



# Causality and Machine Learning (80-816/516)

*Classes 8 & 9* (Feb 6 & 11, 2025)

## Identification of Causal Effects (Causal Inference) & Counterfactual Reasoning

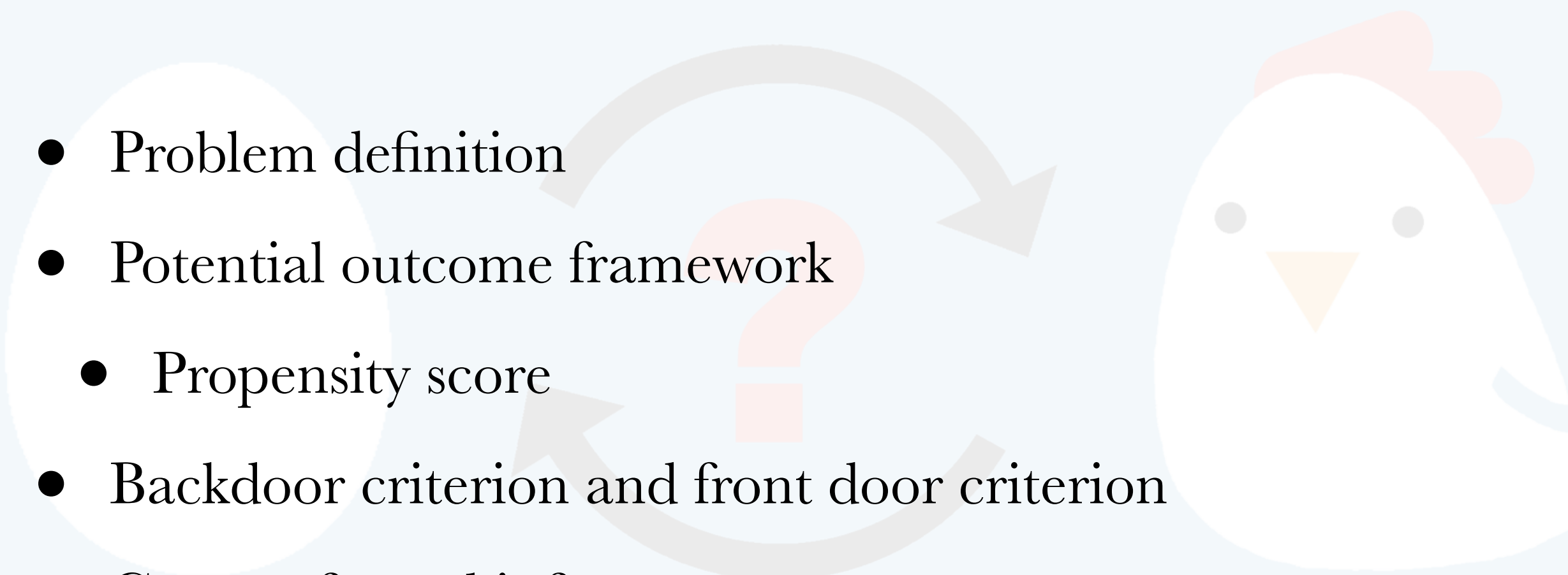
Instructor:

Kun Zhang ([kunz1@cmu.edu](mailto:kunz1@cmu.edu))

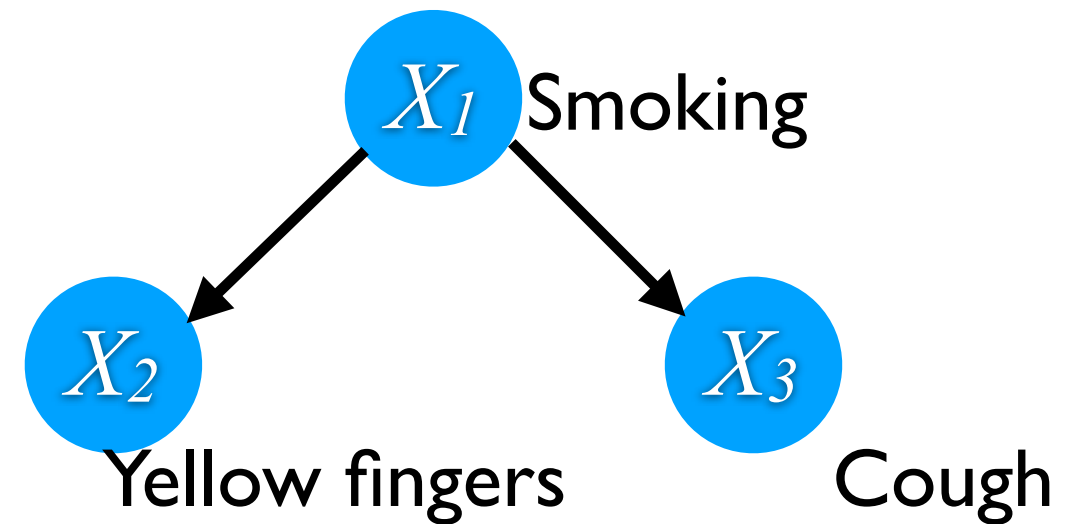
Zoom link: <https://cmu.zoom.us/j/8214572323>

Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by  
appointment

# Identification of Causal Effects & Counterfactual Inference: Outline

- Problem definition
  - Potential outcome framework
    - Propensity score
  - Backdoor criterion and front door criterion
  - Counterfactual inference
- 

# Three Types of Problems in Current AI



- Three questions:

$X_1$	$X_2$	$X_3$
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction:** Would the person cough if we *find* he/she has yellow fingers?

$$P(X_3 \mid X_2=1)$$

- **Intervention:** Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X_3 \mid \text{do}(X_2=1))$$

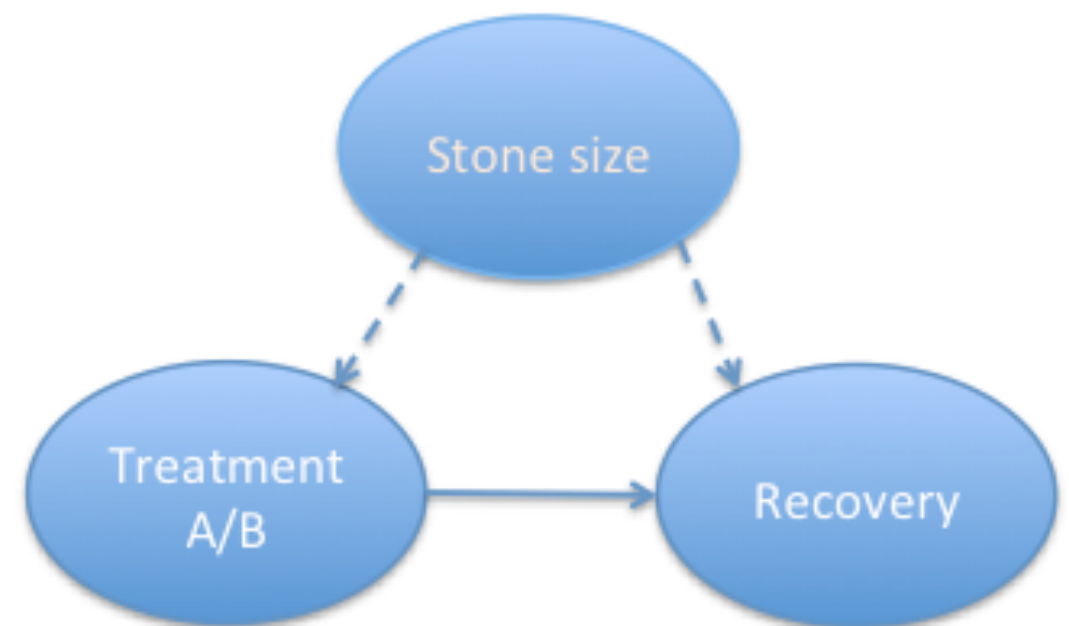
- **Counterfactual:** Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X_3_{X_2=1} \mid X_2=0, X_3=1)$$

# Identification of Causal Effects

$$P(\text{Recovery} \mid \text{do}(\text{Treatment}=A)) ?$$

- “Gold standard”: randomized controlled experiments
- **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable



- Usually expensive or impossible to do!

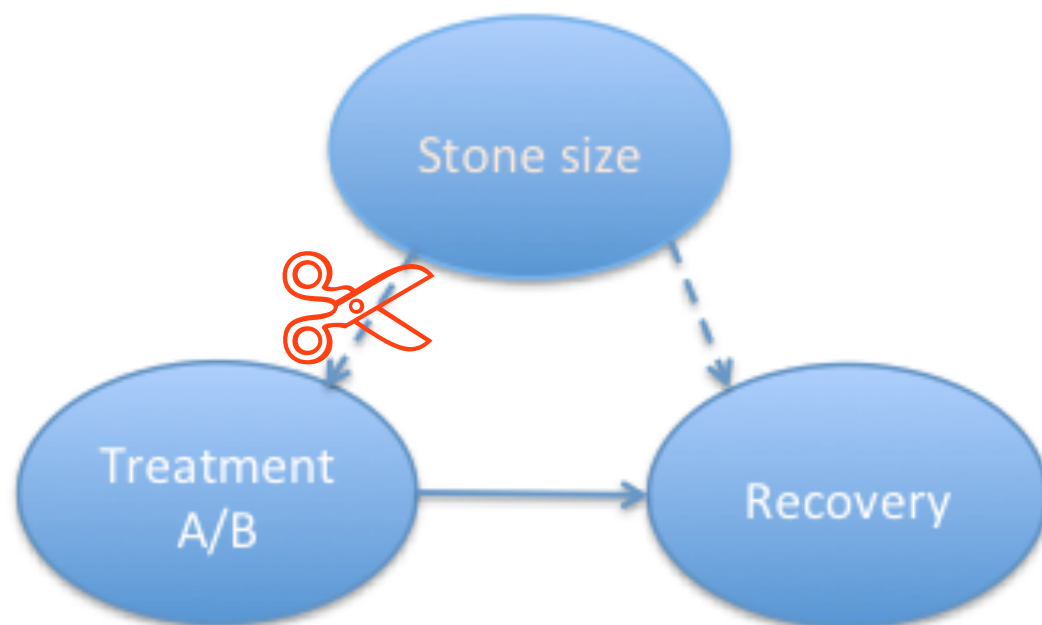


# Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large Stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)

$$P(R|T) = \sum_S P(R|T, S)P(S|T)$$

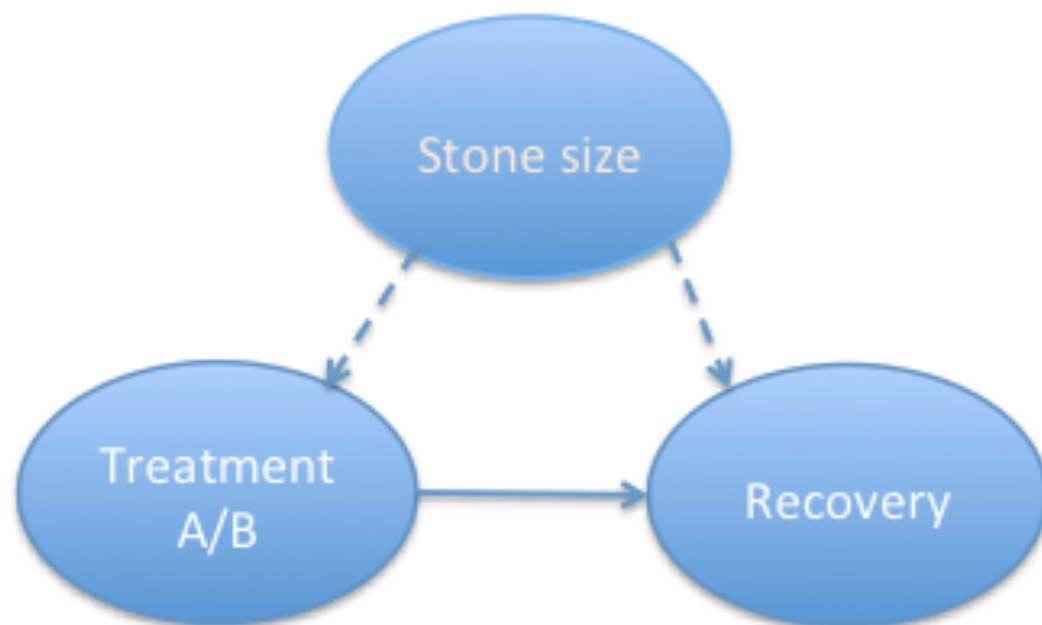
$$P(R | do(T)) = \sum_S P(R | T, S)P(S)$$



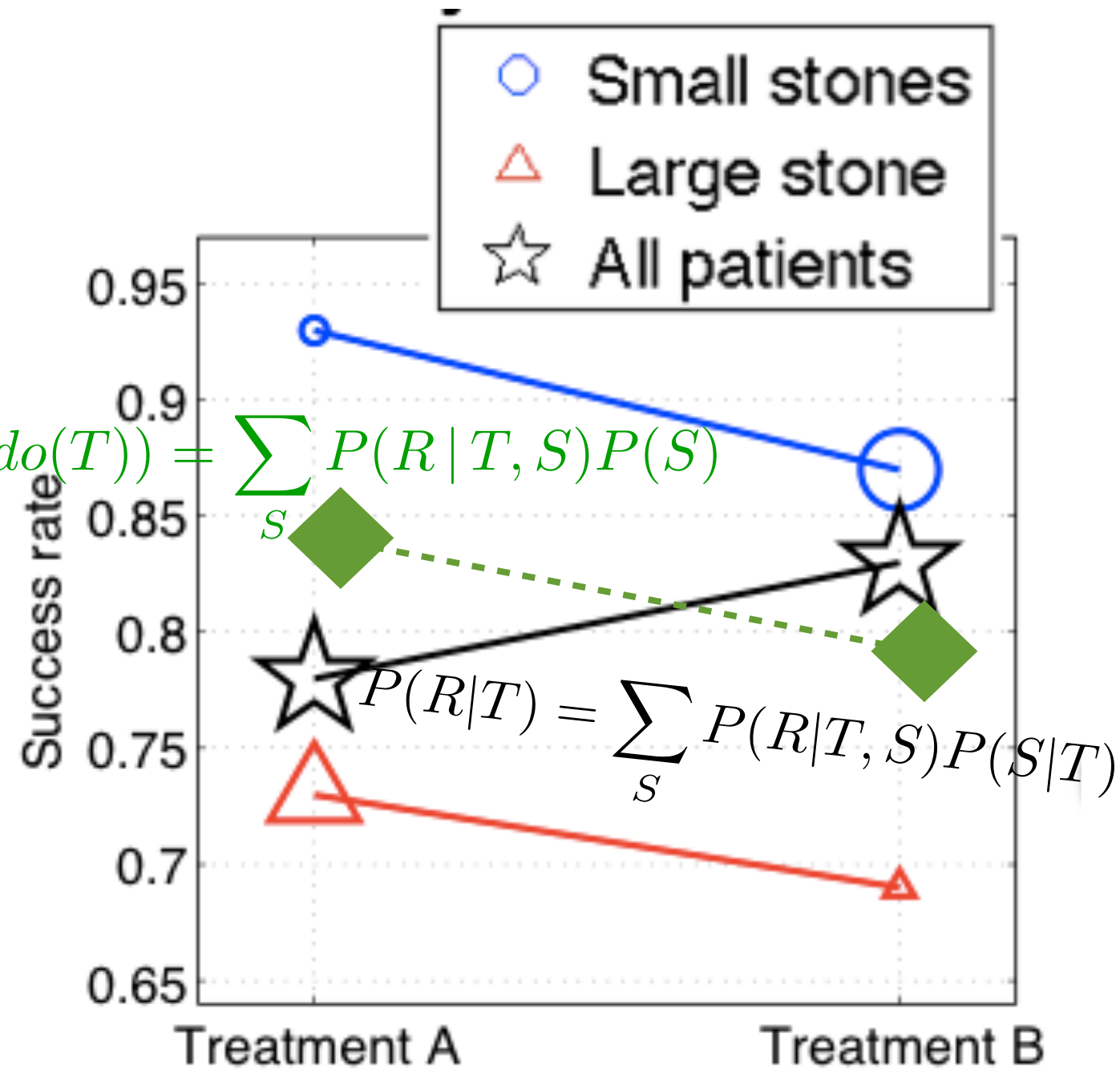
conditioning vs. **manipulating**

# Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large Stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)



$$P(R | do(T)) = \sum_S P(R | T, S) P(S)$$



conditioning vs. **manipulating**

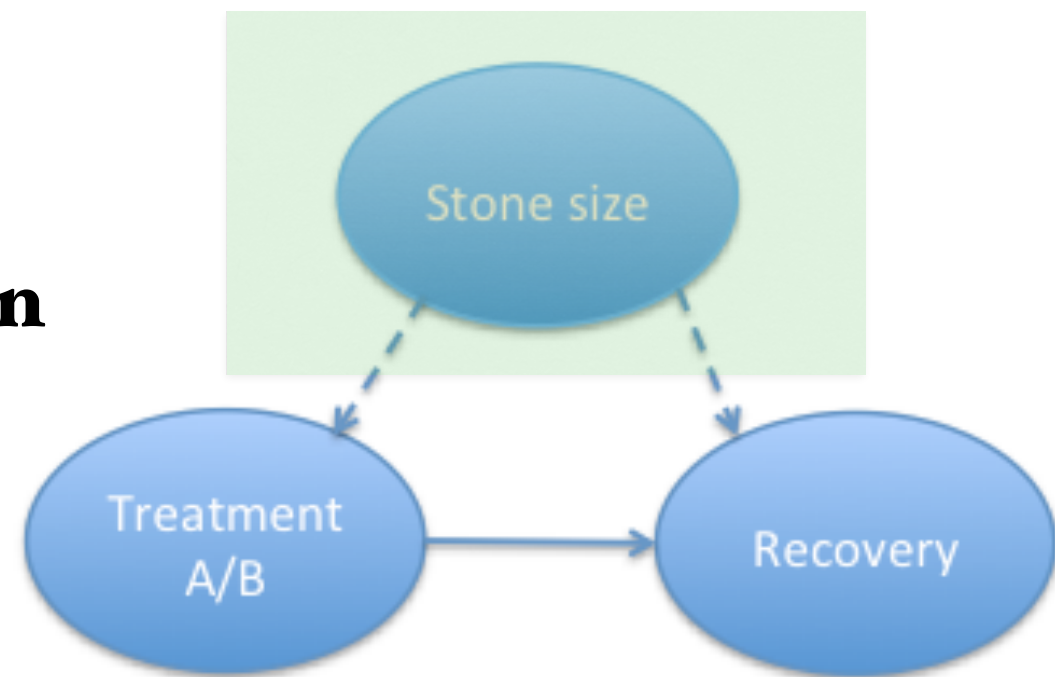
# Identifiability of Parameters in Statistical Models

- Identifiability, in simple words, means that different values of a parameter must produce different probability distributions.
- Mathematically, a parameter  $\theta$  is said to be identifiable if and only  
 $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$ , or equivalently  $P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$
- Is the mean of a Gaussian distribution identifiable?

