Introduction to Data Science and Analytics (DSC510)

University of Cyprus

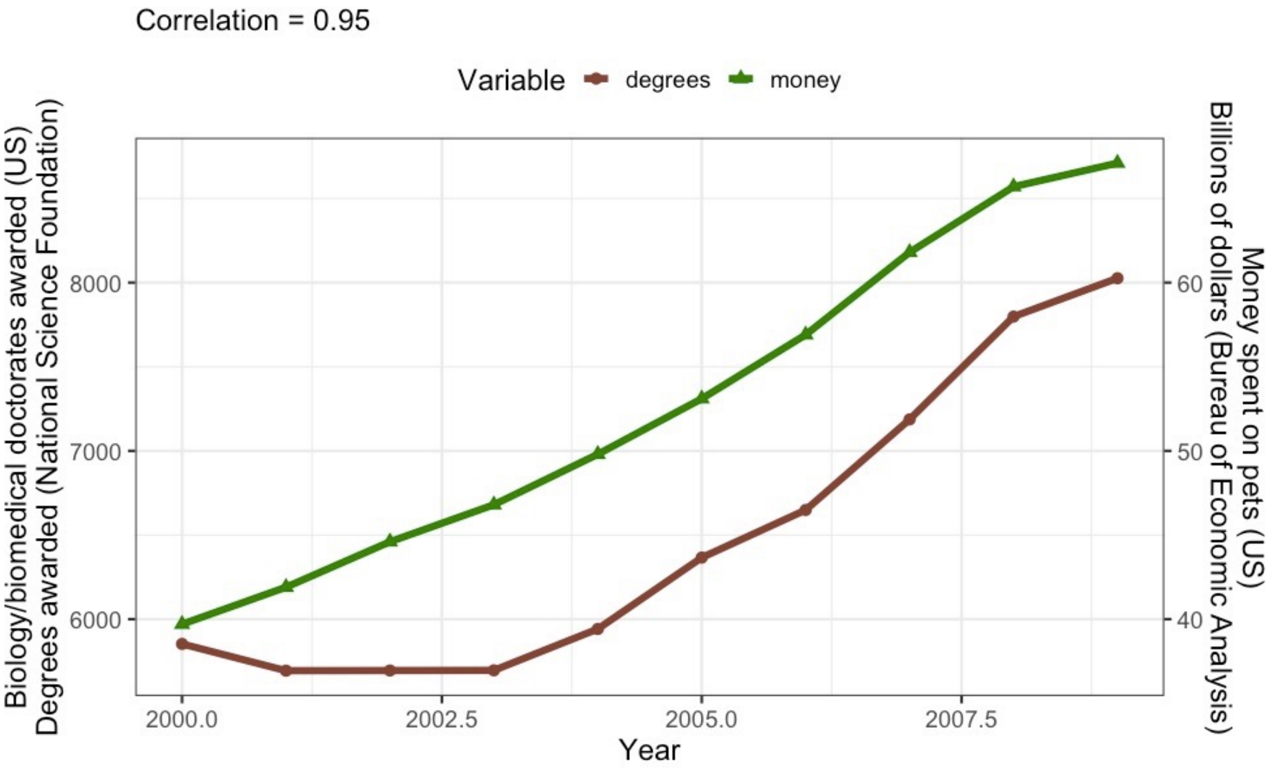
Observational Studies

George Pallis

Goals of this lecture

- Clarify difference between **experimental** and **observational** studies
- Highlight **pitfalls** of observational studies
- Motivate you to **read** Rosenbaum’s great book “Design of Observational Studies” (in particular Chapters 1, 2, and 3; or this book) and Pearl’s eye-opening “Book of Why”

Observation: Can We Learn Anything?

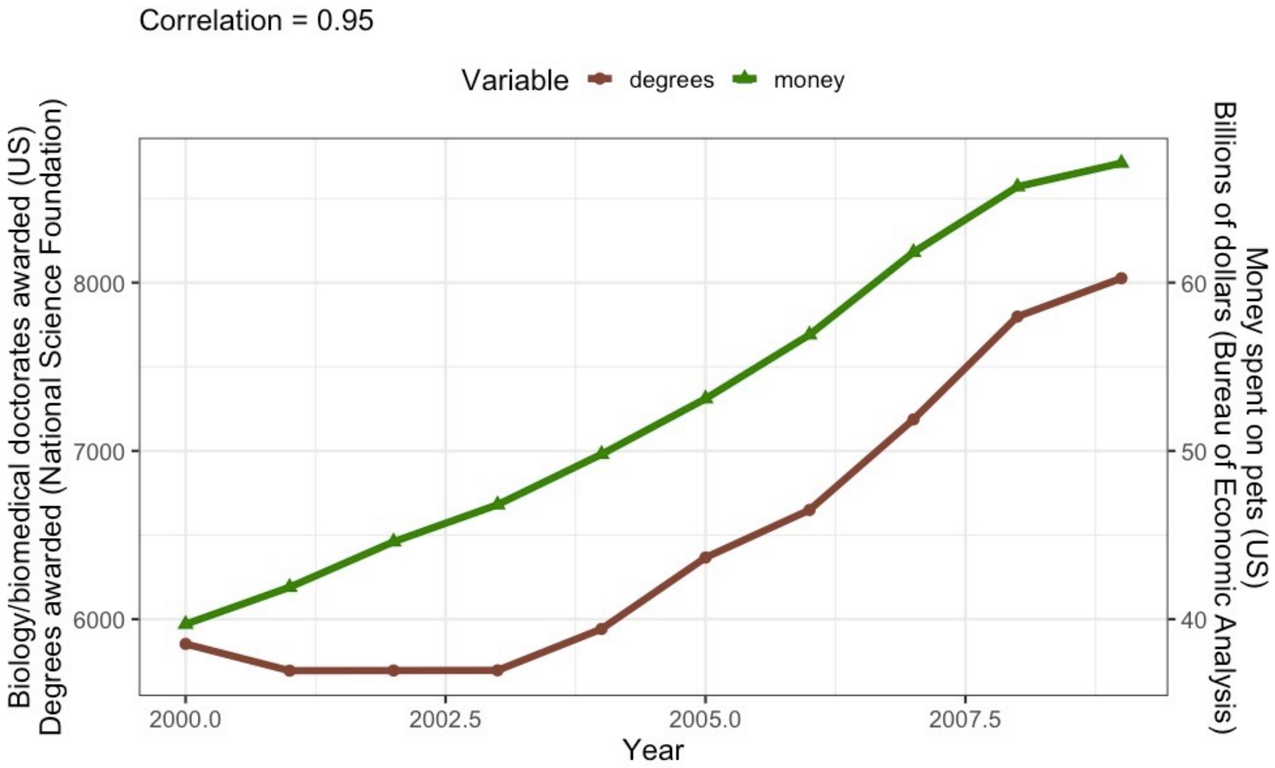


http://tylervigen.com/

Correlation does not equal causation…
but where there’s smoke, there’s fire.

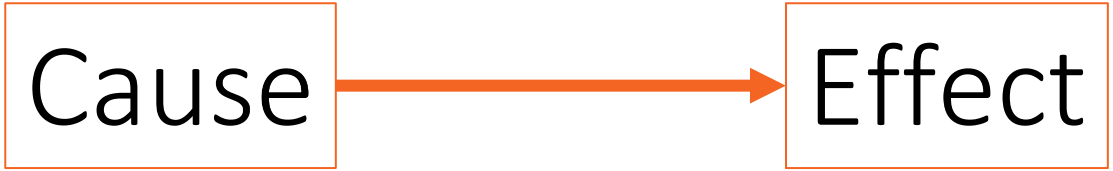
-Jim Grace

Where's the Fire?



http://tylervigen.com/

What We Want to Evaluate



Cause

Effect

What is causal inference?

Inferring the effects of any treatment/policy/intervention/etc.

What is causal inference?

Inferring the effects of any treatment/policy/intervention/etc.

Examples:

- Effect of treatment on a disease

What is causal inference?

Inferring the effects of any treatment/policy/intervention/etc.

Examples:

- Effect of treatment on a disease
- Effect of climate change policy on emissions

What is causal inference?

Inferring the effects of any treatment/policy/intervention/etc.

Examples:

- Effect of treatment on a disease
- Effect of climate change policy on emissions
- Effect of social media on mental health

What is causal inference?

Inferring the effects of any treatment/policy/intervention/etc.

Examples:

- Effect of treatment on a disease
- Effect of climate change policy on emissions
- Effect of social media on mental health
- Many more (effect of X on Y)

Simpson’s paradox: COVID-27

New disease: COVID-27

Simpson’s paradox: COVID-27

New disease: COVID-27

Treatment T: A (0) and B (1)

Simpson's paradox: COVID-27

New disease: COVID-27

Treatment T: A (0) and B (1)

Simpson's paradox: COVID-27

New disease: COVID-27

Treatment T: A (0) and B (1)

Condition C: mild (0) or severe (1)

Simpson's paradox: COVID-27

New disease: COVID-27

Treatment T: A (0) and B (1)

Condition C: mild (0) or severe (1)

Outcome Y: alive (0) or dead (1)

Simpson’s paradox: mortality rate table

Treatment

	Total
A	**16%** (240/1500)
B	19% (105/550)

$\mathbb{E}[Y \mid T]$

Simpson’s paradox: mortality rate table

		Condition		
		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

$\mathrm{E}[Y \mid T, C = 0]$

Simpson's paradox: mortality rate table

Condition

Treatment

	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

$\mathrm{E}[Y \mid T, C = 0]$

$$\frac{1400}{1500}(0.15) + \frac{100}{1500}(0.30) = 0.16$$

$$\frac{50}{550}(0.10) + \frac{500}{550}(0.20) = 0.19$$

Simpson’s paradox: mortality rate table

Condition

		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	16% (240/1500)
	B	10% (5/50)	20% (100/500)	19% (105/550)

$\mathbb{E}[Y \mid T, C = 0]$

$$\boxed{\frac{1400}{1500}}(0.15) + \boxed{\frac{100}{1500}}(0.30) = 0.16$$

$$\frac{50}{550}(0.10) + \frac{500}{550}(0.20) = 0.19$$

Simpson’s paradox: mortality rate table

Condition

Treatment

	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	16% (240/1500)
B	10% (5/50)	20% (100/500)	19% (105/550)

$$\boxed{\frac{1400}{1500}}(0.15) + \boxed{\frac{100}{1500}}(0.30) = 0.16$$

$$\boxed{\frac{50}{550}}(0.10) + \boxed{\frac{500}{550}}(0.20) = 0.19$$

$\mathrm{E}[Y \mid T, C \; = 0]$

Simpson’s paradox: mortality rate table

Condition

Treatment

	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

$\mathrm{E}[Y \mid T, C = 0]$

$$\boxed{\frac{1400}{1500}}(0.15) + \boxed{\frac{100}{1500}}(0.30) = 0.16$$

$$\boxed{\frac{50}{550}}(0.10) + \boxed{\frac{500}{550}}(0.20) = 0.19$$

Which treatment should you choose?

