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

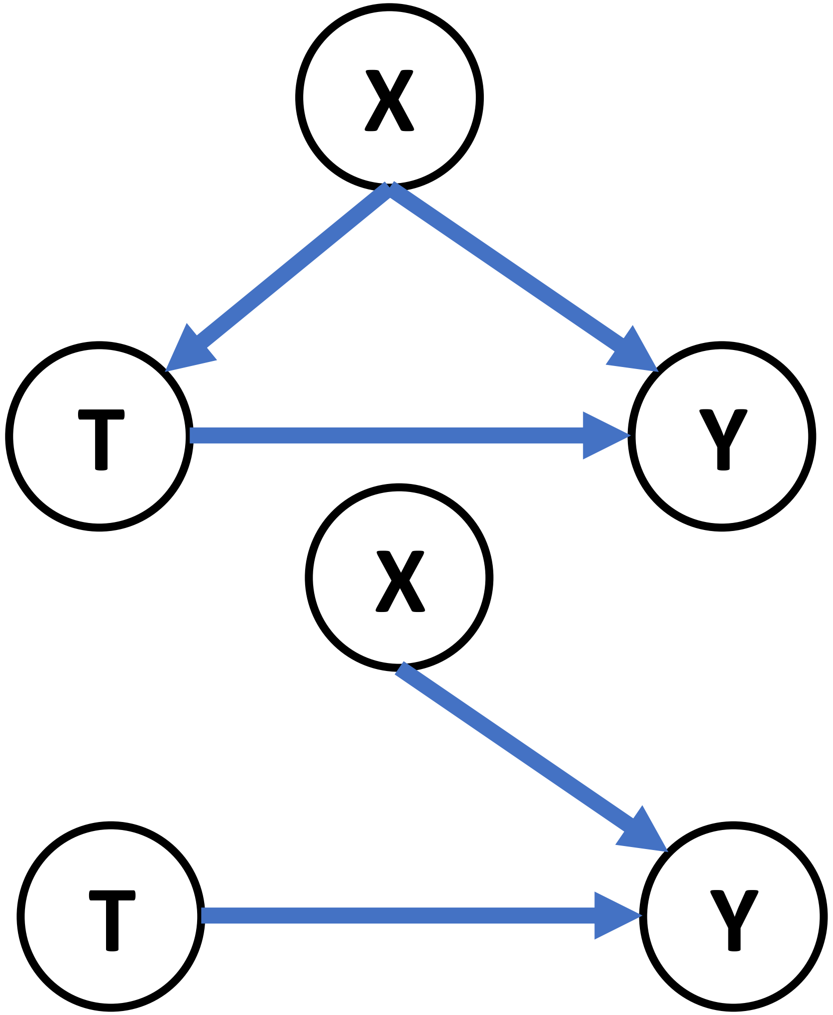
But, confound X influences both T and Y

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

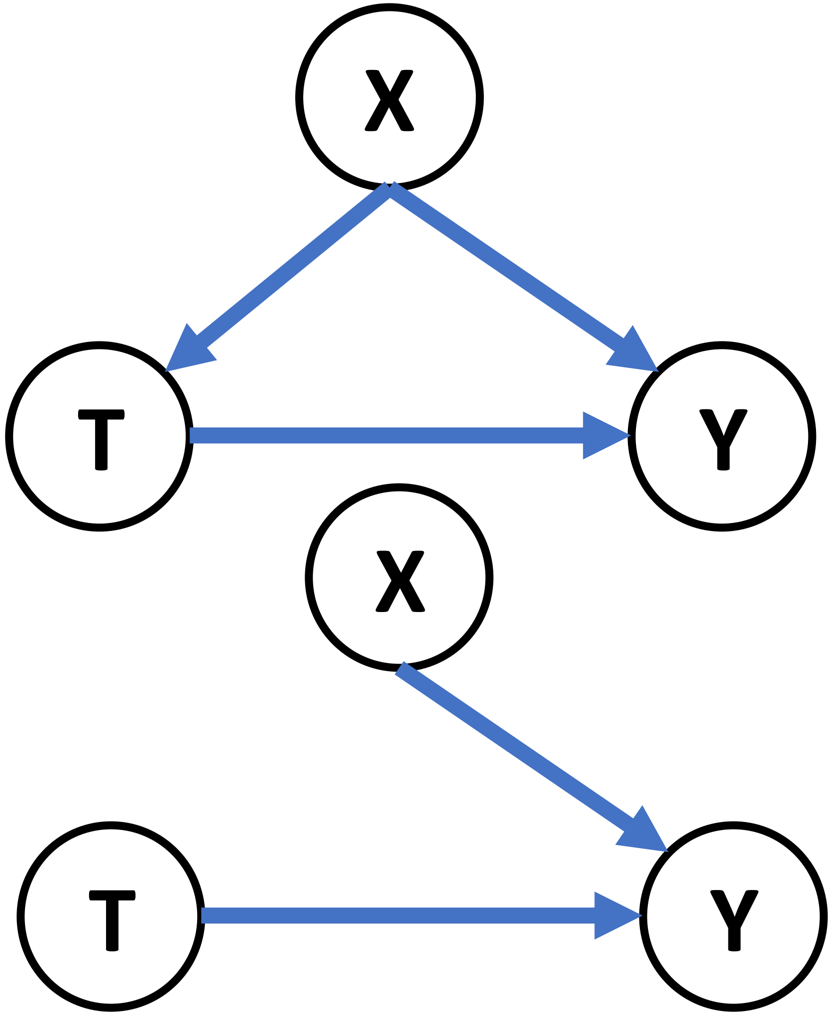
Goal: Estimate effect of a treatment T on an outcome Y

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

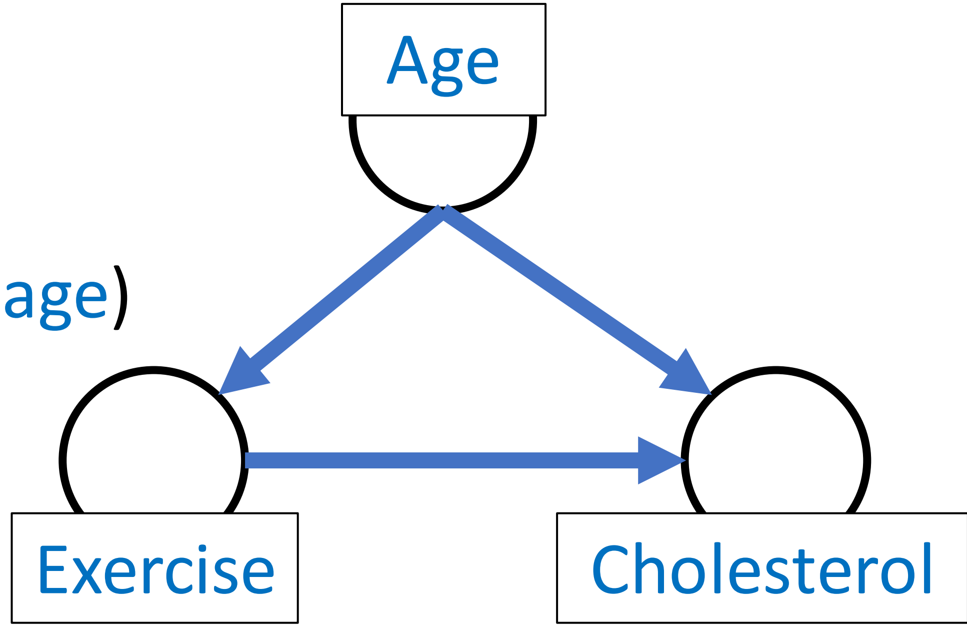
Goal: Estimate effect of exercise on cholesterol

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



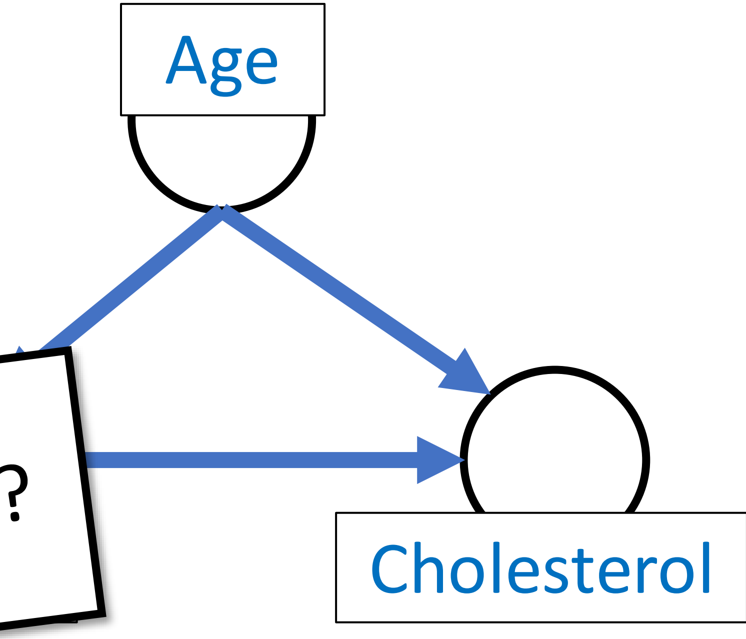
Running example

~~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 exp

exerc

Thus, ⫫ age

But, what if we cannot actively intervene?

Part II.A. Observational Studies

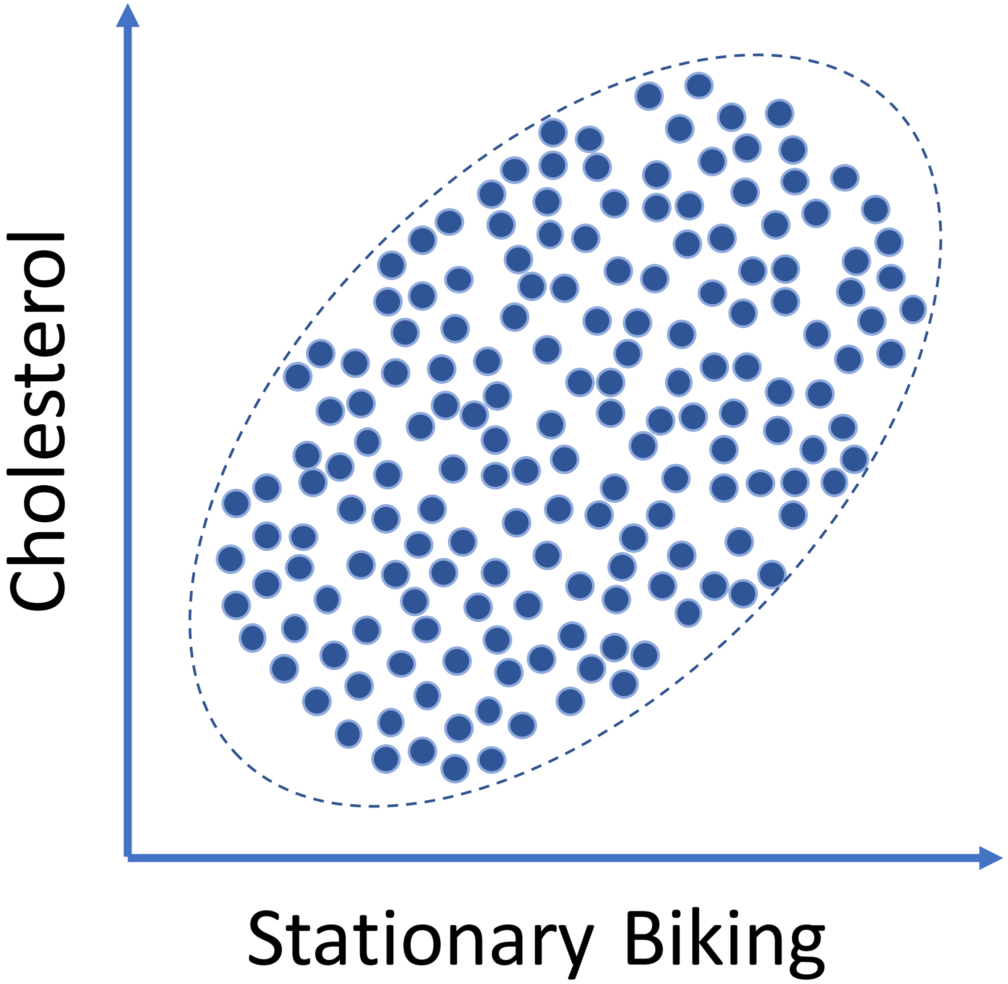
“Simulating randomized experiments"

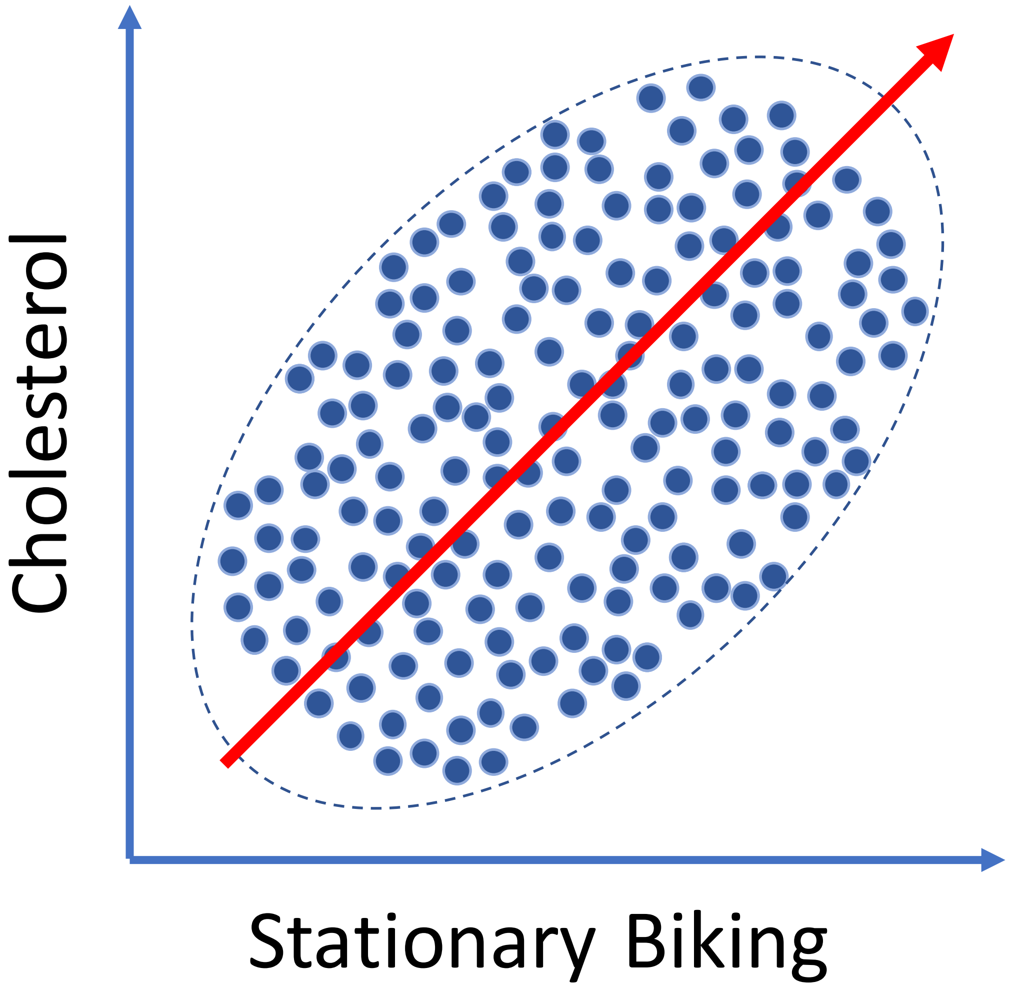
- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Regression
- Doubly Robust
- Synthetic Controls