# Identifiability of Causal Effects

Sometimes written as  $P(y | \hat{x})$

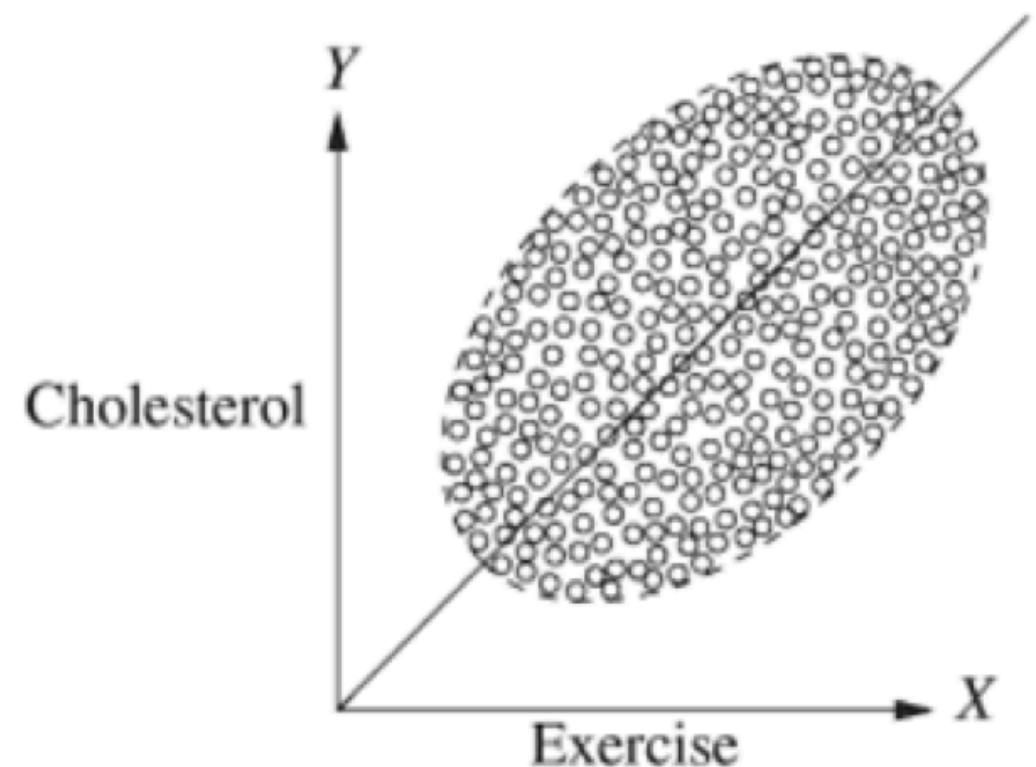
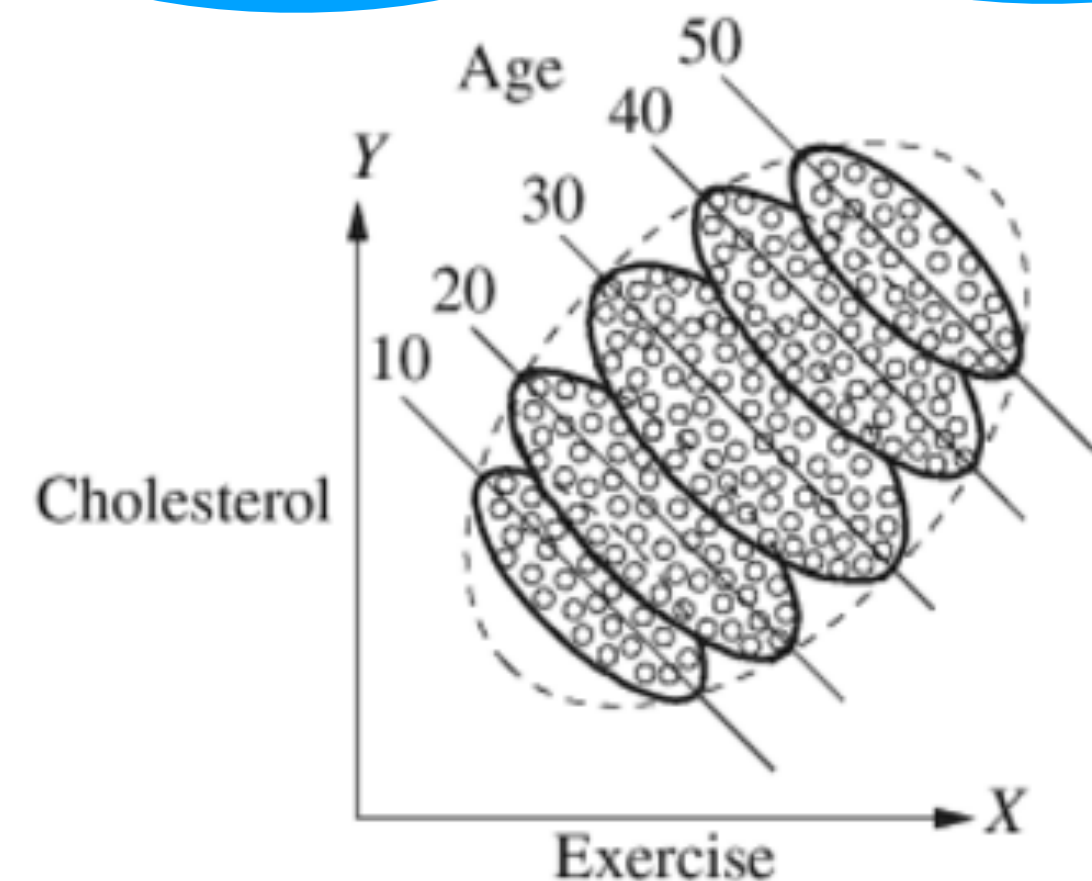
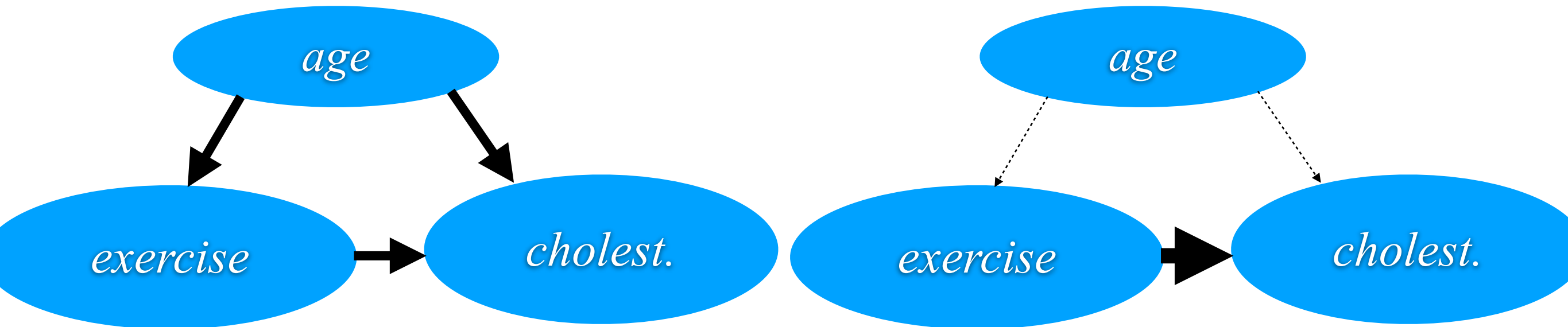
- Is causal effect, denoted by  $P(Y | do(X))$ , identifiable given complete or partial causal knowledge?
  - Two models with **the same causal structure** and **the same distribution for the observed variables** give the same causal effect?
- How?
- Key issue: Controlling confounding effects



*Examples: Average causal effect (ACE)...*

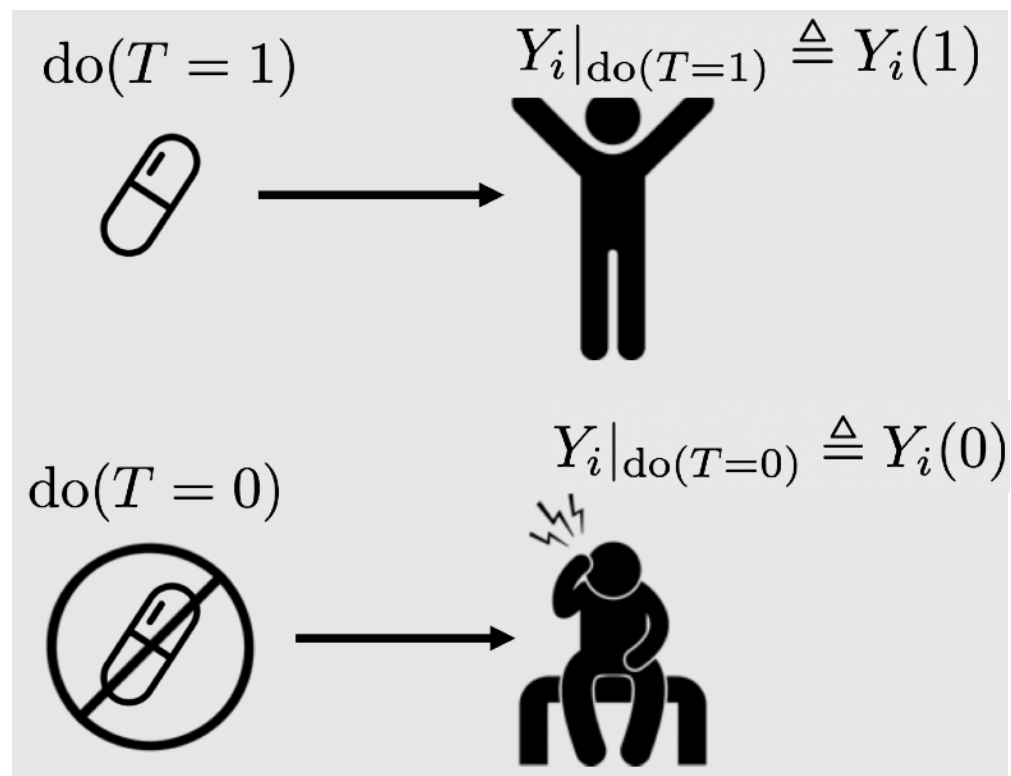
# Key Issue: Controlling Confounding Bias

- Exercise-cholesterol study: identifiable if age is not observed?



# Potential Outcome

- Causal inference: Inferring the effect of treatment/policy on some outcome



*(Adapted from Brady Neal, 2020)*

Causal effect:  
 $Y_i(1) - Y_i(0)$

$T$ : observed treatment

$Y$ : observed outcome

$i$ : denote a specific subject or unit

$Y_i(1)$ : **potential outcome if the patient had been treated**

$$Y_i |_{\text{do}(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$ : **potential outcome if the patient had not been treated**

$$Y_i |_{\text{do}(T=0)} \triangleq Y_i(0)$$

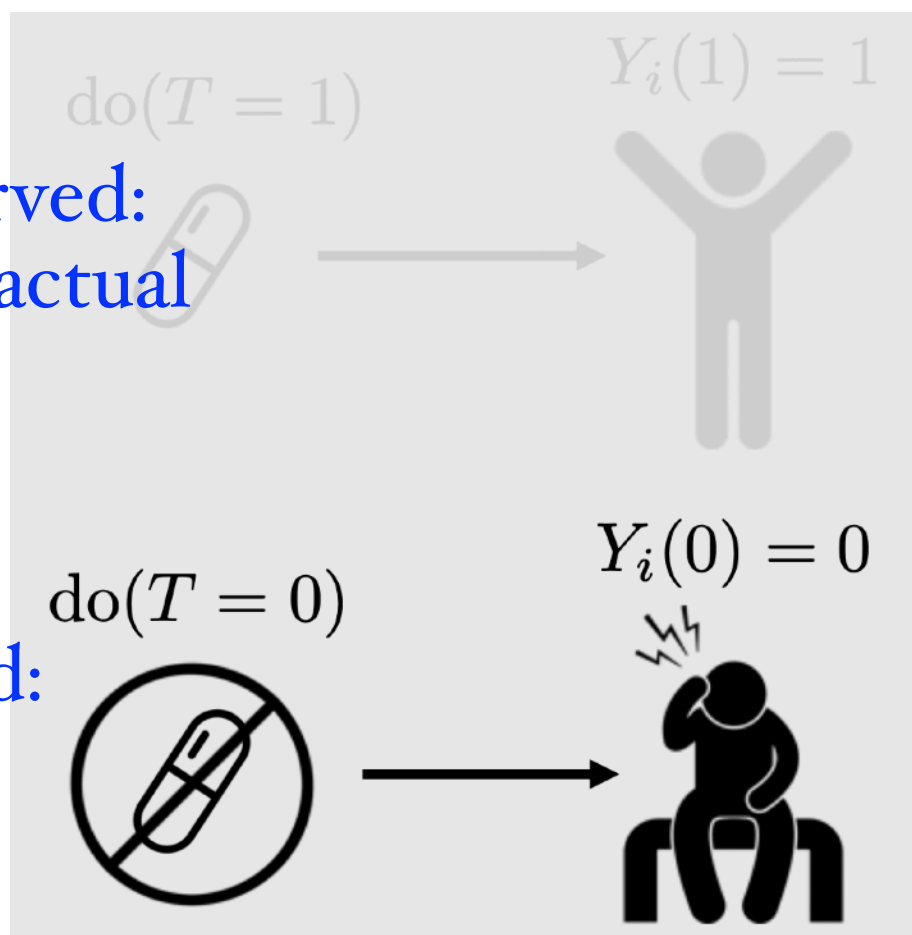


# Fundamental Problem of Causal Inference

- Missing data issue

Unobserved:  
Counterfactual

Observed:  
Factual



T: observed treatment

Y: observed outcome

i: denote a specific subject or unit

$Y_i(1)$ : potential outcome under treatment

$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$ : potential outcome without treatment

