Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 11: Causal Discovery

March 4, 2025

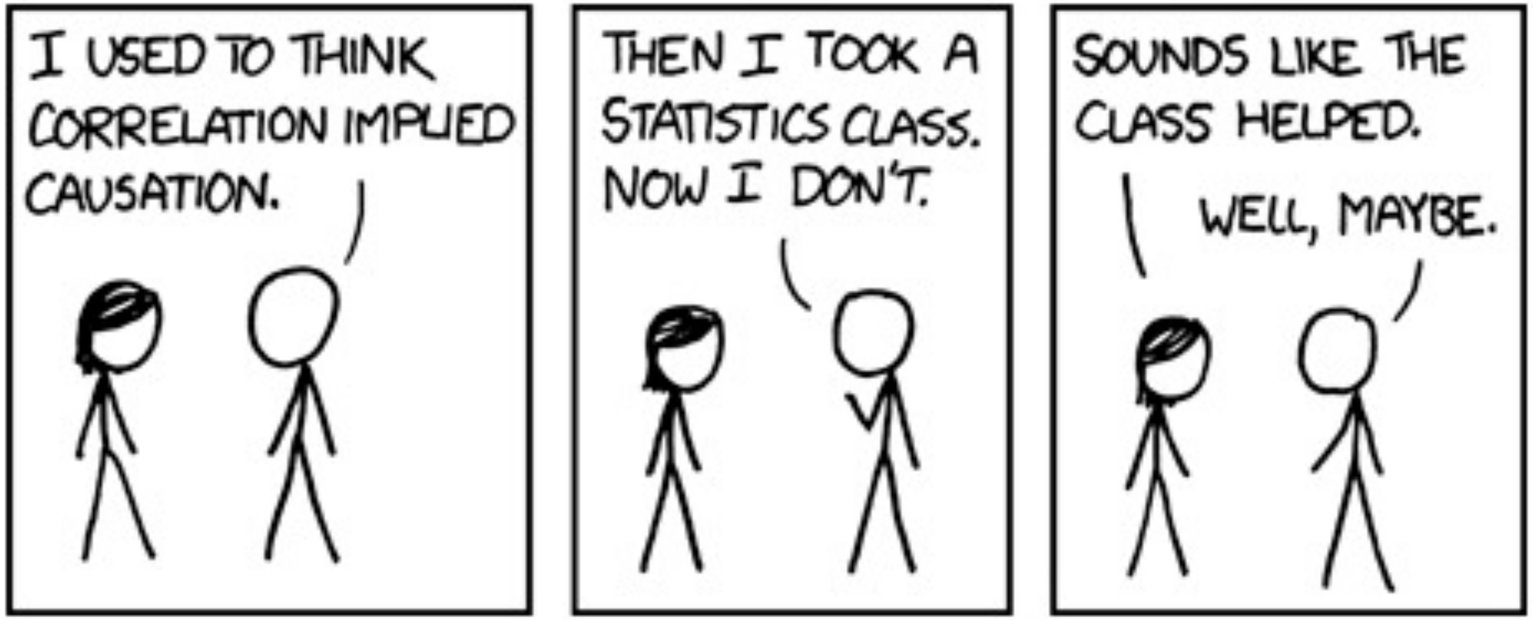
Reading: See course homepage

Today

- Causal Thinking
- Identification of causal effects
- Causal Discovery
- Causality in Practice

Causal Thinking

Association vs. Dependence


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

(http://imgs.xkcd.com/comics/correlation.png)

X and Y are associated iff

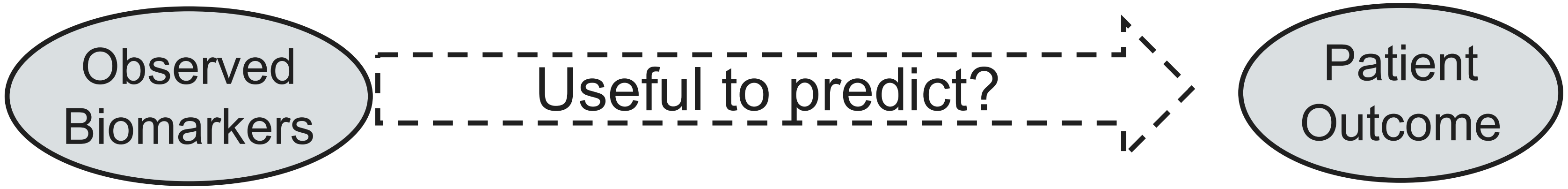
$\exists x_1 \neq x_2 \; P(Y|X{=}x_1) \neq P(Y|X{=}x_2)$

X is a cause of Y iff

$\exists x_1 \neq x_2 \; P(Y|do\,(X{=}x_1)) \neq P(Y|do\,(X{=}x_2))$

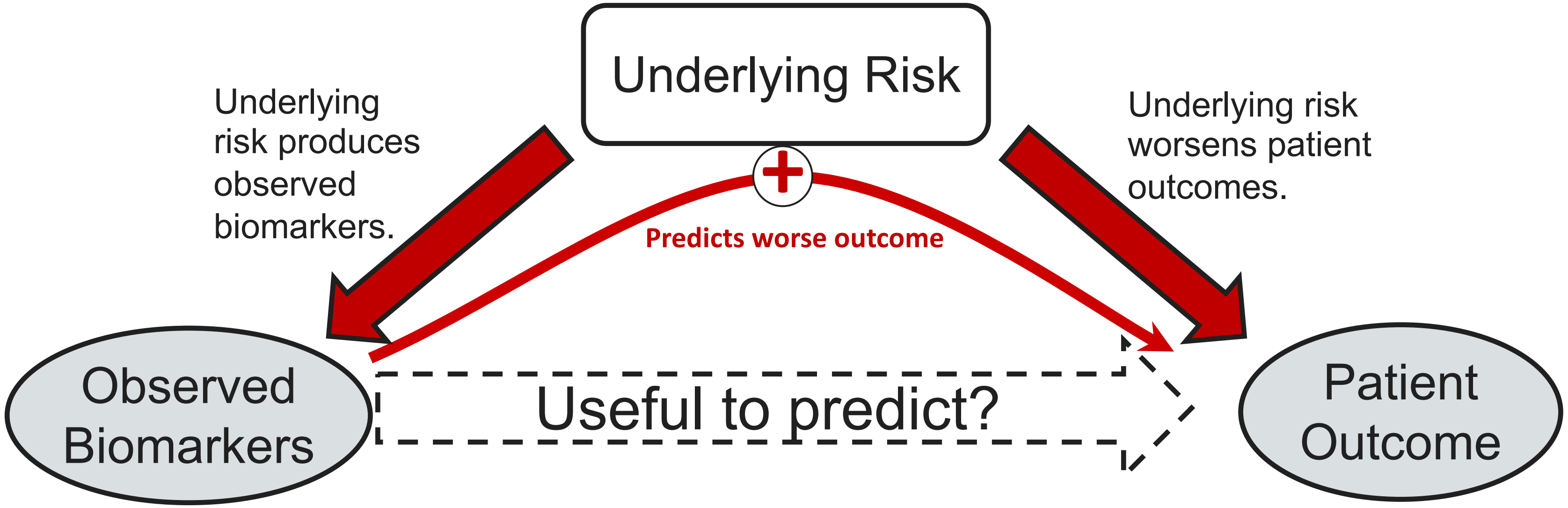
Example 1 of Causal Thinking: Learning from Medical Data

- Can we learn causal effects from real-world observations?



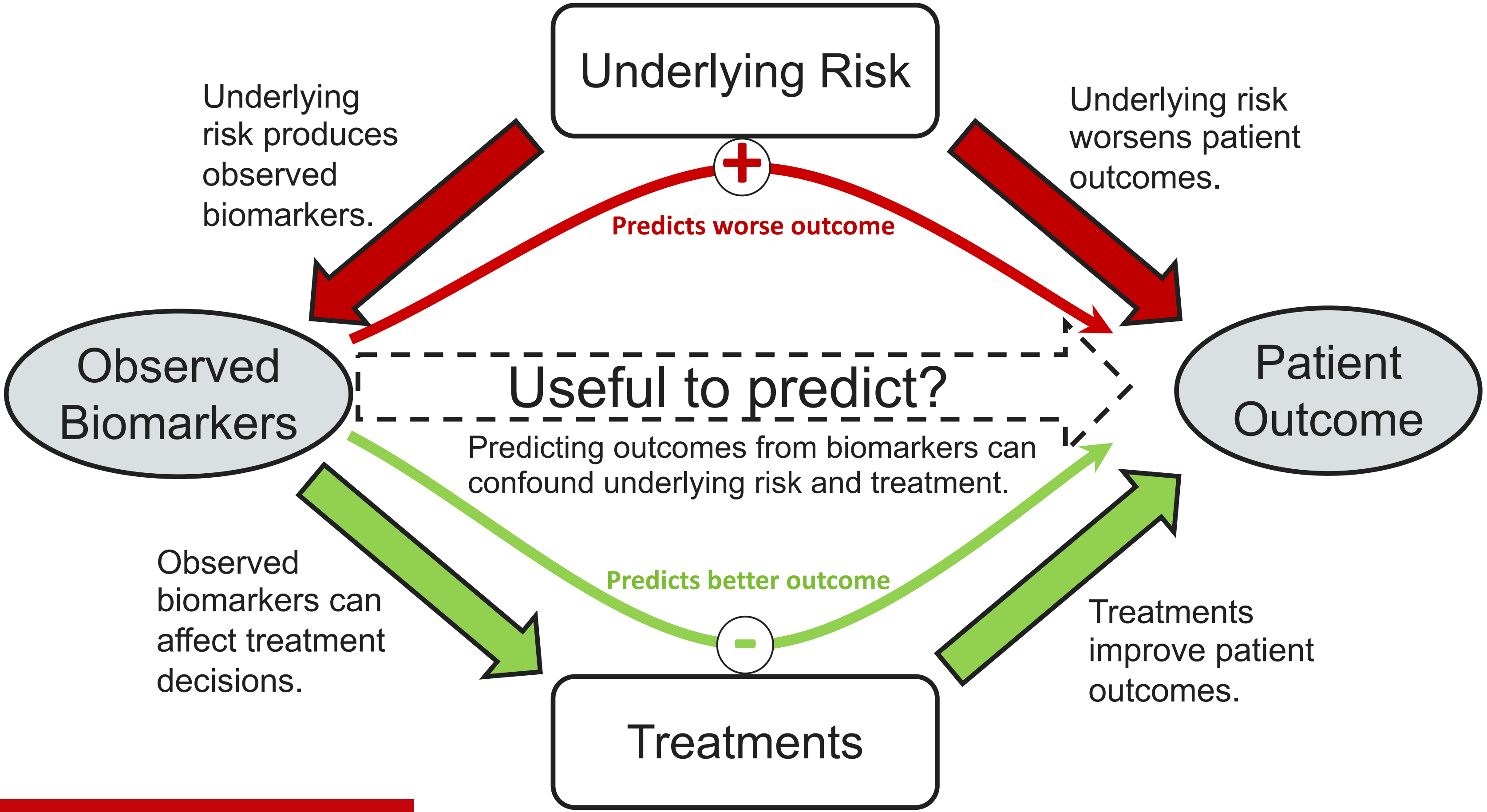
Example 1 of Causal Thinking: Learning from Medical Data

- Can we learn causal effects from real-world observations?



Example 1 of Causal Thinking: Learning from Medical Data

- Can we learn causal effects from real-world observations?



Underlying Risk

Underlying risk produces observed biomarkers.

Underlying risk worsens patient outcomes.

+

Predicts worse outcome

Observed Biomarkers

Useful to predict?

Patient Outcome

Predicting outcomes from biomarkers can confound underlying risk and treatment.

Observed biomarkers can affect treatment decisions.

