Part2. Basic Causal Effect Identification

Hang Wu

Dr. May D. Wang

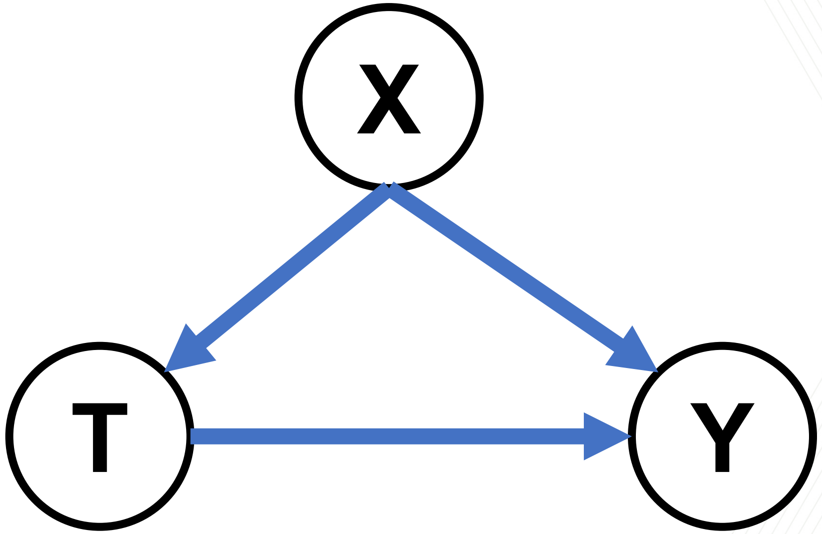
Dept. Biomedical Engineering, Georgia Tech

Link: https://github.com/hang-wu/CI

Georgia Tech
CREATING THE NEXT

Problem Setup

- Notations:
  - T_i: treatment $\{0,1\}$
  - X_i: features
  - $Y_i(t)$: the potential outcome under treatment t
  - Y_i: observed treatment outcome



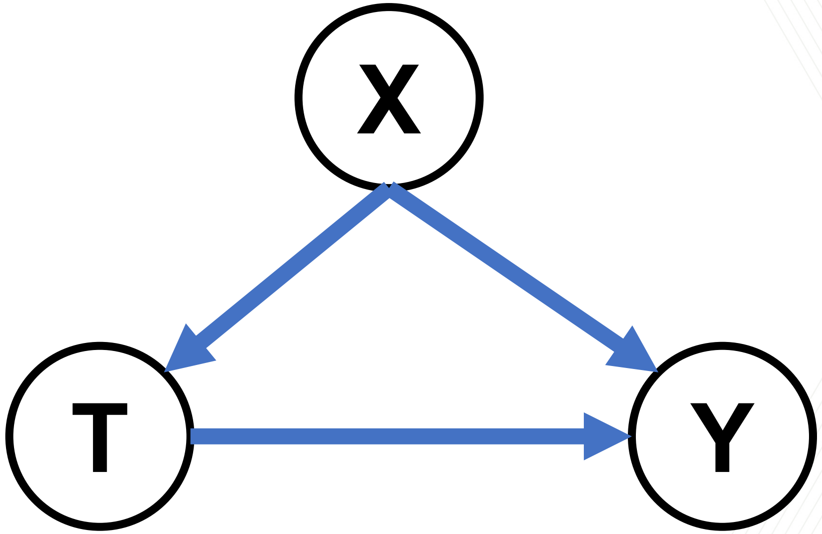
$$Y_i = T_i Y_i(T=1) + (1-T_i) Y_i(T=0)$$

Problem Setup

- Objective:
  - Estimating the average effect of treatment T on Y:

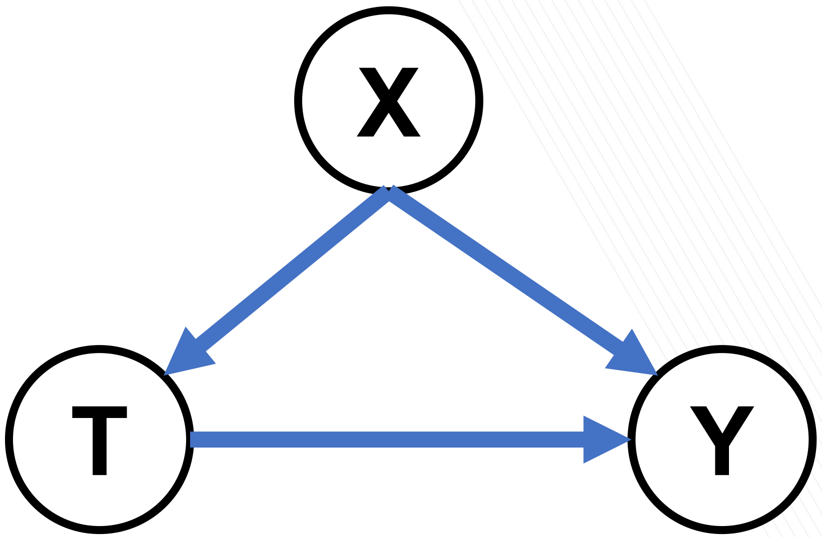
$$\mathrm{ATT} = E[Y_i(1) - Y_i(0)]$$

- Q: So why this is challenging?



A simple numerical example

- In our example
  - $X \sim some\ random\ distribution$
  - $T = 2X + 0.01 \ * \ N_T(0,1)$
  - $Y = 4X + 3T + N_Y(1,1) = 5T + Noise$




```
import numpy as np

X = np.random.randint(0, 10, size=6)
T = 2*X + np.random.randn(6) * 0.01
Y = 4. * X + 3. * T + np.random.rand(6) * 0.01
```

A simple numerical example: When we only observe (T, Y)

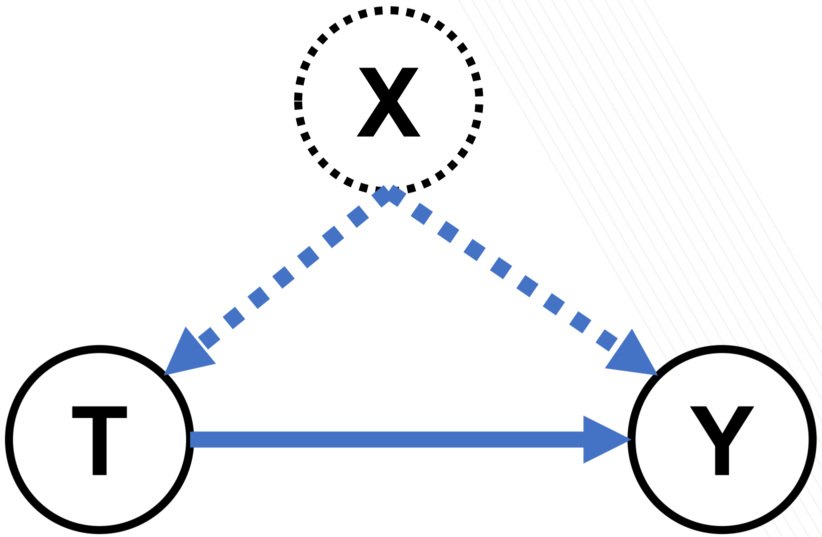
Patient	T	Y(:.1f)
P1	3	30.0
P2	1	10.0
P3	4	40.0
P4	1	9.9
P5	4	3.9
P6	0	0.0

