An Overview of the Proof of Fermat's Last Theorem

Glenn Stevens

The principal aim of this article is to sketch the proof of the following famous assertion.

Fermat's Last Theorem. *For $n > 2$, we have*

$$\mathbf{FLT}(n): \quad \left.\begin{array}{r} a^n + b^n = c^n \\ a, b, c \in \mathbf{Z} \end{array}\right\} \Longrightarrow abc = 0.$$

Many special cases of Fermat's Last Theorem were proved from the 17th through the 19th centuries. The first known case is due to Fermat himself, who proved FLT(4) around 1640. FLT(3) was proved by Euler between 1758 and 1770. Since FLT(d) $\Longrightarrow$ FLT(n) whenever $d|n$, the results of Euler and Fermat immediately reduce our theorem to the following assertion.

Theorem. *If $p \geq 5$ is prime, and $a, b, c \in \mathbf{Z}$, then $a^p + b^p + c^p = 0 \Longrightarrow abc = 0$.*

The proof of this theorem is the result of the combined efforts of innumerable mathematicians who have worked over the last century (and more!) to develop a rich and powerful arithmetic theory of elliptic curves, modular forms, and galois representations. It seems appropriate to emphasize the names of five individuals who had the insight to see how this theory could be used to prove Fermat's Last Theorem and to supply the final crucial ingredients of the proof:

Gerhart Frey (1985), who first suggested that the existence of a solution of the Fermat equation might contradict the Modularity Conjecture of Taniyama, Shimura, and Weil;

Jean-Pierre Serre (1985-6), who formulated and (with J.-F. Mestre) tested numerically a precise conjecture about modular forms and galois representations mod p and who showed how a small piece of this conjecture—the so-called *epsilon conjecture*—together with the Modularity Conjecture would imply Fermat's Last Theorem;

Ken Ribet (1986), who proved Serre's *epsilon conjecture*, thus reducing the proof of Fermat's Last Theorem to a proof of the Modularity Conjecture for semistable elliptic curves;

Richard Taylor (1994), who collaborated with Wiles to complete the proof of Wiles's numerical criterion in the *minimal case*;

Andrew Wiles (1994), who had the vision to identify the crucial numerical criterion from which the Modularity Conjecture for semistable elliptic curves would follow, and who finally supplied a proof of this criterion, thus completing the proof of Fermat's Last Theorem.

To prove the theorem we follow the program outlined by Serre in [16]. Fix a prime $p \geq 5$ and suppose $a, b, c \in \mathbf{Z}$ satisfy $a^p + b^p + c^p = 0$ but $abc \neq 0$. The triple (a^p, b^p, c^p) is what Gerhard Frey has called a "remarkable" triple of integers, so remarkable in fact, that we suspect it does not exist. To derive a contradiction, we will transform this triple into another object with remarkable properties, namely a very special modular form f_{a^p,b^p,c^p}, something firmly rooted in the fertile grounds of modern number theory. The construction of this modular form is a two-step process. First, by a simple but insightful construction due independently to Yves Hellegouarch and Gerhard Frey, we obtain a certain semistable elliptic curve E_{a^p,b^p,c^p} defined over $\mathbf{Q}$. Then, by Wiles's semistable modularity theorem, we deduce the existence of a modular form f_{a^p,b^p,c^p} associated to E_{a^p,b^p,c^p} by the correspondence of Eichler and Shimura.

With f_{a^p,b^p,c^p} in hand, we seek a contradiction within the realm of modular forms. The crucial ingredients that finally lead to a contradiction are encoded in a certain galois representation $\overline{\rho}_{a^p,b^p,c^p} : G \longrightarrow \mathrm{GL}_2(\mathbf{F}_p)$ mod p associated to f_{a^p,b^p,c^p}. As noted by Frey and Serre, the remarkableness of the triple (a^p, b^p, c^p) is reflected by some remarkable local properties of $\overline{\rho}_{a^p,b^p,c^p}$. Indeed, they noted that $\overline{\rho}_{a^p,b^p,c^p}$ can ramify only at 2 and p, and that the ramification at p is rather mild (what Serre called *peu ramifiée*). But experience with galois representations shows that it is difficult to make large galois representations with so little ramification. As Serre conjectured and Ribet proved, the existence of such a galois representation has untenable consequences in the theory of modular forms. Fermat's Last Theorem follows.

§1. A Remarkable Elliptic Curve

In this section we describe the crucial construction of an elliptic curve E_{a^p,b^p,c^p} out of a hypothetical solution of the Fermat equation $a^p+b^p+c^p = 0$. For any triple (A, B, C) of coprime integers satisfying $A + B + C = 0$, Gerhart Frey [8] considered the elliptic curve $E_{A,B,C}$ defined by the Weierstrass equation

$$E_{A,B,C} : y^2 = x(x - A)(x + B)$$

and explained some of the ways in which the arithmetic properties of $E_{A,B,C}$ are related to the diophantine properties of the triple (A, B, C). Especially interesting are the connections with the Masser-Oesterle A-B-C conjecture and its generalizations. For a discussion of this line of thought including connections with modular curves, we refer the reader to [7] and to Frey's article in this volume (chapter XX).

For our purposes it suffices to consider only the special case where $(A, B, C) = (a^p, b^p, c^p)$ corresponds to a hypothetical solution of the Fermat equation. Without loss of generality, we may assume $a \equiv -1$ modulo 4 and $2|b$. It is not hard to calculate both the minimal discriminant Δ_{a^p,b^p,c^p} and the conductor N_{a^p,b^p,c^p} of the elliptic curve E_{a^p,b^p,c^p}.

(1.1) Proposition. *Let $p \geq 5$ be prime and let a, b, c be coprime integers satisfying $abc \neq 0$, $a \equiv -1$ modulo 4, $2|b$, and $a^p + b^p + c^p = 0$. Then E_{a^p,b^p,c^p} is a semistable elliptic curve whose minimal discriminant and conductor are given by the formulas*
(a) $\Delta_{a^p,b^p,c^p} = 2^{-8} \cdot (abc)^{2p}$, *and*
(b) $N_{a^p,b^p,c^p} = \prod_{\ell|abc} \ell$.

