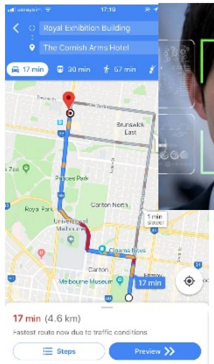
Causal Inference and Stable Learning

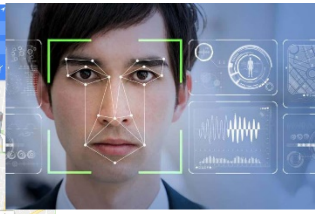
Peng Cui
Tsinghua University

Tong Zhang
Hong Kong University of Science and Technology

ML techniques are impacting our life

- A day in our life with ML techniques





8:00 am

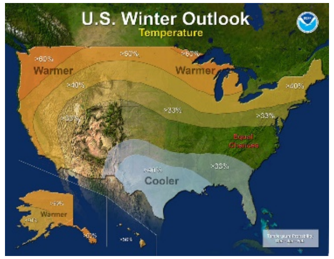
10:00 am

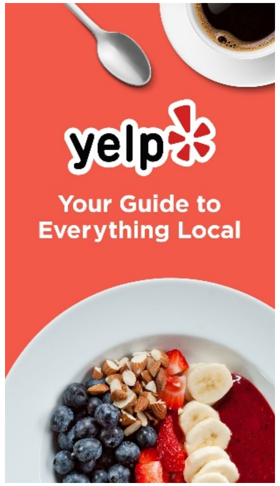
6:00 pm

8:30 am

4:00 pm

8:00 pm





Now we are stepping into risk-sensitive areas

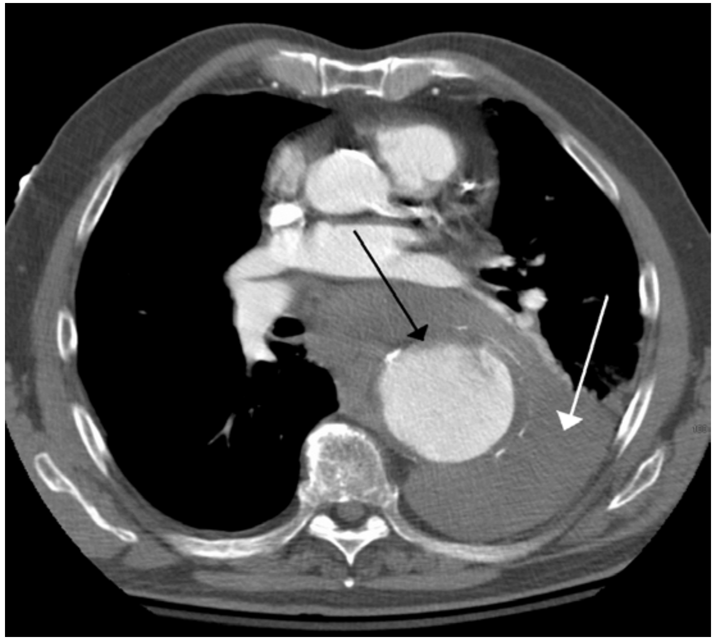
Human

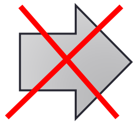
Shifting from ***Performance Driven*** to ***Risk Sensitive***

Problems of today's ML - *Explainability*

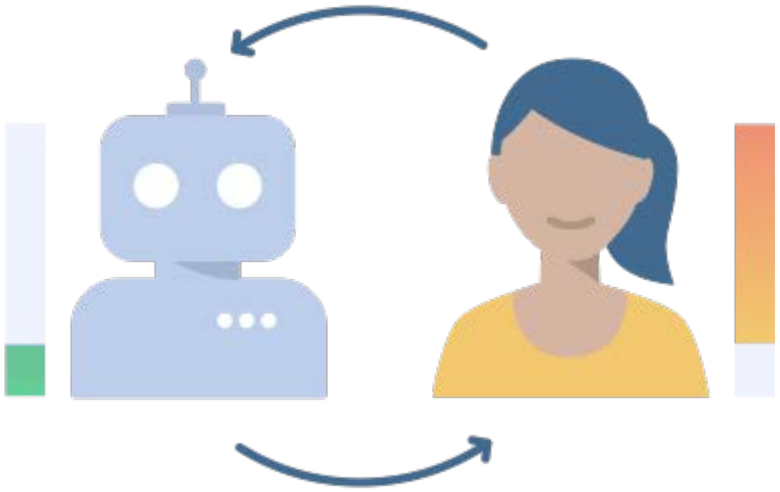
Most machine learning models are black-box models

Unexplainable





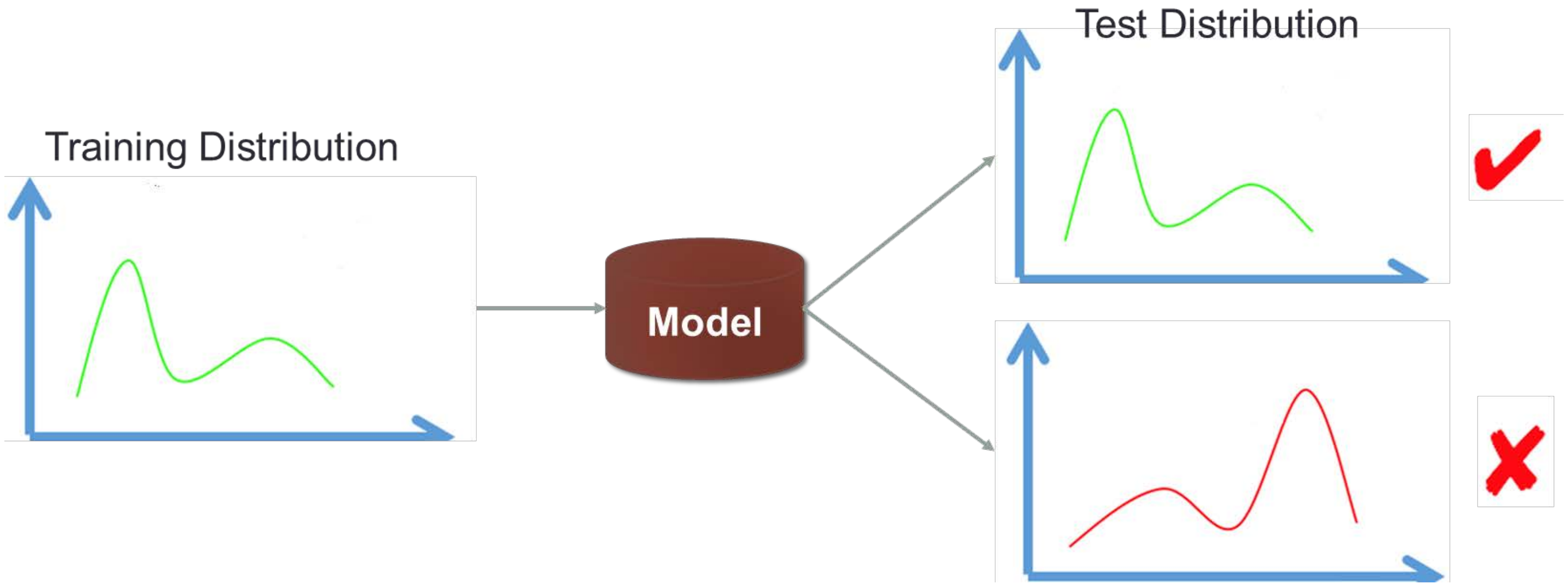
Human in the loop



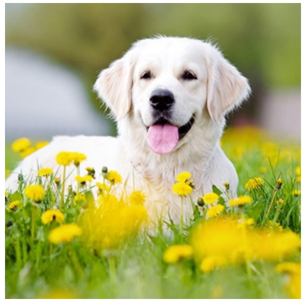
Health Military Finance Industry

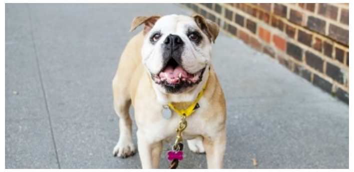
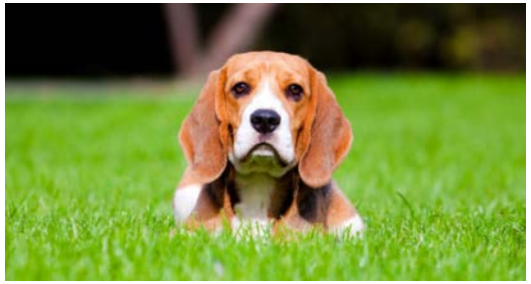
Problems of today's ML - ***Stability***

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



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

Yes

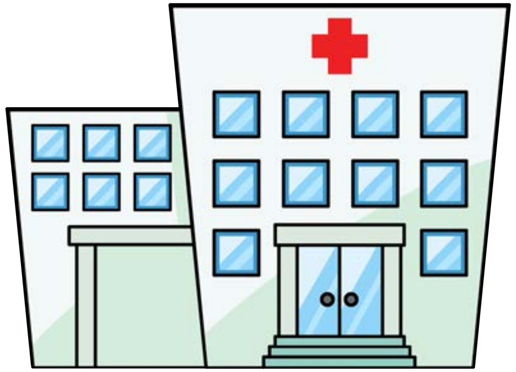
Maybe

No

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

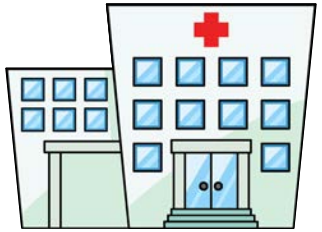
- Cancer survival rate prediction

Training Data



Predictive Model

Testing Data



City Hospital

City Hospital

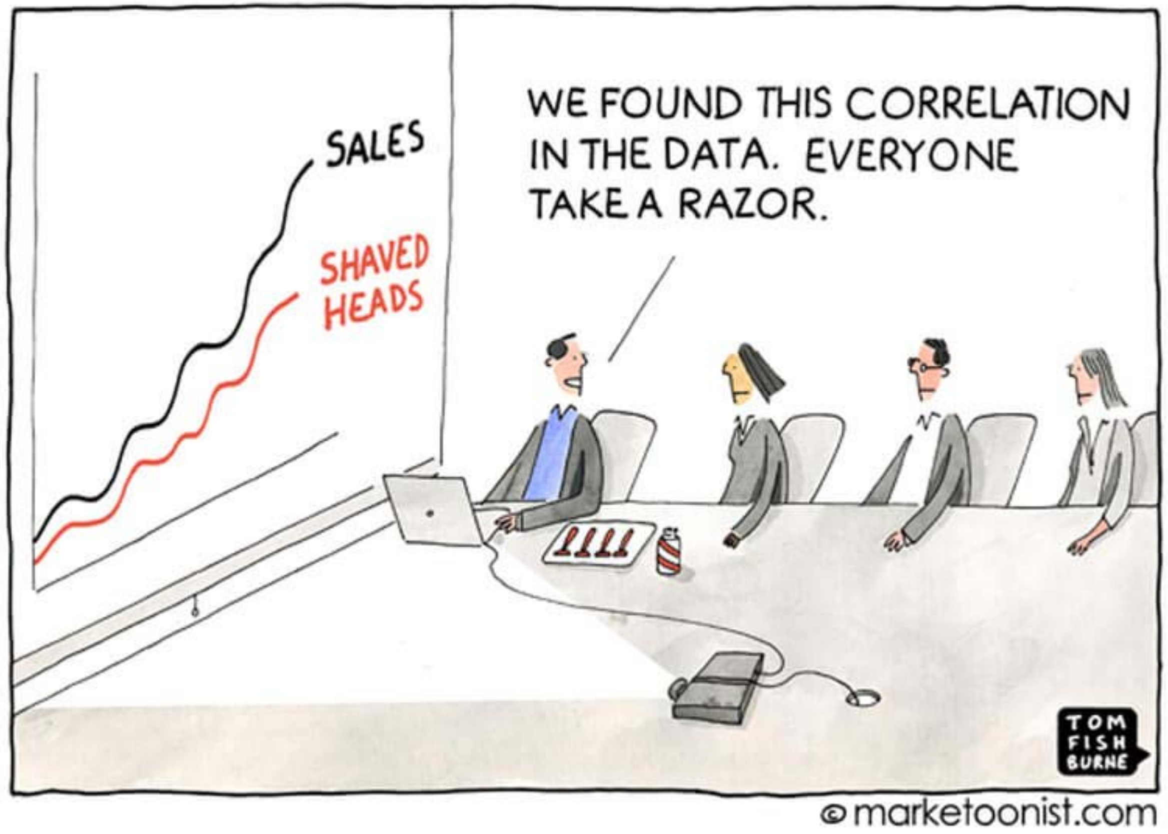
Higher income, higher survival rate.

University Hospital

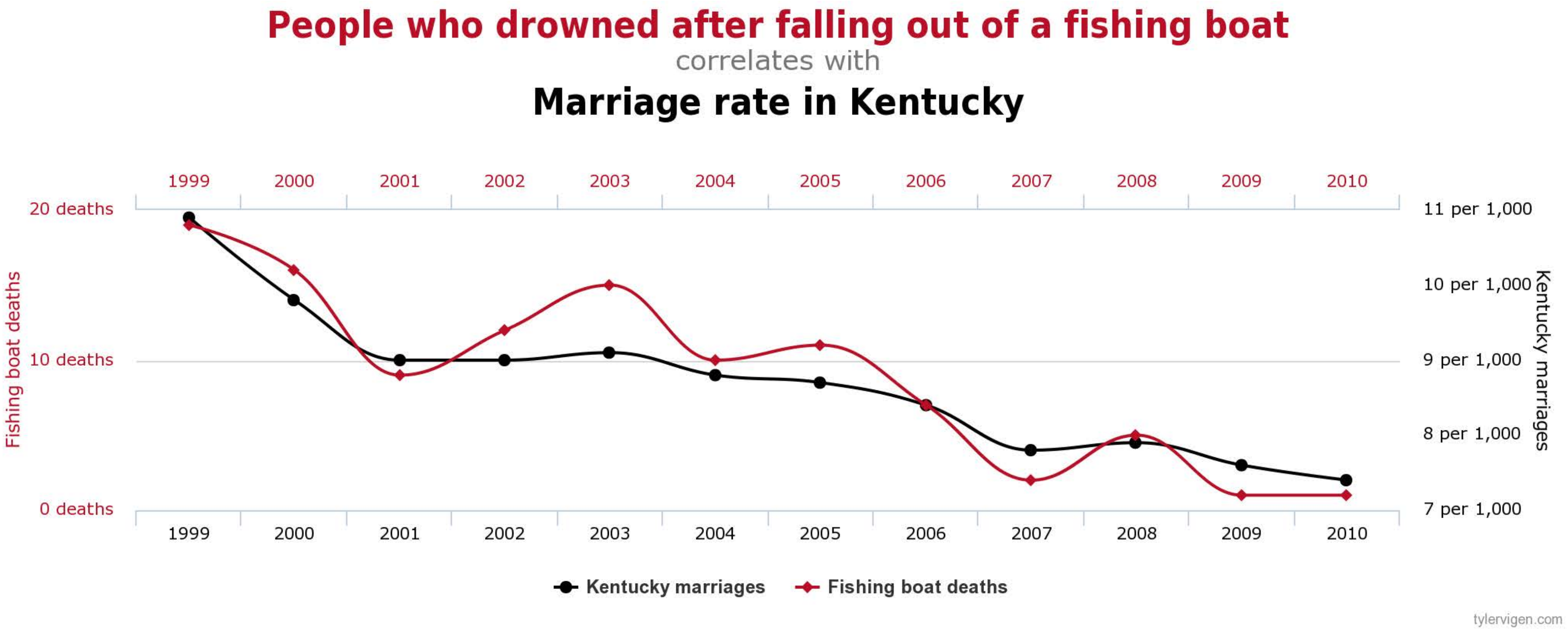
Survival rate is not so correlated with income.

