CAUSAL INFERENCE AND STABLE LEARNING

Peng Cui, Tsinghua University

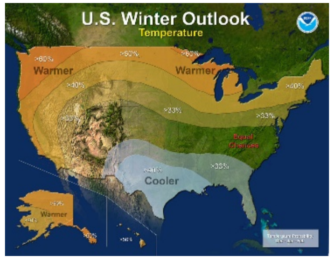
Kun Kuang, Zhejiang University

Bo Li, Tsinghua University

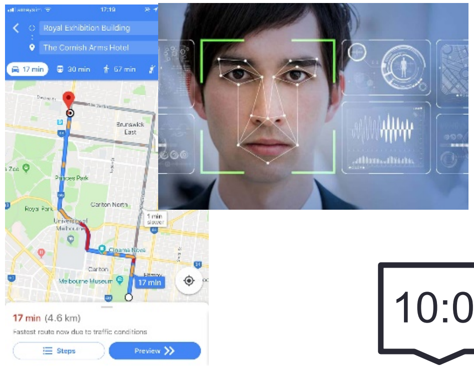
ML techniques are impacting our life

- A day in our life with ML techniques

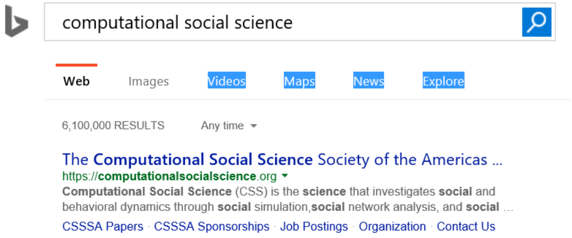
8:00 am



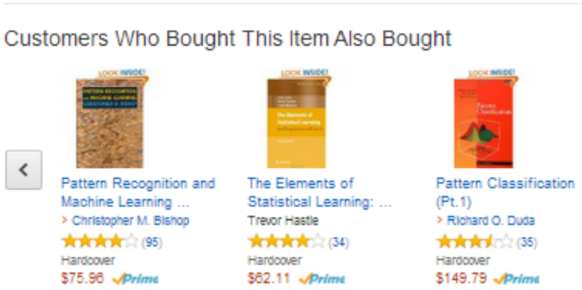
8:30 am



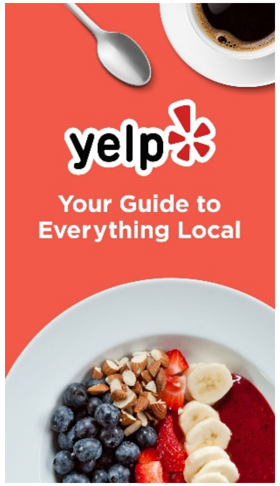
10:00 am



4:00 pm



6:00 pm



8:00 pm

Now we are stepping into risk-sensitive areas

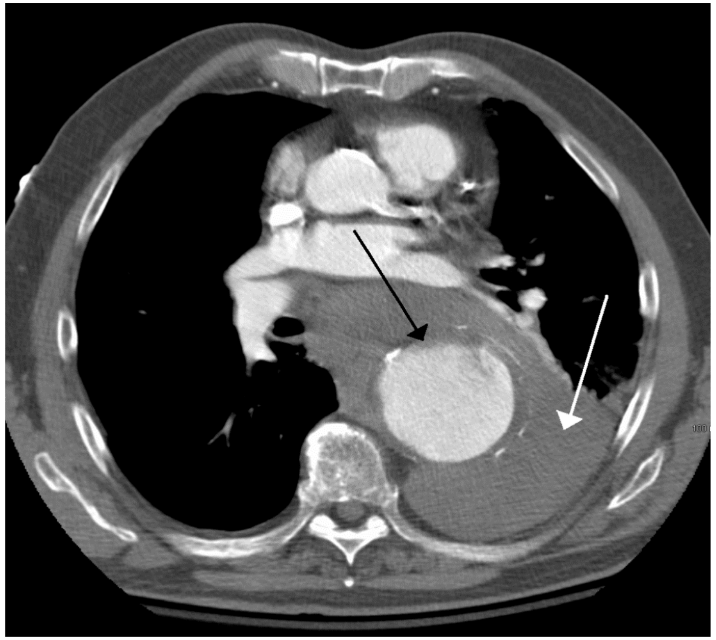
Human

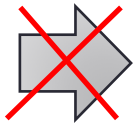
Shifting from ***Performance Driven*** to ***Risk Sensitive***

Problems of today's ML - *Explainability*

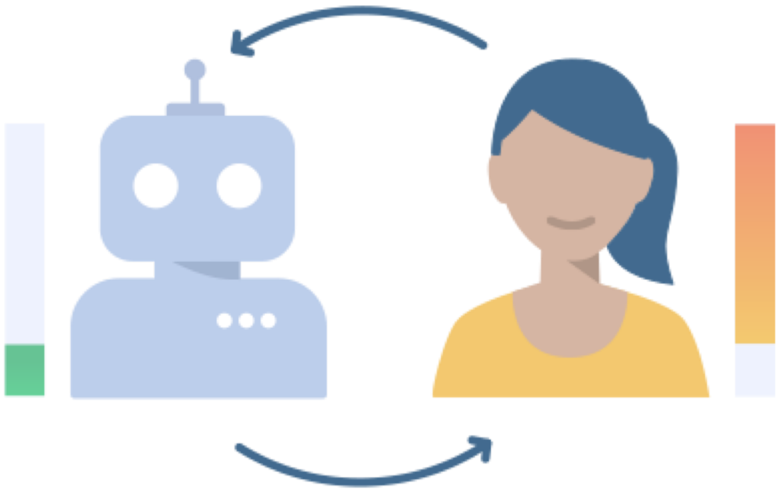
Most machine learning models are black-box models

Unexplainable





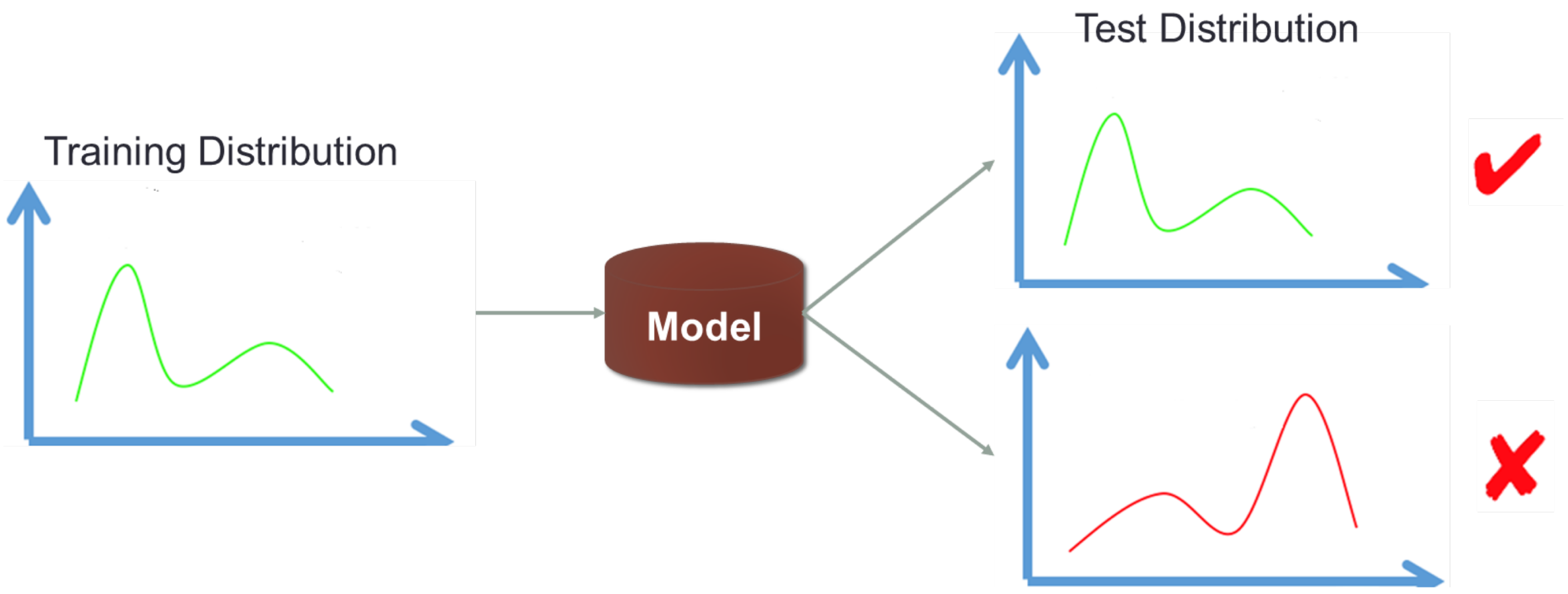
Human in the loop



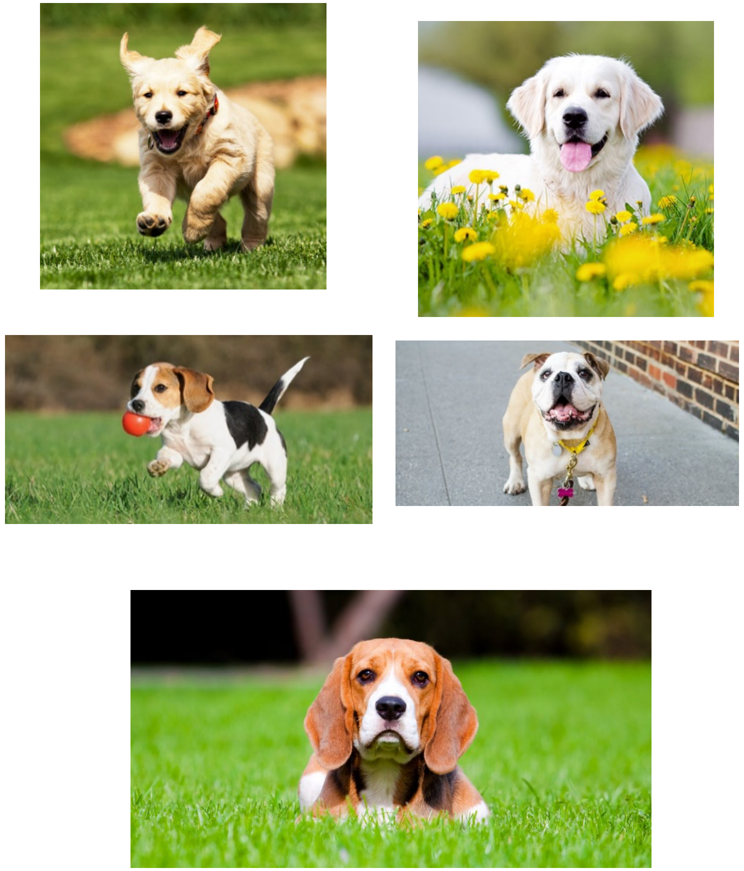
Health Military Finance Industry

Problems of today's ML - ***Stability***

Most ML methods are developed under I.I.D hypothesis



Problems of today's ML - ***Stability***



Yes

Maybe

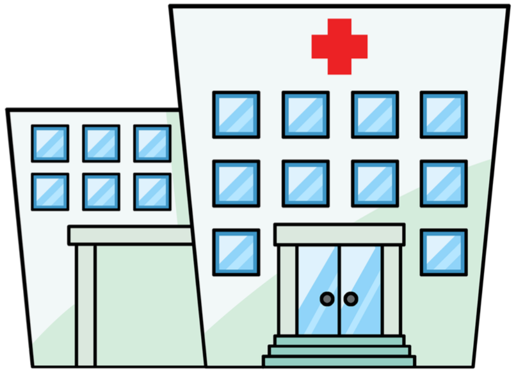
No

Problems of today's ML - ***Stability***

- Cancer survival rate prediction

Testing Data

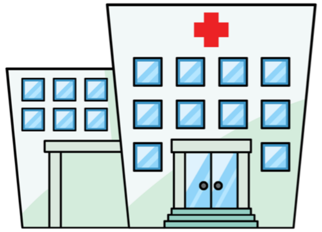
Training Data



City Hospital

Higher income, higher survival rate.

Predictive Model



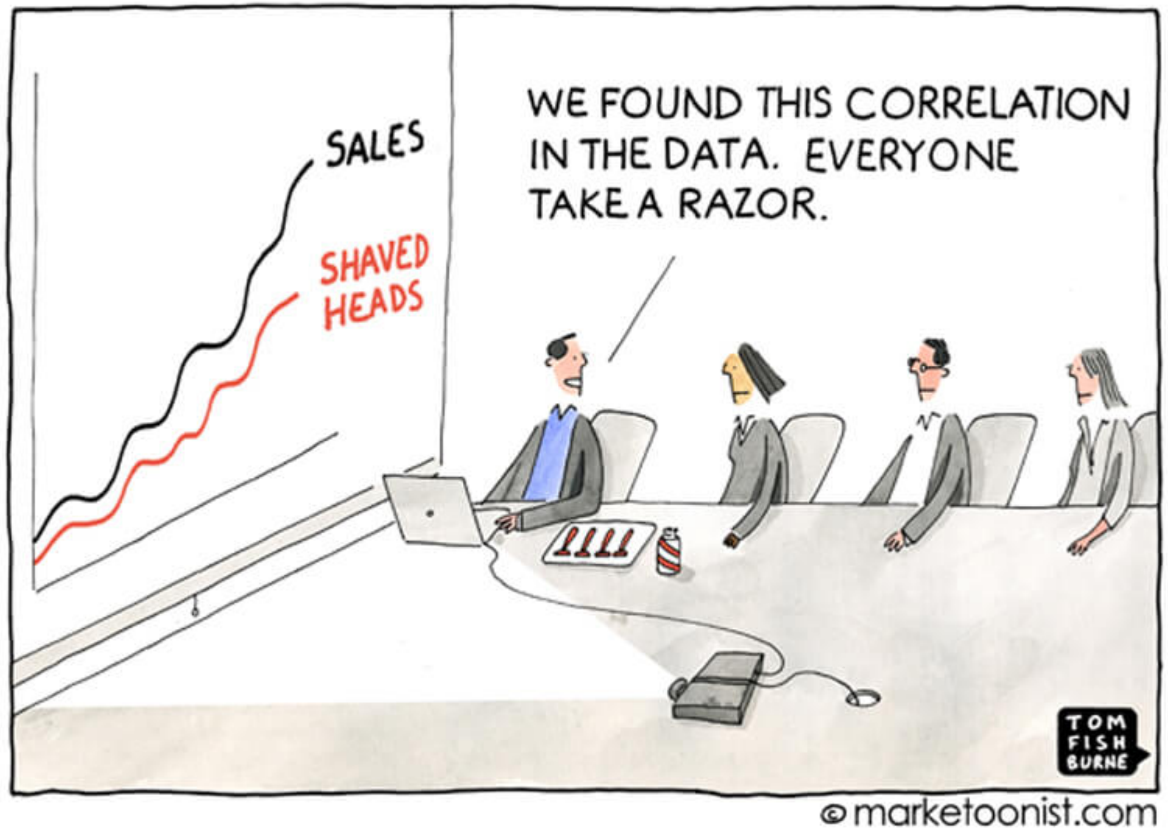
City Hospital

University Hospital

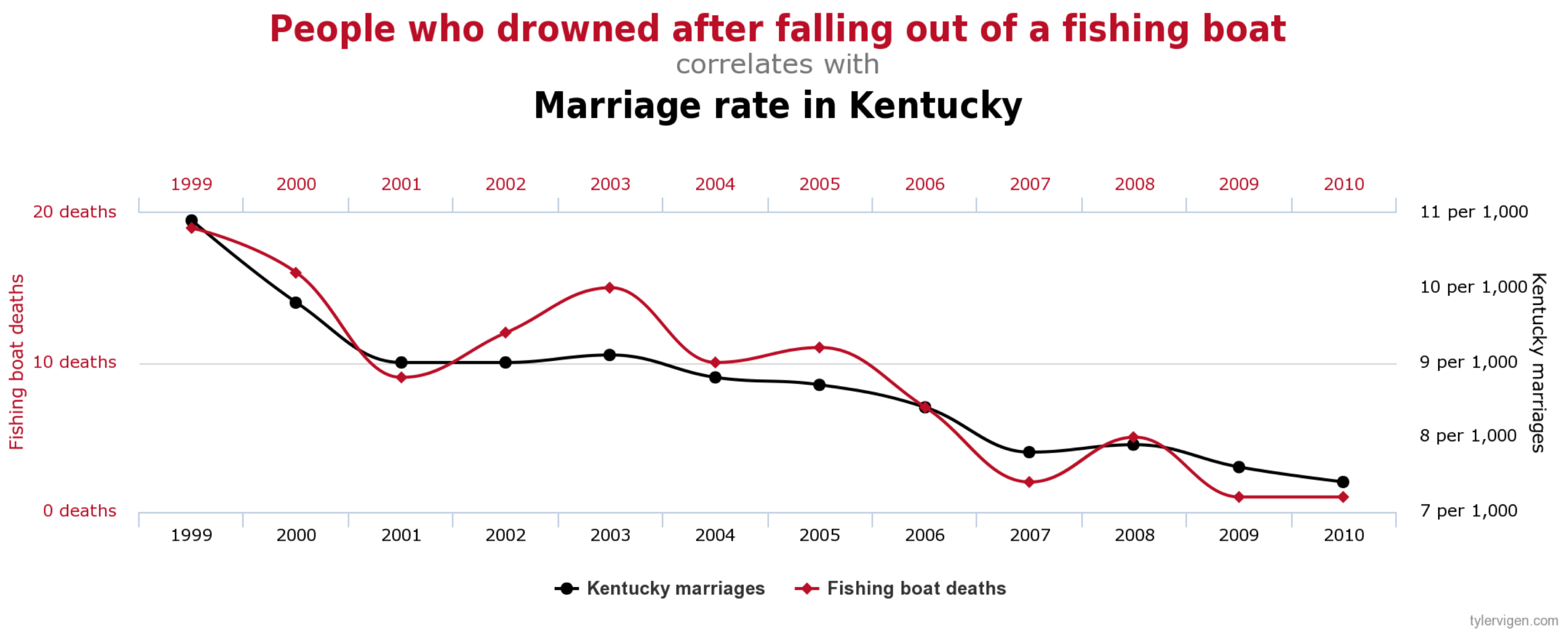
Survival rate is not so correlated with income.