For definitions of semistability and of the conductor and minimal discriminant see Silverman's article in this volume (chapter II, especially §14 and §17). In general the primes dividing the minimal discriminant of an elliptic curve over $\mathbf{Q}$ are the same as those dividing the conductor and this might lead us to suspect that the discriminant and conductor should be close to one another. Indeed, Szpiro has formulated the following conjecture (see [19] where a slightly stronger form of the conjecture is formulated).

Conjecture. (Szpiro) *For any $\epsilon > 0$ there is a constant $C > 0$ such that the minimal discriminant Δ_E and conductor N_E of any elliptic curve $E_{/\mathbf{Q}}$ satisfy the inequality*

$$|\Delta_E| < C \cdot N_E^{6+\epsilon}.$$

On the other hand, proposition 1.1 shows that a counterexample to $FLT(p)$ for sufficiently large p gives rise to an elliptic curve whose minimal discrimant and conductor are so far apart that they would contradict Szpiro's conjecture. We might thus hope to uncover a contradiction within the field of diophantine geometry. We will follow a different but related path and examine certain galois representations attached to E_{a^p,b^p,c^p}.

The idea of using elliptic curves to study Fermat's Last Theorem and vice versa goes back at least to the work of Y. Hellegouarch [9] (1972) who studied connections between the Fermat equation and torsion points on elliptic curves. Gerhart Frey seems to have been the first to suspect that a counterexample to Fermat's Last Theorem might contradict the Modularity Conjecture and to investigate various approaches based on this idea.

§2. Galois Representations.

In this section we collect the basic definitions and conventions from the theory of galois representations that we will need later. For more details we refer the reader to the article by Mazur in this volume (chapter VIII).

Let $\overline{\mathbf{Q}}$ be the algebraic closure of $\mathbf{Q}$ in $\mathbf{C}$. We endow the galois group $G_{\mathbf{Q}} := Gal(\overline{\mathbf{Q}}/\mathbf{Q})$ with the Krull topology in which a basis of neighborhoods of the origin is given by the collection of subgroups $H \subseteq G_{\mathbf{Q}}$ of finite index in $G_{\mathbf{Q}}$. With this topology, $G_{\mathbf{Q}}$ is a profinite group and in particular is a compact topological group.

By a two dimensional galois representation over a topological ring A we mean a continuous group homomorphism

$$\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A).$$

In this paper, the topological ring A will always be what Mazur calls a *coefficient ring* (in chapter VIII). Since these rings will play an important role in what follows, we make a formal definition.

(2.1) Definition. A *coefficient ring* is a complete noetherian local ring with finite residue field of characteristic p (our fixed prime).

Whenever we write that $\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$ is a galois representation, it is understood that A is a coefficient ring and that ρ is continuous.

(2.2) Residual representations and deformations. Let A be a coefficient ring with maximal ideal m_A and let $k_A := A/m_A$ be the residual field. We define the *residual representation* of a galois representation $\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$ to be the representation

$$\overline{\rho} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(k_A)$$

obtained by composing ρ with the reduction map $\mathrm{GL}_2(A) \longrightarrow \mathrm{GL}_2(k_A)$. Conversely, if $\rho_0 : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(k)$ is a two dimensional galois representation over a finite field k, then we say that ρ is a *lifting* of ρ_0 to A if $k = k_A$ and $\overline{\rho} = \rho_0$. Two liftings ρ, ρ' of ρ_0 to A are said to be *equivalent* if ρ' can be conjugated to ρ by a matrix in $\mathrm{GL}_2(A)$ that is congruent to the identity matrix modulo m_A.

A *deformation* of ρ_0 to A is an equivalence class of liftings of ρ_0 to A. For a given lifting ρ of ρ_0, we will abuse notation and also write ρ to denote the deformation to which it belongs. This should not cause confusion in our discussion.

(2.3) The determinant of a galois representation. If ρ is a two dimensional galois representation over A then

$$\det(\rho) : G_{\mathbf{Q}} \longrightarrow A^{\times}$$

will denote the composition of ρ with the determinant homomorphism $\det : \mathrm{GL}_2(A) \longrightarrow A^{\times}$. In the applications it is sometimes convenient to restrict our attention to representations with prescribed determinant.

For example, let $\chi_p : G_{\mathbf{Q}} \longrightarrow \mathbf{Z}_p^{\times}$ denote the *cyclotomic character*, which is characterized by the property $\sigma(\zeta) = \zeta^{\chi_p(\sigma)}$ for any p-power root of unity ζ and any $\sigma \in G_{\mathbf{Q}}$. Any coefficient ring A admits a unique continuous ring homomorphism $\mathbf{Z}_p \longrightarrow A$ and we therefore have a canonical group homomorphism $\mathbf{Z}_p^{\times} \longrightarrow A^{\times}$. We say that ρ has determinant χ_p if $\det(\rho)$ is the composition of χ_p with the canonical homomorphism $\mathbf{Z}_p^{\times} \longrightarrow A^{\times}$.

