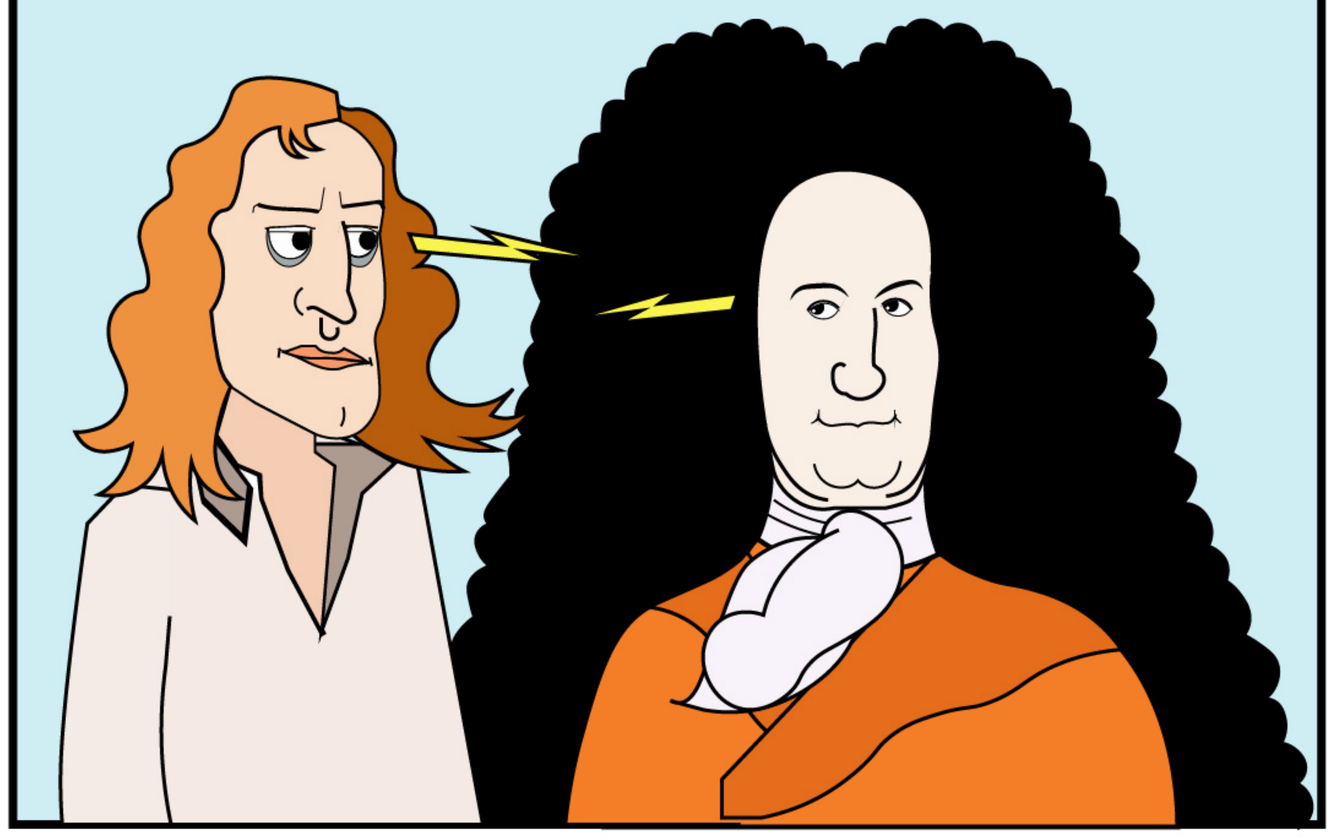
Gauss, la estadística y el planeta enano Ceres

Cuando pensamos en agrias disputas matemáticas siempre nos viene a la memoria el enfrentamiento que hubo entre Sir **Isaac Newton** y **Gottfried Wilhelm Leibniz** en el siglo XVII por la primacía en el descubrimiento del cálculo infinitesimal. En este cómic os contaremos la historia de otra rivalidad entre otros insignes matemáticos y en otro campo de las matemáticas: La **ESTADÍSTICA**.



