Estimating Identifiable Causal Effects and its Application toward Interpretable ML/AI

Yonghan Jung

Purdue University

yonghanjung.me

Korea Summer Workshop
on Causal Inference 2022

Outline of the talk

In this talk, I will explain how to estimate any identifiable causal effects.

1. [*Estimating Identifiable Causal Effects through Double Machine Learning*] Y.Jung, J. Tian, E. Bareinboim. **AAAI-21.**

2. [*Estimating Identifiable Causal Effects on Markov Equivalence Class through Double Machine Learning*] Y.Jung, J. Tian, E. Bareinboim. **ICML-21**.

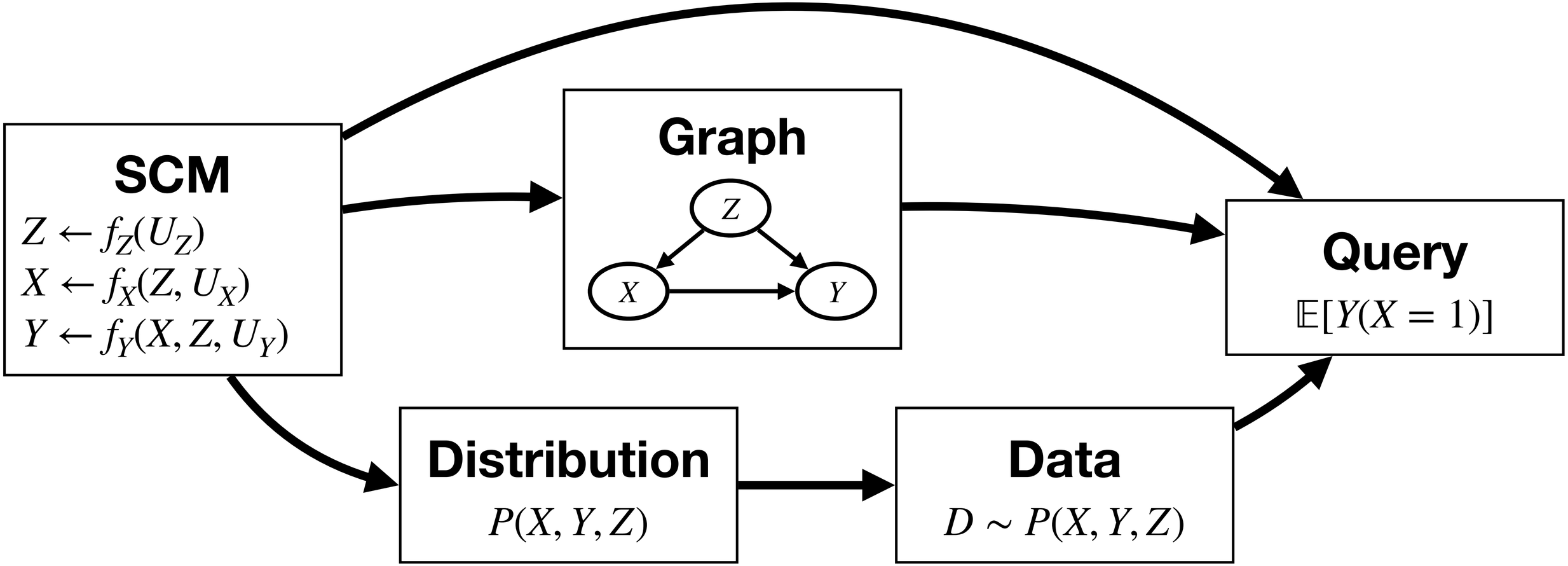
Then, I will give an example of how our task (“Estimating identifiable causal effects”) is applied in the trustworthy-AI domain.

3. [*On Measuring Causal Contribution via do-intervention]* Y. Jung, S. Kasiviswanathan, J. Tian, D. Janzing, P. Blöbaum, E. Bareinboim. **ICML-22**

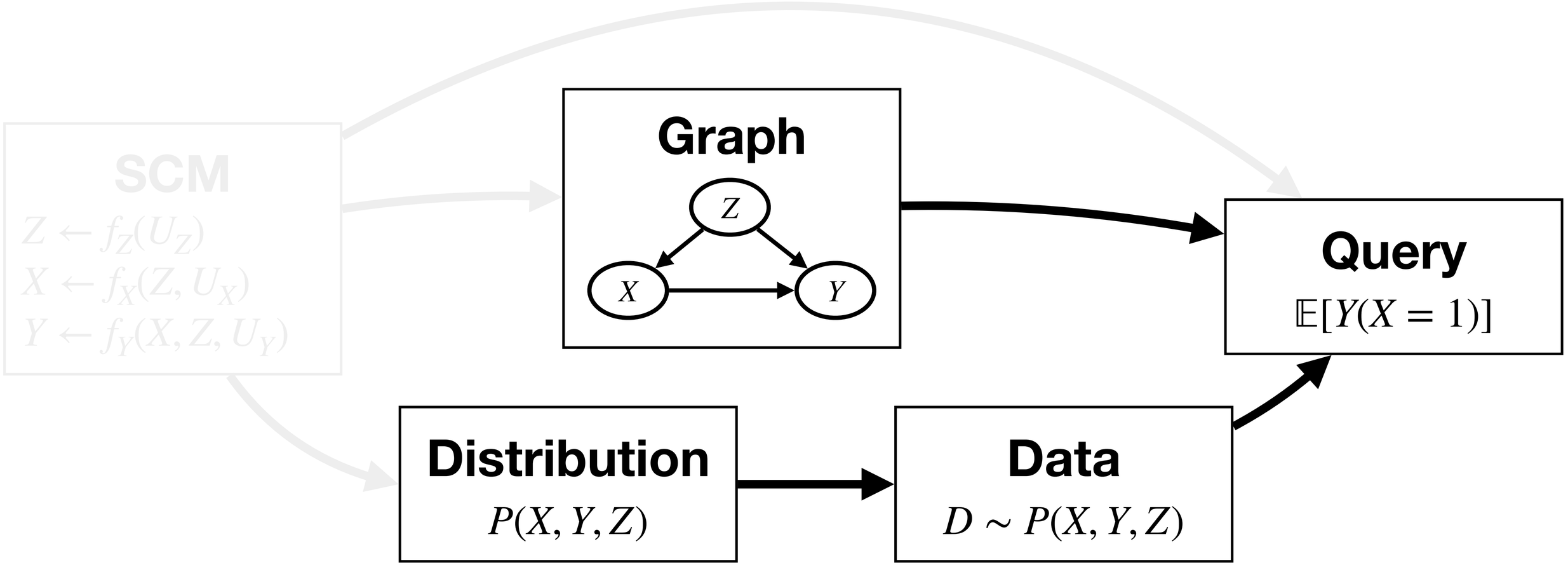
Two Tasks in Causal Inference

Causal Effect Identification and Estimation

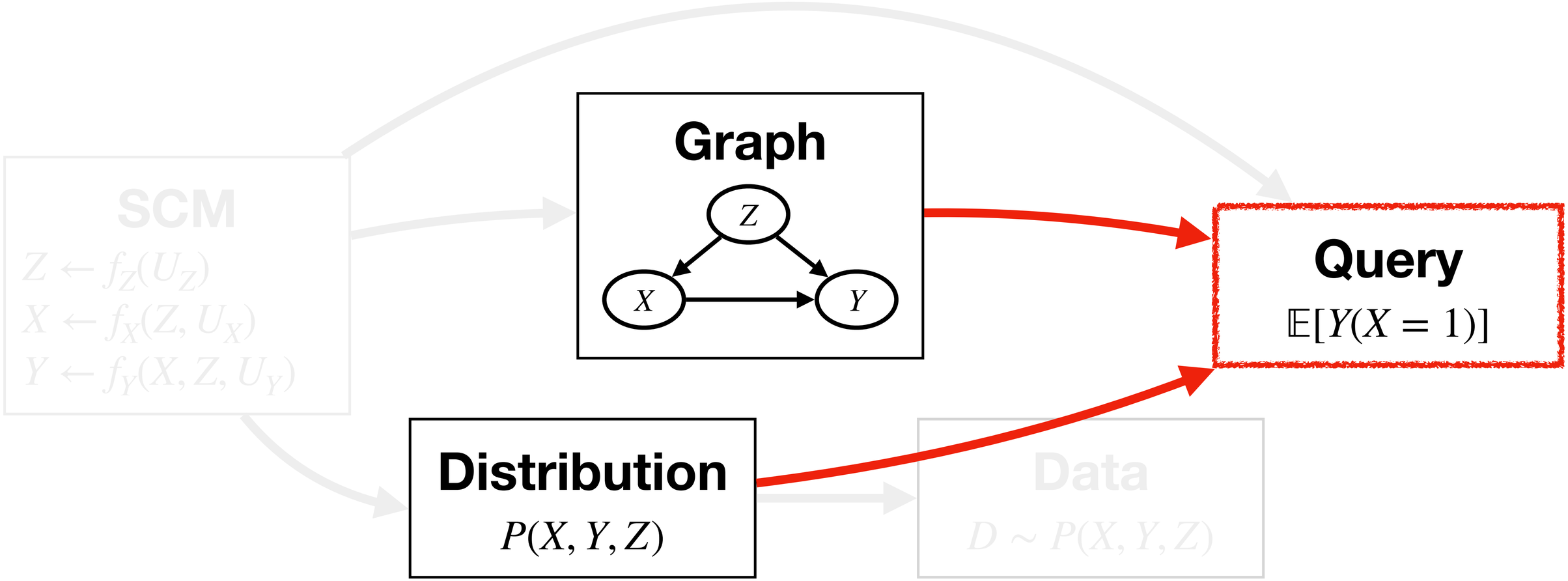
Big Picture for Causal Inference



Big Picture for Causal Inference

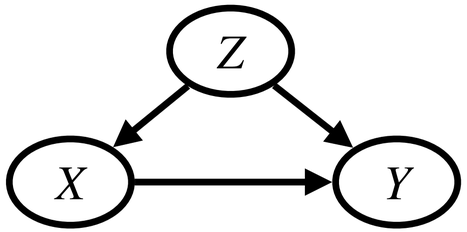


Task 1. Causal Effect Identification



Task 1. Causal Effect Identification

Graph (G)



Distribution (P)

(e.g., $P(Z, X, Y)$)

Query (Q)

(e.g., $\mathbb{E}[Y(x)]$)

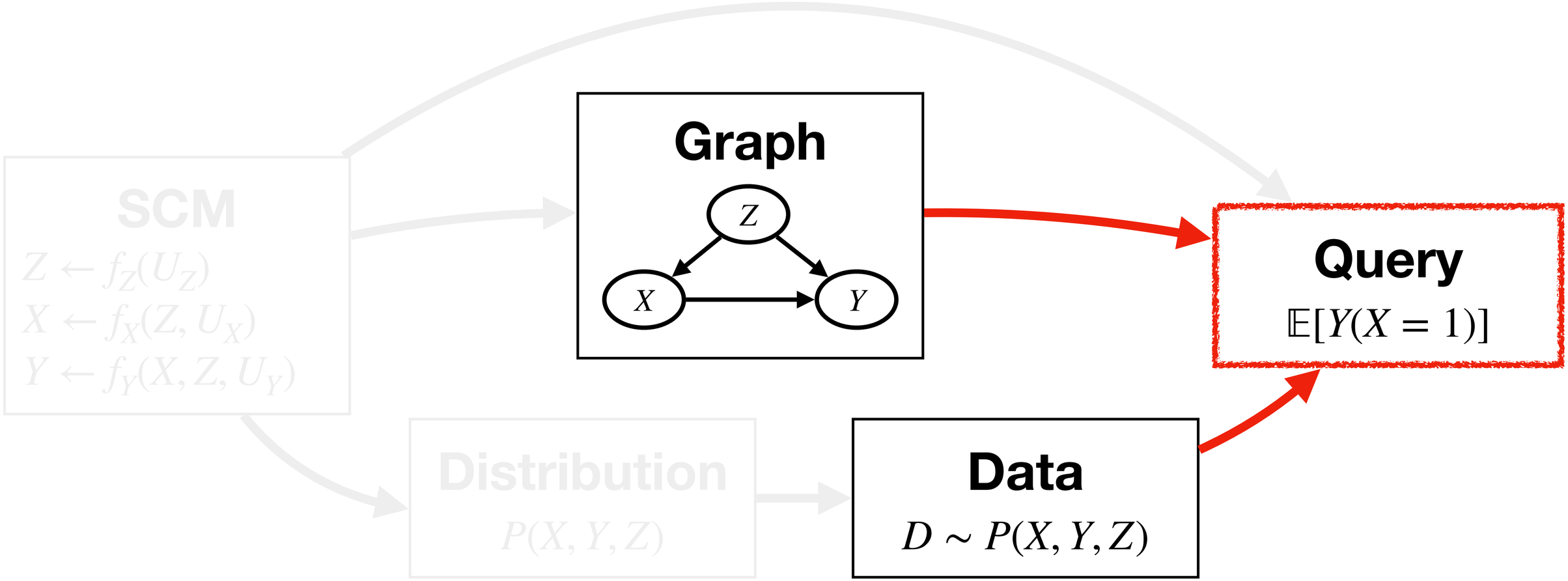
Causal Effect Identification (ID)

