

Potential outcomes, counterfactuals & conditional treatment effects

Causal Inference & Deep Learning

MIT IAP 2018

Fredrik D. Johansson



Everyone wants to make better decisions.

Predicting effects of decisions
requires **causal** reasoning

Causal inference problems are becoming

High-dimensional



What are the effects of genetics?

Causal inference problems are increasingly

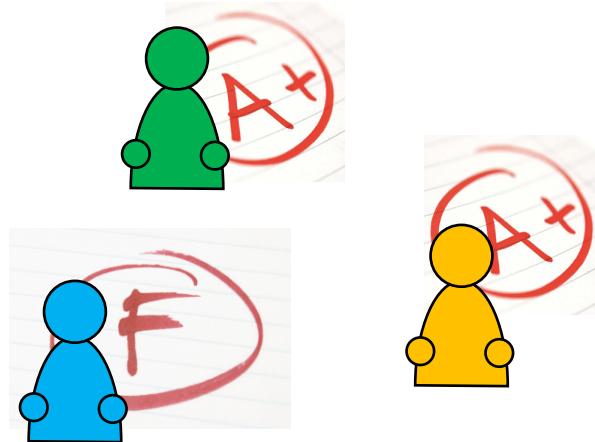
Observational (non-experimental)



How should we treat our patients?

Causal inference problems are becoming

Personalized



How should we teach our students?

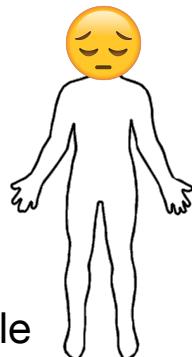
Why do **I** care about decision making?

- ▶ Clinical machine learning group (clinicalml.com)
- ▶ Many health care problems concern decision making
 - ▶ **Who** should receive which treatment?
 - ▶ **Which** lab tests should be ordered?
 - ▶ **When** should we intervene?
 - ▶ (**What** was the cause of death?)



May 15

Anna



Age = 54

Gender = Female

Race = Asian

Blood pressure = 150/95

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

Temperature = $36.7^\circ C$

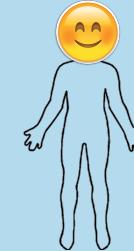
Blood sugar = High

Medication A
“Control”

$t = 0$



Sep 15

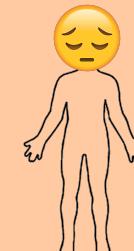


Blood sugar = ?

$Y(0)$

Medication B
“Treated”

$t = 1$



Blood sugar = ?

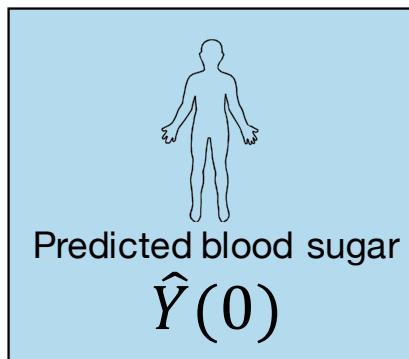
$Y(1)$

Treatment effects

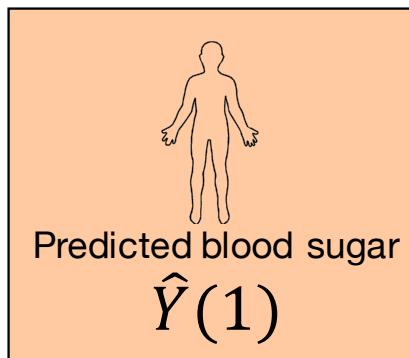
- ▶ The **effect** of treatment is the **difference** between what would have happened under each path, $\tau = Y(1) - Y(0)$
- ▶ Can be used for **population-wide** policy-making:
Does smoking have an effect on cancer rate?
- ▶ or for **personalized** intervention:
Which medication works best for Anna, specifically?

Machine learning approach

- **Predict** both outcomes $Y(0)$ and $Y(1)$



If $\hat{Y}(0) < \hat{Y}(1)$ give treatment $T = 0$



If $\hat{Y}(1) < \hat{Y}(0)$ give treatment $T = 1$

Supervised learning

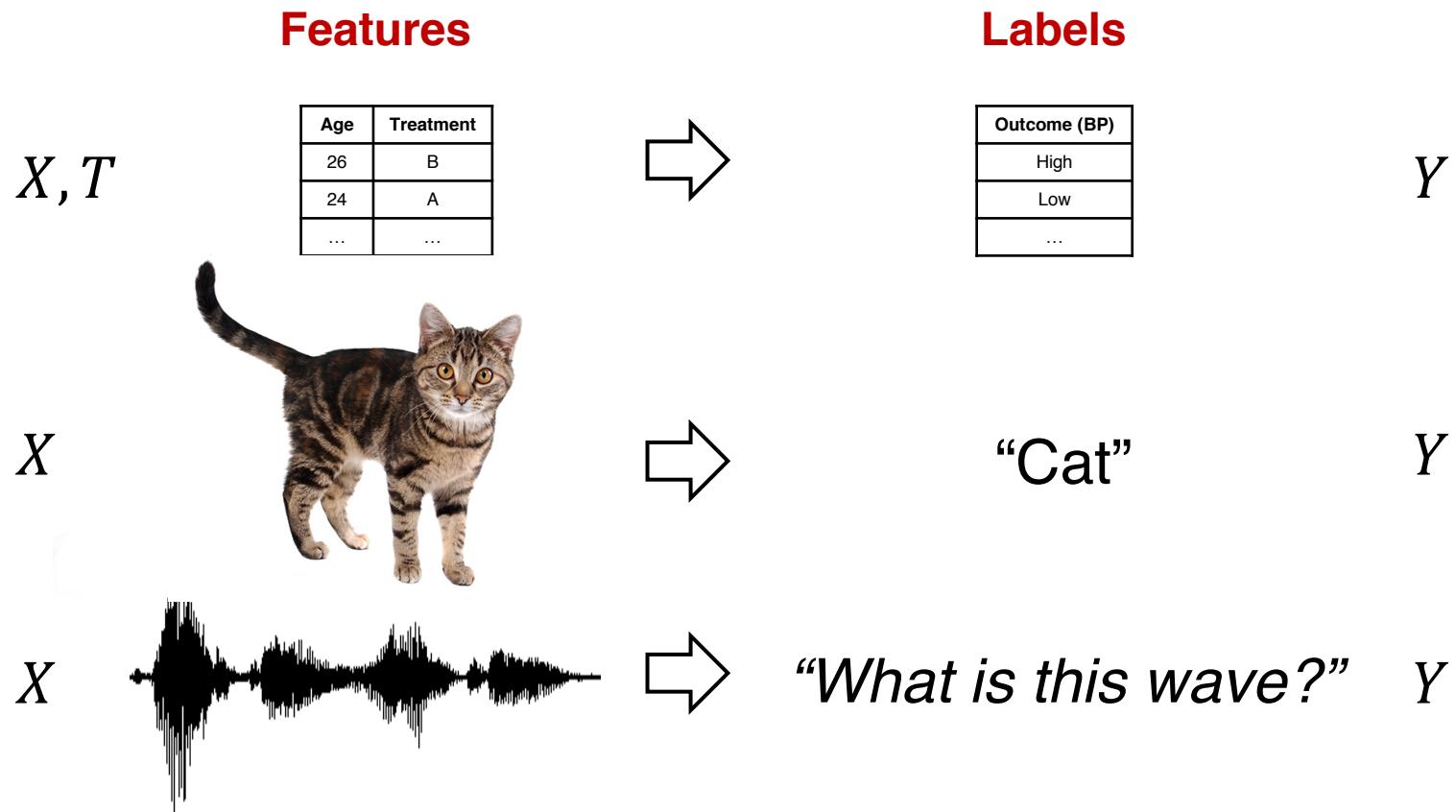
- ▶ Fit function to historical records of patients

X	T	Y
Age	Treatment	Outcome (BP)
26	B	High
24	A	Low
...

