因果启发的学习和推理

崔鹏，沈哲言
清华大学

AI is stepping into risk-sensitive areas

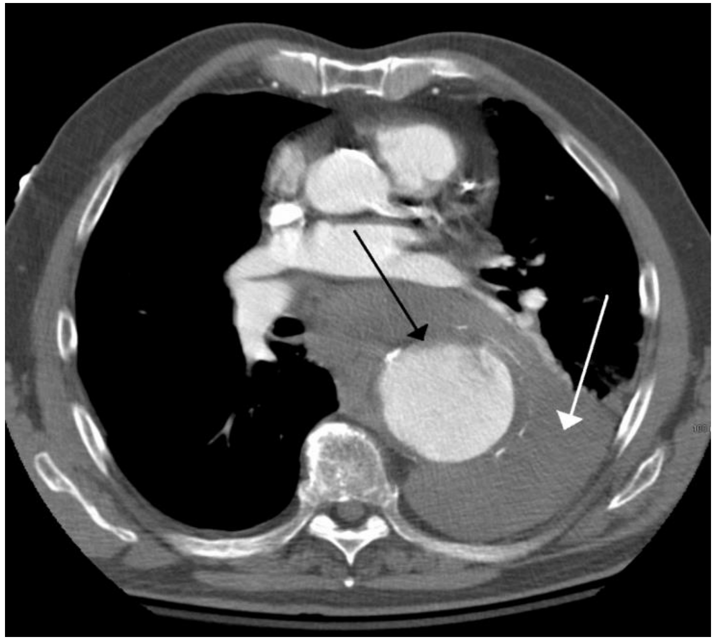
Human

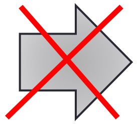
Shifting from ***Performance Driven*** to ***Risk Sensitive***

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

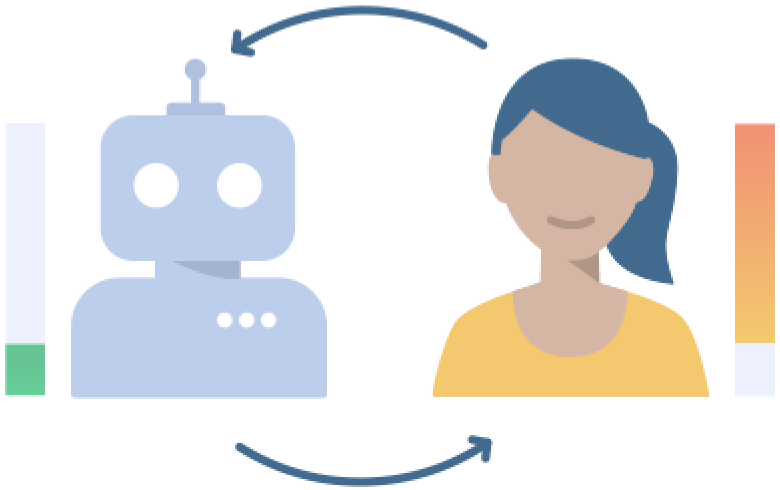
Most machine learning models are black-box models