- If we fit a linear regression model using OLS

$$\beta = (T'T)^{-1}TY$$

- We get $\hat{\beta} \approx 5$
- A biased result

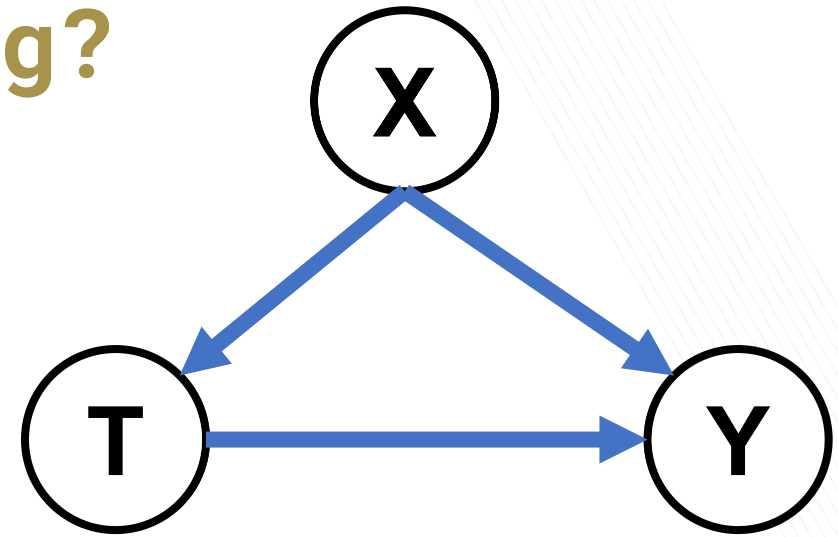
A simple numerical example



- In our example
  - $X \sim N(0, 1)$
  - $T = 2X + N_T(0.5, 1)$
  - $Y = 4X + 3T + N_Y(1, 1) = 5T + Noise$
- The treatment effect of T on Y should be 3 (i.e., when we keep X unchanged, changing T from 0 to 1 changes Y by 3 units)
- Confounding of X influences both T and Y

How to deal with confounding?

- Ideal case: We can break the dependence of T on X
  - Randomly assign T
  - $E[Y_i(1) - Y_i(0)] = E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$

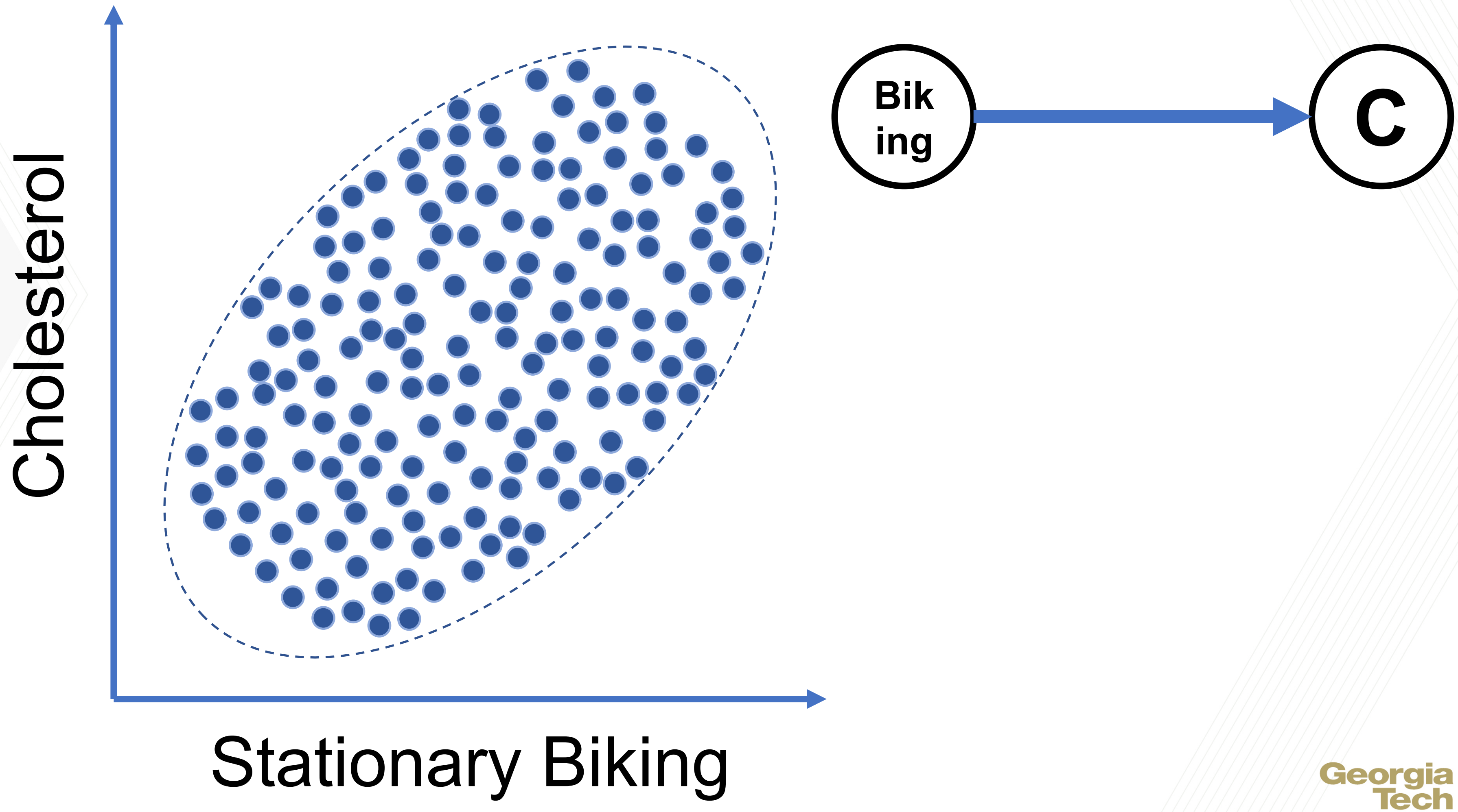


Q: What if we cannot intervene?