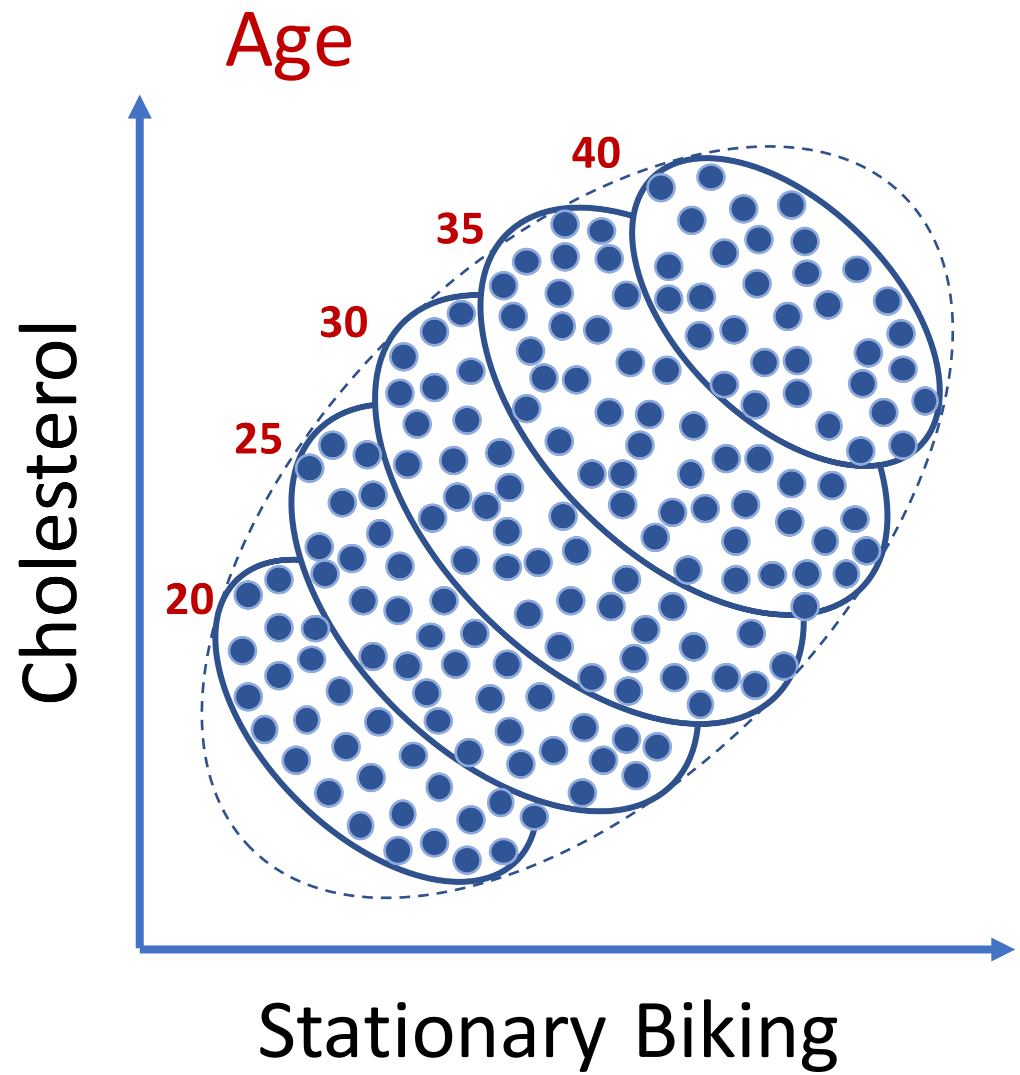
Part II.A. Observational Studies

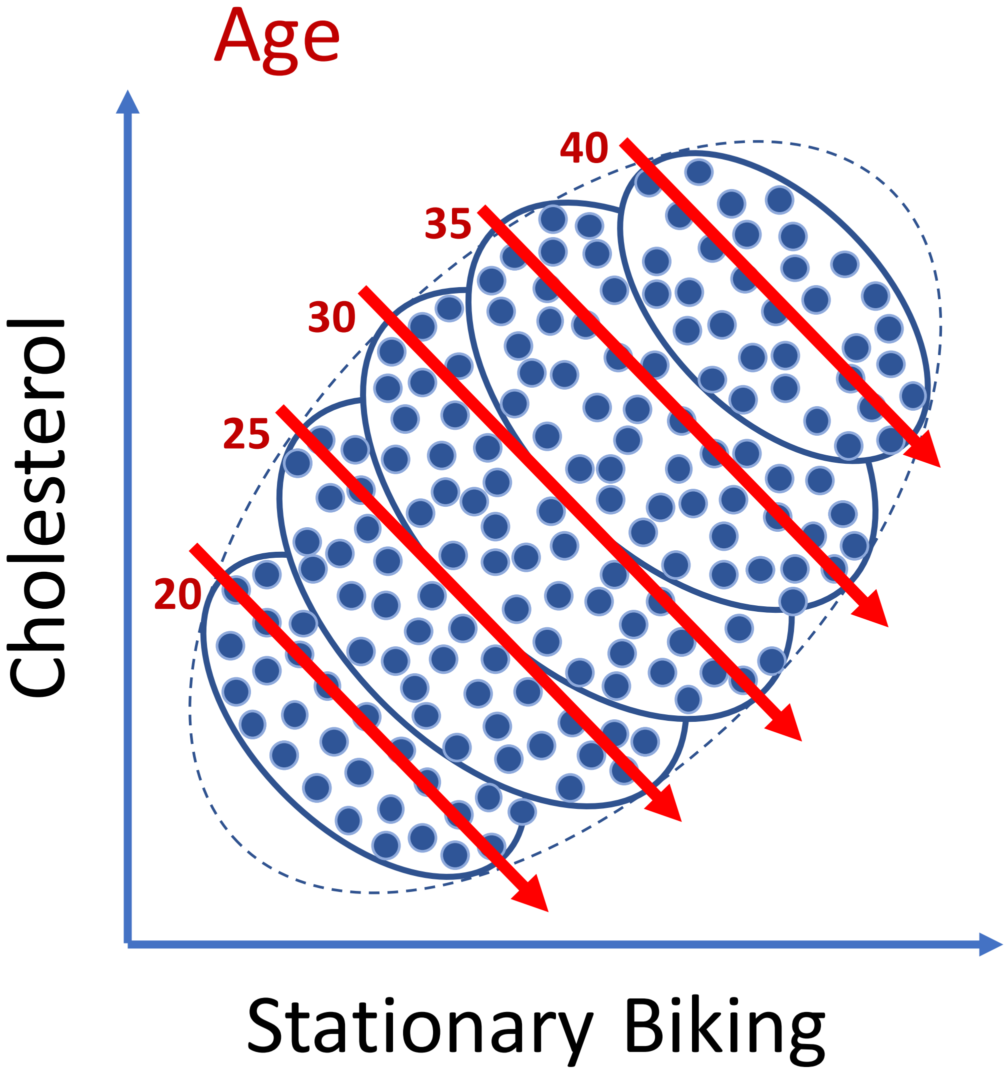
“Simulating randomized experiments"

- 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
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.
- 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

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

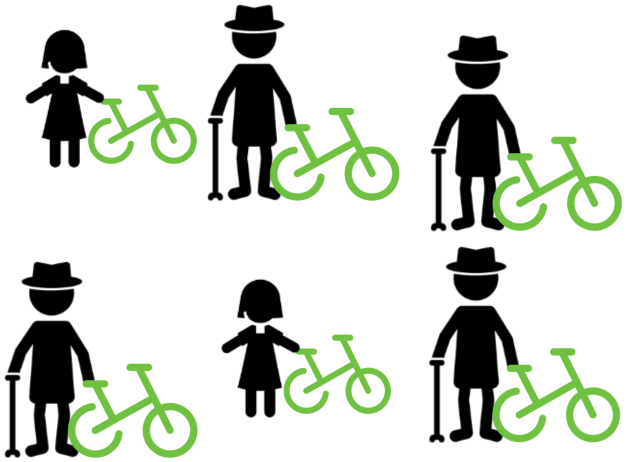
Example: What is the effect of giving a fax machine to an individual?

- It depends on whether or no

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

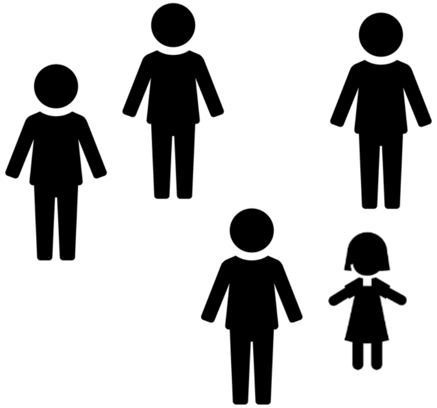
A3. Common support

- 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.



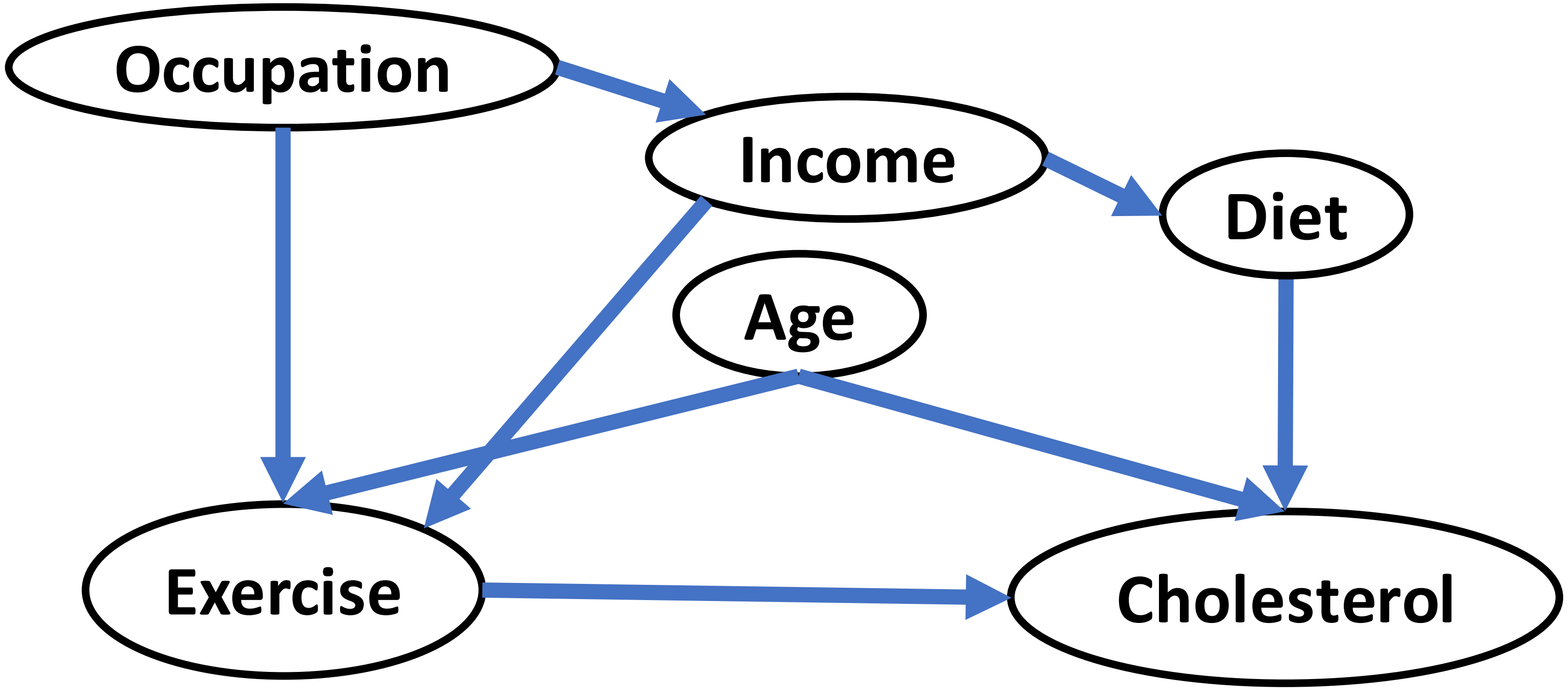
Common support

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



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.

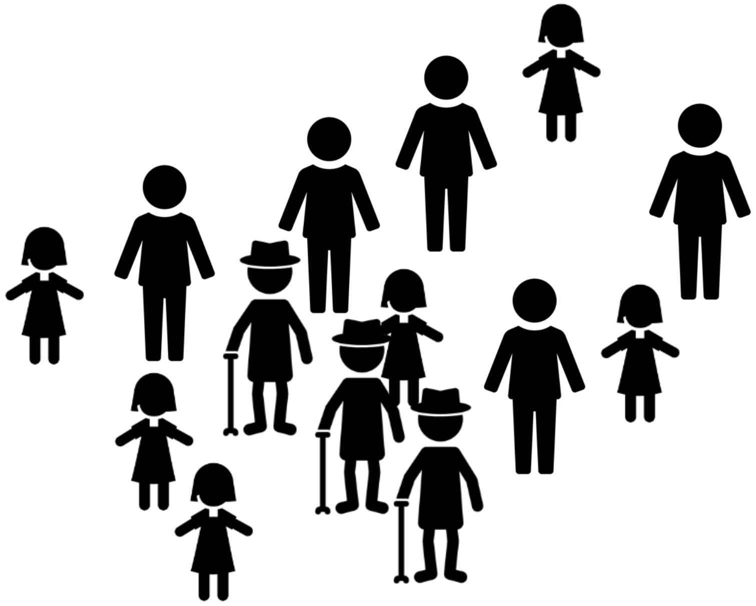
Keep in mind How do we know what to condition on?

Grouping becomes harder as dimensionality of X increases

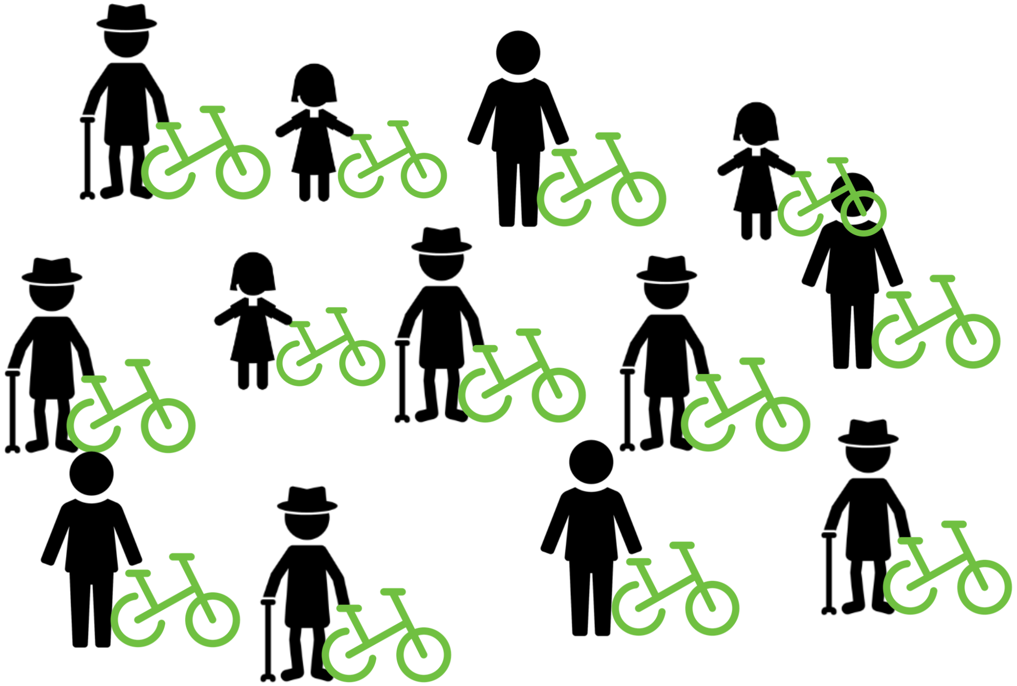
Part II.A. Observational Studies

“Simulating randomized experiments"

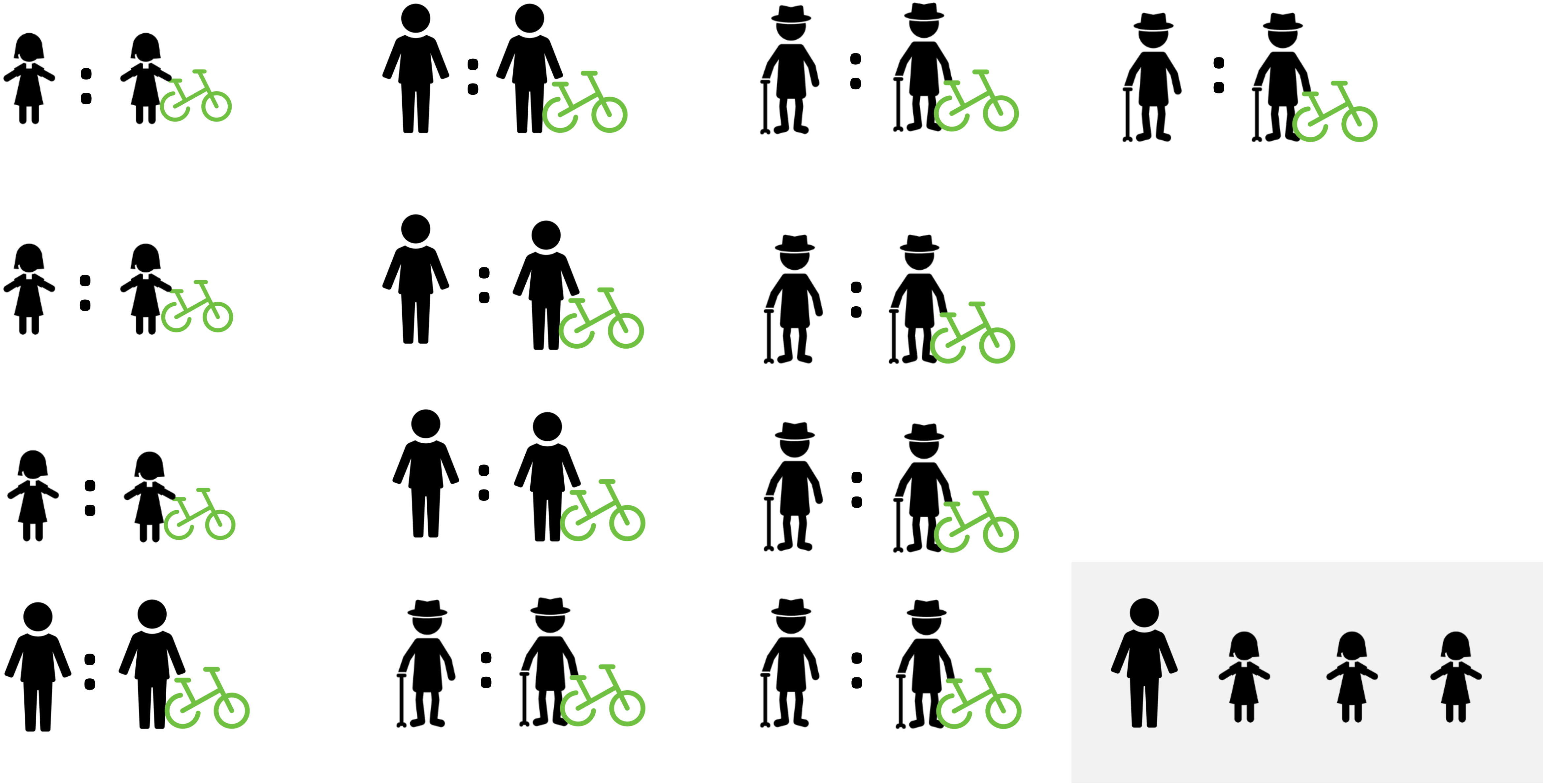
- Conditioning on Key Variables
- Matching and Stratification
- Weighting
- Simple Regression
- Doubly Robust
- Synthetic Controls



Avg Cholesterol = 200



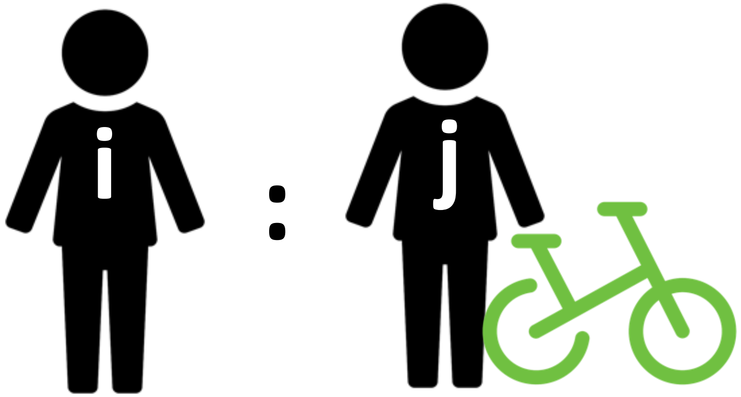
Avg Cholesterol = 206



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:

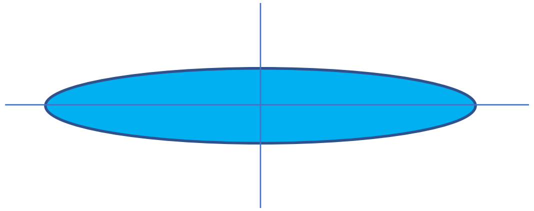
$$Distance(\vec{x}_i, \vec{x}_j) = \begin{cases} 0, & \vec{x}_i = \vec{x}_j \\ \infty, & \vec{x}_i \neq \vec{x}_j \end{cases}$$

Use this in low-dimensional settings when overlap is abundant

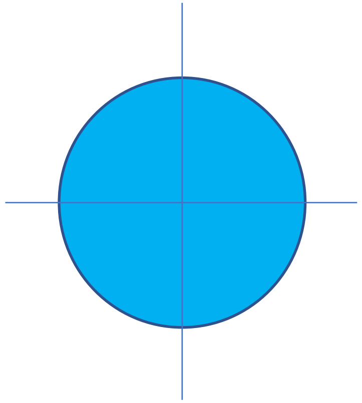
But in most cases, there will be too few exact matches …

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.