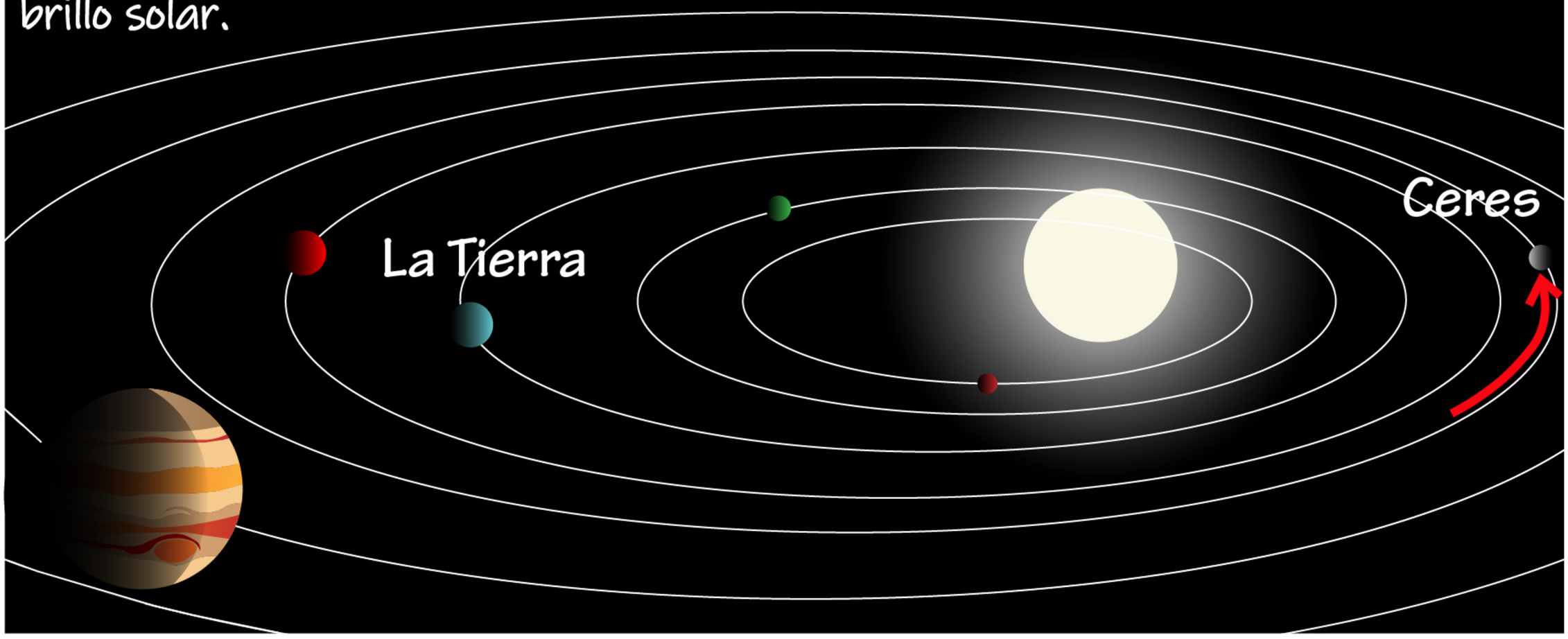
El 1 de enero de 1801, recién estrenado el siglo XIX, el astrónomo italiano **Giuseppe Piazzi** anunció el descubrimiento de un nuevo planeta que giraba alrededor del Sol en una órbita entre las de Marte y Jupiter. El planeta, al que bautizaron **CERES** como la diosa romana de la agricultura y la fecundidad, era mucho más pequeño que los siete planetas conocidos hasta entonces.

Durante 40 días, hasta la noche del 11 de febrero, Piazzi siguió al objeto en su viaje por el espacio.

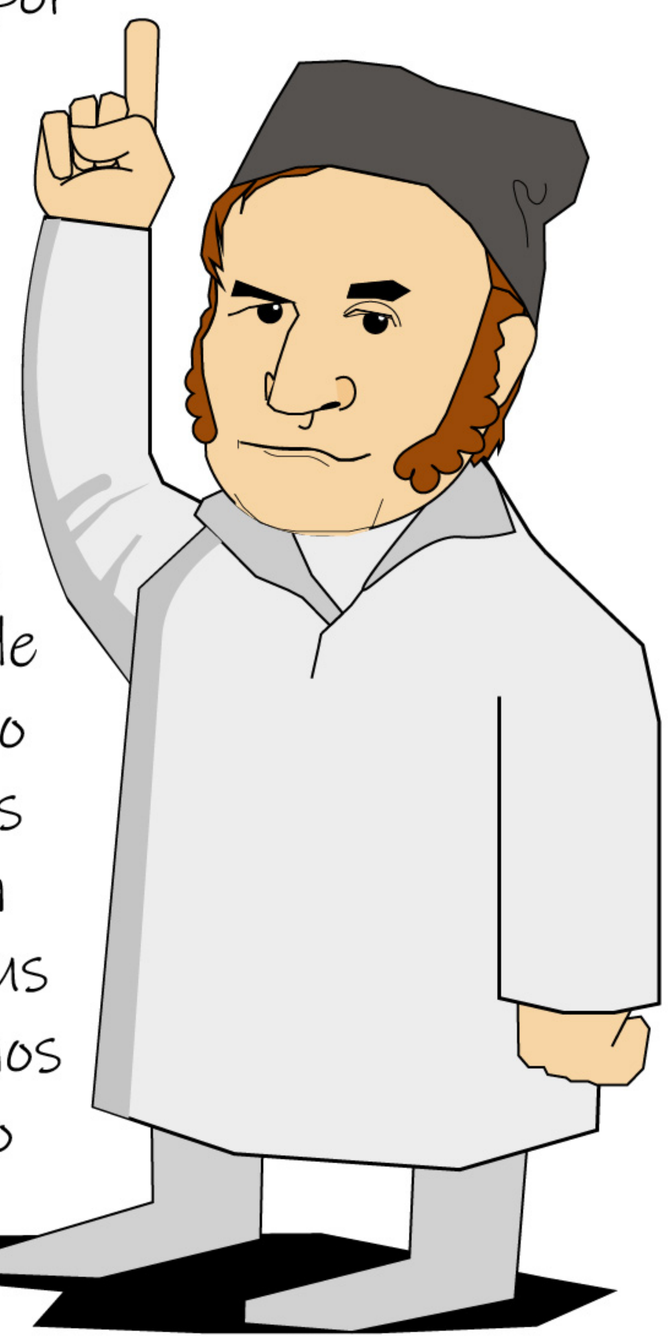
¡Ejem! tengo que recordar que yo inventé un telescopio reflector usando un espejo primario cóncavo y un espejo secundario de plano diagonal.

¡Si Newton!, pero ya se habían fabricado otros telescopios. Algunos por el mismísmo **Galileo Galilei** a quien se le ocurrió no utilizar estos artilugios para fines topográficos o militares ¡sino para observar el cielo! .

