Causal Inference II

Methods for Causal Inference

CS547 Machine Learning for Big Data

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Course Evaluation Announcement

- Course evaluation is out
  - https://uw.iasystem.org/survey/224031
  - Also see link on Ed (pinned)
  - Please fill out the form before June 7. Thanks!!!
- We appreciate your feedback!

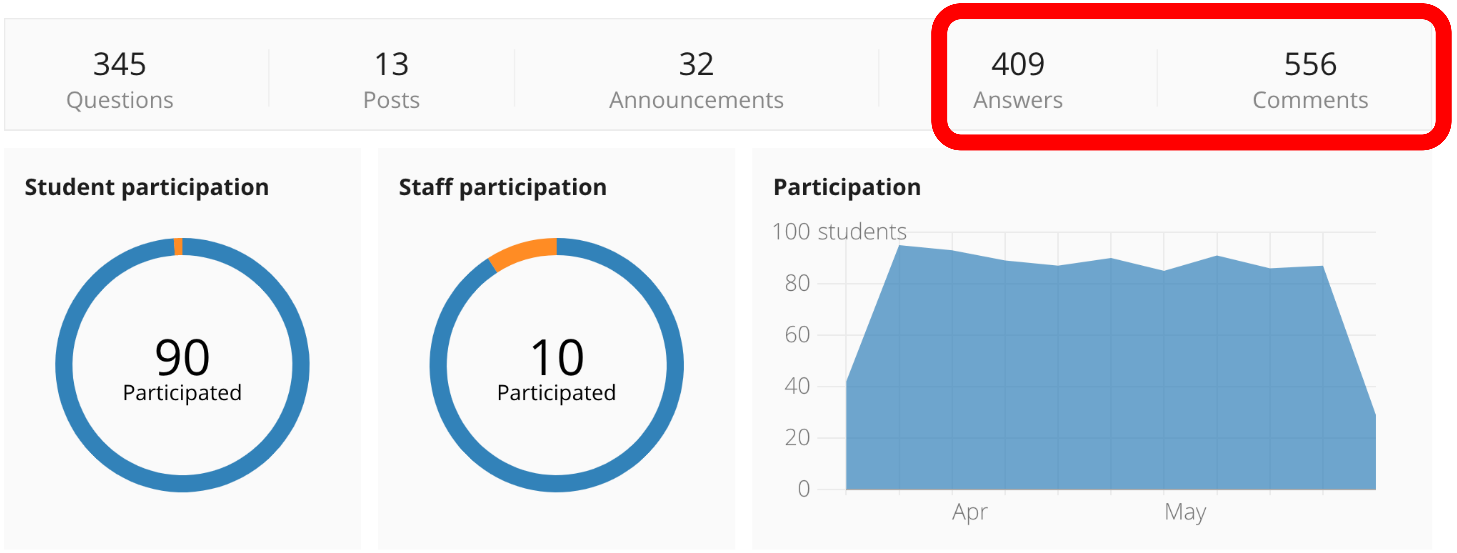
Acknowledgements

Summary of changes

- Less homework problems (implemented already)
- Colab 9 optional – extra credit only
- Final Project Presentation optional – extra credit only
  - We assume you will present iff presentation is uploaded.
- Final Project Report still mandatory and due by Sunday night

- Bottom line: If you skip anything your grade will not change. Whatever you do will only improve it.

Thanks!



345 Questions

13 Posts

32 Announcements

409 Answers

556 Comments

Student participation

90 Participated

Staff participation

10 Participated

Participation

Teaching Assistants

Ashish Sharma
(Head TA)

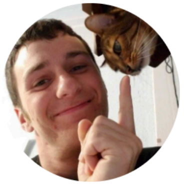
Kristof Glauninger

Qifan Huang

Stephen J. Jonany

Jack Khuu

Alon Milchgrub



Galen Weld

Chi-Hui Yen

Plan for today: Methods for Causal Inference

Observational Studies

How to simulate randomized experiments?

Sensitivity Analyses

How to consider violations of assumptions?

Example

Example: Treatment, Outcome and Confound

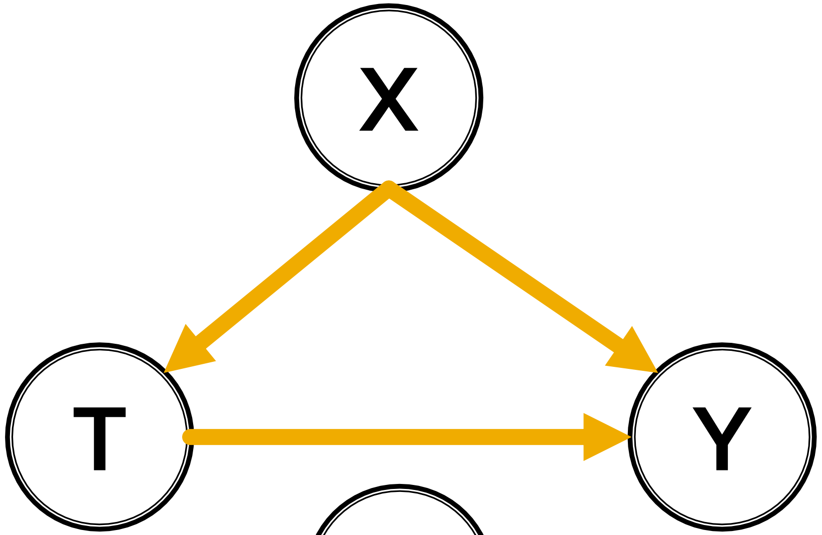
Goal: Estimate effect of a treatment T on an outcome Y

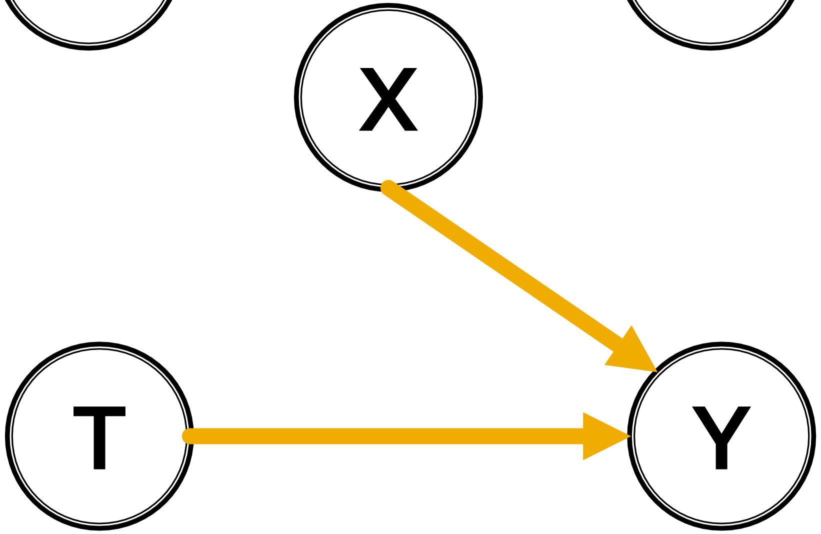
But, confound X influences both T and Y

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

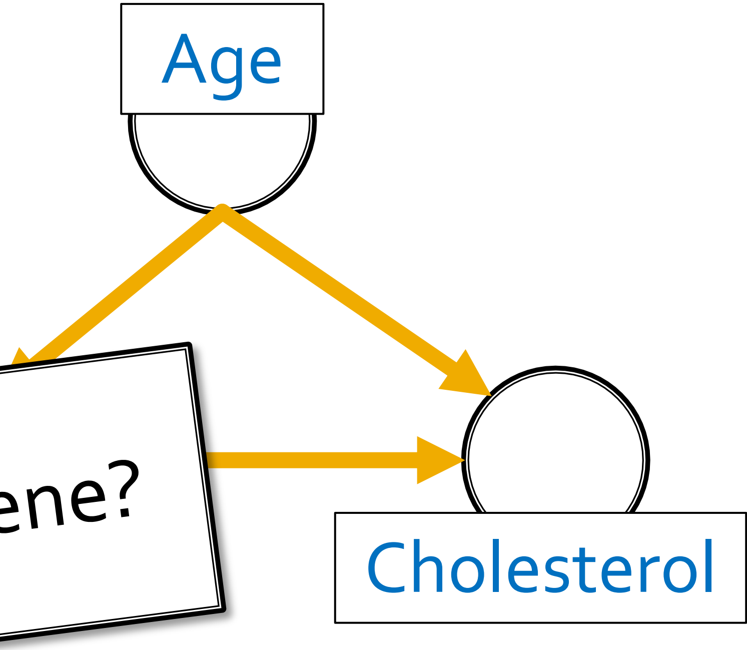
Example: Exercise, Cholesterol, and Age

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 (th

exercise ⫫



But what if we cannot actively intervene?