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 randomiz
  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!

Propensity score matching python code

```
# learn propensity score model
psmodel = linear_model.LinearRegression()
psmodel.fit(covariates, treatment_status)
data['ps'] = psmodel.predict(covariates)
# find nearest neighbor matches
controlMatcher = NearestNeighbors().fit(untreated['ps’])
distances, matchIndex = controlMatch.kneighbors(treated['ps'])
# iterate over matched pairs and sum difference in outcomes
for i in range(numtreatedunits):
    treated_outcome = treated.iloc[i][outcome_name].item()
    untreated_outcome = untreated.iloc[matchIndex[i]][outcome_name].item()
    att += treated_outcome - untreated_outcome
# normalize
att /= numtreatedunits
```

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.
- Treatment should be a binary point treatment
  - Advanced variants allow multi-dose, and other treatment regimens

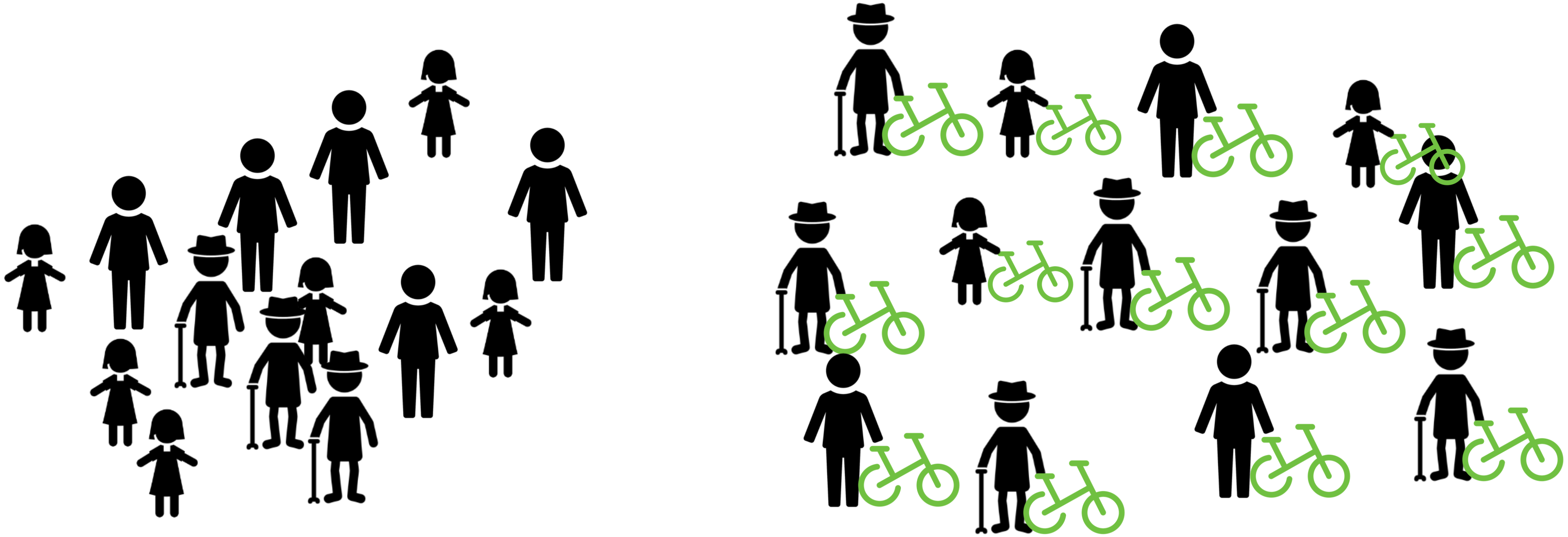
What we just learned: Matching

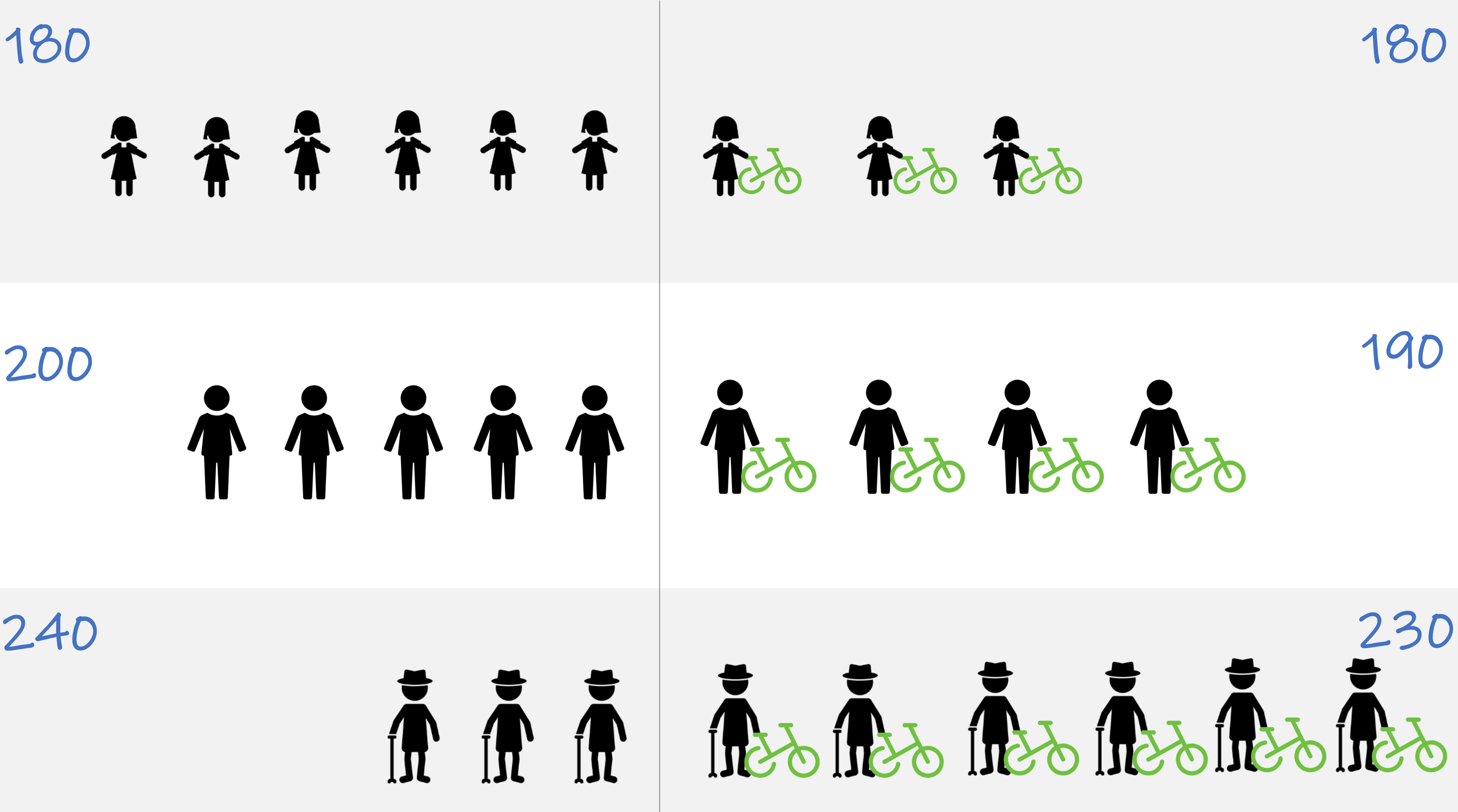
Definition Matching calculates treatment effects by identifying pairs of similar individuals, where one is treated and the other is not.

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. 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

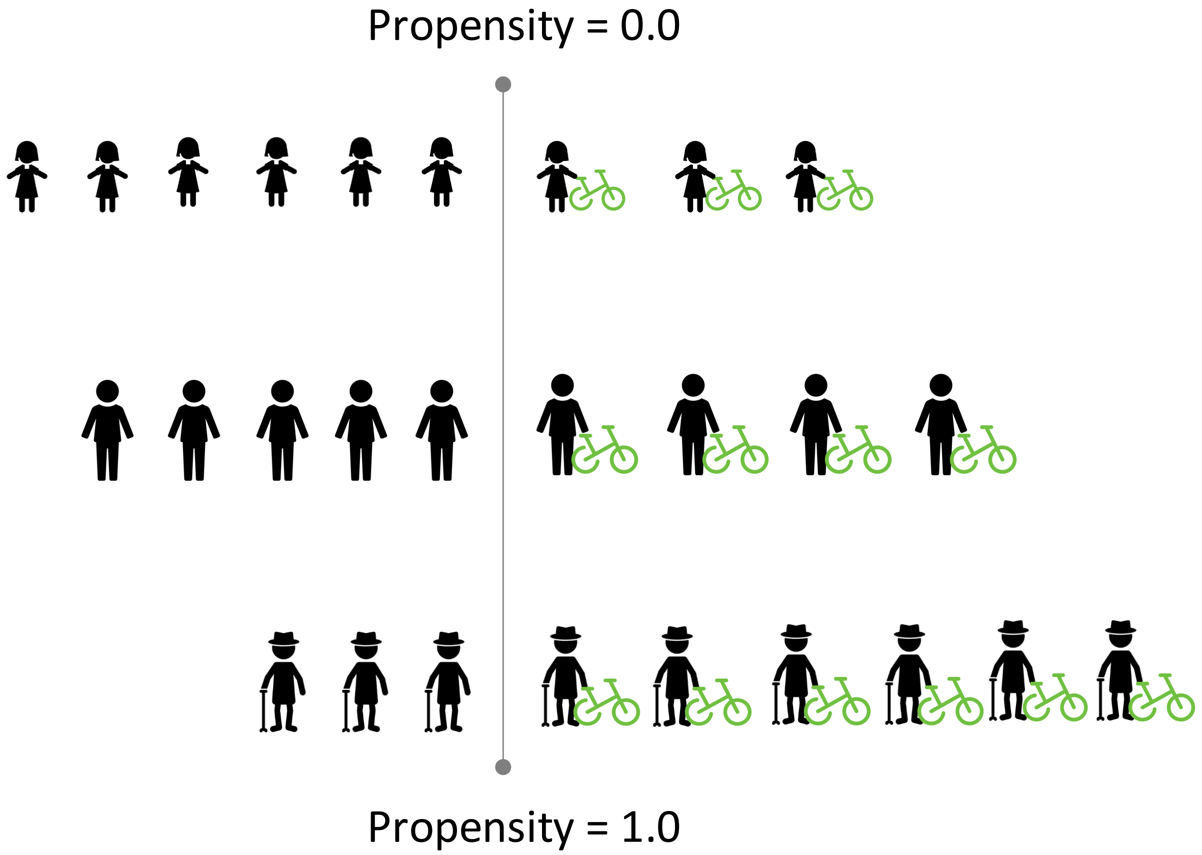
2. Distance is the difference between propensity scores

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

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.



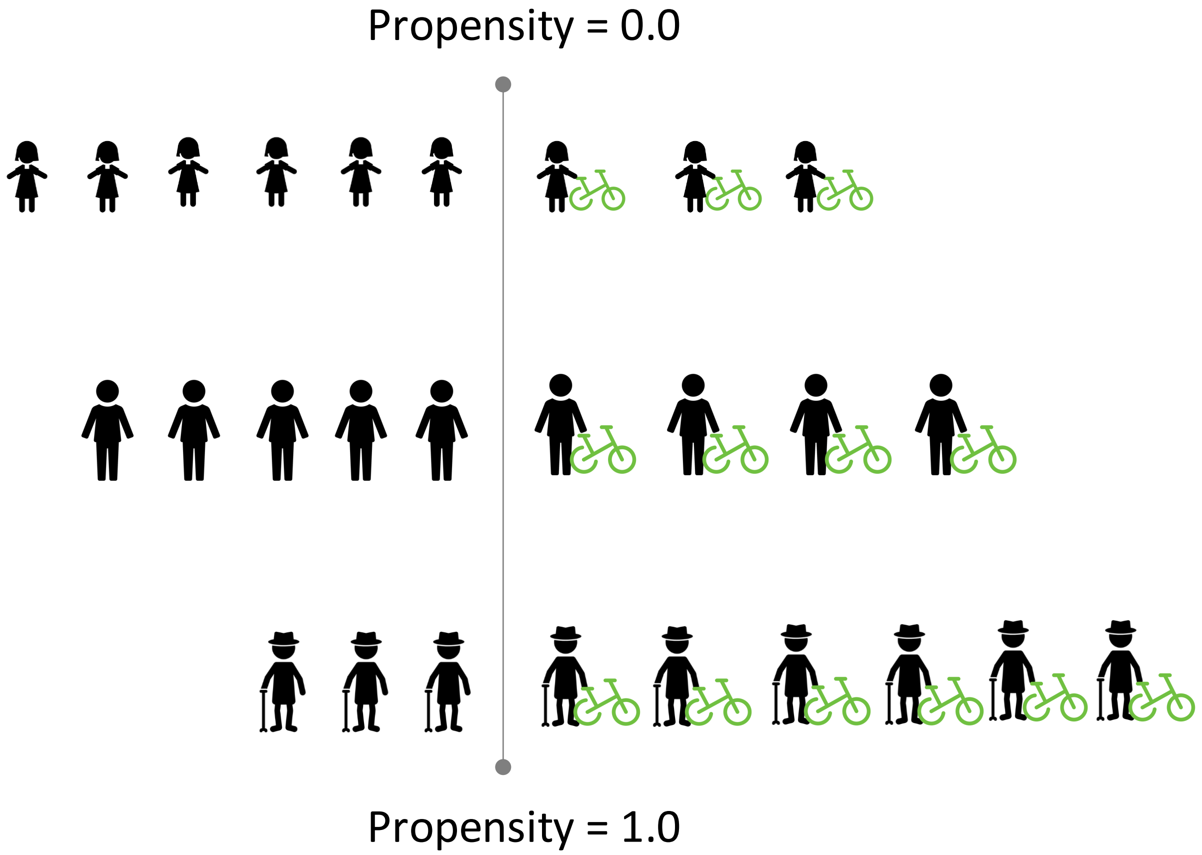
Propensity Score Stratification

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

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 score stratification python code

```
# build propensity score model and assign each item a score as earlier…

# create a column 'strata' for each element that marks what strata it belongs to
data['strata'] = ((data['ps'].rank(ascending=True) / numrows) * numStrata).round(0)
data['T_y'] = data['T'] * data['outcome’]            # T_y = outcome iff treated
data['Tbar'] = 1 - data['treated’]                   # Tbar = 1 iff untreated
data['Tbar_y'] = data['Tbar'] * data['outcome']      # Tbar_y = outcome iff untreated
stratified = data.groupby('strata')
# sum weighted outcomes over all strata  (weight by treated population)
outcomes = stratified.agg({'T':['sum'],'Tbar':['sum'],'T_y':['sum'],'Tbar_y':['sum']})
# calculate per-strata effect
outcomes[‘T_y_mean'] = outcomes[‘T_y_sum'] / outcomes['T']
outcomes[‘Tbar_y_mean'] = outcomes[‘Tbar_y_sum'] / outcomes['dbar_sum']
outcomes['effect'] = outcomes[‘T_y_mean'] - outcomes[‘Tbar_y_mean’]
# weighted sum of effects over all strata
att = (outcomes['effect'] * outcomes['T']).sum() / totaltreatmentpopulation
```

