

Nordic Probabilistic AI School

Causal Machine Learning

Fredrik Johansson
June 16, 2023

Why did my business fail?

What if I have surgery?

If I didn't do a physics BSc, would I still be a professor?

Which genes affect which traits?

Part I

Causal machine
learning

Part II

Coding example

Part III

Reflections

IncomeSim

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

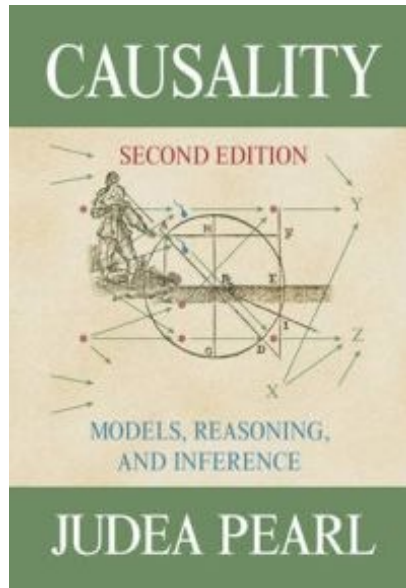
Based on the well-known Adult dataset from UCI

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!

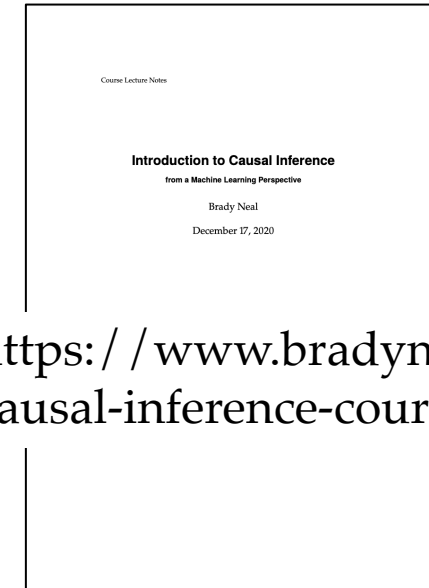
Causality
Pearl, 2009



Counterfactuals and
Causal Inference (2nd Ed)
Morgan & Winship, 2014



Brady Neal's
Causal inference course

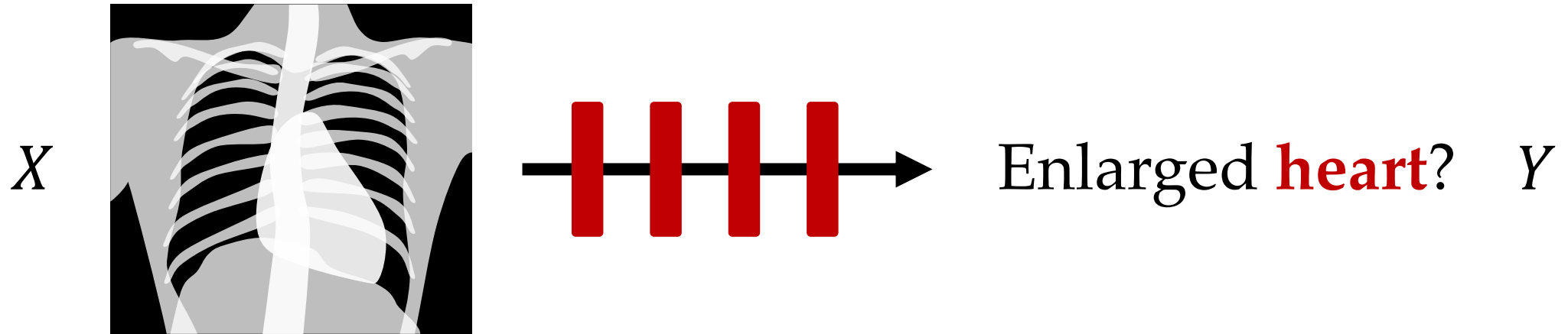


[https://www.bradyneal.com/
causal-inference-course](https://www.bradyneal.com/causal-inference-course)

Part I: Causal machine learning

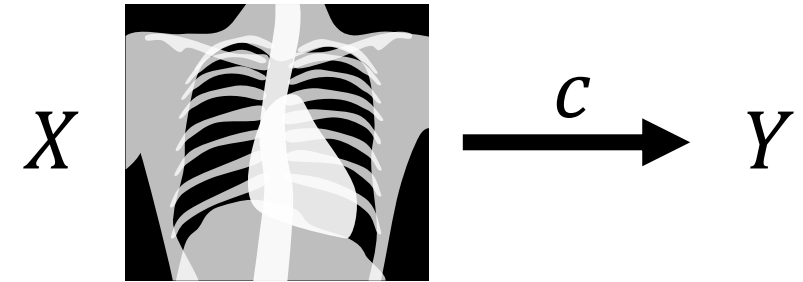
Part I.a: Supervised learning recap

ML example: Chest X-ray classification



Supervised machine learning

Supervised learning is the task of finding a **concept** $c : \mathcal{X} \rightarrow \mathcal{Y}$ which maps an **input** $X \in \mathcal{X}$ to an **outcome** $Y \in \mathcal{Y}$



Example (cont.)

- $\mathcal{X} \subseteq \mathbb{R}^{w \times h}$ is the set of possible chest X-ray images, X is an image
- $\mathcal{Y} = \{1, \dots, K\}$ is a set of K possible diagnoses, Y is a diagnosis
- c is the labelling “function” of a trained radiologist

¹C.f. Mohri, FoML, Chapter 2.1

How do we learn?

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

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

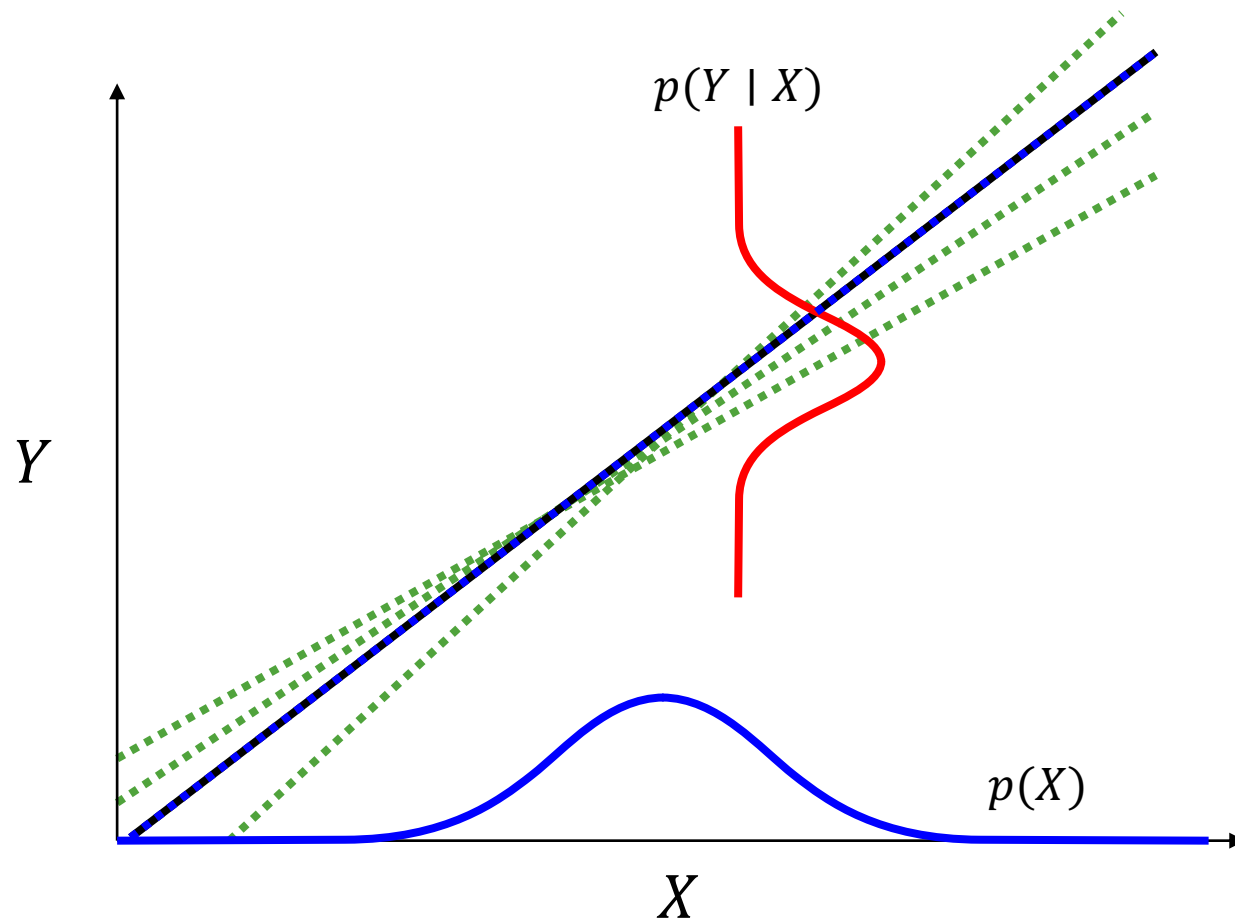
and $L : Y \times Y \rightarrow \mathbb{R}_+$ is a **loss function**

R is called “risk” or “generalization error”

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

* 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

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

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

Empirical risk minimization (ERM)

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

The “training error”

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

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

In ML, we care about sample generalization

From empirical risk (training error) to population risk (test error)

How much larger is $R(h)$ than $\hat{R}(h)$?

How much larger is $R(\hat{h})$ than $R(h^*)$?

The differences measure the **generalization error**

Here, h^* is the distribution-optimal model: $h^* = \arg \min_h R(h)$

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}

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

The test error tells us something about our model's performance

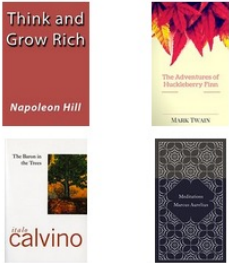







Is this answering the right questions?

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?

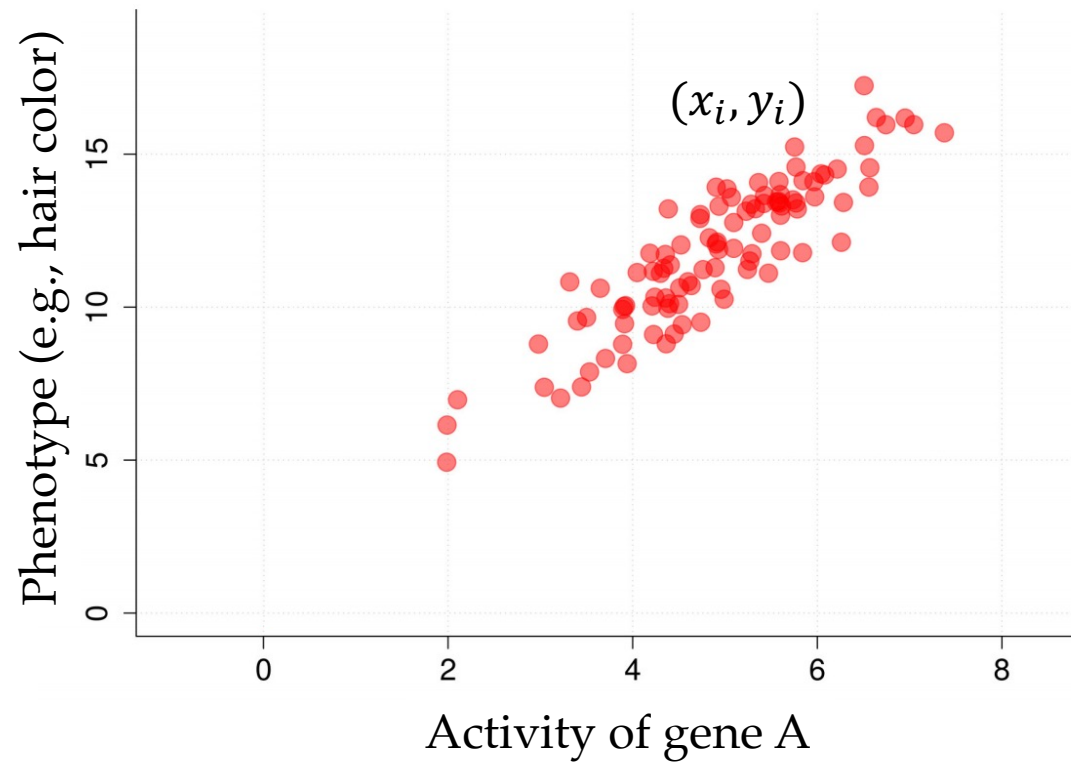
Are these answered by the test error?

If we recommend it, will you buy it?

 <p>Literature & Fiction 62 ITEMS</p>	 <p>Exercise & Fitness Equipment 8 ITEMS</p>	 <p>Health, Fitness & Dieting Books 37 ITEMS</p>	 <p>Tableware 12 ITEMS</p>
 <p>Prime Video – Unlimited Streaming for Prime Members 12 ITEMS</p>	 <p>Coffee, Tea & Espresso 98 ITEMS</p>	 <p>Biographies & Memoirs 17 ITEMS</p>	 <p>Engineering Books 7 ITEMS</p>

Part I.b: Machine learning & causality

Example: Gene knock out



Example from Jonas Peters.

Genetic data

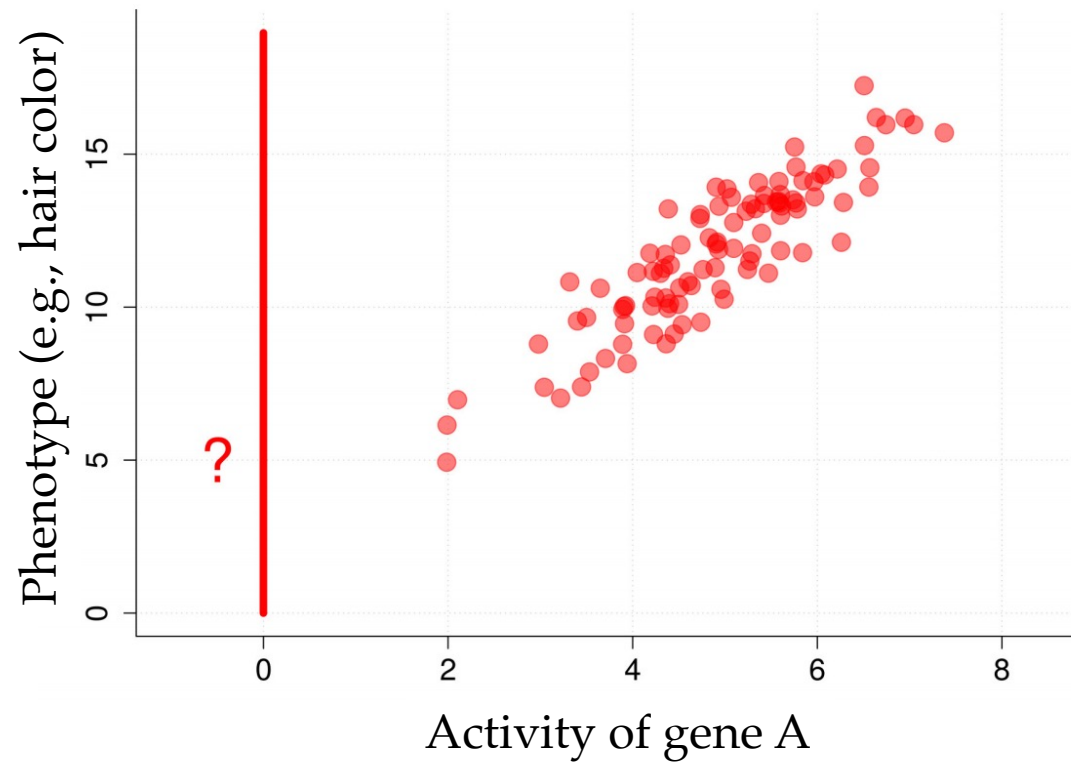
Assume we have sequenced the **genome** of 100 individuals drawn from a population p

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

Can we influence hair color?