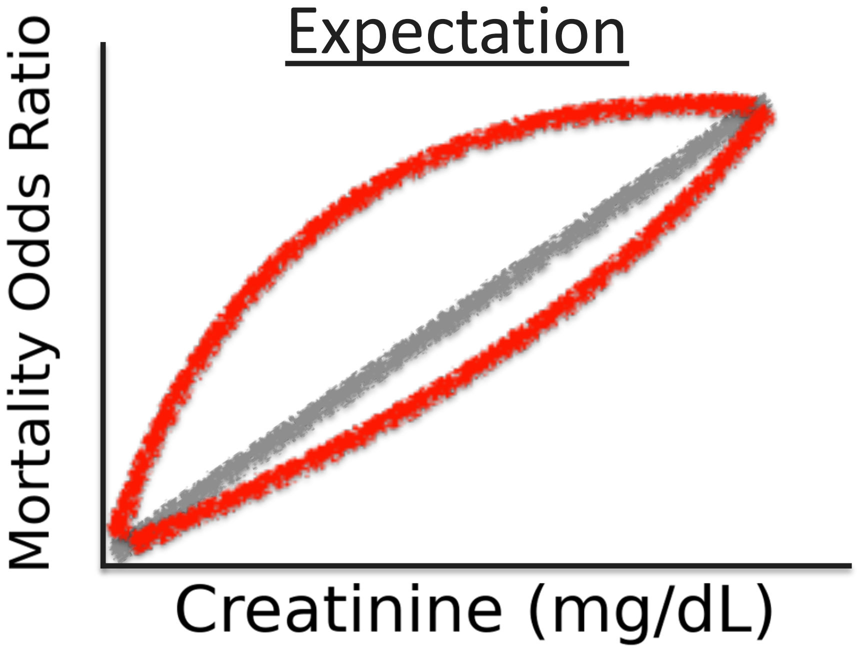
Predicts better outcome

-

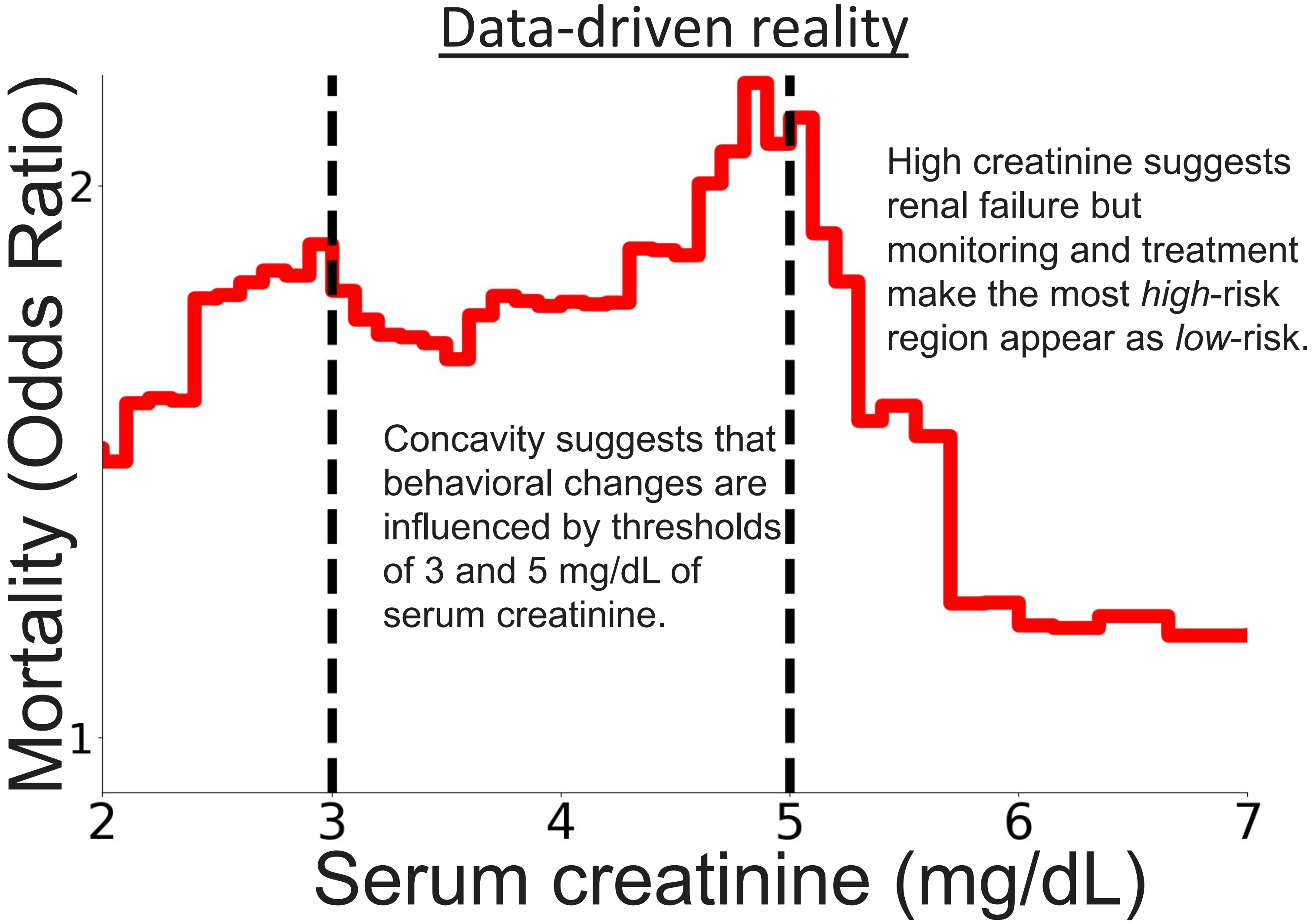
Treatments improve patient outcomes.

Treatments

Example 1 of Causal Thinking: Learning from Medical Data



Elevated creatinine levels are an indicator of renal failure, so we may expect mortality risk to **increase** with creatinine.



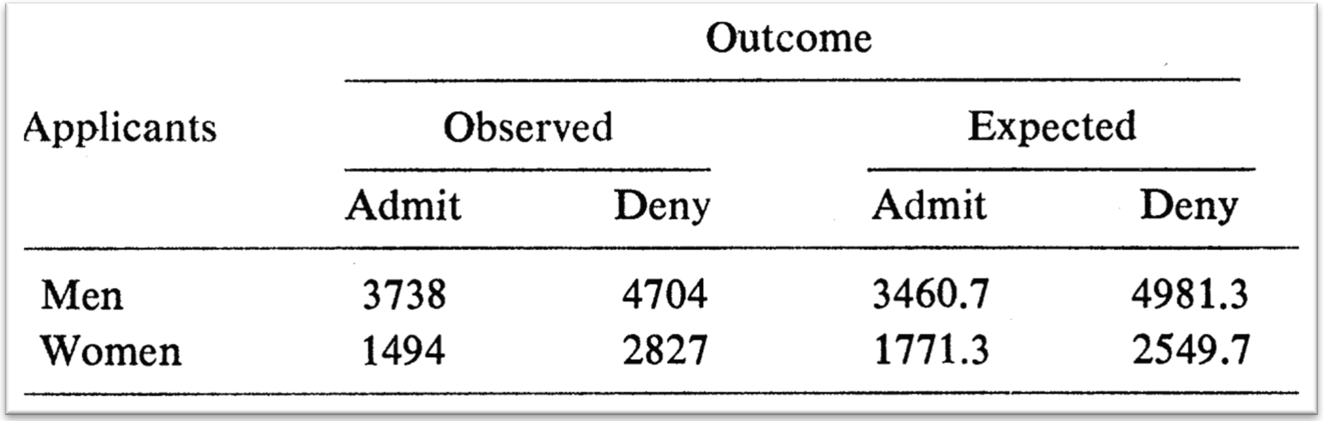
Concavity suggests that behavioral changes are influenced by thresholds of 3 and 5 mg/dL of serum creatinine.

High creatinine suggests renal failure but monitoring and treatment make the most *high*-risk region appear as *low*-risk.

Lengerich et al 2022

Example 2 of Causal Thinking: Simpson’s Paradox

- Graduate admissions at UC Berkeley in 1973



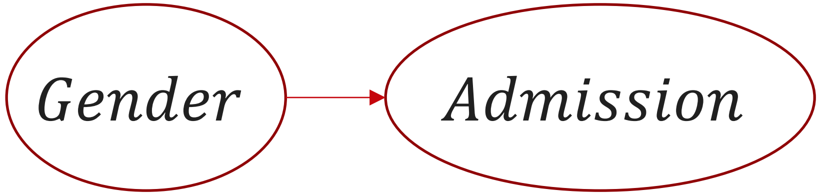
Applicants	Outcome			
	Observed		Expected	
	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3
Women	1494	2827	1771.3	2549.7

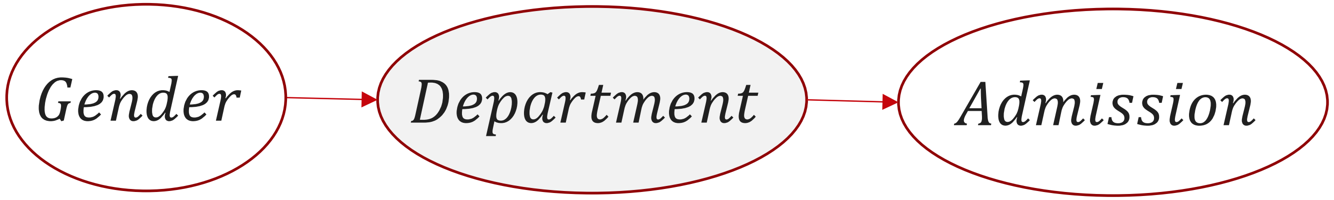
[Bickel]

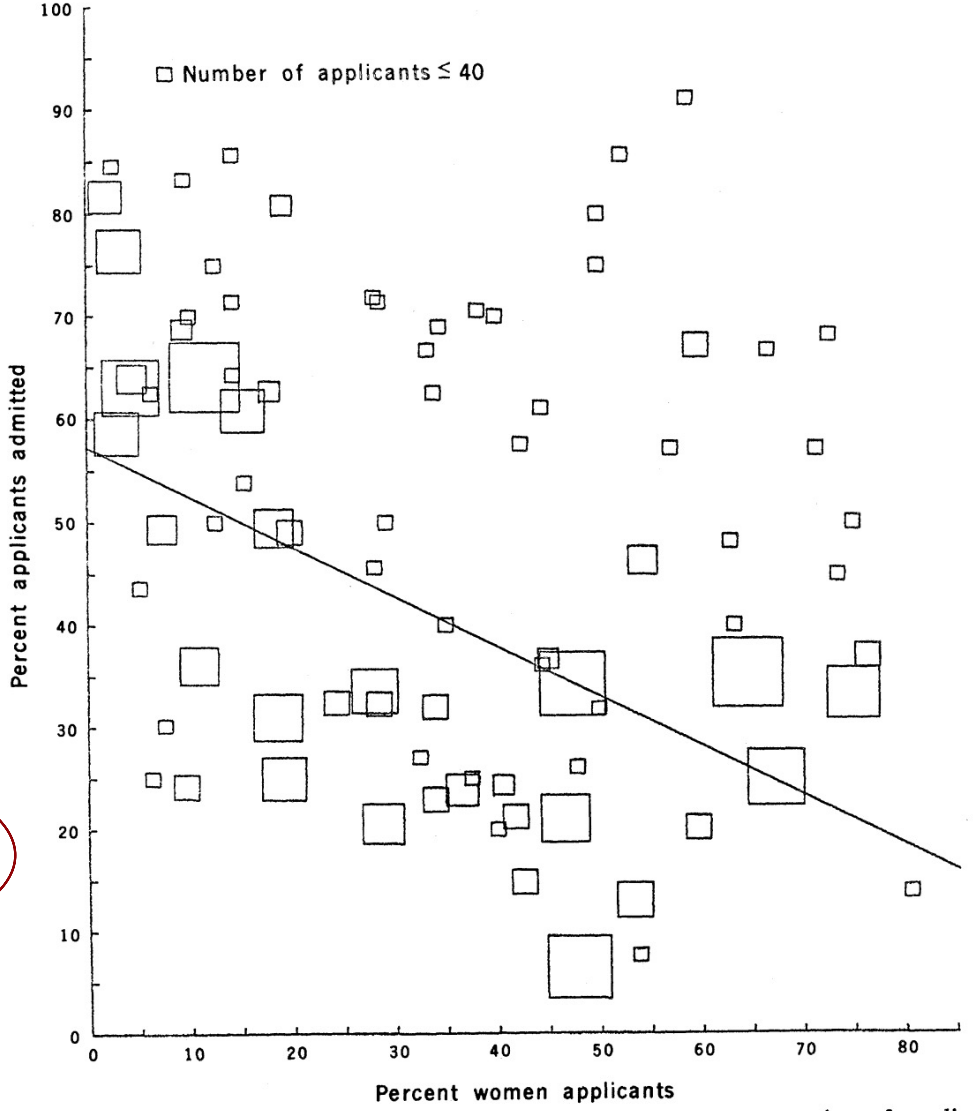
Gender bias?

Example 2 of Causal Thinking: Simpson’s Paradox

- More women were applying to departments with lower admissions rates:


Gender
Admission

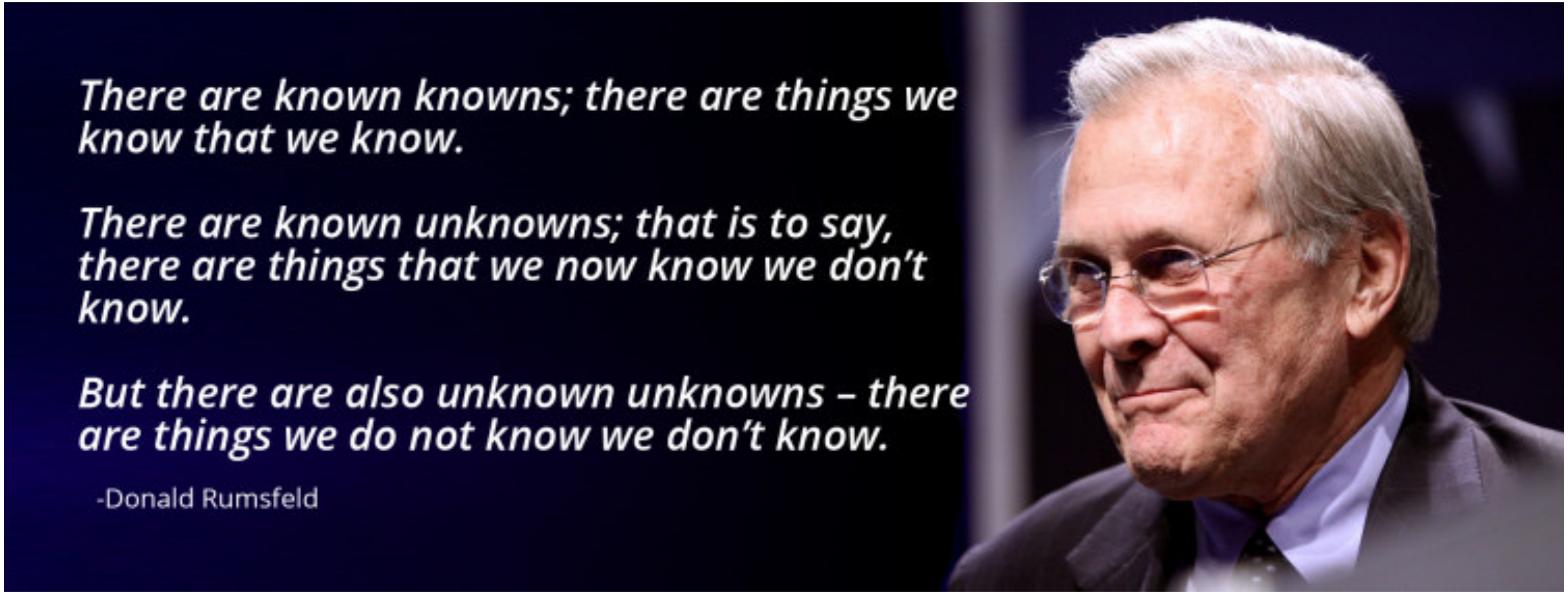

Gender
Department
Admission



[Bickel]

The Fundamental Problem of Causal Learning

- We don’t know if we have unobserved confounders.



