown raises the serious question of how science and mathematics are perceived by the public. To many, these subjects are seen as lying within some impenetrable fortress or in an ivory tower, surrounded by an air of incomprehensible mystery. The people within this fortress are viewed as living in a world of their own, speaking to each other in some strange language and out of touch with the reality of the world outside.

Scientists themselves have contributed to this view by failing to reach out to the public. This state of affairs is unfortunate and even dangerous, as any communication gap is dangerous. The great library of Alexandria, the Institute for Advanced Study of ancient times, was burnt to the ground by a frenzied crowd unable to understand the work being done there and enraged at the apparently privileged life of the scholars. The health of scientific activity depends on our bridging the gulf of understanding to the public.

Both technology and fundamental research serve society and play an important role in it. Moreover they are human activities and enrich the human experience. As music can be appreciated even by those who are not musicians, so can mathematics be appreciated. If science is portrayed as the artistic and creative experience which it is, it will be more accessible to the public.

To return to our narrative, Kummer tried to approach Fermat's last theorem by generalizing the approach of Gauss of $n = 3$. One observes in the first place that it suffices to consider the case when $n = p$ is prime. Kummer developed the arithmetic of the field $\mathbb{Q}(\zeta_p)$ where ζ_p is a primitive pth root of 1. Assuming that unique factorization held in it, he could show that

$$X^p + Y^p = Z^p \qquad (7)$$

has no non-degenerate solutions by following the same line of attack as in the case $p = 3$.

Unfortunately, unique factorization does *not* hold in this field in general. One can measure how far it deviates by a group called the ideal class group. A fundamental theorem of algebraic number theory is that the class group is finite. Let us denote its order by h_p. What Kummer proved is that if $h_p = 1$ (or even if p does not divide h_p) then (7) has no non-degenerate solutions. The condition $h_p = 1$ can be interpreted Galois-theoretically using class field theory.

Denote by $\overline{\mathbb{Q}}$ an algebraic closure of $\mathbb{Q}$. Then, it is equivalent to asserting that any homomorphism

$$\chi : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}(\zeta_p)) \to \mathbb{C}^*, X \neq 1$$

is ramified somewhere. Equivalently, any extension field $L/\mathbb{Q}(\zeta_p)$ is ramified somewhere. Summarizing, Kummer showed that if (7) has a non-degenerate solution, then there exist a χ and an L which are unramified everywhere.

Since it *is* possible to construct such χ and L, Kummer's approach did not prove that (7) has no non-degenerate solutions. Many years later, Frey re-examined the problem and asked whether it would be possible to construct two-dimensional Galois representations with restricted ramification starting from a solution to (7). He made two modifications: he considered representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and he replaced $\mathbb{C}$ with a finite field. Thus, he was looking for homomorphisms

$$\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Z}/p).$$

As soon as this is written, it at once suggests elliptic curves.

What is an elliptic curve? Let us return to Diophantus and the equation (1). The procedure we used to solve it will work if we replace the circle with any conic section. So in the program of 'solving all diophantine equations', that is, all equations of the form

$$F(X_1, \ldots, X_n) = 0$$

where F is a polynomial with rational coefficients in n-variables, the next step is curves which are not conics and the 'simplest' of these are elliptic curves.

Topologically, if rational curves look like a plane, elliptic curves are tori, that is, doughnut-shaped. More specifically, over the complex numbers $\mathbb{C}$, an elliptic curve is given by $\mathbb{C}/L$ where $L = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$ is a lattice in $\mathbb{C}$. Algebraically, an elliptic curve over say a field K of characteristic zero can be given by an equation

$$E : y^2 = x^3 + ax + b, \qquad a, b \in K.$$

where the cubic on the right has distinct roots.