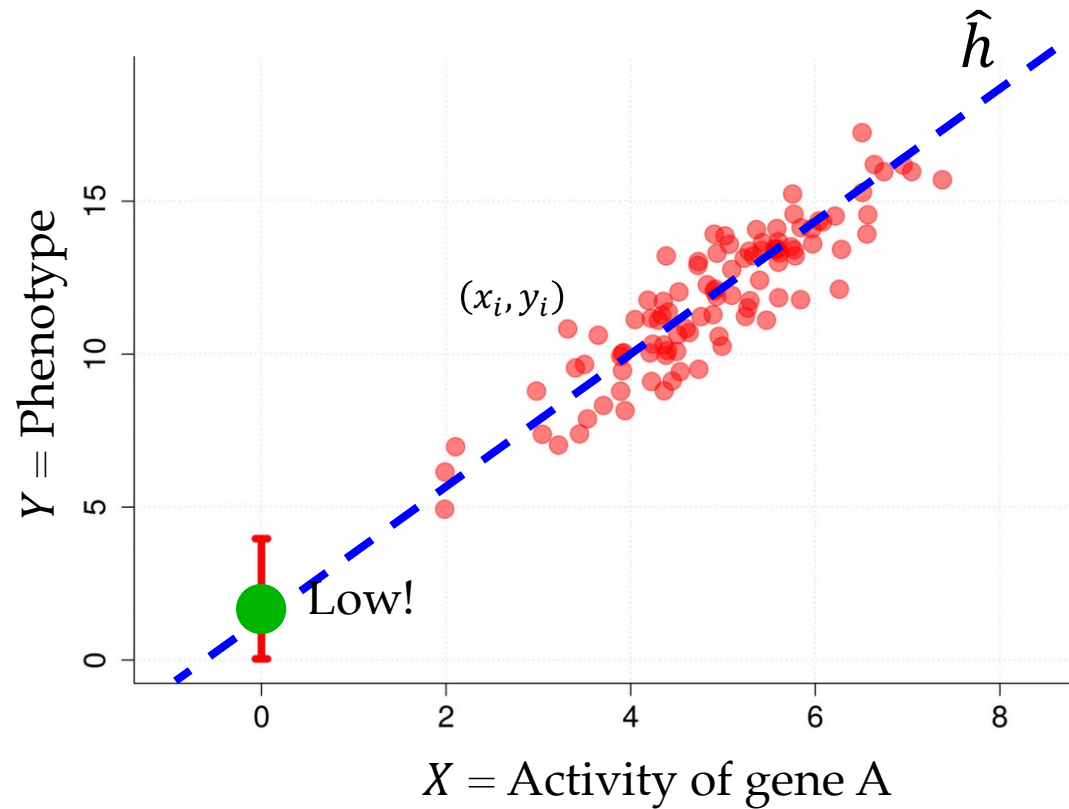
Example: Gene knock out

What if we “knock out” Gene A (**intervene** to set its activity to 0)?



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

Out-of-the-box ML solution



Example from Jonas Peters.

Step 2.

Pick a learning principle

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

Step 3.

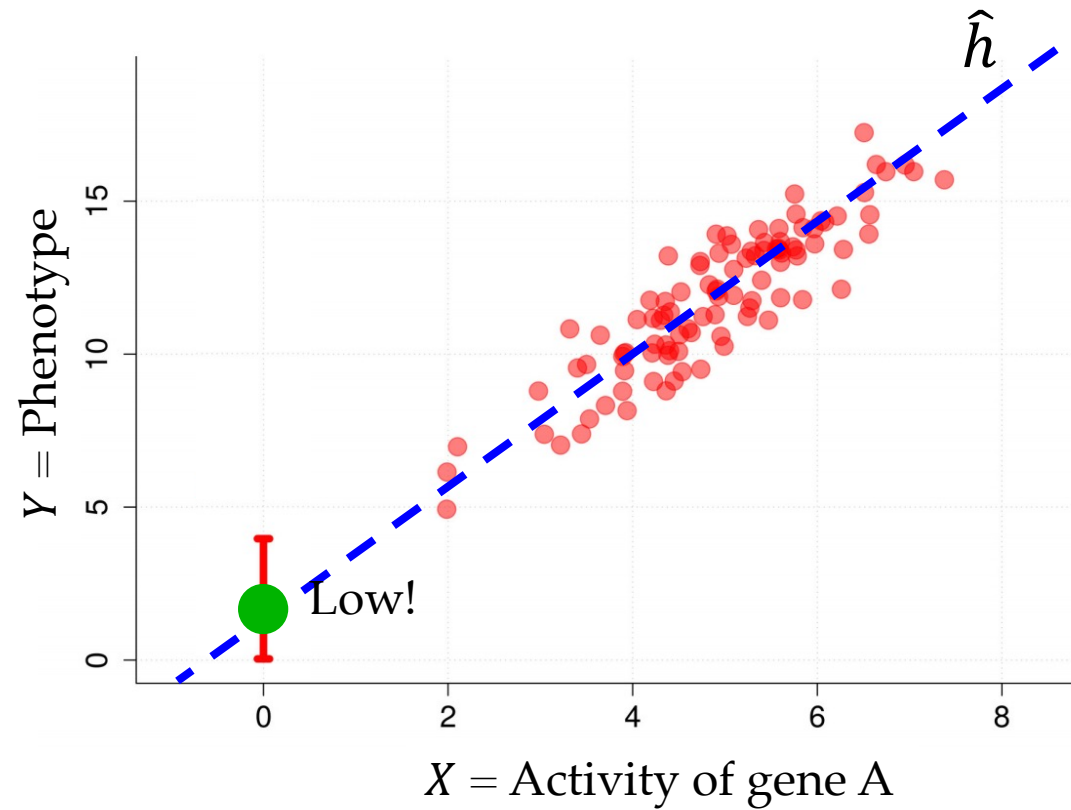
Learn a hypothesis from data

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

Step 4.

Evaluate test error. Good fit!

ML would probably do this



Question:
Is this correct?
How do you know?

Example from Jonas Peters.

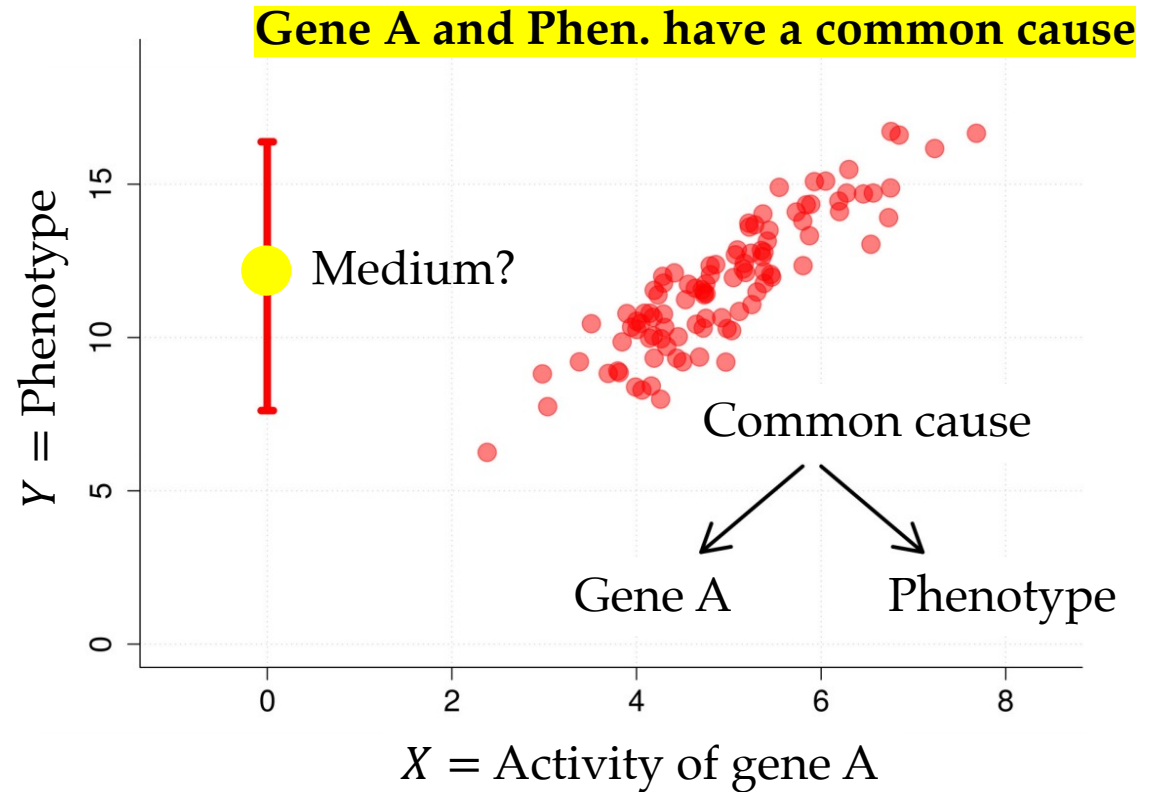
If there is a **common cause** which influences both gene activity X and phenotype Y

... the phenotype may not change when we knock out Gene A!

The association is **confounded**!

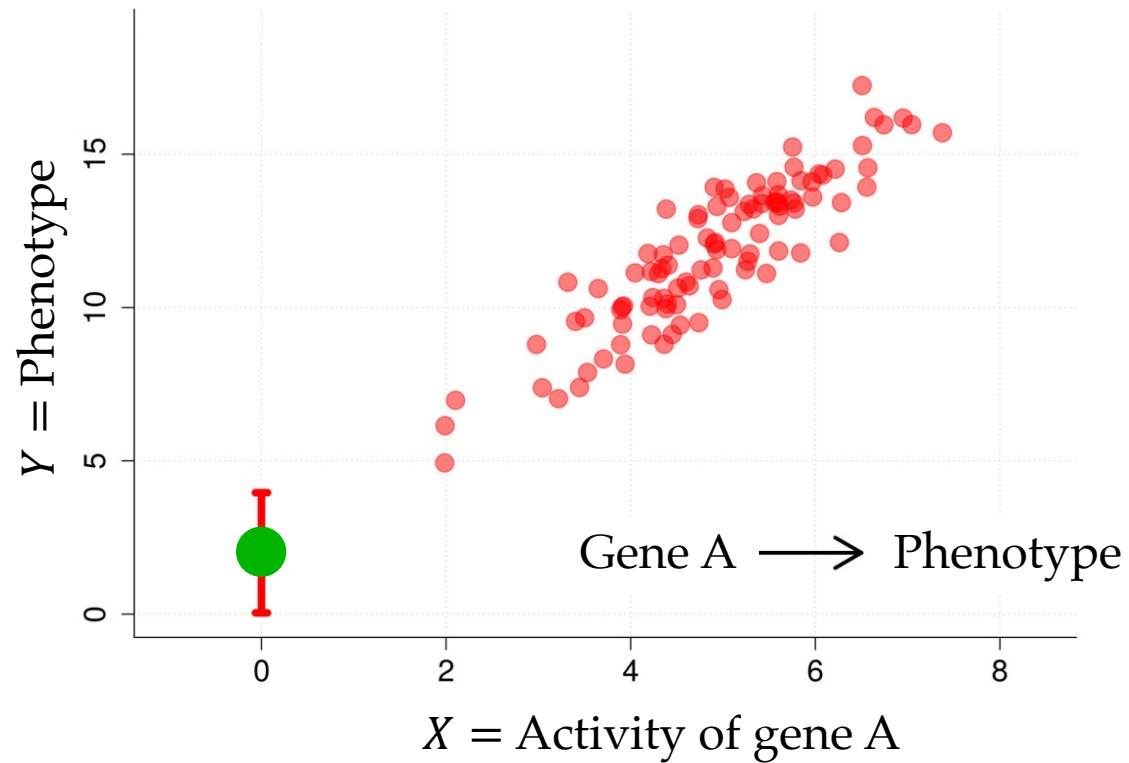
Example from Jonas Peters.