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

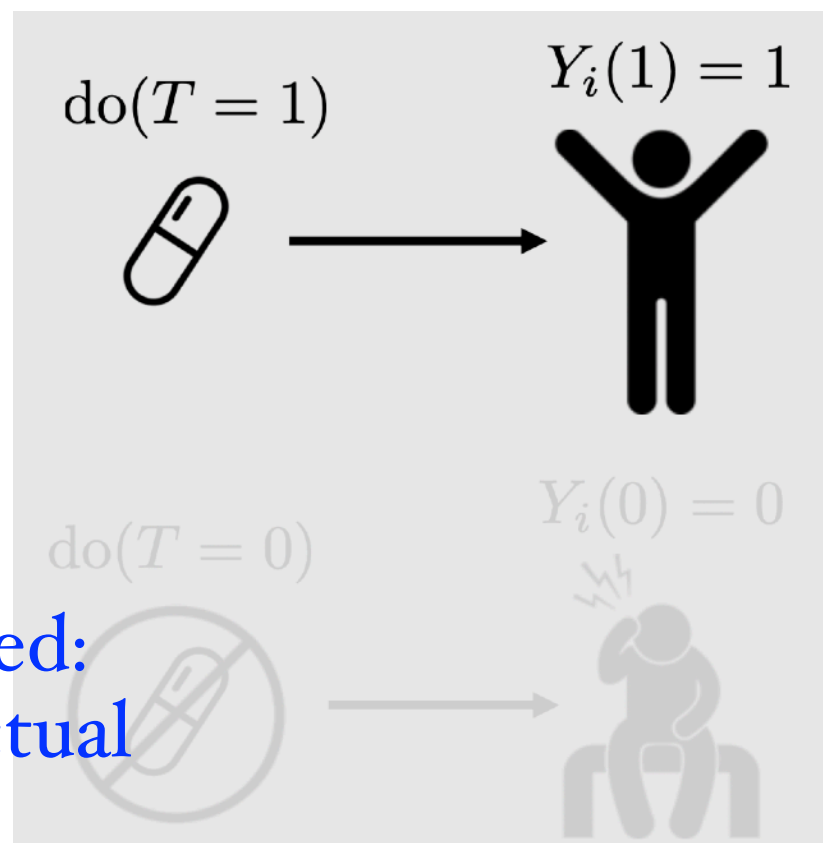
Causal effect:

$$Y_i(1) - Y_i(0)$$

# Fundamental Problem of Causal Inference

- Missing data issue

Observed:  
Factual



Unobserved:  
Counterfactual

T: observed treatment

Y: observed outcome

i: denote a specific subject or unit

Y<sub>i</sub>(1): potential outcome under treatment

$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

Y<sub>i</sub>(0): potential outcome without treatment

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

Causal effect:

$$Y_i(1) - Y_i(0)$$



# Fundamental Problem of Causal Inference

- Missing data issue

$i$	$T$	$Y$	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

$T$ : observed treatment

$Y$ : observed outcome

$i$ : denote a specific subject or unit

$Y_i(1)$ : potential outcome under treatment

$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$ : potential outcome without treatment

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

Causal effect:

$$Y_i(1) - Y_i(0)$$

# Formally, Potential Outcome Framework

- For a set of i.i.d. subjects  $i = 1, \dots, n$ , we observed a tuple  $(X_i, T_i, Y_i)$ , comprised of
  - A **feature vector**  $X_i \in \mathbb{R}^p$
  - A **treatment assignment**  $T_i \in \{0, 1\}$
  - A **response**  $Y_i \in \mathbb{R}$
- $Y_i(1)$  and  $Y_i(0)$  are **potential outcomes** in that they represent the outcomes for individual  $i$  had they received the treatment or control respectively.
- Missing data issue: we only get to see  $Y_i$ , with

$$Y_i = Y_i(T_i) = Y_i(0)(1 - T_i) + Y_i(1)T_i$$

# Potential Outcome Framework

- Our first goal is to estimate the **average treatment effect (ATE)**

$$\tau = E_i[Y_i(1) - Y_i(0)] \quad (\text{formally, } i \text{ should be dropped.})$$

- However, we cannot find  $Y_i(1) - Y_i(0)$  because of the unobserved potential outcome
- Then what assumptions do we need in order to estimate ATE from observational data?

# \* Assumptions in the Potential-Outcome Framework

*Assumptions that make the ATE be estimated from observational data*

- Ignorability:  $\{Y(0), Y(1)\} \perp\!\!\!\perp T$

Conditional ignorability:  $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$

- Positivity:  $0 < P(T = 1 \mid X = x) < 1$

- No interference:  $Y_i(t_1, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_n) = Y_i(t_i)$
- Consistency:  $T = t \implies Y = Y(t)$

} Stable Treatment Value Assumption (SUTVA)

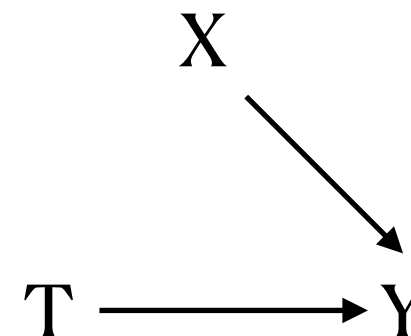


# Assumption 1: Ignorability

- The **ignorability** assumption:  $\{Y(0), Y(1)\} \perp\!\!\!\perp T$

*That is, the potential outcomes of subjects had they been treated or not does not depend on whether they have really been (observable) treated or not*

- Corresponding graphical model: there is no other path from  $T$  to  $Y$ , except the direct edge



- $ATE = E[Y(1)] - E[Y(0)]$   
 $= E[Y(1) | T = 1] - E[Y(0) | T = 0]$  (*ignorability*)

$= E[Y | T = 1] - E[Y | T = 0]$

 (*Consistency*)

Only contains observable moments



# Assumption 1: Ignorability

- The **ignorability** assumption:  $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$

$$E[Y(1)] - E[Y(0)] = E[Y(1) | T = 1] - E[Y(0) | T = 0]$$

$$= E[Y | T = 1] - E[Y | T = 0]$$

$$= 100.52 - 100.59$$

$Y_i(0)$	$Y_i(1)$	$\tau_i$
154.68	—	—
135.67	—	—
—	117.68	—
—	95.08	—
—	146.73	—
117.89	—	—
—	75.59	—
—	65.68	—
100.07	—	—
—	82.30	—
...	...	...
110.59	100.52	—

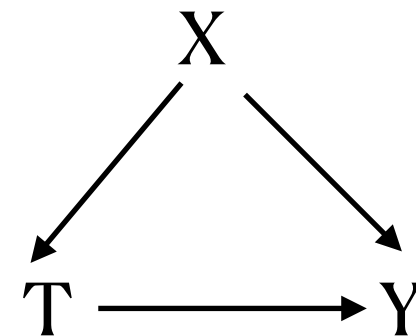


# Assumption 1: Conditional ignorability

- The **conditional ignorability** assumption:  $\{Y(0), Y(1)\} \perp\!\!\!\perp T | X$

*That is, given the covariates, the potential outcomes of subjects had they been treated or not does not depend on whether they have really been (observable) treated or not*

- Corresponding graphical model:  $X$  blocks all paths from  $T$  to  $Y$ , except the direct edge



- Conditional average treatment effect:

$$CATE = E[Y(1) - Y(0) | X]$$

$$= E[Y(1) | X] - E[Y(0) | X]$$

$$= E[Y(1) | T = 1, X] - E[Y(0) | T = 0, X] \quad (\text{Conditional ignorability})$$

$$= \boxed{E[Y | T = 1, X] - E[Y | T = 0, X]} \quad (\text{Consistency})$$

Only contains observable moments



# From CATE to ATE

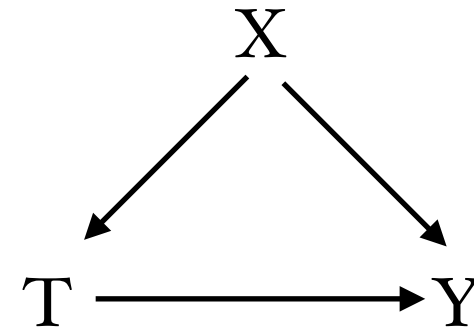
- Adjustment formula to identifying ATE

$$ATE = E[Y(1) - Y(0)]$$

$$= E_X E[Y(1) - Y(0) | X]$$

$$= E_X [E[Y | T = 1, X] - E[Y | T = 0, X]]$$

*We will see how to estimate it later (why?)*





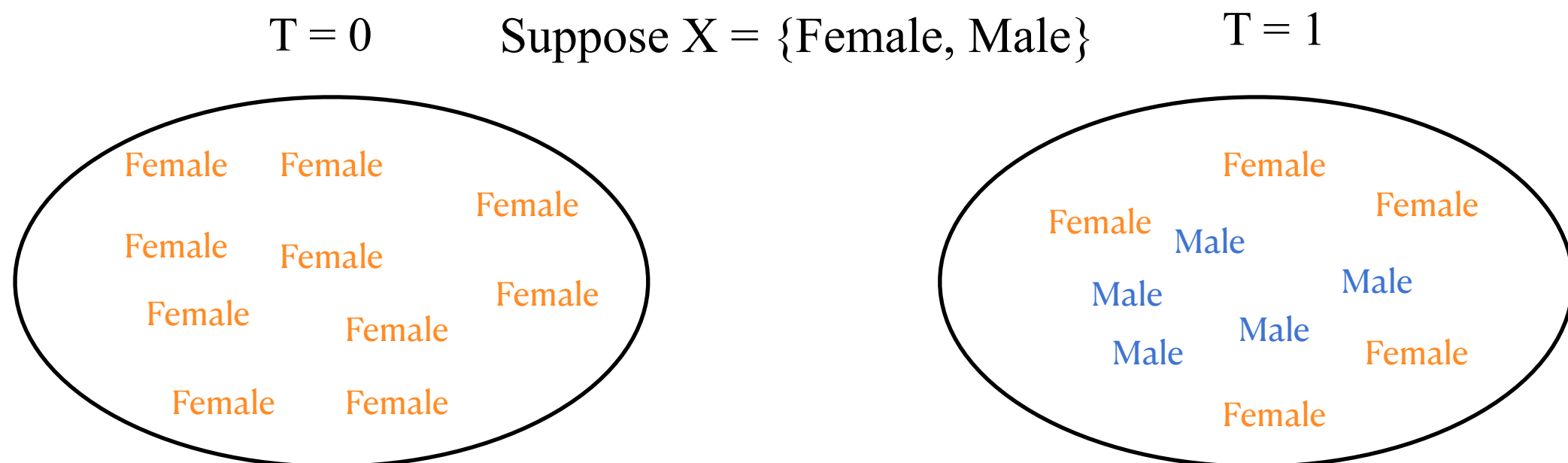
# Assumption 2: Positivity

- The **positivity** assumption

For all values of covariates  $x$  present in the population of interest (i.e.,  $x$  such that  $P(X = x) > 0$ ),

$$0 < \underline{P(T = 1 | X = x)} < 1$$

- A case where the positivity assumption violates



# Assumption 3: No Interference

- The **no interference** assumption: treatments of other units do not affect one's potential outcome, so

$$Y_i(t_1, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_n) = Y_i(t_i)$$

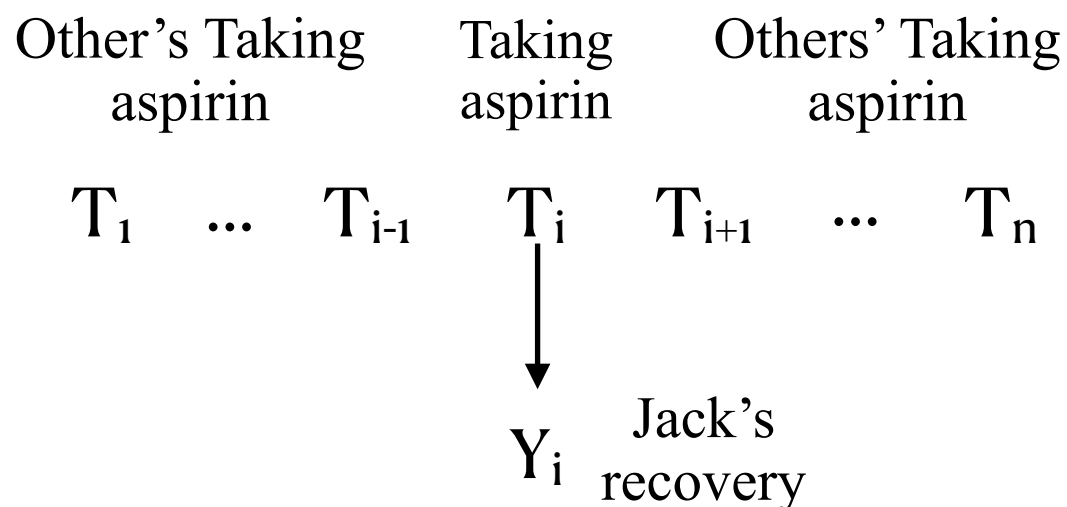
*That is, unit  $i$ 's potential outcome is only a function of its own treatment, but will not be affected by other units' treatment*

- *A case where the assumption holds:*

Jack's recovery is not affected by others' taking aspirin.

- *Violation:*

Job training for too many people may flood the market with qualified job applicants (interference)

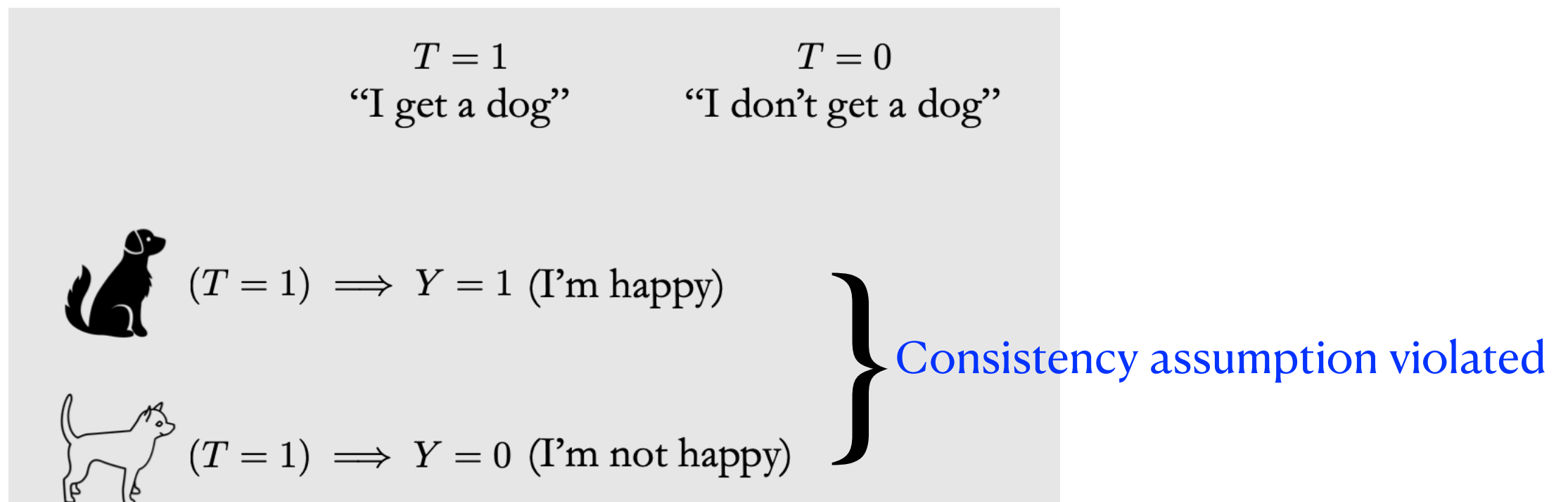


# Assumption 4: Consistency

- The **consistency** assumption: the potential outcome under treatment  $T=t$ ,  $Y(t)$ , is equal to the observed outcome if the actual treatment received is  $T=t$ , i.e.,

$$T = t \implies Y = Y(t), \text{ for all } t$$

*That is, the observed outcome is equal to the potential outcome  $Y(t)$ , when the actual treatment received is  $T=t$ ; there is no variation in treatment*



*(Adapted from Brady Neal, 2020)*

# Recall the Assumptions

*Assumptions that make the ATE be estimated from observational data*

- Ignorability:  $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$   
Conditional ignorability:  $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$
  - Positivity:  $0 < P(T = 1 \mid X = x) < 1$
  - No interference:  $Y_i(t_1, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_n) = Y_i(t_i)$
  - Consistency:  $T = t \implies Y = Y(t)$
- } Stable Unit Treatment Value Assumption (SUTVA)

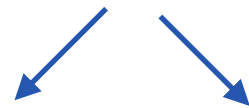
Stable Unit Treatment Value Assumption (SUTVA): No interference assumption + Consistency assumption

SUTVA allows to write potential outcome for the  $i^{\text{th}}$  person in terms of only that person's treatments



# Derivation of ATE

No interference:



$$ATE = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)] \quad (\text{Linearity of expectation})$$

$$= E_X[E[Y(1) | X] - E[Y(0) | X]] \quad (\text{Law of iterated expectations})$$

$$= E_X[E[Y(1) | T = 1, X] - E[Y(0) | T = 0, X]] \quad (\text{Ignorability and Positivity})$$

$$= E_X[E[Y | T = 1, X] - E[Y | T = 0, X]] \quad (\text{Consistency})$$

*We will see how to estimate it later (why?)*



# Estimands, Estimates, and Estimation

- Estimand: any quantity we want to estimate
  - Causal estimand (e.g.  $E[Y(1) - Y(0)]$ )
  - Statistical estimand (e.g.  $E_X[E[Y | T = 1, X] - E[Y | T = 0, X]]$ )
- Estimate: approximation of some estimand, using data
- Estimation: process for getting from data + estimand to estimate

## *The Identification-Estimation Flowchart*



# Example: Effect of Sodium Intake on Blood Pressure

Data (Epidemiological example taken from Luque-Fernandez et al. (2018)):

- Outcome Y: (systolic) blood pressure (continuous)
- Treatment T: sodium intake (1 if above 3.5 mg and 0 if below)
- Covariates X: age and amount of protein excreted in urine
- Simulation: so we know the “true” ATE is 1.05

## Estimation of ATE

True ATE:  $\mathbb{E}[Y(1) - Y(0)] = 1.05$

Identification:  $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]]$

Estimation:  $\frac{1}{n} \sum_x \underbrace{\mathbb{E}[Y | T = 1, x]}_{\text{Model (linear regression)}} - \underbrace{\mathbb{E}[Y | T = 0, x]}_{\text{Model (linear regression)}}$  *Why?*

Estimate: 0.85  $\frac{|0.85 - 1.05|}{1.05} \times 100\% = 19\%$

Naive:  $\mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0]$

Naive estimate: 5.33  $\frac{|5.33 - 1.05|}{1.05} \times 100\% = 407\%$

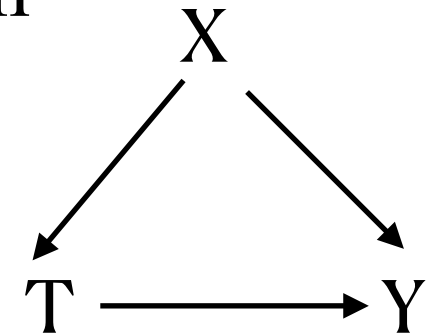
*(Adapted from  
Brady Neal, 2020)*



# How to estimate causal effect in the Presence of confounders?

- In many cases, treatment assignment is associated with covariates  $X$ , so

$$E[Y(1) - Y(0)] \neq E[Y | T = 1] - E[Y | T = 0]$$

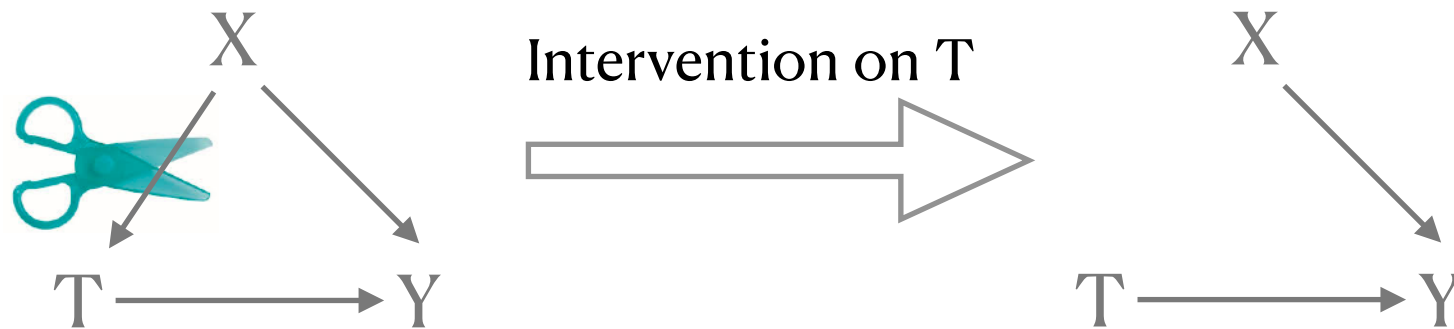


- We want to estimate average treatment effects with associational difference
  - How can we flexibly “control” for  $X$
  - Under what conditions is “controlling” for  $X$  enough
- The **Assumption**: Controlling for  $X$  is enough if treatment is as good as random
- The **Question**: What methods enable **inference** about the ATE given this assumption

# How to Estimate Causal Effect With Confounders?

## 1) Randomization

$$E[Y(1) - Y(0)] = E[Y | T = 1] - E[Y | T = 0]$$



## 2) Statistical adjustment

$$ATE = E_X[E[Y | T = 1, X] - E[Y | T = 0, X]]$$

# Covariates Adjustments

$$ATE = E_X[E[Y | T = 1, X] - E[Y | T = 0, X]]$$

- Regression adjustments
- Matching
  - Mahalanobis distance matching
  - Propensity Score matching
- Inverse propensity score reweighting
- Doubly robust method

# Covariates Adjustments

$$ATE = E_X[E[Y | T = 1, X] - E[Y | T = 0, X]]$$

- **Regression adjustments**
- Matching
  - Mahalanobis distance matching
  - Propensity Score matching
- Inverse propensity score reweighting
- Doubly robust method

# Regression Adjustments

- Regression adjustments under ignorability / unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$$

- We can express the ATE in terms of conditional response,

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

$$= E[E[Y_i(1) \mid X_i] - E[Y_i(0) \mid X_i]]$$

$$= E[E[Y_i(1) \mid T_i = 1, X_i] - E[Y_i(0) \mid T_i = 0, X_i]]$$

$$= E[E[Y_i \mid T_i = 1, X_i] - E[Y_i \mid T_i = 0, X_i]]$$

$$= E[\mu_{(1)}(X_i)] - E[\mu_{(0)}(X_i)]$$

*(“i” should be dropped; it is kept for clarity.)*

where  $\mu_{(t)}(x) = E[Y_i \mid T_i = t, X_i = x]$

# Regression Adjustments

- Given ignorability, we have  $\tau = E[\mu_{(1)}(X_i)] - E[\mu_{(0)}(X_i)]$ ,  
with  $\mu_{(t)}(x) = E[Y_i | X_i = x, T_i = t]$ 
  - Fit  $\hat{\mu}_t(x)$  via linear regression
  - Fit  $\hat{\mu}_t(x)$  via non-parametric approach
- One may use the following estimation strategy
  1. Learn  $\hat{\mu}_0(x)$  by predicting  $Y$  from  $X$  on controls
  2. Learn  $\hat{\mu}_1(x)$  by predicting  $Y$  from  $X$  on treated units
  3. Estimate  $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$

$\hat{\tau}$  is consistent if  $\hat{\mu}_t(x)$  is consistent for  $\mu_t(x)$ ...

# Regression Adjustments with Linear Regression

- A classical approach to the ATE is to estimate  $\mu_0(x)$  and  $\mu_1(x)$  via ordinary least-squares regression (OLS)

- We first posit a linear model,

$$\mu_{(w)}(x) = x\beta_{(w)}$$

and estimate  $\hat{\beta}_{(0)}$  and  $\hat{\beta}_{(1)}$  by fitting the model

- Finally, we make predictions  $\hat{\mu}_{(w)}(x) = x\hat{\beta}_{(w)}$ , and the estimated ATE is

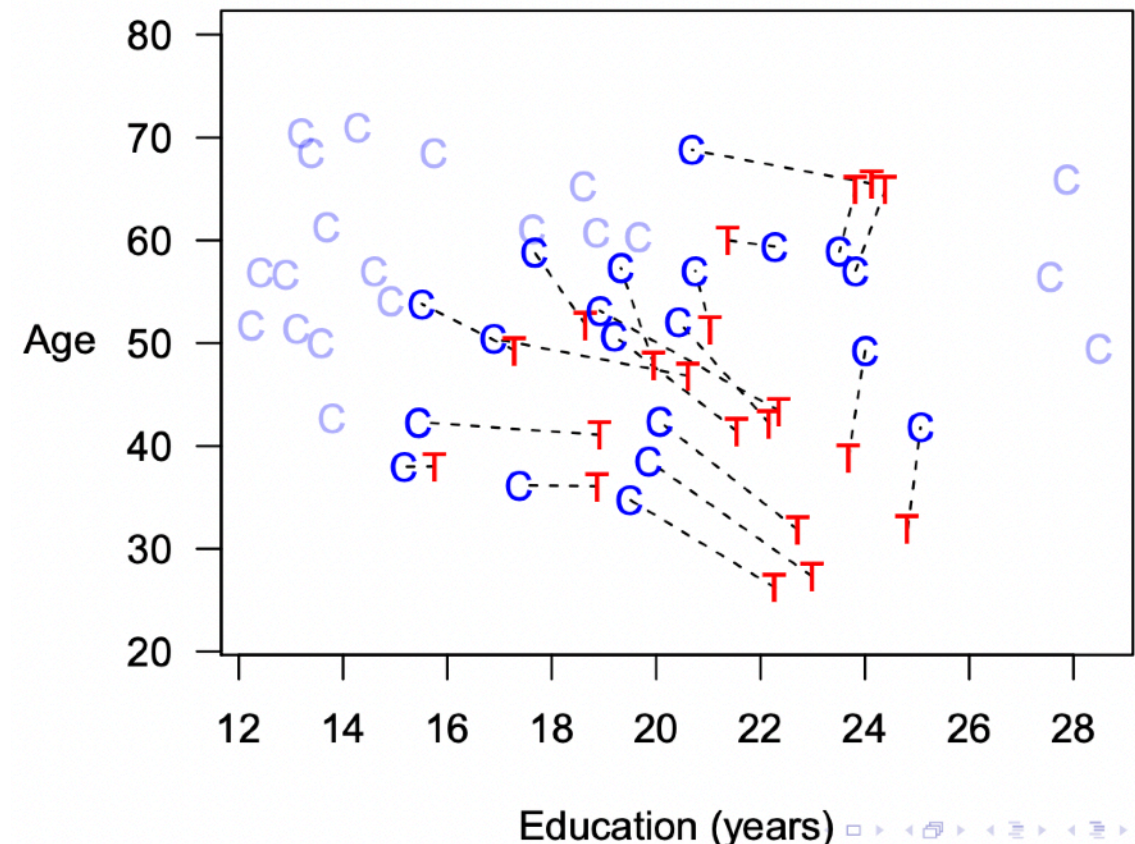
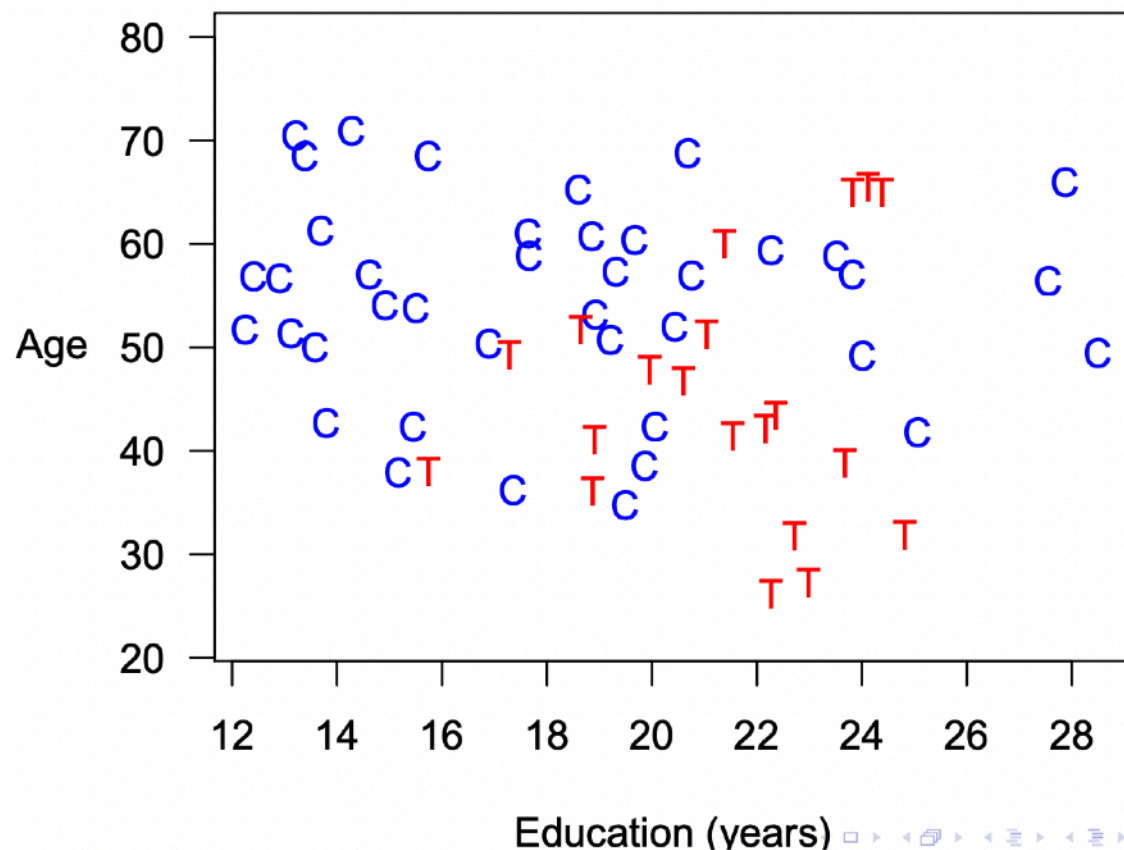
$$\begin{aligned}\hat{\tau} &= \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)) \\ &= (\hat{\beta}_{(1)} - \hat{\beta}_{(0)})\bar{X}\end{aligned}$$

# Matching 1: Mahalanobis Distance Matching

- **Mahalanobis distance matching:** match the feature of each treated unit to the nearest control unit, with the distance

$$D(X_i, X_j) = \sqrt{((X_i - X_j)^T S^{-1} (X_i - X_j))}$$

- Control units: pruned if unused
- Prune matches if distance > threshold





# Propensity Score

- The propensity score measures the probability of being treated conditionally on  $X_i$ , i.e.,

$$e(x) = P(T_i = 1 | X_i = x)$$

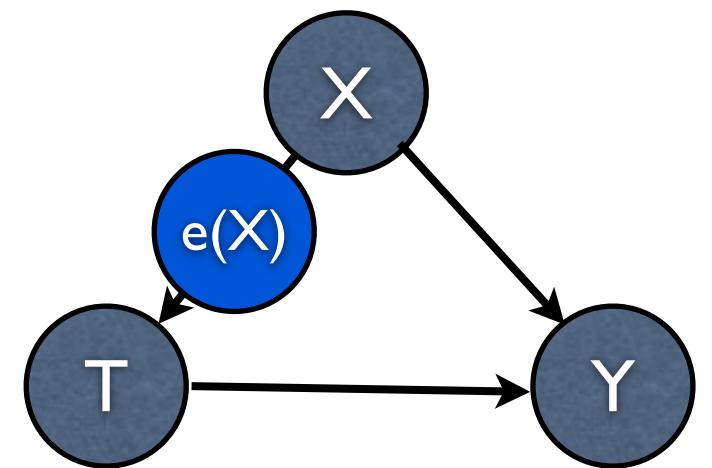
- In a randomized trial, the propensity score is constant

$$e(x) = e_0 \in (0,1)$$

- At least qualitatively, the variability of the propensity score gives a measure of how far we are from a randomized trial

# Propensity Score Matching

- One way is to match covariates  $X$ , *but it is hard especially for high-dimensional  $X$*
- Propensity Score
  - Let  $e(X) = P(T=1 | X)$ ;  $T \perp\!\!\!\perp X | e(X)$
  - Then  $e(X)$  and  $X$  are (confounding)-equivalent
  - $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i | X_i \implies \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i | e(X_i)$ 
    - Unconfoundness given  $X$  implies unconfoundness given  $e(X)$
    - $X$  may be high-dimensional, while  $e(X)$  is one-dimension



# Propensity Score Matching: Equivalence

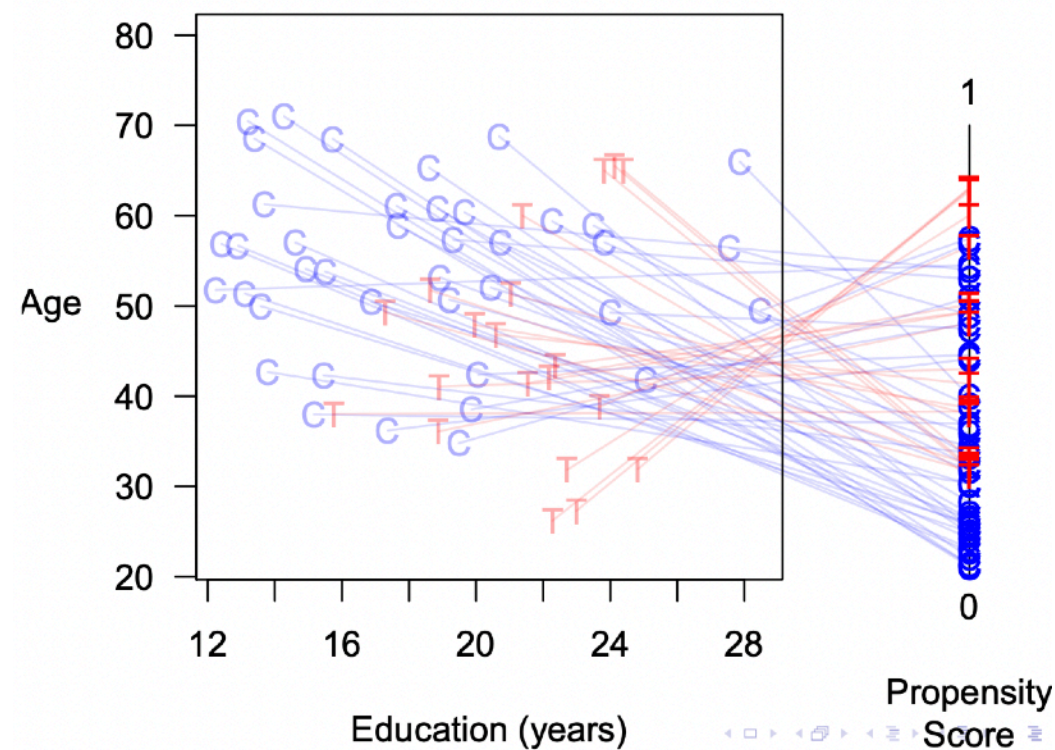
- Propensity Score
- Let  $e(X) = P(T=1 | X)$ ;  $T \perp\!\!\!\perp X | e(X)$
- Then  $e(X)$  and  $X$  are (confounding)-equivalent:

$$\begin{aligned}\sum_x P(Y|t, x)P(x) &= \sum_x \sum_e P(Y|t, x)P(e)P(x|e) \\ &= \sum_x \sum_e P(Y|t, x, e)P(e)P(x|t, e) = \sum_x \sum_e P(Y, x|t, e)P(e) \\ &= \sum_e P(Y|t, e)P(e)\end{aligned}$$

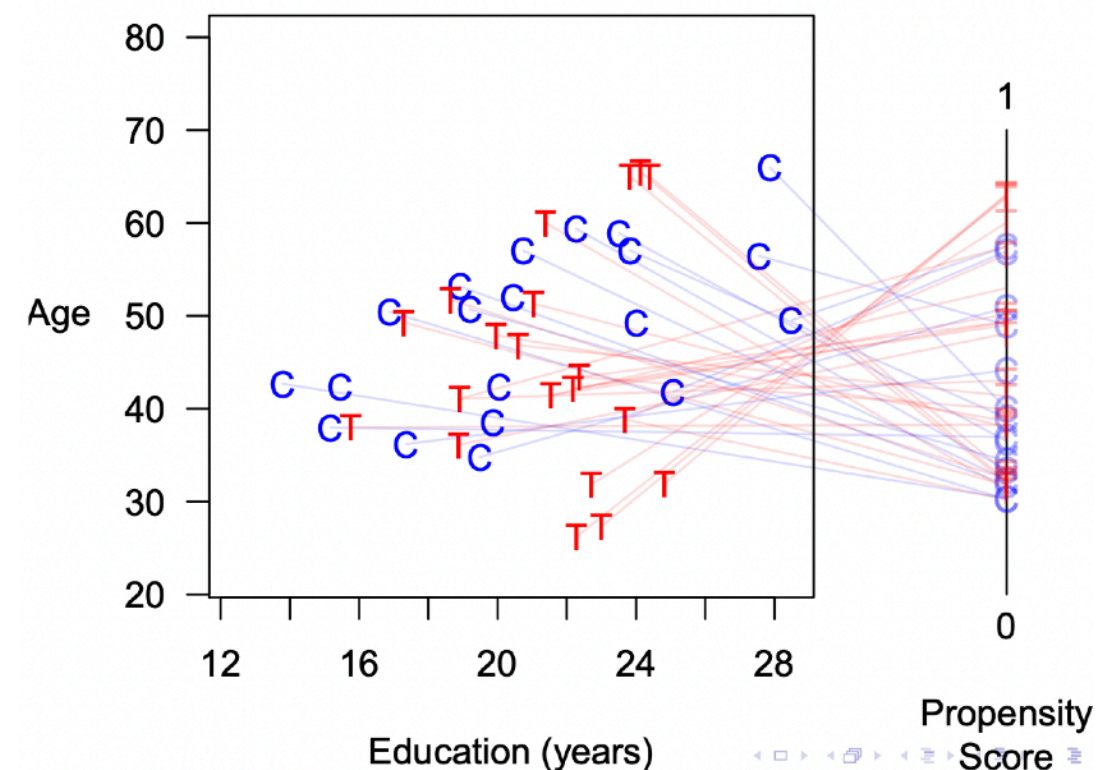
# Propensity Score Matching: Procedure

General procedures of propensity score matching:

1. Estimate propensity scores  $c(X) = P(T=1 | X)$ , e.g. with logistic regression
2. Match each treated to the nearest untreated on propensity score
  - Nearest neighbor matching
  - Optimal full matching ...

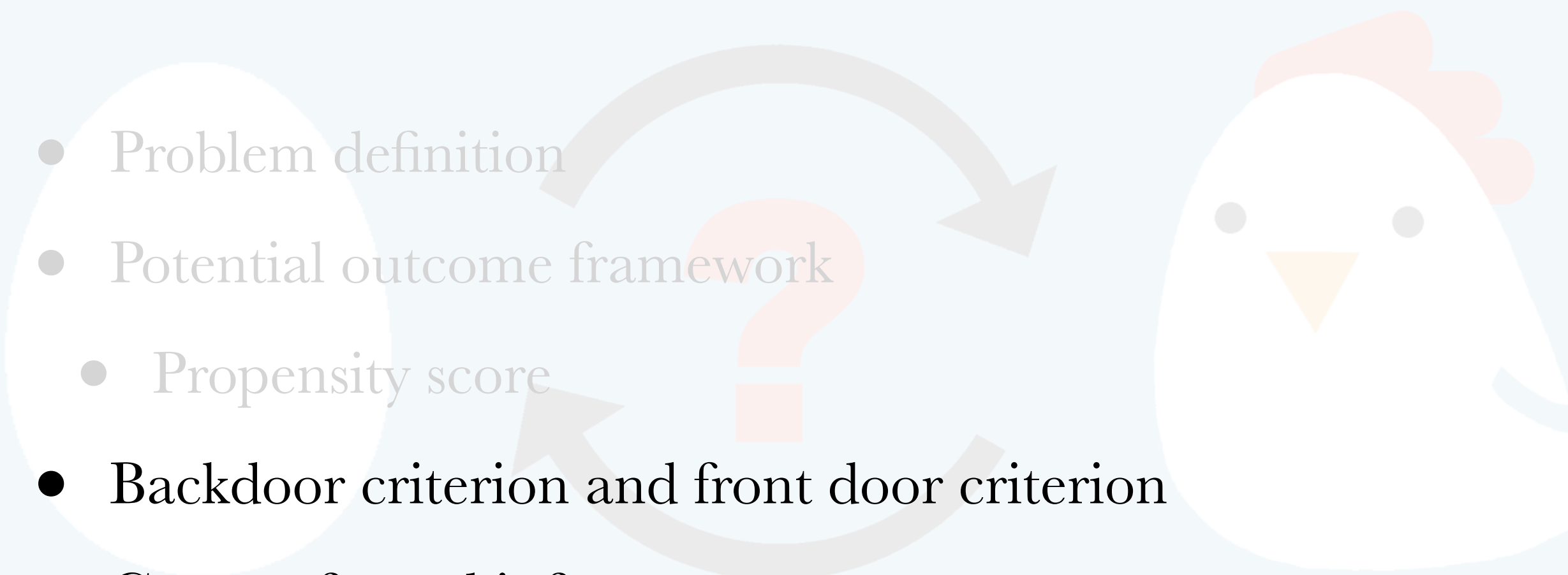


*Estimate propensity scores*



*Matching*

# Identification of Causal Effects & Counterfactual Inference: Outline

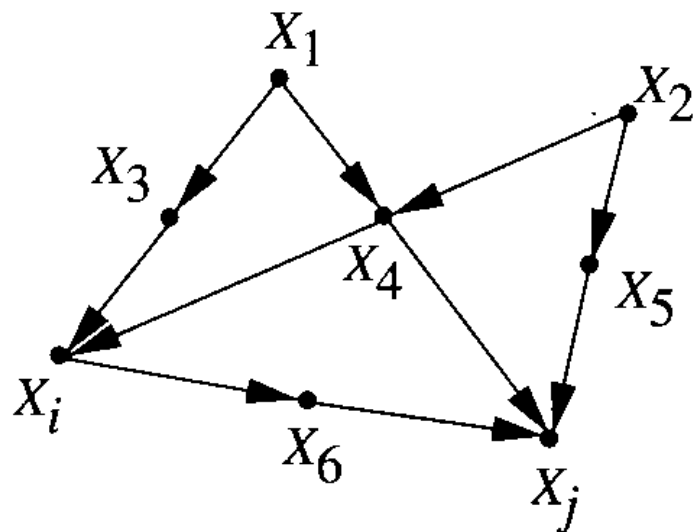
- Problem definition
  - Potential outcome framework
  - Propensity score
  - Backdoor criterion and front door criterion
  - Counterfactual inference
- 
- The diagram illustrates a circular process for identifying causal effects and counterfactual inference. It consists of five steps arranged in a circle, connected by curved arrows. The first three steps are faded: 'Problem definition', 'Potential outcome framework', and 'Propensity score'. The last two steps are bold: 'Backdoor criterion and front door criterion' and 'Counterfactual inference'. A large red question mark is positioned in the center of the circle. To the right of the circle is a cartoon illustration of a white chicken with a red comb and wattle.

# Graphical Criterion: Back-Door Criterion

## Definition 3.3.1 (Back-Door)

A set of variables  $Z$  satisfies the back-door criterion relative to an ordered pair of variables  $(X_i, X_j)$  in a DAG  $G$  if:

- (i) no node in  $Z$  is a descendant of  $X_i$ ; and
- (ii)  $Z$  blocks every path between  $X_i$  and  $X_j$  that contains an arrow into  $X_i$ .



- What if  $Z = \{X_3, X_4\}$ ?

$Z = \{X_4, X_5\}$ ?

$Z = \{X_4\}$ ?

- What if there is a confounder?

## Theorem 3.3.2 (Back-Door Adjustment)

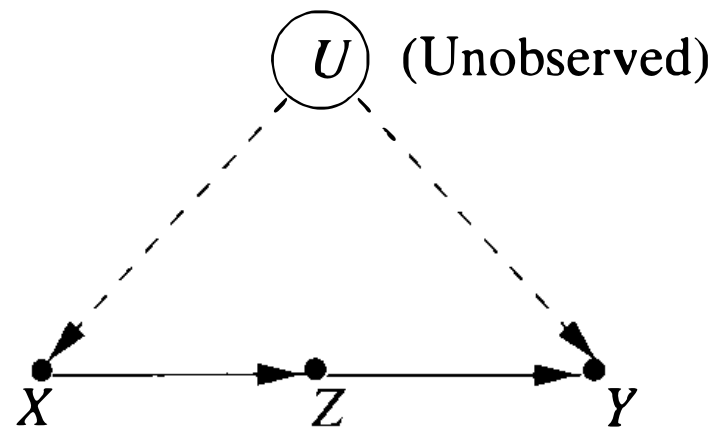
If a set of variables  $Z$  satisfies the back-door criterion relative to  $(X, Y)$ , then the causal effect of  $X$  on  $Y$  is identifiable and is given by the formula

$$\boxed{P(y \mid \hat{x})} = \sum_z P(y \mid x, z) P(z).$$

Or  $P(Y=y \mid \text{do}(X=x))$



# Front-Door Criterion



## Definition 3.3.3 (Front-Door)

A set of variables  $Z$  is said to satisfy the front-door criterion relative to an ordered pair of variables  $(X, Y)$  if:

- (i)  $Z$  intercepts all directed paths from  $X$  to  $Y$ ;
- (ii) there is no back-door path from  $X$  to  $Z$ ; and
- (iii) all back-door paths from  $Z$  to  $Y$  are blocked by  $X$ .

## Theorem 3.3.4 (Front-Door Adjustment)

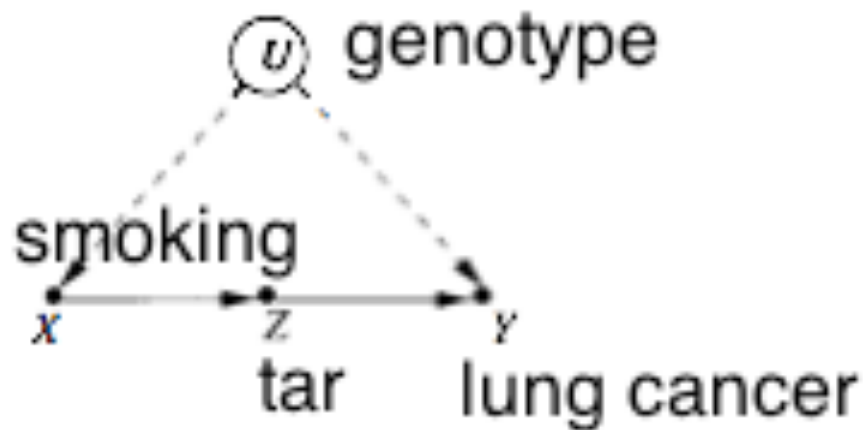
If  $Z$  satisfies the front-door criterion relative to  $(X, Y)$  and if  $P(x, z) > 0$ , then the causal effect of  $X$  on  $Y$  is identifiable and is given by the formula

$$P(y \mid \hat{x}) = \sum_z P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \quad (3.29)$$





# Example: Smoking & Genotype Theory



Group Type		$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0, Z = 0$	Nonsmokers, No tar	47.5	10
$X = 1, Z = 0$	Smokers, No tar	2.5	90
$X = 0, Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1, Z = 1$	Smokers, Tar	47.5	85

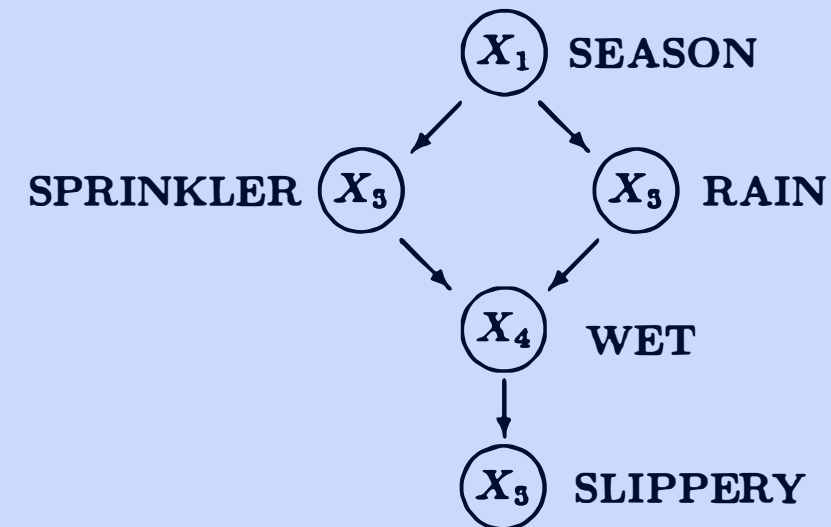
$$\begin{aligned}P(Y = 1 \mid do(X = 1)) &= .05(.10 \times .50 + .90 \times .50) \\&\quad + .95(.05 \times .50 + .85 \times .50) \\&= .05 \times .50 + .95 \times .45 = .4525,\end{aligned}$$

$$\begin{aligned}P(Y = 1 \mid do(X = 0)) &= .95(.10 \times .50 + .90 \times .50) \\&\quad + .05(.05 \times .50 + .85 \times .50) \\&= .95 \times .50 + .05 \times .45 = .4975.\end{aligned}$$



# Remember Structural Causal Models?

- For simplicity, suppose we have  $X$  and  $Y$ :
  - SEM:  $X = E_X$ ;  $Y = f(X, E_Y)$
  - A particular experimental unit (e.g., a patient)  $u$  has its values for exogenous variables  $E_X$  and  $E_Y$ , say,  $e_x$  and  $e_y$
  - Do intervention on  $X$ :  $X = x$ ;  $Y = f(x, E_Y)$
  - **Potential outcome**  $Y(x, u)$  or  $Y_x(u)$
  - $Y(x)$ : **counterfactual variable**



$$PA_i \longrightarrow X_i$$

$$\begin{aligned} X_1 &= E_1, \\ X_2 &= f_2(X_1, E_2), \\ X_3 &= f_3(X_1, E_3), \\ X_4 &= f_4(X_2, X_3, E_4), \\ X_5 &= f_5(X_4, E_5) \end{aligned}$$

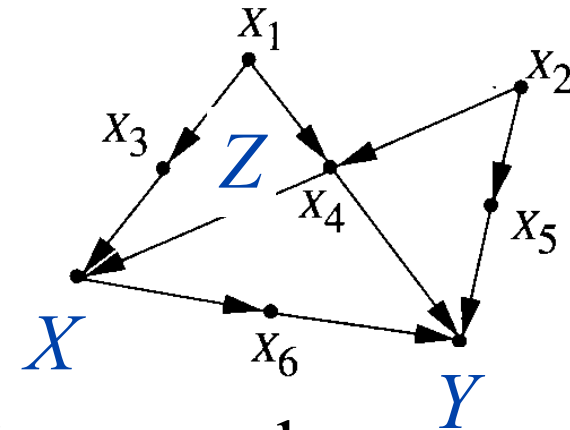


# Relation to Ignorability (Potential Outcome Framework)

## Definition 3.3.1 (Back-Door)

A set of variables  $Z$  satisfies the back-door criterion relative to an ordered pair of variables  $(X_i, X_j)$  in a DAG  $G$  if:

- (i) no node in  $Z$  is a descendant of  $X_i$ ; and
- (ii)  $Z$  blocks every path between  $X_i$  and  $X_j$  that contains an arrow into  $X_i$ .



- (Conditional) ignorability assumption in the potential outcome framework:

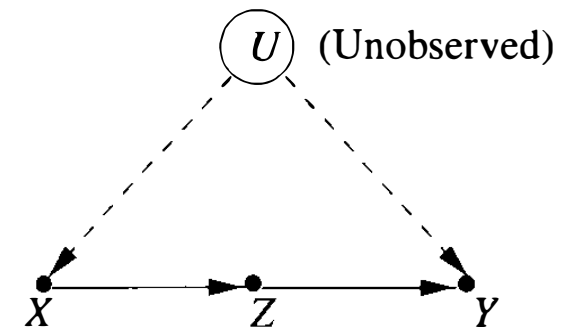
$$Y(x) \perp\!\!\!\perp X \mid Z.$$

$Y(x, u)$ : the value attained by  $Y$  in unit  $u$  under intervention  $\text{do}(x)$ ;  
 $Y(x)$ : counterfactual variable ( $u$  is treated as a variable)

## Definition 3.3.3 (Front-Door)

A set of variables  $Z$  is said to satisfy the front-door criterion of variables  $(X, Y)$  if:

- (i)  $Z$  intercepts all directed paths from  $X$  to  $Y$ ;
- (ii) there is no back-door path from  $X$  to  $Z$ ; and
- (iii) all back-door paths from  $Z$  to  $Y$  are blocked by  $X$ .