A plausible reason: ***Correlation***

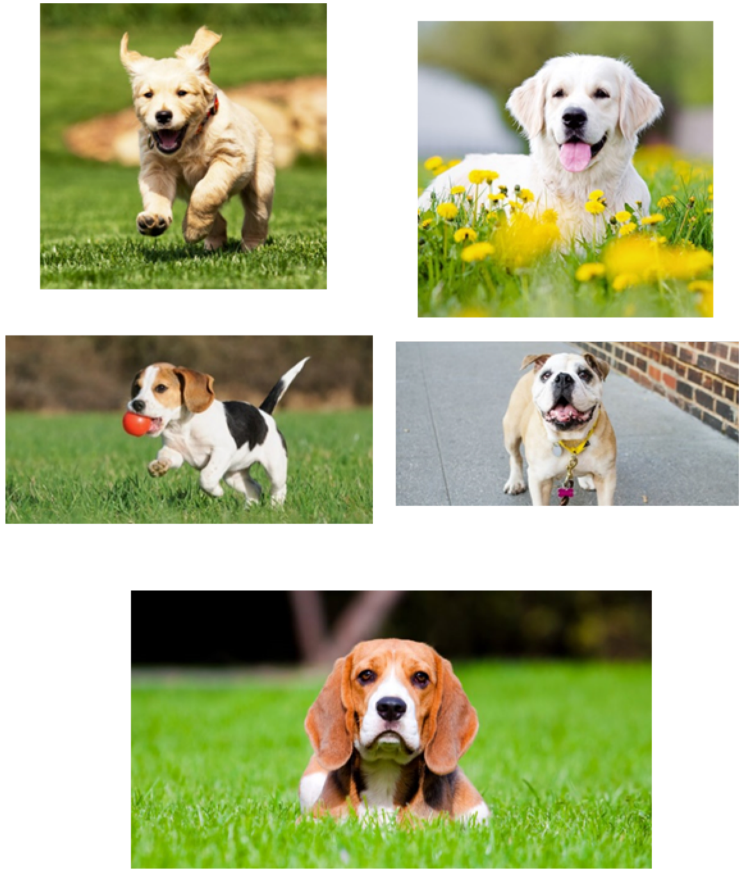
Correlation is the very basics of machine learning.

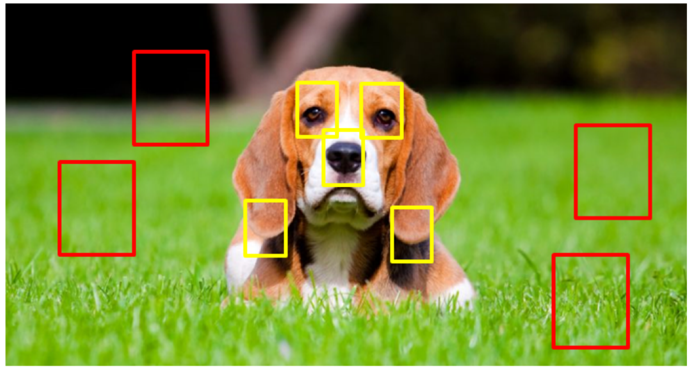


Correlation is not explainable



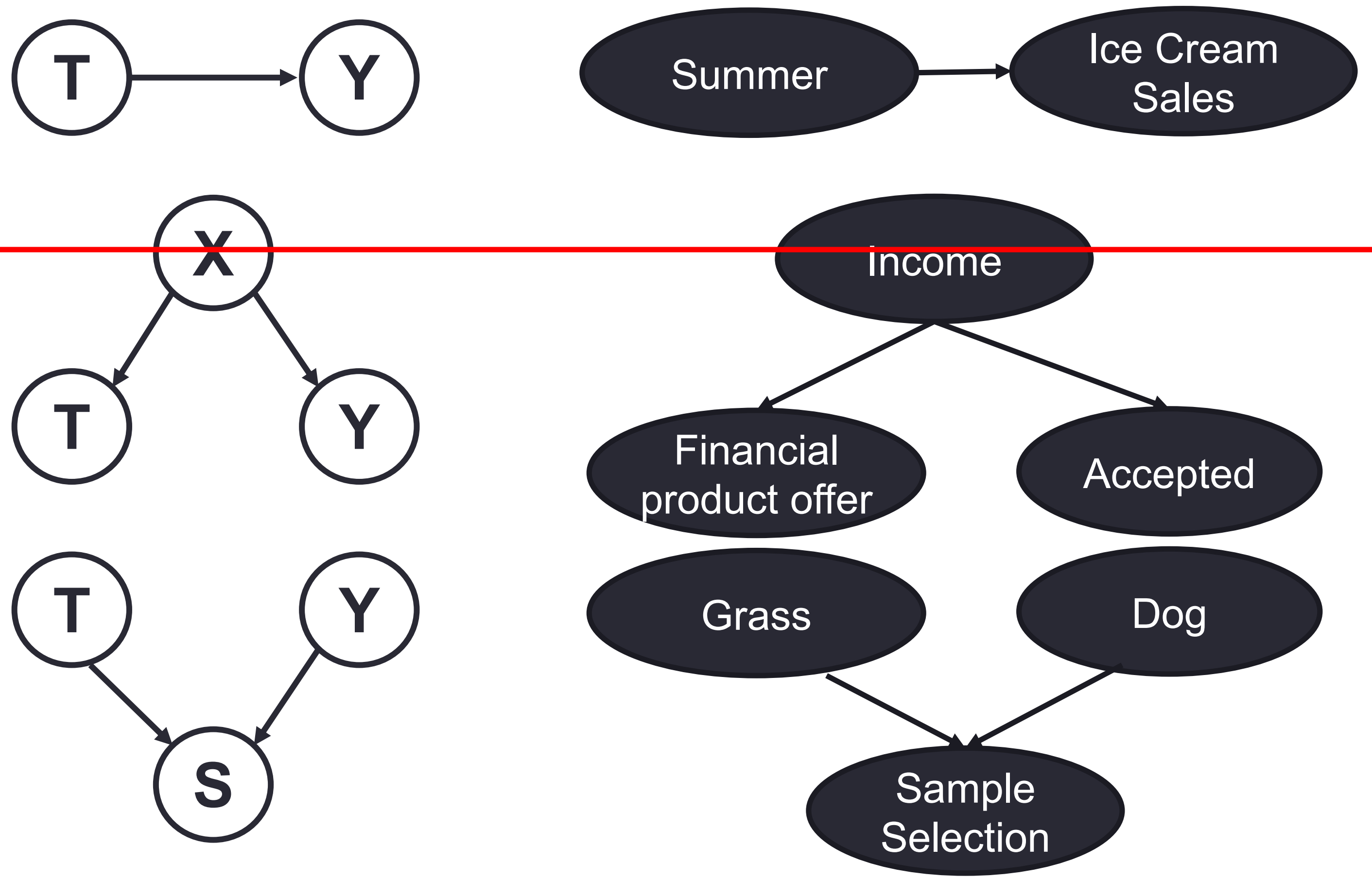
Correlation is ‘*unstable*’





It's not the fault of ***correlation***, but the way we use it

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

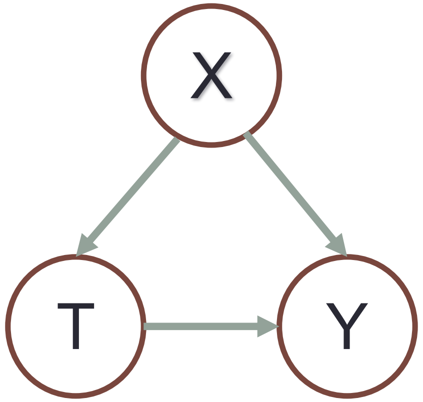


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



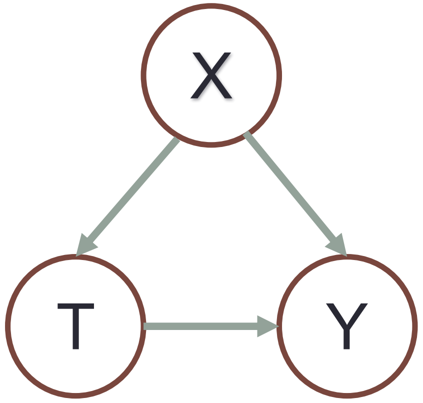
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

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

The *benefits* of bringing causality into learning

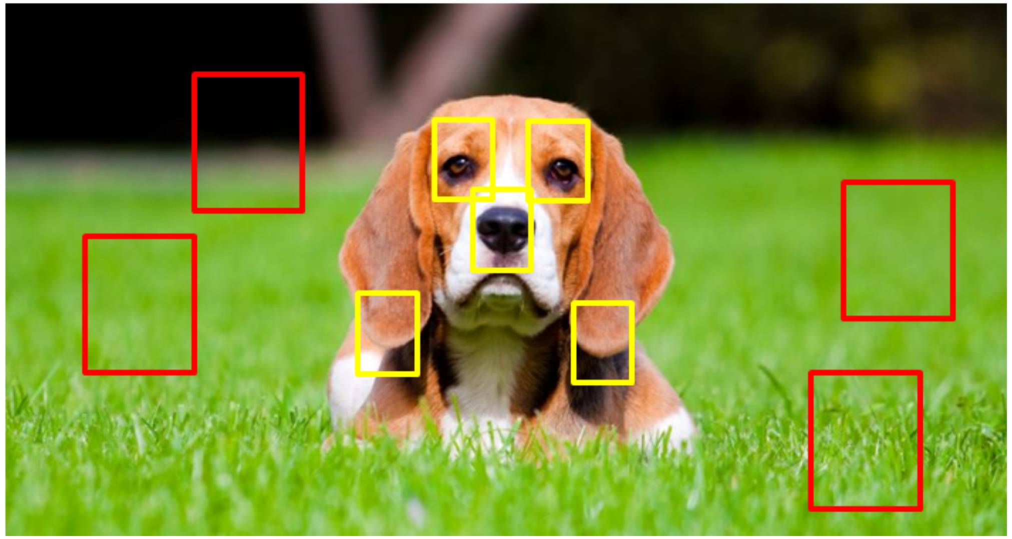
Causal Framework



T： grass
X： dog nose
Y： label

Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation



More ***Explainable*** and More ***Stable***

The *gap* between causality and learning

- How to evaluate the outcome?
- Wild environments
  - High-dimensional
  - Highly noisy
  - Little prior knowledge (model specification, confounding structures)
- Targeting problems
  - Understanding v.s. Prediction
  - Depth v.s. Scale and Performance

How to bridge the gap between *causality* and *(stable) learning*?

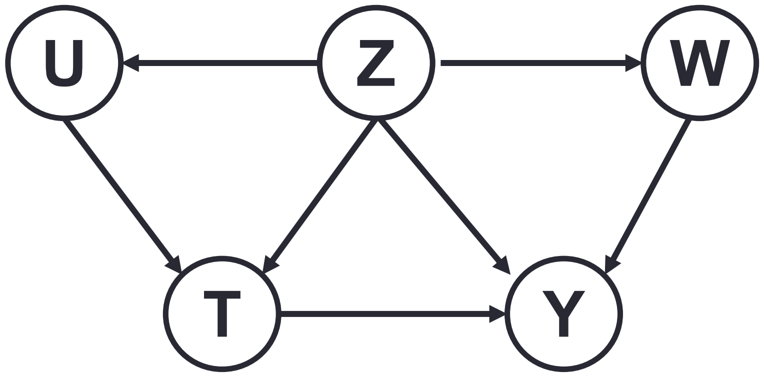
Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Conclusions

Paradigms - Structural Causal Model

A graphical model to describe the causal mechanisms of a system

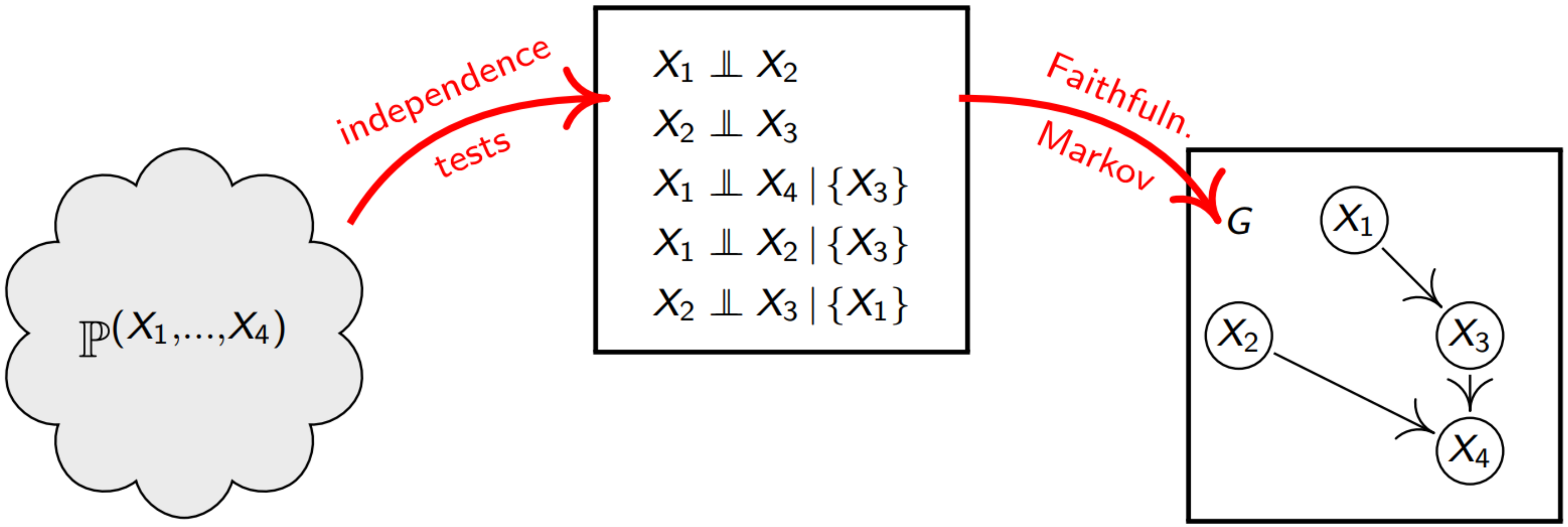
- Causal Identification with back door criterion
- Causal Estimation with do calculus



How to discover the causal structure?

Paradigms – Structural Causal Model

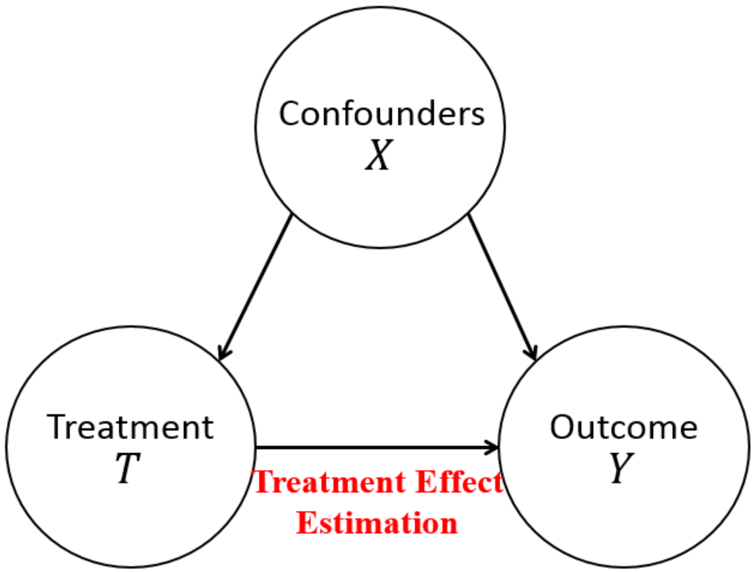
- Causal Discovery
  - Constraint-based: conditional independence
  - Functional causal model based



A ***generative*** model with strong expressive power.
But it induces high complexity.

Paradigms - Potential Outcome Framework

- A simpler setting
  - Suppose the confounders of T are known a priori
- The computational complexity is affordable
  - Under stronger assumptions
  - E.g. all confounders need to be observed

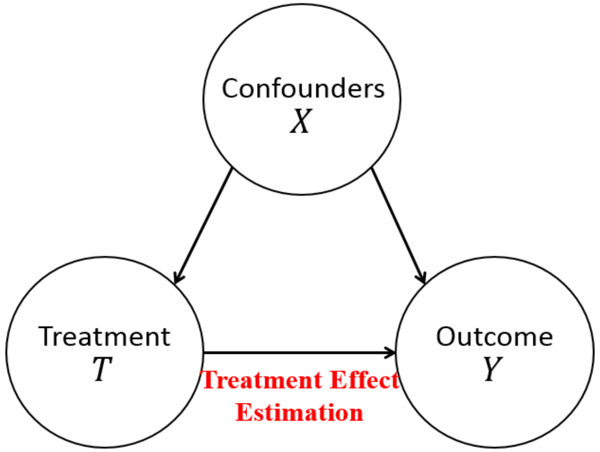


More like a ***discriminative*** way to estimate treatment's partial effect on outcome.