Find $f(x, t) \approx y$

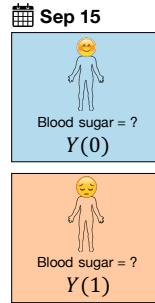
Supervised learning

- ▶ Superficially similar to many supervised learning problems



Supervised learning

- ▶ Observe only outcome of prescribed treatment



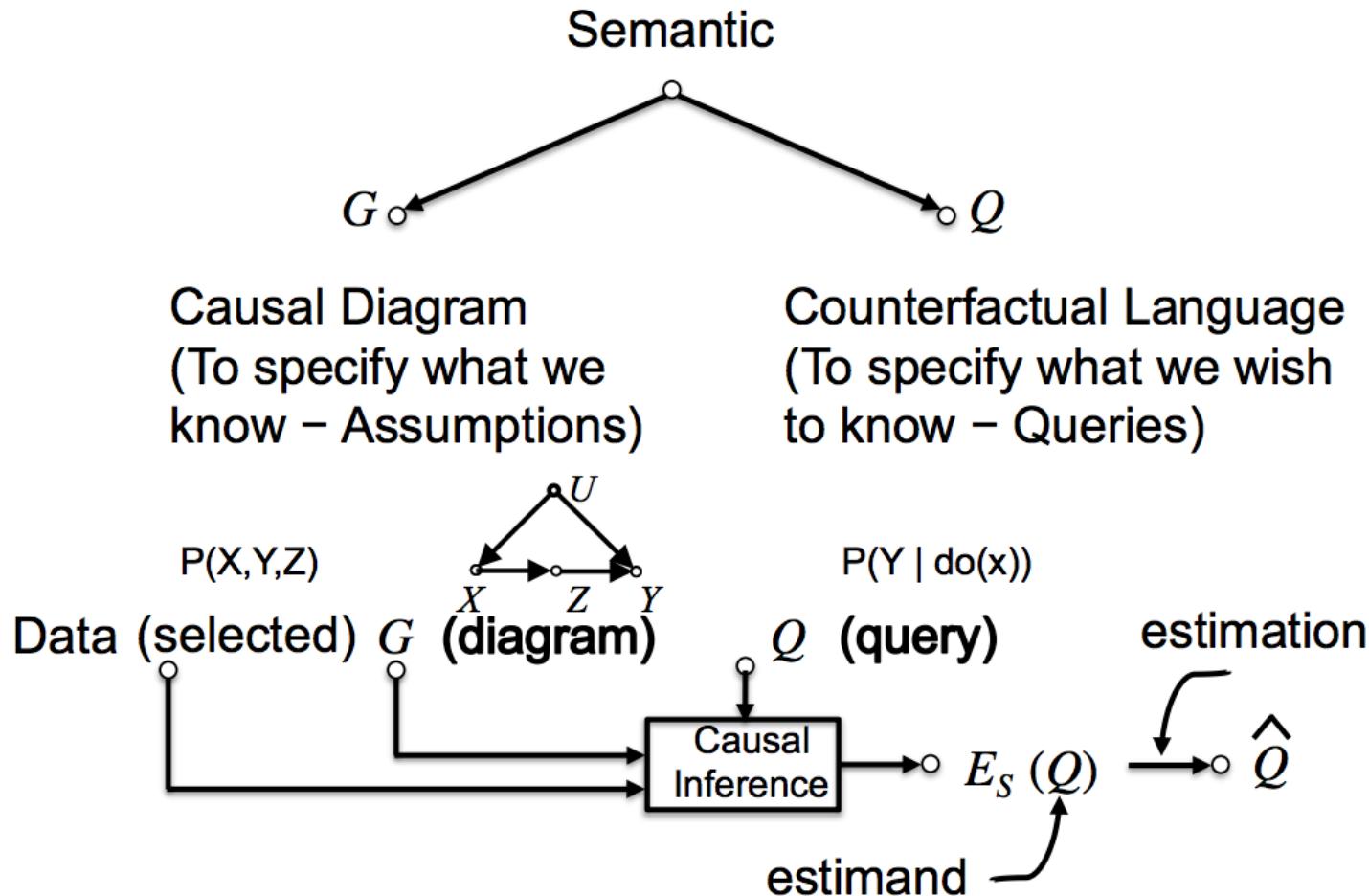
X	T	Y	Y_{cf}
Age	Treatment	Outcome (BP)	Counterfactual
26	B	High	?
24	A	Low	?
...

Find $f(x, t) \approx Y(t)$

Why is supervised learning not enough?

- ▶ No supervision for counterfactual outcomes!
- ▶ Counterfactual inference thought of as a **missing variables** problem or as **domain adaptation**
- ▶ **Treatment** is **unlike** other features—we do not want to e.g. remove it in variable selection

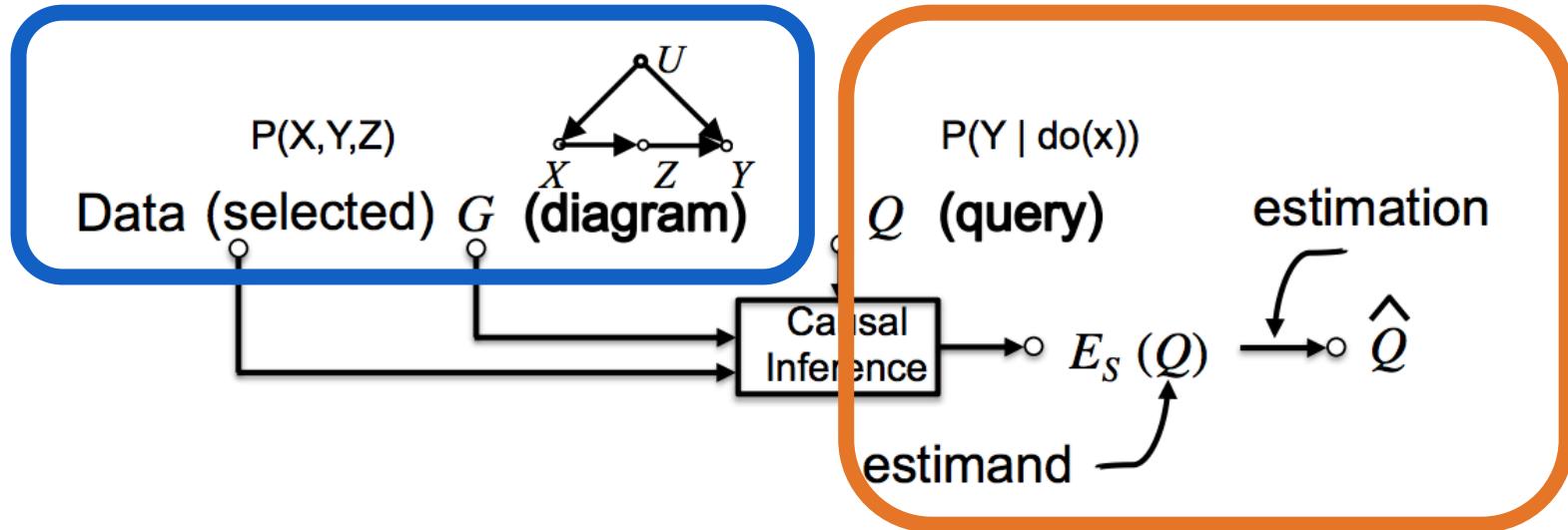
Causality and machine learning



Causality and machine learning

Max's part

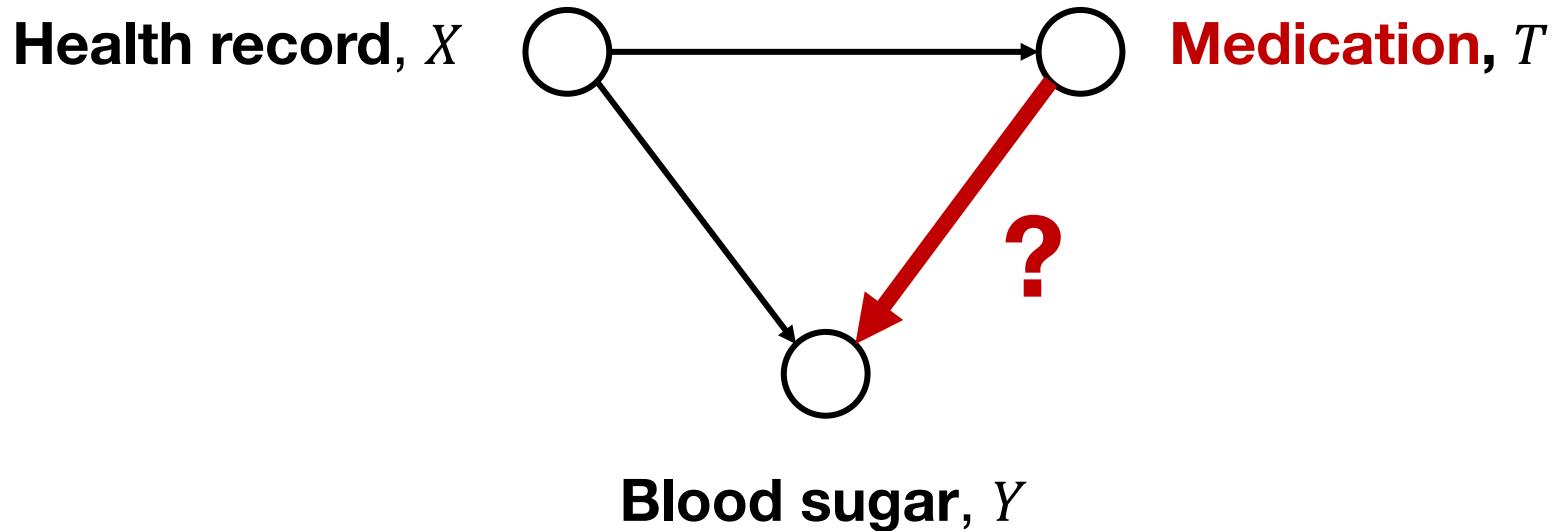
My part



- 1. Potential outcomes framework**
- 2. Supervised learning (risk minimization)**
- 3. Adjusting for distributional shift**

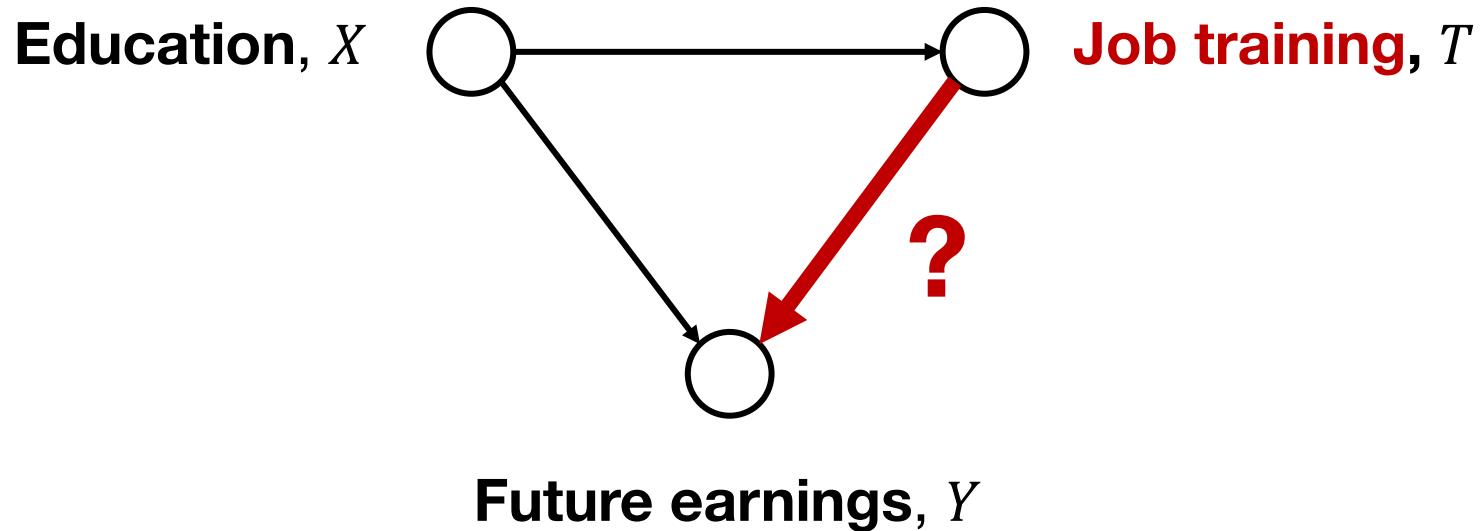
Anna's causal graph

- ▶ Does **Medication B** reduce **blood sugar**?