(2.4) Local galois groups. For each prime ℓ, we let $\mathbf{Q}_\ell$ denote the field of ℓ-adic rationals, i.e. the completion of $\mathbf{Q}$ with respect to the ℓ-adic absolute value $|\cdot|_\ell$. We fix once and for all an algebraic closure $\overline{\mathbf{Q}}_\ell$ of $\mathbf{Q}_\ell$ as well as an embedding of $\overline{\mathbf{Q}}$ into $\overline{\mathbf{Q}}_\ell$. For $\ell = \infty$ we let $\mathbf{Q}_\infty := \mathbf{R}$, the completion

of $\mathbf{Q}$ with respect to the usual absolute value $|\cdot|_\infty$, and we take $\overline{\mathbf{Q}}_\infty := \mathbf{C}$. For each ℓ (ℓ prime, or $\ell = \infty$), the *local galois group* at ℓ is the group

$$G_{\mathbf{Q}_\ell} := Gal(\overline{\mathbf{Q}}_\ell/\mathbf{Q}_\ell).$$

For $\ell = \infty$, we have

$$G_{\mathbf{Q}_\infty} := Gal(\mathbf{C}/\mathbf{R}) = \langle c \rangle,$$

the cyclic group of order 2 generated by complex conjugation c. It is well-known that for each ℓ there is a unique absolute value $|\cdot|_\ell$ on $\overline{\mathbf{Q}}_\ell$ extending the given absolute value on $\mathbf{Q}_\ell$. From this it follows easily that the elements of $G_{\mathbf{Q}_\ell}$ are *continuous* automorphisms of $\overline{\mathbf{Q}}_\ell$.

Using our fixed embeddings $\overline{\mathbf{Q}} \subseteq \overline{\mathbf{Q}}_\ell$, we may restrict any automorphism of $\overline{\mathbf{Q}}_\ell$ to obtain an automorphism of $\overline{\mathbf{Q}}$. Since $\overline{\mathbf{Q}}$ is dense in $\overline{\mathbf{Q}}_\ell$, the induced homomorphisms $G_{\mathbf{Q}_\ell} \to G_{\mathbf{Q}}$ are injective and we will regard them as inclusions:

$$G_{\mathbf{Q}_\ell} \subseteq G_{\mathbf{Q}}.$$

These subgroups are often called the decomposition subgroups of $G_{\mathbf{Q}}$. Of course, strictly speaking, they are not well-defined since their definition depends on our choice of the fixed embeddings of $\overline{\mathbf{Q}}$ into $\overline{\mathbf{Q}}_\ell$. However, changing any one of these embeddings has the effect of conjugating the corresponding decomposition subgroup by an element of $G_{\mathbf{Q}}$. This ambiguity will not be important to us.

(2.5) Inertia groups. For $\ell \neq \infty$, $G_{\mathbf{Q}_\ell}$ preserves the ring $\overline{\mathbf{Z}}_\ell$ of integers in $\overline{\mathbf{Q}}_\ell$ and also preserves the maximal ideal $\lambda \subseteq \overline{\mathbf{Z}}_\ell$. Thus, $G_{\mathbf{Q}_\ell}$ acts naturally on the residual field $\overline{\mathbf{F}}_\ell = \overline{\mathbf{Z}}_\ell/\lambda$ and we obtain a natural map $G_{\mathbf{Q}_\ell} \longrightarrow Gal(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell)$, which is easily seen to be surjective. Its kernel I_ℓ is called the inertia group at ℓ. Thus for each $\ell \neq \infty$, we have an exact sequence

$$1 \longrightarrow I_\ell \longrightarrow G_{\mathbf{Q}_\ell} \longrightarrow Gal(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell) \longrightarrow 1.$$

(2.6) Local properties of galois representations. Given a *global* galois representation $\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$, we may restrict ρ to the decomposition groups $G_{\mathbf{Q}_\ell}$ and obtain the family $\{\rho|_{G_\ell}\}$ of *local* galois representations

$$\rho|_{G_\ell} : G_{\mathbf{Q}_\ell} \longrightarrow \mathrm{GL}_2(A).$$

In many important examples from number theory one knows that the global representation ρ is determined up to isomorphism by the family of local representations $\{\rho|_{G_\ell}\}_{\ell \notin S}$, where ℓ ranges over the complement of any finite set S of primes. By the local properties at ℓ of a galois representation ρ we mean the properties of the local representation $\rho|_{G_\ell}$. The next three definitions describe three local properties that play a special role in what follows.

(2.7) Definition. We say that ρ is *odd* if $\det\rho(c) = -1$, where c is the complex conjugation generating $G_{\mathbf{Q}_\infty}$.

(2.8) Definition. We say that ρ is *unramified* at a prime ℓ if $I_\ell \subseteq \ker\rho|_{G_\ell}$.

Since the galois group $Gal(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell)$ is a topologically cyclic group generated by the ℓth power Frobenius automorphism $Frob_\ell$, when ρ is unramified at ℓ, $\rho|_{G_\ell}$ may be viewed as a homomorphism $Gal(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell) \longrightarrow \mathrm{GL}_2(A)$ and is thus determined by its value on any representative of $Frob_\ell$ in $G_{\mathbf{Q}_\ell}$.

When $\ell = p$ we need the following weaker condition.

(2.9) Definition. We say that ρ is *flat* at p if, for every ideal $I \subseteq A$ for which A/I is finite, the representation $G_{\mathbf{Q}_p} \longrightarrow \mathrm{GL}_2(A/I)$, obtained by reducing $\rho|_{G_{\mathbf{Q}_p}}$ mod I, extends to a finite flat group scheme over $\mathbf{Z}_p$ (see Tate's article in this volume (chapter V)).

(2.10) Examples from number theory. The Galois representations that arise naturally in number theory have the especially nice property of being unramified almost everywhere, that is, they are unramified at all but finitely many primes ℓ. For example, let $E_{/\mathbf{Q}}$ be an elliptic curve. Then for each $n \geq 0$ the galois group $G_{\mathbf{Q}}$ acts on the group $E[p^n] \cong (\mathbf{Z}/p^n\mathbf{Z})^2$ of p^n-torsion points on E. Since the action of $G_{\mathbf{Q}}$ commutes with multiplication by p on E, $G_{\mathbf{Q}}$ acts naturally on the *Tate module* $Ta_p(E) := \varprojlim E[p^n] \cong \mathbf{Z}_p^2$ and we obtain the p-adic galois representation

$$\rho_{E,p} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathbf{Z}_p)$$

associated to E. The residual representation $\overline{\rho}_{E,p} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathbf{F}_p)$. describes the action of $G_{\mathbf{Q}}$ on $E[p] \cong \mathbf{F}_p^2$. We have the following basic result concerning the properties of these representations.