The connection between the two descriptions is given by the Weierstrass $\mathcal{P}$-function. Indeed, given the lattice L, let us set

$$\mathcal{P}(z) = \frac{1}{z^2} + \sum_{\omega \in L\setminus\{0\}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right).$$

Then $\mathcal{P}$ satisfies $\mathcal{P}(z+\omega) = \mathcal{P}(z)$ for any $\omega \in L$. Moreover, we have the algebraic relation

$$(\mathcal{P}')^2 = 4\mathcal{P}^3 - g_2\mathcal{P} - g_3,$$

where $g_2 = 60\ \Sigma\omega^{-4}$ and $g_3 = 140\ \Sigma\omega^{-6}$ where the summation is over $\omega \in L\backslash\{0\}$.

The curve E has several numerical invariants attached to it. The first is the discriminant

$$\Delta = \Delta_E = 16(4a^3 - 27b^2) \neq 0.$$

If e_1, e_2, e_3 are the roots of the cubic,

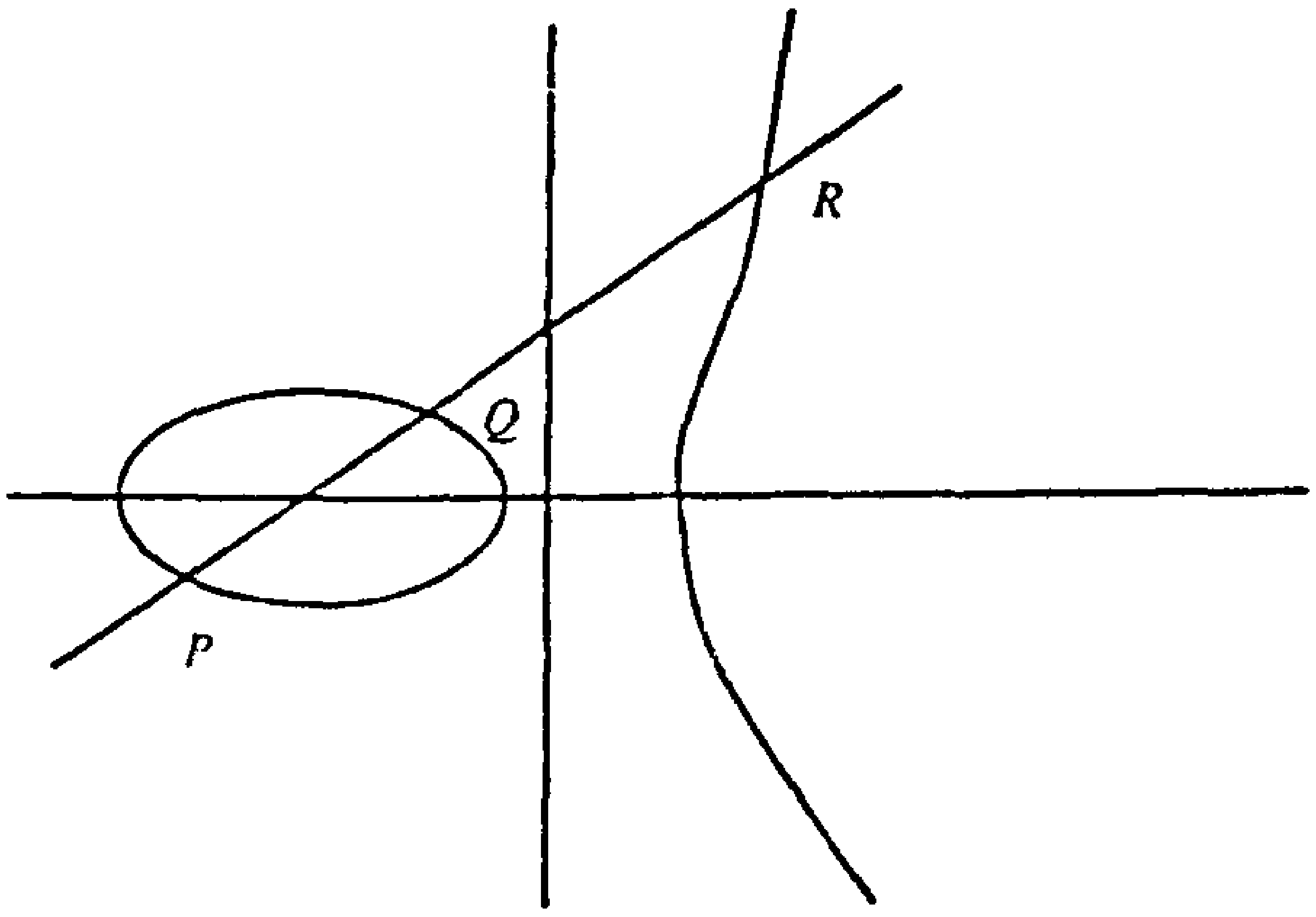
$$\Delta = (e_1 - e_2)^2 (e_1 - e_3)^2 (e_2 - \grave{e}_3)^2.$$

