Photo credit: Dave DiCello

17-803 Empirical Methods

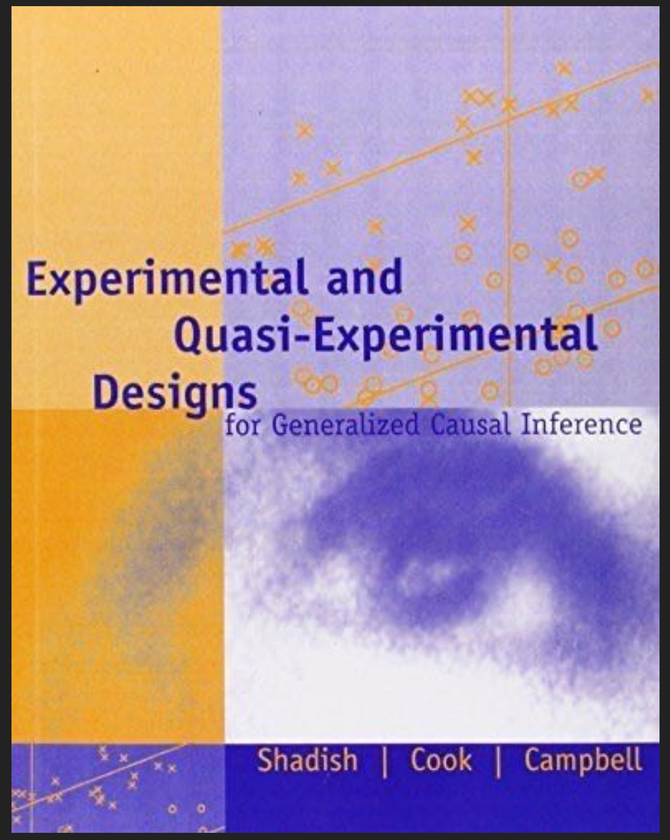
Bogdan Vasilescu, Institute for Software Research

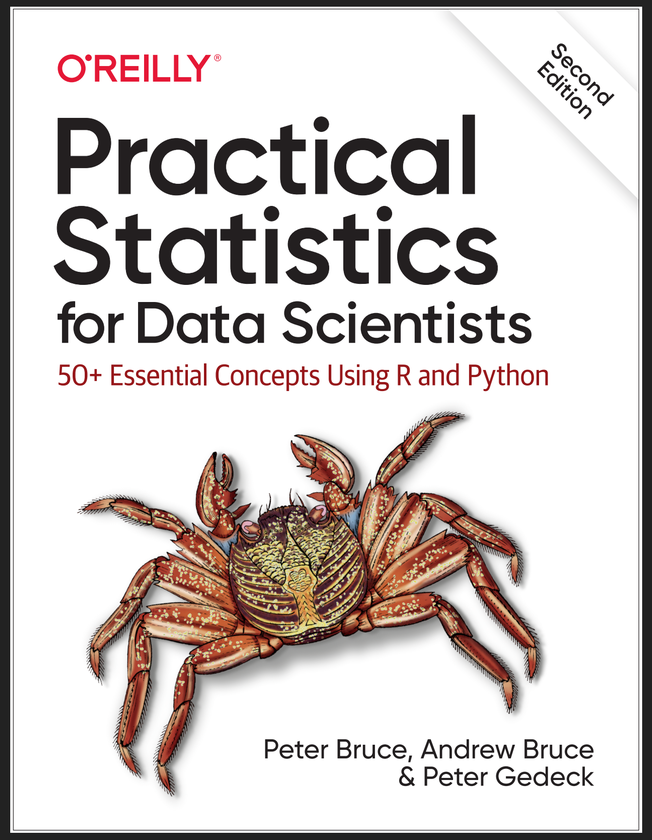
Designing Experiments (Part II)

