Causal Inference from Observational Data

Shalmali Joshi
Vector Institute

Motivational Questions

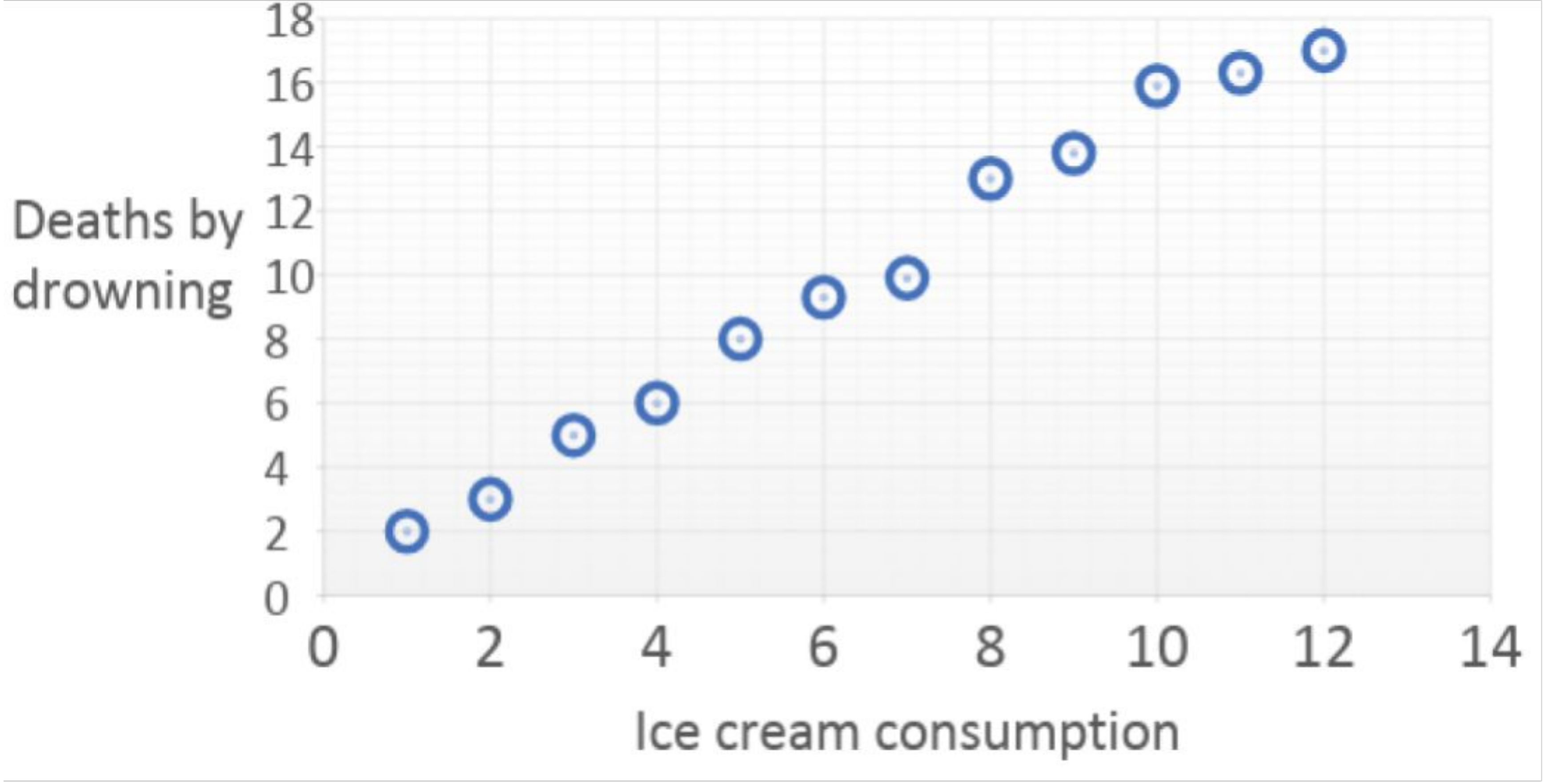
- Find which medication A/B is best for diabetics?
- Should I deploy this new feature in company's product?
- Would this person be rejected for the job had their name been different?

Based on Uri Shalit's slides, 2018

Bring in the Machine Learning Hammer

- Supervised Classification only learns “associations” $p(y|x)$
- X = [lab_tests, diagnoses, medications]
- y = [severely_diabetic]
- Mostly just correlations

But then many things are correlated



Thanks to Lan Liu, U. Minnesota

Do we really want to learn p(death by drowning| # ice-creams)?

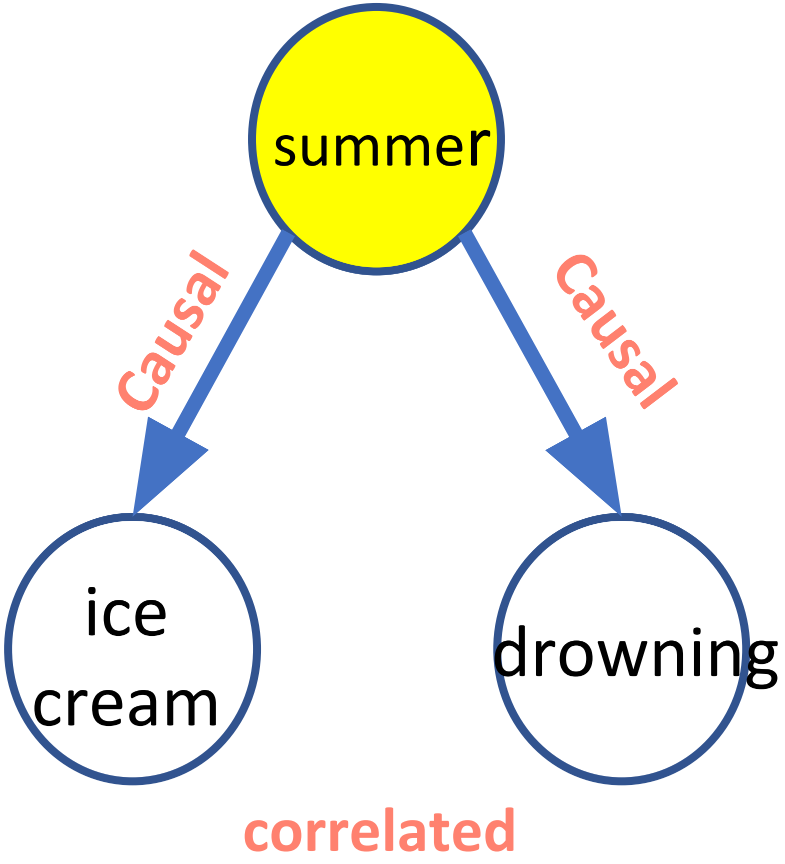
Questions:

1. Does eating ice-cream cause death by drowning?

2. Is something else causing both these phenomena

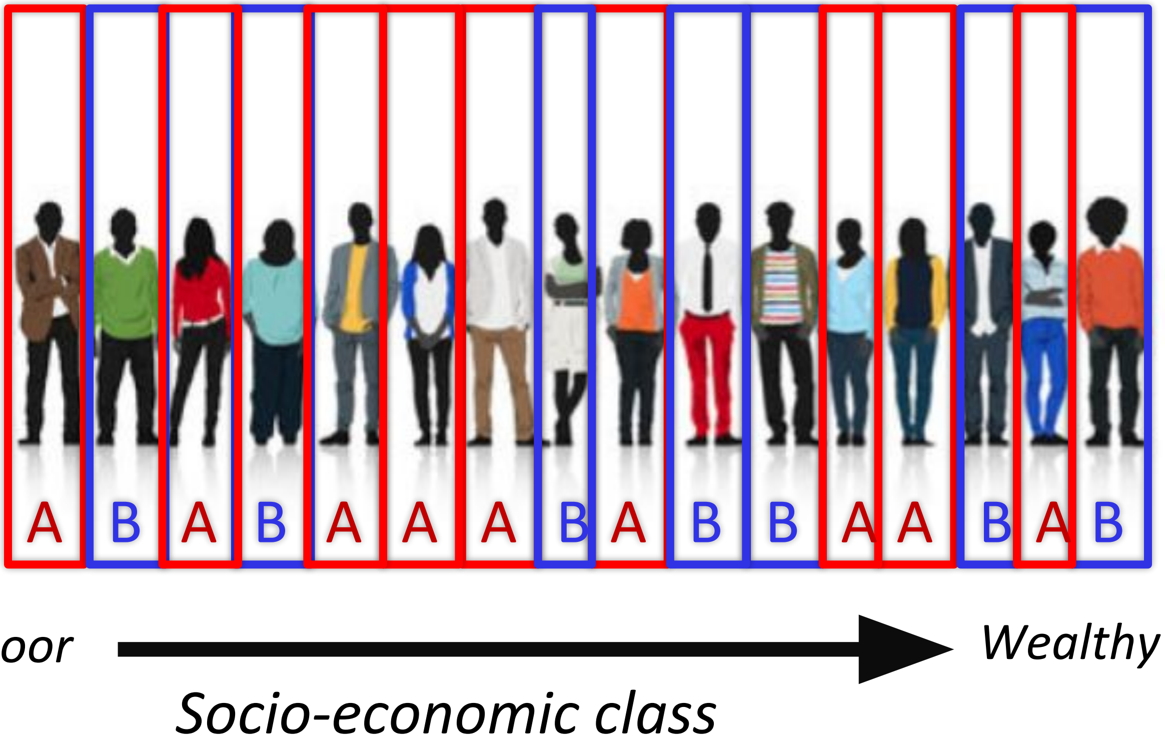
3. Could we realistically have some randomly chosen humans eat lots of ice-cream and see if what happens?

