CDSF07, CDSF08

Recap III

DS Academy

2018

Elon Musk: 'A.I. will make jobs kind of pointless' — so study this

Published Thu, Aug 29 2019·10:47 AM EDT • Updated Fri, Aug 30 2019·10:54 AM EDT

CNBC

Bias and Regression

Variables

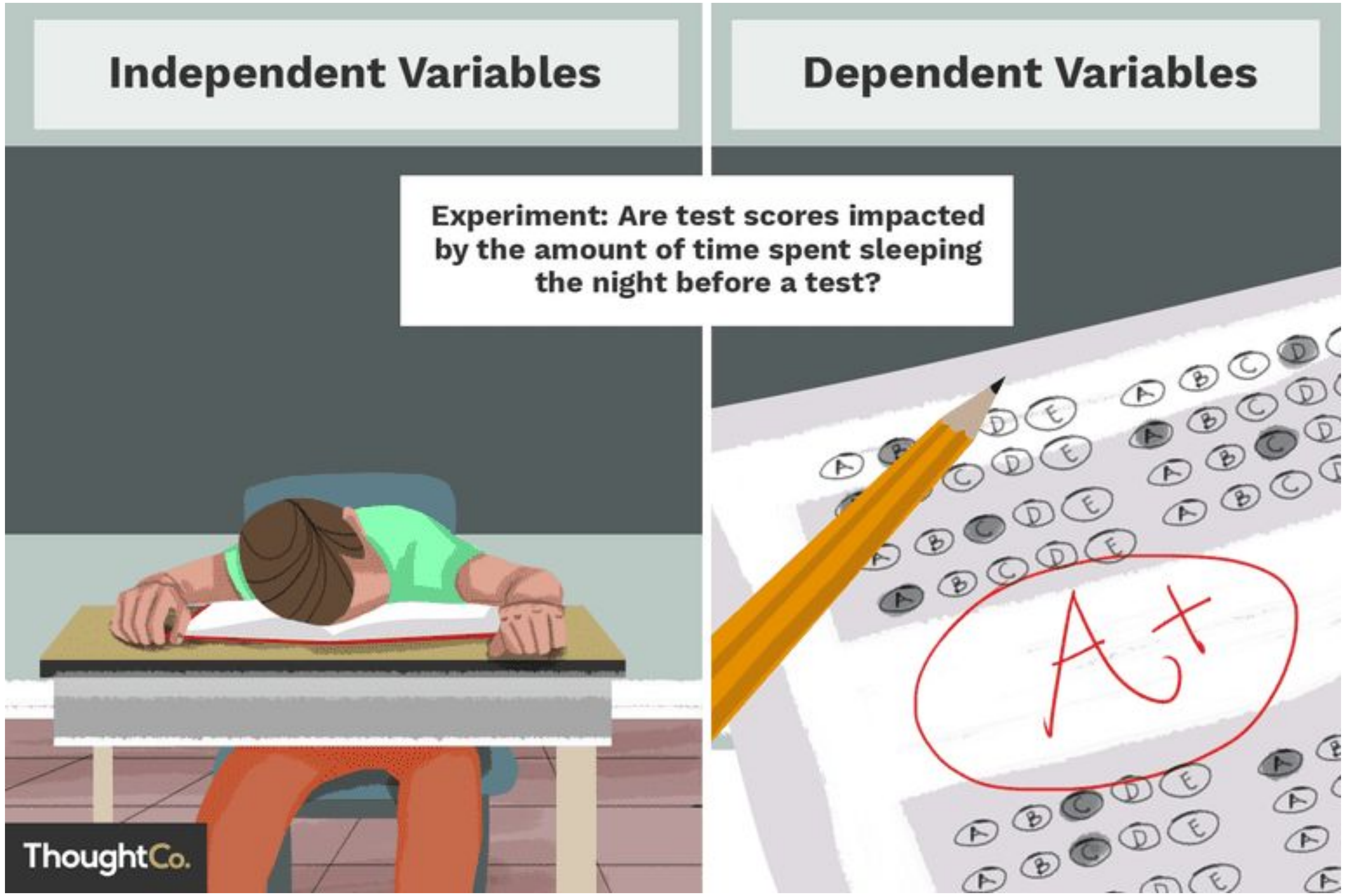


Image from ThoughtCo.

Variables

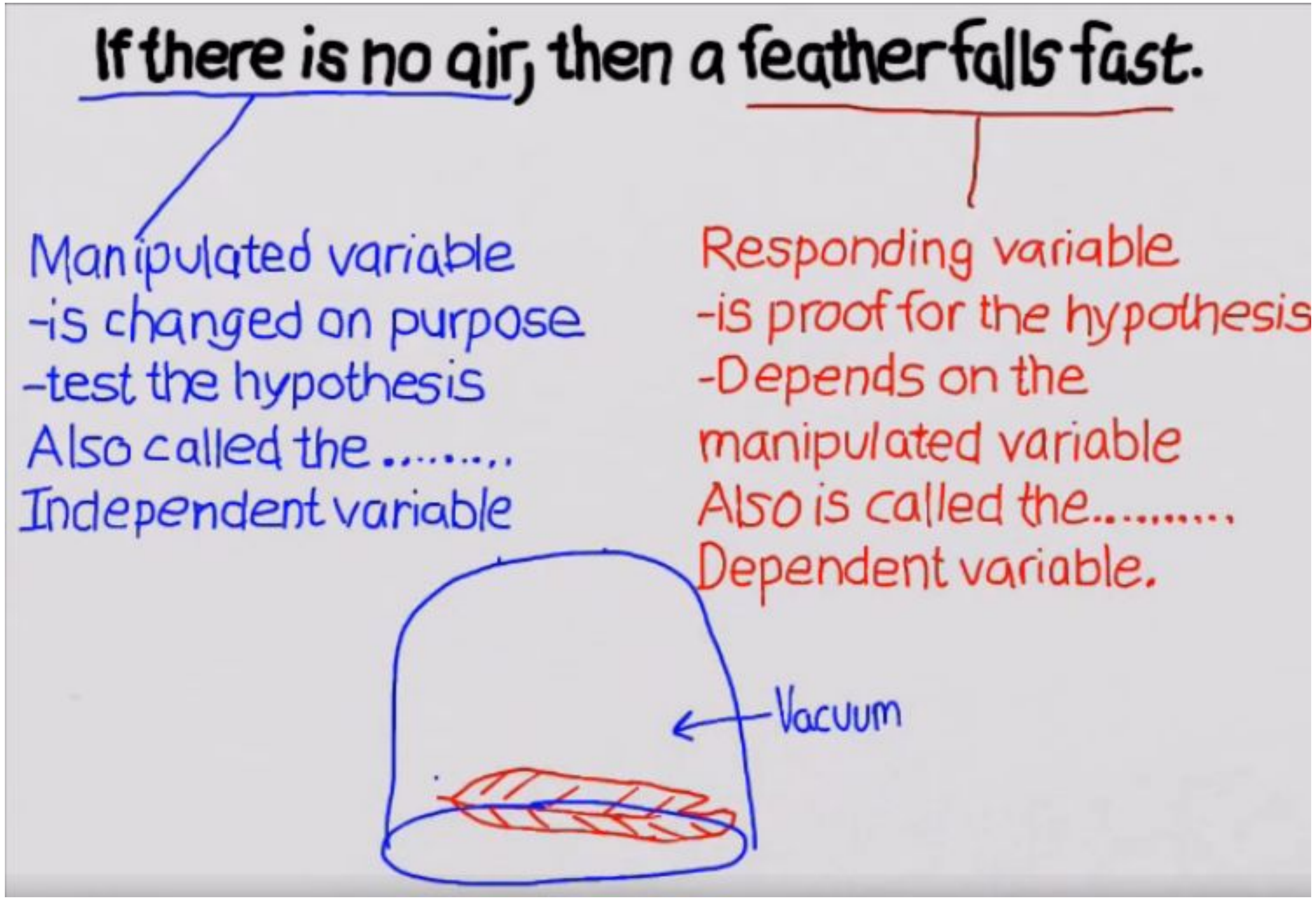
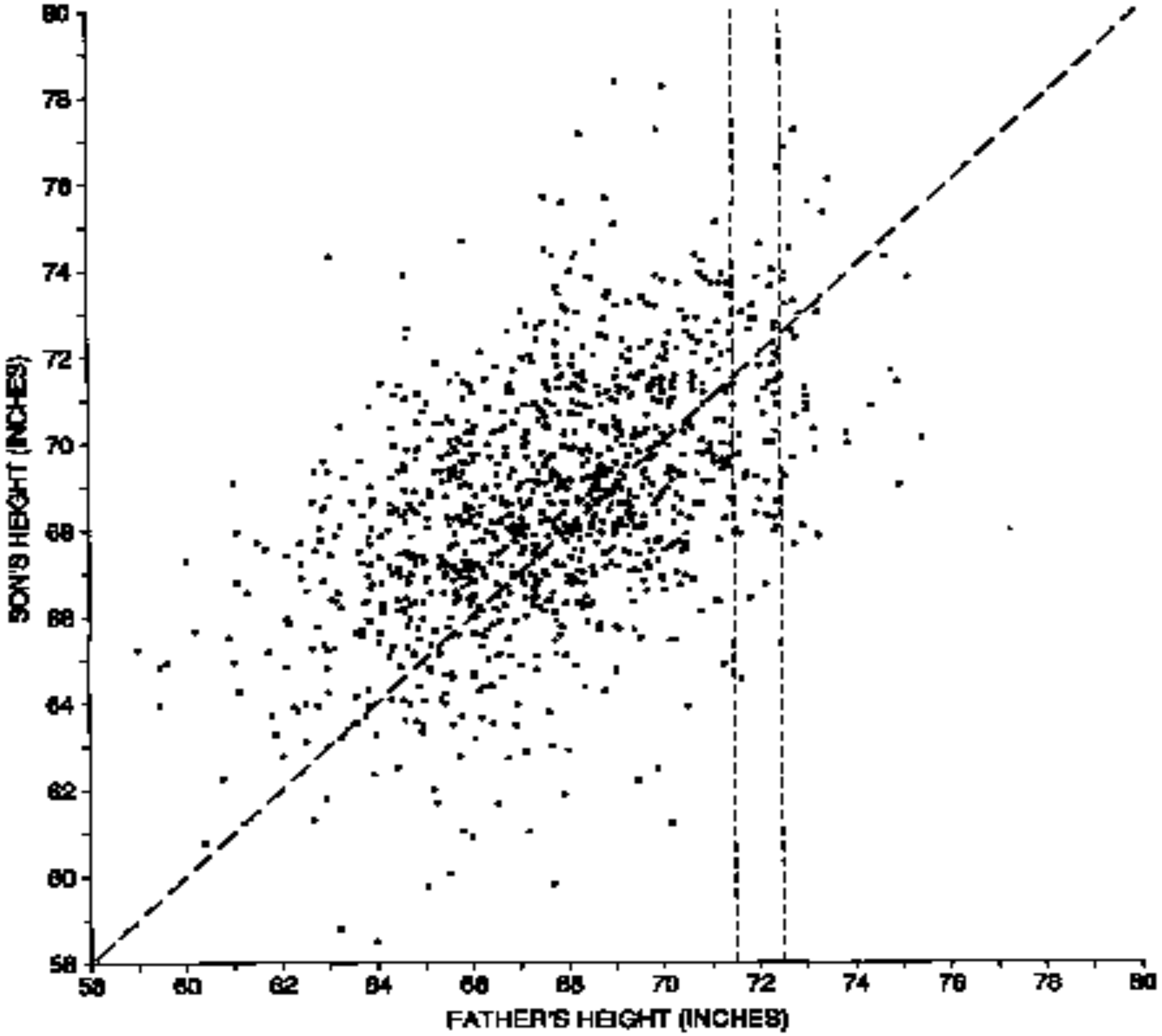


Image from Andres Robotics and Science

Correlation

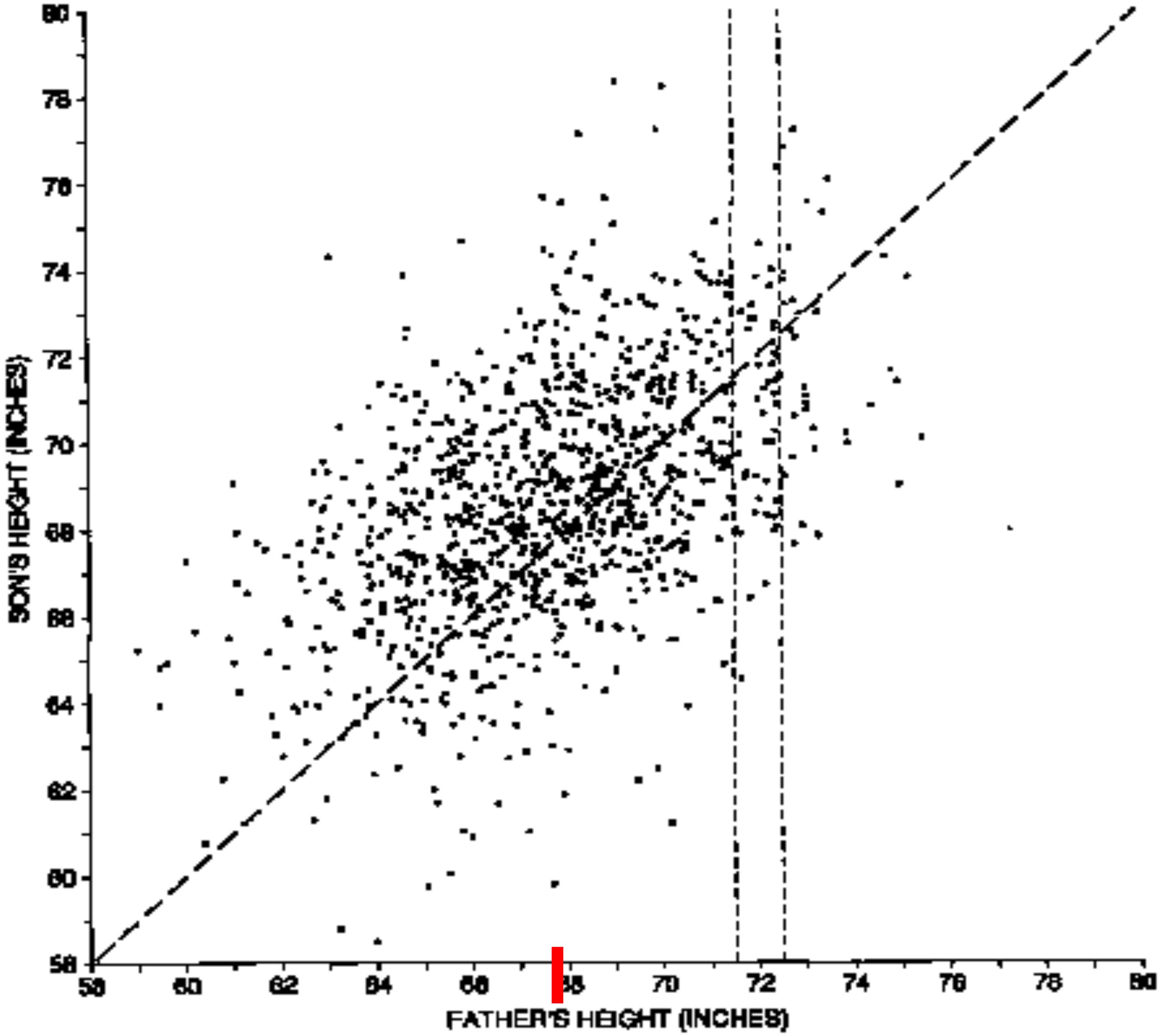


How can we summarize?