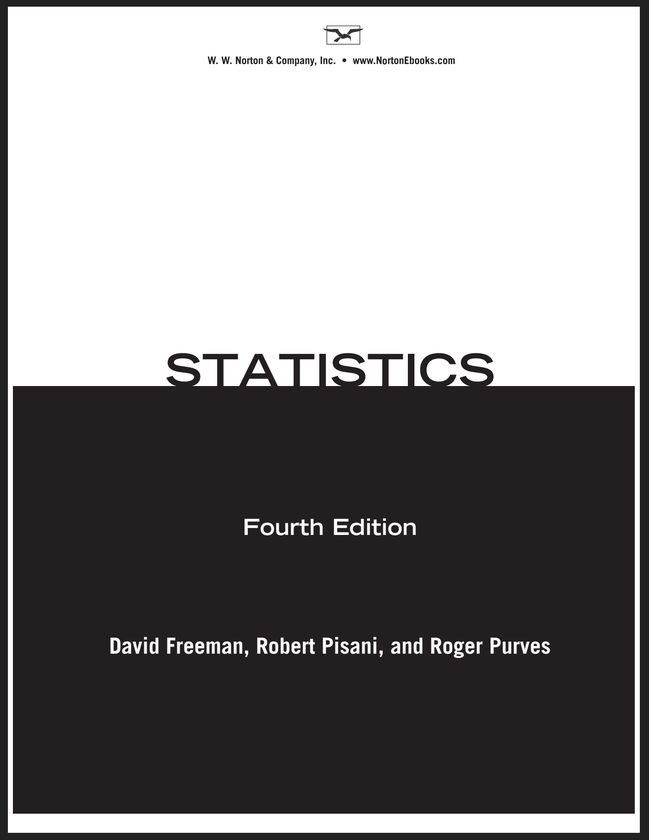
Tuesday, March 16, 2021

Outline for Today

- Example experiment papers
- Experimental design







Example papers

WSDM (Conference on Web Search and Data Mining) Experiment

- Setup
  - Four committee members reviewed each paper
  - Two single blind, two double blind

- Results
  - “Reviewers in the single-blind condition [...] preferentially bid for papers from top universities and companies.”
  - “Single-blind reviewers are significantly more likely than their double-blind counterparts to recommend for acceptance papers from famous authors [odds multiplier 1.64], top universities [1.58], and top companies [2.10].”

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. Proceedings of the National Academy of Sciences, 114(48), 12708-12713.

NeurIPS (Conference on Neural Information Processing Systems) Experiment

- Setup
  - Organizers split the program committee down the middle
  - Most submitted papers were assigned to a single side
  - 10% of submissions (166) were reviewed by both halves of the committee

- Results
  - “most papers [57%] at NeurIPS would be rejected if one reran the conference review process (with a 95% confidence interval of 40-75%)”

http://blog.mrtz.org/2014/12/15/the-nips-experiment.html

Teaching Formal Methods Experiment

- Two classes of students at Miami University of Ohio that studied object-oriented (OO) design in a one semester course:
  - Control group (random sample): OO design class
  - Treatment group (volunteers): OO design class + formal methods
    - No statistical difference between the abilities of the two groups on standardized ACT pre-tests
- As project, both classes were assigned the development of an elevator system
  - Students had to hand in:
    - functioning executable + source code
    - (+ formal specification written using first-order logic)

Sobel, A. E. K., & Clarkson, M. R. (2002). Formal methods application: An empirical tale of software development. IEEE Transactions on Software Engineering, 28(3), 308-320.

Teaching Formal Methods Experiment

- Standard set of test cases:
  - 45.5% of control teams passed all tests
  - 100% of treatment teams

- Conclusions:
  - “formal methods students had increased complex-problem solving skills”
  - “the use of formal methods during software development produces ‘better’ programs”

Sobel, A. E. K., & Clarkson, M. R. (2002). Formal methods application: An empirical tale of software development. IEEE Transactions on Software Engineering, 28(3), 308-320.

Teaching Formal Methods Experiment

- “Unfortunately, the paper contains several subtle problems. The reader unfamiliar with the basic principles of experimental psychology may easily miss them and interpret the results incorrectly. Not only do we wish to point out these problems, but we also aim to illustrate what to look for when drawing conclusions from controlled experiments.”

Berry, D. M., & Tichy, W. F. (2003). Comments on “Formal methods application: an empirical tale of software development". IEEE Transactions on Software Engineering, 29(6), 567-571.

Teaching Formal Methods Experiment

- Confounding variables:
  - differences in motivation (treatment group volunteers more motivated)
  - differences in exposure (treatment group more instruction)
  - differences in learning style (treatment group better learners)
  - differences in skills (outside of ACT)
- Novelty effects
- …

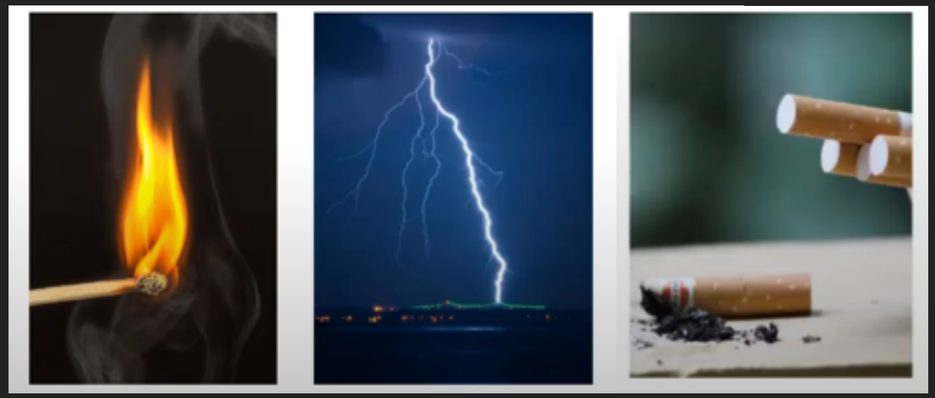
Berry, D. M., & Tichy, W. F. (2003). Comments on “Formal methods application: an empirical tale of software development". IEEE Transactions on Software Engineering, 29(6), 567-571.

