



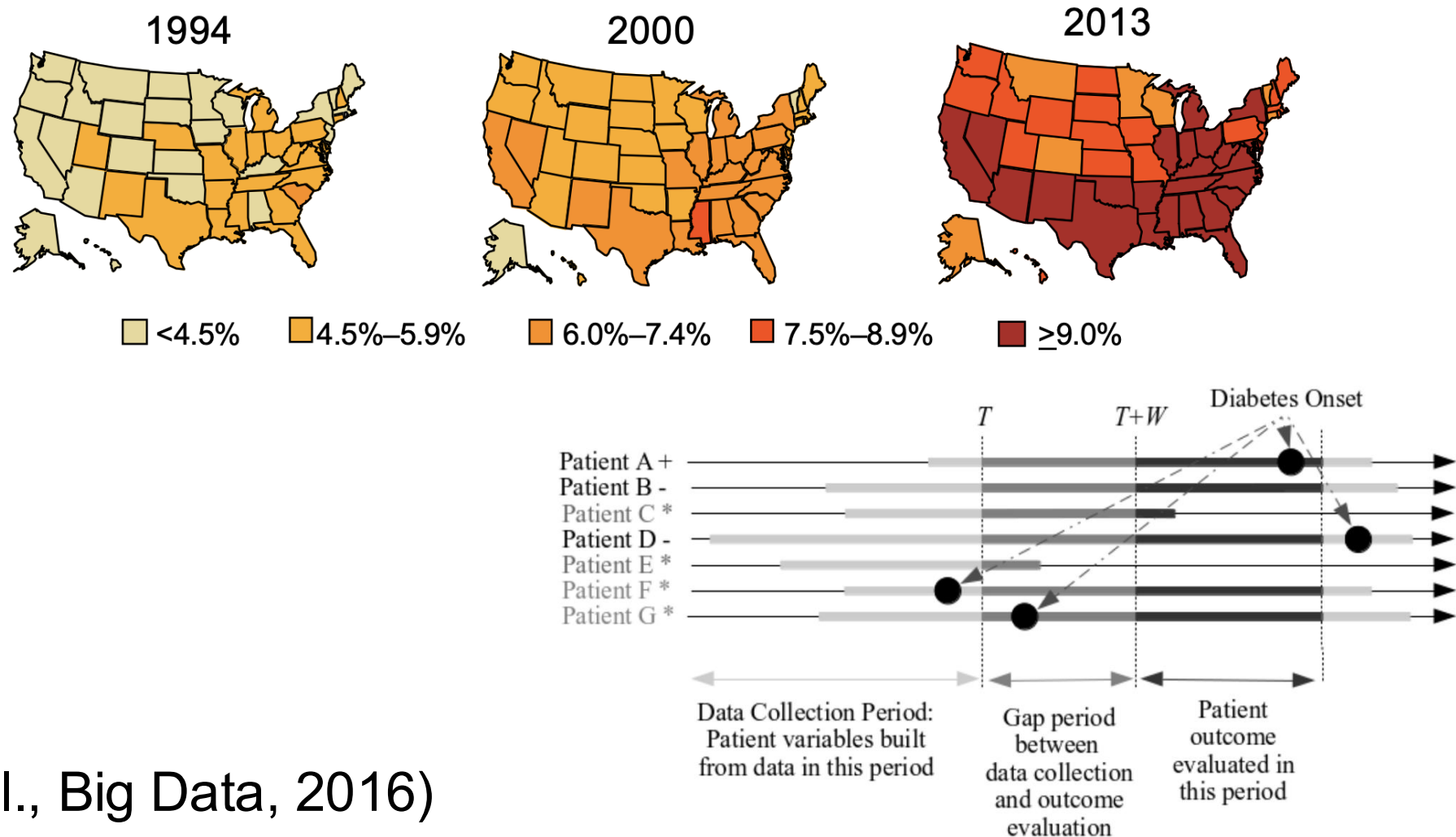
CPH 100: Foundations for CPH

Causal Inference

Irene Y. Chen

Slides adapted from David
Sontag and Uri Shalit

Previously: Risk stratification for type 2 diabetes



(Razavian et al., Big Data, 2016)

Highly weighted features

Impaired Fasting Glucose (Code 790.21)

Abnormal Glucose NEC (790.29)

Hypertension (401)

Obstructive Sleep Apnea (327.23)

Obesity (278)

Abnormal Blood Chemistry (790.6)

Hyperlipidemia (272.4)

Shortness Of Breath (786.05)

Esophageal Reflux (530.81)

O

Additional Disease Risk

Factors Include:

Pituitary dwarfism (253.3),
Hepatomegaly(789.1), Chronic
Hepatitis C (070.54), Hepatitis
(573.3), Calcaneal Spur(726.73),
Thyrotoxicosis without mention
of goiter(242.90), Sinoatrial
Node dysfunction(427.81), Acute
frontal sinusitis (461.1),
Hypertrophic and atrophic
conditions of skin(701.9),
Irregular menstruation(626.4), ...

1.05
(1.78 1.93)

(Razavian et al., Big Data, 2016)

Thinking about interventions

1. Do highly weighted features suggest avenues for preventing onset of diabetes?
 - **Example: Gastric bypass surgery.** Highest negative weight (9th most predictive feature)
 - What is the mathematical justification for thinking of highly weighted features in this way?
2. What happens if the patient did not get diabetes because an intervention made in the gap?
 - How do we **deconvolve** effect of interventions from the prediction task?

Outline

1. What is causal inference? (15 mins)
2. Potential outcomes framework (20 mins)
3. Tools of the trade (15 mins)



How can we make Data
146 better for you?

Learning Objective: Identify and solve causal inference problems

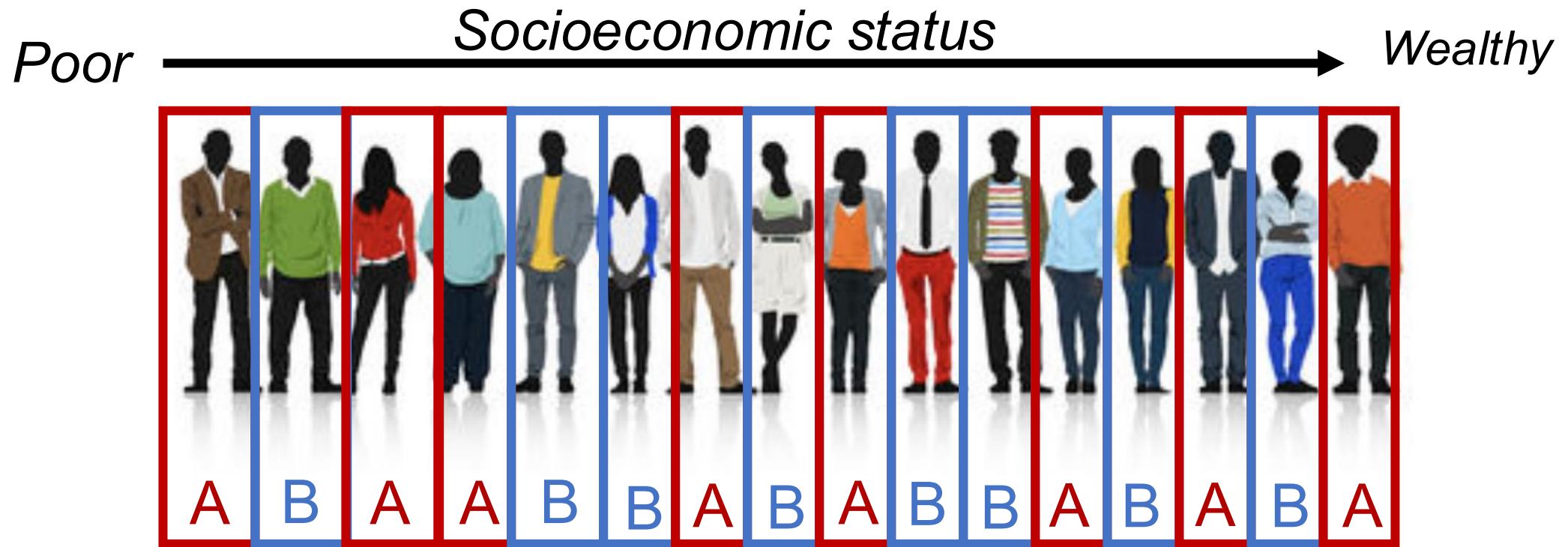
Randomized trials vs observational studies



Which treatment works better?

A or **B**?

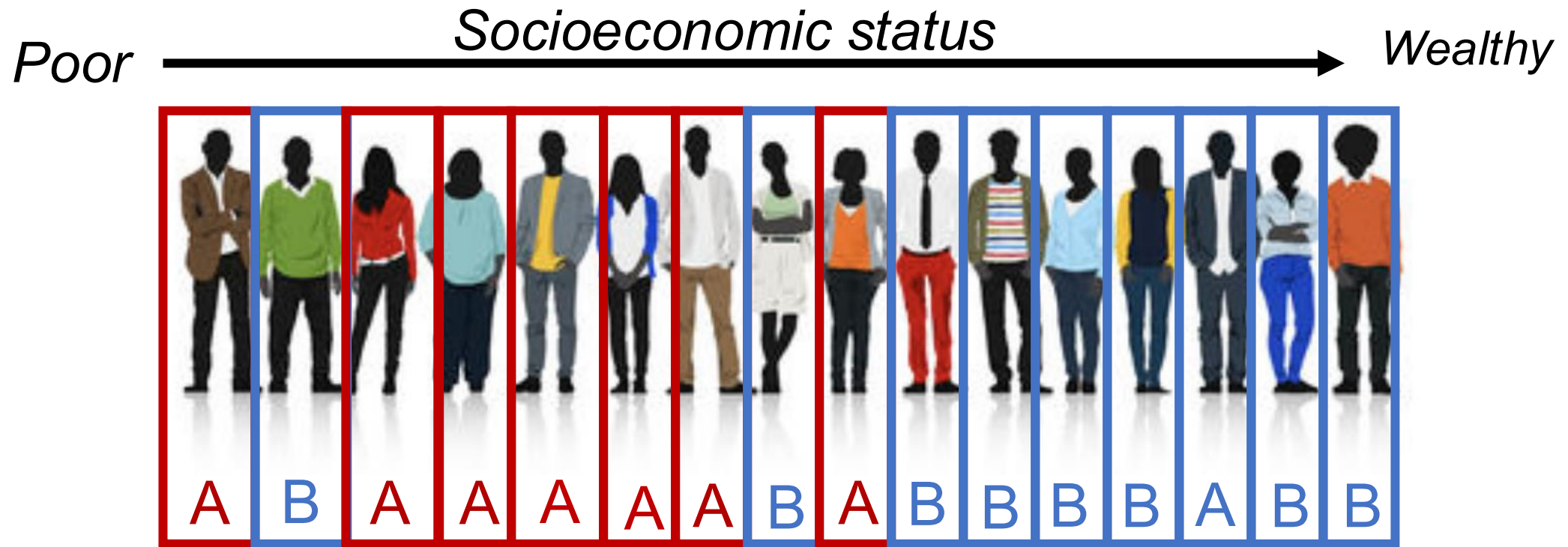
Randomized control trial (RCT)



Which treatment works better?

