CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Kun Kuang (况琨)
Assistant Professor
College of Computer Science
Zhejiang University
Homepage: https://kunkuang.github.io/

Research Interests

- Causal Inference and Machine Learning
  - Machine Learning for Causal Inference
    - High Dimensional, Big Data Era
  - Causal Inference for Machine Learning
    - Interpretable prediction, Stable Learning
  - Causally Interpretable AI + X (司法，医疗，教育)

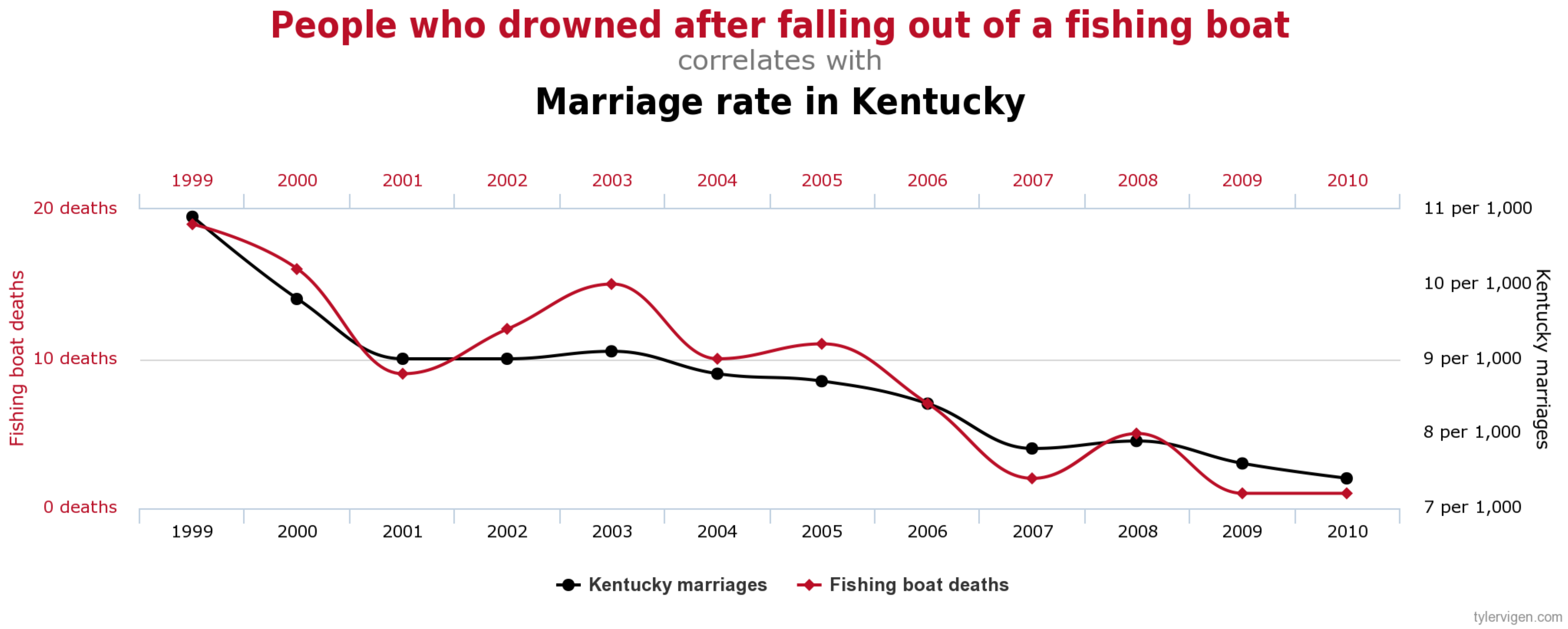
Causal Inference: Cause and Effect

- Cause: The REASON why something happened
- Effect: The RESULT of what happened

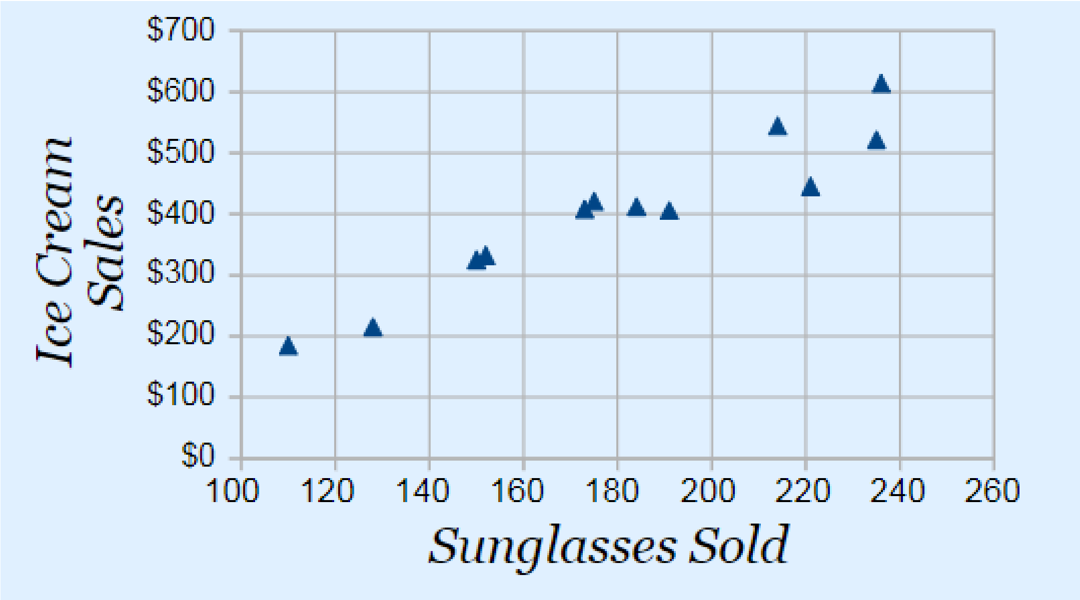
- Questions of cause and effect:
  - Medicine: drug trials, effect of a drug
  - Social science: effect of a policy
  - Marketing: effect of a marketing strategy
  - …
- **What is causality?**

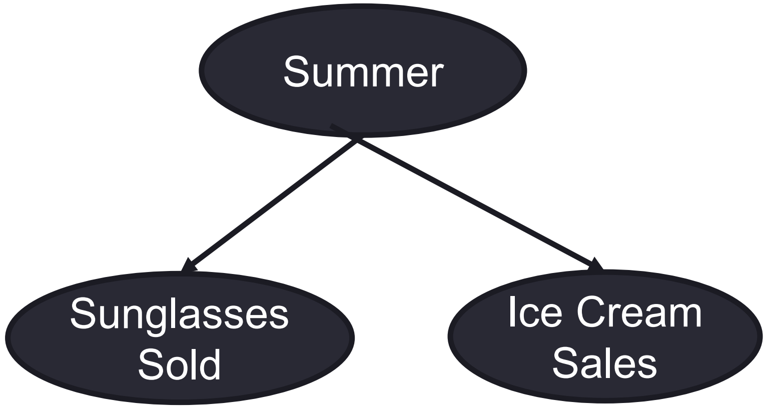
Correlation v.s. Causality: Explainability

- Correlation is not explainable



Correlation v.s. Causality: Explainability

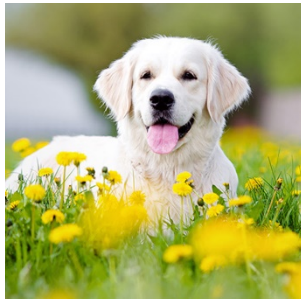


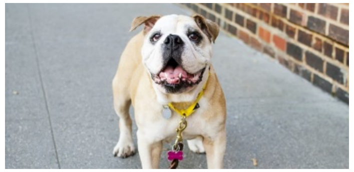


Spurious Correlation !

Correlation does not imply causation!

Correlation v.s. Causality: Stability



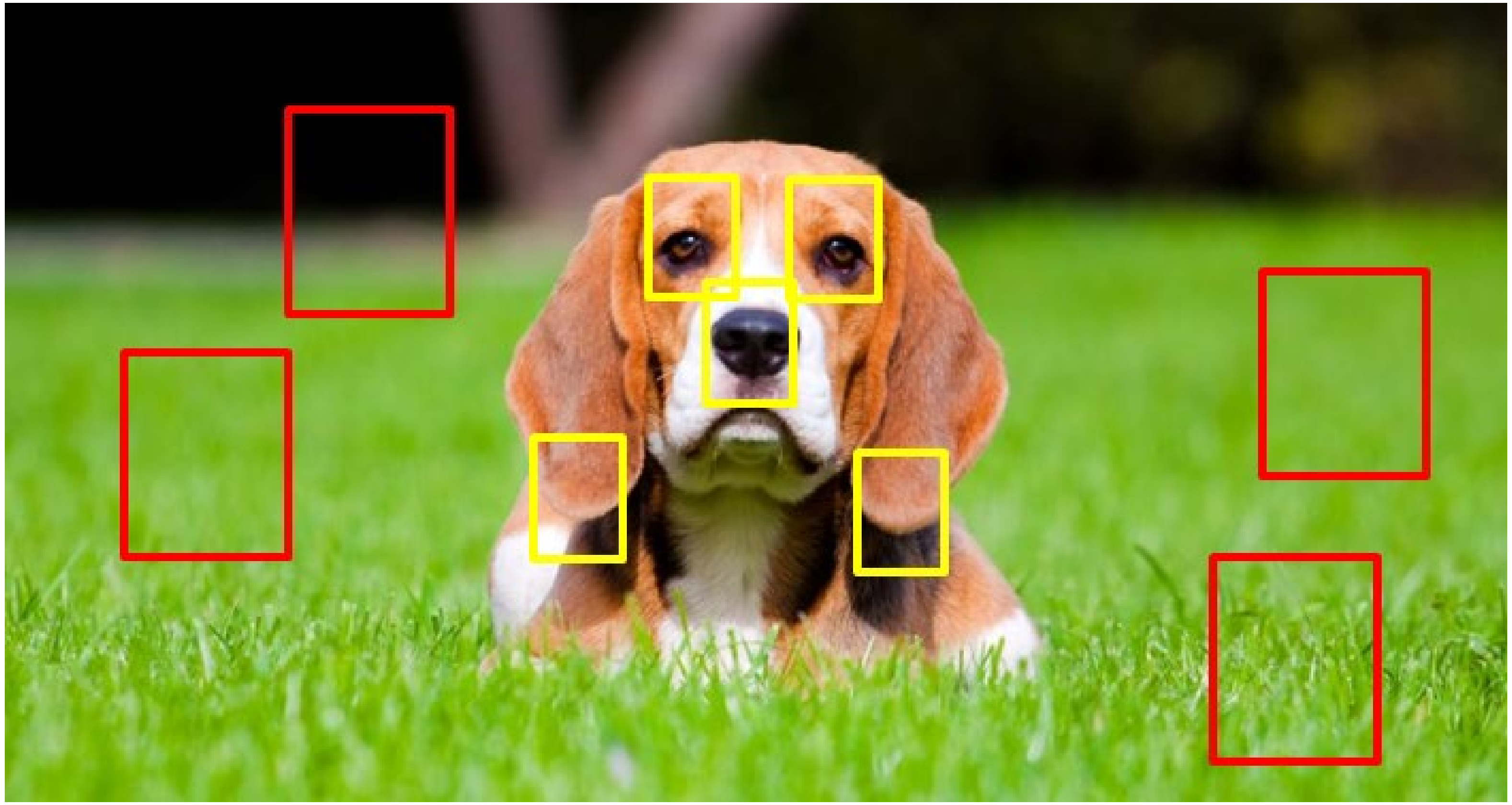

Yes

Maybe

No

Correlation v.s. Causality: Stability

Correlation v.s. Causality



Correlation v.s. Causality: Actionability

- Does predictive models guide decision making?
- System changes algorithm from A to B at some point.
- Is the new algorithm B better?
- Say algorithm that provides promotion or discount link to a different customers

Algorithm A

Algorithm B

Correlation v.s. Causality: Actionability

- Measure success rate (SR)

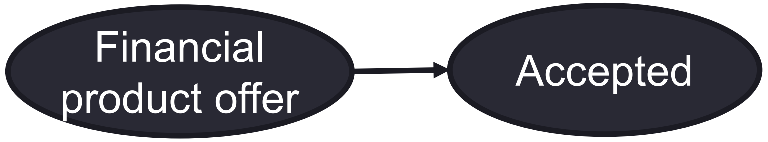
Old Algorithm (A)	New Algorithm (B)
50/1000 **(5%)**	54/1000 **(5.4%)**

New algorithm increases overall success rate, so it is better?

