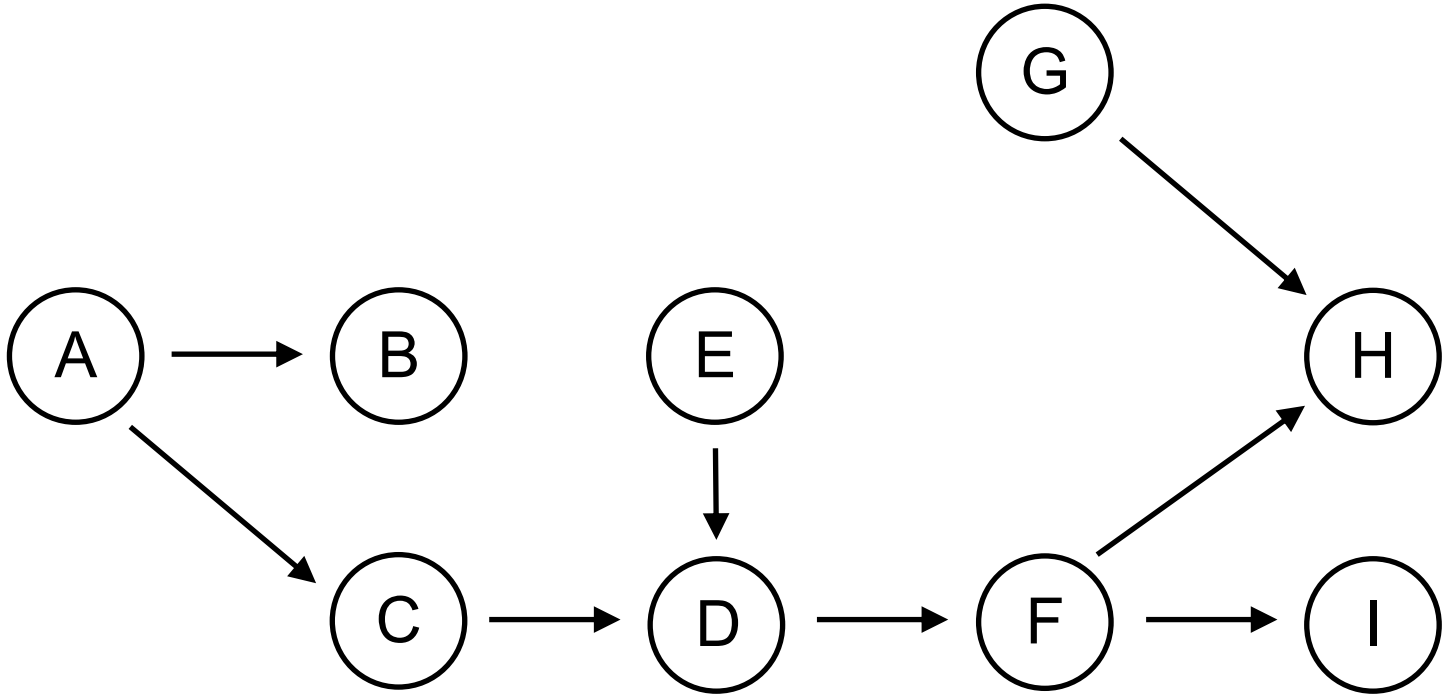


Introduction to causal inference

Roni Kobrosly, PhD

SciPy 2022

"I would rather discover one causal law than be King of Persia"

- Democritus (460-370 B.C.)

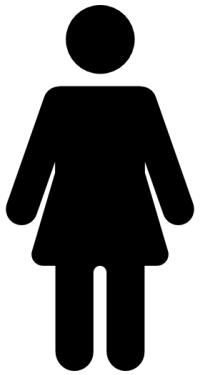
By the end of this tutorial, you should be able to

- Understand the pitfalls of observational data analysis
- Understand the various types of causal relationships
- Describe the hierarchy of statistical analyses, causal inference, and experiments
- Start conducting preliminary causal analyses on your own data
- Confidently explore the topic on your own (now that you have a solid foundational understanding of causal thinking)

Does Vitamin D supplementation prevent severe covid symptoms?

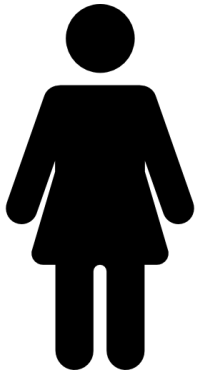
The alternative universe example

Our Observed reality



The alternative universe example

Our Observed reality

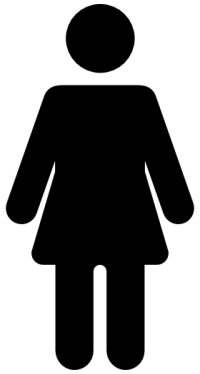


Early 2020

Starts Taking Vitamin D

The alternative universe example

Our Observed reality



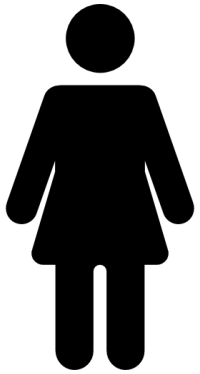
Early 2020

Starts Taking Vitamin D

Lives life

The alternative universe example

Our Observed reality



Early 2020

Starts Taking Vitamin D

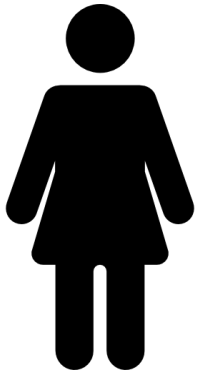
Lives life

Mid 2020

Gets covid with Mild symptoms

The alternative universe example

Our Observed reality



Early 2020

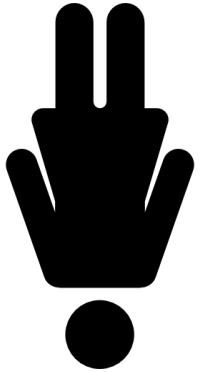
Starts Taking Vitamin D

Lives life

Mid 2020

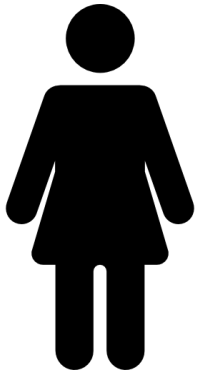
Gets covid with Mild symptoms

An alternative reality



The alternative universe example

Our Observed reality



Early 2020

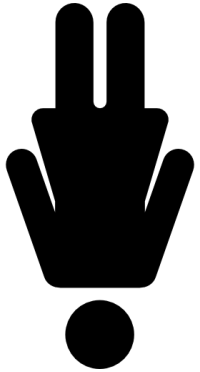
Starts Taking Vitamin D

Lives life

Mid 2020

Gets covid with Mild symptoms

An alternative reality

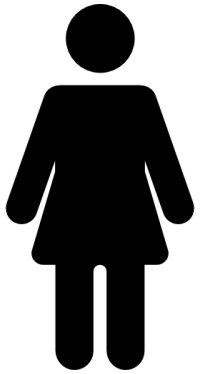


Early 2020

Never takes vitamin D

The alternative universe example

Our Observed reality



Early 2020

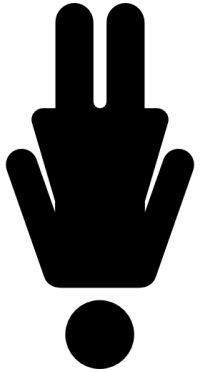
Starts Taking Vitamin D

Lives life

Mid 2020

Gets covid with Mild symptoms

An alternative reality



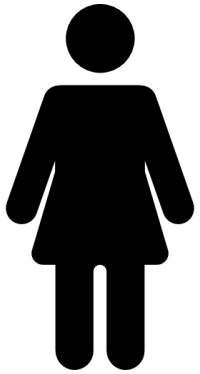
Early 2020

Never takes vitamin D

Lives life

The alternative universe example

Our Observed reality



Early 2020

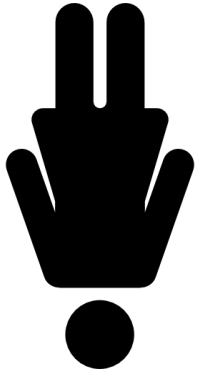
Starts Taking Vitamin D

Lives life

Mid 2020

Gets covid with Mild symptoms

An alternative reality



Early 2020

Never takes vitamin D

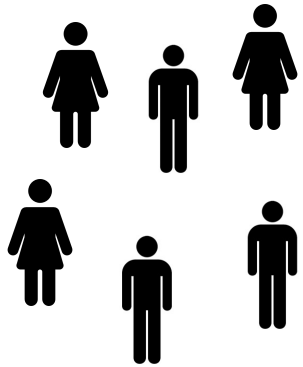
Lives life

Mid 2020

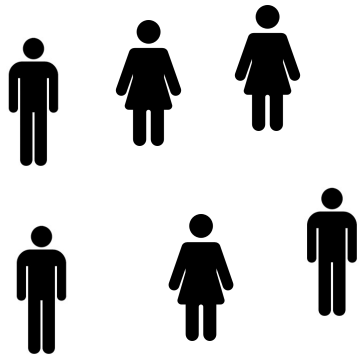
???

Experiments (AKA A/B Tests, AKA Randomized Controlled Trials)

Treatment group

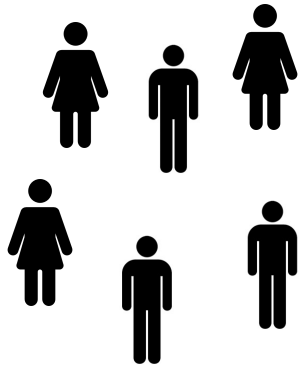


Control group



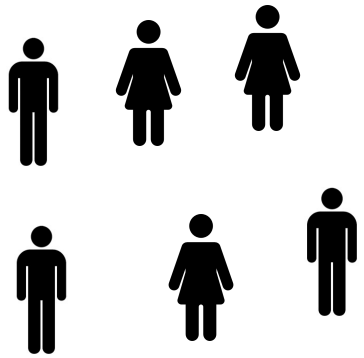
Experiments (AKA A/B Tests, AKA Randomized Controlled Trials)

Treatment group



given blank vitamin D pills

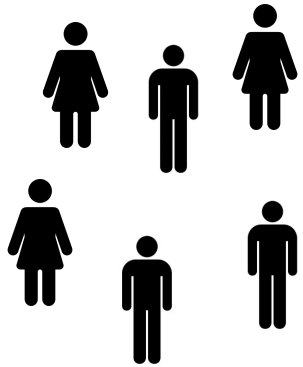
Control group



given blank sugar pills

Experiments (AKA A/B Tests, AKA Randomized Controlled Trials)

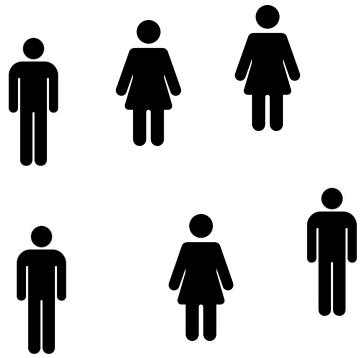
Treatment group



given blank vitamin D pills

See how many people get severe covid symptoms over next 2 months

Control group



given blank sugar pills

See how many people get severe covid symptoms over next 2 months

Experiments won't always save us

NOT ETHICAL: randomly assign some people to be exposed to lead paint while others are not, then see which group is more likely to develop neurological disorders.

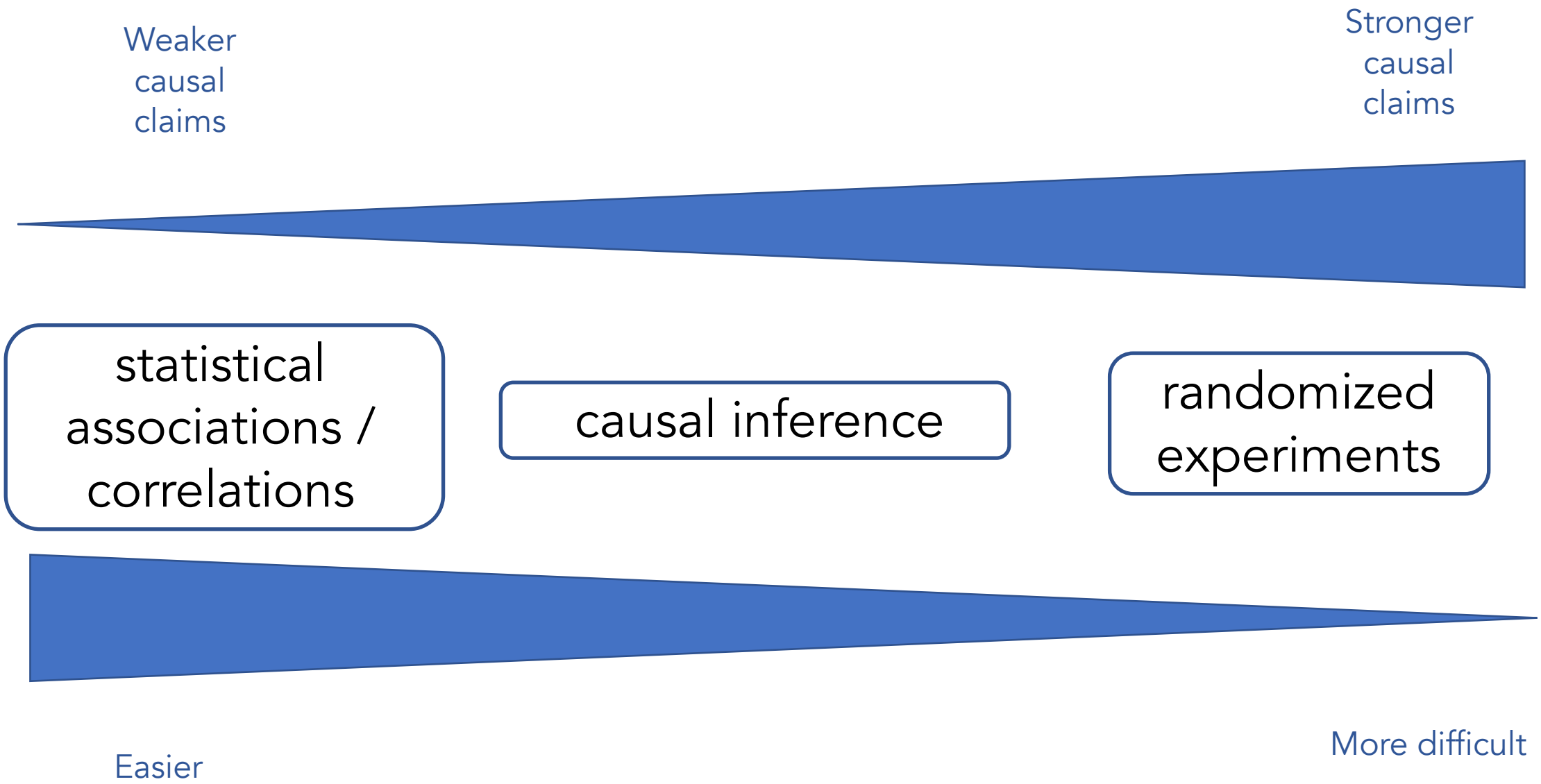
NOT FEASIBLE: modify household incomes in neighborhoods, to see if changing a neighborhood's income inequality improves local crime rate.

Pearl's causal hierarchy

Level	Typical Activity	Examples
1) Association	Seeing	• Is increased income inequality in a city correlated with more violent crime?
2) Intervention	Doing, intervening	• What happens if we ban the sale of cigarettes in this county?
3) Counterfactual	Imagining, Retrospection	• If Lucy hadn't been smoking cigarettes the last 10 years, would she still have developed cancer? • Was it the aspirin that stopped my headache?

J. Pearl. 2018. "The Seven Tools of Causal Inference with Reflections on Machine Learning"

A simplified hierarchy…



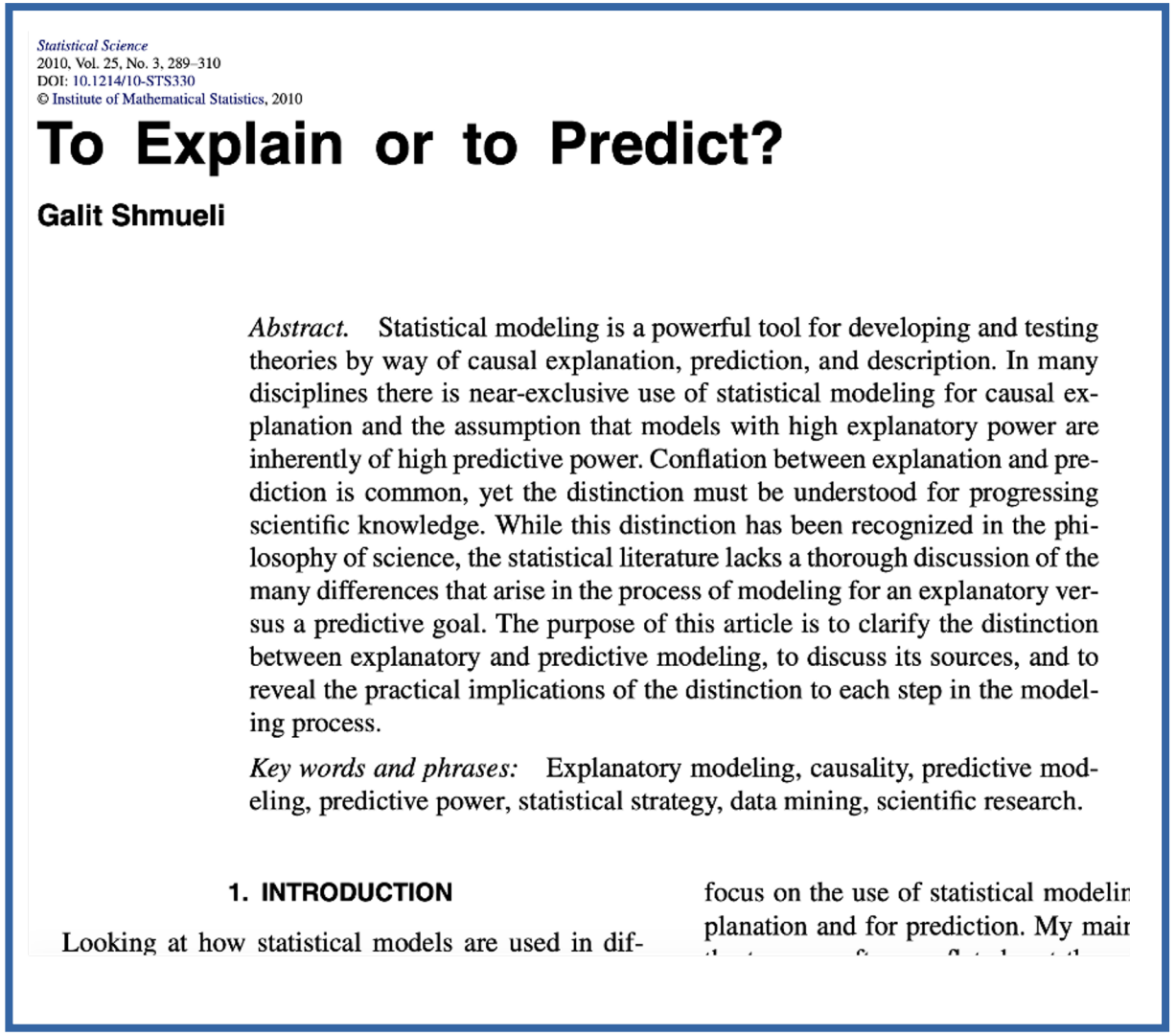
Inference vs prediction in modeling

Inference:

- What is the sample mean of X?
- Is X associated with Y?
- What is strength of that association?

Prediction:

- How can I best predict Y?
- Between A, B, and C, what variable is the best predictor of Y?
- Why did the model make a particular prediction (AKA interpretability)



Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

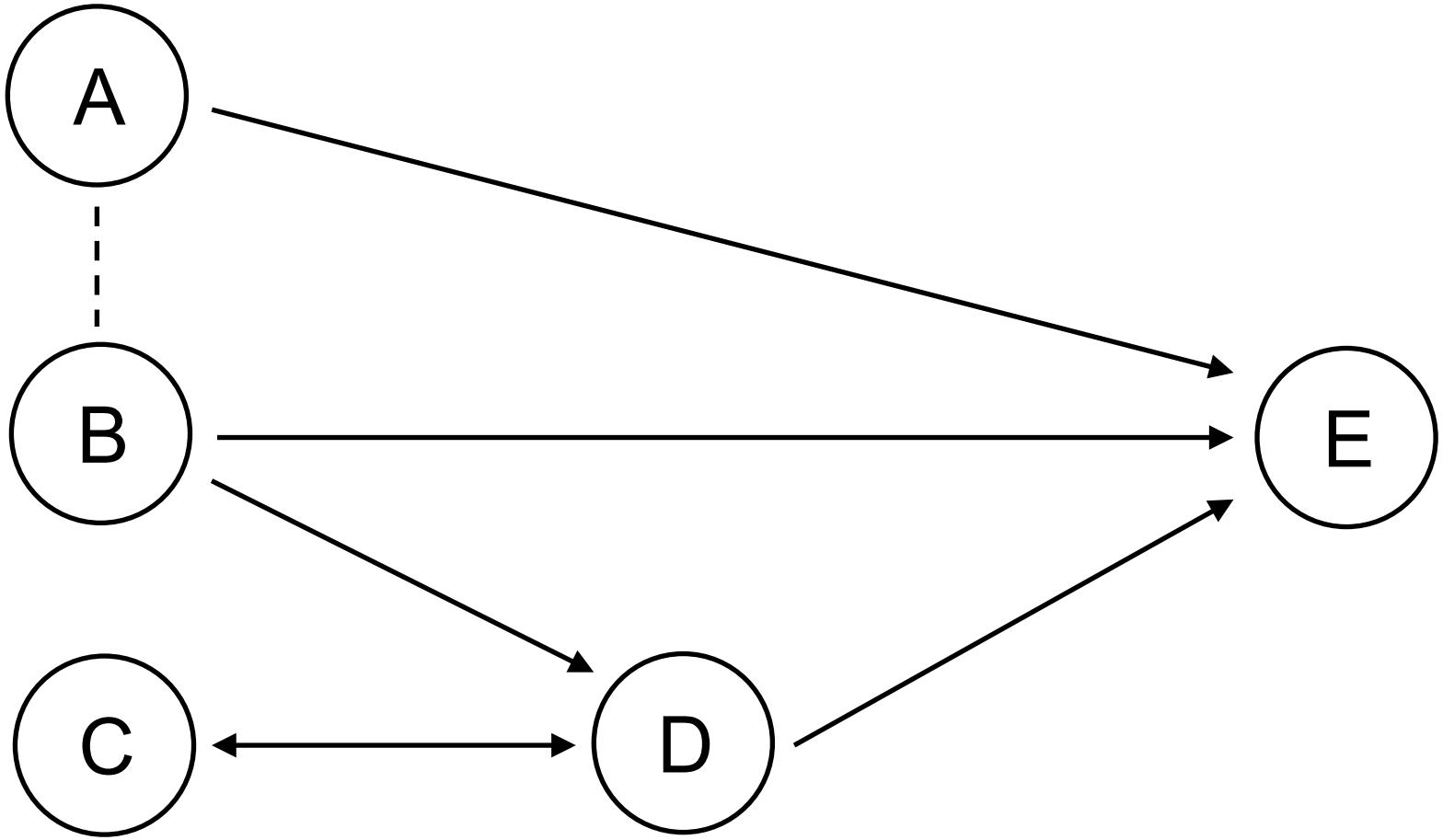
Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

1. INTRODUCTION

Looking at how statistical models are used in dif-

focus on the use of statistical modelir
planation and for prediction. My mair

A causal graph



Car insurance exercise

Make & model

Antilock

Car safety rating

Socioeconomics

Theft history

Vehicle Year

Age

Advanced airbag

Car value

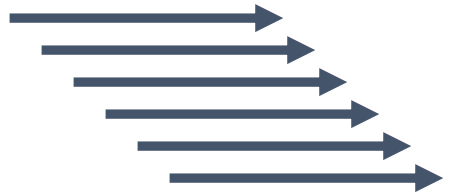
Good student?

