Transfer Learning and Transportability

Brady Neal

causalcourse.com

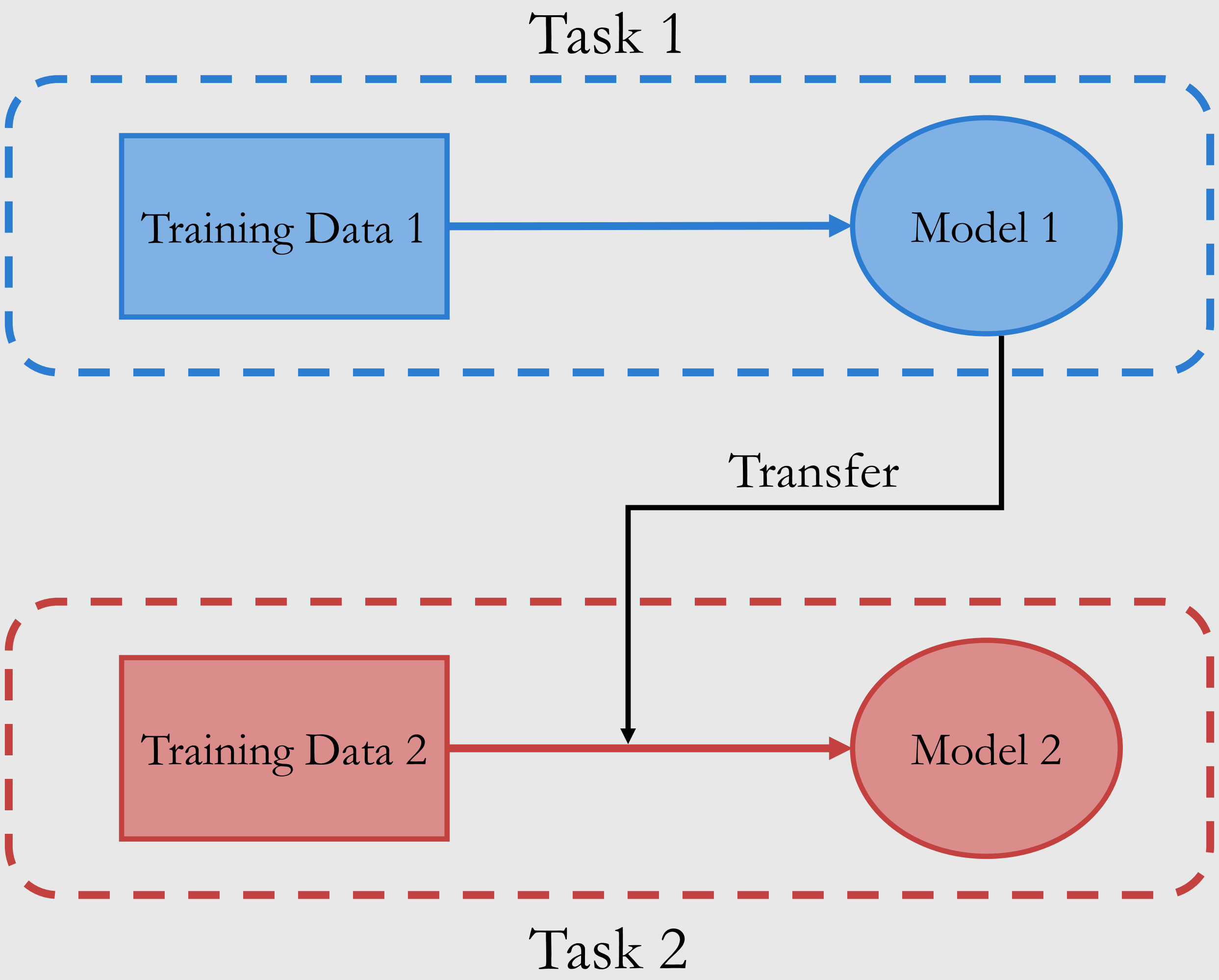
Causal Insights for Transfer Learning

Transportability of Causal Effects Across Populations

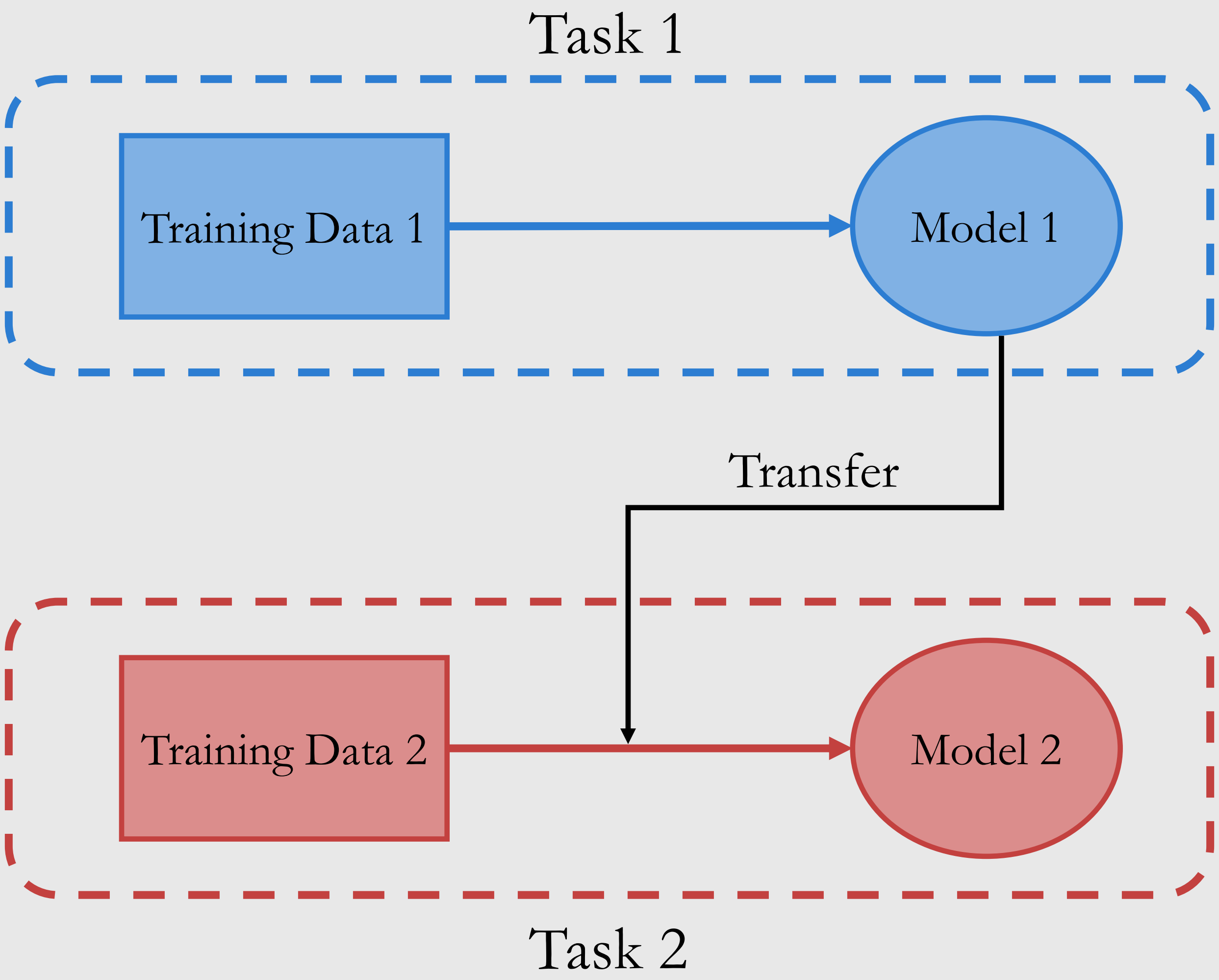
Causal Insights for Transfer Learning

Transportability of Causal Effects Across Populations

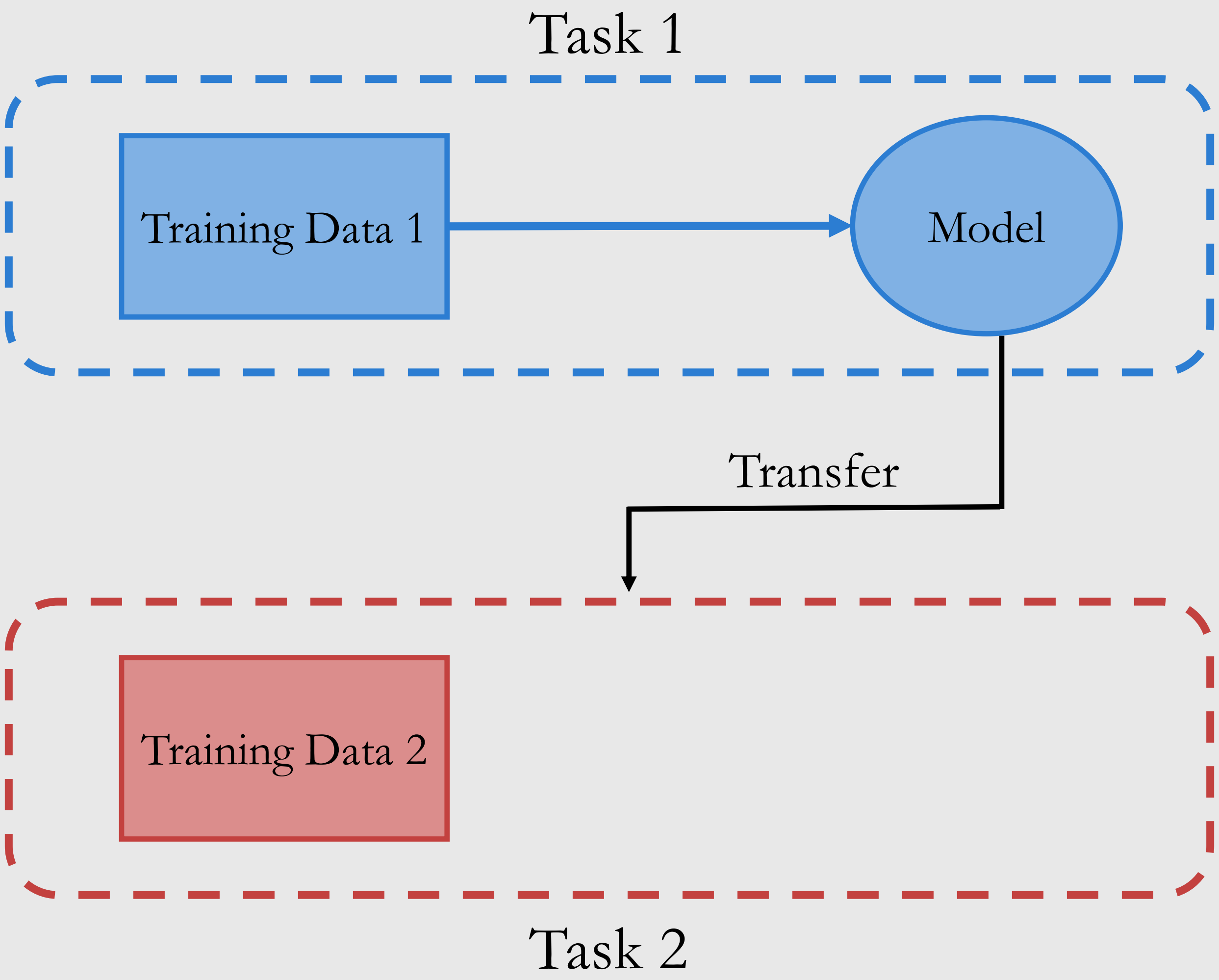
Transfer Learning



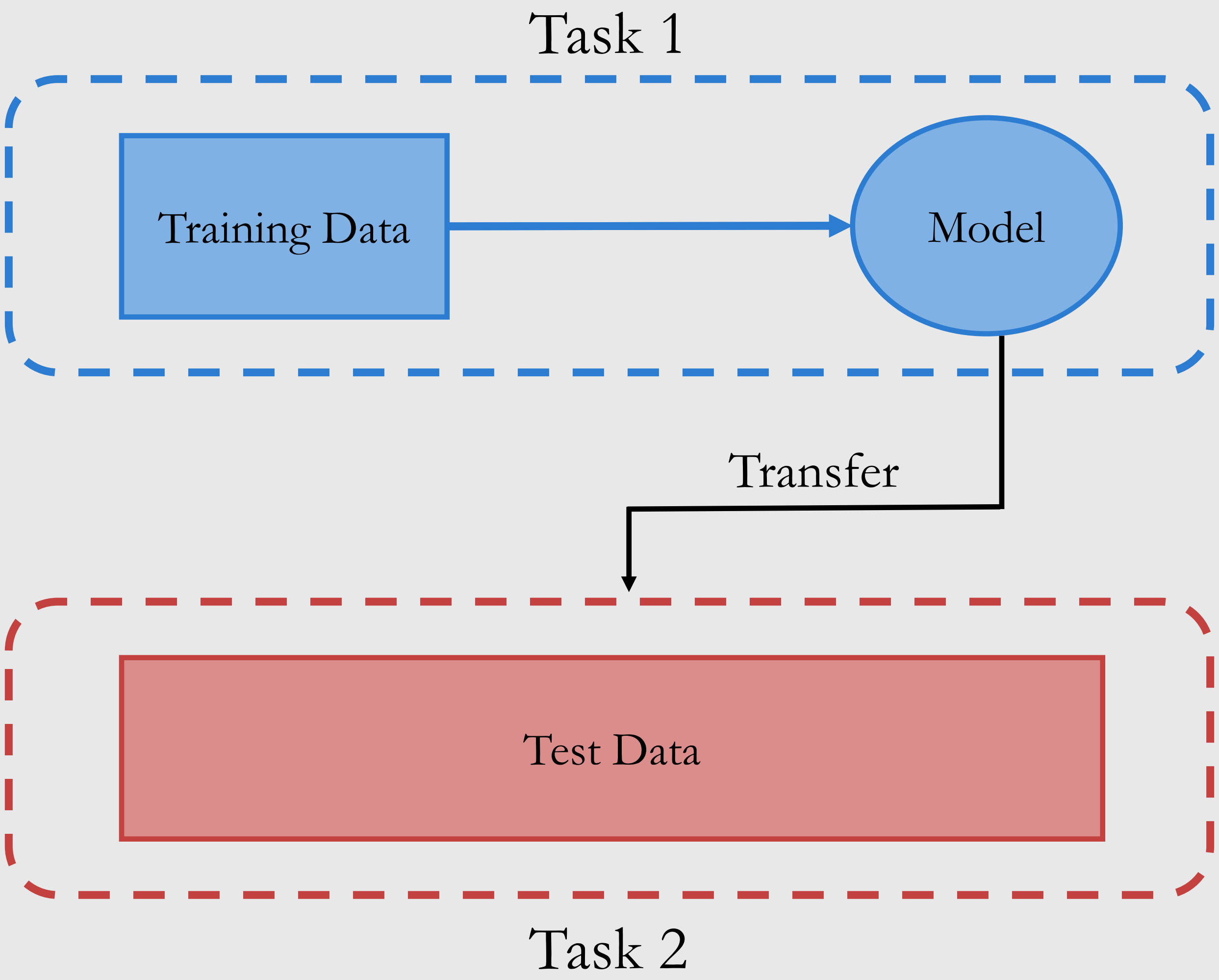
Domain Generalization



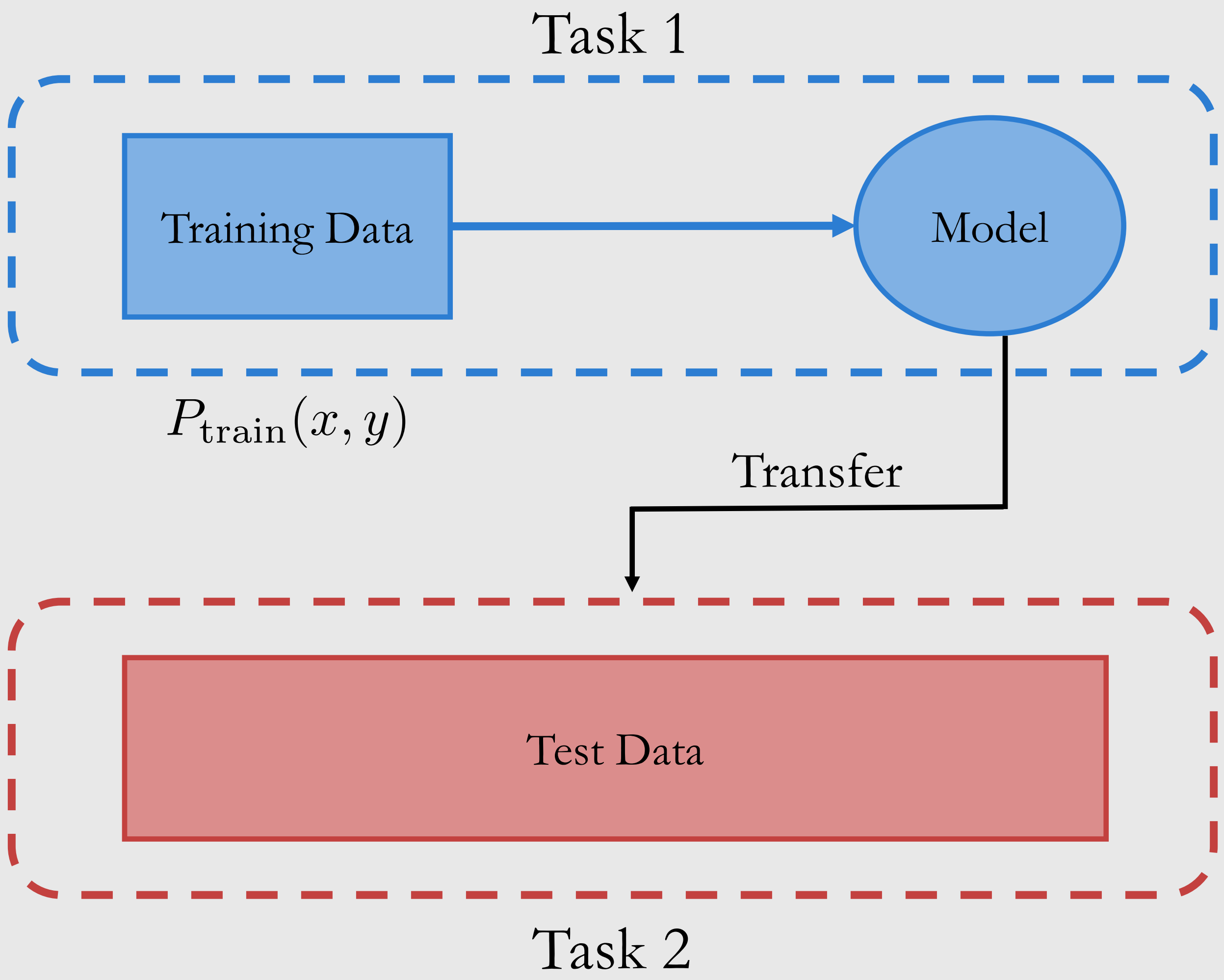
Domain Generalization



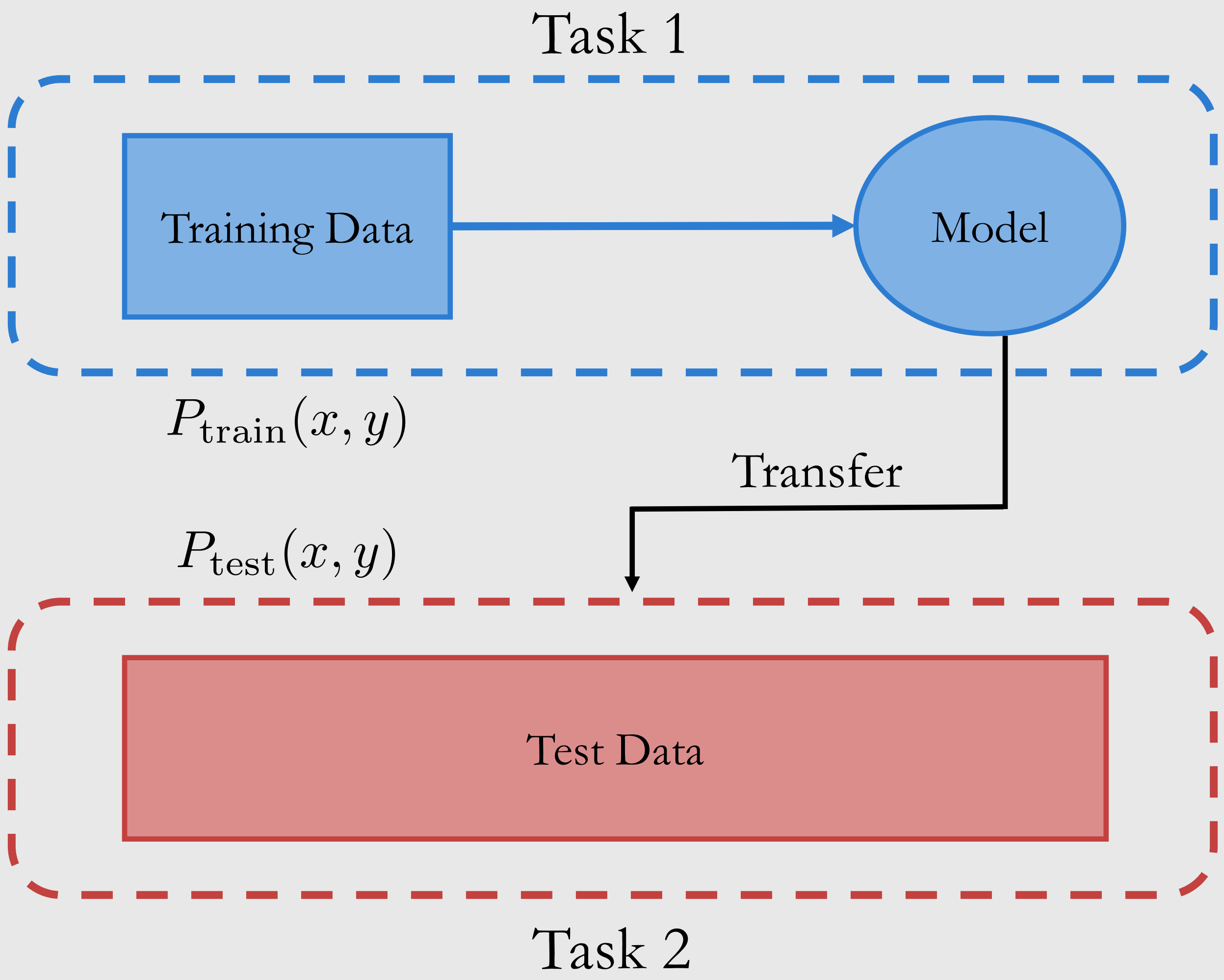
Domain Generalization



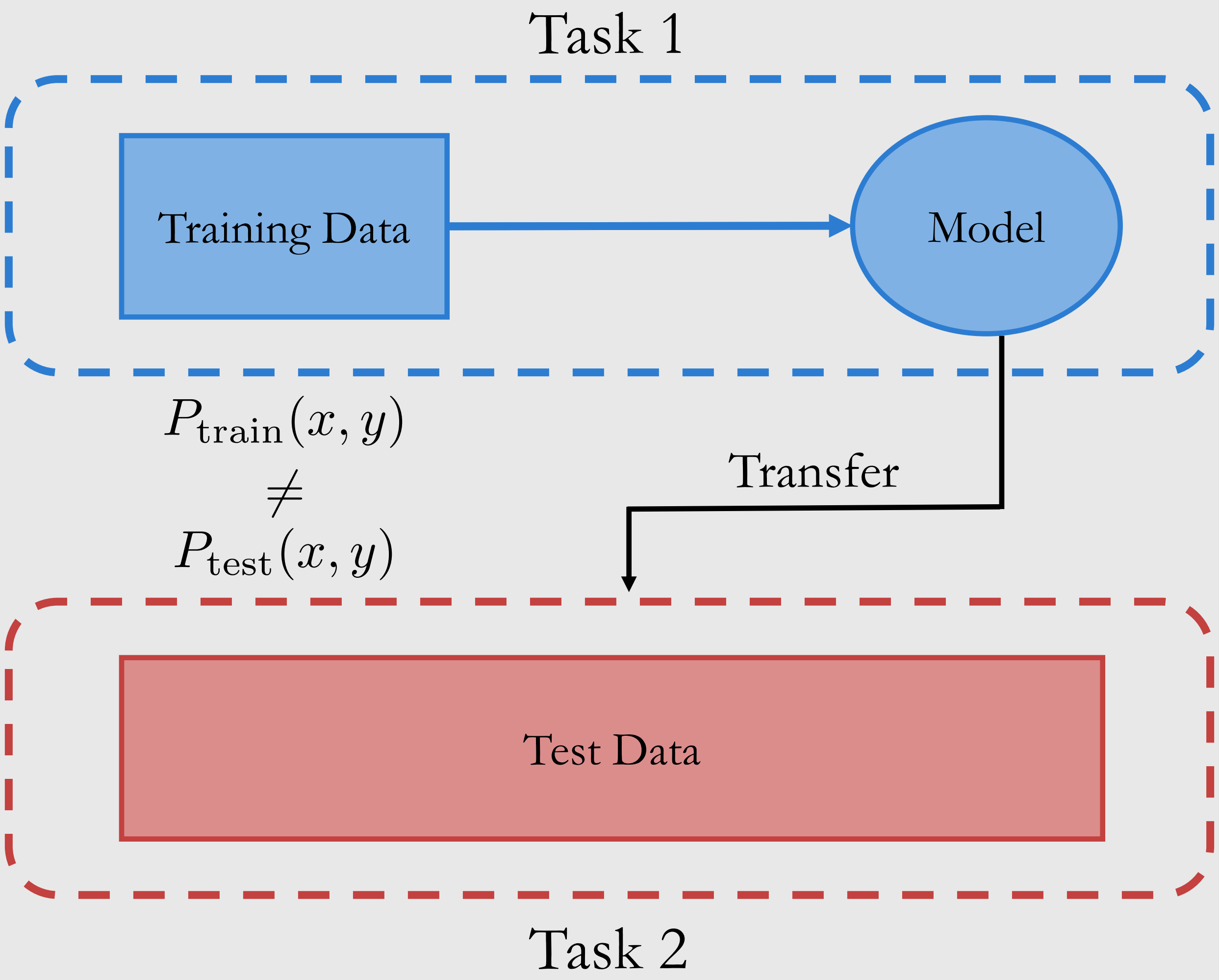
Domain Generalization



Domain Generalization



Domain Generalization



Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$

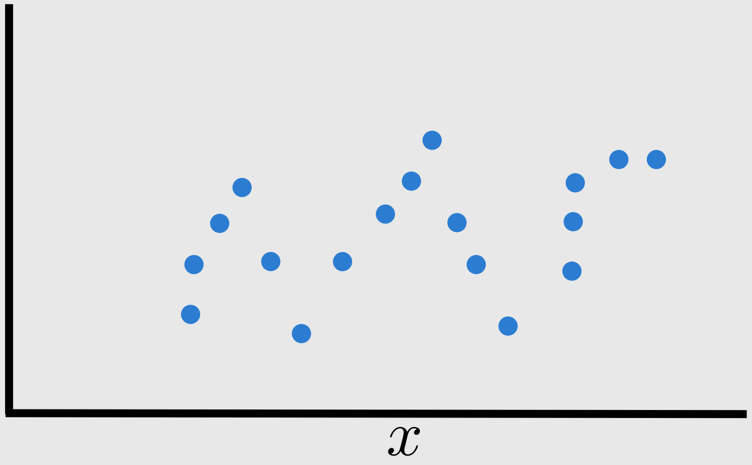
Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$

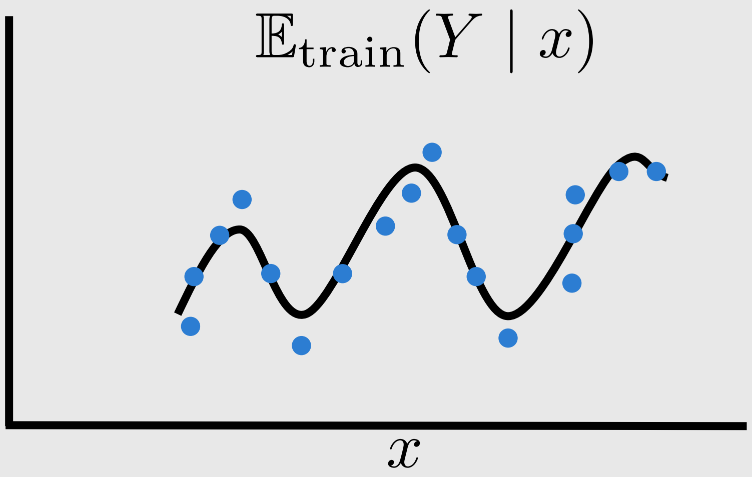


Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$
$P_{\text{train}}(x) \neq P_{\text{test}}(x)$



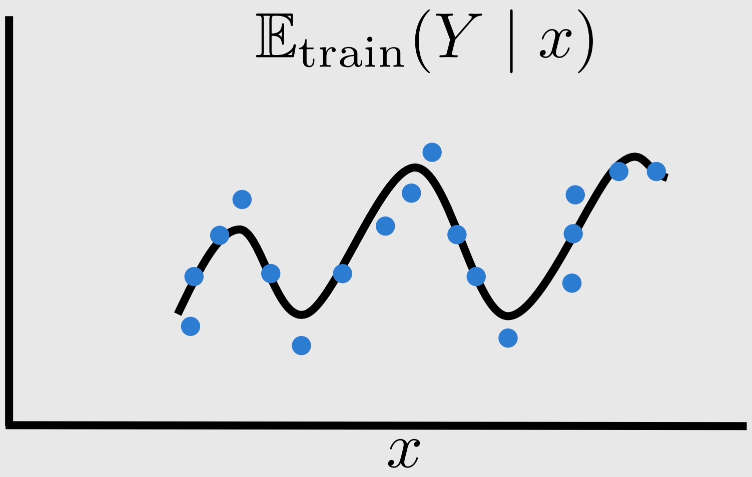
Covariate Shift

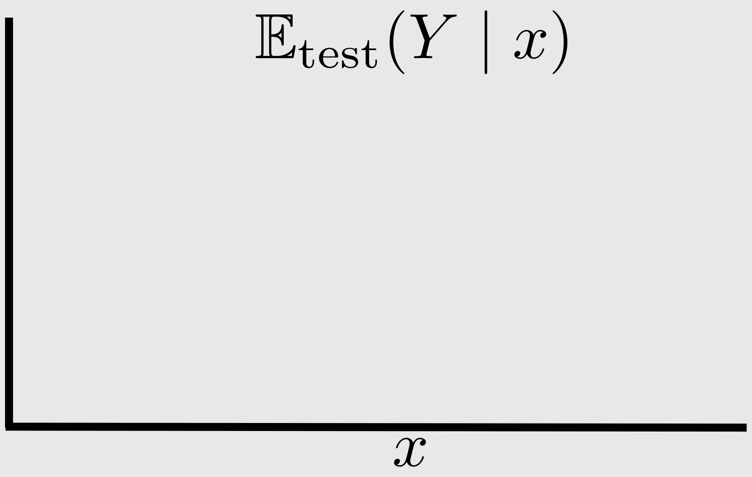
Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$





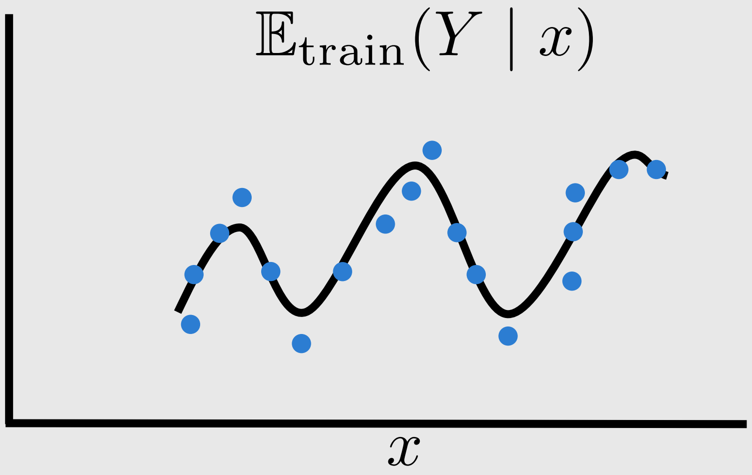
Covariate Shift

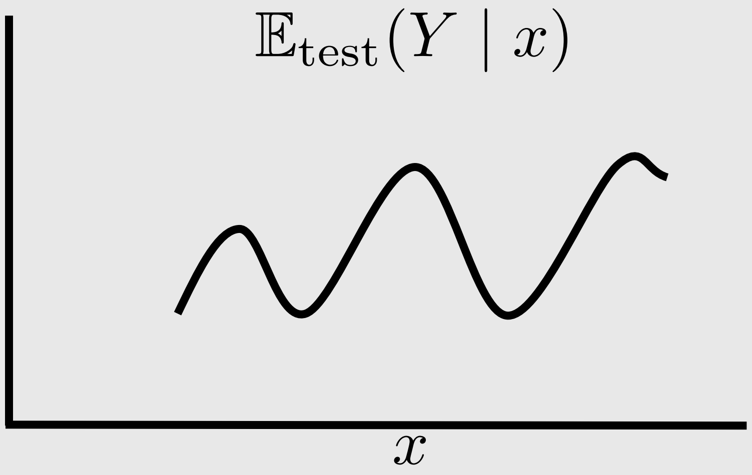
Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$





Covariate Shift

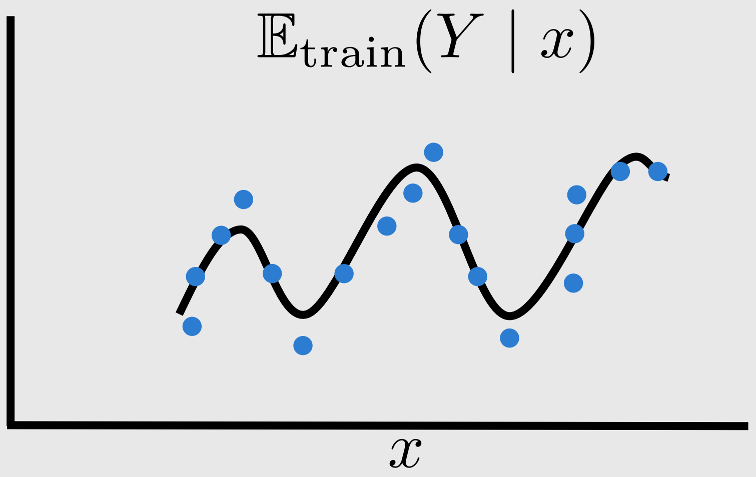
Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

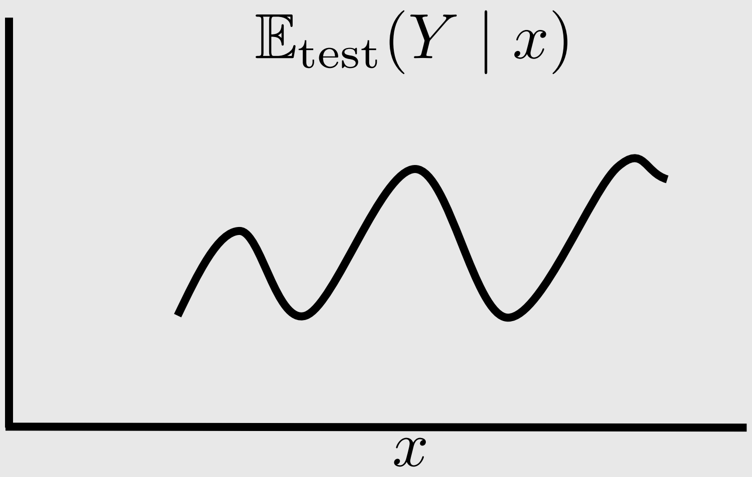
Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$

$$\text{supp}_{\text{train}}(x) = \text{supp}_{\text{test}}(x)$$



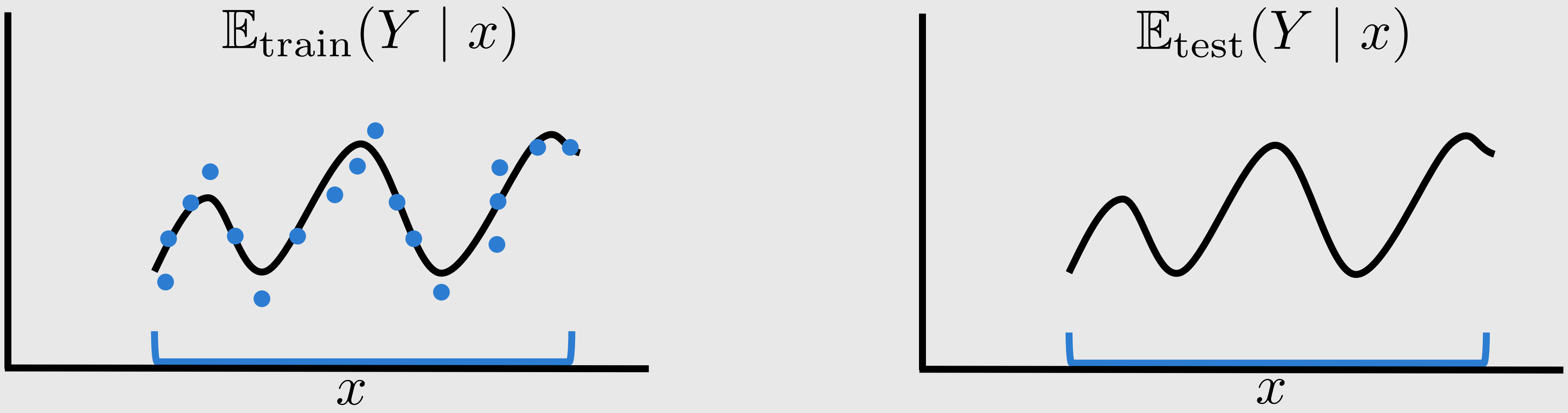


Covariate Shift

Setting: $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$

Goal: Model $\mathbb{E}_{\text{test}}(Y \mid x)$ only given access to $P_{\text{train}}(x, y)$

Covariate Shift Assumption:
$$P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$$
$$P_{\text{train}}(x) \neq P_{\text{test}}(x)$$
$$\text{supp}_{\text{train}}(x) = \text{supp}_{\text{test}}(x)$$

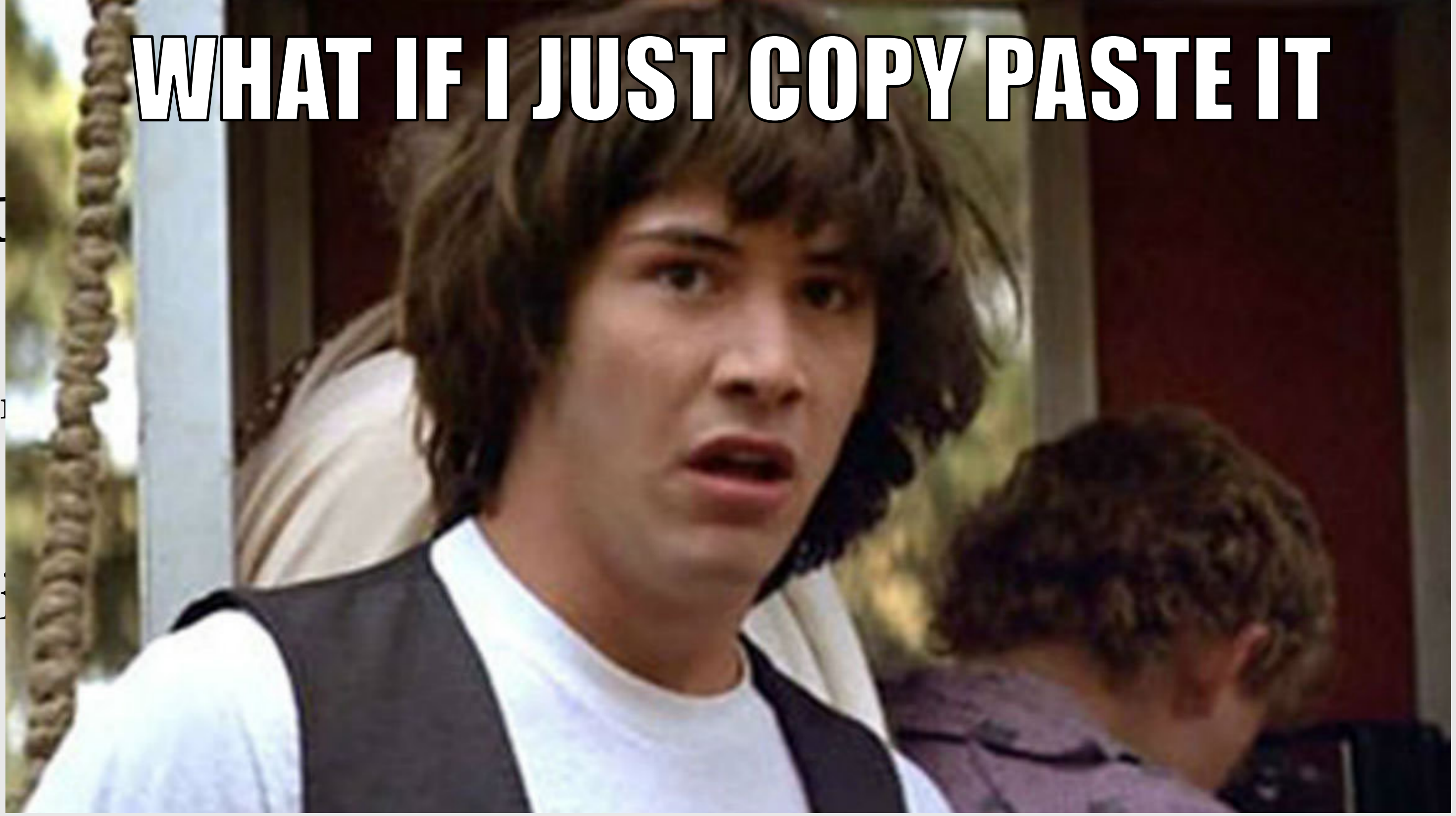


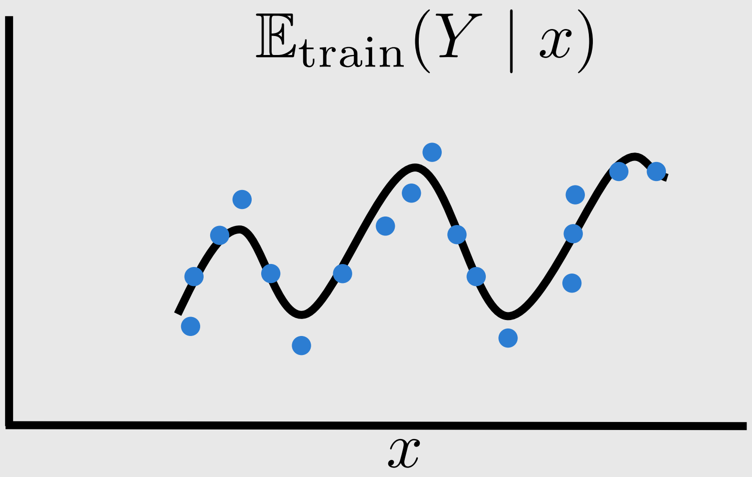
Covariat

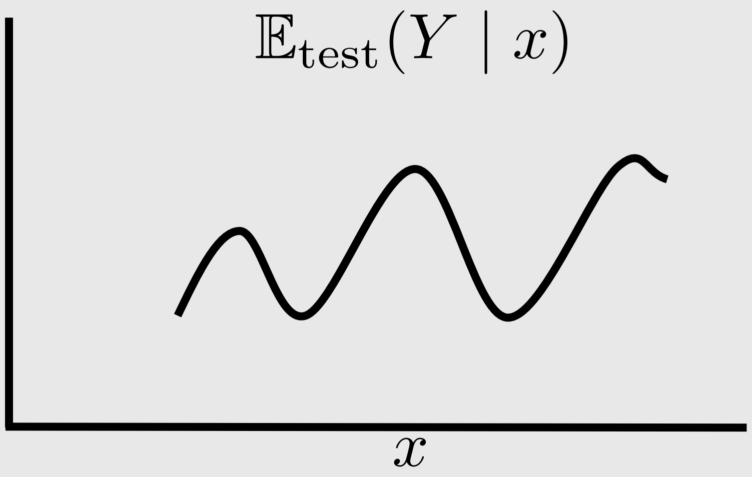
Setting: P_{trai}

Goal: Model

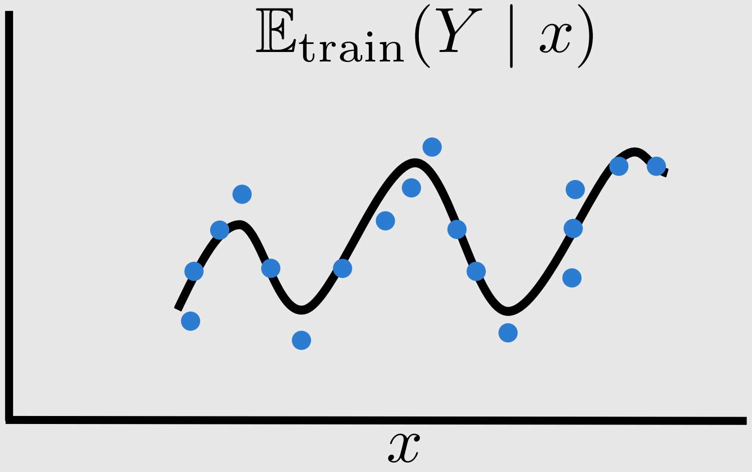
Covariate Shi



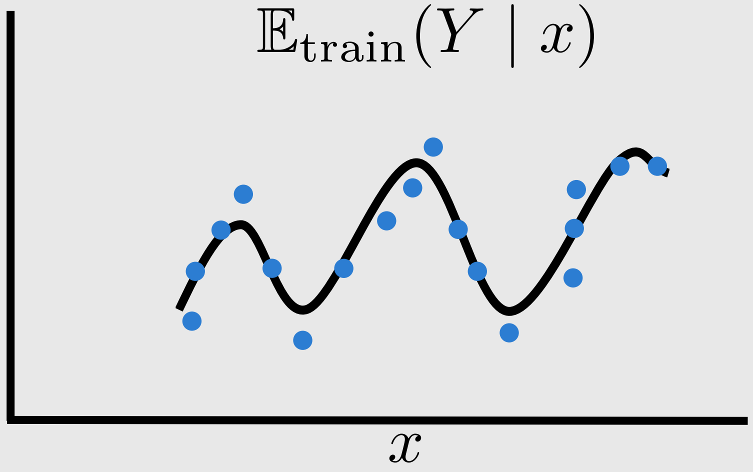


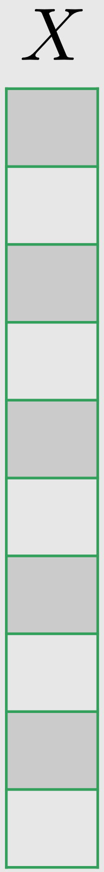


Predicting Y from an Unstructured Vector

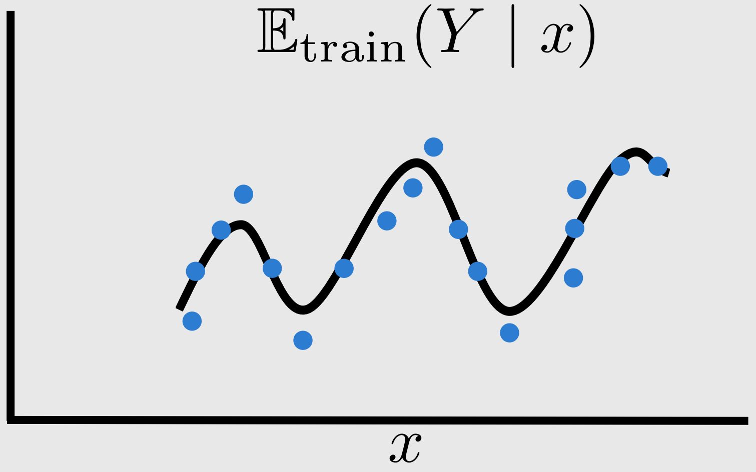


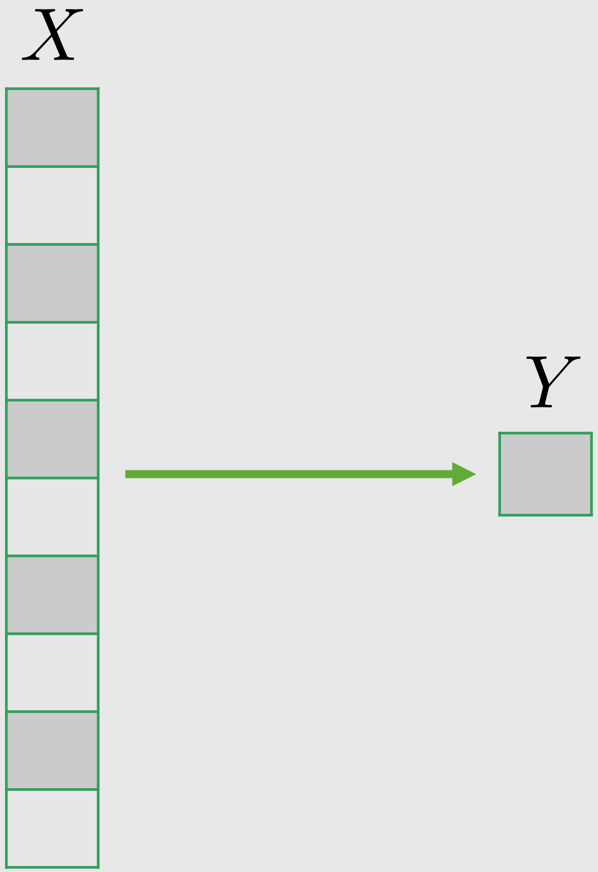
Predicting Y from an Unstructured Vector



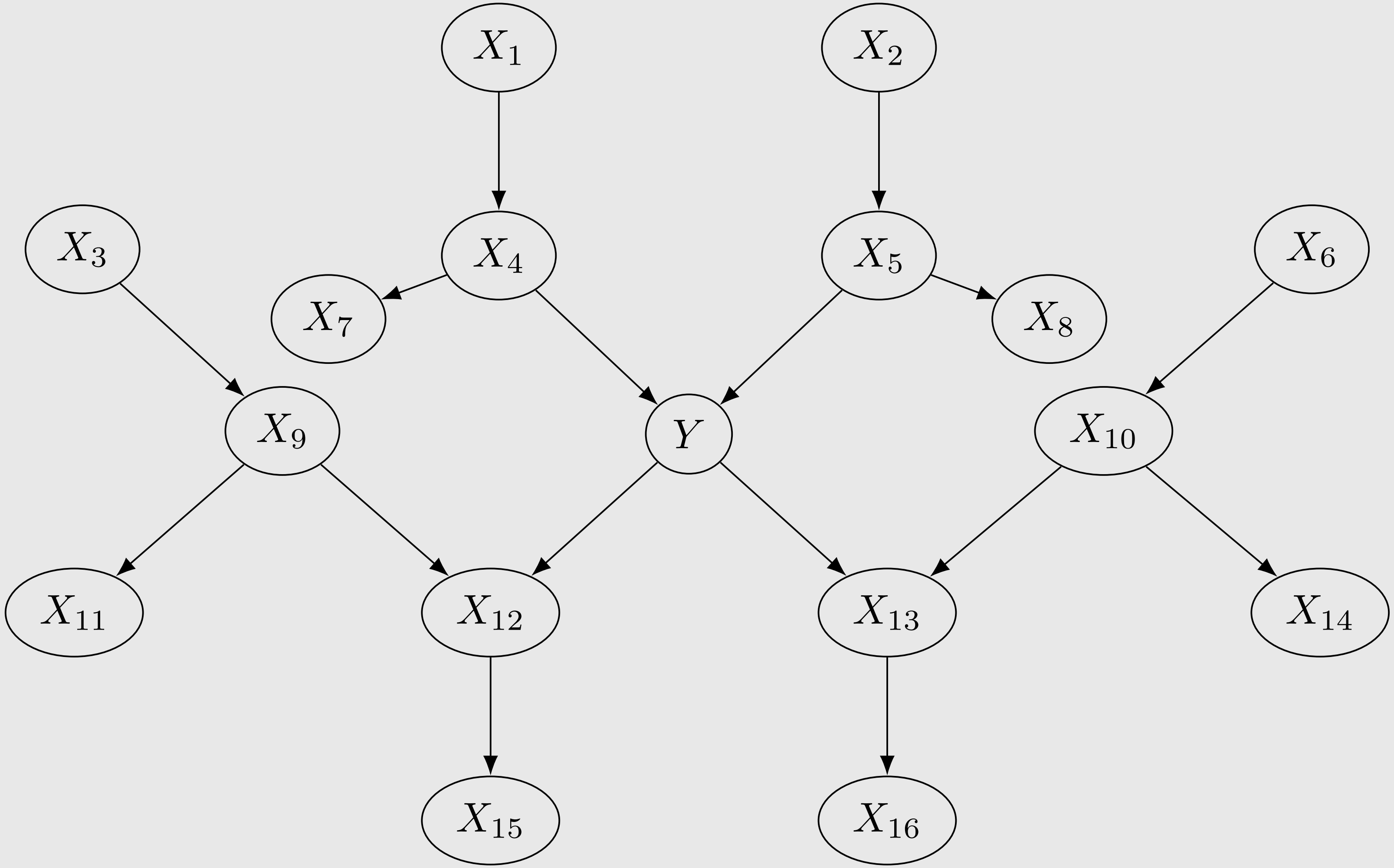


Predicting Y from an Unstructured Vector

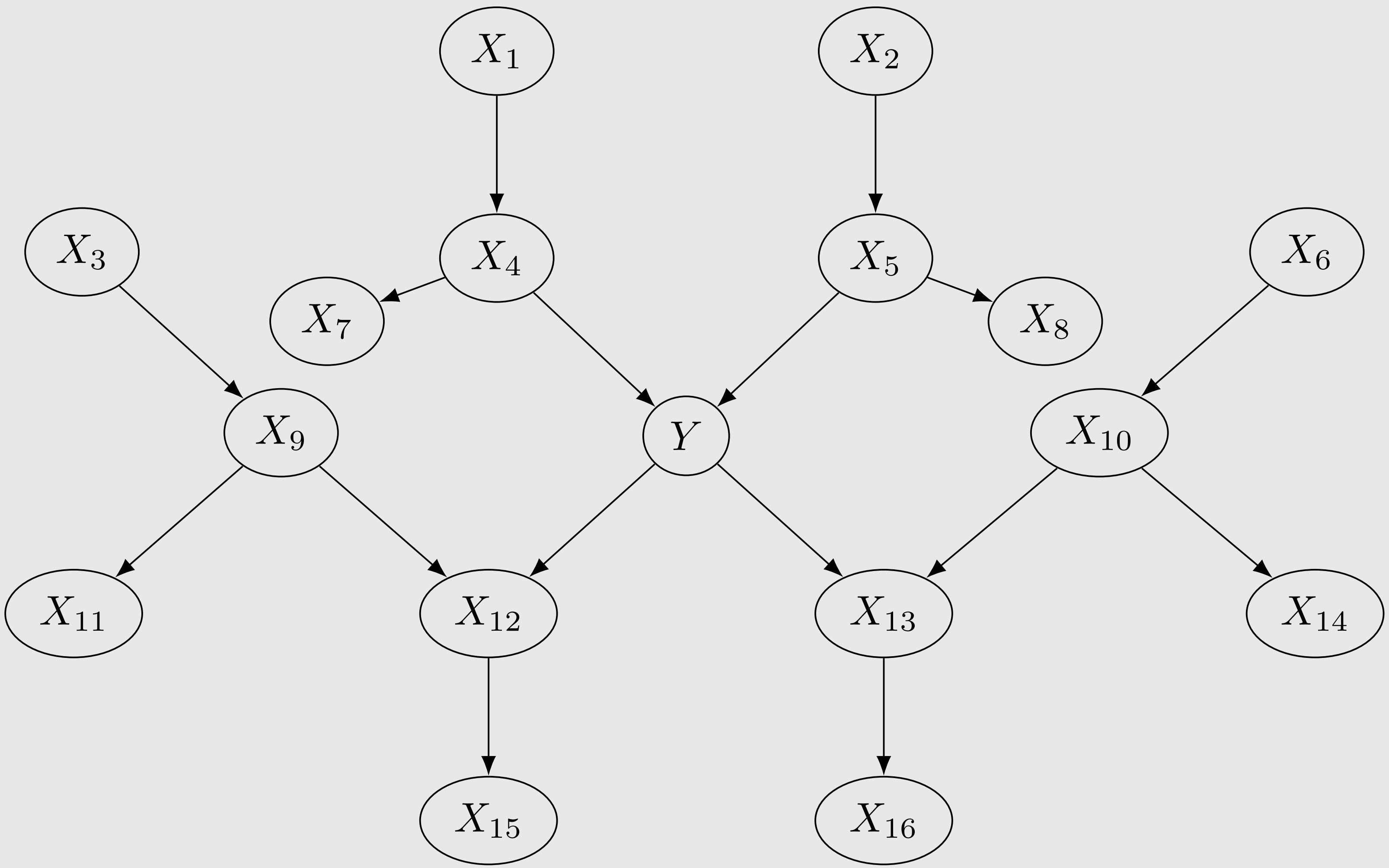




Use the Causal Structure

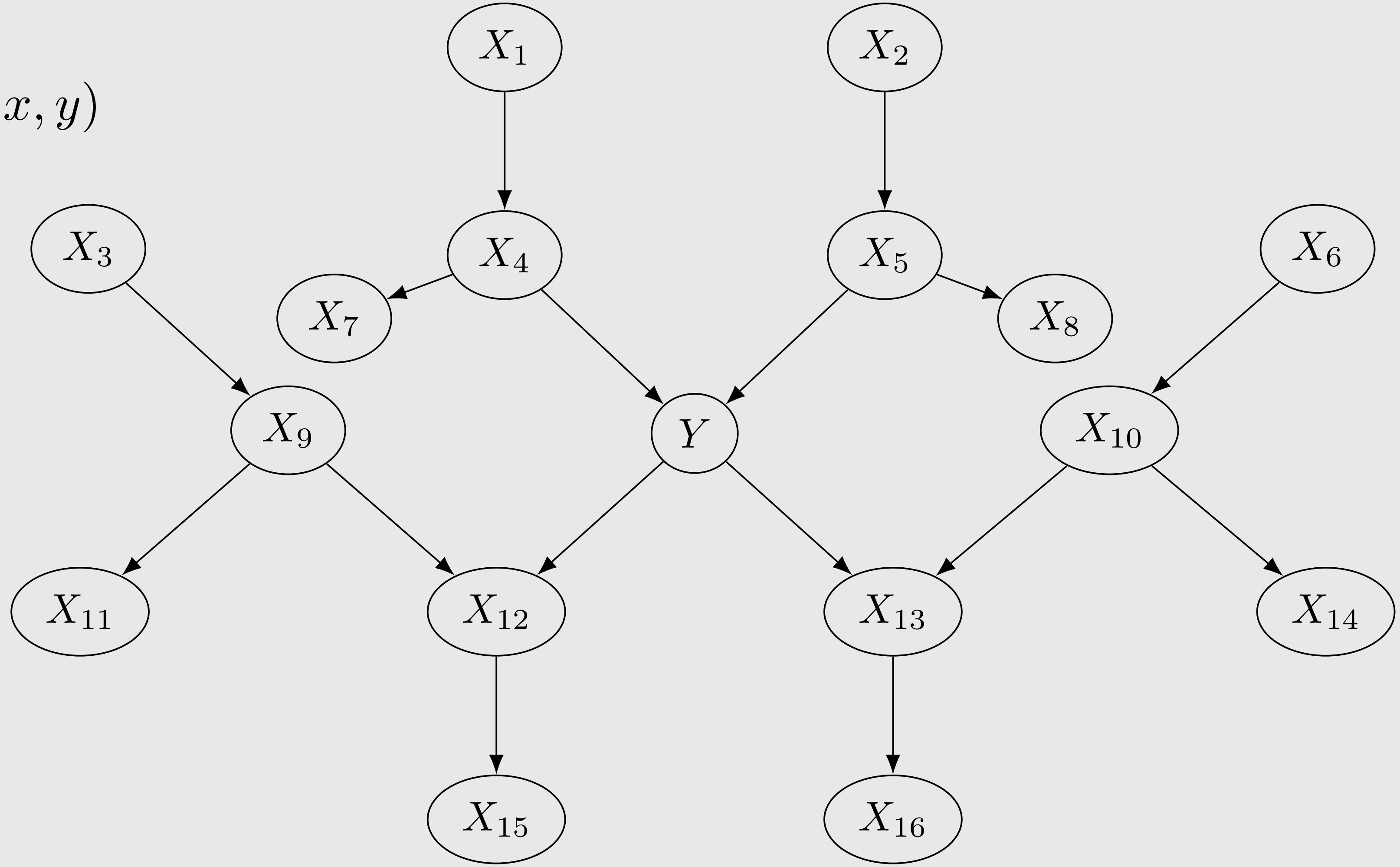


In-distribution Prediction of Y – Markov Blanket



In-distribution Prediction of Y – Markov Blanket

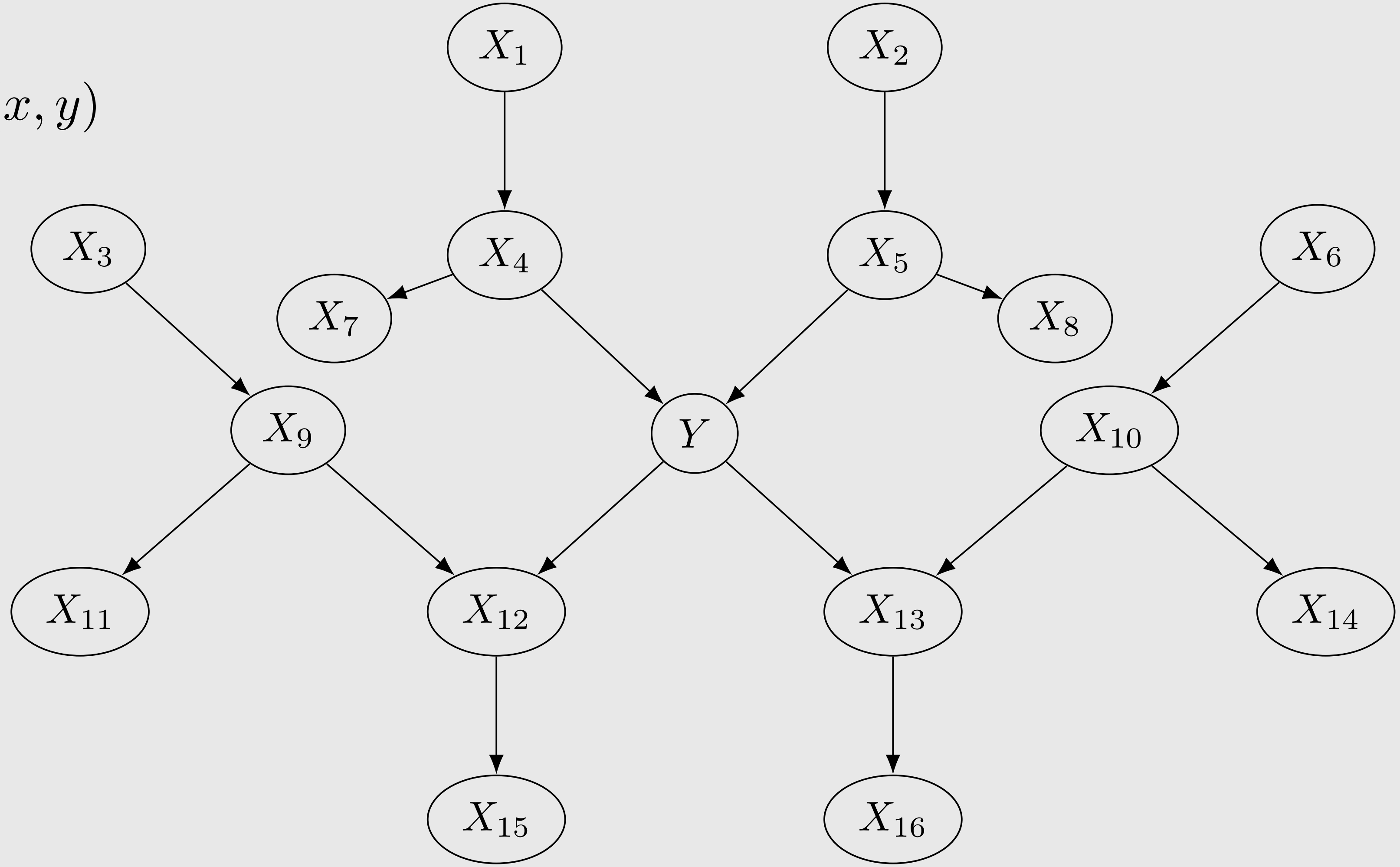
Training data from $P_{\text{train}}(x, y)$



In-distribution Prediction of Y – Markov Blanket

Training data from $P_{\text{train}}(x, y)$

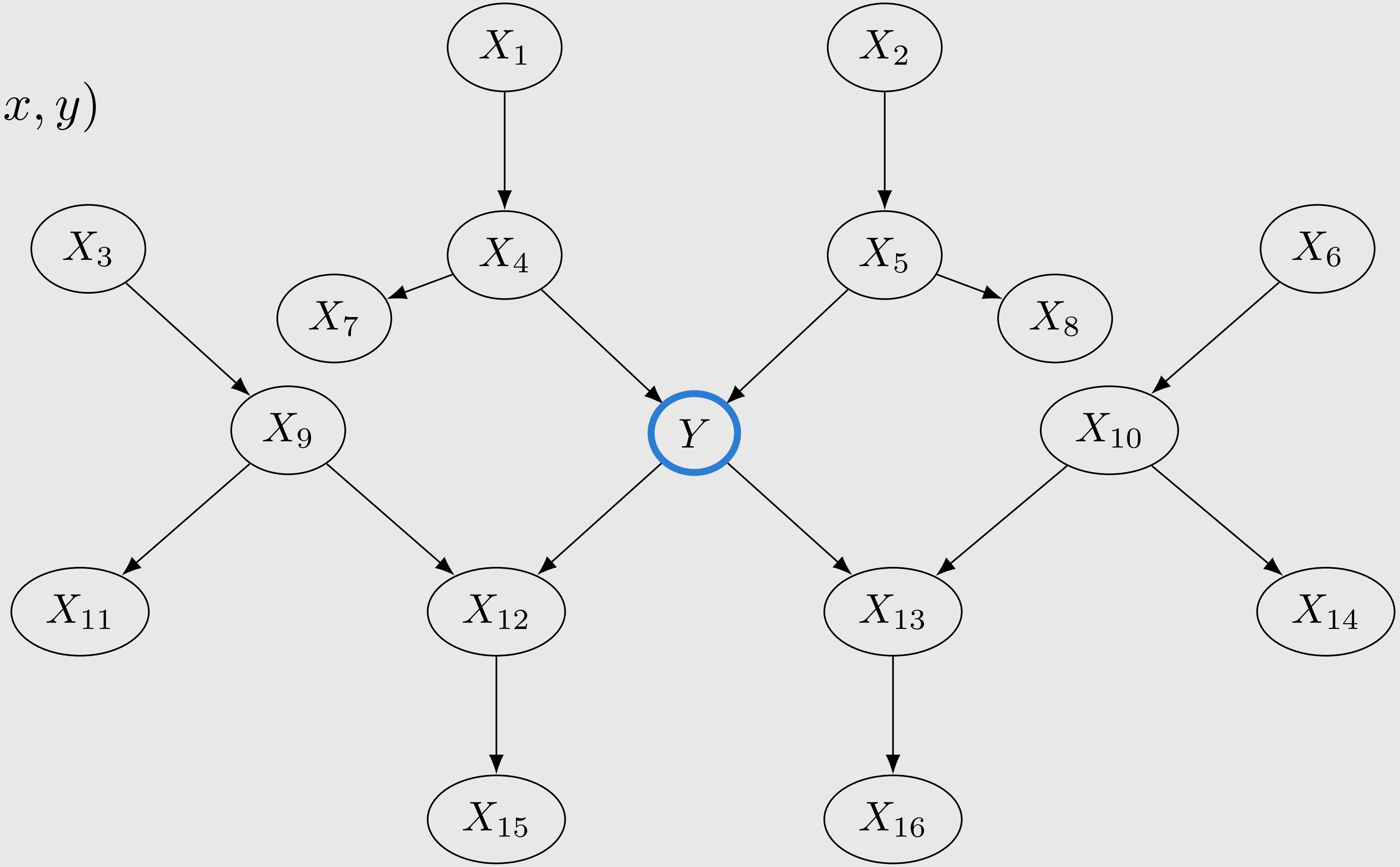
Goal: in-distribution prediction of Y from out-of-sample data for X



In-distribution Prediction of Y – Markov Blanket

Training data from $P_{\text{train}}(x, y)$

Goal: in-distribution prediction of Y from out-of-sample data for X

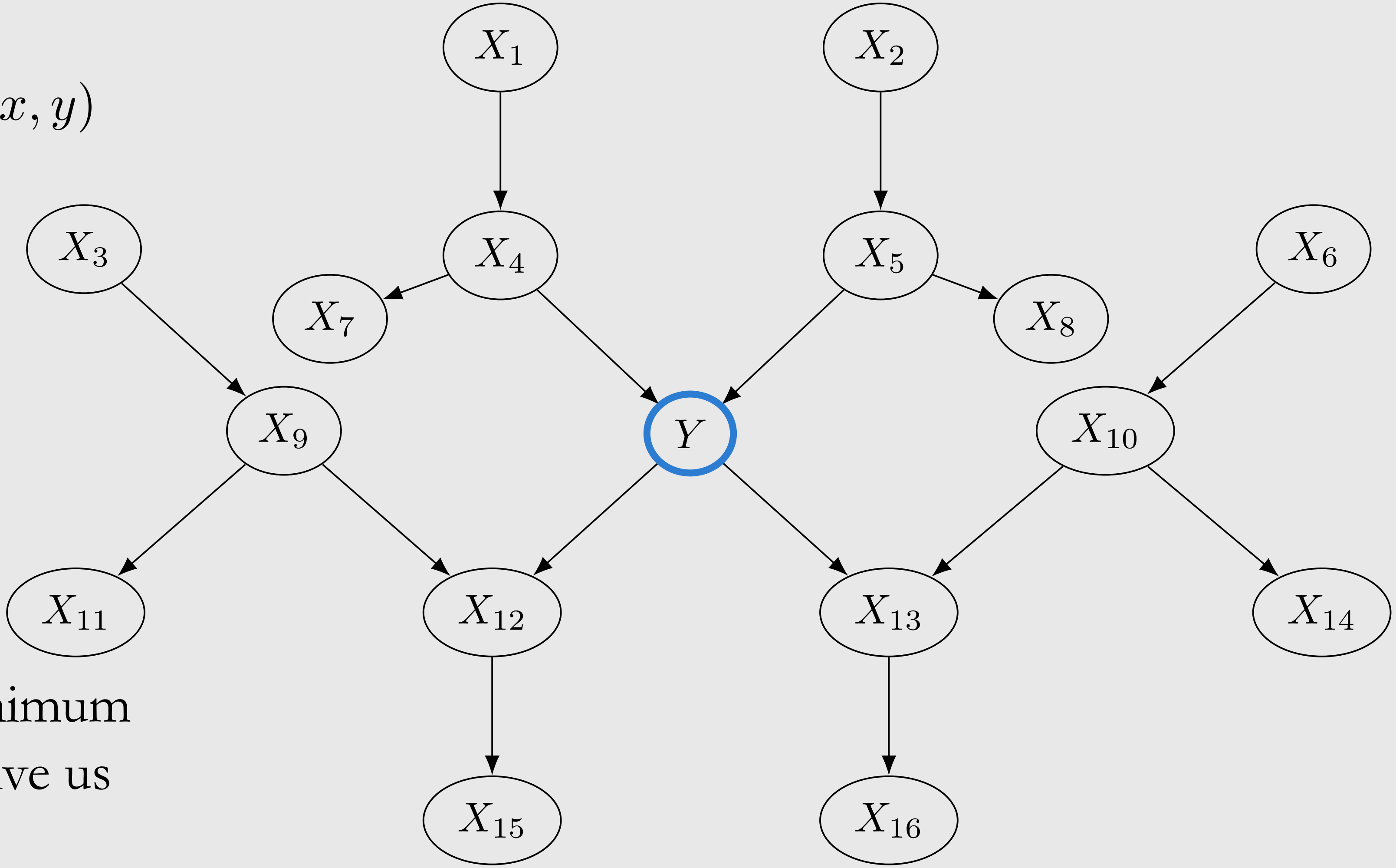


In-distribution Prediction of Y – Markov Blanket

Training data from $P_{\text{train}}(x, y)$

Goal: in-distribution prediction of Y from out-of-sample data for X

Question: What is the minimum set of variables that will give us optimal prediction?

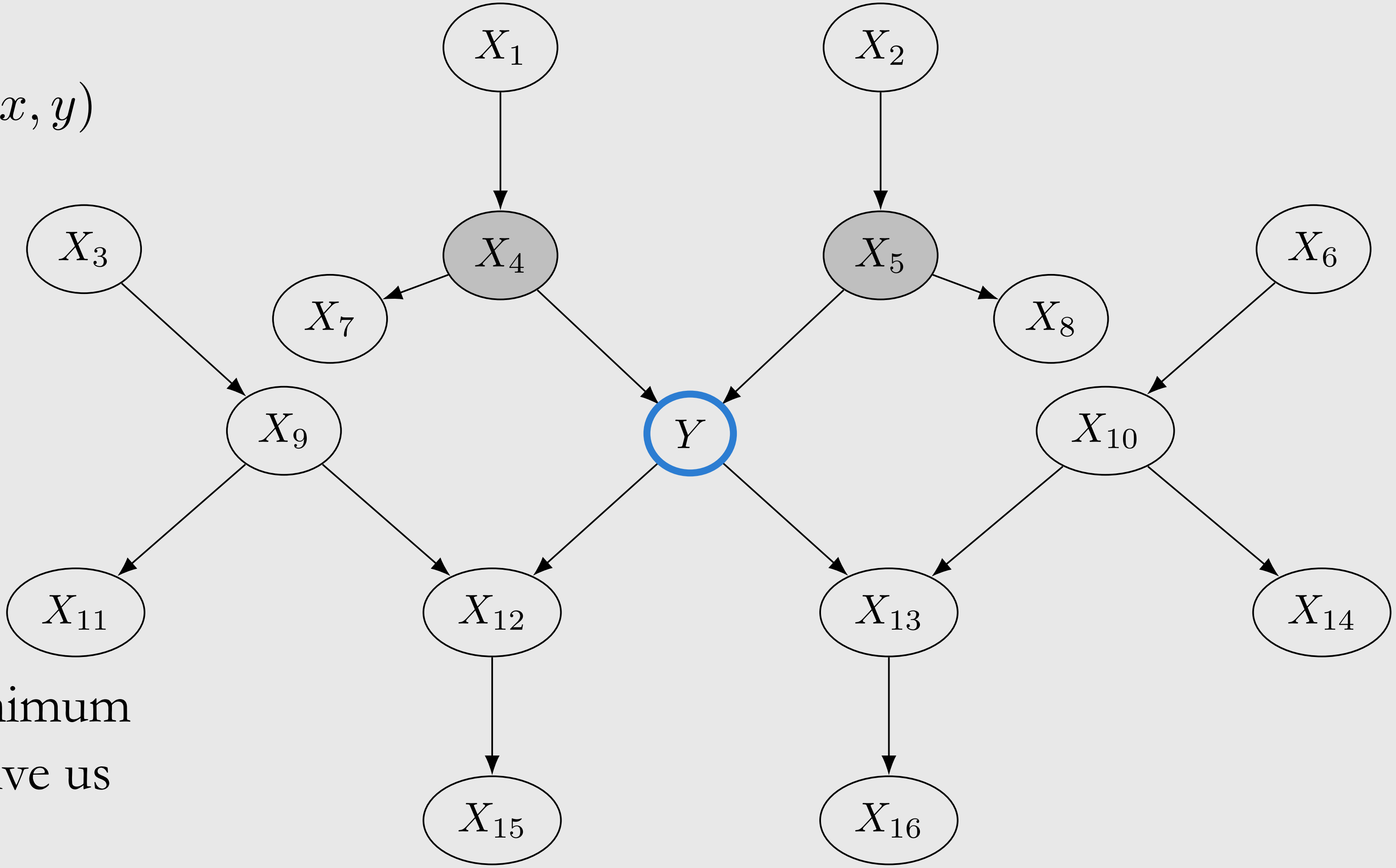


In-distribution Prediction of Y – Markov Blanket

Training data from $P_{\text{train}}(x, y)$

Goal: in-distribution prediction of Y from out-of-sample data for X

Question: What is the minimum set of variables that will give us optimal prediction?

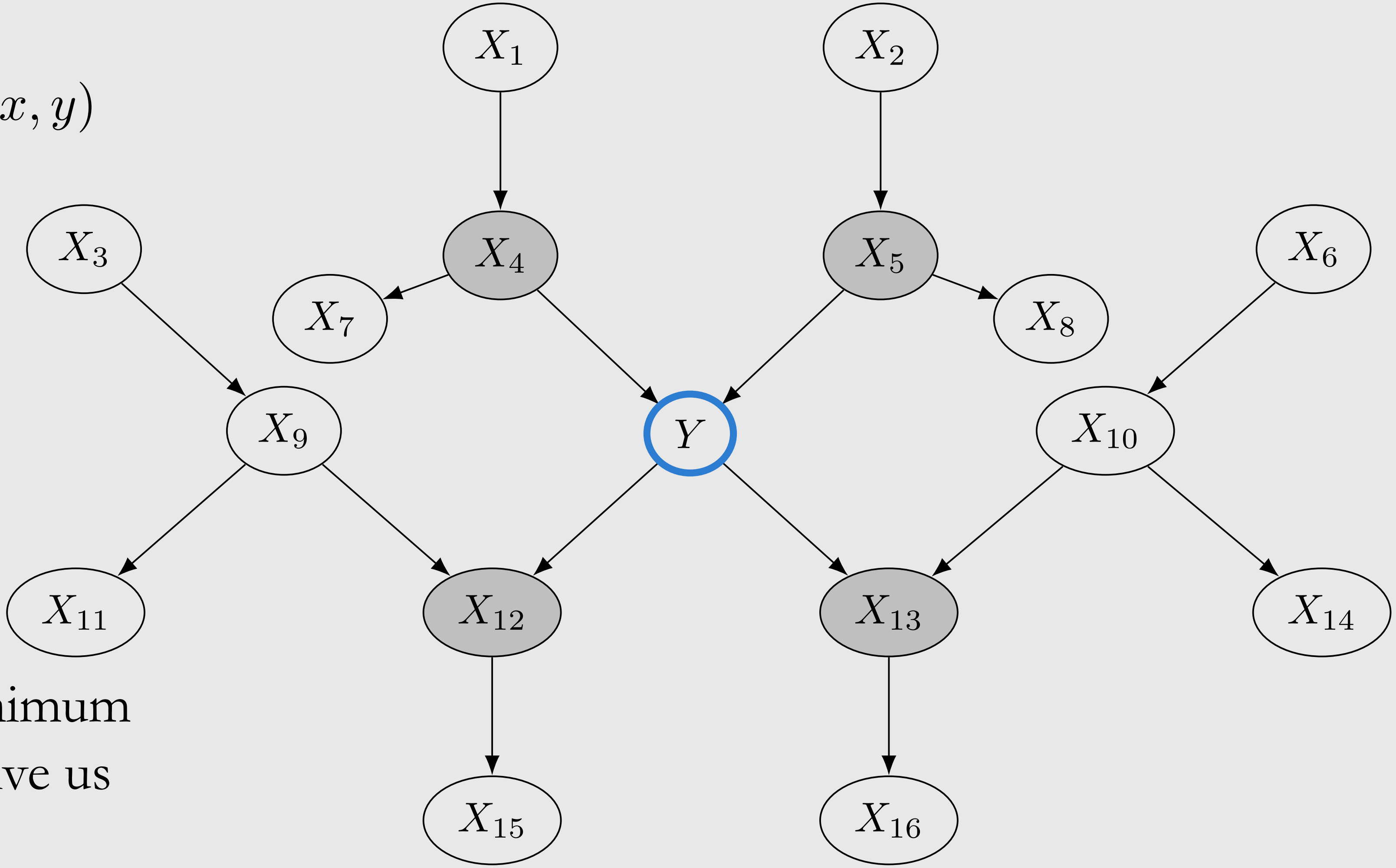


In-distribution Prediction of Y – Markov Blanket

Training data from $P_{\text{train}}(x, y)$

Goal: in-distribution prediction of Y from out-of-sample data for X

Question: What is the minimum set of variables that will give us optimal prediction?

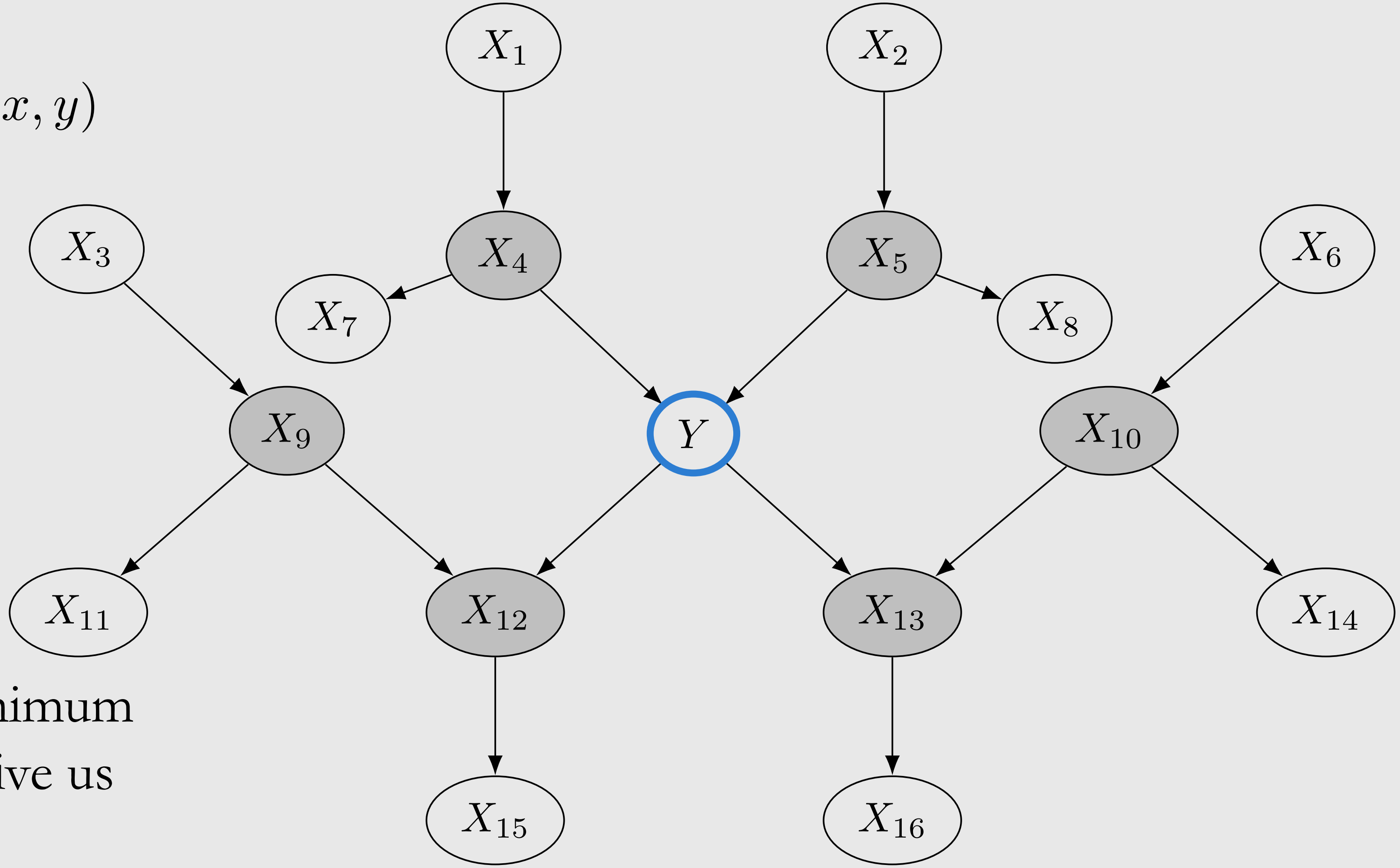


In-distribution Prediction of Y – Markov Blanket

Training data from $P_{\text{train}}(x, y)$

Goal: in-distribution prediction of Y from out-of-sample data for X

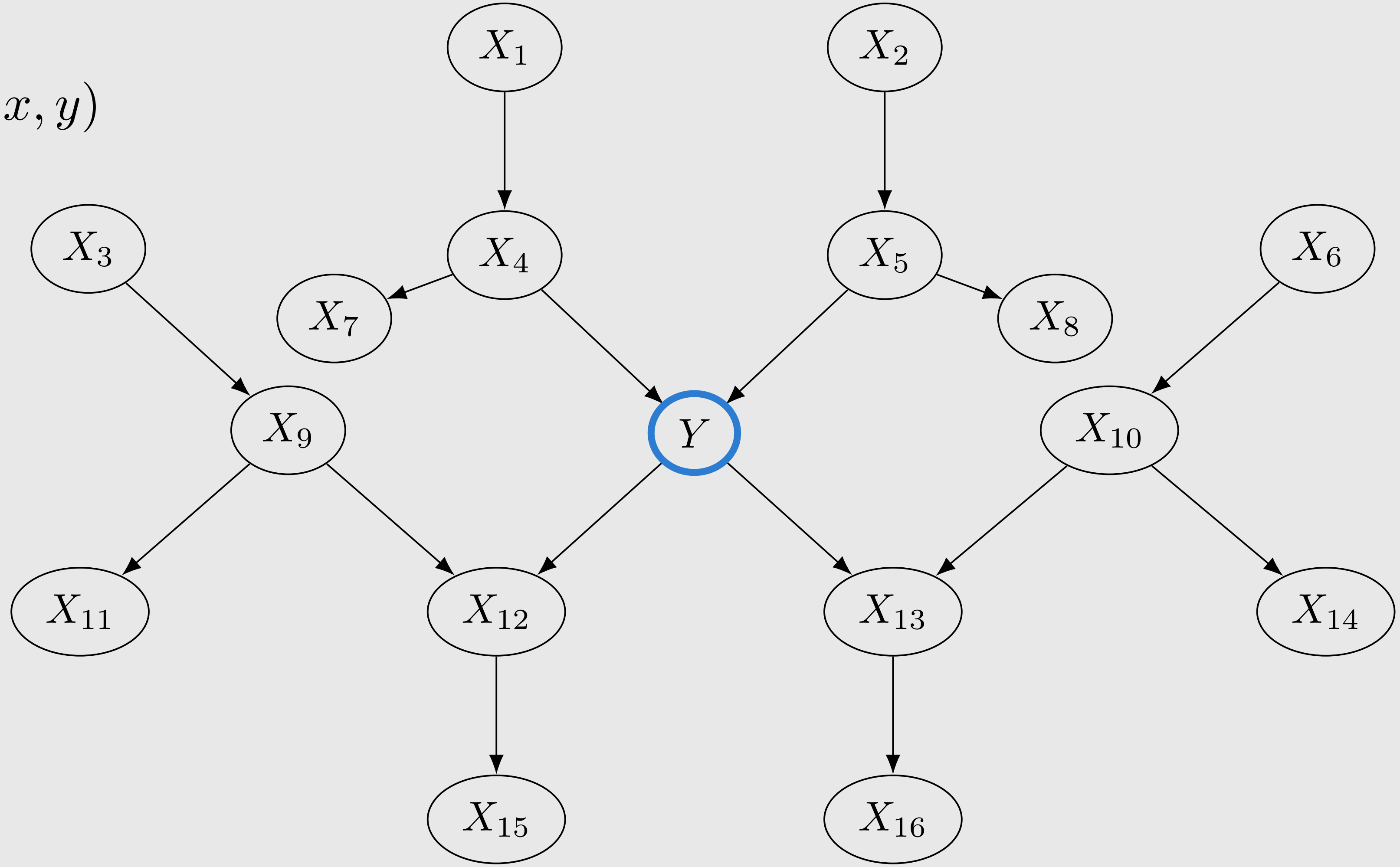
Question: What is the minimum set of variables that will give us optimal prediction?



Other Tasks Generated via Interventions

Training data from $P_{\text{train}}(x, y)$

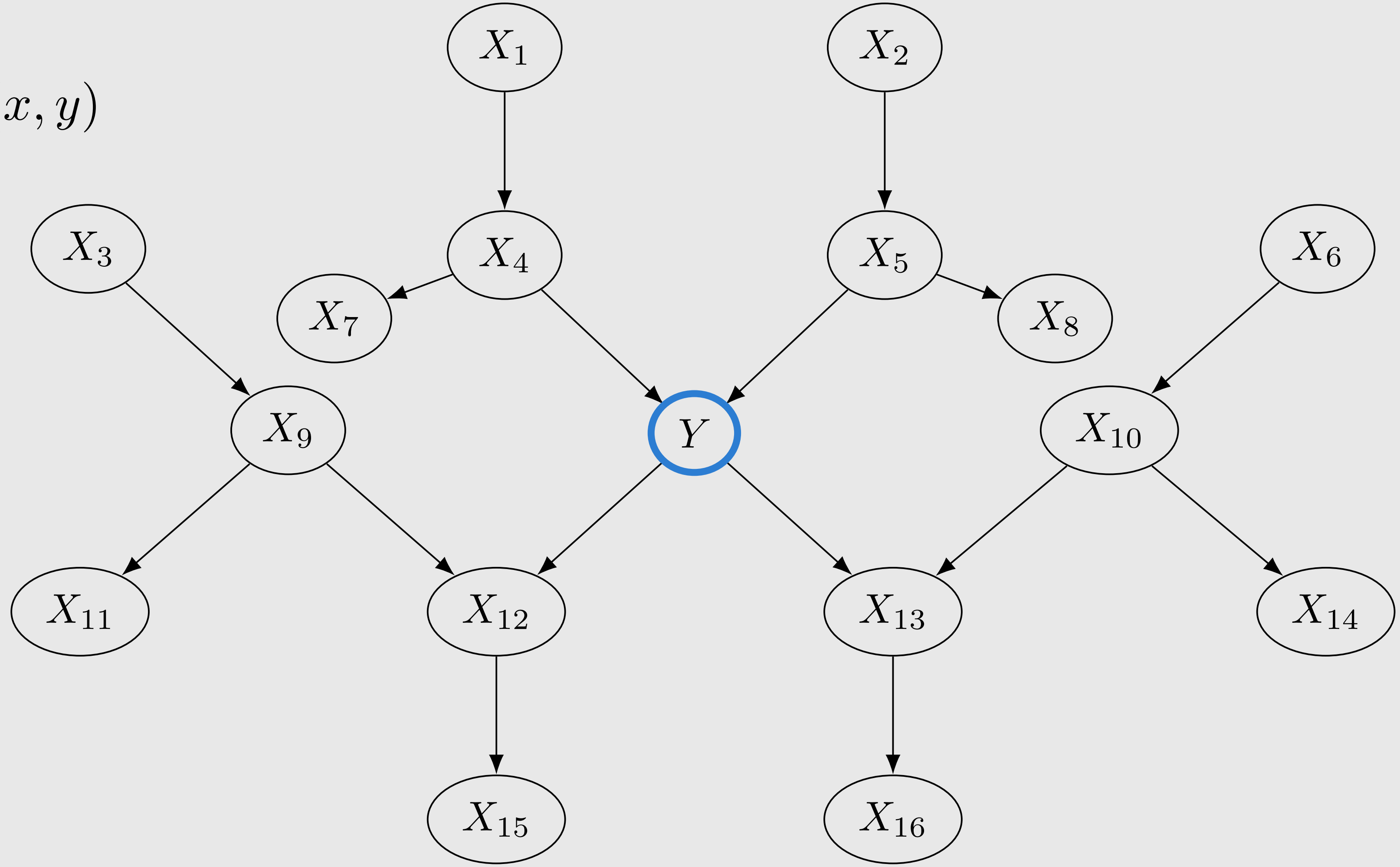
Goal: in-distribution prediction of Y from out-of-sample data for X



Other Tasks Generated via Interventions

Training data from $P_{\text{train}}(x, y)$

Goal: prediction of Y from X sampled from $P_{\text{test}}(x, y)$

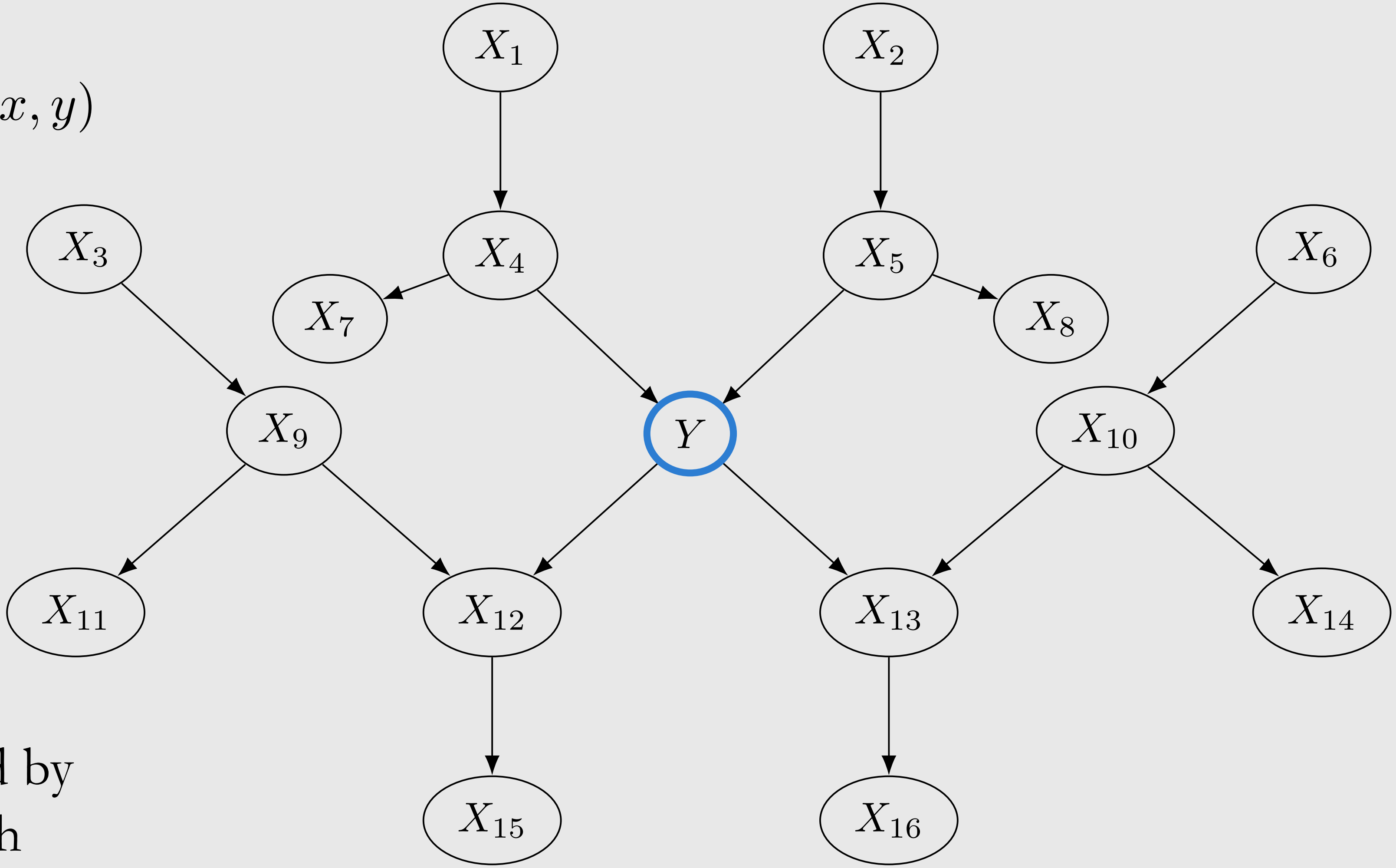


Other Tasks Generated via Interventions

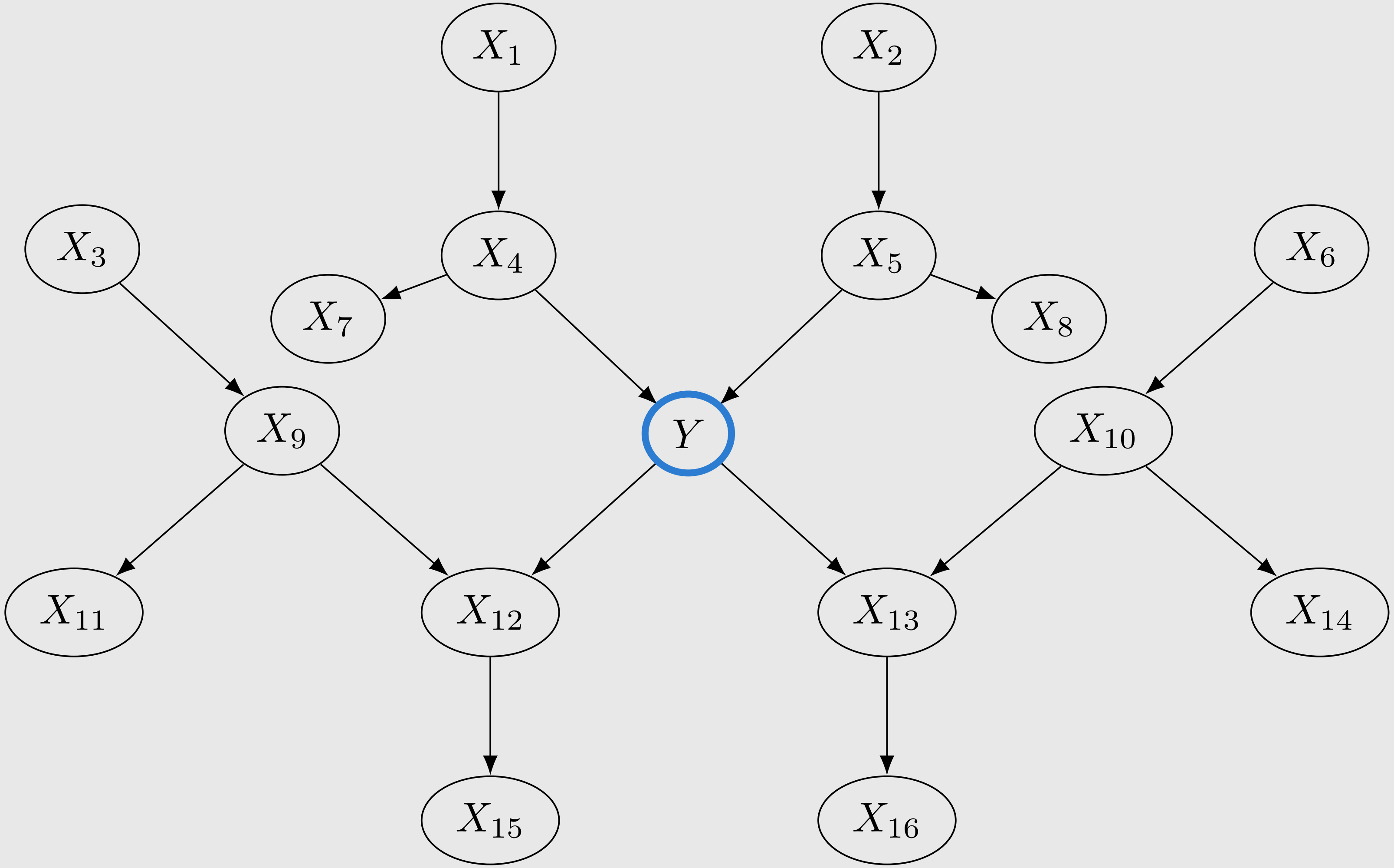
Training data from $P_{\text{train}}(x, y)$

Goal: prediction of Y from X sampled from $P_{\text{test}}(x, y)$

Consider that are all test distributions are generated by interventions on this graph

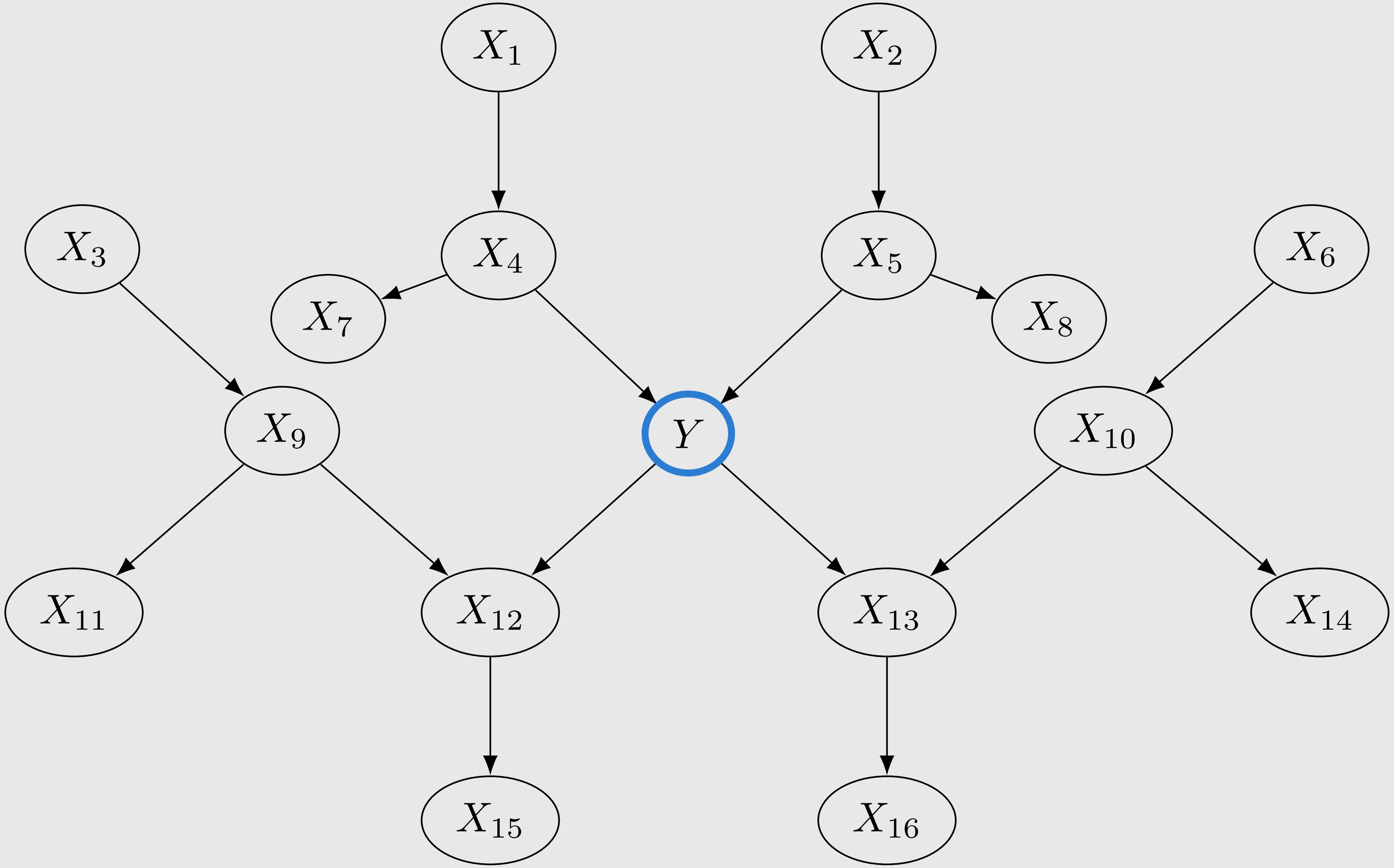


Recall Modularity



Recall Modularity

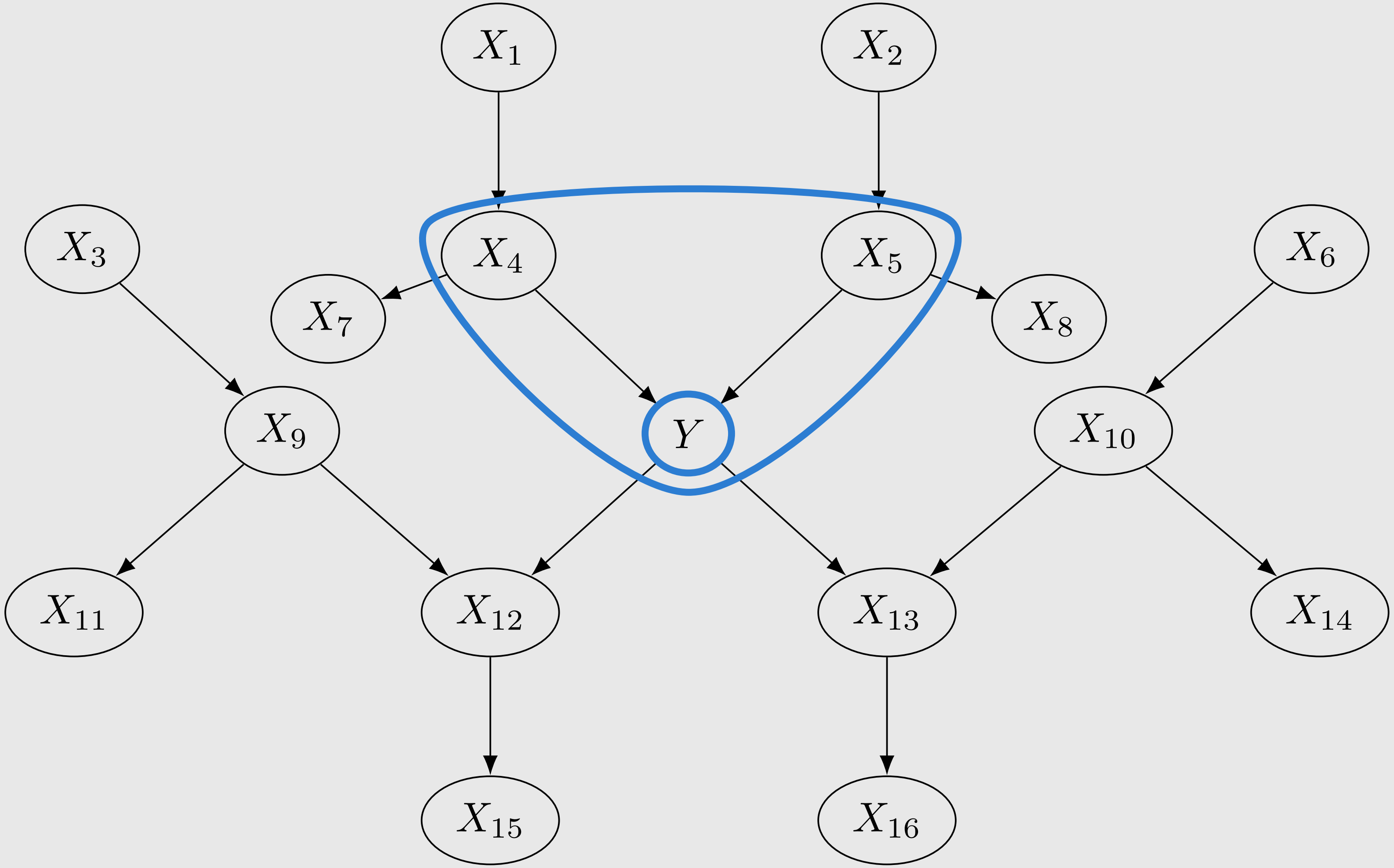
Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.



Recall Modularity

Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.

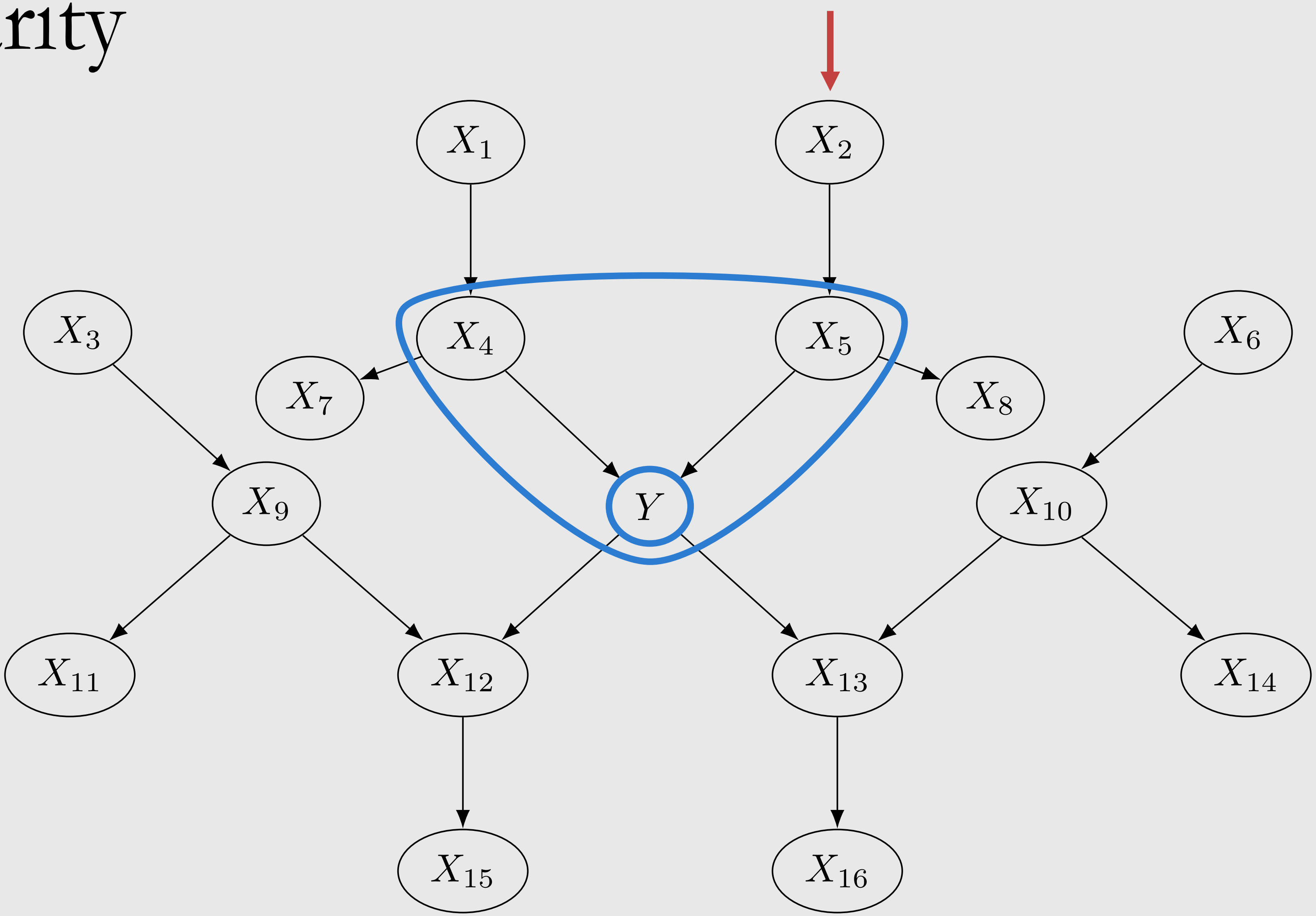
Mechanism for Y:
$P(y \mid x_4, x_5)$



Recall Modularity

Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.

Mechanism for Y:
$P(y \mid x_4, x_5)$

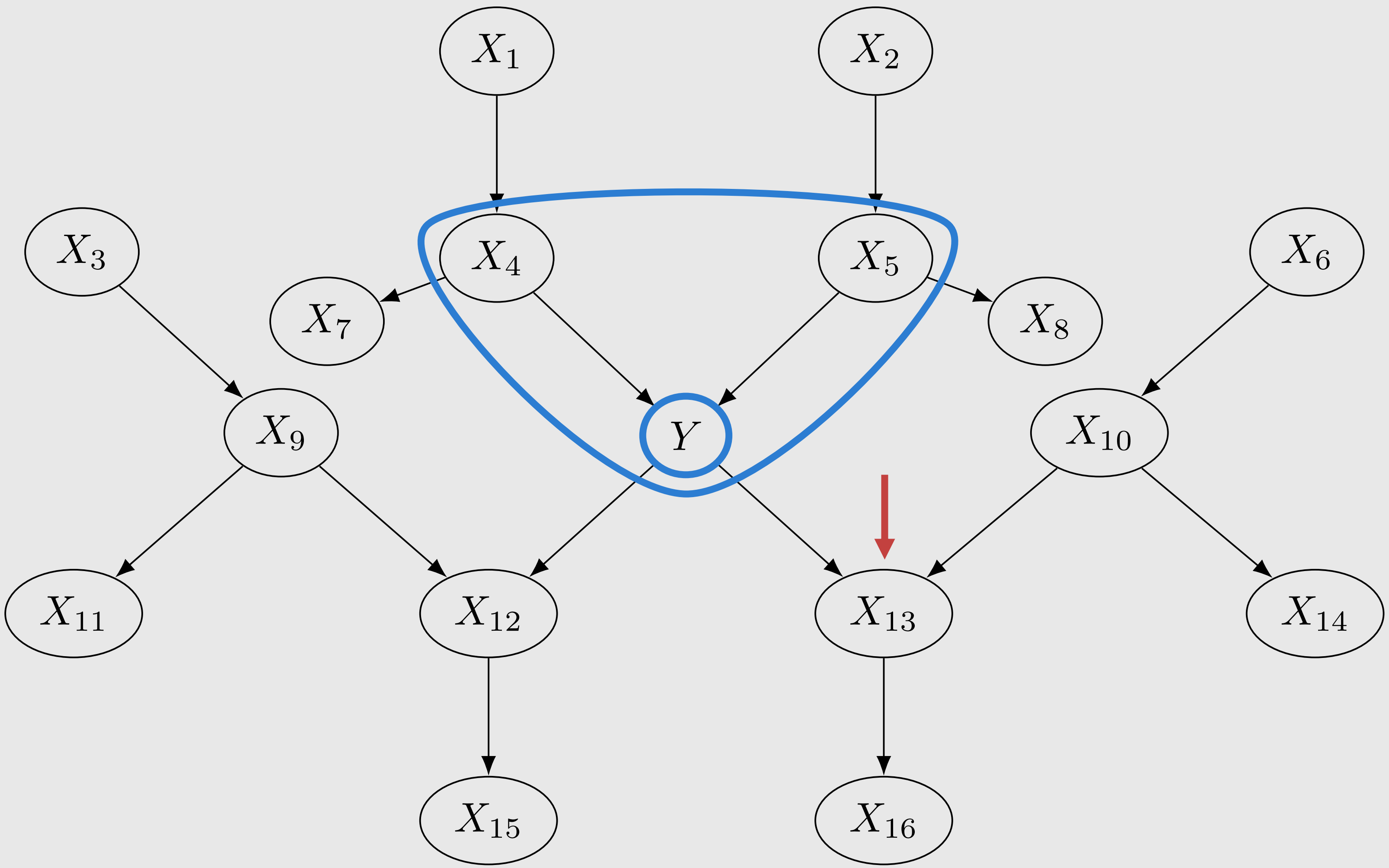


Recall Modularity

Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.

Mechanism for Y:

$$P(y \mid x_4, x_5)$$

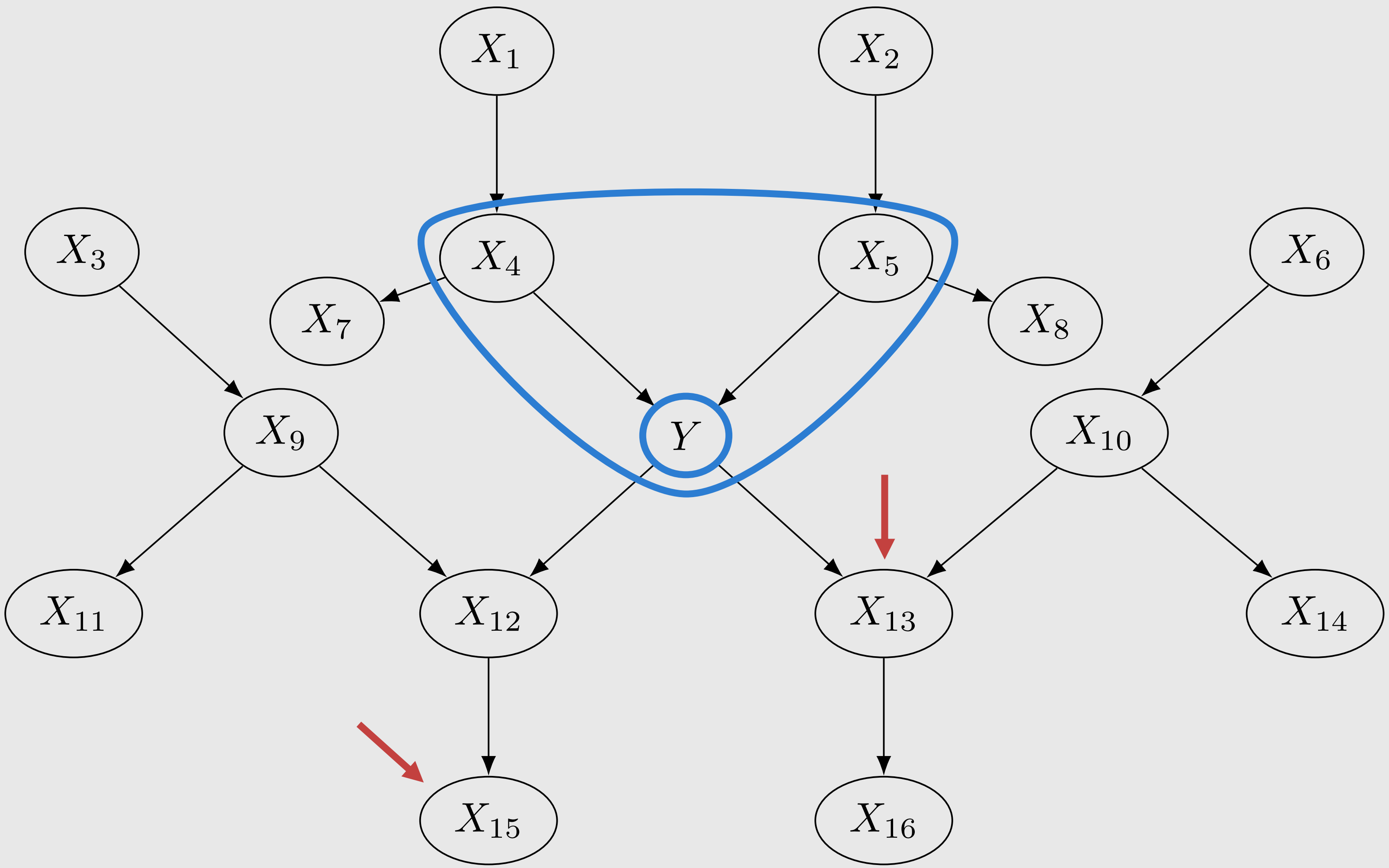


Recall Modularity

Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.

Mechanism for Y:

$P(y \mid x_4, x_5)$

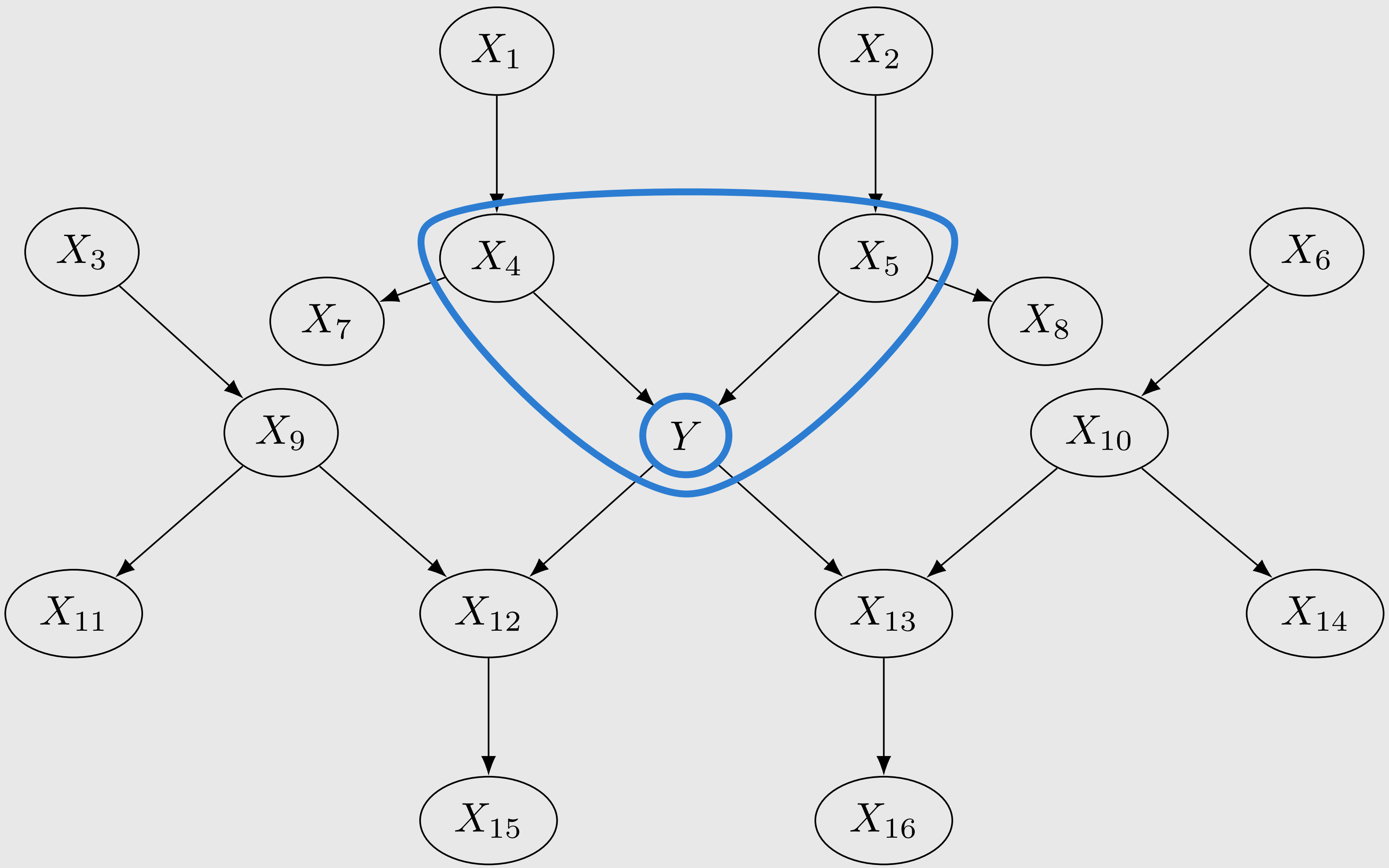


Recall Modularity

Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.

Mechanism for Y:
$$P(y \mid x_4, x_5)$$

Non-causal conditional:
$$P(y \mid x_4, x_5, x_{12})$$



Recall Modularity

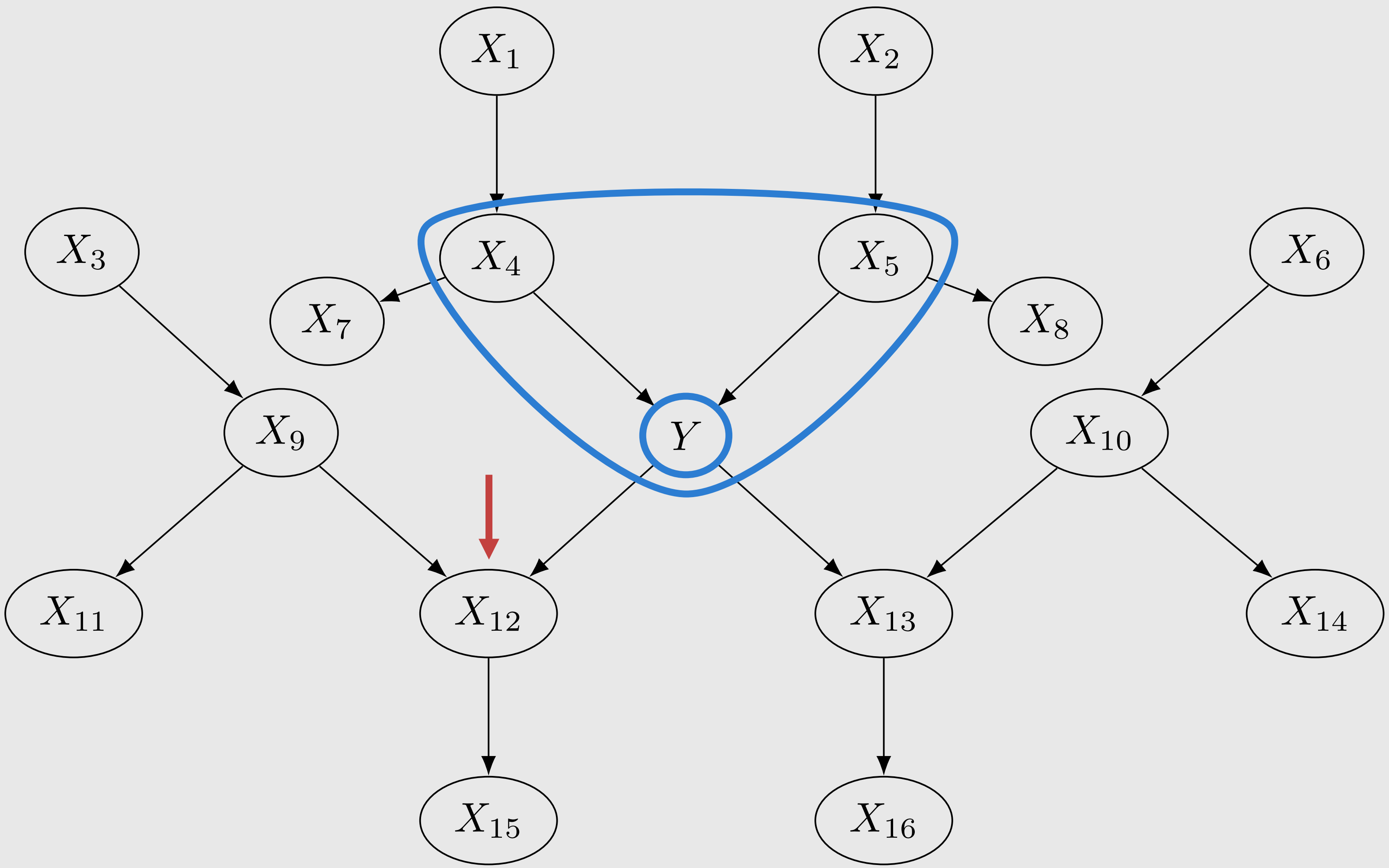
Intervening on a variable only changes the causal mechanism (structural equation) for that variable. All other causal mechanisms remain unchanged.

Mechanism for Y:

$P(y \mid x_4, x_5)$

Non-causal conditional:

$P(y \mid x_4, x_5, x_{12})$



Causal Mechanism is Optimal in Robust Sense

Causal Mechanism is Optimal in Robust Sense

$$\mathbb{E}[Y \mid \mathrm{pa}(Y)] = \arg\min_f \max_{P_{\text{test}}} \mathbb{E}_{(X,Y)\sim P_{\text{test}}} (Y - f(X))^2$$

(see, e.g., Rojas-Carulla et al., (2018, Appendix A.1))

Causal Mechanism is Optimal in Robust Sense

$$\mathbb{E}[Y \mid \mathrm{pa}(Y)] = \arg\min_f \max_{P_{\mathrm{test}}} \mathbb{E}_{(X,Y)\sim P_{\mathrm{test}}} (Y - f(X))^2$$

(see, e.g., Rojas-Carulla et al., (2018, Appendix A.1))

Still requires common support: $\mathrm{supp}_{\mathrm{train}}(\mathrm{pa}(Y)) = \mathrm{supp}_{\mathrm{test}}(\mathrm{pa}(Y))$

Causal Mechanism is Optimal in Robust Sense

$$\mathbb{E}[Y \mid \mathrm{pa}(Y)] = \arg\min_f \max_{P_{\mathrm{test}}} \mathbb{E}_{(X,Y)\sim P_{\mathrm{test}}} (Y - f(X))^2$$

(see, e.g., Rojas-Carulla et al., (2018, Appendix A.1))

Still requires common support: $\mathrm{supp}_{\mathrm{train}}(\mathrm{pa}(Y)) = \mathrm{supp}_{\mathrm{test}}(\mathrm{pa}(Y))$

Or that we can extrapolate well from $\mathrm{supp}_{\mathrm{train}}(\mathrm{pa}(Y))$ to $\mathrm{supp}_{\mathrm{test}}(\mathrm{pa}(Y))$

Relaxation of Covariate Shift

Relaxation of Covariate Shift

Covariate shift: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

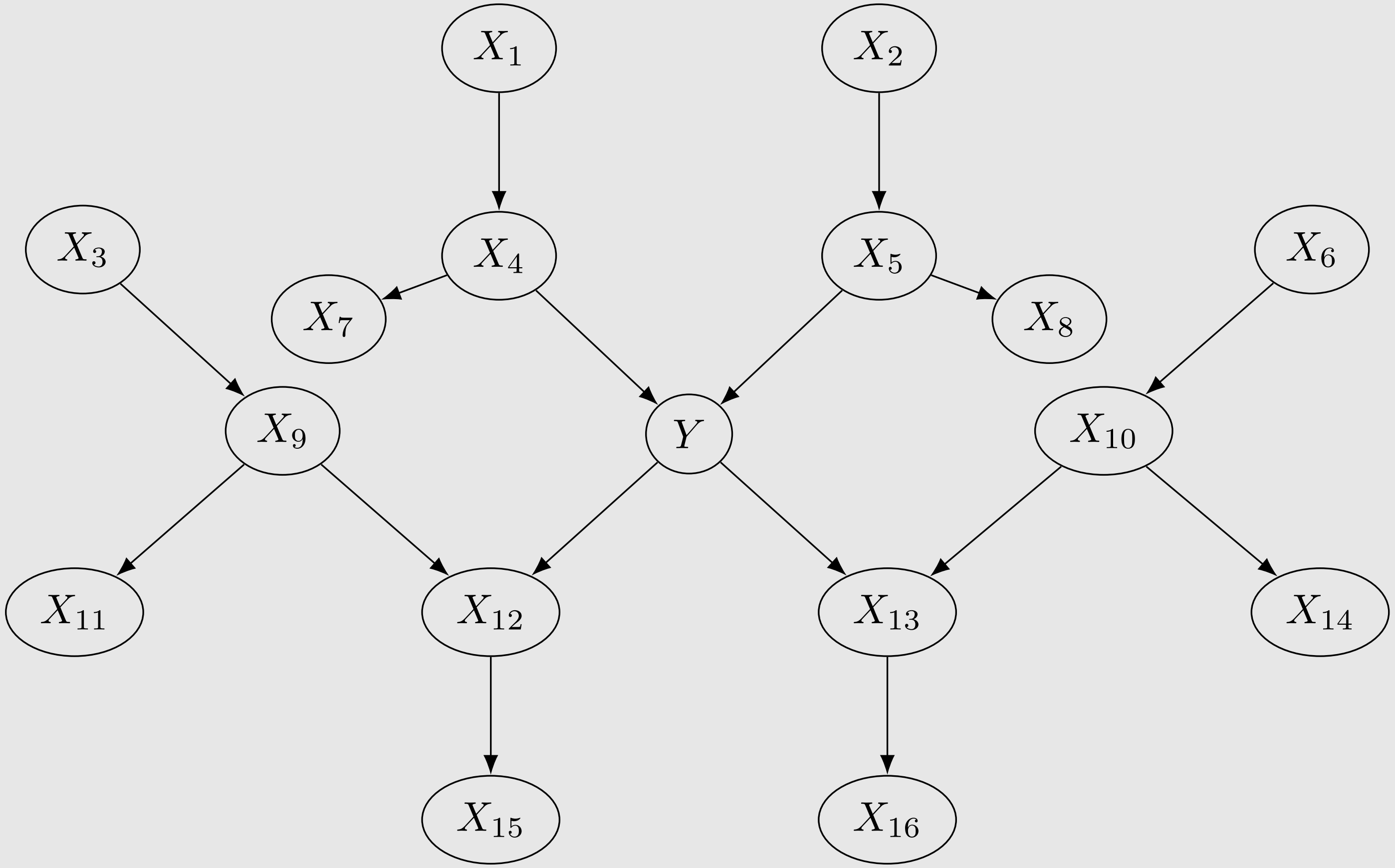
Relaxation of Covariate Shift

Covariate shift: $P_{\text{train}}(y \mid x) = P_{\text{test}}(y \mid x)$

Modularity: $P_{\text{train}}(y \mid \text{pa}(Y)) = P_{\text{test}}(y \mid \text{pa}(Y))$

Questions:

1. What is the Markov blanket of Y in this graph?
2. What task is the Markov blanket good for?
3. What input variables should we use for optimal robust prediction?



Causal Insights for Transfer Learning

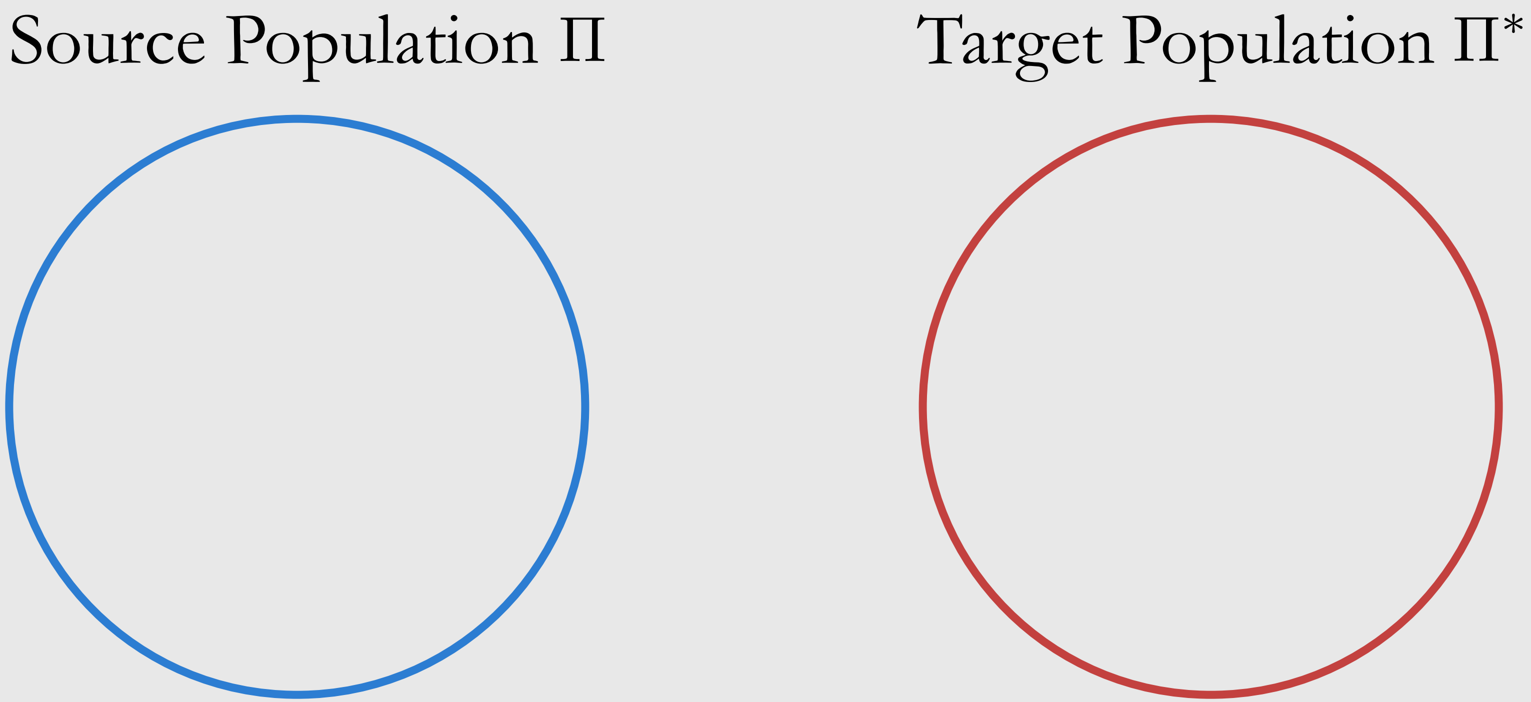
Transportability of Causal Effects Across Populations

Transportability Problem

Transportability Problem

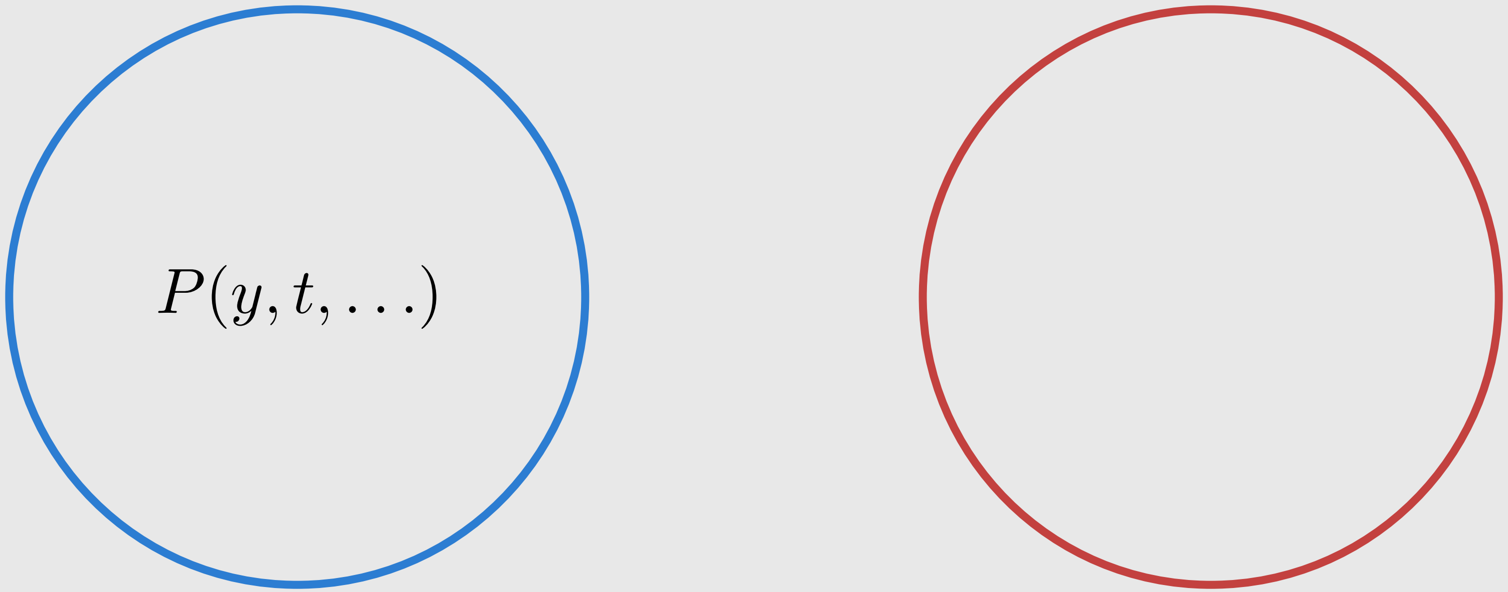
Source Population Π

Transportability Problem

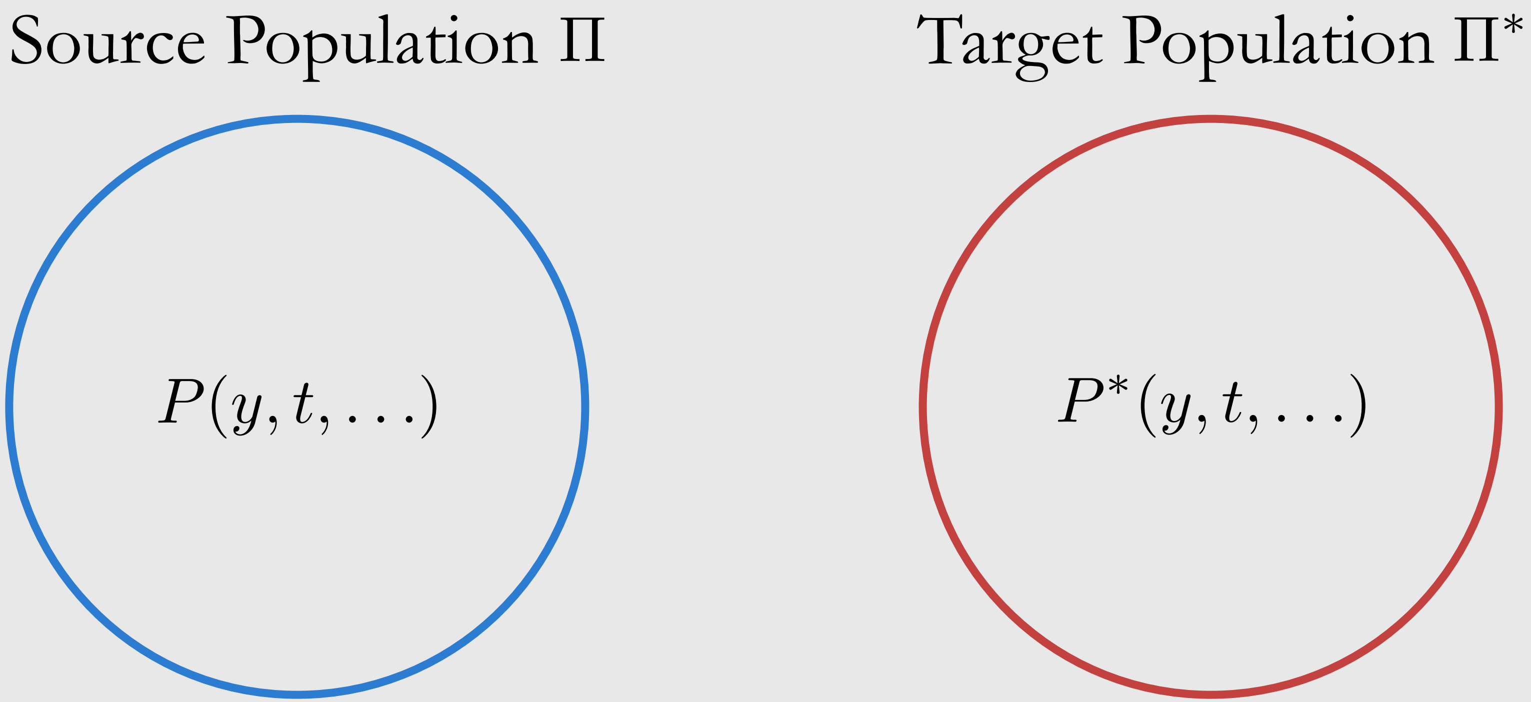


Transportability Problem

Source Population Π | Target Population Π^*

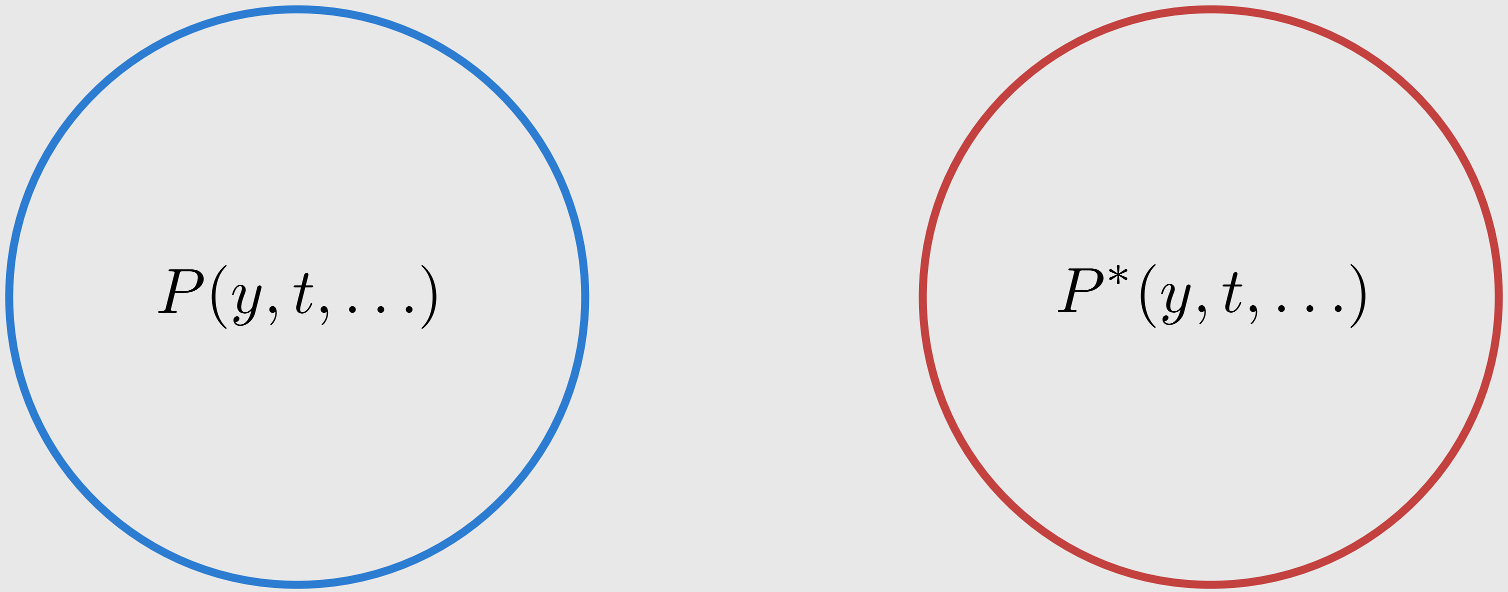


Transportability Problem



Transportability Problem

Source Population Π Target Population Π^*

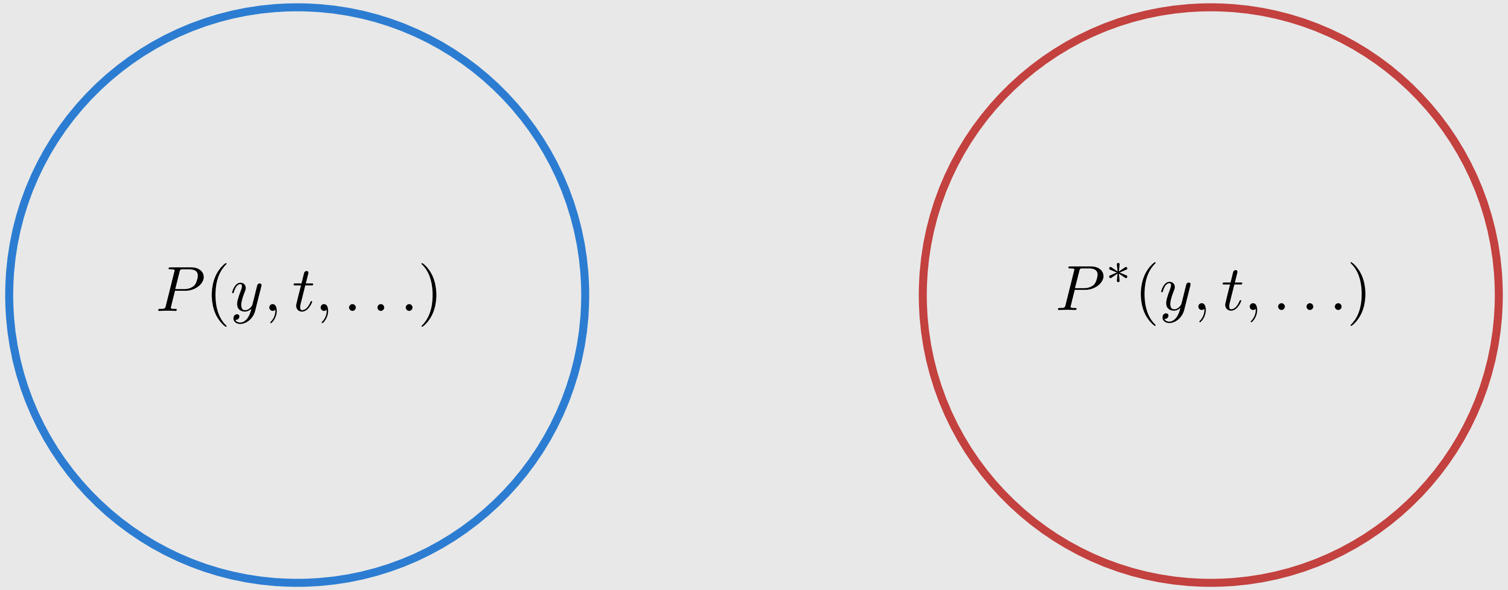


Given $P(y \mid do(t), x)$

Transportability Problem

Source Population Π

Target Population Π^*



Given $P(y \mid do(t), x)$

Goal: $P^*(y \mid do(t), x)$

Transportability Problem

Source Population Π

Target Population Π^*

$P(y, t, \ldots)$

$P^*(y, t, \ldots)$

Given $P(y \mid do(t), x)$

$P(y \mid do(t), x) \stackrel{?}{=} P^*(y \mid do(t), x)$

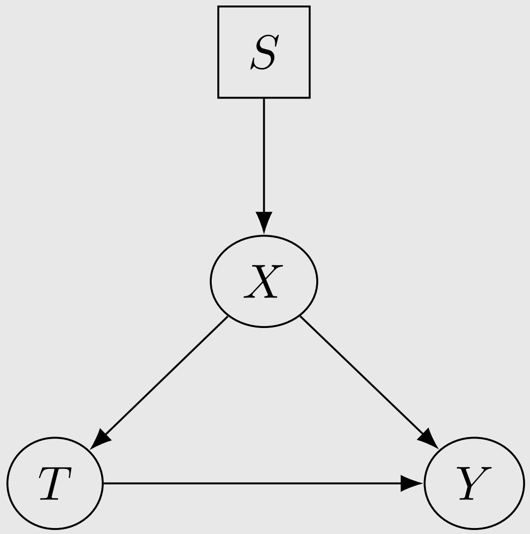
Selection Diagrams

Selection Diagrams

Allow for different causal mechanisms across the two distributions

Selection Diagrams

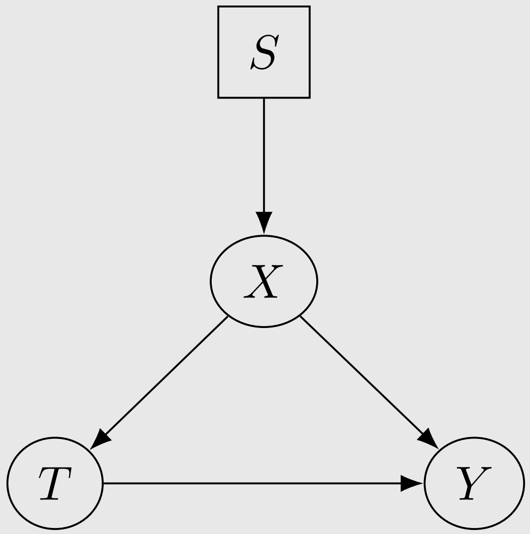
Allow for different causal mechanisms across the two distributions



Selection Diagrams

Allow for different causal mechanisms across the two distributions

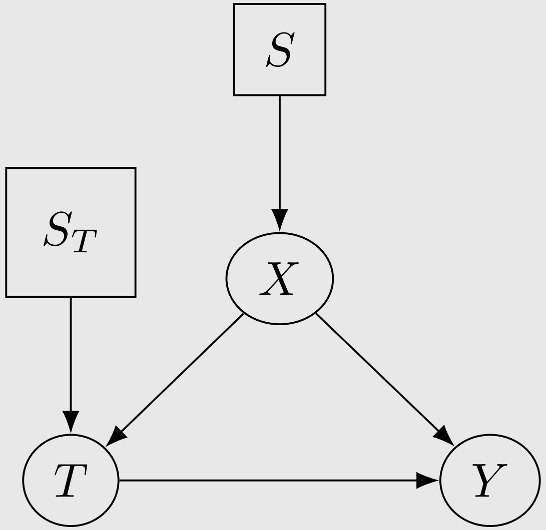
Mechanism that differs between Π and Π^*



Selection Diagrams

Allow for different causal mechanisms across the two distributions

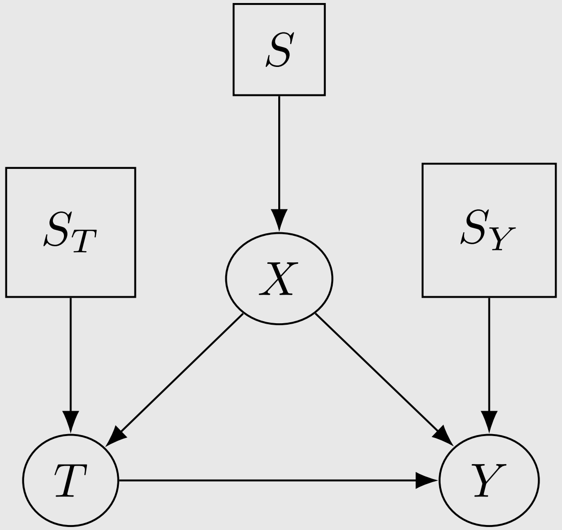
Mechanism that differs between Π and Π^*



Selection Diagrams

Allow for different causal mechanisms across the two distributions

Mechanism that differs between Π and Π^*

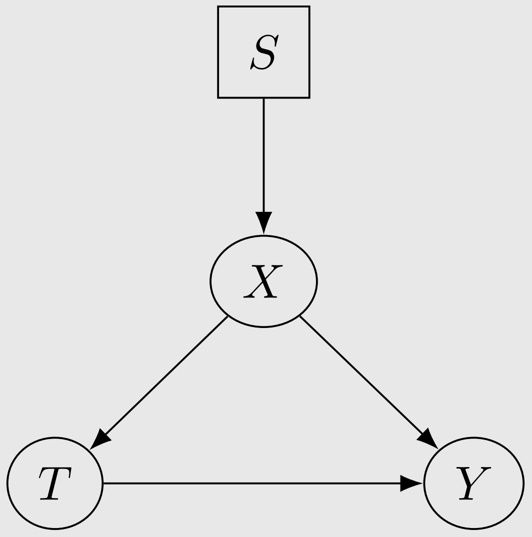


Selection Diagrams

Allow for different causal mechanisms across the two distributions

Mechanism that differs between Π and Π^*

Absence of S encodes invariance

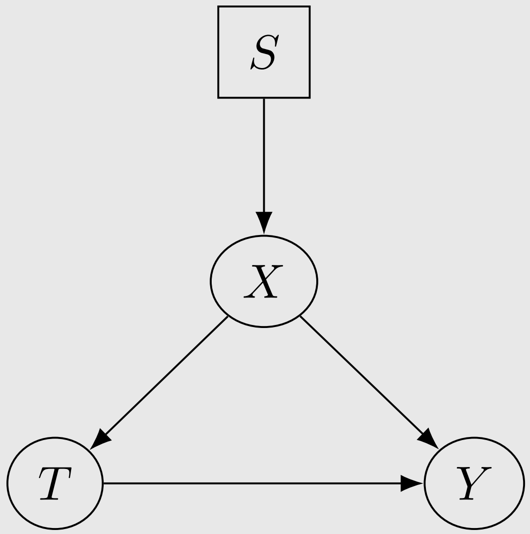


Selection Diagrams

Allow for different causal mechanisms across the two distributions

Mechanism that differs between Π and Π^*

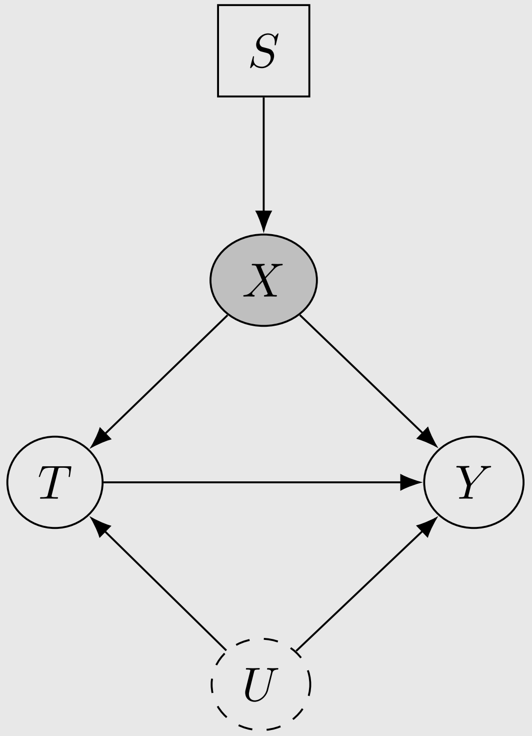
Absence of S encodes invariance



$$P^*(y \mid do(t), x) \triangleq P(y \mid do(t), x, s^*)$$

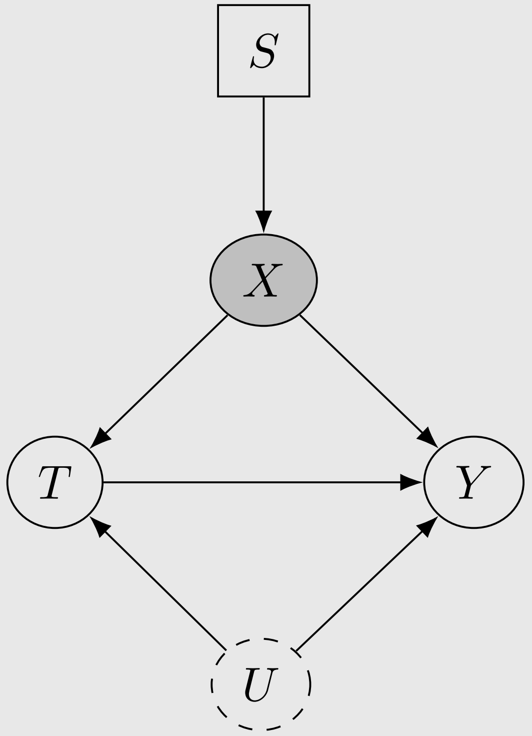
Direct Transportability (External Validity)

$$P(y \mid do(t), x) \stackrel{?}{=} P^*(y \mid do(t), x)$$



Direct Transportability (External Validity)

$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$
$$\text{if } Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, X$$

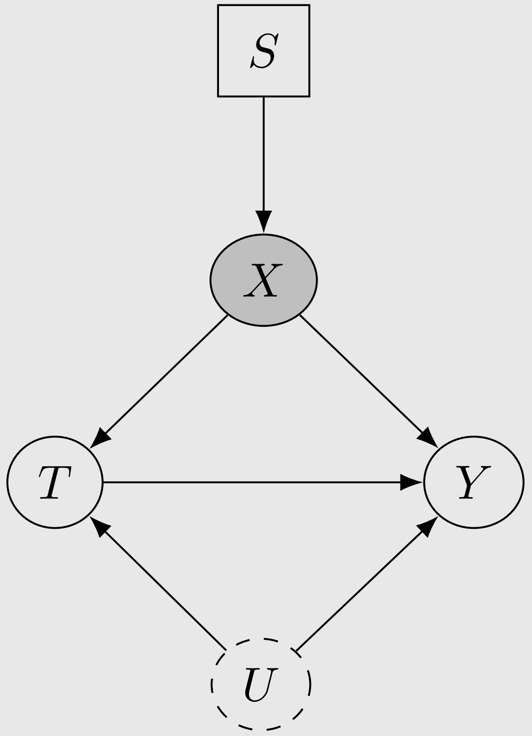


Direct Transportability (External Validity)

$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$

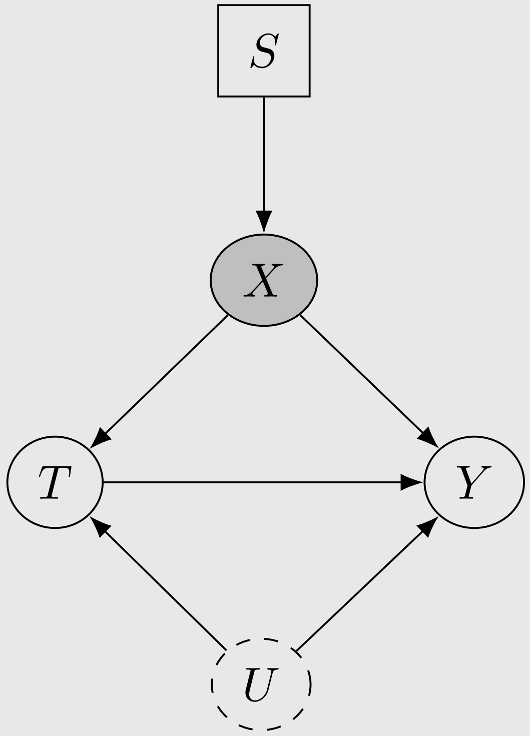
$$\text{if } Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, X$$

Proof:



Direct Transportability (External Validity)

$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$
$$\text{if } Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, X$$



Proof:

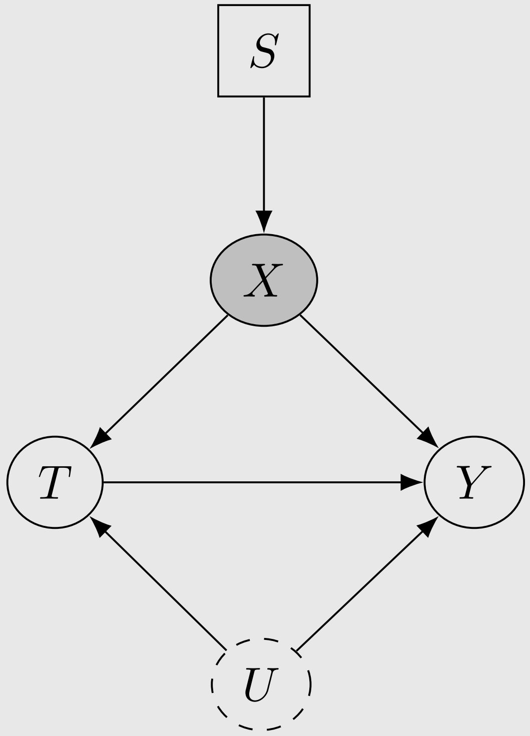
$$P^*(y \mid do(t), x) = P(y \mid do(t), x, s^*)$$

Direct Transportability (External Validity)

$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$
$$\text{if } Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, X$$

Proof:

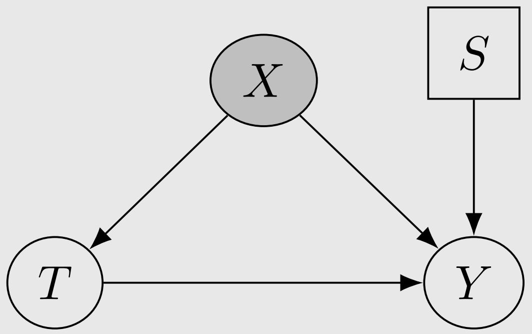
$$P^*(y \mid do(t), x) = P(y \mid do(t), x, s^*)$$
$$= P(y \mid do(t), x)$$



Direct Transportability (External Validity)

$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$

$$\text{if } Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, X$$



Proof:

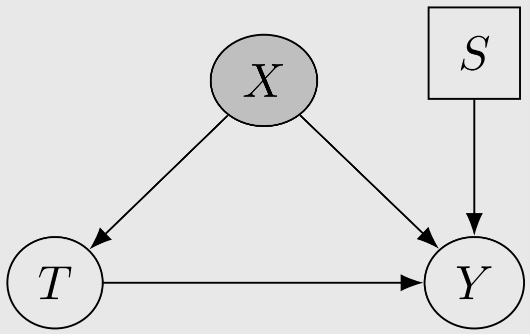
$$\begin{aligned} P^*(y \mid do(t), x) &= P(y \mid do(t), x, s^*) \\ &= P(y \mid do(t), x) \end{aligned}$$

Direct Transportability (External Validity)

$$P(y \mid do(t), x) = P^*(y \mid do(t), x)$$

if $Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, X$

$$P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$$

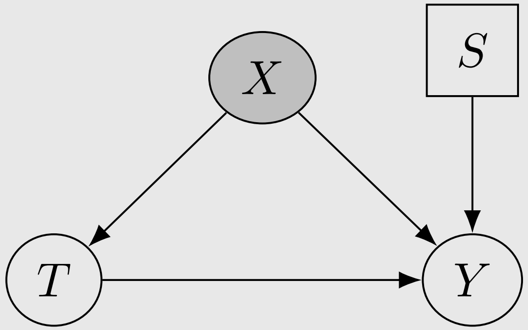


Proof:

$$P^*(y \mid do(t), x) = P(y \mid do(t), x, s^*)$$
$$= P(y \mid do(t), x)$$

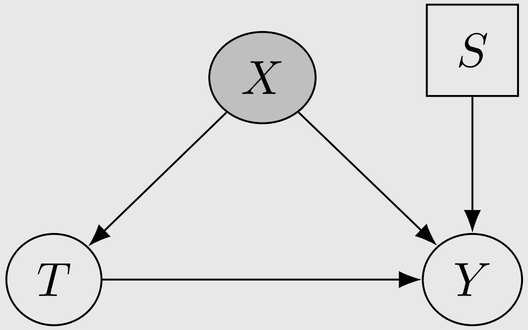
Trivial Transportability

- Don't have direct transportability: $P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$



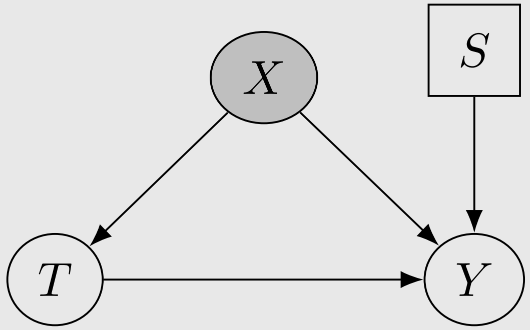
Trivial Transportability

- Don't have direct transportability: $P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$
- Have access to observational data from target population: $P^*(y, t, x)$



Trivial Transportability

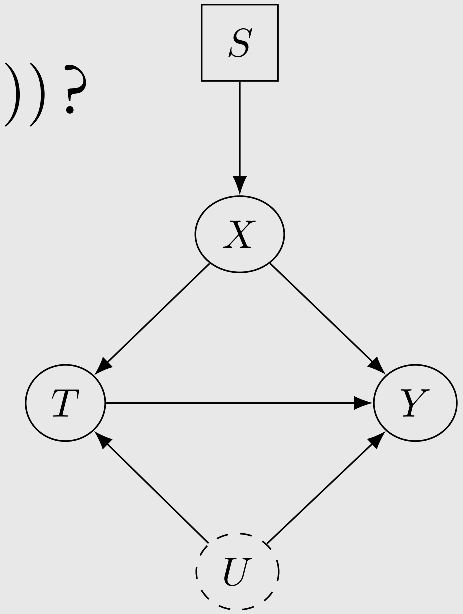
- Don't have direct transportability: $P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$
- Have access to observational data from target population: $P^*(y, t, x)$
- Can identify estimand using only target data: $P^*(y \mid do(t), x) = P^*(y \mid t, x)$

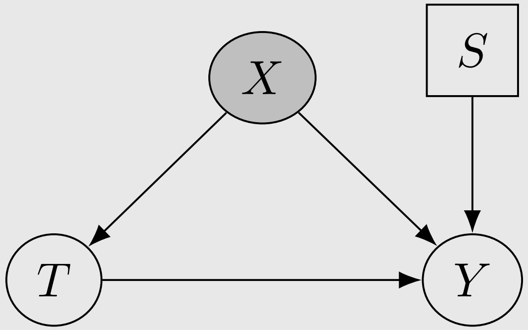


Trivial Transportability

- Don't have direct transportability: $P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$
- Have access to observational data from target population: $P^*(y, t, x)$
- Can identify estimand using only target data: $P^*(y \mid do(t), x) = P^*(y \mid t, x)$

Identify $P^*(y \mid do(t))$?



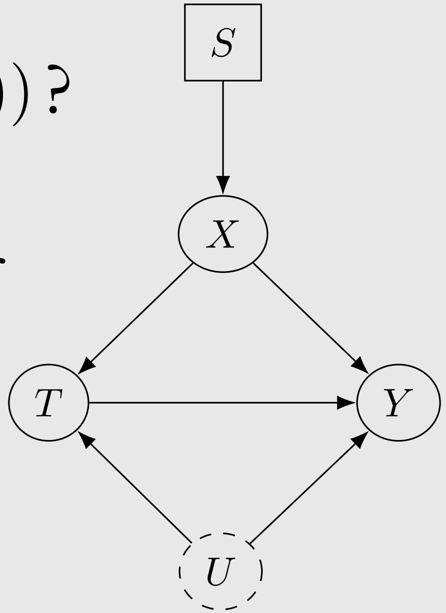


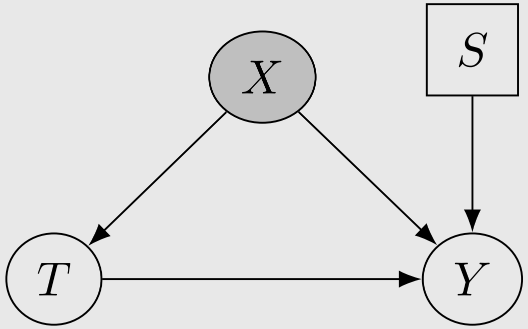
Trivial Transportability

- Don't have direct transportability: $P(y \mid do(t), x) \neq P^*(y \mid do(t), x)$
- Have access to observational data from target population: $P^*(y, t, x)$
- Can identify estimand using only target data: $P^*(y \mid do(t), x) = P^*(y \mid t, x)$

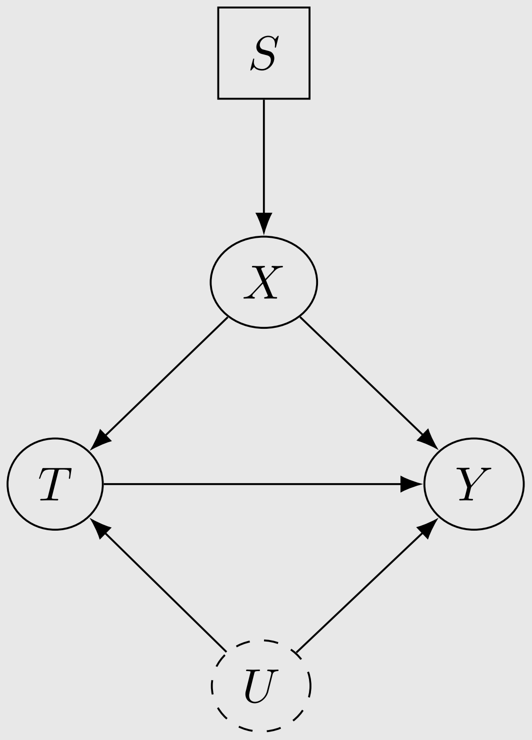
Identify $P^*(y \mid do(t))$?

Combine aspects of trivial and direct transportability



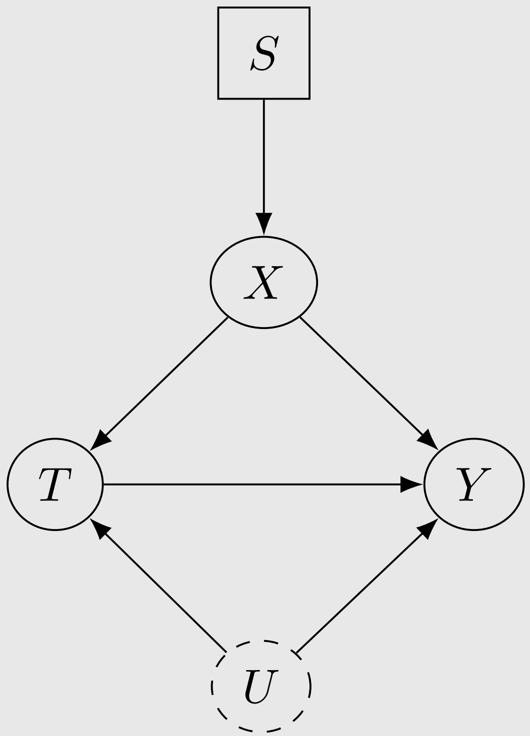


S-Admissibility and Transport Formula



S-Admissibility and Transport Formula

S-Admissibility: A set of variables W is S-admissible if $Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, W$

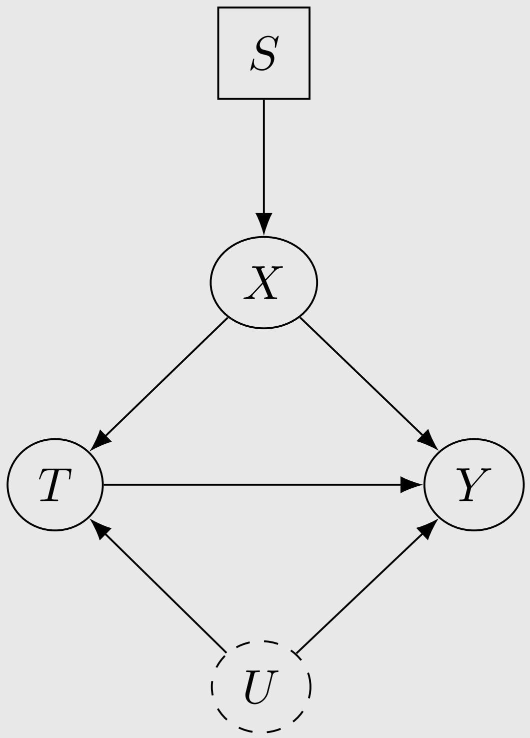


S-Admissibility and Transport Formula

S-Admissibility: A set of variables W is S-admissible if $Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, W$

Transport Result: If W is S-admissible, then

$$P^*(y \mid do(t)) = \sum_w P(y \mid do(t), w) \, P^*(w)$$



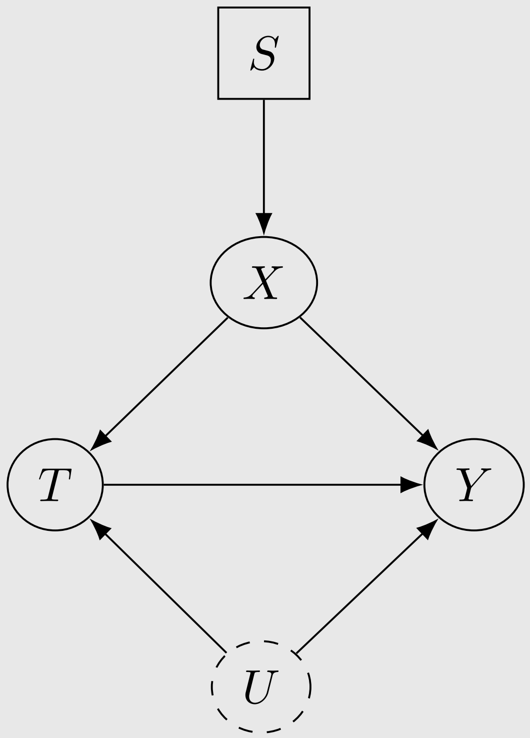
S-Admissibility and Transport Formula

S-Admissibility: A set of variables W is S-admissible if $Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, W$

Transport Result: If W is S-admissible, then

$$P^*(y \mid do(t)) = \sum_w P(y \mid do(t), w)\, P^*(w)$$

Note: Another word for "sufficient adjustment set" from week 4 is "admissible set."

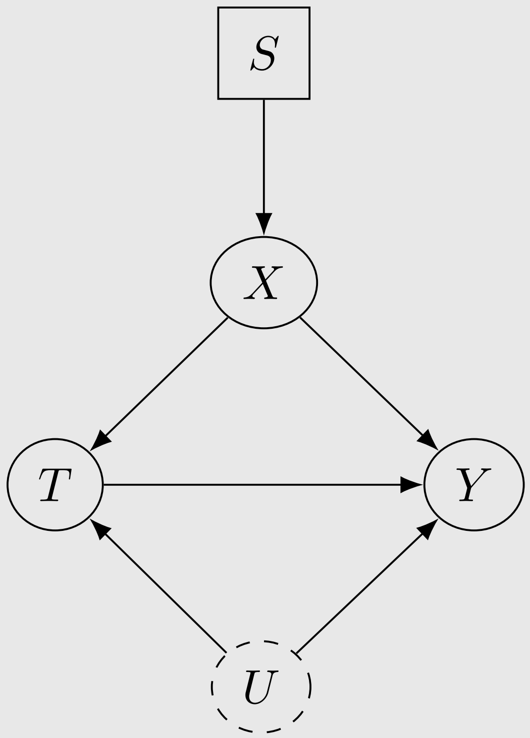


S-Admissibility and Transport Formula

S-Admissibility: A set of variables W is S-admissible if $Y \perp\!\!\!\perp_{G_{\overline{T}}} S \mid T, W$

Transport Result: If W is S-admissible, then

$$P^*(y \mid do(t)) = \sum_w P(y \mid do(t), w)\, P^*(w)$$



Note: Another word for "sufficient adjustment set" from week 4 is "admissible set."

Main Paper: Pearl & Bareinboim (2014)

Questions:

1. Describe direct transportability in your own words.
2. Describe trivial transportability in your own words.
3. Prove the transport result on the previous slide.