Introduction to Causal Inference

Jennifer Hill
Applied Statistics, Social Sciences, and Humanities
PRIISM Center
New York University

Introductions

Jennifer Hill

Professor of Applied Statistics

Co-Chair: Department of Applied Statistics, Social Science, and Humanities

Co-Director: PRIISM Center

New York University

jennifer.hill@nyu.edu

Areas of Focus:
Bayesian machine learning
Causal inference
Missing Data
Multilevel models

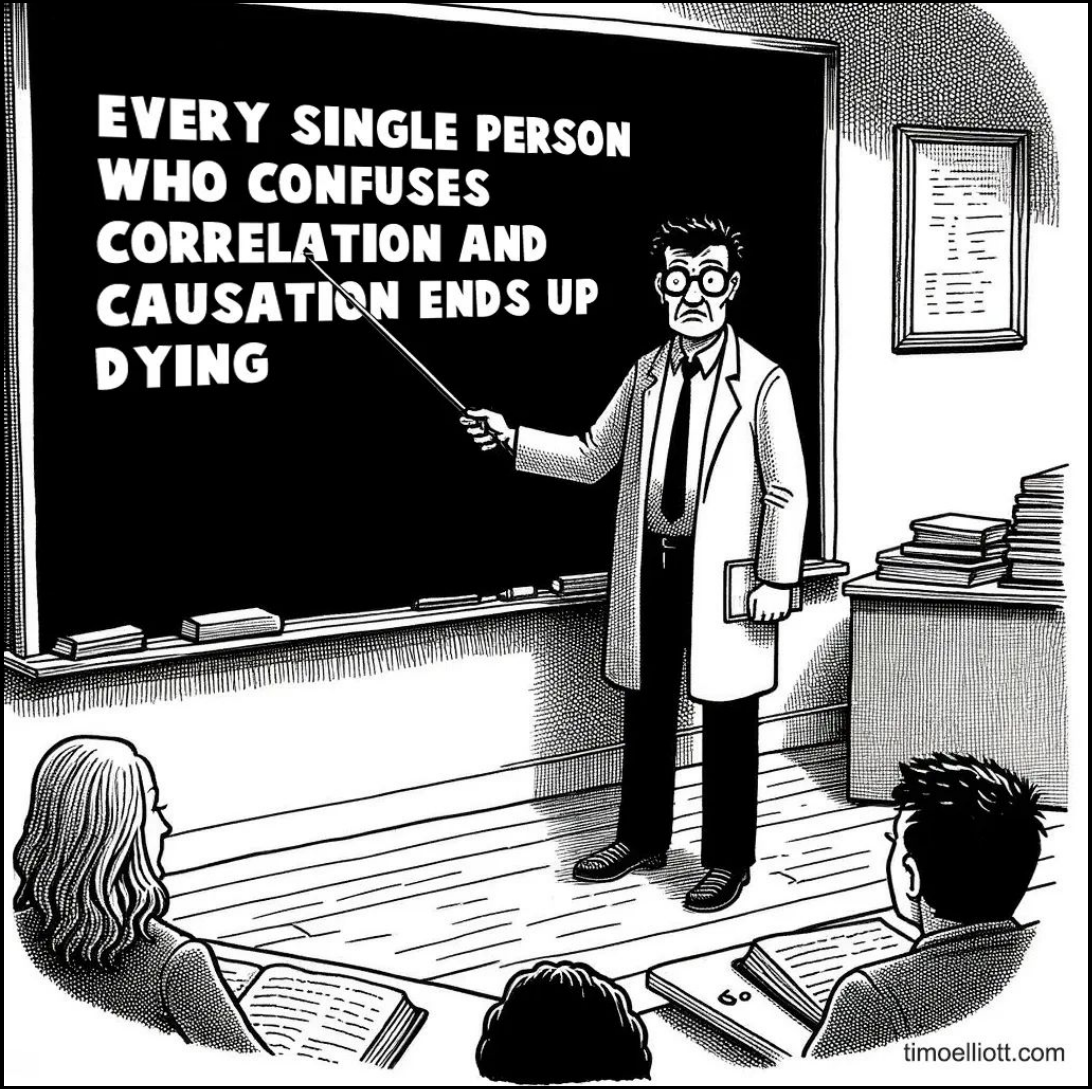
Agenda

- Foundations of Causal Inference
- Quasi-experimental designs and methods

Foundations of Causal Inference

Foundations: agenda

- Motivation
- Counterfactuals and causal estimands
- Randomized Trials


EVERY SINGLE PERSON WHO CONFUSES CORRELATION AND CAUSATION ENDS UP DYING
timoelliott.com

Motivation

Motivation and Concepts

- Cautionary tales
- Counterfactuals
- Causal Estimands

Why do we care about causal inference?

Social Policy questions are CAUSAL questions!

Social Policy questions…..

Does Abstinence-only Education Work?

Social Policy questions…..

Does Abstinence-only Education Work?

Can Gay Marriage Solve Our Adoption Problem?

Social Policy questions.....

Does Abstinence-only Education Work?

Can Gay Marriage Solve Our Adoption Problem?

Does exposing preschoolers to music make them smarter?

Social Policy questions.....

Does Abstinence-only Education Work?

Can Gay Marriage Solve Our Adoption Problem?

Does exposing preschoolers to music make them smarter?

Did the introduction of CitiBike make New Yorkers healthier?

Social Policy questions.....

Does Abstinence-only Education Work?

Can Gay Marriage Solve Our Adoption Problem?

Does exposing preschoolers to music make them smarter?

Did the introduction of CitiBike make New Yorkers healthier?

Does the death penalty reduce crime?

Social Policy questions.....

Does Abstinence-only Education Work?

Can Gay Marriage Solve Our Adoption Problem?

Does exposing preschoolers to music make them smarter?

Did the introduction of CitiBike make New Yorkers healthier?

Does the death penalty reduce crime?

Would a 'Medicare for All' plan help you save money on your family's health-care costs?

Social Policy questions.....

Does Abstinence-only Education Work?

Can Gay Marriage Solve Our Adoption Problem?

Does exposing preschoolers to music make them smarter?

Did the introduction of CitiBike make New Yorkers healthier?

Does the death penalty reduce crime?

Would a 'Medicare for All' plan help you save money on your family's health-care costs?

What Happens When the Poor Receive a Stipend?

How likely are we to get the wrong answers to these questions?

How likely are we to get the wrong answers to these questions?

What is the cost if we do?

Causal Inference is Important!

Failing to carefully think through causal issues can cost time, money, lives......

Cautionary Tales

- Salk Vaccine
- Internet ads

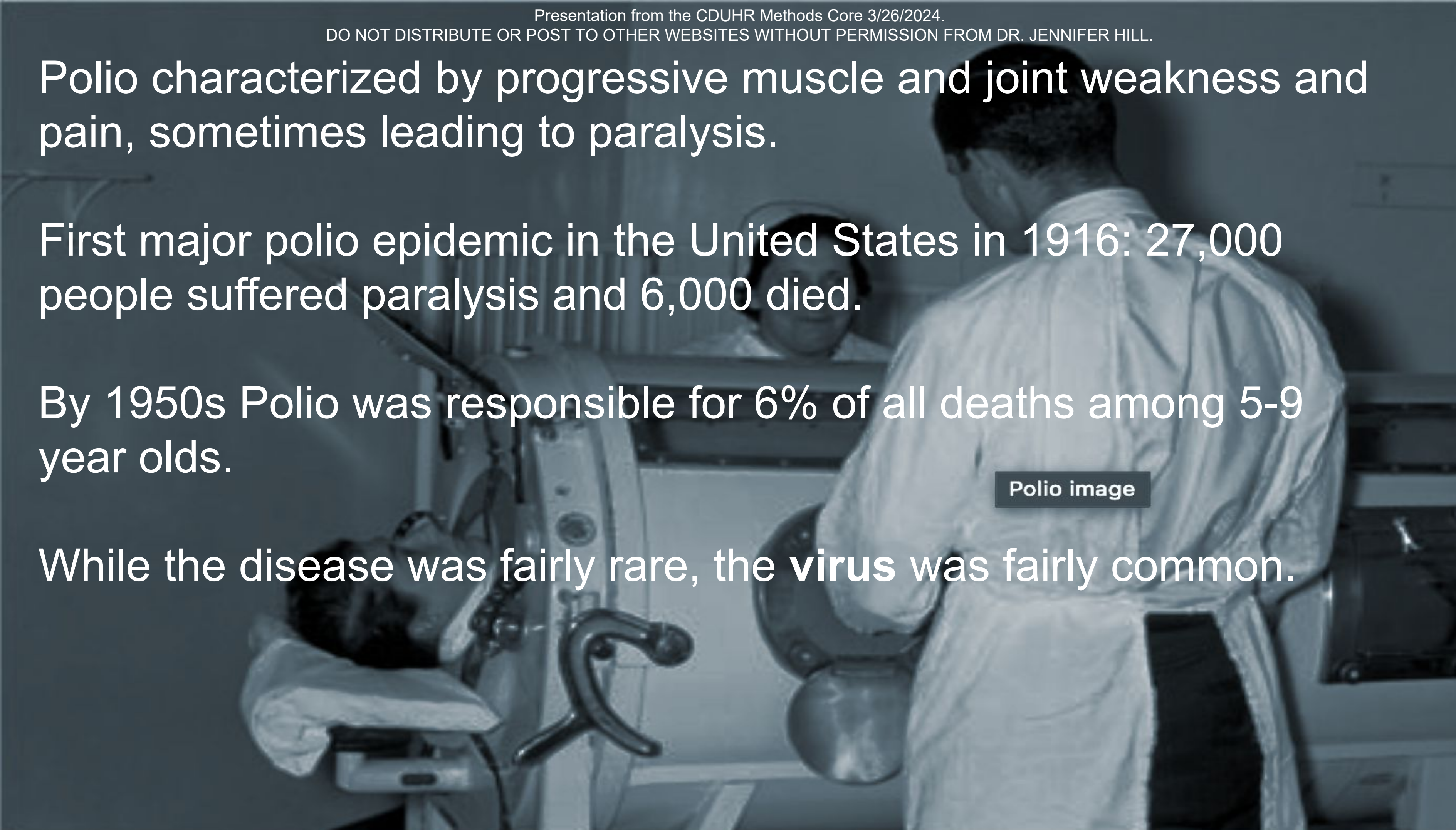
Polio and the Salk Vaccine

Polio characterized by progressive muscle and joint weakness and pain, sometimes leading to paralysis.

First major polio epidemic in the United States in 1916: 27,000 people suffered paralysis and 6,000 died.

By 1950s Polio was responsible for 6% of all deaths among 5-9 year olds.

While the disease was fairly rare, the **virus** was fairly common.



Patient in iron lung, Rhode Island polio epidemic, 1960


POLIO cripples family savings
because big Polio expenses usually continue for months and months
POLIO INSURANCE
THE ANN ARBOR NEWS
$10.00 A YEAR


The Daily Tar Heel
POLIO FORCES CANCELATION OF TWO GAMES
Douglas To Be Speaker Here
Eisenhower, McGrath Also May Give Talks
BULLETIN
NEWS IN BRIEF
One Grid Player, 3 Others Stricken
Coeds Note Rush Change
Status Of Cobb Is Left Vague By Legislature
Yugoslavian Cosmopolitan Prexy Elect

"By the mid-20th century, the poliovirus could be found all over the world and killed or paralysed over half a million people every year. With no cure, and epidemics on the rise, there was an urgent need for a vaccine."

www.who.int/news-room/spotlight/history-of-vaccination/history-of-polio-vaccination

Could the Salk vaccine eradicate the disease?

- 1954: US Public Health Service wants to investigate the effectiveness of a vaccine invented by Salk

Could the Salk vaccine eradicate the disease?

- 1954: US Public Health Service wants to investigate the effectiveness of a vaccine invented by Salk
- Disappointingly, observational evidence comparing those vaccinated with those not vaccinated did not demonstrate convincing success!

Could the Salk vaccine eradicate the disease?

- 1954: US Public Health Service wants to investigate the effectiveness of a vaccine invented by Salk
- Disappointingly, observational evidence comparing those vaccinated with those not vaccinated did not demonstrate convincing success!
- A randomized experiment was then conducted that suggested the vaccine was effective!

Why was the observational evidence misleading?

Which type of kid had more access to the vaccine?

Which type of kid had more access to the vaccine?

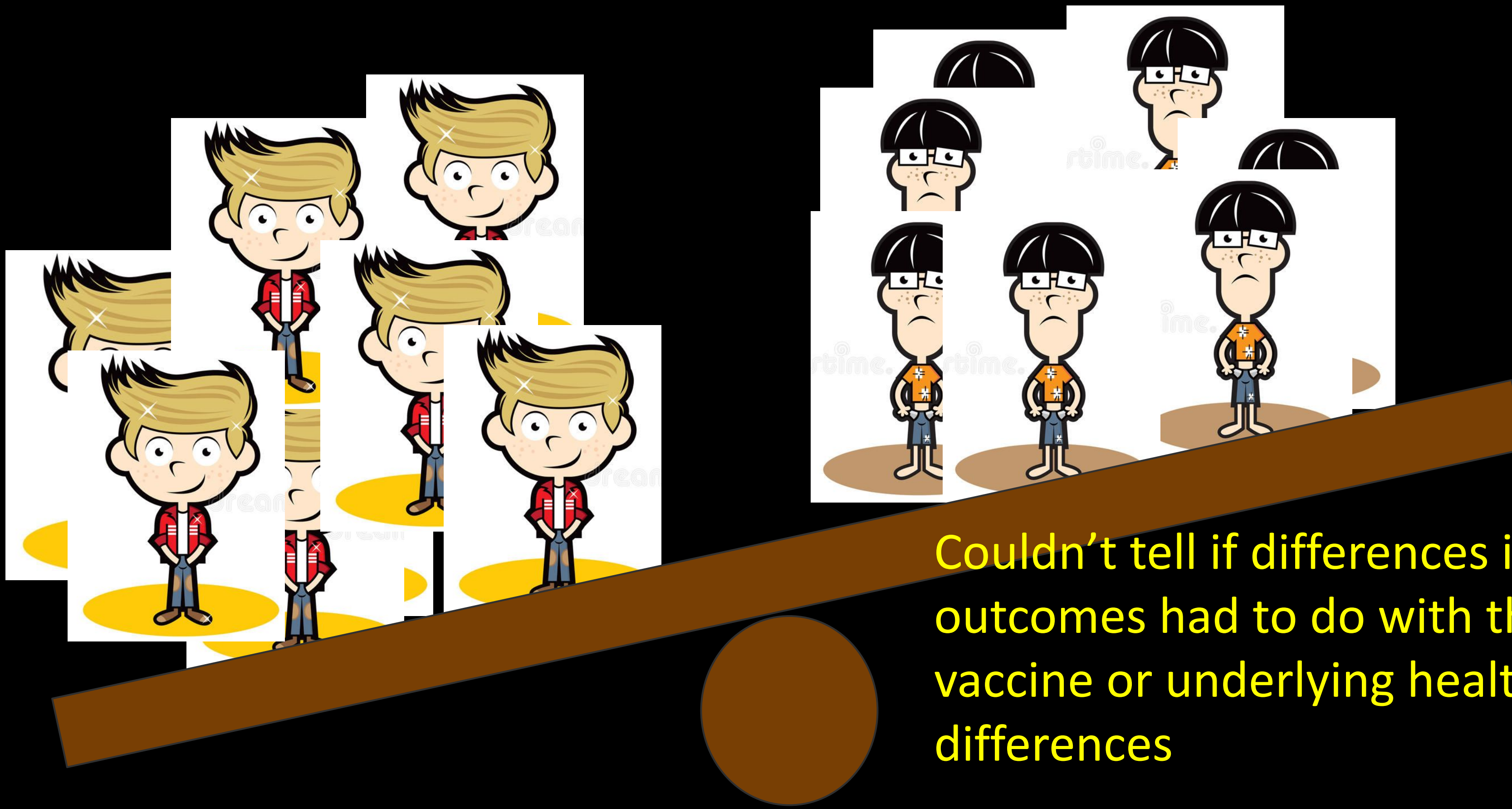
Which type of kid had more resistance to the virus?

Which type of kid had more resistance to the virus?

In the absence of the vaccine who would have been more likely to survive?

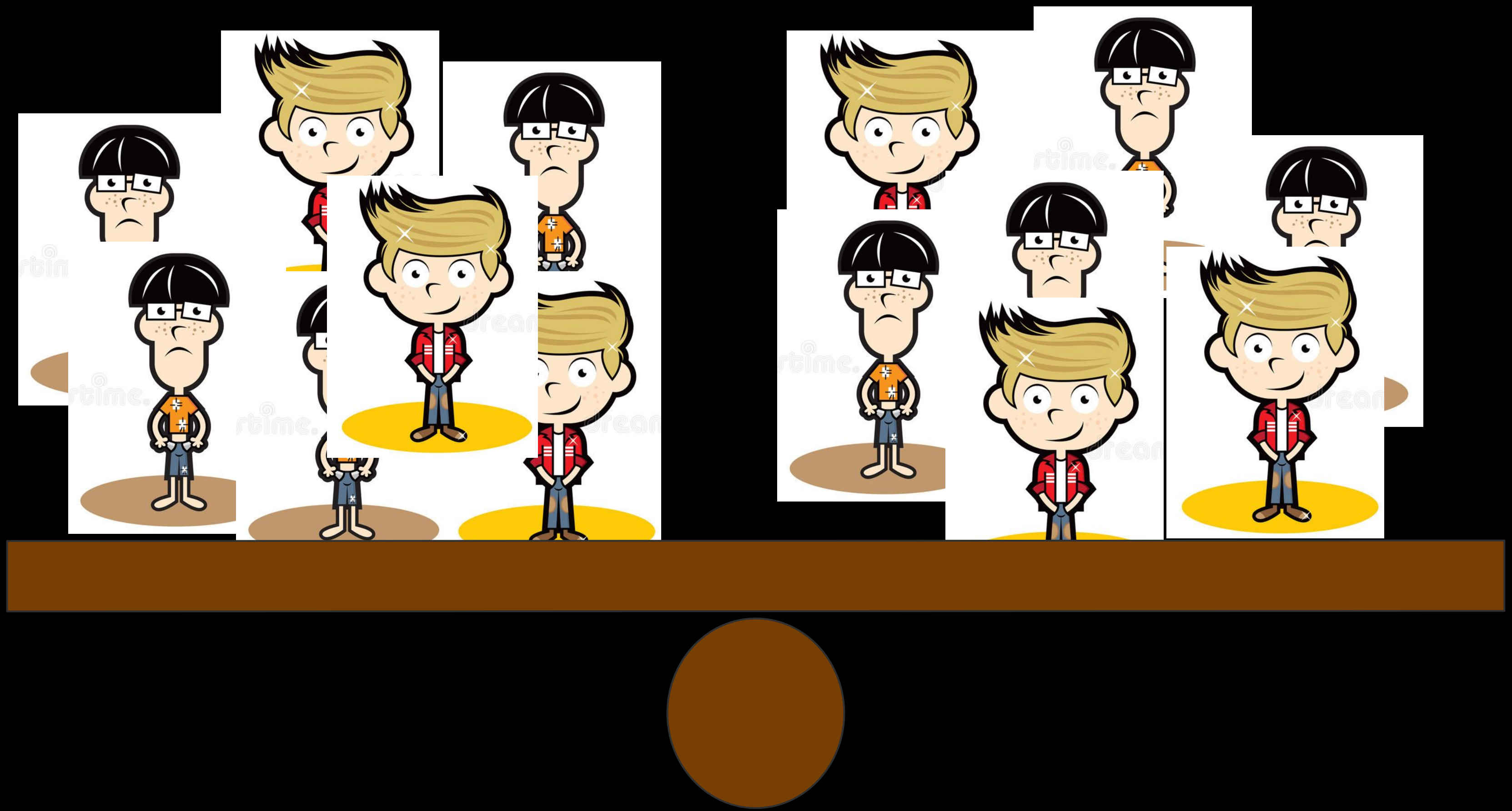
In the absence of the vaccine who would have been more likely to survive?

The observational comparison wasn't fair!

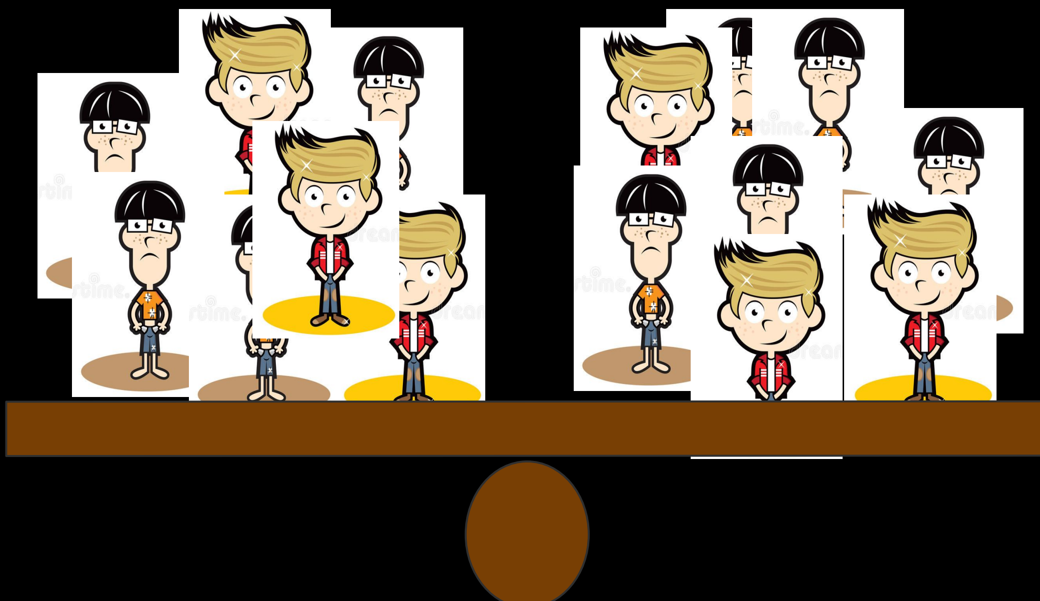


Couldn't tell if differences in outcomes had to do with the vaccine or underlying health differences

The randomized comparison **was** fair!



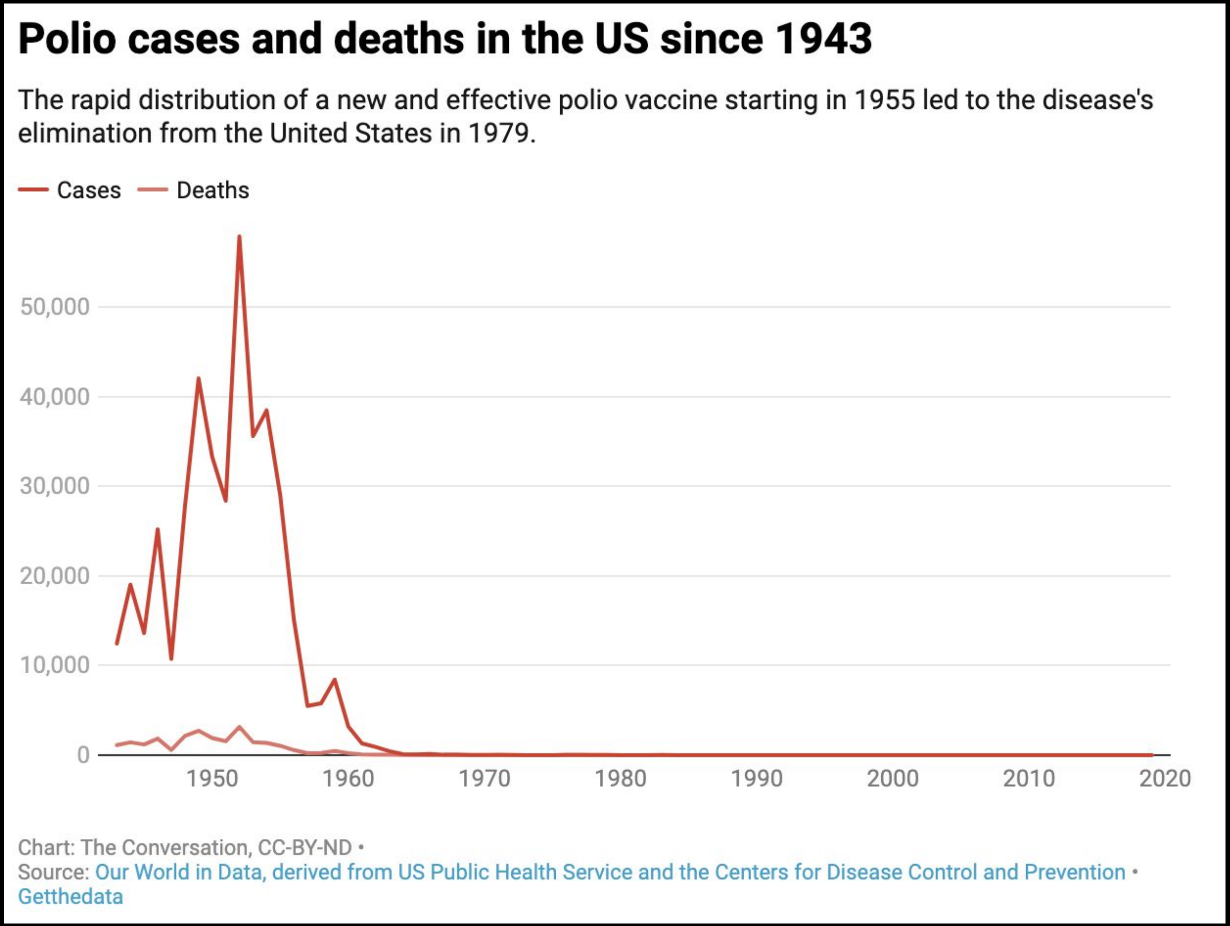
The randomized comparison **was** fair!



Groups were balanced (similar) both on observed and unobserved characteristics.

Differences in outcomes could be attributed to the vaccine.

Lives saved because of evidence from a randomized experiment!



NEW YORK, WEDNESDAY, APRIL 13, 1955. FIVE CENTS

SALK POLIO VACCINE PROVES SUCCESS; MILLIONS WILL BE IMMUNIZED SOON; CITY SCHOOLS BEGIN SHOTS APRIL 25

GH COURT HEARS OUTH WILL DEFY UICK END TO BIAS

dual Approaches Urged Integration of Schools— egro Lawyers Opposed

By LUTHER A. HUSTON

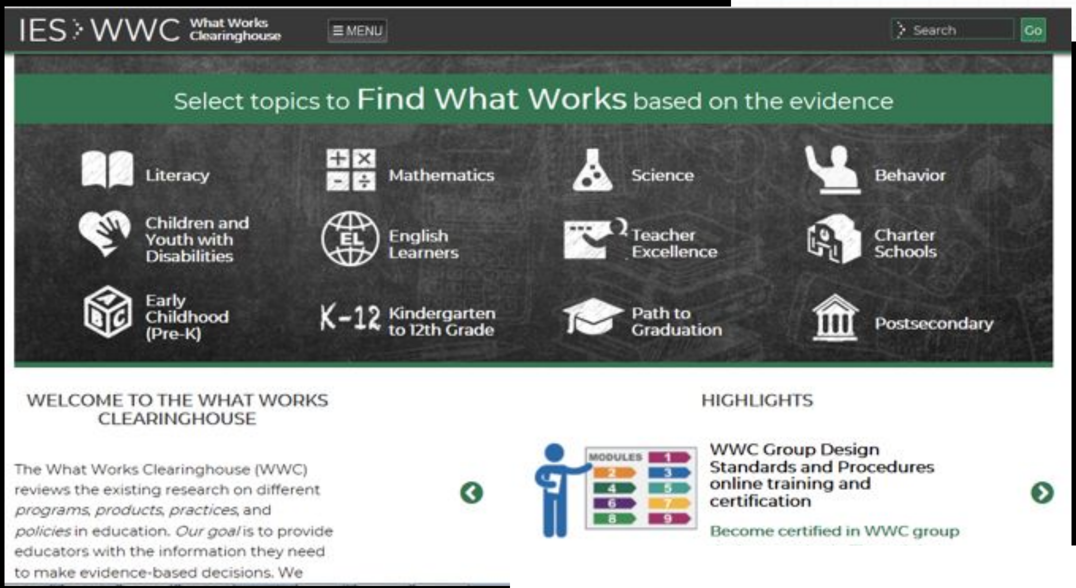
TRIAL DATA GIVEN

Efficacy of 80 to 90% Shown—Salk Sees Further Advance

By WILLIAM L. LAURENCE

ANN ARBOR, Mich., April 12 —The world learned today that its hopes for finding an effective weapon against paralytic polio had been realized.

FDA U.S. FOOD & DRUG ADMINISTRATION



Cochrane Trusted evidence. Informed decisions. Better health.

After that we had learned our lesson about the importance of thinking carefully about causality, right….?

60 years later...

We have BIG DATA

We have BIG DATA

We have fancy machine learning methods to analyze it

Do big data and machine learning make it ***easier*** or ***harder*** to understand causal relationships?

WIRED MAGAZINE: 16.07

Science : Discoveries

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

The Petabyte Age:

There is now a better way. Petabytes allow us to say: "Correlation is enough."

Big Data = Big Hubris

WIRED MAGAZINE: 16.07

Science : Discoveries

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chr[illegible] 23.08

Illustration: Marian Bantjes

The Petabyte Age:

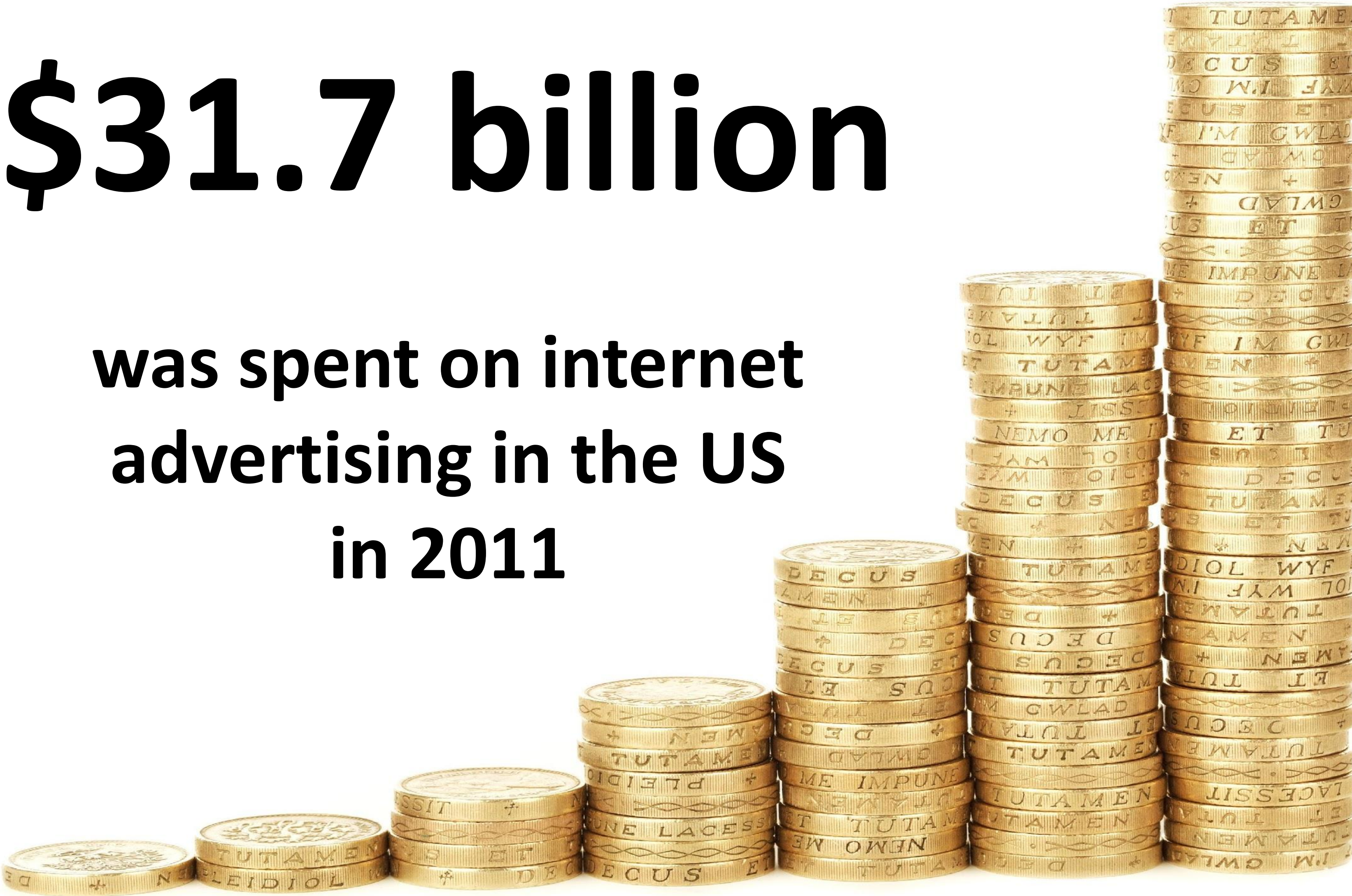
There is now a better way. Petabytes allow us to say: "Correlation is enough."

sigh…

Internet Ads

$31.7 billion

was spent on internet advertising in the US in 2011



Do click throughs → $$$?

Do click throughs → $$$?