Estimation using Observation Data

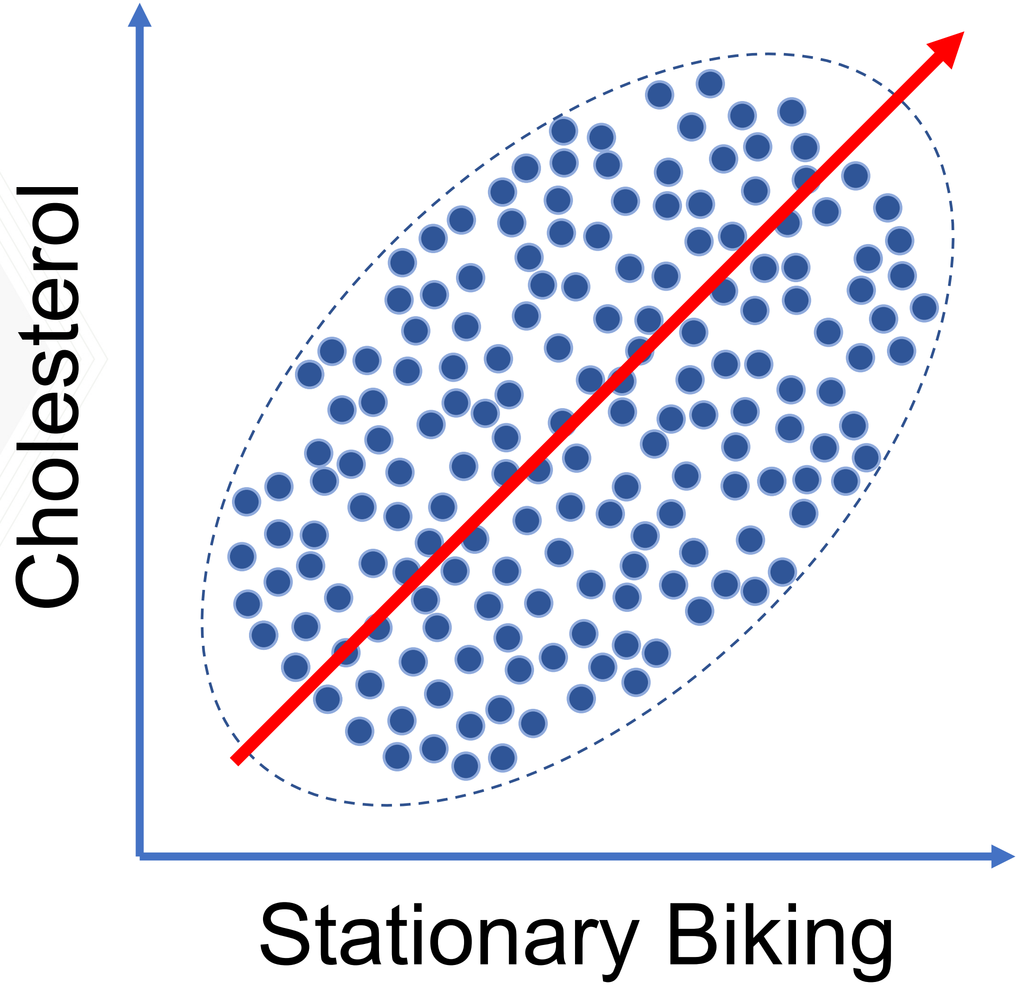
Georgia Tech
CREATING THE NEXT

How to deal with confounding?

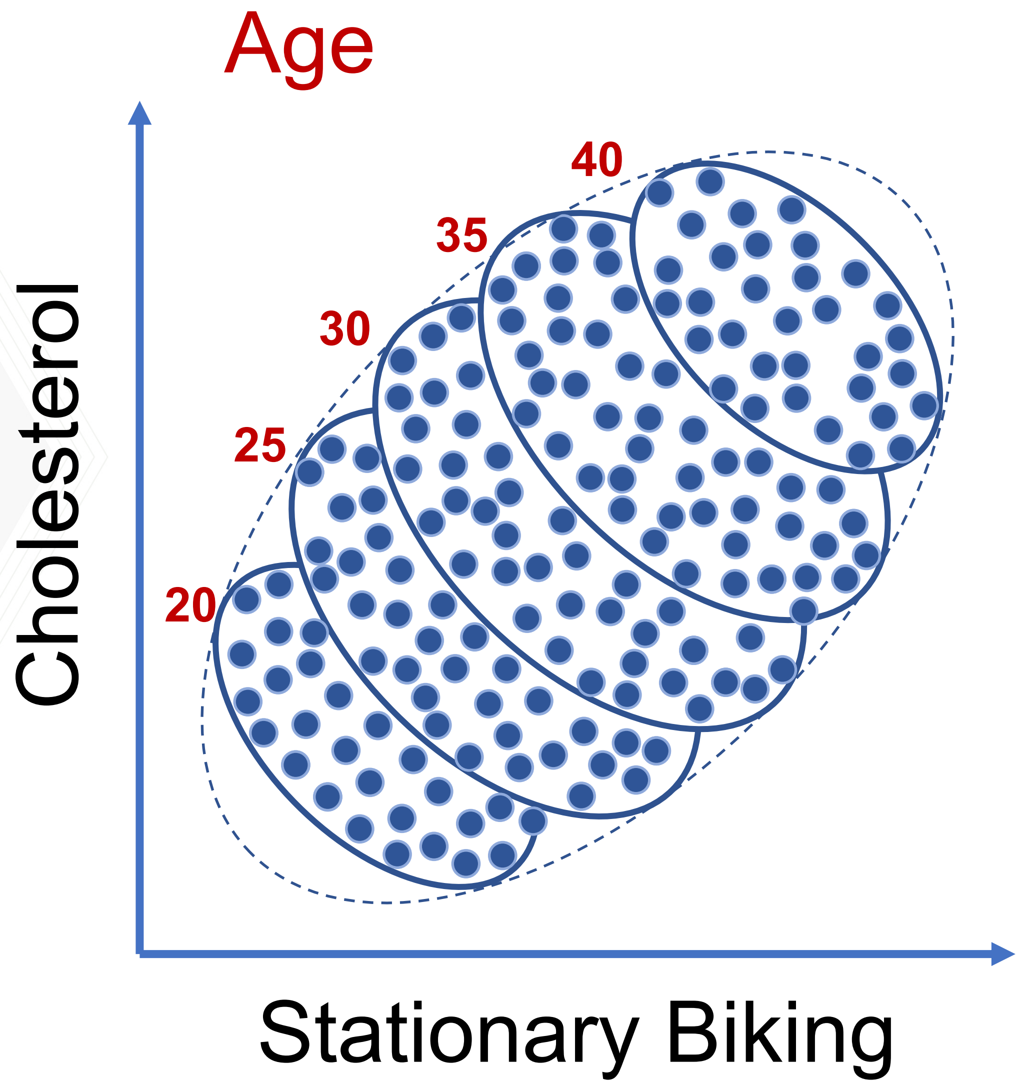


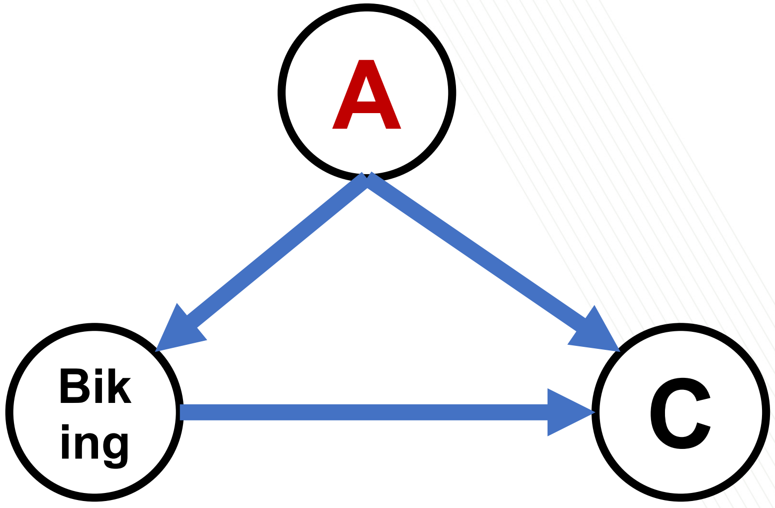
Example/Slides credit: [Amit Sharma](#) ([@amt_shrma](#)), [Emre Kiciman](#) ([@emrek](#))