Risk aversion

Driving course?

Accident history

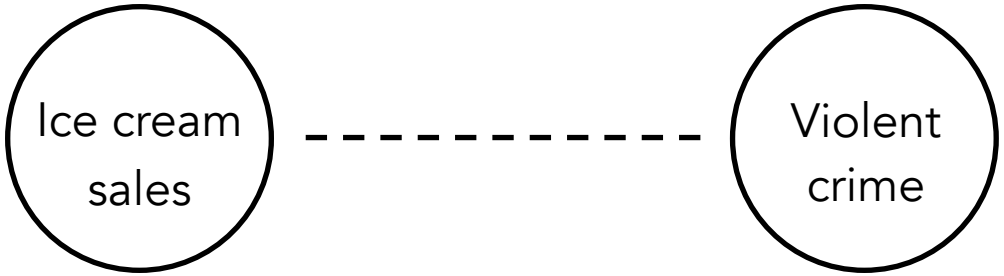
Anti-theft tech



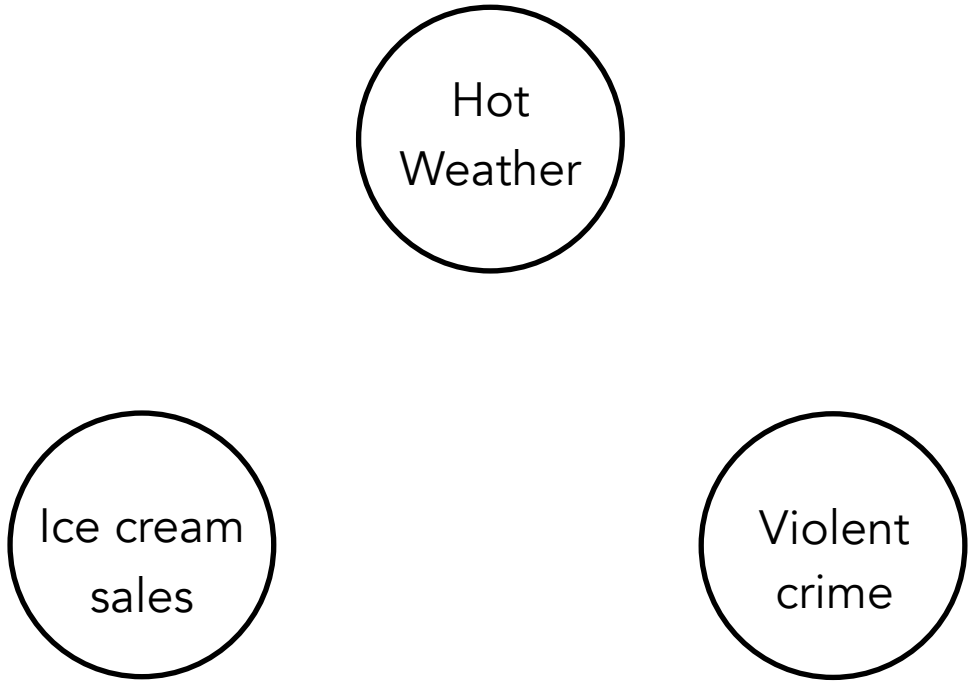
Medical Cost

Property Cost

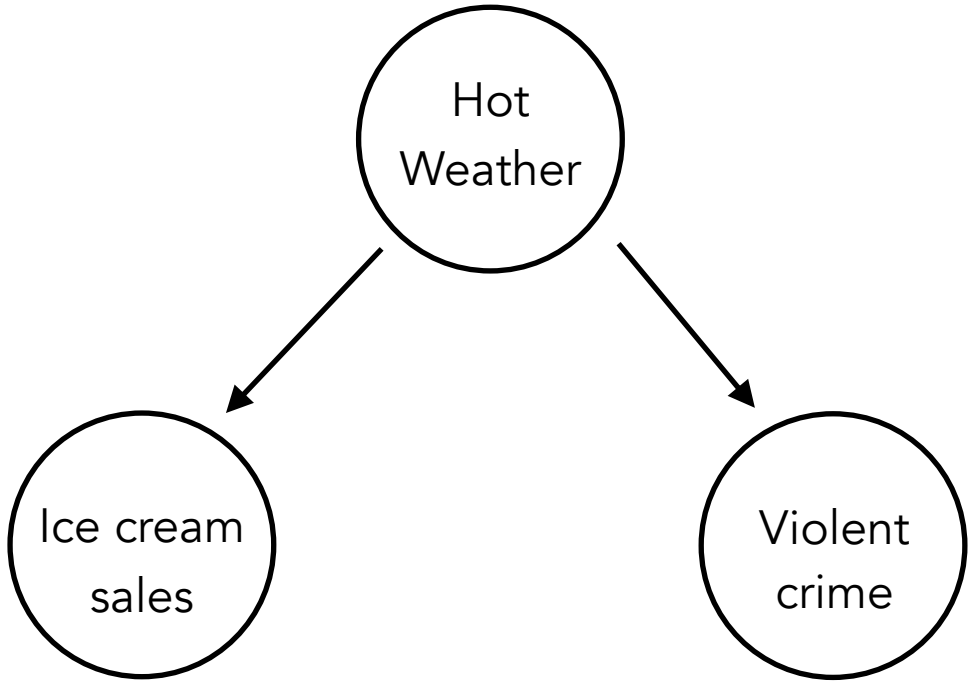
Ice cream and violent crime



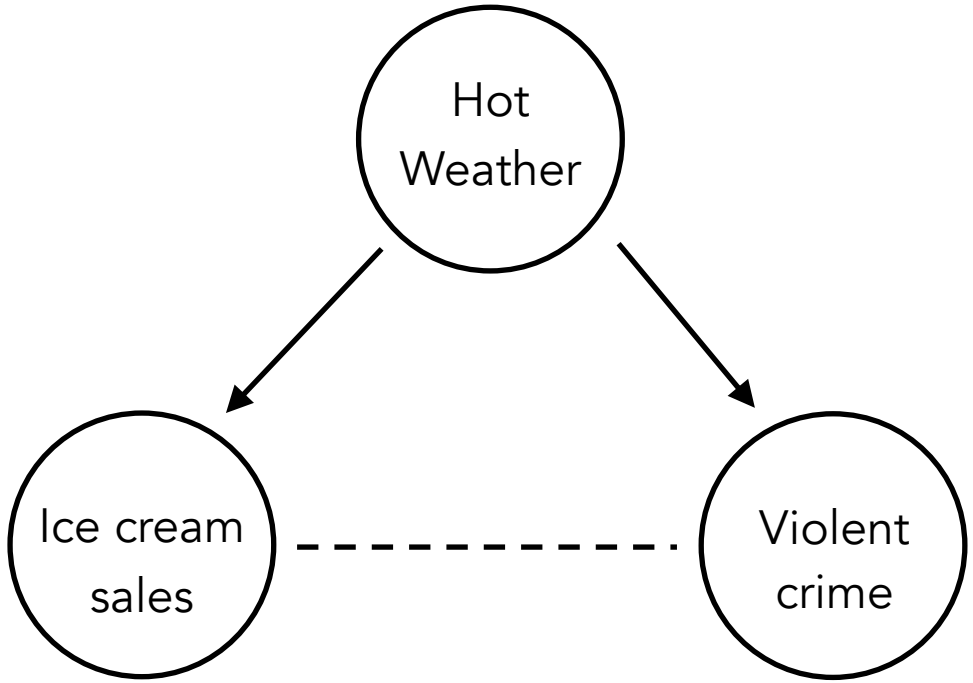
Summer weather induces a false association between ice cream sales and violent crime



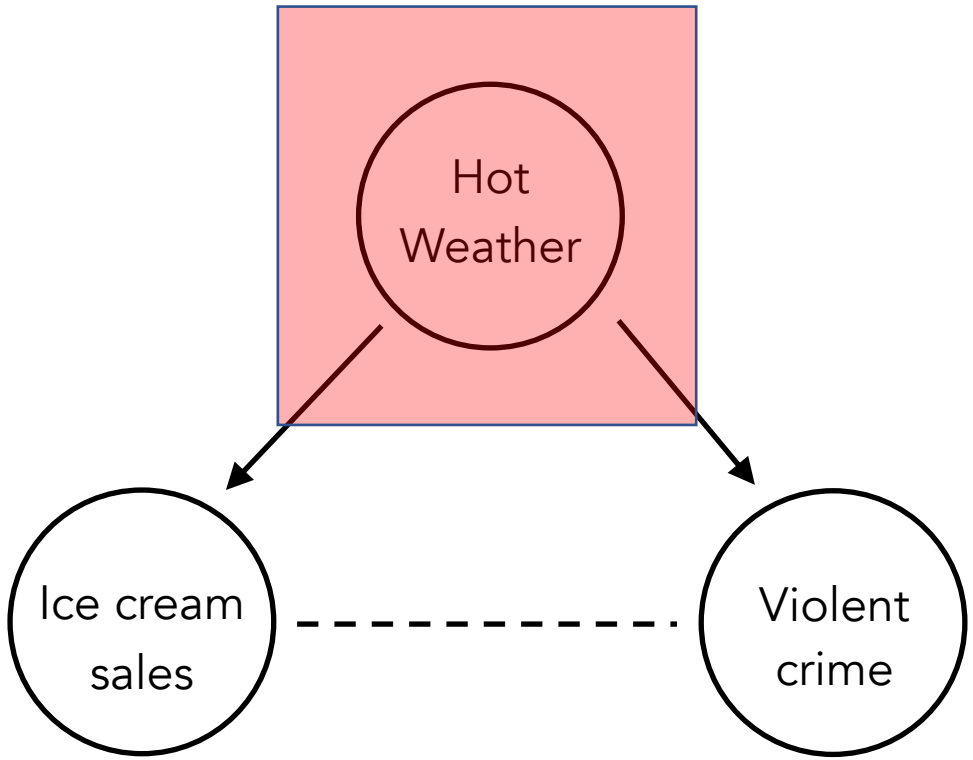
Summer weather induces a false association between ice cream sales and violent crime



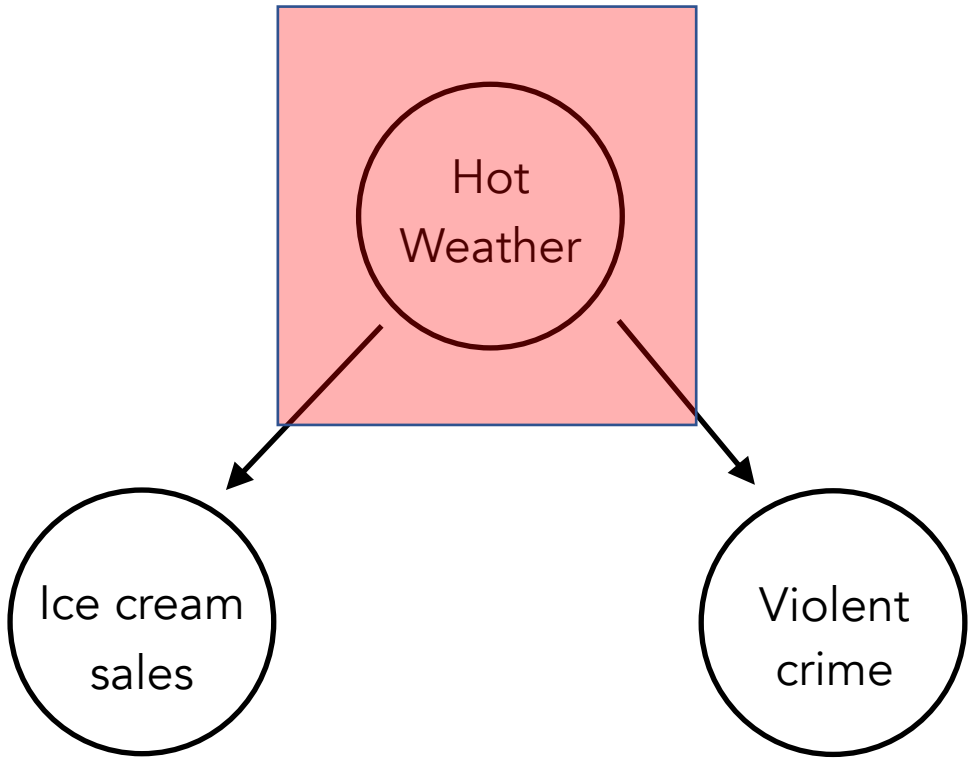
Summer weather induces a false association between ice cream sales and violent crime



Control for the season and then the ice cream-violent crime association disappears

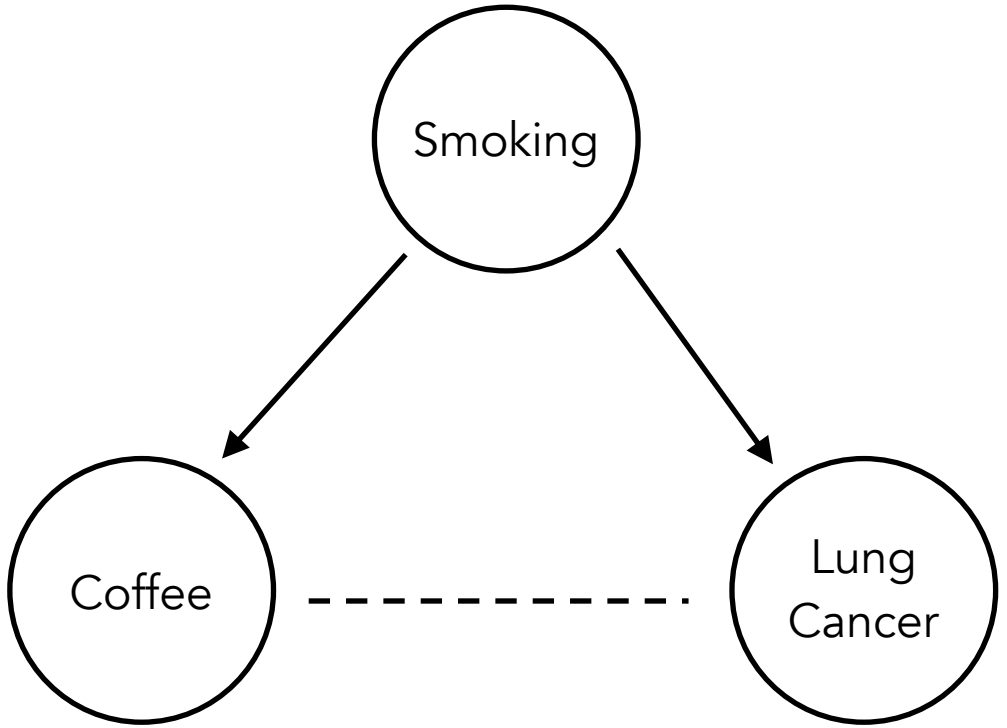


Control for the season and then the ice cream-violent crime association disappears



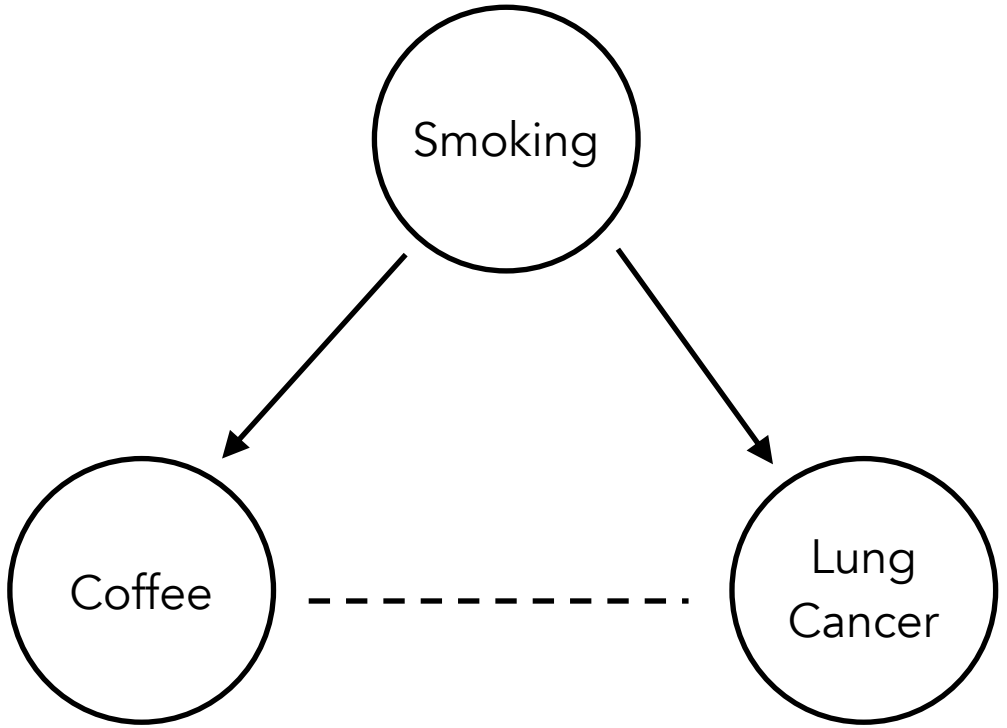
Confounders

- Always want to control for / condition on confounders in inferential modeling
- Confounding changes the effect size and possibly statistical significance of your association of interest
- Confounders can flip the sign of your association of interest
- Leftover confounding in a model is named “residual confounding”

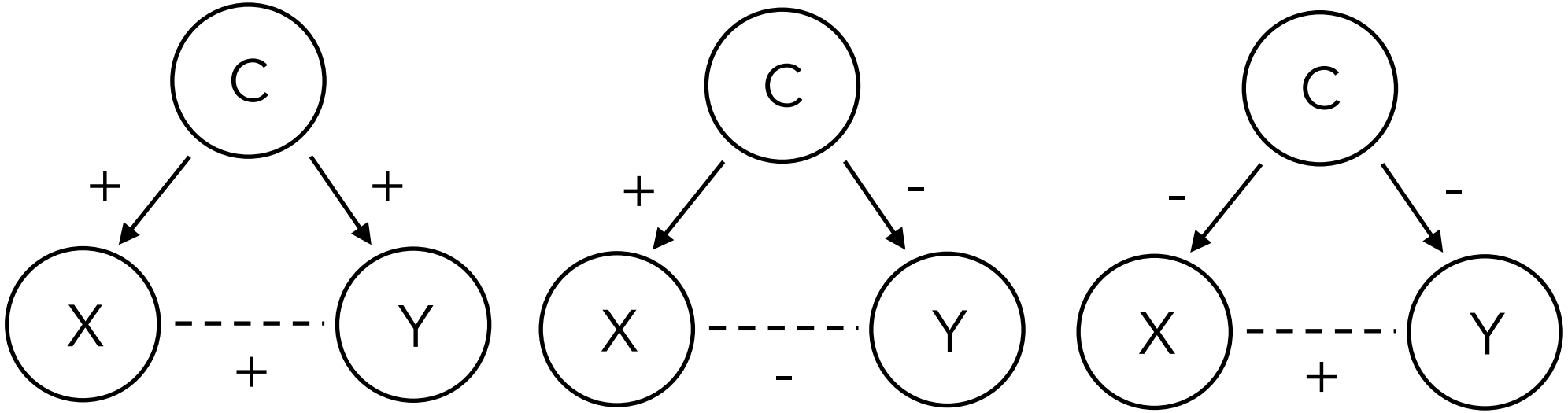


Confounders

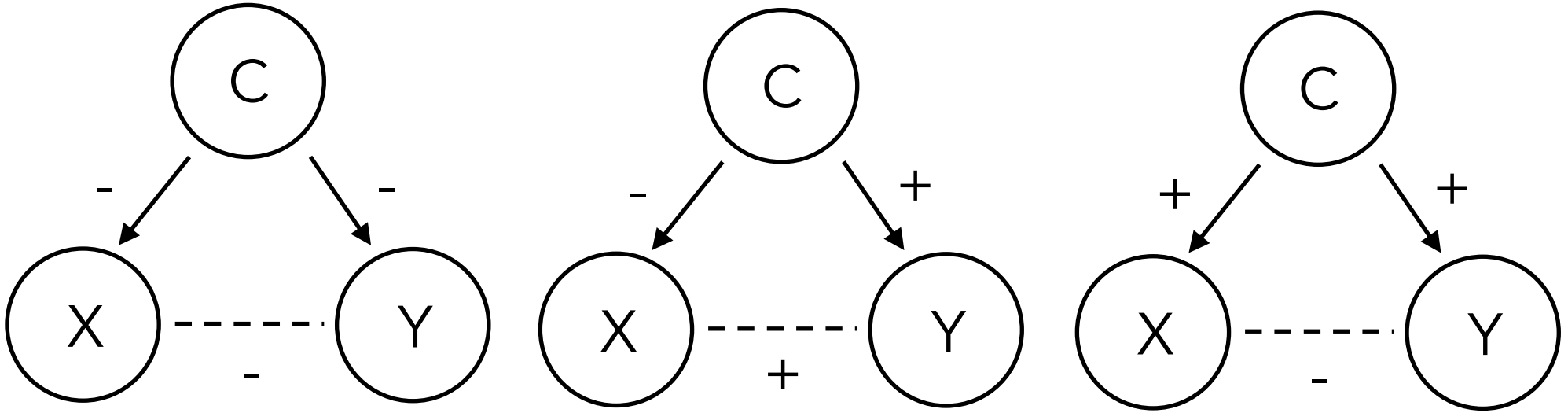
- Positive confounding: confounder introduces a bias that pushes association of interest away from the “null”
- Negative confounding: confounder biases association towards the “null”