(2.11) Theorem. *Let $\rho_{E,p}$ be the p-adic galois representation associated to an elliptic curve $E_{/\mathbf{Q}}$ and let N_E be the conductor of E. Then*

- *the determinant of $\rho_{E,p}$ is χ_p, and*
- *$\rho_{E,p}$ is unramified outside of pN_E.*

In particular, $\rho_{E,p}$ is odd. If E is semistable with minimal discriminant Δ_E, then the residual representation $\overline{\rho}_{E,p}$ has the following local properties.

- *If $\ell \neq p$, then $\overline{\rho}_{E,p}$ is unramified at $\ell \iff p|ord_\ell(\Delta_E)$.*
- *$\overline{\rho}_{E,p}$ is flat at $p \iff p|ord_p(\Delta_E)$.*

§3. A Remarkable Galois Representation.

Let $E := E_{a^p,b^p,c^p}$ be as in §1 and consider the galois representation

$$\overline{\rho}_{a^p,b^p,c^p} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathbf{F}_p)$$

given by $\overline{\rho}_{a^p,b^p,c^p} = \overline{\rho}_{E,p}$. Gerhart Frey [7,8] and Jean-Pierre Serre [16] noted that this representation has some remarkable local properties. More precisely they proved the following theorem.

(3.1) Theorem *Let $p \geq 5$ be prime and $a, b, c \in \mathbf{Z}$ satisfy $a^p + b^p + c^p = 0$ and $abc \neq 0$. Assume further that $a \equiv -1$ modulo 4 and $2|b$. Then*

(a) $\overline{\rho}_{a^p,b^p,c^p}$ *is absolutely irreducible;*

(b) $\overline{\rho}_{a^p,b^p,c^p}$ *is odd;*

(c) $\overline{\rho}_{a^p,b^p,c^p}$ *is unramified outside $2p$ and is flat at p.*

One suspects that there are no galois representations $\rho_0 : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathbf{F}_p)$ satisfying properties (a), (b) and (c), but this suspicion remains unproven. On the other hand, by a theorem of Ribet, we *do* know that no such galois representation lives in the world of modular forms, in a sense that we will make precise in the next section.

§4. Modular Galois Representations.

The theory of modular forms offers a rich source of galois representations. Using the Hecke operators, these "modular" galois representations can be constructed out of the torsion groups on the modular jacobians $J_1(N)$, $N > 0$ by the method of Eichler and Shimura. For an introduction to the theory of modular forms and the Eichler-Shimura theory, see David Rohrlich's article in this volume (chapter III).

(4.1) Galois representations associated to newforms. Fix, once and for all, a prime $\wp$ of $\overline{\mathbf{Q}}$ lying over p. Let $f = \sum_{n\geq 1} a_n q^n$ be a weight two (normalized) newform of conductor N and character ϵ (in (3.5) of chapter III, newforms are called primitive forms). We let K_f be the completion at $\wp$ of the number field generated by the values of ϵ and the fourier coefficents a_n ($n \geq 1$), and we let $\mathcal{O}_f \subseteq K_f$ be the ring of integers in K_f. The theory of Eichler and Shimura associates to f an odd two dimensional galois representation

$$\rho_f : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathcal{O}_f)$$

such that for all sufficiently large primes ℓ, ρ_f is unramified at ℓ and

$$\mathrm{Trace}\big(\rho_f(Frob_\ell)\big) = a_\ell \quad \text{and} \quad \det\big(\rho_f(Frob_\ell)\big) = \epsilon(\ell)\ell.$$

For the details of the Eichler-Shimura construction, we refer to section 3.7 of Rohrlich's chapter III in this volume, where ρ_f appears as ρ_λ. By the work of Carayol and others, we now have a good understanding of the local structure of ρ_f at all primes. In particular we know that ρ_f is unramified outside pN and that the above conditions on the trace and determinant of $\rho_f(Frob_\ell)$ are satisfied for these primes.

By the work of Deligne [3] and Deligne-Serre [4], we know that similar assertions hold for newforms of any weight $w \geq 1$. Indeed, if f is a weight w newform of conductor N then Deligne has constructed an odd two dimensional p-adic galois representation ρ_f, which is unramified outside pN and satisfies $\mathrm{Trace}(\rho_f(Frob_\ell)) = a_\ell$ and $\det(\rho_f(Frob_\ell)) = \epsilon(\ell)\ell^{w-1}$ for all $\ell \not| pN$. In this paper, we will be concerned almost exclusively with the case $w = 2$.

(4.2) Hecke algebras. Let $N > 0$ be an integer and let $S_2(N)$ denote the space of weight 2 cusp forms for $\Gamma_1(N)$ (see (3.2) of chapter III). We let

$$\mathbf{T}'(N) := \mathbf{Z}[T_\ell, \langle d \rangle] \subseteq \mathrm{End}\big(S_2(N)\big)$$

be the $\mathbf{Z}$-subalgebra of $\mathrm{End}(S_2(N))$ generated by the Hecke operators T_ℓ and the diamond operators $\langle d \rangle$ where ℓ runs over all primes not dividing pN, and d runs over $(\mathbf{Z}/N\mathbf{Z})^\times$ (see (3.3) of chapter III).

(4.3) Modularity of galois representations. Motivated by (4.1) we say that a galois representation

$$\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$$

over a coefficient ring A is *modular* if there exists an integer $N > 0$ and a homomorphism $\pi : \mathbf{T}'(N) \longrightarrow A$ such that ρ is unramified outside Np and for every prime $\ell \not| pN$ we have

$$\mathrm{Trace}(\rho(Frob_\ell)) = \pi(T_\ell) \quad \text{and} \quad \det(\rho(Frob_\ell)) = \pi(\langle \ell \rangle)\ell.$$

Remark: In view of the above restriction on the determinant it might be more appropriate to call these modular representations of weight 2. However, since all of our representations will have weight 2, we will drop that modifier from our language.