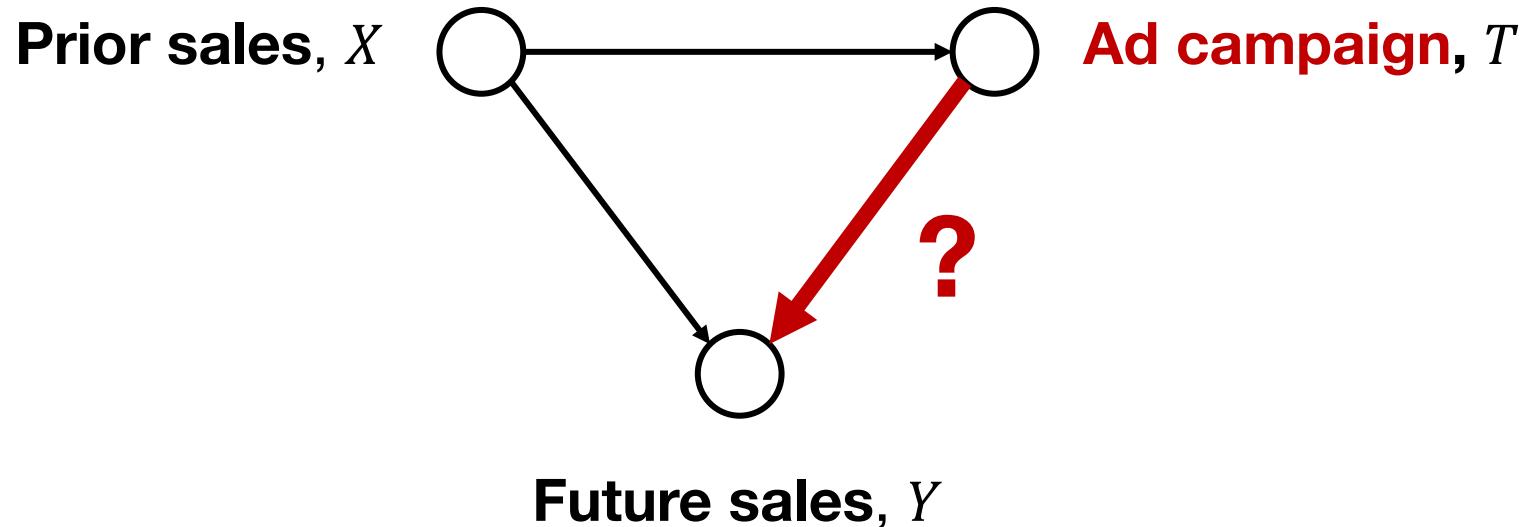
Job training

- ▶ Does **job training** increase future **earnings**?



Advertising

- ▶ Will running an **ad campaign** increase **sales**?



Potential outcomes

- ▶ The (X, T, Y) context-treatment-outcome triple is a common pattern—same graph, same queries of interest
- ▶ **Potential outcomes** (Neyman-Rubin causal model) are convenient in this setting
- ▶ Let $Y(t)$ denote the **potential** outcome under intervention t . *What would happen if we intervened with $T = t$?*
- ▶ Potential outcomes are complementary to causal graphs
- ▶ Interventions are often called “treatment” for historical reasons

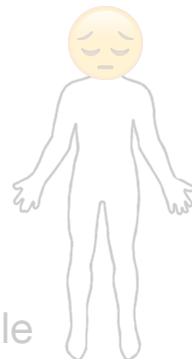
Potential & conditional outcomes

- ▶ We use potential outcomes to differentiate between conditioning and intervening
- ▶ **Conditional** outcome: $\mathbb{E}[Y | T = 1]$
What is the expected outcome for subjects who we would currently treat with $T = 1$?
- ▶ **Potential** outcome: $\mathbb{E}[Y(1)]$
*What is the expected outcome if we were to treat **everyone**?*



May 15

Anna



Age = 54

Gender = 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



Blood sugar = ?

$Y(0)$

Medication B “Treated”

$t = 1$



Blood sugar = ?

$Y(1)$

Potential outcomes – Binary treatment

- ▶ Let treatment be binary: $T \in \{0, 1\}$
and call $t = 0$ **control**, $t = 1$ **treated**
- ▶ $Y(1)$ treated outcome, $Y(0)$ control outcome
- ▶ **Factual** outcome: $Y = (1 - T)Y(0) + TY(1)$
- ▶ Joint distribution $X, T, Y \sim p(X, T, Y)$

¹Average is taken over noise in Y

Two effects of interest

- ▶ Conditional Average Treatment Effect (**CATE**)¹

$$\tau(x) = \mathbb{E}_p[Y(1) | X = x] - \mathbb{E}_p[Y(0) | X = x]$$

- ▶ The average effects on subjects with features $X = x$
- ▶ Can be used to **personalize** treatment:
Should I be taking Medication B, given my current state?

¹Average is taken over noise in Y

Two effects of interest

- ▶ (Population) Average Treatment Effect (**PATE/ATE**)

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

$$= \mathbb{E}_p[\tau(x)]$$

- ▶ Studied for hundreds of years
- ▶ Often the target of drug trials:
Does Medication B work better on average?

¹Average is taken over noise in Y

Potential outcomes & CATE

- ▶ Tabular records of patients

Age	Treatment	Outcome (BS)
26	A	High
24	B	Low
...