Unexplainable



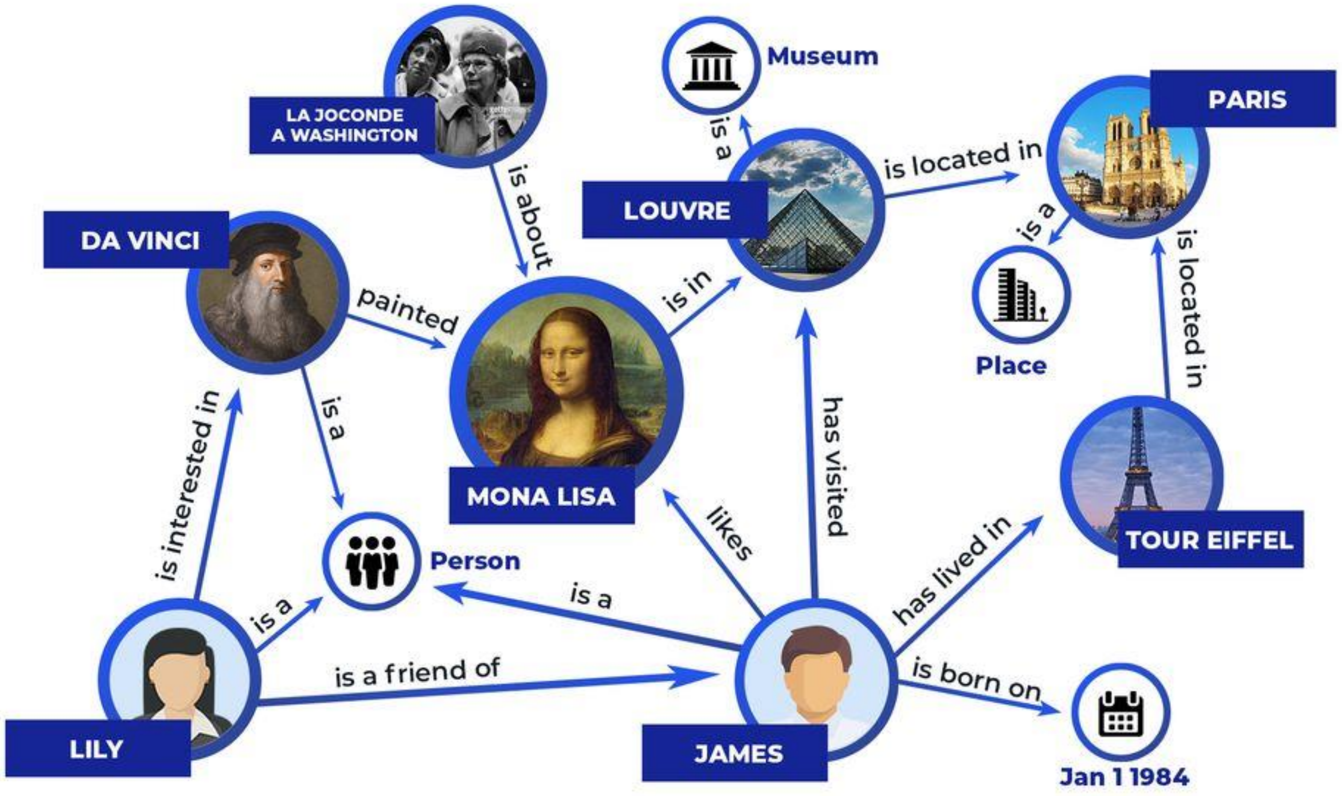


Human in the loop



Health Military Finance Industry

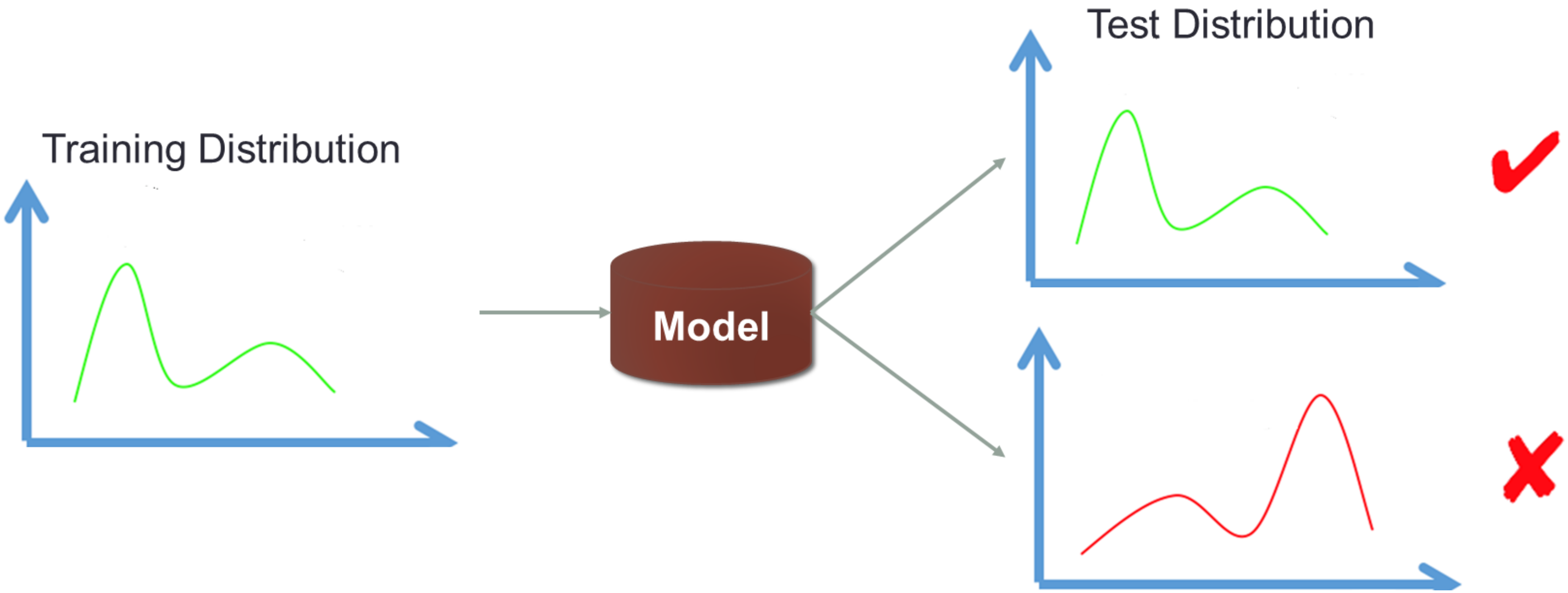
Problems of today's ML - *Explainability*



Embedding-based methods for knowledge graph acquisition are unexplainable

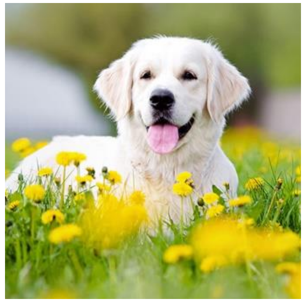
Problems of today's ML - ***Stability***

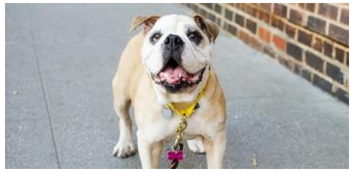
Most ML methods are developed under I.I.D hypothesis



OOD Generalization Problem

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




Yes

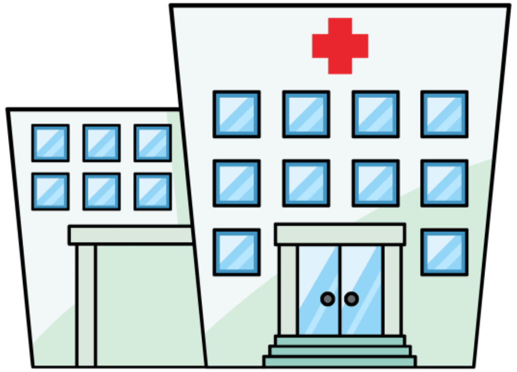
Maybe

No

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

- Cancer survival rate prediction

Training Data

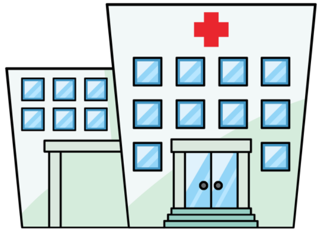


City Hospital

Higher income, higher survival rate.

