Microsoft

Tutorial @ **KDD 2018**
http://causalinference.gitlab.io/kdd-tutorial/

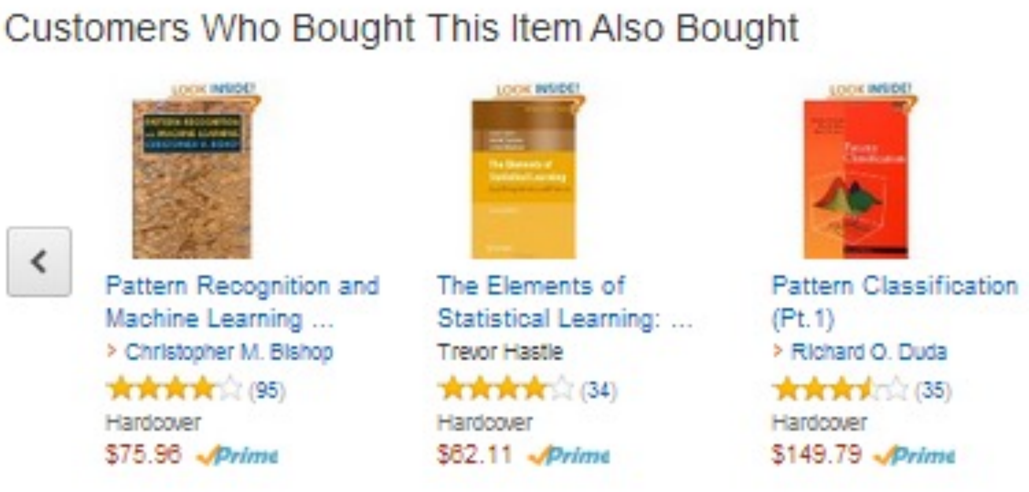
Causal Inference and Counterfactual Reasoning

Emre Kıcıman and Amit Sharma

emrek@microsoft.com, amshar@microsoft.com

Causal Inference and Counterfactual Reasoning at Microsoft Research

Predictive systems are impacting our lives

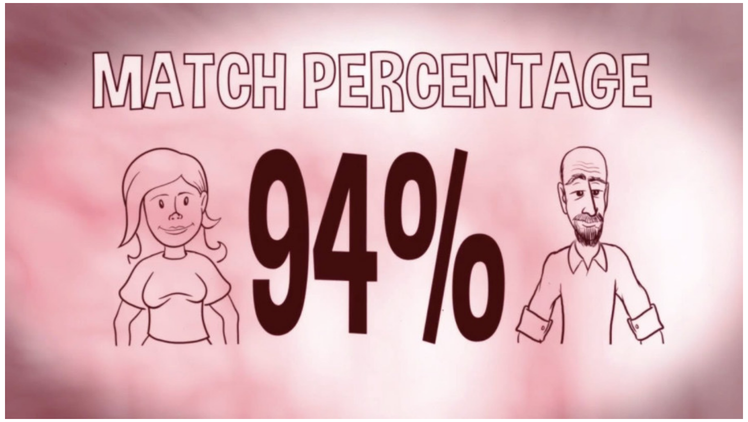


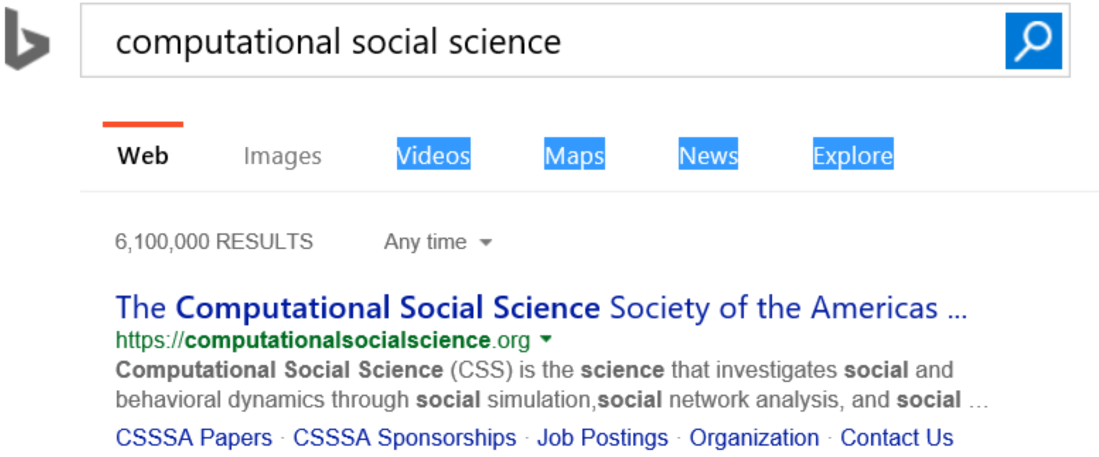
Customers Who Bought This Item Also Bought

Pattern Recognition and Machine Learning ...
Christopher M. Bishop
(95)
Hardcover
$75.96 Prime

The Elements of Statistical Learning: ...
Trevor Hastie
(34)
Hardcover
$62.11 Prime

Pattern Classification (Pt.1)
Richard O. Duda
(35)
Hardcover
$149.79 Prime





computational social science

6,100,000 RESULTS

The Computational Social Science Society of the Americas ...
https://computationalsocialscience.org
Computational Social Science (CSS) is the science that investigates social and behavioral dynamics through social simulation,social network analysis, and social ...
CSSSA Papers · CSSSA Sponsorships · Job Postings · Organization · Contact Us

Why should we care about causality?

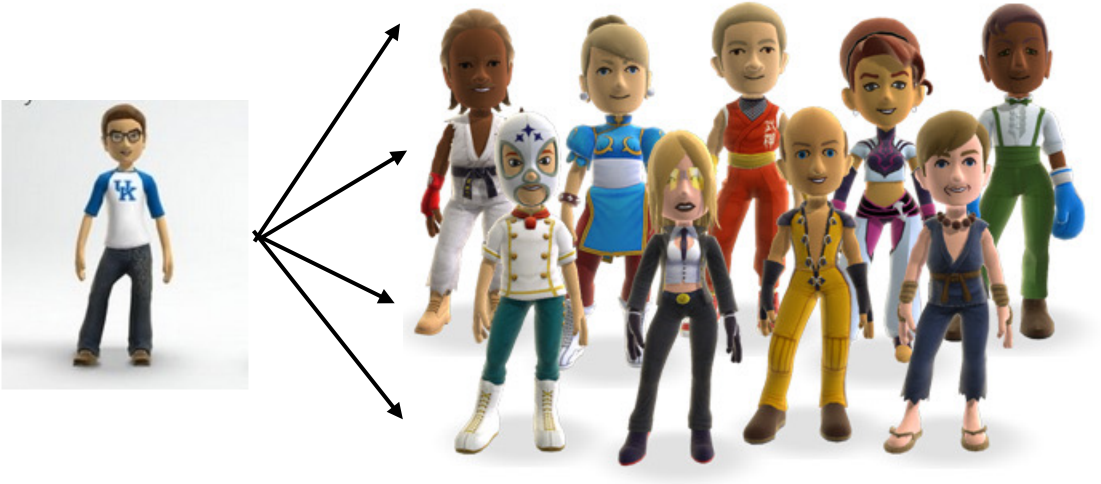
We have increasing amounts of data and highly accurate predictions.

How is causal inference useful?

1) Do prediction models guide decision-making?

From data to prediction

Can we predict a user's future activity based on exposure to their social feed?



Use the social feed to predict a user's future activity.

- Future Activity -> *f*(items in social feed) + ϵ

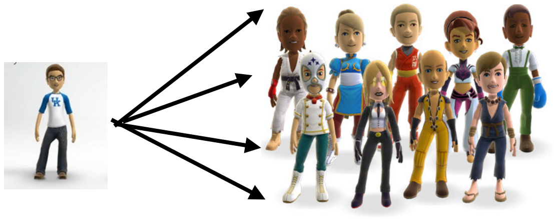
Highly predictive model.

Does it mean that feeds are influencing us significantly?

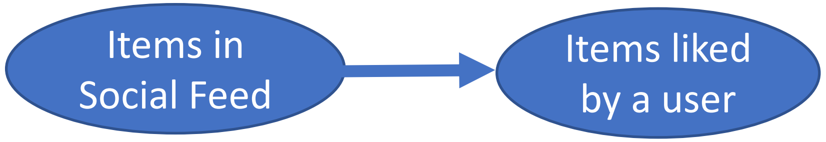
From prediction to decision-making

Would changing what people see in the feed affect what a user likes?

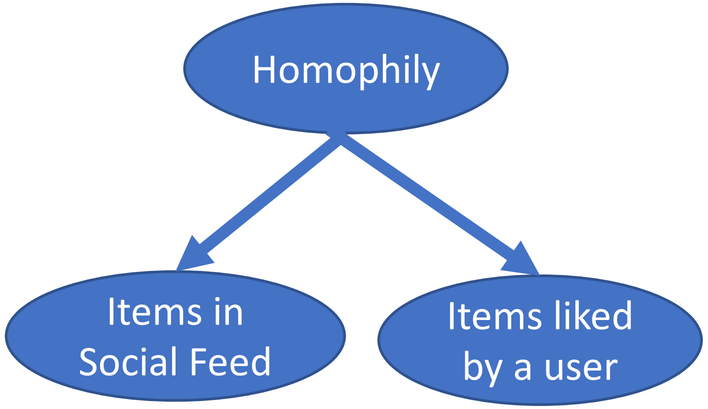
Maybe, maybe not (!)



Predictability due to
feed influence



Predictability due to
homophily



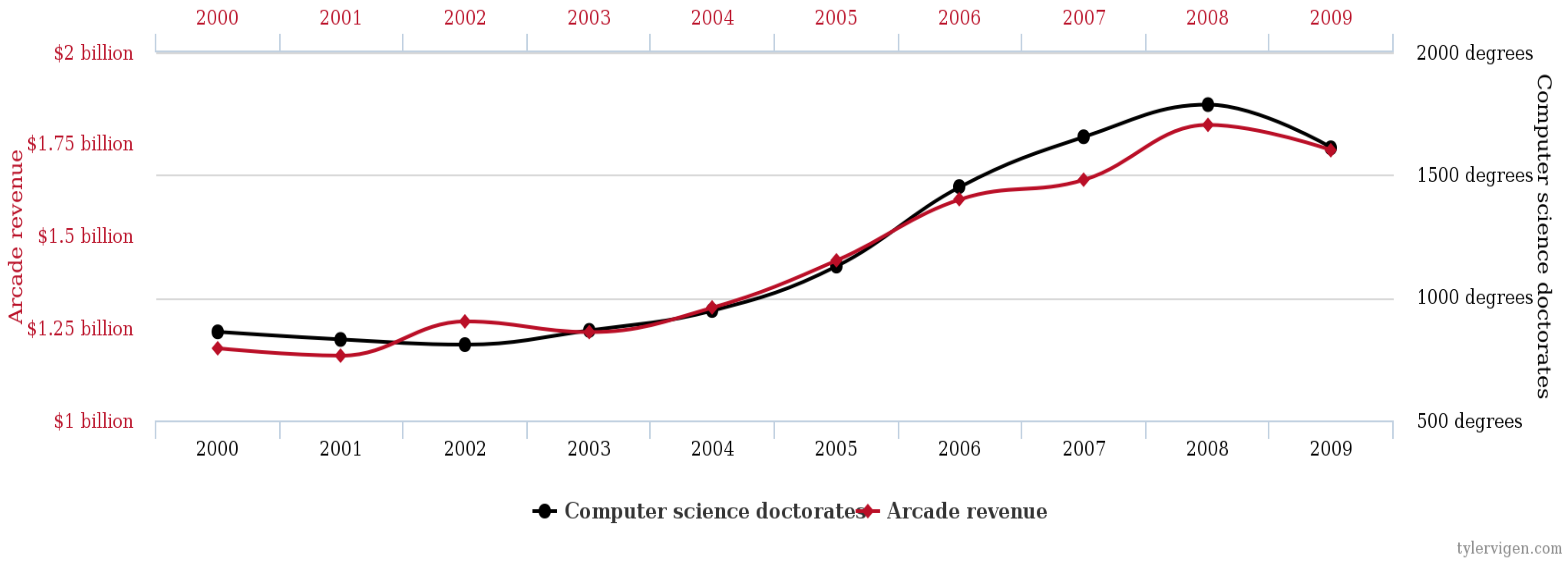
Friends’ activity can predict a person’s activity with high accuracy.
But that tells us *nothing* about the effect of the social feed.

2) Will the predictions be robust tomorrow, or in new contexts?

Total revenue generated by arcades

correlates with

Computer science doctorates awarded in the US



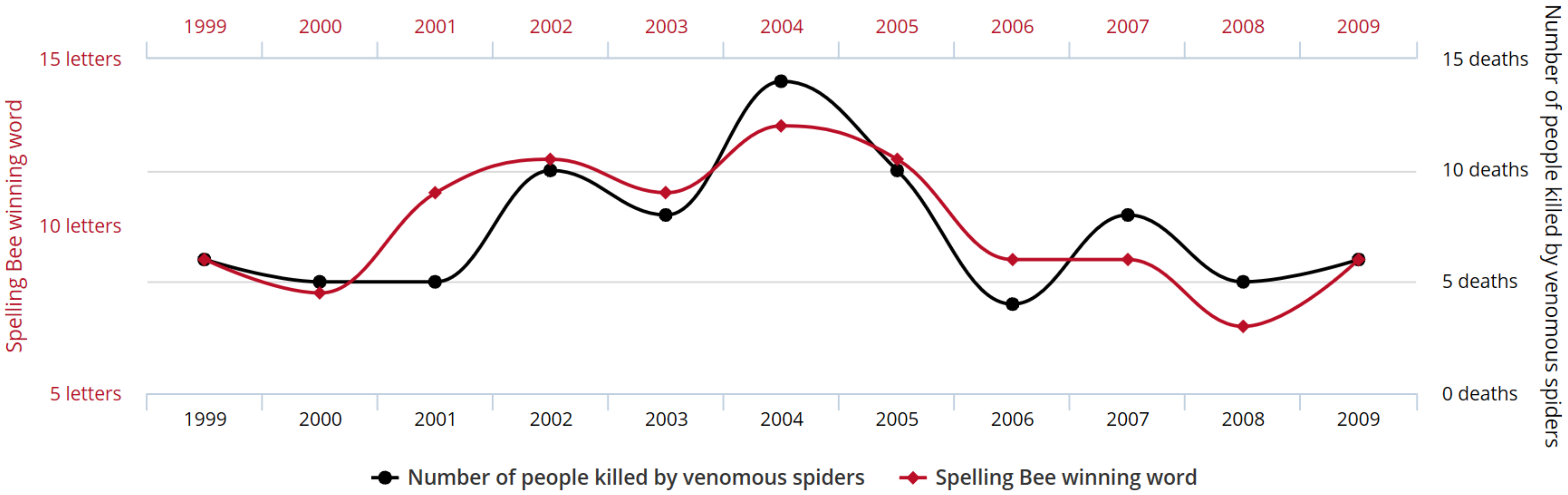
http://www.tylervigen.com/spurious-correlations

Letters in Winning Word of Scripps National Spelling Bee

correlates with

Number of people killed by venomous spiders

Correlation: 80.57% (r=0.8057)



tylervigen.com

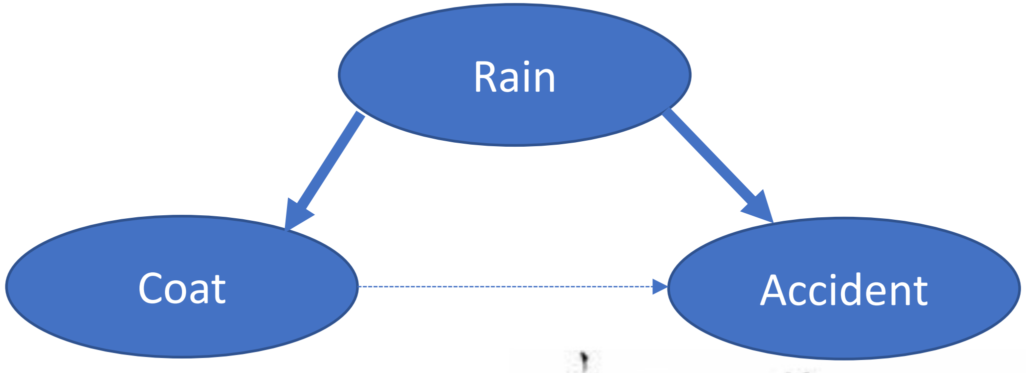
Data sources: National Spelling Bee and Centers for Disease Control & Prevention

Story: London Taxi Drivers

- Examples:

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Decision based on the causality?