A or **B**?

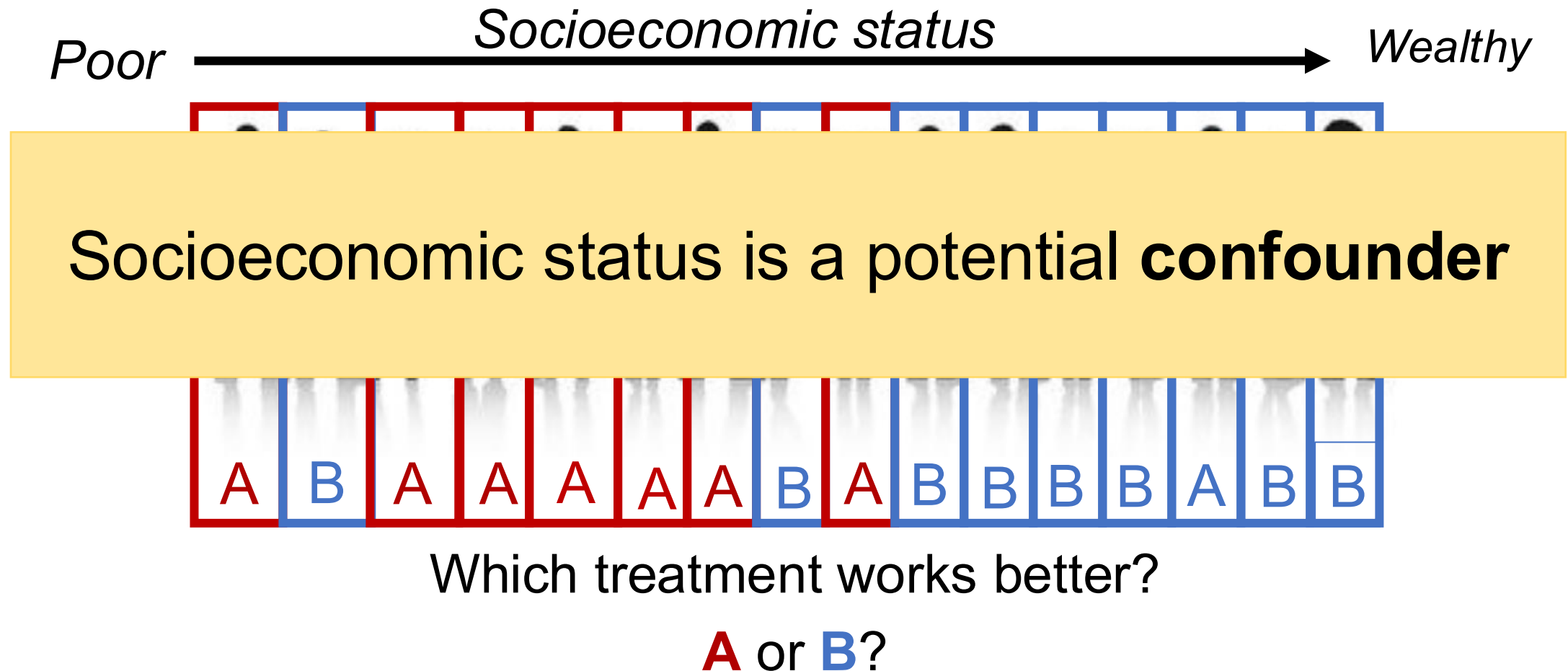
Observational Study



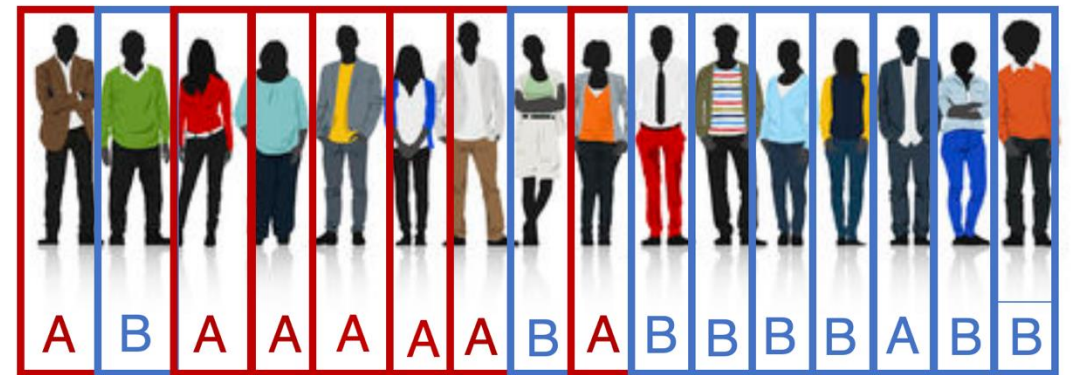
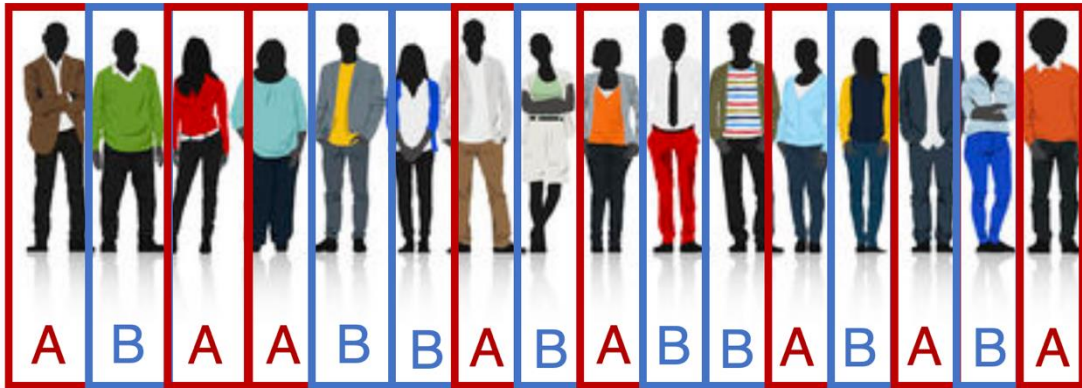
Which treatment works better?

A or **B**?

Observational Study



In many fields, randomized studies are the gold standard for causal inference, but

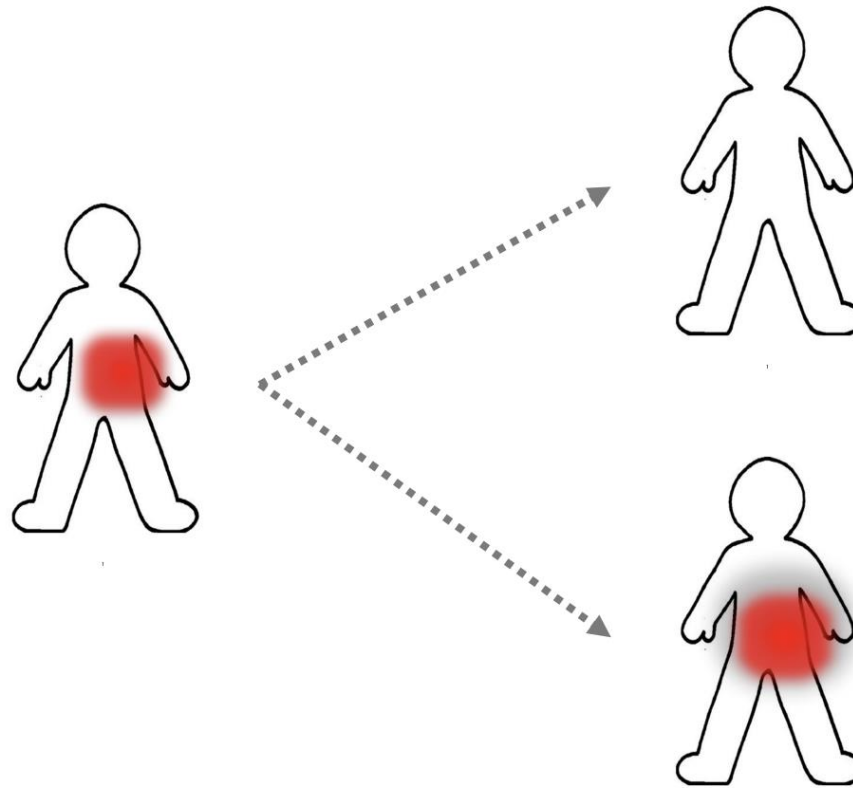


- Does inhaling asbestos cause cancer?
- Does decreasing the interest rate reinvigorate the economy?
- We have a budget for **one** new antidiabetic drug experiment. Can we use past health records of 100,000 diabetics to guide us?

Even randomized controlled trials have flaws

- Not personalized – only population effect
- Study population might not represent true population
 - Recruiting is hard
 - People might drop out of study
 - Study in one company/hospital/state/country could fail to generalize to others
 - Large studies are expensive

Example 1: Individual Treatment Effect (ITE)



Which treatment is best **for me**?

- Which anti-hypertensive treatment?
 - Calcium channel blocker (A)
 - ACE inhibitor (B)
- Current situation:
 - Clinical trials
 - Doctor's knowledge & intuition
- Use datasets of patients and their histories



*Blood pressure = 150/95
WBC count = $6 \cdot 10^9/L$
Temperature = 98°F
HbA1c = 6.6%
Thickness of heart artery
plaque = 3mm
Weight = 65kg*

Which treatment is best **for me**?

- Which anti-hypertensive treatment?
 - Calcium channel blocker (A)
 - ACE inhibitor (B)
- Potential *confounder*: maybe rich patients got more often, and poor patients got medication



Example 2: Average Treatment Effect (ATE)



Should the government fund job-training programs?

- Existing job training programs seem to help unemployed and underemployed find better jobs
- Should the government fund such programs?
- Maybe training helps but only marginally? Is it worth the investment?
- **Average Treatment Effect (ATE):** Potential confounder: Maybe only motivated people go to job training? Maybe they would have found better jobs anyway?

A major challenge in causal inference from observational studies is how to **control** or **adjust** for the confounding factors

Counterfactuals and causal inference

- Does treatment T cause outcome Y ?
- If T had not occurred, Y would not have occurred (David Hume)
- Counterfactual: Kim received job training (T), and her income one year later (Y) is 20,000\$ What would have been Kim's income had she not had job training?

Counterfactuals and causal inference

- Does treatment T cause outcome Y ?
- If T had not occurred, Y would not have occurred (David Hume)
- Counterfactual: Kim received job training (T), and her income one year later (Y) is 20,000\$ What would have been Kim's income had she not had job training?
- ... But you may never observe the counterfactual!