El entusiasmo del descubrimiento pronto se tornó en decepción cuando la órbita de Ceres lo condujo al otro lado del Sol, quedando oculto por el deslumbrante brillo solar.



Pero tras regresar del otro lado del Sol Ceres desapareció del cielo nocturno perdido en la inmensidad del firmamento. A pesar de disponer de Leyes que describían las órbitas y movimiento de los planetas los astrónomos del siglo XIX no disponían de suficientes herramientas matemáticas para calcular su trayectoria a partir de los escasos datos recogidos por Piazzi.



Tras los intentos infructuosos del equipo de astrónomos capitaneados por el húngaro-alemán Franz Xaver Von Zach, el 9 de diciembre, decidió observar la posición indicada por Gauss y allí, como por arte de magia, apareció un punto brillante: ¡CERES!

¿Cómo pudo saber Gauss el punto exacto en el que reaparecería el astro?

Nuestro héroe no hizo públicos hasta 1809 los procedimientos que utilizó para su predicción de la órbita.

El matemático noruego **Niels Henrik Abel** diría en una ocasión de Gauss:

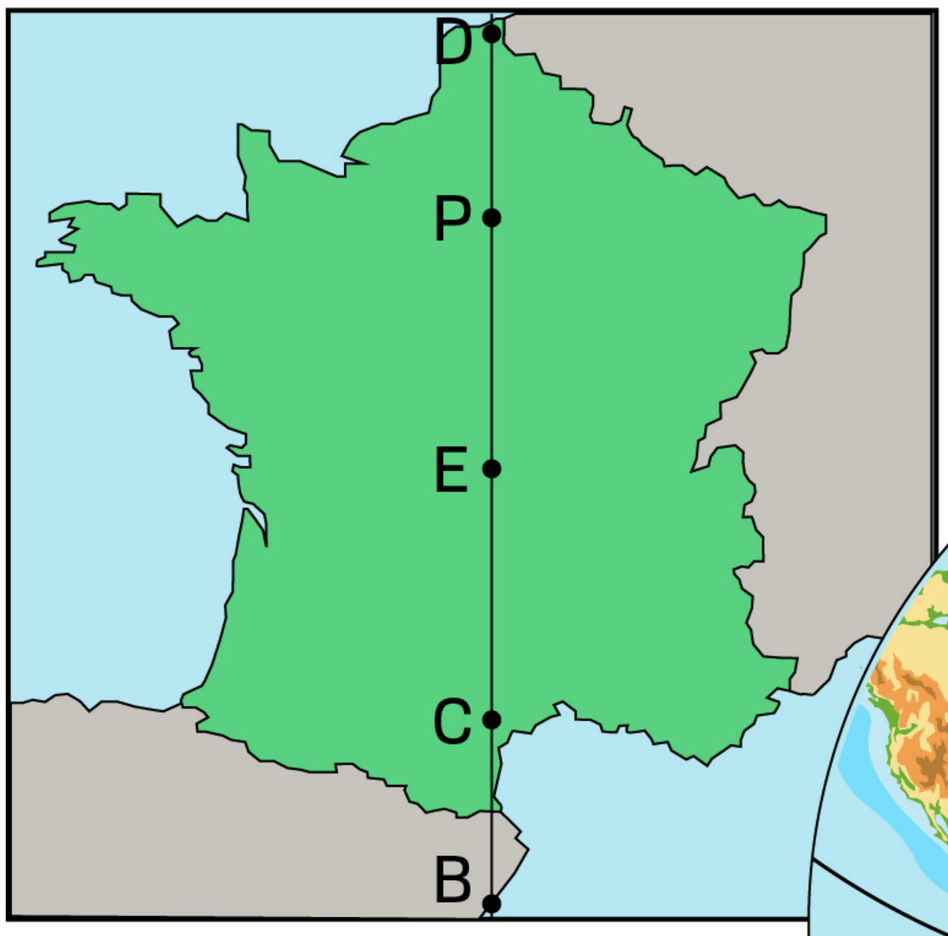
Gauss, como el zorro, borra con la cola la senda que sigue, para no dejar pista alguna de sus trabajos.

Yo escribo lento. Principalmente porque nunca estoy satisfecho hasta que he dicho todo lo posible, en pocas palabras; escribir en forma breve toma mucho más tiempo que escribir largo y tendido