Solution

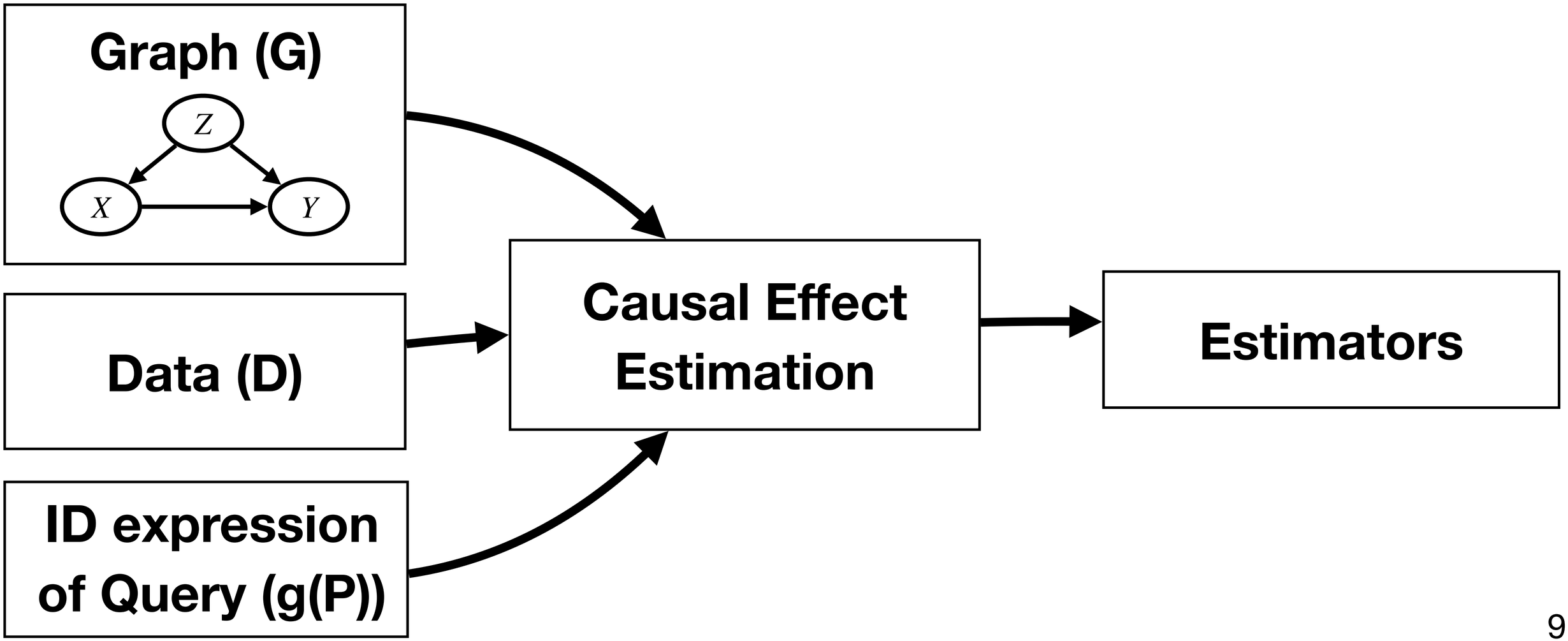
YES: A function $g(P)$ s.t. $g(P) = Q$.

No: If no such functions exist.

Task 2. Causal Effect Estimation: Big Picture



Task 2. Causal Effect Estimation

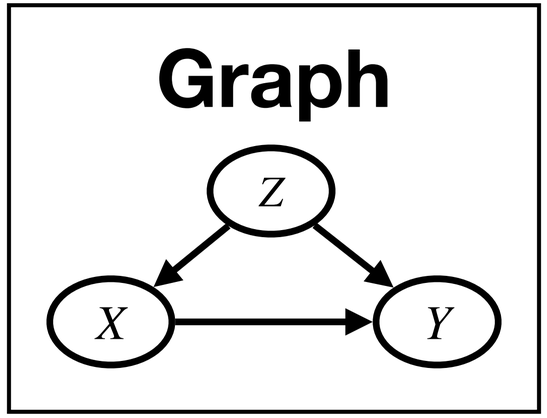


Causal Effect Estimation: Current Status

For a small portion of the identification, the causal effect estimation problem has been solved.

↳ Causal effect estimators have been developed only for the “Back-door adjustment” (also known as “ignorability”, “no unmeasured confounders” assumption).

Causal Effect Estimation: Back-door Adjustment



Back-door Adjustment

↳ If there exists Z s.t. (1) Z is non-descendent of $\{X, Y\}$ and (2) $(Y \perp\!\!\!\perp X \,|\, Z)_{\underline{X}}$, then

$$\mathbb{E}[Y(x)] = \sum_z \mathbb{E}[Y \,|\, x, z] P(z).$$

Causal Effect Estimation: Example of a BD estimator

$$\mathbb{E}[Y(x)] = g(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z).$$

Double/Debiased Machine Learning Estimator T for $\widehat{g(P)}$.

$$T := \mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \left(Y - \hat{\mathbb{E}}[Y \mid X, Z] \right) + \hat{\mathbb{E}}[Y \mid x, Z] \right]$$

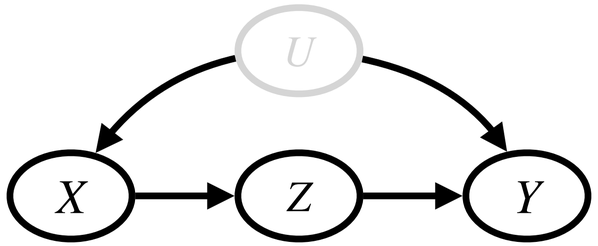
↳ T converges fast even when either $\hat{\mathbb{E}}[Y \mid X, Z]$ or $\hat{P}(X \mid Z)$ is correct; or $\hat{\mathbb{E}}[Y \mid X, Z]$ and $\hat{P}(X \mid Z)$ converge slowly.

Causal Effect Estimation: Limitation

When the identification expression is beyond the back-door adjustment, virtually no estimators are available.

↳ For example, when the graph satisfies the “front-door criteria”, the expression is not a BD adjustment, so the existing estimators are not applicable.

Front-door



$$\mathbb{E}[Y(x)] = \sum_z P(z \mid x) \sum_{x'} \mathbb{E}[Y \mid x', z] P(x')$$

Causal Effect Estimation: Our Tasks

My research interest centers around developing estimators for any identifiable causal effects.

[Estimating Identifiable Causal Effects through Double Machine Learning] Y.Jung, J. Tian, E. Bareinboim. AAAI-21.

[Estimating Identifiable Causal Effects on Markov Equivalence Class through Double Machine Learning] Y.Jung, J. Tian, E. Bareinboim. ICML-21.

DML Estimator for Any Identifiable Causal Effects

Y.Jung, J. Tian, E. Bareinboim. **AAAI-21.**

Back-door adjustment and Double/Debiased ML (DML)

Goal of This Section

We will understand the mechanism of the DML estimator by constructing the estimator for

$$g(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z).$$

We assume the followings in the lecture.

- $X \in \{0,1\}$ a binary treatment variable.
- $P(\mathbf{v}) > 0$ for any $\mathbf{v}$.
- Y is 1-dimensional variable (continuous/discrete); Z can be multivariate (continuous/discrete)

Amenable Expression of the BD Adjustment

$$g(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z).$$

$$= \mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)} \left\{ Y - \mathbb{E}[Y \mid X, Z] \right\} + \mathbb{E}[Y \mid x, Z] \right]$$

Amenable Expression of the BD Adjustment: Proof (1)

$$\sum_z \mathbb{E}[Y \mid x, z] P(z) = \mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)} \left\{Y - \mathbb{E}[Y \mid X, Z]\right\} + \mathbb{E}[Y \mid x, Z]\right]$$

This holds because (1)

$$\mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)} \left\{Y - \mathbb{E}[Y \mid X, Z]\right\}\right] = \mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)} \left\{\mathbb{E}[Y \mid X, Z] - \mathbb{E}[Y \mid X, Z]\right\}\right] = 0$$

Amenable Expression of the BD Adjustment: Proof (2)

$$\sum_z \mathbb{E}[Y \mid x, z] P(z) = \mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)} \left\{Y - \mathbb{E}[Y \mid X, Z]\right\} + \mathbb{E}[Y \mid x, Z]\right]$$

This holds because (2)

$$\mathbb{E}\left[\mathbb{E}[Y \mid x, Z]\right] = \sum_z \mathbb{E}[Y \mid x, z] P(z).$$

Double/Debiased Machine Learning (DML) Estimator for the BD Adjustment

$$\sum_z \mathbb{E}[Y|x,z]P(z) = \mathbb{E}\left[\frac{I_x(X)}{P(X|Z)}\left\{Y - \mathbb{E}[Y|X,Z]\right\} + \mathbb{E}[Y|x,Z]\right]$$

$$T := \mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X|Z)}\left\{Y - \hat{\mathbb{E}}[Y|X,Z]\right\} + \hat{\mathbb{E}}[Y|x,Z]\right]$$

Robustness Property of the DML Estimator

$$T := \mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X|Z)} \left\{ Y - \hat{\mathbb{E}}[Y|X,Z] \right\} + \hat{\mathbb{E}}[Y|x,Z] \right]$$

error

$$:= \mathbb{E}[T] - \mathbb{E} \left[\frac{I_x(X)}{P(X|Z)} \left\{ Y - \mathbb{E}[Y|X,Z] \right\} + \mathbb{E}[Y|x,Z] \right]$$

$$= O_P \left(\| \, \mathbb{E}[Y|X,Z] - \hat{\mathbb{E}}[Y|X,Z] \, \| \| \, P(X|Z) - \hat{P}(X|Z) \, \| \right)$$

where $O_P(\| \, f(\mathbf{X}) \, \|) := \sqrt{\mathbb{E}[f^2(\mathbf{X})]}$.

Properties of DML Estimators: Derivation

$$\mathbb{E}[T] - \mathbb{E}\left[\frac{I_x(X)}{P(X|Z)}\left\{Y - \mathbb{E}[Y|X,Z]\right\} + \mathbb{E}[Y|x,Z]\right]$$

$$= \mathbb{E}\left[\frac{I_x(X)}{\hat{P}(X|Z)}\left\{Y - \hat{\mathbb{E}}[Y|X,Z]\right\} + \hat{\mathbb{E}}[Y|x,Z] - \mathbb{E}[Y|x,Z]\right]$$

$$= \mathbb{E}\left[\frac{1}{\hat{P}(X|Z)}\left\{P(X|Z) - \hat{P}(X|Z)\right\}\left\{\hat{\mathbb{E}}[Y|X,Z] - \mathbb{E}[Y|X,Z]\right\}\right]$$

$$= O_P\left(\| \mathbb{E}[Y|X,Z] - \hat{\mathbb{E}}[Y|X,Z] \| \| P(X|Z) - \hat{P}(X|Z) \|\right)$$

Properties of DML Estimators

$$\text{error} = O_P\left(\| \mathbb{E}[Y|X,Z] - \hat{\mathbb{E}}[Y|X,Z] \| \| P(X|Z) - \hat{P}(X|Z) \| \right)$$

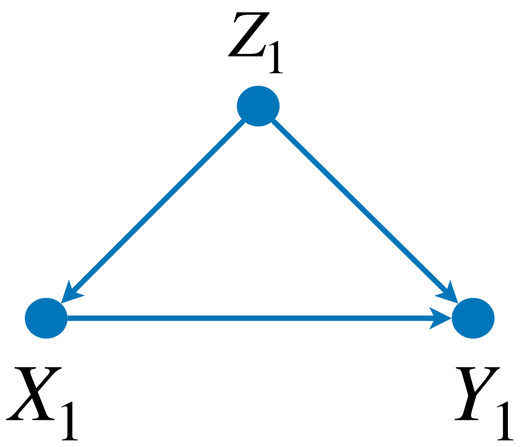
Doubly Robustness: The error is 0 (i.e., T is unbiased) if $\hat{\mathbb{E}}[Y|X,Z] = \mathbb{E}[Y|X,Z]$; or $\hat{P}(X|Z) = P(X|Z)$.

Debiasedness: The error converges at $N^{-1/2}$ rate if $\hat{\mathbb{E}}[Y|X,Z]$ and $\hat{P}(X|Z)$ converges to $\mathbb{E}[Y|X,Z]$ and $P(X|Z)$ at slower $N^{-1/4}$ rate.

