Nordic Probabilistic AI School

Causal Machine Learning

Fredrik Johansson

June 16, 2023

Healthy AI Lab www.healthyai.se

Machine learning for improved data-driven decision-making with applications in healthcare

- Decision making & causality
- Generalization & transfer learning
- Healthcare applications

Fredrik Johansson
fredrik.johansson@chalmers.se

Anton Matsson

Adam Breitholtz

Newton Mwai

Lena Stempfle

Starting
Sep 1

Ahmet Balcioglu

Starting
Sep 1

Herman Bergström

Current PhD students

Many who do ML ask

- We fit a supervised learning model—how can we **use** it?
- We have our results; how should we **interpret** them?
- If we act on our model's predictions, are our **decisions** better?

Let's help answer them!

Part I

Causal machine learning

Part II

Coding example (IncomeSim)

Part III

Reflections

IncomeSim

Simulator of causal effects of **studies** on **future income**

Based on the well-known Adult dataset from UCI

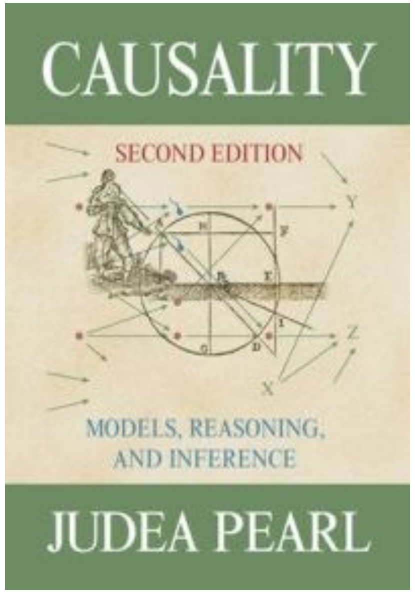
On the repository: https://github.com/Healthy-AI/IncomeSim/

You can find a link to a Colab notebook there!

We can't cover everything today!

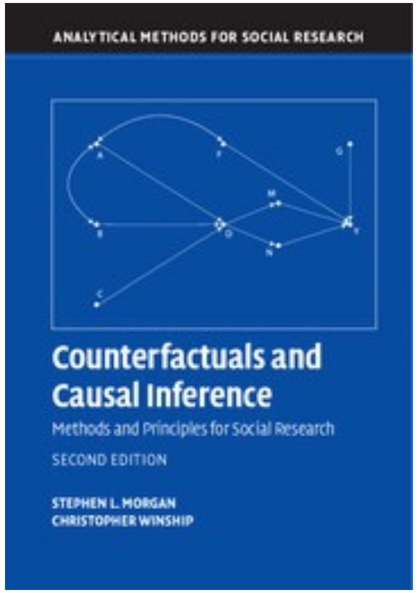
Causality

Pearl, 2009

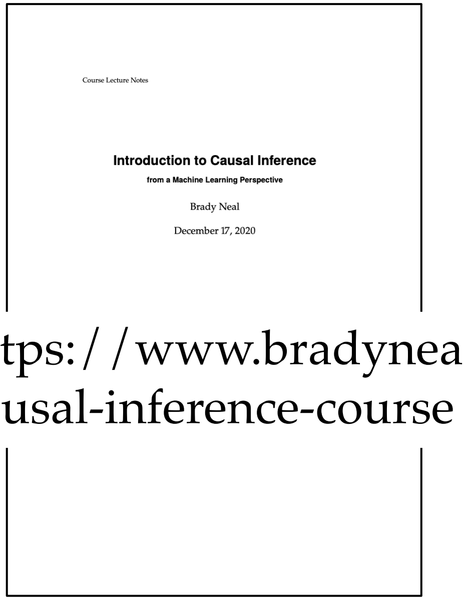


Counterfactuals and Causal Inference (2nd Ed)

Morgan & Winship, 2014


ANALYTICAL METHODS FOR SOCIAL RESEARCH
Counterfactuals and Causal Inference
Methods and Principles for Social Research
SECOND EDITION
STEPHEN L. MORGAN
CHRISTOPHER WINSHIP

Brady Neal's
Causal inference course



Course Lecture Notes

Introduction to Causal Inference

from a Machine Learning Perspective

Brady Neal

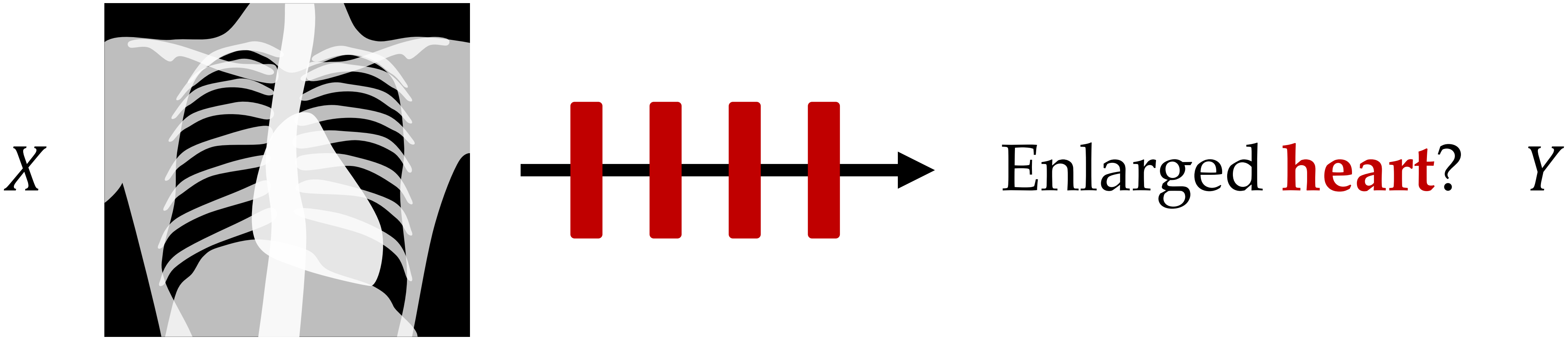
December 17, 2020

https://www.bradyneal.com/causal-inference-course

Part I: Causal machine learning

Part I.a: Supervised learning recap

ML example: Chest X-ray classification



X

Enlarged **heart**? Y

How do we learn?

One of the most common* learning principles for selecting hypotheses is **risk minimization**

$$\min_{h \in \mathcal{H}} R(h) \quad \text{where} \quad R(h) \coloneqq \mathbb{E}_{X,Y \sim p}[L(h(X), Y)]$$

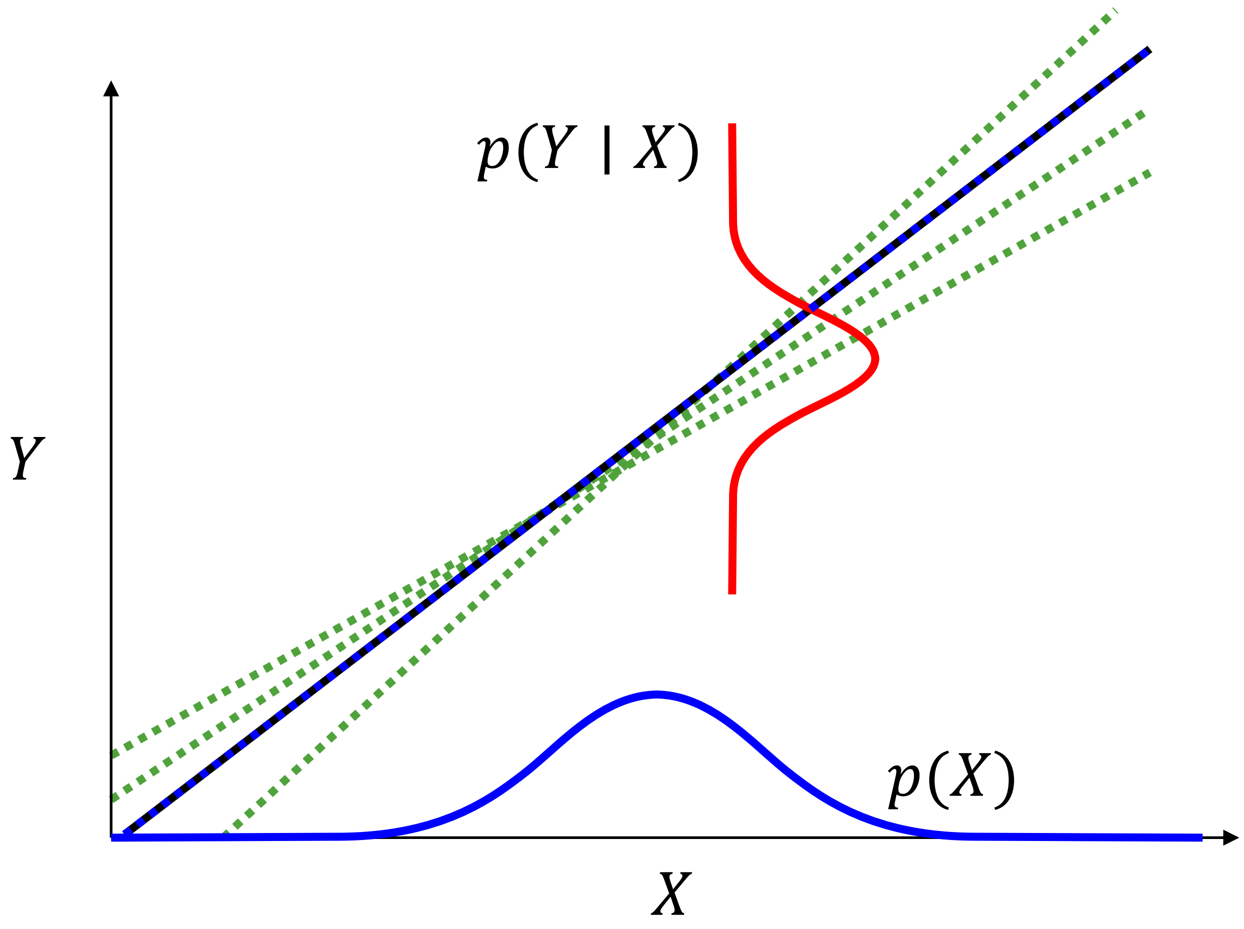
R is called "risk" or "generalization error"

and $L : Y \times Y \to \mathbb{R}_+$ is a **loss function**

For example, **mean squared error**: $R(h) \coloneqq \mathbb{E}_{X,Y \sim p}[(h(X) - Y)^2]$

* Another example is maximum likelihood estimation. These are *sometimes* equivalent

Example: Least-squares linear regression



If $Y = kx + b + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$

and

$$h^*(x) = \arg\min_h \mathbb{E}_{X,Y}[(h(X) - Y)^2]$$

then

$$h^*(x) = \mathbb{E}[Y \mid X = x] = kx + b$$

The best possible model is the conditional expectation $\mathbb{E}[Y \mid x]$

Empirical risk minimization (ERM)

The "training error"

Since the distribution of inputs and labels is unknown,
instead of minimizing R, we minimize the **empirical risk** $\hat{R}(h)$

$$\min_{h\in\mathcal{H}} \hat{R}(h) \quad \text{where} \quad \hat{R}(h) := \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i)$$

Find the hypothesis $\hat{h}$ which fits **observed examples** the best

Sample splitting for evaluation

Almost any ML course will teach sample splitting for evaluation

Example: Fit data to 80% of your data, evaluate on unseen 20%

This allows to estimate the generalization error using a test set D_{te}

$$R(h) = \mathbb{E}_{\boldsymbol{D}_{\mathrm{te}}}\left[\frac{1}{n_{\mathrm{te}}} \sum_{(x_i, y_i) \in \boldsymbol{D}_{\mathrm{te}}} L(h(x_i), y_i)\right] \approx \frac{1}{n_{\mathrm{te}}} \sum_{(x_i, y_i) \in D_{\mathrm{te}}} L(h(x_i), y_i) = \hat{R}(h)$$

The test error tells us **something** about our model’s performance

Is this answering the right questions?

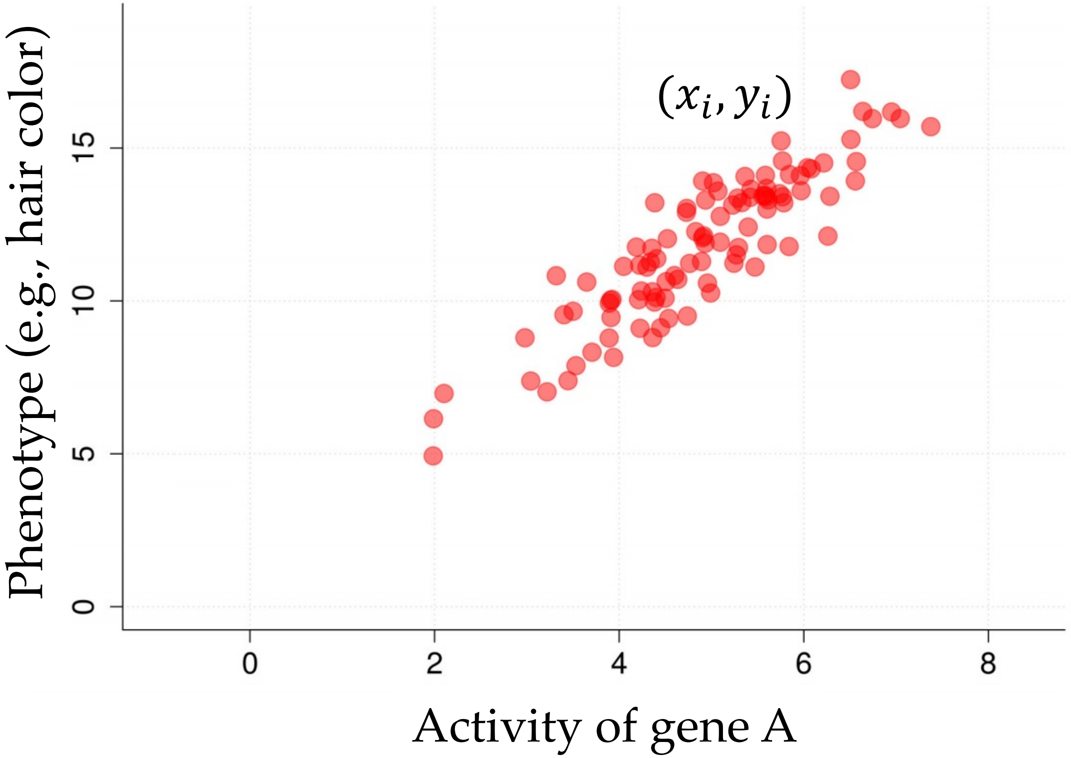
- We fit a supervised learning model—how can we **use** it?
- We have our results; how should we **interpret** them?
- If we act on our model's predictions, are our **decisions** better?

Are these answered by the test error?

Does smaller test error imply better decisions?

Part I.b: Machine learning & causality

Example: Gene knock out



Genetic data

Assume we have sequenced the **genome** of 100 individuals drawn from a population p

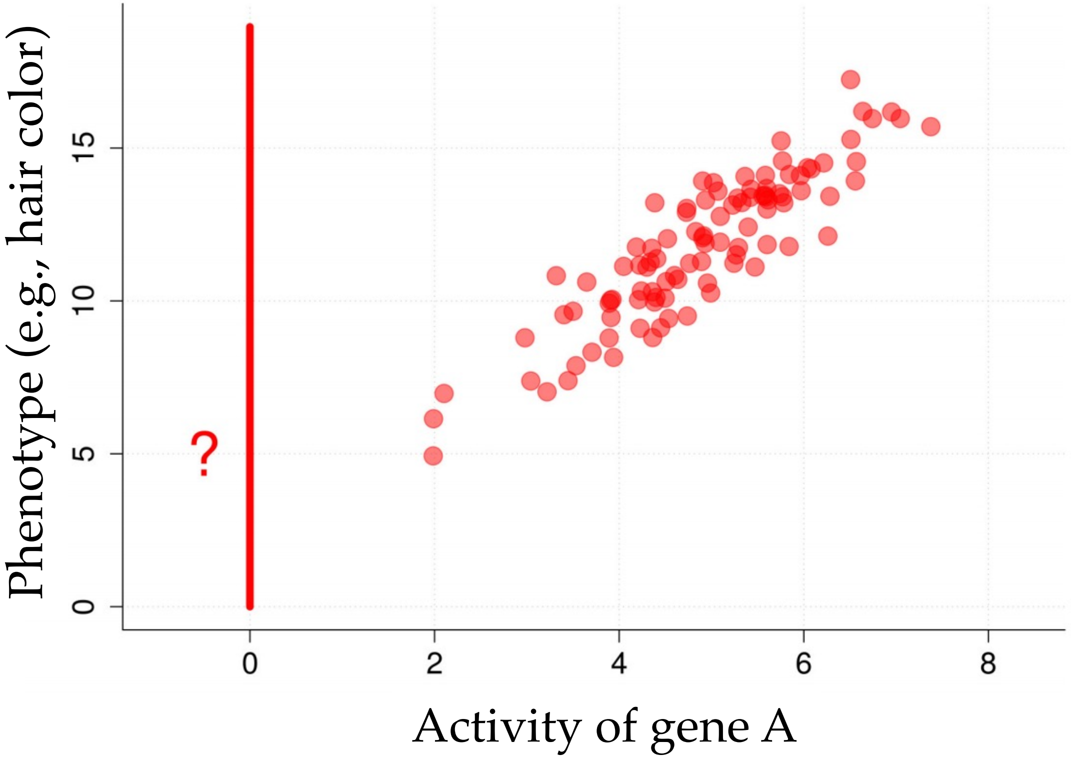
We are interested in the relationship between Gene A (X) and the phenotype **hair color** (Y)

Can we influence hair color?

Example from Jonas Peters.

Example: Gene knock out

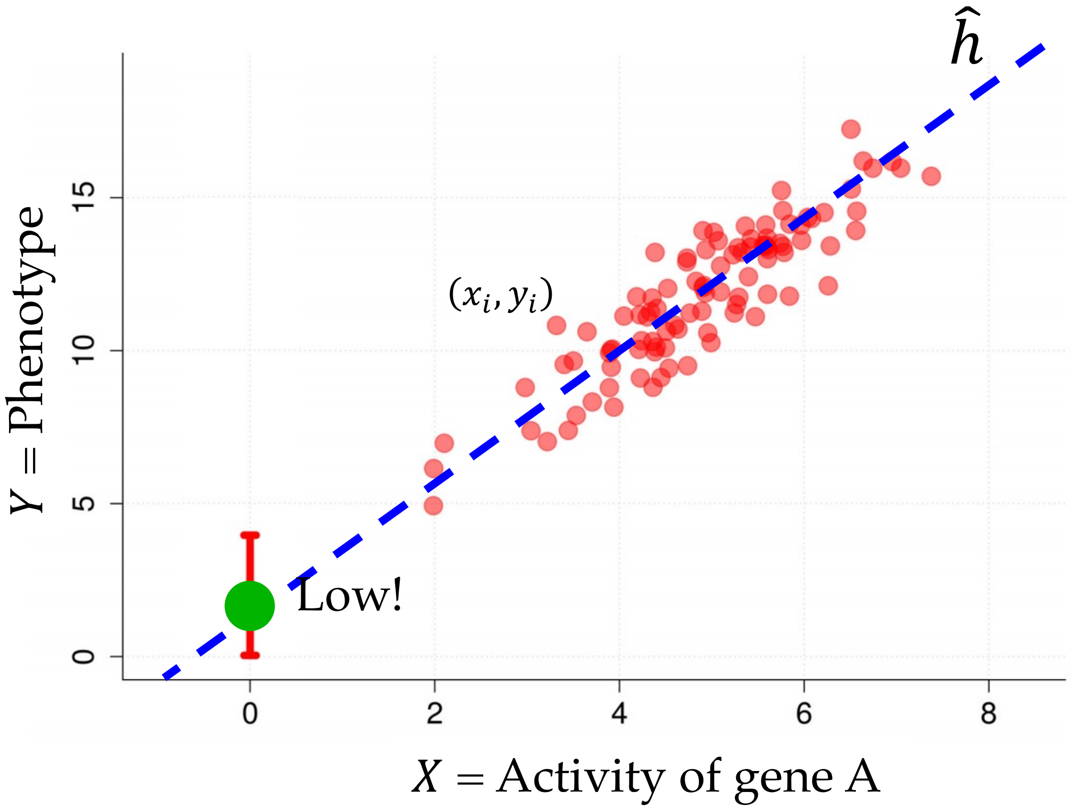
What if we "knock out" Gene A (**intervene** by setting its activity to 0)?