A plausible reason: ***Correlation***

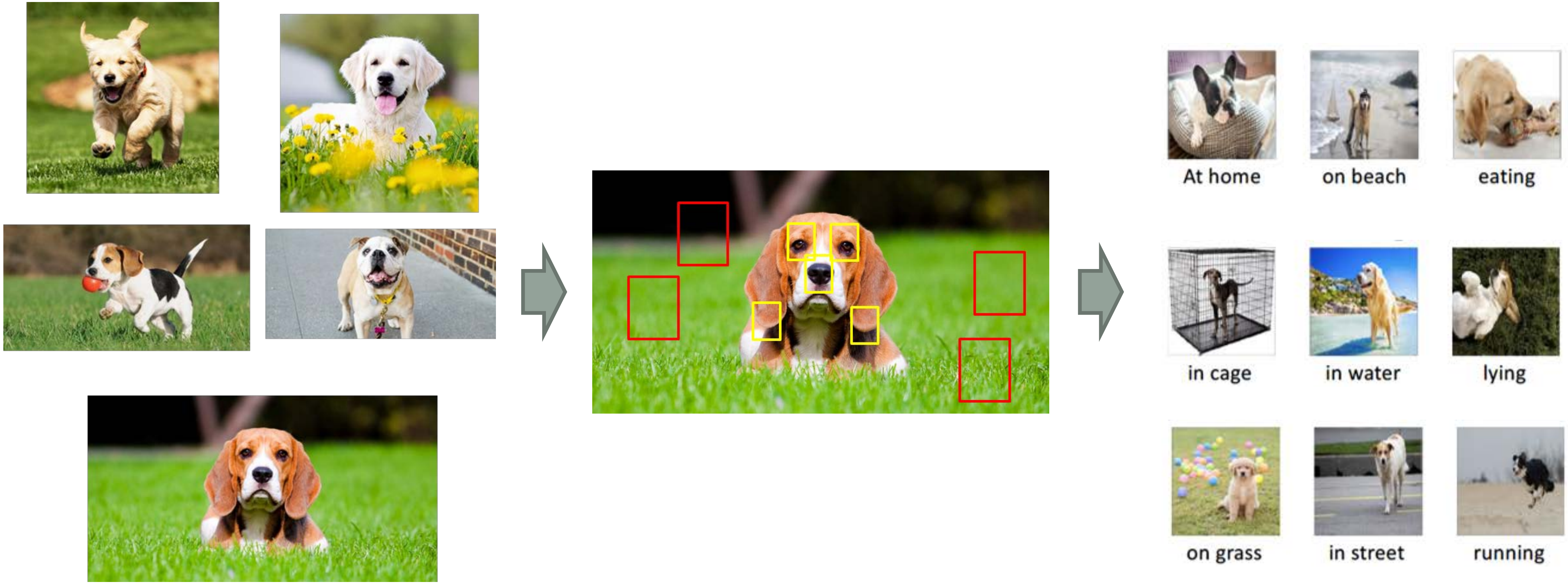
Correlation is the very basics of machine learning.



Correlation is not explainable

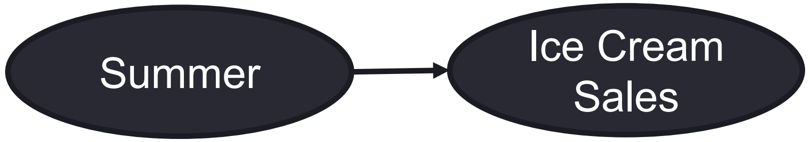


Correlation is ‘*unstable*’

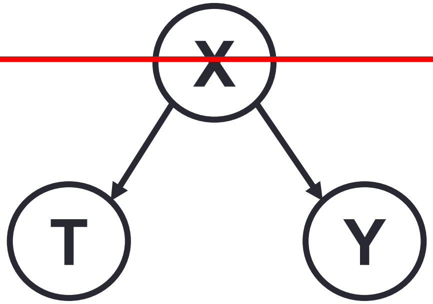


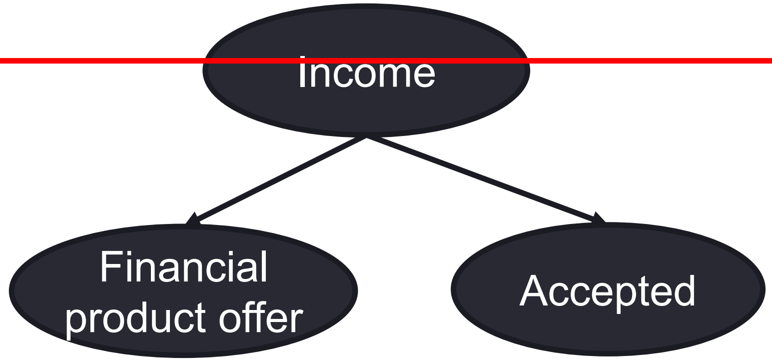
It’s not the fault of ***correlation***, but the way we use it

- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and explainable

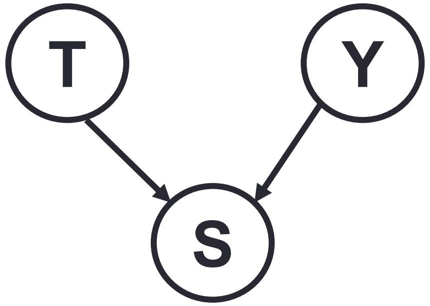


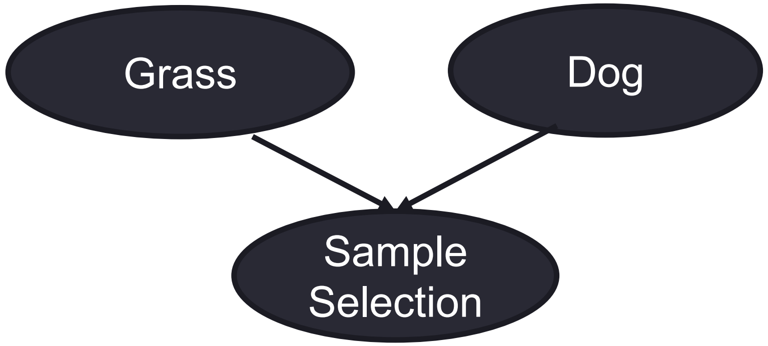
  - Confounding
    - Ignoring X
    - Spurious Correlation





  - Sample Selection Bias
    - Conditional on S
    - Spurious Correlation



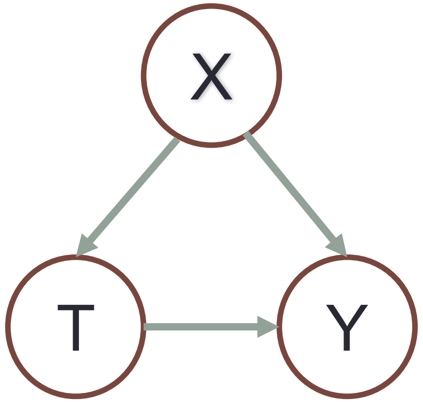


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



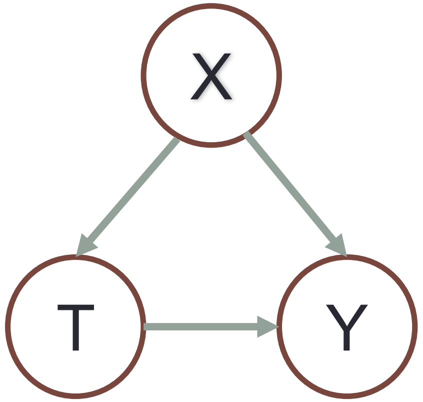
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

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

The ***benefits*** of bringing causality into learning

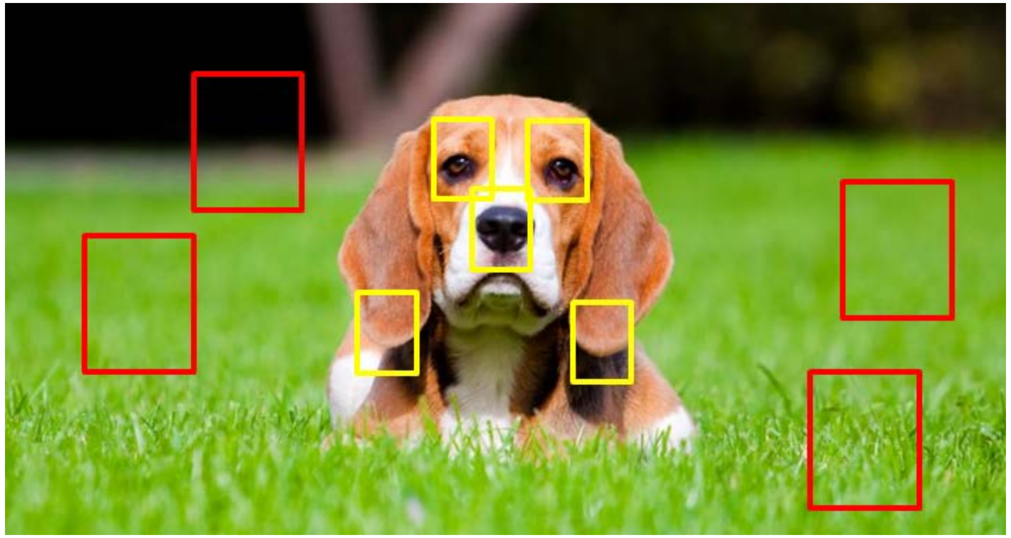
Causal Framework



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*?

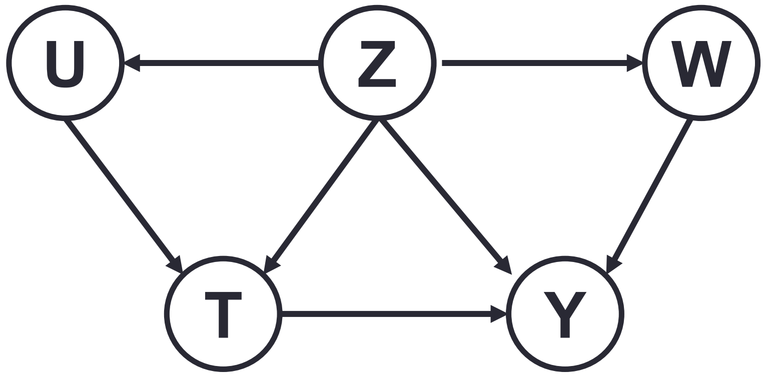
Outline

- 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

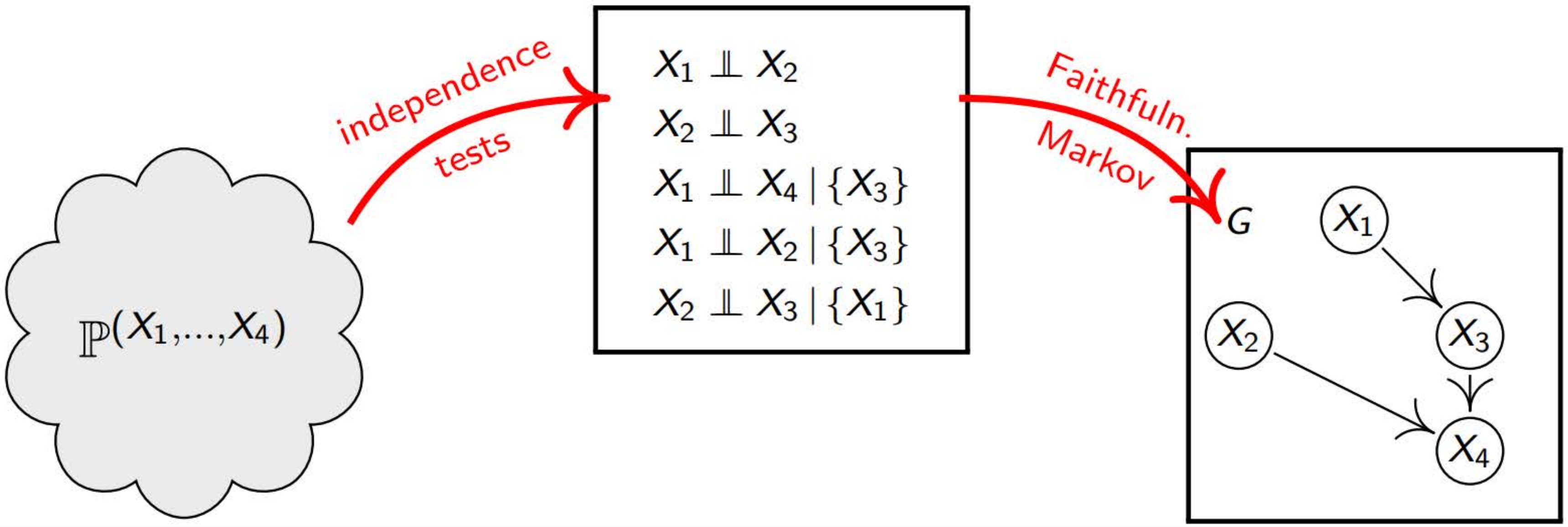
- Causal Identification with back door criterion
- Causal Estimation with do calculus



How to discover the causal structure?