- Common wisdom:
  - internet advertising is highly effective

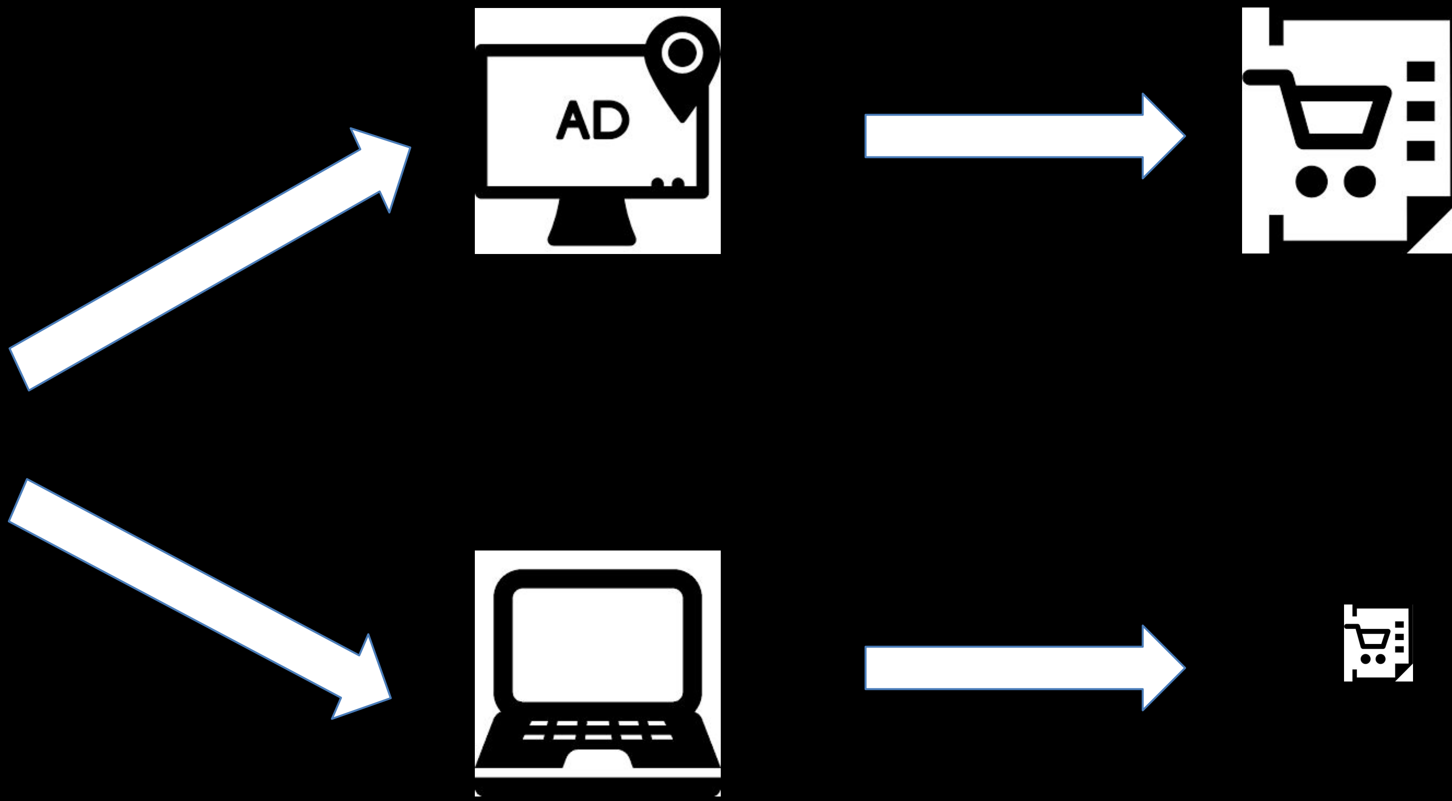
Do click throughs → $$$?

- Common wisdom:
  - internet advertising is highly effective
- Data:
  - did you click on ad?
  - did you buy the product?

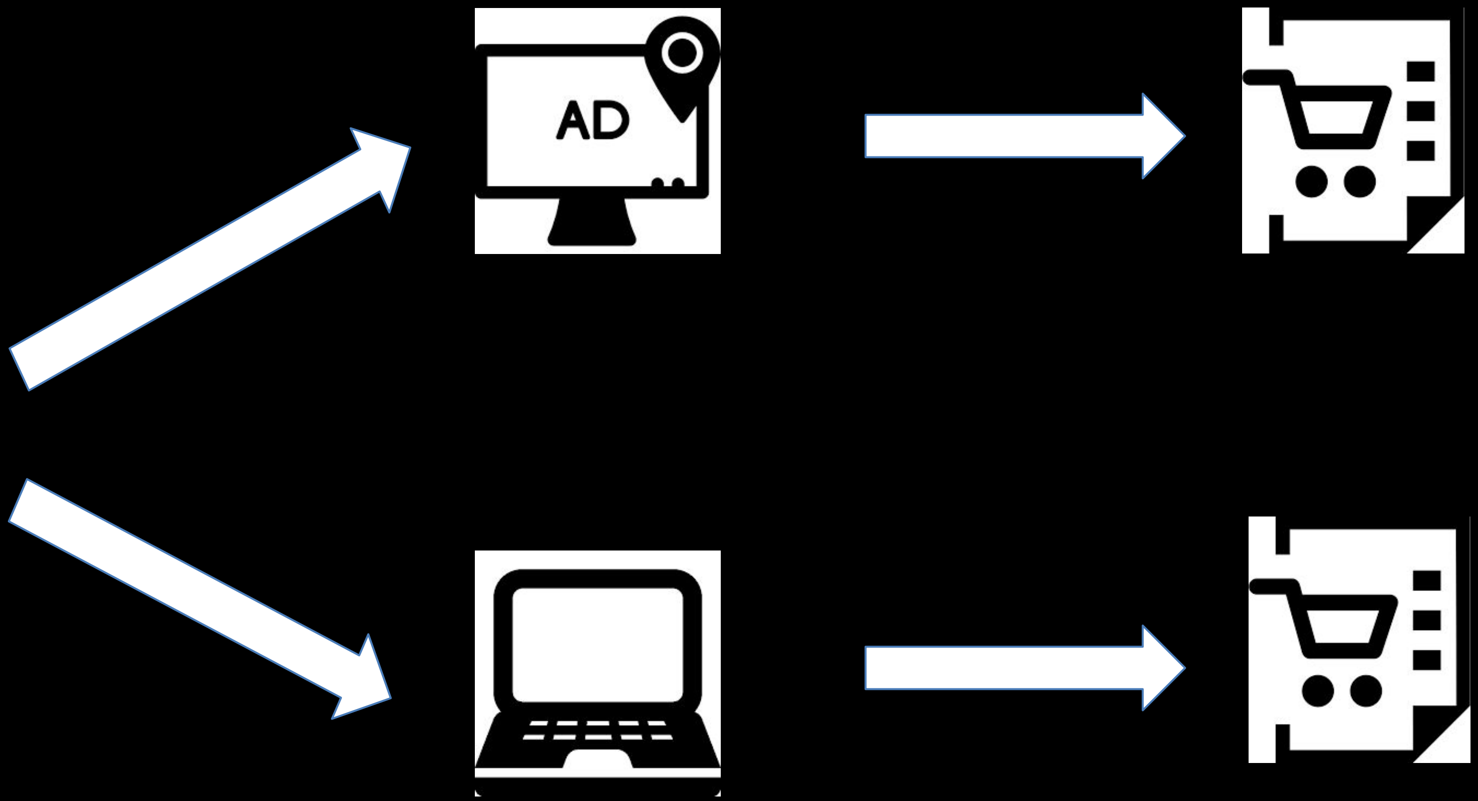
Do click throughs → $$$?

- Common wisdom:
  - internet advertising is highly effective
- Data:
  - did you click on ad?
  - did you buy the product?
- Methods:
  - machine learning algorithms that predict purchases from clicks (i.e. big data + machine learning)

Marketers wants you to believe…



But what if the truth is….?



Causal question:

What if shoppers would have bought the product *anyway*?

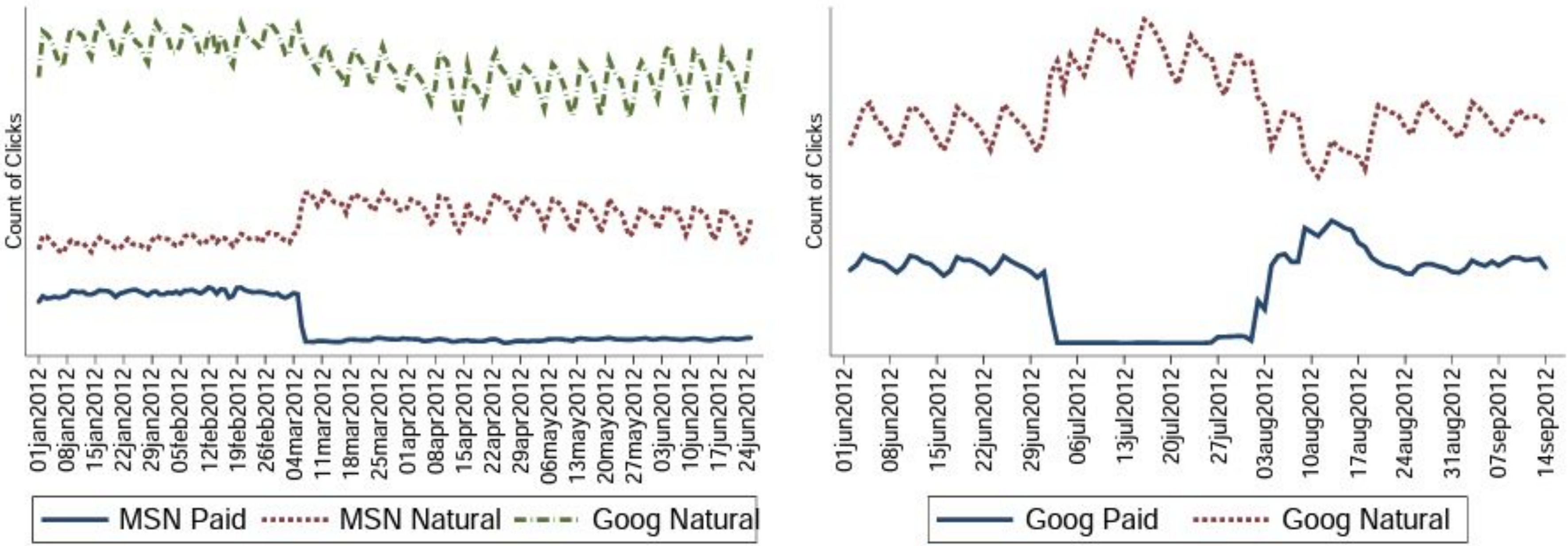
Ebay performed a quasi experimental study

Compared

- click through traffic with ads on and off on one search engine
- click through traffic with no ads on other engines

Blake, T., Nosko, C., and S. Tadelis (2013) "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment"

99.5% of purchases happened without an ad!

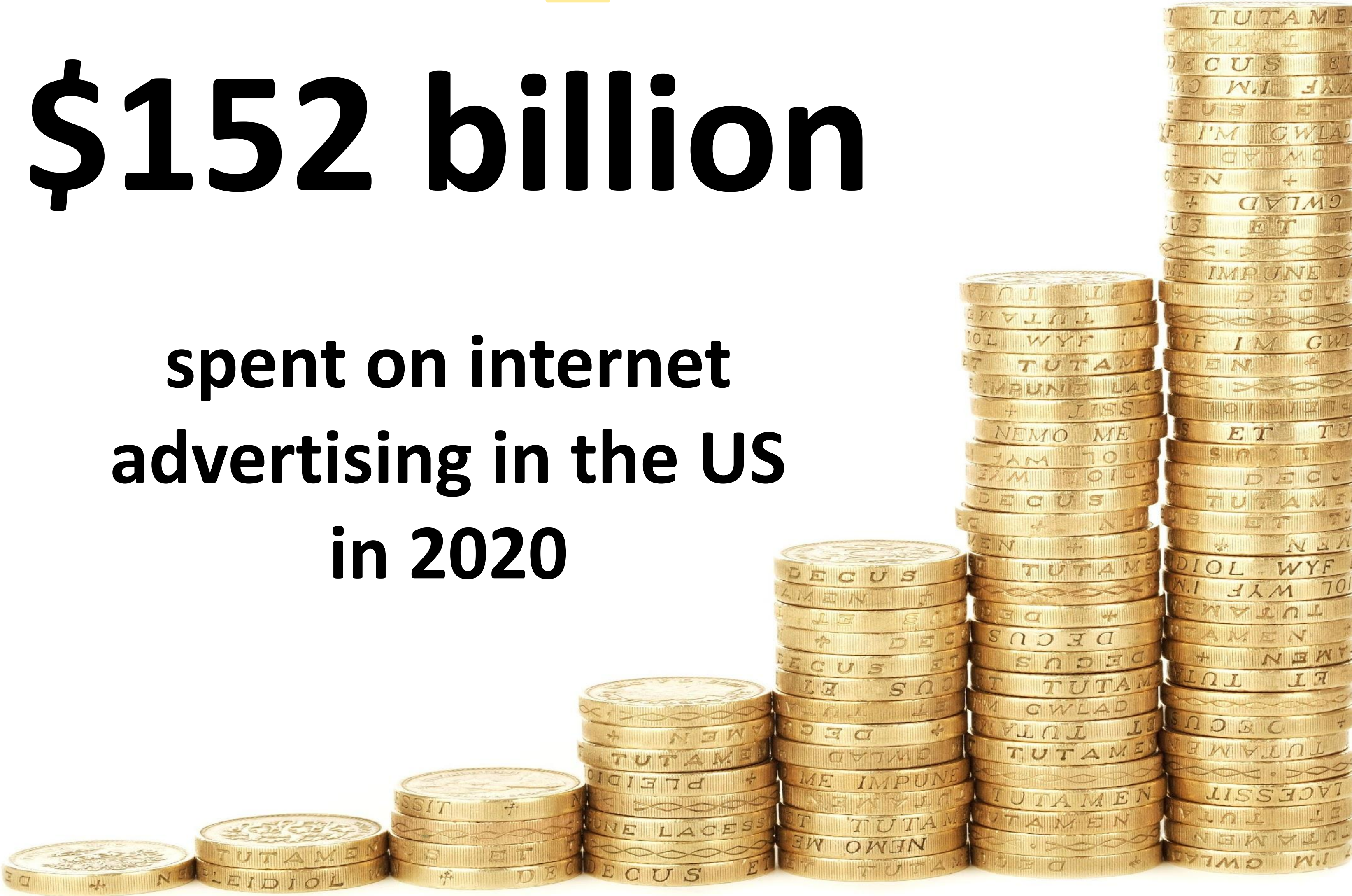


(a) **MSN Test** (b) **Google Test**

Note: MSN and Google click traffic is shown for two events where paid search was suspended (Left) and suspended and resumed (Right).

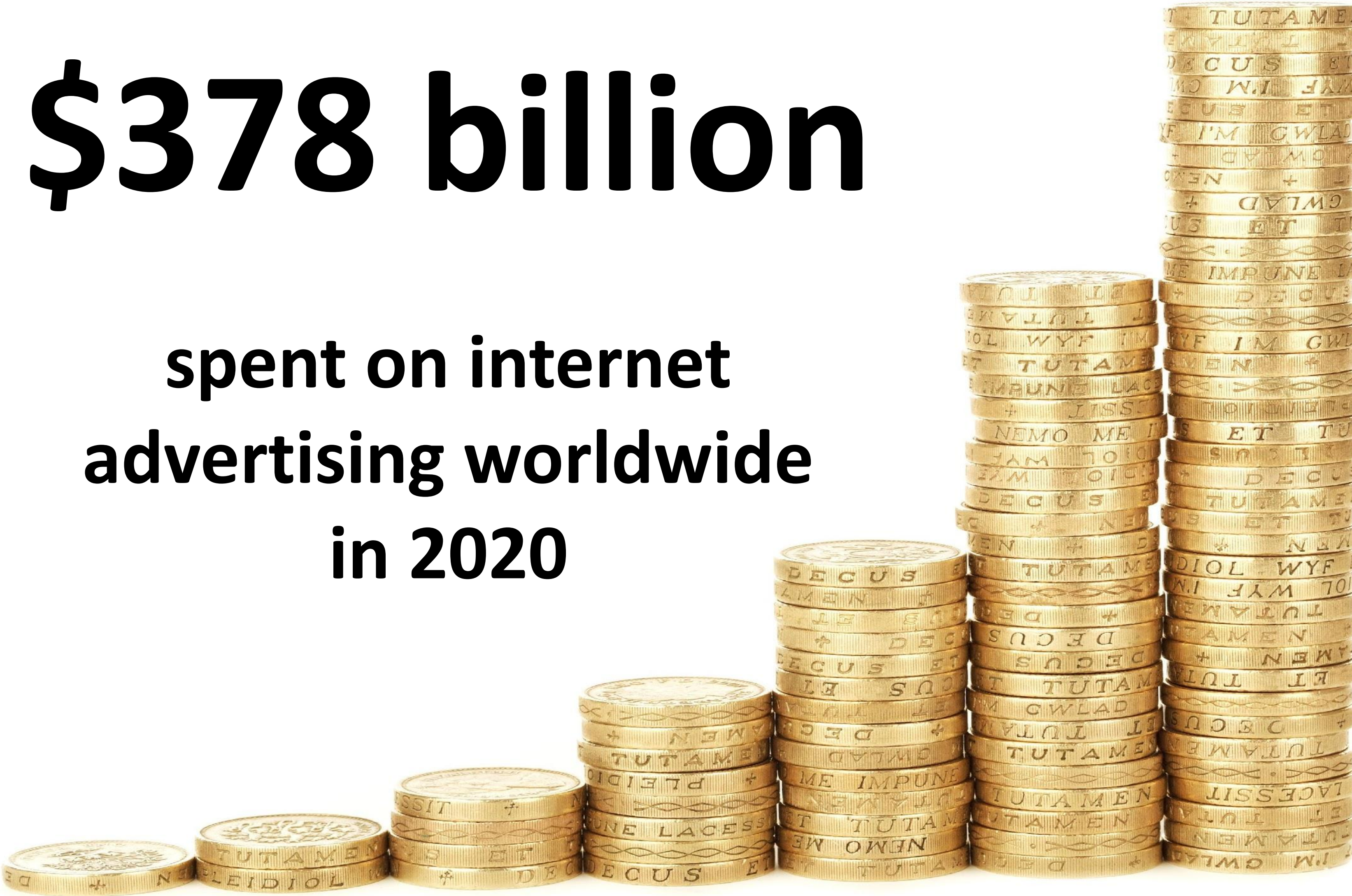
$152 billion

spent on internet advertising in the US in 2020



$378 billion

spent on internet advertising worldwide in 2020



We ignore causal inference at our peril!

Failing to carefully think through causal issues can cost time, money, lives......

so…….

What's going on: Selection Bias!

Selection bias

- when different types of observations are selected or self-selected into different treatments

and

- these differences across observations are also predictive of outcomes.

Is there a solution?

Is there a solution?

Maybe…….?????

Is there a solution?

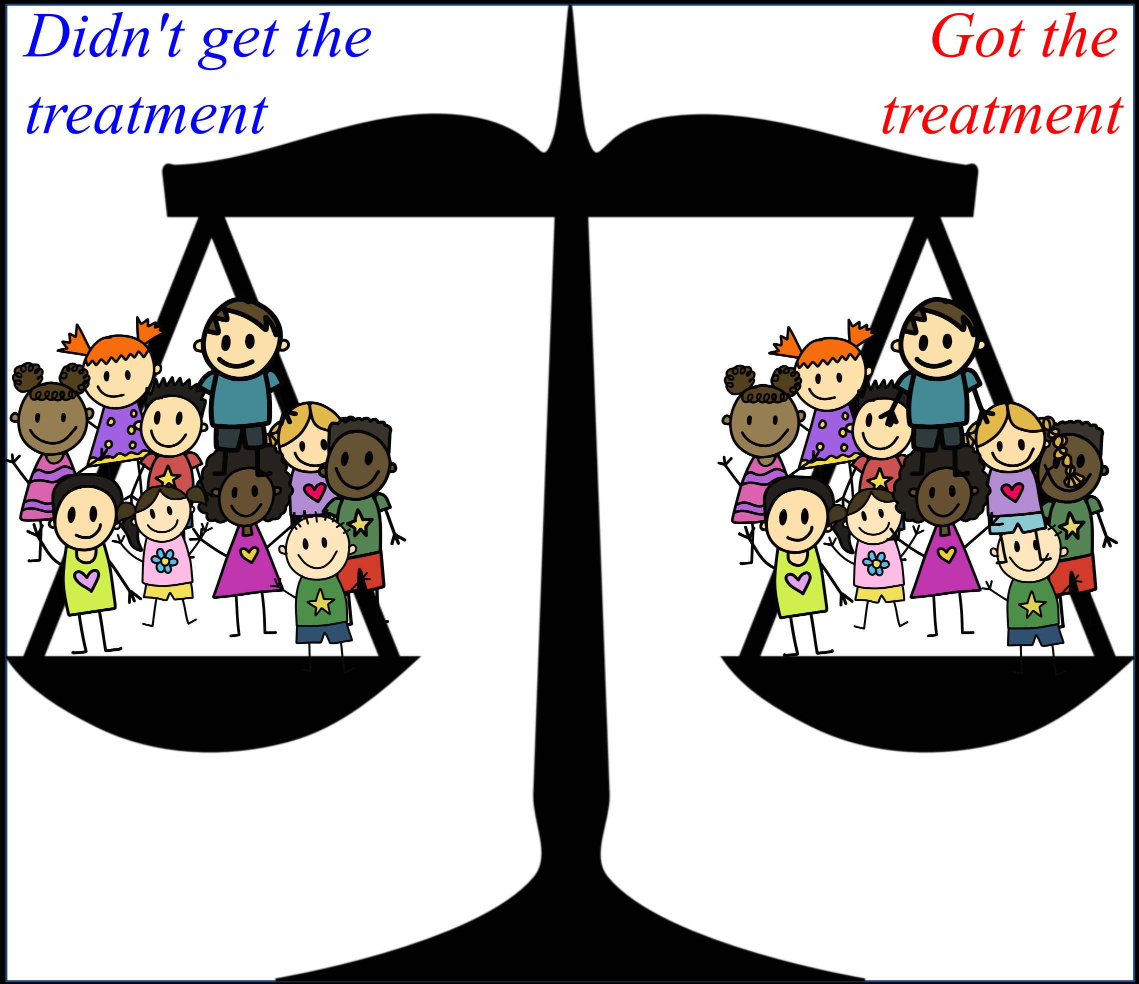
Maybe…….?????

- Design
- Modeling
- Transparency about assumptions

First, let's formalize the **problem**……

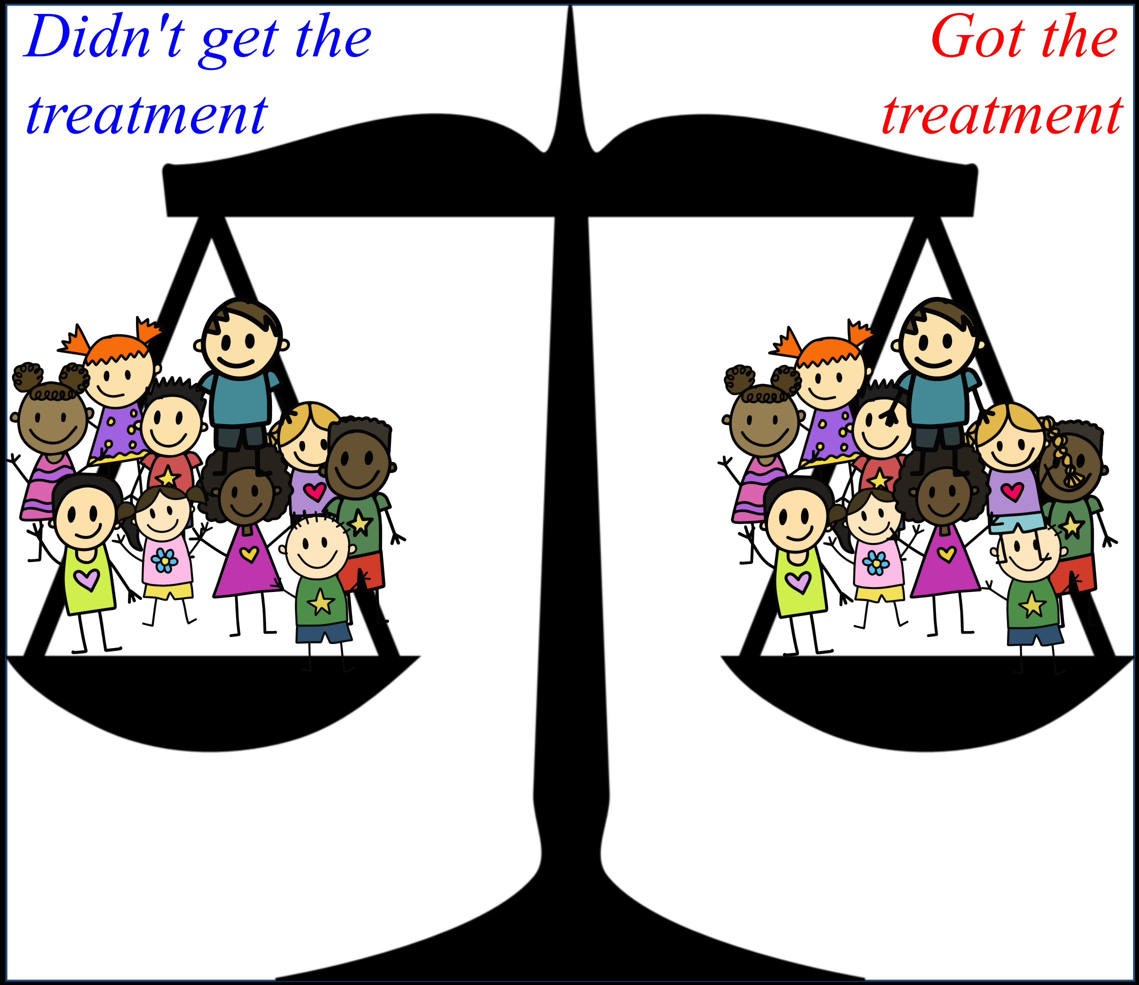
Causal Inference is hard

Causal inference is about making fair comparisons



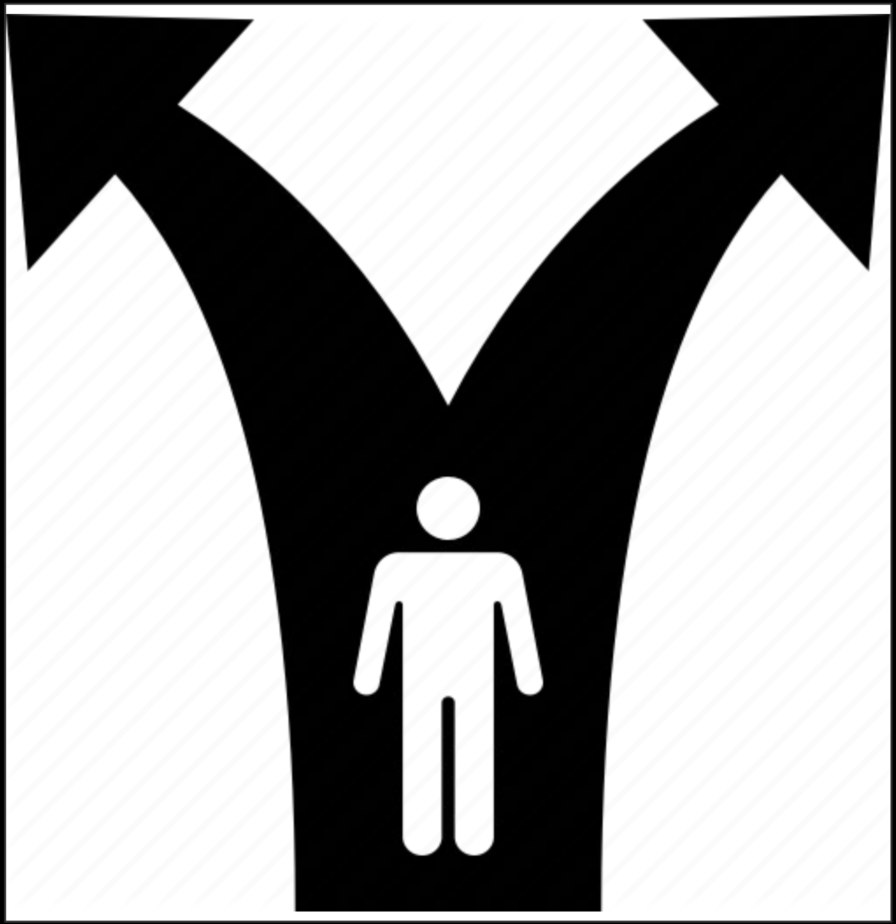
But often we aren't given a level playing field

Or more insidiously, the two groups LOOK the same



But in fact they are different in ways we haven't measured

Let's make this idea of fair comparisons more concrete!



Counterfactuals and Causal Estimands

How do we define a causal effect?

To understand causal inference....

we need to understand....

Counterfactuals

Why do we need counterfactuals?

Why do we need counterfactuals?

Consider the following....

- Jo is struggling in math

Why do we need counterfactuals?

Consider the following....

- Jo is struggling in math
- Jo uses an online tool for extra help with the material

Why do we need counterfactuals?

Consider the following....

- Jo is struggling in math
- Jo uses an online tool for extra help with the material
- Jo scores poorly on the subsequent math test

Why do we need counterfactuals?

Consider the following....

- Jo is struggling in math
- Jo uses an online tool for extra help with the material
- Jo scores poorly on the subsequent math test

Did the online tool cause the low test score?

Why do we need counterfactuals?

Consider the following....

- Jo is struggling in math
- Jo uses an online tool for extra help with the material
- Jo scores poorly on the subsequent math test

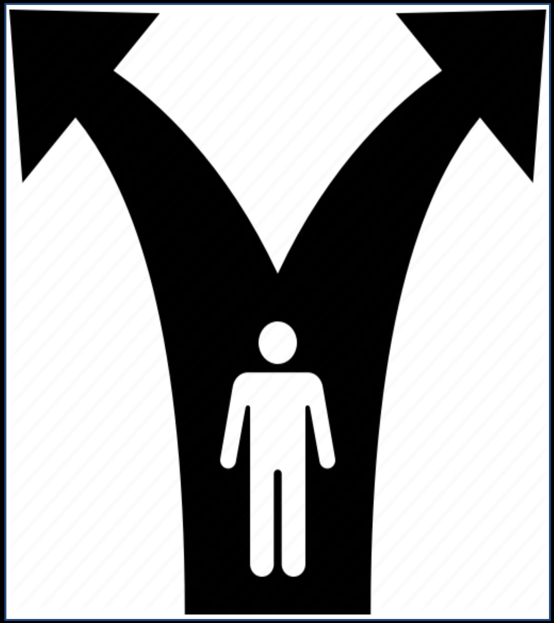
Did the online tool cause the low test score?

Q: What would have happened if Jo had not used the tool?

Jo after classroom instruction alone Y(0)

Jo after classroom instruction + tool Y(1)

Causal inference requires a comparison of counterfactual states



Effect of the online tool for Jo: Y(1)- Y(0)

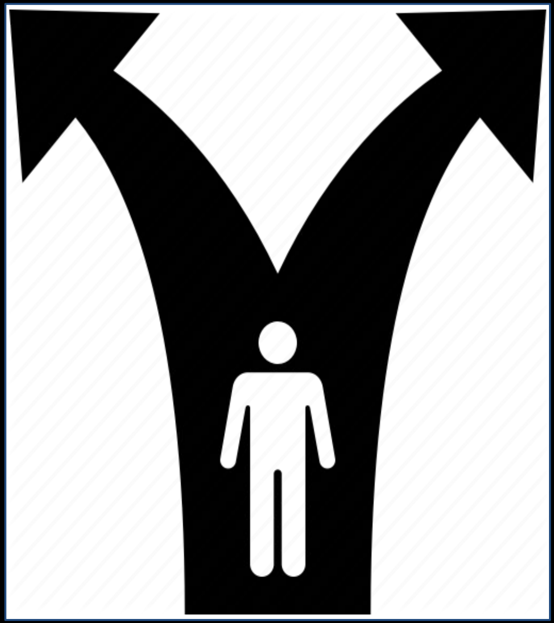
But we can't see BOTH potential outcomes at the same time!

We have a missing data problem !!!!!!

?

Jo after classroom instruction + tool Y(1)

Causal inference requires a comparison of counterfactual states



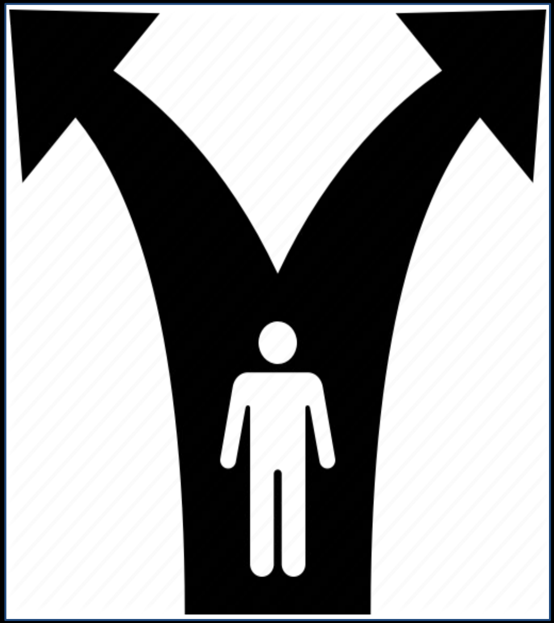
Effect of the online tool for Jo: Y(1)- Y(0)

We have a missing data problem !!!!!!

Jo after classroom instruction alone Y(0)

?

Causal inference requires a comparison of counterfactual states



Effect of the online tool for Jo: $Y(1) - Y(0)$

The Estimand

The quantity we are trying to estimate

The **estimand**: What we are trying to estimate?

The **estimand** is *the quantity we are trying to estimate.*

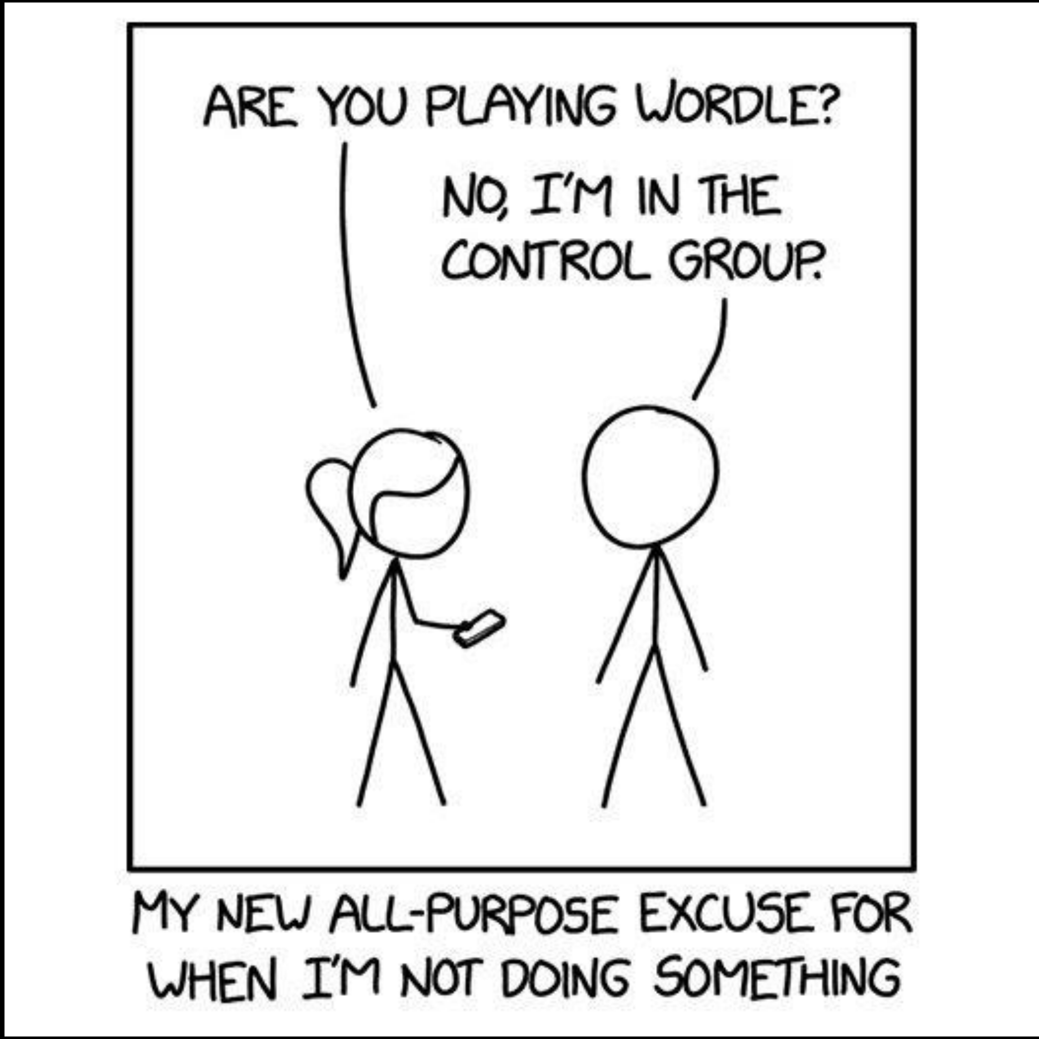
We often focus on estimating *average* causal effects.