Predictive Model

Testing Data



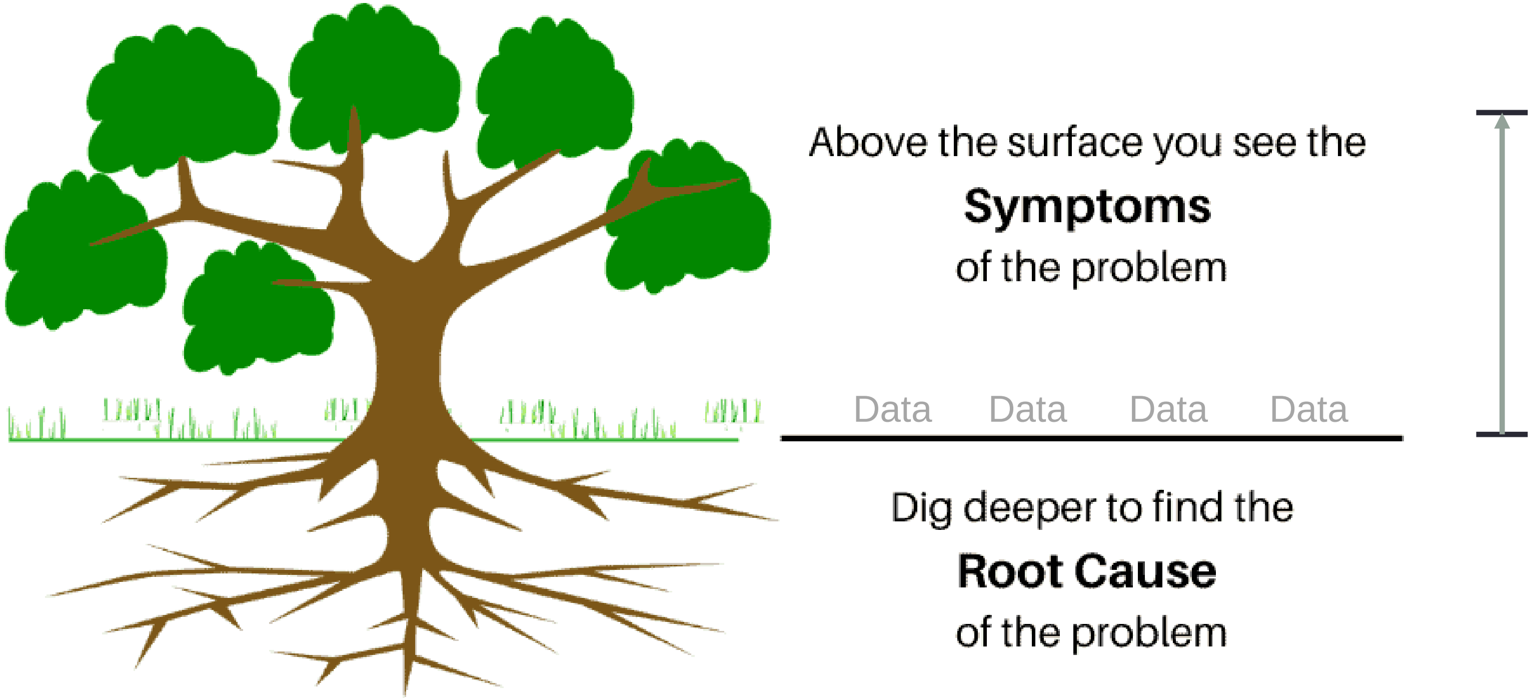
City Hospital

University Hospital

Survival rate is not so correlated with income.