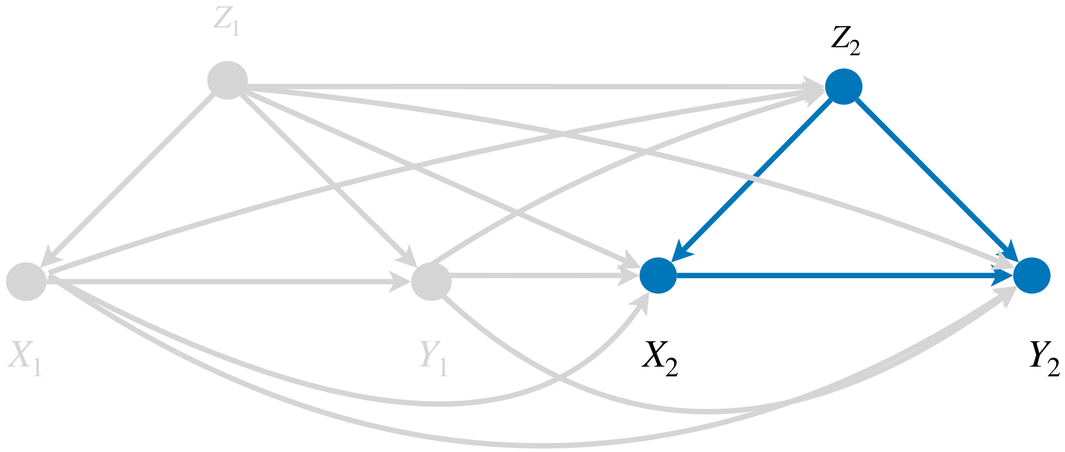
Multi-outcome Sequential Back-door adjustment

Multi-outcome sequential BD (mSBD)

Multi-outcome sequential BD (mSBD): An extension of BD, where, at ith round, Z_i satisfies the BD criterion relative to $\{X_i, Y_i\}$ conditioned on previous variables $\{(X_j, Y_j, Z_j)\}_{j=1}^{i-1}$.



$i = 1$



$i = 2$

Multi-outcome sequential BD (mSBD)

mSBD adjustment: If $\mathbf{Z} = \{Z_1, \cdots, Z_n\}$ satisfies the mSBD criterion relative to $(\mathbf{X}, \mathbf{Y})$,

$$g(P) := P(\mathbf{y} \mid do(\mathbf{x}))$$

$$= \sum_{\mathbf{z}} \prod_{Y_i \in \mathbf{Y}} P(y_i \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i-1)}) \prod_{Z_i \in \mathbf{Z}} P(z_i \mid \mathbf{x}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{y}^{(i-1)})$$

$$\mathbf{X}^{(i)} := \{X_1, \cdots, X_i\}$$

Amenable Representation for the mSBD Adjustment (1)

$$g(P) = \mathbb{E}\left[H_0^1(x_1) + \sum_{k=1}^{n} W_0^k \{ H_0^{k+1}(x_{k+1}) - H_0^k(X_k) \} \right]$$

$H_0^{n+1}(x_{n+1}) := I_{\mathbf{y}}(\mathbf{Y})$ and for all $k = n, n-1, \cdots, 1,$

$$H_0^k(X_k) = \mathbb{E}[H_0^{k+1}(x_{k+1}) \,|\, X_k, \mathbf{X}^{(k-1)}, \mathbf{Y}^{(k-1)}, \mathbf{Z}^{(k)}]$$

$$H_0^k(x_k) = \mathbb{E}[H_0^{k+1}(x_{k+1}) \,|\, x_k, \mathbf{X}^{(k-1)}, \mathbf{Y}^{(k-1)}, \mathbf{Z}^{(k)}]$$

Amenable Representation for the mSBD Adjustment (2)

$$g(P) = \mathbb{E}\left[H_0^1(x_1) + \sum_{k=1}^{n} W_0^k \{ H_0^{k+1}(x_{k+1}) - H_0^k(X_k) \} \right]$$

$$W_0^k = \prod_{p=1}^{k} \frac{I_{x_p}(X_p)}{\pi_0^p(\mathbf{X}^{(p-1)}, \mathbf{Z}^{(p)}, \mathbf{Y}^{(p-1)})}, \text{ where}$$

$$\pi_0^p(\mathbf{X}^{(p-1)}, \mathbf{Z}^{(p)}, \mathbf{Y}^{(p-1)}) := P(X_p \,|\, \mathbf{X}^{(p-1)}, \mathbf{Z}^{(p)}, \mathbf{Y}^{(p-1)})$$

DML Estimator for the mSBD adjustment

$$g(P) = \mathbb{E}\left[H_0^1(x_1) + \sum_{k=1}^{n} W_0^k\{H_0^{k+1}(x_{k+1}) - H_0^k(X_k)\}\right]$$

$$T := \mathbb{E}_D\left[\hat{H}_0^1(x_1) + \sum_{k=1}^{n} \hat{W}_0^k\{\hat{H}_0^{k+1}(x_{k+1}) - \hat{H}_0^k(X_k)\}\right]$$

Robustness Property of the DML Estimator

error

$$:= \mathbb{E}[T] - \mathbb{E}\left[H_0^1(x_1) + \sum_{k=1}^{n} W_0^k \{ H_0^{k+1}(x_{k+1}) - H_0^k(X_k) \} \right]$$

$$= O_P\left(\sum_{i=1}^{n} \| \hat{H}^i - H_0^i \| \| \hat{\pi}^i - \pi_0^i \| \right)$$

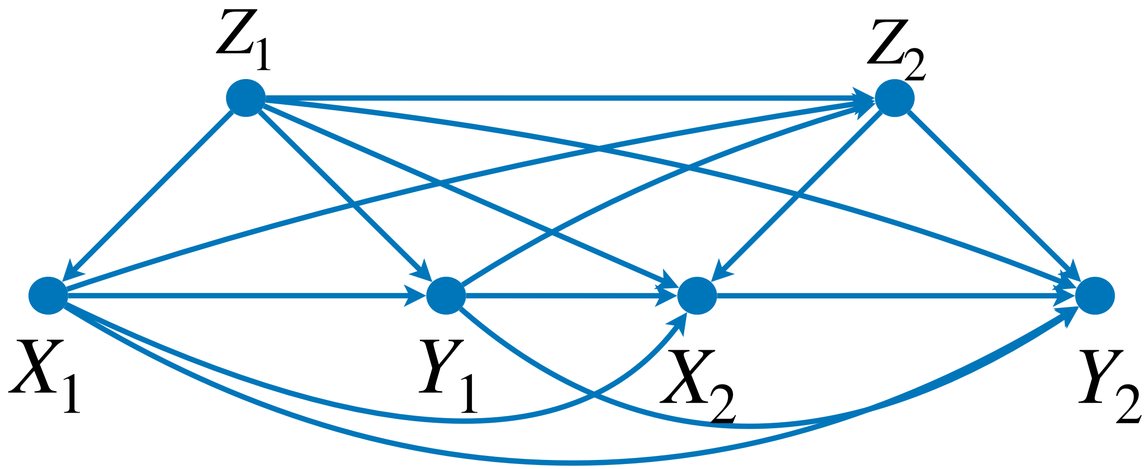
Robustness Property of the DML Estimator

$$\text{error} = O_P\left(\sum_{i=1}^{n} \| \hat{H}^i - H_0^i \| \| \hat{\pi}^i - \pi_0^i \|\right)$$

Doubly Robustness: The error is 0 (i.e., T is unbiased) if $\hat{H}^i = H_0^i$; or $\hat{\pi}^i = \pi_0^i$ for all $i = 1,2,\cdots,n$.

Debiasedness: The error converges at $N^{-1/2}$ rate if $\hat{H}^i$ and $\hat{\pi}^i$ converges to H_0^i and π_0^i at slower $N^{-1/4}$ rate.

DR Estimand for mSBD Example 1

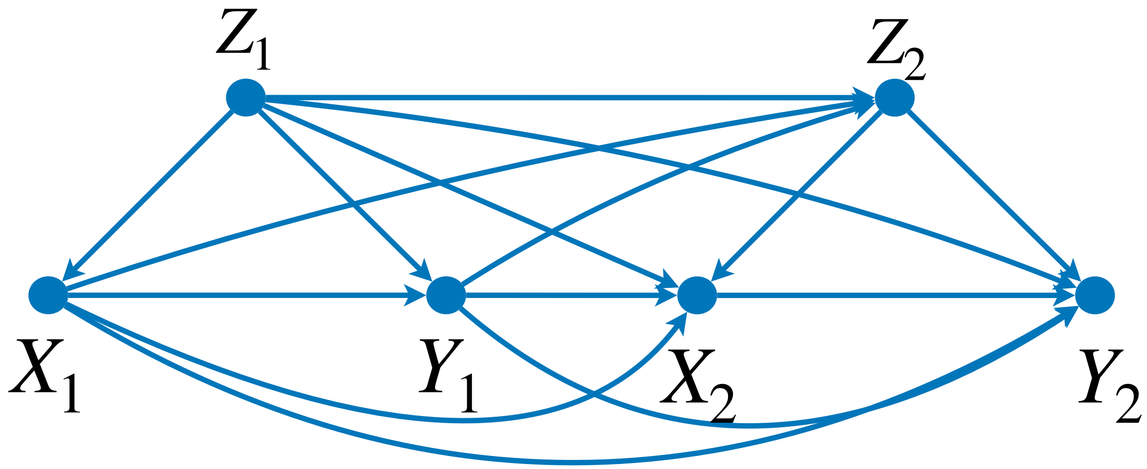


$$g(P) = \mathbb{E}\left[H_0^1(x_1) + W_0^1(H_0^2(x_2) - H_0^1(X_1)) + W_0^2(I_{\mathbf{y}}(\mathbf{Y}) - H_0^2(X_2))\right]$$

$$H_0^2(X_2) = \mathbb{E}[I_{\mathbf{y}}(\mathbf{Y}) \,|\, \mathbf{X}^{(2)}, Y_1, \mathbf{Z}^{(2)}] \qquad H_0^1(X_1) = \mathbb{E}[H_0^2(x_2) \,|\, X_1, Z_1]$$

$$H_0^2(x_2) = \mathbb{E}[I_{\mathbf{y}}(\mathbf{Y}) \,|\, x_2, X_1, Y_1, \mathbf{Z}^{(2)}] \qquad H_0^1(x_1) = \mathbb{E}[H_0^2(x_2) \,|\, x_1, Z_1]$$

DR Estimand for mSBD Example 1



$$W_0^2 = \frac{I_{x_1}(X_1)}{\pi_0^1(Z_1)} \frac{I_{x_2}(X_2)}{\pi_0^2(X_1, \mathbf{Z}^{(2)}, Y_1)}$$

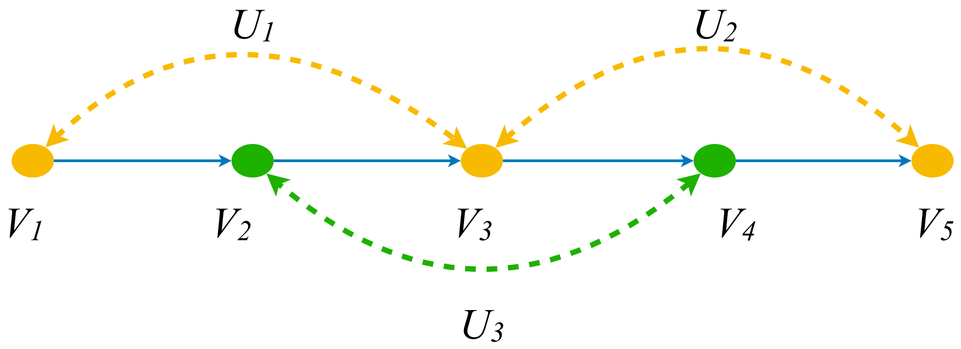
$$W_0^1 = \frac{I_{x_1}(X_1)}{\pi_0^1(Z_1)}$$

$$\pi_0^1(Z_1) = P(X_1 \,|\, Z_1)$$

$$\pi_0^2(X_1, \mathbf{Z}^{(2)}, Y_1) = P(X_2 \,|\, X_1, \mathbf{Z}^{(2)}, Y_1)$$

Recap: the Causal ID algorithm

C-component and C-factors



- **C-component**: A set of variables connected by bi-directed edges (e.g., $\{V_1, V_3, V_5\}$ and $\{V_2, V_4\}$.
- **C-factor**: $Q[\mathbf{C}] := P(\mathbf{c} \mid do(\mathbf{v} \backslash \mathbf{c}))$

- A distribution can be factorized w.r.t. C-factors.

$$P(\mathbf{v}) = Q[V_2, V_4]Q[V_1, V_3, V_5]$$

C-factor Algebra - Summary

We have three basic operations over c-factors

1. Identification of c-factor:

$$Q[\mathbf{C}] = \prod_{V_i \in \mathbf{C}} P(v_i \mid v^{(i-1)})$$ where $\mathbf{C}$ is a C-component in G

$$Q[\mathbf{W}] = \sum_{\mathbf{c} \setminus \mathbf{w}} Q[\mathbf{C}]$$ If $\boldsymbol{W}$ is ancestral in $G(\mathbf{C})$

3. Factorize into c-components

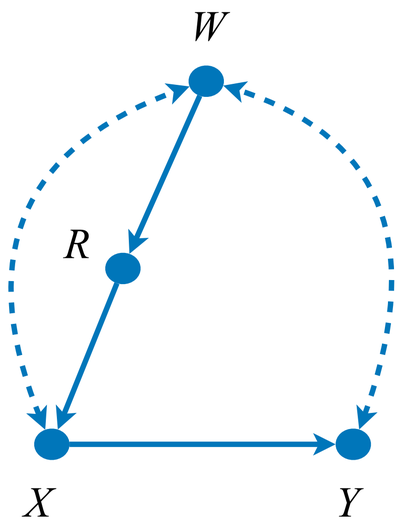
$$Q[\mathbf{H}] = \prod_{j} Q[H_j]$$ Where $\mathbf{H}_1, \ldots, \mathbf{H}_k$, are the c-components in $\mathcal{G}[\mathbf{H}]$

Recap: Complete ID algorithm

ID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the C-components of G.
2. $Q[\mathbf{S}_i] = \prod_{V_k \in \mathbf{S}_i} P(v_k \mid v^{(k-1)})$ by the 1st algebra.
3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be C-components of $G(\mathbf{D})$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4. Identify $Q[\mathbf{D}_j]$ from $Q[\mathbf{S}]$ by recursively applying 2nd and 3rd C-factor algebra
5. $P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_j Q[\mathbf{D}_j]$ if all $Q[\mathbf{D_j}]$ is defined, FAIL otherwise.

