Jest to zapis odczytu wygłoszonego podczas XXI Szkoły Matematyki Poglądowej, *Zdarzyło się w XX wieku*. Red.

O Wielkim Twierdzeniu Fermata

Zbigniew MARCINIAK, Warszawa

Historia jest dobrze znana: studiując w dziele Diofantosa przepis na trójki liczb naturalnych x, y, z spełniające równanie Pitagorasa $x^2 + y^2 = z^2$, Pierre Fermat odnotował na marginesie, że równanie $x^n + y^n = z^n$ nie ma dla wykładników $n \geq 3$ rozwiązań w liczbach całkowitych dodatnich x, y, z. Choć miało to miejsce w XVII wieku, do bardzo niedawna nie potrafiliśmy rozstrzygnąć, czy Fermat miał rację. Dzięki pracy Andrew Wilesa, korzystającej z dorobku wielu innych uczonych, dziś wiemy, że Fermat się nie mylił. Celem tego artykułu jest przedstawienie idei leżących u podstaw dowodu Wilesa.

To też dobrze wiadomo: jeśli wykładnik n dzieli się przez d, to wystarczy udowodnić Twierdzenie Fermata dla wykładnika d, a dla n wyniknie ono automatycznie. Gdyby bowiem liczby całkowite a, b, c spełniały równość $a^n + b^n = c^n$, to liczby $a_1 = a^{n/d}$, $b_1 = b^{n/d}$, $c_1 = c^{n/d}$ będą spełniać równość $a_1^d + b_1^d = c_1^d$, co jest niemożliwe. Stąd w prosty sposób wynika, że wystarczy udowodnić Twierdzenie Fermata dla $n = 4$ (co uczynił sam Fermat) oraz dla wszystkich nieparzystych liczb pierwszych.

Poszukiwanie rozwiązań równania $x^n + y^n = z^n$ w zbiorze liczb naturalnych $\mathbb{N}$ nie jest łatwe także dlatego, że zbiór ten ma zbyt ubogą strukturę algebraiczną. Zbiór liczb całkowitych $\mathbb{Z}$ jest pod tym względem znacznie lepszy: liczby te można nie tylko dodawać i mnożyć, ale także odejmować. Dla liczb całkowitych, Problem Fermata można sformułować jak następuje:

$$\text{jeżeli } x^n + y^n = z^n \text{ i } x, y, z, n \in \mathbb{Z} \text{ oraz } n \geq 3, \text{ to } xyz = 0.$$

Jeszcze prościej można go sformułować dla ciała $\mathbb{Q}$ liczb wymiernych, które oprócz dodawania, mnożenia i odejmowania można także dzielić:

$$\text{jeżeli } X^n + Y^n = 1 \text{ i } X, Y \in \mathbb{Q} \text{ oraz } n \geq 3, \text{ to } XY = 0.$$

Zauważmy, że w ostatnim sformułowaniu występują tylko dwie niewiadome: X i Y, co pozwala przedstawić zbiór opisany równaniem $X^n + Y^n = 1$ na płaszczyźnie. Dowodząc twierdzenia Fermata mamy wykazać, że jedynymi punktami tej krzywej o obu współrzędnych wymiernych są punkty przecięcia krzywej z osiami układu współrzędnych.

Dokładnie analizując własności tych krzywych, Gerd Faltings potrafił w 1983 r. wykazać, że posiadają one skończenie wiele punktów wymiernych. Znaczy to, że równanie Fermata ma co najwyżej skończenie wiele rozwiązań w liczbach naturalnych. Był to ogromny postęp, ale w ciągu następnych 10 lat nikt nie potrafił poprawić tego wyniku i „dokończyć" dowodu Wielkiego Twierdzenia Fermata.

Matematycy od dawna badali własności krzywych, zwłaszcza tych opisanych równaniem niskiego stopnia. Na przykład, krzywe opisane równaniem stopnia drugiego to *stożkowe*. Do najlepiej poznanych krzywych stopnia trzeciego należą krzywe opisane równaniem postaci $y^2 = x^3 + ax + b$.