There are known knowns; there are things we know that we know.

There are known unknowns; that is to say, there are things that we now know we don’t know.

But there are also unknown unknowns – there are things we do not know we don’t know.

-Donald Rumsfeld

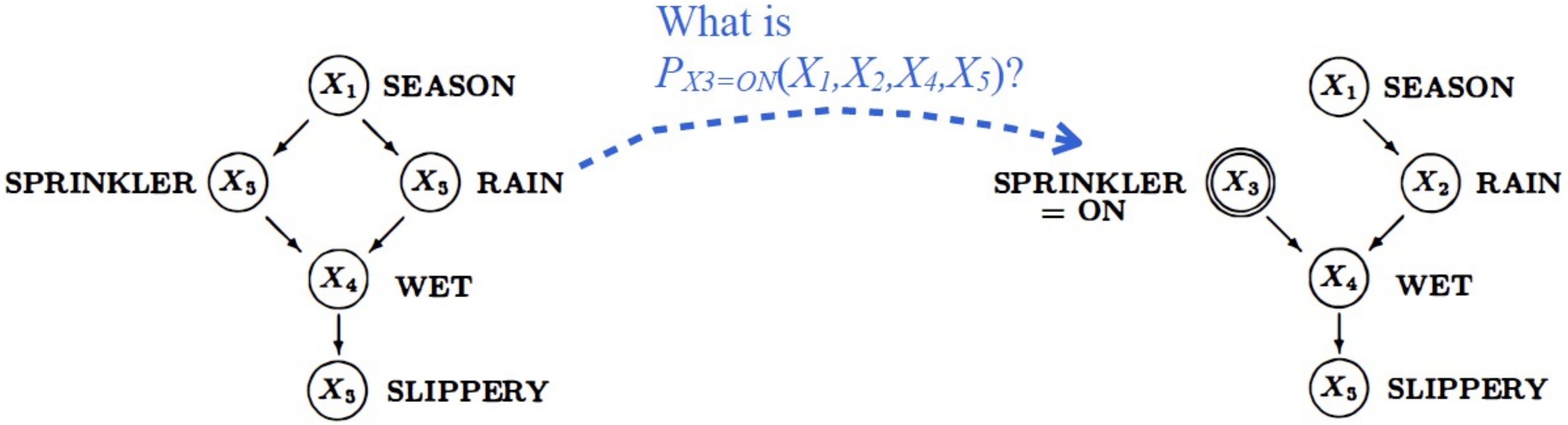
The Mindset of Causal Learning from Observational Data

- Given a fixed set of variables X, observational data **doesn't prove causality**; it **rules out non-causal** explanations.

Causal Models

Causal Models

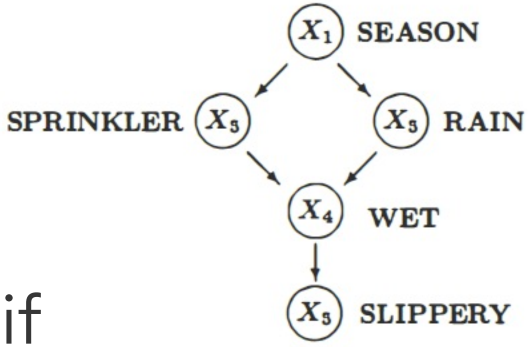
- Infer effect of interventions:



What is $P_{X3=ON}(X_1,X_2,X_4,X_5)$?

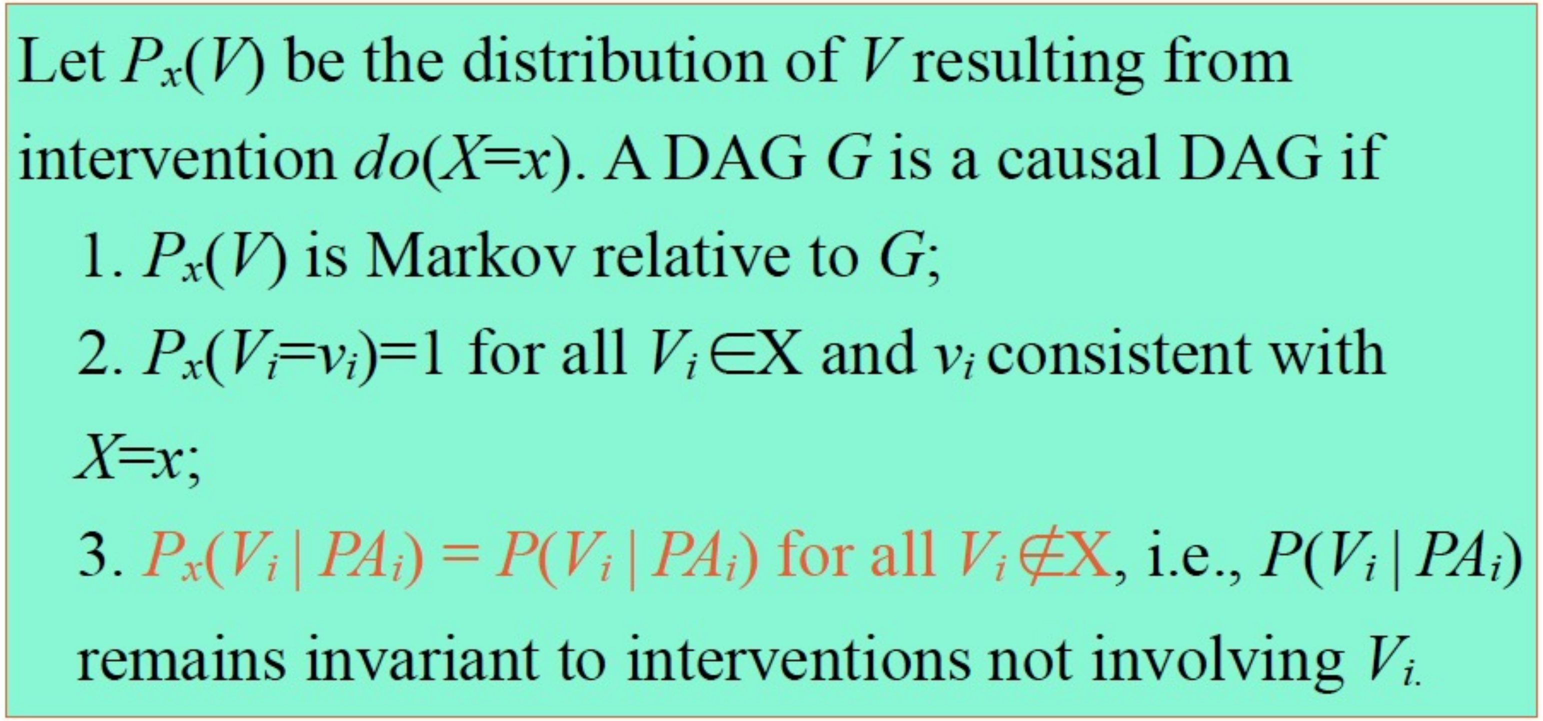
Kinds of questions we ask with Causal Models

- **Prediction:** Would the pavement be slippery if we *find* the sprinkler off?
  - $P(Slippery \mid Sprinkler = off)$
- **Intervention:** Would the pavement be slippery if we *make sure* that the sprinkler is off?
  - $P(Slippery \mid do(Sprinkler = off))$
- **Counterfactual:** Would the pavement be slippery had the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?
  - $P(Slippery_{\{Sprinkler=off\}} \mid Sprinkler = on, Slippery = no)$



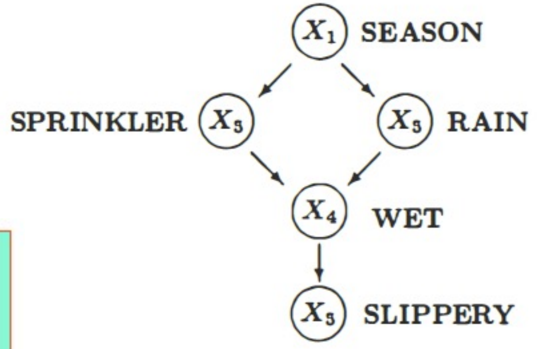
Causal DAGs

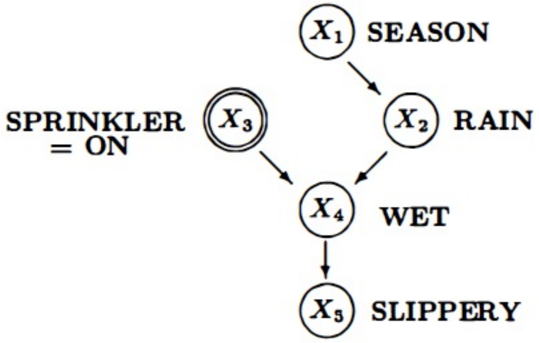
- Able to represent and respond to external or spontaneous changes



Let $P_x(V)$ be the distribution of V resulting from intervention $do(X=x)$. A DAG G is a causal DAG if

1. $P_x(V)$ is Markov relative to G;
2. $P_x(V_i=v_i)=1$ for all $V_i \in X$ and v_i consistent with $X=x$;
3. $P_x(V_i \mid PA_i) = P(V_i \mid PA_i)$ for all $V_i \notin X$, i.e., $P(V_i \mid PA_i)$ remains invariant to interventions not involving V_i.

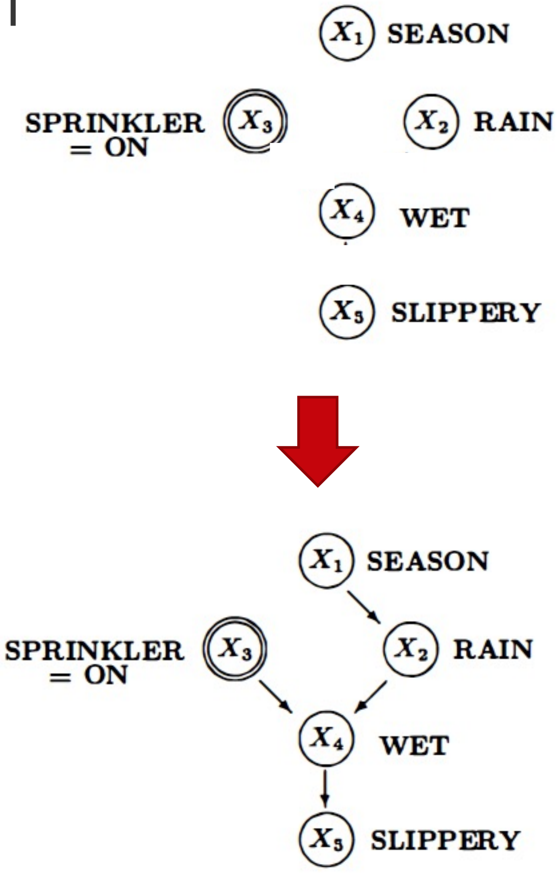




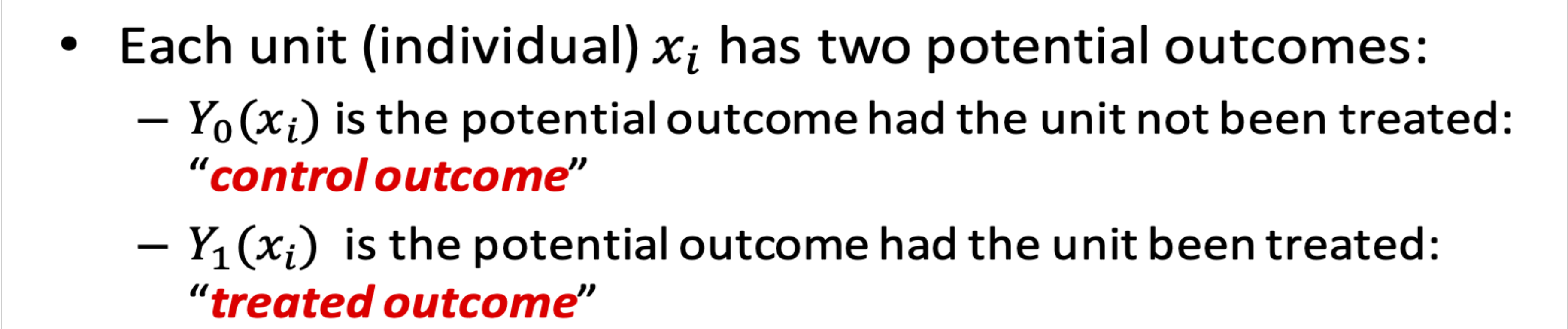
What is $P_{X_3=ON}(X_1, X_2, X_4, X_5)$?