Example of ID: Napkin



$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}.$

- $Q[\mathbf{S}_1] \equiv P(w, x, y \,|\, do(r)) = P(w)P(x, y \,|\, r, w)$
- $\{W\}$ is an ancestral set in $G(\mathbf{S}_1)$,

$$Q[X, Y] = \sum_w Q[\mathbf{S}_1] = \sum_w P(w)P(x, y \,|\, r, w)$$

- $\{Y\}$ is a descendent set in $G(\{X, Y\})$.

$$Q[Y] = \sum_x Q[X, Y] = \sum_w P(w)P(x \,|\, r, w)$$

- $Q[Y] = \dfrac{Q[X, Y]}{Q[X]} = \dfrac{\sum_w P(w)P(x, y \,|\, r, w)}{\sum_w P(w)P(x \,|\, r, w)}$

DML estimation for ID functional

Expressing C-factor as mSBD

$Q[\mathbf{S}_i]$, where $\mathbf{S}_i$ is the i'th C-component in G, can be expressed as a mSBD adjustment.

Expressing C-factor as mSBD adjustment

Let $\mathbf{C}$ be a C-component in G, $\mathbf{W}$ denote the ancestral set of $\mathbf{C}$ (i.e., $\mathbf{W} = An(\mathbf{W})_{G(\mathbf{C})}$) and $\mathbf{R} \equiv Pa(\mathbf{W})$. Then, $\mathbf{Z} = (\mathbf{C} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W})$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$, and

$$Q[\mathbf{W}] = M[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$$

mSBD adjustment where $\mathbf{Z}$ is mSBD admissible w.r.t. $(\mathbf{R}, \mathbf{W})$

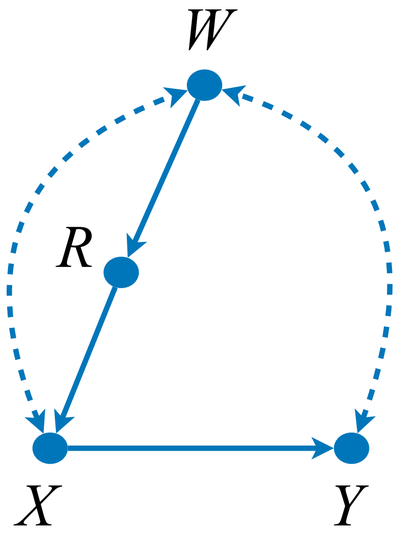
ID algorithm

DML-ID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the c-components of G.
2. Let $Q[\mathbf{S}_i] = M[\mathbf{s}_i \mid Pa(\mathbf{s}_i); \varnothing]$.
3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be the c-components of $G(\mathbf{D})$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4. $Q[\mathbf{D}_j] = A^j(\{M^j_\ell\}) = \text{Identify}(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$.
5. $P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_j A^j(\{M^j_\ell\})$ if all $A^j(\{M^j_\ell\})$ have been defined, FAIL otherwise.

$Q[\mathbf{D}_j]$ is expressed as an algebraic operation (A^j) of mSBD adjustments.

Example of DML-ID: Napkin



$$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}.$$

- $Q[\mathbf{S}_1] \equiv P_r(w, x, y) = M[\mathbf{s}_1 \,|\, r; \varnothing]$
- $\{W\}$ is an ancestral set in $G(\mathbf{S}_1)$,

$$Q[X, Y] = \sum_w M[\mathbf{s}_1 \,|\, r; \varnothing] = M[(x, y) \,|\, r; w]$$

- $\{Y\}$ is an descendent set in $G(\{X, Y\})$.

$$Q[X] = \sum_y M[(x, y) \,|\, r; w] = M[x \,|\, r; w]$$

- $$P_x(y) = Q[Y] = \frac{Q[X, Y]}{Q[X]} = \frac{M[x, y \,|\, r; w]}{M[x \,|\, r; w]}$$

Expressing a Causal Effect as a Function of mSBDs

Any identifiable causal effect can be represented as a function of mSBDs:

$$P_{\mathbf{x}}(\mathbf{y}) = A(\{M_a\}),$$

where M_a is the mSBD adjustment, and A is a multiplication/ division/marginalization of M_a.

DML-ID Estimator for Any Identifiable Causal Effects

$$P_{\mathbf{x}}(\mathbf{y}) = A(\{M_a\}).$$

$T := A\left(\{\hat{M}^a\}\right)$, where $\hat{M}^a$ denotes the DML estimator for the mSBD adjustment M^a.

Robustness Property of the DML Estimator

error

$$:= \mathbb{E}[T] - A(\{M^a\})$$

$$= O_P\left(\sum_a \sum_{i=1}^{n} \| \hat{H}^{i,a} - H_0^{i,a} \| \| \hat{\pi}^{i,a} - \pi_0^{i,a} \| \right)$$

where $H^{i,a}$, $\pi^{i,a}$ denotes the nuisance of the mSBD adjustment M^a.

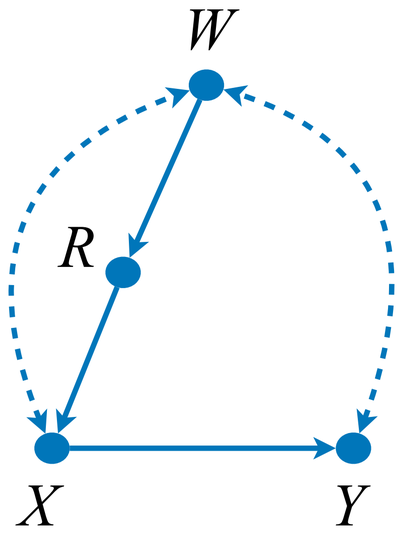
Properties of DML Estimators

$$\text{error} = O_P\left(\sum_a \sum_{i=1}^{n} \| \hat{H}^{i,a} - H_0^{i,a} \| \| \hat{\pi}^{i,a} - \pi_0^{i,a} \|\right)$$

Doubly Robustness: The error is 0 (i.e., T is unbiased) if, for all a, $\hat{H}^{i,a} = H_0^{i,a}$; or $\hat{\pi}^i = \pi_0^i$ for$i = 1,2,\cdots,$

Debiasedness: The error converges at $N^{-1/2}$ rate if $\hat{H}^{i,a}$ and $\hat{\pi}^{i,a}$ converges to $H_0^{i,a}$ and $\pi_0^{i,a}$ at slower $N^{-1/4}$ rate.

Example of DML-ID: Napkin



$$P(y \mid do(x)) = \frac{M[x, y \mid r; w]}{M[x \mid r; w]} = \frac{M^a}{M^b}$$

$T := \frac{\hat{M}^a}{\hat{M}^b}$ where $\hat{M}^i$ denote the DML estimator.

$$\hat{M}^a := \mathbb{E}_D \left[\frac{I_r(R)}{\hat{\pi}(R \mid W)} \{ I_{x,y}(X, Y) - \hat{H}^a(R) \} + \hat{H}^a(r) \right]$$

$$\hat{M}^b := \mathbb{E}_D \left[\frac{I_r(R)}{\hat{\pi}(R \mid W)} \{ I_x(X) - \hat{H}^b(R) \} + \hat{H}^b(r) \right]$$

Example of DML-ID: Napkin

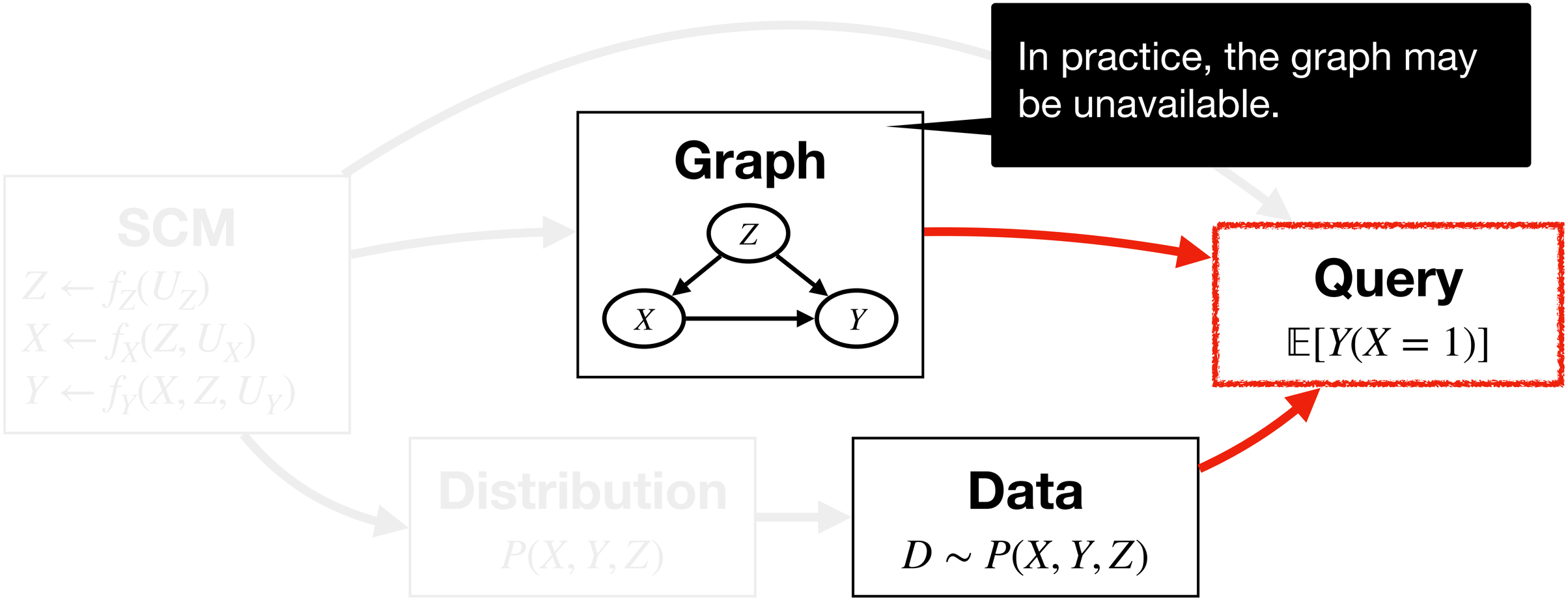
$$T - P(y \mid do(x)) = O_P(\| M^a - \hat{M}^a \| + \| M^b - \hat{M}^b \|)$$
$$= O_P(\| H^a - \hat{H}^a \| \| \pi - \hat{\pi} \| + \| H^b - \hat{H}^b \| \| \| \pi - \hat{\pi} \|)$$

T achieves doubly robustness and debiased w.r.t its parameter.

