What Can We Conclude from Data?

Zhiling Gu
February 3, 2026

Yale SCHOOL OF PUBLIC HEALTH

MISSISSIPPI STATE UNIVERSITY

Are the following Scientific claims the same?

- "Higher yield is linked to using fertilizer A" → **Association**
- "We predicts crop yield = X if given a unit dose of fertilizer A" → **Prediction**
- "Fertilizer A increase crop yield" → **Causation**
- "The fertilizer A study could be replicated" → **Reproducibility**

✳ **Data do not speak for themselves!**

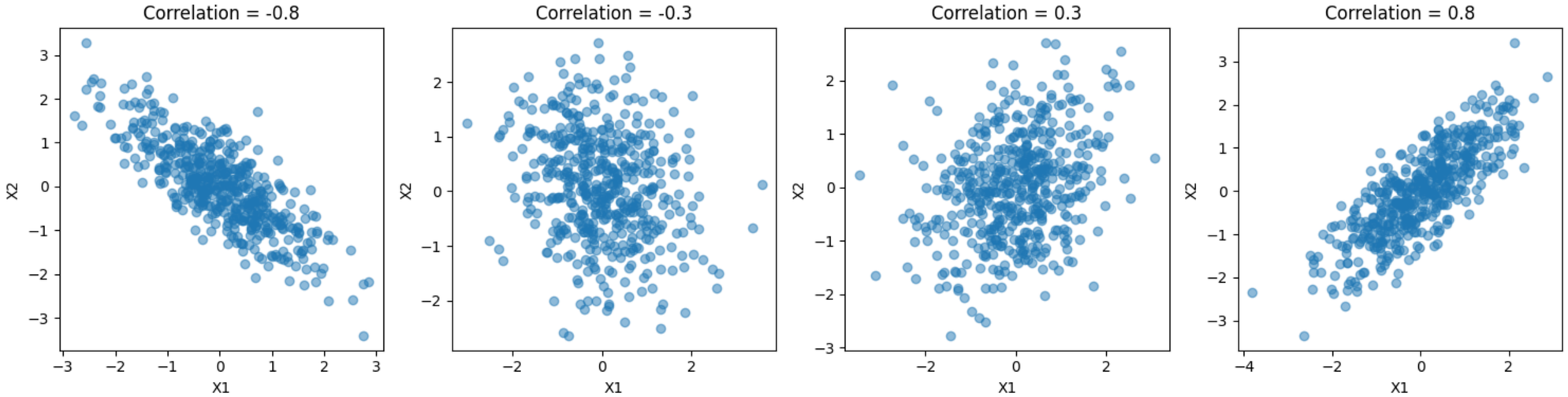
The conclusions we can draw depend on:

(i) The question, (ii) How data were generated, (iii)The analysis, (iv)Whether results hold up again

- Google Colab Code:

Association

Two variables are **associated** if they tend to **vary together.**



Q1: How would you compute the strength of the association?

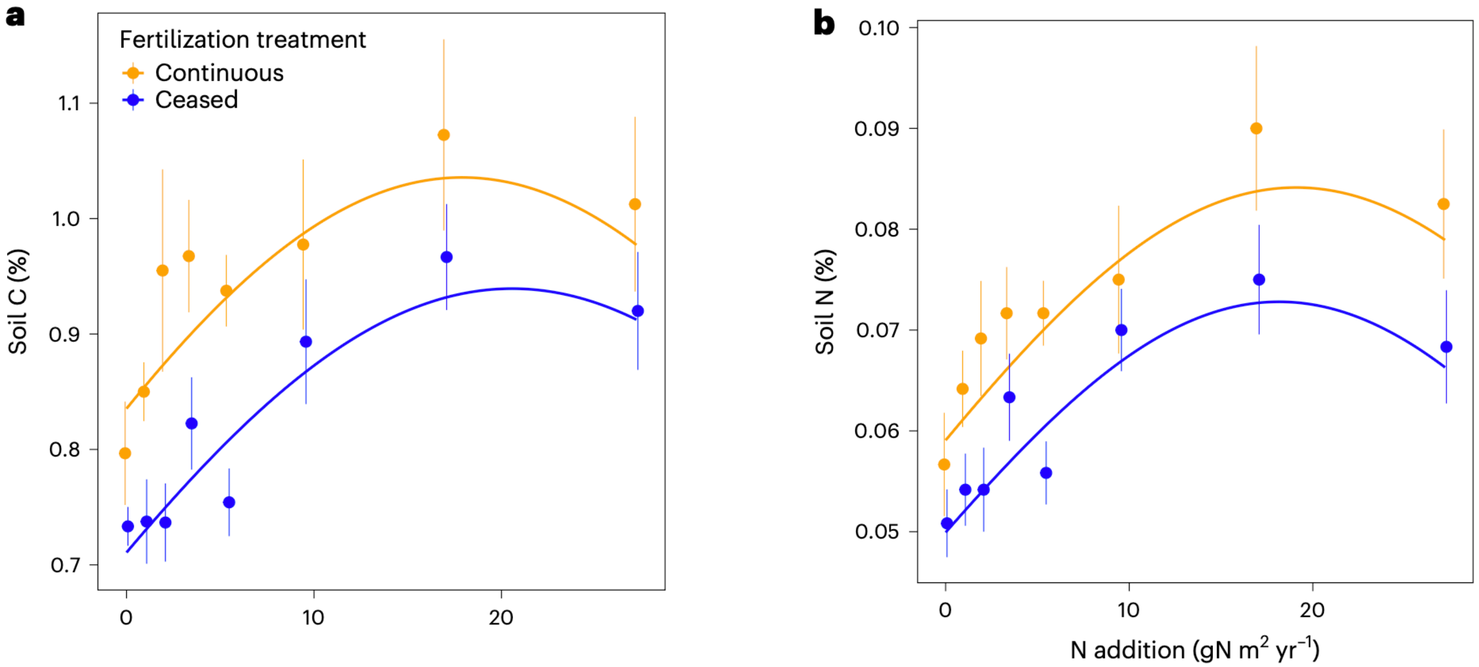
Correlation coefficient $\rho = \mathbb{E}\left(\left(X_1 - \mu_1\right)\left(X_2 - \mu_2\right)\right)/\sigma_1\sigma_2,\ -1 \leq \rho \leq 1$

✳ $\rho < 0$**: negatively correlated;** $\rho > 0$**: positively correlated;**

✳ $|\rho| \to 1$**: more strongly correlated**

Association

Two variables are **associated** if they tend to **vary together.**



✳ **Fertilizer** use *is associated with* **soil carbon** (C) & **nitrogen** (N) gains

Q2: Are they positively or negatively associated?

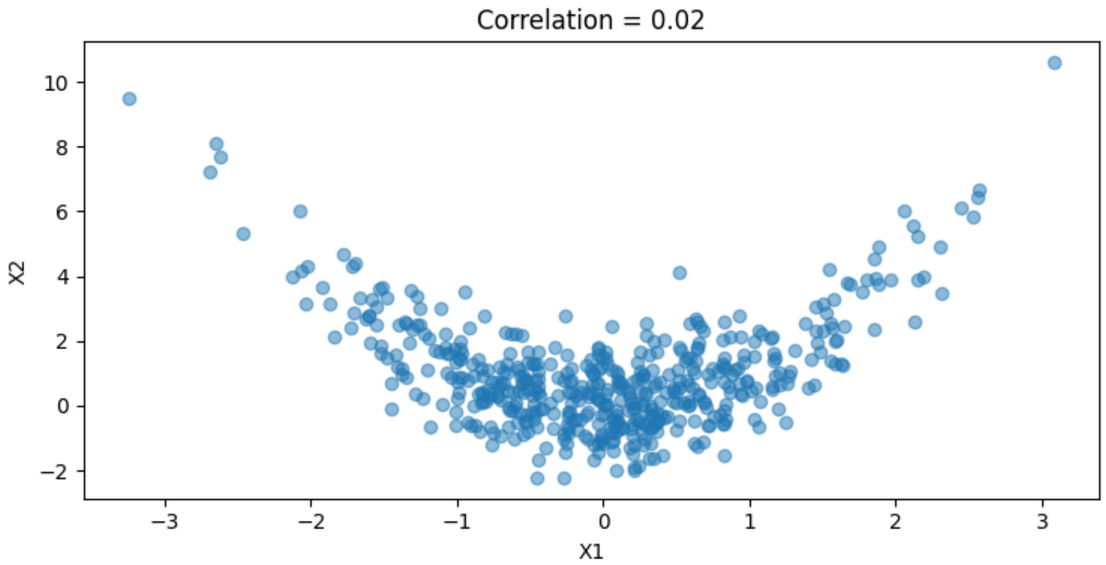
✳ **Correlation coefficient only measures linear association!**