There is also the j-invariant

$$j = j_E = {}^*a^3/\Delta,$$

where * is an explicit, but (for us) unimportant, constant. The value of j_E is unchanged if E is replaced by an isomorphic curve, whereas Δ_E depends on the choice of model. Besides j and Δ, there is a subtler invariant $N = N_E$ called the conductor of E. It is more technical to define.

One new structure that appears in the case of elliptic curves is that it is possible to define a group law $\oplus$ on the points.



The identity $\mathcal{O}$ is the point at infinity. The group law is determined by requiring that

$$P \oplus Q \oplus R = \mathcal{O}$$

if the three points P, Q, R lie on a straight line. If we are given the coordinates of P and Q, it is clear that we can determine the coordinates of R using rational functions.

Now suppose that p is an odd prime, $A, B, C \in \mathbb{Z}$, $ABC \neq 0$, $(A, B, C) = 1$ and that

$$A^p + B^p = C^p$$

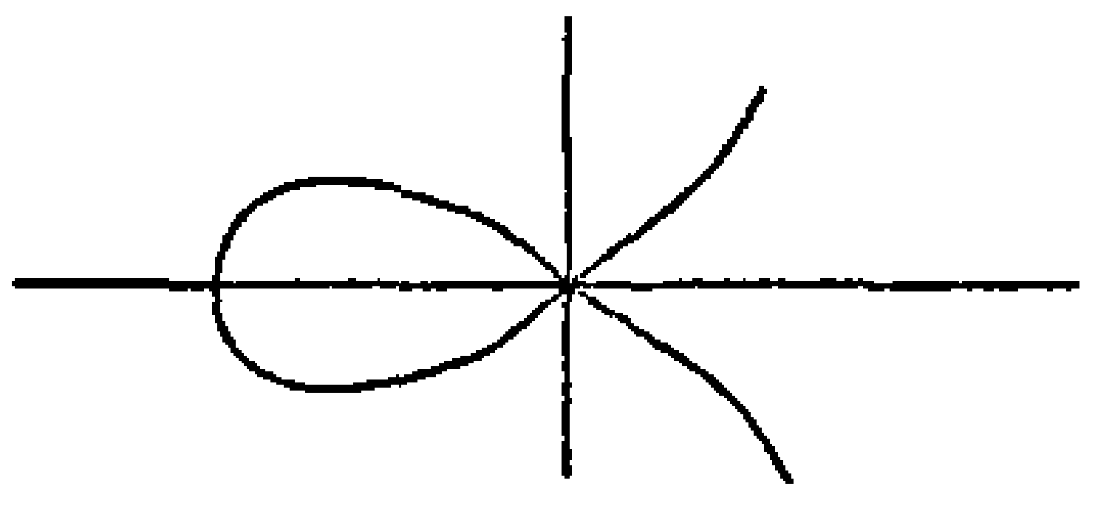
In other words, suppose we had a non-degenerate solution (A, B, C) to the Fermat equation of degree p. Consider