Positive Confounding:

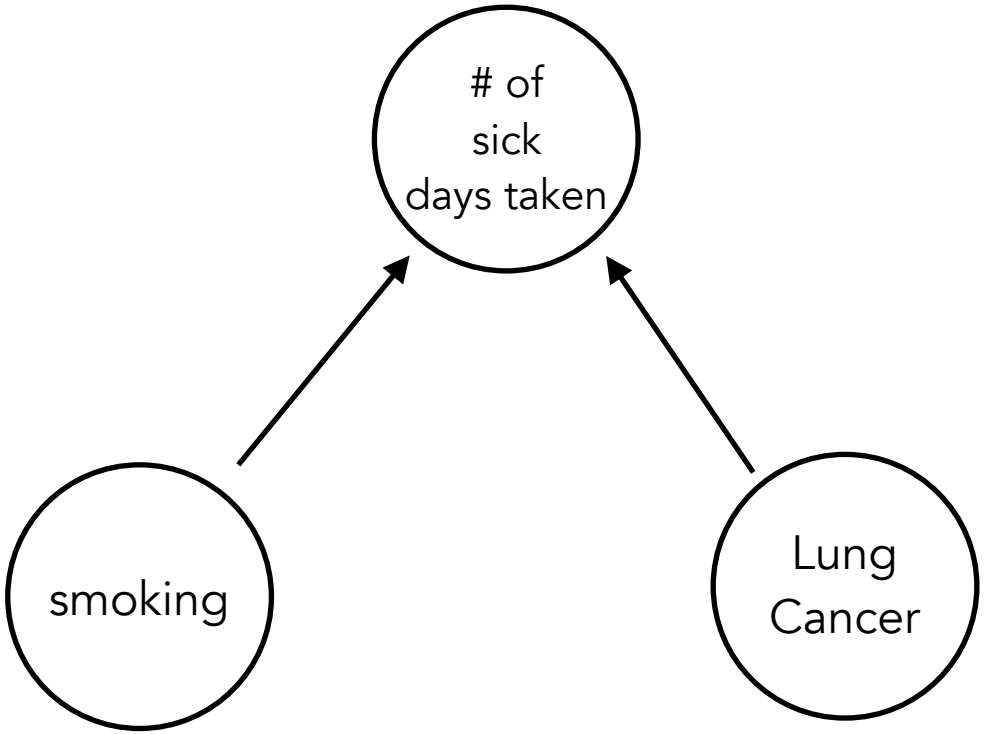


Negative: Confounding:



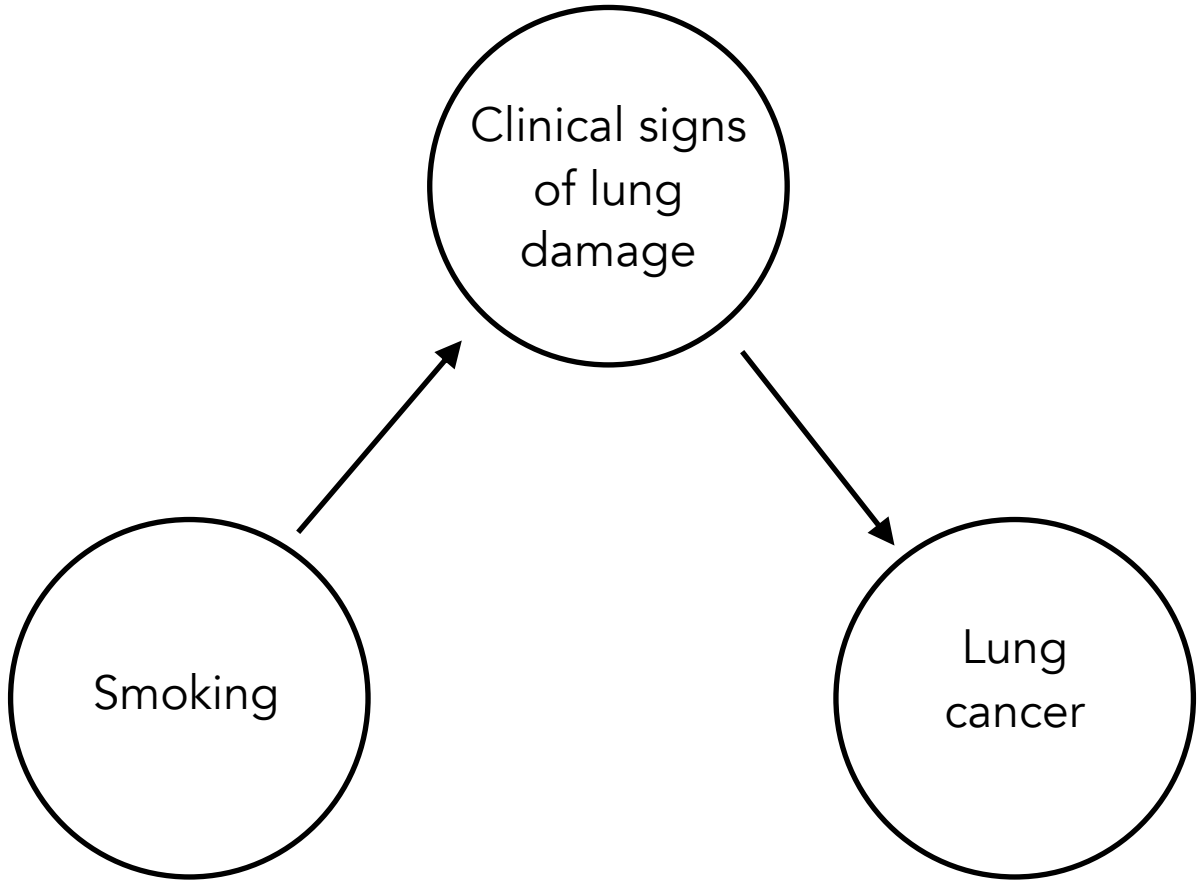
Colliders

- Never want to control for / condition on colliders
- Conditioning on a common effect causes collider bias, which can be in positive or negative direction



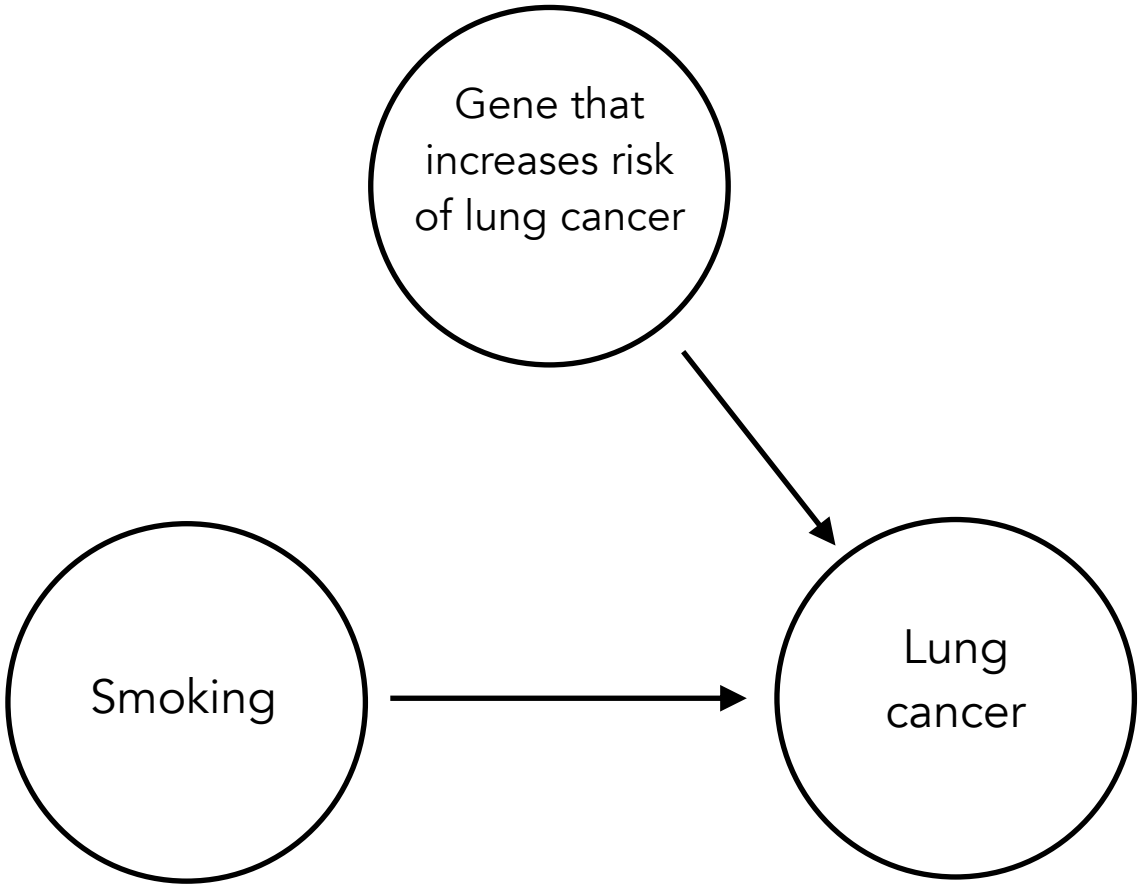
Mediators

- Controlling for a mediator will nullify associations of interest
- There are statistical tests of mediation you can use to help determine causal relationships in observational data

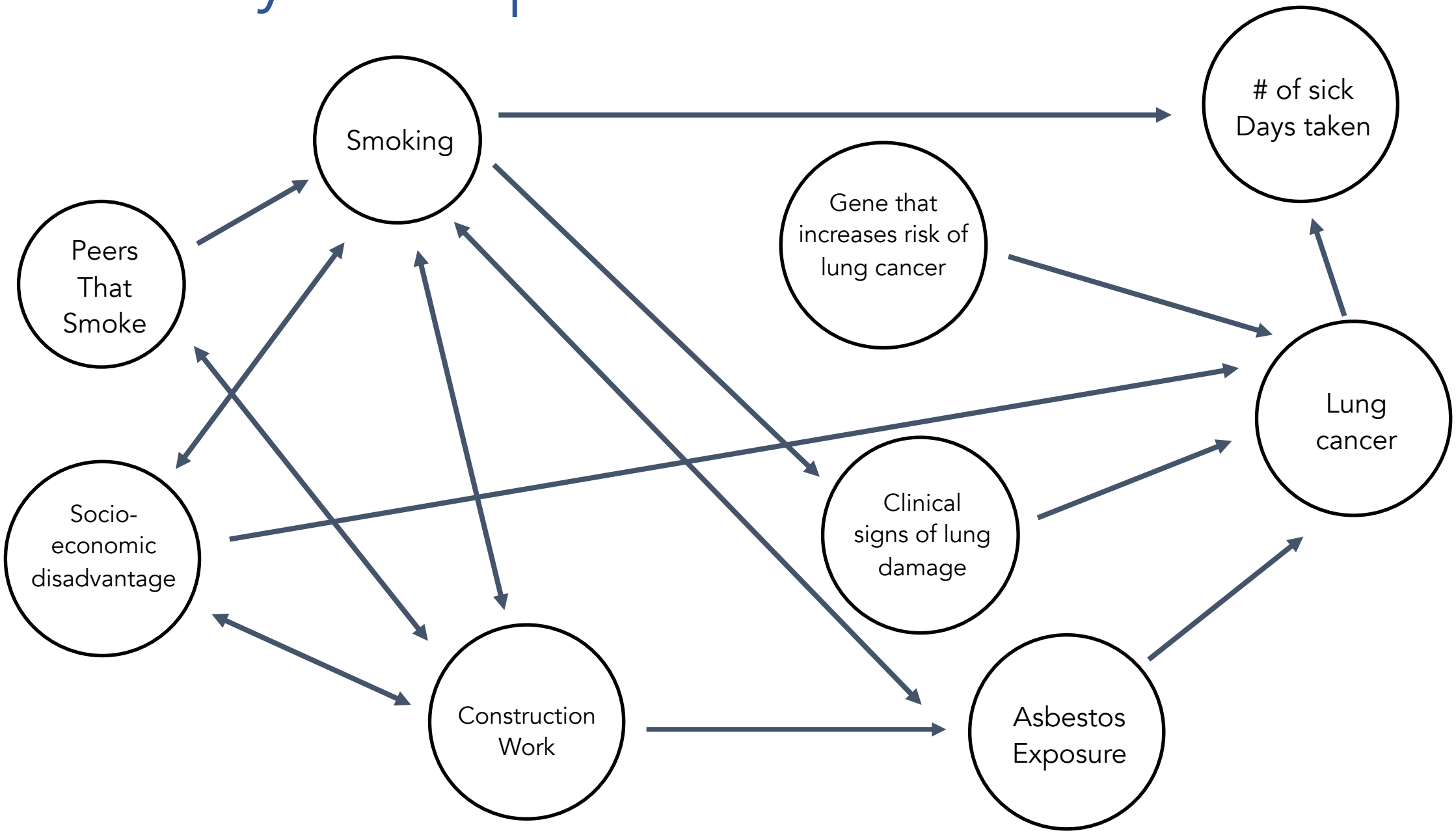


Unrelated predictors

- If you are working with a large N, incorporating them in your model will improve overall model performance and actually shrink variance of other effect estimates.
- If working with a small N, this can take up degrees of freedom, so be careful.



Causality is complicated!



Bias and fairness

- Pretty clear by now that applying black box models towards decision making in any broad social process is a terrible idea
- e.g. predicting crime location, recidivism, who gets approved for a credit card, which job applicant's resume should be looked at, etc…

ARTICLES

Sex Bias in Graduate Admissions: Data from Berkeley

P. J. Bickel[1], E. A. Hammel[1], J. W. O'Connell[1]

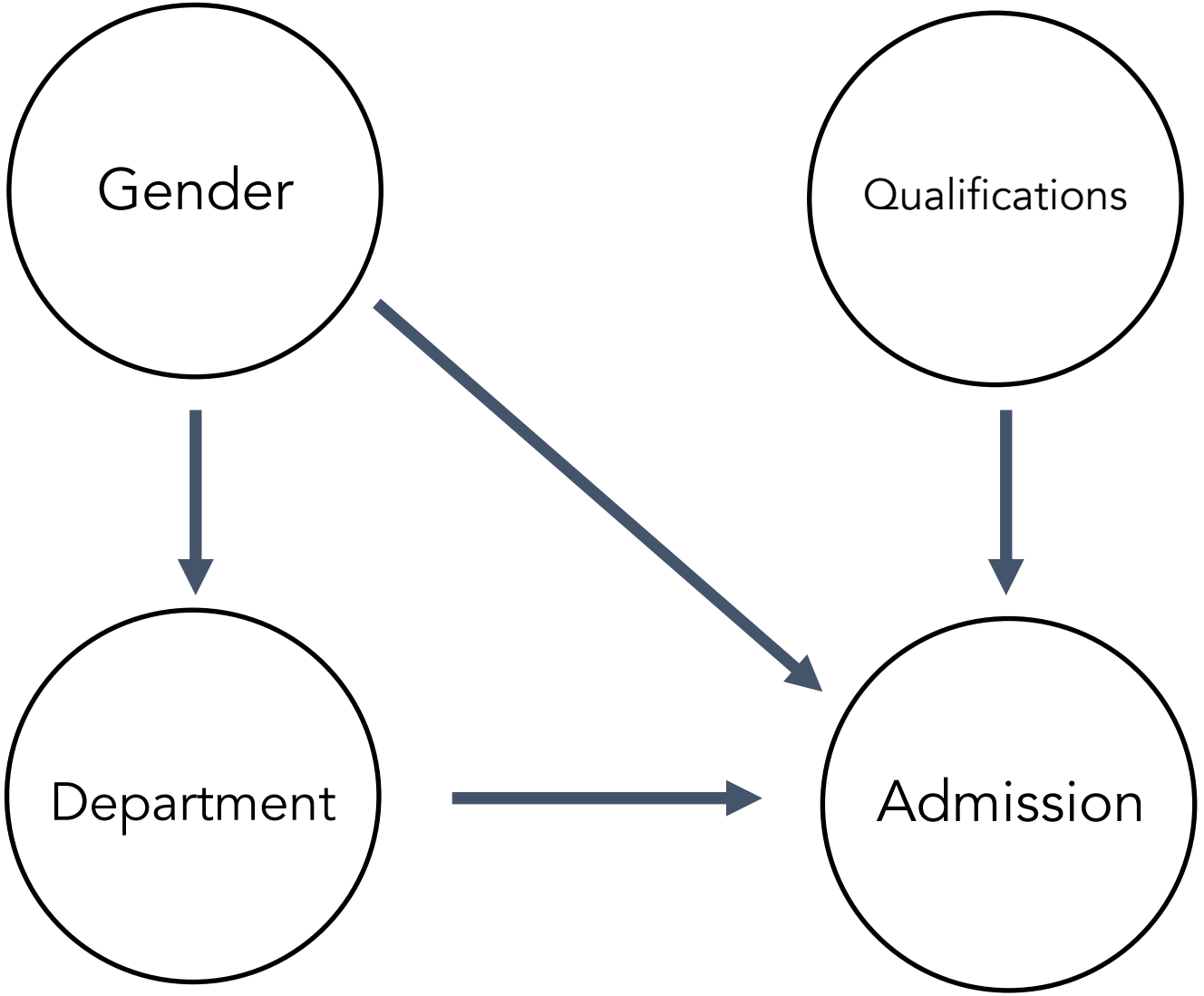
+ See all authors and affiliations

Science 07 Feb 1975:
Vol. 187, Issue 4175, pp. 398-404
DOI: 10.1126/science.187.4175.398

Abstract

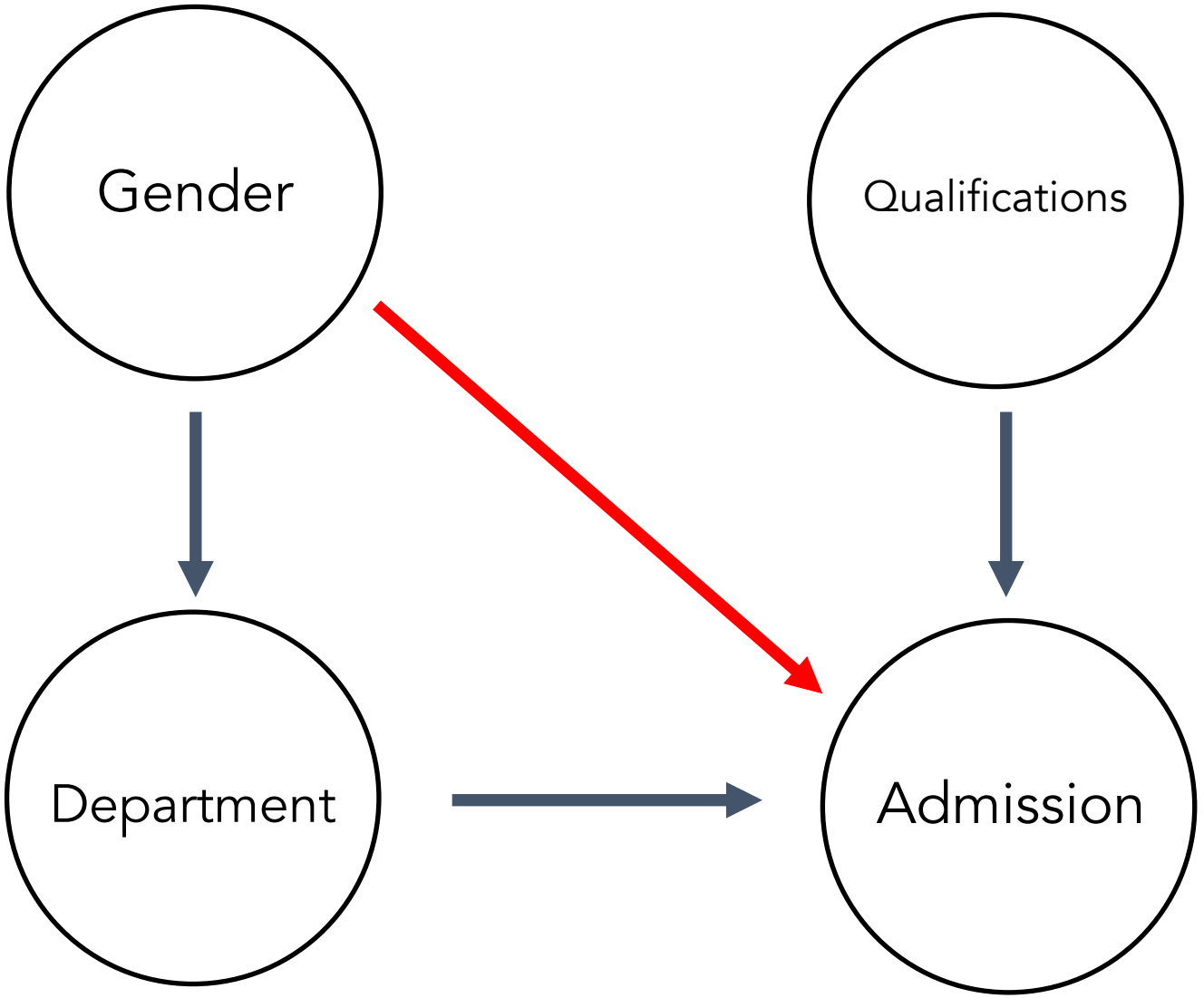
Examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. Examination of the disaggregated data reveals few decision-making units that show