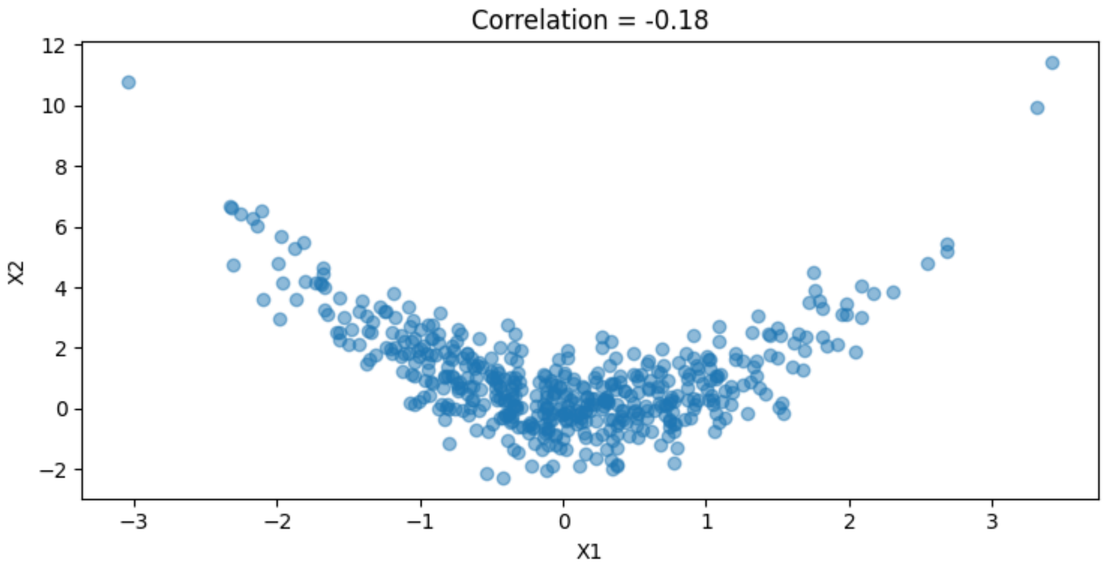
Seabloom, E.W., Hobbie, S.E., MacDougall, A.S. et al. Multidecadal persistence of soil carbon gains on retired cropland following fertilizer cessation. Nat. Geosci. 18, 1014–1019 (2025). https://doi.org/10.1038/s41561-025-01801-5

Association

Two variables are **associated** if they tend to **vary together.**

- In fact, we can design $X_2 = X_1^2 + \epsilon$, such that both signs of correlation can occur.

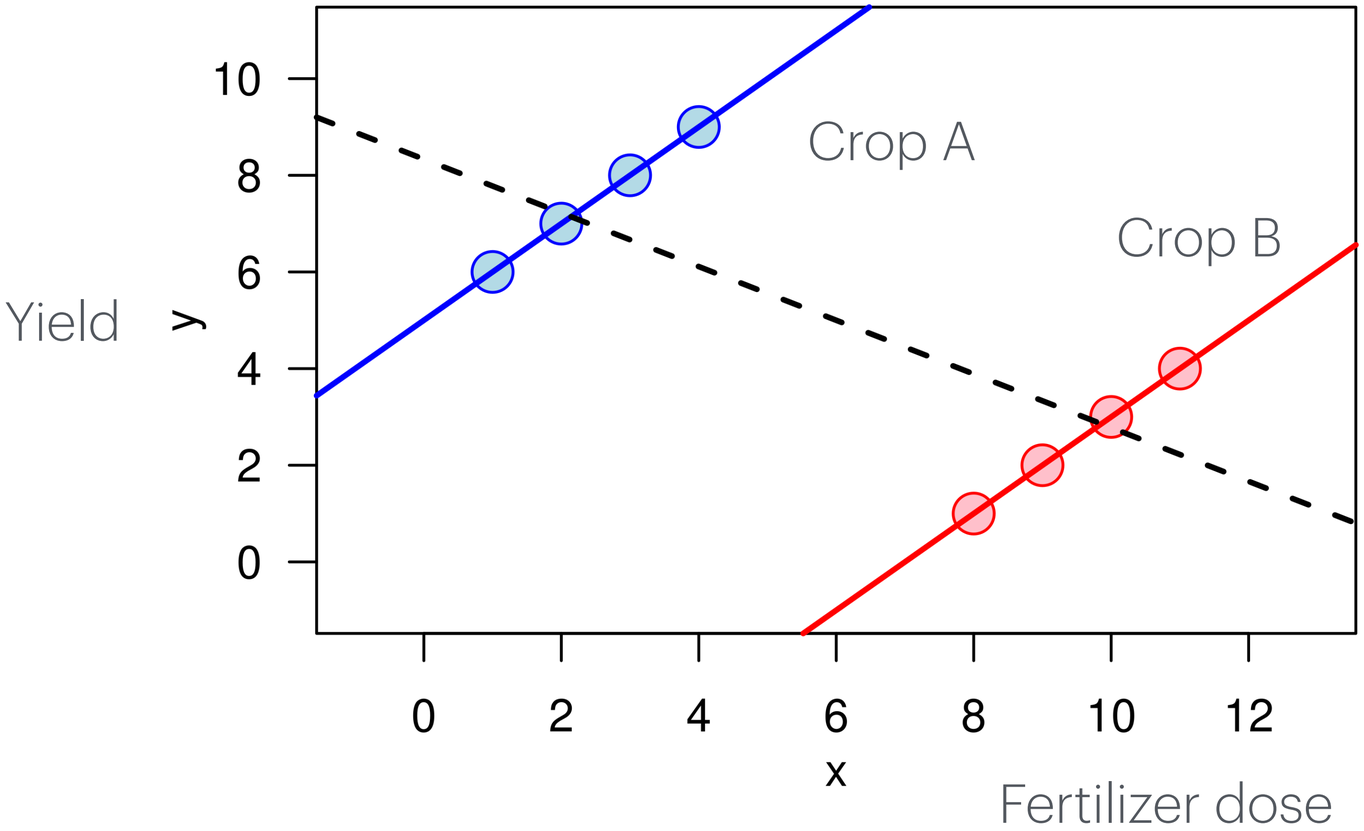




✳ **Watch out for nonlinear relationships - Do NOT only rely on correlation coefficient.**

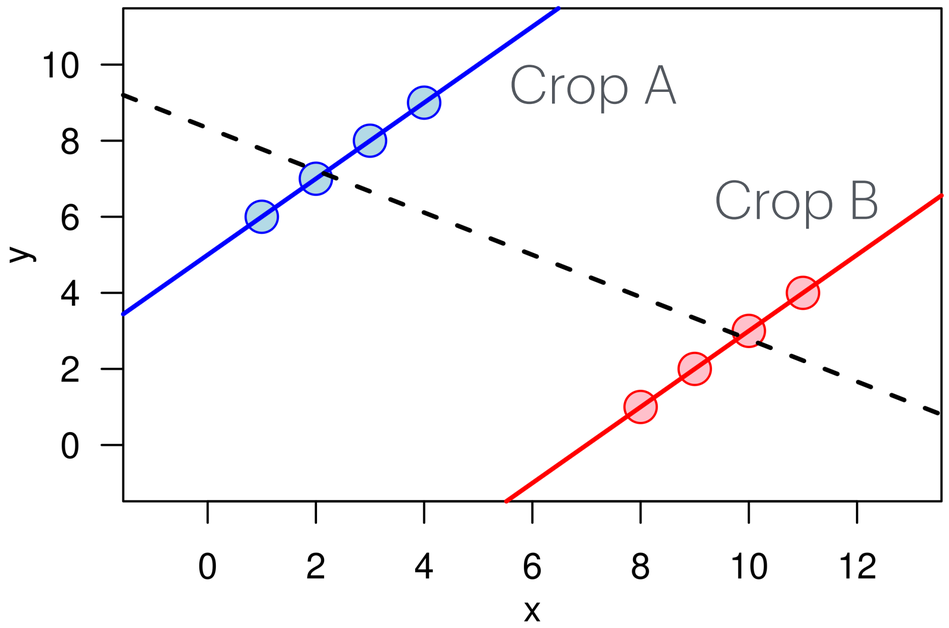
Association

Q3: Are Yield and Fertilizer use positively or negatively associated?