We have defined an individual level treatment effect as the difference between two potential outcomes

$$Y_i(1) - Y_i(0)$$

The average treatment effect (ATE) can be defined as

$$\text{Avg}[Y(1)-Y(0)]$$



Randomized Experiments

Randomized experiments

- Definition
- Intuition
- Assumptions
- Estimation
- IHDP example
- Compliance and ethics

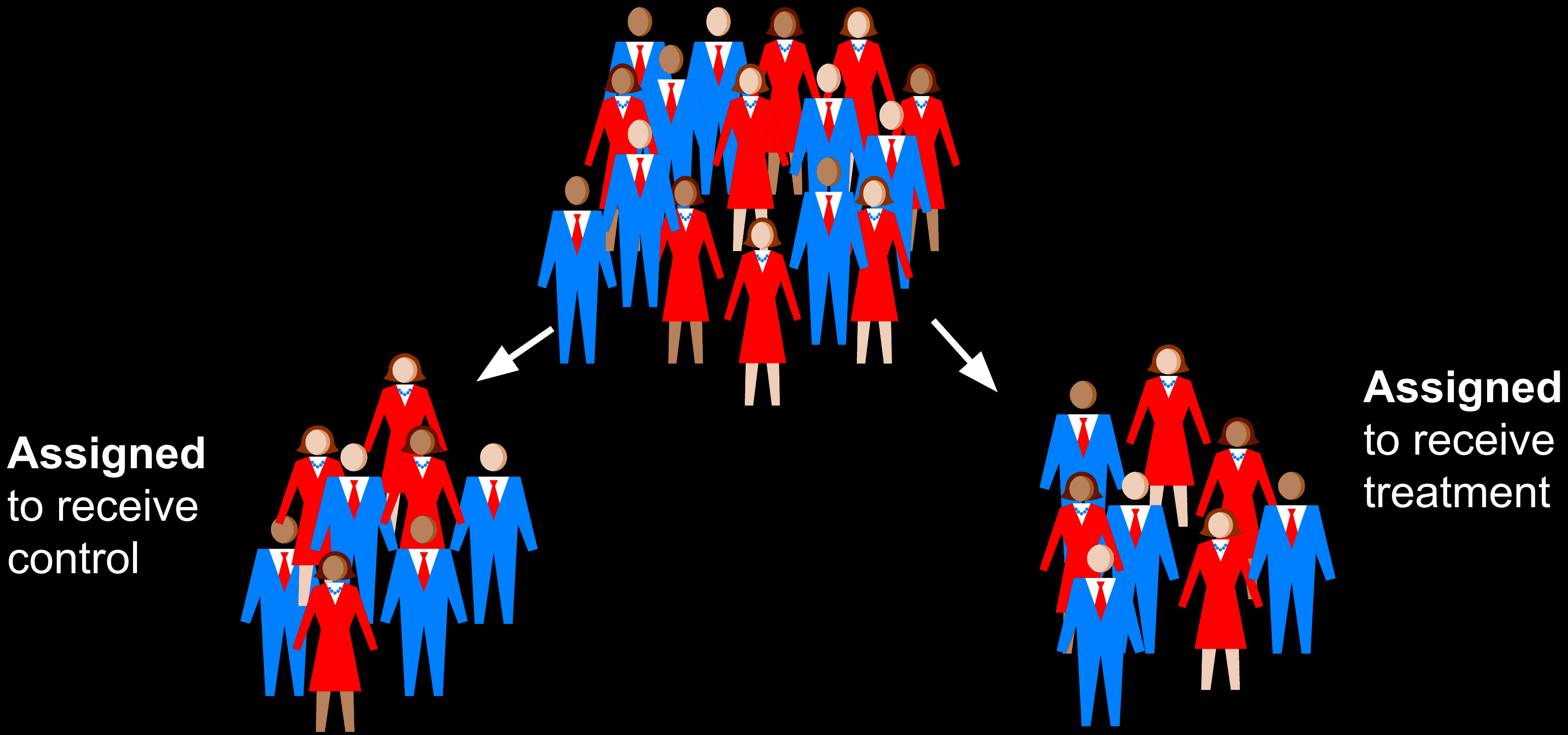
Randomized experiments: the gold standard

- Randomized experiments: "gold standard" for answering causal questions

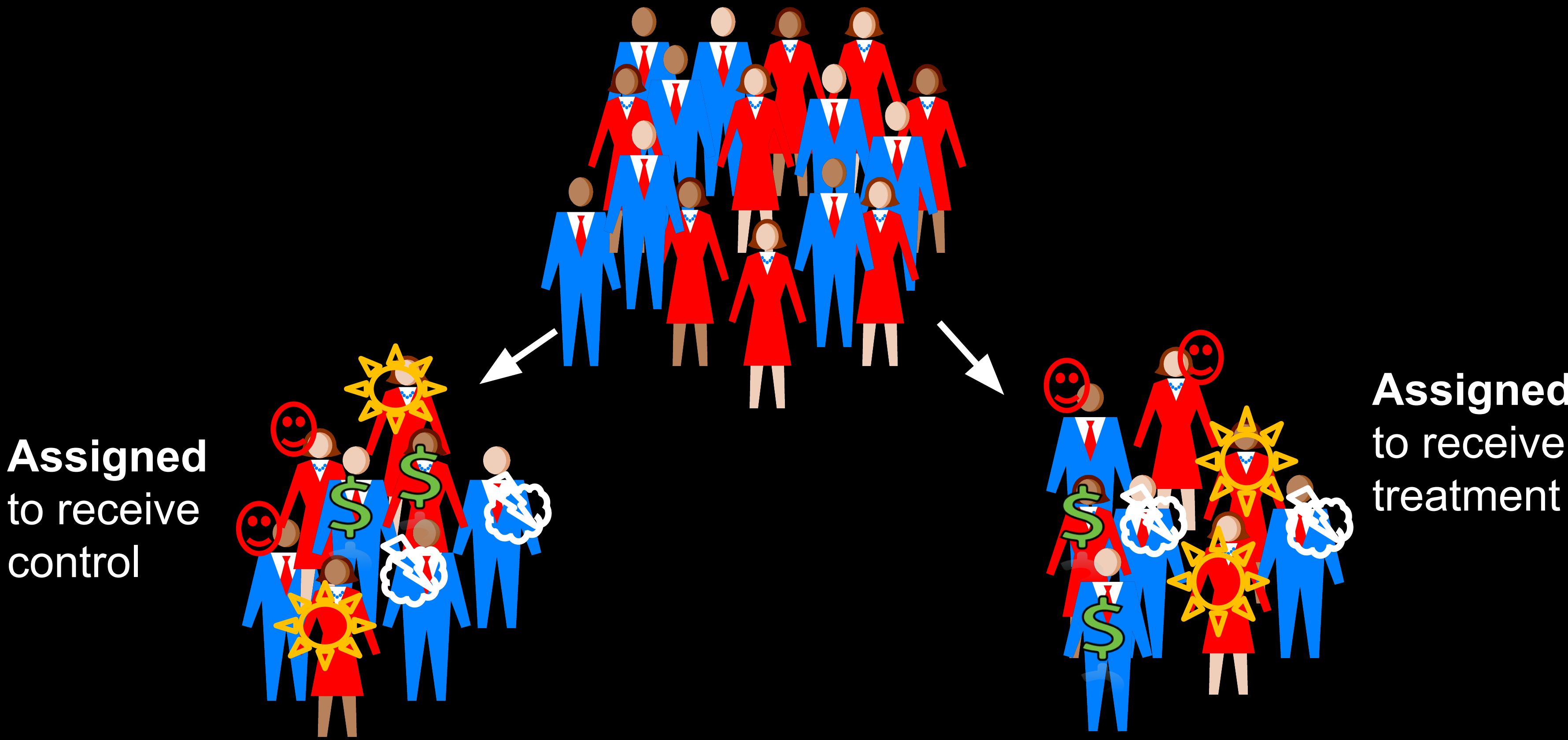
- They create two (or more) groups that are virtually identical to each other on average

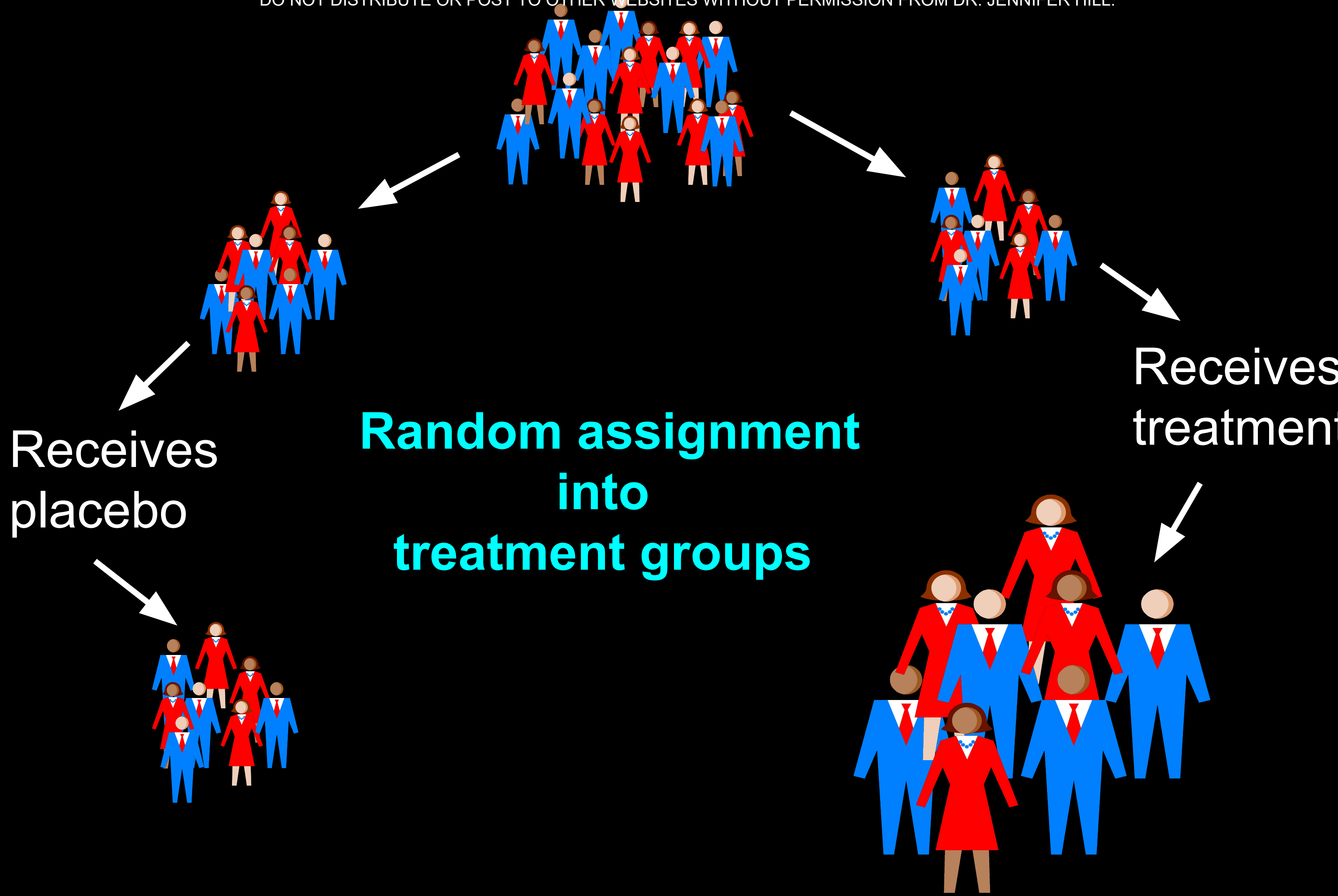
- If each group receives a different treatment, we can safely attribute any difference in outcomes to the different treatments

Randomized experiments: creating balance



Randomized experiments: balance on observed AND unobserved features of the observations




Random assignment into treatment groups
Receives placebo
Receives treatment

Randomized experiments: defining characteristics

- Each unit assigned to treatment using a known probabilistic rule
- Each unit has nonzero probability of being allocated to each treatment
- Let's focus on two types of randomized experiments
  - completely randomized experiment
  - randomized block experiment

Completely randomized experiment: properties (assumptions that are satisfied by design)

- Since treatments are allocated by a known probabilistic mechanism we know that

  $$\Pr(Z \mid Y(0), Y(1)) = \Pr(Z)$$

- Equivalently: $Z \perp Y(0), Y(1)$
- This is referred to under many names including:
  - no hidden bias
  - ignorability

Example: Infant Health & Development Program

- Observations on ~1000 children; random assignment:
  - ⅓ were randomly assigned to participate in IHDP (Z=1)
  - ⅔ assigned to receive no intervention (Z=0)

- Covariates (X) were recorded. For example,
  - Age
  - Mom's education level (high school graduate or not)

- IQ score of each child (Y) a year after program ends

$\text{Avg(age)}_{Z=1} = 23.83$

(Hypothetical) observed data from IHDP

$\text{Avg(age)}_{Z=0} = 25.17$

Person	Treat	Educ.	Age	Y(0)	Y(1)	Y
1	1	1	26	?	114	114
2	1	1	21	?	112	112
3	1	1	30	?	116	116
4	1	1	19	?	112	112
5	1	0	25	?	110	110
6	1	0	22	?	108	108
7	0	1	26	110	?	110
8	0	1	21	108	?	108
9	0	1	42	116	?	116
10	0	1	15	102	?	102
11	0	0	26	106	?	106
12	0	0	21	104	?	114

Information are we missing if we want to **calculate** ATE

Completely randomized experiment: implications

Consider the average treatment effect,
$\text{Avg}[Y(1)\text{-}Y(0)] = \text{Avg}[Y(1)] - \text{Avg}[Y(0)]$?

How do we estimate $\text{Avg}[Y(1)]$ even though we are missing half of the values?

How do we estimate $\text{Avg}[Y(0)]$ even though we are missing half of the values?

Recall our IHDP example

Person	Treat	Y(0)	Y(1)	Y
1	1	110	114	114
2	1	108	112	112
3	1	112	116	116
4	1	108	112	112
5	1	106	110	110
6	1	104	108	108
7	0	110	114	110
8	0	108	112	108
9	0	116	120	116
10	0	102	106	102
11	0	106	110	106
12	0	104	108	104

Goal is to estimate

ATE = Avg[Y(1) - Y(0)]

= Avg[Y(1)] - Avg[Y(0)]

Recall our IHDP example

Person	Treat	Y(0)	Y(1)	Y
1	1	110	114	114
2	1	108	112	112
3	1	112	116	116
4	1	108	112	112
5	1	106	110	110
6	1	104	108	108
7	0	110	114	110
8	0	108	112	108
9	0	116	120	116
10	0	102	106	102
11	0	106	110	106
12	0	104	108	104

Goal is to estimate

ATE = Avg[Y(1) - Y(0)]

= Avg[Y(1)] - Avg[Y(0)]

If we want to estimate **Avg[Y(1)]**

We can get an unbiased estimate by just using the treated sample!

The randomized experiment ensured that they are a **random sample** of the full sample.

Recall our IHDP example

Person	Treat	Y(0)	Y(1)	Y
1	1	110	114	114
2	1	108	112	112
3	1	112	116	116
4	1	108	112	112
5	1	106	110	110
6	1	104	108	108
7	0	110	114	110
8	0	108	112	108
9	0	116	120	116
10	0	102	106	102
11	0	106	110	106
12	0	104	108	104

Goal is to estimate

ATE = Avg[Y(1) - Y(0)]

= Avg[Y(1)] - Avg[Y(0)]

If we want to estimate **Avg[Y(0)]**

We can get an unbiased estimate by just using the treated sample!

The randomized experiment ensured that they are a **random sample** of the full sample.

Completely randomized experiment: implications

Consider the average treatment effect,

Avg[Y(1)-Y(0)] = Avg[Y(1)] - Avg[Y(0)]?

We can estimate Avg[Y(1)] using the mean of the Y's in the treatment group. *Because those units are a random sample from the full sample.*

We can estimate Avg[Y(0)] using the mean of the Y's in the control group. *Because those units are a random sample from the full sample.*

Completely randomized experiment: estimation

Consider the average treatment effect,

Avg[Y(1)-Y(0)] = Avg[Y(1)] - Avg[Y(0)]?

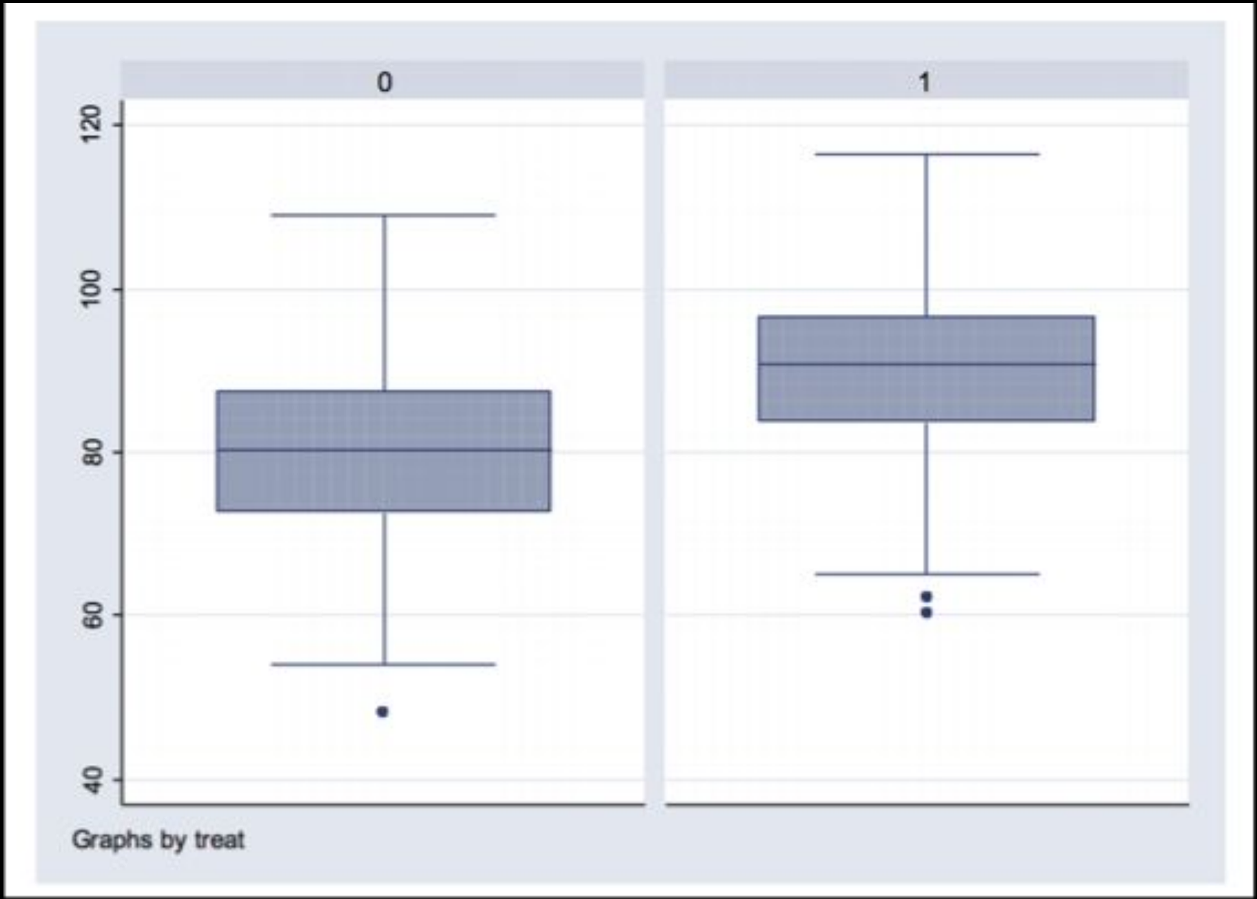
We can estimate **Avg[Y(1)]** using the mean of the Y's in the treatment group, $\bar{Y}_1$. *Because those units are a random sample from the full sample.*

We can estimate **Avg[Y(0)]** using the mean of the Y's in the control group, $\bar{Y}_0$. *Because those units are a random sample from the full sample.*

Estimating treatment effects, options

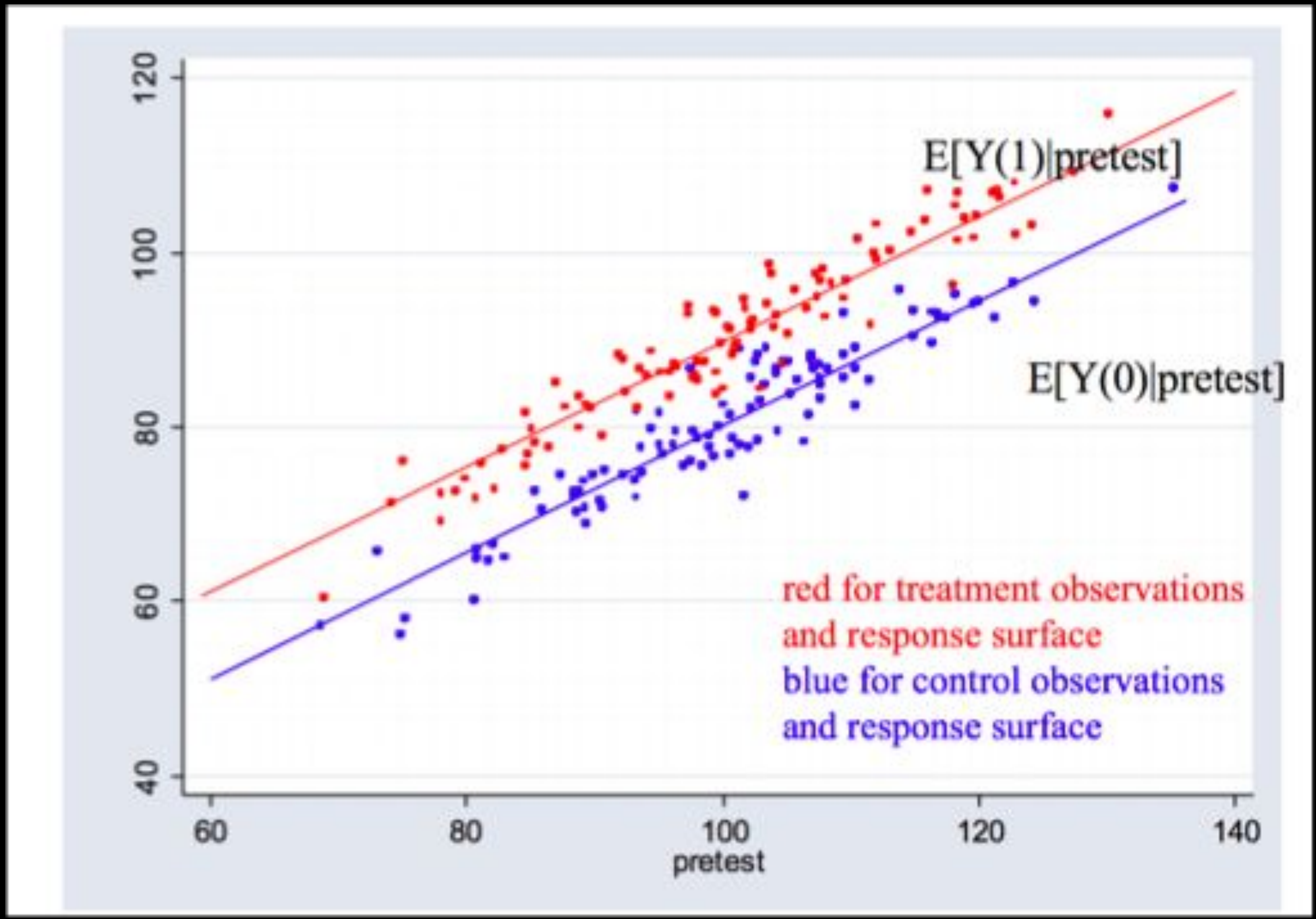
- Difference in means: $\bar{Y}_1 - \bar{Y}_0$
- Regression with:
  - an indicator for treatment (but nothing else)
  - an indicator for treatment + pre-treatment variables
  - ~~Post-treatment variables~~

Randomized experiment



Randomized experiment

regression modeling for more precision



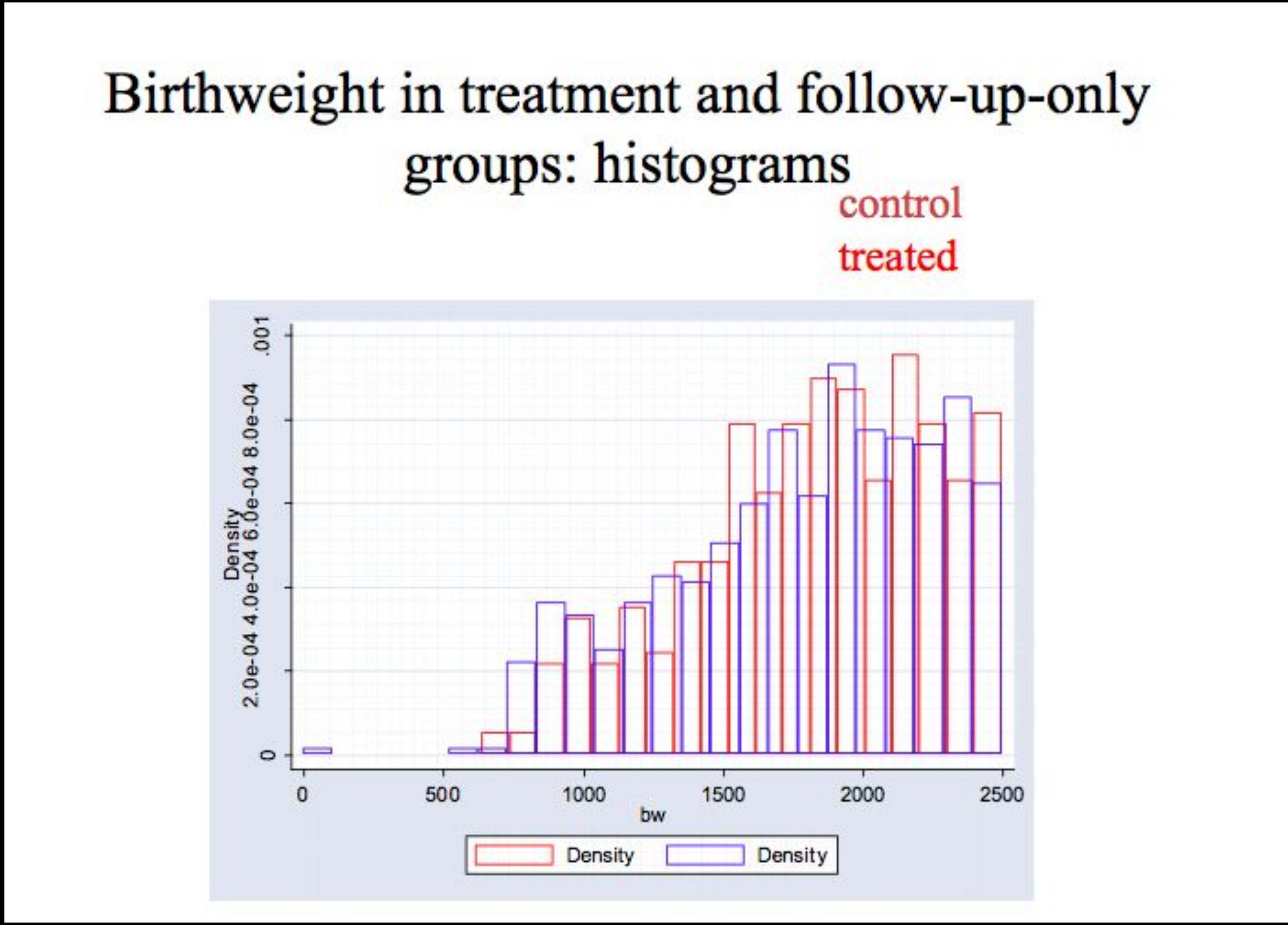
Results: IHDP

What would we expect the **distribution** of any given outcome variable to look like for the treatment group relative to the control group?

Did randomization work?

Variables	FU	IHDP	Variables	FU	IHDP
Mother			*Child*		
Age	25.0	24.7	Birth weight	1787	1816
Black	0.52	0.55	Head circ (birth)	29.5	29.5
Hispanic	0.12	0.09	Sex	0.52	0.50
White	0.36	0.36	Weeks pre-term	7.0	7.0
Married (birth)	0.49	0.43	Birth order	1.9	1.9
< high school	0.37	0.43	Neonatal health	99.6	100.9
High school	0.27	0.28	Twin	0.17	0.19
Some college	0.22	0.17			
College grad	0.13	0.13	*Father*		
Cigarettes (preg)	0.35	0.35	Black	0.52	0.55
Alcohol (preg)	0.13	0.11	Hispanic	0.12	0.10
Drugs (preg)	0.03	0.04	White	0.36	0.35
Worked (preg)	0.59	0.60			
Prenatal care	0.96	0.94			

Balance across treatment and control groups



Estimated impact: age 3 test scores

- *Regress:*

```
Y ~ treat + covariates
```

- *Estimated impact:* +6.4 (se = 1.2)

Increase precision through design?

Randomized Block Experiment

Randomized Block Experiments

- Divide data set into "blocks" (groups, strata…)
  - → Based on age, education, etc.
- Randomize **separately** within each group

Randomized Block Experiments

By grouping the subjects, one can ensure that subjects are "balanced" across groups with respect to these variables.

Particularly useful when ….

- sample size is small
- treatment effects vary across these covariates
- the probability of being assigned to treatment varies across blocks

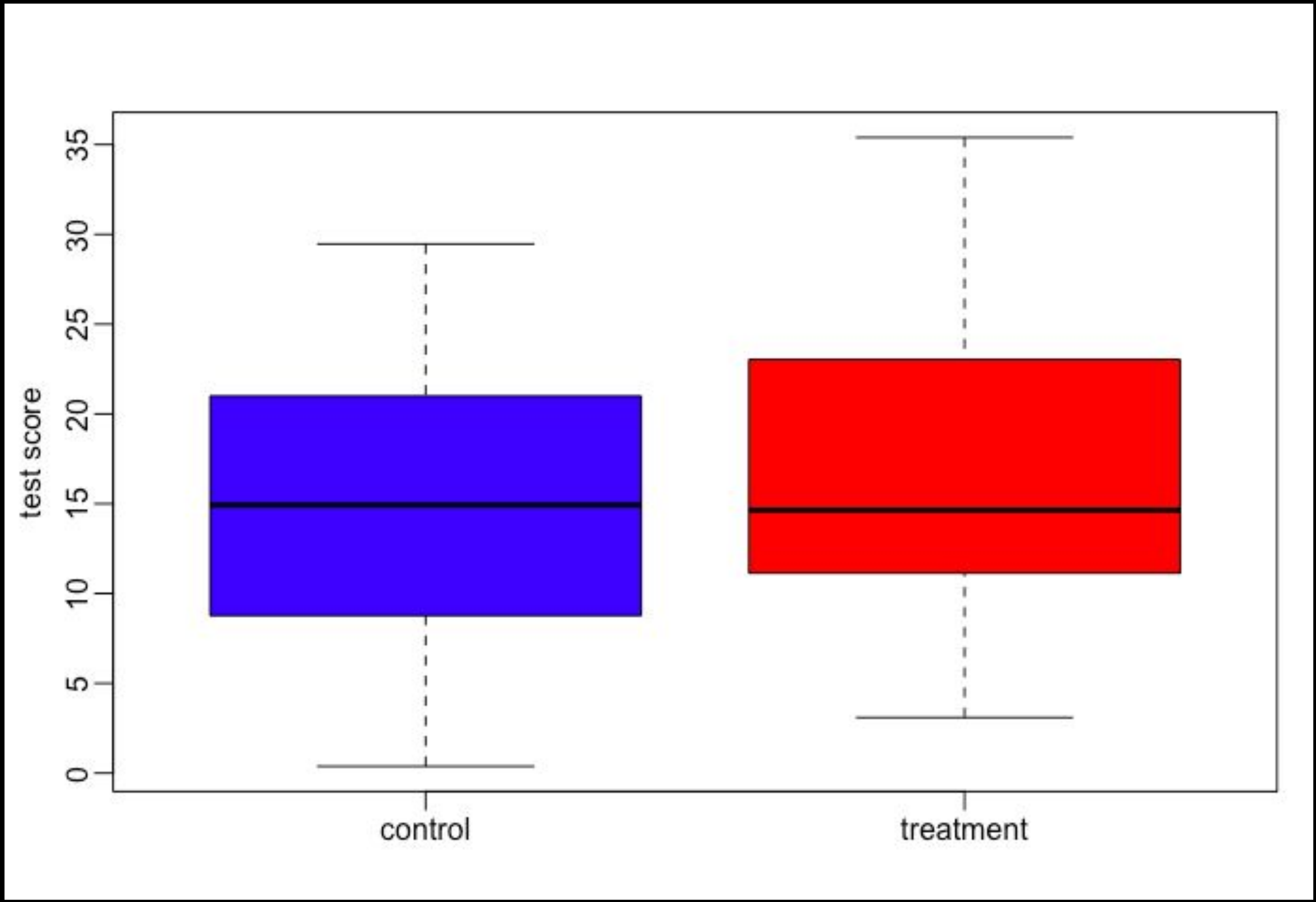
Randomized Block Experiments

By grouping the subjects, one can ensure that subjects are "balanced" across groups with respect to these variables.

Particularly useful when ….