Question?
What do you think would be the value of the phenotype?

Example from Jonas Peters.

Out-of-the-box ML solution



Step 2.
Pick a learning principle

$$h^* = \arg\min_h \mathbb{E}_p[(h(X) - Y)^2]$$

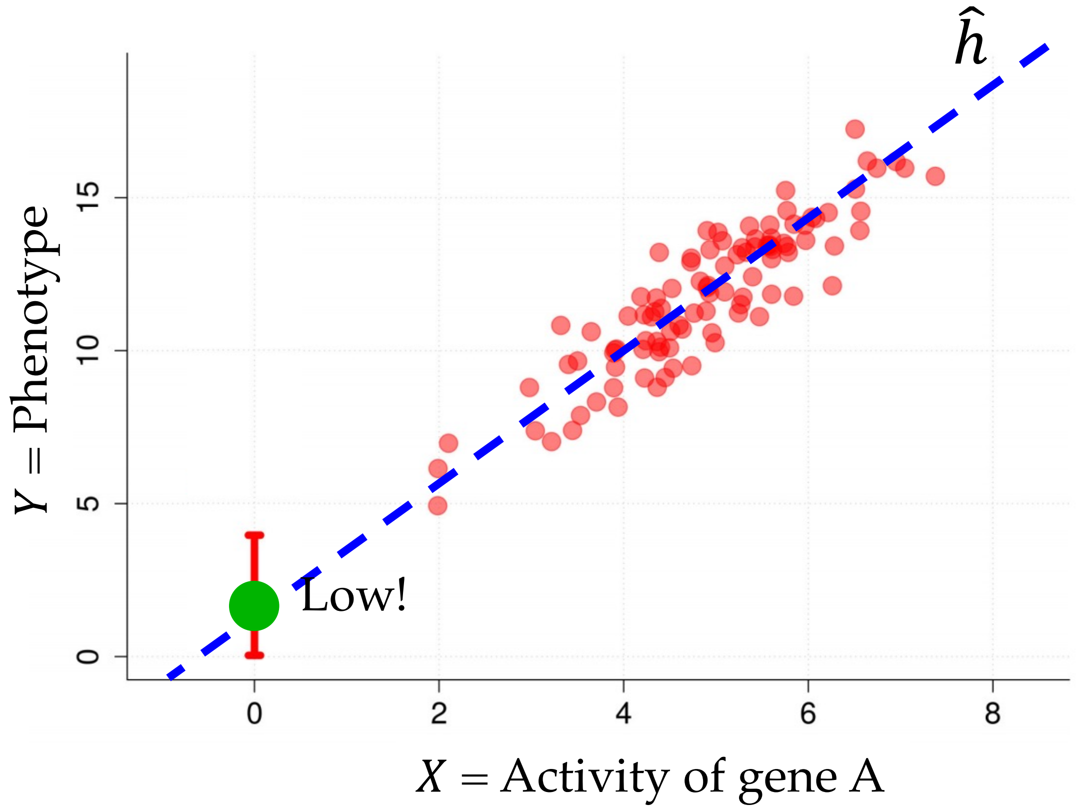
Step 3.
Learn a hypothesis from data

$$\hat{h} = \arg\min_h \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)^2$$

Step 4.
Evaluate test error. **Good fit!**

Example from Jonas Peters.

ML which regresses Y on X would return something like this



Question:
Is this correct?
How do you know?

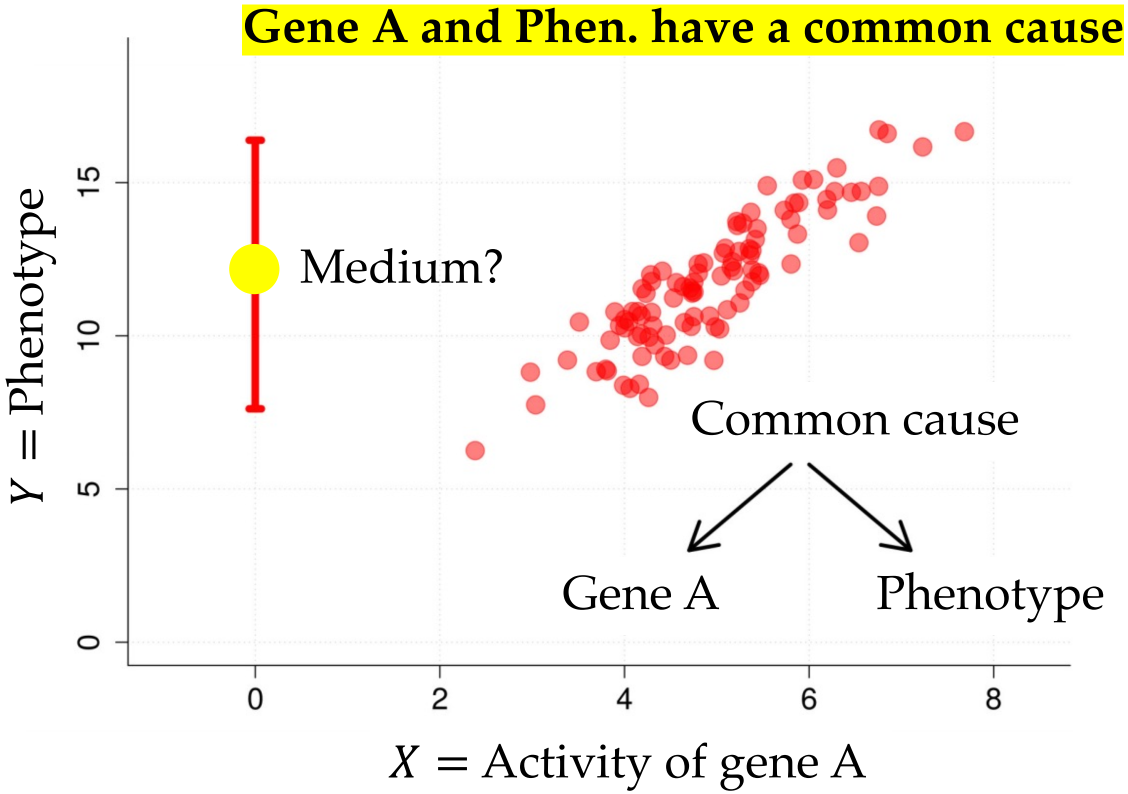
Example from Jonas Peters.

Explanation B

If there is a **common cause** which influences both gene activity X and phenotype Y

… the phenotype may not change when we knoch out Gene A!

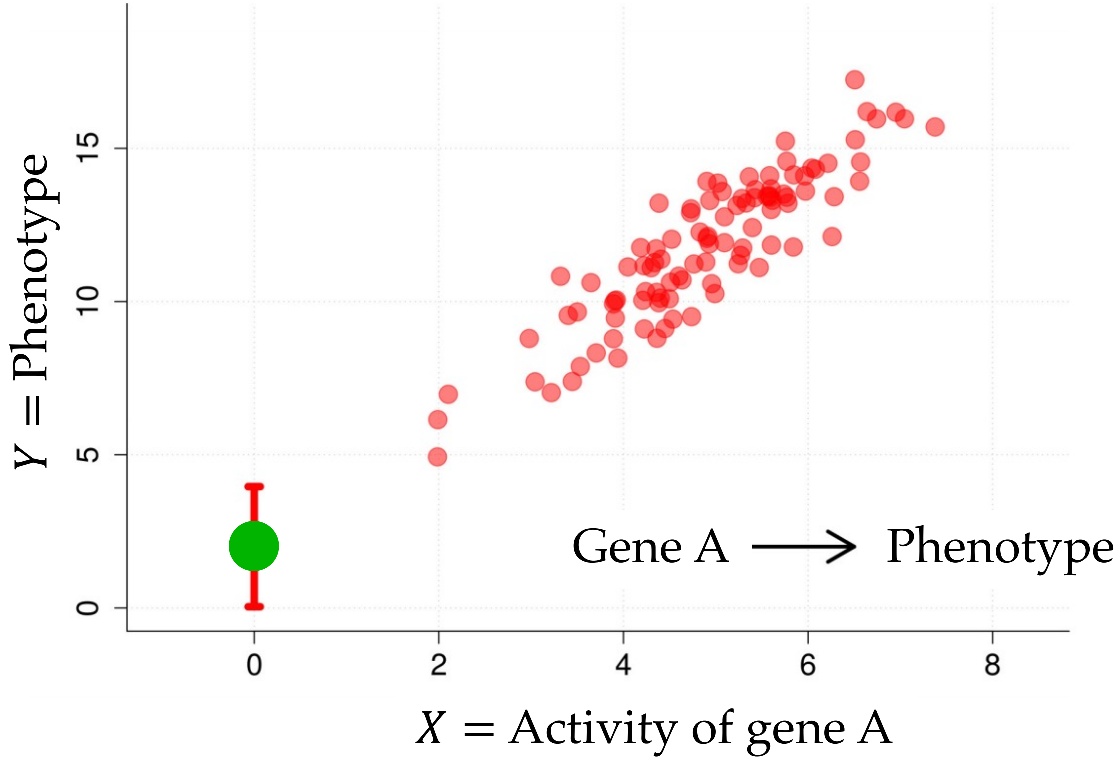
The association is **confounded**!



Example from Jonas Peters.

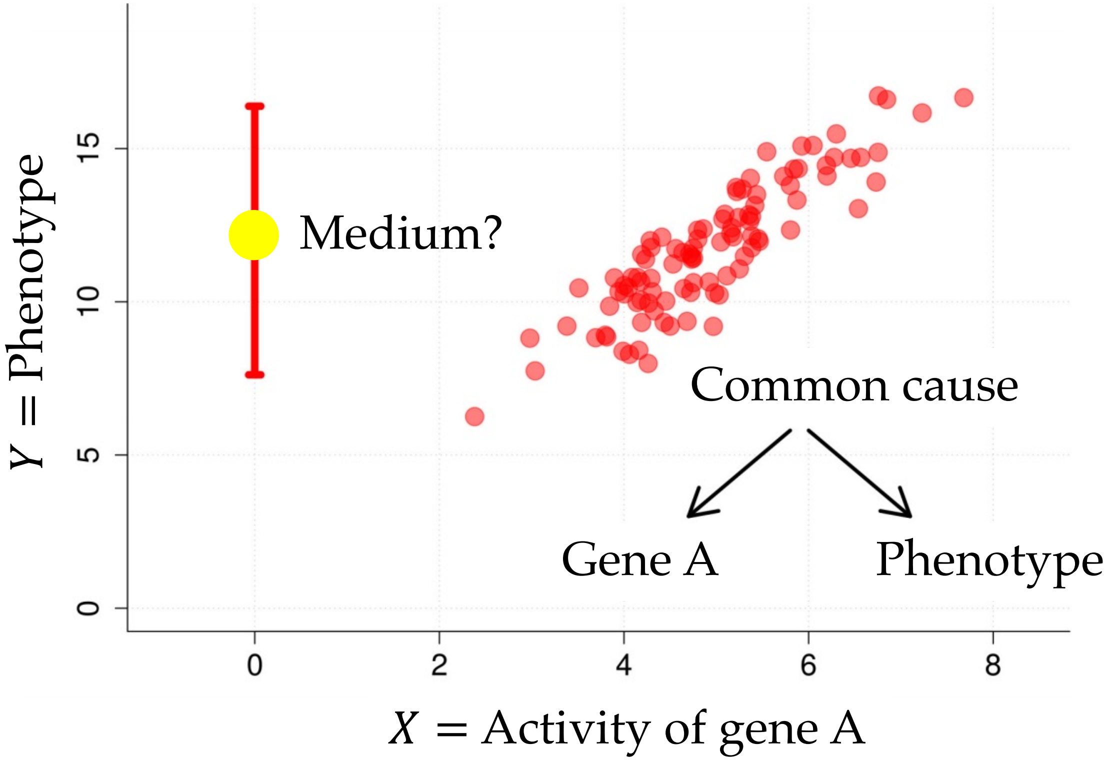
Explanation A

Gene A causes phenotype



Explanation B

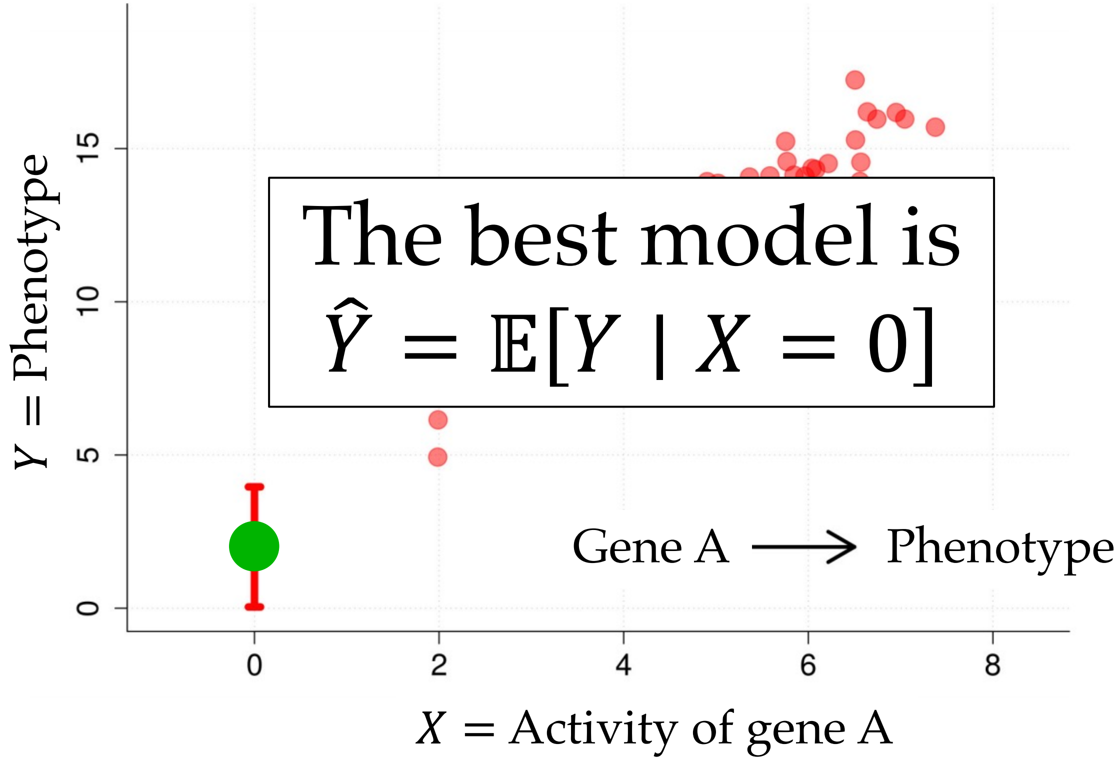
Gene A and Phen. have a common cause



Example from Jonas Peters

Explanation A

Gene A causes phenotype

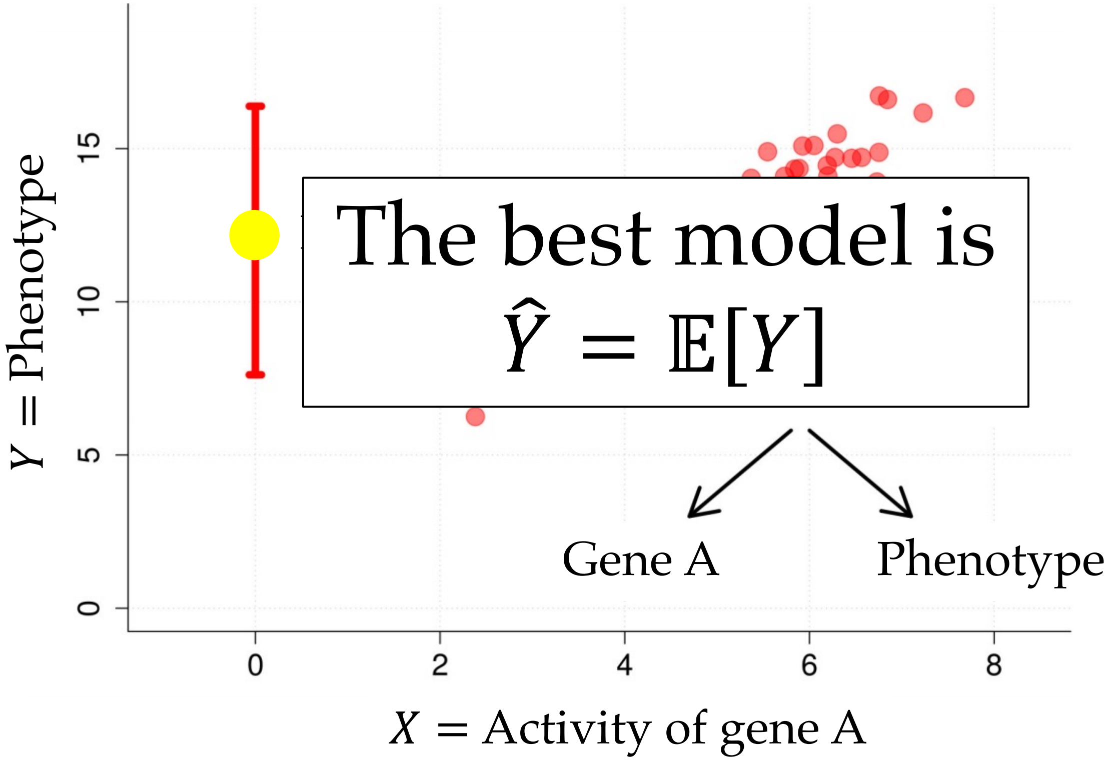


The best model is

$$\hat{Y} = \mathbb{E}[Y \mid X = 0]$$

Explanation B

Gene A and Phen. have a common cause



The best model is

$$\hat{Y} = \mathbb{E}[Y]$$

Example from Jonas Peters

Causation ≠ association

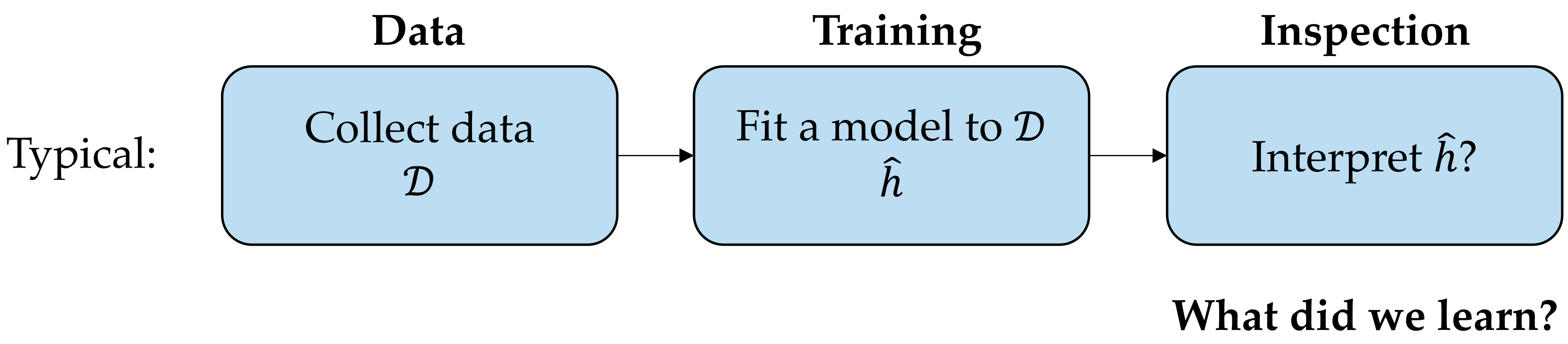
The two scenarios are equally plausible given *only* the data,
but *the right learning task* depends on causality

$$\mathbb{E}[Y \mid X = 0] \neq \mathbb{E}[Y \mid do(X = 0)] \neq \mathbb{E}[Y] \qquad \text{in general}$$

This is not a question of learning (or data)!