$$E : y^2 = x(x + A^p)(x - B^p).$$

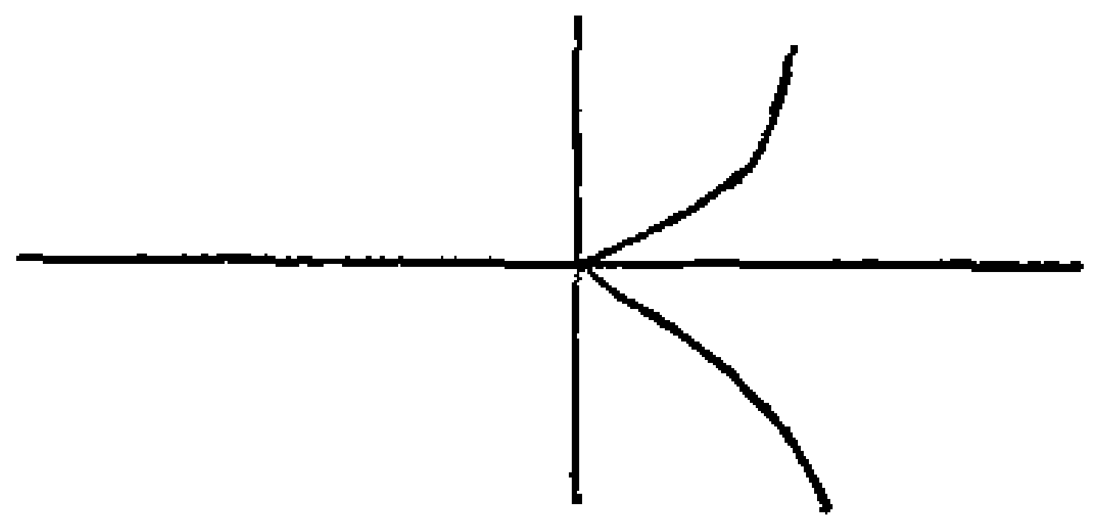
This is not quite in the (Weierstrass) form described above, but can easily be put in that form. When one does so, one finds that

$$\Delta_E = (ABC)^{2p}.$$

The curve E is *semistable* in the following sense. Consider the curves



$E_0 . y^2 = x^2(x+1)$



$E_1 \cdot y^2 = x^3$

These are not elliptic curves, but degenerate versions where two or three of the roots of the cubic coincide (so $\Delta = 0$). The first is called a nodal curve and the second a cuspidal curve.

In number theory, it is often easier to study equations modulo a varying prime l. Given E we may consider it modulo l (that is, view it as a curve defined over the finite field $\mathbb{Z}/l$). If 0, A^p, $-B^p$ stay distinct modulo l, then we still get an elliptic curve. But if l divides A^p or B^p or $A^p + B^p$ then we get a nodal or cuspidal curve. In fact, we cannot get a cuspidal curve for that would require l to divide A, B and C and we had assumed (without loss of generality) that $(A, B, C) = 1$. This is the meaning of E being semistable: for *any* prime l the reduction of E modulo l is either an elliptic curve or a nodal curve.

Knowing that an elliptic curve has semistable reduction everywhere has Galois-theoretic implications. So let E be any such curve defined over $\mathbb{Q}$. For a prime l, we consider

$$E[l] = \{(x, y) \in E, x, y \in \mathbb{C} : l \cdot (x, y) = 0\}$$

(Here

$$l \cdot (x,y) = \underbrace{(x,y) \oplus \cdots \oplus (x,y)}_{l \text{ times}}.).$$

The set $E[l]$ is an Abelian group and in fact $E[l] \cong \mathbb{Z}/l \oplus \mathbb{Z}/l$. Moreover x, y must lie in $\overline{\mathbb{Q}}$, and if $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, then $(\sigma(x), \sigma(y))$ is again in $E[l]$. Moreover, if (x, y), $(u, v) \in E[l]$ then $(s, t) = (x, y) \oplus (u, v) \in E[l]$ and $(\sigma(s), \sigma(t)) = (\sigma(x), \sigma(y)) \oplus (\sigma(u), \sigma(v))$. Thus, we get a representation

$$\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Aut}(E[l]) \cong \mathrm{GL}_2(\mathbb{Z}/l).$$