4. In a healthcare setting, one cannot risk death because of the treatment!



Confounding!

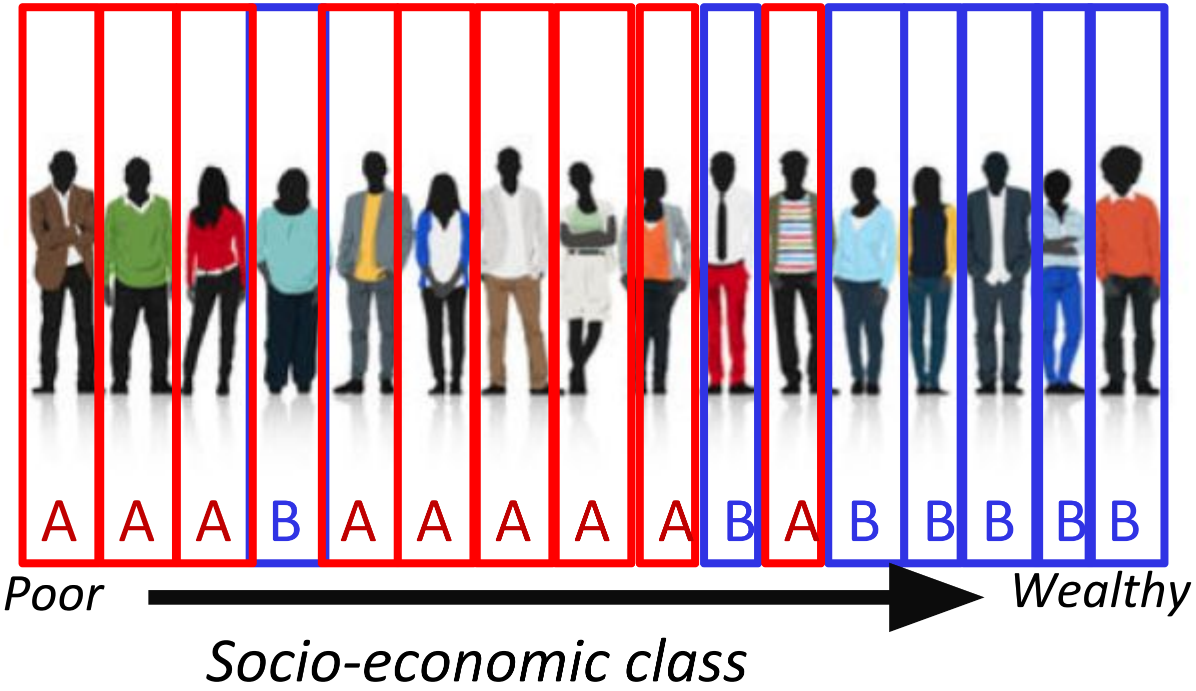
Randomized controlled trial (RCT)



treatment

A **or** B

More Common: Observational Setting



treatment

A **or** B

Clinical setting

- RCTs are also known as “clinical trials”
  - Tens of thousands every year, costing tens of billions of dollars
  - Every new medication must pass several stages of RCTs before approval for human use
- Observational study
  - Use existing data, tracking people’s medications and blood sugar
  - Problem: the space of possible confounders

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***?
- Hidden confounding:
  Maybe using ***B*** is *worse* than ***A***, but rich patients usually take ***B*** and richer people also have better health outcomes.
  If we don’t know whether a patient is rich or not, we might conclude ***B*** is better

Causal Hierarchy (not captured by mere associations)

Observational Questions: “What if we see A”

Action Questions: “What if we do A?”

Counterfactuals Questions: “What if we did things differently?”

Options: “With what probability?”

Two foundational ways to think of Causality

- Potential Outcomes (Rubin, Neyman)
- Causal Graphical Models (Judea Pearl)
- Either framework needs manipulating reality

Potential Outcomes

- Unit: a person, a bacteria, a company, a school, a website, a family, a piece of metal, …
- Treatments / actions / interventions (A/B)
- Potential outcomes

  Y1 : the unit's outcome had they been subjected to treatment t=1

  Y0 : the unit's outcome had they been subjected to treatment t=0. If number of treatments is T, we have T potential outcomes (T possibly infinite)

- In observations, a single unit gets one of the T treatments

Inferring under this framework requires assumptions

SUTVA: Stable Unit Treatment Value Assumption

The potential outcomes for any unit do not vary with the treatments assigned to other units

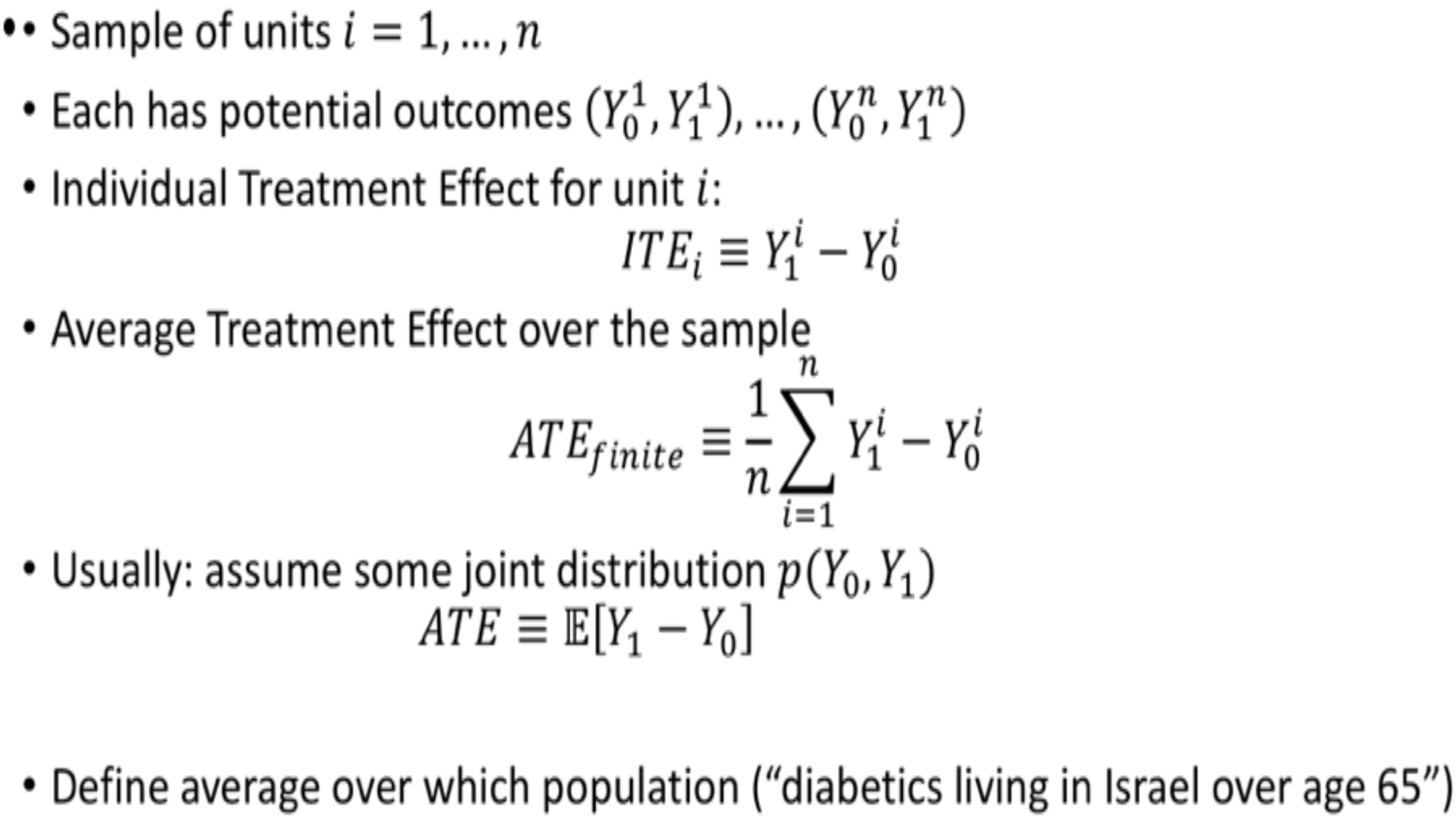
failure example: vaccination, network effects