Explanation B



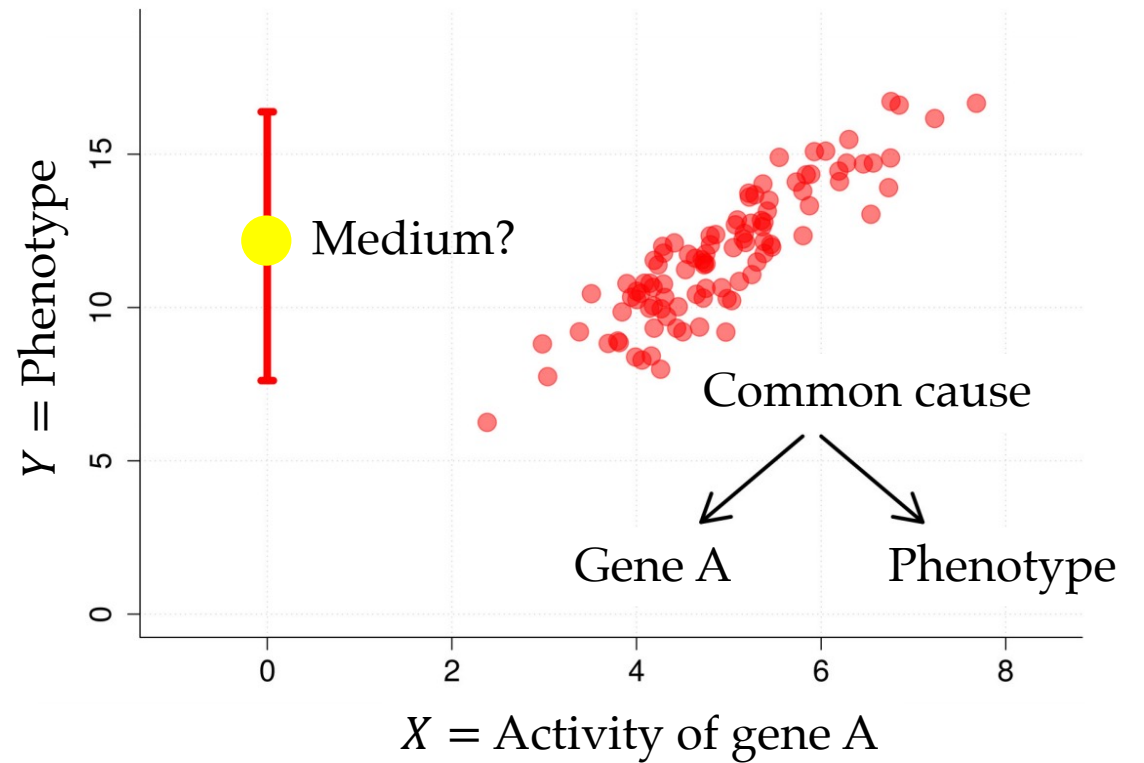
Explanation A

Gene A causes phenotype



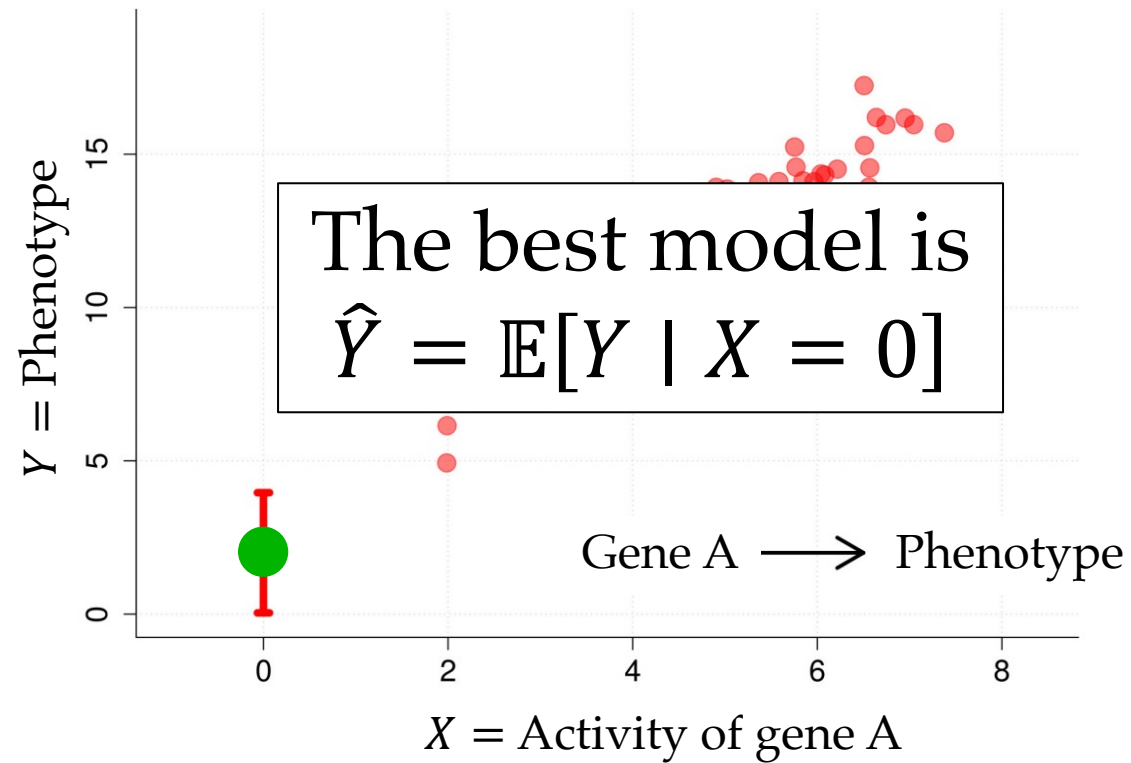
Explanation B

Gene A and Phen. have a common cause



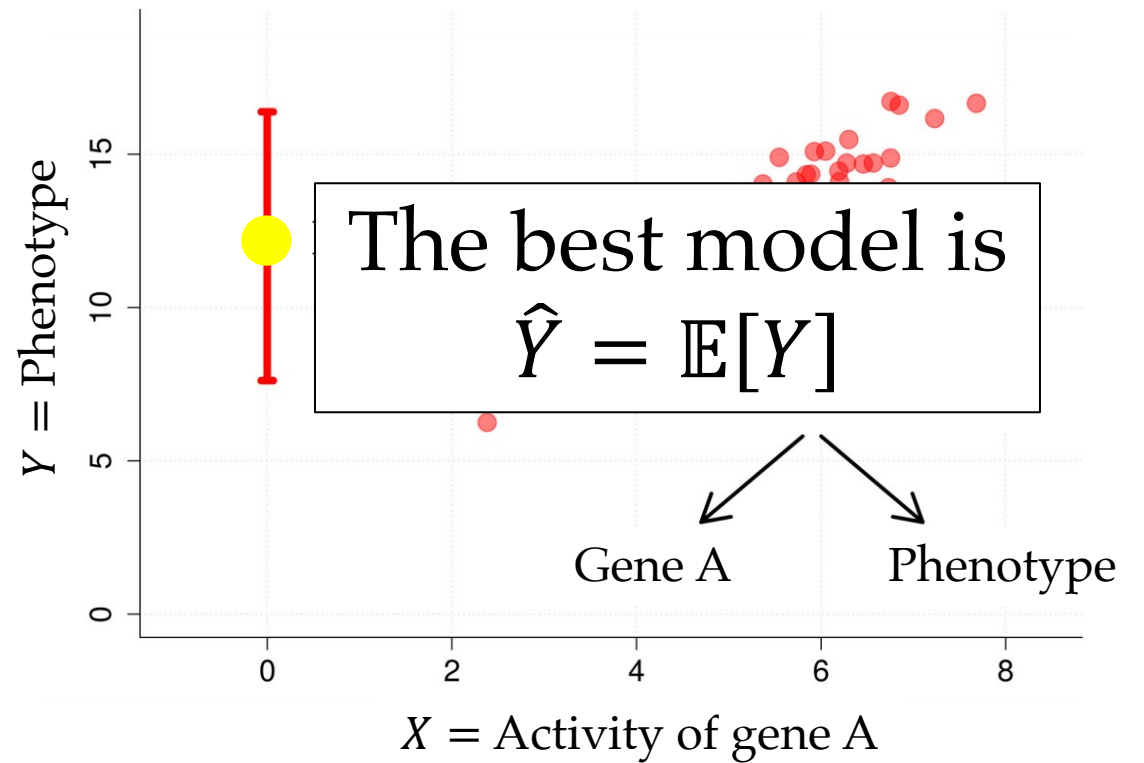
Explanation A

Gene A causes phenotype



Explanation B

Gene A and Phen. have a common cause



Causation \neq association

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

$$\mathbb{E}[Y \mid X = 0] \neq \mathbb{E}[Y] \quad \text{in general}$$

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

Even a perfect regression $f(x) \approx y$ will get the wrong answer!

Identification in causal ML

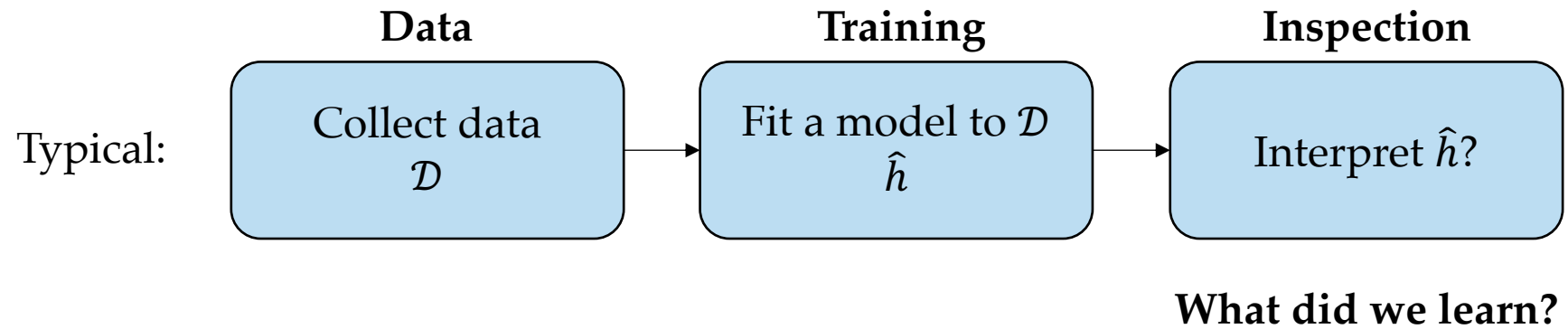
In causal machine learning, there are two “systems”

- The system where we observe data, e.g., $p(X, Y)$
- The system after intervention, e.g., $p_{X \leftarrow x}(Y)$

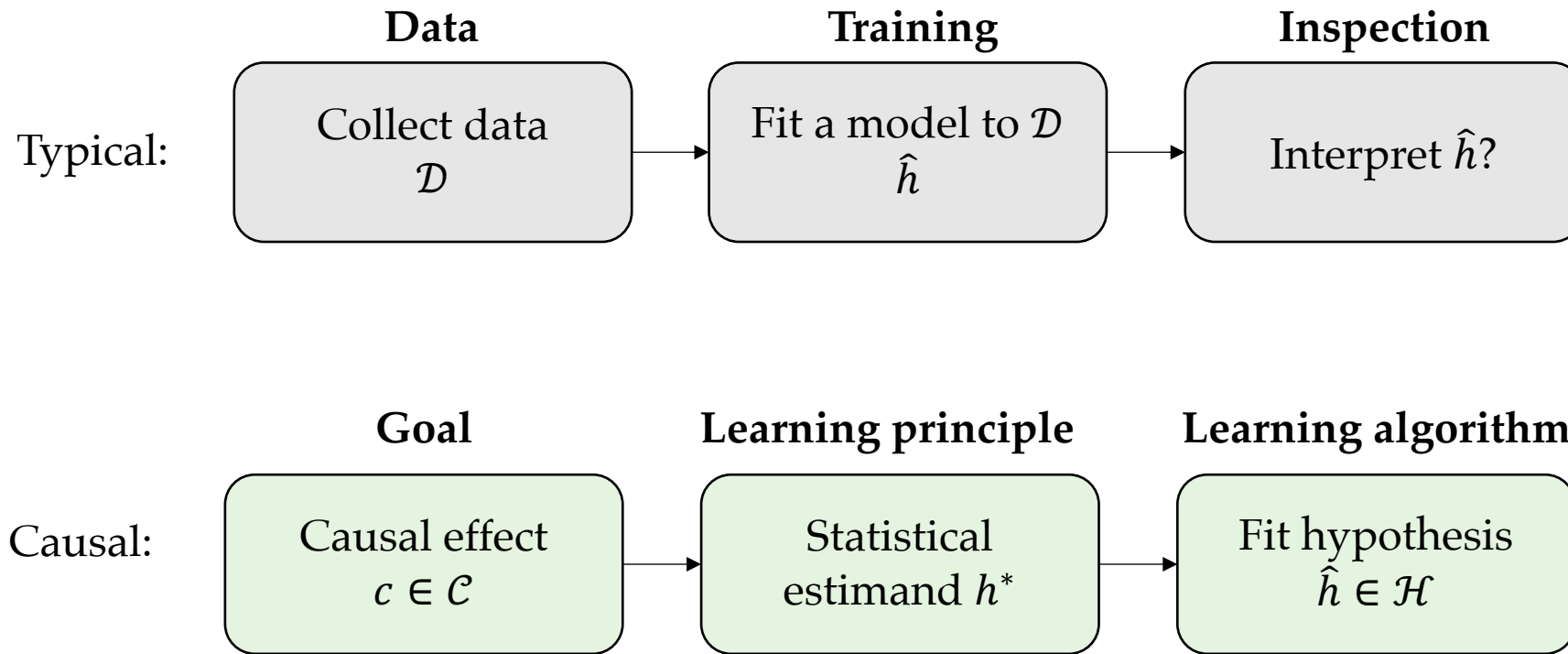
The goal of **causal identification** is to connect the two:

What does the data-generating system say about
the system after intervention?

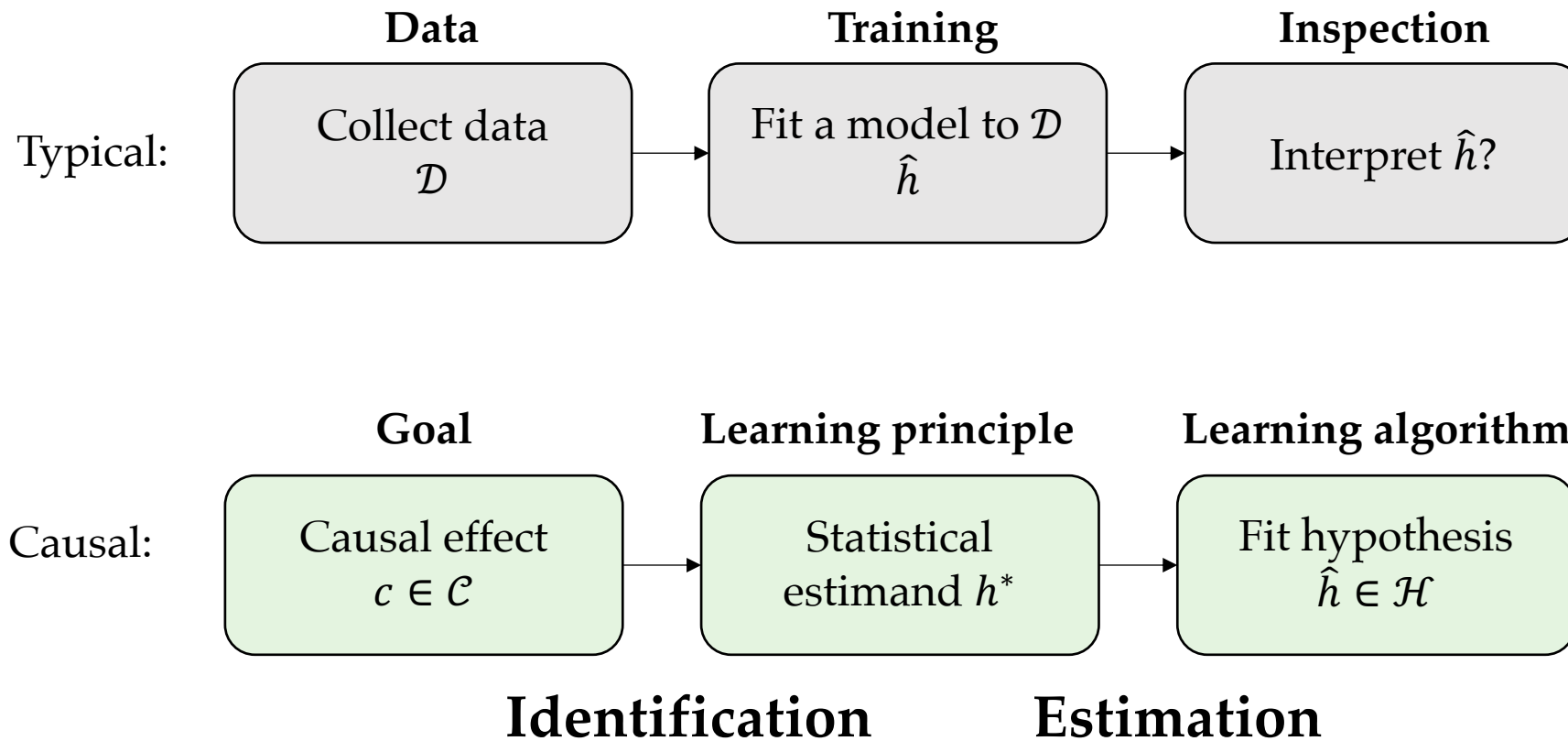
A common ML pattern...



Causal machine learning should be different!



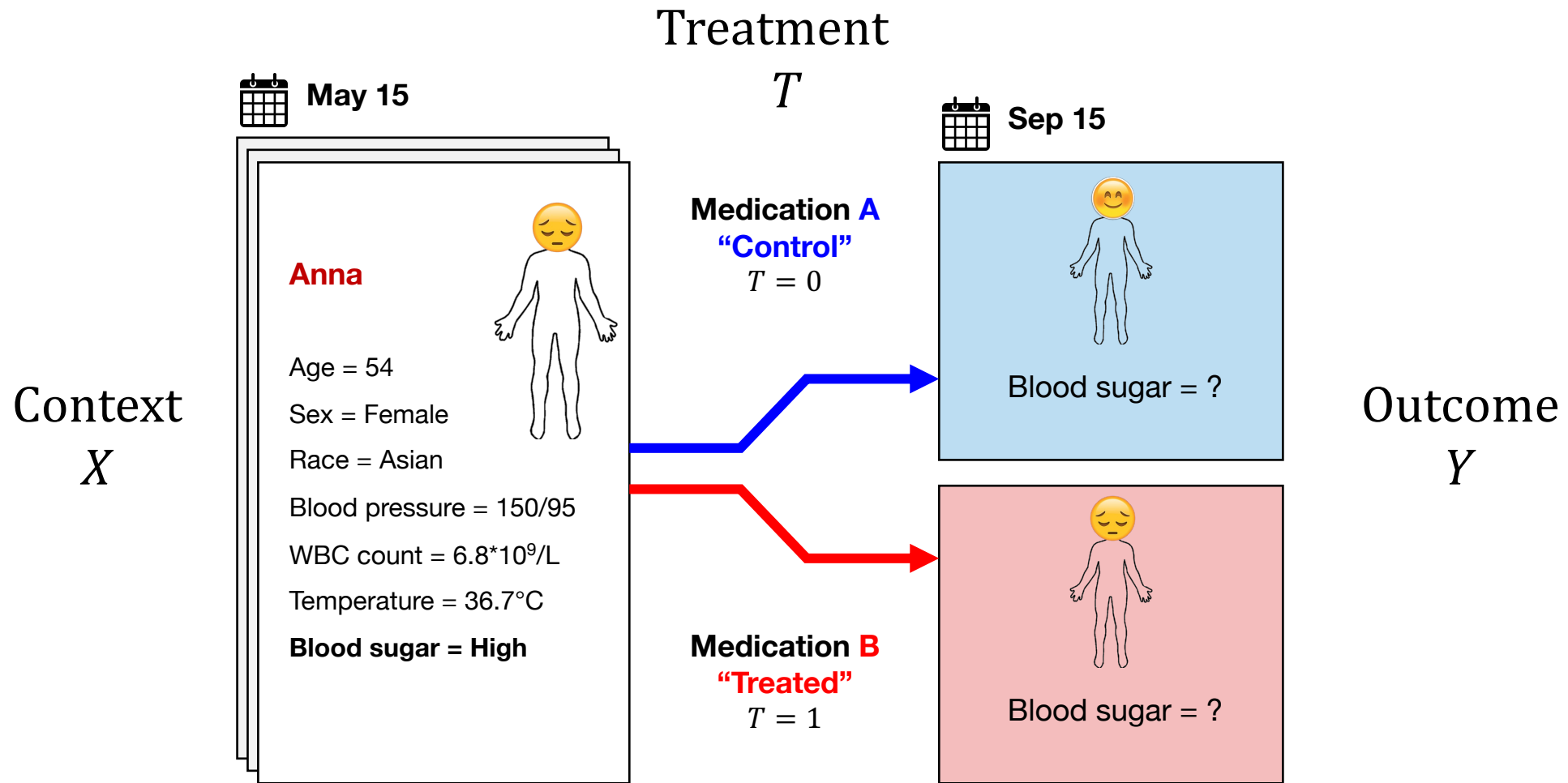
Causal machine learning is different!