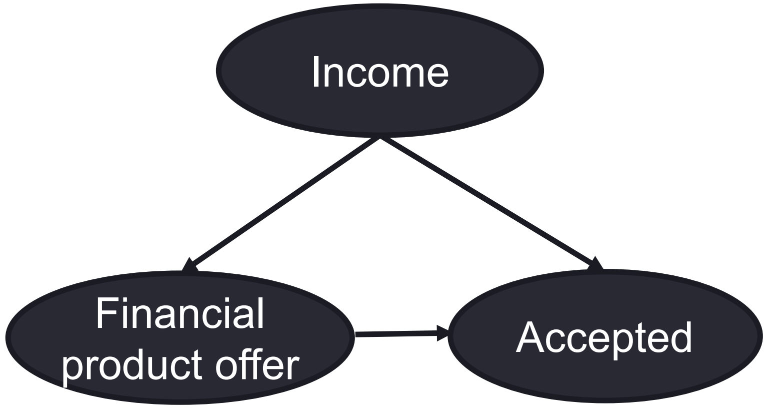
	Old Algorithm (A)	New Algorithm (B)
Low-income Users	10/400 **(2.5%)**	4/200 **(2%)**
High-income Users	40/600 **(6.6%)**	50/800 **(6.2%)**
Overall	50/1000 **(5%)**	54/1000 **(5.4%)**

Which is better?

Correlation v.s. Causality: Actionability



Higher success rate due to algorithm



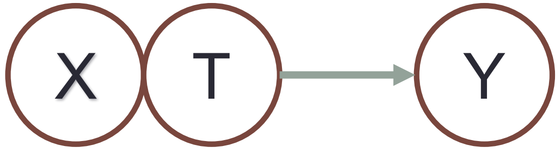
Higher success rate due to confounding bias

Decision making is a counterfactual problem, not a predictive problem!

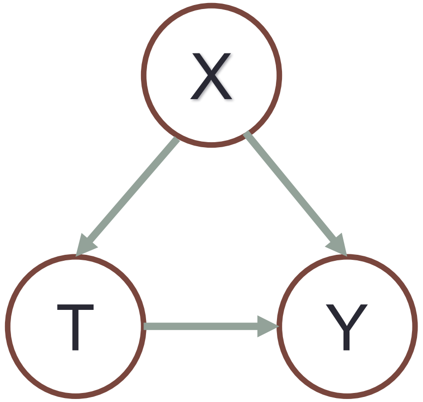
Correlation v.s. Causality: Fairness

Correlation v.s. Causality: Fairness

Correlation Framework



Causal Framework



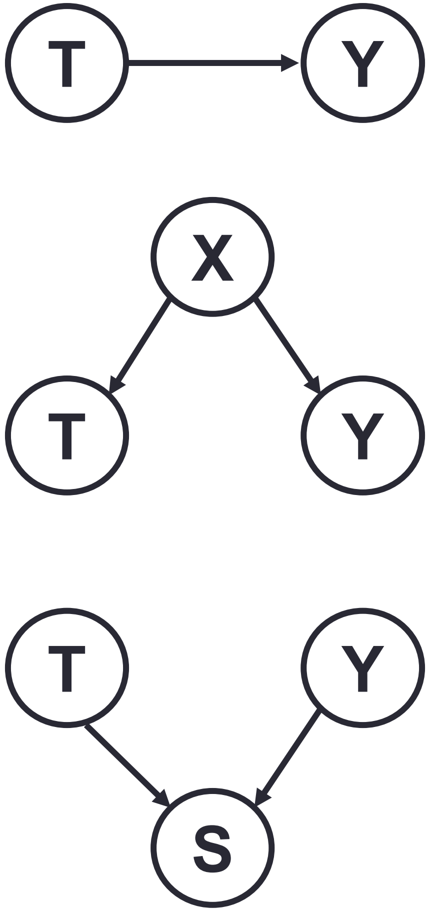
T： skin color
X： income
Y： crime rate

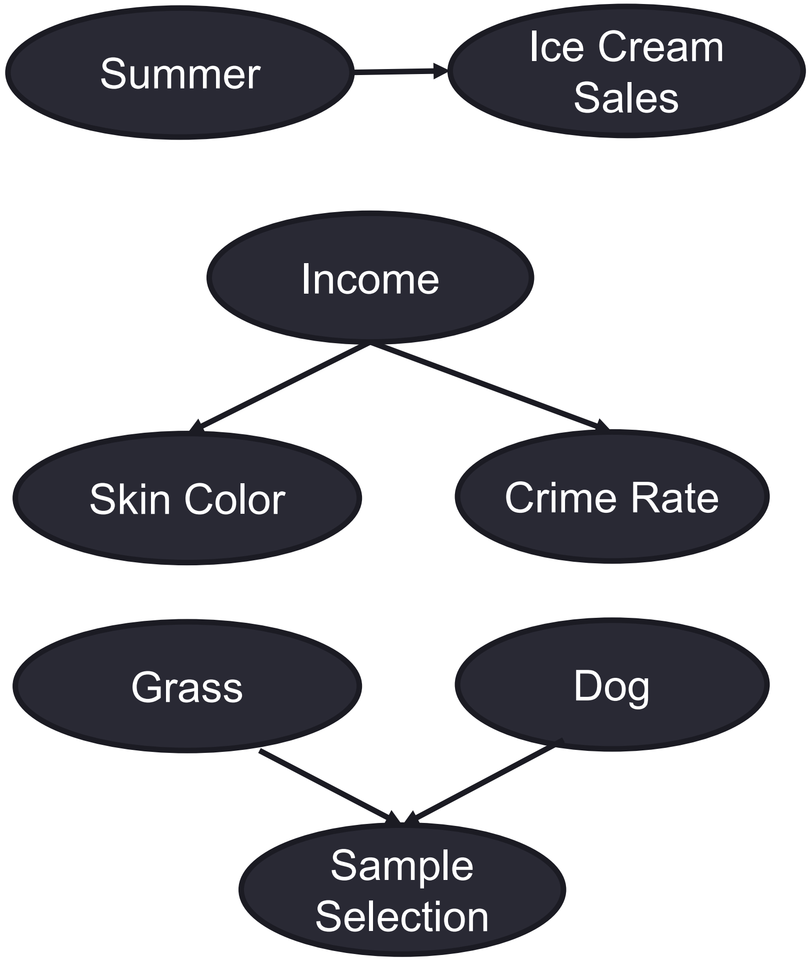
income—crime rate: Strong correlation
skin color—crime rate: Strong correlation

income—crime rate: Strong causation
skin color—crime rate: Weak causation

Correlation V.S. Causation

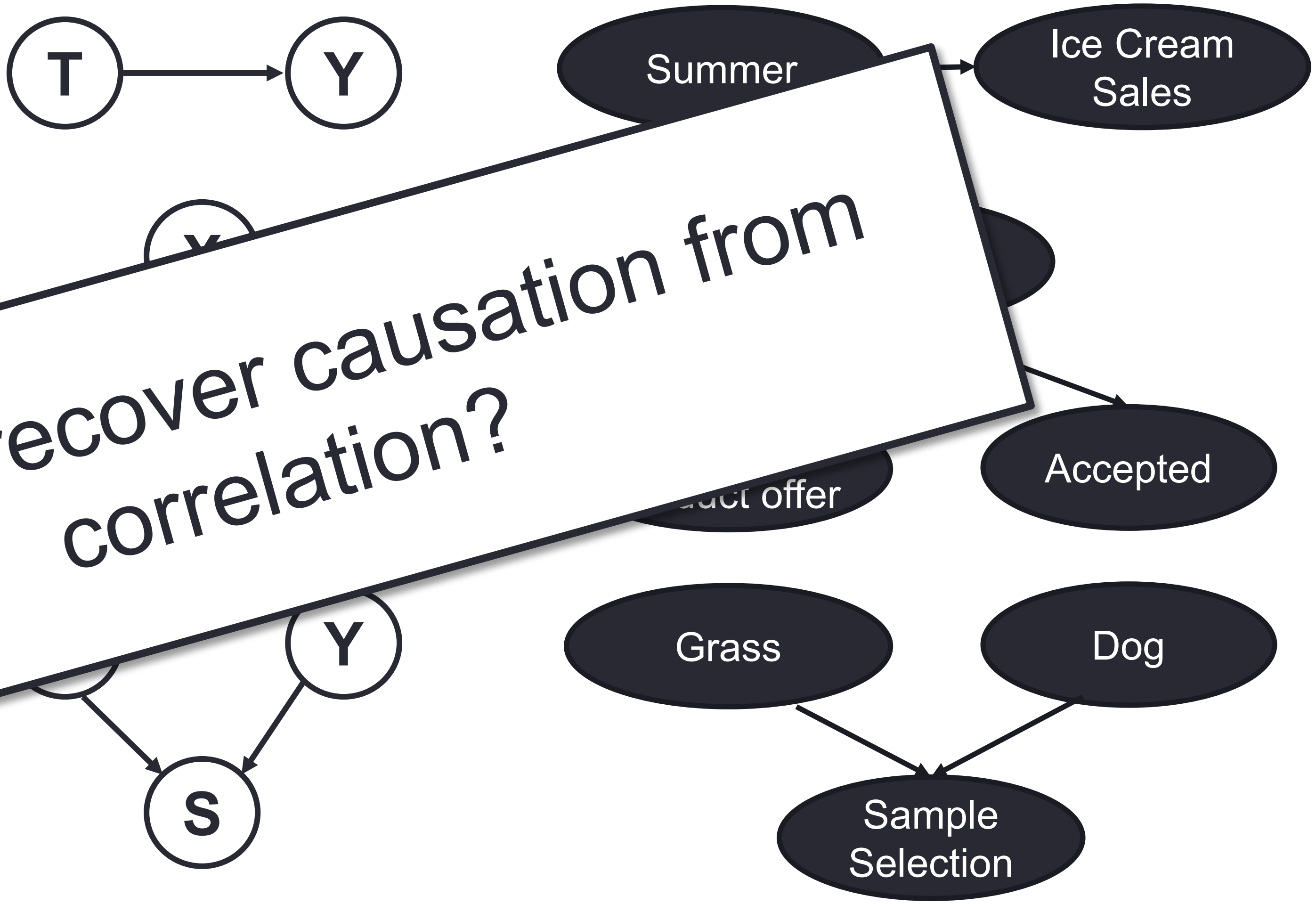
- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and Robust
  - Confounding
    - Ignoring X
    - Spurious Correlation
  - Sample Selection
    - Conditional on S
    - Spurious Correlation





Correlation V.S. Causation

- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and Robust
  - Confounding
    - Ignoring X
    - Spurious
  - Sample Se
    - Conditiona
    - Spurious Correlation

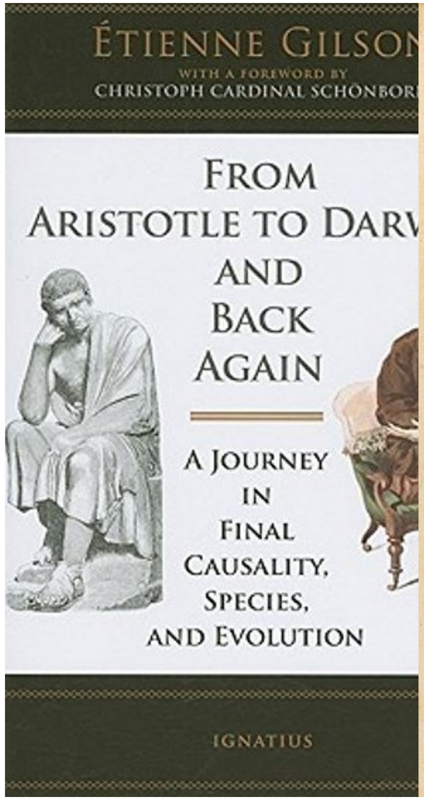


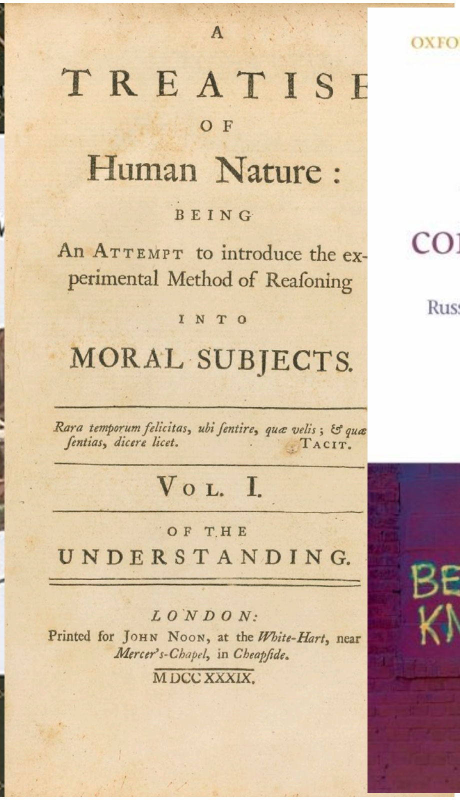
Why should we care about causality?