Potential Outcomes Framework (Rubin-Neyman Causal Model)

- Each unit x_i has two potential outcomes:
 - $Y_0(x_i)$ is the potential outcome had the unit not been treated: “**control outcome**”
 - $Y_1(x_i)$ is the potential outcome had the unit been treated: “**treated outcome**”

- Individual Treatment Effect for unit i :

$$ITE(x_i) = \mathbb{E}_{Y_1 \sim P(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim P(Y_0|x_i)}[Y_0|x_i]$$

- Average Treatment Effect:

$$ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{x \sim p(x)}[ITE(x)]$$

Potential Outcomes Framework (Rubin-Neyman Causal Model)

- Each unit x_i has two potential outcomes:
 - $Y_0(x_i)$ is the potential outcome had the unit not been treated: “**control outcome**”
 - $Y_1(x_i)$ is the potential outcome had the unit been treated: “**treated outcome**”

- Observed factual outcome:

$$y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$$

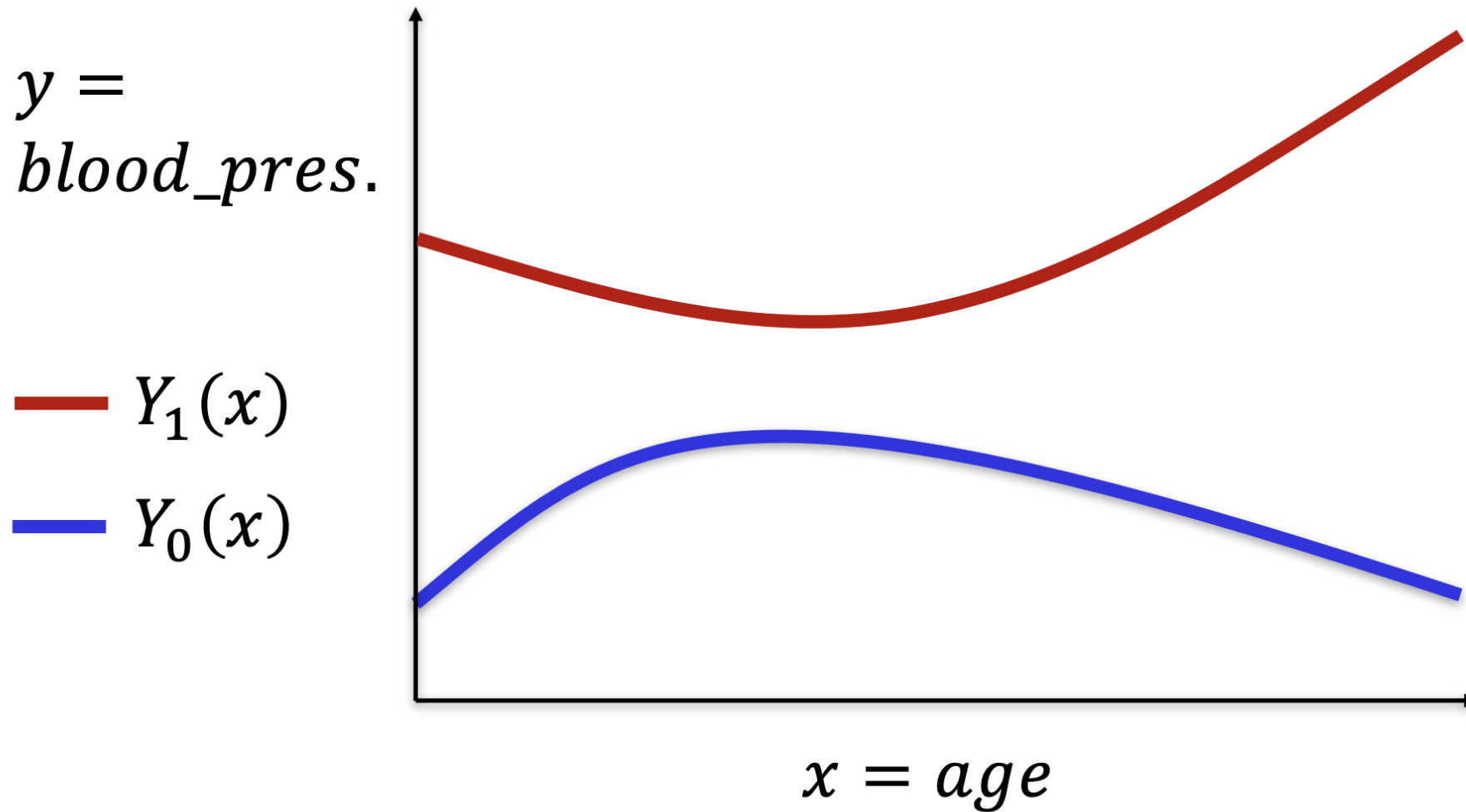
- Unobserved counterfactual outcome:

$$y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$

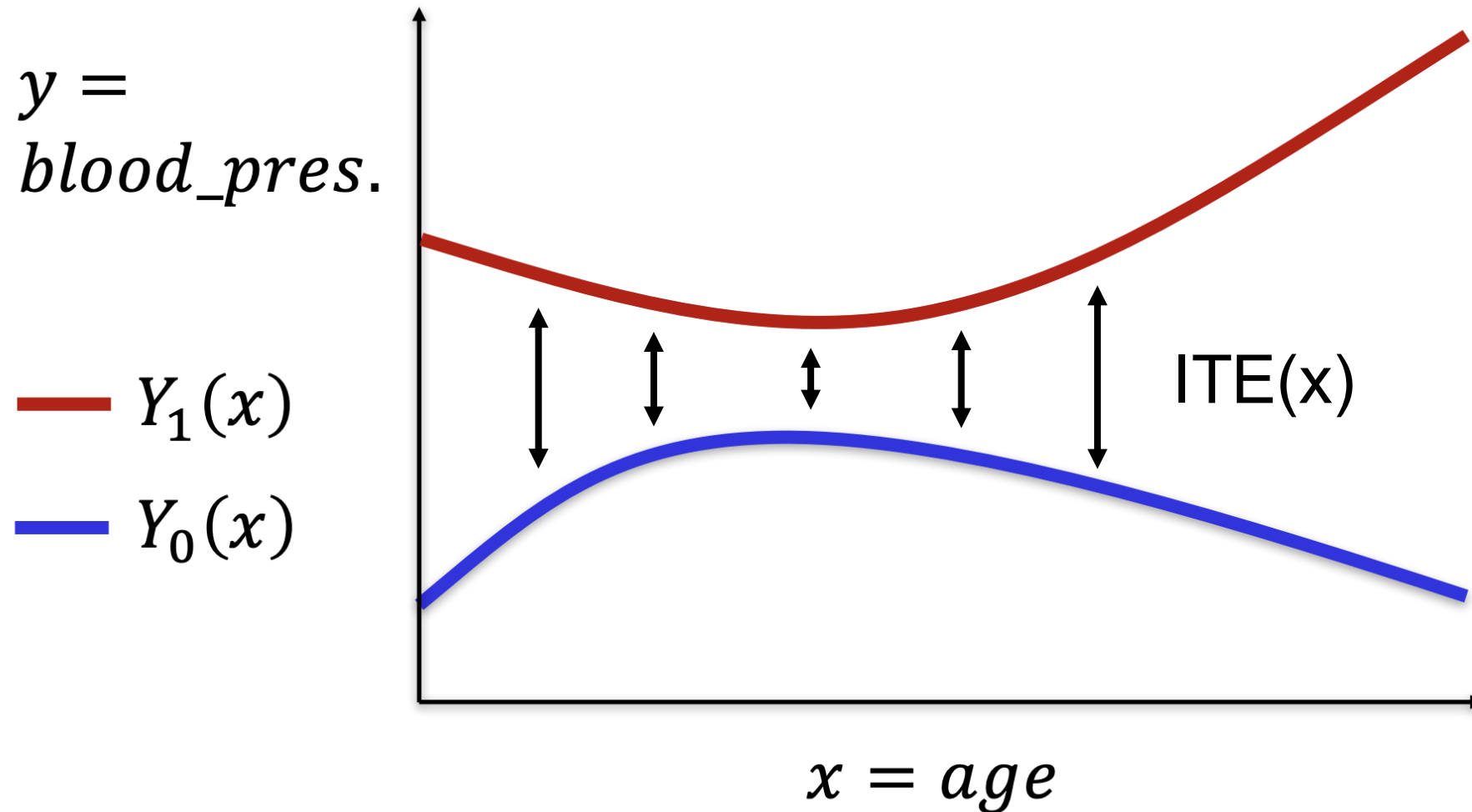
Terminology

- *Unit*: data point, e.g. patient, customer, student
- *Treatment*: binary indicator (in this tutorial)
Also called intervention
- *Treated*: units who received treatment=1
- *Control*: units who received treatment=0
- *Factual*: the set of observed units with their respective treatment assignment
- *Counterfactual*: the factual set with flipped treatment assignment