Causal relationships

Cause

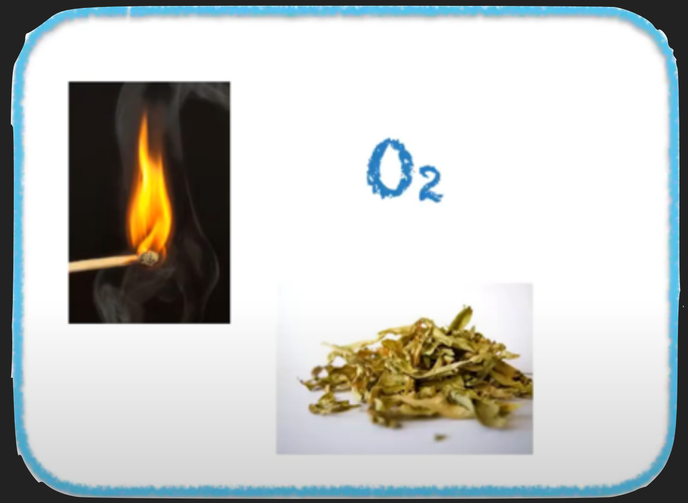
- *inus condition* – "insufficient but nonredundant part of an unnecessary but sufficient condition"

- Example: match to start a forest fire
- Fires can start even without matches
  - → Match is not a necessary condition
- Matches don't always start forest fires (e.g., not on long enough, rainy weather)
  - → Match is not a sufficient condition



Cause

- *inus condition* – "insufficient but nonredundant part of an unnecessary but sufficient condition"
- Match is part of a bigger constellation of conditions without which a fire would not result
  - Insufficient: needs oxygen, dry leaves, etc
  - Nonredundant: needs to add something unique besides oxygen, dry leaves, etc



Effect

- Counterfactual: what would have happened to these subjects had the cause not been present?
  - What did happen when people received a treatment, vs
  - What would have happened to those same people if they simultaneously had not received the treatment ("counterfactual", i.e., contrary to fact)
  - Effect is distance between the two

- Can't observe, must infer / approximate.

Experimental design:

- Creating a high-quality but necessarily imperfect source of counterfactual inference
- Understanding how this source differs from the treatment condition

Ingredients for Establishing a Causal Relationship?

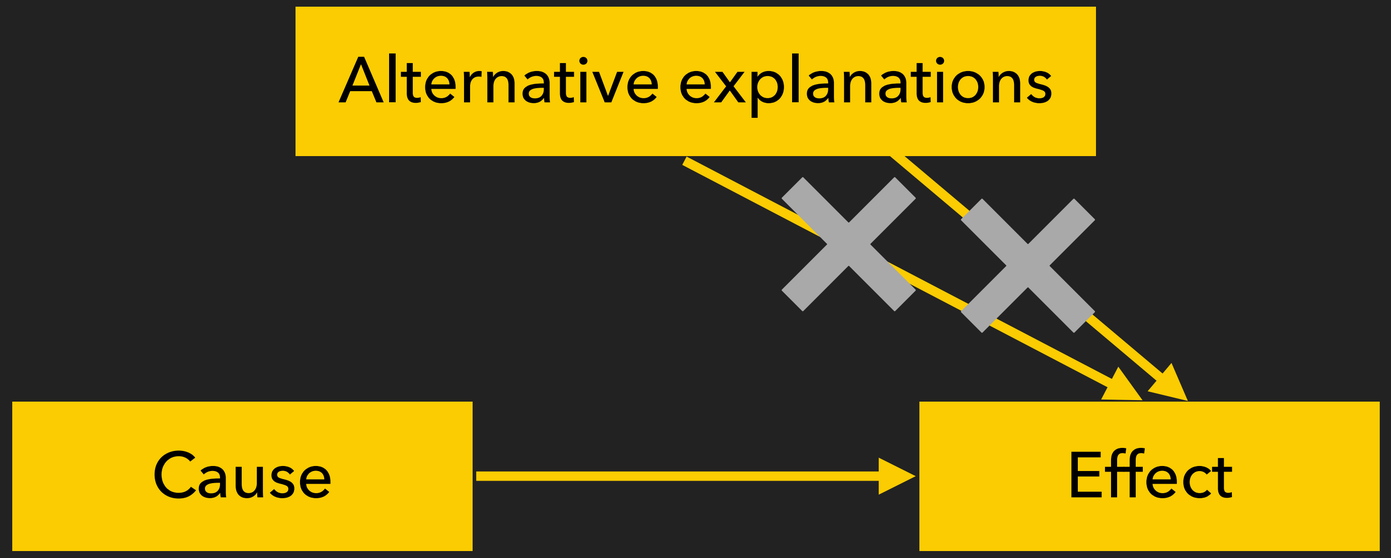
Ingredients for Establishing a Causal Relationship

The cause preceded the effect

The cause was related to the effect

We can find no plausible alternative explanation for the effect other than the cause

Ingredients for Establishing a Causal Relationship



Note how this mirror what happens in experiments.