The college admission process

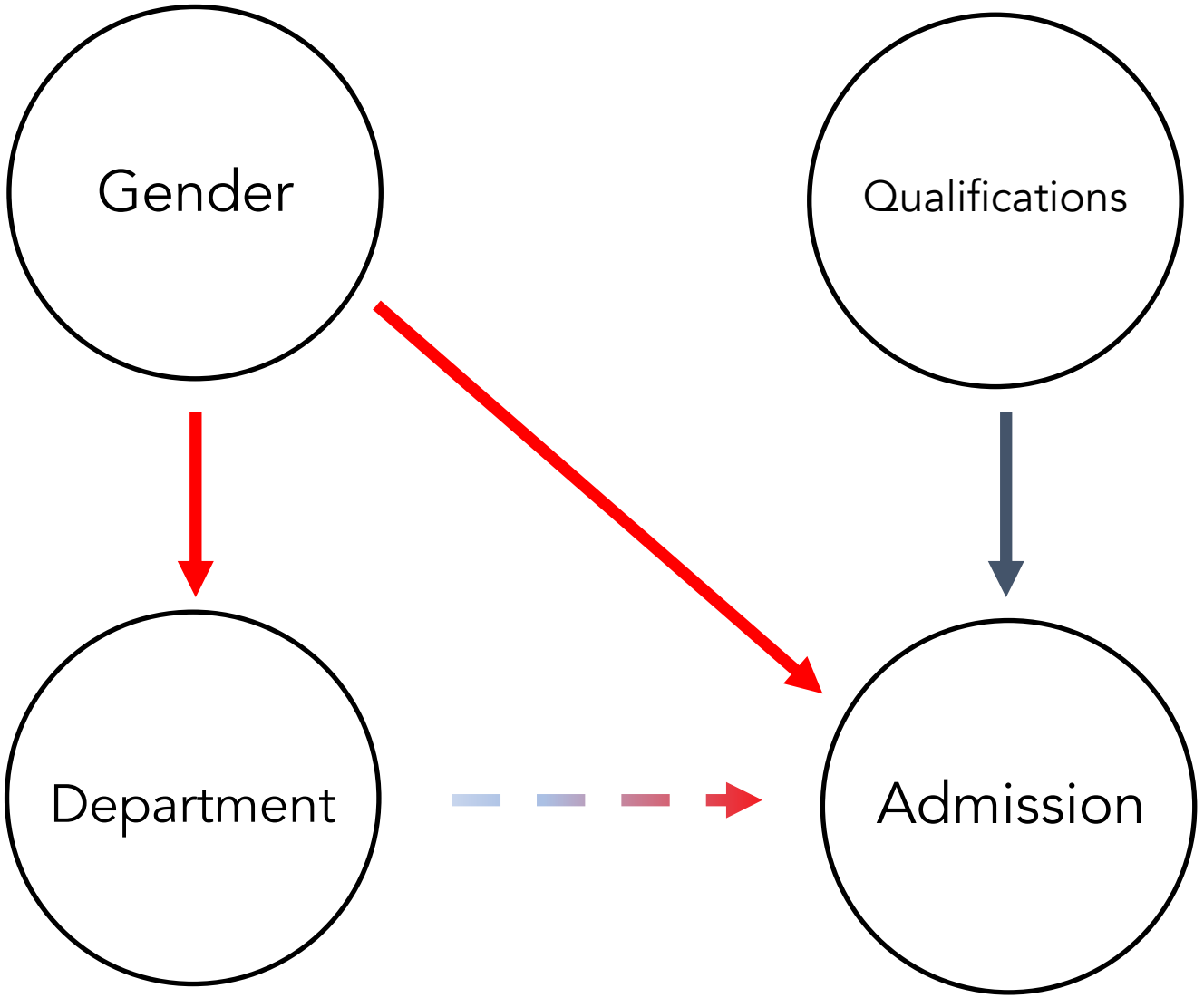


S chiappa et al. deepmind blog: “Causal Bayesian Networks: A flexible tool to enable fairer machine learning”. 2019

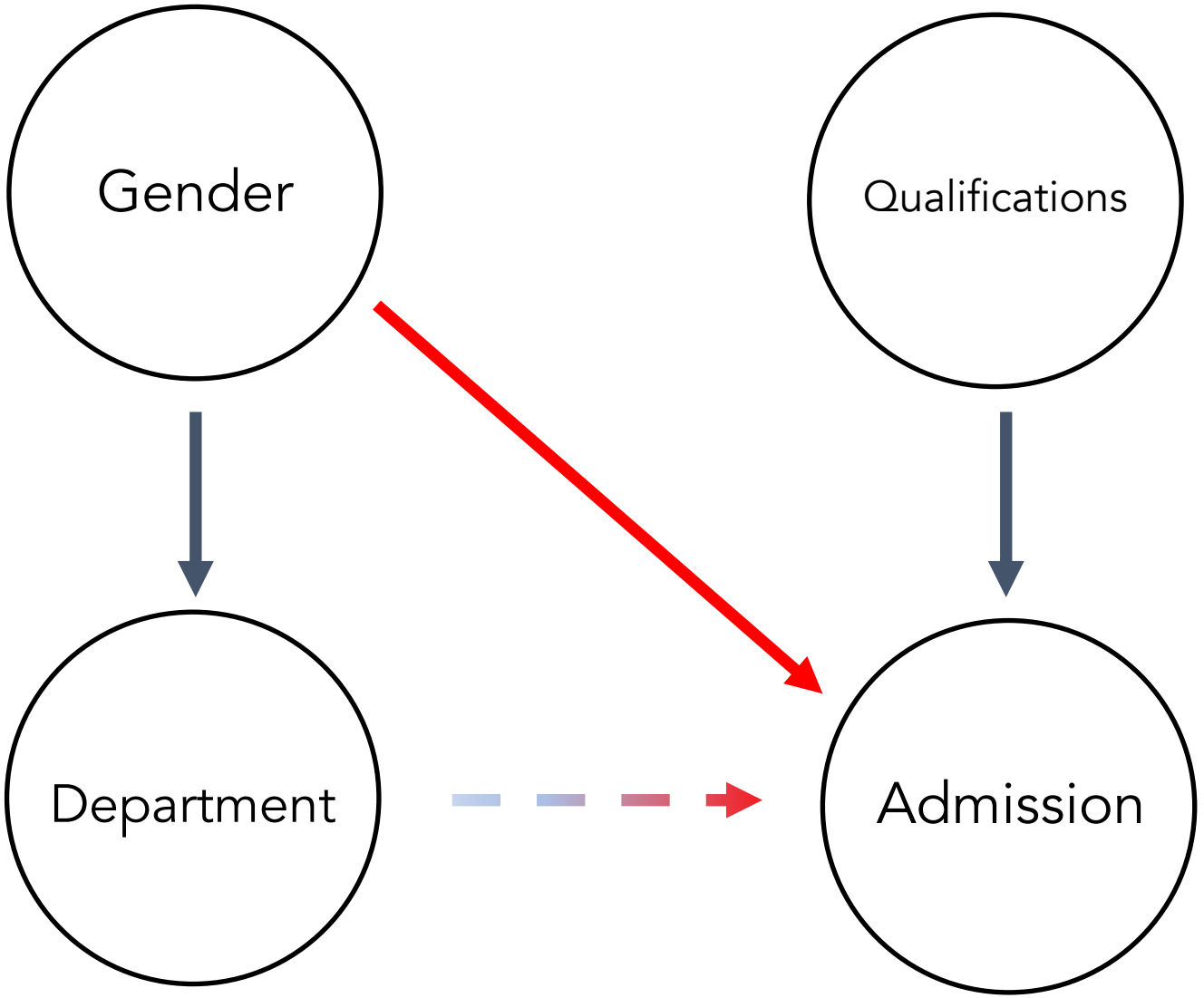
Scenario #1



Scenario #2



Scenario #3

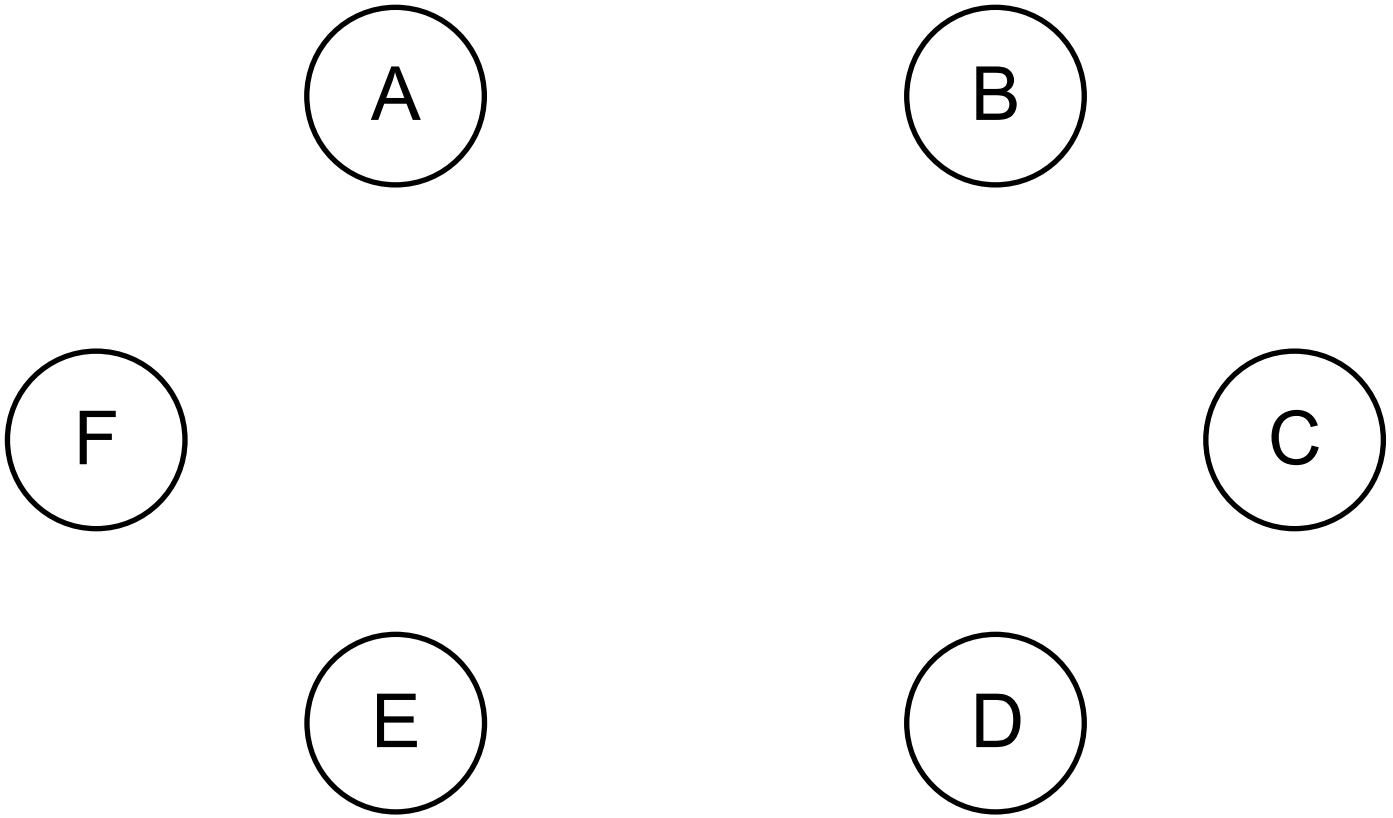


If extra time…
here are some slides on causal structure learning…

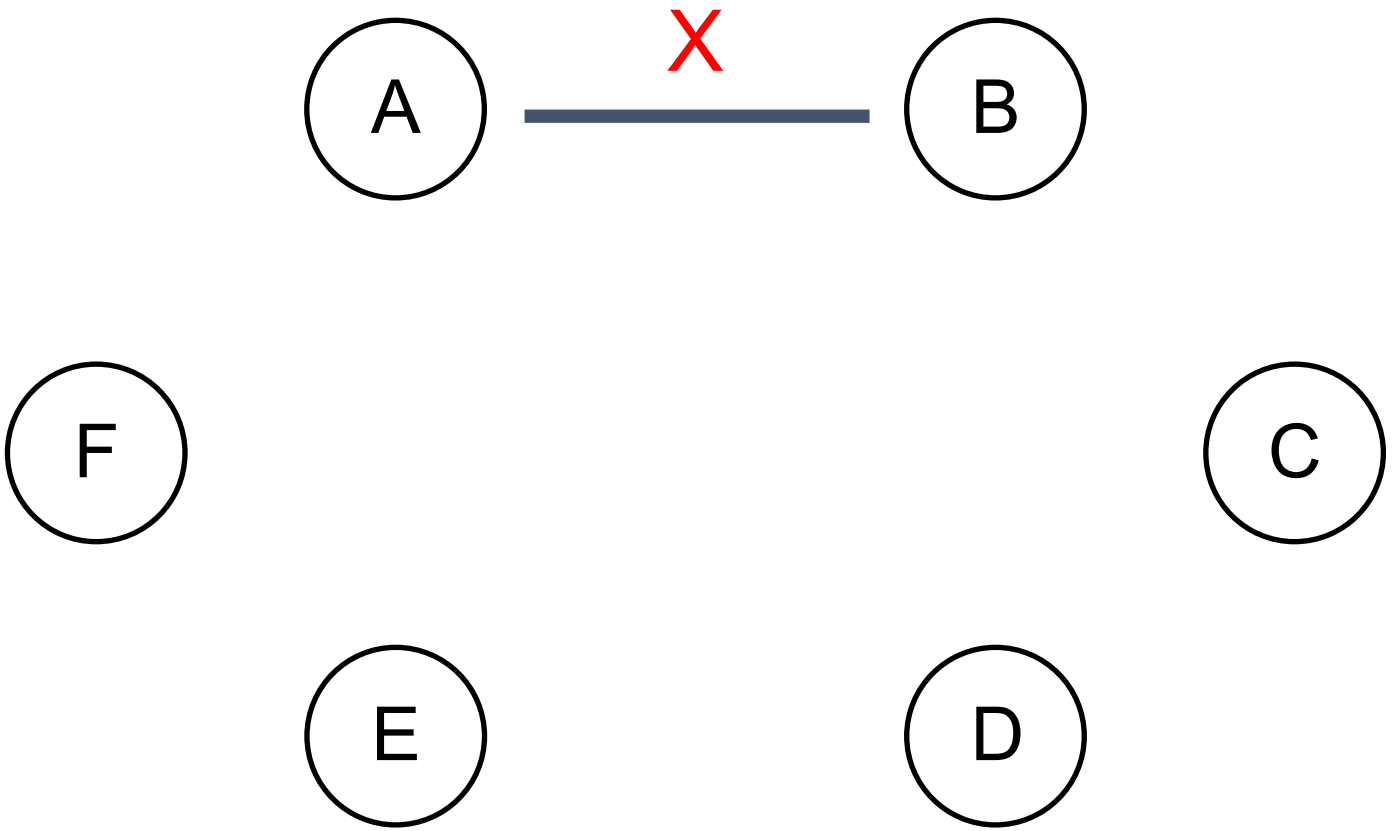
Causal graph learning (is very hard)

- It's a combinatorial nightmare: the number of possible DAGs can grow super-exponentially with the number of nodes.
- Blindly using structure learning on data without having domain knowledge is an awful idea
- A few classes of algorithms to assist with this:
  - Constraint-based algos
  - Score-based algos
  - Newer graph neural network approaches

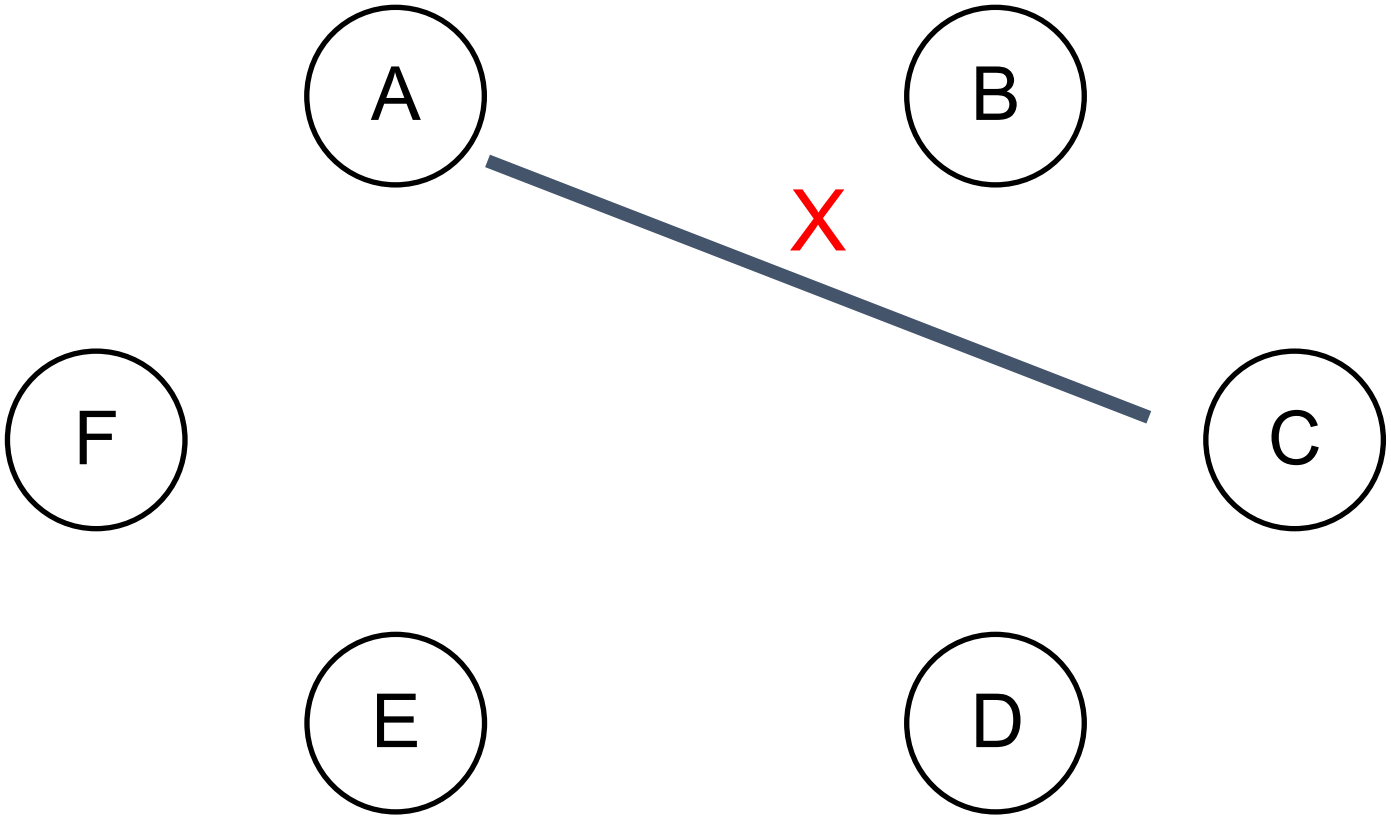
Constraint-based algos (AKA lots of little tests of association)



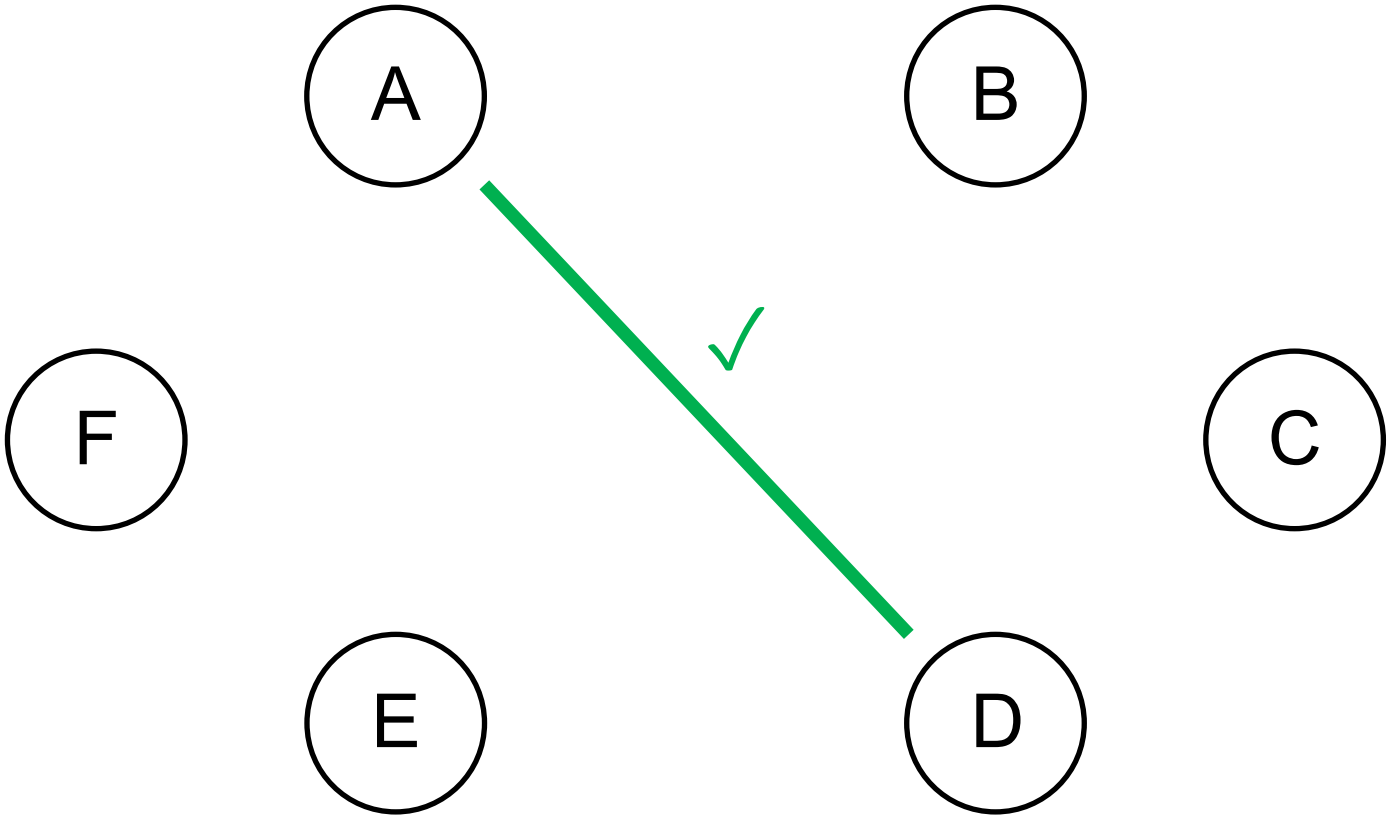
Constraint-based algos (AKA lots of little tests of association)