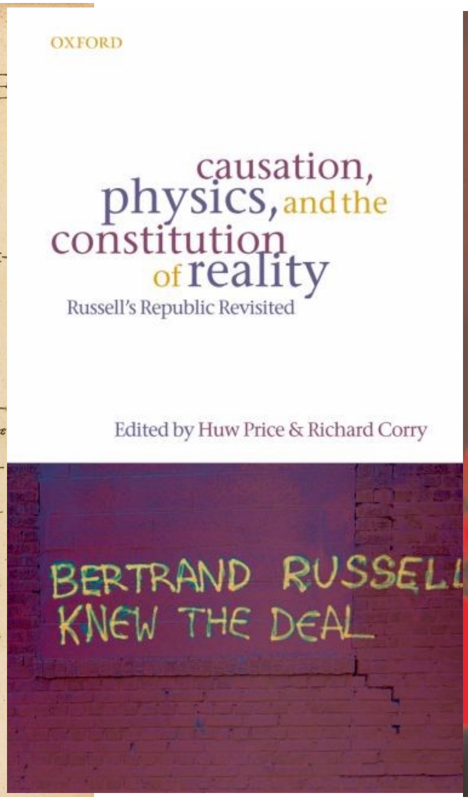
- Recover causation for interpretability
- Help to guide decision making
- Make stable and robust prediction in the future
- Prevent algorithmic bias

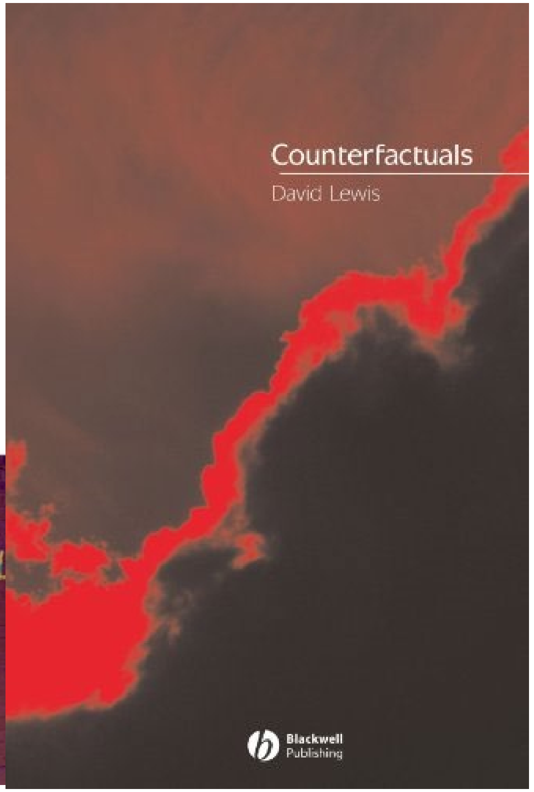
What is causality?

- A big scholarly debate, from Aristotle to Russell









A practical definition

Definition: T causes Y if and only if

changing T leads to a change in Y,

keep everything else constant.

Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Two key points: changing T, everything else constant

**Interventionist* definition [http://plato.stanford.edu/entries/causation-mani/]

Causal Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Average Causal Effect of Treatment (ATE):

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

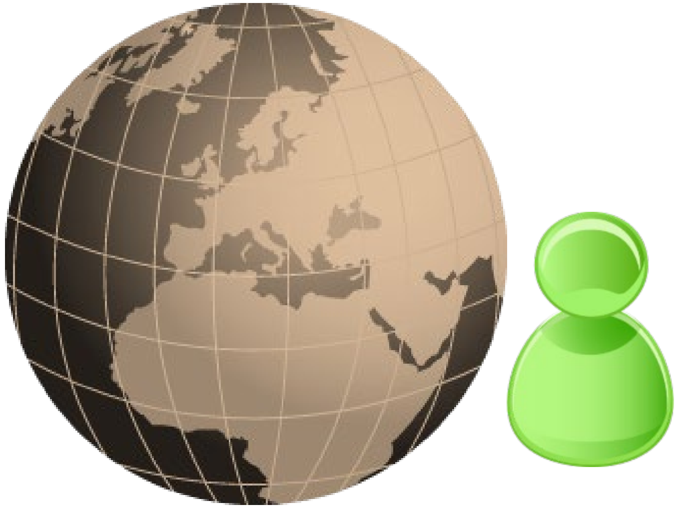
- Counterfactual Problem:

$$Y(T = 1) \quad \text{or} \quad Y(T = 0)$$

Ideal Solution: Counterfactual World

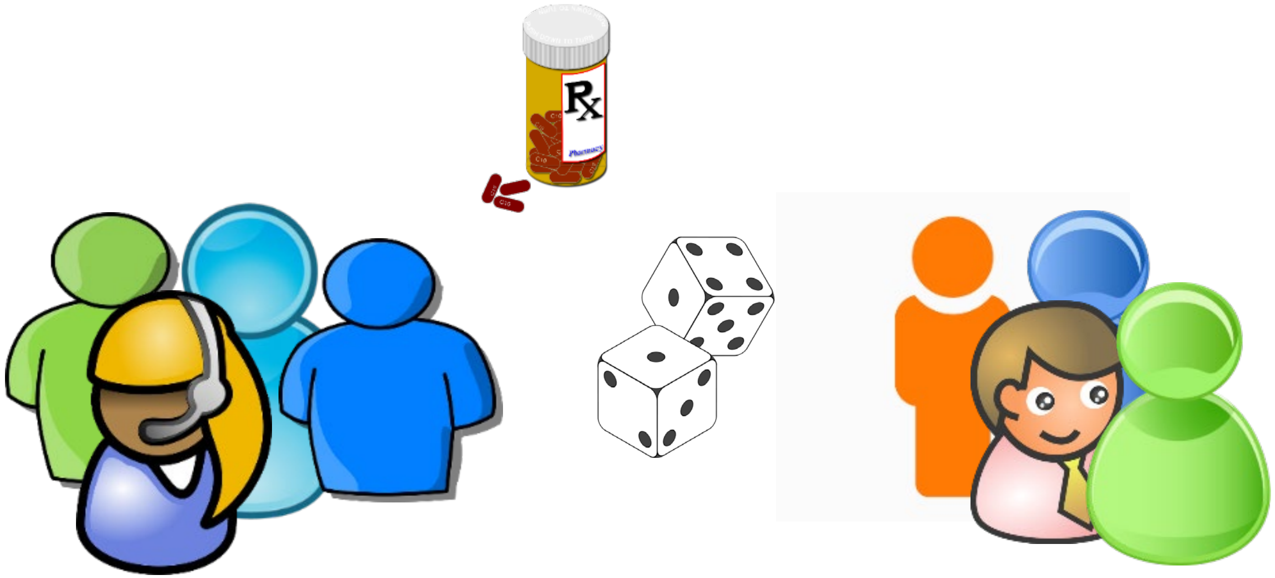
- Reason about a world that does not exist
- Everything is the same on real and counterfactual worlds, but the treatment

$Y(T = 1)$



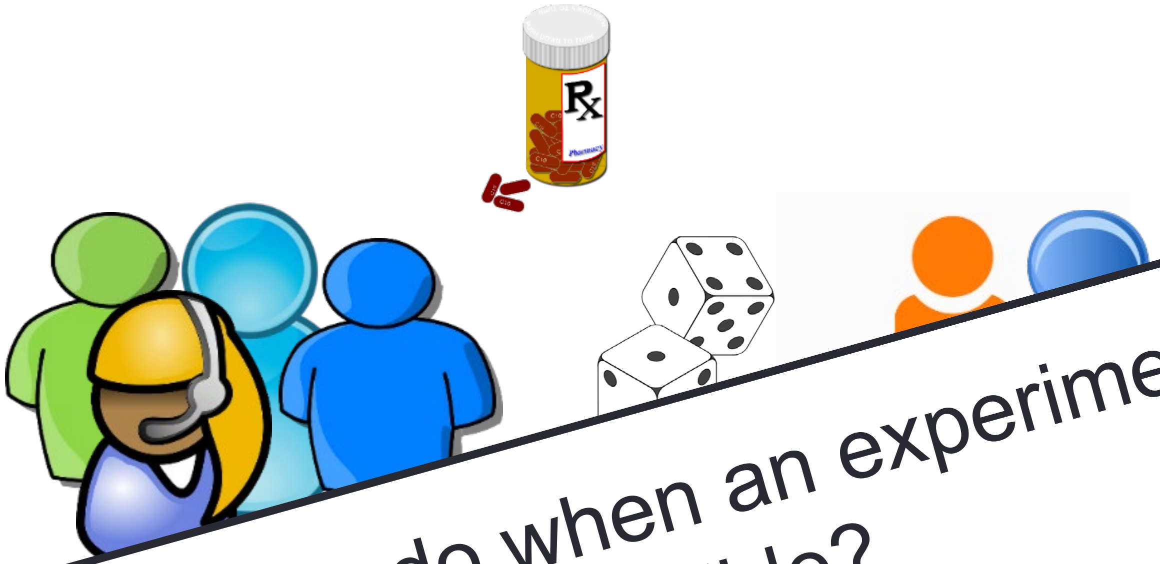
$Y(T = 0)$

Randomized Experiments are the “Gold Standard”



- Drawbacks of randomized experiments:
  - Cost
  - Unethical

Randomized Experiments are the "Gold Standard"



- Drawbac
  - Cost
  - Unethical

What can we do when an experiment is not possible?

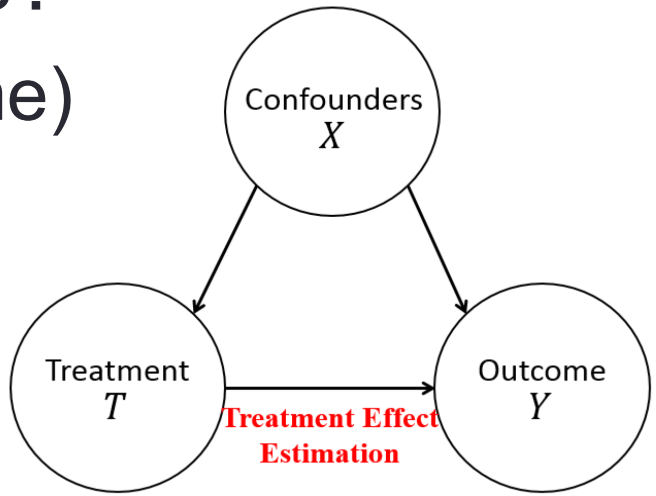
Observational Studies!

Causal Inference with Observational Data

- Counterfactual Problem:

$$Y(T = 1) \quad \text{or} \quad Y(T = 0)$$

- Can we estimate ATE by directly comparing the average outcome between treated and control groups?
  - Yes with randomized experiments (X are the same)
  - No with observational data (X might be different)
- Two key points:
  - Changing T (T=1 and T=0)
  - Keeping everything else (Confounder X) constant

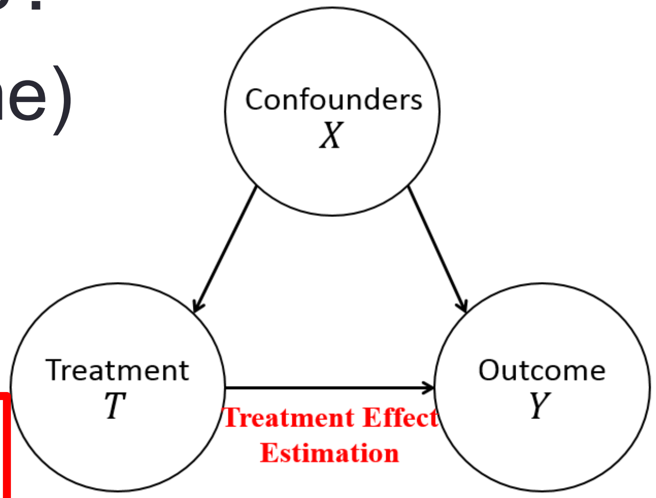


Causal Inference with Observational Data

- Counterfactual Problem:

$$Y(T=1) \quad \text{or} \quad Y(T=0)$$

- Can we estimate ATE by directly comparing the average outcome between treated and control groups?
  - Yes with randomized experiments (X are the same)
  - No with observational data (X might be different)
- Two key points:



Balancing Confounders' Distribution

Methods for Causal Inference

- **Matching**
- **Propensity Score Based Methods**
  - Propensity Score Matching
  - Inverse of Propensity Weighting (IPW)
  - Doubly Robust
  - Data-Driven Variable Decomposition (D^2VD)
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing
  - Differentiated Confounder Balancing

Assumptions of Causal Inference

- **A1: Stable Unit Treatment Value (SUTV):** The effect of treatment on a unit is independent of the treatment assignment of other units

$$P\left(Y_i \middle| T_i, T_j, X_i\right) = P(Y_i | T_i, X_i)$$

- **A2: Unconfounderness:** The distribution of treatment is independent of potential outcome when given the observed variables

$$T \perp \left(Y(0), Y(1)\right) | \, X$$

No unmeasured confounders

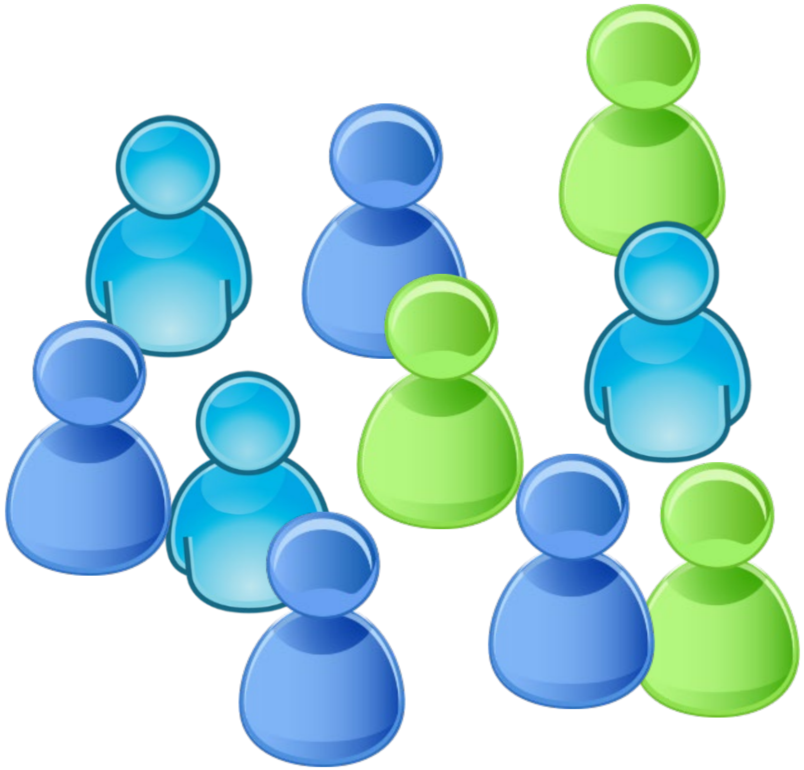
- **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

$$0 < P(T = 1 | X = x) < 1$$

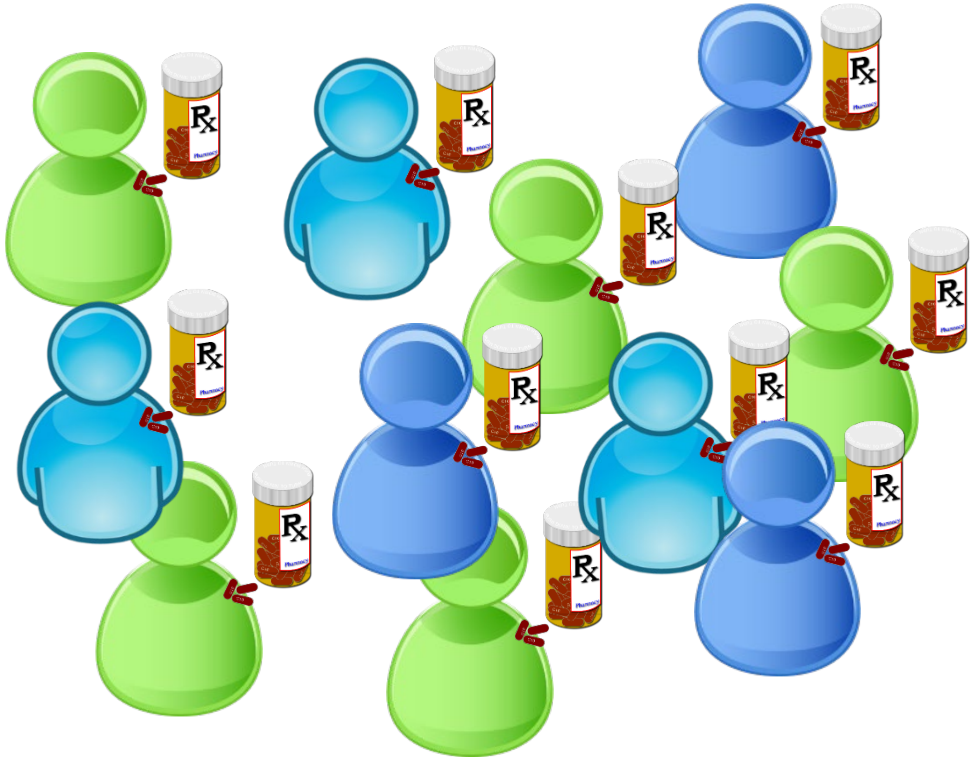
Methods for Causal Inference

- **Matching**
- **Propensity Score Based Methods**
  - Propensity Score Matching
  - Inverse of Propensity Weighting (IPW)
  - Doubly Robust
  - Data-Driven Variable Decomposition (D^2VD)
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing
  - Differentiated Confounder Balancing

Matching

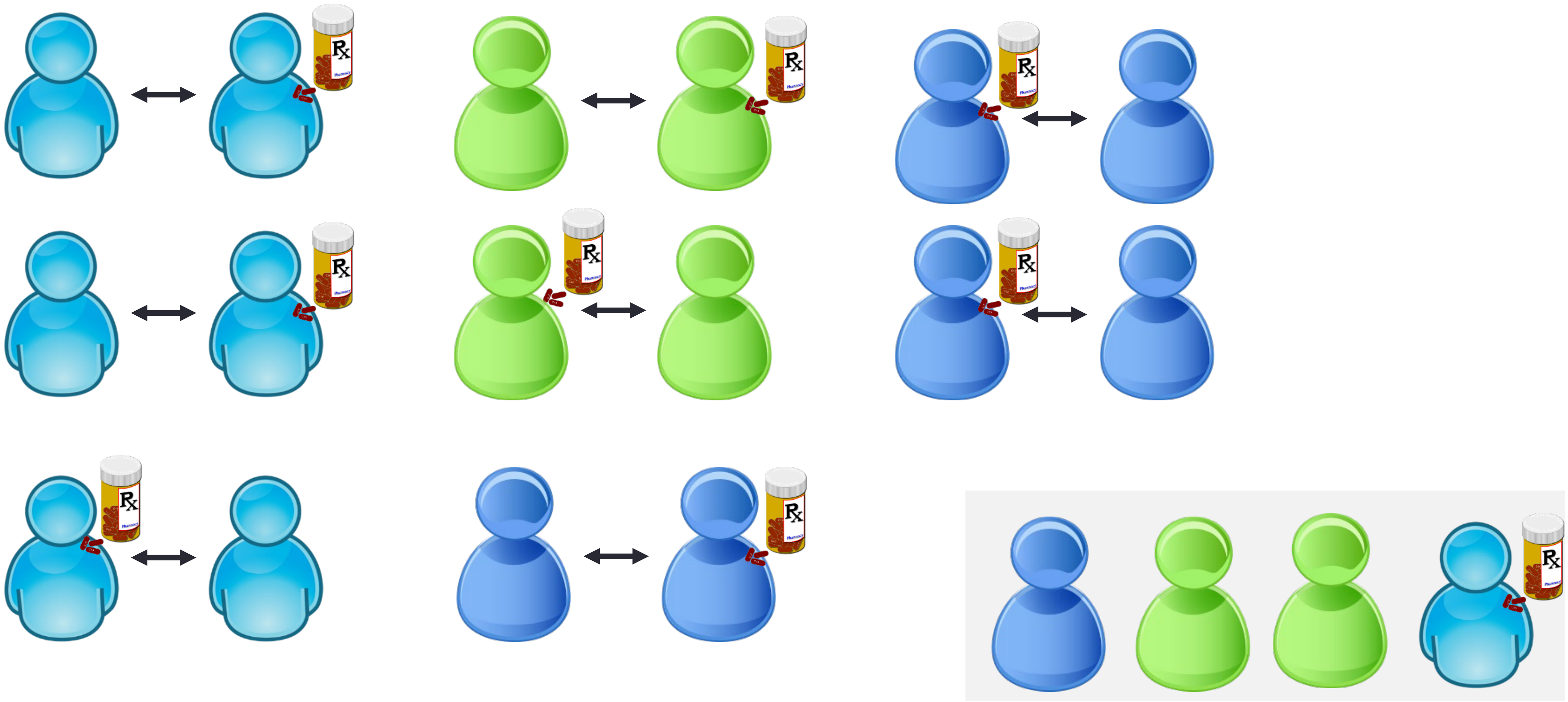


$T = 0$



$T = 1$

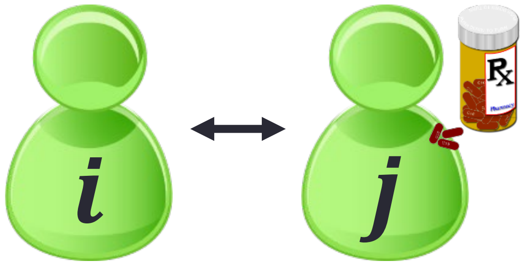
Matching



Matching

- Identify pairs of treated (T=1) and control (T=0) units whose confounders X are similar or even identical to each other

$$Distance(X_i, X_j) \le \epsilon$$



- Paired units provide the everything else (Confounders) approximate constant
- Estimating average causal effect by comparing average outcome in the paired dataset

- Smaller ϵ: less bias, but higher variance

Methods for Causal Inference

- **Matching**
- **Propensity Score Based Methods**
  - Propensity Score Matching
  - Inverse of Propensity Weighting (IPW)
  - Doubly Robust
  - Data-Driven Variable Decomposition (D^2VD)
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing
  - Differentiated Confounder Balancing

Propensity Score Based Methods

- Propensity score $e(X)$ is the probability of a unit to be treated

$$e(X) = P(T = 1|X)$$

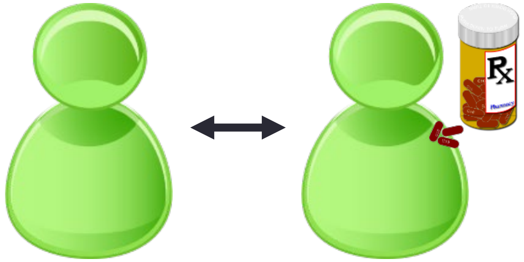
- Then, Rubin shows that the propensity score is sufficient to control or summarized the information of confounders

$$T \perp\!\!\!\perp X \mid e(X) \quad \Rightarrow \quad T \perp\!\!\!\perp (Y(1), Y(0)) \mid e(X)$$

- Propensity score are rarely observed, need to be estimated

Propensity Score Matching

- Estimating propensity score: $\hat{e}(X) = P(T = 1|X)$
  - **Supervised learning**: predicting a known label T based on observed covariates X.
  - Conventionally, use logistic regression



$$Distance(X_i, X_j) \leq \epsilon$$

- Matching pairs by distance between propensity score:

$$Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$$

- High dimensional challenge: transferred from matching to PS estimation

Inverse of Propensity Weighting (IPW)

- Estimating ATE by IPW [1]:

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1 - T_i)Y_i}{1 - \hat{e}(X_i)}$$

  - Interpretation: IPW creates a pseudo-population where the confounders are the same between treated and control groups.

