CAUSAL INFERENCE FOR OBSERVATIONAL STUDIES

Uri Shalit & David Sontag

ICML 2016, New York

June 2016

Tutorial slides and reference list

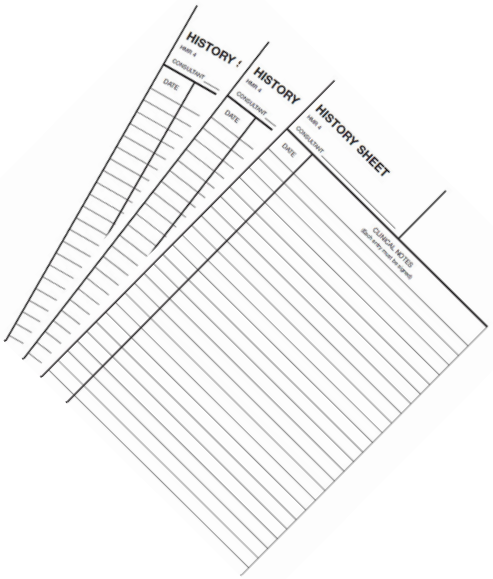
- Tutorial homepage:
  cs.nyu.edu/~shalit/tutorial.html

- Slides:
  cs.nyu.edu/~shalit/slides.pdf

- Reference list:
  cs.nyu.edu/~shalit/slides.pdf

When supervised learning isn't enough

- Dataset of 10,000,000 patients
- Medications, blood tests, past diagnoses, doctors' notes, demographics, genetic testing



When supervised learning isn't enough

- Patient "Anna" comes in with hypertenstion
  - Asian, 54, history of diabetes, blood pressure 150/95, …
- Which medication will better lower her blood pressure:
  - Calcium channel blocker (A)
  - ACE inhibitor (B)
- I have data from 10,000,000 other patients – surely that can help!

When supervised learning isn't enough

- Patient "Anna" comes in with hypertenstion
  - Asian, 54, history of diabetes, blood pressure 150/95, …
- Which medication will better lower her blood pressure:
  - Calcium channel blocker (A)
  - ACE inhibitor (B)
- Approach 1:
  Find patients like Anna, one who received A and one who received B, compare outcomes

When supervised learning isn't enough

- Patient "Anna" comes in with hypertenstion
  - Asian, 54, history of diabetes, blood pressure 150/95, …
- Which medication will better lower her blood pressure:
  - Calcium channel blocker (A)
  - ACE inhibitor (B)
- Approach 2:
  Build a regression model from patient features to blood pressure

When supervised learning isn't enough

- Build regression model from patient features to blood pressure (BP)
- Input:

Output:

Anna's features

A

predicted BP

–

Anna's features

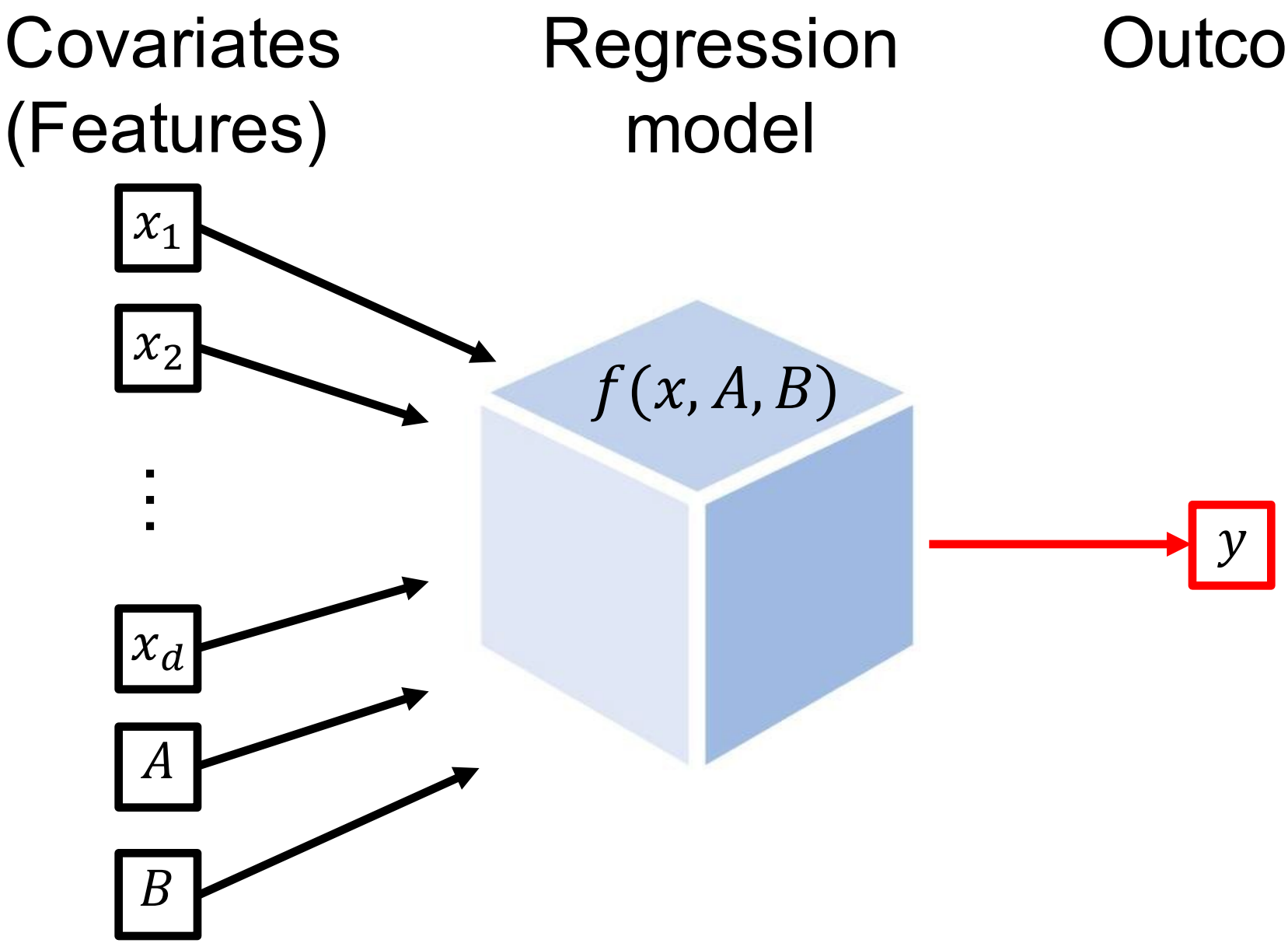
B

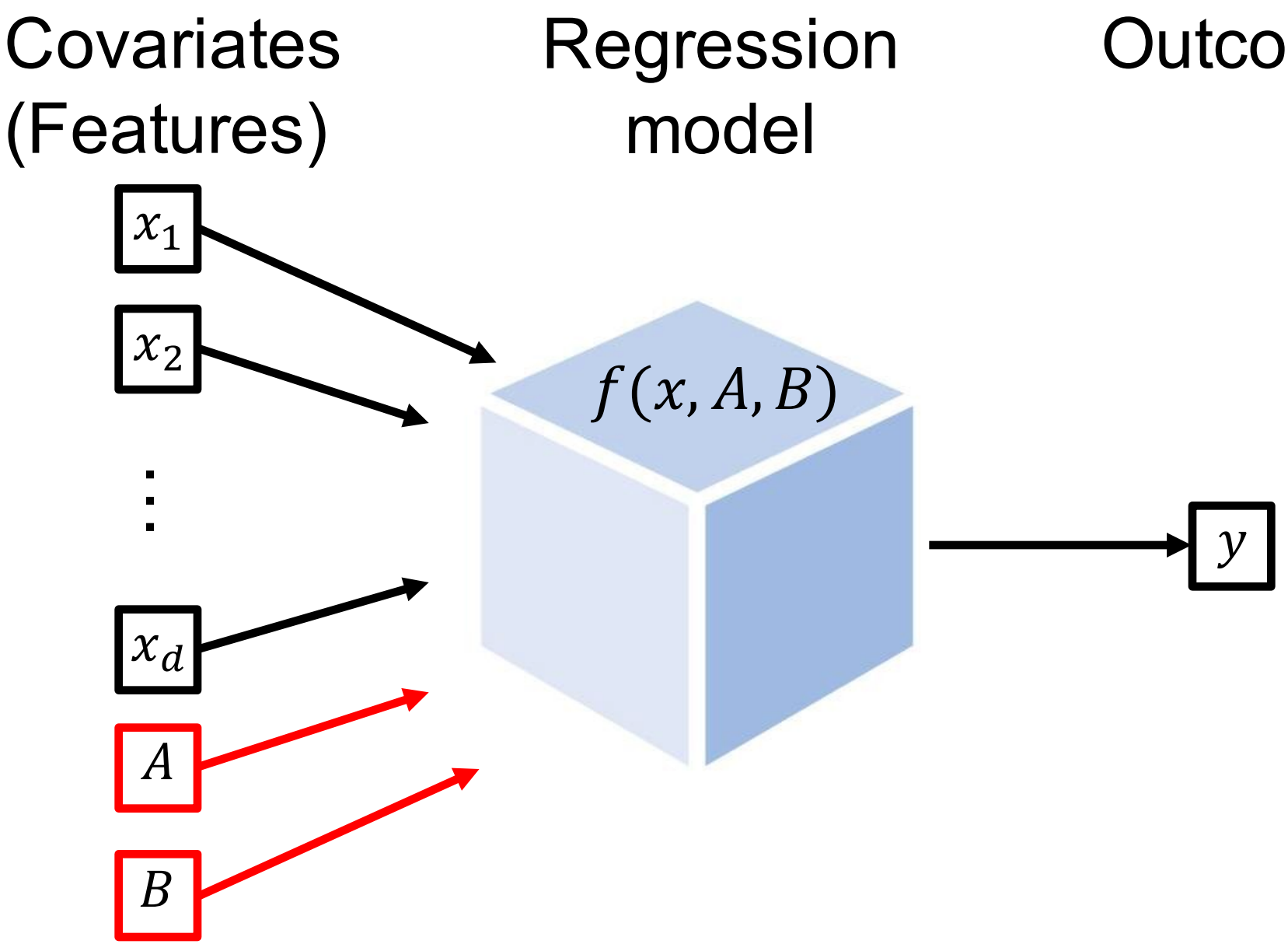
predicted BP

=

?

- Compare


Covariates (Features)
Regression model
Outcome
x_1
x_2
x_d
A
B
$f(x, A, B)$
y


Covariates (Features)
Regression model
Outcome
x_1
x_2
x_d
A
B
$f(x, A, B)$
y

When supervised learning isn't enough

- 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?

Causal inference

- We are interested in the causal influence of medications A and B on blood pressure
- This is not a textbook machine learning prediction problem

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

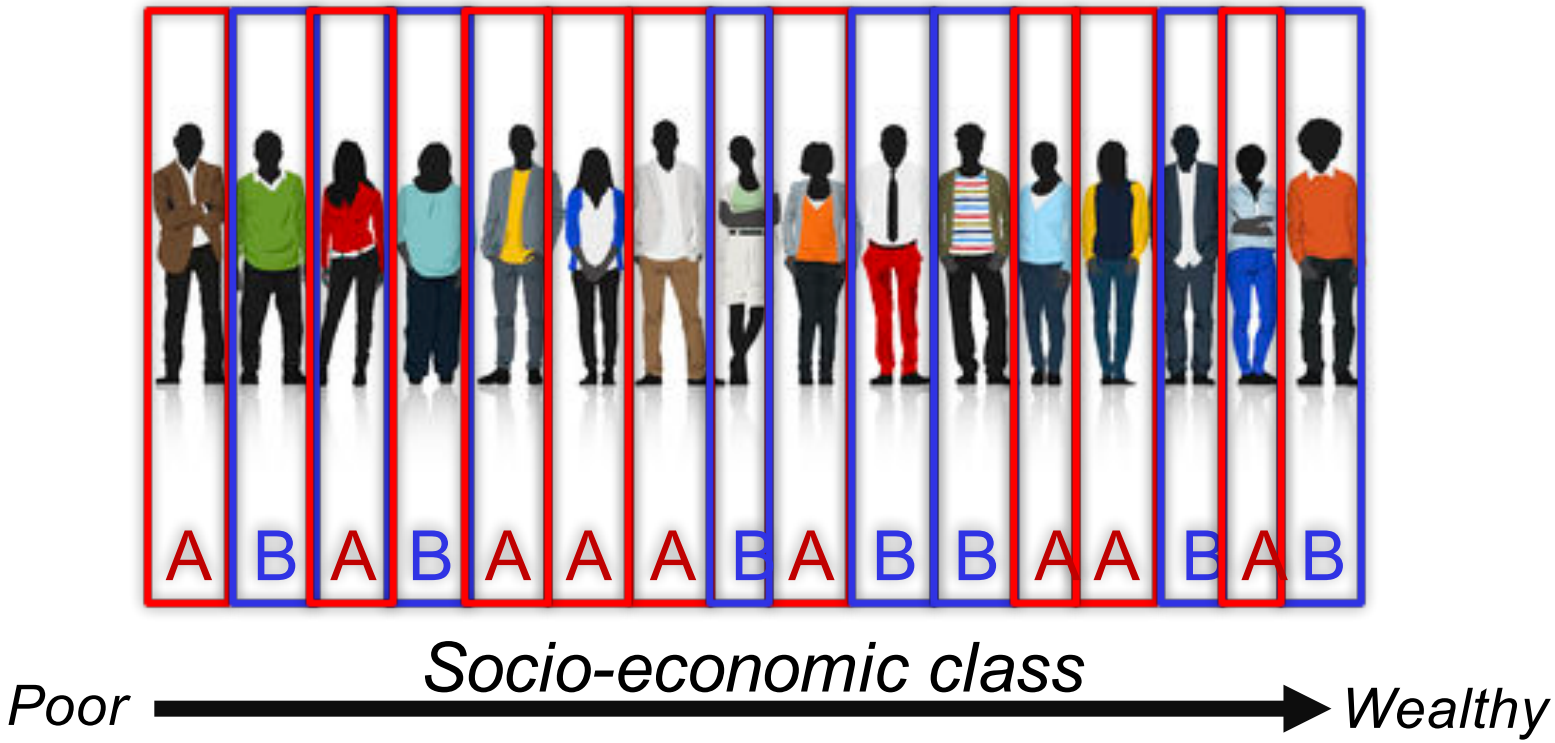
- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

Randomized trials vs. observational studies

treatment

A **or** B

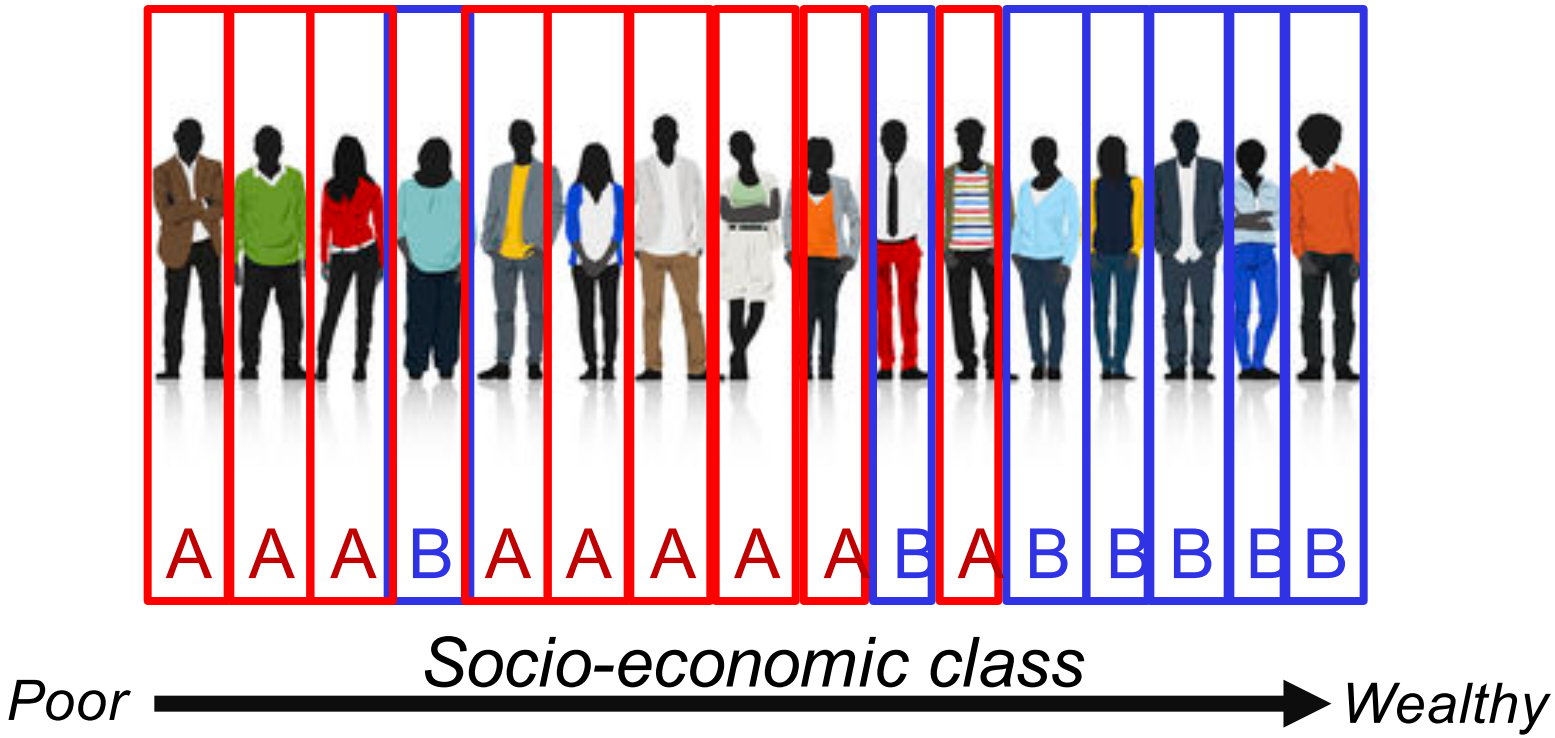
Randomized controlled trial (RCT)



treatment

A **or** B

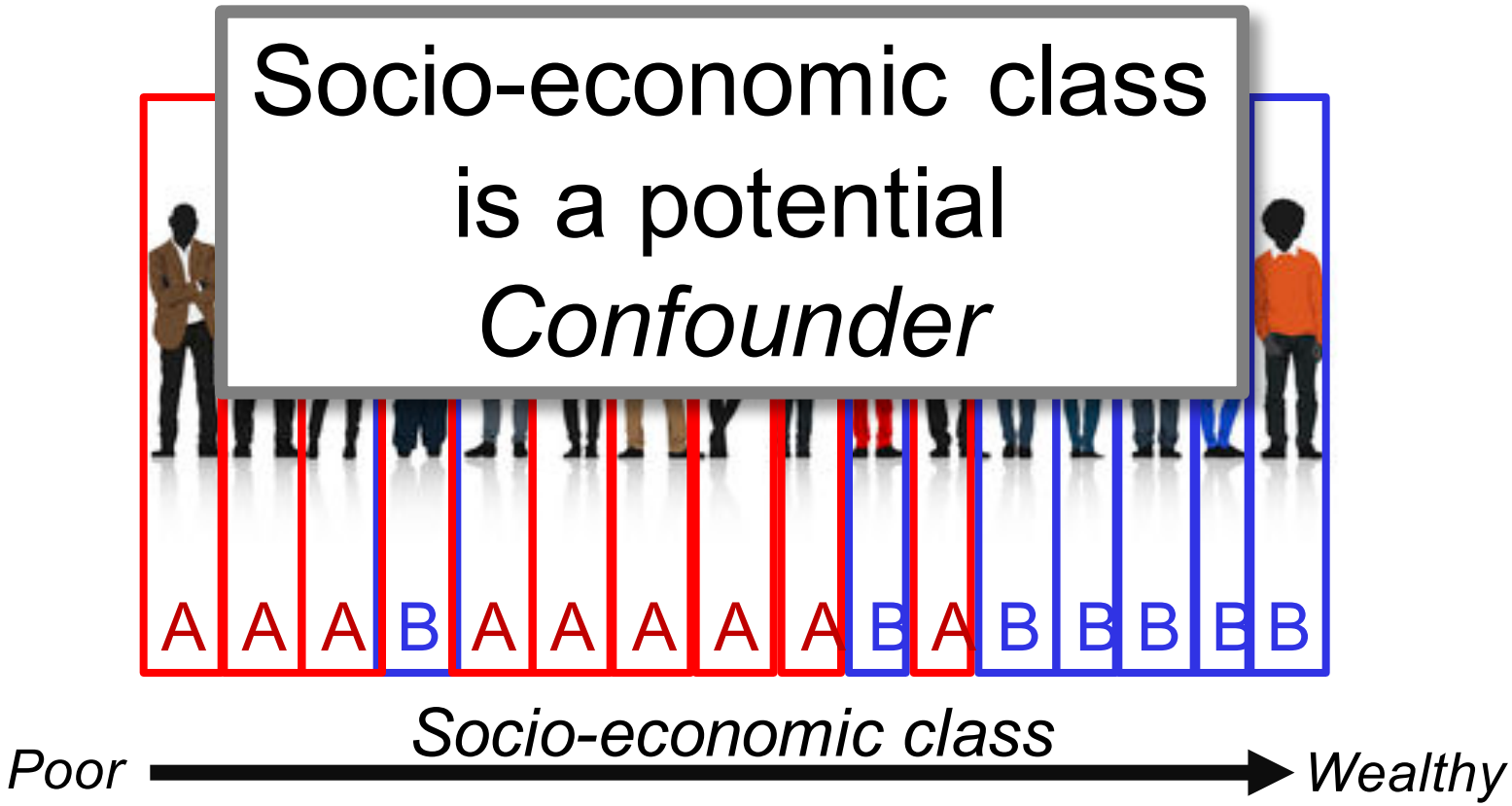
Observational study



treatment

A **or** B

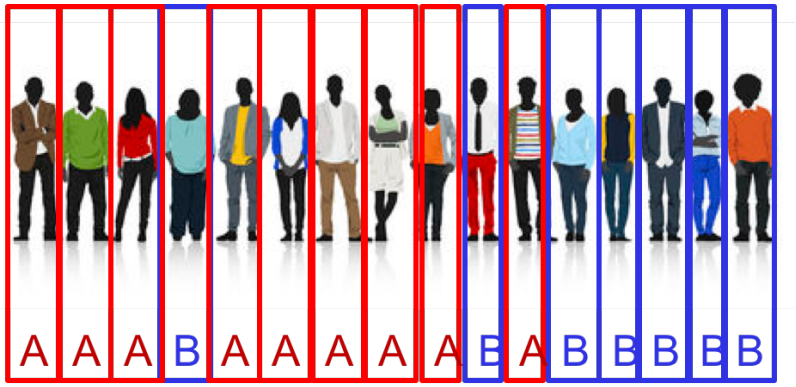
Observational study

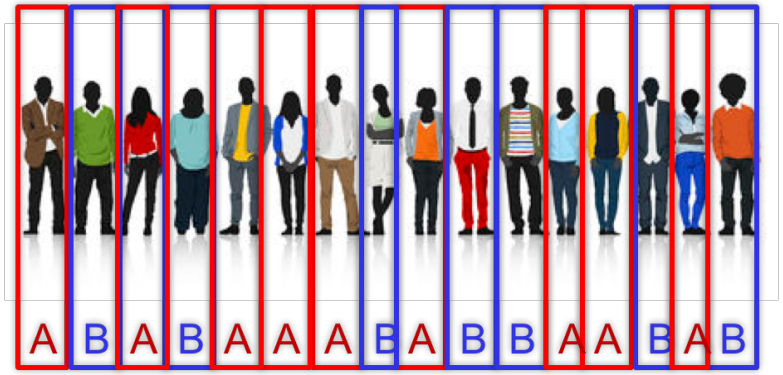


treatment

A **or** B

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



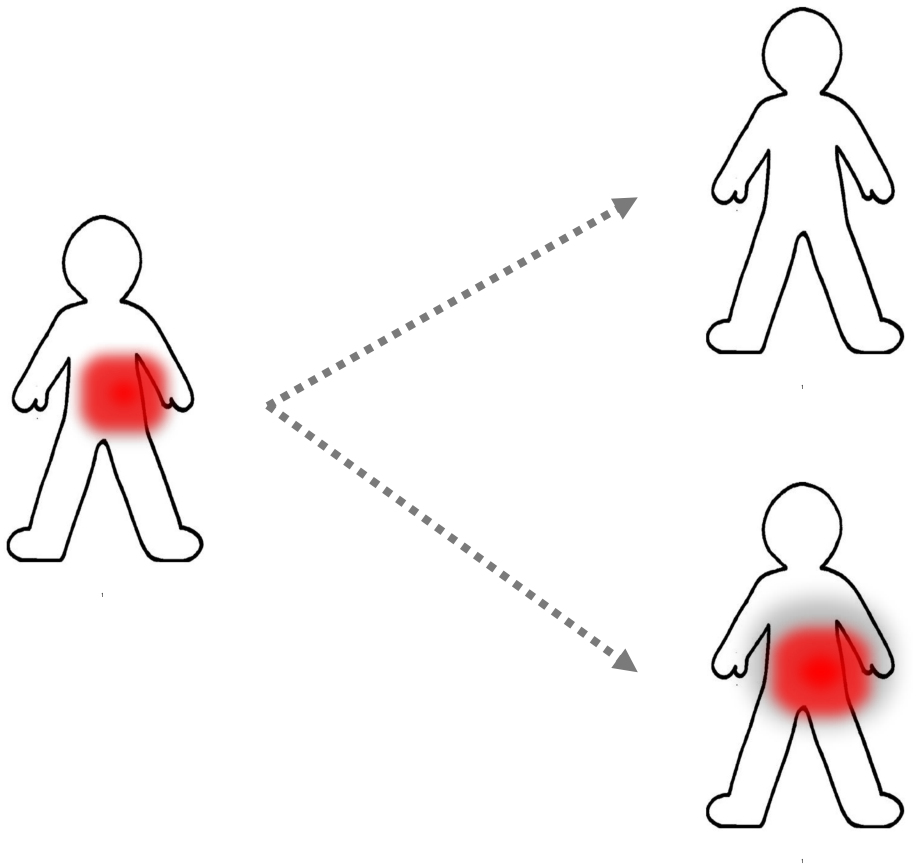


- Does inhaling Asbestos cause cancer?
- Does decreasing the interest rate reinvigorate the economy?
- We have a budget for **one new** anti-diabetic drug experiment. Can we use past health records of 100,000 diabetics to guide us?

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

Example 1
Precision medicine:
Individualized 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*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)

- Future blood pressure: treatment A vs. B
- **Individualized Treatment Effect (ITE)**

Which treatment is best **for me**?

- Which anti-hypertensive treatment?
  - Calcium channel blocker (A)
  - ACE inhibitor (B)

- Potential *confounder*: maybe rich patients got medication A more often, and poor patients got medication B more often

Example 2
Job training:
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?

Example 3
Ad-placement

Ad-placement

- Company X wants to decide which of two actively used advertisements is performing the best
- Using past data, can we know which advertisement has a better click-through rate?
- Potential *confounder*: the choices of the ad-placement algorithm

Average Treatment Effect (ATE)

Observational studies

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

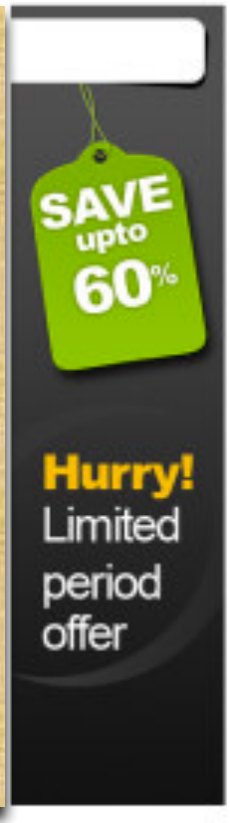
More examples

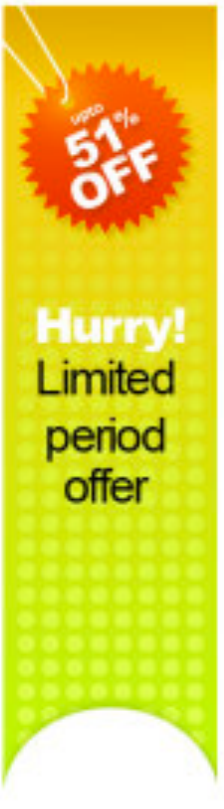

Is smoking
dangerous?

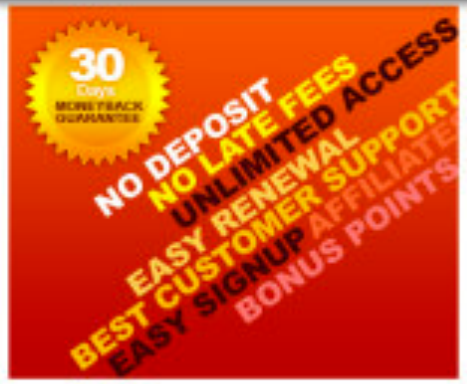
Do stricter gun laws lead to safer communities?

Is pre-kindergarten beneficial for children?

Will running an ad-campaign increase sales?







Did a company discriminate against job applicants?