Potrzeba badania tych krzywych wynikła z problemów, jakie napotykamy przy obliczaniu całek, w których występuje funkcja $\sqrt{x^3 + ax + b}$. Takie całki pojawiają się dość często w mechanice i geometrii, np. wtedy, gdy chcemy obliczyć długość łuku elipsy. Dlatego całki tego rodzaju nazywamy *całkami eliptycznymi*, krzywe zaś opisane równaniem $y^2 = x^3 + ax + b$ – *krzywymi eliptycznymi*.

Na potrzeby dowodu Twierdzenia Fermata to ostatnie pojęcie nieco zawęzimy: krzywą eliptyczną E nazwiemy zbiór opisany równaniem $y^2 = x^3 + ax + b$, gdzie liczby a, b są obie całkowite, a wielomian $w(x) = x^3 + ax + b$ ma trzy różne pierwiastki (w ciele liczb zespolonych $\mathbb{C}$).

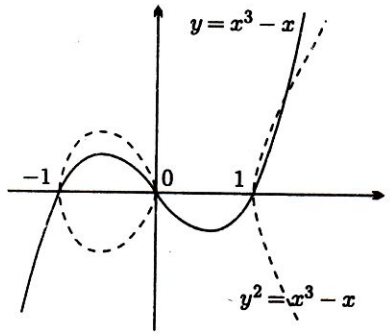
W istocie, powyższe równanie opisuje wiele zbiorów: wszystko zależy od tego, na „której" płaszczyźnie go narysujemy. Zamiast znanej ze szkoły płaszczyzny euklidesowej $\mathbb{R}^2$ możemy bowiem rozważyć płaszczyznę k^2 nad dowolnym ciałem k. Zamiast ciała liczb rzeczywistych $k = \mathbb{R}$ możemy wziąć $k = \mathbb{Q}$, $\mathbb{C}$, $\mathbb{Q}(\sqrt{2}) = \{p + q\sqrt{2} : p, q \in \mathbb{Q}\}$, itp. Możemy się również posłużyć ciałami skończonymi: najprostsze z nich to ciała $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ reszt modulo p, gdzie p jest ustaloną liczbą pierwszą. Ponadto każde z ciał $\mathbb{F}_p$ można powiększyć do ciała $\mathbb{F}_{p^r}$, które ma dokładnie p^r elementów, dla każdego $r \geq 2$.

Ta niejednoznaczność w wyborze płaszczyzny jest korzystna: otrzymujemy dla każdego ciała k inną „emanację" naszej krzywej

$$E(k) = \{(x, y) \in k^2 : y^2 = x^3 + ax + b\},$$

związaną z tym ciałem. Pozwala nam to na spojrzenie na E z wielu różnych stron.

Przykład. Dla krzywej o równaniu $y^2 = x^3 - x$ opiszmy zbiory $E(\mathbb{F}_5)$, $E(\mathbb{R})$ i $E(\mathbb{C})$.



Rys. 1

Aby opisać zbiór $E(\mathbb{F}_5)$, sporządzamy tabelki kwadratów i sześcianów w ciele $\mathbb{F}_5$:

y	0	1	2	3	4
y^2	0	1	4	4	1

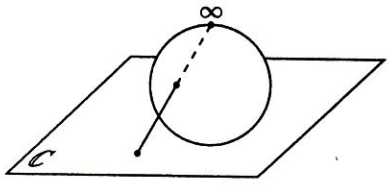
x	0	1	2	3	4
x^3	0	1	2	3	4

Porównując zawartości tych tabelek, otrzymujemy

$$E(\mathbb{F}_5) = \{(x, y) \in \mathbb{F}_5^2 : y^2 = x^3 - x\} =$$
$$= \{(0,0), (1,0), (4,0), (2,1), (2,4), (3,2), (3,3)\}.$$

Aby narysować zbiór $E(\mathbb{R})$, rozpoczynamy od wykresu pomocniczej funkcji $y = x^3 - x$. Następnie dla tych x, które spełniają nierówność $x^3 - x \geq 0$ pierwiastkujemy wartość y i odkładamy ją nad i pod osią poziomą (rys. 1).