Does more *stationary biking* lead to higher *cholesterol?*

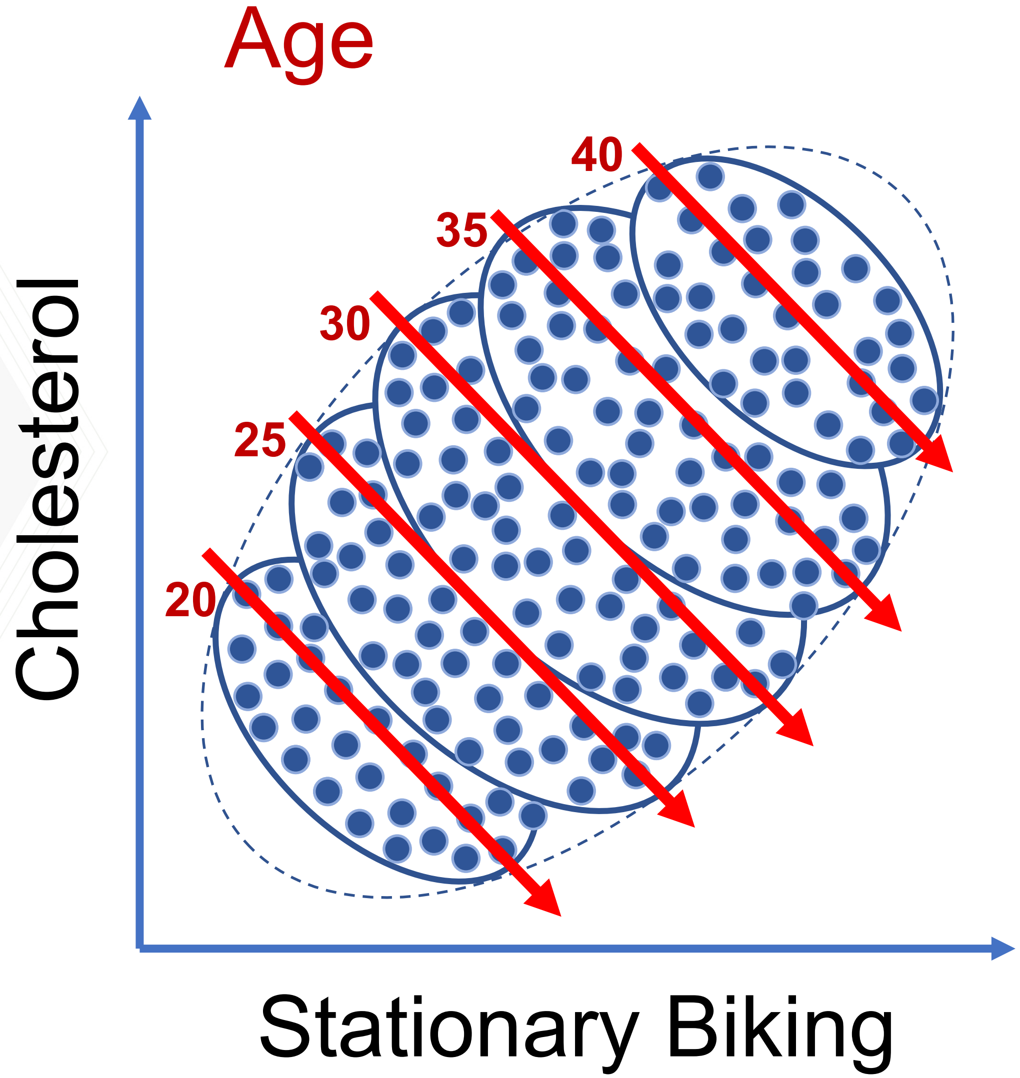


There is a confounder - age





We can condition on age



Recap

- *Age influences both stationary biking and cholesterol => confounder*
- We condition on age (by analyzing each age group separately)
- And find stationary biking now seems to lead to lower cholesterol

Identification vs Estimation

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

- Left hand-side:
  - A causal quantity
- Right hand-side:
  - A statistical quantity
- Using our causal knowledge, causal effect identification => statistical estimation problem

Conditioning

- Key intuition:
  - Conditioning on age, we have random assignments
  - => Lots of small RCTs

Assumptions We Made

- A1. Conditional Ignorability/ Unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp T_i | X_i = x \; for \; any \; x$$

  - We are talking about potential outcomes, not the observed outcomes

$$Y_i = T_i Y_i(T = 1) + (1 - T_i) Y_i(T = 0)$$

  - Among units with identical values of Xi , Ti is “as-if” randomly assigned.

Assumptions We Made

- A2. Common Support/ Positivity

$$0 < \Pr(T_i = 1|X_i = x) < 1 \; for \; any \; x$$

  - With any value of X_i , unit could have received either treatment or control.

Assumptions We Made

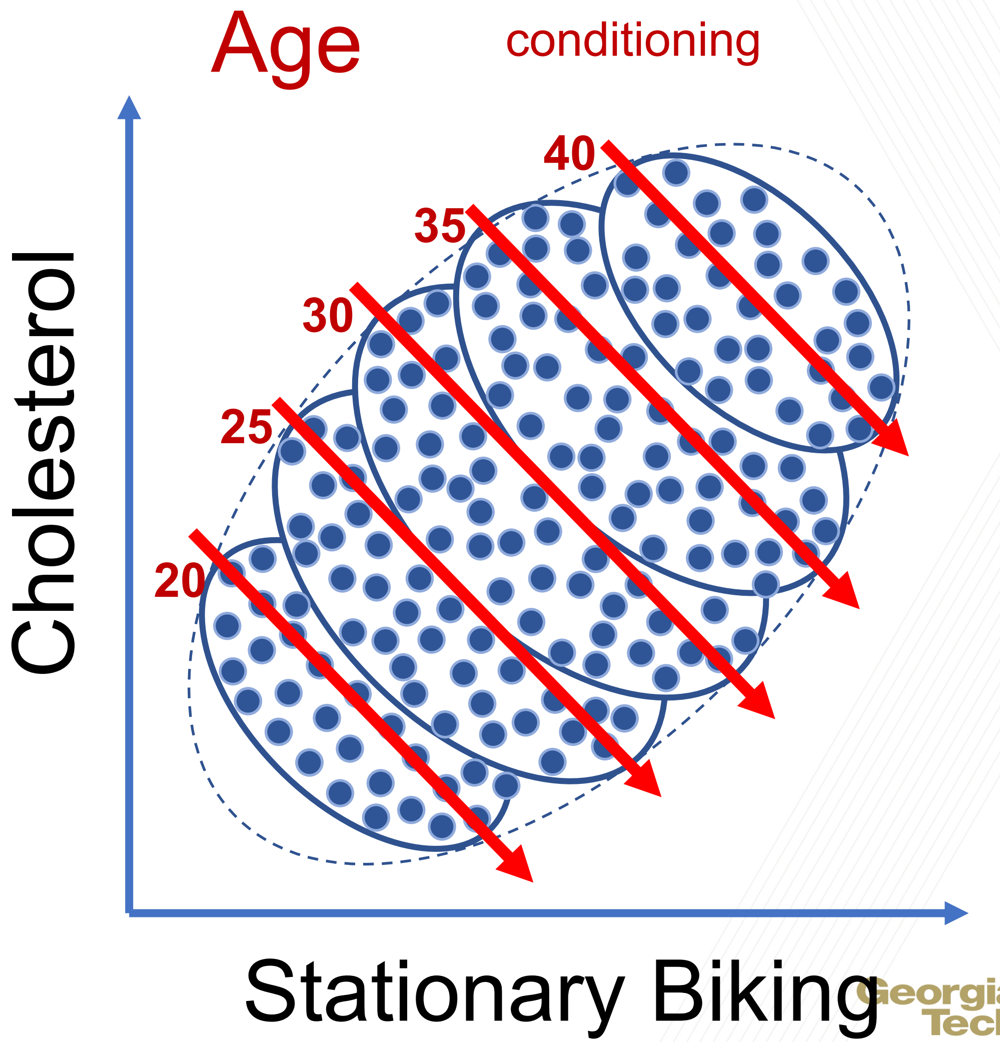
- A3. Stable Unit Treatment Value (SUTVA) assumption
  - the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units
  - No network effects