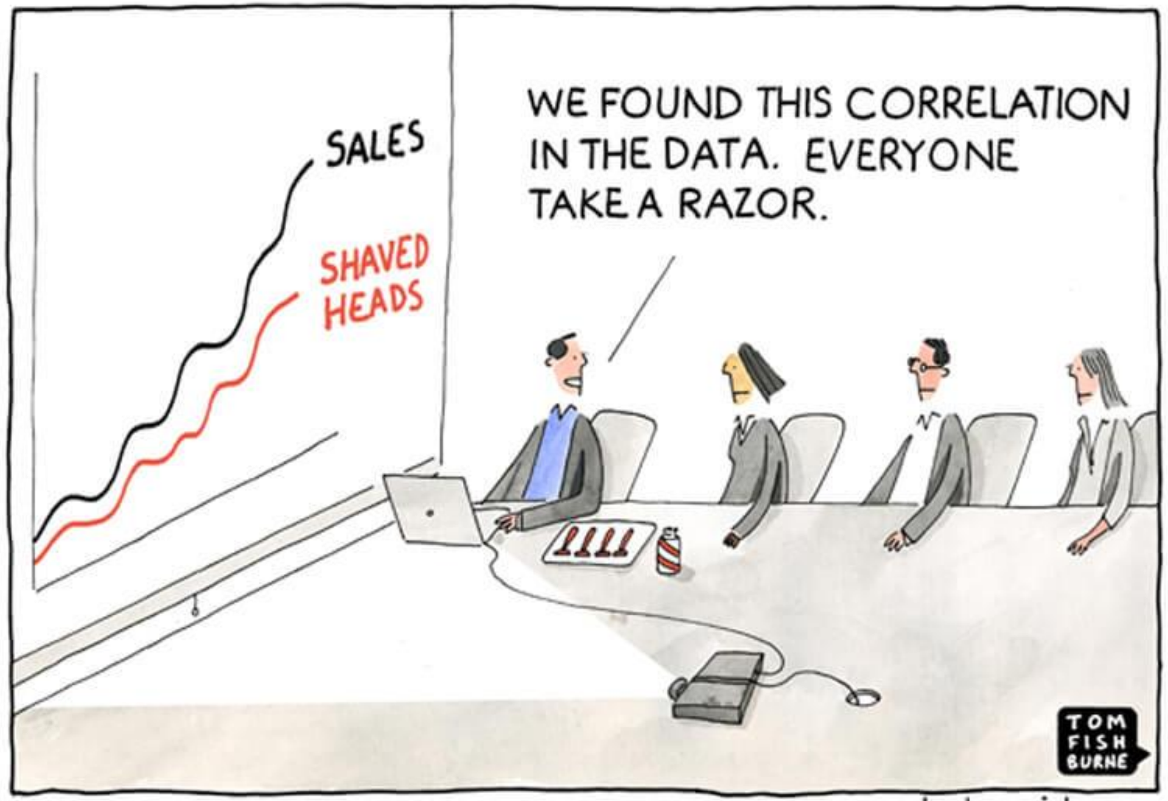
Problems of today's ML - *Fairness*

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



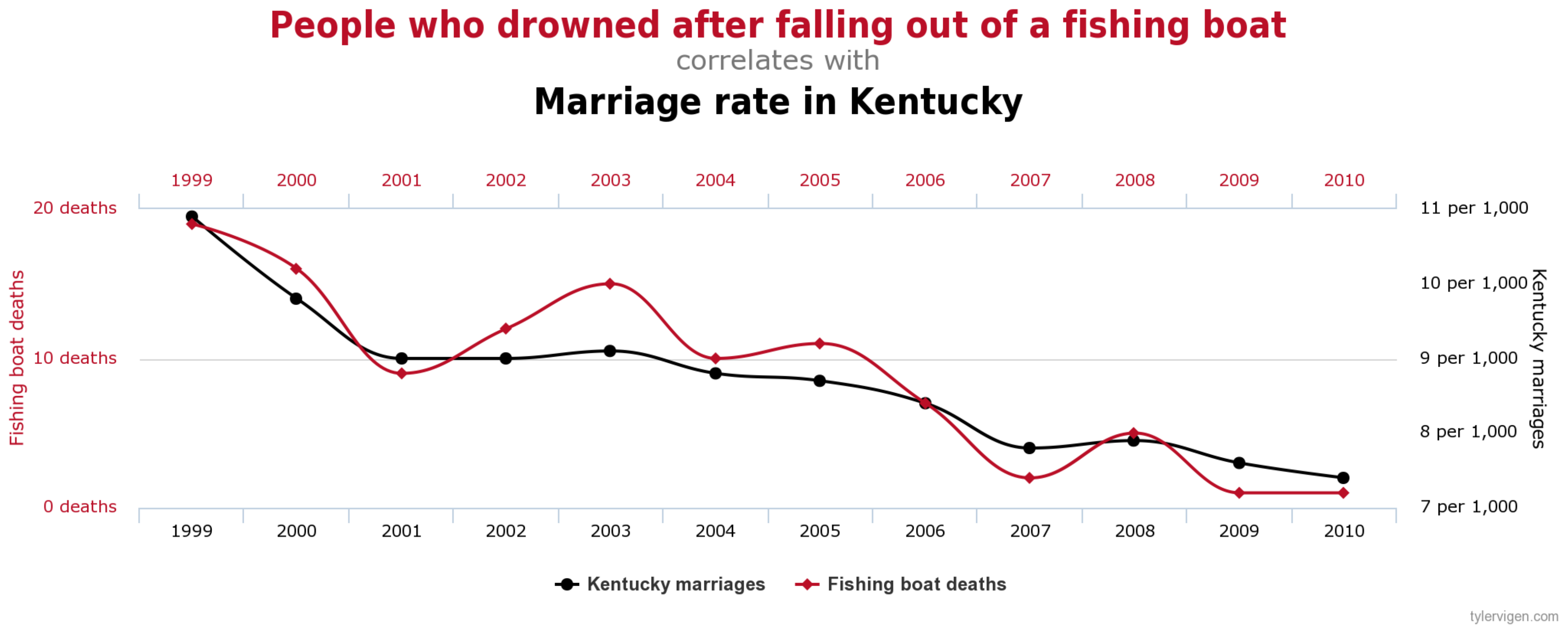
A plausible reason: ***Correlation***

Correlation is the very basics of machine learning.

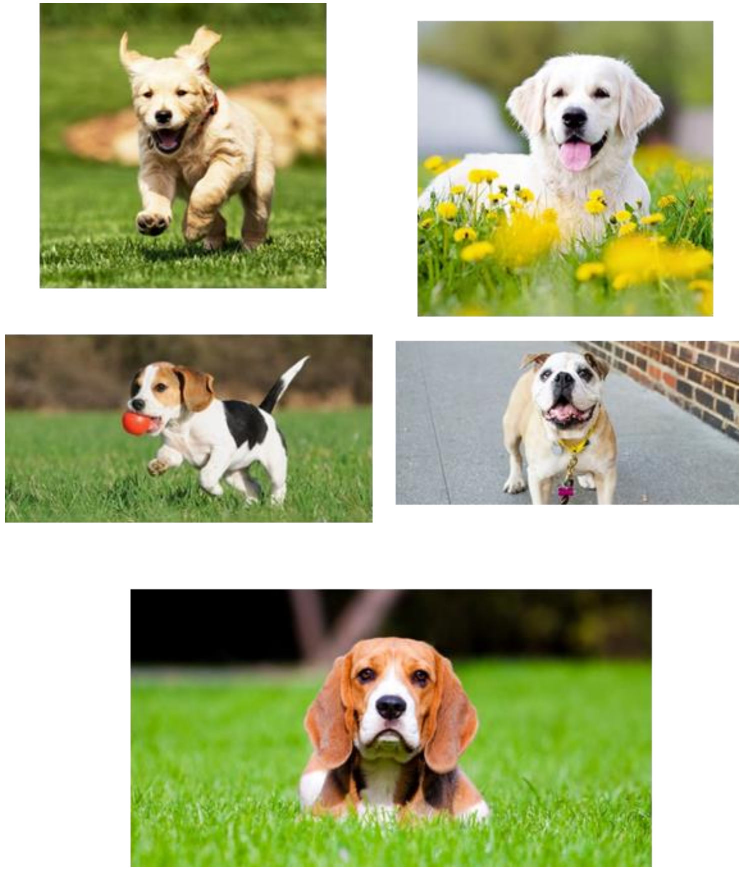


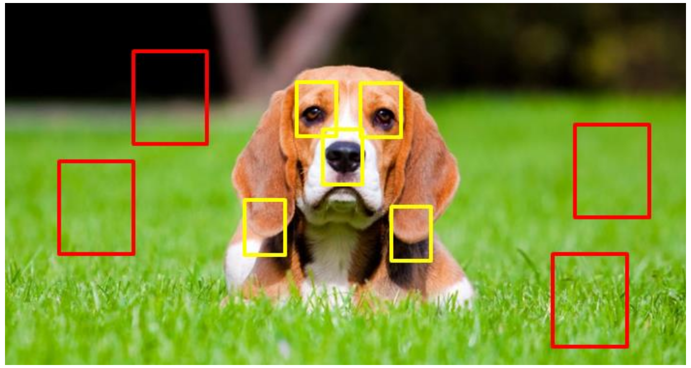
© marketoonist.com

Correlation is not explainable



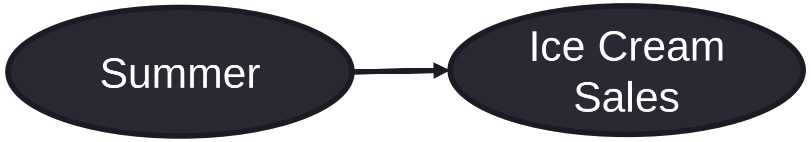
Correlation is ‘*unstable*’