Even a perfect regression $f(x) \approx y$ will get the wrong answer
and the right model may have higher "test error"!

A common ML pattern…



Data

Training

Inspection

Typical:

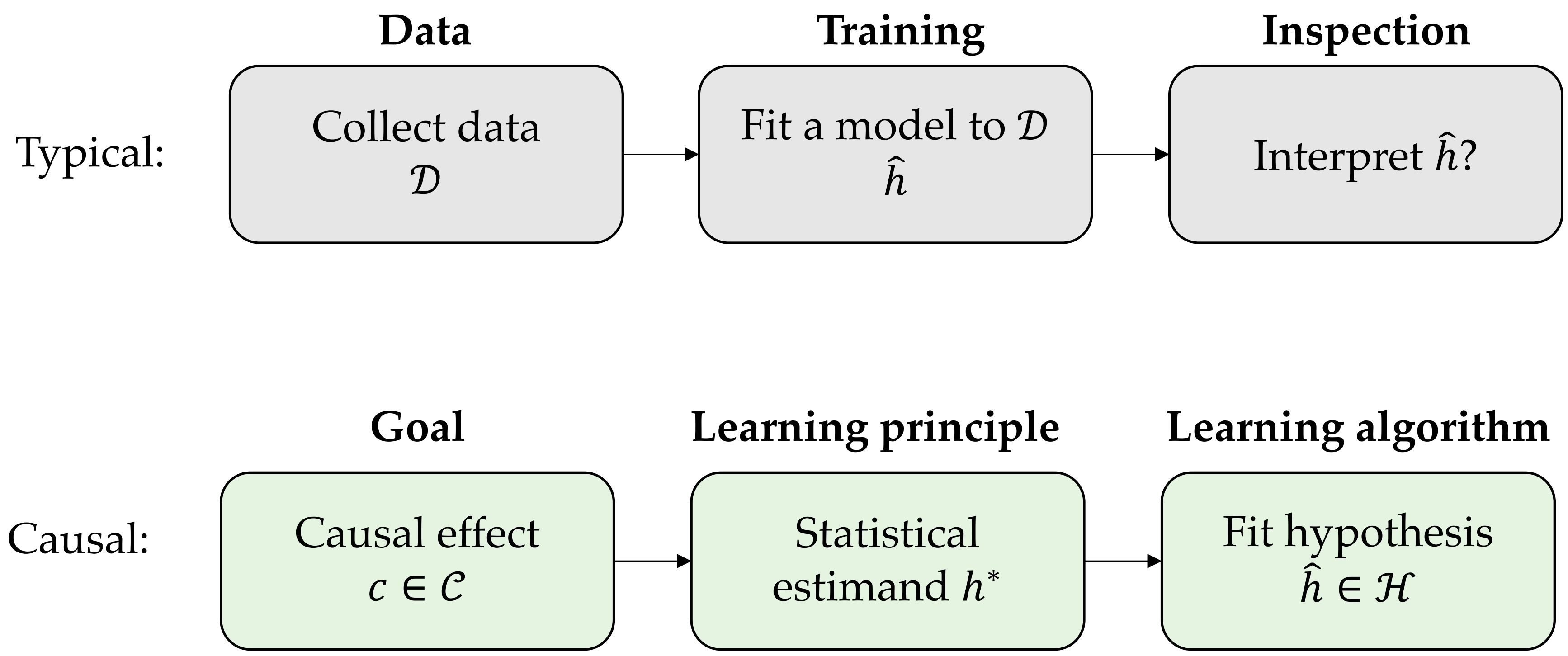
Collect data
$\mathcal{D}$

Fit a model to $\mathcal{D}$
$\hat{h}$

Interpret $\hat{h}$?

What did we learn?

Causal machine learning should be different!



Typical:

Data

Collect data $\mathcal{D}$

Training

Fit a model to $\mathcal{D}$ $\hat{h}$

Inspection

Interpret $\hat{h}$?

Causal:

Goal

Causal effect $c \in \mathcal{C}$

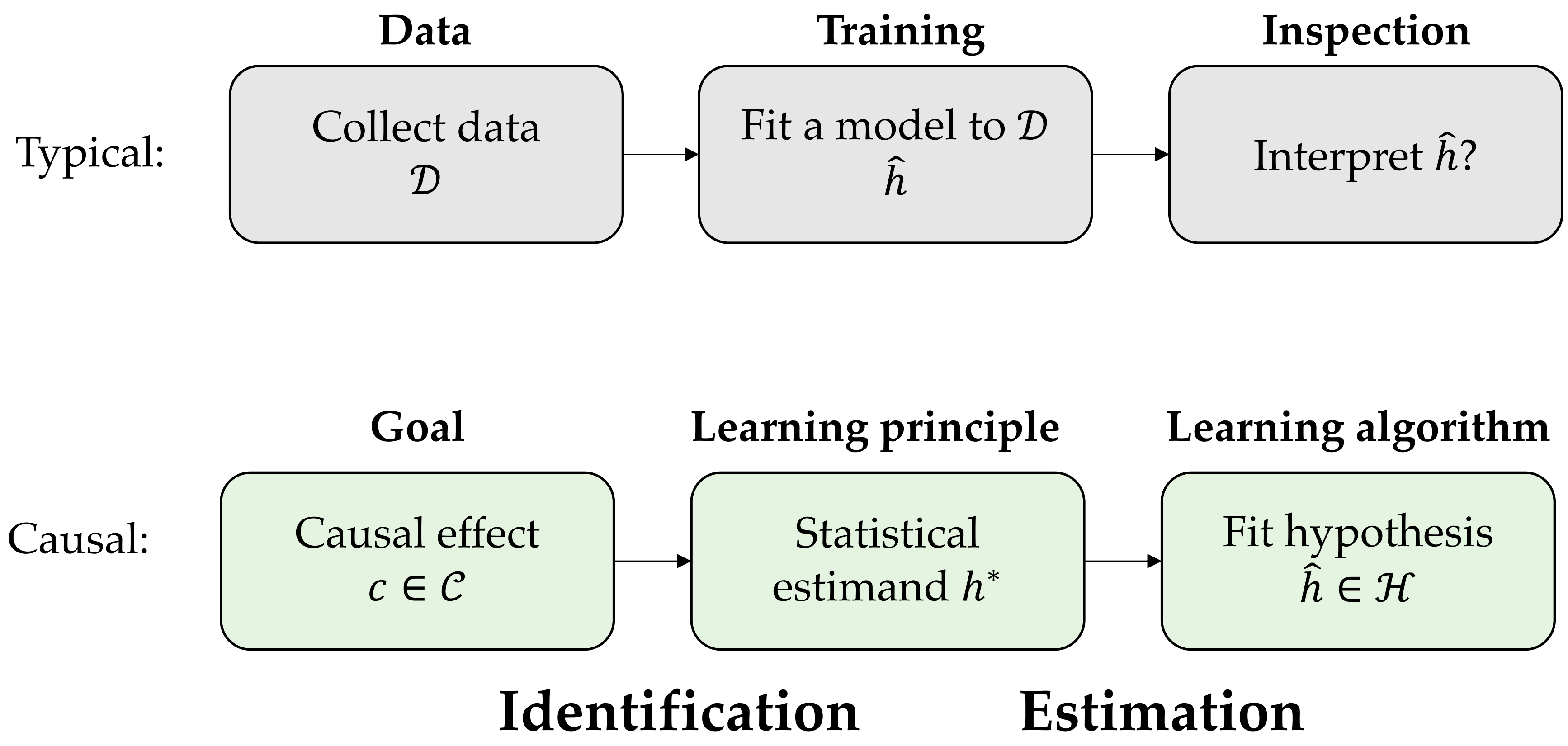
Learning principle

Statistical estimand h^*

Learning algorithm

Fit hypothesis $\hat{h} \in \mathcal{H}$

Causal machine learning should be different!



Typical:

Data

Collect data
$\mathcal{D}$

Training

Fit a model to $\mathcal{D}$
$\hat{h}$

Inspection

Interpret $\hat{h}$?

Causal:

Goal

Causal effect
$c \in \mathcal{C}$

Learning principle

Statistical
estimand h^*

Learning algorithm

Fit hypothesis
$\hat{h} \in \mathcal{H}$

Identification

Estimation

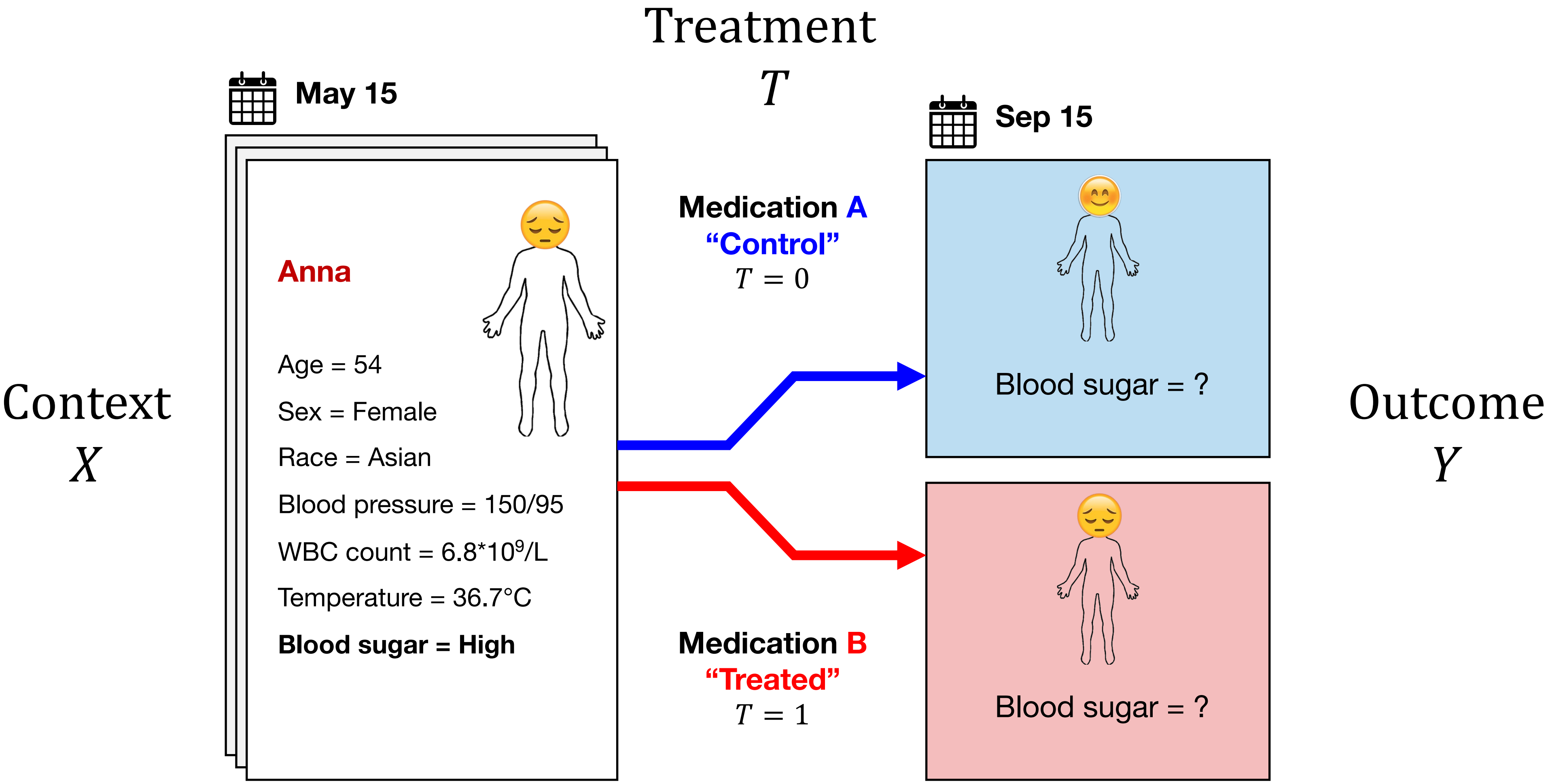
To make *decisions from data,*

we need a way to *specify causal questions*

to identify the *right goal* for machine learning

Part I.c: Potential outcomes of interventions

See e.g., Hernán & Robins, Chapter 1



Treatment

T

Context

X

May 15

Anna

Age = 54

Sex = Female

Race = Asian

Blood pressure = 150/95

WBC count = 6.8*10^9/L

Temperature = 36.7°C

Blood sugar = High

Medication A

“Control”

$T = 0$

Medication B

“Treated”

$T = 1$

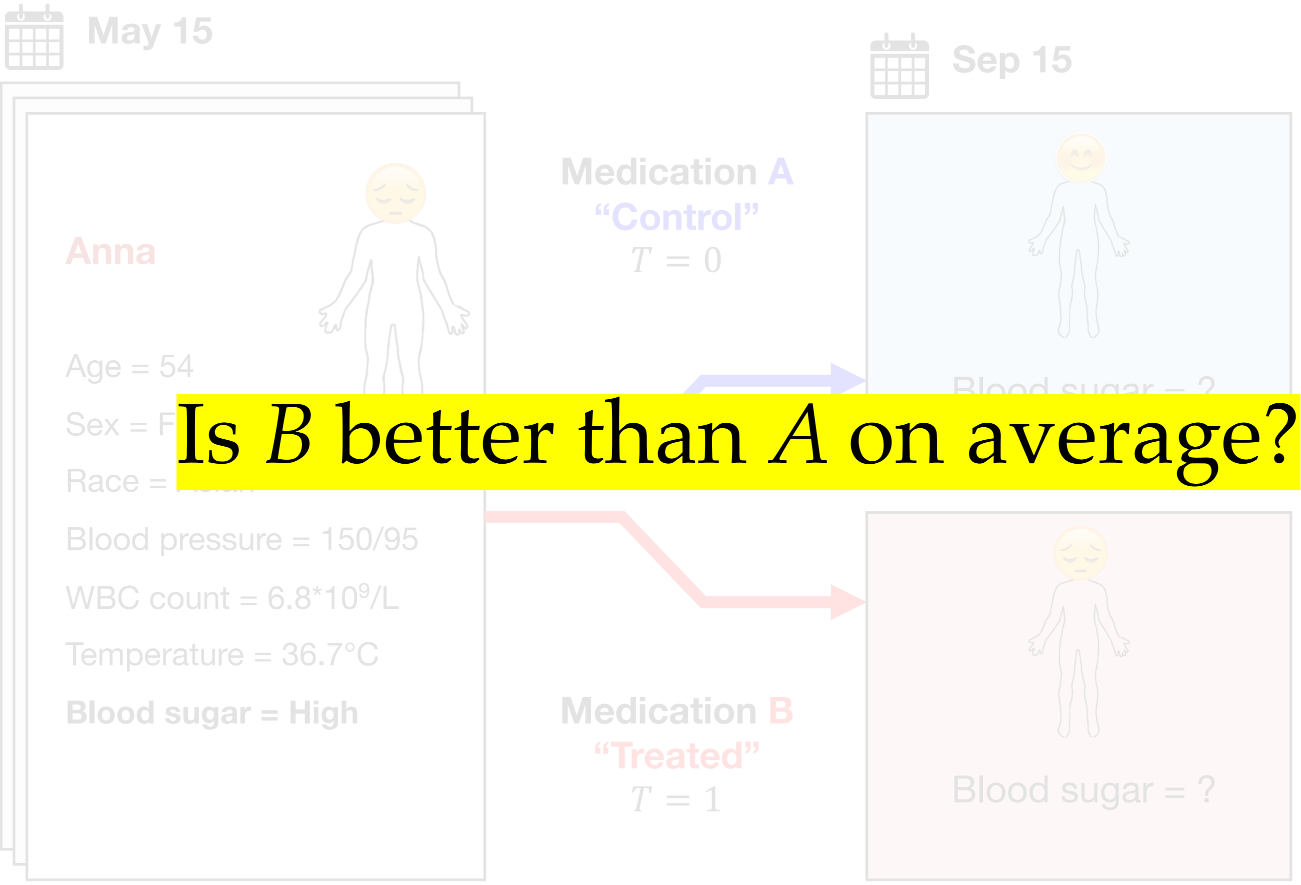
Sep 15

Blood sugar = ?

Blood sugar = ?