(4.4) Serre's Conjectures. In the special case where $A = k$ is a finite field, Serre [16] has formulated some precise conjectures about modularity of galois representations over k. One consequence of Serre's conjectures is the following conjecture.

Conjecture. *Every odd absolutely irreducible galois representation*

$$\rho_0 : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(k)$$

is modular (in the sense of (4.3)).

In fact, Serre's conjectures are much more precise. They predict—in terms of the local structure of ρ—the optimal weight, conductor and character of a newform f for which $\overline{\rho}_f = \rho_0$. For precise statements of Serre's conjectures and an account of what is known about them today, see the article by Edixhoven in this volume (chapter VII). An important special case of these conjectures, which Serre called the *epsilon conjecture* in [16], is the following theorem of Ribet [13] (see §3 of chapter VII for a sketch of the proof).

(4.5) Ribet's Theorem. *Let f be a weight two newform of conductor $N\ell$ where $\ell \not| N$ is a prime. Suppose $\overline{\rho}_f$ is absolutely irreducible and that one of the following is true:*

- *$\overline{\rho}_f$ is unramified at ℓ; or*

- $\ell = p$ *and* $\overline{\rho}_f$ *is flat at* p.

Then there is a weight two newform g *of conductor* N *such that* $\overline{\rho}_f \cong \overline{\rho}_g$.

§5. The Modularity Conjecture and Wiles's Theorem.

We say that an elliptic curve $E_{/\mathbf{Q}}$ is modular if there is a weight two newform f of conductor N_E and trivial character for which

$$L(f, s) = L(E, s).$$

There are a number of equivalent ways of defining modularity of elliptic curves. Here are a few.

(5.1) Theorem. *The following assertions are equivalent for an elliptic curve* $E_{/\mathbf{Q}}$.

(a) *E is modular;*
(b) *for some prime p, $\rho_{E,p}$ is modular;*
(c) *for every prime p, $\rho_{E,p}$ is modular;*
(d) *there is a non-constant morphism $\pi : X_0(N_E) \longrightarrow E$ of algebraic curves defined over $\mathbf{Q}$;*
(e) *E is isogenous to the modular abelian variety A_f associated to some weight two newform f of conductor N_E.*

We have the following profound conjecture developed between 1957 and 1967 by Shimura, Taniyama, and Weil.

(5.2) The Modularity Conjecture. *Every elliptic curve over* $\mathbf{Q}$ *is modular.*

The Modularity Conjecture is still open in general, but thanks to the work of Wiles [20] and Taylor–Wiles [18], we know at least that it is true for a large and important class of elliptic curves, namely the semistable ones.

(5.3) Wiles's Theorem. *Every semistable elliptic curve over* $\mathbf{Q}$ *is modular.*

We will sketch the proof in §7. In fact, by improving Wiles's methods, Fred Diamond [5] has proven the much stronger result that every elliptic curve $E_{/\mathbf{Q}}$ that is semistable at 3 and 5 is modular. The proof is outlined in chapter XVII by Diamond.

§6. The proof of Fermat's Last Theorem.

Returning to the situation of §1 and §3 we suppose $p \geq 5$ and assume $a, b, c \in \mathbf{Z}$ satisfy $a^p + b^p + c^p = 0$ but $abc \neq 0$. We derive a contradiction by the method described in [16] (see also [8]). Without loss of generality, we may assume $a \equiv -1 \pmod 4$ and $2|b$. Let E_{a^p,b^p,c^p} be the elliptic curve $y^2 = x(x - a^p)(x + b^p)$ and let ρ_{a^p,b^p,c^p} be the associated p-adic galois representation.

By proposition 1.1, E_{a^p,b^p,c^p} is semistable and has conductor $N_{a^p,b^p,c^p} = \prod_{\ell|abc} \ell$. Hence, by Wiles's theorem, E_{a^p,b^p,c^p} is modular and there is a weight two newform f_{a^p,b^p,c^p} of conductor N_{a^p,b^p,c^p} associated to E_{a^p,b^p,c^p}. In particular, we have $\rho_{a^p,b^p,c^p} \cong \rho_{f_{a^p,b^p,c^p}}$. But according to theorem 2.11 $\overline{\rho}_{a^p,b^p,c^p}$ is absolutely irreducible and is unramified outside $2p$ and flat at p. Applying Ribet's Theorem we conclude that there is a weight two newform g of conductor 2 such that $\overline{\rho}_g \cong \overline{\rho}_{a^p,b^p,c^p}$. But the dimension of $S_2(\Gamma_0(2))$ is equal to the genus of $X_0(2)$, which is easily seen to be zero. Thus there are no weight two newforms of conductor 2. This is a contradiction and Fermat's Last Theorem is proved.

§7. The proof of Wiles's Theorem.

In this final section, we describe the structure of the proof of Wiles's Theorem [18,20]. For other surveys of the proof, we recommend [2,12,14,17]. Here we assume that the distinguished prime p is ≥ 3. Let k be a finite field of characteristic p and let

$$\rho_0 : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(k)$$

be a galois representation. As we move through this section we will impose a number of cumulative hypotheses on ρ_0. The first of these is the following.

Hypothesis A. ρ_0 has determinant χ_p.

(7.1) Semistable galois representations. We say that a galois representation

$$\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$$

is *ordinary* at p if the restriction of ρ to the inertia group I_p at p has the form $\rho|_{I_p} = \begin{pmatrix} \chi_p & * \\ 0 & 1 \end{pmatrix}$ for a suitable choice of basis. We say that ρ is *semistable* at a prime ℓ if one of the following two conditions is satisfied.

- $\ell = p$ and ρ is either flat at p or ordinary at p (or both).
- $\ell \neq p$ and $\rho|_{I_\ell} = \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}$ for a suitable choice of basis.

We say that a two dimensional galois representation ρ is semistable if it is semistable at every prime. From now on, we impose the following additional hypothesis on ρ_0.

Hypothesis B. ρ_0 is semistable.