Randomize **eriments** actively assign exercise independent of any age

Thus, by construction: exercise ⫫ age

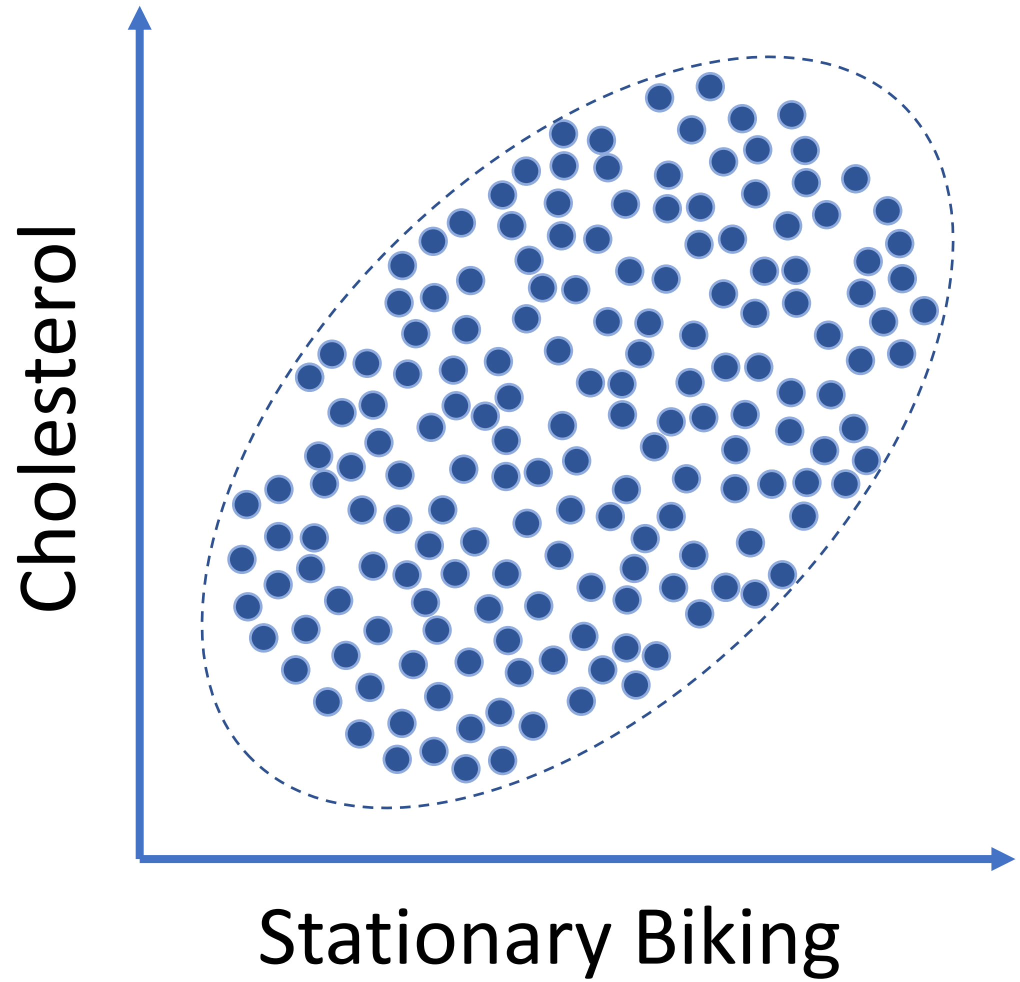
Observational Studies Methods

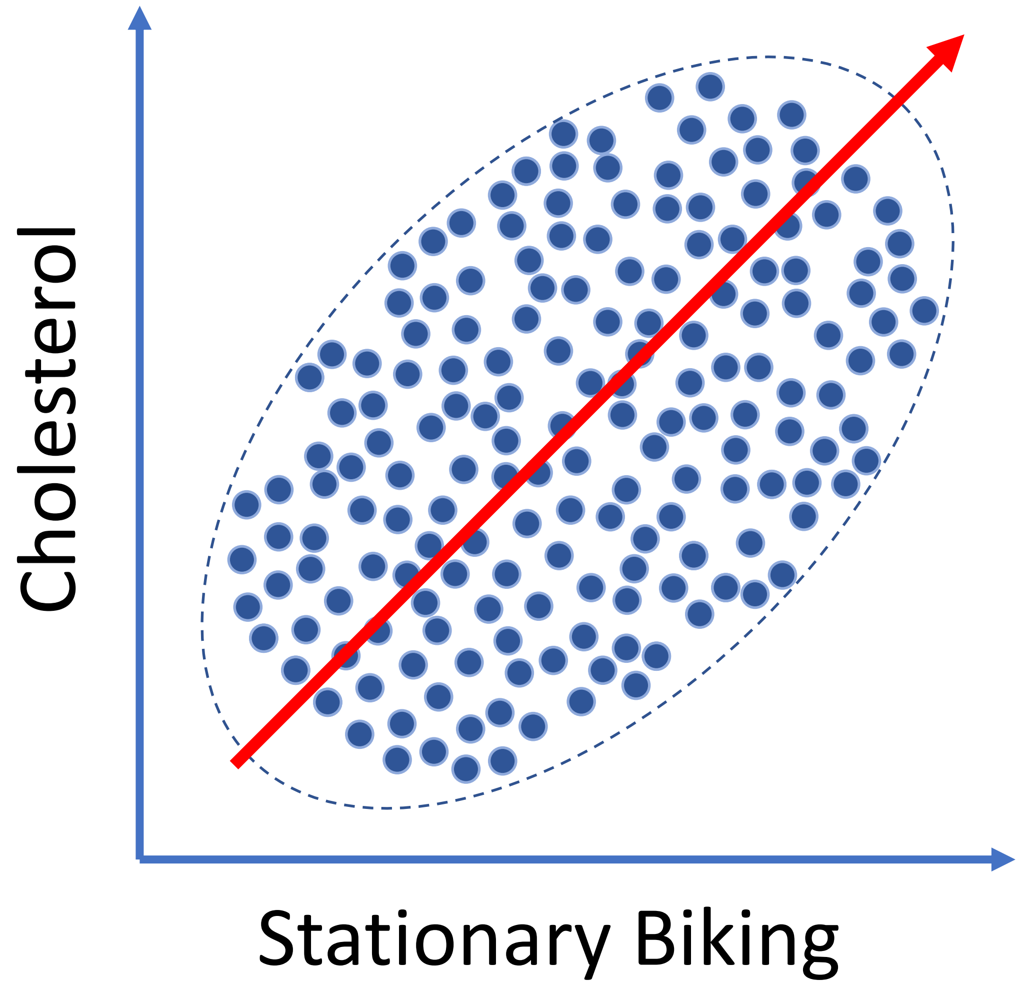
“Simulating randomized experiments"

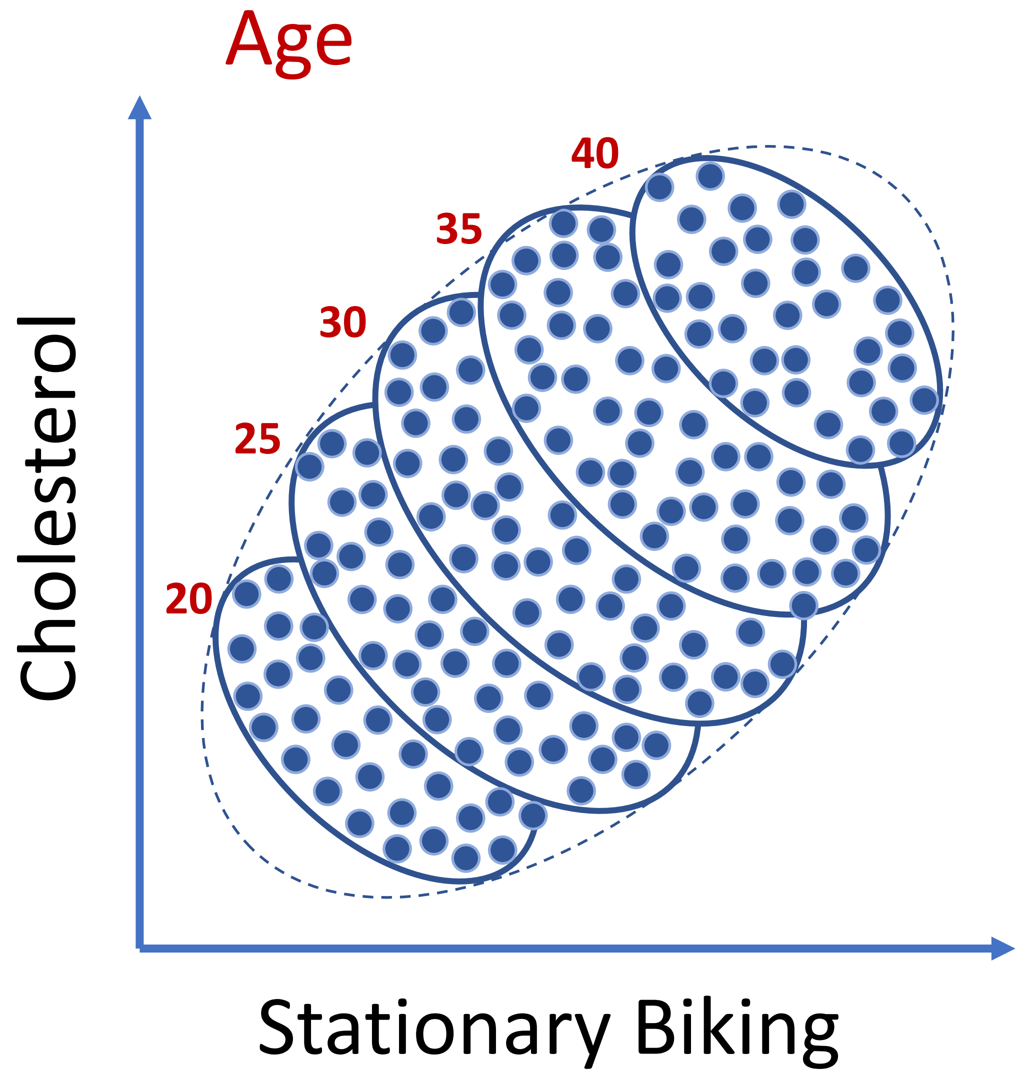
- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression

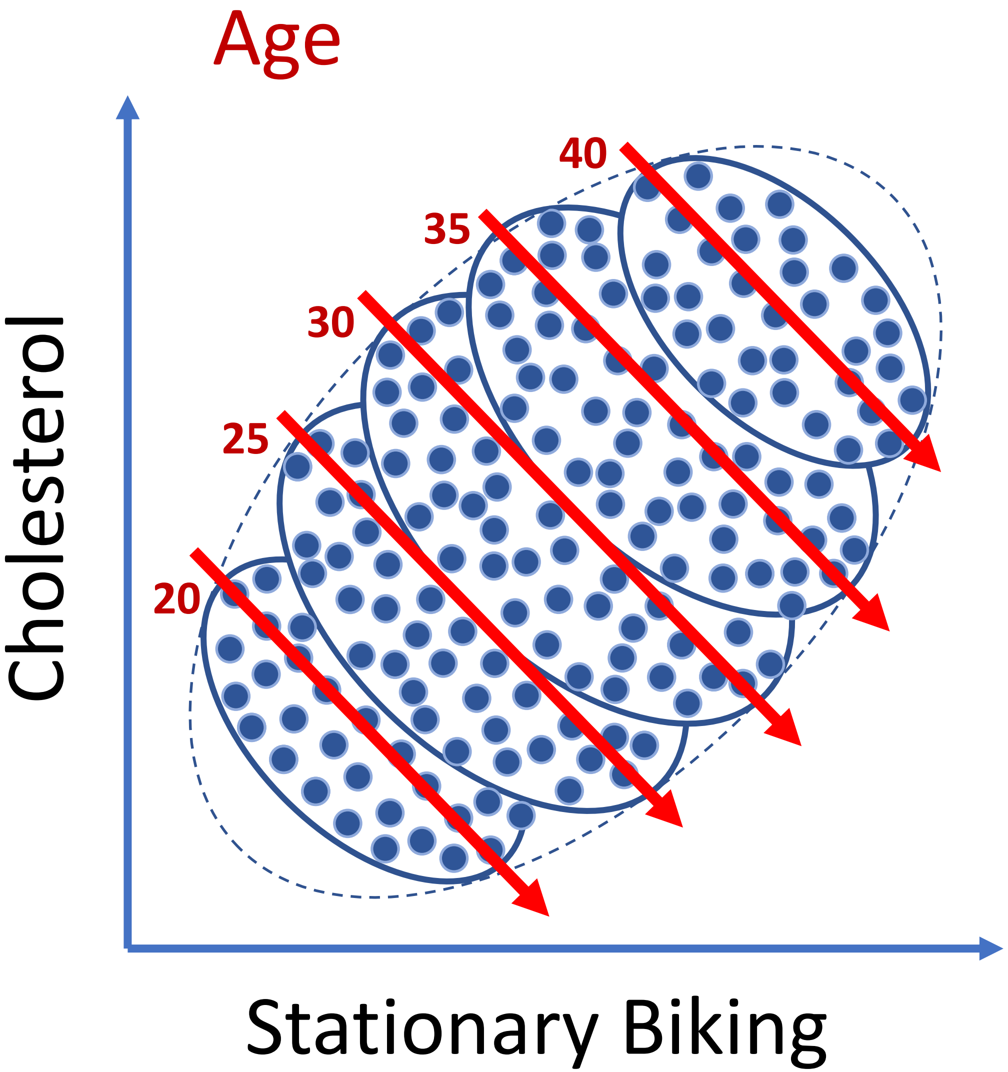
“Simulating randomized experiments"

- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression


Cholesterol
Stationary Biking


Cholesterol
Stationary Biking


Age
40
35
30
25
20
Cholesterol
Stationary Biking


Age
40
35
30
25
20
Cholesterol
Stationary Biking

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:

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

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 (and not for unobserved covariates!)
- So we have to assume that after conditioning on observed covariates, any unmeasured covariates are irrelevant.

- **Ignorability**
  - $(Y_1, Y_0) \perp T \mid X = x$ for all x $[where\ \mathrm{Y_T} = \mathrm{Y}|do(T)]$

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

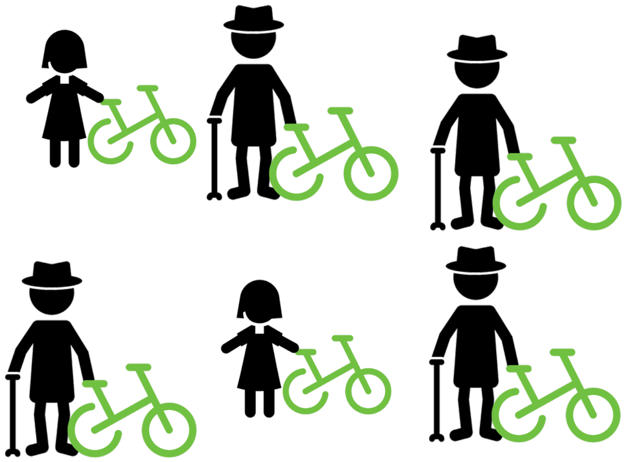
SUTVA

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

Example: What is the effect of giving a fax machine to an individual?
It depends on whether or not other people have fax machines!

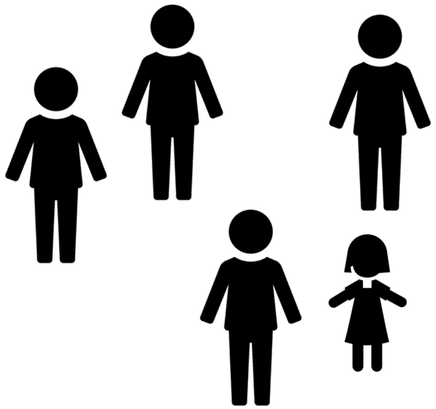
A3. Common support

- 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.



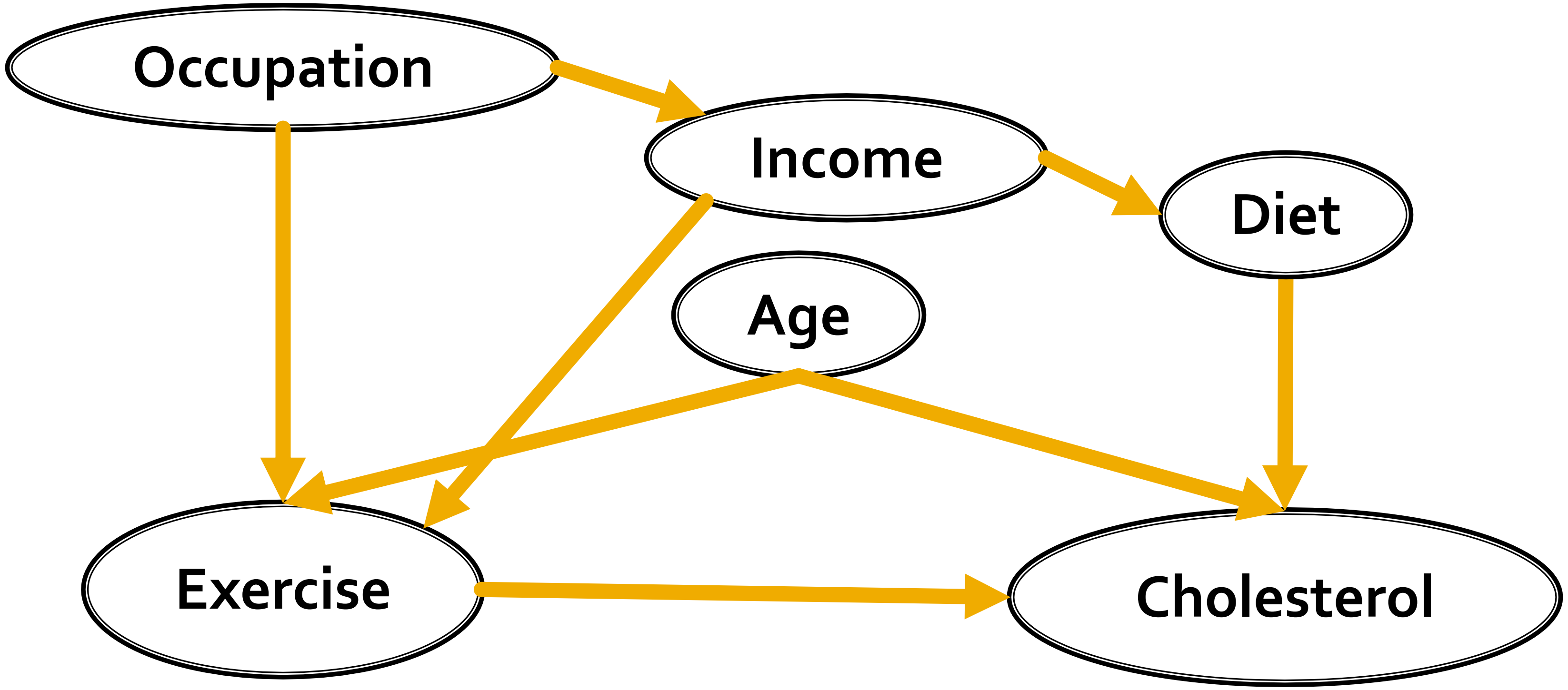
Common support

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



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.

