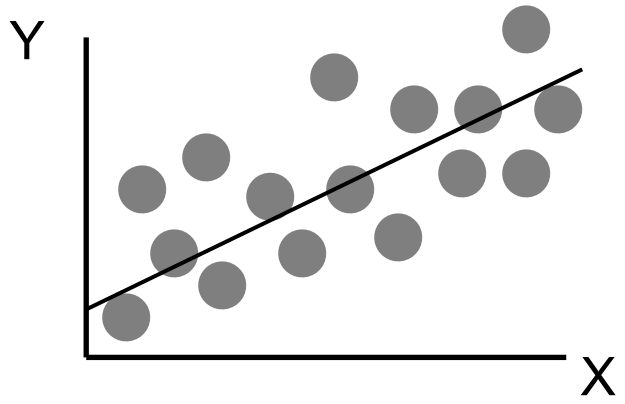
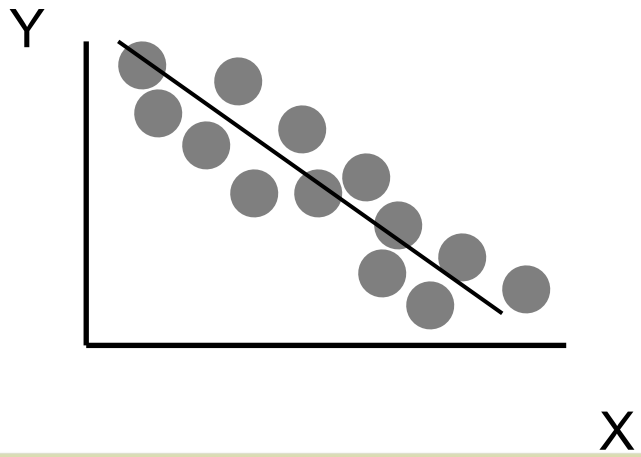
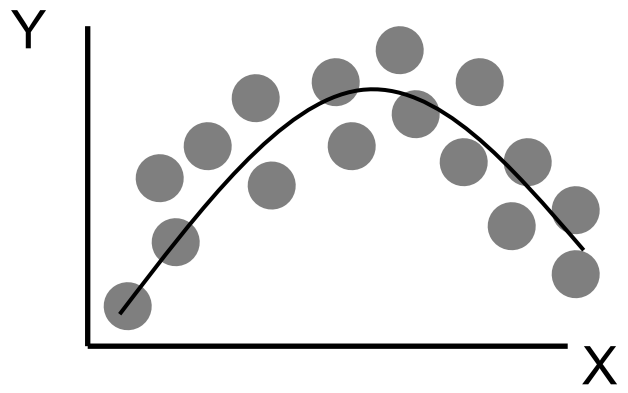
Introduction to Data Analysis in Physics