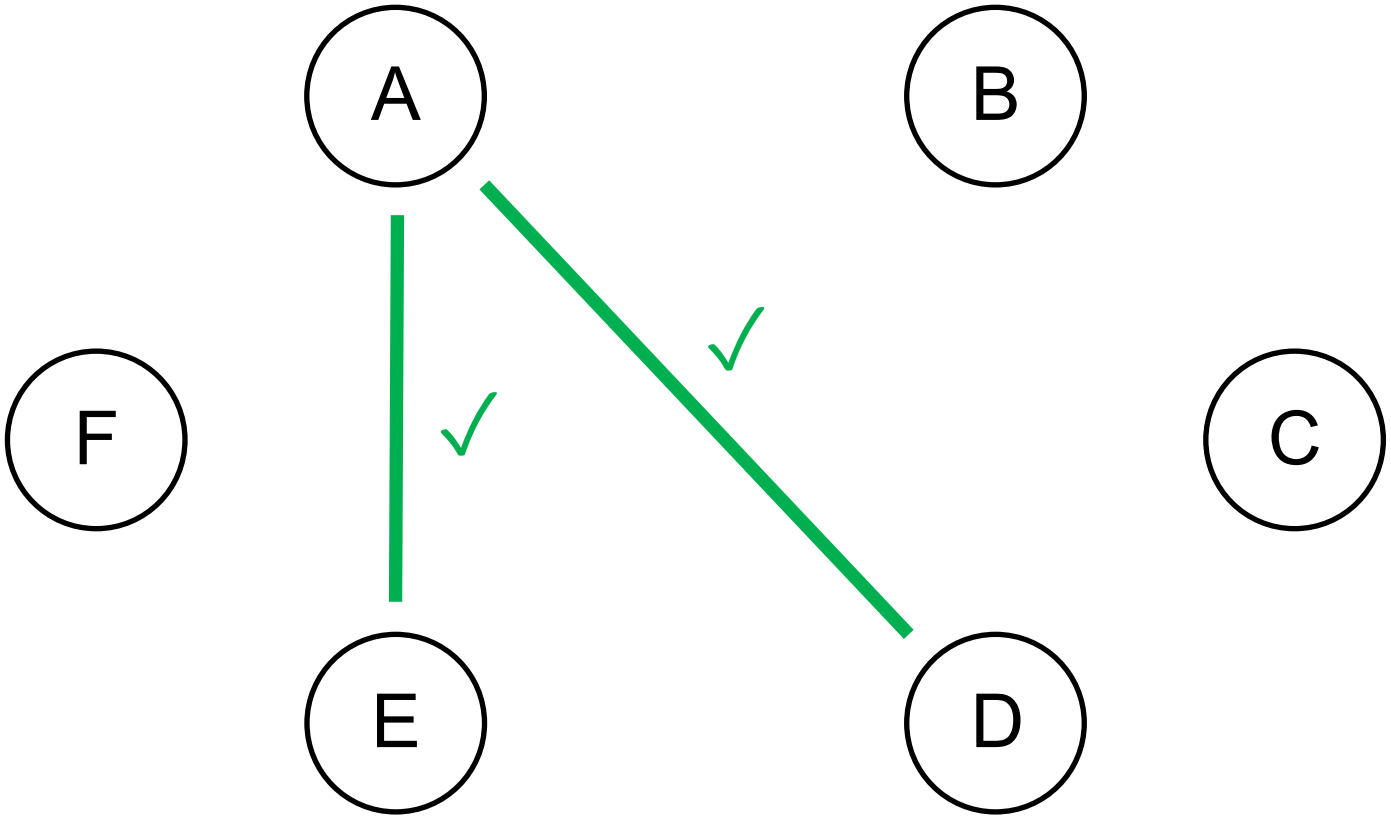
Constraint-based algos (AKA lots of little tests of association)



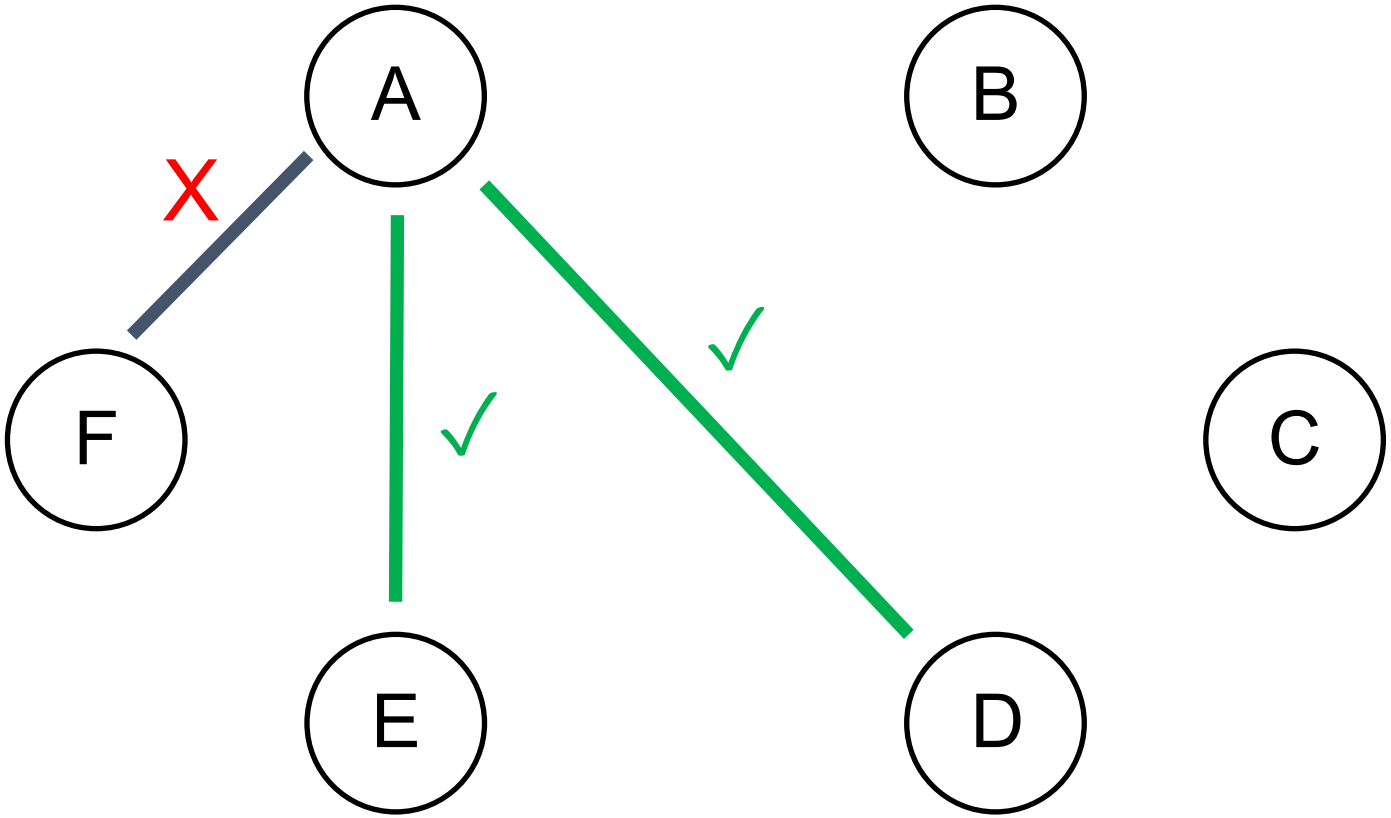
Constraint-based algos (AKA lots of little tests of association)



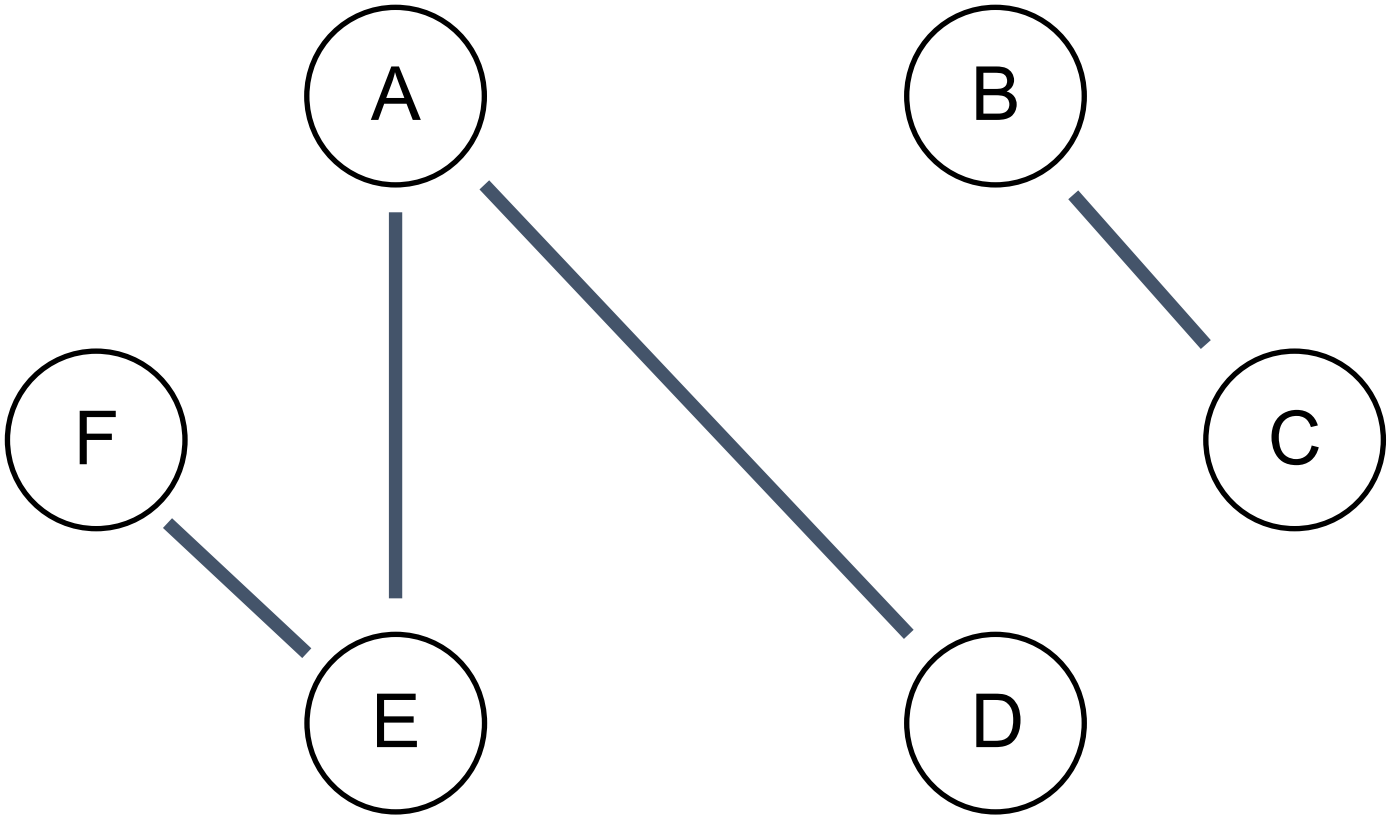
Constraint-based algos (AKA lots of little tests of association)



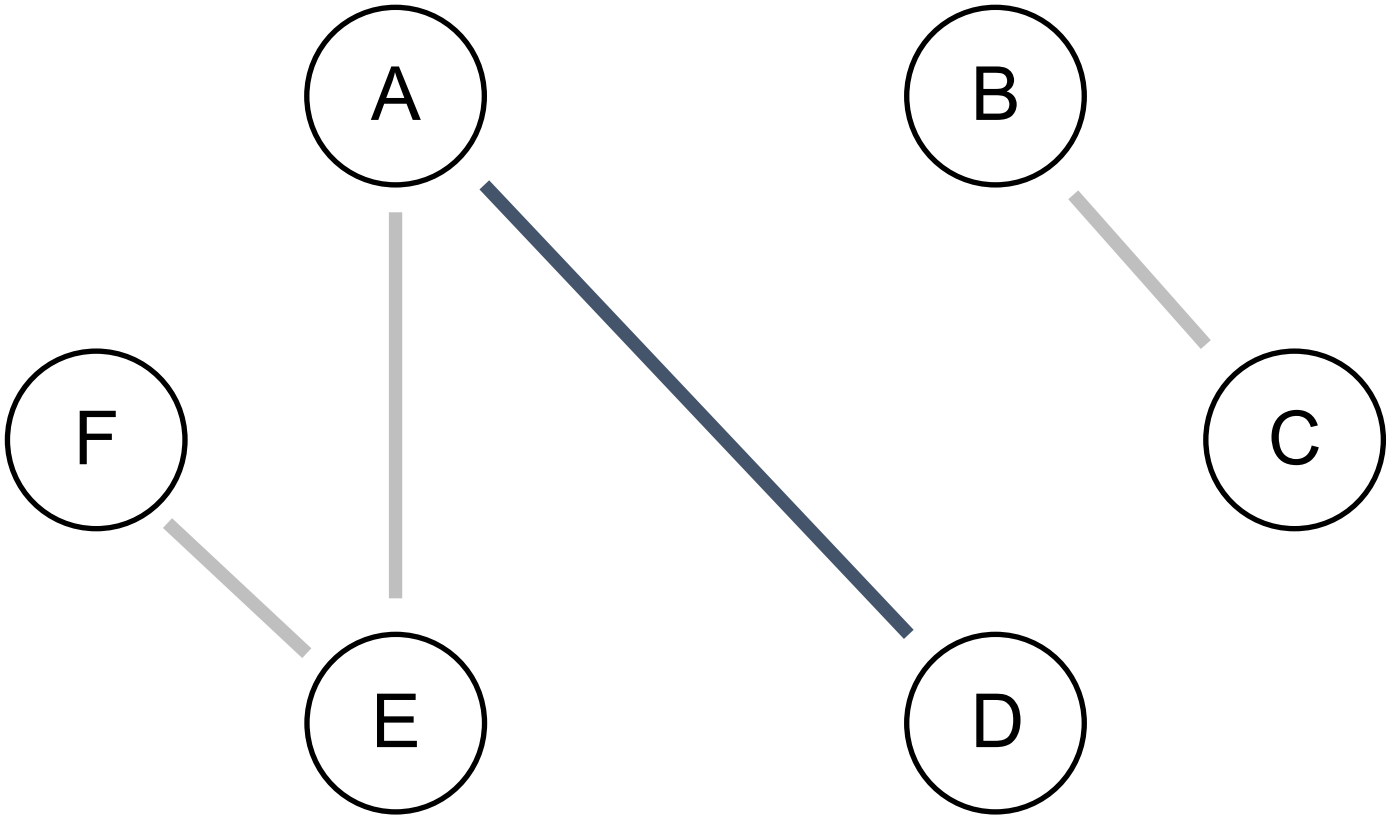
Constraint-based algos (AKA lots of little tests of association)



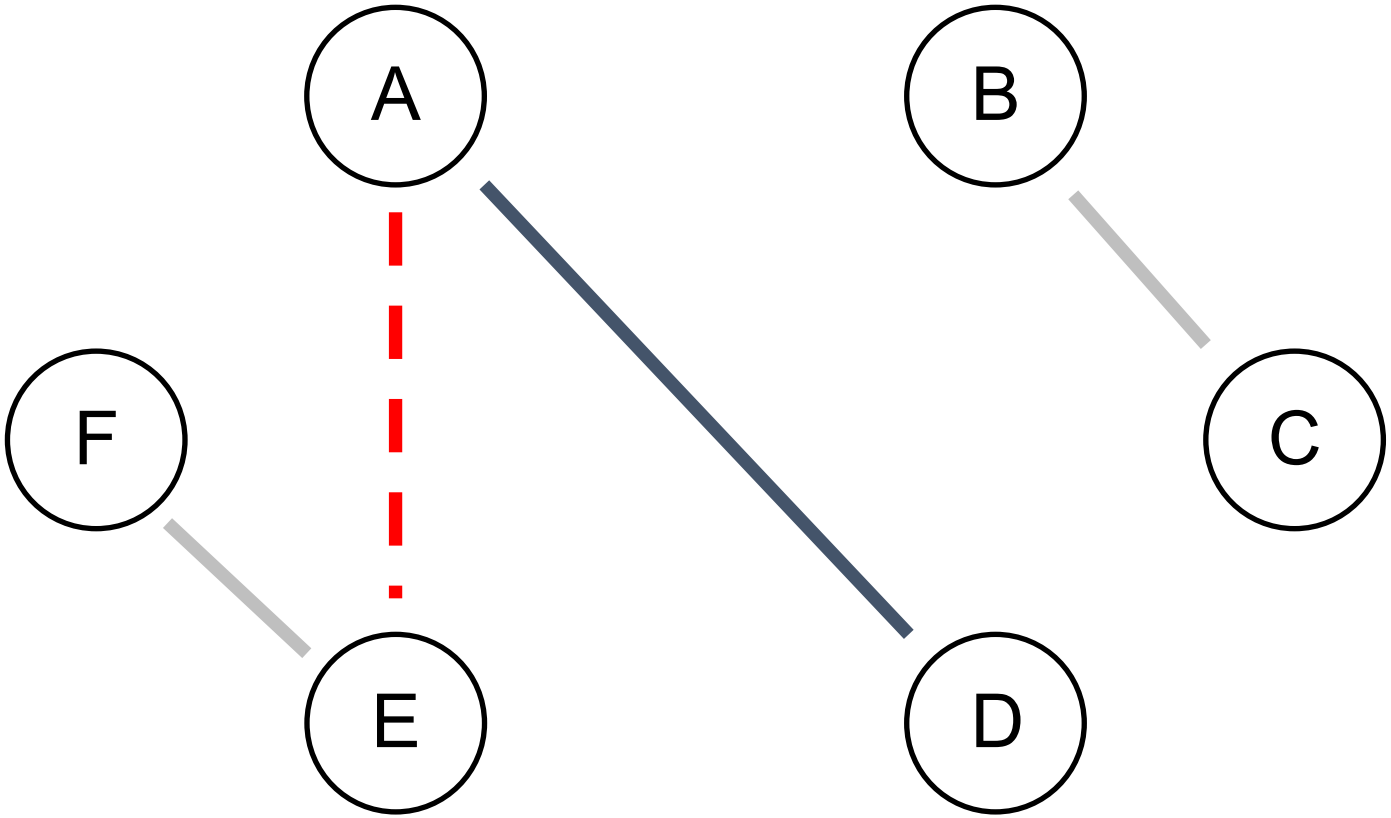
Constraint-based algos (AKA lots of little tests of association)



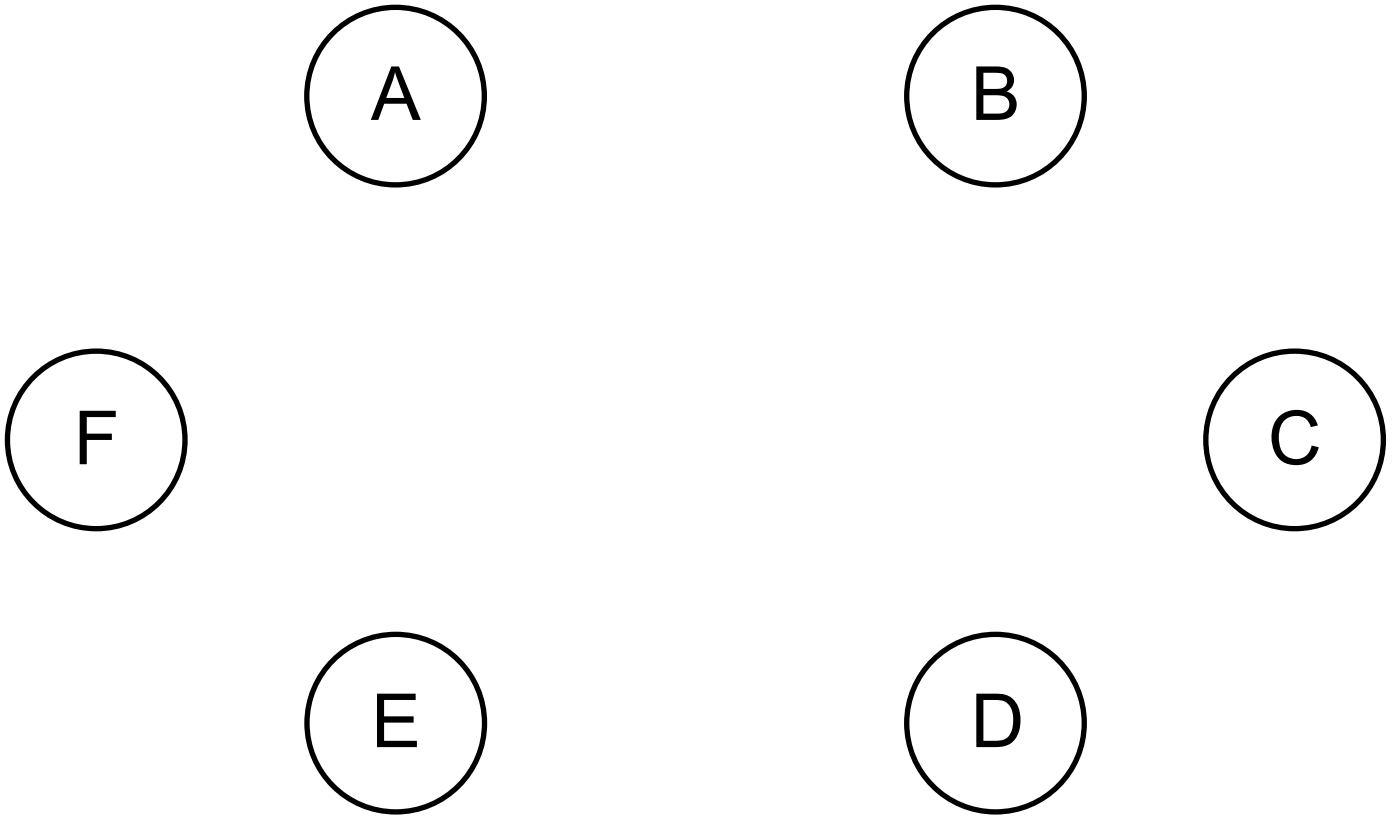
Constraint-based algos (AKA lots of little tests of association)



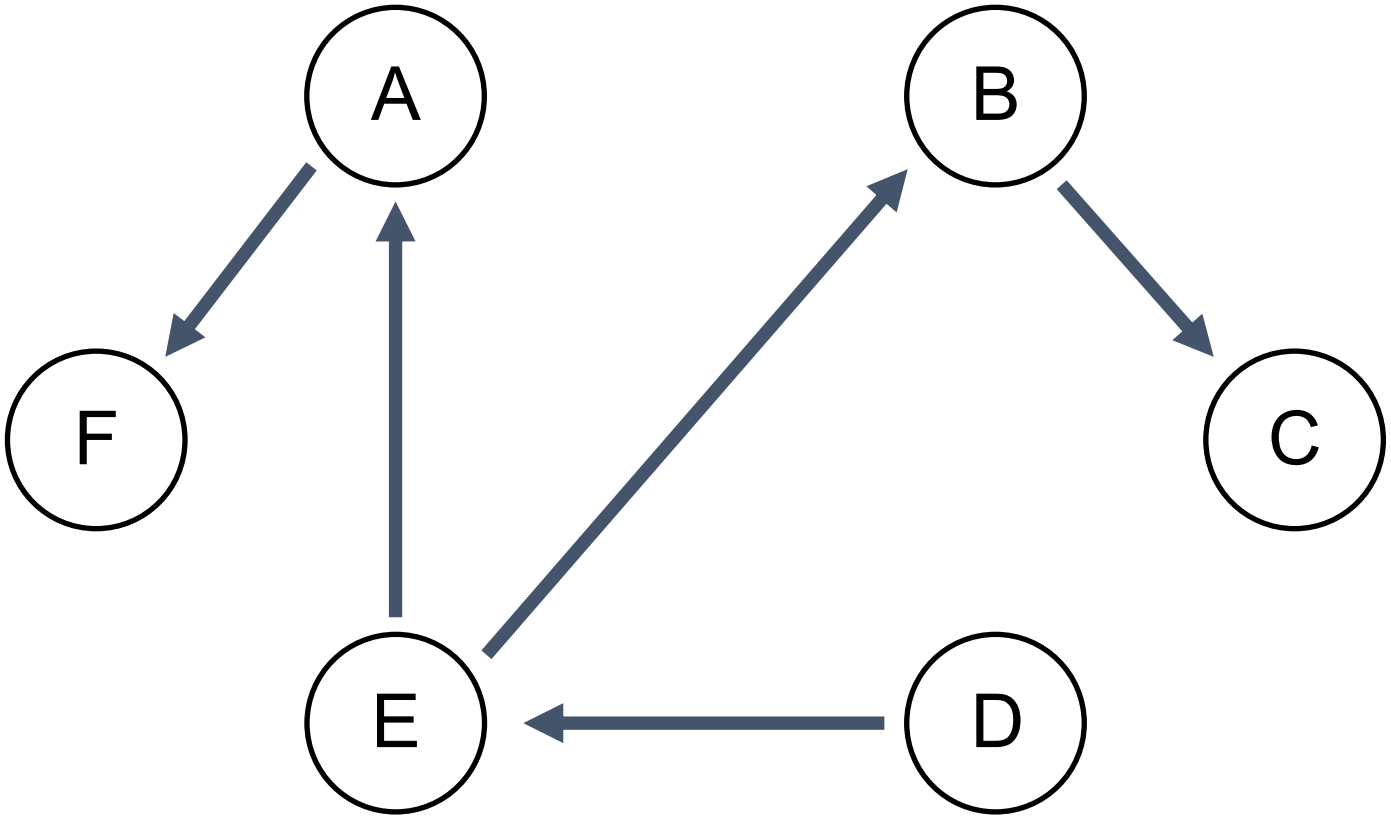
Constraint-based algos (AKA lots of little tests of association)