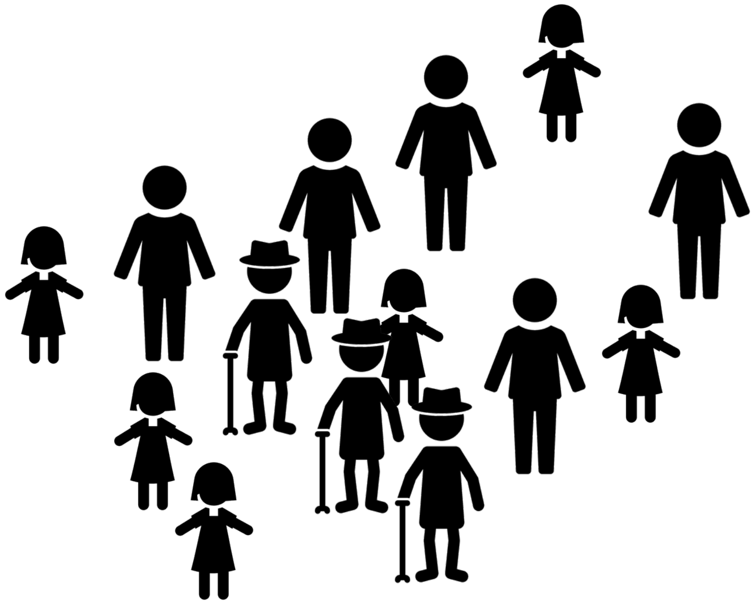
By conditioning analysis on age, we can identify the effect of exercise.

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

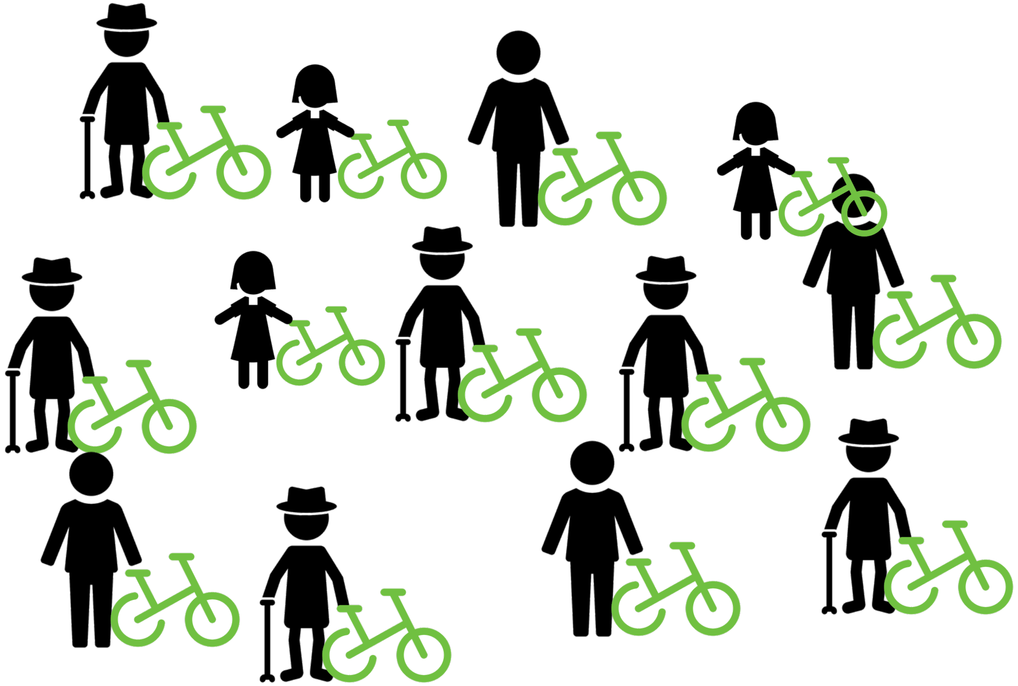
Grouping becomes harder as dimensionality of X increases

“Simulating randomized experiments"

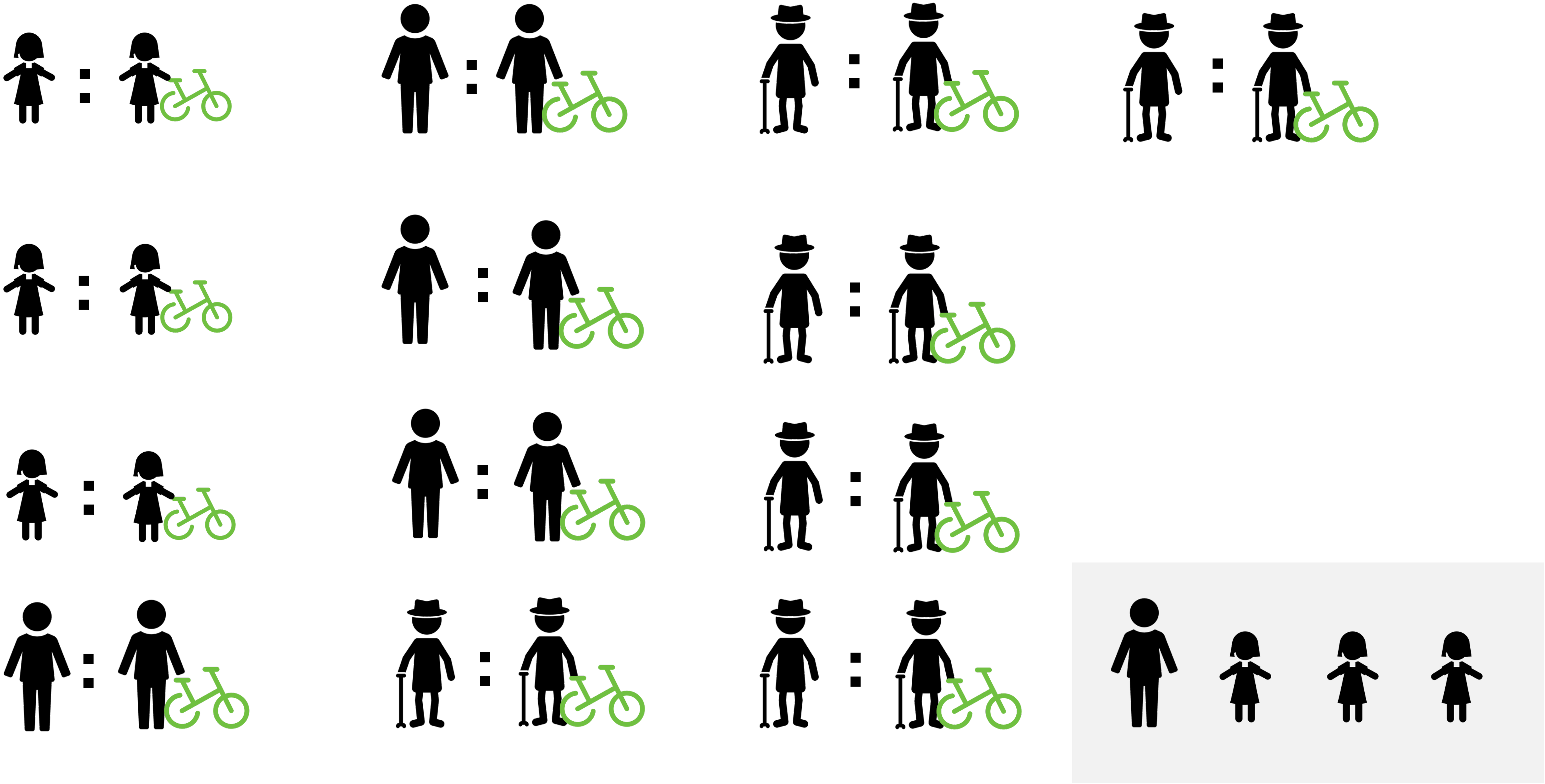
- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression



Avg Cholesterol = 200



Avg Cholesterol = 206



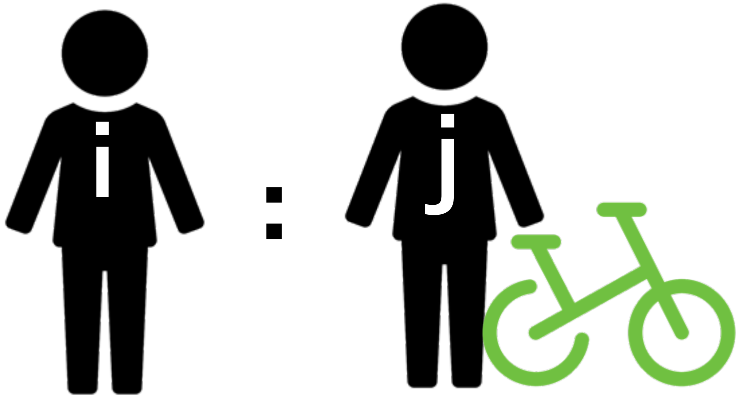
Matching

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 (ATT)*



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

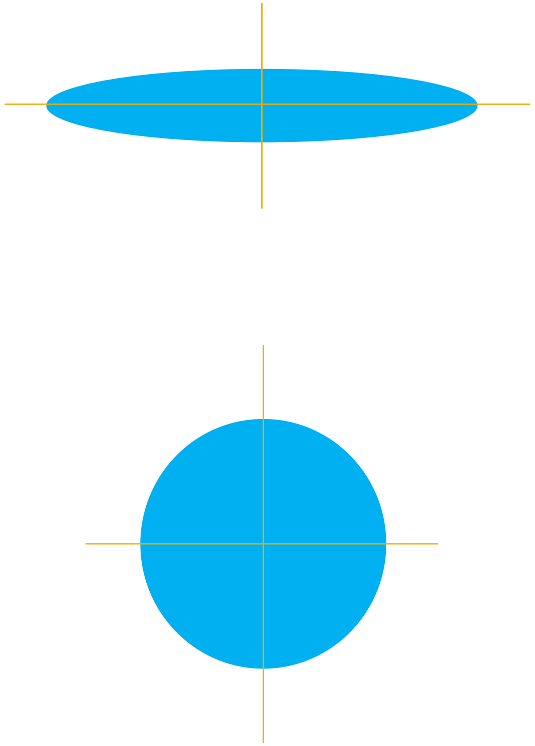
But in most cases, there will be too few exact matches …

Reminder: 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.