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

Part II.A. Observational Studies

“Simulating randomized experiments"

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

Weighting: An alternative to conditioning

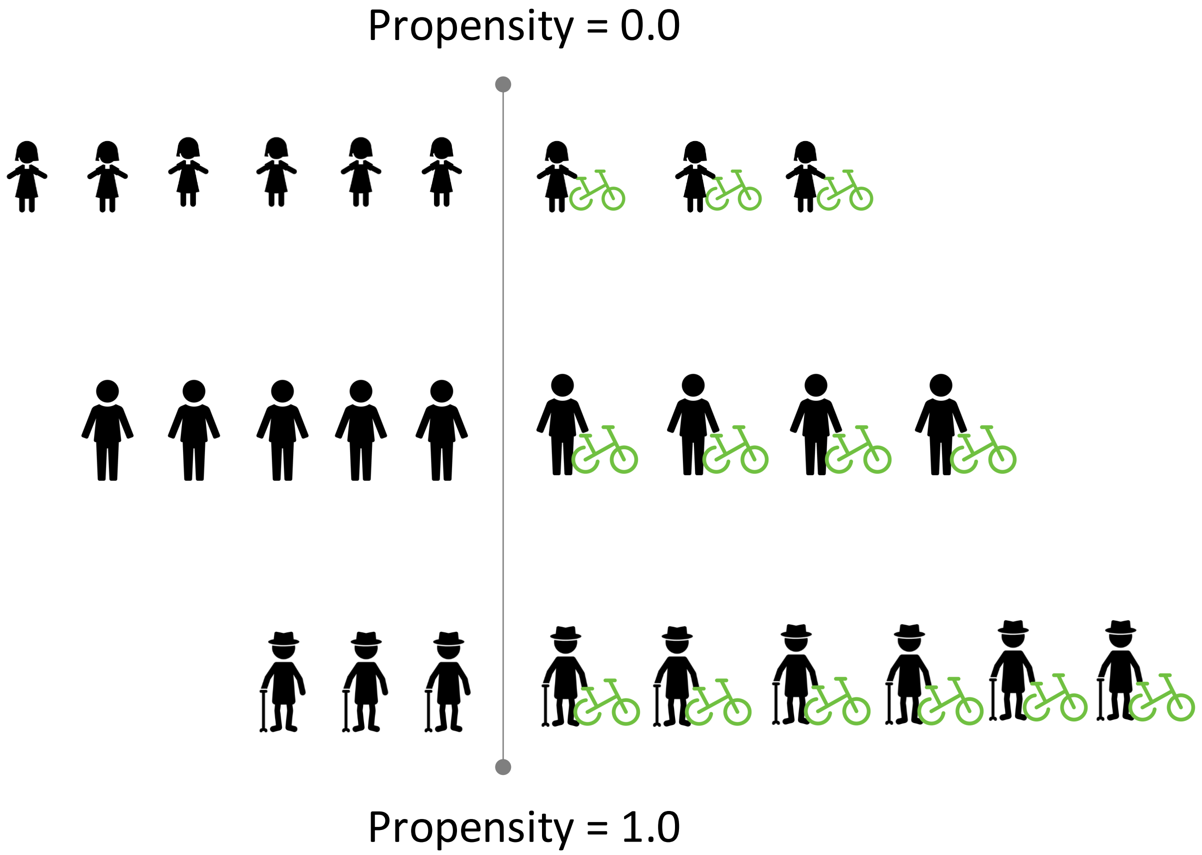
What if we assign weights to observations to simulate randomized experiment?

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



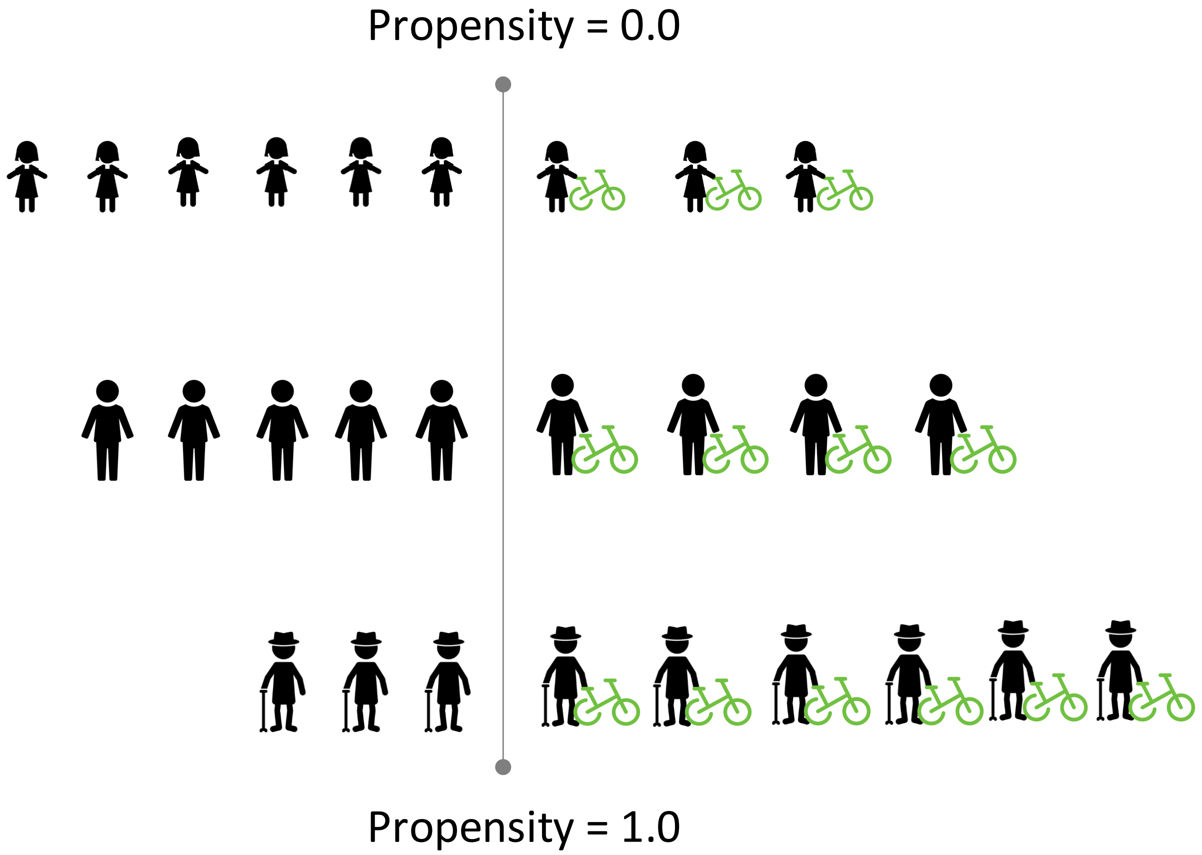
Weighting

Stratification weights strata results by number of treated

Weighting by treated population ~ weighting by propensity score.

Generalized weighting: Calculate effect by weighted sum over all individual outcomes

Many weighting methods to generate a balanced dataset



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)

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

$$\mathrm{N}_{\mathrm{T=1}} = \sum \frac{T}{e}; \quad N_{T=0} = \sum \frac{1-T}{1-e}$$

Weighting: Caveats and Practical notes

- High variance when e 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

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

Part II.A. Observational Studies

"Simulating randomized experiments"

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

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.

Part II.A. Observational Studies

“Simulating randomized experiments"

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

Doubly robust: Best of both worlds?

- Both propensity score weighting and regression models require correctly specified models
    - E.g., if propensity score or regression is modeled as a linear combination, but is non-linear, than it is not correctly specified
- Doubly robust methods combine “best of” propensity score and regression methods
- If either propensity score *or* regression is correctly specified, then doubly robust is correct.

DR: Combines 3 components

Learn 3 models:

1,2: Models of outcome given treatment and covariates: $\hat{Y}_{T=0}$, $\hat{Y}_{T=1}$

3: Propensity of treatment given covariates: $\hat{e}$

Combine to calculate doubly robust estimators, DR_1 and DR_0, for each individual:

$$DR_1 = \begin{cases} \dfrac{Y}{\hat{e}} - \dfrac{\hat{Y}_{T=1}(1-\hat{e})}{\hat{e}}, & T = 1 \\ \hat{Y}_{T=1}, & T = 0 \end{cases}$$

$$DR_0 = \begin{cases} \hat{Y}_{T=0}, & T = 1 \\ \dfrac{Y}{1-\hat{e}} - \dfrac{\hat{Y}_{T=1}\hat{e}}{1-\hat{e}}, & T = 0 \end{cases}$$

Finally, calculate mean $\overline{DR_1}$ and $\overline{DR_0}$ over the whole study population, and take difference as the causal effect of T

Doubly Robust: Caveat

If either propensity score *or* regression is correctly specified, then doubly robust is unbiased.

Seems like doubly robust should be strictly better (less biased) than either propensity score weighting or regression

But, if both propensity score or regression are *slightly* incorrect, then doubly robust estimator may become *very* biased

What we just learned: Doubly Robust

Intuition	Combine propensity score weighting and regression models to provide unbiased estimate when either propensity score or regression is correctly specified
Keep in mind	Fundamental assumptions (ignorability, etc) must still hold. If both models are slightly incorrect, doubly robust estimator can be more biased

Part II.A. Observational Studies

“Simulating randomized experiments"

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

Synthetic control

All previous methods require that we observe both *treated* and *untreated* individuals

What if we are analyzing a scenario where everyone is treated?

E.g., effect of a large marketing campaign, or a global policy change?

Pre/Post comparison is option, but not robust to dynamics, seasonality, …

Alternative: Build *synthetic controls* that estimate what $\bar{Y}_{T=0}$ would have been for a population were it not for treatment

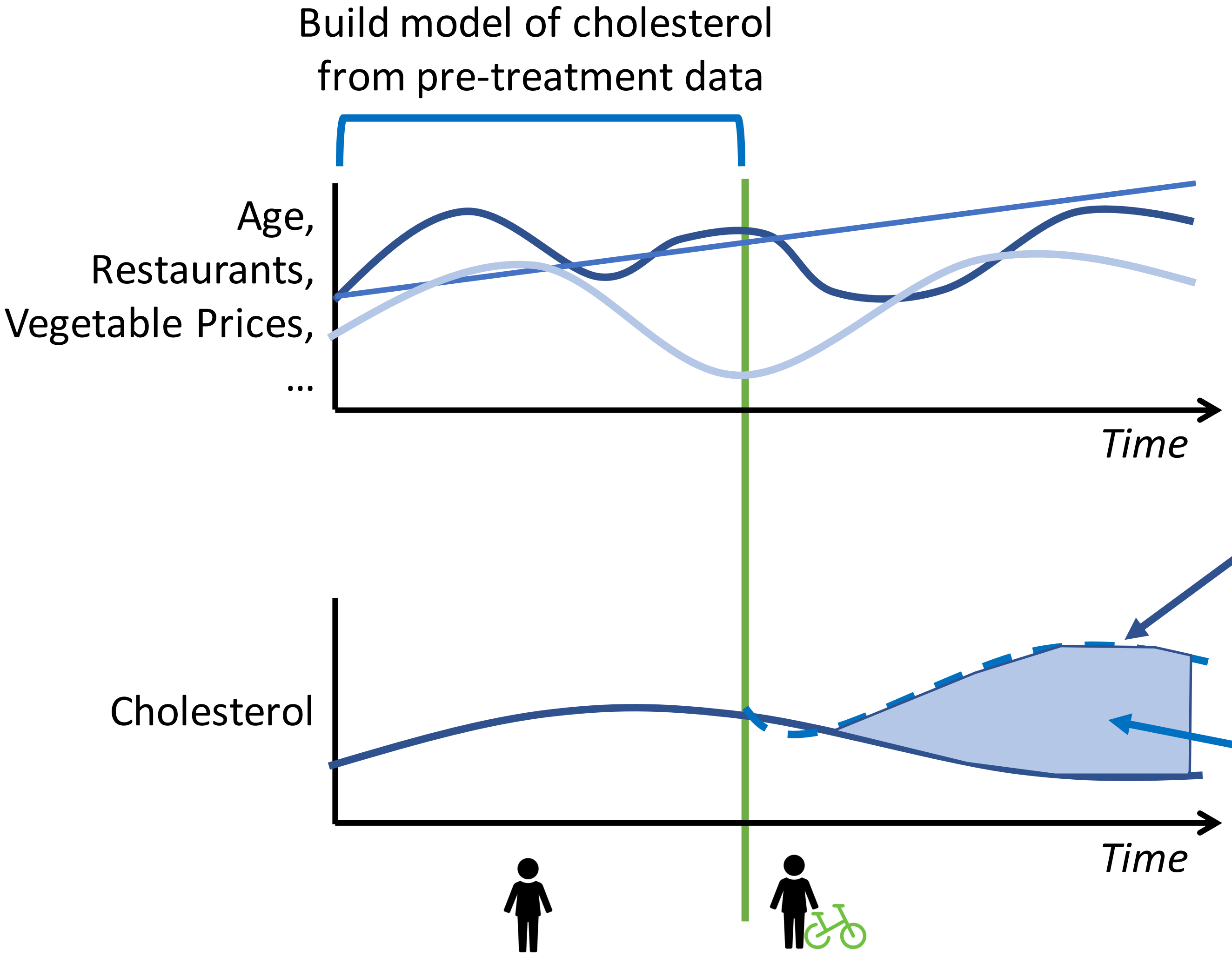
Synthetic controls: Intuition

1. *Decide what the treatment will be*

2. *Pre-treatment stage:* Observe the world for awhile
    - Record the outcome we care
    - Record covariates that can help us predict our observed outcome, but will not be effected by the treatment. Use domain-knowledge / theory to identify these covariates.
    - Learn a model that predicts outcome based on covariates.

3. *Post-treatment stage:*
    - Keep recording outcome. This is now the treated outcome.
    - Predict untreated outcome using learned model and current covariates
    - ATE = Difference between observed outcome and prediction of untreated outcome

Example: policy change to encourage exercise



Synthetic control (untreated) predicted from covariates

Treatment effect is the difference between predicted control and observed (treated) outcome

What we just learned: Synthetic Controls

Definition Calculate treatment effect by comparing observed outcomes of treated population with synthetic (predicted) outcomes of an untreated population

Intuition If we can measure covariates that are unaffected by the treatment and predictive of untreated outcomes, then we can build a synthetic control

Example Predicting effect of global policy change to encourage exercise on population-wide cholesterol

Keep in mind Ignorability assumption must still hold;
Relatedly, be concerned about generalizability/robustness of learned outcome model

PART II. Methods for Causal Inference

- Observational Studies
- Natural Experiments
- Refutations

Part II.B. Natural Experiments

Simple natural experiment

Instrumental Variables

Regression Discontinuities

Natural experiments: What can we do without ignorability?

Rather than assume ignorability over the entire dataset, find data subsets that approximate an experiment.

“Natural” → as if Nature *conducted an experiment* for you

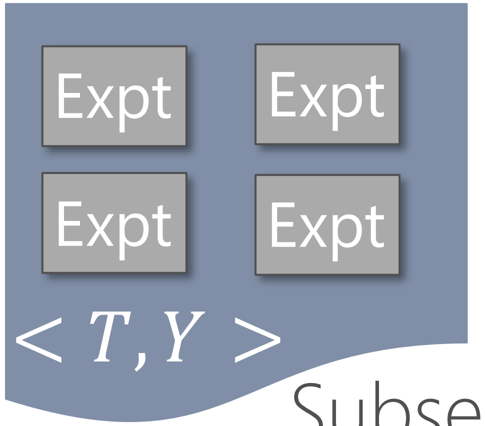
Common sources: Prior A/B tests, Lottery, any randomized policy, an external shock to the treatment.

Allows common causes of T and Y, as long as the source is not affected by them.