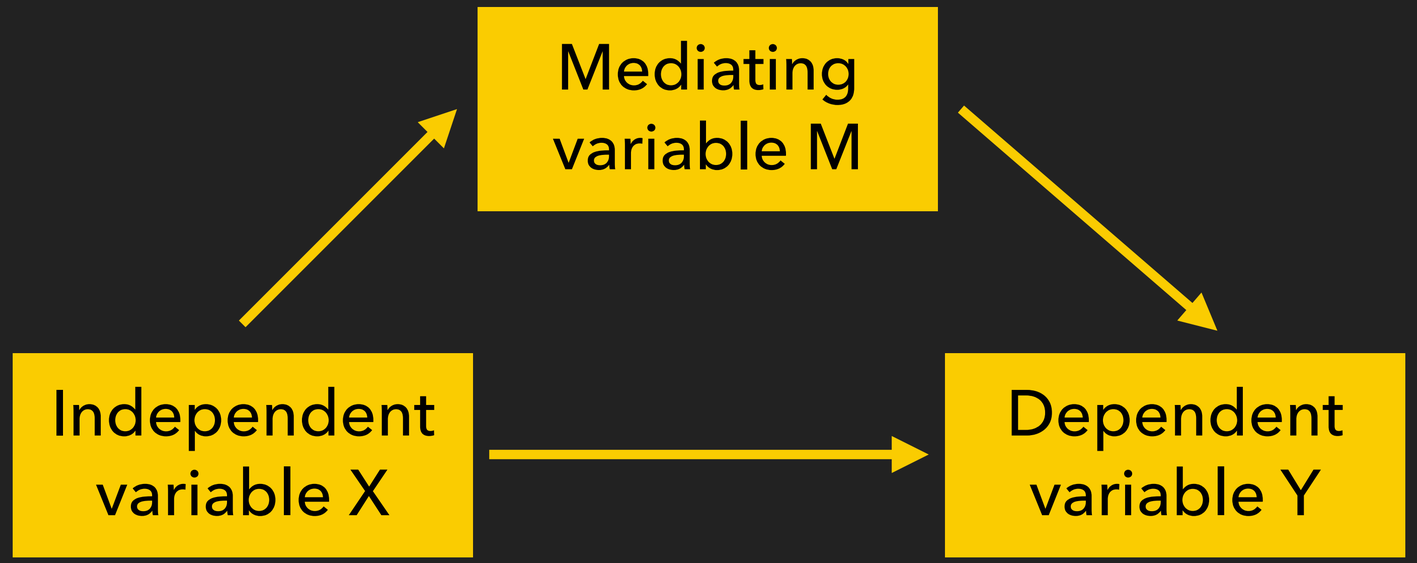
No other scientific method regularly matches the characteristics of causal relationships so well.