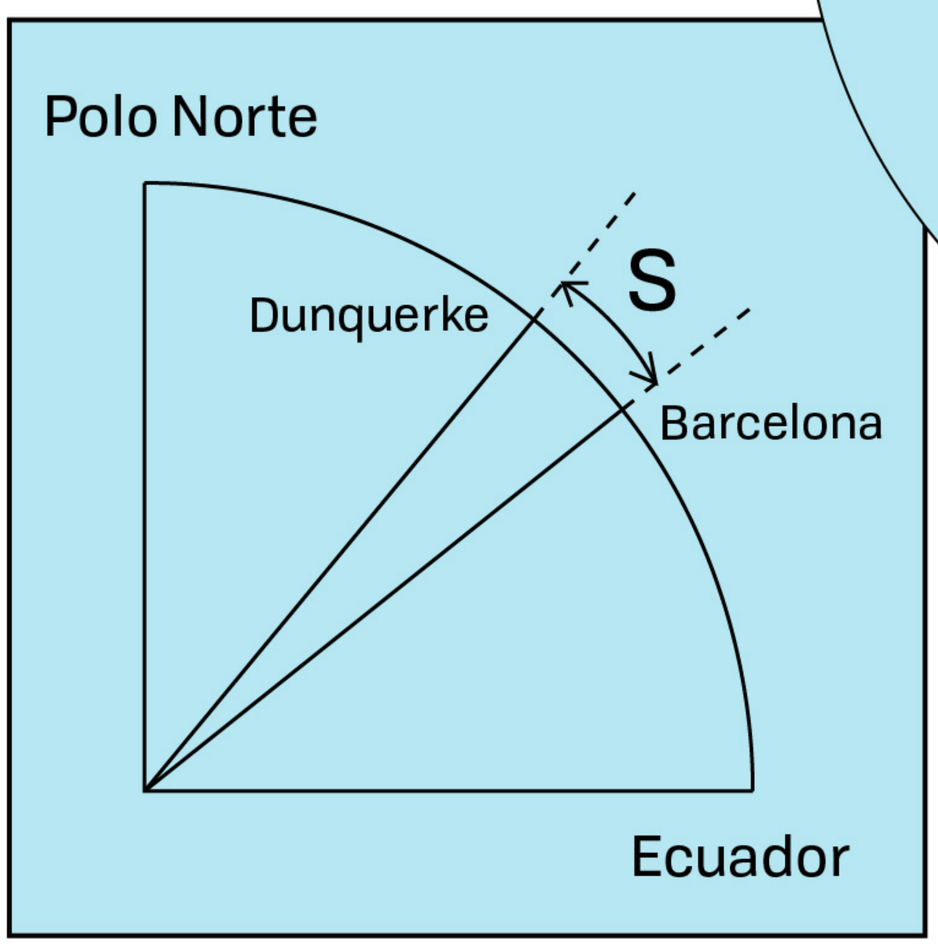
El hecho de no publicar su método tuvo dos consecuencias. Por una parte su predicción asombrosa, sin precedentes en la astronomía, tuvo un cariz sobrehumano que lanzó a Gauss al estrellato como paradigma del genio matemático.

La segunda consecuencia no fue tan positiva para Gauss, pues en 1805 el legendario matemático francés **August-Marie LEGENDRE** publicó un artículo titulado "Nouvelles méthodes pour la determination des orbites des comètes", utilizando la misma técnica que Gauss y bautizándolo como el método de los **MINIMOS CUADRADOS.**

¡¡¡ Yo soy el primero en publicar la técnica de los mínimos cuadrados y a mi me corresponde su autoría !!!



Gauss trató de probar que había usado la técnica de los mínimos cuadrados con anterioridad a 1805. En concreto en la medición del arco de meridiano que va de Dunquerke a Barcelona.



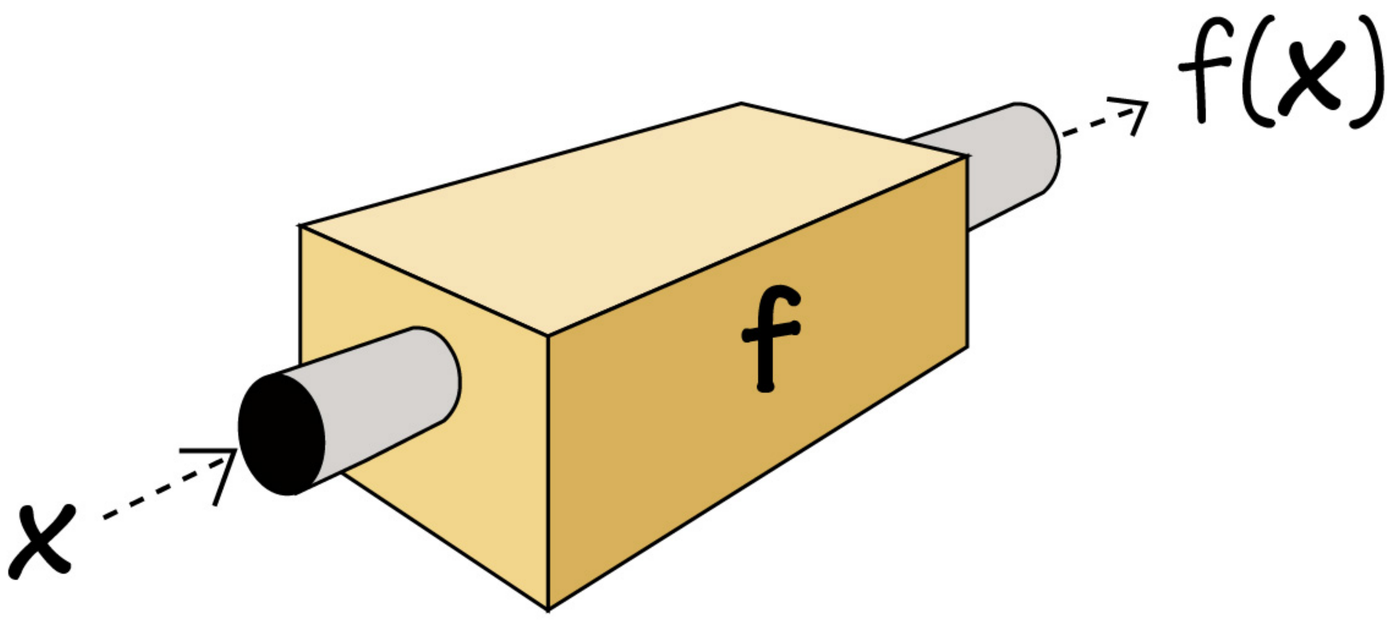
Este trabajo surgió a raíz de que la **Academia de Ciencias Francesa** decidiera en 1793 basar el sistema métrico en una unidad, **el metro**, que se definiría como la 10.000.000-ésima parte del cuadrante de meridiano.