Causal Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Treated Group $(T = 1)$ and Control Group $(T = 0)$
- 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

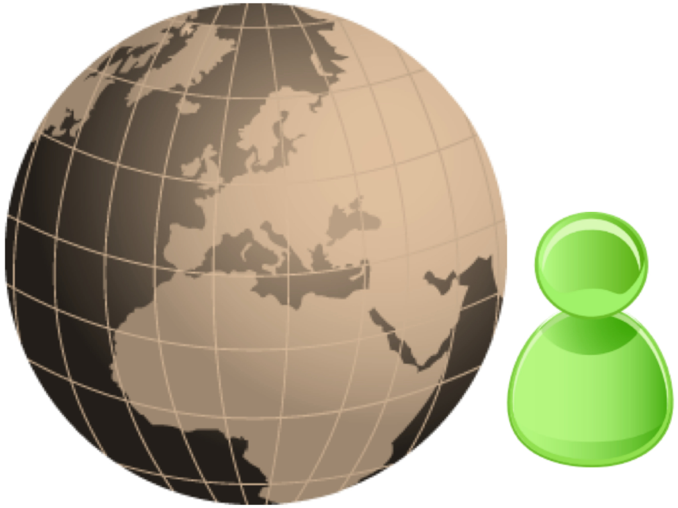
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	?
P2	0	?	0.6
P3	1	0.3	?
P4	0	?	0.1
P5	1	0.5	?
P6	0	?	0.5
P7	0	?	0.1

- 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

Ideal Solution: Counterfactual World

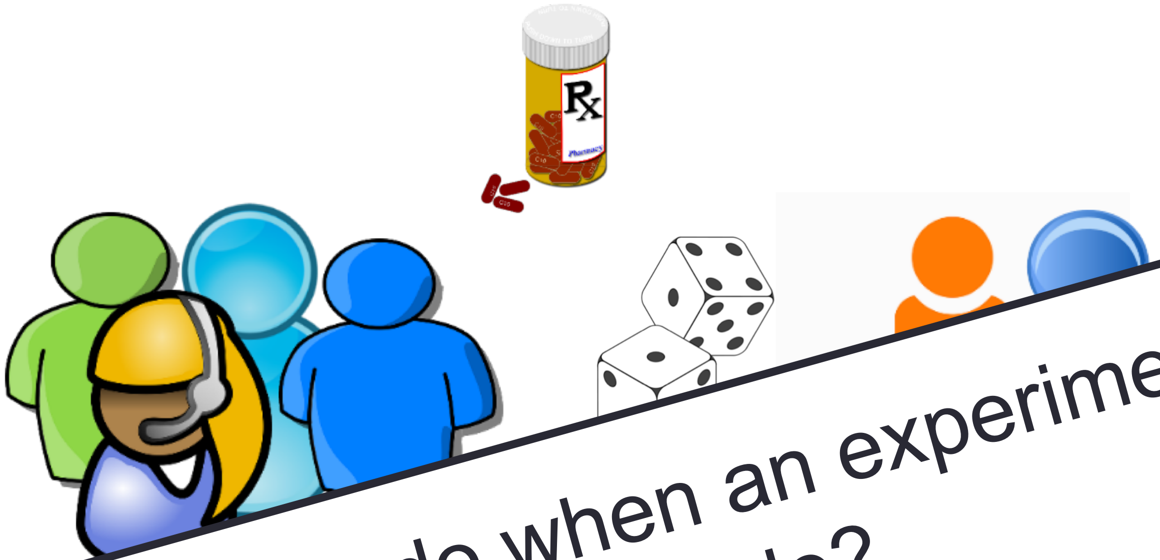
- Reason about a world that does not exist
- Everything in the counterfactual world is the same as the real world, except the treatment

$Y(T = 1)$



$Y(T = 0)$

Randomized Experiments are the "Gold Standard"



- Drawbac
  - Cost
  - Unethical
  - Unrealistic

What can we do when an experiment is not possible?

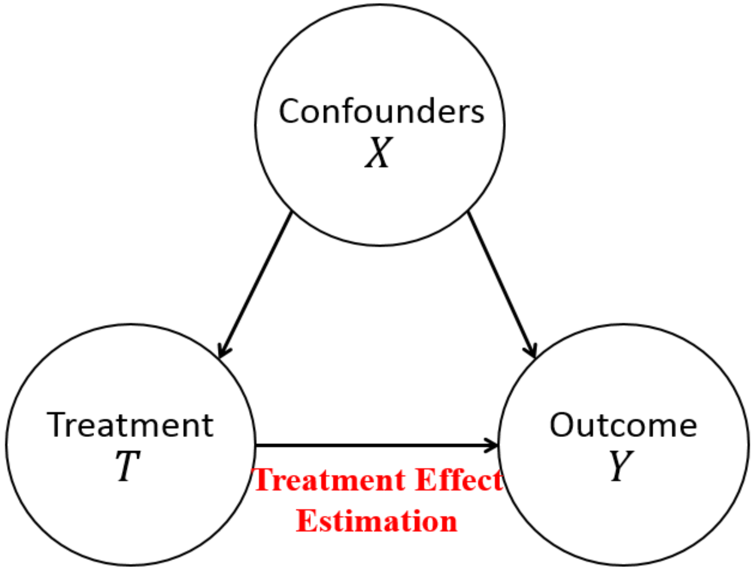
Observational Studies!

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 next Section.



Outline

- Correlation v.s. Causality
- Causal Inference
  - Methods for Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Conclusions

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

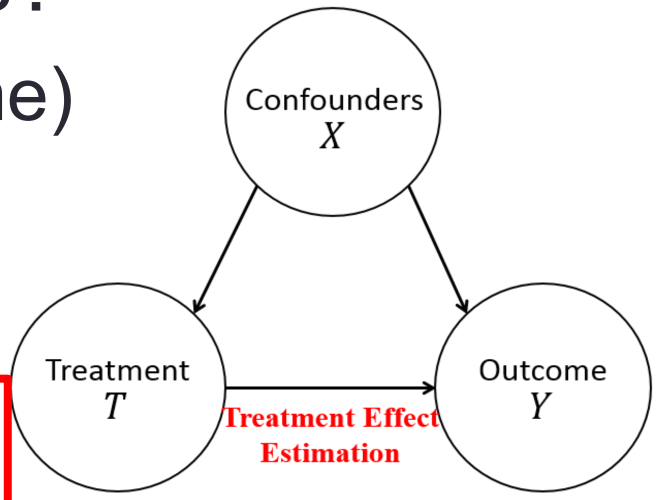
Causal Inference with Observational Data

- 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:

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

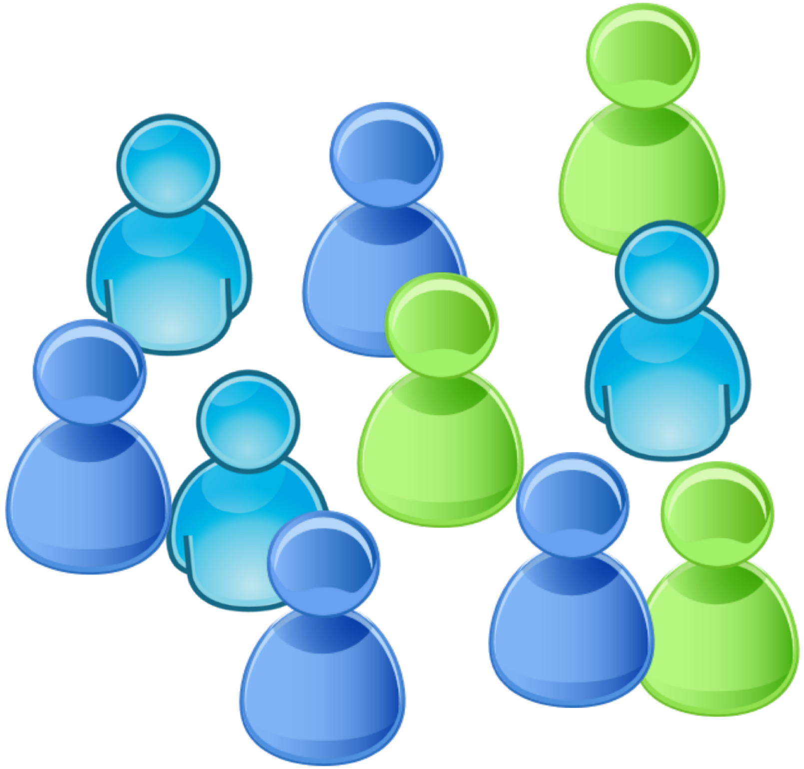
- **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

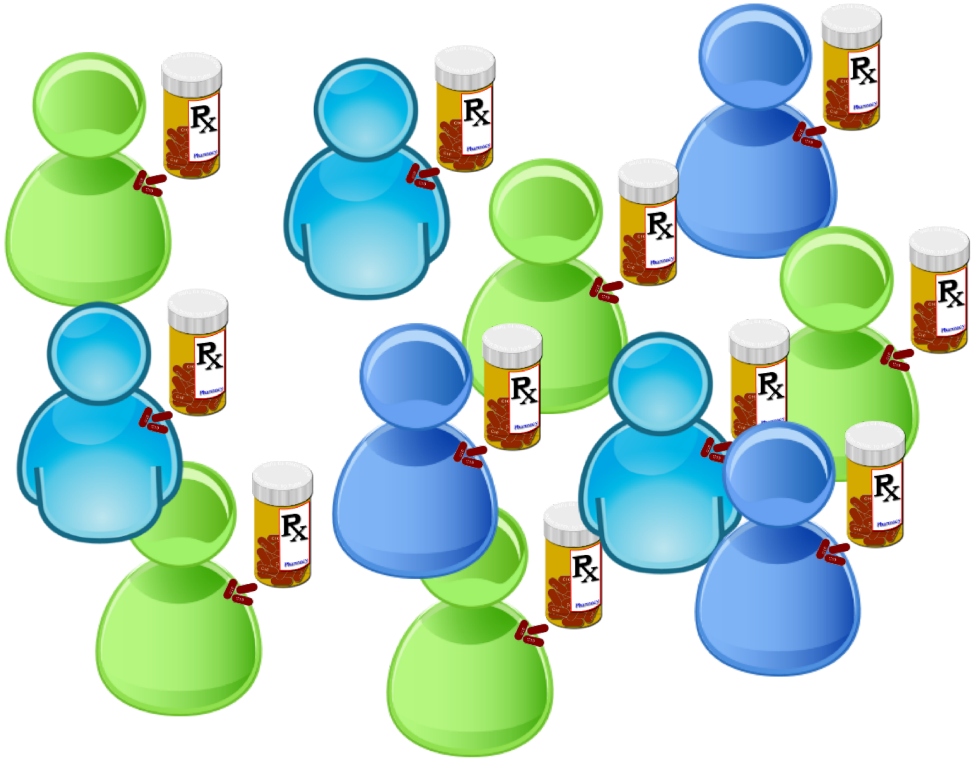
$$0 < P(T = 1 | X = x) < 1$$

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

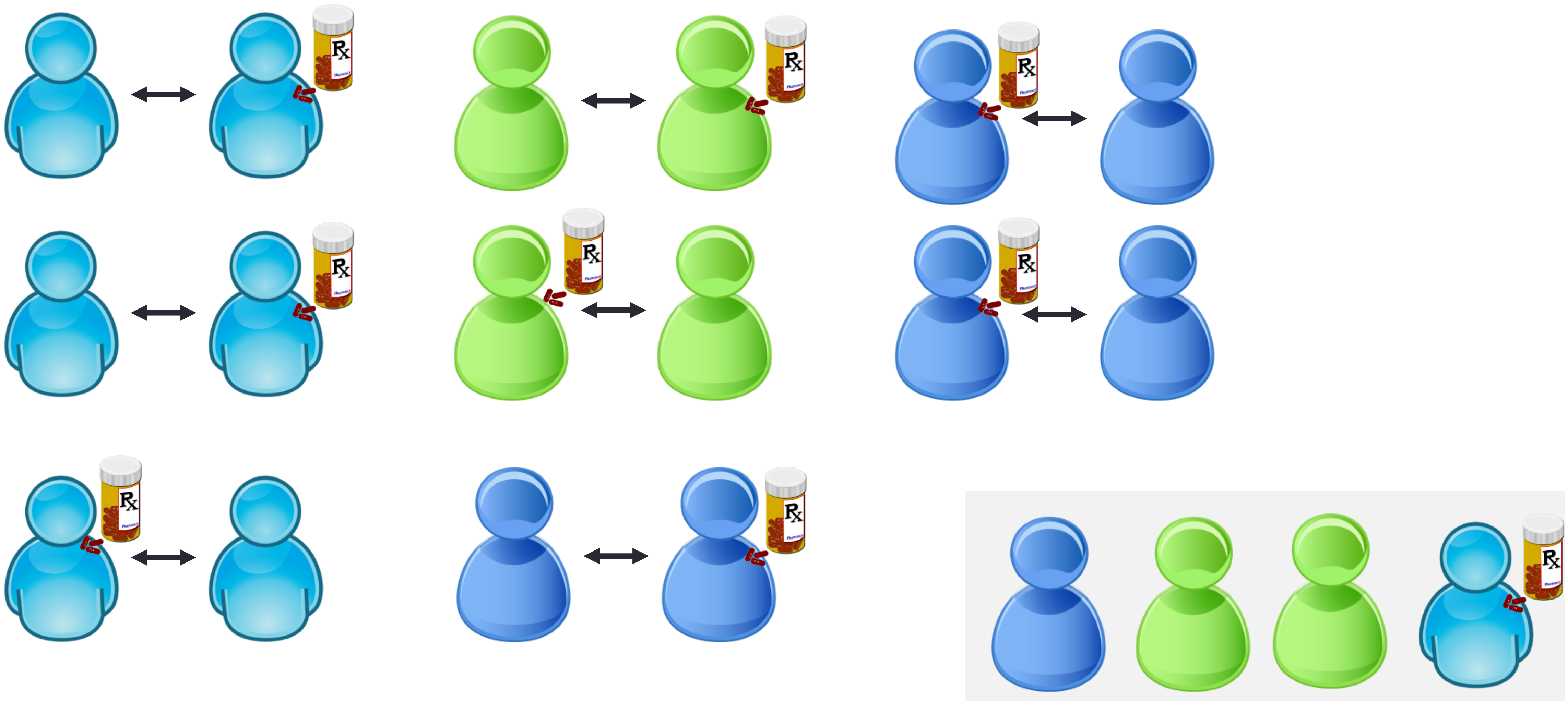
Matching





$T = 0$ $T = 1$

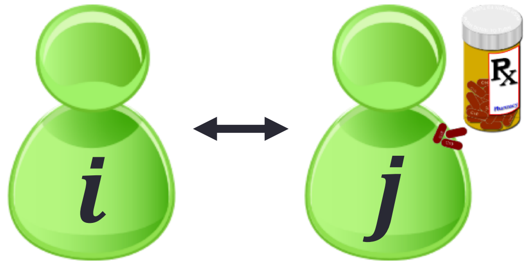
Matching



Matching

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

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



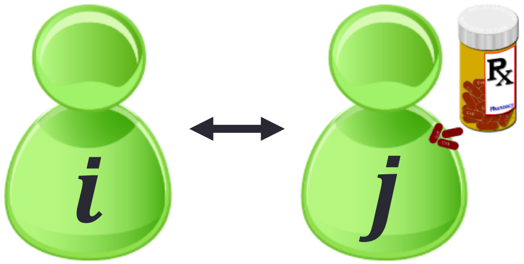
- Paired units provide the everything else (Confounders) approximate constant
- Estimating average causal effect by comparing average outcome in the paired dataset

- Smaller ϵ: less bias, but higher variance

Matching

- Exactly Matching:

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



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

- Easy to implement, but limited to low-dimensional settings
- Since 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

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

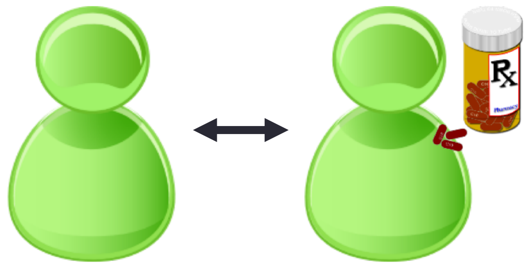
- Then, Rubin shows that the propensity score is sufficient to control or summarize the information of confounders

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

- 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



- Matching pairs by distance between propensity score:

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

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

- 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*

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

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

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

(1) $Y = T * Y_1 + (1 - T) * Y_0$

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

(3) $e(X) = E(T|X)$

- 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:

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

  - 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

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

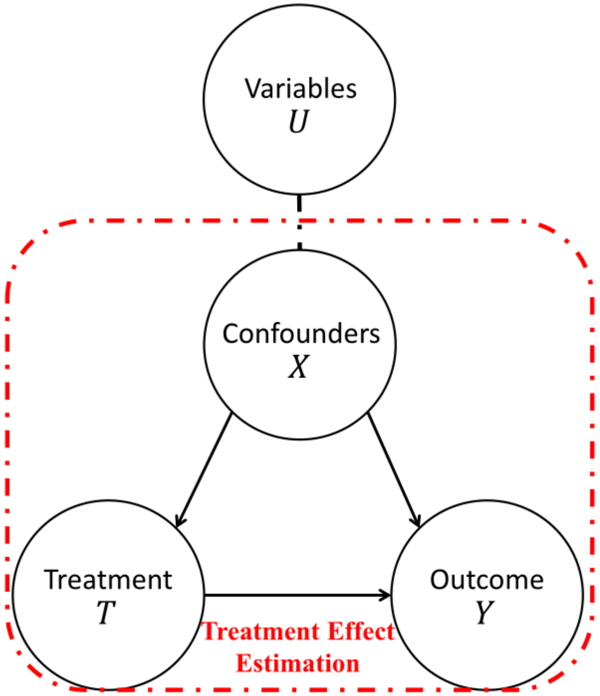
- Estimating ATE with Doubly Robust estimator:

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

  - *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

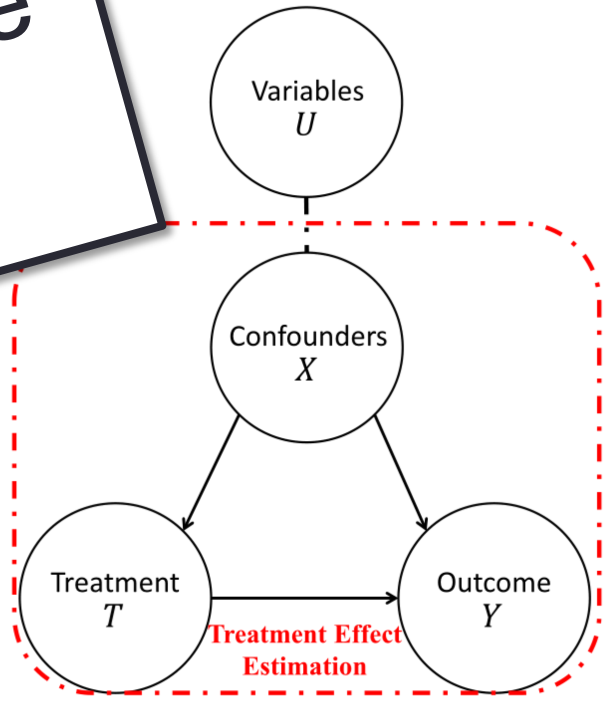


(a) Previous Causal Framework.