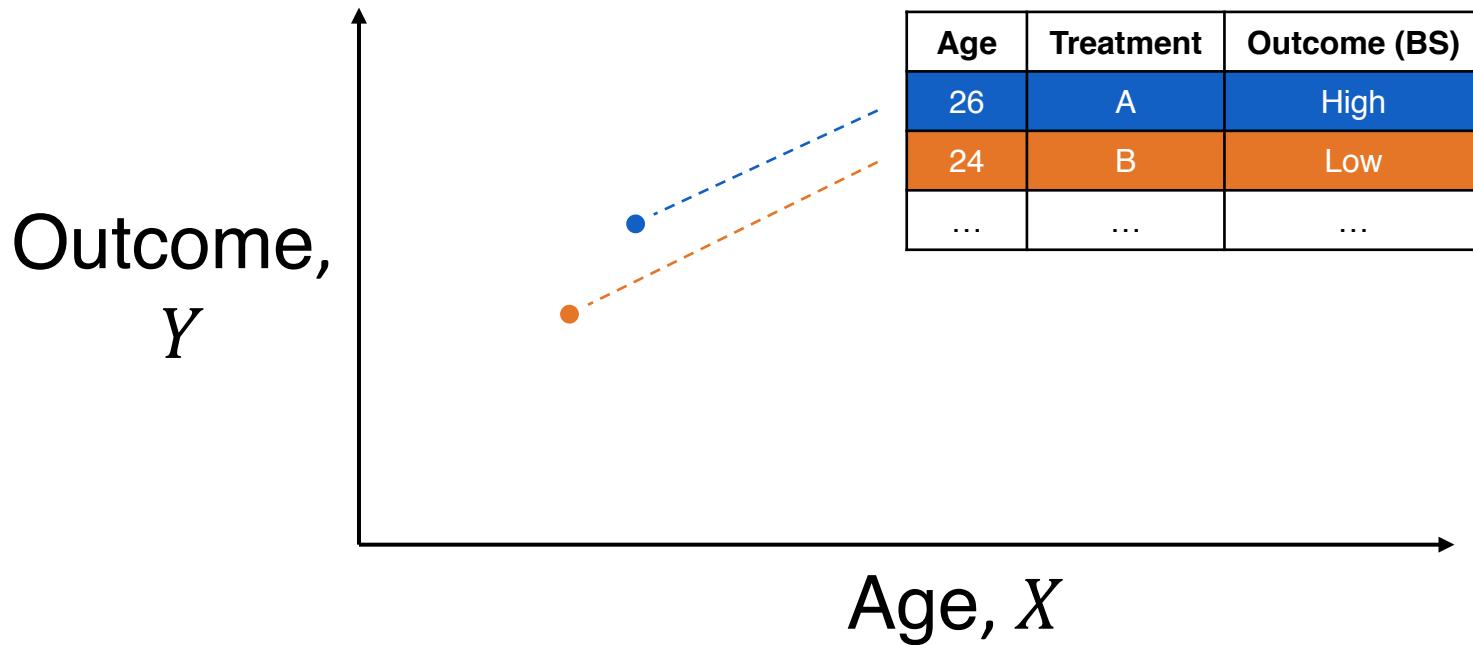
Potential outcomes & CATE

- ▶ Represent features and outcome in a plot



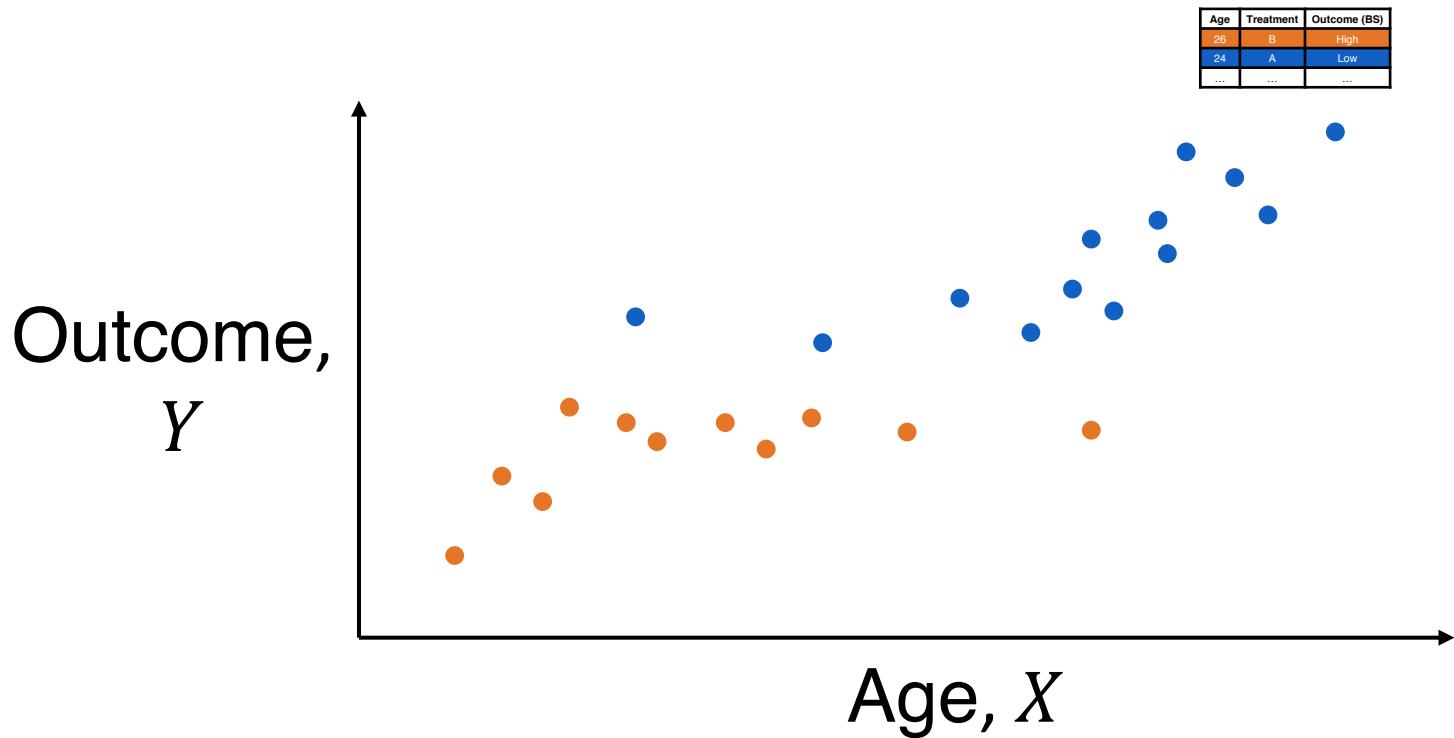
Potential outcomes & CATE

- ▶ Let color represent treatment groups



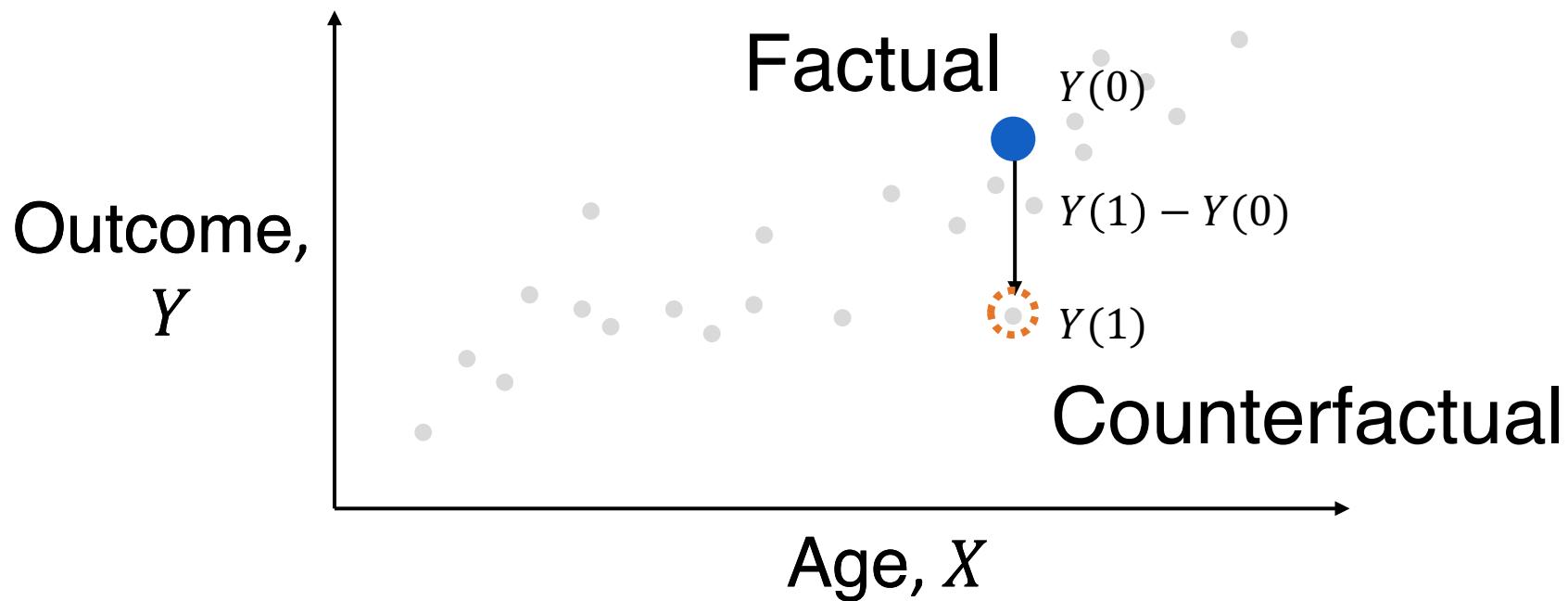
Potential outcomes & CATE

- Treatment groups need not be identically distributed (see later)