For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes

failure example: some people get out-of-date medication

Consistency: $p(Y_t=y|X=x, T=t) = p(Y = y| X=x, T=t)$

Potential Outcomes Formalized



- Sample of units $i = 1, \dots, n$
- Each has potential outcomes $(Y_0^1, Y_1^1), \dots, (Y_0^n, Y_1^n)$
- Individual Treatment Effect for unit i:

$$ITE_i \equiv Y_1^i - Y_0^i$$

- Average Treatment Effect over the sample

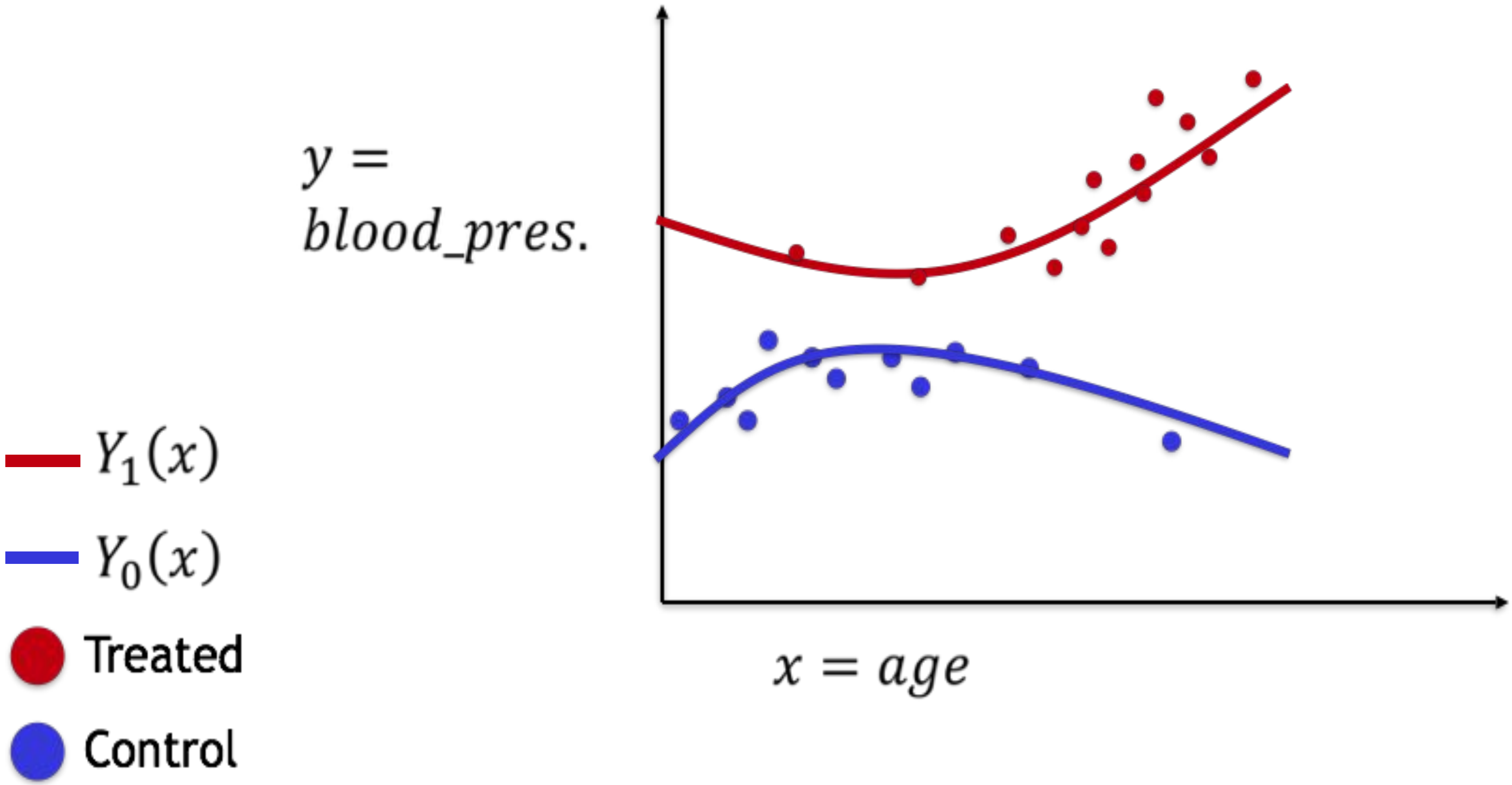
$$ATE_{finite} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_1^i - Y_0^i$$

- Usually: assume some joint distribution $p(Y_0, Y_1)$

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

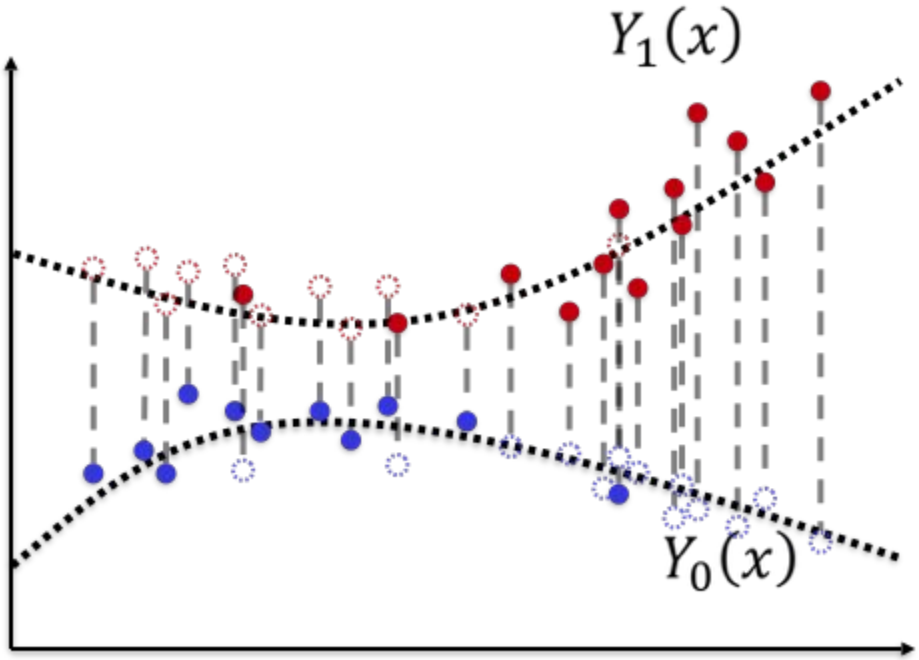
- Define average over which population ("diabetics living in Israel over age 65")

Example: Blood Pressure and Age



Example: Blood Pressure and Age

$y =$ $blood\_pres.$



Treated

$x = age$

● Treated

● Control

○ Counterfactual treated

○ Counterfactual control

━ $Y_1(x)$

━ $Y_0(x)$

Estimation Example

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

- True treatment effect:
  $\mathbb{E}[Y_1 - Y_0] = 2$

  $\mathbb{E}[Y|t=1] - \mathbb{E}[Y|t=0] =$
  $\frac{1}{4}(10+6+6+6) +$
  $\frac{1}{4}(8+8+8+4) =$

  $7 - 7 = 0$

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Within each group we get the true treatment effect!

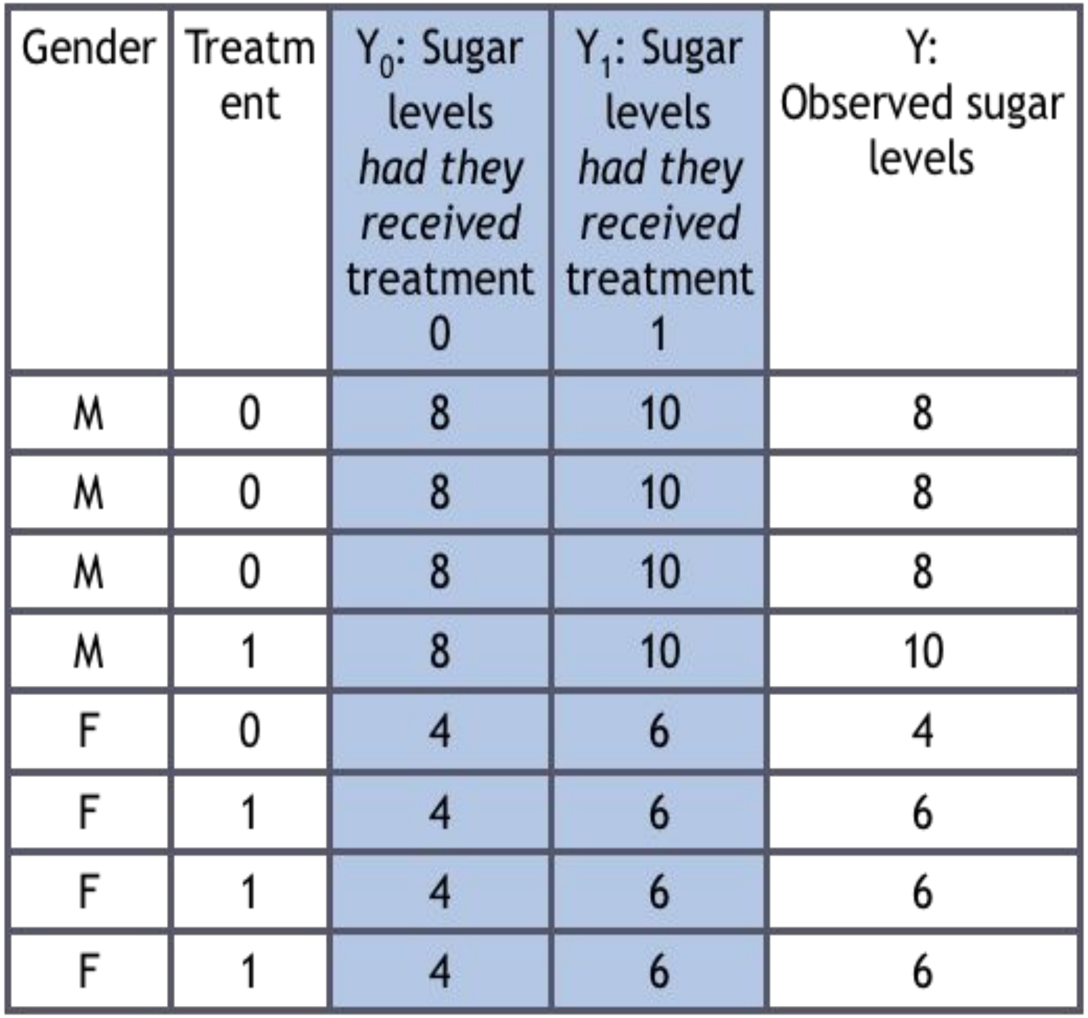
Estimation

- True treatment effect:
  $\mathbb{E}[Y_1 - Y_0] = 2$

$\mathbb{E}[Y|t = 1] = 7$
$\mathbb{E}[Y|t = 0] = 7$

$\mathbb{E}[Y|t = 0, Gender = M] = 8$
$\mathbb{E}[Y|t = 1, Gender = M] = 10$

$\mathbb{E}[Y|t = 0, Gender = F] = 4$
$\mathbb{E}[Y|t = 1, Gender = F] = 6$



Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment assignment mechanism

- G=0 if gender=F,
  G=1 if gender=M

  $Y_0 = 4+4*G$
  $Y_1 = 4+4*G+2$

- $p(t=1|G=1) = 0.25$
  $p(t=1|G=0) = 0.75$

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Random Treatment Assignments

They work because it allows to get expectations from observations!