Outcome

Y



May 15

Anna

Age = 54

Sex = F

Race =

Blood pressure = 150/95

WBC count = $6.8 \cdot 10^9$/L

Temperature = 36.7°C

Blood sugar = High

Medication A
“Control”
$T = 0$

Medication B
“Treated”
$T = 1$

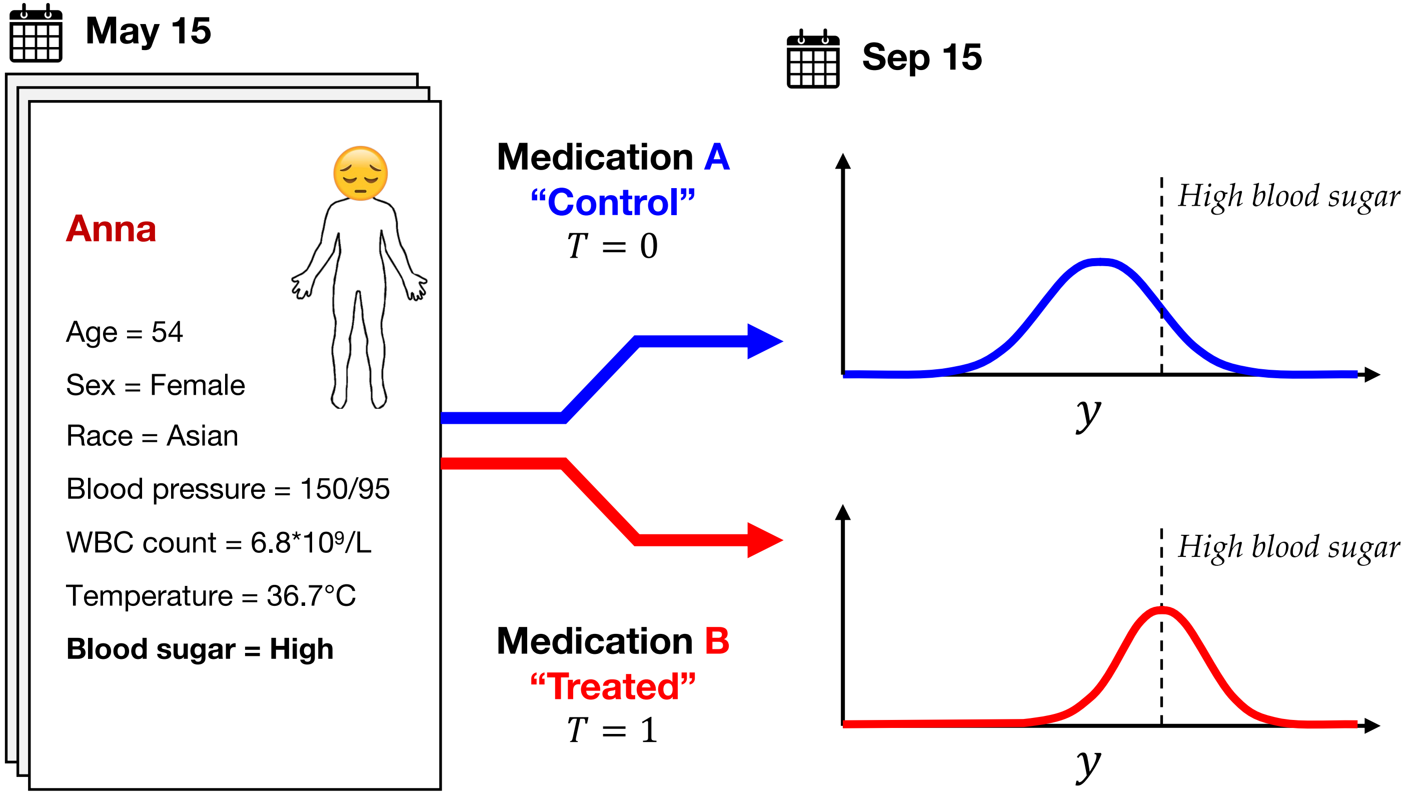
Sep 15

Blood sugar = ?

Blood sugar = ?

Is B better than A on average?

We can imagine two **potential** scenarios for our patients



May 15

Anna

Age = 54

Sex = Female

Race = Asian

Blood pressure = 150/95

WBC count = $6.8*10^9$/L

Temperature = 36.7°C

Blood sugar = High

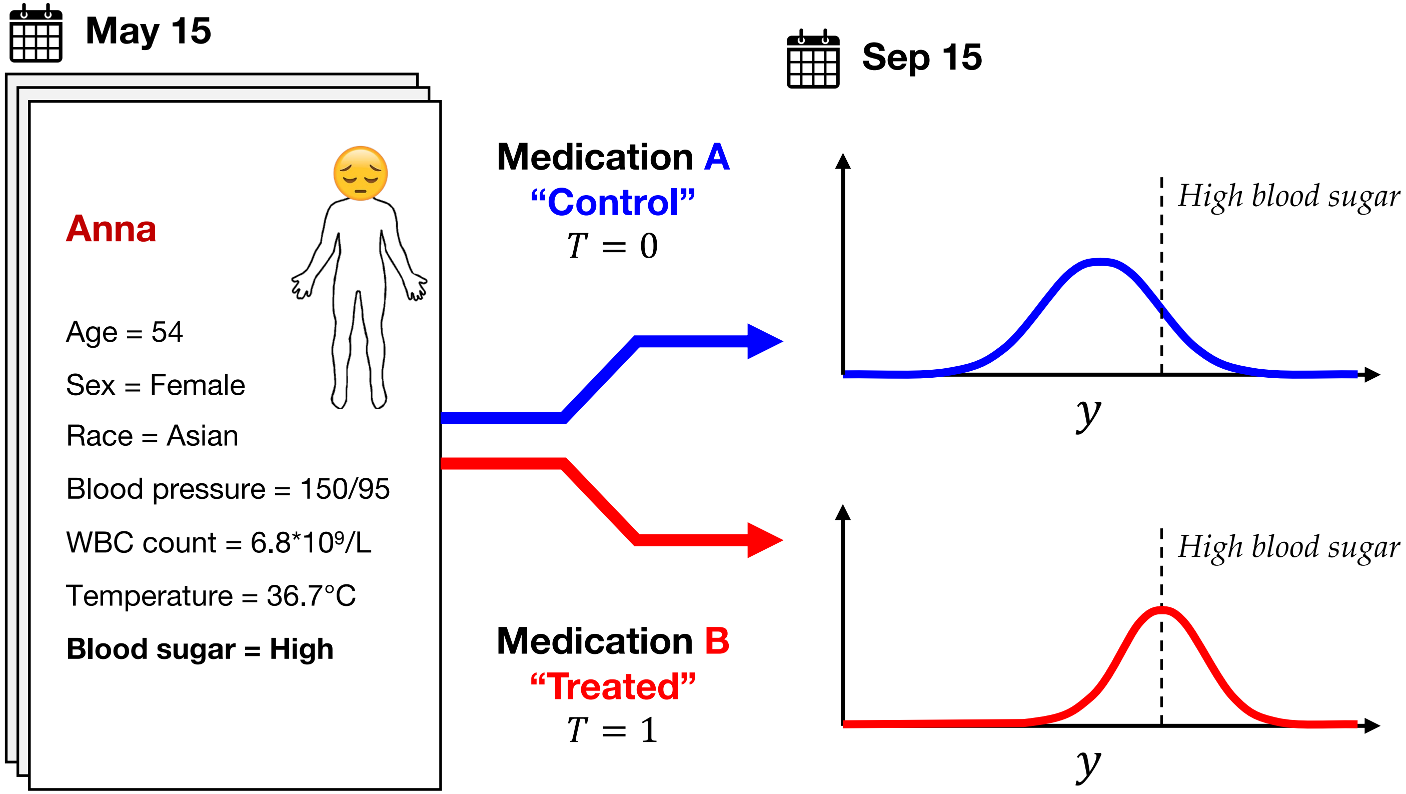
Medication A
"Control"
$T = 0$

Medication B
"Treated"
$T = 1$

Sep 15

Which is better?

Two “potential” random variables



May 15

Anna

Age = 54

Sex = Female

Race = Asian

Blood pressure = 150/95

WBC count = $6.8*10^9$/L

Temperature = 36.7°C

Blood sugar = High

Medication A
“Control”
$T = 0$

Medication B
“Treated”
$T = 1$

Sep 15

$$Y(0) \sim p_0(Y)$$

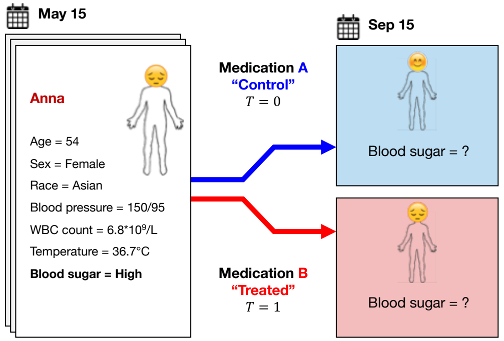
$$Y(1) \sim p_1(Y)$$

Potential outcomes[1] (Neyman-Rubin model)

$\boldsymbol{Y(0)}$: What *would* happen under treatment 0 (“control”)

$\boldsymbol{Y(1)}$: What *would* happen under treatment 1 (“treatment”)

We call $Y(0), Y(1)$ “**potential outcomes**”.



[1] See e.g., Hernan & Robins, Chapter 1

Causal effects of interventions

Potential outcomes let us define the causal effect* Δ of a binary intervention $T \in \{0,1\}$ as

$$\Delta = Y(1) - Y(0)$$

Δ is a *random variable*, just like $Y(0), Y(1)$

Compare with $\mathbb{E}[Y \mid do(X = x)] - \mathbb{E}[Y \mid do(X = x')]$ from earlier!

*Causal effects are also called "treatment effects"

Interpreting causal effects

We should understand Δ as the increase in the outcome as we intervene with $T \leftarrow 1$ compared to $T \leftarrow 0$

If $\Delta > 0$, and higher outcome is better, treatment 1 is preferred

If $\Delta < 0$, and higher outcome is better, treatment 0 is preferred

The causal effect can tell us which *action*, which *decision*, is best

How can we learn Δ from data?

SCMs vs potential outcomes

SCMs which you learned about in the morning is a more general model than potential outcomes

Potential outcomes is a convenient tool for many problems where only a certain intervention is of interest

Observational studies: Using historical data

Example:

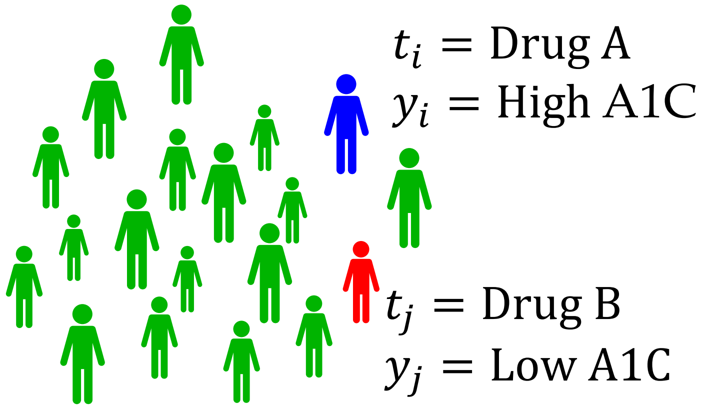
We have been treating pre-diabetic patients for 10 years

We know **which drugs** T were given to which patients and we know what **their blood sugar** Y was 6 months later, as well as some **context** X

We have data $D = \{(x_i, t_i, y_i)\}_{i=1}^{m}$

Dataset *D*

Context *X*	**Treatment *T***	**Outcome *Y***
Older male	Drug A	High A1C
Young female	Drug B	Low A1C
Young male	Drug B	Low A1C
…	…	…

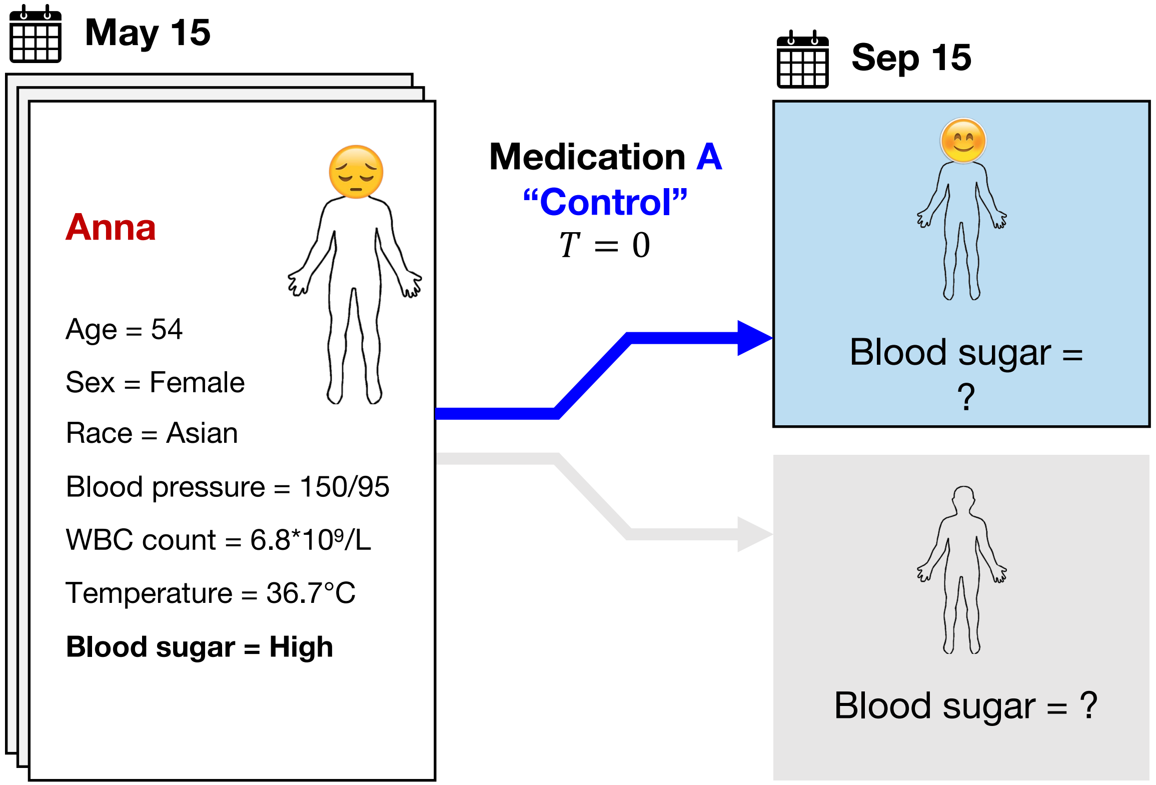


t_i = Drug A
y_i = High A1C

t_j = Drug B
y_j = Low A1C

Fundamental problem of causal inference

We can only observe the outcome of one intervention



May 15

Anna

Age = 54

Sex = Female

Race = Asian

Blood pressure = 150/95

WBC count = 6.8*10⁹/L

Temperature = 36.7°C

Blood sugar = High

Medication A

"Control"

$T = 0$

Sep 15

Blood sugar = ?

Blood sugar = ?

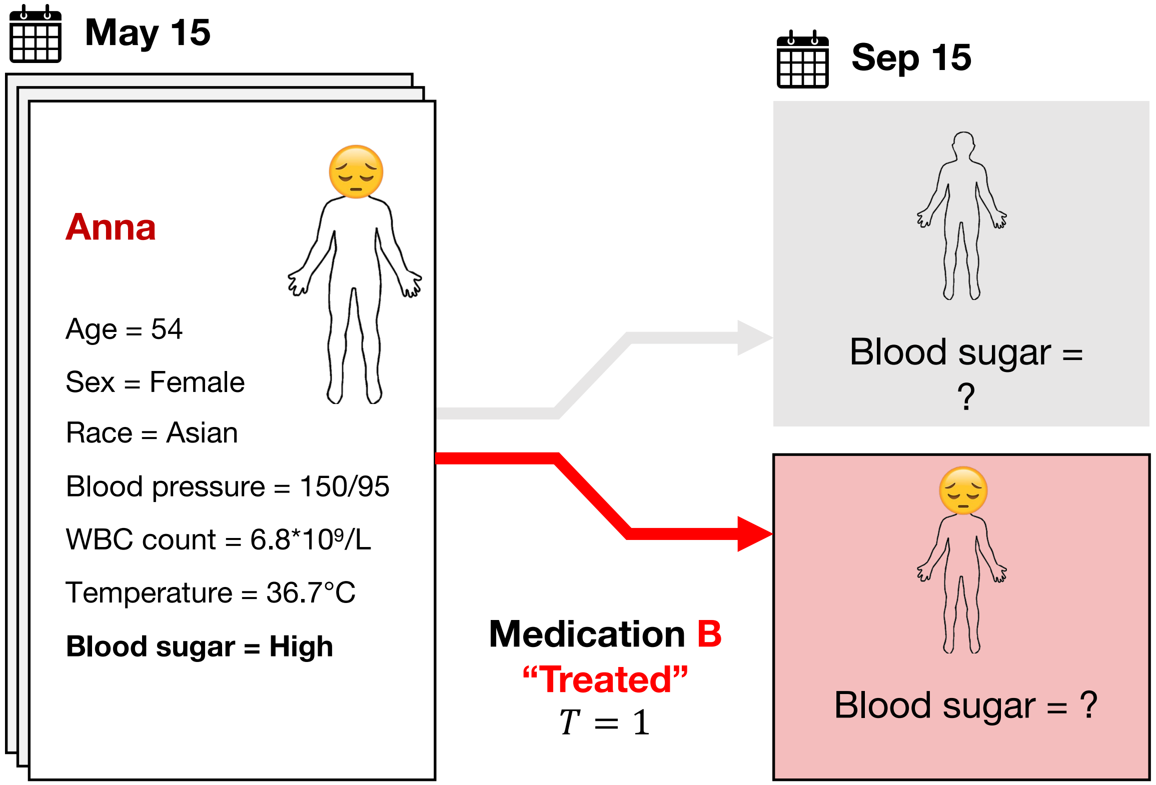
*If we give subject i $T_i \leftarrow 0$, we see $Y_i(0)$

Context X	**Outcome $Y(0)$**	**Outcome $Y(1)$**
Older male	High A1C	**?**
Young female	**?**	Low A1C
Young male	**?**	Low A1C
…	…	…

* This is an implicit assumption called "Consistency". $Y = Y(T)$

Fundamental problem of causal inference

We can only observe the outcome of one intervention



May 15

Anna

Age = 54

Sex = Female

Race = Asian

Blood pressure = 150/95

WBC count = 6.8*10⁹/L

Temperature = 36.7°C

Blood sugar = High

Sep 15

Blood sugar = ?

Medication B

“Treated”

$T = 1$

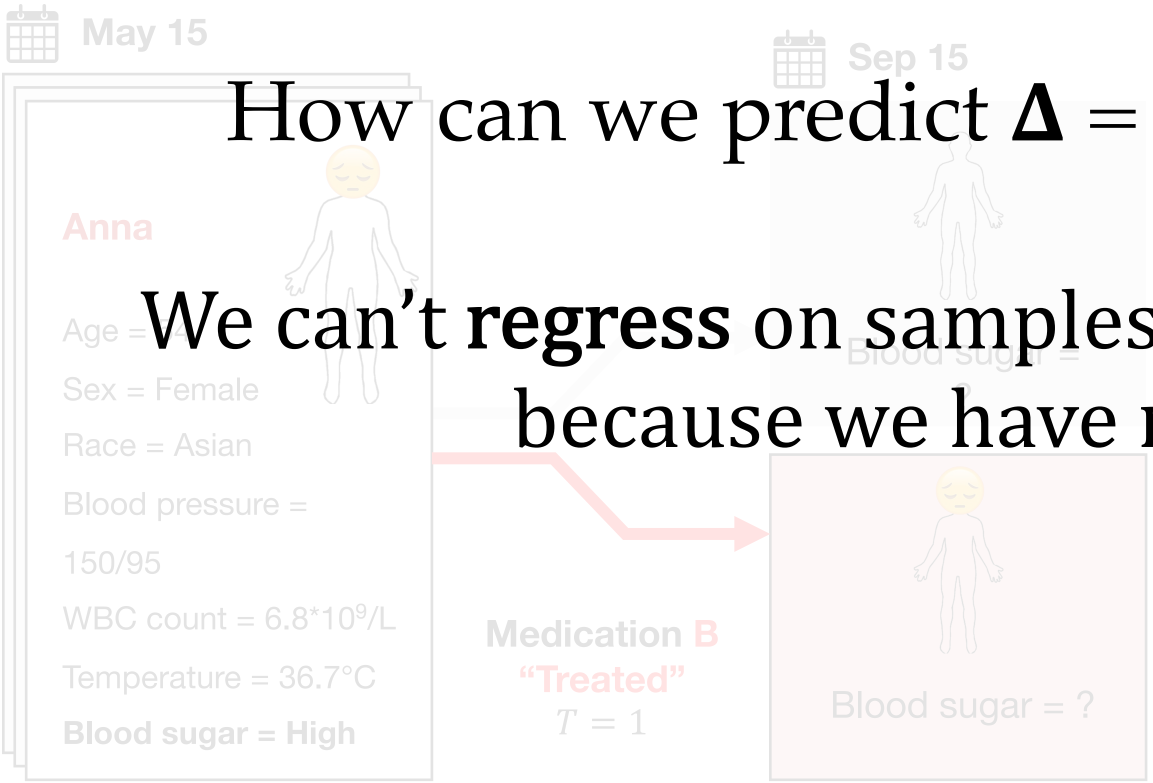
Blood sugar = ?