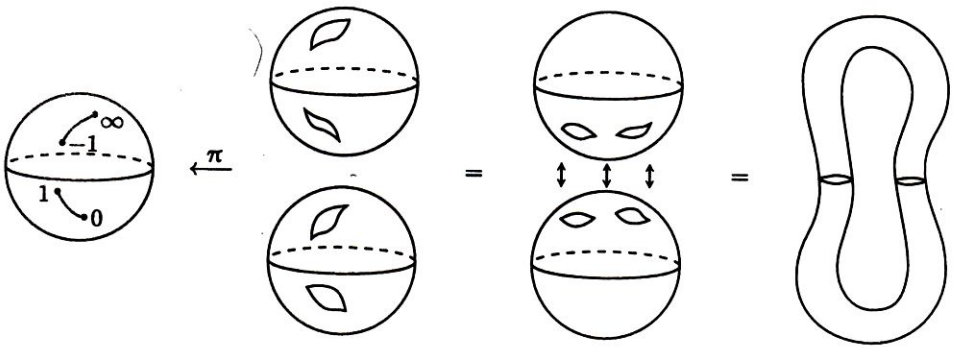
Otrzymany rysunek składa się z dwóch składowych: krzywej zamkniętej i nieograniczonego łuku. Każdą krzywą $E(k)$ korzystnie jest domknąć (w płaszczyźnie rzutowej), uzupełniając ją jednym punktem „w nieskończoności". Domkniętą krzywą dalej także będziemy oznaczać $E(k)$. Dla ciała $k = \mathbb{R}$ nieograniczony łuk stanie się częścią drugiej krzywej zamkniętej. Zatem (topologicznie) $E(\mathbb{R})$ jest sumą dwóch rozłącznych okręgów.



Rys. 2

Aby opisać zbiór $E(\mathbb{C})$, rozważmy płaszczyznę zespoloną $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ z dołączonym punktem w nieskończoności, którą wygodnie jest wyobrażać sobie jako sferę (rys. 2).

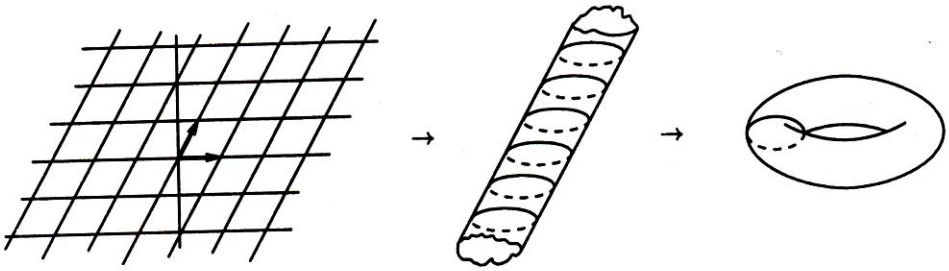
Rozważmy teraz przekształcenie $\pi : E(\mathbb{C}) \to \overline{\mathbb{C}}$, dane wzorem $\pi(x, y) = x$. Na sferze $\overline{\mathbb{C}}$ leżą cztery punkty: -1, 0, 1, ∞, na które π przekształca dokładnie po jednym punkcie $(x, y) \in E(\mathbb{C})$. Nad każdym innym punktem leży para punktów krzywej. Niech S będzie sferą $\overline{\mathbb{C}}$, z której usunięto dwa rozłączne łuki łączące pary punktów -1, 0 oraz 1, ∞. Wtedy ta część krzywej $E(\mathbb{C})$, którą π przekształca na zbiór S, wygląda jak dwie rozłączne kopie zbioru S. Dodanie brakujących łuków w $\mathbb{C} \cup \{\infty\}$ odpowiada w krzywej zlepieniu tej pary podzbiorów otworami w taki sposób, że powstanie *torus* (rys. 3).