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?

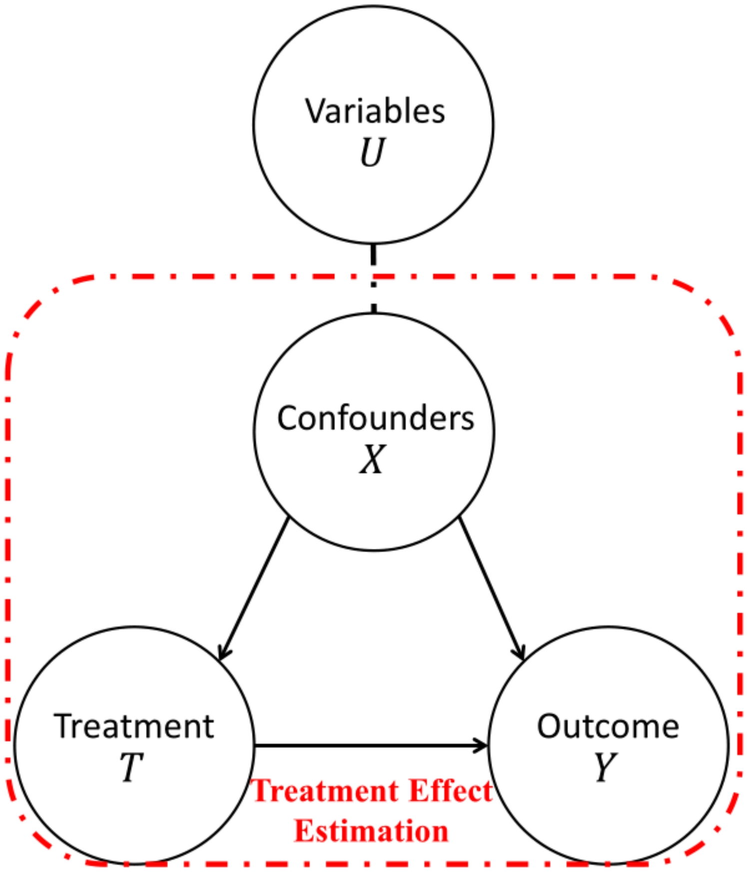


(a) Previous Causal Framework.

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:

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

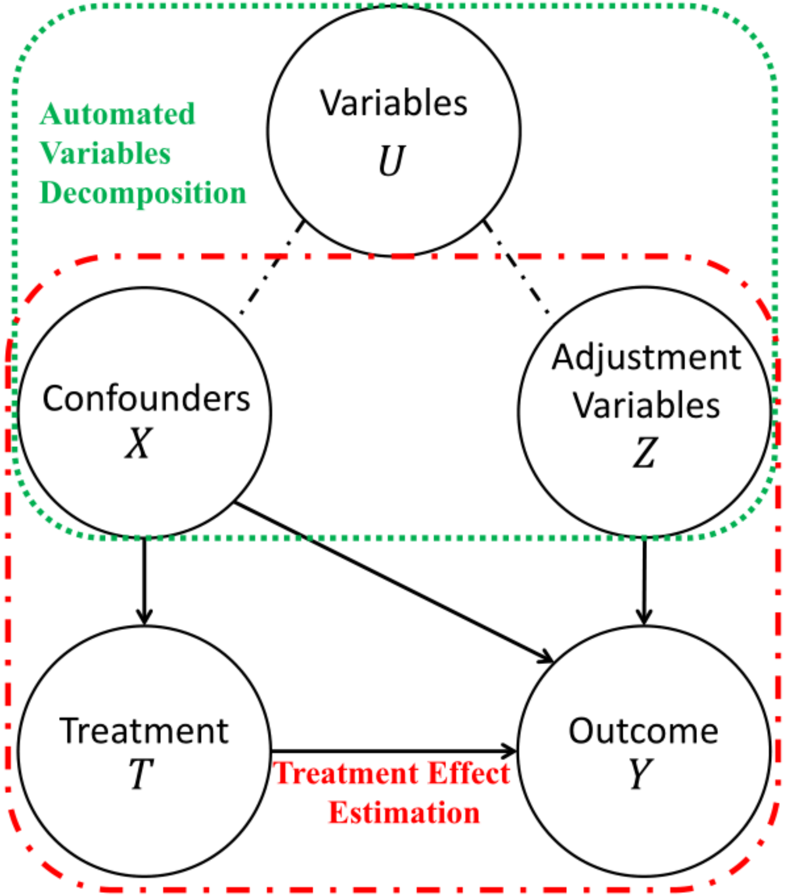
- Adjusted Outcome:

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

- 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:

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

Data-Driven Variable Decomposition (D²VD)

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

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

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

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

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

$$\|\alpha\|_1 \leq \lambda, \ \|\beta\|_1 \leq \delta, \ \|\gamma\|_1 \leq \eta, \ \|\alpha \odot \beta\|_2^2 = 0.$$

α, β, γ

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

Data-Driven Variable Decomposition (D^2VD)

Bias Analysis:

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

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

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

Variance Analysis:

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

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

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

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 (D^{2}VD)

- 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

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

  - 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)

- Dataset generation:

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

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

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

- 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: $Bias = |\widehat{ATE} - ATE|$

T/m	n / Estimator	n = 50 Bias	SD	MAE	RMSE	n = 100 Bias	SD	MAE	RMSE	n = 200 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 $ATE|$

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.

	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 ⟵ Y = 1, if LIKE; Y = 0, if DISLIKE
  - Treatment T: one feature
  - Observed Variables U: other features

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

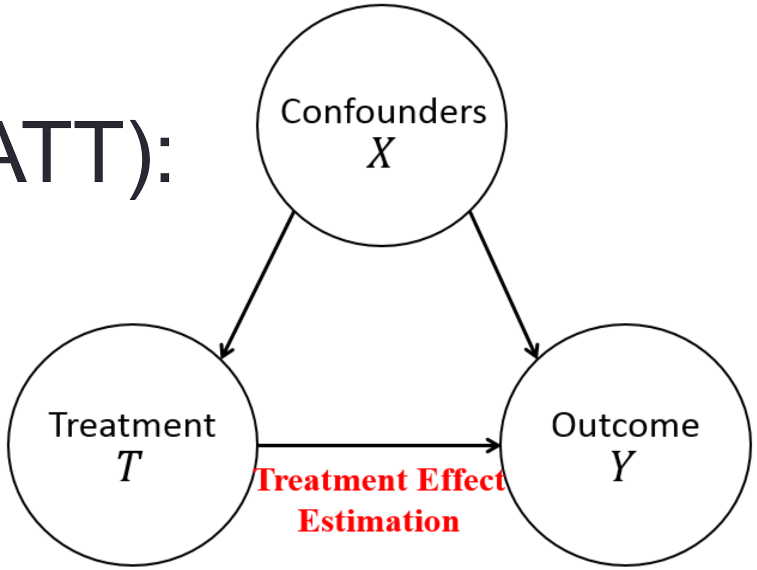
- Average Treatment Effect (ATE):

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

- 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):

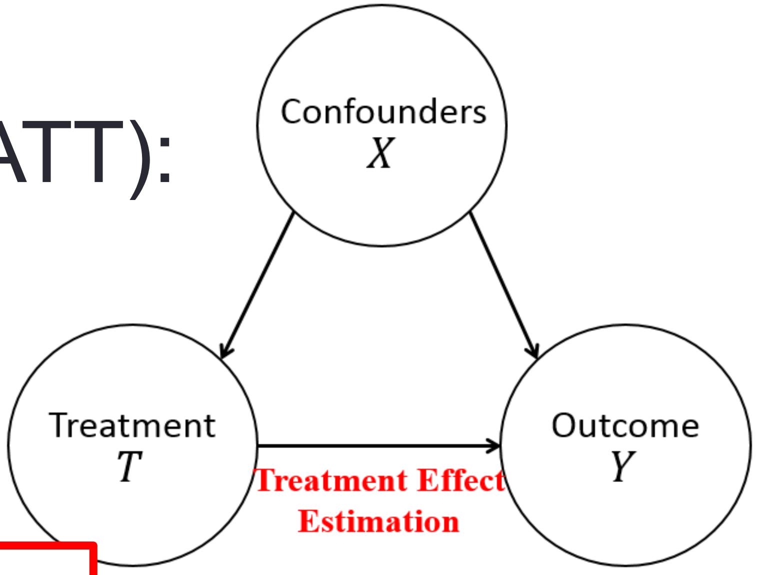
$$ATE = E[Y(T=1) - Y(T=0)]$$

- Average Treatment effect on the Treated (ATT):

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

- Two key points:

Balancing Confounders' Distribution



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

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

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

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

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

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

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

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)

Entropy Balancing

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

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

Approximate Residual Balancing

- 1. compute approximate balancing weights W as

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

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

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

- 3. Estimate the ATT as

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

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

Directly Confounder Balancing

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

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

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?

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

- But, treat ... 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 β*** and ***sample weighs W***.

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

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

How to learn the confounder weights?

Confounder Weights Learning

- General relationship among X, T, and Y:

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

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

Confounder weights

Confounding bias

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

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.

