STAT 113
Correlation II

Colin Reimer Dawson

Oberlin College

September 19, 2017


Talk
like
a
Pirate
Day
September 19

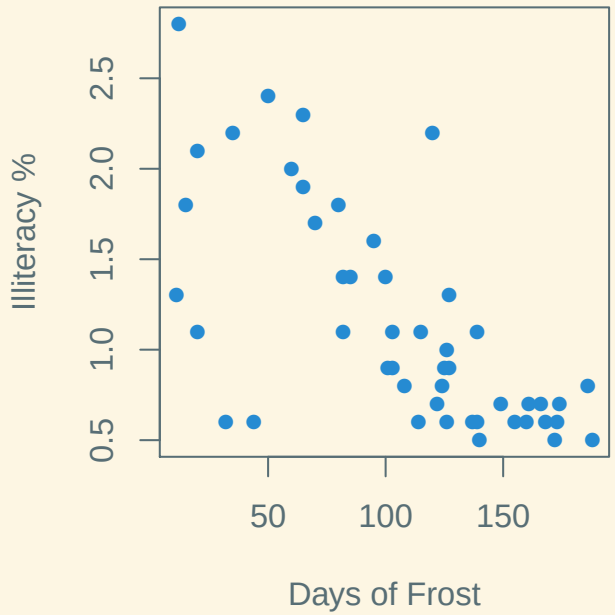
Outline

Correlation Math

Pearson's r

Warmup: Frost and Illiteracy

What do you expect a scatterplot relating a state's frost rate (days of frost per year) and its illiteracy rate (% of population who can't read) to look like?
$r = -0.68$



Correlation vs. Causation

Correlation $\neq$ Causation!

A correlation between two variables does not mean that one caused the other or that they are necessarily related in real life.