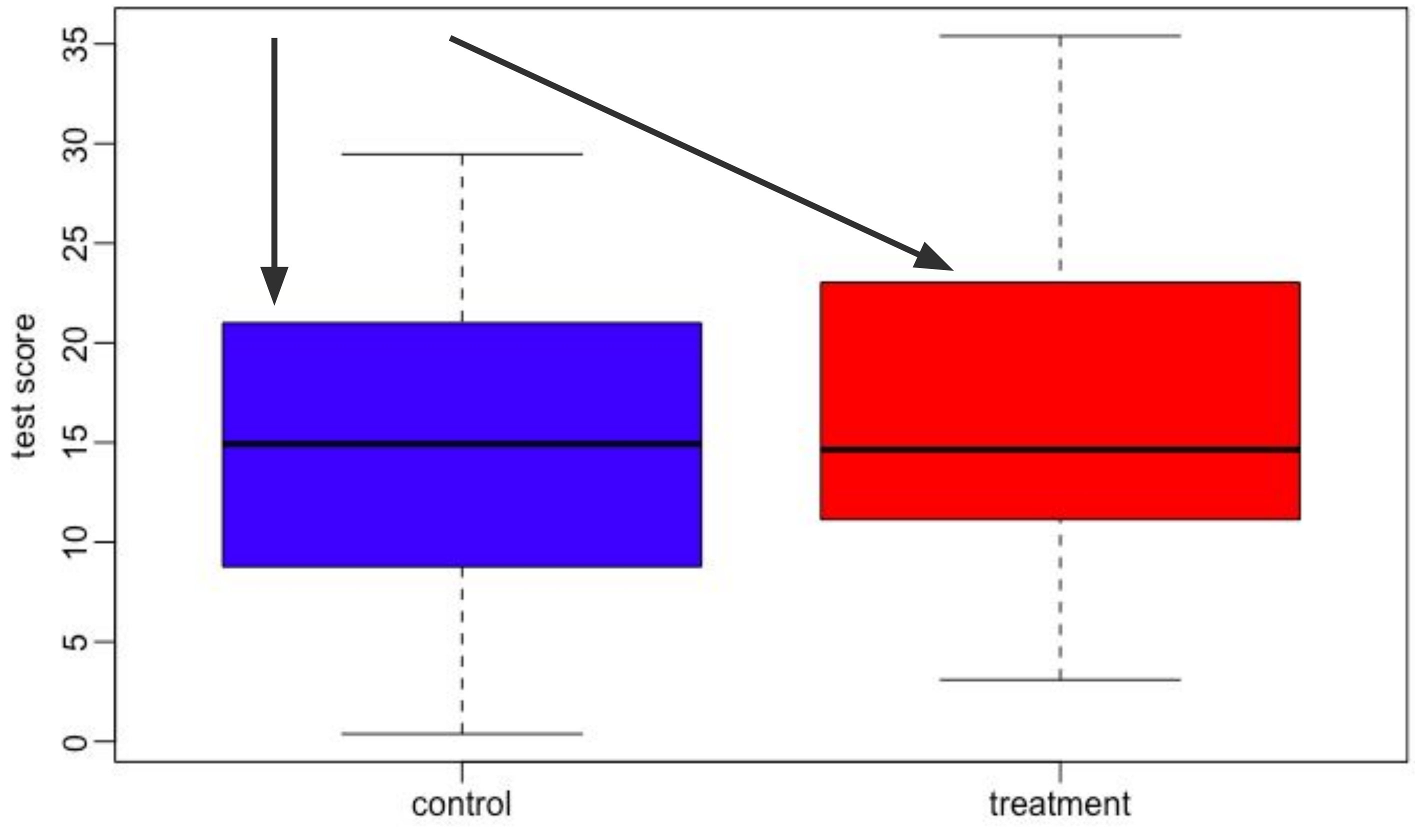
- sample size is small
- blocks are predictive of outcomes
- it's important to give greater access to some groups
- treatment effects are expected to vary across groups

Randomized Experiment without Blocking



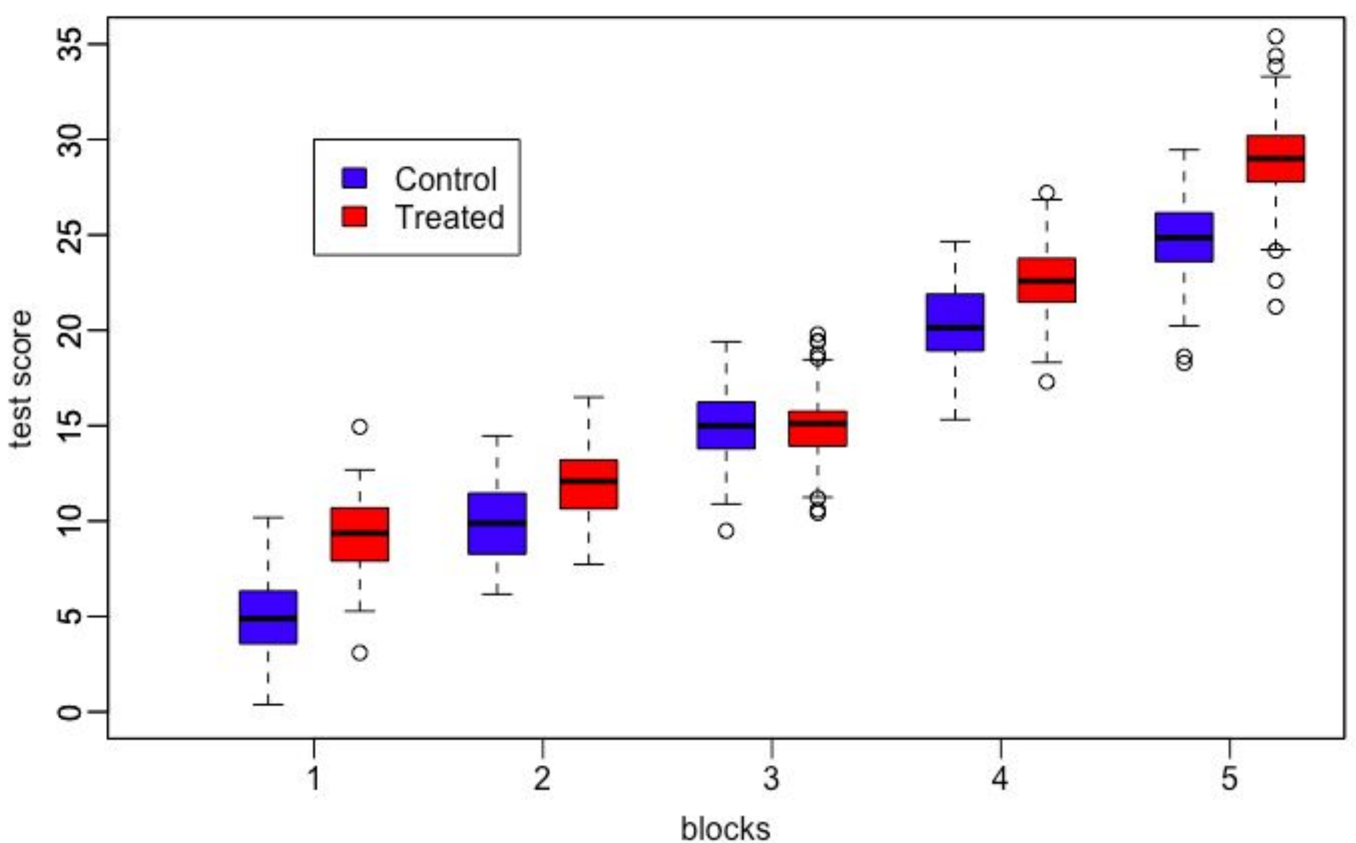
Randomized Experiment without Blocking

Look at all the unexplained variance -- that's what is feeding the standard error of our estimate!

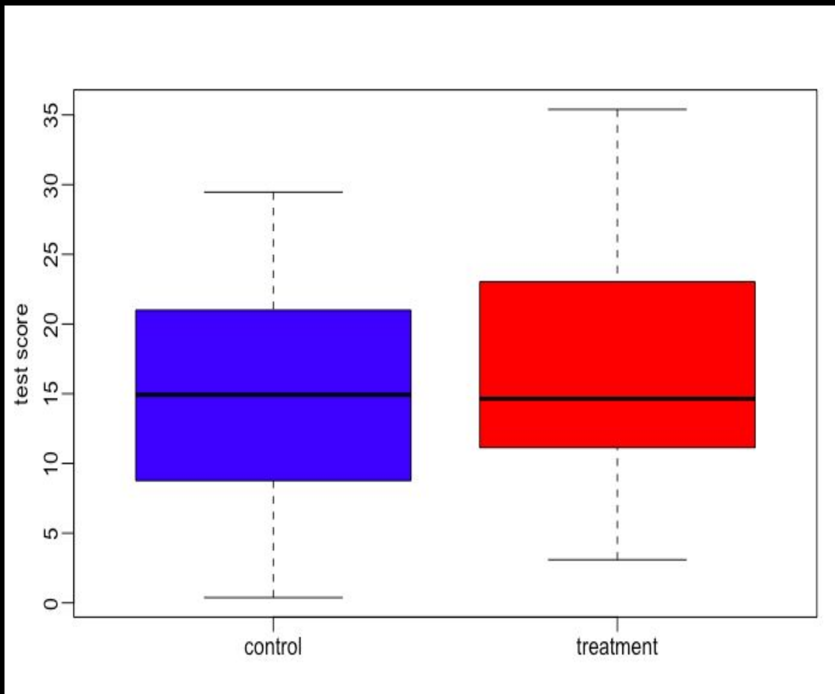


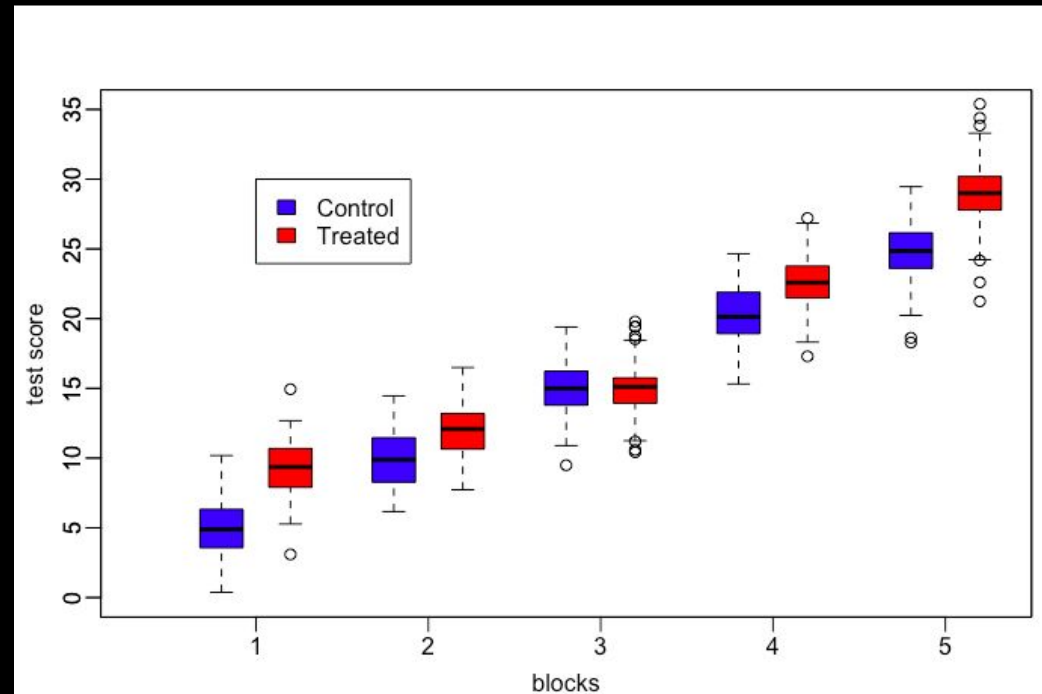
Randomized Block Experiments

See how the unexplained variance has been drastically reduced!



Compare experiments with and without blocking





Randomized Block Experiment: Assumptions

Formally we say that within any block the distribution of potential outcomes is the same across treatment groups,

$$Z \perp Y(0), Y(1) \mid W$$

where W denotes blocks.

It is *not* necessarily true that: $Z \perp Y(0), Y(1)$

Randomized Block Experiment: Assumptions

Colloquially we say that within any block the groups are balanced (on average) in all pre-treatment variables.

There should be no systematic differences between groups.

Terms that capture this idea: ignorability, no hidden bias, all confounders measured, selection on observables, exchangeability. These are more often used with observational studies.

Estimation

To estimate the average treatment effect, we can

average up block-specific treatment effects (different weights for different estimands)

run a regression on treatment and block indicators (possibly with interactions)

Randomized experiment:
friend or foe?

Advantages of randomized experiments

- Unbiased estimate of the treatment effect (assuming no additional complications)
- Fair (if oversubscribed/insufficient resources for all)
- Simpler (at least to analyze)
- Can reduce need for data collection
- More convincing evidence to funders, policy makers

Disadvantages of randomized experiments

- Cost
- Administrative burden
- Ethical?
- Necessarily prospective
- Requires a higher level of buy-in from subjects and practitioners
- Can trade-off “internal validity” for “external validity”
- “But I already know my program works!”

Ethical arguments against randomization

Feels unfair to withhold from some people

Benefits don't necessarily go to the most needy

People receiving a treatment they deem to be beneficial will eventually lose access to that

Do we have to keep the study going if we can tell before the scheduled end of study that the treatment is beneficial

Ethical arguments in favor of randomization

Giving some things to some people may be better than giving nothing to anyone

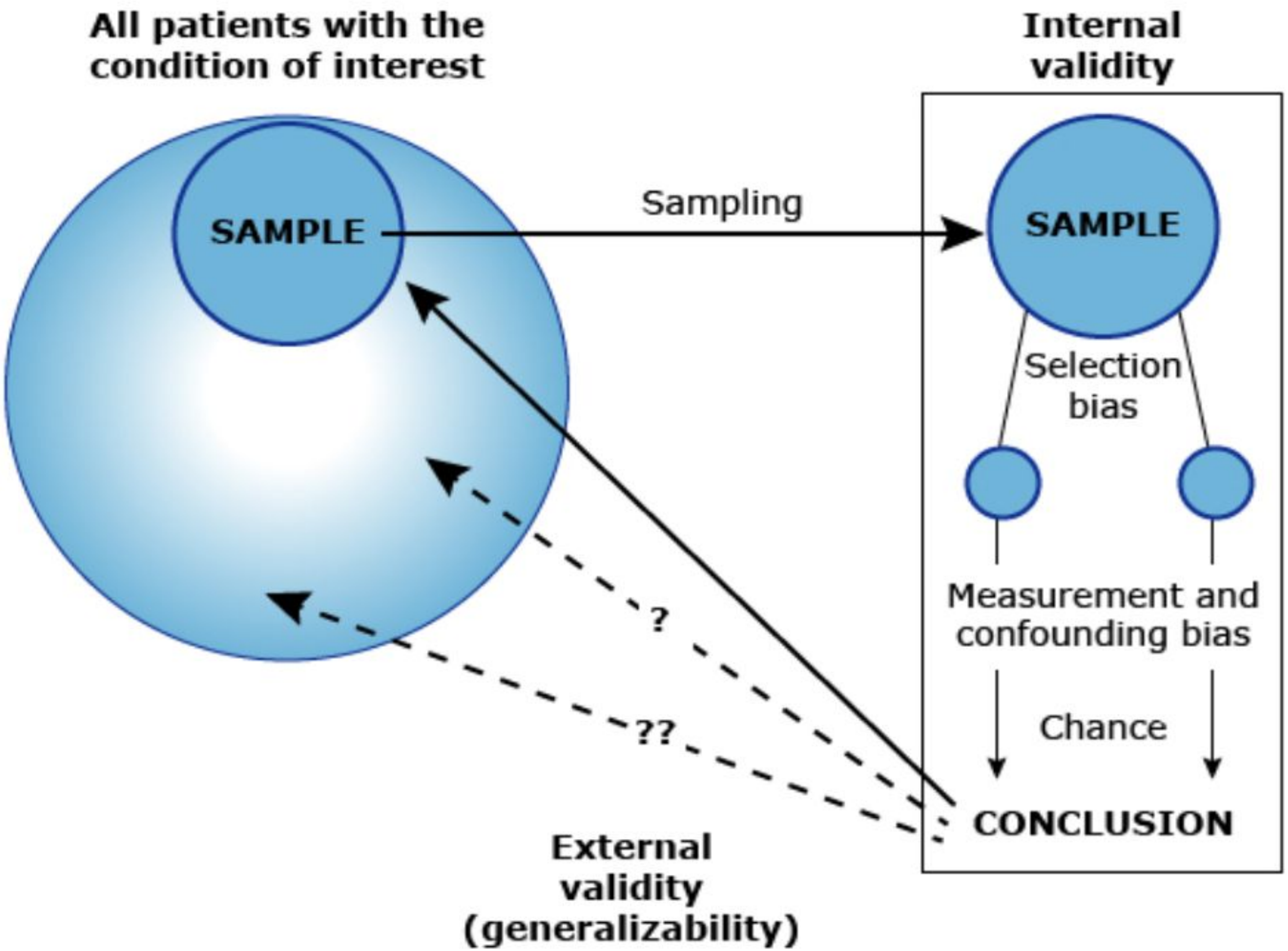
Strong evidence that might influence adoption of a program

Don't have resources for everyone to get the treatment it could be the most ethical choice

You don't know if something is effective

Can stop the study if you find treatment is very effective (but then lose the ability for looking at the impact of long-term outcomes)

Internal and external validity



internal and external validity

What are my other options?

Variations on traditional randomized experiments

Alternatives to traditional randomized experiments

Hold out groups (100% of folks in need get services, everyone else randomized)

Waitlist controls designs (Those randomized to the control group are guaranteed to receive the services after a specified amount of time)

Randomized encouragement designs (Randomize encouragement or incentives)

Randomized block designs (higher probability for those in most need)

Randomized encouragement designs: estimation

Suppose you randomize encouragement

- those not encouraged can still get the treatment
- those encouraged not forced to take up the tx

Cleanest estimation is for the effect of encouragement

Can also estimate the effect of the treatment, but need to make additional assumptions (*instrumental variables*)

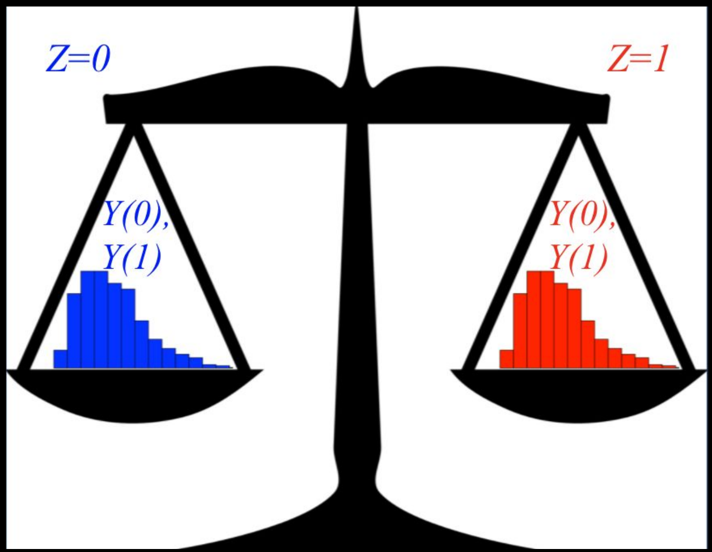

WHOA. YOU'RE ONLY 30. GOT TO SAY YOU LOOK A LOT OLDER.
I EXPERIMENTED A LOT IN COLLEGE.
LETS JUST SAY IT TOOK ITS TOLL.
A FEW YEARS EARLIER
GOD DAMN YOU, WHY WON'T YOU WORK.
goneintoRapture.com

BREAK

Design Solution!!

Randomized experiments: Ignorability satisfied (with blocks, X)

Randomized experiment

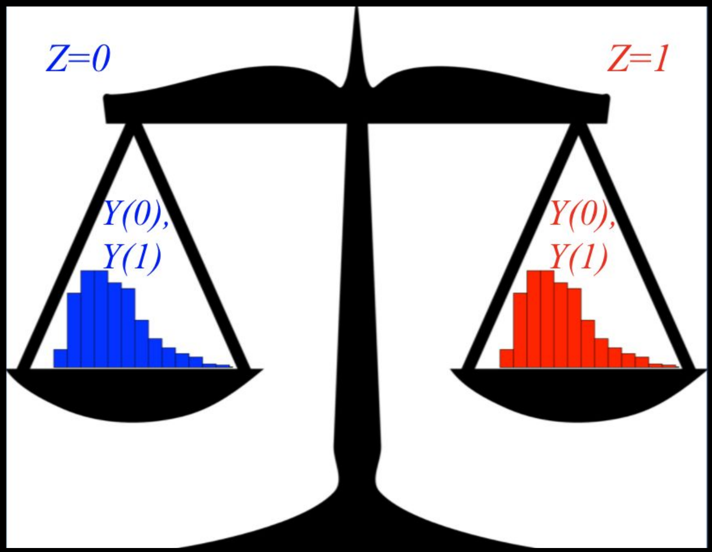


$$Y(0), Y(1) \perp Z$$

Design Solution!!

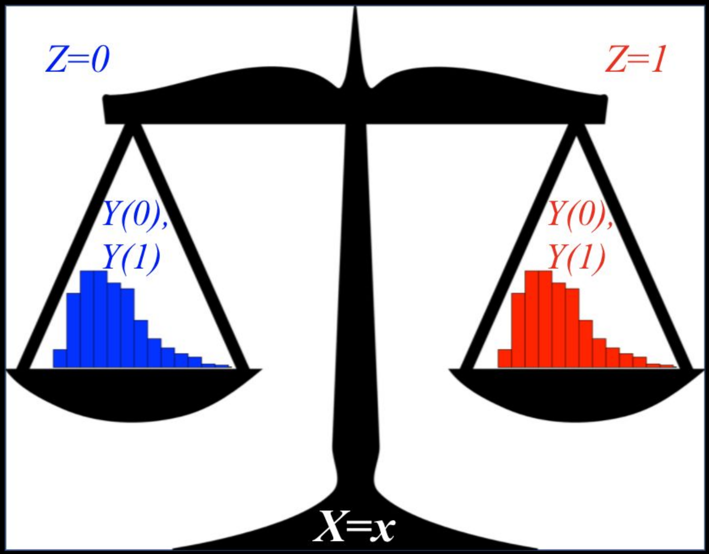
Randomized experiments:
Ignorability satisfied (with blocks, X)

Randomized experiment



$$Y(0), Y(1) \perp Z$$

Randomized block experiment



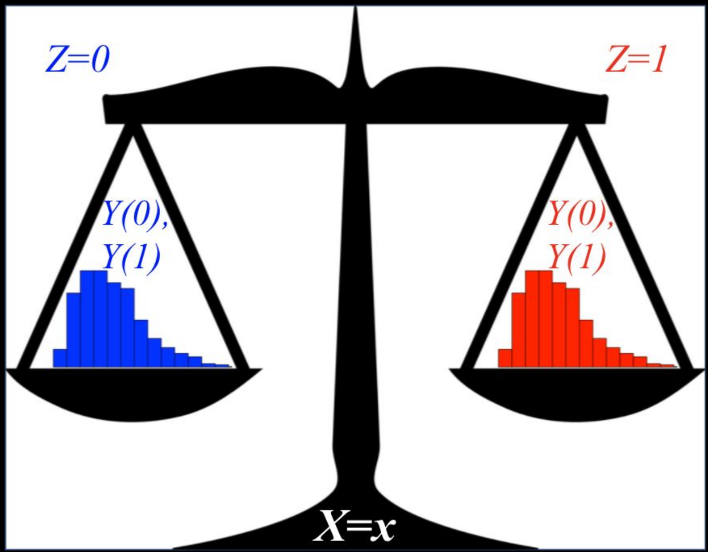
$$Y(0), Y(1) \perp Z \mid X$$

Observational study: Ignorability **ASSUMED** conditional on covariates X

Observational studies

We hope our observational study is like a complicated randomized block experiment.

This requires measuring the right set of confounders, X



$$Y(0), Y(1) \perp Z \mid X$$

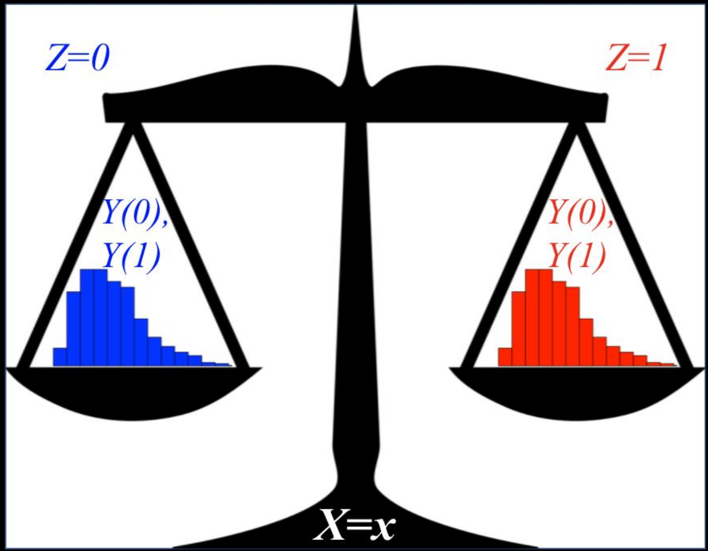
Observational study: Ignorability **ASSUMED** conditional on covariates X

Leap of faith "solution"!!

We hope our observational study is like a complicated randomized block experiment.

This requires measuring the right set of confounders, X

Observational studies



$$Y(0), Y(1) \perp Z \mid X$$

Design summary

Randomized (or natural) experiments

- great but rare
- may be limited to narrow questions or populations
- still challenging to understand when, why, and for whom

Design summary

Randomized (or natural) experiments

- great but rare
- may be limited to narrow questions or populations
- still challenging to understand when, why, and for whom

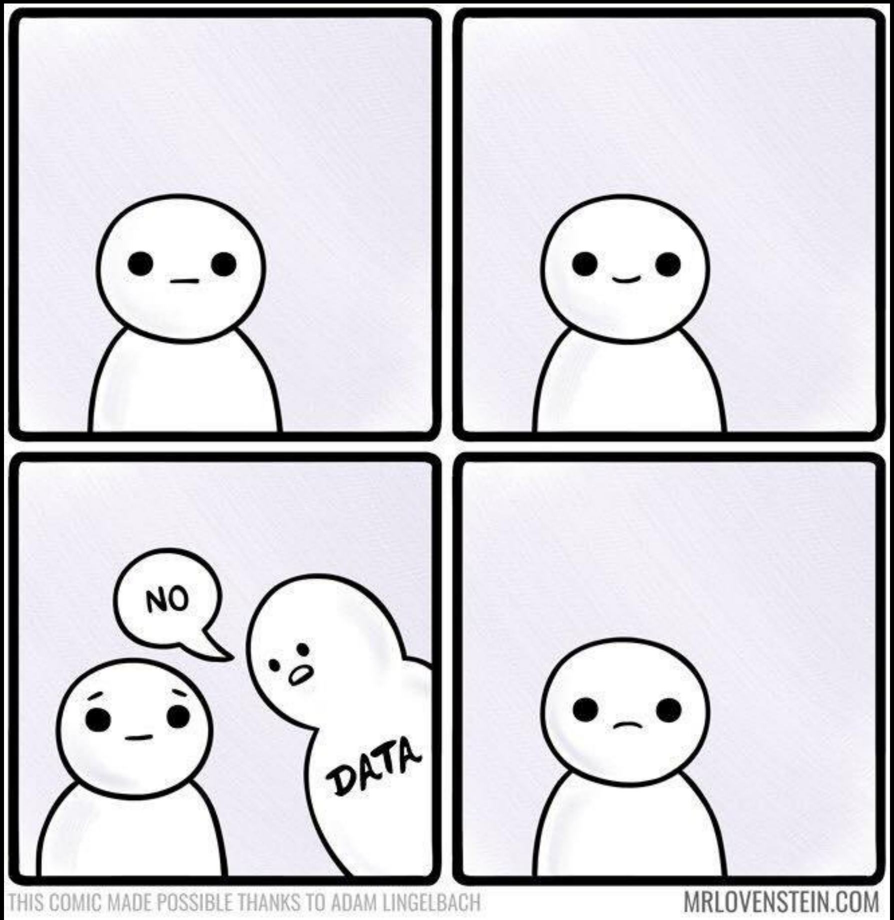
Observational studies and quasi-experiments

- often necessary due to ethics, logistics, time, money....
- often requires appropriately conditioning on many covariates (proxies for potential outcomes) to satisfy ignorability **(the more covariates the stronger the parametric assumptions)**
- alternately we need to capitalize on particular data structures

Agenda

Quasi-experimental designs and methods

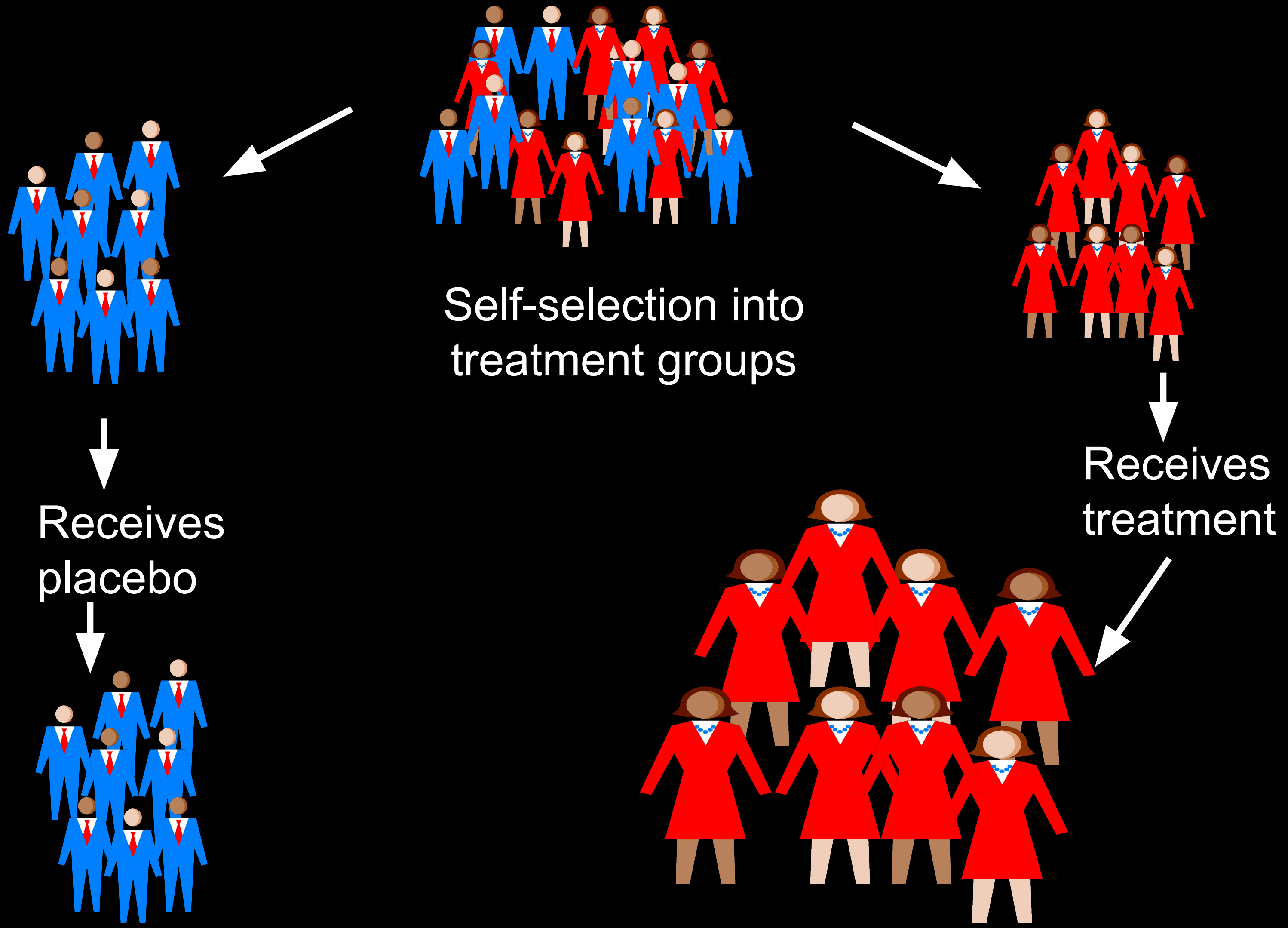
- Matching
- Difference In Differences (DID)
- Interrupted Time Series
- Regression Discontinuity Designs(RDD)
- Machine learning?



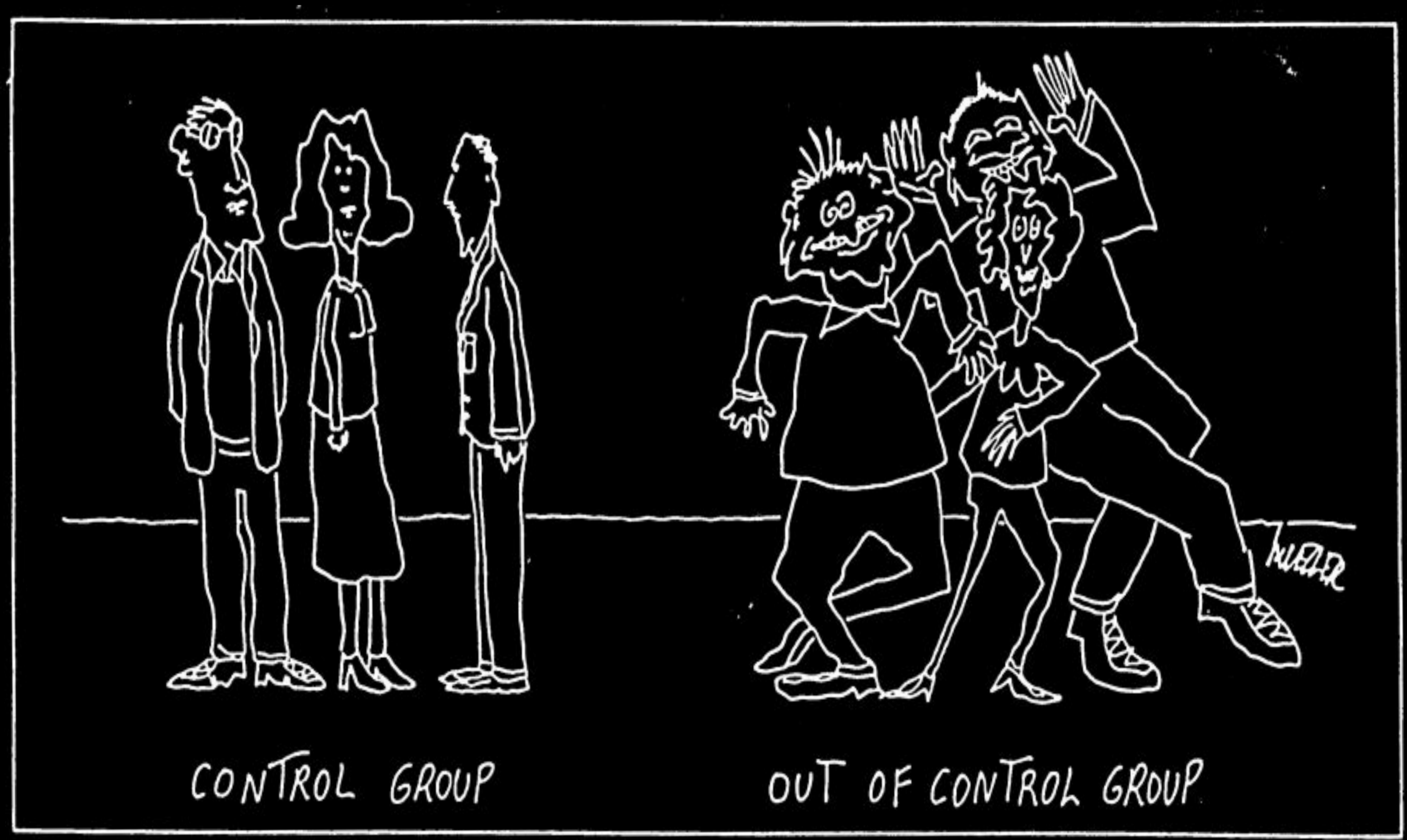
Quasi-experimental designs and methods

What happens in the absence of randomization?