If we give patient i $T_i \leftarrow 1$, we see $Y_i(1)$

* This is an implicit assumption called “Consistency”. $Y = Y(T)$

Fundamental problem of causal inference

We can only observe the outcome of **one intervention**



If we give the patient $T = 1$, we see $Y(1)$

How can we predict **Δ** = Y(1) – Y(0)?

We can't **regress** on samples of the effect **Δ**, because we have none!

Average treatment effect

First, let's consider the **population average effect** instead

Definition.

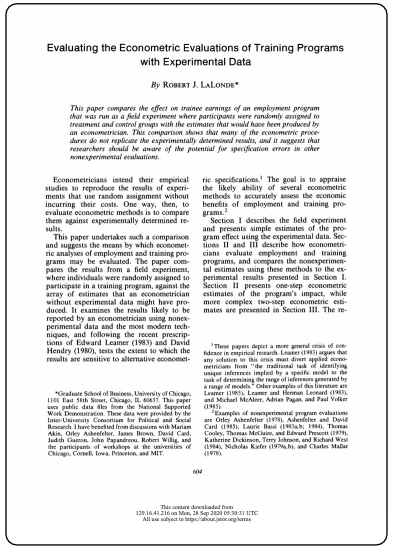
The *average causal effect* (ATE) of a binary treatment $t \in \{0,1\}$ is

$$\tau \coloneqq \mathbb{E}[\Delta] = \mathbb{E}[Y(1) - Y(0)]$$

The ATE averages over individual variation

The "average treatment effect" (ATE) is also called the "average causal effect" (ACE)

Example: LaLonde's study (1986)



Evaluating the Econometric Evaluations of Training Programs with Experimental Data

By Robert J. LaLonde*

This paper compares the effect on trainee earnings of an employment program that was run as a field experiment where participants were randomly assigned to treatment and control groups with the estimates that would have been produced by an econometrician. This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations.

Econometricians intend their empirical studies to reproduce the results of experiments that use random assignment without incurring their costs. One way, then, to evaluate econometric methods is to compare them against experimentally determined results.

This paper undertakes such a comparison and suggests the means by which econometric analyses of employment and training programs may be evaluated. The paper compares the results from a field experiment, where individuals were randomly assigned to participate in a training program, against the array of estimates that an econometrician without experimental data might have produced. It examines the results likely to be reported by an econometrician using nonexperimental data and the most modern techniques, and following the recent prescriptions of Edward Leamer (1983) and David Hendry (1980), tests the extent to which the results are sensitive to alternative econometric specifications.[1] The goal is to appraise the likely ability of several econometric methods to accurately assess the economic benefits of employment and training programs.[2]

Section I describes the field experiment and presents simple estimates of the program effect using the experimental data. Sections II and III describe how econometricians evaluate employment and training programs, and compares the nonexperimental estimates using these methods to the experimental results presented in Section I. Section II presents one-step econometric estimates of the program's impact, while more complex two-step econometric estimates are presented in Section III. The re-

*Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637. This paper uses public data files from the National Supported Work Demonstration. These data were provided by the Inter-University Consortium for Political and Social Research. I have benefited from discussions with Mariam Akin, Orley Ashenfelter, James Brown, David Card, Judith Gueron, John Papandreou, Robert Willig, and the participants of workshops at the universities of Chicago, Cornell, Iowa, Princeton, and MIT.

[1] These papers depict a more general crisis of confidence in empirical research. Leamer (1983) argues that any solution to this crisis must divert applied econometricians from "the traditional task of identifying unique inferences implied by a specific model to the task of determining the range of inferences generated by a range of models." Other examples of this literature are Leamer (1985), Leamer and Herman Leonard (1983), and Michael McAleer, Adrian Pagan, and Paul Volker (1985).

[2] Examples of nonexperimental program evaluations are Orley Ashenfelter (1978), Ashenfelter and David Card (1985), Laurie Bassi (1983a,b; 1984), Thomas Cooley, Thomas McGuire, and Edward Prescott (1979), Katherine Dickinson, Terry Johnson, and Richard West (1984), Nicholas Kiefer (1979a,b), and Charles Mallar (1978).

The LaLonde study looked at the effectiveness of a job training program, T, on the future real earnings of an individual, Y

Subjects were randomized into job training or not

The average yearly real earnings of subjects following job training was \$851 (\$886) higher for females (males) than without training*

$$\widehat{ATE}_{\text{female}} = \$851, \quad \widehat{ATE}_{\text{male}} = \$886$$

*The job training program **cost** \$6800–\$9100

Conditional average treatment effect

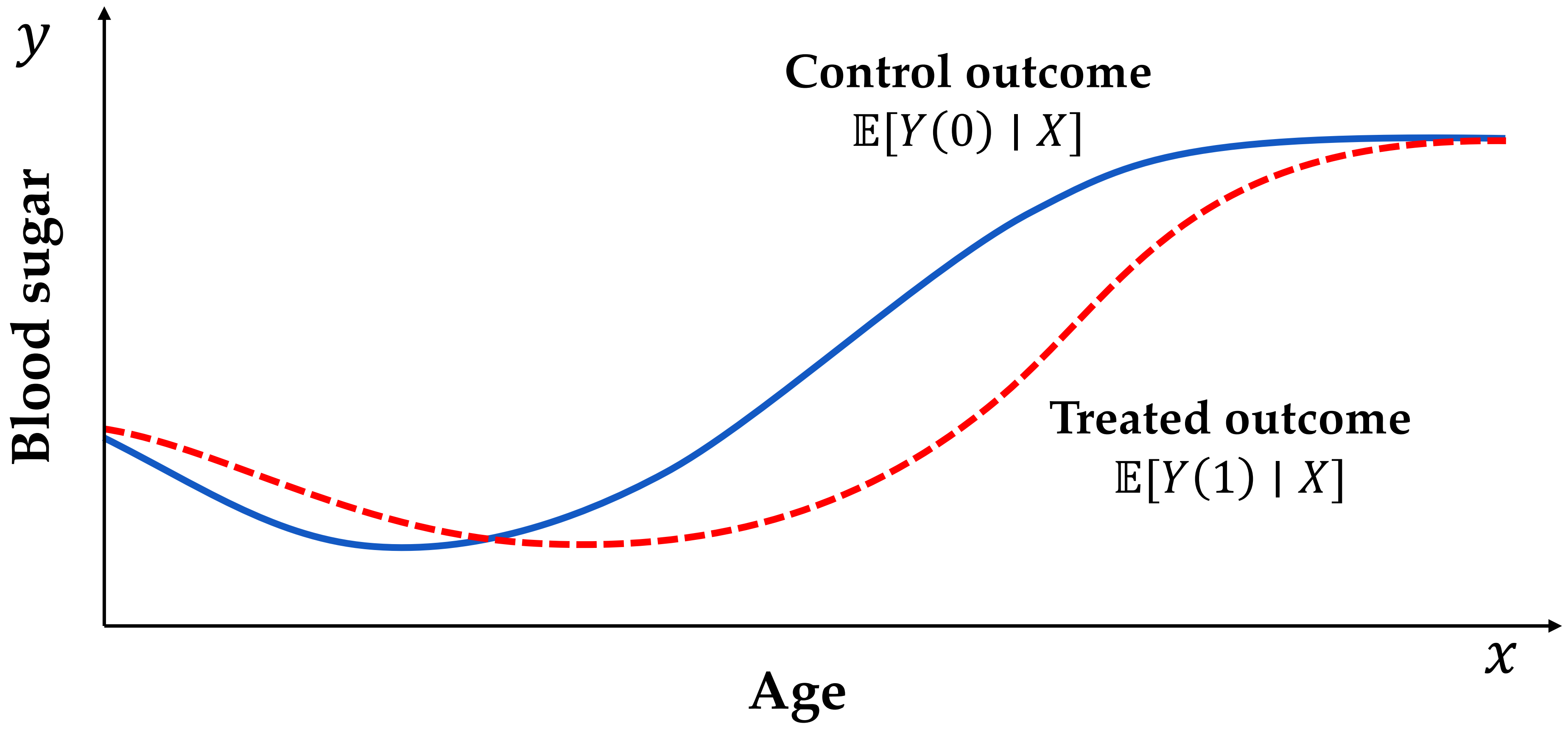
Let X represent a **context variable**, such as age, medical history, etc

Conditional average treatment effect (CATE) w.r.t. X:

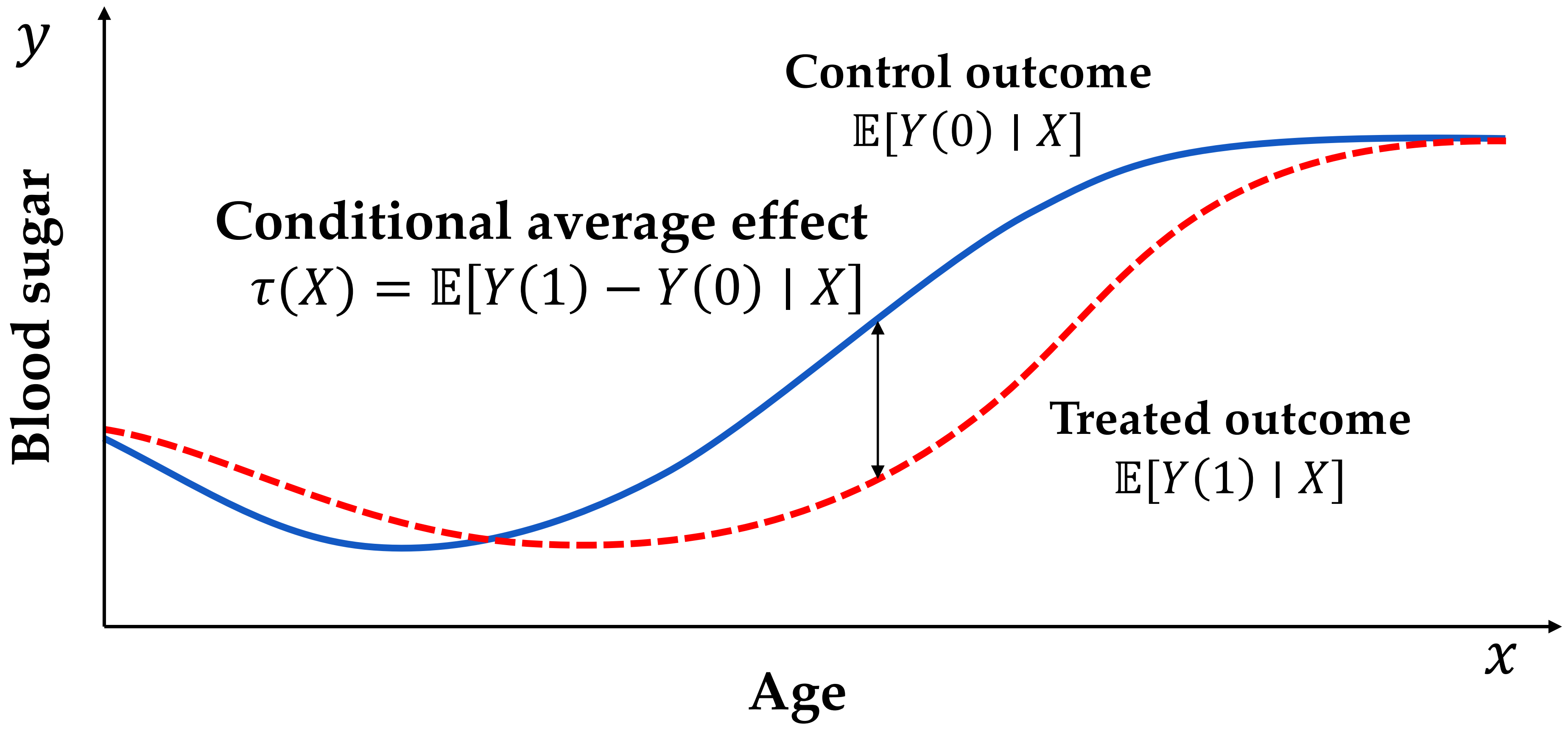
$$\tau(x) \coloneqq \mathbb{E}_Y[\,\Delta \mid X = x\,] = \mathbb{E}_Y[Y(1) - Y(0) \mid X = x]$$

The ATE is the expected CATE, i.e., $\tau = \mathbb{E}_X[\tau(X)]$

Potential outcomes and CATE



Potential outcomes and CATE

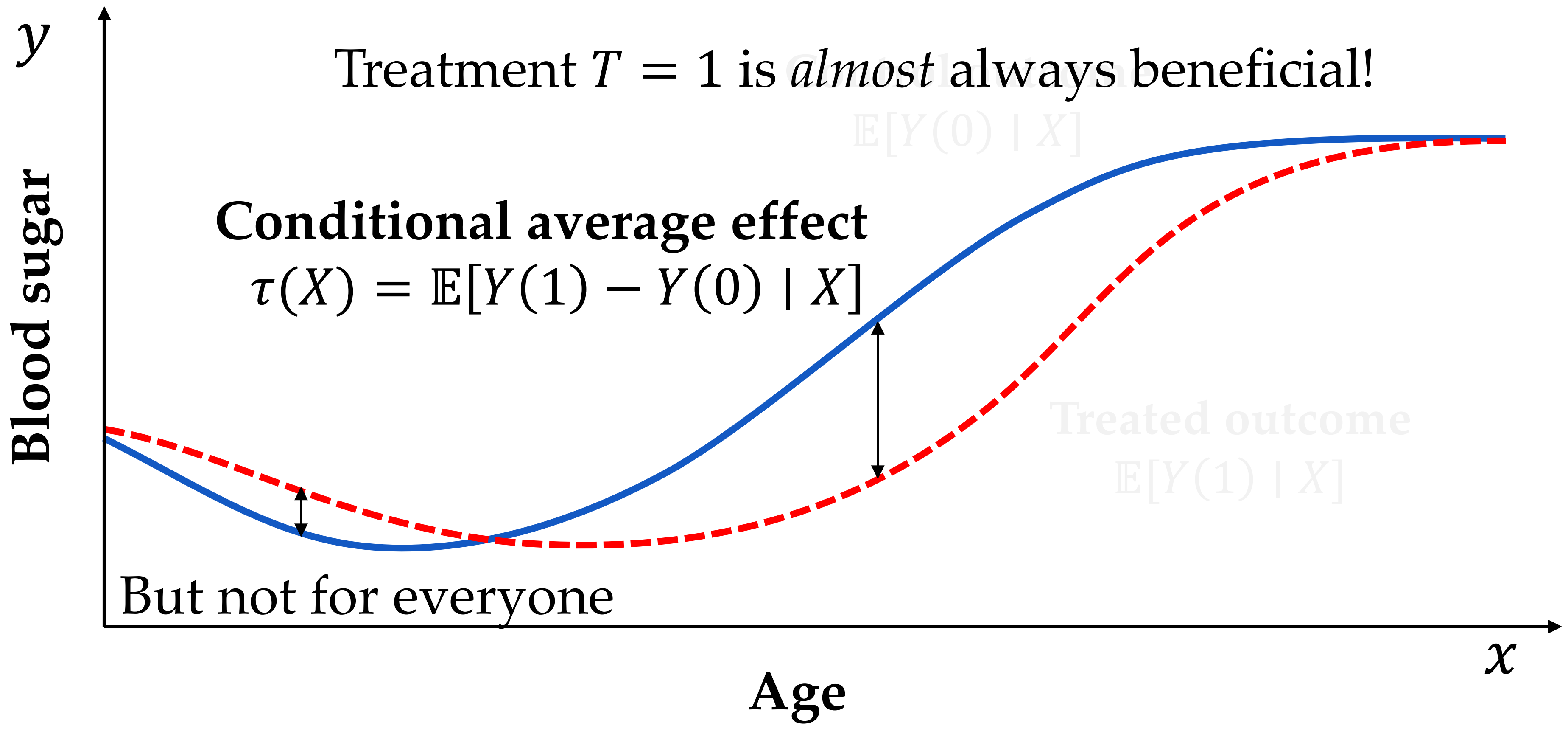


Control outcome
$\mathbb{E}[Y(0) \mid X]$

Conditional average effect
$\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$

Treated outcome
$\mathbb{E}[Y(1) \mid X]$

Potential outcomes and CATE



Treatment $T = 1$ is *almost* always beneficial!