Simpson’s paradox: scenario 1 (treatment B)

Condition

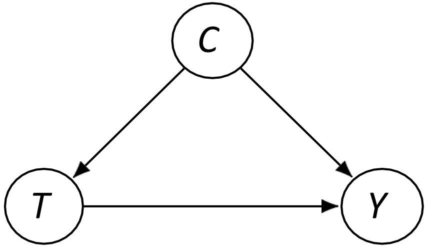
Treatment

	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

Simpson’s paradox: scenario 1 (treatment B)

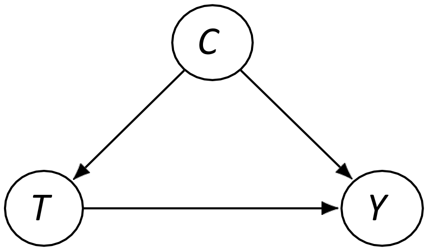
Condition

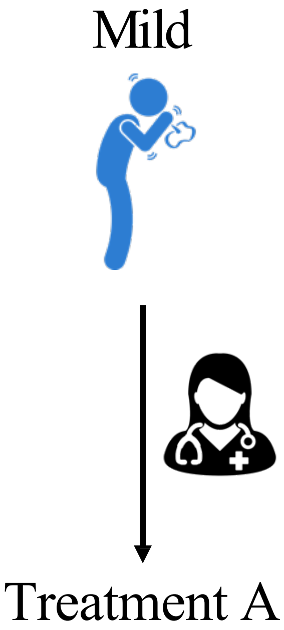
Treatment	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
B	**10%** (5/50)	**20%** (100/500)	19% (105/550)



Simpson’s paradox: scenario 1 (treatment B)

		Condition		
		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)

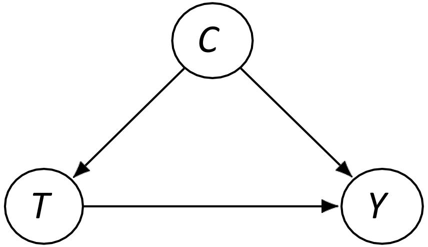


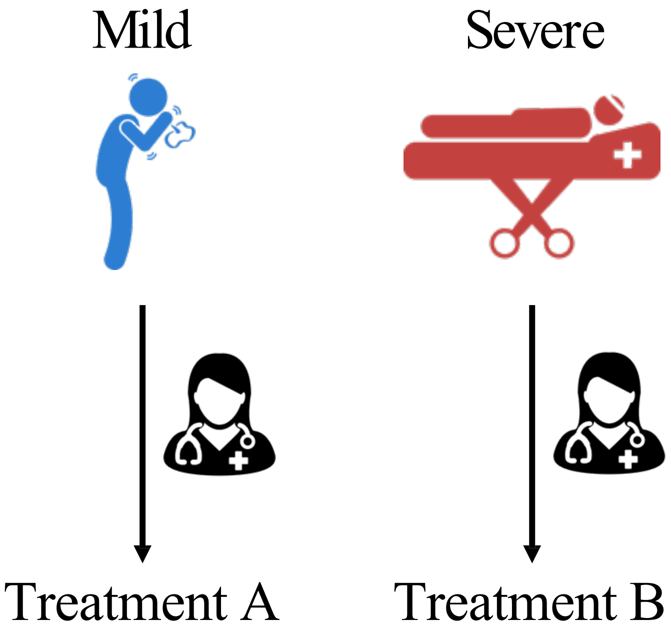


Simpson’s paradox: scenario 1 (treatment B)

Condition

		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	**16%** (240/1500)
	B	**10%** (5/50)	**20%** (100/500)	19% (105/550)



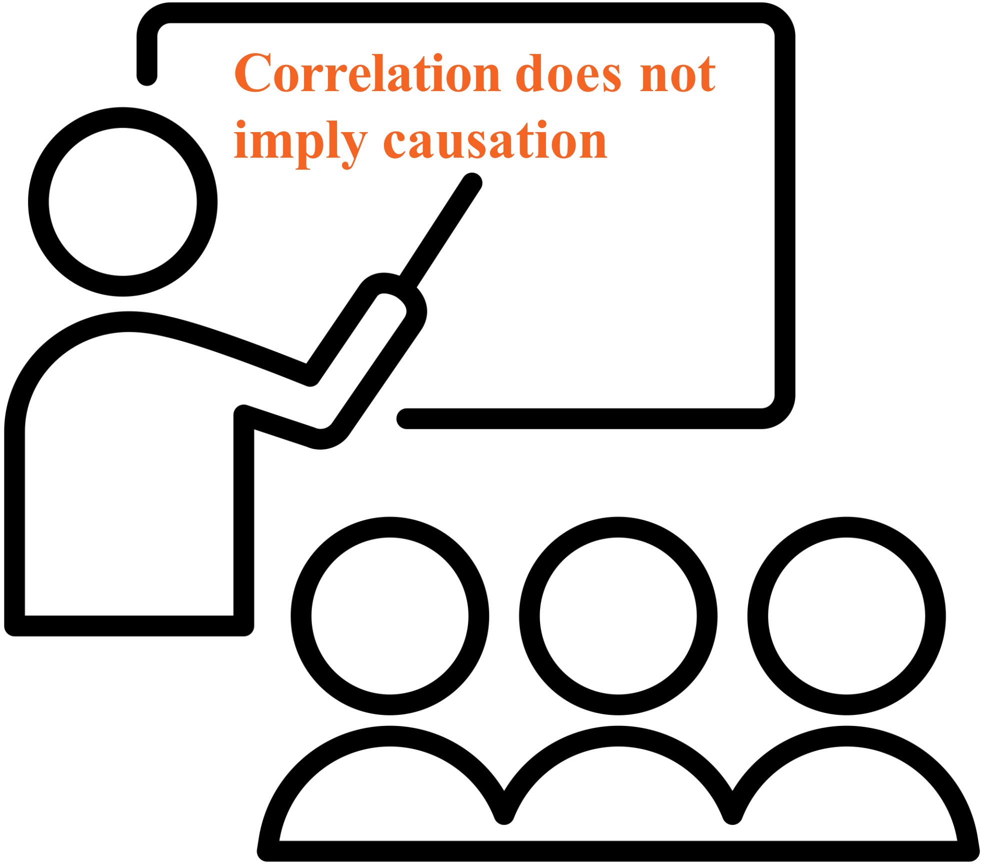


Mild

Severe

Treatment A

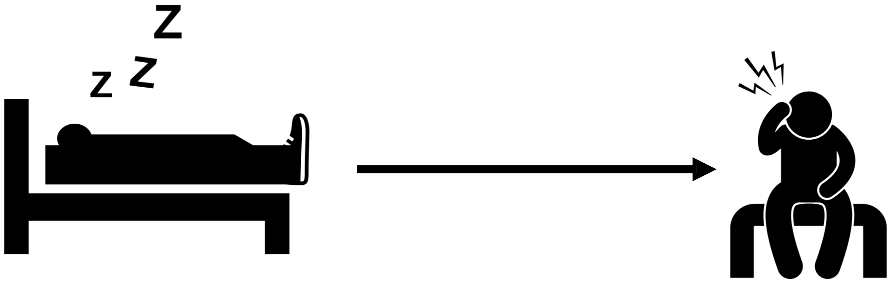
Treatment B



Correlation does not imply causation

Correlation does not imply causation

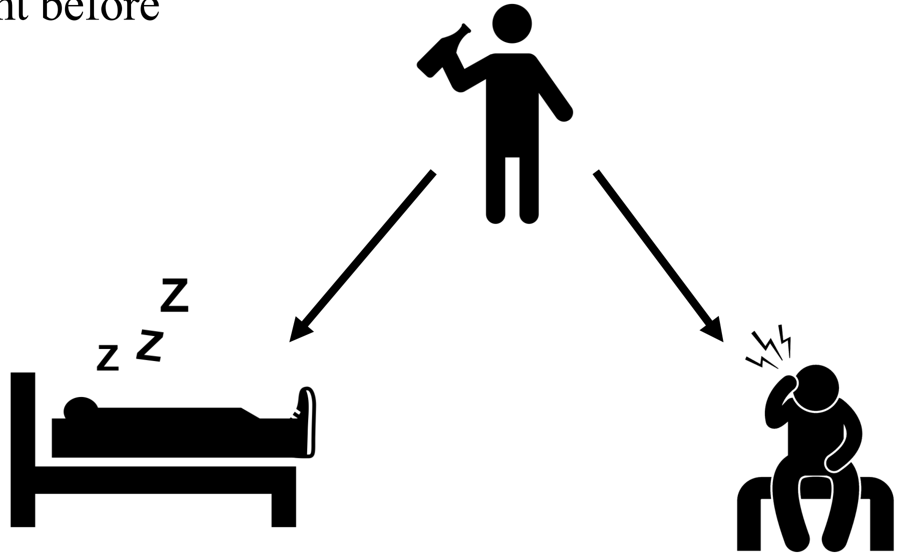
Sleeping with shoes on is strongly correlated with waking up with a headache