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 treatment assignment

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

→ **Conditioning on propensity score breaks influence of confound X, allowing estimate of $\boldsymbol{T \rightarrow Y}$**

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, GAM, NN 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: Big problem! Means we cannot disentangle covariates from treatment status. Treated and untreated units are too different. 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!

Advanced: Matching

- When matching, should we allow replacement?
  - It’s a bias / variance trade-off
- When matching, what if nearest neighbor is far away?
  - Use a caliper threshold to limit acceptable distance
- What if not all treated individuals are matched to untreated?
  - This will bias results. Consider redefining original cohort / population to cleanly exclude treated who won’t have matches in untreated population.
- What if treatment is not binary?
  - Advanced variants allow multi-dose, and other treatment regimens

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.

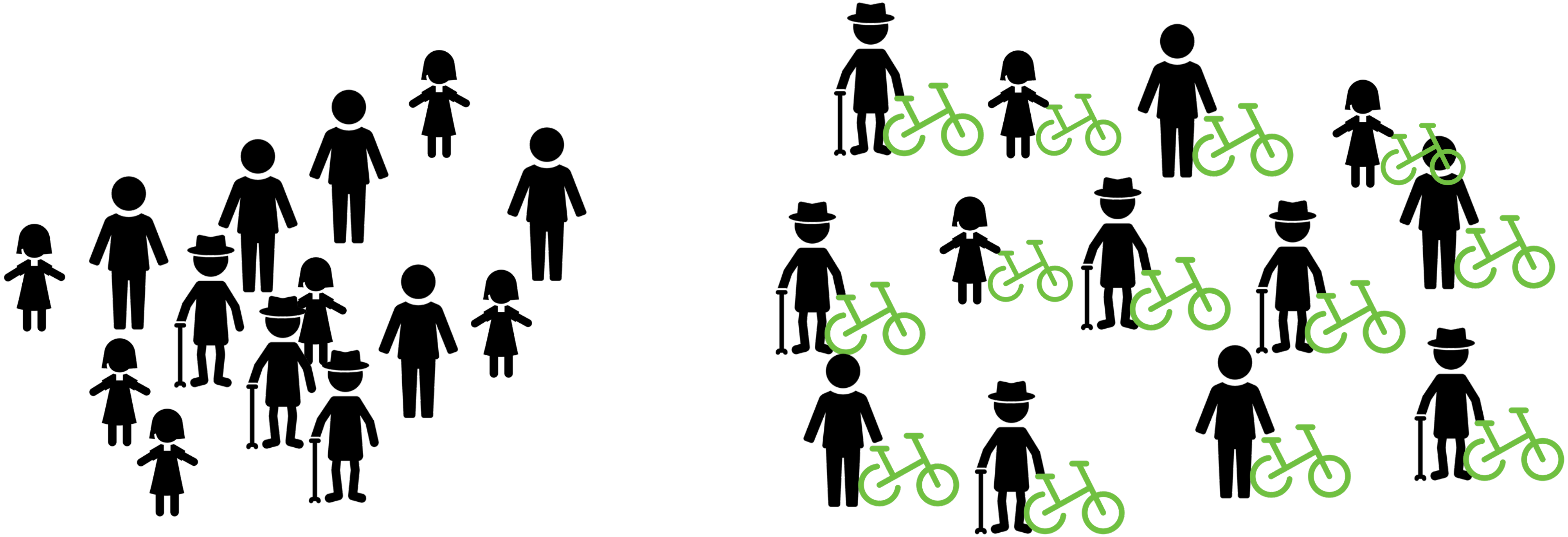
Intuition The paired individuals stand-in as the counterfactual observations for one another.

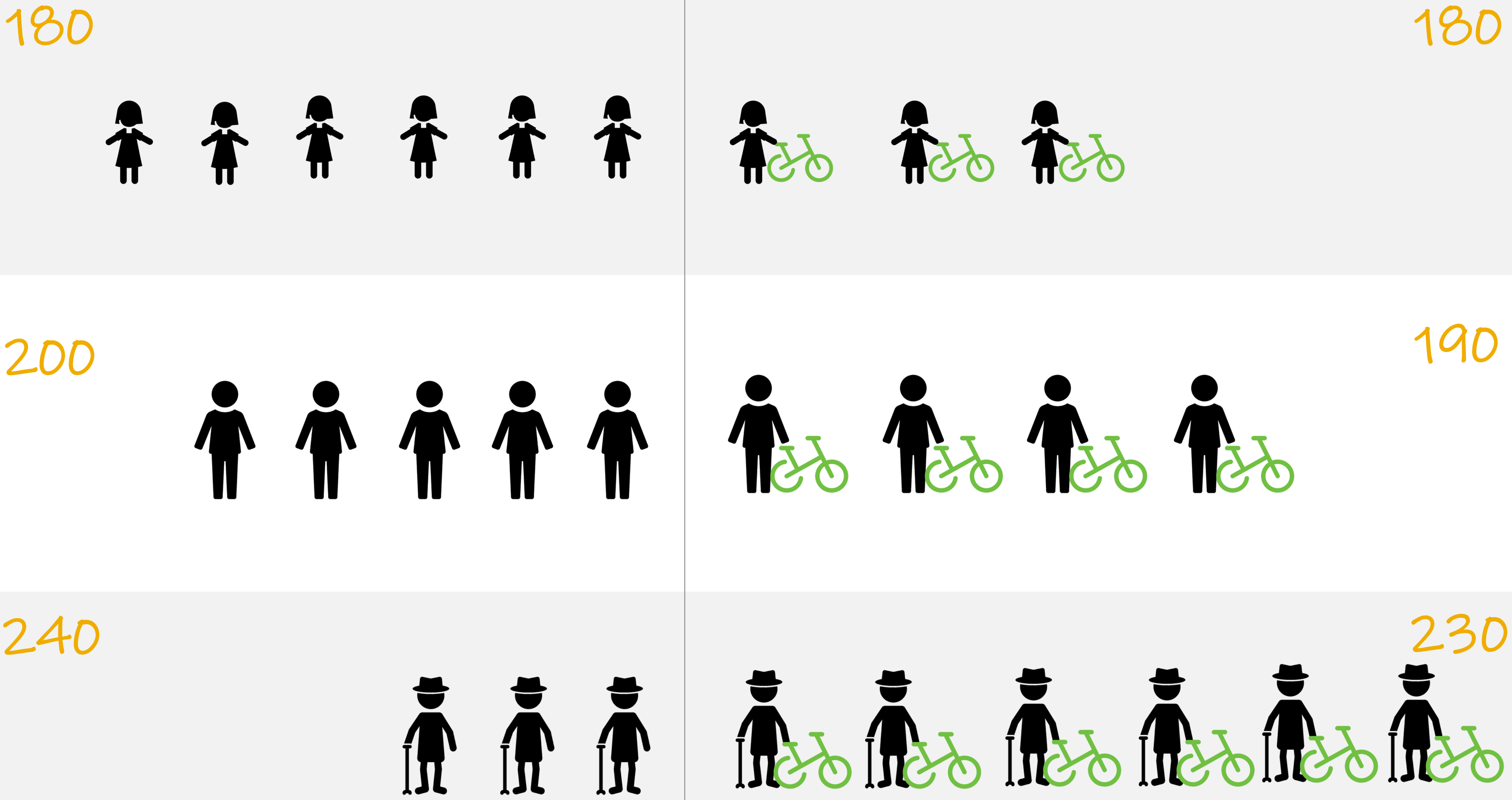
Example In our cartoon, we create pairs of individuals matched exactly on their age. More generally, we can use Mahalanobis distance or propensity score matching to find similar individuals to be matched.

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

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.