La mala fortuna quiso que los cálculos de Gauss se perdieran y en 1831 se le sugirió la necesidad de que repitiera dichos cálculos con el fin de probar que utilizó para ellos el método de los mínimos cuadrados.
Gauss se negó categóricamente:

Archimedes' Tub 4

¿En qué consiste exáctamente el método de los MÍNIMOS CUADRADOS?

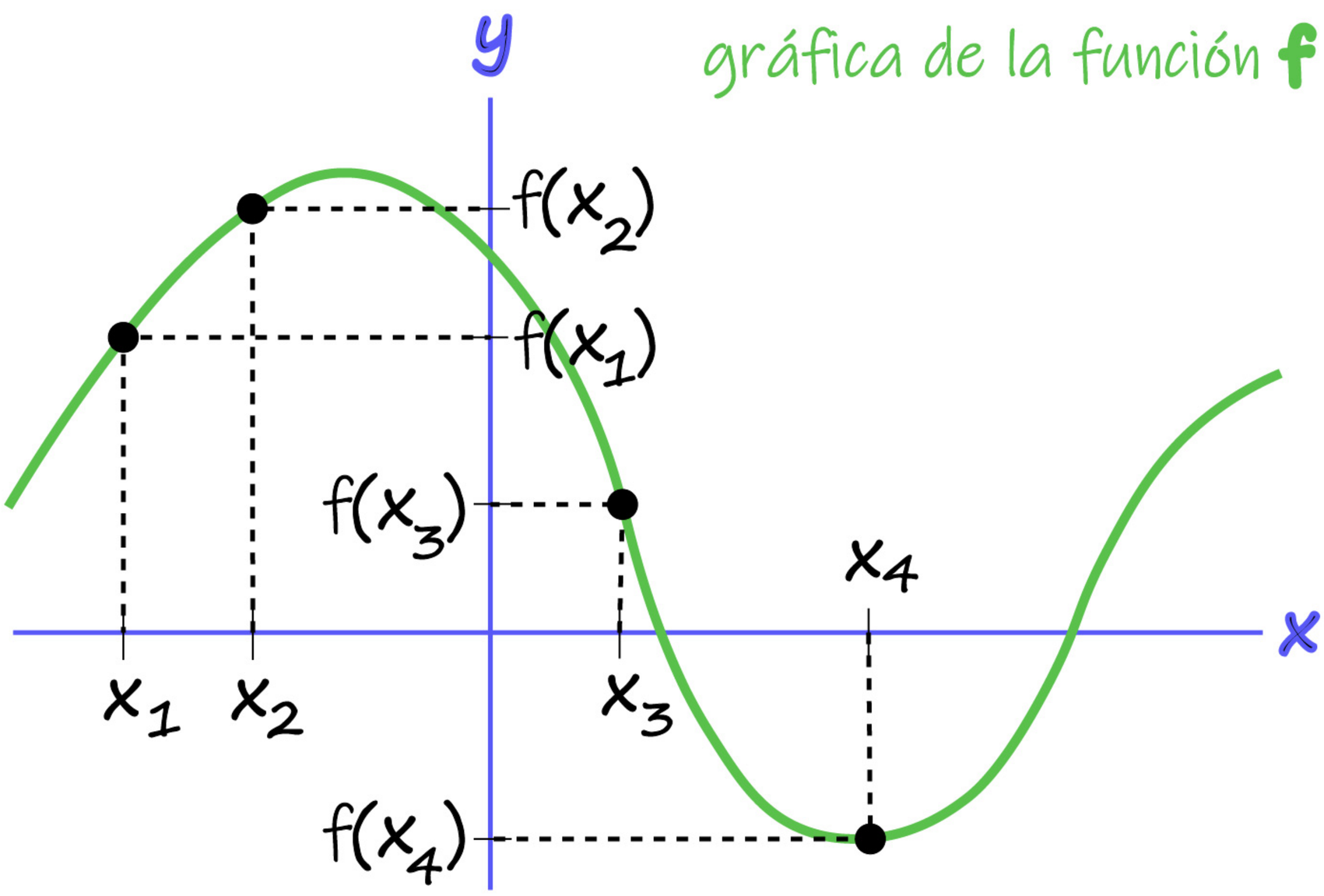
Todos estamos familiarizados con el concepto de **función**, que no es otra cosa que una regla, que denotaremos por **f**, que permite asociar a un cierto valor numérico **x** un único valor **f(x)**.