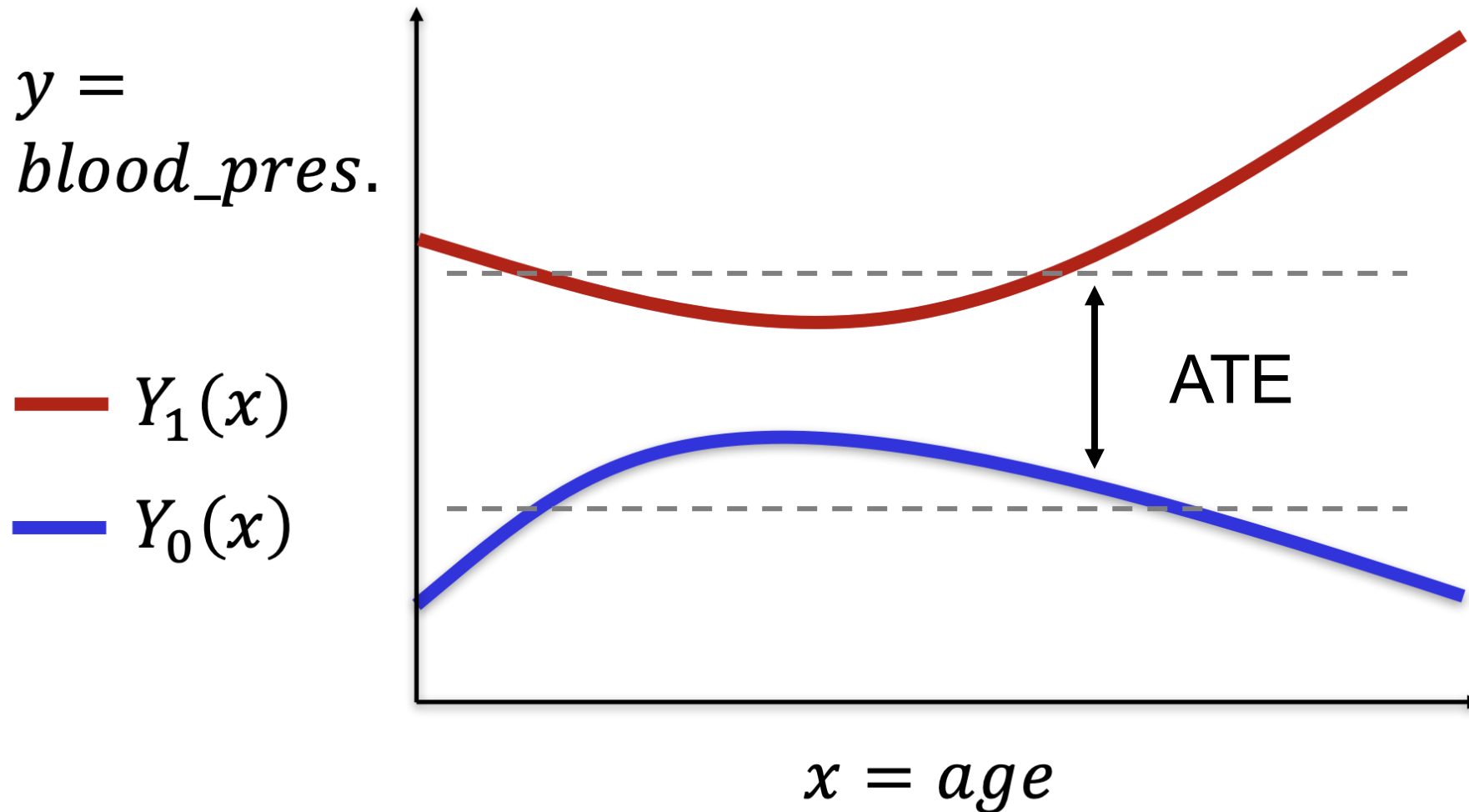
Example: Blood pressure and age



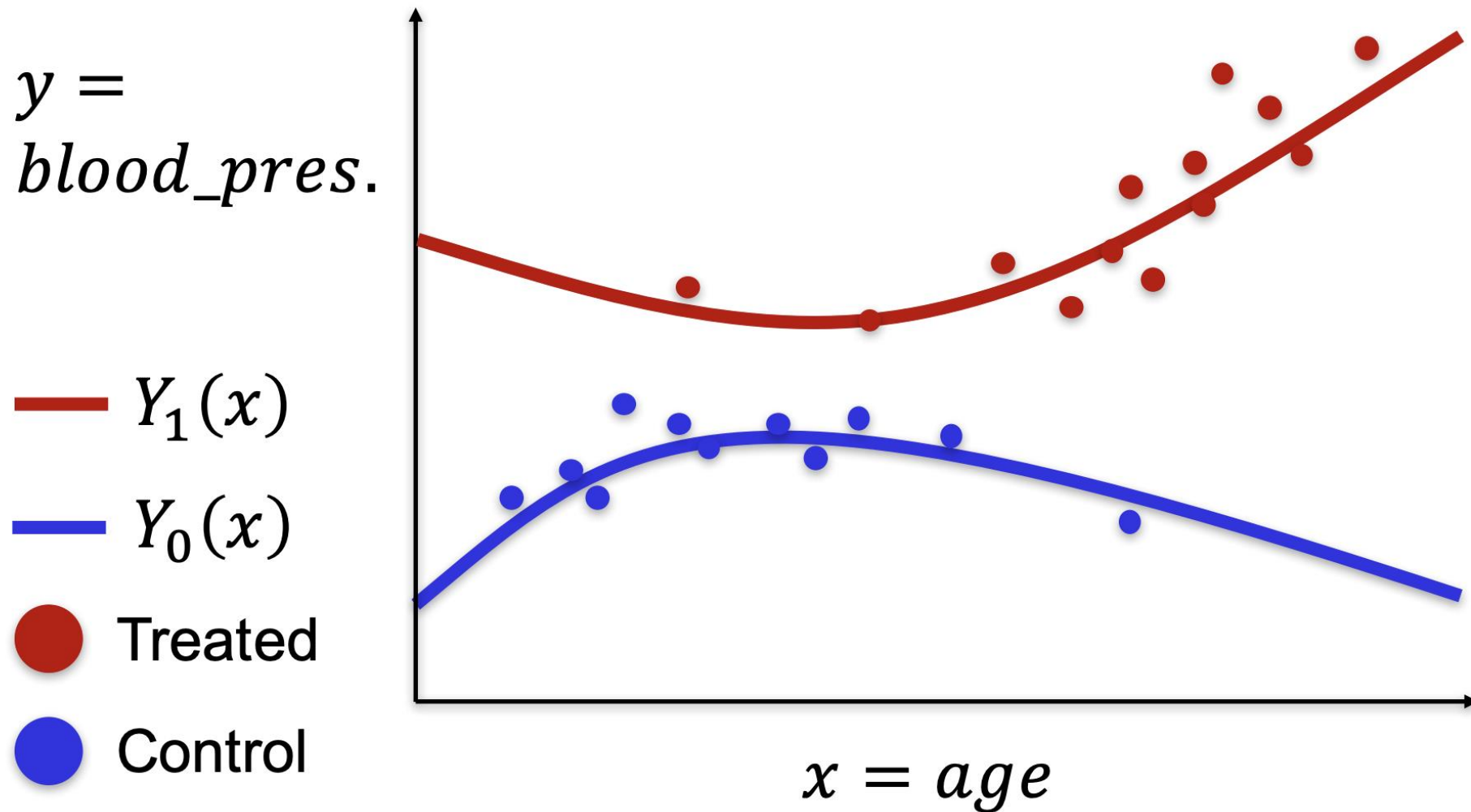
Example: Blood pressure and age



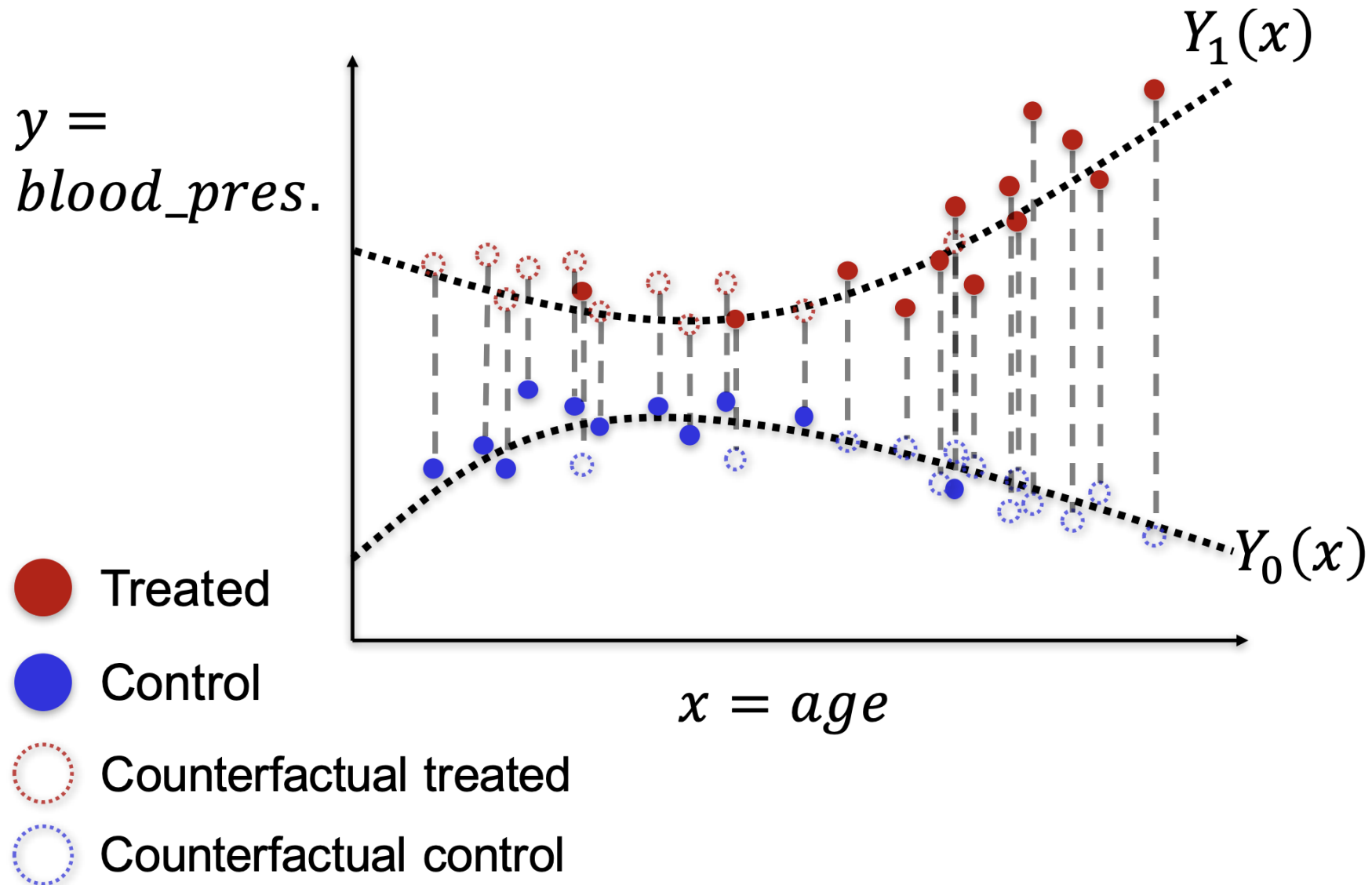
Example: Blood pressure and age



Example: Blood pressure and age



Example: Blood pressure and age



“The fundamental problem” of causal inference is that we only observe one of the two outcomes

Key Assumption 1: No unmeasured confounders

- Y_0, Y_1 : potential outcomes for control and treated
- x : unit covariates (features)
- T : treatment assignment

- We assume:

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

- The potential outcomes are independent of treatment assignment, conditioned on covariates x

Key Assumption 1: No unmeasured confounders

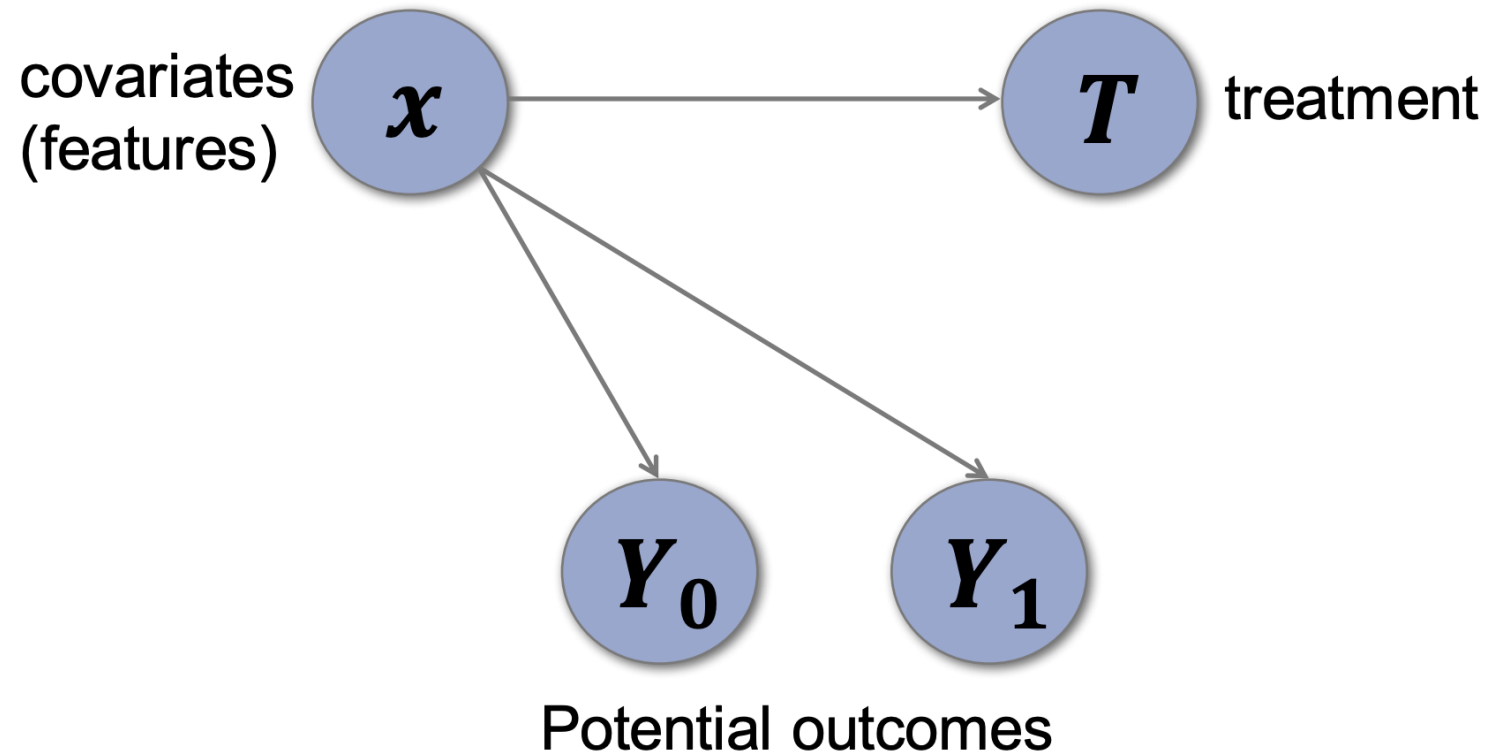
- Y_0, Y_1 : potential outcomes for control and treated
- x : unit covariates (features)
- T : treatment assignment