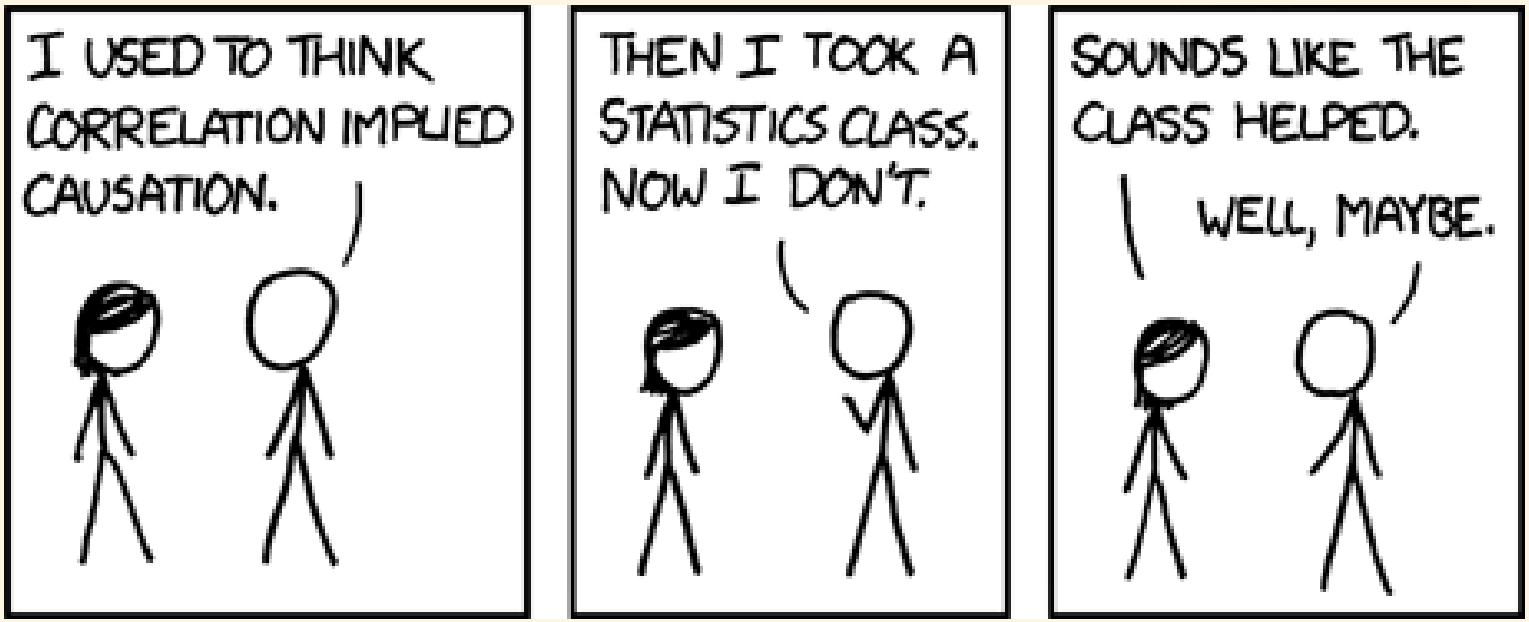


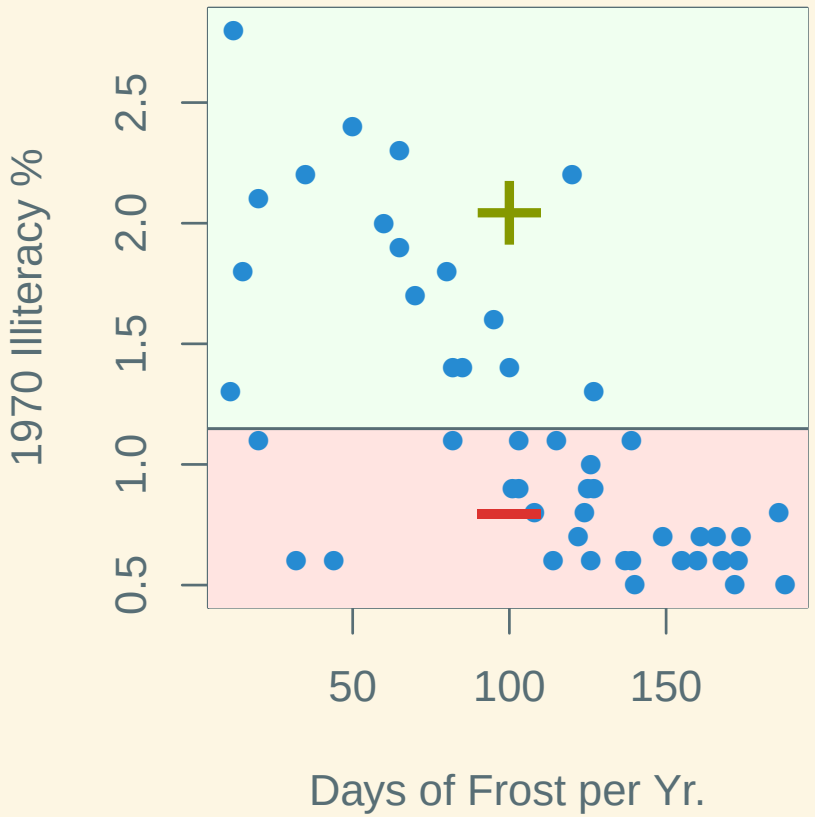
Figure: xkcd.com/552/

Measuring Relationships

- The idea: Two variables have a "positive" relationship if one has "high values" at the same time the other is high, and "low values" when the other is low
- If the opposite is true, there is a "negative" relationship.
- What counts as "high" or "low"?

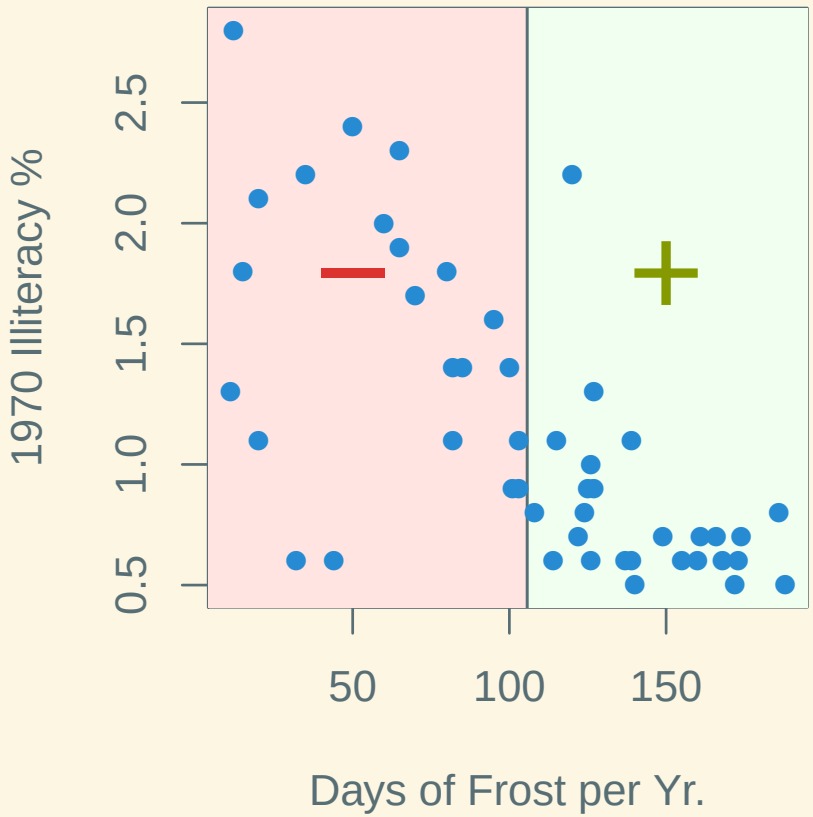
What's High and Low?