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

Differentiated Confounder Balancing

- Objective Function

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

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

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.
```

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

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

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:

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

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}

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

    - 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}

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

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!

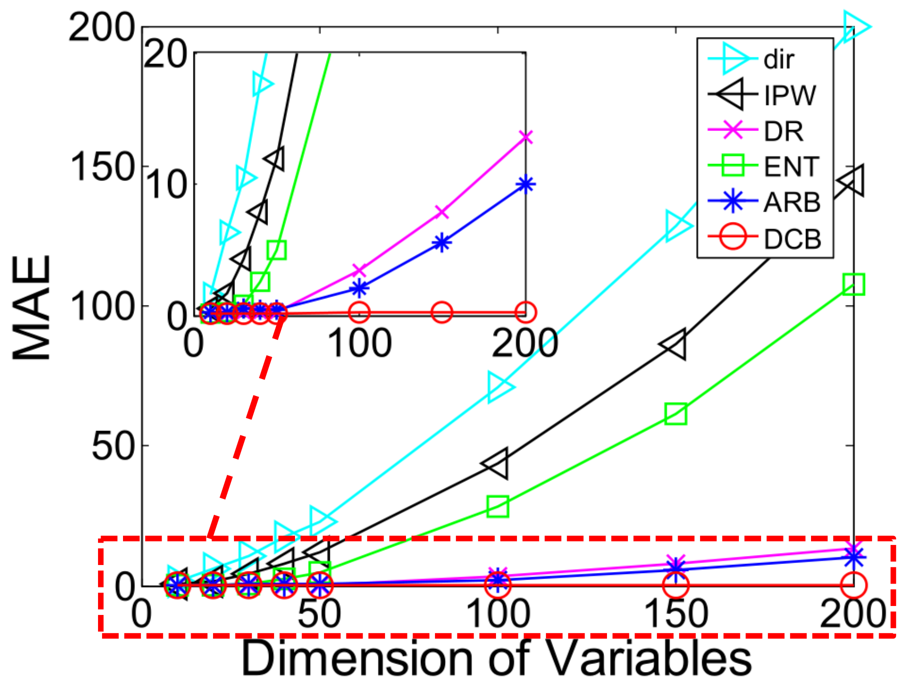
	n/p	$n=2000, p=50$			$n=2000, p=100$		
r_c	Estimator	$Bias$ (SD)	MAE	RMSE	$Bias$ (SD)	MAE	RMSE
$r_c = 0.8$	$\widehat{ATT}_{dir}$	51.06 (3.725)	51.06	51.19	143.0 (9.389)	143.0	143.3
	$\widehat{ATT}_{IPW}$	29.99 (4.048)	29.99	30.26	98.24 (8.462)	98.24	98.60
	$\widehat{ATT}_{DR}$	0.345 (0.253)	0.367	0.428	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

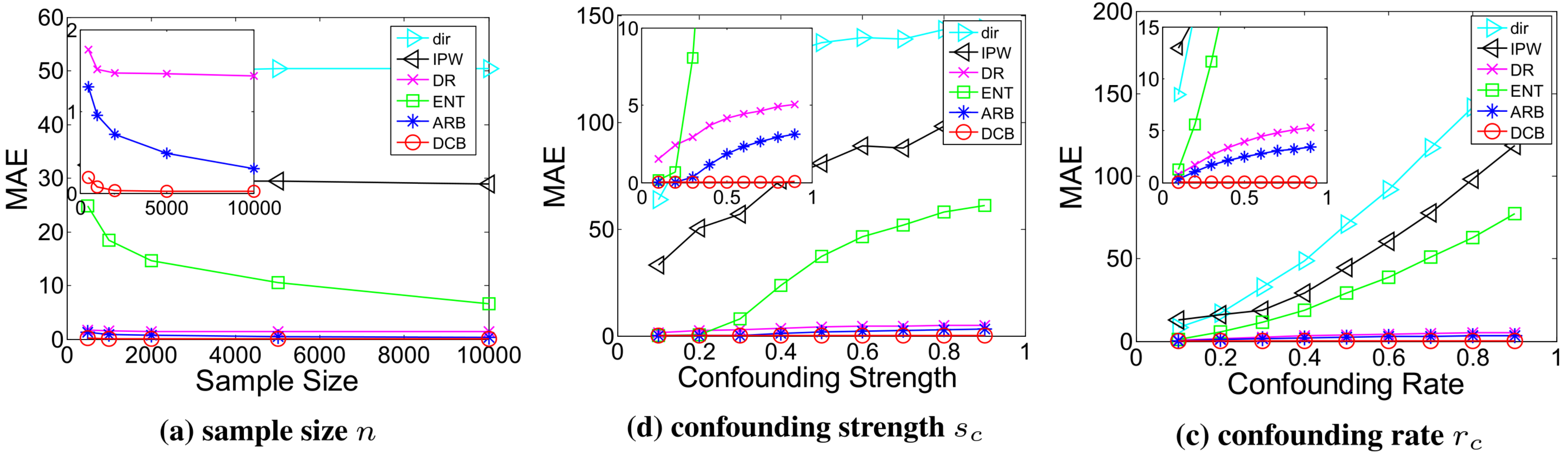
- Sample Size
- Dimension of variables
- Confounding rate
- Confounding strength



(b) dimension of variables p

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

Experiments - Robustness Test



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

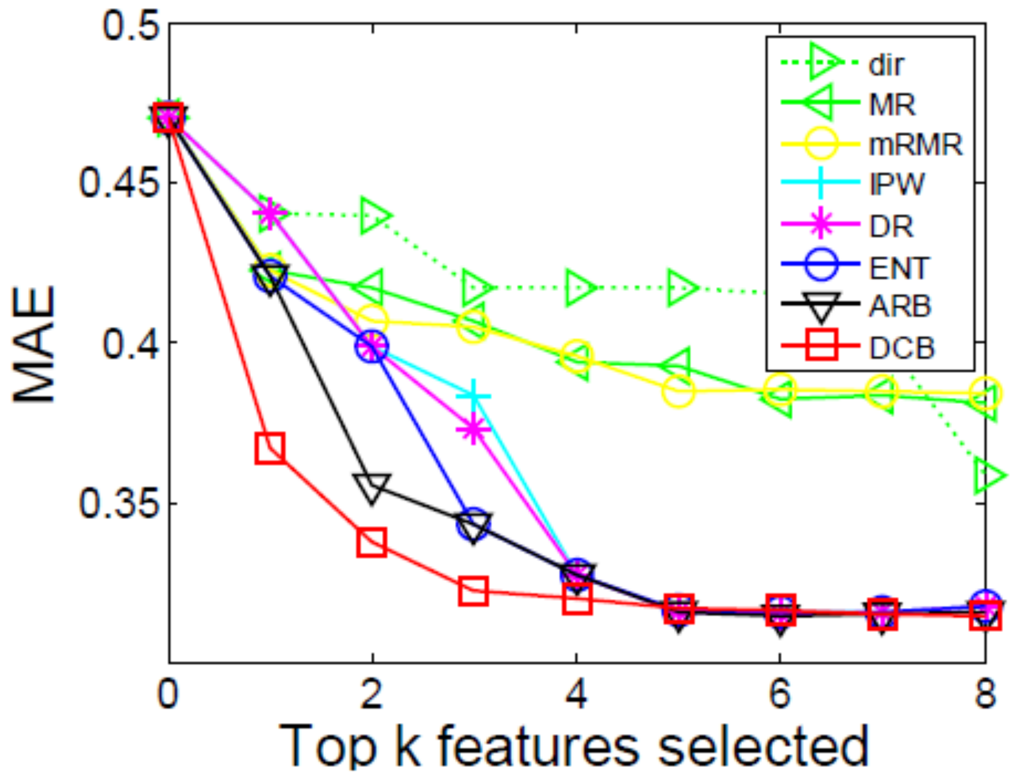
2015 WeChat

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

Sectional 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

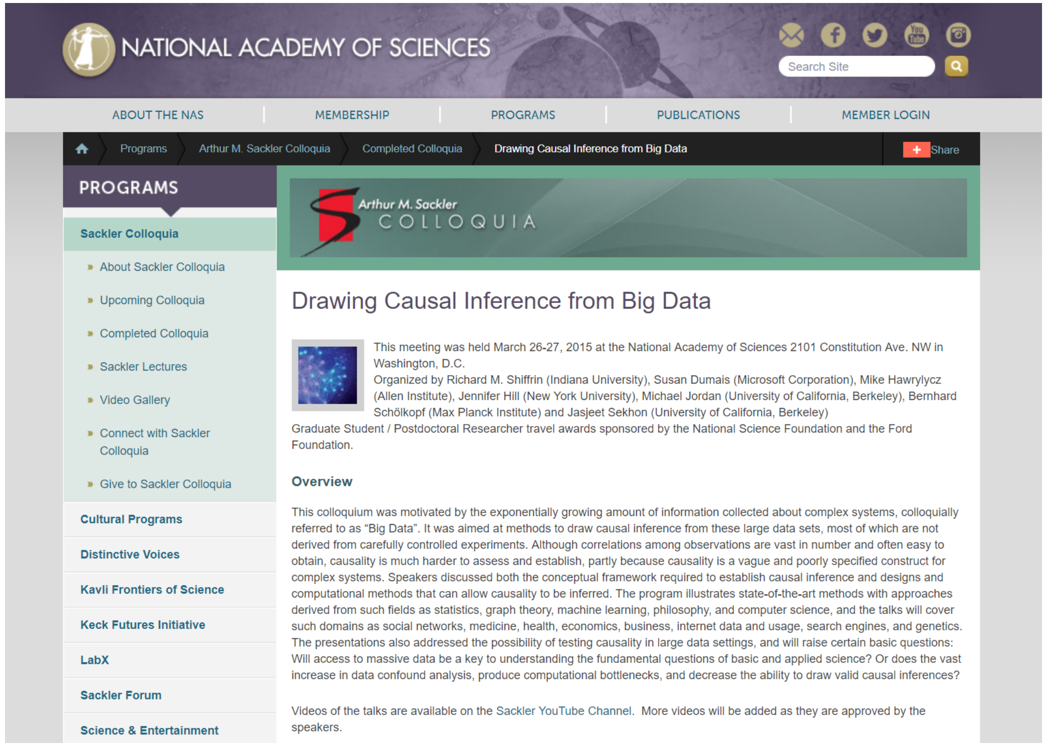
Balance all confounder equally

Different confounders make different bias

Sectional Summary: Methods for Causal Inference

- Progress has been made to draw causality from big data.
- From single to group
- From binary to continuous
- Weak assumptions

Ready for Learning?



Outline

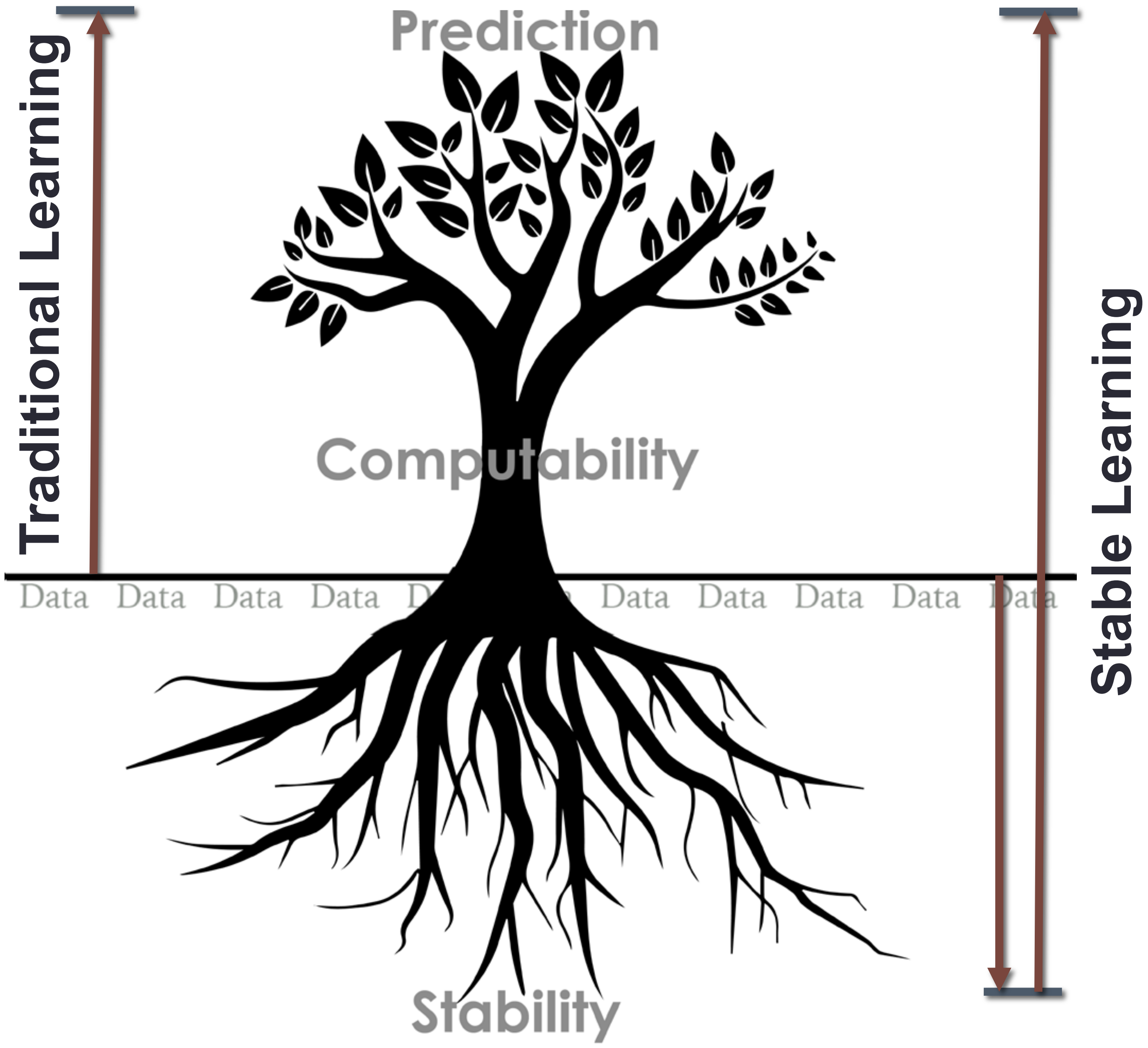
- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Future Directions and Conclusions

Stability and Prediction

Prediction Performance

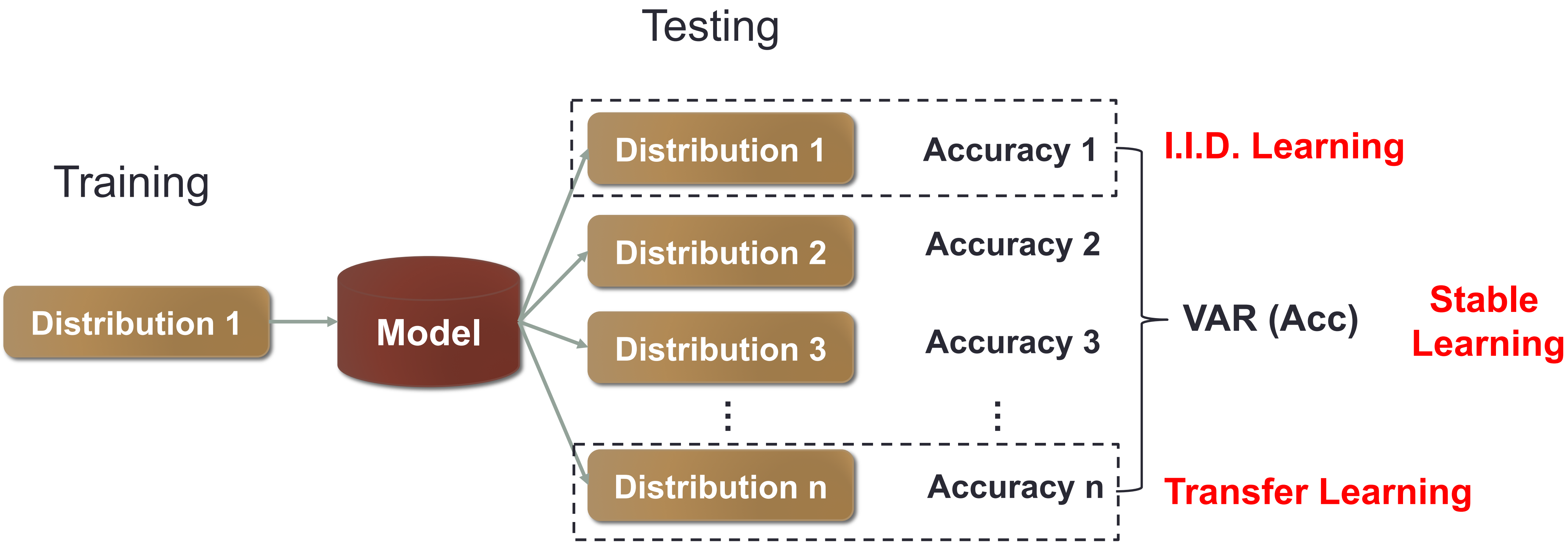
Learning Process

True Model



Bin Yu (2016), Three Principles of Data Science: predictability, computability, stability

Stable Learning

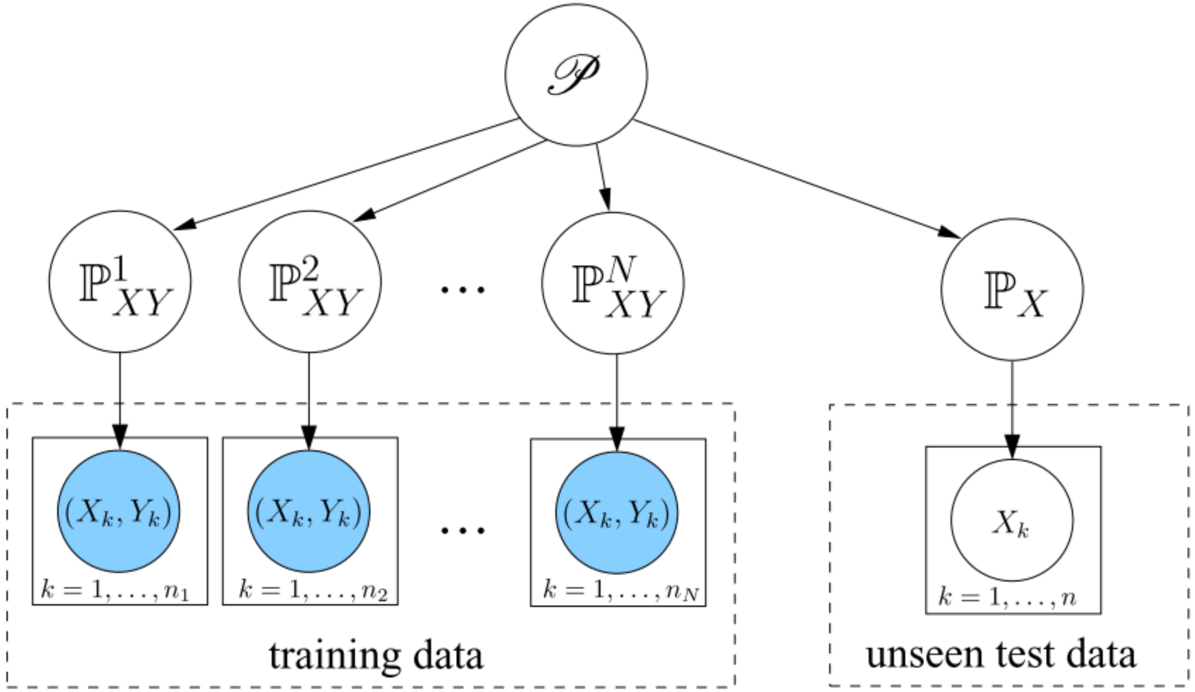


Stability and Robustness

- Robustness
  - More on prediction performance over data perturbations
  - ***Prediction*** performance-driven
- Stability
  - More on the true model
  - Lay more emphasis on ***Bias***
  - Sufficient for robustness

Stable learning is a (intrinsic?) way to realize robust prediction

Domain Generalization / Invariant Learning



- Given data from different observed environments $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

- The task is to predict Y given X such that the prediction works well (is “robust”) for “all possible” (including unseen) environments

Domain Generalization

- **Assumption**: the conditional probability P(Y|X) is stable or invariant across different environments.
- **Idea**: taking knowledge acquired from a number of related domains and applying it to previously unseen domains
- **Theorem**: Under reasonable technical assumptions. Then with probability at least $1-\delta$

$$\sup_{\|f\|_{\mathcal{H}}\leq 1}\left|\mathbb{E}^{*}_{\mathscr{P}}\mathbb{E}_{\mathbb{P}}\ell(f(\tilde{X}_{ij}),Y_i)-\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}),Y_i)\right|^2$$

$$\leq c_1\cdot\underbrace{\mathbb{V}_{\mathcal{H}}(\mathbb{P}^1,\mathbb{P}^2,\ldots,\mathbb{P}^N)}_{\textit{distributional variance}}+\underbrace{c_2\frac{N\cdot(\log\delta^{-1}+2\log N)}{n}+c_3\frac{\log\delta^{-1}}{N}+\frac{c_4}{N}}_{\textit{vanish as } N,n\to\infty}$$

Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature. ICML 2013.

Invariant Prediction

- **Invariant Assumption:** There exists a subset $S \in X$ is causal for the prediction of Y, and the conditional distribution P(Y|S) is stable across all environments.

for all $e \in \mathcal{E}$, X^e has an arbitrary distribution and

$$Y^e = g(X^e_{S^*}, \varepsilon^e), \qquad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}$$

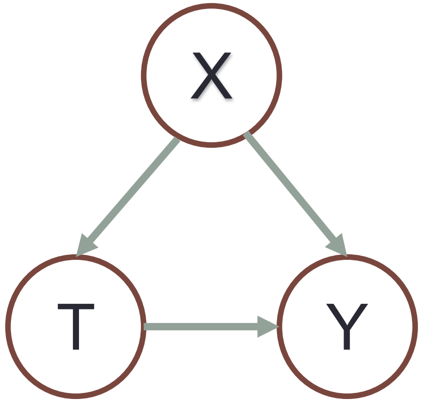
- **Idea: Linking to causality**
  - Structural Causal Model (Pearl 2009):

$$Y^e \leftarrow \sum_{k \in \mathrm{pa}(Y)} \underbrace{\beta_{Y,k}}_{\forall e} X^e_k + \underbrace{\varepsilon^e_Y}_{\sim F_\varepsilon \forall e \in \mathcal{G}}$$

  - The parent variables of Y in SCM satisfies Invariant Assumption
  - The causal variables lead to invariance w.r.t. “all” possible environments

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016*

From *Variable Selection* to *Sample Reweighting*



Typical Causal Framework