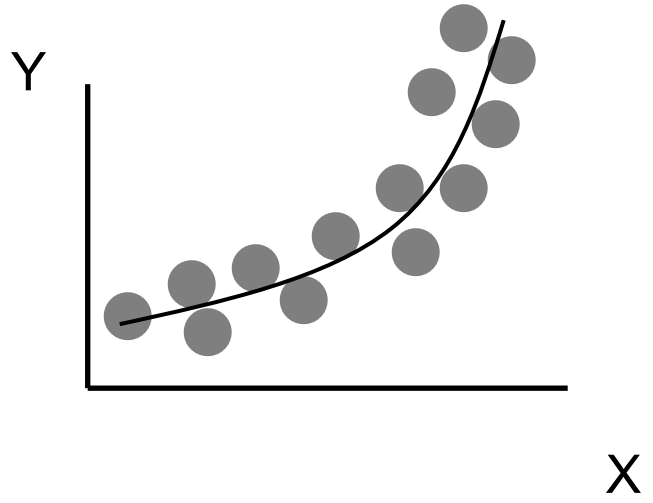
Linear relationships



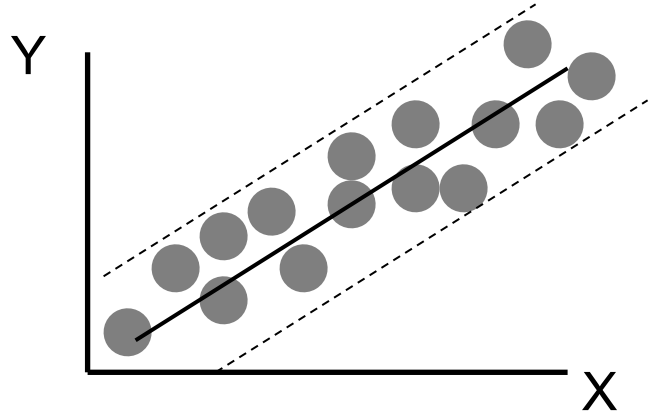


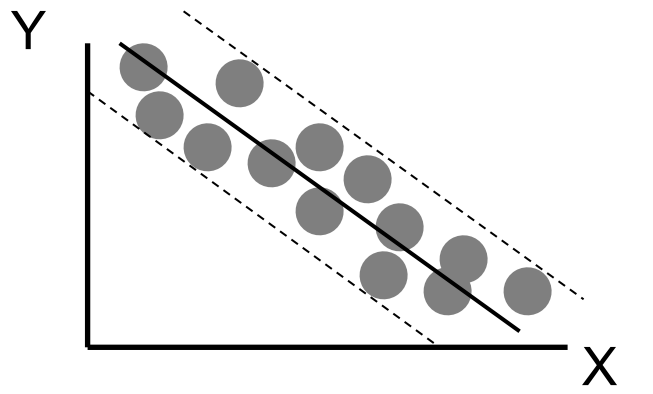
Curvilinear relationships



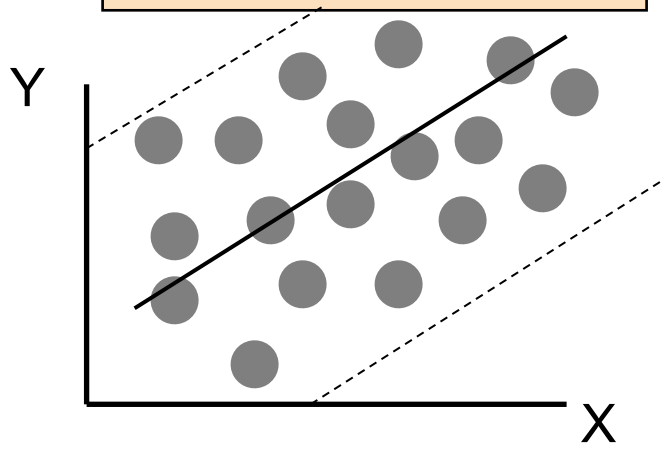