The use of the word *semistable* in this context is motivated by the simple fact that if $E_{/\mathbf{Q}}$ is a semistable elliptic curve, then the p-adic galois representation $\rho_{E,p} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathbf{Z}_p)$ is semistable in the above sense.

(7.2) Deformation types. A deformation type $\mathcal{D}$ is a list of conditions to be imposed on deformations of ρ_0, satisfying certain properties. Using more sophisticated terminology, a deformation type may be regarded as

a functor from the category of coefficient rings to the category of sets, where, for a given coefficient ring A, $\mathcal{D}(A)$ is the set of two dimensional galois representations over A that satisfy the conditions of $\mathcal{D}$. For more discussion of deformation types we refer the reader to Mazur's chapter VIII in this volume.

Wiles considers a variety of different deformation types, but for the application to the semistable modularity conjecture it suffices to restrict to the following special cases. Let $S := \{\, \ell \neq p \,|\, \rho_0 \text{ is ramified at } \ell \,\}$. A deformation type $\mathcal{D}$ is associated to a finite set of primes $\Sigma_{\mathcal{D}}$ disjoint from $S \cup \{\, p \,\}$. We say that a deformation ρ of ρ_0 is of type $\mathcal{D}$ if the following conditions are satisfied.

- ρ has determinant χ_p,
- ρ is unramified outside $S \cup \{\, p \,\} \cup \Sigma_{\mathcal{D}}$,
- ρ is semistable outside $\Sigma_{\mathcal{D}}$, and
- if $p \notin \Sigma_{\mathcal{D}}$ and if ρ_0 is flat at p, then ρ is also flat at p.

Roughly speaking, the last three conditions say that ρ has the same local properties as ρ_0 at primes not in $\Sigma_{\mathcal{D}}$. We remark that in any case, if ρ_0 is ordinary at p then ρ is also ordinary at p.

(7.3) Universal deformation rings and Hecke rings. In addition to hypotheses A and B above we suppose ρ_0 satisfies the following hypothesis.

Hypothesis C. ρ_0 is absolutely irreducible.

Using Mazur's theory of deformations of galois representations [10], Wiles associates to each deformation type $\mathcal{D}$ a *universal deformation ring* $R_{\mathcal{D}}$ (which is, in particular, a coefficient ring) and a *universal deformation*

$$\rho_{\mathcal{D}} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(R_{\mathcal{D}})$$

of ρ_0 of type $\mathcal{D}$. The representation $\rho_{\mathcal{D},mod}$ satisfies the following universal property: for every deformation $\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$ of ρ_0 of type $\mathcal{D}$ there is a unique homomorphism $\pi_A : R_{\mathcal{D}} \longrightarrow A$ such that the diagram

$$\begin{array}{ccc} G_{\mathbf{Q}} & \xrightarrow{\rho_{\mathcal{D}}} & \mathrm{GL}_2(R_{\mathcal{D}}) \\ & \rho \searrow \quad \swarrow \pi_A & \\ & \mathrm{GL}_2(A) & \end{array}$$

is commutative. For details on the properties and construction of $R_{\mathcal{D}}$ see chapter VIII by Mazur and chapter XIII by Brian Conrad. An explicit approach to constructing deformation rings is given in chapter IX by de Smit, Rubin, and Schoof.

Hypothesis D. ρ_0 is modular, and $\rho_0|_{G_{\mathbf{Q}(\sqrt{-3})}}$ is absolutely irreducible.

Under this hypothesis, Wiles defines another coefficient ring $\mathbf{T}_{\mathcal{D}}$, the *universal modular deformation ring* and a *universal modular deformation*

$$\rho_{\mathcal{D},mod} : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathbf{T}_{\mathcal{D}})$$

of ρ_0 of type $\mathcal{D}$. The representation $\rho_{\mathcal{D},mod}$ satisfies the analogous universal property for modular deformations of type $\mathcal{D}$. Namely, for every *modular* deformation $\rho : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(A)$ of ρ_0 of type $\mathcal{D}$ there is a unique homomorphism $\pi_A : \mathbf{T}_{\mathcal{D}} \longrightarrow A$ such that the obvious diagram commutes.

The constructions of $\mathbf{T}_{\mathcal{D}}$ and $\rho_{\mathcal{D},mod}$ are quite difficult. The algebra $\mathbf{T}_{\mathcal{D}}$ is defined in chapter XII by Diamond and Ribet. It's existence depends on the highly non-trivial fact (described in chapter VII by Edixhoven) that ρ_0 admits at least one modular deformation of type $\mathcal{D}$. The representation $\rho_{\mathcal{D},mod}$ is cut out of the Tate module of a modular Jacobian using the Hecke operators. Wiles's proof that this representation is a free rank two $\mathbf{T}_{\mathcal{D}}$-module depends on the Gorenstein property of $\mathbf{T}_{\mathcal{D}}$ (see Tilouine's chapter X in this volume). Later, other proofs of this fact were given that do not make explicit use of the Gorenstein property, but rather have the Gorenstein property as a by-product (for example, see [6]).

(7.4) The main theorem. By the universal property of $\rho_{\mathcal{D}}$ there is a unique homomorphism $\varphi_{\mathcal{D}} : R_{\mathcal{D}} \longrightarrow \mathbf{T}_{\mathcal{D}}$ such that $\rho_{\mathcal{D},mod} = \varphi_{\mathcal{D}} \circ \rho_{\mathcal{D}}$. The following theorem is a special case of the main theorem of Wiles [20].

Theorem. *Suppose ρ_0 satisfies hypotheses* A-D. *Then the canonical map $\varphi_{\mathcal{D}} : R_{\mathcal{D}} \longrightarrow \mathbf{T}_{\mathcal{D}}$ is an isomorphism of complete intersection rings.*

For the definition of complete intersection rings, we refer to chapter IX by Schoof, Rubin, and de Smit in this volume. For our purposes what matters is the conclusion that $\varphi_{\mathcal{D}}$ is an isomorphism. The proof of the theorem is based on the numerical criterion of Wiles described in the next secion, which reduces the proof to an inequality between two numbers. The theorem has the following important corollary as an immediate consequence.