Score-based algos (AKA randomly tweak structure and check fit)

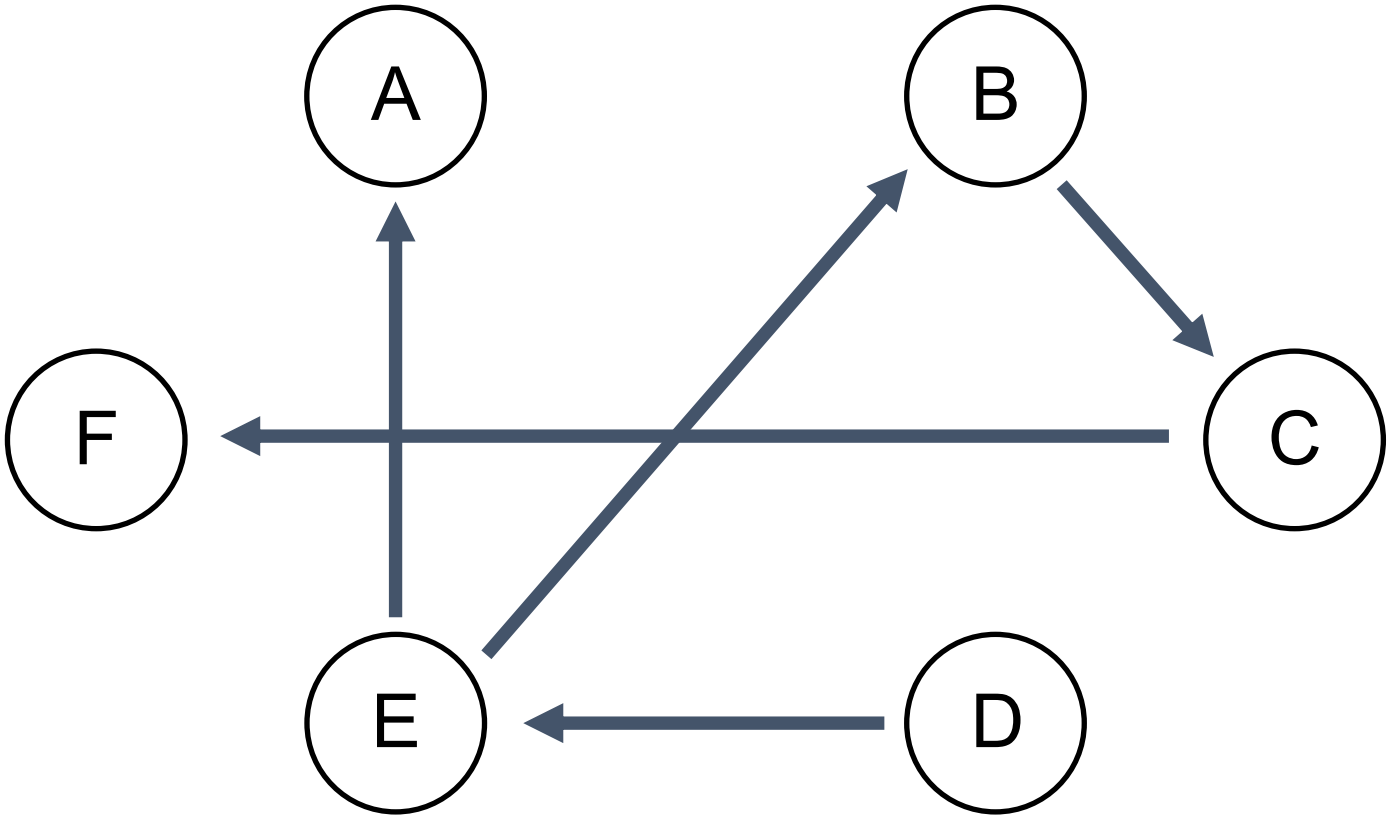


Score-based algos (AKA randomly tweak structure and check fit)



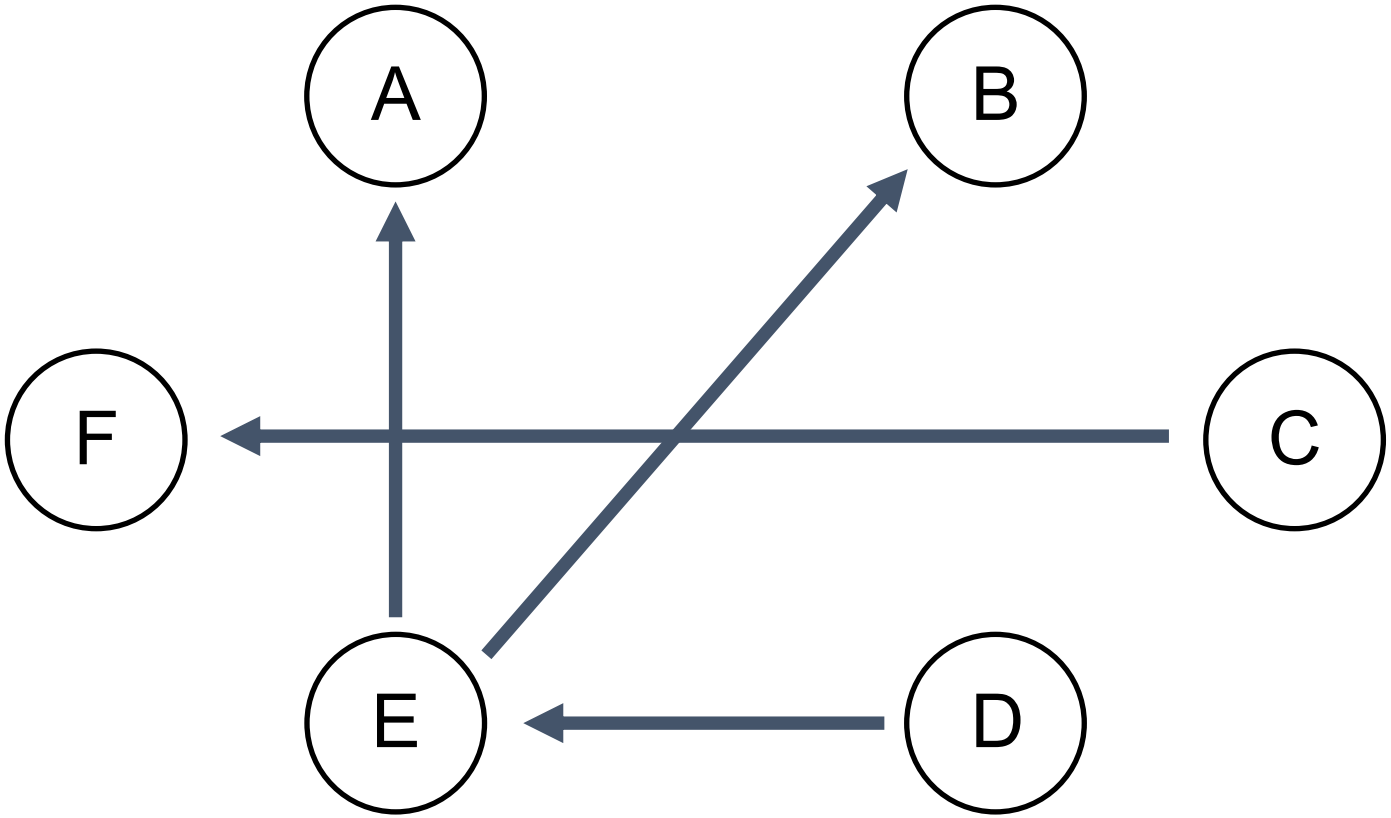
Fit score: 58

Score-based algos (AKA randomly tweak structure and check fit)



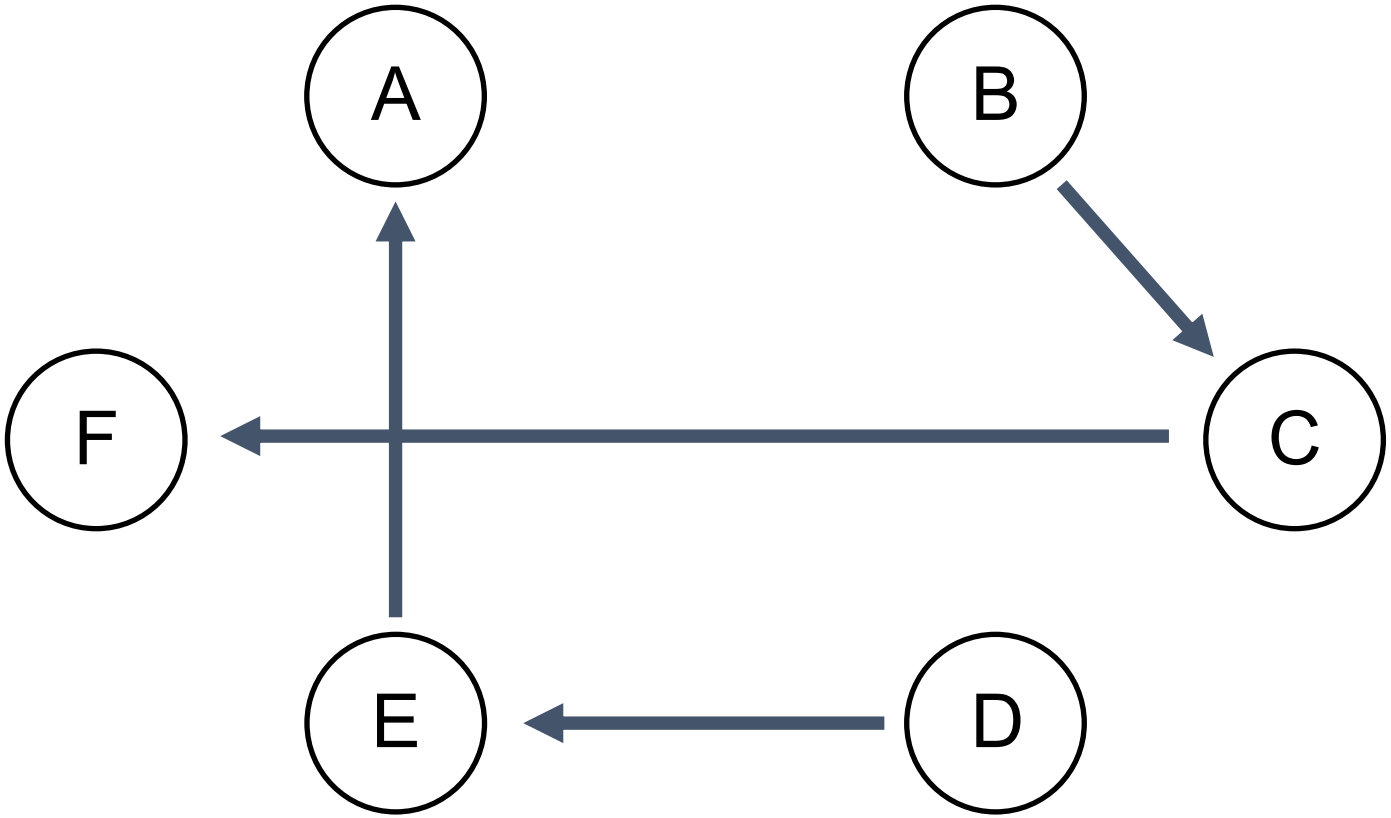
Fit score: 55

Score-based algos (AKA randomly tweak structure and check fit)



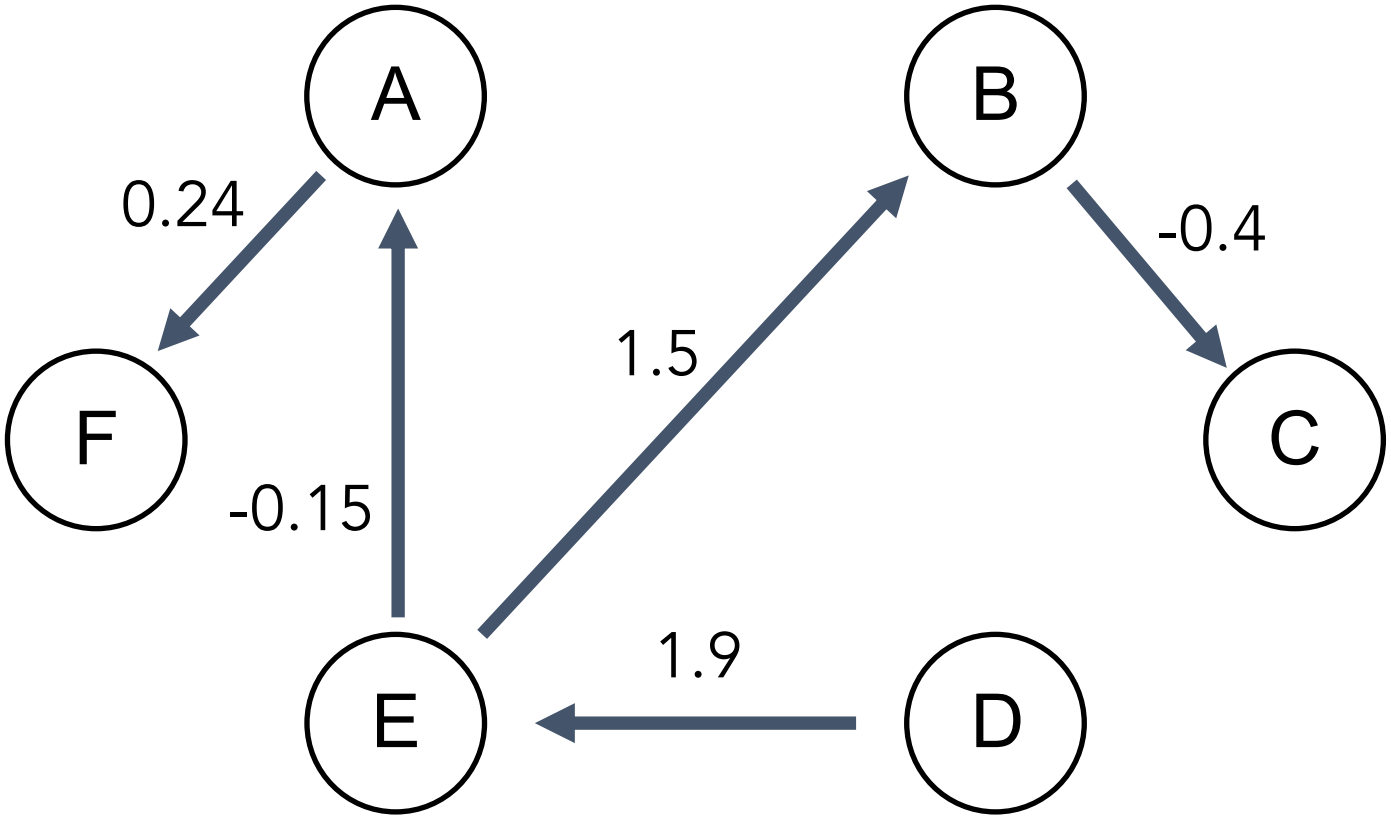
Fit score: 62

Score-based algos (AKA randomly tweak structure and check fit)



Fit score: 54

Graph Neural Network approach



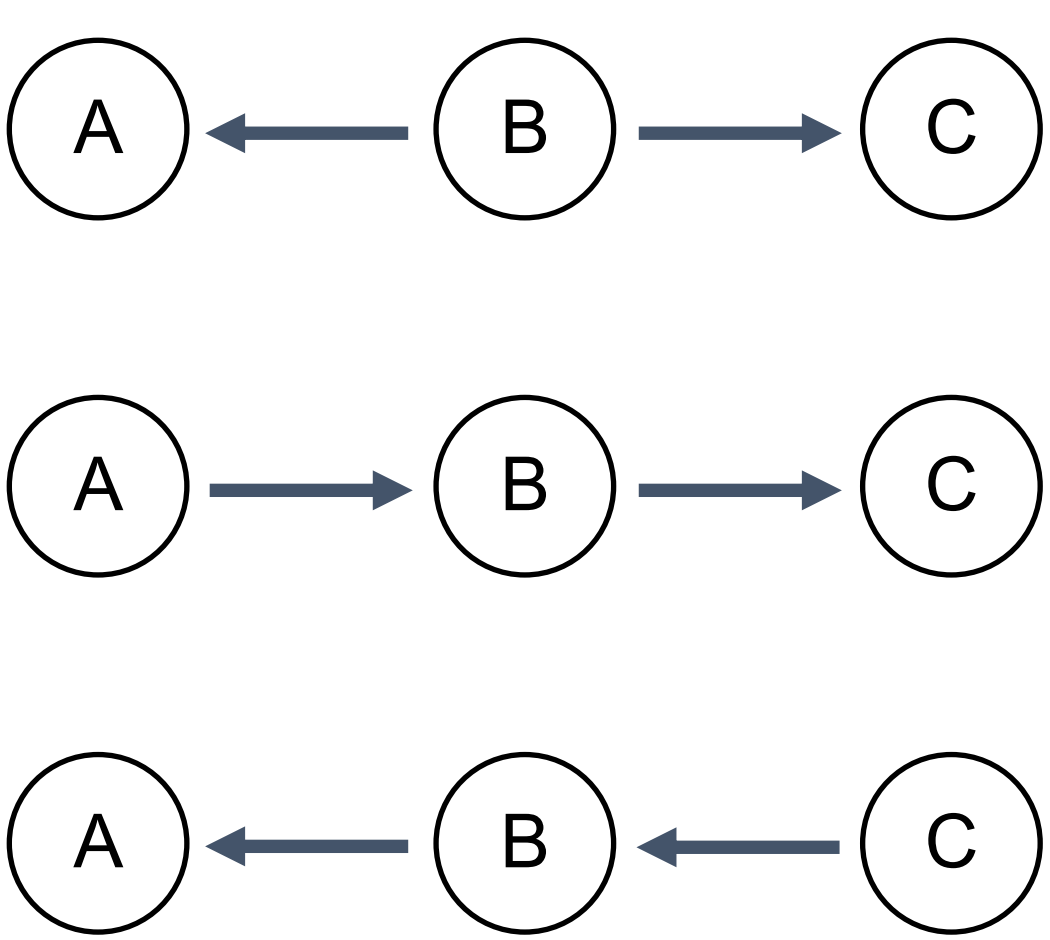
Graph Neural Network approach

Source \ Destination	A	B	C	D	E	F
A	-0.3	-2.2	0.001	1.1	0.2	0.3
B	2.5	1.3	0.002	0.01	0.4	0.35
C	0.03	0.2	0.09	-0.3	-0.3	1.04
D	0.01	0.4	0.093	-0.41	1.1	-0.2
E	0.003	0.031	-0.05	0.07	0.03	-0.1
F	2.3	0.6	-0.01	0	0.1	0.033

Graph Neural Network approach

Source \ Destination	A	B	C	D	E	F
A	0	1	0	1	0	0
B	1	0	0	0	0	0
C	0	0	0	0	0	1
D	0	0	0	0	1	0
E	0	0	0	0	0	0
F	1	1	0	0	0	0

Structure learning ain't easy



These three graphs belong to the same "Markov Equivalence Class" and are indistinguishable with observational data!

Back to the regularly scheduled program…

If you are doing causal modeling…

- First, think carefully about quantities of interest and their relationships before looking at any data - this requires domain knowledge
- Before modeling, understand bivariate relationships between independent vars, also between independent vars and dependent var
- Identify potential confounders and identify covariates not to control for…

Assumptions of causal inference

- **Temporality**. Causes always occur before effects: The treatment variable needs to occur before measured outcome. Covariates should occur before treatment (prevents you from controlling on colliders).
- **Stable Unit Treatment Value**. The treatment status of a given individual does not affect the potential outcomes of any other individuals.
- **Positivity.** For each level of each covariate in your data, there needs to be some variability of the treatment and outcome variables.
- **Ignorability.** All major confounding variables are included in your data. This is a tough one, but necessary unless you want a biased estimate of the treatment effect.

Shout out which assumptions are violated!

Example #1

I want to understand whether frequent emails to customers might impact customer satisfaction.

I have survey data with customer, self-reported satisfaction from a year ago, and I use this past month's number of emails for each customer as a proxy for how often we email them generally.

Example #2

I want to see the causal impact of a neighborhood's cleanliness on crime rates, controlling for 20 known confounders.

I pull up an academic dataset with data on 40 different neighborhoods.

Example #3

I want to see how releasing a new in-app, multiplayer game through my social media app impacts user engagement. I only want to give it to some test users initially.