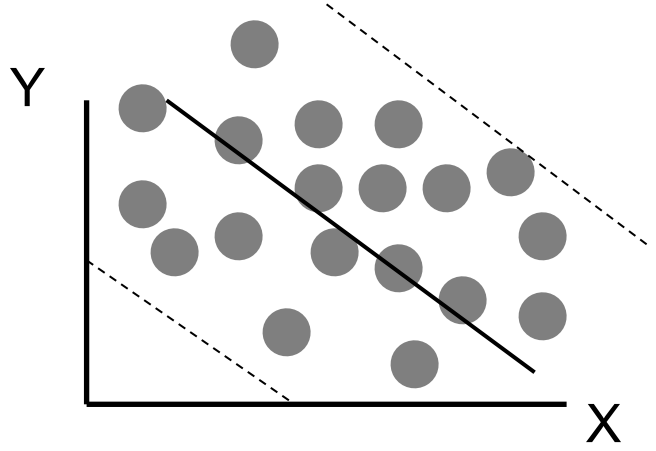
Strong relationships

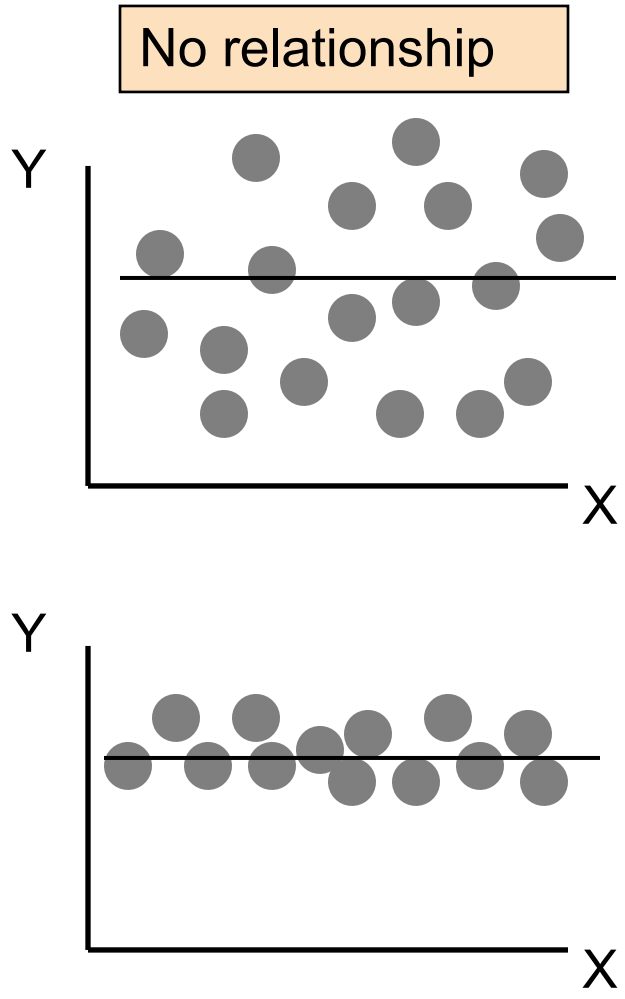




Weak relationships



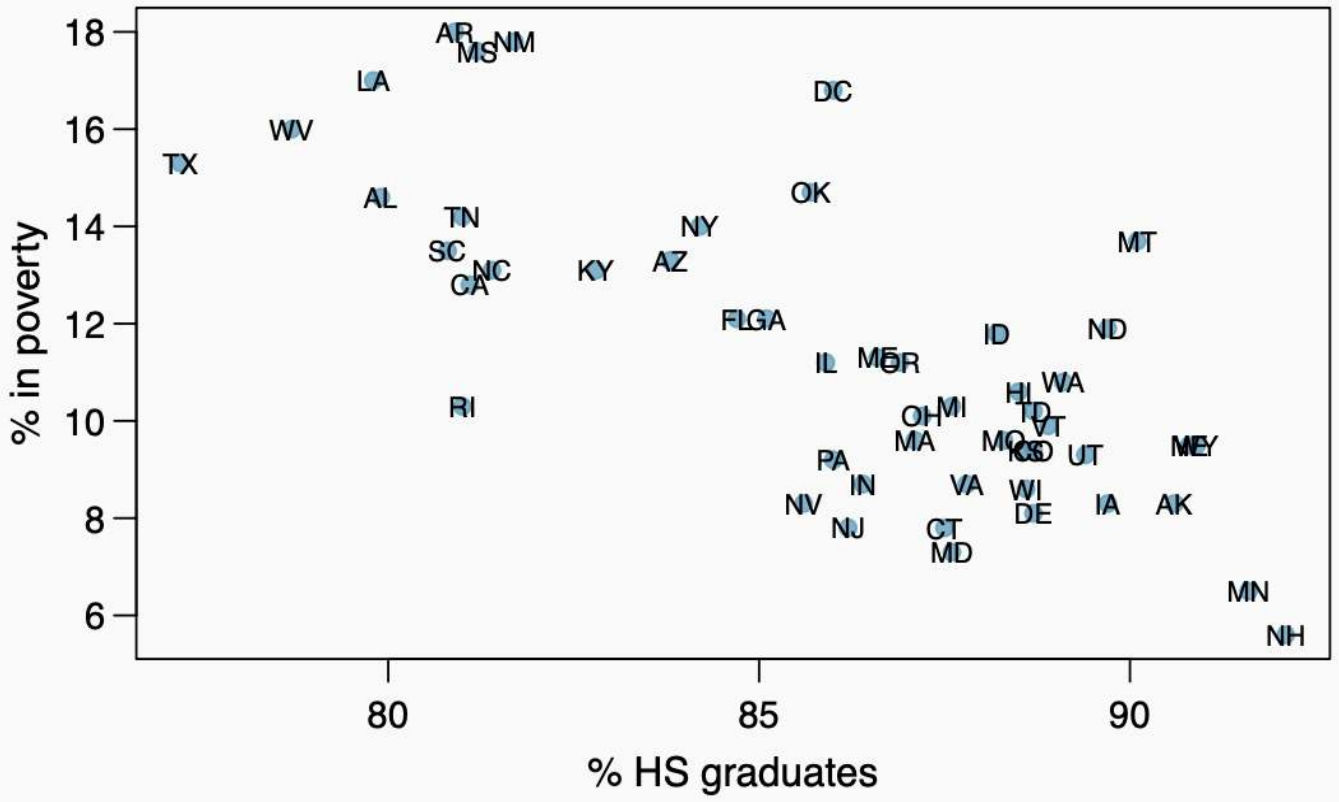



No relationship
Y
X
Y
X

How Physics Works?