Estimation Under Unconfoundeness

- Case 1: Subclassification/Conditioning
  - When we have discrete variables

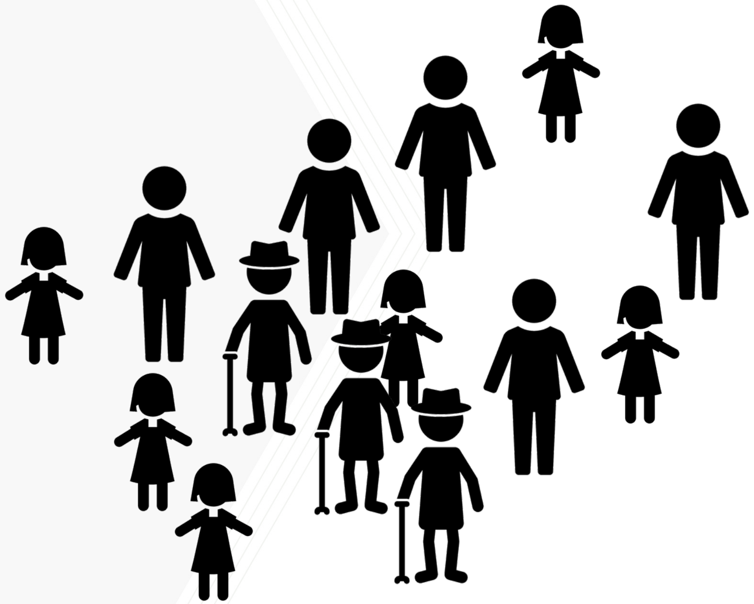
Discrete Features

- $ATE = \sum\{E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x]\}\Pr(X_i = x)$
- That is, we can
  1. Group units into strata by values of X_i
  2. For each strata, compute the difference in outcome between treated and untreated
  3. Calculate the weighted average of Step 2.