Directly Confounder Balancing

Given a feature T

↓

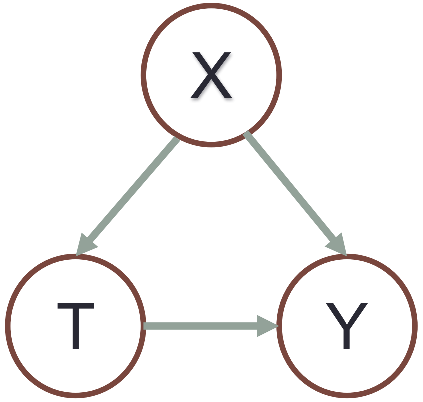
Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

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

Sample reweighting can make a variable independent of other variables.

Global Balancing: Decorrelating Variables



Typical Causal Framework

Global Balancing

Given **ANY** 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)

Partial effect can be regarded as causal effect. Predicting with causal variables is stable across different environments.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Theoretical Guarantee

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^*.

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

↓

0

PROOF. Since $\|\cdot\| \geq 0$, Eq. (8) can be simplified to $\forall j, \forall k \neq j$

$$\lim_{n\to\infty} \left(\frac{\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i}{\sum_{i:\mathbf{X}_{i,j}=1} W_i} - \frac{\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=0} W_i}{\sum_{i:\mathbf{X}_{i,j}=0} W_i} \right) = 0$$

with probability 1. For W^*, from Lemma 3.1, $0 < P(\mathbf{X}_i = x) < 1$, $\forall x, \forall i, t = 1$ or 0,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,j}=t} W_i^* &= \lim_{n\to\infty} \frac{1}{n} \sum_{x:x_j=t} \sum_{i:\mathbf{X}_i=x} W_i^* \\ &= \lim_{n\to\infty} \sum_{x:x_j=t} \frac{1}{n} \sum_{i:\mathbf{X}_i=x} \frac{1}{P(\mathbf{X}_i=x)} \\ &= \lim_{n\to\infty} \sum_{x:x_j=t} P(\mathbf{X}_i = x) \cdot \frac{1}{P(\mathbf{X}_i=x)} = 2^{p-1} \end{aligned}$$

with probability 1 (Law of Large Number). Since features are binary,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i^* = 2^{p-2}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,j}=0} W_i^* = 2^{p-1}, \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=0} W_i^* = 2^{p-2}$$

and therefore, we have following equation with probability 1:

$$\lim_{n\to\infty} \left(\frac{\mathbf{X}_{\cdot,k}^T (W^* \odot \mathbf{X}_{\cdot,j})}{W^{*T} \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^T (W^* \odot (1 - \mathbf{X}_{\cdot,j}))}{W^{*T} (1 - \mathbf{X}_{\cdot,j})} \right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

□

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Causal Regularizer

Set feature *j* as treatment variable

$$\sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2,$$

All features excluding treatment *j*

Sample Weights

Indicator of treatment status

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM****, 2018*.

Causally Regularized Logistic Regression

$$\min \quad \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (x_i \beta))),$$

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

$$W \ge 0, \ \ \|W\|_2^2 \le \lambda_2, \ \ \|\beta\|_2^2 \le \lambda_3, \ \ \|\beta\|_1 \le \lambda_4,$$

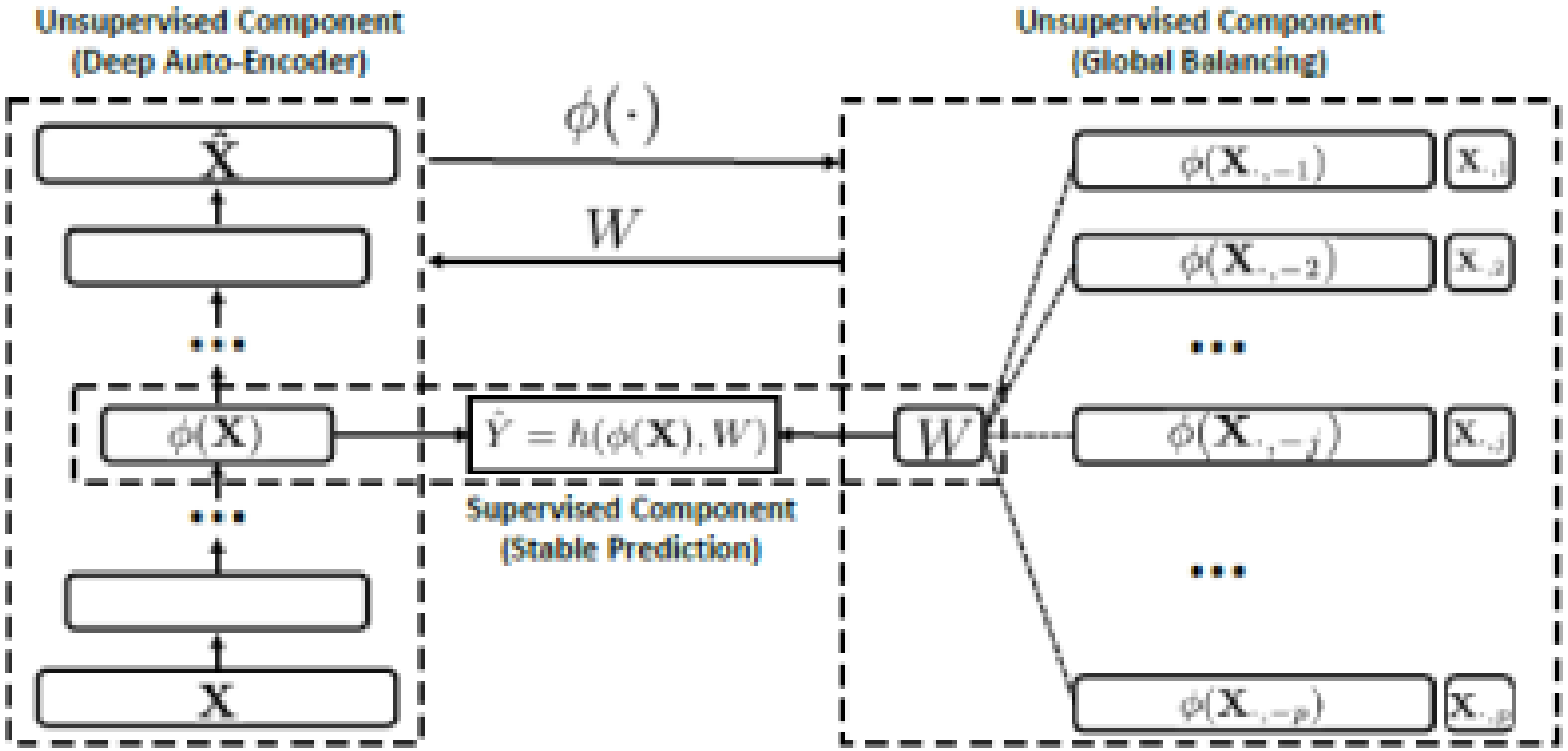
$$(\sum_{k=1}^{n} W_k - 1)^2 \le \lambda_5,$$

Sample reweighted logistic loss

Causal Contribution

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM***, *2018*.

From Shallow to Deep - DGBR

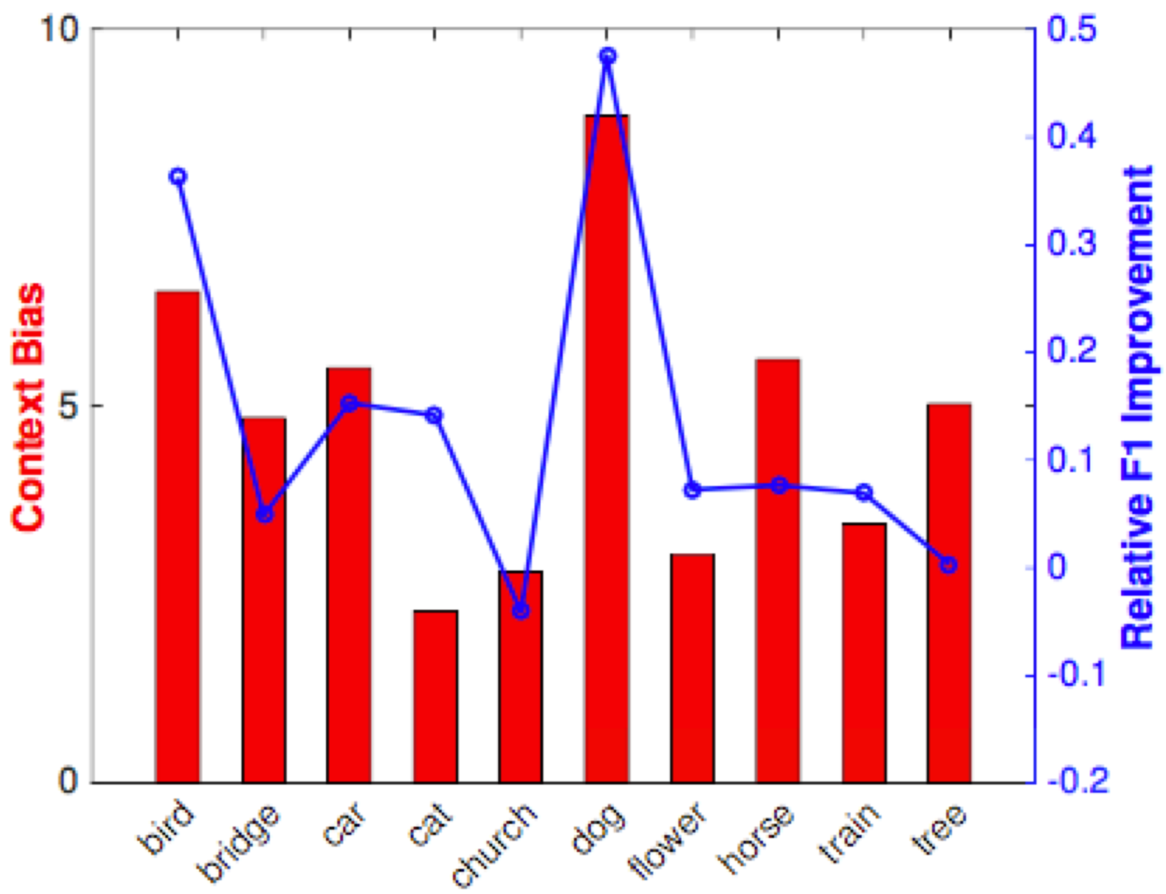


Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

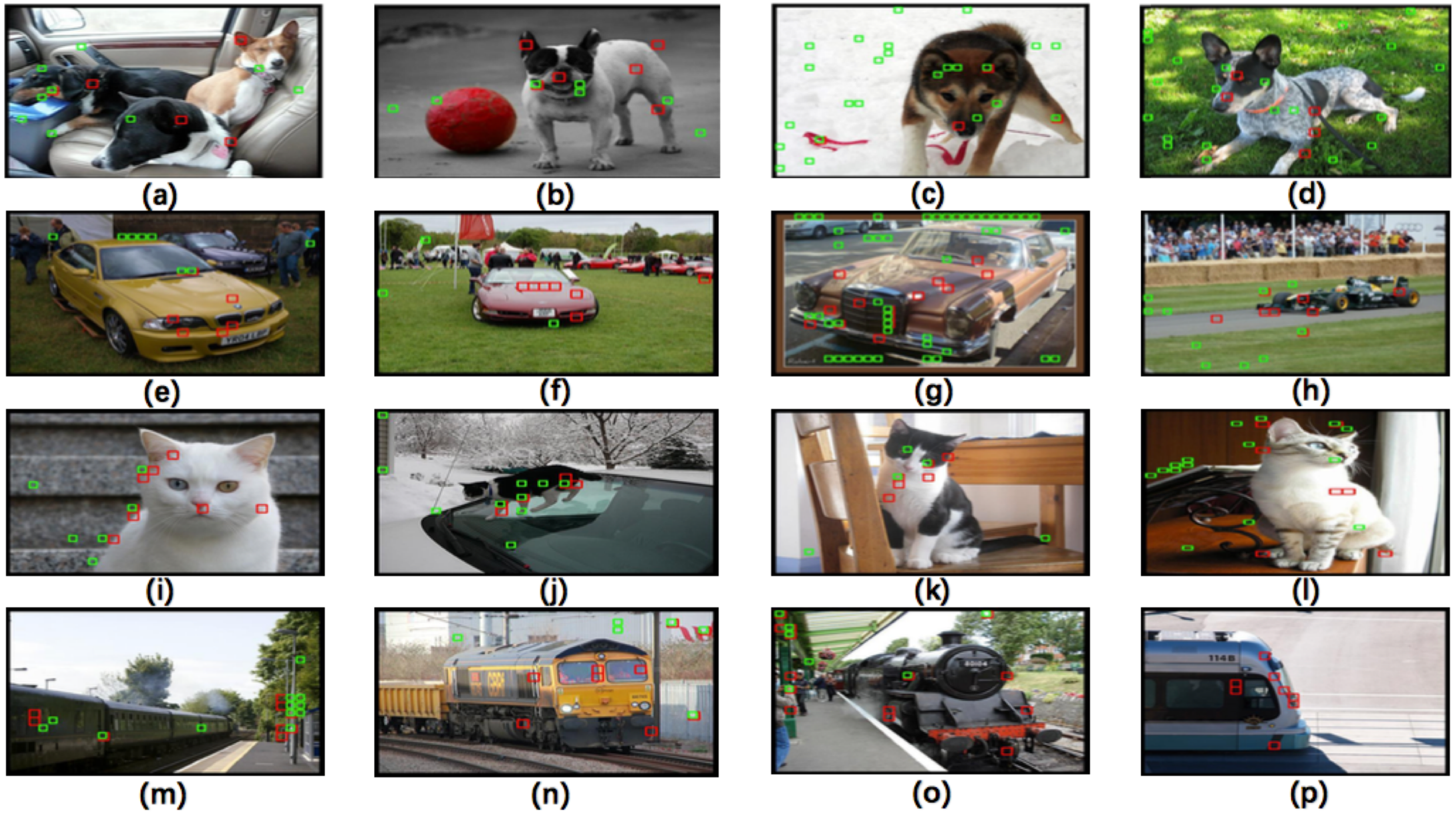
Experiment 1 – non-i.i.d. image classification

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

Experimental Result - insights

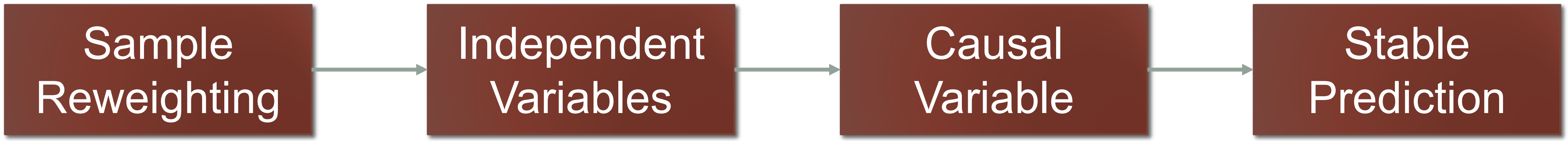


Experimental Result - insights

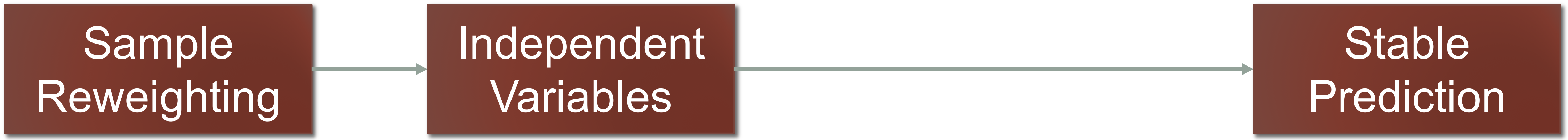


From ***Causal*** problem to ***Learning*** problem

- Previous logic:

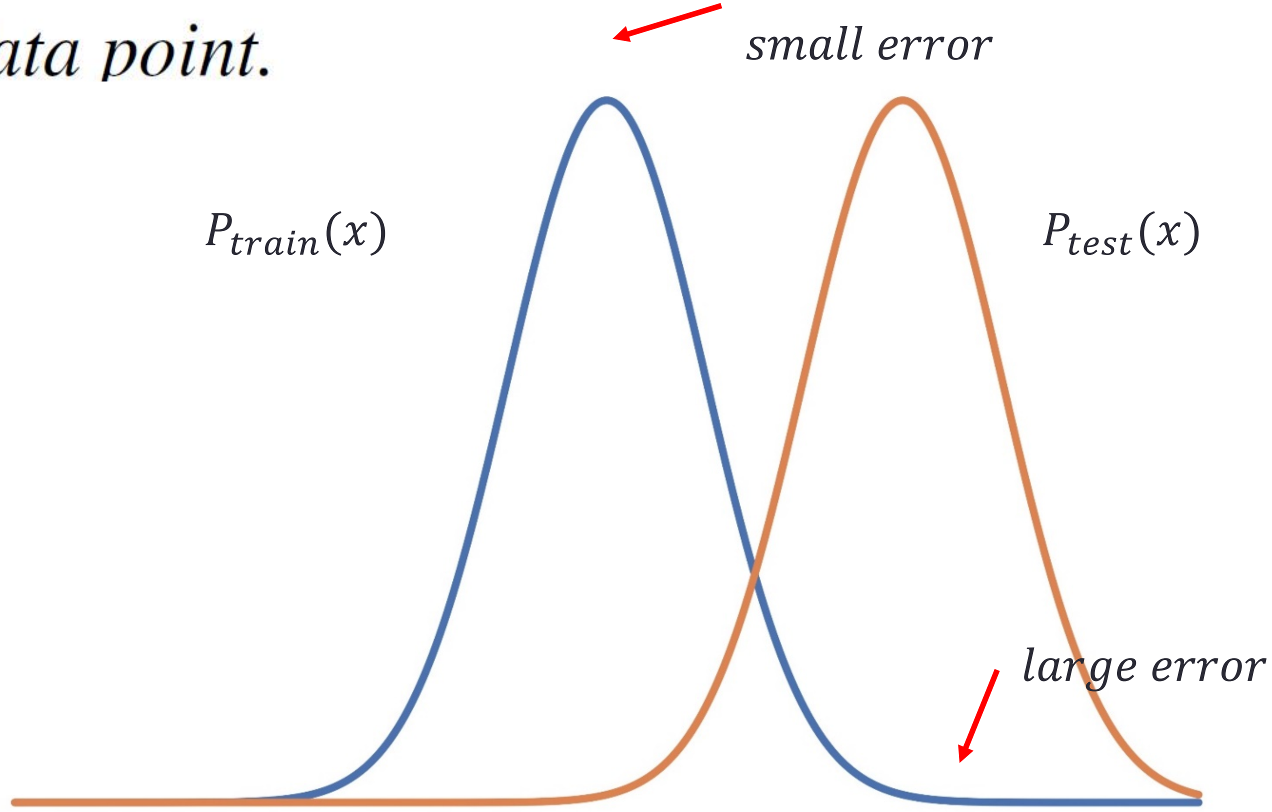


- More direct logic:



Thinking from the ***Learning*** end

Problem 1. ***(Stable Learning)*** *: Given the target y and p input variables $x = [x_1, \ldots, x_p] \in \mathbb{R}^p$, the task is to learn a predictive model which can achieve* ***uniformly*** *small error on* ***any*** *data point.*



Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Stable Learning of Linear Models

- Consider the linear regression with misspecification bias

$$y = x^\top \overline{\beta}_{1:p} + \overline{\beta}_0 + b(x) + \epsilon$$

Goes to infinity when perfect collinearity exists!

Bias term with bound $b(x) \leq \delta$

- By accurately estimating $\overline{\beta}$ with the property that $b(x)$ is uniformly small for all x, we can achieve stable learning.
- However, the estimation error caused by misspecification term can be as bad as $\|\hat{\beta} - \overline{\beta}\|_2 \leq 2(\delta/\gamma) + \delta$, where γ^2 is the smallest eigenvalue of centered covariance matrix.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Toy Example

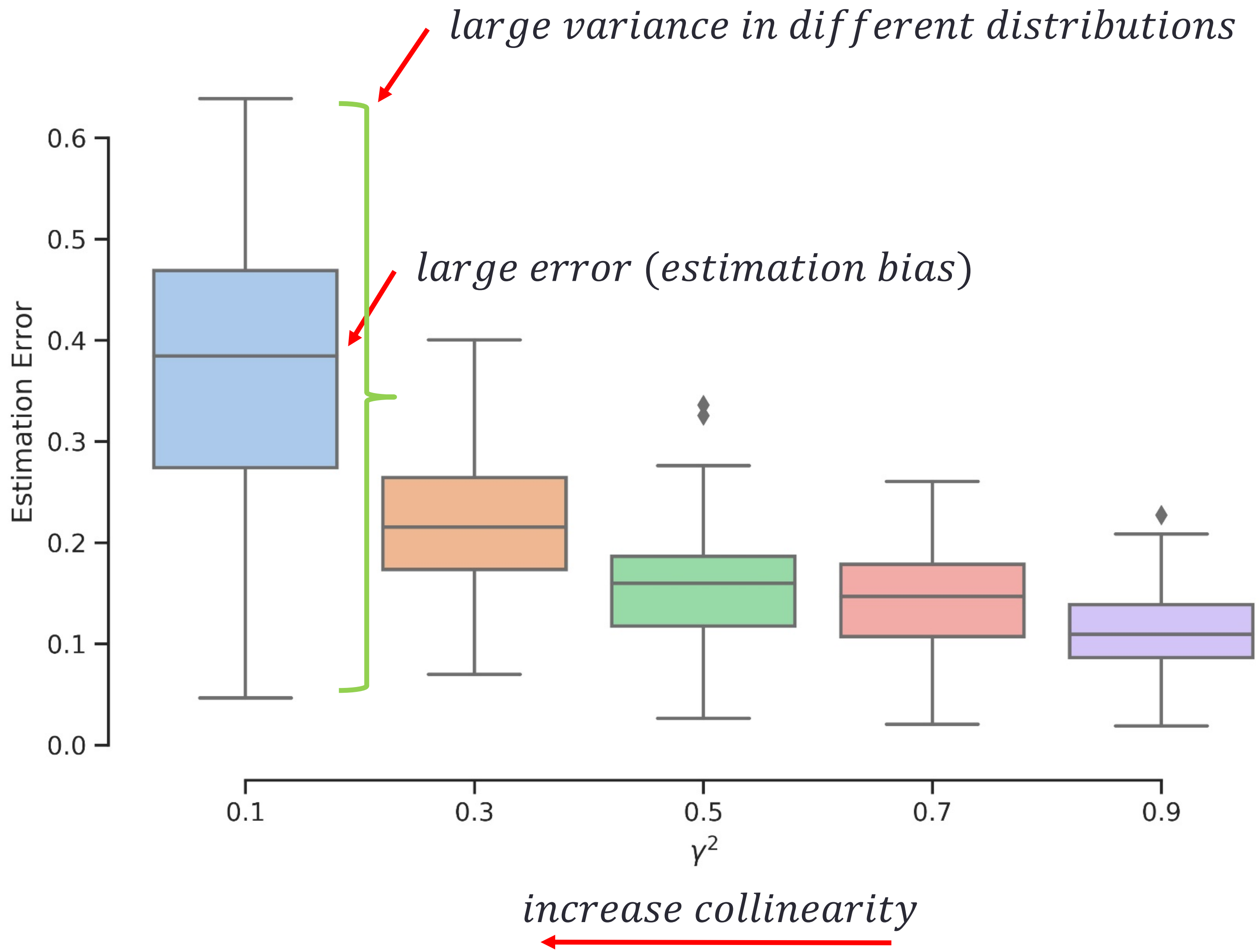
- Assume the design matrix X consists of two variables X_1, X_2, generated from a multivariate normal distribution:

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- By changing ρ, we can simulate different extent of collinearity.