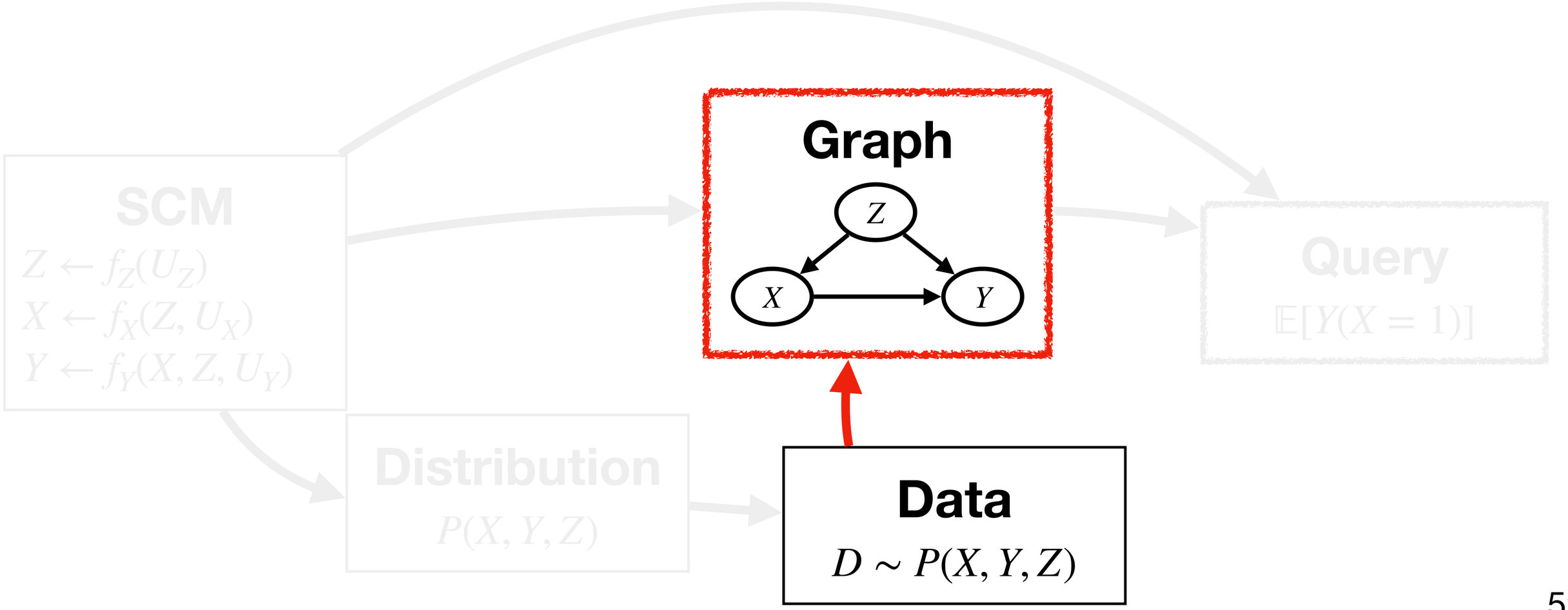
DML Estimator for Any Identifiable Causal Effects in Markov Equivalence Class

Y.Jung, J. Tian, E. Bareinboim. **ICML-21.**

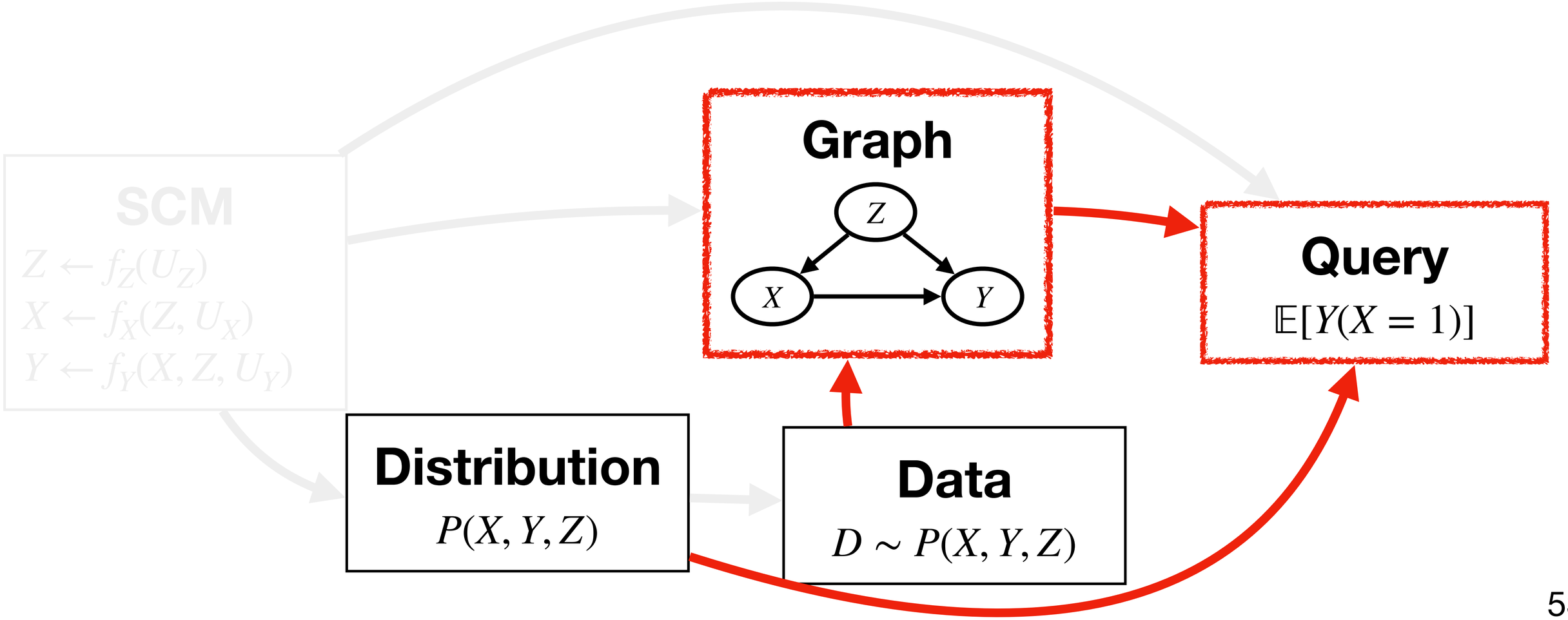
Data-Driven Causal Inference: Motivation



Causal Effect Discovery

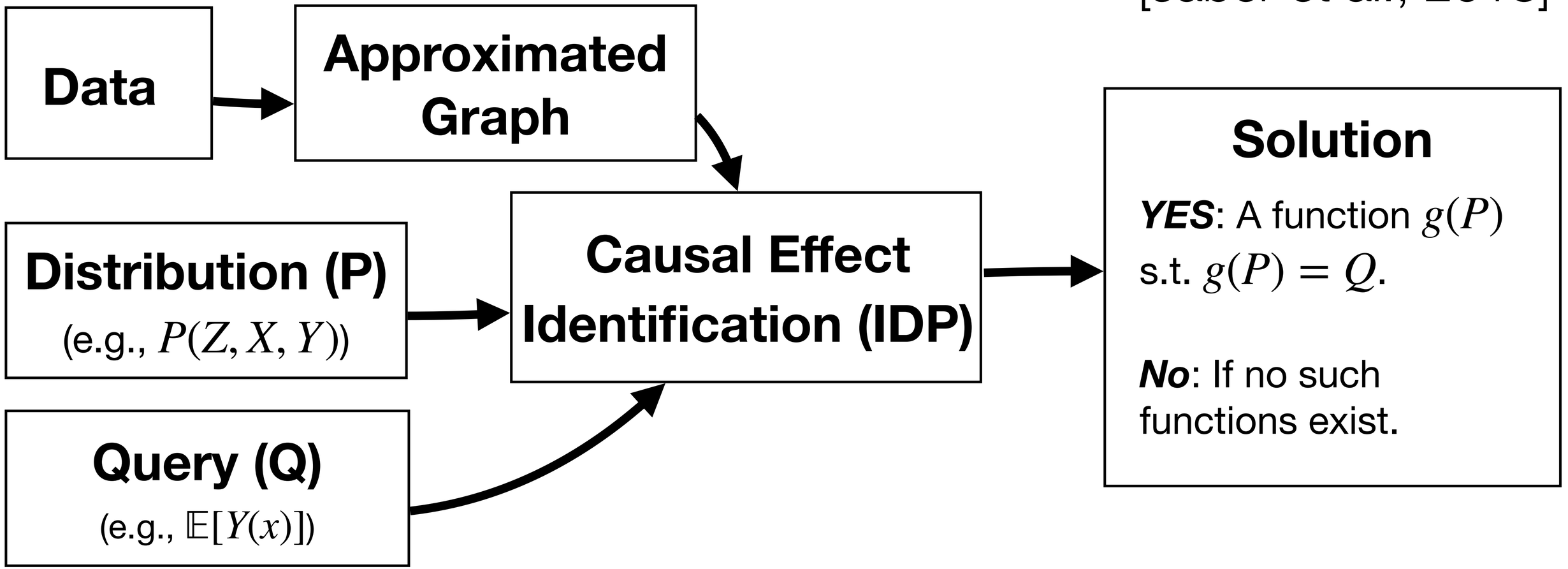


Data-Driven Causal Inference: Causal Effect Identification



Data-Driven Causal Inference: Causal Effect Identification

[Jaber et al., 2018]



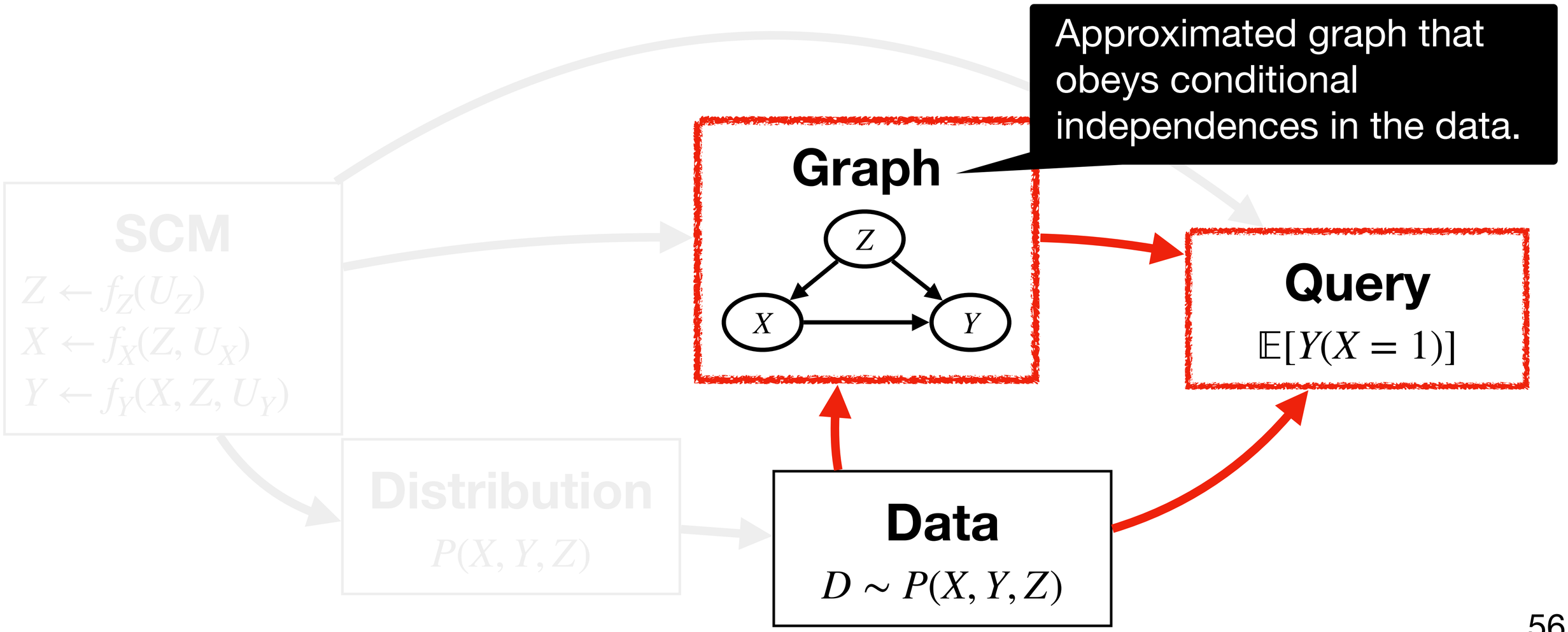
Data-Driven Causal Inference: Causal Effect Identification

The complete solution for the causal effect identification exists.

↳An algorithm for the causal effect identification exists. The algorithm states “YES” ***if-and-only-if*** the causal effect is identifiable.

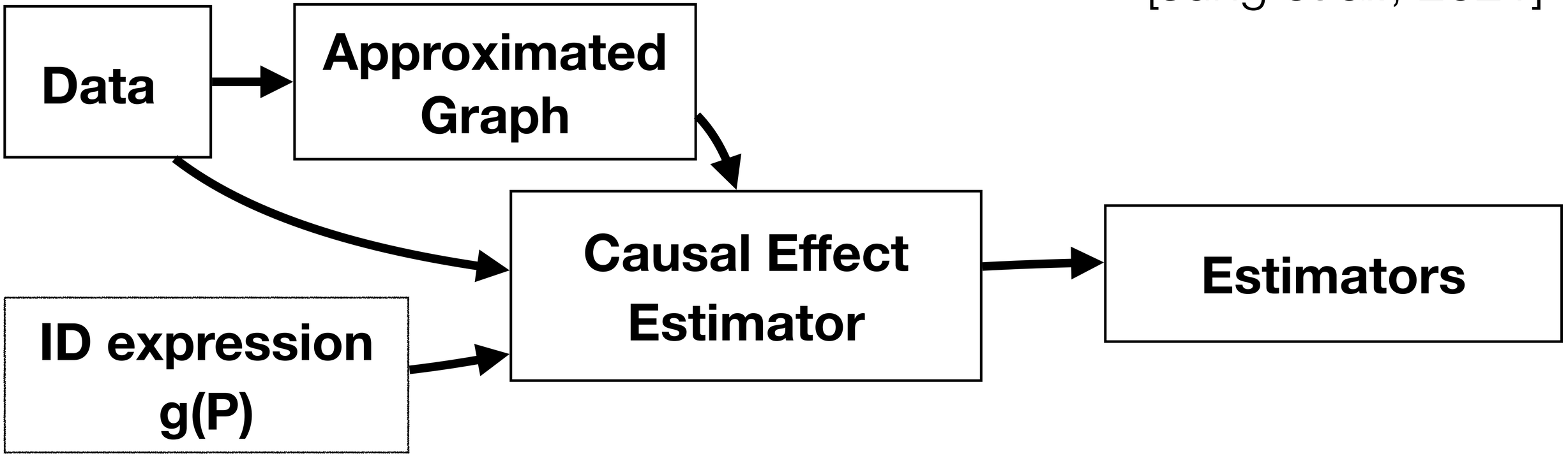
↳Jaber et al., (2018, 2019)

Data-Driven Causal Inference: Big Picture



Data-Driven Causal Inference: Our Task

[Jung et al., 2021]



Data-Driven Causal Inference: Our Task

In [ICML-21], we developed the DML estimator for generally identifiable causal effects in the approximated graph from causal discovery algorithms.