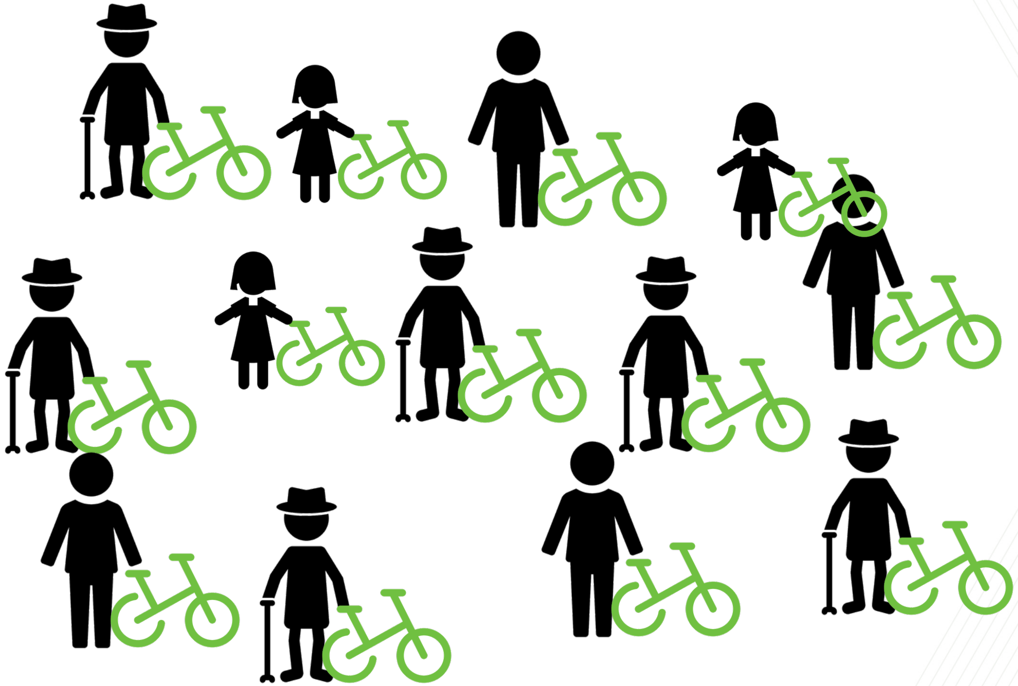


Estimation Under Unconfoundeness

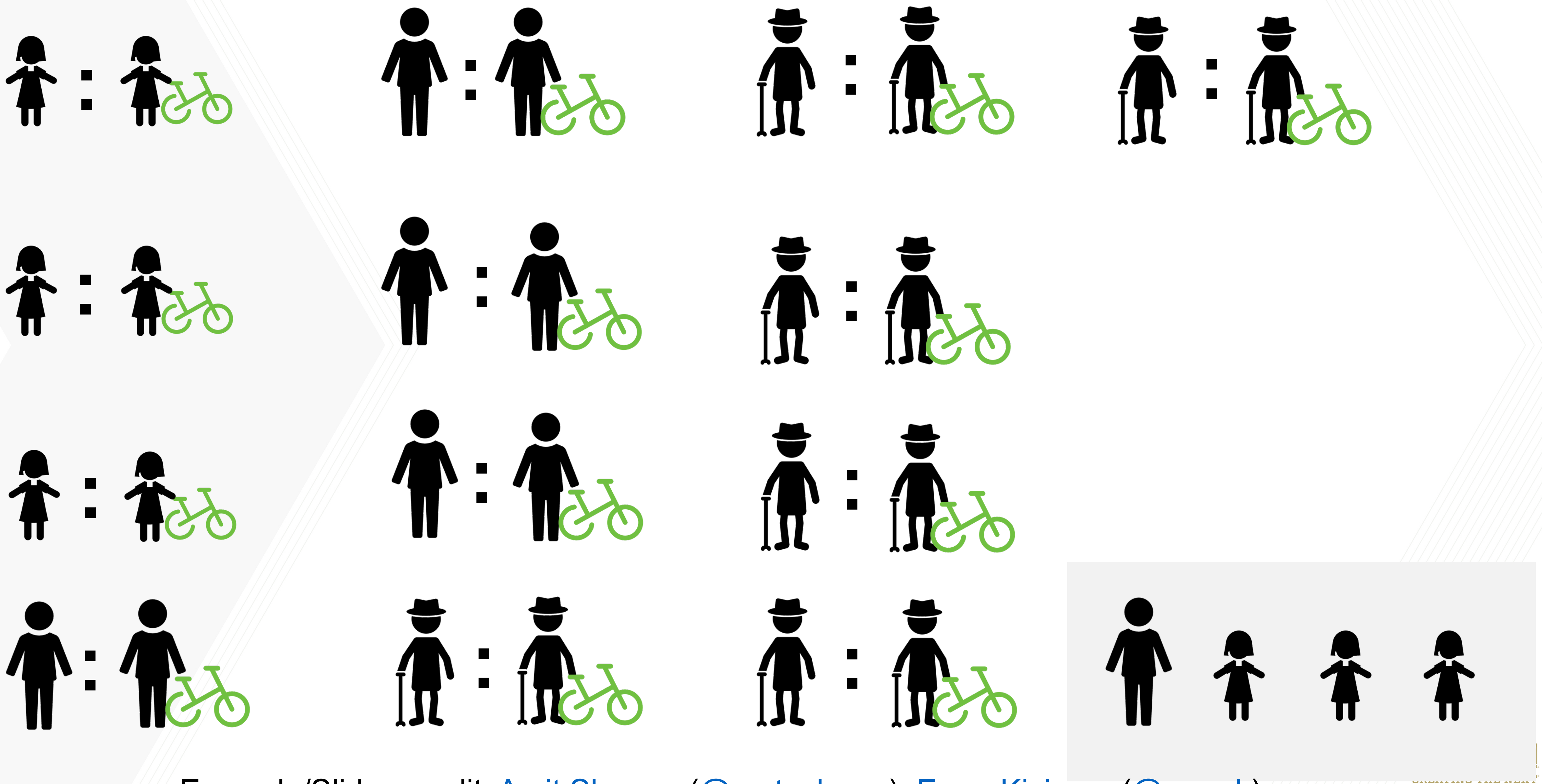
- Case 1: Subclassification/Conditioning
  - When we have discrete variables
- Case 2: Matching
  - When we have some/all continuous variables
  - Intuition: Find a pair of twins with opposite treatment



Avg Cholesterol = 200



Avg Cholesterol = 206



Example/Slides credit: Amit Sharma (@amt_shrma), Emre Kiciman (@emrek)

Matching

1. For each observation i, find an observation $\tilde{\imath}$
   - in the opposite group
   - with the most similar values of X: $Distance(X_i, X_j) < \epsilon$
2. Estimate ATE by the average difference between the pairs:

$$\tau_{ATT} = \frac{1}{n}\sum_{i}(Y_i - Y_{\tilde{\imath}})(-1)^{T_i+1}$$

where $\tilde{\imath}$ is the matched closest unit to the unit i with contrary treatment

Note: can match to multiple

Distance metrics for matching

- Mahalanobis distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma_X^{-1} (X_i - X_j)}$$

  - where Σ_X^{-1} is the (sample) variance-covariance matrix

Propensity Score

- Propensity score is an individual's *probability to be treated*

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

- Propensity scores are estimated or modeled, *not observed*.
- $\{Y_i(0), Y_i(1)\} \perp T_i |\, e(X_i)$

Distance metrics for matching

- Mahalanobis distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma_X^{-1} (X_i - X_j)}$$

  - where Σ_X^{-1} is the (sample) variance-covariance matrix
- Propensity scores:

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

  - The probability of a unit being treated
  - $D(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$

Estimation Under Unconfoundedness

- Case 1: Subclassification/Conditioning
  - When we have discrete variables
- Case 2: Matching
  - When we have some/all continuous variables
  - Intuition: Find a pair of twins with opposite treatment
- Case 3: Weighting
  - We can think of as a continuous version of matching
  - Intuition: for each i, a proportion of $j \neq i$ is matched to i

Weighting

- Under the conditional ignorability and common support assumptions

$$ATE = E[Y_i \frac{T_i - e(X_i)}{e(X_i)(1 - e(X_i))}]$$

- These can be thus estimated using sample averages:

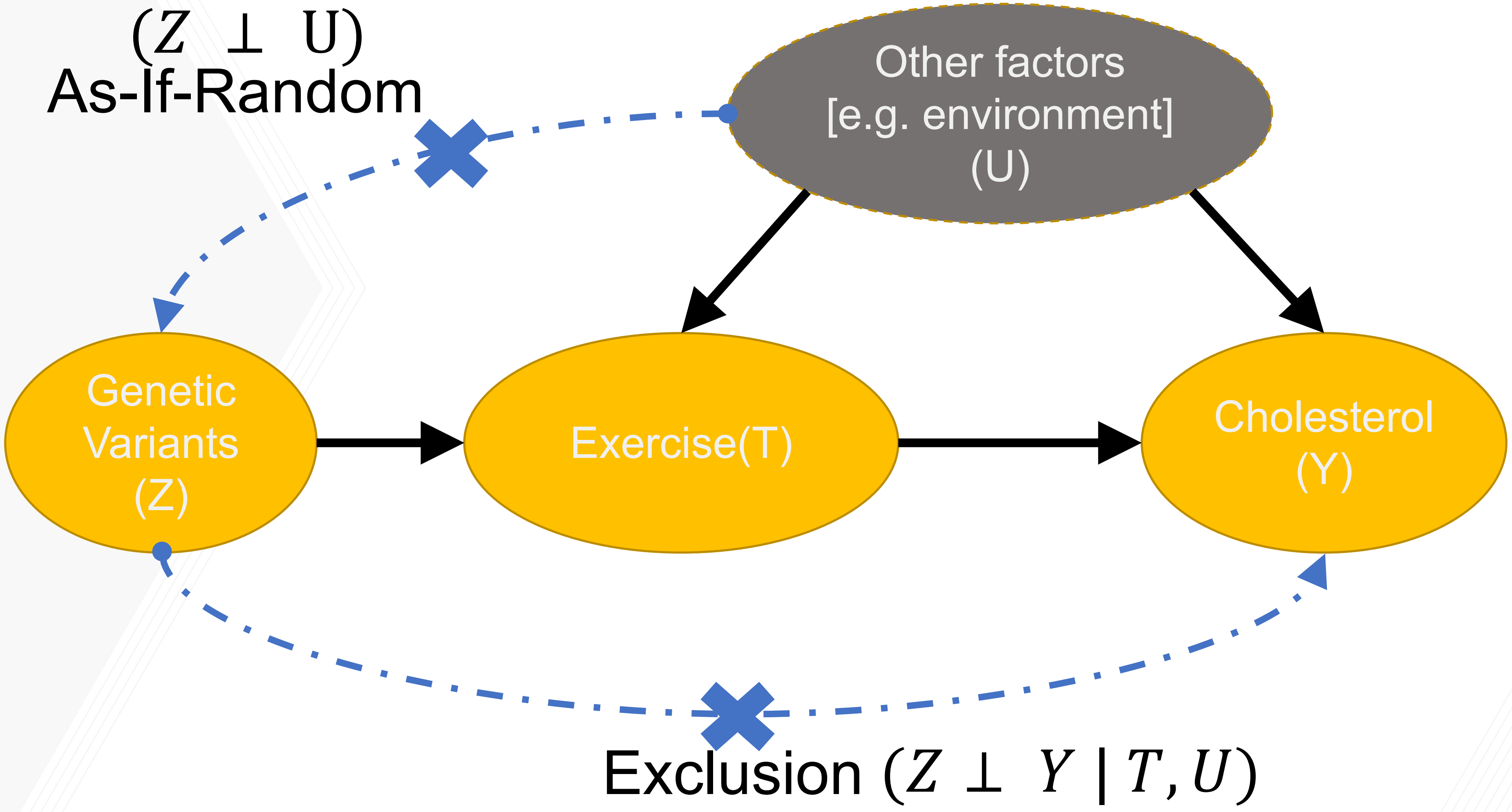
$$\tau_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \frac{T_i - \hat{e}(X_i)}{\hat{e}(X_i)(1 - \hat{e}(X_i))}$$