Aside: Mediators & Moderators

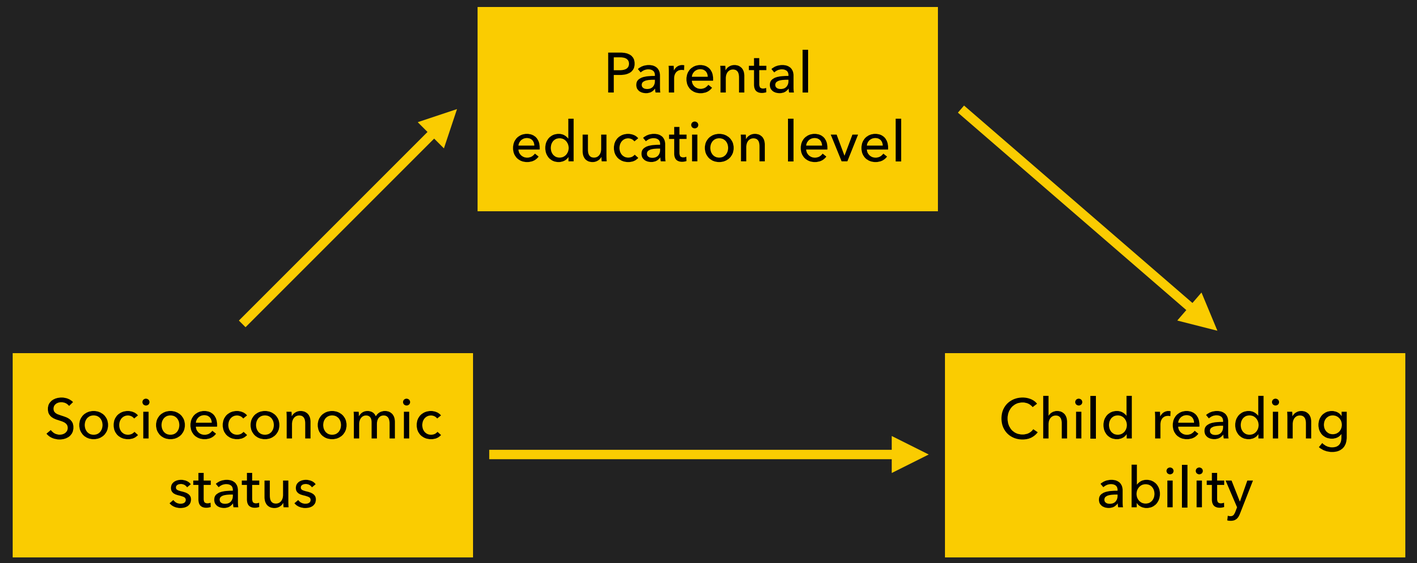
Mediators and Moderators

Mediators and Moderators

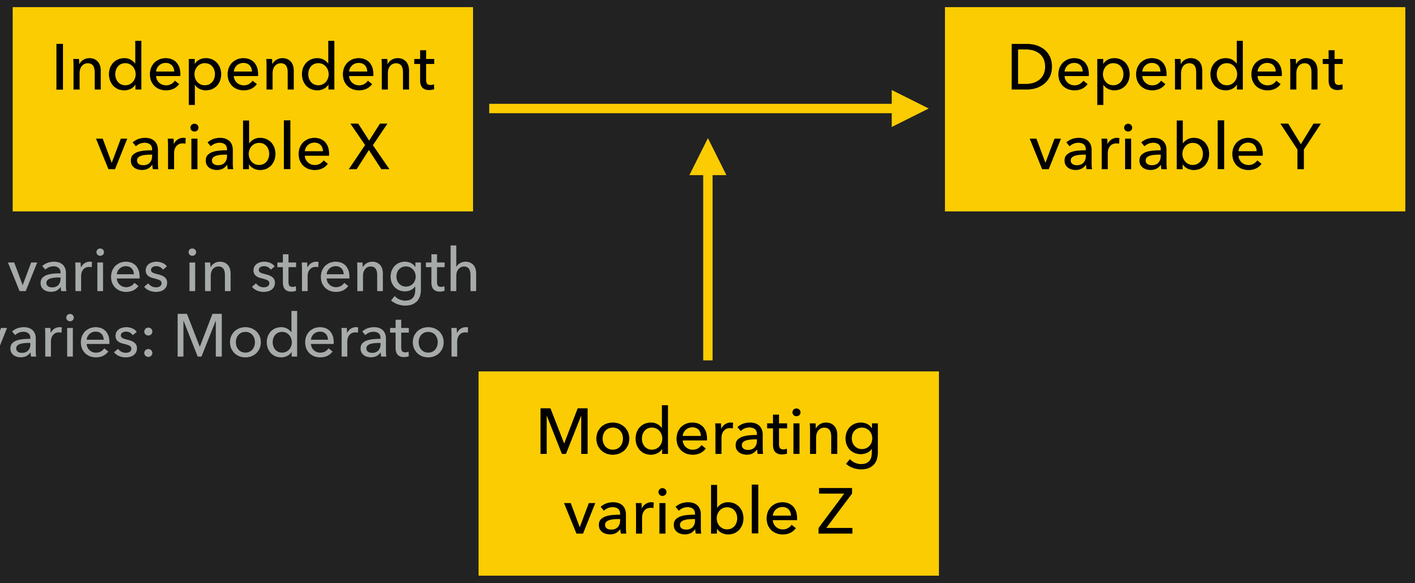
Links in the explanatory chain: Mediator