How can we summarize?

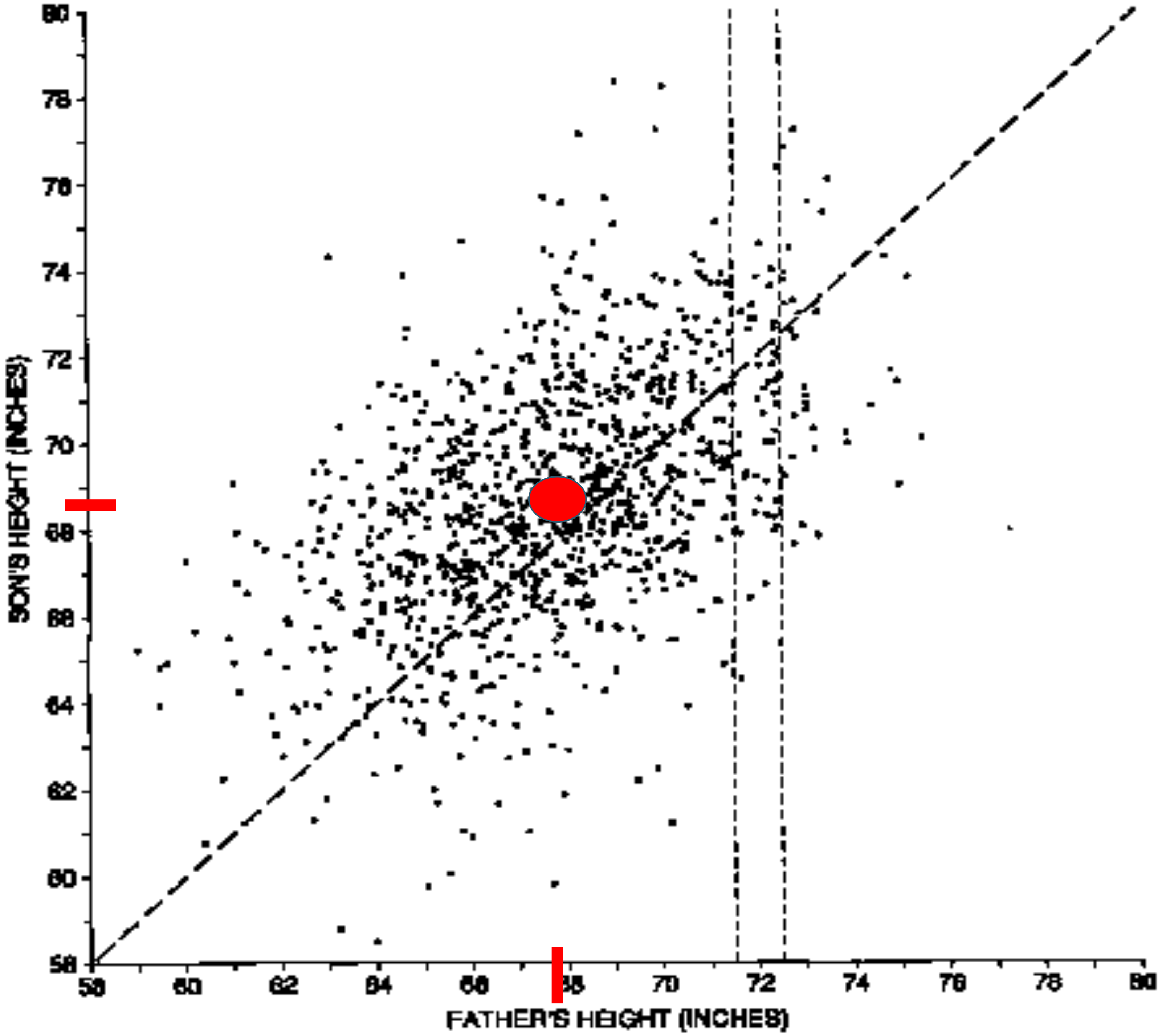
Average of X

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



How can we summarize?

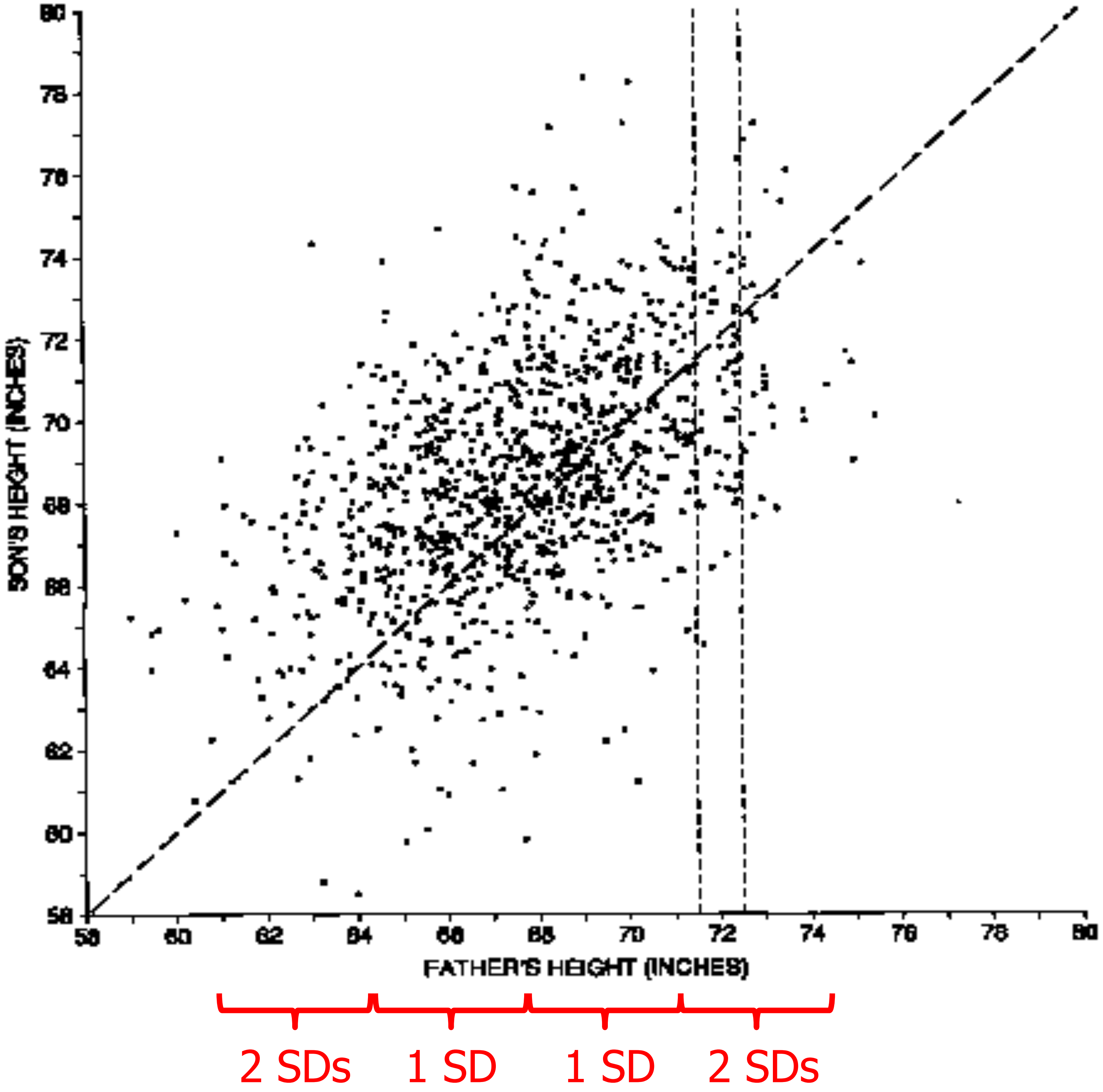
Average of X
Average of Y
Center



How can we summarize this data?

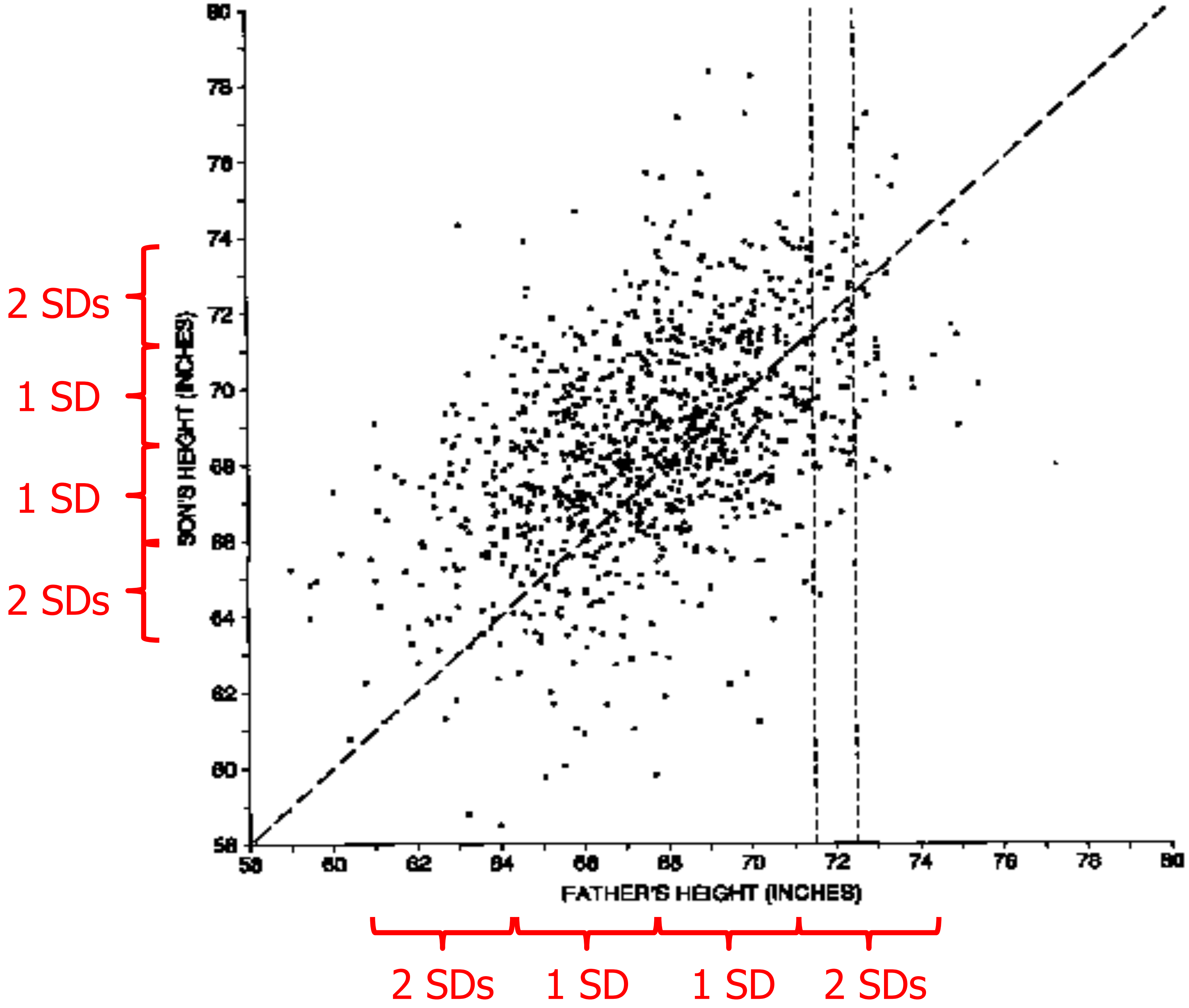
SD of X

$$\sigma = \sqrt{\frac{\Sigma (X - \overline{X})^2}{n - 1}}$$



How can we summarize this data?

SD of X
SD of Y
Spread



How can we summarize this data?