Finding a natural experiment

Full dataset

$$y = f(t, x)$$
$$t = g(x)$$



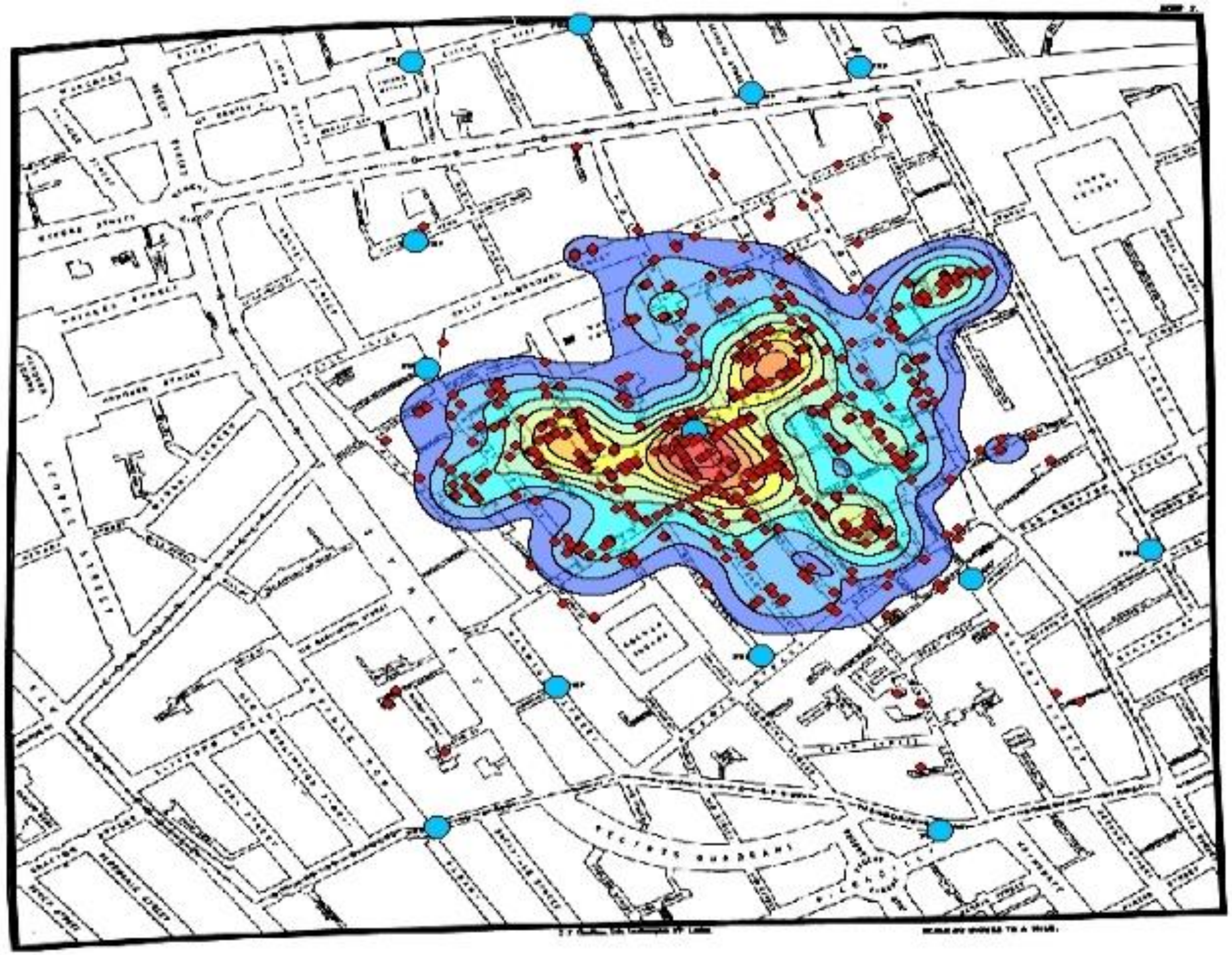
Subsets of the data

$$y = f(t, u)$$
$$t = g(r)$$

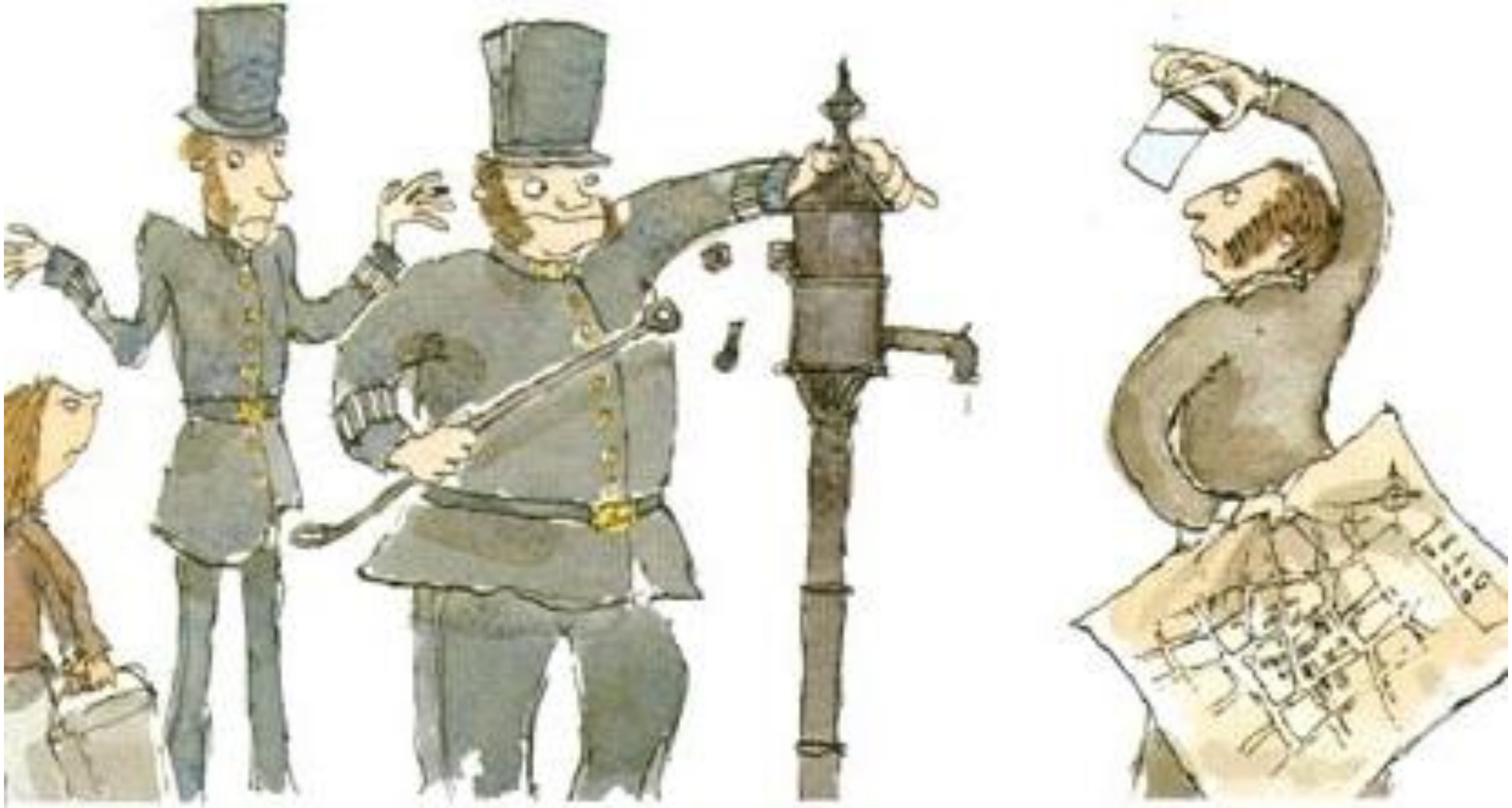
$\boldsymbol{r}\text{:}\ randomized$

How to find such experiments?

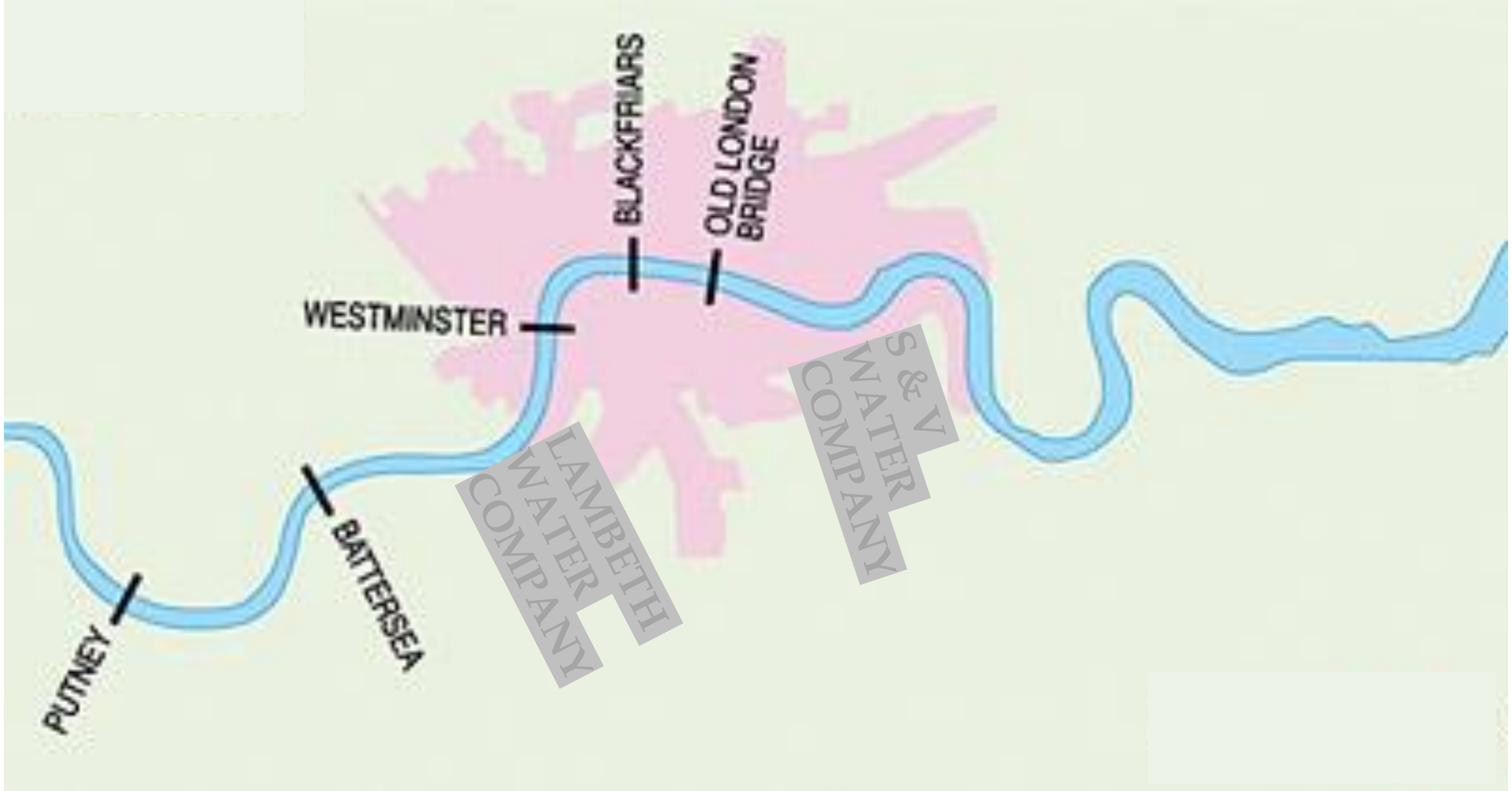
Example: Cholera cause estimation in 1850s.



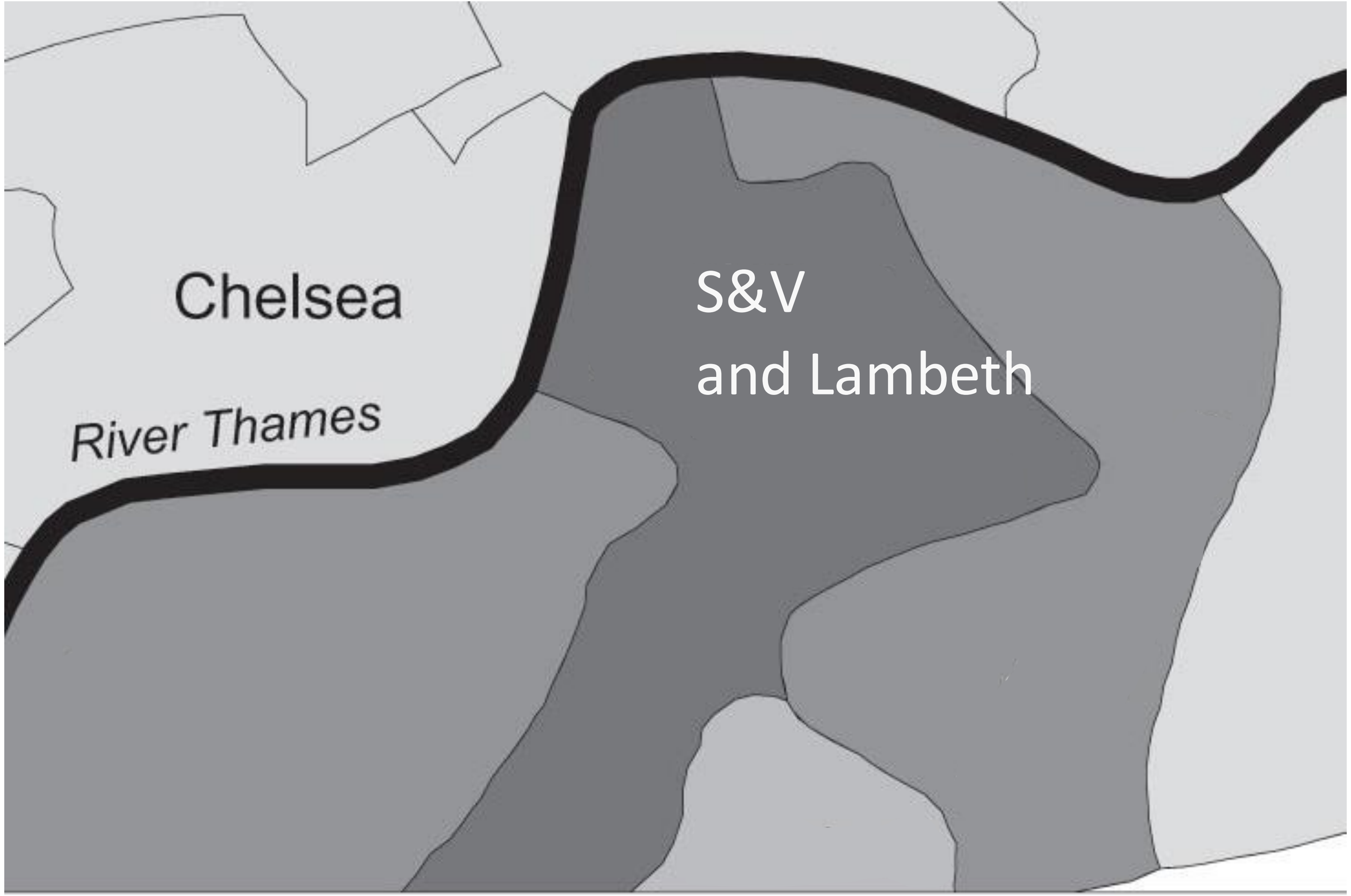
1854: London was having a devastating cholera outbreak



Enter John Snow. He found higher cholera deaths near a water pump, but could be just correlational.



New Idea: Two major water companies for London:
one upstream and one downstream.
Customers of each company distributed throughout city



No difference in neighborhood, still an 8-fold increase in cholera with the downstream company.

“Natural” experiments: exploit variation in observed data

Can exploit naturally occurring **as-if random** variation in data.

Since data is not actively randomized, as-if-random remains an assumption.

Also need **exclusion**: the source of variation should not affect the outcome directly, only the treatment.

Dunning (2002), Rosenzweig-Wolpin (2000)

What we just learned: Simple natural experiment

Definition Exploit “as-if random” assignment of treatments to measure outcome.

Intuition When assignment of treatment is unrelated to the measured outcome and their common causes, we can treat it as if it is a randomized experiment to estimate treatment effect.

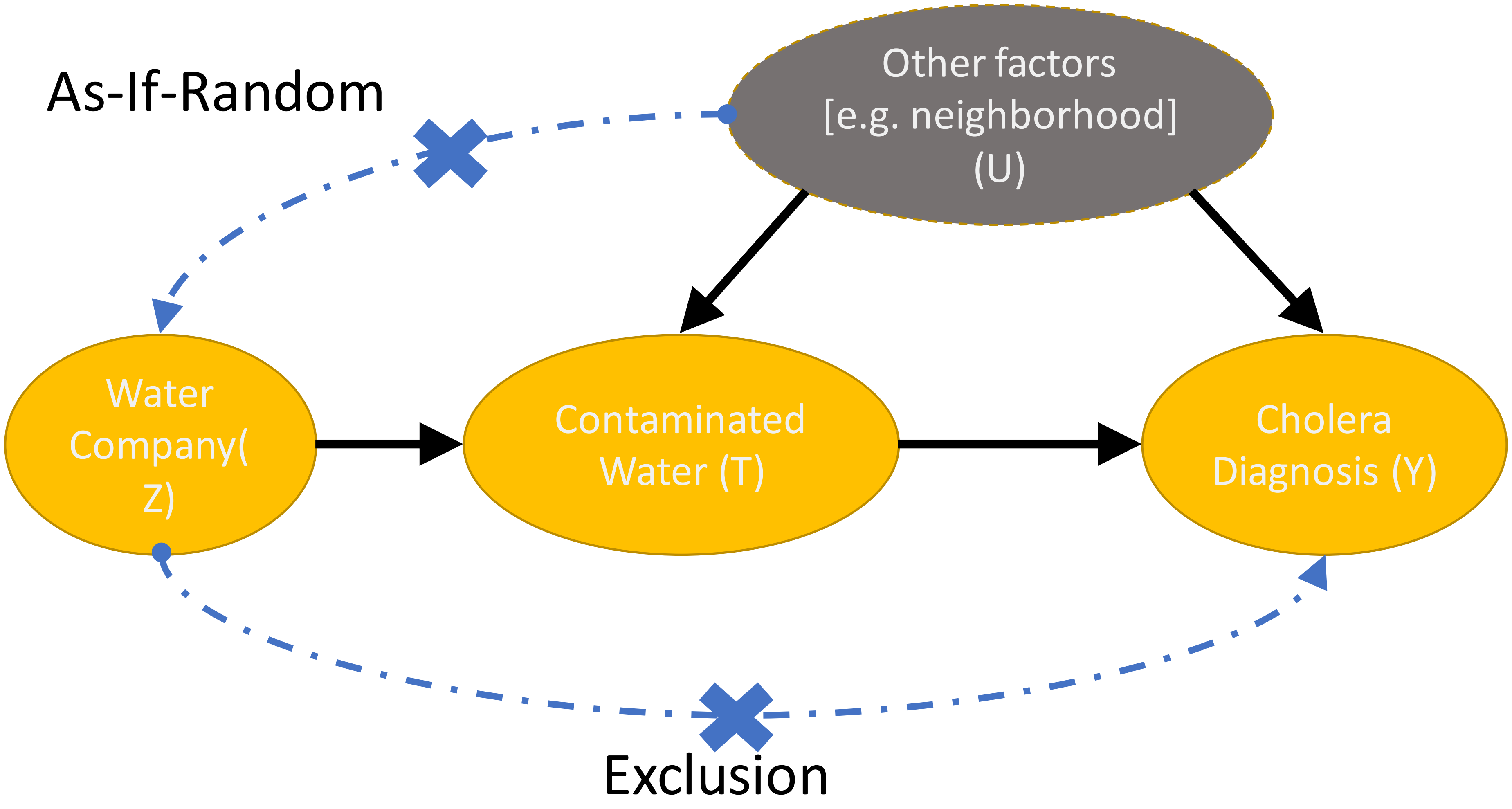
Example What water company do you buy from?