Mediators and Moderators

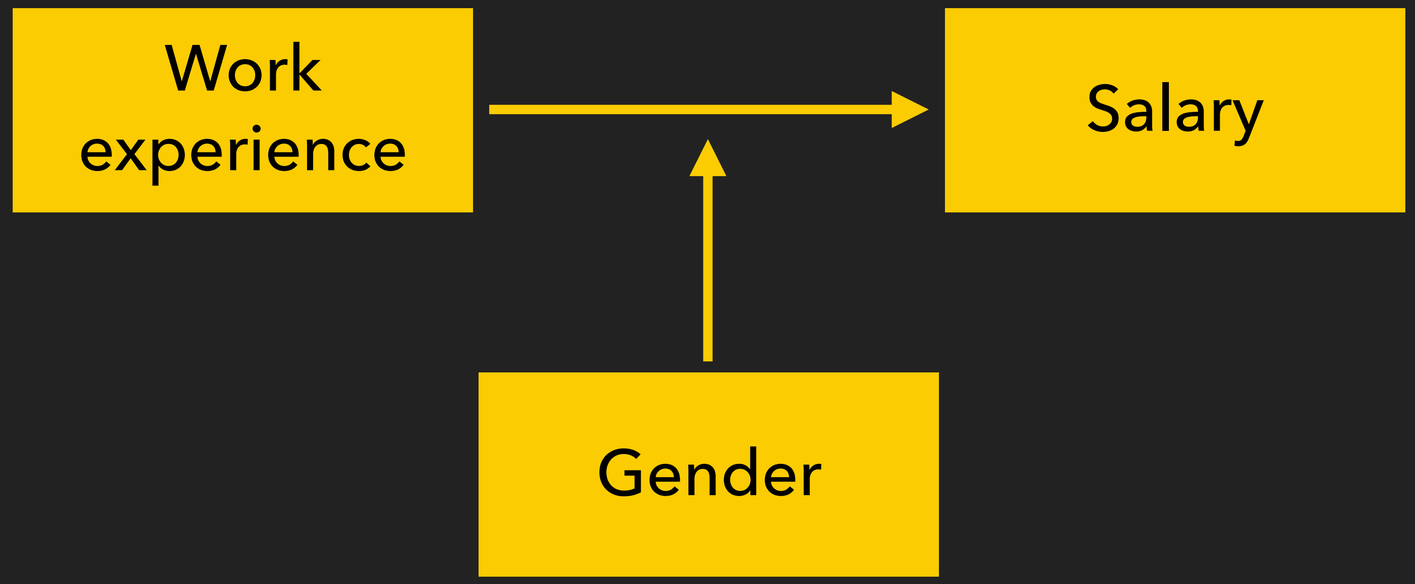


Mediators and Moderators



Causal relationship varies in strength (or direction) as Z varies: Moderator

Mediators and Moderators

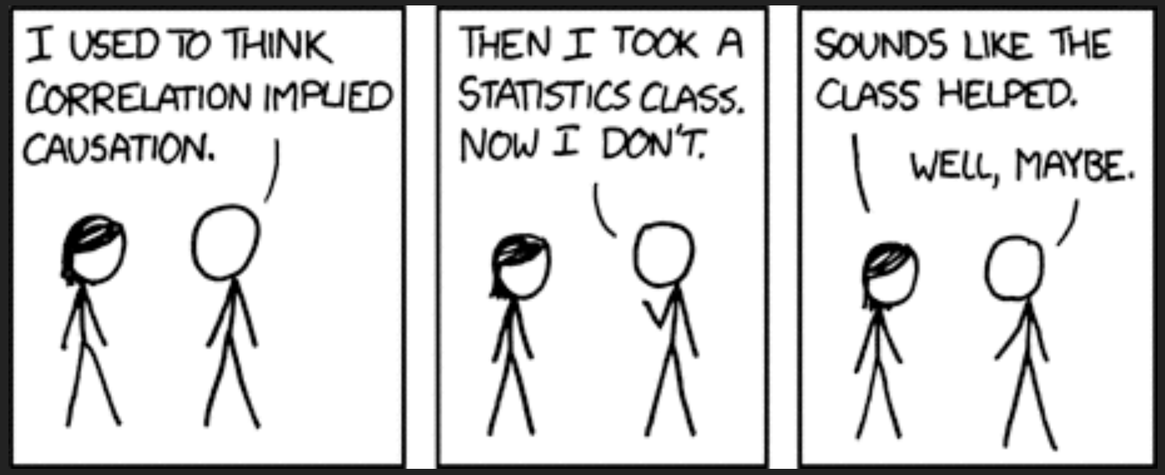


Aside: Correlation is not enough!