This homomorphism factors through a finite quotient, so in fact we get an extension field K_l (say) of finite degree over $\mathbb{Q}$. This field is obtained by adjoining to $\mathbb{Q}$ the coordinates of all points in $E[l]$ and it is the fixed field of the kernel of ρ. A theorem of Tate tells us that if E is semistable, then $K/\mathbb{Q}$ is unramified at any odd prime $q \neq l$ with $\min(v_q(j_E), 0) \equiv 0 \pmod{l}$. Applying it to the elliptic curve we got from a solution to Fermat, one deduces the following: K_p is unramified at all primes $l \neq 2, p$. Moreover, $\min(v_p(j_l), 0) \equiv 0 \pmod{p}$. Thus we have achieved our initial objective. Starting from a solution to Fermat, we have produced an extension field which has limited ramification and we obtained this extension from points of finite order on a certain elliptic curve.

We remark that for a semistable curve E, we have for any prime l

$$v_l(\Delta_E) = -\min(v_l(j_E), 0)$$

and so the above condition may be written as

$$v_p(\Delta_E) \equiv 0 \pmod{p}.$$

Moreover, the conductor N_E is given by

$$N_l = \prod_{v_l(\Delta_l)>0} l.$$

Using this, one can deduce a kind of converse to the construction of E. Indeed, Frey showed that for $p \geq 5$, the following are *equivalent*:

(a) there is a non-degenerate solution of $X^p + Y^p = Z^p$

(b) there is a semistable elliptic curve E defined over $\mathbb{Q}$ such that

(i) $\mathbb{Q}(E[2]) = \mathbb{Q}$

(ii) $\mathbb{Q}(E[p])$ is unramified outside $2p$

(iii) $v_l(\Delta_E) \equiv 0 \pmod{p}$, $v_2(\Delta_E) \equiv 8 \pmod{p}$.

Frey's result (which built on earlier work of Hellegouarch) was a very significant turning point. For the first time, it showed clearly that Fermat's last theorem was *equivalent* to a problem in the theory of elliptic curves. Now the search was on for a method to show that curves satisfying (b) do not exist.

One way of doing this is to find a list of *all* elliptic curves over $\mathbb{Q}$ and check that the Frey curves are not there. So next, one looks for a parametrization of all elliptic curves over $\mathbb{Q}$. Taniyama found some elliptic curves 'in nature'. Start with the upper half plane

$$\mathfrak{H} = \{z \in \mathbb{C}: \operatorname{Im} z > 0\}.$$

$\mathrm{SL}_2(\mathbb{R})$ acts on $\mathfrak{H}$ by fractional linear transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}.$$

(Here $\mathrm{SL}_2(\mathbb{R})$ denotes 2×2 matrices with real entries and having determinant equal to 1.) To each elliptic curve E over the complex numbers, we can associate a point $z \in \mathfrak{H}$. Indeed, write $E = \mathbb{C}/L$ with $L = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$, and $\operatorname{Im}(\omega_1/\omega_2) > 0$. Then we map

$$E \mapsto \omega_1/\omega_2.$$

If E_1 and E_2 are isomorphic and correspond to $z_1, z_2 \in \mathfrak{H}$, then there is an element $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $\gamma z_1 = z_2$. So in a natural way, the quotient (or orbit) space $\mathrm{SL}_2(\mathbb{Z})\backslash\mathfrak{H}$ parametrizes isomorphism classes of elliptic curves.

We may also consider subgroups

$$\Gamma_0(N) = \left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix} = g \in \mathrm{SL}_2(\mathbb{Z}) : g \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \bmod N\right\}.$$

Then the orbit space $\Gamma_0(N)\backslash\mathfrak{H}$ can be shown to be a parameter space for isomorphism classes of pairs (E, C_N), E an elliptic curve and C_N a cyclic subgroup of order N. So when we are interested in 'enumerating' elliptic curves, it is natural to consider these parameter spaces.

These spaces are open Riemann surfaces and they can be compactified by adding a finite number of points. It is a fundamental result of Shimura that the resulting compact Riemann surfaces are algebraic curves having a model (that is, a defining equation), denoted $X_0(N)$, which is defined over $\mathbb{Q}$.