Average of X

Average of Y

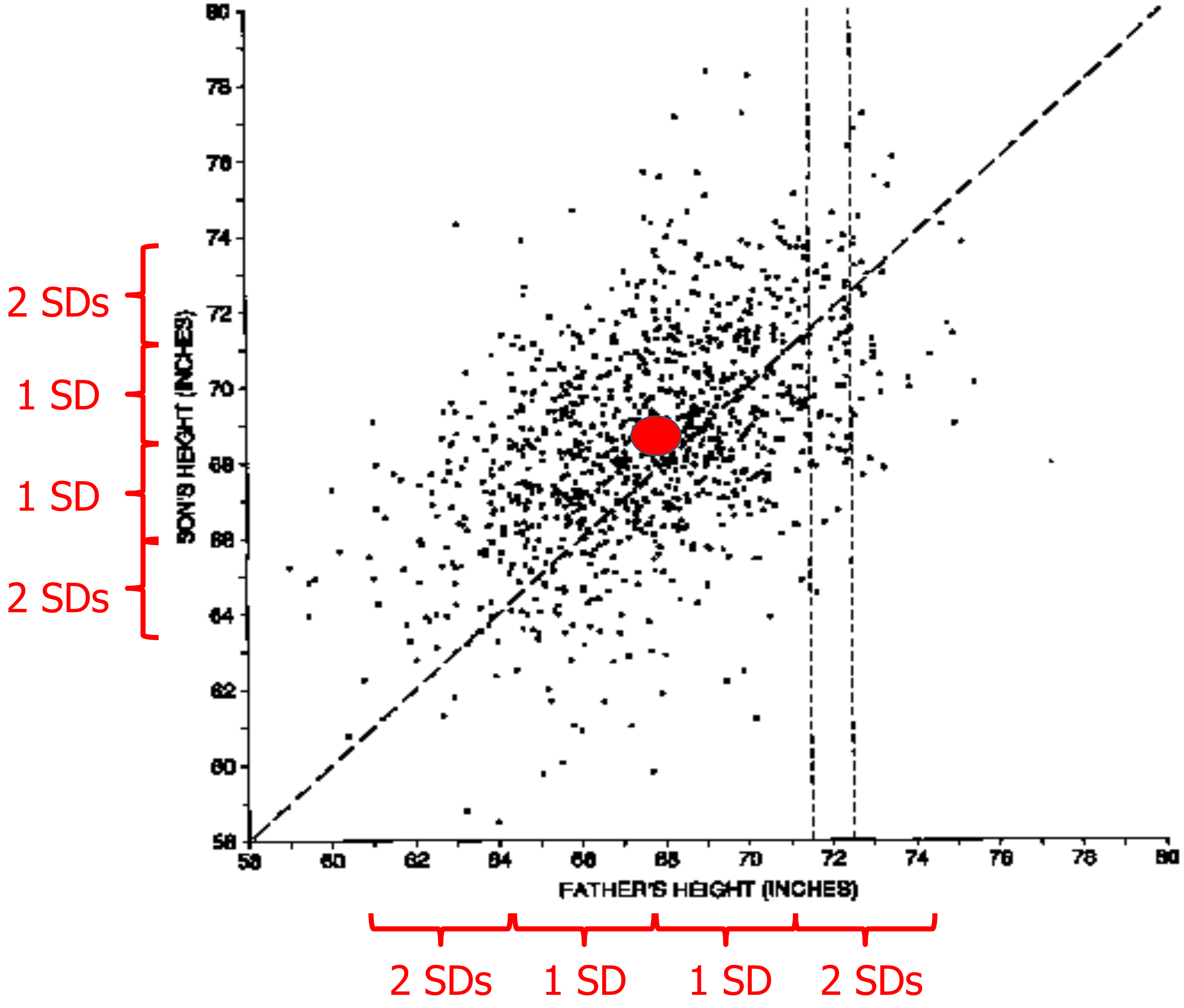
Center

SD of X

SD of Y

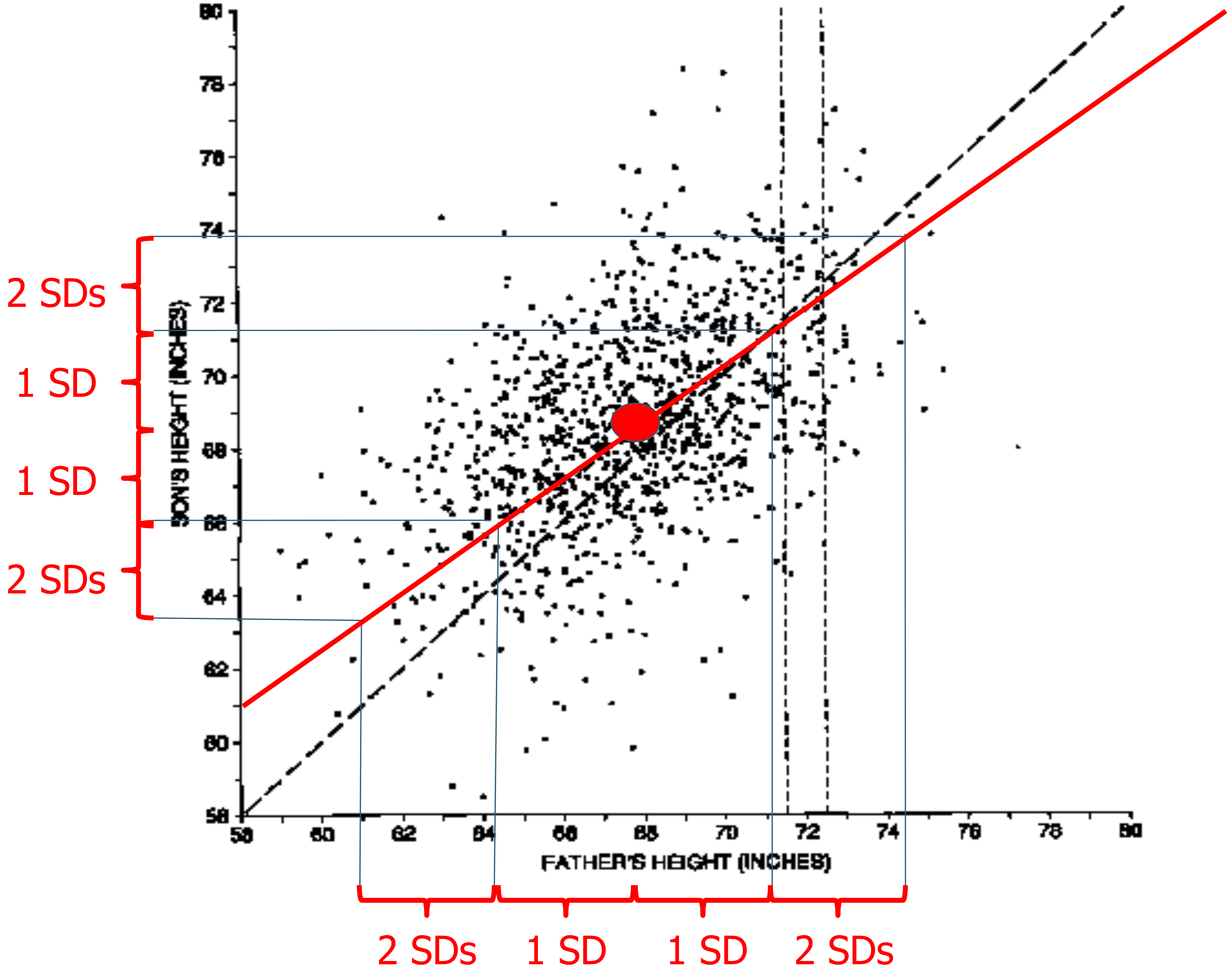
Spread

How are the two variables related?



Convert each variable to standard units.

The average of the products gives the **correlation coefficient**



Pearson Correlation

$$Correlation = \frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

For Population

$$Cov(x,y) = \frac{\Sigma\,(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma\,(x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

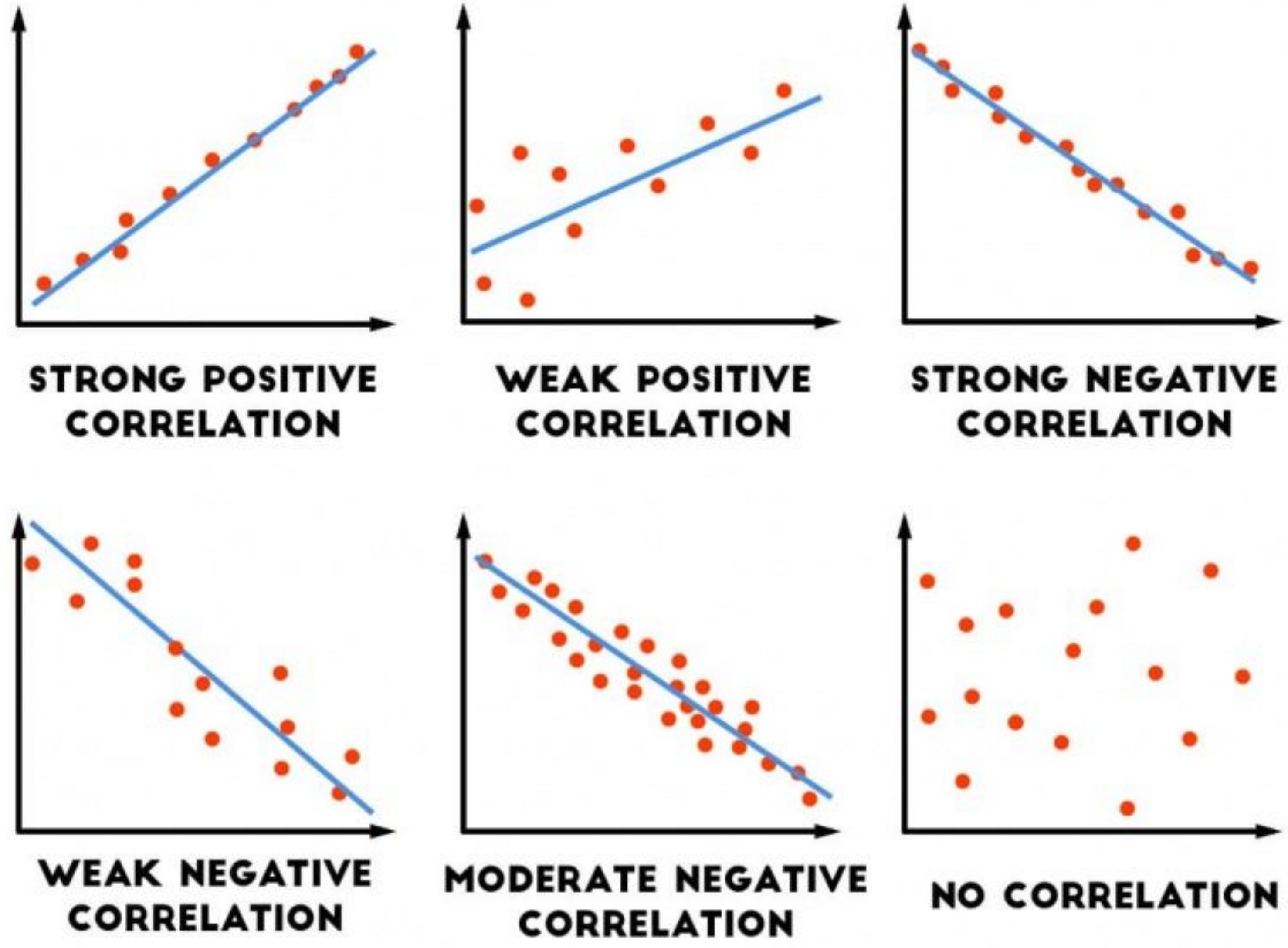


Image from Andres Robotics and Science

Pearson as Standard Unit

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}\frac{(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$$\frac{1}{n-1}\sum_{i=1}^{n}\mathrm{su}_{xi}\mathrm{su}_{yi}$$

Pearson Correlation

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

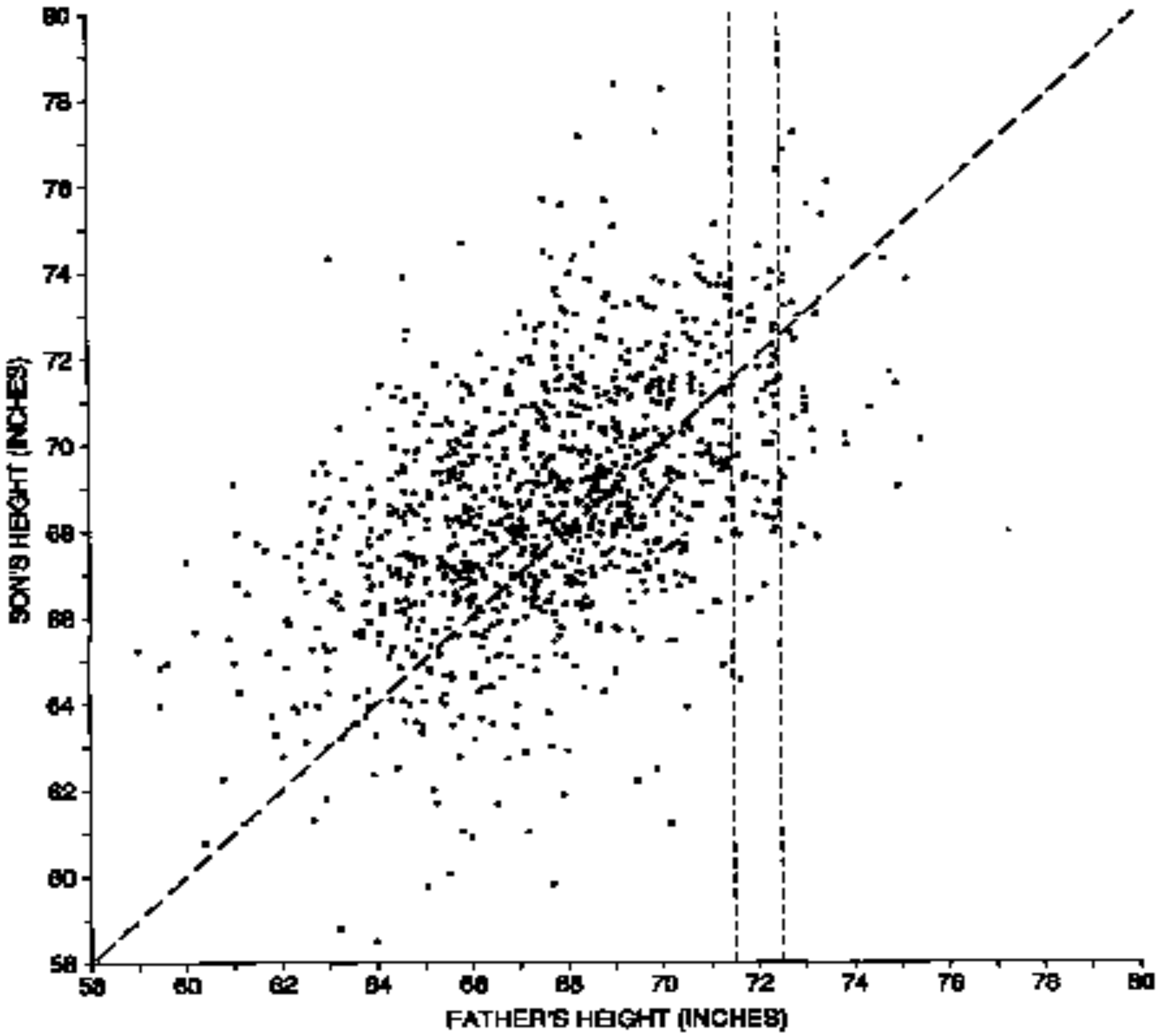
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