Rys. 3

Okazuje się, że dla *dowolnej* krzywej eliptycznej E zbiór $E(\mathbb{C})$ jest torusem. Zauważmy, że przecięcie tego torusa płaszczyzną $\mathbb{R}^2 \subset \mathbb{C}^2$ daje parę okręgów $E(\mathbb{R})$.

Torus można otrzymać z płaszczyzny zespolonej $\mathbb{C}$ przez jej zwinięcie kolejno w dwóch niezależnych kierunkach (rys. 4.).



Rys. 4

Proces zwijania można opisać za pomocą pary niezależnych wektorów. Sumy i różnice wszystkich wielokrotności tych wektorów tworzą regularną kratę punktów $\Lambda \subset \mathbb{C}$, które po zwinięciu reprezentują jeden z punktów torusa. Każda krzywa $E(\mathbb{C})$ jest trochę inaczej zwinięta z płaszczyzny $\mathbb{C}$. Na przykład krzywa $y^2 = x^3 - x$ odpowiada kracie $\Lambda = \{p + qi : p, q \in \mathbb{Z}\} \subset \mathbb{C}$.

Związek kraty Λ z równaniem krzywej eliptycznej należy do klasycznej analizy i przedstawia się pokrótce jak następuje. Dla ustalonej kraty $\Lambda \subset \mathbb{C}$ rozważmy zbiór wszystkich funkcji okresowych $f : \mathbb{C} \to \mathbb{C}$, których okresy leżą w Λ: $f(z + \lambda) = f(z)$ dla wszystkich $z \in \mathbb{C}$ i $\lambda \in \Lambda$. Takie funkcje można dodawać, odejmować, mnożyć i dzielić – tworzą one ciało $\mathcal{E}_\Lambda$ *funkcji eliptycznych* związanych z kratą Λ. Można udowodnić, że każdą funkcję eliptyczną można przedstawić jako funkcję wymierną od funkcji Weierstrassa

$$\wp(z) = \frac{1}{z^2} + \sum_{\lambda \in \Lambda \setminus \{0\}} \left(\frac{1}{(z-\lambda)^2} - \frac{1}{\lambda^2} \right)$$

oraz jej pochodnej. Inaczej mówiąc, $\mathcal{E}_\Lambda = \mathbb{C}(\wp, \wp')$. W końcu, funkcje $\wp(z)$ i $\wp'(z)$ związane są relacją

$$\wp'(z) = f(\wp(z)), \text{ gdzie } f(x) = 4x^3 - g_2 x - g_3.$$

Stałe g_2, g_3 są wyznaczone przez kratę Λ za pomocą wzorów:

$$g_2 = 60 \cdot \sum_{\lambda \in \Lambda \setminus 0} \lambda^{-4}, \qquad g_3 = 140 \cdot \sum_{\lambda \in \Lambda \setminus 0} \lambda^{-6}.$$

Jest to, z dokładnością do prostego podstawienia, równanie odpowiedniej krzywej eliptycznej E. Funkcja $\mathbb{C} \to \mathbb{C}^2$, dana wzorem $z \mapsto (\wp(z), \wp'(z))$ wyznacza utożsamienie $\mathbb{C}/\Lambda \approx E(\mathbb{C}) \subset \mathbb{C}^2$.

Powyższy opis ma ważne konsekwencje. Na przykład, płaszczyzna $\mathbb{C}$ z działaniem dodawania wektorów jest grupą przemienną, krata Λ zaś jest jej podgrupą. Zatem krzywa $E(\mathbb{C})$ posiada strukturę grupy (ilorazowej) $\mathbb{C}/\Lambda$.