Paradigms – Structural Causal Model

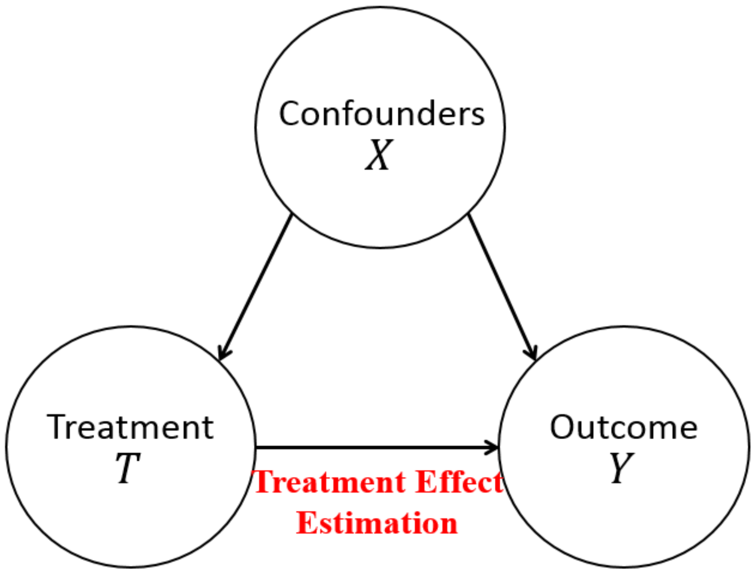
- Causal Discovery
  - Constraint-based: conditional independence
  - Functional causal model based



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

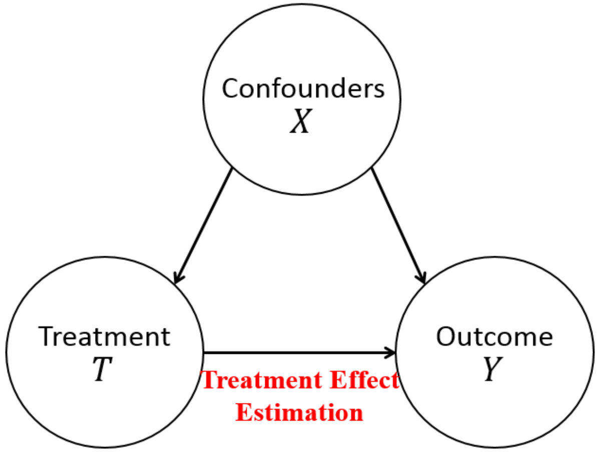


More like a ***discriminative*** way to estimate treatment's partial effect on outcome.

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

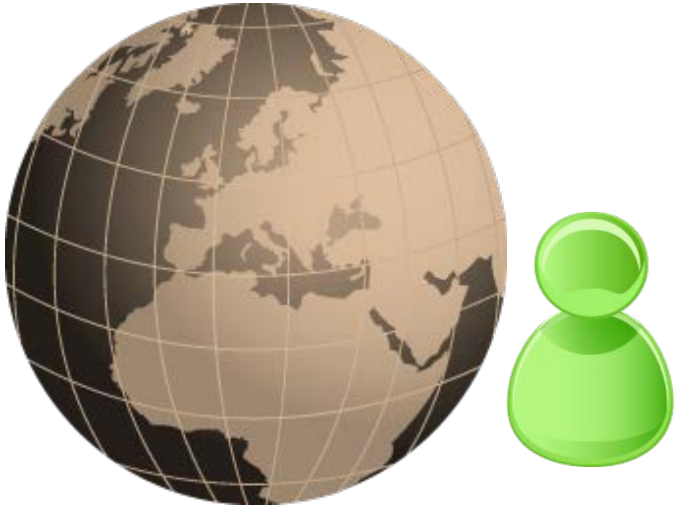
- Two key points for causal effect estimation
  - Changing T
  - Keeping everything else constant

- 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”

Observational Studies!

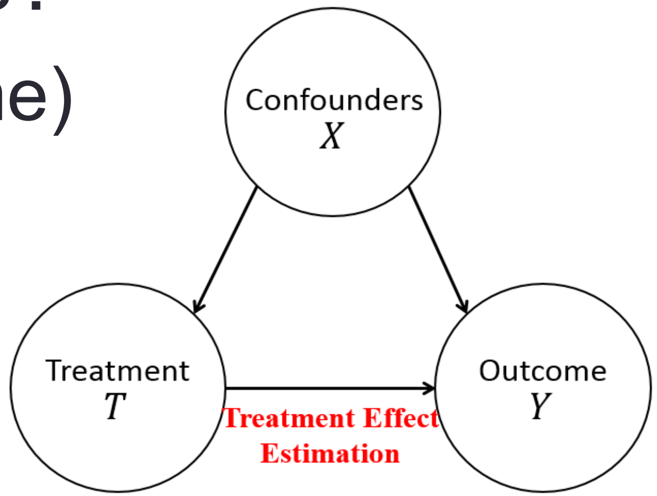
- Drawbacks … nts:
  - Cost
  - Unethical
  - Unrealistic

Causal Inference with Observational Data

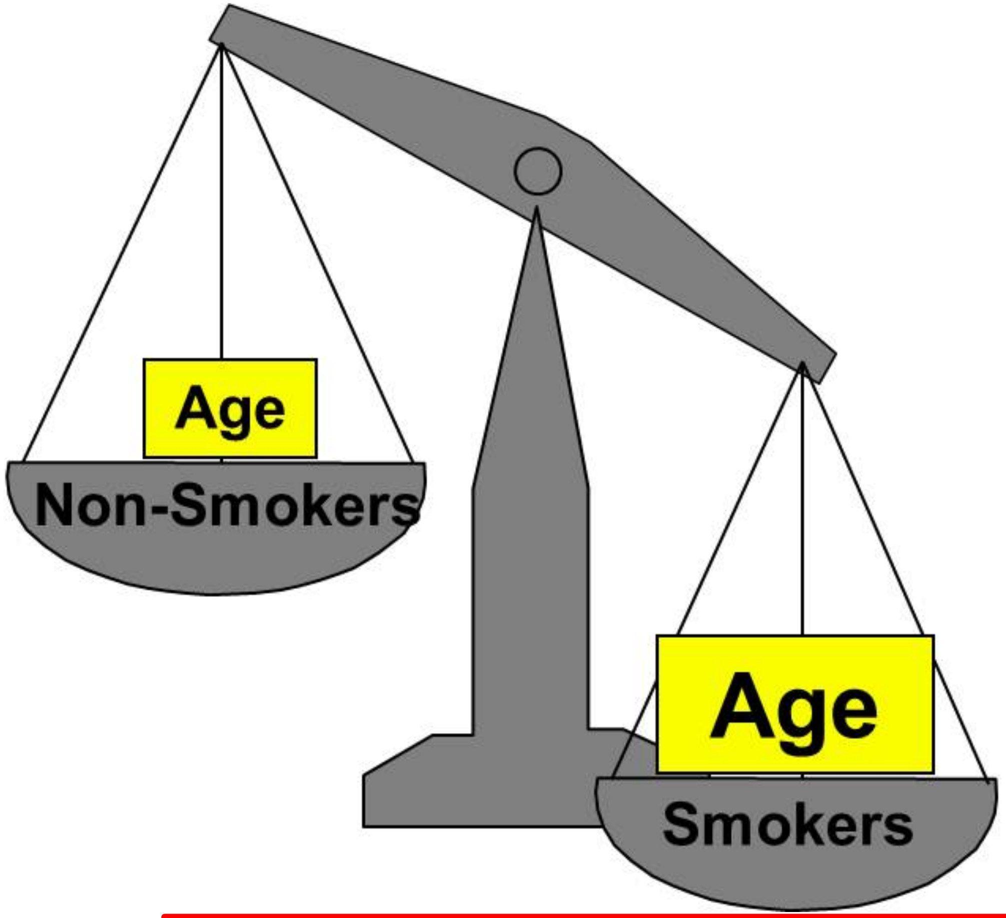
- Counterfactual Problem:

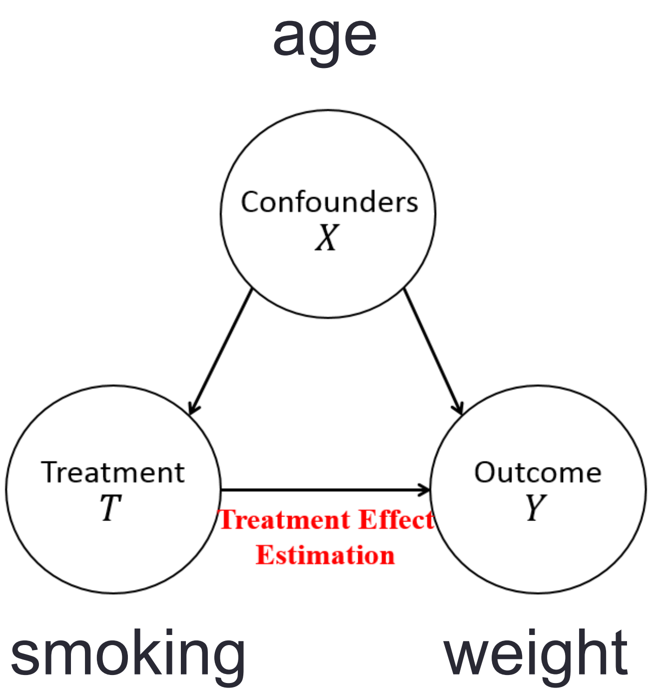
$$Y(T=1) \quad \text{or} \quad Y(T=0)$$

- 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)