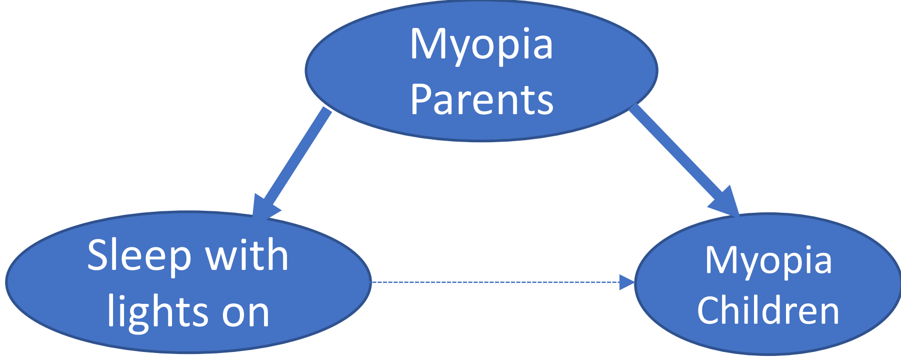
Examples:

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…

Correlation is not causality
Causality really matters

Another example: Myopia study



- A study published in Nature made the causal conclusion that children who sleep with the light on are more likely to develop myopia later in life.

G. E. Quinn, C. H. Shin, M. G. Maguire, and R. A. Stone, “Myopia and ambient lighting at night,” Nature, vol. 399, no. 6732, pp. 113–113, 1999

- However, as it turns out, myopic parents tend to leave the light on more often, as well as pass their genetic predisposition to myopia to their children. Accounting for the confounding variable of parent’s myopia, the causal results were subsequently invalidated or substantially weakened.

Gwiazda J, Ong E, Held R*, et al*. Myopia and ambient night-time lighting. *Nature* 2000;**404**:144.
Zadnik K, Jones LA, Irvin BC*, et al*. Myopia and ambient night-time lighting. *Nature* 2000;**404**:143–4.

Recap: Prediction is insufficient for choosing interventions

How often do they lead us to the right decision?

- Unclear, predictive algorithms provide no insight on effects of decisions

Will the predictions be robust tomorrow, or in new contexts?

- Correlations can change
- Causal mechanisms more robust

What if the prediction accuracy is really high? Does that help?

- Active interventions change correlations

PART I. Introduction to Counterfactual Reasoning

PART II. Methods for Causal Inference

PART III. Large-scale and Network Data

PART IV. Broader Landscape

PART I. Introduction to Counterfactual Reasoning

What is causality?

Potential Outcomes Framework

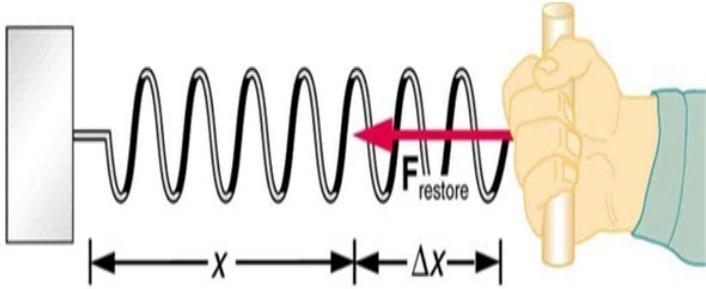
Unobserved Confounds / Simpson’s Paradox

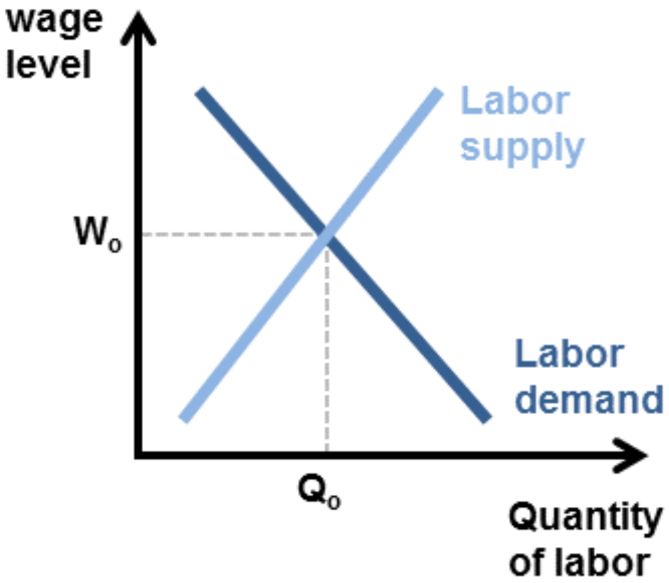
Structural Causal Model Framework

Cause and Effect

- Questions of cause and effect common in biomedical and social sciences
- Such questions form the basis of almost all scientific inquiry
  - Medicine: drug trials, effect of a drug
  - Social sciences: effect of a certain policy
  - Genetics: effect of genes on disease

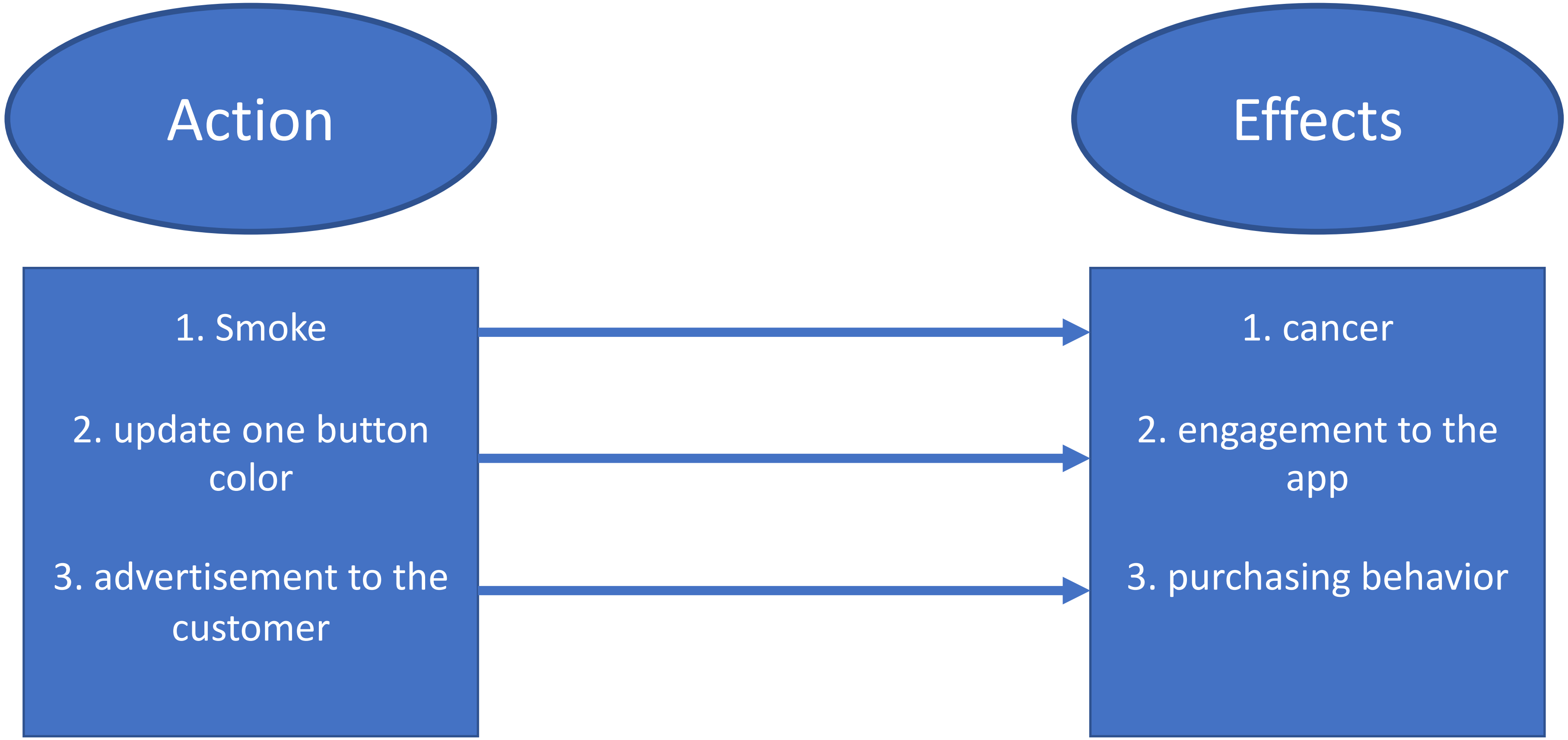
- **So what is causality?**
- **What does it mean to *cause* something?**





Causality examples （A causes B ）

- Exposure/Action/Decision Effects



Action

1. Smoke

2. update one button color

3. advertisement to the customer

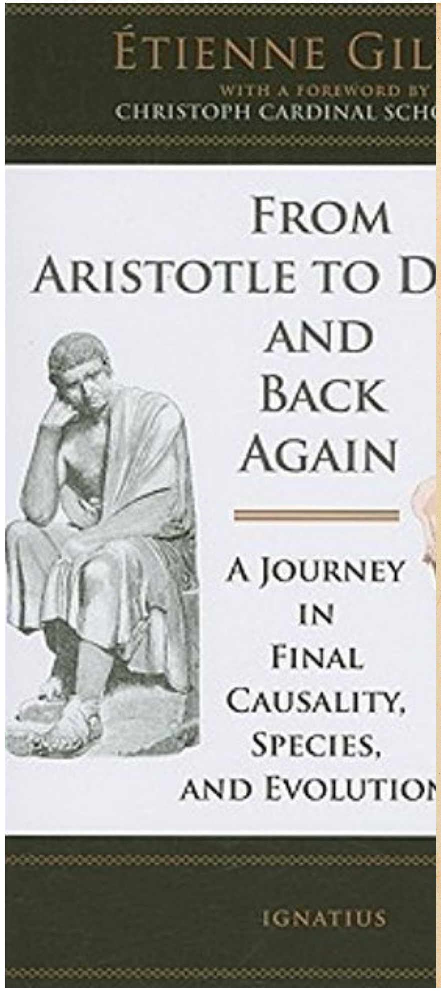
Effects

1. cancer

2. engagement to the app

3. purchasing behavior

A big scholarly debate, from Aristotle to Russell



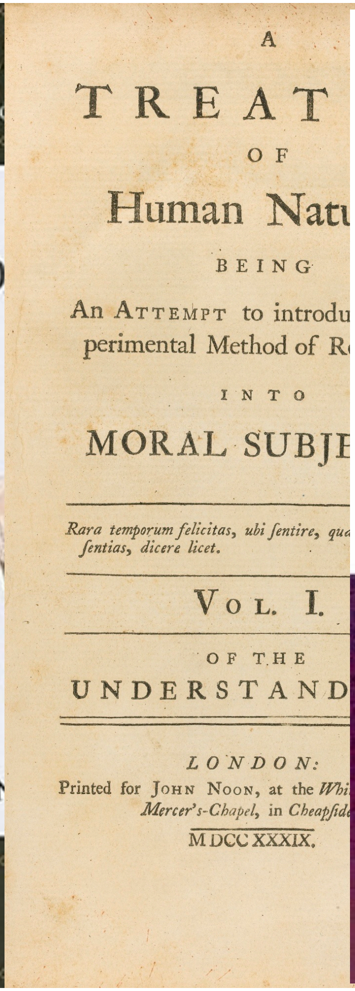
ÉTIENNE GIL

WITH A FOREWORD BY
CHRISTOPH CARDINAL SCHO

FROM
ARISTOTLE TO D
AND
BACK
AGAIN

A JOURNEY
IN
FINAL
CAUSALITY,
SPECIES,
AND EVOLUTION

IGNATIUS



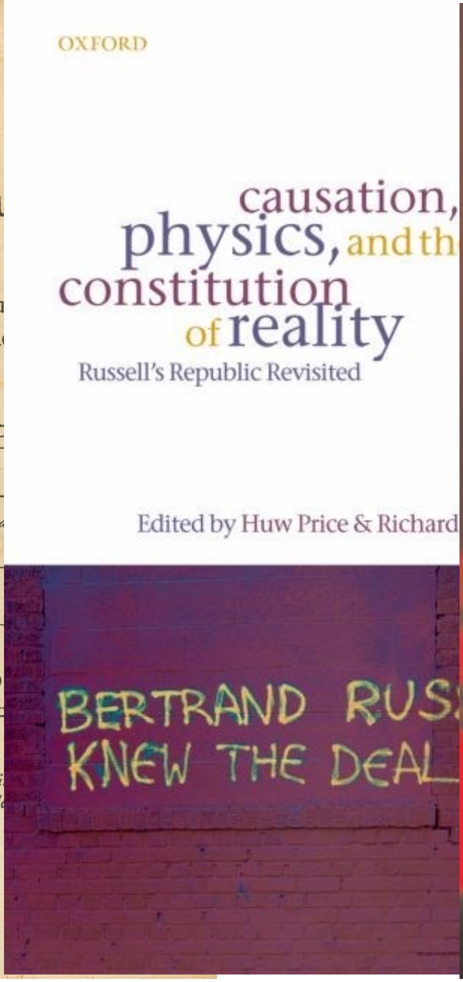
A
TREAT
OF
Human Natu
BEING
An ATTEMPT to introdu
perimental Method of R
INTO
MORAL SUBJE

Rara temporum felicitas, ubi ſentire, qu
ſentias, dicere licet.

VOL. I.

OF THE
UNDERSTAND

LONDON:
Printed for JOHN NOON, at the *Whi*
Mercer's-Chapel, in *Cheapſide*
MDCCXXXIX.

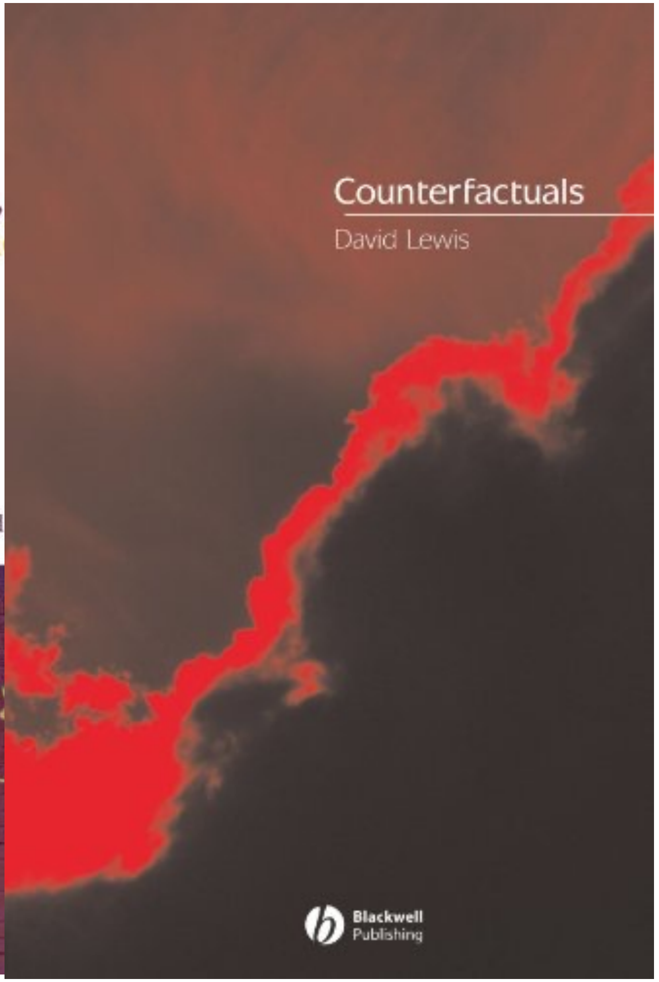


OXFORD

causation,
physics, and th
constitution
of reality

Russell's Republic Revisited

Edited by Huw Price & Richard



Counterfactuals

David Lewis

Blackwell
Publishing

What is causality?

- A fundamental question
- Surprisingly, until very recently---maybe the last 30+ years---we have not had a mathematical language of causation. We have not had an arithmetic for representing causal relationships.

"More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history"

--Gary King, Harvard University

The Three Layer Causal Hierarchy

Pearl, Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution, arXiv:1801.04016v1. 11 Jan 2018

Level	Typical Activity	Typical Question	Examples
1. Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y \mid do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x \mid x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

A practical definition

Definition: T causes Y iff
changing T leads to a change in Y,
keeping everything else constant.

The **causal effect** is the magnitude by which Y is changed by a unit change in T.

Called the "interventionist" interpretation of causality.