Correlation does not imply causation

Sleeping with shoes on is strongly correlated with waking up with a headache

Common cause: drinking the night before

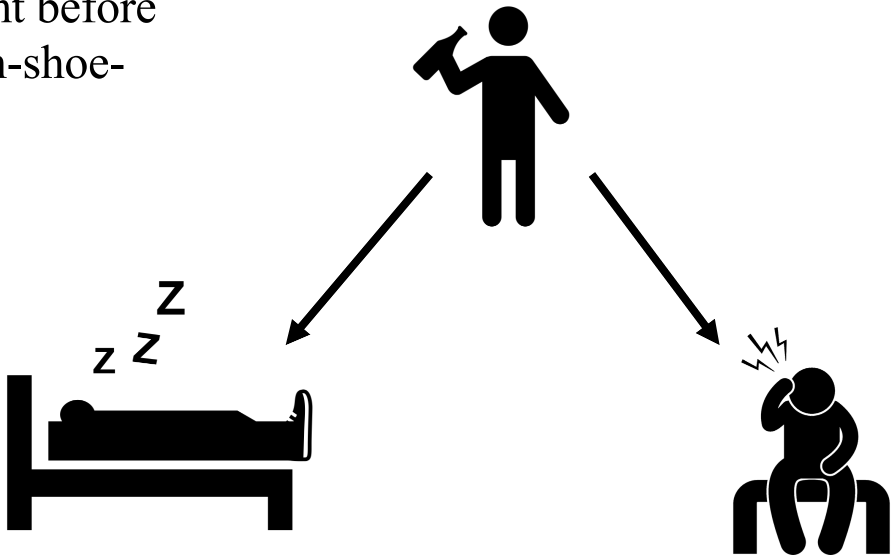


Correlation does not imply causation

Sleeping with shoes on is strongly correlated with waking up with a headache

Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way

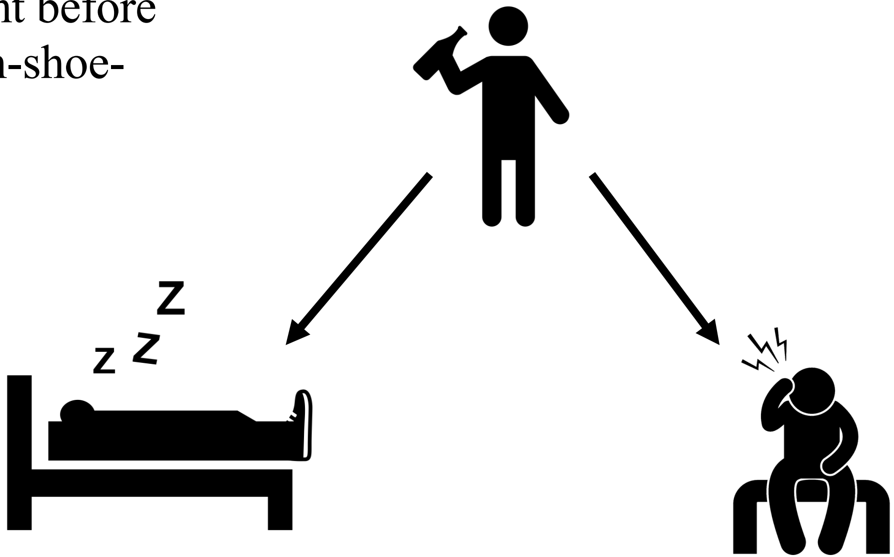


Correlation does not imply causation

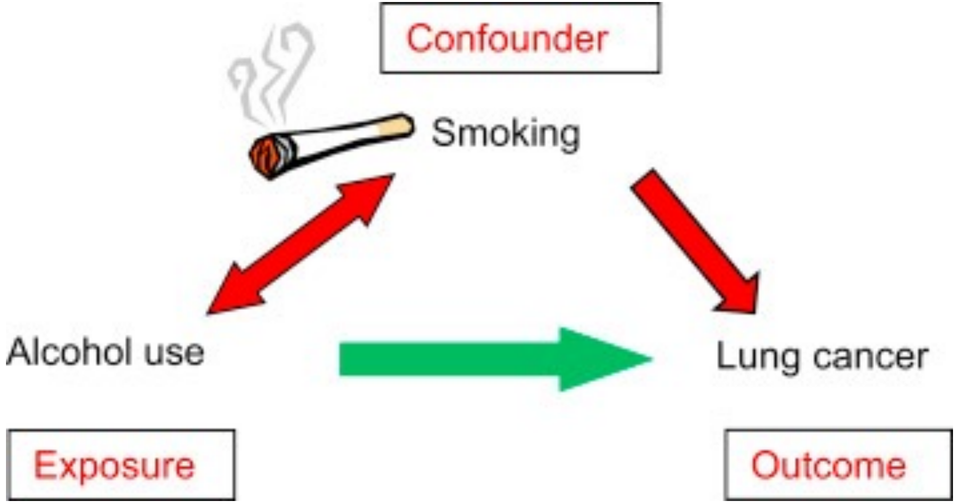
Sleeping with shoes on is strongly correlated with waking up with a headache

Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way
2. Confounding



A confounder is a variable that is associated or has a relationship with both the exposure and the outcome of interest

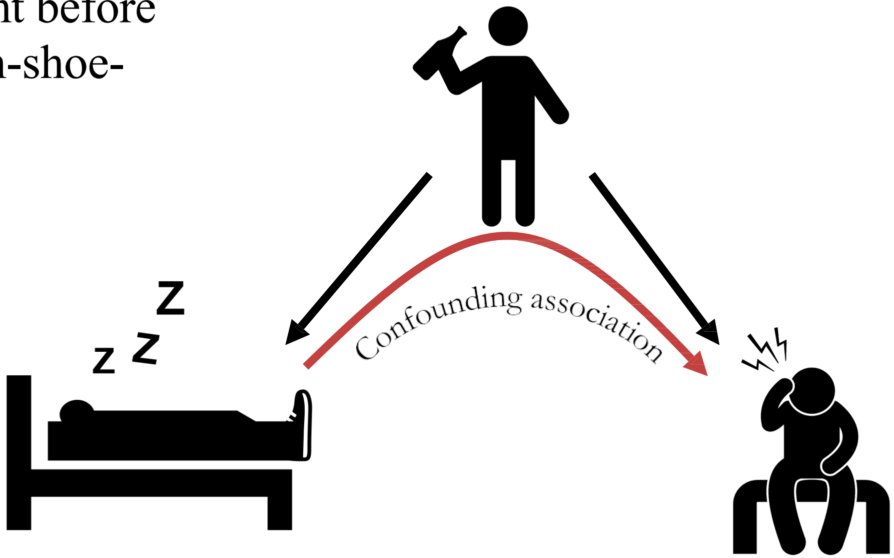


Correlation does not imply causation

Sleeping with shoes on is strongly correlated with waking up with a headache

Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way
2. Confounding

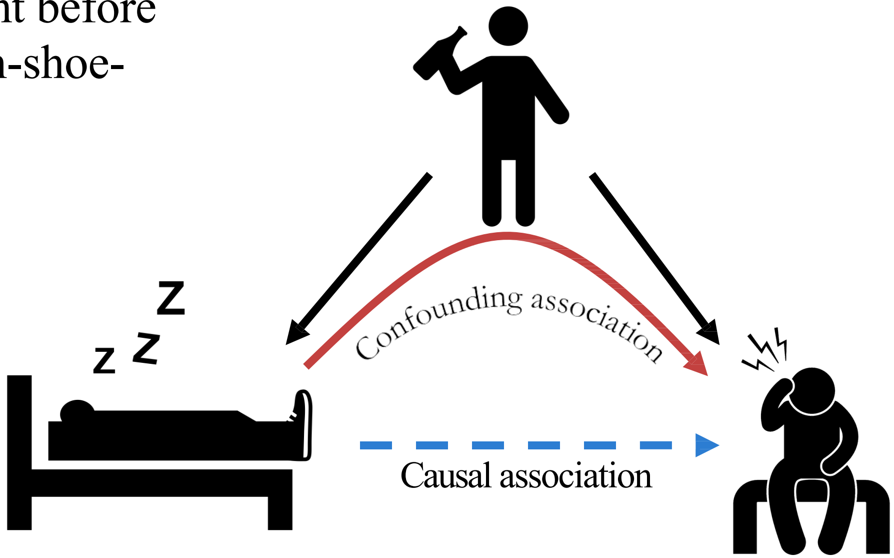


Correlation does not imply causation

Sleeping with shoes on is strongly correlated with waking up with a headache

Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way
2. Confounding



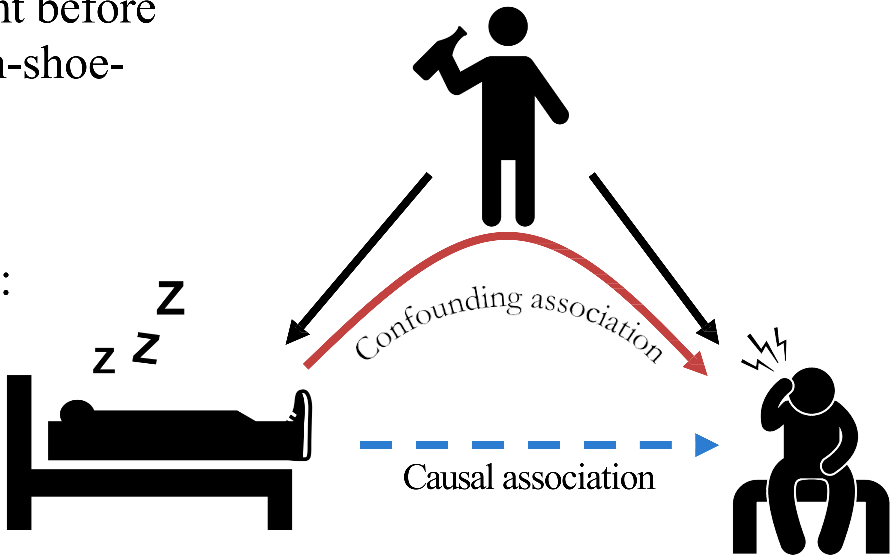
Correlation does not imply causation

Sleeping with shoes on is strongly correlated with waking up with a headache

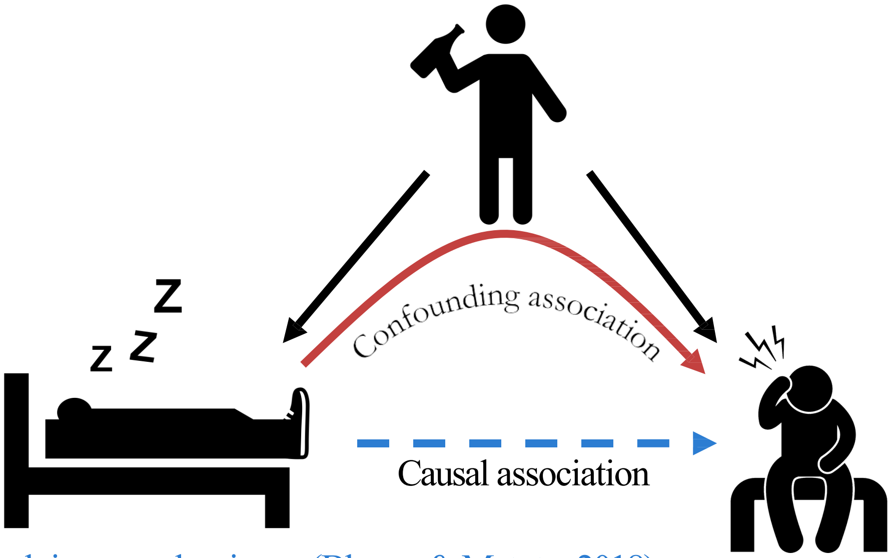
Common cause: drinking the night before

1. Shoe-sleepers differ from non-shoe-sleepers in a key way
2. Confounding

Total association (e.g. correlation): mixture of causal and confounding association