- Observations “self-select” into treatment groups
- Treatment and control groups are likely to be different in important ways (age, income, race, “motivation”, health)
- If characteristics that differ across groups also predict outcomes we can’t distinguish whether differences in outcomes are caused by the treatment or covariates.
- Accordingly these are called *confounding covariates*
- The bias caused by this self-selection is often referred to as *selection bias* or *confounding*


Self-selection into treatment groups
Receives placebo
Receives treatment

Designs



Design our observational study

- Design: Focus on approximating randomized trial

Design our observational study

- Design: Focus on approximating randomized trial

Emulate design of randomized trials → no outcomes

Restructure data so treated and control units are similar

- How do we do this with many covariates?

Propensity score: a useful one-number summary

$$e(X) = \mathbb{P}(Z \mid X)$$

Conditional probability of treatment given X

- e.g., prob of treatment given age and education

Propensity score: a useful one-number summary

$$e(X) = \mathbb{P}(Z \mid X)$$

Conditional probability of treatment given X

- e.g., prob of treatment given age and education

Propensity score theorem

$$Z \perp Y(0), Y(1) \mid e(X) \quad \leftrightarrow \quad Z \perp Y(0), Y(1) \mid X$$

Propensity score: a useful one-number summary

$$e(X) = \mathbb{P}(Z \mid X)$$

Balancing score for X

If two groups of observations have similar values of e(X), they should have similar distributions of X

Match/weight units based on e(X) → similarity wrt X

Propensity score: a useful one-number summary

$$e(X) = \mathbb{P}(Z \mid X)$$

Propensity score is **known** in RCTs; here we must **estimate** it

NO MAGIC -- still **assume away** unmeasured confounders

$$Z \perp Y(0), Y(1) \mid X$$

"Simple" Template for Using Propensity Scores

Design phase: (without outcomes)

- Define treatment, select potential confounders
- Repeat until convergence:
  - Estimate propensity score
  - "Restructure" data set (matching/weighting)
  - Check balance between treated and pseudo-control units

Analysis phase: (with outcomes)

- Estimate causal effects → difference in means, "regression," ...

Classic example: National Supported Work (NSW)

Randomized evaluation of NSW in 1970s

- Training program for job skills to disadvantaged workers
- Large, positive effect on wages

Constructed observational study combines

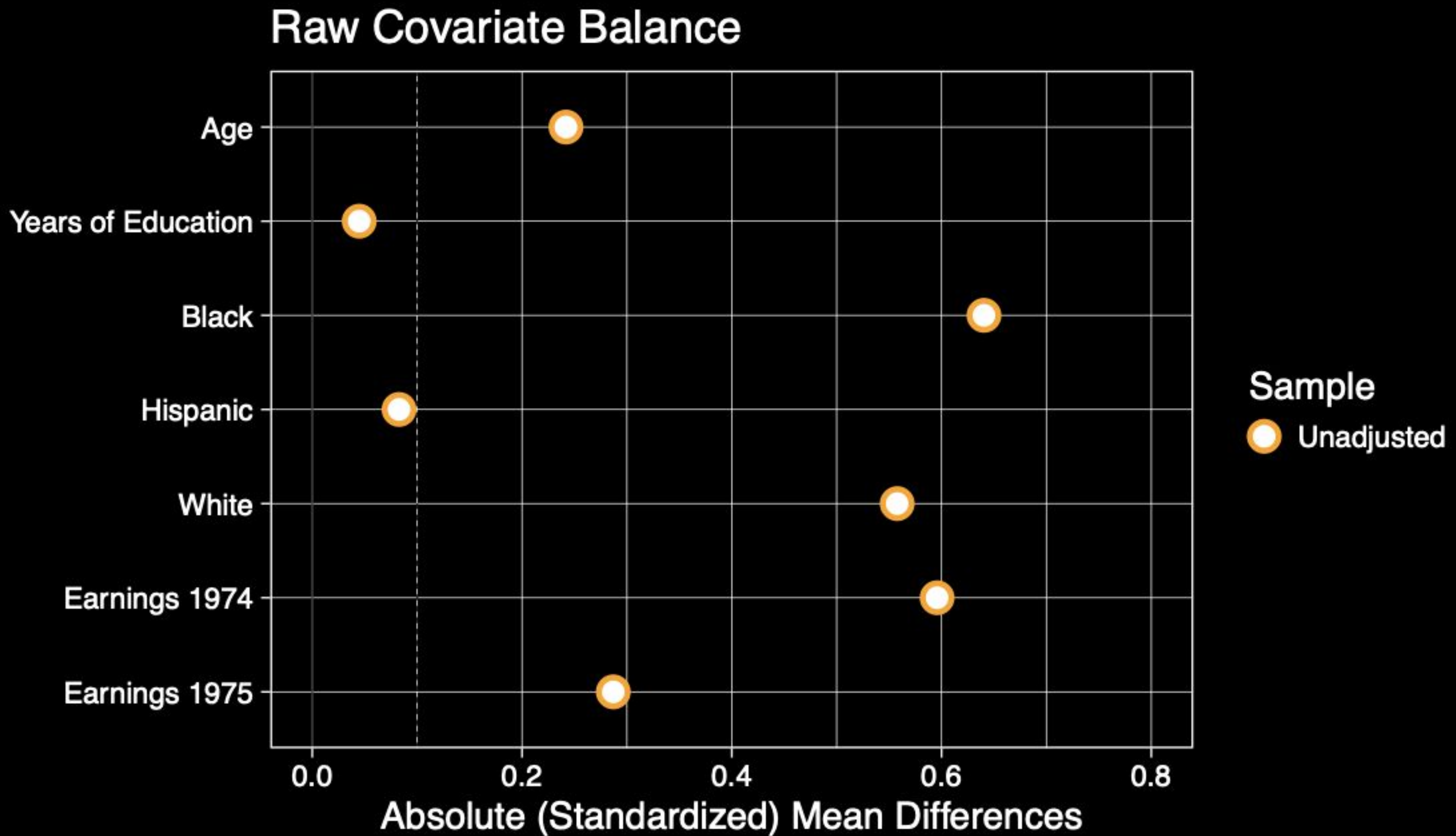
- the treatment group from NSW with
- a comparison groups from a separate survey

Can we recover the experimental estimate?

Pre-treatment Data

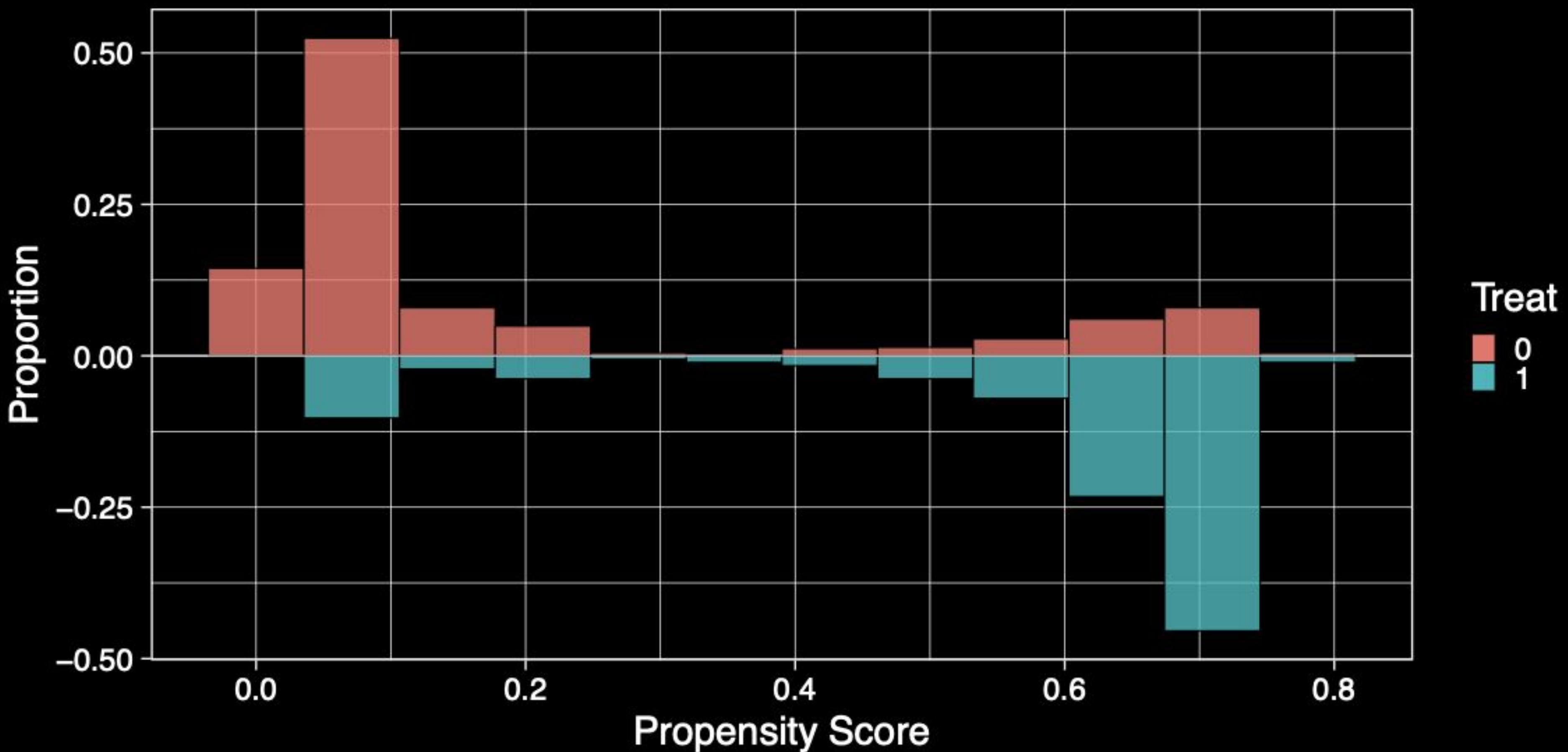
(variables that could be collected across both datasets)

- Worker demographics:
  - Age
  - Years of education
  - Race/ethnicity, coded {Black, Hispanic, White}
- Prior earnings: in 1974, in 1975


Raw Covariate Balance
Age
Years of Education
Black
Hispanic
White
Earnings 1974
Earnings 1975
0.0
0.2
0.4
0.6
0.8
Absolute (Standardized) Mean Differences
Sample
Unadjusted

Raw Data: Prop. Score Balance

Unadjusted Sample



Restructuring the data to make groups similar

Matching

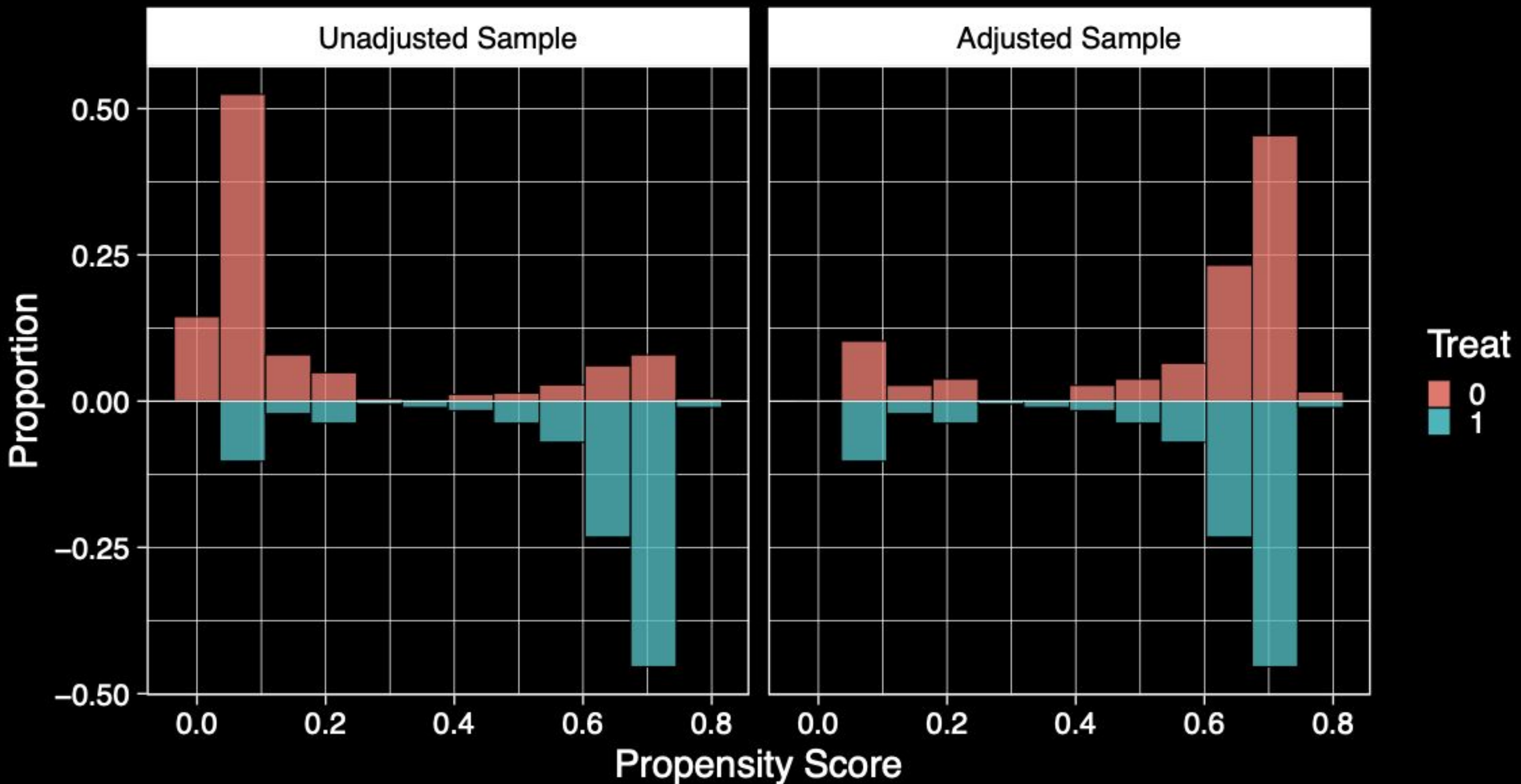
→ **For each treated unit:** find the control unit with closest estimated propensity score

Weighting for the effect of the treatment on the treated

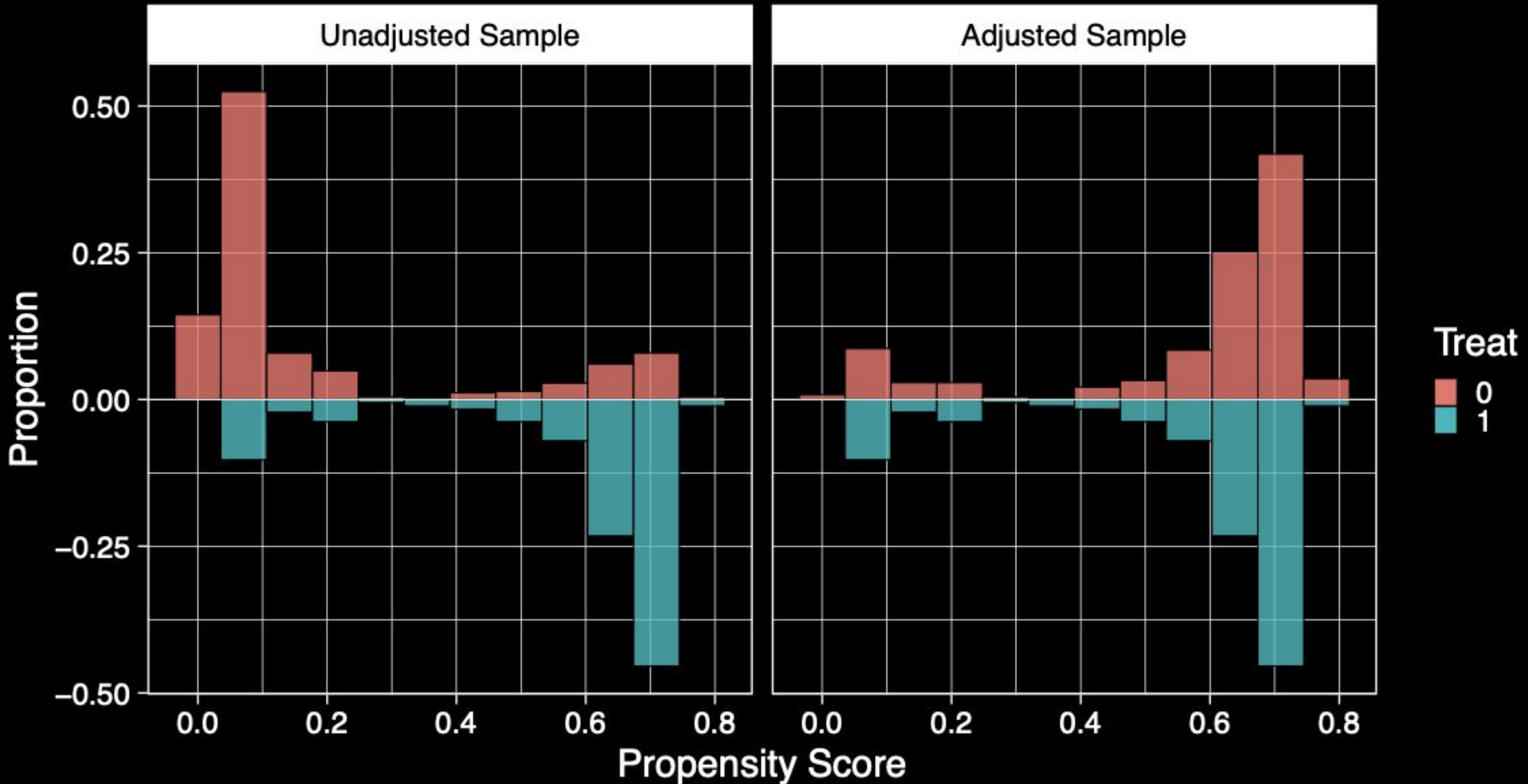
→ Assign each treated unit weight 1

→ Assign each **control** unit weight: $\frac{\hat{e}(X)}{1-\hat{e}(X)}$

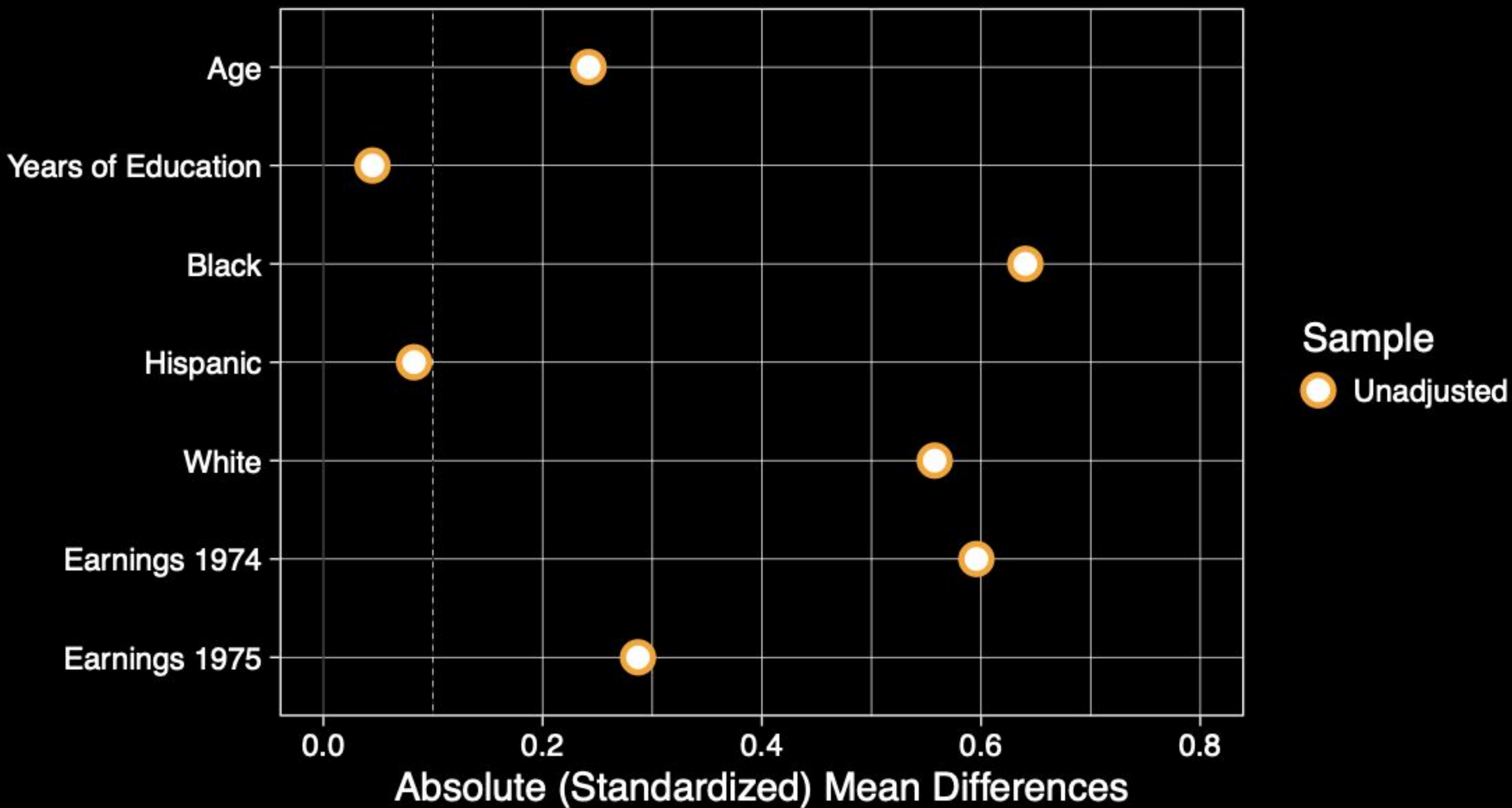
Prop. Score Matching: Prop. Score Balance



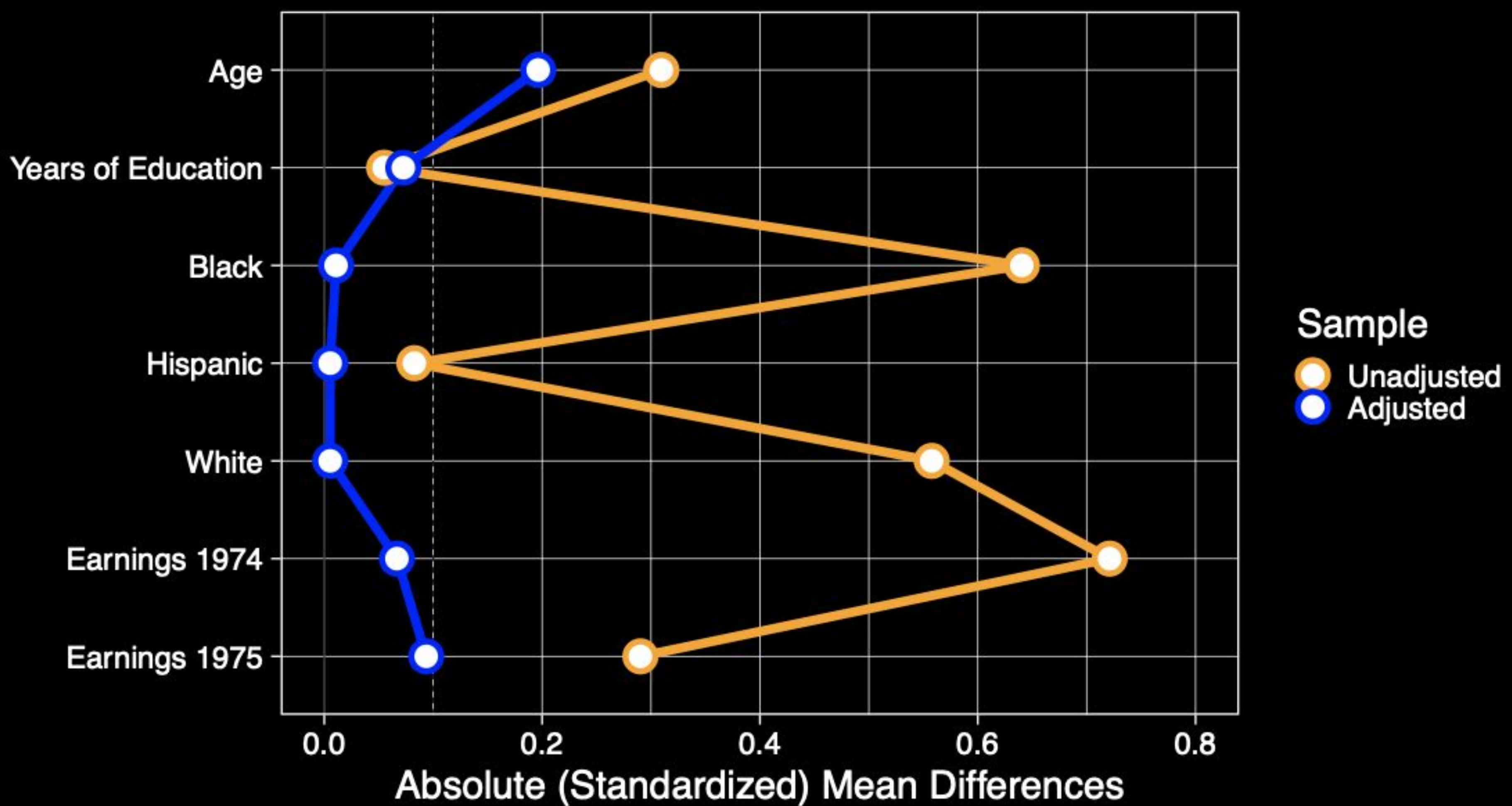
Prop. Score Weighting: Prop. Score Balance



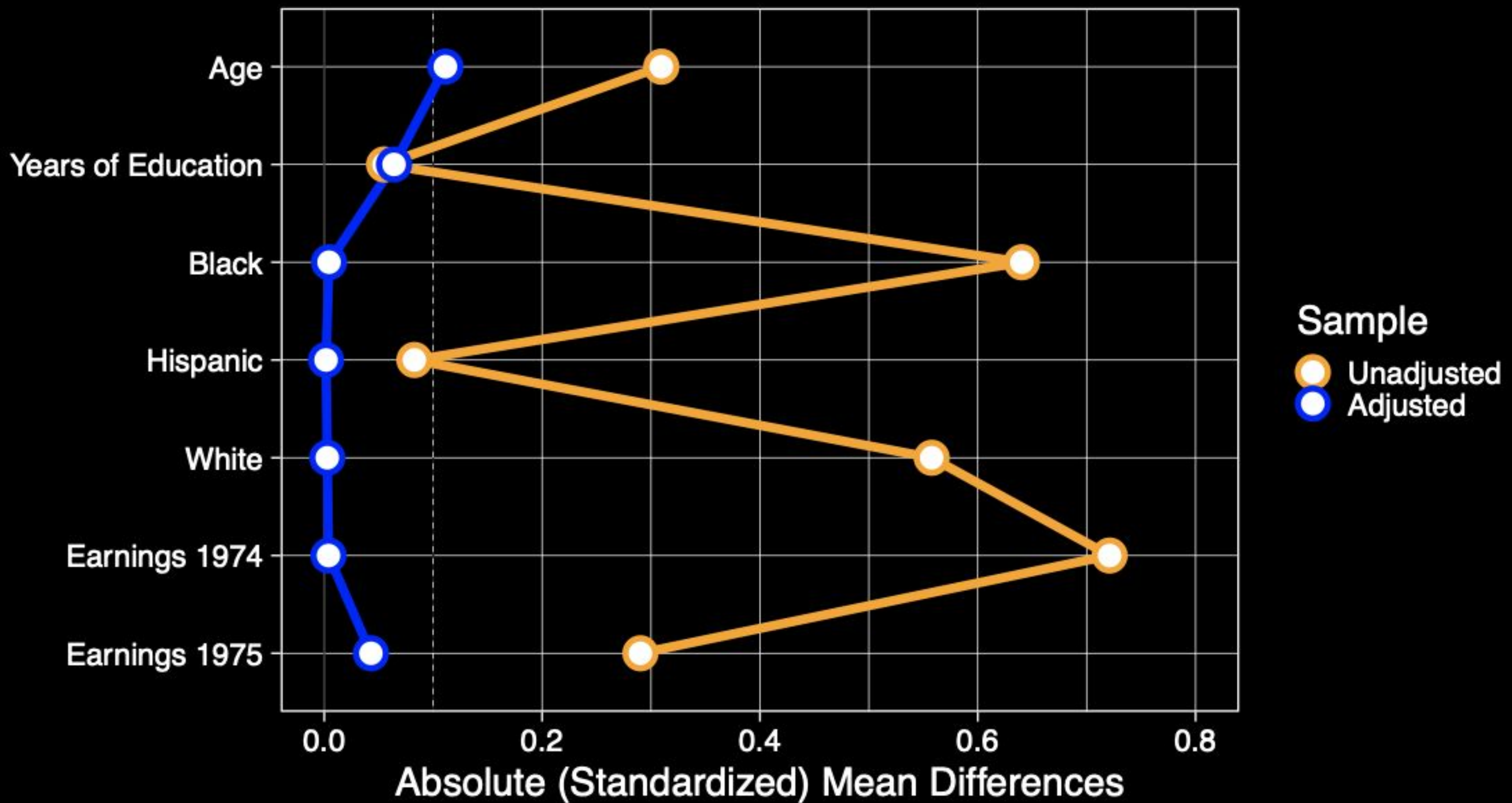
Raw Covariate Balance



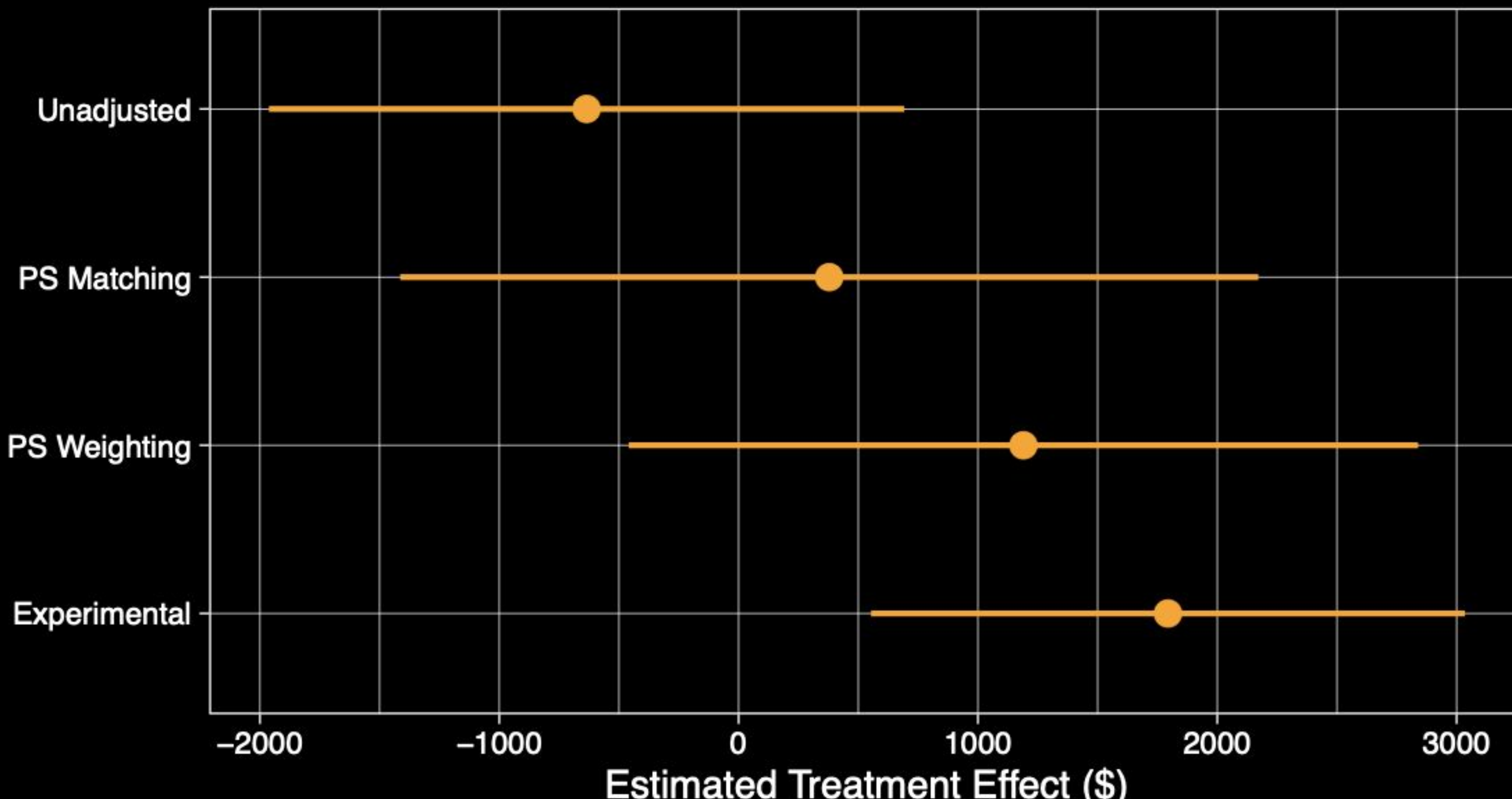
Prop. Score Matching: Covariate Balance



Prop. Score Weighting: Covariate Balance



Estimated Treatment Effect By Method



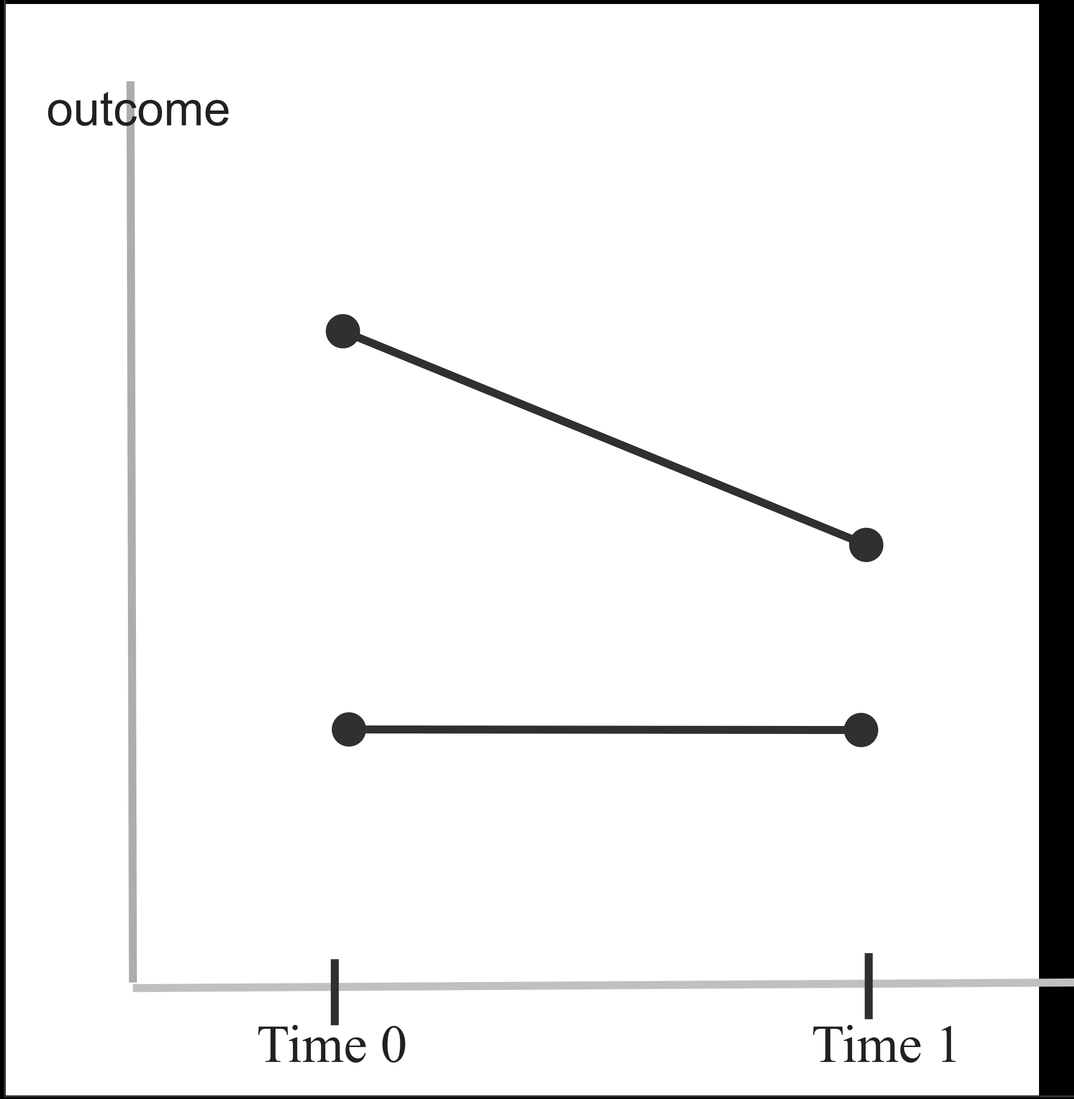
But wait, there's more!