- These inverse PS weighting (IPW) estimators are consistent, but not unbiased.

Inference without Conditional Ignorability

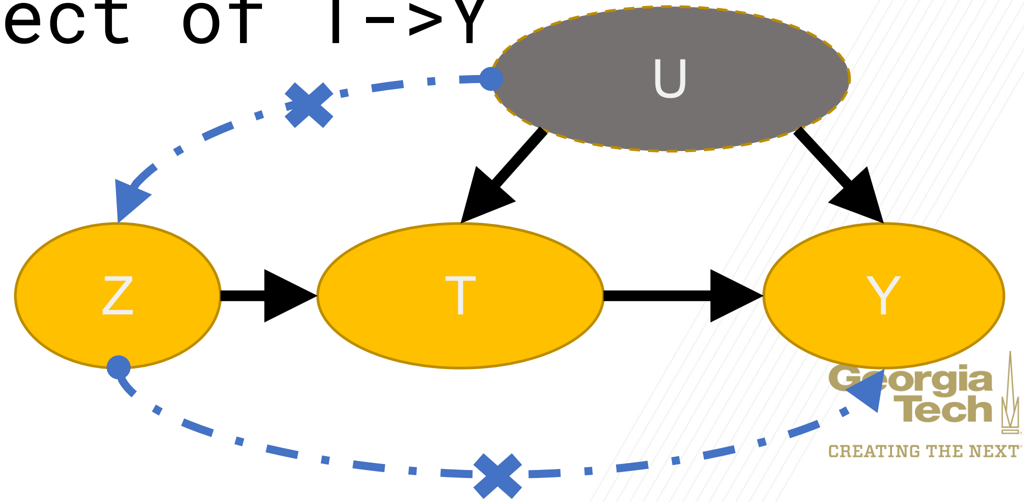
- Natural Experiment
  - Instrument variables: find extra variables

Instrument Variables



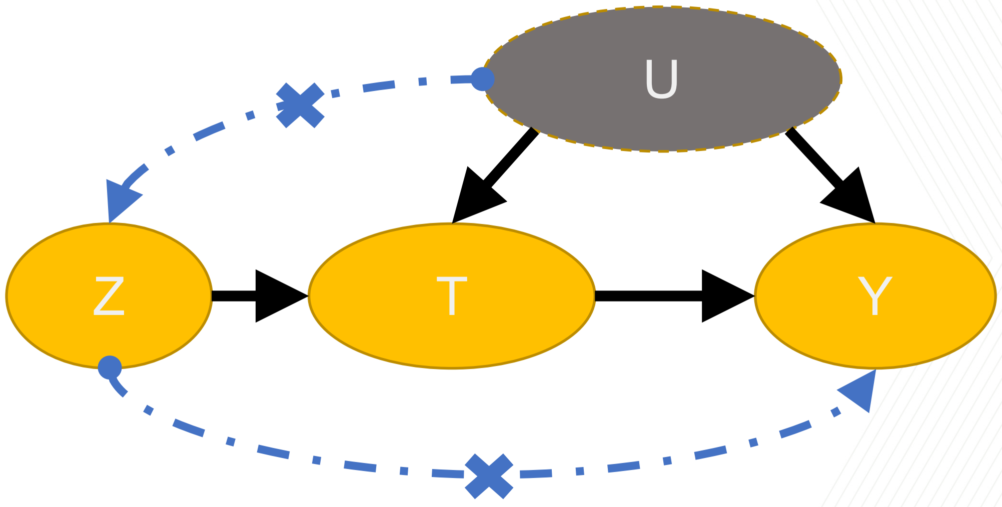
Intuition

- An increase in Z can lead to a change in Y *only through* T.
- So change in Y is a product of change in Z->T and T->Y arrows.
- If we identify:
  - Effect of Z->T
  - Effect of Z->Y
- We can identify the causal effect of T->Y



A simple example

Patient	Z	T	U	*Y*
P1	0.5	3	2.00	5.1
P2	1	6	4.01	9.9
P3	0	0.05	0.1	0.01
P4	0.5	3.01	1.99	4.95
P5	1	5.99	3.98	10.32
P6	1.5	9.01	6.02	15.01



T= U + 2Z + Noise

Y= T + U + Noise

Direct Estimation Y~T

```
import numpy as np

Z = np.array([.5, 1, .0, .5, 1., 1.5])
T = np.array([3., 6., .05, 3.01, 5.99, 9.01])
U = np.array([2., 4.01, .1, 1.99, 3.98, 6.02])
Y = np.array([5.01, 9.9, 0.01, 4.95, 10.32, 15.01])
```

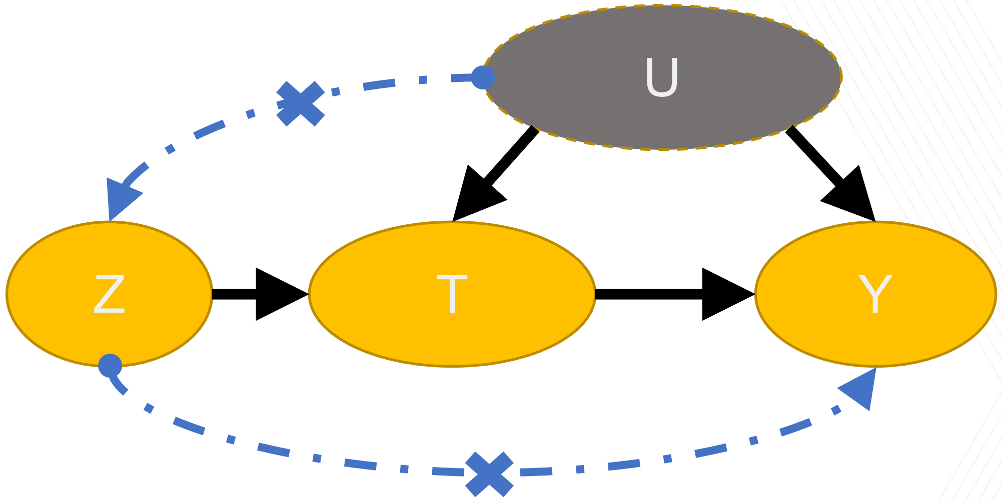
```
[2] # Ordinary Least Squares

beta_OLS = 1./ np.dot(T.T, T) * np.dot(T.T, Y)
print(beta_OLS)
```