**Interventionist* definition [http://plato.stanford.edu/entries/causation-mani/]

Keeping everything else constant: Imagine a *counterfactual* world

"What-if" questions

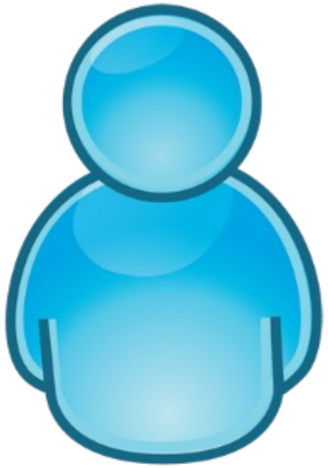
Reason about a world that does not exist.

- What if a system intervention was not done?
- What if an algorithm was changed?
- What if I gave a drug to a patient?

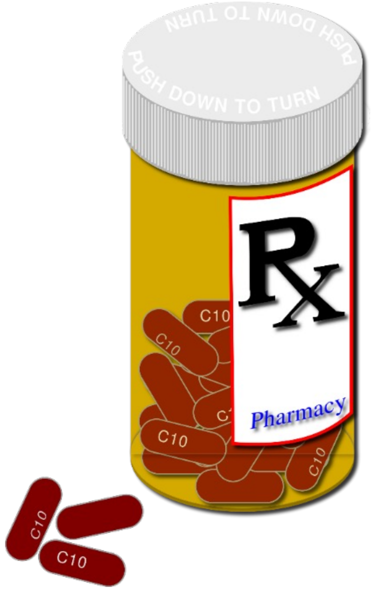
PART I. Introduction to Counterfactual Reasoning

- What is causality?
- Potential Outcomes Framework
- Unobserved Confounds / Simpson’s Paradox
- Structural Causal Model Framework

Potential Outcomes framework

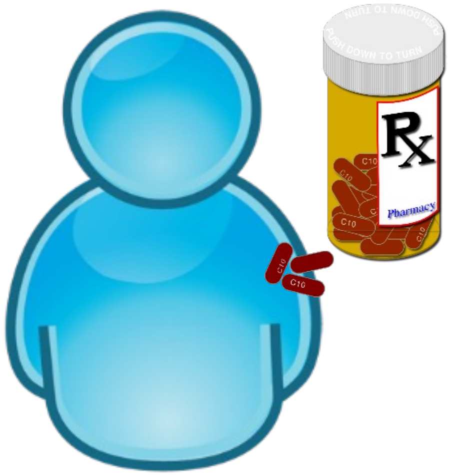


Alice



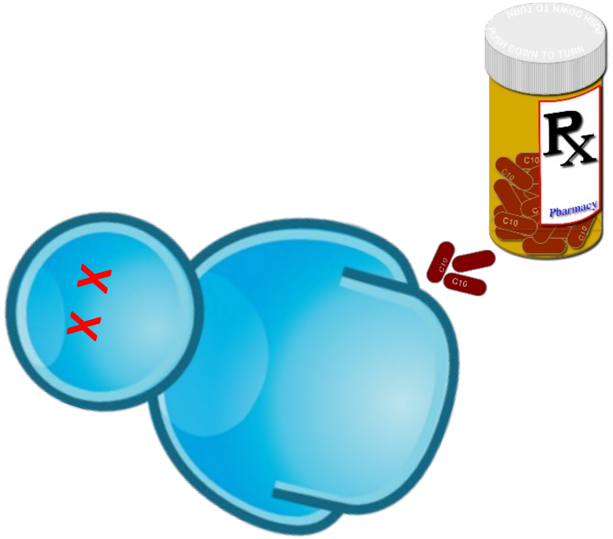
Treatment

Potential Outcomes framework



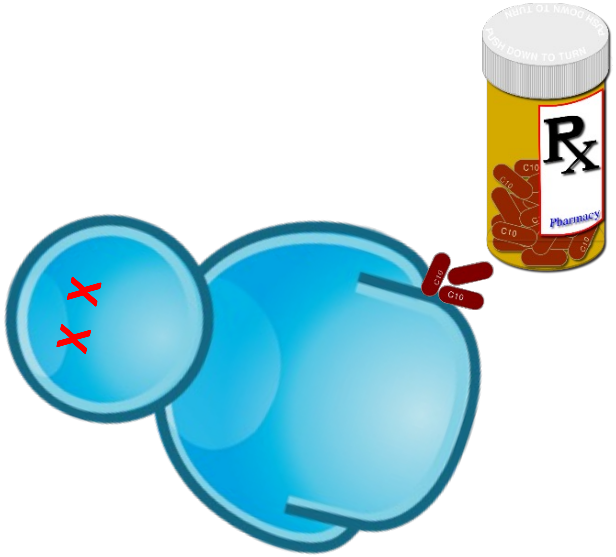
Alice

Potential Outcomes framework

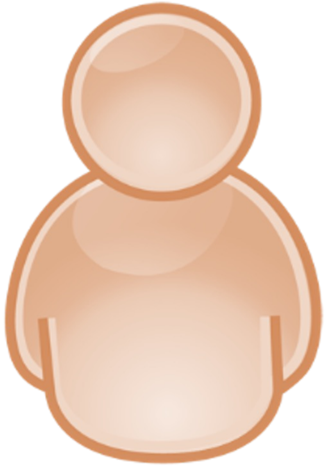


Alice

Potential Outcomes framework: Introduce a counterfactual quantity



$$Y_{T=1}$$



$$Y_{T=0}$$

Causal effect of treatment =

$$E[Y_{T=1} \quad - Y_{T=0}]$$

Causal inference is the problem of estimating the counterfactual $Y_{t=\sim t}$

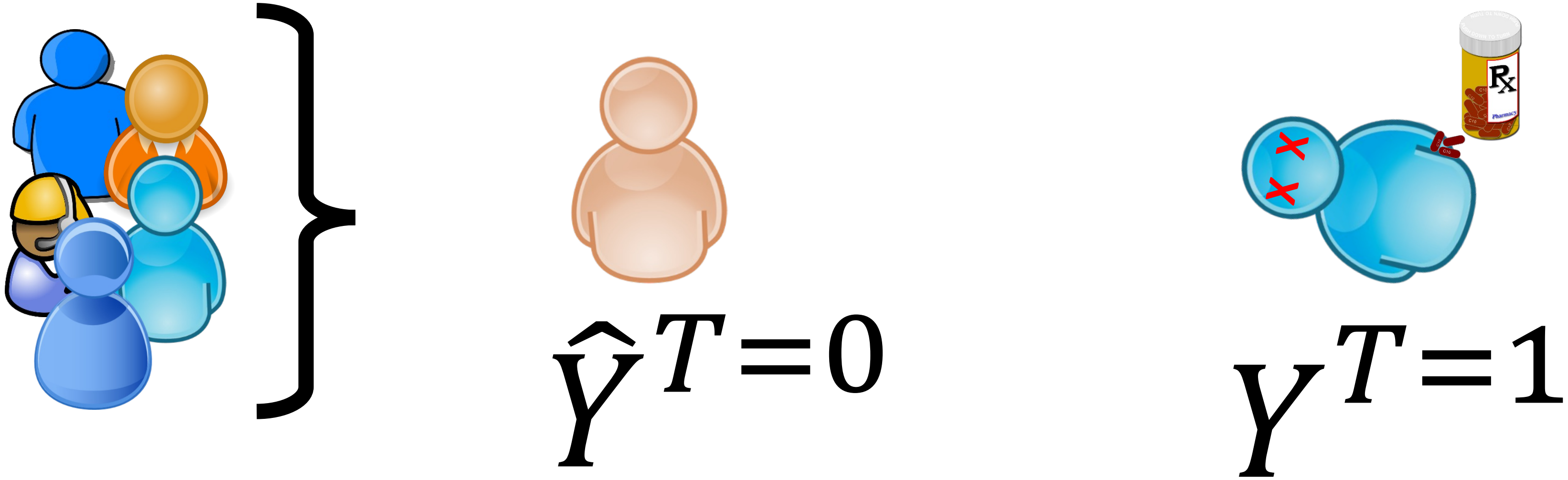
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	0.3
P2	0	0.8	0.6
P3	1	0.3	0.2
P4	0	0.3	0.1
P5	1	0.5	0.5
P6	0	0.6	0.5
P7	0	0.3	0.1

Causal effect: $E[Y_{t=1} - Y_{t=0}]$

Fundamental problem of causal inference: For any person, observe only one: either $Y_{t=1}$or $Y_{t=0}$

Fundamental problem: counterfactual outcome is not observed

- “Missing data” problem
- Estimate missing data values using various methods
- $Y_{T=0}$ now becomes an estimated quantity, based on outcomes of other people who did not receive treatment



$$\hat{Y}^{T=0} \qquad Y^{T=1}$$

Randomized Experiments are the “gold standard”

One way to estimate counterfactual

Cost: Possibly risky, unethical

Unethical to deny useful treatment or administer risky treatment.

Infeasible or costly in other situations.

What can we do when an experiment is not possible?

Coming soon in Section 2

Recap: Potential Outcomes Framework

- Potential outcomes reasons about causal effects by comparing outcome of treatment to outcome of no-treatment
- For any individual, we cannot observe both treatment and no-treatment.
- Randomized experiments are one solution
- We’ll discuss others in tutorial Section 2

PART I. Introduction to Counterfactual Reasoning

- What is causality?
- Potential Outcomes Framework
- Unobserved Confounds / Simpson’s Paradox
- Structural Causal Model Framework

Example: Auditing the effect of an algorithm

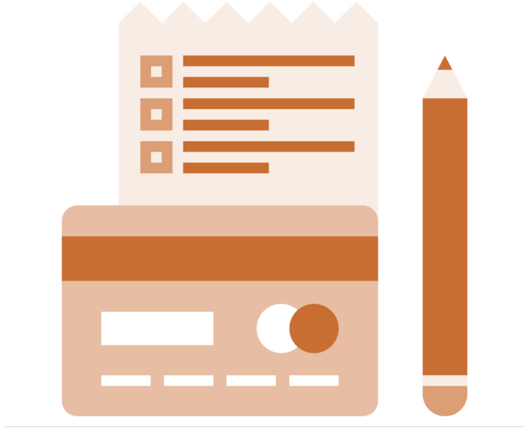
System changes algorithm from A to B at some point.

Is the new algorithm B better?

Say a feature that provides information or discount for a financial product.

Success Rate=ρ

Algorithm A



?

Algorithm B

New algorithm increases overall success rate

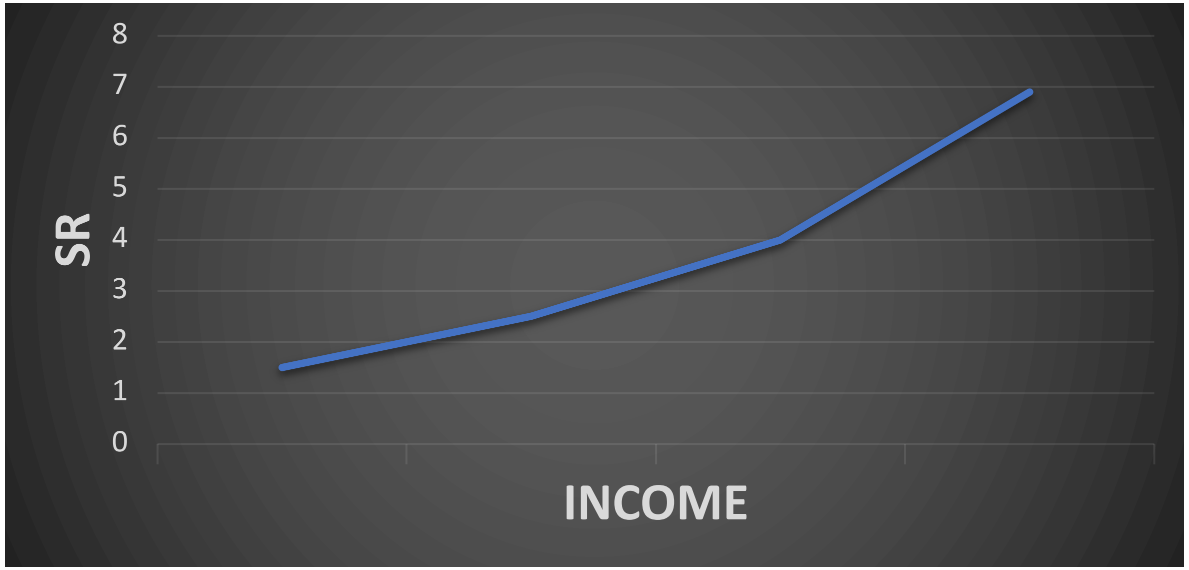
Two algorithms, A (old) and B (new) running on the system.

From system logs, collect data for 1000 sessions for each.
Measure Success Rate (SR).

Old Algorithm (A)	New Algorithm (B)
50/1000 **(5%)**	54/1000 **(5.4%)**

New algorithm is better?

Unobserved Confounds



What if there are unobserved features of audience that matter?

Old Algorithm (A)	New Algorithm (B)
10/400 **(2.5%)**	4/200 **(2%)**

Low-income Users

Old Algorithm (A)	New Algorithm (B)
40/600 **(6.6%)**	50/800 **(6.2%)**

High-income Users

The Simpson’s paradox: New algorithm is better overall, but worse for each subgroup

	Old algorithm (A)	New Algorithm (B)
CTR for Low-income users	10/400 (2.5%)	4/200 (2%)
CTR for High-income users	40/600 (6.6%)	50/800 (6.2%)
Total CTR	**50/1000 (5%)**	**54/1000 (5.4%)**

So, which is better?

Simpson (1951)

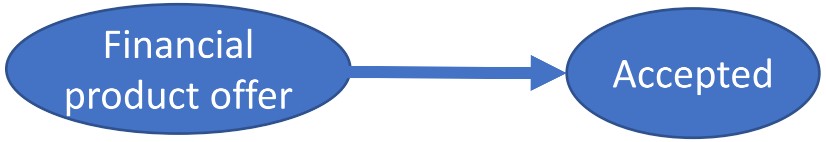
From metrics to decision-making

Did the change to new Algorithm increase success rate for the system?

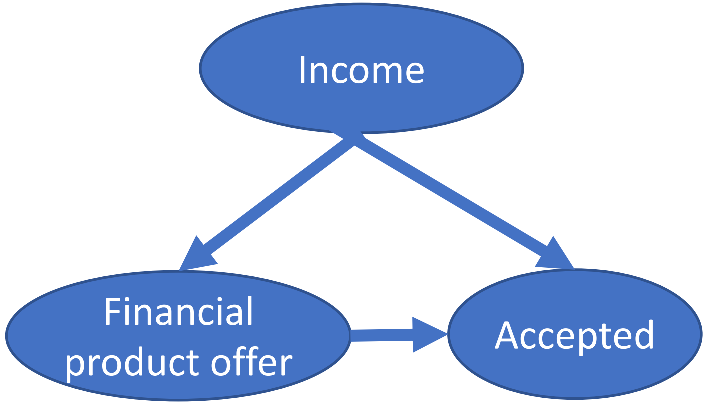
Answer (as usual):

Maybe, maybe not (!)

Higher success rate due to
new algorithm



Higher success rate due to
selection effects

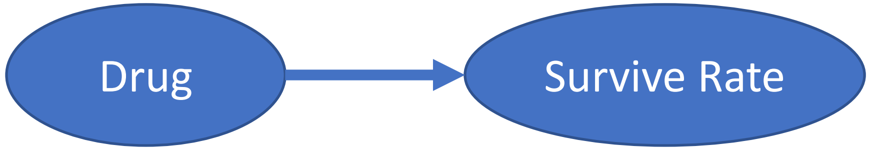


E.g., Algorithm B is shown at a different time than A.

There could be other hidden causal variations.