Identification of Causal Effects

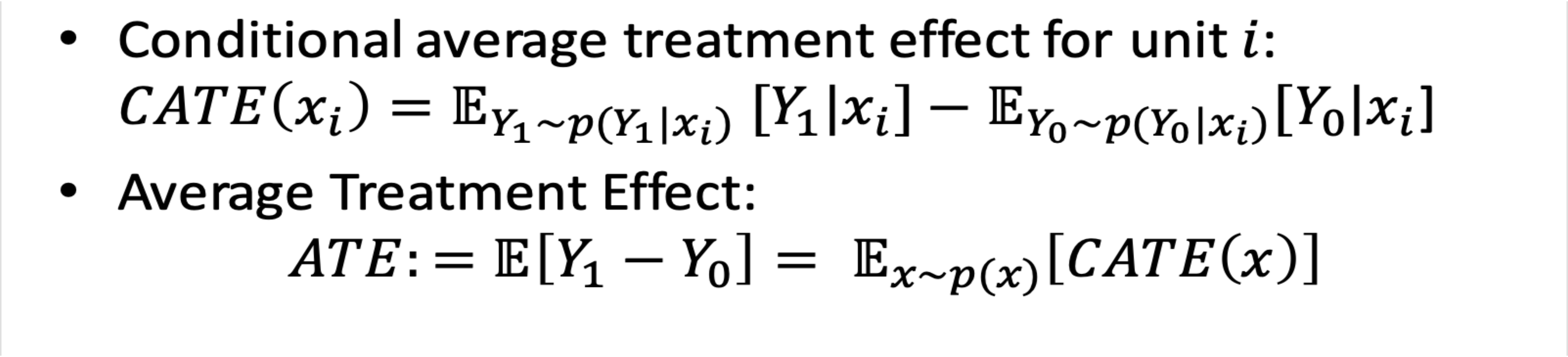
- **Intervention:** Would the pavement be slippery if we *make sure* that the sprinkler is off?
  - $P(Slippery \mid do(Sprinkler = off))$
- **Gold standard:** Randomized controlled experiments.
- Often expensive or impossible/unethical to do.



Potential Outcomes Framework (Rubin-Neyman)



- Each unit (individual) 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***"



- Conditional average treatment effect for unit i:
  $CATE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$
- Average Treatment Effect:

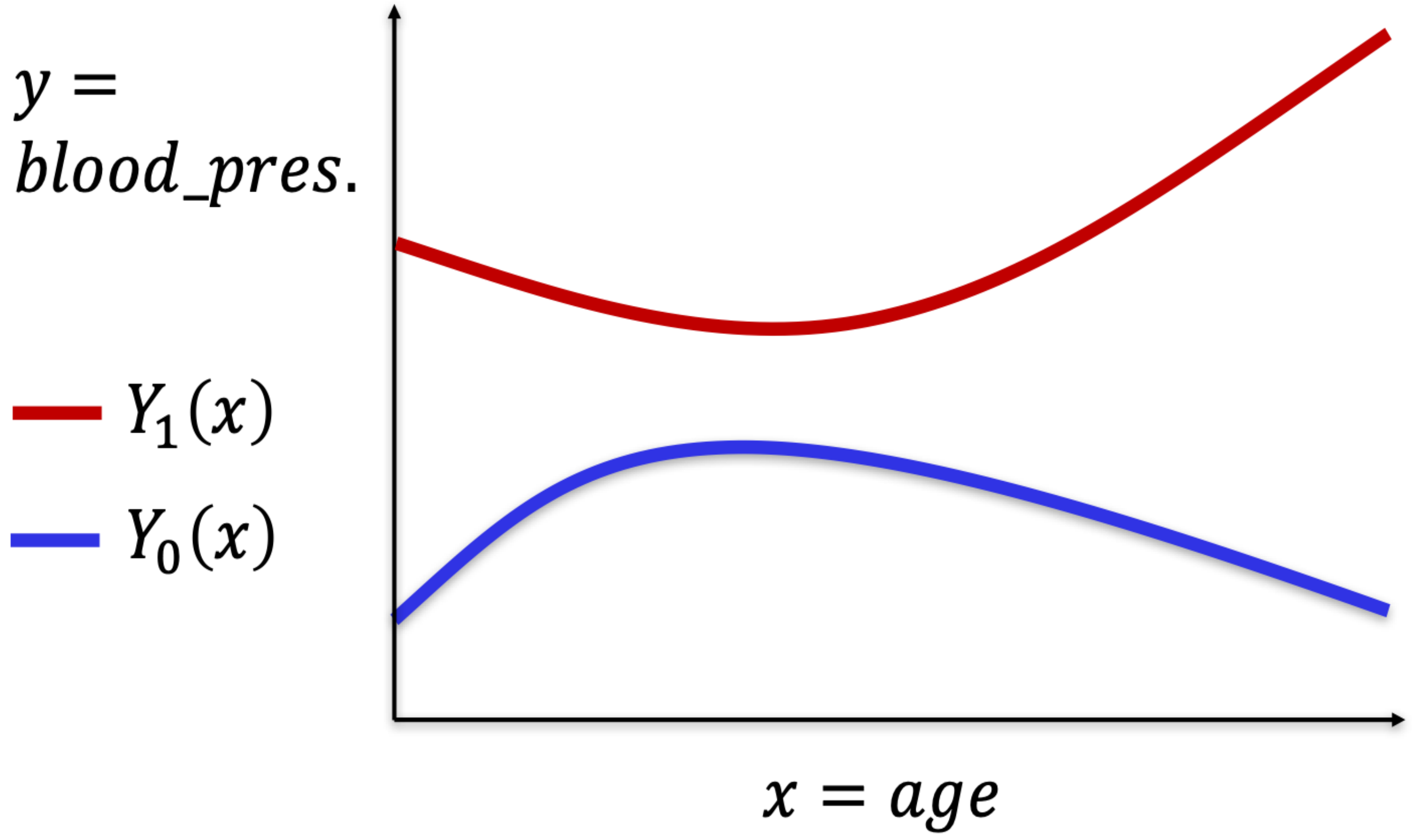
$$ATE := \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{x \sim p(x)}[CATE(x)]$$

- In RCT, $E[Y_1] = E[\,Y \mid do(Treatment)\,]$ and
  $E[Y_0] = E[\,Y \mid do(NoTreatment)\,]$

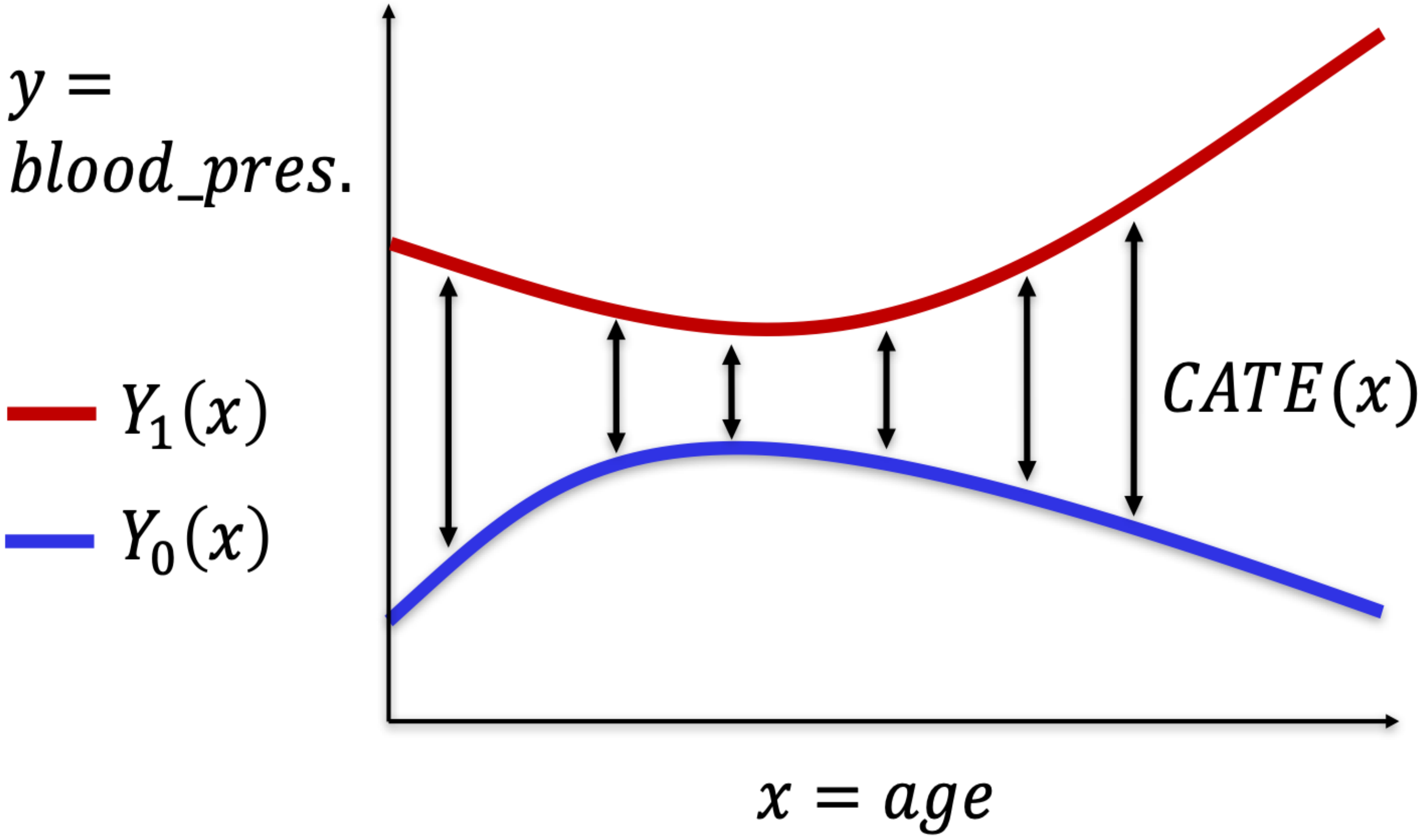
“The fundamental problem of causal inference”