Confounding Effect



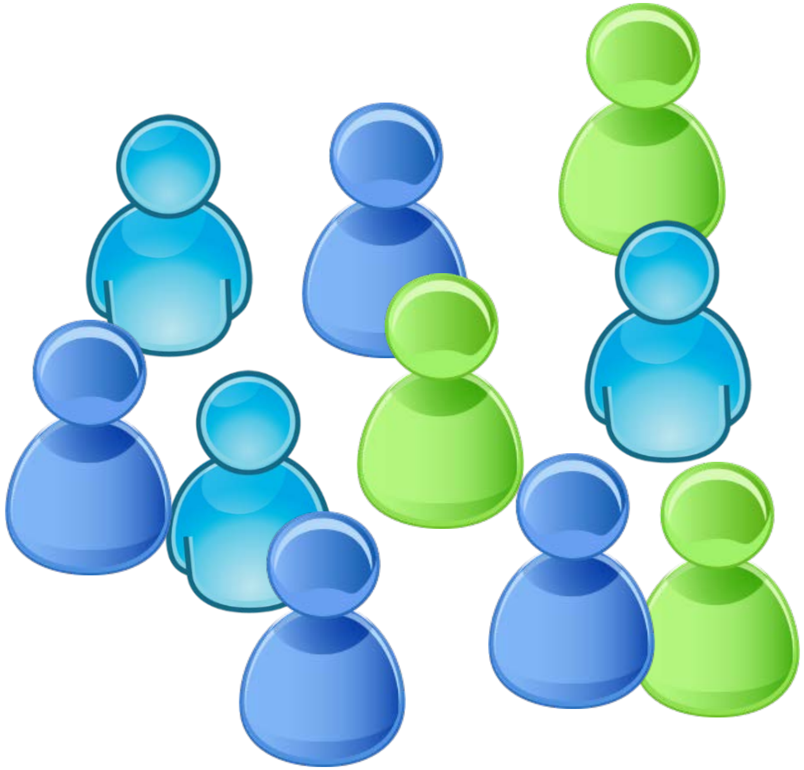


Balancing Confounders' Distribution

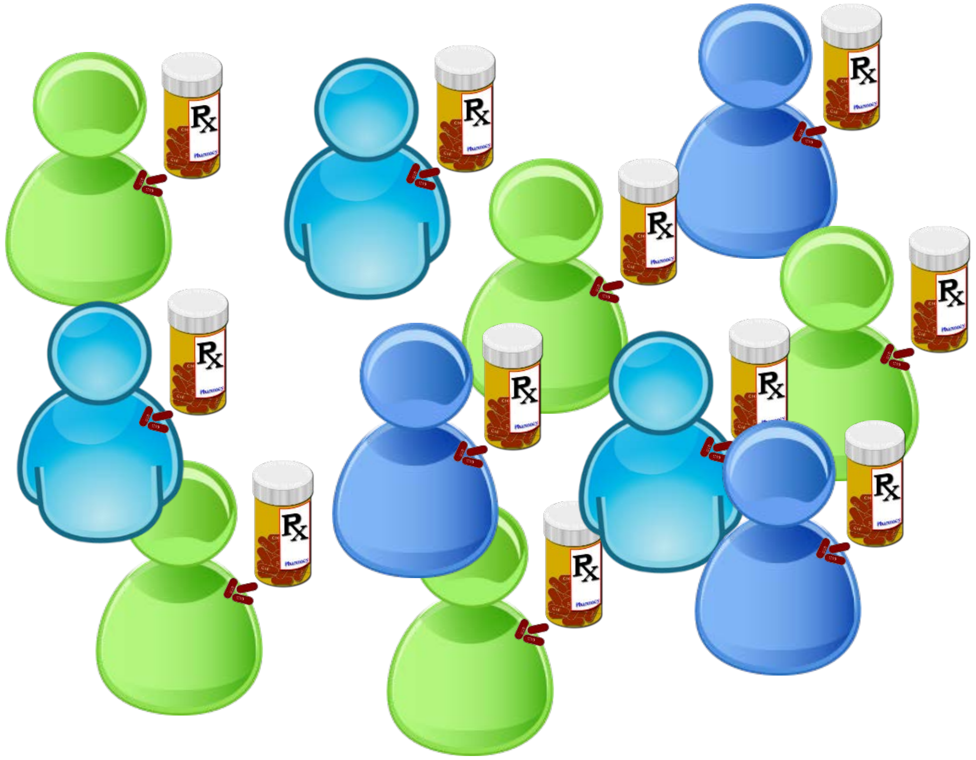
Methods for Causal Inference

- Matching
- Propensity Score
- Directly Confounder Balancing

Matching

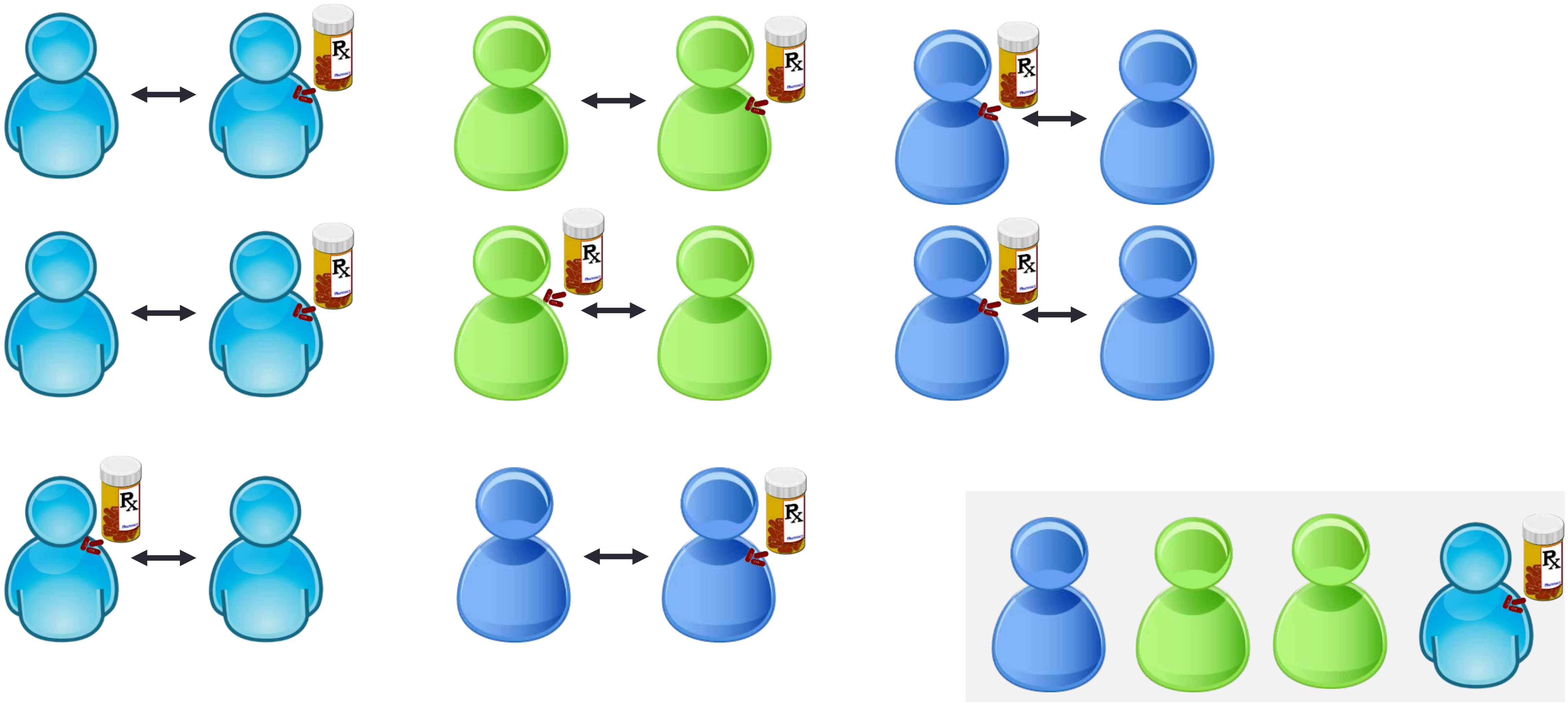


$T = 0$



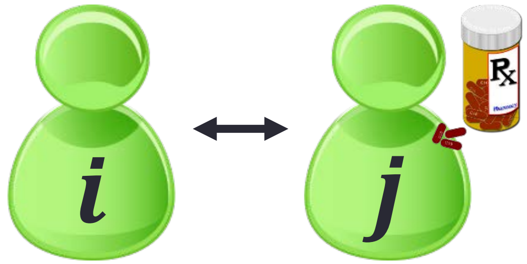
$T = 1$

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 guarantee that the everything else (Confounders) approximate constant
- Small ϵ: less bias, but higher variance
- Fit for low-dimensional settings
- But in high-dimensional settings, there will be few exact matches

Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Propensity Score Based Methods

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

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

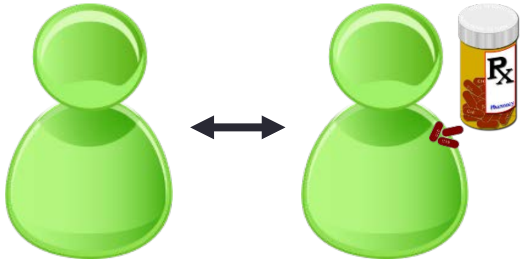
- Then, Donald Rubin shows that the propensity score is sufficient to control or summarize the information of confounders

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

- Propensity scores cannot be 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: from matching to PS estimation

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3):399–424, 2011.

Inverse of Propensity Weighting (IPW)

- Why weighting with inverse of propensity score?
  - 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: $w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

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

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Non-parametric solution

- Model specification problem is inevitable
- Can we directly learn sample weights that can balance confounders’ distribution between treated and control groups?

Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

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 (ATT problem)

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

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.

J. Hainmueller. Entropy balancing for causal effects: A mul- tivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1):25–46, 2012.

Entropy Balancing

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

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

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

- Directly confounder balancing by sample weights W
- Minimize the entropy of sample weights W

Either know confounders a priori or regard all variables as confounders . All confounders are balanced equally.

Athey S, et al. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B, 2018, 80(4): 597-623.

Differentiated Confounder Balancing

- **Idea**: Different confounders make different confounding bias
- 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.

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

Differentiated Confounder Balancing

- General relationship among X, T, and Y:

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

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

Confounder weights

Confounding bias

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

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

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

Differentiated Confounder Balancing

- **Ideas**: simultaneously learn ***confounder weights β*** and ***sample weighs W***.

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

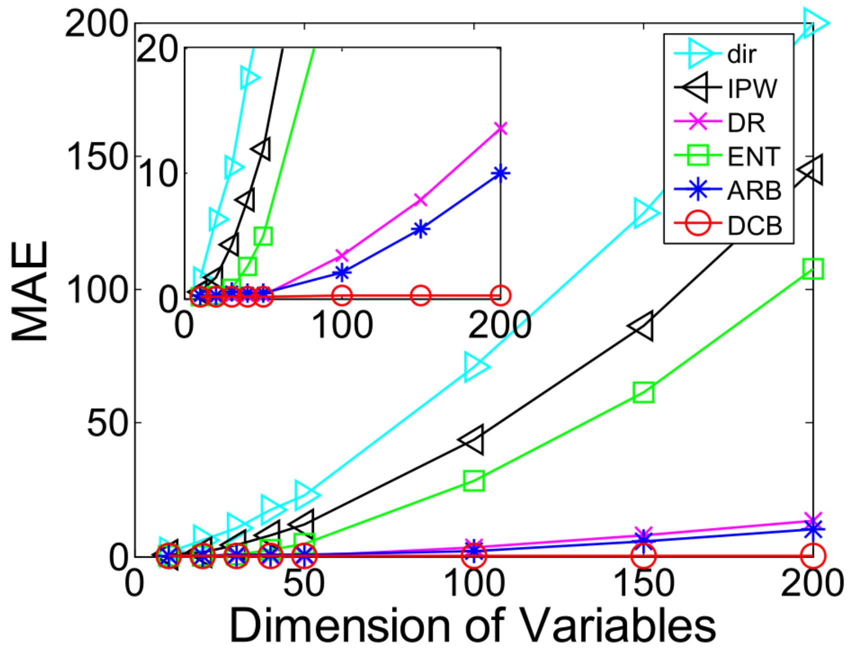
$$s.t. \quad \|W\|_2^2 \le \delta, \ \ \|\beta\|_2^2 \le \mu, \ \ \|\beta\|_1 \le \nu, \mathbf{1}^T W = 1 \ \ and \ \ W \succeq 0$$

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

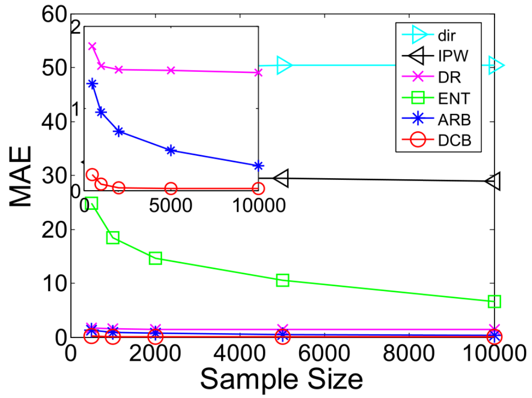
- The ENT algorithm is a special case of DCB algorithm by setting the confounder weights as unit vector.

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

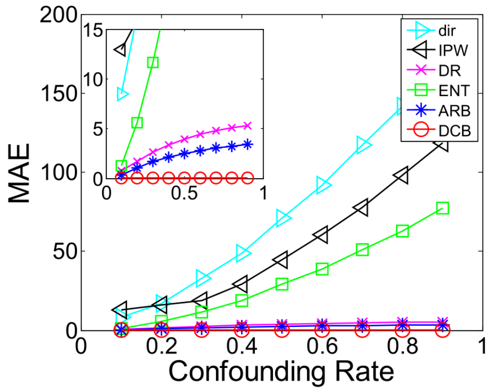
Experiments



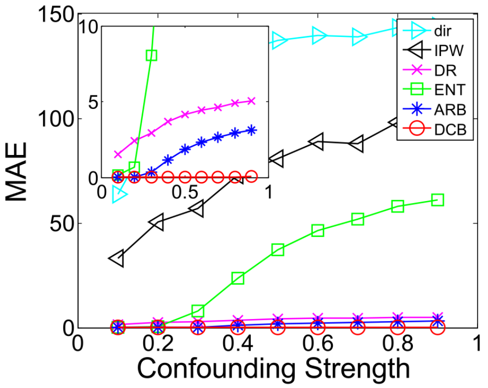
(b) dimension of variables p



(a) sample size n



(c) confounding rate r_c



(d) confounding strength s_c

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)

LaLonde

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

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

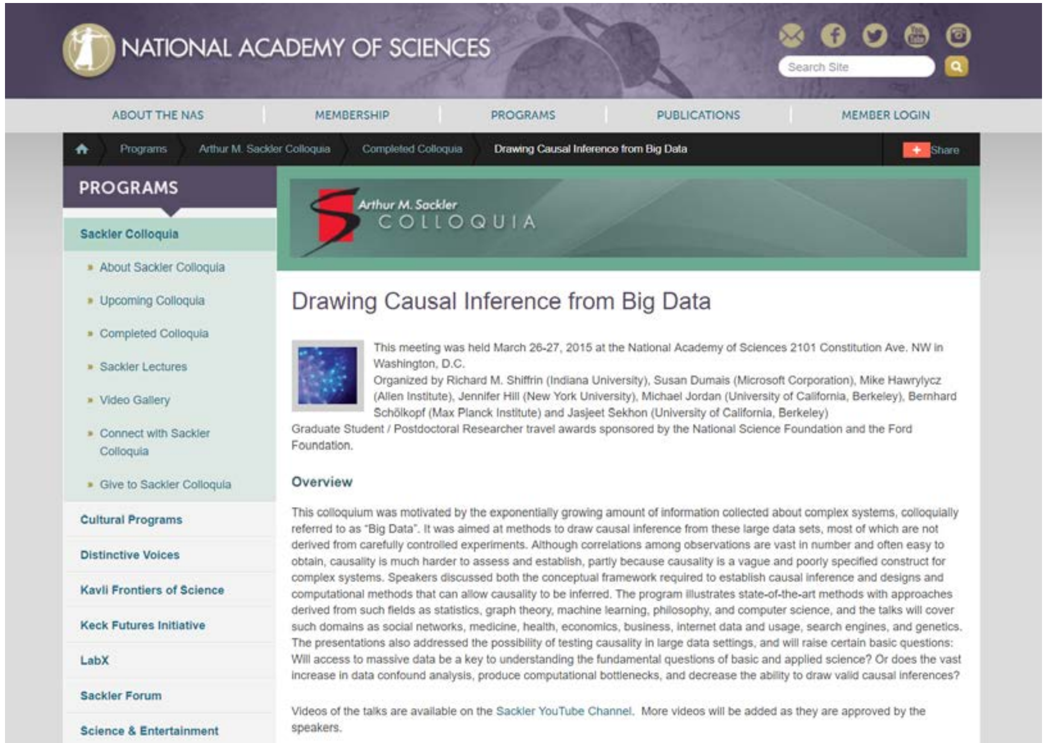
- **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

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

Sectional Summary

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

Ready for Learning?



Outline

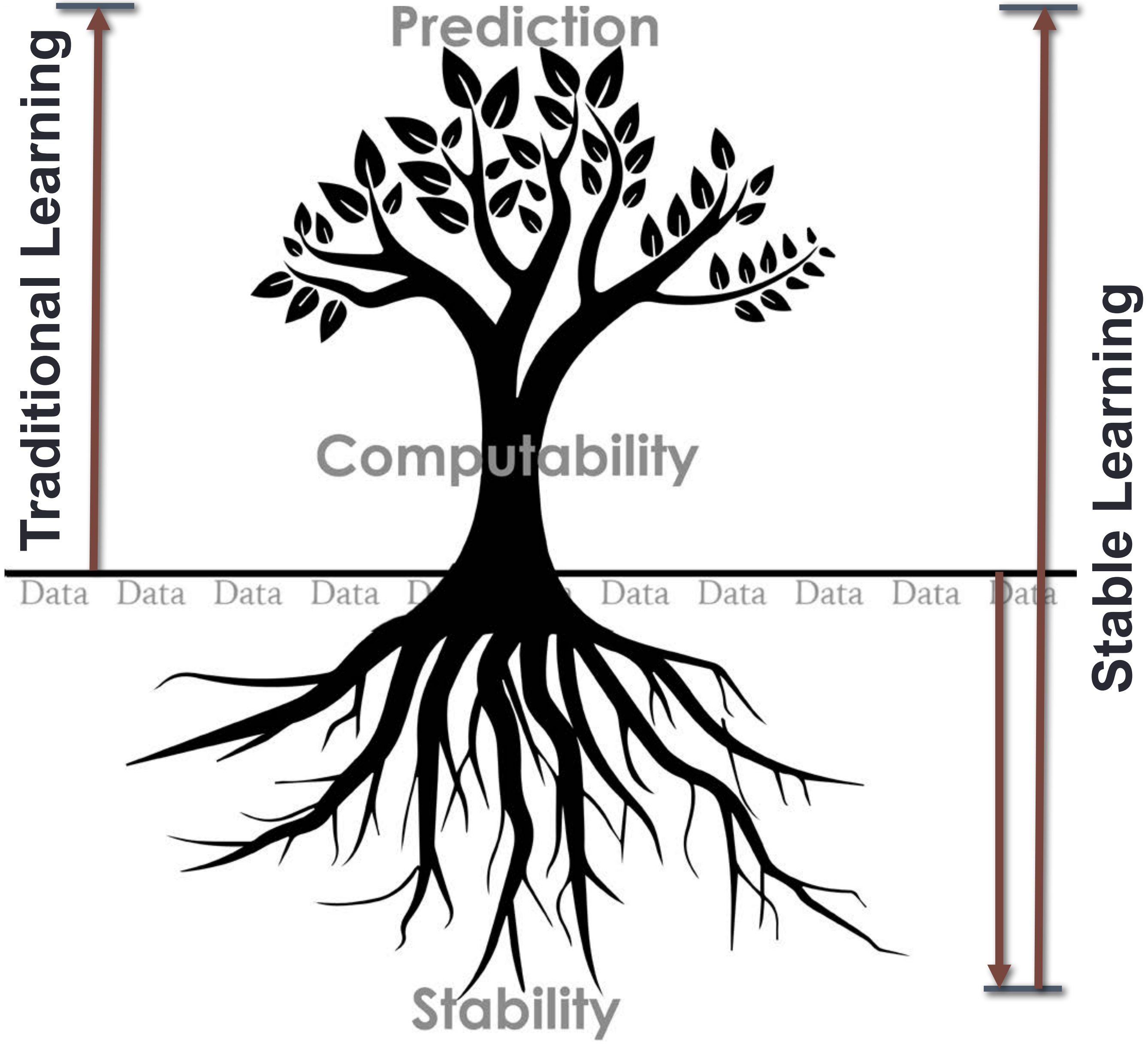
- Correlation v.s. Causality
- Causal Inference
- Stable Learning
- NICO: An Image Dataset for Stable Learning
- Future Directions and Conclusions

Stability and Prediction

Prediction Performance

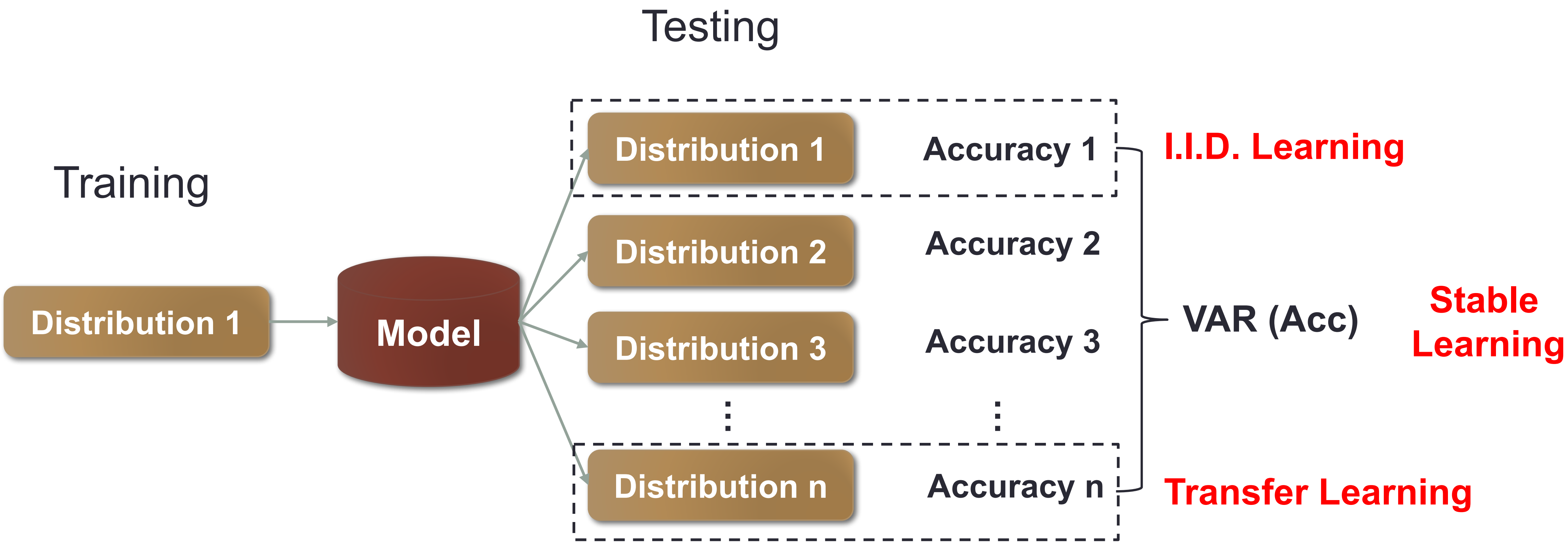
Learning Process

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

Stability

- Statistical stability holds if statistical conclusions are robust to appropriate perturbations to data.
  - Prediction Stability
  - Estimation Stability

Bernoulli **19**(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: binyu@stat.berkeley.edu

Prediction Stability

- Lasso

$$\hat{\beta}(\lambda) = \arg_{\beta \in R^p} \{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\},$$

- Prediction Stability by Cross-Validation
  - n data units are randomly partitioned into V blocks, each block has d = [n/V] units.
  - Leave one out: training on (n-d) units, validating on d units.
  - CV does not provide a good interpretable model because Lasso+CV is unstable.

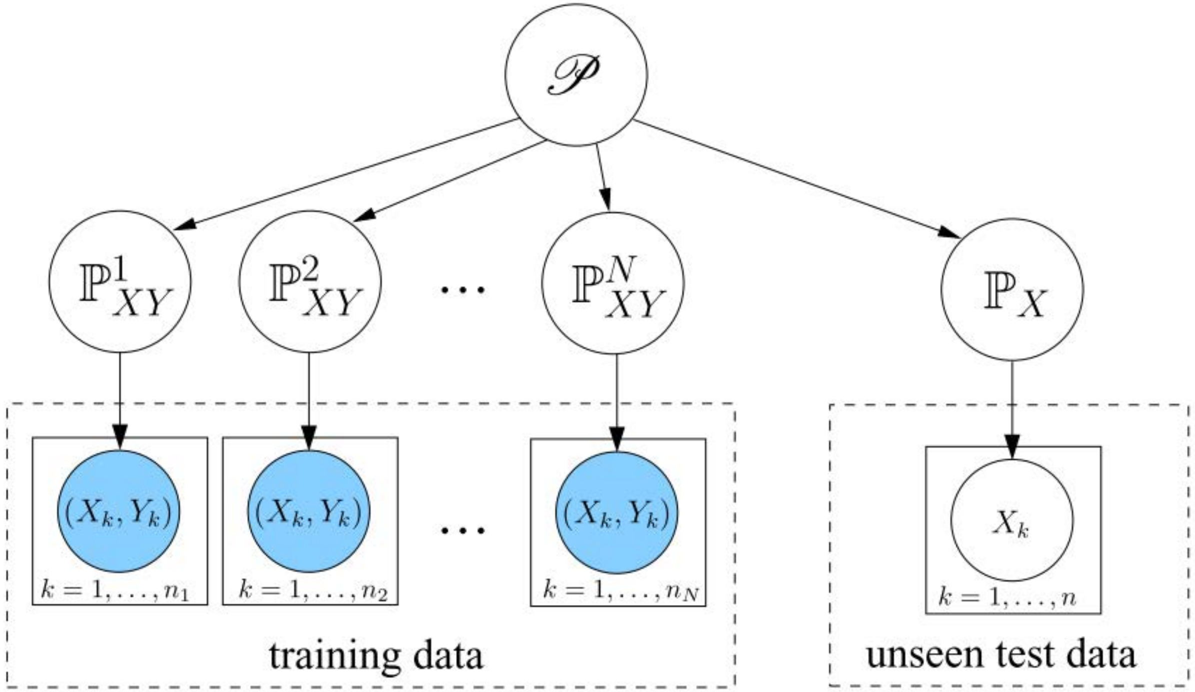
Estimation Stability

- Estimation Stability:
  - Mean regression function: $$\hat{m}(\tau) = \frac{1}{V}\sum_v X\hat{\beta}_v(\tau),$$
  - Variance of function m: $$\hat{T}(\tau) = \frac{n-d}{d}\frac{1}{V}\sum_v \left(\|X\hat{\beta}_v(\tau) - \hat{m}(\tau)\|^2\right).$$
  - Estimation Stability:

$$ES(\tau) = \frac{1/V \sum_v \|X\hat{\beta}_v(\tau) - \hat{m}(\tau)\|^2}{\hat{m}^2(\tau)} = \frac{d}{n-d}\frac{\hat{T}(\tau)}{\hat{m}^2(\tau)}$$

ES+CV is better than Lasso+CV

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.

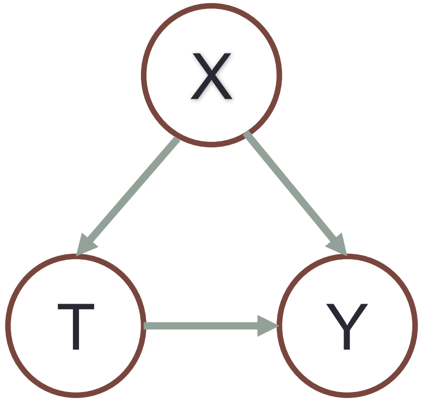
$$\text{for all } e \in \mathcal{E}, X^e \text{ has an arbitrary distribution and}$$

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

- **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

↓

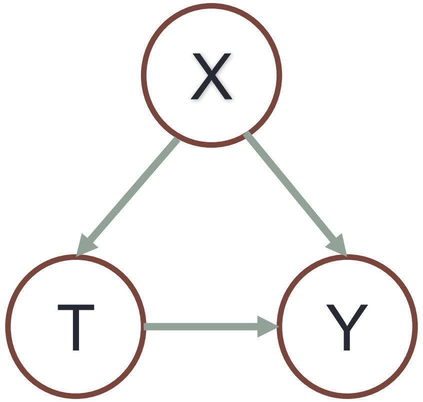
Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

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^*.

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

$$\downarrow$$

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:

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

$\square$

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

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

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

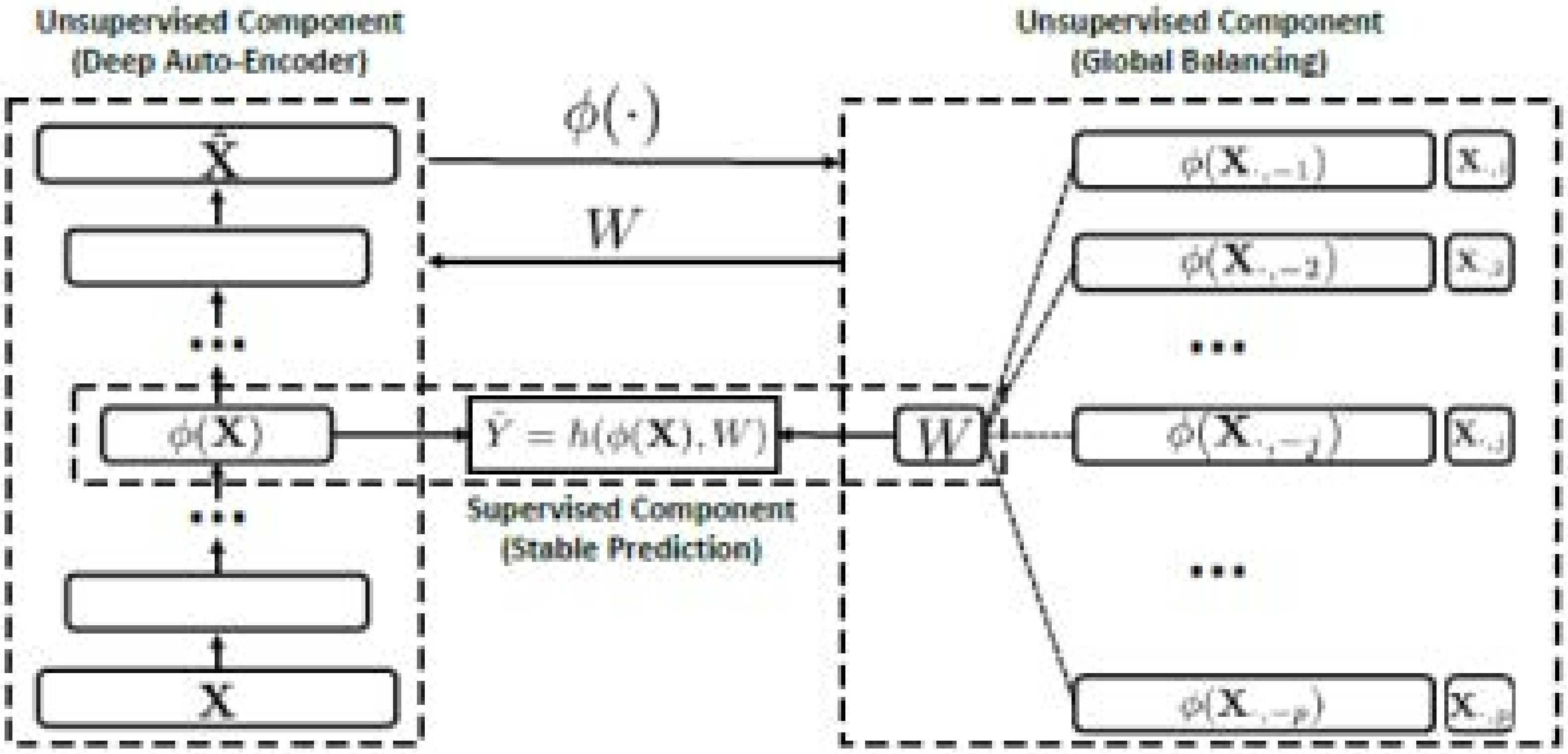
$$
\begin{aligned}
\min \quad & \boxed{\sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (x_i \boxed{\beta})))}, \\
s.t. \quad & \sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{-j}^T \cdot (W \odot \mathbf{I}_j)}{W^T \cdot \mathbf{I}_j} - \frac{\mathbf{X}_{-j}^T \cdot (W \odot (1 - \mathbf{I}_j))}{W^T \cdot (1 - \mathbf{I}_j)} \right\|_2^2 \leq \lambda_1, \\
& W \succeq 0, \ \|W\|_2^2 \leq \lambda_2, \ \|\beta\|_2^2 \leq \lambda_3, \ \|\beta\|_1 \leq \lambda_4, \\
& (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_5,
\end{aligned}
$$