Podobnie jak w problemie Fermata, możemy pytać o to, ile punktów o obu współrzędnych całkowitych leży na danej krzywej eliptycznej. Zauważmy, że każdy punkt $(x, y) \in E(\mathbb{Z})$ wyznacza pewien punkt $(\overline{x}, \overline{y}) \in E(\mathbb{F}_p)$ dla każdej liczby pierwszej p: liczby $\overline{x}$, $\overline{y}$ są resztami z dzielenia x, y przez p. Warto wobec tego zainteresować się krzywymi $E(\mathbb{F}_p)$.

Zauważmy też, że jeśli k jest ciałem skończonym, to krzywa $E(k)$ składa się ze skończonej liczby punktów. Pojawia się naturalne pytanie: z ilu punktów składa się krzywa $E(\mathbb{F}_p)$? Ogólniej, z ilu punktów składa się krzywa $E(\mathbb{F}_{p^r})$ dla $r = 1, 2, \ldots$?

Niech N_r oznacza moc zbioru $E(\mathbb{F}_{p^r})$. Liczby naturalne $N_1, N_2, N_3, \ldots$ posiadają bardzo interesującą własność: jeśli użyjemy ich jako współczynników

w nieskończonym szeregu

$$Z_p(T) = \exp\Big(\sum_{r=1}^{\infty} N_r \frac{T^r}{r}\Big), \text{ gdzie } \exp(u) = e^u = \sum \frac{u^k}{k!},$$

to okazuje się, że tak otrzymana funkcja zmiennej T jest wymierna. Dokładniej, dla dowolnej krzywej eliptycznej E odpowiadająca jej funkcja jest postaci

$$Z_p(T) = \frac{1 - a_p T + pT^2}{(1-T)(1-pT)}$$

dla odpowiednio dobranej liczby całkowitej a_p. Zauważmy, że a_p jest jedynym współczynnikiem we wzorze, zależącym od naszej krzywej. Co oznacza ta liczba?

Łatwo to obliczyć, badając współczynnik przy T w szeregu $Z_p(T)$. Z jednej strony mamy

$$\begin{aligned} Z_p(T) &= \exp\Big(N_1 T + \frac{N_2}{2} T^2 + \cdots\Big) = \\ &= 1 + \frac{\big(N_1 T + \frac{N_2}{2} T^2 + \cdots\big)}{1!} + \frac{\big(N_1 T + \frac{N_2}{2} T^2 + \cdots\big)^2}{2!} + \cdots = \\ &= 1 + N_1 T + \cdots. \end{aligned}$$

Z drugiej zaś strony:

$$\begin{aligned} Z_p(T) &= \frac{1 - a_p T + pT^2}{(1-T)(1-pT)} = \\ &= (1 - a_p T + pT^2) \cdot (1 + T + T^2 + \cdots) \cdot (1 + pT + p^2 T^2 + \cdots) = \\ &= 1 + (-a_p + 1 + p)T + \cdots. \end{aligned}$$

Zatem

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

Liczba a_p jest zatem ściśle związana z liczbą punktów skończonej krzywej eliptycznej $E(\mathbb{F}_p)$. Na pierwszy rzut oka jest ona łatwa do obliczenia – porównaj nasz przykład. Jednakże dla dużych p jest to dość kłopotliwe. Co więcej, jeśli interesują nas punkty całkowite na naszej krzywej, to warto znać wartości liczb a_p dla wszystkich liczb pierwszych p. Chodzi o to, że dwa różne rozwiązania całkowite $(x, y) \in \mathbb{Z}^2$ mogą się redukować do tego samego punktu $E(\mathbb{F}_p)$ dla jednej liczby pierwszej, lecz być doskonale rozróżnialne przez inną.