We only ever observe one of the two outcomes

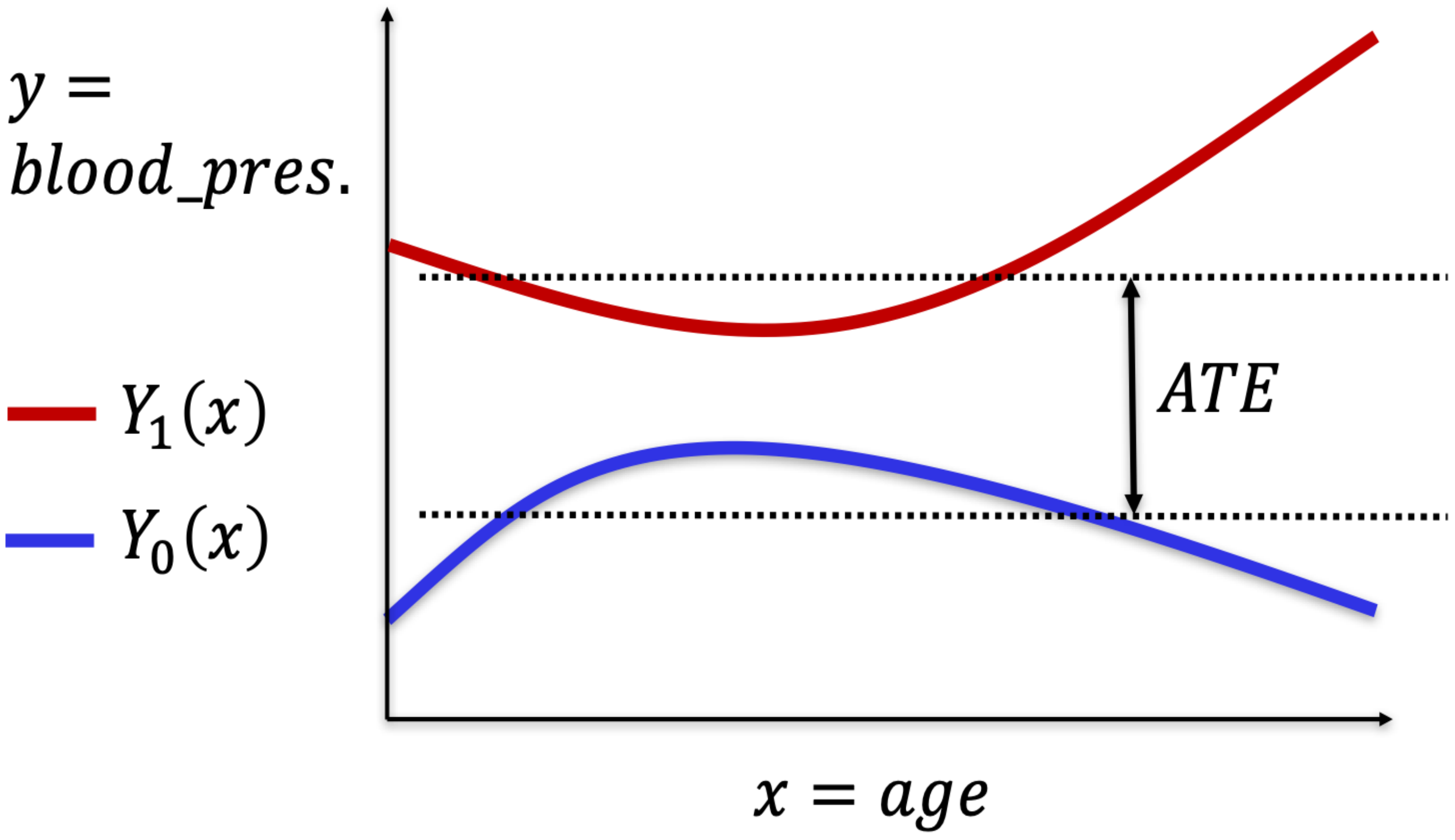
Example – Blood pressure and age



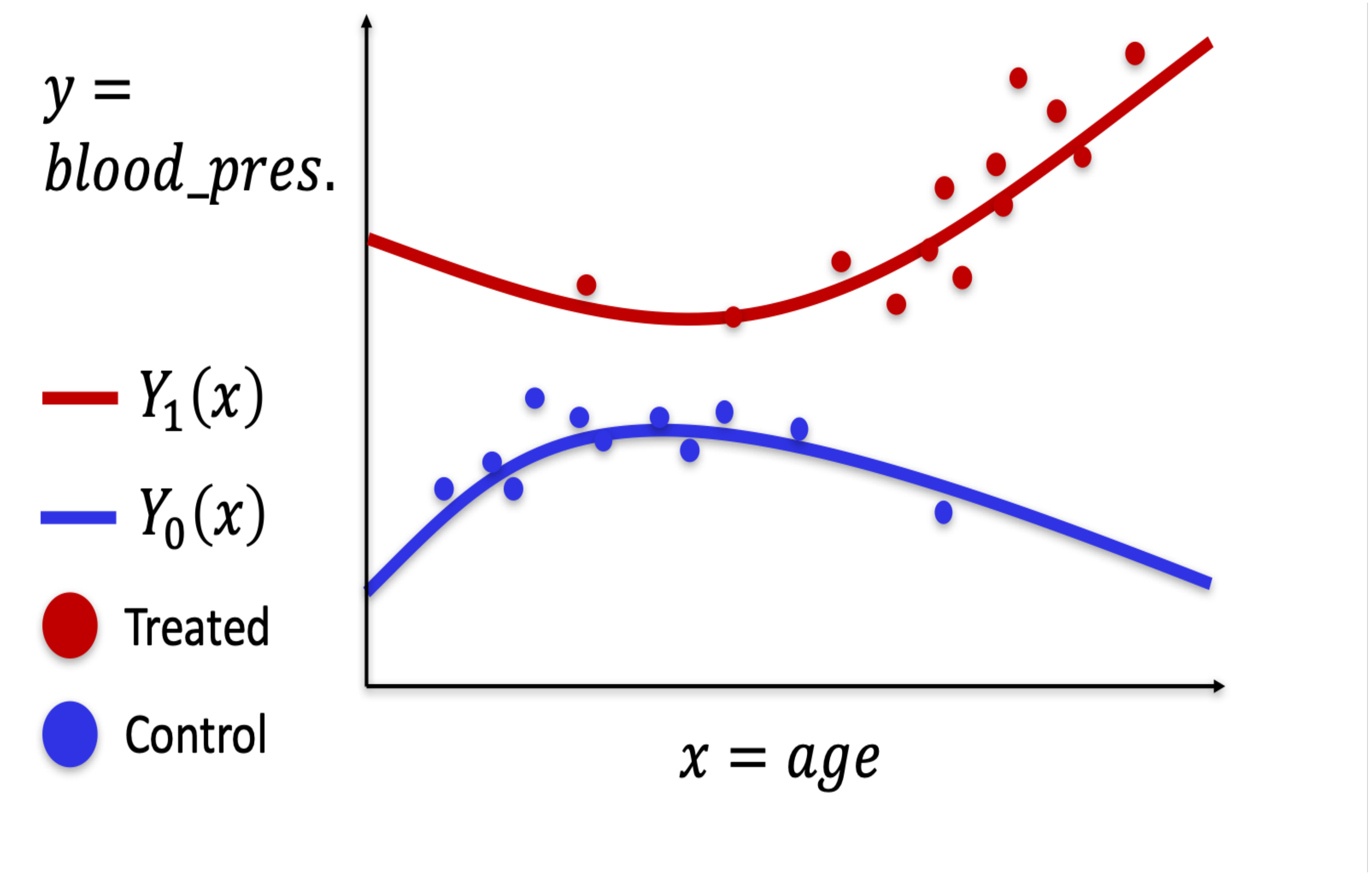
Example – Blood pressure and age



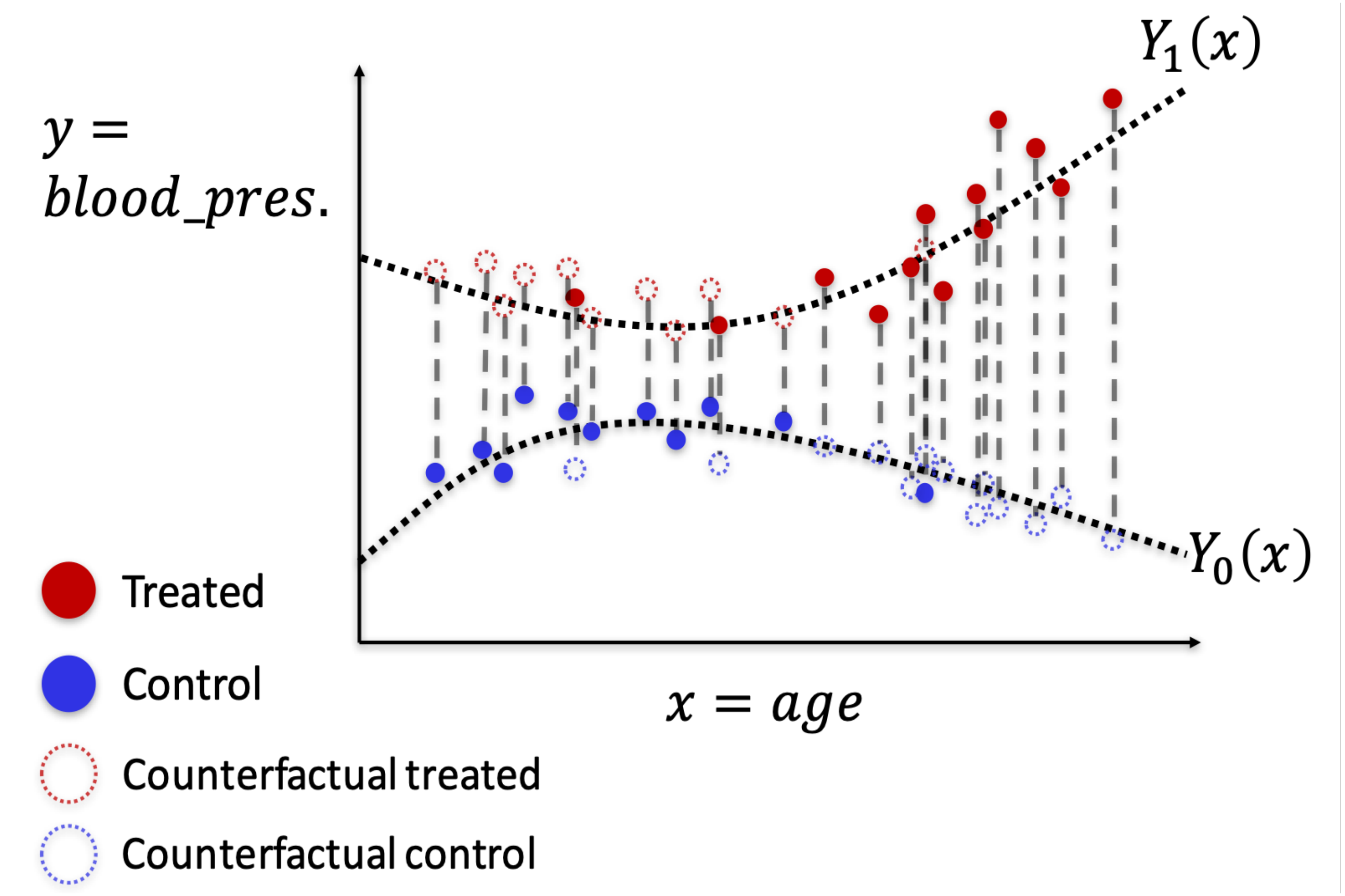
Example – Blood pressure and age



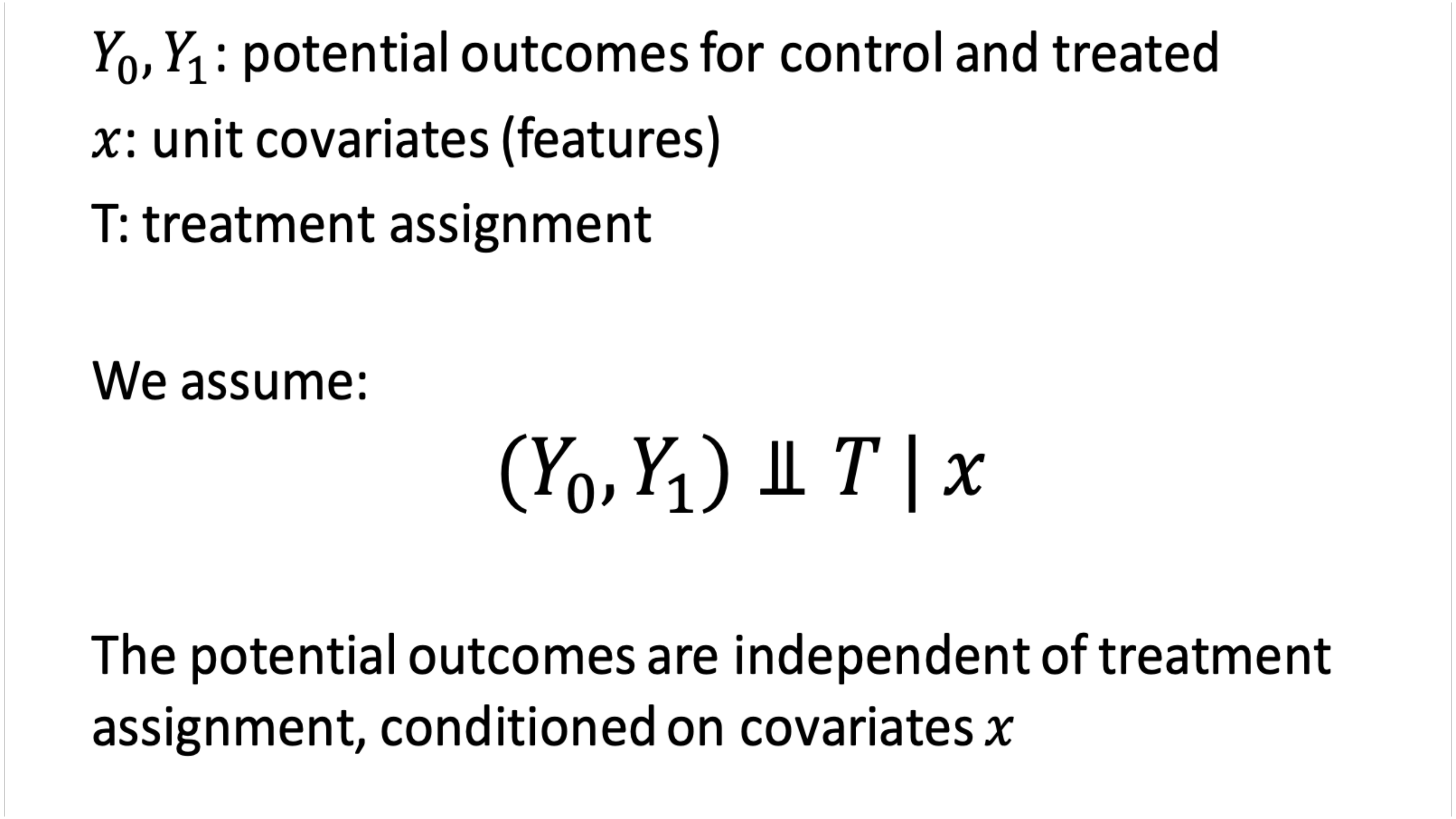
Example – Blood pressure and age



Example – Blood pressure and age



Typical Assumption – No unmeasured confounders



Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

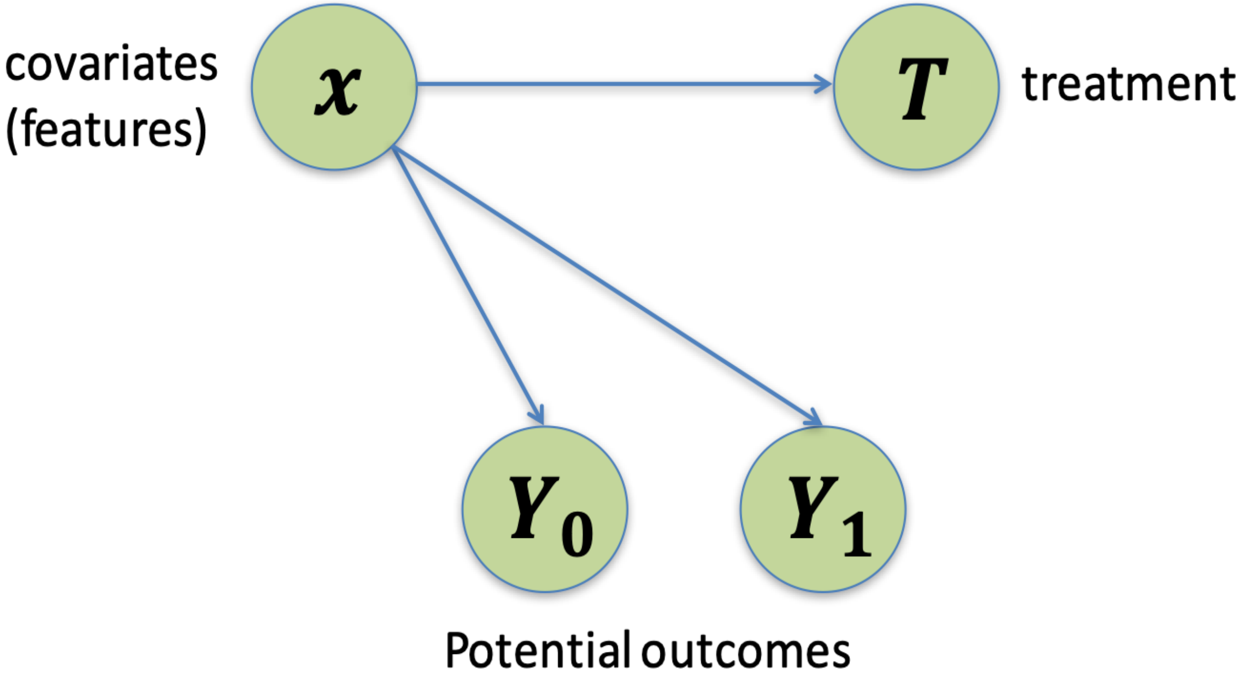
T: treatment assignment

We assume:

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

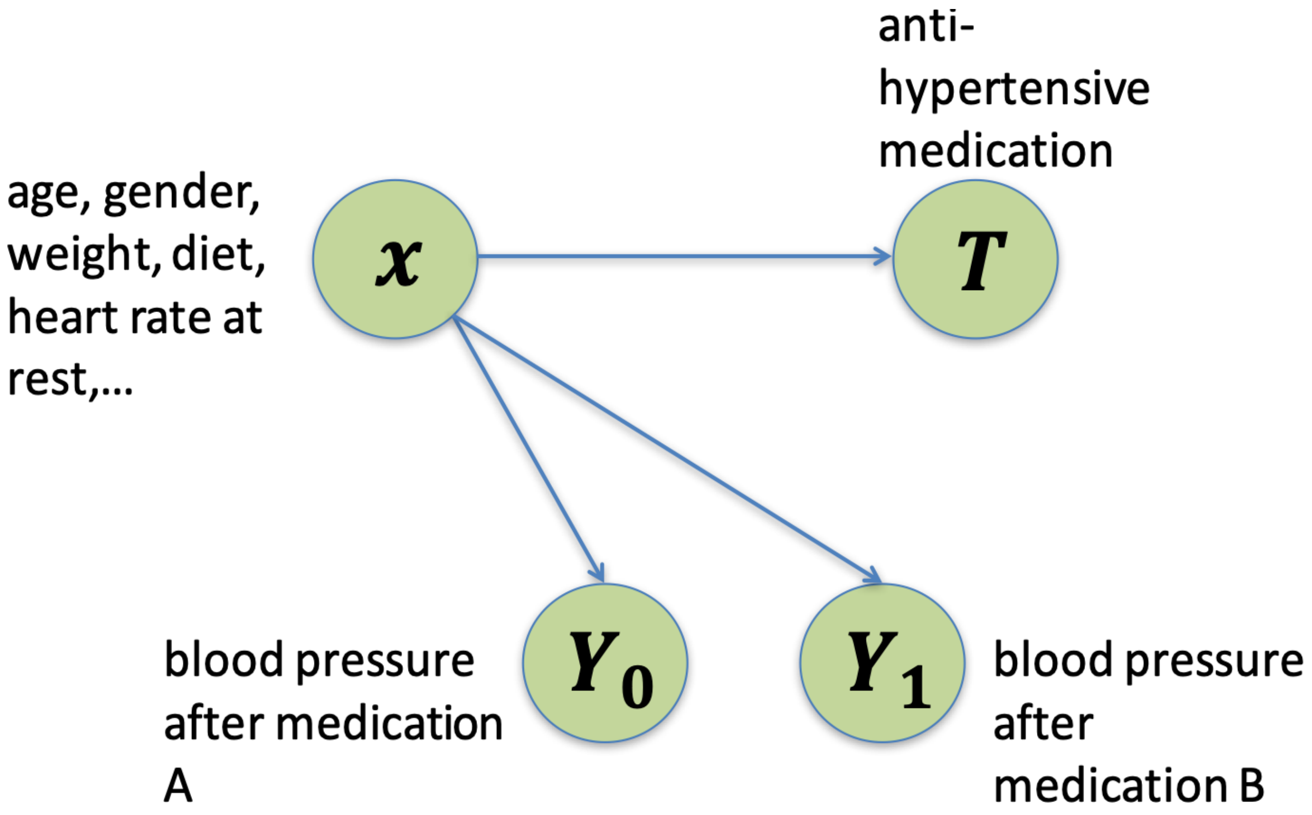
The potential outcomes are independent of treatment assignment, conditioned on covariates x

Typical Assumption – Ignorability



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

Typical Assumption – Ignorability



anti-hypertensive medication

age, gender, weight, diet, heart rate at rest,...

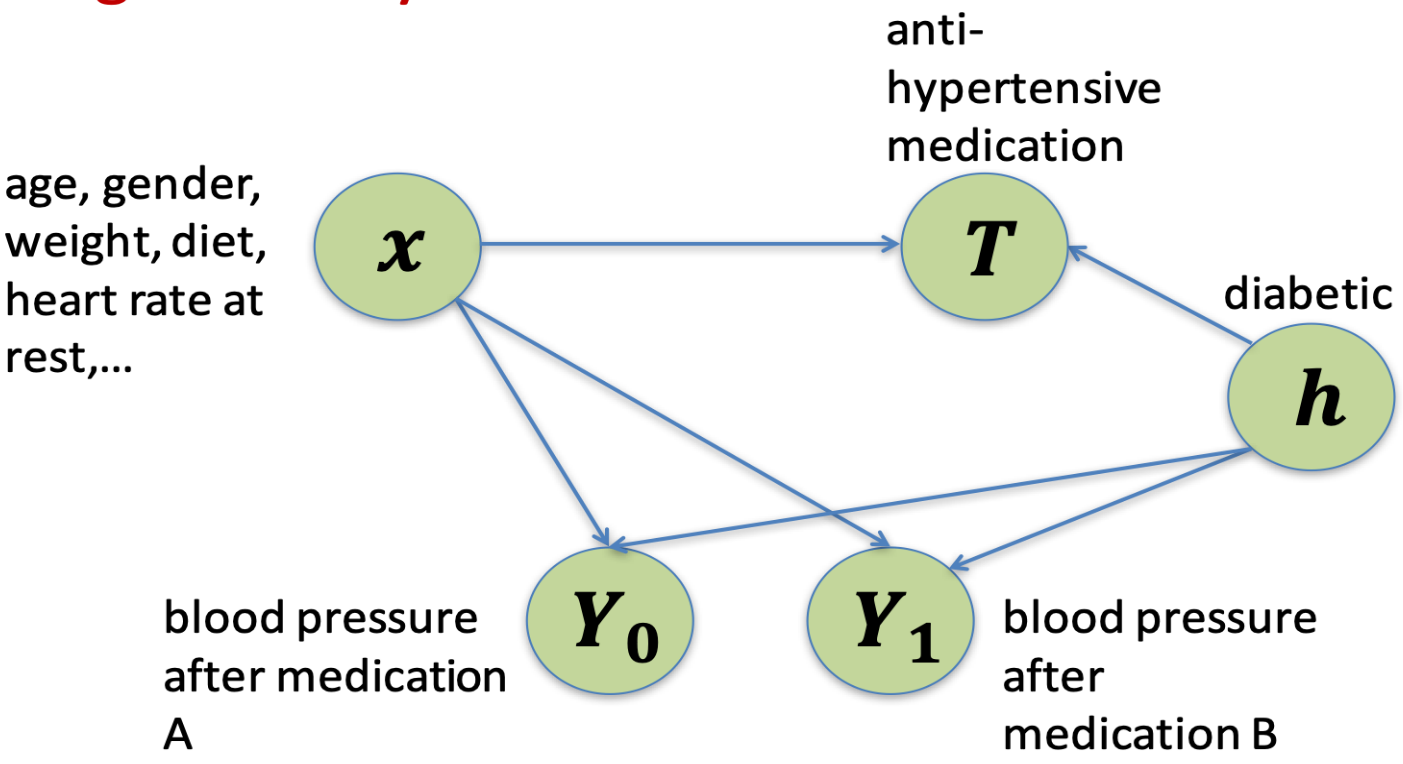
blood pressure after medication A

blood pressure after medication B

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

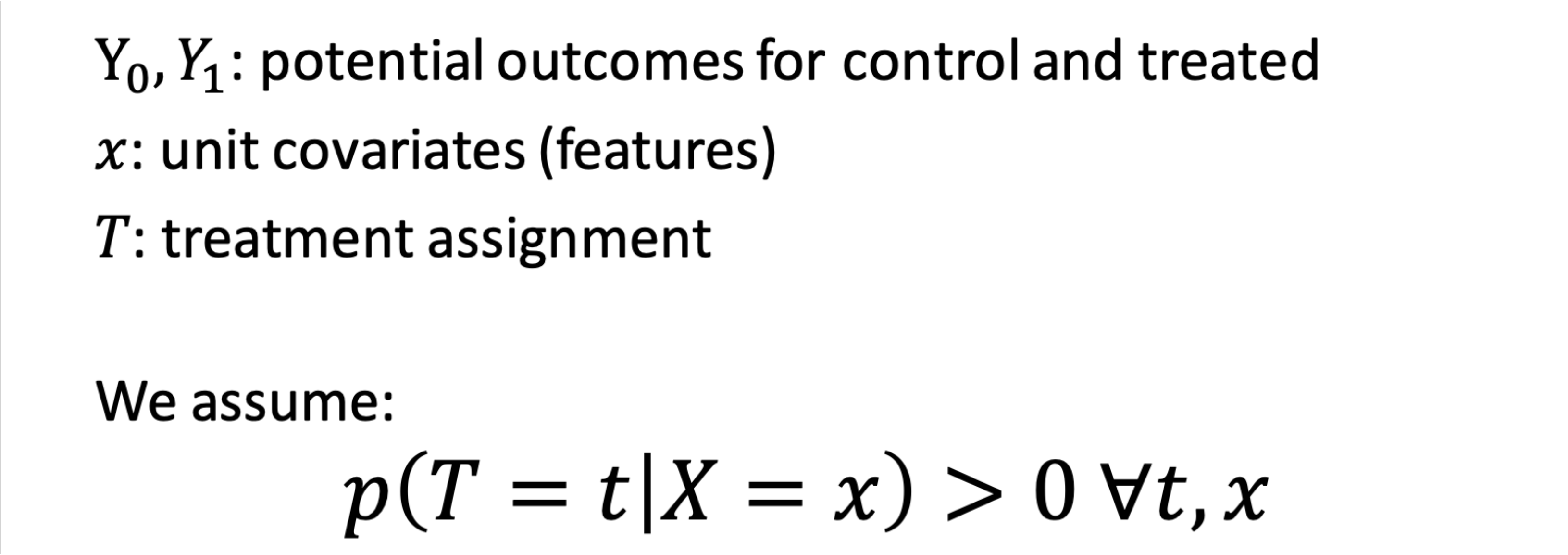
Typical Assumption – Ignorability

No Ignorability


anti-hypertensive medication
age, gender, weight, diet, heart rate at rest,...
x
T
diabetic
h
blood pressure after medication A
Y_0
Y_1
blood pressure after medication B

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

Typical Assumption – Common Support



Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

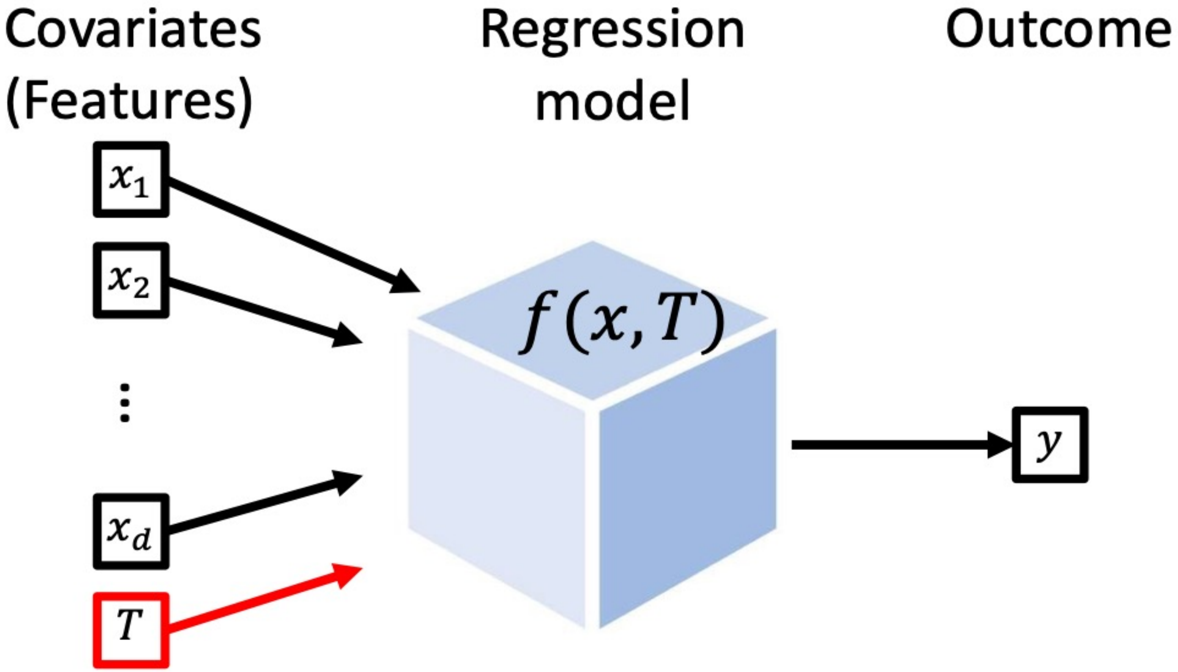
T: treatment assignment

We assume:

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

Covariate Adjustment

Explicitly model the relationship between treatment, confounders, and outcome:

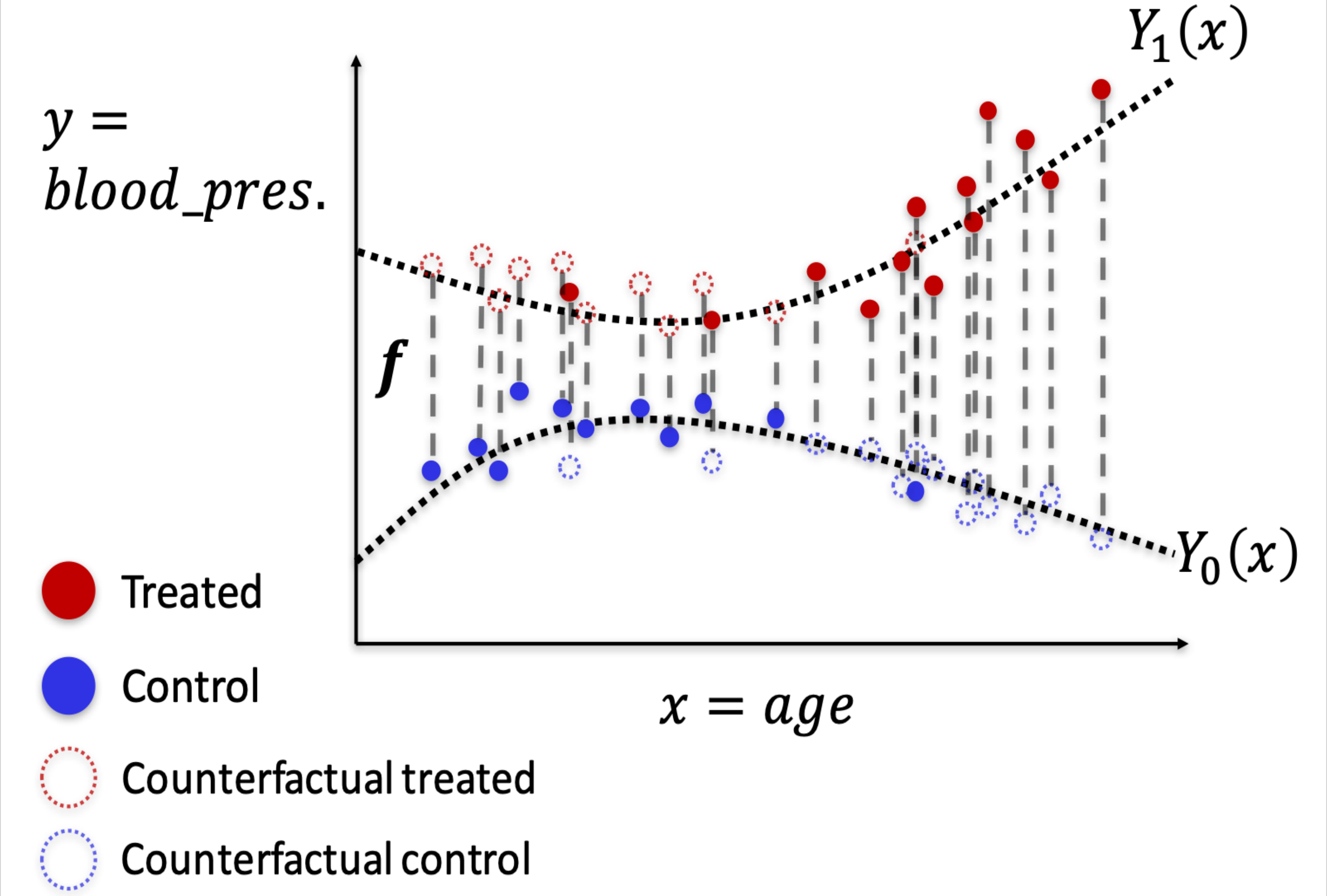


Covariate Adjustment

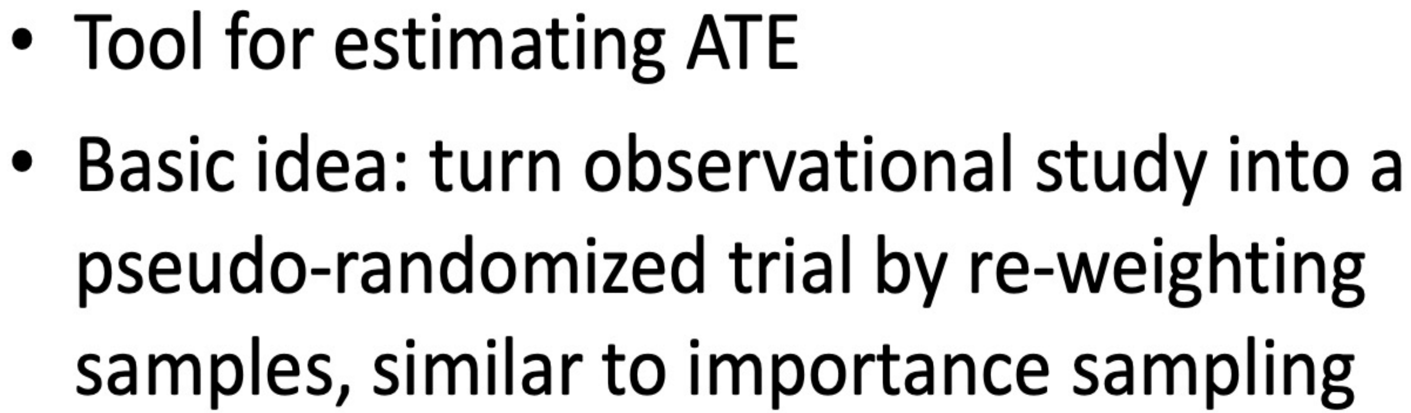
- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:
  $\mathbb{E}_{x\sim p(x)}\left[\, \mathbb{E}[Y_1|T=1,x] - \mathbb{E}[Y_0|T=0,x]\right]$
- Fit a model $f(x,t) \approx \mathbb{E}[Y_t|T=t,x]$

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

Covariate Adjustment



Propensity scores

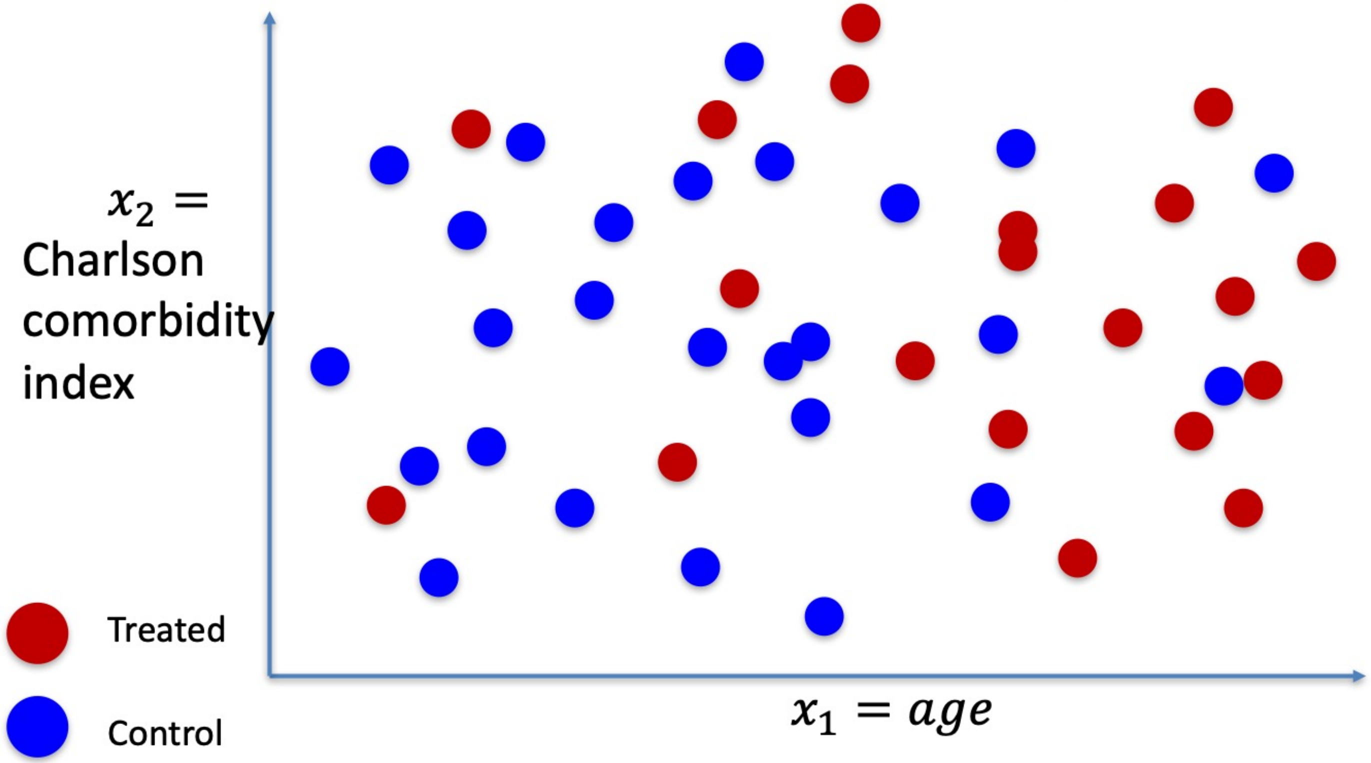


- Tool for estimating ATE
- Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples, similar to importance sampling