For Population

$$Cov(x,y) = \frac{\Sigma\ (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

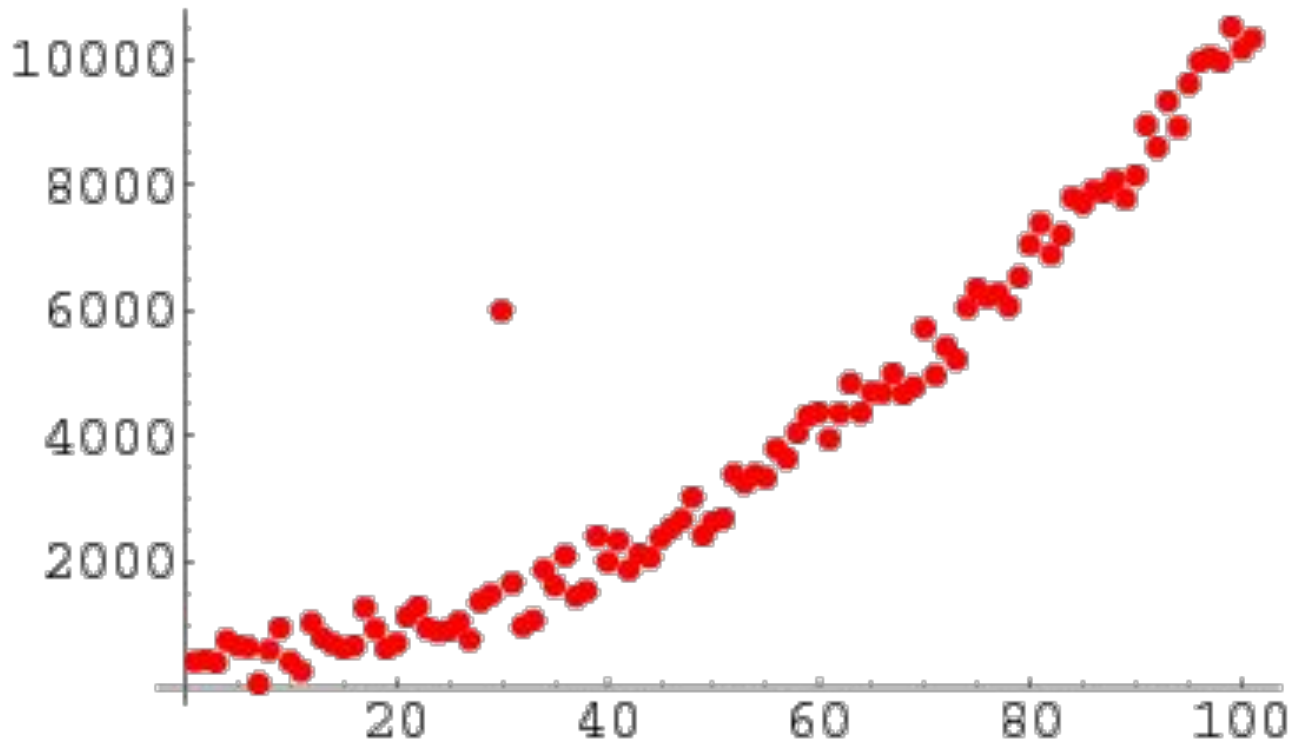
For Sample

$$Cov(x,y) = \frac{\Sigma\ (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

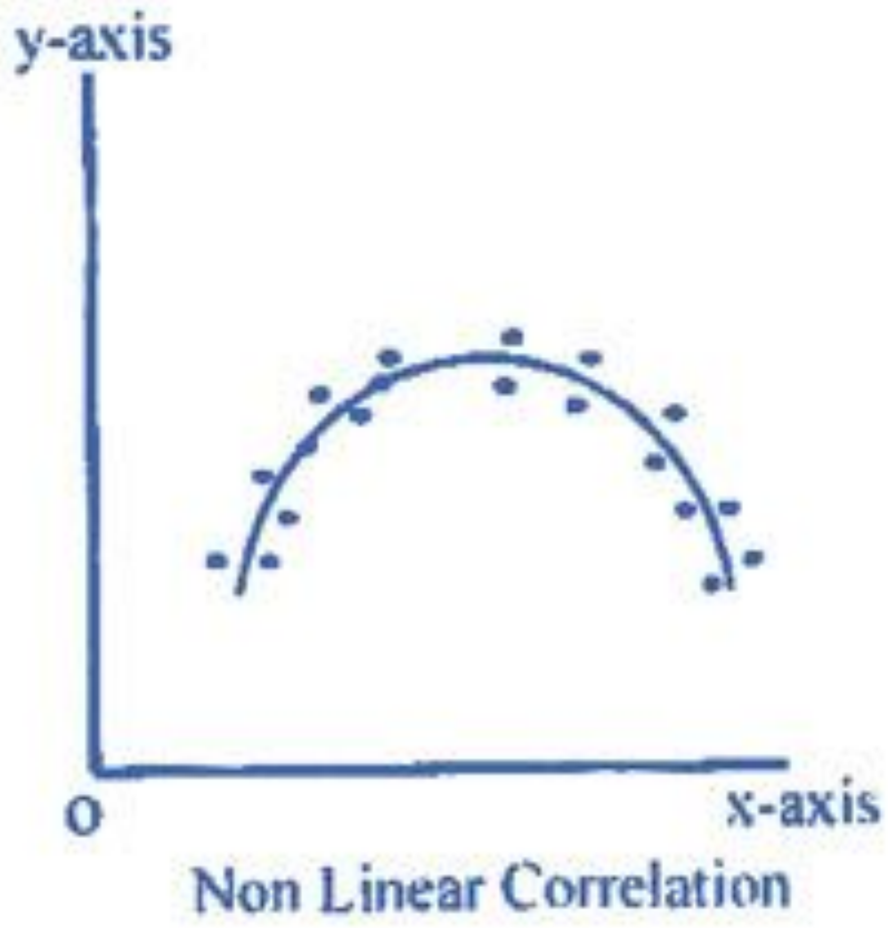


Correlation - Exceptional Cases

Outliers



Non Linear



Correlation is not Causation!

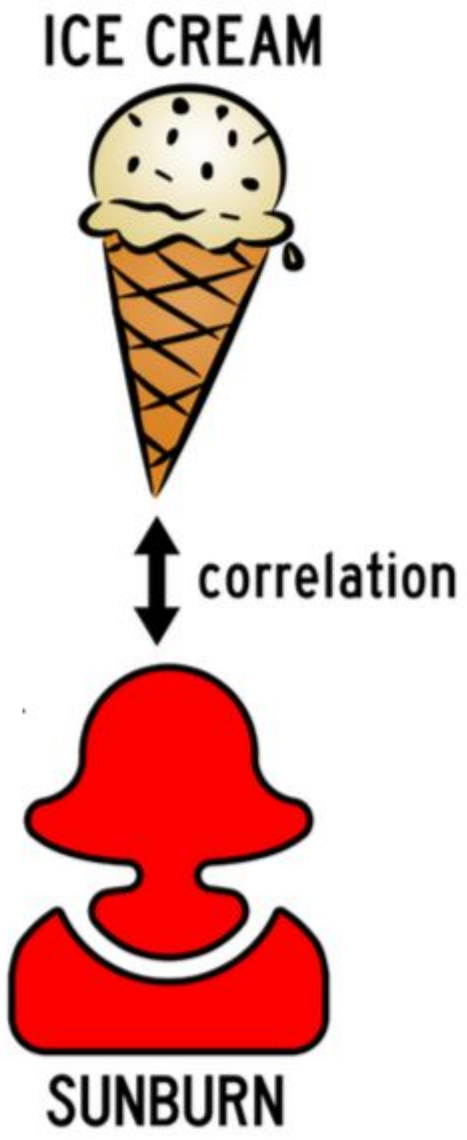


Image from Towards Data Science

Correlation is not Causation!

