CAUSALLY REGULARIZED MACHINE LEARNING

Peng Cui, Tsinghua University

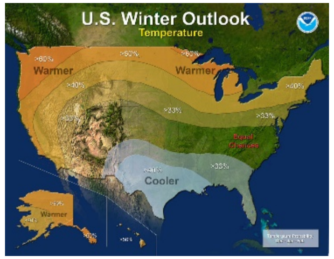
Kun Kuang, Tsinghua University

Bo Li, Tsinghua University

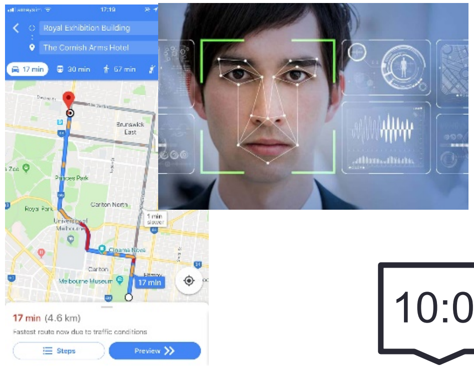
Predictive systems are impacting our life

- A day in our life with predictive analytics

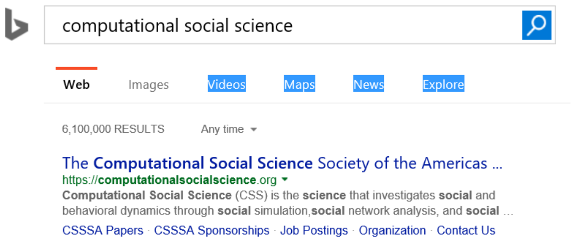
8:00 am

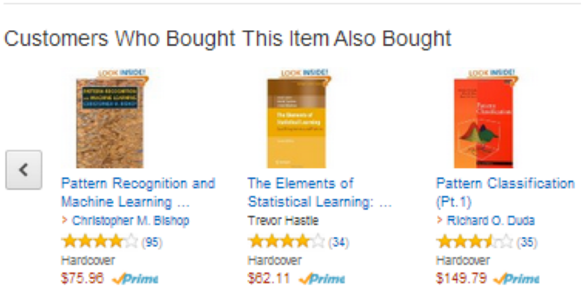


8:30 am



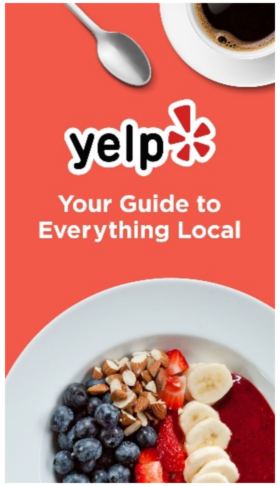
10:00 am





4:00 pm

6:00 pm



8:00 pm

Even in risk-sensitive areas

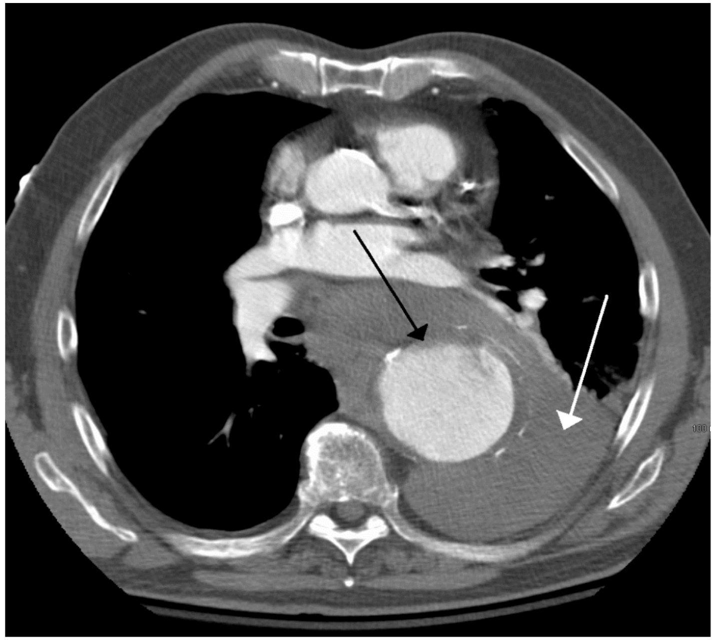
Human

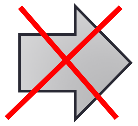
Human's risk-sensitive sense brings new challenges to today's AI

Explainability

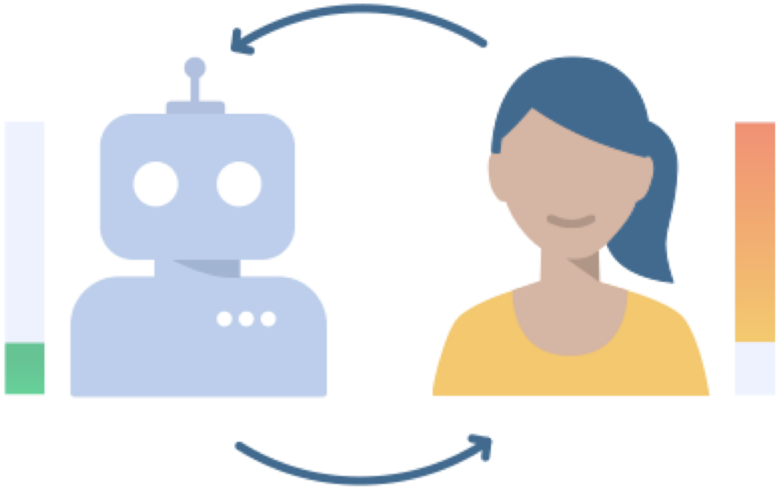
Most machine learning models are black-box models

Unexplainable





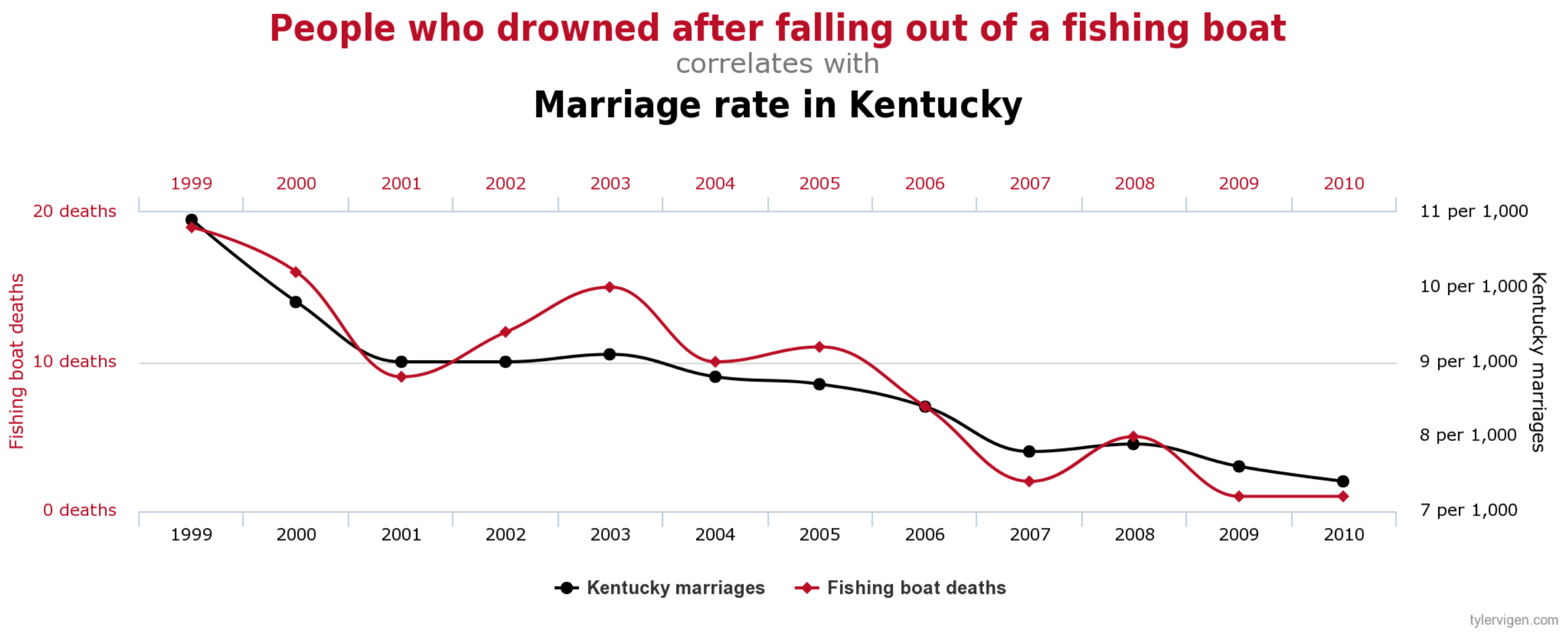
Human in the loop



Health Military Finance Industry

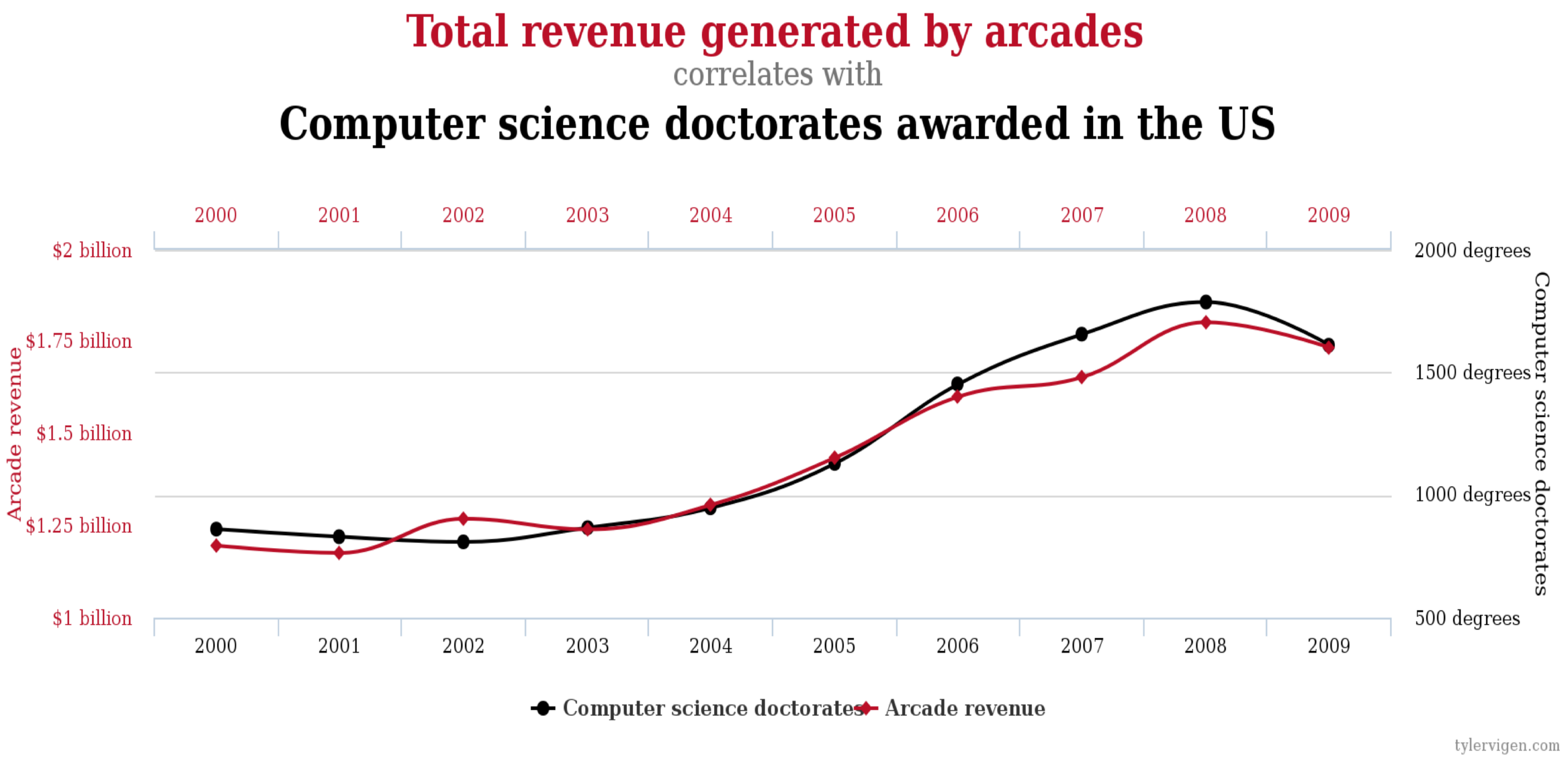
Explainability

- Correlation is not explainable

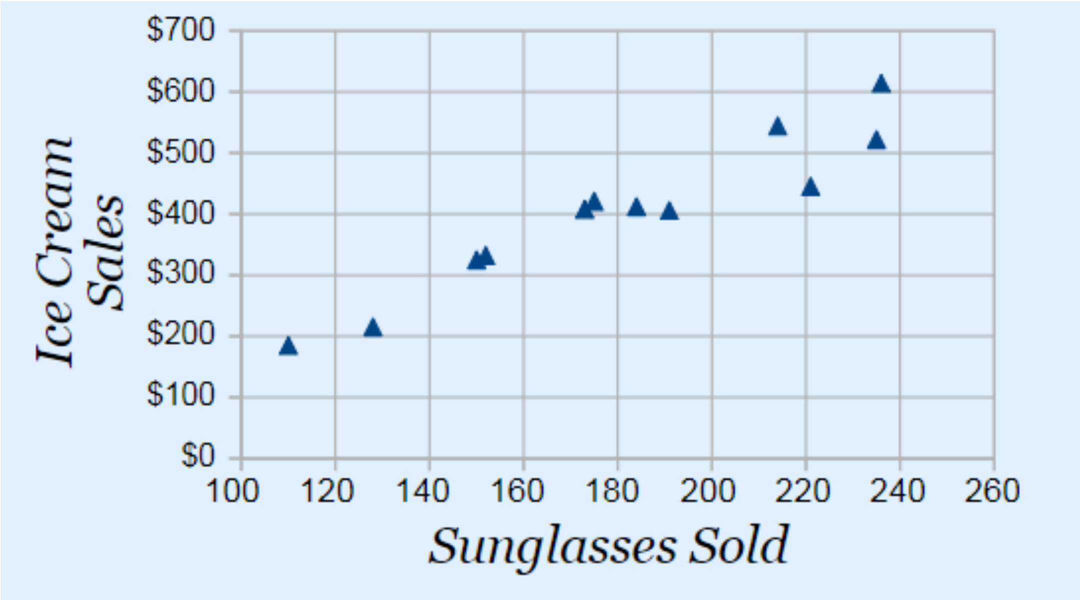


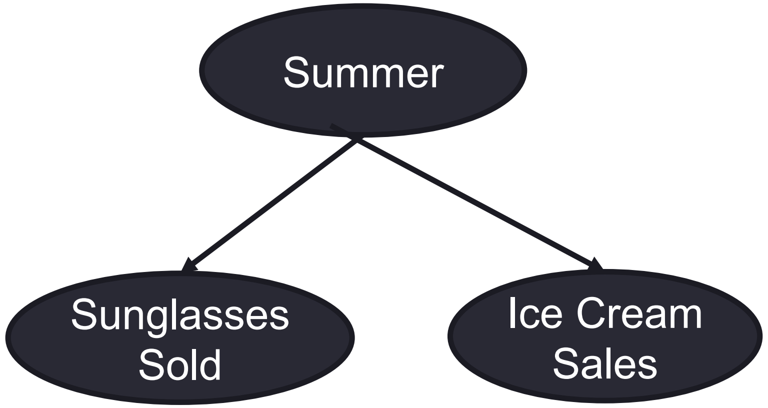
Explainability

- Correlation is not explainable



Explainability



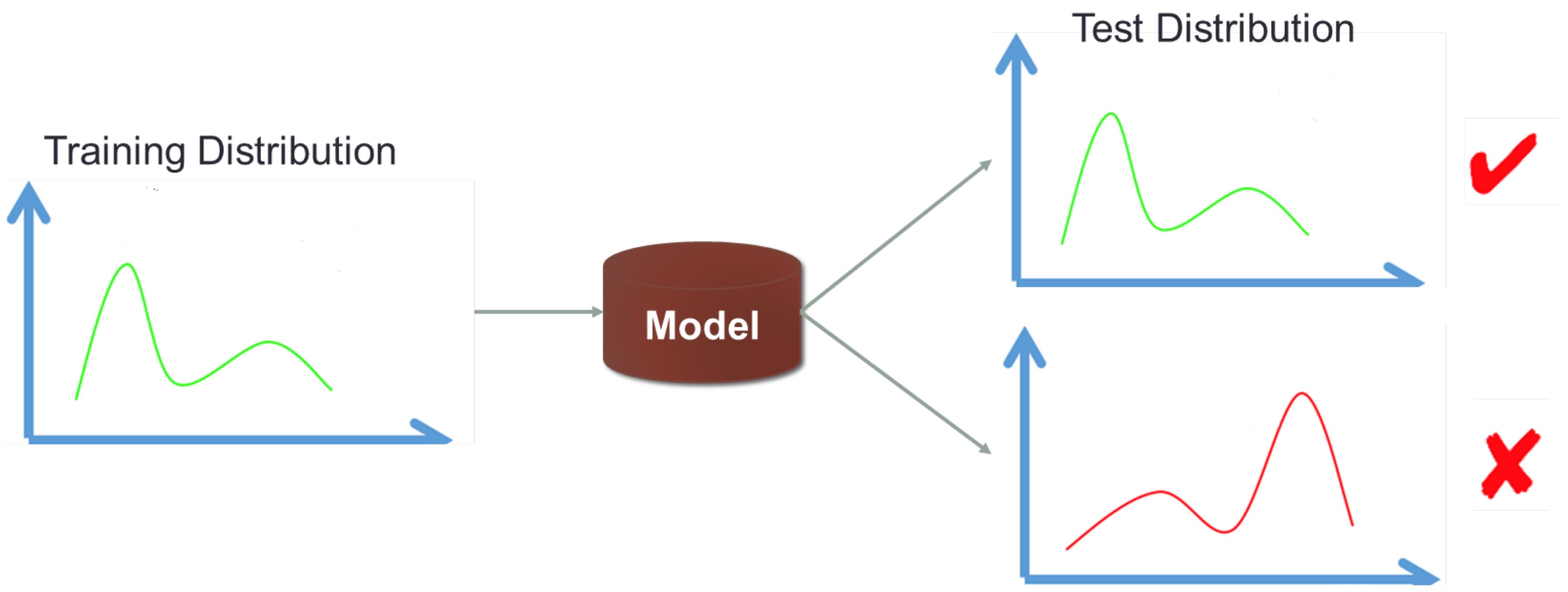


Spurious Correlation !

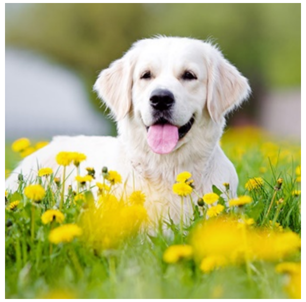
Correlation does not imply causation!

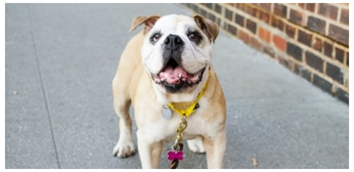
Stability

Most ML methods are developed under IID hypothesis



Stability



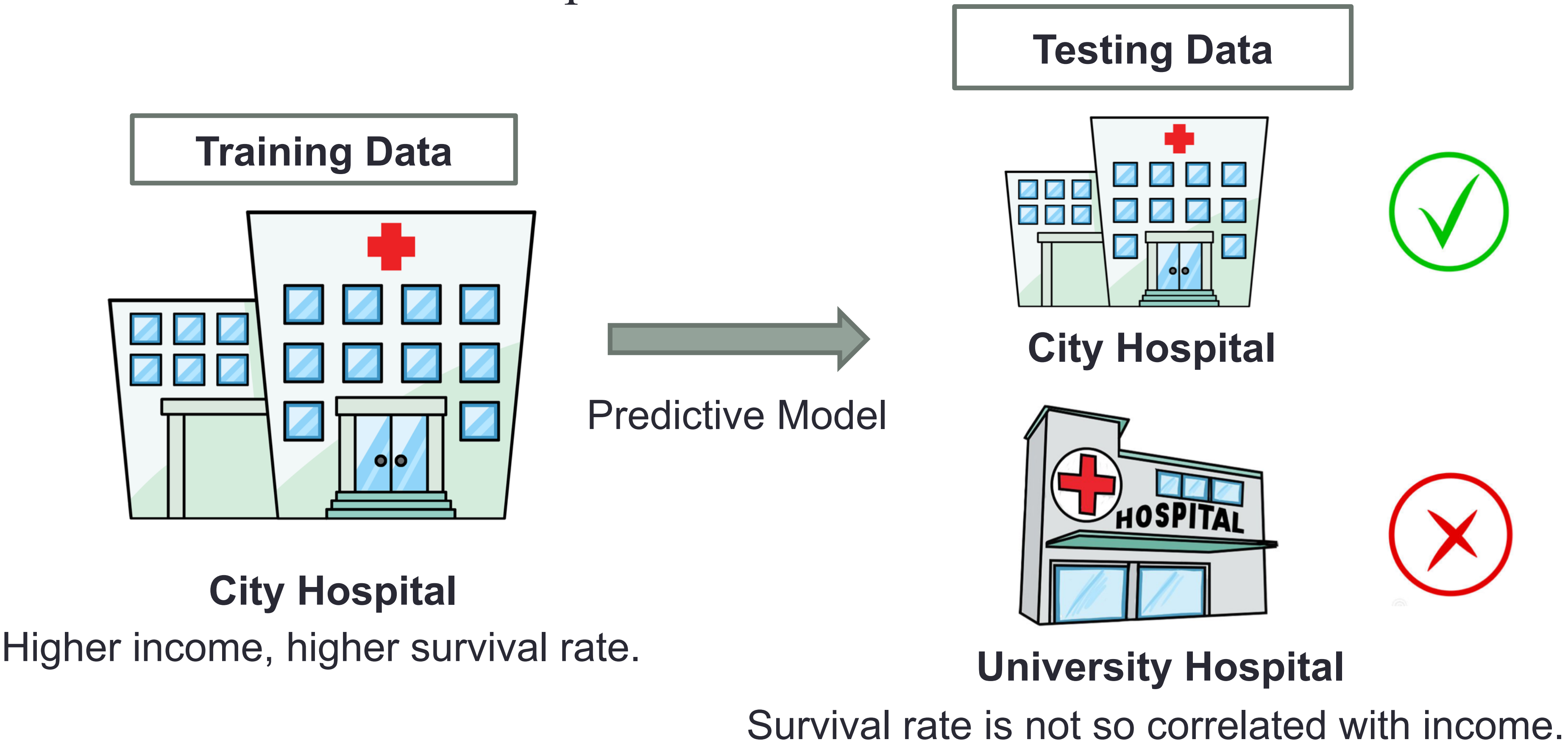

Yes

Maybe

No

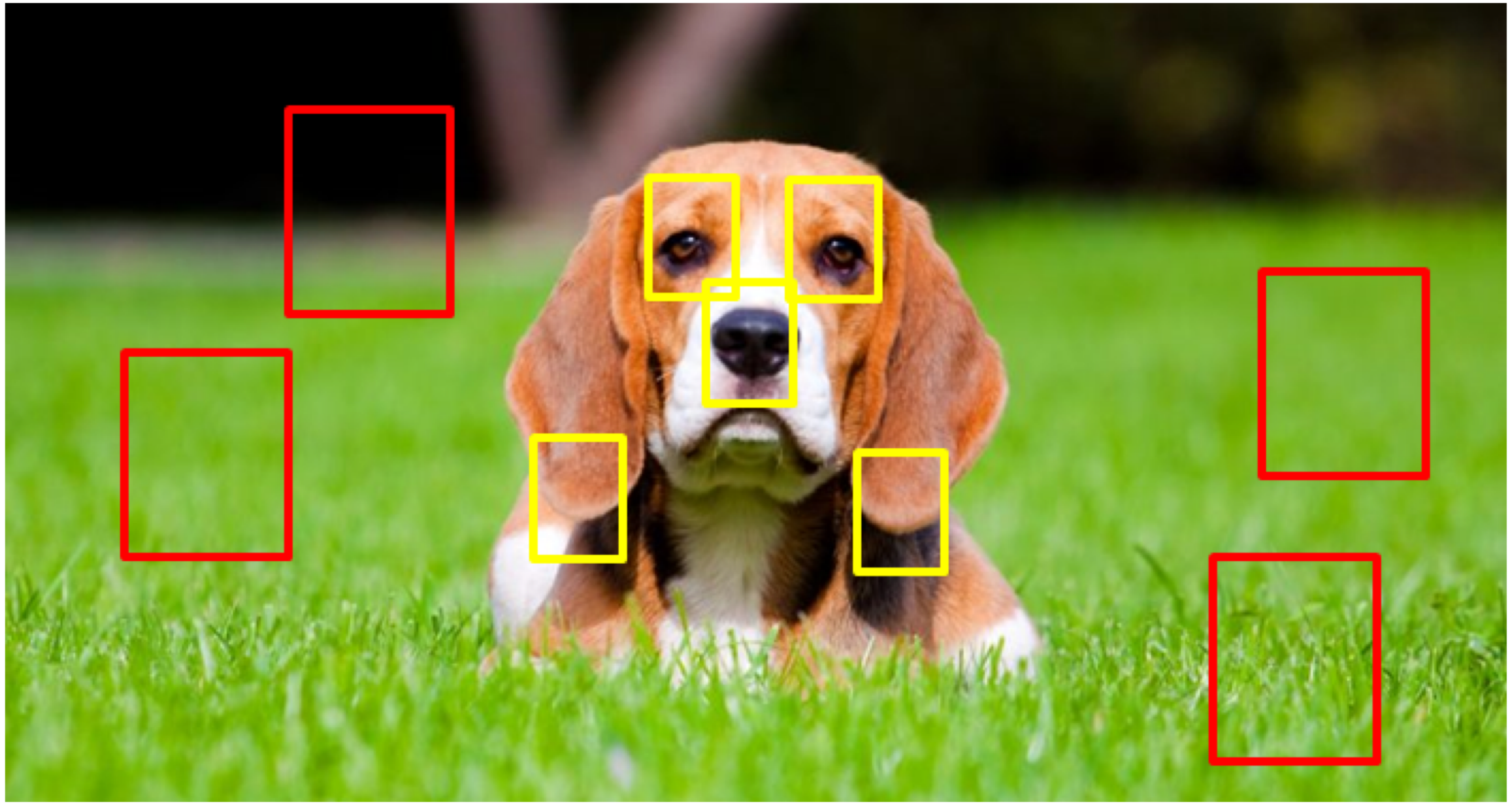
Stability

- Cancer survival rate prediction



Stability

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

Actionability

- Measure success rate (SR)

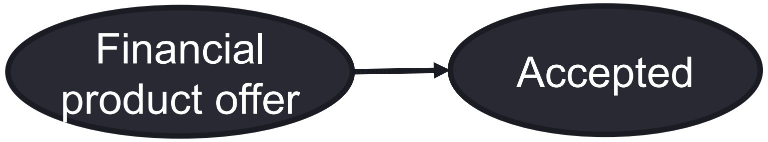
Old Algorithm (A)	New Algorithm (B)
50/1000 **(5%)**	54/1000 **(5.4%)**

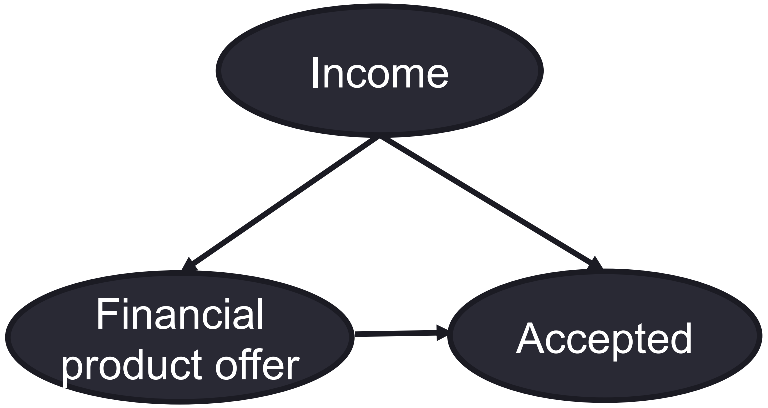
New algorithm increases overall success rate, so it is better?

	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?