Keep in mind As-if random assignments of treatments are hard to find.
Estimates very sensitive to violation of exclusion assumption.

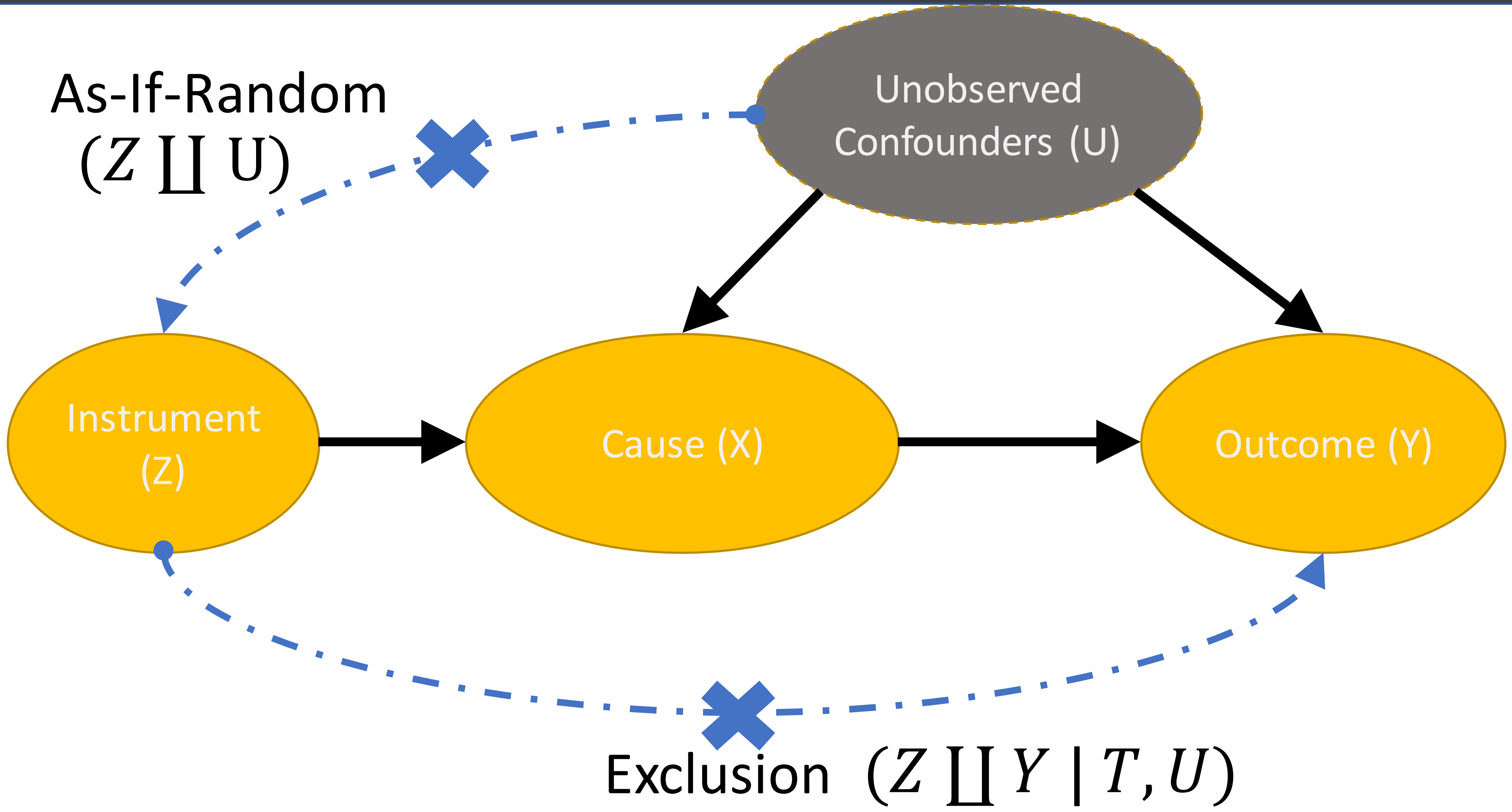
Part II.B. Natural Experiments

- As-if Random
- Instrumental Variables
- Regression Discontinuities

Prior setup can be generalized as search for an “instrumental variable”



Prior setup can be generalized as search for an “instrumental variable”



Intuition: Can use this variation to compute causal effect

An increase in Z can lead to a change in Y *only through* X.

So change in Y is a product of change in Z->X and X->Y arrows.

Compare the extent by which random assignment affects X versus Y.

Causal effect (X->Y) = $\frac{Y_{Z=1} - Y_{Z=0}}{X_{Z=1} - X_{Z=0}}$

A generalized natural experiment: Instrumental Variables

Can look at *as-if random* variations due to external events.

E.g.,

Experimental: Encouraging randomly selected users of an app to exercise.

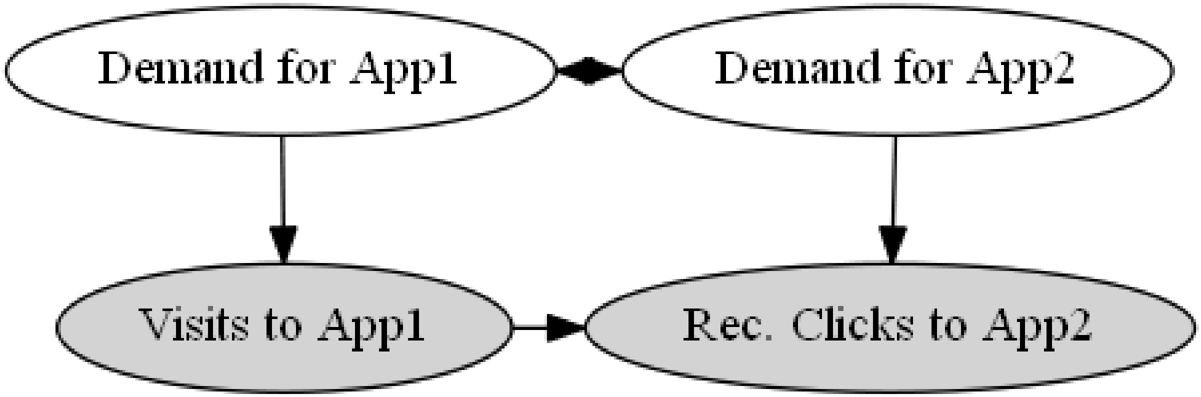
Observational: Looking at a past A/B test intervention that increased chances of exercise.

Example: What is the effect of recommendations on an app store?

Instrumental Variable: External sources that drive sudden, large traffic to an app.

Angrist-Pischke (2008)

Example: Effect of store recommendations

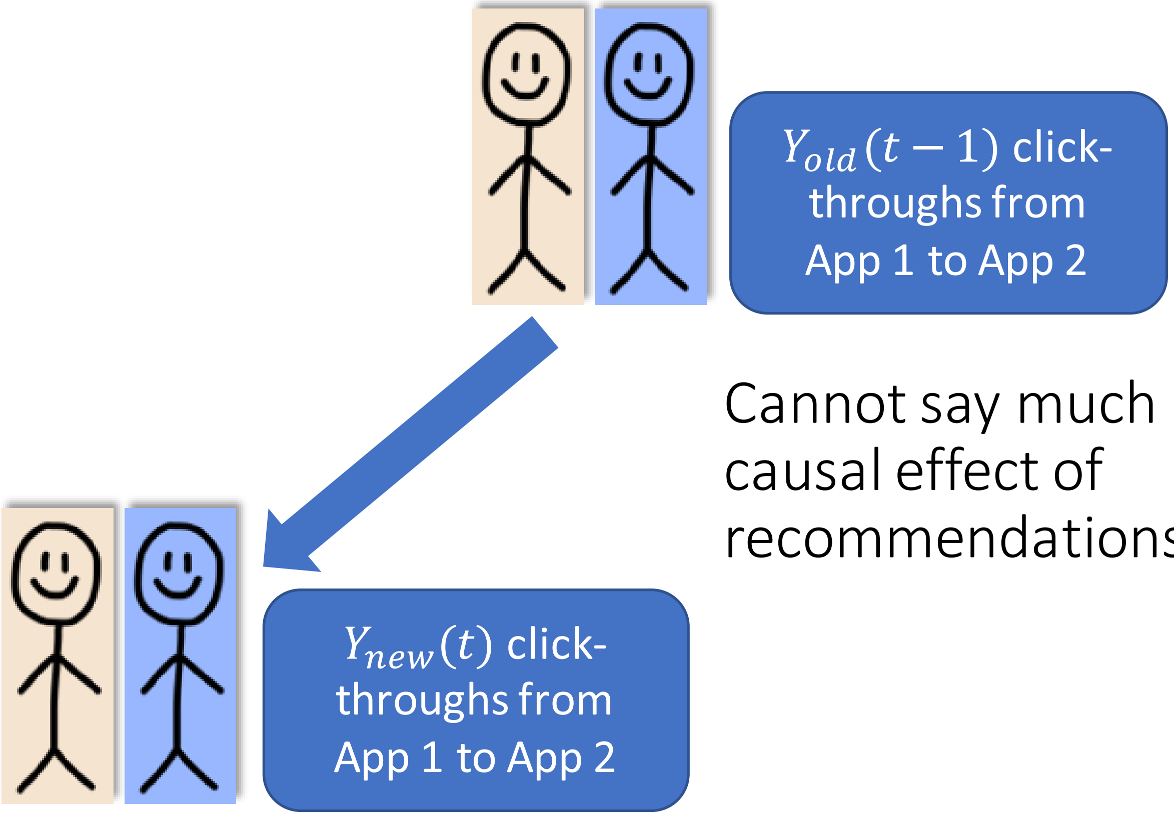


How many new visits are *caused* by the recommender system?

Demand for App 1 is correlated with demand for App 2.

⇒ Users would most likely have visited App 2 even without recommendations.

Traffic on normal days to App 1

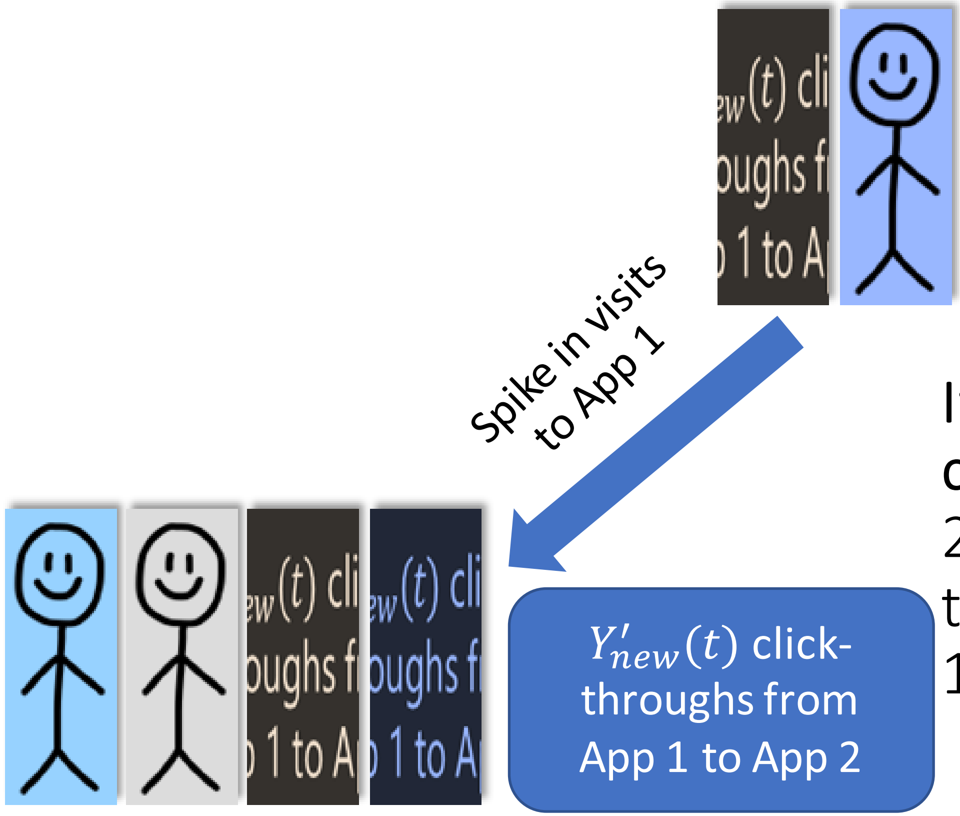


$Y_{old}(t-1)$ click-throughs from App 1 to App 2

$Y_{new}(t)$ click-throughs from App 1 to App 2

Cannot say much about the causal effect of recommendations from App 1.

External shock brings as-if random users to App1



$Y_{old}(t-1)$ click-throughs from App 1 to App 2

$Y'_{new}(t)$ click-throughs from App 1 to App 2

If demand for App 2 remains constant, additional views to App 2 would not have happened had these new users not visited App 1.

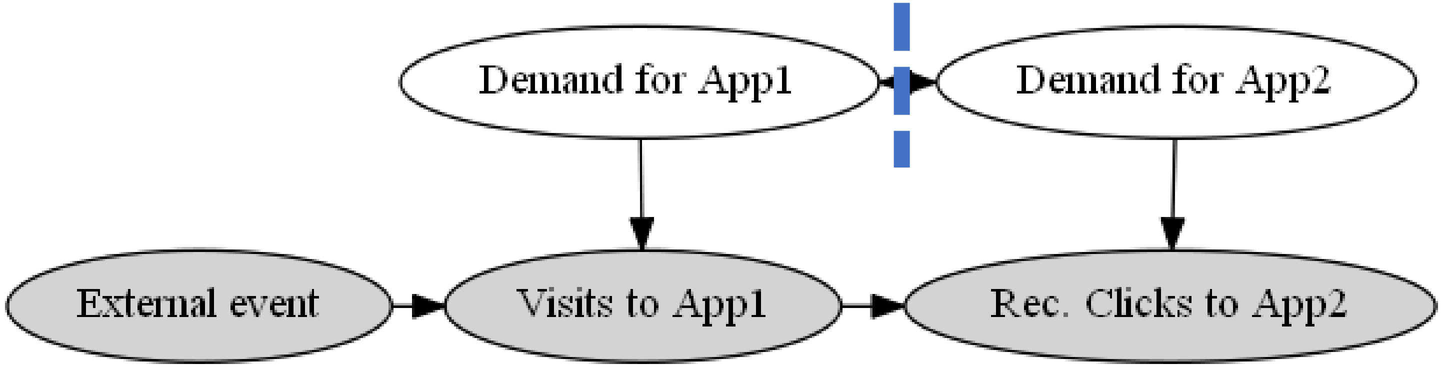
$$Causal\ clicks = Y'_{new}(t) - Y_{old}(t-1)$$

Sharma-Hofman-Watts (2015)

Exploiting sudden variation in traffic to App 1

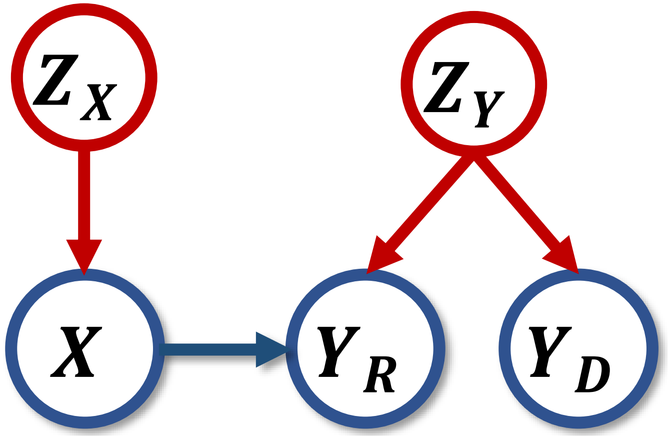
To compute Causal CTR of Visits to App1 on Visits to App2:

- Compare observed effect of external event separately on Visits to App1, and on Rec. Clicks to App2.
- Causal click-through rate = $\dfrac{\Delta(\text{Rec. Click–throughs from App1 to App2})}{\Delta(\text{Visits to App1})}$