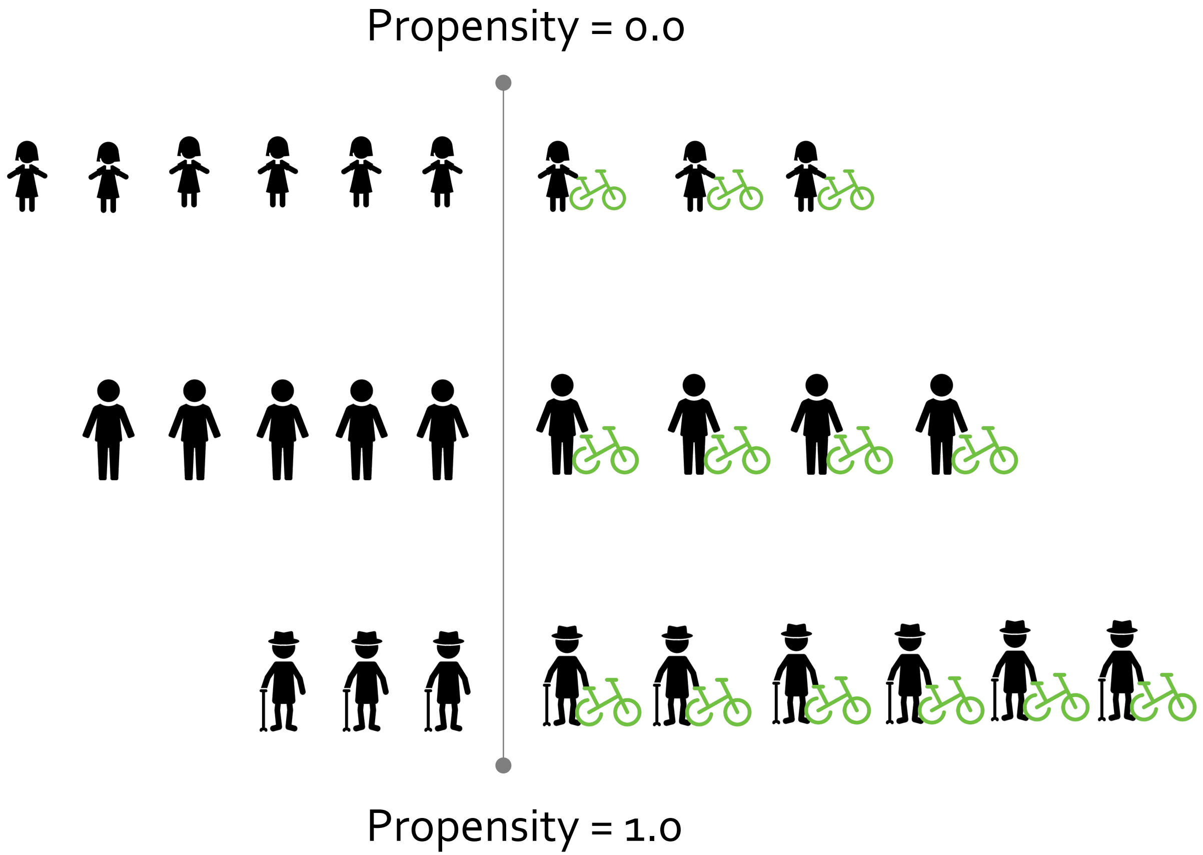



180
180
200
190
240
230

Propensity Score Stratification

We can use propensity score to stratify populations

1. Calculate propensity scores per individual as in matching (supervised ML problem)
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 ATT

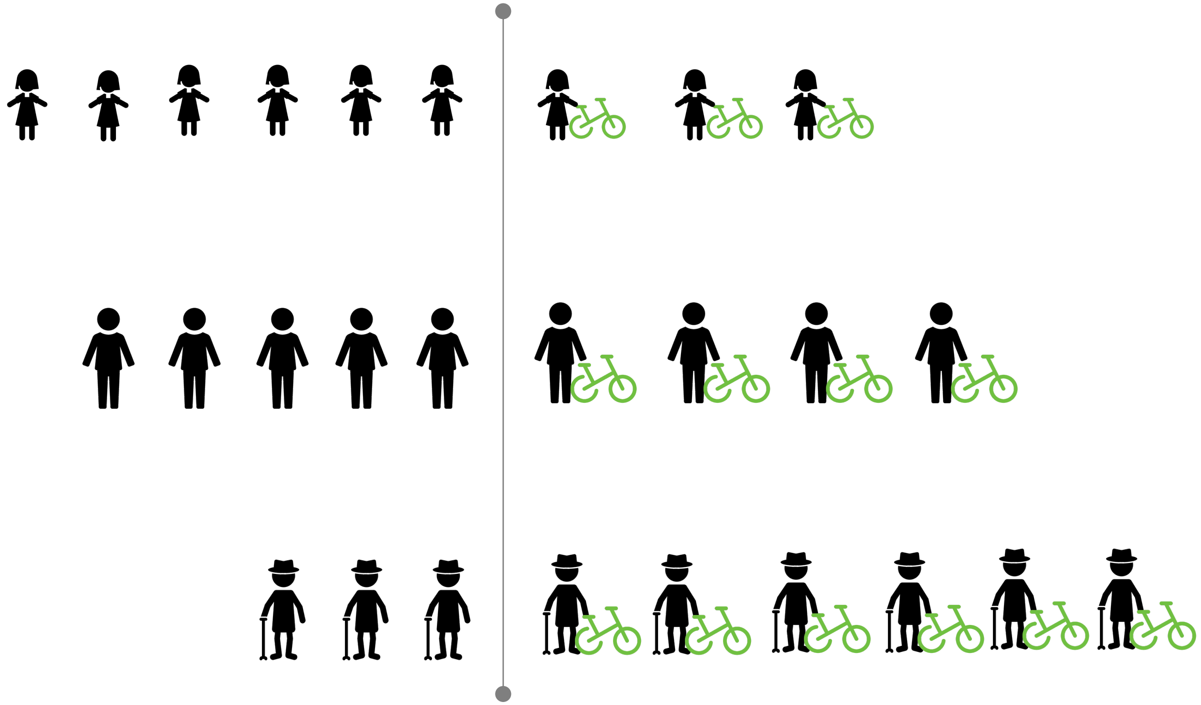


Propensity Score Stratification

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

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

Propensity = 0.0



Propensity = 1.0

PScore 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, cans pick 100 to 1000 strata.
  - Set strata boundaries to split observed population evenly
  - Aside: why not always pick a small number of strata?

- 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
  - This challenges our “overlap” assumption.
  - You can drop (“Clip”) these strata from analysis. But technically, you are now calculating a local-average-treatment-effect.
  - Better: Redefine population to avoid this issue.

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. ATT 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

“Simulating randomized experiments"

- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression

Weighting: An alternative to conditioning

What if we assign weights to observations to simulate randomized experiment?

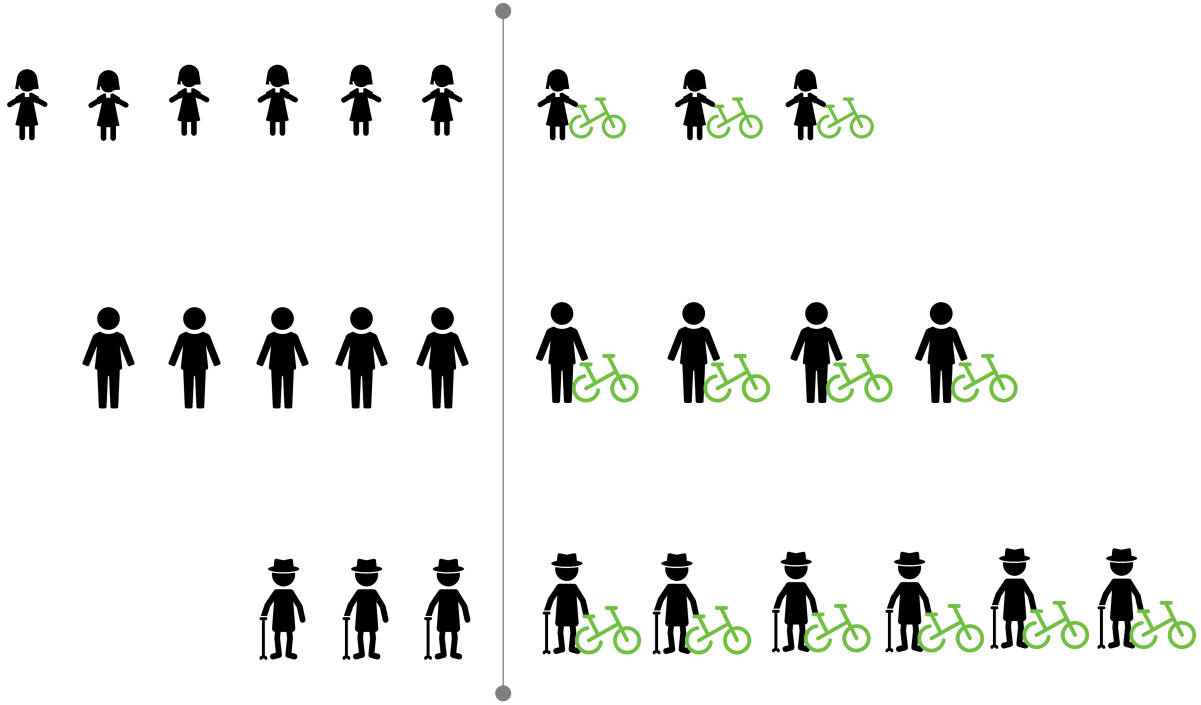
Stratification weights strata results by number of treated ~

Weighting by treated population ~ weighting by propensity score.