- To induce bias related to collinearity, we generate bias term $b(X)$ with $b(X) = Xv$, where v is the eigenvector of centered covariance matrix corresponding to its smallest eigenvalue γ^2.
- The bias term is sensitive to collinearity.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Simulation Results



Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Reducing collinearity by sample reweighting

Idea: Learn a new set of ***sample weights*** $w(x)$ to decorrelate the input variables and increase the smallest eigenvalue

- Weighted Least Square Estimation

$$\hat{\beta} = \arg\min_{\beta} \mathbf{E}_{(x)\sim D} w(x) \left(x^\top \beta_{1:p} + \beta_0 - y\right)^2$$

which is equivalent to

$$\hat{\beta} = \arg\min_{\beta} \mathbf{E}_{(x)\sim \tilde{D}} \left(x^\top \beta_{1:p} + \beta_0 - y\right)^2$$

So, how to find an “oracle” distribution $\tilde{D}$ which holds the desired property?

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Sample Reweighted Decorrelation Operator (cont.)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \xrightarrow{\text{Decorrelation}} \tilde{\mathbf{X}} = \begin{pmatrix} x_{i1} & \cdots & x_{rl} & \cdots \\ x_{j1} & \cdots & x_{sl} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{tl} & \cdots \end{pmatrix}$$

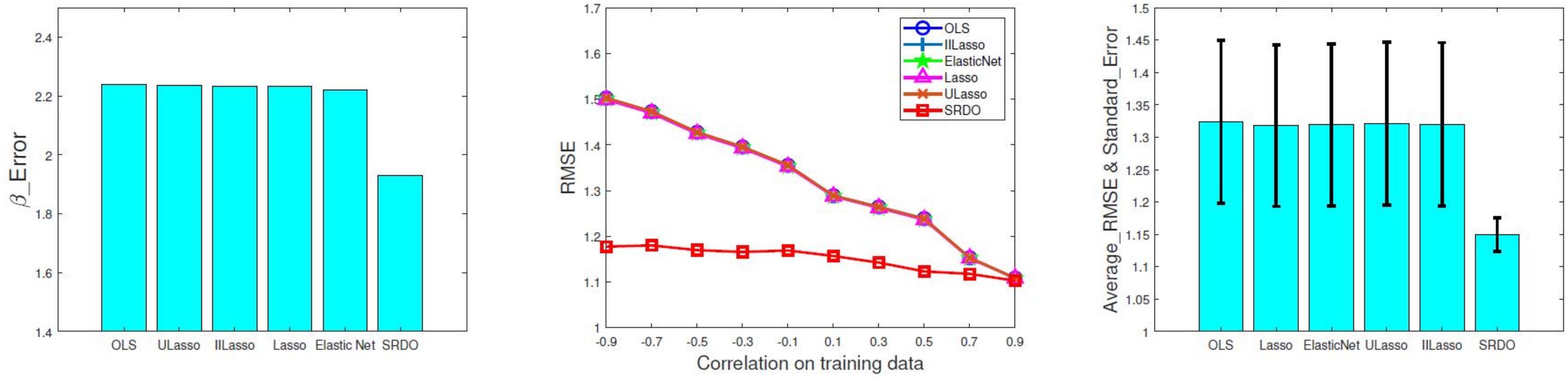
where i, j, k, r, s, t are drawn from 1 … n at random

- By treating the different columns independently while performing random resampling, we can obtain a column-decorrelated design matrix with the same marginal as before.
- Then we can use density ratio estimation to get $w(x)$.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Experimental Results

- Simulation Study

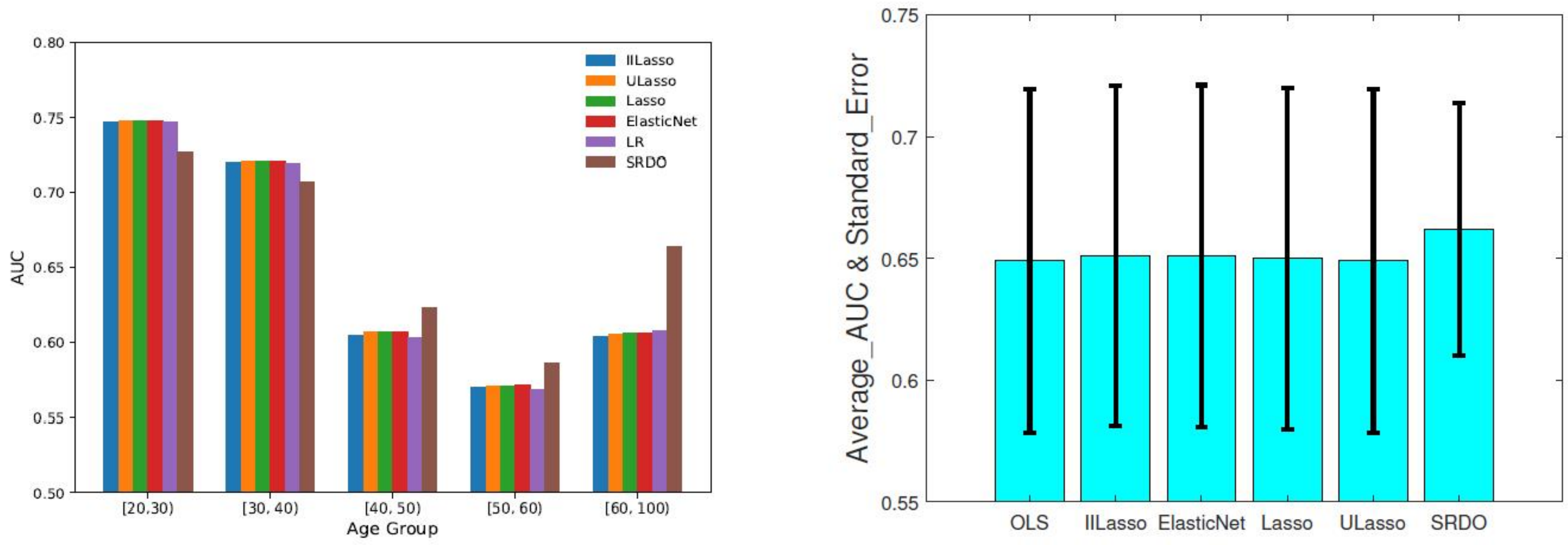


(a) Estimation error (b) Prediction error over different test environments (c) Average prediction error&stability

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Experimental Results

- Regression
- Classification



(a) AUC over different test environments. (b) Average AUC of all the environments and stability.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Disentanglement Representation Learning

From decorrelating input variables to learning disentangled representation

- Learning Multiple Levels of Abstraction
  - The big payoff of deep learning is to allow learning higher levels of abstraction
  - Higher-level abstractions disentangle the factor of variation, which allows much easier generalization and transfer

Yoshua Bengio, From Deep Learning of Disentangled Representations to Higher-level Cognition. (2019). YouTube. Retrieved 22 February 2019.

Disentanglement for Causality

- Causal / mechanism independence
  - Independently Controllable Factors *(Thomas, Bengio et al., 2017)*

selectively change

correspond to value

A policy π_k

A representation f_k

$$sel(s,a,k) = \mathbb{E}_{s' \sim \mathcal{P}^a_{ss'}} \left[\frac{|f_k(s') - f_k(s)|}{\sum_{k'} |f_{k'}(s') - f_{k'}(s)|} \right]$$

  - Optimize both π_k and f_k to minimize

$$\underbrace{\mathbb{E}_s[\tfrac{1}{2}||s - g(f(s))||_2^2]}_{\mathcal{L}_{ae} \text{ the reconstruction error}} - \lambda \underbrace{\sum_k \mathbb{E}_s[\sum_a \pi_k(a|s) sel(s,a,k)]}_{\mathcal{L}_{sel} \text{ the disentanglement objective}}.$$

Require subtle design on the policy set to guarantee causality.

Sectional Summary

- Causal inference provide valuable insights for stable learning
- Complete causal structure means data generation process, necessarily leading to stable prediction
- Stable learning can also help to advance causal inference
- Performance driven and practical applications

Benchmark is important!

Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Future Directions and Conclusions

Non-I.I.D. Image Classification

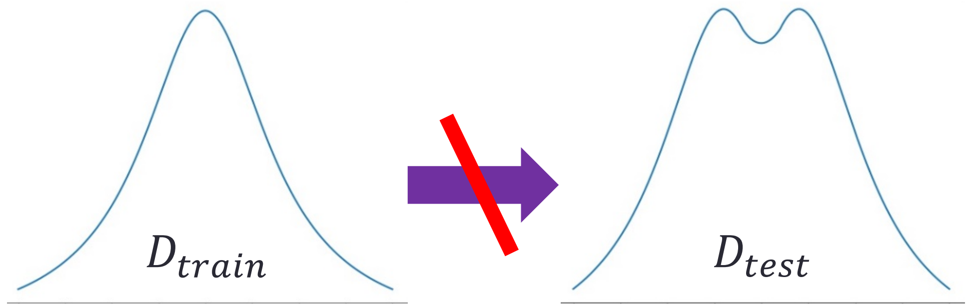
- Non I.I.D. Image Classification

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

- 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

known

unknown



Existence of Non-I.I.Dness

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

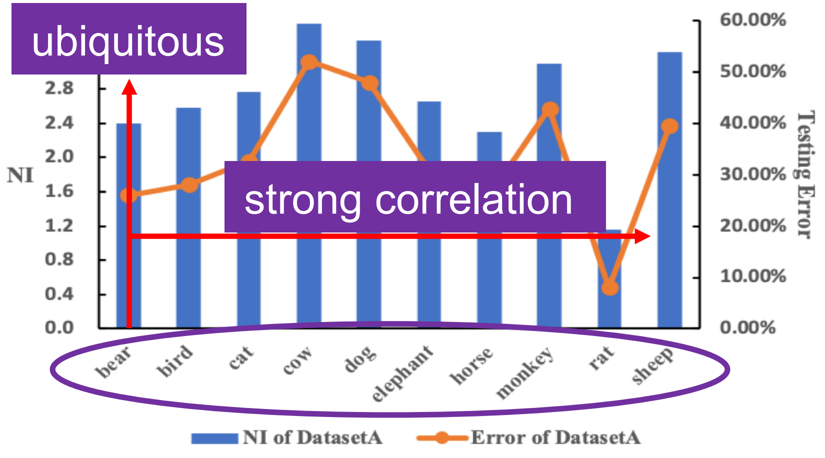
Definition 1 ***Non-I.I.D. Index (NI)*** *Given a feature extractor $g_\varphi(\cdot)$ and a class C,* ***the degree of distribution shift*** *between training data D^C_{train} and testing data D^C_{test} is defined as:*

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

Distribution shift

For normalization

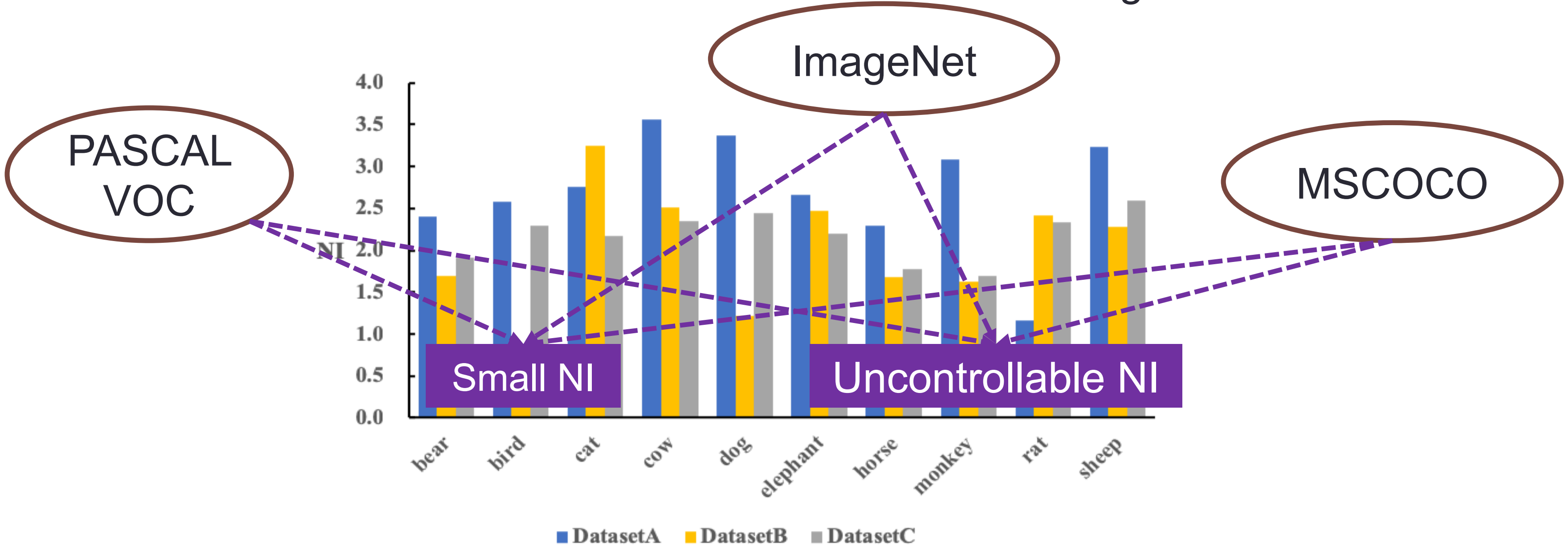
- Existence of Non-I.I.Dness on Dataset consisted of 10 subclasses from ImageNet
- For each class
  - Training data
  - Testing data
  - CNN for prediction



Related Datasets

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

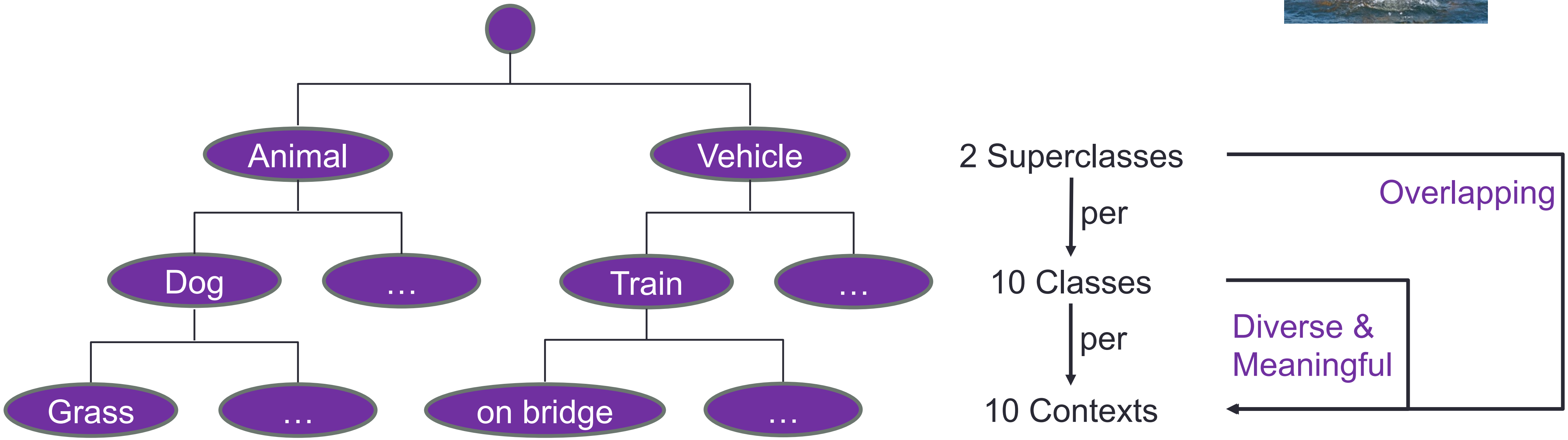
Average NI: 2.7