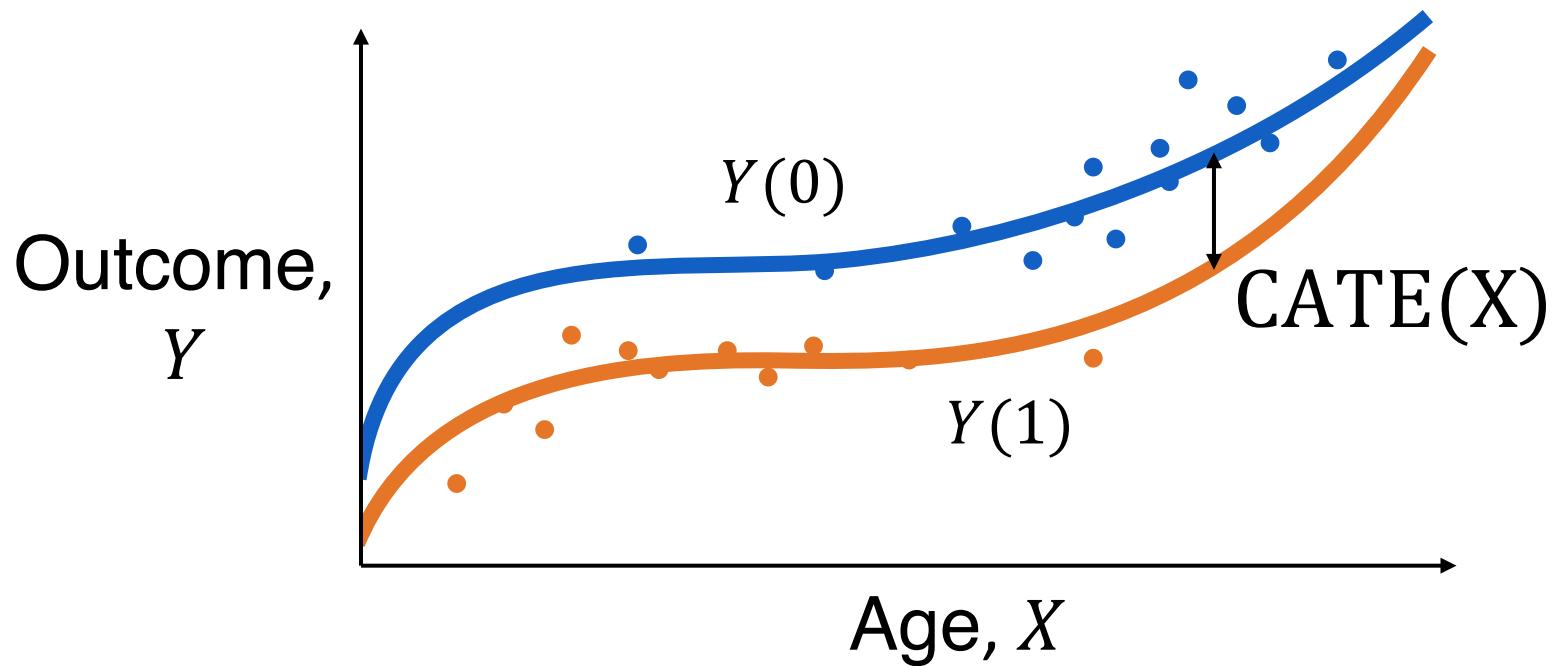
Counterfactuals

- ▶ Effect on a unit determined by (unobserved) **counterfactual**



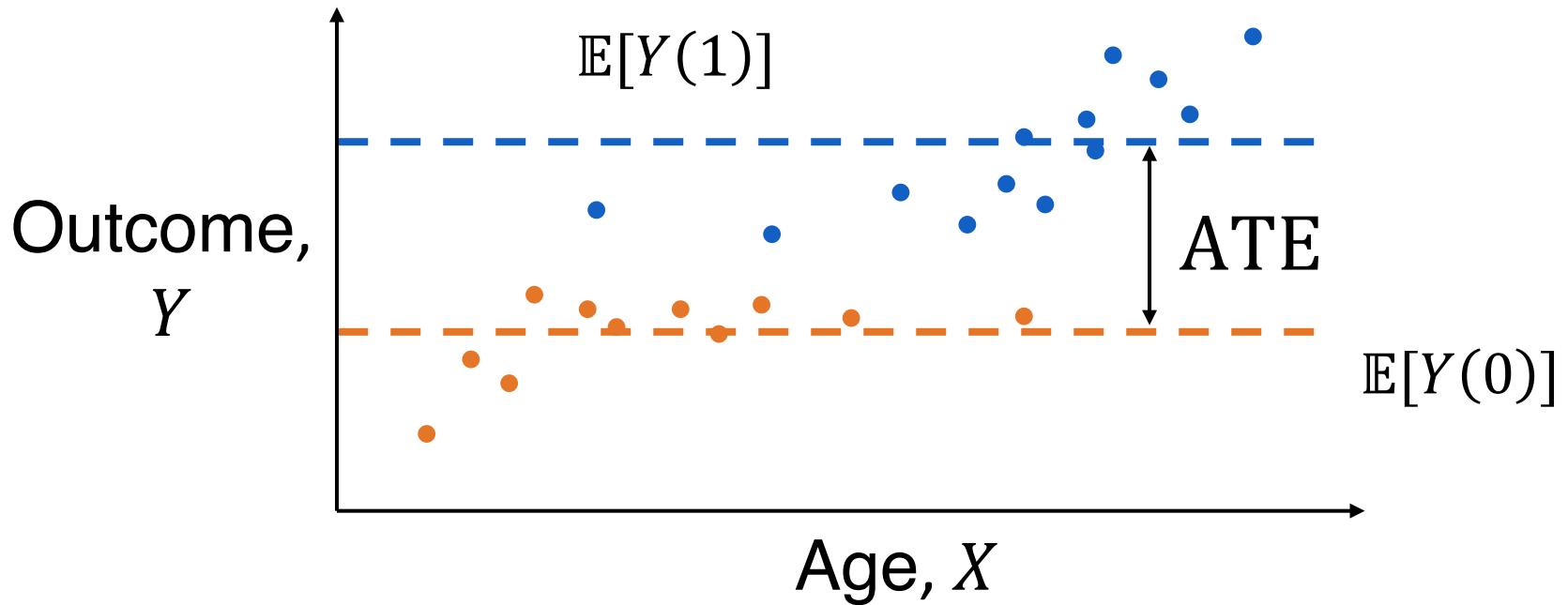
Potential outcomes & CATE

- ▶ Expected potential outcomes $Y(0), Y(1)$ are the goal



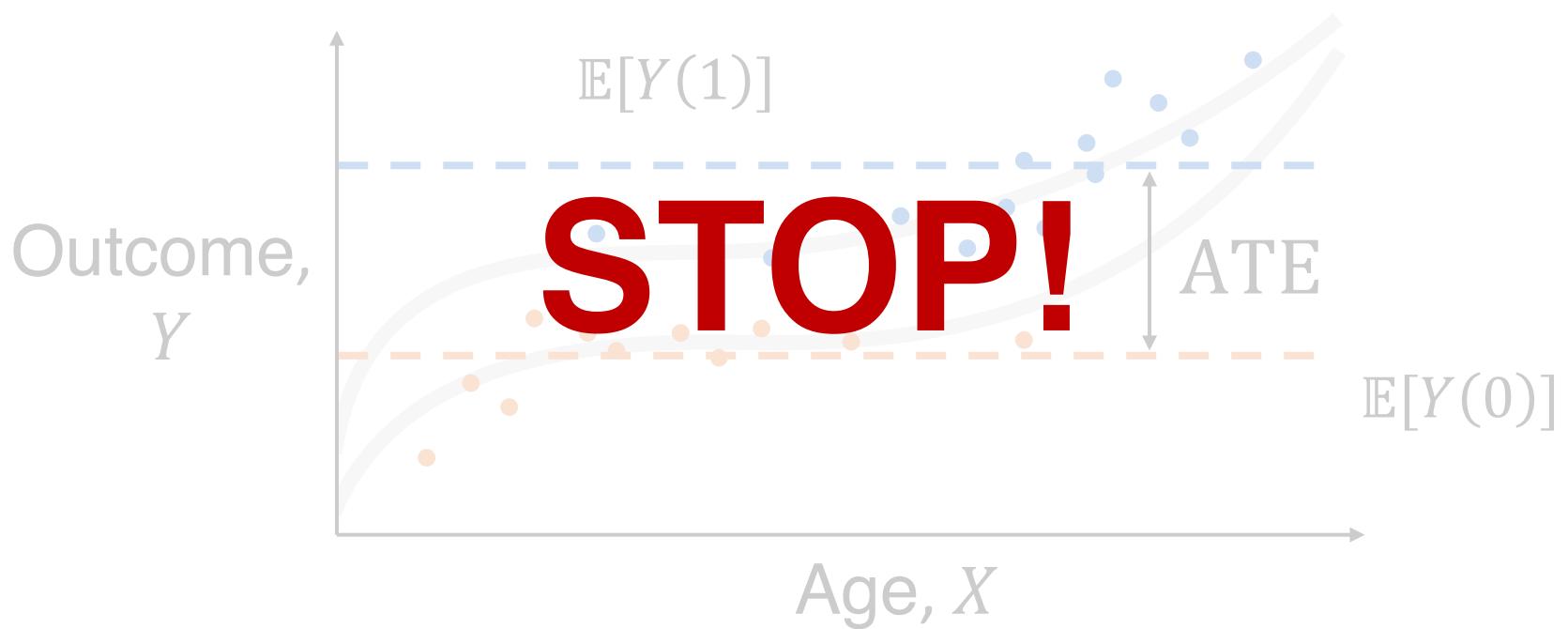
Potential outcomes & ATE

- ▶ Average treatment effect



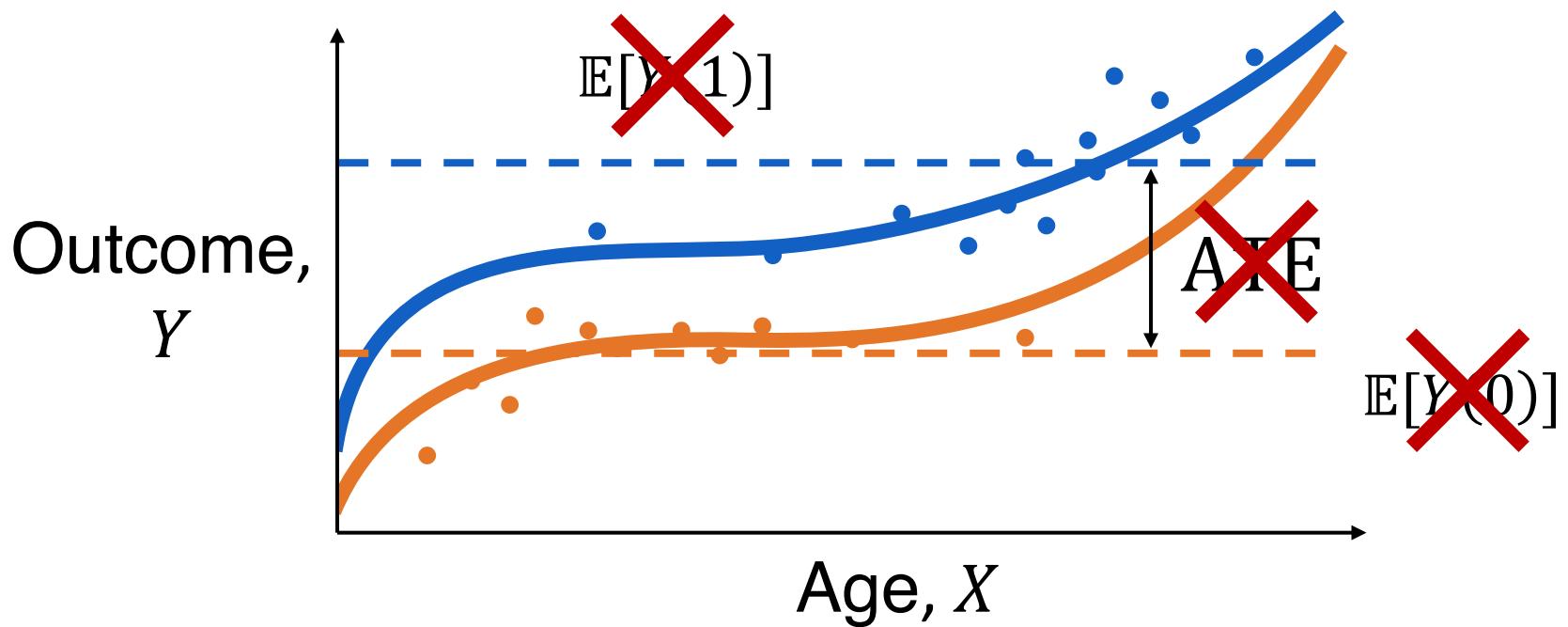
Potential outcomes & ATE

- Average treatment effect



Potential outcomes & ATE

- Average effect can't be larger than maximum conditional!