Propensity scores are conceptually useful, but we can often do better in practice

- Find matches/weights that **directly balance** covariates
- Go beyond difference-in-means
- Adjust for covariates using a flexible model.... machine learning!

"There's a flaw in your experimental design. All the mice are scorpios."

Quasi experiments:
DID
ITS (etc)
RDD



Difference In Differences

Difference In Differences overview

DID implemented in scenarios in which

1) there are at least two groups, at least one of which received the treatment

2) there are measurements of the outcome both before and after potential treatment exposure/implementation for both groups

DID example: Litigation and bullying

RQ: "Does litigation related to sexual orientation–based harassment and discrimination in schools reduce rates of homophobic bullying?"

Context: ≅1.5 million students in 499 California high schools

DID example: Litigation and bullying

RQ: "Does litigation related to sexual orientation–based harassment and discrimination in schools reduce rates of homophobic bullying?"

Treatment: ?

DID example: Litigation and bullying

RQ: "Does litigation related to sexual orientation–based harassment and discrimination in schools reduce rates of homophobic bullying?"

Treatment: "litigation addressing alleged violations of the rights of students who are (or are perceived to be) lesbian, gay, bisexual, or transgender (LGBT) under laws prohibiting harassment or discrimination in California schools after 2000"

DID example: Litigation and bullying

RQ: "Does litigation related to sexual orientation–based harassment and discrimination in schools reduce rates of homophobic bullying?"

Treatment:

- successful LGBT harrassment/discrimination litigation
- unsuccessful LGBT harrassment/discrimination litigation

DID example: Litigation and bullying

Litigation and bullying

RQ: "Does litigation related to sexual orientation–based harassment and discrimination in schools reduce rates of homophobic bullying?"

Outcome?

DID example: Litigation and bullying

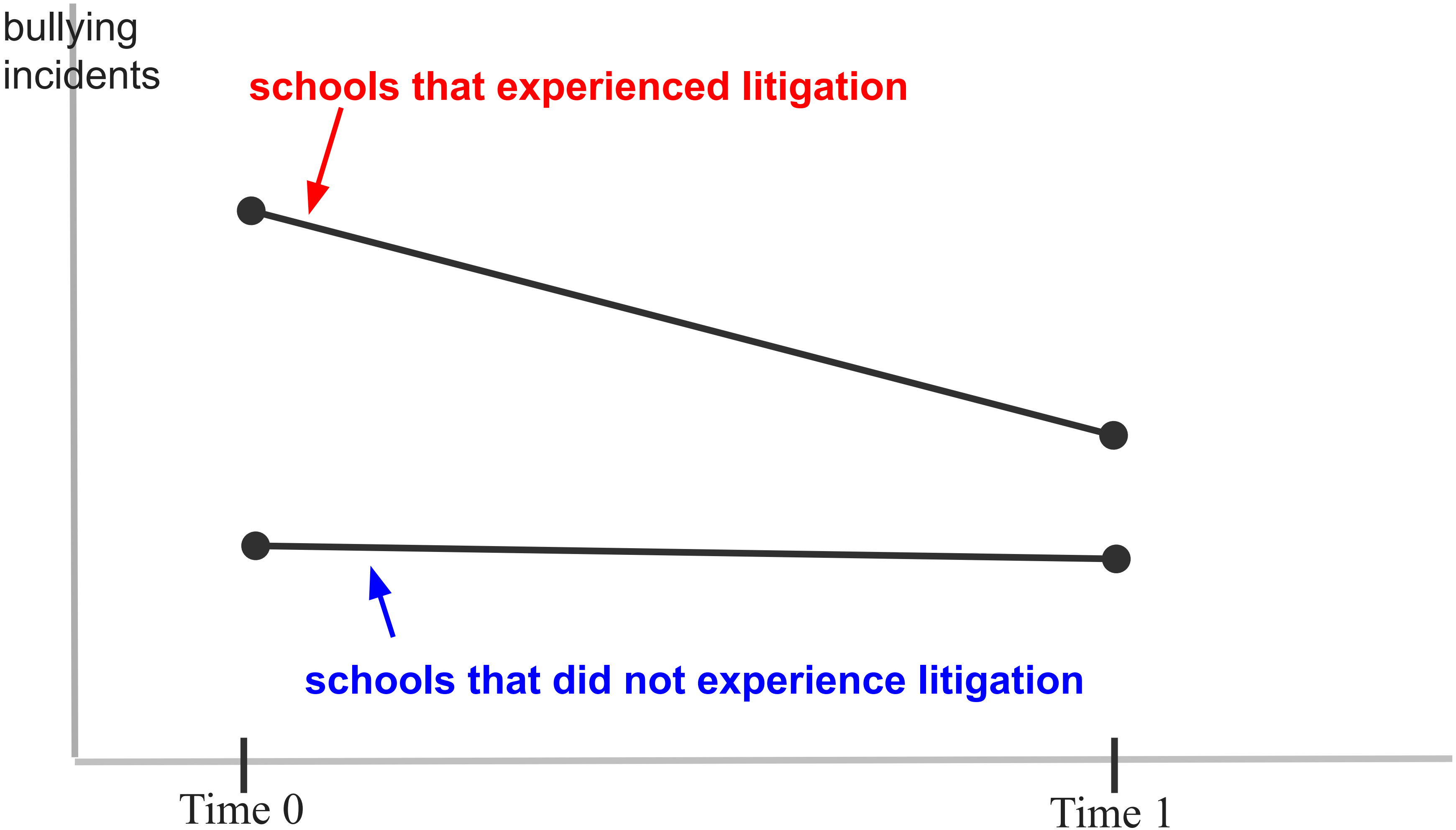
Litigation and bullying

RQ: "Does litigation related to sexual orientation–based harassment and discrimination in schools reduce rates of homophobic bullying?"

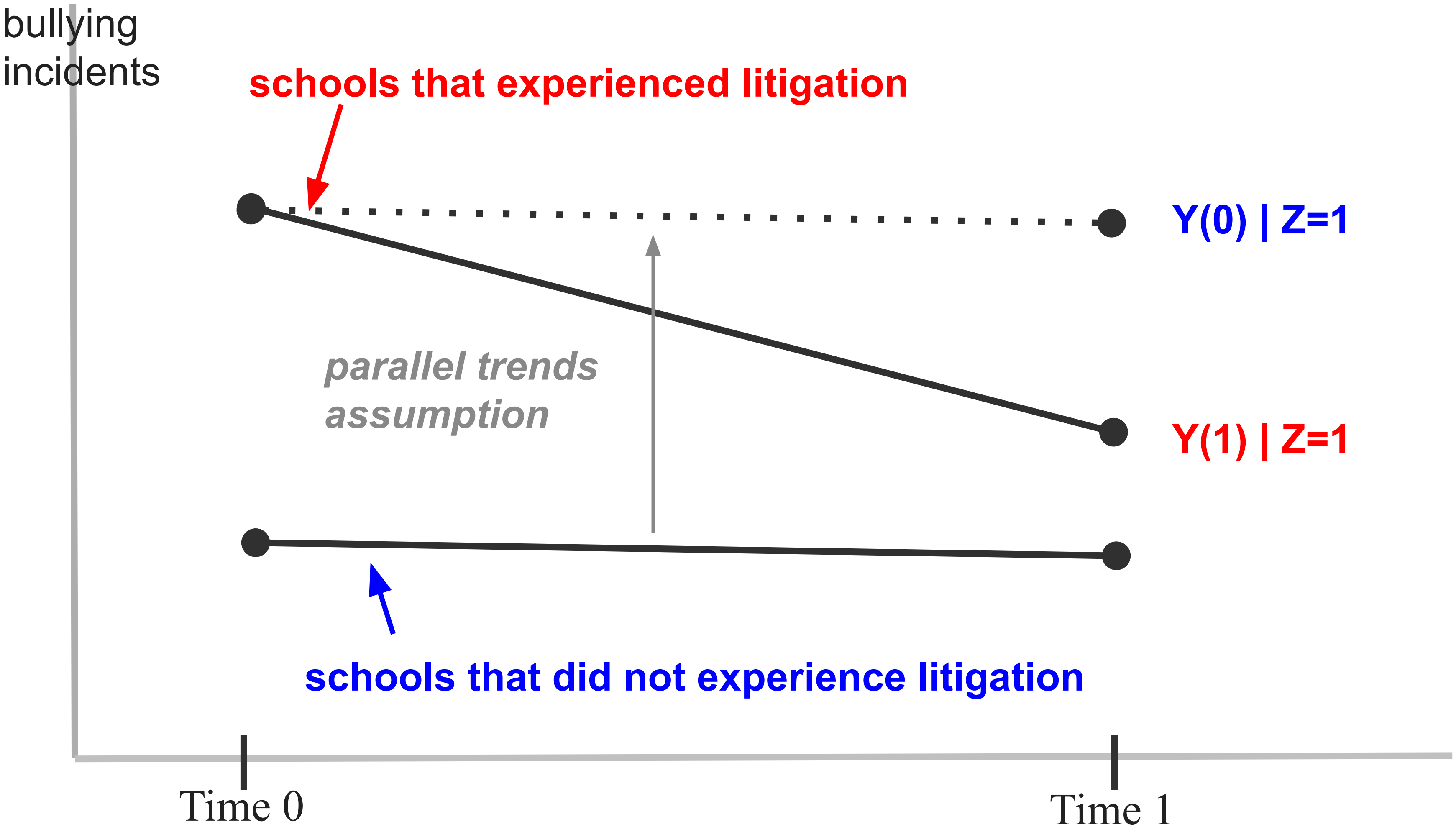
Homophobic bullying: "survey data on homophobic bullying from 1,448,778 California high school students in 499 schools."

15 consecutive waves of data from the California Healthy Kids Survey (CHKS)...collected between the 2001- 2002 and 2015-2016 academic years.

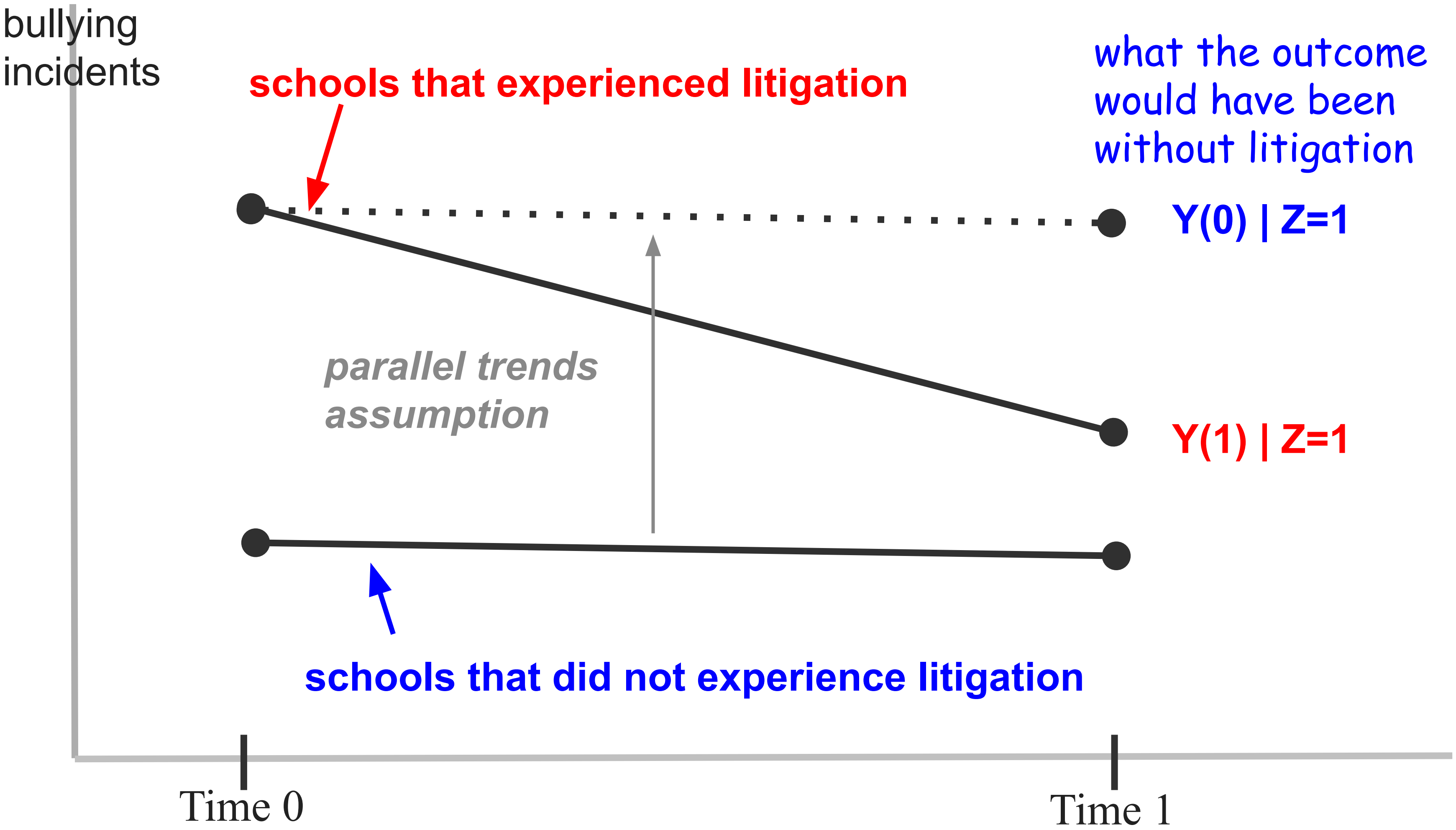
Difference in Differences: Bullying example (illustrative)



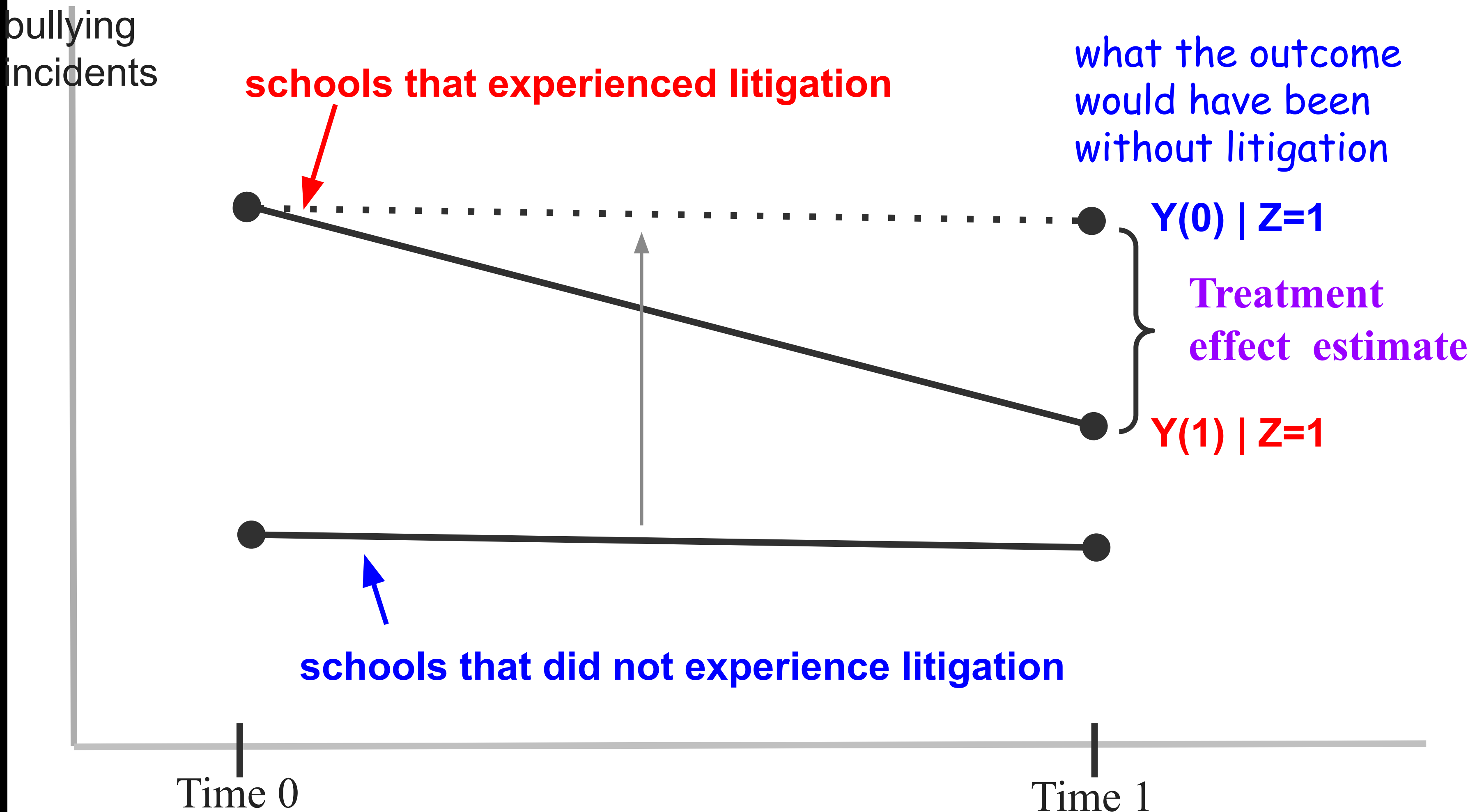
Difference in Differences: Bullying example (illustrative



Difference in Differences: Bullying example (illustrative



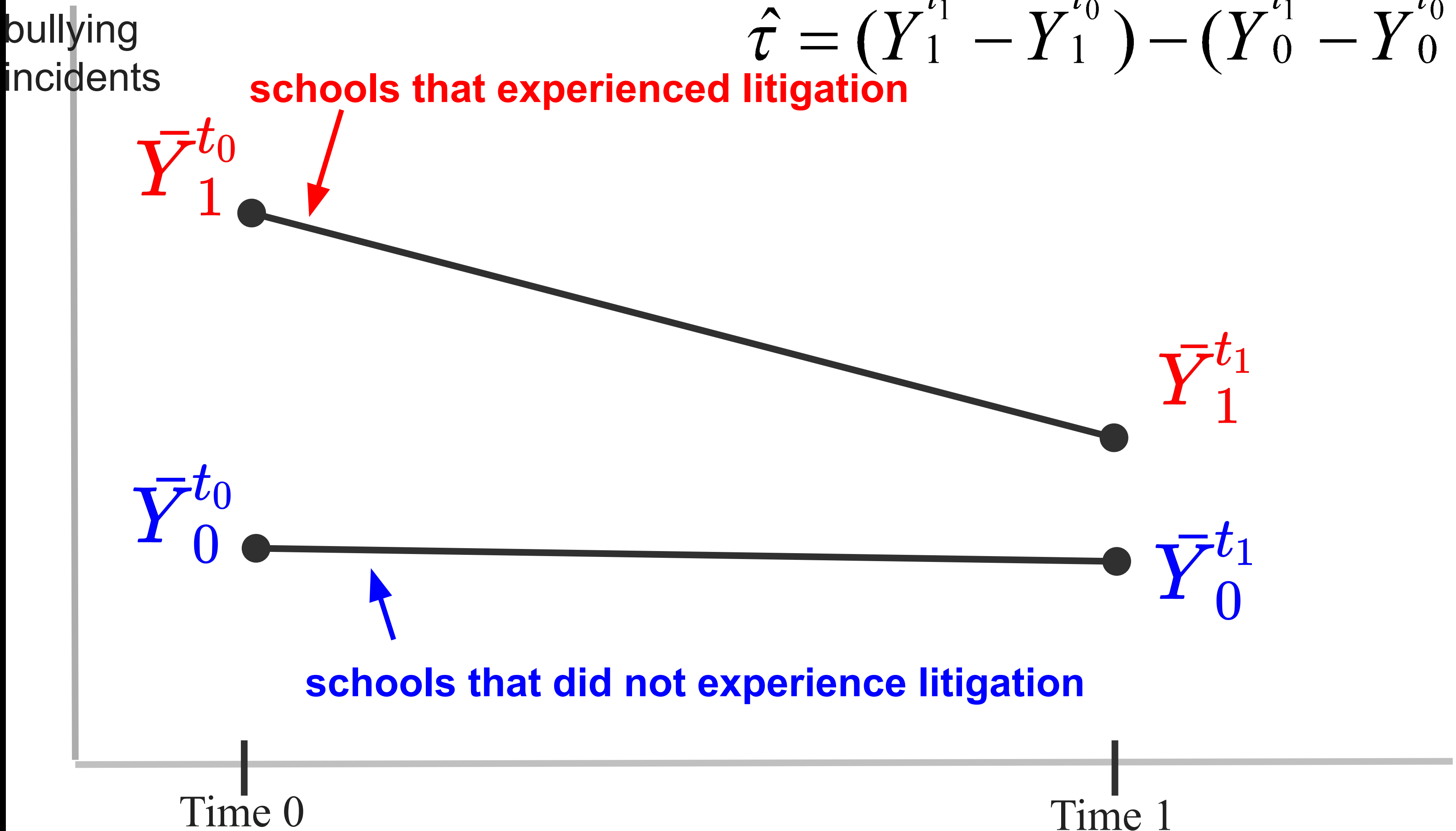
Difference in Differences: Bullying example (illustrative)



DID: estimation

Difference in Differences: Estimation

$$\hat{\tau} = (\overline{Y}_1^{t_1} - \overline{Y}_1^{t_0}) - (\overline{Y}_0^{t_1} - \overline{Y}_0^{t_0})$$



DID Estimation

If we have the individual data points we can estimate the DID effect using the following regression model.

$$E[Y|Z,T] = \alpha_0 + \lambda_0 Z_i + \delta_0 T_i + \beta Z_i T_i$$

Z_i = exposure group (school that experienced litigation or not)

T_i = time period (bullying measured pre- or post-litigation)

DID: assumptions

Parallel trends

The critical assumption for difference in difference analysis is that the change in outcomes over time for the control group represents the same change that ***would have happened*** for the treatment group if they hadn't been exposed to the treatment

Difference in Differences: Parallel trends

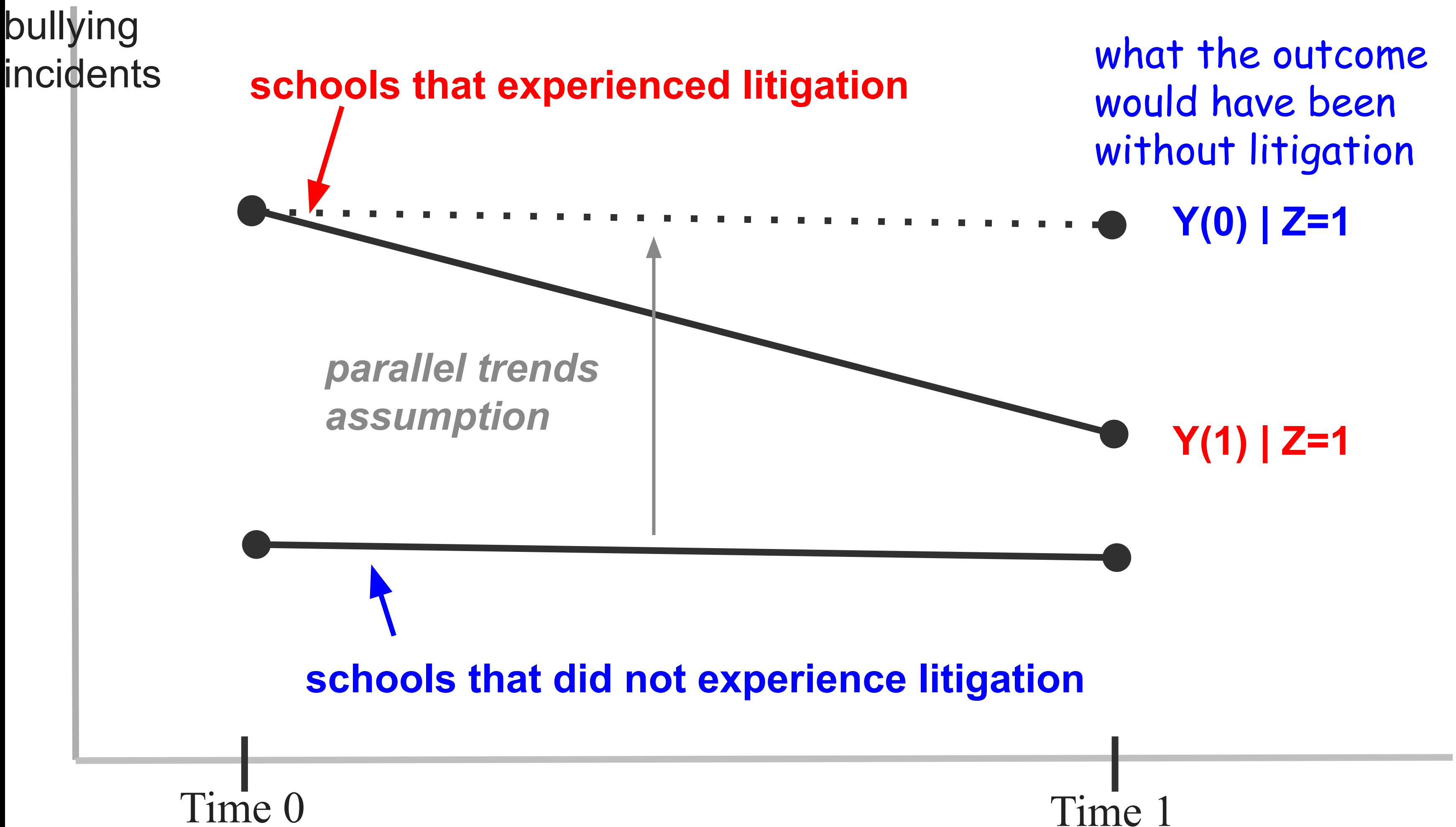
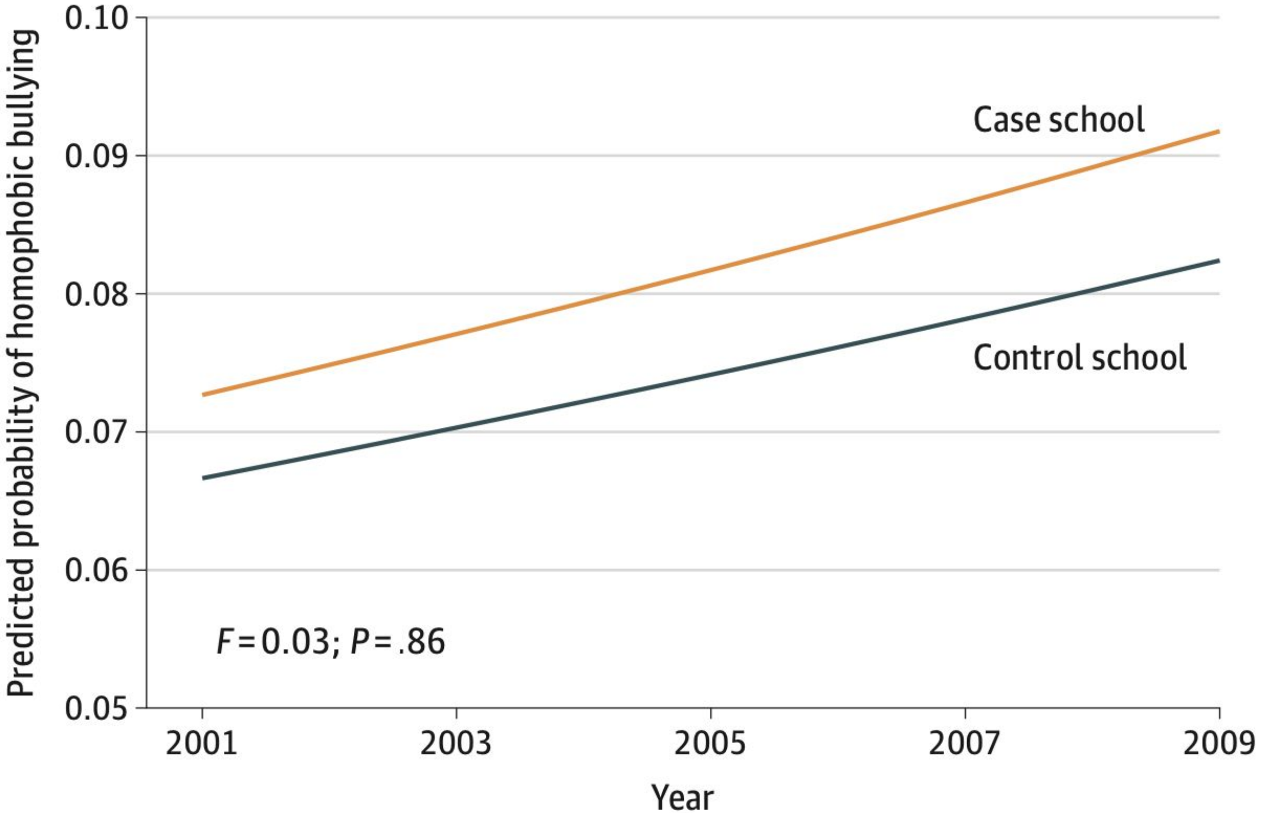


Figure 2. Test of Parallel Trends Assumption Comparing Homophobic Bullying in Case Schools With Control Schools in the Years Prior to the Case

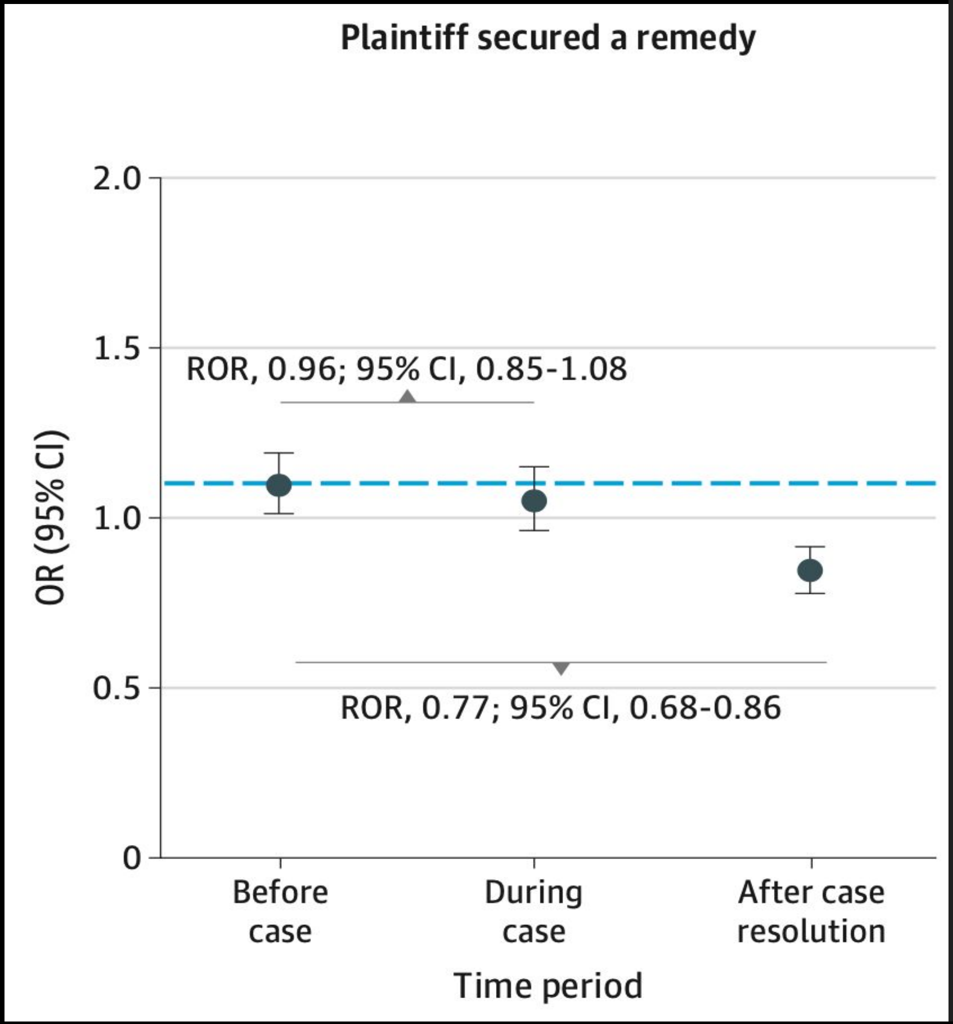


How can we try to justify the **Parallel Trends** assumption?