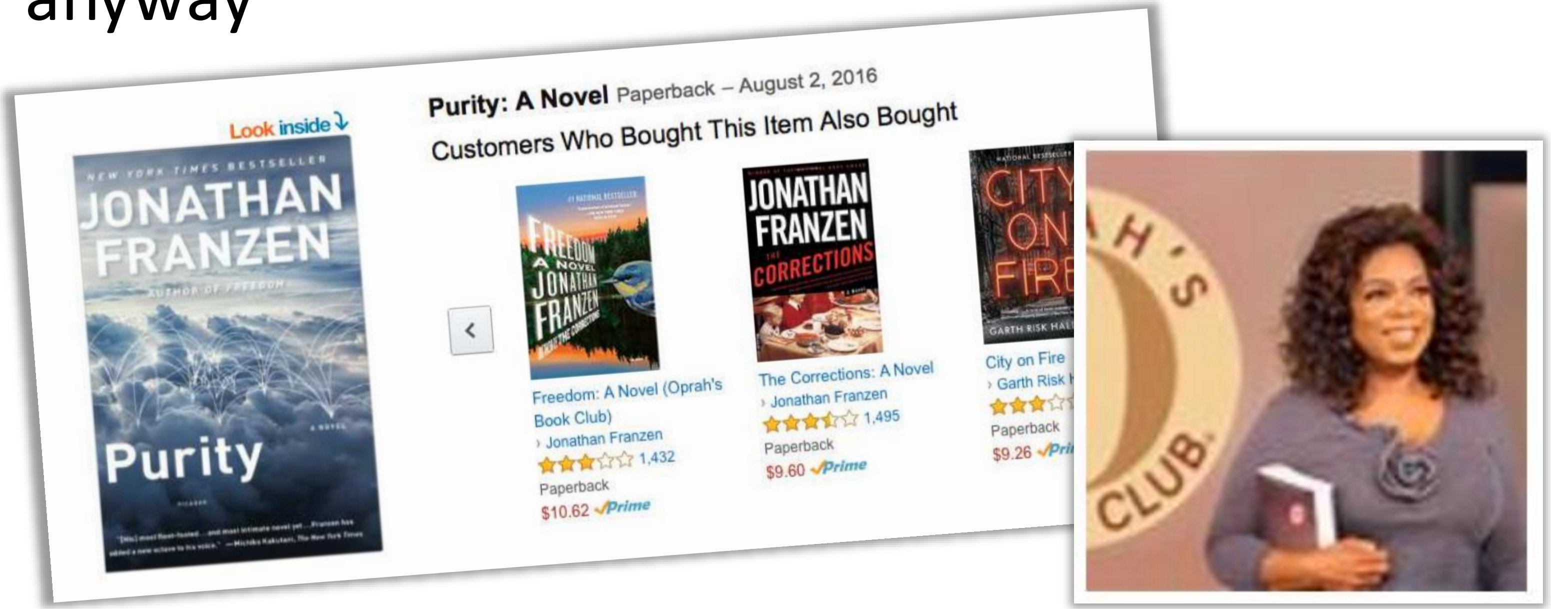


Automatically Identifying Natural Experiments

Split-Door Criterion



- Finds 7,000 natural experiments, instead of 133
- Result: Across 10 product categories, half of recommendation clicks would have happened anyway


Look inside
NEW YORK TIMES BESTSELLER
JONATHAN FRANZEN
AUTHOR OF FREEDOM
Purity
FREEDOM A NOVEL JONATHAN FRANZEN
JONATHAN FRANZEN THE CORRECTIONS
CITY ON FIRE
GARTH RISK HAL
H'S CLUB

Purity: A Novel Paperback – August 2, 2016

Customers Who Bought This Item Also Bought

Freedom: A Novel (Oprah's Book Club)
› Jonathan Franzen
★★★☆☆ 1,432
Paperback
$10.62 ✓Prime

The Corrections: A Novel
› Jonathan Franzen
★★★½☆ 1,495
Paperback
$9.60 ✓Prime

City on Fire
› Garth Risk
★★★☆☆
Paperback
$9.26 ✓Pri

Sharma et al. Split-door criterion for causal identification: Automatic search for natural experiments, 2016

$< T, Y >$

Examples of Instrumental Variables

▨ Lottery

▨ Weather

▨ Shocks

▨ Discontinuities

■ Hard-to-find variations

Dunning (2012)

But there are so many natural variations.

$< T, Y >$

Change in access of digital services

Change in train stops in a city

Change in medicines at a hospital

...

▨ Lottery

▨ Weather

▨ Shocks

▨ Discontinuities

■ Hard-to-find variations

What we just learned: Instrumental Variables

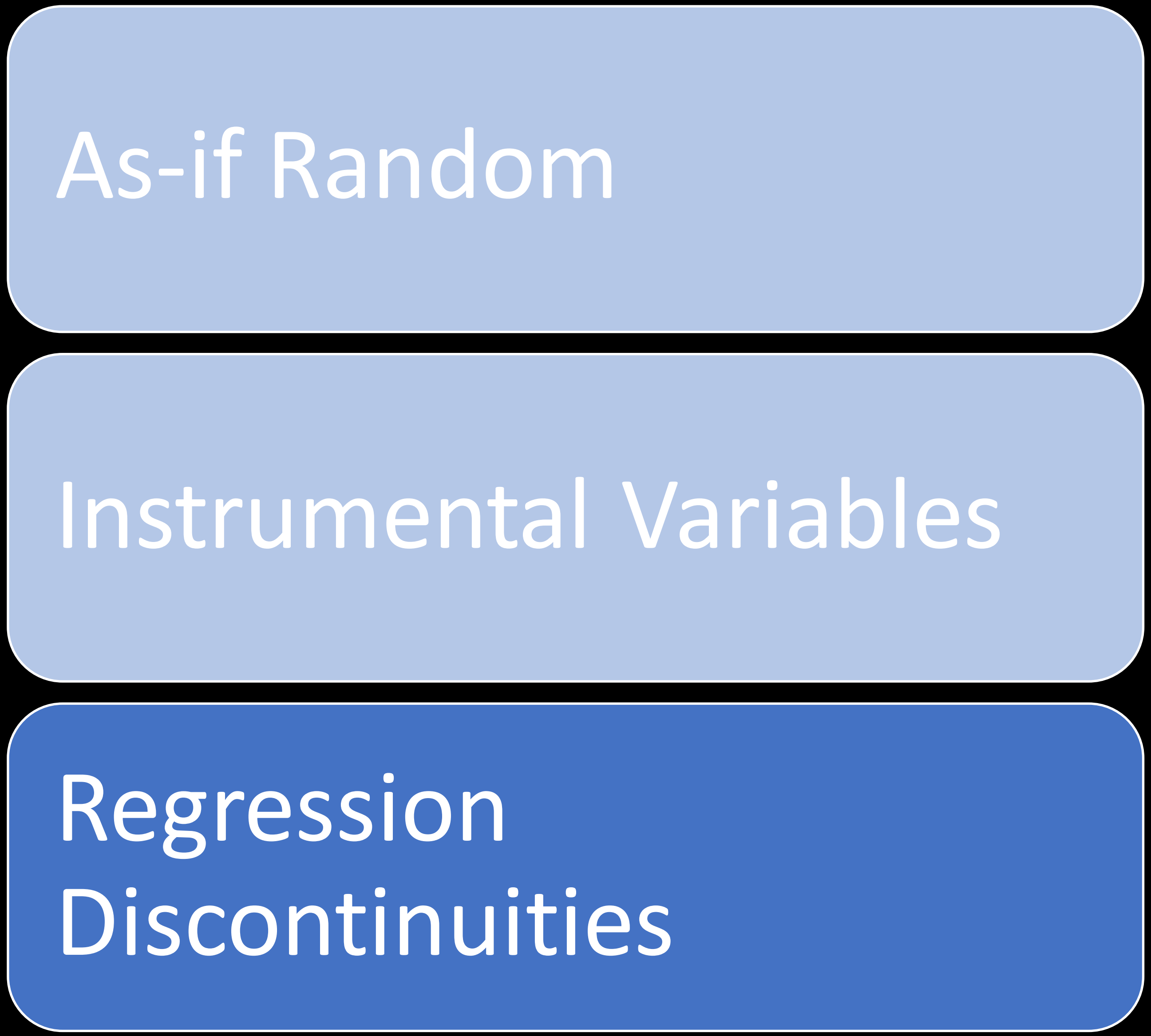
Definition Instrumental variables (IV) introduce "as-if random" noise into treatment assignment, and are used to estimate treatment effect

Intuition Because IVs are not influenced by confounds, IVs' indirect effect on outcome Y is independent of confounds too.
Because IVs do not directly influence outcome, their effect must be due to the effect of the treatment.

Examples Encouraging people to exercise at random.
Sudden increase in page visits to a product.

Keep in Mind Causal Estimate may not generalize to full population.
Estimate very sensitive to the violations of IV assumptions.

Part II.B. Natural Experiments



As-if Random

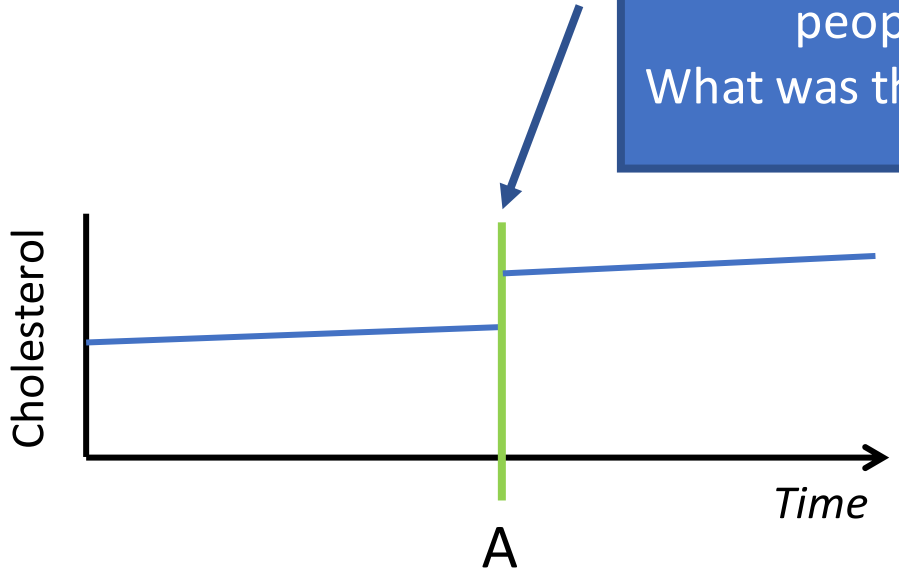
Instrumental Variables

Regression Discontinuities

Regression discontinuities: Look for arbitrary changes to treatment

Instead of an IV changing the distribution of treatment over individuals, an arbitrary change decides the treatment deterministically.

At time t, a cholesterol drug A is banned, and people switch to another drug B.
What was the relative effect of drug A over B ?

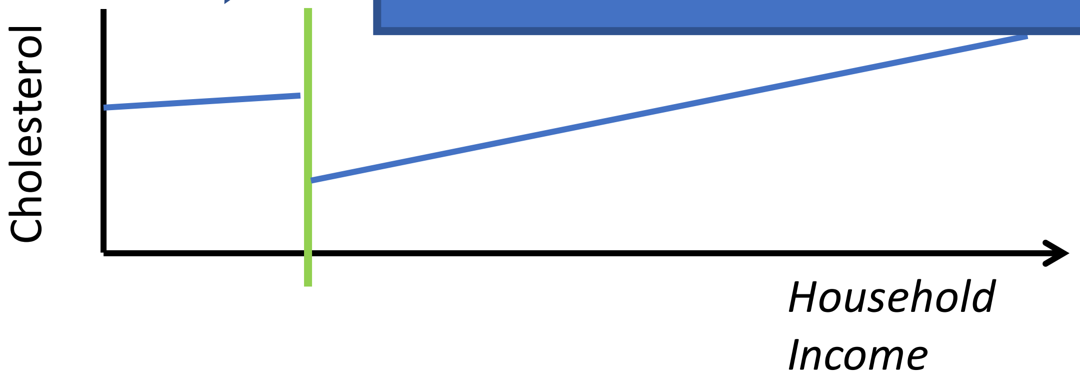


Due to selection effects, people taking drug A are different from those taking drug B.

But within [t-1, t+1] duration, patients of A and B can be assumed to be similar.

Regression discontinuities

Below income threshold t, free health insurance. What is effect of health insurance on cholesterol?



Due to selection effects, people with health insurance different from those without.

But within [t-1, t+1] income, people with or without health insurance are similar.

Regression discontinuities also depend on as-if-random and exclusion

As-if-random: People near the threshold are similar to each other, as if Nature randomized them on either side of the threshold.

Exclusion: Merely being on one side of the threshold does not affect the outcome.

Very common: Many decisions in organizations, arbitrary decisions in software are examples.

Can be thought of as a special case of an instrumental variable.

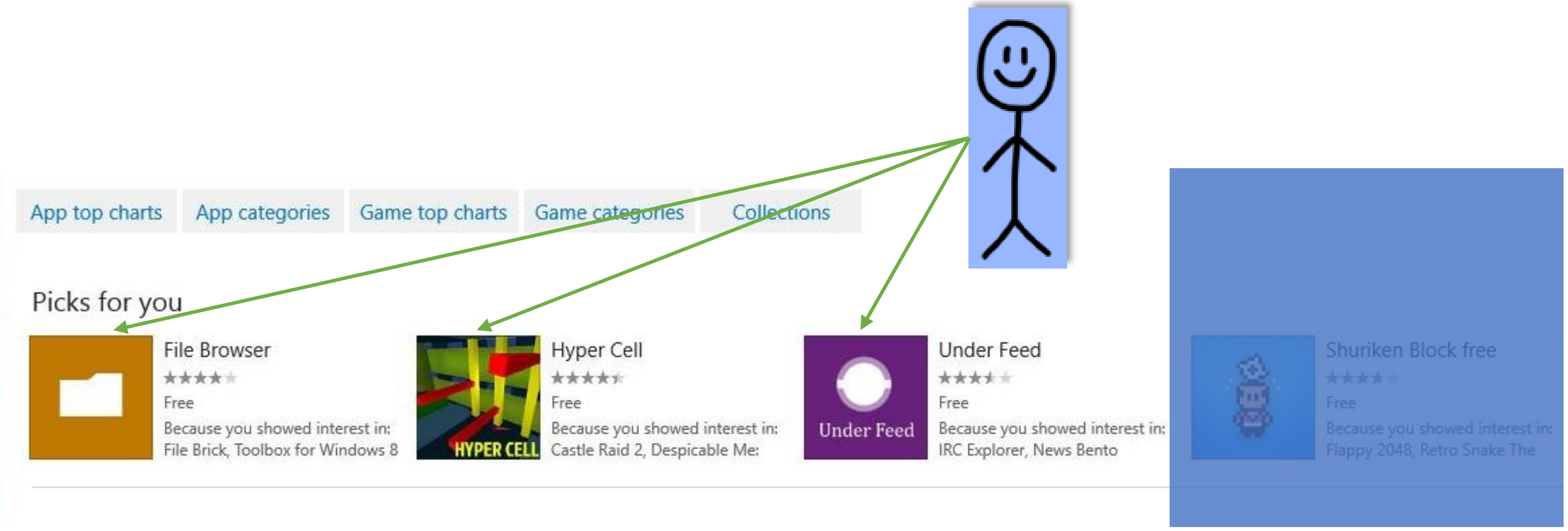
Example: Effect of Store recommendations

Suppose instead of comparing recommendation algorithms, we want to estimate the causal effect of showing *any* algorithmic recommendation.

Can be used to benchmark how much revenue a recommendation system brings, and allocate resources accordingly.

(and perhaps help analyze the tradeoff with users' privacy)

Exploiting arbitrary cutoffs to recommendations



Picks for you

File Browser
Free
Because you showed interest in: File Brick, Toolbox for Windows 8

Hyper Cell
Free
Because you showed interest in: Castle Raid 2, Despicable Me:

Under Feed
Free
Because you showed interest in: IRC Explorer, News Bento

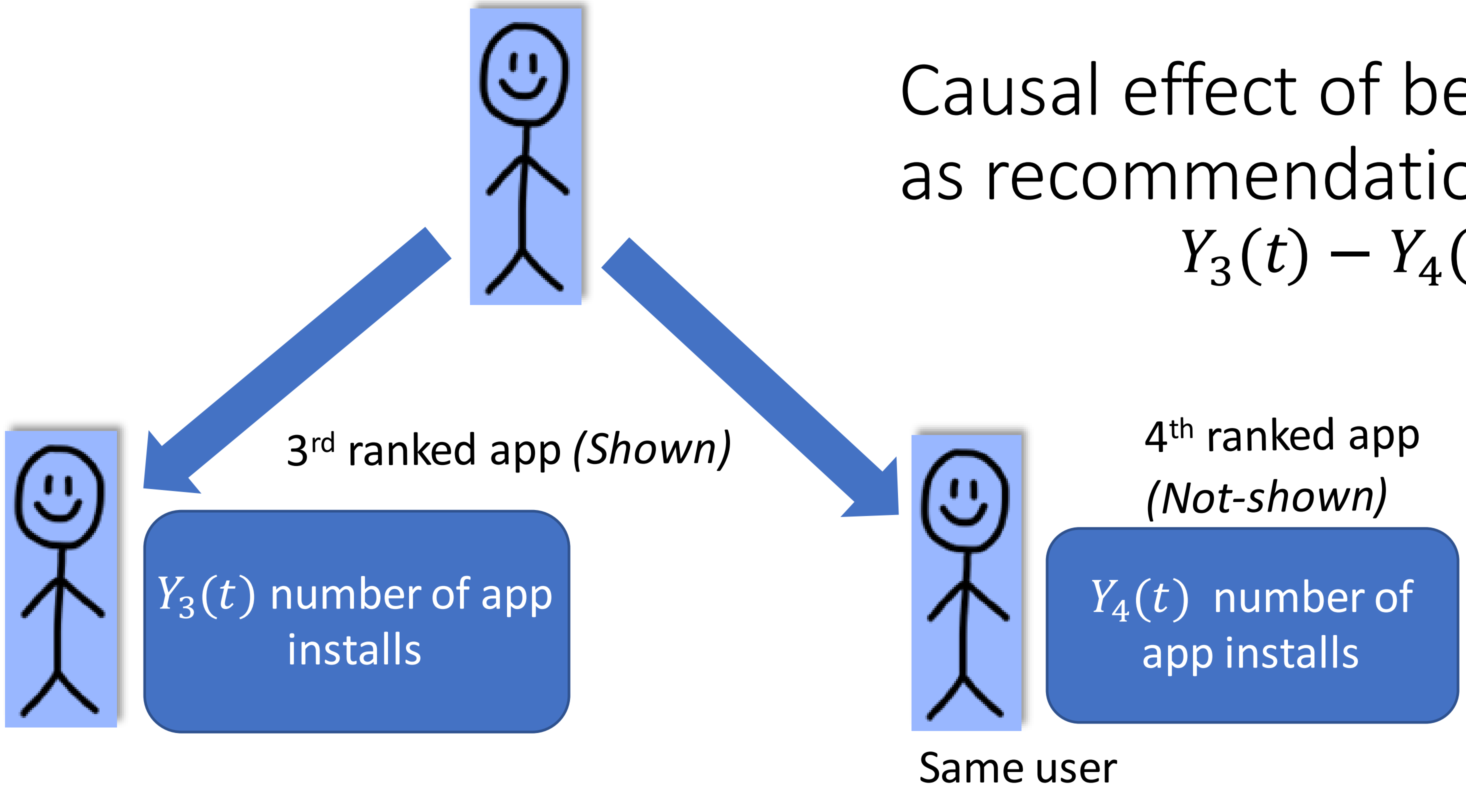
Shuriken Block free
Free
Because you showed interest in: Flappy 2048, Retro Snake The

Only 3 recommendations shown to user.