<u>Generalized weighting</u>: Calculate effect by weighted sum over all individual outcomes

Many weighting methods to generate a balanced dataset

Propensity = 0.0



Propensity = 1.0

Weighting

$$ATE = \frac{1}{N_{T=1}} \sum_{i \in treated} w_i Y_i - \frac{1}{N_{T=0}} \sum_{j \in untreated} w_j Y_j$$

Inverse Probability of Treatment Weighting (IPTW) for ATE

$$w_i = \frac{T}{e} + \frac{1-T}{1-e};$$

e is propensity score

$$\mathrm{N_{T=1}} = \sum T; \quad N_{T=0} = \sum 1 - T$$

Weighting: Caveats and Practical notes

- High variance when propensity score e is close to 0 or 1
  A single value can derail the estimate!
- Many heuristics for clipping weights; stabilizing weights; etc.
- Assumes propensity score model is correctly specified (i.e., that e is correctly estimated for all individuals)

- Variants of weighting: Calculate average treatment effect on treated (ATT) instead of ATE

What we just learned: Weighting

Definition Weighting calculates average treatment effect as the difference between the weighted sum of the treated and untreated populations

Intuition Weights on each individual act to balance the distribution of covariates in the treated and untreated groups. (i.e., break the dependence between treatment status and covariates)

Keep in mind High variance when propensity scores are very high or very low
Many variants of weighting schemes

"Simulating randomized experiments"

- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression

Regression (or supervised learning)

In regression analysis, we build a model of Y as a function of covariates X and T, and interpret coefficients of X and T causally:

$$E(Y|X,T) = \alpha_1 X_1 + \alpha_2 X_2 + \cdots \alpha_n X_n + \alpha_T T$$

Example:

$$Cholesterol = \alpha_{age} Age + \alpha_{exercise} Exercise$$

Model is fit with standard methods (e.g., MLE)

The bigger α is, the stronger the causal relationship to Y

Regression warnings

Causal interpretation of regressions requires **many assumptions.** Threats to validity include:

- **Model correctness:** e.g., what if we use a linear model and causal relationship is non-linear
- **Multicollinearity:** if covariates are correlated, can’t get accurate coefficients
- **Ignorability (Omitted variables):** Omission of confounds will invalidate findings

What we just learned: Regression

Definition Use a regression-based causal analysis, we interpret coefficients as the strength of causal relationship

Example *Modeling cholesterol as a function of exercise and age*

Keep in mind Analysis must be carefully designed to ensure causal interpretability, avoiding collinearity and including all relevant confounds

Avoid unless you are absolutely sure of what you are doing.

Context of Causal Inference Methods

- There are many other causal inference methods
  - with different assumptions (e.g. instrumental variables)
  - or for specific situations (e.g. time series)
- Examples
  - Natural experiments
  - Instrumental variables
  - Regression Discontinuity
  - Difference-in-Differences

- Check out UW Econ 488 or Stat 566 if you are interested!

Sensitivity Analyses

Causal inference is only possible with assumptions

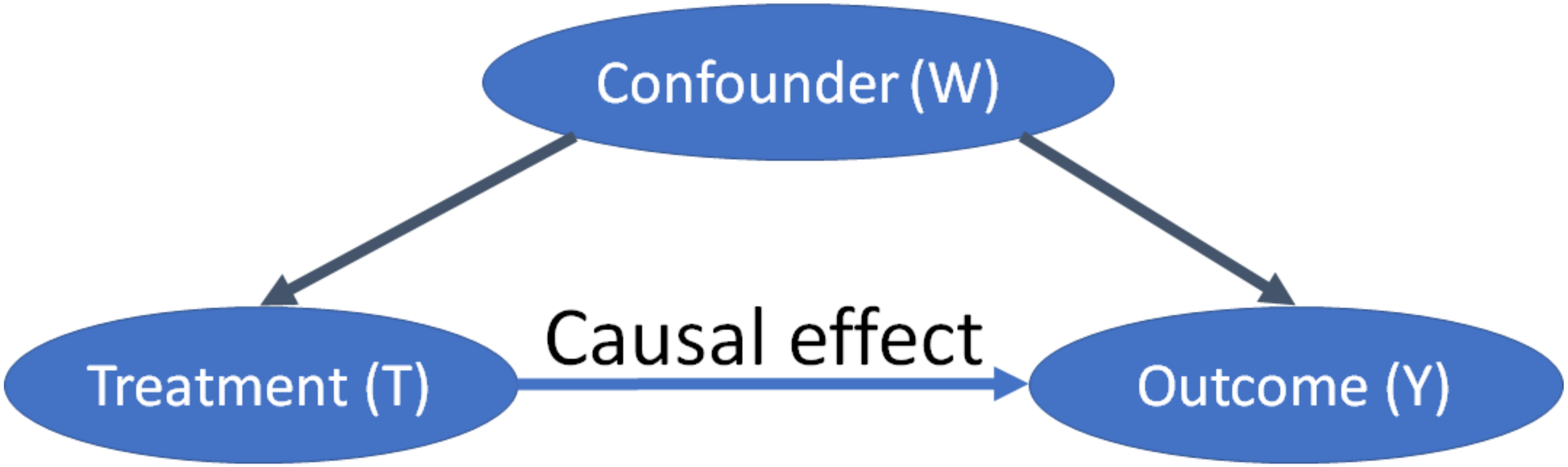
“Causal” part does not come from the data.

It comes from your **assumptions** that lead to ***identification.***

The data is simply used for statistical ***estimation.***

Critical to verify your assumptions. But how?

(Step 1): Making explicit the difference between identification and estimation

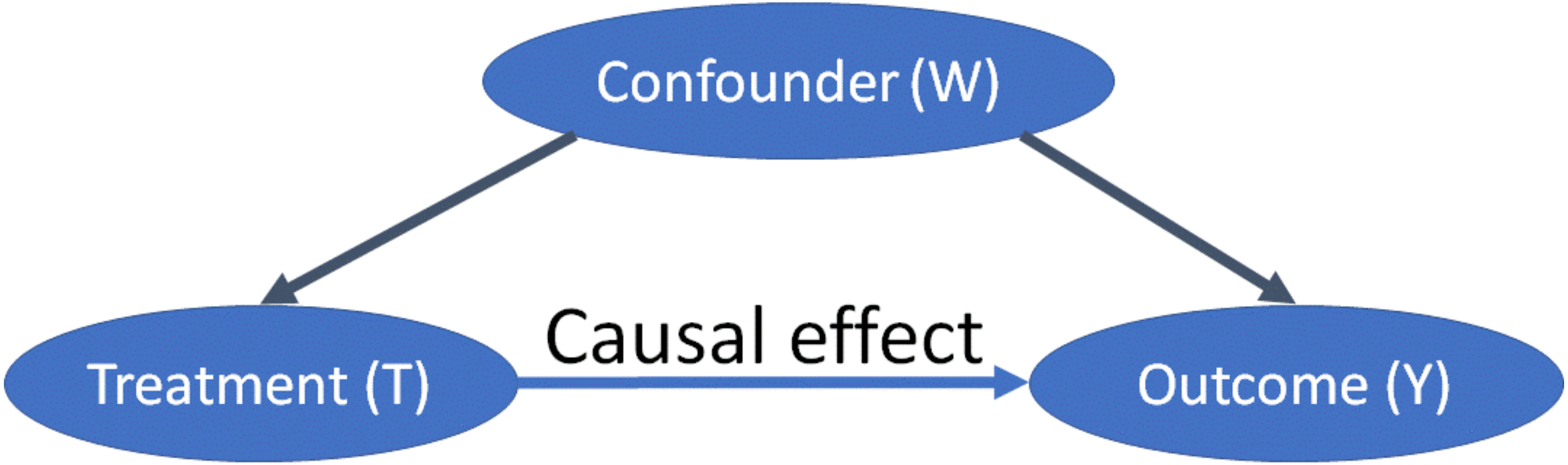


Identification: Causal effect → Observed effect conditioned on W, $\mathrm{E}[Y|T,W]$

Estimation: $\mathrm{E}[Y|T,W]$ → Propensity Score Stratification

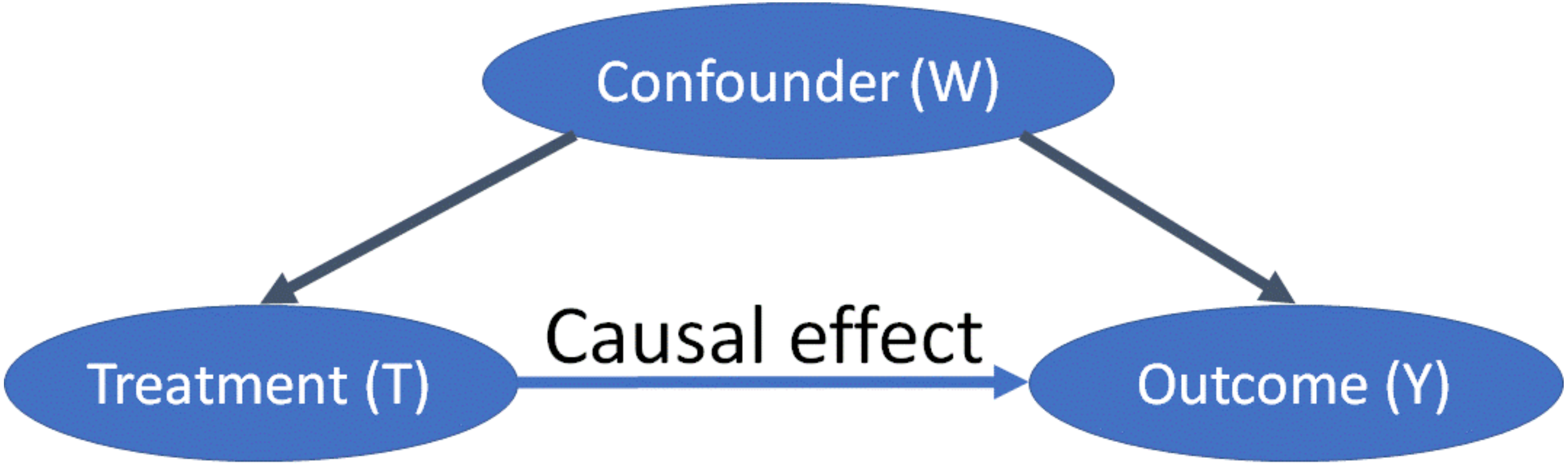
Why do observational studies fail? Most likely due to errors in identification.
--Estimation is a statistical problem, and often easier than correct identification!