Actionability





Higher success rate due to algorithm

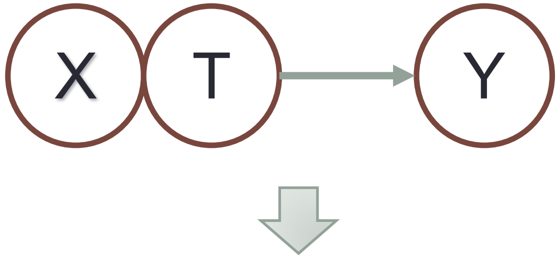
Higher success rate due to confounding bias

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

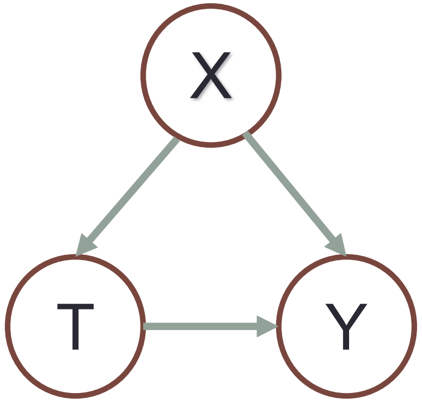
Fairness

The source of these problems: Correlation

Correlation Framework



Causal Framework



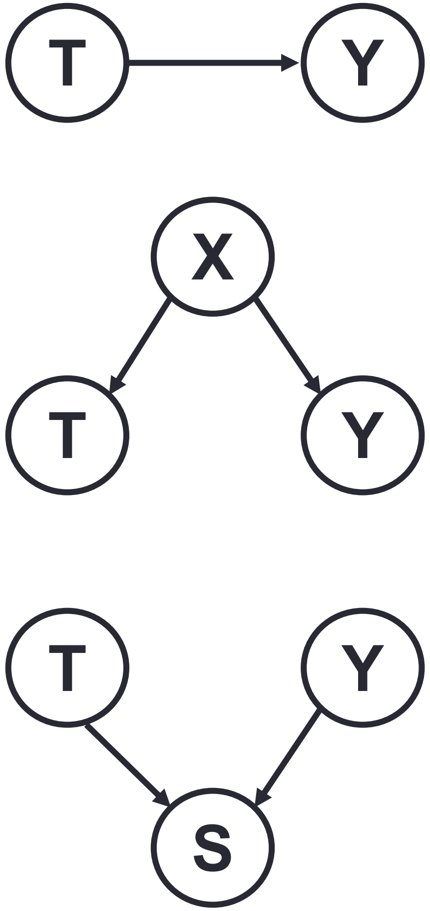
T: skin color
X: income
Y: crime rate

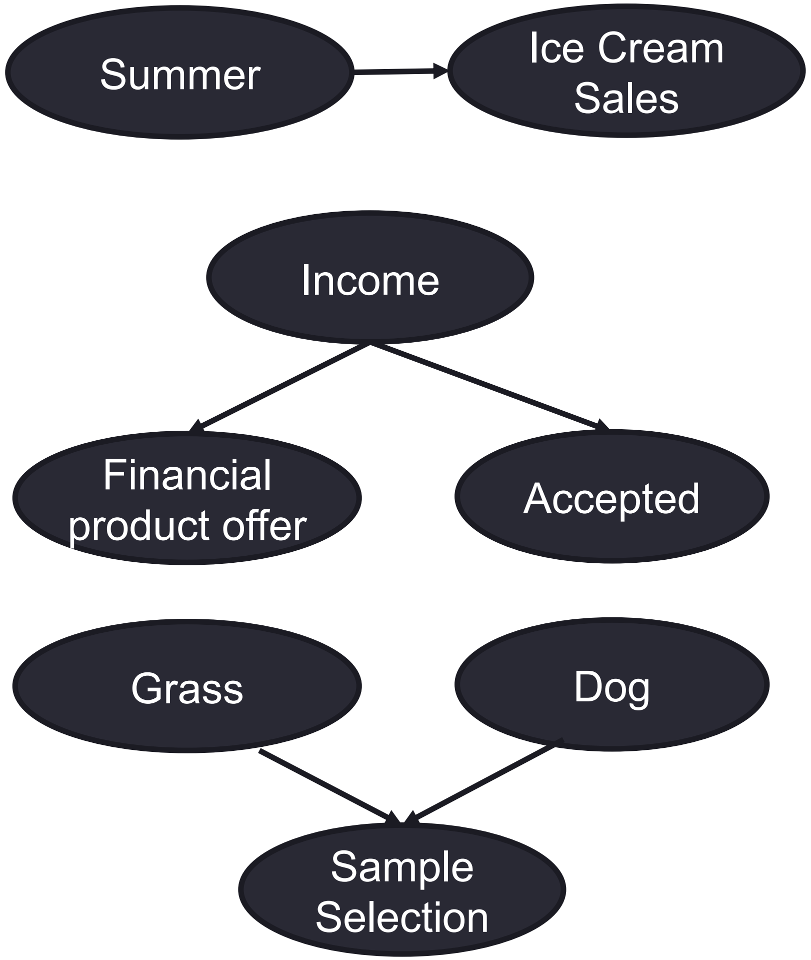
income—crime rate: Strong correlation
skin color—crime rate: Strong correlation

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

Correlation V.S. Causation

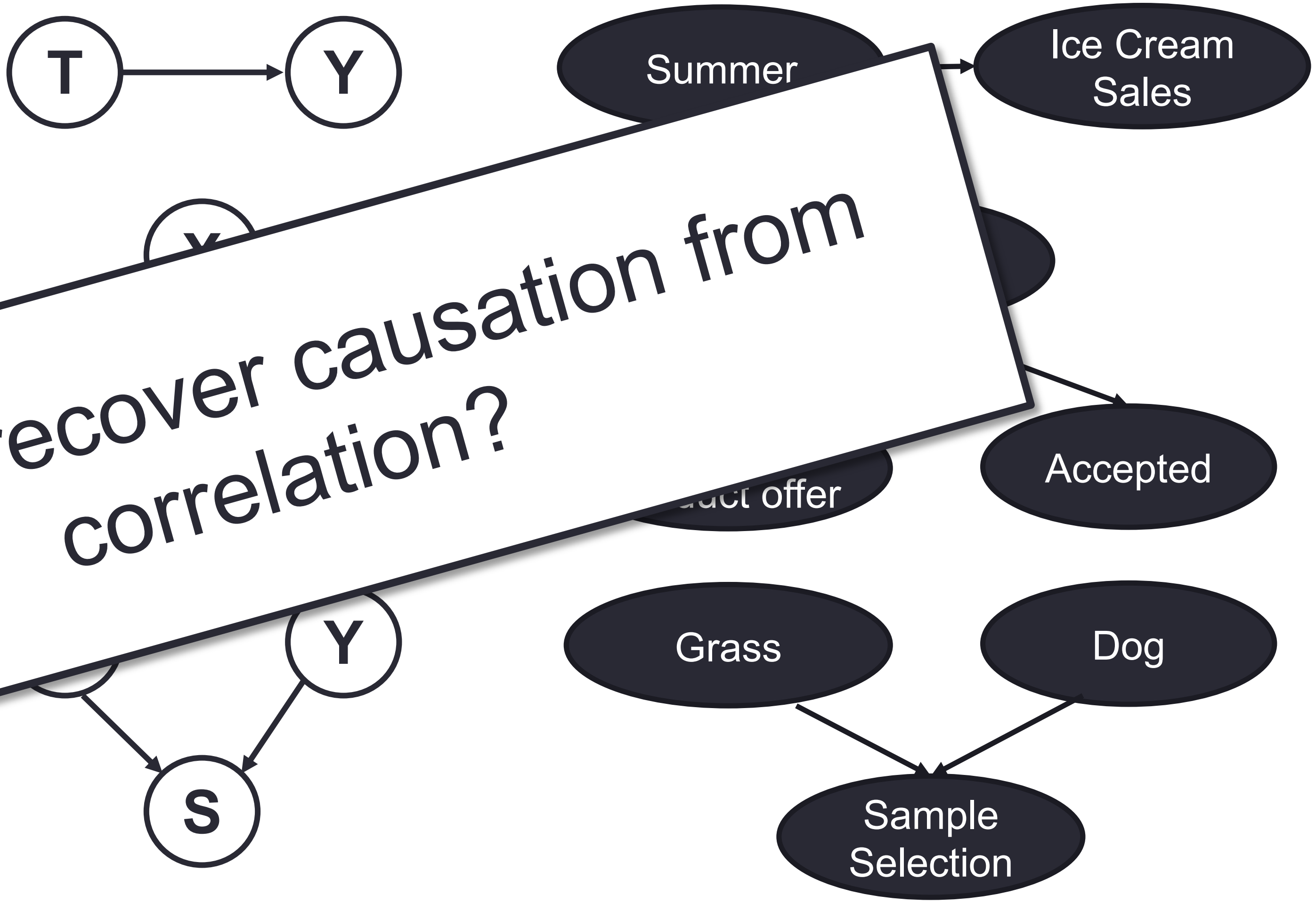
- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and Robust
  - Confounding
    - Ignoring X
    - Spurious Correlation
  - Sample Selection
    - Conditional on S
    - Spurious Correlation





Correlation V.S. Causation

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



Can we recover causation from 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

OUTLINE

PART I. Introduction to Causal Inference

PART II. Methods for Causal Inference

PART III. Causally Regularized Machine Learning

PART IV. Benchmark and Open Datasets

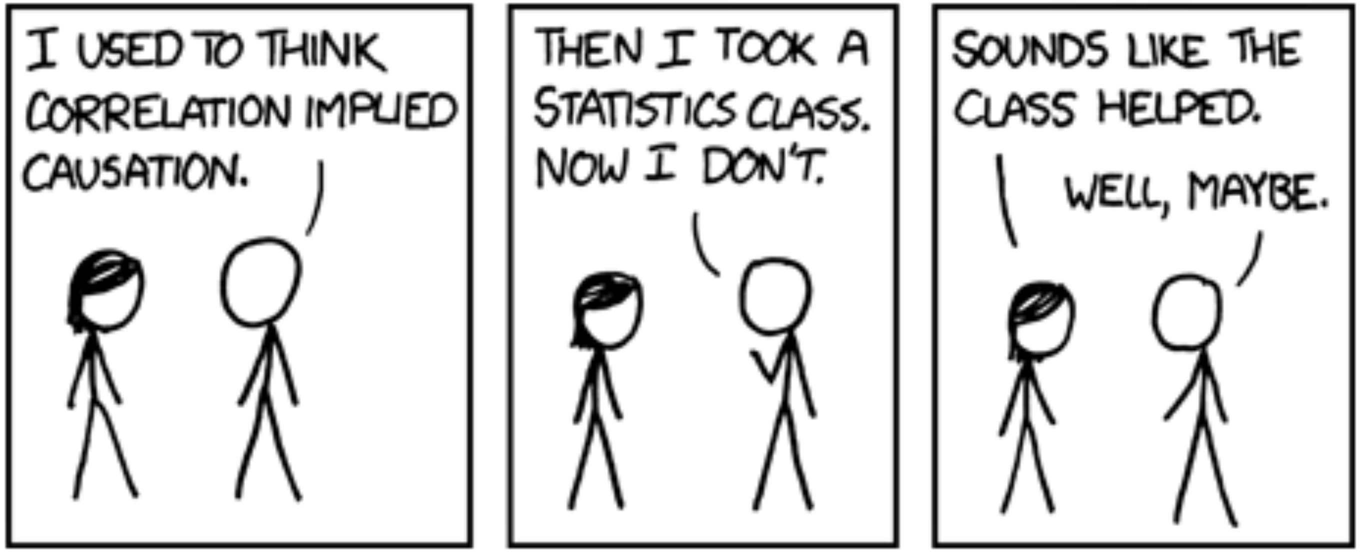
PART V. Conclusion and Discussion

Cause and Effect

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

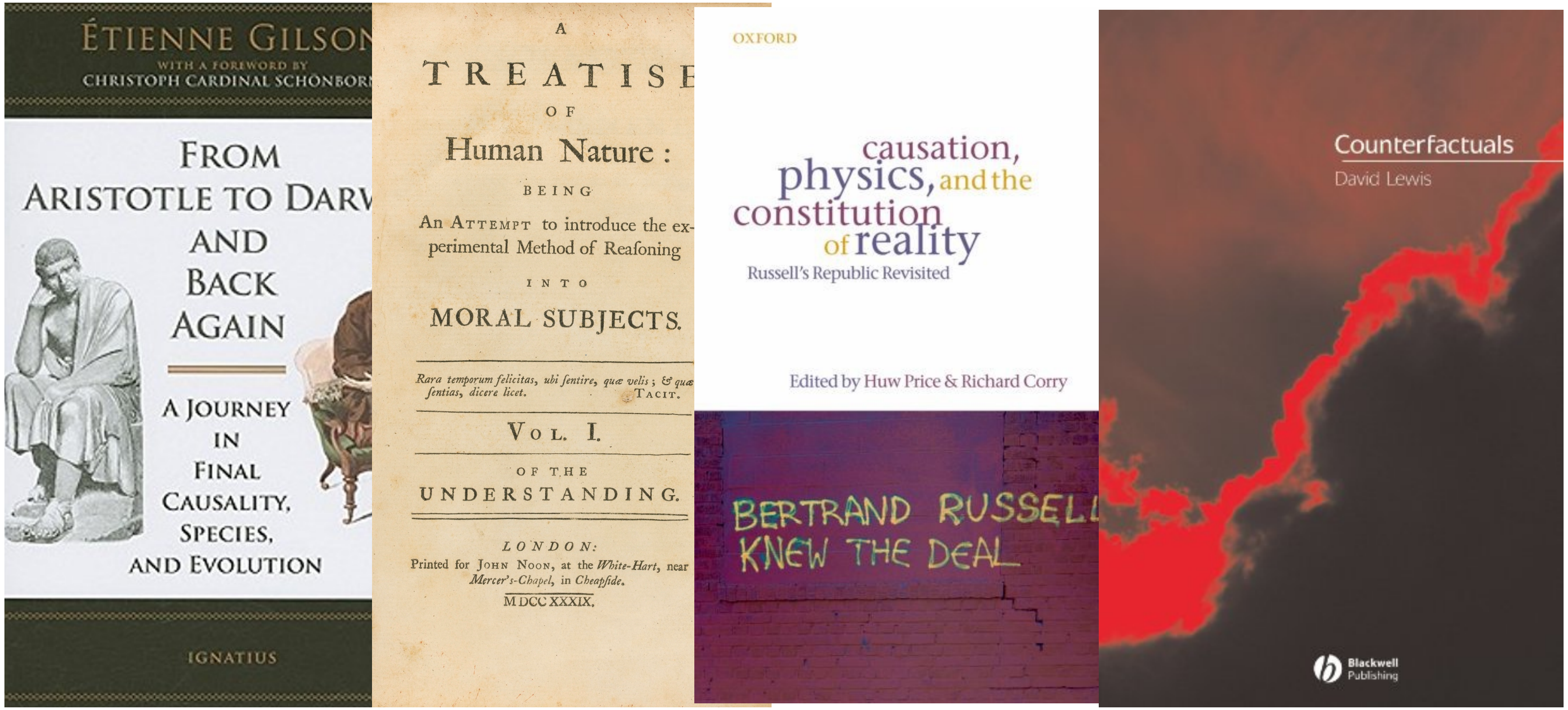
- Questions of cause and effect:
  - Medicine: drug trials, effect of a drug
  - Social science: effect of a policy
  - Marketing: effect of a marketing strategy
  - …
- **What is causality?**

What is causality?



What is causality?

- A big scholarly debate, from Aristotle to Russell



The Three Layer Causal Hierarchy

Level	Typical Activity	Typical Question	Examples
1. Association $P(y \mid x)$ Observational Questions	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y \mid do(x), z)$ Action Questions	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x \mid x', y')$ Counterfactuals Questions	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What I had not been smoking the past 2 years?

Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution[J]. arXiv preprint arXiv:1801.04016, 2018.

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.

Called the “interventionist” interpretation of causality.

**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:

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

Counterfactual Problem

Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	?
P2	0	?	0.6
P3	1	0.3	?
P4	0	?	0.1
P5	1	0.5	?
P6	0	?	0.5
P7	0	?	0.1

- Two key points for causal effect estimation
  - Changing T
  - Keeping everything else constant

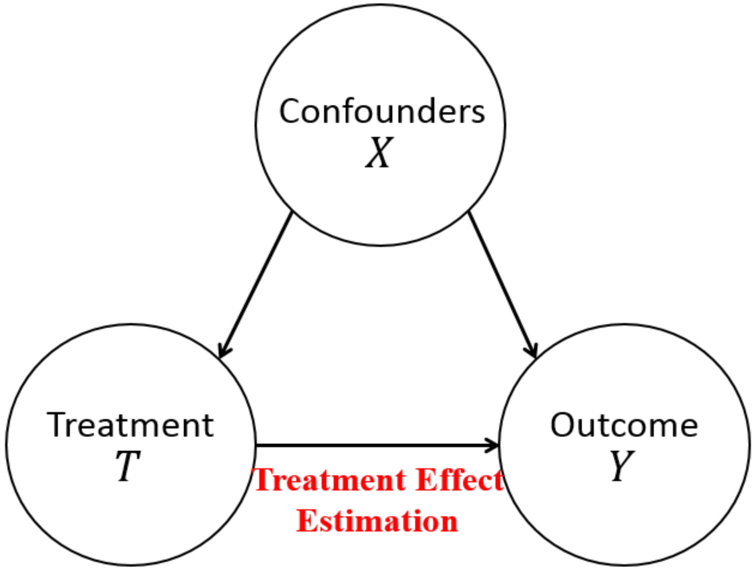
- For each person, observe only one: either $Y_{t=1}$ or $Y_{t=0}$
- For different group (T=1 and T=0), something else are not constant

Potential Outcome Framework

- Confounders X: everything else

- Why keep everything else constant:
  - Confounders X influences both T and Y
  - Y's change could be induced by change of T or since X changed both T and Y?

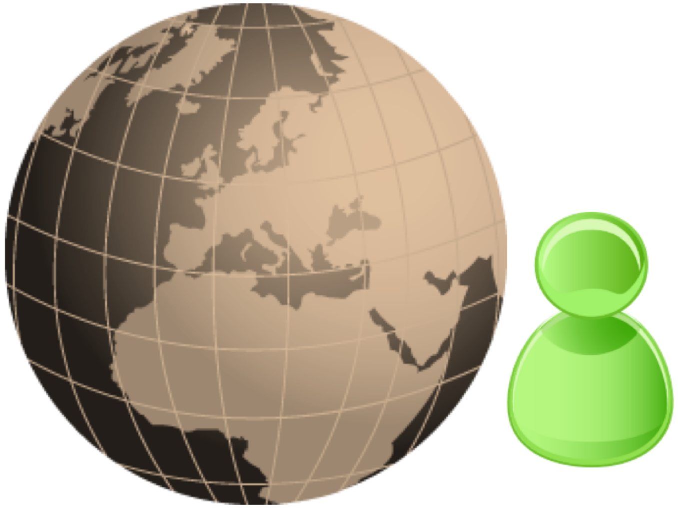
- In different group, keep confounders the same!



Ideal Solution: Counterfactual World

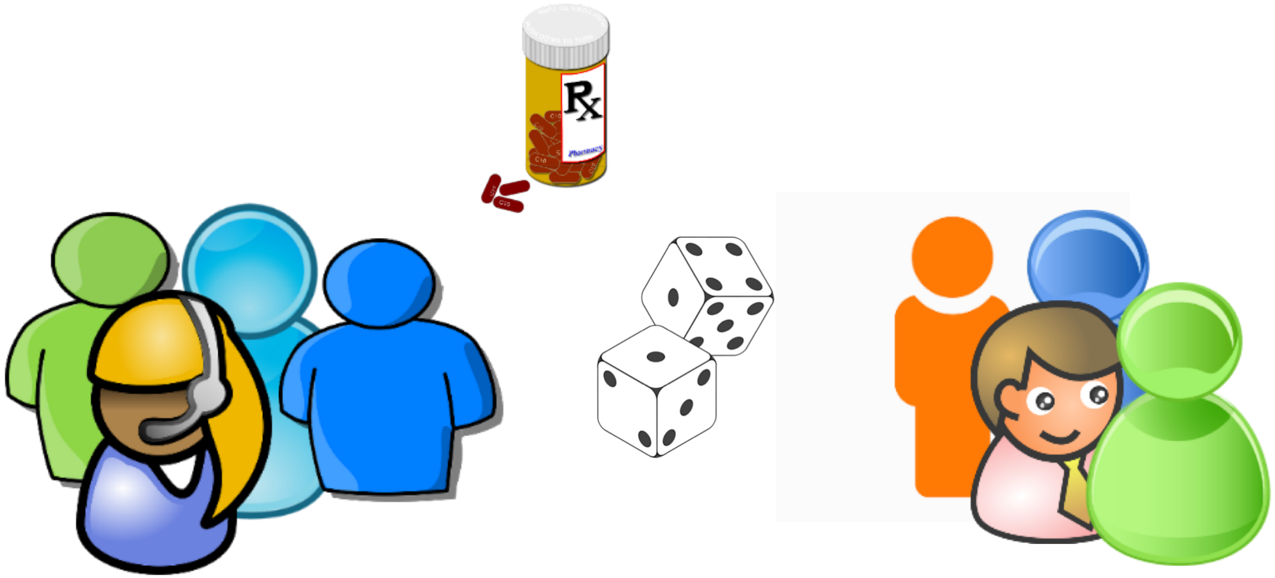
- Reason about a world that does not exist
- Everything is the same on real and counterfactual worlds, but the treatment

$Y(T = 1)$



$Y(T = 0)$

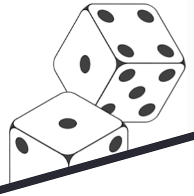
Randomized Experiments are the “Gold Standard”



- 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?

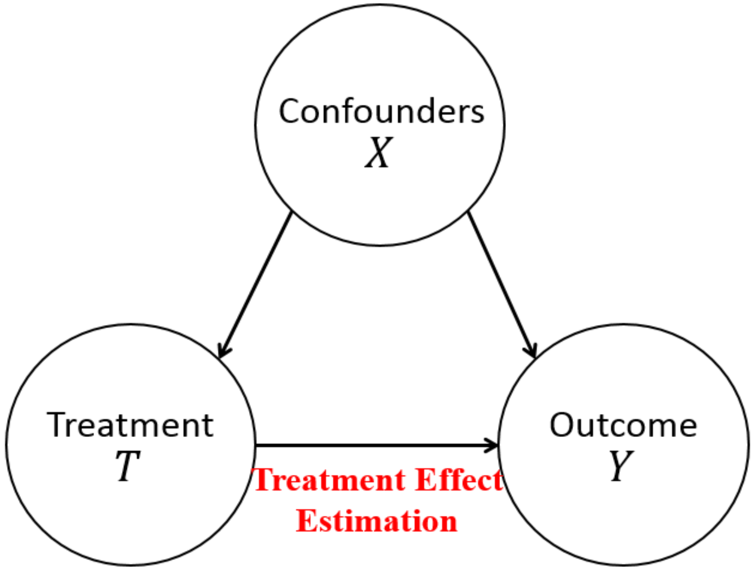
Observational Studies!

Recap: Causal Effect and Potential Outcome

- Two key points for causal effect estimation
  - Changing T
  - Keeping everything else (X) constant
- Counterfactual Problem

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

- Ideal Solution: Counterfactual World
- “Gold Standard”: Randomized Experiments
- We will discuss other solutions in Section 2.



OUTLINE

PART I. Introduction to Causal Inference

PART II. Methods for Causal Inference

PART III. Causally Regularized Machine Learning

PART IV. Benchmark and Open Datasets

PART V. Conclusion and Discussion

Causal Inference with Observational Data

- **Average Treatment Effect (ATE)** represents the mean (average) difference between the potential outcome of units under treated (T=1) and control (T=0) status.

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

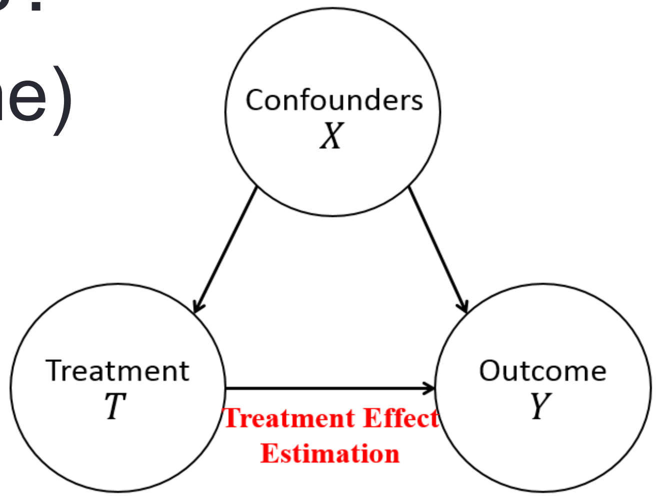
- Treated (T=1): taking a particular medication
- Control (T=0): not taking any medications
- **ATE:** the causal effect of the particular medication

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



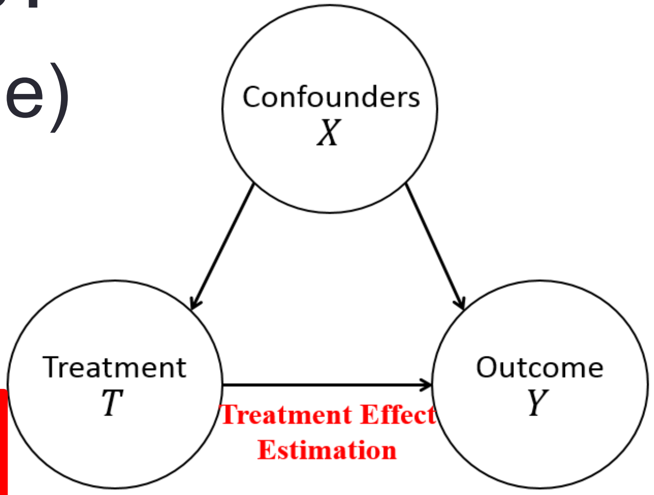
Causal Inference with Observational Data

- 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 (DCB)

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

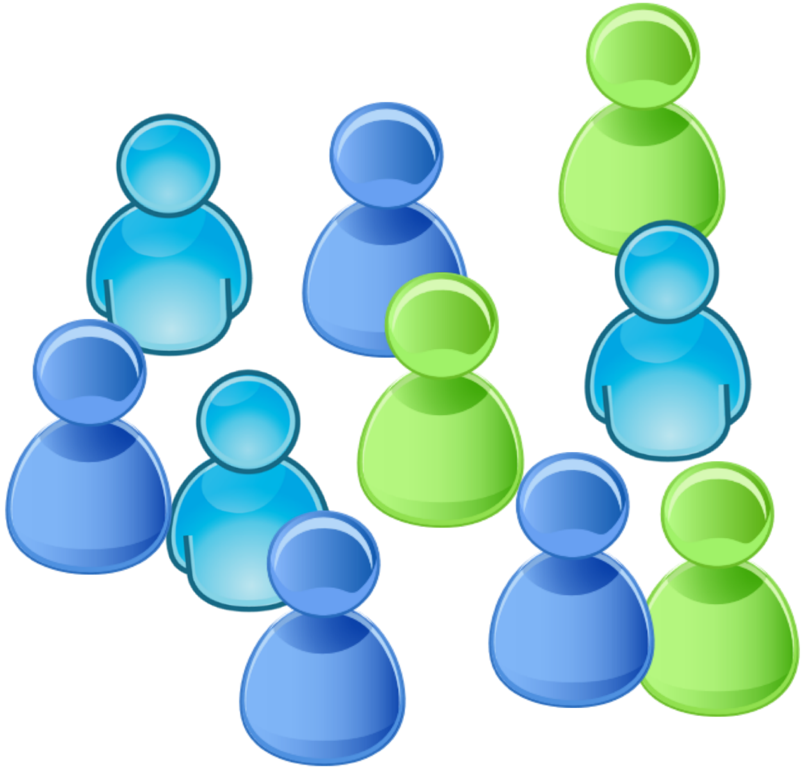
- **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

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

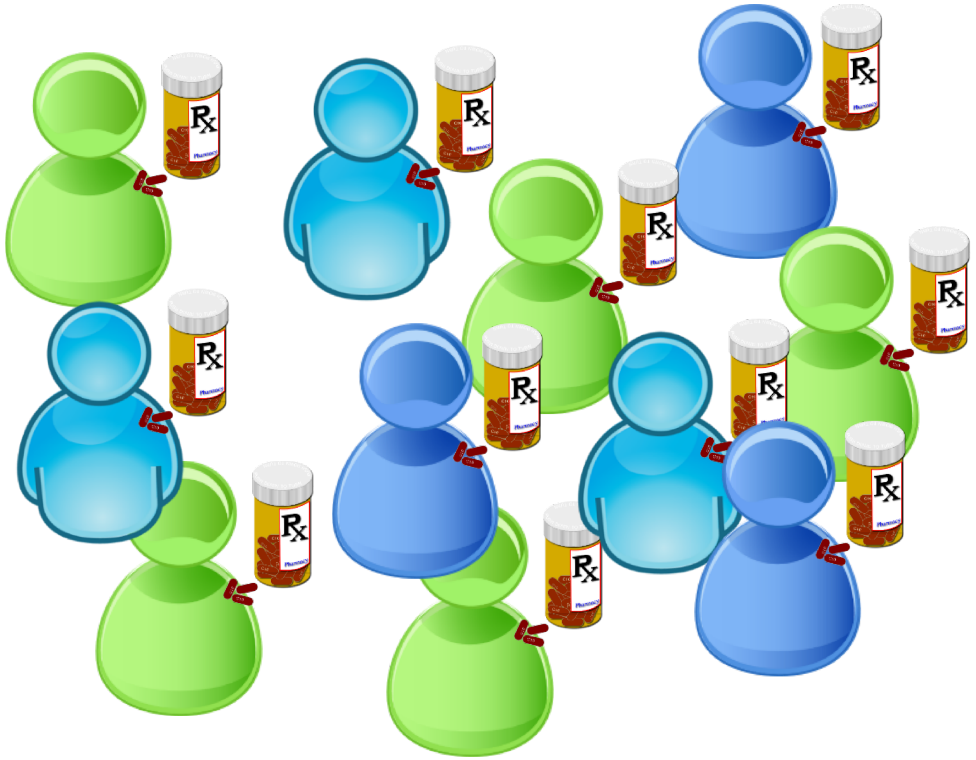
Methods for Causal Inference

- **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

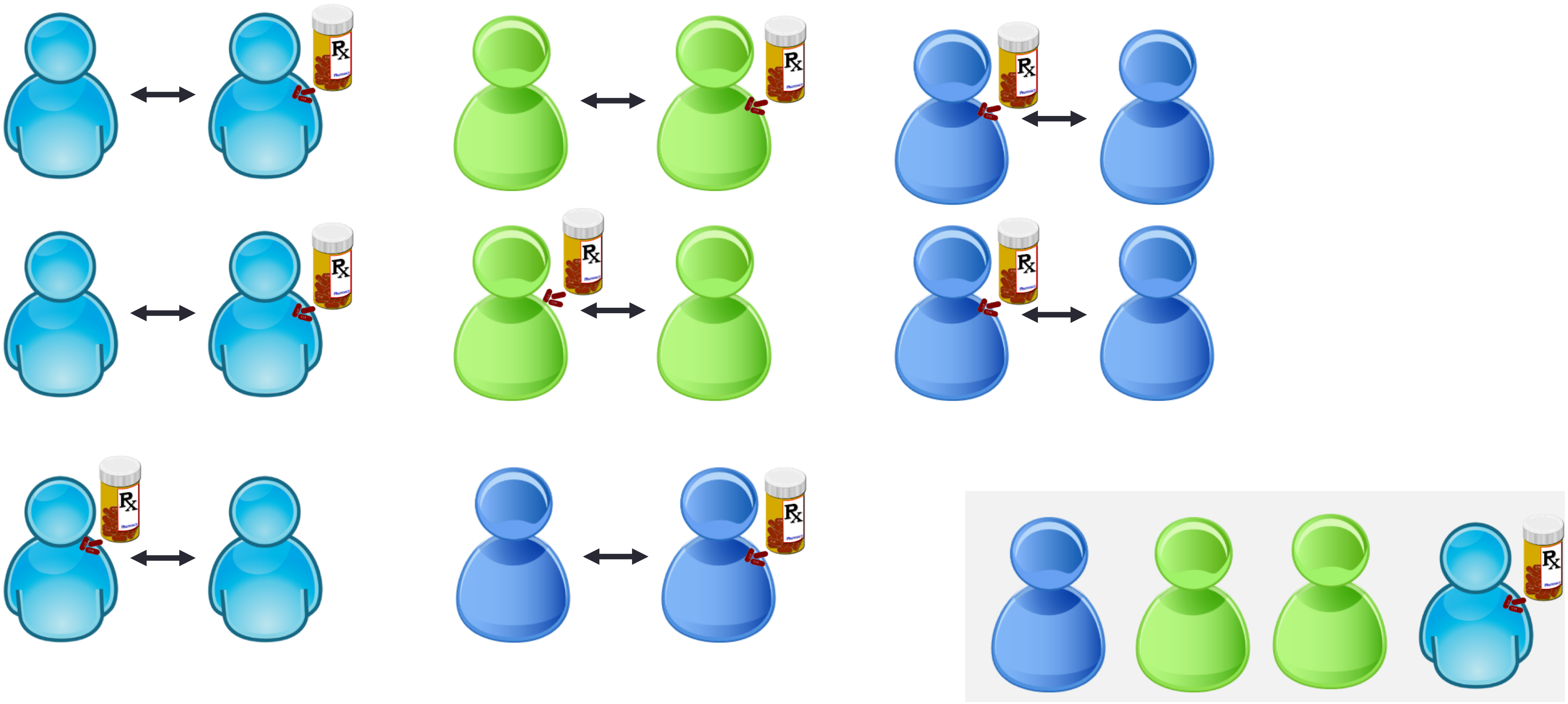


$T = 0$



$T = 1$

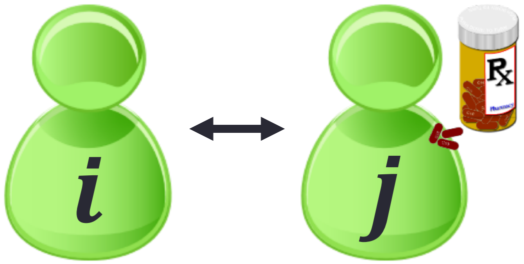
Matching



Matching

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

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

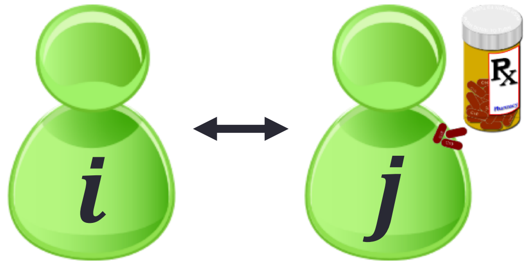


- Paired units provide the everything else (Confounders) approximate constant
- Average the difference in outcomes with in pairs to calculate the average causal effect
- Smaller ϵ: less bias, but higher variance

Matching

- Exactly Matching:

$$Distance(X_i, X_j) = \begin{cases} 0, & X_i = X_j \\ \infty, & X_i \neq X_j \end{cases}$$



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

- Use this in low-dimensional settings

- But in high-dimensional settings, there will be few exact matches

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

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

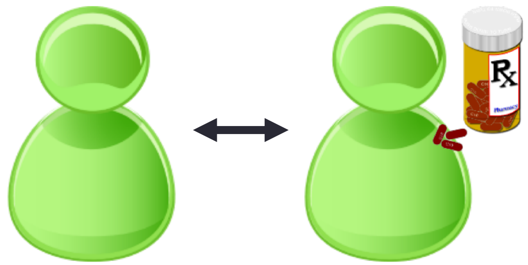
- Then, Rubin shows that the propensity score is sufficient to control or summarized the information of confounders

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

- 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

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

Inverse of Propensity Weighting (IPW)

- Why weighting with inverse of propensity score is helpful?
  - Propensity score induces the distribution bias on confounders X

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

Unit	$e(X)$	$1 - e(X)$	#units	#units (T=1)	#units (T=0)
A	0.7	0.3	10	7	3
B	0.6	0.4	50	30	20
C	0.2	0.8	40	8	32

Distribution Bias

Unit	#units (T=1)	#units (T=0)
A	10	10
B	50	50
C	40	40

Confounders are the same!

Reweighting by inverse of propensity score: $w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$

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

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

Doubly Robust

- Recap: $ATE = E[Y(T=1) - Y(T=0)]$
- Simple outcome regression:

  $$m_1 = E(Y|T=1, X) \quad \text{and} \quad m_0 = E(Y|T=0, X)$$

  - Unbiased if the regression models are correct
- IPW estimator:
  - Unbiased if the propensity score model is correct
- Doubly Robust [2]: combine both approaches

Doubly Robust

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

- Estimating ATE with Doubly Robust estimator:

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

  - *Unbiased* if either propensity score or regression model is correct
  - This property is referred to as *double robustness*

Doubly Robust

- Theoretical Proof:

$$
\begin{aligned}
& E\left[\frac{TY}{\hat{e}(X_i)} - \frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right] \\
= \quad & E\left[\frac{TY_1}{\hat{e}(X_i)} - \frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\hat{m}_1(X_i)\right] \\
= \quad & E\left[Y_1 + \frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\{Y_1 - \hat{m}_1(X_i)\}\right] \\
= \quad & E\left(Y_1\right) + \boxed{E\left[\frac{\{T-\hat{e}(X_i)\}}{\hat{e}(X_i)}\{Y_1 - \hat{m}_1(X_i)\}\right]}
\end{aligned}
$$

Doubly Robust

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

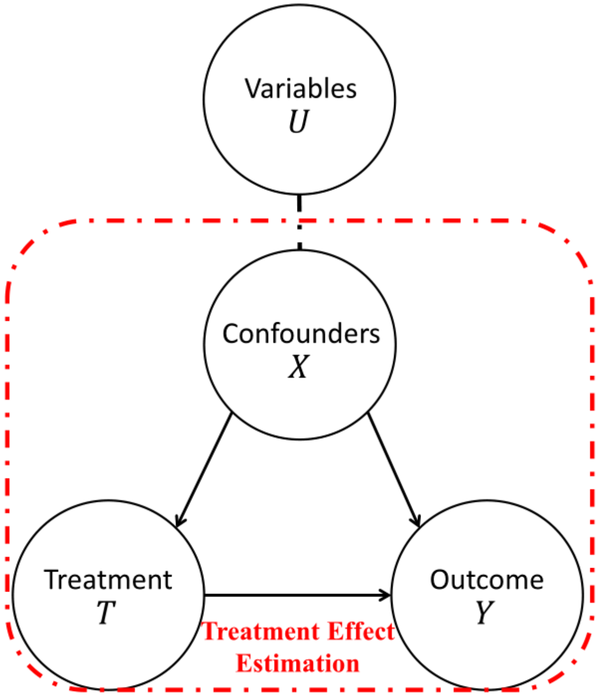
- Estimating ATE with Doubly Robust estimator:

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

  - *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

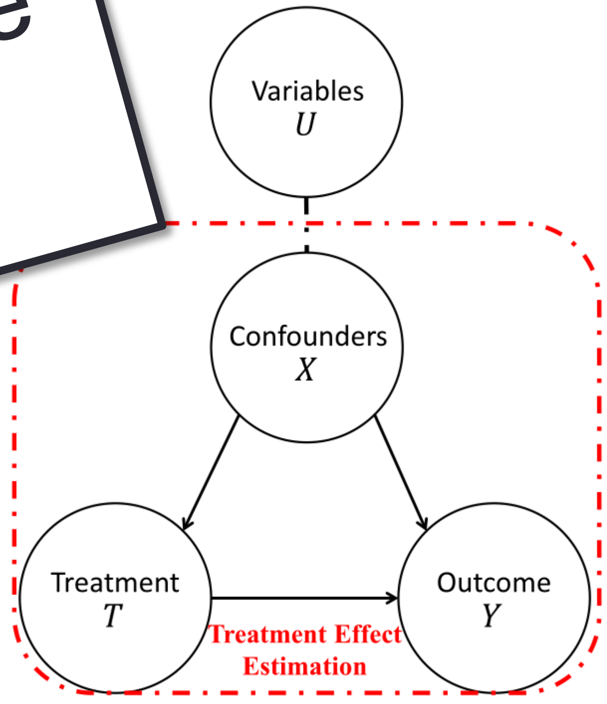


(a) Previous Causal Framework.

Propensity Score based Methods

- Recap:
  - Propensity Score Matching
  - Inverse of Propensity Weig
  - Doubly Robust
- Need to
  - Treat al as confounders
  - In Big Da , High dimensional data
  - But, not all variables are confounders

How to automatically separate the confounders?

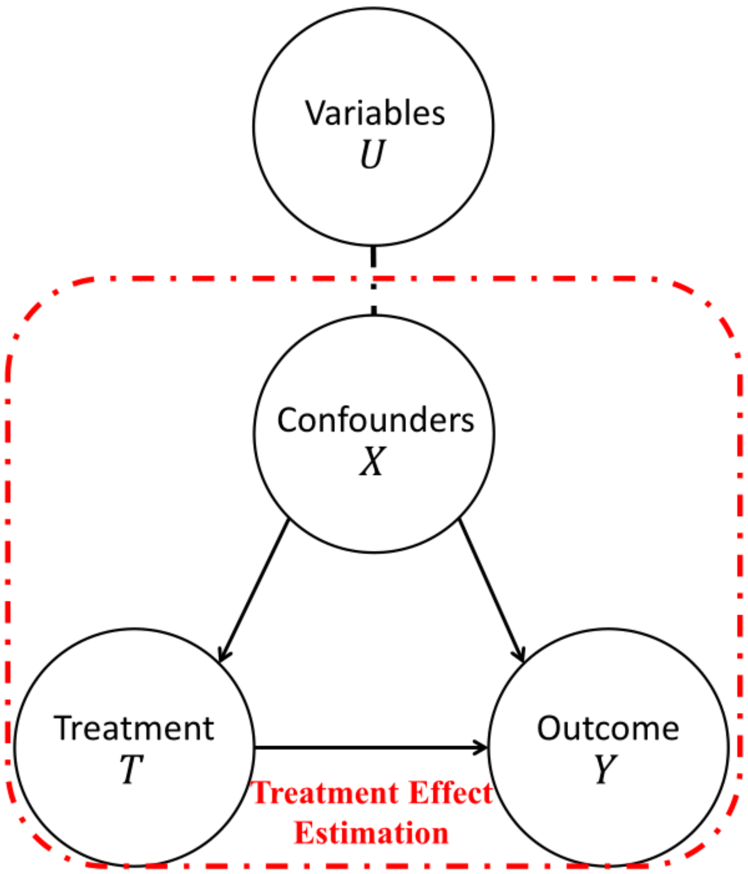


(a) Previous Causal Framework.

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)

Inverse of Propensity Weighting (IPW)



(a) Previous Causal Framework.

- Treat all observed variables **U** as confounders **X**
- Propensity Score Estimation:

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

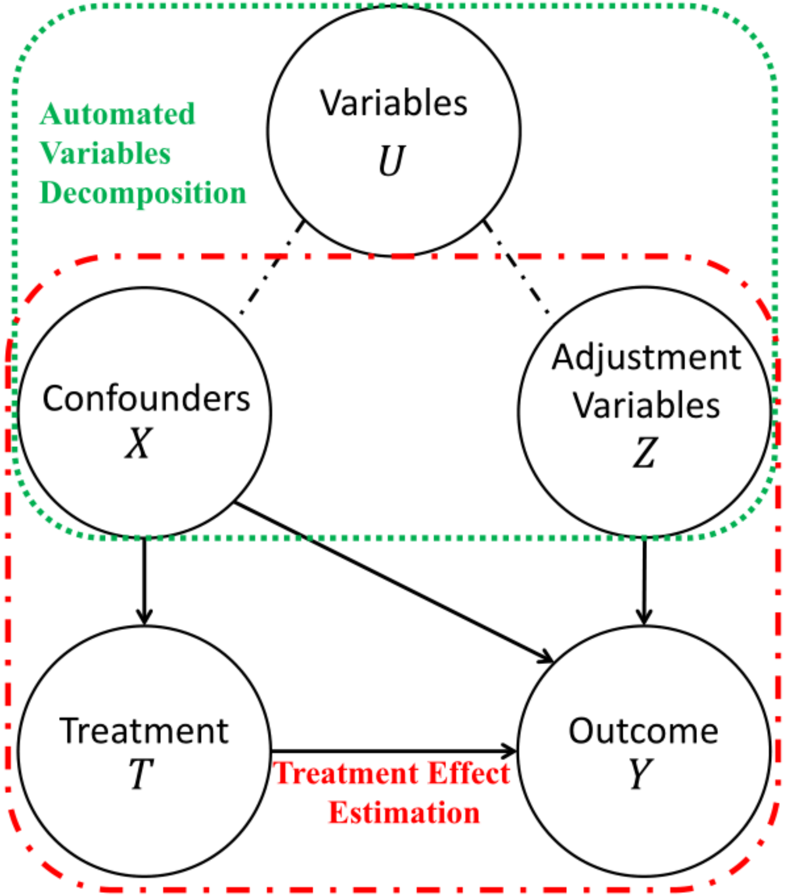
- Adjusted Outcome:

$$Y^{\star} = Y^{obs} \cdot \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} = Y^{obs} \cdot \frac{T - e(\mathbf{X})}{e(\mathbf{X}) \cdot (1 - e(\mathbf{X}))}$$

- IPW ATE Estimator:

$$\widehat{ATE}_{IPW} = \hat{E}(Y^{\star})$$

Data-Driven Variable Decomposition (D^2VD)



(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 D^2VD ATE Estimator:

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

Data-Driven Variable Decomposition (D²VD)

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

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

- By minimizing following objective function:

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

- We can estimate the ATE as:

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

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 $\quad h(\mathbf{U}) = \mathbf{U}\gamma,$

$$\begin{aligned} minimize \quad & \|(Y^{obs} - \mathbf{U}\alpha) \odot W(\beta) - \mathbf{U}\gamma\|_2^2, \quad \text{Where } W(\beta) := \frac{T - e(\mathbf{U})}{e(\mathbf{U}) \cdot (1 - e(\mathbf{U}))} \\ s.t. \quad & \sum_{i=1}^{m} \log(1 + \exp((1 - 2T_i) \cdot U_i\beta)) < \tau, \\ & \|\alpha\|_1 \le \lambda,\ \|\beta\|_1 \le \delta,\ \|\gamma\|_1 \le \eta,\ \boxed{\|\alpha \odot \beta\|_2^2 = 0.} \end{aligned}$$

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 $\widehat{ATE}_{IPW}$:*

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

Data-Driven Variable Decomposition (D^2VD)

- OUR: *Data-Driven Variable Decomposition* (**D^2VD**)
- Baselines
  - *Directly Estimator* (dir): ignores confounding bias
  - *IPW Estimator* (IPW): treats all variables as confounders
  - *Doubly Robust Estimator* (DR): IPW+regression
  - *Non-Separation Estimator* (D^2VD-): no variables separation

Data-Driven Variable Decomposition (D²VD)

- Dataset generation:
  - Sample size m={1000,5000}
  - Dimension of observed variables n={50,100,200}
  - Observed variables: $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I})$

    $$\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}, \mathbf{z}_1, \cdots, \mathbf{z}_{n_z}, \mathbf{i}_1, \cdots, \mathbf{i}_{n_i} \overset{iid}{\sim} \mathcal{N}(0,1),$$
  - Treatment: logistic and misspecified

    $$T_{logit} \sim Bernoulli(1/(1+\exp(-\textstyle\sum_{i=1}^{n_x} x_i))) \text{ and}$$
    $$T_{missp} = 1\; if\; \textstyle\sum_{i=1}^{n_x} x_i > 0.5,\; T_{missp} = 0\; otherwise.$$
  - Outcome:

    $$Y = \textstyle\sum_{j=\frac{nx}{2}}^{n_x} \mathbf{x}_j \cdot \omega_j + \sum_{j=1}^{n_z} \mathbf{z}_k \cdot \rho_k + T + \mathcal{N}(0,2),$$

Data-Driven Variable Decomposition (D²VD)

- Dataset generation:

The **true treatment effect** in synthetic data is **1**.

- Observed variables: $\mathbf{U} = (\mathbf{X}, \mathbf{Z}, \mathbf{I})$

$$\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}, \mathbf{z}_1, \cdots, \mathbf{z}_{n_z}, \mathbf{i}_1, \cdots, \mathbf{i}_{n_i} \overset{iid}{\sim} \mathcal{N}(0,1),$$

- Treatment: logistic and misspecified

$$T_{logit} \sim Bernoulli(1/(1+\exp(-\sum_{i=1}^{n_x} x_i))) \text{ and}$$
$$T_{missp} = 1 \; if \; \sum_{i=1}^{n_x} x_i > 0.5, \; T_{missp} = 0 \; otherwise.$$

- Outcome:

$$Y = \sum_{j=\frac{nx}{2}}^{n_x} \mathbf{x}_j \cdot \omega_j + \sum_{j=1}^{n_z} \mathbf{z}_k \cdot \rho_k + T + \mathcal{N}(0,2),$$

Data-Driven Variable Decomposition (D²VD)

- Experimental Results on Synthetic Data: $Bias = |\widehat{ATE} - ATE|$

	n	$n=50$				$n=100$				$n=200$			
T/m	Estimator	$Bias$	SD	MAE	RMSE	$Bias$	SD	MAE	RMSE	$Bias$	SD	MAE	RMSE
$T=T_{logit}$ $m=1000$	$\widehat{ATE}_{dir}$	0.418	0.409	0.479	0.582	0.302	0.490	0.472	0.571	0.405	0.628	0.574	0.720
	$\widehat{ATE}_{IPW}+lasso$	0.078	0.310	0.252	0.317	0.097	0.356	0.295	0.366	0.073	0.328	0.267	0.320
	$\widehat{ATE}_{DR}+lasso$	0.060	0.181	0.152	0.189	0.067	0.190	0.155	0.199	0.081	0.181	0.169	0.190
	$\widehat{ATE}_{D^2VD(-)}$	0.053	0.138	0.124	0.146	0.064	0.130	0.117	0.144	**0.018**	0.170	0.128	0.162
	$\widehat{ATE}_{D^2VD}$	**0.045**	**0.108**	**0.091**	**0.116**	**0.019**	**0.114**	**0.093**	**0.115**	0.067	**0.144**	**0.130**	**0.152**
$T=T_{logit}$ $m=5000$	$\widehat{ATE}_{dir}$	0.418	0.170	0.418	0.451	0.659	0.181	0.659	0.681	0.523	0.412	0.555	0.653
	$\widehat{ATE}_{IPW}+lasso$	0.036	0.201	0.163	0.202	0.034	0.222	0.194	0.213	**0.032**	0.341	0.274	0.325
	$\widehat{ATE}_{DR}+lasso$	0.051	0.079	0.071	0.094	0.106	0.075	0.114	0.127	0.055	0.084	0.086	0.096
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.080	0.118	0.137	0.114	0.102	0.121	0.150	0.164	0.076	0.164	0.179
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.072**	**0.061**	**0.078**	**0.023**	**0.073**	**0.061**	**0.073**	0.042	**0.068**	**0.062**	**0.076**
$T=T_{missp}$ $m=1000$	$\widehat{ATE}_{dir}$	0.664	0.387	0.670	0.766	0.273	0.445	0.436	0.518	0.380	0.766	0.691	0.848
	$\widehat{ATE}_{IPW}+lasso$	0.266	0.279	0.319	0.384	0.298	0.295	0.328	0.417	0.191	0.482	0.403	0.514
	$\widehat{ATE}_{DR}+lasso$	0.138	0.187	0.174	0.231	0.253	0.197	0.269	0.320	**0.050**	0.218	0.170	0.222
	$\widehat{ATE}_{D^2VD(-)}$	0.269	0.162	0.270	0.313	0.129	0.162	0.170	0.206	0.175	0.207	0.236	0.269
	$\widehat{ATE}_{D^2VD}$	**0.066**	**0.113**	**0.102**	**0.129**	**0.019**	**0.119**	**0.101**	**0.120**	0.059	**0.177**	**0.149**	**0.184**
$T=T_{missp}$ $m=5000$	$\widehat{ATE}_{dir}$	0.446	0.180	0.446	0.480	0.587	0.323	0.587	0.662	0.778	0.246	0.778	0.812
	$\widehat{ATE}_{IPW}+lasso$	0.148	0.133	0.161	0.198	0.172	0.167	0.199	0.239	0.142	0.224	0.206	0.263
	$\widehat{ATE}_{DR}+lasso$	0.119	0.073	0.123	0.139	0.100	0.067	0.107	0.120	0.127	0.079	0.127	0.148
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.070	0.119	0.132	0.058	**0.067**	0.069	0.086	0.068	0.055	0.073	0.086
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.055**	**0.052**	**0.063**	**0.039**	0.068	**0.066**	**0.075**	**0.032**	**0.047**	**0.049**	**0.055**

Data

- Exp

1. The direct estimator is failed under all settings.
2. IPW and DR estimators are good when $T=T_{logit}$, but poor when $T=T_{missp}$.
3. $D^2VD(-)$ has no variables separation, get similar results with DR estimator.
4. D^2VD can improve accuracy and reduce variance for ATE estimation.

$ATE|$

	n	$n=50$				$n=100$				$n=200$			
T/m	Estimator	$Bias$	SD	MAE	RMSE	$Bias$	SD	MAE	RMSE	$Bias$	SD	MAE	RMSE
$T=T_{logit}$ $m=1000$	$\widehat{ATE}_{dir}$	0.418	0.409	0.479	0.582	0.302	0.490	0.472	0.571	0.405	0.628	0.574	0.720
	$\widehat{ATE}_{IPW}+lasso$	0.078	0.310	0.252	0.317	0.097	0.356	0.295	0.366	0.073	0.328	0.267	0.320
	$\widehat{ATE}_{DR}+lasso$	0.060	0.181	0.152	0.189	0.067	0.190	0.155	0.199	0.081	0.181	0.169	0.190
	$\widehat{ATE}_{D^2VD(-)}$	0.053	0.138	0.124	0.146	0.064	0.130	0.117	0.144	**0.018**	0.170	0.128	0.162
	$\widehat{ATE}_{D^2VD}$	**0.045**	**0.108**	**0.091**	**0.116**	**0.019**	**0.114**	**0.093**	**0.115**	0.067	**0.144**	**0.130**	**0.152**
$T=T_{logit}$ $m=5000$	$\widehat{ATE}_{dir}$	0.418	0.170	0.418	0.451	0.659	0.181	0.659	0.681	0.523	0.412	0.555	0.653
	$\widehat{ATE}_{IPW}+lasso$	0.036	0.201	0.163	0.202	0.034	0.222	0.194	0.213	**0.032**	0.341	0.274	0.325
	$\widehat{ATE}_{DR}+lasso$	0.051	0.079	0.071	0.094	0.106	0.075	0.114	0.127	0.055	0.084	0.086	0.096
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.080	0.118	0.137	0.114	0.102	0.121	0.150	0.164	0.076	0.164	0.179
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.072**	**0.061**	**0.078**	**0.023**	**0.073**	**0.061**	**0.073**	0.042	**0.068**	**0.062**	**0.076**
$T=T_{missp}$ $m=1000$	$\widehat{ATE}_{dir}$	0.664	0.387	0.670	0.766	0.273	0.445	0.436	0.518	0.380	0.766	0.691	0.848
	$\widehat{ATE}_{IPW}+lasso$	0.266	0.279	0.319	0.384	0.298	0.295	0.328	0.417	0.191	0.482	0.403	0.514
	$\widehat{ATE}_{DR}+lasso$	0.138	0.187	0.174	0.231	0.253	0.197	0.269	0.320	**0.050**	0.218	0.170	0.222
	$\widehat{ATE}_{D^2VD(-)}$	0.269	0.162	0.270	0.313	0.129	0.162	0.170	0.206	0.175	0.207	0.236	0.269
	$\widehat{ATE}_{D^2VD}$	**0.066**	**0.113**	**0.102**	**0.129**	**0.019**	**0.119**	**0.101**	**0.120**	0.059	**0.177**	**0.149**	**0.184**
$T=T_{missp}$ $m=5000$	$\widehat{ATE}_{dir}$	0.446	0.180	0.446	0.480	0.587	0.323	0.587	0.662	0.778	0.246	0.778	0.812
	$\widehat{ATE}_{IPW}+lasso$	0.148	0.133	0.161	0.198	0.172	0.167	0.199	0.239	0.142	0.224	0.206	0.263
	$\widehat{ATE}_{DR}+lasso$	0.119	0.073	0.123	0.139	0.100	0.067	0.107	0.120	0.127	0.079	0.127	0.148
	$\widehat{ATE}_{D^2VD(-)}$	0.112	0.070	0.119	0.132	0.058	**0.067**	0.069	0.086	0.068	0.055	0.073	0.086
	$\widehat{ATE}_{D^2VD}$	**0.033**	**0.055**	**0.052**	**0.063**	**0.039**	0.068	**0.066**	**0.075**	**0.032**	**0.047**	**0.049**	**0.055**

Data-Driven Variable Decomposition (D^2VD)

- Experimental Results on Synthetic Data:

Table 3: Separation results of confounders **X** and adjustment variables **Z**. The closer to **1** for TPR and TNR is better.

$\mathbf{T} = \mathbf{T}_{\text{logit}}$							
		$n = 50$		$n = 100$		$n = 200$	
m		TPR	TNR	TPR	TNR	TPR	TNR
$m = 1000$	**X**	1.000	0.917	0.977	0.948	0.966	0.906
	Z	1.000	0.973	1.000	0.983	1.000	0.984
$m = 5000$	**X**	1.000	0.923	1.000	0.887	0.994	0.989
	Z	1.000	0.975	1.000	0.987	1.000	0.994
$\mathbf{T} = \mathbf{T}_{\text{missp}}$							
$m = 1000$	**X**	1.000	0.844	0.997	0.866	0.867	0.977
	Z	1.000	0.982	1.000	0.987	1.000	0.983
$m = 5000$	**X**	1.000	0.843	1.000	0.837	0.998	0.965
	Z	1.000	0.986	1.000	0.990	1.000	0.994

TPR: true positive rate
TNR: true negative rate

Our D^2VD algorithm can precisely separate the confounders and adjustment variables.

Experiments on Real World Data

- Dataset Description:
  - Online advertising campaign (LONGCHAMP)
  - Users Feedback: 14,891 LIKE; 93,108 DISLIKE
  - 56 Features for each user
    - Age, gender, #friends, device, user setting on WeChat

WeChat

2015

- Experimental Setting:
  - Outcome Y: users feedback
  - Treatment T: one feature
  - Observed Variables U: other features

Y = 1, if LIKE
Y = 0, if DISLIKE

Experiments Results

- ATE Estimation.

No.	Features	$\widehat{ATE}_{D^2VD}$ (SD)	$\widehat{ATE}_{IPW}$ (SD)	$\widehat{ATE}_{DR}$ (SD)	$ATE_{matching}$
1	No. friends (> 166)	0.295 (0.018)	0.240 (0.026)	0.297(0.021)	0.276
2	Age (> 33)	-0.284 (0.014)	-0.235 (0.029)	-0.302(0.068)	-0.263
3	Share Album to Strangers	0.229 (0.030)	0.236 (0.030)	-0.034(0.021)	n/a
4	With Online Payment	0.226 (0.019)	0.260 (0.029)	0.244(0.028)	n/a
5	With High-Definition Head Portrait	0.218 (0.028)	0.203 (0.032)	0.237(0.046)	n/a
6	With WeChat Album	0.191 (0.014)	0.237 (0.021)	0.097(0.050)	n/a
7	With Delicacy Plugin	0.124 (0.038)	-0.253 (0.037)	0.067(0.051)	0.099
8	Device (iOS)	0.100 (0.024)	0.206 (0.012)	0.060(0.021)	0.085
9	Add friends by Drift Bottle	-0.098 (0.012)	0.016 (0.019)	-0.115(0.015)	-0.032
10	Gender (Male)	-0.073 (0.017)	-0.240 (0.029)	0.065(0.055)	-0.097

1. Our D^2VD estimator evaluate the ATE more accuracy.
2. Our D^2VD estimator can reduce the variance of estimated ATE.
3. Younger Ladies are with higher probability to like the LONGCHAMP ads.

Experiments Results

- Variables Decomposition.

Table 4: Confounders and adjusted variables when we set feature "Add friends by Shake" as treatment.

Confounders	Adjustment Variables
Add friends by Drift Bottle	No. friends
Add friends by People Nearby	Age
Add friends by QQ Contacts	With WeChat Album
Without Friends Confirmation Plugin	Device

1. The confounders are many other ways for adding friends on WeChat.
2. The adjustment variables have significant effect on outcome.
3. Our D^2VD algorithm can precisely separate the confounders and adjustment variables.

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)

Causal Inference with Observational Data

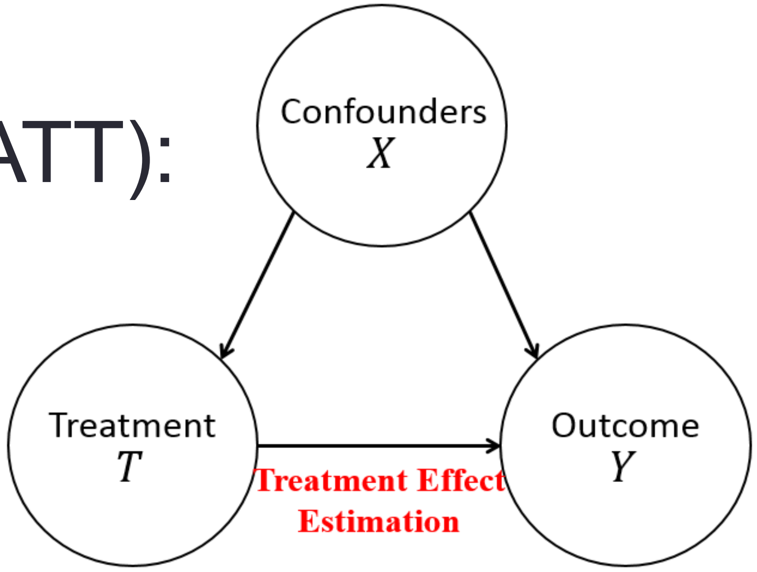
- Average Treatment Effect (ATE):

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

- Average Treatment effect on the Treated (ATT):

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

- Two key points:
  - Changing T (T=1 and T=0)
  - Keeping everything else (Confounder X) constant



Causal Inference with Observational Data

- Average Treatment Effect (ATE):

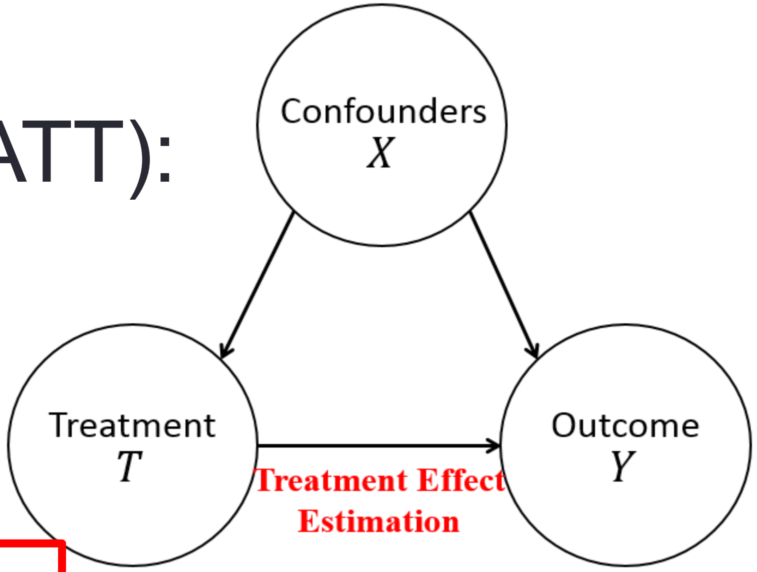
$$ATE = E[Y(T=1) - Y(T=0)]$$

- Average Treatment effect on the Treated (ATT):

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

- Two key points:

$\boxed{\textbf{Balancing Confounders' Distribution}}$ t



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

$$\min_{W} \quad W \log(W)$$
$$s.t. \quad \boxed{\|\overline{\mathbf{X}}_t - \mathbf{X}_c^T W\|_2^2 = 0}$$
$$\textstyle\sum_{i=1}^{n} W_i = 1, W \succeq 0$$

- 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

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

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

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

- 3. Estimate the ATT as

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

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

Directly Confounder Balancing

- Recap:
  - *Entropy Balancing*, *Approximate Residual Balancing* etc.
  - Moments uniquely determine variables' distribution
  - Learning sample weights by balancing confounders' moments

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

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

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

- But, treat all variables as confounders, and balance them equally
- Different confounders make different confounding bias

Directly Confounder Balancing

- Recap:
  - *Entropy Balancing*, *Approximate Residual Balancin*... etc.
  - Moments uniquely determine variables'...
  - Learning sample weights b... ...oments

How to differentiated confounders and their bias?

...p

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

- But, treat a... variables as confounders, and balance them equally
- Different confounders make different confounding bias

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.

How to learn the these weights?

Confounder Weights Learning

- General relationship among X, T, and Y:

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

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

Confounder weights

Confounding bias

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

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.

$$\min \quad \Big(\beta^T \cdot \big(\boxed{\overline{\mathbf{M}}_t} - \boxed{\mathbf{M}_c^T W}\big)\Big)^2$$

Confounders' moments on the Treated Group

Confounders' moments on the Control Group

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.

Differentiated Confounder Balancing

- Algorithm

Algorithm 1 Differentiated Confounder Balancing (DCB)

Input: Tradeoff parameters $\lambda > 0$, $\delta > 0$, $\mu > 0$, $\nu > 0$, Augmented Variables Matrix on treat units $\mathbf{M}_t$, Augmented Variables Matrix on control units $\mathbf{M}_c$ and Outcome Y.
Output: Confounder Weights β and Sample Weights W

```
Initialize Confounder Weights β^(0) and Sample Weights W^(0)
Calculate the current value of J(W, β)^(0) = J(W^(0), β^(0)) with Equation (11)
Initialize the iteration variable t ← 0
repeat
    t ← t + 1
    Update β^(t) by solving J(β^(t−1)) in Equation (12)
    Update W^(t) by solving J(W^(t−1)) in Equation (13)
    Calculate J(W, β)^(t) = J(W^(t), β^(t))
until J(W, β)^(t) converges or max iteration is reached
return β, W.
```

$$\begin{aligned}\mathcal{J}(\beta) &= \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2 + \mu\|\beta\|_2^2 + \nu\|\beta\|_1 \\ &\quad + \lambda \textstyle\sum_{j:T_j=0}(1 + W_j) \cdot (Y_j - M_j \cdot \beta)^2\end{aligned} \tag{12}$$

$$\begin{aligned}\mathcal{J}(W) &= \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2 + \delta\|W\|_2^2 \\ &\quad + \lambda \textstyle\sum_{j:T_j=0}(1 + W_j) \cdot (Y_j - M_j \cdot \beta)^2, \\ s.t. &\quad \mathbf{1}^T W = 1 \quad and \quad W \succeq 0.\end{aligned} \tag{13}$$

In each iteration, we first update β by fixing W, and then update W by fixing β

- Training Complexity: $O(np)$
  - n: sample size, p: dimensions of variables

Experiments

- Experimental Tasks:
  - Robustness Test (high-dimensional and noisy)
  - Accuracy Test (real world dataset)
  - Predictive Power Test (real ad application)

Experiments

- Baselines:
  - **Directly Estimator**: comparing average outcome between treated and control units.
  - **IPW Estimator** [1]: reweighting via inverse of propensity score
  - **Doubly Robust Estimator** [2]: IPW + regression method
  - **Entropy Balancing Estimator** [3]: directly confounder balancing with entropy loss
  - **Approximate Residual Balancing** [4]: confounder balancing + regression

- Evaluation Metric:

$$\begin{aligned}
Bias &= \left|\frac{1}{K}\sum_{k=1}^{K}\widehat{ATT}_k - ATT\right| \\
SD &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATT}_k - \frac{1}{K}\sum_{k=1}^{K}\widehat{ATT}_k)^2} \\
MAE &= \frac{1}{K}\sum_{k=1}^{K}\left|\widehat{ATT}_k - ATT\right| \\
RMSE &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{ATT}_k - ATT)^2}
\end{aligned}$$

Experiments - Robustness Test

- Dataset
  - Sample size: $n = \{2000, 5000\}$
  - Variables' dimensions: $p = \{50,100\}$
  - **Observed Variables**: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p \overset{iid}{\sim} \mathcal{N}(0,1),$$

  - **Treatment**: from logistic function T_{logit} and misspecified function T_{missp}

$$T_{logit} \sim Bernoulli(1/(1+\exp(-\textstyle\sum_{i=1}^{p\cdot r_c} s_c \cdot x_i + \mathcal{N}(0,1)))), and$$
$$T_{missp} = 1\ if\ \textstyle\sum_{i=1}^{p\cdot r_c} s_c \cdot x_i + \mathcal{N}(0,1) > 0,\ T_{missp} = 0\ otherwise$$

    - Confounding rate r_c: the ratio of confounders to all observed variables.
    - Confounding strength s_c: the bias strength of confounders
  - **Outcome**: from linear function Y_{linear} and nonlinear function Y_{nonlin}

$$Y_{linear} = T + \textstyle\sum_{j=1}^{p}\{I(mod(j,2) \equiv 0) \cdot (\frac{j}{2} + T) \cdot \mathbf{x}_j\} + \mathcal{N}(0,3),$$
$$Y_{nonlin} = T + \textstyle\sum_{j=1}^{p}\{I(mod(j,2) \equiv 0) \cdot (\frac{j}{2} + T) \cdot \mathbf{x}_j\} + \mathcal{N}(0,3)$$
$$+ \textstyle\sum_{j=1}^{p-1}\{I(mod(j,10) \equiv 1) \cdot \frac{p}{2} \cdot (x_j^2 + x_j \cdot x_{j+1})\},$$

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!

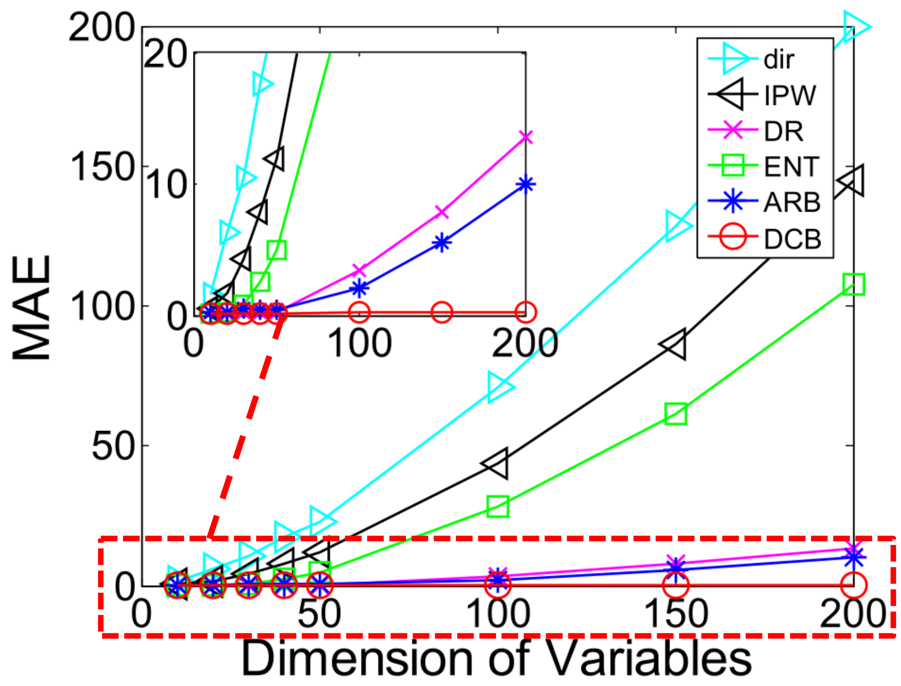
	n/p	$n = 2000, p = 50$			$n = 2000, p = 100$		
r_c	Estimator	$Bias$ (SD)	MAE	RMSE	$Bias$ (SD)	MAE	RMSE
$r_c = 0.8$	$\widehat{ATT}_{dir}$	51.06 (3.725)	51.06	51.19	143.0 (9.389)	143.0	143.3
	$\widehat{ATT}_{IPW}$	29.99 (4.048)	29.99	30.26	98.24 (8.462)	98.24	98.60
	$\widehat{ATT}_{DR}$	0.345 (0.253)	0.367	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 - Robustness Test

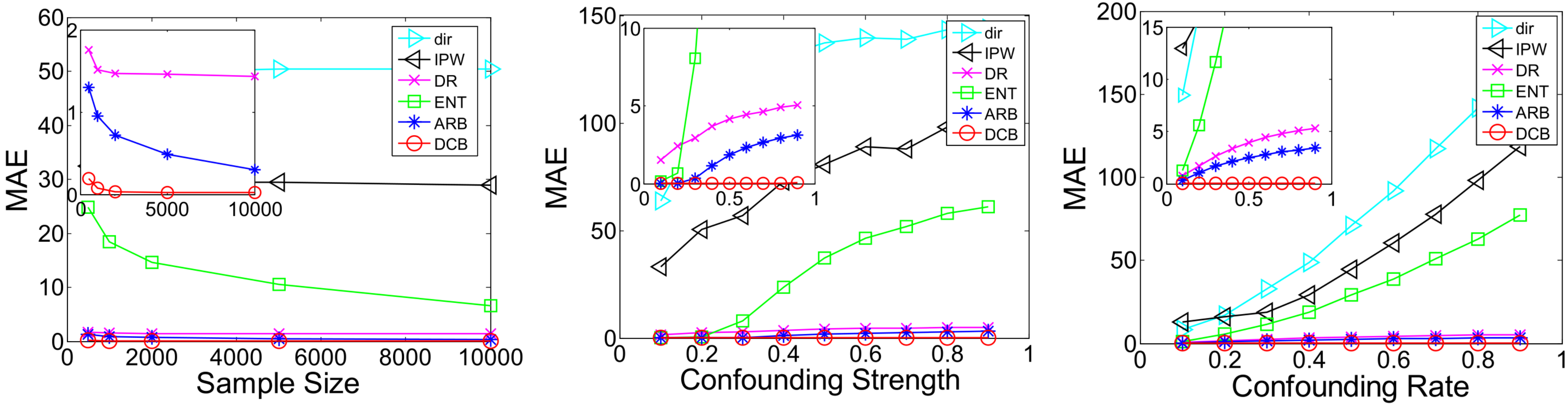
- Sample Size
- Dimension of variables
- Confounding rate
- Confounding strength



(b) dimension of variables p

The MAE of our DCB estimator is consistent stable and small.

Experiments - Robustness Test



(a) sample size n **(d) confounding strength s_c** **(c) confounding rate r_c**

Our DCB algorithm is very robust for treatment effect estimation.

Experiments - Accuracy Test

- LaLonde Dataset [5]: *Would the job training program increase people's earnings in the year of 1978?*
  - **Randomized experiments**: provide ground truth of treatment effect
  - **Observational studies**: check the performance of all estimators

- Experimental Setting:
  - **V-RAW**: variables set of 10 raw observed variables, including employment, education, age ethnicity and married status.
  - **V-INTERACTION**: variables set of raw variables, their pairwise one way interaction and their squared terms.

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.

Experiments - Predictive Power

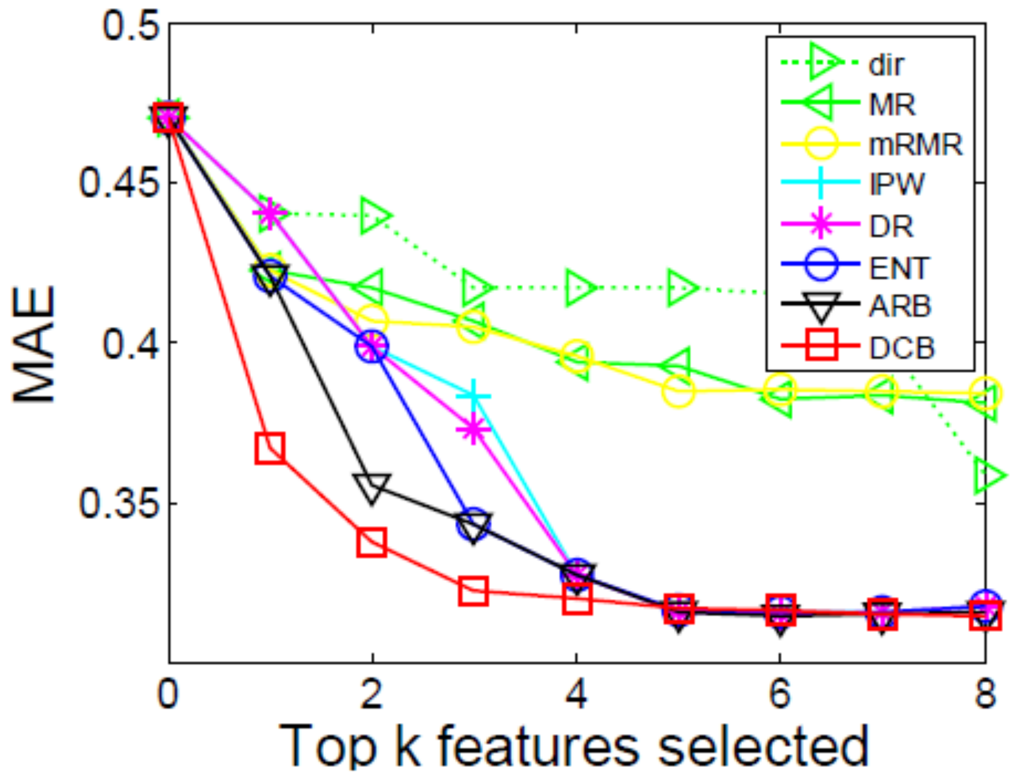
2015 WeChat

- Dataset Description:
  - Online advertising campaign (LONGCHAMP)
  - Users Feedback: 14,891 LIKE; 93,108 DISLIKE
  - 56 Features for each user
    - Age, gender, #friends, device, user setting on WeChat

- Experimental Setting:
  - Outcome Y: users feedback ⟵ $Y = 1$, if LIKE; $Y = 0$, if DISLIKE
  - Treatment T: one feature

Select the top k features with high causal effect for prediction

Experiments - Predictive Power



- Two correlation-based feature selection baselines:
  - ***MRel*** ***[6]:*** maximum relevance
  - ***mRMR*** ***[7]:*** Maximum relevance and minimum redundancy.

➢Our DCB estimator achieves the best prediction accuracy.
➢Correlation based methods perform worse than causal methods.

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

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
  - Data-Driven Variable Decomposition (D^2VD)
- **Directly Confounder Balancing**
  - Entropy Balancing
  - Approximate Residual Balancing
  - Differentiated Confounder Balancing (DCB)

Treat all observed variables as confounder

Not all observed variables are confounders

Balance all confounder equally

Different confounders make different bias

OUTLINE

OUTLINE

PART I. Introduction to Causal Inference

PART II. Methods for Causal Inference

PART III. Causally Regularized Machine Learning

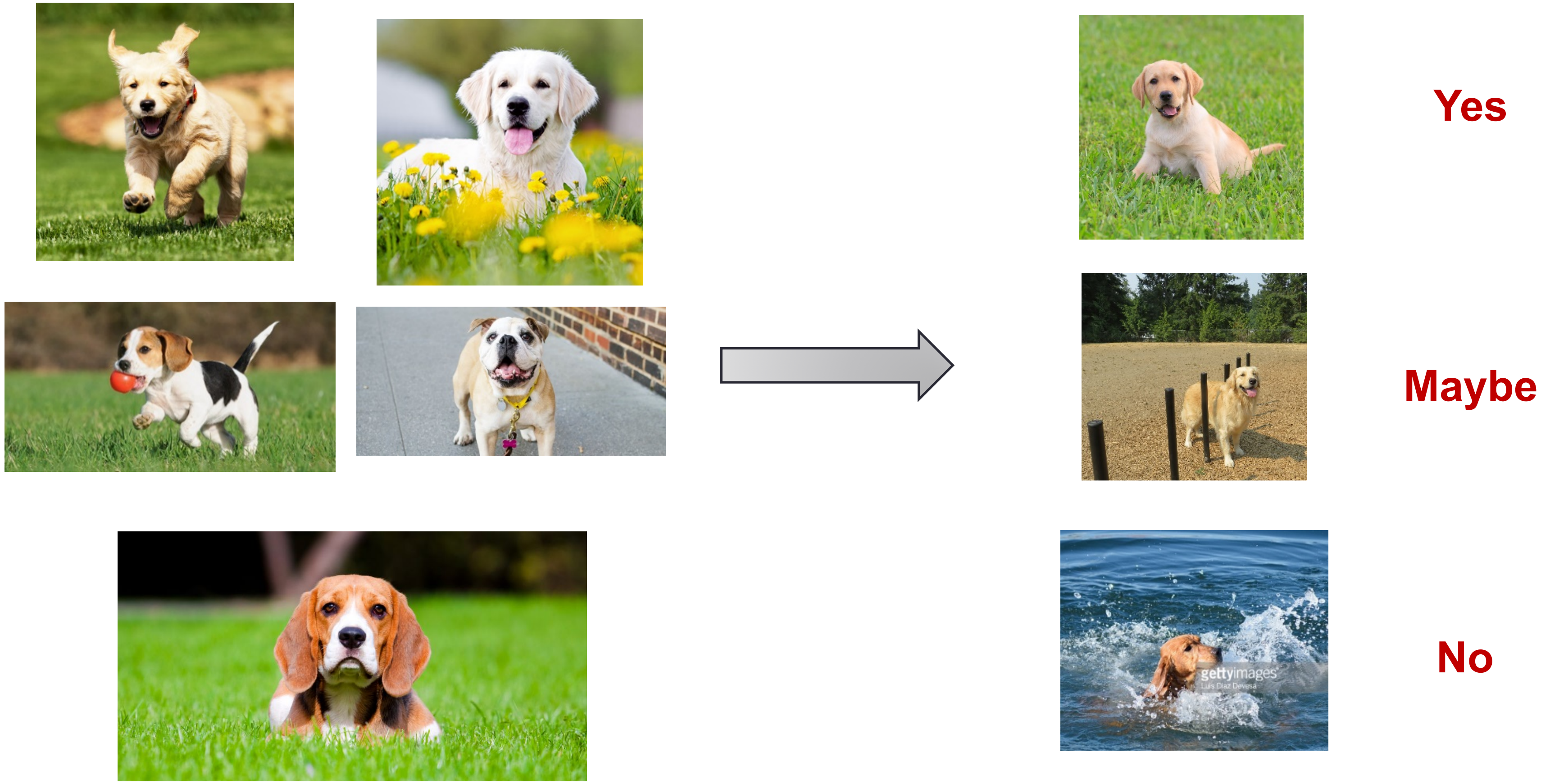
- **Causal Inference for Stable Prediction**
- **Causal Inference for Offline Policy Evaluation**

PART IV. Benchmark and Open Datasets

PART V. Conclusion and Discussion

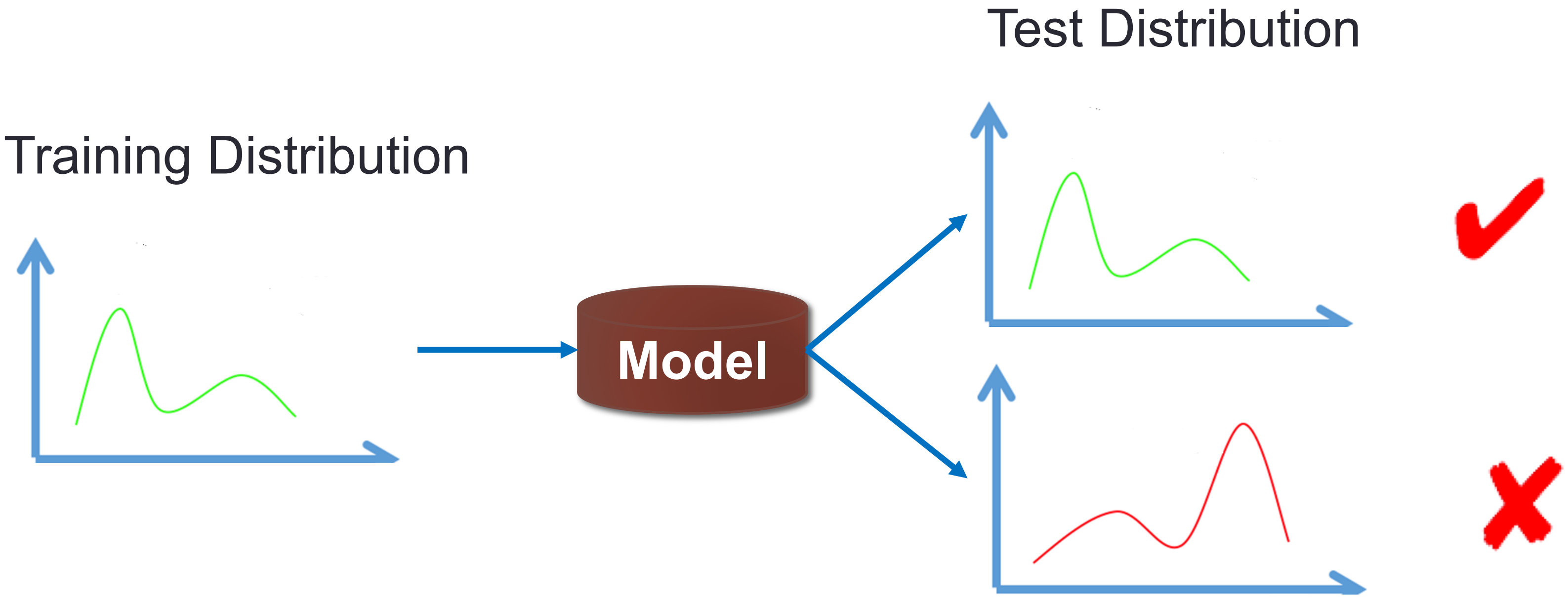
Causal Inference for Stable Prediction

- CAN and CANNOT of predictive models



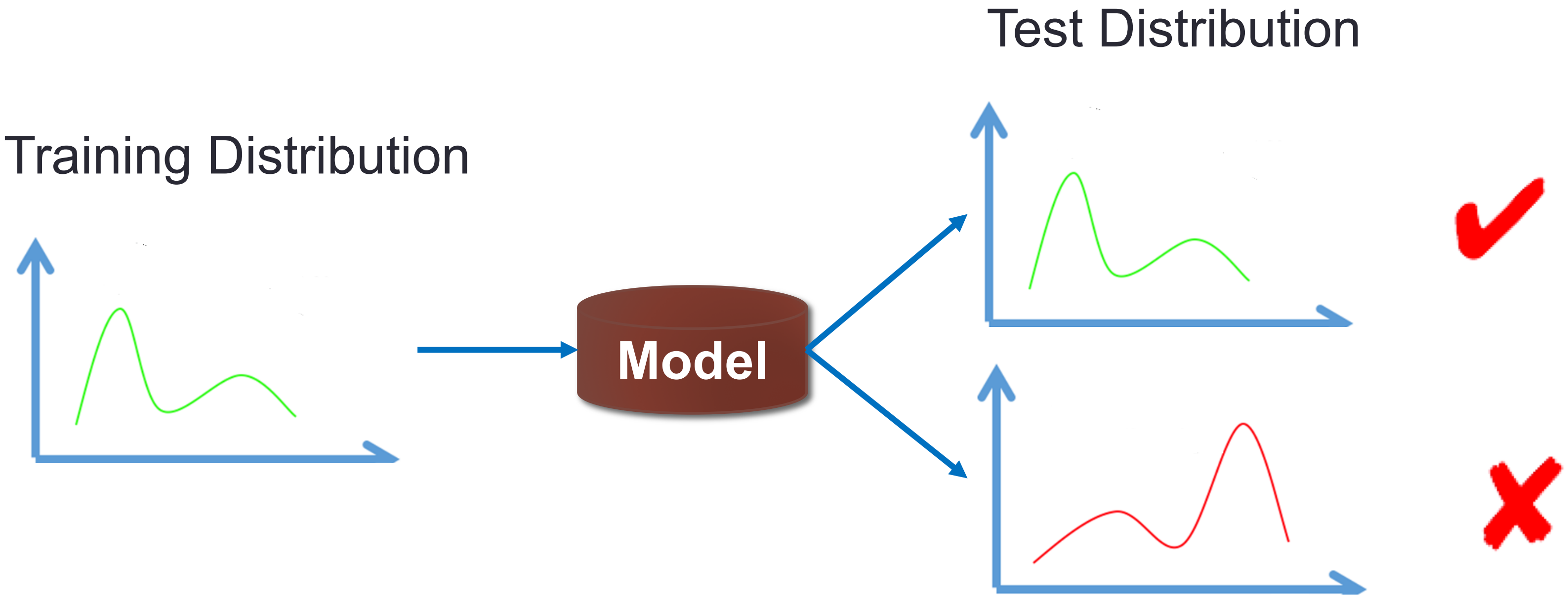
Why they fail?

- The fault of Data
  - IID hypothesis (violated often)
  - Sample selection bias result in distribution shift
  - More serious in small-sample learning
  - We CANNOT control the generation of testing data



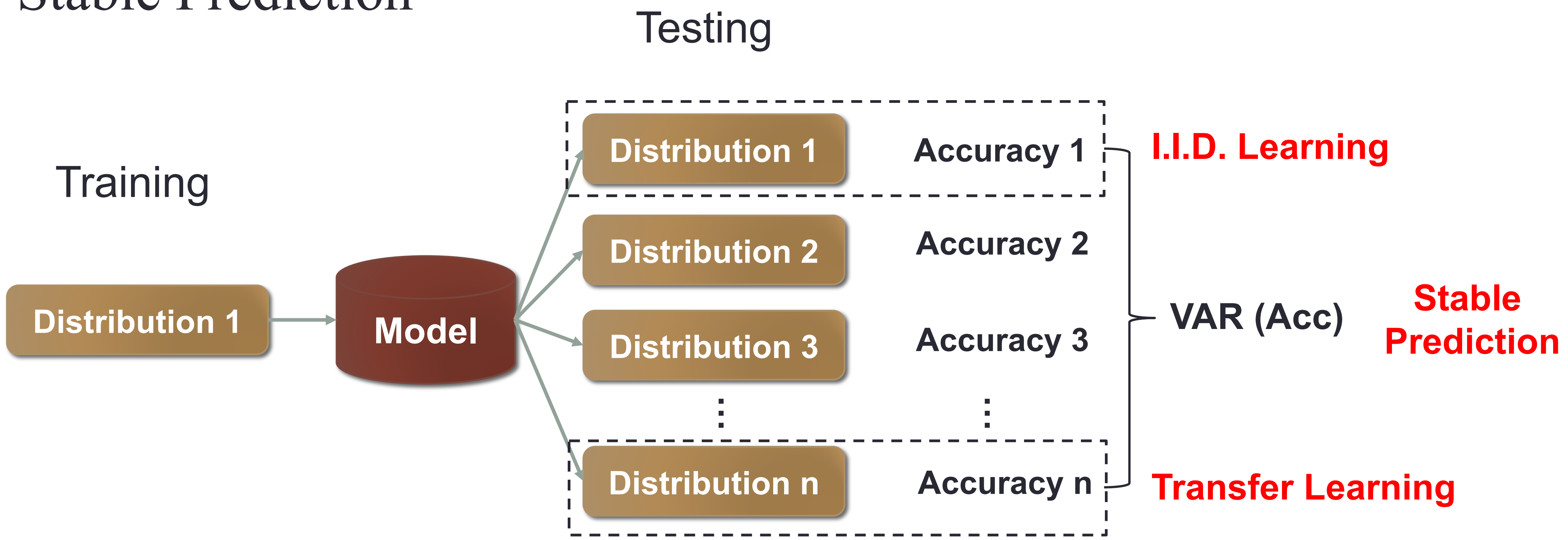
Why they fail?

- The fault of Model
  - Correlation based model
  - Three sources of correlation: **Causation**, Confounding, and Selection Bias (**Invariant Causation** and Spurious Correlation)
  - Idea: Causally Regularized Stable Learning



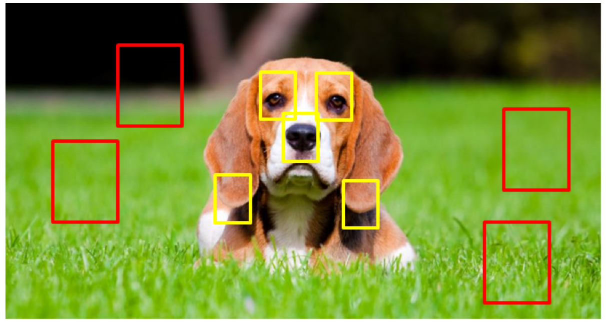
Stable Prediction

- Stable Prediction



Stable Prediction across Unknown Testing Data

Why would a predictive model not be stable?



- Prediction / Classification
  - X: vector of features; $Y = \{0,1\}$
  - Environment: joint distribution of X and Y, denoted as $P(XY)$
- Suppose $X = \{S, V\}$, and $Y = f(S) + \varepsilon$
  - S: set of stable (causal) features
  - V: set of non-causal features
  - $P(Y|S)$ is stable, but $P(Y|V)$ is not stable
- **Why would a predictive model not be stable?**
  - **Dependence issue**, Y is not independent with V (Spurious Correlation)
  - **Environment shift issue**, $P(XY)_{training} \neq P(XY)_{testing}$

Why would a predictive model not be stable?

- **Dependence issue**
  - $X = \{S, V\}$, and $Y = f(S) + \varepsilon$
  - Diagram (b) & (c):
    - Y is not independent with V
  - Diagram (a): $Y \perp V$
    - Selection bias, leading to Y is not independent with V
  - **Some v $\subseteq V$ would be learned as important predictors**
- **Environment shift issue**
  - $P(XY) = P(Y|X)P(X) = P(Y|S)P(X)$ (assume P(Y|S) is stable)
  - Selection bias → $P(X)_{training} \neq P(X)_{testing}$

    Y is not independent with V ⇨ $Corr(V_{training}, Y_{training}) \neq Corr(V_{testing}, Y_{testing})$

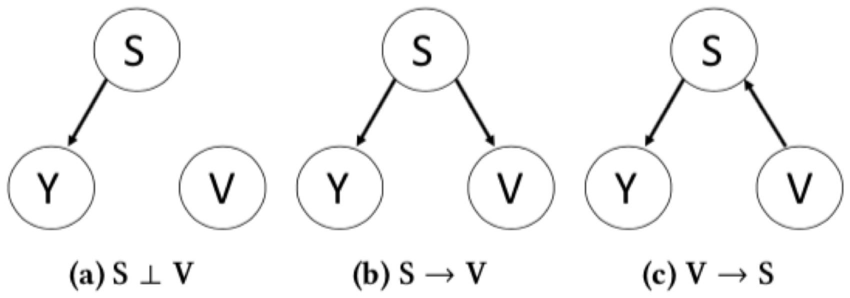


Figure 1: Three diagrams for stable features S, noisy features V, and response variable Y.