- We assume:

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

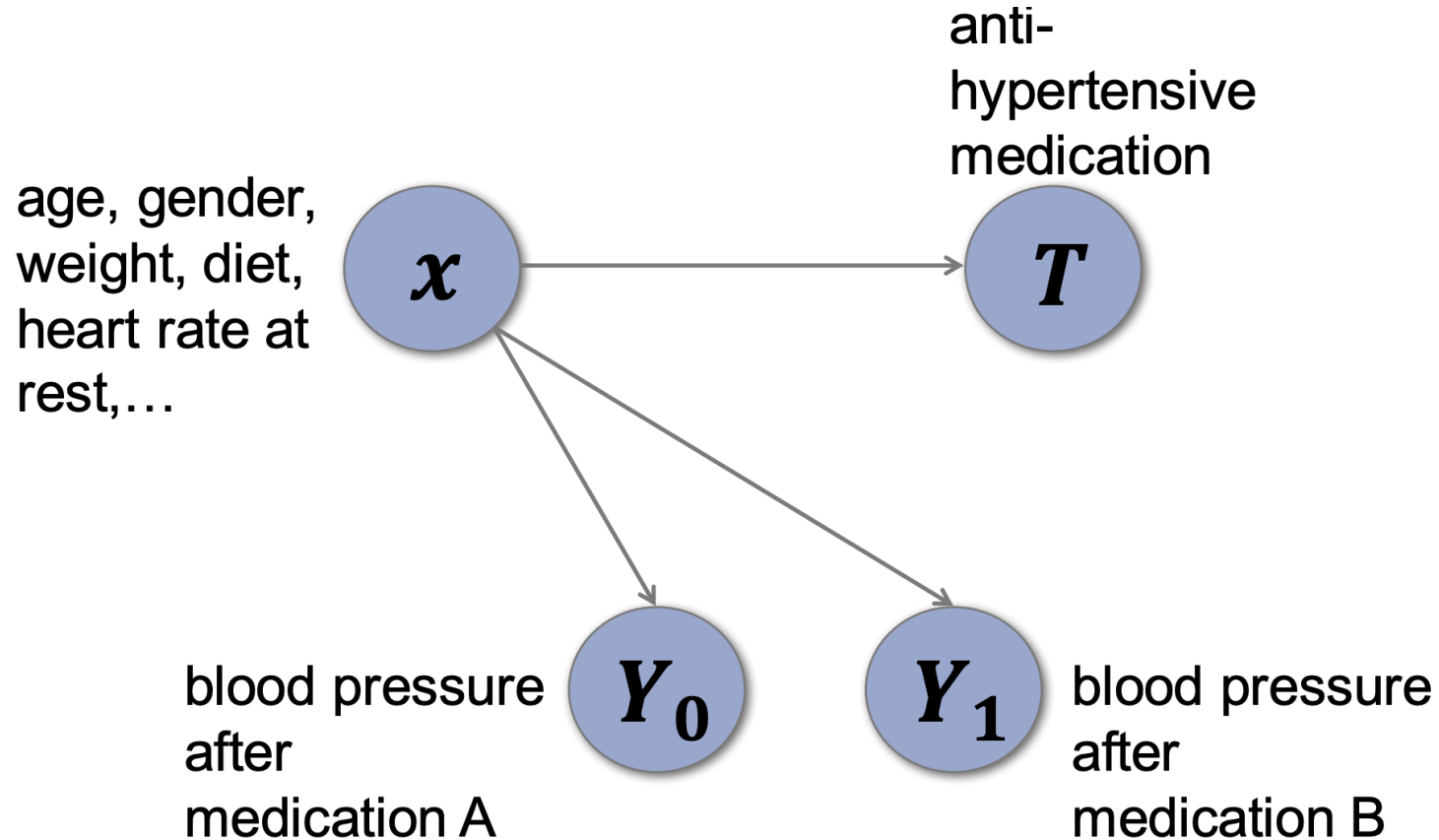
- Defined as **ignorability**

Ignorability



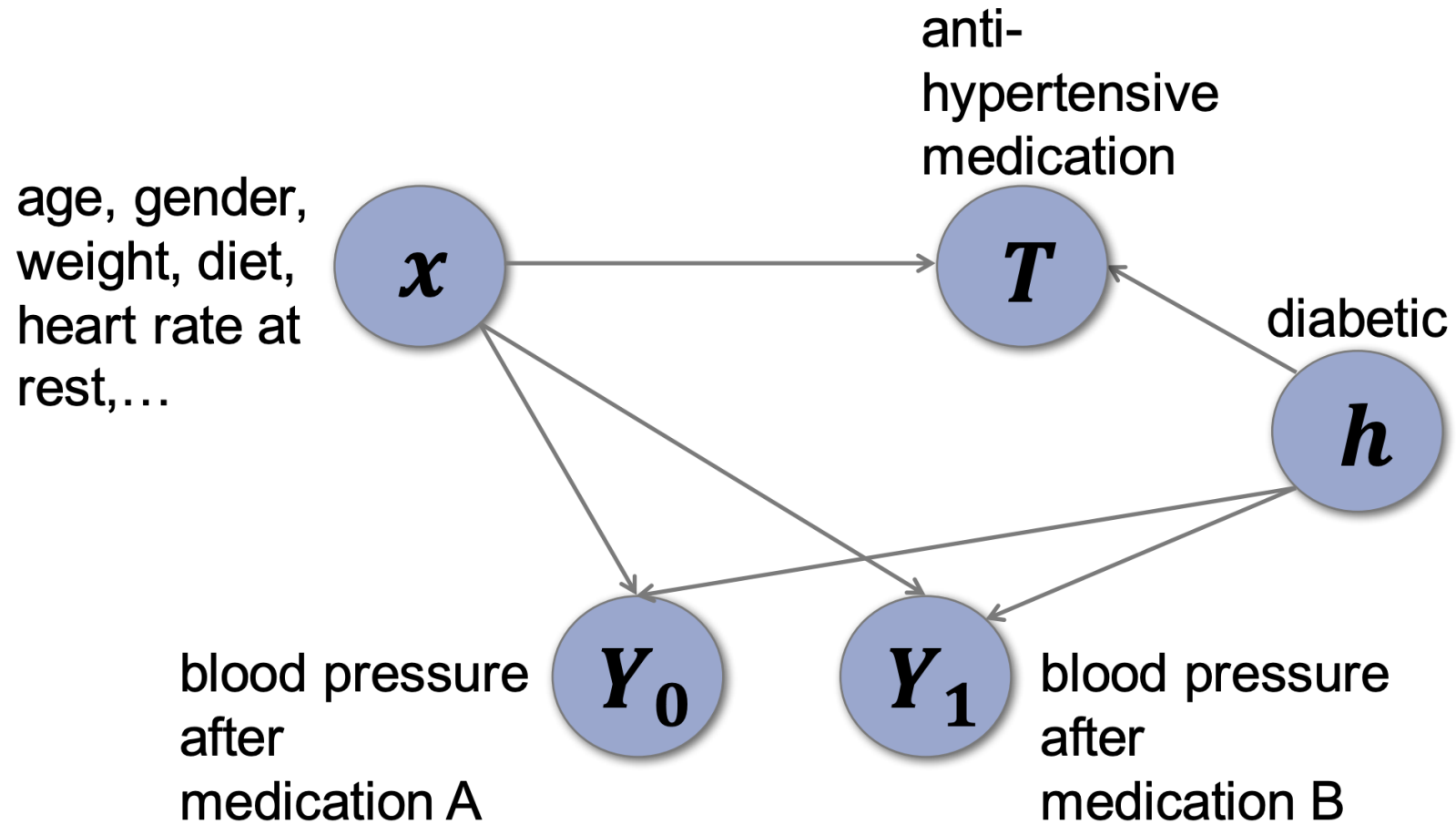
$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

Ignorability



$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

No Ignorability



$$(Y_0, Y_1) \not\perp T \mid x$$

Key Assumption 2: Common support

- Y_0, Y_1 : potential outcomes for control and treated
- x : unit covariates (features)
- T : treatment assignment
- We assume:

$$p(T = t | X = x) > 0 \quad \forall t, x$$

Average Treatment Effect: the adjustment formula

- Assuming ignorability, we will derive *the adjustment formula* (Hernán & Robins 2010, Pearl 2009)
- The adjustment formula is extremely useful in causal inference
- Also called *G-formula*

Average Treatment Effect

$$ATE := \mathbb{E}[Y_1 - Y_0]$$

Average Treatment Effect

$$ATE := \mathbb{E}[Y_1 - Y_0]$$

$$\mathbb{E}[Y_1] = \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_1 \sim p(Y_1|x)}[Y_1|x] \right] \text{ Law of total expectation}$$

Average Treatment Effect

$$ATE := \mathbb{E}[Y_1 - Y_0]$$

$$\mathbb{E}[Y_1] = \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1 | x] \right]$$

$$= \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1 | x, T = 1] \right]$$

Ignorability
 $(Y_0, Y_1) \perp T | x$

Average Treatment Effect

$$ATE := \mathbb{E}[Y_1 - Y_0]$$

$$\mathbb{E}[Y_1] = \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1 | x] \right]$$

$$= \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1 | x, T = 1] \right]$$

$$= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_1 | x, T = 1]]$$

Shorter
notation

Average Treatment Effect

$$ATE := \mathbb{E}[Y_1 - Y_0]$$

$$\begin{aligned}\mathbb{E}[Y_0] &= \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_0 \sim p(Y_0|x)} [Y_0|x] \right] \\ &= \mathbb{E}_{x \sim p(x)} \left[\mathbb{E}_{Y_0 \sim p(Y_0|x)} [Y_0|x, T = 0] \right] \\ &= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_0|x, T = 1]]\end{aligned}$$

The adjustment formula

Under the assumption of ignorability, we have:

$$\begin{aligned}ATE &:= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_1 | x, T = 1] - \mathbb{E}[Y_0 | x, T = 0]]\end{aligned}$$

The adjustment formula

Under the assumption of ignorability, we have:

$$\begin{aligned} ATE &:= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_1 | x, T = 1] - \mathbb{E}[Y_0 | x, T = 0]] \end{aligned}$$

We can estimate these two quantities from data!