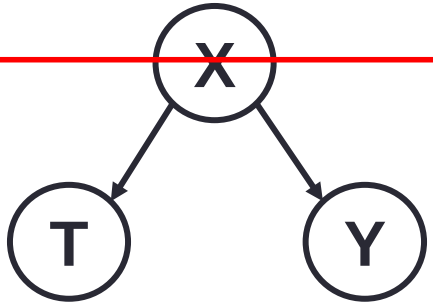


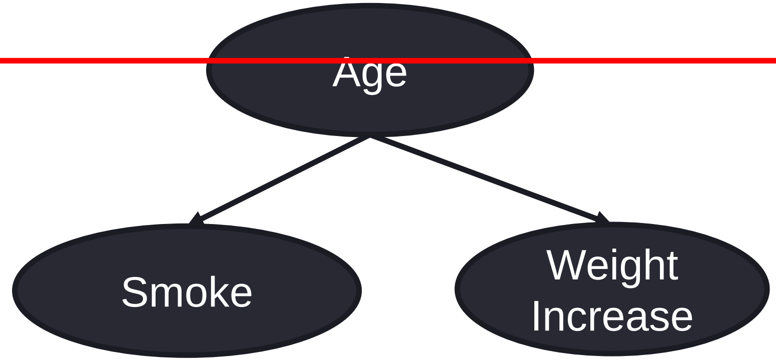
It’s not the fault of ***correlation***, but the way we use it

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

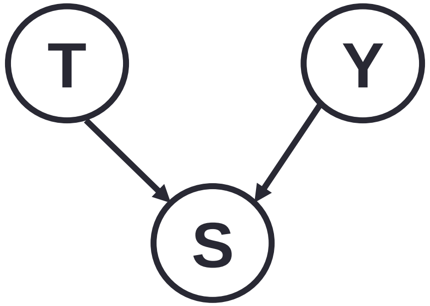


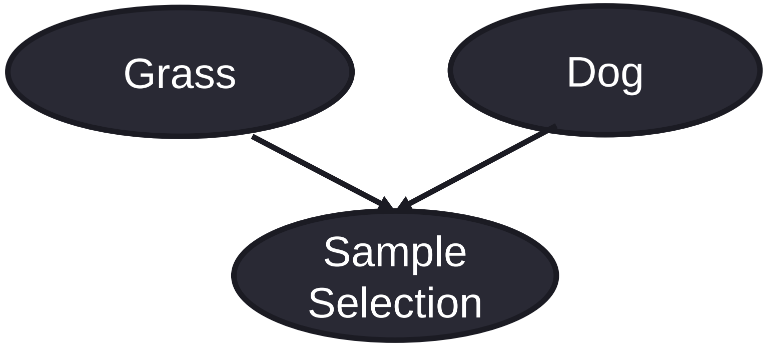
  - Confounding
    - Ignoring X
    - Spurious Correlation





  - Sample Selection Bias
    - Conditional on S
    - Spurious Correlation



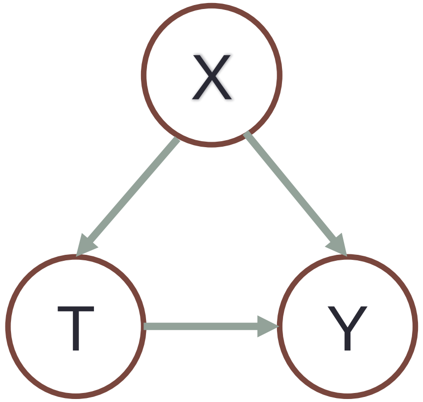


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



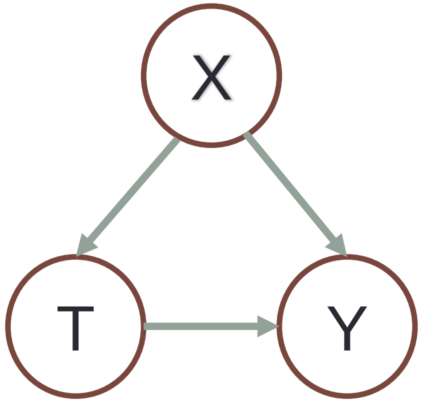
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the "interventionist" interpretation of causality.

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

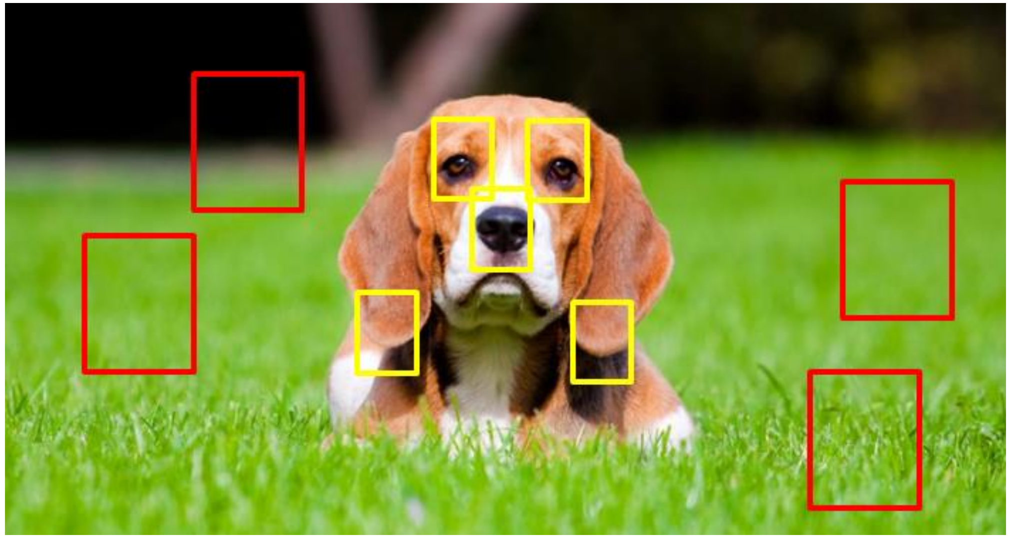
The *benefits* of bringing causality into learning