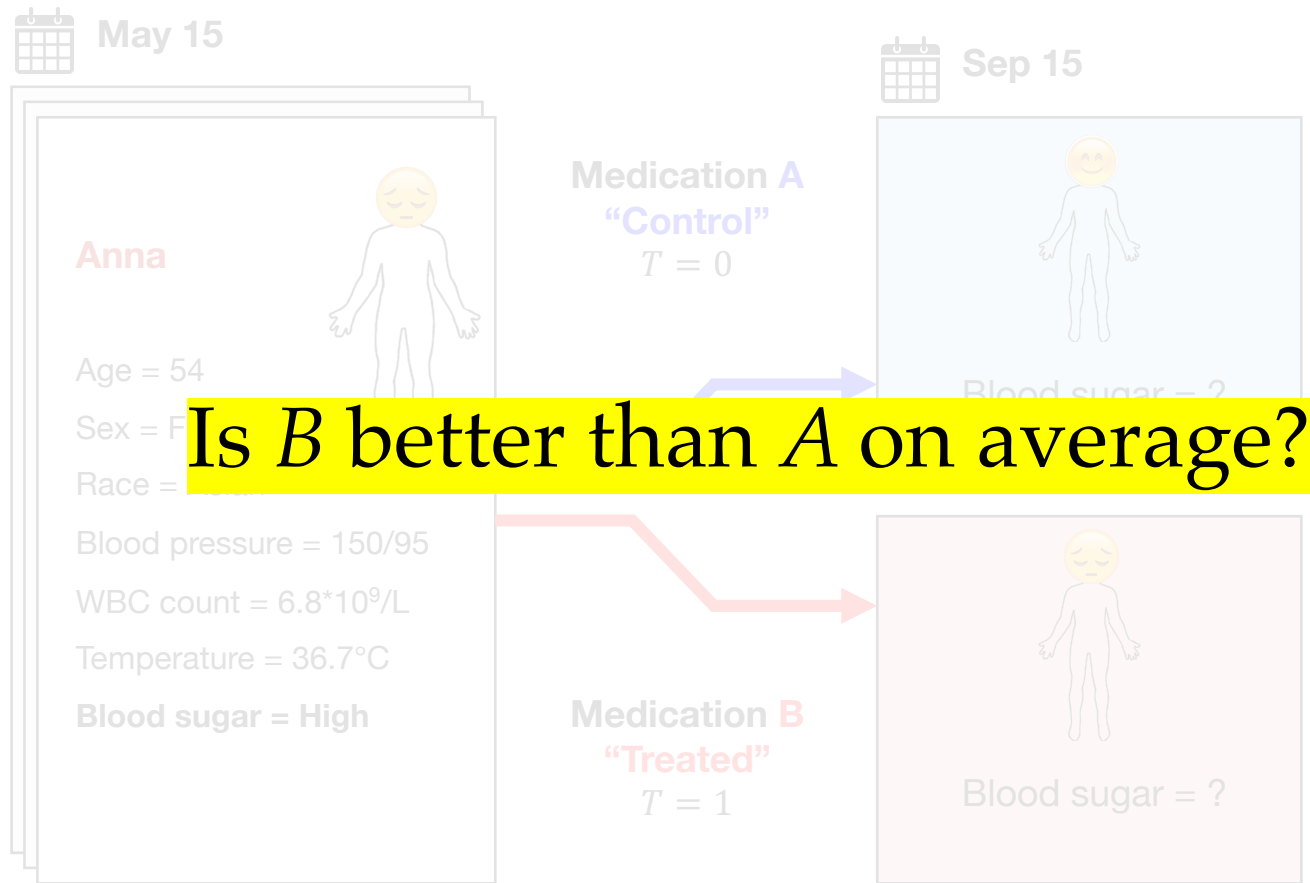


We need a way to *specify causal questions*
to identify the *right goal* for machine learning

Part I.c: Potential outcomes of interventions

See e.g., Hernán & Robins, Chapter 1

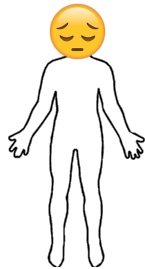




We can imagine two **potential** scenarios for our patients

 May 15

Anna



Age = 54

Sex = Female

Race = Asian

Blood pressure = 150/95

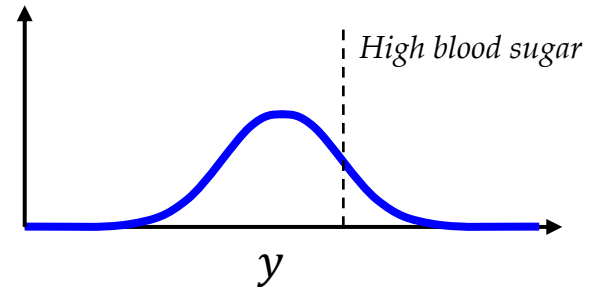
WBC count = $6.8 \times 10^9/L$

Temperature = $36.7^\circ C$

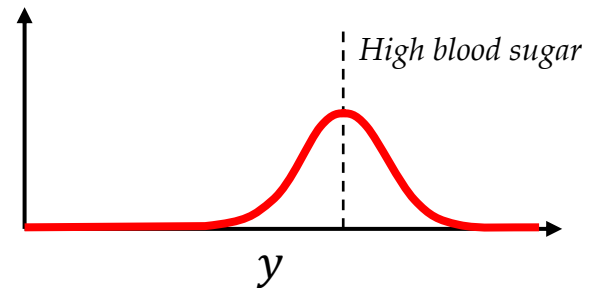
Blood sugar = High

Medication A
“Control”
 $T = 0$

 Sep 15



Medication B
“Treated”
 $T = 1$

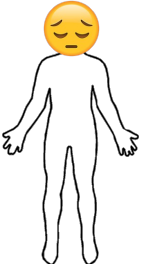


Which is better?

Two “potential” random variables

May 15

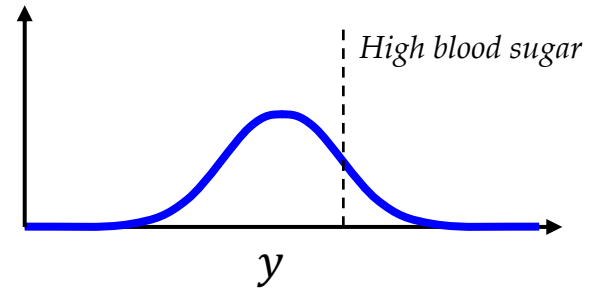
Anna



Age = 54
Sex = Female
Race = Asian
Blood pressure = 150/95
WBC count = $6.8 \times 10^9/\text{L}$
Temperature = 36.7°C
Blood sugar = High

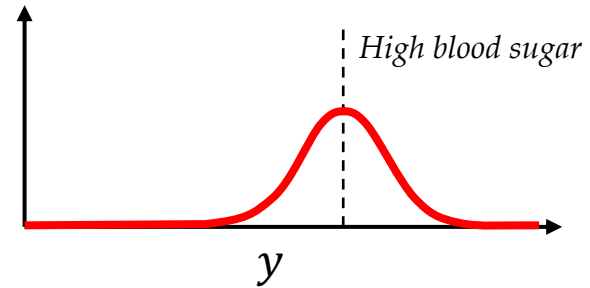
Medication A
“Control”
 $T = 0$

Sep 15



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

Medication B
“Treated”
 $T = 1$



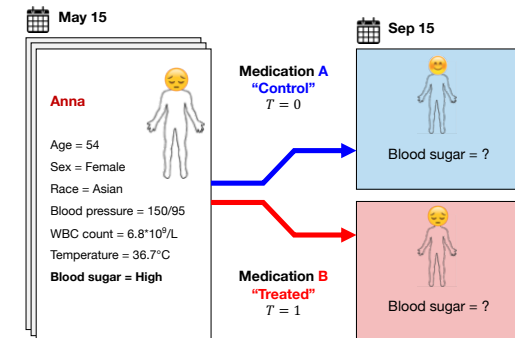
$$Y(1) \sim p_1(Y)$$

Potential outcomes¹ (Neyman-Rubin model)

$Y(0)$: What *would* happen under treatment 0 (“control”)

$Y(1)$: What *would* happen under treatment 1 (“treatment”)

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



¹See e.g., Hernan & Robins, Chapter 1

Causal effects

Potential outcomes lets 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 $Y(0), Y(1)$

*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 (to predict) Δ ?

Observational studies: Using historical data

Example:

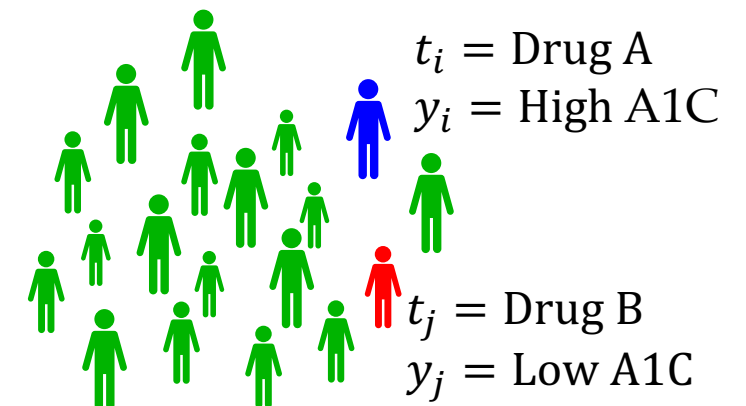
We have been treating pre-diabetic patients for 10 years

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

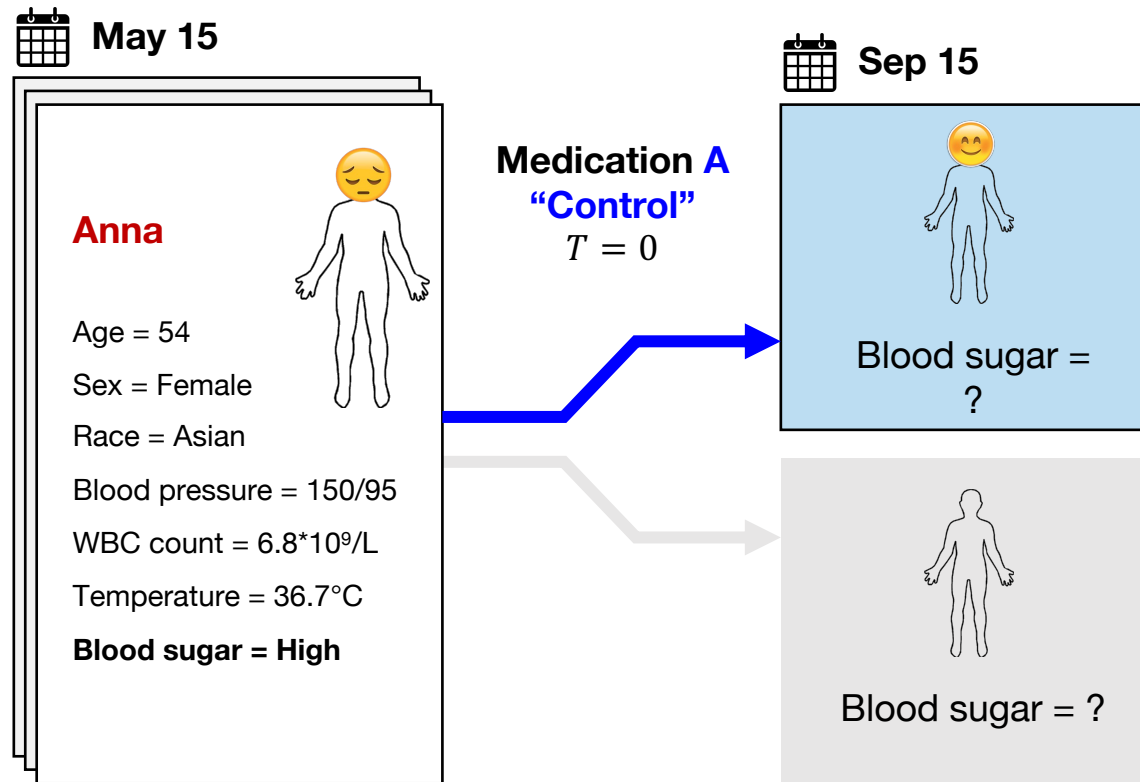
Dataset D

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



Fundamental problem of causal inference

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



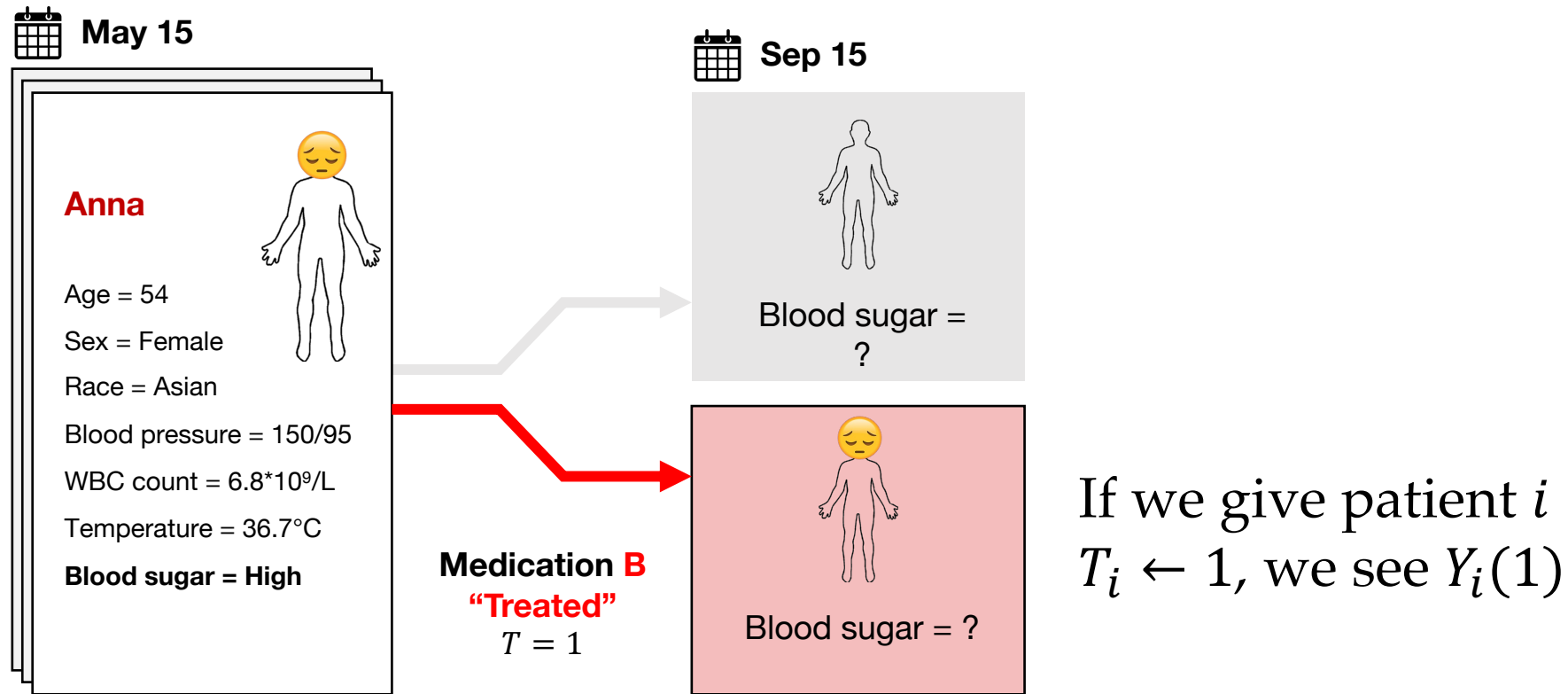
*If we give subject i
 $T_i \leftarrow 0$, we see $Y_i(0)$