Conditional average effect
$\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$

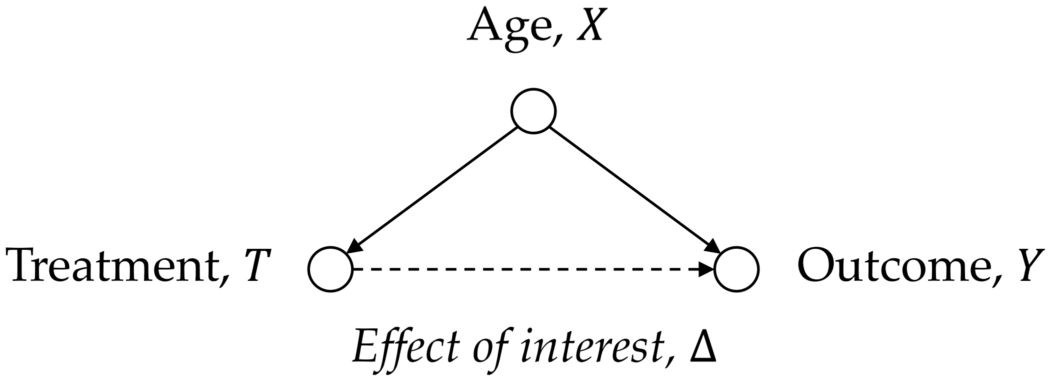
But not for everyone

Part I.d: Partial observability & confounding

Confounding

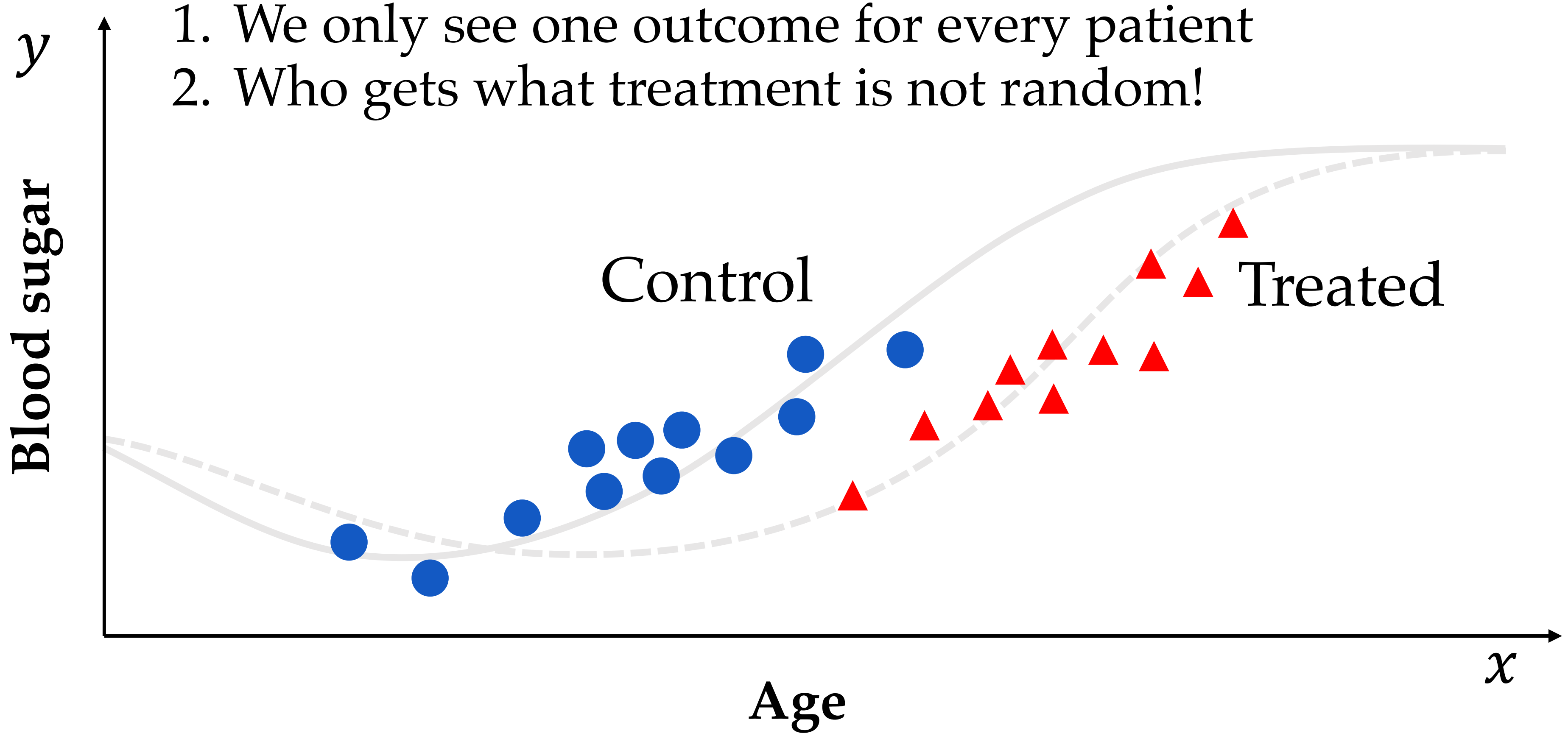
Confounding biases are differences in the outcome between treatment groups which are not related to the treatment's effect

Example: Drug $T = 1$ is given to older patients than Drug B and older patients are at higher risk ⇒ Drug $T = 0$ looks worse!



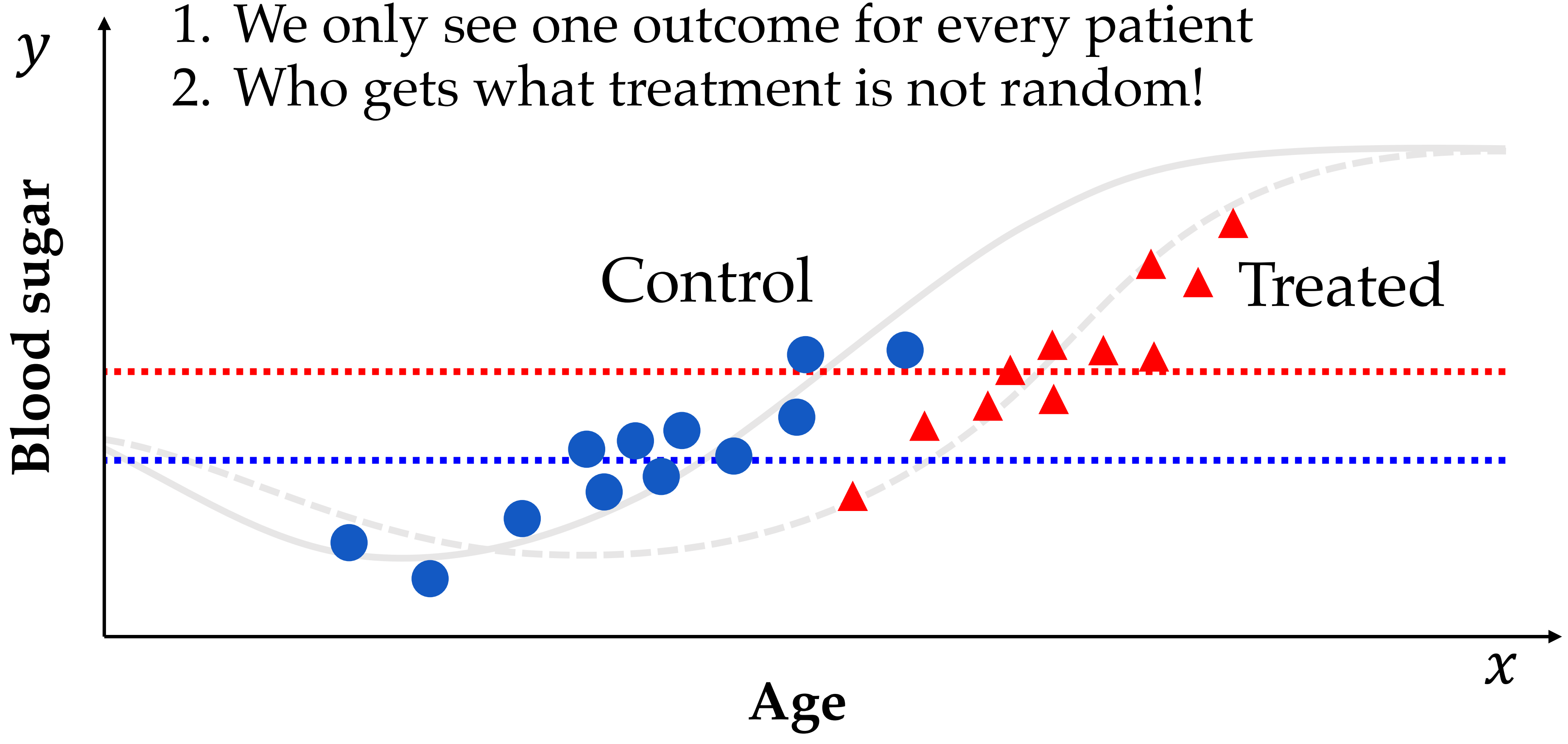
Observational studies

1. We only see one outcome for every patient
2. Who gets what treatment is not random!

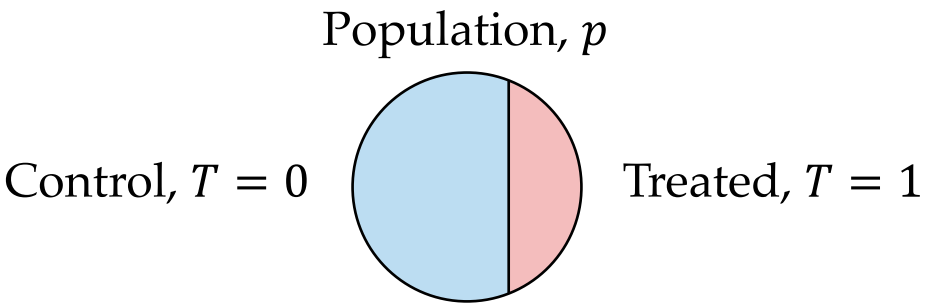


Observational studies

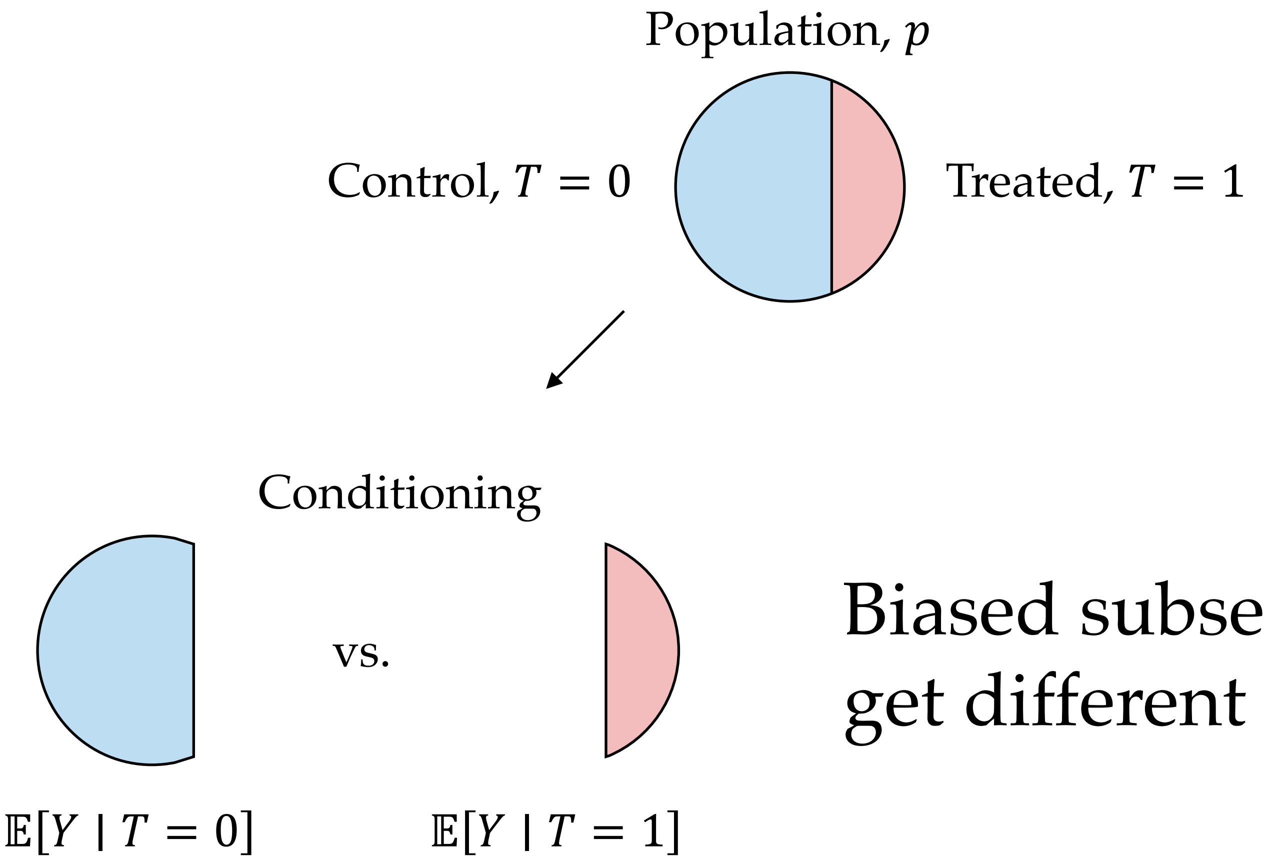
1. We only see one outcome for every patient
2. Who gets what treatment is not random!



The discs represent types of individuals, the color their treatment status



Inspired by Figure 1.1 from Hernan & Robins, *What if*, 2020



Population, p

Control, $T = 0$

Treated, $T = 1$

Conditioning

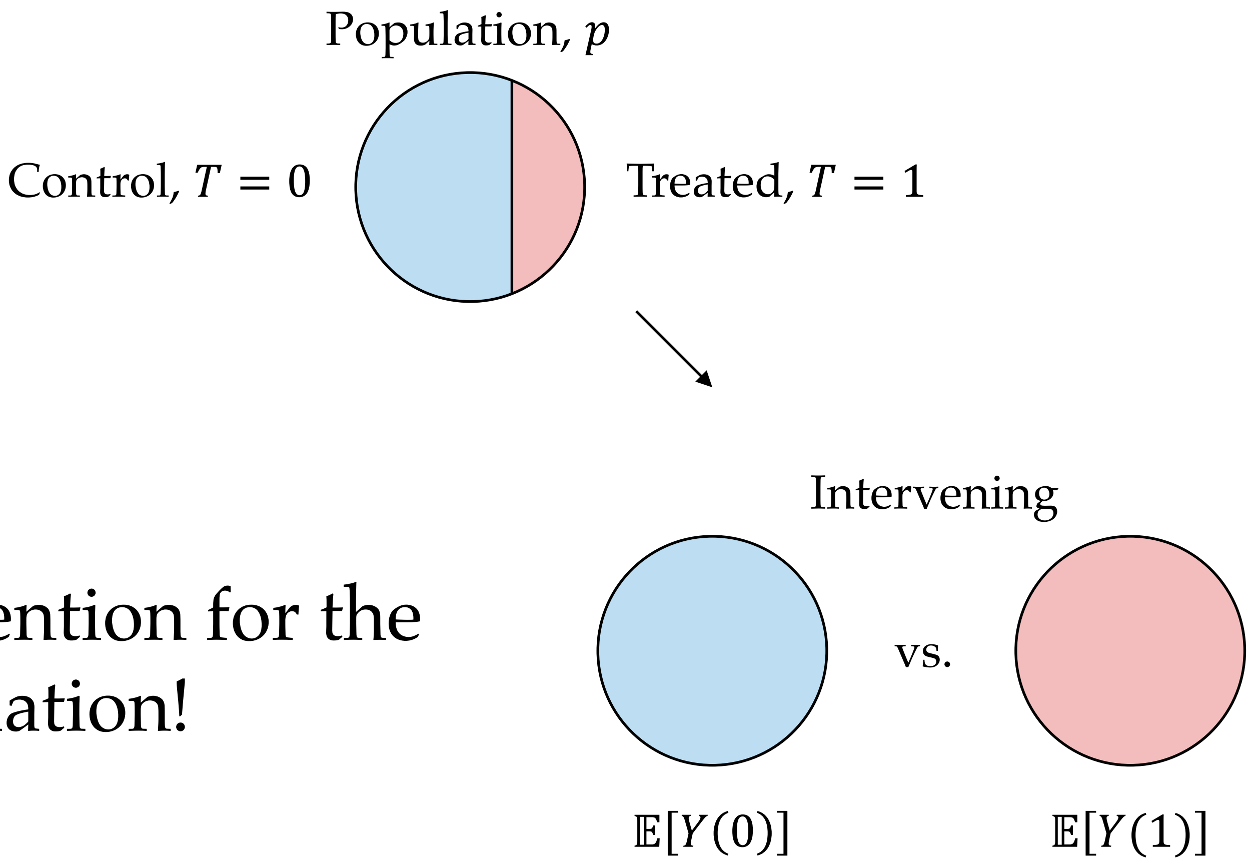
$\mathbb{E}[Y \mid T = 0]$

vs.

$\mathbb{E}[Y \mid T = 1]$

Biased subsets of the population get different treatments

Inspired by Figure 1.1 from Hernan & Robins, *What if*, 2020



Same intervention for the whole population!

Inspired by Figure 1.1 from Hernan & Robins, *What if*, 2020

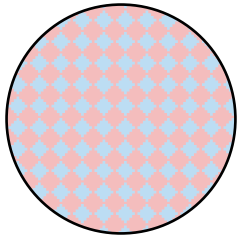
Exchangeability

The best-case scenario is if there is *no pattern* in how treatments are assigned in the data-generating process related to the outcome,

$$Y(t) \perp T$$ — *Treatment groups are "exchangeable"!*

In this case, and $\mathbb{E}[Y(t)] = \mathbb{E}[Y \mid T = t]$

Population, p



We have identified $\mathbb{E}[\boldsymbol{Y}(\boldsymbol{t})]$ as a function of $p(X, T, Y)$

Randomized experiments & identifiability

Experiments ensure exchangeability through **randomization**

Example: $T \sim \text{Bernoulli}(0.5) \quad \Rightarrow \quad T \perp Y(t)$

In observational studies, **can we still have exchangeability**?

Adjustment sets

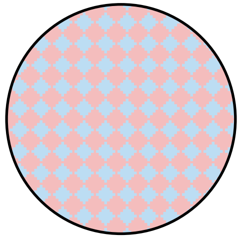
Our strategy will be to *remove as much of the pattern* in treatment assignment as possible by *adjusting* for confounders

If we can find a group of subjects with context variables X that *would respond the same* if given the same treatment, we have

$$Y(t) \perp T \mid X$$

Conditional exchangeability given X*

Subpopulation, $p(\cdot \mid x)$

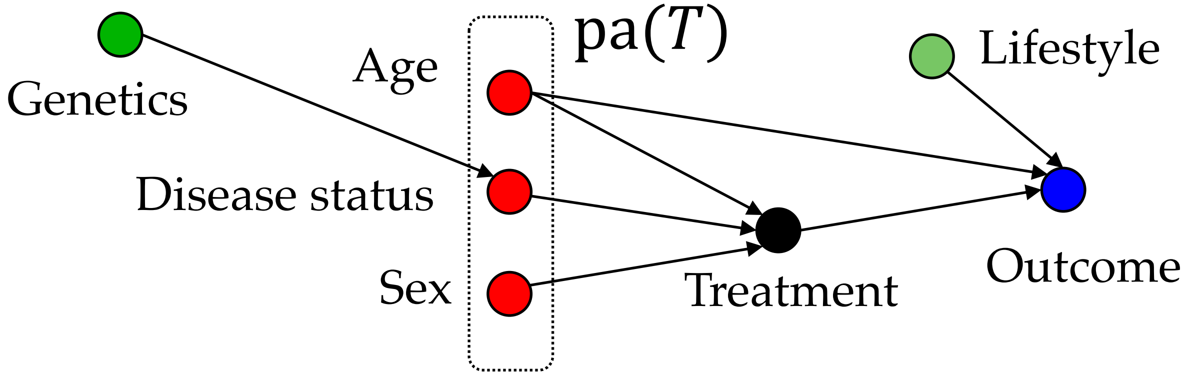


Such a variable X is called an **adjustment set**

* Sometimes called "ignorability"

Parent adjustment

A special case is when we know all direct causes of treatment — all parents in the *causal graph*. (Recall from this morning)



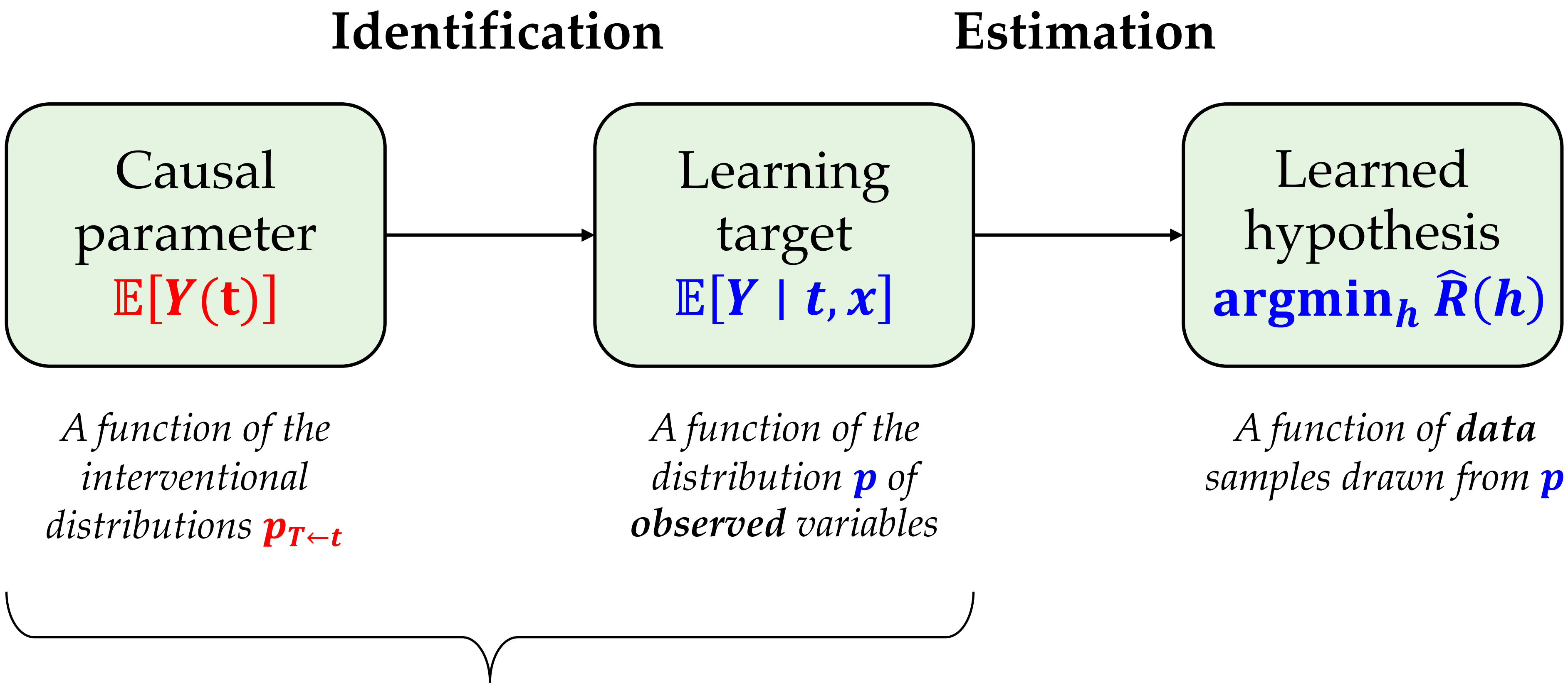
The full set of parents pa(T) satisfies conditional exchangeability*!