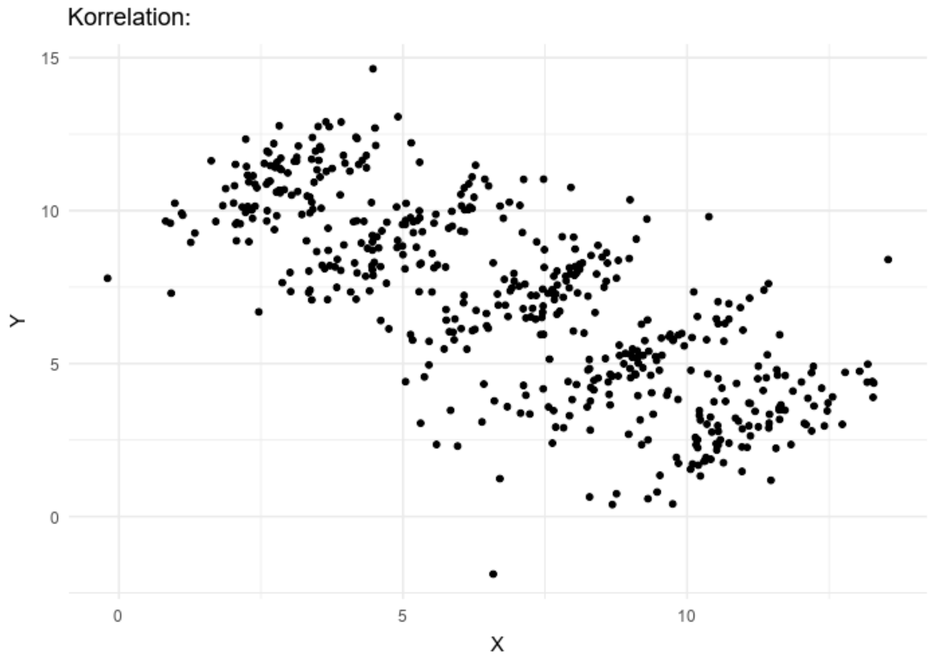


Association

Simpson's paradox

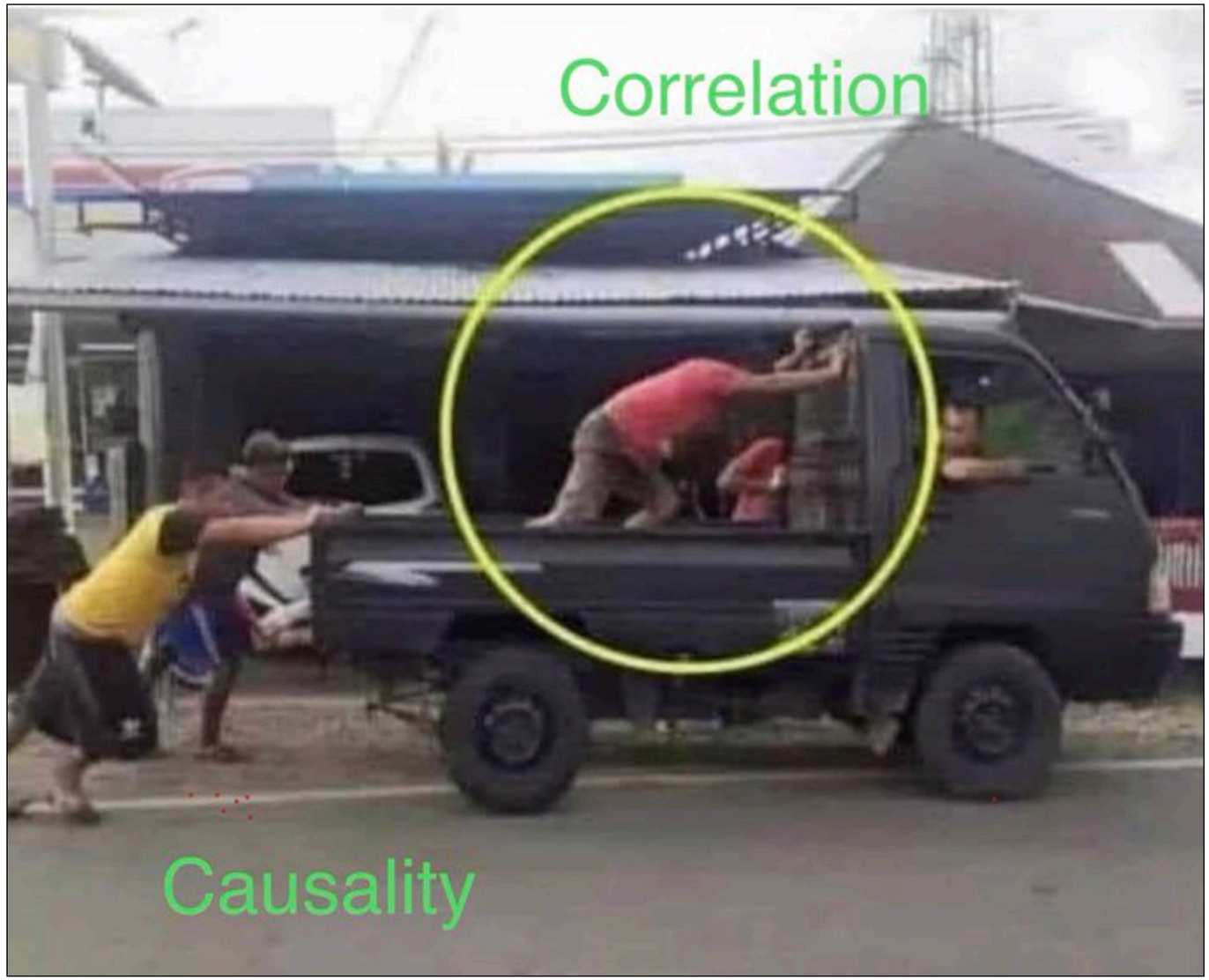


- A trend **appears** in several groups of data (positive asso. between x and Y)
- This trend **disappears or reverses** (negative asso. between X and Y) when the groups are combined



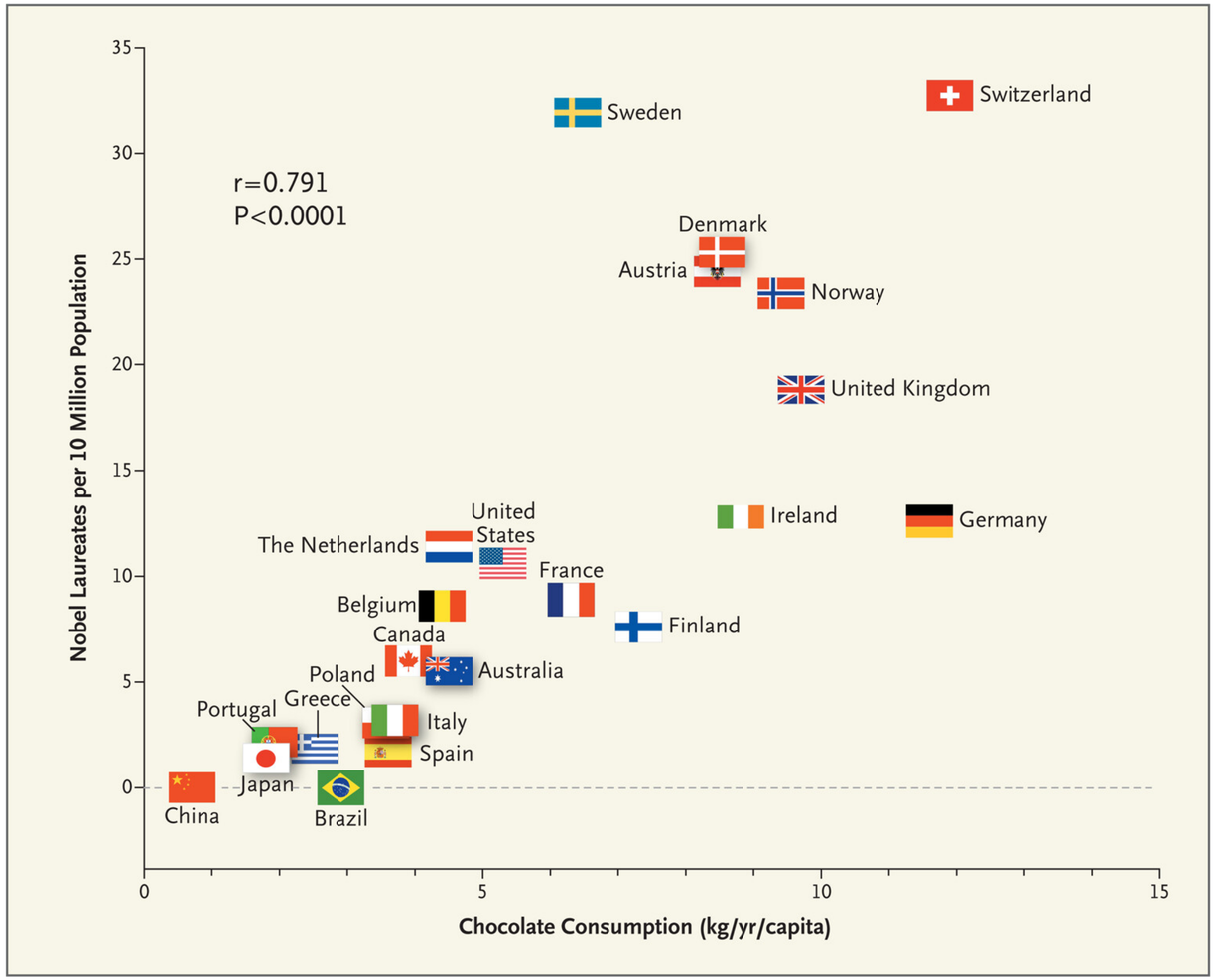
✳**Associations depend on how data are grouped**

✳**Ignoring important variables can lead to completely wrong conclusions.**



Association ≠ Causation

Example 1: Chocolate v.s. Nobel Prize



Choose your theory:

A. Eating chocolates produces Nobel Prize winners (improves cognitive functions)

B. Geniuses are more likely to eat chocolates

C. Wealthy people tend to receive higher education, and consume chocolates

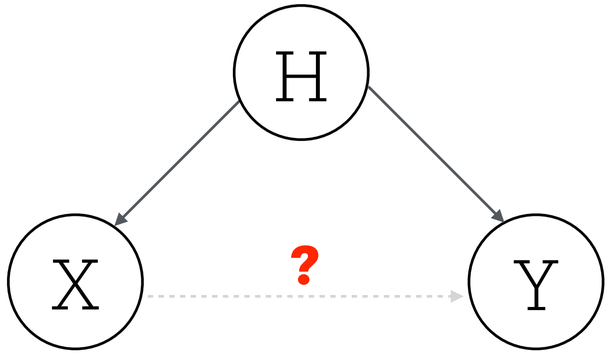
F. H. Messerli: Chocolate Consumption, Cognitive Function, and Nobel Laureates, N Engl J Med 2012

well... you might have your own theories...