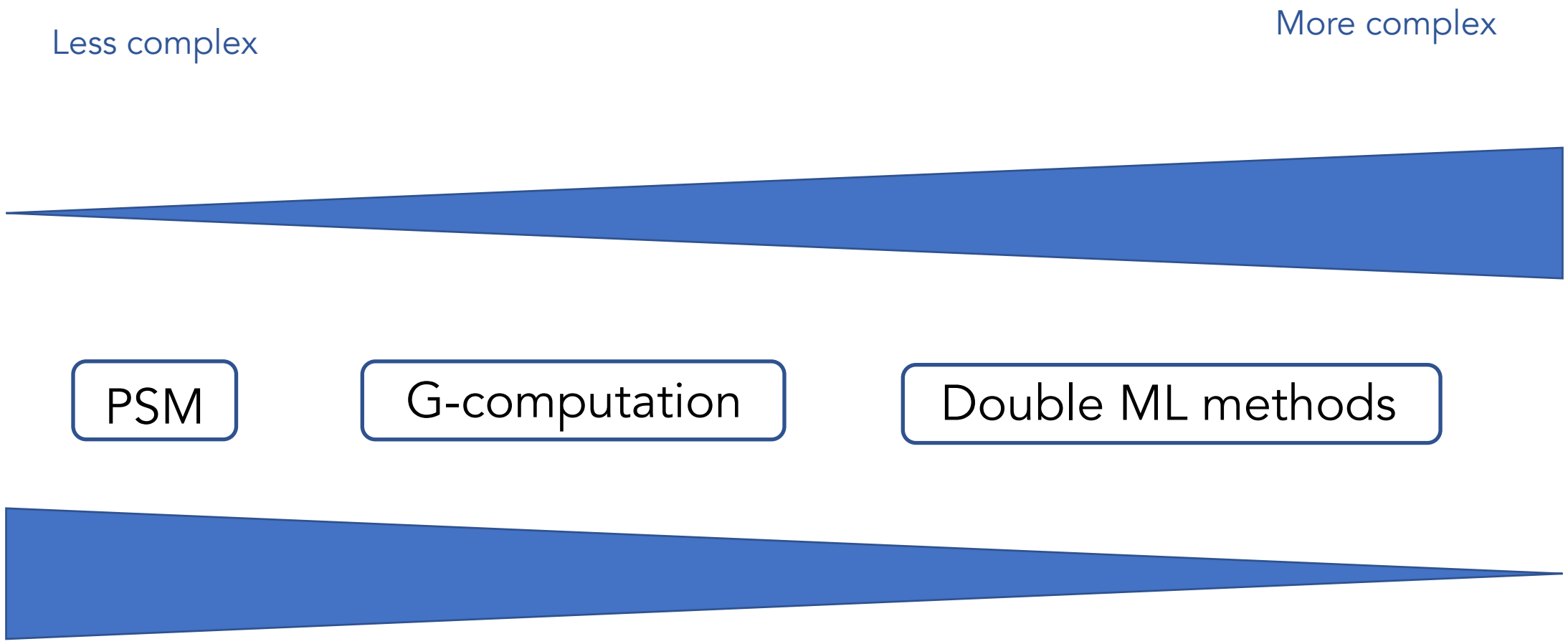
With this multiplayer game you can play with anyone who has the social media app by sending them invites. Accidentally, our test users can invite non-test users.

Example #4

We’re curious how a job training program could impact a person’s income 3 years in the future.

We don’t have lots of data so we perform a causal inference analysis only controlling for the person’s age.

The familiar modeling spectrum…



Some empirical studies have been conducted, but jury is still out. Largely depends on your audience? (PSM easier to explain)

scientific reports

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › scientific reports › articles › article

Article | Open Access | Published: 08 June 2020

G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study

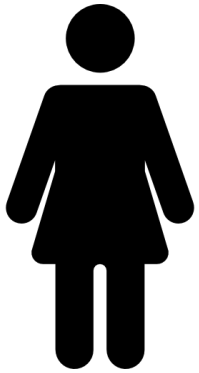
Arthur Chatton, Florent Le Borgne, Clémence Leyrat, Florence Gillaizeau, Chloé Rousseau, Laetitia Barbin, David Laplaud, Maxime Léger, Bruno Giraudeau & Yohann Foucher ✉

Scientific Reports **10**, Article number: 9219 (2020) | Cite this article

7763 Accesses | **5** Citations | **12** Altmetric | Metrics

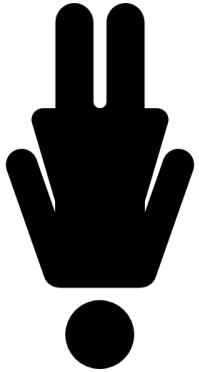
Counterfactuals (with a binary treatment)

Our Observed reality



10-year smoker

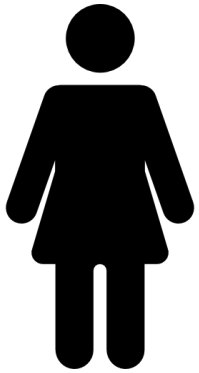
An alternative reality



Never smokes

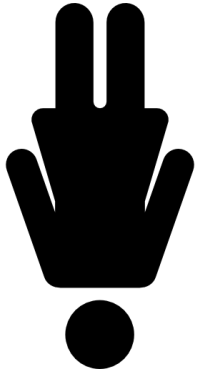
Counterfactuals (with a binary treatment)

Our Observed reality



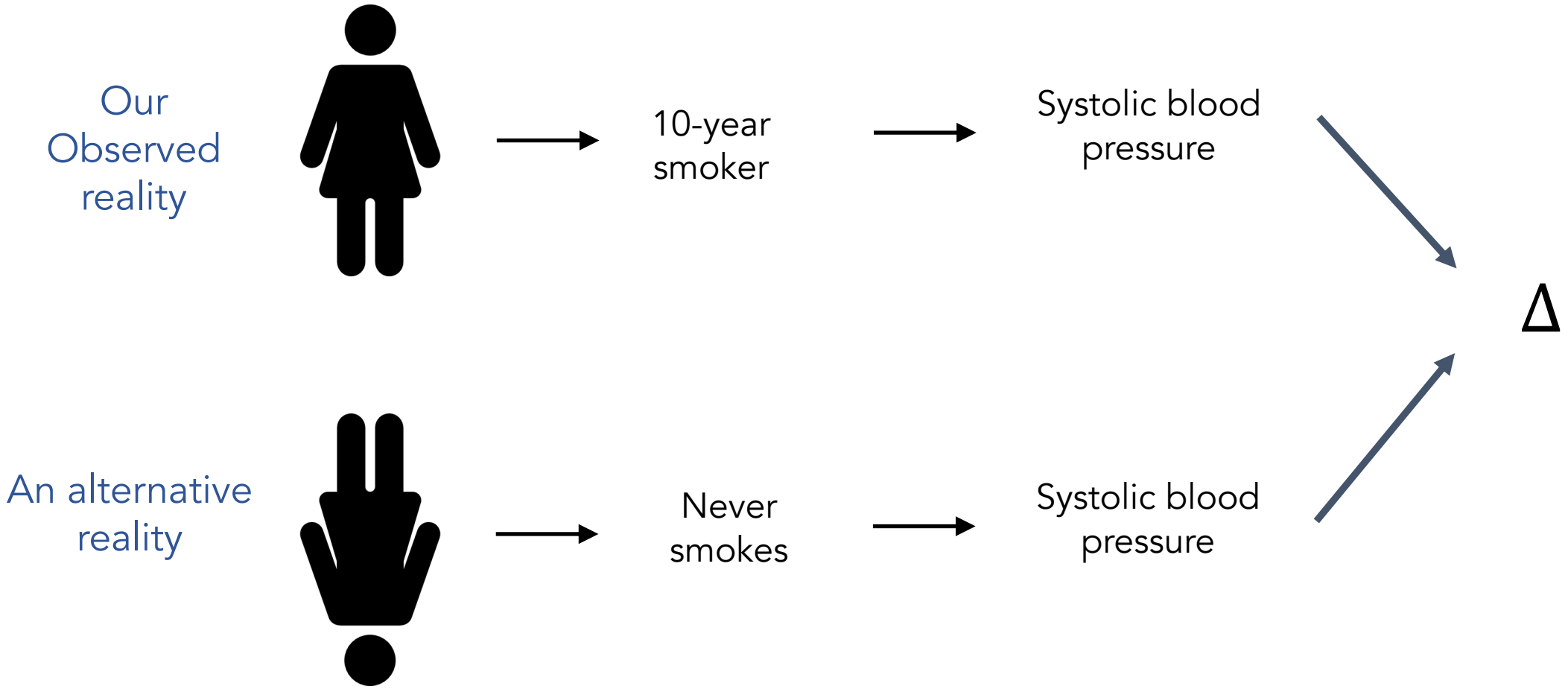
10-year smoker → Systolic blood pressure

An alternative reality

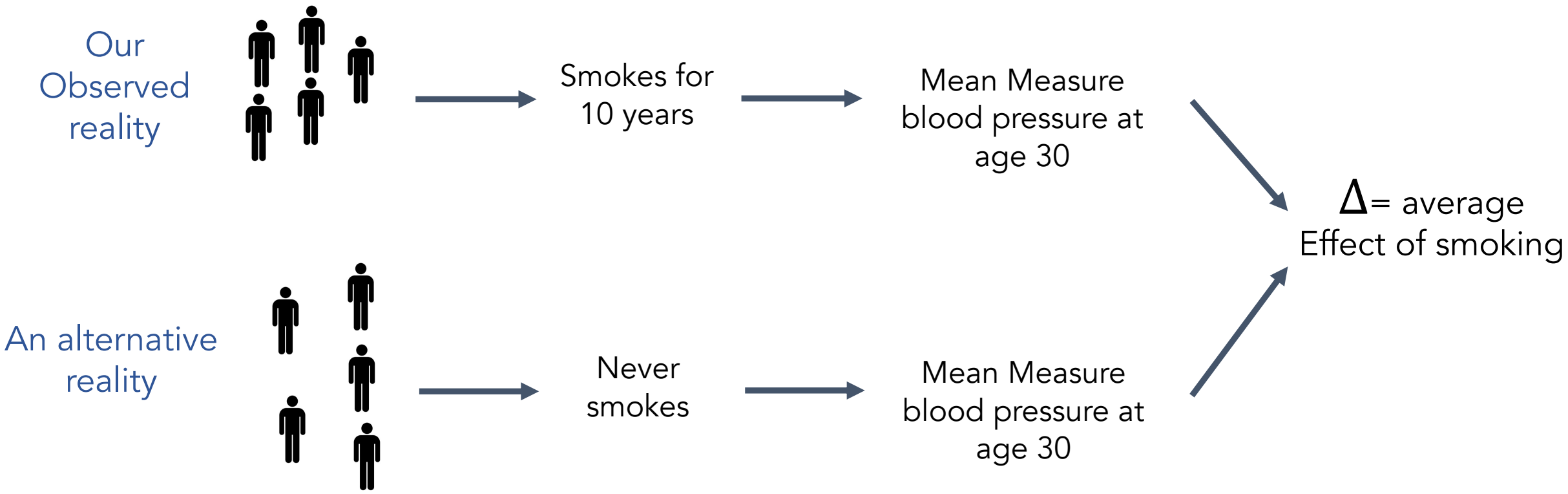


Never smokes → Systolic blood pressure

Counterfactuals (with a binary treatment)



Counterfactuals (with a binary treatment)

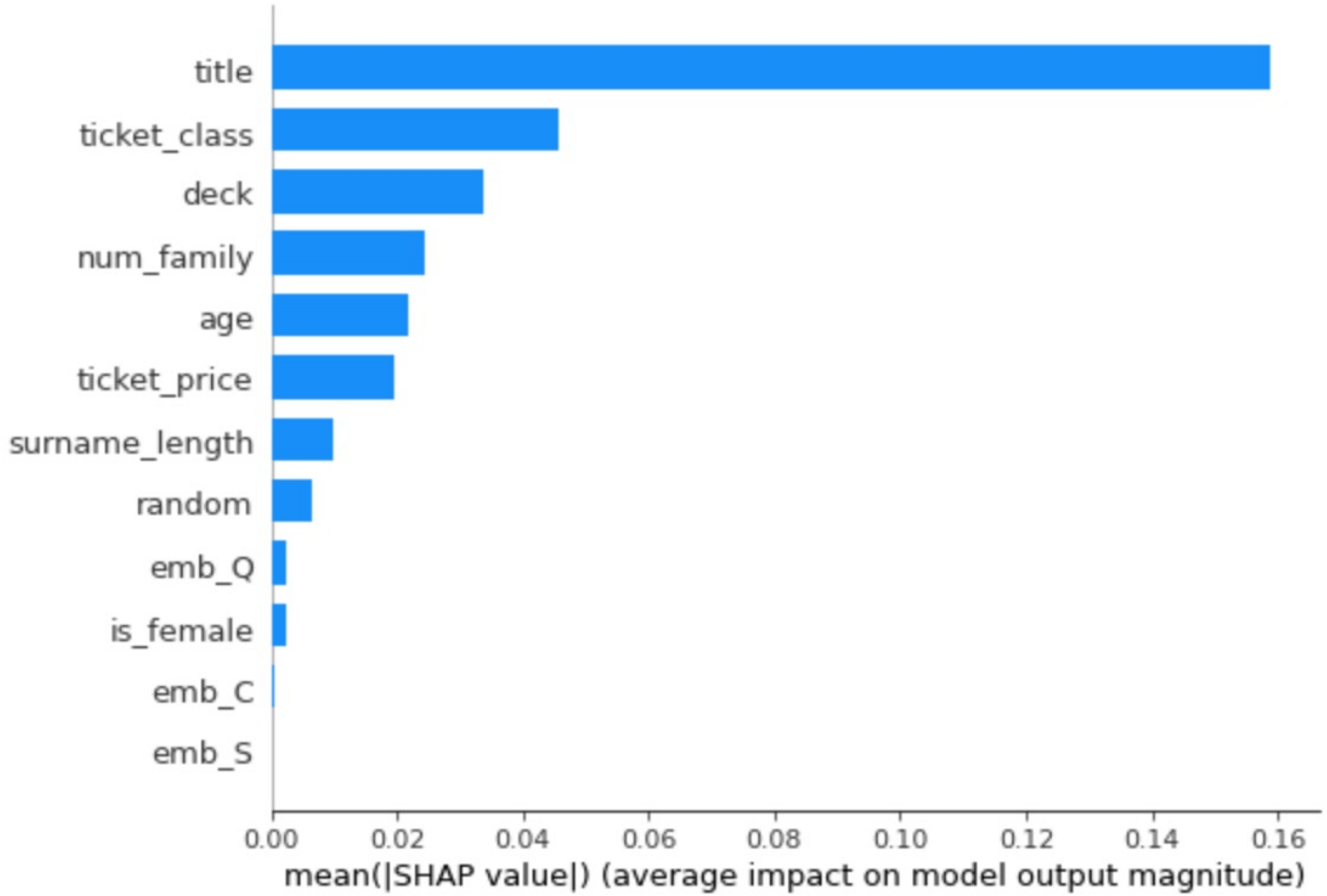


Causal inference metrics

- Average treatment effect (ATE)
- Average treatment effect among treated (ATT)
- Average treatment effect among untreated (ATU)

nota bene!

- Traditional variable importance methods don't tell you anything about causality!



Propensity score matching (PS)

1) Start with a set of participants for whom we have complete treatment, outcome, and covariate data

ID#	Covar 1	Covar 2	treat	outcome
1	…	…	1	20
2	…	…	1	15
3	…	…	0	10
4	…	…	0	10
5	…	…	1	20

2) for all participants, calculate probability of them receiving treatment, based on covariate data (a propensity score)

ID#	Covar 1	Covar 2	treat	ps	outcome
1	…	…	1	0.65	20
2	…	…	1	0.33	15
3	…	…	0	0.64	10
4	…	…	0	0.33	10
5	…	…	1	0.97	20

3) Take sub-sample of treated participants and match to sub-sample of control participants, based on similar ps values

ID#	Covar 1	Covar 2	treat	ps	outcome
1	…	…	1	0.65	20
3	…	…	0	0.64	10

ID#	Covar 1	Covar 2	treat	ps	outcome
2	…	…	1	0.33	15
4	…	…	0	0.33	10

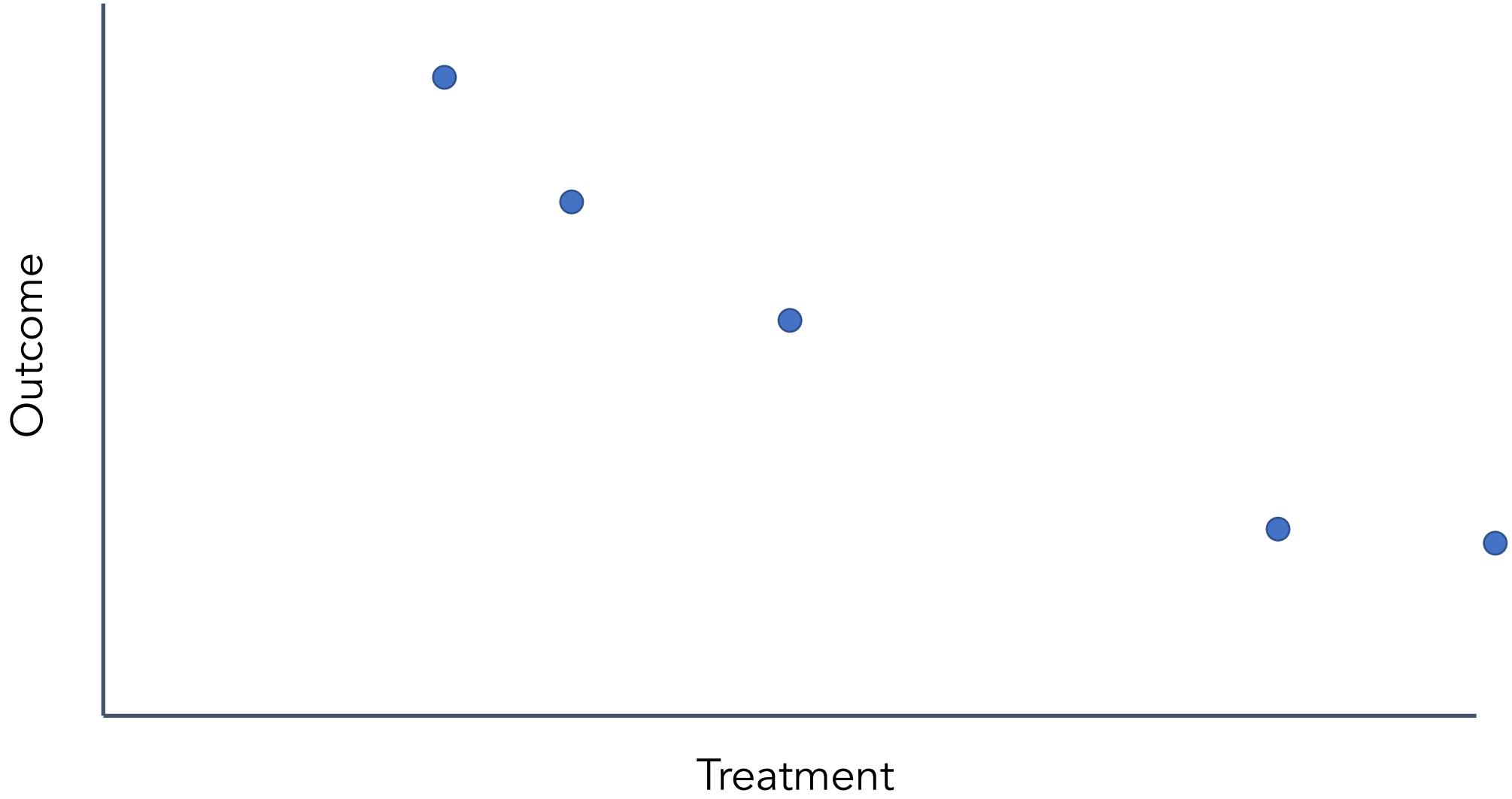
4) Calculate treatment effect on these two sub-samples using standard approaches

ID#	Covar 1	Covar 2	treat	ps	outcome
1	…	…	1	0.65	20
2	…	…	1	0.33	15
3	…	…	0	0.64	10
4	…	…	0	0.33	10

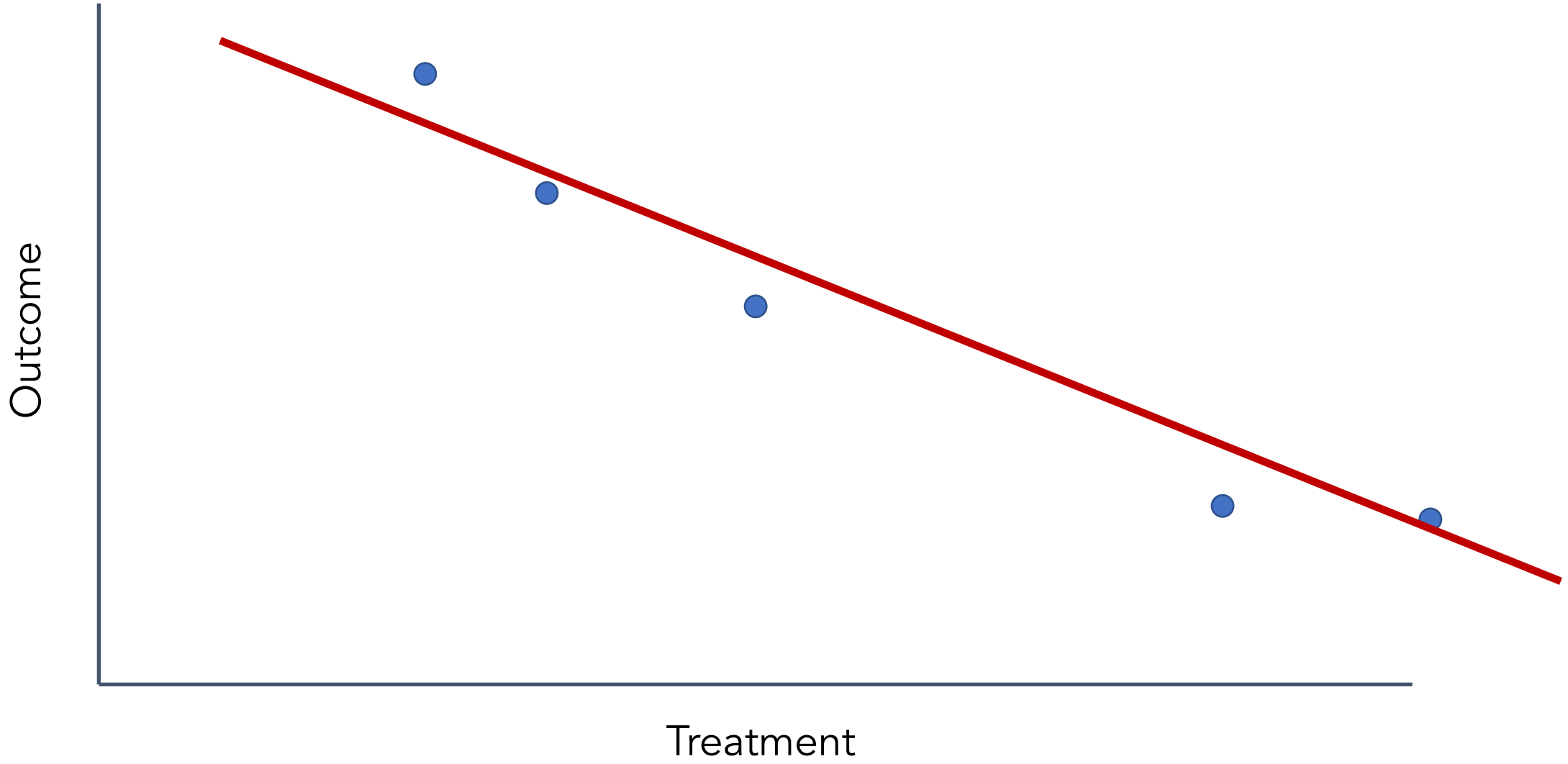
$$ATT = \bar{x}_{treated} - \bar{x}_{untreated} = \frac{(20+15)}{2} - \frac{(10+10)}{2} = 7.5$$