Intuition: count "above average/center" as "high"; "below average/center" as "low".



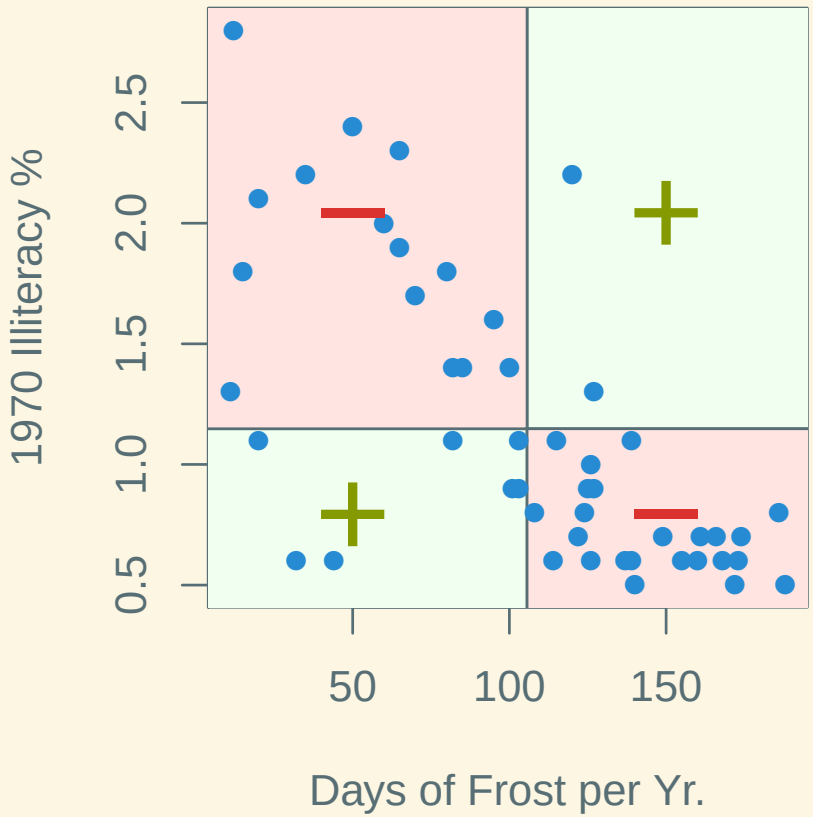
What's High and Low?

Intuition: count "above average/center" as "high"; "below average/center" as "low".



Positive vs. Negative Association

If deviations have the same sign, the observation suggests positive association. Opposite $\rightarrow$ negative.



x and y deviations

Each data point has two coordinates. We can write the i^{th} data point as (x_i, y_i).

$$\text{Deviation in } x \text{ direction} = x_i - \bar{x}$$

$$\text{Deviation in } y \text{ direction} = y_i - \bar{y}$$

From Deviations to Association

- What can we do so that we get a positive number when both deviation scores have the same sign, and a negative otherwise?
- Multiply them together!
- We can average these products to get a measure of association.

$$\text{Average Product of Devations} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

But Wait!

- What would happen if we change units?
- If we rescale one of the variables, the product of deviations will be rescaled as well. So Average Product of Deviations is sensitive to units.

A Solution

To remove units, use z-**scores** instead of raw deviations:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Measure deviation from mean using a one-standard-deviation length as the "yardstick".