Outline

Introduction

- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

Outline

Introduction

- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

Causal inference – the challenge

- Train-test paradigm doesn't apply
  (Leon Bottou ICML 2015 keynote)

Causal inference – the challenge

- Automating scientific discovery is inherently hard
- Impossible without several important Assumptions

Outline

Introduction

- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

Outline

Introduction

- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

“The Assumptions”

- Necessary conditions for causal inference to be possible

1) *No unmeasured confounders*

2) *Common support*

No unmeasured confounders

Common support

Did not receive job training

Received job training

"The Assumptions"

- Include more than these two, will discuss in *mathematical foundations* later
- Go by many different names
  - *ignorability*, *exchangeability, selection on observables*
  - "common support" is also called *overlap, positivity*, related to *balance*

Outline

Introduction

- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

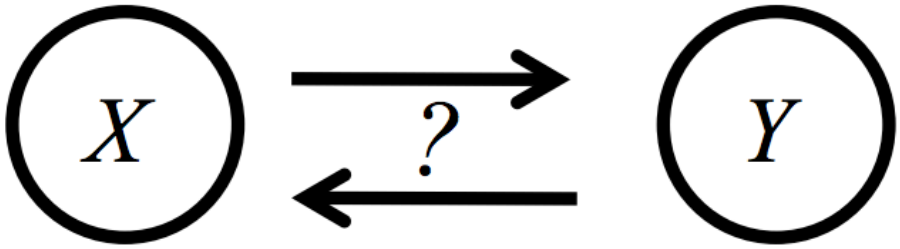
Outline

Introduction

- Examples
- Challenges of causal inference
- “The Assumptions”
- Causal inference and the ML community

Machine learning "causals" that we won't discuss:

- Identifying causal direction between two variables:



Bernhard Schölkopf

- Assumptions on noise process
- Work by Schölkopf, Janzing, Guyon, Mooij, Peters, Geiger, Lopez-Paz and others

Machine learning "causals" that we won't discuss:

- Learning causal graph structure from data:
  - Distinguishes between direct and indirect effects
  - Makes different set of assumptions, such as "faithfulness"
- Bühlmann, Geng, Maathuis, Pearl, Meinshausen, Tsamardinos...
- UAI tutorial next week

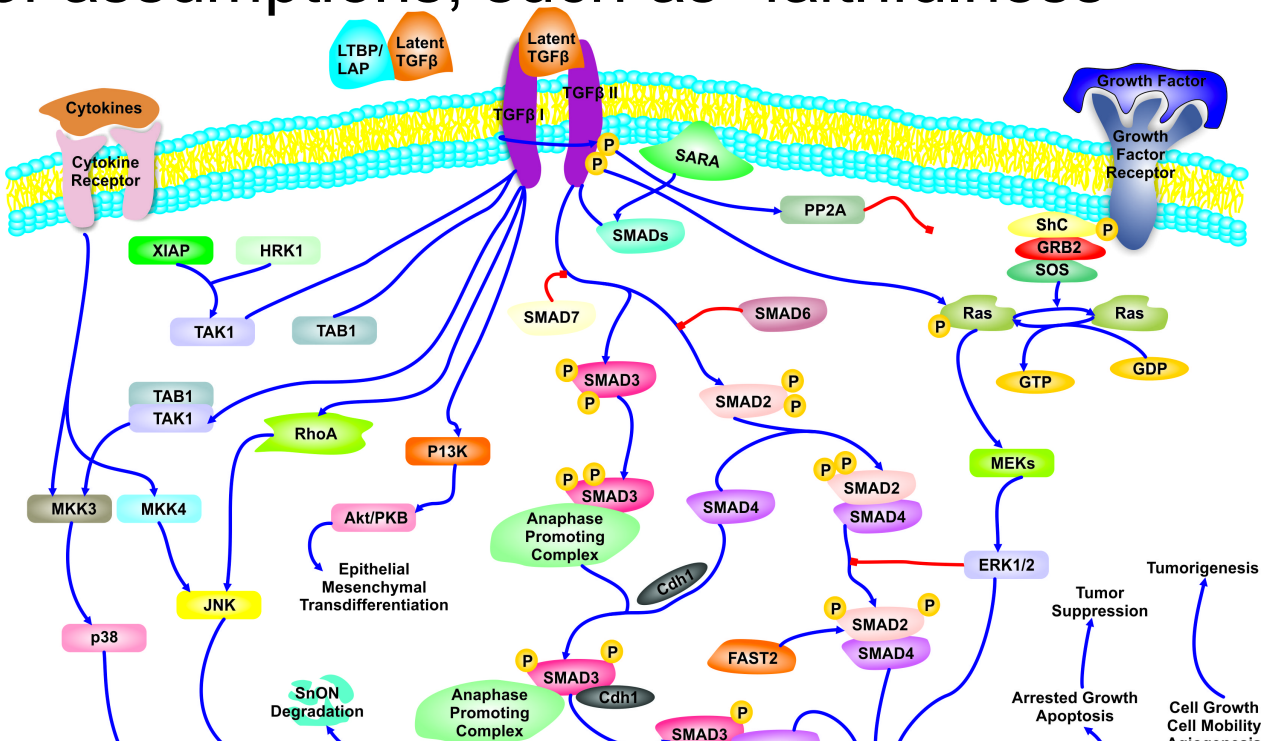


Image from:
http://www.mensxmachina.org/causalpath/state.html

Connections to other ML problems

- Off-policy learning and evaluation
- Learning with bandit feedback and contextual bandits
- Domain adaptation and covariate shift

We will outline these connections throughout the talk

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Counterfactuals and causal inference

- Does treatment T cause outcome Y?
- If T had not occurred, Y would not have occurred (David Hume)

- Counterfactuals:
  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

- Counterfactuals:
  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?
- If her income would have been $18,000, we say that job training caused an increase of $2,000 in Kim's income
- The problem: you never know what might have been

Sliding Doors

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 – patient blood pressure

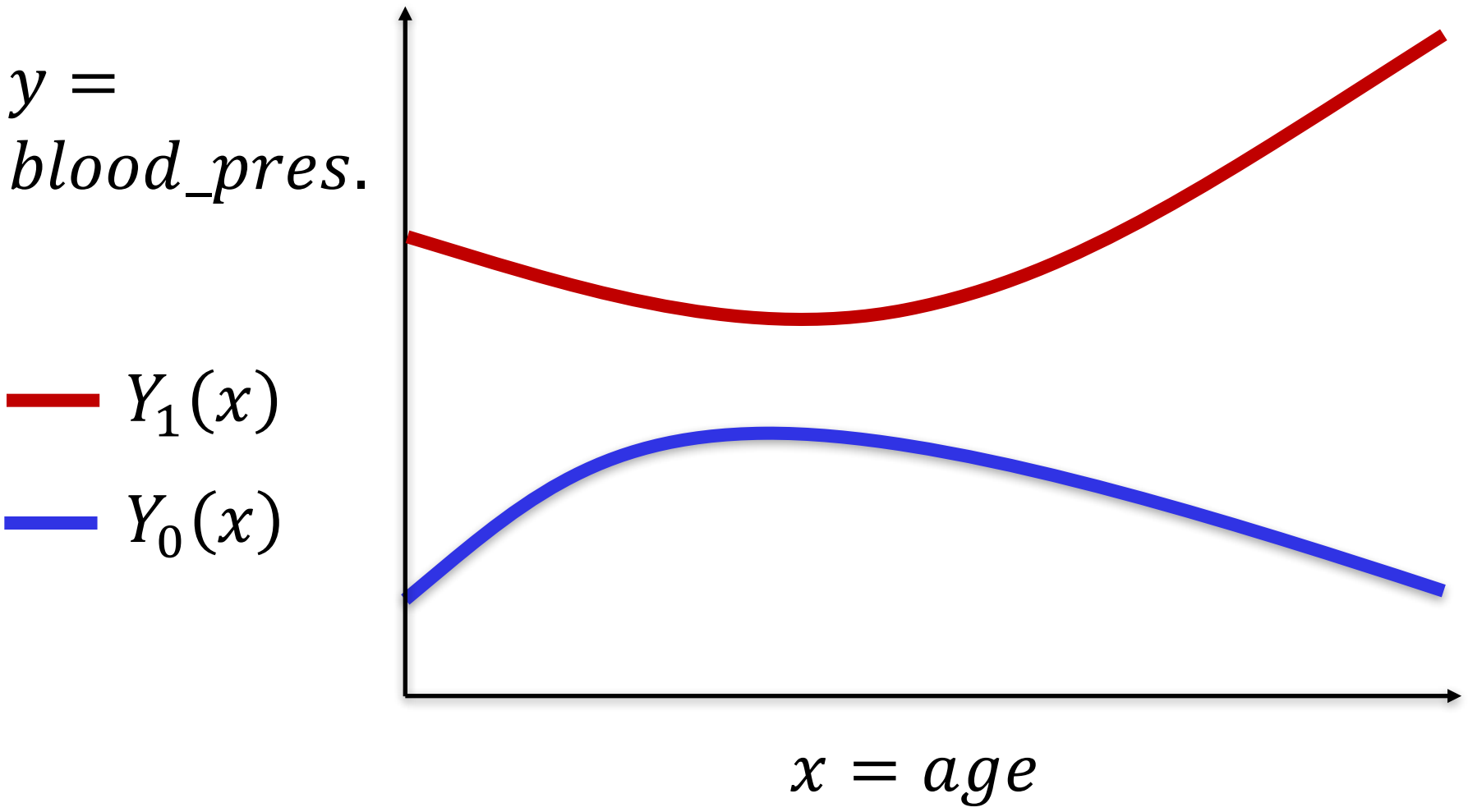
- Features x = (age, gender)
  treatment=medication A, control=medication B

Observed set	
(age,gender, treatment)	Blood pressure after medication
(40, F, **1**)	140
(40, M, **1**)	145
(65, F, **0**)	170
(65, M, **0**)	175
(70, F, **0**)	165

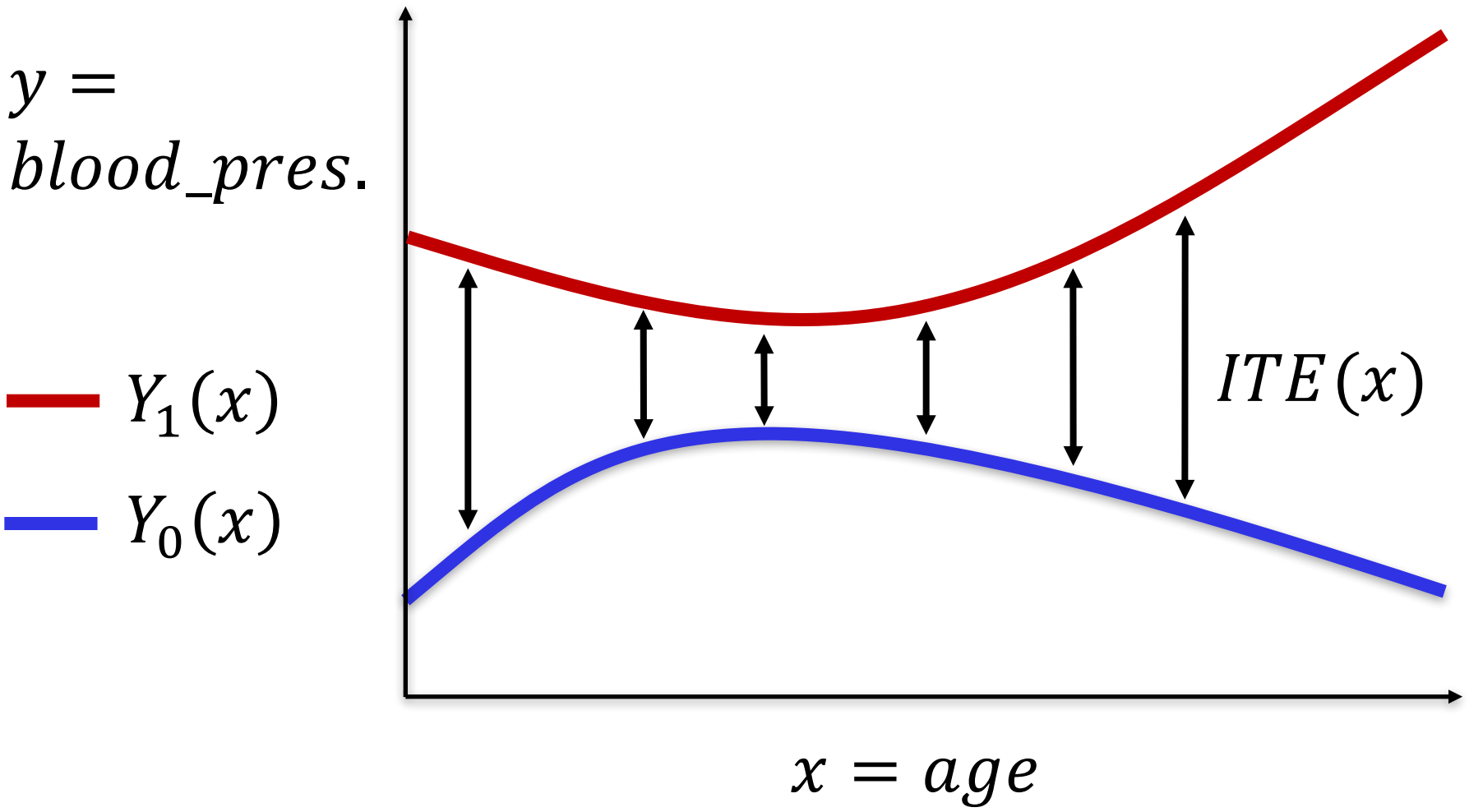
Prediction set	
(age,gender, treatment)	Blood pressure after medication
(40, F, **0**)	?
(40, M, **0**)	?
(65, F, **1**)	?
(65, M, **1**)	?
(70, F, **1**)	?

Blood pressure and age

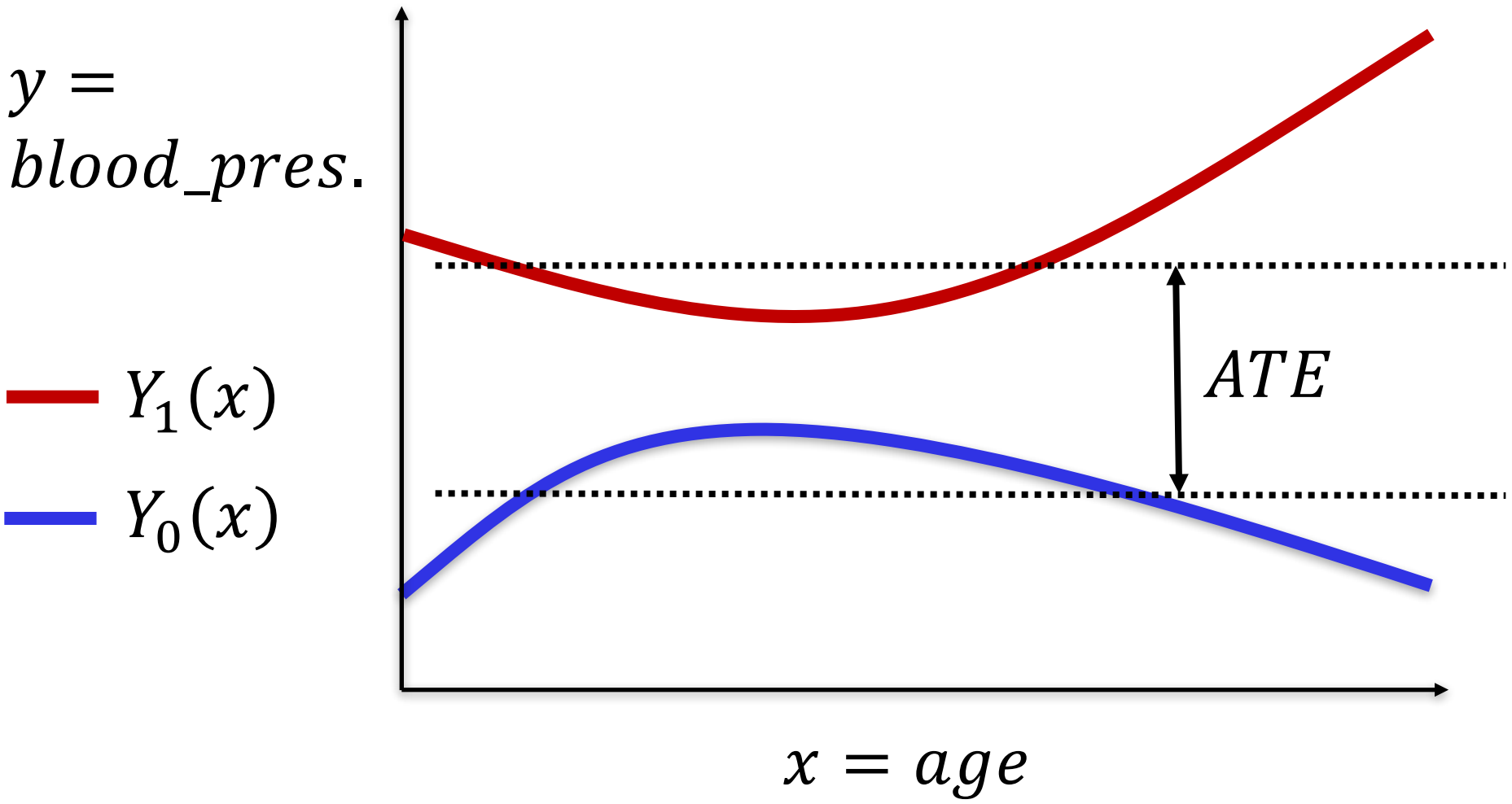
$y = blood\_pres.$



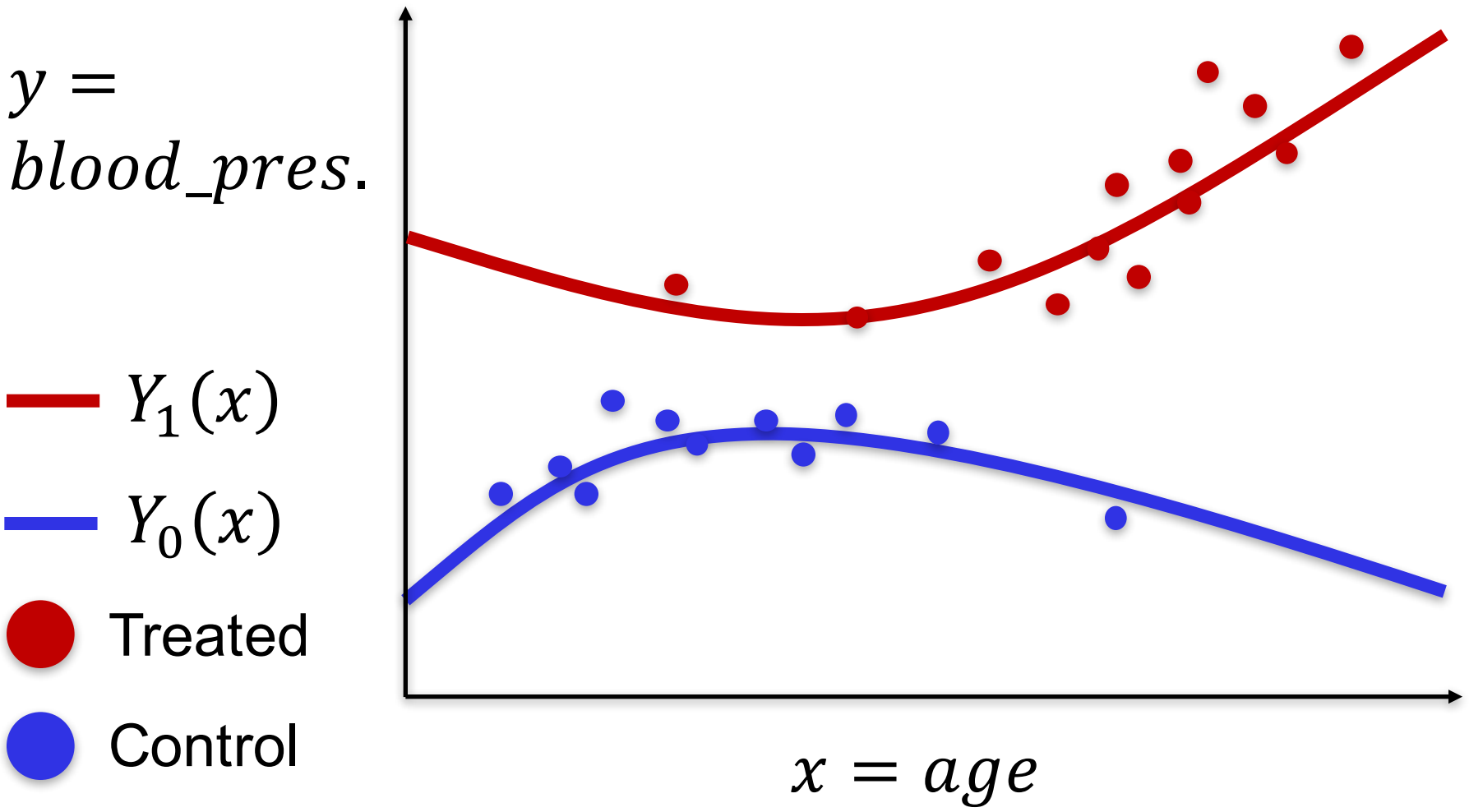
Blood pressure and age



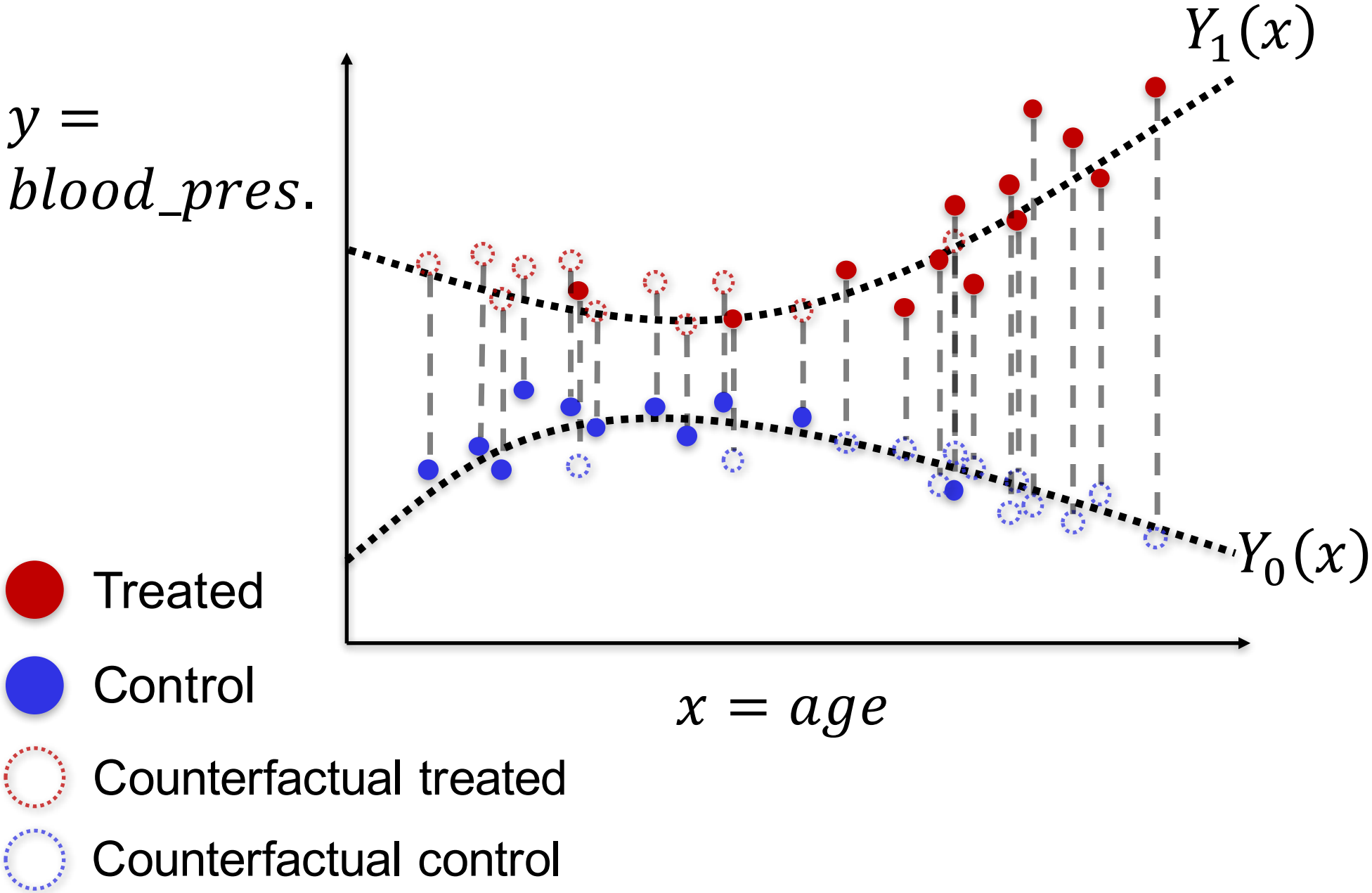
Blood pressure and age



Blood pressure and age



Blood pressure and age



“The fundamental problem of causal inference”

We only ever observe one of the two outcomes

"The Assumptions" – 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

"The Assumptions" – no unmeasured confounders

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

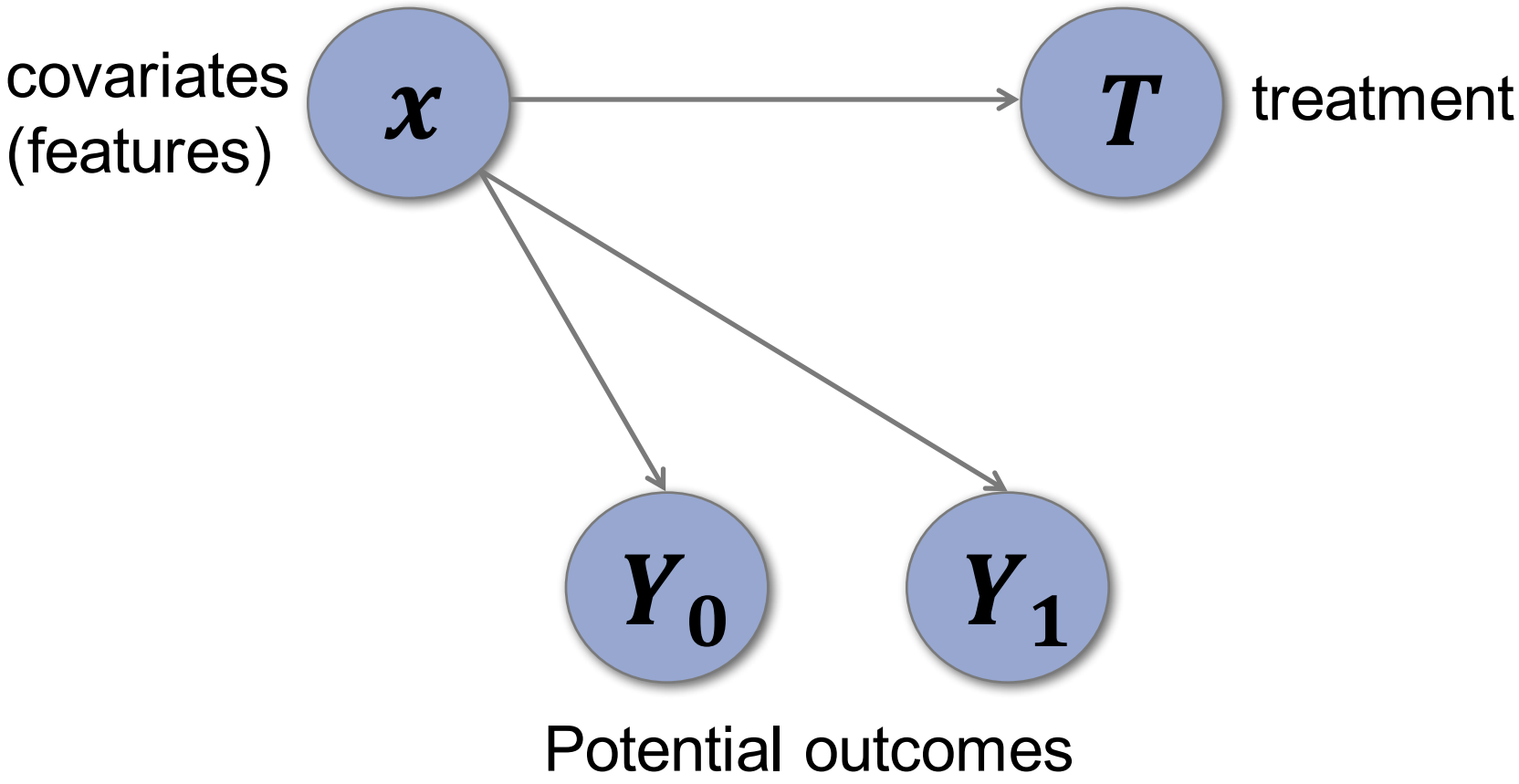
T: treatment assignment

We assume:

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

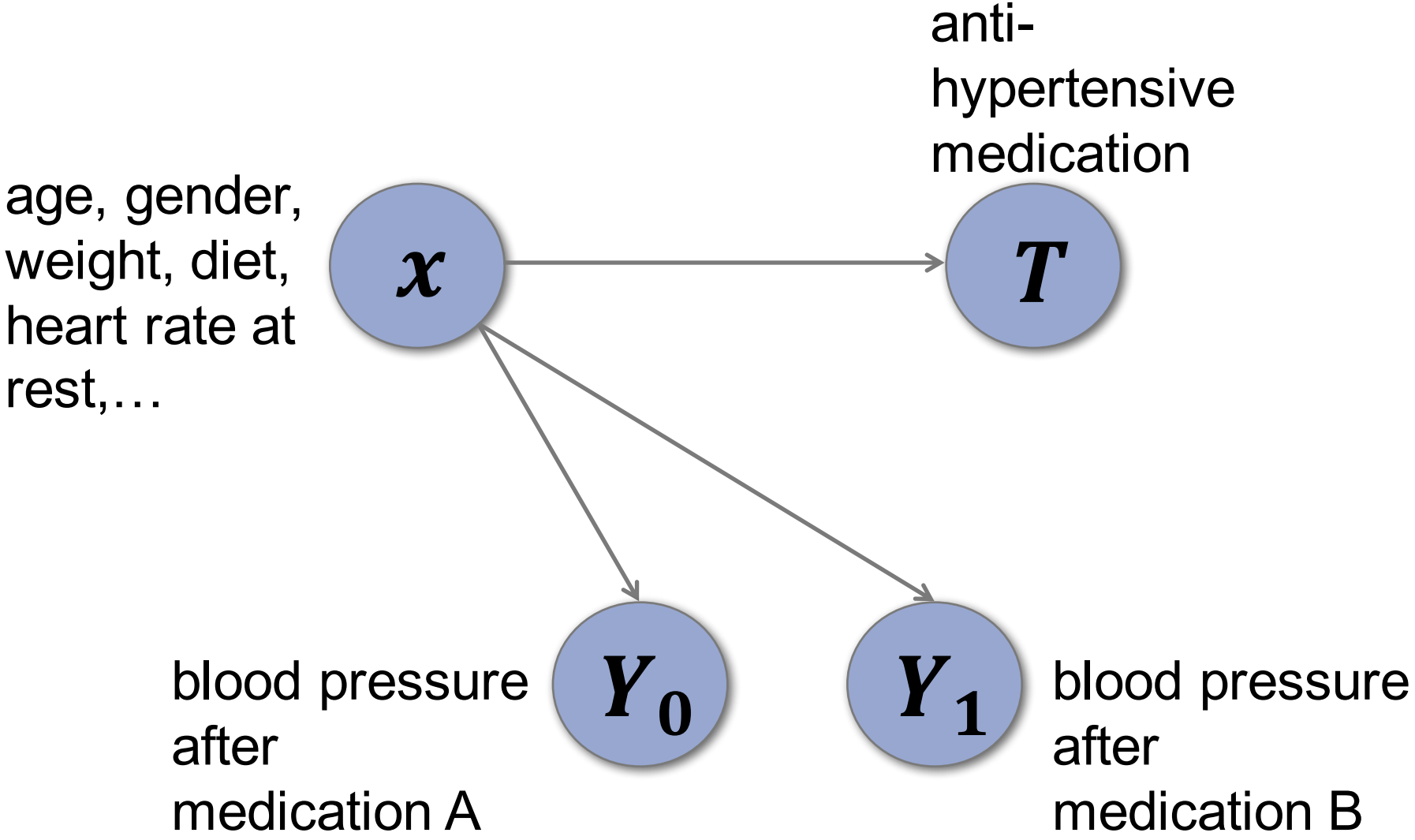
Ignorability

Ignorability



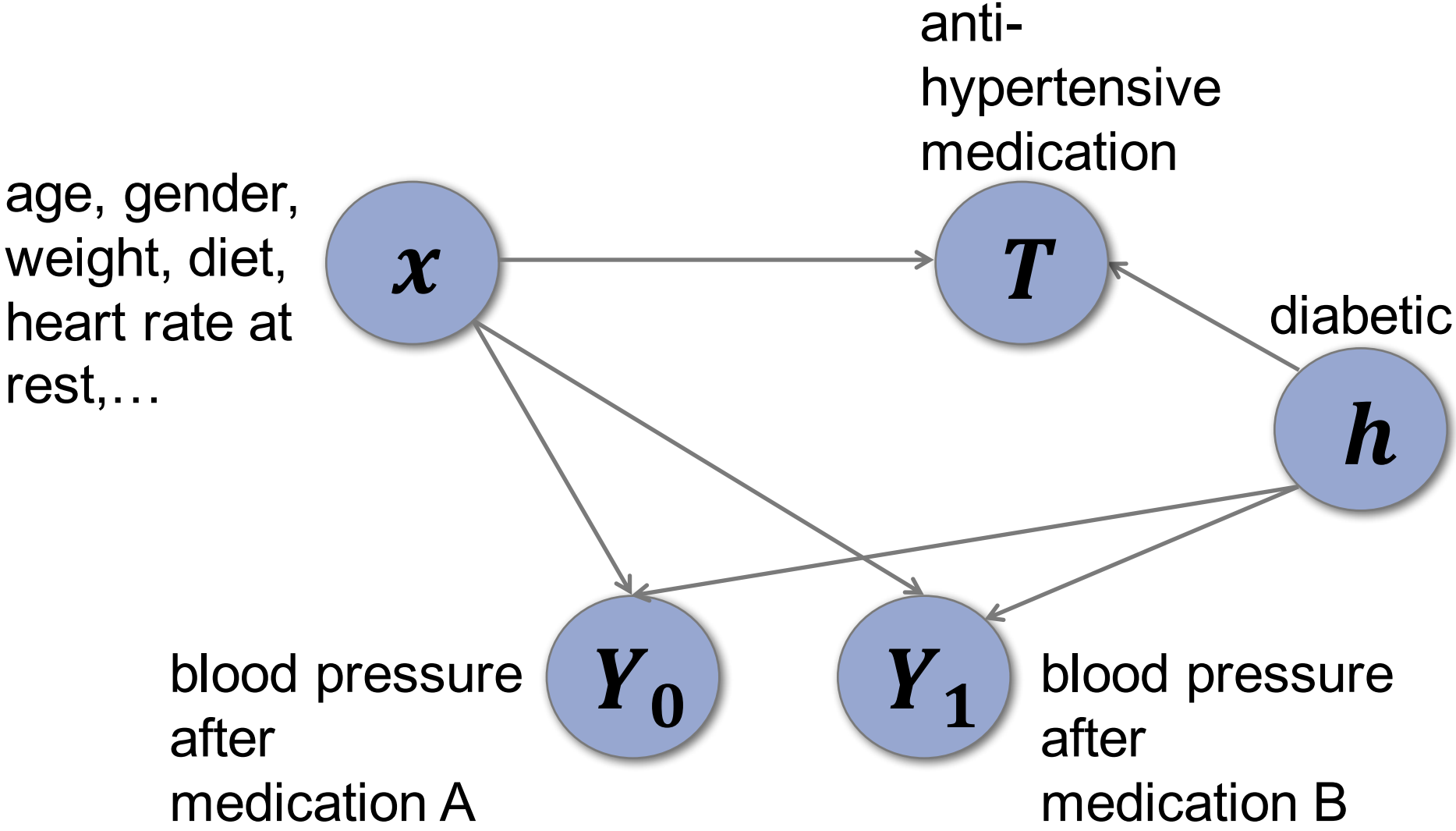
$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

Ignorability



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

No Ignorability



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

“The Assumptions” – common support

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

We assume:

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

Average Treatment Effect

The expected causal effect of T on Y:

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

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

The expected causal effect of T on Y:

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

Average Treatment Effect

The expected causal effect of T on Y:

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

law of total expectation

$$\mathbb{E}\left[Y_1\right] =$$

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

Average Treatment Effect

The expected causal effect of T on Y:

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

$$\mathbb{E}\left[Y_1\right] =$$

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

ignorability

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

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

Average Treatment Effect

The expected causal effect of T on Y:

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

$$\mathbb{E}\left[Y_1\right] =$$

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

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

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

shorter notation

Average Treatment Effect

The expected causal effect of T on Y:

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

$$\mathbb{E}\left[Y_0\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_0 \sim p(Y_0|x)}\left[Y_0|x\right]\right] =$$

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

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

The adjustment formula

Under the assumption of ignorability, we have that:

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

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

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

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

} Quantities we can estimate from data

The adjustment formula

Under the assumption of ignorability, we have that:

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

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

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

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

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

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

} Quantities we *cannot* directly estimate from data

The adjustment formula

Under the assumption of ignorability, we have that:

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

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

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

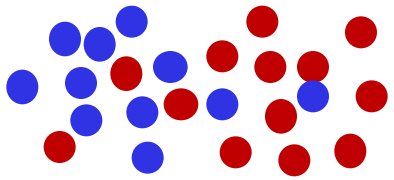
$\mathbb{E}[Y_0|x, T=0]$

} Quantities we can estimate from data

Empirically we have samples from $p(x|T=1)$ or $p(x|T=0)$.
Extrapolate to $p(x)$

Potential outcomes and domain adaptation

Factual

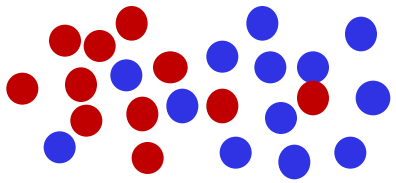


$p_F(x,t) = p_F(x)p_F(t|x)$

the joint *factual* distribution over covariates and treatment assignment

labeled y_i

Counterfactual



$p_{CF}(x,t) \coloneqq p_F(x)p_F(1-t|x)$

the joint *counterfactual* distribution over covariates and treatment assignment

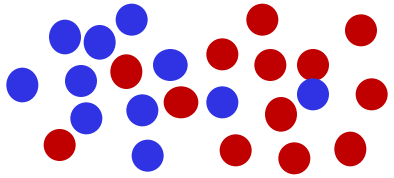
unlabeled

Treated

Control

Potential outcomes and domain adaptation

Source

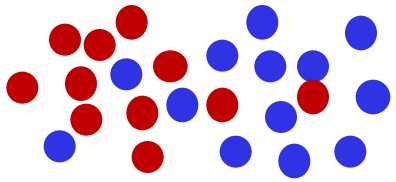


$p_{source}(x)$

the *source*
distribution over covariates

labeled

Target



$p_{target}(x)$

the *target*
distribution over covariates

Treated

Control

unlabeled

Example – patient blood pressure

- Features x = (age, gender)
  treatment=medication A, control=medication B

Observed set	
(age,gender, treatment)	Blood pressure after medication
(40, F, **1**)	140
(40, M, **1**)	145
(65, F, **0**)	170
(65, M, **0**)	175
(70, F, **0**)	165

Prediction set	
(age,gender, treatment)	Blood pressure after medication
(40, F, **0**)	?
(40, M, **0**)	?
(65, F, **1**)	?
(65, M, **1**)	?
(70, F, **1**)	?

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Set up

- Samples: $x_1, x_2, \ldots, x_n$
- Observed binary treatment assignments: $t_1, t_2, \ldots, t_n$
- Observed outcomes: $y_1, y_2, \ldots, y_n$

$$x = (age, gender, married, education, income\_last\_year, \ldots)$$
$$t = (no\_job\_training, job\_training)$$
$$y = income\_one\_year\_after\_training$$

- Does job training raise average future income?

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

Matching

- Find each unit's long-lost counterfactual identical twin, check up on his outcome

Matching

- Find each unit's long-lost counterfactual identical twin, check up on his outcome

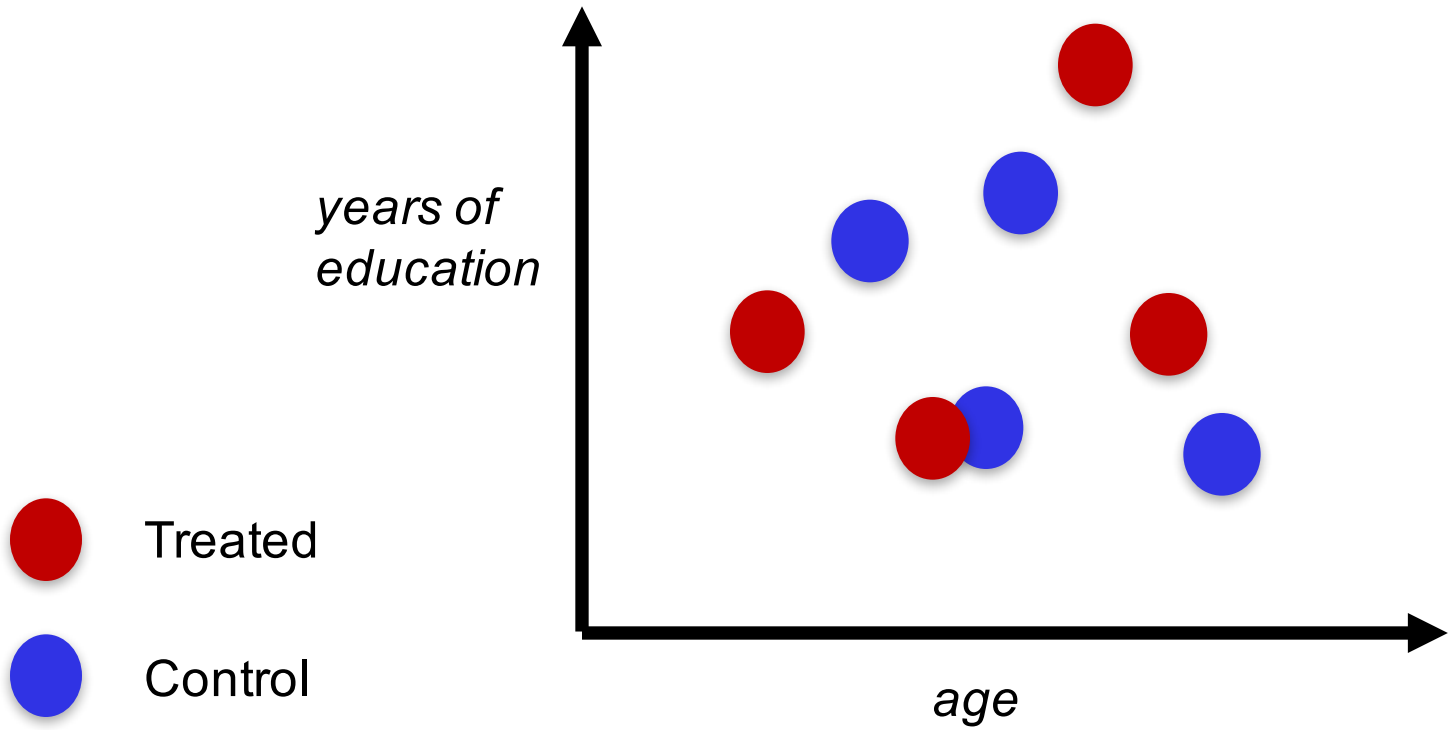
Obama, had he gone to law school

Obama, had he gone to business school

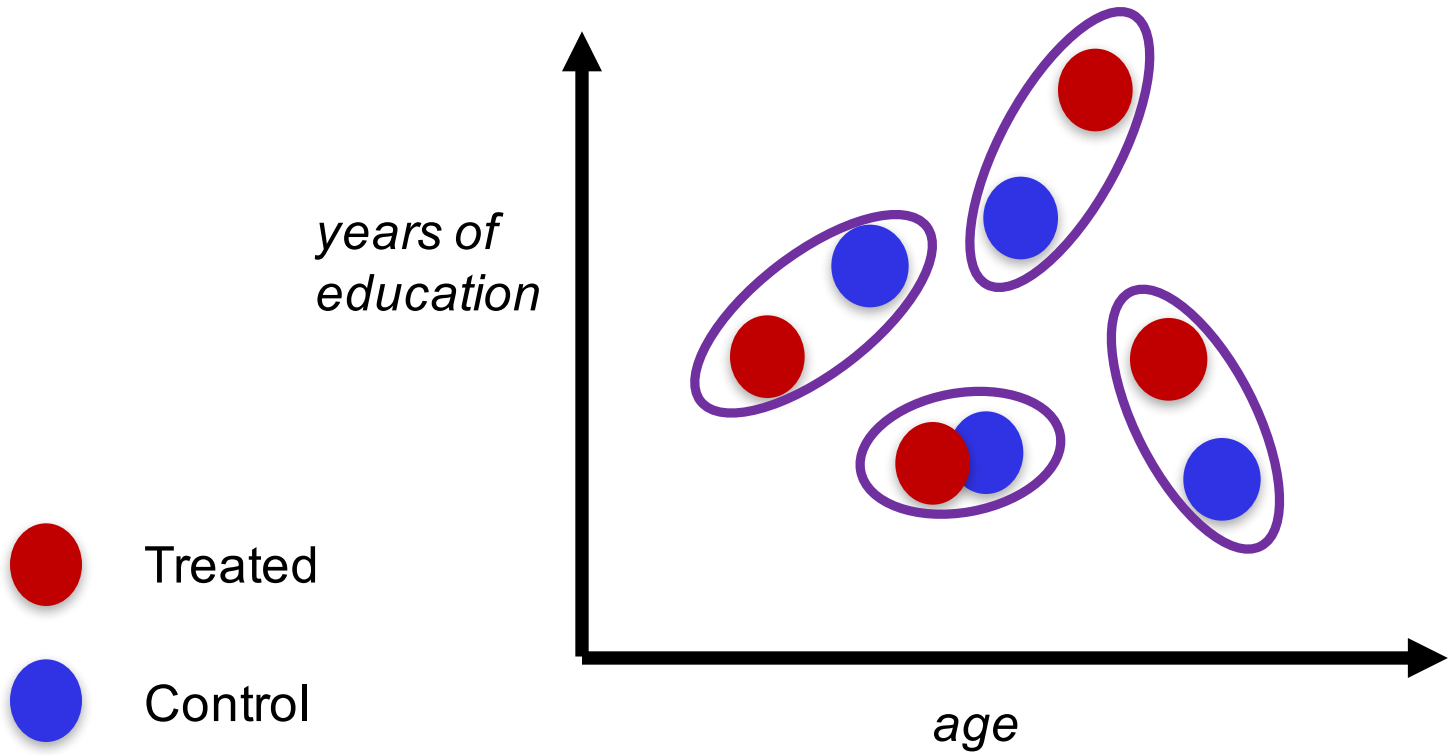
Matching

- Find each unit's long-lost counterfactual identical twin, check up on his 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 $d(\cdot,\cdot)$ be a metric between x's
- For each i, define $j(i) = \underset{j \text{ s.t. } t_j \neq t_i}{\operatorname{argmin}} \; d(x_j, x_i)$

  $j(i)$ is the nearest counterfactual neighbor of i
- $t_i = 1$, unit i is treated:
- $\widehat{ITE}(x_i) = y_i - y_{j(i)}$
- $t_i = 0$, unit i is control:
- $\widehat{ITE}(x_i) = y_{j(i)} - y_i$

1-NN Matching

- Let $d(\cdot,\cdot)$ be a metric between x's
- For each i, define $j(i) = \underset{j \text{ s.t. } t_j \neq t_i}{\operatorname{argmin}} \; d(x_j, x_i)$

  $j(i)$ is the nearest counterfactual neighbor of i

- $\widehat{ITE}(x_i) = (2t_i - 1)(y_i - y_{j(i)})$
- $\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} \widehat{ITE}(x_i)$

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

Matching

- Many other matching methods we won't discuss:
  - 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)

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

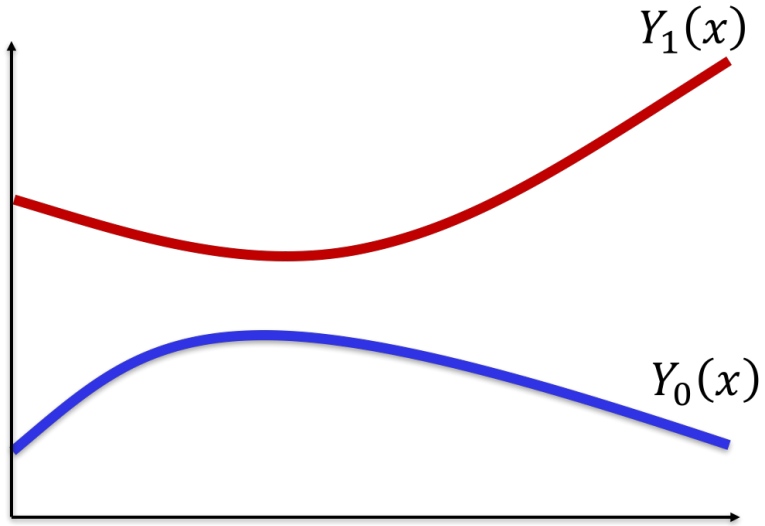
Outline

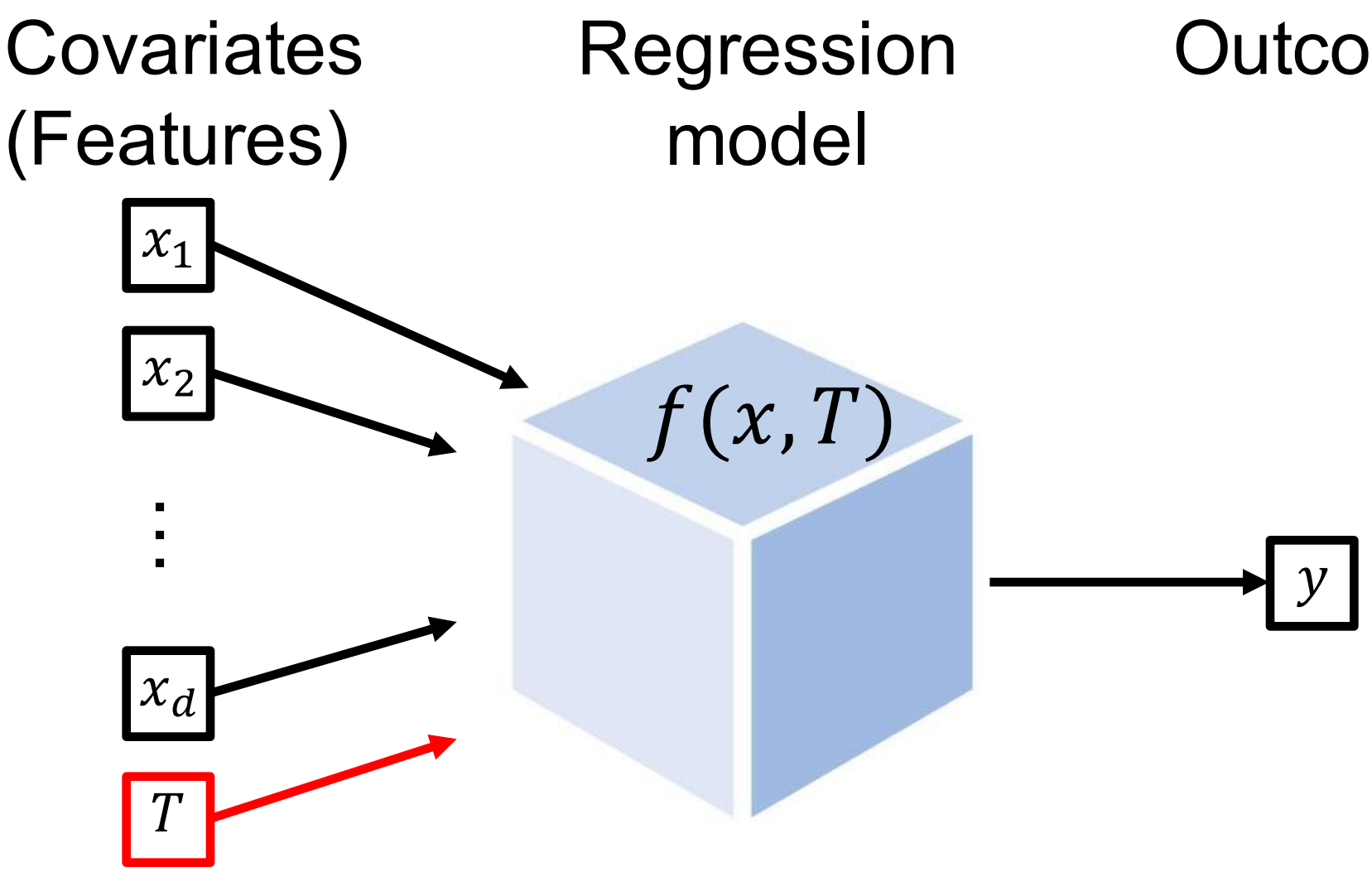
Tools of the trade

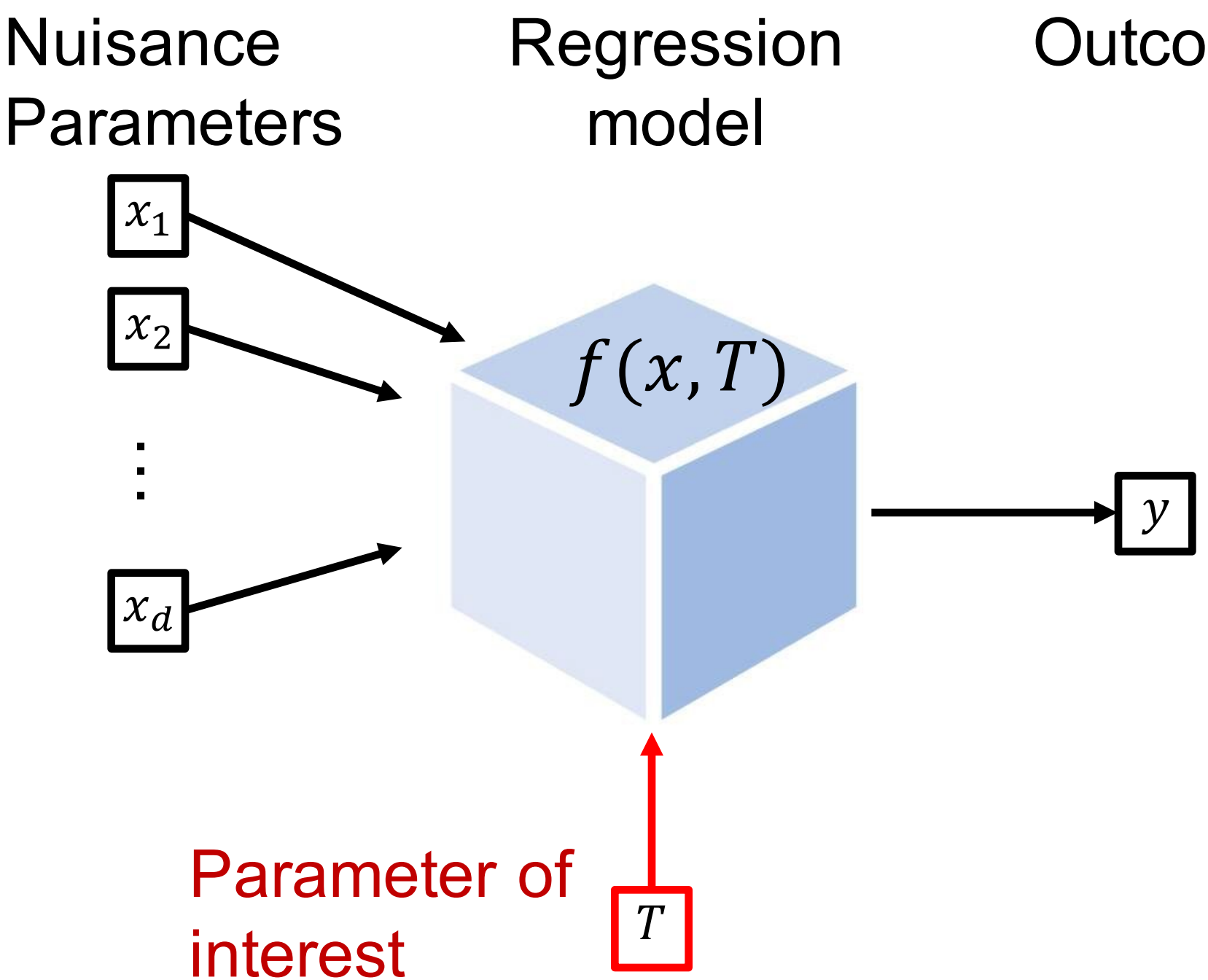
- Matching
- Covariate adjustment
- Propensity score
- Double robustness

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)
Regression
model
Outcome
x_1
x_2
x_d
T
$f(x, T)$
y


Nuisance Parameters
Regression model
Outcome
x_1
x_2
x_d
$f(x, T)$
y
T
Parameter of interest

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

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

Covariate adjustment (parametric g-formula)

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

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

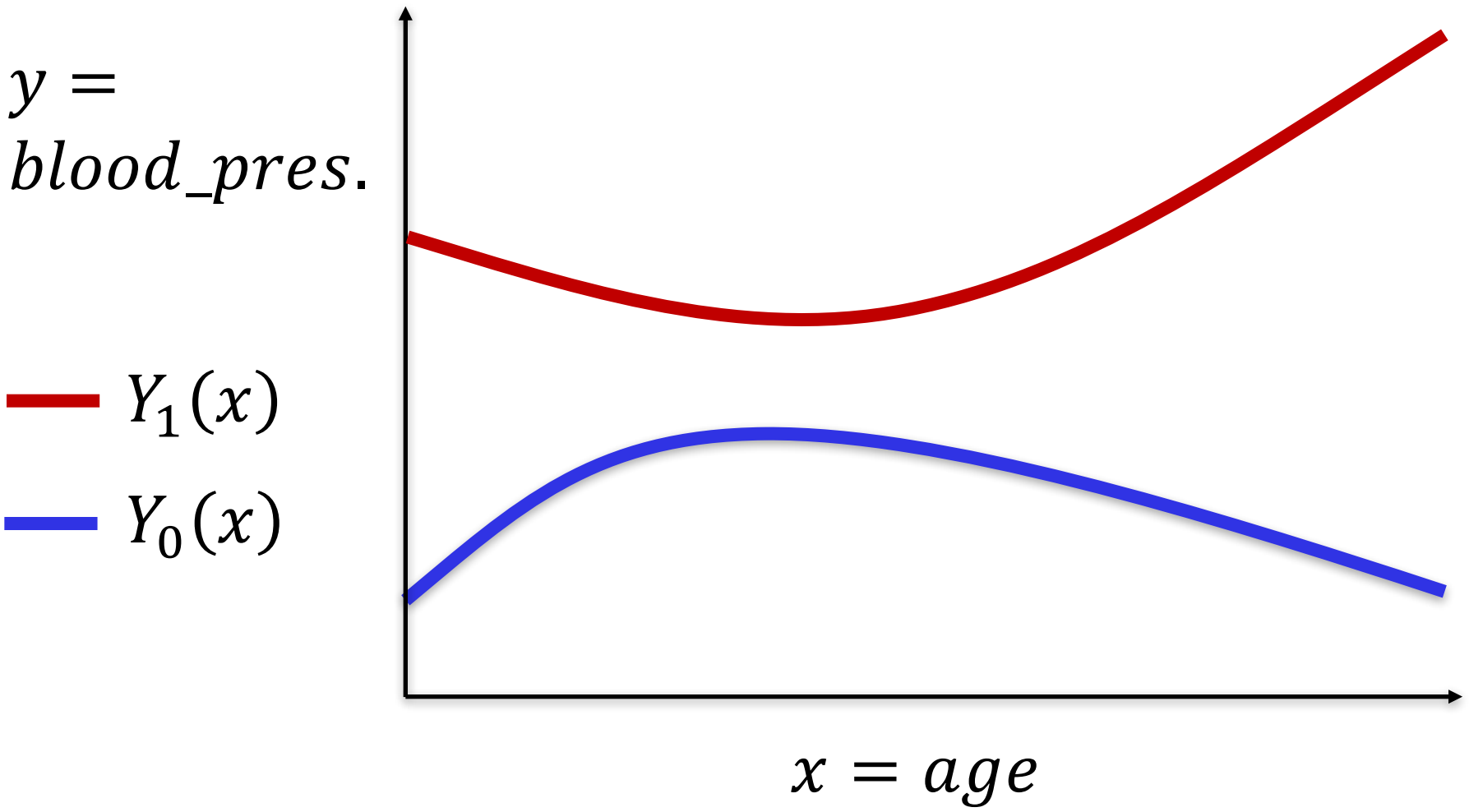
- Fit a model $f(x,t)\approx\mathbb{E}[Y_t|T=t,x]$

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

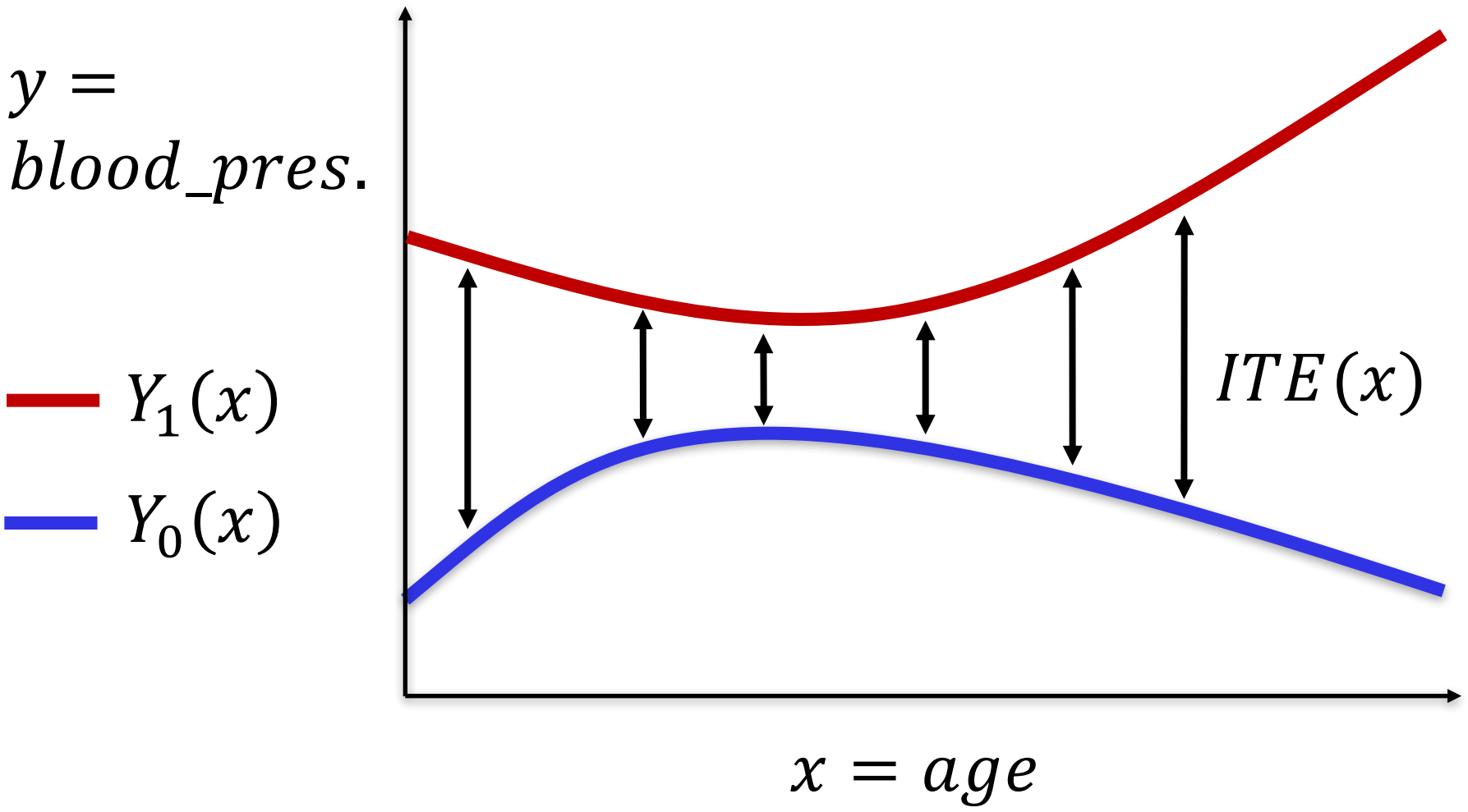
Covariate adjustment - consistency

- 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 $\widehat{ATE}$ will converge to the true ATE
- A sufficient condition: overlap and well-specified model

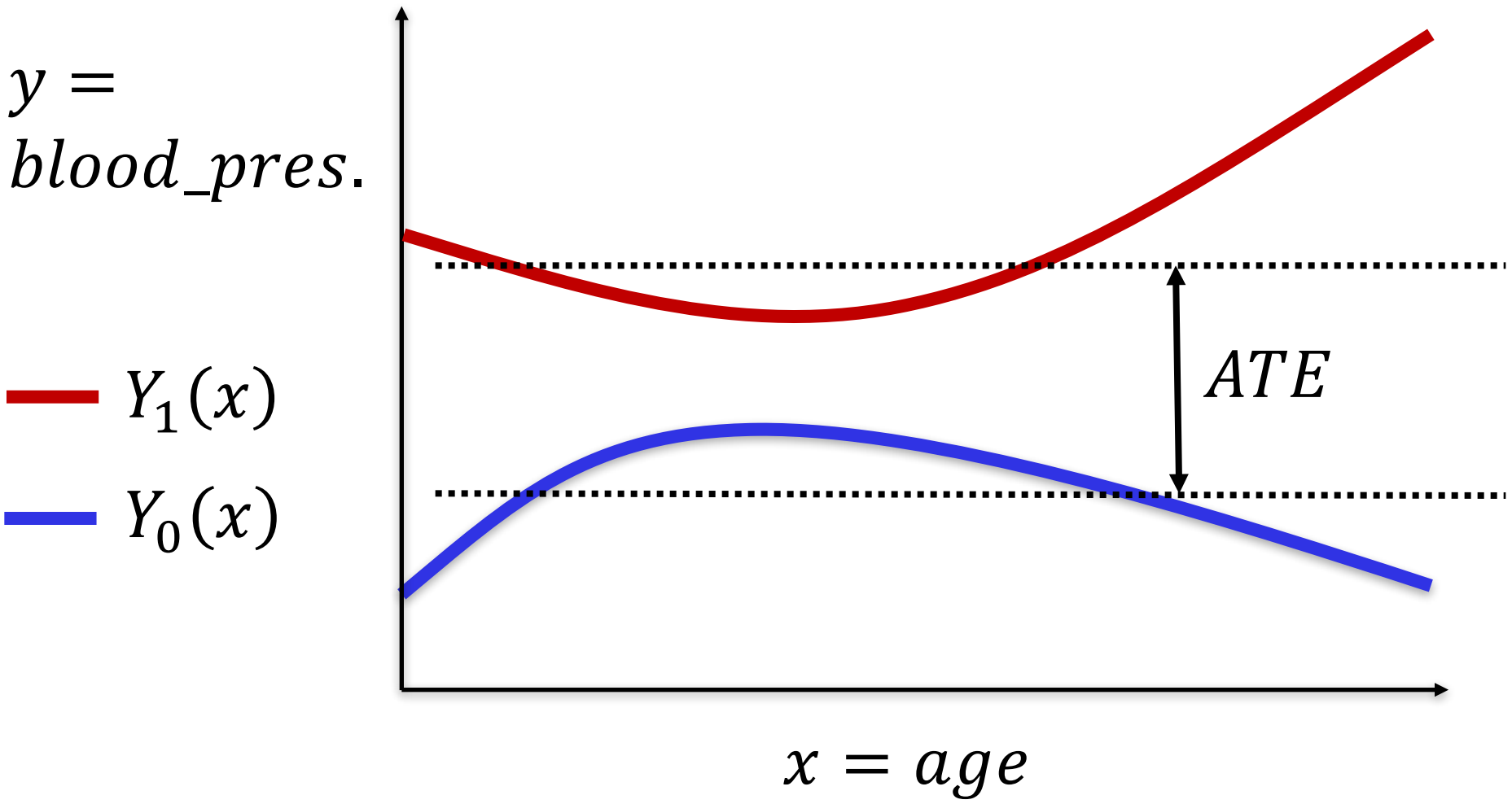
Covariate adjustment



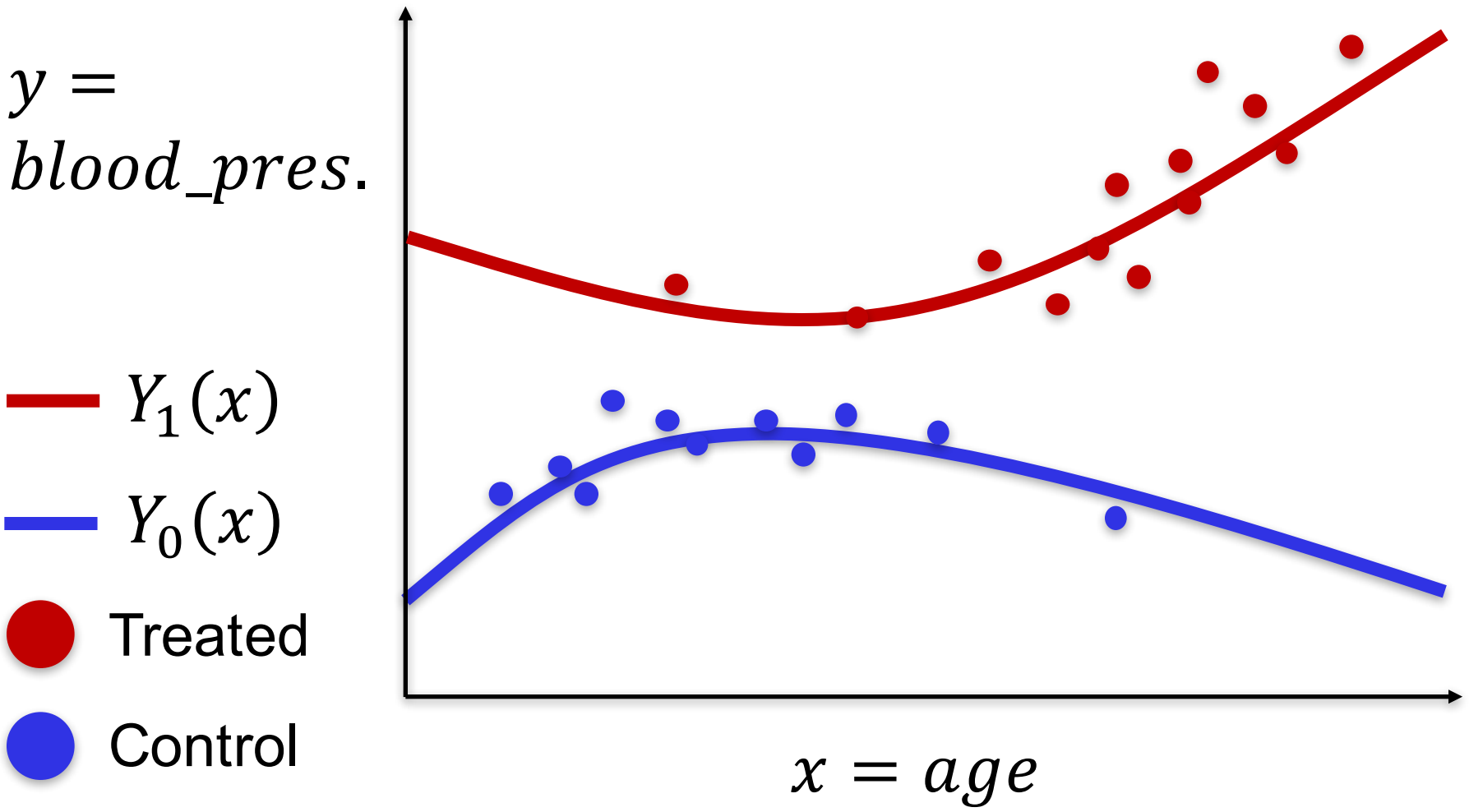
Covariate adjustment



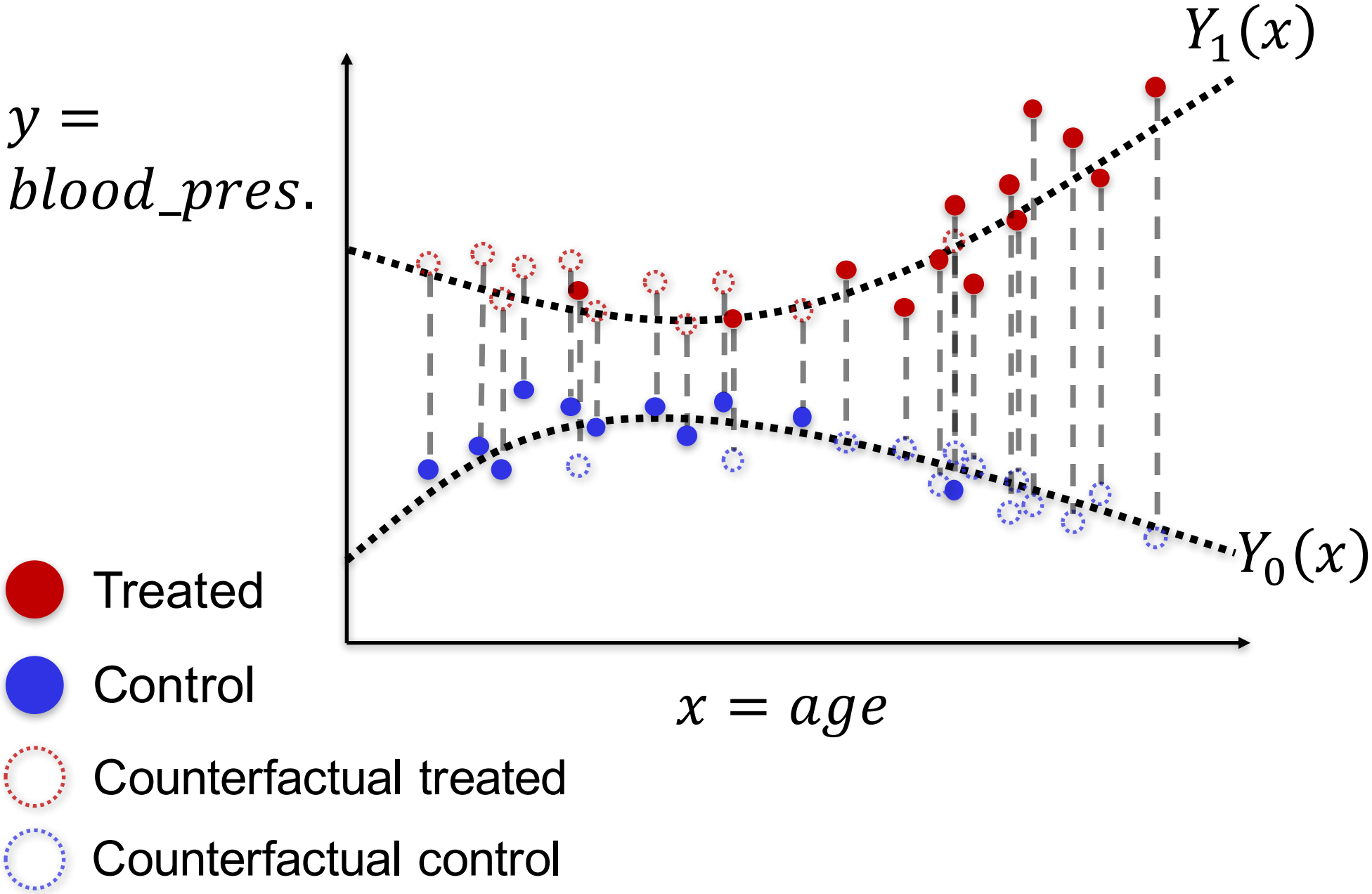
Covariate adjustment



Covariate adjustment



Covariate adjustment



Linear model

- Assume that:

Blood pressure **age** **medication**

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

- Then:

$$ITE(x) := Y_1(x) - Y_0(x) =$$

$$ATE := \mathbb{E}[Y_1(x) - Y_0(x)] = \gamma + \mathbb{E}[\epsilon_1] - \mathbb{E}[\epsilon_0]$$

Linear model

- Assume that:

$$Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

$$ATE = \mathbb{E}[Y_1(x) - Y_0(x)] = \gamma$$

- We care about γ, not about $Y_t(x)$
  Identification, not prediction

Linear model

blood pressure **age,weight,…** **medication**

- $Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t$

  Hypertension is affected by many variables: lifestyle, weight, genetics, age
- Each of these often stronger **predictor** of blood-pressure, compared with type of medication taken
- Regularization (e.g. Lasso) might remove the treatment variable!
- Features → (“nuisance parameters”, “variable of interest”)

Regression - misspecification

- True data generating process, $x \in \mathbb{R}$:

$$Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2$$

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

- Hypothesized model:

$$\widehat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t$$

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

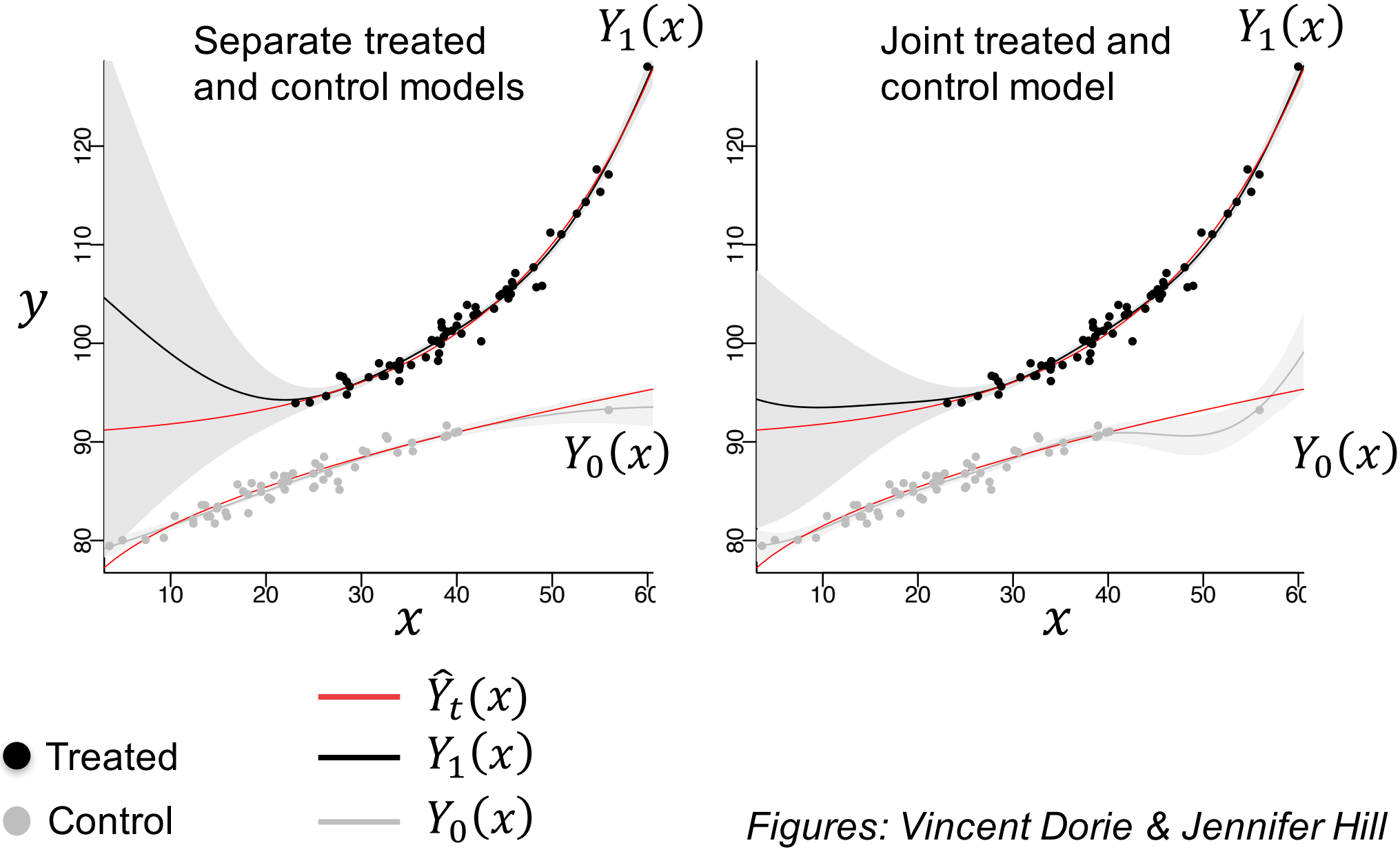
Using machine learning for causal inference

- 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)

Using machine learning for causal inference

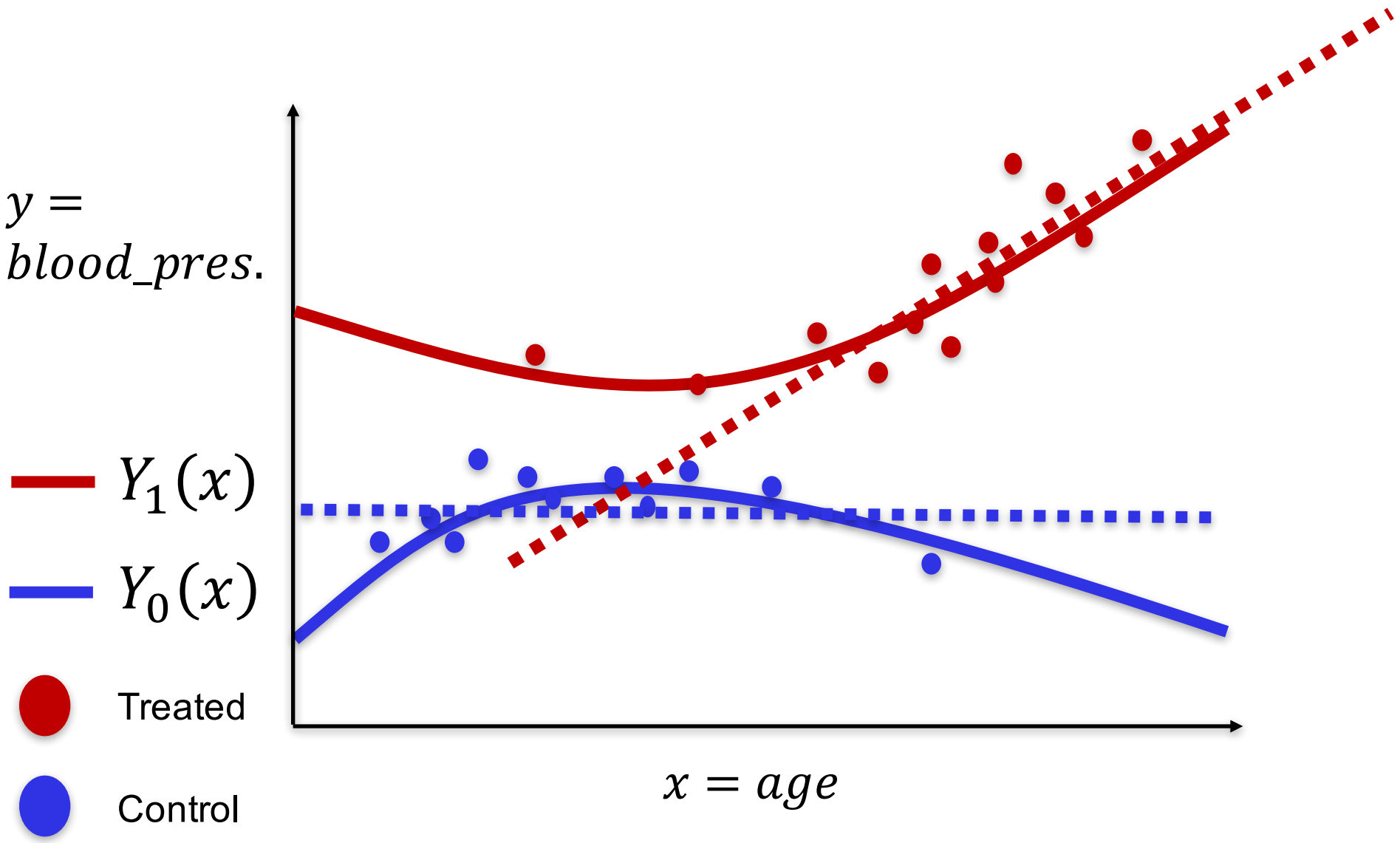
- Machine learning techniques can be very useful and have recently seen wider adoption
- How is the treatment variable used:
  - Fit two different models for treated and control?
  - Not regularized?
  - Privileged

Example: Gaussian process



Figures: Vincent Dorie & Jennifer Hill

Covariate adjustment: no overlap



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 in general inconsistent, though only with small bias (Imbens 2004)

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

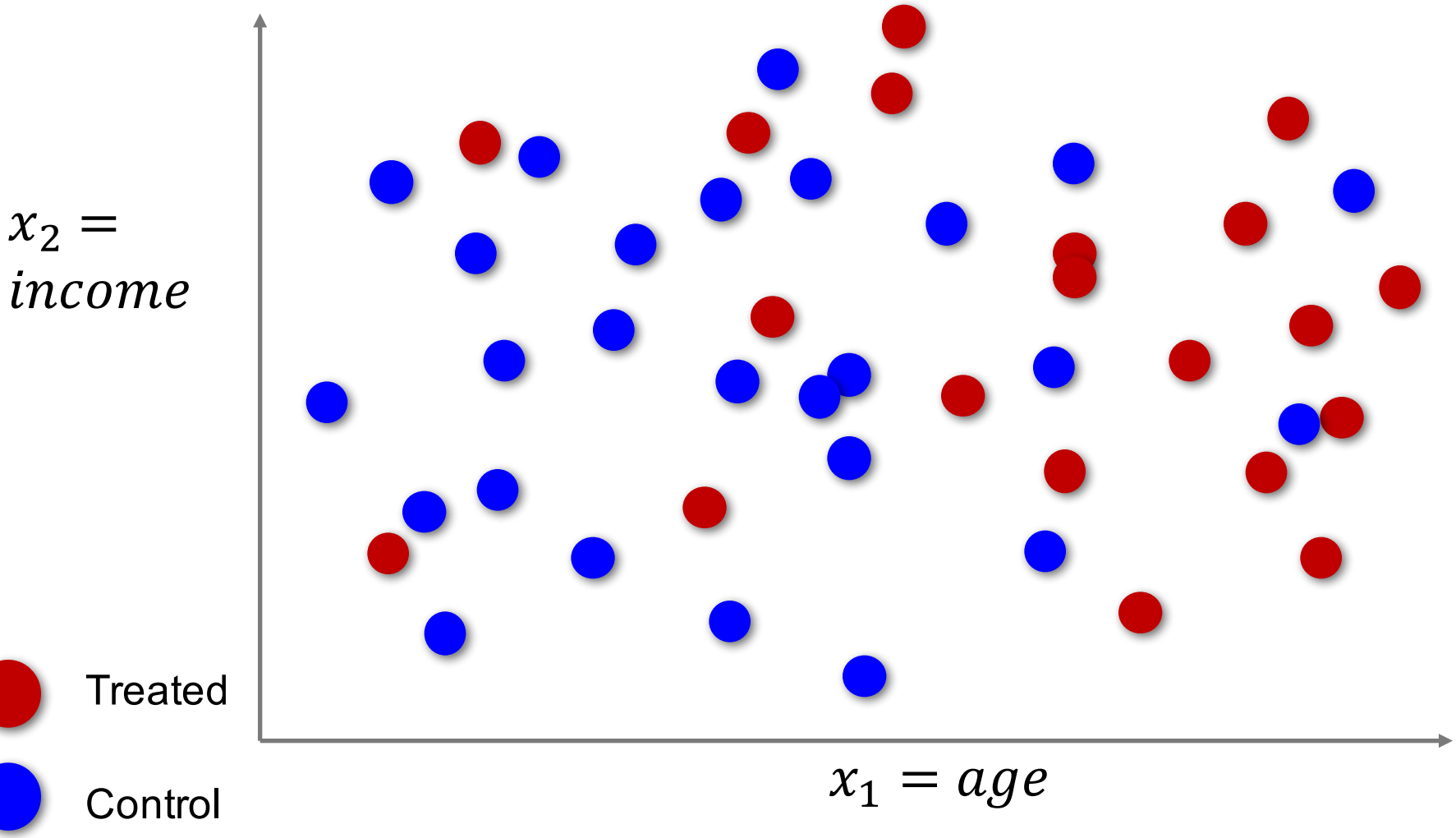
Propensity score

- Tool for estimating ATE
- Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples, similar to importance sampling

Inverse propensity score re-weighting

$p(x|t=0) \cdot w_0(x) \neq p(x|t=1) \cdot w_1(x)$

reweighted control *reweighted treated*



Propensity score

- Propensity score: $p(T = 1|x)$, using machine learning tools
- Samples re-weighted by the inverse propensity score of the treatment they received

How to obtain ATE with propensity score

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$
2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

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

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$
2. $\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

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

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

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

$$\frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Sum over $\sim \frac{n}{2}$ terms

Propensity scores - derivation

- Recall average treatment effect:

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

- We only have samples for:

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

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

Propensity scores - derivation

- We only have samples for:

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

Propensity scores - derivation

- We only have samples for:

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

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

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot \quad ? \quad = p(x)$$

Propensity scores - derivation

- We only have samples for:

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

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

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot \frac{p(T=1)}{p(T=1|x)} = p(x)$$

Propensity score

Propensity scores - derivation

- We only have samples for:

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

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

- We need to turn $p(x|T=0)$ into $p(x)$:

$$p(x|T=0) \cdot \frac{p(T=0)}{p(T=0|x)} = p(x)$$

Propensity score

- We only have samples for:

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

- We want:

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

- We know that:

$$p(x|T=1)\cdot\frac{p(T=1)}{p(T=1|x)}=p(x)$$

- Then:

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

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

Calculating the propensity score

- If $p(T = t|x)$ is known, then propensity scores re-weighting is consistent
- Example: ad-placement algorithm samples $T = t$ based on a known algorithm
- Usually the score is unknown and must be estimated
- Example: use logistic regression to estimate the probability that patient x received medication $T = t$
- Calibration: must estimate the probability correctly, not just the binary assignment variable

“The Assumptions” – ignorability

- If ignorability doesn’t hold then the average treatment effect is **not** $\mathbb{E}_{x\sim p(x)}\left[\mathbb{E}[Y_1|T=1,x]-\mathbb{E}[Y_0|T=0,x]\right]$, invalidating the starting point of the derivation

“The Assumptions” – overlap

- If there’s not much overlap, propensity scores become non-informative and easily miscalibrated

- Sample variance of inverse propensity score re-weighting scales with $\sum_{i=1}^{n} \frac{1}{\hat{p}(T=1|x_i)\hat{p}(T=0|x_i)}$, which can grow very large when samples are non-overlapping
(Williamson et al., 2014)

Propensity score in machine learning

- 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)

Propensity scores

- The propensity score is the "coarsest" function of x such that :

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x \Rightarrow$$
$$(Y_0, Y_1) \perp\!\!\!\perp T \mid p(T = 1|x)$$

(Rubin & Rosenbaum, 1983)

Propensity scores - other uses

- Use propensity score as a covariate in covariate adjustment
- Match on propensity score:
  unbiased because 1-NN is unbiased in one-dimensonal space (Imbens, 2004)
- Doubly robust methods – see next section
- Give insight into treatment assignment mechanism

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

Outline

Tools of the trade

- Matching
- Covariate adjustment
- Propensity score
- Double robustness

Doubly robust methods

- 1) Covariate adjustment: model outcome as function of treatment and covariates
- 2) Propensity score: ignores outcome, model treatment assignment as function of covariates
- 1) and 2) are sensitive to model misspecification
- Doubly robust: combine 1) and 2) and create an estimator which is consistent ***if at least one of the models is well-specified***

Doubly robust methods

- Basic method: combine regression estimate with an inverse propensity weighted estimate of the regression residuals
- Many sophisticated methods exist, e.g. Targeted Maximum Likelihood (van der Laan & Rubin, 2006)
- Used in machine learning for off-policy evaluation (Langford et al. 2011) and covariate shift adjustment (Reddi et al. 2015)

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

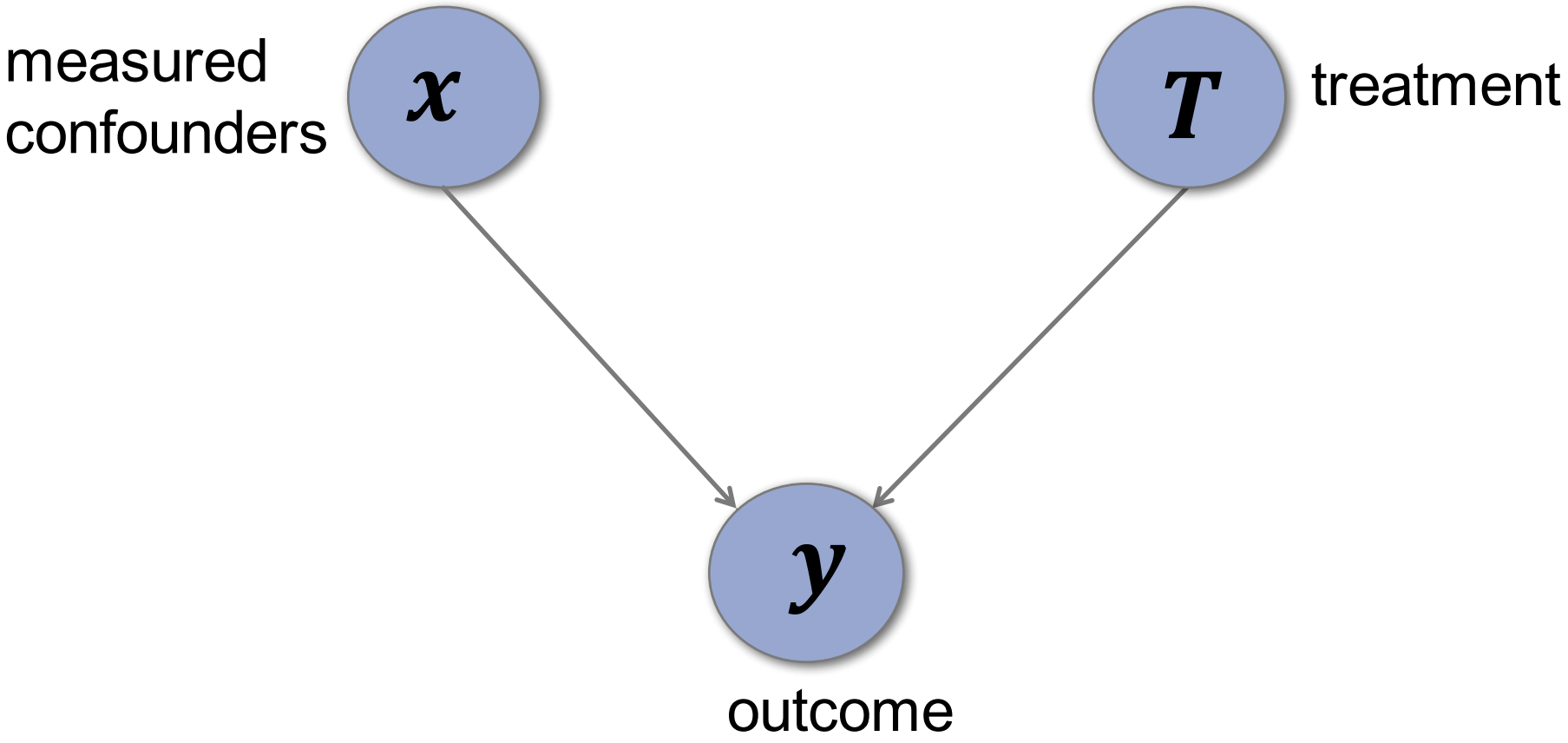
Causal graphs (Wright, Spirtes, Pearl)

- Graphical models with well defined causal interpretation and causal assumptions
- Causal graphs framework
  - Address causal inference questions when there is prior knowledge about the graph structure, particularly about hidden variables
  - Identifiability: can the causal effect be estimated from data?

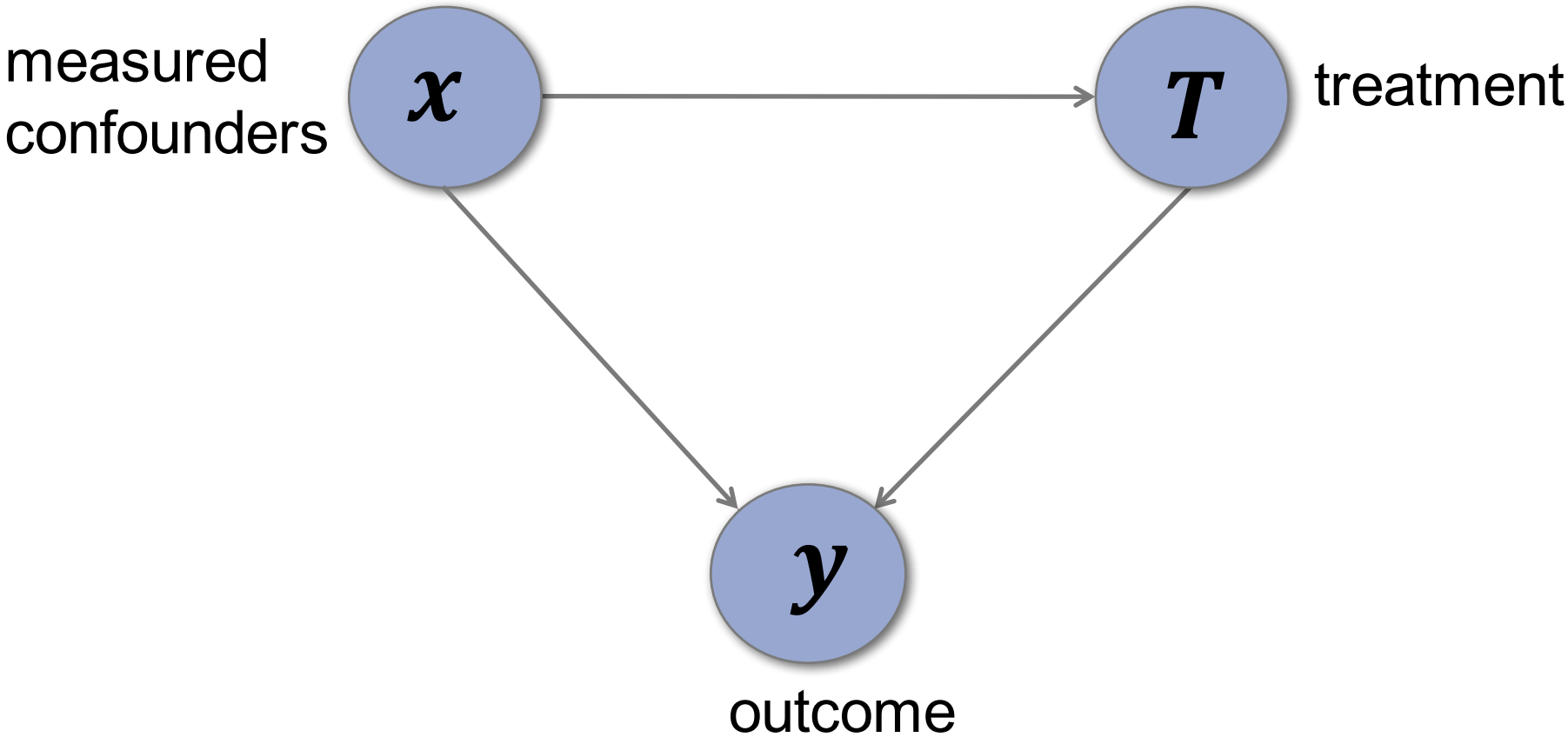
Causal graphs (Wright, Spirtes, Pearl)

- Causal graphs are a great way of formalizing when is correct causal inference *possible* in face of unmeasured confounders

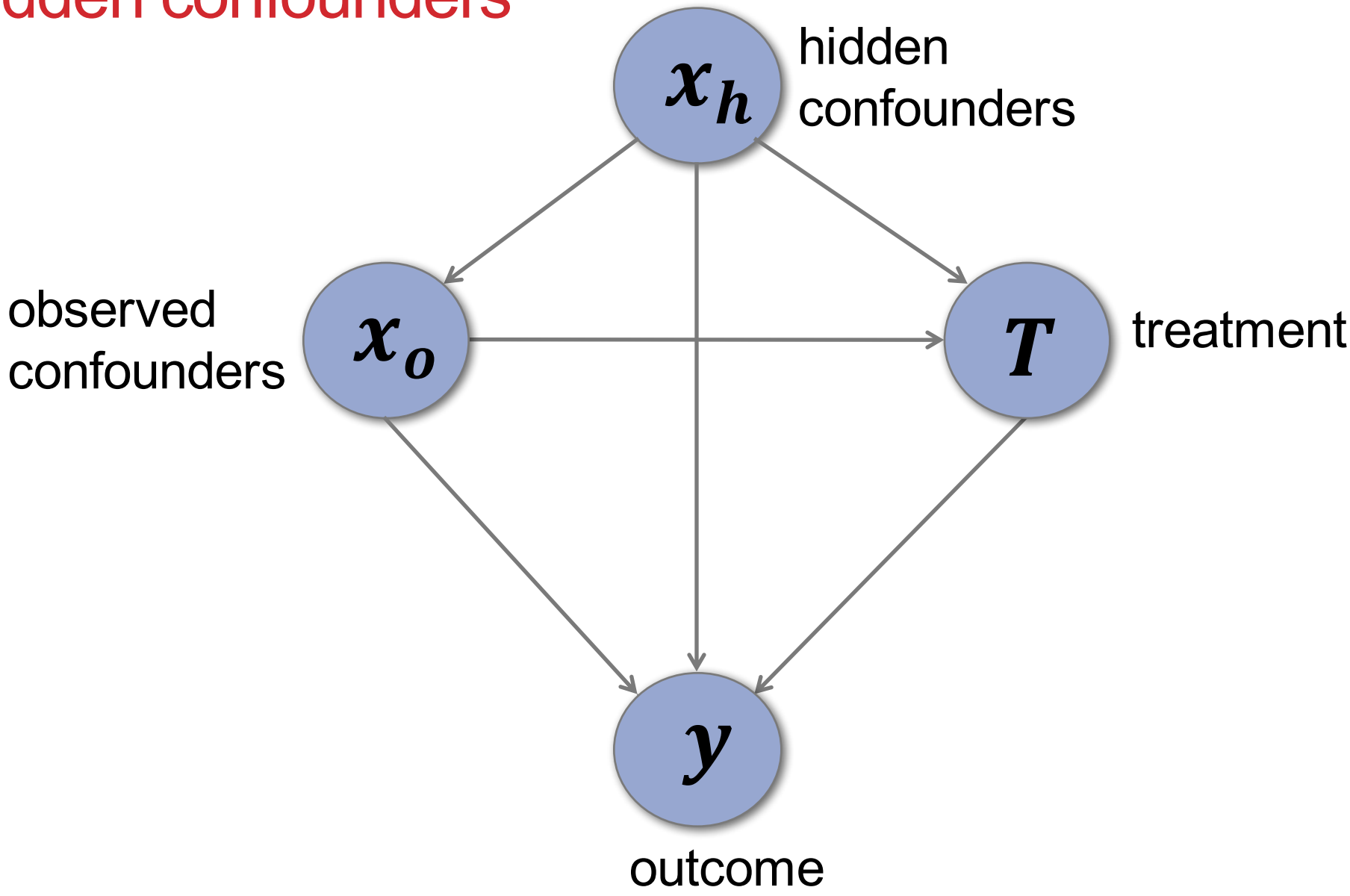
Randomized controlled study



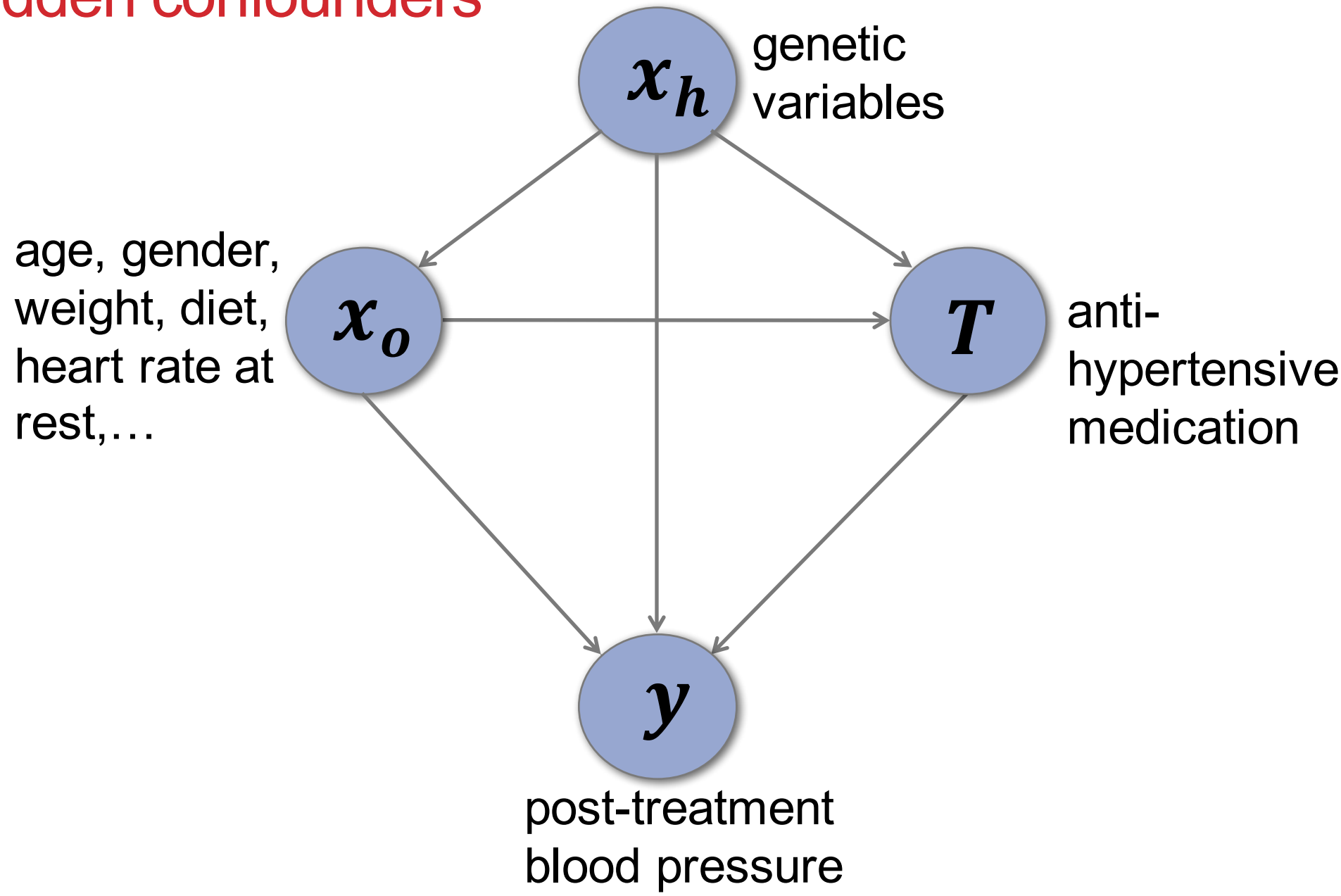
Observational study with no hidden confounders



Observational study with hidden confounders



Observational study with hidden confounders

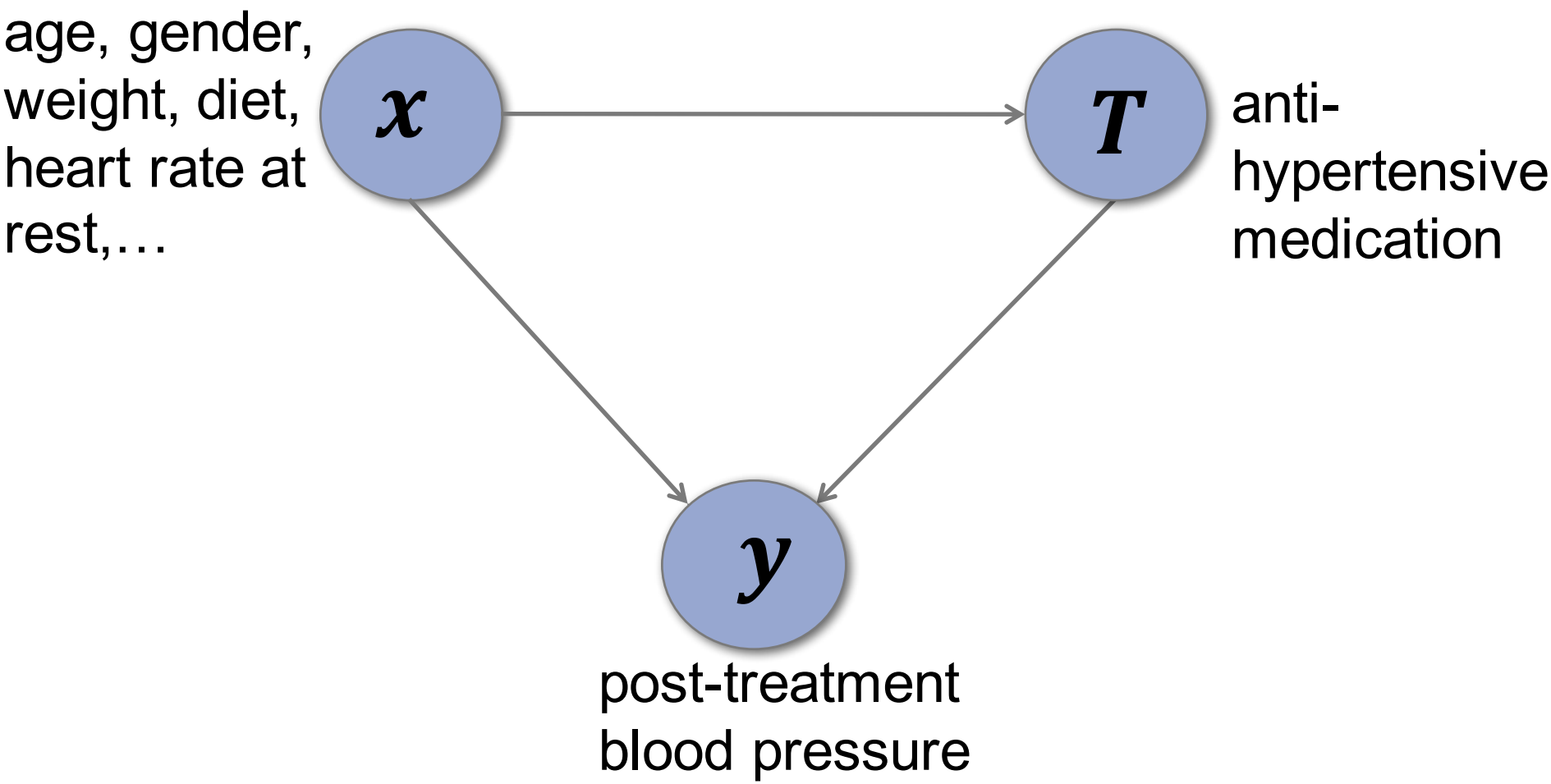


Hidden (unmeasured) confounders

- Examples:
  - Patient response to medication depends on unmeasured genetic variation with no other observed phenotype
  - Customers click on ads because they're next to another ad which we do not know of
  - Students succeed or fail based on home conditions which we do not measure
- **Cannot in general be tested from data!**

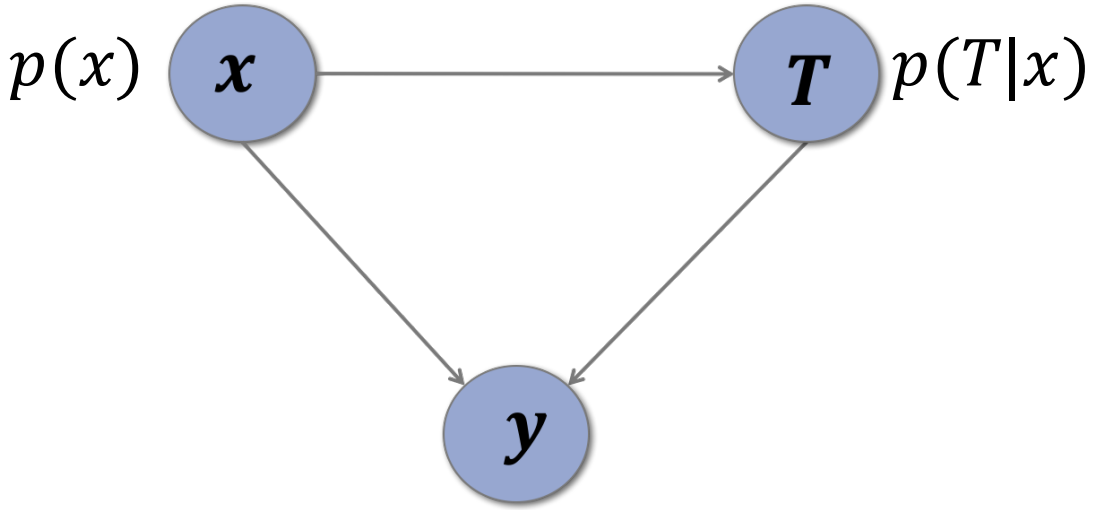
How to think about interventions: The do operator

(Pearl, 2009)



The do operator

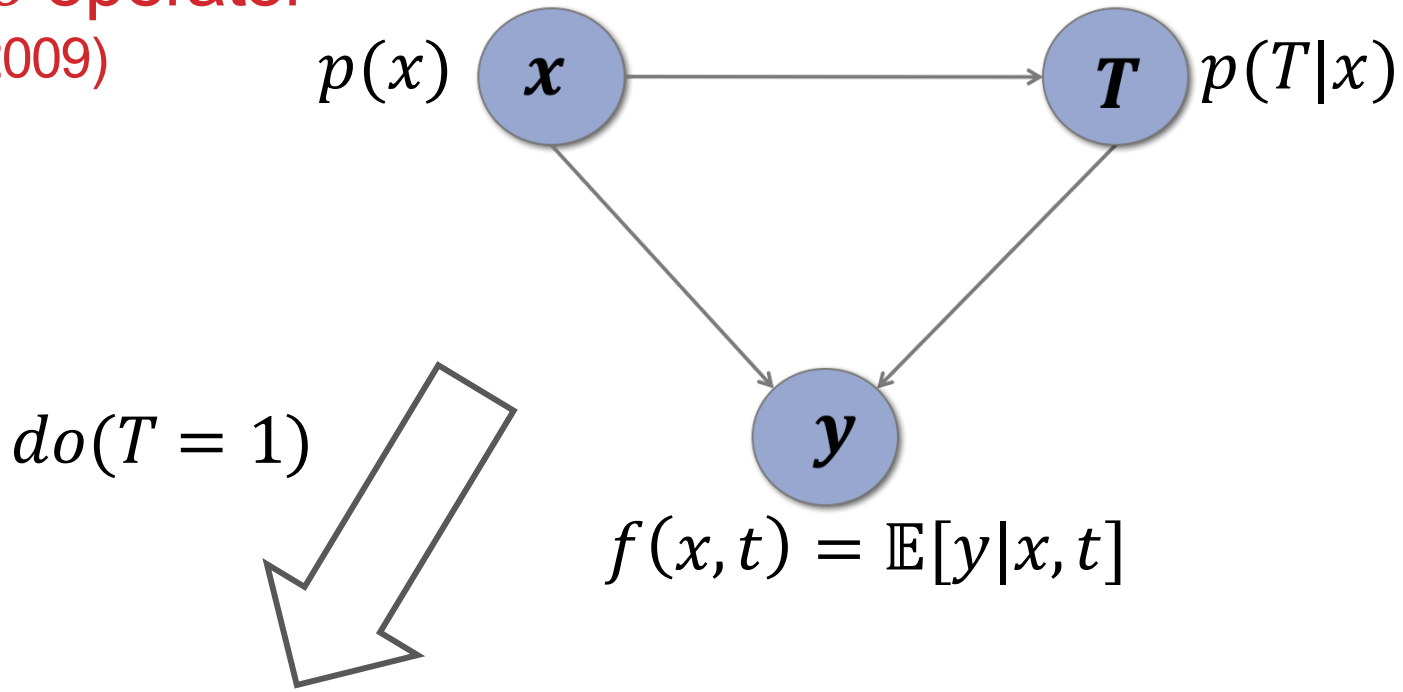
(Pearl, 2009)



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

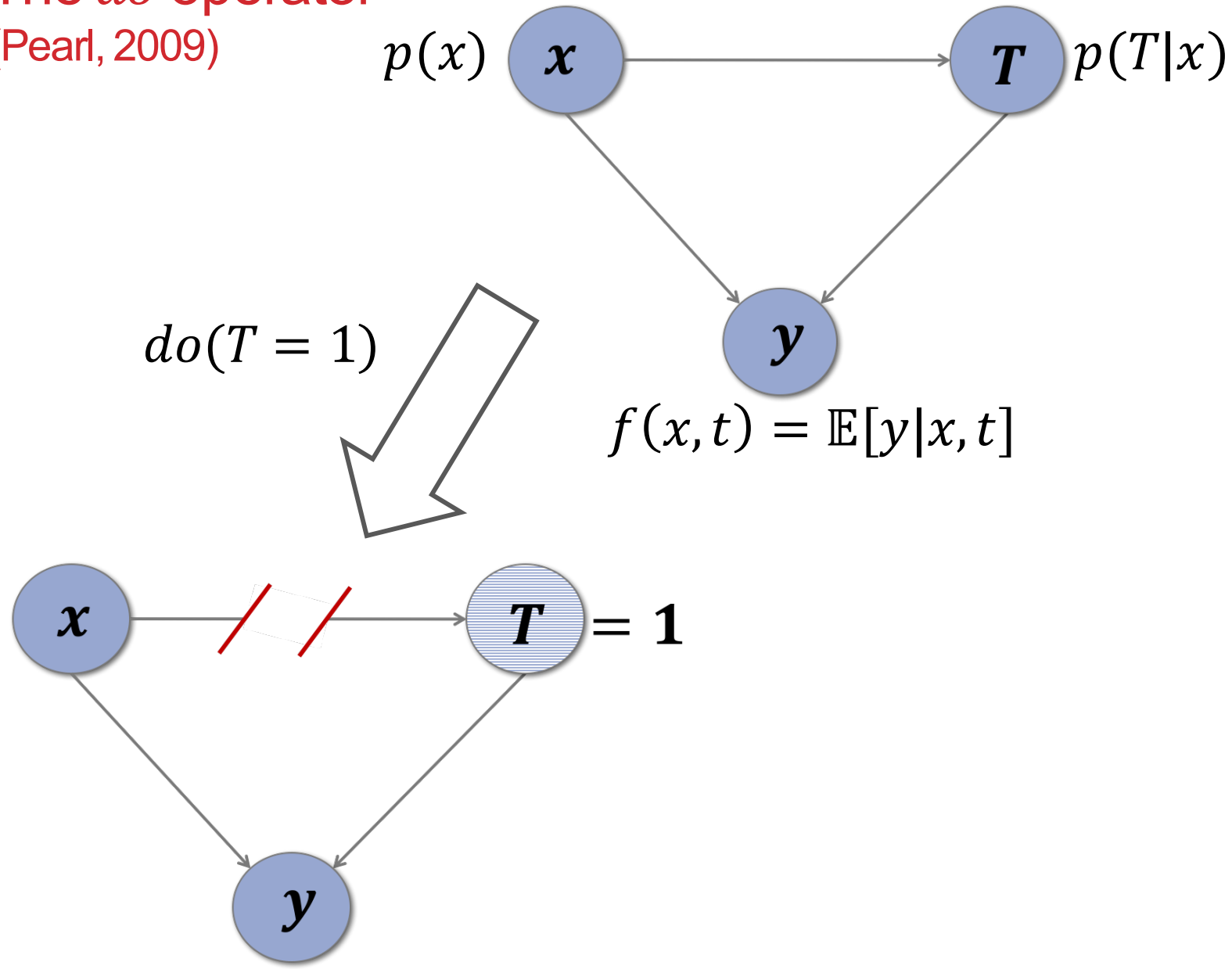
The do operator

(Pearl, 2009)



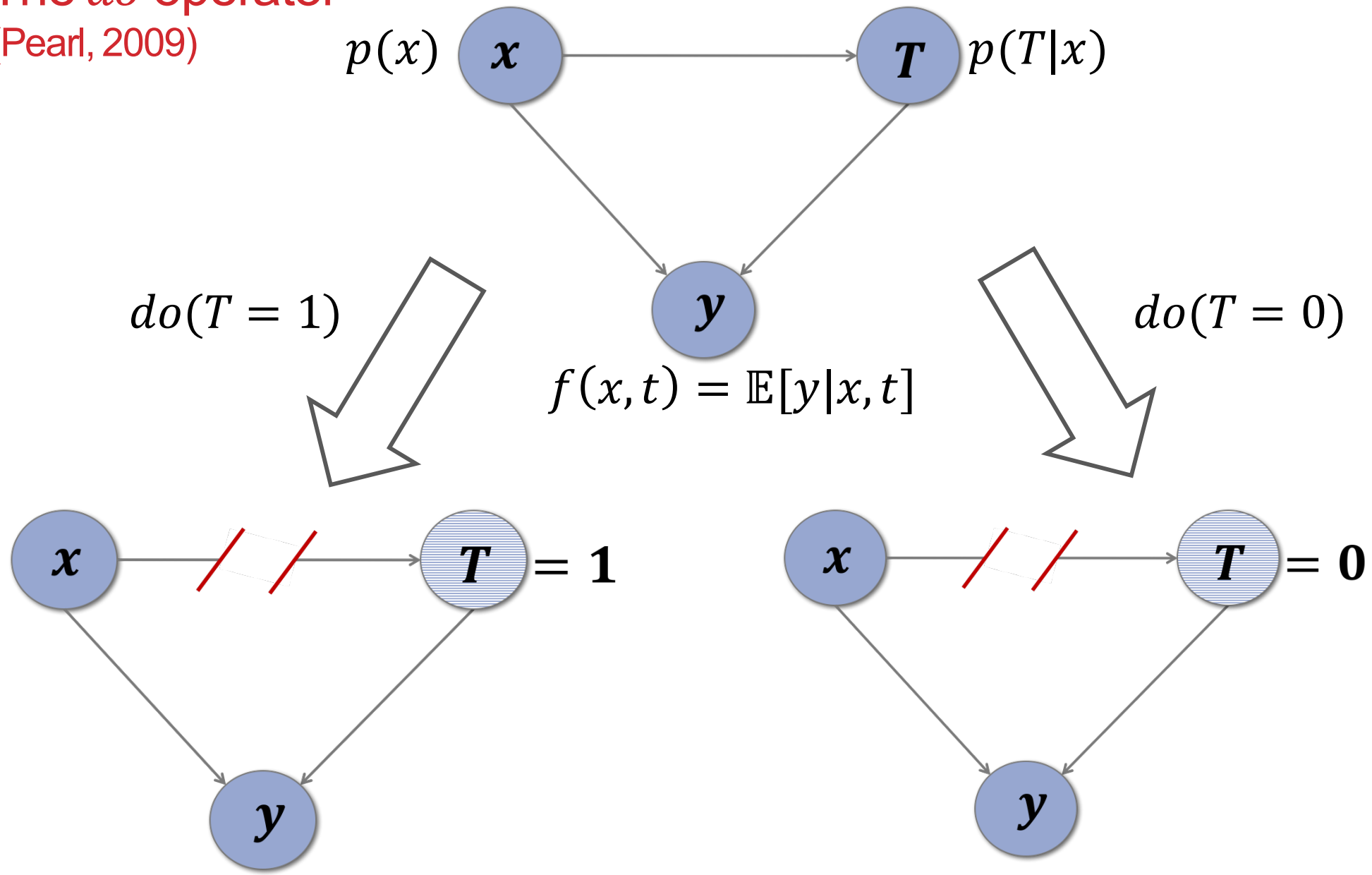
The *do* operator

(Pearl, 2009)



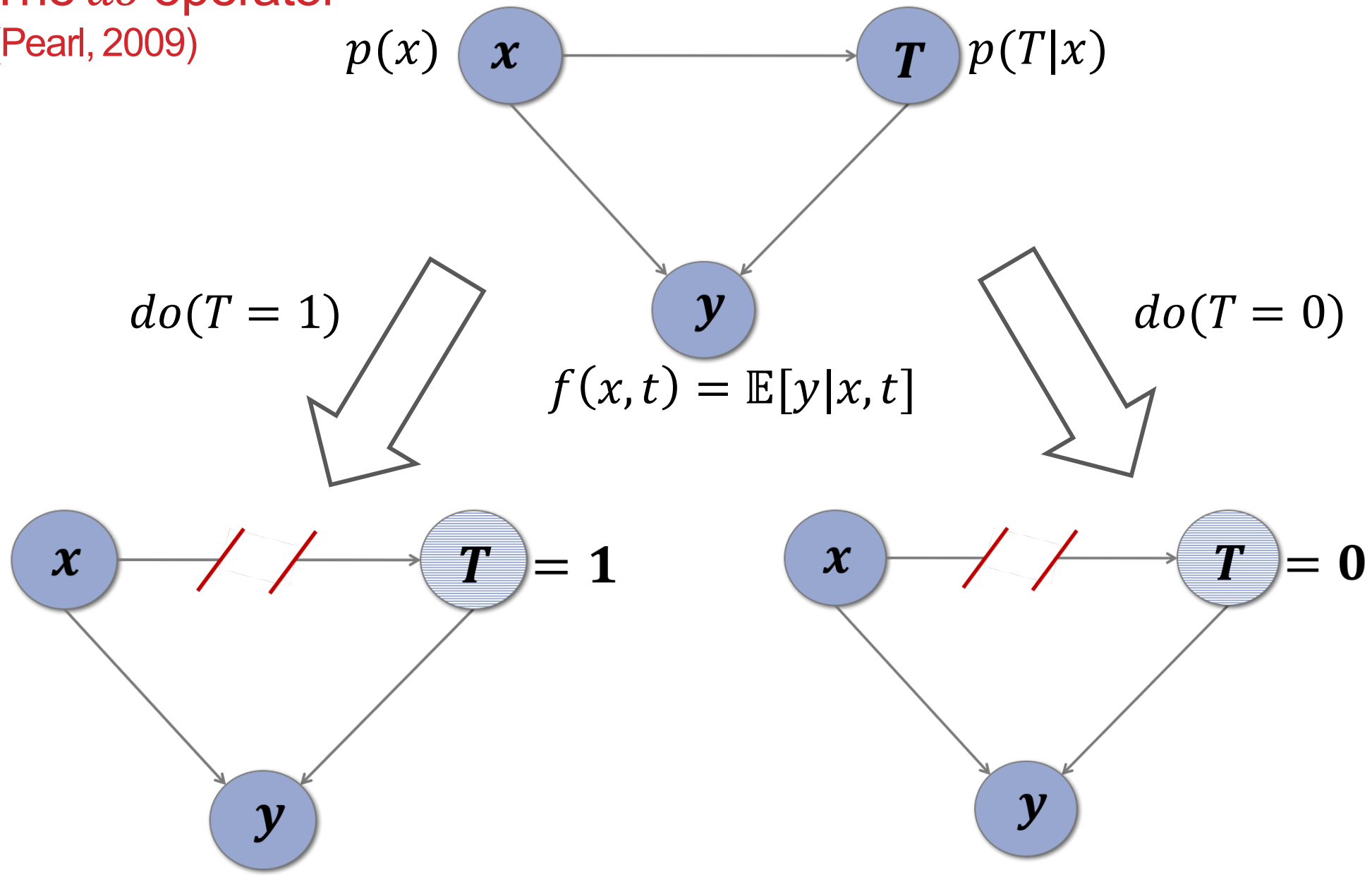
The *do* operator

(Pearl, 2009)



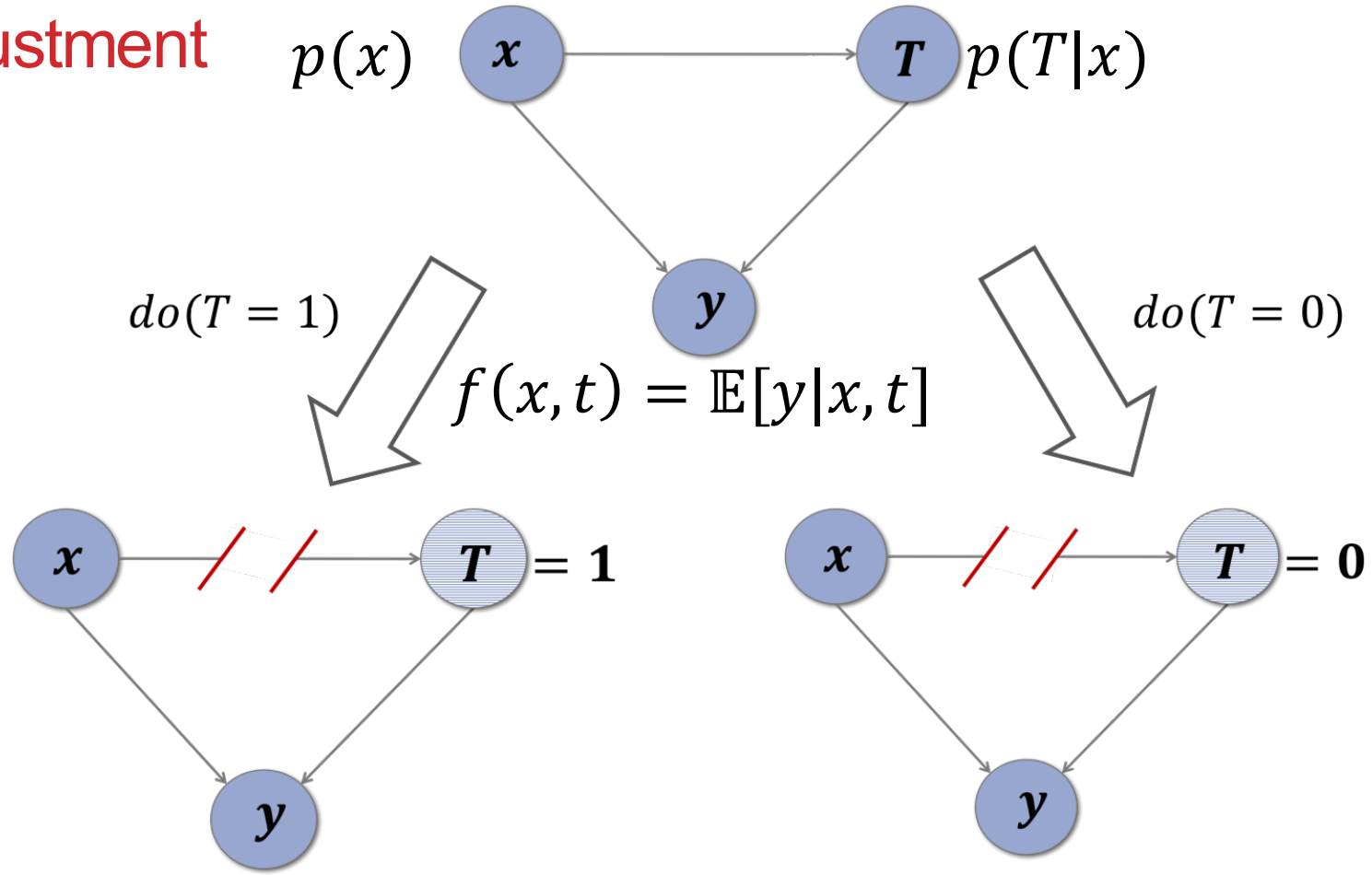
The *do* operator

(Pearl, 2009)



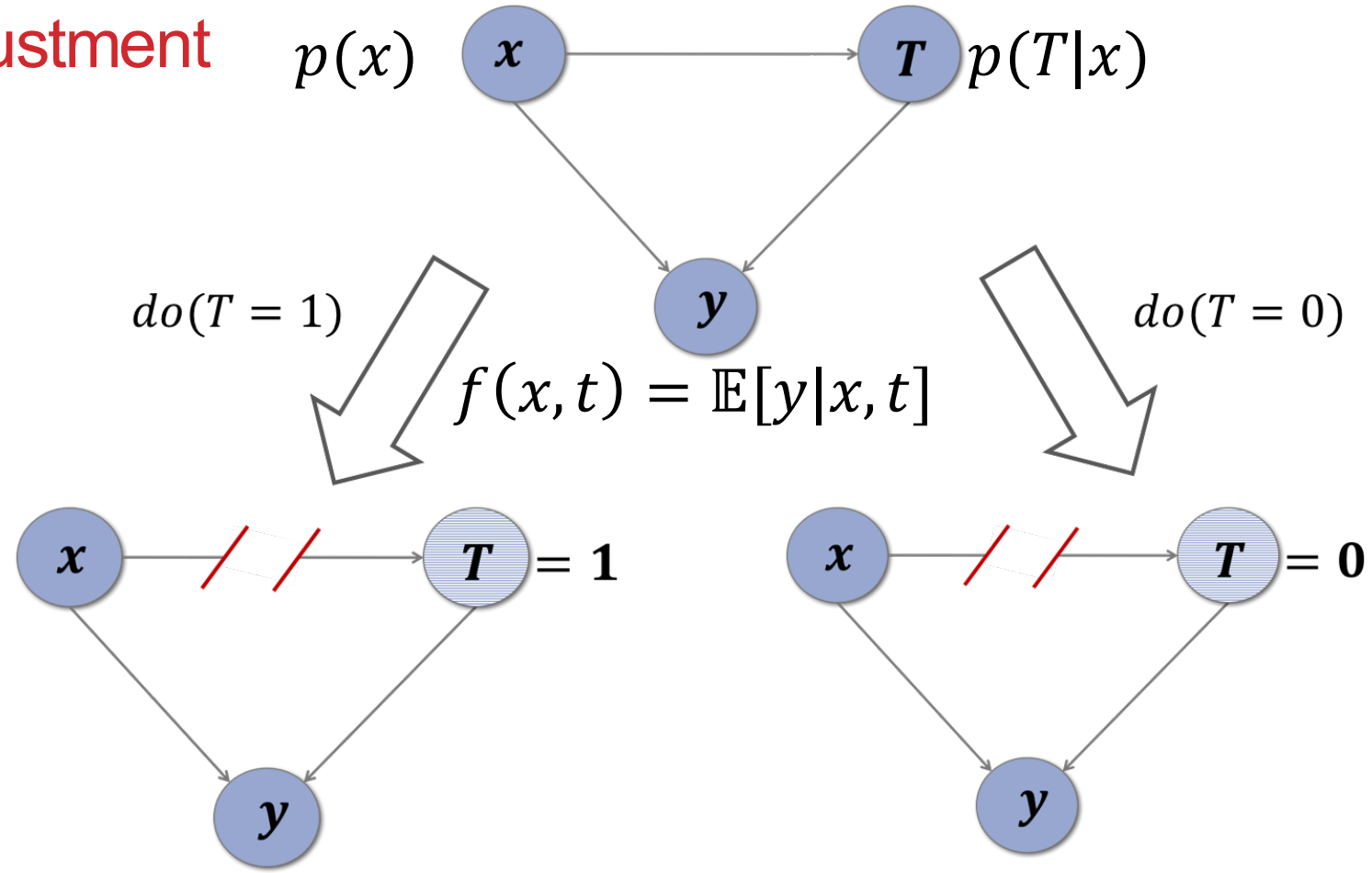
$$ATE := \mathbb{E}[y|do(T=1)] - \mathbb{E}[y|do(T=0)]$$

The *do* operator and adjustment formula



$$ATE := \mathbb{E}[y|do(T=1)] - \mathbb{E}[y|do(T=0)] =$$

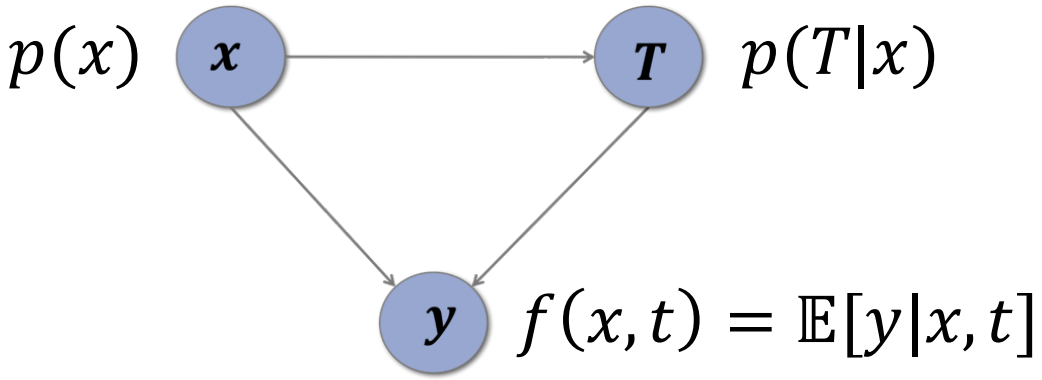
The *do* operator and adjustment formula



$$ATE := \mathbb{E}[y|do(T=1)] - \mathbb{E}[y|do(T=0)] =$$

$$\mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=1)]] - \mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=0)]]$$

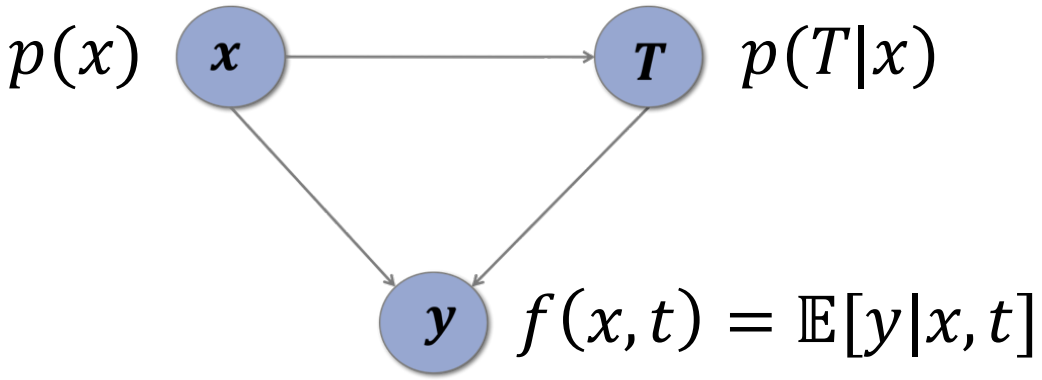
The *do* operator and adjustment formula



$$ATE := \mathbb{E}[y|do(T=1)] - \mathbb{E}[y|do(T=0)] =$$

$$\mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=1)]] - \mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=0)]]$$

The *do* operator and adjustment formula



$$ATE := \mathbb{E}[y|do(T=1)] - \mathbb{E}[y|do(T=0)] =$$

$$\mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=1)]] - \mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=0)]]$$

Covariate adjustment formula:

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

How to think about interventions: The do operator

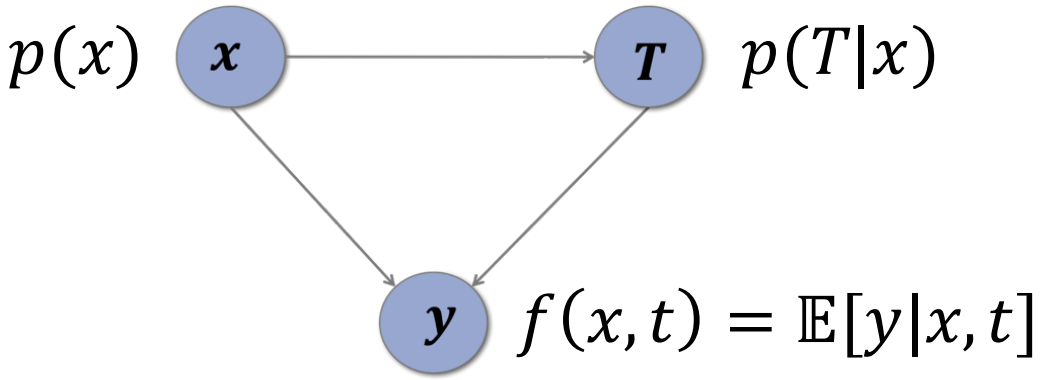
(Pearl, 2009)

In a simple observational study with *no unmeasured confounding*:

- T is intervened on
- x is observed
- Only need to model y as a function of (x, T)

- In this case, do-operator causal effect is equivalent to covariate adjustment

The *do* operator and adjustment formula

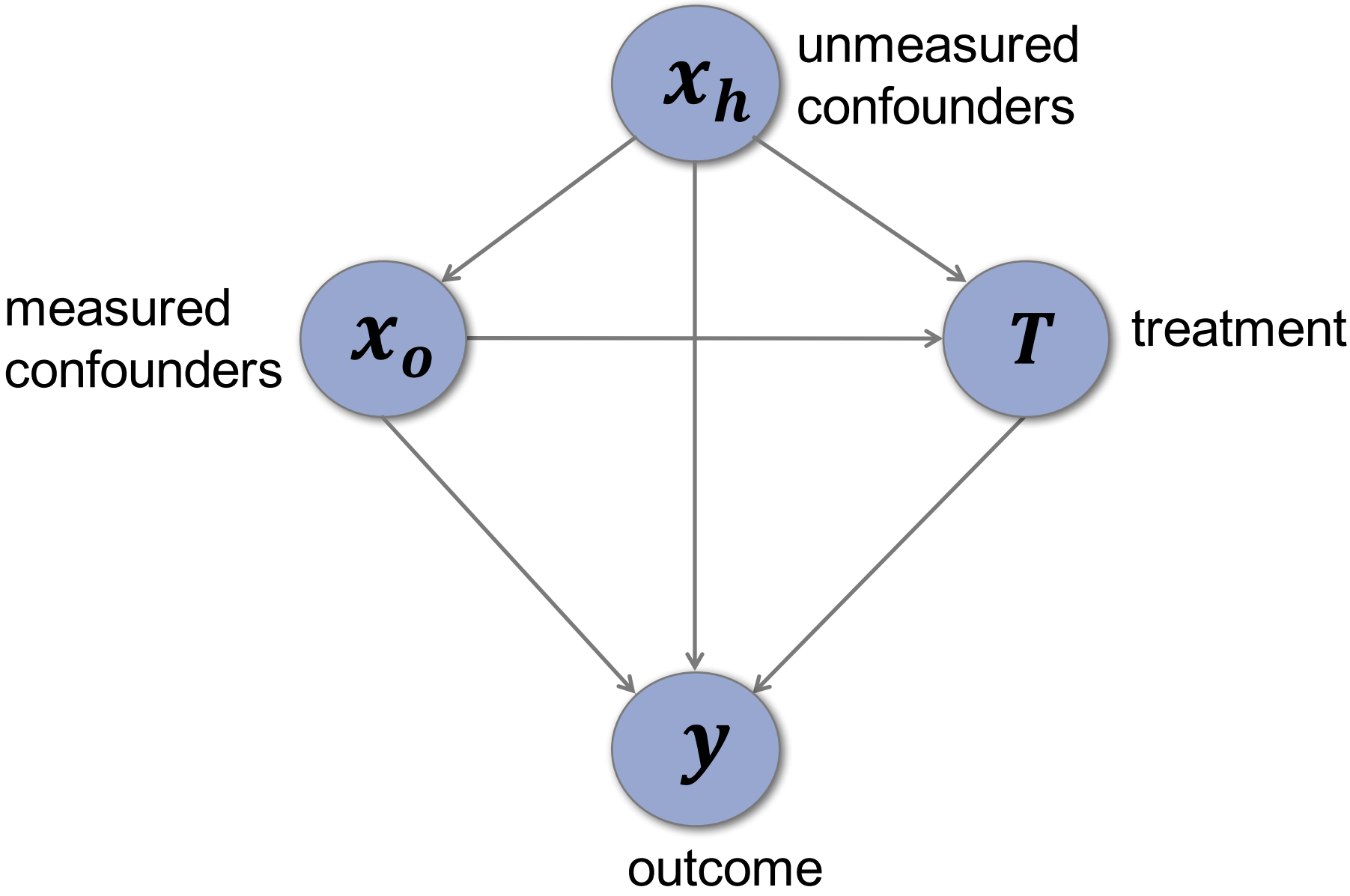


$$ATE := \mathbb{E}[y|do(T=1)] - \mathbb{E}[y|do(T=0)] =$$

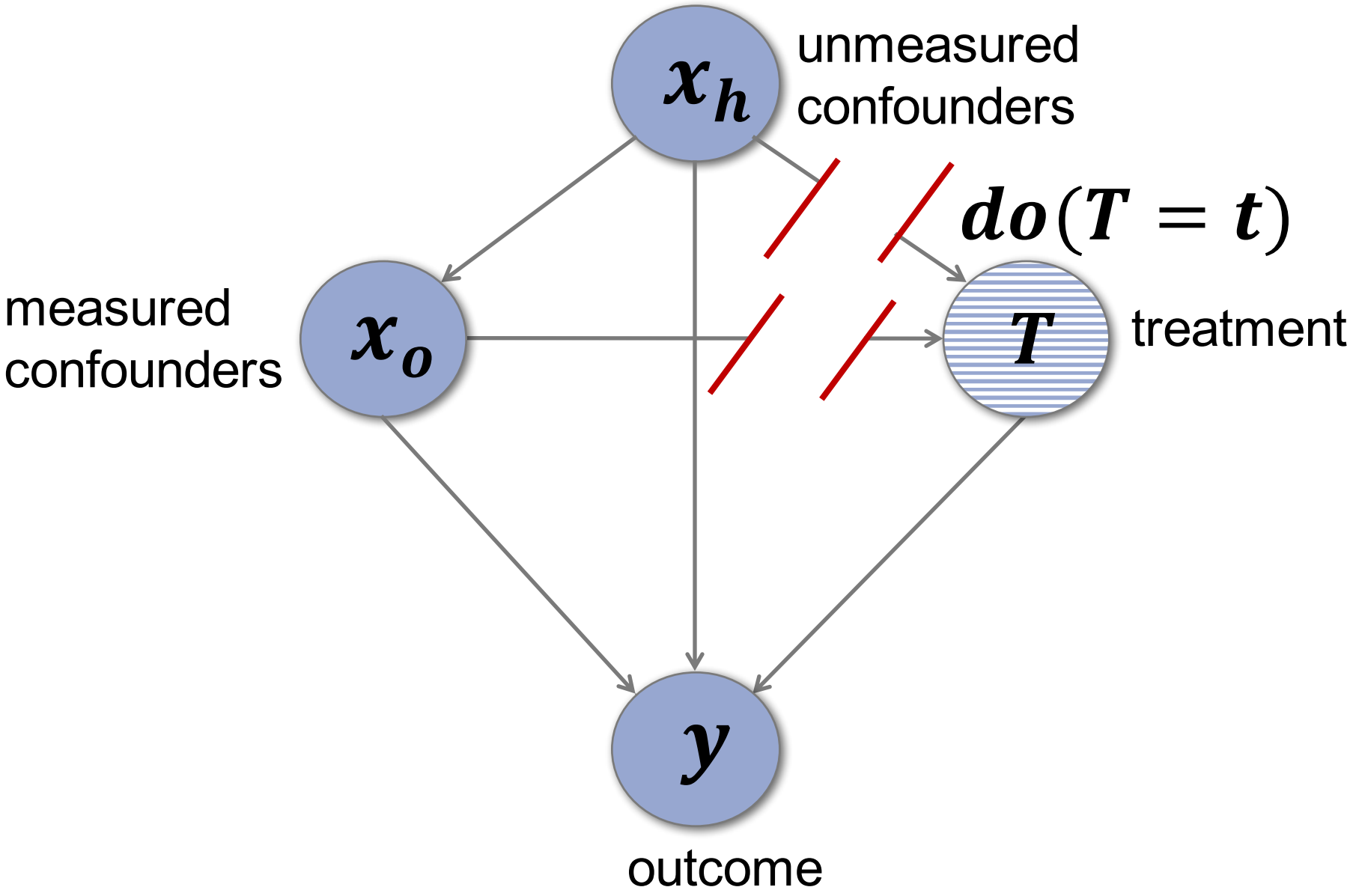
$$\mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=1)]] - \mathbb{E}_{x\sim p(x)}[\mathbb{E}[y|x,do(T=0)]]$$

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

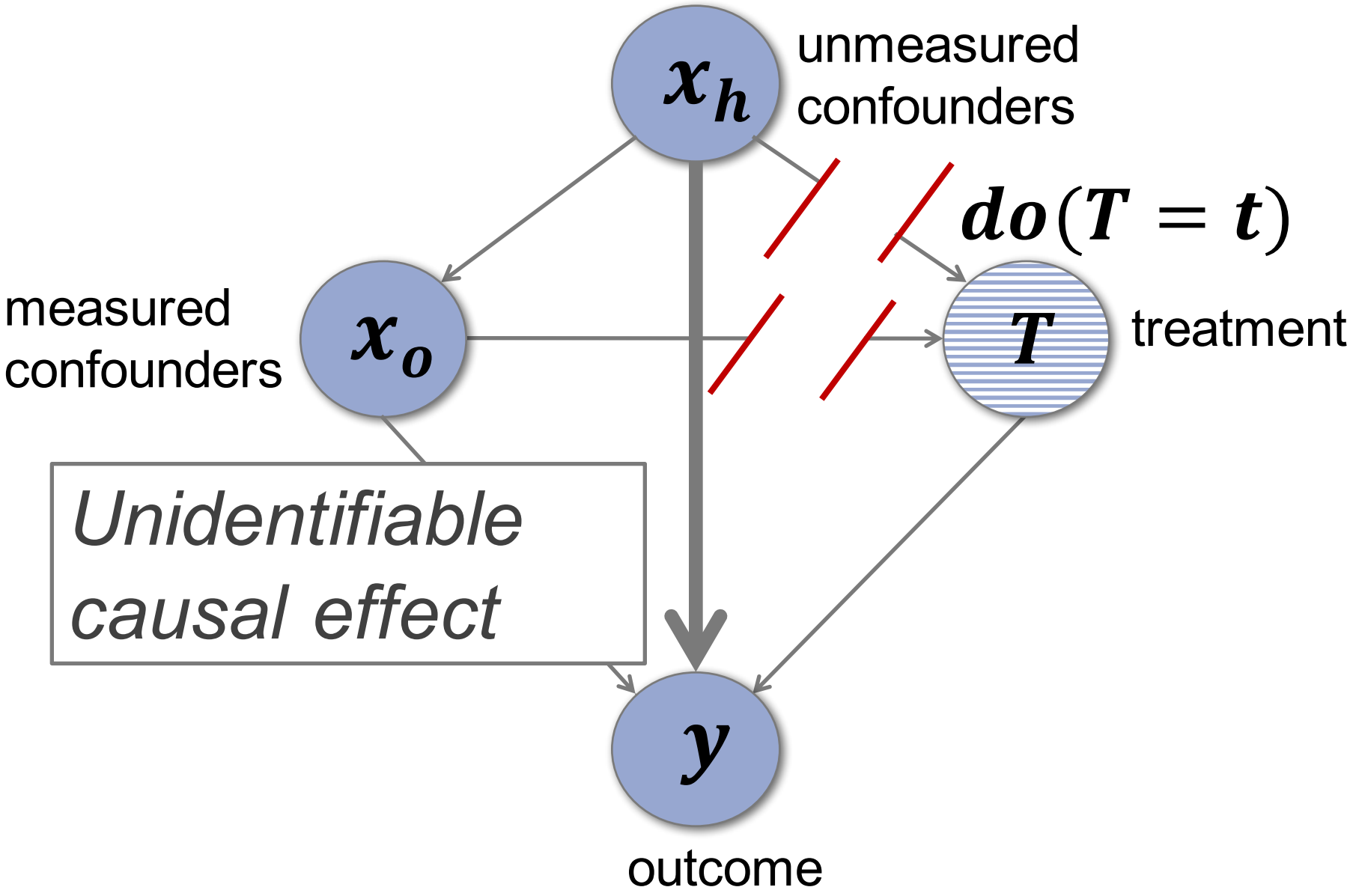
Hidden confounding



Hidden confounding



Hidden confounding



The Assumptions: causal identifiability

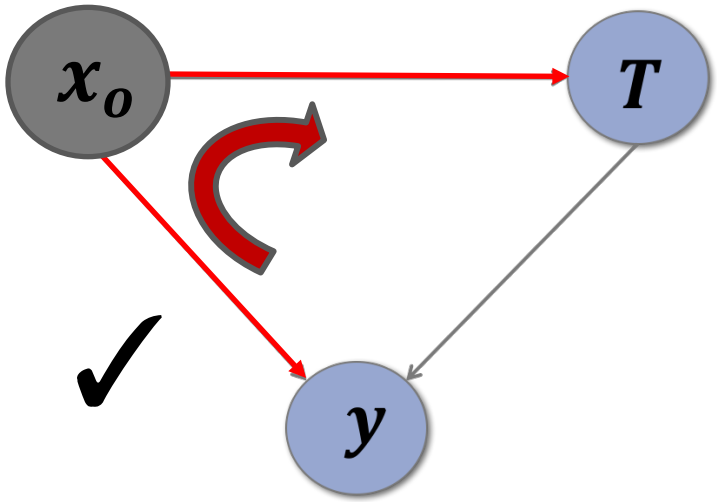
- A very useful sufficient condition for causal identifiability is the *back-door criterion*

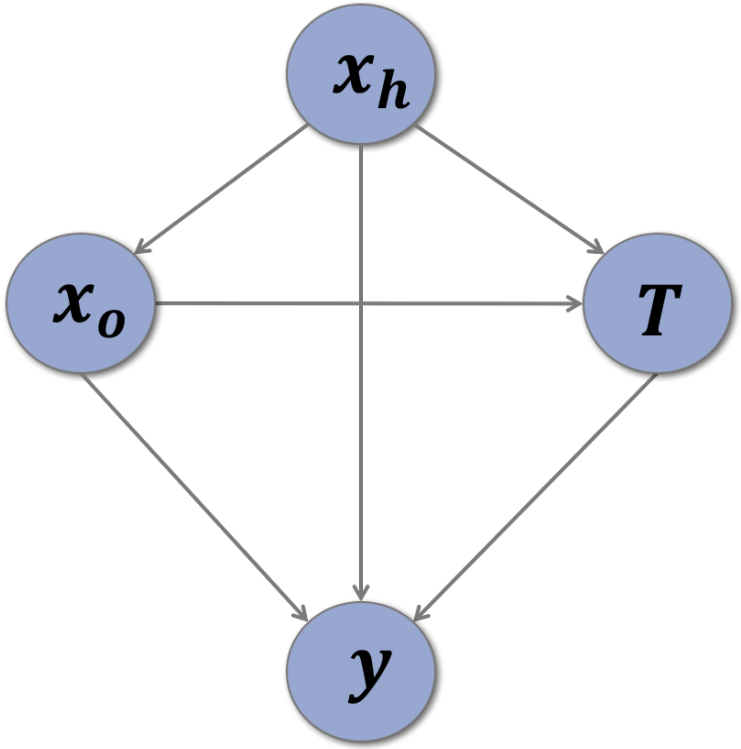
The Assumptions: causal identifiability

- Back-door criterion (Pearl, 1993, 2009):
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T
- Tells us what can we measure that will ensure causal identifiability
- There are other useful sufficient conditions, for example the “front-door criterion” (Pearl, 2009)

The Assumptions: causal identifiability

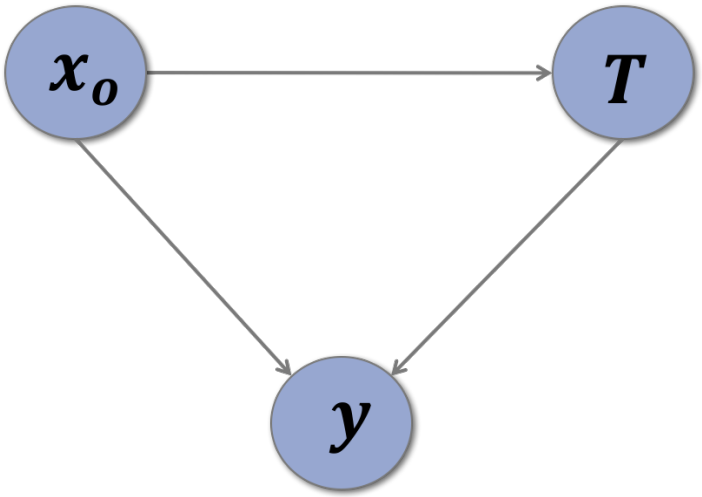
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

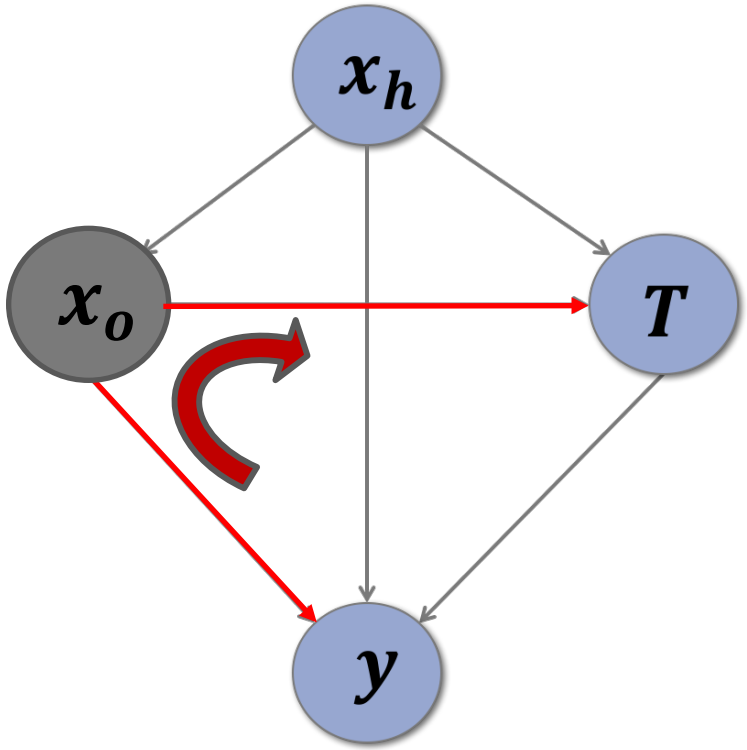




The Assumptions: causal identifiability

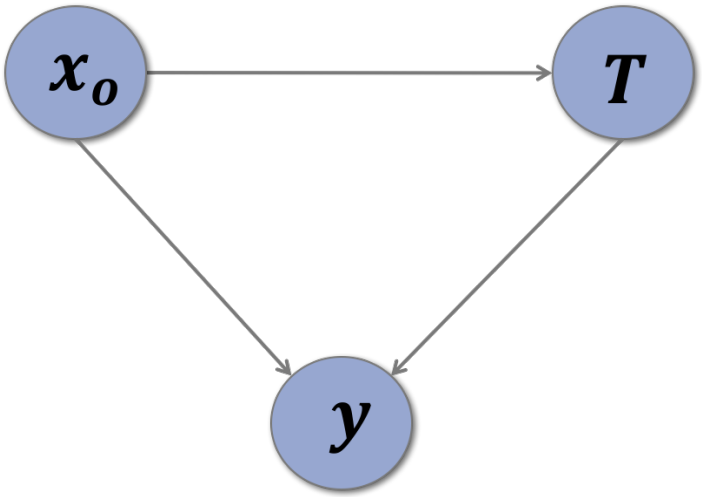
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

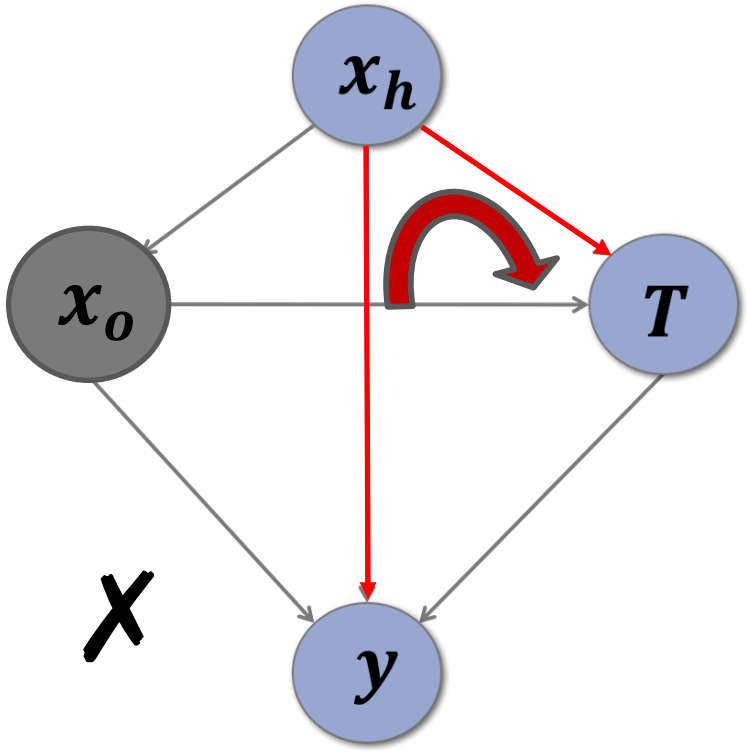




The Assumptions: causal identifiability

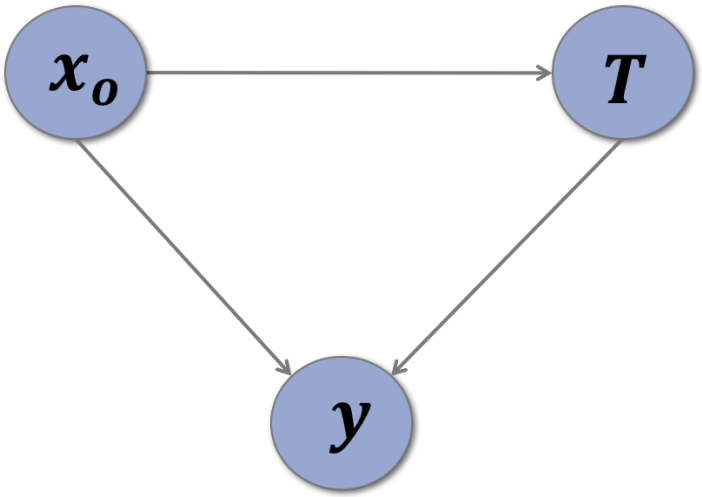
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

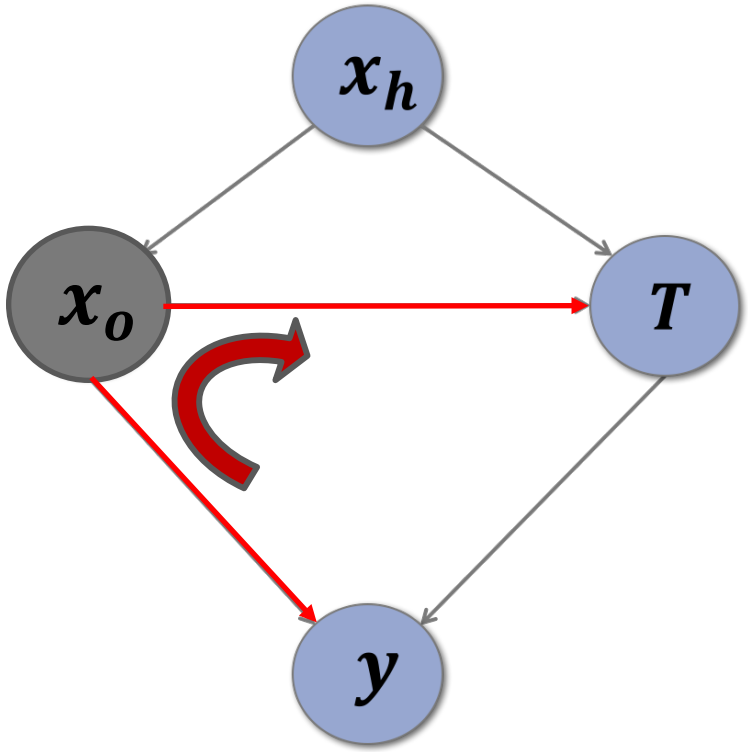




The Assumptions: causal identifiability

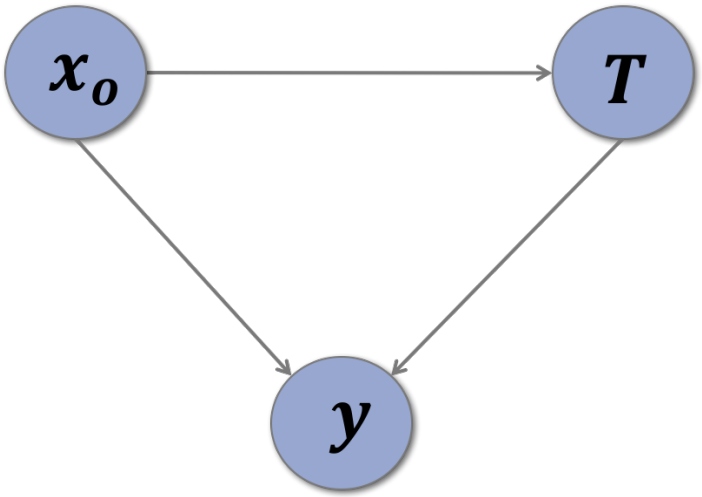
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

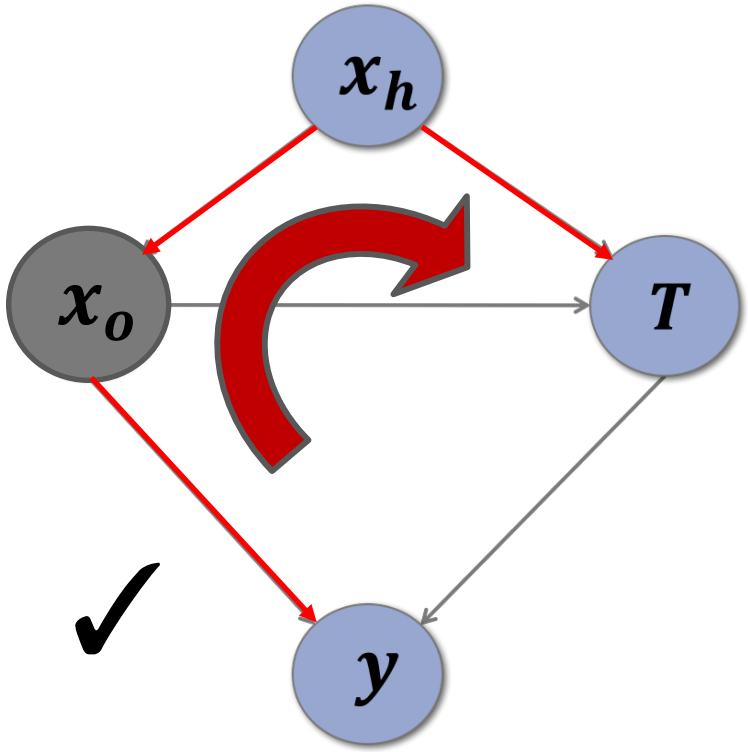




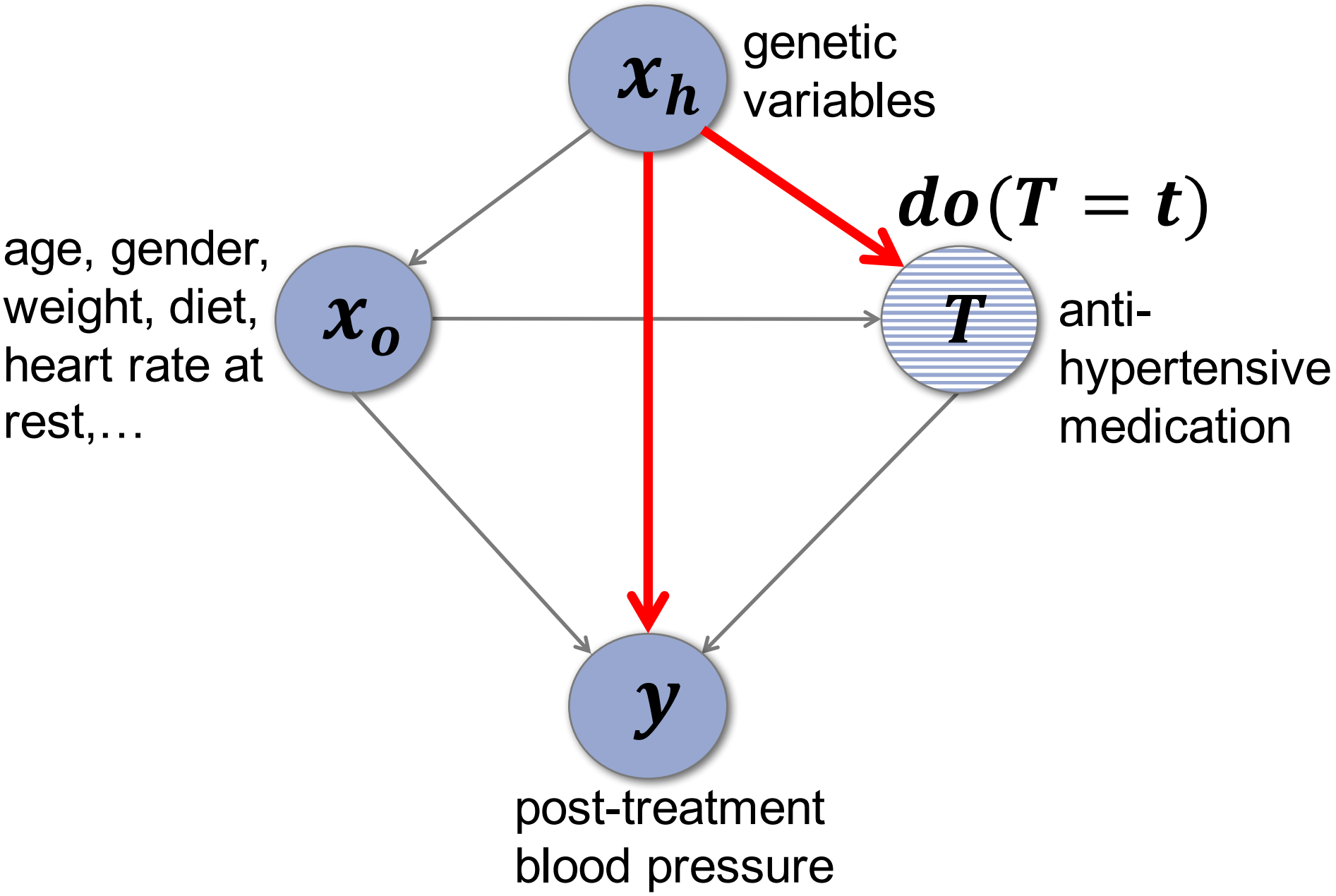
The Assumptions: causal identifiability

- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

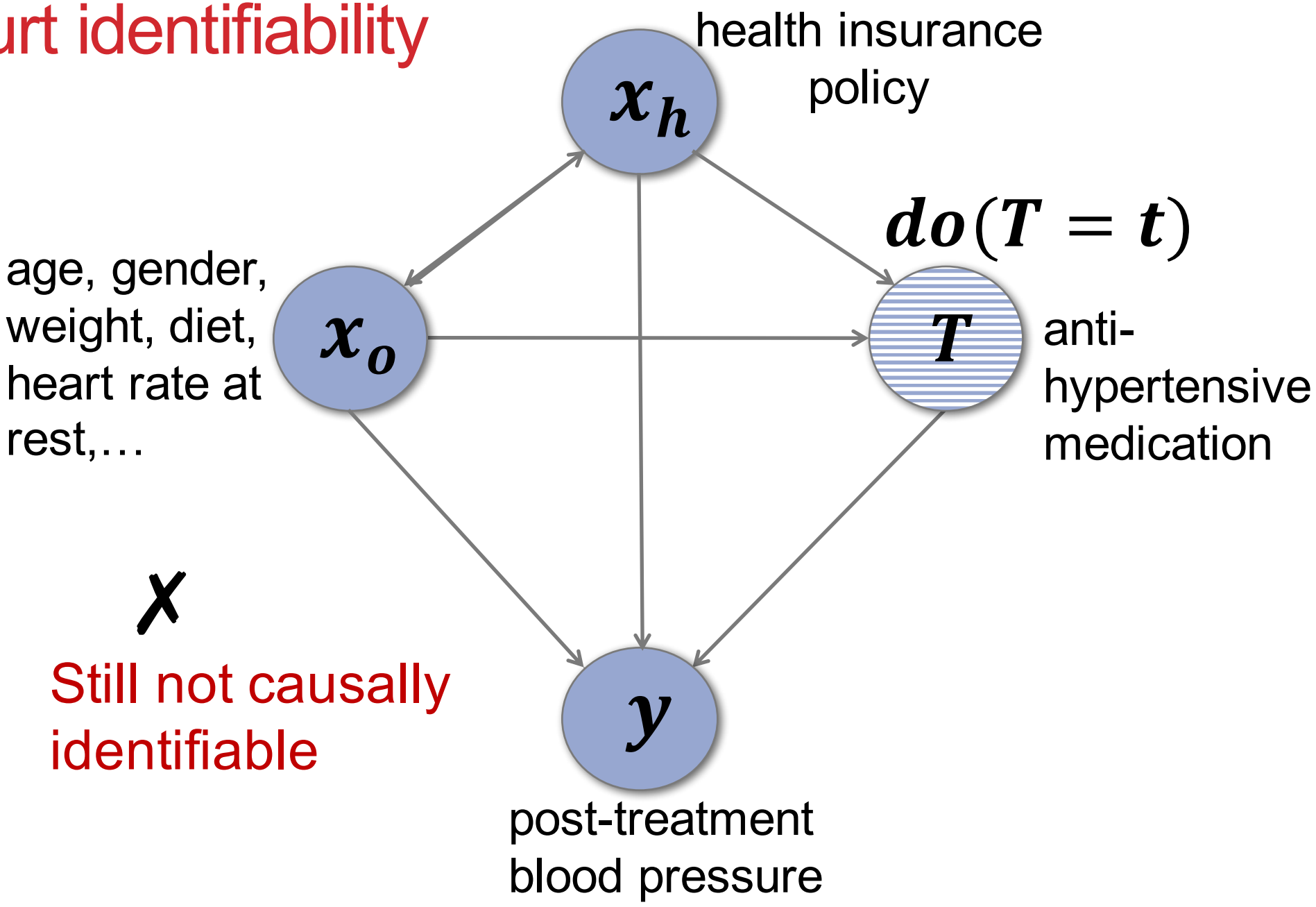




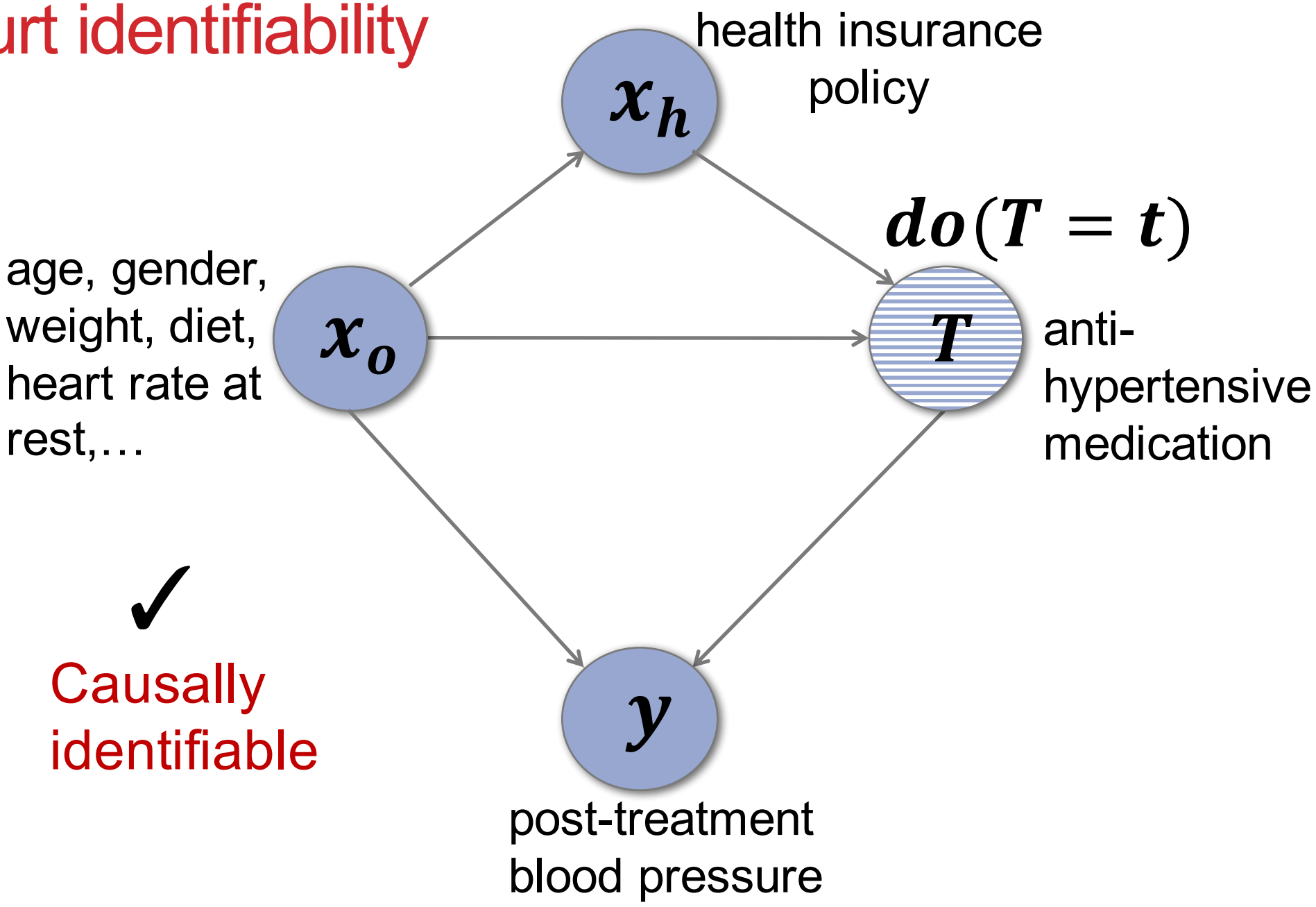
Unidentifiable causal effect



Sometimes hidden confounding does not hurt identifiability



Sometimes hidden confounding does not hurt identifiability



Potential outcomes and causal graphs

- In simple binary treatment with no hidden confounders, the two frameworks are equivalent
- Not mutually exclusive, and unifications exist (Richardson & Robins, 2013)
- Some people feel passionately about which framework to use

Potential outcomes and causal graphs

- Our opinion (subjective!):
  - Potential outcomes are conceptually easier to work with in unconfounded binary treatment studies
  - Potential outcomes closer in spirit to regression problems and domain adaptation

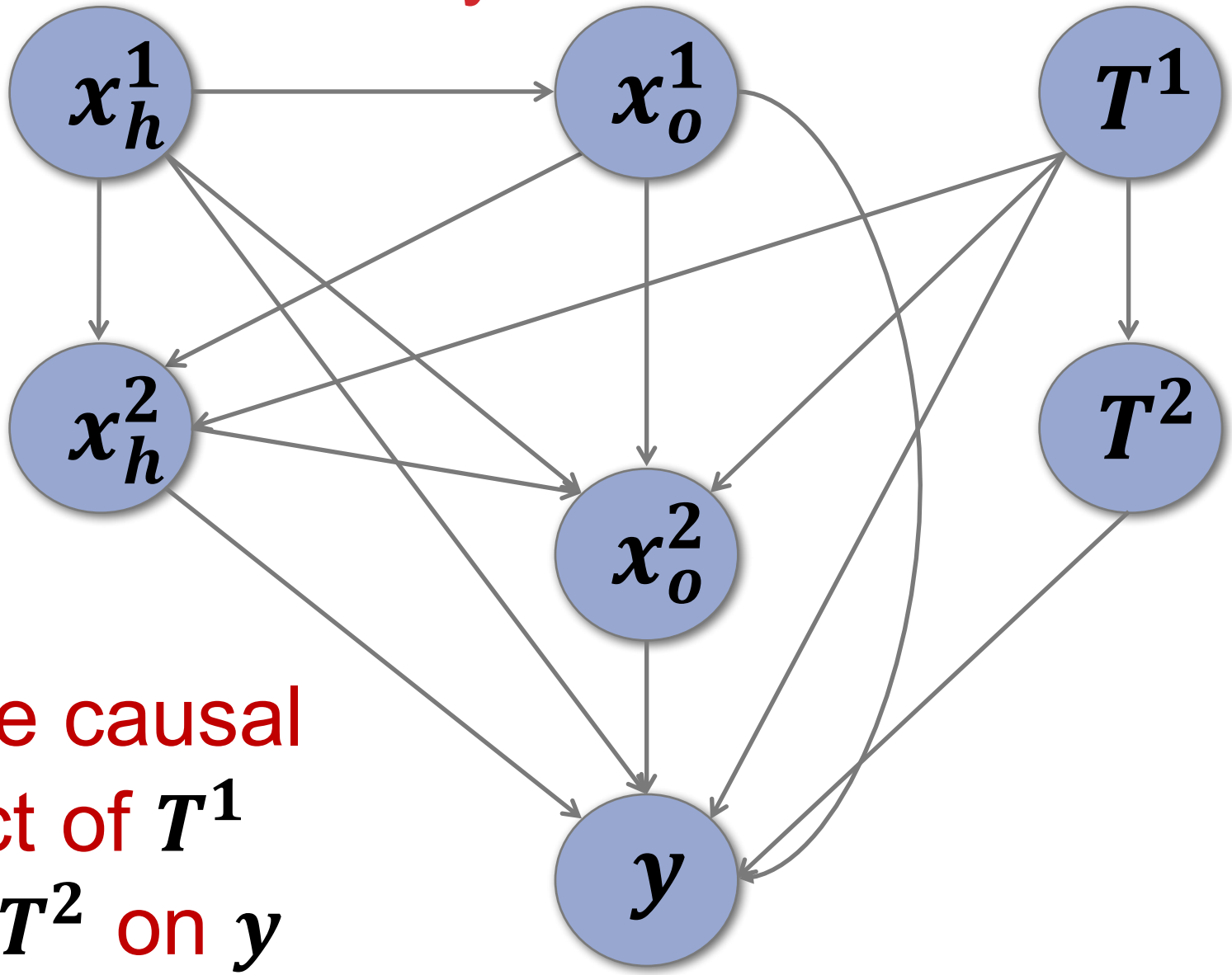
Potential outcomes or causal graphs?

- However, when the causal structure is more complicated, causal graphs are extremely useful
- Using causal graphs can help us determine which covariates we'd like to measure, or which interventions could lead to causal identifiability

Potential outcomes or causal graphs?

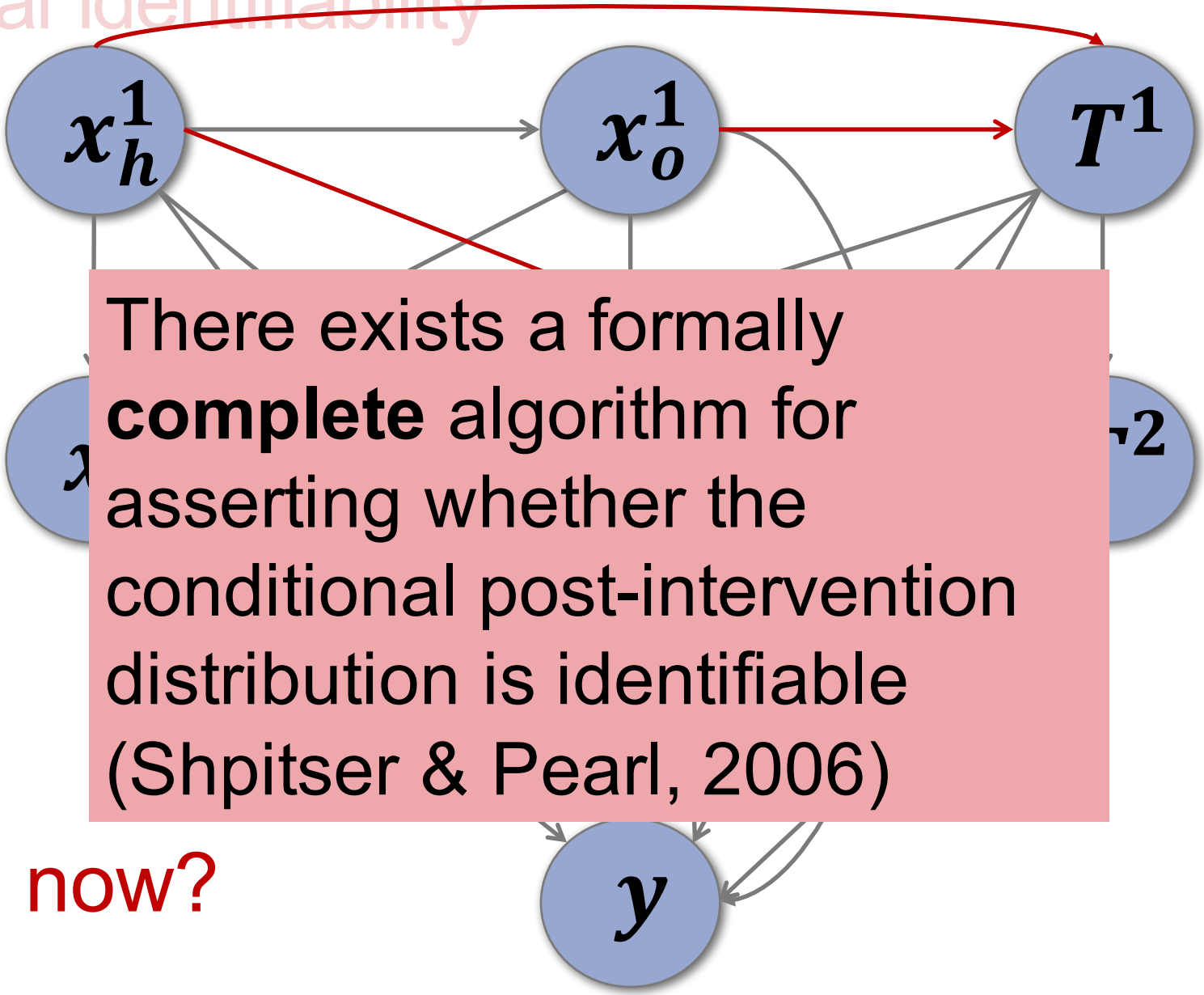
- However, when the causal structure is more complicated, causal graphs are extremely useful
- Example of non-trivial identifiability: ongoing anti-hypertensive treatment, where the doctor changes medication dosage based on previous dosage and changing blood pressure and background variables

Causal identifiability



Is the causal effect of T^1 and T^2 on y identifiable?

Causal identifiability



There exists a formally **complete** algorithm for asserting whether the conditional post-intervention distribution is identifiable (Shpitser & Pearl, 2006)

And now?

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Running an observational study

- Check your assumptions and design:
  - Is there reason to believe no unmeasured confounding holds? Use domain knowledge
  - More generally, do you believe back-door / ignorability holds?
  - If not - change the design:
    - Add more variables
    - Measure treatment differently
    - Measure outcome differently

Checking the assumptions

- Comparing effectiveness of two anti-hypertensive medications
- Treatment: first administration of medication
- Outcome: blood pressure 3 months after first treatment
  - Is outcome only measured for some of the patients?

Checking the assumptions

- Did we measure the important known causes of hypertension? Literature survey
- Example: high alcohol use is known to be a cause of hypertension
- Doctors know this, and might use this information in deciding on treatment
- If we don’t measure alcohol use, it becomes hidden confounder which might bias our conclusions

Checking the assumptions

- Did we include a post-treatment covariate? example: weight measured after treatment
- Measuring the post-treatment weight could explain away some of the causal effect, inducing bias in our estimation of the causal effect
- Conditioning on post-treatment covariates violates ignorability

Running an observational study - overlap

- Check for overlap between treated and control:
  - Reduce dimension and plot populations
  - Check overlap on important univariate and bivariate variables, e.g. age, gender, weight in a medical study

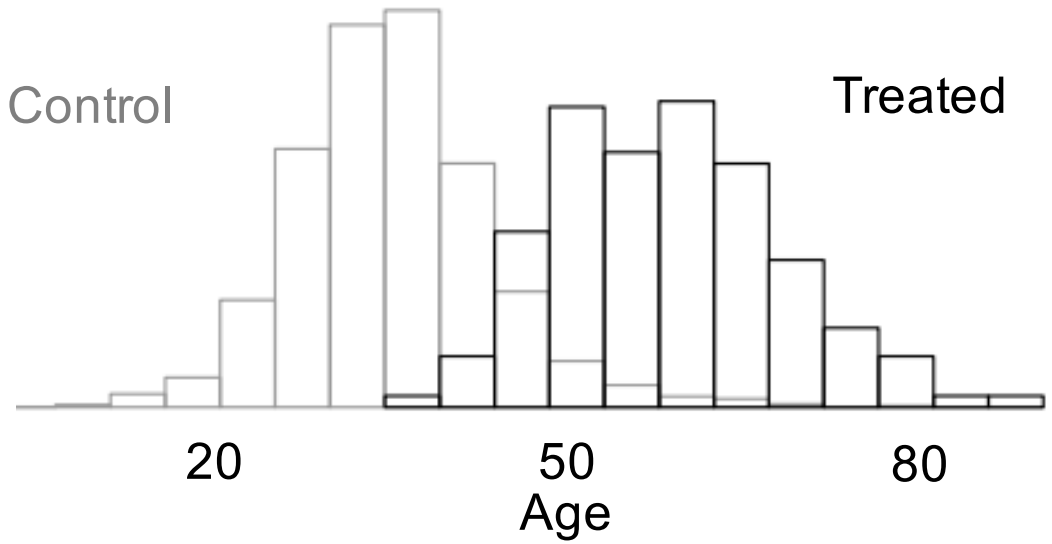


Figure:
Hill & Gelman

Running an observational study - overlap

- If no overlap:
  - Redefine study population, e.g. only people ages 40-60
  - More risky: check if outcome is sensitive to unbalanced variables.
    Example: treated and control might differ in commute distance from hospital, but distance from hospital is not associated with any important socio-economic factors and has no observed association with outcome

Running an observational study – checking the model

- Example: in a hypertension study using covariate adjustment, we identify a sub-population with high error in predicting the factual outcome
- Further examination indicates these are mostly cancer patients
- Solution:
  - Create a more sophisticated model
  - Exclude sub-population

Running an observational study - iterate

- Apply method(s) of choice
- Examine overlap, model predictions
- Change study design and/or model
- Iterate

Running an observational study

- Rosenbaum, Paul R. “Design of observational studies.” (2010)

Which method should I use?

- Unlike standard machine learning, can't compare by cross-validation error!
- Choice depends on domain knowledge and problem attributes:
  - How well can the outcome or treatment assignment model be specified?
  - How many confounders are there?
  - How much overlap is there?

Summary of methods

	ITE?	High-dim?	Model-sensitive?	Interpretable?
Matching	Yes	No	No (sensitive to metric)	Yes
Covariate adjustment (regression)	Yes	Yes	Yes	Only for interpretable regression models
Propensity score	No	No	Yes	Less so
Doubly robust	No	Partial	Less	Less so

Many more ideas and methods

- Natural experiments
- Instrumental variables

Many more ideas and methods – Natural experiments

- Does stress during pregnancy affect later child development?
- Confounding: genetic, mother personality, economic factors…
- Natural experiment: the Cuban missile crisis of October 1962. Many people were afraid a nuclear war is about to break out.
- Compare children who were in utero during the crisis with children from immediately before and after

Many more ideas and methods – Instrumental variables

- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools?
- Confounding: different student population, different teacher population
- Can’t force people which school to go to

Many more ideas and methods – Instrumental variables

- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools?
- Can't force people which school to go to
- Can *randomly* give out vouchers to some children, giving them an opportunity to attend private schools
- The voucher assignment is the instrumental variable

Do observational studies work?

- Success story: comparing 21 observational studies to 17 RCTs about mortality in patients suffering from Acute Coronary Syndrome (Dahabreh et al., 2012)
  Similar studies in healthcare by Kitsios et al. 2015, Anglemeyer et al. 2014
- Failure story: 12 ad-exposure RCTs at Facebook, compared with observational studies on the same population with very rich datasets and large samples. Found high variance in performance of various observational methods
  (Gordon et al., 2016)

Do observational studies work?

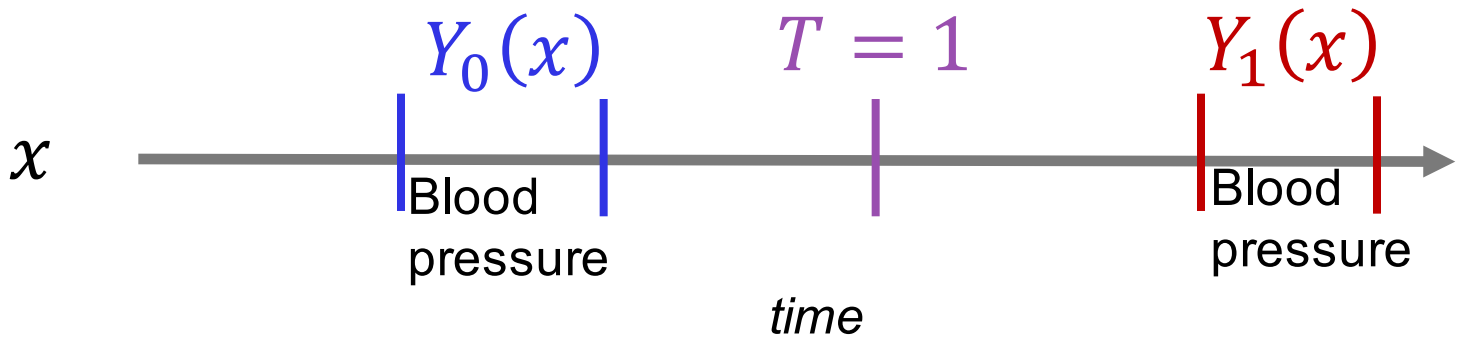
- The Observational Medical Outcomes Partnership (OMOP), Ryan et al. (2012, 2013)
- Identifying acute rare adverse side-effects of drugs from health claims data: hospital and medical insurance records

- 399 known treatment-outcome associations, either negative or positive
  - Positive: antibiotics and acute liver injury
  - Negative: Beta-blockers and gastro-intestinal bleeding

- Compared study design and inference methods

Do observational studies work?

- Choice of study design outweighed specific methods choices
- Best results obtained using a *self-control design*: look at only treated units, and compare before and after treatment:



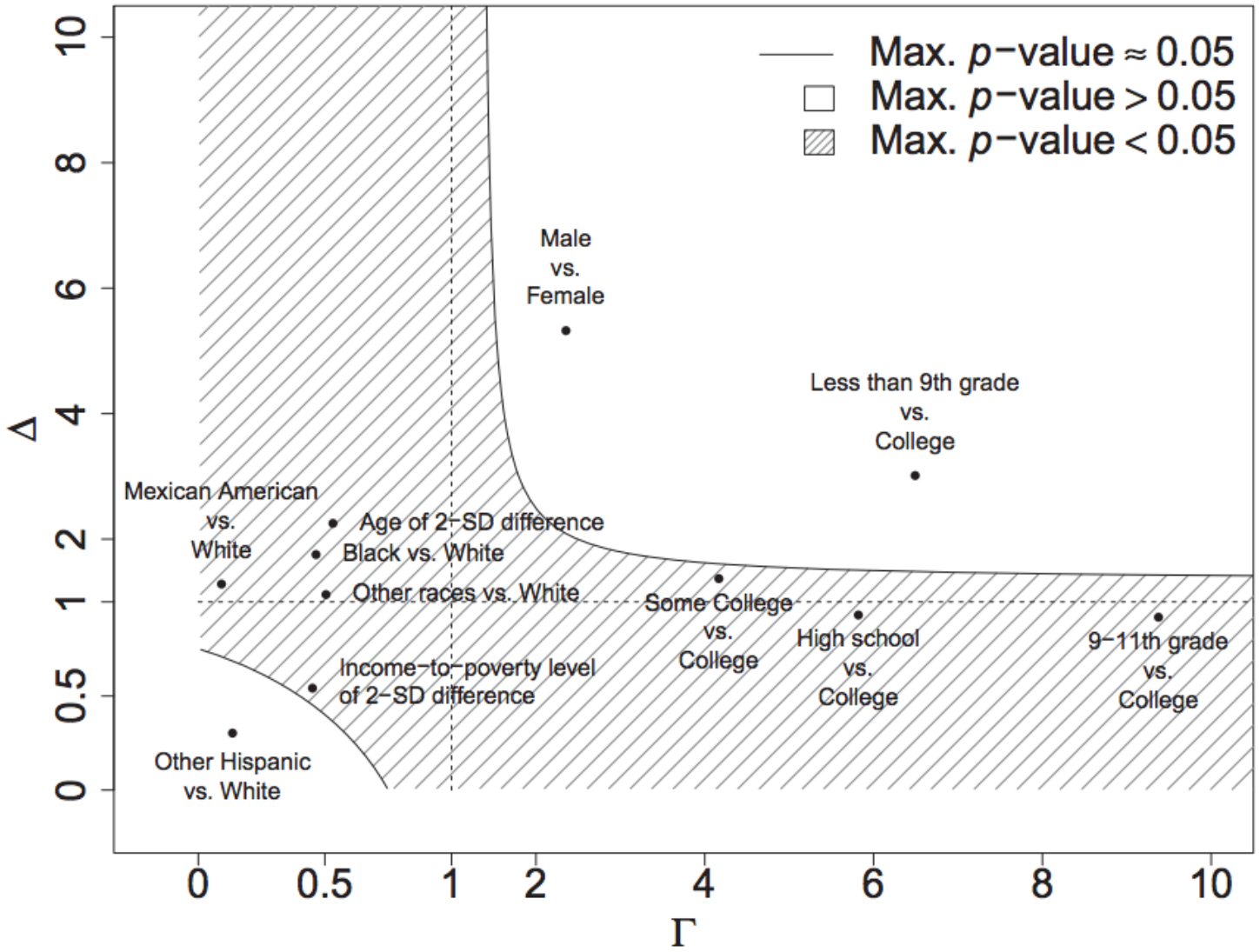
- Achieved AUCs of 0.76-0.94 over the 399 positive or negative treatment-outcome associations

Sensitivity analysis

- How much unmeasured confounding to flip our conclusions?

Does cigarette smoking increase blood lead?

Unmeasured confounding in outcome model $h \rightarrow y$



Hsu & Small, 2013

Unmeasured confounding in treatment assignment $h \rightarrow t$

Atlantic Causal Inference Conference: Causal inference challenge 2016

- 77 causal inference datasets with varying degrees of overlap, non-linearity, and others aspects
- No hidden confounding
- Real data, simulated outcomes and treatment assignments

Atlantic Causal Inference Conference: Causal inference challenge 2016

- Data generating process still confidential! But similar in spirit to Hill (2011):
- Infant Health and Development Program (**IHDP**)
  - Randomized trial following infants and their development with or without intense child-care assistance.
- 25 covariates such as birth weight, mother's education, prenatal care, race etc.
- Randomized → observational by removing treated children with non-white moms from the study
- $Y_0(x) \sim N(\beta^T(x + \vec{a}), 1), \quad Y_1(x) \sim N(\beta^T x + b, 1)$ with a sparse random β and predefined $\vec{a}$ and b

Atlantic Causal Inference Conference: Causal inference challenge 2016

- Varying degrees of non-linearity, sparsity, correlation between treatment assignment and outcome, non-linearity of treatment effect, overlap
- Challenge to researchers: http://jenniferhill7.wix.com/acic-2016#!competition/ttela
- Still open to submissions!

Atlantic Causal Inference Conference: Causal inference challenge 2016

Results so far

Covariate adjustment methods perform well:

1. Bayesian Additive Regression Trees (BART)
2. Random forests + kernel ridge regression

Superlearner, a doubly-robust ensemble regression method also performs very well

ACIC competition

- No hidden confounding: sufficiently flexible covariate adjustment is expected to work
- Most submitted methods incorporate covariate adjustment
- Using propensity score or matching does not seem to convey any particular advantage

Comparison studies: broad picture

ACIC competition (2016), Hill (2011), Kang & Schafer, (2007), Todd & Smith (2005)

- Comparison studies for the case of no hidden confounders show flexible regression methods perform very well
- Real world study on job training dataset (Todd & Smith 2005), with embedded randomized trial, has matching methods as best performing
- Propensity score alone: underperforming
- Doubly robust: vanilla version underperforming, more sophisticated methods might see some gains

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Causal inference methods in ML

- Bandits and reinforcement learning
- Recent work on ML and observational studies

Outline

Causal inference methods in ML

- Bandits and reinforcement learning
- Recent work on ML and observational studies

Bandit problems, reinforcement learning, and causal inference

- Who controls the treatment assignment?
- How big is the “action” space?
- How is the information acquired?
- When is the decision needed?

Who controls the actions?

	Actions determined by mechanism designer	**Actions mostly determined externally**
Domains	Robotics, ad-placement	Medicine, education
How is the problem called?	Off-policy evalution, learning from logged bandit feedback (Li et al. 2011, Bottou et al. 2013)	Observational studies (Rosenbaum 2010)
Data acquisition	Exploration, often sequential (online)	(Quasi)-randomized experiments, natural experiments
Research	Computer scientists	Statisticians, economists

Who controls the actions?

	Actions determined by mechanism designer	**Actions mostly determined externally**
Domains	Robotics, ad-placement	Medicine, education
How is the problem called?	Off-policy evalution, learning from logged bandit feedback (Li et al. 2011, Bottou et al. 2013)	Observational studies (Rosenbaum 2010)
Data acquisition	Exploration, often sequential (online)	(Quasi)-randomized experiments, natural experiments
Research	Computer scientists	Statisticians, economists

Bandit problems and potential outcomes

- Bandit feedback / off-policy
- Instance/context x
- Action space $a \in \mathcal{A}$, where $\mathcal{A}$ is usually discrete but often large
- Reward vector $r(x) = \left(r_1(x), \ldots, r_{|\mathcal{A}|}(x)\right)$
- Choose action a using algorithm, observe feedback $r_a(x)$

- Potential outcomes
- Confounder x
- Treatment T, often binary
- Potential outcomes $(Y_0(x), Y_1(x))$
- "Someone" chose treatment $T = t$, observe outcome $Y_t(x)$

Relevant theoretical results

- Low-variance inverse propensity score weighting methods for large action spaces, with known propensity scores (Swaminathan & Joachims, 2015)
- Bias and variance of doubly robust estimators for large action spaces (Langford et al., 2011)

Outline

Causal inference methods in ML

- Bandits and reinforcement learning
- Recent work on ML and observational studies

Outline

Causal inference methods in ML

Bandits and reinforcement learning

Recent work on ML and observational studies

Recent advances in ML for causal inference in observational studies

- Driven by big data and need to make individual predictions, e.g. precision medicine
- See reference list for many papers not mentioned here

Recent advances in ML for causal inference in observational studies

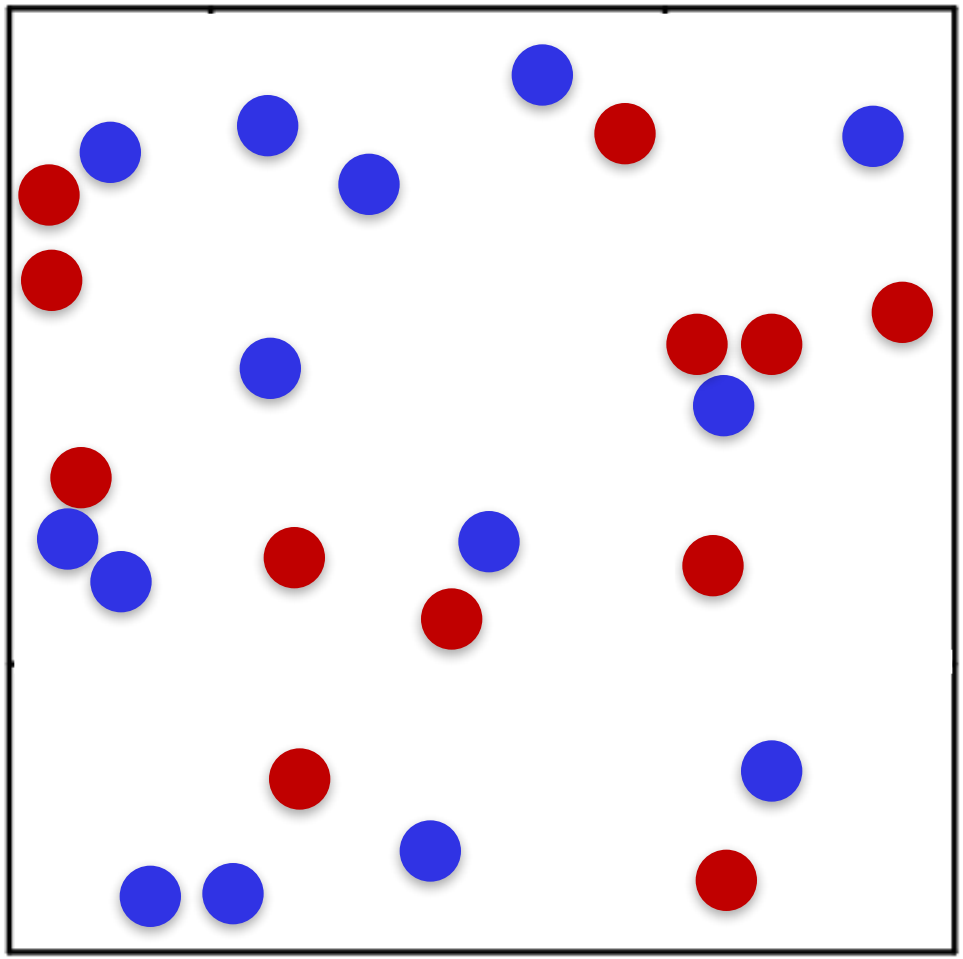
- **Causal trees**
  Athey & Imbens. "*Recursive Partitioning for Heterogeneous Causal Effects*." *arXiv:1504.01132* (2015)
- **Causal forest**
  Wager & Athey. "*Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*." *arXiv:1510.04342* (2015)
- **Sparse causal inference**
  Athey, Imbens & Wager. "*Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing*." *arXiv:1604.07125* (2016).

Causal forests
(Wager & Athey, 2015)

- Random forest adapted for causal inference
- Proven to be consistent estimators of average treatment effect (!)

Causal forests

(Wager & Athey, 2015)



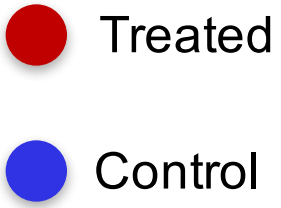
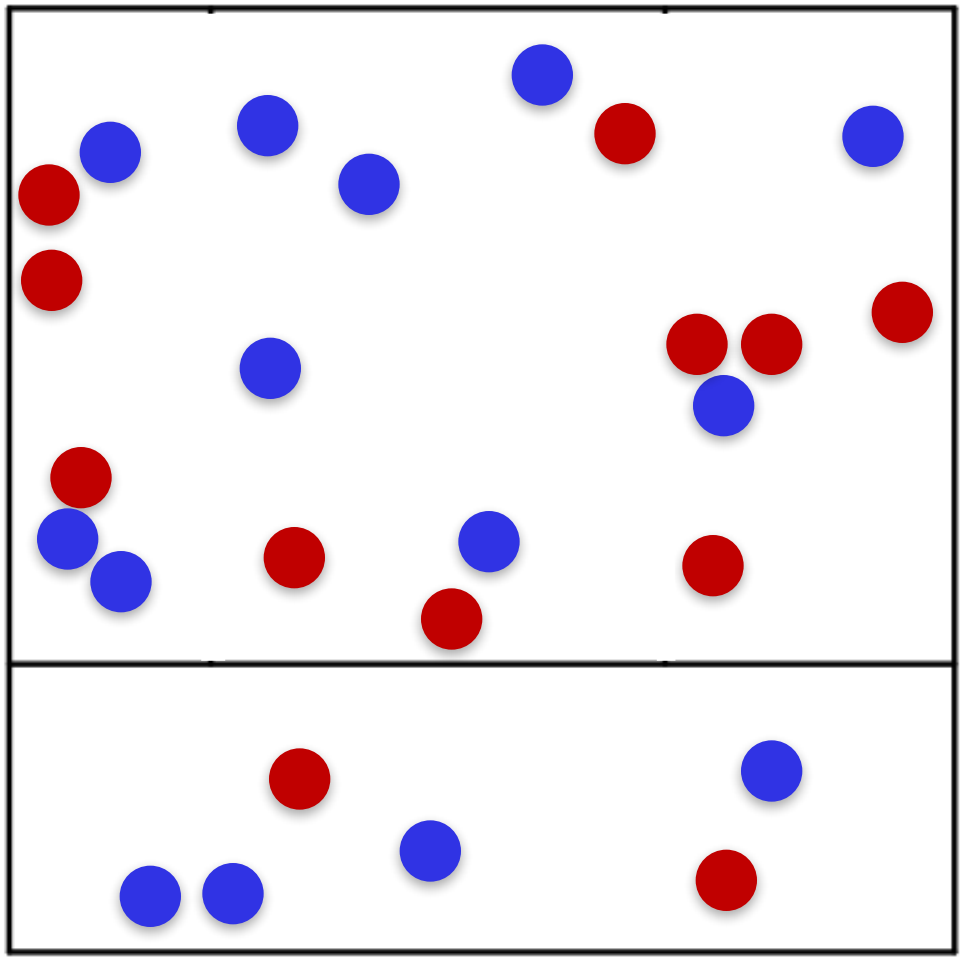


Figure:
Stefan Wager

Causal forests

(Wager & Athey, 2015)



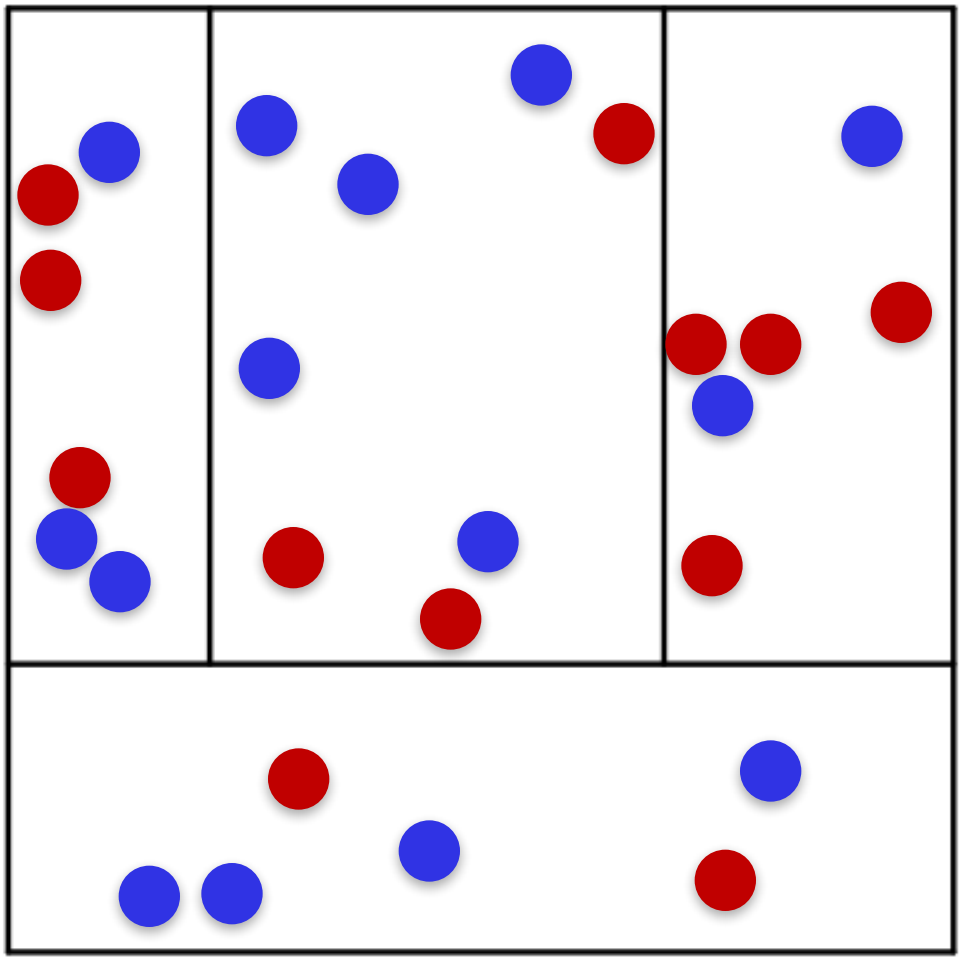
Treated

Control

Figure:
Stefan Wager

Causal forests

(Wager & Athey, 2015)



Treated

Control

Figure:
Stefan Wager

Causal forests

(Wager & Athey, 2015)

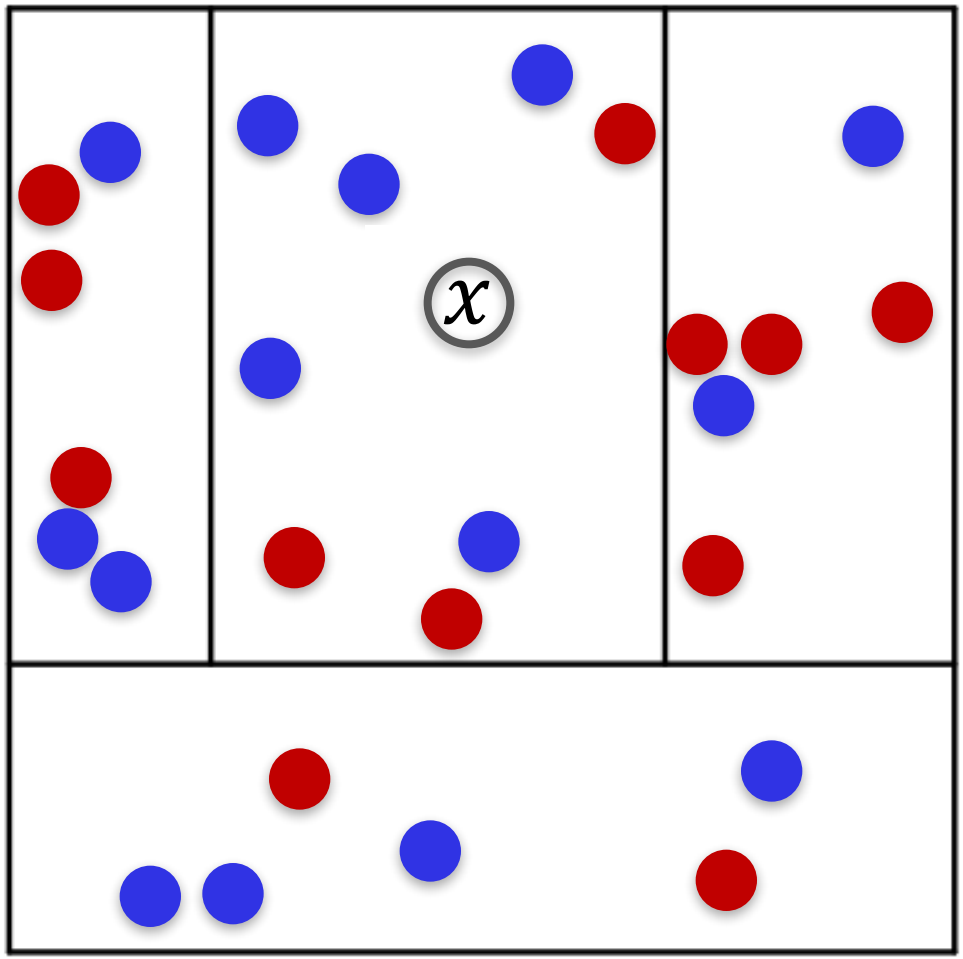
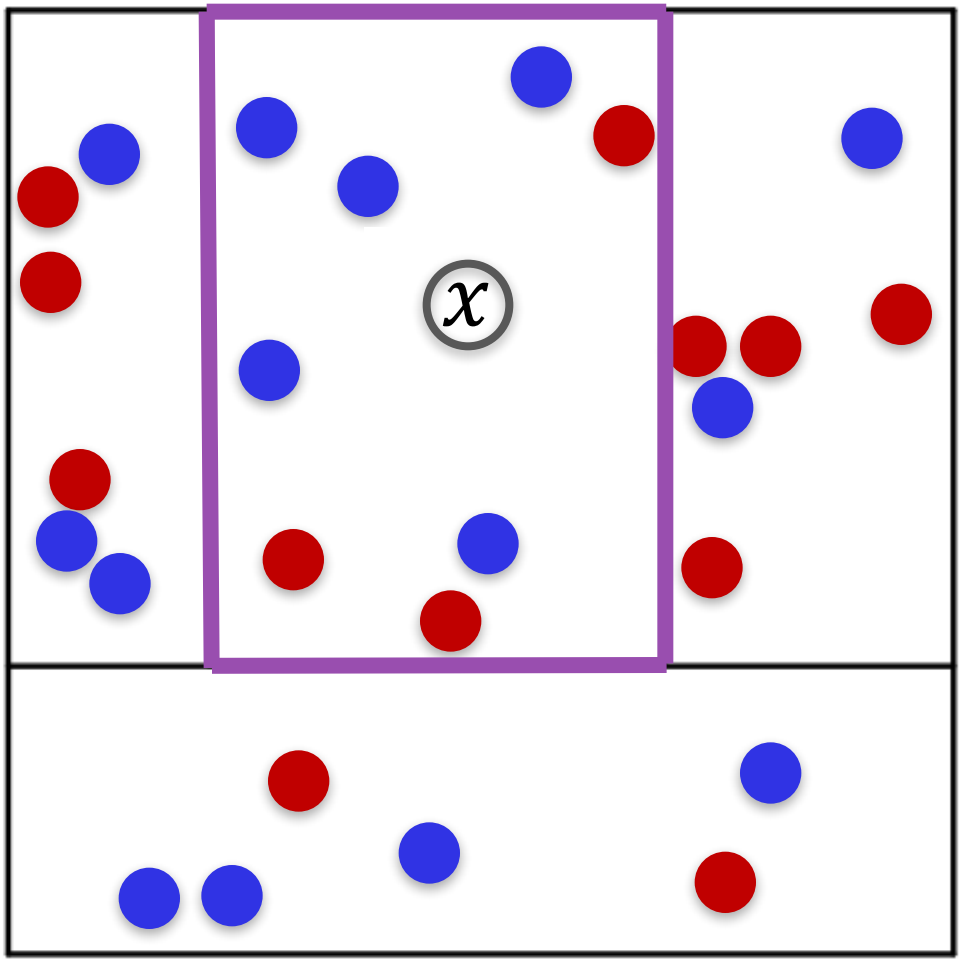


Figure:
Stefan Wager

Causal forests

$$\hat{ITE}(x) = \frac{1}{\#\text{treat in leaf}(x)} \sum_{\substack{i \in \text{leaf}(x) \\ t_i = 1}} y_i - \frac{1}{\#\text{control in leaf}(x)} \sum_{\substack{i \in \text{leaf}(x) \\ t_i = 0}} y_i$$



Treated

Control

Figure:
Stefan Wager

Causal forests

(Wager & Athey, 2015)

- Divide training sample in two:
  One half used for learning tree structure
  Other half used for estimating ITE

- Split rule:
  Maximize variance of $\widehat{ITE}$, instead of variance of outcome y as done in ordinary regression trees

Recent advances in ML for causal inference in observational studies

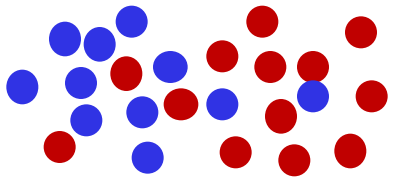
- **Representation learning**

Johansson, Shalit & Sontag. "*Learning Representations for Counterfactual Inference*." ICML 2016

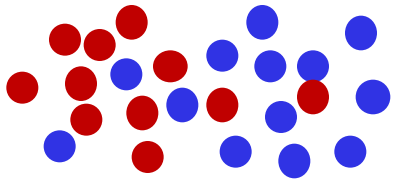
Shalit, Johansson & Sontag. "*Bounding and Minimizing Counterfactual Error*." *arXiv:1606.03976*

Causal inference as domain adaptation

Factual = Source domain



Counterfactual= Target domain



Treated

Control

$p_F(x,t) = p_F(x)p_F(t|x)$

the joint *factual* distribution over covariates and treatment assignment

labeled y_i

$p_{CF}(x,t) \coloneqq p_F(x)p_F(1-t|x)$

the joint *counterfactual* distribution over covariates and treatment assignment

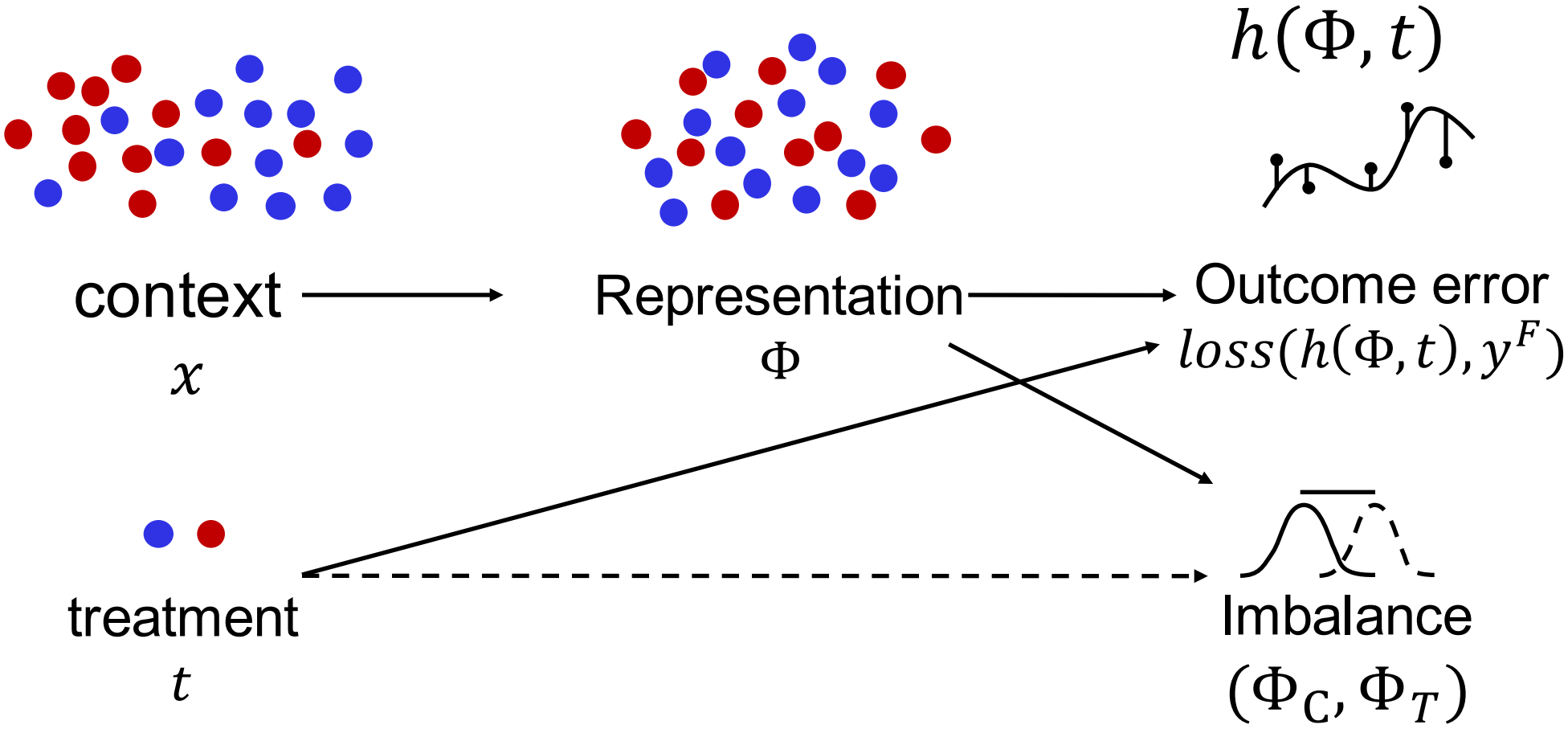
unlabeled

Causal inference as domain adaptation

- Causal inference as prediction on the counterfactual set
- Learn a representation Φ that minimizes a weighted sum of:
  - prediction loss while using Φ
  - A measure of distance between treated and control distributions as represented in Φ

Learning balanced representation

(Johansson, Shalit & Sontag, 2016)



Causal inference as domain adaptation

(Shalit, Johansson & Sontag, 2016)

- Theorem:

$$\mathbb{E}\left[\left(\hat{\mathrm{ITE}}(\Phi, h) - \mathrm{ITE}\right)^2\right] \leq$$

$$\mathbb{E}\left[\left(h\left(\Phi(x), t\right) - Y_t(x)\right)^2\right] + K_\Phi \cdot \mathrm{IPM}\left(p_\Phi^{treated}, p_\Phi^{control}\right)$$

- $\widehat{ITE}(x_i) = h(\Phi(x), 1) - h(\Phi(x), 0)$
- IPM: integral probability metric, e.g. *Wasserstein* distance or *Maximum Mean Discrepancy (MMD)*
- K_Φ: depends on condition number of Jacobian of Φ

Causal inference as domain adaptation

(Shalit, Johansson & Sontag, 2016)

- Theorem:

$$\mathbb{E}\left[\left(\hat{\mathrm{ITE}}(\Phi, h) - \mathrm{ITE}\right)^2\right] \leq$$

$$\mathbb{E}\left[\left(h\left(\Phi(x), t\right) - Y_t(x)\right)^2\right] + K_\Phi \cdot \mathrm{IPM}\left(p_\Phi^{treated}, p_\Phi^{control}\right)$$

- We upper bound the expected error in estimating ITE
- The upper bound has two terms:
  - The ordinary machine learning expected test loss
  - A distance measure between the treated and control distributions induced by Φ

Causal inference as domain adaptation

(Shalit, Johansson & Sontag, 2016)

- Theorem: for representation Φ and hypothesis h

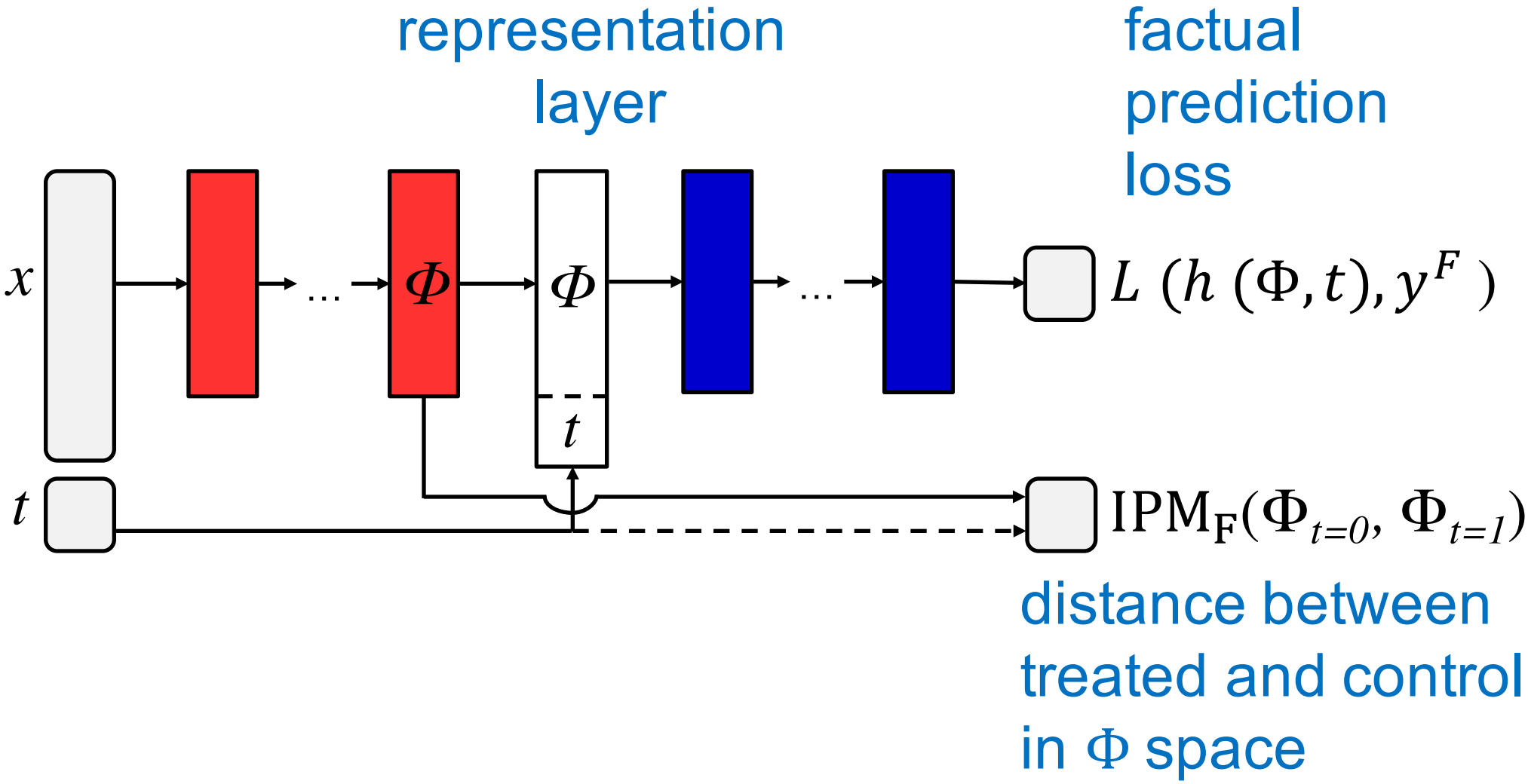
$$\mathbb{E}\left[\left(\hat{\mathrm{ITE}}(\Phi, h) - \mathrm{ITE}\right)^2\right] \leq$$

$$\mathbb{E}\left[\left(h\left(\Phi(x), t\right) - Y_t(x)\right)^2\right] + K_\Phi \cdot \mathrm{IPM}\left(p_\Phi^{treated}, p_\Phi^{control}\right)$$

- The upper bound has two terms:
  - The ordinary machine learning expected test loss
  - A distance measure between the treated and control distributions induced by Φ
- We can minimize the upper bound over Φ and h, using either the Wasserstein or MMD distance

Learning balanced representations

(Johansson, Shalit & Sontag, 2016)



Experimental results – IHDP dataset

- IHDP: real confounders and treatment, simulated outcomes (Hill, 2011)
- CFR-2-2: our model, with 2 layers before Φ and 2 after Φ

Method	ATE error	ITE error
Lasso+Ridge regression	0.2	2.8 ± 0.1
BART (Chipman et al.)	0.2	2.1 ± 0.2
Causal forests (Wager et al.)	0.2	2.2 ± 0.1
CFR-2-2 no IPM	0.3	1.8 ± 0.0
CFR-2-2 MMD	0.3	1.7 ± 0.1
CFR-2-2 Wass	**0.2**	**1.6 ± 0.0**

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Outline

Introduction

Counterfactuals and potential outcomes

Tools of the trade

BREAK

Mathematical foundations: causal graphs

Practical lessons

Causal inference methods in ML

Conclusion

Machine learning and causal inference for observational studies

- Ideas we are very familiar with are extremely relevant
- Covariate adjustment is a function approximation problem, estimate $\mathbb{E}[Y_t|x,t]$
- Propensity score, estimate $p(T=t|x)$
- Strong connections to domain adaptation, reinforcement learning and bandit problems

Machine learning and causal inference for observational studies: open problems

- What are the best ways to use the treatment variable in regression models?
- How can we do model selection without cross-validation?
- How to best use the propensity score?
- What kind of theoretical statements can we make beyond consistency?

Where ML should head

- Policy question in healthcare, education, social policy and more require people who know how to handle big, complicated, real-world data
- These policy questions often hinge on causality

WE WANT YOU! TO THINK ABOUT CAUSAL INFERENCE

Acknowledgments

- Fredrik Johansson (U. Chalmers)
- Dylan Small (UPenn)
- Thomas Richardson (U. Washington)
- Aaron Baum (Mt. Sinai)
- Ilya Shpitser (Johns Hopkins)
- NYU Causal inference support group:
  - Jennifer Hill
  - Vince Dorie
  - Marc Scott
  - Dan Cervone

Thank you
and questions