“Correlation = Causation” is a cognitive bias[1]



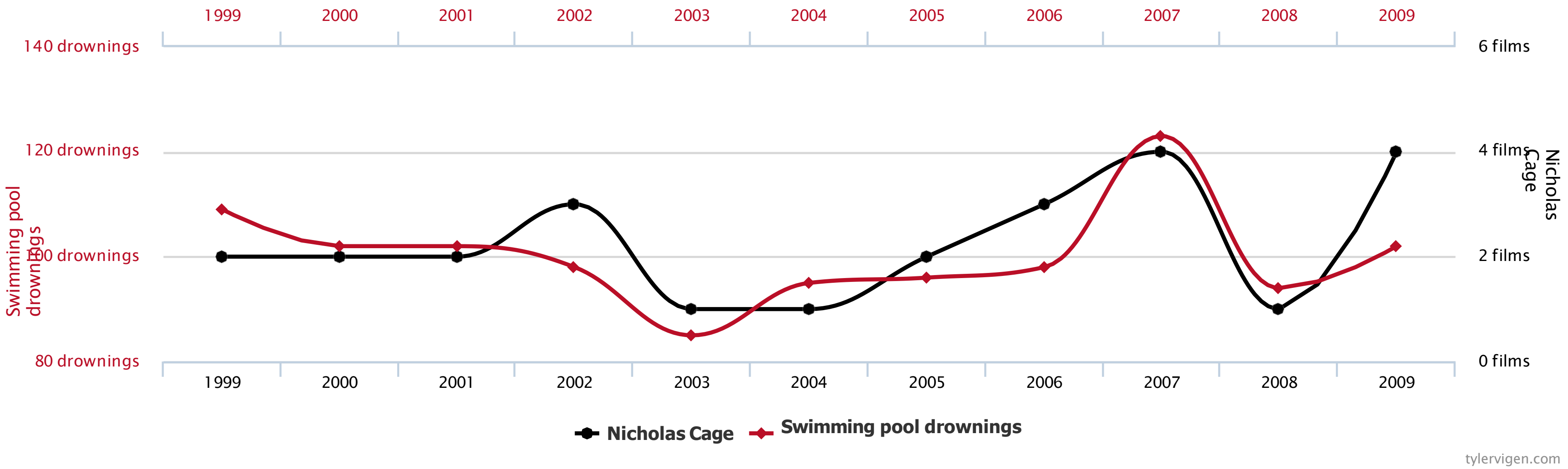
[1] The illusion of causality: A cognitive bias underlying pseudoscience (Blanco & Matute, 2018)

Nicolas Cage drives people to drown themselves

Number of people who drowned by falling into a pool

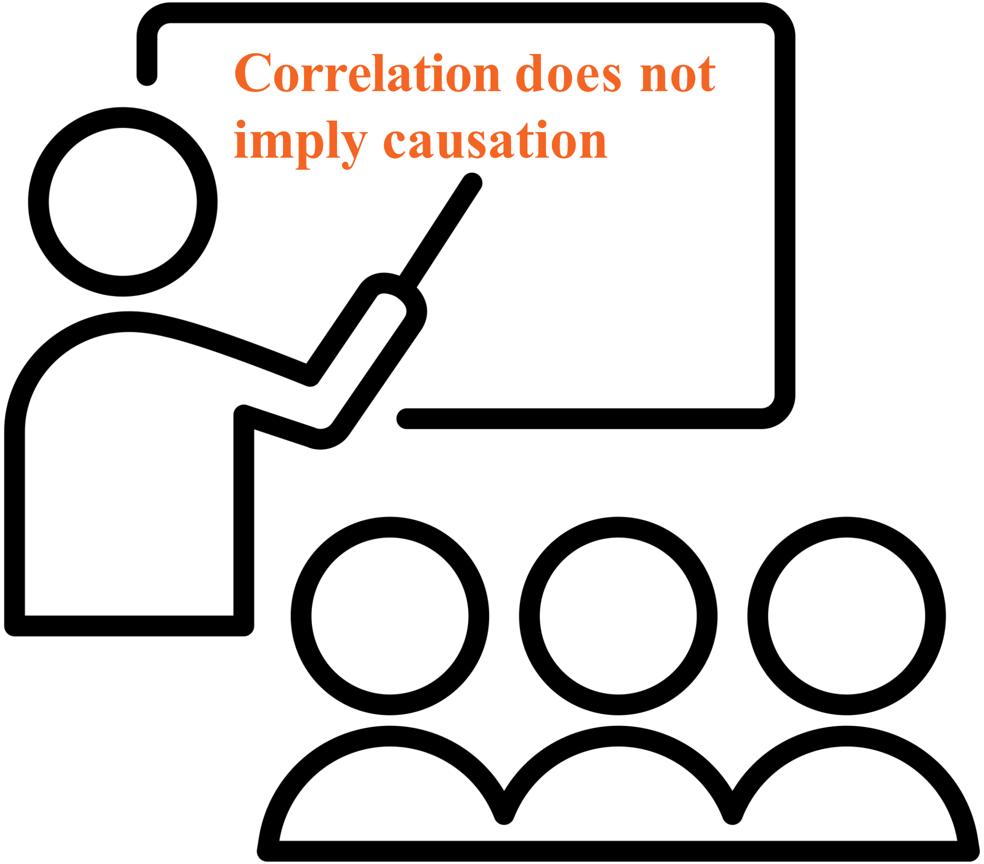
correlates with

Films Nicolas Cage appeared in



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

Correlation does not imply causation



Correlation does not imply causation

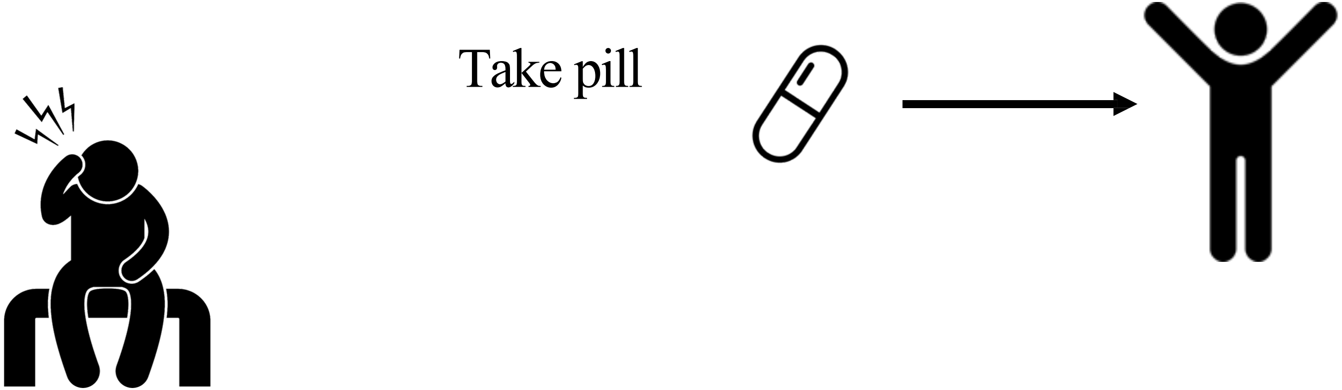
Then, what does imply causation?

Potential outcomes: intuition

Inferring the effect of treatment/policy on some outcome

Potential outcomes: intuition

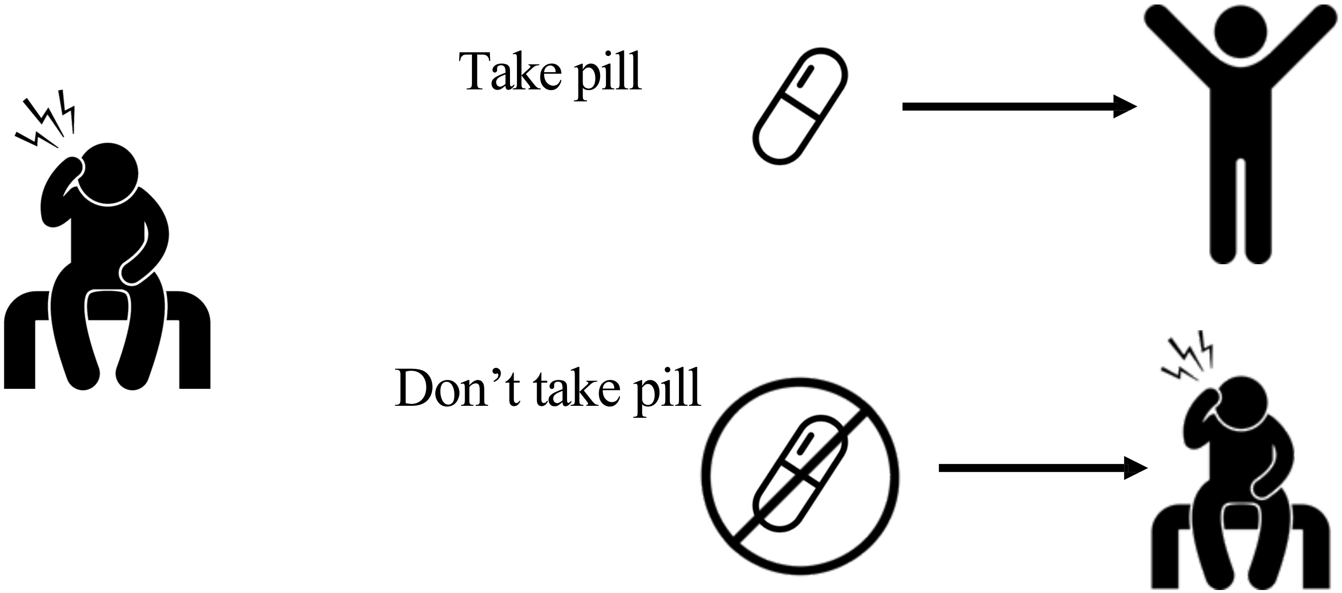
Inferring the effect of treatment/policy on some outcome