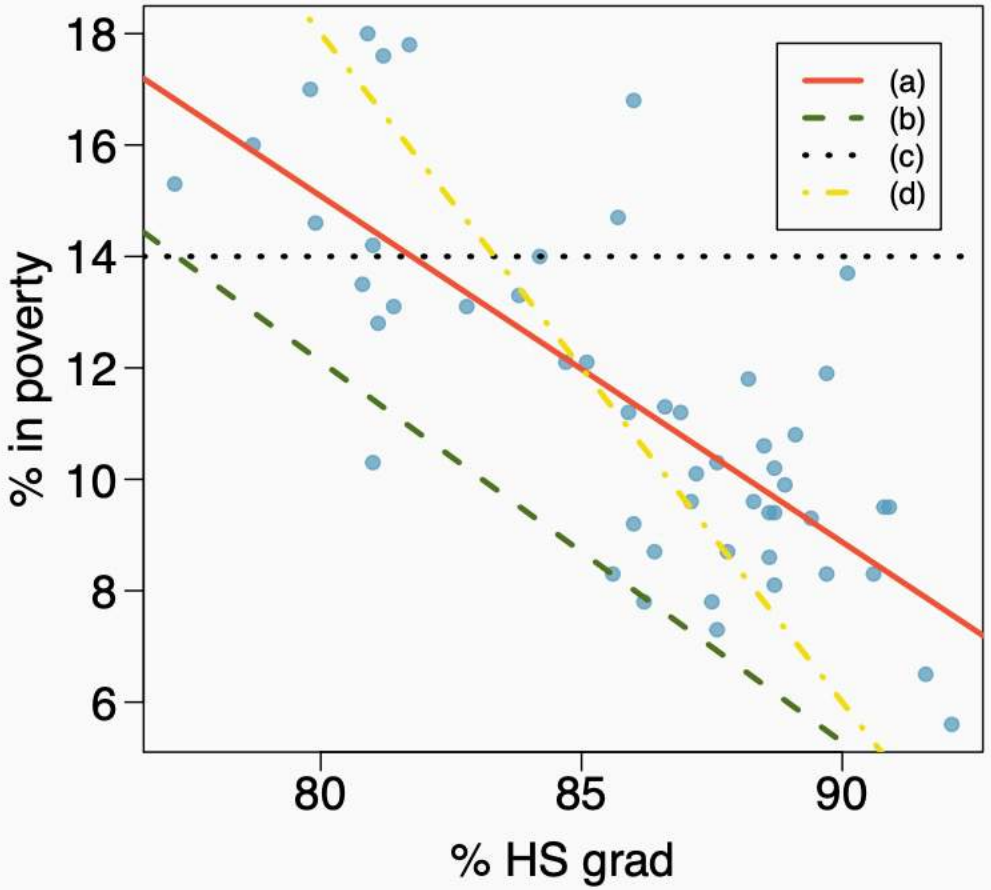
- Model laws of physics based on some fundamental principles
- Apply the model to data
- If the model correctly predicts the data, the model works
- Else find a better model!