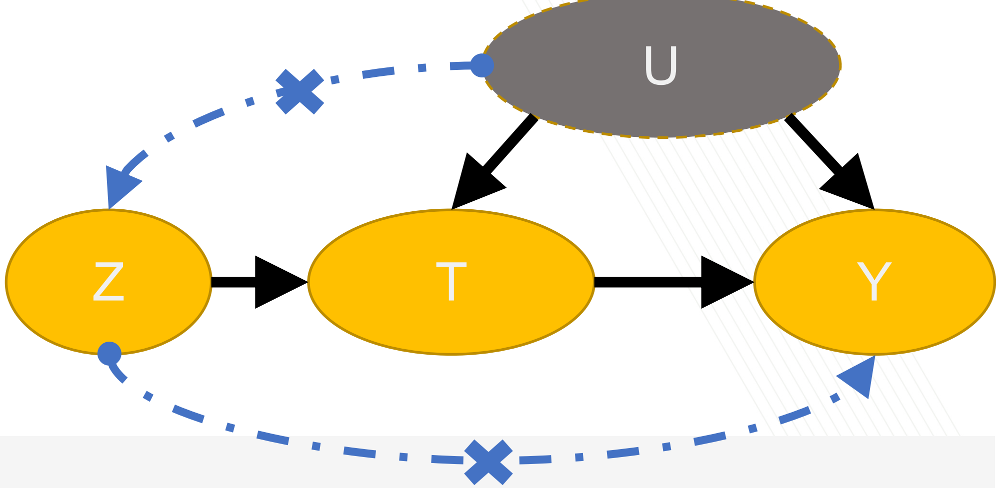
```
1.6735753505669613
```

Two-Stage Least Squares: A linear example

- $Y = \alpha T + \delta U + N_Y$
- $T = \beta Z + \gamma U + N_T$
- Stage1:
  - Regress $T \sim Z$ gives $\hat{\beta}$
- Stage 2:
  - Regress $Y \sim \hat{\beta} Z$
  - Why: $Y = \alpha T + \delta U + N_Y = \alpha(\beta Z) + (\alpha\gamma + \delta)U + N_Y$
  - => we get α



Georgia Tech
CREATING THE NEXT




```
[3] # 2 Stage Least Squares
    delta = 1./ np.dot(Z.T, Z) * np.dot(Z.T, T)
    print(delta)
    That = Z * delta

    beta_2LS = 1./ np.dot(That.T, That) * np.dot(That.T, T)
    print(beta_2LS)
```

```
6.002105263157894
1.0000000000000002
```

T= U + 2Z + Noise

Y= T + U + Noise

Georgia Tech

CREATING THE NEXT

Estimation without unconfoundedness

- Natural Experiment
  - Instrument variables
  - Regression discontinuity design (Skipped due to time constraint)
  - Difference in difference (Skipped due to time constraint)
- Nonparametric bounds and Sensitivity Analysis

Question: how much can we learn τ from data

- $\tau = E[Y_i(1) - Y_i(0)]$
- $\quad = E[Y_i(1)|T_i = 1]\Pr(T_i = 1) + E[Y_i(1)|T_i = 0]\Pr(T_i = 0)$
- $\quad -E[Y_i(0)|T_i = 1]\Pr(T_i = 1) - E[Y_i(0)|T_i = 0]\Pr(T_i = 0)$
- The two quantities in red are unobserved
  - =>need to make assumptions

Georgia Tech
CREATING THE NEXT

Assumptions on the unknown

	T_i	$Y_i(0)$	$Y_i(1)$
$\Pr(T_i = 0)$	0	$E[Y_i(0) \mid T_i = 0]$	?
$\Pr(T_i = 1)$	1	?	$E[Y_i(1) \mid T_i = 1]$

Case 1: Randomized Controlled Trials

	T_i	$Y_i(0)$	$Y_i(1)$
$\Pr(T_i = 0)$	0	$E[Y_i(0)\|T_i = 0]$	$E[Y_i(1)\|T_i = 1]$
$\Pr(T_i = 1)$	1	$E[Y_i(0)\|T_i = 0]$	$E[Y_i(1)\|T_i = 1]$

- In RCTs, since the data are missing at random
- Treatment and control groups are identical in expectation
- PO of treatment and control groups identical in expectation

Case 2: Lower Bound

	T_i	$Y_i(0)$	$Y_i(1)$
$\Pr(T_i = 0)$	0	$E[Y_i(0) \mid T_i = 0]$	$\underline{Y}$
$\Pr(T_i = 1)$	1	$\bar{Y}$	$E[Y_i(1) \mid T_i = 1]$

- Assume the worst possible outcome
- Treated units would have best possible outcome $\bar{Y}$ if untreated
- Control units would have had worst possible outcome $\underline{Y}$ if treated

This results in a lower bound:

$$E[Y_i(1) \mid T_i = 1] \Pr(T_i = 1) + \underline{Y} \Pr(T_i = 0) - \bar{Y} \Pr(T_i = 1) - E[Y_i(0) \mid T_i = 0] \Pr(T_i = 0)$$

Case 3: Upper Bound

	T_i	$Y_i(0)$	$Y_i(1)$
$\Pr(T_i = 0)$	0	$E[Y_i(0)\|T_i = 0]$	$\bar{Y}$
$\Pr(T_i = 1)$	1	$\underline{Y}$	$E[Y_i(1)\|T_i = 1]$

- Assume the best possible outcome
- Treated units would have worst possible outcome $\underline{Y}$ if untreated
- Control units would have had best possible outcome $\bar{Y}$ if treated

This results in a lower bound:

$$E[Y_i(1)|T_i = 1]\Pr(T_i = 1) + \bar{Y}\Pr(T_i = 0) - \underline{Y}\Pr(T_i = 1) - E[Y_i(0)|T_i = 0]\Pr(T_i = 0)$$

Case 4: Adding Assumptions

- Example: Monotone treatment selection
  - units who select the treatment have higher expectation of outcome under either condition on average (e.g. sicker)
  - $E[Y_i(0)|T_i = 0] \leq E[Y_i(0)|T_i = 1]$
  - $E[Y_i(1)|T_i = 0] \geq E[Y_i(1)|T_i = 1]$
  - We can obtain a tighter upper bound

Agenda

Break

Part 1

- Introduction of Causal Inference

3:00 – 3:30

Part 2

- Basic of Causal Effect Estimation Algorithm

3:30 – 4:15

Part 3

- Recent Advances
- Challenges and Opportunities

4:35 – 5:15

I'll attend the NSF student award from 4:20 – 4:30 CNB-MAC