Causal dose-response curve estimation (AKA estimating the causal curve)

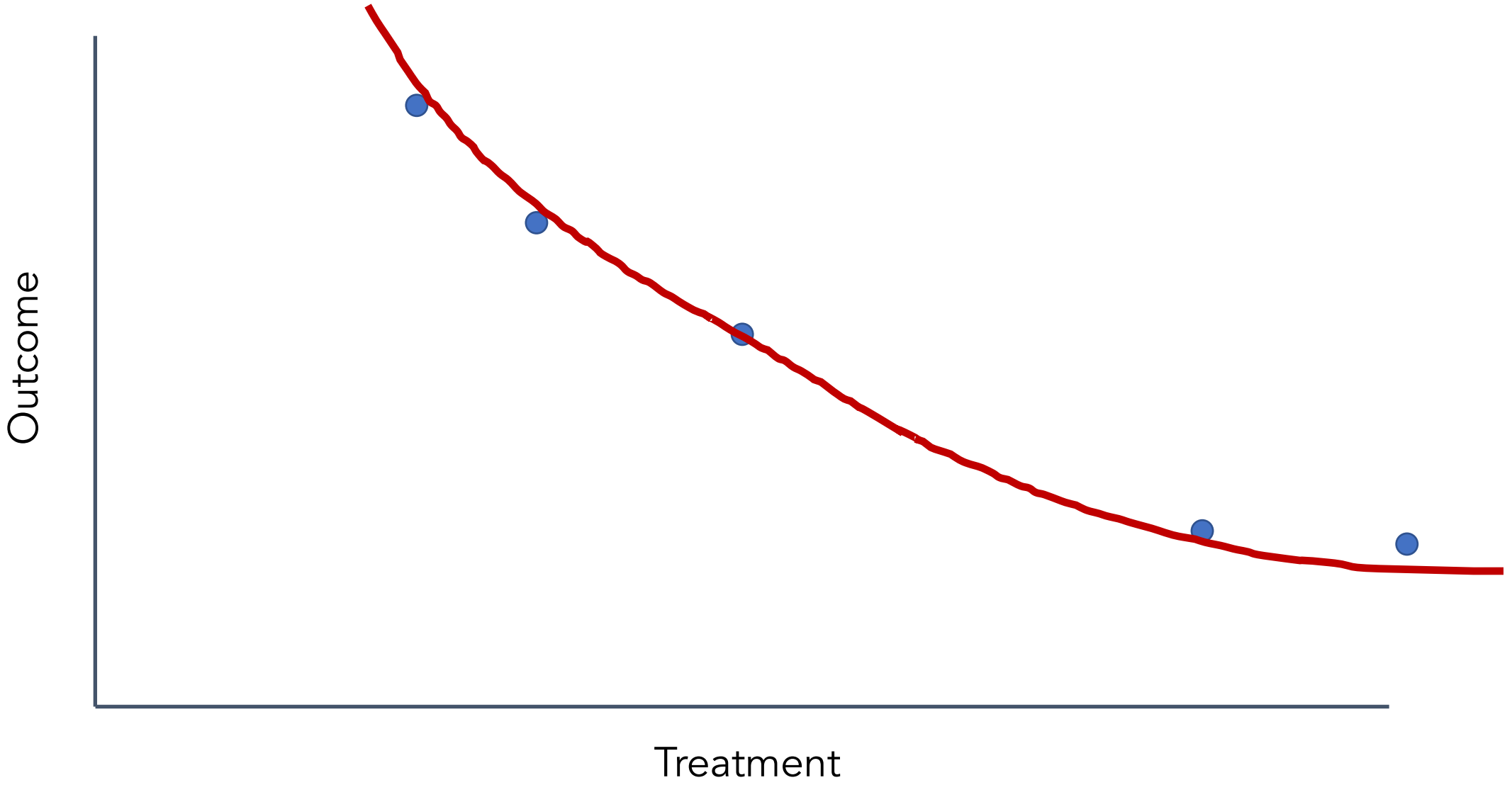
Counterfactuals (with a continuous treatment)



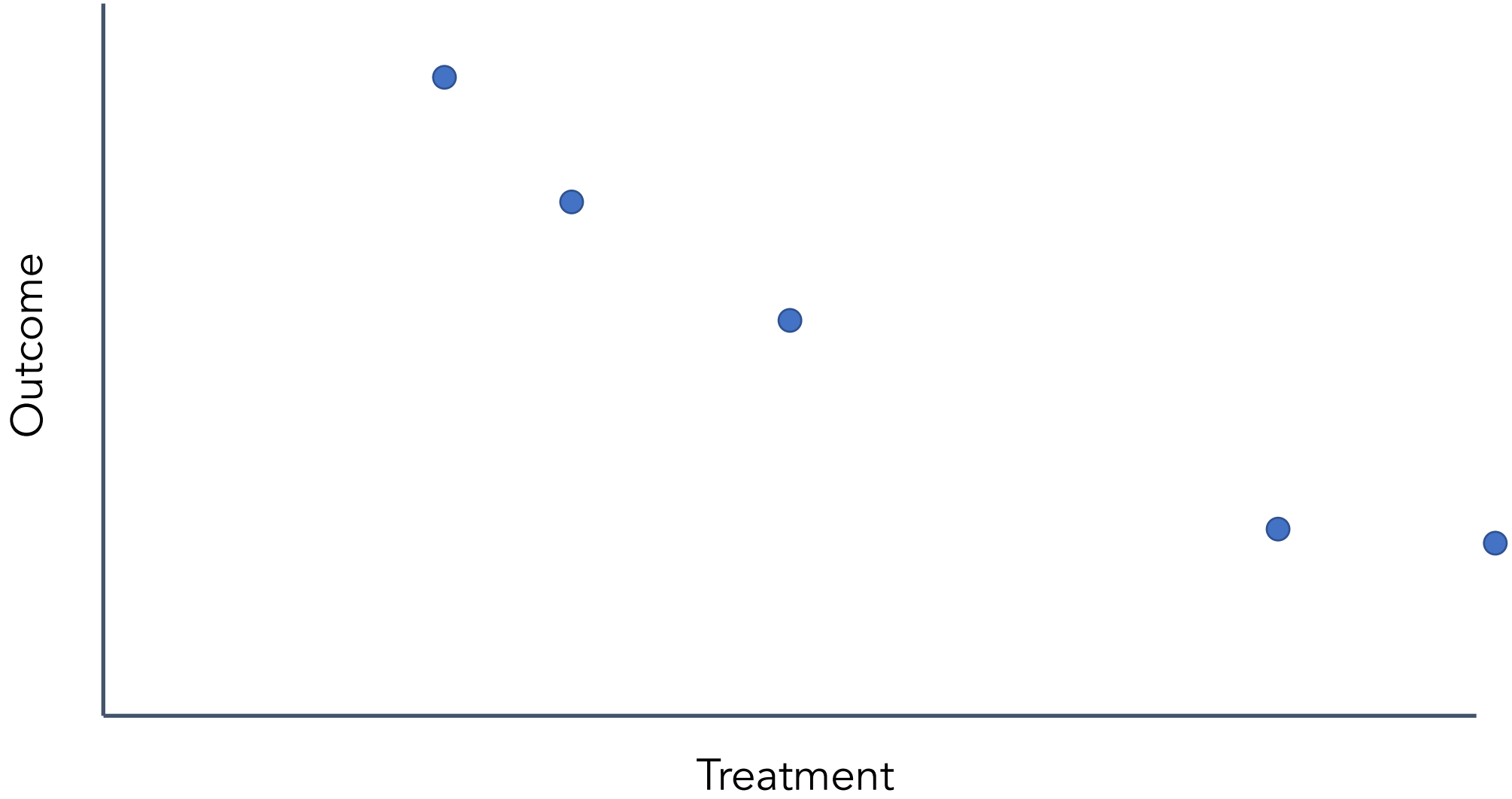
Counterfactuals (with a continuous treatment)



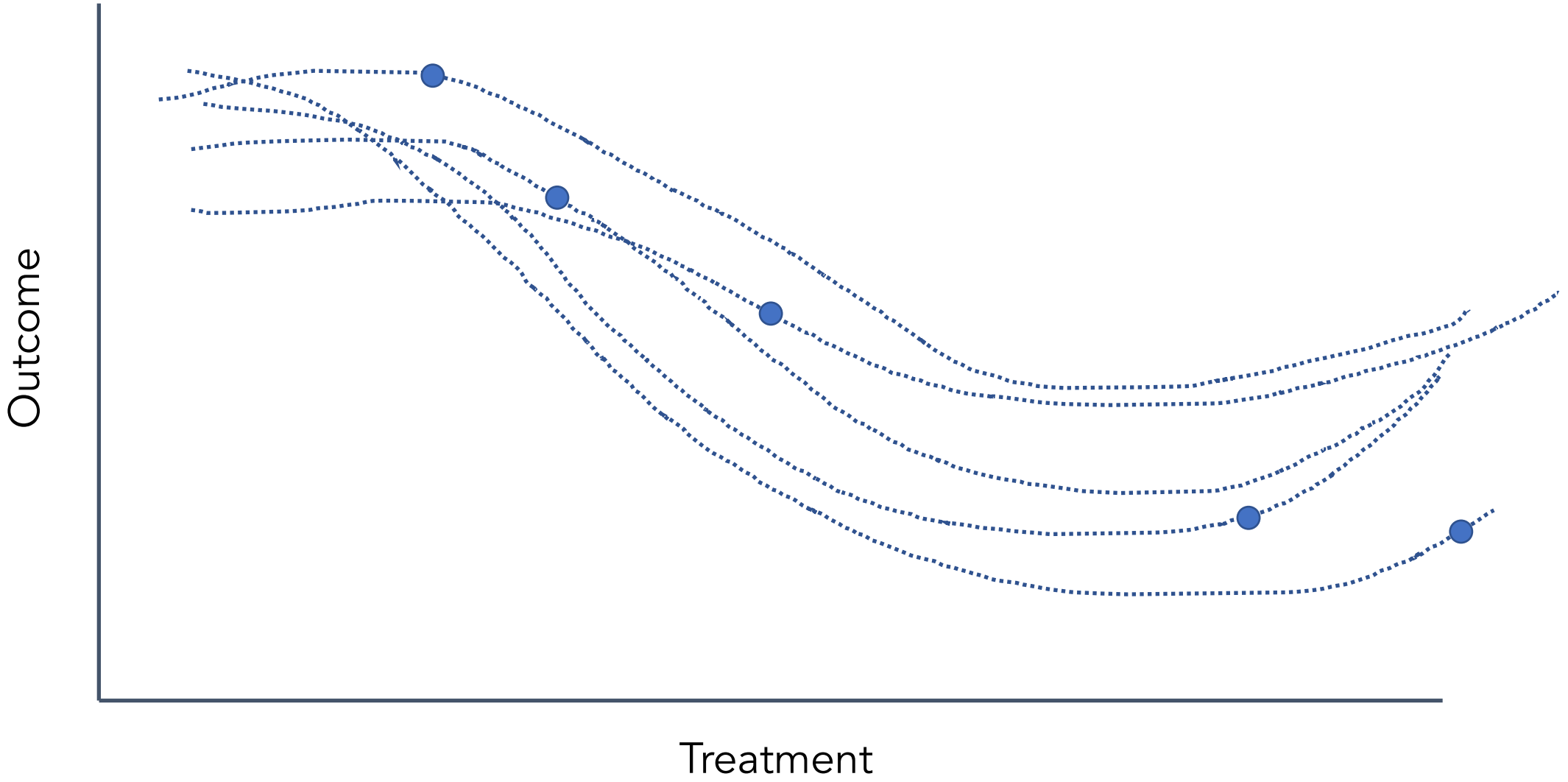
Counterfactuals (with a continuous treatment)



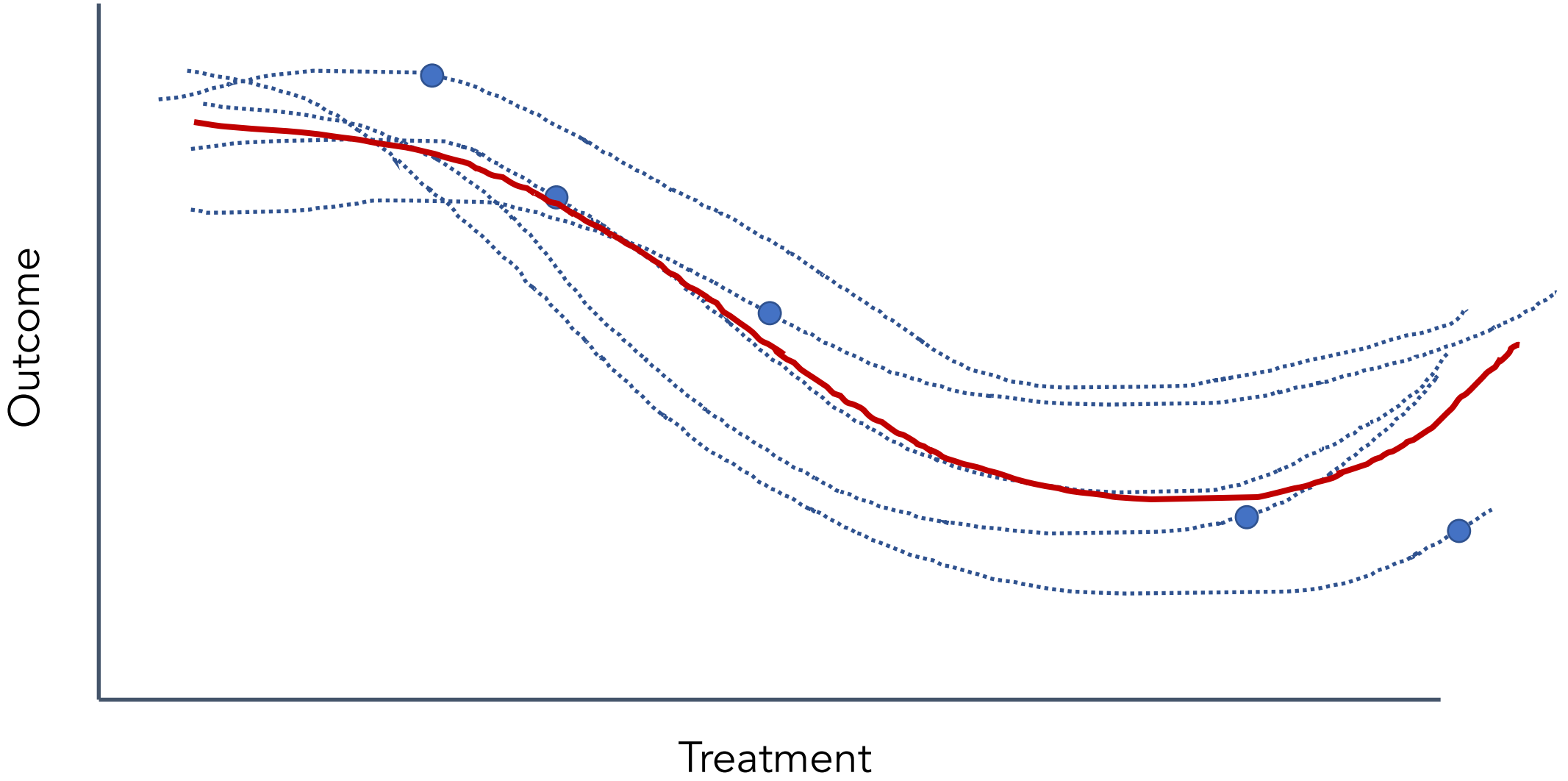
Counterfactuals (with a continuous treatment)



Counterfactuals (with a continuous treatment)



Counterfactuals (with a continuous treatment)



Estimating the “causal curve”

GPS is an extension of the standard propensity score method. It is the treatment assignment density calculated at a particular treatment value

1) Calculate the GPS associated with each treatment value observation
2) Fit a curve of treatment values predicting outcome values, adjusted for the GPS
3) The resulting treatment against outcome curve is your causal dose response curve (AKA your causal curve)

G-computation

1) Start with a set of participants for whom we have complete treatment, outcome, and covariate data

ID#	Covar 1	Covar 2	treat	outcome
1	…	…	1	20
2	…	…	1	15
3	…	…	0	10
4	…	…	0	10
5	…	…	1	20

2) train a model that predicts the outcome from all covariates and treatment variable. Aim for high recall and precision.

ID#	Covar 1	Covar 2	treat	outcome
1	…	…	1	20
2	…	…	1	15
3	…	…	0	10
4	…	…	0	10
5	…	…	1	20

3) “force” every observation in the dataset to receive the treatment

ID#	Covar 1	Covar 2	treat	outcome
1	…	…	1	20
2	…	…	1	15
3	…	…	1	10
4	…	…	1	10
5	…	…	1	20

4) Predict outcome values with these covariate and treatment values

ID#	Covar 1	Covar 2	treat	outcome	$\hat{O}_{treat}$
1	…	…	1	20	22.5
2	…	…	1	15	16.0
3	…	…	1	10	14.0
4	…	…	1	10	17.0
5	…	…	1	20	22.5

5) Now "force" every observation to not receive treatment, And make outcome predictions again

ID#	Covar 1	Covar 2	treat	outcome	$\hat{O}_{treat}$	$\hat{O}_{untreat}$
1	…	…	0	20	22.5	18.5
2	…	…	0	15	16.0	14.0
3	…	…	0	10	14.0	11.5
4	…	…	0	10	17.0	13.0
5	…	…	0	20	22.5	19.5

6) Calculate the average difference between treated and untreated outcome estimates

ID#	$\hat{O}_{treat}$	$\hat{O}_{untreat}$	Δ
1	22.5	18.5	4.0
2	16.0	14.0	2.0
3	14.0	11.5	2.5
4	17.0	13.0	4.0
5	22.5	19.5	3.0

$$\mu_{\Delta} = 3.1$$

Double ML
And
Targeted maximum likelihood estimation (TMLE)

The double modeling approaches

- Using any machine learning method (or an ensemble of them), make an initial model to predict the outcome given the treatment and covariates: Y ~ A + W1+ W2 + Wi. With this, you are able to calculate an initial, crude estimate of the treatment effect by artificially setting A = 1 for all observations, and then by setting A = 0 for all individuals, and observing the difference in Y.
- Using the same or different machine learning method, make a model to predict treatment assignment using the covariates: A ~ W1+ W2 + Wi. For each individual, estimate the probability of A = 1, given the covariates, and the probability of A = 0, given the covariates
- Use the model and probabilities from step 2, update the initial model in step 1 in a "targeting step". In step 2 we exploit information about the relationship between the treatment and covariates to reduce bias of the estimate from step 1.

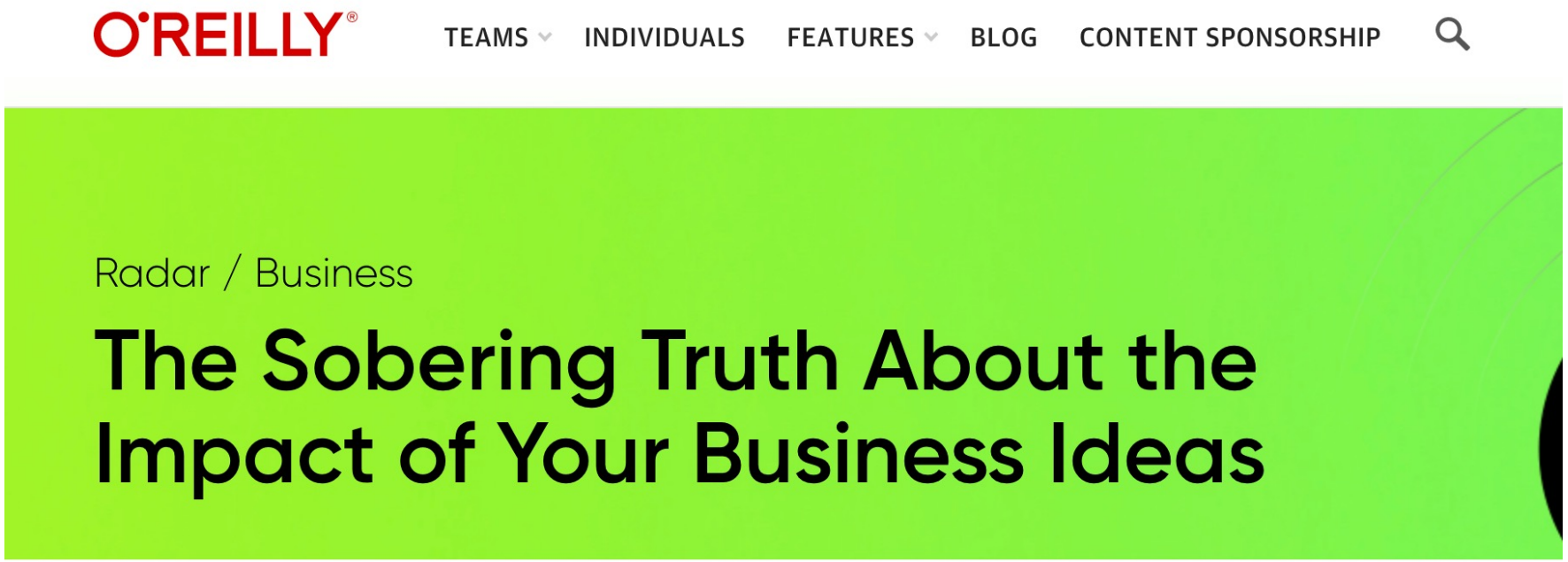
The double modeling approaches

- Pros
  - Double-robust
  - Unlike other approaches, Can handle as many covariates as needed (as long as there are enough observations)
  - Can handle more complex causal inference scenarios (e.g. control for time-varying confounding in panel data)
  - Doesn’t require bootstrapping to estimate confidence bounds
- Cons
  - Slower than other methods (depending on what ML you use)
  - Least explainable relative to other methods

Closing thoughts: troubleshooting

- Understanding the data-generating process is often way more valuable than employing an algo
- There is value in trying multiple techniques to understand their range of estimates (use p-value correction if you're running lots of analyses)
- You'll never be able to capture all confounders, but aim to capture the major ones
- If your results don't make sense, you're probably missing a big source of bias
- Causal inference is still second to proper experiments. Approach all results with healthy skepticism

Closing thoughts: be humble, it's likely your research or business idea doesn't work at all!



By Eric Colson, Daragh Sibley and Dave Spiegel

October 26, 2021