It covers all patterns in treatment which may or may not be related to the outcome

[*] See much more on this in Pearl, *Causality*, 2009, including more general criteria such as the backdoor criterion

Part II: Causal effect estimation

Regression & propensity adjustment



Identification **Estimation**

Causal parameter $\mathbb{E}[\boldsymbol{Y}(\mathbf{t})]$

A function of the interventional distributions $\boldsymbol{p}_{\boldsymbol{T}\leftarrow\boldsymbol{t}}$

Learning target $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{t}, \boldsymbol{x}]$

A function of the distribution $\boldsymbol{p}$ *of* ***observed*** *variables*

Learned hypothesis $\mathbf{argmin}_{h}\, \widehat{\boldsymbol{R}}(\boldsymbol{h})$

A function of ***data*** *samples drawn from* $\boldsymbol{p}$

What assumptions can we use here?

Identifying assumptions

We will make the following **identifying assumptions**

A set of variables X is known which satisfies, for all t and x:

1. **Conditional exchangeability:** $Y(t) \perp T \mid X = x$
2. **Treatment group overlap:** $p(T = t \mid X = x) > 0$
3. **Consistency:** $Y = Y(T)$

With these we can prove identifiability of ATE and derive methods!

Two common methods

Regression adjustment

identifies the effects of treatment by accounting for the effect of confounding variables on the **outcome**

Use machine learning to model *Y*!

Propensity adjustment

identifies the effects of treatment by accounting for the effect of confounding variables on the **treatment**

Use machine learning to model *T*!

Part II.a: Regression adjustment

Proof of identifiability w. regression adjustment

Under **conditional exchangeability**, **consistency**

$$Y(t) \perp T \mid X \quad \text{and} \quad Y = Y(T)$$

we have

$$p(\,Y(t) \mid X\,) \underset{\substack{\textit{Cond. exchangeability} \\ Y(t) \perp T \mid X}}{=} p(\,Y(t) \mid T = t, X\,)$$

Proof of identifiability w. regression adjustment

Under **conditional exchangeability**, **consistency**

$$Y(t) \perp T \mid X \quad \text{and} \quad Y = Y(T)$$

we have

$$p(Y(t) \mid X) \underset{\textit{Cond. exchangeability}}{=} p(Y(t) \mid T = t, X) \underset{\substack{\textit{Consistency} \\ Y = Y(T)}}{=} p(Y \mid T = t, X)$$

Proof of identifiability w. regression adjustment

Under **conditional exchangeability**, **consistency**

$$Y(t) \perp T \mid X \quad \text{and} \quad Y = Y(T)$$

we have

$$p(\,Y(t) \mid X\,) = p(\,Y(t) \mid T = t, X\,) = p(Y \mid T = t, X)$$

It follows that

Interventional $$\mathbb{E}[Y(t) \mid X = x] = \sum_y y\, p(Y(t) = y \mid X = x)$$

$$= \sum_y y\, p(Y \mid X = x, T = t) = \mathbb{E}[Y \mid X = x, T = t]$$ *Observational*

Final steps of identification

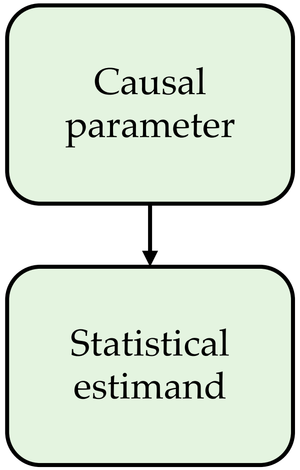
On the last slide, we proved $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid X = x, T = t]$ under conditional exchangeabiltiy and consistency

To get the ATE, we can average these conditionals

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

$$= \mathbb{E}_X\big[\mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0]\big]$$

Can estimate with regression (need overlap to get right!)



Causal parameter

Statistical estimand

Regression adjustment with “T-learner”*

1. Identify a valid adjustment set X
2. Collect data $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$
3. Fit regression models $\hat{m}_t(x) \approx \mathbb{E}[Y \mid X = x, T = t]$

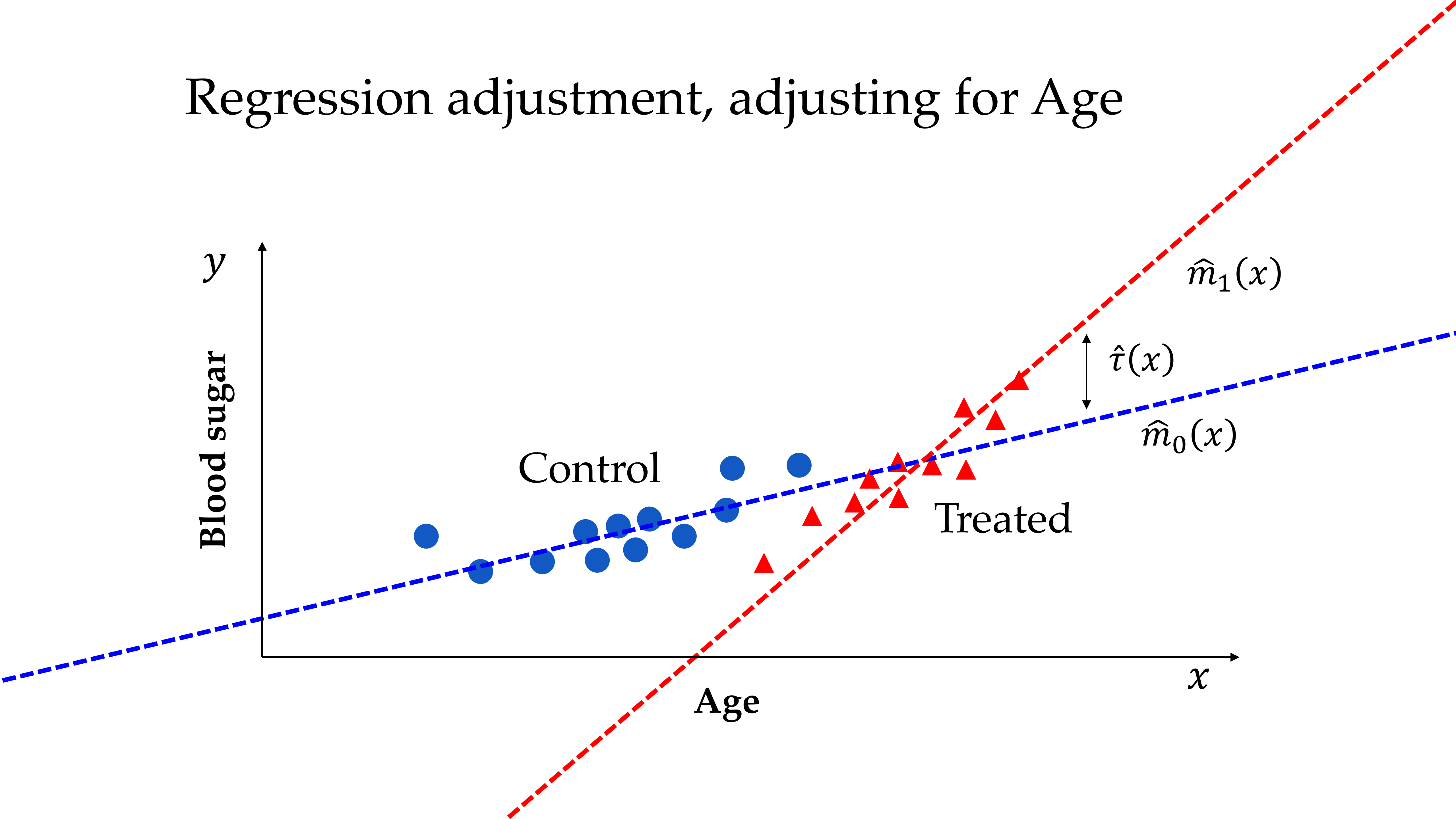
   e.g., using empirical risk minimization (ERM)*:

   $$\hat{m}_t(x) = \arg\min_f \frac{1}{n_t} \sum_{i:t_i=t} (f(x_i) - y_i)^2$$

4. Return $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \big(\hat{m}_1(x) - \hat{m}_0(x)\big)$

*Künzel et al., *PNAS*, 2019

Regression adjustment, adjusting for Age



Part II.b: Inverse-propensity weighting

Propensity score

The propensity score $e(x) = p(T = 1 \mid X = x)$ is the probability that a subject with covariates x receives treatment[1]

It is a summary of the adjustment set with an interesting property:

Assuming that X is an adjustment set, $e(X)$ is also one

[1] Rosenbaum & Rubin, *The central role of the propensity score in observational studies of causal effects*, 1983

Inverse propensity score weighting (IPW)

Proof of identifiability. By definition, we have that

$$\mathbb{E}[Y(t)] = \sum_{x,y} p(X = x) p(Y(t) = y \mid X = x) y = (*)$$

Overall population

Inverse propensity score weighting (IPW)

Proof of identifiability. By definition, we have that

$$\mathbb{E}[Y(t)] = \sum_{x,y} p(X = x) p(\, Y(t) = y \mid X = x\,) y = (*)$$

Multiplying and dividing by $p(X = x \mid T = t)$, we get

$$(*) = \sum_{x,y} \underset{\substack{| \\ \textit{Treatment group } t}}{p(X = x \mid T = t)} \frac{p(X = x)}{p(X = x \mid T = t)} \underset{\substack{| \\ \textit{Potential outcome}}}{p(\, Y(t) = y \mid X = x\,)}\, y$$

Inverse propensity score weighting (IPW)

By **consistency**, **overlap, conditional exchangeability**, **Bayes rule**,

$$(*) = \sum_{x,y} p(X = x \mid T = t) \frac{p(X = x)}{p(X = x \mid T = t)} p(\underset{\textit{Observational outcome}}{Y} = y \mid X = x, T = t\,)y$$

$$= \mathbb{E}_{X,Y}\left[\frac{p(T = t)}{\underset{> 0 \textit{ due to overlap}}{p(\,T = t \mid X = x\,)}} Y \,\middle|\, T = t \right] \square \quad \textit{Weighted average}$$

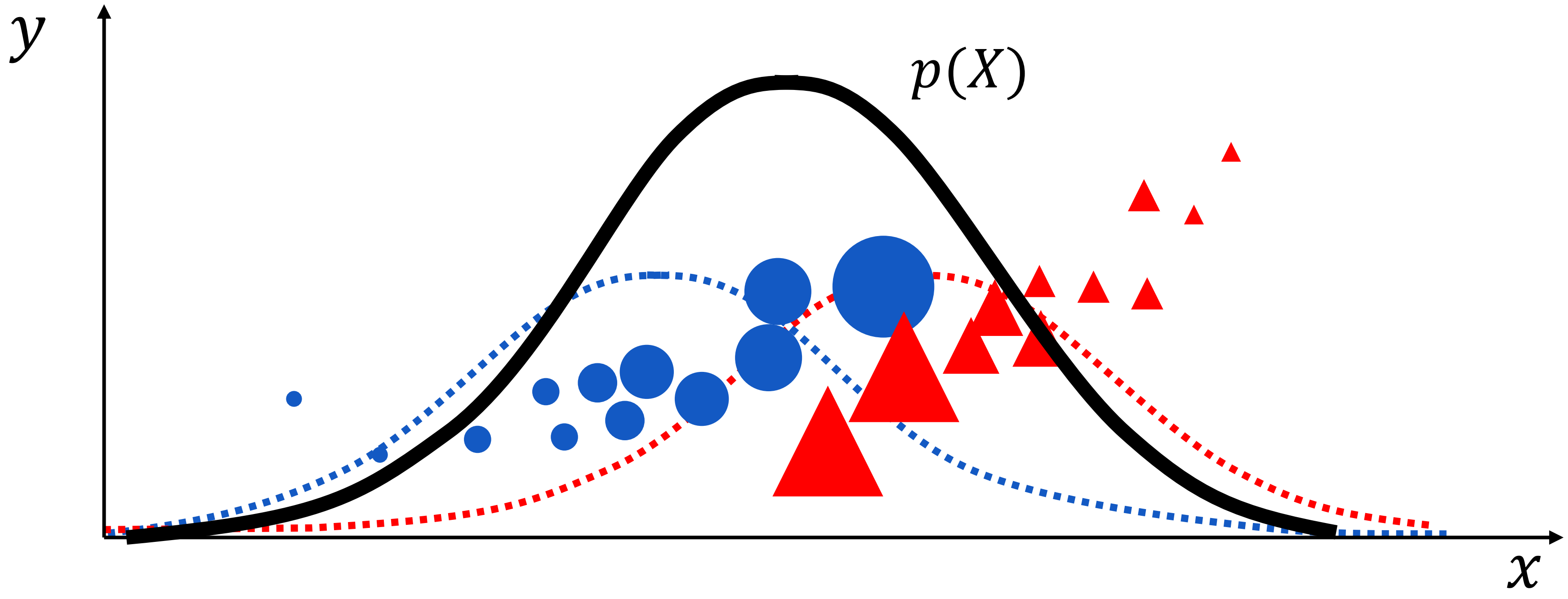
$\Rightarrow \mathbb{E}[Y(t)]$ can be identified by Inverse Propensity Weighting (IPW)

Inverse-Propensity Weighting Adjustment

1. Identify a valid adjustment set X
2. Collect data $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$
3. Fit model $\hat{e}(x) \approx p(\,T = 1 \mid X = x\,)$ of the propensity score
4. Estimate $\hat{\tau} = \frac{1}{n}\sum_{i:t_i=1} \frac{y_i}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i:t_i=0} \frac{y_i}{1-\hat{e}(x_i)}$
5. Return $\hat{\tau}$

IPW example

IPW emphasizes points to look like they came from $p(X)$



Demo: Causal effects of studies

IncomeSim

Simulator of causal effects of **studies** on **future income**

Based on the well-known Adult dataset from UCI

On the repository: https://github.com/Healthy-AI/IncomeSim/

You can find a link to a colab notebook there!

Today’s task

We will work with the question

“What is the causal effect of starting a year of **studies** at on **income** after 10 years?”

We will work with samples generated by the simulator

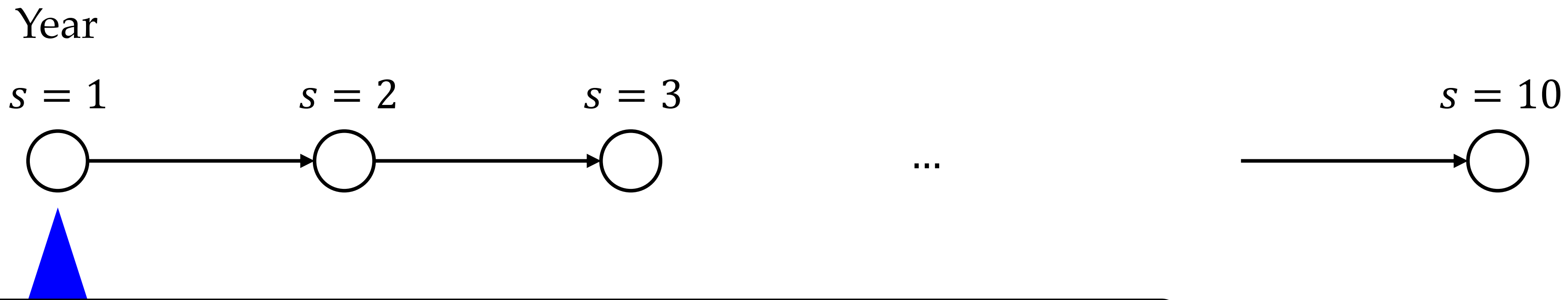
[Let’s have a look at the data]

The simulator takes the variables from Adult and fits an autoregressive model of its variables in a particular order

Example: ‘workclass’ a function of ‘age’, ‘education’, ‘race’, ‘sex’

A variable ‘training’ is added by hand to capture the effects of studies on the other variables (through the ‘education’ marker)

Markov model



Starting state distributions ($s = 1$):

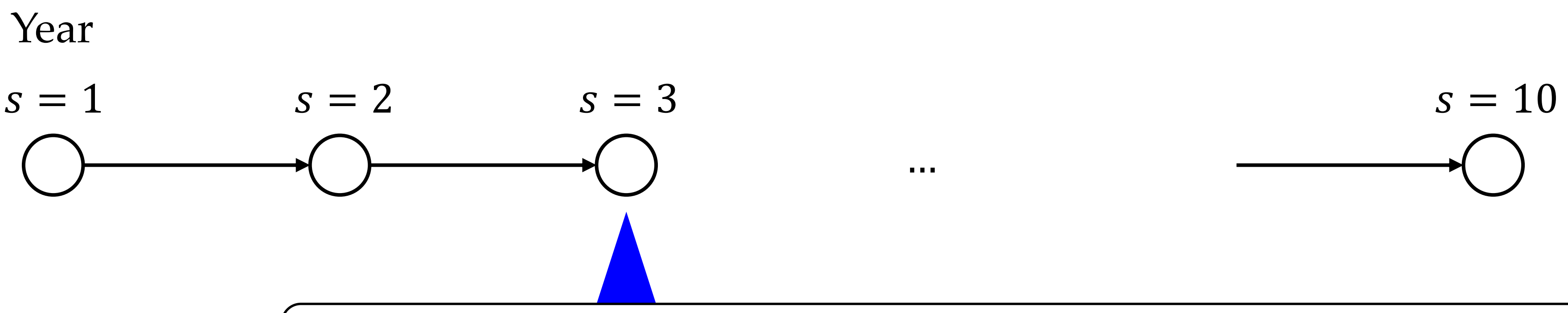
$\text{age}_0 \sim \hat{p}(\text{age}_0)$

$\ldots$

$\text{workclass}_0 \sim \hat{p}(\text{workclass}_0 \mid \text{sex}_0, \text{race}_0, \text{age}_0, \text{education}_0)$

$\ldots$

Markov model



Transition distributions $(s > 0)$:

$\text{age}_s \sim p(\text{age}_s \mid \text{age}_{s-1})$
$\ldots$
$\text{workclass}_s \sim p(\text{workclass}_s \mid \text{workclass}_{s-1}, \text{sex}_s, \text{race}_s, \text{age}_s, \text{education}_s)$
$\ldots$

Causal effect of studies

The data is a typical cross-sectional dataset:

- Some context variables X
- A treatment variable T
- An outcome variable Y

Values from simulator at $s = 1$

Value from simulator at $s = 10$

We could attempt to estimate the average treatment effect

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

The meaning of ATE?