X = chocolate consumption, Y = Nobel Prize

A. Chocolates produces Nobel Prize

B. Geniuses eat more chocolates



C. Hidden Confounder: H = Wealth

F. H. Messerli: Chocolate Consumption, Cognitive Function, and Nobel Laureates, N Engl J Med 2012

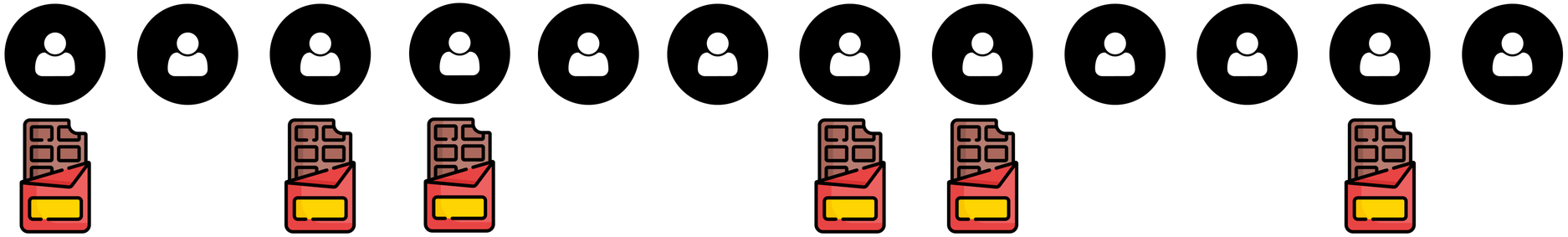
well... you might have your own theories...

It would be most helpful to do:

- an experiment
- a randomized controlled trial (RCT)
  - (often considered as) the gold-standard
  - forcing some people to eat lots and lots of chocolate!

Gold-standard: a randomized controlled trial (RCT)

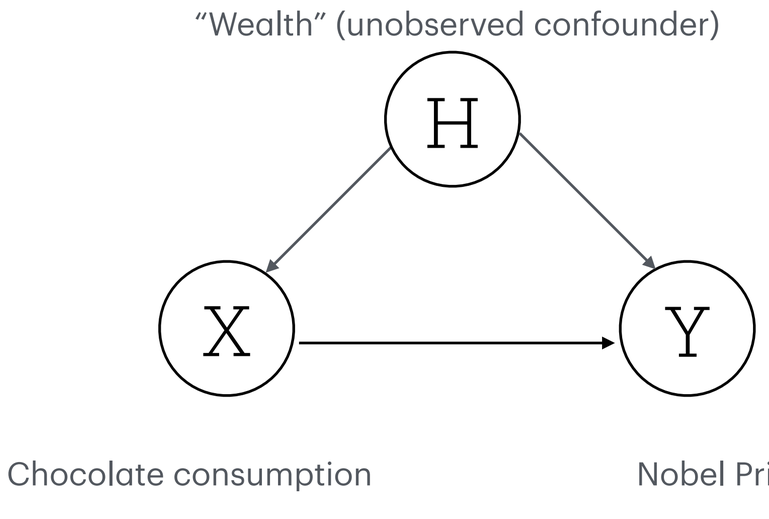
- two groups at random
  - (at random: to break dependencies to hidden variables)
  - force one group to eat lots of chocolate
  - ban the other group from eating chocolate at all
  - wait a lifetime to see what happens; and compare!

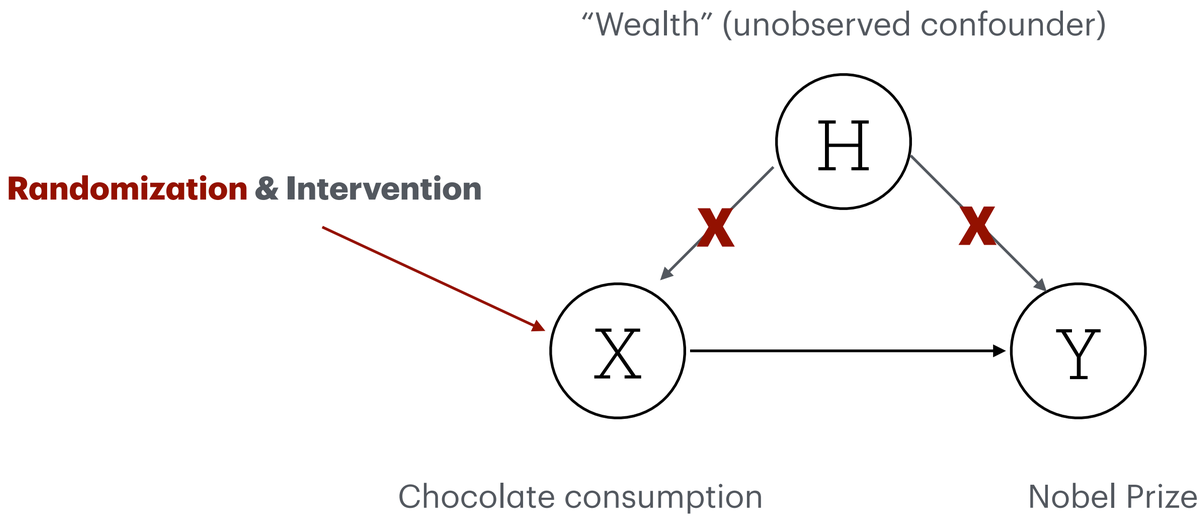


Like gold, RCT is expensive and heavy…

Gold-standard: a randomized controlled trial (RCT)

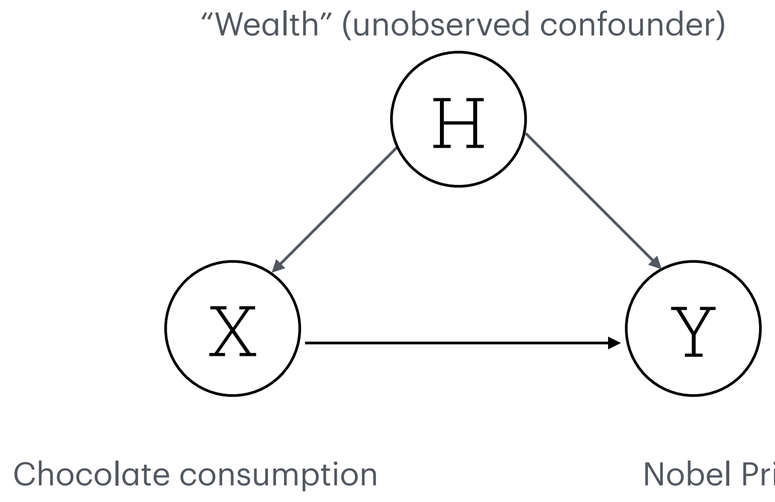
Why Randomization?

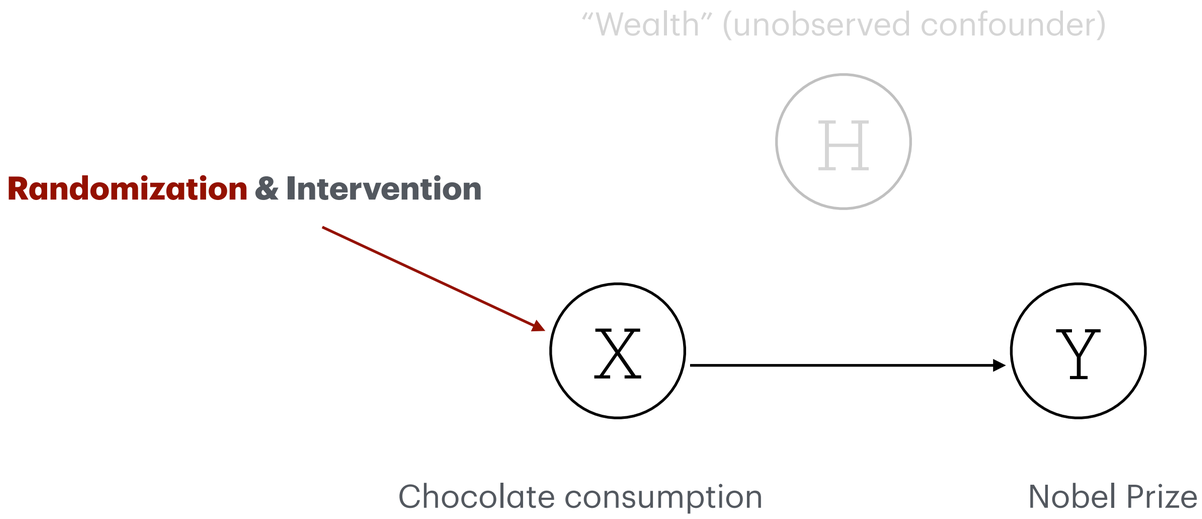




Gold-standard: a randomized controlled trial (RCT)

Why Randomization?