Take pill

Potential outcomes: intuition

Inferring the effect of treatment/policy on some outcome

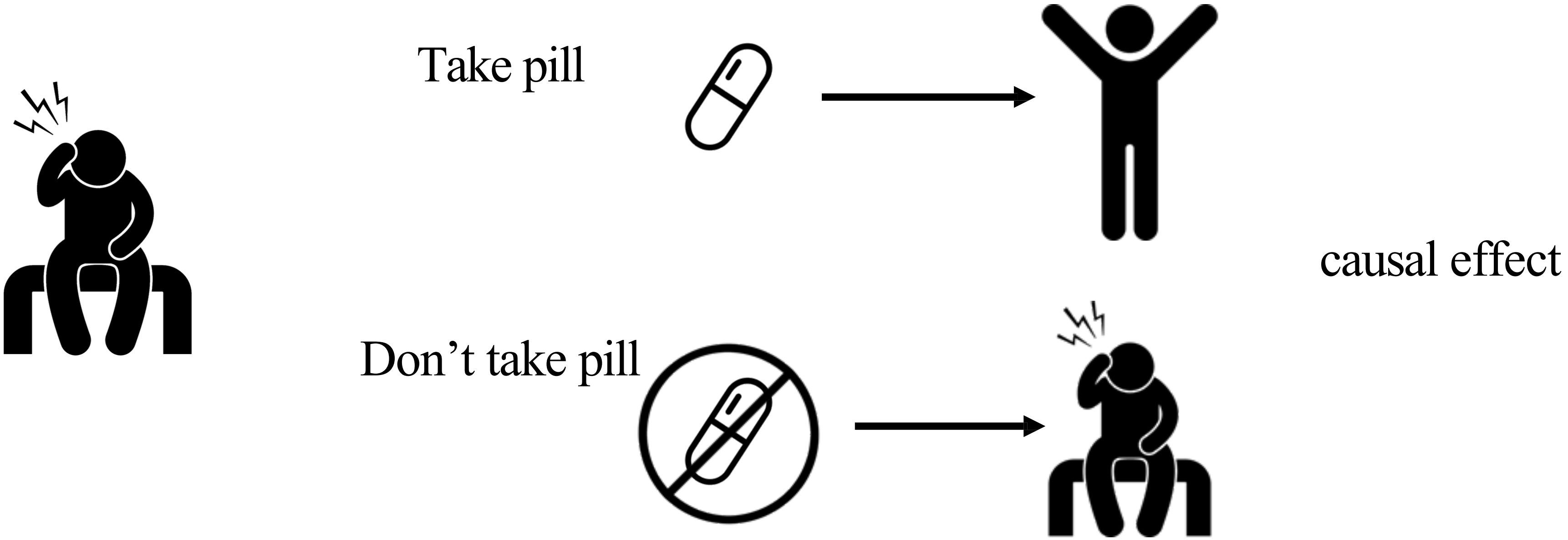


Take pill

Don’t take pill

Potential outcomes: intuition

Inferring the effect of treatment/policy on some outcome



Take pill

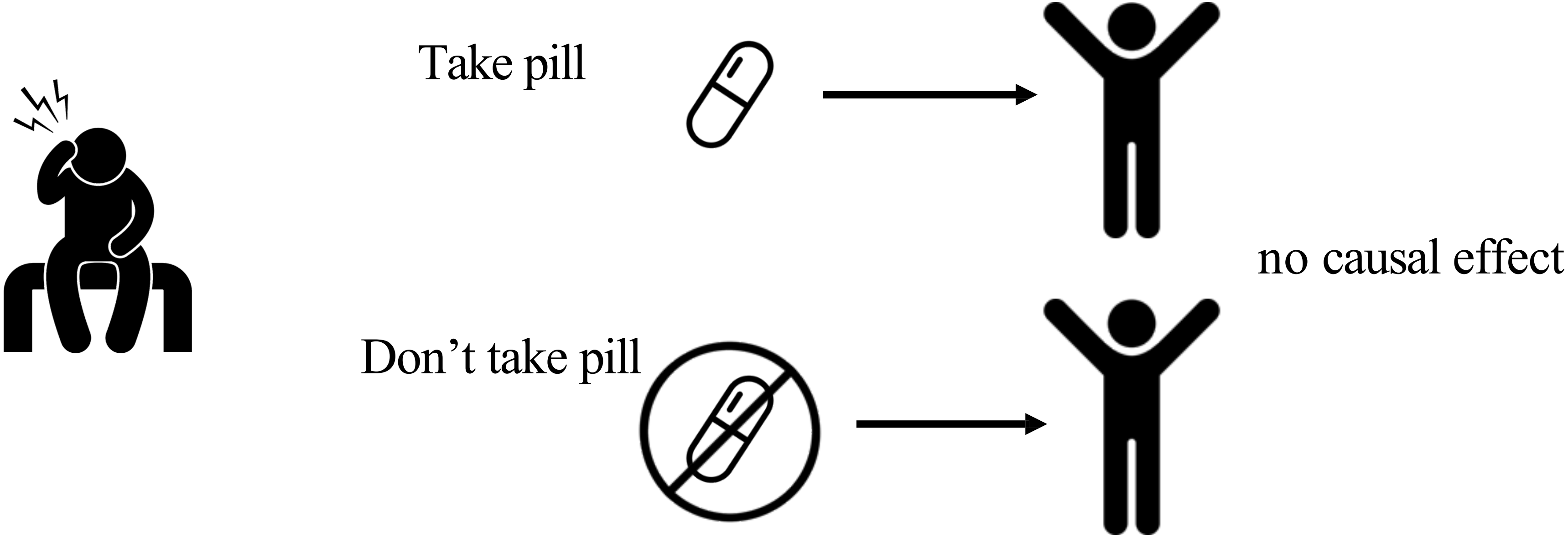
Don’t take pill

causal effect

Then, what does imply causation?

Potential outcomes: intuition

Inferring the effect of treatment/policy on some outcome



Take pill

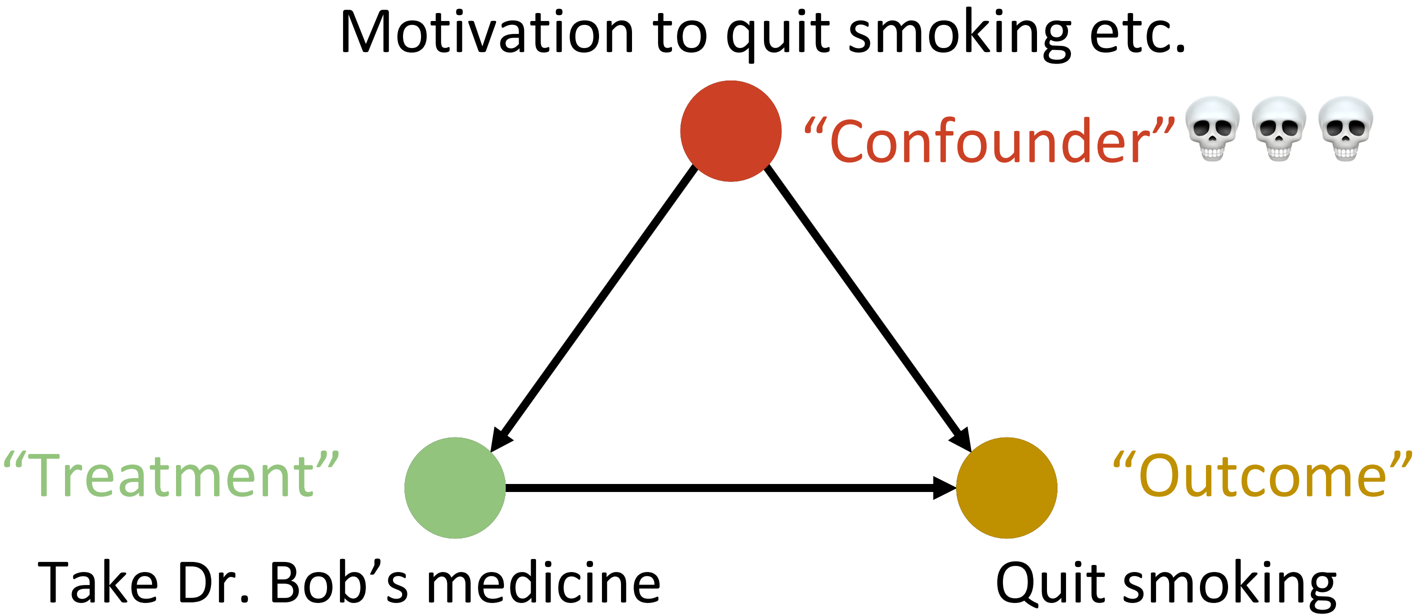
Don't take pill

no causal effect

Dr. Bob’s smoking cure

- Dr. Bob claims to have developed a medicine that helps you quit smoking
- He asks all smokers: “Do you want to try my medicine?”
- Smokers = treated smokers ∪ untreated (“control”) smokers
- Fraction of successful quitters is higher in the treated group
- He conclude: “My medicine helps you quit smoking! Buy it!”
- Do you believe him?

Dr. Bob's "experiment" as a causal diagram



Motivation to quit smoking etc.

"Confounder"💀💀💀

"Treatment"

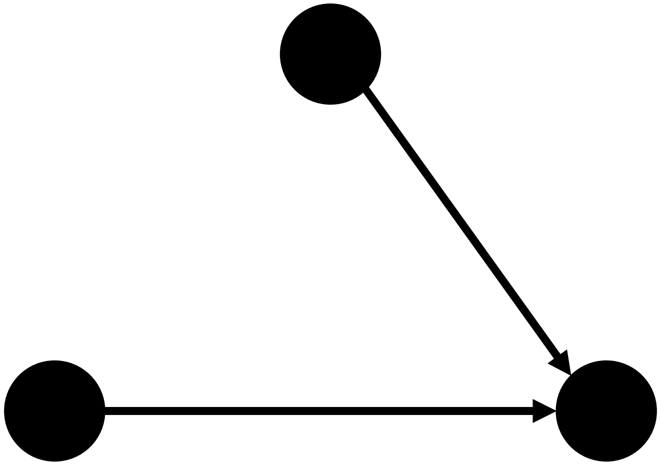
Take Dr. Bob's medicine

"Outcome"

Quit smoking

Ideal setting as a causal diagram

Motivation to quit smoking etc.