Corollary. *Suppose ρ_0 satisfies hypotheses* A-D. *Then every deformation of ρ_0 of type $\mathcal{D}$ is modular.*

(7.5) Wiles's numerical criterion. *Let R and T be coefficient rings and suppose we have a commutative diagram*

$$\begin{array}{ccc} R & \xrightarrow{\varphi} & T \\ & {\scriptstyle \pi_R}\searrow \quad \swarrow{\scriptstyle \pi_T} & \\ & \mathcal{O} & \end{array}$$

in which $\mathcal{O}$ is a complete discrete valuation ring and all the arrows are surjective. Let $I_R := \ker \pi_R$, $I_T := \ker \pi_T$, and let $\eta_T := \pi_T\big(\mathrm{Ann}_T(I_T)\big)$. Then the following three assertions are equivalent.

- *φ is an isomorphism of complete intersection rings;*
- *I_R/I_R^2 is finite and $\#(I_R/I_R^2) \leq \#(\mathcal{O}/\eta_T)$;*
- *I_R/I_R^2 is finite and $\#(I_R/I_R^2) = \#(\mathcal{O}/\eta_T)$.*

This is a special case of Criterion I given in chapter IX by Schoof, Rubin, and de Smit.

(7.6) Selmer groups and congruence modules. Now let f be a weight two newform and suppose $\rho_f : G_{\mathbf{Q}} \longrightarrow \mathrm{GL}_2(\mathcal{O}_f)$ is a deformation of ρ_0 of type $\mathcal{D}$. By the universality of $\mathbf{T}_{\mathcal{D}}$ there is a unique homomorphism $\pi_{\mathbf{T}_{\mathcal{D}}} : \mathbf{T}_{\mathcal{D}} \longrightarrow \mathcal{O}_f$ such that $\rho_f = \pi_{\mathbf{T}_{\mathcal{D}}} \circ \rho_{\mathcal{D},mod}$. Let $\pi_{R_{\mathcal{D}}} := \pi_{\mathbf{T}_{\mathcal{D}}} \circ \varphi_{\mathcal{D}}$ so that we have the following commutative diagram:

$$\begin{array}{ccc} R_{\mathcal{D}} & \xrightarrow{\varphi_{\mathcal{D}}} & \mathbf{T}_{\mathcal{D}} \\ & {\scriptstyle \pi_{R_{\mathcal{D}}}} \searrow \quad \swarrow {\scriptstyle \pi_{\mathbf{T}_{\mathcal{D}}}} & \\ & \mathcal{O}_f. & \end{array}$$

To prove that $\varphi_{\mathcal{D}}$ is an isomorphism, Wiles establishes the middle inequality in the above criterion. For this, he first interprets the two sides of the inequality in terms of other objects that have been studied in some detail in the literature. More precisely, Wiles interprets the "tangent space" $Hom_{\mathcal{O}}(I_{R_{\mathcal{D}}}/I_{R_{\mathcal{D}}}^2, K/\mathcal{O})$ as a *Selmer group* $H^1_{\mathcal{D}}(G_{\mathbf{Q}}, ad^0(\rho_f) \otimes K/\mathcal{O})$ (i.e. as a certain subgroup of the galois cohomology group $H^1(G_{\mathbf{Q}}, ad^0(\rho_f) \otimes K/\mathcal{O})$ determined by local conditions associated to $\mathcal{D}$), and he interprets $\mathcal{O}/\eta_{\mathbf{T}_{\mathcal{D}}}$ as a *congruence module* classifying congruences between f and other newforms of type $\mathcal{D}$. For precise definitions, see chapter XII by Diamond and Ribet, sections 4.2 and 4.3, chapter VIII by Mazur, and chapter IV by Washington. The isomorphism between tangent spaces and Selmer groups is described in chapter VIII.

The proof of the crucial numerical inequality divides into two parts. The case where $\Sigma_{\mathcal{D}} = \phi$, which is called the *minimal case*, is proved by Wiles with Taylor in [18]. Their original proof has been simplified by making use of another criterion due to Faltings, a generalization of which is given as criterion II in chapter XI. This is the method followed by de Shalit in chapter XIV. The *non-minimal case* is proved by induction on the number of primes in $\Sigma_{\mathcal{D}}$. The proof is accomplished by analyzing how the Selmer groups and congruence modules grow as $\Sigma_{\mathcal{D}}$ is enlarged to conclude that if the numerical inequality is satisfied for one $\mathcal{D}$ then it is also satisfied when more primes are included in $\Sigma_{\mathcal{D}}$. See chapter XII by Diamond and Ribet for more details.

(7.7) The Proof of Wiles's Theorem. We prepare for the proof by noting that hypotheses A and B are satisfied by $\overline{\rho}_{E,p}$ for every prime p. Indeed hypothesis A is contained in theorem 2.11 and hypothesis B is a consequence of the semistability of E.

Moreover, by a theorem of Serre ([15], prop. 21, and [17], §3.1), the semistability of E guarantees that $\overline{\rho}_{E,p}$ is either surjective or reducible for every prime $p \geq 3$. Hence for $p \geq 3$, absolute irreducibility of $\overline{\rho}_{E,p}$ is equivalent to irreducibility of $\overline{\rho}_{E,p}$, and if $p = 3$ this is equivalent to absolute

irreducibility of $\overline{\rho}_{E,3}|_{G_{\mathbf{Q}(\sqrt{-3})}}$. Thus the following lemma is a consequence of corollary 7.4.

(7.8) Lemma. *Let $E_{/\mathbf{Q}}$ be a semistable elliptic curve and suppose $\overline{\rho}_{E,p}$ is both modular and irreducible for some prime $p \geq 3$. Then E is modular.*

Wiles gave an ingenious argument to show that for E semistable, the hypotheses of the lemma are satisfied by either $p = 3$ or $p = 5$. The proof is based on the following three theorems.