- Why does this work? Consider $\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)}$

Inverse of Propensity Weighting (IPW)

- If: $\hat{e}(X) = e(X)$, the *true propensity score*

$$\boxed{E\left\{\frac{TY}{e(X)}\right\} = E\left\{\frac{TY_1}{e(X)}\right\}} = E\left[E\left\{\frac{TY_1}{e(X)}|Y_1, X\right\}\right]$$

$$= \boxed{E\left\{\frac{Y_1}{e(X)}E(T|Y_1, X)\right\} = E\left\{\frac{Y_1}{e(X)}\boxed{E(T|X)}\right\}}$$

$$= E\left\{\frac{Y_1}{e(X)}e(X)\right\} = E\left(Y_1\right)$$

(1) $\boldsymbol{Y = T * Y_1 + (1 - T) * Y_0}$

(2) $\boldsymbol{T \perp (Y_1, Y_0) \mid X}$

(3) $\boldsymbol{e(X) = E(T|X)}$

- Similarly: $E\left\{\frac{(1-T)Y}{1-e(X)}\right\} = E(Y_0)$

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

Inverse of Propensity Weighting (IPW)

- If: $\hat{e}(X) = e(X)$, the *true propensity score,* the IPW estimator is *unbiased*

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(X_i)} = E(Y_1 - Y_0)$$

- Wildly used in many applications

- But requires the propensity score model is correct
- High variance when e is close to 0 or 1

Doubly Robust

$$m_0 = E(Y|T=0,X)$$
$$m_1 = E(Y|T=1,X)$$

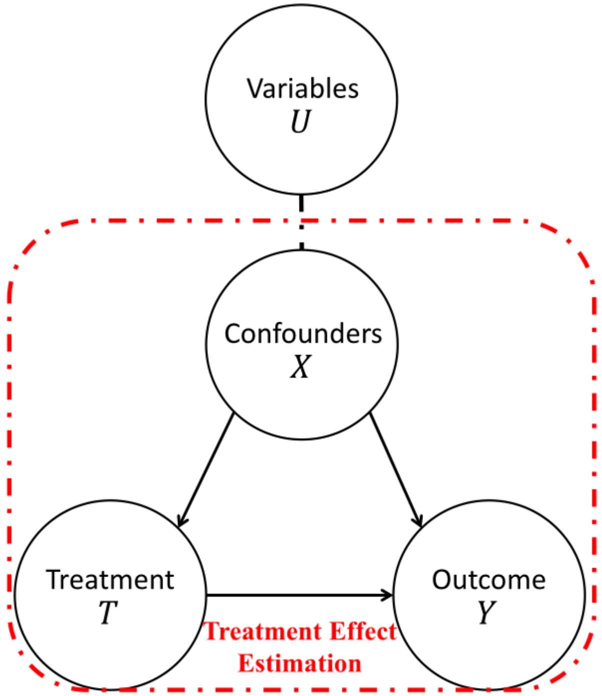
- Estimating ATE with Doubly Robust estimator:

$$\begin{aligned} ATE_{DR} \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{\{T_i - \hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right] \\ &- \quad \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)\,Y_i}{1-\hat{e}(X_i)} + \frac{\{T_i - \hat{e}(X_i)\}}{1-\hat{e}(X_i)}\hat{m}_0(X_i)\right] \end{aligned}$$

  - *Unbiased* if propensity score or regression model is correct
  - This property is referred to as *double robustness*
- But may be very biased if both models are incorrect

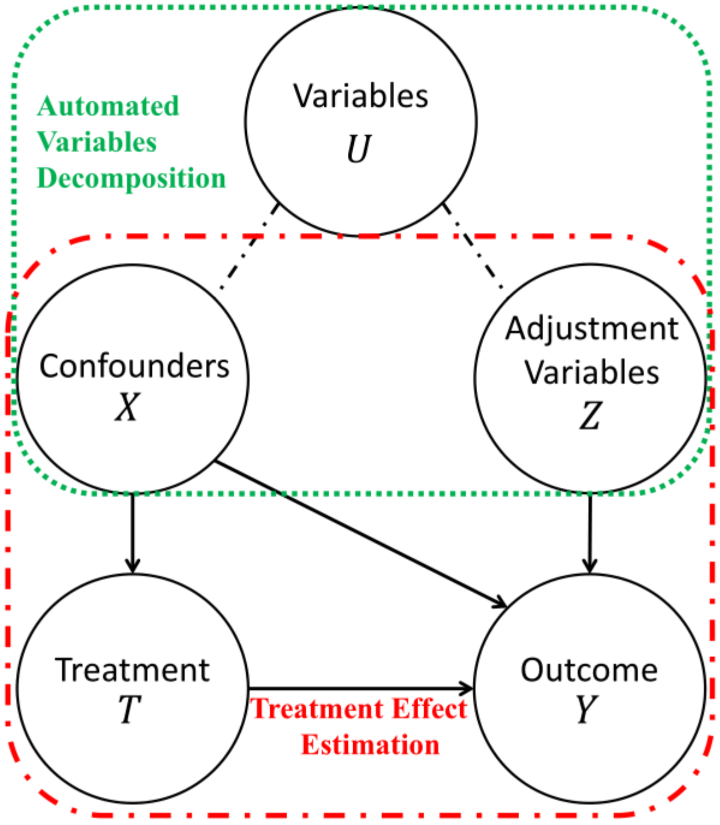
Propensity Score based Methods

- Recap:
  - Propensity Score Matching
  - Inverse of Propensity Weighting
  - Doubly Robust
- Need to estimate propensity score
  - Treat all observed variables as confounders
  - In Big Data Era, High dimensional data
  - But not all variables are confounders



(a) Previous Causal Framework.

Data-Driven Variable Decomposition (D²VD)



(b) Our Causal Framework.

- Separateness Assumption:
  - All observed variables U can be decomposed into three sets: Confounders **X**, Adjustment Variables **Z**, and Irrelevant variables **I** (Omitted).
- Propensity Score Estimation:

$$e(\mathbf{X}) = p(T = 1|\mathbf{X})$$

- Adjusted Outcome:

$$Y^{+} = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

- Our D²VD ATE Estimator:

$$\widehat{ATE}_{D^2VD} = \hat{E}(Y^{+})$$

Kuang K, Cui P, Li B, et al. Treatment effect estimation with data-driven variable decomposition[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

Data-Driven Variable Decomposition (D²VD)

- **Confounders Separation** & **ATE Estimation**.
- With our D²VD estimator:

$$\widehat{ATE}_{D^2VD} = \hat{E}(Y^+) \;=\; E\left(\left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))} \right)$$

- By minimizing following objective function:

$$minimize \quad \|Y^+ - h(\mathbf{U})\|^2.$$

- We can estimate the ATE as:

$$\widehat{ATE}_{D^2VD} = \hat{E}(h(\mathbf{U}))$$

Data-Driven Variable Decomposition (D²VD)

$$minimize \quad \|Y^+ - h(\mathbf{U})\|^2 \qquad \text{Where} \quad Y^+ = \left(Y^{obs} - \phi(\mathbf{Z})\right) \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

$$\boxed{e(\mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\beta)} \qquad \phi(\mathbf{Z}) = \mathbf{Z}\alpha,}$$

Replace X, Z with U $\quad h(\mathbf{U}) = \mathbf{U}\gamma,$

$$minimize \quad \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2, \qquad \text{Where} \quad W(\beta) := \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))}$$

$$s.t. \quad \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau,$$

$$\|\alpha\|_1 \le \lambda, \ \|\beta\|_1 \le \delta, \ \|\gamma\|_1 \le \eta, \ \boxed{\|\alpha \odot \beta\|_2^2 = 0.}$$

α, β, γ

- Adjustment variables: $\mathbf{Z} = \{\mathbf{U}_i : \hat{\alpha}_i \neq 0\}$
- Confounders: $\mathbf{X} = \{\mathbf{U}_i : \hat{\beta}_i \neq 0\}$
- Treatment Effect: $\widehat{ATE}_{D^2VD} = E(\mathbf{U}\hat{\gamma})$

Data-Driven Variable Decomposition (D^2VD)

Bias Analysis:

Our D^2VD algorithm is unbiased to estimate causal effect

THEOREM 1. *Under assumptions 1-4, we have*

$$E(Y^+|\mathbf{X},\mathbf{Z}) = E(Y(1) - Y(0)|\mathbf{X},\mathbf{Z}).$$

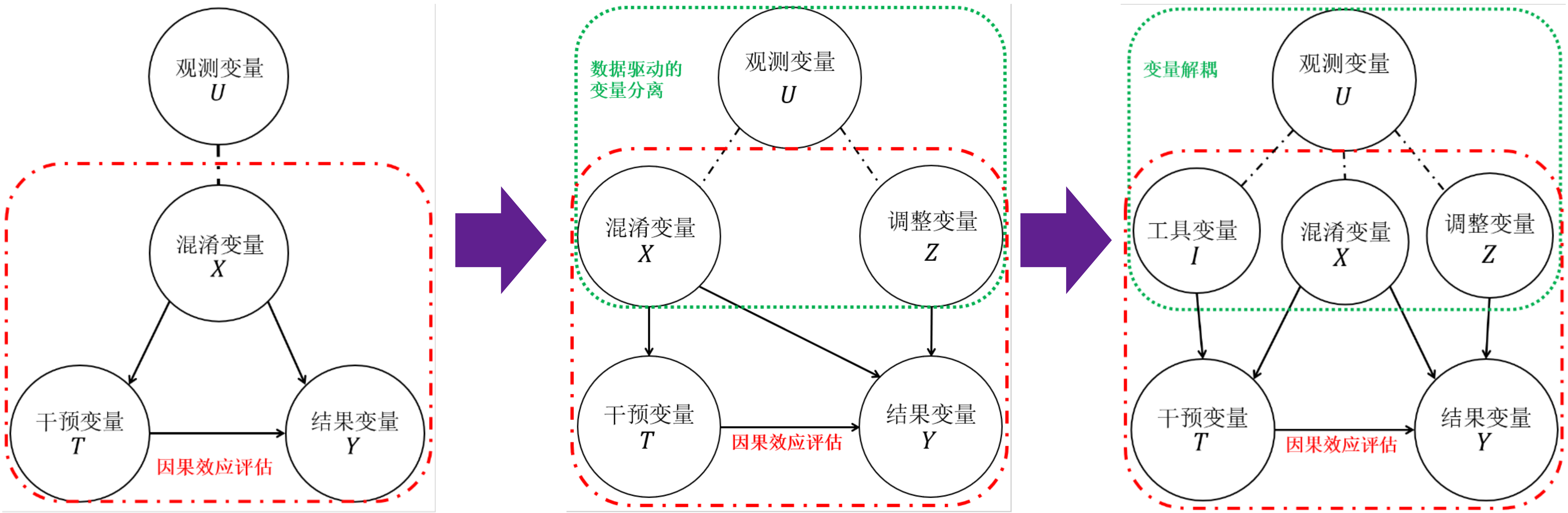
Variance Analysis:

The asymptotic variance of Our D^2VD algorithm is smaller

THEOREM 2. *The asymptotic variance of our adjusted estimator $\widehat{ATE}_{adj}$ is no greater than IPW estimator $\widehat{ATE}_{IPW}$:*

$$\sigma^2_{adj} \leq \sigma^2_{IPW}.$$

Learning Decomposed Representation for Counterfactual Inference

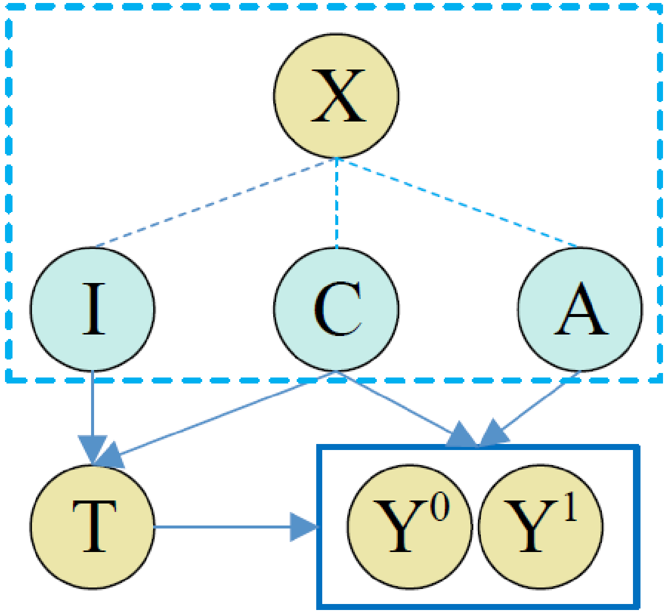


Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Learning Decomposed Representation for Counterfactual Inference