Not just theory. Differences in interpretations can attract lawsuits (UC Berkeley admissions, 1973)

Simpson’s Paradox in naturally generated data



Drug

Survive Rate

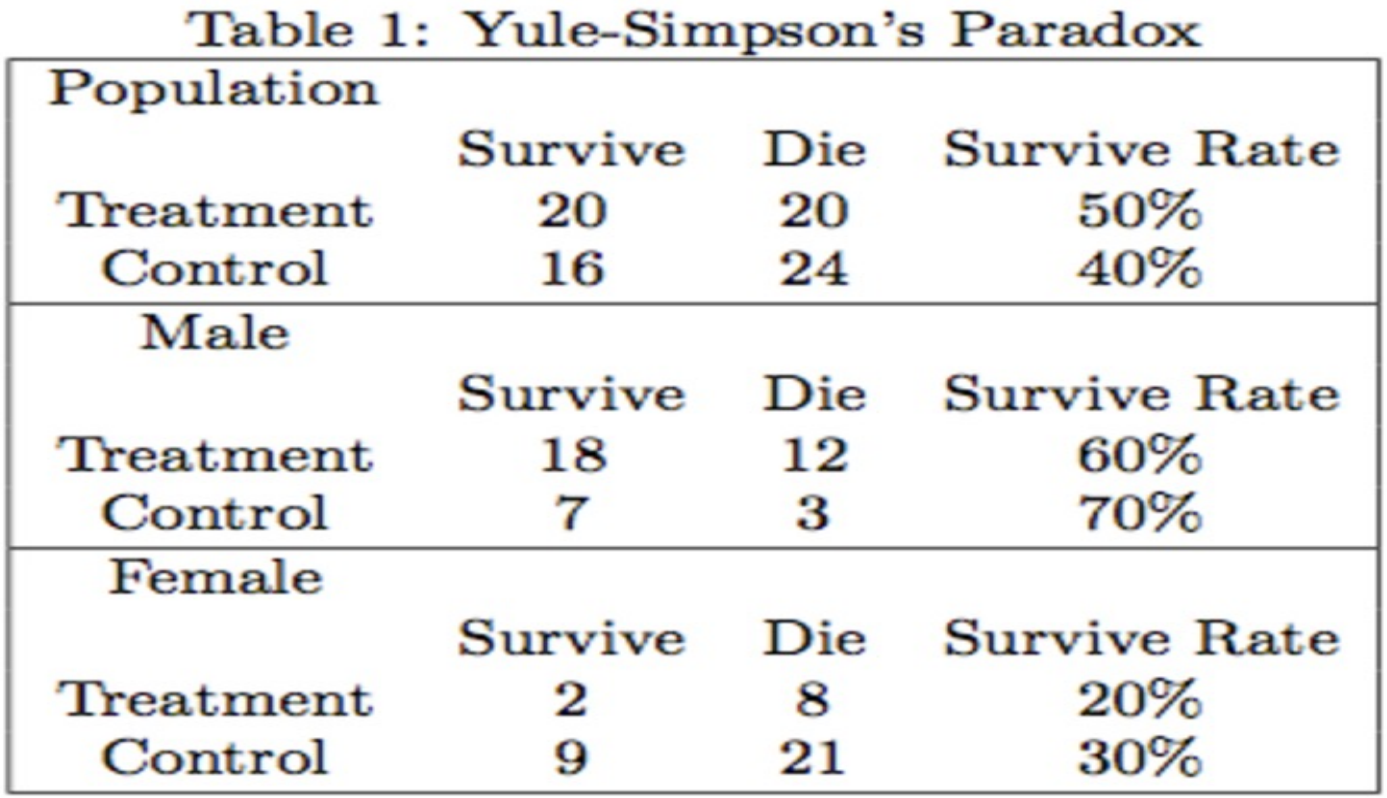


Table 1: Yule-Simpson’s Paradox

Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

Treatment is better

Control is better

Control is better

Simpson’s Paradox

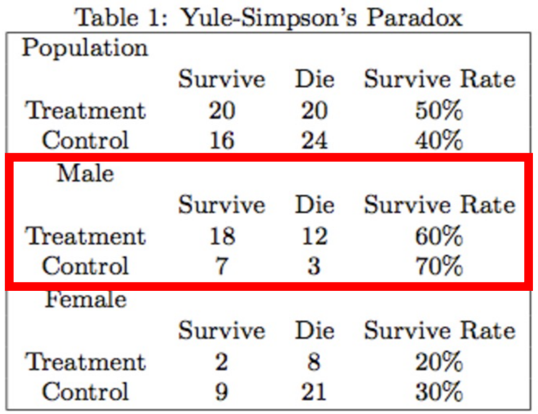
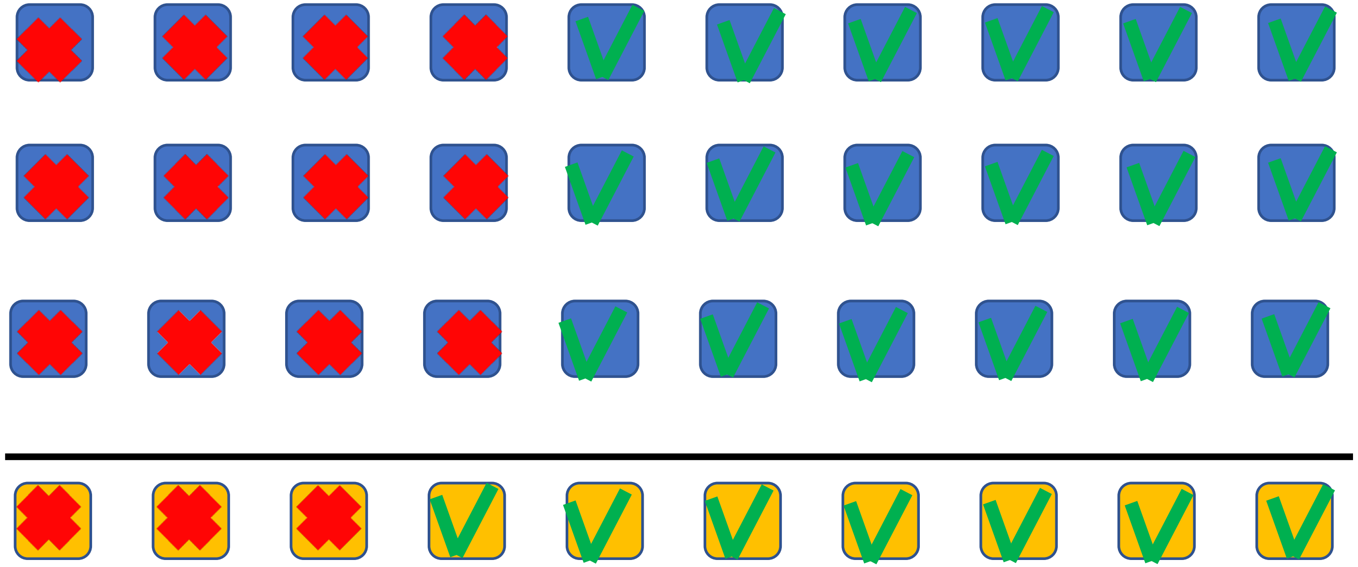


Table 1: Yule-Simpson’s Paradox

Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%



🟦 Male treatment

🟨 Male control

Simpson’s Paradox

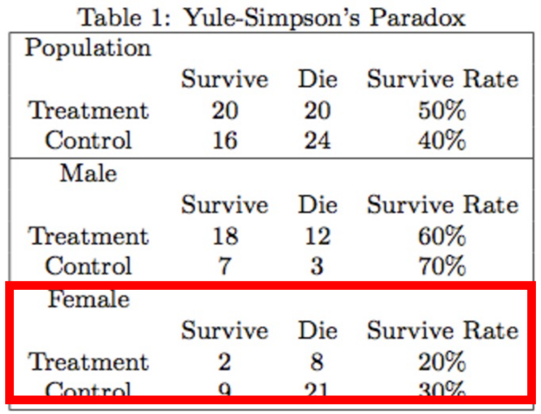
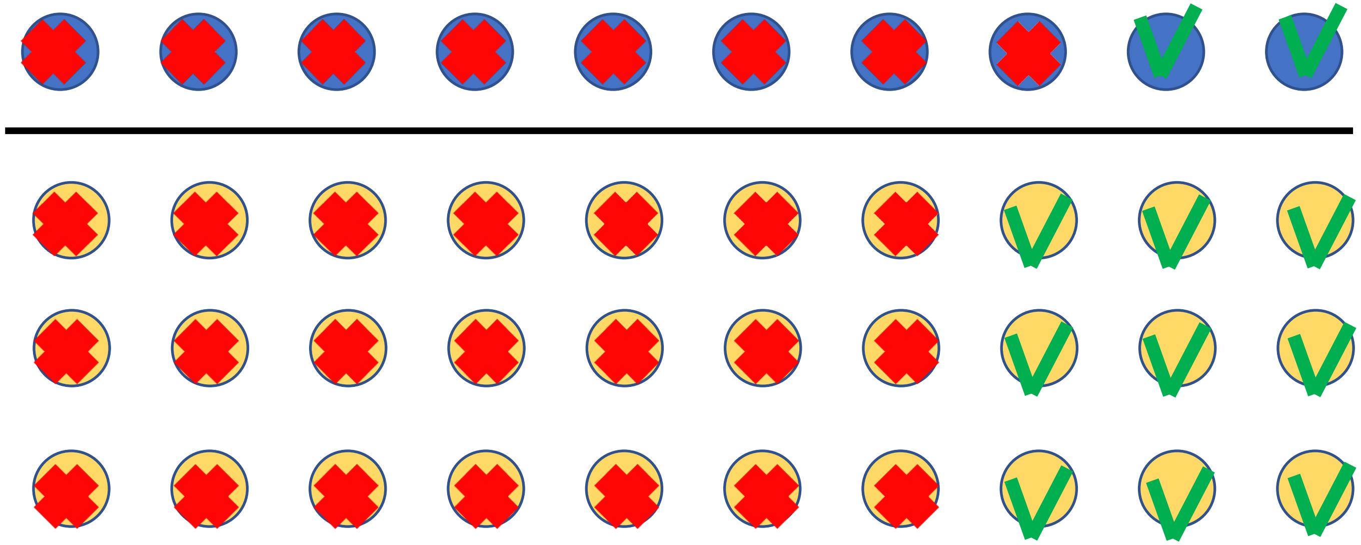


Table 1: Yule-Simpson’s Paradox

Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

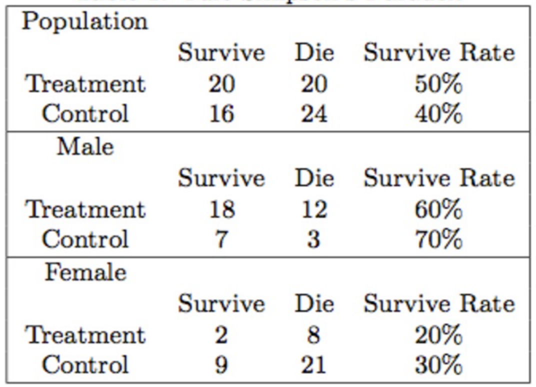


● Female treatment

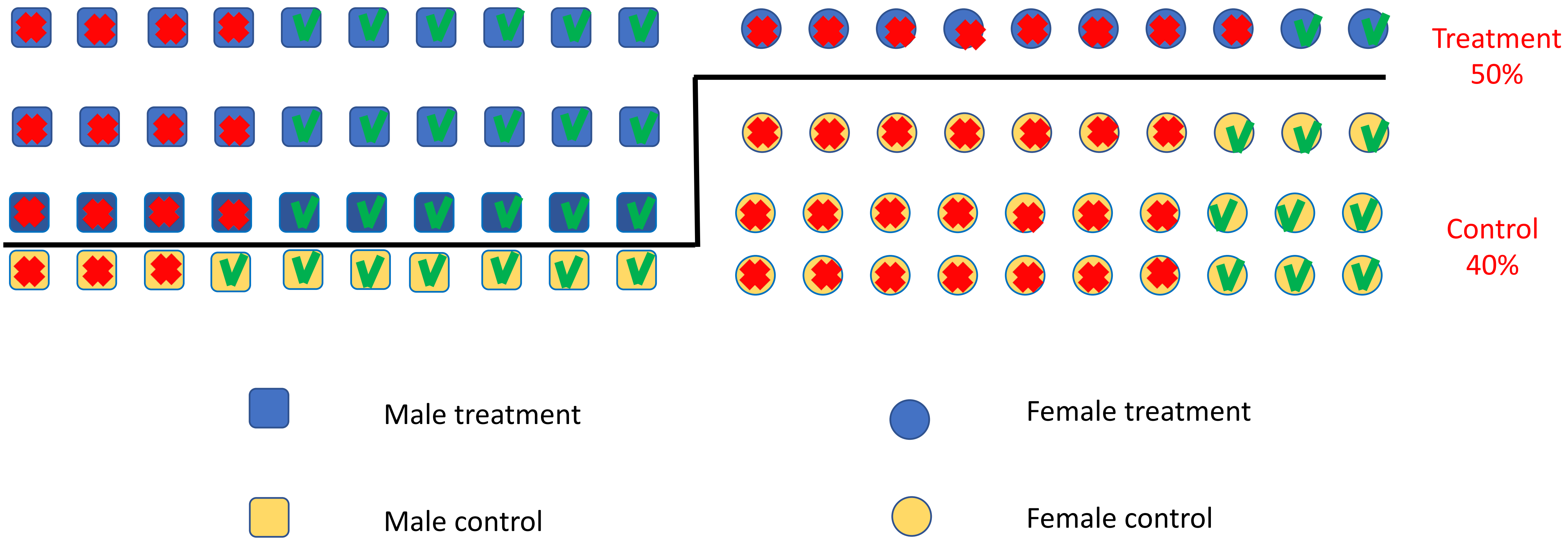
● Female control

Simpson’s Paradox

Table 1: Yule-Simpson’s Paradox

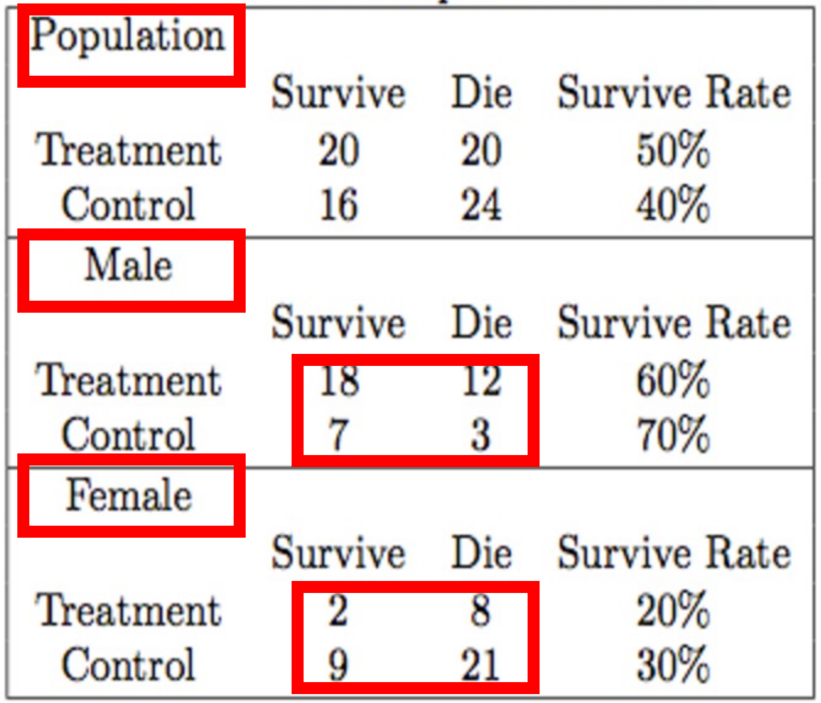


Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%



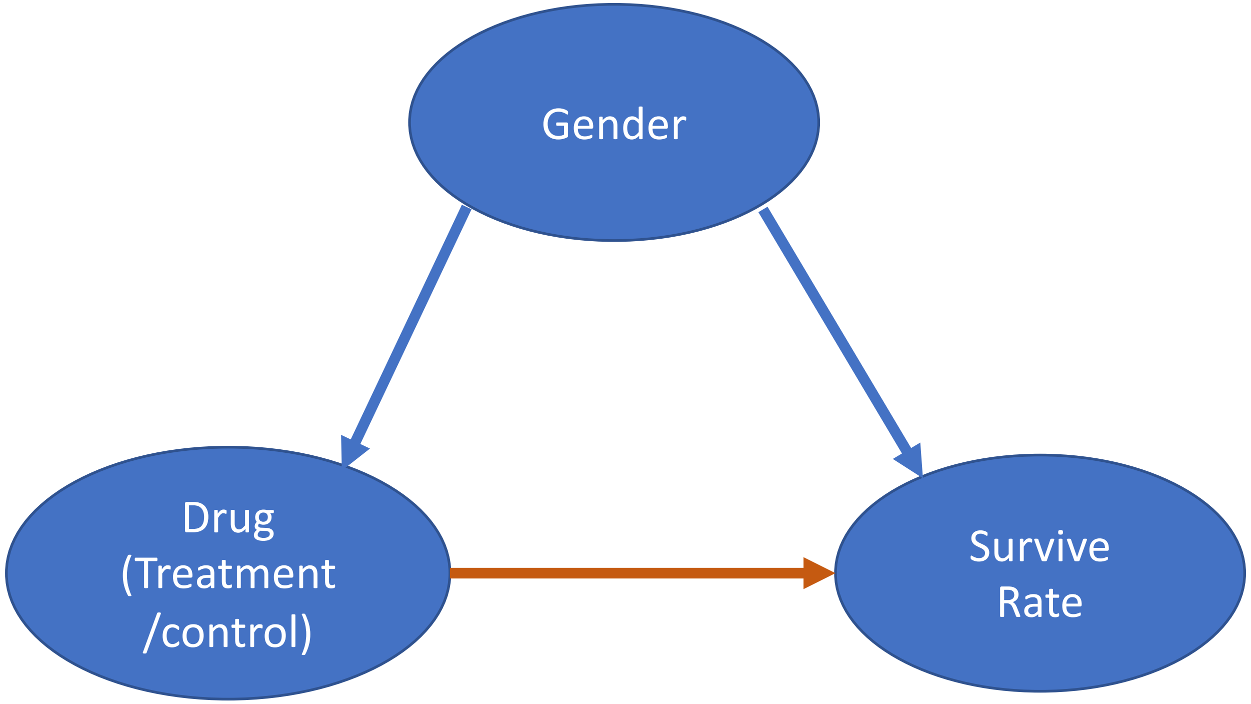
Confounding factor: Gender

Table 1: Yule-Simpson's Paradox



Population			
	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male			
	Survive	Die	Survive Rate
Treatment	18	12	60%
Control	7	3	70%
Female			
	Survive	Die	Survive Rate
Treatment	2	8	20%
Control	9	21	30%

Confounding factor



Making sense of such data can be too complex.