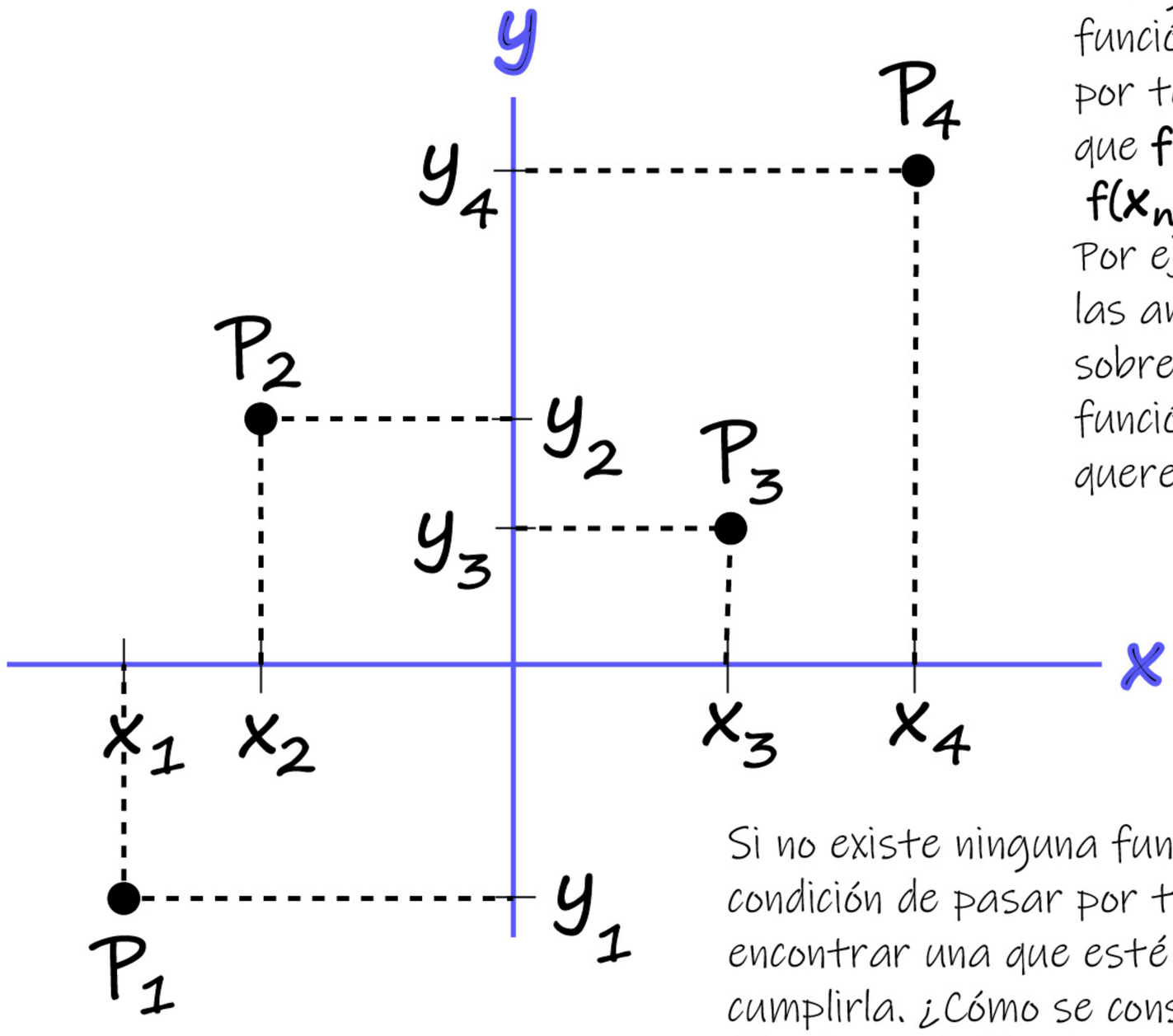
Podemos visualizar una funcion **f** como si de una "máquina" se tratase. En ella introducimos un número **x**, que tras ciertas manipulaciones se convierte en otro número **f(x)**.

Otra forma de visualizar una función es a través de su gráfica.

Si representamos en un eje horizontal. los valores **x** que se introducen en la función **f**, y en un eje vertical los valores **f(x)** que se van obteniendo, podemos dibujar cada par **(x, f(x))** como un punto en el plano. Todos estos puntos determinan la **gráfica** de la función **f**.



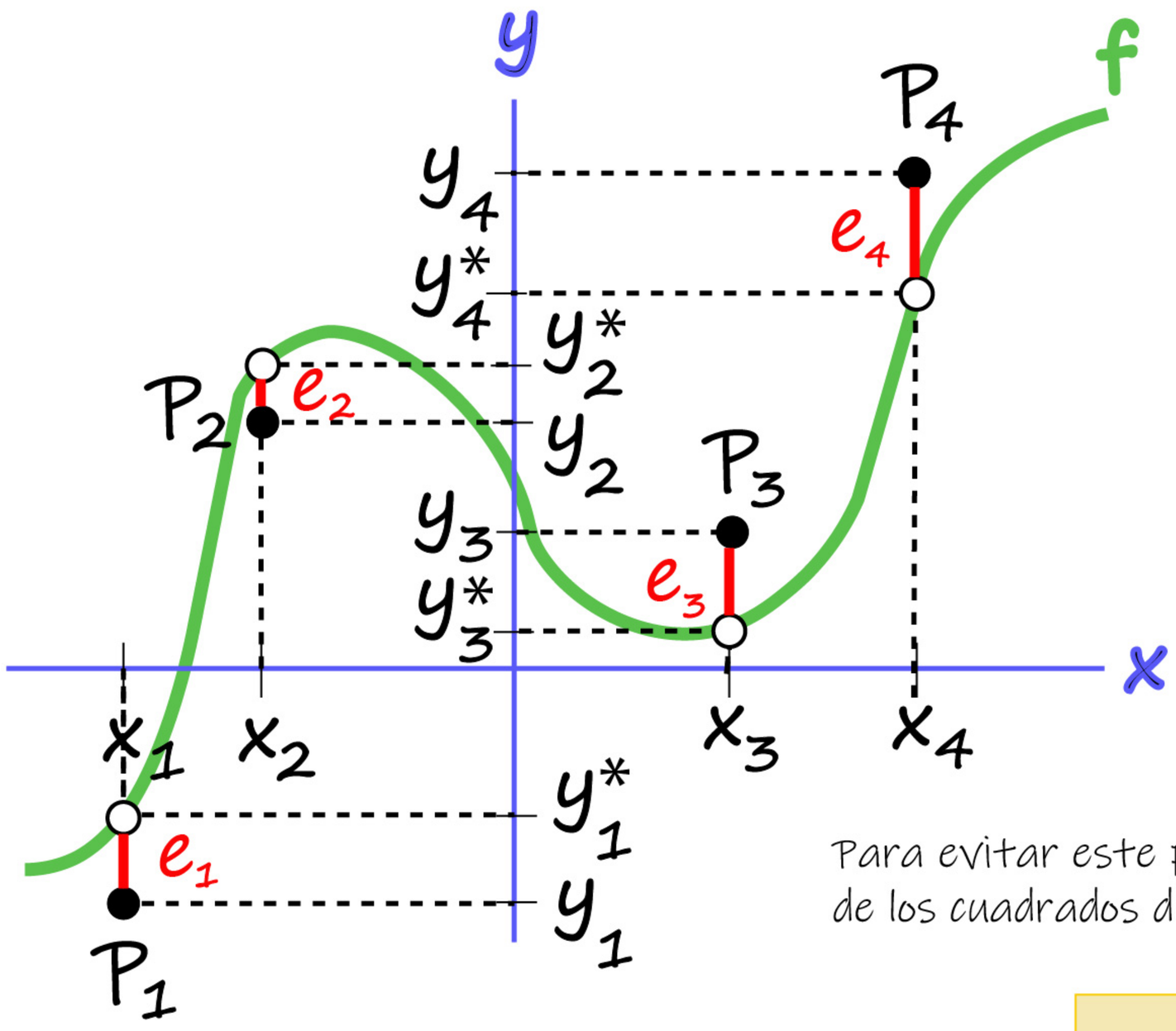
Supongamos que tenemos una colección de puntos del plano $P_1 = (x_1, y_1), P_2 = (x_2, y_2), \ldots, P_n = (x_n, y_n)$, y una determinada familia de funciones.