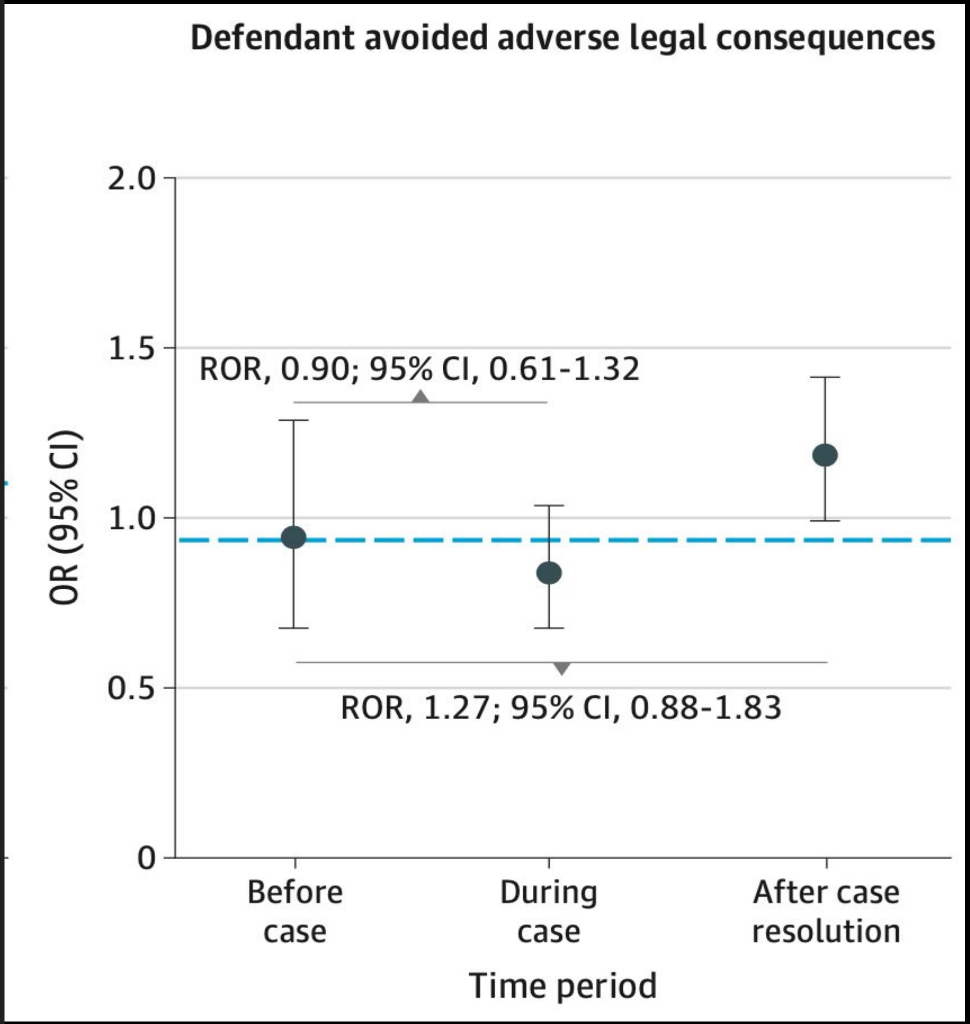
Use evidence from *pre-treatment* time periods

Supports, doesn't guarantee

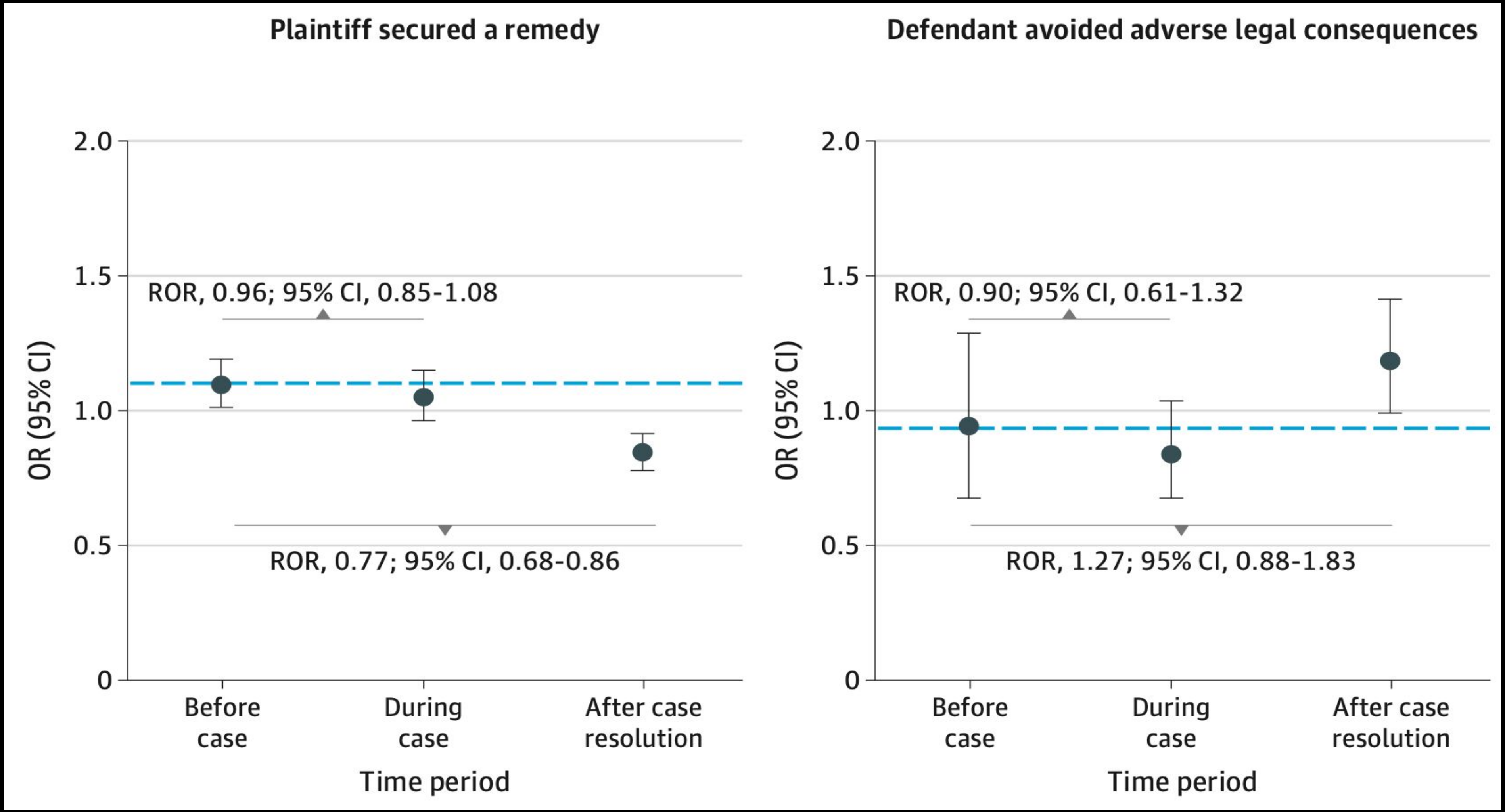
Results


Plaintiff secured a remedy
2.0
1.5
1.0
0.5
0
OR (95% CI)
ROR, 0.96; 95% CI, 0.85-1.08
ROR, 0.77; 95% CI, 0.68-0.86
Before case
During case
After case resolution
Time period

Results


Defendant avoided adverse legal consequences
OR (95% CI)
2.0
1.5
1.0
0.5
0
ROR, 0.90; 95% CI, 0.61-1.32
ROR, 1.27; 95% CI, 0.88-1.83
Before case
During case
After case resolution
Time period

Results


Plaintiff secured a remedy
Defendant avoided adverse legal consequences
2.0
1.5
1.0
0.5
0
OR (95% CI)
ROR, 0.96; 95% CI, 0.85-1.08
ROR, 0.77; 95% CI, 0.68-0.86
ROR, 0.90; 95% CI, 0.61-1.32
ROR, 1.27; 95% CI, 0.88-1.83
Before case
During case
After case resolution
Time period

DID features

1) Requires a comparison group which may help to create similarity across groups (though may not)

2) Treatment will likely be manipulable but may not be as "well-defined" as you'd like

3) Does enforce temporal ordering of treatment and outcomes (not necessarily covariates depending on analysis)

4) Makes a VERY STRONG assumption (parallel trends)

Interrupted Time Series
(and friends)

Change in trend in smoking prevalence in Pennsylvania before and after cigarette tax increase

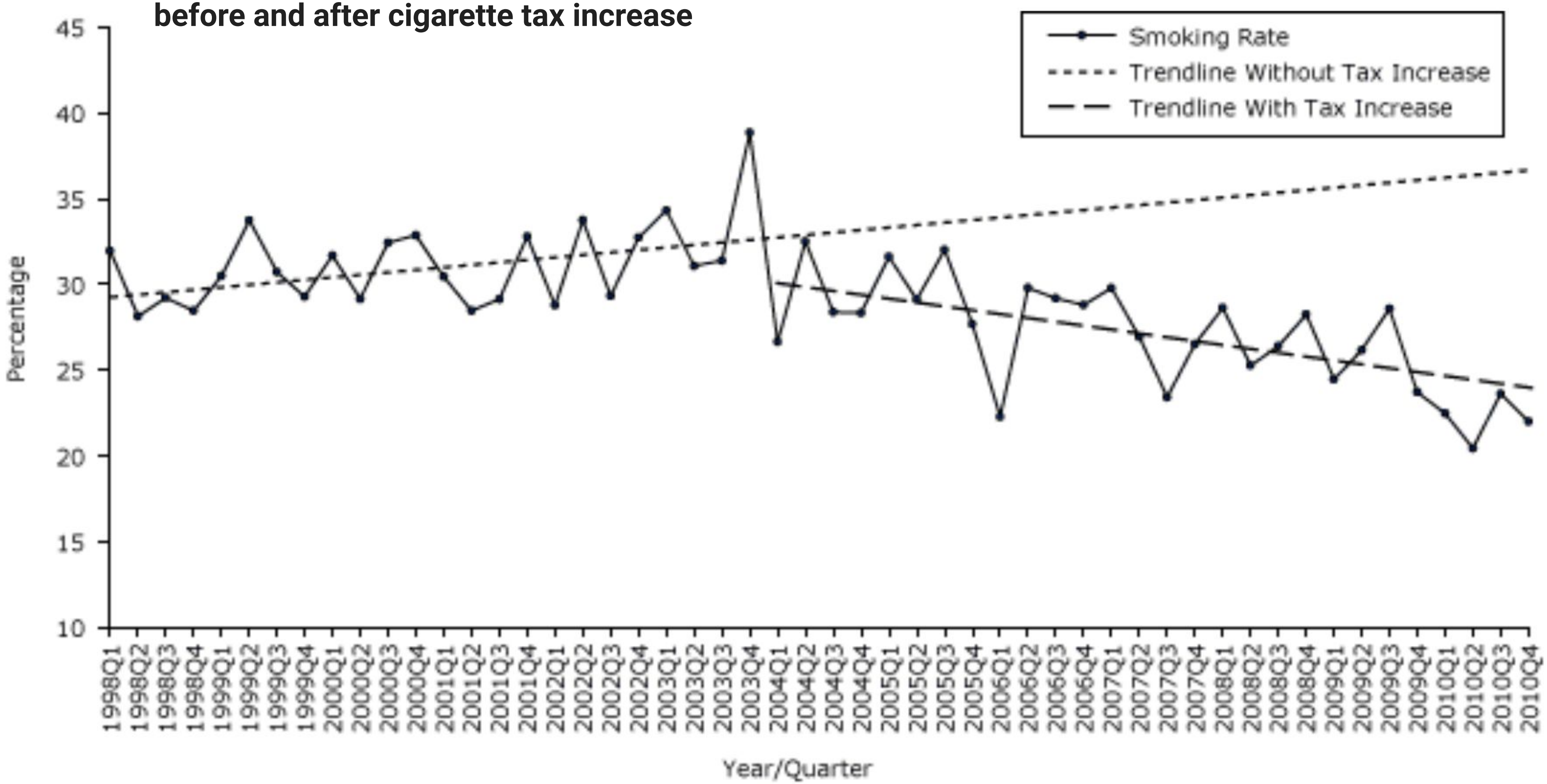
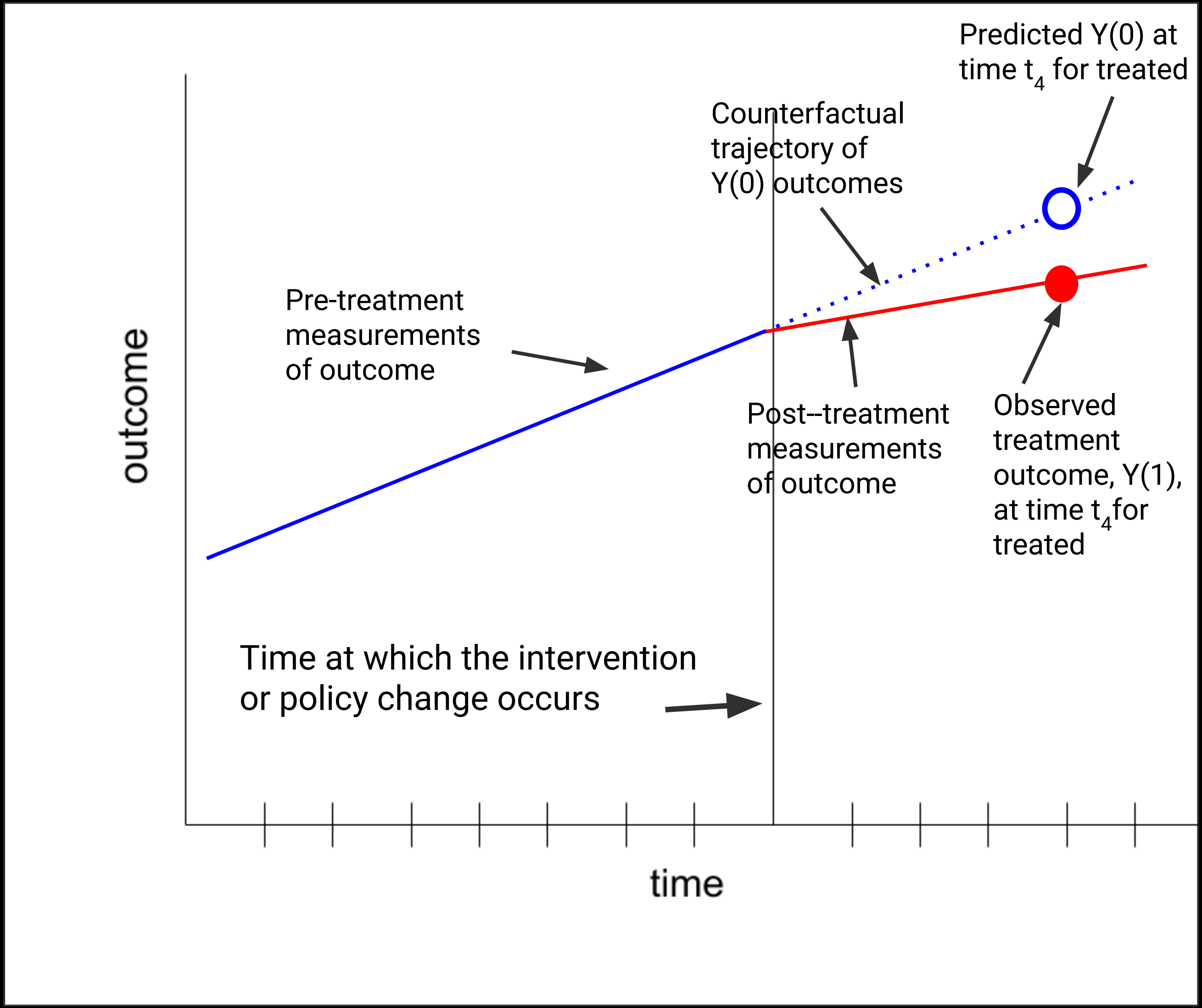
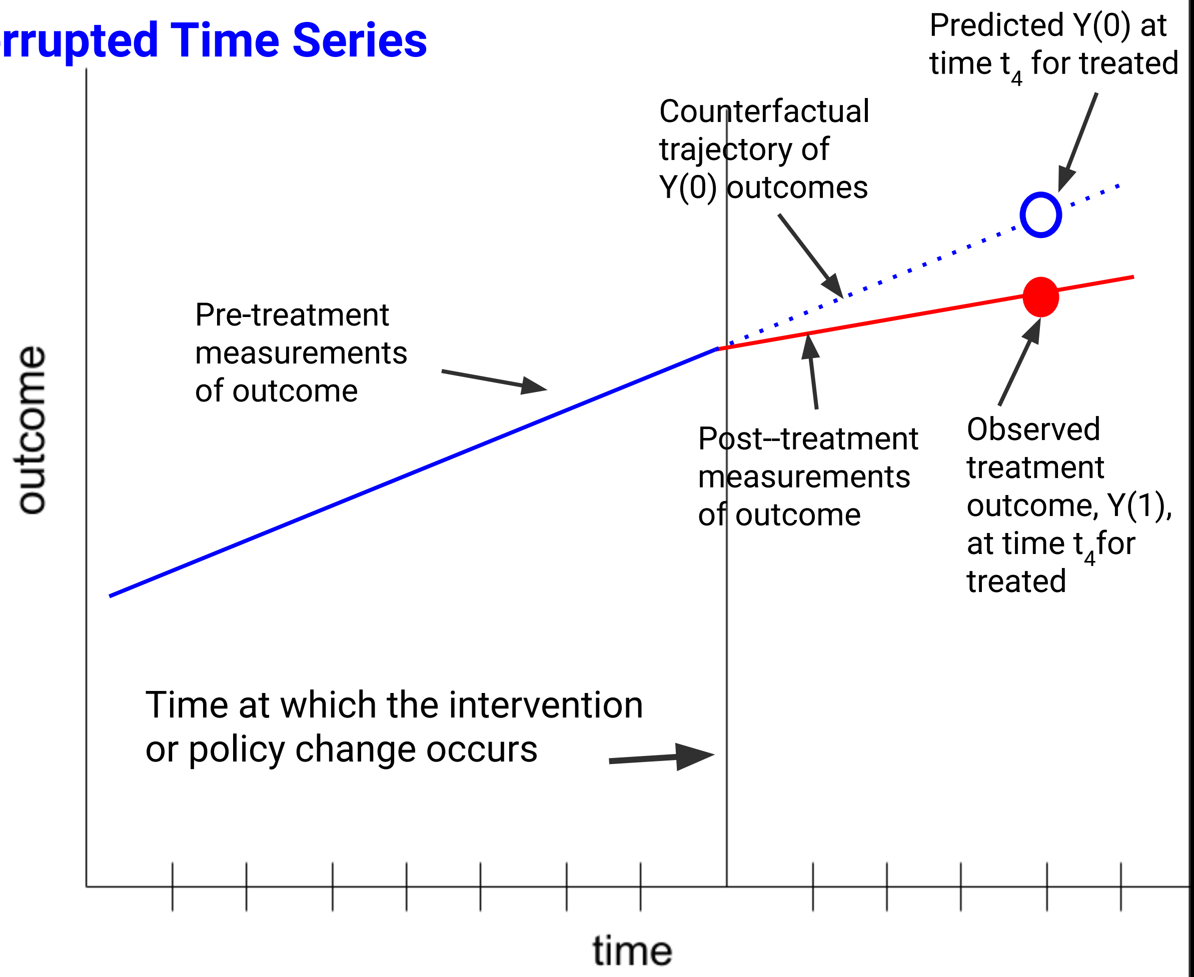


Figure 1a. Quarterly smoking prevalence for adults aged 18–39 years, Pennsylvania, 1998–2010. Source: 1998–2010 Behavioral Risk Factor Surveillance System survey data.

Interrupted Time Series



Interrupted Time Series



Typical strategy to construct a trajectory of counterfactuals

1) Model the pre-intervention trend

2) Extrapolate that model beyond the intervention timeline as displayed by the dotted blue line.

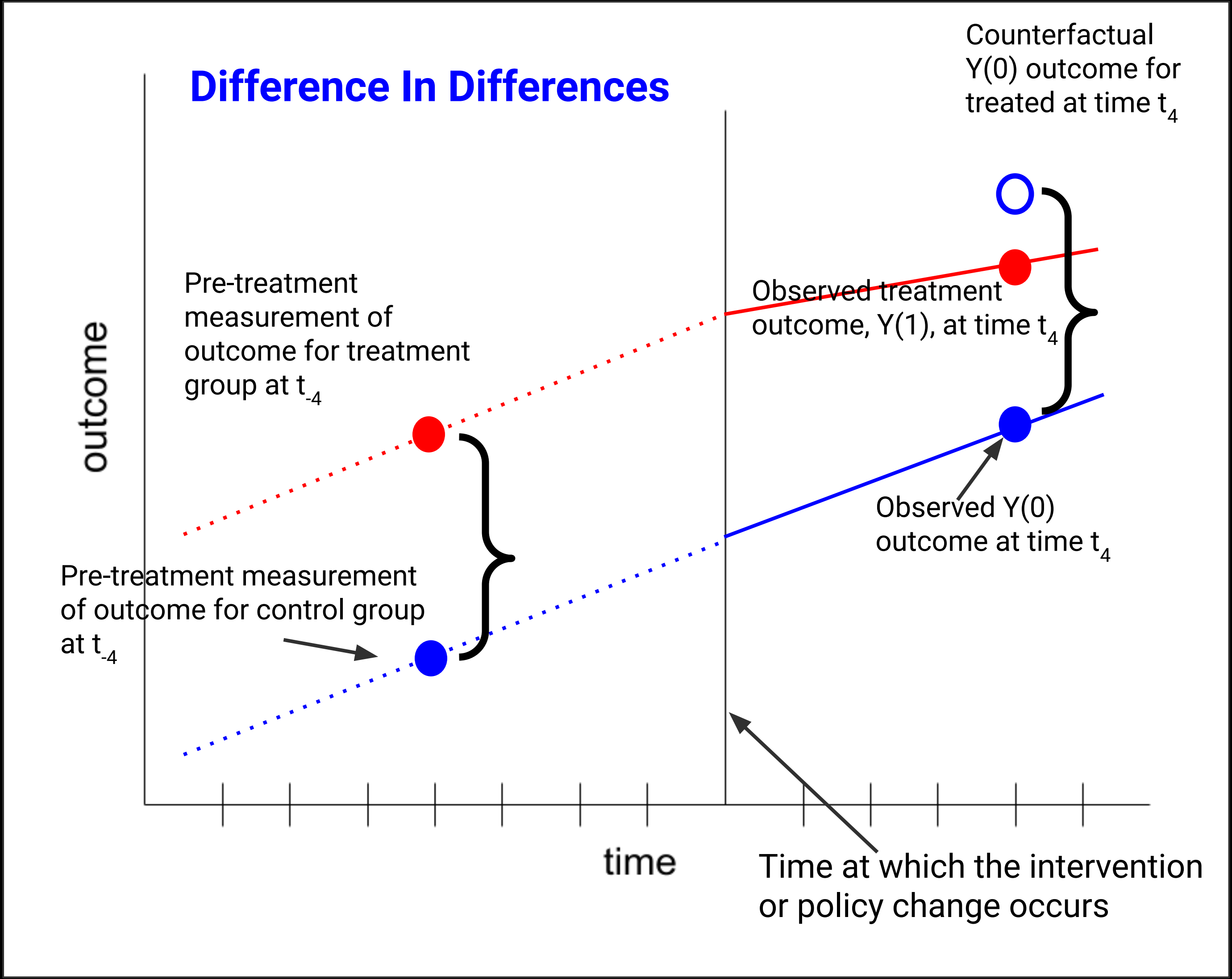
3) Estimate the treatment effect as the difference between the observed outcome for the treated and the corresponding point on the projected trend line

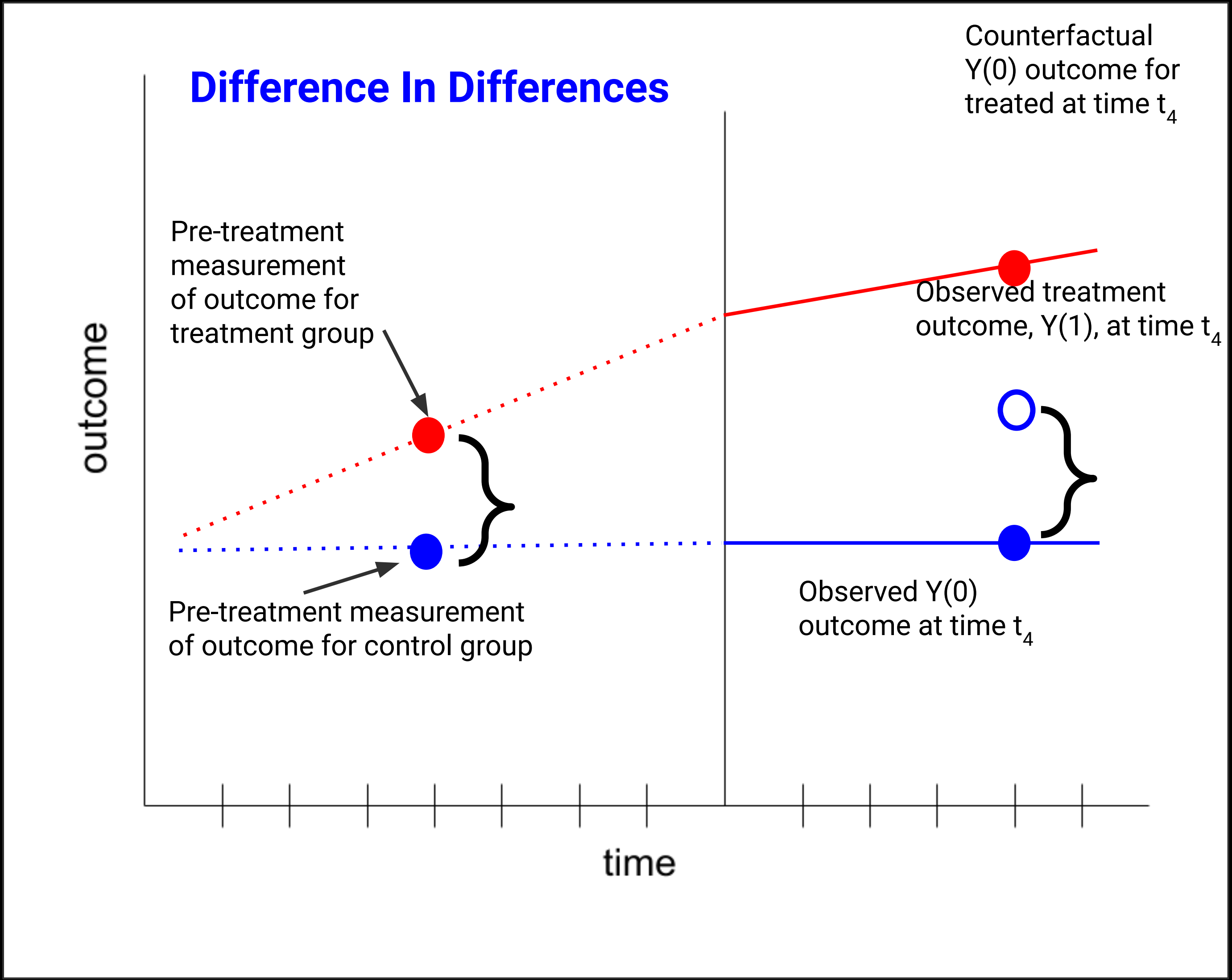
ITS versus DID

Downside of ITS is that we don’t really know how the trajectory in time might be evolving if the “treatment” (e.g. change in policy) had never occurred.

DID on the other hand uses a comparison group to make an educated guess at that trajectory.

DID


Difference In Differences
Counterfactual Y(0) outcome for treated at time t4
Pre-treatment measurement of outcome for treatment group at t-4
Observed treatment outcome, Y(1), at time t4
Observed Y(0) outcome at time t4
Pre-treatment measurement of outcome for control group at t-4
outcome
time
Time at which the intervention or policy change occurs


Difference In Differences
Counterfactual Y(0) outcome for treated at time t_4
Pre-treatment measurement of outcome for treatment group
Observed treatment outcome, Y(1), at time t_4
outcome
Pre-treatment measurement of outcome for control group
Observed Y(0) outcome at time t_4
time

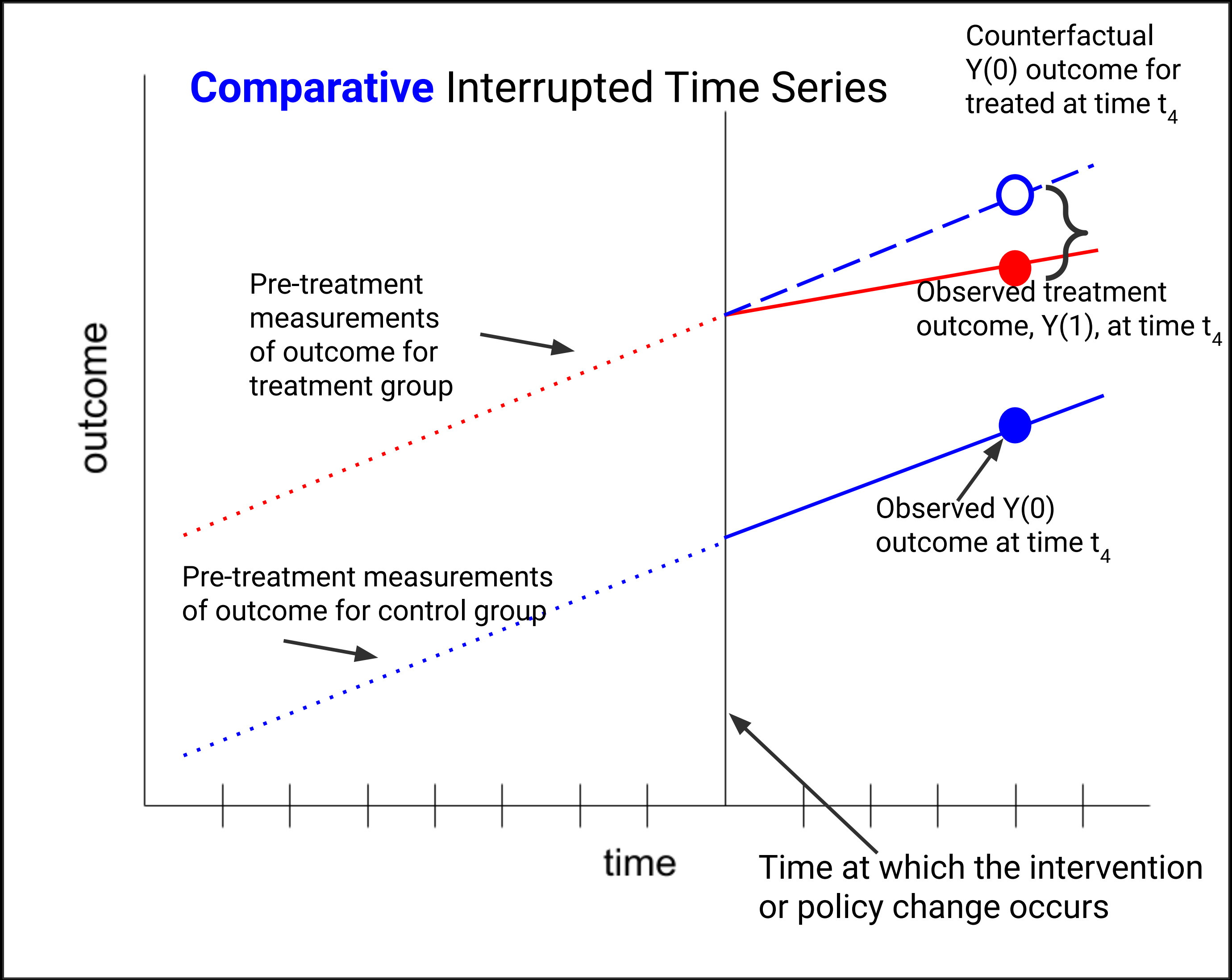
CITS (best of both worlds?)

Often framed as a more complicated version of DID, but there are important distinctions.

In CITS, the counterfactual is constructed with these steps:
1) fit linear models to the control outcome in each of the pre- and post-intervention periods,
2) compute the pre- to post-period changes in the intercepts and slopes,
3) fit a linear model to the treated outcomes in the pre-intervention period, and
4) assume the comparison group's intercept and slope changes computed in step (2) would have held in the treated group in the absence of intervention.