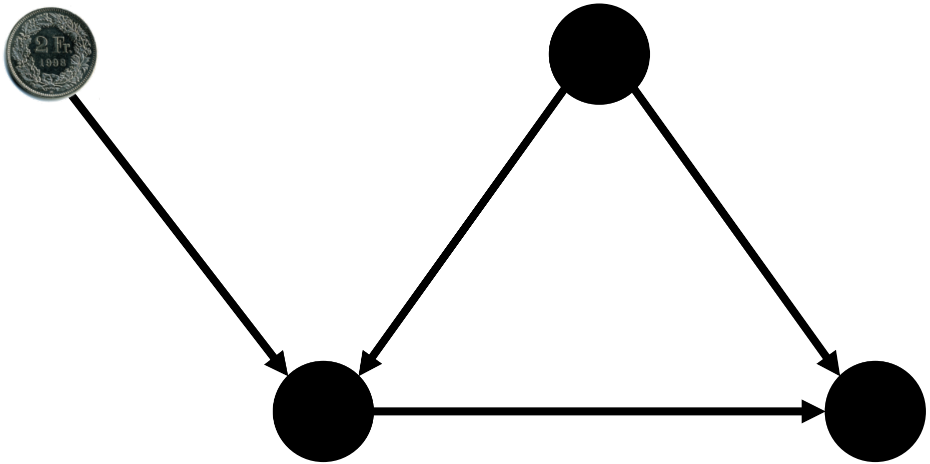
Take Dr. Bob’s medicine

Quit smoking

Randomized controlled experiments

- Two experimental conditions:
  - Treatment (e.g., medicine)
  - Control (e.g., placebo [fun fact])
- Assignment of participants to conditions is random
  - Probability of receiving treatment same for everyone
- Treatment and control groups are indistinguishable
  - E.g., determination to quit smoking is not systematically higher in the treated group

Randomized controlled experiment as a causal diagram



Coin

Motivation to quit smoking etc.

Take Dr. Bob’s medicine

Quit smoking

Limits of randomization

- Do seat belts save lives?
- Experiment:
  - Flip coin at birth to assign to treatment (always wear seat belt for entire life) or control (never wear seat belt)
  - Measure fraction of traffic deaths in each group
- Randomized experiments aren’t always feasible
  - Unethical (see above), expensive, fundamentally impossible (e.g., do earthquakes decrease life spans?)
  - Most modern “big data” is “found data”
- Sometimes, observational studies are even better suited

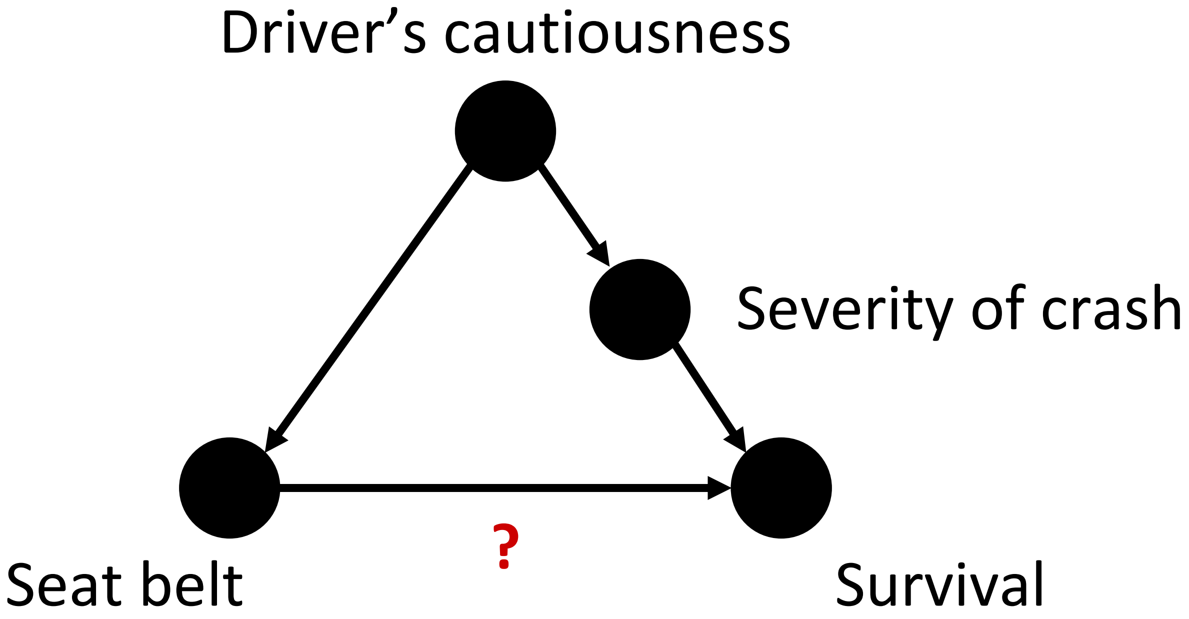
Alternative: observational studies

- Fundamentally different from experiment:
  - Researcher can’t control who goes to which condition
  - Researcher is merely an observer, not a tinkerer
  - Much less problematic w.r.t. ethics, price, feasibility
  - Much more problematic w.r.t. validity of conclusions
- All advantages of randomized experiment are gone
  - Subjects self-select to be treated
  - Treatment assignment and response may be caused by same hidden correlate (a.k.a. confounders; e.g., resolve to quit smoking)

Example: seat belts revisited

- Recall: experiment infeasible because unethical
- Observational study:
    - Dataset: all traffic accidents in a given time span
    - Two treatment conditions:
        - Treated: seat-belt wearers
        - Control: non-seat-belt wearers
    - Compare fraction dead in treated vs. control
- What problems do you see?

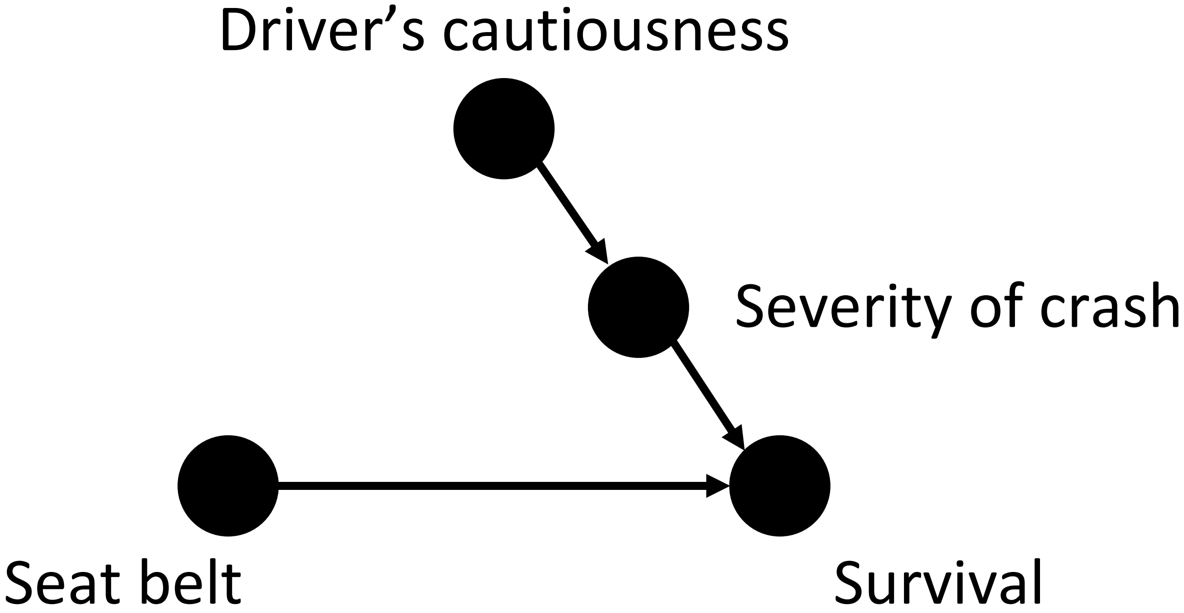
As a causal diagram



A matched observational study

- Consider only particular subset of accident cars:
    - 2 people in car: driver + passenger
    - Exactly one of them died in accident
    - Exactly one of them wearing seat belt at time of accident (i.e., 1 treated + 1 control per car)
- As before: compare fraction dead in treated vs. control
- New: everything else is controlled for, incl. type of car, speed, severity of accident
- Fundamental concept: **matching**

As a causal diagram



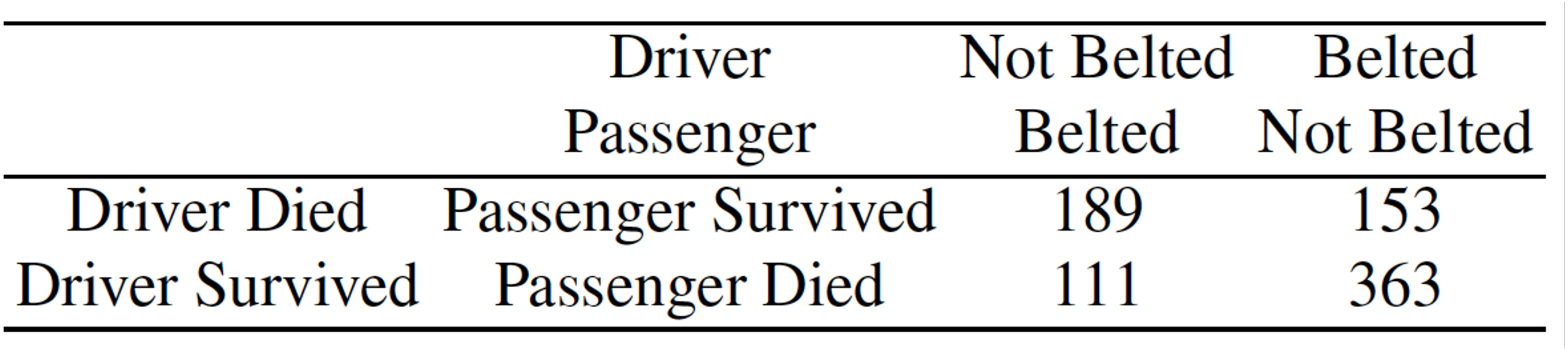
Driver’s cautiousness

Severity of crash

Seat belt

Survival

Settling the seat-belt question



	Driver	Not Belted	Belted
	Passenger	Belted	Not Belted
Driver Died	Passenger Survived	189	153
Driver Survived	Passenger Died	111	363

Nature didn’t flip a coin for me – should I just go home and weep?