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***

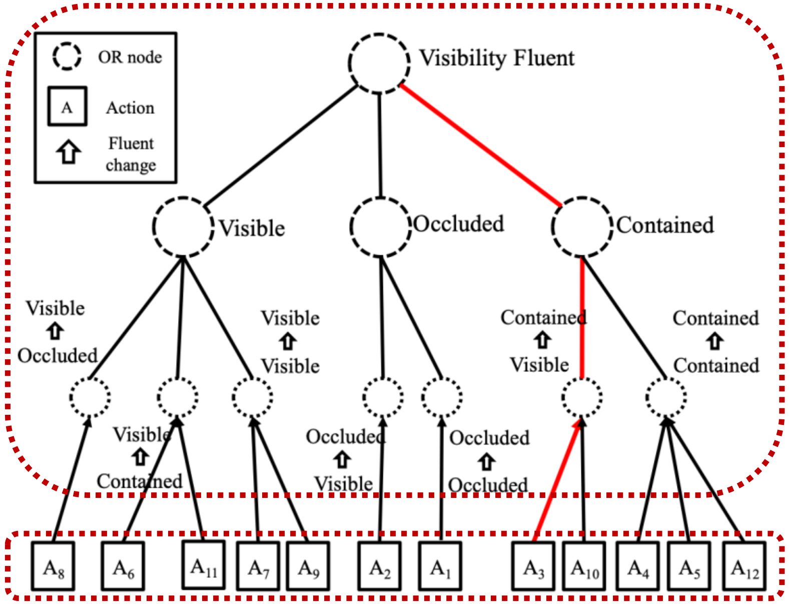
Explainability with Causality

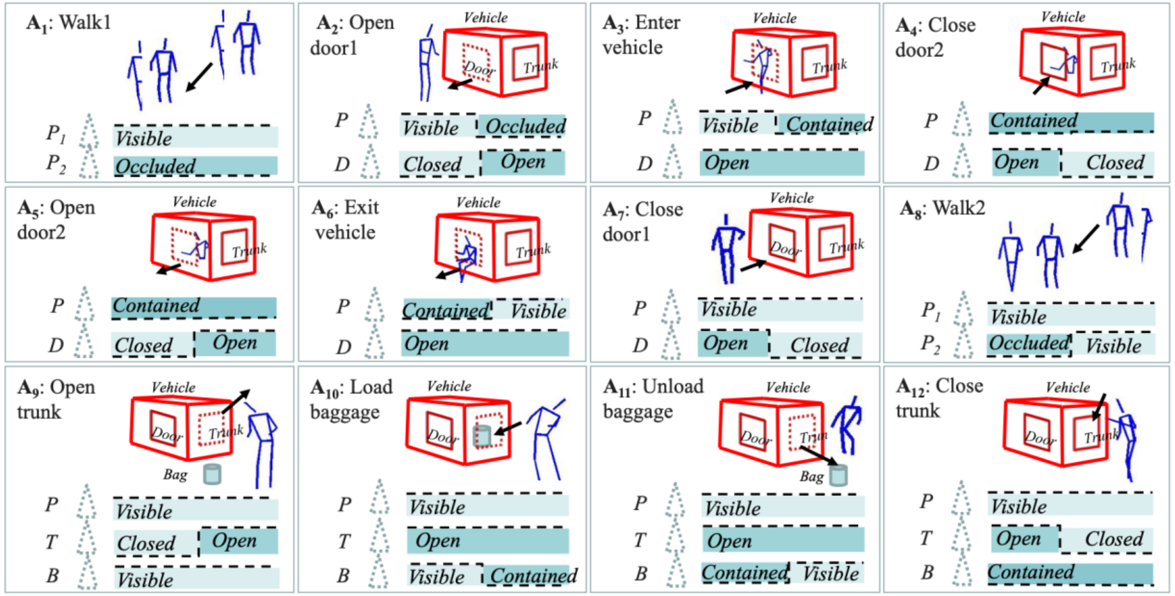
Application --- visibility fluent reasoning

- introduce a Causal And-Or Graph (C-AOG) to represent the causal-effect relations between an object’s visibility fluent and its actions

Visibility fluent

Atomic actions





Causal And-Or Graph

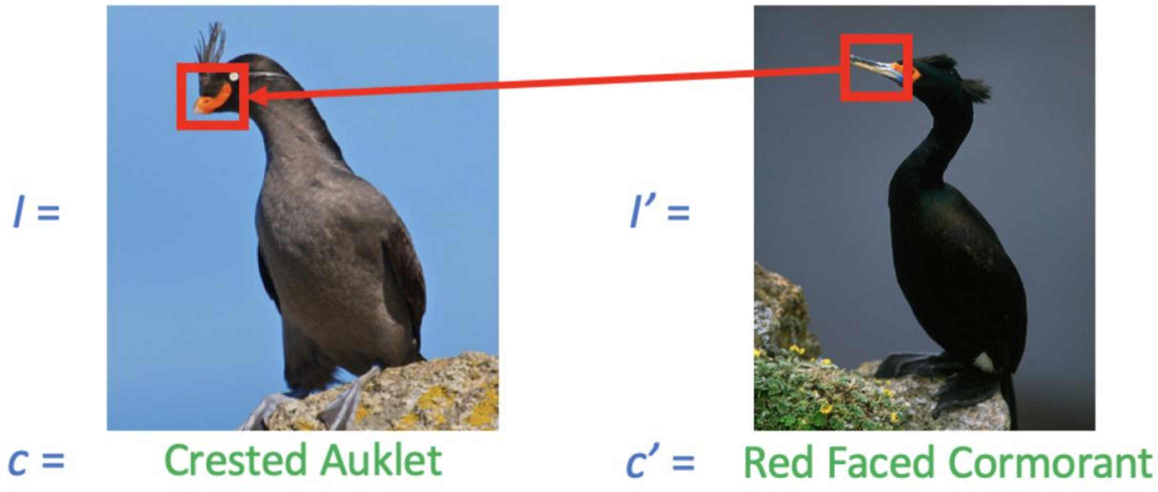
Atomic actions

Xu, Yuanlu, et al. "A causal and-or graph model for visibility fluent reasoning in tracking interacting objects." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Explainability with Causality

Application --- counterfactual visual explanations

- A causal explanation: why the example image was classified as class c instead of c'?
  - If the bird on the left had a similar beak to that on the right, then the system would have output the right class.