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

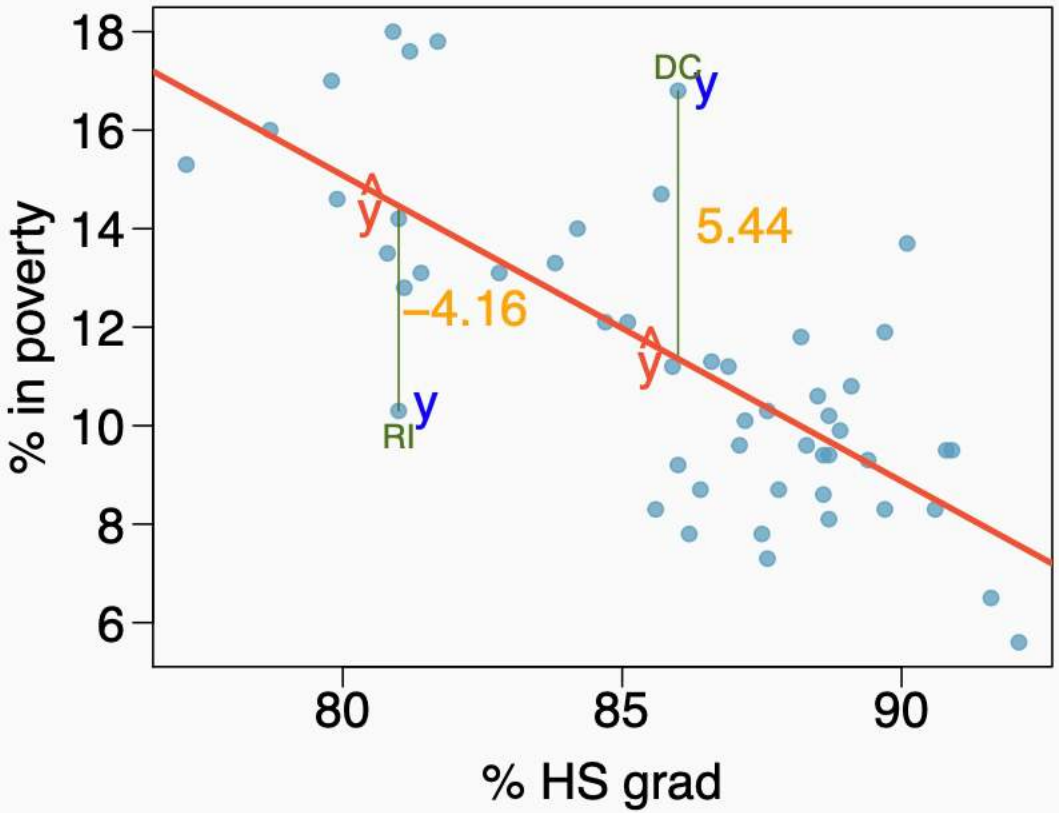


http://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L23.pdf

Which line appears to best fit the linear relationship between % in poverty and % HS grad?



http://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L23.pdf



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

http://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L23.pdf

We want a line $y = a + bx$ having small residuals:

Option 1: Minimize the sum of absolute values of residuals

$$|e_1| + |e_2| + \cdots + |e_n| = \sum_i |y_i - \widehat{y}_i| = \sum_i |y_i - a - bx_i|$$

- Difficult to compute. Nowadays possible by computer technology
- Giving less penalty to large residuals.
  The line selected is less sensitive to outliers

Option 2: Minimize the sum of squared residuals – *least square method*

$$e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

- Easier to compute by hand and using software
- Giving more penalty to large residuals.
  A residual 2x as large as another is often more than 2x as bad.
  The line selected is more sensitive to outliers

http://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L23.pdf

The *least-square (LS) regression line* is the line $y = a + bx$ that minimizes the sum of squared errors:

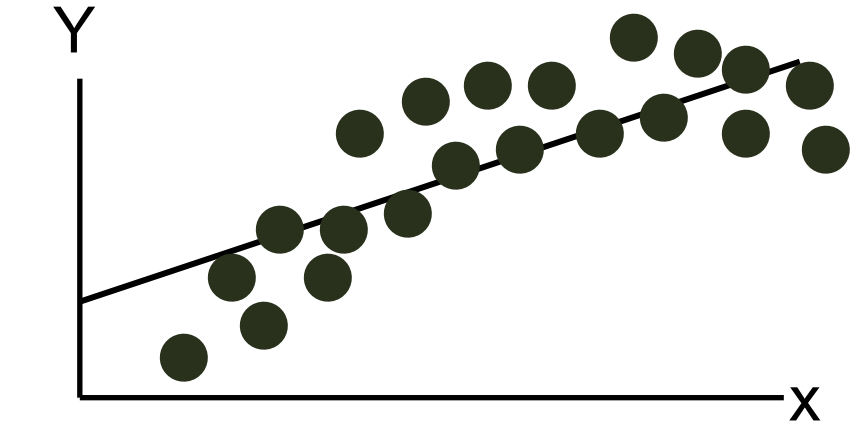
$$\sum_i e_i^2 = \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

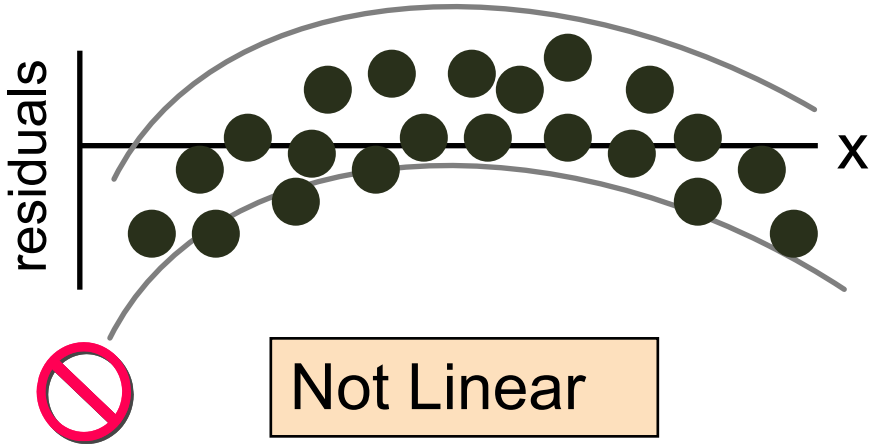
The slope and intercept of the LS regression line can be shown by math to be