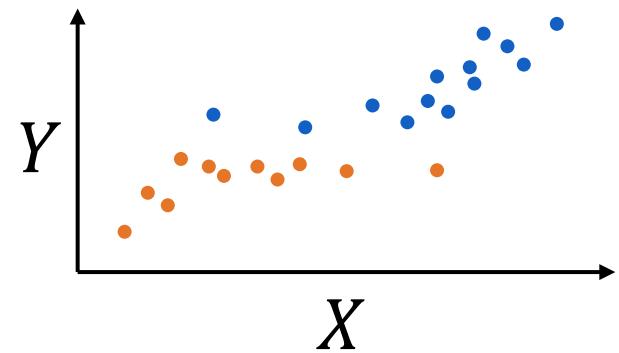
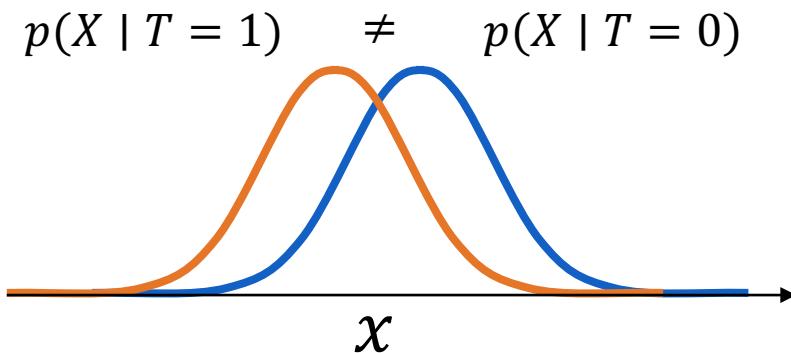




What happened here?

Treatment group imbalance

- ▶ Treatment groups are not identically distributed



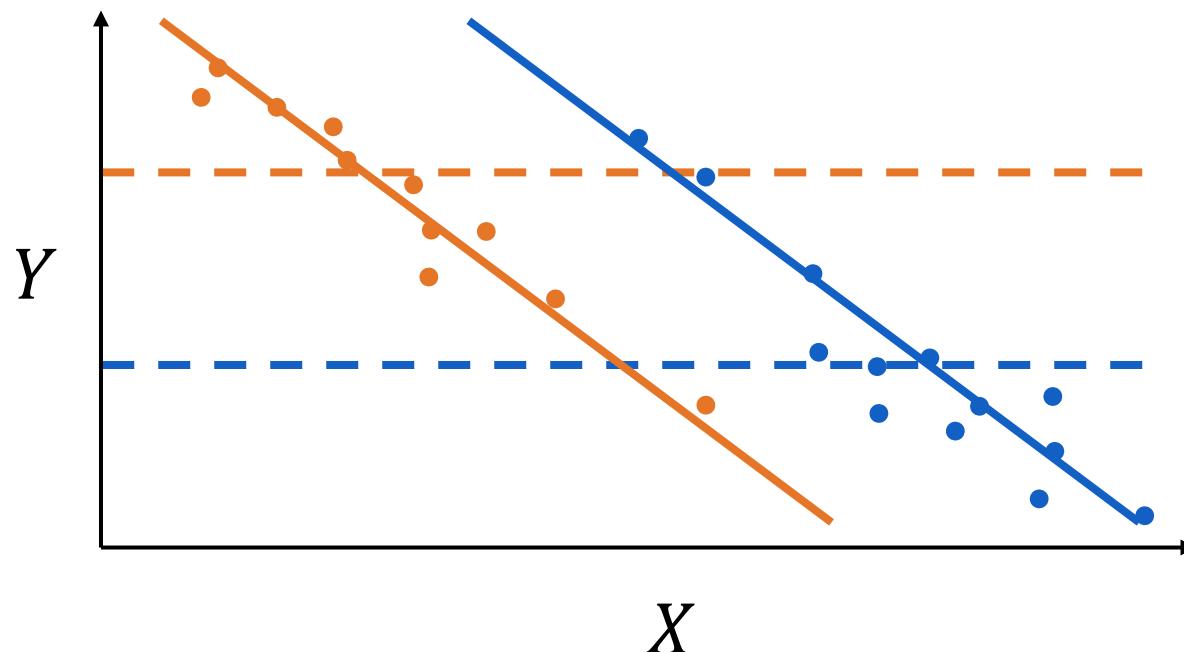
$$\underbrace{\mathbb{E}_p[Y | T = 1] - \mathbb{E}_p[Y | T = 0]}_{\text{What we computed}} \neq \underbrace{\mathbb{E}_p[Y(1) - Y(0)]}_{\text{What we want}}$$

What we computed

What we want

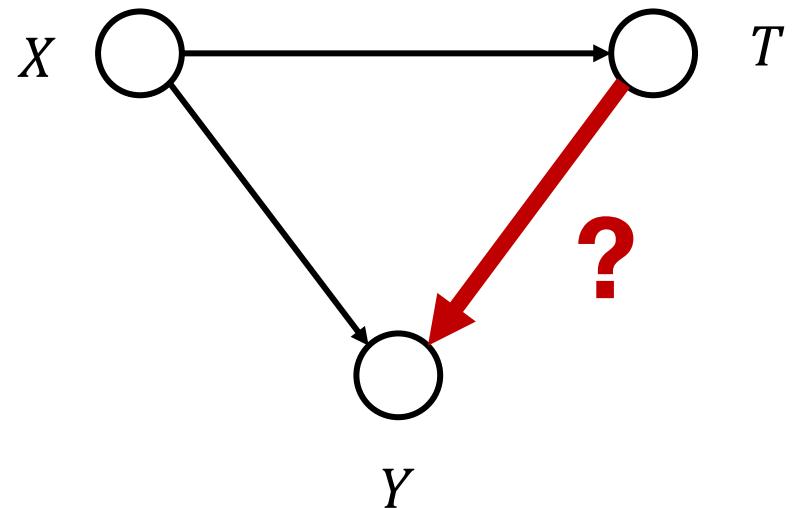
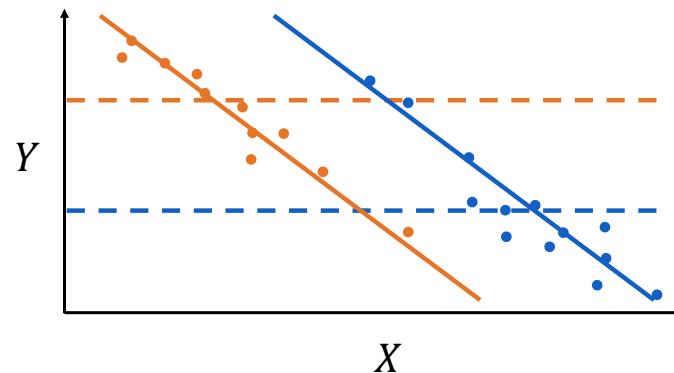
Simpsons paradox

- ▶ There are examples where for **every** $x \in \mathcal{X}$, CATE is positive, but the naive ATE estimate is negative



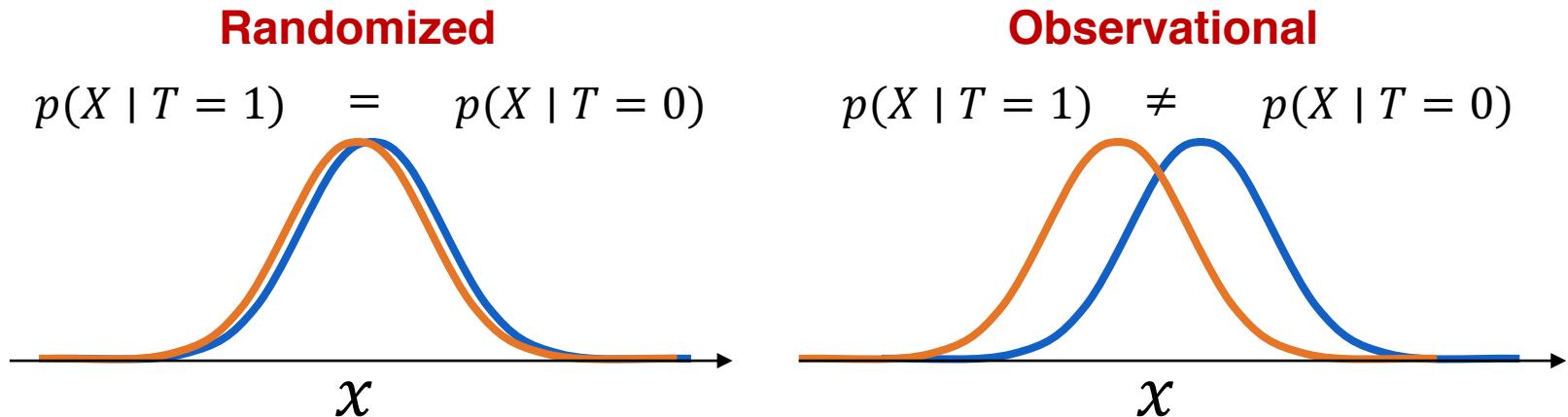
Confounding

- ▶ X is a **confounder!**
- ▶ It affects both the treatment assignment T and the outcome Y
- ▶ Confounds the naïve estimate of treatment effect



Randomized controlled trials (RCT)

- ▶ Gold standard. Removes confounding!
- ▶ Treatment assignment is randomized: $p(T | X) = p(T)$



Observational data

- ▶ **Randomized** controlled trials are expensive and sometimes unethical or impractical
We can't force people to smoke
- ▶ **Observational** data is plentiful and cheap
How are patients treated in hospitals today?
- ▶ So, how can we learn from observational data?

