On Measuring Causal Contributions via do-interventions

based on:
[ICML-22, Jung et al.,]

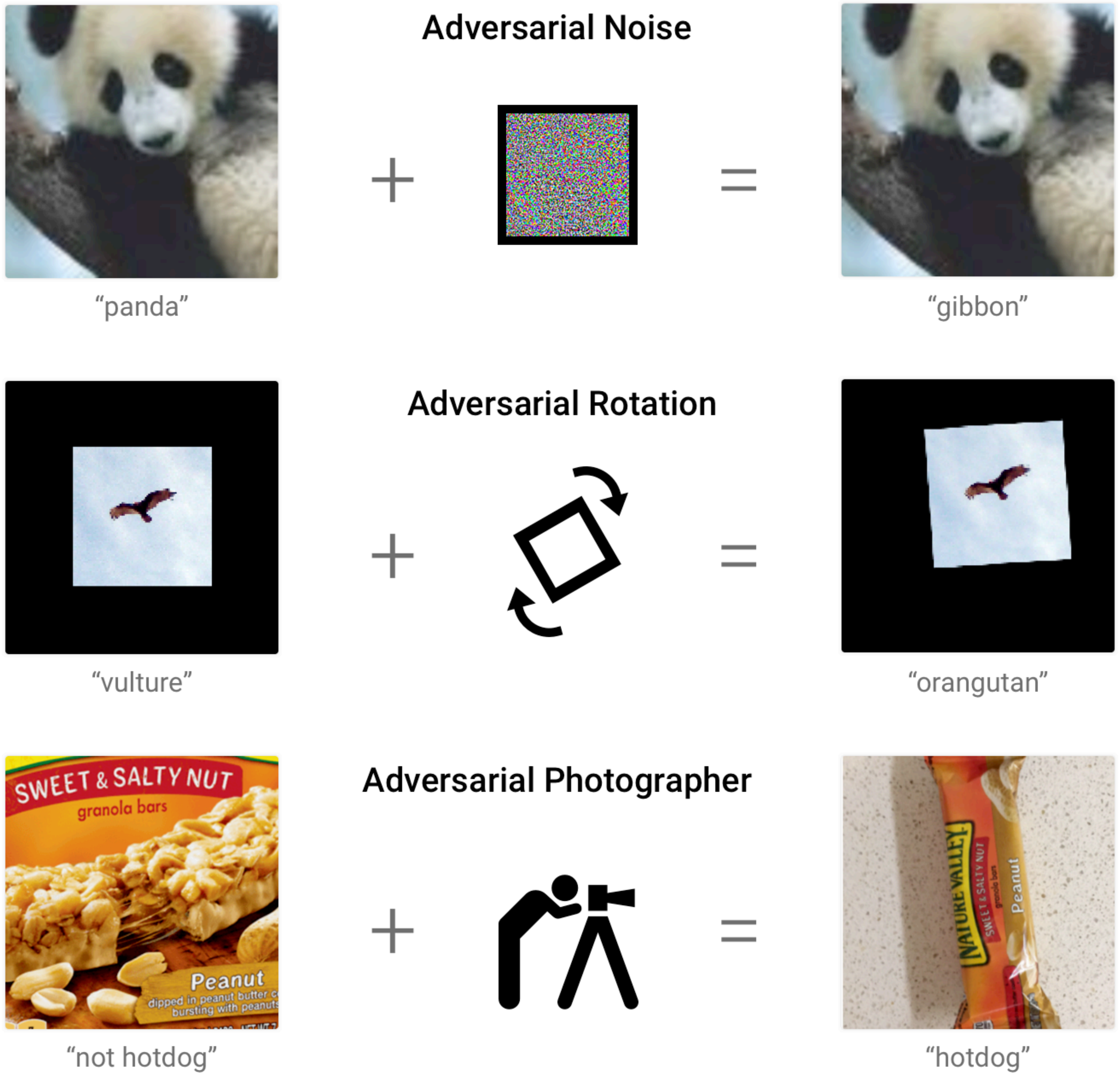
Yonghan Jung

Dept. of Computer Science, Purdue University

yonghanjung.me

May 2024

Importance of Interpretability



What is interpretability?

Common consensus on the definition of the interpretability are:

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]

"Feature attribution task" + "Causality"

Feature Attribution

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

- **Input**: A pair of $(\mathbf{x}, f(\mathbf{x}))$, where $f(\mathbf{x})$ is an ML output for some input $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ (where x_i means the ith feature).
- **Output**: A vector $attr(f, \mathbf{x}) \equiv \{\phi_1, \cdots, \phi_n\}$ where ϕ_i is interpreted as an importance of x_i.

Example: $f(x_1, x_2, x_3) = \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3$

Shapley value-based Attribution

- For any subset $\mathbf{x}_S \subseteq \{x_1, x_2, \cdots, x_n\}$, let $\nu(S) := f(\mathbf{x}_S)$ denote ML results using $\mathbf{x}_S$.

- **Shapley value**: The *weighted-average of the marginal contribution of ith feature*

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

Marginal contribution of x_i given $\mathbf{x}_S$

Axiomatic characterization of feature attribution

Shapley value is the *unique attribution method* satisfying some desirable properties.

- **Efficiency**: $\sum_{i=1}^{n} \phi_i = f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$;

  Centralized $f(\mathbf{x})$ is perfectly explained by $attr(f, \mathbf{x})$.

- **Dummy**: If $v(S \cup \{i\}) - v(S) = 0$ for all $S \subseteq [n] \backslash \{i\}$, then $\phi_i = 0$.

  If the ***marginal contribution of the player*** i ***in the team*** S, $v(S \cup \{i\}) - v(S)$, *is zero for all team* S, *then* $\phi_i = 0$

- **Symmetry** : If $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq [n] \backslash \{i, j\}$, then $\phi_i = \phi_j$.

  If the ***marginal contribution of the player*** i, j ***in the team*** S *are the same for all team, then* $\phi_i = \phi_j$.

- **Linearity** : If $f = af_1 + bf_2$, then $\phi_i(f) = a\phi_i(f_1) + b\phi_i(f_2)$.

  If the ***marginal contribution of the player*** i, j ***in the team*** S *are the same for all team, then* $\phi_i = \phi_j$.

Choice of $v(S)$ in Shapley

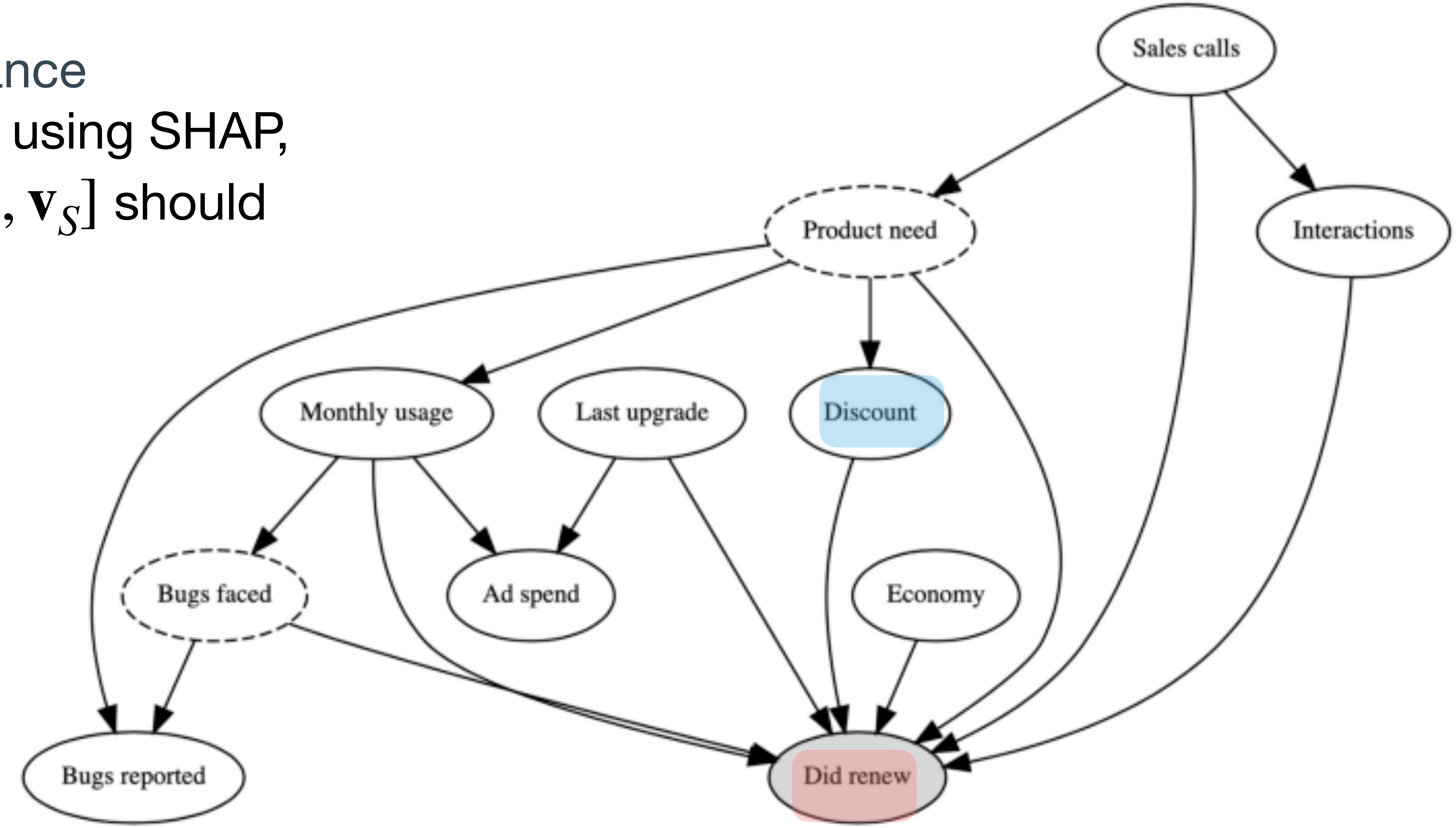
- $v(S) \equiv f(\mathbf{x}_S)$ is unclear in practice, because most ML model f is designed to take a full input $\mathbf{x}$. prediction result using a subset of features $\mathbf{x}_S \equiv \{x_i, i \in S\}$

- To address, [Lundberg & Lee, 2017] proposed $v_{cond}(S) \equiv \mathbb{E}[f(\mathbf{X}) \;\; \mathbf{x}_S]$ as a proxy of $f(\mathbf{x}_S)$. Shapley values induced by v_{cond} is "SHAP" or "Conditional Shapley"

- SHAP becomes one of the most popular feature attribution method. However, many pointed out that results of SHAP doesn't match with the human intuition [Janzing et al., 2020, Sundararajan and Najmi, 2020].

Failure on practical examples - 1

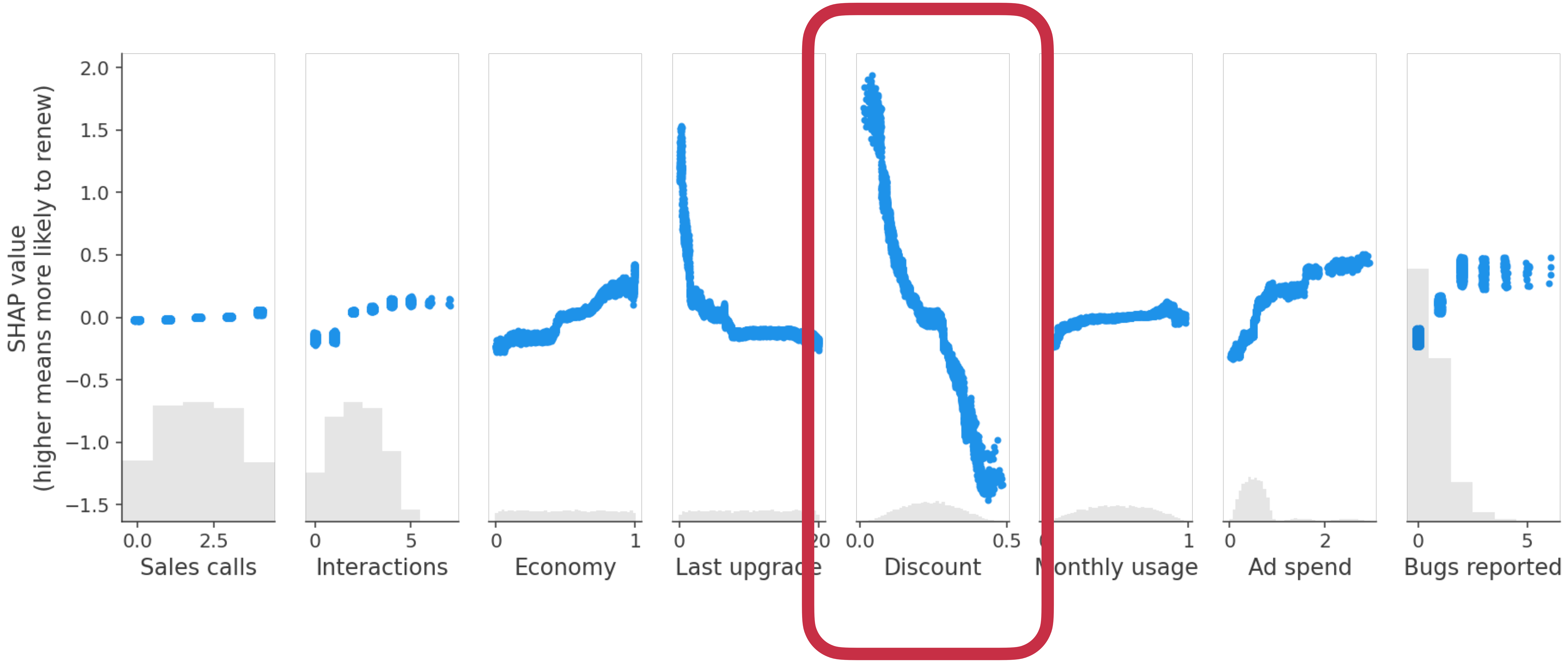
- To compute the importance "Discount" to Retention using SHAP, $\mathbb{E}[\text{Rentention} \;\; \text{Discount}, \mathbf{v}_S]$ should be computed.



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

Failure on practical examples - 2

As Discount value increases, it gives less explainability for Retention



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

Failure on practical examples - 3

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

$\mathbb{E}[\text{Rentention} \quad \text{Discount}, \mathbf{v}_S]$ models an `*association*’ between Discount and Retention, rather than the `causation’ of Discount to Retention.

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

Outline

We develop ***causally*** interpretable ***feature attribution*** *method*.

1. We *axiomatize* a causally interpretable feature attribution method, and propose do-Shapley values.

2. We provide *identifiability* condition where the do-Shapley values can be inferred from the observational data.

3. We construct a *double/debiased machine learning (DML)* based do-Shapley estimator for practical settings.

Outline

We develop ***causally*** interpretable ***feature attribution*** *method*.

1. We *axiomatize* a causally interpretable feature attribution method, and propose do-Shapley values.

2. We provide *identifiability* condition where the do-Shapley values can be inferred from the observational data.

3. We construct a *double/debiased machine learning (DML)* based do-Shapley estimator for practical settings.

Structural Causal Model

Structural Causal Model $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{u}) \rangle$

- $\mathbf{V}$: A set of endogenous (observable) variables.
- $\mathbf{U}$: A set of exogenous (latent) variables.
- $\mathbf{F}$: A set of structural equations $\{f_{V_i}\}_{V_i \in \mathbf{V}}$ determining the value of $V_i \in \mathbf{V}$, where $V_i \leftarrow f_{v_i}(PA_{V_i}, U_{V_i})$ for some $PA_{V_i} \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$.
- $P(\mathbf{u})$: A probability measure for $\mathbf{U}$.

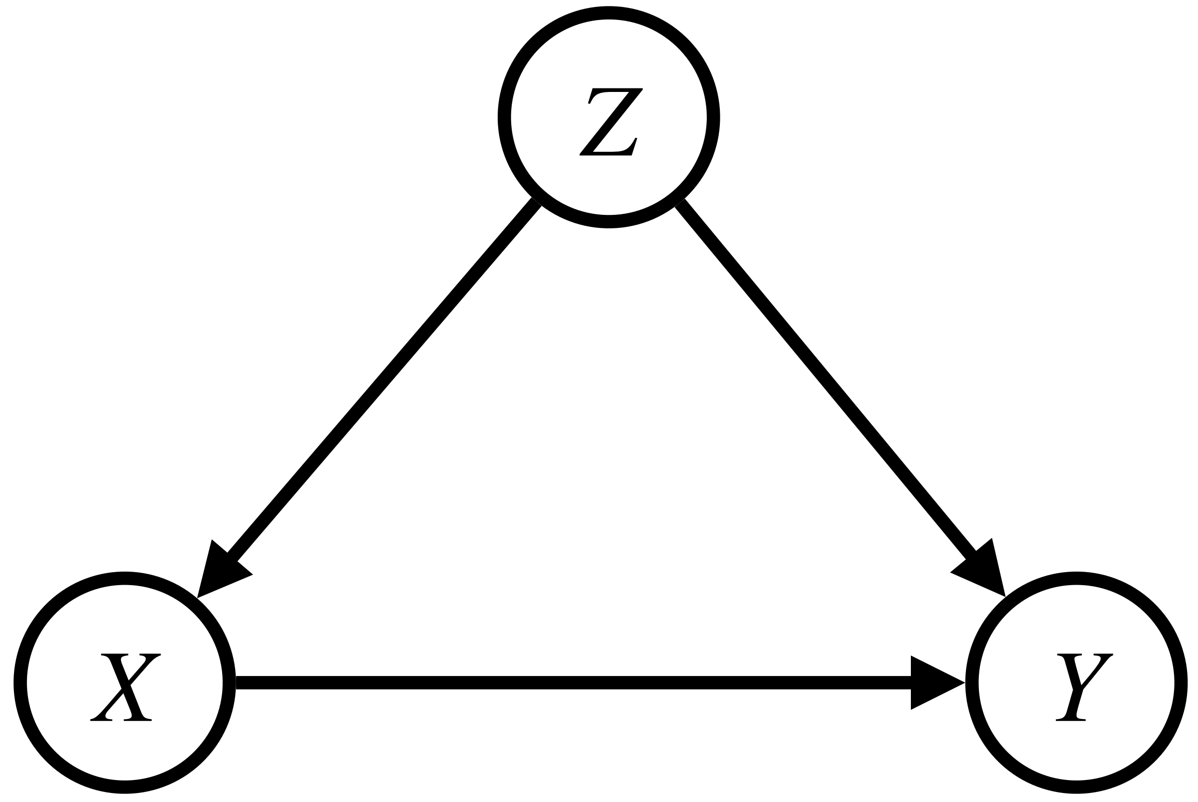
💡 An SCM induced a a "**causal graph**" $G \equiv G(\mathscr{M})$.

Causal Graphical Model

SCM

$$U_Z, U_X, U_Y \sim \text{normal}(0,1)$$
$$Z \leftarrow f_Z(U_Z)$$
$$X \leftarrow f_X(Z, U_X)$$
$$Y \leftarrow f_Y(X, Z, U_Y)$$

Graph



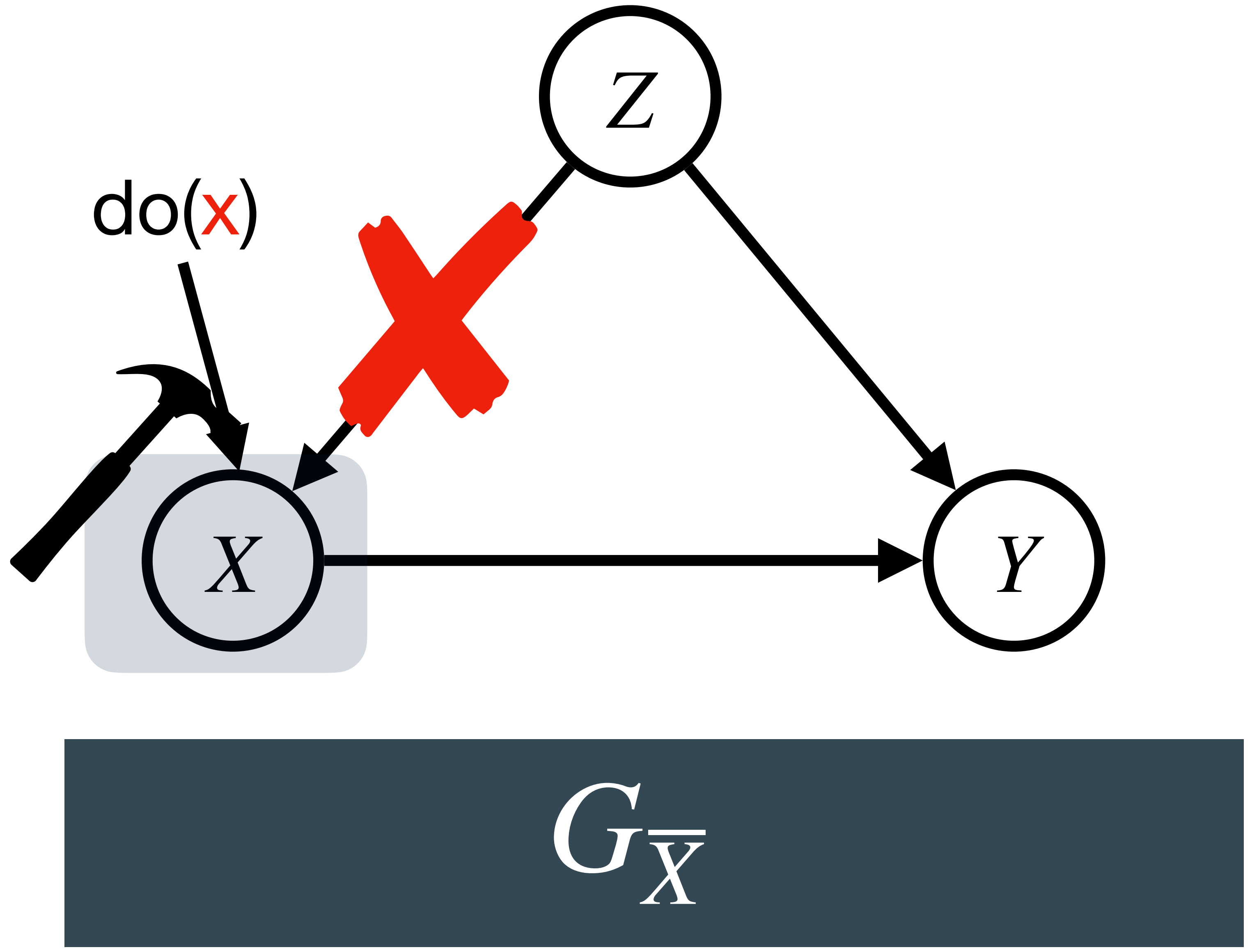
Intervention: do-operator

SCM

$U_Z, U_X, U_Y \sim \text{normal}(0,1)$
$Z \leftarrow f_Z(U_Z)$

$X \leftarrow x$ (= do(x))
$Y \leftarrow f_Y(x, Z, U_Y)$

Graph



Task: Application to ML Interpretation

For the ML model $Y := f(\mathbf{V})$, the problem is to measure the importance of $v_i \in \mathbf{v}$ to explain $f(\mathbf{v}) = \mathbb{E}[f(\mathbf{V}) \mid \mathrm{do}(\mathbf{v})]$

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

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

Output of the ML model $f(\mathbf{v})$

Feature Attribution

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

Axiom for Causal Feature Attribution

- **Perfect assignment**: $\sum_{v_i \in \mathbf{v}} \phi_{v_i} = f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$.

  Centralized $f(\mathbf{x})$ is perfectly explained by $attr(f, \mathbf{x})$.

- **Causal Irrelevance**: If V_i is *causally irrelevant* to $Y = f(\mathbf{X})$, then $\phi_{v_i} = 0$.

  $P(y \quad do(v_i)) = P(y)\ \forall y, v_i$ for $V_i \in \mathbf{V}$.

- **Causal Symmetry**: If $v_i, v_j \in \mathbf{v}$ have the same *causal explanatory power* to Y, then $\phi_{v_i} = \phi_{v_j}$.

  $P(Y \quad do(v_i), do(\mathbf{w})) = P(Y \quad do(v_j), do(\mathbf{w}))$ for $\mathbf{W} \subseteq \mathbf{V} \backslash \{V_i, V_j\}$.

- **Linearity** : If $f = af_1 + bf_2$, then $\phi_i(f) = a\phi_i(f_1) + b\phi_i(f_2)$.

do-Shapley as a desirable causal IML method

Thm. 1. Axiomatic characterization of do-Shapley

A following attribution method $attr(f, \mathbf{v}) = \{\phi_{v_i}\}_{v_i \in \mathbf{v}}$, named do-Shapley, is **uniquely satisfying the Axiom**.

$$\phi_{v_i} = (1/n) \sum_{S \subseteq [n] \backslash \{i\}} \binom{n-1}{S}^{-1} \mathbb{E}[Y \;\; do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \;\; do(\mathbf{v}_S)]\},$$

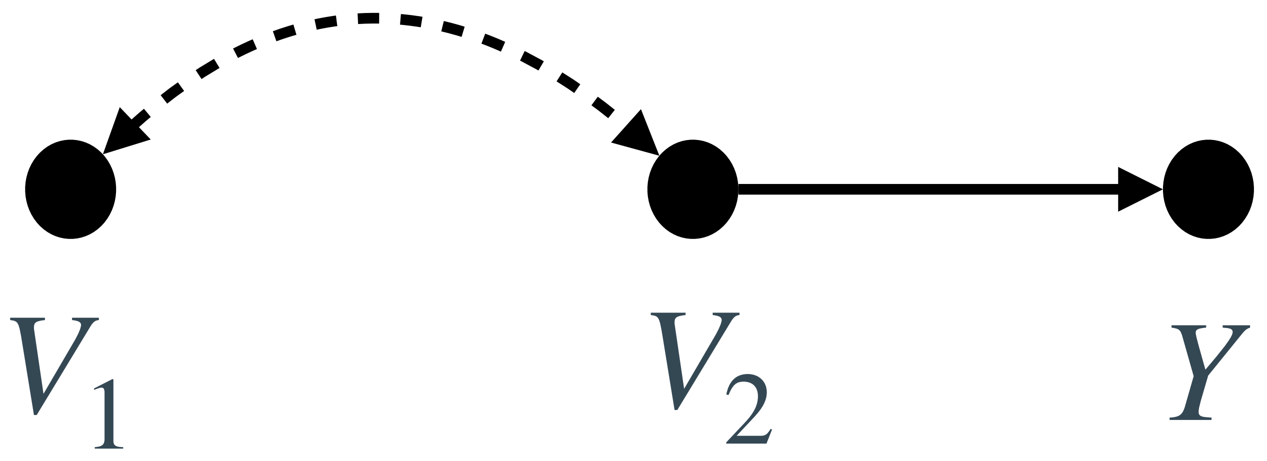
Comparison with previous works

- Two types of broadly used Shapley values:
  - Conditional Shapley uses $v(S) = v_{cond}(S) \equiv \mathbb{E}[f(\mathbf{V})\ \ \mathbf{v}_S]$; ([Lundburg & Lee, 2017])
  - Marginal Shapley uses $v(S) = v_{mar}(S) \equiv \mathbb{E}[f(\mathbf{v}_S, \mathbf{V}_{\bar{S}})]$, ([Janzing et al., 2020, Sundararajan and Najmi, 2020, Frye et al, 2021])

- [Heskes et al., 2020] propose to use $v_{do}(S) \equiv \mathbb{E}[f(\mathbf{V})\ \ do(\mathbf{v}_S)]$.
  - Assumes no latent variables.
  - Assumes that $f(\cdot)$ is accessible (i.e., can evaluate $f(\mathbf{x}')$ for any input $\mathbf{x}'$), which may be infeasible in practice.

vs. Conditional Shapley

Conditional Shapley can assign a non-zero importance to the causally-irrelevant variables.

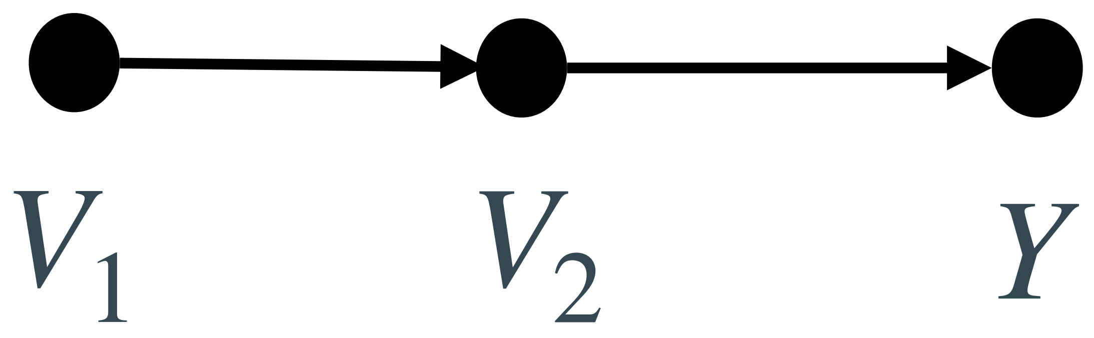


- V_1 is causally irrelevant to Y (i.e., $P(y \mid do(v_1)) = P(y)$).
- $\phi_{V_1}(\nu_{do}) = 0$, because $\nu_{do}(\{1\}) - \nu_{do}(\{\}) = \nu_{do}(\{1,2\}) - \nu_{do}(\{2\}) = 0$,
- However, it's possible that $\phi_{V_1}(\nu_{cond}) \neq 0$ [Janzing et al., 2020].

Causal Irrelevance axiom does not hold in Conditional Shapley.

vs. Marginal Shapley

Marginal Shapley always assigns zero contributions to indirect variables even if they may be root-causes of the predictions.



- It's possible that v_1 and v_2 are equally important (i.e., $\mathbb{E}[Y\ \ do(v_1)] = \mathbb{E}[Y\ \ do(v_2)]$), which leads $\phi_{V_1}(\nu_{do}) = \phi_{V_2}(\nu_{do})$, but $\phi_{V_i}(\nu_{do}) \neq 0$.

- Since $\nu_{mar}(\{1\}) - \nu_{mar}(\{\}) = \nu_{mar}(\{1,2\}) - \nu_{mar}(\{2\}) = 0$, $\phi_{V_1}(\nu_{mar}) = 0$.

⛔ **Causal Symmetry** axiom does not hold in Marginal Shapley.

Outline

We develop ***causally*** interpretable ***feature attribution*** *method*.

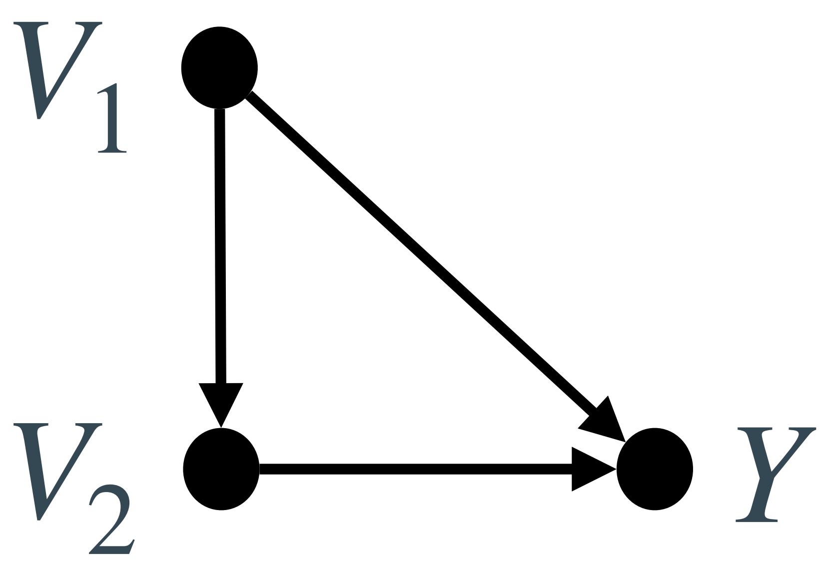
1. We *axiomatize* a causally interpretable feature attribution method, and propose do-Shapley values.

2. We provide *identifiability* condition where the do-Shapley values can be inferred from the observational data.

3. We construct a *double/debiased machine learning (DML)* based do-Shapley estimator for practical settings.

Identifiability of do-Shapley

"**Causal effect identifiability**" — Determining if $\mathbb{E}[Y \;\; do(\mathbf{v}_S)]$ can be represented as a function of $P(\mathbf{v})$; If so, $\mathbb{E}[Y \;\; do(\mathbf{v}_S)]$ can be computed using data $\mathscr{D} \sim P(\mathbf{v})$, the observational distribution.



A function of the observational distribution P

$$\mathbb{E}[Y \;\; do(v_2)] = \sum_{v_1} \mathbb{E}[Y \;\; v_1, v_2] P(v_1)$$

- The r.h.s. is a function of P, so that it's computable using data $\mathscr{D} \sim P(\mathbf{v})$.

do-Shapley Identifiability - Challenge

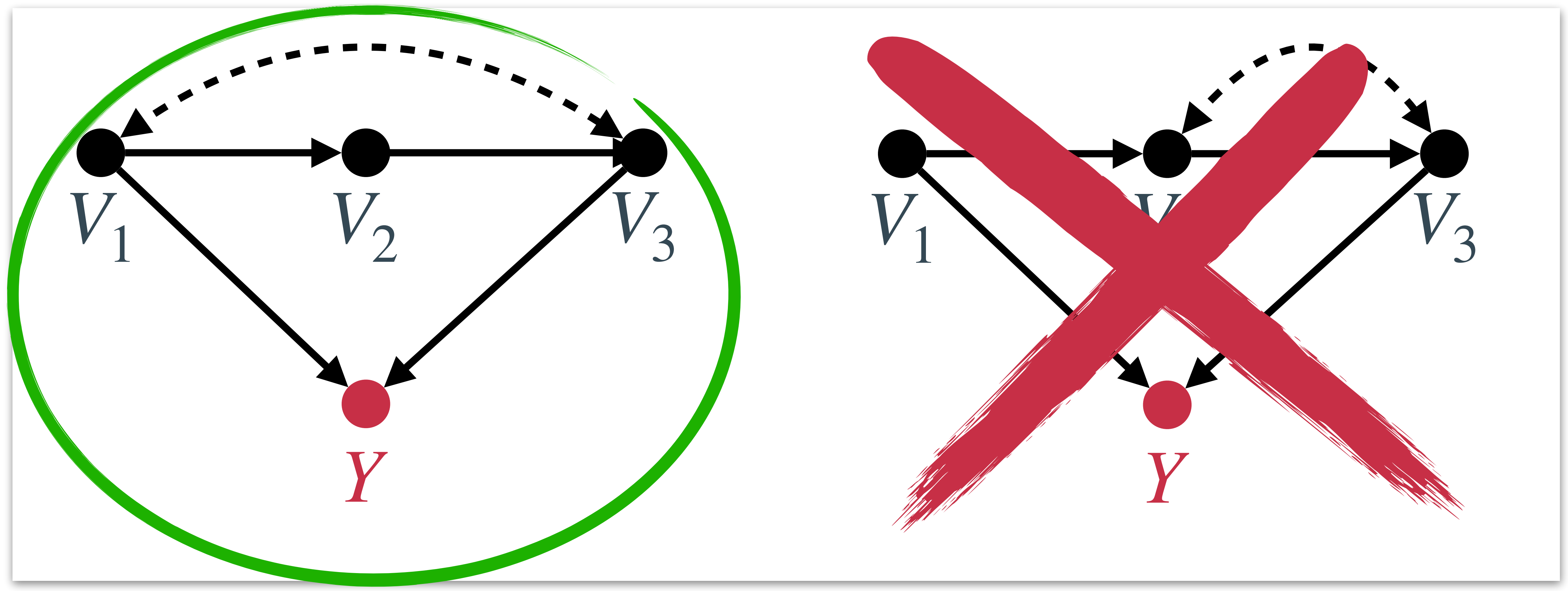
$$\phi_{v_i} := \frac{1}{n} \sum_{S \subseteq [n]} \binom{n-1}{S}^{-1} \left\{ \mathbb{E}[Y \;\; do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \;\; do(\mathbf{v}_S)] \right\}$$

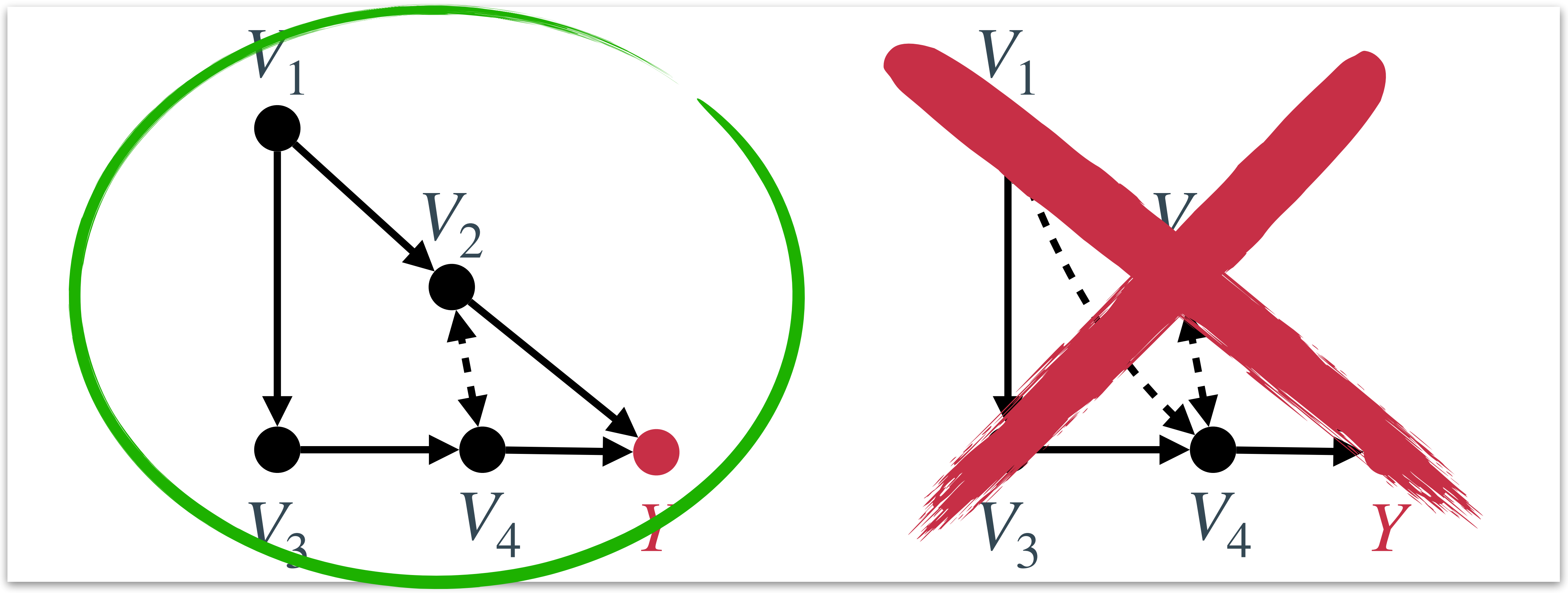
- We have to determine the identifiably of $\mathbb{E}[Y \;\; do(\mathbf{v}_S)]$ for all $\mathbf{V}_S \subseteq \mathbf{V}$.
- This might take exponential computational time.

do-Shapley Identifiability

When the unmeasured confounders exist

do-Shapley is identifiable if and only if there are no $V_i \in \mathbf{V}$ that is connected to $Ch(V_i)$ by bidirected paths.





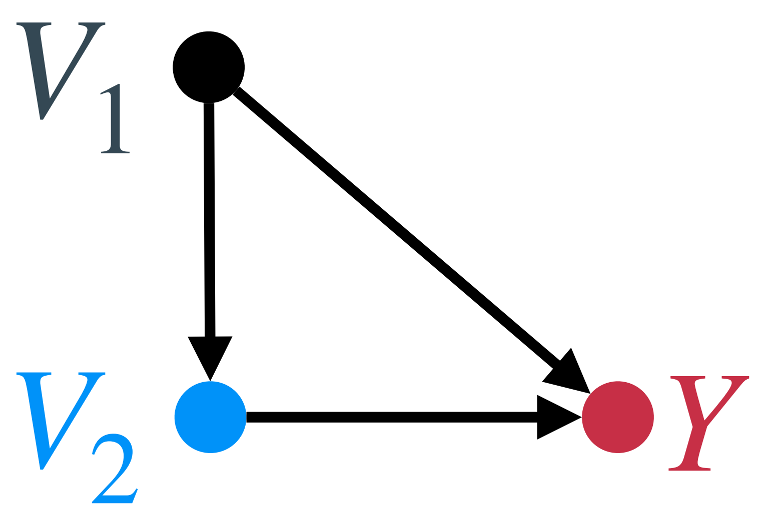
do-Shapley Identifiability

When the unmeasured confounders doesn't exist

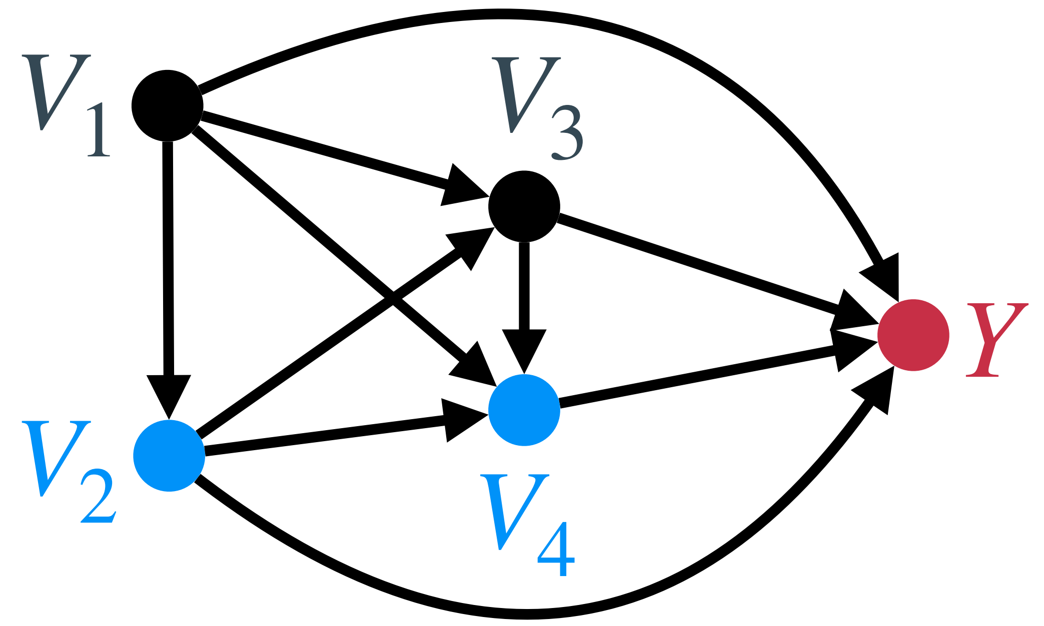
If there are no unmeasured confounders (i.e., DAG), then

$$\mathbb{E}[Y \ do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathbb{E}[Y \ \mathbf{v}_S, \mathbf{v}_{\overline{S}}] \prod_{V_i \in \mathbf{V}_{\overline{S}}} P(v_i \ pre(v_i)),$$

where $pre(V_i)$ is a predecessor of V_i given topological order on G.



$$\mathbb{E}[Y \ do(v_2)] = \sum_{v_1} \mathbb{E}[Y \ v_1, v_2] P(v_1)$$



$$\mathbb{E}[Y \ do(v_2, v_4)] = \sum_{v_1, v_3} \mathbb{E}[Y \ \mathbf{v}] P(v_3 \ v_1, v_2) P(v_1)$$

Outline

We develop ***causally*** interpretable ***feature attribution*** *method*.

1. We *axiomatize* a causally interpretable feature attribution method, and propose do-Shapley values.

2. We provide *identifiability* condition where the do-Shapley values can be inferred from the observational data.

3. We construct a *double/debiased machine learning (DML)* based do-Shapley estimator for practical settings.

Two components in do-Shapley estimation

$$\phi_{v_i} = (1/n) \sum_{S \subseteq [n] \backslash \{i\}} \binom{n-1}{S}^{-1} \mathbb{E}[Y \;\; do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \;\; do(\mathbf{v}_S)]\},$$

Computing the Shapley value requires

1. Exploring all possible subsets in $[n] \backslash \{i\}$;

Takes exponential computational time!

Random Permutation based approximation

2. Estimating $\nu_{do}(S)$ from finite samples $\mathscr{D}$.

A robust estimator to the finite sample bias is desirable!

Double/Debiased Machine Learning (DML) [Chernozhukov, 2018]

Two components in do-Shapley estimation

$$\phi_{v_i} = (1/n) \sum_{S \subseteq [n] \backslash \{i\}} \binom{n-1}{S}^{-1} \mathbb{E}[Y \ do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \ do(\mathbf{v}_S)]\},$$

Computing the Shapley value requires

1. Exploring all possible subsets in $[n] \backslash \{i\}$;

Takes exponential computational time!

Random Permutation based approximation

2. Estimating $\nu_{do}(S)$ from finite samples $\mathscr{D}$.

A robust estimator to the finite sample bias is desirable!

Double/Debiased Machine Learning (DML) [Chernozhukov, 2018]

Monte-Carlo approximation for do-Shapley (1)

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

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

[Štrumbelj and Kononenko, 2014]

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 $\pi_{(m)}$),
- Compute $\nu(v_i, \text{pre}_{\pi_{(m)}}(v_i)) - \nu(\text{pre}_{\pi_{(m)}}(v_i))$ and take an average.
- The computation time is $O(N \times \mathbf{V}\ )$

Random permutation-based algorithm

1. Initiate $\phi_{V_i} = 0$ for all $V_i \in \mathbf{V}$.

2. Generate M randomly generated permutations of $\mathbf{V}$. The permuted variables are $\mathbf{V}_\pi = \{V_{\pi,1}, \cdots, V_{\pi,n}\}$, where $V_{\pi,i}$ is the ith variable in the permutation π.

3. For each $i = 1,2,\cdots,n$, compute

$$\phi_{V_i} \leftarrow \phi_{V_i} + \left\{\mathbb{E}[Y \mid \mathrm{do}(v_{\pi,i}, pre_\pi(v_{\pi,i})] - \mathbb{E}[Y \mid \mathrm{do}(pre_\pi(v_{\pi,i})]\right\}$$

4. For each $i = 1,2,\cdots,n$, $\phi_{V_i} \leftarrow (1/M) \cdot \phi_{V_i}$.

Two components in do-Shapley estimation

$$\phi_{v_i} = (1/n) \sum_{S \subseteq [n] \setminus \{i\}} \binom{n-1}{S}^{-1} \mathbb{E}[Y \;\, do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \;\, do(\mathbf{v}_S)]\},$$

Computing the Shapley value requires

1. Exploring all possible subsets in $[n] \setminus \{i\}$;

Takes exponential computational time!

Random Permutation based approximation

2. Estimating $\nu_{do}(S)$ from finite samples $\mathscr{D}$.

A robust estimator to the finite sample bias is desirable!

Double/Debiased Machine Learning (DML) [Chernozhukov, 2018]

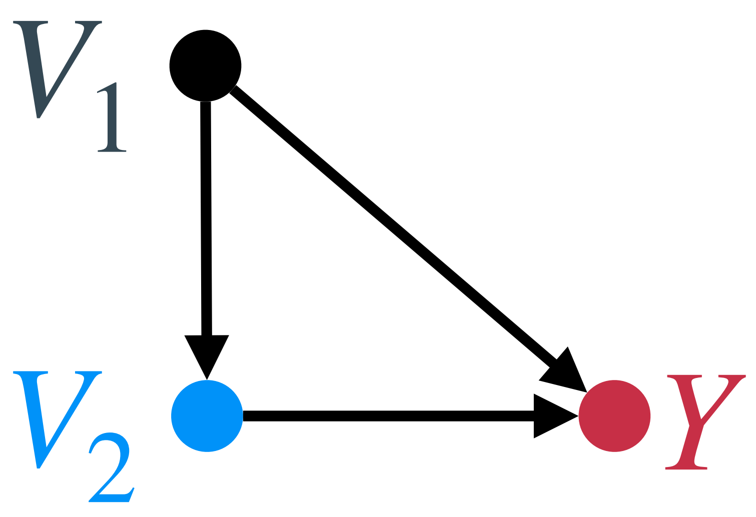
Estimation of Causal Coalition

- To estimate do-Shapley, we have to estimate $\nu_{do}(S) \equiv \mathbb{E}[Y \mid do(\mathbf{v}_S)]$ from finite samples.

$$\mathbb{E}[Y \mid do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathbb{E}[Y \mid \mathbf{v}_S, \mathbf{v}_{\overline{S}}] \prod_{V_i \in \mathbf{V}_{\overline{S}}} P(v_i \mid pre(v_i)).$$

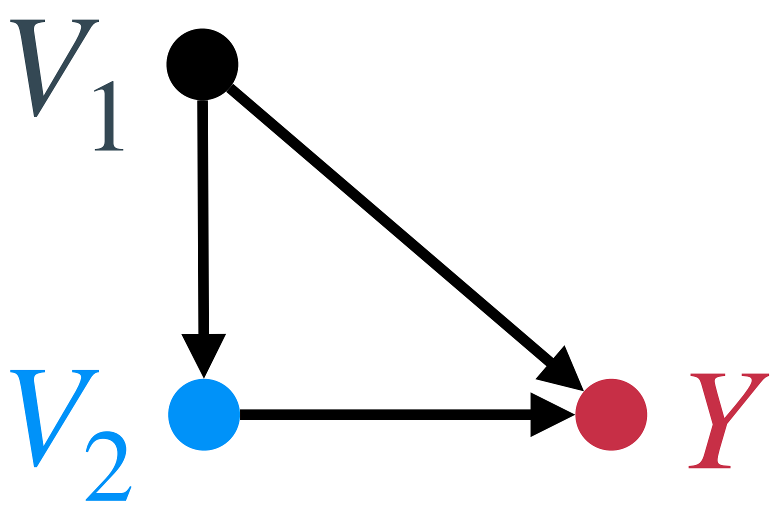
- Which estimator should we choose?

- In the presentation, we will focus on the canonical working example:



$$\mathbb{E}[Y \mid do(v_2)] = \sum_{v_1} \mathbb{E}[Y \mid v_1, v_2] P(v_1)$$

Estimation of Causal Coalition — Plug-in



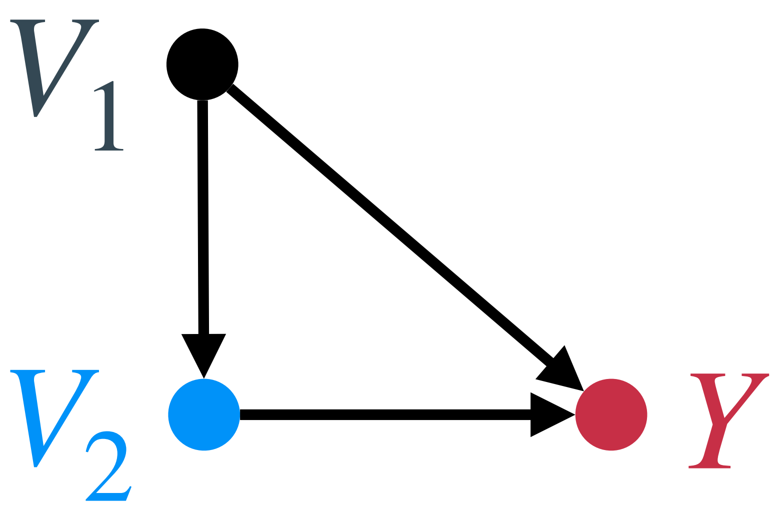
$$\mathbb{E}[Y \;\; do(v_2)] \; = \sum_{v_1} \mathbb{E}[Y \;\; v_1, v_2] P(v_1)$$

“Plug-in estimator” — estimates the functions composing the estimand (“*nuisance*") and then plug nuisances into the functional.

$$\widehat{\mathbb{E}}[Y \;\; do(v_2)] = \sum_{v_1} \widehat{\mathbb{E}}[Y \;\; v_1, v_2] \, \widehat{P}(v_1)$$

- (+) Easy; If nuisances are correct, then achieves smallest variance asymptotically.
- (−) Summation takes exponential computation time.
- (−) If nuisances are misspecified or converging slow, then the estimator is so.

Estimation of Causal Coalition — IPW



$$\mathbb{E}[Y \;\; do(v_2)] = \mathbb{E}_P\left[\frac{I_{v_2}(V_2)\cdot Y}{P(V_2 \;\; V_1)}\right]$$

"**IPW** (Inverse Probability Weighting) **estimator**"

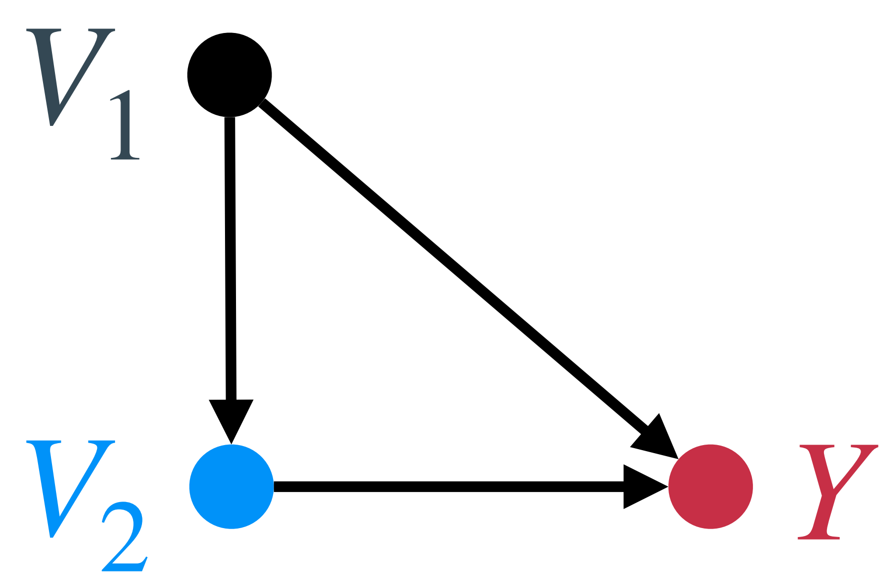
$$\widehat{\mathbb{E}}\,[Y \;\; do(v_2)] = \mathbb{E}_{\mathscr{D}}\left[\frac{I_{v_2}(V_2)\cdot Y}{\widehat{P}\,(V_2 \;\; V_1)}\right]$$

Empirical average over samples $\mathscr{D}$

Nuisances

- Can be evaluated in polynomial time.
- If $\widehat{P}\,(V_2 \;\; V_1)$ is misspecified or converging slowly, then the IPW estimator is so.

Estimation of Causal Coalition — DML



$$\mathbb{E}[Y \ do(v_2)] = \mathbb{E}_P\left[g(\mathbf{V};\eta)\right] \text{ where } \eta \equiv \{P(V_2 \ V_1), \mathbb{E}[Y \ V_1, V_2]\}$$

$$g(\mathbf{V};\eta) = \frac{I_{v_2}(V_2)}{P(V_2 \ V_1)}\left\{Y - \mathbb{E}_P[Y \ V_1, V_2]\right\} + \mathbb{E}_P[Y \ v_1, V_2]$$

"**DML** (Double/Debiased Machine Learning, [Chernozhukov, 2018]) **estimator**" T

1. Randomly split the dataset $\mathcal{D} = \{\mathcal{D}_a, \mathcal{D}_b\}$,

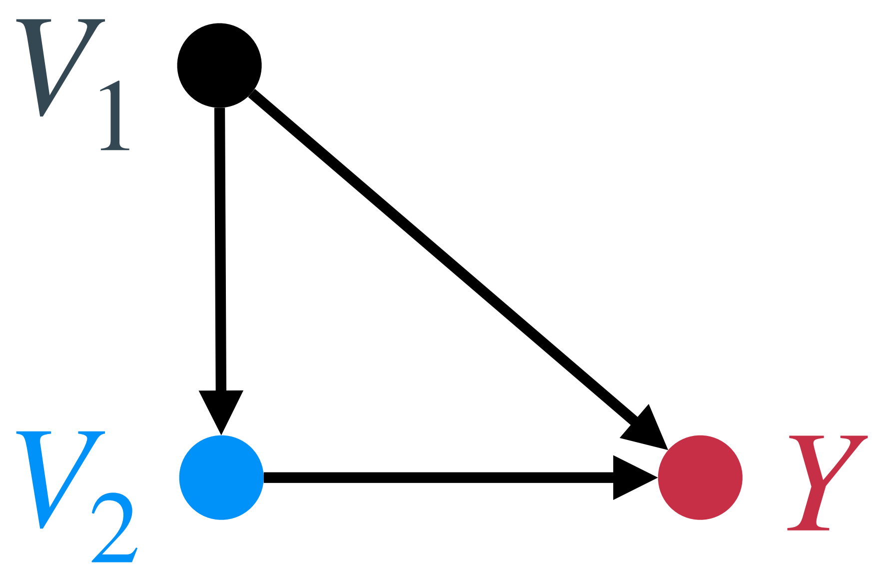
2a. Train the estimator for η using $\mathcal{D}_a$. Denote the trained estimator as $\widehat{\eta^a}$.

3a. Evaluate $g(\mathbf{V}; \hat{\eta}^a)$ using $\mathcal{D}_b$ as a test dataset; i.e., $T_a \equiv \mathbb{E}_{\mathcal{D}_b}\left[g(\mathbf{V}; \widehat{\eta^a})\right]$.

2b, 3b. Repeat (2a, 3a) with switching $\{\mathcal{D}_a, \mathcal{D}_b\}$ and $\{\hat{\eta}^a, \hat{\eta}^b\}$

Return $T \equiv (T_a + T_b)/2$

Doubly Robustness



$$\mathbb{E}[Y \;\; do(v_2)] = \mathbb{E}_P \left[\frac{I_{v_2}(V_2)}{P(V_2 \;\; V_1)} \left\{ Y - \mathbb{E}_P[Y \;\; V_1, V_2] \right\} + \mathbb{E}_P[Y \;\; v_1, V_2] \right]$$

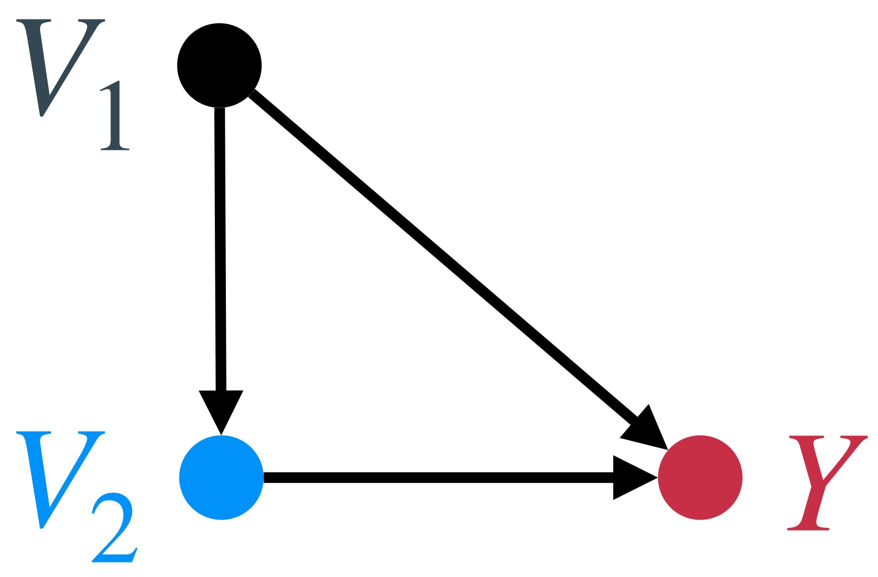
If $P(V_2 \;\; V_1)$ is misspecified so that $\tilde{P}(V_2 \;\; V_1)$ is used, we can check

$$\mathbb{E}[Y \;\; do(v_2)] = \mathbb{E}_P \left[\frac{I_{v_2}(V_2)}{\tilde{P}(V_2 \;\; V_1)} \left\{ Y - \mathbb{E}_P[Y \;\; V_1, V_2] \right\} + \mathbb{E}_P[Y \;\; v_1, V_2] \right]$$

If $\mathbb{E}[Y \;\; V_1, V_2]$ is misspecified so that $\tilde{\mathbb{E}}[Y \;\; V_1, V_2]$ is used, we can check

$$\mathbb{E}[Y \;\; do(v_2)] = \mathbb{E}_P \left[\frac{I_{v_2}(V_2)}{P(V_2 \;\; V_1)} \left\{ Y - \tilde{\mathbb{E}}_P[Y \;\; V_1, V_2] \right\} + \tilde{\mathbb{E}}_P[Y \;\; v_1, V_2] \right]$$

Debiasedness



$$\mathbb{E}[Y \mid do(v_2)] = \mathbb{E}_P\left[\frac{I_{v_2}(V_2)}{P(V_2 \mid V_1)}\left\{Y - \mathbb{E}_P[Y \mid V_1, V_2]\right\} + \mathbb{E}_P[Y \mid v_1, V_2]\right]$$

One can show that **the error** b/w DML estimator T vs. $\nu_{do}(\{2\}) = \mathbb{E}[Y \mid do(v_2)]$ is

$$T - \mathbb{E}[Y \mid do(v_2)] = O_P(N^{-1/2}) + \| P(V_2 \mid V_1) - \widehat{P}(V_2 \mid V_1) \| \cdot \| \mathbb{E}[Y \mid V_1, V_2] - \widehat{\mathbb{E}}[Y \mid V_1, V_2] \|$$

💡 Even if $\widehat{P}(V_2 \mid V_1)$ and $\widehat{\mathbb{E}}[Y \mid V_1, V_2]$ converges slowly (say $O_P(N^{-1/4})$), DML estimator converges ***doubly*** faster at a rate $O_P(N^{-1/2})$.

do-DML-Shapley

$$\phi_{v_i} = (1/n) \sum_{S \subseteq [n] \setminus \{i\}} \binom{n-1}{S}^{-1} \mathbb{E}[Y \;\; do(\mathbf{v}_S, v_i)] - \mathbb{E}[Y \;\; do(\mathbf{v}_S)]\},$$

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

do-DML-Shapley

$$\widehat{\phi}_{V_i}(T) = \frac{1}{M} \sum_{m=1}^{M} \left\{ T(v_i, \mathrm{pre}_{\pi_{(m)}}(v_i)) - T(\mathrm{pre}_{\pi_{(m)}}(v_i)) \right\}$$

Property for do-DML-Shapley

do-DML-Shapley

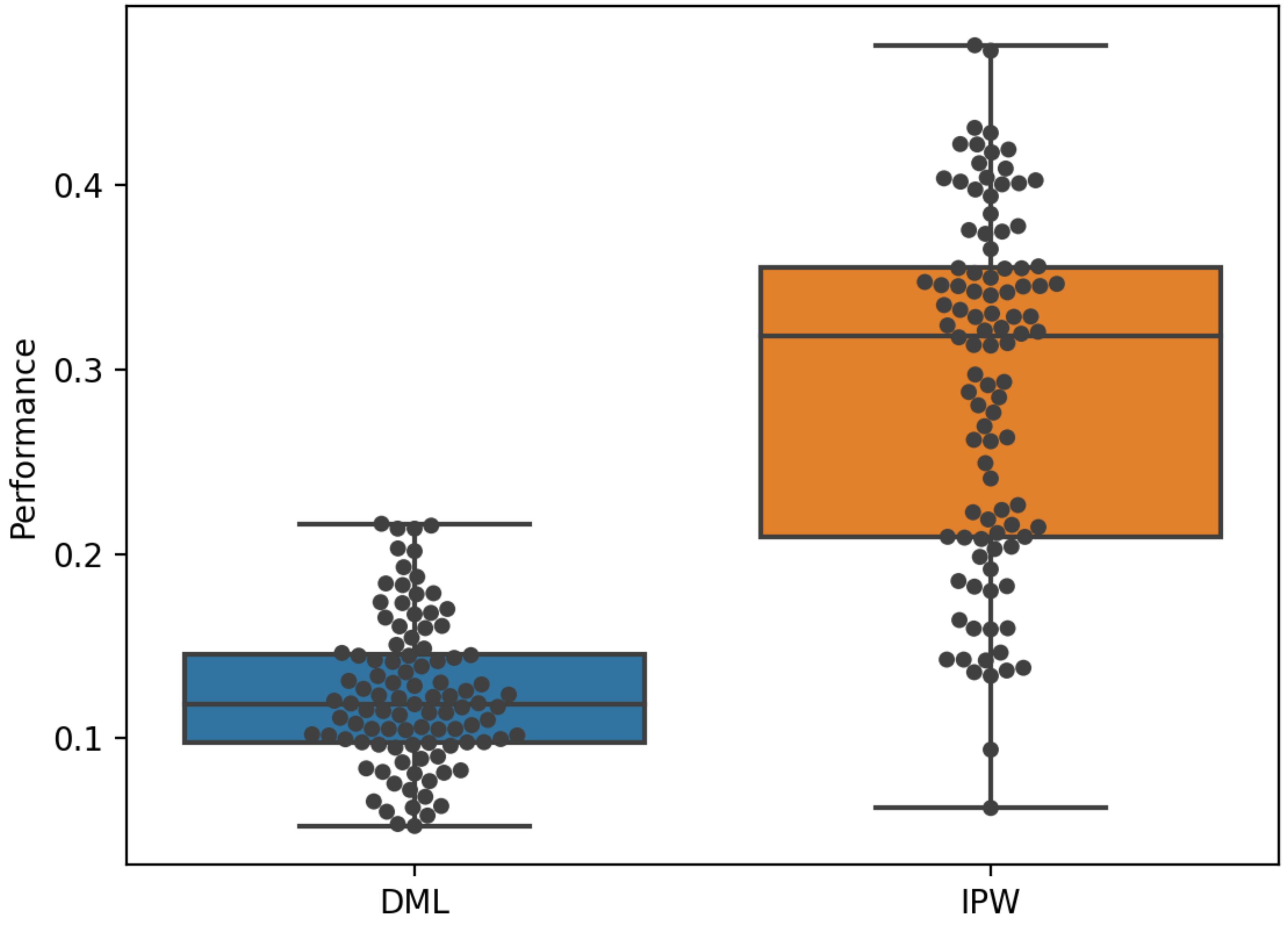
$$\widehat{\phi}_{V_i}(T) = \frac{1}{M}\sum_{m=1}^{M}\left\{T(v_i, \text{pre}_{\pi_{(m)}}(v_i)) - T(\text{pre}_{\pi_{(m)}}(v_i))\right\}$$

Robustness of do-DML-Shapley

do-DML-Shapley $\widehat{\phi}_{V_i}(T)$ achieves Doubly Robustness (DR) and Debiasedness (DB) with respect to nuisance functionals $\widehat{\mathbb{E}}[Y \;\; \mathbf{v}_S, \mathbf{v}_{\overline{S}}]$ and $\{\widehat{P}(V_k \;\; pa(V_k))\}_{V_k \in \mathbf{V}}$.

Simulation: Robustness

We compare the do-DML-Shapley with do-IPW-Shapley, estimates for do-Shapely where causal coalition $\nu_{do}(S)$ is estimated using the IPW.



For 100 random samples $(\mathbf{x}, f(\mathbf{x}))$ from $\mathscr{D}$,

For each features,

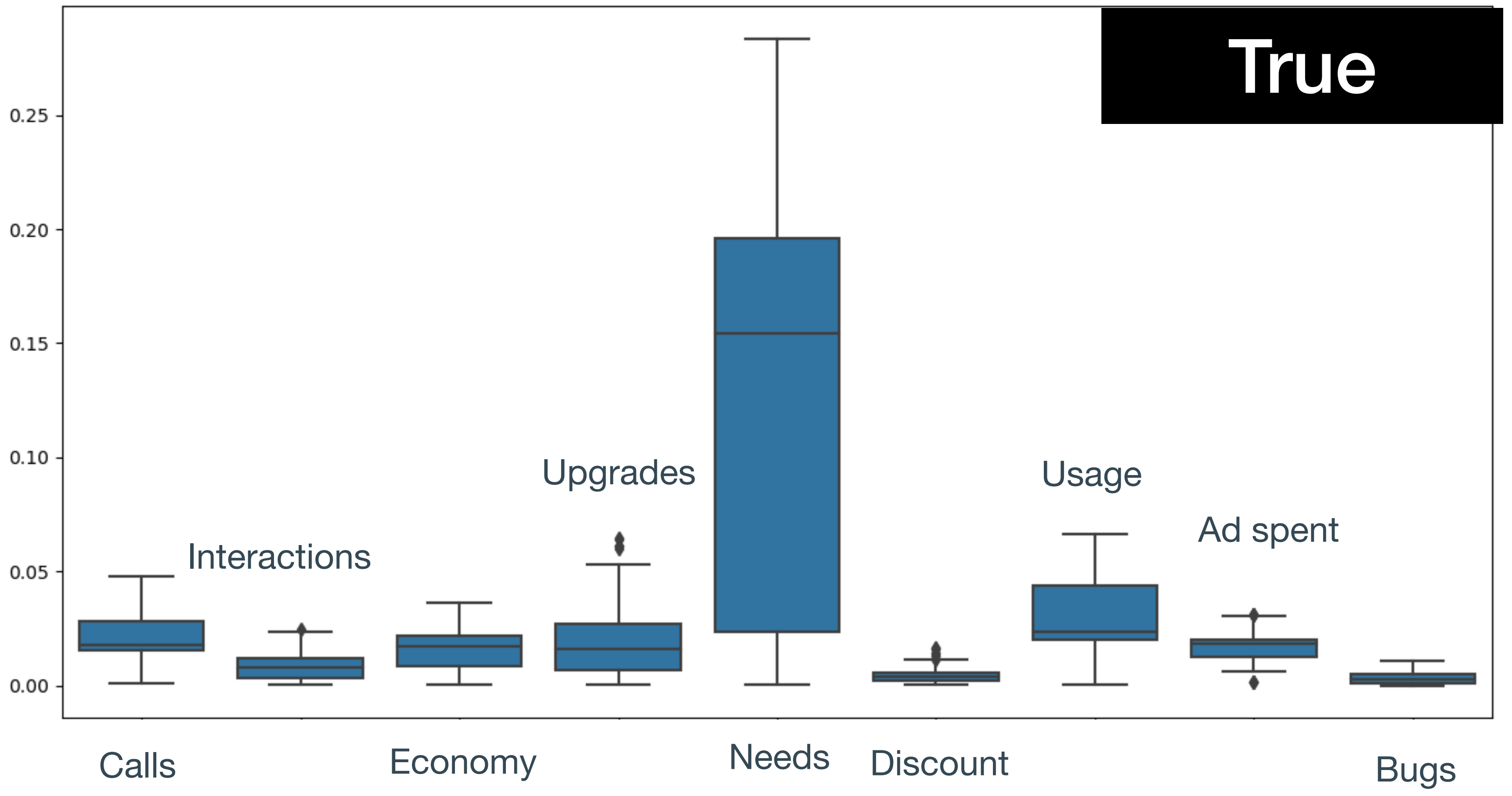
compute do-DML (or IPW)-Shapley values;

compute the gap with the do-True-Shapley;

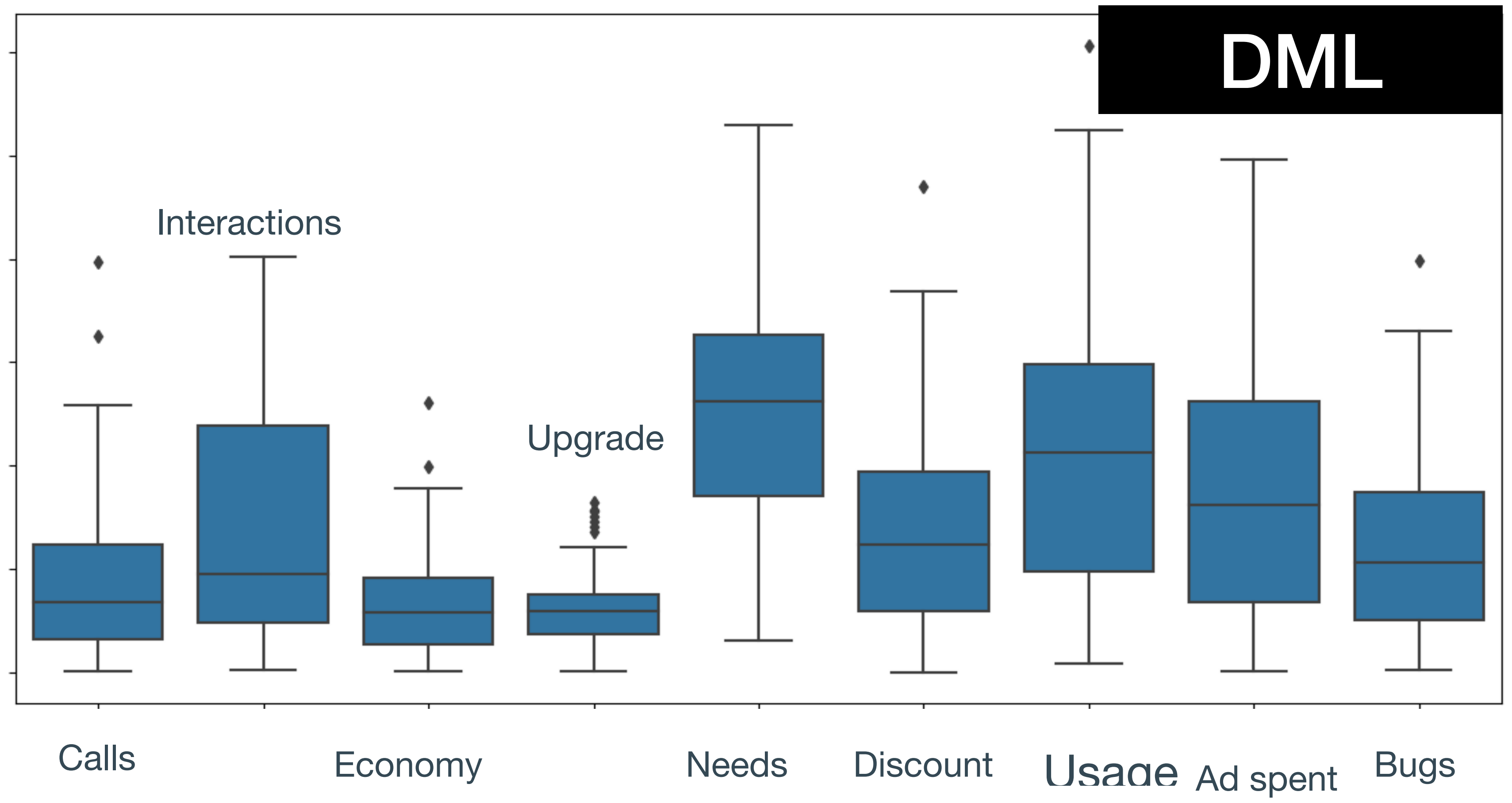
Take an average of this gap over features.

do-DML-Shapely provides accurate & robust estimates!

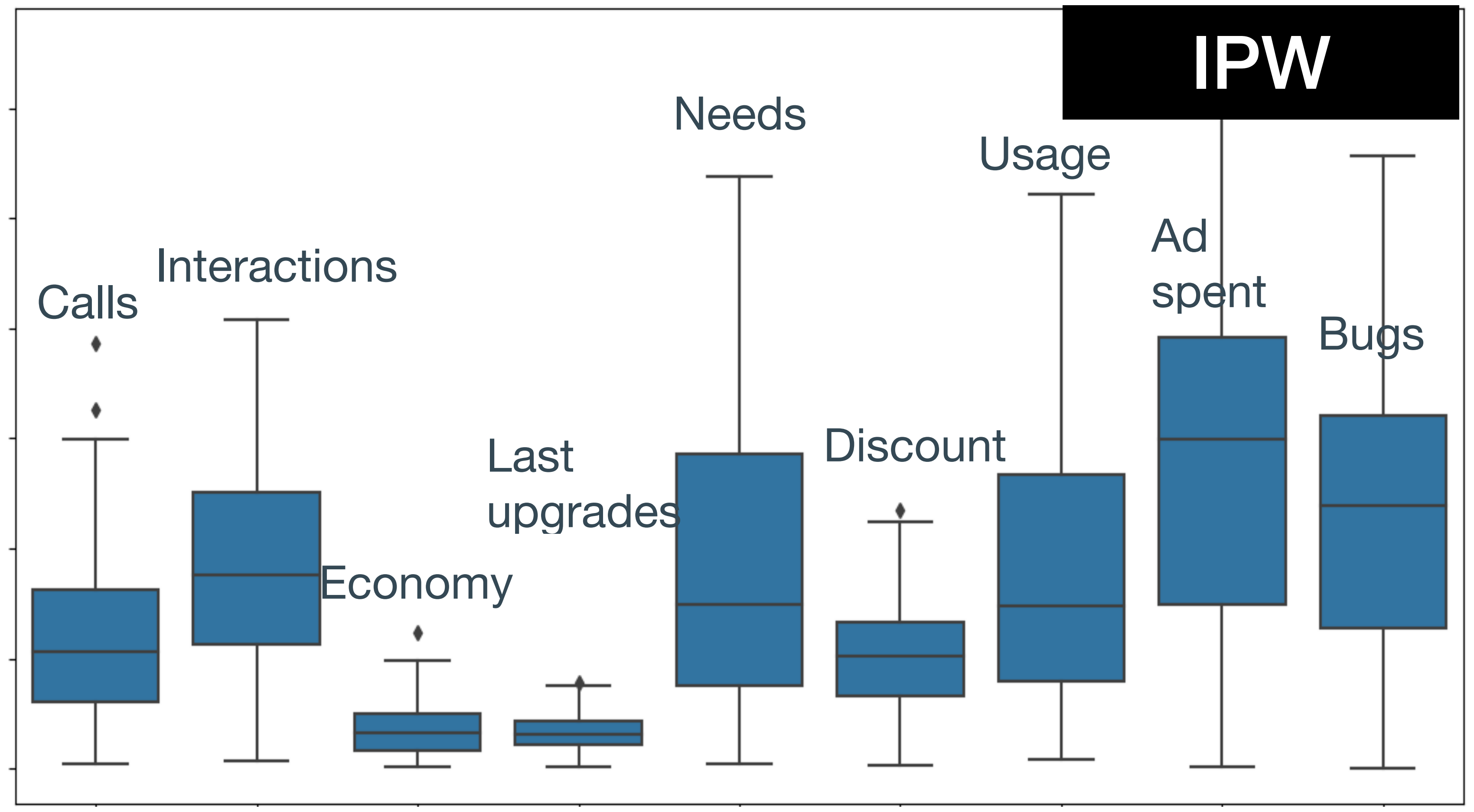
Simulation: Better Interpretability



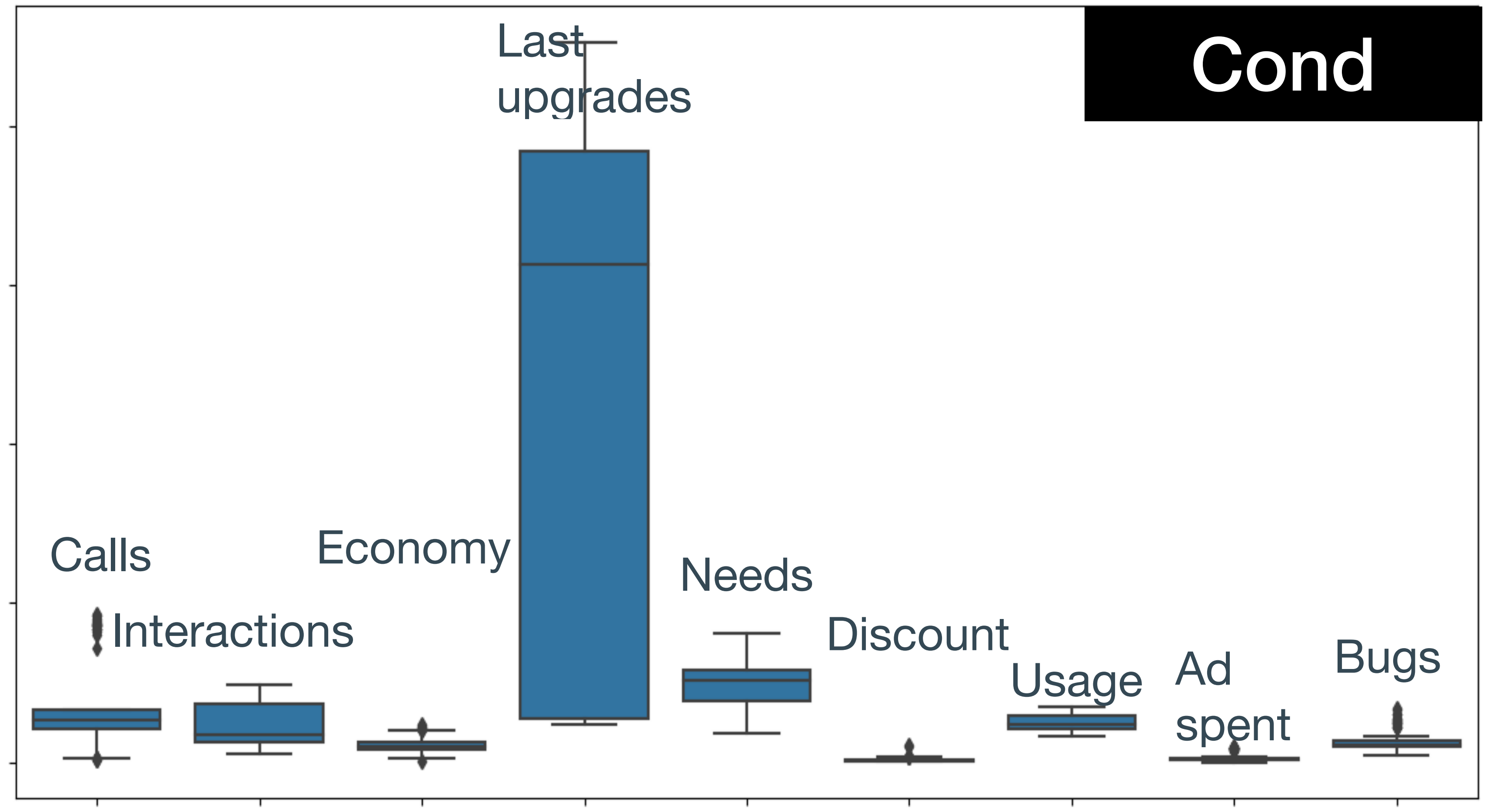
Top 3 = {Needs, Usage, Calls}



Top 3 = {Needs, Usage, Ad}



Top 3 = {Ad, Needs, Interaction}



Top 3 = {Upgrade, Needs, Calls}

Conclusion

We develop ***causally*** *interpretable* ***feature attribution*** *method*.

1. We *axiomatize* a causally interpretable feature attribution method, and propose do-Shapley values.

2. We provide *identifiability* condition where the do-Shapley values can be inferred from the observational data.

3. We construct a *double/debiased machine learning (DML)* based do-Shapley estimator for practical settings.

Thank you

Time for Questions