Inverse propensity score re-weighting

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

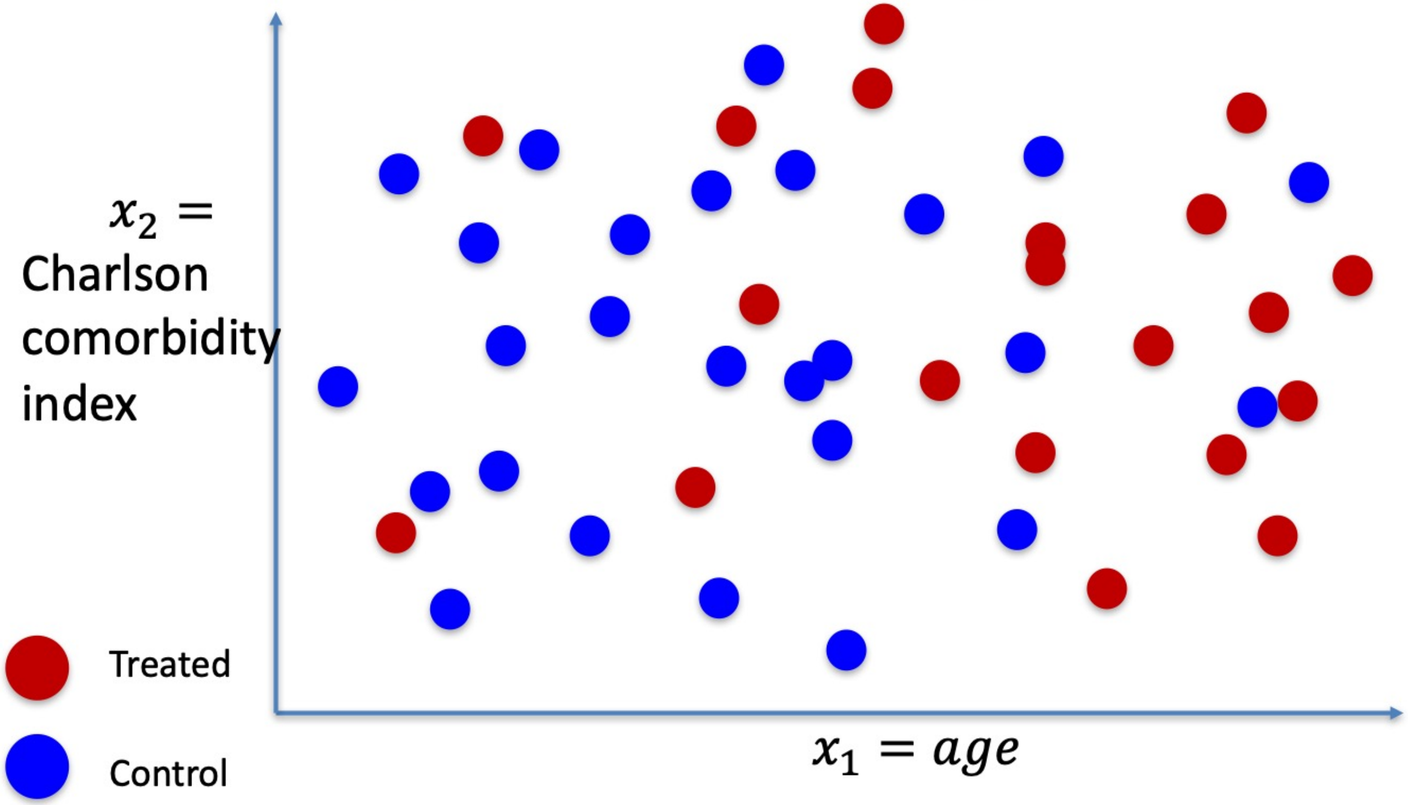
control *treated*



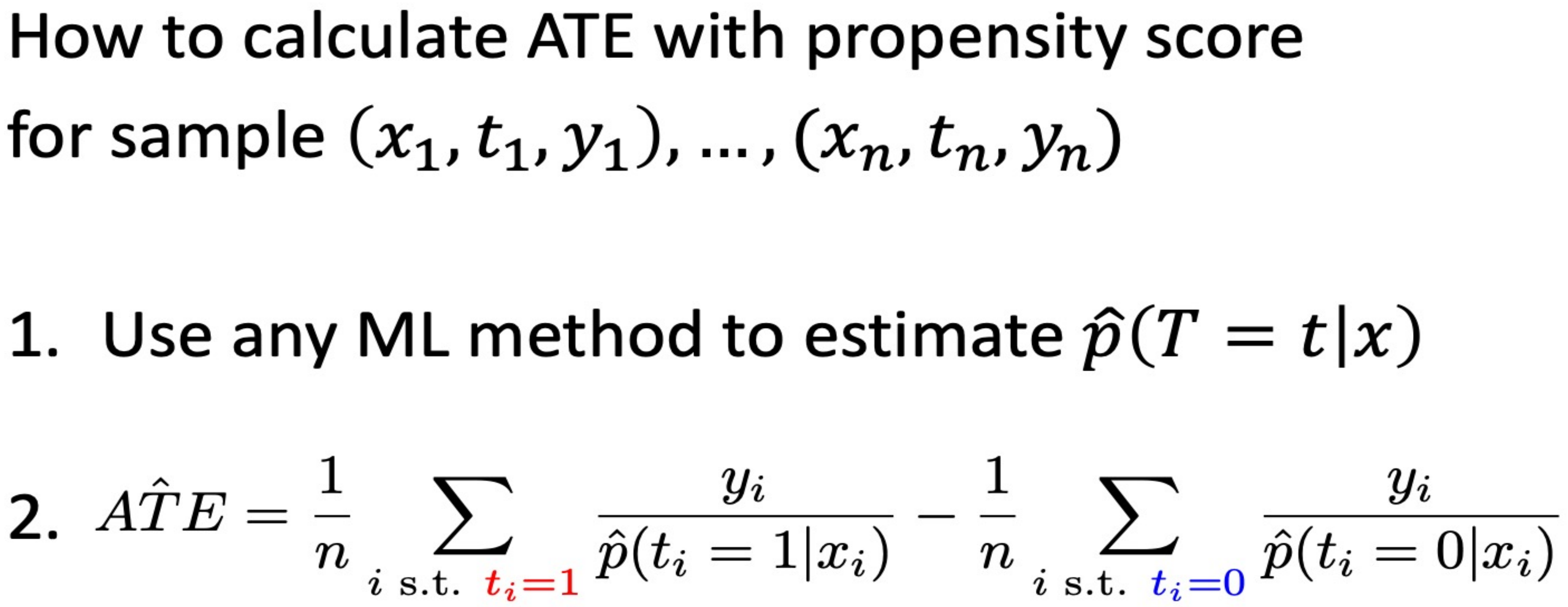
Inverse propensity score re-weighting

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

reweighted control *reweighted treated*



Inverse propensity score re-weighting



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

Inverse propensity score re-weighting

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

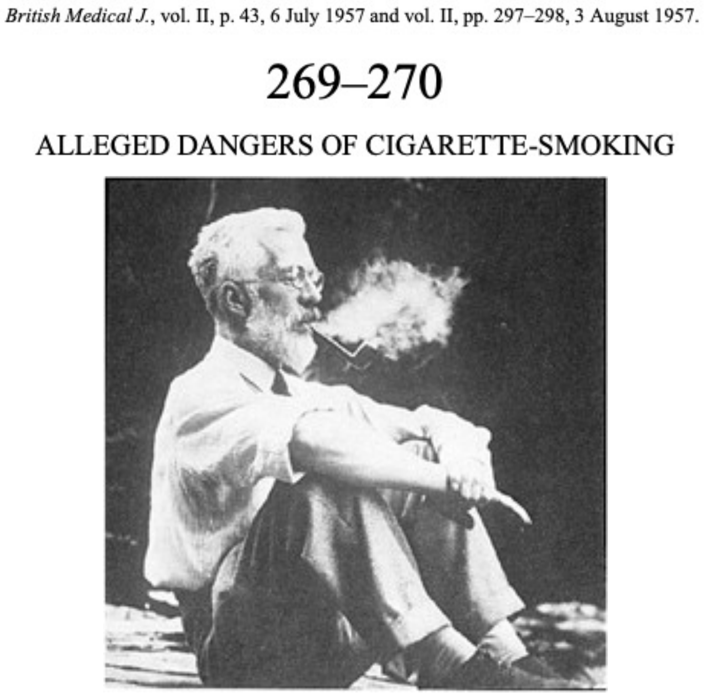
Problems with inverse propensity scores

- Need to estimate propensity score (problem in all propensity score methods)
- If there’s not much overlap, propensity scores become non-informative and easily mis-calibrated
- Weighting by inverse can create large variance and large errors for small propensity scores
  - Exacerbated when more than two treatments

Causality in Practice

Causality in Practice

- RA Fisher: famous statistician, rejected smoking->cancer causality
- His claim: Only associational studies have been run so far.
  - Monozygotic twins have more similar smoking patterns than dizygotic twins, so maybe a genetic propensity to smoke instead of a causal link?
- How many cancers were caused by this wrong interpretation?

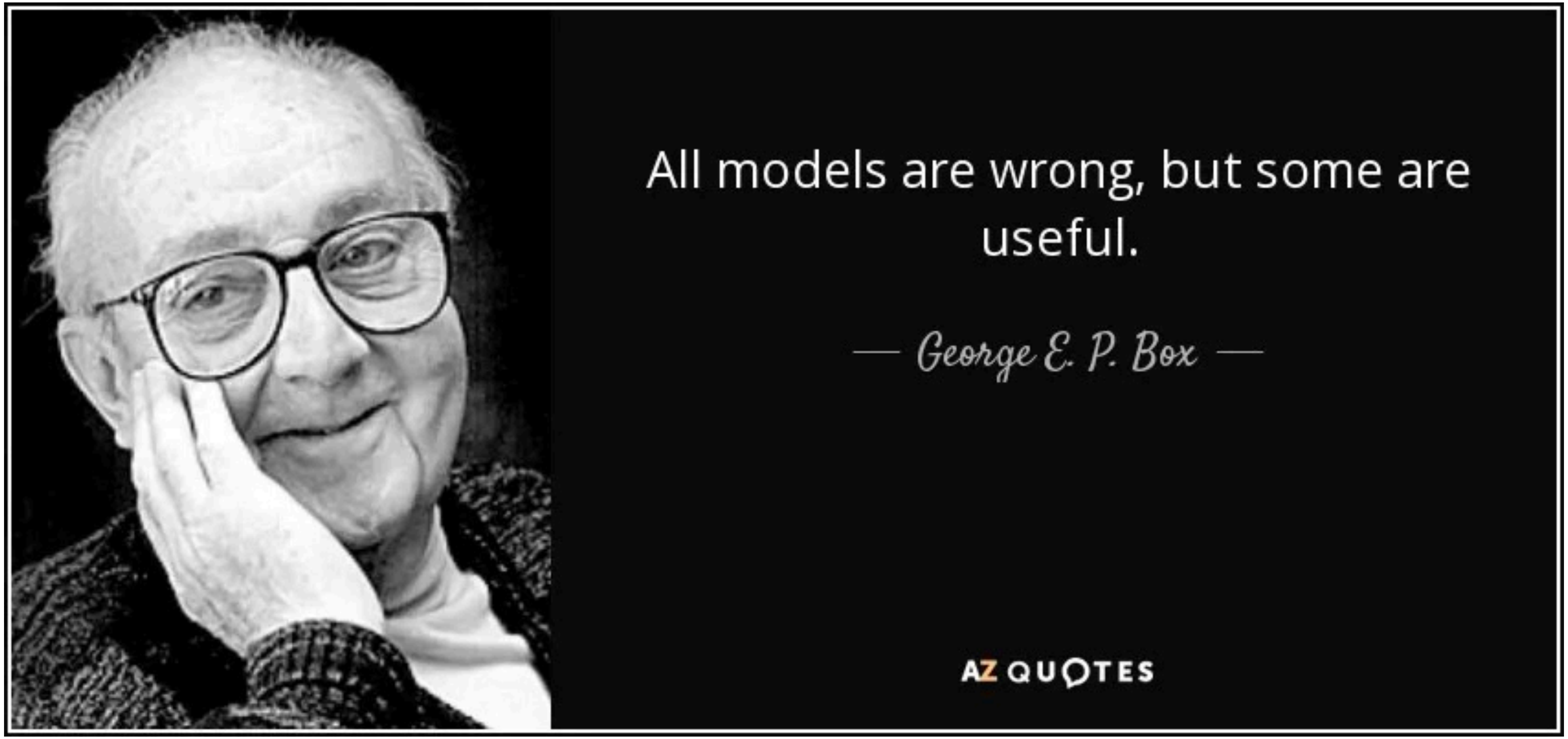


British Medical J., vol. II, p. 43, 6 July 1957 and vol. II, pp. 297–298, 3 August 1957.

269–270

ALLEGED DANGERS OF CIGARETTE-SMOKING

Causality in Practice



All models are wrong, but some are useful.

— George E. P. Box —

Causality in Practice: What is our model's use?

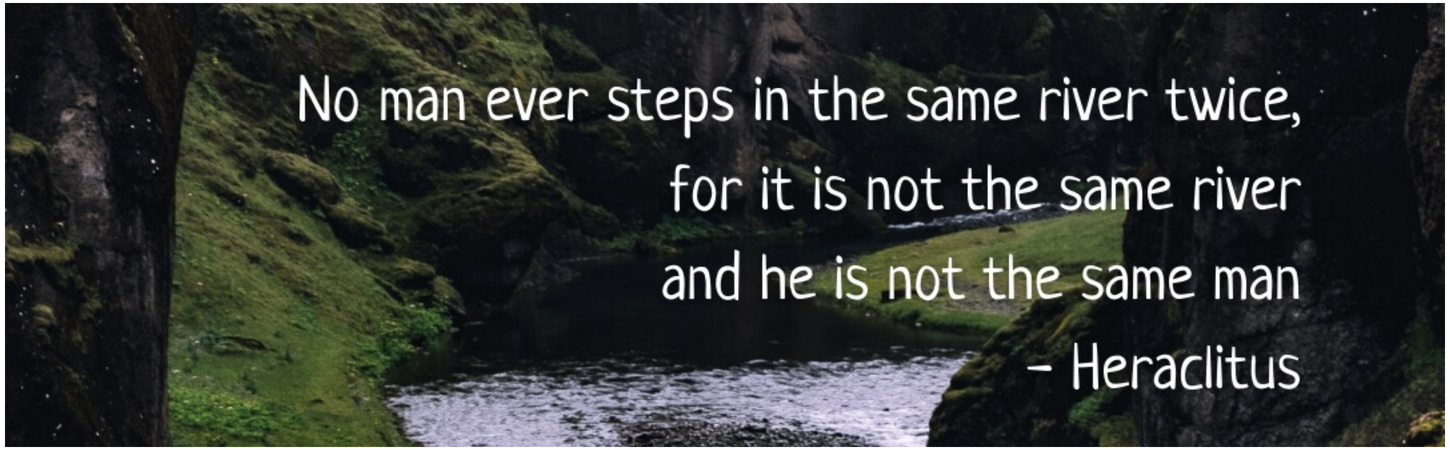
- **Models are simplifications** of reality—they can never be entirely correct.
- **The key question is:**
  - How can we use models to make **better decisions**?
- **Causal inference vs. Prediction:**
  - **Prediction models** optimize accuracy but may not reveal **why** outcomes occur.
  - **Causal models** aim to uncover mechanisms, guide interventions, and inform policy.

Example: Sensitive features

- Suppose we have access to a sensitive feature (e.g. race, gender) that we don’t want to make decisions based on.
- Should we exclude this feature from our model training?
- But holding it out won’t get rid of the effect:
  - Indirect bias, hide disparities rather than eliminate them.
- Better strategy: Learn the causal effect of the sensitive feature, then choose what to do with it:
  - Throw out the effect of the feature (counterfactual fairness)
  - Sweep over all possible values of the sensitive feature
  - Learn an invariant representation

Example: Process-based decisions in medicine

- Medicine is a continuous process, not a one-time prediction.



> No man ever steps in the same river twice,
> for it is not the same river
> and he is not the same man
> – Heraclitus

- Dropping into the river of treatment:
  - Upstream influences are missing not-at-random.
  - Correcting for missing not-at-random can drive us toward biological causality.
  - BUT if the missing not-at-random will persist in the real world, then the causal model is LESS useful than the model biased by upstream influences.

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON