ICML | 2022

On Measuring Causal Contributions via *do*-interventions

Yonghan Jung
Purdue University

Shiva Kasiviswanathan
Amazon

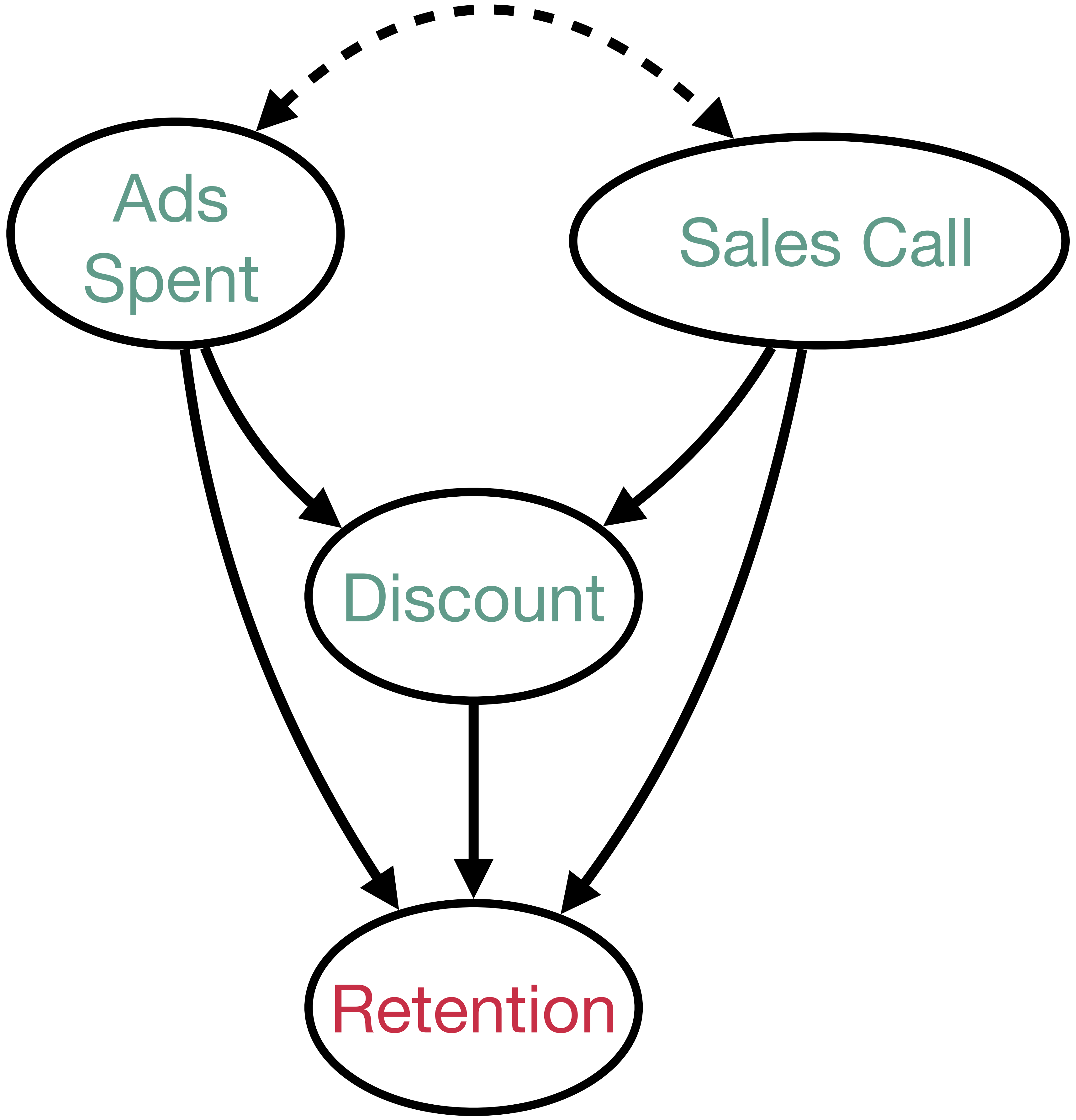
Jin Tian
Iowa State University

Dominik Janzing
Amazon

Patrick Blöbaum
Amazon

Elias Bareinboim
Columbia University

Motivational Example



This causal diagram depicts the data generating process of customers' retention decisions for a video streaming service.

Given the hypothetical sales policy (e.g., higher Ads spent and sales calls but lower discounts), suppose a royal customer, Alice, is expected to discontinue the service.

What are the *contributions* of each input to the customer's decision (discontinuation)?

Task: Causality-based Feature Attribution

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

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

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

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

Feature
Attribution

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

Application to ML Interpretation

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.

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 $f(\mathbf{v})$ based on causality.

Results 1. Feature Attribution based on Desirable Properties (Axiom)

1. We provide desirable properties that the causality-based feature attribution method, taking account of causality, should satisfy [Axiom 1]. We propose the do-Shapley value ϕ_{v_i} as a unique attribution method that satisfies the properties [Theorem 1].

$$\phi_{v_i} = \frac{1}{n} \sum_{S \subseteq [n] \setminus i} \binom{n-1}{|S|}^{-1} \{\mathbb{E}[Y | do(\mathbf{v}_{S \cup i}) - \mathbb{E}[Y | do(\mathbf{v}_S)\}.$$

Results 2. Identification of do-Shapley

$$\phi_{v_i} = \frac{1}{n} \sum_{S \subseteq [n] \setminus i} \binom{n-1}{|S|}^{-1} \{\mathbb{E}[Y \mid do(\mathbf{v}_{S \cup i}) - \mathbb{E}[Y \mid do(\mathbf{v}_S)\}.$$

2. To estimate ϕ_{v_i} from samples $D \sim P(\mathbf{V})$, all $\mathbb{E}[Y \mid do(\mathbf{v}_S)]$ must be identified (i.e., expressed as a function of P). We provide graphical criteria where identifying $\mathbb{E}[Y \mid do(\mathbf{v}_S)]$ can be done in polynomial time [Theorem 2, Corollary {1,2}]

Results 3. Estimation of do-Shapley

3. We propose an estimator for the do-Shapley value, which exhibits robustness against bias. This includes the "*debiasedness*" property, which guarantees a fast convergence rate of the do-Shapley [Theorem 3].

Summary

Task: Develop the feature attribution method which measure the contribution of $v_i \in \mathbf{v}$ to the target effect $\mathbb{E}[Y \mid do(\mathbf{v})]$ based on causality.

Results

1. We developed the do-Shapley value, which is a contribution measure that uniquely satisfies certain desirable properties.

2. We provided a graphical criterion where the do-Shapley value can be expressed as a function of observational distribution in poly-time.

3. We developed an estimator, which exhibits robustness property against bias.