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

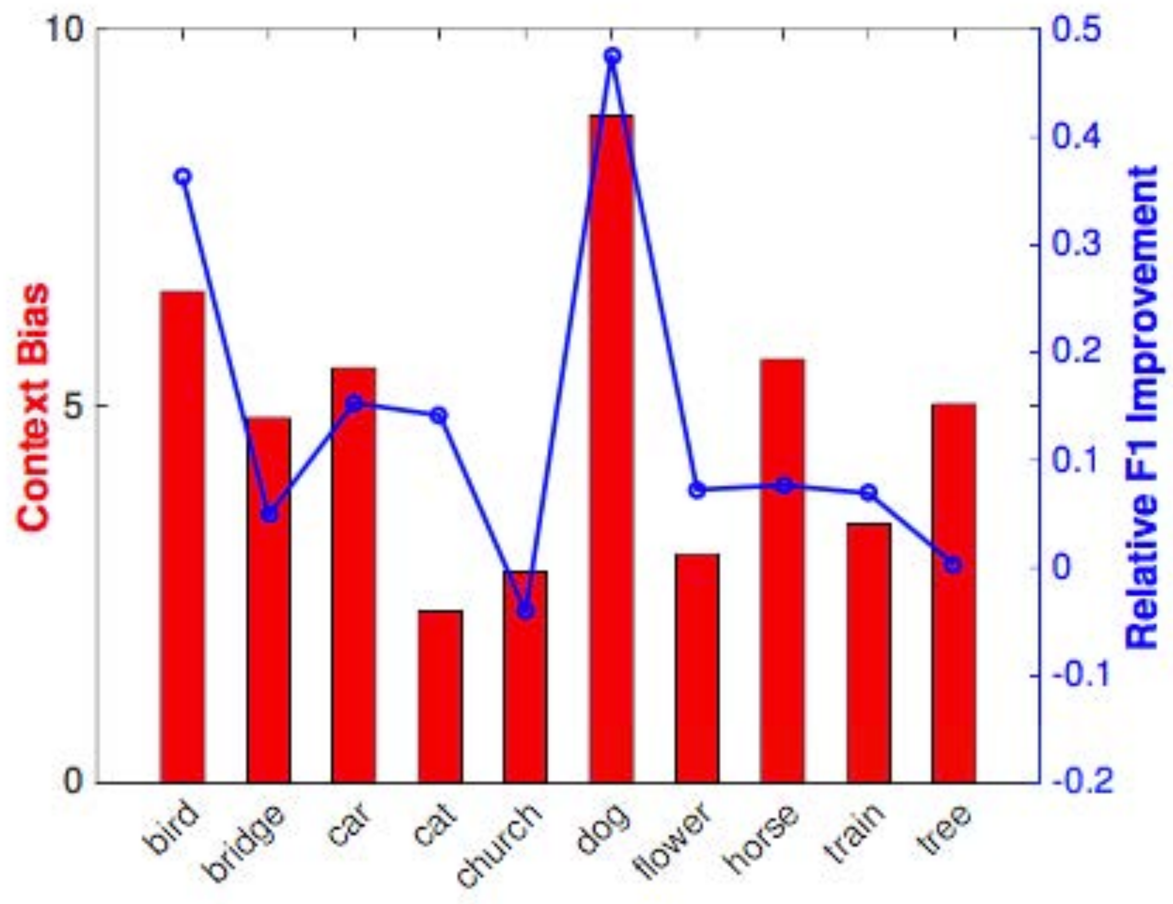


Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

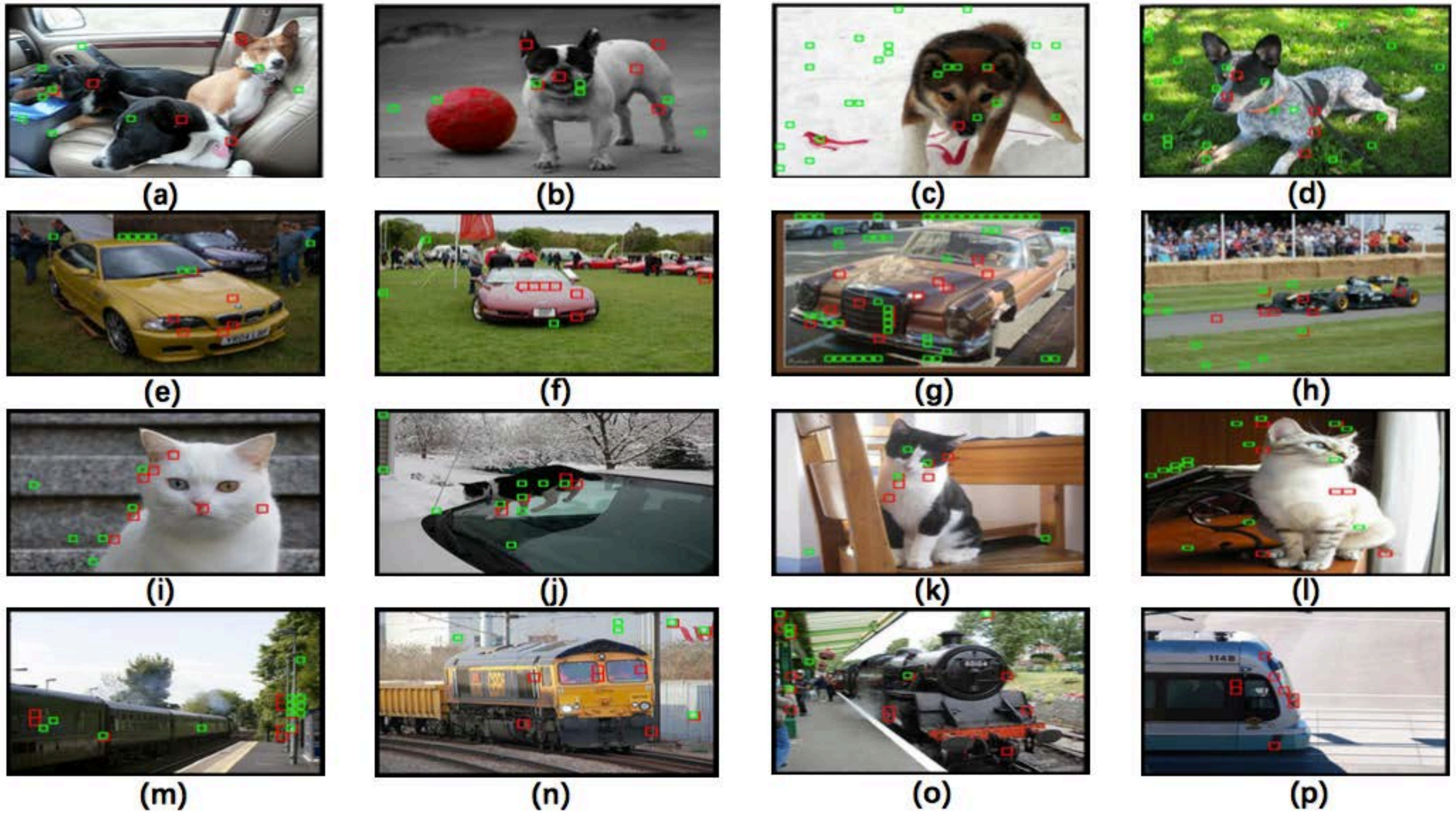
Experiment 1 – non-i.i.d. image classification

- 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



Experiment 2 – 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 $Age \in [50,100)$.

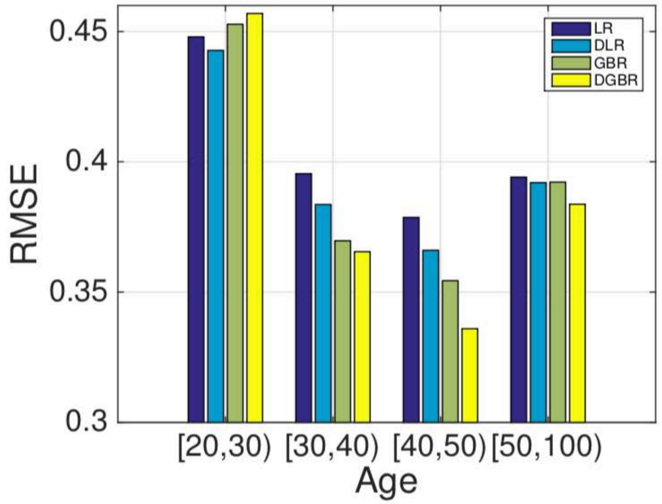


Fig. 15: Prediction across environments separated by age. The models are trained on dataset where uses' $Age \in [20, 30)$, but tested on various datasets with different users' age range.

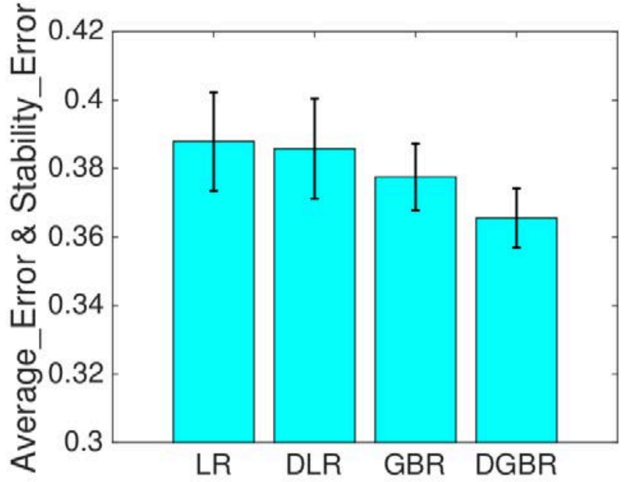


Fig. 16: $Average\_Error$ and $Stability\_Error$ of all algorithms across environments after fixing $P(Y)$ as the same with its value on global dataset.

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

- Previous logic:

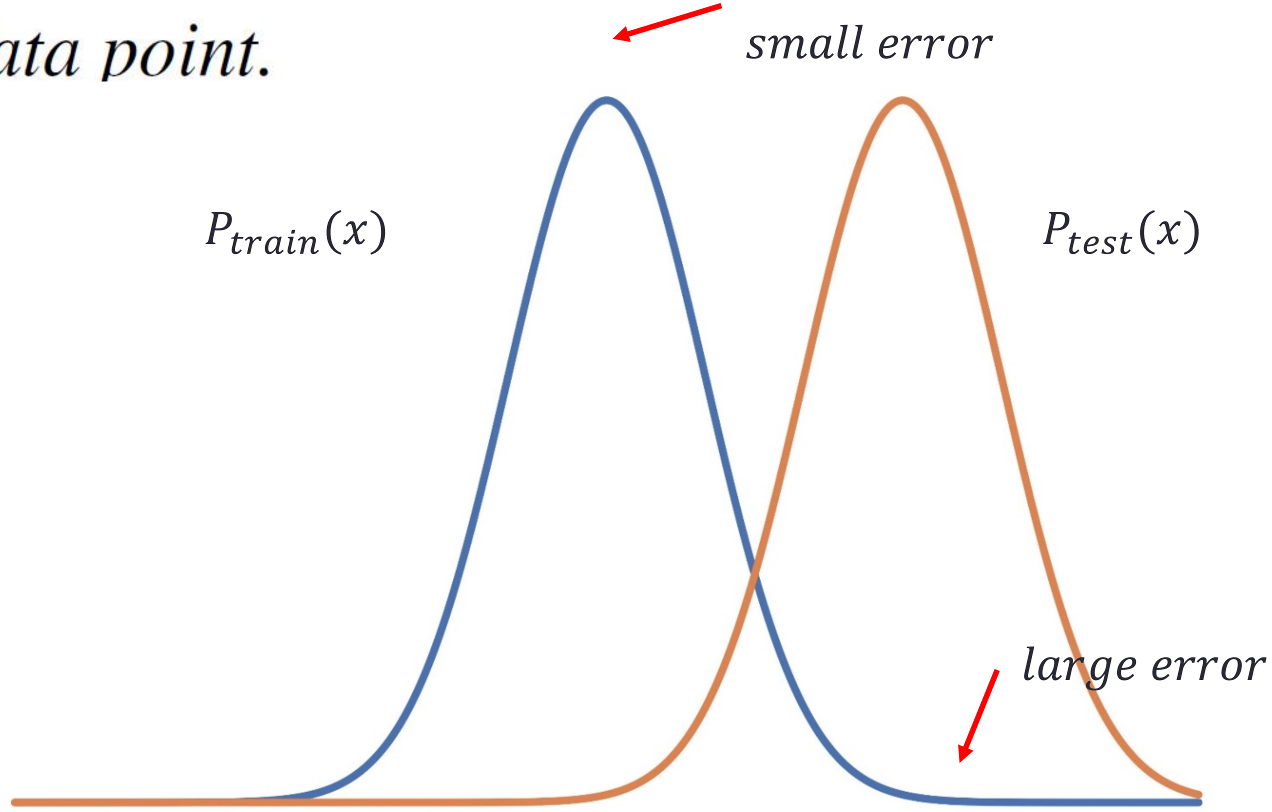
Sample Reweighting → Independent Variables → Causal Variable → Stable Prediction

- More direct logic:

Sample Reweighting → Independent Variables → Stable Prediction

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

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

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 $\|\hat{\beta} - \overline{\beta}\|_2 \leq \boxed{2(\delta/\gamma) + \delta}$, 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

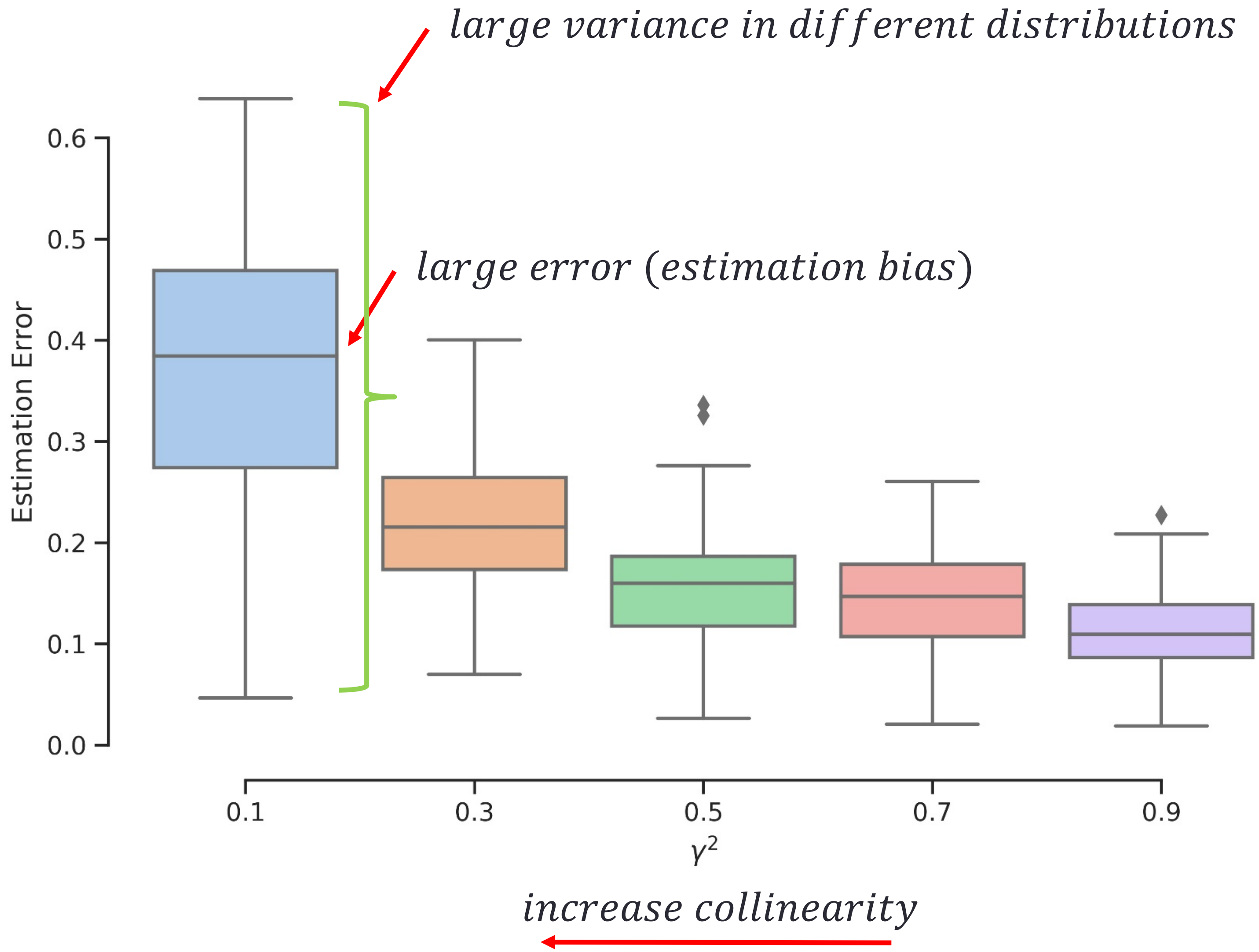
- Assume the design matrix X consists of two variables X_1, X_2, generated from a multivariate normal distribution:

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

- 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.)

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

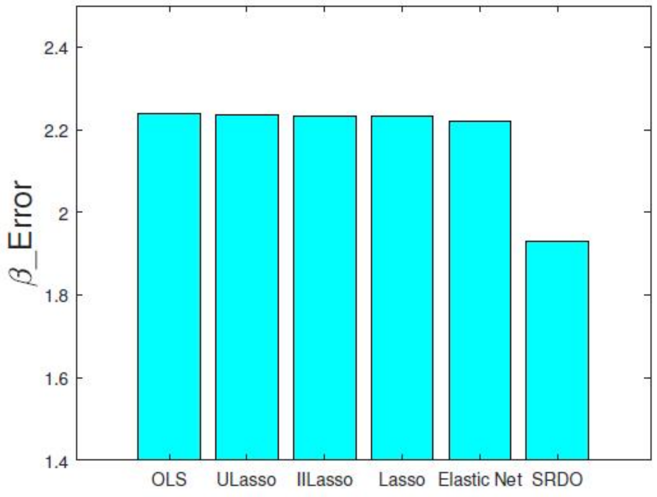
where i, j, k, r, s, t are drawn from 1 ... n at random

- 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 $w(x)$.

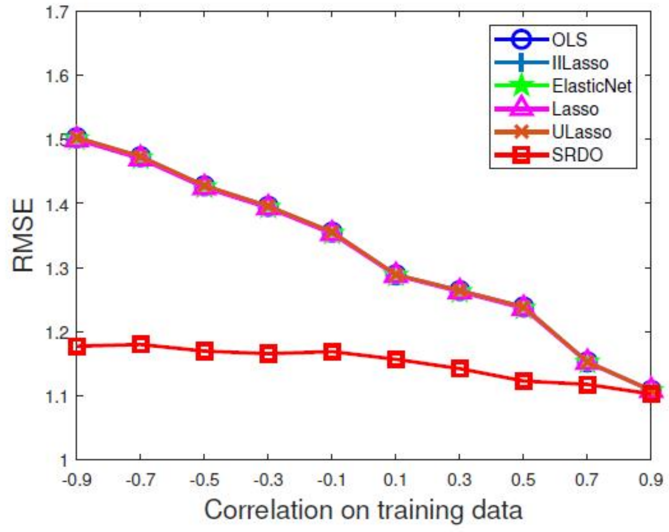
Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Experimental Results

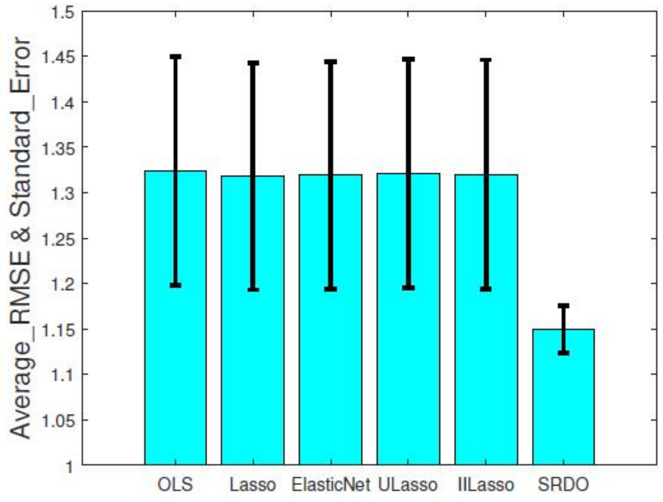
- Simulation Study



(a) Estimation error



(b) Prediction error over different test environments

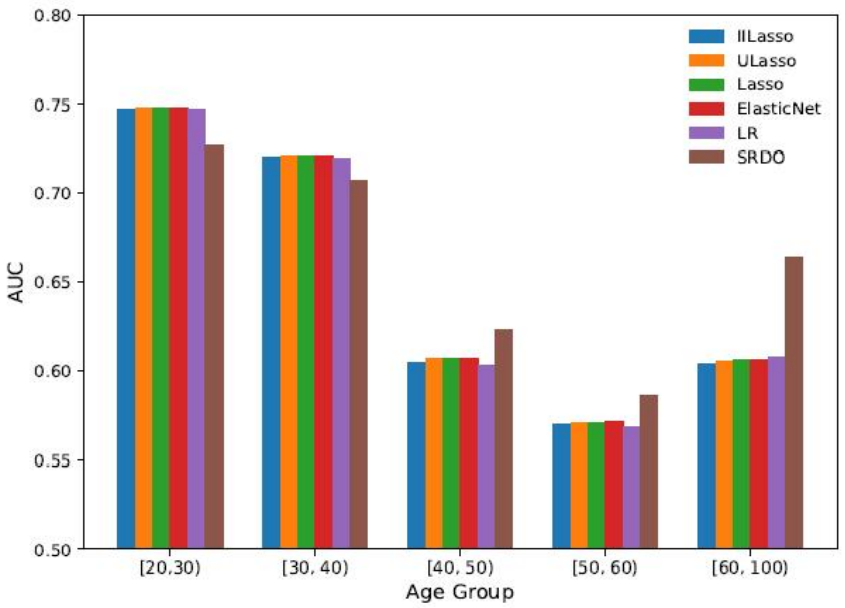


(c) Average prediction error&stability

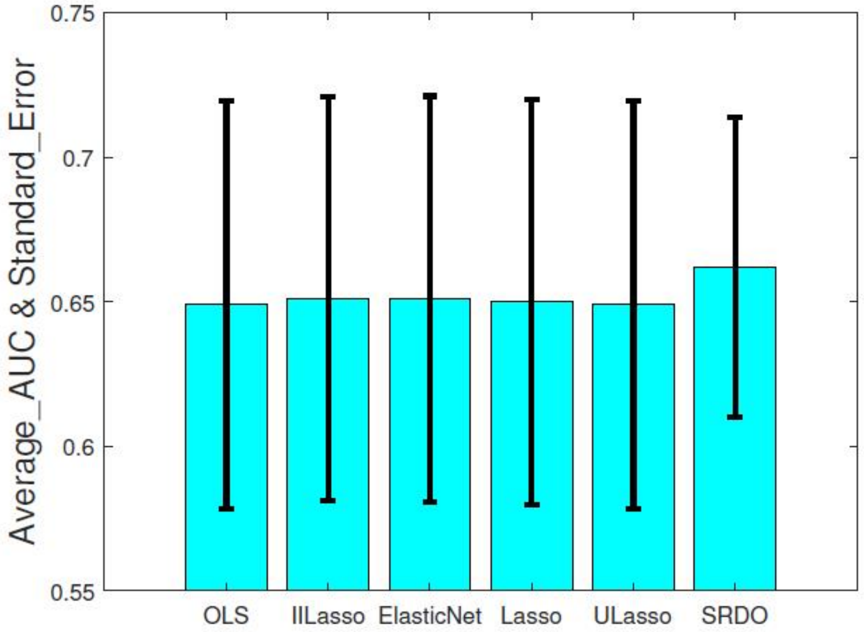
Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

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

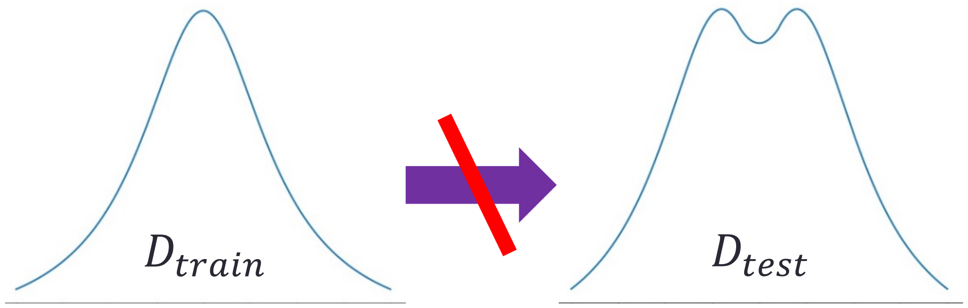
- Non I.I.D. Image Classification

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

- 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

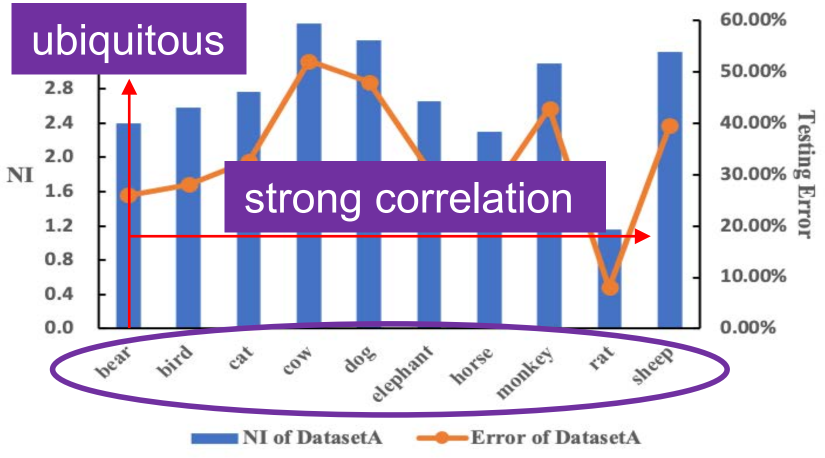
Definition 1 ***Non-I.I.D. Index (NI)*** *Given a feature extractor $g_\varphi(\cdot)$ and a class C,* ***the degree of distribution shift*** *between training data D^C_{train} and testing data D^C_{test} is defined as:*