- Treatment is random: $(Y_0, Y_1) \perp\!\!\!\perp T$
- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1 | T = 1] =$
- $\mathbb{E}[Y_{obs} | T = 1]$

- Treatment is random: $(Y_0, Y_1) \perp\!\!\!\perp T$
- $\mathbb{E}[Y_0] =$
- $\mathbb{E}[Y_0 | T = 0] =$
- $\mathbb{E}[Y_{obs} | T = 0]$

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

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

$$\mathbb{E}[Y_{obs} | T = 1] - \mathbb{E}[Y_{obs} | T = 0]$$

Treatment assignment not random!

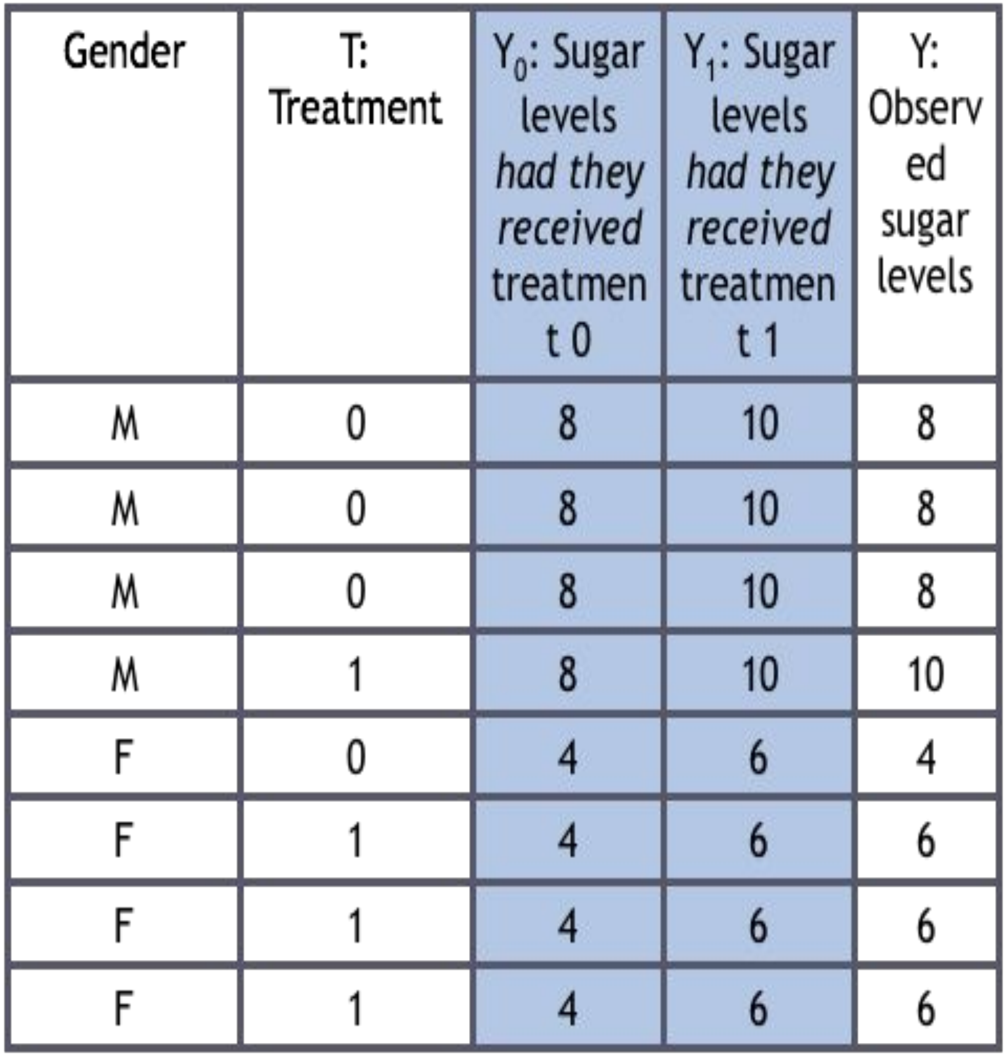
Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$P(Y_0 = 8|T = 0) = 0.75$$
$$P(Y_0 = 8|T = 1) = 0.25$$
$$P(Y_1 = 10|T = 0) = 0.75$$
$$P(Y_1 = 10|T = 1) = 0.25$$

(Y_0, Y_1) ***are not*** independent of T

Gender	T: Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$P(Y_0 = 4|T = 0, G = F) = 1$$
$$P(Y_0 = 4|T = 1, G = F) = 1$$
$$P(Y_1 = 6|T = 0, G = F) = 1$$
$$P(Y_1 = 6|T = 1, G = F) = 1$$



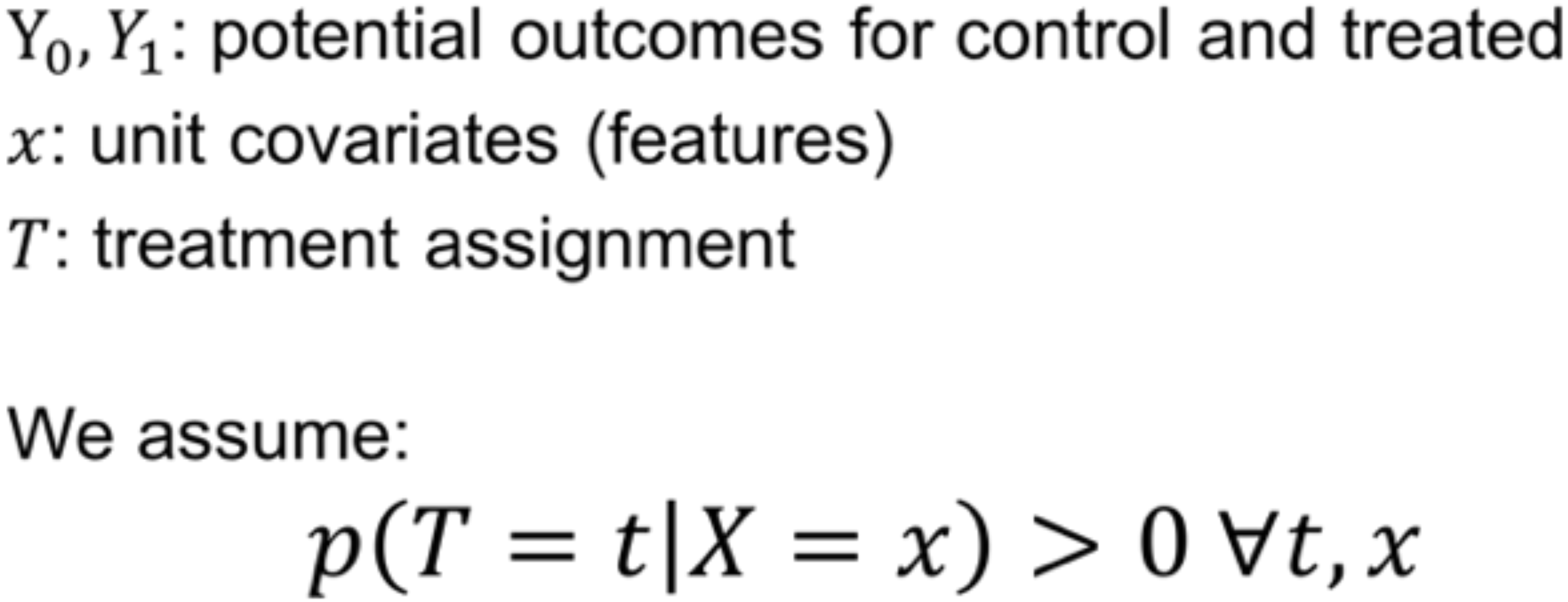
Gender	T: Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