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

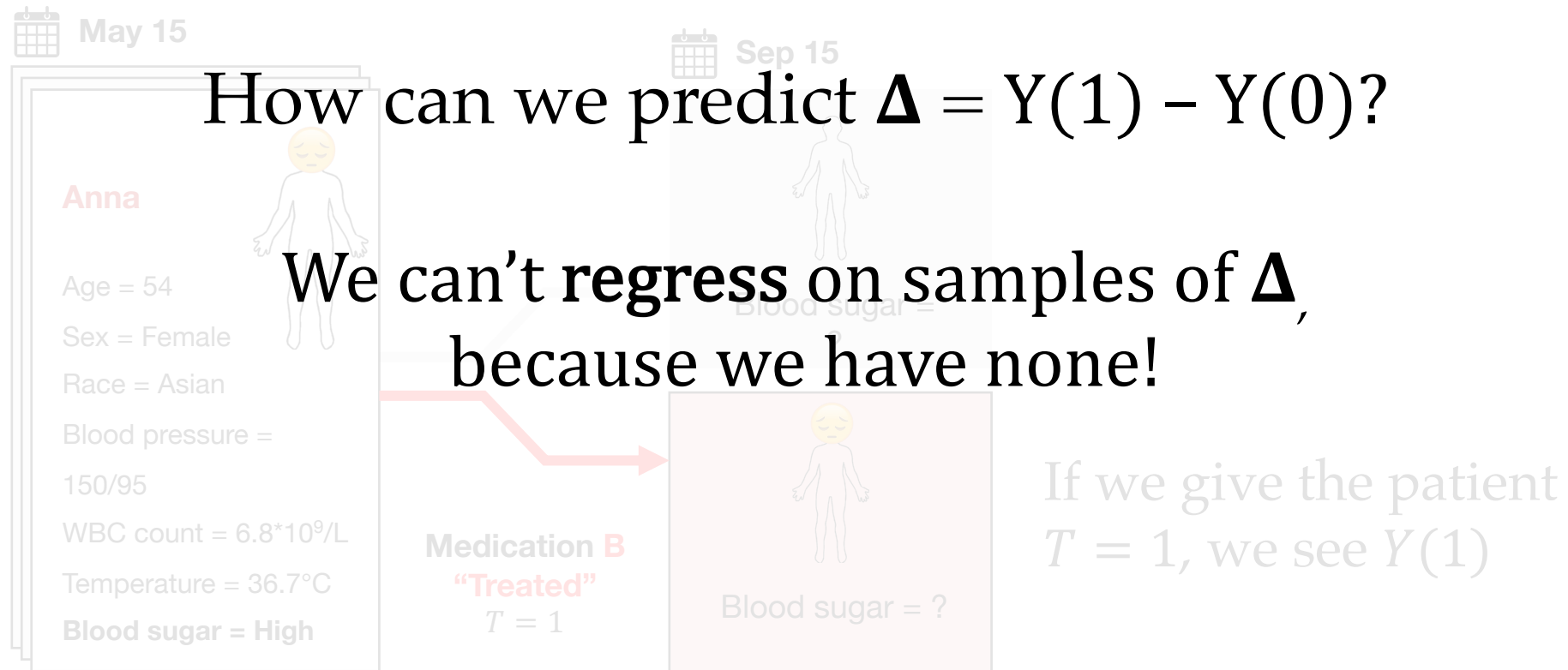
We can only observe the outcome of **one intervention**



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

Fundamental problem of causal inference

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



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

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

The ATE averages over individual variation

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

Example: LaLonde's study (1986)

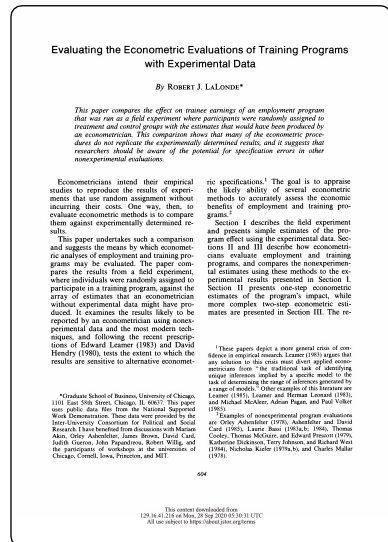
The LaLonde study looked at the effectiveness of a job training program, T , on the future real earnings of an individual, Y

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**

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

* More on this later. **The job training program cost \$6800–\$9100



Conditional average treatment effect

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

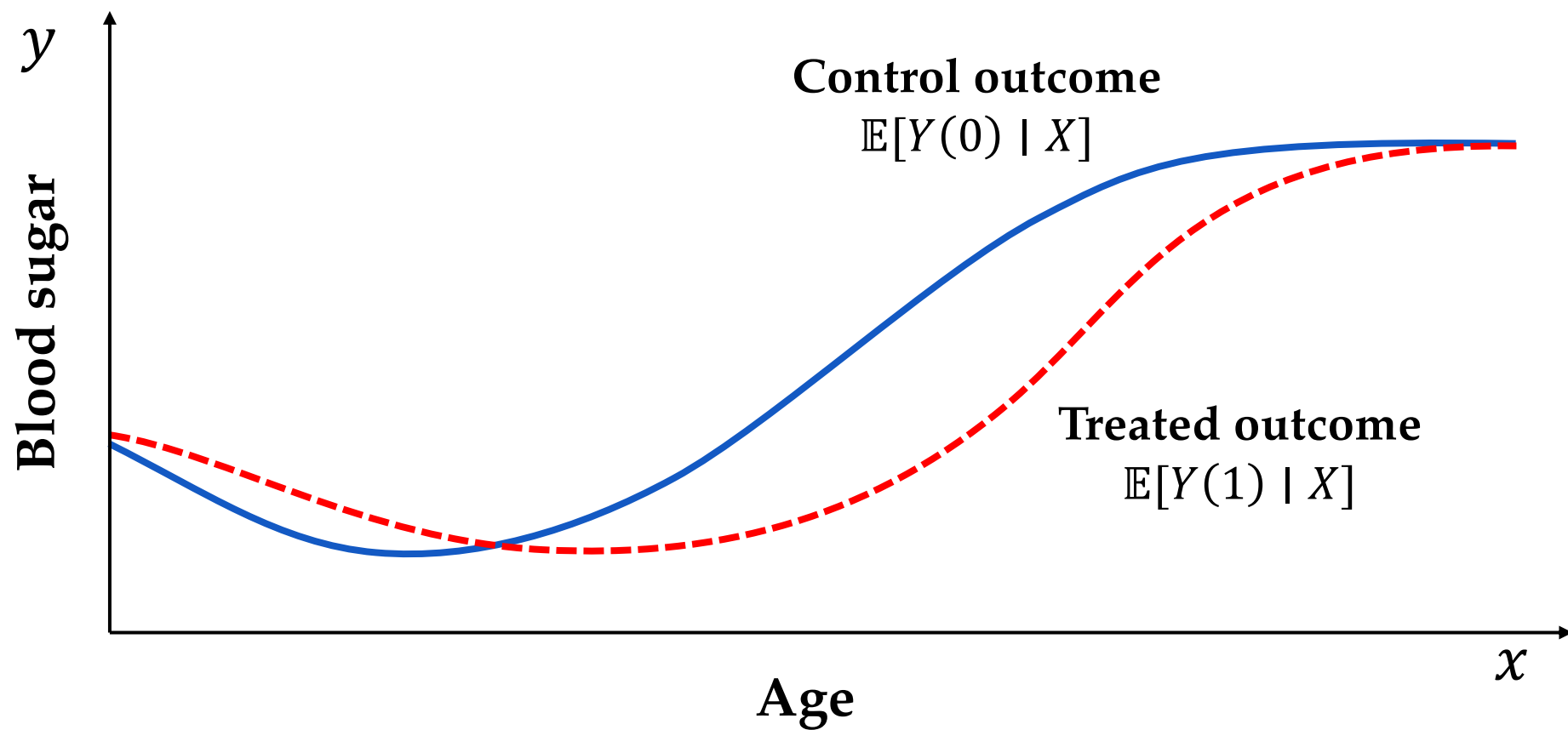
Conditional average treatment effect (CATE) w.r.t. X :

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

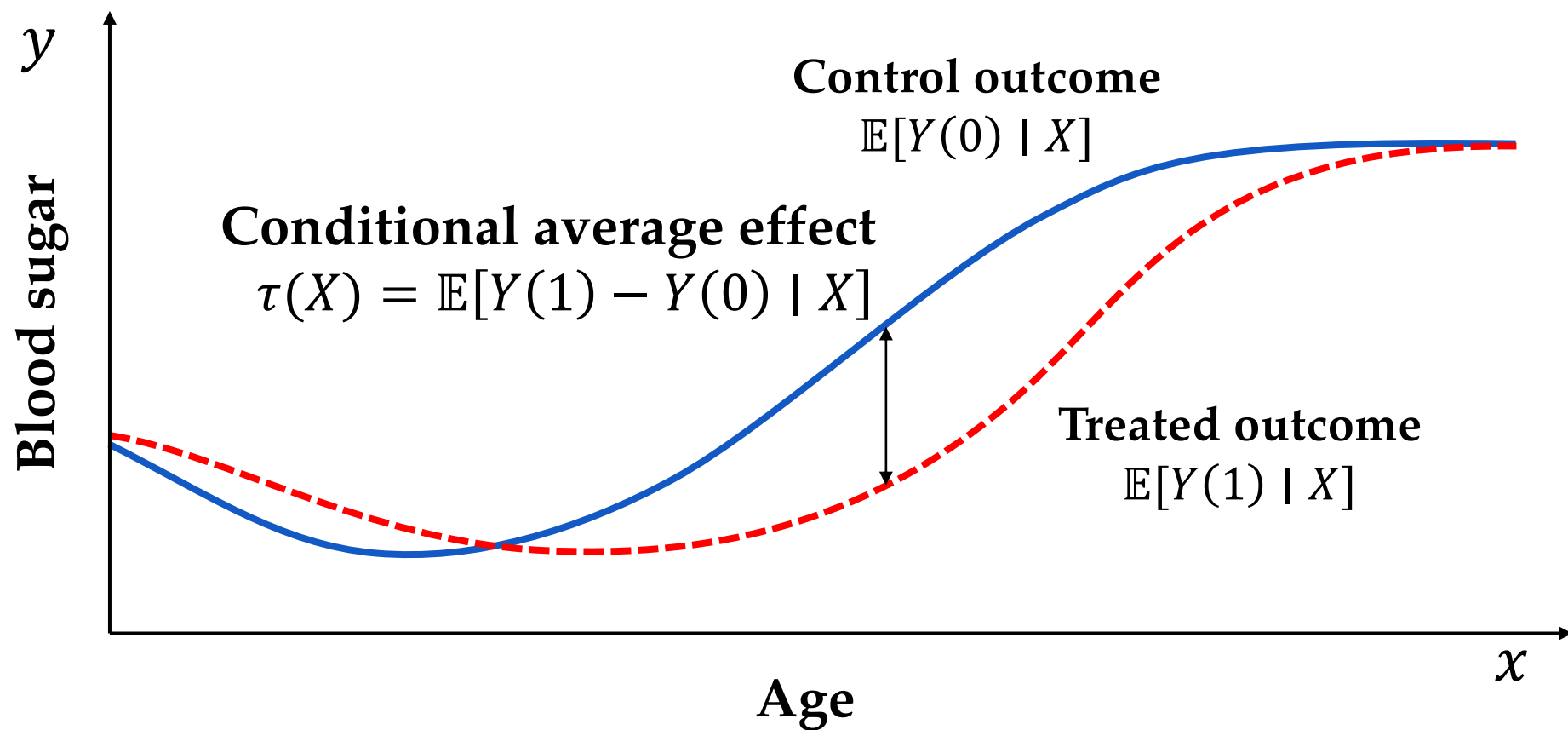
The ATE is the expected CATE, i.e., $\tau = \mathbb{E}_X[\tau(X)]$

* For consistency, “CACE” would have been more appropriate, but I rarely see that used.

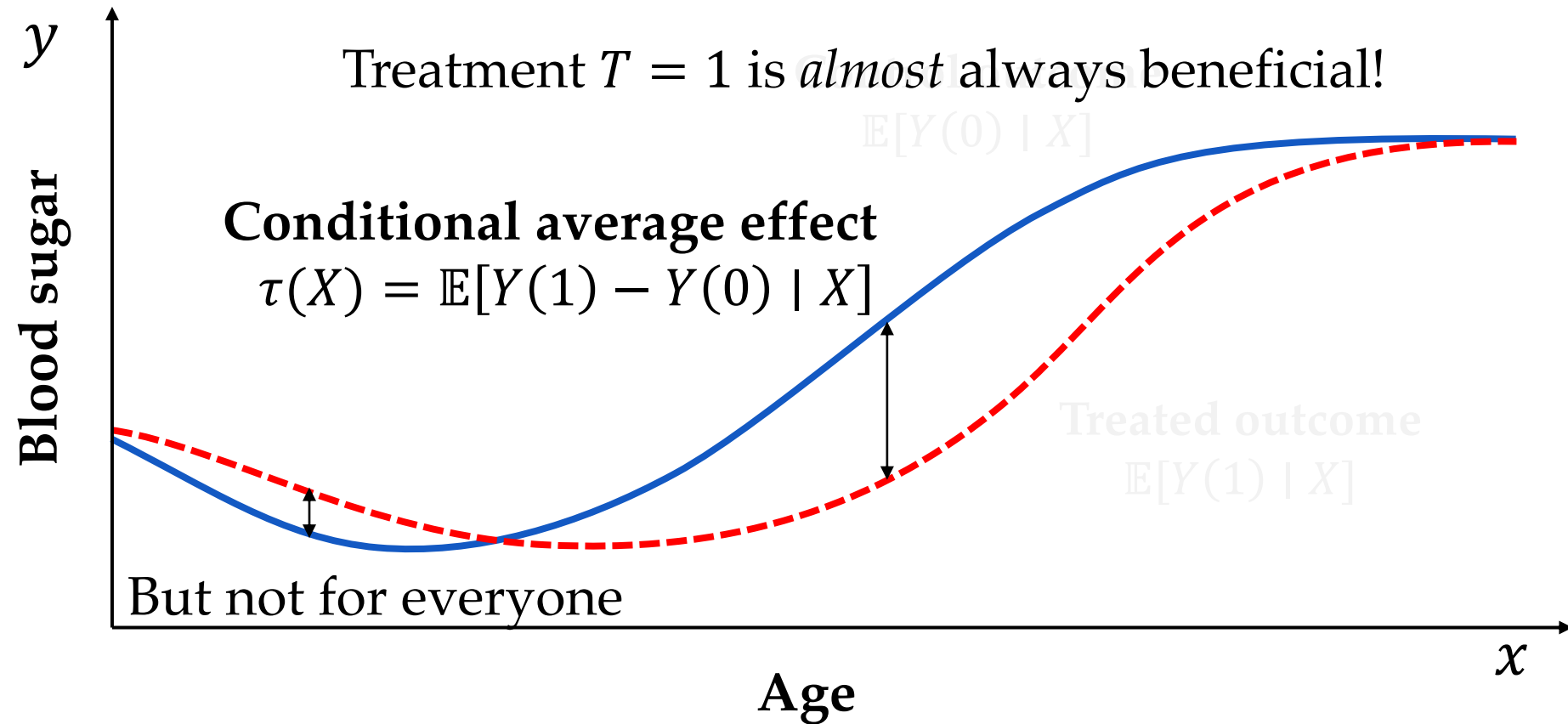
Potential outcomes and CATE



Potential outcomes and CATE



Potential outcomes and CATE



Part I.d: Partial observability & confounding

Confounding

Confounding bias comes from differences in the outcome which is not related to the treatment's effect itself

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

Everyone gets drug 1!

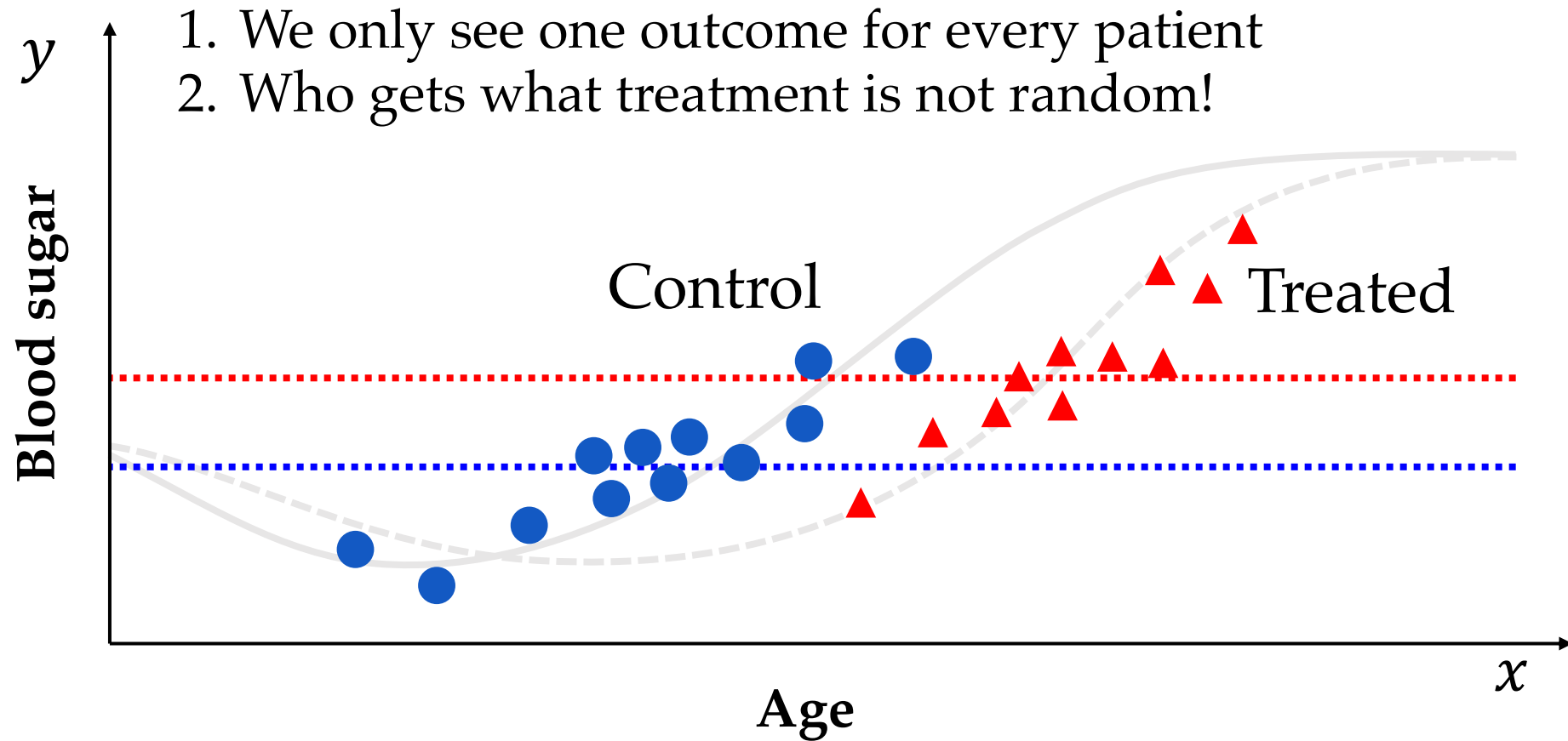
$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \neq \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$$

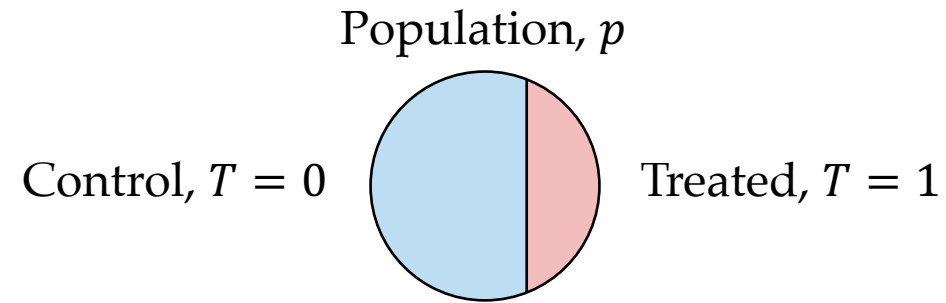
Interventional

Drug 1 is given to older patients!

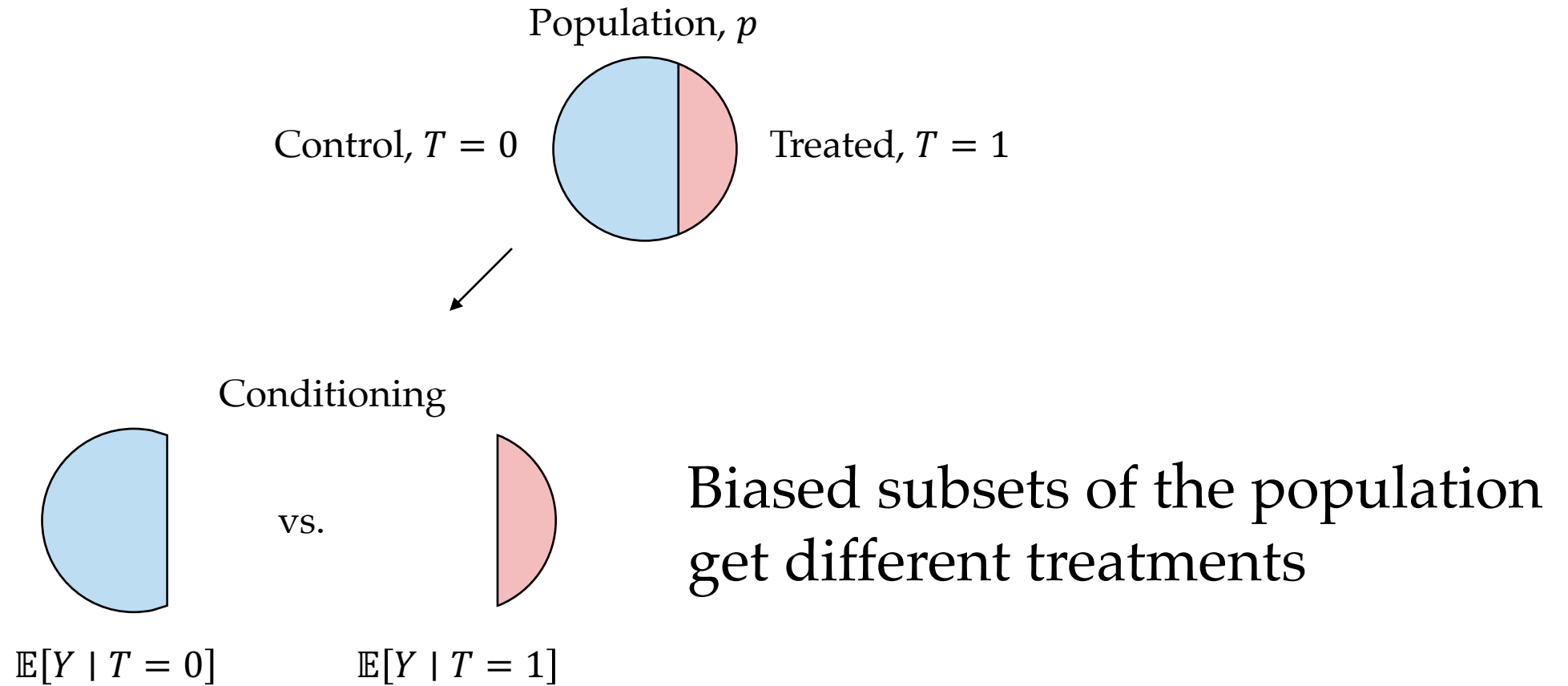
Conditional / Observational

Observational studies

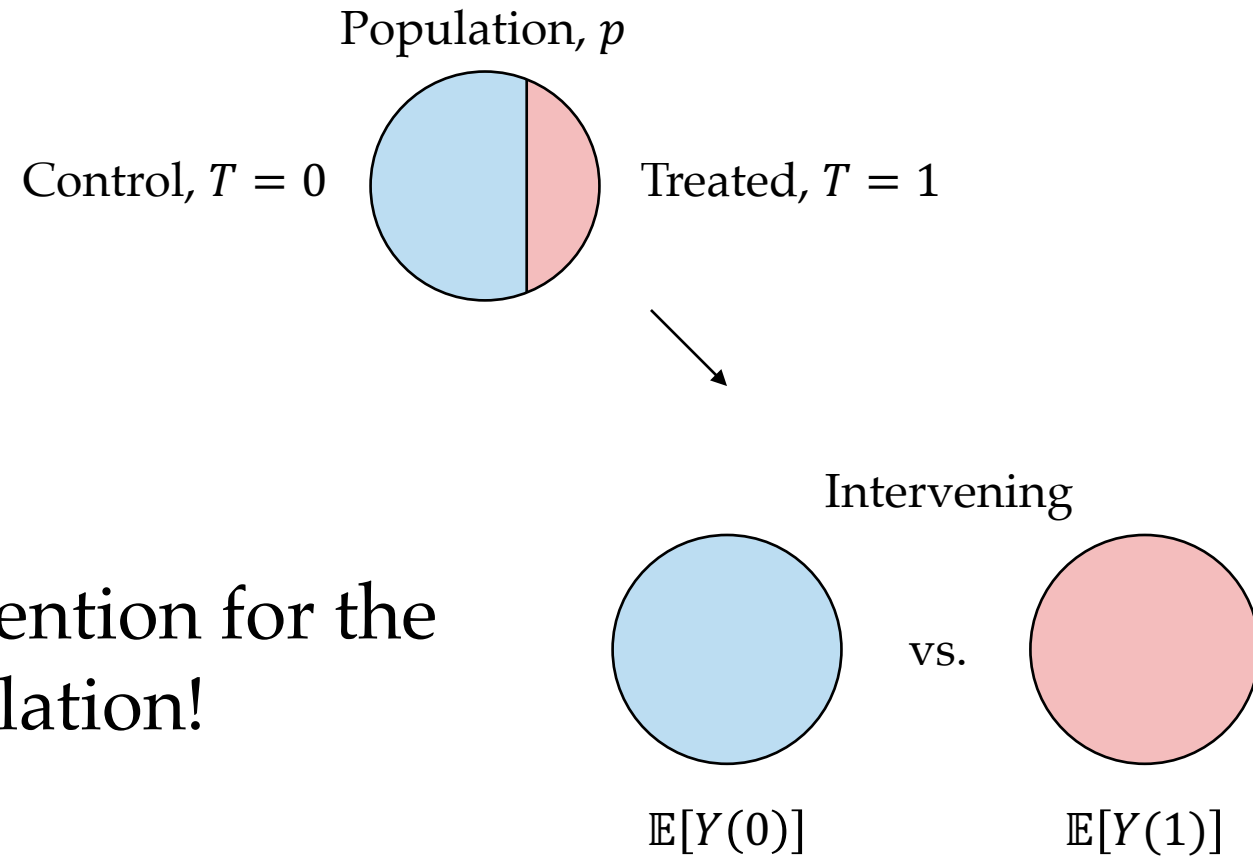




The discs represent
types of individuals,
the color their
treatment
status



Same intervention for the whole population!



Identifiability of causal effects

Identifiability of causal effects is to connect an

interventional quantity, e.g., $\mathbb{E}[Y(1)]$

with an

observational quantity, e.g., $\mathbb{E}[Y \mid T = 1]$

To do this, we must
make assumptions!

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,

$$Y(t) \perp T \text{ ——— } \textit{Treatment groups are "exchangeable"}$$

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

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

Randomized experiments & identifiability

Experiments ensure exchangeability through **randomization**

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

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

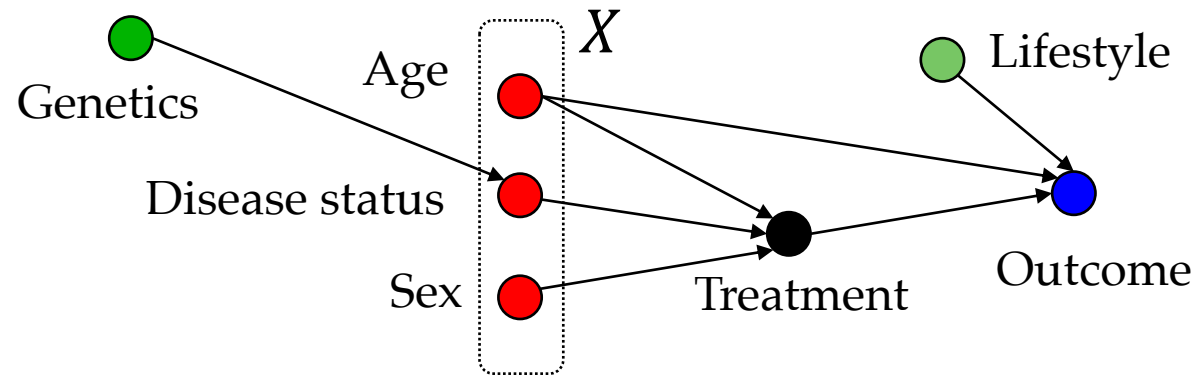
$$Y(t) \perp T \mid X$$

Conditional exchangeability given X

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

Parent adjustment

A special case is when we know all direct causes of treatment — all parents in the *causal graph**



The full set of parents X 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

Identifying assumptions

We will make the following **identifying assumptions**

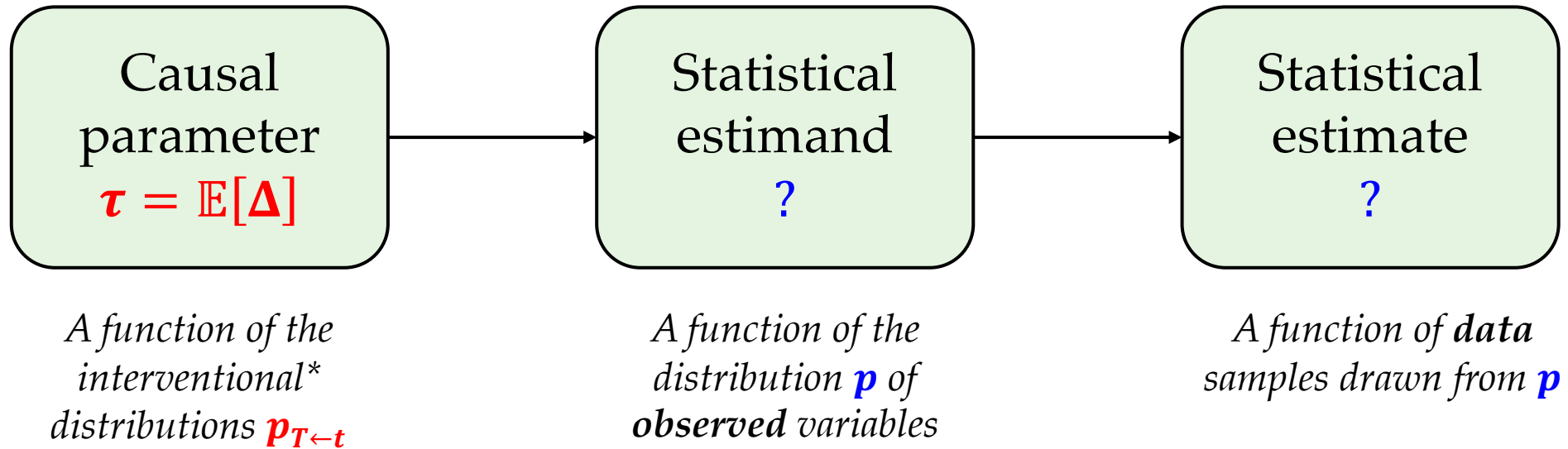
A set of variables X is known which satisfies, for all t and x :

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!

Identification

Estimation



* Or counterfactual

Two common methods

Regression adjustment

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

Propensity adjustment

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

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

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

Cond. exchangeability

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

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

Cond. exchangeability

Consistency

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

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

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

$$\begin{aligned} \text{Interventional} \quad \mathbb{E}[Y(t) \mid X = x] &= \sum_y y p(Y(t) = y \mid X = x) \\ &= \sum_y y p(Y \mid X = x, T = t) = \mathbb{E}[Y \mid X = x, T = t] \quad \text{Observational} \end{aligned}$$

Final steps

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

To get the ATE, we can average these conditionals

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}_X \left[\underbrace{\mathbb{E}[Y \mid X = x, T = 1]}_{\text{Can estimate with regression}} - \underbrace{\mathbb{E}[Y \mid X = x, T = 0]}_{\text{need overlap to get right!}} \right] \end{aligned}$$

Regression adjustment with “T-learner”*

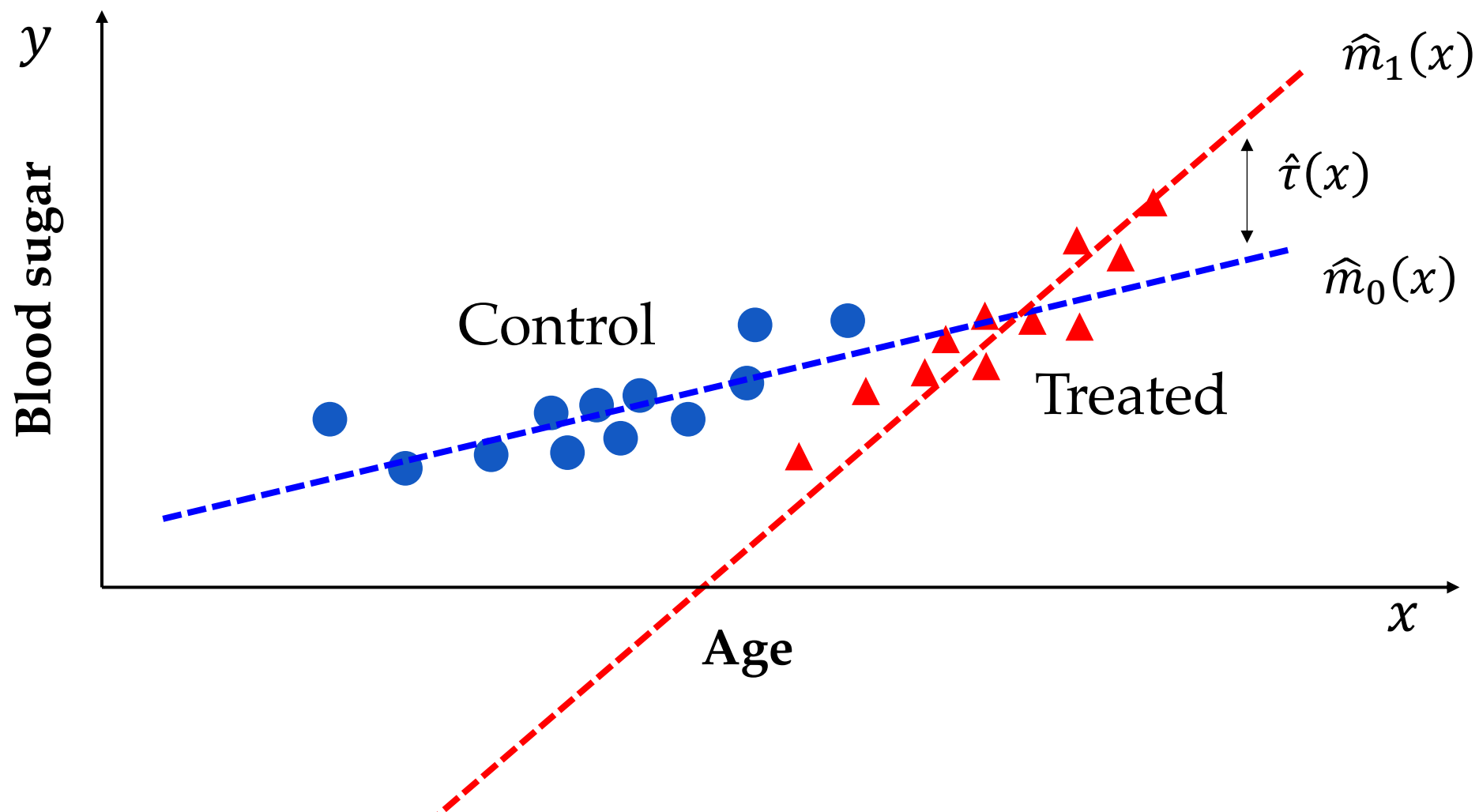
1. Identify a valid adjustment set X
2. Collect data $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$
3. Fit regression models $\hat{m}_t(x) \approx \mathbb{E}[Y \mid X = x, T = t]$
e.g., using empirical risk minimization (ERM)*:

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

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

*Künzel et al., *PNAS*, 2019

Regression adjustment



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 treatment¹

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

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

¹Rosenbaum & 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

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

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

$$(*) = \sum_{x,y} \underbrace{p(X = x | T = t)}_{\text{Treatment group } t} \frac{p(X = x)}{p(X = x | T = t)} \underbrace{p(Y(t) = y | X = x)}_{\text{Potential outcome}} y$$

Inverse propensity score weighting (IPW)

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

$$\begin{aligned} (*) &= \sum_{x,y} p(X = x \mid T = t) \frac{p(X = x)}{p(X = x \mid T = t)} \underbrace{p(Y = y \mid X = x, T = t)}_{\text{Observational outcome}} y \\ &= \mathbb{E}_{X,Y} \left[\underbrace{\frac{p(T = t)}{p(T = t \mid X = x)}}_{> 0 \text{ due to overlap}} Y \mid T = t \right] \quad \square \quad \text{Weighted average} \end{aligned}$$

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

Inverse-Propensity Weighting Adjustment

1. Identify a valid adjustment set X
2. Collect data $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$
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 $\hat{\tau}$

Weighted average

The IPW estimator is a weighted average. With

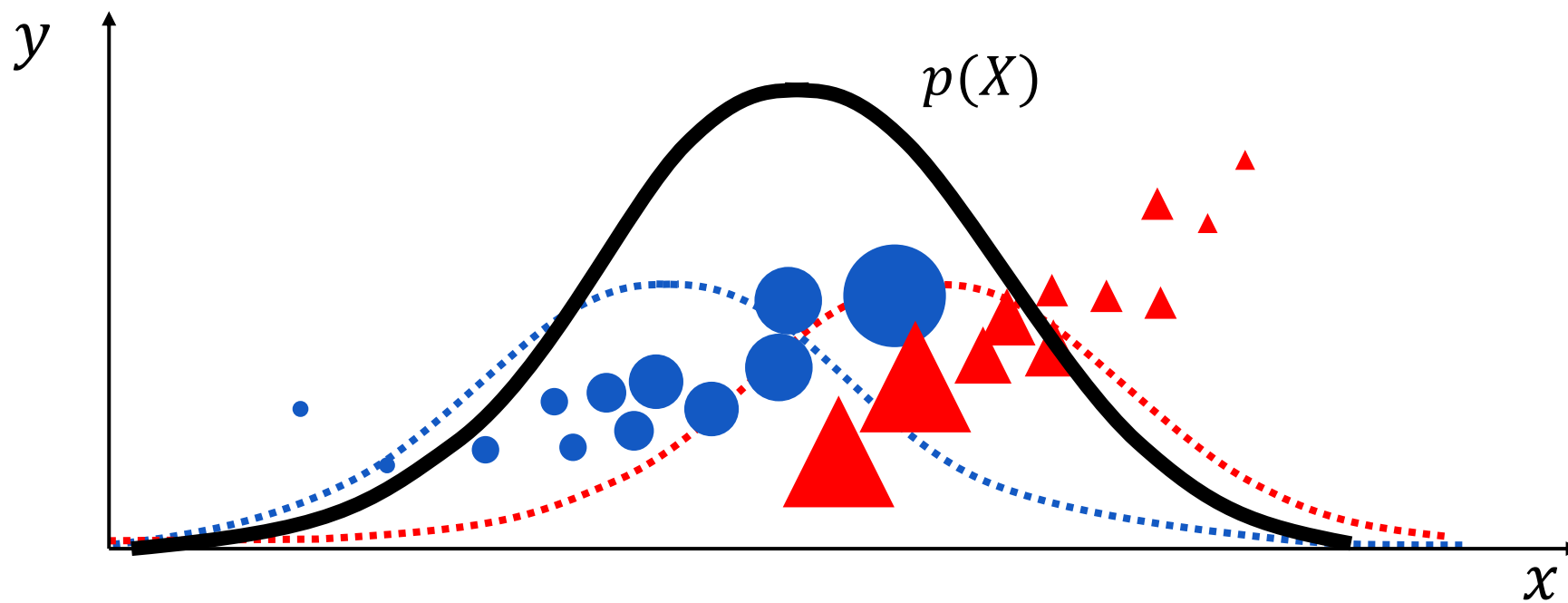
$$w_i = \frac{p(T = t)}{p(T = t \mid X = x_i)}$$

we have

$$\hat{\mu}_t = \frac{1}{n_t} \sum_{i:t_i=t} y_i w_i$$

IPW example

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



IPW example

Regression adjustment and inverse-propensity weighting are two of the most commonly used methods to estimate ATE / CATE

There are **many** more methods!

Many of these combine both approaches and are “doubly robust”^{*}
This can improve sample efficiency

^{*} See e.g., Nie & Wager, *Biometrika*, 2021 who introduced the “R-learner”

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

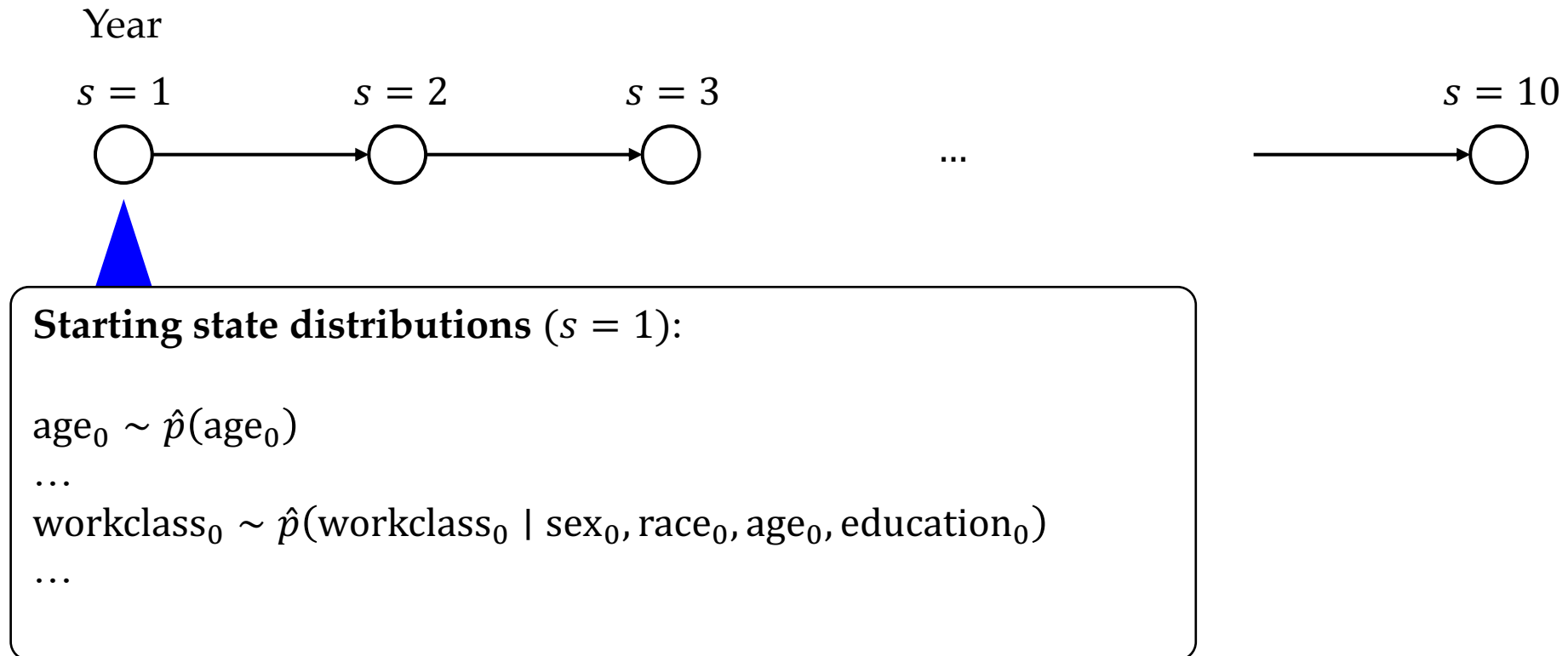
[Let's have a look at the data]

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

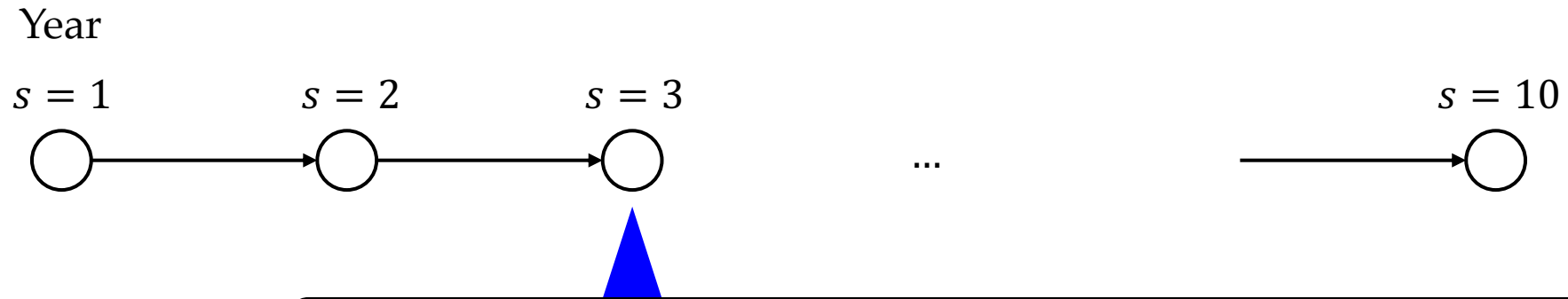
Example: 'workclass' a function of 'age', 'education', 'race', 'sex'

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

Markov model



Markov model



Transition distributions ($s > 0$):

$$\text{age}_s \sim p(\text{age}_s \mid \text{age}_{s-1})$$

...