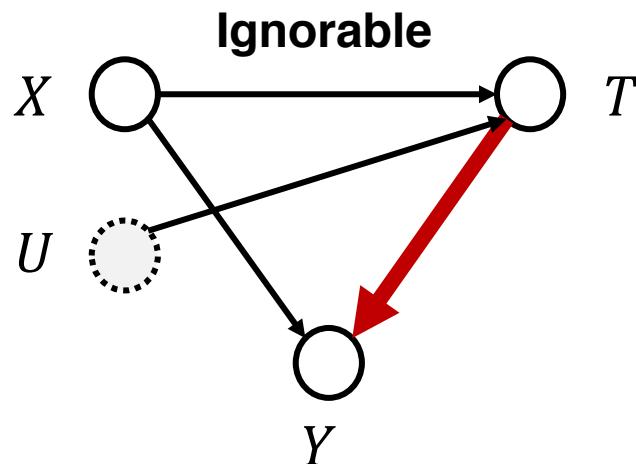
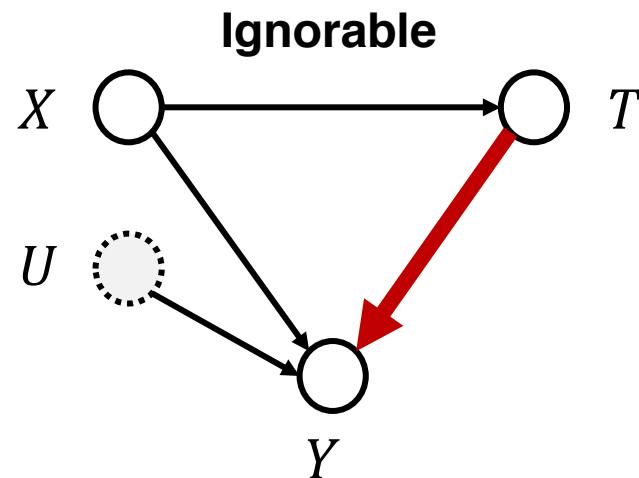
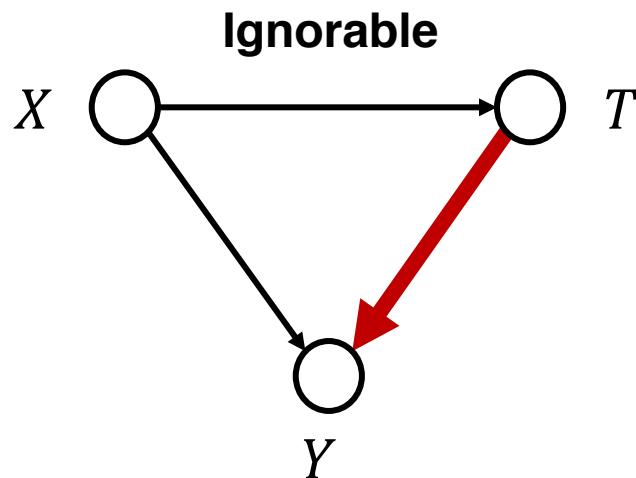
Assumptions in observational case

- ▶ Sufficient conditions for **identifiability**:
- ▶ **Strong ignorability**¹: $Y_0, Y_1 \perp\!\!\!\perp T \mid X$ “No hidden confounders”
All variables that affect both Y_t and T are measured
- ▶ **Overlap**: $\forall x, t: p(T = t \mid X = x) > 0$
All treatments have non-zero probability of being observed
- ▶ **Stable Unit Treatment Value Assumption** (SUTVA):
Treatments and outcomes of different subjects are independent

Ignorability

○ Measured

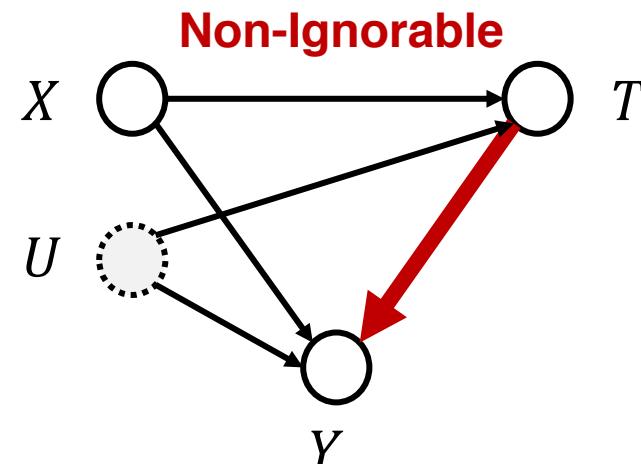
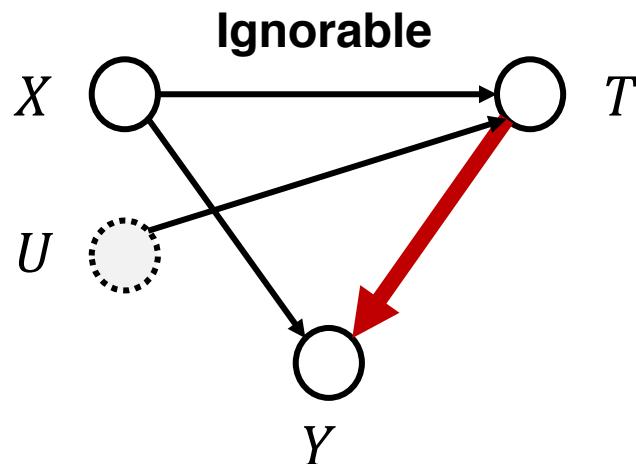
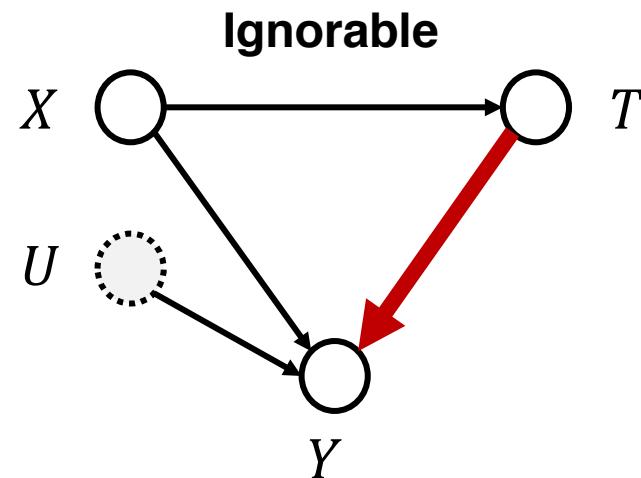
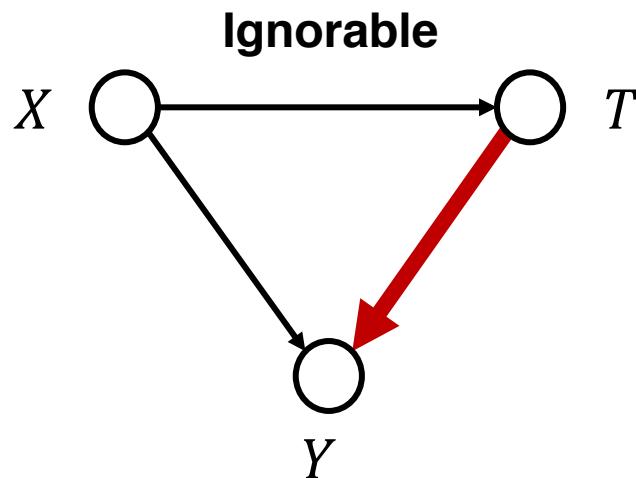
○ Unmeasured / hidden



Ignorability

○ Measured

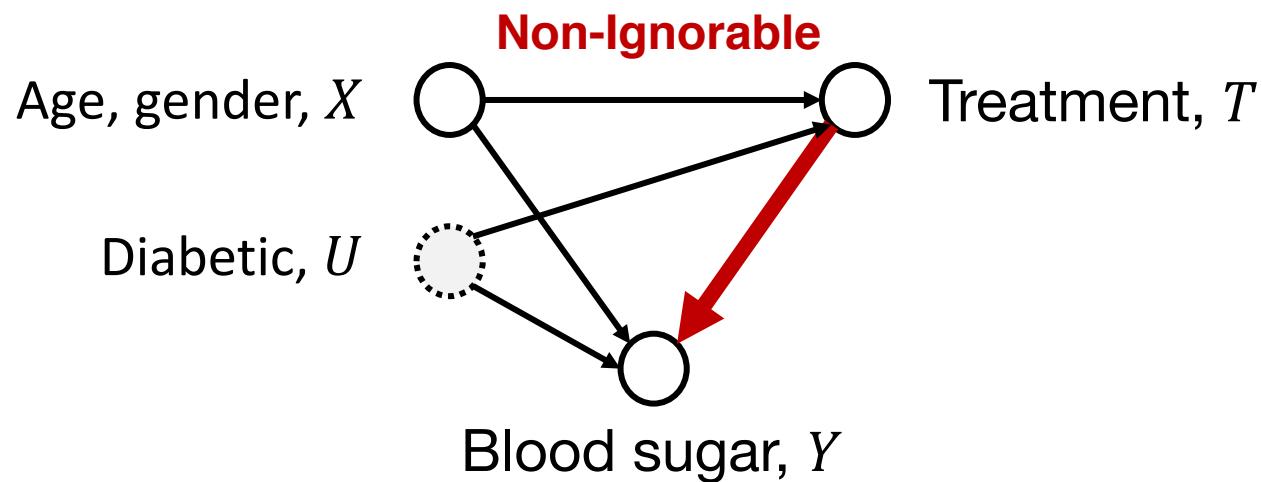
○ Unmeasured / hidden



Ignorability

○ Measured

○ Unmeasured / hidden



Covariate adjustment

- ▶ It can be shown (on a couple of lines) that under the assumptions listed above

$$\mathbb{E}[Y(t) | x] = \mathbb{E}[Y | x, t]$$

- ▶ and as a result

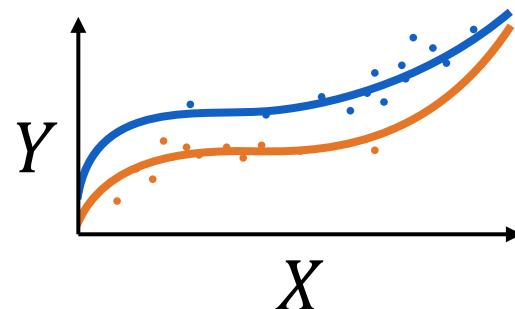
$$ATE = \mathbb{E}[\mathbb{E}[Y | x, T = 1] - \mathbb{E}[Y | x, T = 0]]$$

- ▶ We find (regression) models for $\mathbb{E}[Y | x, t]$ and extrapolate

Estimating both CATE and ATE

- ▶ In fact, $\mathbb{E}[Y | x, t]$ is sufficient to estimate both ATE and CATE under these assumptions:

$$\tau(x) = \mathbb{E}[Y | x, T = 1] - \mathbb{E}[Y | x, T = 0]$$



- ▶ Our job is now to estimate $\mathbb{E}[Y | x, t]$!