Aby zapanować nad wszystkimi liczbami a_p naraz, konstruujemy tzw. *L–funkcję Hasse-Weila* krzywej E. Uzyskujemy ją, umiejętnie zlepiając funkcje $Z_p(T)$ dla różnych liczb p. Dokładniej,

$$L_E(s) = \frac{\zeta(s)\zeta(s-1)}{\prod_p Z_p(p^{-s})} \quad \text{dla} \quad s \in \mathbb{C},\ \mathrm{Re}(s) > 1.$$

W powyższym wzorze $\zeta(s) = \prod_p \frac{1}{1-p^{-s}}$ jest zeta-funkcją Riemanna. Zauważmy, że

$$\zeta(s-1) = \prod_p \frac{1}{1 - p \cdot p^{-s}} \quad \text{oraz} \quad Z_p(p^{-s}) = \frac{1 - a_p p^{-s} + p \cdot p^{-2s}}{(1-p^{-s})(1-p\cdot p^{-s})},$$

czyli licznik funkcji $L_E(s)$ jest zaprojektowany tak, by pochłonąć mało dla nas interesujące mianowniki funkcji Z_p.

Nietrudno zauważyć, że

$$\zeta(s) = \prod_p \frac{1}{1-p^{-s}} = \prod_p \Big(1 + \Big(\frac{1}{p}\Big)^s + \Big(\frac{1}{p^2}\Big)^s + \Big(\frac{1}{p^3}\Big)^s + \cdots\Big) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

gdyż każda liczba naturalna ma (jednoznaczny) rozkład na iloczyn potęg liczb pierwszych.

Postępując podobnie, otrzymujemy wzór

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

gdzie współczynniki a_n, o numerkach n będących liczbami pierwszymi, pokrywają się z rozważanymi wcześniej liczbami a_p. Chcielibyśmy umieć prosto wyznaczać współczynniki a_n szeregu L_E.

Sposób na wyznaczanie liczb a_n można znaleźć w teorii *form modularnych*. Są to funkcje zespolone, określone na górnej półpłaszczyźnie $H = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ i posiadające dość szczególne własności.

Aby opisać te własności należy przypomnieć, że na półpłaszczyźnie H działa grupa macierzy

$$SL_2(\mathbb{Z}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} : a, b, c, d \in \mathbb{Z}, ad - bc = 1 \right\}$$

w taki sposób, że macierz $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ przekształca punkt z na punkt $g(z) = \frac{az+b}{cz+d}$. Łatwo obliczyć, że $\mathrm{Im}(g(z)) = \mathrm{Im}(z)/|cz + d|^2$. Stąd wynika, że jeśli $z \in H$, to także $g(z) \in H$.

Wraz z grupą $SL_2(\mathbb{Z})$ działają na zbiorze H wszystkie jej podgrupy. Szczególnie interesujące dla nas będą grupy

$$\Gamma(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL_2(\mathbb{Z}) : N \mid c \right\},$$

określone dla wszystkich liczb naturalnych $N \in \mathbb{N}$.

Funkcję $f : H \to \mathbb{C}$ nazwiemy *formą modularną wagi 2 dla grupy* $\Gamma(N)$, jeśli jest meromorficzna na H, holomorficzna w nieskończoności oraz spełnia warunek

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^2 f(z) \text{ dla dowolnej macierzy } \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \Gamma(N).$$

Ten ostatni warunek gwarantuje, że forma różniczkowa $f(z)dz^2$ jest niezmiennicza względem działania grupy $\Gamma(N)$.

Zauważmy, że dla macierzy $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \Gamma(N)$ powyższy warunek przyjmuje postać $f(z+1) = f(z)$, a zatem każda forma modularna jest funkcją okresową o okresie 1. Skoro tak, to można ją rozwinąć w szereg Fouriera:

$$f(z) = \sum a_n e^{2\pi i z}.$$

Założenie, że f jest holomorficzna w nieskończoności implikuje, że $a_n = 0$ dla $n < 0$. Jeśli ponadto zachodzi równość $a_0 = 0$, to f nazywamy *formą paraboliczną* wagi 2 dla grupy $\Gamma(N)$.