Assumption: Closely-ranked not-shown apps are as relevant as shown apps

Causal effect of being shown as recommendation

$$Y_3(t) - Y_4(t)$$



3rd ranked app *(Shown)*

$Y_3(t)$ number of app installs

4th ranked app *(Not-shown)*

$Y_4(t)$ number of app installs

Same user

Algorithm: Regression discontinuity

For any top-k recommendation list:

- Using logs, identify apps that were similarly ranked but could not make it to the top-k shown apps.
- Measure difference in app installs between **shown and not-shown apps** for each user.

What we just learned: Regression Discontinuities

Definition Regression discontinuities identify arbitrary boundaries between treated and untreated populations, measure treatment effect as difference in outcomes at the boundary

Intuition Regression discontinuities approximate randomized experiments as long as no substantial differences between people just on one side or the other. That is, at the boundary, $\mathrm{T} \perp\!\!\!\perp X, U$

Example Policy decisions based on income or time; exogenous shocks; and are all common sources of regression discontinuities

Keep in mind Only estimates treatment effect at the boundary. Effect may vary elsewhere!

PART II. Methods for Causal Inference

- Observational Studies
- Natural Experiments
- Refutations

Causal inference is only possible with assumptions

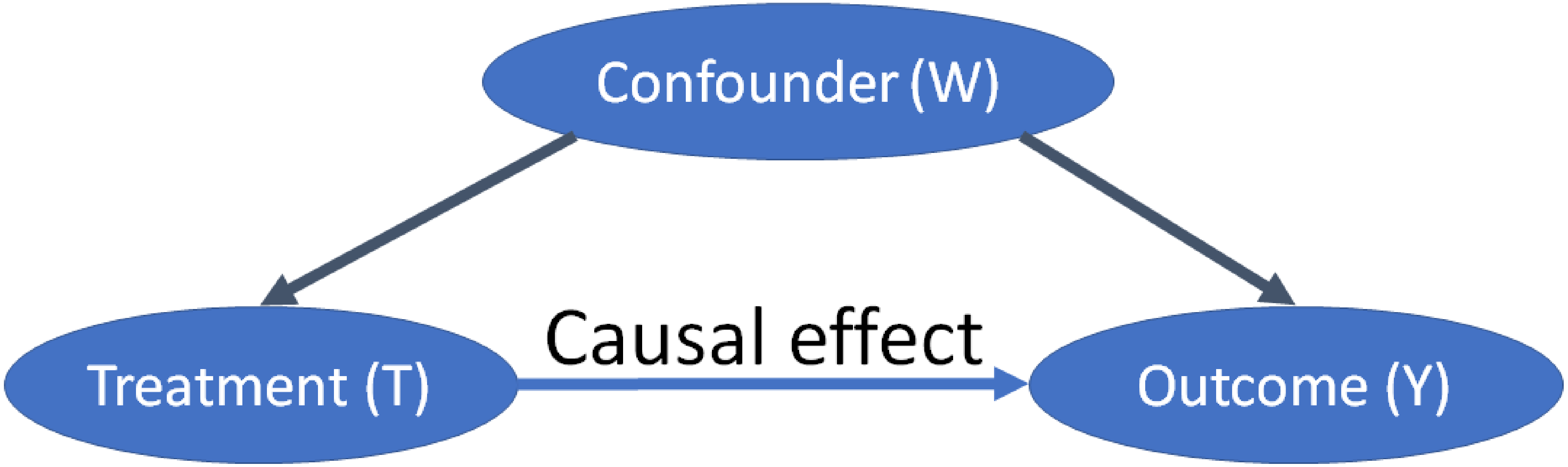
“Causal” part does not come from the data.

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



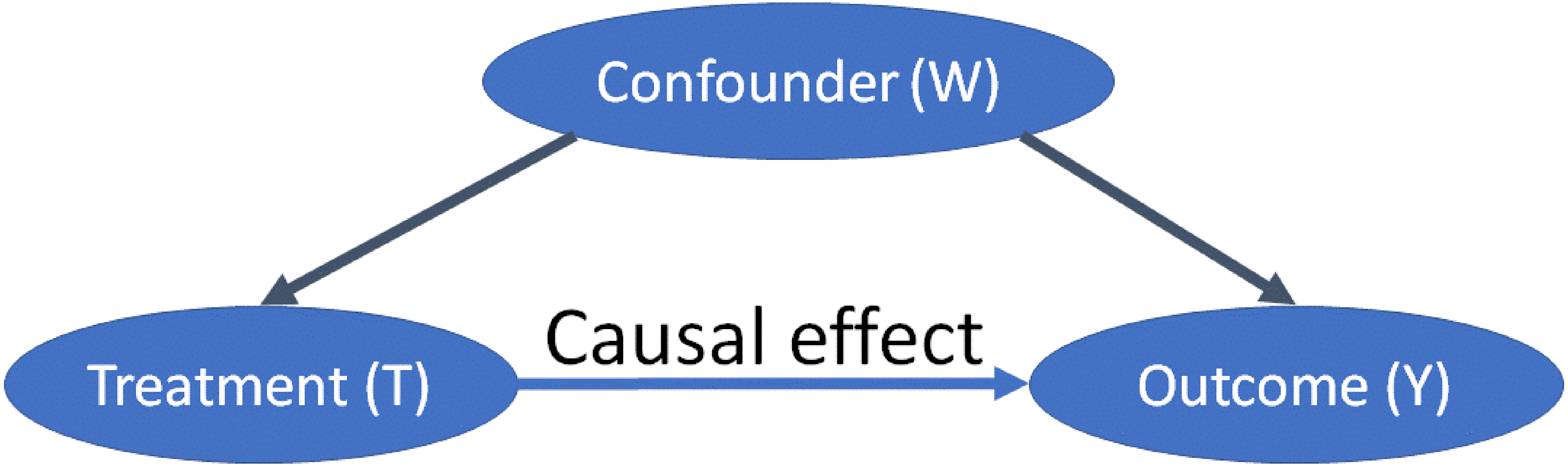
Identification: Causal effect → Observed effect conditioned on W, $\mathrm{E}[Y|T, W]$

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, relatively easy.

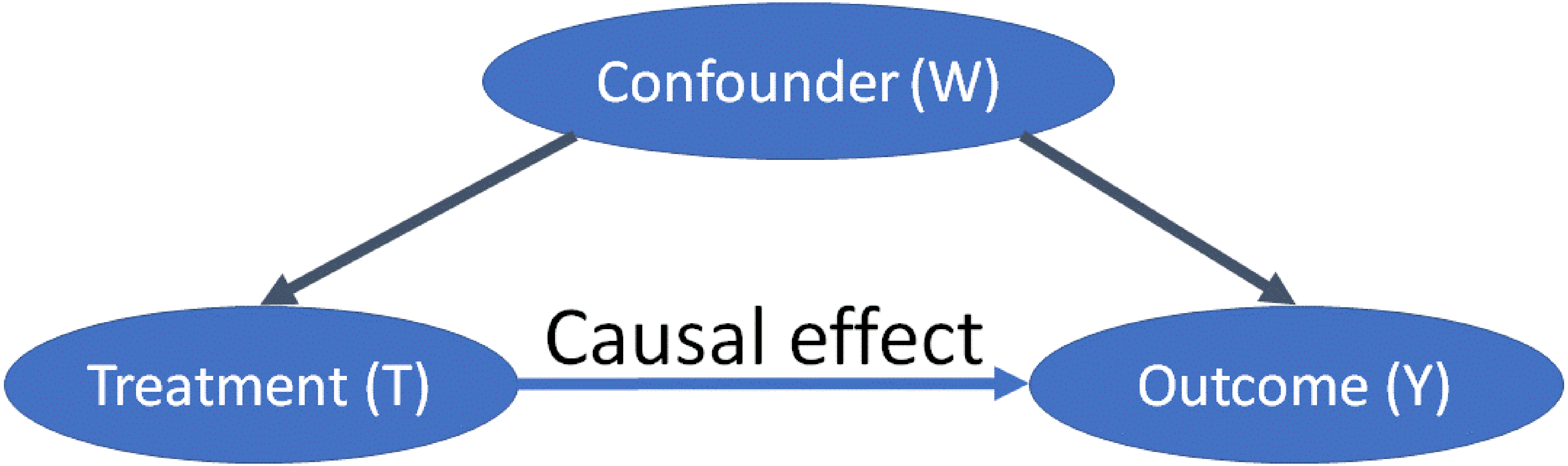
(Step 2): Explicitly represent your identifying and estimating assumptions.



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 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?

To make these steps easy, we created DoWhy: a python library for causal inference

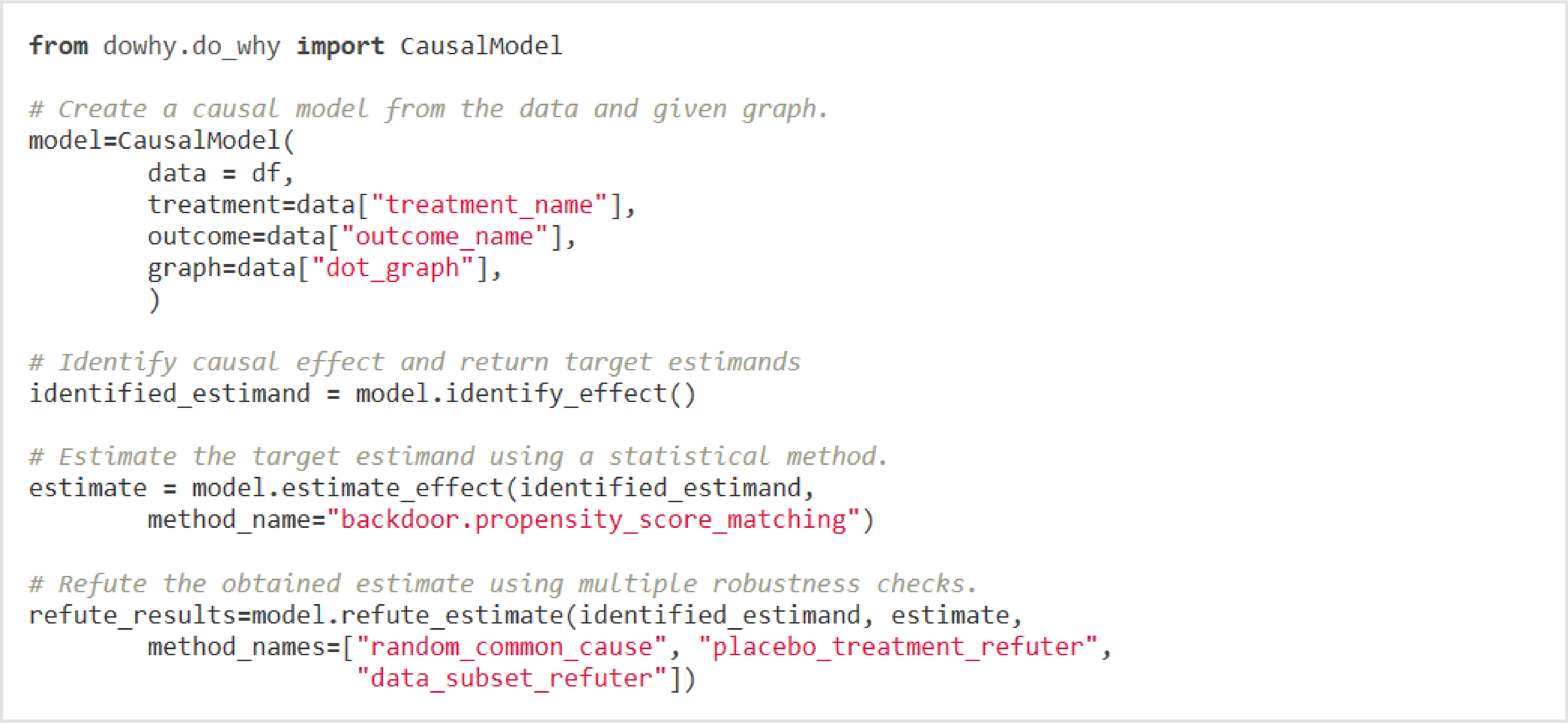
DoWhy focuses attention on the **assumptions** required for causal inference.

Provides estimation methods such as matching and IV so that you can focus on the identifying assumptions.

-- Models assumptions explicitly using causal graphical model.
-- Provides an easy way to test them (if possible) or analyze sensitivity to violations.

Unifies all methods to yield **four verbs** for causal inference:
-- Model
-- Identify
-- Estimate
-- Refute

DoWhy: Sample causal inference analysis in 4 lines



```
from dowhy.do_why import CausalModel

# Create a causal model from the data and given graph.
model=CausalModel(
        data = df,
        treatment=data["treatment_name"],
        outcome=data["outcome_name"],
        graph=data["dot_graph"],
        )

# Identify causal effect and return target estimands
identified_estimand = model.identify_effect()

# Estimate the target estimand using a statistical method.
estimate = model.estimate_effect(identified_estimand,
        method_name="backdoor.propensity_score_matching")

# Refute the obtained estimate using multiple robustness checks.
refute_results=model.refute_estimate(identified_estimand, estimate,
        method_names=["random_common_cause", "placebo_treatment_refuter",
                      "data_subset_refuter"])
```

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.

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

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

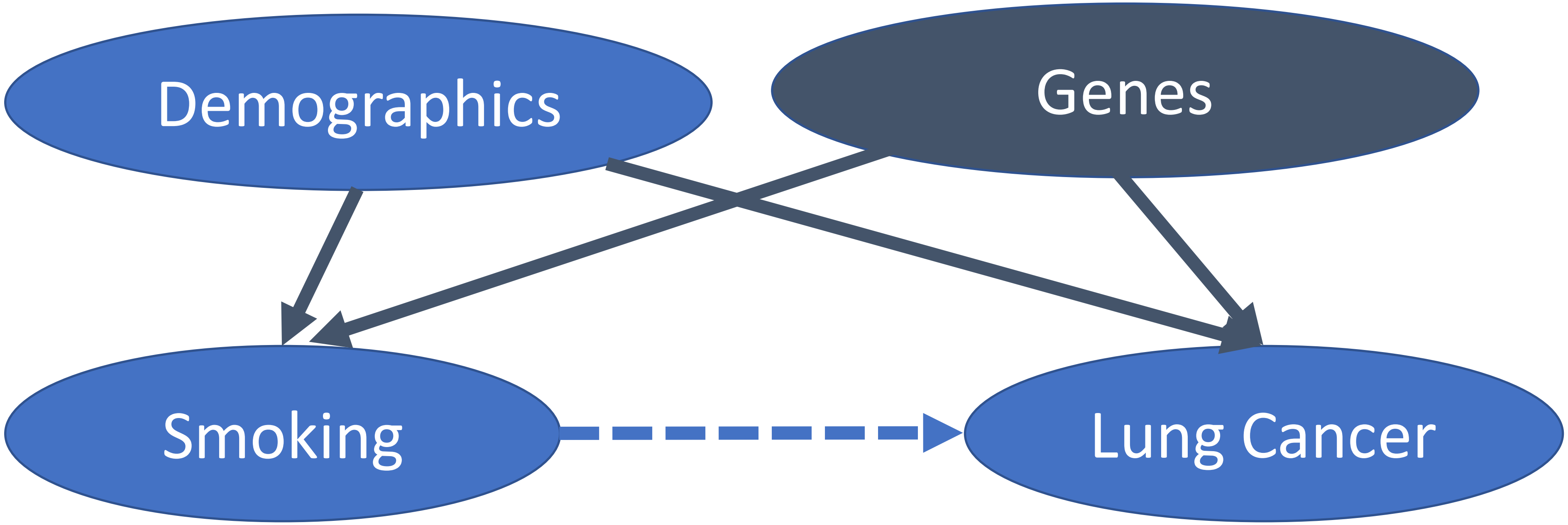
E.g. How big should the effect of a confounder be so that your estimate reverses in direction?

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.

Try out DoWhy to see best practices in action

DoWhy: A Python Library for Causal Inference

Principled: Converts prior knowledge to a formal causal graph

Simple: Automated analysis of many assumptions, one line of code for powerful causal inference algorithms

Robust: Battery of tests to refute obtained estimates

Modest: No estimate if the data is insufficient

- **Input:** Observational data, Causal graph
- **Output:** Causal effect between desired variables, "What-if" analysis

Code: **https://github.com/Microsoft/dowhy**

Docs: http://causalinference.gitlab.io/dowhy

PART I. Introduction to Counterfactual Reasoning

PART II. Methods for Causal Inference

PART III. Large-scale and Network Data

PART IV. Broader Landscape