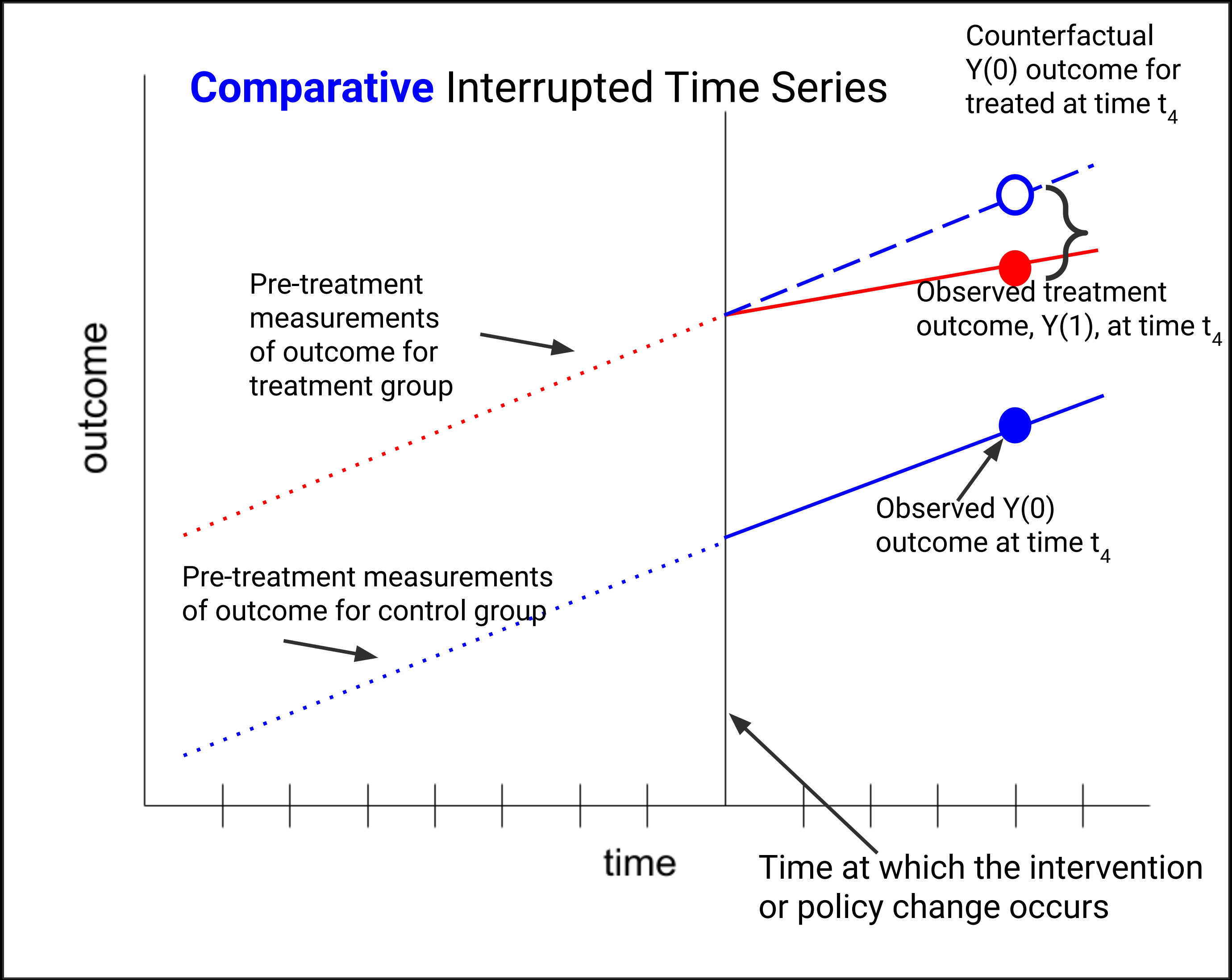
Material augmented by : https://diff.healthpolicydatascience.org/#cits

CITS



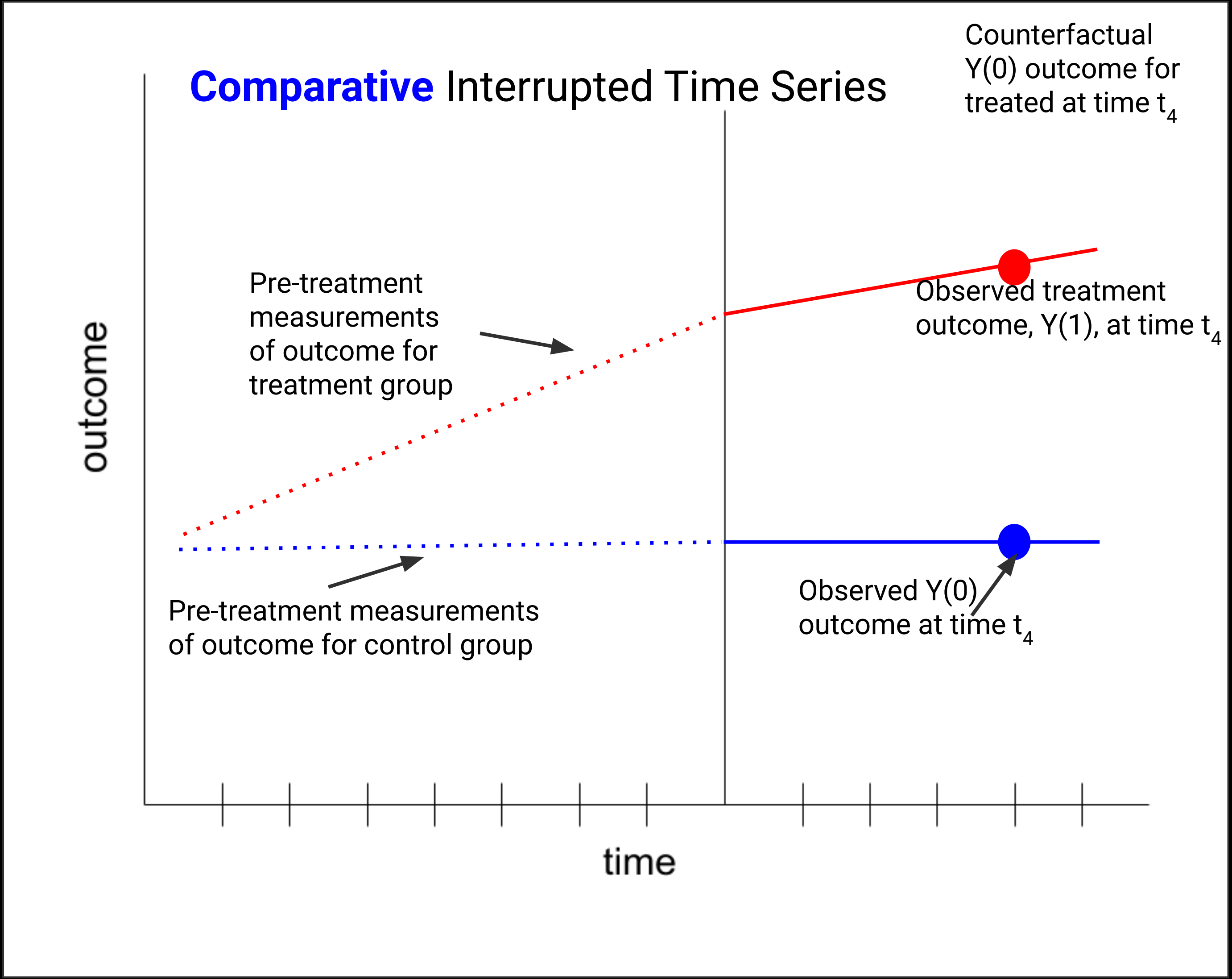
1) fit linear models to the control outcomes in each of the pre- and post-intervention periods,
2) compute the pre- to post-period changes in intercepts and slopes,
3) fit a linear model to treated outcomes in the pre-intervention period,
4) **assume comparison group's intercept and slope changes computed in step (2) would have held in the treated group in the absence of intervention.**

CITS

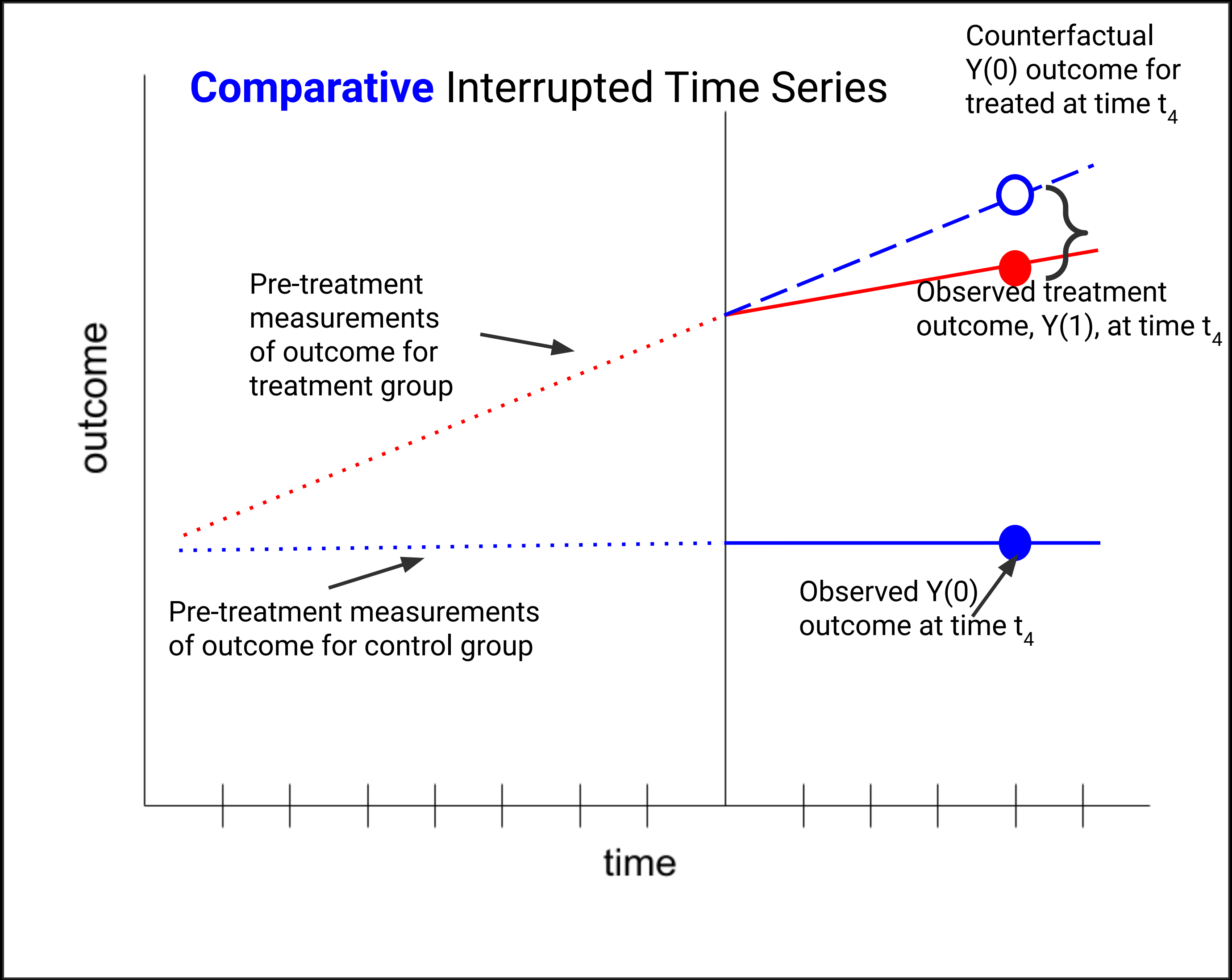


Control group doesn't change slope or intercept at the intervention time point …. so counterfactual mimics this when extrapolating the treatment line

CITS



This is a very different trajectory for the control group ..... **what would the new Y(0) look like?**



CITS

This is a very different trajectory for the control group but the story remains the same.

Control group doesn't change slope or intercept at the intervention time point …. so counterfactual mimics this when extrapolating the treatment line

ITS, DID, CITS

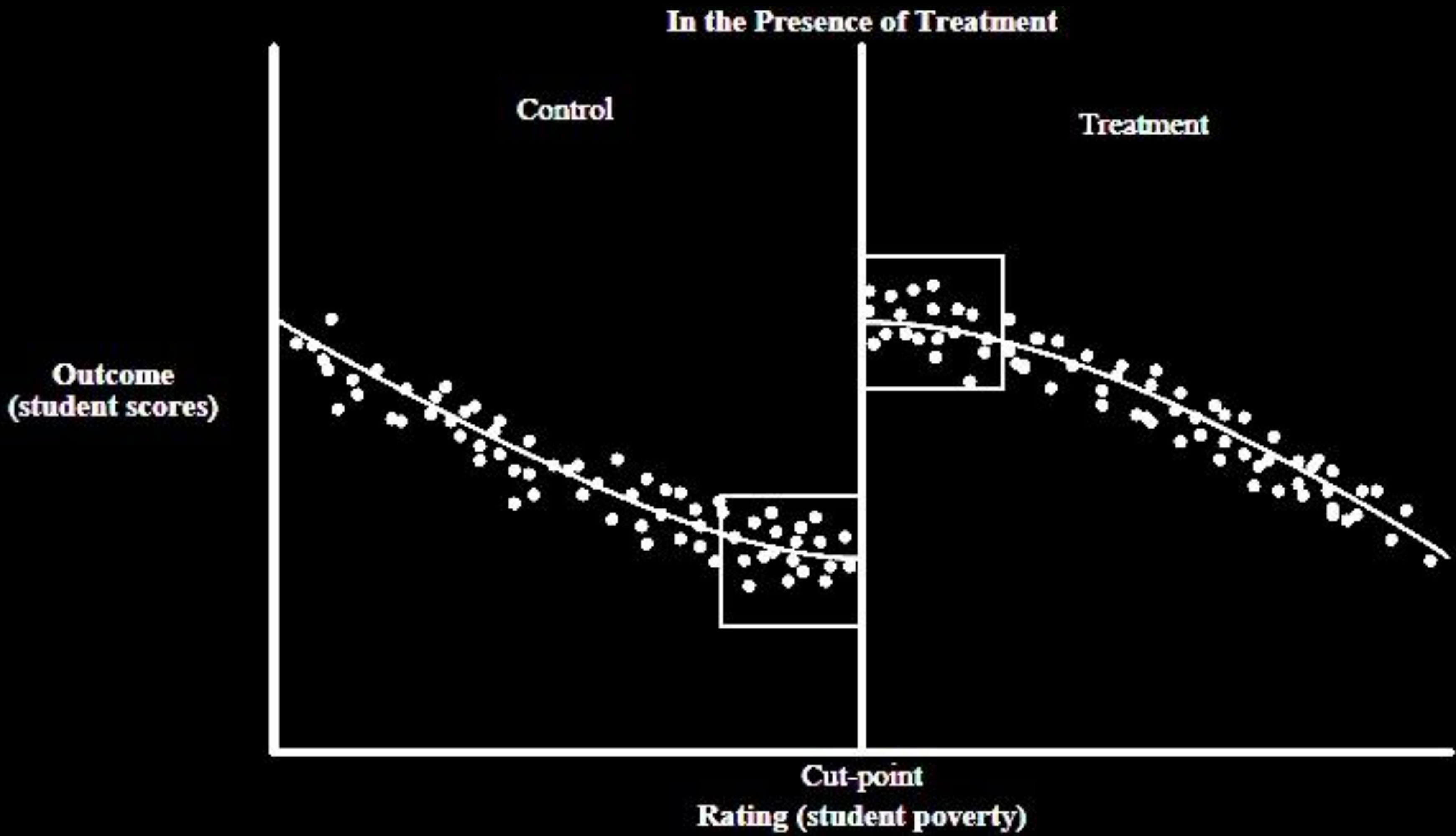
Each capitalizes on (strong, untestable) assumptions about similarities in trajectories over time.

Each is sensitive to departures from the assumptions

Generally preferable to have a comparison group (DID and CITS). The more similar that group is to the treatment group at the outset the more confidence we typically have.

Regression Discontinuity

Regression Discontinuity Design



Regression Discontinuity Design

Arbitrary cutoffs are common in practice

- Test score cutoff for winning a college scholarship
- Birth weight cutoff for sending newborn to ICU
- Program officers assessment of risk for housing program
- Income threshold for means-tested social supports

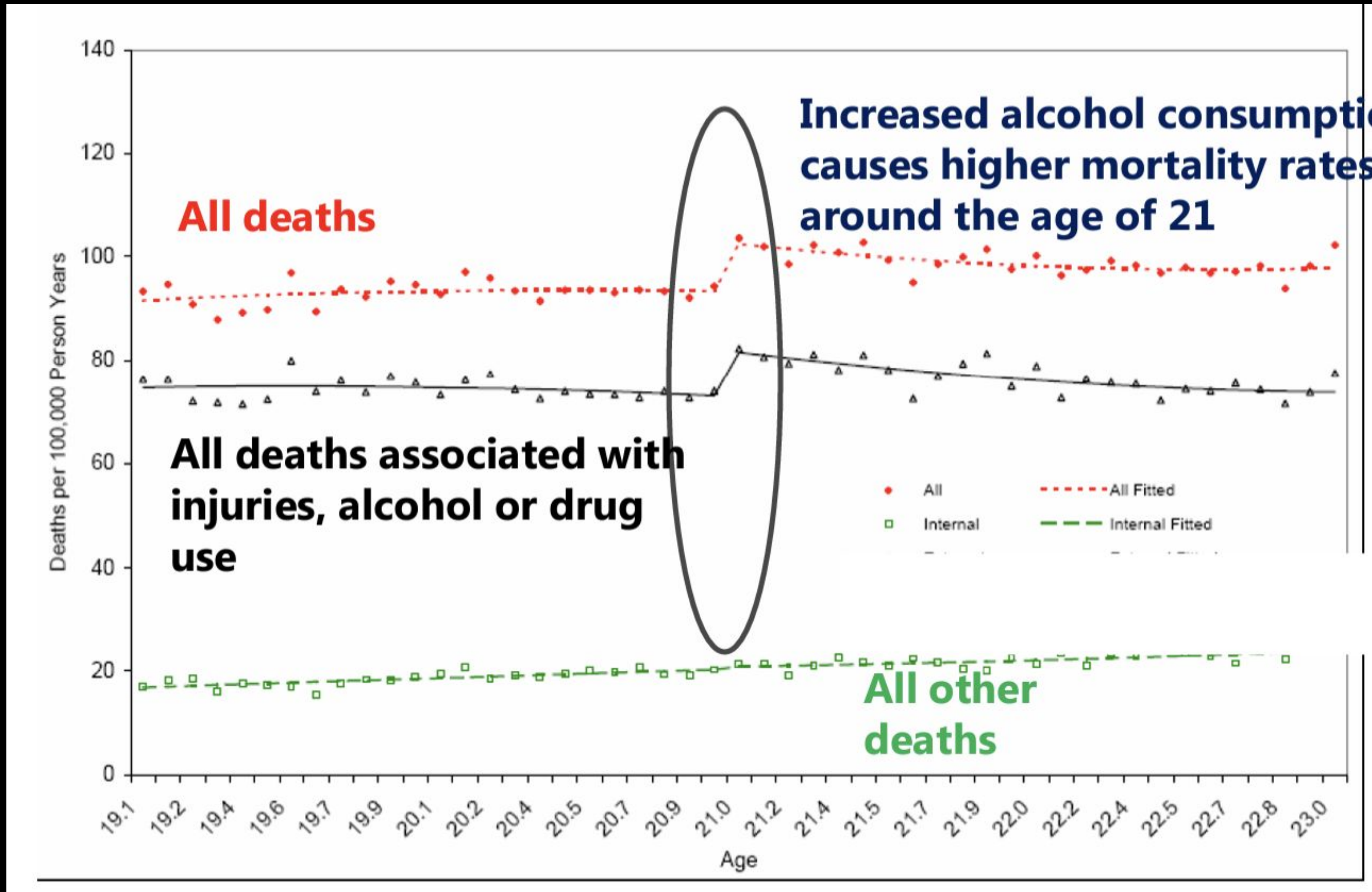
Advantages of RDD

- We know the assignment rule (which means we know the true confounders)

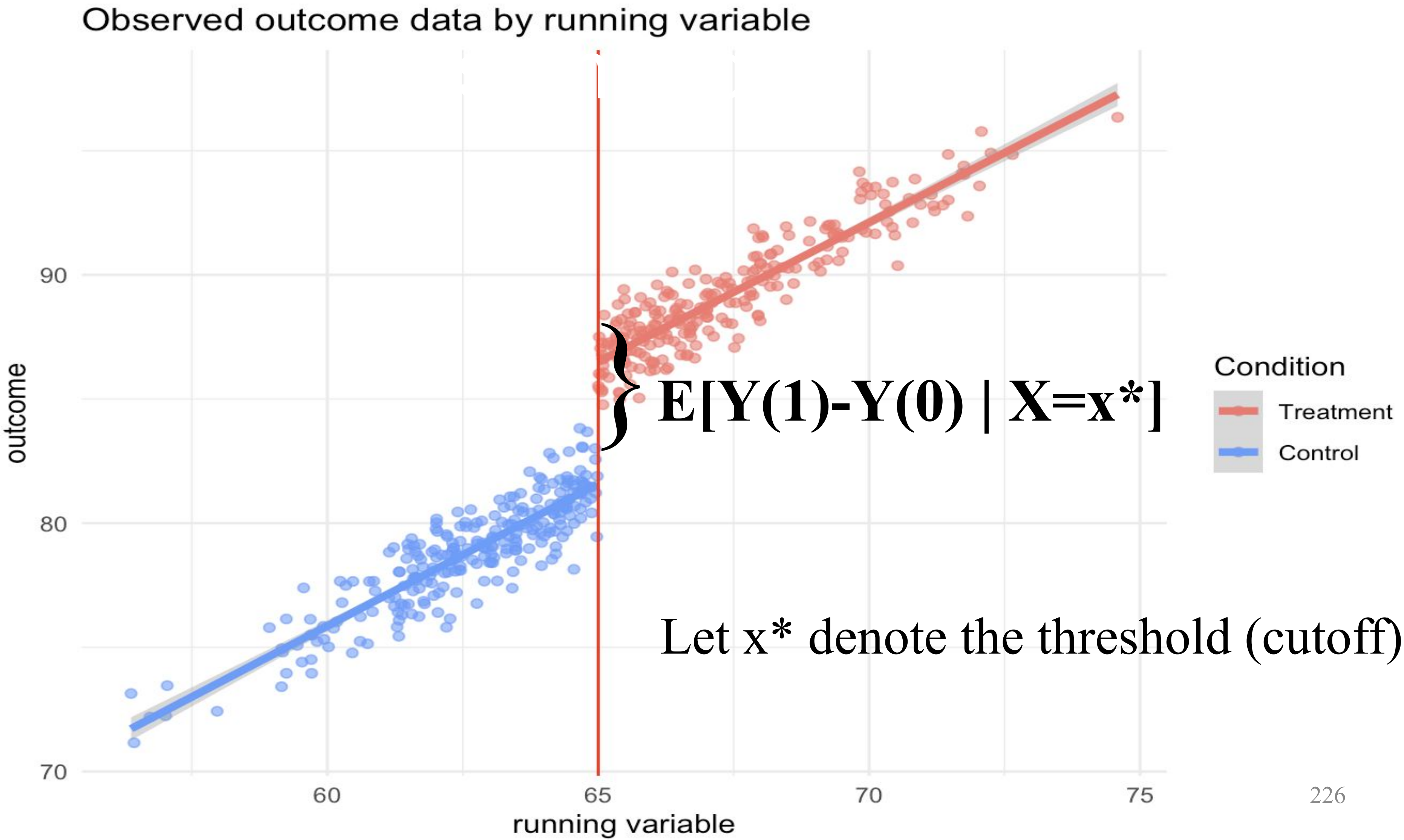
But **many statistical challenges**

- No **overlap**
- Need to estimate impacts at a boundary

Regression Discontinuity Design



Source: World Bank: 16-Technical-Track-Regression-Discontinuity.pdf



$E[Y(1)-Y(0) \mid X=x^*]$

Let x* denote the threshold (cutoff)

The RD estimator

Most popular estimators use the following models **fit to data in a selected bandwidth**

$$\mathrm{E}[Y \mid Z, X] = \beta_0 + \beta_1 X^c + \tau Z + \beta_2 X^c Z$$

$$\begin{aligned}\mathrm{E}[Y \mid Z, X] &= \gamma_0 + \gamma_1 X^c + \gamma_2 (X^c)^2 + \tau Z \\ &+ \gamma_3 X^c Z + \gamma_4 (X^c)^2 Z\end{aligned}$$

where, for simplicity, we let $X^c = X - x^*$

Y = outcome

Z = treatment assignment

X = running variable; X* = cutoff;

Regression Discontinuity Design: Ethics

Regression discontinuity is sometimes proposed as a **more ethical alternative** to a randomized experiment

If the score that determines the cutoff / treatment eligibility is a measure of "need" then it might help ensure that the most needy receive the treatment/program

Sometimes leads to unethical behavior at the threshold (artificially inflating test scores or deflating income to allow someone to be eligible)

Veil of Darkness

Using causal inference to assess discrimination in traffic stops

Understanding the causal effect of discrimination

Why is it hard to assess the impact of discrimination?

RDD to understand the impact of discrimination

Consider the following research question….

Is there a causal effect of race on the probability that a driver is pulled over by the police?

Idea 1

Compare the percent of people pulled over for traffic stops across racial groups.

Problem?

Idea 2

In essence then it would be nice to compare to a situation where the officers making the stops had no information about race… when would this happen?

Idea 2

In essence then it would be nice to compare to a situation where the officers making the stops had no information about race… when would this happen?

How about when it’s too dark to be able observe race clearly?

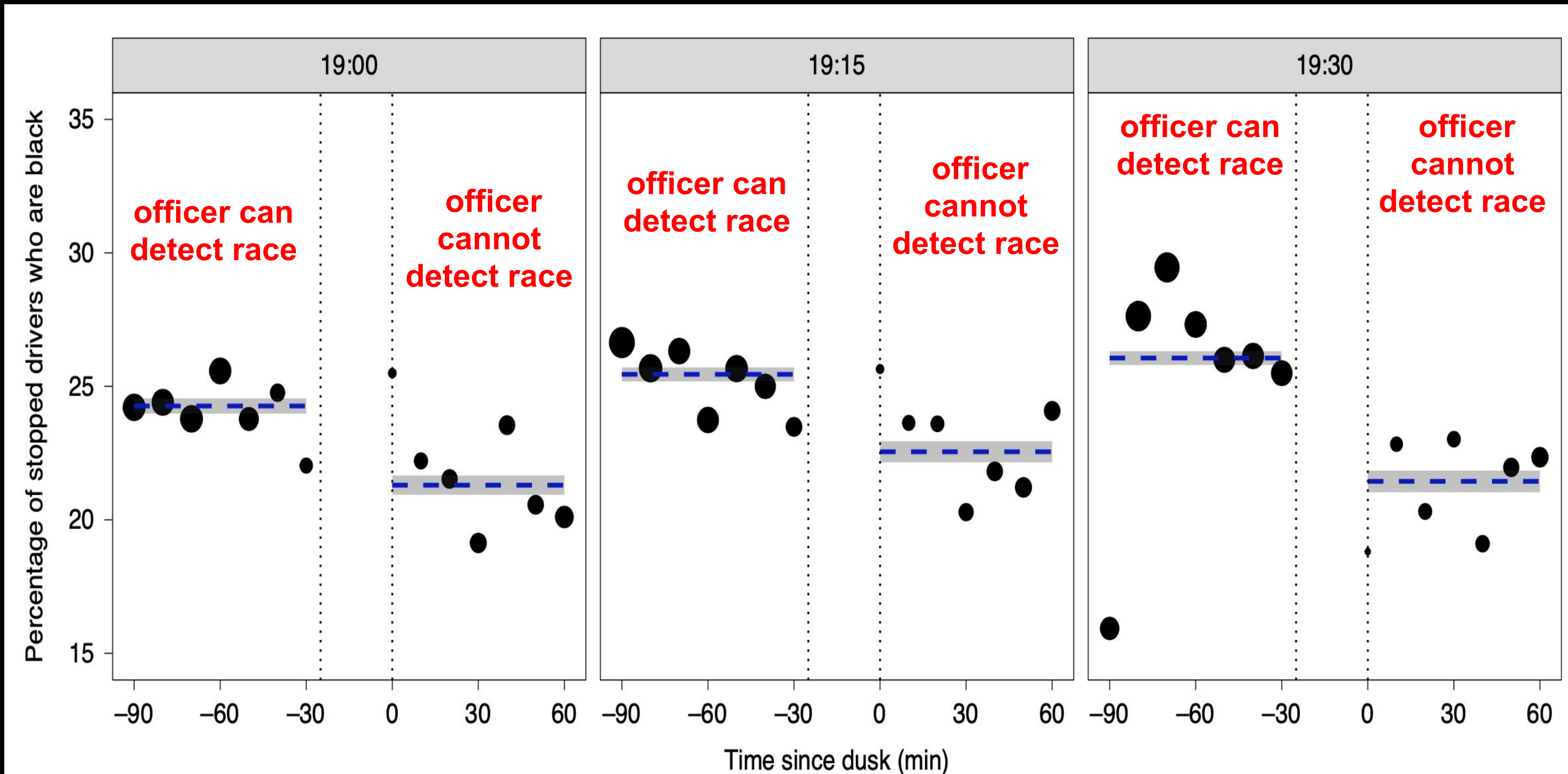
Idea 2

How about comparing stops by racial group at two different times of day:

-When it's light enough for the officer to see the driver's race

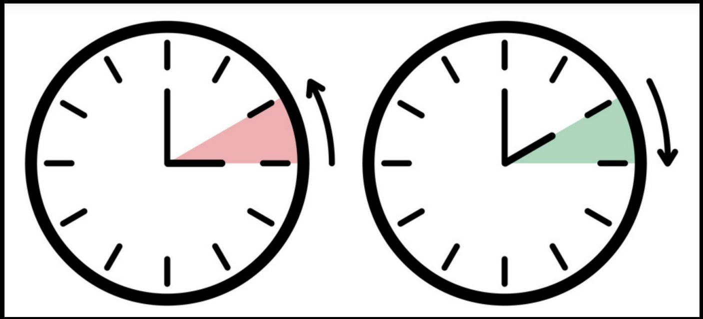
-When it's dark enough to mask the driver's race

Stops occurring in three short time windows in a single state, Texas



BIG IDEA: Daylight savings!

Daylight savings in the US creates a situation where if we make comparisons at the same time of day on the day (week) before and after the time change one one day it will be light and on the next it will be dark.



Idea 3

How about we compare stop races by group at the same time of day but across days that are separated by the time change that occurs due to daylight savings time?

Idea 3

How about we compare stop races by group at the same time of day but across days that are separated by the time change that occurs due to daylight savings time?

Sounds good!

Model

$$\Pr(\text{black} \mid t, g, p, d, s, c) = \operatorname{logit}^{-1} \left(\alpha_s \times s \times d + \alpha_c \times c \times d + \beta^T \times \text{ns}_6(t) + \gamma[g] + \delta[p]\right)$$

models the probability that a stopped driver is black at a given point in time, *t*, location, *g*, and period, *p* (start or end of daylight savings). *d* denotes after dusk or before sunset. *c* denotes city police versus state patrol, *s*. $ns_6(t)$ is a spline.

Model and results

$$\Pr(\text{black}|t, g, p, d, s, c) = \text{logit}^{-1}\left(\alpha_s \times s \times d + \alpha_c \times c \times d + \beta^T \times \text{ns}_6(t) + \gamma[g] + \delta[p]\right)$$

α_s = -.033 (-.039, -.027)
α_c = -.039 (-.045, -.022)

The representation of black drivers among those stopped decreases strongly when the officers have a more difficult time assessing the race of the driver.

This is powerful evidence of discrimination!


We're just starting to plan our evaluation. Which methods should we consider?
All of them.

Causal inference is important but tricky...

LEARN MORE!!!


CAUSALITY
SECOND EDITION
Springer Series in Statistics
Paul R. Rosenbaum
Observational Studies
Second Edition
MOSTLY HARMLESS ECONOMETRICS
An Empiricist's Companion
Analytical Methods for Social Research
Regression and Other Stories
Andrew Gelman, Jennifer Hill and Aki Vehtari
ANALYTICAL METHODS FOR SOCIAL RESEARCH
Counterfactuals and Causal Inference
Methods and Principles for Social Research
STEPHEN L. MORGAN
CHRISTOPHER WINSHIP
CAMBRIDGE
EXPLANATION IN CAUSAL INFERENCE
Methods for Mediation and Interaction
TYLER J. VANDERWEELE
OXFORD
CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES
AN INTRODUCTION
GUIDO W. IMBENS
DONALD B. RUBIN
CAUSAL INFERENCE
What If
Miguel A. Hernán and James M. Robins

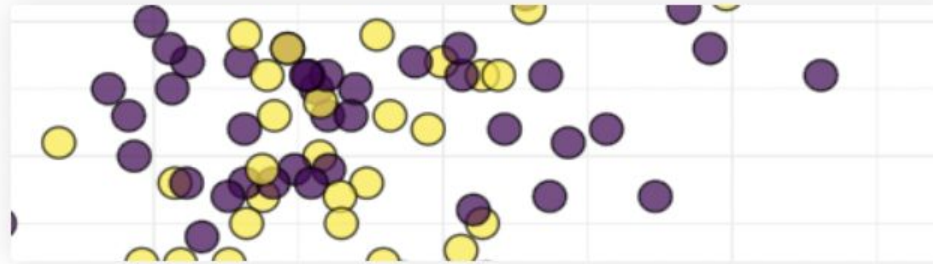
Learn Analyze Reproduce

thinkCausal

Scaffolded, user-friendly access to sophisticated causal inference tools

Educational components for "just in time" learning

apsta.shinyapps.io/thinkcausal/



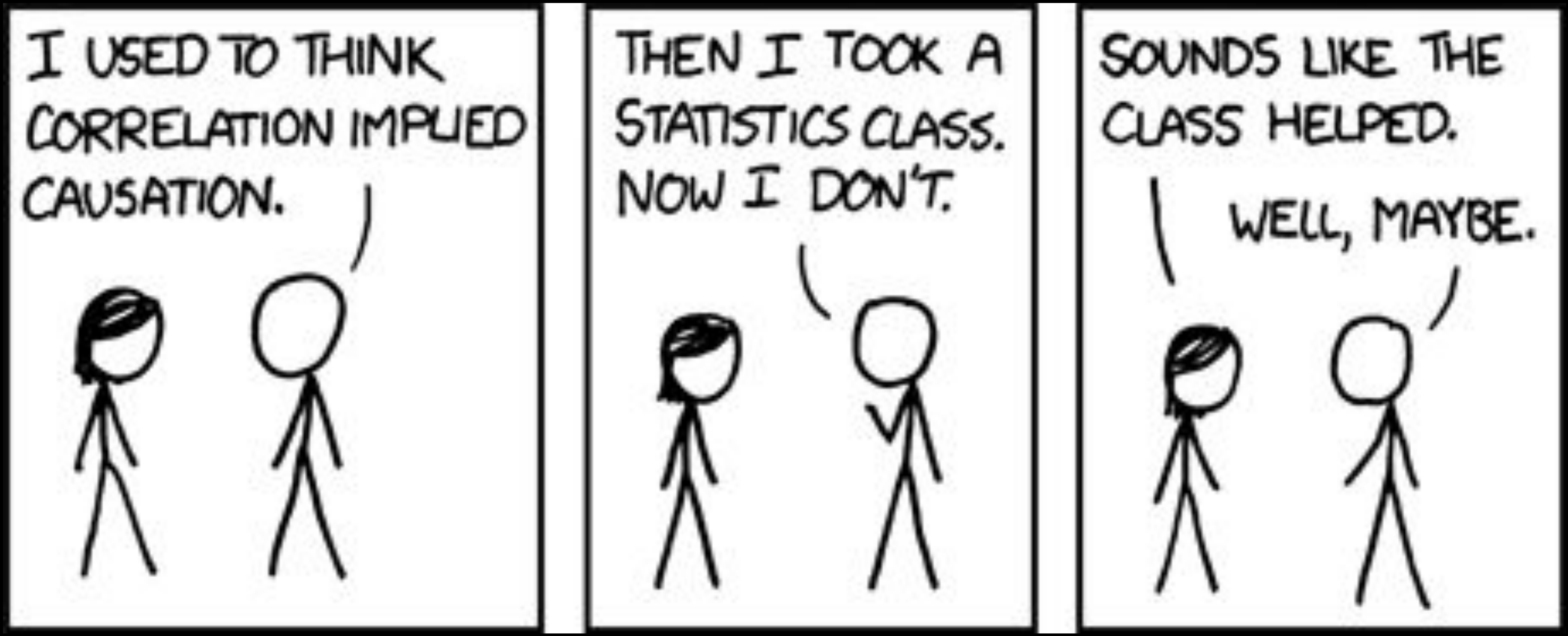
Learn

Interactively learn the foundational concepts of casual inference.

Analyze

Utilize modern causal inference methods. Easily implement Bayesian Additive Regression Trees.

Thank You!



jennifer.hill@nyu.edu