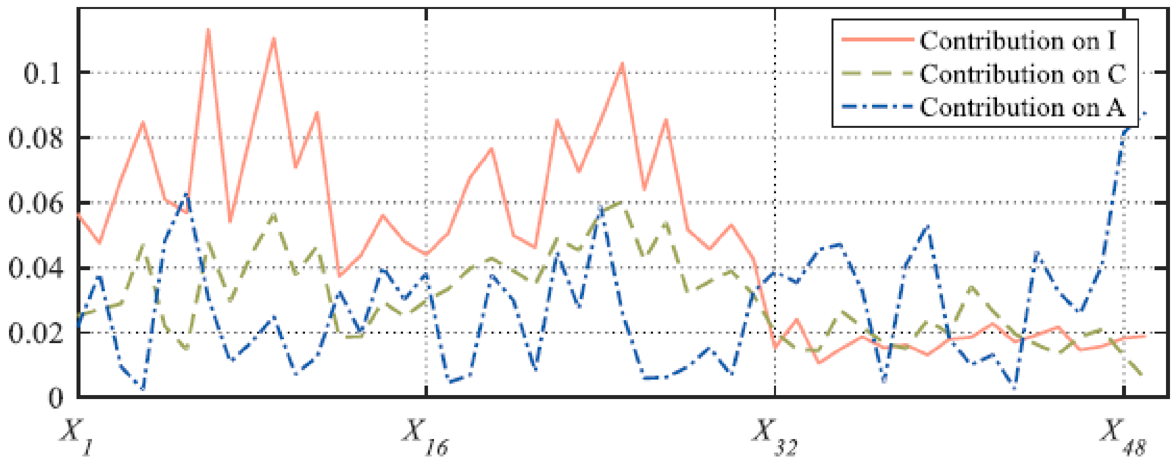
- Three decomposed representation networks
  - $I(X),\ C(X),\ A(X)$
- Three decomposition and balancing regularizers
  - Confounder identification: $A(X) \perp T, I(X) \perp Y \mid T$
  - Confounder balancing: $w \cdot C(X) \perp T$
- Two regression networks
  - $Y(T=1),\ Y(T=0)$
- Orthogonal Regularizer for Decomposition

$$\mathcal{L}_O = \bar{I}_W^T \cdot \bar{C}_W + \bar{C}_W^T \cdot \bar{A}_W + \bar{A}_W^T \cdot \bar{I}_W$$

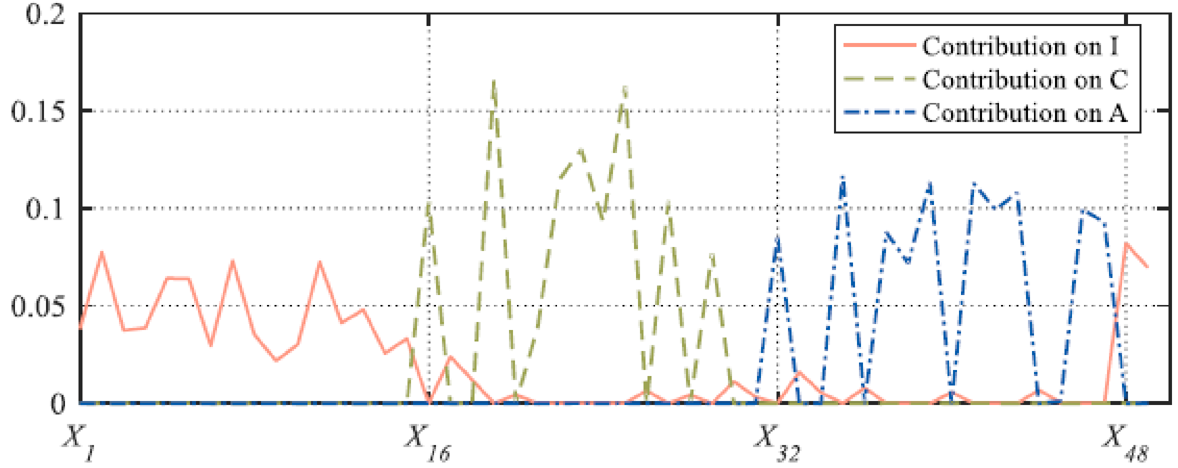


Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Learning Decomposed Representation for Counterfactual Inference



(a) DR-CFR in Syn_16_16_16_3000



(b) DeR-CFR in Syn_16_16_16_3000

Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Learning Decomposed Representation for Counterfactual Inference

Table 1: The results on IHDP.

IHDP				
Mean +/- Std	**Within-sample**		**Out-of-sample**	
Methods	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
CFR-MMD	0.702 +/- 0.037	0.284 +/- 0.036	0.795 +/- 0.078	0.309 +/- 0.039
CFR-WASS	0.702 +/- 0.034	0.306 +/- 0.040	0.798 +/- 0.088	0.325 +/- 0.045
CFR-ISW	0.598 +/- 0.028	0.210 +/- 0.028	0.715 +/- 0.102	0.218 +/- 0.031
SITE	0.609 +/- 0.061	0.259 +/- 0.091	1.335 +/- 0.698	0.341 +/- 0.116
DR-CFR	0.657 +/- 0.028	0.240 +/- 0.032	0.789 +/- 0.091	0.261 +/- 0.036
DeR-CFR	**0.444 +/- 0.020**	**0.130 +/- 0.020**	**0.529 +/- 0.068**	**0.147 +/- 0.022**

Table 2: Ablation studies of DeR-CFR.

$\mathcal{L}_A$	$\mathcal{L}_I$	$\mathcal{L}_{C\_B}$	$\mathcal{L}_O$	PEHE Within-sample	PEHE Out-of-sample
✓	✓	✓	✓	**0.444 +/- 0.020**	**0.529 +/- 0.068**
✓	✓	✓		0.478 +/- 0.033	0.542 +/- 0.053
✓	✓		✓	0.482 +/- 0.039	0.565 +/- 0.075
✓		✓	✓	0.479 +/- 0.030	0.560 +/- 0.071
	✓	✓	✓	0.635 +/- 0.035	0.858 +/- 0.133

Wu A, Kuang K, Yuan J, et al. Learning Decomposed Representation for Counterfactual Inference[J]. arXiv preprint arXiv:2006.07040, 2020.

Summary: Propensity Score based Methods

$$e(X) = P(T = 1|X)$$

- Propensity Score Matching (PSM):
  - Units matching by their propensity score
- Inverse of Propensity Weighting (IPW):
  - Units reweighted by inverse of propensity score
- Doubly Robust (DR):
  - Combing IPW and regression

Treat all observed variables as confounder, ignoring non-confounders

- **Data-Driven Variable Decomposition (D^2VD):**
  - Automatically separate the confounders and adjustment variables
  - Confounder: estimate propensity score for IPW
  - Adjustment variables: regression on outcome for reducing variance
  - Improving accuracy and reducing variance on treatment effect estimation
- But these methods need propensity score model is correct

Methods for Causal Inference

- **Matching**
- **Propensity Score Based Methods**
  - Propensity Score Matching
  - Inverse of Propensity Weighting (IPW)
  - Doubly Robust
  - Data-Driven Variable Decomposition (D^2VD)
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing
  - Differentiated Confounder Balancing (DCB)

Directly Confounder Balancing

- Recap: Propensity score based methods
  - Sample reweighting for confounder balancing
  - But need propensity score model is correct
  - Weights would be very large if propensity score is close to 0 or 1

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

- Can we directly learn sample weight that can balance confounders' distribution between treated and control?

Yes!

Directly Confounder Balancing

- **Motivation**: The collection of all the moments of variables uniquely determine their distributions.
- **Methods**: Learning sample weights by directly balancing confounders' moments as follows

$$\min_{W} \|\boxed{\overline{\mathbf{X}}_t} - \boxed{\mathbf{X}_c^T W}\|_2^2$$

The first moments of X on the **Treated** Group

The first moments of X on the **Control** Group

With moments, the sample weights can be learned without any model specification.

Directly Confounder Balancing

- **Motivation**: The collection of all the moments of variables uniquely determine their distributions.
- **Methods**: Learning sample weights by directly balancing confounders' moments as follows

$$\min_{W} \|\boxed{\overline{\mathbf{X}}_t} - \boxed{\mathbf{X}_c^T W}\|_2^2$$

The first moments of X on the Treated Group

The first moments of X on the Control Group

- Estimating ATT by: $$\widehat{ATT} = \sum_{i:T_i=1} \frac{1}{n_t} Y(1) - \sum_{j:T_j=0} W_j Y(0)$$

Entropy Balancing

$$\min_{W} \quad W \log(W)$$

$$s.t. \quad \boxed{\|\overline{\mathbf{X}}_t - \mathbf{X}_c^T W\|_2^2 = 0}$$

$$\sum_{i=1}^{n} W_i = 1, W \succeq 0$$

- Maximum the entropy of sample weights W
- Directly confounder balancing by sample weights W
- But, treat all variables as confounders and balance them equally

Approximate Residual Balancing

- 1. compute approximate balancing weights W as

$$W = \operatorname{argmin}_W \left\{ (1-\zeta)\|W\|_2^2 + \boxed{\zeta \left\| \overline{X}_t - \mathbf{X}_c^\top W \right\|_\infty^2} \text{ s.t. } \sum_{\{i:T_i=0\}} W_i = 1 \text{ and } W_i \geq 0 \right\}$$

- 2. Fit β_c in the linear model using a lasso or elastic net,

$$\hat{\beta}_c = \operatorname{argmin}_\beta \left\{ \sum_{\{i:W_i=0\}} \left(Y_i^{\text{obs}} - X_i \cdot \beta \right)^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \right\}$$

- 3. Estimate the ATT as

$$\widehat{ATT} = \overline{Y}_t - \left(\overline{X}_t \cdot \hat{\beta}_c + \sum_{\{i:T_i=0\}} W_i \left(Y_i^{\text{obs}} - X_i \cdot \hat{\beta}_c \right) \right)$$

- Double Robustness: Exact confounder balancing or regression is correct.
- But, treats all variables as confounders and balance them equally

Methods for Causal Inference

- **Matching**
- **Propensity Score Based Methods**
  - Propensity Score Matching
  - Inverse of Propensity Weighting (IPW)
  - Doubly Robust
  - Data-Driven Variable Decomposition (D^2VD)
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing
  - **Differentiated Confounder Balancing (DCB)**

Differentiated Confounder Balancing

- **Ideas**: simultaneously learn ***confounder weights*** $\boldsymbol{\beta}$ and ***sample weighs*** $\boldsymbol{W}$.

$$\min \quad (\underline{\beta^T} \cdot (\underline{\overline{\mathbf{X}}_t - \mathbf{X}_c^T W}))^2$$

- ***Confounder weights*** determine which variable is confounder and its contribution on confounding bias.
- ***Sample weights*** are designed for confounder balancing.

Kuang K, Cui P, Li B, et al. Estimating treatment effect in the wild via differentiated confounder balancing[C]//Proceedings of the 23rd ACM SIGKDD. 2017: 265-274.

Confounder Weights Learning

- General relationship among X, T, and Y:

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon \quad \Rightarrow \quad \begin{aligned} ATT &= E(g(\mathbf{X}_t)) \\ Y(0) &= f(\mathbf{X}) + \epsilon \end{aligned}$$

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{a}_1\mathbf{X} + \sum_{ij} a_{ij}X_iX_j + \sum_{ijk} a_{ijk}X_iX_jX_k + \cdots + R_n(\mathbf{X}) \\ &= \alpha\mathbf{M}. \end{aligned} \qquad \mathbf{M} = (\mathbf{X}, X_iX_j, X_iX_jX_k, \cdots).$$

Confounder weights

Confounding bias

$$\widehat{ATT} = ATT + \underline{\sum_{k=1}^{p} \boxed{\alpha_k} \boxed{\left(\sum_{i:T_i=1} \frac{1}{n_t} M_{i,k} - \sum_{j:T_j=0} W_j M_{j,k}\right)}} + \phi(\epsilon).$$

If $\alpha_k = 0$, then M_k is not confounder, no need to balance.
Different confounders have different confounding weights.

Confounder Weights Learning

Propositions:

- In observational studies, **not all** observed variables are confounders, and different confounders make **unequal** confounding bias on ATT with their own weights.
- The **confounder weights** can be learned by regressing potential outcome $Y(0)$ on augmented variables M.

$$\mathbf{M} = (\mathbf{X}, X_i X_j, X_i X_j X_k, \cdots).$$

Sample Weights Learning

$$\mathbf{M} = (\mathbf{X}, X_i X_j, X_i X_j X_k, \cdots).$$

- Any variable's distribution can be uniquely determined by the collection of all its moments.
- Learning the sample weights W by directly confounder balancing with confounders' moments.

$$\min \quad \left(\beta^T \cdot (\underbrace{\boxed{\overline{\mathbf{M}}_t}}_{\text{Confounders' moments on the Treated Group}} - \underbrace{\boxed{\mathbf{M}_c^T W}}_{\text{Confounders' moments on the Control Group}})\right)^2$$

With moments, the sample weights can be learned without any model specification.

Differentiated Confounder Balancing

- Objective Function

$$\min \quad \boxed{\left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2} + \boxed{\lambda \sum_{j:T_j=0} (1+W_j)\cdot(Y_j - M_j \cdot \beta)^2,}$$

$$s.t. \quad \|W\|_2^2 \le \delta, \ \ \|\beta\|_2^2 \le \mu, \ \ \|\beta\|_1 \le \nu, \mathbf{1}^T W = 1 \ \ and \ \ W \succeq 0$$

The ENT[3] and ARB[4] algorithms are special case of our DCB algorithm by setting the confounder weights as unit vector.

Our DCB algorithm is more generalize for treatment effect estimation.

Experiments - Robustness Test

More results see our paper!

	n/p	$n=2000, p=50$			$n=2000, p=100$		
r_c	Estimator	$Bias$ (SD)	MAE	RMSE	$Bias$ (SD)	MAE	RMSE
$r_c = 0.8$	$\widehat{ATT}_{dir}$	51.06 (3.725)	51.06	51.19	143.0 (9.389)	143.0	143.3
	$\widehat{ATT}_{IPW}$	29.99 (4.048)	29.99	30.26	98.24 (8.462)	98.24	98.60
	$\widehat{ATT}_{DR}$	0.345 (0.253)	0.367	0.428	4.492 (0.333)	4.492	4.504
	$\widehat{ATT}_{ENT}$	15.06 (1.745)	15.06	15.16	63.02 (4.551)	63.02	63.19
	$\widehat{ATT}_{ARB}$	0.231 (0.645)	0.553	0.685	2.909 (0.491)	2.909	2.951
	$\widehat{ATT}_{DCB}$	**0.003** (0.127)	**0.102**	**0.127**	**0.020** (0.135)	**0.114**	**0.136**

- ***Directly estimator*** fails in all settings, since it ignores confounding bias.
- ***IPW and DR estimators*** make huge error when facing high dimensional variables or the model specifications are incorrect.
- ***ENT and ARB estimators*** have poor performance since they balance all variables equally.

Experiments - Robustness Test

More results see our paper!

	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$		
r_c	Estimator	$Bias$ (SD)	MAE	RMSE	$Bias$ (SD)	MAE	RMSE
$r_c = 0.8$	$\widehat{ATT}_{dir}$	51.06 (3.725)	51.06	51.19	143.0 (9.389)	143.0	143.3
	$\widehat{ATT}_{IPW}$	29.99 (4.048)	29.99	30.26	98.24 (8.462)	98.24	98.60
	$\widehat{ATT}_{DR}$	0.345 (0.253)	0.367	0.428	4.492 (0.333)	4.492	4.504
	$\widehat{ATT}_{ENT}$	15.06 (1.745)	15.06	15.16	63.02 (4.551)	63.02	63.19
	$\widehat{ATT}_{ARB}$	0.231 (0.645)	0.553	0.685	2.909 (0.491)	2.909	2.951
	$\widehat{ATT}_{DCB}$	**0.003** (0.127)	**0.102**	**0.127**	**0.020** (0.135)	**0.114**	**0.136**

Our DCB estimator achieves significant improvements over the baselines in different settings.

Our DCB estimator is very robust!

Experiments - Accuracy Test

Results of ATT estimation

Variables Set	V-RAW		V-INTERACTION	
Estimator	$\widehat{ATT}$	$Bias$ (SD)	$\widehat{ATT}$	$Bias$ (SD)
$\widehat{ATT}_{dir}$	-8471	10265 (374)	-8471	10265 (374)
$\widehat{ATT}_{IPW}$	-4481	6275 (971)	-4365	6159 (1024)
$\widehat{ATT}_{DR}$	1154	639 (491)	1590	204 (812)
$\widehat{ATT}_{ENT}$	1535	259 (995)	1405	388 (787)
$\widehat{ATT}_{ARB}$	1537	257 (996)	1627	167 (957)
$\widehat{ATT}_{DCB}$	1958	**164** (728)	1836	**43** (716)

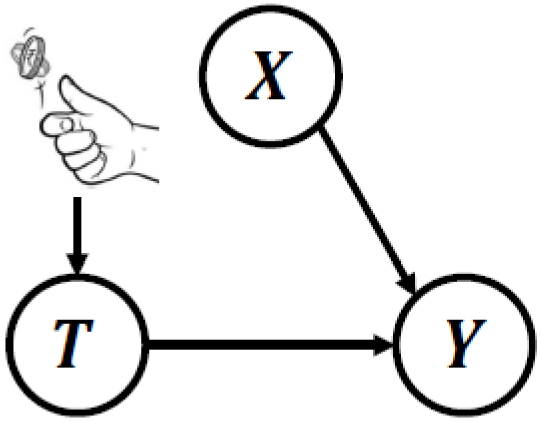
Our DCB estimator is more **accurate** than the baselines.

Our DCB estimator achieve a better confounder balancing under V-INTERACTION setting.

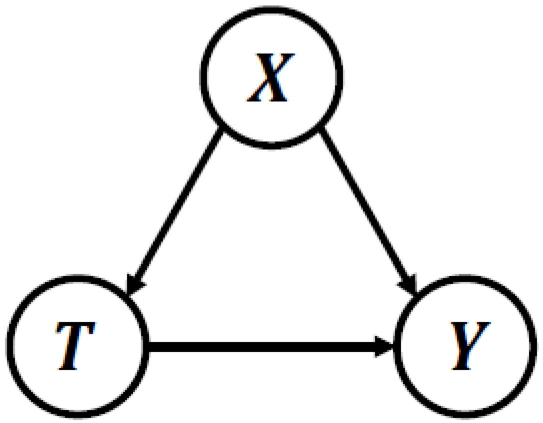
Summary: Directly Confounder Balancing

- **Motivation:** Moments can uniquely determine distribution
- Entropy Balancing
  - Confounder balancing with maximizing entropy of sample weights
- Approximate Residual Balancing
  - Combine confounder balancing and regression for doubly robust
- Treat all variables as confounders, and balance them equally
- But different confounders make different bias
- **Differentiated Confounder Balancing (DCB)**
  - Theoretical proof on the necessary of differentiation on confounders
  - Improving the accuracy and robust on treatment effect estimation

Continuous Treatment Effect Estimation



(a) Randomized Controlled Trial (RCT)



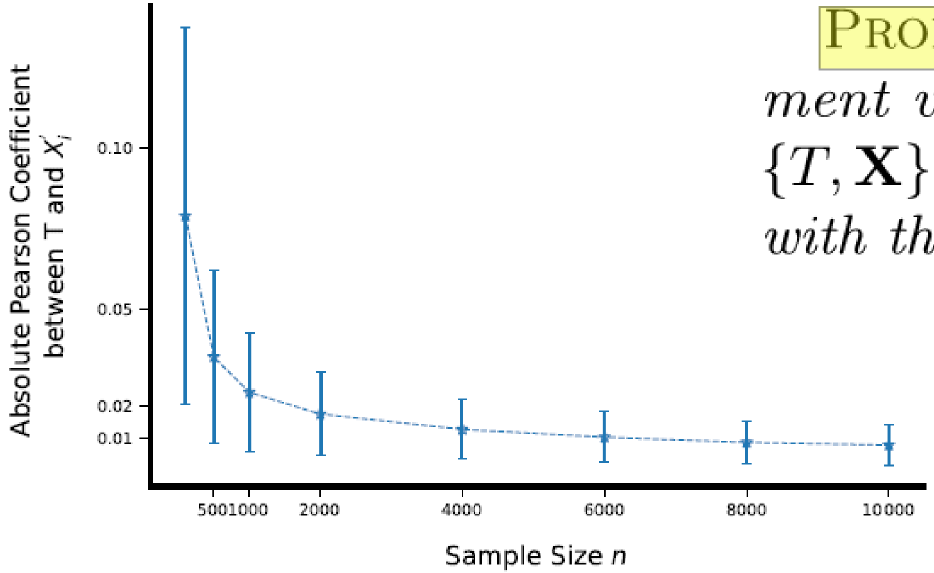
(b) Observational Studies

- Binary Treatment
  - T=0 or T=1
  - $T \perp X$: confounder balancing
- Multi-valued Treatment
  - T=0,1,2,…
  - $T \perp X$: confounder balancing
- Continuous Treatment
  - How to make $T \perp X$?

Li R, Kuang K, Li B, et al. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding[C]//KDD workshop 2020.

Continuous Treatment Effect Estimation

- Our goal: $T \perp X$
- Variable randomly shuffle to achieve independence



PROPOSITION 1. *By randomly shuffle the value of the treatment variable T over all samples in observed data $\mathbf{D}_{obs} = \{T, \mathbf{X}\}$, the shuffled treatment T would become independent with the covariates $\mathbf{X}$ if sample size $n \to \infty$.*

Li R, Kuang K, Li B, et al. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding[C]//KDD workshop 2020.

Continuous Treatment Effect Estimation

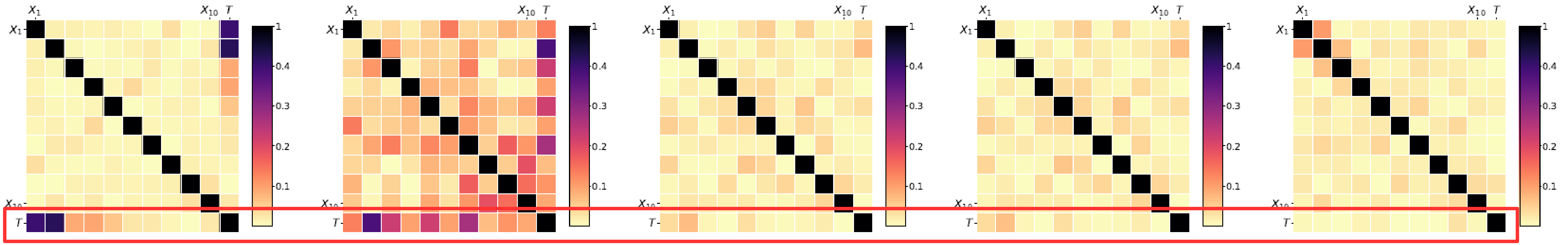
- Our goal: $T \perp X$
- "calibration" distribution generation

$$\mathbf{D}_{cal} = \{T', \mathbf{X}'\}$$

- "calibration" distribution approximation
  - Learning sample weights for distribution matching $\mathbf{D}_{obs} = \{T, \mathbf{X}\}$
  - GAN based methods: Generative Adversarial De-confounding (**GAD**)

$$\begin{aligned} L(\mathbf{w}, d) &= \mathbb{E}_{(t,x)\sim\mathbf{D}_{cal}}[l(d(t,x), \boxed{1})] \\ &+ \mathbb{E}_{(t,x)\sim\mathbf{D}_{obs}}[\boxed{w_{(t,x)}} \cdot l(d(t,x), \boxed{0})], \\ s.t. \quad & \mathbb{E}_{(t,x)\sim\mathbf{D}_{obs}}[w_{(t,x)}] = 1, \mathbf{w} \succeq 0, \end{aligned}$$

Continuous Treatment Effect Estimation



(a) Raw Data (b) $\mathbf{IPW}_{unstable}$ (c) $\mathbf{IPW}_{stable}$ (d) CBGPS (e) GAD

Method	*TWINS*		
	BIAS_{MTEF}	RMSE_{MTEF}	RMSE_{ADRF}
OLS	0.208(0.079)	0.236(0.089)	0.686(0.350)
$\text{IPW}_{unstable}$	1.385(0.757)	1.532(0.890)	5.506(2.061)
IPW_{stable}	1.693(1.599)	1.878(1.849)	6.982(4.453)
ISMW	0.165(0.062)	0.181(0.069)	0.962(0.214)
CBGPS	0.187(0.137)	0.216(0.158)	0.683(0.380)
GAD	**0.127(0.039)**	**0.144(0.046)**	**0.383(0.091)**

Summary: Methods for Causal Inference

- **Matching** Limited to low-dimensional settings
- **Propensity Score Based Methods**
  - Propensity Score Matching
  - Inverse of Propensity Weighting (IPW)
  - Doubly Robust

  Treat all observed variables as confounder

  - Data-Driven Variable Decomposition (D^2VD)

  Not all observed variables are confounders

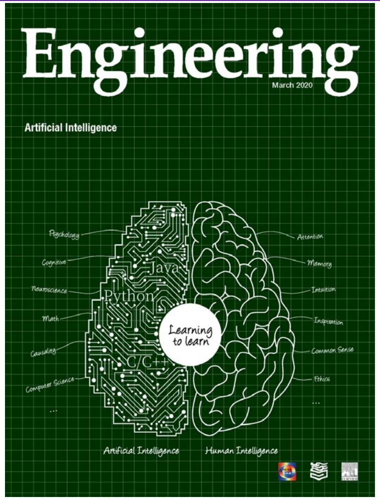
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing

  Balance all confounder equally

  - Differentiated Confounder Balancing (DCB)

  Different confounders make different bias

- Generative Adversarial De-confounding



Engineering 综述论文解读：因果推理（Causal Inference）

况琨，李廉，耿直，徐雷，张坤，廖备水，
黄华新，丁鹏，苗旺，蒋智超

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z. (2020). Causal Inference. *Engineering*.
http://www.engineering.org.cn/ch/10.1016/j.eng.2019.08.016

具体内容

- 况琨：平均因果效应评估-简要回顾与展望
- 李廉：反事实推理的归因问题
- 耿直：辛普森悖论和替代指标悖论
- 徐雷：因果发现CPT（因果势理论）方法
- 张坤：从观测数据中发现因果关系
- 廖备水，黄华新：形式论辩在因果推理和解释中的作用
- 丁鹏：复杂实验中的因果推断
- 苗旺：观察性研究中的工具变量和阴性对照方法
- 蒋智超：有干扰下的因果推断

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z. (2020). Causal Inference. *Engineering*. http://www.engineering.org.cn/ch/10.1016/j.eng.2019.08.016

De-biased Court's View Generation with Causality (EMNLP20)

PLAINTIFF'S CLAIM	The plaintiff A claimed that the defendant B should return the loan of $29,500 *Principle Claim* and the corresponding interest *Interest Claim*.
FACT DESCRIPTION	After the hearing, the court held the facts as follows: The defendant B borrowed $29,500 from the plaintiff A, and agreed to return after one month. After the loan expired, the defendant failed to return *Fact*.
COURT'S VIEW	The court concluded that the loan relationship between the plaintiff A and the defendant B is valid. The defendant failed to return the money on time *Rationale*. Therefore, the plaintiff's claim on principle was supported *Acceptance* according to law. The court did not support the plaintiff's claim on interest *Rejection* because the evidence was insufficient *Rationale*.

Input:

- ☐ Plaintiff's claim
- ☐ Fact description

Output:

- ☐ Court's View, which consists of
  - ☐ Rationale
  - ☐ Judgment

Court's view generation is a specific text generation task

Yiquan Wu, Kun Kuang*, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si and Fei Wu . De-biased Court' s View Generation with Causality, EMNLP, 2020

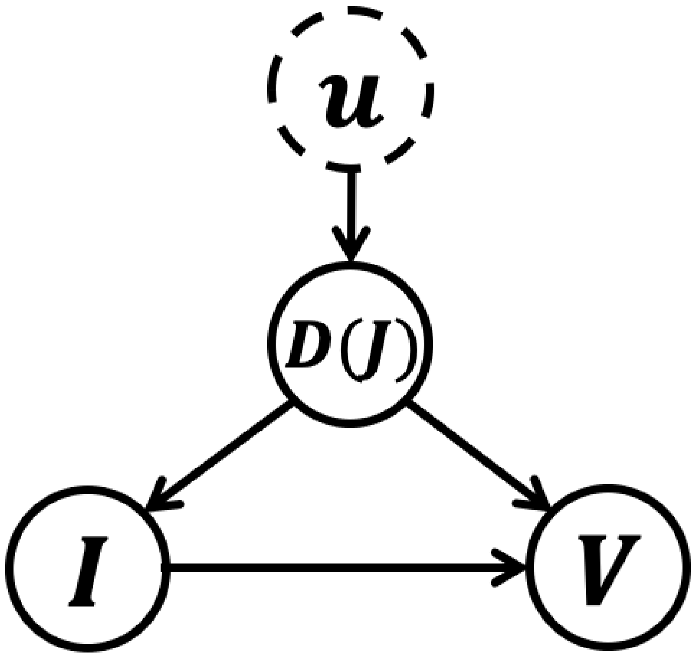
Challenges

PLAINTIFF'S CLAIM	The plaintiff A claimed that the defendant B should return the loan of $29,500 *Principle Claim* and the corresponding interest *Interest Claim*.
FACT DESCRIPTION	After the hearing, the court held the facts as follows: The defendant B borrowed $29,500 from the plaintiff A, and agreed to return after one month. After the loan expired, the defendant failed to return *Fact*.
COURT'S VIEW	The court concluded that the loan relationship between the plaintiff A and the defendant B is valid. The defendant failed to return the money on time *Rationale*. Therefore, the plaintiff's claim on principle was supported *Acceptance* according to law. The court did not support the plaintiff's claim on interest *Rejection* because the evidence was insufficient *Rationale*.

- There exists '**no claim, no trial**' principle in civil legal systems
  - court's view should only focus on the facts related to the claims
- The **imbalance** of judgment in civil cases
  - over 76% of cases were supported in private lending
  - would blind the training of the model by focusing on the supported cases while ignoring the non-supported cases

Imbalance: Mechanism Confounding Bias

- Imbalance between supported and non-supported cases
  - Lead to confounding bias during model training
- Understanding confounding bias with a causal graph:
  - u: unobserved data generation mechanism
  - D(J): judgment in dataset
  - I: input (i.e., plaintiff's claim and fact description)
  - V: court's view

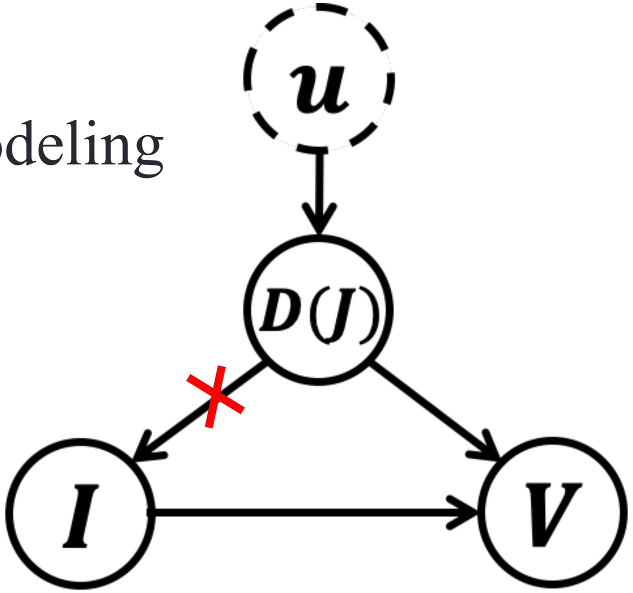


- Understanding confounding bias mathematically
  - j: judgment (support and non-support):

$$\boxed{P(V|I) = \sum_j P(V|I,j)P(j|I)} \xrightarrow{P(j=1|I)\approx 1} \boxed{P(V|I) \approx P(V|I,j=1)}$$

Attentional and Counterfactual based NLG

- Attentional encoder:
  - Claim-aware attention
- Counterfactual decoder:
  - Back-door adjustment: from observation to intervention
  - Cut the dependence between D(J) and I via counterfactual modeling



$$P(V|I) = \sum_j P(V|I,j)P(j|I)$$

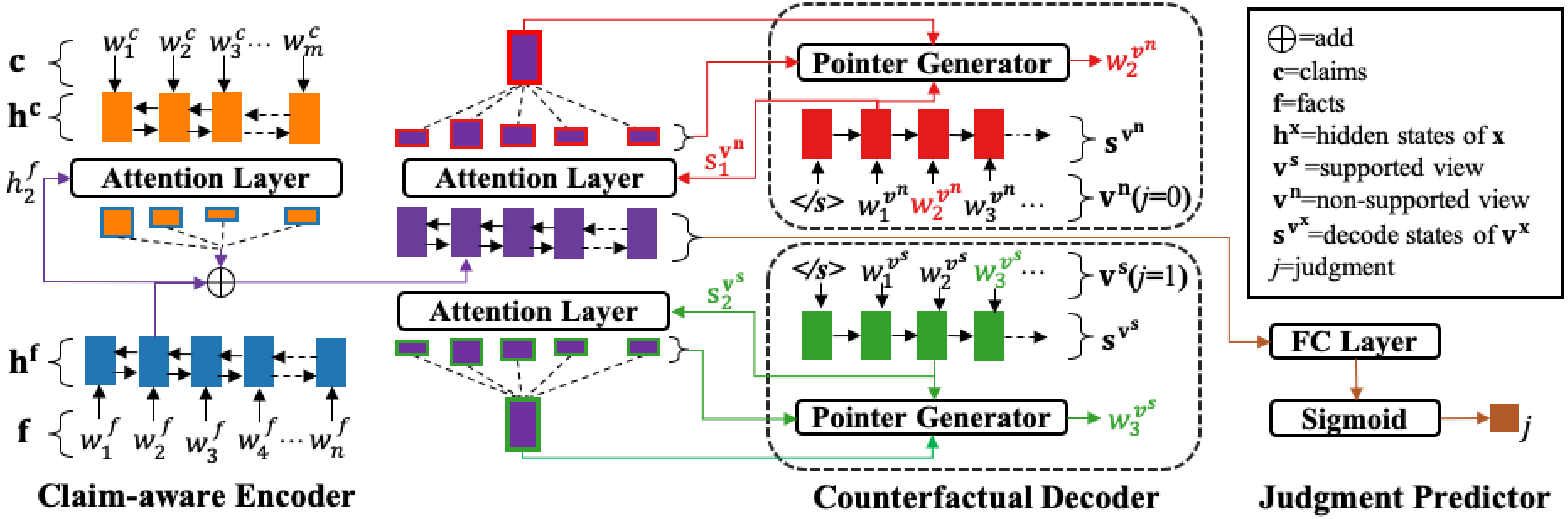
Back-door

$$P(V|do(I)) = \sum_j P(V|I,j)P(j)$$

Binary j

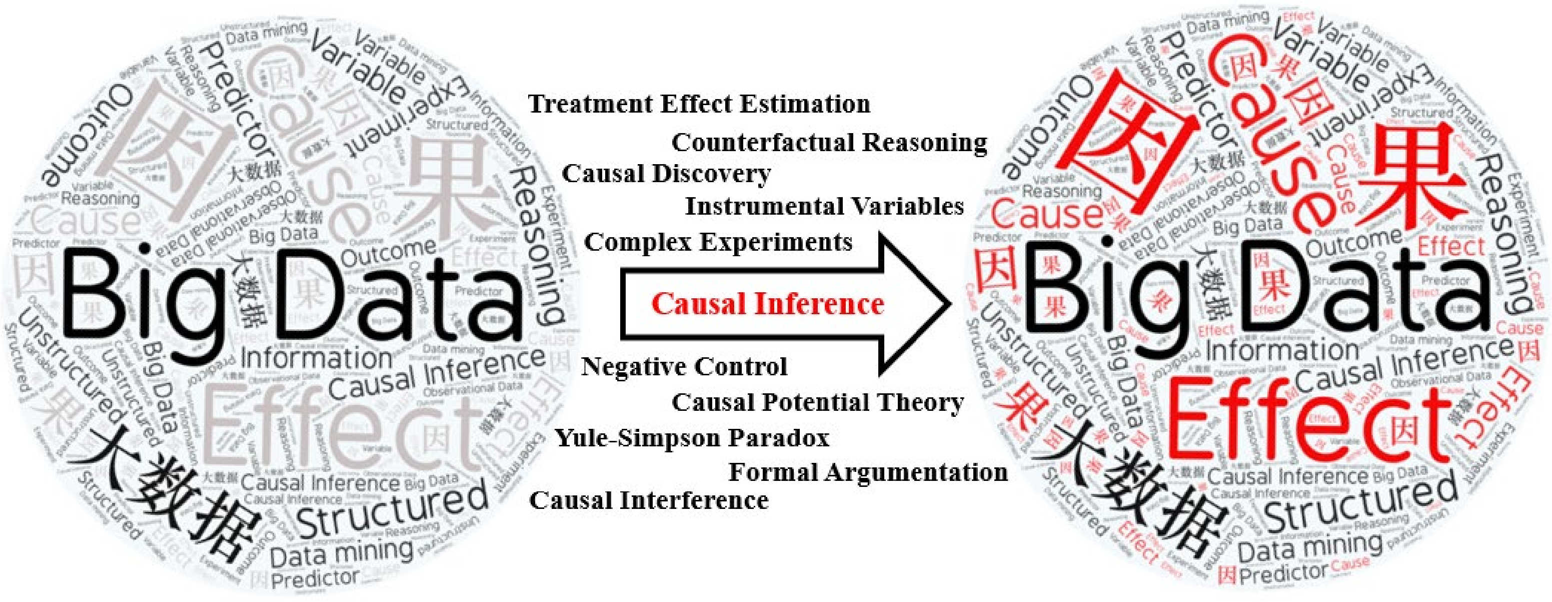
$$P(V|do(I)) = P(V|I,j=0)P(j=0) + P(V|I,j=1)P(j=1)$$

Our Framework



AC-NLG is a multi-task model with:

- Claim-aware encoder
  - Claim embedding
  - Fact embedding
  - Claim-Fact attention
- Counterfactual decoders
  - Supportive court's view generation
  - Non-supportive court's view generation
- Judgment predictor



Thank You!

Kun Kuang
kunkuang@zju.edu.cn
Homepage: https://kunkuang.github.io/