Nos gustaría saber si existe una función **f** de esa familia que pase por todos los puntos, esto es, tal que $f(x_1) = y_1$, $f(x_2) = y_2$, ... , $f(x_n) = y_n$.

Por ejemplo, los puntos pueden ser las anotaciones tomadas por Piazzi sobre la posición de Ceres, y la función **f** podría ser la órbita que queremos averiguar.

Si no existe ninguna función en la familia que cumpla la condición de pasar por todos los puntos, nos gustaría encontrar una que esté lo más "cerca" posible de cumplirla. ¿Cómo se consigue esto? Si tenemos que $f(x_1) = y_1^*$, $f(x_2) = y_2^*$, ... , $f(x_n) = y_n^*$. Podemos considerar el error como la diferencia $e_i = y_i^* - y_i$.



De este modo buscamos la función **f** , de la familia dada que minimize los errores e_i .

Para ello nótese que no sería muy inteligente minimizar la suma

$$\sum_{i=1}^{n} e_i$$

pues los errores pueden ser de diferente signo y darse el caso de que el sumatorio es nulo a pesar de que la función no pase por ningún punto P_i .

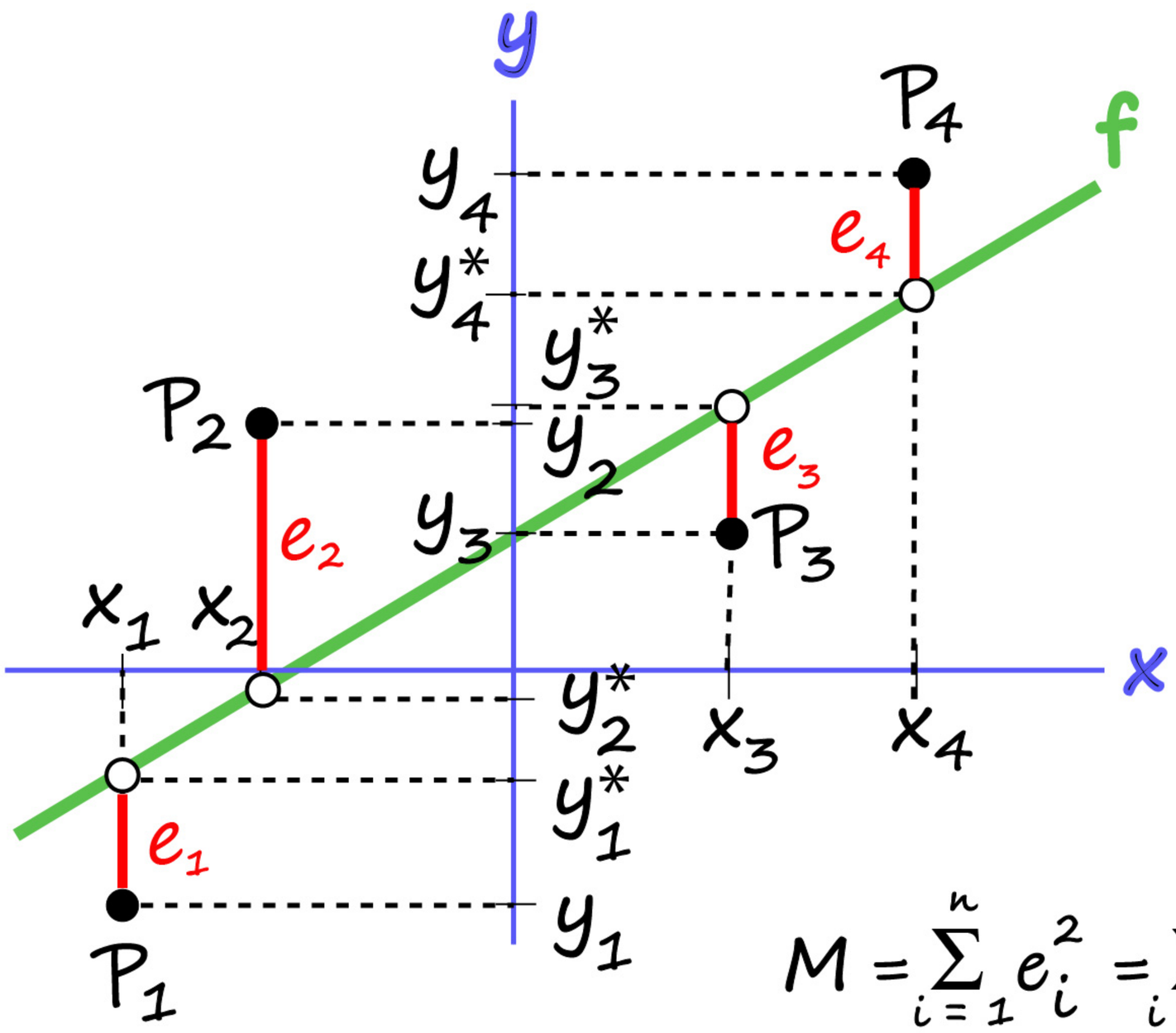
Para evitar este problema se minimizará la suma de los cuadrados de los errores, esto es:

$$M = \sum_{i=1}^{n} e_i^2$$

Archimedes' Tub 6

REGRESIÓN LINEAL

El caso particular en el que aproximamos nuestros datos utilizando una familia de funciones lineales, esto es, rectas, recibe el nombre de **Regresión Lineal,** y es de especial relevancia en el campo de la **Estadística**.



Tendríamos que las funciones de la familia son de la forma

$$f(x_i) = y_i^* = a + b\,x_i,$$

y por tanto los errores que nos aparecen son

$$e_i = a + b\,x_i - y_i.$$

Siguiendo el método de los mínimos cuadrados, la expresión que hay que minimizar es

$$M = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (a + b\,x_i - y_i)^2$$

Esta expresión depende de dos parametros a y b, por tanto, para minimizarla derivaremos M respecto de a y b, e igualaremos a 0. El sistema resultante tendrá como incógnitas a y b.

Por simplicidad en la notación los sumatorios los escribiremos simplemente como Σ.

$$\left.\begin{aligned} \frac{\partial M}{\partial a} &= \Sigma\, 2(a + b\,x_i - y_i) = 0 \\ \frac{\partial M}{\partial b} &= \Sigma\, 2(a + b\,x_i - y_i)\,x_i = 0 \end{aligned}\right\}$$

Sacando del sumatorio los factores que no tengan índices, operando y teniendo en cuenta que

$$\sum_{i=1}^{n} a = a\,n,$$

obtenemos el siguiente sistema de ecuaciones lineales:

$$\left.\begin{aligned} a\,n + b\Sigma\, x_i &= y_i \\ a\,\Sigma\, x_i + b\,\Sigma\, x_i^2 &= \Sigma\, x_i y_i \end{aligned}\right\}$$

Este sistema recibe el nombre de **Ecuaciones Normales**.

El sistema de Ecuaciones Normales podemos resolverlo fácilmente por la regla de **Cramer**. En efecto, el determinante de la matriz de coeficientes es

$$\Delta = \begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix} = n\sum x_i^2 - \left(\sum x_i\right)^2$$

Sustituyendo la primera y segunda columna de este determinante por la columna de términos independientes y cocientando por el determinante D obtenemos los valores de las incógnitas a y b respectivamente:

$$a = \frac{1}{\Delta}\begin{vmatrix} \boxed{\begin{matrix}\sum y_i \\ \sum x_i y_i\end{matrix}} & \begin{matrix}\sum x_i \\ \sum x_i^2\end{matrix} \end{vmatrix} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{1}{\Delta}\begin{vmatrix} \begin{matrix} n \\ \sum x_i \end{matrix} & \boxed{\begin{matrix}\sum y_i \\ \sum x_i y_i\end{matrix}} \end{vmatrix} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Podemos simplificar considerablemente estas expresiones para a y b utilizando las nociones de media, varianza y covarianza.
Recordemos que si x_i son los datos correspondientes a una variable aleatoria x, una medida de concentración de la variable está dada por la media

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde n es el número de datos recogidos.

Dividiendo por n^2 en el numerador y denominador de la expresión obtenida para b nos queda:

$$b = \frac{\frac{1}{n}\sum x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum x_i^2 - \overline{x}^2}$$

Si además tenemos en cuenta cómo estaban definidas las **medidas de dispersión**, todavía podemos escribir la expresión anterior de una forma más compacta.

La **varianza** de la variable aleatoria **x**, mide cómo se desvían los datos de la variable respecto de la media $\overline{x}$, y se define por la fórmula:

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 .$$

Una definición similar a la de la varianza pero mezclando la desviación de dos variables **x** e **y** es la **covarianza**, definida por la fórmula:

$$\sigma_{xy}^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{x}\,\overline{y} ,$$

vemos que la expresión para b puede escribirse en término de media, varianza y covarianza simplemente como:

$$\boxed{b = \frac{\sigma_{xy}^2}{\sigma_x^2}}$$

Finalmente, si dividimos entre **n** la primera ecuación del sistema de Ecuaciones Normales obtenemos:

$$a + b\,\overline{x} = \overline{y}$$

de donde obtenemos despejando, la fórmula para calcular a en términos de b y las medias $\overline{x}$ e $\overline{y}$:

$$\boxed{a = \overline{y} - b\,\overline{x}}$$