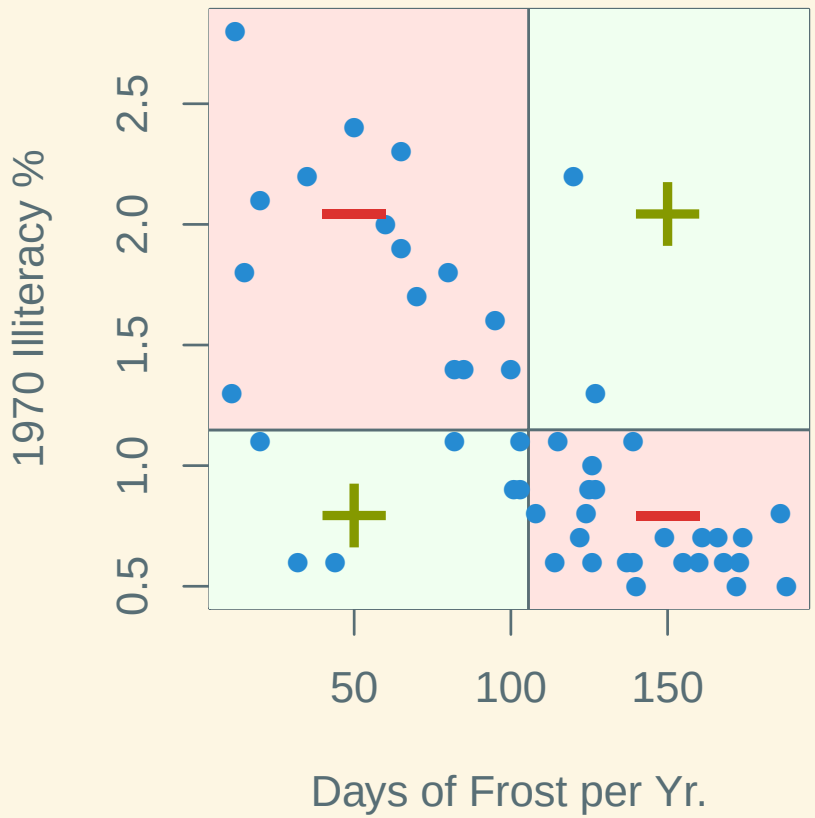
Pearson's Correlation Coefficient

Pearson's Correlation Coefficient (notation: r) is defined as:

$$r = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n-1}$$

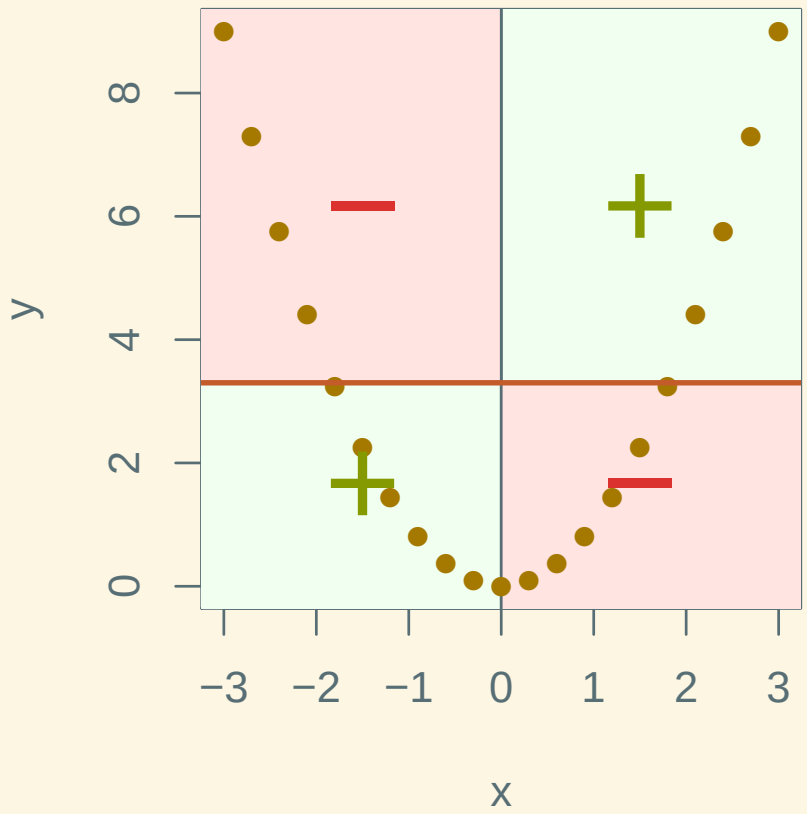
where z_{x_i} represents the z-score for the i^{th} data point in the x direction, and z_{y_i} the same in the y direction.