Not Simpson's Paradox

Recap: Unobserved Confounds

- Unobserved confounds are a threat to causal reasoning

Recap: Section 1 - Introduction

- Causality is important for decision-making and study of effects
- Potential Outcomes Framework gives practical method for estimating causal effects
  - Translates causal inference into counterfactual estimation
- Unobserved confounds are a critical challenge
- Structural Causal Model Framework gives language for expressing and reasoning about causal relationships

PART I. Introduction to Counterfactual Reasoning

PART II. Methods for Causal Inference

PART III. Large-scale and Network Data

PART IV. Broader Landscape

PART II.
Methods for Causal Inference

PART II. Methods for Causal Inference

- Observational Studies
- Natural Experiments
- Refutations

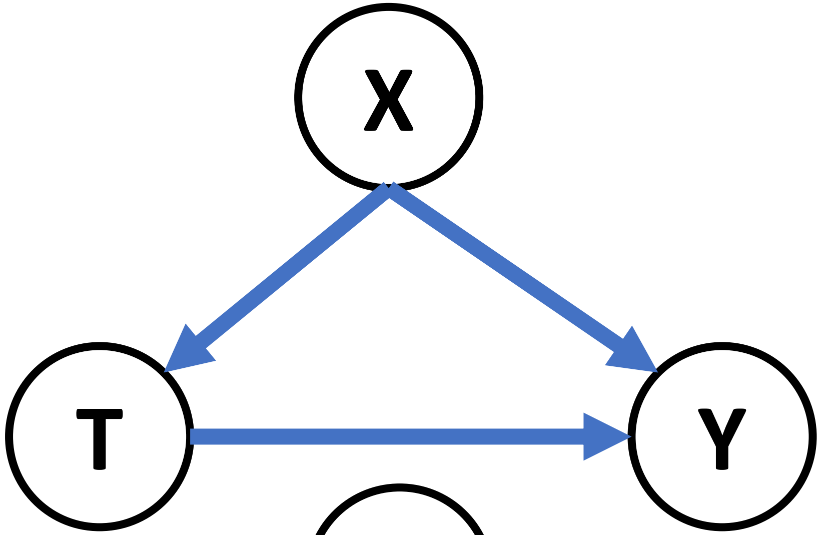
Review: Treatment, Outcome and Confound

Goal: Estimate effect of a treatment T on an outcome Y

But, confound X influences both T and Y

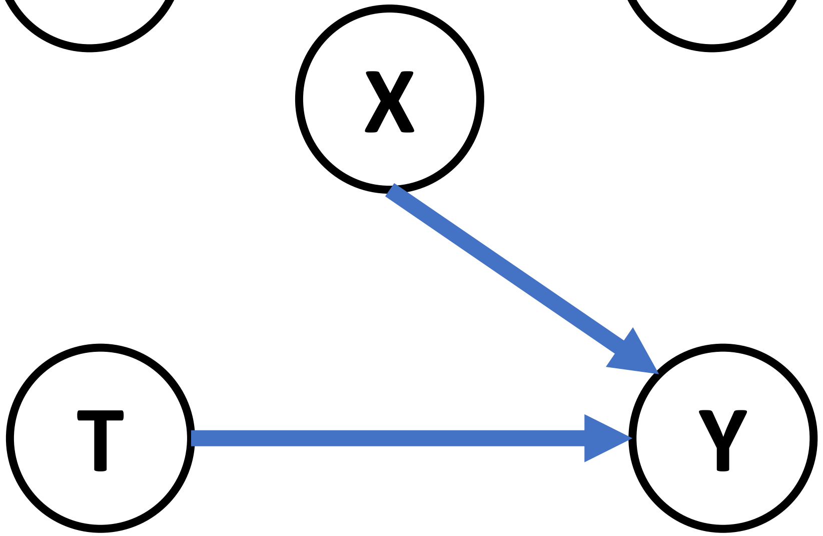
To estimate $T \rightarrow Y$, break the dependence $X \rightarrow T$ (that is, $T \perp\!\!\!\perp X$)

- Y ⫫ X also works, but much less practical.



Randomized experiments actively assign treatment T independent of any confound X

Thus, by construction: $T \perp\!\!\!\perp X$



Review: Treatment, Outcome and Confound

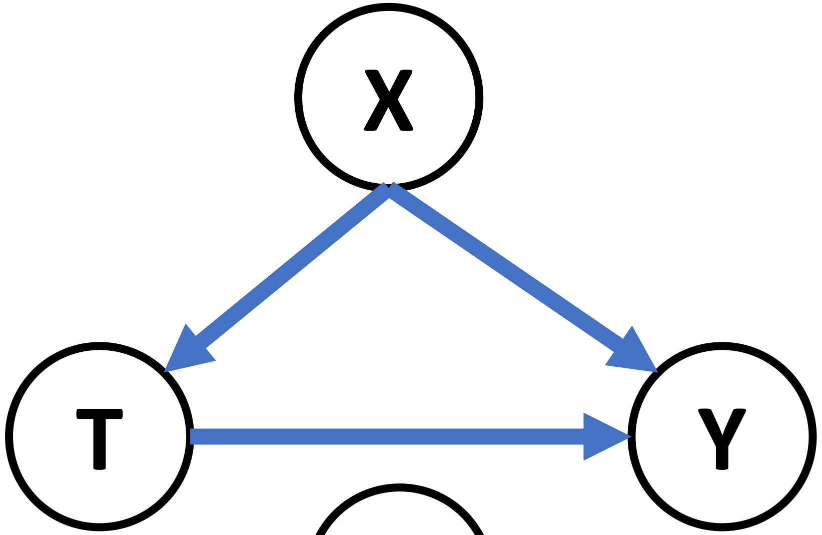
Goal: Estimate effect of a treatment T on an outcome Y

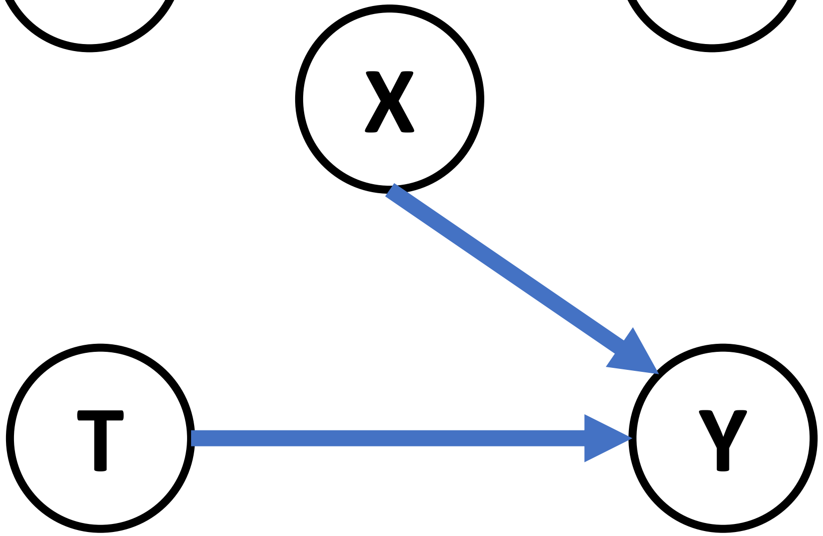
But, confound X influences both T and Y

To estimate $T \rightarrow Y$, break the dependence $X \rightarrow T$ (that is, $T \perp\!\!\!\perp X$)

Randomized experiments actively assign treatment T independent of any confound X

Thus, by construction: $T \perp\!\!\!\perp X$





Running example ~~Review~~: Exercise, Cholesterol, and Age

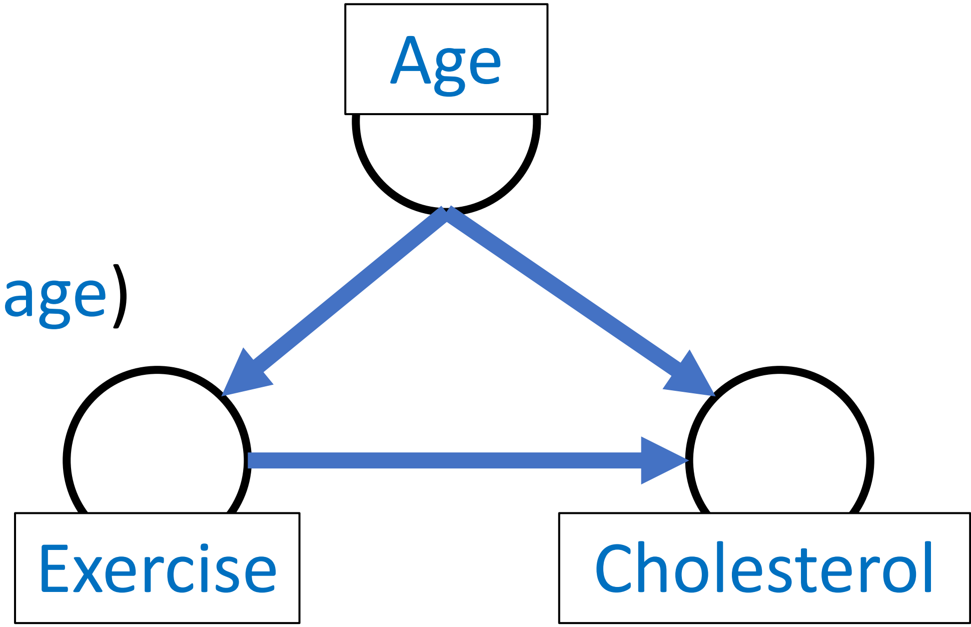
Goal: Estimate effect of exercise on cholesterol

But, one's age influences both exercise and cholesterol

To estimate exercise→cholesterol, break the dependence age→exercise (that is, exercise ⫫ age)

Randomized experiments actively assign exercise independent of any age

Thus, by construction: exercise ⫫ age



Running example ~~Review~~: Exercise, Cholesterol, and Age

Goal: Estimate effect of exercise on cholesterol

But, one's age influences both exercise and cholesterol

To estimate exercise→cholesterol, break the dependence age→exercise (that is, exercise ⫫ age)

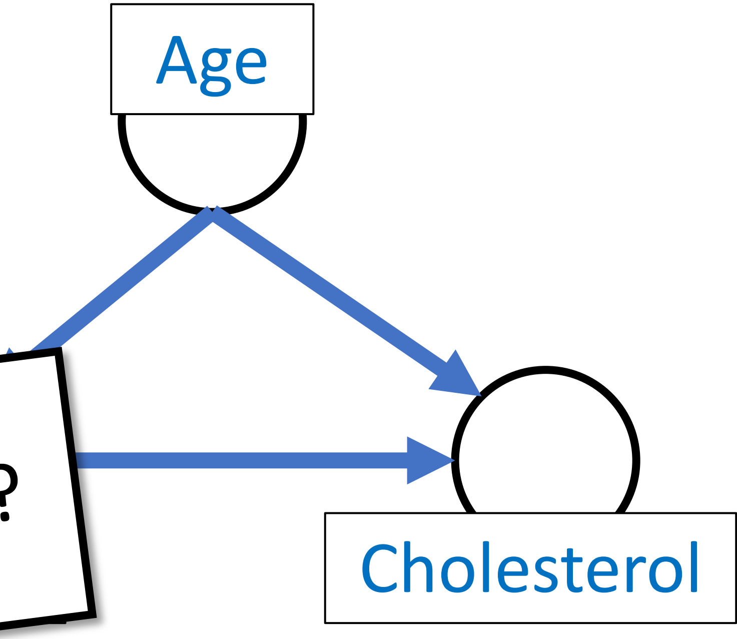
Randomized expe

exerc

Thus,

⫫ age

But, what if we cannot actively intervene?



Part II.A. Observational Studies

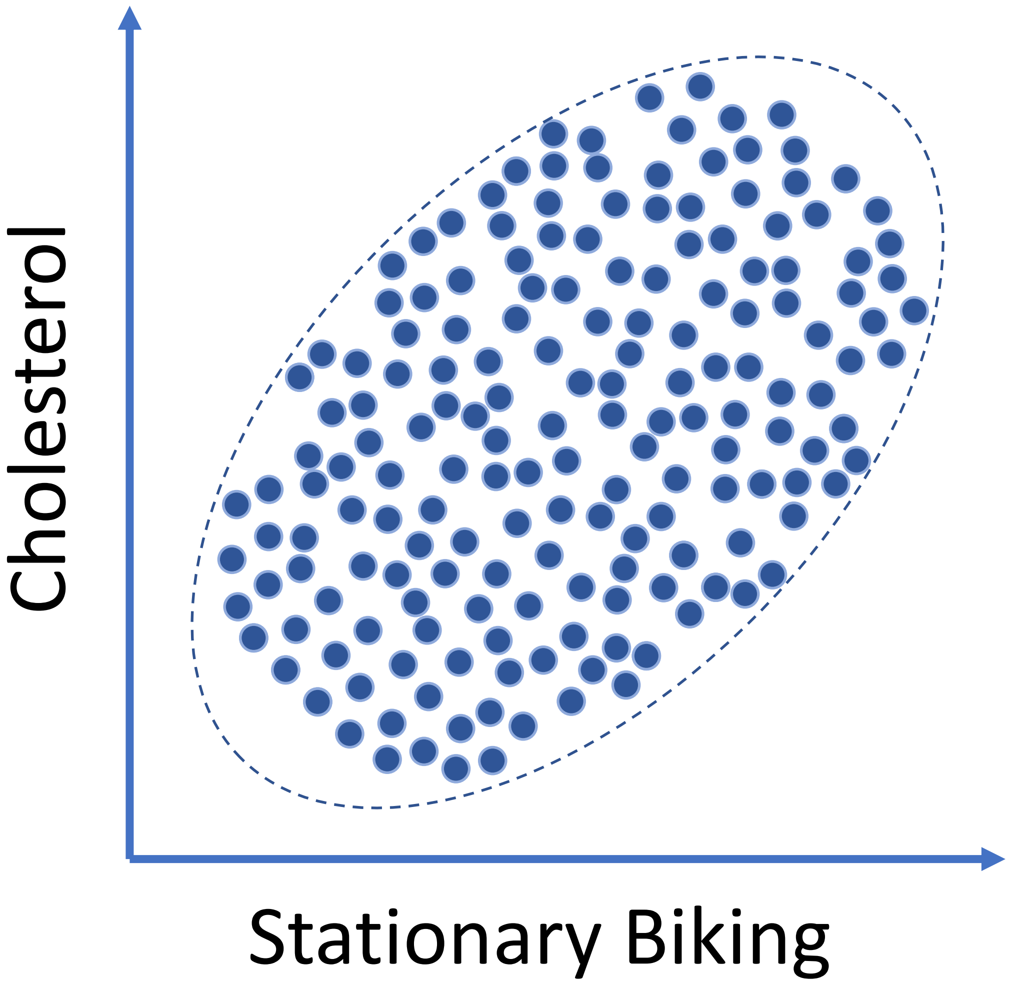
“Simulating randomized experiments"