These curves have many interesting properties. In particular, they have many 'correspondences' which can be used to decompose the Jacobian variety

$$\mathrm{Jac}\,(X_0(N)) \sim A_1 \times \cdots \times A_r.$$

The A_i are Abelian varieties (as is $\mathrm{Jac}\,(X_0(N))$) and $\sim$ denotes isogeny. (An *Abelian variety* is a higher dimensional generalization of an elliptic curve. If A_1 and A_2 are Abelian varieties, a morphism $\phi: A_1 \to A_2$ is called an *isogeny* if $\ker \phi$ is finite.) If $\dim A_i = 1$, we say that it is a *modular elliptic curve*. As I said, they appear in nature. The 'smallest' example is $X_0(11)$ which is itself an elliptic curve. It is given by the equation $y^2 + y = x^3 - x^2$. (The problem of finding equations for modular curves is an old one and it is continuing to attract attention even now.)

There are several equivalent ways of defining a modular elliptic curve. One is to say that E is modular if for some N there is a non-constant map

$$X_0(N) \to E.$$

The conjecture of Taniyama-Shimura is that *any* elliptic curve defined over $\mathbb{Q}$ is modular.

On the one hand, given a modular elliptic curve E, we can choose a prime p and associate a Galois representation

$$\rho: \mathrm{Gal}\,(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Aut}\, E[p] \cong \mathrm{GL}_2(\mathbb{Z}/p).$$

On the other hand, given such a representation ρ, we can try to see if it comes from a modular elliptic curve. A given ρ may not arise from an elliptic curve, or it may arise from many different elliptic curves i.e. ρ does *not* determine E.

There are several pieces of data involved here:

(a) the ramification of ρ: as we saw, if ρ comes from a solution to the Fermat equation of degree p, it has very restricted ramification properties, namely it is unramified outside $2p$ and $v_p(\Delta_F) \equiv 0 \pmod p$.
(b) if ρ arises from an elliptic curve E, we can consider its conductor N. If E is modular, it is a result of Carayol that E is of level N, i.e. occurs in $\mathrm{Jac}(X_0(N))$.

A fundamental conjecture of Serre asserts a relationship between these two pieces of information. Namely, if E is modular of level lM, $l \nmid M$ and ρ is 'finite' at l, then there is another modular elliptic curve E' (or more generally, a modular form f) of level M which also gives rise to ρ. The condition of being 'finite' at l is satisfied if ρ is unramified at l. For semistable curves, it is also satisfied if $v_l(\Delta_F) \equiv 0 \pmod p$.

The Taniyama-Shimura conjecture and Serre's conjecture together imply Fermat's Last Theorem. Indeed, starting with a solution (a, b, c) to F_p, construct E and ρ. By Taniyama-Shimura, E is modular. We know that E has conductor

$$N = \prod_{l \mid abc} l$$

as E is semistable. Note that N is squarefree (that is, it is not divisible by the square of any prime). By Carayol's result, E occurs in $\mathrm{Jac}(X_0(N))$. Since ρ is 'finite' at any $l \mid N$, $l \neq 2$, Serre's conjecture implies that there must be a modular form f of level 2 giving rise to ρ. But $\mathrm{Jac}(X_0(2)) = 0$ (as $X_0(2)$ has genus zero) so f does not exist.

In 1987, K. Ribet proved Serre's conjecture thereby showing that the Taniyama-Shimura conjecture implies Fermat's last theorem. The work announced by Wiles in June 1993 claimed to prove the Taniyama-Shimura conjecture for all semistable elliptic curves. By our discussion above, this would be enough to imply Fermat's last theorem. However at present, there seems to be a gap in the argument. This was spotted by one of the referees selected to study the manuscript. Wiles' manuscript was never released to the mathematical public and was only made available to a small group of experts. The e-mail networks,,however, were kept buzzing with every small piece of information that leaked out and it was especially news of the 'gap' which provided most of the grist for the rumour mills.