Linear Association



- r measures how well data fits a *straight line*.
- $r = 1$ when the data falls *exactly* on an upward sloping line; $r = -1$ when *exactly* on a downward sloping line.
- Here, $r = -0.68$

Nonlinear Association



$r = 0$

What Matters for Pearson's Correlation?

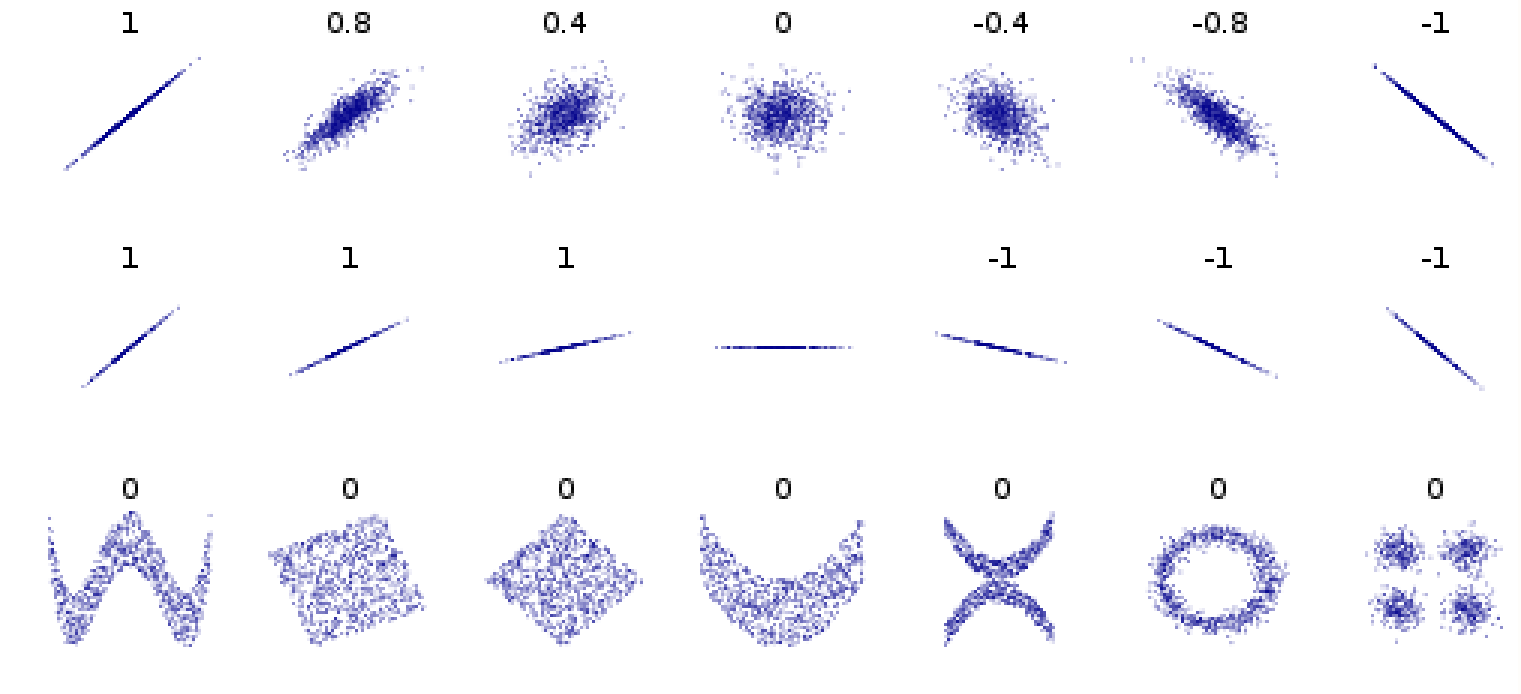


Figure: Hypothetical bivariate data and the corresponding Pearson's correlation coefficient

Explanatory vs. Response

$$r = \frac{1}{n-1} \sum_i z_{x_i} z_{y_i}$$

What will happen to r if we reverse the roles of X and Y? Nothing: Correlation is symmetric: only measures strength of association

Summary

- Correlation is one way to measure the **relationship** between two numeric variables.
- Ranges between -1 (perfect negative correlation) to +1 (perfect positive correlation)
- Measures **linear** relationships (does not capture more complex relationships)
  - Lack of correlation does *not* mean variables are unrelated, just that they don't always move in the same direction
- Correlation is *symmetric* (does not care which is explanatory and which is response), and *dimensionless* (invariant to choice of units).

Correlation $\neq$ Causation

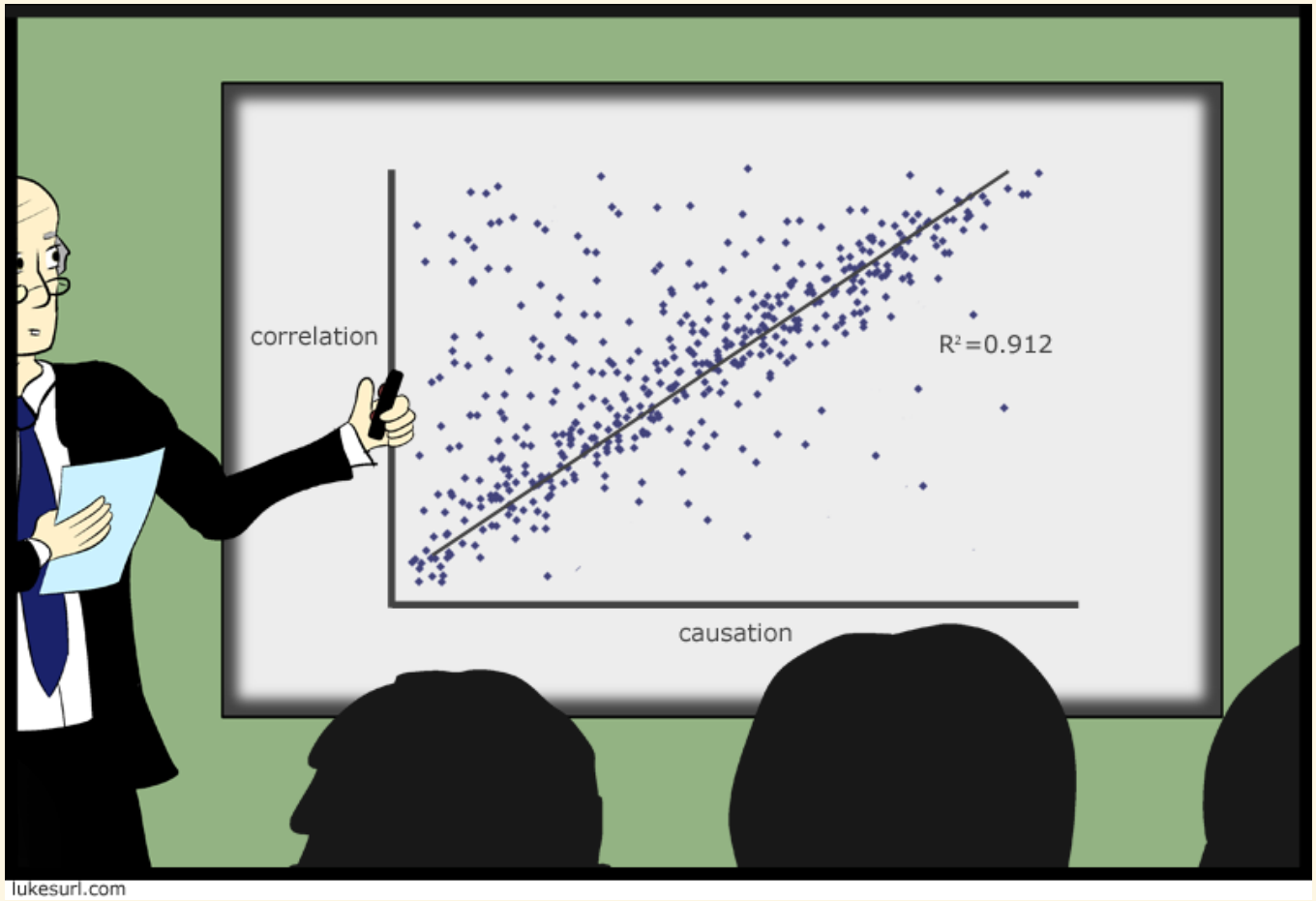


Figure: `http://www.lukesurl.com/`