(Y_0, Y_1) ***are independent*** of T ***conditioned*** on G=M, and conditioned on G=F

$$(Y_0, Y_1) \perp\!\!\!\perp T|G$$

No Unmeasured Confounding! Or Ignorability

Common support assumption



- Y_0, Y_1: potential outcomes for control and treated

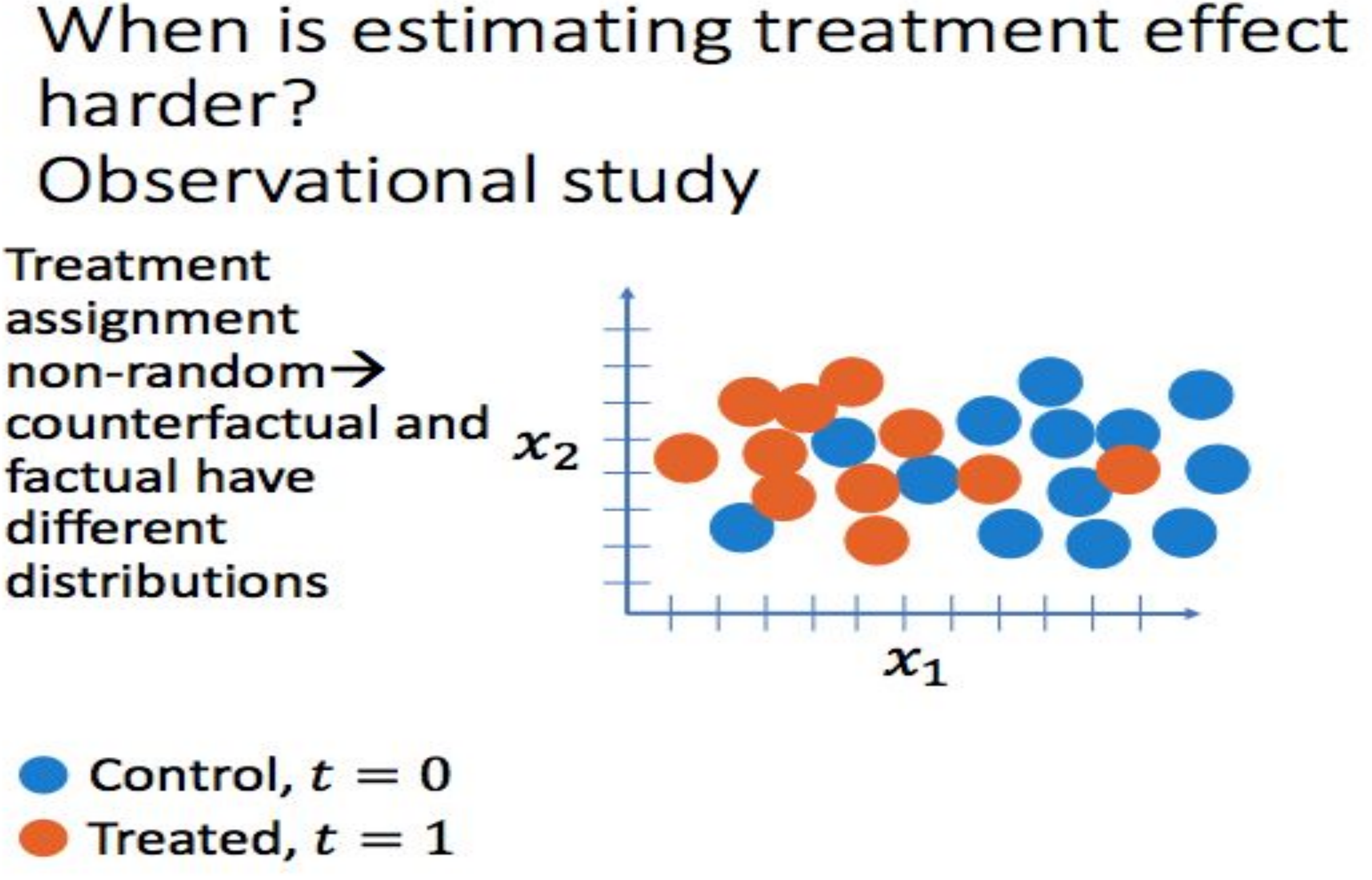
  x: unit covariates (features)

  T: treatment assignment

  We assume:

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

Propensity Score



When is estimating treatment effect harder?

Observational study

Treatment assignment non-random→ counterfactual and factual have different distributions

● Control, $t = 0$

● Treated, $t = 1$

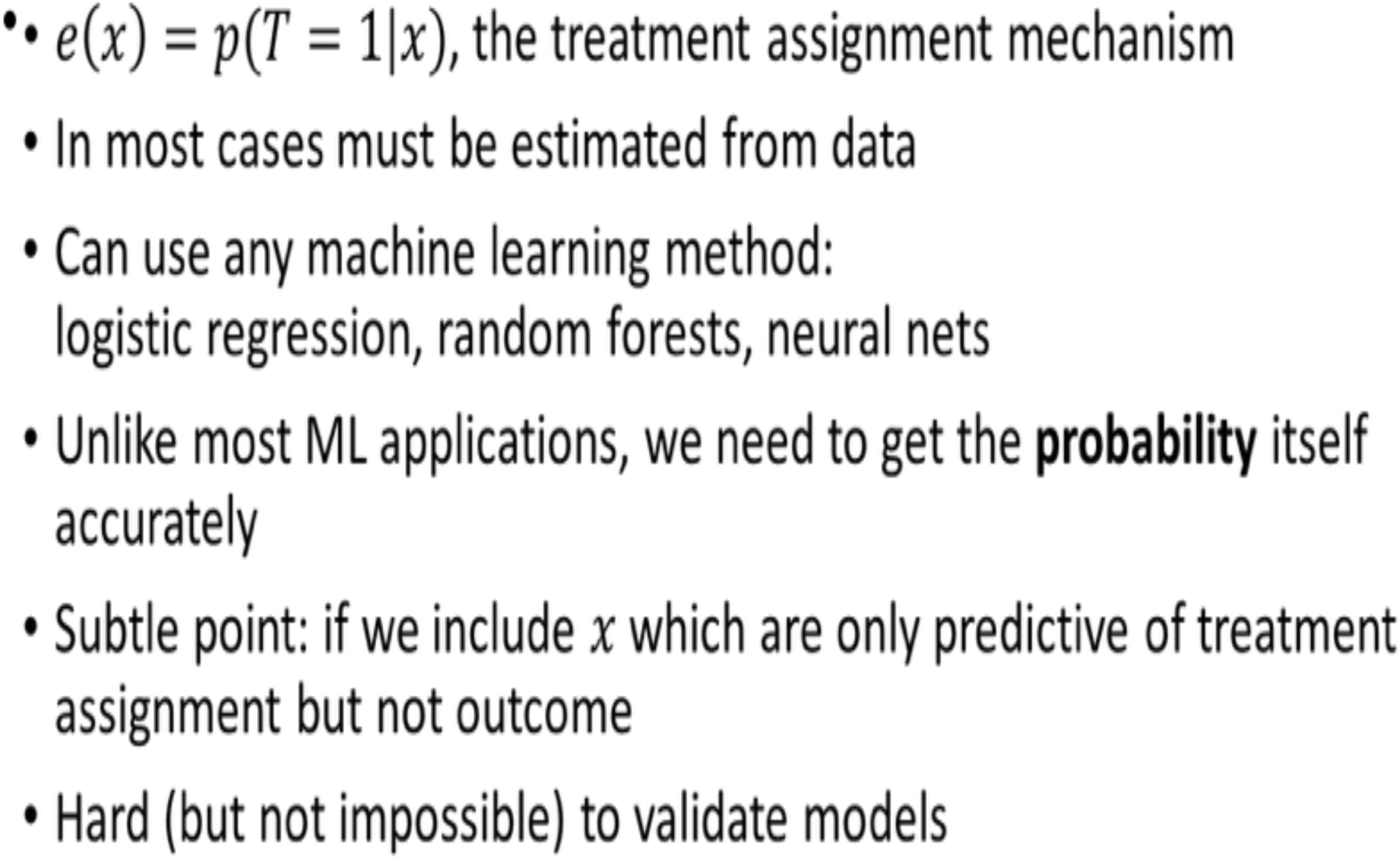
Propensity score

- Extremely widely used tool
- Basic idea: turn observational study into a pseudo-randomized trial by correcting for non-random sampling

Ignorability

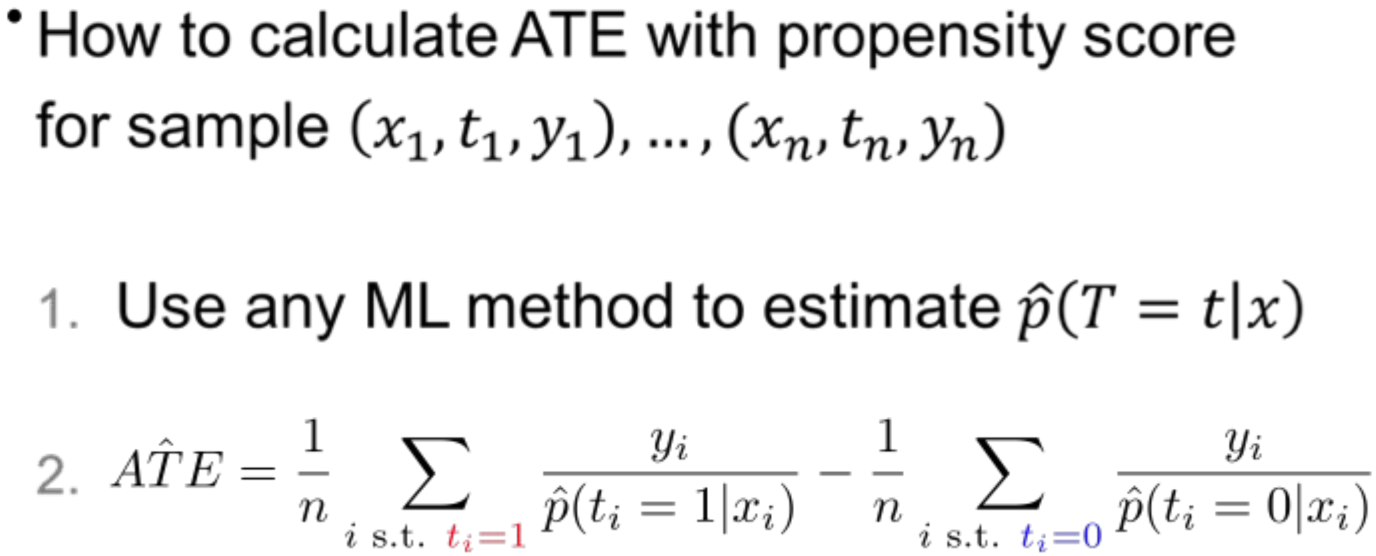
- $(Y_0, Y_1) \perp\!\!\!\perp T \mid x$
- What functions of $f(x)$ will still allow $(Y_0, Y_1) \perp\!\!\!\perp T \mid f(x)$?
- Theorem:
  Let $e(x) = p(T = 1|x)$, also called the ***propensity score.***
  If ignorability holds for x, then $e(x)$ is the coarsest function of x for which ignorability still holds