- No! You can still get good mileage if you’re smart about it
- Fundamental concept: matching
- Idea: Pair up 2 “similar” people: 1 treated, 1 control
- Ideal (rather: Utopian): “similar” := “identical”
- Also sufficient (phew!): “similar” := equal probability to receive treatment (given the state of the world before the study)

Time for some notation

$$\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$$

ℓ : a subject participating in the study

π_ℓ : probability of being treated, *given full knowledge of the world*

Z_ℓ : treatment assignment (1 := treated; 0 := control)

$r_{T\ell}$: response if subject is treated (observed iff $Z = 1$)

$r_{C\ell}$: response if subject is control (observed iff $Z = 0$)

$\mathbf{x}_\ell$: observed covariates (a.k.a. features)

u_ℓ : unobserved covariates

The ideal matching

- Recall: we match 1 treated with 1 control subject
- Ideal matching: equal probability to be treated:
  $\pi_k = \pi_\ell$ for all matched pairs (k, ℓ)
- Why is this ideal? Because it entails that each individual is equally likely to be the treated one in the pair

$$
\begin{aligned}
&\Pr\left(Z_k = 1, Z_\ell = 0 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell, Z_k + Z_\ell = 1\right) \\
&= \frac{\Pr\left(Z_k = 1, Z_\ell = 0 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell\right)}{\Pr\left(Z_k + Z_\ell = 1 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell\right)} \\
&= \frac{\pi_\ell^{1+0}\,(1-\pi_\ell)^{(1-1)+(1-0)}}{\pi_\ell^{1+0}\,(1-\pi_\ell)^{(1-1)+(1-0)} + \pi_\ell^{0+1}\,(1-\pi_\ell)^{(1-0)+(1-1)}} \\
&= \frac{\pi_\ell\,(1-\pi_\ell)}{\pi_\ell\,(1-\pi_\ell) + \pi_\ell\,(1-\pi_\ell)} = \frac{1}{2}
\end{aligned}
$$

Ok, so are we done?

- Problem: You generally don't know the probabilities to treat:

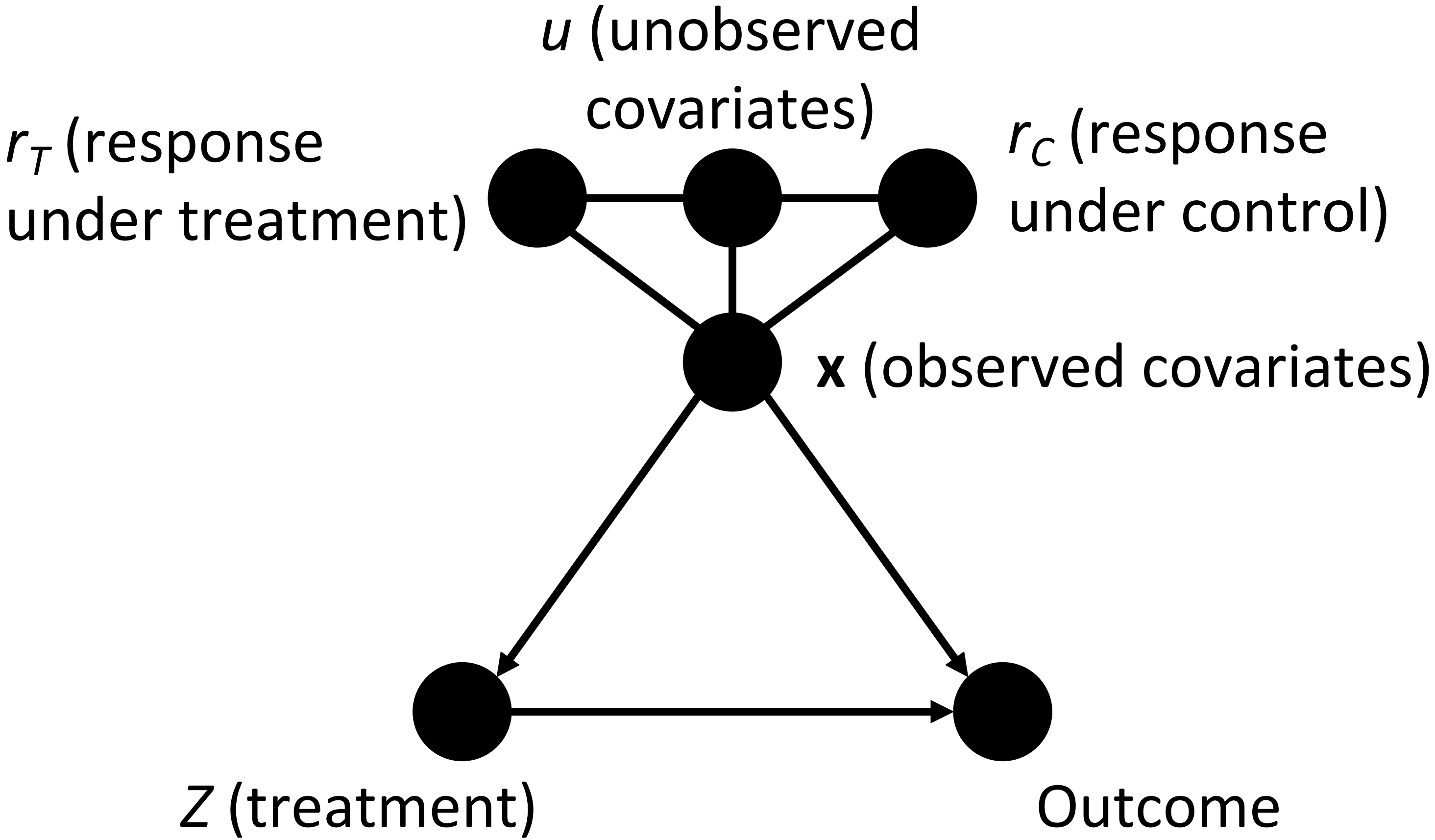
$$\pi_\ell = \Pr\left(Z_\ell = 1 \,\middle|\, \boxed{r_{T\ell}}, \boxed{r_{C\ell}}, \mathbf{x}_\ell, \boxed{u_\ell}\right)$$

- A **naive model:** "People who look comparable are comparable", or "Only observed variables determine treatment assignment":

$$\pi_\ell = \Pr\left(Z_\ell = 1 \,\middle|\, \cancel{r_{T\ell}}, \cancel{r_{C\ell}}, \mathbf{x}_\ell, \cancel{u_\ell}\right) = \Pr\left(Z_\ell = 1 \,\middle|\, \mathbf{x}_\ell\right)$$

i.e., $Z \perp\!\!\!\perp (r_T, r_C, u) \,\big|\, \mathbf{x}.$

Naive model as a causal diagram



u (unobserved covariates)

r_T (response under treatment)

r_C (response under control)

x (observed covariates)

Z (treatment)

Outcome

Slides by Dr. Robert West, Applied Data Analysis

If the naive model were true…

- … you could “simulate” a randomized experiment:
    - Simply match subjects with identical observed variables **x**
    - Subjects in a pair have the same probability to treat
    - So who gets treated is up to chance, as in experiment
    - Analysis: compare outcome for treated to outcome of control (e.g., mean difference treated-minus-control)

Two problems

1. Even if naive model were true: matching on **x** exactly may not be possible
   - E.g., if **x** contains 20 binary features: 1 million possible instantiations of **x**, so likely no match
   - Solution: propensity scores
2. The naive model is naive and rarely true
   - Solution: sensitivity analysis

Two problems

1. Even if naive model were true: matching on **x** exactly may not be possible
    - E.g., if **x** contains 20 binary features: 1 million possible instantiations of **x**, so likely no match
    - Solution: propensity scores
2. The naive model is naive and rarely true
    - Solution: sensitivity analysis

Propensity scores

- Matching on observed covariates **x** is rarely feasible
- Idea: reduce the information to a single number
- Definition: propensity score $\boxed{e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})}$
- Defined even when naive model not true
- But if naive model is true, it equals the probability to treat

$$\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell) = e(\mathbf{x}_\ell)$$

- Typically computed via logistic regression
  - Features: **x**; label: *Z*

Balancing property of propensity score

- Fact: all subjects (treated and control) with equal propensity score have equal distribution of observed covariates **x**:

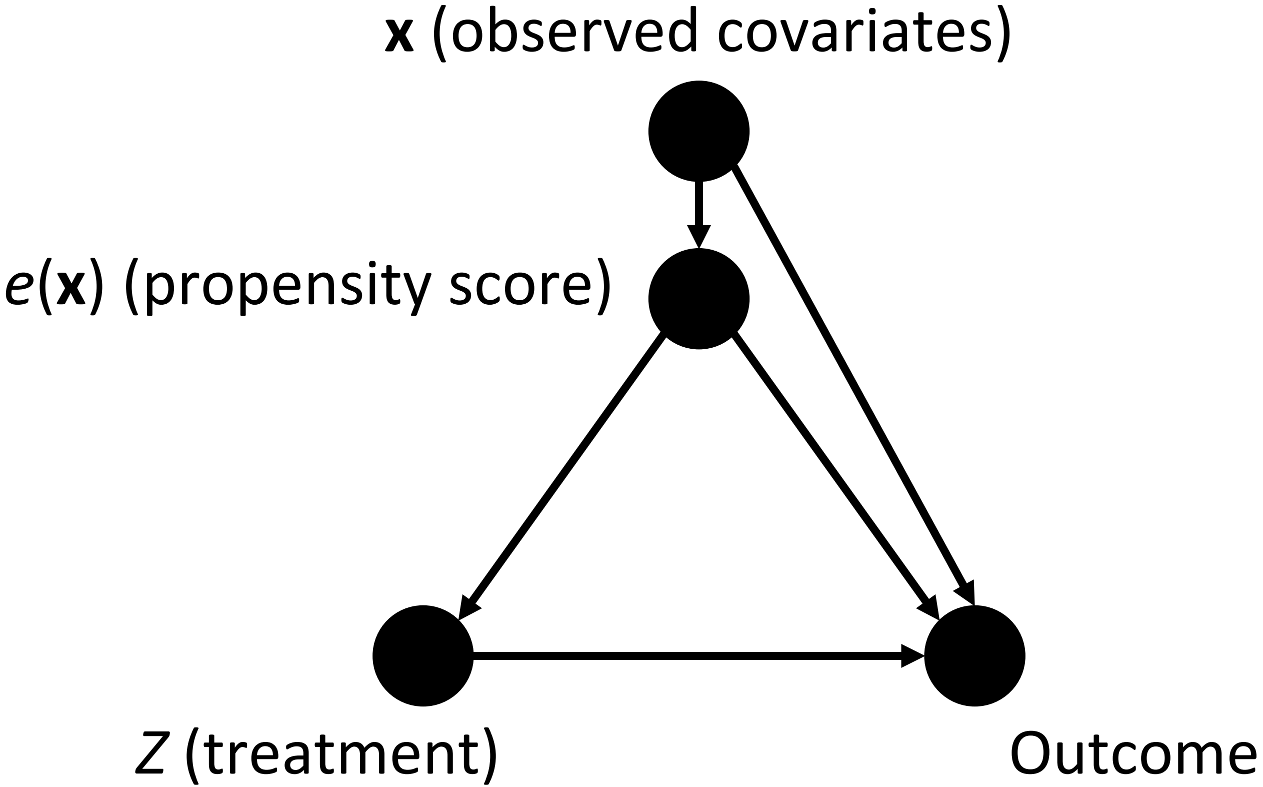
$$\Pr\{\mathbf{x} \mid Z = 1, e(\mathbf{x})\} = \Pr\{\mathbf{x} \mid Z = 0, e(\mathbf{x})\}$$

or equivalently

$$Z \perp\!\!\!\perp \mathbf{x} \,\big|\, e(\mathbf{x}) \,,$$

- Subjects in a matched pair might not have equal **x**, but treated and control groups will have similar distributions of **x**
- That is, matching on *e*(**x**) is just as good as matching on **x**