(7.9) Theorem. *Let E be an arbitrary elliptic curve and suppose $\overline{\rho}_{E,3}$ is irreducible. Then $\overline{\rho}_{E,3}$ is modular.*

This follows from a deep theorem of Langlands and Tunnell and depends in a crucial way on the theory of Langlands for GL_2. For an exposition of the Langlands theory and the proof of Theorem 7.9, see chapter VI by Stephen Gelbart in this volume.

(7.10) Theorem. *Let $E_{/\mathbf{Q}}$ be a semistable elliptic curve and suppose $\overline{\rho}_{E,5}$ is irreducible. Then there is another semistable elliptic curve $E'_{/\mathbf{Q}}$ for which*
(a) *$\overline{\rho}_{E',3}$ is irreducible, and*
(b) *$\overline{\rho}_{E',5} \cong \overline{\rho}_{E,5}$.*

Indeed, proposition 11 and the argument in section 4 of Rubin's chapter XVI in this volume provide us with a family of elliptic curves $E'_{/\mathbf{Q}}$ satisfying conditions (a) and (b). All of these curves are semistable away from 5. By taking E' in this family sufficiently close 5-adically to E, we obtain the desired semistable curve.

(7.11) Theorem. *Let $E_{/\mathbf{Q}}$ be a semistable elliptic curve. Then at least one of the representations $\overline{\rho}_{E,3}$ or $\overline{\rho}_{E,5}$ is irreducible.*

Indeed, if both $\overline{\rho}_{E,3}$ and $\overline{\rho}_{E,5}$ were reducible then $E[15]$ would contain a galois invariant subgroup of order 15. This contradicts Lemma 9 (iv) of chapter XVI by Karl Rubin (see also [11]).

(7.12) Conclusion of the proof. Let $E_{/\mathbf{Q}}$ be a semistable elliptic curve. If $\overline{\rho}_{E,3}$ is irreducible then, according to theorem 7.9, $\overline{\rho}_{E,3}$ is also modular, so E is modular by lemma 7.8. If $\overline{\rho}_{E,3}$ is not irreducible, then by theorem 7.11, $\overline{\rho}_{E,5}$ is irreducible. Then there is another semistable elliptic curve $E'_{/\mathbf{Q}}$ satisfying (a) and (b) of theorem 7.10. In particular, $\rho_{E',3}$ is irreducible. Repeating the above argument we see that E' is modular. Hence $\rho_{E',5}$ is modular and by (b) of 7.10, $\overline{\rho}_{E,5}$ is modular. Once again we use lemma 7.8 to conclude E is modular.

References

[1] Carayol, H.: Sur les représentations galoisiennes modulo ℓ attachées aux formes modulaires. *Duke Math. J.* **59** (1989), 785-801.

[2] Darmon, H., Diamond, F., Taylor, R. L.: Fermat's Last Theorem. In *Current Developments in Mathematics, 1995*, International Press. To appear.

[3] Deligne, P.: Formes modulaires et représentation ℓ-adiques. Sém. Bourbaki, 1968/69, Exposé 355. *Lect. Notes in Math.* **179** (1971), 139-172.

[4] Deligne, P., Serre, J.-P.: Formes modulaires de poids 1. *Ann. Sci. E.N.S.* **7** (1974), 507-530.

[5] Diamond, F.: On deformations rings and Hecke rings. *Ann. of math.*. To appear.

[6] Diamond, F.: The Taylor-Wiles construction and multiplicity one. *Invent. Math.*. To appear.

[7] Frey, G.: Links between solutions of $A - B = C$ and elliptic curves. In *Number Theory, proceedings of the Journees arithmetiques, held in Ulm, 1987*, H.P. Schlickewei, E. Wirsing, editors. Lecture notes in mathematics **1380**. Springer-Verlag, Berlin, New York, 1989.

[8] Frey, G.: Links between stable elliptic curves and certain Diophantine equations. *Ann. Univ. Saraviensis, Ser. Math.* **1** (1986), 1-40.

[9] Hellegouarch, Y.: Points d'ordre $2p^h$ sur les courbes elliptiques. *Acta. Arith.* **26** (1974/75), 253-263.

[10] Mazur, B.: Deforming Galois representations. In *Galois groups over* **Q**: proceedings of a workshop held March 23-27, 1987, Y. Ihara, K. Ribet, J.-P. Serre, editors. Mathematical Sciences Research Institute publications **16**. Springer-Verlag, New York,1989, pp. 385-437.

[11] Mazur, B.: Modular curves and the Eisenstein ideal. *Publ. Math. I.H.E.S.* **47** (1977), 33-186.

[12] Murty, V.K.: Modular elliptic curves. in *Seminar on Fermat's Last Theorem.* Canadian Math. Soc. Conf. Proc. **17**, 1995.

[13] Ribet, K.A.: On modular representations of $Gal(\overline{\mathbf{Q}}/\mathbf{Q})$ arising from modular forms. *Invent. math.* **100** (1990), 431-476.

[14] Oesterlé, J.: Travaux de Wiles (et Taylor, ...), Partie II. *Asterisque* **237** (1996), 333-355.

[15] Serre, J.-P.: Propriétés galoisiennes des points d'ordre fini des courbes elliptiques. *Invent. Math.* **15** (1972), 259-331.

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

[17] Serre, J.-P.: Travaux de Wiles (et Taylor, ...), Partie I. *Asterisque* **237** (1996), 319-332.

[18] Taylor, R. L., Wiles, A.: Ring theoretic properties of certain Hecke algebras. *Annals of Math.* **141** (1995), 553-572.

[19] Vojta, P.: *Diophantine Approximations and Value Distribution Theory.* Lect. Notes in Math. **1239**, 1987

[20] Wiles, A.: Modular elliptic curves and Fermat's Last Theorem. *Annals of Math.* **141** (1995), 443-551.