$$b = \text{slope} = r \cdot \frac{s_y}{s_x}$$

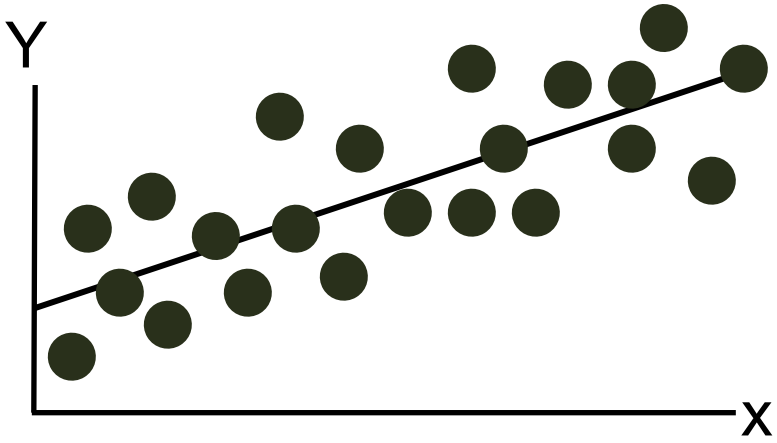
$$a = \text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

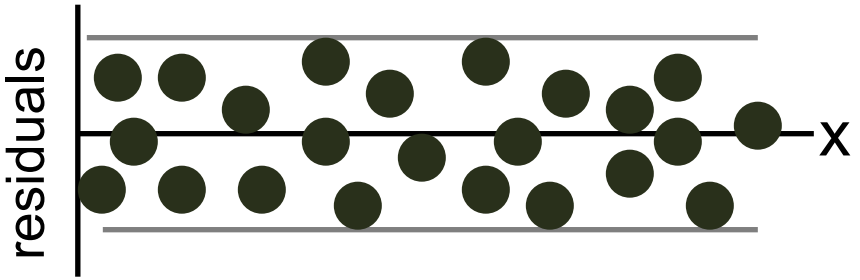
http://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L23.pdf