The adjustment formula

$$\begin{aligned}ATE &:= \mathbb{E}[Y_1 - Y_0] \\&= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_1 | x, T = 1] - \mathbb{E}[Y_0 | x, T = 0]]\end{aligned}$$

We can estimate:

$$\mathbb{E}[Y_1 | x, T = 1]$$

$$\mathbb{E}[Y_0 | x, T = 0]$$

The adjustment formula

$$\begin{aligned}ATE &:= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_1 | x, T = 1] - \mathbb{E}[Y_0 | x, T = 0]]\end{aligned}$$

We can estimate:

$$\mathbb{E}[Y_1 | x, T = 1]$$

$$\mathbb{E}[Y_0 | x, T = 0]$$

We cannot estimate:

$$\mathbb{E}[Y_0 | x, T = 1]$$

$$\mathbb{E}[Y_1 | x, T = 0]$$

$$\mathbb{E}[Y_1 | x]$$

$$\mathbb{E}[Y_0 | x]$$

The adjustment formula

$$\begin{aligned}ATE &:= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y_1 | x, T = 1] - \mathbb{E}[Y_0 | x, T = 0]]\end{aligned}$$

We can estimate:

$$\mathbb{E}[Y_1 | x, T = 1]$$

$$\mathbb{E}[Y_0 | x, T = 0]$$

Instead, we can collect

samples of $p(x|T = 1)$

and $p(x|T = 0)$ and

extrapolate to $p(x)$

Tools of the trade

1. Matching
2. Covariate adjustment
3. Propensity scoring

Matching setup

- Samples $x_1, x_2, \dots, x_N$
- Observed binary treatment assignments $t_1, t_2, \dots, t_N$
- Observed outcomes: $y_1, y_2, \dots, y_N$
- $x = (\text{age, gender, married, education, income_last_year, ...})$
- $t = (\text{job_training, no_job_training})$
- $y = \text{income_one_year_after_job_training}$
- Does job training increase future income?

Tools of the trade

1. Matching
2. Covariate adjustment
3. Propensity scoring

Matching

- For each unit, find its long-lost counterfactual twin, check up on his outcome



Laufey if she had gone into cello

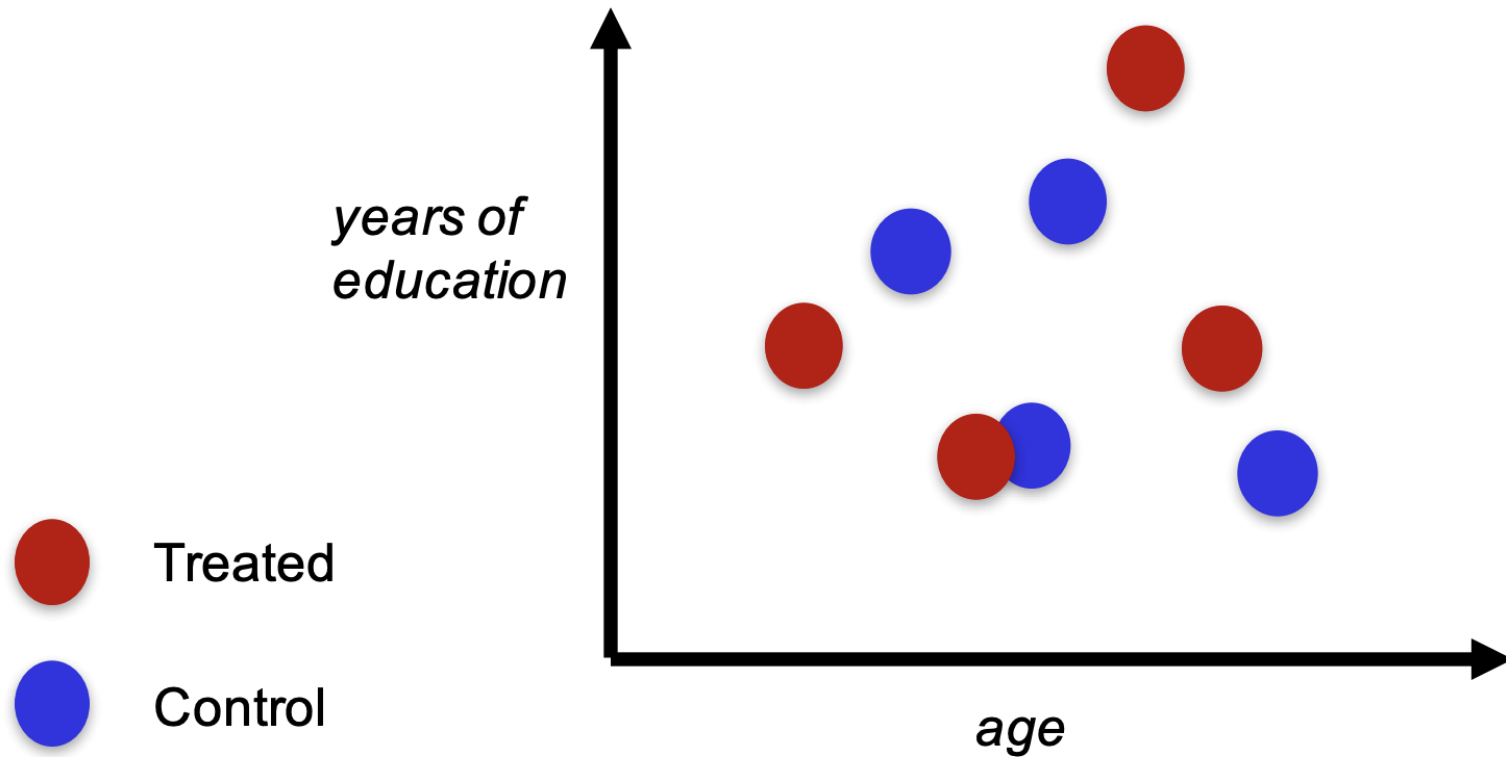


Laufey if she had gone into jazz pop

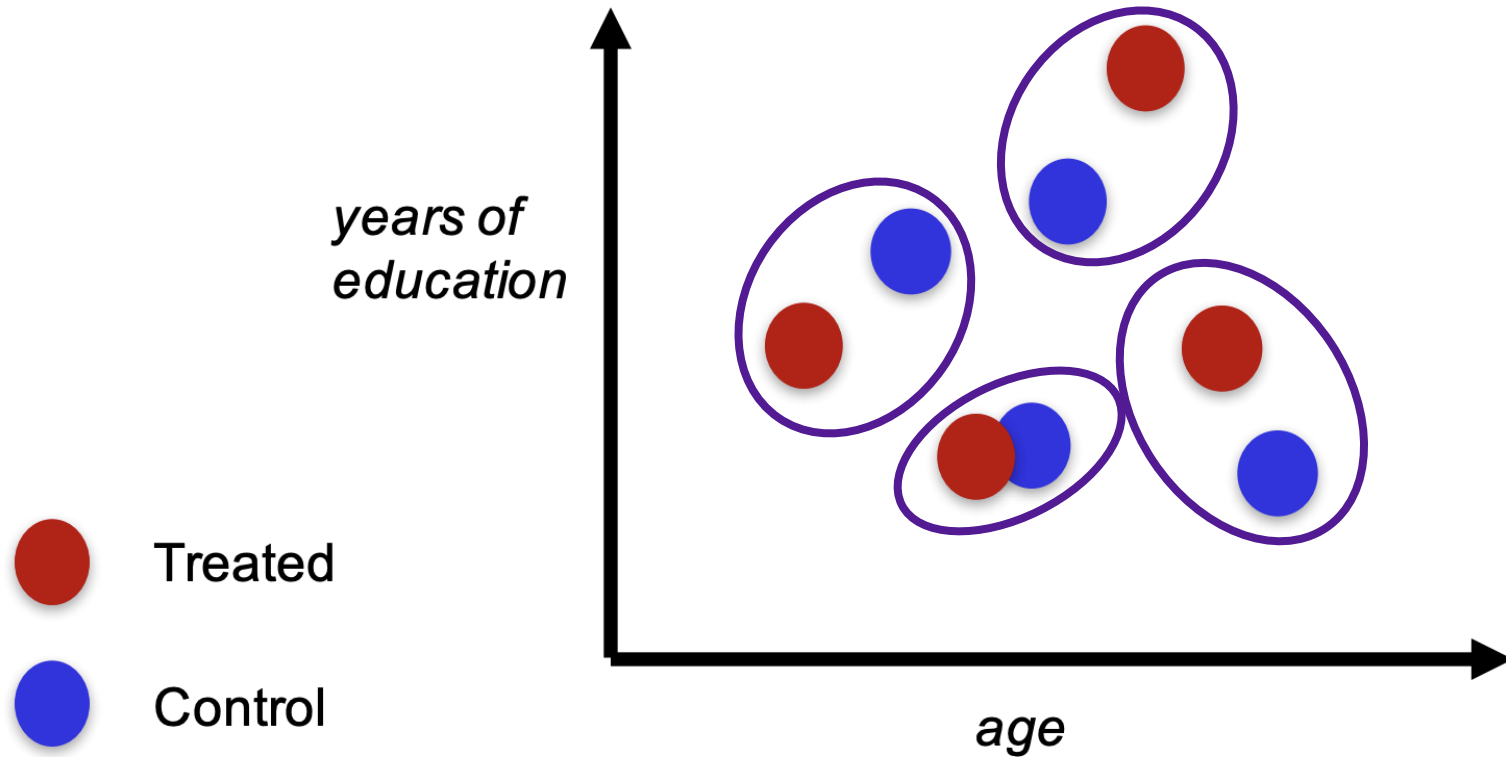
Matching

- Find each unit's long-lost counterfactual identical twin, check up on her outcome
- Used for estimating both ATE and ITE

Match to nearest neighbor from opposite group



Match to nearest neighbor from opposite group



1-NN Matching

- Let $d(.,.)$ be the metric between two x values
- For each i , define $j(i) = \operatorname{argmin}_{j \text{ s.t. } t_j \neq t_i} d(x_i, x_j)$
 - $j(i)$ is the nearest counterfactual neighbor of i
- If $t_i = 1$, unit i is treated:
 - $ITE(x_i) = y_i - y_{j(i)}$
- If $t_i = 0$, unit i is untreated:
 - $ITE(x_i) = y_{j(i)} - y_i$

1-NN Matching

- Let $d(.,.)$ be the metric between two x values
- For each i , define $j(i) = \operatorname{argmin}_{j \text{ s.t. } t_j \neq t_i} d(x_i, x_j)$
 - $j(i)$ is the nearest counterfactual neighbor of i
- $ITE(x_i) = (2t_i - 1)(y_i - y_{j(i)})$
- $ATE = \frac{1}{n} \sum_{i=1}^n ITE(x_i)$