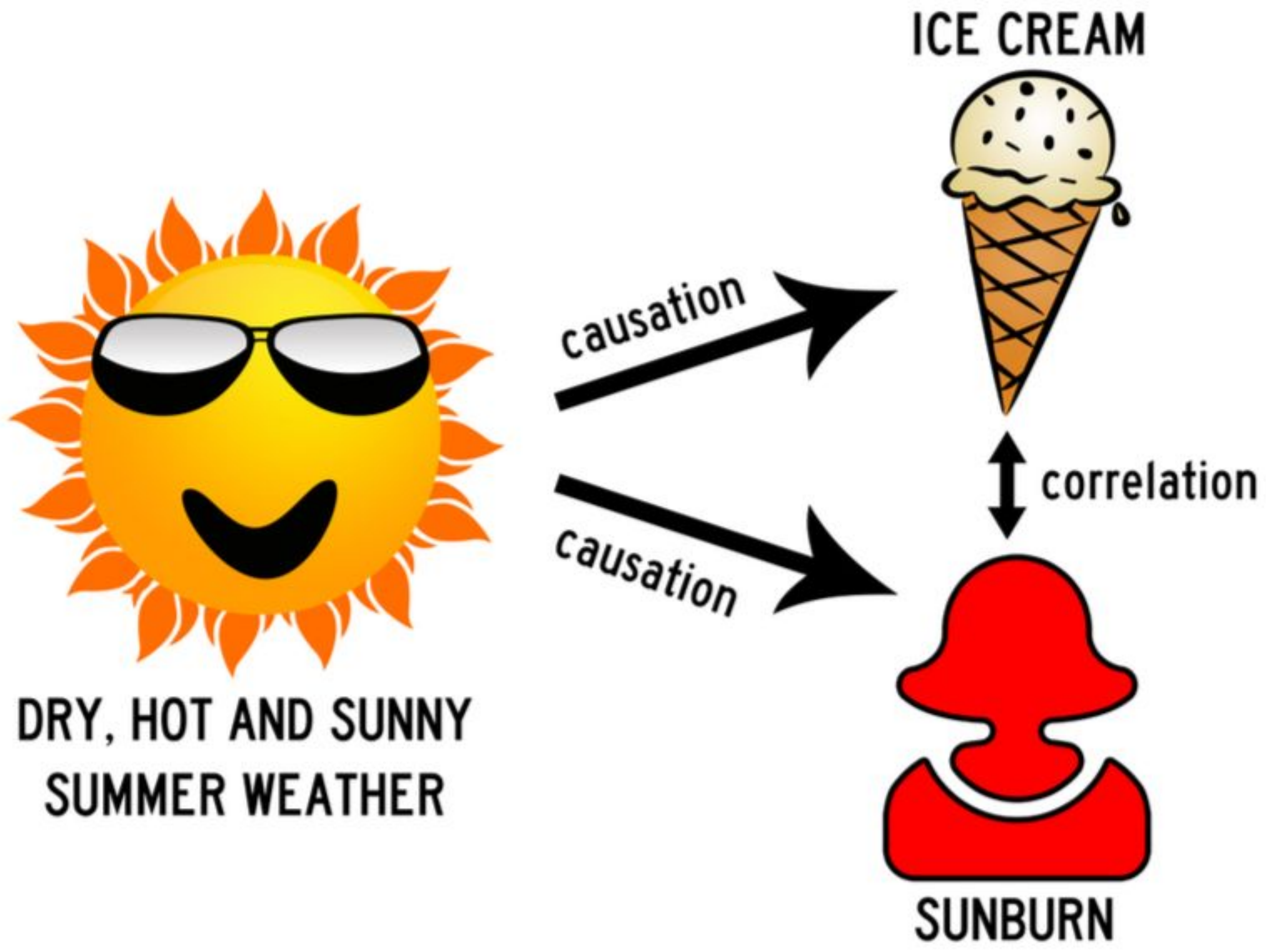
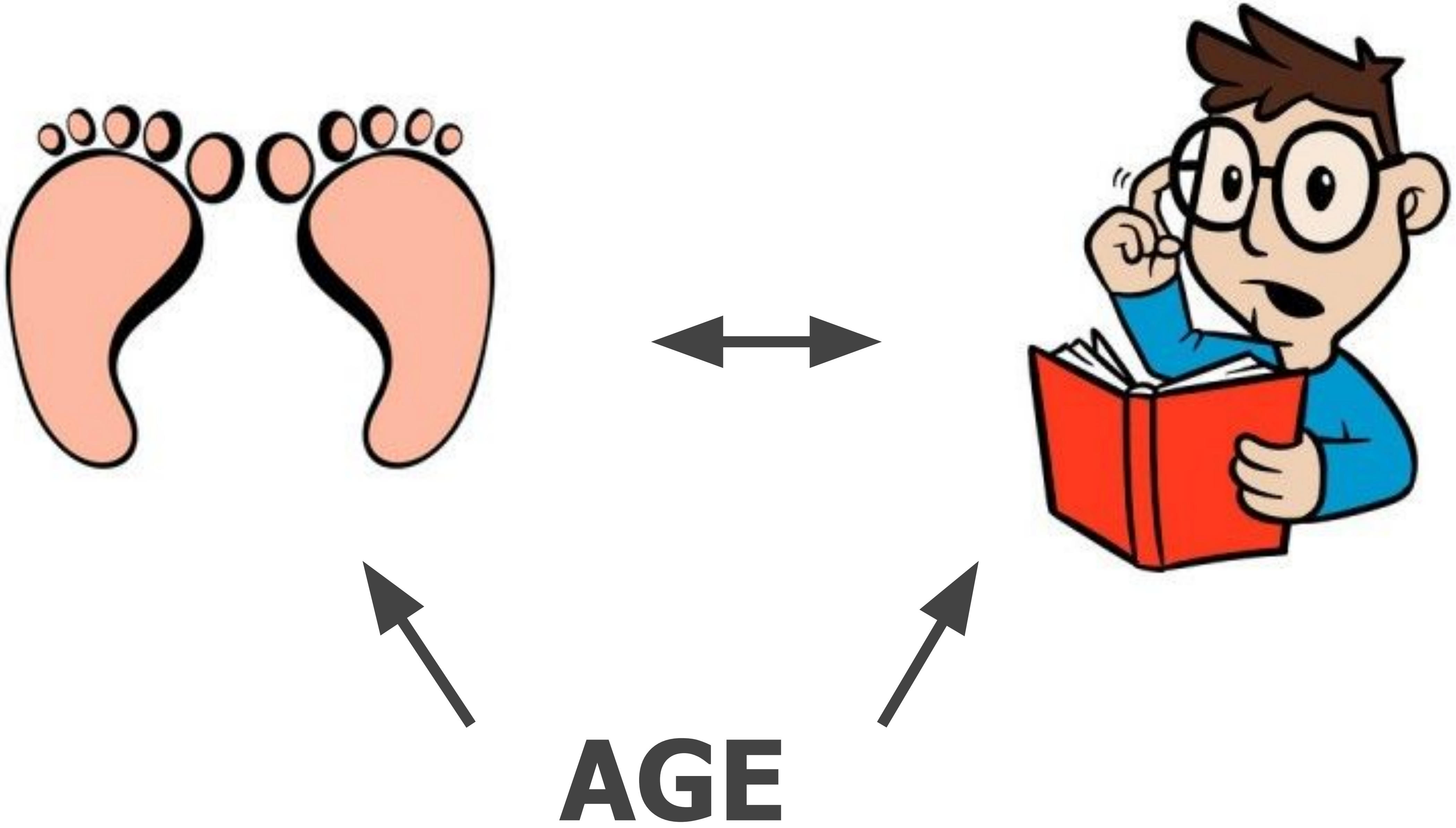
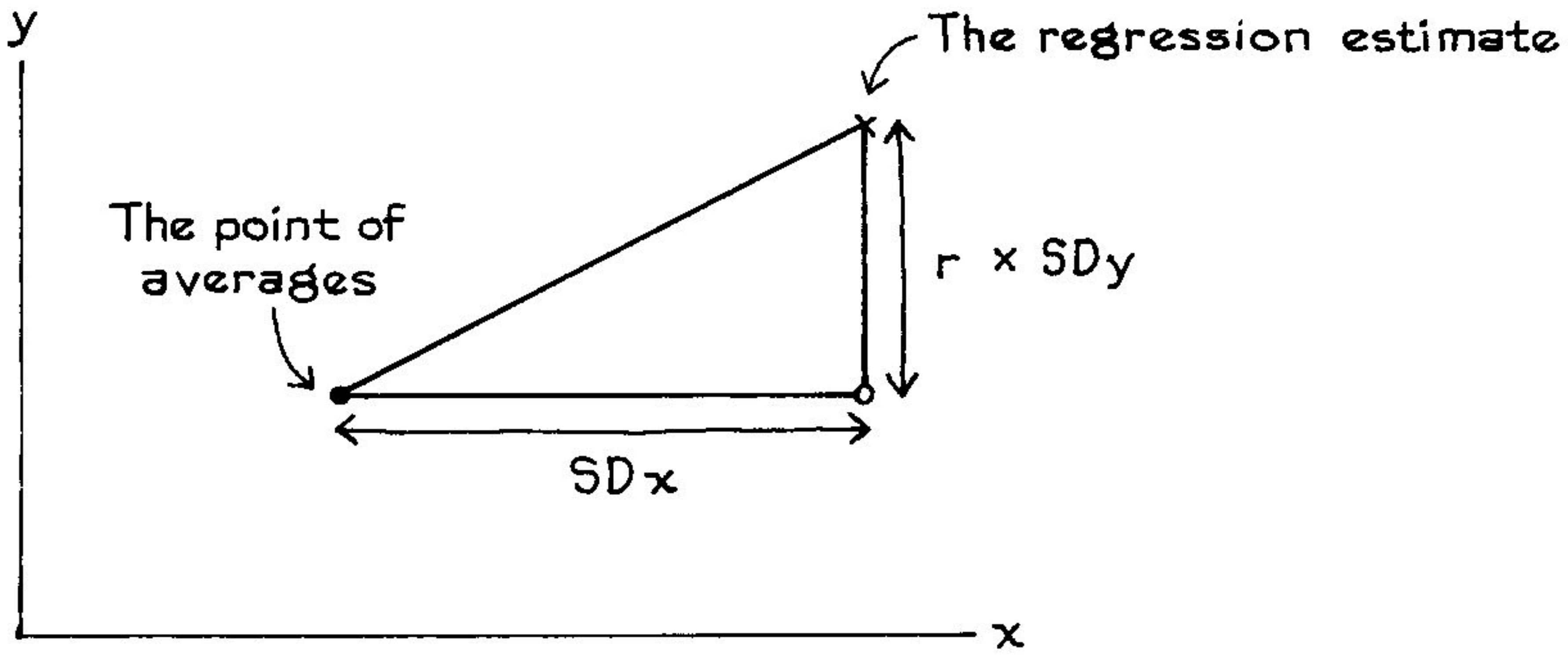


Image from Towards Data Science

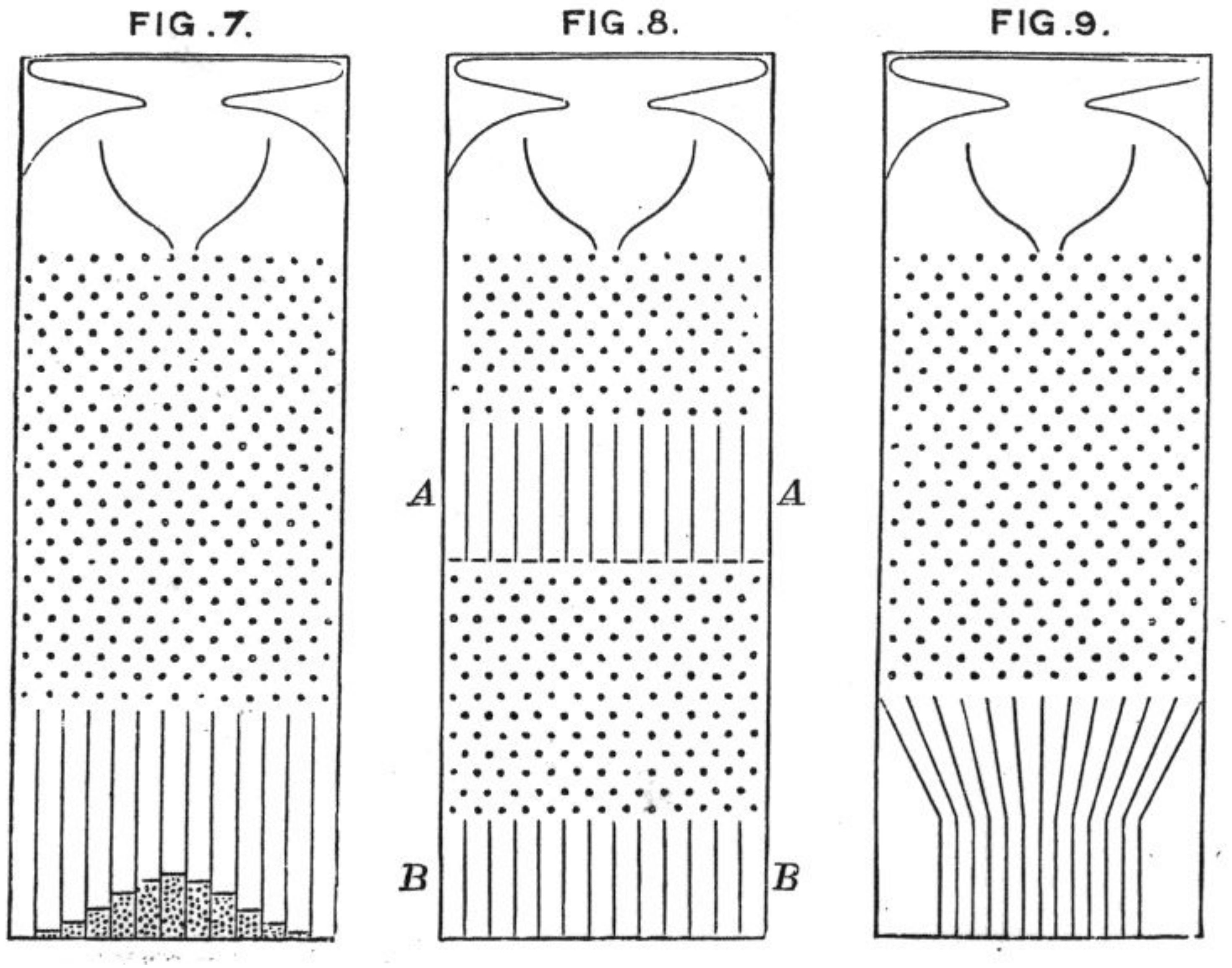
Correlation is not Causation!



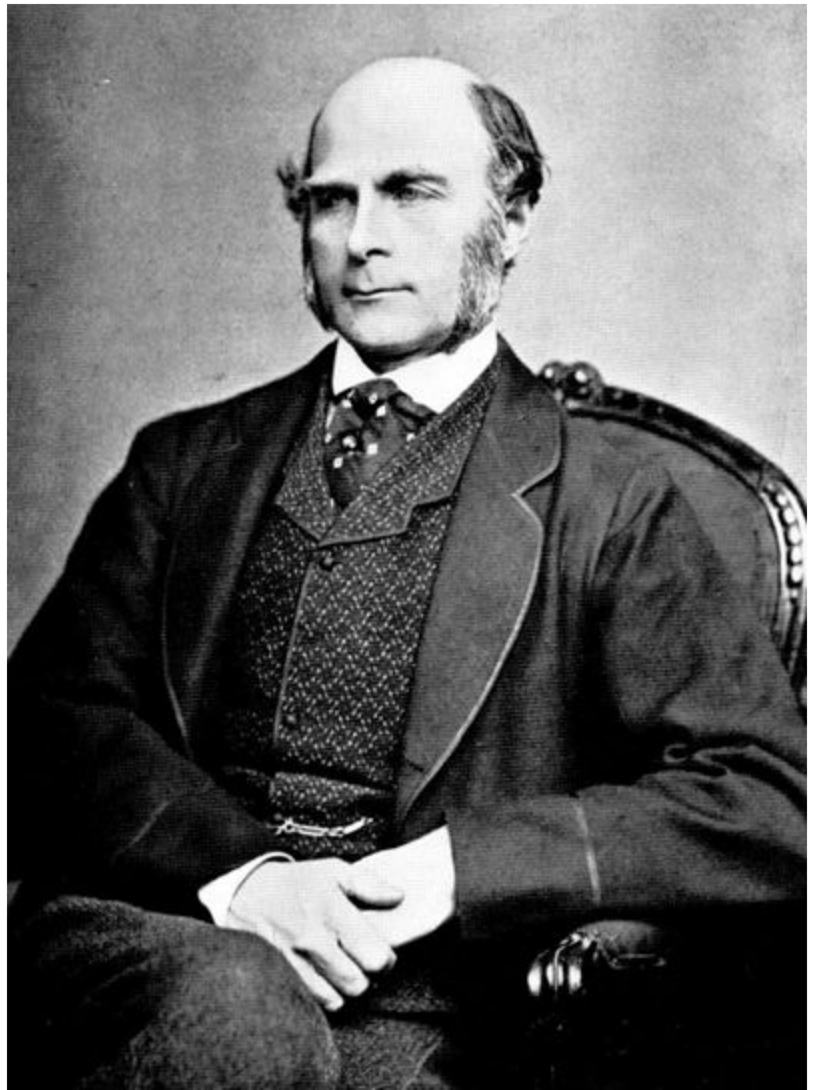
With each increase of one SD in X there is an increase of only r SDs in Y, on the average.



Regression Towards the Mean



Sir Francis Galton



The Regression Fallacy

"... it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them."