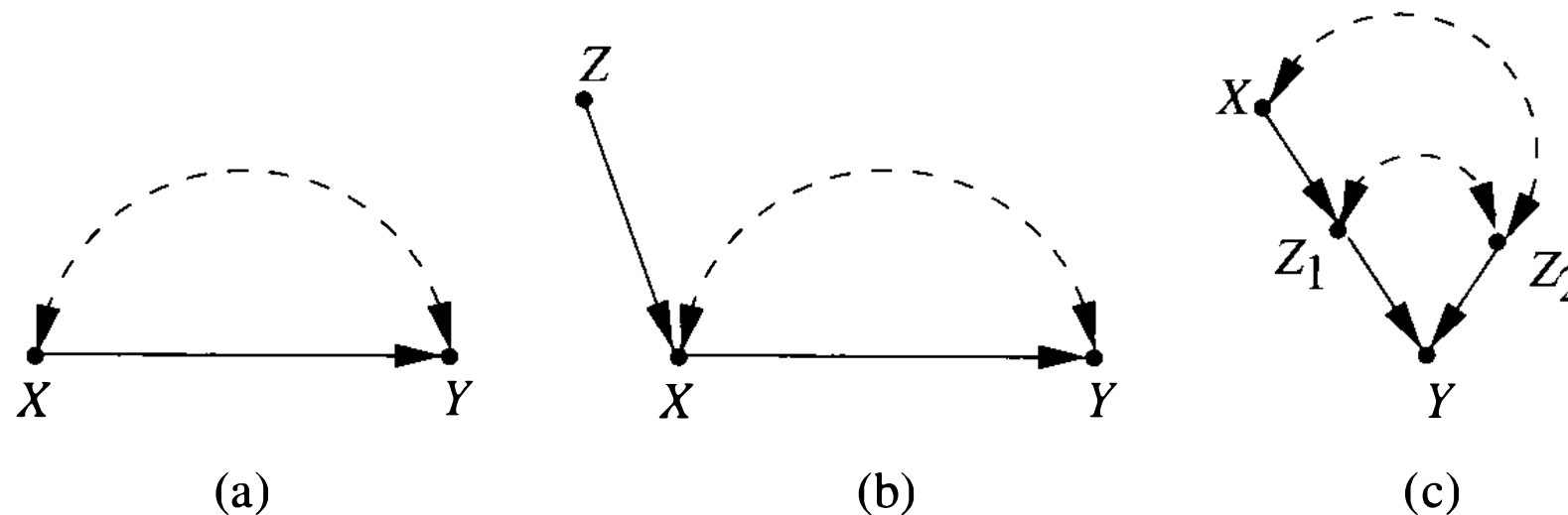


$$- Y(z, x) = Y(z); \{Y(z), X\} \perp\!\!\!\perp Z(x).$$



# A Unification of the Graphical Criteria

- (Pear & Tian, 2002) A **sufficient** condition for identifying the causal effect  $P(y \mid do(x))$  is that **there exists no bi-directed path** (i.e., a path composed entirely of bi-directed arcs) **between  $X$  and any of its children**.
- Necessary & sufficient conditions also exist (e.g., Shpitser and Pearl, 2008)...
- Examples:

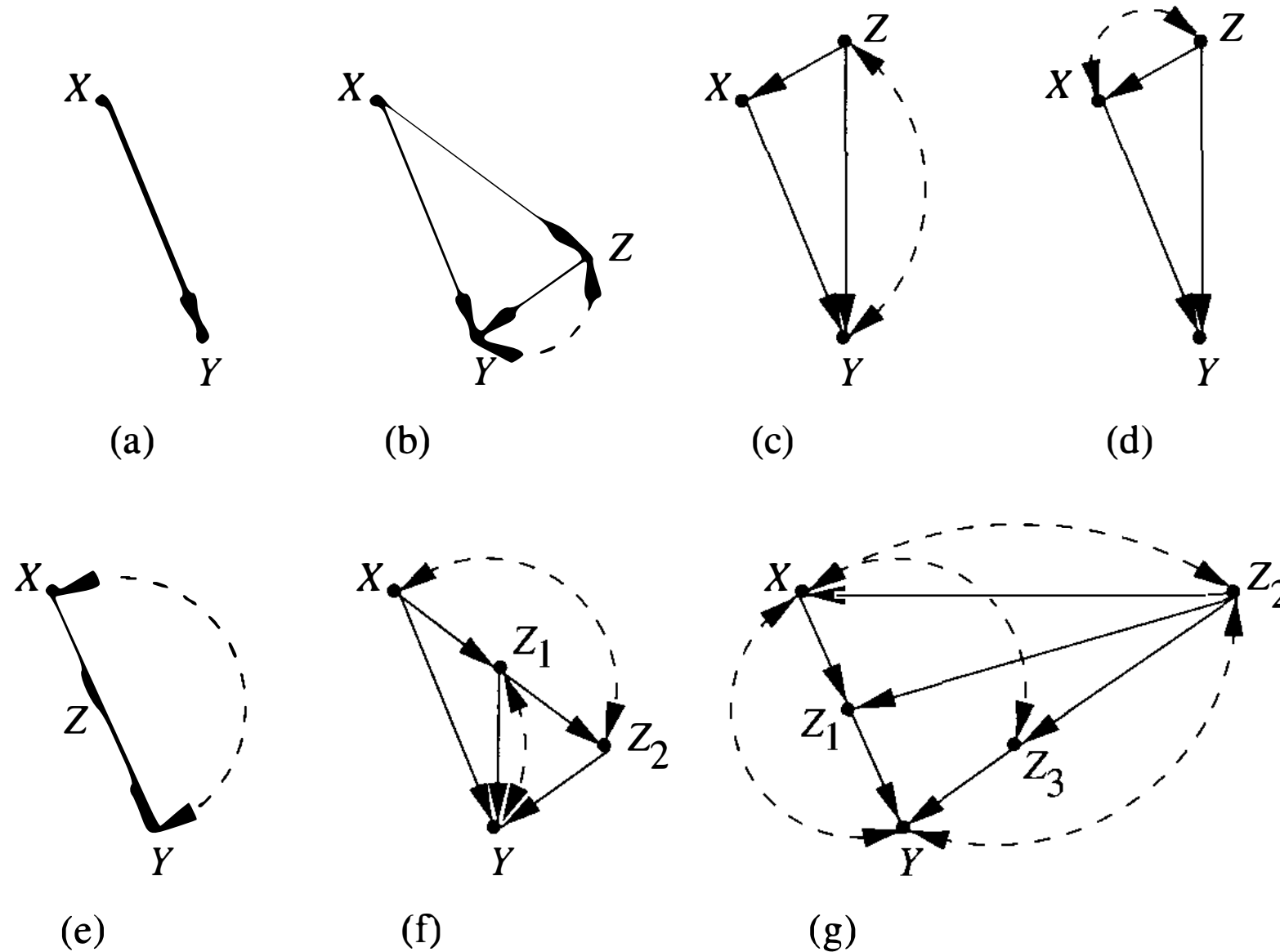


**Figure 3.7** (a) A bow pattern: a confounding arc embracing a causal link  $X \rightarrow Y$ , thus preventing the identification of  $P(y \mid \hat{x})$  even in the presence of an instrumental variable  $Z$ , as in (b). (c) A bowless graph that still prohibits the identification of  $P(y \mid \hat{x})$ .



# A Unification: Examples

- Examples:

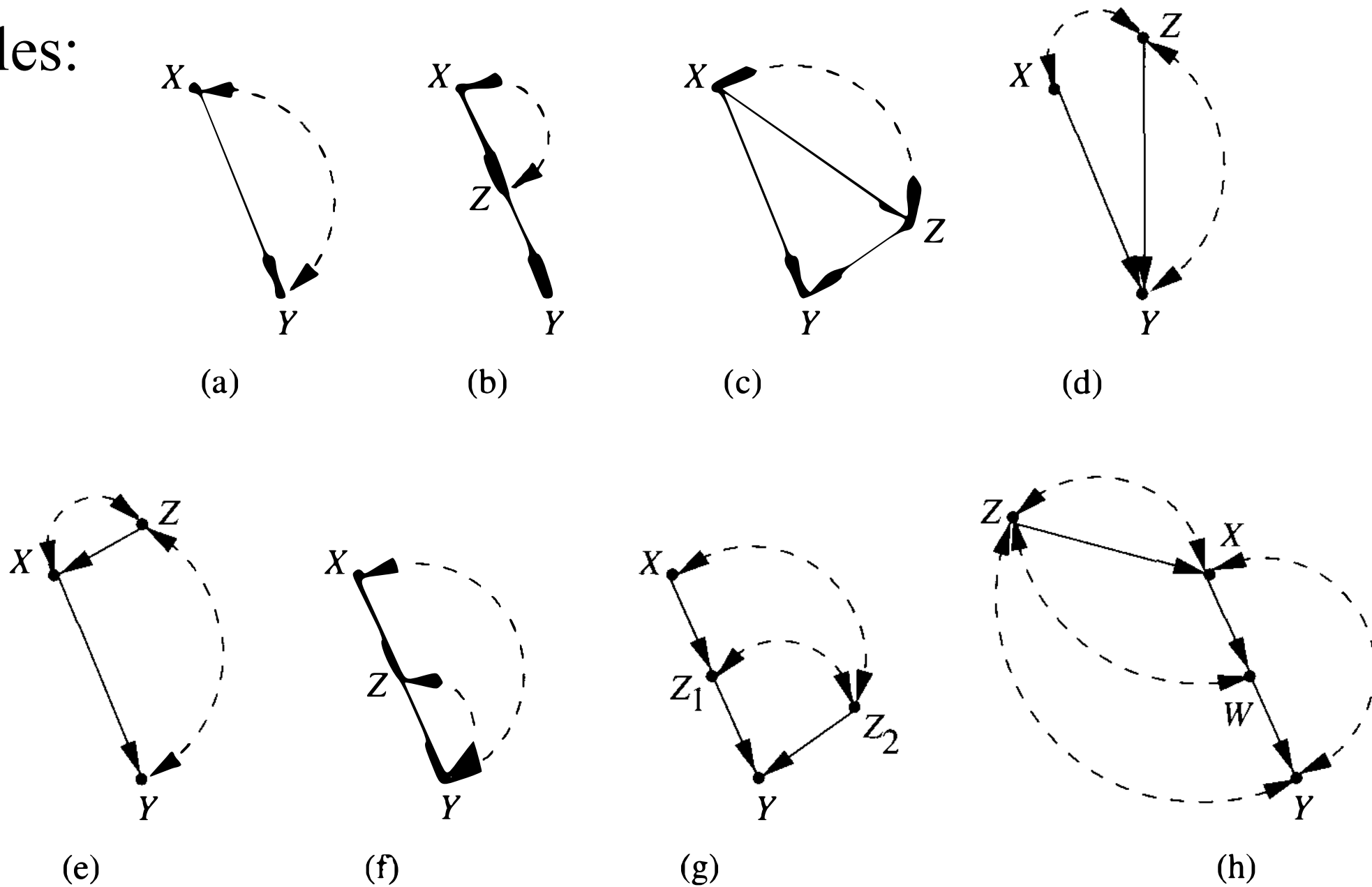


**Figure 3.8** Typical models in which the effect of  $X$  on  $Y$  is identifiable. Dashed arcs represent confounding paths, and  $Z$  represents observed covariates.



# A Unification: Examples

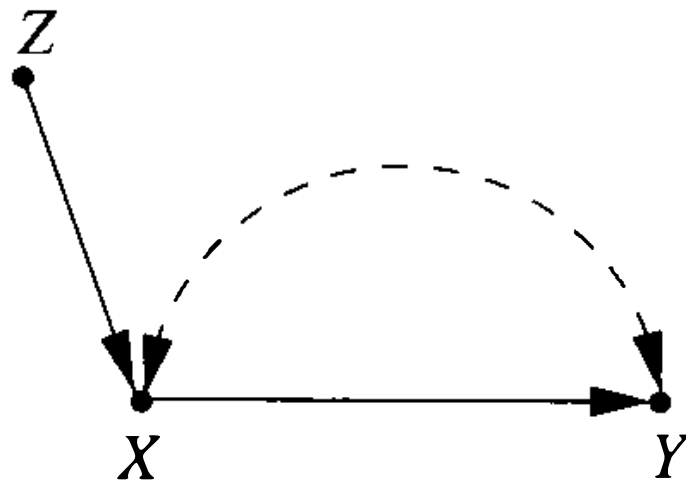
- Examples:



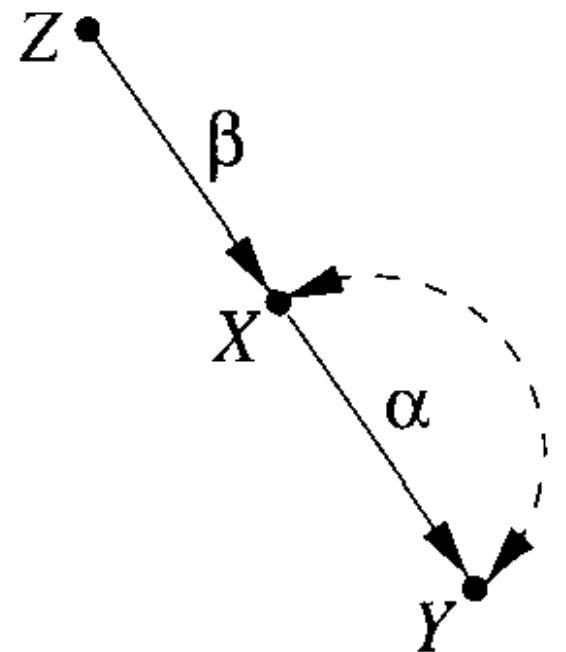
**Figure 3.9** Typical models in which  $P(y | \hat{x})$  is not identifiable.



# Nonparametric vs. Parametric



- What if the causal relations are linear?




$\beta = r_{XZ}$  (regression coefficient of regressing  $X$  on  $Z$ )

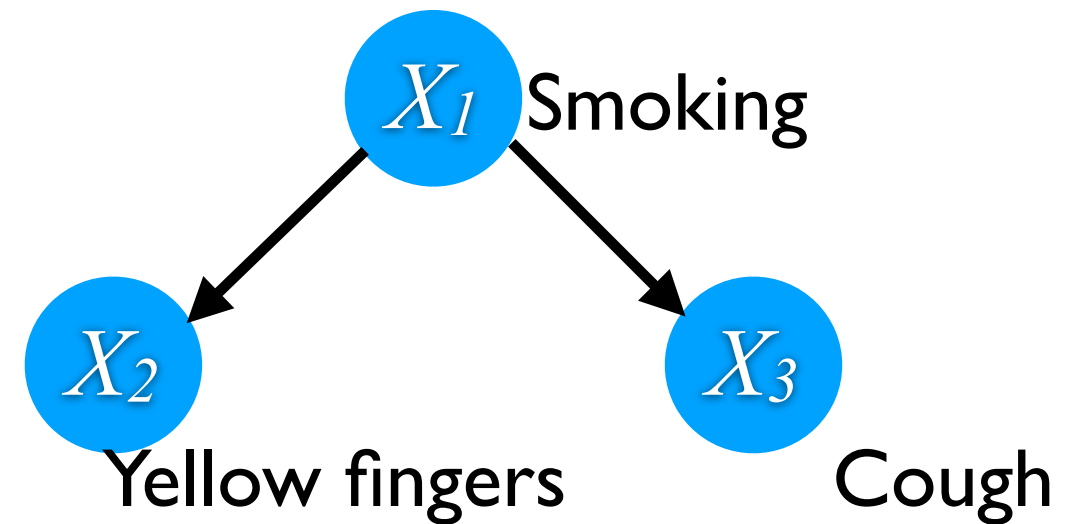
$$\alpha\beta = r_{YZ}$$

so  $\alpha = r_{YZ}/r_{XZ}$ .

# Identification of Causal Effects & Counterfactual Inference: Outline

- Problem definition
  - Potential outcome framework
  - Propensity score
  - Backdoor criterion and front door criterion
  - Counterfactual inference
- 

# Three Types of Problems in Current AI



- Three questions:

$X_1$	$X_2$	$X_3$
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction:** Would the person cough if we *find* he/she has yellow fingers?

$$P(X_3 \mid X_2=1)$$

- **Intervention:** Would the person cough if we *make sure* that he/she has yellow fingers?