Balancing as a causal diagram



x (observed covariates)

e(**x**) (propensity score)

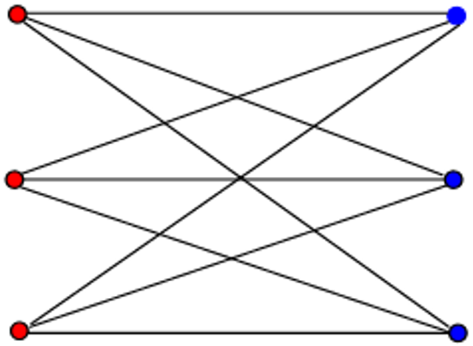
Z (treatment)

Outcome

Slides by Dr. Robert West, Applied Data Analysis

Matching

- Unlikely that 2 subjects have identical propensity scores *e*(**x**)
- → Matching
- Bipartite graph: each subject connected to all other subjects
- Edge weights: absolute (or squared) difference of propensity scores
- Find minimum matching, e.g., via Hungarian algorithm



Two problems

1. Even if naive model were true: matching on **x** exactly may not be possible
    - E.g., if **x** contains 20 binary features: 1 million possible instantiations of **x**, so likely no match
    - Solution: propensity scores
2. The naive model is naive and rarely true
    - Solution: sensitivity analysis

The sensitivity analysis model

- Idea: Quantify the degree to which the naive model is wrong
- More concretely, model assumes that treatment odds of two identically-looking subjects (i.e., identical observed covariates **x**) differ by a bounded factor *Γ*
- Then reasoning: "To change the conclusions of my study, two identically-looking people (1 treated, 1 control) would have hugely different treatment odds (i.e., huge *Γ*). Common sense (or domain knowledge) suggests that this is not the case, so my conclusions stand."

The sensitivity analysis model

$$\frac{1}{\Gamma} \leq \frac{\pi_k/(1-\pi_k)}{\pi_\ell/(1-\pi_\ell)} \leq \Gamma \quad \text{whenever} \quad \mathbf{x}_k = \mathbf{x}_\ell \,. \qquad \Gamma \geq 1.$$

- Bounded odds ratio
- Odds isomorphic to probabilities
  - e.g., prob 2/3 = odds 2/1; prob 1/2 = odds 1/1
- Sensitivity $\Gamma = 1$ → naive model is true
- Sensitivity $\Gamma = 2$ → subject with same observed covariates **x** up to twice as likely to be the one to receive treatment
- Sensitivity $\Gamma = \infty$ → void statement

Example: smoking and lung cancer

- Under naive model: matching on observed covariates gives a very small *p*-value for the null hypothesis, which states that smoking does not increase lung cancer risk (using an appropriate hypothesis test)
- Tobacco lobby: “The naive model isn’t true! There may be hidden (e.g., genetic) correlates that increase both the probability to enjoy smoking and the probability of lung cancer. They, not smoking, cause cancer!”

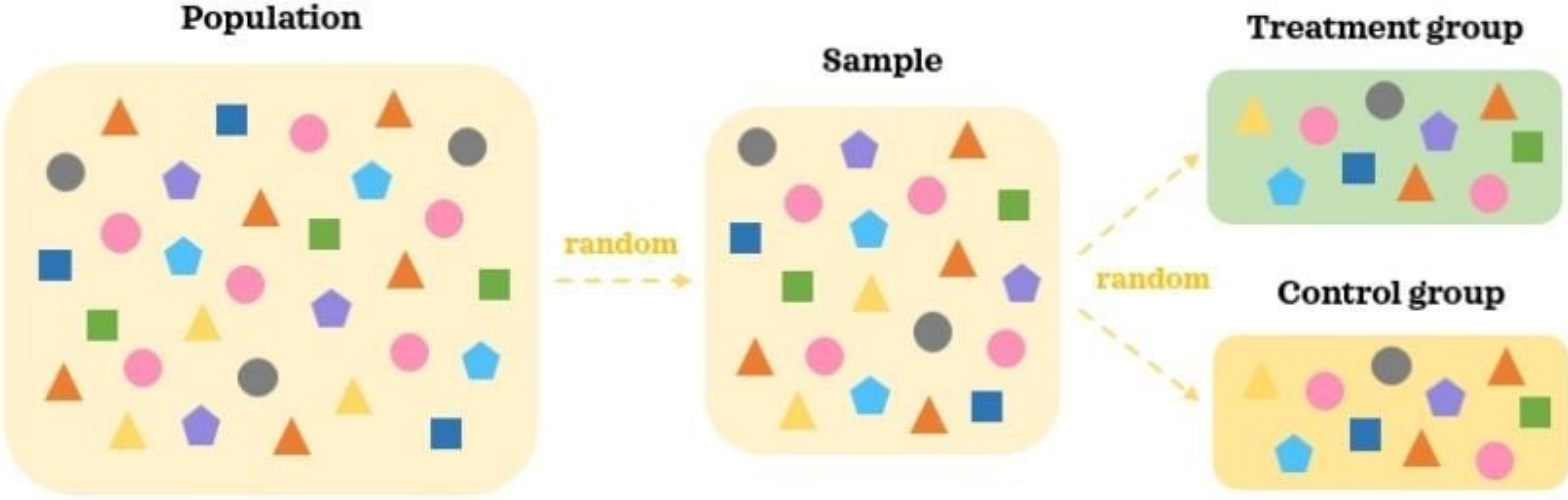
Example: smoking and lung cancer

- Under sensitivity analysis model, increasing sensitivity *Γ* increases the *p*-value for null hypothesis
- Anti-tobacco lobby: “But making *p* > 0.05 would require *Γ* > 6; i.e., the odds of being a smoker would need to be six times higher for one of two people with the exact same observed features (age, gender, education, income, ...). It’s unlikely that any unobserved covariate would have such a large effect on smoking habits. So smoking causes cancer!”

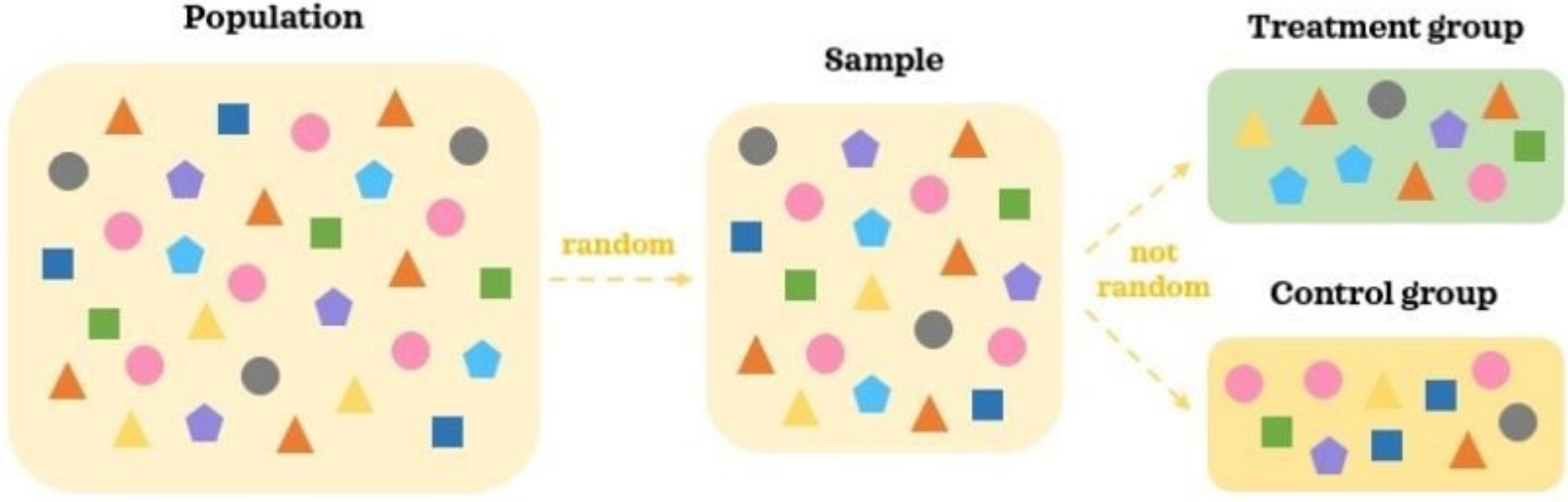
Two parts: mechanical vs. scientific

- Mechanical part:
  - Create pairs (1 treated + 1 control) with similar observed covariates (using exact or propensity-score matching)
- Scientific (i.e., fun) part:
  - Mitigate concerns that your findings might be caused by unobserved covariates, rather than treatment (e.g., using sensitivity analysis, ad-hoc arguments, natural experiments)

Experimental data



Observational data



Summary

- Holy grail: randomized experiment
- When experiment not possible: observational study
- Crucial problem: treatment assignment not random (biases!)
- Semi-holy grail: natural experiment
- Matched studies: pair up treated/control based on observed covariates
- Problem: still, treatment assignment not random (biases via unobserved covariates)
- Solution: sensitivity analysis
- Keep this lecture (more here) in mind for your projects!

Credits

- Much of the material is based on Paul Rosenbaum's amazing book "Design of Observational Studies", available for free here