Daniel Kahneman, winner of the 2002 Nobel Memorial Prize in Economic Sciences

Control Group

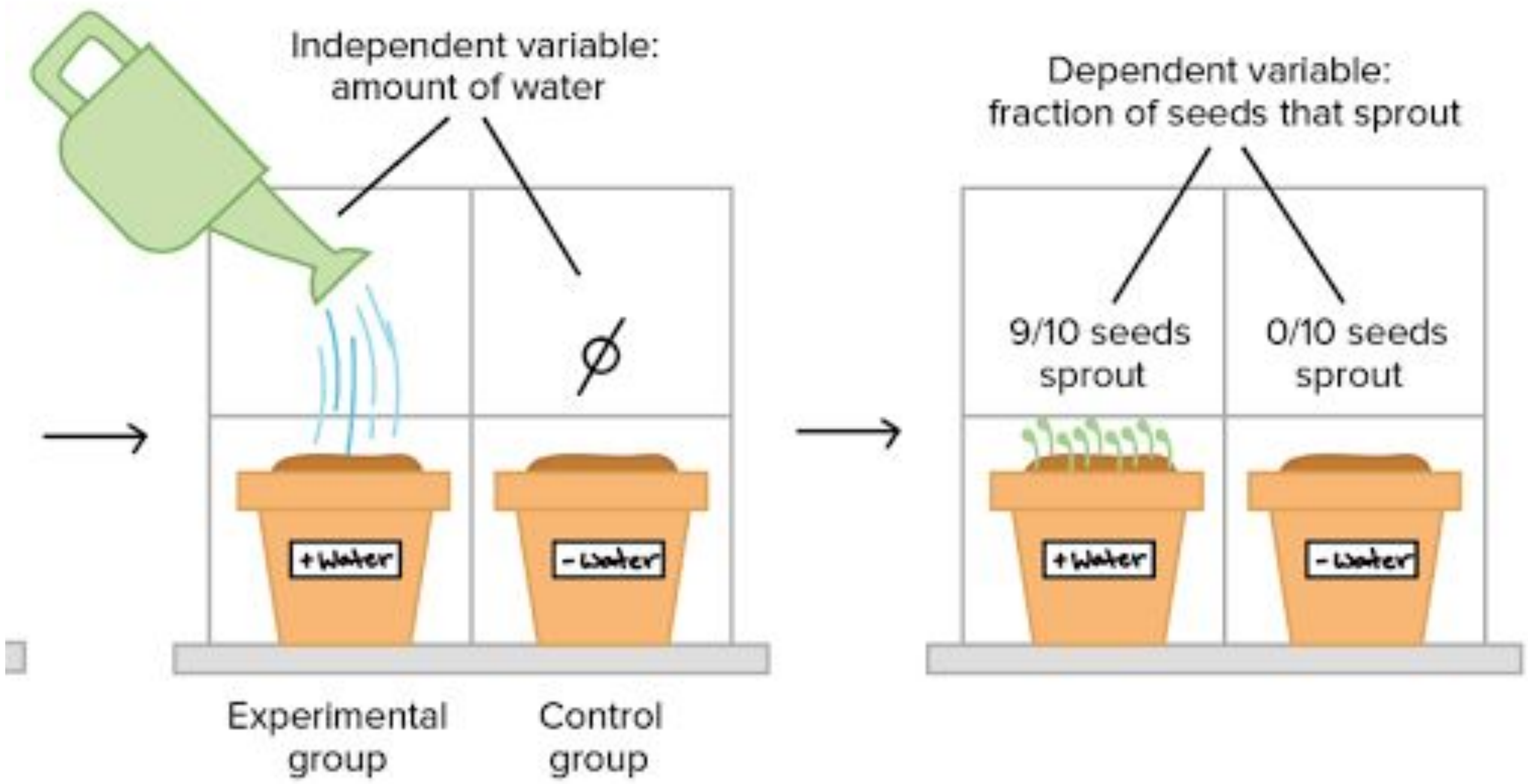


Image from Khan Academy

Sampling Bias

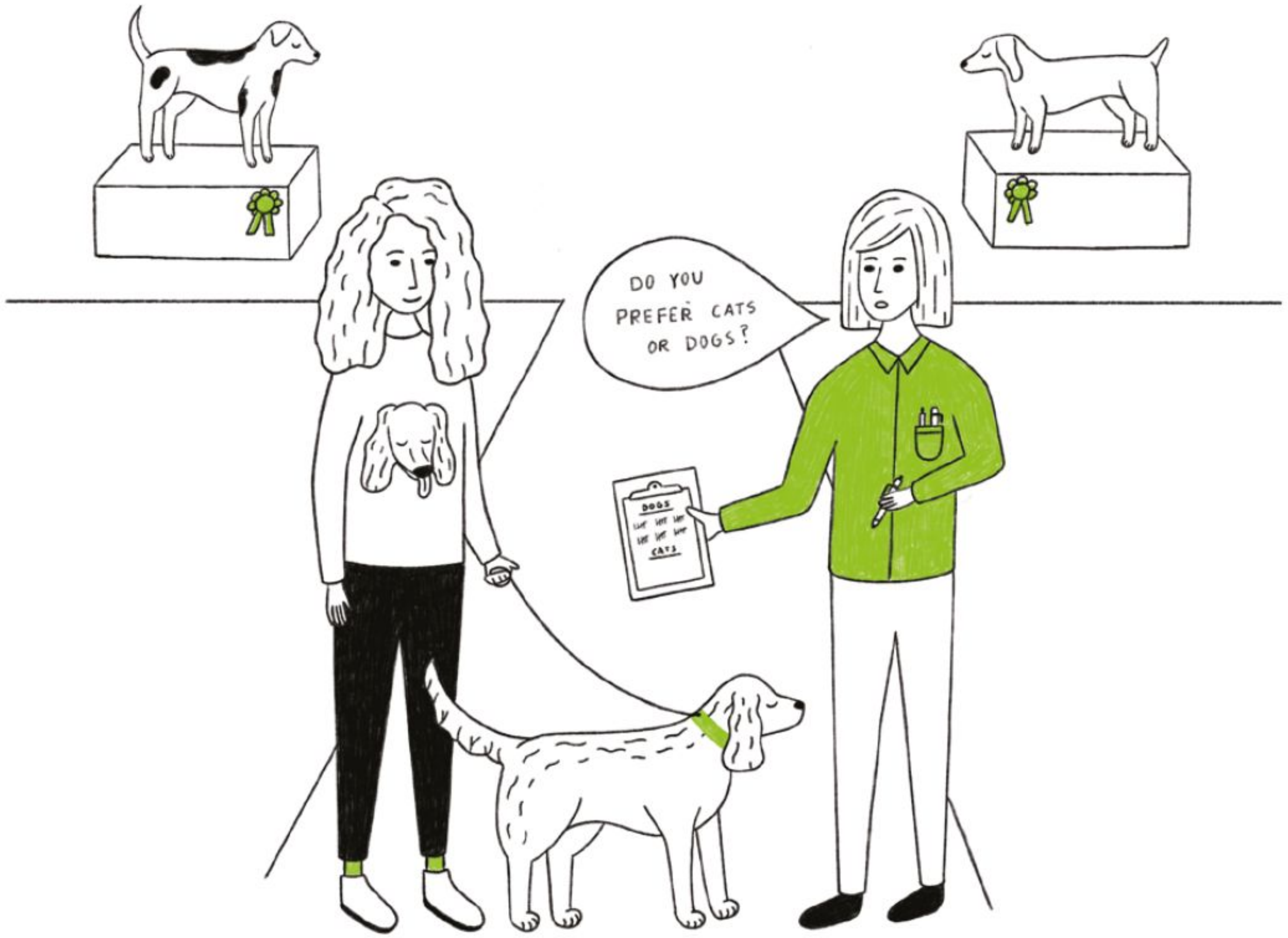


Image from Geckoboard

Sampling Bias

Image from Amazon.com

P-value

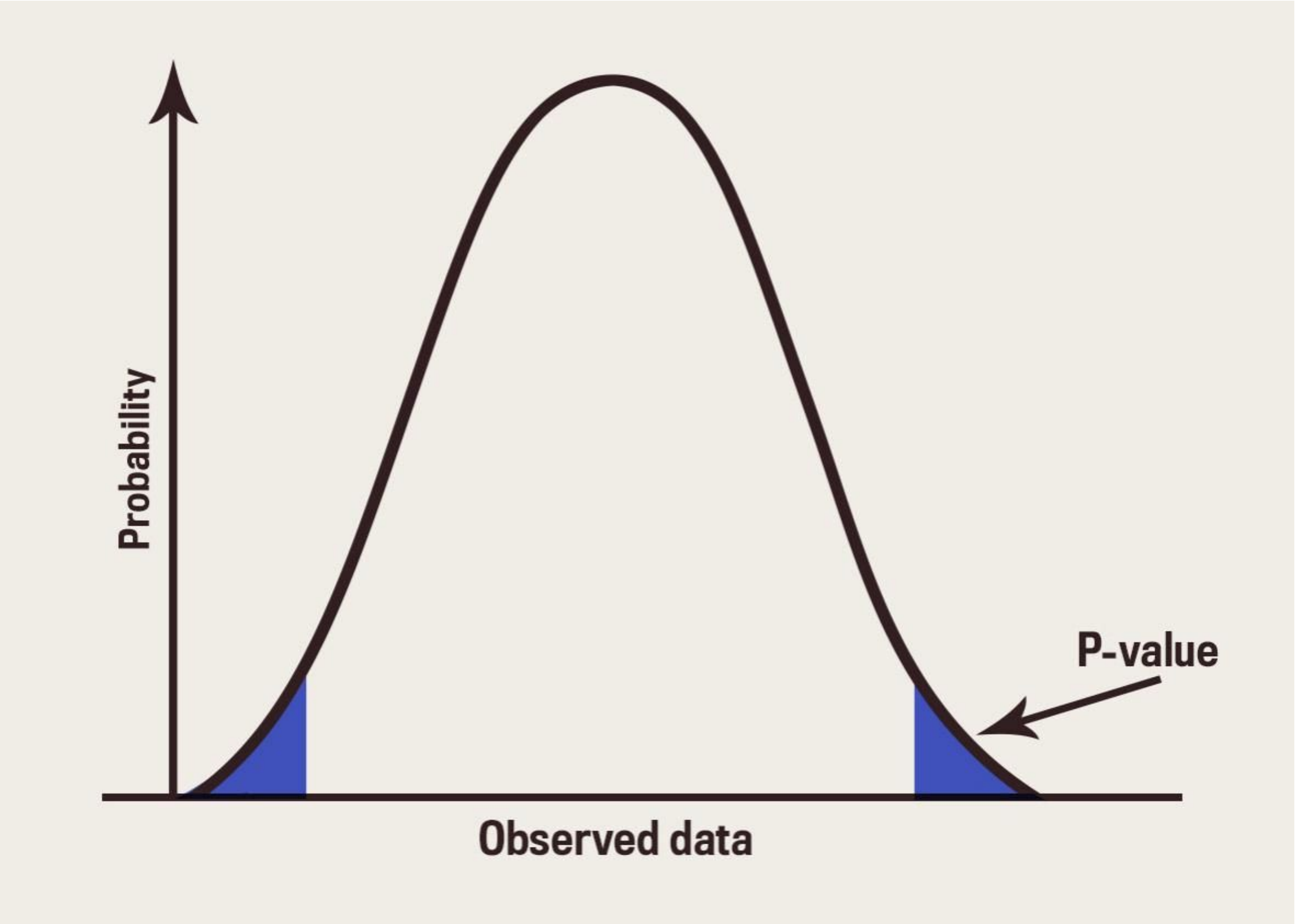


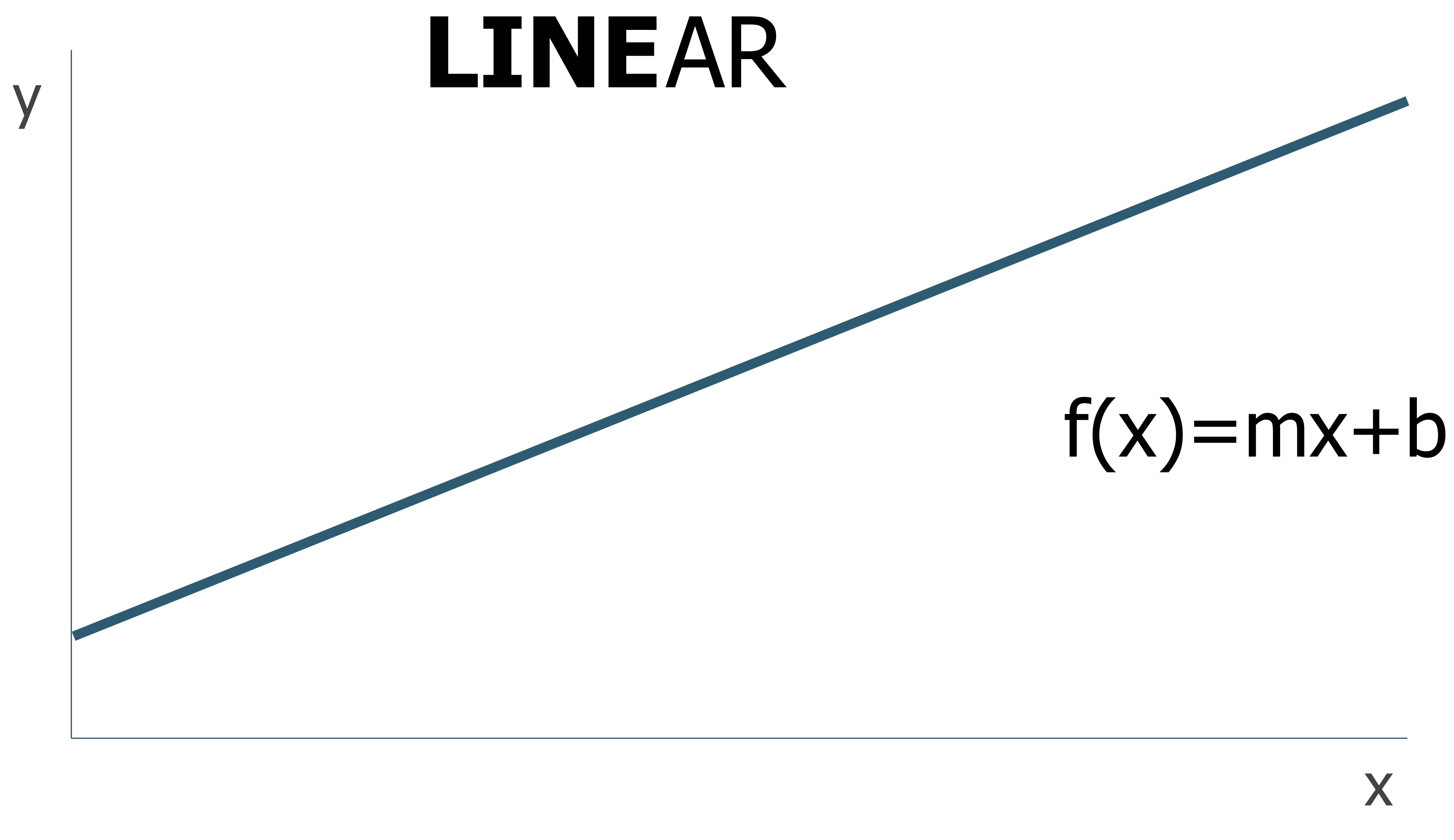
Image from Slate.com

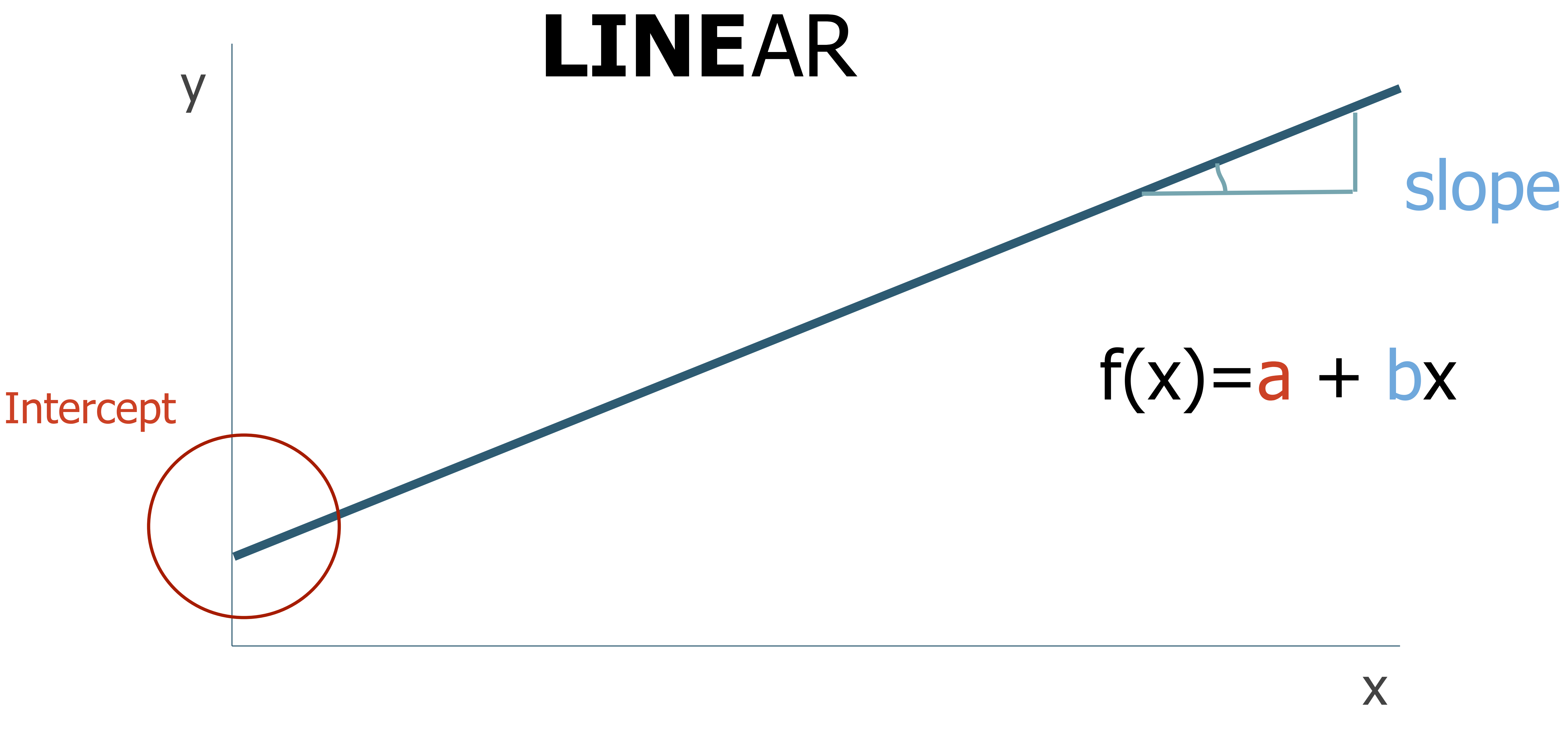
A **linear function** is one for which

$$f(x+y)=f(x)+f(y)$$

and

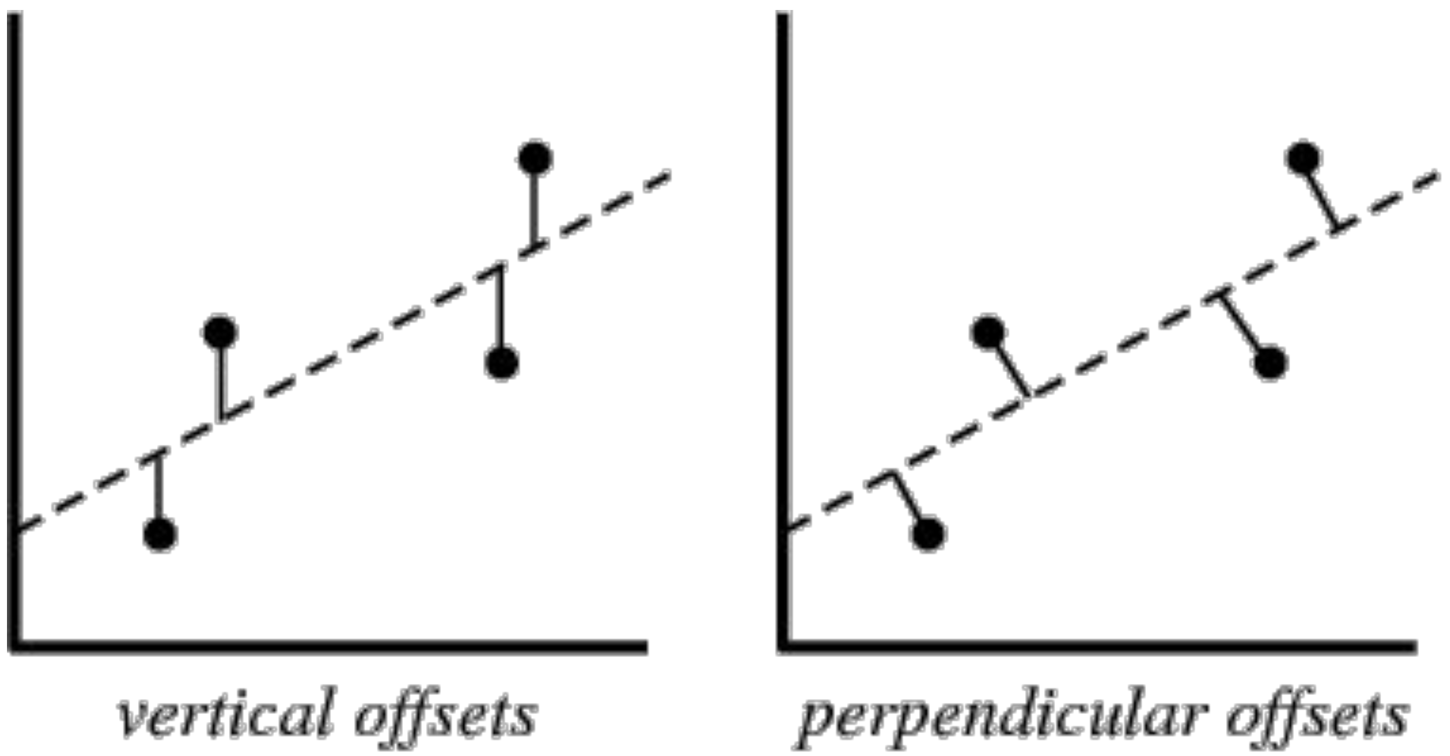
$$f(ax)=af(x)$$


LINEAR
y
f(x)=mx+b
x


LINEAR
y
slope
f(x)=a + bx
Intercept
x

Least Square

$$S = \sum_{i=1}^{N} r_i = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2$$

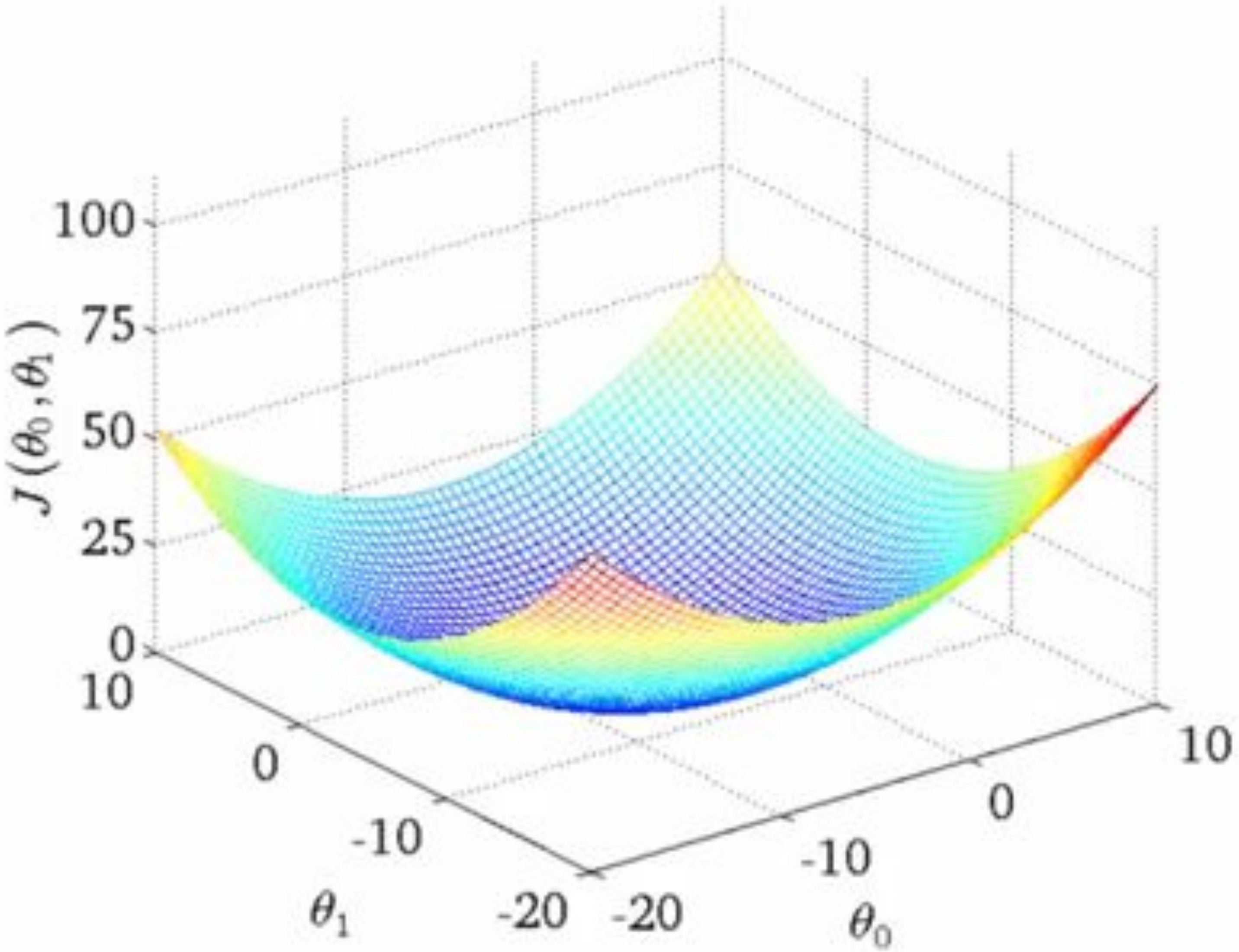


Least Square

$$\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

$$\frac{\partial}{\partial \hat{\beta}_0} \mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = 0$$

$$\frac{\partial}{\partial \hat{\beta}_1} \mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = 0$$



Least Square

$$\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

$$\frac{\partial}{\partial \hat{\beta}_0} \mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{n} y_i + 2n\hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$= \boxed{-2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1 \overline{x}}$$

$$\frac{\partial}{\partial \hat{\beta}_0} \mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = 0$$

$$-2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1 \overline{x} = 0$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Chain Rule

If f and g are both differentiable and $F(x)$ is the composite function defined by $F(x) = f(g(x))$ then F is differentiable and F' is given by the product

$$F'(x) = f'(g(x))\, g'(x)$$

Differentiate outer function

Differentiate inner function

Full Explanation

http://stat.math.uregina.ca/~kozdron/Teaching/Regina/252Winter05/Handouts/least_squares.pdf