DML Estimator for any Identifiable Causal Effects in Markov Equivalence Class

Like the DML-ID, any ID expression derived by the IDP algorithm is also composed of mSBD adjustments.

$$P_{\mathbf{x}}(\mathbf{y}) = A(\{M^a\}_{a=1}^{m})$$

DML Estimator for any Identifiable Causal Effects in Markov Equivalence Class

Let $\hat{M}$ denote the DML estimator for the mSBD adjustment. Then,

$$T := A(\{\hat{M}^a\}_{a=1}^m)$$

The same error analysis with the previous can be applied.

On Measuring Causal Contribution via do-intervention

Y. Jung, S. Kasiviswanathan, J. Tian, D. Janzing,
P. Blöbaum, E. Bareinboim. **ICML-22**

What is Interpretability?

Interpretability is the degree to which a human can

1. consistently predict the model's result [Kim et al., 2016]
2. understand the cause of a prediction [Miller, 2019]

This leads “Feature attribution task”
taking account of Causality!

Task of Interpretable Machine Learning

Feature attribution given $(\mathbf{v}, f(\mathbf{v}))$

- **Input**: A pair of $(\mathbf{v}, f(\mathbf{v}))$, where $f(\mathbf{v})$ is a black-box machine learning model prediction for some input $\mathbf{v} = \{v_1, v_2, \cdots, v_n\}$ (where v_i means the ith feature).

- **Output**: A vector $attr(f, \mathbf{v}) \equiv \{\phi_{v_1}, \cdots, \phi_{v_n}\}$ where ϕ_i is an importance of v_i.

Preliminary

Task of the Shapley Value

Let $\nu([n])$ denote the value made by the coalition of $[n] := \{1,2,\cdots,n\}$ players.

Let $\nu(S)$ denote the value made by the coalition of $S \subseteq [n]$.

The task of the Shapley value is to attribute an individual player i to the target value $\nu([n])$.

Shapley value

The Shapley value is a weighted average of the marginal contribution of the player i (i.e., $\nu(S \cup i) - \nu(S)$) under all possible coalition S.

$$\phi_i(\nu) := \frac{1}{n} \sum_{S \subseteq [n] \setminus i} \binom{n-1}{|S|}^{-1} \{\nu(S \cup i) - \nu(S)\}$$

Characterization of the Shapley value

The Shapley value is the unique attribution satisfying these four properties!

- Its sum equals to the total value $\nu([n])$ (“*Perfect Assignment*”)
- $\phi_i = 0$, if $\nu(S \cup i) = \nu(S)$ for all $S \subseteq [n] \backslash i$ (“*Dummy player*”)
- $\phi_i = \phi_j$ if $\nu(S \cup i) = \nu(S \cup j)$ for all $S \subseteq [n] \backslash \{i, j\}$ (“*Symmetry*”)
- ϕ_i is a linear function of $\{\nu(S)\}_{S \subseteq [n]}$ (“Linearity”)

Existing Shapley Value-based Attribution Method

[Lundberg & Lee, 2017] propose "SHAP" or "Conditional Shapley", which is defined as follow

$$\phi_i := \frac{1}{n} \sum_{S \subseteq [n] \setminus i} \binom{n-1}{|S|}^{-1} \left\{ \mathbb{E}[Y \,|\, v_i, \mathbf{v}_S] - \mathbb{E}[Y \,|\, \mathbf{v}_S] \right\}$$

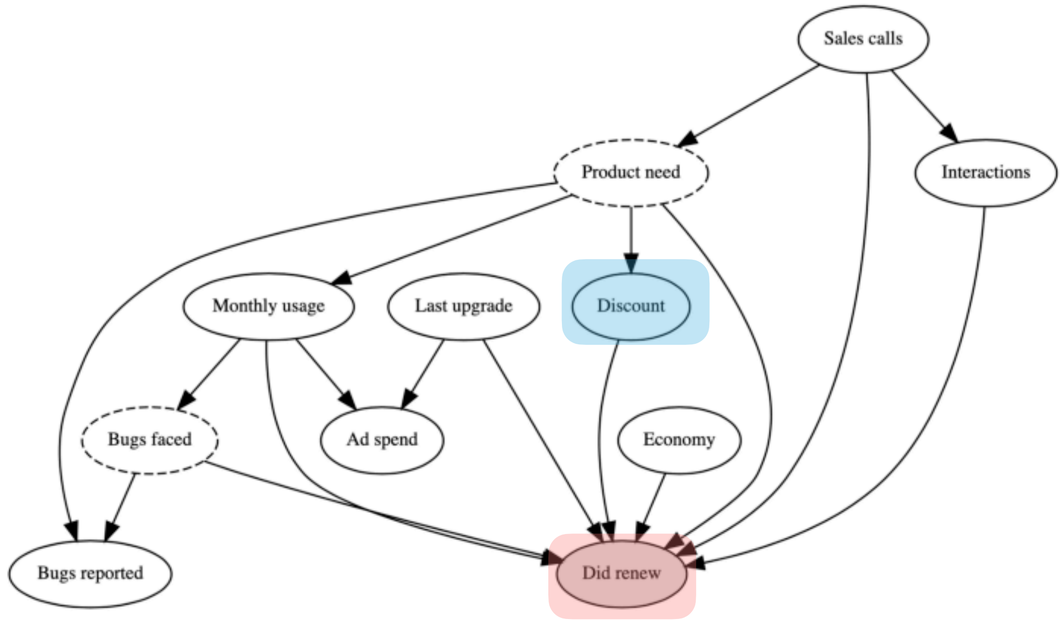
Failure of Conditional Shapley - (1)

Scenario: Predict customers' retention rate.

The data-generating process is here:

We measure the feature importance of "Discount" to explain Retention.

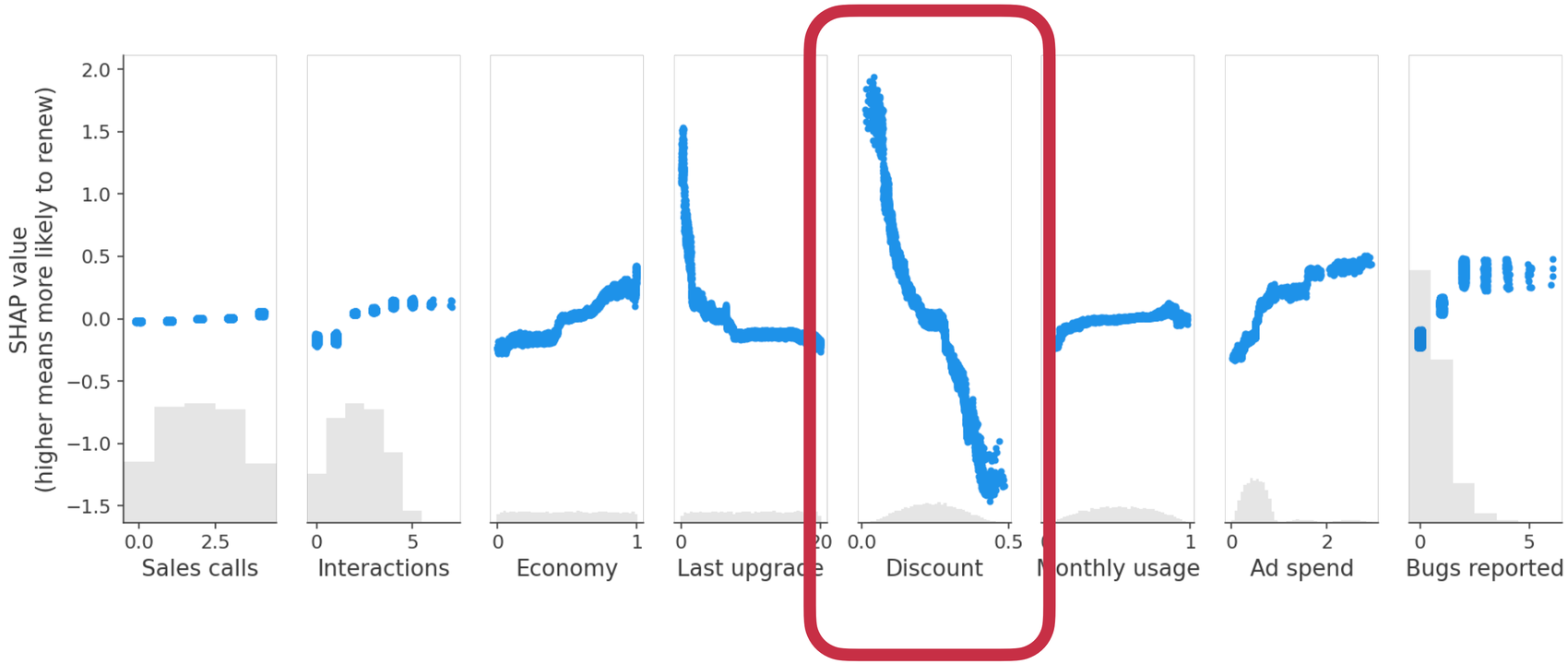
$$\mathbb{E}[\text{Rentention} \mid \text{Discount}, \mathbf{v}_S]$$



https://towardsdatascience.com/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6

Failure on practical examples - 2

The results state that providing more discount leads to less retention. 🤷‍♀️

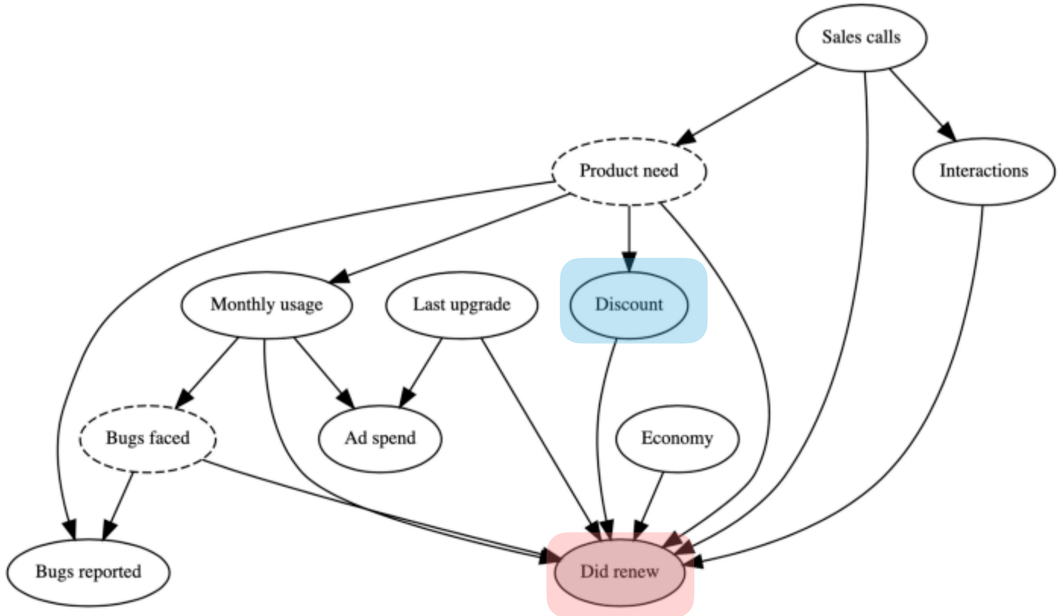


https://towardsdatascience.com/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6

Failure on practical examples - 3

Lundberg, who developed SHAP, diagnosed this model fails due to the lack of considering causality.

“interpreting a normal predictive model as causal are often unrealistic.”



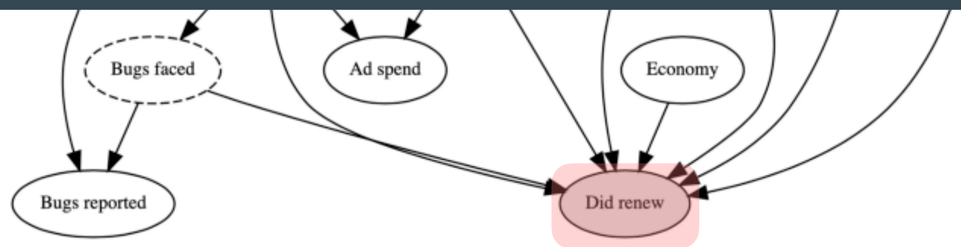
https://towardsdatascience.com/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6

Failure on practical examples - 3

Lundberg, who developed SHAP, diagnosed this model fails due to the lack of consider

“interp

Feature attribution method must take account of causality!



https://towardsdatascience.com/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6

Introduction to do-Shapley

Task: Feature Attribution based on Causality

Causal Graph
$G = G(\mathbf{V})$

Samples $D \sim P(\mathbf{V})$
where P is
compatible w/ G

Target
$Q := \mathbb{E}[Y \mid do(\mathbf{v})]$

Feature
Attribution

Contribution of
$v_i \in \mathbf{v}$ to Q based
on causality

Our task is to measure the contribution of $v_i \in \mathbf{v}$ to the target effect $\mathbb{E}[Y \mid do(\mathbf{v})]$ based on causality.

Task: Application to ML Interpretation

Causal Graph
$G = G(\mathbf{V})$

Samples
$D \sim P(\mathbf{V})$

Output of the ML model $f(\cdot)$
$Q := f(\mathbf{v})$

Feature Attribution

Contribution of inputs $v_i \in \mathbf{v}$ to the ML output $f(\mathbf{v})$

When the outcome is a ML model output $Y := f(\mathbf{V})$, the target is reduced to $Q := f(\mathbf{v}) = \mathbb{E}[Y \mid do(\mathbf{v})]$, and the problem reduces to measuring the importance of inputs.

do (causal) -Shapley value

The do-Shapley value is a weighted average of the marginal contribution of the variable v_i (i.e., $\mathbb{E}[Y|do(\mathbf{v}_{S\cup i})] - \mathbb{E}[Y|do(\mathbf{v}_S)]$) among all possible coalition S.

$$\phi_{v_i} := \frac{1}{n} \sum_{S \subseteq [n]} \binom{n-1}{|S|}^{-1} \left\{ \mathbb{E}[Y|do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y|do(\mathbf{v}_S)] \right\}$$

Characterizing Properties of do-Shapley value

The do-Shapley value is a unique measure satisfying these four causality properties!

"**Assignment**": Its sum equals to $f(\mathbf{x}) = \sum_{x_i \in \mathbf{x}} \phi_{x_i}$.

"**Causal Irrelevance**": $\phi_{v_i} = 0$, if $\mathbb{E}[Y \,|\, do(x_i, \mathbf{x}_S)] = \mathbb{E}[Y \,|\, do(x_i', \mathbf{x}_S)]$ for all $\mathbf{X}_S \subseteq \mathbf{X}$.

"**Causal Symmetry**": $\phi_{v_i} = \phi_{v_j}$ if $\mathbb{E}[Y \,|\, do(v_i), do(\mathbf{w})] = \mathbb{E}[Y \,|\, do(v_j), do(\mathbf{w})]$ for all $\mathbf{W} \subseteq \mathbf{V}$

"**Linearity**": ϕ_{v_i} is a linear function of $\mathbb{E}[Y \,|\, do(\mathbf{x}_S)]$ $\forall \mathbf{X}_S \subseteq \mathbf{V}$.

do-Shapley Identifiability

do-Shapley Identifiability - Challenge

$$\phi_{v_i} := \frac{1}{n} \sum_{S \subseteq [n]} \binom{n-1}{|S|}^{-1} \left\{ \mathbb{E}[Y \,|\, do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \,|\, do(\mathbf{v}_S)] \right\}$$

- We have to determine the identifiably of $\mathbb{E}[Y \,|\, do(\mathbf{v}_S)]$ for all $\mathbf{V}_S \subseteq \mathbf{V}$.
- This might take exponential computational time.

do-Shapley Identifiability - Challenge

Identification of do-Shapley

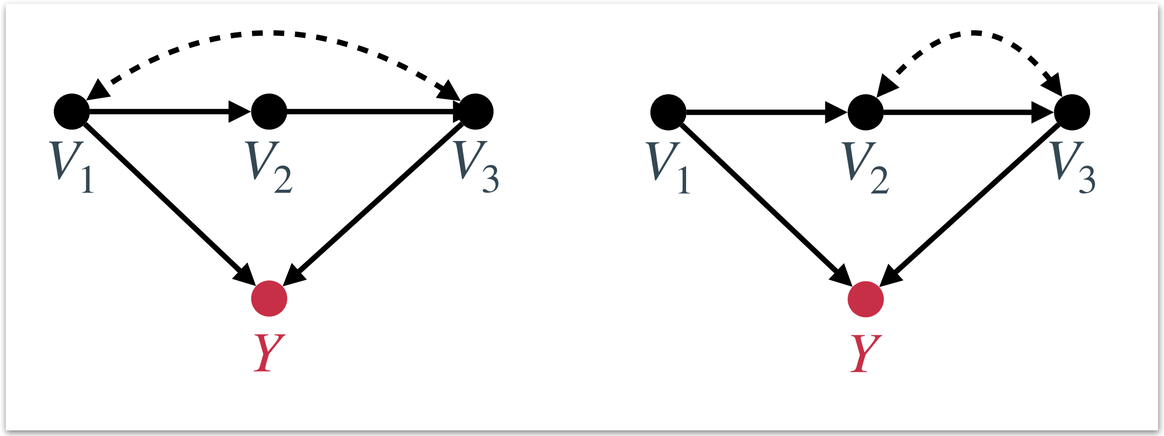
Assume Y is not connected by bidirected paths. If any variables are not connected to its children by bidirected paths (i.e., V_i and $Ch(V_i)$ are not in the same C-component), then the do-Shapley is identifiable (i.e., $\mathbb{E}[Y|do(\mathbf{v}_S)]$ for all $\mathbf{V}_S \subseteq \mathbf{V}$ is identifiable).

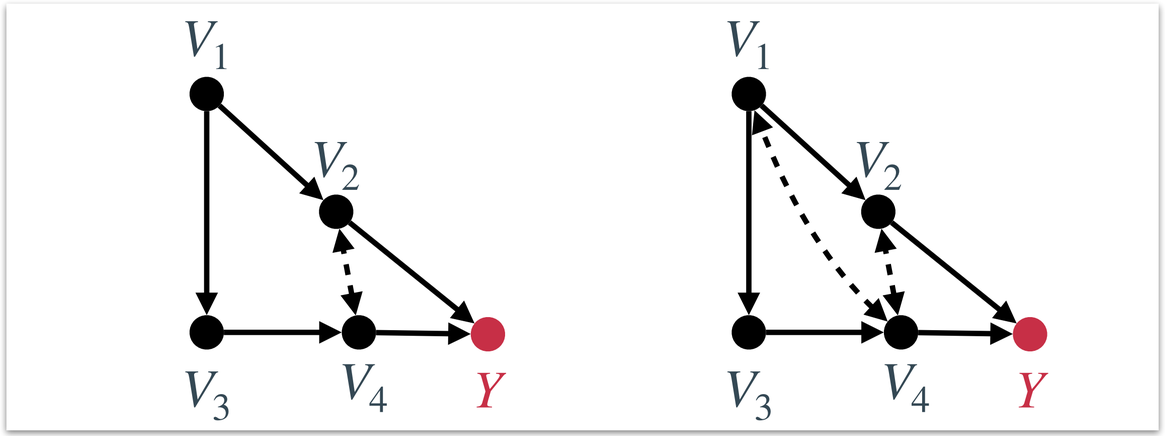
Specifically,

$$\mathbb{E}[Y|do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\bar{S}}} \mathbb{E}[Y|\mathbf{v}] \frac{P(\mathbf{v})}{\prod_{V_a \in C(\mathbf{V}_S)} P(v_a|pre(v_a))} \prod_{k=1}^{c} \sum_{\mathbf{s}_k} \prod_{V_b \in C(\mathbf{S}_k)} P(v_b|pre(v_b))$$

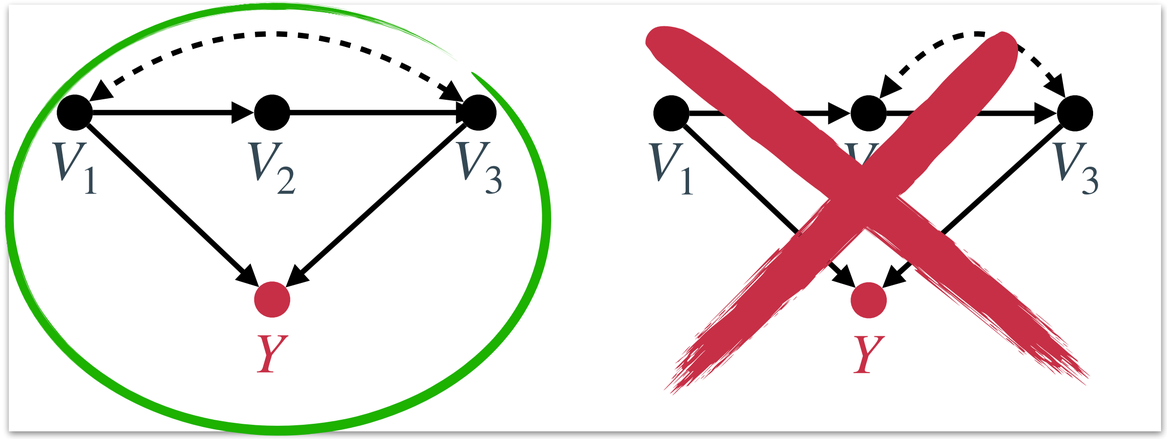
where $\mathbf{S}_k$ is some partition of $\mathbf{V}_S$.

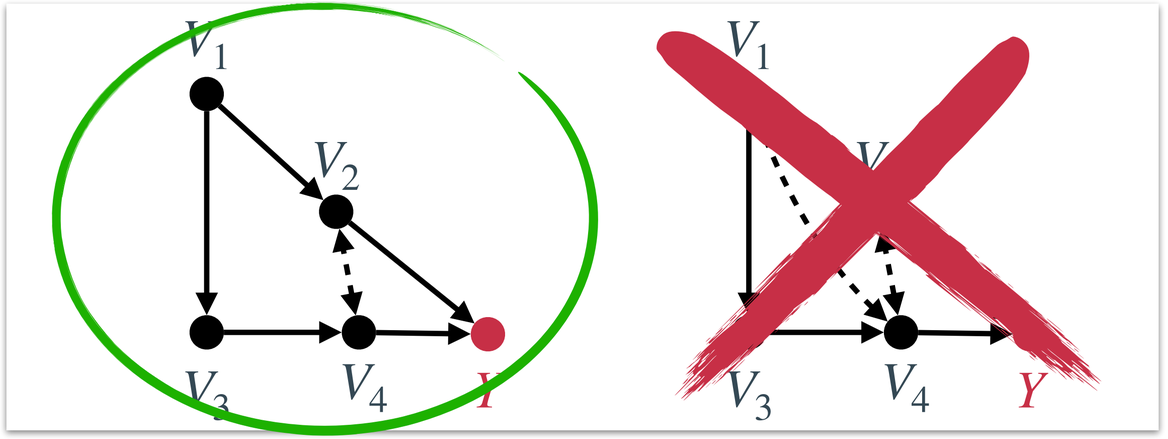
do-Shapley Identifiability: Examples



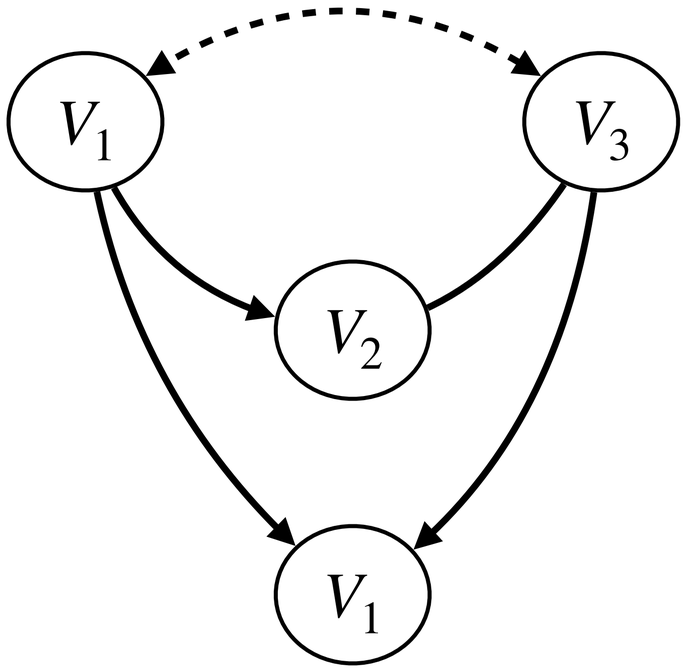


do-Shapley Identifiability: Examples





do-Shapley Identifiability: Examples



$$\mathbb{E}\left[Y|do(\mathbf{v}_S)\right]$$
$$= \begin{cases} \sum_{\mathbf{v}_{\overline{S}}} \mathbb{E}\left[Y|\mathbf{v}\right] P(v_2|v_1,v_3)P(\mathbf{v}_S), \text{ if } S \in \{1,3\}, \\ \sum_{\mathbf{v}_{\overline{S}}} \mathbb{E}\left[Y|\mathbf{v}\right] P(\mathbf{v}_{\overline{S}}), \text{ if } S \in \{\emptyset, 2, \{1,2\}, \{2,3\}\}, \\ \sum_{\mathbf{v}_{\overline{S}}} \mathbb{E}\left[Y|\mathbf{v}\right] P(\mathbf{v}_{\overline{S}}|\mathbf{v}_S), \text{if } S \in \{\{1,3\}\}, \\ \mathbb{E}\left[Y|\mathbf{v}\right] \text{ if } S = \{1,2,3\}. \end{cases}$$

do-Shapley Estimation

Monte-Carlo approximation for do-Shapley (1)

Let $\nu(S) := \mathbb{E}[Y \mid do(\mathbf{v}_S)]$, where $\mathbf{V}_S \subseteq \mathbf{V}$

$$\phi_i \equiv \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \binom{n-1}{|S|}^{-1} \{v(S \cup \{i\}) - v(S)\}.$$

$$= \frac{1}{n!} \sum_{\pi(\mathbf{V}) \in \text{perm}(\mathbf{V})} \{\nu(v_i, \text{pre}_\pi(v_i)) - \nu(\text{pre}_\pi(v_i))\}$$

all possible permutation of $\mathbf{V} = \{V_i\}_{i=1}^n$

Predecessor of V_i given the fixed permutation $\pi(\mathbf{V})$.

$$= \mathbb{E}_{\pi(\mathbf{V})} \left[\nu(v_i, \text{pre}_\pi(v_i)) - \nu(\text{pre}_\pi(v_i))\right]$$

The expectation is over the probability for each permutation order $\pi(\mathbf{V})$, where $P(\pi) = \frac{1}{n!}$.

Monte-Carlo approximation for do-Shapley (2)

$$\phi_i = \mathbb{E}_{\pi(\mathbf{V})}\left[\nu(v_i, \text{pre}_\pi(v_i)) - \nu(\text{pre}_\pi(v_i))\right].$$

$$\tilde{\phi}_i = \frac{1}{M}\sum_{m=1}^{M}\left\{\nu(v_i, \text{pre}_{\pi_{(m)}}(v_i)) - \nu(\text{pre}_{\pi_{(m)}}(v_i))\right\}$$

- For M number of randomly generated permutations of $\mathbf{V}$ (where each permutations are denoted $\pi_{(m)}$),

Monte-Carlo approximation for do-Shapley (2)

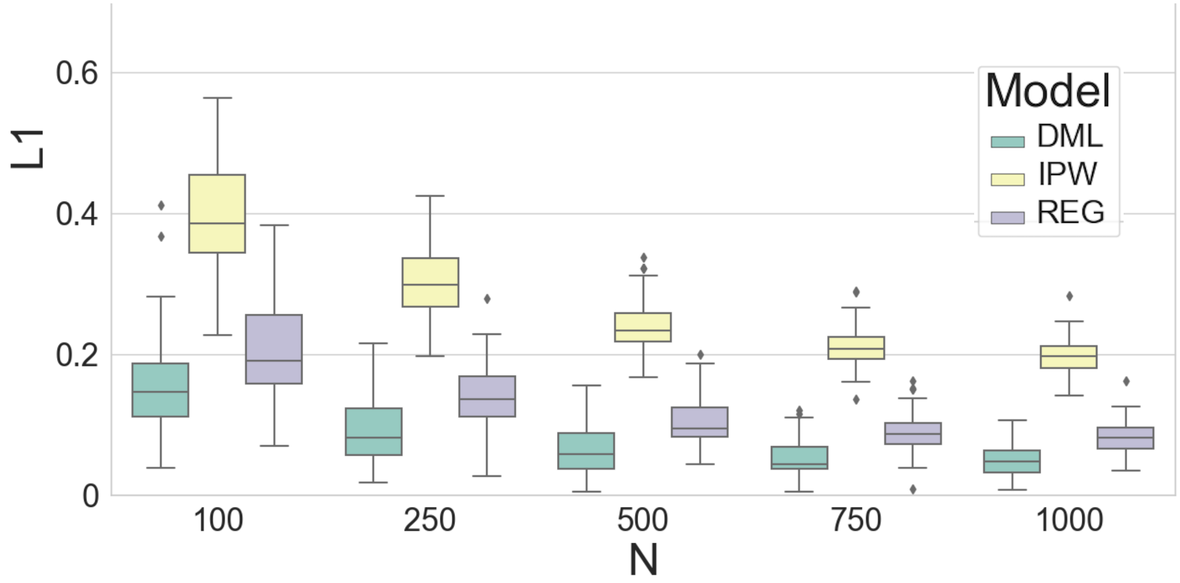
$$\tilde{\phi}_i = \frac{1}{M} \sum_{m=1}^{M} \left\{ \hat{\nu}(v_i, \text{pre}_{\pi_{(m)}}(v_i)) - \hat{\nu}(\text{pre}_{\pi_{(m)}}(v_i)) \right\}$$

- where $\hat{\nu}(S)$ is a DML estimator for $\nu(S) := \mathbb{E}[Y \,|\, do(\mathbf{v}_S)]$

Simulation

Empirical Study: DML Property

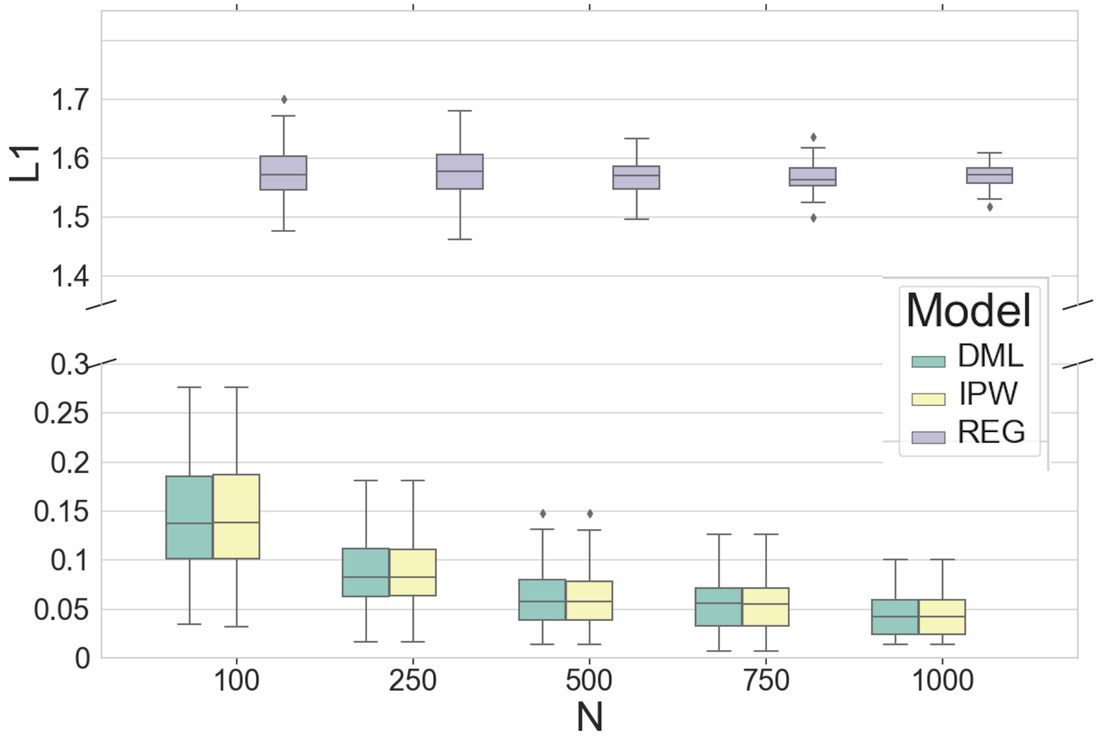
We compared the DML-based do-Shapley estimator with other existing estimators when the $\mathbb{E}[Y \mid do(\mathbf{v}_S)]$ is given as mSBD adjustment:

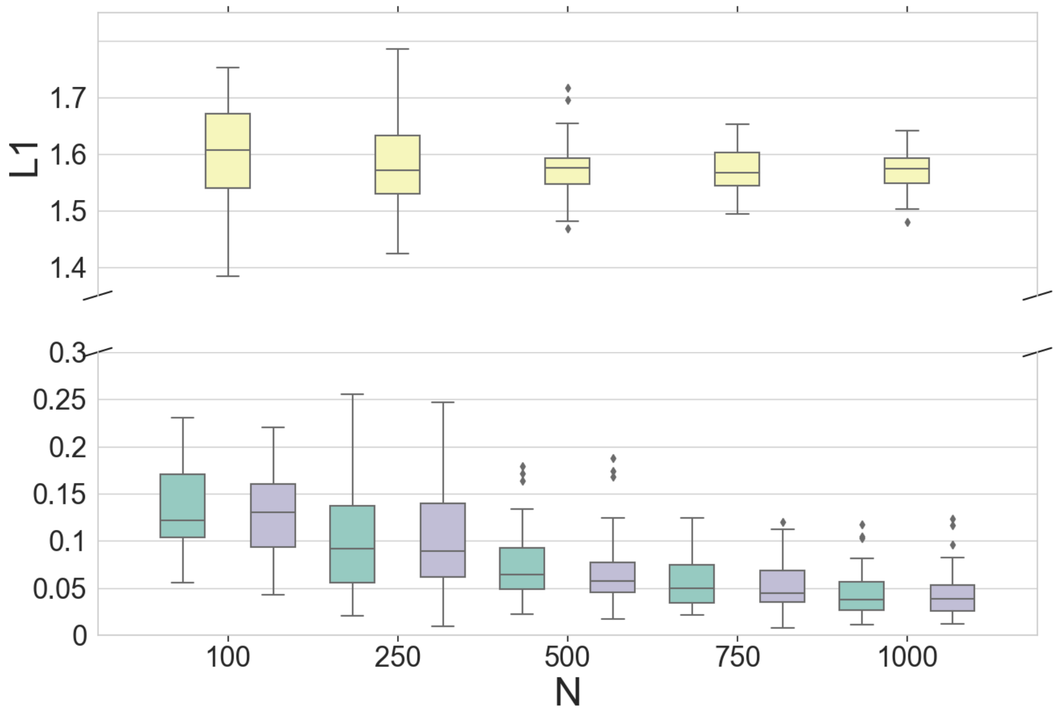


The DML estimator converges faster than competing estimators.

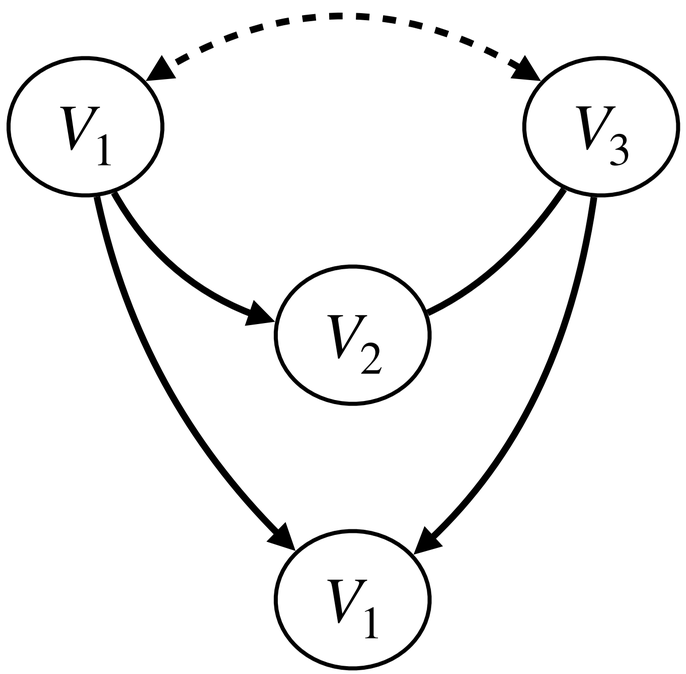
Empirical Study: DML Property

When nuisances corresponding to the IPW, REG estimators are misspecified, the DML estimator converges fast.





A simulation result



$$Y = 3V_1 + 0.4V_2 + V_3 + U_Y$$

We designed the DGP s.t. the importances are ordered as $V_1 > V_3 > V_2$.

We compared the DML-based do-Shapley based method with the conditional-Shapley.

The DML-based do-Shapley ranks $V_1 > V_3 > V_2$, while the conditional Shapley ranks V_2 as the most important one, in our scenario.

Conclusion

Conclusion

We overviewed

1. [*Estimating Identifiable Causal Effects through Double Machine Learning*] Y.Jung, J. Tian, E. Bareinboim. **AAAI-21.**

2. [*Estimating Identifiable Causal Effects on Markov Equivalence Class through Double Machine Learning*] Y.Jung, J. Tian, E. Bareinboim. **ICML-21**.

3. [*On Measuring Causal Contribution via do-intervention]* Y. Jung, S. Kasiviswanathan, J. Tian, D. Janzing, P. Blöbaum, E. Bareinboim. **ICML-22**