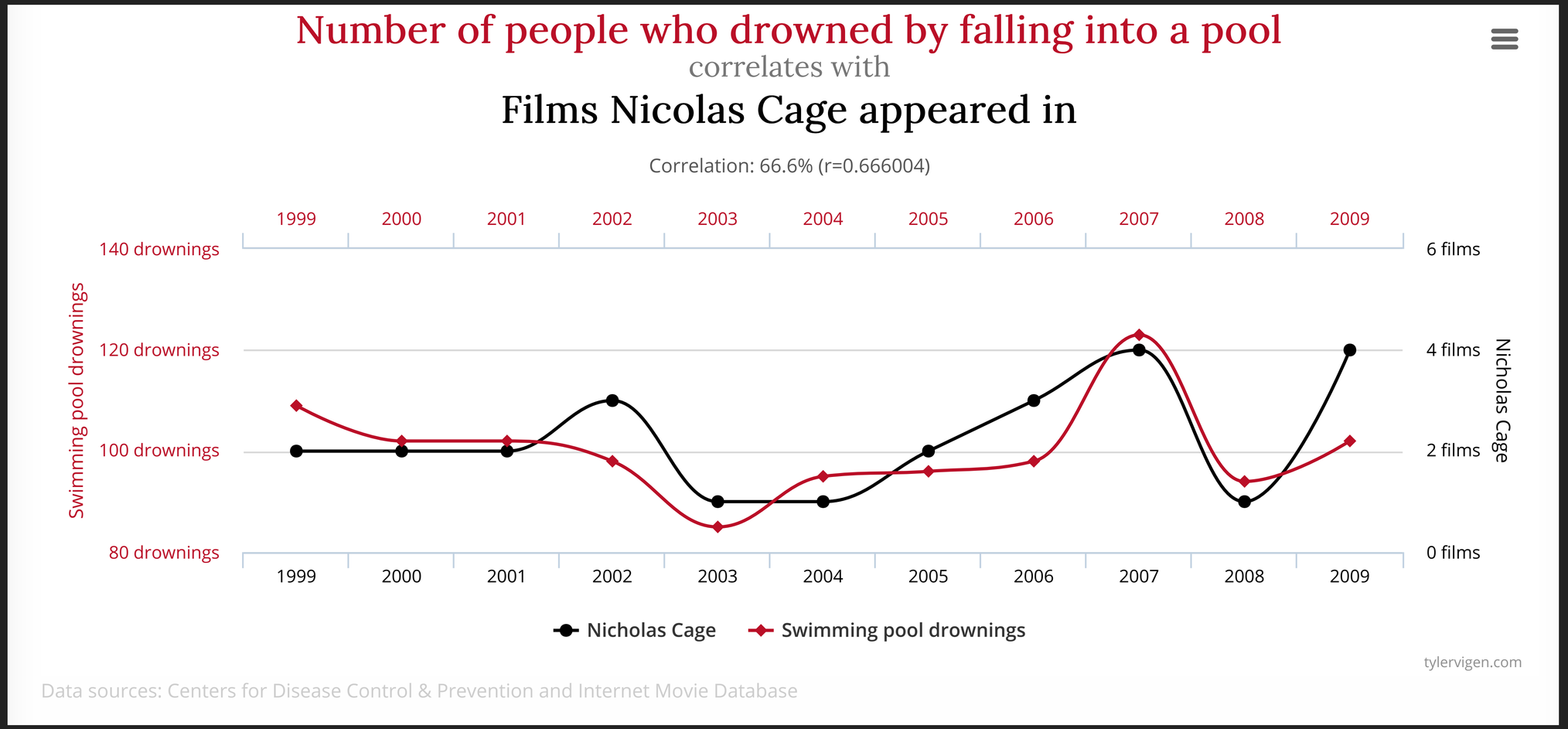
Correlation Does Not Prove Causation

- Which variable came first?
- Are there alternative explanations for the presumed effect?