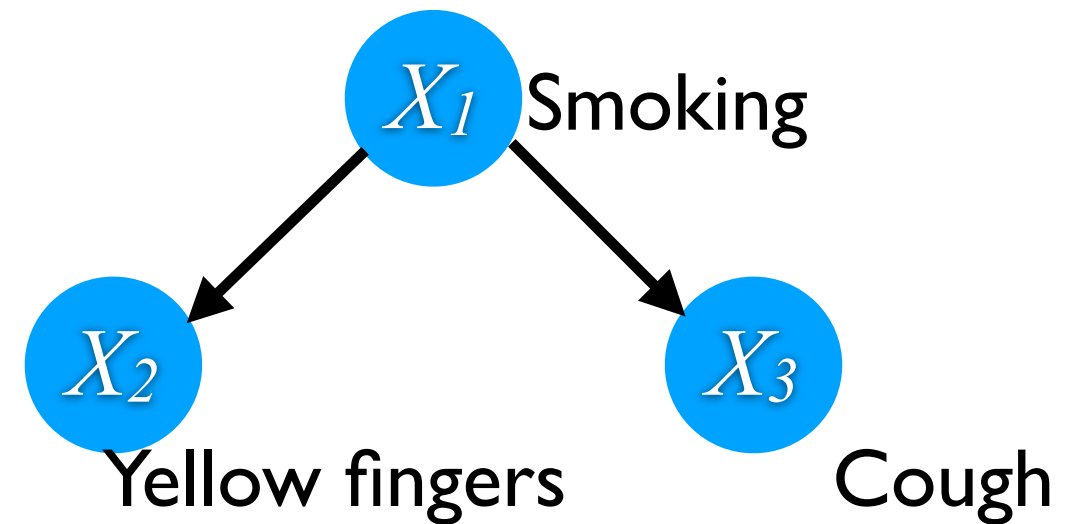
$$P(X_3 \mid \text{do}(X_2=1))$$

- **Counterfactual:** Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X_3_{X_2=1} \mid X_2 = 0, X_3 = 1)$$



# Three Types of Problems in Current AI



- Three questions:

$X_1$	$X_2$	$X_3$
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
...	...	...

- **Prediction:** Would the person cough if we *find* he/she has yellow fingers?

$$P(X_3 \mid X_2=1)$$

- **Intervention:** Would the person cough if we *make sure* that he/she has yellow fingers?

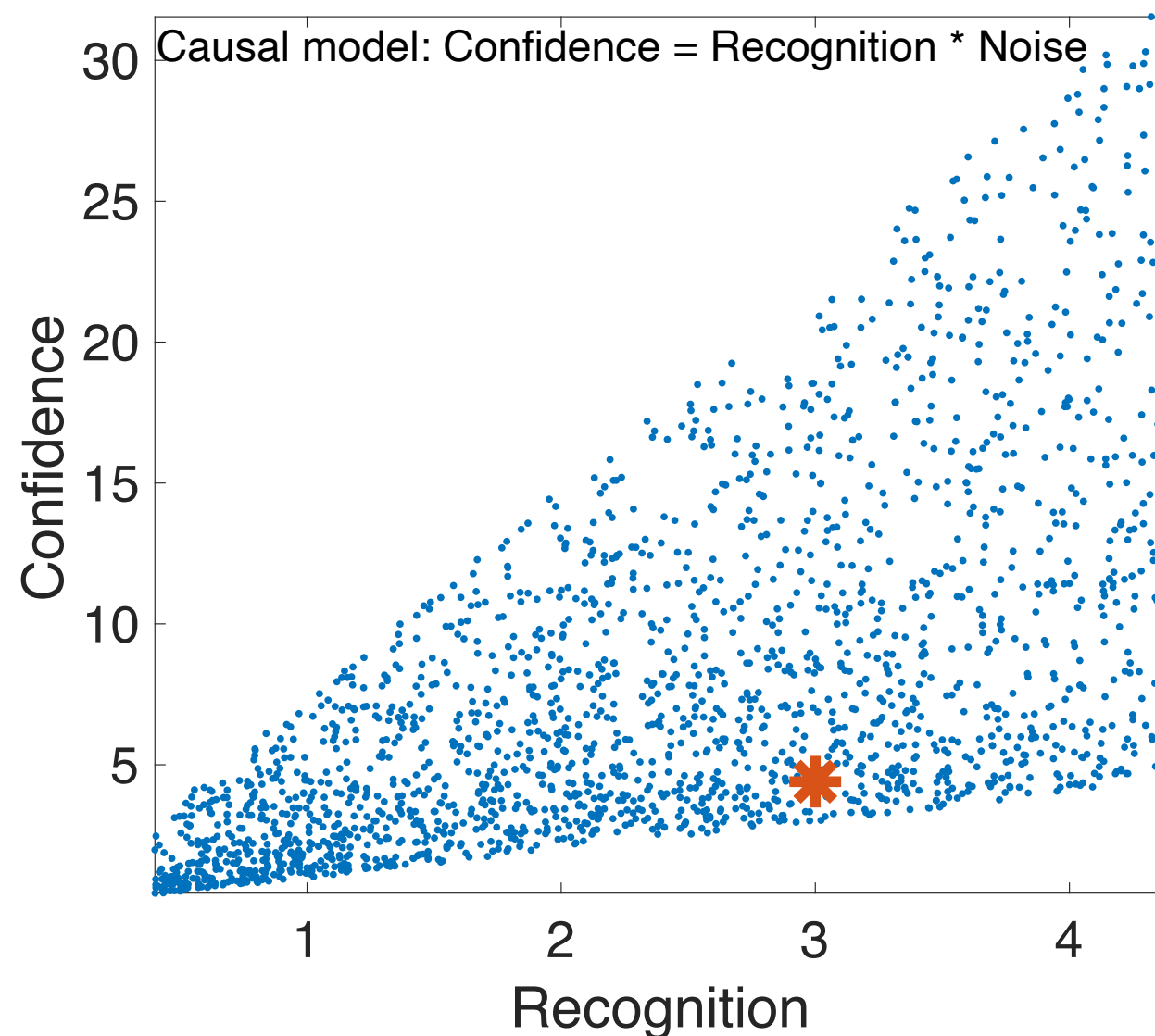
$$P(X_3 \mid \text{do}(X_2=1))$$

- **Counterfactual:** Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X_3_{X_2=1} \mid X_2 = 0, X_3 = 1)$$

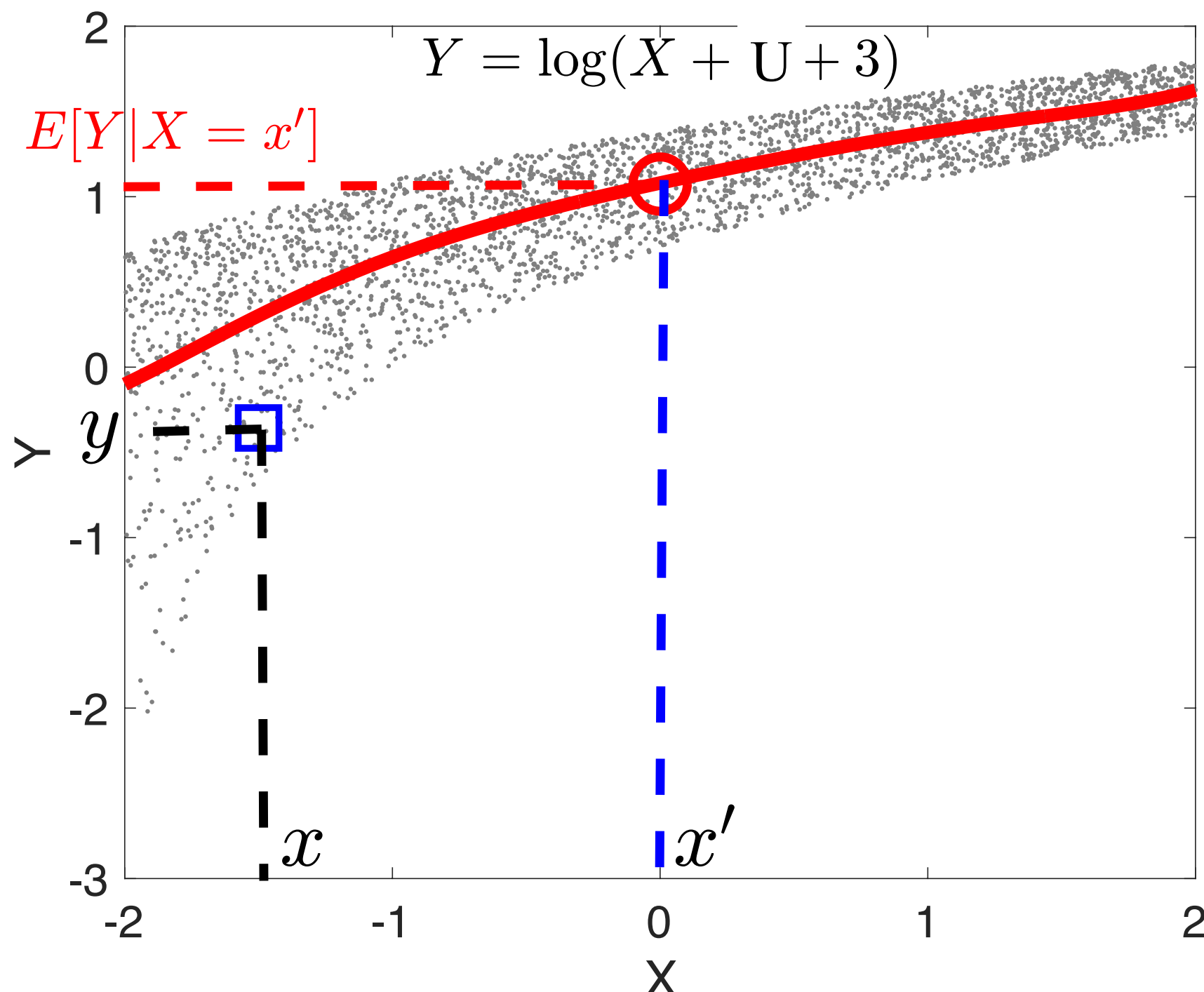
# Counterfactual Inference vs. Prediction

Example 1: Suppose for human beings, confidence is an effect of recognition with causal influence  $\text{confidence} = \text{recognition} * \text{noise}$ . Suppose I receive recognition 3 and that my confidence is 4.4. What would my confidence be if my recognition were 4?



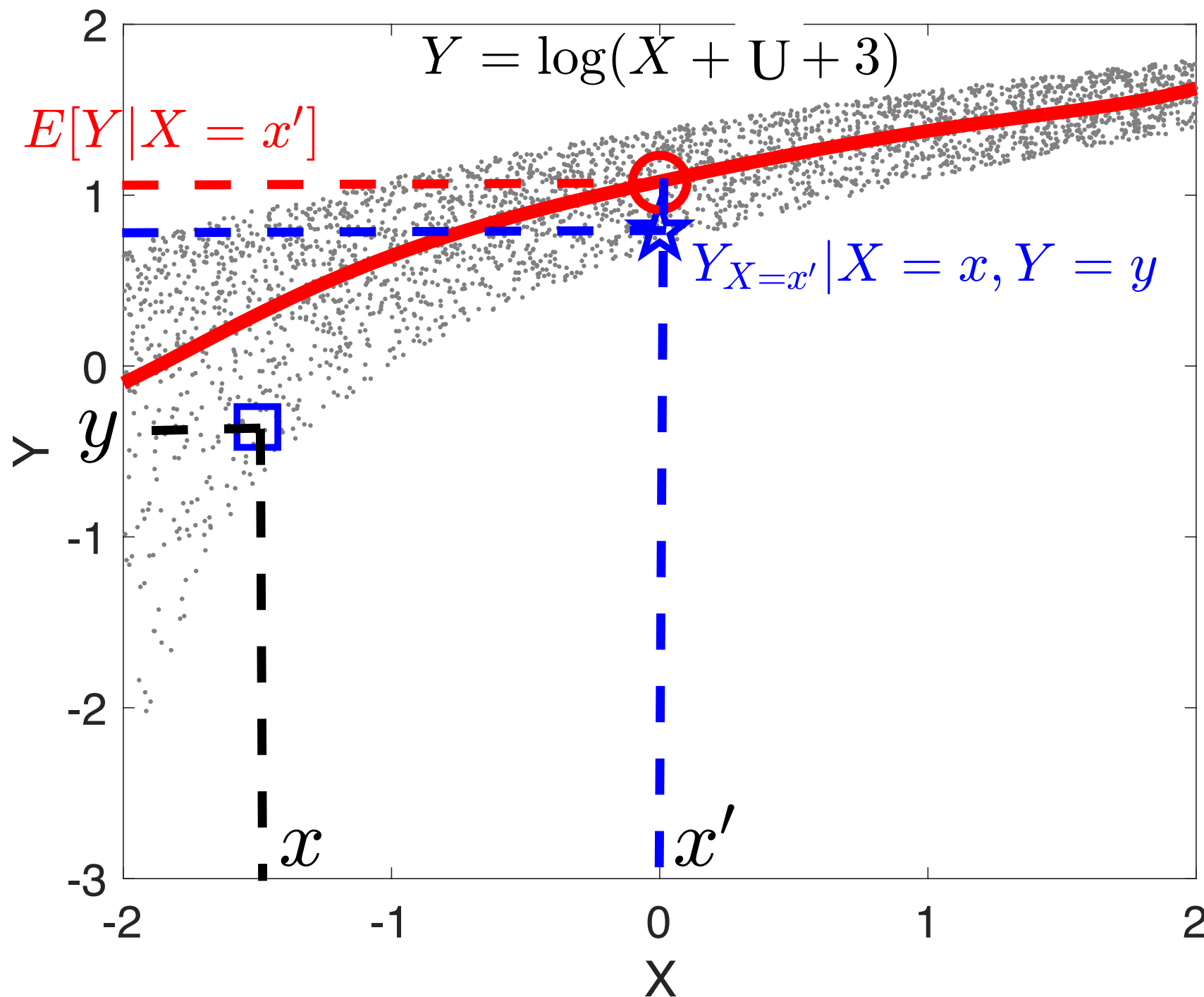
# Counterfactual Inference vs. Prediction

- Suppose  $\overset{\text{attendance}}{X} \rightarrow \overset{\text{grade}}{Y}$  with  $Y = \log(X + U + 3)$ . For an individual with  $(x, y)$ , what would  $Y$  be if  $X$  had been  $x'$ ?



# Counterfactual Inference vs. Prediction

- Suppose  $X \rightarrow Y$  with  $Y = \log(X + U + 3)$ . For an individual with  $(x, y)$ , what would  $Y$  be if  $X$  had been  $x'$ ?

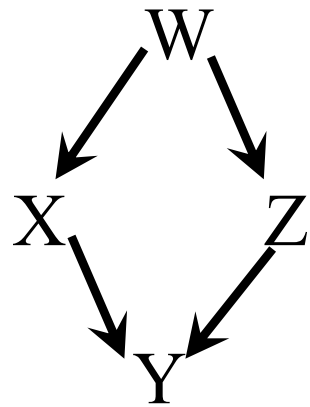


# Standard Counterfactual Questions

- We talk about a particular situation (or unit)  $U = u$ , in which  $X = x$  and  $Y = y$
- What value would  $Y$  be had  $X$  been  $x'$  in situation  $u$ ?  
I.e., we want to know  $Y_{X=x'}(u)$ , the value of  $Y$  in situation  $u$  if we do( $X=x'$ )
- $u$  is not directly observable, so  $P(Y_{X=x'} \mid X = x, Y = y)$  instead

*For identification of causal effects,  $U$  is randomized. It is fixed for counterfactual inference.*