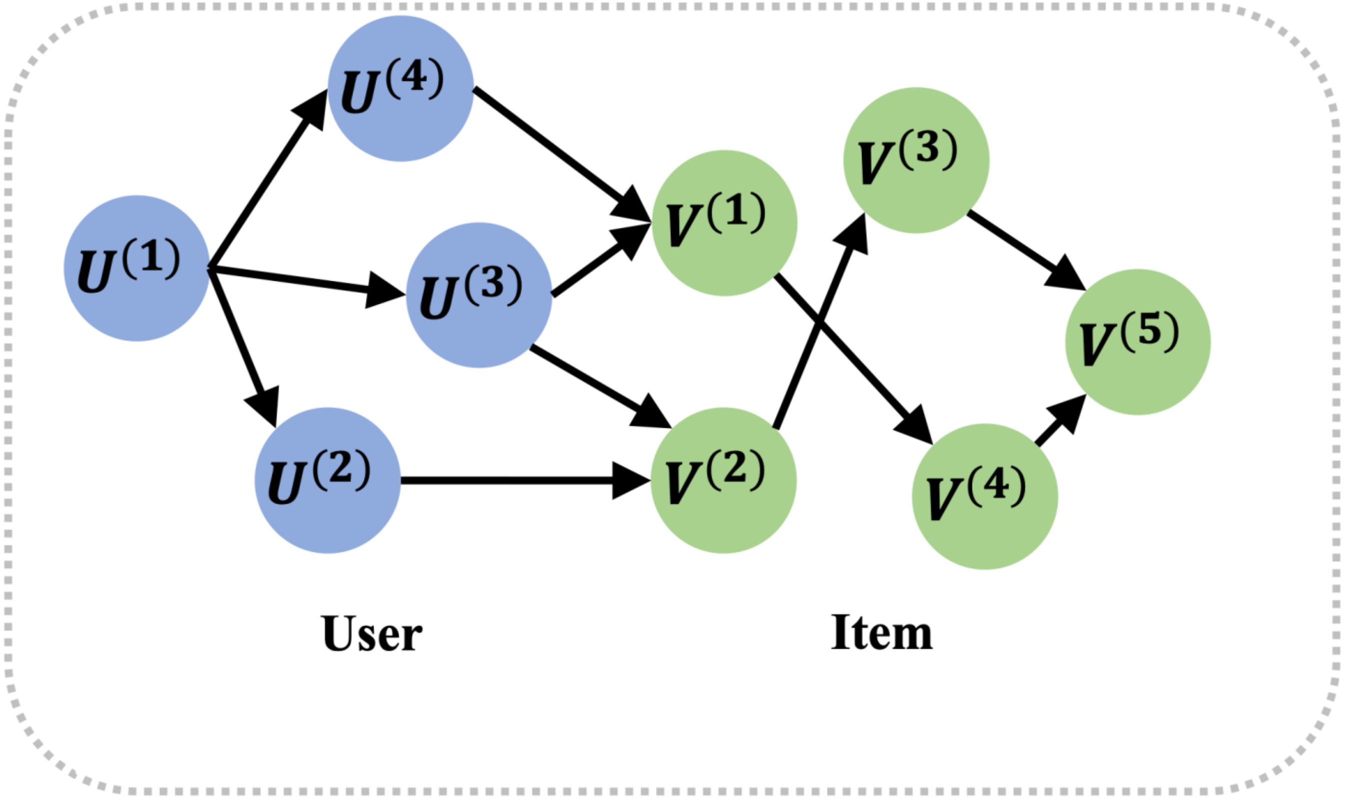


Goyal, Yash, et al. "Counterfactual visual explanations." International Conference on Machine Learning. PMLR, 2019.

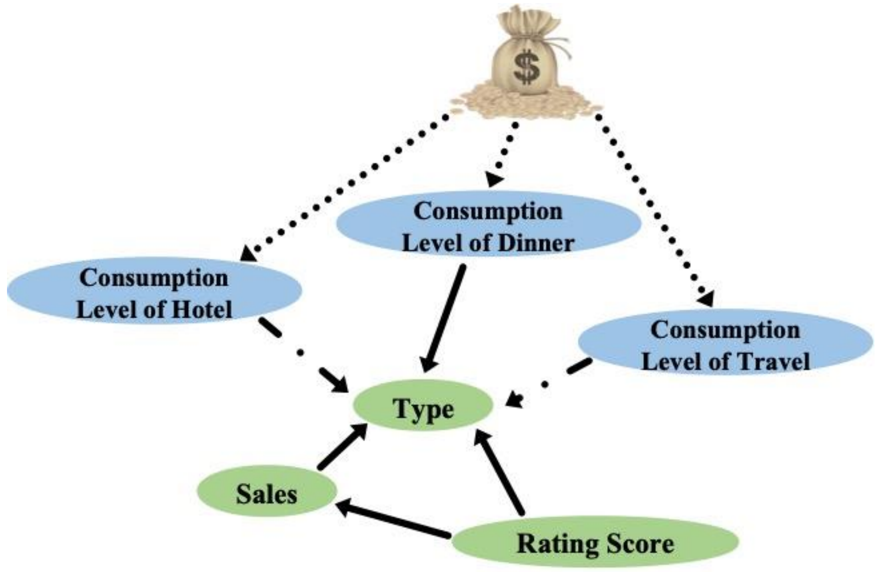
Explainability with Causality

Application --- causal recommendation

Caual structure among user features and item features



Example



He et al. "Collaborative Causal Filtering for Out-of-Distribution Recommendation." *Under review.*

Explainability and OOD

$$\text{OOD} \longleftarrow \text{Causality} \longrightarrow \text{Explainability}$$

- Explainability would be a side product when pursuing OOD with causality

Knowledge Graph and Causality

- Representation and Construction
  - 知 (fact) 识 (causality)
  - 格物致知 -》 格数致知
- Inference
  - Know Why -> Know How -> Know What
  - What Known -> What Unknown *BIAS!* *(the target of **stable learning**)*
- Utility
  - Prediction (we are here!)
  - Inference
  - Decision