Takie formy tworzą przestrzeń wektorową $S(N)$, na której działa pewna algebra $\mathcal{H}$, zwana *algebrą operatorów Hekkego*. Jeśli „wektor" $f \in S(N)$ (tj. pewna forma paraboliczna) nie porusza się pod działaniem żadnego elementu $a \in \mathcal{H}$, to współczynniki a_n jego szeregu Fouriera można bardzo łatwo wyznaczyć rekurencyjnie. Zmieniając nieco długość wektora można założyć, że $a_1 = 1$, a wtedy obowiązują wzory

$$\begin{cases} a_n = a_{np} + p \cdot a_{n/p} & \text{dla } p \nmid N, \\ a_n = a_{np} & \text{dla } p \mid N. \end{cases}$$

Chciałoby się równie łatwo wyznaczać współczynniki a_n funkcji L_E dla krzywej eliptycznej E. Wprowadźmy więc definicję:

Definicja. Krzywa eliptyczna jest *modularna*, jeśli dla pewnej liczby naturalnej N istnieje forma paraboliczna $f \in S(N)$, $f = \sum a_n e^{2\pi i z}$, której współczynniki a_n pokrywają się ze współczynnikami występującymi w szeregu $L_E(s) = \sum a_n \frac{1}{n^s}$.

Sens tej definicji jest taki: dla tych krzywych nie mamy problemu z wyznaczeniem L_E.

Następująca śmiała, lecz bardzo trudna hipoteza prosto rozprawia się z problemem wyznaczania współczynników szeregu L_E w ogólnym przypadku.

Hipoteza Taniyamy–Weila: Każda krzywa eliptyczna jest modularna.

A gdzie w tym wszystkim jest schowane Wielkie Twierdzenie Fermata? Przypuśćmy, że dla nieparzystej liczby pierwszej p potrafilibyśmy znaleźć

trzy całkowite liczby dodatnie a, b, c, które spełniają warunek $a^p + b^p = c^p$. Frey zaproponował w takiej sytuacji rozważenie krzywej eliptycznej $y^2 = x(x - a^p)(x - c^p)$. Okazało się, że krzywa Freya nie może być modularna!

Gdybyśmy wiedzieli, że hipoteza Taniyamy–Weila jest prawdziwa, to mielibyśmy dowód Twierdzenia Fermata. Istotnie, gdyby równanie $x^p + y^p = z^p$ miało rozwiązanie w liczbach naturalnych, to otrzymalibyśmy parę zdań sprzecznych:

- Krzywa Freya jest krzywą eliptyczną i nie jest modularna.
- Każda krzywa eliptyczna jest modularna.

Ta sprzeczność dowodzi, że równanie Fermata rozwiązań nie ma.

Pozostaje jeszcze jeden drobiazg: nie umiemy udowodnić hipotezy Taniyamy–Weila!

Ale jest z tego impasu sprytne wyjście. Okazuje się, że o krzywej Freya można powiedzieć także coś dobrego: jest *półstabilna*, tj. jej redukcje do wszystkich ciał $\mathbb{F}_p$ są „dobre". W przybliżeniu oznacza to, że po redukcji współczynników dostajemy równanie postaci $y^2 = w(x)$, gdzie $w(x)$ jest wielomianem stopnia trzeciego, który ma co najmniej 2 różne pierwiastki.

W końcu możemy powiedzieć, co udało się udowodnić A.Wilesowi: udowodnił on, że każda półstabilna krzywa eliptyczna jest modularna. Ponieważ krzywa Freya jest eliptyczna, półstabilna i nie jest modularna, to – istnieć nie może. Wobec tego równanie Fermata nie ma rozwiązań w liczbach całkowitych dodatnich, c.b.d.o.