(Step 2): Explicitly represent your identifying and estimating assumptions.



Identifying assumption: All the arrows missing in the causal graphical model. E.g. No other common cause exists -> Untestable in general!
Estimating assumption: Overlap between treated and untreated population. Can be checked empirically. Can be solved by collecting more data.

(Step 3): Refute your assumptions, and analyze your estimate's sensitivity to violations



Identifying assumption: All the arrows missing in the causal graphical model. E.g. No other common cause exists -> Untestable in general.

-- *What happens* when another common cause exists?

-- *What happens* when treatment is placebo?

Refutation 1: Add random variables to your model

Can add randomly drawn covariates into data

Rerun your analysis.

Does the causal estimate change? *(Hint: it shouldn’t)*

Refutation check 2: Replace treatment by a placebo (A/A test)

Randomize or permute the treatment.

Rerun your analysis.

Does the causal estimate change? *(Hint: it should become 0)*

Refutation Check 3: Divide data into subsets (cross-validation)

Create subsets of your data.

Rerun your analysis.

Does the causal estimate vary across subsets?
(Hint: it shouldn’t vary significantly)

Refutation Check 4: Test Balance of Covariates

Many methods (e.g., matching, stratification, weighting, regression discontinuity) depend on balancing of covariates

Can test this! In fact, we absolutely need to!

Approaches include statistical tests (t-test, KS statistic, standardized mean difference).

When refutations are not possible? Sensitivity Analysis to violations of assumptions

Question: *How sensitive is your estimate to minor violations of assumptions?*

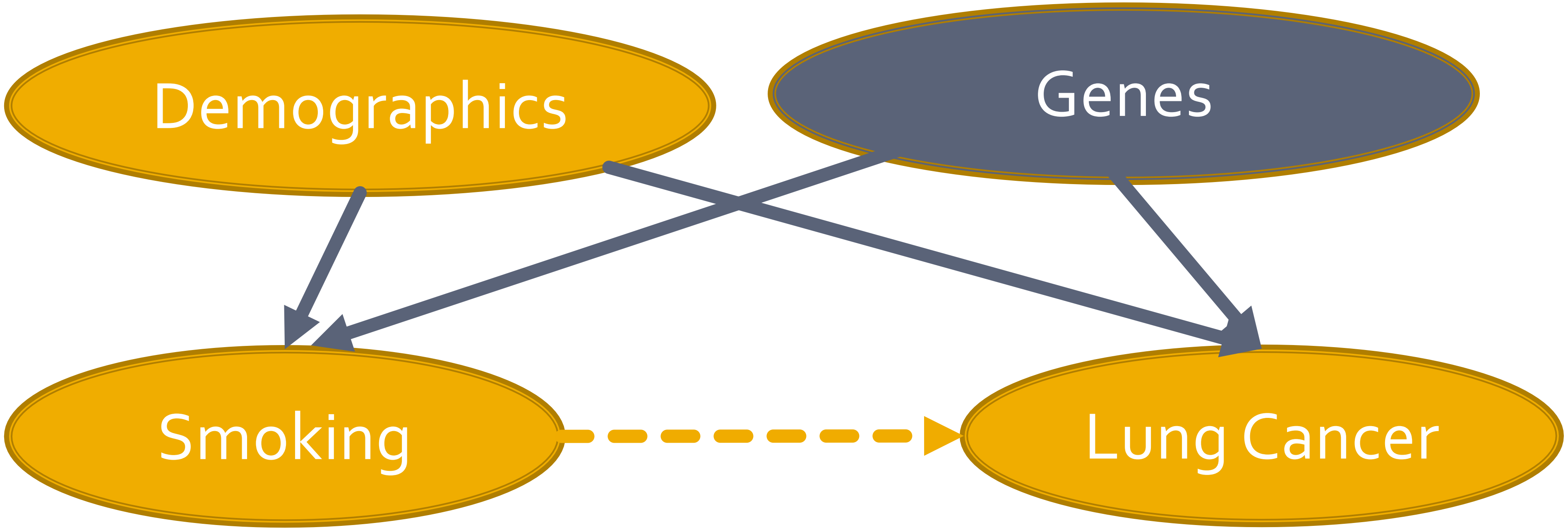
E.g. How big should the effect of a confounder be so that your estimate reverses in direction?

Use simulation to add effect of unknown confounders.

Domain knowledge helps to guide reasonable values of the simulation.

Make comparisons to other known estimates.

Example: Does smoking cause lung cancer?



Cornwell (1959) showed that the effect of Genes had to be 8 times any known confounder for the effect to go to zero.

Observational causal inference: Best practices

Always follow the four steps: *Model, Identify, Estimate, Refute.*

Refute is the most important step.

Aim for simplicity.

If your analysis is too complicated, it is most likely wrong.

Try at least two methods with different assumptions.

Higher confidence in estimate if both methods agree.

Recap of today:

- Many methods for statistical estimation of causal effects exist
  - Conditioning
  - Matching
  - Stratification
  - Weighting
  - Regression
- The main idea is to attempt to simulate a randomized experiment with observational data.
- Causal inference works through making assumptions
  - Make sure to check them and attempt to refute your models!

Thank you for your class participation!

Teaching Assistants

Ashish Sharma
(Head TA)

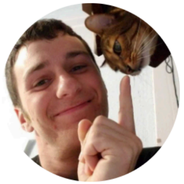
Kristof Glauninger

Qifan Huang

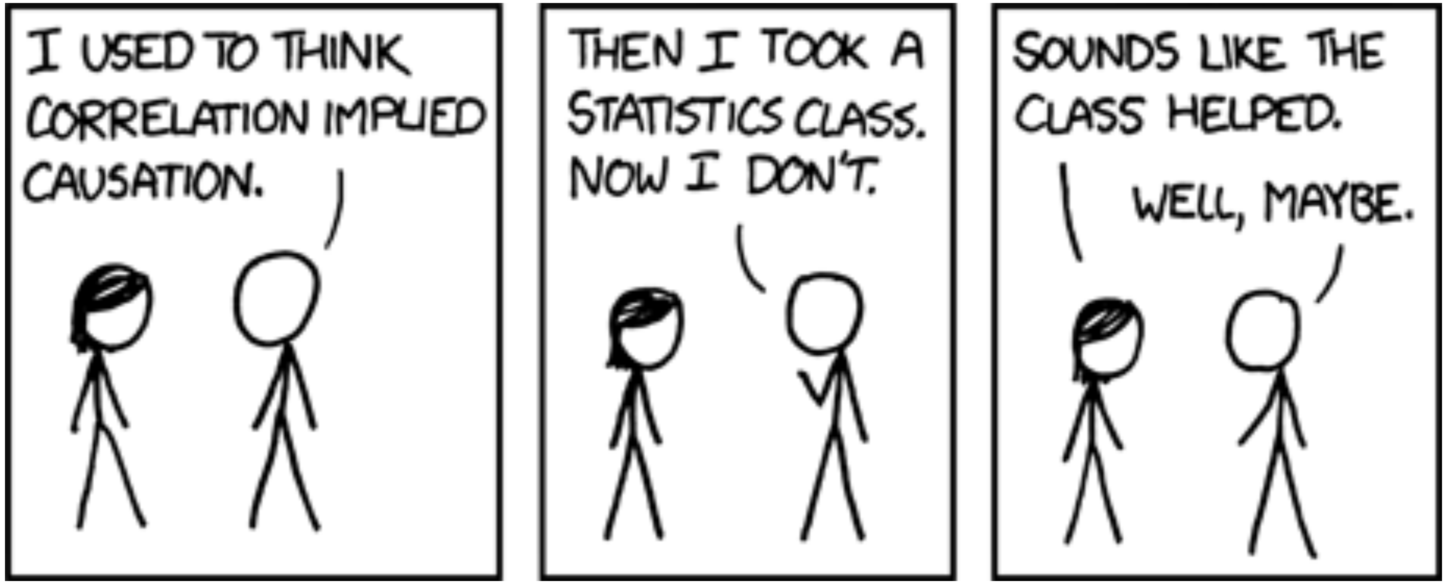
Stephen J. Jonany

Jack Khuu

Alon Milchgrub


Galen Weld

Chi-Hui Yen


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

Let’s do the course evaluation

- https://uw.iasystem.org/survey/224031

- Your feedback makes a difference!
- Thank you!