Propensity Score



- $e(x) = p(T = 1|x)$, the treatment assignment mechanism
- In most cases must be estimated from data
- Can use any machine learning method:
  logistic regression, random forests, neural nets
- Unlike most ML applications, we need to get the **probability** itself accurately
- Subtle point: if we include x which are only predictive of treatment assignment but not outcome
- Hard (but not impossible) to validate models

Propensity Score - Algorithm for ATE estimation

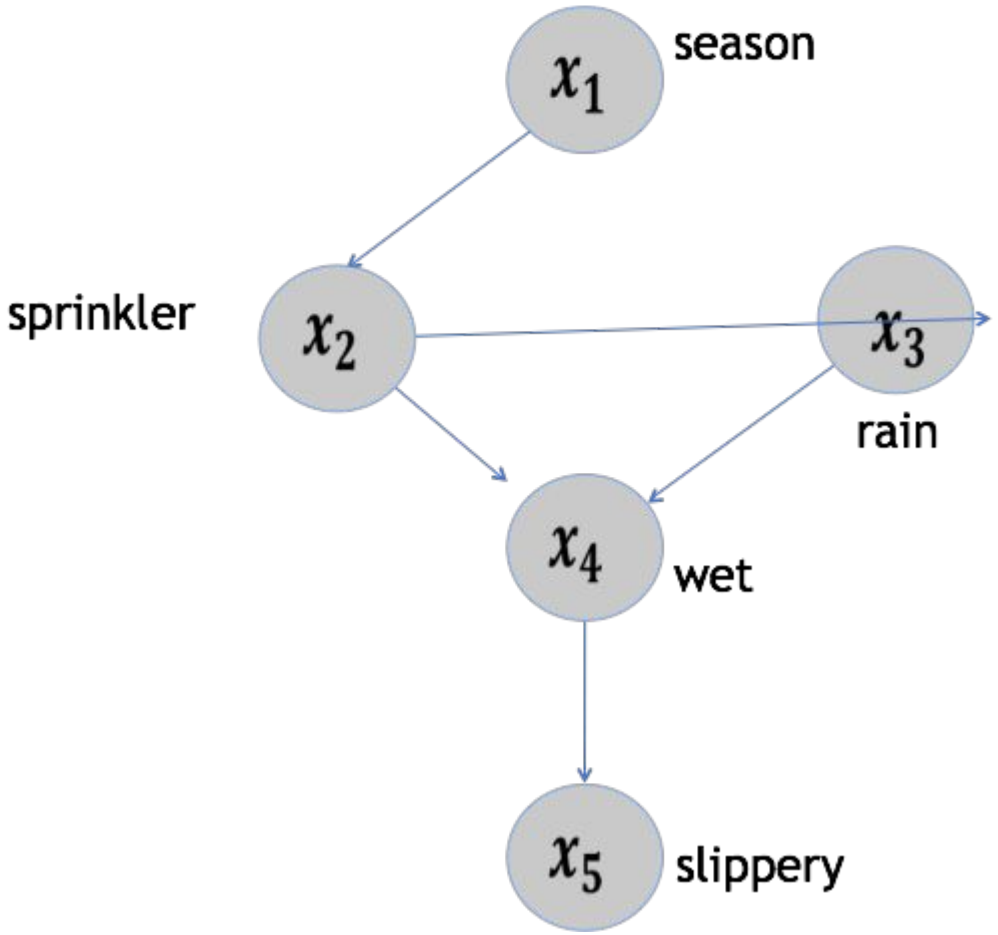


- How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

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

Not Covered: Propensity Score Matching

Pearlean Causal Framework



$$p(x_1, x_2, x_3, x_4, x_5) =$$
$$p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3, x_2)p(x_5|x_4)$$

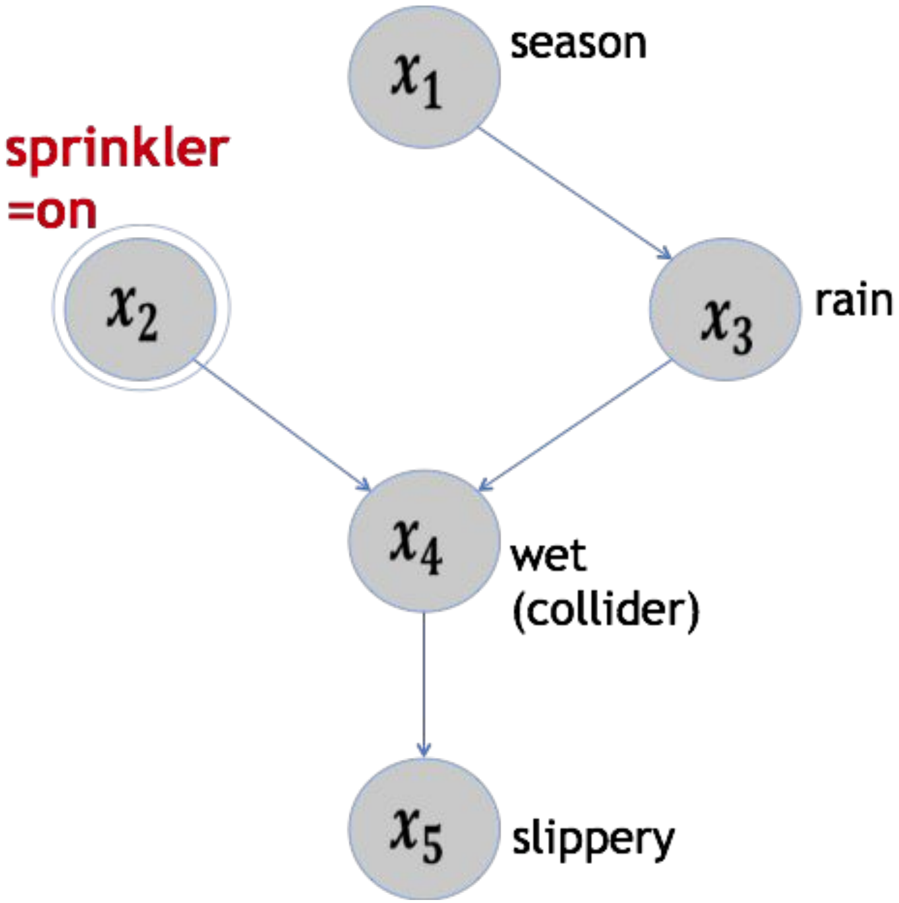
Intervention

- Turn the sprinkler on, please

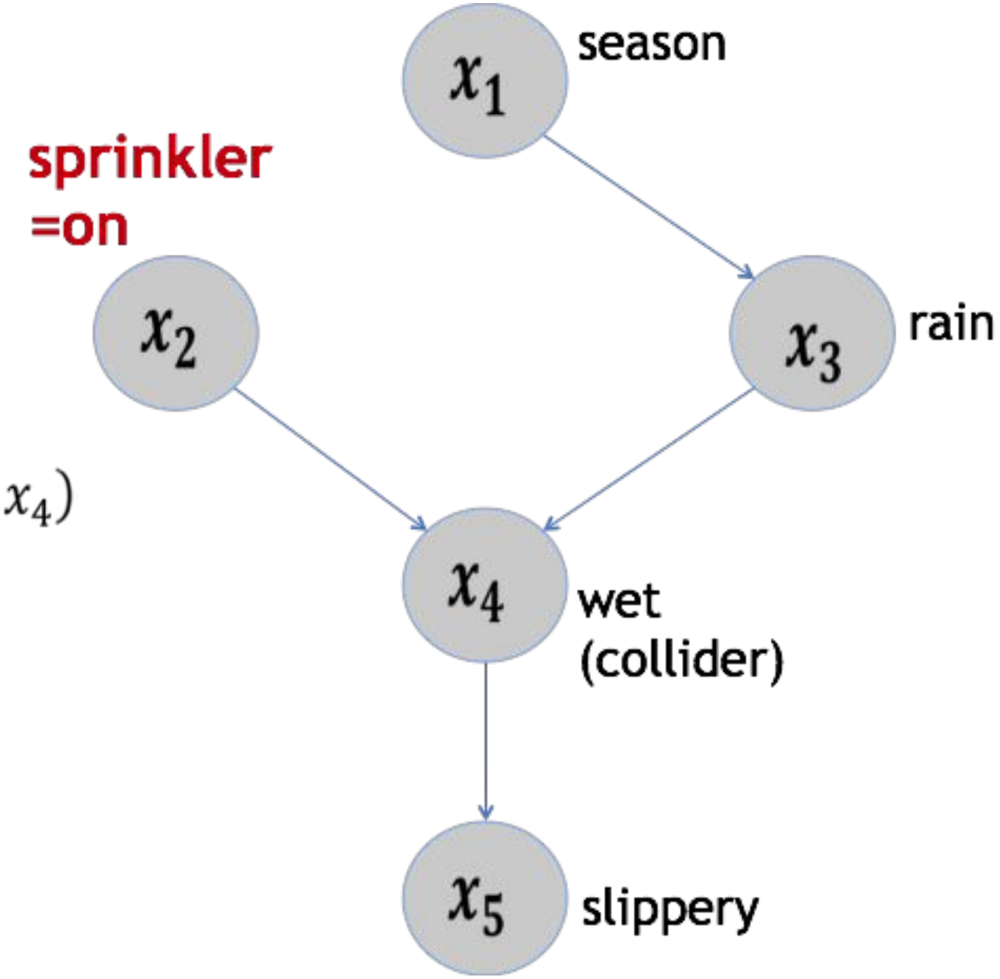
- We removed the association between season and sprinkler