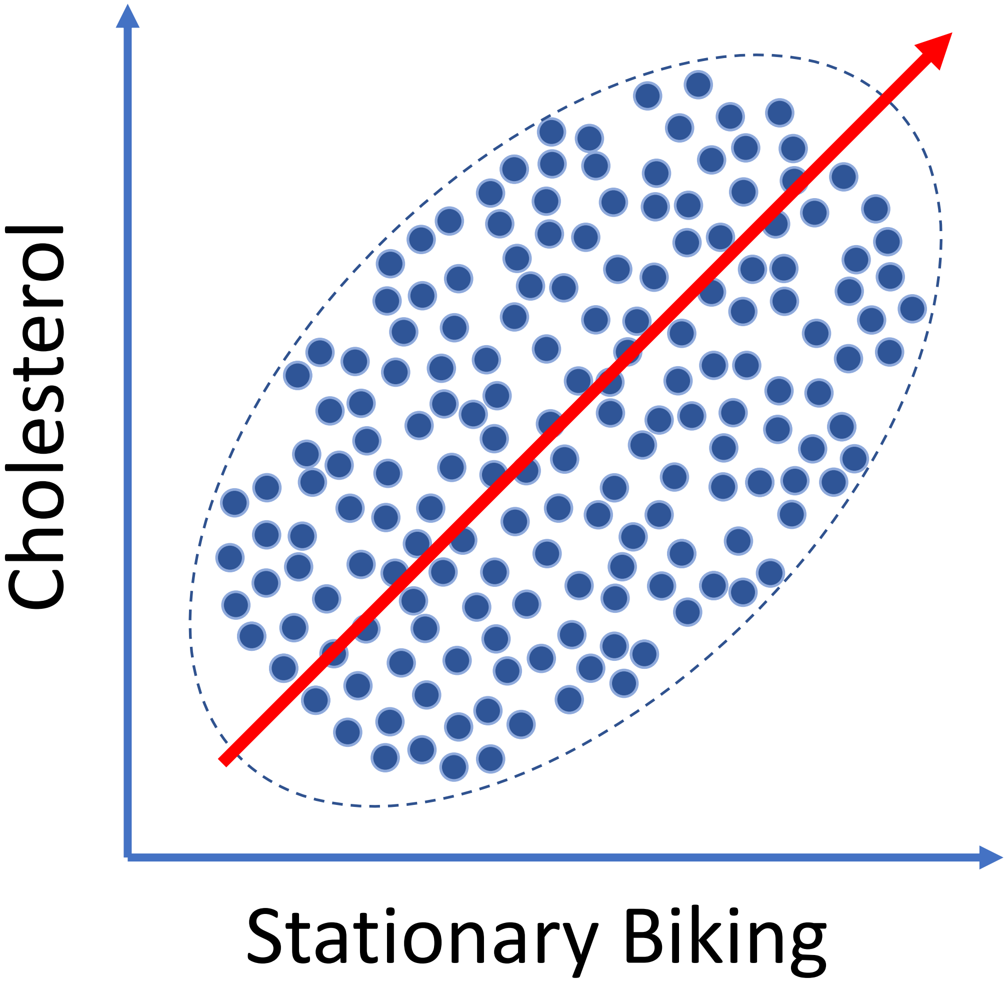
- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression
- Doubly Robust
- Synthetic Controls

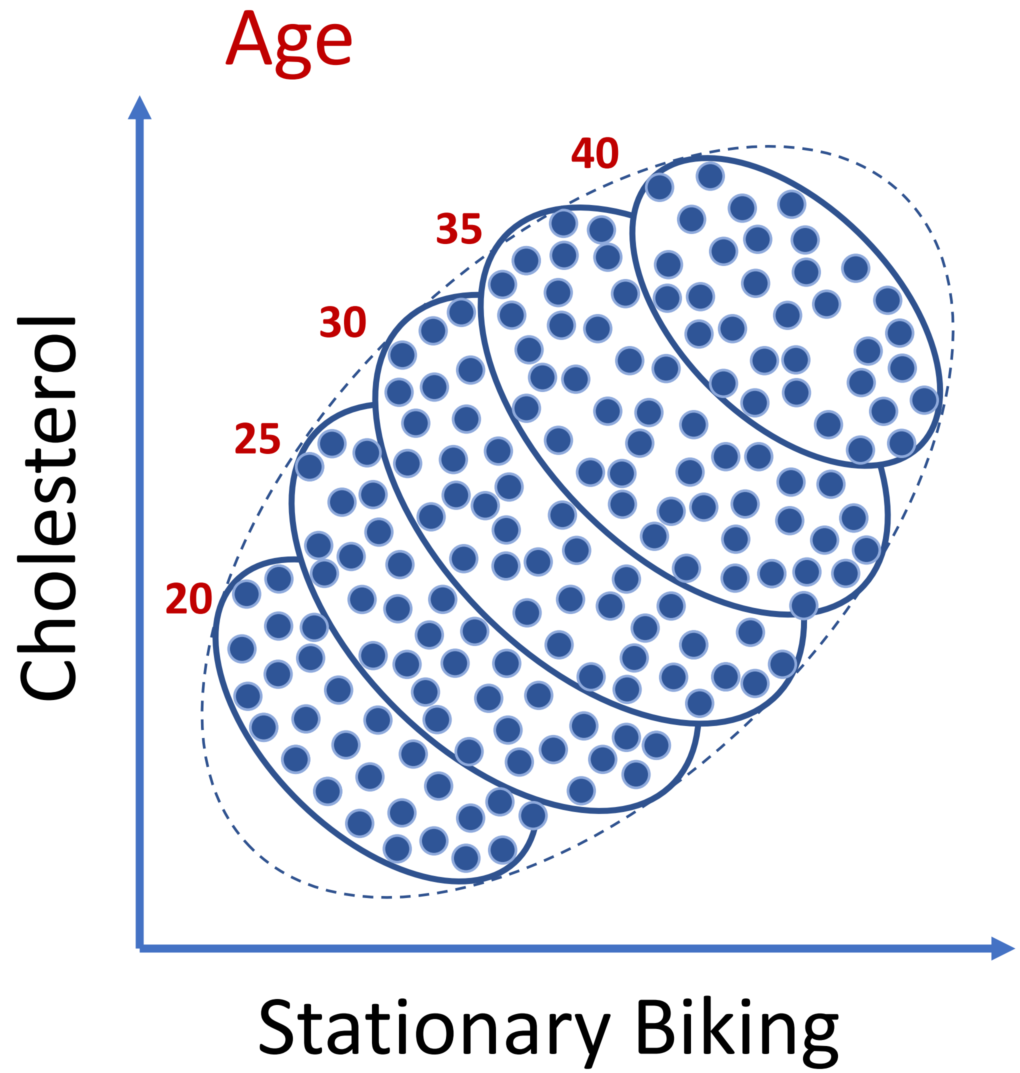
Part II.A. Observational Studies

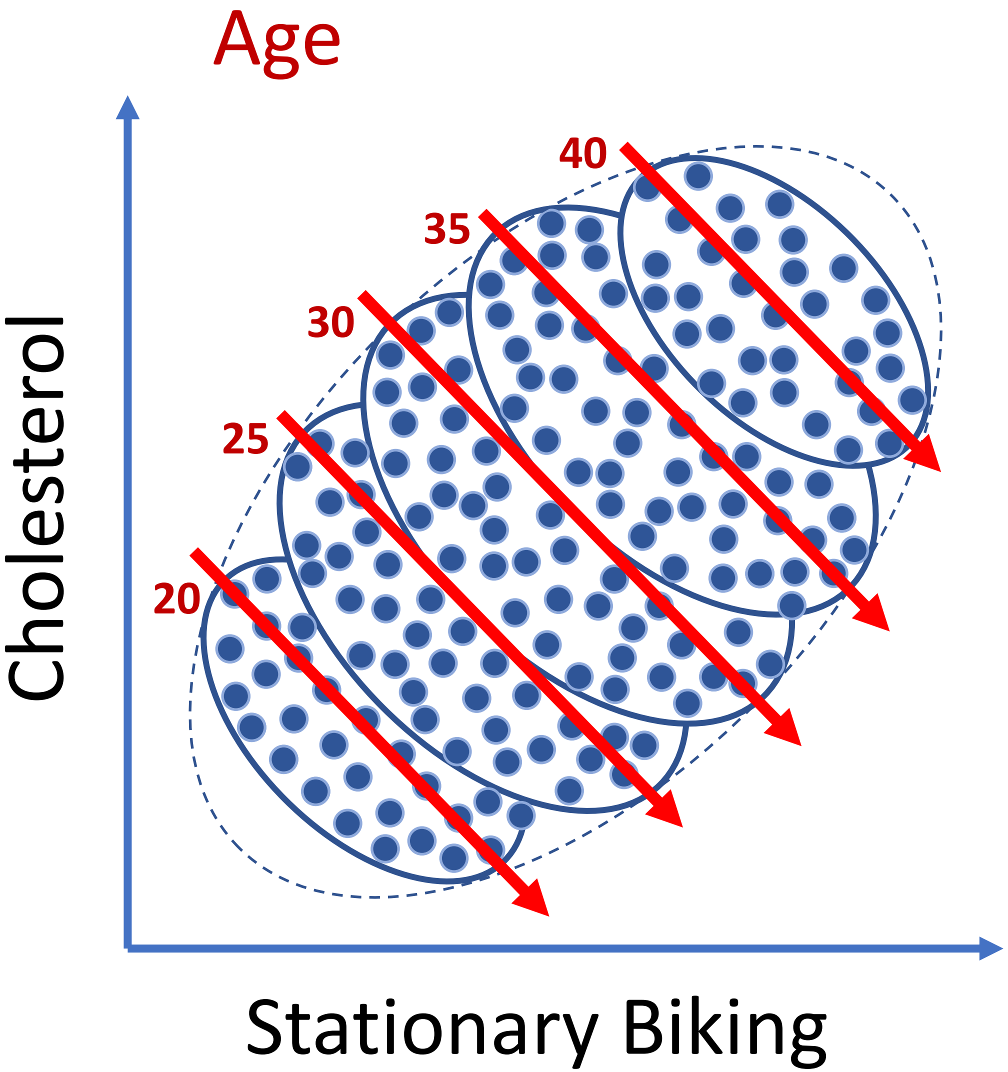
“Simulating randomized experiments"

- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression
- Doubly Robust
- Synthetic Controls


Cholesterol
Stationary Biking


Cholesterol
Stationary Biking


Age
40
35
30
25
20
Cholesterol
Stationary Biking


Age
Cholesterol
Stationary Biking
40
35
30
25
20

Recapping what just happened

- At first, more *stationary biking* seems to lead to higher *cholesterol*
- But, we realize that there is a confounder, *age, that influences both stationary biking and cholesterol*
- We condition on age (by analyzing each age group separately)
- And find stationary biking now seems to lead to lower cholesterol

Conditioning:

$$P(Cholesterol \mid do(S\_Biking)) = \sum_{age} P(Cholesterol \mid S\_Biking, age)\, P(age)$$

Conditioning

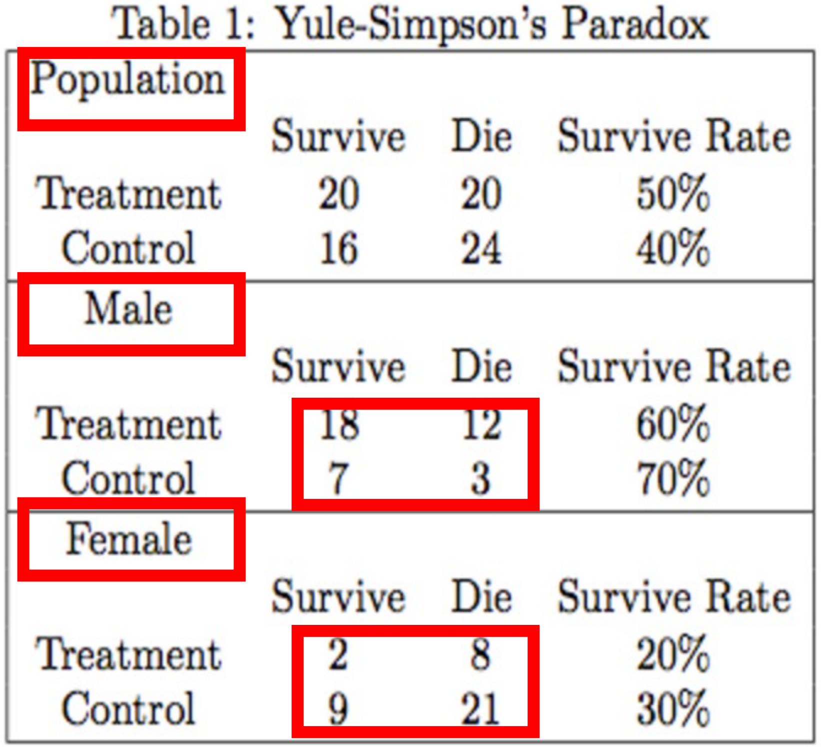


Table 1: Yule-Simpson's Paradox

Population	Survive	Die	Survive Rate
Treatment	20	20	50%
Control	16	24	40%
Male	**Survive**	**Die**	**Survive Rate**
Treatment	18	12	60%
Control	7	3	70%
Female	**Survive**	**Die**	**Survive Rate**
Treatment	2	8	20%
Control	9	21	30%

$$
\begin{aligned}
\widehat{ACE}_{unadj} &= \widehat{P}(Y=1 \mid Z=1) - \widehat{P}(Y=1 \mid Z=0) \\
&= 0.50 - 0.40 = 0.10 > 0.
\end{aligned}
$$

male

female

$$
\begin{aligned}
&\widehat{ACE}_{adj} \\
&= \{\widehat{P}(Y=1 \mid Z=1, X=1) - \widehat{P}(Y=1 \mid Z=0, X=1)\}\widehat{P}(X=1) \\
&\quad + \{\widehat{P}(Y=1 \mid Z=1, X=0) - \widehat{P}(Y=1 \mid Z=0, X=0)\}\widehat{P}(X=0) \\
&= (0.60 - 0.70) \times 0.5 + (0.20 - 0.30) \times 0.5 \\
&= -0.10 < 0.
\end{aligned}
$$

What are the assumptions we made?

- **Assumption:** *age* is the only confounder
  - *“Ignorability*” or *“selection on observables*” assumption
  - How do we know what we must condition on?
- **Assumption:** effect of *stationary biking* doesn’t depend on friends’ exercise
  - Stable Unit Treatment Value (SUTVA) assumption
  - Are there network effects?
- **Assumption:** our observations of exercise/no-exercise cover similar people
  - *“Common support”* or *“Overlap”* assumption
- **Also:** data is not covering all combinations of age and levels of exercise
  - Will our lessons generalize beyond the observed region?

A1: Ignorability

- Conditional Independence Assumption (CIA)
  - Under random experiments, $T \perp X$ for both observed and unobserved covariates
  - But conditioning and related techniques can only construct $T \perp X$ for observed covariates.
- So assume that after conditioning on observed covariates, any unmeasured covariates are irrelevant.

Ignorability

- Let $X = \{X_{obs}, X_{unobs}\}$
- Then $P(Y_T |\, X_{obs}) = P(Y_T | X_{obs}, T)$ $[where\ \mathrm{Y_T} = \mathrm{Y}|do(T)]$

A2. Stable Unit Treatment Value

The effect of treatment on an individual is independent of whether or not others are treated.

I.e., no spillover or network effects

SUTVA

$$P(Y_i|do(T_i, T_j)) = P(Y_i|do(T_i))$$

Example: What is the effect of giving a fax machine to an individual?

- It depends on whether or n

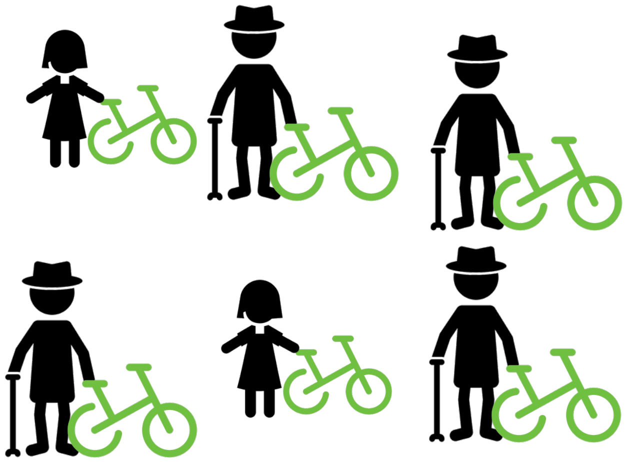
Do people here know / remember what a fax machine is?