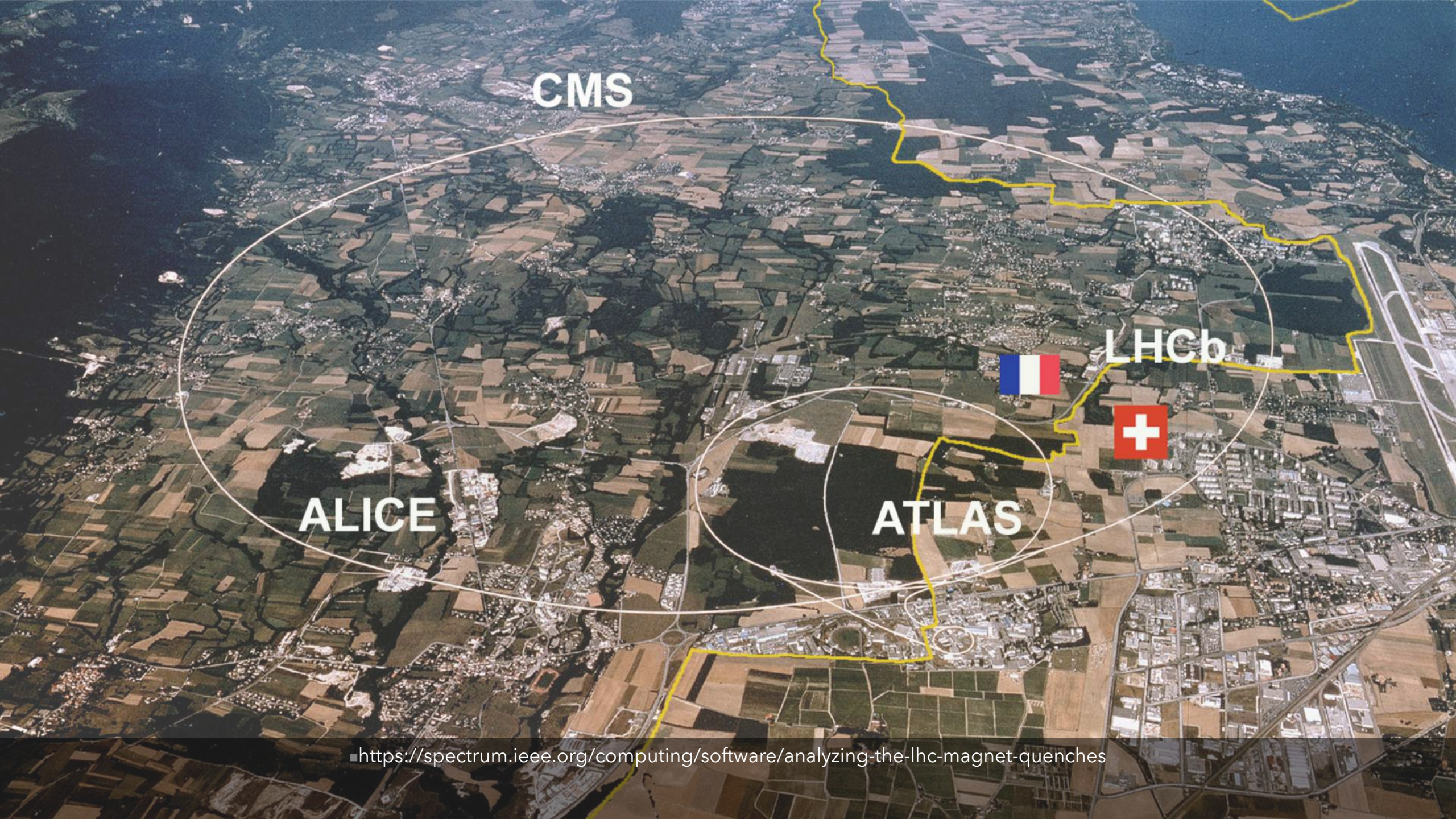


Not Linear





Linear



https://spectrum.ieee.org/computing/software/analyzing-the-lhc-magnet-quenches

Detecting New Particle at Particle Accelerators

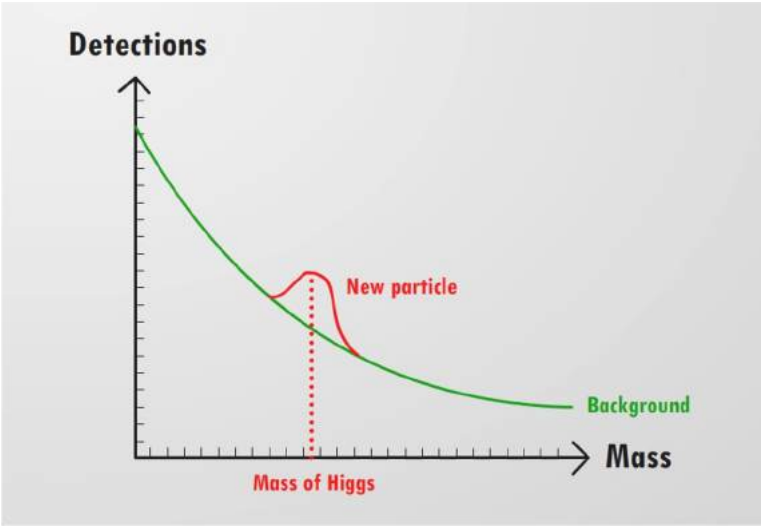
$$M = \sqrt{(E_1 + E_2)^2 - (\vec{p_1} + \vec{p_2})^2},$$

INVARIANT MASS

Invariant mass M is a value that can be calculated from the results of measurements of the detectors. Invariant mass is a mathematical concept, not a physical mass.

a particle A decays to two particles B and C. The invariant mass of the two particles B and C is determined by the equation

HIGGS 2012!



- The invariant mass can be used to examine the existence of the particle A. If particles B and C stem from the decay of the particle A, the invariant mass of them equals the physical mass of the particle A. If particles B and C stem from some other process than decay of A (there are enormous number of processes in particle collisions), the invariant mass of B and C is something else!

$$M = \sqrt{(E_1 + E_2)^2 - (\vec{p_1} + \vec{p_2})^2},$$

https://higgshunters.wordpress.com/2014/12/01/the-higgs-boson-part-2-discovery-of-the-higgs/