- We are now in a new world, where the sprinkler is set to on

- This is the do-operator



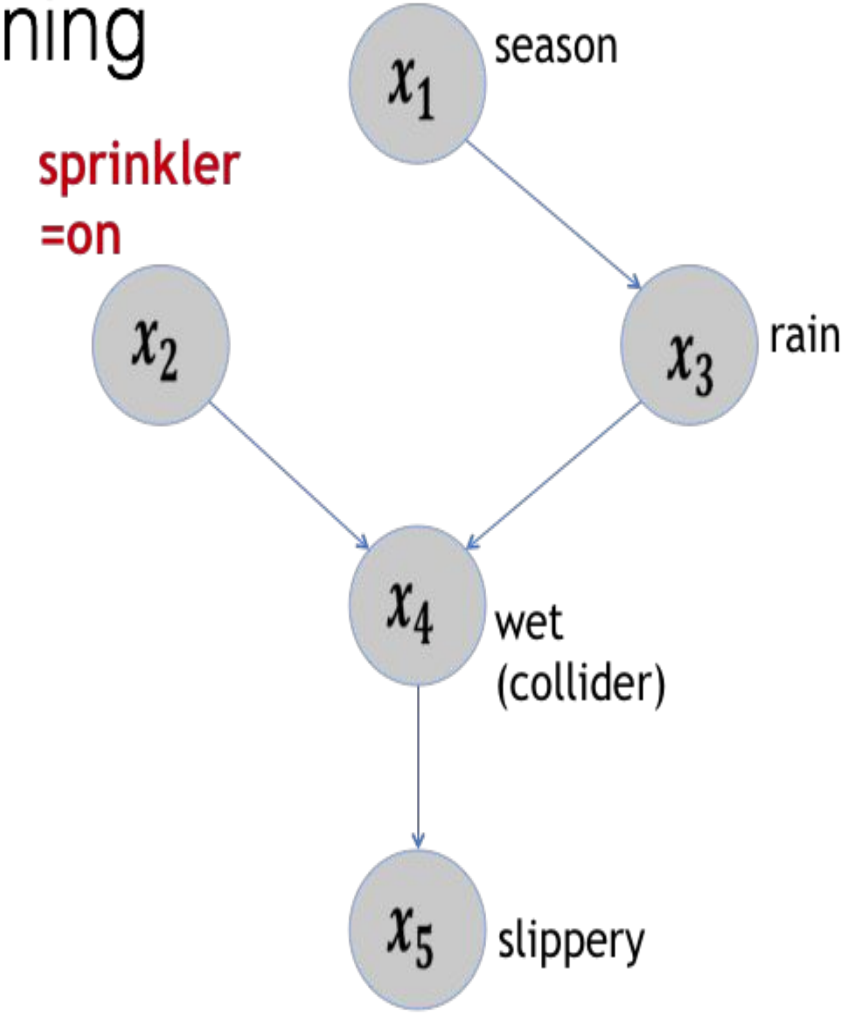
Intervention (do-Calculus)



- $p_{do(x_2=on)}(x_1, x_3, x_4, x_5\ ) = p(x_1)p(x_3|x_1)p(x_4|x_3, x_2 = on)p(x_5|x_4)$

- $p(x_1, x_3, x_4, x_5|x_2 = on) = p(x_1|x_2 = on)p(x_3|x_1, x_2 = on) \cdot p(x_4|x_3, x_2 = on)p(x_5|x_4, x_2 = on)$

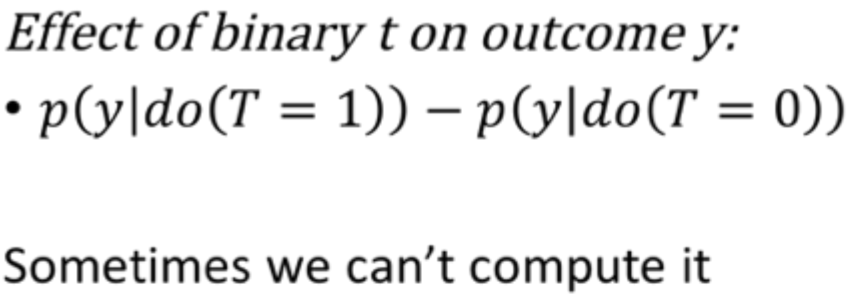
do-operator vs. conditioning



- $p(x_1, x_3, x_4, x_5 | do(x_2) = on)$
  distribution under an ***action***

- $p(x_1, x_3, x_4, x_5 | x_2 = on)$
  distribution given ***evidence***

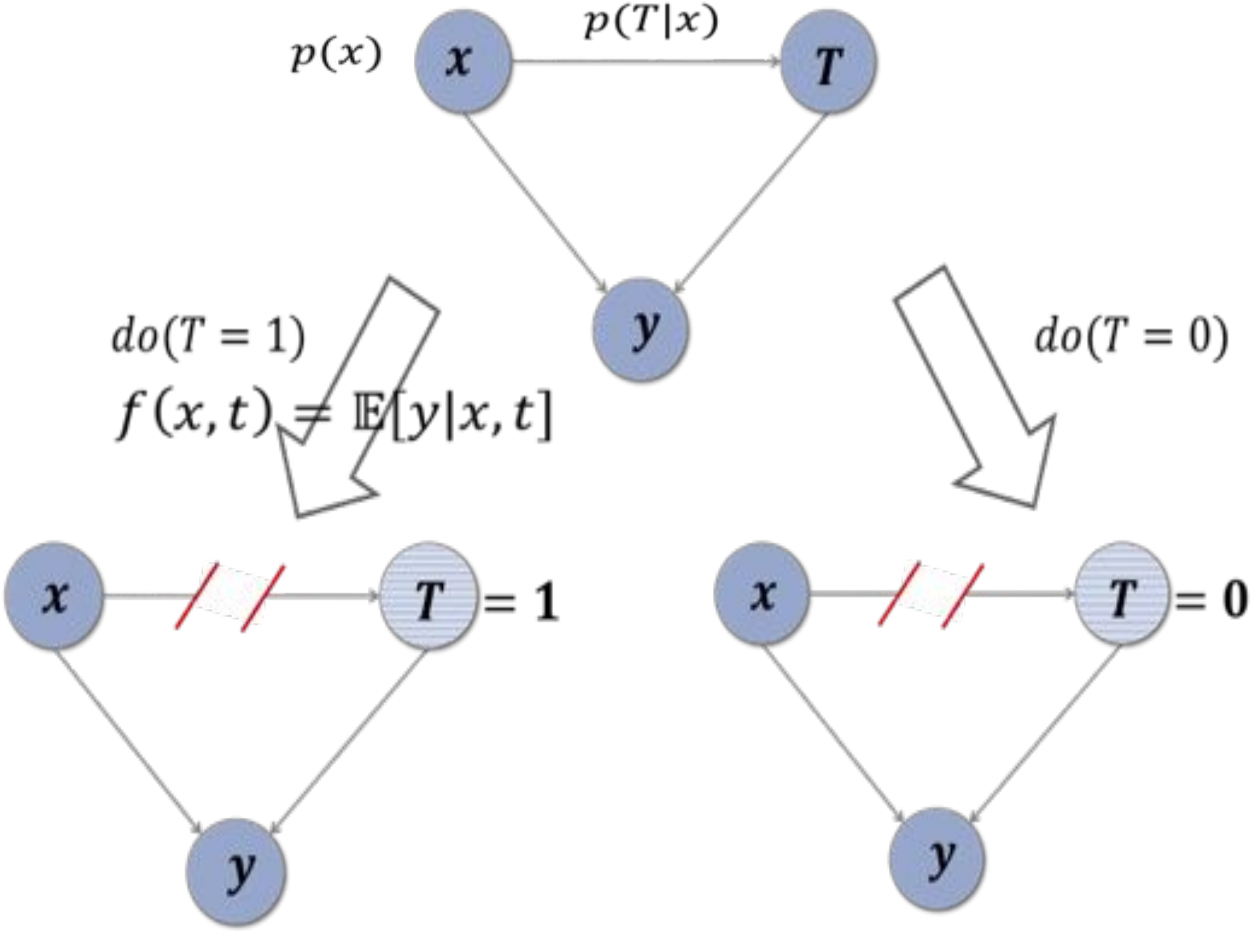
What is cause-effect here?