Matching

- Interpretable, especially in small-sample regime
- Nonparametric
- Heavily reliant on the underlying metric (however see below about propensity score matching)
- Could be misled by features which don't affect the outcome

Other matching methods

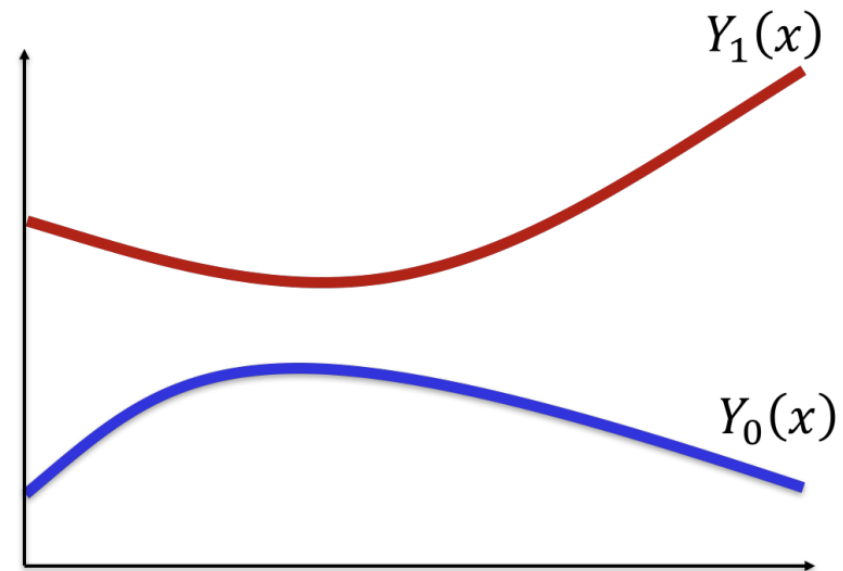
- Coarsened exact matching
 - Iacus et al. (2011)
- Optimal matching
 - Rosenbaum (1989, 2002)
- Propensity score matching
 - Rosenbaum & Rubin (1983), Austin (2011)
- Mahalanobis distance matching
 - Rosenbaum (1989, 2002)

Tools of the trade

1. Matching
2. Covariate adjustment
3. Propensity scoring

Covariate adjustment

- Explicitly model the relationship between treatment, confounders, and outcome
- Also called “Response Surface Modeling”
- Used for both ITE and ATE
- A regression problem



Covariates
(Features)

x_1

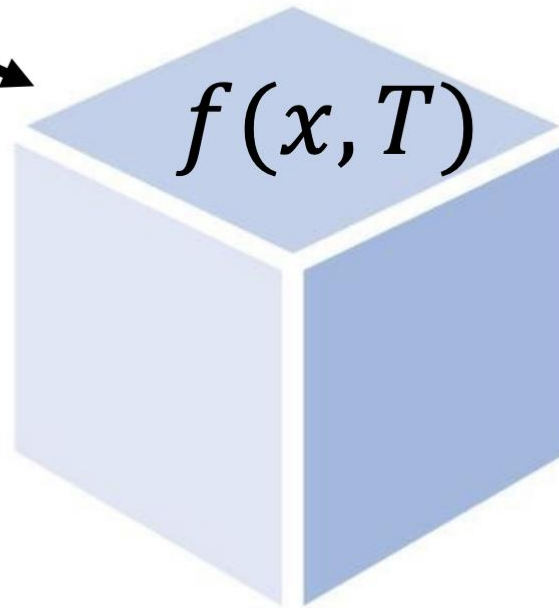
x_2

$\vdots$

x_d

T

Regression
model



Outcome

y

Covariates
(Features)

x_1

x_2

$\vdots$

x_d

Regression
model

$f(x, T)$

Outcome

y

Parameter of
interest

T

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :

$$\mathbb{E}_{x \sim p(x)} \left[\mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right]$$

- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \left(f(x_i, 1) - f(x_i, 0) \right)$$

Covariate adjustment (parametric g-formula)

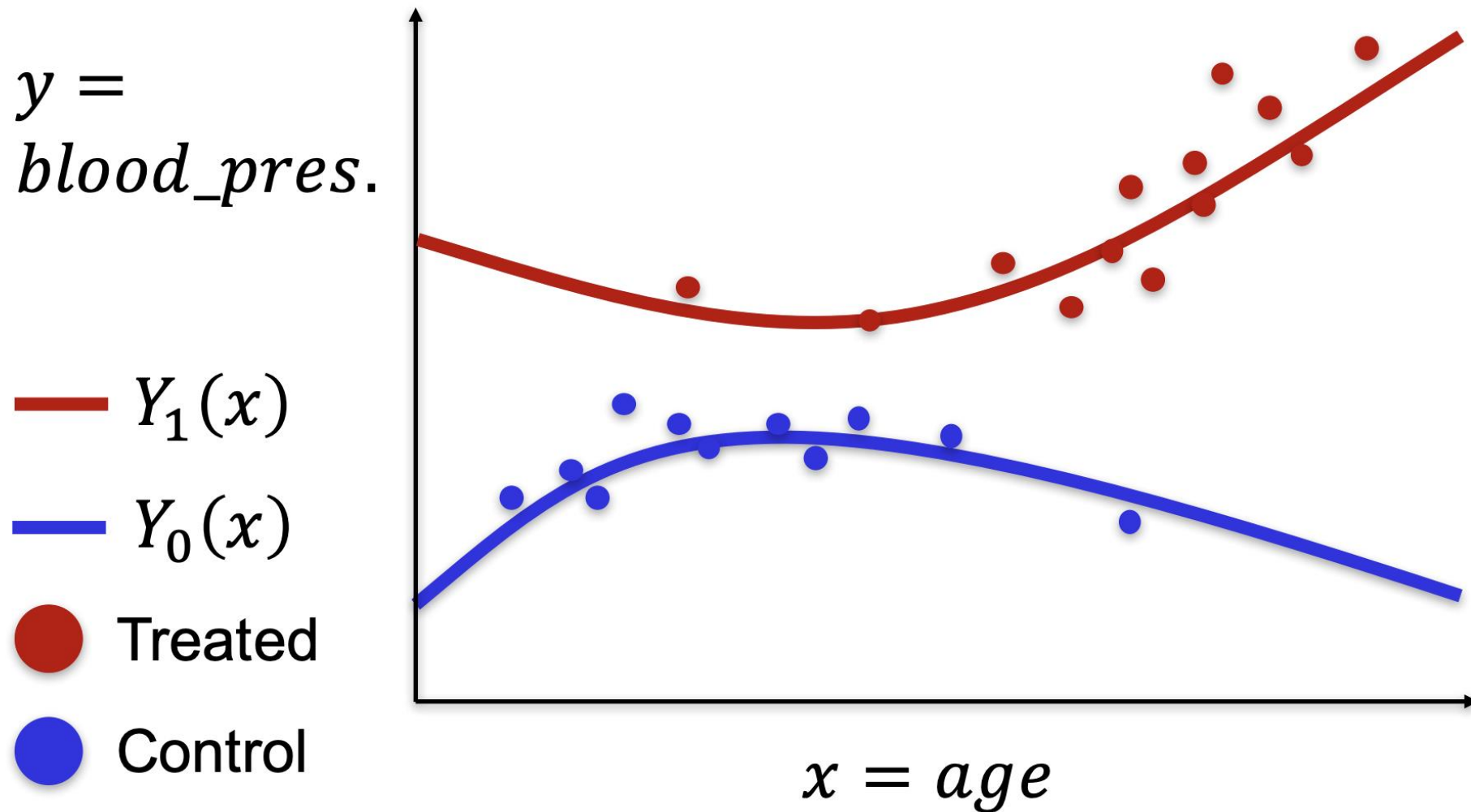
- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :

$$\mathbb{E}_{x \sim p(x)} \left[\mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right]$$

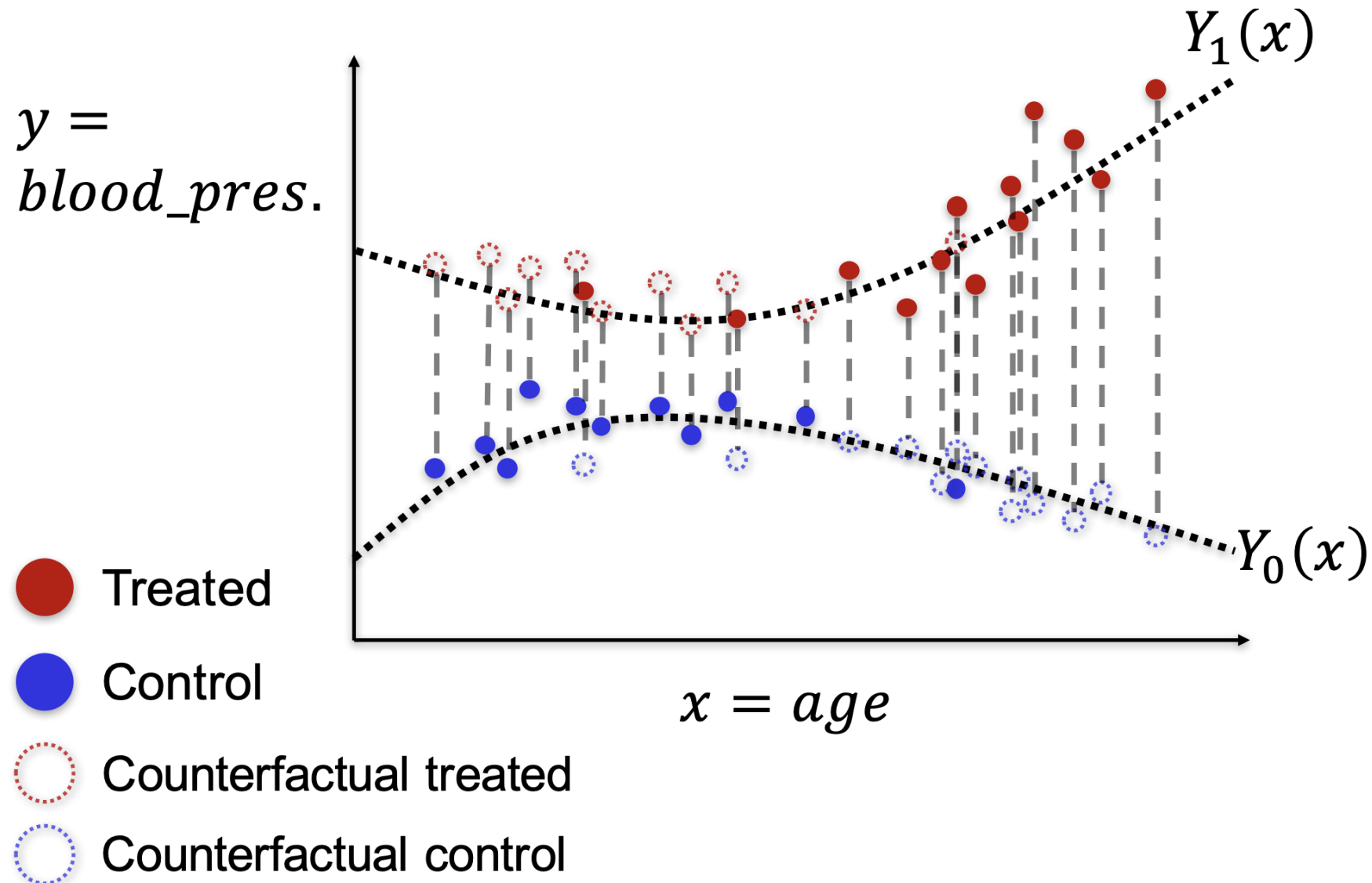
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{ITE}(x_i) = f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment



Covariate adjustment



Warning: this is **not** a classic supervised learning problem

- Our model was optimized to predict outcome, not to differentiate the influence of A vs. B
- What if our high-dimensional model threw away the feature of medication A/B?
- Maybe the model never saw a patient like Anna get medication A? Maybe there's a reason patients like Anna never get A?