Related Work – address env. shift problem

- Covariate shift
  - Kernel mean matching [1], maximum entropy [2], robust bias-aware [3]
  - Importance weights: mimic the distribution of testing data to training data

$$\lim_{n\to\infty} \min_h \mathbb{E}_{f^{(n)}_{\mathrm{training}}(x)\tilde{f}(y|x)} \left[\boxed{\frac{f_{\mathrm{testing}}(\mathbf{X})}{f_{\mathrm{training}}(\mathbf{X})}} (Y - h(\mathbf{X}))^2 \right]$$

$$= \min_h \mathbb{E}_{f_{\mathrm{testing}}(x)\tilde{f}(y|x)} \left[(Y - h(\mathbf{X})^2) \right]$$

  - These methods require prior knowledge of testing data
  - These methods ignore the dependence issue

Related Work

- Invariant Component Learning
  - Invariant prediction [4], domain generalization [5]
  - Assume $P(Y|S)$ is stable across environments
  - Finding a subset/representation of features S', such that $P(Y|S')$ is invariant across all observed **multiple** environments

  - Their performance depends on the diversity of their training data
  - They could still have dependence issue on V', if $P(Y|V')$ is also invariant across observed environments

Challenges

- **Dependence challenge**
  - Y is not independent with V
  - **Some v $\subseteq V$ would be learned as important predictors**
- **Environment shift challenge**
  - The joint distribution $P(XY)$ is different across environments.
  - $\boldsymbol{Corr(V_{training}, Y_{training}) \neq Corr(V_{testing}, Y_{testing})}$
  - Can be addressed if $V \perp Y$ on training environment
- **Unknown testing environments challenge**
  - No prior knowledge on future testing data.
  - Can be addressed if $V \perp Y$ on training environment

Challenges

- **Dependence challenge**
  - Y is not independent with V
  - **Some v $\subseteq$ V would be learned as important predictors**
- **Environment shift challenge**
  - The joint distribution $P(XY)$ is different across environments.
  - $\boldsymbol{Corr(V_{training}, Y_{training}) \neq Corr(V_{testing}, Y_{testing})}$
  - Can be addressed if $V \perp Y$ on training environment
- **Unknown testing environments challenge**

Key Challenge: How to make $V \perp Y$

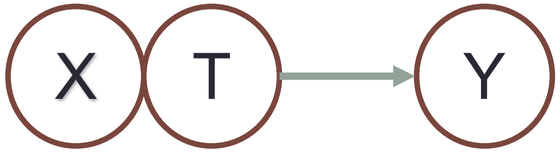
Linking to Causality

- Outcome generating mechanism
  - $Y = f(S) + \varepsilon,\ X = \{S, V\}$
- Difference between S and V
  - S has causal effect on Y,
  - but V has no causal effect on Y.

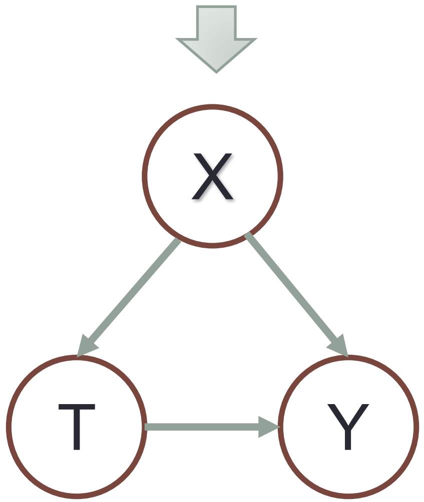
- **Our idea**: Recover causation between X and Y, such that $V \perp Y$, and only S is correlated with Y

Towards stable prediction

- **Discard spurious correlation and embrace causality.**



Typical Correlation Framework

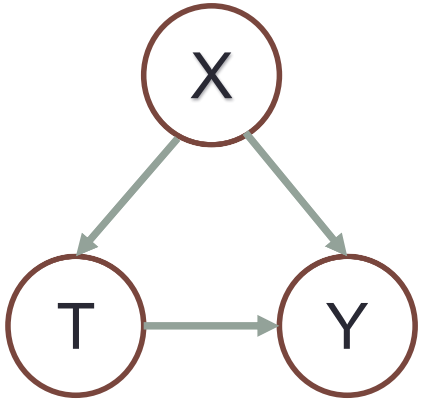


Typical Causal Framework

Estimate the correlation effect of variable *T* and output *Y* without evaluating the relationships between *X* and *T*

Estimate the causal effect of variable *T* on output *Y* With balanced confounder *X* (A/B Testing)

Causal Inference by Exactly Matching



Typical Causal Framework

Analogy of A/B Testing

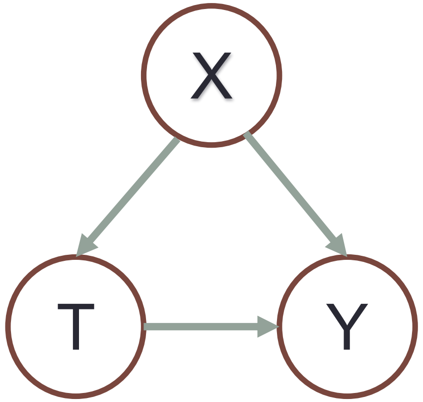
Given a feature T

Find out the sample pairs that one contains T while the other don't, but they are similar in all other features.

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

The requirement is too strong and we can hardly find satisfied groups of samples.

Causal Inference by Confounder Balancing



Typical Causal Framework

Analogy of A/B Testing

Given a feature T

↓

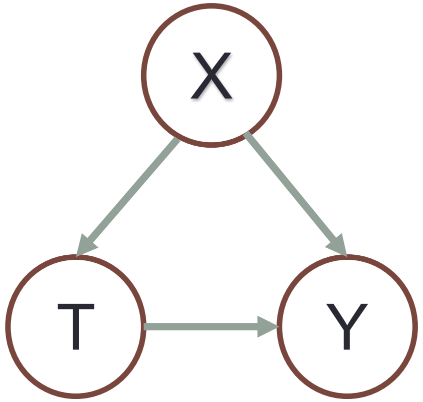
Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Too many parameters. For N samples and K features, we need to learn K*N parameters. Not learning-friendly.

Global Balancing: bridging causality and prediction



Typical Causal Framework

Analogy of A/B Testing

Given **ANY** feature T

Assign global sample weights to samples so that the samples with T and the samples without T have similar distributions in X

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Reduce the parameter number from K*N to N.

Causal Regularizer and Theoretical Guarantee

- **Causal Regularizer** (Approximate global balancing)
  - Making any two variables in X become independent by learning a global sample weights W:

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^{T} \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^{T} \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_{2}^{2}, \quad (4)$$

$$\downarrow$$

$$0$$

Proposition 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

Causally Regularized Logistic Regression

- Global Balancing Regression (GBR) Algorithm

$$\min \quad \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (\mathbf{X}_i \boxed{\beta}))), \qquad (5)$$

$$s.t. \quad \sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^T \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_2^2 \le \lambda_1, \quad W \ge 0,$$

$$\|W\|_2^2 \le \lambda_2, \quad \|\beta\|_2^2 \le \lambda_3, \quad \|\beta\|_1 \le \lambda_4, \quad (\sum_{k=1}^{n} W_k - 1)^2 \le \lambda_5$$

Sample re-weighted logistic loss

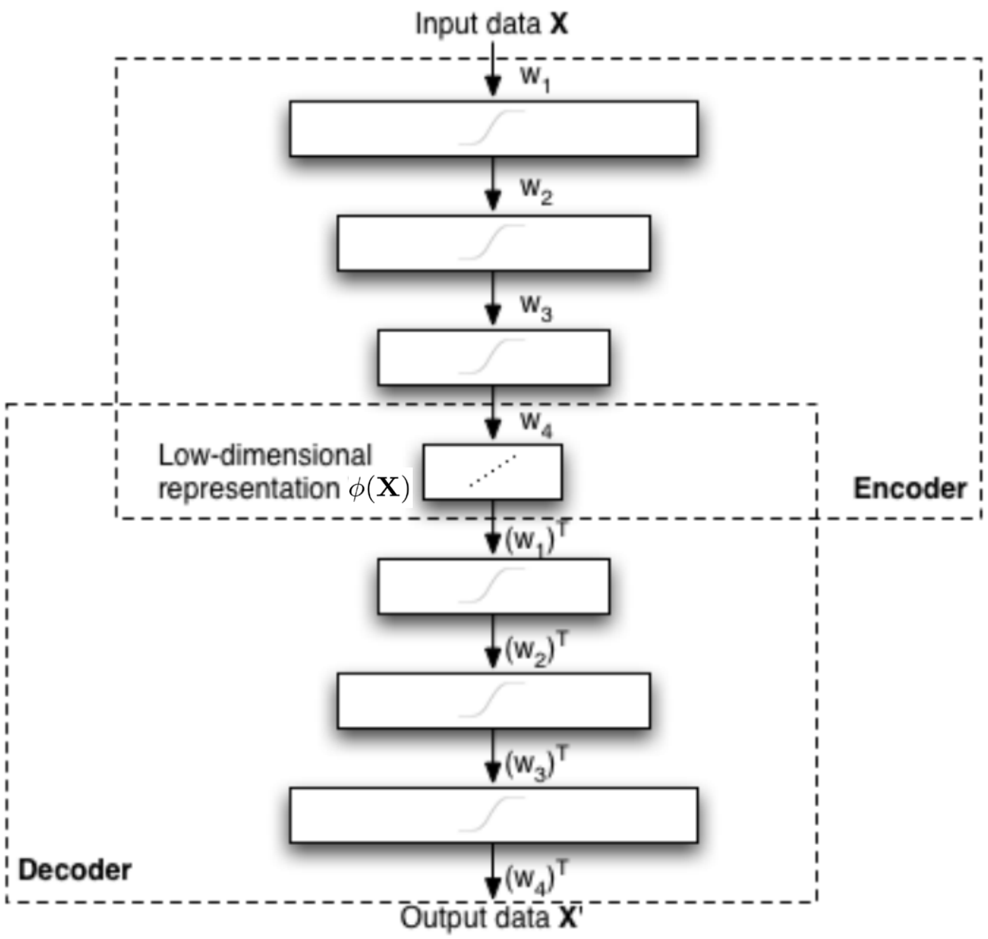
Causal Regularizer

Causality Coefficients

- Causality Coefficients: explainable and stable
- Linear model

Challenges from the Wild Big Data Era

- High dimensional predictors
  - Hundred and thousand variables
  - Dimension reduction
- Non-linear predictions
  - Non-linear relationship between predictors and outcome variable
  - Non-linear function
- Deep Auto-Encoder



From Shallow to Deep - DGBR

- Deep Global Balancing Regression Algorithm

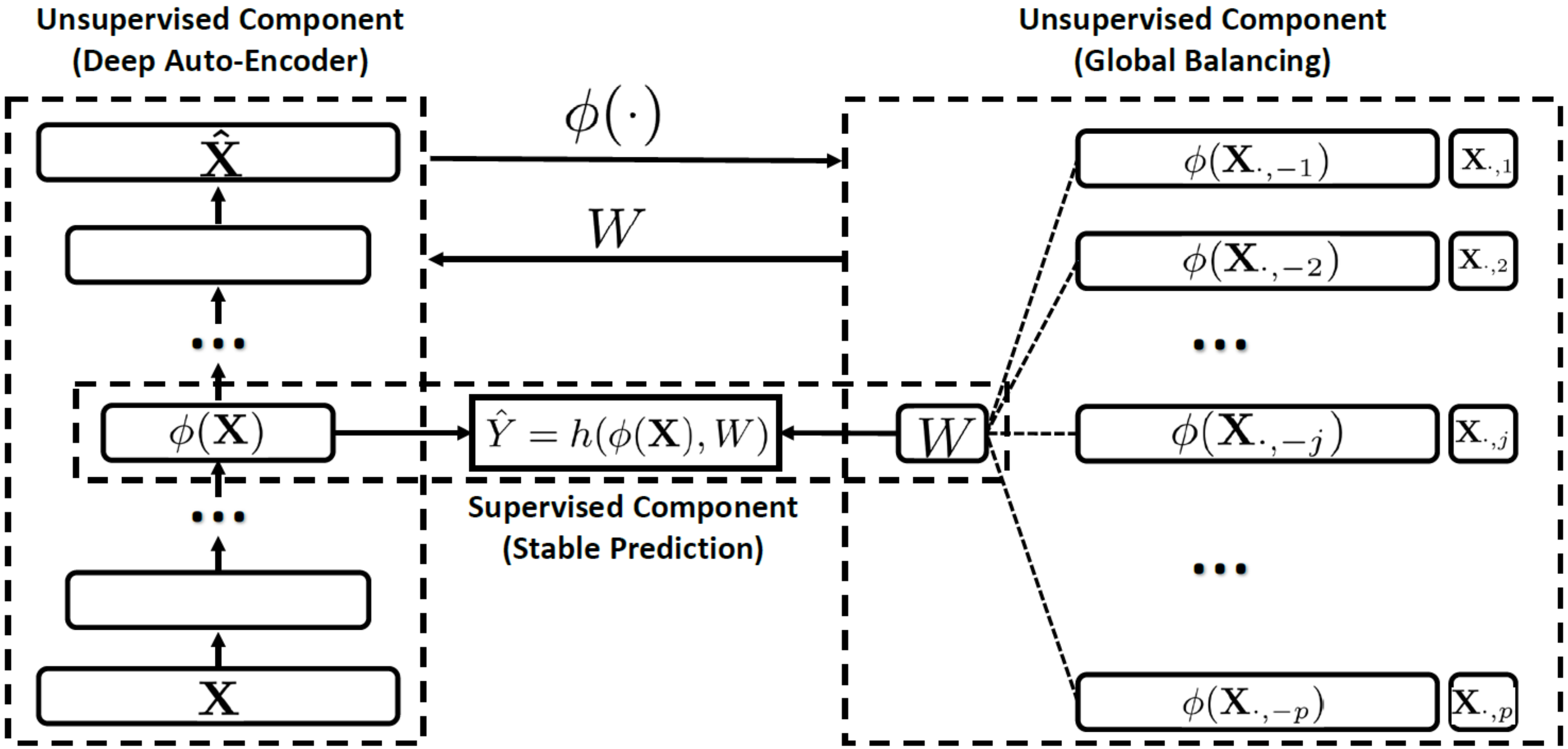


Figure 2: The framework of our proposed DGBR model.

Theoretical Analysis

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^{T} \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^{T} \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2, \quad (4)$$

- The components of X could be mutually independent in the reweighted data.

PROPOSITION 1. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

- **Our GBR algorithm can make $V \perp Y$**

PROPOSITION 2. *If* $0 < \hat{P}(\mathbf{X}_i^e = x) < 1$ *for all* x *in environment* e, $Y^{e'}$ *and* $\mathbf{V}^{e'}$ *are independent when the joint probability mass function of* $(\mathbf{X}^{e'}, Y^{e'})$ *is given by reweighting the distribution from environment* e *using weights* W^*, *so that* $p^{e'}(x, y) = p^e(y|x) \cdot (1/|\mathcal{X}|)$.

Theoretical Analysis

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^{T} \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^{T} \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_{2}^{2}, \quad (4)$$

- The components of X could be mutually independent in the reweighted data.

PROPOSITION 1. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n} \sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

- **Our GBR algorithm can make $V \perp Y$**

PROPOSITION 2. *If* $0 < \hat{P}(\mathbf{X}_i^e = x) < 1$ *for all* x *in environ-*

Propositions 1&2 suggest that our GBR algorithm can make a stable prediction across unknown environments

Theoretical Analysis

- Our DGBR algorithm can preserve all properties of the GBR algorithm while making the overlap property easier to satisfy and reducing the variance of balancing weights.

- Our DGBR algorithm can enable more accurate estimation of $P(Y|S)$.

- More details could be found in our paper.

Experiments

- Baselines:
  - Logistic Regression (LR)
  - Deep Logistic Regression (DLR): LR + Deep Auto Encoder
- Evaluation Metric:
  - RMSE, Average_Error, Stability_Error

$$Average\_Error = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} Error(D^e), \quad (1)$$

$$Stability\_Error = \sqrt{\frac{1}{|\mathcal{E}|-1} \sum_{e \in \mathcal{E}} (Error(D^e) - Average\_Error)^2}, \quad (2)$$

Experiments on Synthetic Data

- Data generating
  - $X = \{S, V\}$ is binary.
  - $Y = h(f(S) + \epsilon)$ is also binary.

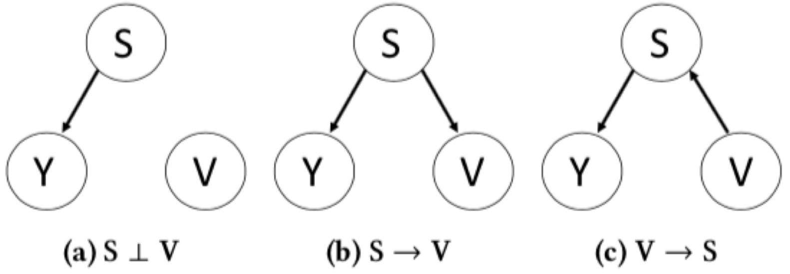


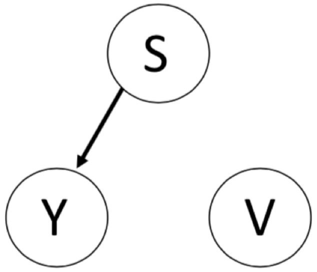
Figure 1: Three diagrams for stable features S, noisy features V, and response variable Y.

- Environments generating
  - Changing P_{XY} by sample selection with the **bias rate: $\boldsymbol{r}$**
  - **Varying $\boldsymbol{P(Y|V)}$:**
    - if $V = Y$, then $p(selected) = r$, otherwise $p(selected) = 1 - \mathrm{r}$.
  - **Different $\boldsymbol{r}$ means different environments**
  - Note that: $\boldsymbol{r > 0.5}$ **implies $\boldsymbol{Corr(V, Y)}$ is positive**

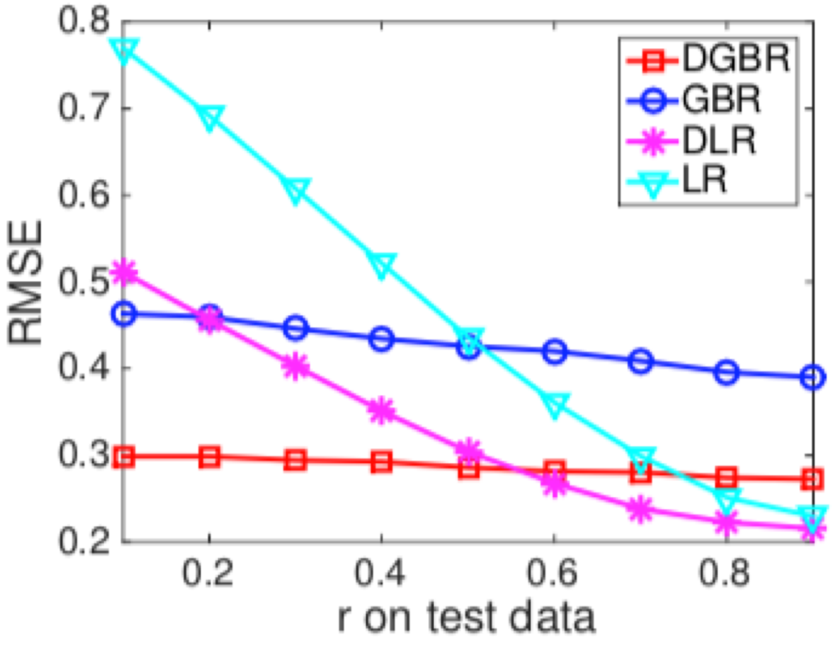
Experiments on Synthetic Data

- Setting $S \perp V$
  - **Trained** on one environment $\boldsymbol{r = 0.85}$, and **tested** on all environments $\boldsymbol{r = \{0.1, \dots, 0.9\}}$
  - **Different r means different environment**
  - $\boldsymbol{r > 0.5}$ **implies** $\boldsymbol{Corr(V, Y)}$ **is positive**

- Traditional LR and DLR failed
- GBR (dark blue) is more stable than LR
- DGBR (Red) is more stable than DLR
- DGBR is more stable and precise than GBR



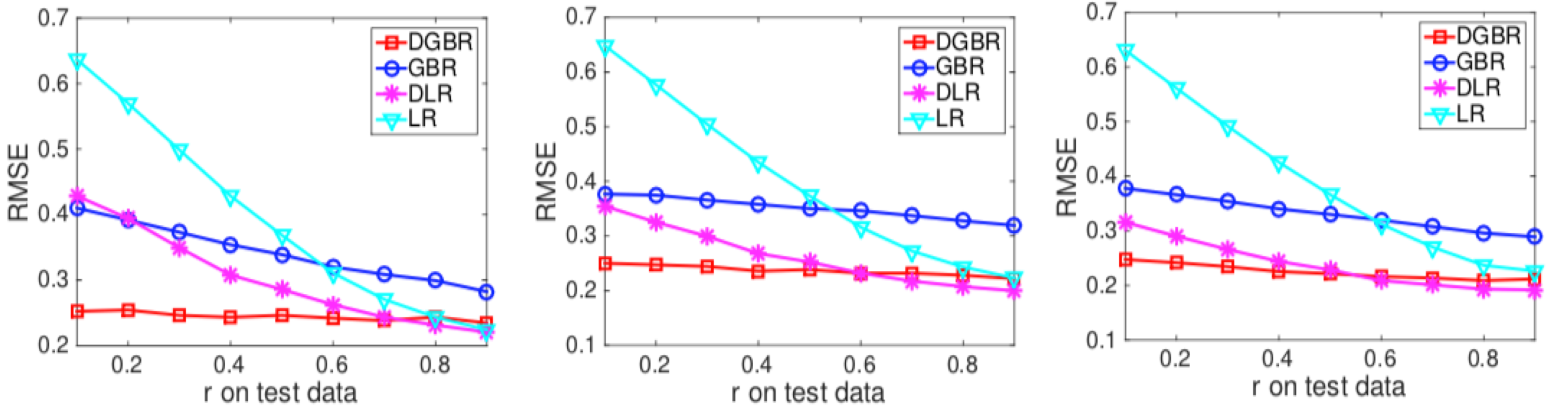
(a) S ⊥ V



(f) Trained on $n = 2000$, $p = 20$, $r = 0.85$

Experiments on Synthetic Data

- More settings: varying n, p, and r

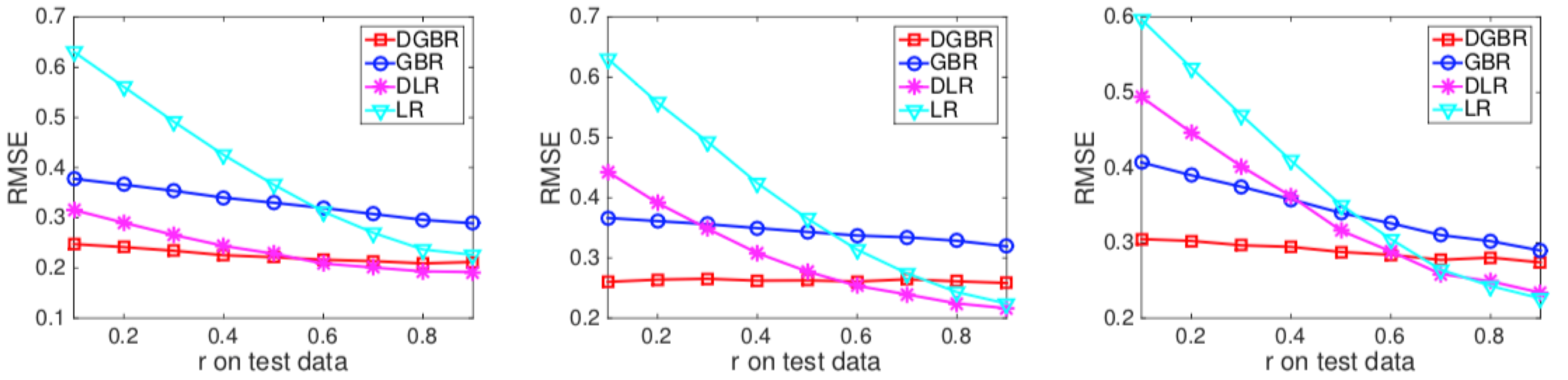


(b) Trained on $n = 1000$, $p = 20$, $r = 0.75$ **(e)** Trained on $n = 2000$, $p = 20$, $r = 0.75$ **(h)** Trained on $n = 4000$, $p = 20$, $r = 0.75$

Vary sample size n

Experiments on Synthetic Data

- More settings: varying n, p, and r

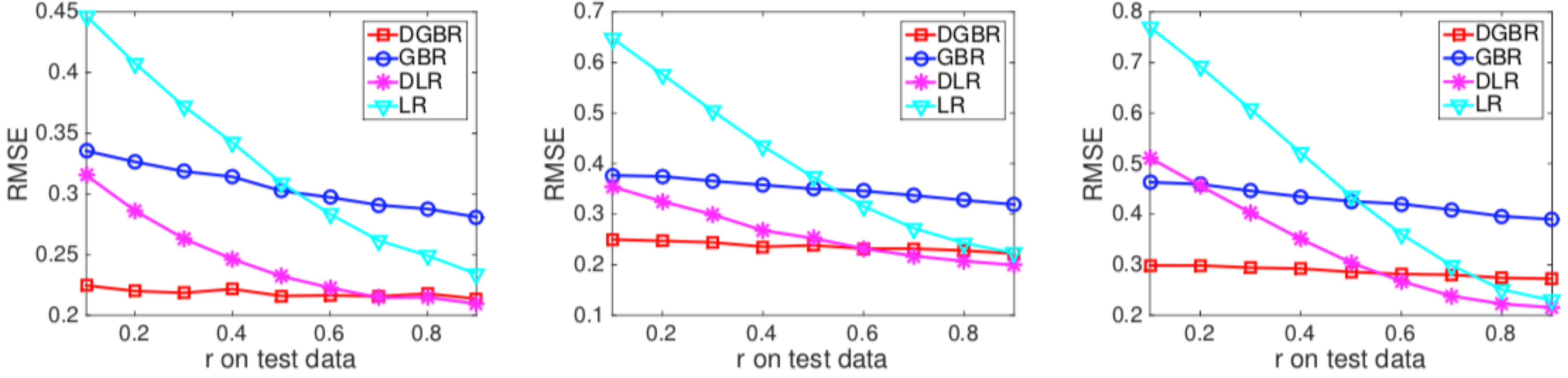


(a) Trained on $n = 4000$, $p = 20$, $r = 0.75$ (b) Trained on $n = 4000$, $p = 40$, $r = 0.75$ (c) Trained on $n = 4000$, $p = 80$, $r = 0.75$

Vary variables' dimension p

Experiments on Synthetic Data

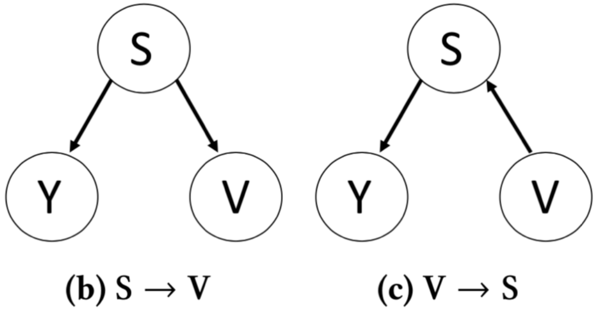
- More settings: varying n, p, and r



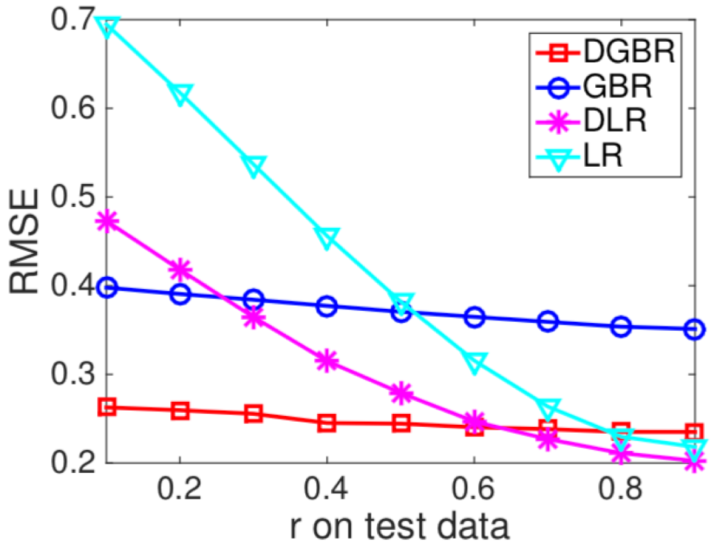
(d) Trained on $n = 2000$, $p = 20$, $r = 0.65$ **(e) Trained on** $n = 2000$, $p = 20$, $r = 0.75$ **(f) Trained on** $n = 2000$, $p = 20$, $r = 0.85$

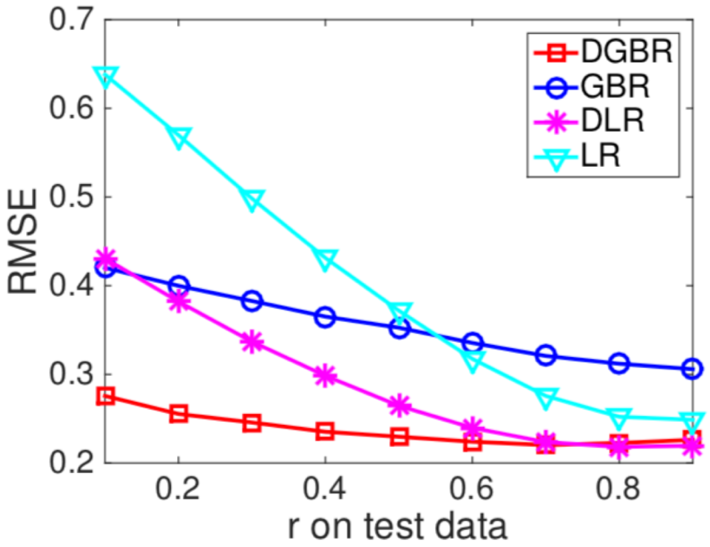
Vary bias rate r on training environment

Experiments on Synthetic Data



- More settings: setting S→V (S is the cause of V)
  setting V→S (V is the cause of S)





The RMSE of DGBR is consistently stable and small across environments under all settings.

Experiments on online advertising

2015 WeChat

- Dataset Description:
  - Online advertising campaign (LONGCHAMP)
  - Users Feedback: 14,891 LIKE; 93,108 DISLIKE
  - 56 Features for each user
    - Age, gender, #friends, device, user setting on WeChat
- Experimental Setting:
  - Outcome Y: users feedback ⟵ Y = 1, if LIKE; Y = 0, if DISLIKE
  - Setting: generating environment with users' age.

Experiments on online advertising

- Environments generating:
  - Separate the whole dataset into 4 environments by users' age, including $Age \in [20,30)$, $Age \in [30,40)$, $Age \in [40,50)$, and $Age \in [50,100)$.

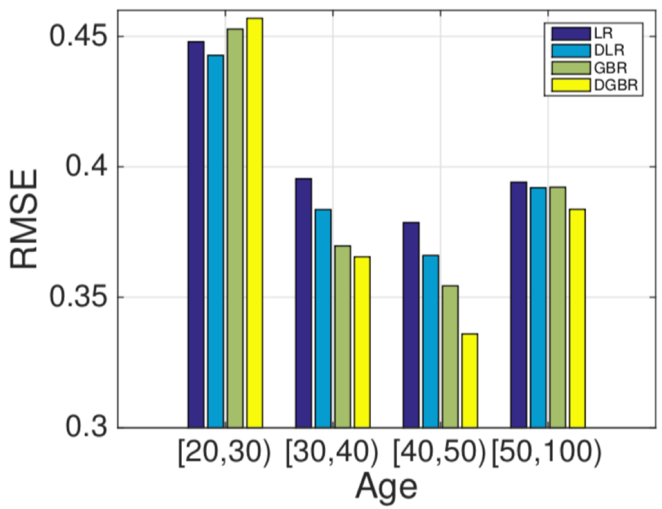


Fig. 15: Prediction across environments separated by age. The models are trained on dataset where uses' $Age \in [20, 30)$, but tested on various datasets with different users' age range.