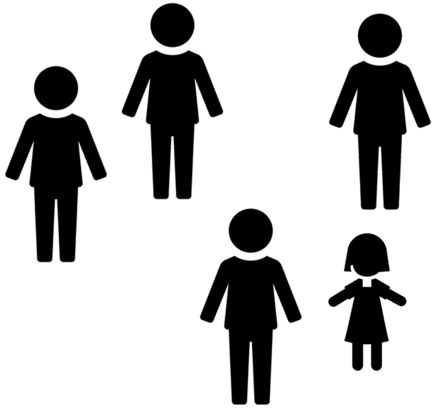
A3. Common support

- The treated and untreated populations have to be similar.
- That is, there should be overlap on observed covariates between treated and untreated individuals.
- Otherwise, cannot estimate counterfactual outcomes.

Common support

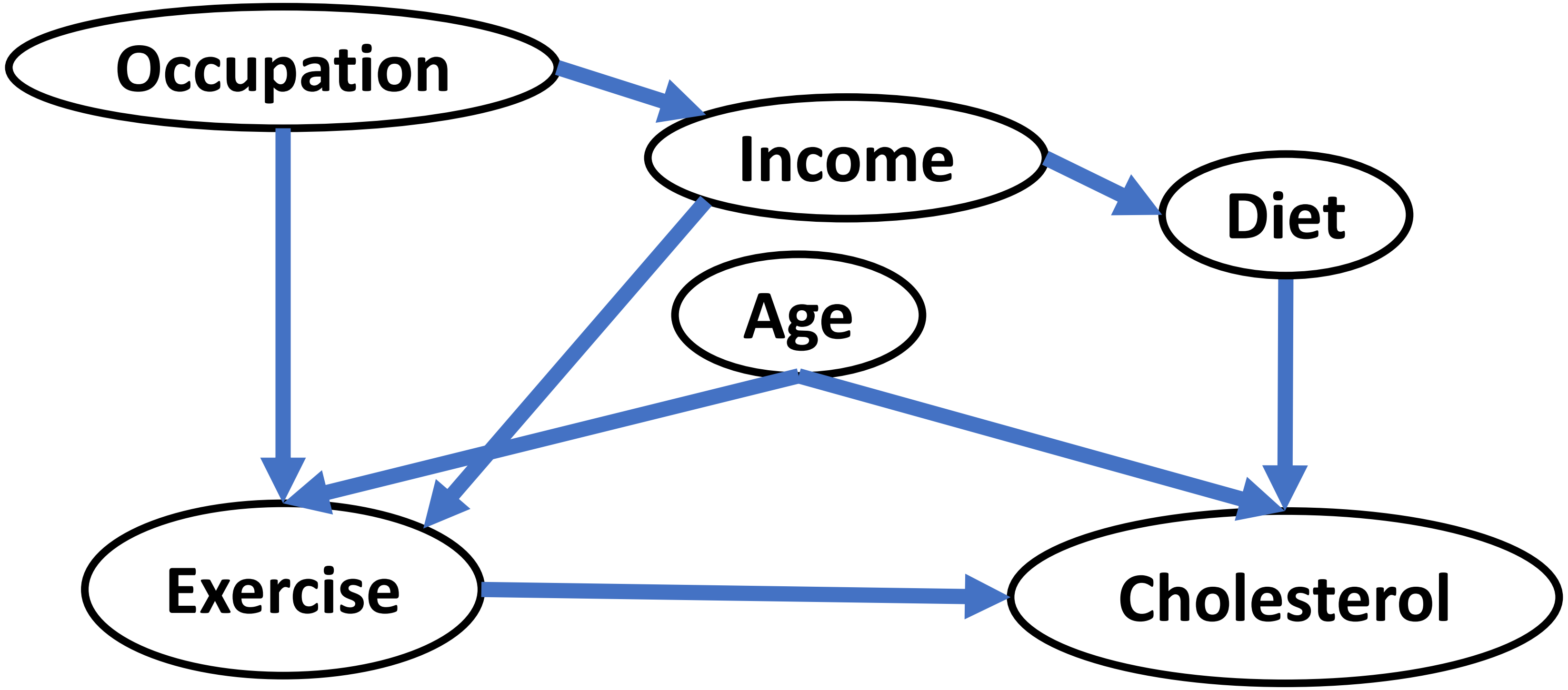
$$0 < P(T = 1|X = x) < 1$$





Advanced: How to know we have the right variables? *Backdoor criterion*

1. Use domain knowledge to build a model of the causal graph
2. Condition on enough variables to cover all backdoor paths



Caveat: Causal effect only if assumed graphical model is correct

What we just learned: Simple Conditioning

Definition Conditioning calculates treatment effects by identifying groups of individuals with the same covariates, where individuals in one group are treated and in the other group are not.

Intuition Conditioning our analysis of $T \rightarrow Y$ on X breaks the dependence between confounds X and the treatment T

Example In the cartoon relationship between exercise and cholesterol, age is a confounder, as it influences both levels of exercise and cholesterol.

By conditioning analysis on age, we can identify the effect of exercise.

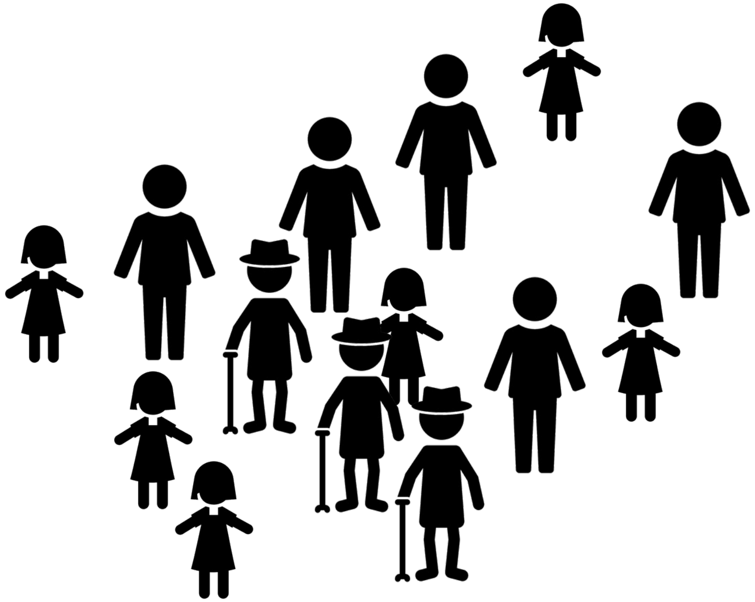
Keep in mind How do we know what to condition on?

Grouping becomes harder as dimensionality of X increases

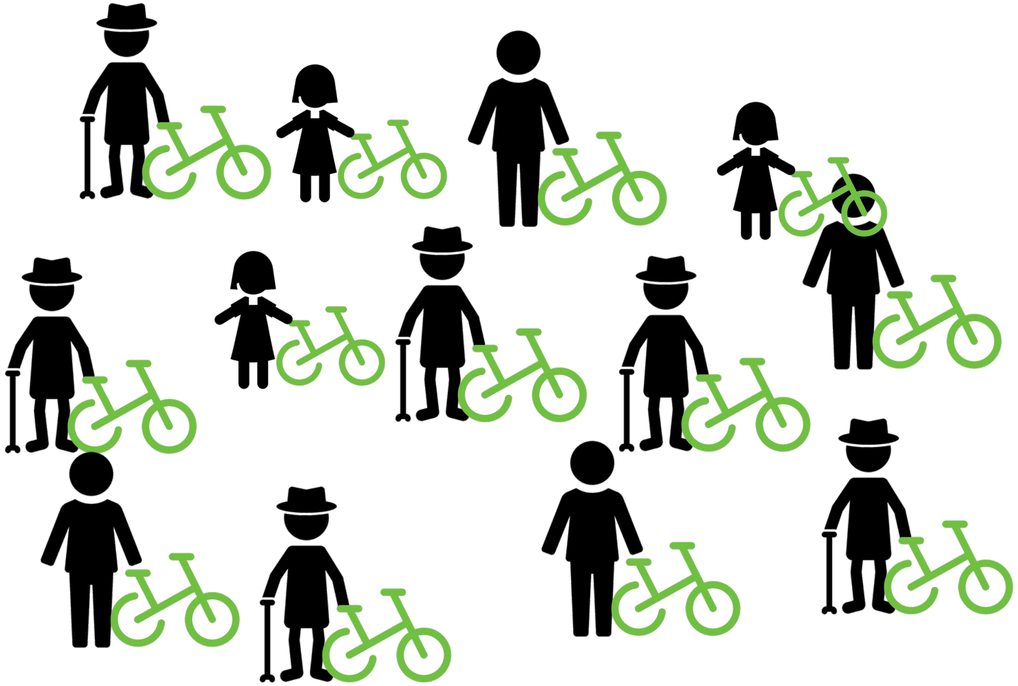
Part II.A. Observational Studies

“Simulating randomized experiments"

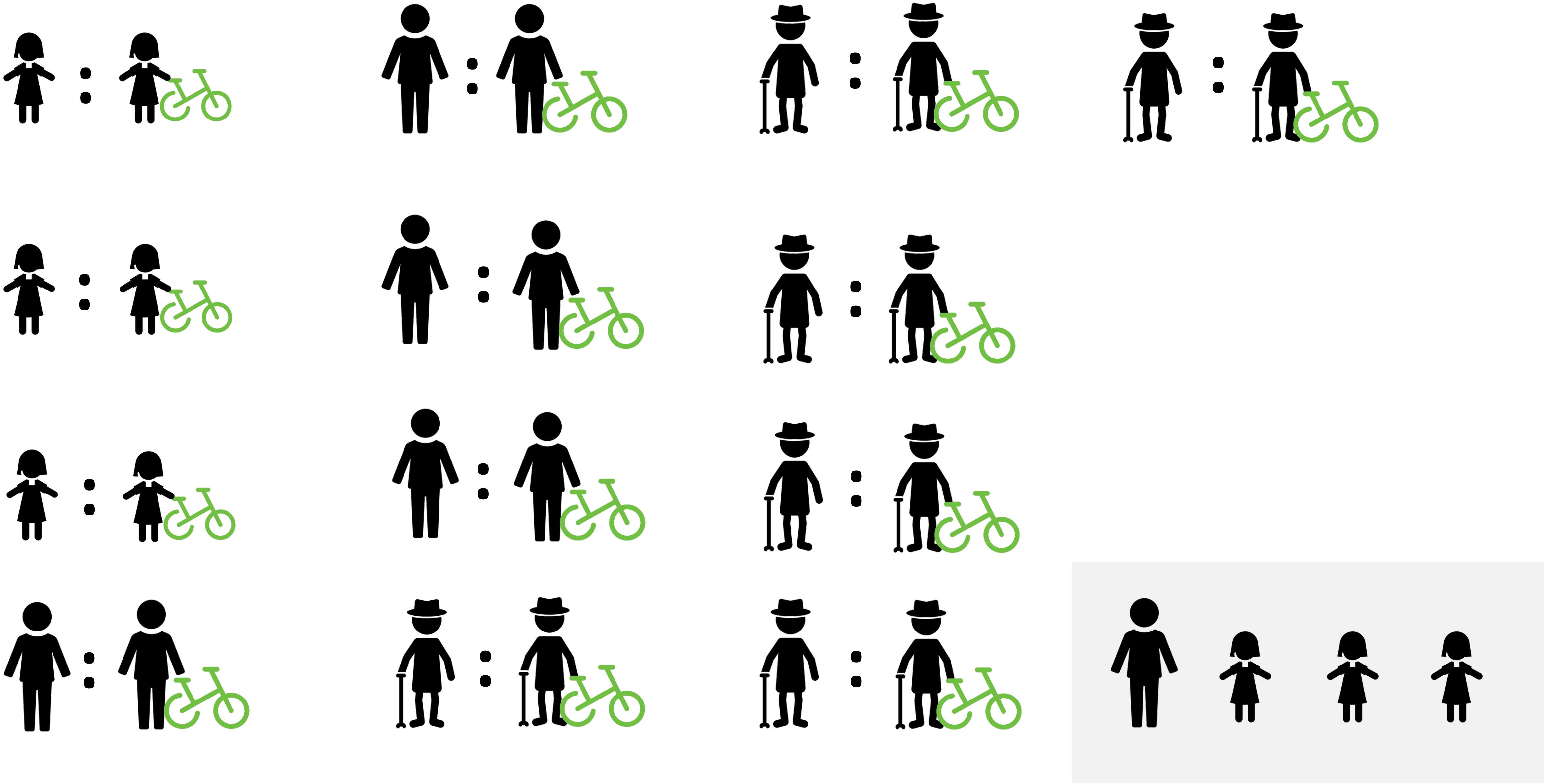
- Conditioning on Key Variables
- **Matching and Stratification**
- Weighting
- Simple Regression
- Doubly Robust
- Synthetic Controls



Avg Cholesterol = 200



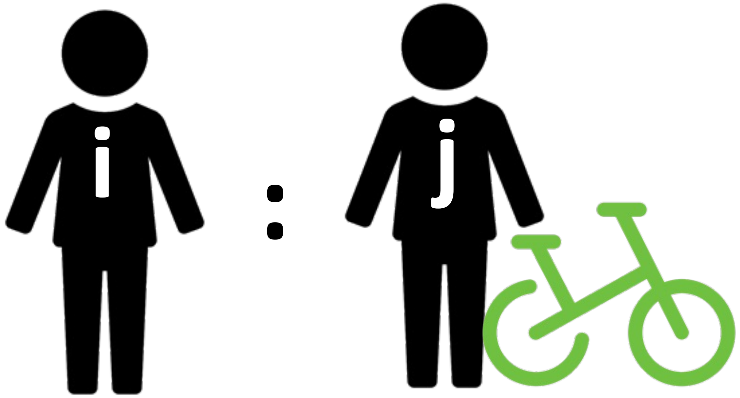
Avg Cholesterol = 206



Matching

Identify pairs of treated and untreated individuals who are very similar or even identical to each other

Very similar ::= $Distance(X_i, X_j) < \epsilon$



Paired individuals provide the counterfactual estimate for each other.

Average the difference in outcomes within pairs to calculate the *average-treatment-effect on the treated*

Exact Match

Simple:

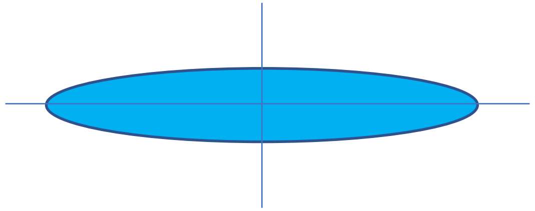
$$Distance(\vec{x}_i, \vec{x}_j) = \begin{cases} 0, & \vec{x}_i = \vec{x}_j \\ \infty, & \vec{x}_i \neq \vec{x}_j \end{cases}$$

Use this in low-dimensional settings when overlap is abundant

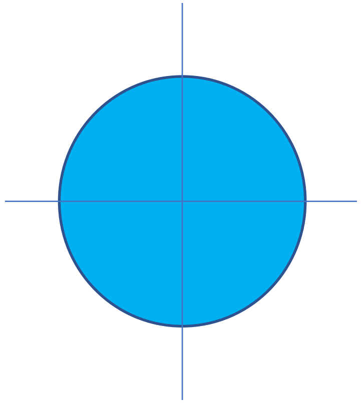
But in most cases, there will be too few exact matches …

Mahalanobis Distance

Mahalanobis distance accounts for unit differences by normalizing each dimension by the standard deviation.



$$Mahalanobis(\overrightarrow{x_i}, \overrightarrow{x_j}) = \sqrt{(\overrightarrow{x_i} - \overrightarrow{x_j})^T S^{-1} (\overrightarrow{x_i} - \overrightarrow{x_j})}$$



And S is the covariance matrix of some distributions that x_i and x_j follows.

Propensity Score

Propensity score is an individual's *propensity to be treated*

$$\hat{e}(X) = P(T = 1|X)$$

- Propensity scores are estimated or modeled, *not observed*.
- Rare exception is if you know likelihood of randomized assignment

Breaks influence of confound X, allowing estimate of $T \rightarrow Y$

Propensity scores subdivide observational data s.t. $T \perp\!\!\!\perp X \mid score$

How to match with propensity score

1. Train a machine learning model to predict treatment status
   - **Supervised learning:** We are trying to predict a known label (treatment status) based on observed covariates.
   - Conventionally, use a logistical regression model, but SVM, GAMs, are fine
   - But score must be well-calibrated. I.e., $(100 * p)\%$ of individuals with score of p are observed to be treated

2. Distance is the difference between propensity scores

$$Distance\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = |\hat{e}(\overrightarrow{x_i}) - \hat{e}\left(\overrightarrow{x_j}\right)|$$

Propensity score, FAQ

Q: Wait, why does this work?