In our case, the meaning of the ATE is

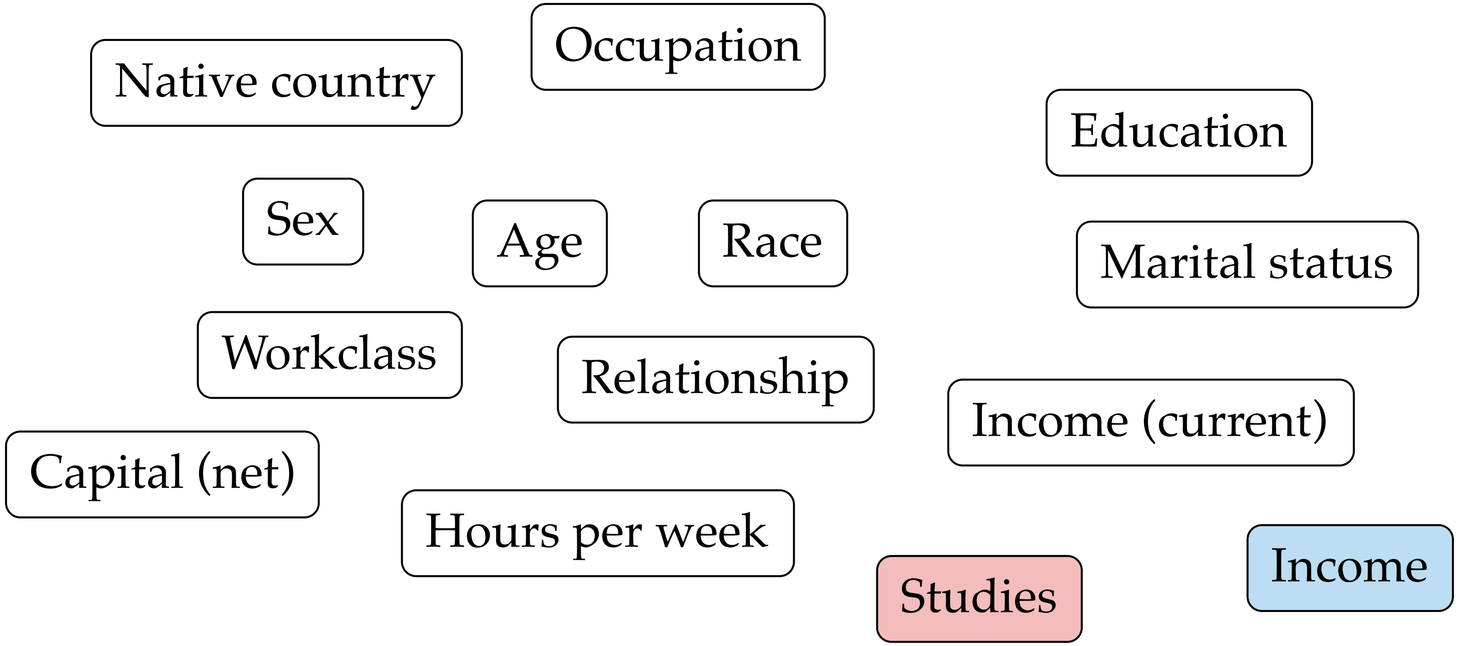
”If we made the population study according to $T = 1$ for a year
and then did what they wanted for 9 years,
how much higher/lower would their yearly income be
compared to if we made them study according to $T = 0$?”

So what does $T = 1$ and $T = 0$ mean?

Recap: Steps to estimate ATE

1. Identify causal parameter	*Average treatment effect*
2. Pick an identification strategy	*Backdoor adjustment*
3. Identify a statistical estimand	*e.g., conditional expectation*
4. Execute an estimation strategy	*e.g., random forest regression*

What is the causal graph, the SCM?



Native country

Occupation

Education

Sex

Age

Race

Marital status

Workclass

Relationship

Income (current)

Capital (net)

Hours per week

Studies

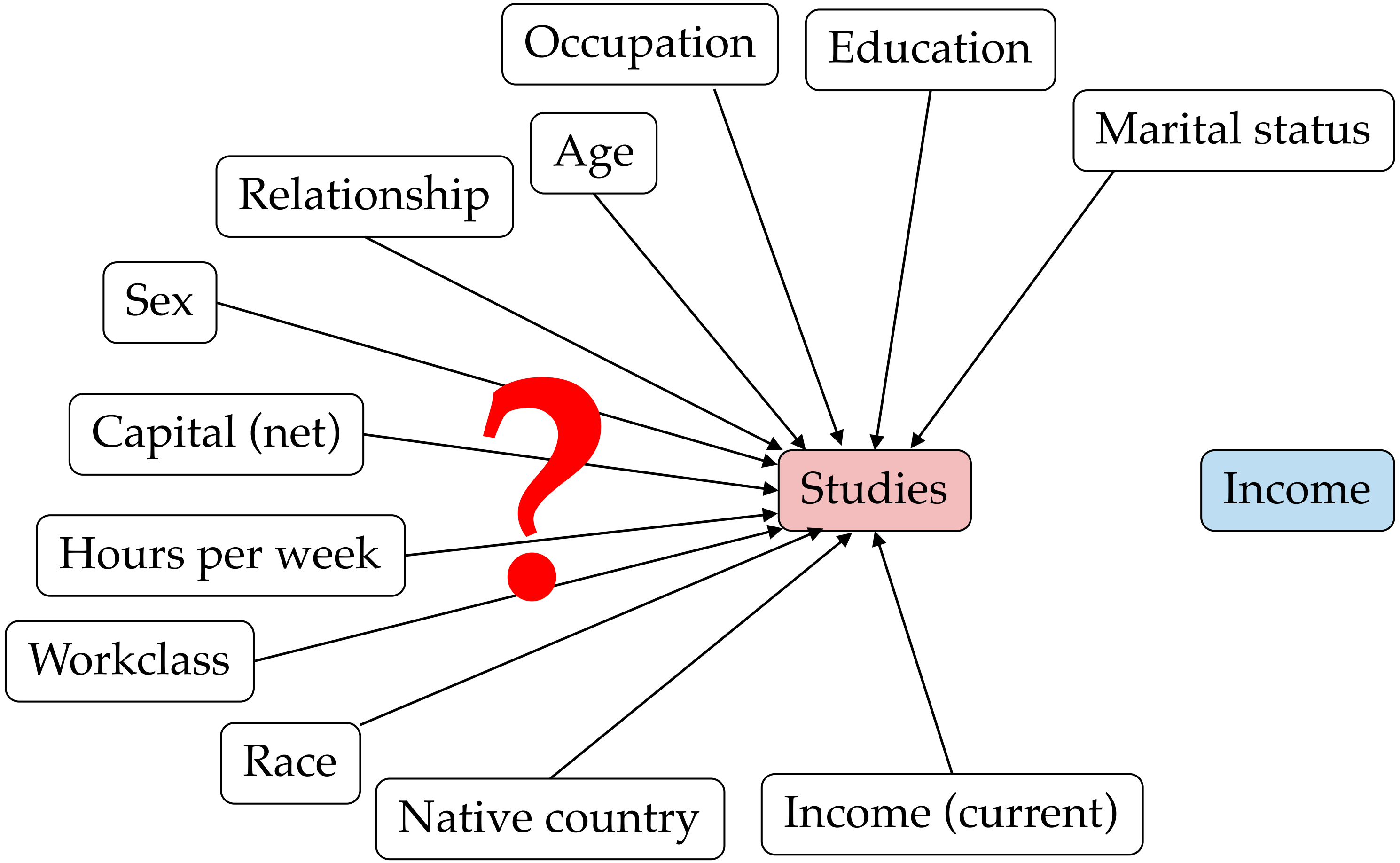
Income

Recall: Parent adjustment

If we know **all direct causes** of the treatment (studies), these satisfy the backdoor criterion

⇒ They are an adjustment set

We don't need the whole graph!

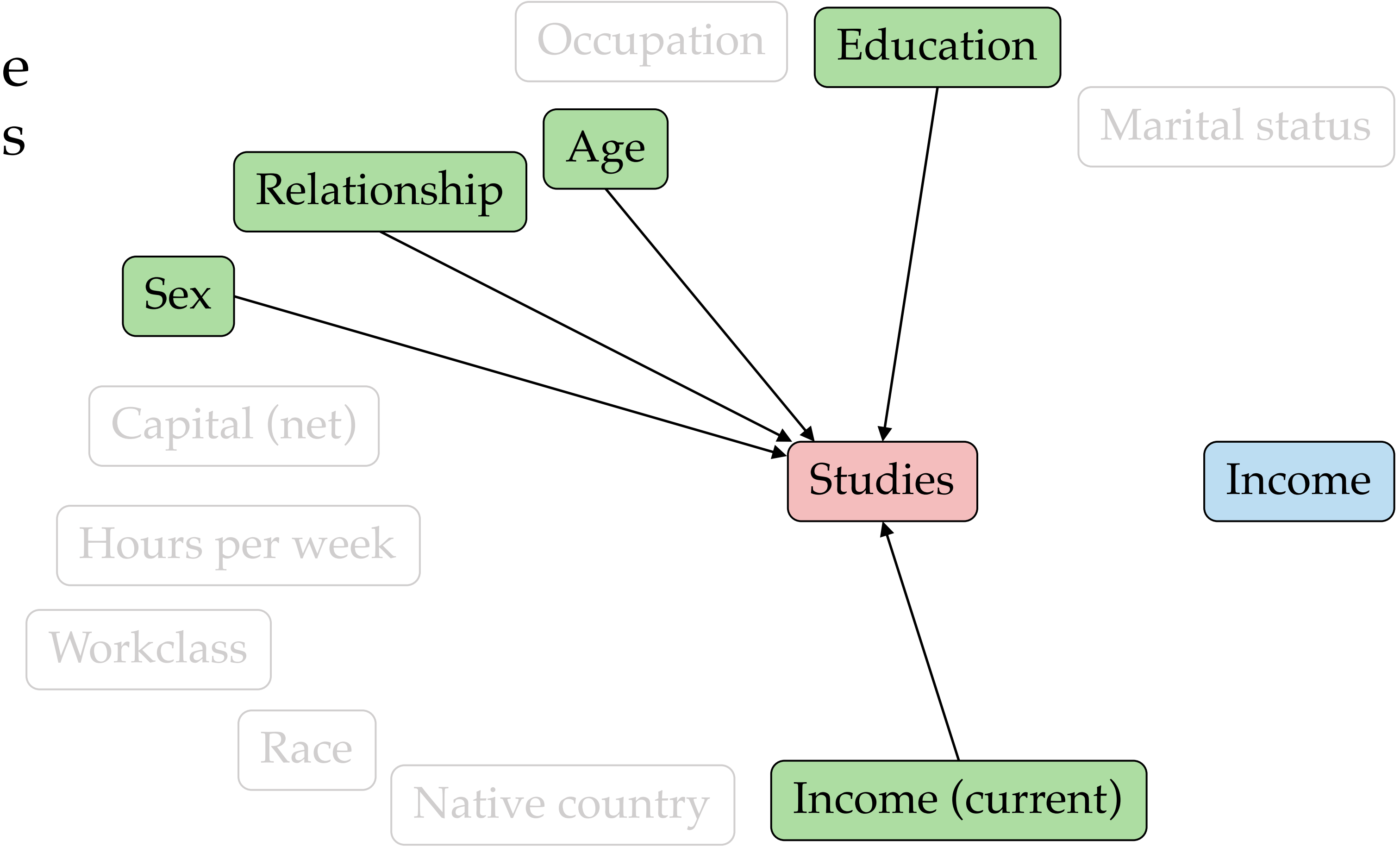


Parent adjustment

Let’s assume that these are *all* the direct causes

- ’Age’,
- ’Sex’,
- ’Relationship’
- ’Education’
- ’Income (current)’

= adjustment set A



[OK, let’s do it!]

Reflections

Where are we now?

We have derived and tested two techniques for causal estimation

- Regression adjustment
- Propensity adjustment

by identifying adjustment sets which satisfy assumptions:
conditional exchangeability, **overlap** and **consistency**

Advanced methods

There are **many** more methods for estimating ATE/CATE!

Many of these combine propensity/regression approaches and are "doubly robust". This can also improve sample efficiency[1]

Many neural network architectures[2,3]! [This is where I started…]

[1] See e.g., Nie & Wager, *Biometrika*, 2021 who introduced the "R-learner", [2] Johansson et al., *ICML*, 2016,
[3] https://github.com/AliciaCurth/CATENets

Many stones left unturned

Knowing *all* the direct causes of treatment *and* being able to measure them is often infeasible

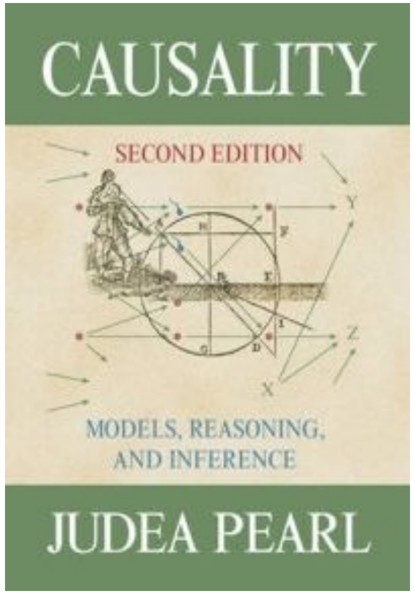
If we leave out some, there may be **unobserved confounders**

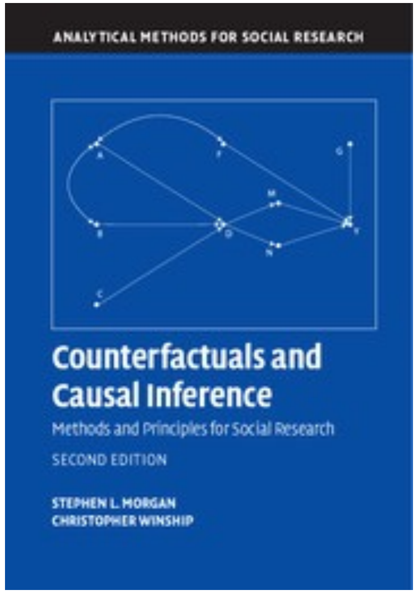
Our primary options are:

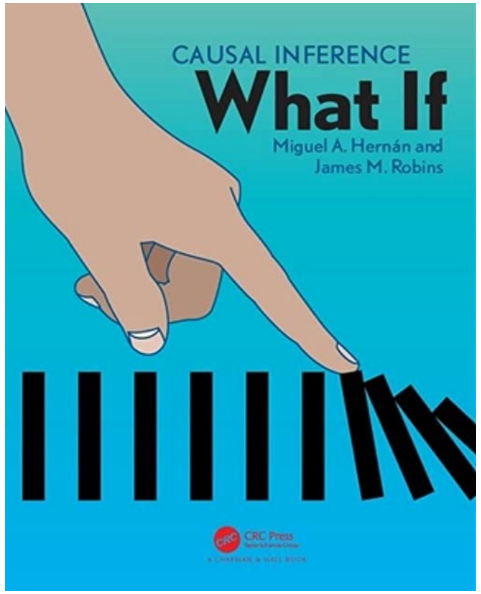
1. Look for other criteria for identifiability (e.g., backdoor criterion)
2. Aim only for partial identifiability (e.g., using sensitivity analysis)

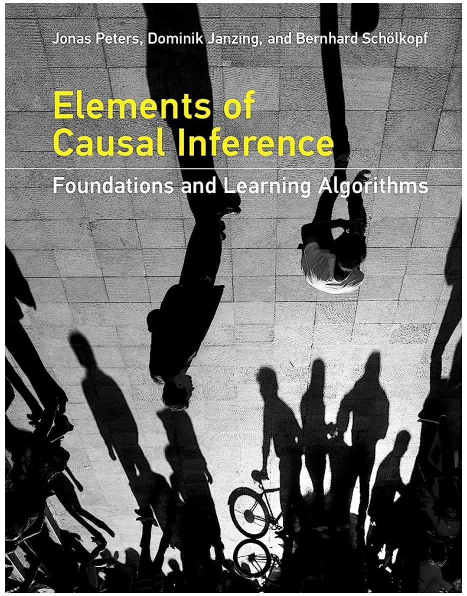
Going further

Causal machine learning is relatively young, but there is already a rich related literature—most of which was not covered today!



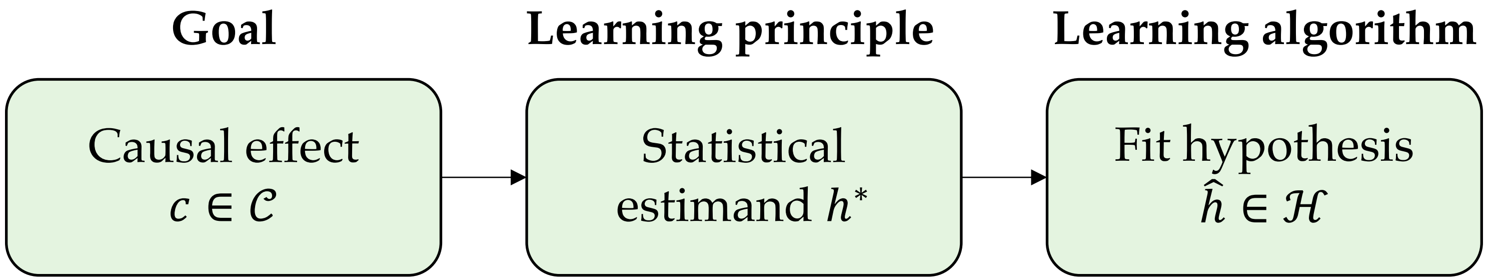

ANALYTICAL METHODS FOR SOCIAL RESEARCH
Counterfactuals and Causal Inference
Methods and Principles for Social Research
SECOND EDITION
STEPHEN L. MORGAN
CHRISTOPHER WINSHIP




Jonas Peters, Dominik Janzing, and Bernhard Schölkopf
Elements of Causal Inference
Foundations and Learning Algorithms

If you know that you will use your ML model to make decisions, make use of this fact!

Causal analysis is used to pick the right learning goal


Goal
Learning principle
Learning algorithm
Causal effect $c \in \mathcal{C}$
Statistical estimand h^*
Fit hypothesis $\hat{h} \in \mathcal{H}$

Identification **Estimation**

fredrik.johansson@chalmers.se

www.healthyai.se