To clarify the situation, on 4 December 1993, Wiles posted the following message on the network sci.math

> In view of the speculation on the status of my work on the Taniyama-Shimura conjecture and Fermat's Last Theorem I will give a brief account of the situation During the review process a number of problems emerged, most of which have been resolved, but one in particular I have not yet settled. The key reduction of (most cases of) the Taniyama-Shimura conjecture to the calculation of the Selmer group is correct. However, the final calculation of a precise upper bound for the Selmer group in the semistable case (of the symmetric square representation associated to a modular form) is not yet complete as it stands. I believe that I will be able to finish this in the near future using the ideas explained in my Cambridge lectures.
>
> The fact that a lot of work remains to be done on the manuscript makes it still unsuitable for release as a preprint. In my course in Princeton beginning in February I will give a full account of this work.
>
> *Andrew Wiles*

In fact Wiles has now begun lecturing on his work. It should be pointed out that there seems to be agreement that Wiles has proved infinitely many new cases of the Taniyama-Shimura conjecture.

We shall close this article by briefly discussing two of the key techniques in Wiles' argument, namely Selmer groups and deformations of Galois representations.

Given as elliptic curve E defined over $\mathbb{Q}$, and a prime l, we have already seen that there is an associated representation

$$\bar{\rho}.\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Aut}\, E[l] \cong \mathrm{GL}_2(\mathbb{Z}/l).$$

In fact, this is the reduction mod l of a representation

$$\rho\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Z}_l),$$

where $\mathbb{Z}_l$ denotes the ring of l-adic integers. Indeed, this representation arises by considering not just $E[l]$, but simultaneously all $E[l^n]$, $n = 1, 2, \ldots$. Though $\bar{\rho}$ does not determine E, it is a theorem of Faltings that ρ does (at least up to isogeny).

On the other hand, we may consider *all* representations into a local ring $\mathcal{O}$ of residue field $\mathbb{Z}/l$

$$\rho.\ \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathcal{O})$$

whose reduction mod l is isomorphic to $\bar{\rho}$. Such a ρ is called a *deformation* of $\bar{\rho}$ and the problem of classifying the deformations of a given $\bar{\rho}$ has been studied extensively by Mazur and others. Actually, to make the problem meaningful, further conditions have to be imposed on $\bar{\rho}$ and ρ.

One way of constructing representations ρ is from modular forms. A *modular deformation* of $\bar{\rho}$ is a deformation ρ which arises from a modular form. A conjecture of Mazur and Fontaine asserts that under certain assumptions on $\bar{\rho}$ (having to do with ramification and behaviour on certain decomposition groups) every deformation is modular. The bulk of Wiles' argument is to prove that if there exists one modular deformation, then all deformations are modular, that is, the Mazur-Fontaine conjecture holds.

Let us briefly consider Wiles' approach to proving this result. Mazur proved that there exists a ring R and a universal representation

$$\rho\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(R)$$

such that any deformation ρ arises by composing ρ with a morphism $R \to \mathcal{O}$. There is also a ring T and a universal modular representation (that is a representation λ such that any modular representation factors through λ). There is a surjective morphism $\varphi\colon R \to T$ and we want to prove that $R = T$. Wiles reduces this to a local computation. We are given one modular lift so that it corresponds to a morphism $\pi\colon T \to \mathbb{Z}_l$. Let $\mathcal{P}_T$ denote its kernel. Let $\mathcal{P}_R$ denote the kernel of the composition $R \to T \to \mathbb{Z}_l$. The cotangent space to the deformation space at the point given by π is $\mathcal{P}_R / \mathcal{P}_R^2$ and in the modular deformation space it is $\mathcal{P}_T/\mathcal{P}_T^2$. The map $\varphi\colon R \to T$ gives a map

$$\overline{\varphi}: \mathcal{P}_R / \mathcal{P}_R^2 \to \mathcal{P}_T/\mathcal{P}_T^2.$$

Wiles uses commutative algebra to show that if T is a local complete intersection and $\overline{\varphi}$ is an isomorphism then $R \cong T$.