1. Potential outcomes framework
2. Supervised learning (risk minimization)
3. Adjusting for distributional shift

Supervised learning

- ▶ Machine learning in general, and deep learning in particular, is great at “function fitting” or supervised learning
- ▶ **Input:** features $X \in \mathbb{R}^d$
Output: label $Y \in \mathbb{R}$ (Think of treatment as in X for now.)
- ▶ Learn from n samples $(x_1, y_1), \dots, (x_n, y_n) \sim p(X, Y)$

Risk minimization

- ▶ Most supervised learning attempt to find models f such that

$$f(x) \approx \mathbb{E}_p[Y \mid X = x]$$

- ▶ often by the **risk minimization** principle

$$f = \operatorname{argmin}_h R_p(h), \quad R_p(h) := \mathbb{E}_p[(h(x) - y)^2]$$

- ▶ Unfortunately, typically only samples are available, not the full expectation
- ▶ Loss is specific to application (e.g. regression). Here, square loss

Empirical risk minimization

- ▶ A common proxy is **Empirical Risk Minimization** (ERM)

$$f = \operatorname{argmin}_h \hat{R}_S(h), \quad \hat{R}_S(h) := \frac{1}{n} \sum_{i=1}^n (h(x_i) - y_i)^2$$

- ▶ where S is a sample from p , $S = (x_1, y_1), \dots, (x_n, y_n)$
- ▶ Supported by learning theory. W.h.p. we generalize from sample

$$R_p(h) \leq \hat{R}_S(h) + \frac{\mathcal{C}}{\sqrt{n}}$$

← Model complexity

Deep learning

- ▶ Deep learning is amazing at empirical risk minimization
- ▶ Incredibly flexible models that are trained by simple algorithms
- ▶ (Opinions highly divided on how/why they generalize and don't overfit)

Estimating potential outcomes

- ▶ We consider learning hypotheses $f(x, t) \approx \mathbb{E}[Y | x, t]$
- ▶ Under ignorability, $f(x, t)$ estimates the potential outcome $Y(t)$
- ▶ Repeat the **risk minimization** principle for each outcome

$$f(\cdot, 0) = \operatorname{argmin}_{h_0} R_p^0(h_0), \quad R_p^0(h_0) := \mathbb{E}_p \left[(h_0(x) - Y(0))^2 \right]$$

$$f(\cdot, 1) = \operatorname{argmin}_{h_1} R_p^1(h_1), \quad R_p^1(h_1) := \mathbb{E}_p \left[(h_1(x) - Y(1))^2 \right]$$



Why is empirical risk minimization
not quite appropriate here?

Estimating potential outcomes

- We don't have samples of $\mathbb{E}_p \left[(h_0(x) - Y(0))^2 \right]$ from $p(X, T)$, only from $p(X | T = 0)$! *We only see the treated outcome for treated*

- No guarantee that

$$\mathbb{E}_p \left[(h_0(x) - Y(0))^2 \right] \approx \frac{1}{n} \sum_{i:t_i=0} (h_0(x_i) - y_i)^2$$

↑
Only control group

- More on this **tomorrow!**

CATE risk minimization

- ▶ Ignoring the issue of empirical risk minimization for now, we seek an objective for **CATE**, and our estimate $\hat{\tau}_f(x)$

$$\hat{\tau}_f(x) = f(x, 1) - f(x, 0)$$

- ▶ We wish to minimize the risk, the **CATE MSE**,

$$R_p^\tau(f) = \mathbb{E}_p \left[(\hat{\tau}_f(x) - \tau(x))^2 \right]$$

- ▶ **Problem!** $\tau(x)$ is never fully observed! (And neither is $R_p^\tau(f)$)

Bounding CATE MSE

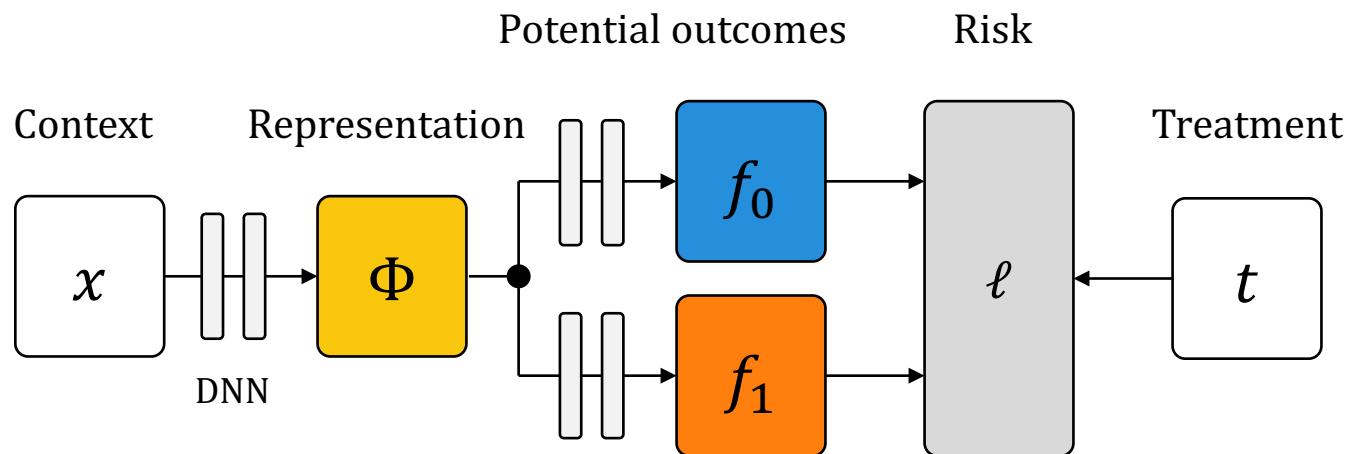
- ▶ $R_p^\tau(f)$ is not observed, but we can bound it!
- ▶ Intuitively, if $f(x, 1)$ and $f(x, 0)$ are good estimates, so is $\hat{\tau}_f(x)$
- ▶ In fact, it is easy to show that (w. relaxed triangle inequality)

$$R_p^\tau(f) \leq 2 \left(R_p^0(f) + R_p^1(f) \right) + \sigma$$

- ▶ Fitting each potential outcome well is sufficient (but not necessary)

Deep learning architecture

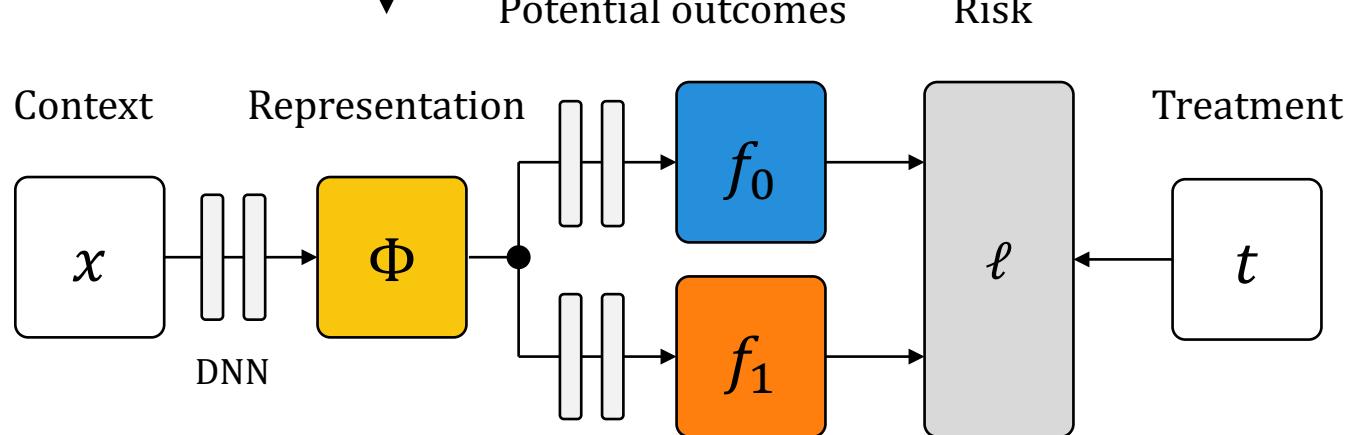
- ▶ Shalit, J., Sontag came up with the following architecture



Deep learning architecture

Shared representation for shared statistical power between groups

Separate heads for different treatments to avoid washing away T



- ▶ This halved the error on a widely used causal effect benchmark!

- 1. Potential outcomes framework**
- 2. Supervised learning (risk minimization)**
- 3. Adjusting for distributional shift**

Tomorrow!