Regression in sklearn and statsmodels

R-squared =

$$\frac{\text{Explained variation}}{\text{Total variation}}$$

$$\frac{\text{var(mean)-var(line)}}{\text{var(mean)}}$$

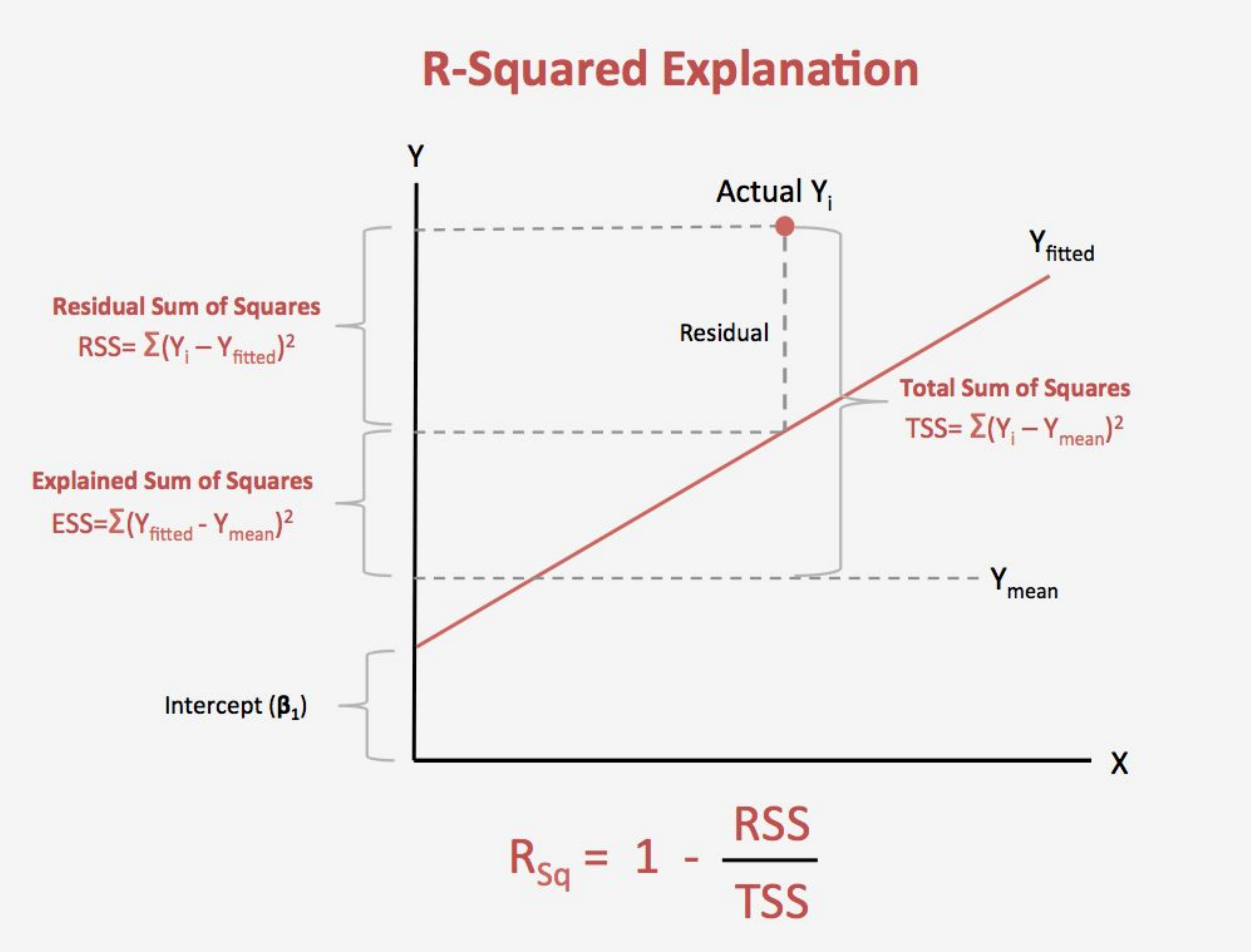


Image from machinelearningplus.com

Understanding Linear Regression Results

```
Optimization terminated successfully.
         Current function value: 0.441635
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                Failure   No. Observations:                   23
Model:                          Logit   Df Residuals:                       21
Method:                           MLE   Df Model:                            1
Date:                Fri, 18 Oct 2019   Pseudo R-squ.:                  0.2813
Time:                        17:57:08   Log-Likelihood:                -10.158
converged:                       True   LL-Null:                       -14.134
                                        LLR p-value:                  0.004804
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      15.0429      7.379      2.039      0.041       0.581      29.505
Temperature    -0.2322      0.108     -2.145      0.032      -0.444      -0.020
===============================================================================
```

Bias vs. Variance

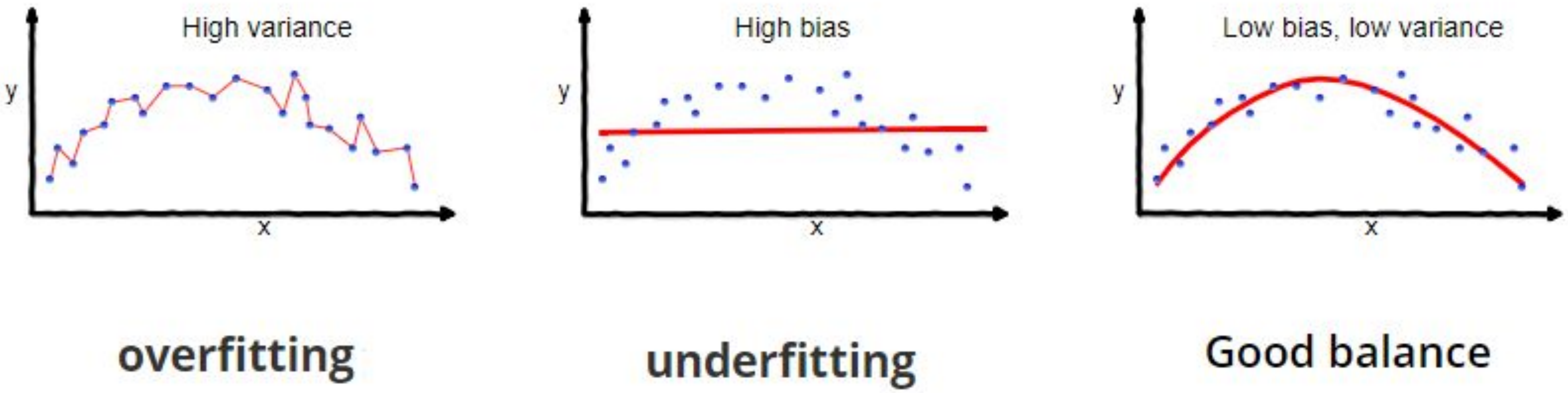


Image from TowardsDataScience

Bias vs. Variance

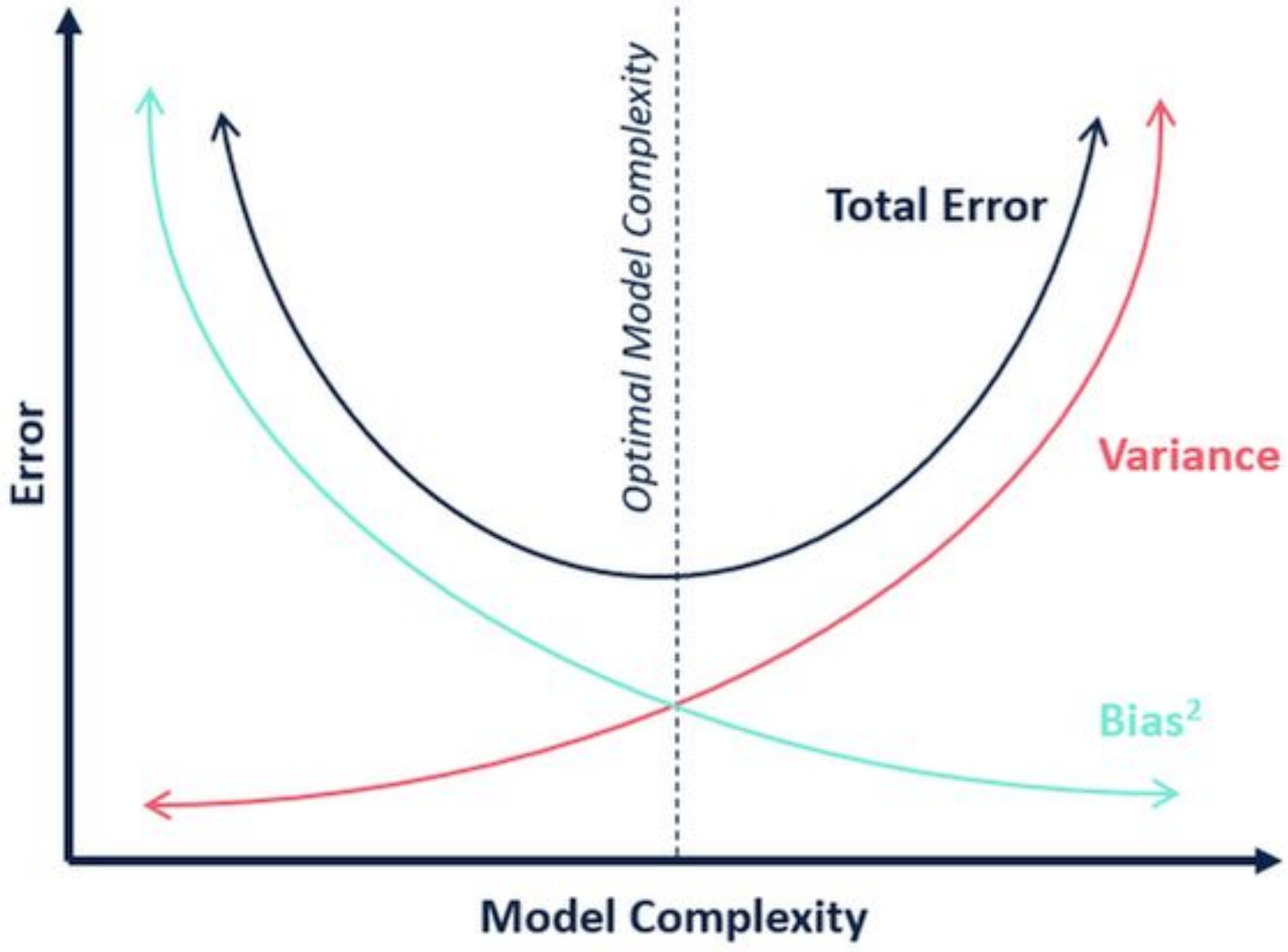


Image from Alteryx Community

Train / Validation / Test

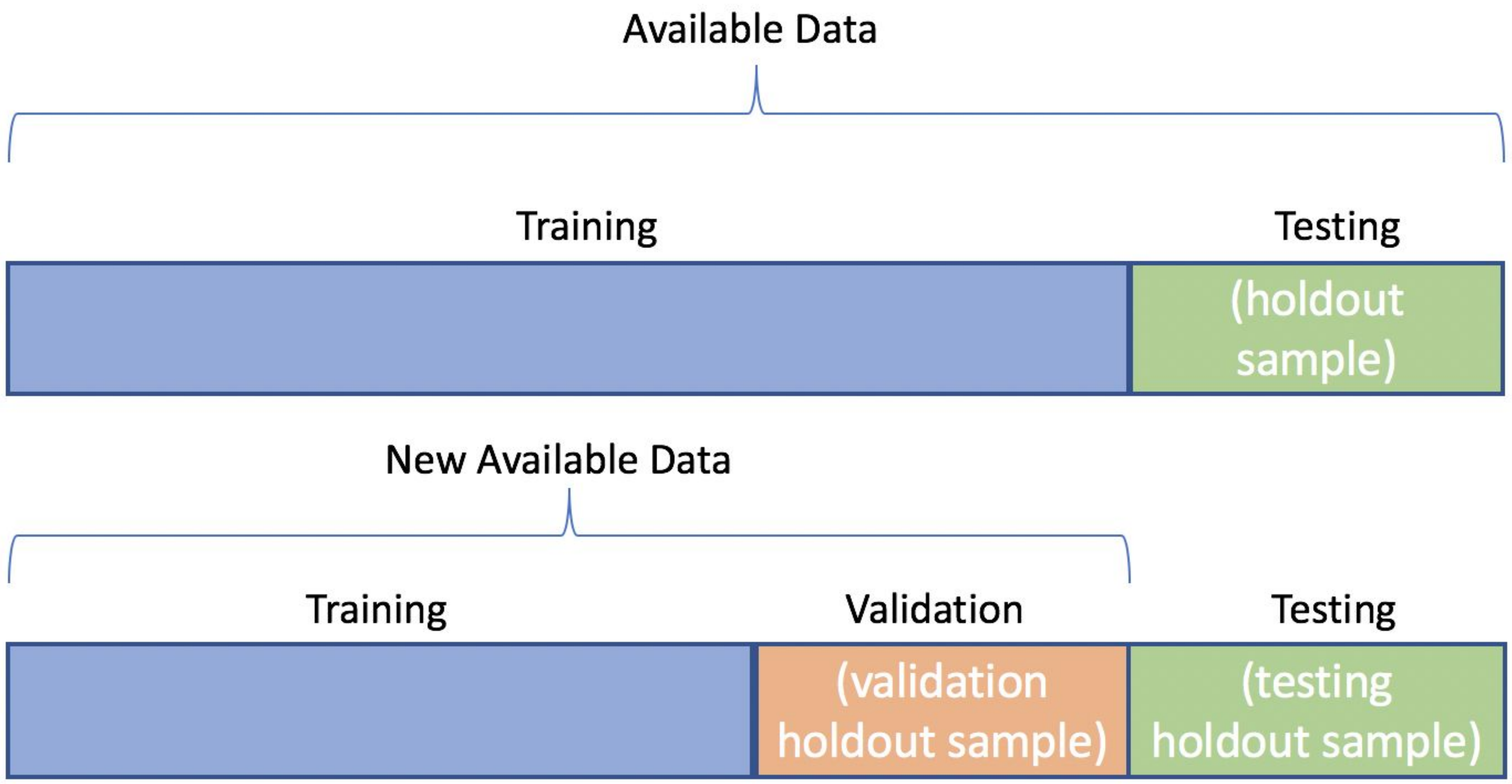
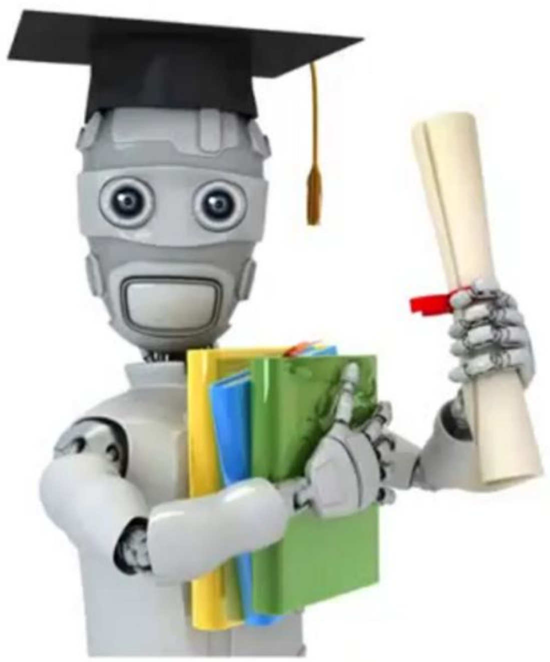


Image from datacamp.com

Useful Resources

Useful Resources



Machine Learning

Machine Learning

by Andrew Ng

coursera

3 Blue 1 Brown

THANK YOU

GABRIEL MAGALHÃES

Data Scientist

gabriel.magalhaes@totvs.com.br

Tecnologia + Conhecimento são nosso DNA.

O sucesso do cliente é o nosso sucesso.

Valorizamos gente boa que é boa gente.

#SOMOSTOTVERS

totvs.com

company/totvs

@totvs

fluig.com