The ring T is shown to be Gorenstein which means that

$$T \cong \mathrm{Hom}_{\mathbb{Z}_l}(T, \mathbb{Z}_l).$$

Thus, associated to $\pi\colon T \to \mathbb{Z}_l$ there is an element $\eta \in \mathbb{Z}_l$ so that the composition of the natural map $\mathbb{Z}_l \to T$ with π is multiplication by η. Wiles uses Fitting ideals to show that

$$|\mathcal{P}_T/\mathcal{P}_T^2| \geq |\mathbb{Z}_l/\eta|.$$

Hence, the problem is to show that T is a local complete intersection and that

$$|\mathcal{P}_R / \mathcal{P}_R^2| \leq |\mathbb{Z}_l / \eta|.$$

To do this, he interprets $\mathcal{P}_R / \mathcal{P}_R^2$ as an analogue of a Selmer group. In recent years, work of Kolyvagin has shed new light on the problem of estimating the size of components of the Selmer group. Wiles tries to adapt these methods and recent work of M. Flach to prove the required bound. To do this, he has to construct a certain distinguished family of cohomology classes (called an Euler system) in a dual Selmer group. It is in this construction that the above mentioned gap occurs.

Assuming that this gap can be filled, one still needs one modular lifting in order to be able to apply the theorem. An amazing aspect of Wiles' theory is that one needs a modular lift for only one prime l. Wiles chooses $l = 3$ where the sought-for modular lift was proved to exist by a theorem of Langlands and Tunnell.

It is hoped that even this brief description of Wiles' work conveys a sense of the grand unification of ideas that it embodies.

General reference on algebraic number theory

K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, Springer-Verlag, 1982.

Kummer's approach to Fermat's Last Theorem

P. Ribenboim, *13 Lectures on Fermat's Last Theorem*, Springer-Verlag, 1979.

Elliptic curves

J. Silverman, *The Arithmetic of Elliptic Curves*, Springer-Verlag, 1986.

S. Lang, *Elliptic Functions*, Addison-Wesley, 1973.

Modular curves and Abelian varieties

V. Kumar Murty, *Introduction to Abelian Varieties*, CRM Monograph Series #3, Am Math Soc., 1993.

Work of Frey, Ribet and Serre

G. Frey, Links between stable elliptic curves and certain Diophantine equations, *Annales Univ. Saraviensis*, **1** (1986), 1–40.

K. Ribet, On modular representations of Gal $(\overline{\mathbb{Q}}/\mathbb{Q})$ arising from modular forms, *Invent. Math.*, **100** (1990), 431–476.

J.-P. Serre, Sur les représentations modulaires de degré 2 de Gal $(\overline{\mathbb{Q}}/\mathbb{Q})$, *Duke Math. J.*, **54** (1987), 179–230.

Euler systems

B. Gross, Kolyvagin's work on modular elliptic curves, in *L-Functions Arithmetic* (Eds. J. Coates and M. Taylor), LMS Lecture Notes **153** (1991), 235–256.

V. Kolyvagin, Euler systems, *in The Grothendieck Festschrift* (Ed. P. Cartier *et al*), vol. 2, pp. 435–483, Birkhauser, 1990.

K. Rubin, Kolyvagin's system of Gauss sums, in *Arithmetic Algebraic Geometry* (Ed. G. van der Geer *et al.*), *Progress in Mathematics*, Birkhauser, 1991, vol. 29, pp. 309–324

Deformations of Galois representations

B Mazur, Deforming Galois representations in *Galois groups over* $\mathbb{Q}$ (Eds. Y. Ihara, K. Ribet and J.-P. Serre), MSRI Publications, 1989, vol. 16, Springer-Verlag, pp. 385–437.

Work of Langlands and Tunnell

R. P. Langlands, Base Change for GL(2), *Ann. Math*, Studies 96, Princeton University Press, 1980.

J. Tunnell, Artin's conjecture for representations of octahedral type, *Bull. Am. Math. Soc.*, **5** (1981), 173–175.

ACKNOWLEDGEMENT. I thank Prof S. Ramaseshan for the invitation to write this article and NSERC for partial research support.

Received 7 March 1994; accepted 27 April 1994