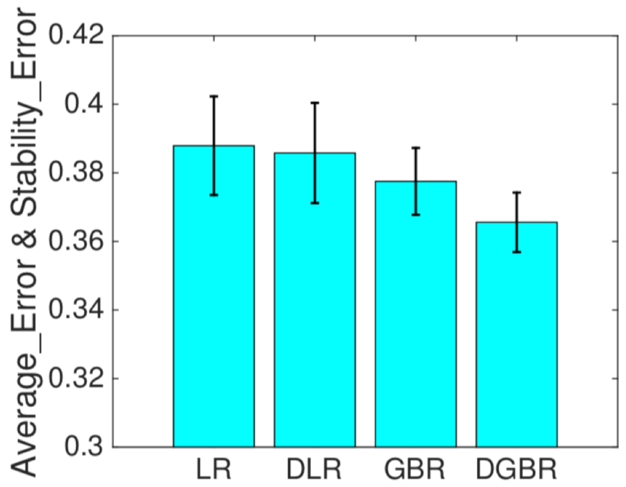
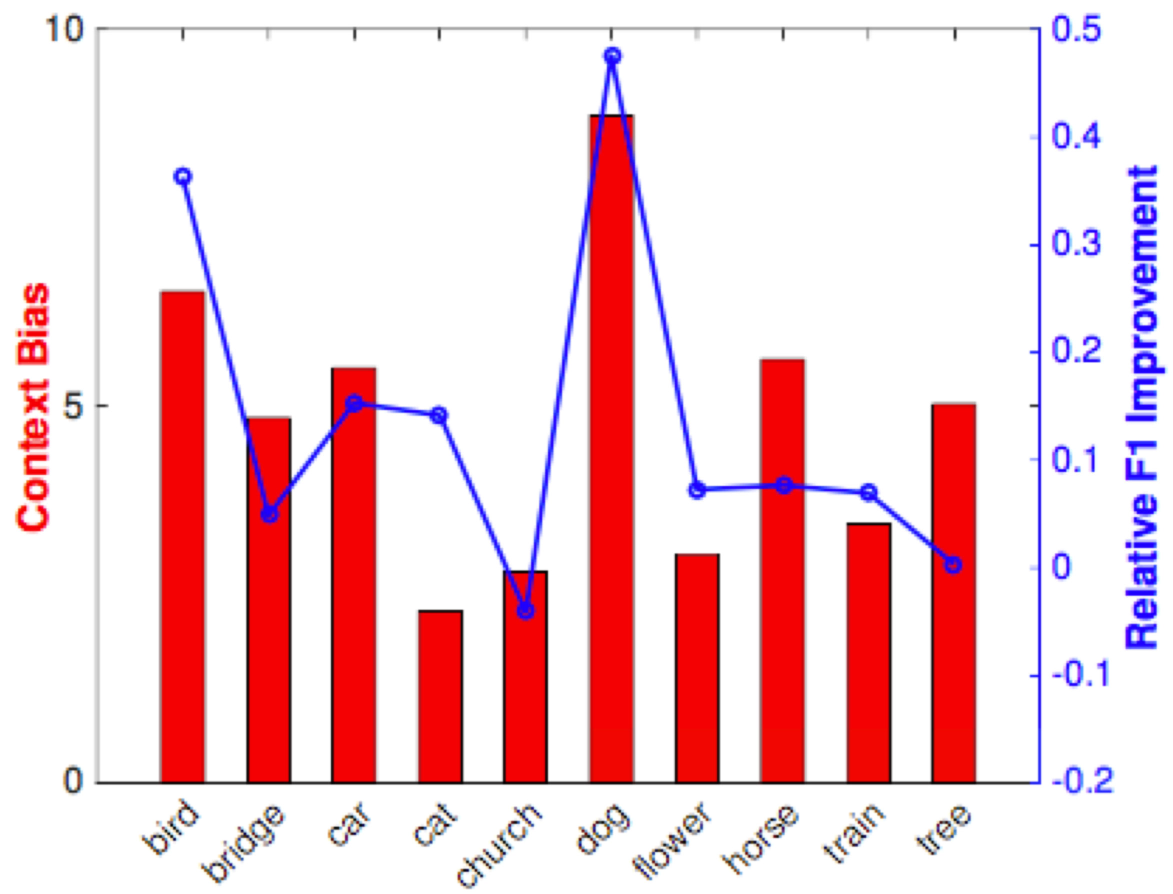


Fig. 16: $Average\_Error$ and $Stability\_Error$ of all algorithms across environments after fixing $P(Y)$ as the same with its value on global dataset.

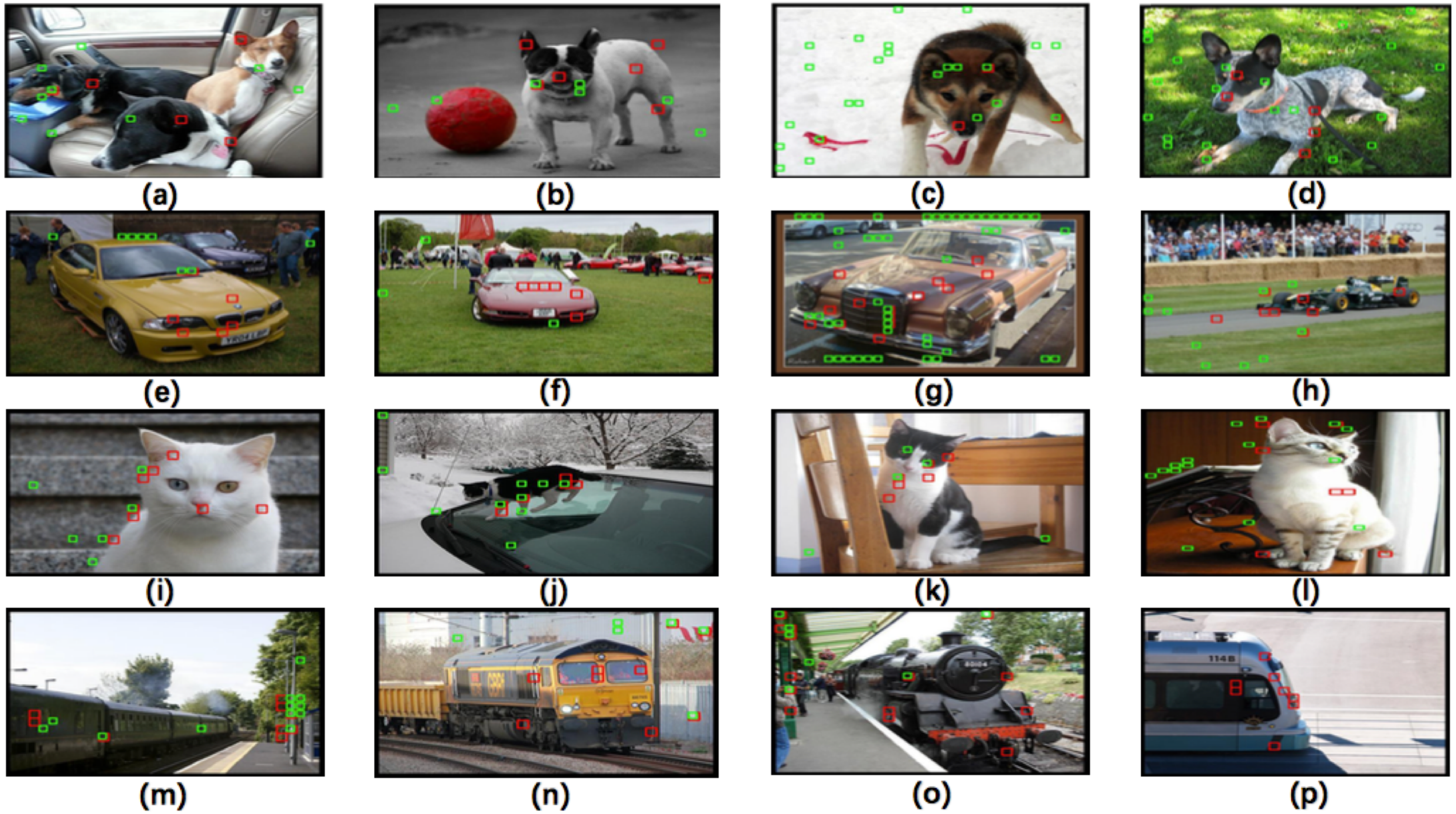
Experiments on image classification

- Source: ***YFCC100M***
- Type: multi-tags and high-resolution
- Scale: 10-category, each with nearly 1000 images
- Method: one ***major object tag*** (as category label) and 5 ***context tags*** which are frequently co-occurred with the major tag

Experiments on image classification



Experiments on image classification



Summary: Causally Regularized Stable Learning

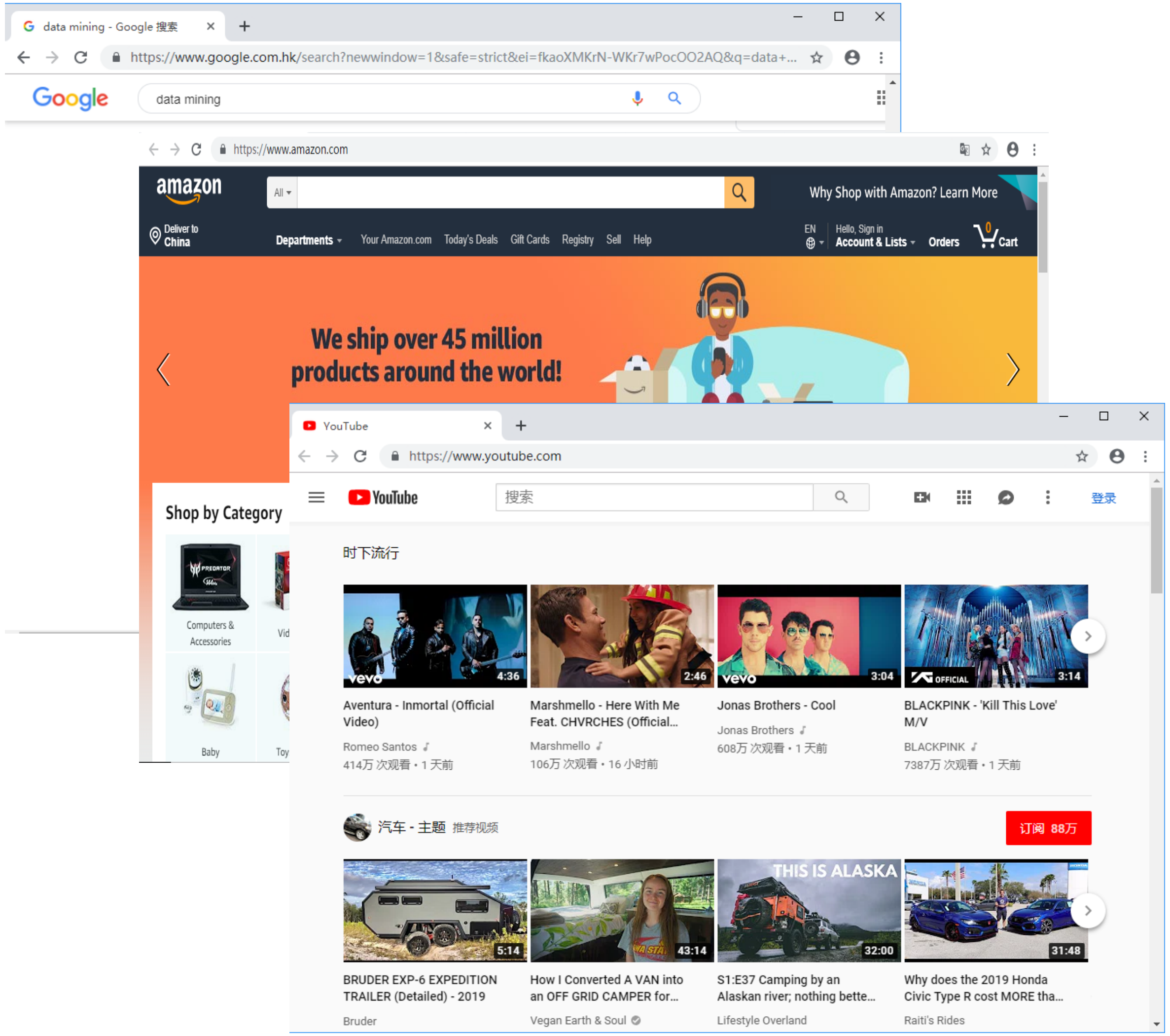
- Today's Machine Learning:
  - Correlation Based
  - Correlation: causation, confounding, selection bias (Spurious Correlation)
  - To know the hows but not the whys
  - 知其然，但不知其所以然

- Causally Regularized Stable Learning
  - Causal regularizer
  - Recover causation from correlation
  - Causation based stable learning
  - Improving interpretability and stability on prediction

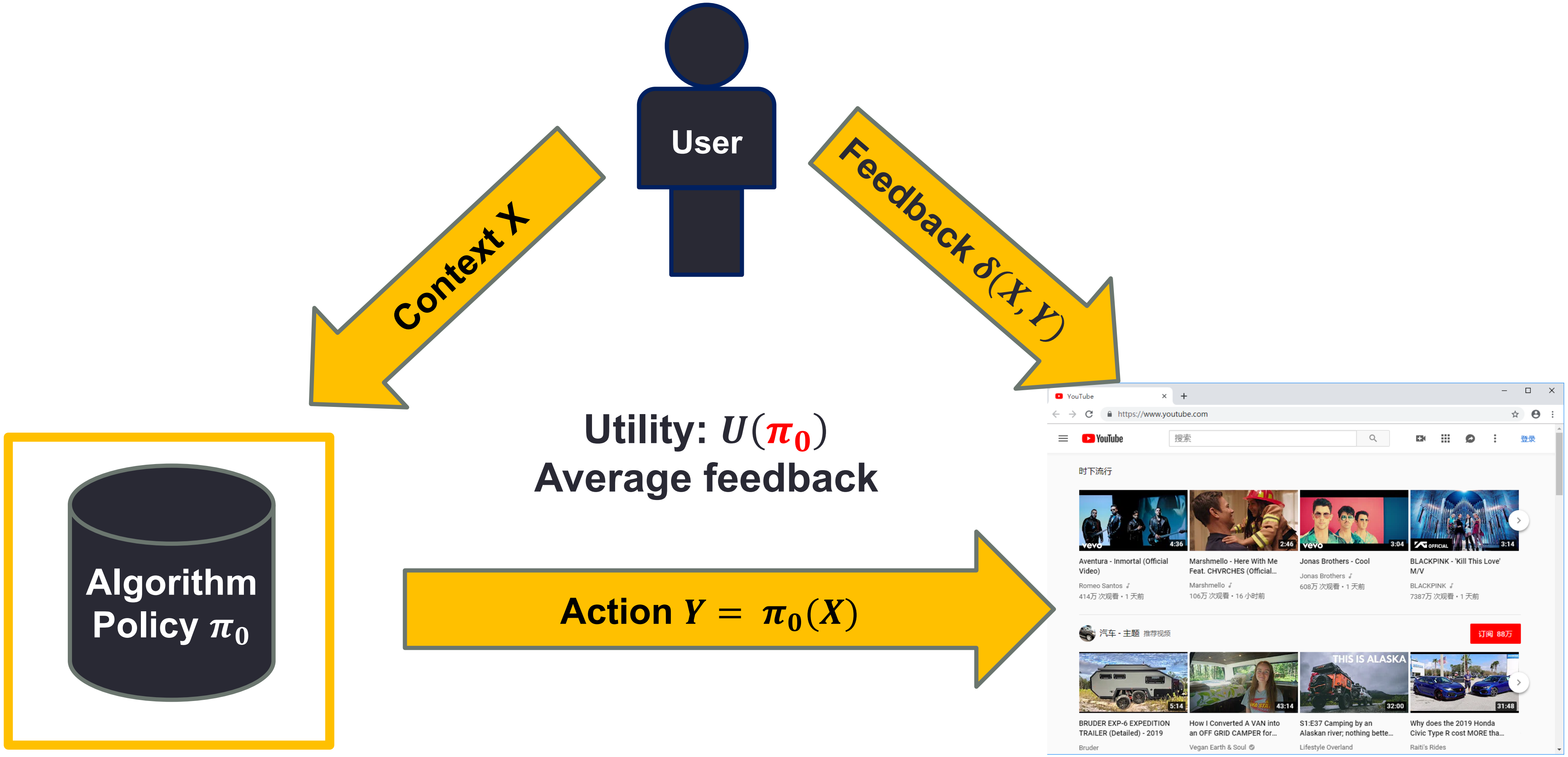
OUTLINE

User Interactive systems

- Examples:
  - Search engines
  - Ads-placement systems
  - Videos recommender systems
- Policy: recommended algorithm
- Logs of user behavior for policy evaluation
  - Evaluate the system performance
  - Improve the policy in the system

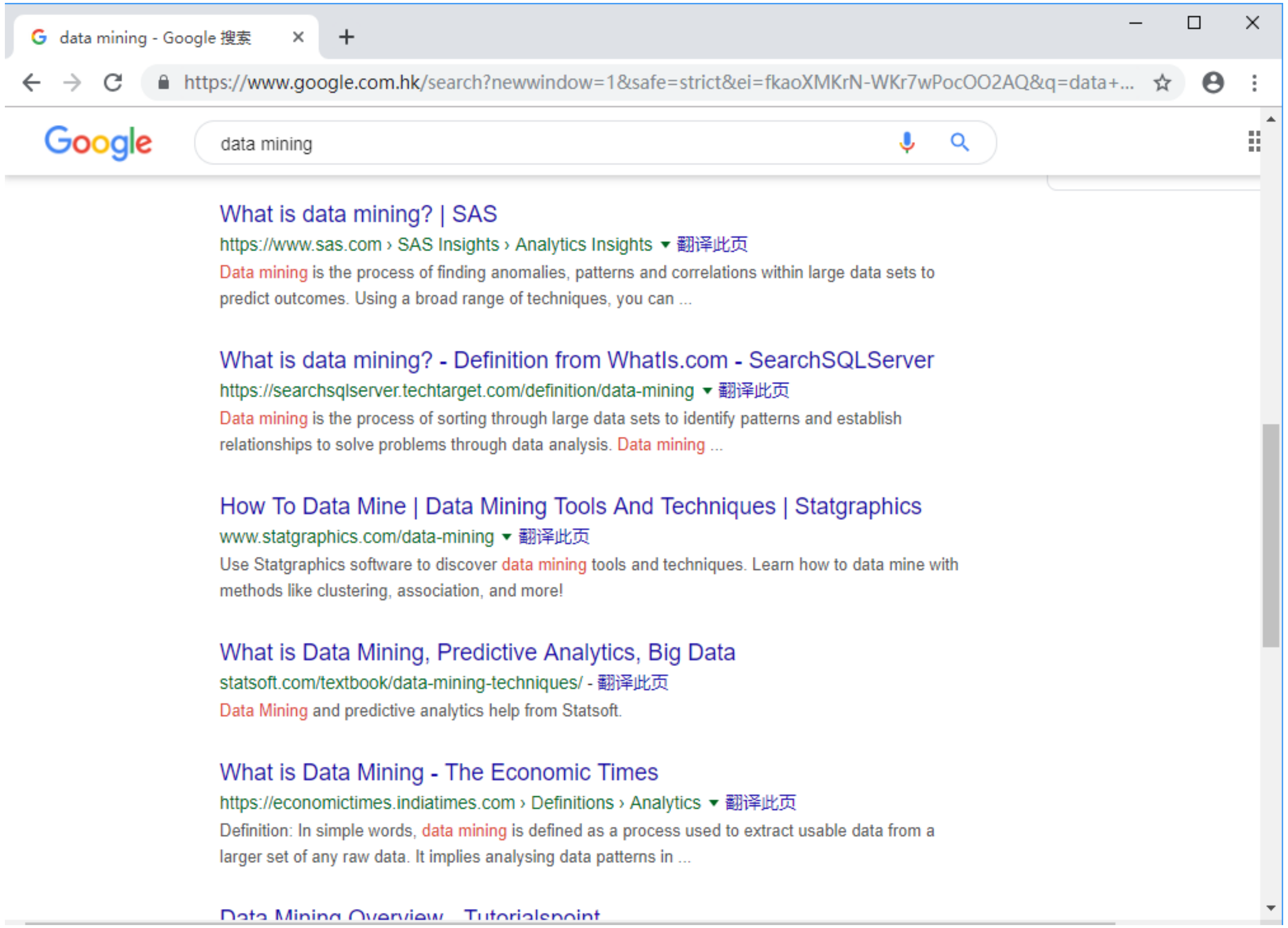


Interactive System Schema



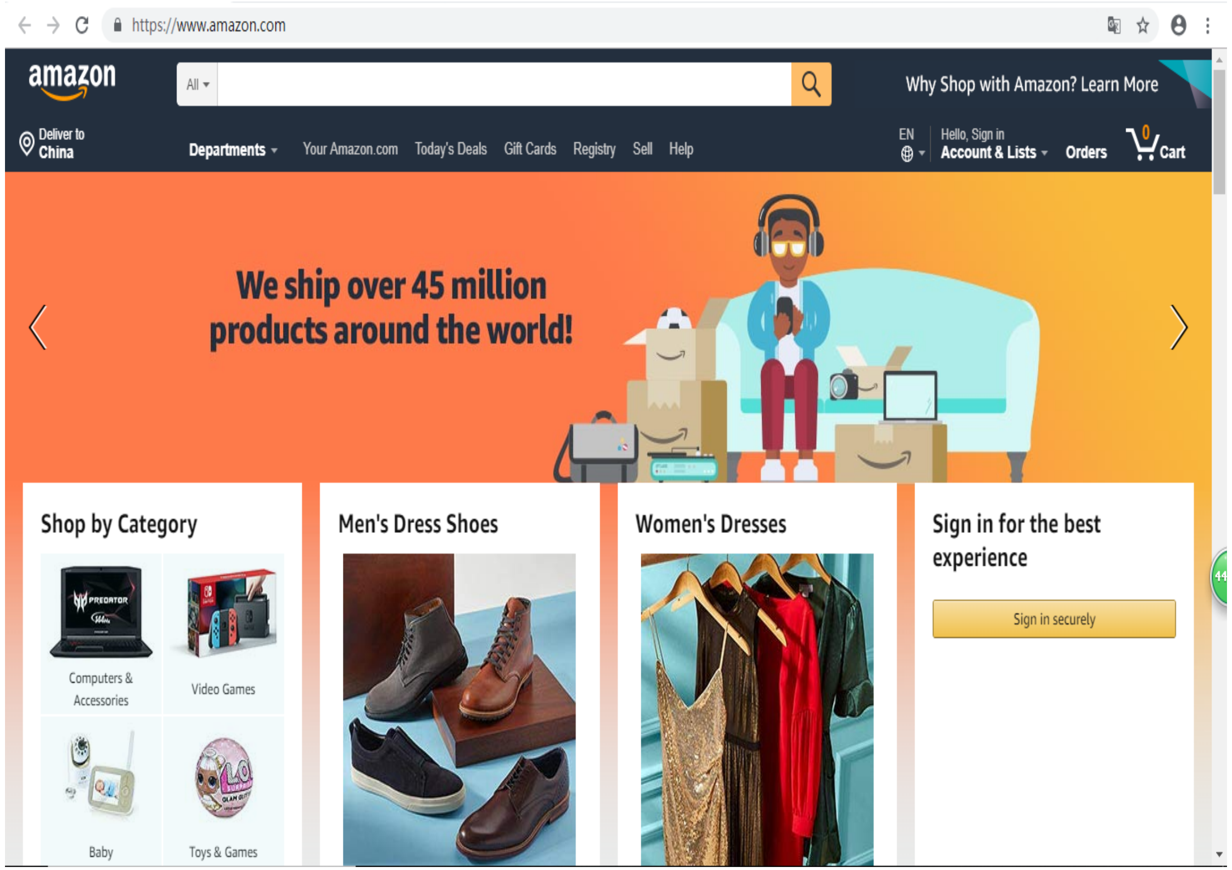
Search engine

- Context X:
  - Query
- Action $Y = \pi_0(X)$:
  - Top-k ranking results
- Feedback $\delta(X, Y)$:
  - Click or not



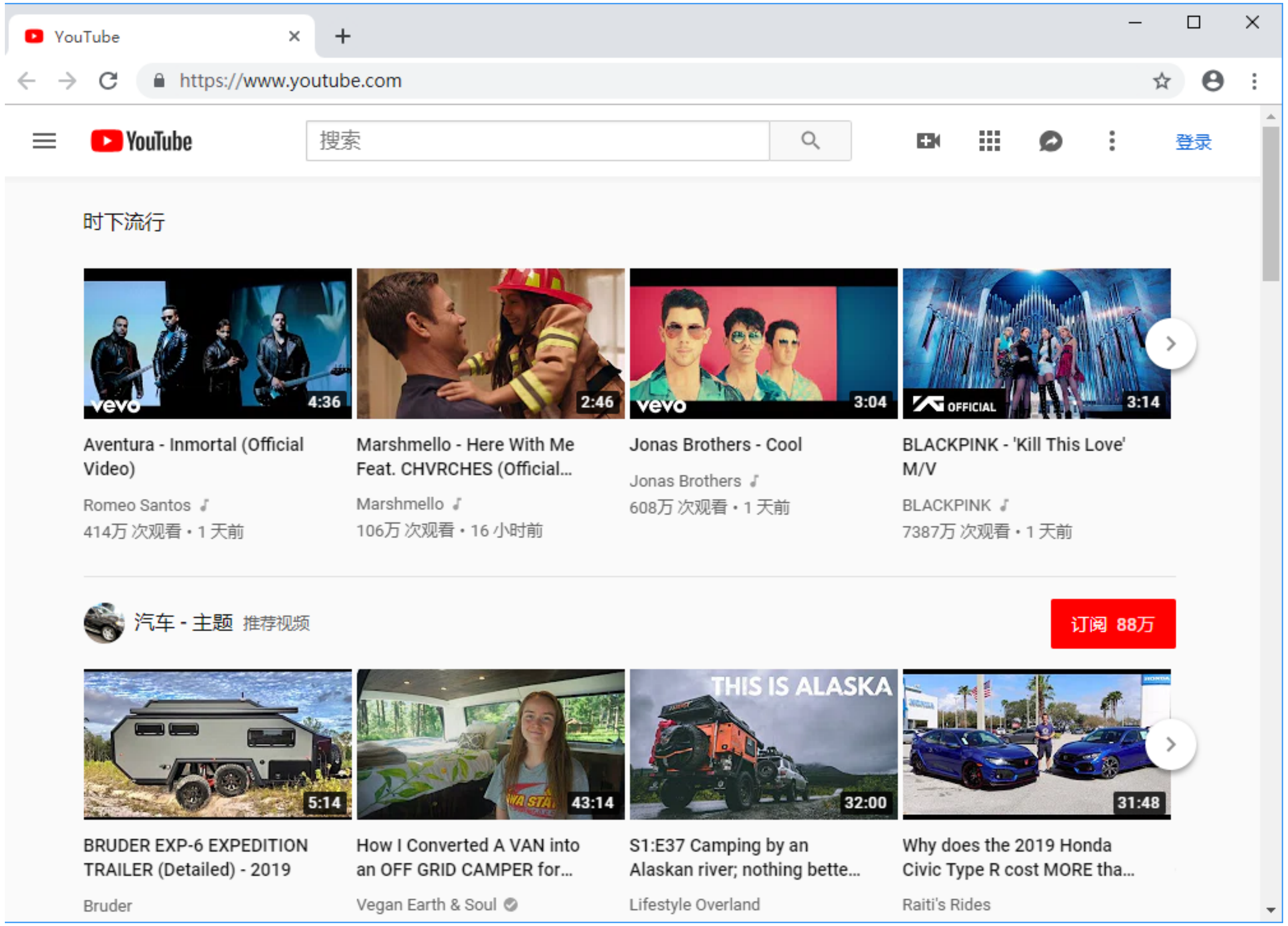
Ads-placement systems

- Context X:
  - Users' features
- Action $Y = \pi_0(X)$:
  - Ads placed
- Feedback $\delta(X, Y)$:
  - Click or not
  - Buy or not



Video Recommender System

- Context X:
  - User features
- Action $Y = \pi_0(X)$:
  - Videos recommend
- Feedback $\delta(X, Y)$:
  - Click or not
  - Watching time



Offline Policy Evaluation

- Log Data from π_0: samples indexed by $1,2,\cdots,n$

$$S = ((X_1, Y_1, \delta_1), (X_2, Y_2, \delta_2), \cdots, (X_n, Y_n, \delta_n))$$

context action feedback

- Properties
  - Contexts X_i are drawn i.i.d from unknown $\Pr(X)$
  - Actions Y_i are decided by the existing policy $\pi_0: X \to Y$
  - Feedback δ_i are from unknown feedback function $\delta: X \times Y \to R$

How to evaluate a new policy π ?

Policy Evaluation: Online A/B Testing

- A/B Testing:
  - Deploy a new policy π in the interactive systems
  - Draw $\mathbf{X} \sim Pr(\mathcal{X})$, select $Y \sim \pi(\mathcal{Y}|X)$, and get $\delta(\mathbf{X}, Y)$
- Drawbacks:
  - Long turn-around time
  - Costly, number of A/B Testing limited
  - May be detrimental to the user experience
- Big Data Era
  - Lots of logged data

How to evaluate a new policy π offline with logged data ?

Offline Policy Evaluation

- **Given** the logged data from a past (existing) policy π_0:

$$S = ((X_1, Y_1, \delta_1), (X_2, Y_2, \delta_2), \cdots, (X_n, Y_n, \delta_n))$$

- **Goal**: to estimate the utility of a new policy π:

$$U(\pi) = \mathbb{E}_{\mathbf{X} \sim Pr(\mathbf{X}), Y \sim \pi(\mathcal{Y} | \mathbf{X})} [\delta(\mathbf{X}, Y)]$$

- Utility: the average feedback of policy over the population

$$U(\pi) = \mathbb{E}_{\mathbf{X}\sim Pr(\mathbf{X}),\, Y\sim\pi(\mathcal{Y}|\mathbf{X})}[\delta(\mathbf{X}, Y)]$$

Challenges of Offline Policy Evaluation

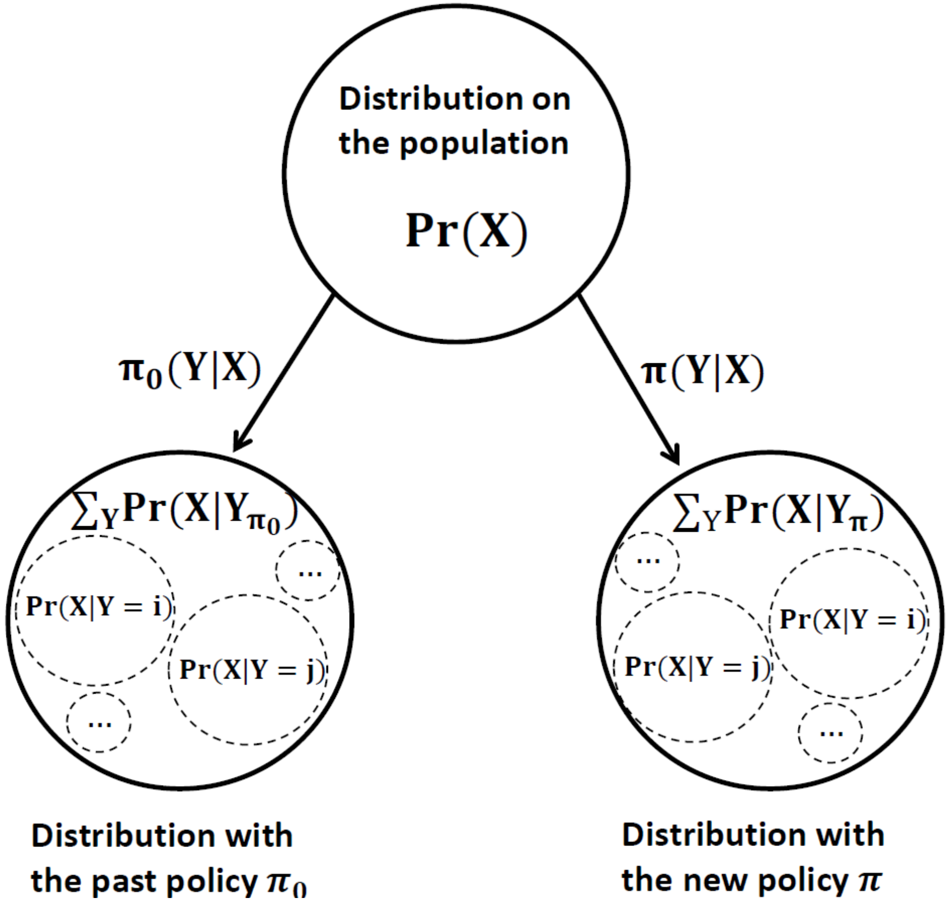
- Distribution shift induced by the past policy π_0
  - Y is assigned based on X through $\pi_0(\mathrm{Y}|\mathrm{X})$

$$Pr(\mathbf{X}|Y_{\pi_0} = i) \neq Pr(\mathbf{X}|Y_{\pi_0} = j) \neq Pr(\mathbf{X})$$

- Action discrepancy induced by the new policy $\pi(Y|X)$: Y is assigned through $\pi(Y|X)$

$$\pi(Y = i|X) \neq \pi(Y = j|X)$$

$\pi(Y = k|X) \approx 0$: No context X will be assigned to Y=k under π, hence distribution shift from action Y=k does not affect results



Focus on the action group with high value of $\pi(Y = i|X)$

Related Work

- Direct method (DM) directly estimate the feedback function $\widehat{\delta}(\mathbf{X}, Y)$ by utilizing the logged data to predict the feedbacks of actions chosen by the new policy π .

$$\widehat{U}_{DM}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{Y_j \in \mathcal{Y}} \widehat{\delta}(X_i, Y_j)\pi(Y_j|X_i)$$

- Direct method is unbiased if the feedback model is correct.

- But we hardly know the real underlying feedback function, and it ignores the distribution shift induced by the past policy.

Related Work

- Inverse propensity score (IPS) estimator use the propensity score (the probability of the chosen action $\widehat{\pi}_0(Y|\mathbf{X})$) to reweight sample:

$$\widehat{U}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\delta_i \frac{\pi(Y_i|X_i)}{\widehat{\pi}_0(Y_i|X_i)}$$

- IPS is unbiased if propensity score model is correct.

- But we have no prior knowledge on propensity score model
- High variance if propensity score is close to 0 and 1
- Ignoring the action discrepancy induced by new policy π

Related Work

- Doubly Robust (DR) estimator combined IPS estimator and direct method:

$$\widehat{U}_{DR}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{Y\in\mathcal{Y}} \pi(Y|X_i)\left[\widehat{\delta}(X_i, Y) + \frac{I(Y = Y_i)}{\widehat{\pi}_0(Y_i|X_i)}(\delta_i - \widehat{\delta}(X_i, Y))\right]$$

- DR estimator is unbiased if either propensity score model or feedback model is correct

- But one cannot guarantee the specified model is correct
- Moreover, it still ignores the action discrepancy induced by new policy π

Summary on Related Work

- Distribution shift induced by the past policy π_0
  - Y is assigned based on $\pi_0(Y|X)$

$$Pr(\mathbf{X}|Y_{\pi_0} = i) \neq Pr(\mathbf{X})$$

Related Work

- Action discrepancy induced by the new policy $\pi(Y|X)$

$$\pi(Y = i|X) \neq \pi(Y = j|X)$$

$\pi(Y = k|X) \approx 0$: No X will be assigned with Y=k under π, hence distribution shift from action Y=k does not affect results

Focus on the action group with high $\pi(Y = i|X)$

Remain Challenges

Models dependency

Context Balancing

- Context Balancing: a non-parametric method based on directly covariate balancing to correct the distribution shift induced by the past policy
- Learning sample weights W in each action group k as follows:

$$W_{Y=k} = \arg\min_{W_{Y=k}} \left\| \boxed{\frac{1}{n}\sum_{i=1}^{n} \mathbf{M}_i} - \boxed{\sum_{j:Y_j=k} W_j \cdot \mathbf{M}_j} \right\|_2^2,$$

The distribution on the population

$$\mathbf{M} = \{\mathbf{X}, \mathbf{X}^2, \mathbf{X}_{,i}\mathbf{X}_{,j}, \mathbf{X}^3, \mathbf{X}_{,i}\mathbf{X}_{,j}\mathbf{X}_{,k}, \cdots\}$$

The corrected distribution

- With sample weights $W = \{W_{Y=1}, W_{Y=2}, \cdots, W_{Y=K}\}$, CB estimator is

$$\widehat{U}_{CB}(\pi) = \sum_{i=1}^{n} \pi(Y_i|X_i) W_i \delta_i$$

Remove the model dependency
But ignore action discrepancy

Focused Context Balancing (FCB) estimator

- Context Balancing: learning sample weights by directly variables balancing

$$W_{Y=k} = \arg\min_{W_{Y=k}} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{M}_i - \sum_{j:Y_j=k} W_j \cdot \mathbf{M}_{j,} \right\|_2^2$$

- Focused Context Balancing: focusing on the action group with high probability when learning sample weights:

$$W_{Y=k} = \arg\min_{W_{Y=k}} \left\| \sum_{i=1}^{n} \frac{1}{n} \boxed{\pi(Y=k|\mathbf{X}_i)} \mathbf{M}_{i,} - \sum_{i:Y_i=k} W_i \boxed{\pi(Y=k|\mathbf{X}_i)} \mathbf{M}_{i,} \right\|_2^2$$

Focused Component

Theoretical Analysis

- Taylor's expansion of feedback function on the context:

$\delta(\mathbf{X}, Y = k) = \alpha_{Y=k} \cdot \mathbf{M}$ where $\mathbf{M} = \{\mathbf{X}, \mathbf{X}^2, \mathbf{X}_{,i}\mathbf{X}_{,j}, \mathbf{X}^3, \mathbf{X}_{,i}\mathbf{X}_{,j}\mathbf{X}_{,k}, \cdots\}$

$$
\begin{aligned}
\widehat{U}(\pi) &= \textstyle\sum_{i=1}^{n} W_i \pi(Y = Y_i|\mathbf{X}_{i,})\delta(\mathbf{X}_{i,}, Y_i) \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} \sum_{i:Y_i=k} W_i \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} \left[\sum_{i:Y_i=k} W_i \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} - \frac{1}{n}\sum_{i=1}^{n} \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,}\right] \\
&\quad + \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} \frac{1}{n}\sum_{i=1}^{n} \pi(Y = k|\mathbf{X}_{i,})\mathbf{M}_{i,} \\
&= \textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} B_k + \frac{1}{n}\sum_{i=1}^{n}\sum_{k\in\mathcal{Y}} \delta(\mathbf{X}_{i,}, Y = k)\pi(Y = k|\mathbf{X}_{i,}) \\
&= \boxed{\textstyle\sum_{k\in\mathcal{Y}} \alpha_{Y=k} B_k} + U(\pi)
\end{aligned}
$$

Distribution shift induced by past policy

Action discrepancy from new policy

$$B_k = \underbrace{\textstyle\sum_{i:Y_i=k} W_i \boxed{\pi(Y = k|\mathbf{X}_{i,})}\mathbf{M}_{i,}}_{\text{The adjusted distribution}} - \underbrace{\textstyle\frac{1}{n}\sum_{i=1}^{n} \boxed{\pi(Y = k|\mathbf{X}_{i,})}\mathbf{M}_{i,}}_{\text{The target distribution}}$$

The adjusted distribution

Focused Component

The target distribution

Focused Context Balancing algorithm

- Objective Function:

Focused Component

$$\min_{W_{Y=k}} \Big\| \sum_{i:Y_i=k} W_i \boxed{\pi(Y=k|\mathbf{X}_{i,})} \mathbf{M}_{i,} - \sum_{i=1}^{n} \frac{1}{n} \boxed{\pi(Y=k|\mathbf{X}_{i,})} \mathbf{M}_{i,} \Big\|_2^2$$

$$s.t. \sum_{i:Y_i=k} W_i^2 \le \lambda \quad \sum_{i:Y_i=k} W_i = 1 \;\; and \;\; W \ge 0, \tag{7}$$

- Policy Evaluation:

$$\widehat{U}_{FCB}(\pi) = \sum_{i=1}^{n} \pi(Y_i|X_i) W_i \delta_i.$$

Experiment

- Baselines:
  - Direct Method: regressing on an estimated feedback function to evaluate the effect of new policy.
  - R-IPS: IPS estimator + roughly estimated propensity score not associated with context.
  - E-IPS: IPS estimator with estimated propensity score
  - T-IPS: IPS estimator with the true propensity score
  - SN-IPS: IPS estimator with estimated propensity score + Normalized sample weights
  - Doubly Robust: IPS estimator with estimated propensity score + Direct Method
  - CB: covariate balancing to learn sample weights + ignoring distribution shift induced by new policy.
- Evaluation Metric:

$$
\begin{aligned}
Bias &= \left|\tfrac{1}{T}\textstyle\sum_{i=1}^{T} \widehat{U}(\pi)_i - U(\pi)\right| \\
SD &= \sqrt{\tfrac{1}{T}\textstyle\sum_{i=1}^{T}(\widehat{U}(\pi)_i - \tfrac{1}{T}\textstyle\sum_{i=1}^{T}\widehat{U}(\pi)_i)^2} \\
MAE &= \tfrac{1}{T}\textstyle\sum_{i=1}^{T}|\widehat{U}(\pi)_i - U(\pi)| \\
RMSE &= \sqrt{\tfrac{1}{T}\textstyle\sum_{i=1}^{T}(\widehat{U}(\pi)_i - U(\pi))^2}
\end{aligned}
$$

Experiment - Simulations

- Dataset
  - Sample size: $n = \{5,000, 10,000\}$
  - Context dimension: $p = \{50, 100\}$
  - Observed context: $\mathbf{X} = (x_1, x_2, \ldots., x_p)$ $\quad x_1, x_2, \ldots, x_p \overset{iid}{\sim} Bernoulli(0.5)$
  - Policy to be evaluated: from sigmoid function

    $$\pi_{sig}(Y = 1|\mathbf{X}) = 1/\left(1 + e^{-\sum_{i=1}^{p}(x_i - 0.5)}\right)$$

  - Logged policy: from inverse proportional function, constant function and linear function

    $$\begin{aligned}\pi_{inv}(Y = 1|\mathbf{X}) &= 1/(1 + 3\textstyle\sum_i x_i/p) + \mathcal{N}(0, 0.1)\\ \pi_{uni}(Y = 1|\mathbf{X}) &= 0.5 + \mathcal{N}(0, 0.1)\\ \pi_{lin}(Y = 1|\mathbf{X}) &= \textstyle\sum_i x_i/p + \mathcal{N}(0, 0.1)\end{aligned}$$

  - Feedback function: from linear and non-linear function

    $$\begin{aligned}\delta_{linear} &= Y + \textstyle\sum_{i=1}^{p}\{I(i \bmod 2 = 0)\cdot(\frac{i}{2} + Y)x_i\} + \mathcal{N}(0, 3)\\ \delta_{nonlin} &= Y + \textstyle\sum_{i=1}^{p}\{I(i \bmod 2 = 0)\cdot(\frac{i}{2} + Y)x_i\} + \mathcal{N}(0, 3)\\ &+ \textstyle\sum_{i=1}^{p-1}\{I(i \bmod 5 = 0)\cdot(\frac{i}{5} + Y)x_i x_{i+1}\}\end{aligned}$$

Experiments on Synthetic Data

- Part of simulation results:

Setting 1: $\delta = \delta_{linear}$													
	n/p	$n = 5000, p = 50$			$n = 5000, p = 100$			$n = 10000, p = 50$			$n = 10000, p = 100$		
π_0	Estimator	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
π_{inv}	$\widehat{U}_{R\text{-}IPS}(\pi)$	7.306(1.632)	7.305	7.486	21.03(6.842)	21.03	22.11	7.083(1.399)	7.083	7.220	20.31(6.726)	20.31	21.40
	$\widehat{U}_{DM}(\pi)$	2.168(0.505)	2.168	2.226	3.612(1.274)	3.612	3.832	1.953(0.302)	1.953	1.975	3.439(1.104)	3.439	3.620
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.120(0.923)	0.787	0.927	0.577(3.865)	2.983	3.905	0.102(0.742)	0.641	0.746	**0.012**(3.015)	2.346	3.012
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.111(1.837)	1.496	1.839	0.058(7.736)	5.911	7.741	0.197(1.769)	1.486	1.780	0.360(7.382)	5.885	7.395
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.074(0.654)	0.540	0.659	**0.013**(1.696)	1.252	1.691	0.032(0.438)	0.350	0.438	0.430(1.299)	1.176	1.415
	$\widehat{U}_{DR}(\pi)$	0.056(0.576)	0.476	0.581	0.031(1.531)	1.079	1.512	0.021(0.398)	0.312	0.393	0.364(1.118)	0.974	1.197
	$\widehat{U}_{CB}(\pi)$	0.058(0.938)	0.755	0.942	0.093(3.363)	2.739	3.348	0.164(0.596)	0.499	0.620	0.256(2.681)	2.153	2.709
	$\widehat{U}_{FCB}(\pi)$	**0.008**(0.492)	**0.404**	**0.494**	0.128(1.250)	**0.904**	**1.295**	**0.014**(0.345)	**0.285**	**0.357**	0.213(0.935)	**0.775**	**0.972**

Estimated propensity score is better than true propensity score.
True propensity score is closer to 0 or 1, leading to high variance.

Experiments on Synthetic Data

- Part of simulation results:

Setting 1: $\delta = \delta_{linear}$													
	n/p	$n = 5000, p = 50$			$n = 5000, p = 100$			$n = 10000, p = 50$			$n = 10000, p = 100$		
π_0	Estimator	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
π_{inv}	$\widehat{U}_{R\text{-}IPS}(\pi)$	7.306(1.632)	7.305	7.486	21.03(6.842)	21.03	22.11	7.083(1.399)	7.083	7.220	20.31(6.726)	20.31	21.40
	$\widehat{U}_{DM}(\pi)$	2.168(0.505)	2.168	2.226	3.612(1.274)	3.612	3.832	1.953(0.302)	1.953	1.975	3.439(1.104)	3.439	3.620
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.120(0.923)	0.787	0.927	0.577(3.865)	2.983	3.905	0.102(0.742)	0.641	0.746	**0.012**(3.015)	2.346	3.012
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.111(1.837)	1.496	1.839	0.058(7.736)	5.911	7.741	0.197(1.769)	1.486	1.780	0.360(7.382)	5.885	7.395
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.074(0.654)	0.540	0.659	**0.013**(1.696)	1.252	1.691	0.032(0.438)	0.350	0.438	0.430(1.299)	1.176	1.415
	$\widehat{U}_{DR}(\pi)$	0.056(0.576)	0.476	0.581	0.031(1.531)	1.079	1.512	0.021(0.398)	0.312	0.393	0.364(1.118)	0.974	1.197
	$\widehat{U}_{CB}(\pi)$	0.058(0.938)	0.755	0.942	0.093(3.363)	2.739	3.348	0.164(0.596)	0.499	0.620	0.256(2.681)	2.153	2.709
	$\widehat{U}_{FCB}(\pi)$	**0.008**(0.492)	**0.404**	**0.494**	0.128(1.250)	**0.904**	**1.295**	**0.014**(0.345)	**0.285**	**0.357**	0.213(0.935)	**0.775**	**0.972**

CB estimator performs not very well.
Because it ignores the action discrepancy from the new policy

Experiments on Synthetic Data

- Part of simulation results:

Setting 1: $\delta = \delta_{linear}$													
	n/p	$n = 5000, p = 50$			$n = 5000, p = 100$			$n = 10000, p = 50$			$n = 10000, p = 100$		
π_0	Estimator	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
π_{inv}	$\widehat{U}_{R\text{-}IPS}(\pi)$	7.306(1.632)	7.305	7.486	21.03(6.842)	21.03	22.11	7.083(1.399)	7.083	7.220	20.31(6.726)	20.31	21.40
	$\widehat{U}_{DM}(\pi)$	2.168(0.505)	2.168	2.226	3.612(1.274)	3.612	3.832	1.953(0.302)	1.953	1.975	3.439(1.104)	3.439	3.620
	$\widehat{U}_{E\text{-}IPS}(\pi)$	0.120(0.923)	0.787	0.927	0.577(3.865)	2.983	3.905	0.102(0.742)	0.641	0.746	**0.012**(3.015)	2.346	3.012
	$\widehat{U}_{T\text{-}IPS}(\pi)$	0.111(1.837)	1.496	1.839	0.058(7.736)	5.911	7.741	0.197(1.769)	1.486	1.780	0.360(7.382)	5.885	7.395
	$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.074(0.654)	0.540	0.659	**0.013**(1.696)	1.252	1.691	0.032(0.438)	0.350	0.438	0.430(1.299)	1.176	1.415
	$\widehat{U}_{DR}(\pi)$	0.056(0.576)	0.476	0.581	0.031(1.531)	1.079	1.512	0.021(0.398)	0.312	0.393	0.364(1.118)	0.974	1.197
	$\widehat{U}_{CB}(\pi)$	0.058(0.938)	0.755	0.942	0.093(3.363)	2.739	3.348	0.164(0.596)	0.499	0.620	0.256(2.681)	2.153	2.709
	$\widehat{U}_{FCB}(\pi)$	**0.008**(0.492)	**0.404**	**0.494**	0.128(1.250)	**0.904**	**1.295**	**0.014**(0.345)	**0.285**	**0.357**	0.213(0.935)	**0.775**	**0.972**

With considering the action discrepancy, Our FCB estimator can consistently improve the performance of policy evaluation.

Experiment - Classifier evaluation

- A classifier can be defined as a policy based on a given dataset
  - Features of samples~ context $\mathbf{X}$
  - Predicted label of samples ~ action Y predicted by the classifier
  - Feedback function: $\delta(\mathbf{X}, Y) = I(Y = Y^t)$. (Y^t is the true label)
  - The policy evaluation is equivalent to the evaluation of the classifier accuracy
- Datasets: several multiclass classification bench-mark from UCI-repository.
- The new policy to be evaluated
  - Logistic regression model trained on the training set
- The past policy:
  - A simple function based on one feature variable

Experiments - Classifier evaluation

Estimator	Dataset:glass			Dataset:wilt			Dataset:pageblock			Dataset:particle		
	Bias(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE	*Bias*(SD)	MAE	RMSE
$\widehat{U}_{R\text{-}IPS}(\pi)$	0.711(7.805)	5.961	7.837	0.750(1.090)	1.112	1.323	42.19(2.711)	42.19	42.28	4.093(0.432)	4.093	4.116
$\widehat{U}_{DM}(\pi)$	8.810(4.164)	8.810	9.744	0.096(0.380)	0.309	0.391	2.224(0.444)	2.224	2.267	0.741(0.228)	0.741	0.776
$\widehat{U}_{E\text{-}IPS}(\pi)$	1.648(5.707)	4.739	5.940	0.128(0.323)	0.267	0.347	4.723(3.991)	5.788	6.184	0.230(0.281)	0.287	0.362
$\widehat{U}_{T\text{-}IPS}(\pi)$	1.488(6.162)	4.866	6.339	0.175(1.205)	0.983	1.217	**0.324**(2.327)	1.794	2.348	**0.012**(0.553)	0.447	0.554
$\widehat{U}^{SN}_{E\text{-}IPS}(\pi)$	0.315(5.455)	4.447	5.465	0.121(0.322)	0.265	0.343	1.539(2.326)	2.247	2.788	0.091(0.277)	0.222	0.293
$\widehat{U}_{CB}(\pi)$	**0.094**(6.364)	5.028	6.365	0.165(0.337)	0.318	0.372	4.660(2.810)	5.014	5.442	0.277(0.325)	0.347	0.429
$\widehat{U}_{DR}(\pi)$	1.035(5.334)	4.420	5.434	0.129(0.323)	0.269	0.347	1.734(1.978)	2.152	2.630	0.124(0.276)	0.228	0.303
$\widehat{U}_{FCB}(\pi)$	0.562(5.242)	**4.098**	**5.273**	**0.024**(0.329)	**0.250**	**0.328**	0.747(0.617)	**0.791**	**0.968**	0.080(0.261)	**0.215**	**0.272**

By simultaneously considering the distribution shift and action discrepancy, Our FCB algorithm performs the best for offline policy evaluation.

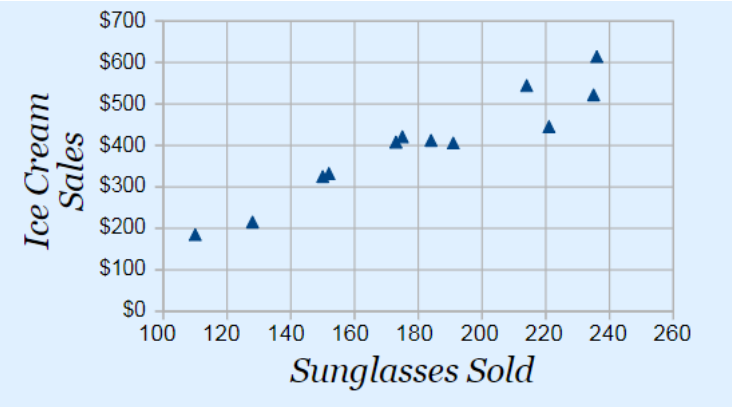
Summary: Causal Inference for Offline Policy Evaluation

- Challenges of offline policy evaluation:
  - Distribution shift induced by the past policy ← **Related work**
  - Action discrepancy induced by the new policy
  - Model dependency

- **Focused Context Balancing**
  - To remove the model dependency
  - Simultaneously consider distribution shift and action discrepancy
  - Significantly improve the accuracy on policy evaluation
  - Supporting for decision making, which policy is the best to deploy

Summary: Causally Regularized Machine Learning

- We have highly accurate predictions, but they are not enough for:
  - Interpretable prediction
  - Stable/Robust prediction in the future
  - Decision making




Algorithm A


Algorithm B

Summary: Causally Regularized Machine Learning

- Causal Inference with Observational Data
  - Recover causation from observed correlation
  - Estimating causal effect for improving interpretability
- Causal Inference for Stable Prediction
  - Disrupt spurious correlation, embrace causation
  - Interpretable and Stable prediction in the future
- Causal Inference for Offline Policy Evaluation
  - Evaluating a new policy based on the log data from a past policy
  - Support decision making with the effect of new policies

OUTLINE

PART I. Introduction to Causal Inference

PART II. Methods for Causal Inference

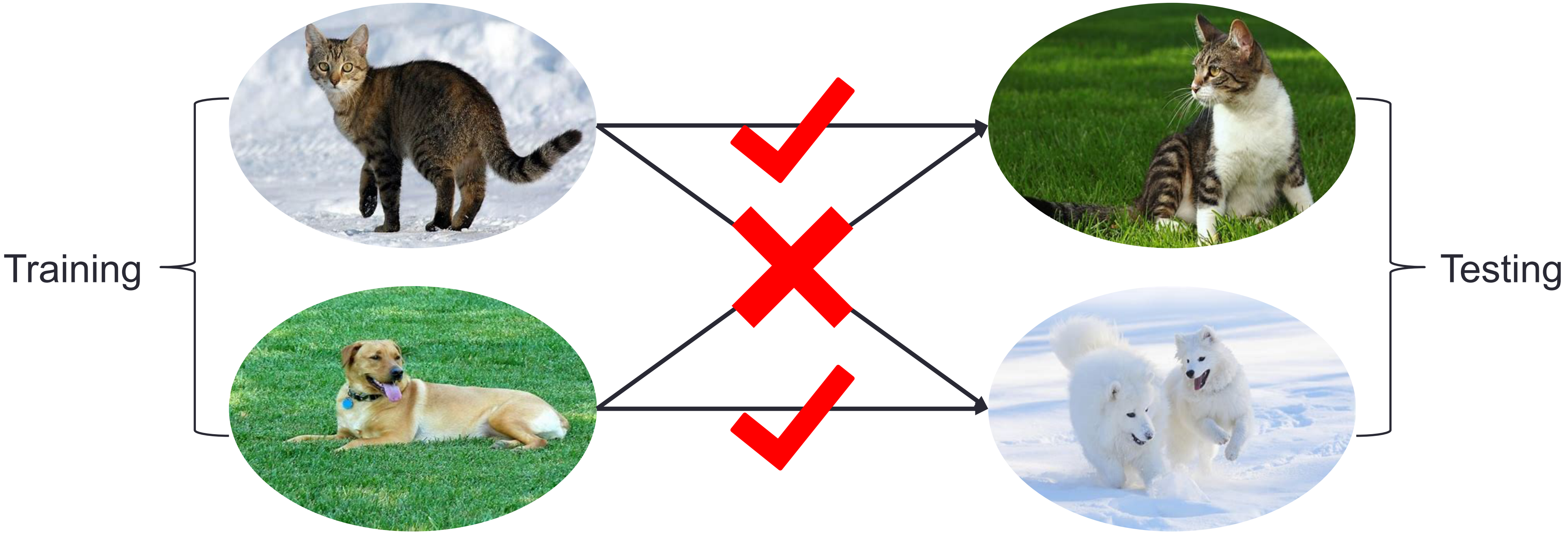
PART III. Causally Regularized Machine Learning

PART IV. Benchmark and Open Datasets

PART V. Conclusion and Discussion

TOWARDS NON-I.I.D. IMAGE CLASSIFICATION: A DATASET AND BASELINES

Correlation V.S. Causation



Non-I.I.D. Image Classification

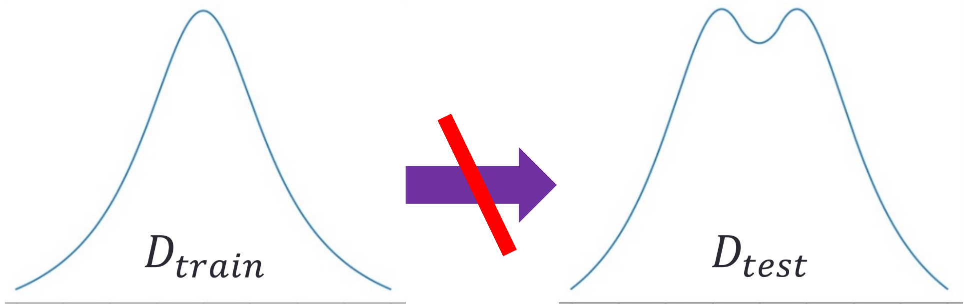
- Non I.I.D. Image Classification

$$\psi(D_{train} = (X_{train}, Y_{train})) \neq \psi(D_{test} = (X_{test}, Y_{test}))$$

known

unknown

- Two tasks
  - Targeted Non-I.I.D. Image Classification
    - Have prior knowledge on testing data
    - e.g. transfer learning, domain adaptation
  - General Non-I.I.D. Image Classification
    - Testing is unknown, no prior
    - more practical & realistic



Existence of Non-I.I.Dness

- One metric (NI) for Non-I.I.Dness

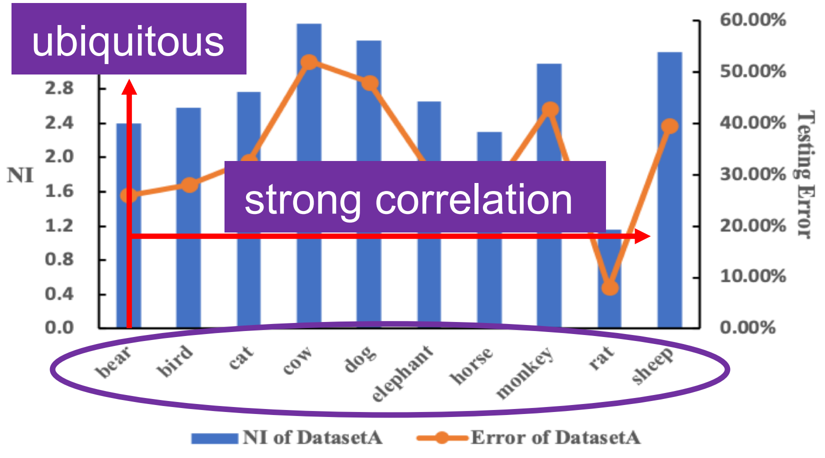
Definition 1 ***Non-I.I.D. Index (NI)*** *Given a feature extractor $g_\varphi(\cdot)$ and a class C,* ***the degree of distribution shift*** *between training data D^C_{train} and testing data D^C_{test} is defined as:*

$$NI(C) = \left\| \frac{\overline{g_\varphi(X^C_{train})} - \overline{g_\varphi(X^C_{test})}}{\sigma(g_\varphi(X^C_{train} \cup X^C_{test}))} \right\|_2,$$

Distribution shift

For normalization

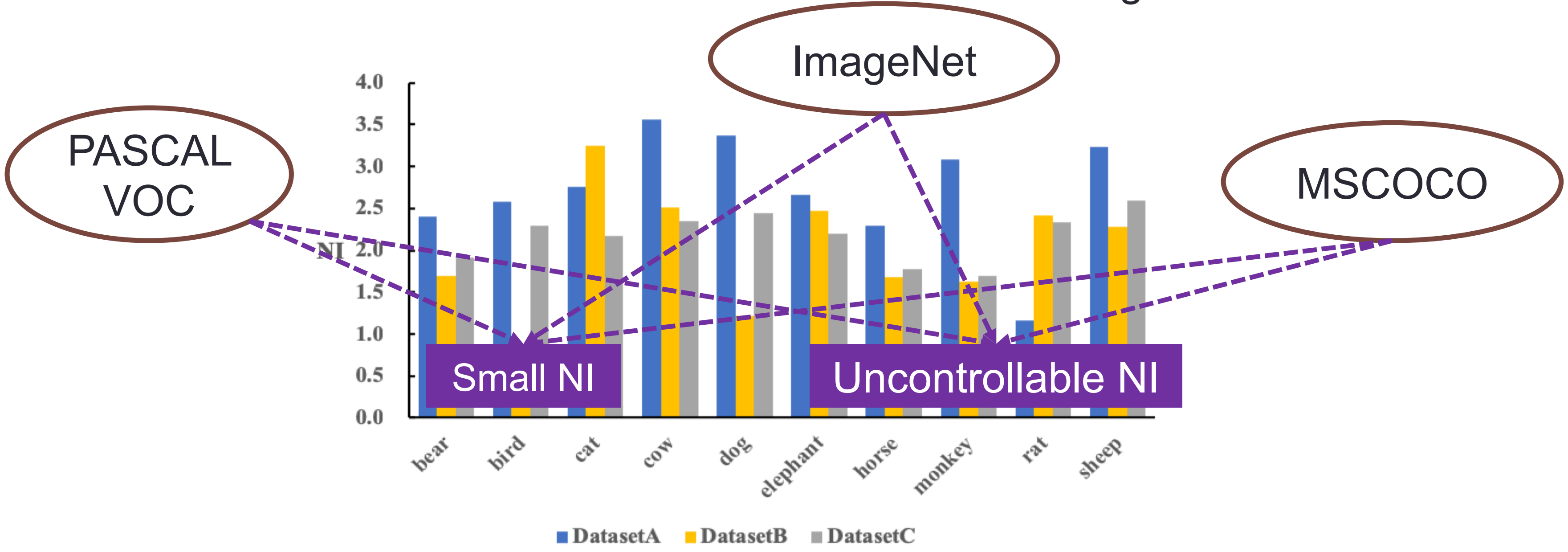
- Existence of Non-I.I.Dness on Dataset consisted of 10 subclasses from ImageNet
- For each class
  - Training data
  - Testing data
  - CNN for prediction



Related Datasets

- DatasetA & DatasetB & DatasetC
  - NI is ubiquitous, but small on these datasets
  - NI is Uncontrollable, not friendly for Non IID setting

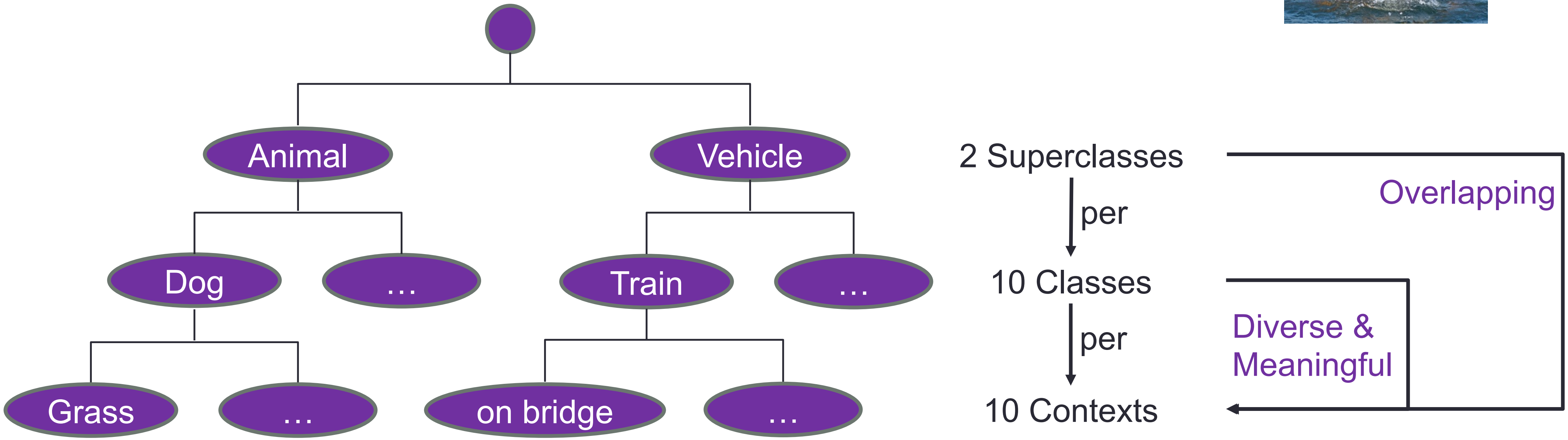
Average NI: 2.7



A dataset for Non-I.I.D. image classification is demanded.