A: Individuals with similar covariates get similar scores, and all individuals mapped to a similar score have similar treatment likelihoods.

Q: What if my propensity score is not accurate? (i.e., can't tell who is treated)

A: That's ok. The role of the model is to balance covariates given a score; not to actually identify treated and untreated.

Q: What if my propensity score is very accurate? (i.e., *can* tell who is treated)

A: Means we cannot disentangle covariates from treatment status. Any effect we observe could be due either to the treatment or to the correlated covariate.

Consider redefining the treatment or general problem statement. Don't dumb down model!

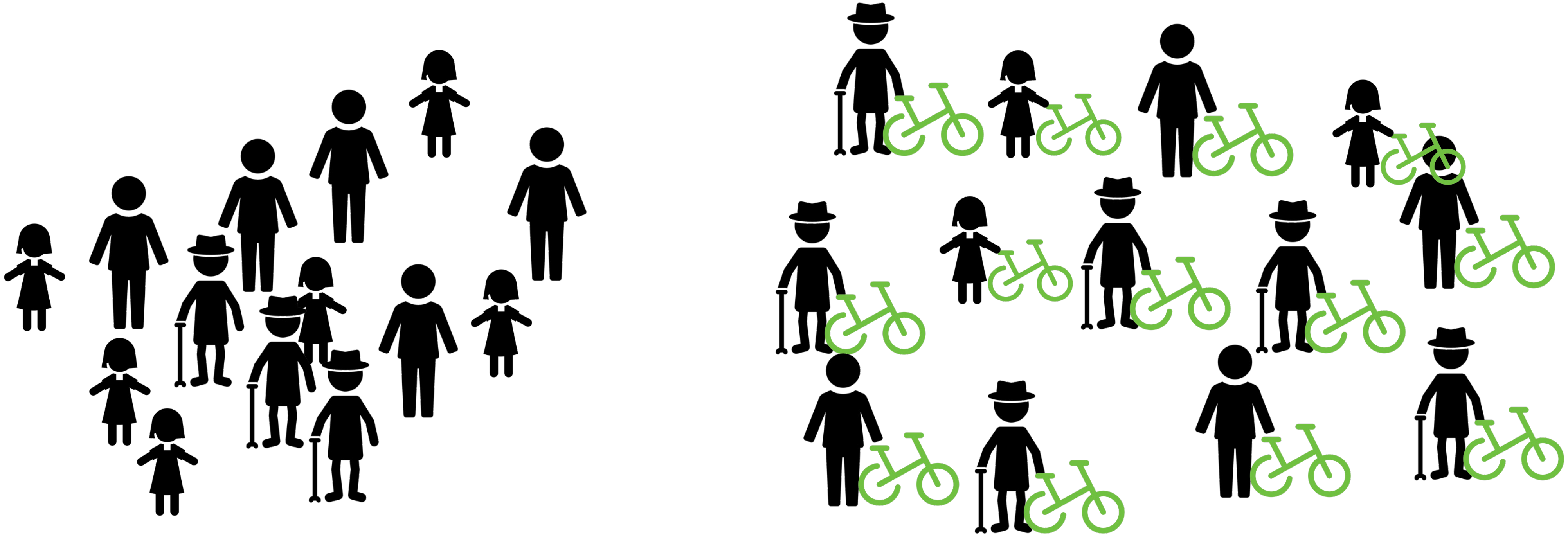
What we just learned: Matching

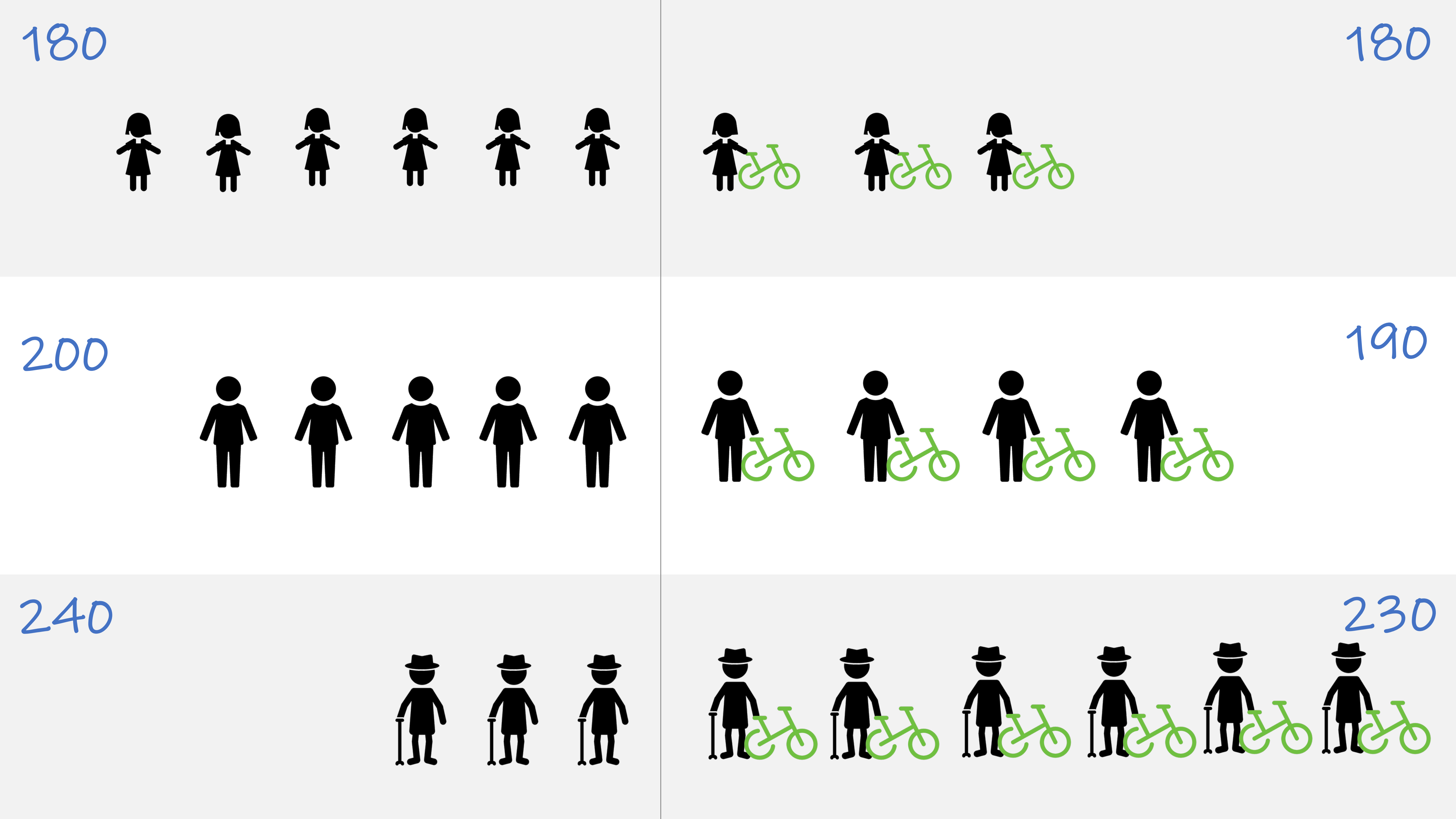
Definition Matching calculates treatment effects by identifying pairs of similar individuals, where one is treated and the other is not.

Intuition The paired individuals stand-in as the counterfactual observations for one another.

Example In our cartoon, we create pairs of individuals matched exactly on their age. More generally, we can use Mahalanobis distance or propensity score matching to find similar individuals to be matched.

Keep in mind Matching calculates the treatment effect on the treated population. We do not know what might happen if people who would never get treatment are suddenly treated.




180
180
200
190
240
230

From Matching to Stratification

- $1:1$ matching generalizes to *many:many* matching.
- Stratification identifies paired *subpopulations* whose covariate distributions are similar.
- There can still be error, if strata are too large.

How to stratify with propensity score

1. Train a machine learning model to predict treatment status
   - **Supervised learning:** We are trying to predict a known label (treatment status) based on observed covariates.
   - Conventionally, use a logistical regression model, but SVM, GAMs, are fine
   - But score must be well-calibrated. I.e., $(100 * p)\%$ of individuals with score of p are observed to be treated

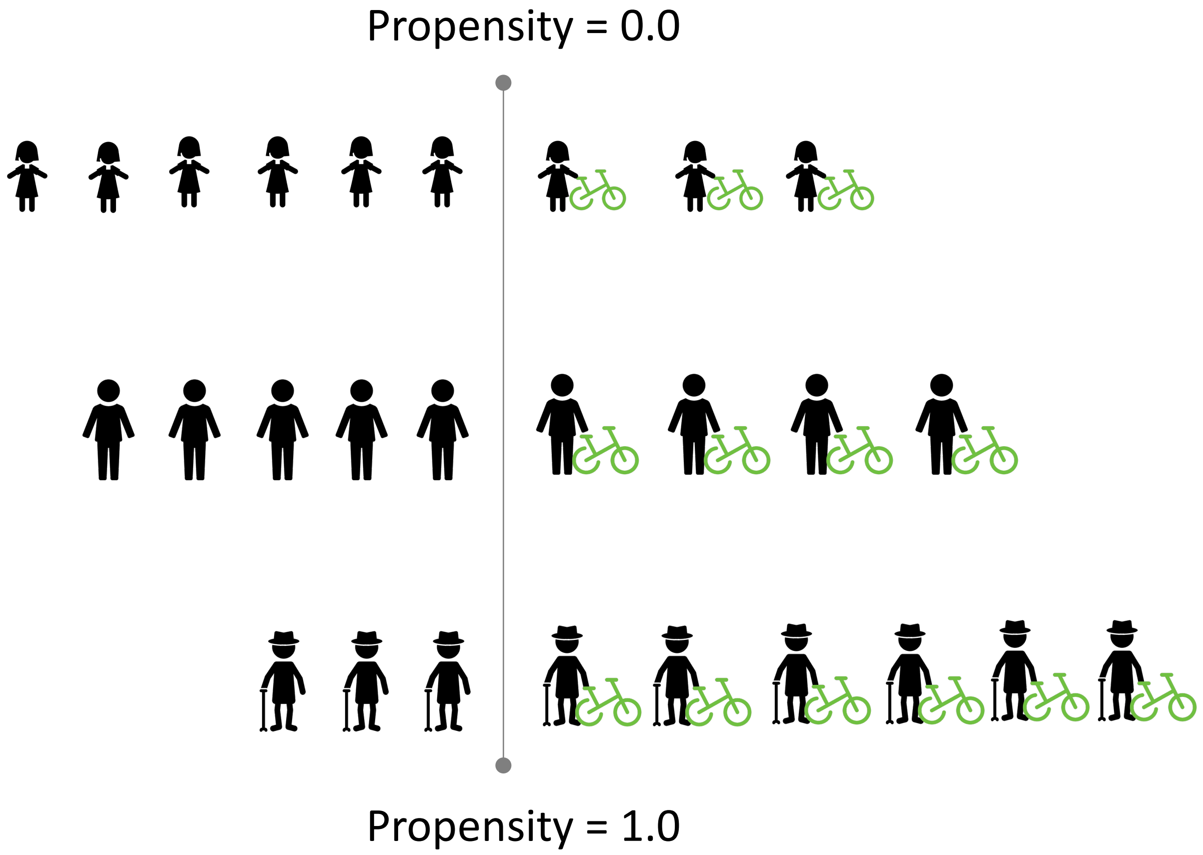
2. Distance is the difference between propensity scores

$$Distance\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = |\hat{e}(\overrightarrow{x_i}) - \hat{e}\left(\overrightarrow{x_j}\right)|$$

Propensity Score Stratification

We can use propensity score to stratify populations

1. Calculate propensity scores per individual as in matching.
2. But instead of matching, stratify based on score.
3. Calculate average treatment effect as weighted average of outcome differences per strata.
4. Weight by number of treated in the population for ATE on treated.



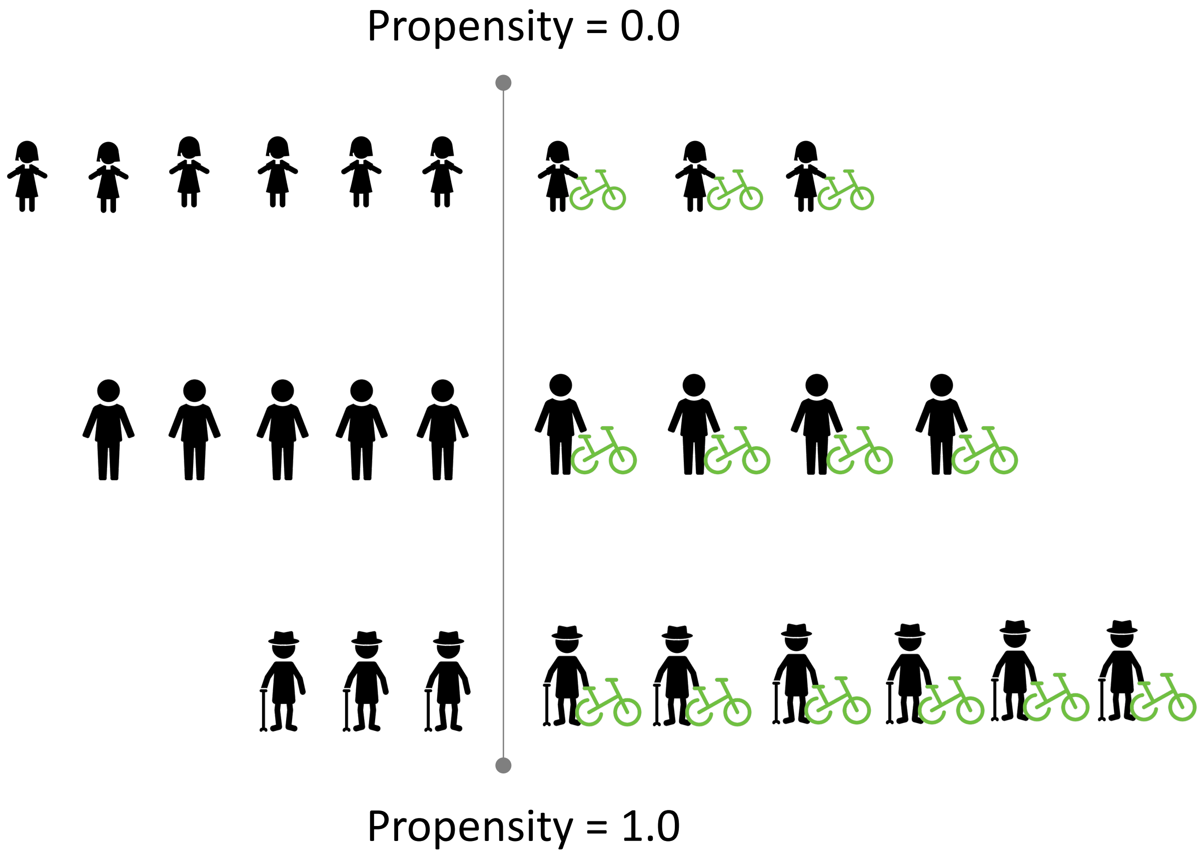
Propensity Score Stratification

$$ATE = \sum_{s \in strata} \frac{1}{N_{s,T=1}} \left(\bar{Y}_{s,T=1} - \bar{Y}_{s,T=0}\right)$$

where,

$\bar{Y}_{s,T}$ is the average outcome at strata s and treatment status T

And $N_{s,T=1}$ is the number of treated individuals in strata s



P.S. Stratification, Practical Considerations

- How many strata do we pick?
  - Scale will depend on data. Want each stratum to have enough data in it.
  - Conventional, small-data literature (e.g., ~100 data points) picked 5.
  - With 10k to 1M or more data points, I pick 100 to 1000 strata.
  - Set strata boundaries to split observed population evenly
  - Aside: why not always pick a small number of strata? It’s a bias-variance trade-off…

- What if there aren’t enough treated or untreated individuals in some of my stratum to make a meaningful comparison?
  - This often happens near propensity score 0.0 and near 1.0
  - Drop (“Clip”) these strata from analysis. Technically, you are now calculating a local-average-treatment-effect.

What we just learned: Stratification

Definition Stratification calculates treatment effects by identifying groups of individuals with similar distributions of covariates, where individuals in one group are treated and in the other group are not.

Intuition The difference in average outcome of paired *groups* tells us the effect of the treatment on that subpopulation. Observed confounds are balanced, due to covariate similarity across paired groups.

Example In our cartoon example, we stratified based on propensity score into 3 strata. ATE is the weighted sum of differences in avg outcomes in each strata.

Keep in mind Make sure there are enough comparable individuals in each strata