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

Distribution shift

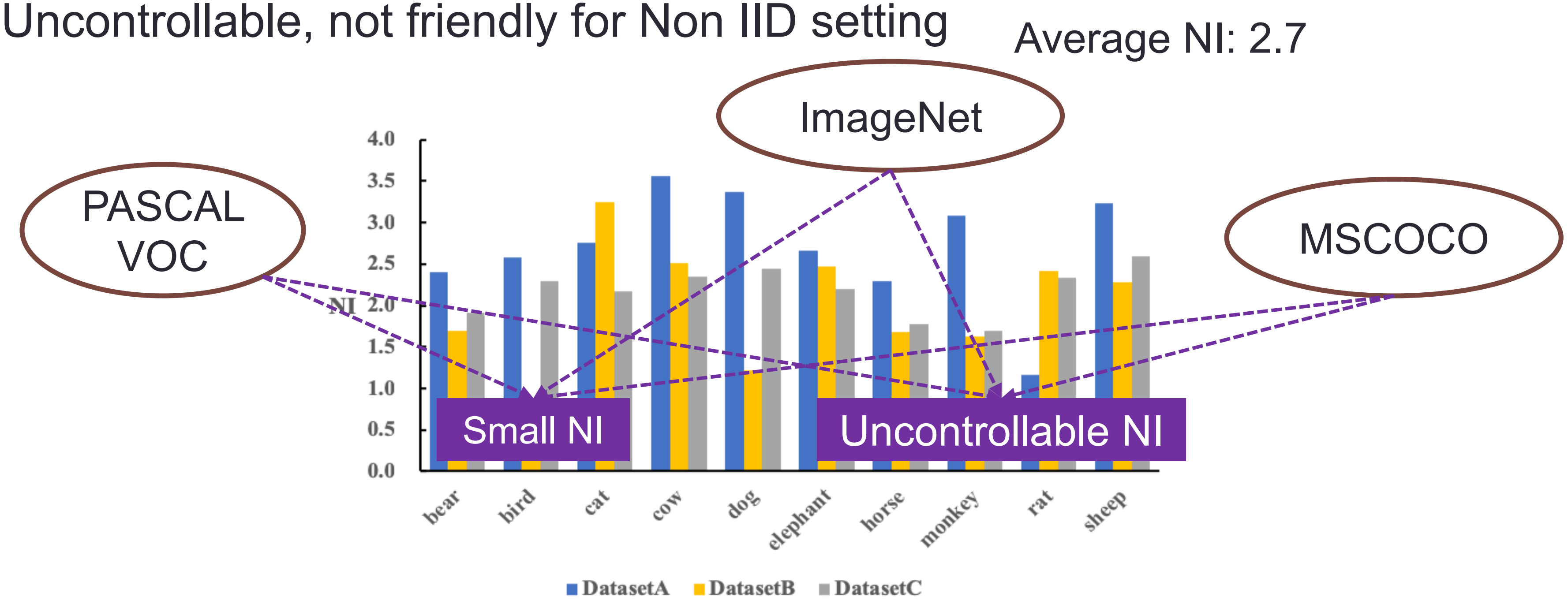
For normalization

- 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

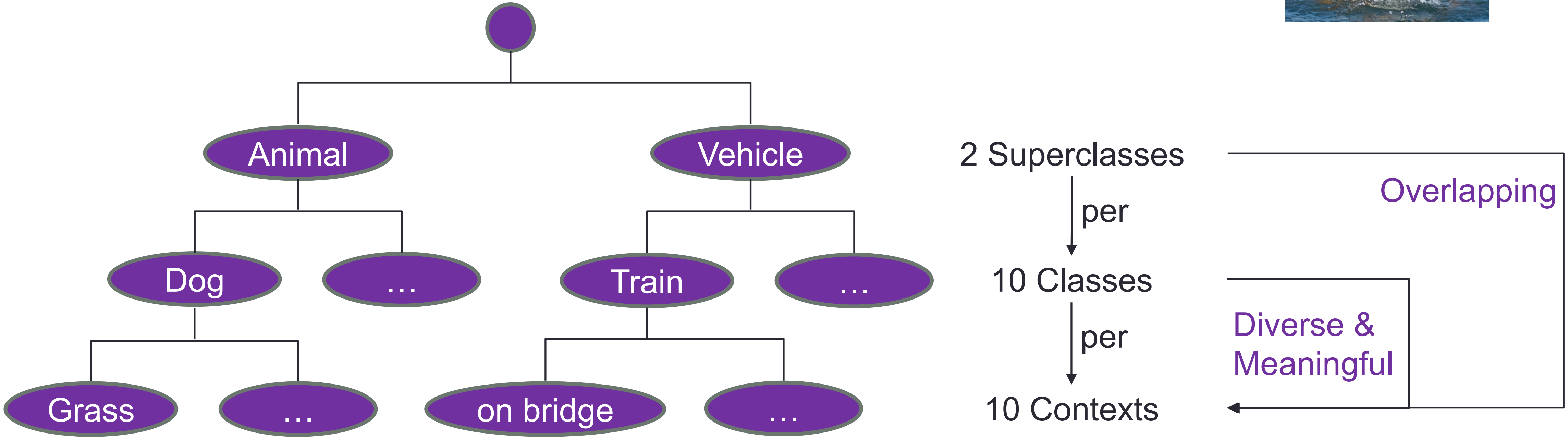
- DatasetA & DatasetB & DatasetC
  - NI is ubiquitous, but small on these datasets
  - NI is Uncontrollable, not friendly for Non IID setting



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

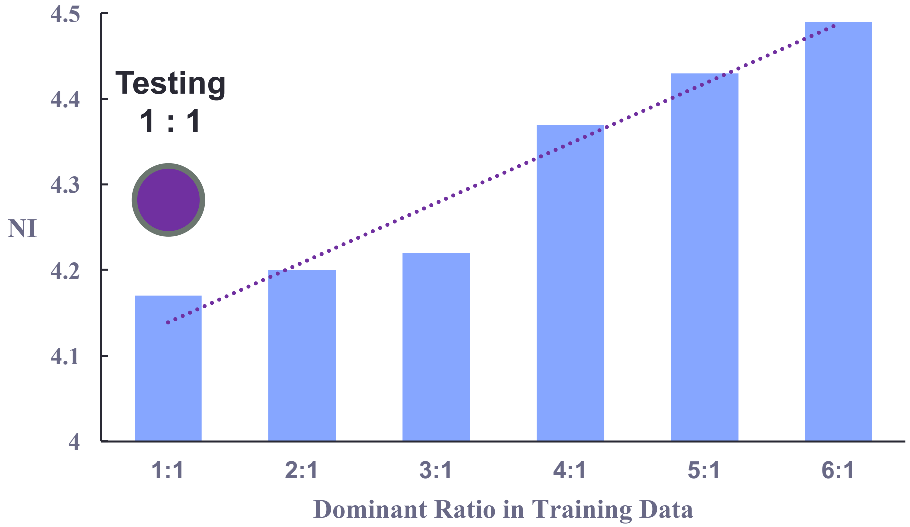
Proportional Bias

- 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

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

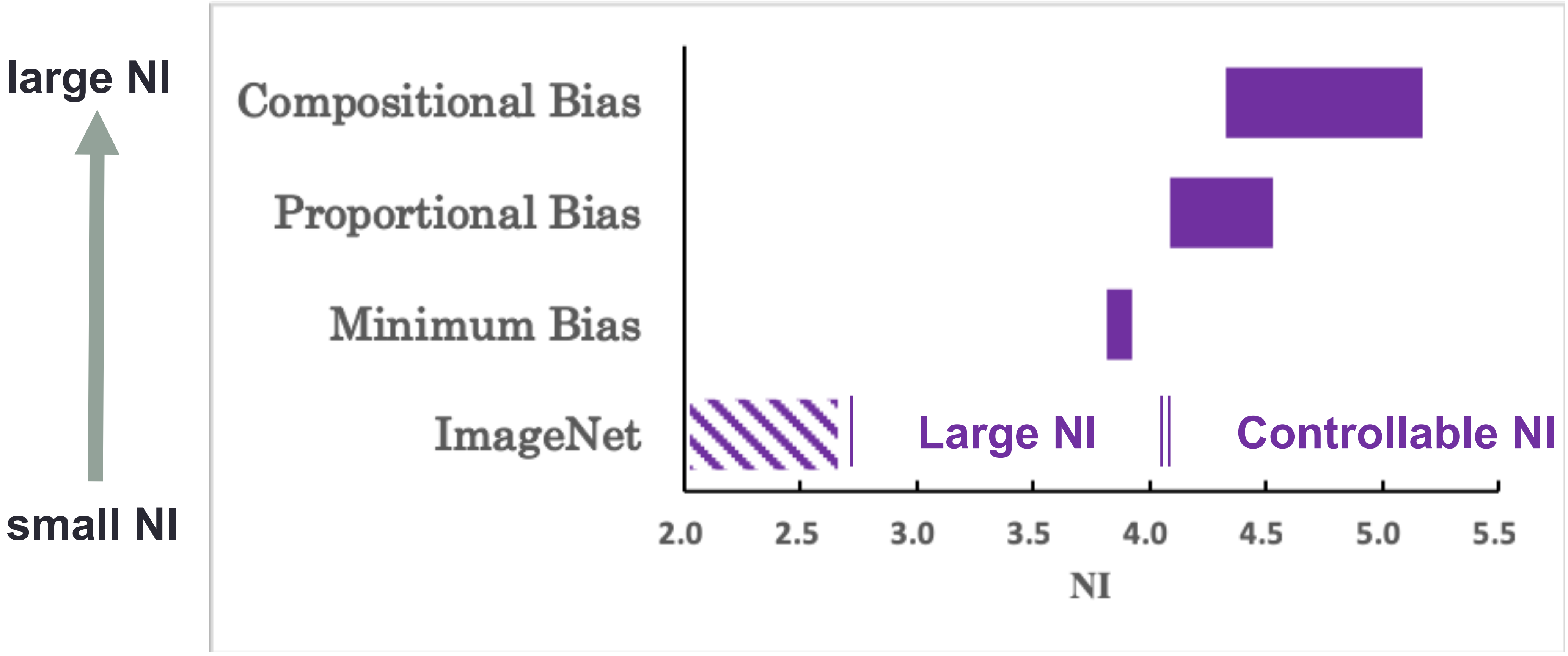
- Given a class, the observed contexts are different between training and testing data.

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

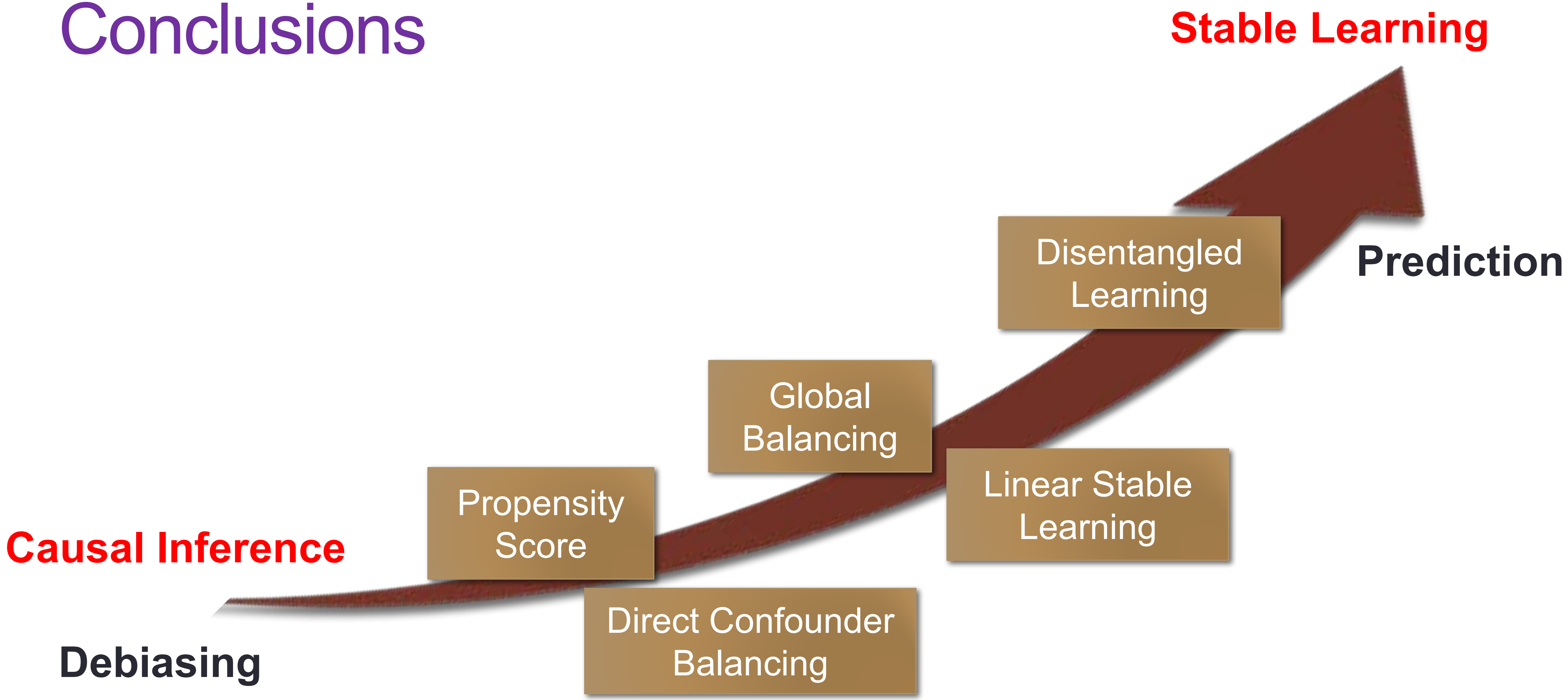
Outline

- 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.

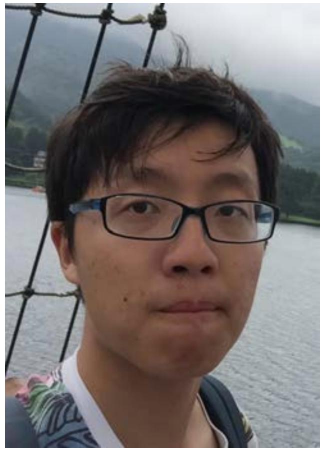
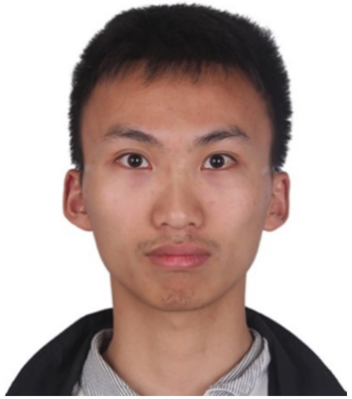
Reference

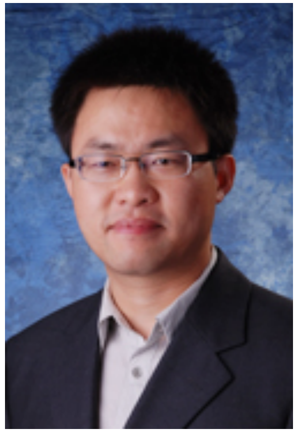
- Pearl J. Causality[M]. Cambridge university press, 2009.
- Austin P C. An introduction to propensity score methods for reducing the effects of confounding in observational studies[J]. Multivariate behavioral research, 2011, 46(3): 399-424.
- Johansson F, Shalit U, Sontag D. Learning representations 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




Kun Kuang
Tsinghua U

Zheyan Shen
Tsinghua U

Hao Zou
Tsinghua U

Yue He
Tsinghua U

Bo Li
Tsinghua U

Susan Athey
Stanford U

Thanks!

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