NICO - Non-I.I.D. Image Dataset with Contexts

- **NICO** Datasets:
- Object label: e.g. dog
- Contextual labels (Contexts)
  - the background or scene of a object, e.g. grass/water
- Structure of NICO



NICO - Non-I.I.D. Image Dataset with Contexts

- Data size of each class in NICO
  - Sample size: thousands for each class
  - Each superclass: 10,000 images
  - Sufficient for some basic neural networks (CNN)

Animal	DATA SIZE	*Vehicle*	DATA SIZE
BEAR	1609	AIRPLANE	930
BIRD	1590	BICYCLE	1639
CAT	1479	BOAT	2156
COW	1192	BUS	1009
DOG	1624	CAR	1026
ELEPHANT	1178	HELICOPTER	1351
HORSE	1258	MOTORCYCLE	1542
MONKEY	1117	TRAIN	750
RAT	846	TRUCK	1000
SHEEP	918		

- Samples with contexts in NICO

Controlling NI on NICO Dataset

- Minimum Bias (comparing with ImageNet)
- Proportional Bias (controllable)
    - Number of samples in each context
- Compositional Bias (controllable)
    - Number of contexts that observed

Minimum Bias

- In this setting, the way of random sampling leads to minimum distribution shift between training and testing distributions in dataset, which simulates a nearly i.i.d. scenario.
  - 8000 samples for training and 2000 sample for testing in each superclass (ConvNet)

	Average NI	Testing Accuracy
Animal	3.85	49.6%
Vehicle	3.20	63.0%

Average NI on ImageNet: 2.7

Images in our NICO are with rich contextual information

more challenging for image classification

Our NICO data is more Non-iid, more challenging

Proportional Bias

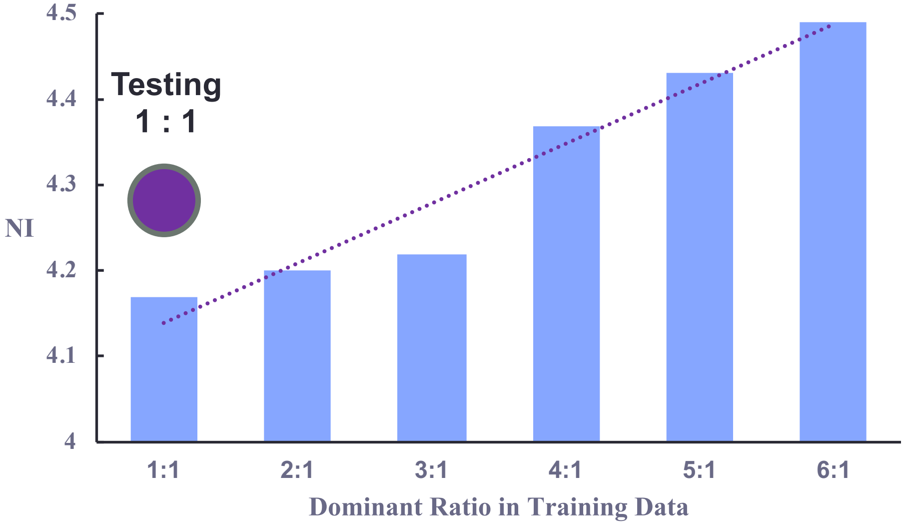
- Given a class, when sampling positive samples, we use all contexts for both training and testing, but the percentage of each context is different between training and testing dataset.

Dog

(5%) (5%) (5%) (5%) (5%) (5%) (5%) (5%) (5%)

Dominate Context (55%)

$$Dominant\ Ratio = \frac{N_{dominant}}{N_{minor}}$$

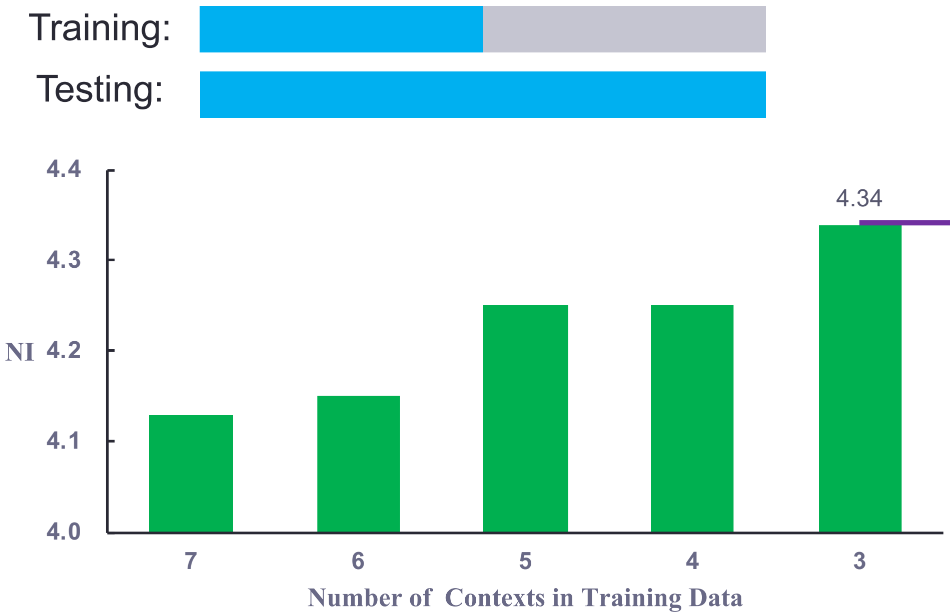


We can control NI by varying dominate ratio

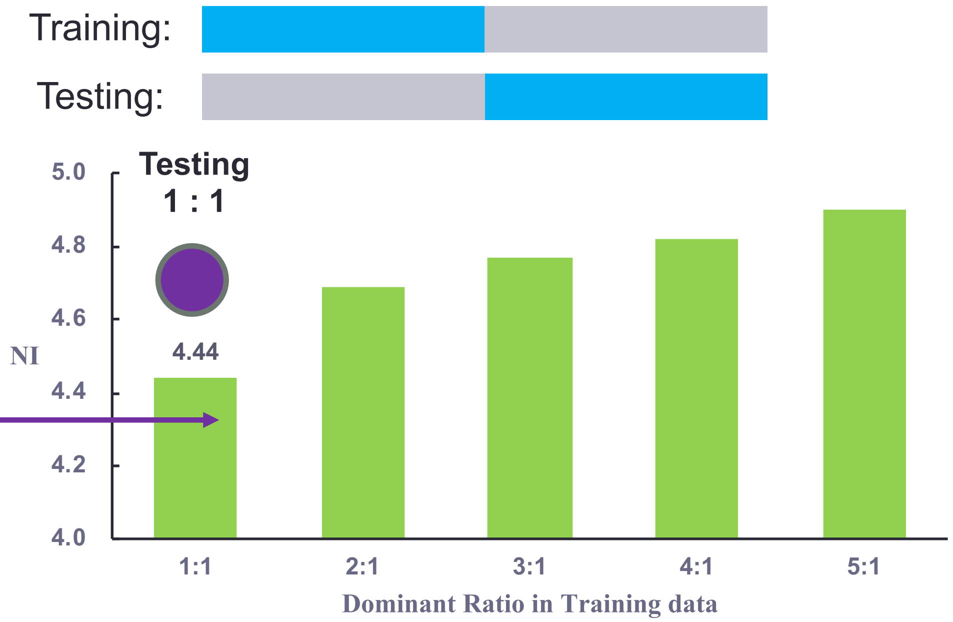
Compositional Bias

$$Dominant\ Ratio = \frac{N_{dominant}}{N_{minor}}$$

- Given a class, the observed contexts are different between training and testing data.



Moderate setting
(Overlap)



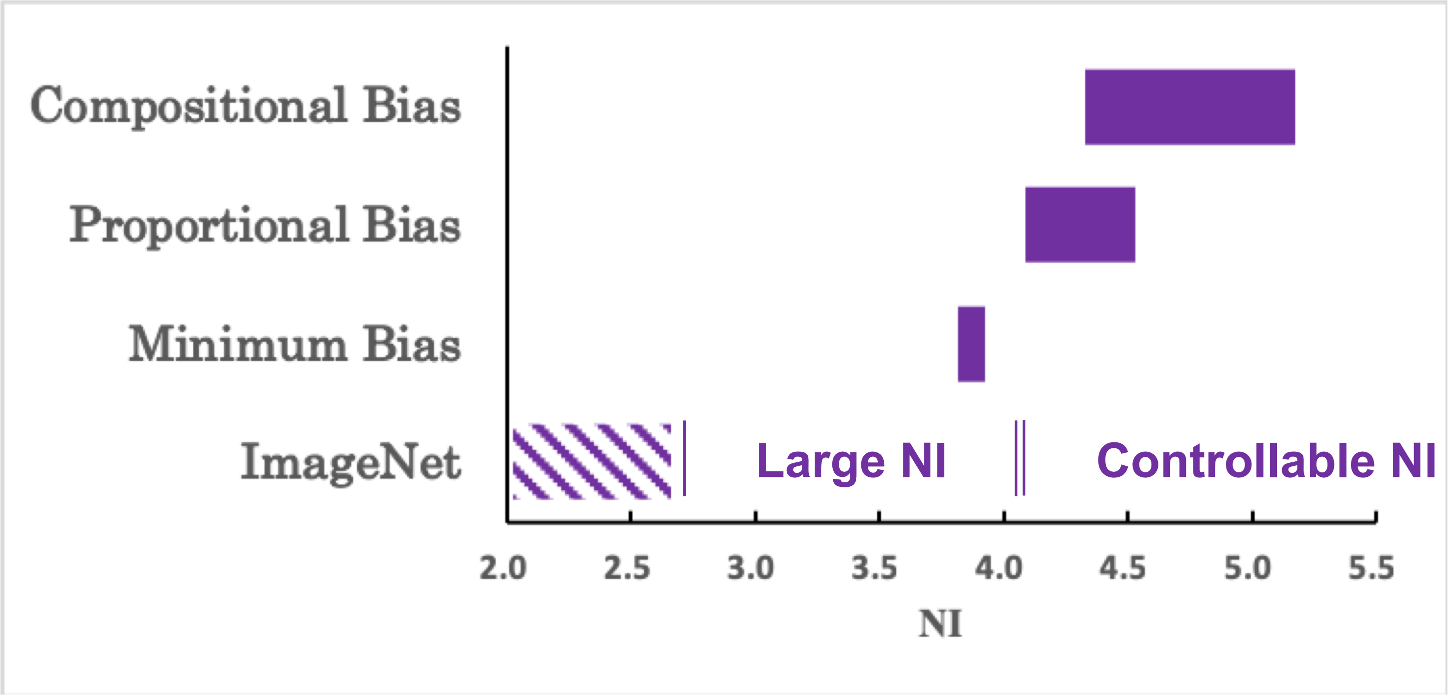
Radical setting
(No Overlap &
Dominant ratio)

NICO - Non-I.I.D. Image Dataset with Contexts

- Summary on Non-iidness on our dataset
- Range of NI value for each method
- Large and controllable NI

large NI

small NI

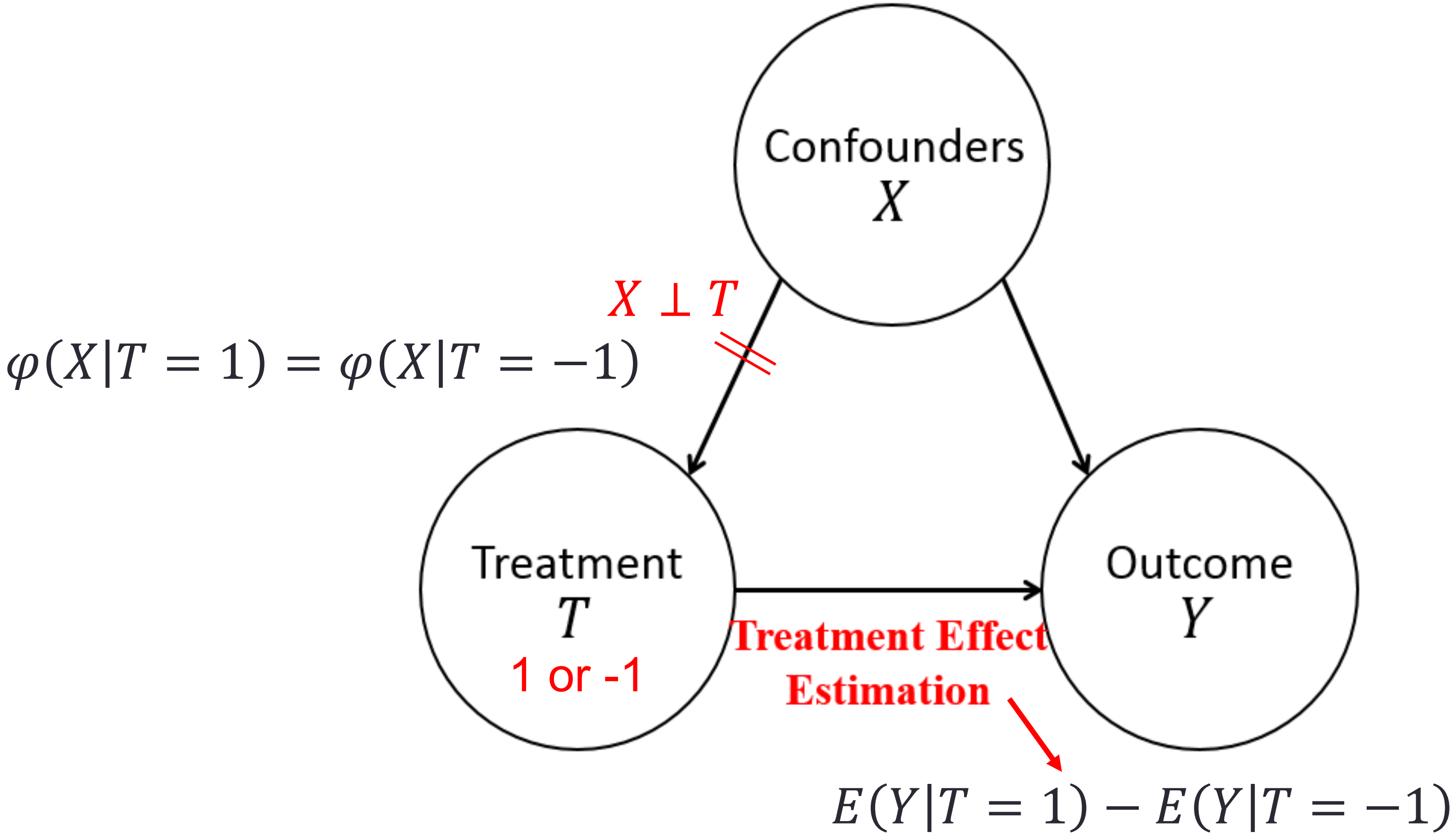


Targeted/General Non-I.I.D. Image Classification

Global Balancing Method

ConvNet with Batch Balancing (CNBB)

- Confounder Balancing in the literature of Causal Inference

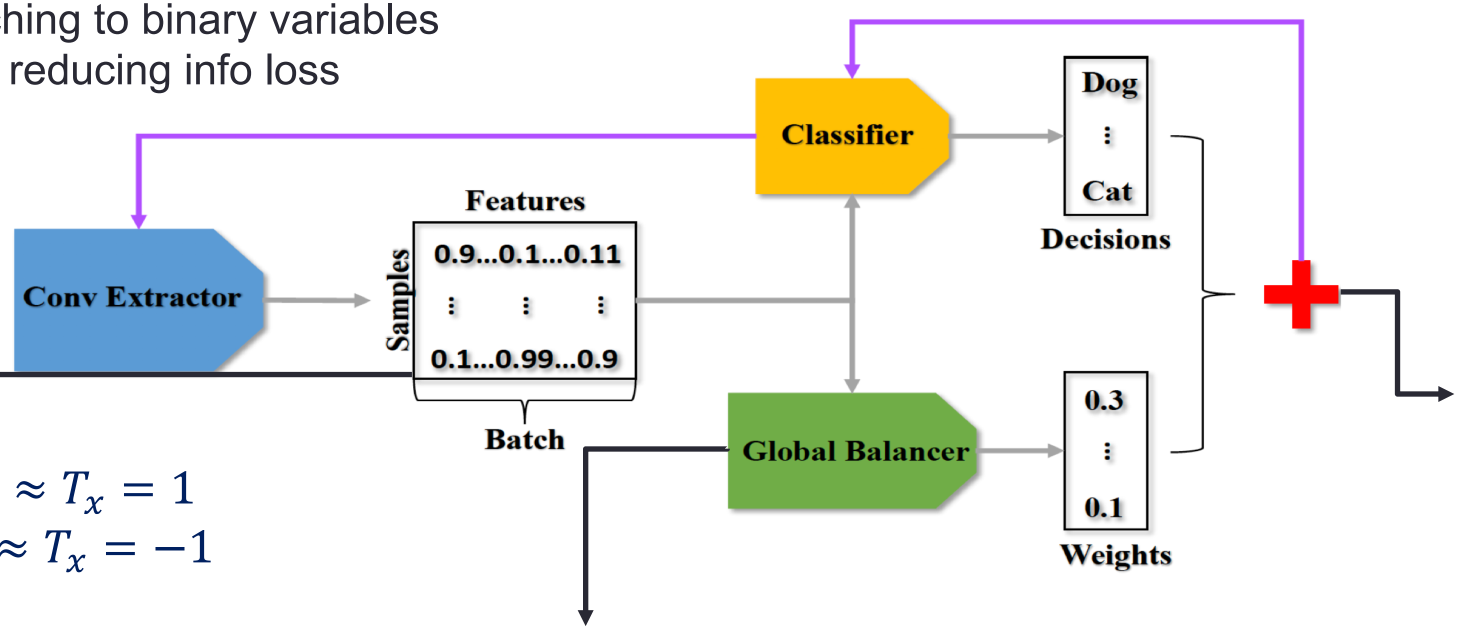


CNBB = Confounder balancing + ConvNet

ConvNet with Batch Balancing (CNBB)

$$Lossq = -\sum_{i=1}^{n} \|g_\varphi(x_i))\|_2^2$$

approaching to binary variables
for reducing info loss



$$\min_{\theta,\varphi} Lossp = \sum_{i=1}^{n} w_i \ln(f_\theta(g_\varphi(x_i)) \cdot y_i) + \lambda Lossq$$

learning causal effects &
causal features jointly

$X > 0 \approx T_x = 1$
$X < 0 \approx T_x = -1$

$$\min_W Lossb = \sum_{j=1}^{p} \left\| \frac{g_\varphi(X)_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{g_\varphi(X)_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2$$

$$+ \alpha \|W\|_2^2 \qquad s.t. \sum_{i=1}^{n} W_i = 1,\ W \geq 0$$

global confounder balancing
for all features

Experiments

- We design four experiments according to the supported Non-I.I.D. settings of NICO:
  - Minimum bias (Exp 1)
    - Nearly I.I.D. in NICO (average improvement **0.33%**)
  - Proportional bias (Exp2)
    - Different dominate ratio
    - fix dominant ratio of training to 5:1
    - vary dominant ratio of testing from 1:5 to 4:1
  - Compositional bias (Exp3)
    - Different observed contexts
    - Testing: with all contexts
    - Training: vary observed contexts from 3 to 7
  - Combined Proportional & Compositional bias (Exp4)
    - No overlap on the observed contexts
    - Different dominate ratio
    - fix dominant ratio of testing to 1:1
    - vary dominant ratio of training from 1:1 to 5:1

always superior

Exp2	1 : 5	1 : 1	2 : 1	3 : 1	4 : 1
CNN	37.17	37.80	41.46	42.50	43.23
CNN+BN	38.70	**39.60**	41.64	42.00	43.85
CNBB	**39.06**	**39.60**	**42.12**	**43.33**	**44.15**

Table 1. Performances of different methods on test accuracy (%) for proportional bias in *Animal* superclass.

Exp3	3	4	5	6	7
CNN	40.61	42.32	43.34	44.03	44.03
CNN+BN	**41.98**	38.85	43.12	44.71	44.31
CNBB	41.41	**43.34**	**44.54**	**45.96**	**45.16**

Table 2. Performances of different methods on test accuracy (%) for composional bias in *Vehicle* superclass.

Exp4	1 : 1	2 : 1	3 : 1	4 : 1	5 : 1
CNN	37.07	35.20	34.53	34.13	33.73
CNN+BN	33.87	32.93	31.20	30.93	30.67
CNBB	**38.98**	**36.89**	**35.87**	**35.33**	**35.02**

Table 3. Performances of different methods of test accuracy (%) for combined proportional & compositional bias in *Vehicle* superclass.

Summary on Experimental Results

- The range of NI with respect to the average improvement of performance to CNN

Experiment	Improvement	NI
Exp1	0.33%	3.81 - 3.93
Exp2	1.22%	4.17 - 4.53
Exp3	1.22%	4.13 - 4.34
Exp4	1.49%	4.44 - 4.90

more effect

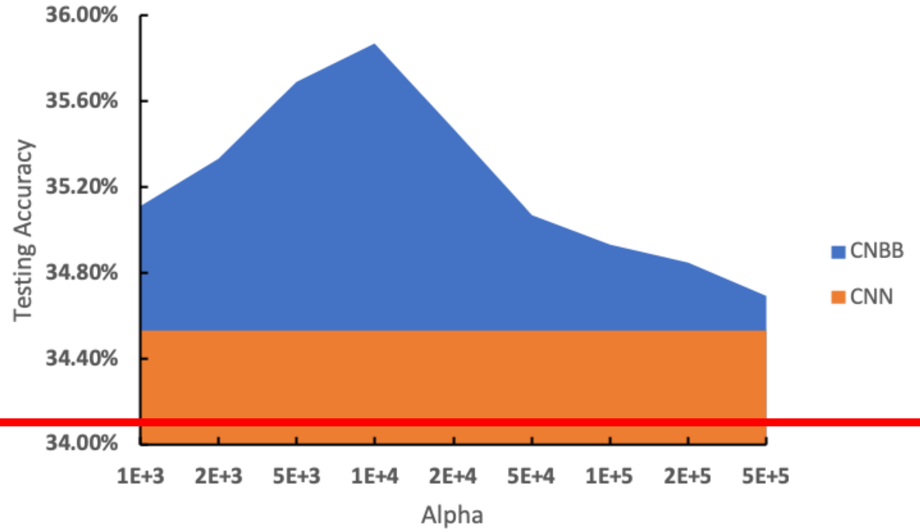
more bias

Analysis

- Insight of Batch Balancing Mechanism

$$\min_{W} Lossb = \sum_{j=1}^{p} \left\| \frac{g_\varphi(X)_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{g_\varphi(X)_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2$$

$$+ \alpha \|W\|_2^2 \qquad s.t. \sum_{i=1}^{n} W_i = 1, \; W \geq 0$$



less valid samples

less balancing degree

Summary: NICO for Non-iid Image Classification

- **NICO**: Non-iid image classification dataset
  - Non-iid Index (NI) to describe the distribution shift
  - Three ways to control NI in NICO Dataset
  - Benchmark for Non-iid image classification

- The performance of benchmark is not so exciting, more work need to do.
- How to use causal knowledge for Non-iid prediction

OUTLINE

Conclusion

- Correlation-based machine learning are not enough for
  - Interpretable learning
  - Decision making
  - Stable/Robust prediction in the future
- Correlation: causation, confounding, selection bias
  - Causation: Invariant and Stable across environments
  - Confounding / Selection bias: Spurious correlation, changeable
- Causally Regularized Machine Learning:
  - Causal regularizer
  - Recover causation from correlation
  - Causation-based machine learning

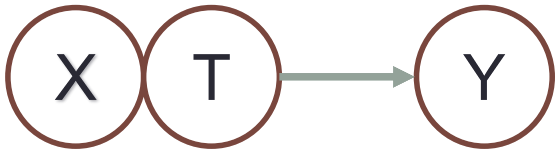
Conclusion

- Causally Regularized Machine Learning: Causation-based
  - Causal Inference for Interpretable learning
  - Policy Evaluation for Decision making
  - Causally Regularized Stable Prediction in the future

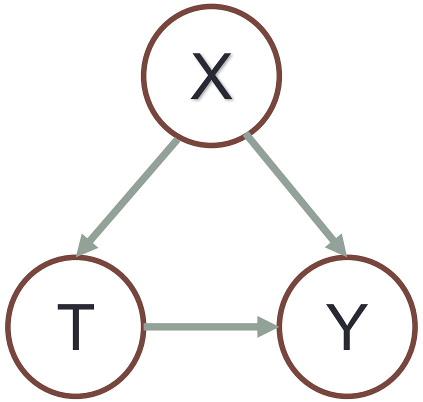
- **NICO**: Non-iid image classification dataset
  - Non-iid Index (NI) to describe the distribution shift
  - Three ways to control NI in NICO Dataset
  - Benchmark for Non-iid image classification

Future Work and Discussion

- Correlation
- Causation



Correlation Framework



Causal Framework

Recover causation from the observed correlation!

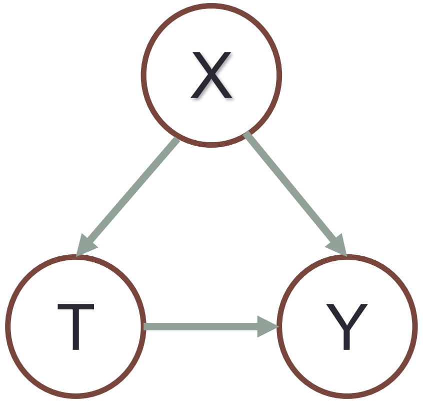
Future Work and Discussion

- With Causality, we can do:
  - Recover causation for interpretability
  - Help to guide decision making (actionable)
  - Make stable and robust prediction in the future
  - Prevent algorithmic bias (Fairness)

- **Discard spurious correlation and embrace causality**
- **Do interpretable, actionable, stable, fairness prediction**

Future Work and Discussion

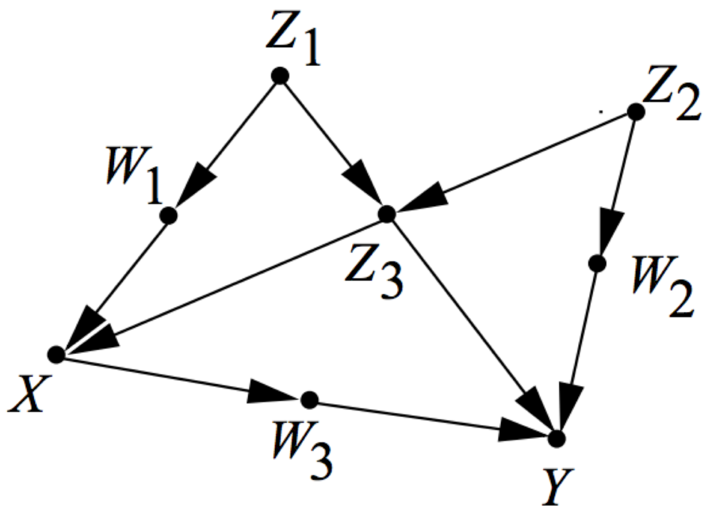
- Potential Outcome Framework
  - Rubin



Potential Outcome Framework

Many untestable assumptions

- Structural Causal Model (SCM)
  - Pearl



SCM

Strong prior knowledge

References

- [1] Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70(1): 41-55.
- [2] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research 46, 3 (2011), 399–424.
- [3] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis 20, 1 (2012), 25–46.
- [4] Athey S, Imbens G W, Wager S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2018, 80(4): 597-623.

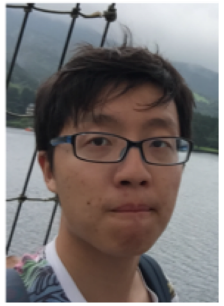
References

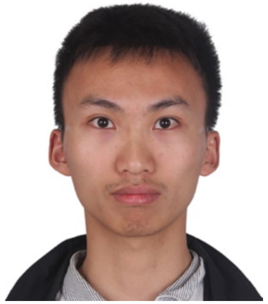
- [5] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. 2017. Treatment Effect Estimation with Data-Driven Variable Decomposition. In AAAI. 140–146.
- [6] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 265–274.
- [7] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 1617-1626.
- [8] Zheyan Shen, Peng Cui, Kun Kuang, et al. Causally regularized learning with agnostic data selection bias[C]//2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 411-419.

Acknowledgement









kkun2010@gmail.com

Thank You!