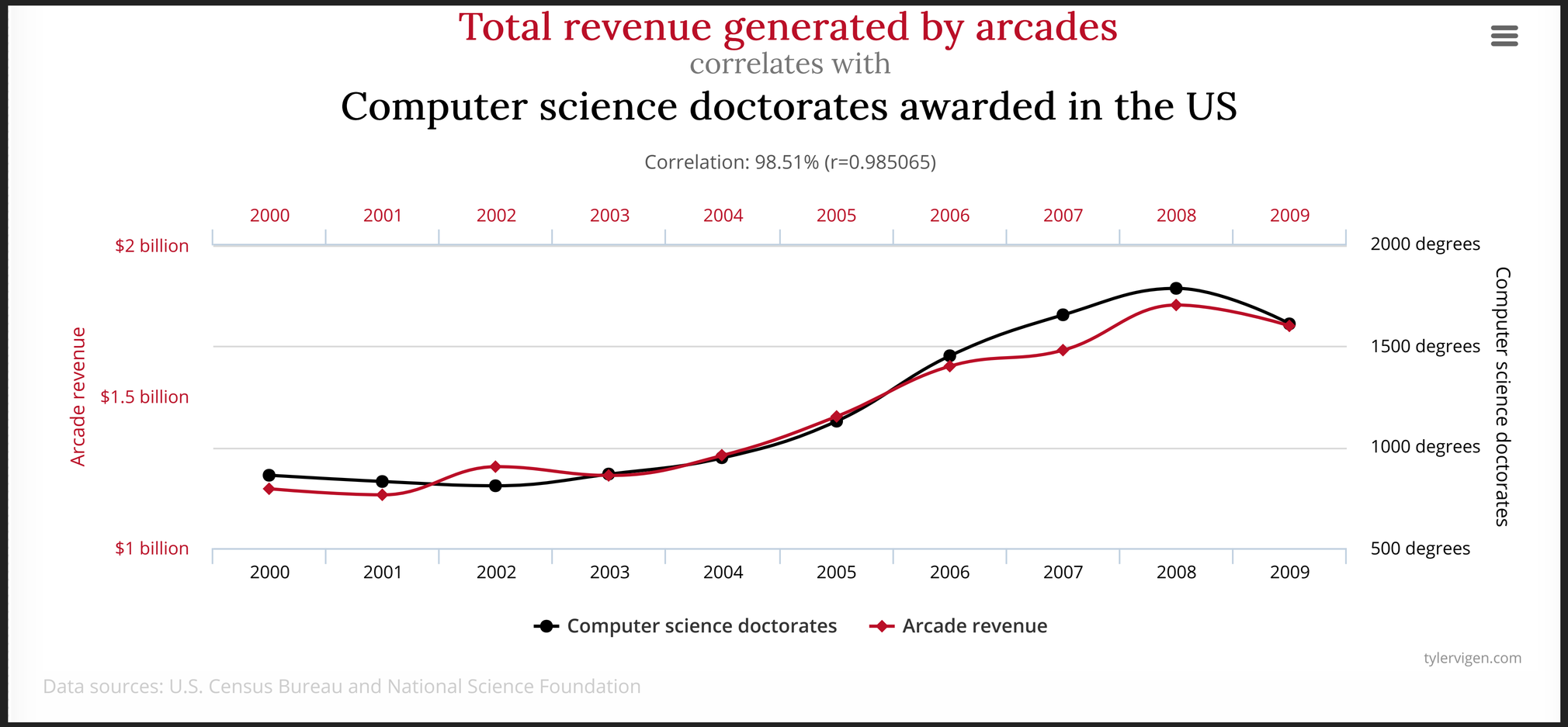
- Example: income ~ education or education ~ income?
  - Confounding variable: intelligence, family socioeconomic status (causes both high education and high income)



http://www.tylervigen.com/spurious-correlations



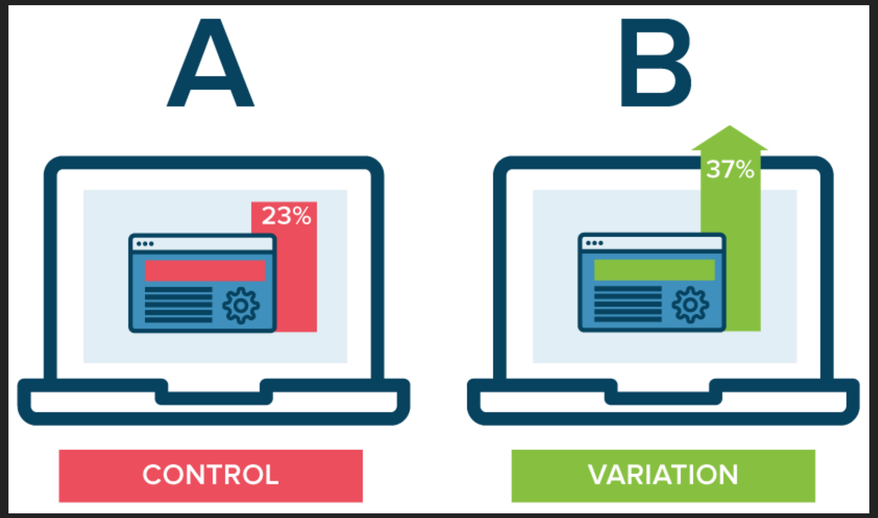
http://www.tylervigen.com/spurious-correlations



Experiments: Summary Pros and Cons

Advantages and Disadvantages of Experiments

- Disadvantages of experiments:
  - Conditions may be unrealistic
  - Tell nothing about how and why effects occurred
  - Cannot deal with cases when we first observe effect and need to look for causes



"O.K., let's slowly lower in the grant money."

Advantages and Disadvantages of Experiments

- Disadvantages of experiments:
  - Conditions may be unrealistic
  - Tell nothing about how and why effects occurred
  - Cannot deal with cases when we first observe effect and need to look for causes

- Unique advantage:
  - Causal description: describe consequences attributable to deliberately varying a treatment
  - (But not causal explanation / mechanisms)

"O.K., let's slowly lower in the grant money."

. . . to be continued

Credits

- Graphics:
  - Dave DiCello photography (cover)
- Content:
  - Chapters from Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Publishing
    - Ch1: Experiments and generalized causal inference
    - Ch2: Statistical conclusion validity and internal validity
    - Ch3: Construct validity and external validity
    - Ch8: Randomized experiments
  - Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.
  - Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (2007). Statistics.