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

...

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

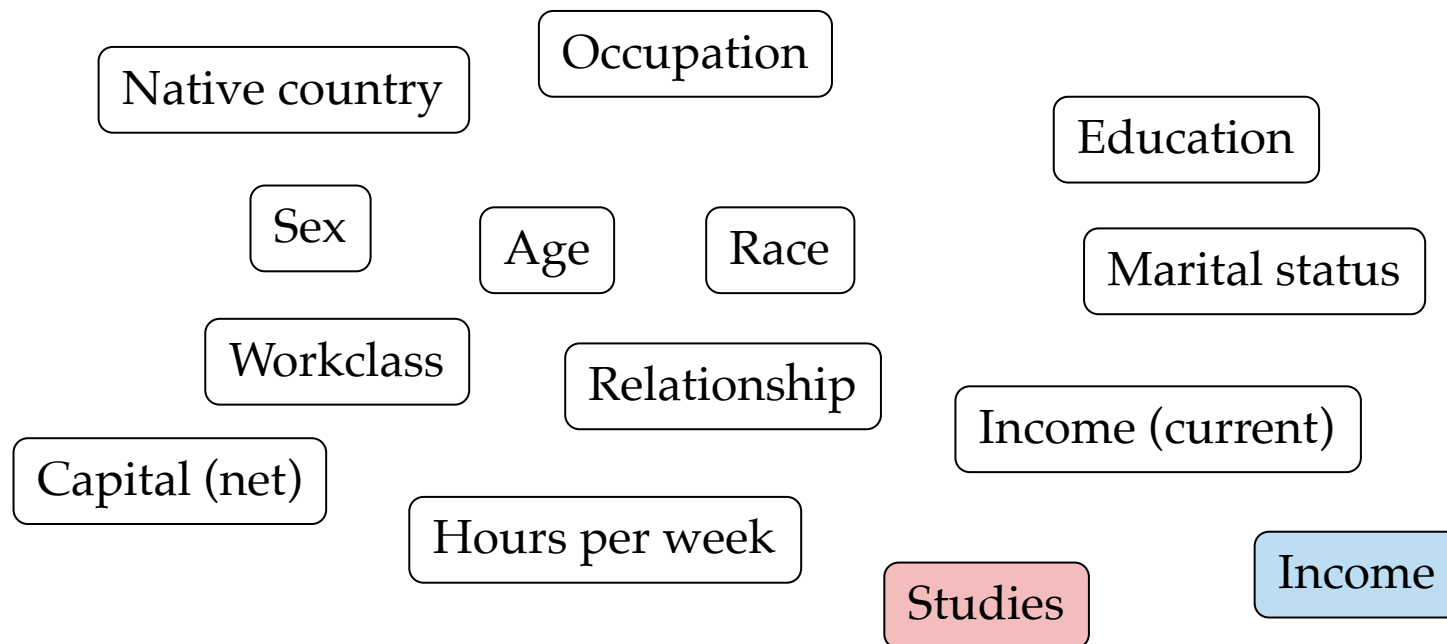
1. Identify causal parameter *e.g., average treatment effect*
2. Pick an identification strategy *e.g., backdoor adjustment*
3. Identify a statistical estimand *e.g., conditional expectation*
4. Execute an estimation strategy *e.g., random forest regression*

Identification

We will use covariate adjustment to identify ATE

To do **covariate adjustment**, we need to establish exchangeability

What is the causal graph?

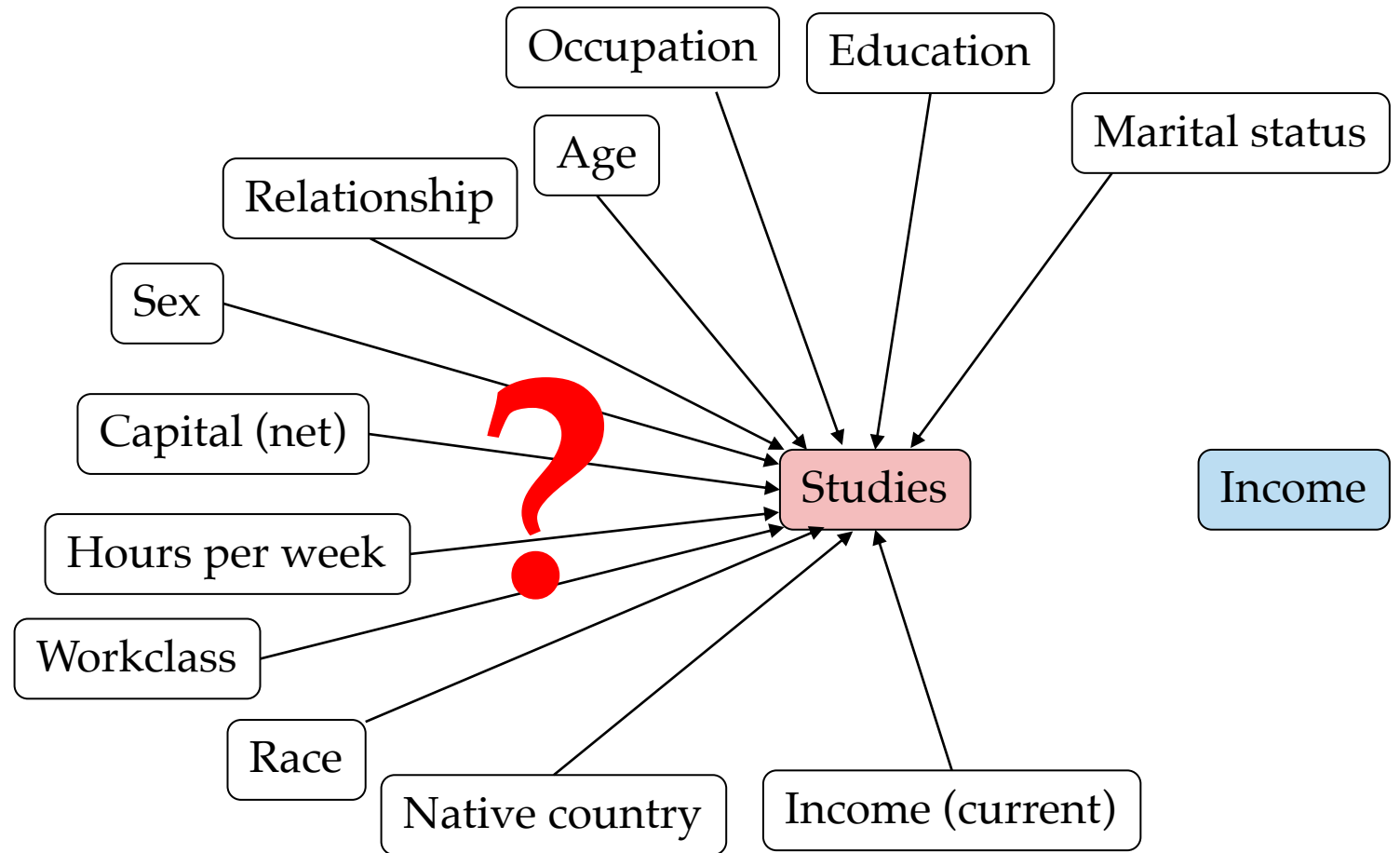


Recall: Parent adjustment

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

⇒ They are an adjustment set

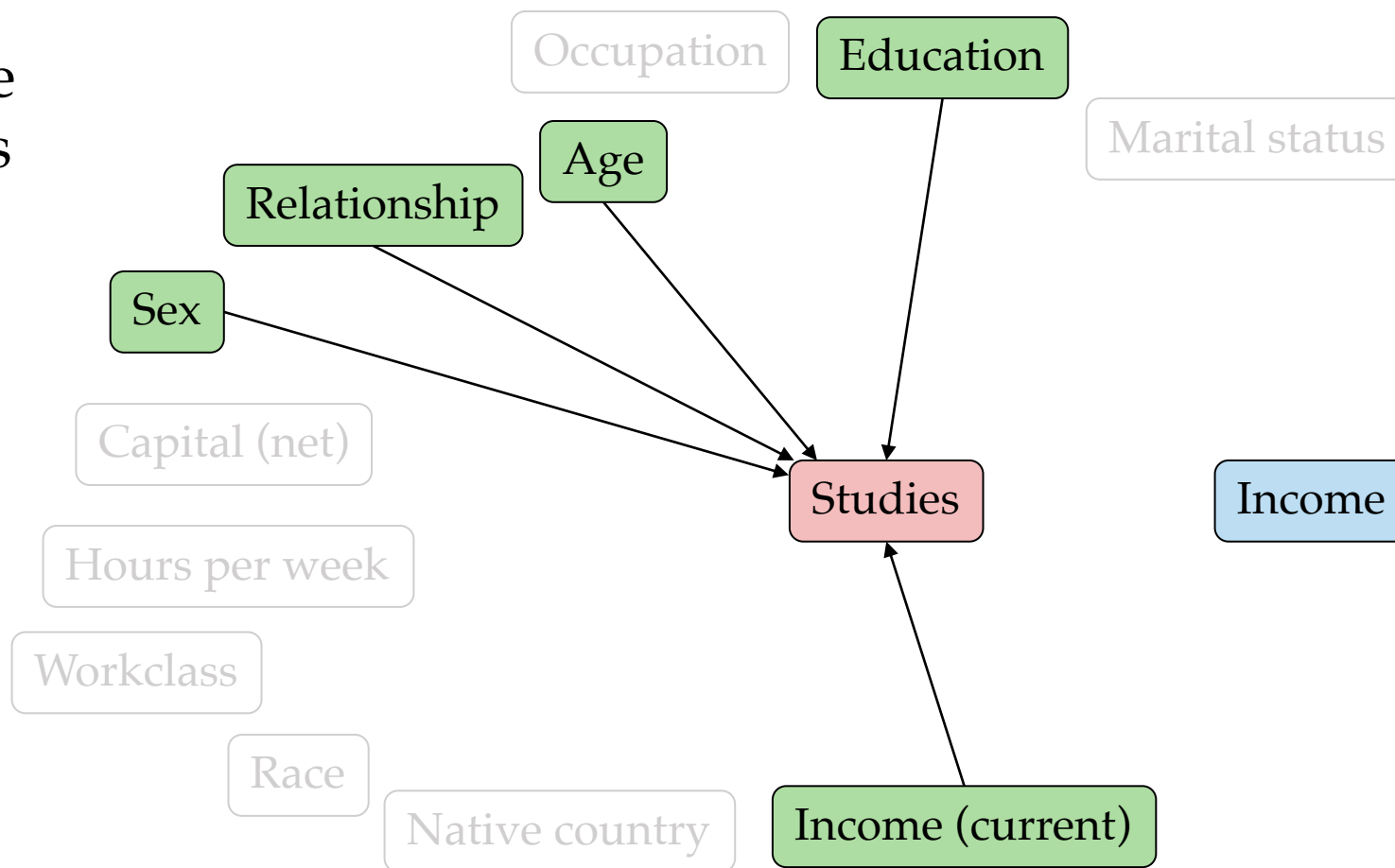
We don't need the whole graph!



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

Many stones left unturned

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

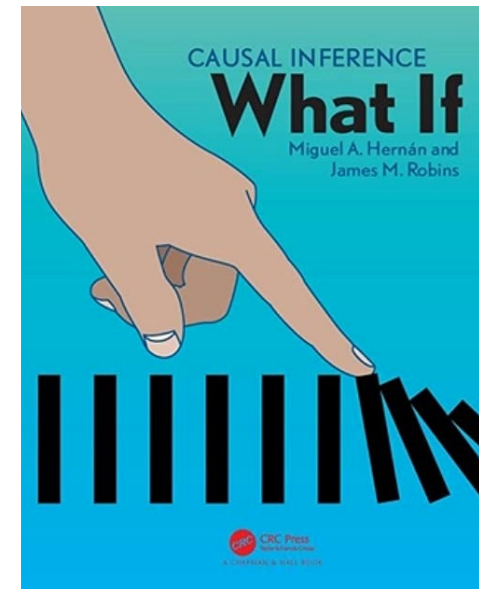
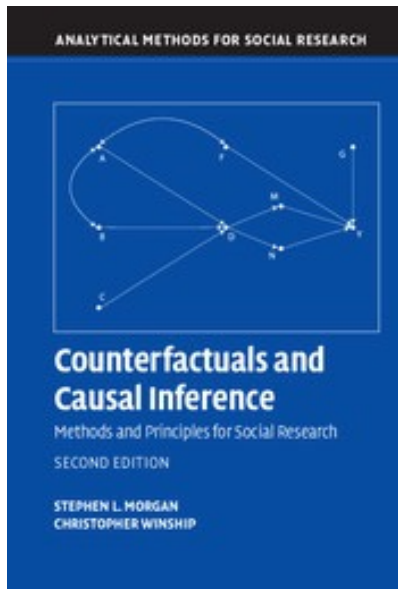
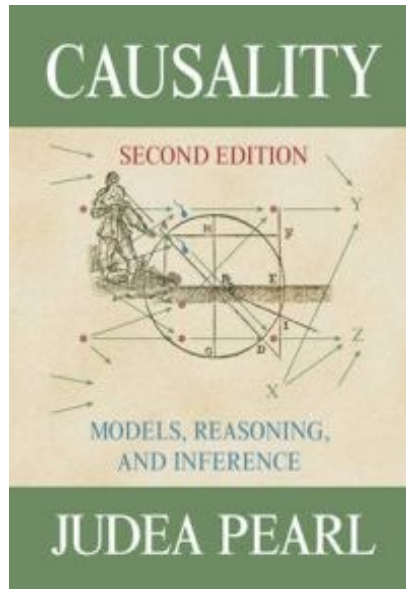
If we leave out some, there may be **unobserved confounders**

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

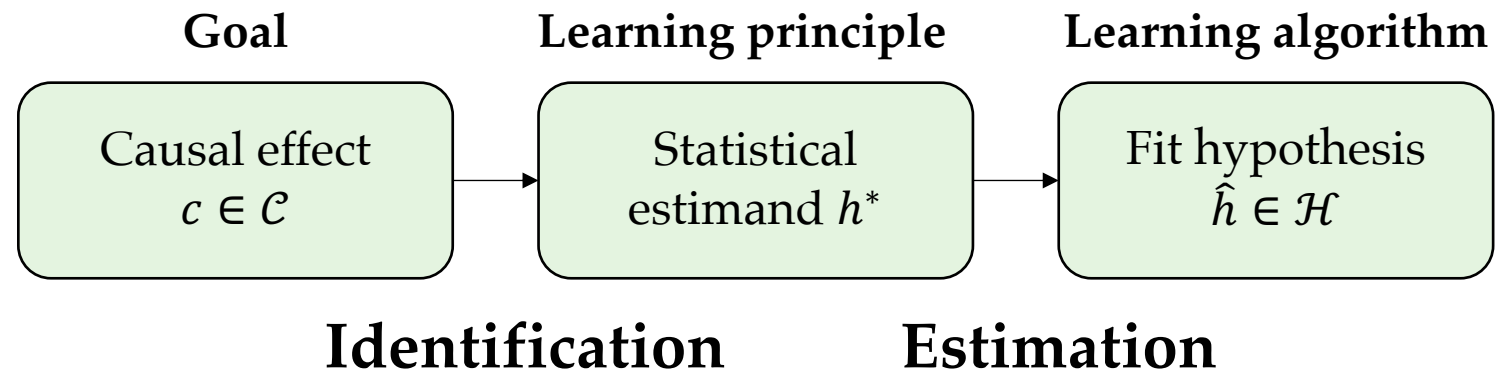
Causal machine learning is relatively young, but there is already a rich literature—most of which was not covered today!



Final remarks

If you know that you will use your ML model to make decisions, make sure your learning algorithm knows this!

Causal analysis is used to pick the right learning goal



fredrik.johansson@chalmers.se
www.healthyai.se