# Counterfactual Inference



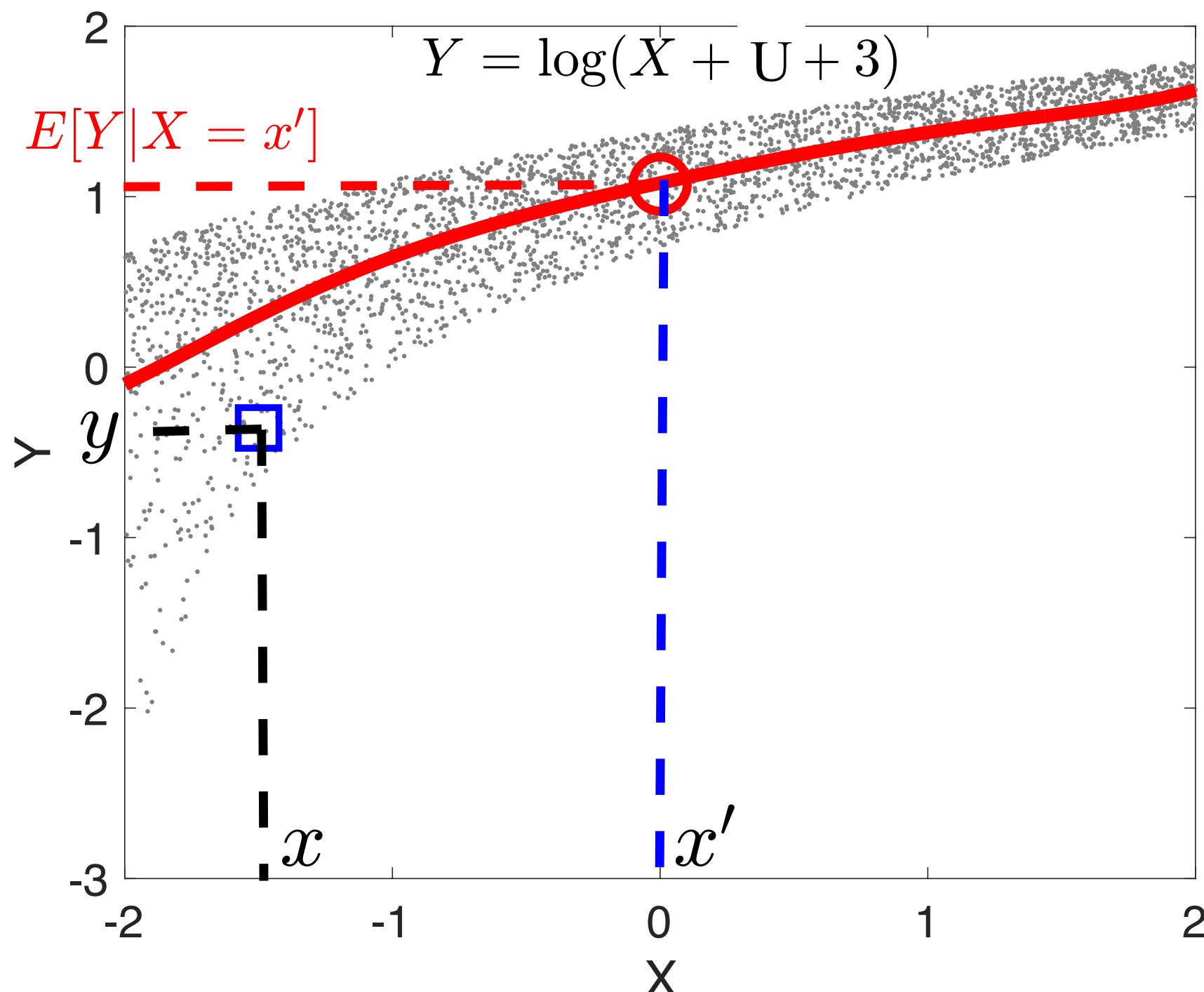
$$\begin{aligned} W &= U_W \\ X &= f_X(W, U_X) \\ Z &= f_Z(W, U_Z) \\ Y &= f_Y(X, Z, U_Y) \end{aligned}$$

$$P(Y_{X=x'} \mid \underbrace{X = x, Y = y, W = w}_{\text{evidence}})$$

- Three steps
  - Abduction: find  $P(U \mid \text{evidence})$
  - Action: Replace the equation for  $X$  by  $X = x'$
  - Prediction: Use the modified model to predict  $Y$

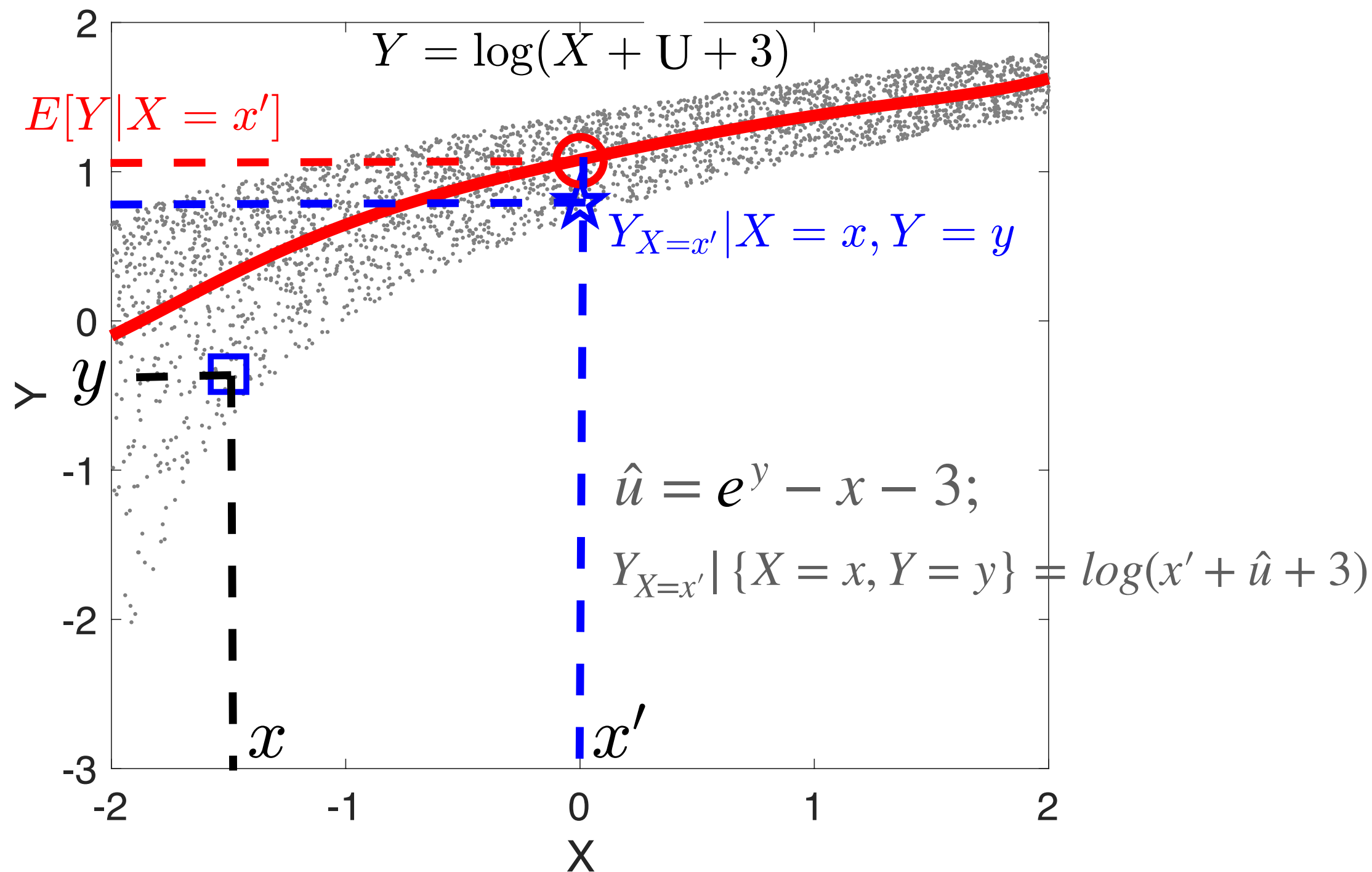
# Counterfactual Inference vs. Prediction

- Suppose  $\overset{\text{attendance}}{X} \rightarrow \overset{\text{grade}}{Y}$  with  $Y = \log(X + U + 3)$ . For an individual with  $(x, y)$ , what would  $Y$  be if  $X$  had been  $x'$ ?



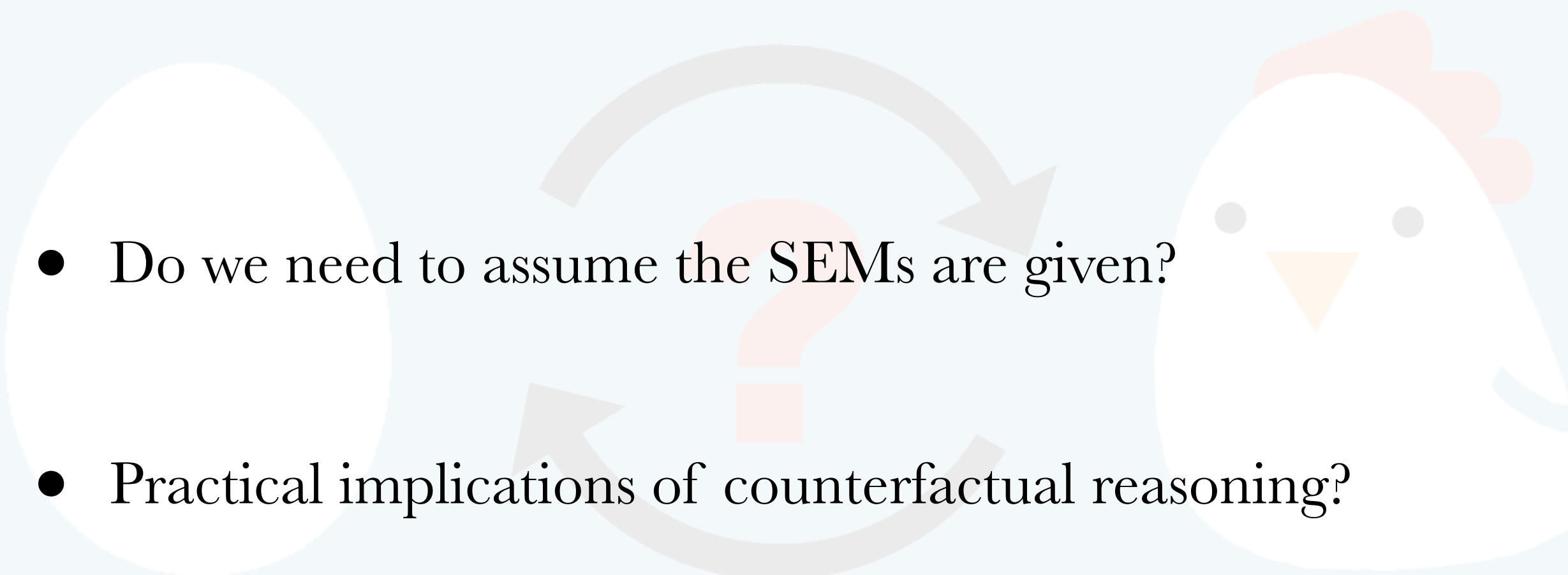
# Counterfactual Inference vs. Prediction

- Suppose  $X \rightarrow Y$  with  $Y = \log(X + U + 3)$ . For an individual with  $(x, y)$ , what would  $Y$  be if  $X$  had been  $x'$ ?





# Recent Advances

- 
- The diagram features a light blue background. On the left is a white egg, and on the right is a white chicken with a pink comb. In the center, a large, faint pink question mark is superimposed over a circular grey arrow that points clockwise from the egg to the chicken and back.
- Do we need to assume the SEMs are given?
  - Practical implications of counterfactual reasoning?

# Counterfactual Reasoning with Nonlinear Quantile Regression

- Feasible without SEMs!
- The same value of the noise corresponds to the quantile function of the same percentile
- Can be easily learned

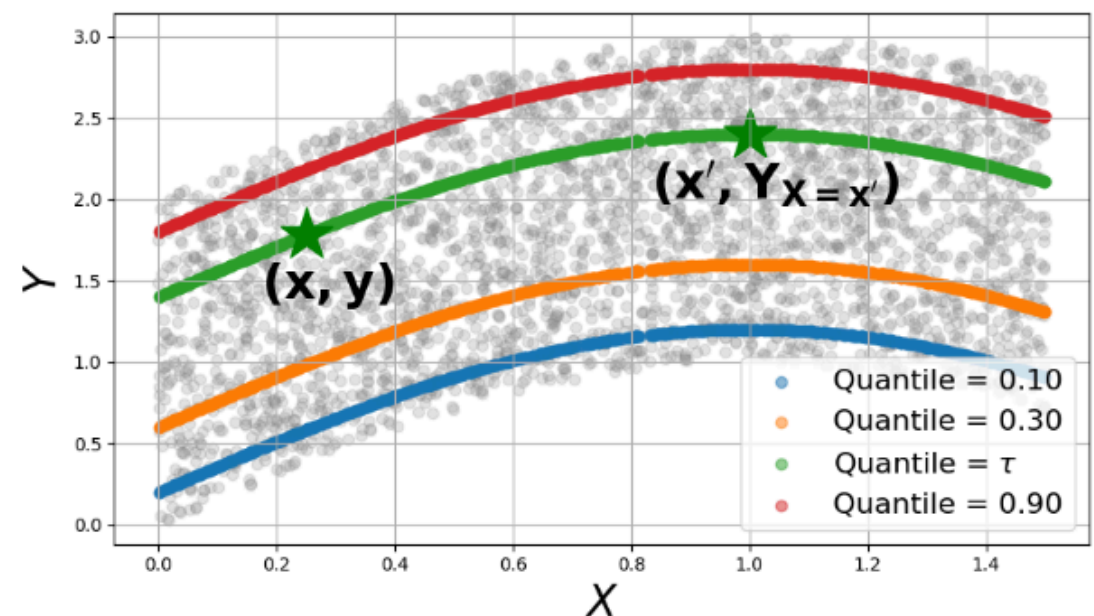
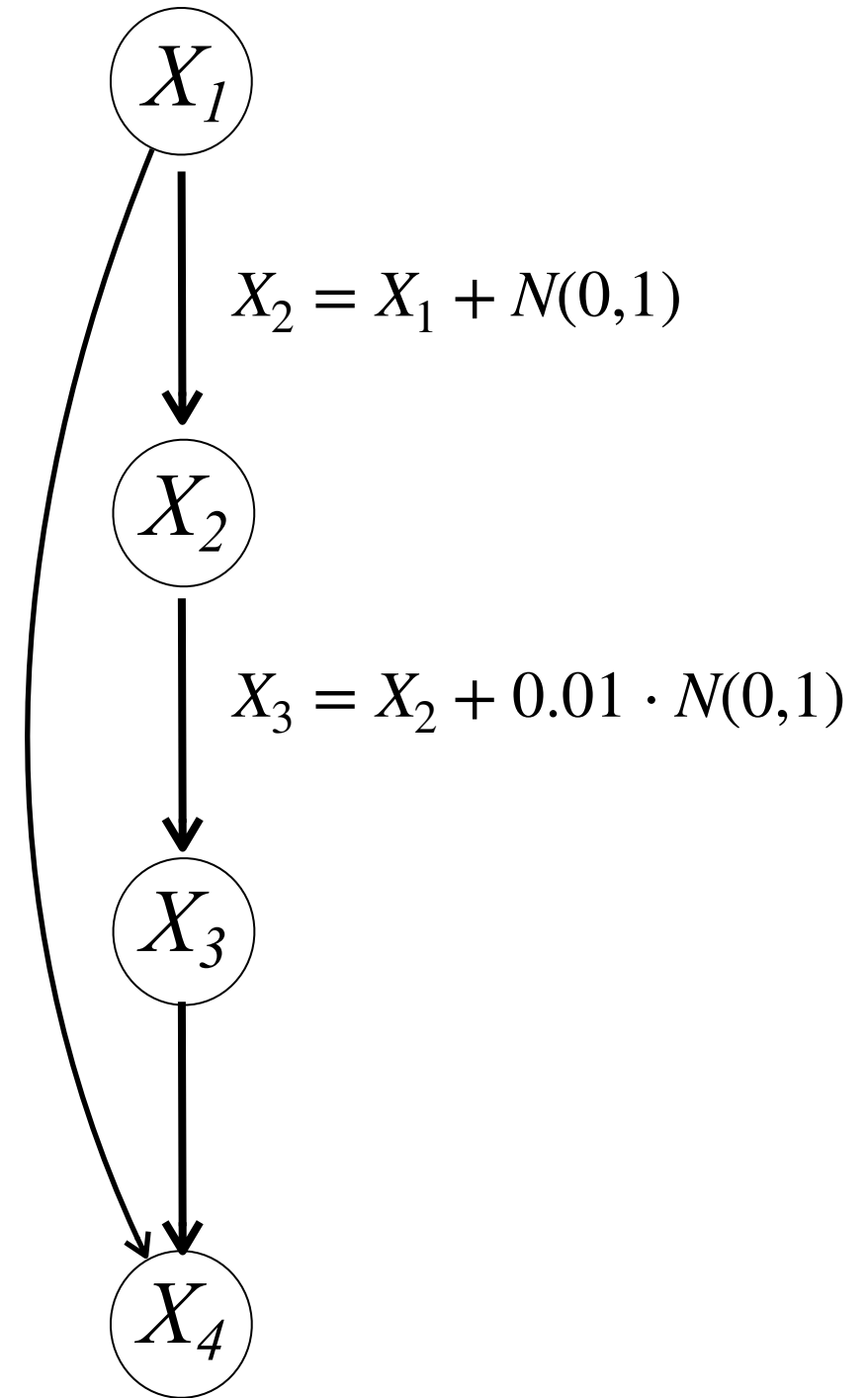


Figure 1: Illustration of our proposed quantile-based counterfactual estimation ( $Z$  is omitted for illustration purpose). For a sample of interest ( $X = x, Z = z, Y = y$ ), we estimate the quantile  $\tau = P(Y \leq y | X = x, Z = z) = 0.70$  with factual observations. Then the counterfactual outcome  $Y_{X=x'}$  is equal to the value  $y'$  which satisfy  $P(Y \leq y' | X = x', Z = z) = \tau$ .

# Is This Counterfactual Reasoning Meaningful?

- Feasible to do intervention on  $X_3$ ?
- Necessary backtracking for feasible changes and practical implications
- Constrained optimization for where to apply (backtracking) interventions



# Summary: Causal Effect Identification & Counterfactual Reasoning

- Causal inference: Classical problem
  - What is taken as input?
  - What does identifiability mean?
  - *Backdoor criterion and unification* \*
- Difference from counterfactual reasoning
  - What do you think of counterfactual reasoning?
  - How to make it more feasible and useful?