- *Effect of binary t on outcome y:*
  - $p(y|do(T=1)) - p(y|do(T=0))$

Sometimes we can't compute it

The *do* operator and adjustment formula



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

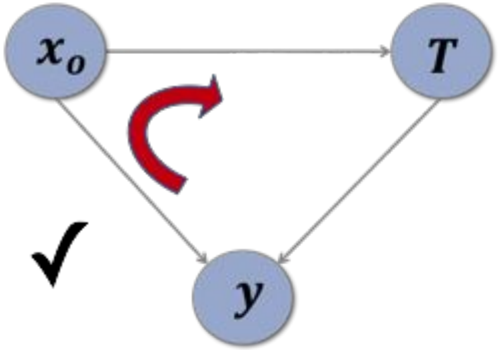
$$ATE \coloneqq \mathbb{E}[y|do(T = 1)] - \mathbb{E}[y|do(T = 0)] =$$

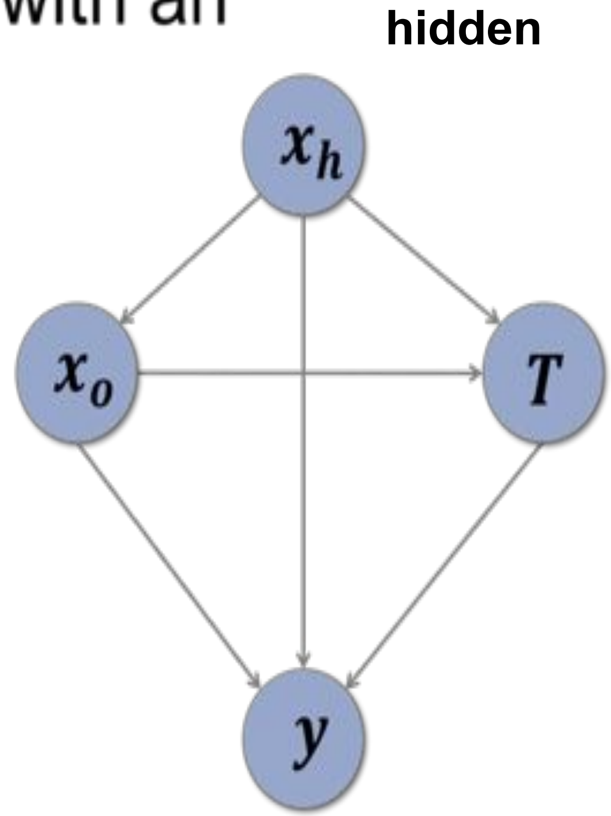
The Assumptions: causal identifiability

- 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

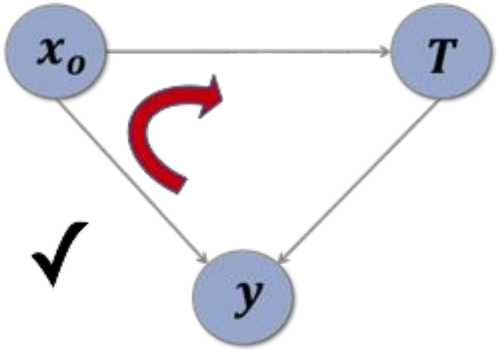
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

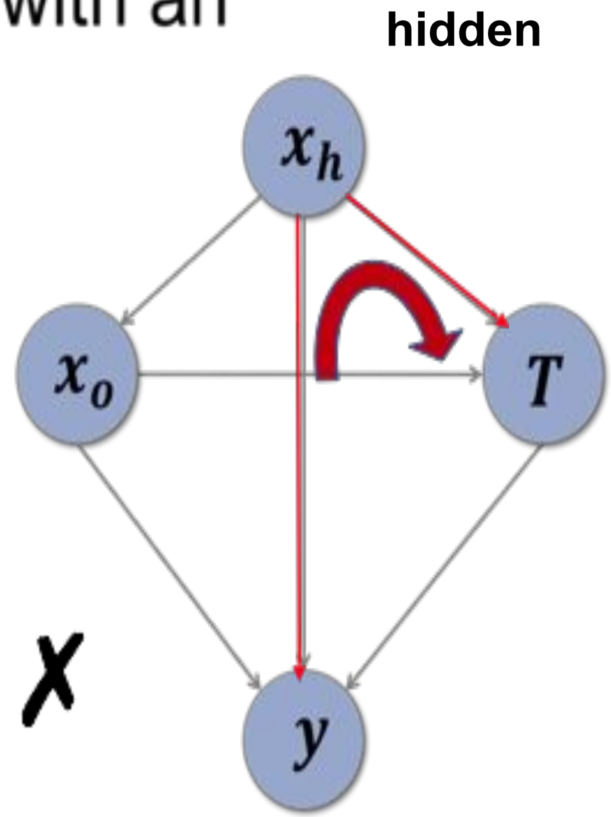




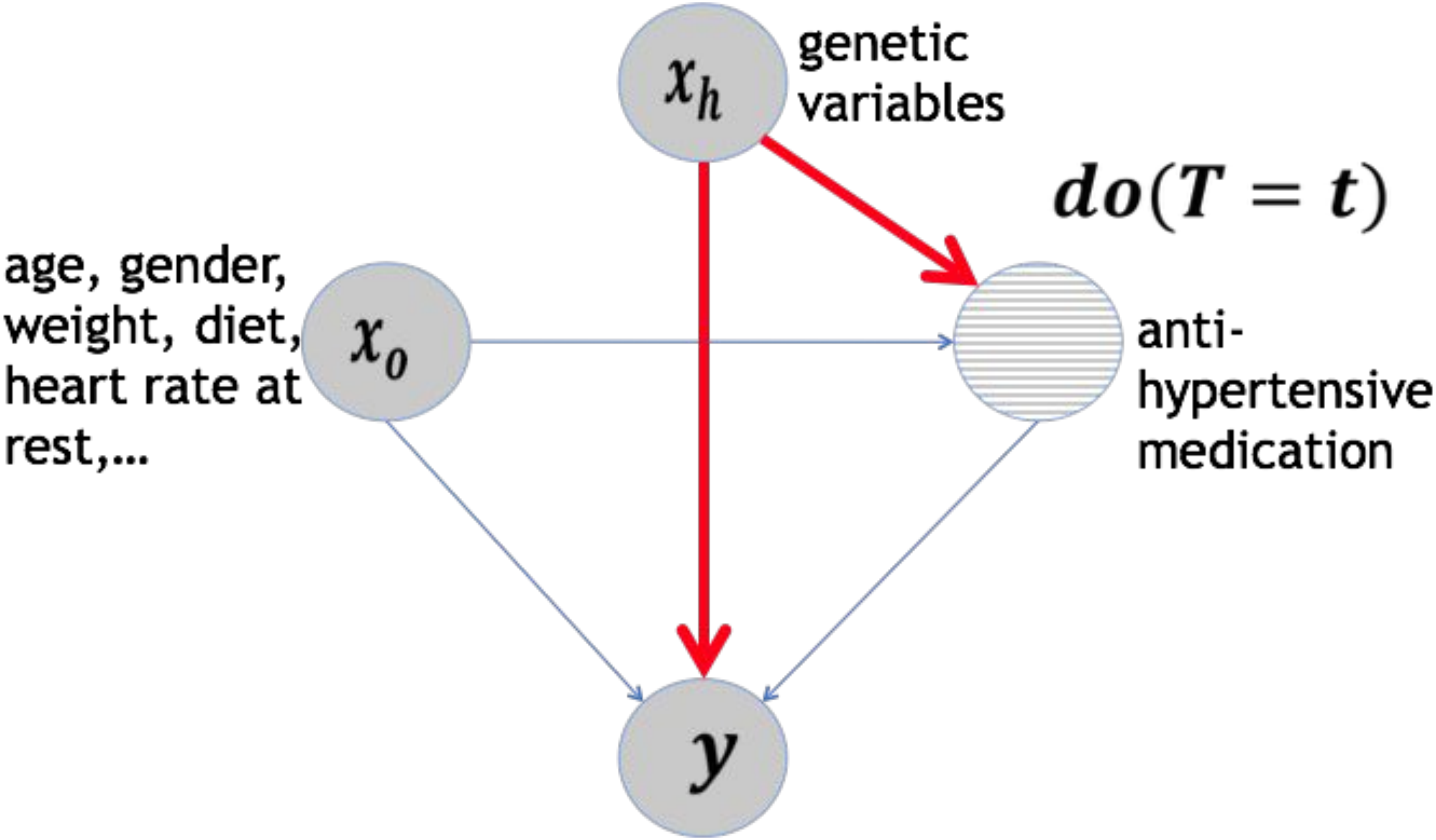
The Assumptions: causal identifiability

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





Unidentifiable Causal Effect



x_h

genetic variables

$\boldsymbol{do(T = t)}$

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

x_o

anti-hypertensive medication

y

Main Takeaways

- Supervised learning has limitations
- RCTs are expensive AND limited
- Ergo, think causally especially for clinical data
- Pearl’s and Rubin’s frameworks provide foundational formalism for causal effect estimation
- Not all effects are identifiable
- Most research questions cater to how to relax all the assumptions we made along the way!