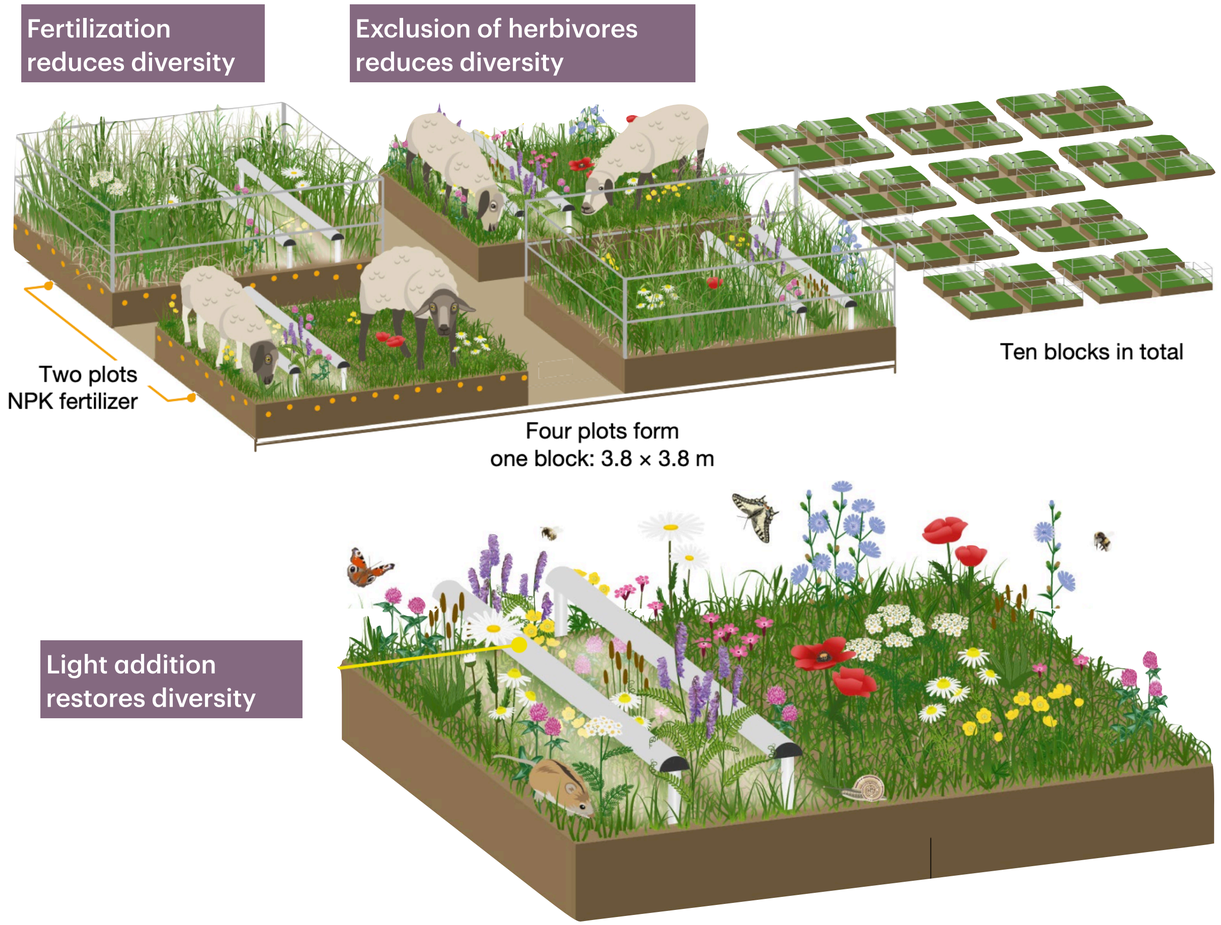




✳Randomization removes confounding because **Treatment** $X \perp$ **Confounder** H

✳Only works when sample size is not small!

Example 2: Would **light competition *drive*** herbivore and nutrient effects on plant diversity?



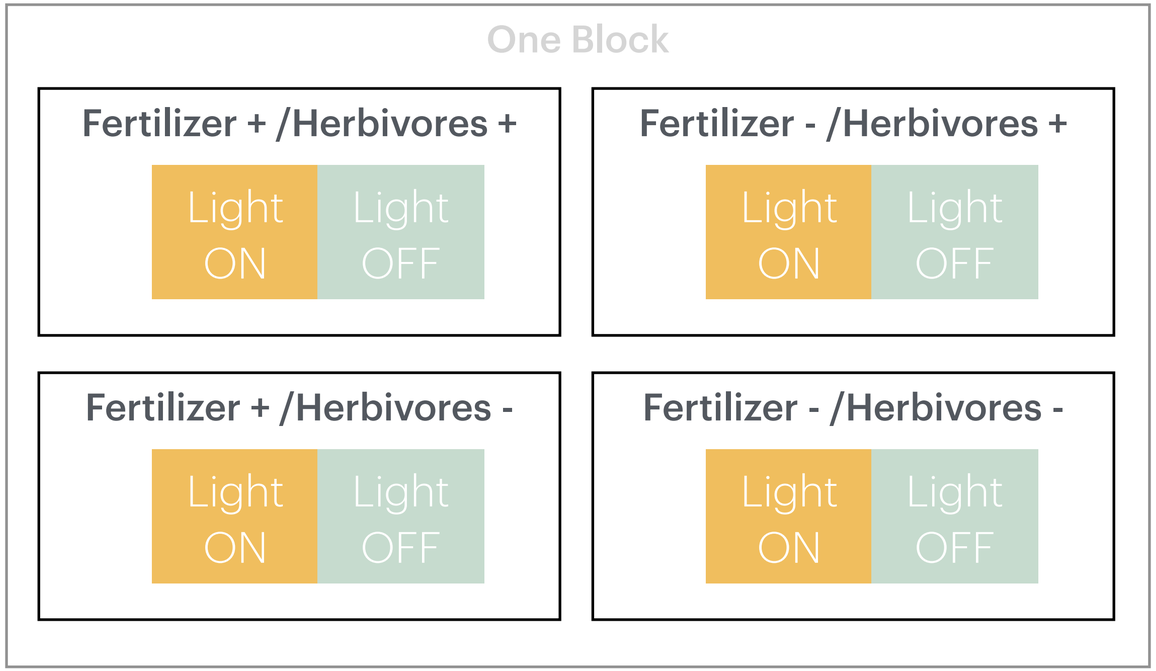
Eskelinen, A., Harpole, W.S., Jessen, MT. *et al.* Light competition drives herbivore and nutrient effects on plant diversity. *Nature* **611**, 301–305 (2022). https://doi.org/10.1038/s41586-022-05383-9

Example 2: light drive fertilizer, herbivore effect on plant diversity

A special RCT: Split-plot Factorial Randomized Design

- Fertilizer (+/-) Herbivore (+/-)
  $2^2 = 4$ plots per block

- Light (+/-)
  2 subplots per plot



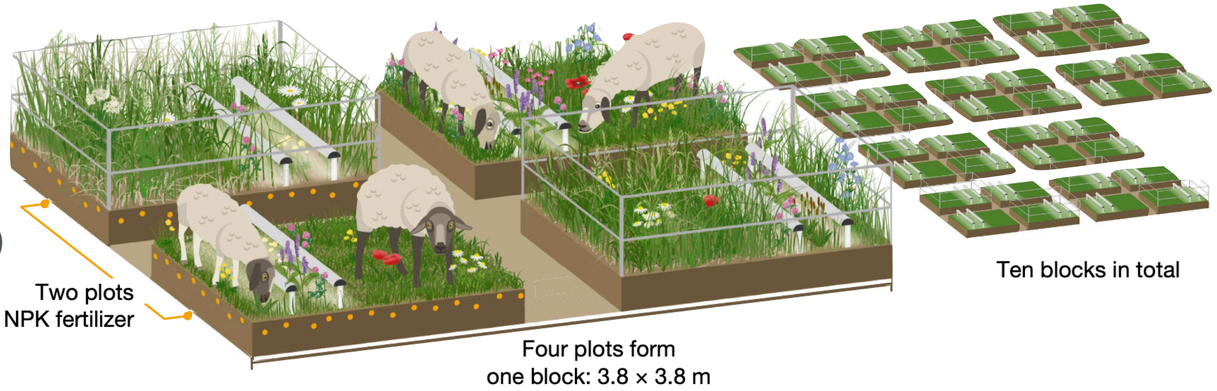
Sometimes we can't randomize everything at the smallest unit — and even when we can, it may be inefficient or impractical.

Eskelinen, A., Harpole, W.S., Jessen, MT. *et al.* Light competition drives herbivore and nutrient effects on plant diversity. *Nature* **611**, 301–305 (2022). https://doi.org/10.1038/s41586-022-05383-9

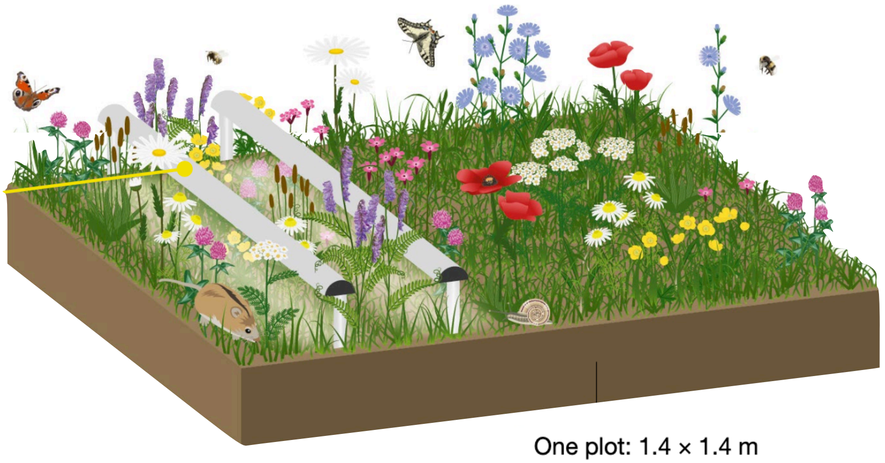
Example 2: light drive fertilizer, herbivore effect on plant diversity

A special RCT: Split-plot Factorial Randomized Design

- Fertilizer (+/-) Herbivore (+/-)
  $2^2 = 4$ plots per block



- Light (+/-)
  2 subplots per plot

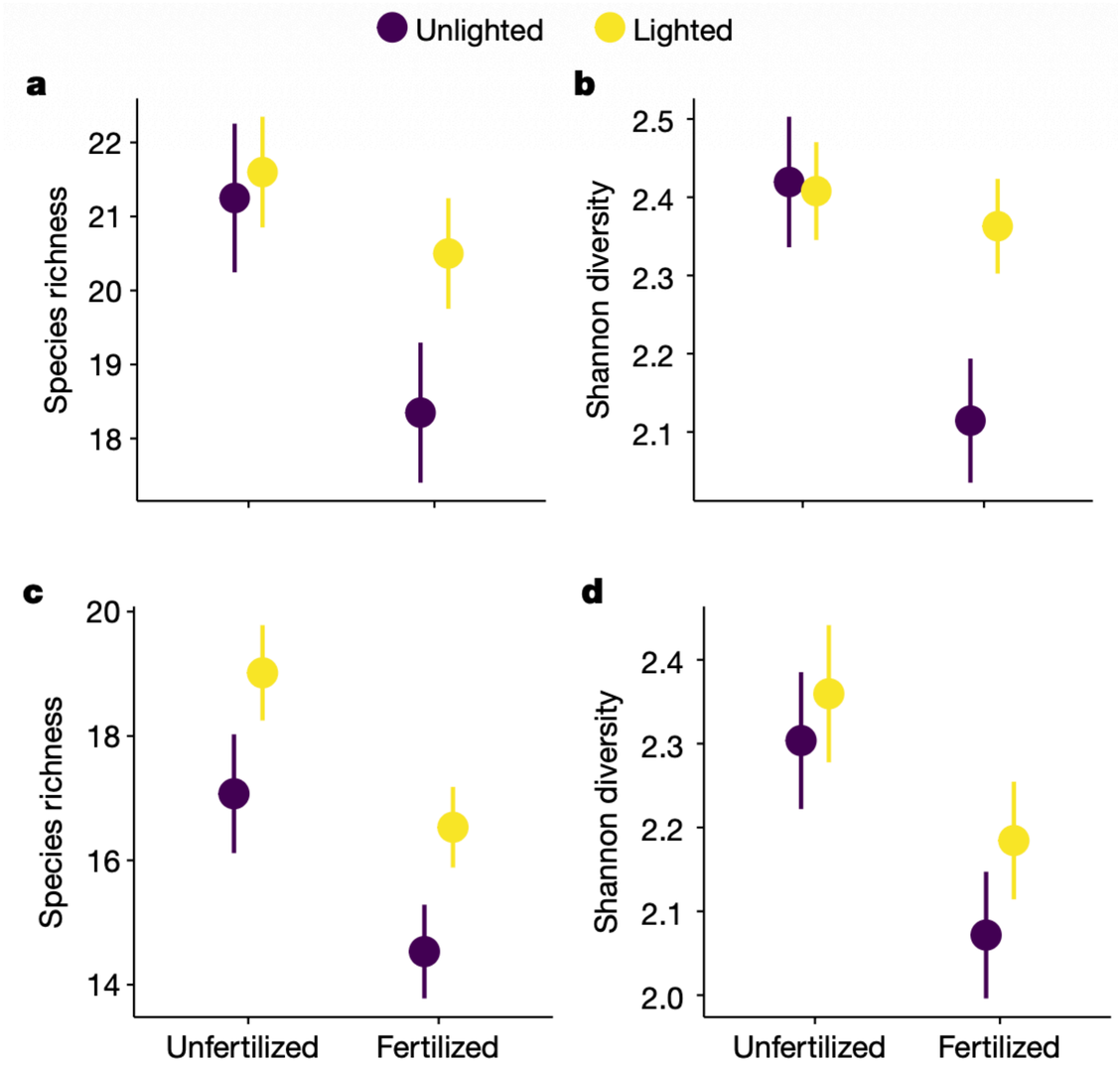


Eskelinen, A., Harpole, W.S., Jessen, MT. *et al.* Light competition drives herbivore and nutrient effects on plant diversity. *Nature* **611**, 301–305 (2022). https://doi.org/10.1038/s41586-022-05383-9

Example 2: light **drive** fertilizer, herbivore effect on plant diversity

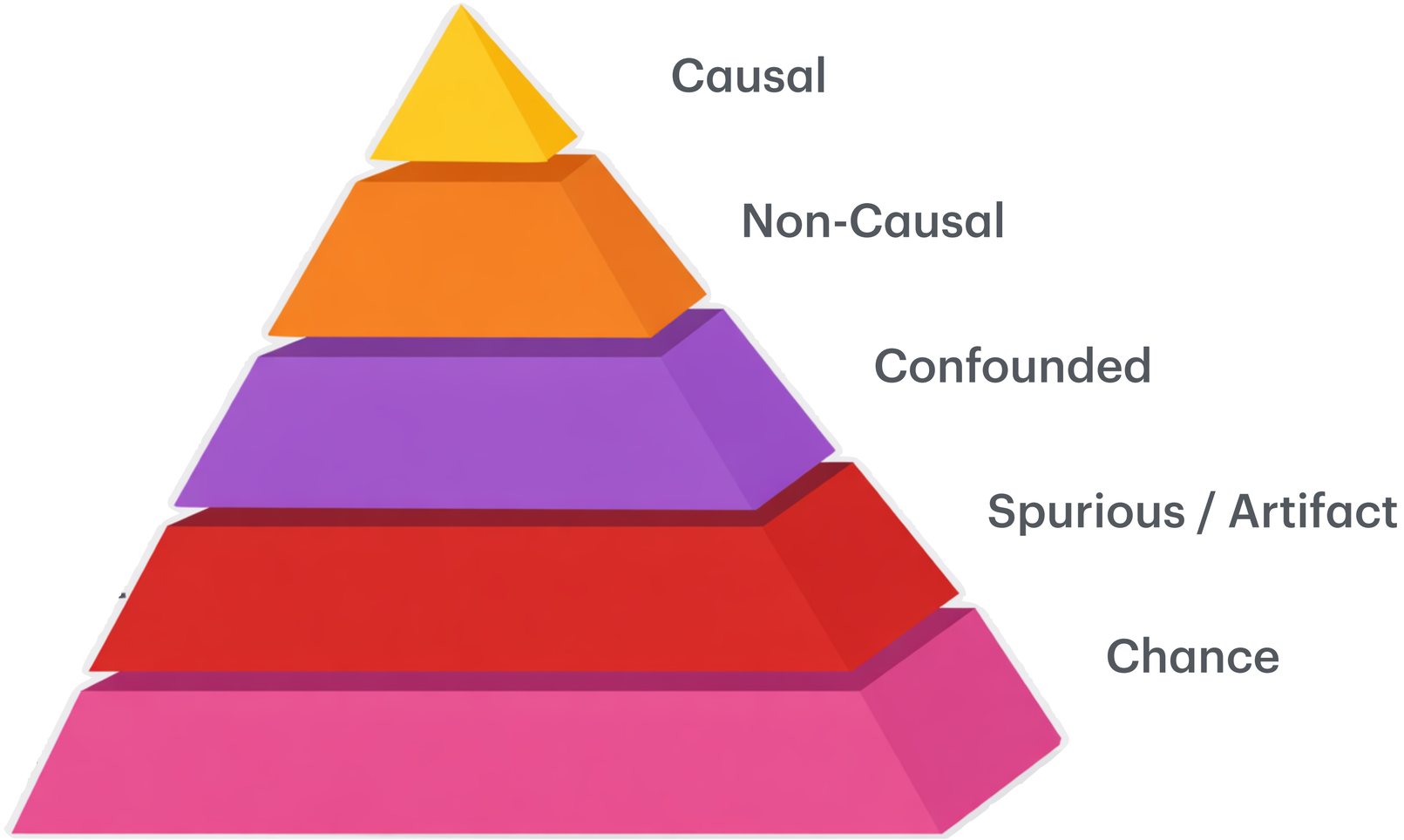
2017

2019



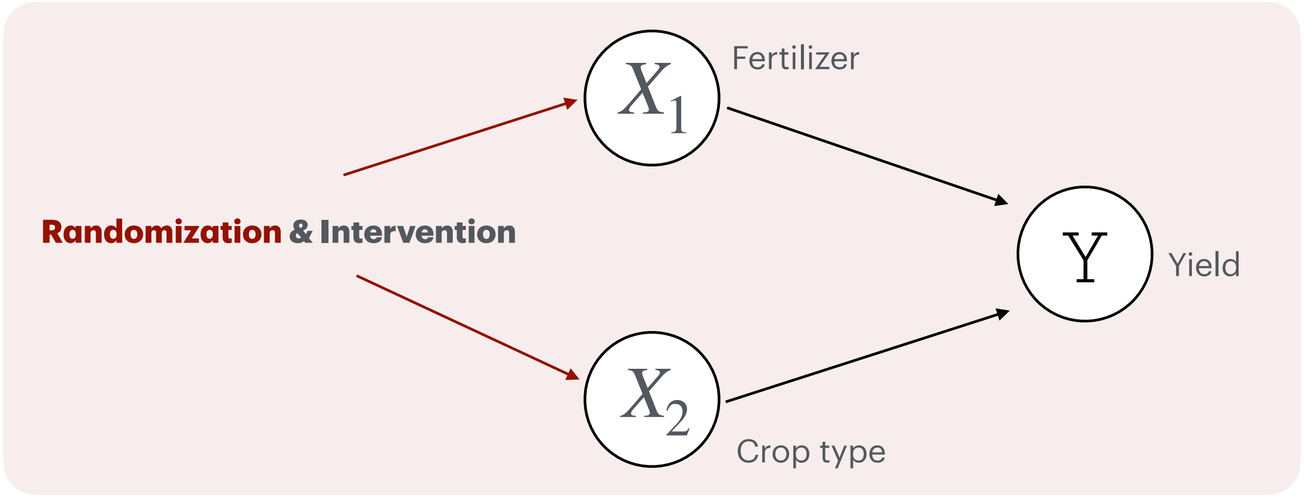
➡ **Restoring light** to understory plants in a natural grassland **mitigates the loss of plant diversity** that is caused by either nutrient enrichment or the absence of mammalian herbivores.

Pyramid of Association → Causality



Prediction

Assuming we have a simple linear regression model:



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Y: Variable to predict, e.g. Crop Yield

X_1: Fertilizer dose

X_2: Crop type

ϵ: Noise with zero mean

Prediction

Assuming we have a simple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Y: Variable to predict, e.g. **Crop Yield**

β_0: Global intercept

β_1: Linear Effect of **Fertilizer dose** X_1

β_2: Linear effect of **Crop type** X_2

ϵ: Noise with zero mean

We can estimate $\beta_1, \beta_2, \beta_3$ by **minimizing** the **Residual Sums of Squares (RSS)** = $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

Here i is the index of n observations: $\{(X_{i1}, X_{i2}, Y_i), i = 1, \ldots, n\}$.

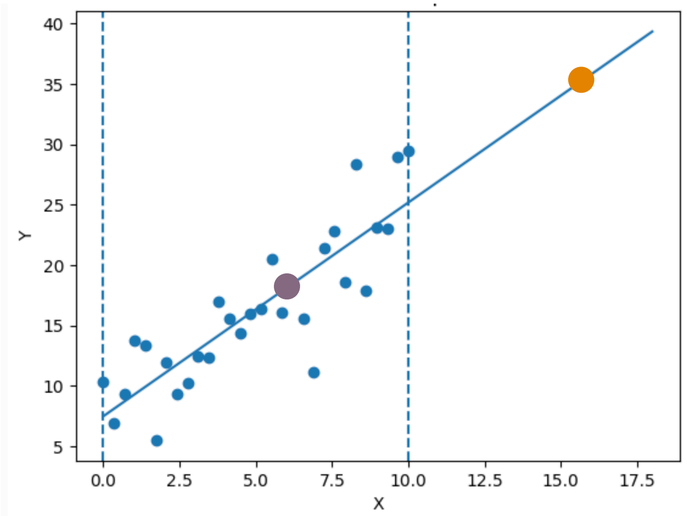
Prediction

For a given fertilizer dose $X_1 = x_1$ and crop type $X_2 = x_2$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$\hat{Y}$: Predicted Crop Yield $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$: Estimated linear coefficients

Prediction ⊇ Extrapolation



You study fertilizer dose (x_1) between 0–10 g N/m^2.

🌱 Predicting response at 6 g N/m^2→ **prediction**

🌱 Predicting at 15 g N/m^2 → **extrapolation (dangerous but sometimes inevitable!)** → additional assumptions needed.

Prediction with Uncertainty

For a given fertilizer dose $X_1 = x_1$ and crop type $X_2 = x_2$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$\hat{Y}$: Predicted Crop Yield

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$: Estimated linear coefficients

Q: $\hat{Y}$ is a point estimate — How certain are we about the estimate?

Prediction interval (PI): Predicting an interval of $Y \mid X_1 = x_1, X_2 = x_2$ **with 95% confidence**,

$$P(\hat{Y} - z_{\alpha/2}\hat{\xi}_n < Y < \hat{Y} + z_{\alpha/2}\hat{\xi}_n) \rightarrow 95\,\%$$

$z_{\alpha/2} = \phi^{-1}(0.975) \approx 1.68$
if noise is normally distributed

$\hat{\xi}_n$: an estimate of uncertainty

Prediction with Uncertainty

For a given fertilizer dose $X_1 = x_1$ and crop type $X_2 = x_2$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$\hat{Y}$: Predicted Crop Yield $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$: Estimated linear coefficients

Prediction interval (PI): Predicting an interval of Y **with 95% confidence**,

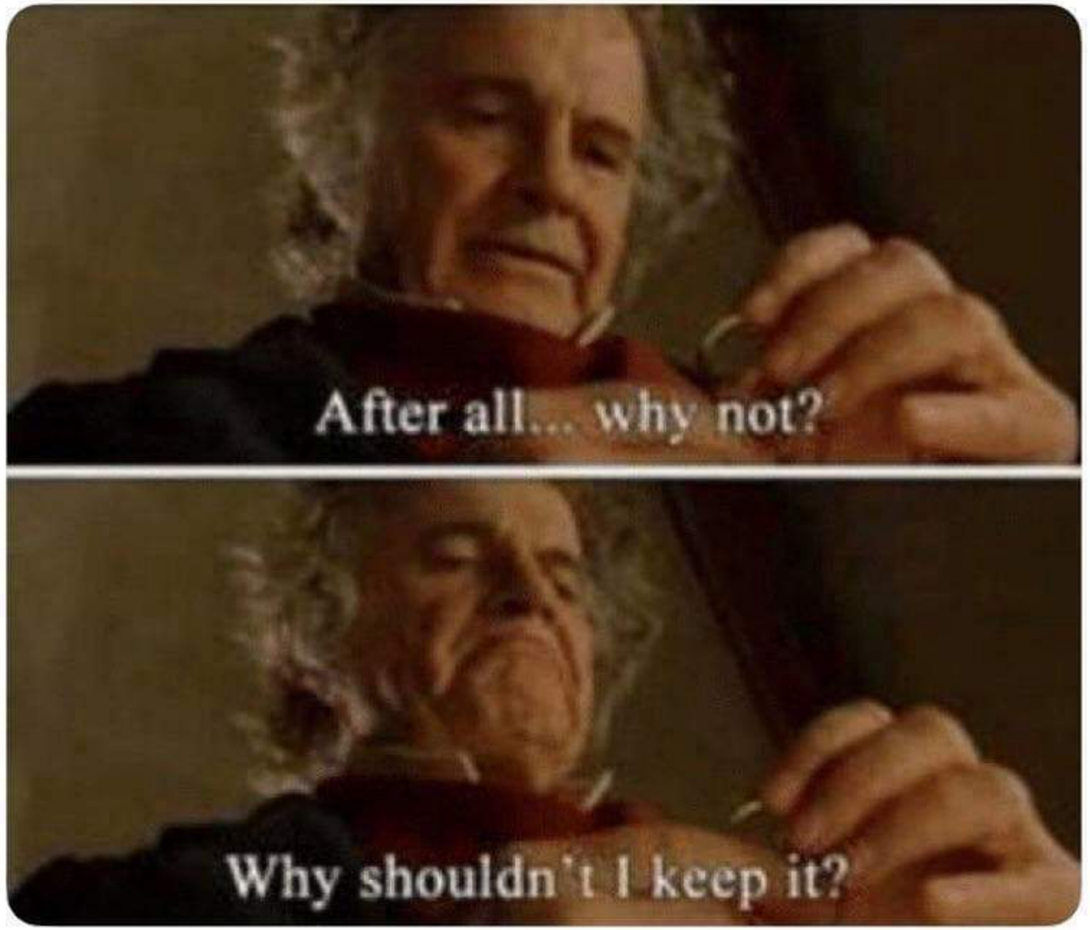
$$P(\hat{Y} - z_{\alpha/2}\hat{\xi}_n < Y < \hat{Y} + z_{\alpha/2}\hat{\xi}_n) \to 95\,\%\,, \text{as}\, n \to \infty$$

✳**This coverage rate is guaranteed asymptotically, when**

✓**Sample size n is large enough**

✓**Our assumption of simple linear model is correct**

When your statistically significant result is clearly driven by a single outlier



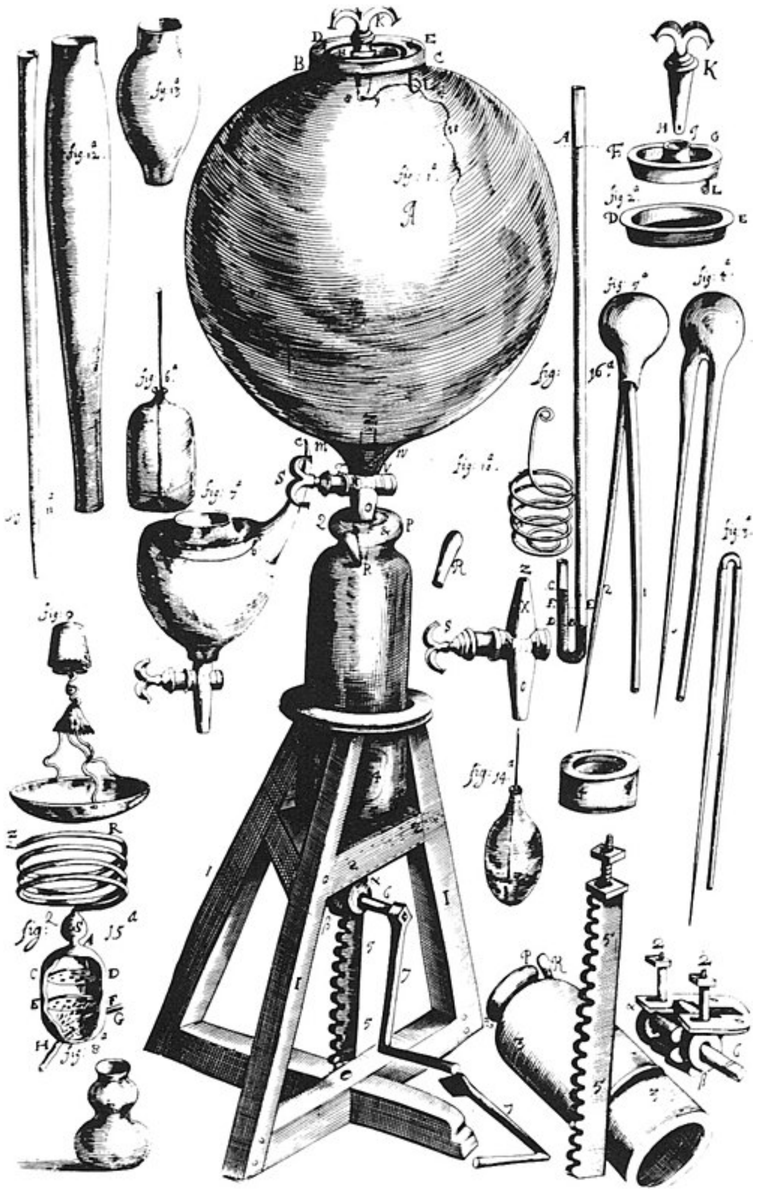
Reproducibility

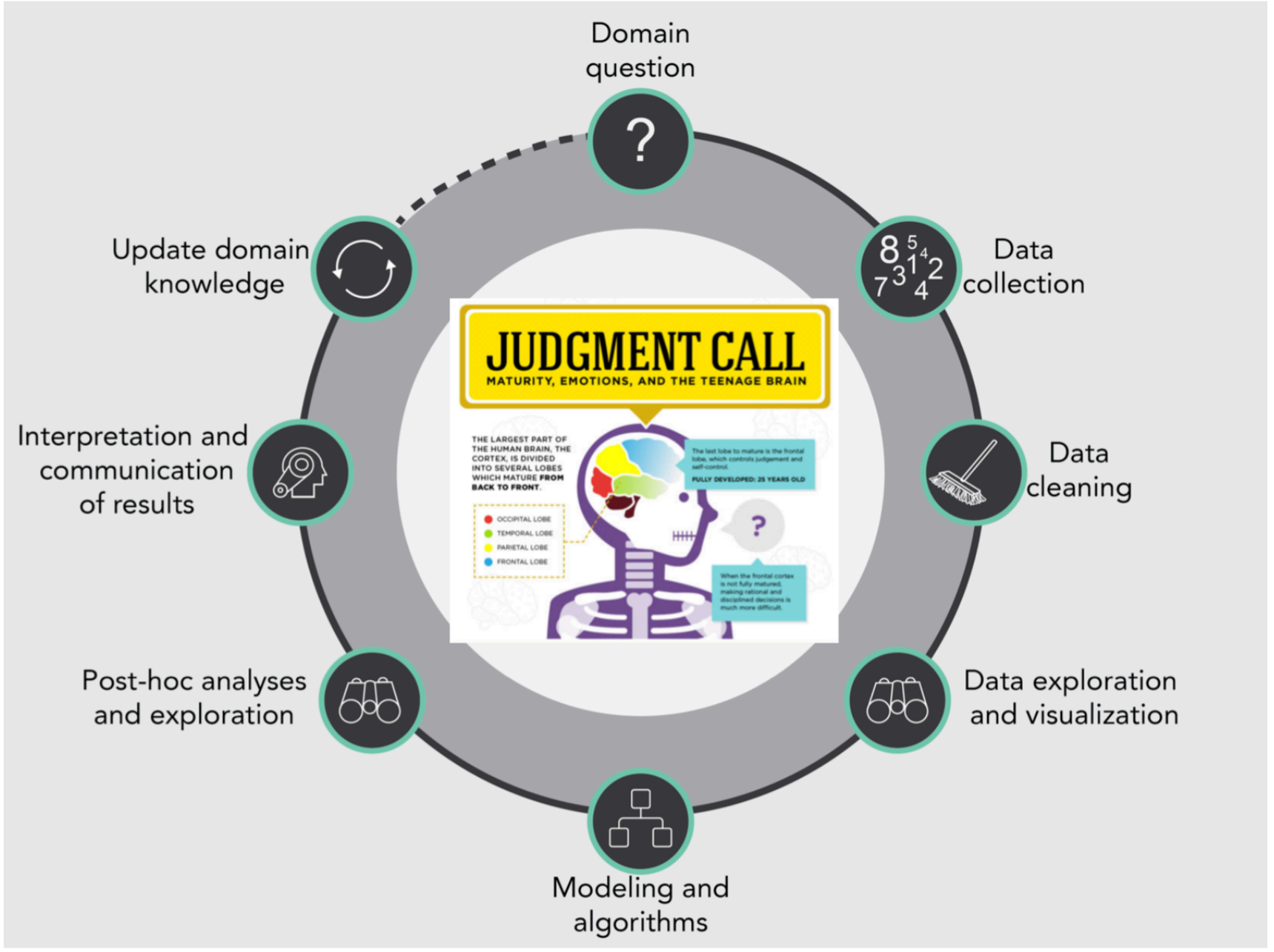
Results obtained by an experiment or an observational study or in a statistical analysis of a data set **should be achieved again with a high degree of reliability when the study is replicated** by an independent group of researchers.

Scientific results should be documented in such a way that their deduction is fully transparent:

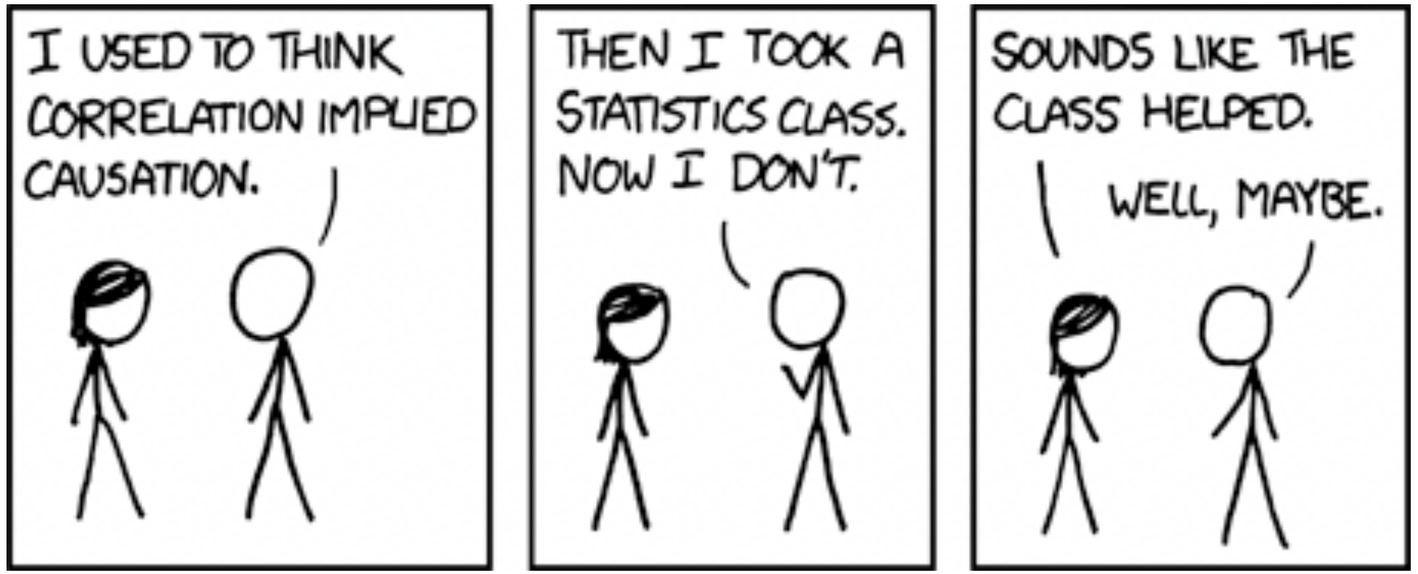
- **a detailed description** of the methods used to obtain the data

- **full dataset and the code** to calculate the results easily accessible




Domain question
Data collection
Data cleaning
Data exploration and visualization
Modeling and algorithms
Post-hoc analyses and exploration
Interpretation and communication of results
Update domain knowledge
JUDGMENT CALL
MATURITY, EMOTIONS, AND THE TEENAGE BRAIN
THE LARGEST PART OF THE HUMAN BRAIN, THE CORTEX, IS DIVIDED INTO SEVERAL LOBES WHICH MATURE FROM BACK TO FRONT.
OCCIPITAL LOBE
TEMPORAL LOBE
PARIETAL LOBE
FRONTAL LOBE
The last lobe to mature is the frontal lobe, which controls judgement and self-control.
FULLY DEVELOPED: 25 YEARS OLD
When the frontal cortex is not fully matured, making rational and disciplined decisions is much more difficult.

Well maybe...


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

References

Notes: https://people.math.ethz.ch/~buhlmann/teaching/mannheim.html

Code: https://colab.research.google.com/drive/1mDyESZUHqfmnxXOBgYMralkMZTa1tV6p?usp=sharing