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 samples 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 NICO are with rich contextual information

more challenging for image classification

Our NICO data is more Non-iid, more challenging

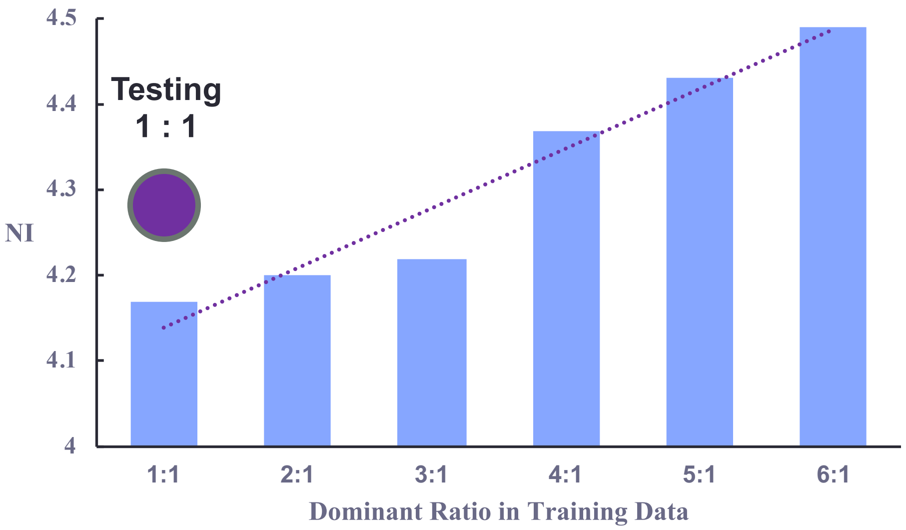
Proportional Bias

- 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.

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

Dominate Context (55%)

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

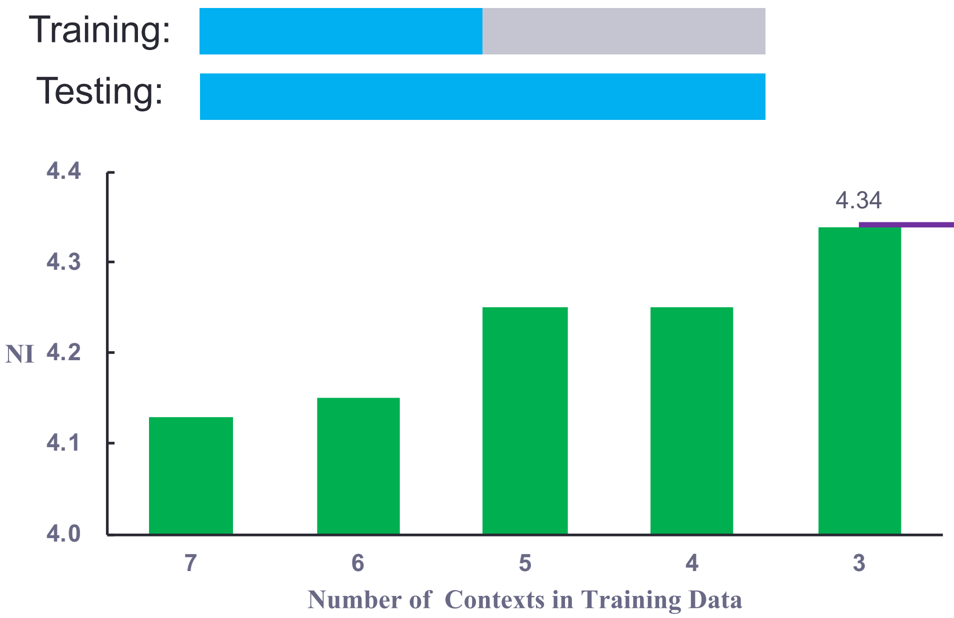


We can control NI by varying dominate ratio

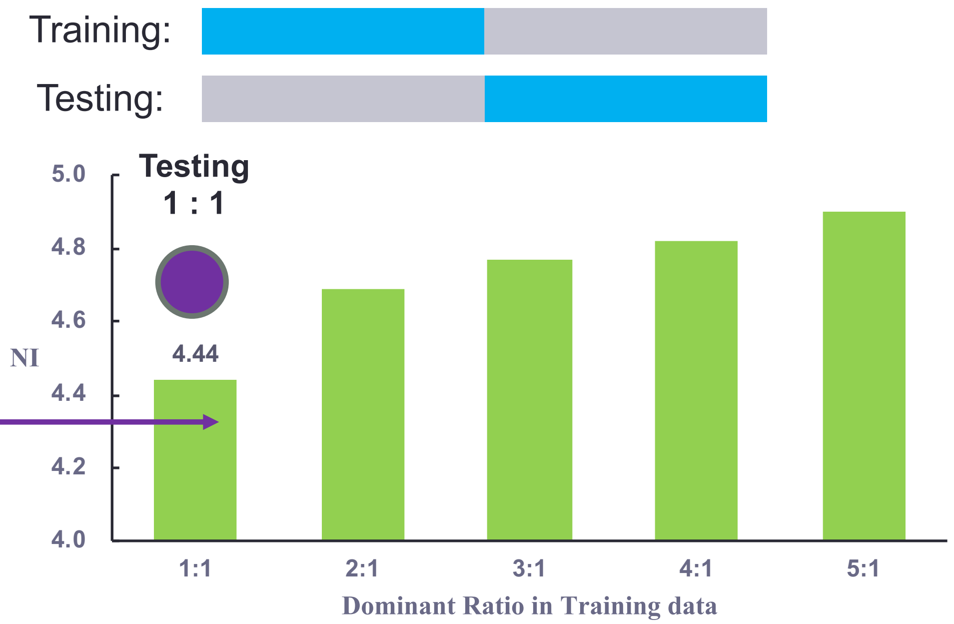
Compositional Bias

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

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



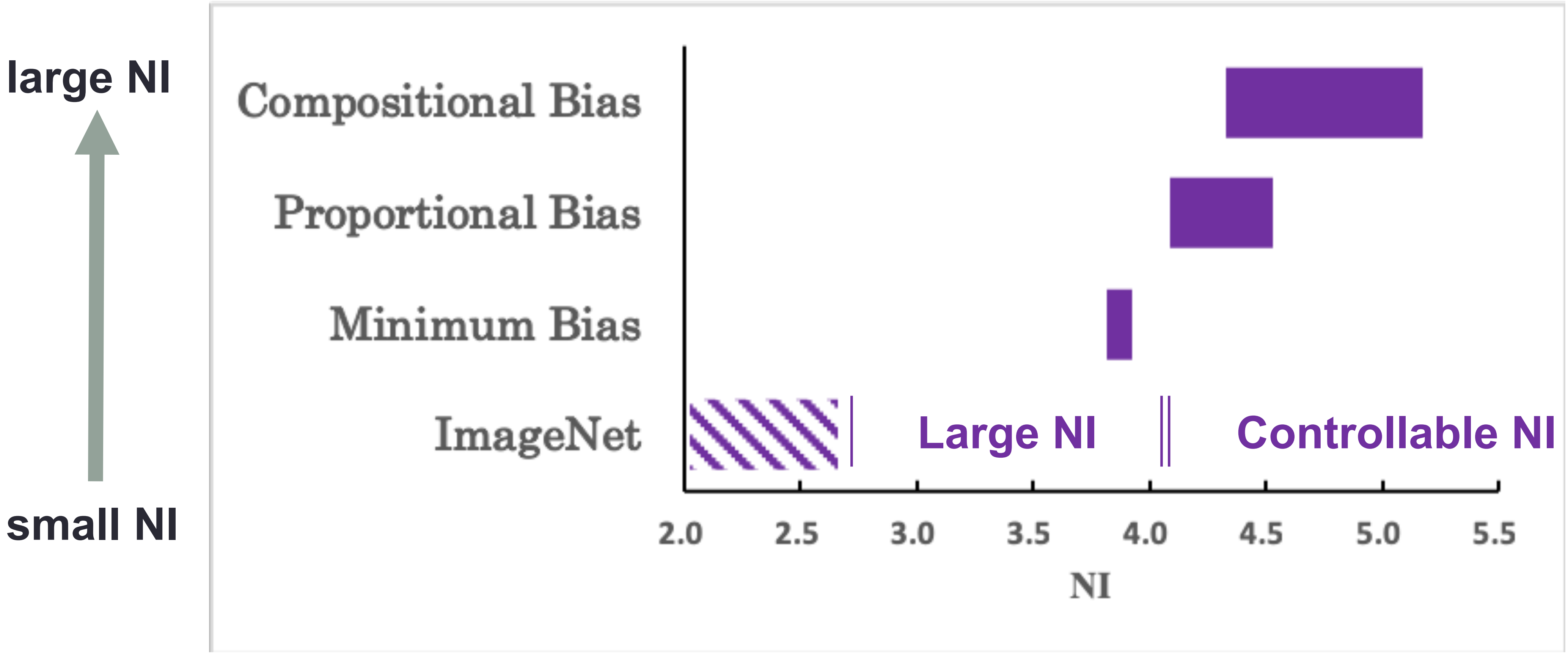
Moderate setting
(Overlap)



Radical setting
(No Overlap &
Dominant ratio)

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

- Large and controllable NI



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

- The dataset can be downloaded from (temporary address):
- https://www.dropbox.com/sh/8mouawi5guaupyb/AAD4fdySrA6fn3PgSmhKwFgva?dl=0

- Please refer to the following paper for details:
- Yue He, Zheyan Shen, Peng Cui. NICO: A Dataset Towards Non-I.I.D. Image Classification. https://arxiv.org/pdf/1906.02899.pdf

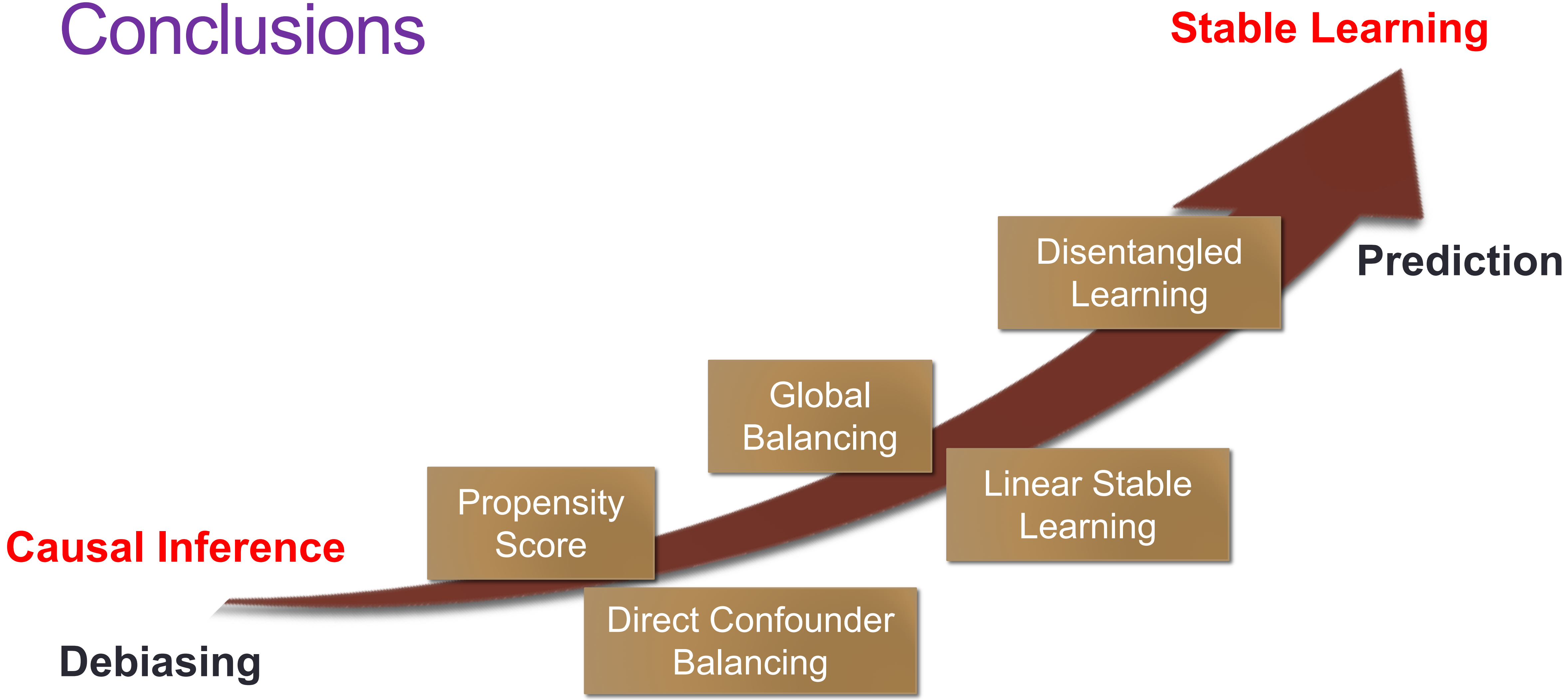
Outline

- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Conclusions

Conclusions

- Predictive modeling is not only about Accuracy.
- **Stability** is critical for us to trust a predictive model.
- Causality has been demonstrated to be useful in stable prediction.
- How to marry causality with predictive modeling effectively and efficiently is still an open problem.

Conclusions



Reference

- Shen Z, Cui P, Kuang K, et al. Causally regularized learning with agnostic data selection bias[C]//2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 411-419.
- Kuang K, Cui P, Athey S, et al. Stable prediction across unknown environments[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 1617-1626.
- Kuang K, Cui P, Li B, et al. Estimating treatment effect in the wild via differentiated confounder balancing[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 265-274.
- Kuang K, Cui P, Li B, et al. Treatment effect estimation with data-driven variable decomposition[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.
- Kuang K, Jiang M, Cui P, et al. Steering social media promotions with effective strategies[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016: 985-990.

Reference

- Pearl J. Causality[M]. Cambridge university press, 2009.
- Austin P C. An introduction to propensity score methods for reducing the effects of confounding in observational studies[J]. Multivariate behavioral research, 2011, 46(3): 399-424.
- Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference[C]//International conference on machine learning. 2016: 3020-3029.
- Shalit U, Johansson F D, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 3076-3085.
- Johansson F D, Kallus N, Shalit U, et al. Learning weighted representations for generalization across designs[J]. arXiv preprint arXiv:1802.08598, 2018.
- Louizos C, Shalit U, Mooij J M, et al. Causal effect inference with deep latent-variable models[C]//Advances in Neural Information Processing Systems. 2017: 6446-6456.
- Thomas V, Bengio E, Fedus W, et al. Disentangling the independently controllable factors of variation by interacting with the world[J]. arXiv preprint arXiv:1802.09484, 2018.
- Bengio Y, Deleu T, Rahaman N, et al. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms[J]. arXiv preprint arXiv:1901.10912, 2019.

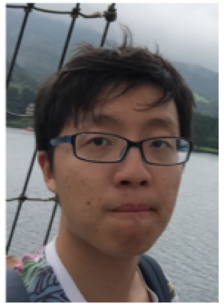
Reference

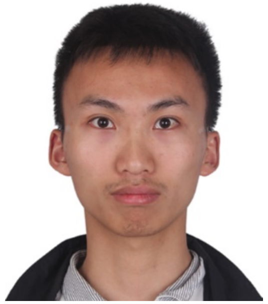
- Yu B. Stability[J]. Bernoulli, 2013, 19(4): 1484-1500.
- Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- Volpi R, Namkoong H, Sener O, et al. Generalizing to unseen domains via adversarial data augmentation[C]//Advances in Neural Information Processing Systems. 2018: 5334-5344.
- Ye N, Zhu Z. Bayesian adversarial learning[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., 2018: 6892-6901.
- Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation[C]//International Conference on Machine Learning. 2013: 10-18.
- Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016, 78(5): 947-1012.
- Rojas-Carulla M, Schölkopf B, Turner R, et al. Invariant models for causal transfer learning[J]. The Journal of Machine Learning Research, 2018, 19(1): 1309-1342.
- Rothenhäusler D, Meinshausen N, Bühlmann P, et al. Anchor regression: heterogeneous data meets causality[J]. arXiv preprint arXiv:1801.06229, 2018.

Acknowledgement







Thanks!

Peng Cui
cuip@tsinghua.edu.cn
http://pengcui.thumedialab.com

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