Covariate adjustment - consistency

- If the model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$ is consistent in the limit of infinite samples, then under ignorability the estimated ATE will converge to the true ATE
- A sufficient condition: overlap and well-specified model

Using ML for causal inference

- Machine learning techniques can be very useful and have recently seen wider adoption
- Random forests and Bayesian trees
 - Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
 - Hoyer et al. (2009), Zigler et al. (2012)
- Neural nets
 - Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)
- “Causal” Lasso
 - Belloni et al. (2013), Farrell (2015), Athey et al. (2016)

Covariate adjustment and matching

- Matching is equivalent to covariate adjustment with two 1-NN classifiers:

$$\hat{Y}_1(x) = y_{NN_1(x)} , \hat{Y}_0(x) = y_{NN_0(x)}$$

where $y_{NN_t(x)}$ is the nearest-neighbor of x among units with treatment assignment $t = 0,1$

- 1-NN matching is generally inconsistent though only with small bias (Imbens 2004)

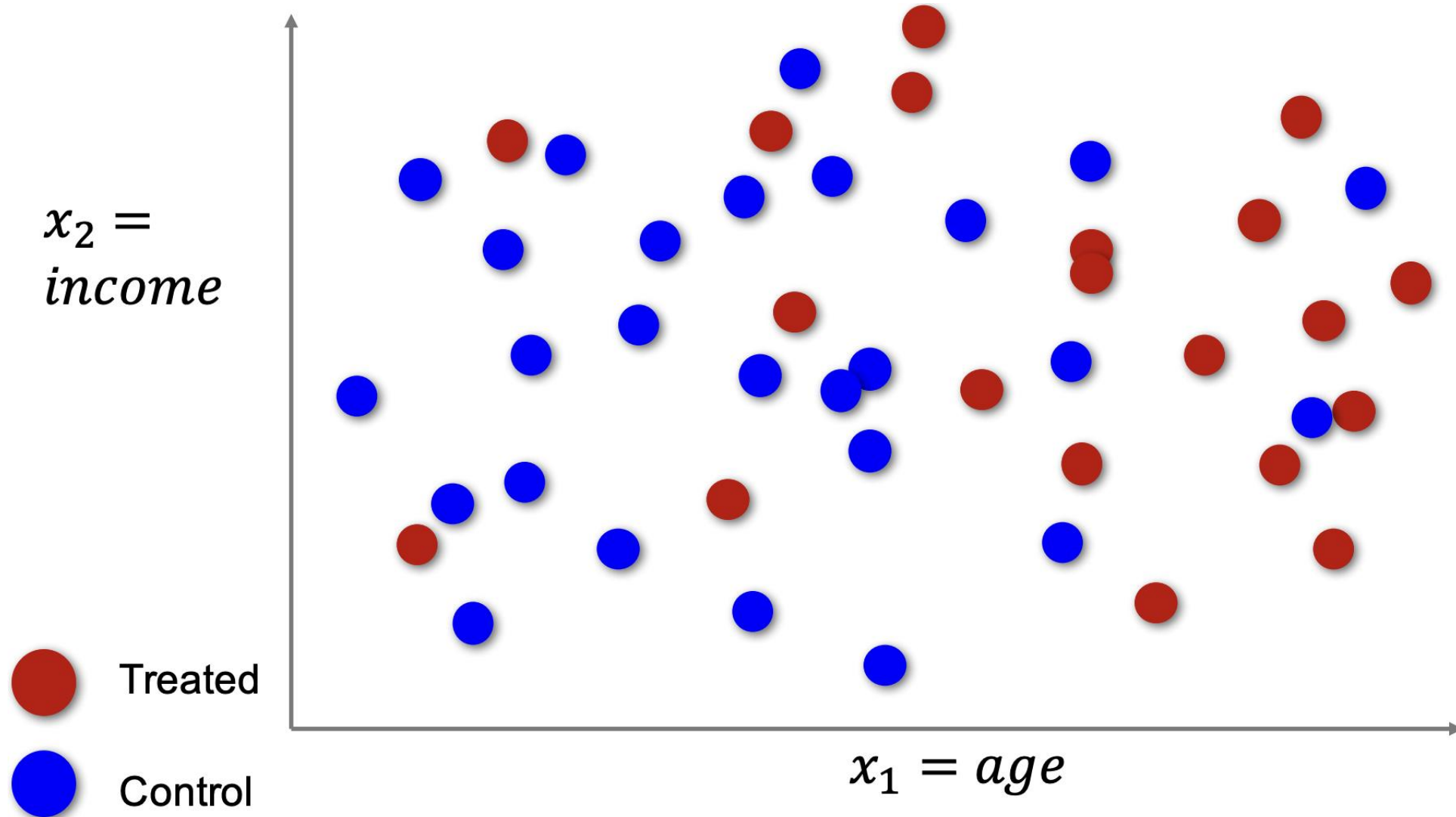
Tools of the trade

1. Matching
2. Covariate adjustment
3. Propensity scoring

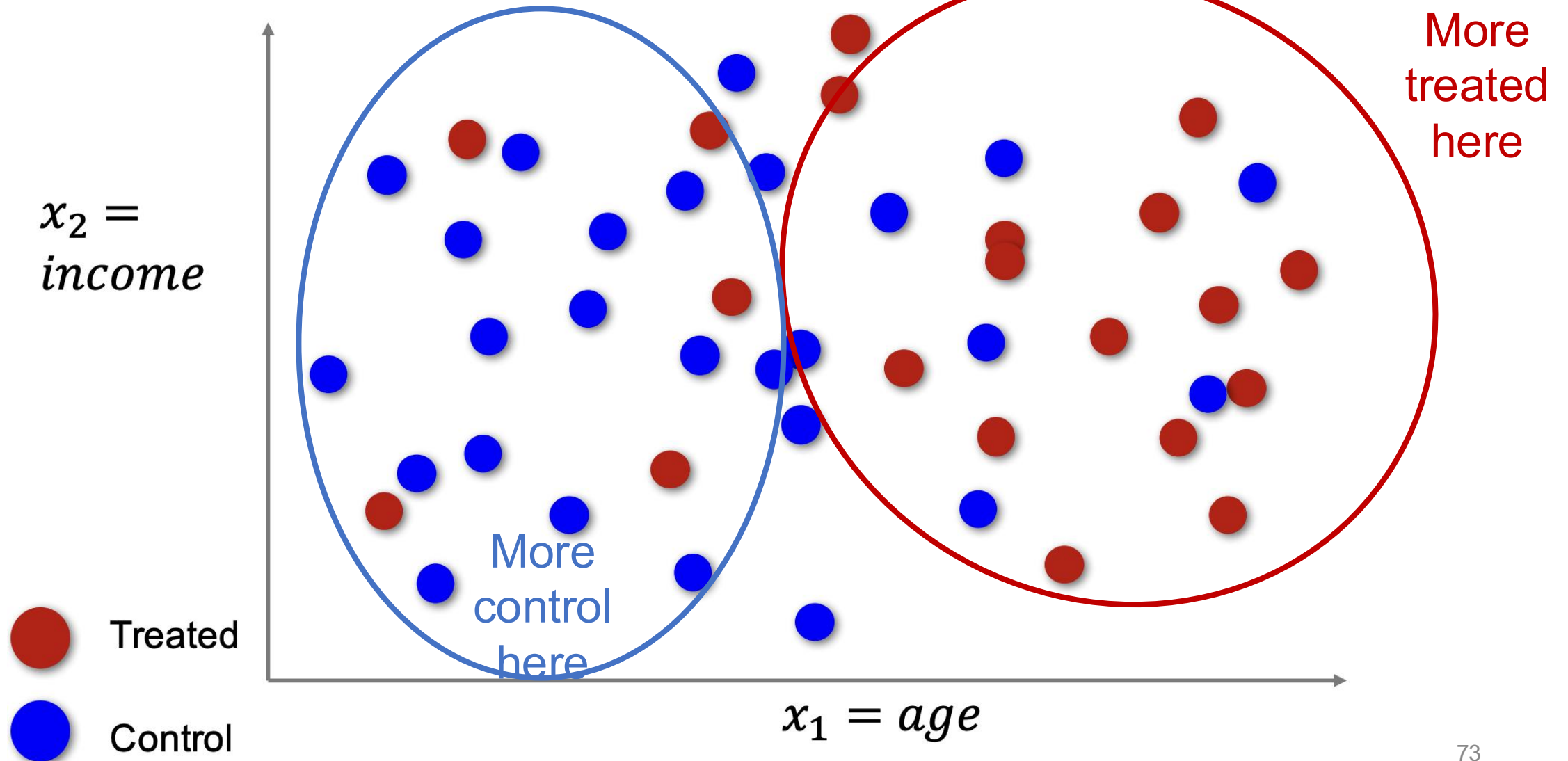
Propensity score

- Tool for estimating ATE
- Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples, similar to importance sampling
- We want to model $p(T|x)$

Inverse propensity score re-weighting



Inverse propensity score re-weighting



Propensity score

- Propensity score: model $p(T = 1, x)$, using machine learning tools
- Re-weight samples according to the inverse of the probability score: “more rare samples for this treatment/control get upweighted”

Propensity scores – algorithm

How to calculate ATE with propensity score for sample $(x_1, y_1, t_1), \dots, (x_N, y_N, t_N)$

1. Use any ML method to estimate $p(T = 1|x)$
2. Calculate:

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

How to calculate ATE with propensity score for sample $(x_1, y_1, t_1), \dots, (x_N, y_N, t_N)$

1. In a randomized control trial, $p(T = 1|x) = 0.5$
2. Calculate:

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$

Propensity scores – algorithm

How to calculate ATE with propensity score for sample $(x_1, y_1, t_1), \dots, (x_N, y_N, t_N)$

1. In a randomized control trial, $p(T = 1|x) = 0.5$
2. Calculate:

$$\begin{aligned} \hat{ATE} &= \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} = \\ &= \frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i \end{aligned}$$

Propensity score in machine learning

- Same idea is in importance sampling!
- Used in off-policy evaluation and learning from logged bandit feedback (Swaminathan & Joachims, 2015)
- Similar ideas used in covariate shift work (Bickel et al., 2009)

Summary

- ✓ What is causal inference?
- ✓ Potential outcomes framework
- ✓ Tools of the trade



How can we make Data
146 better for you?

Next class: LLM evaluation with Prof Monica Agrawal