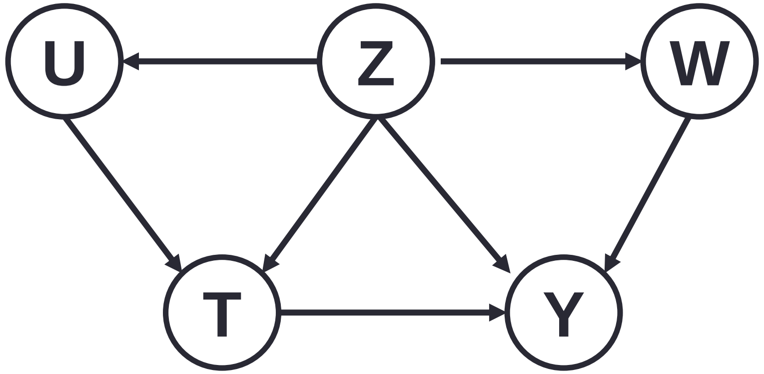
Outline

- Brief introduction to causal inference
- Stable learning and its development
- Positioning stable learning in OOD generalization
- Benchmark and dataset

Paradigms - Structural Causal Model

A graphical model to describe the causal mechanisms of a system

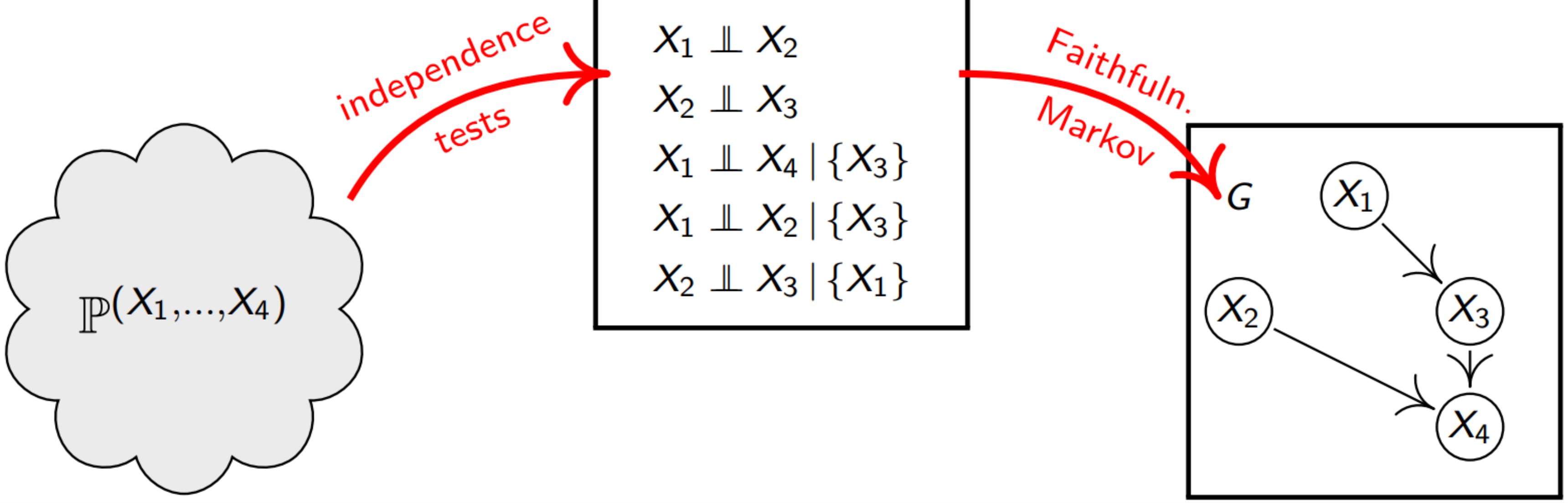
- Causal Identification with back door criterion
- Causal Estimation with do calculus



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.

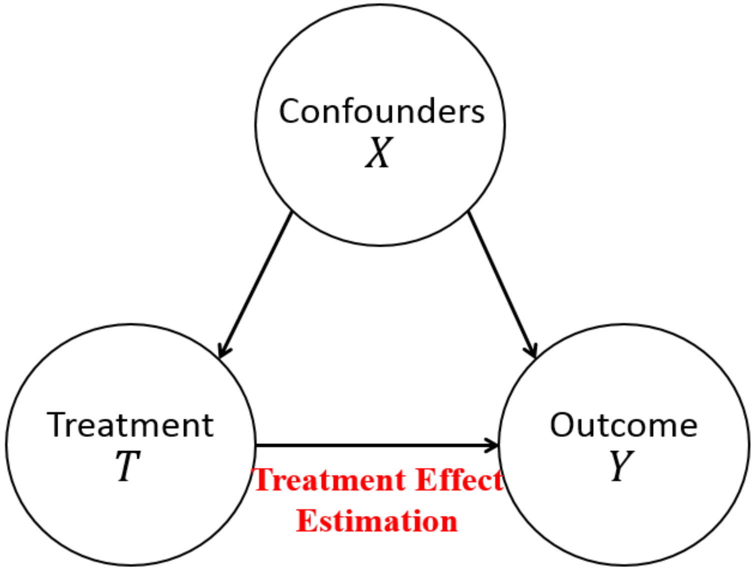
Intractability

d	Number of DAGs with d nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505

Peters et al. *Elements of Causal Inference*. 2017

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

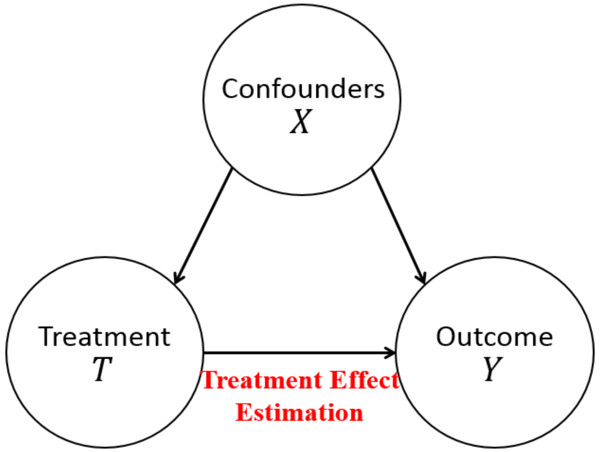


More like a ***discriminative*** way to estimate treatment's partial effect on outcome.

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

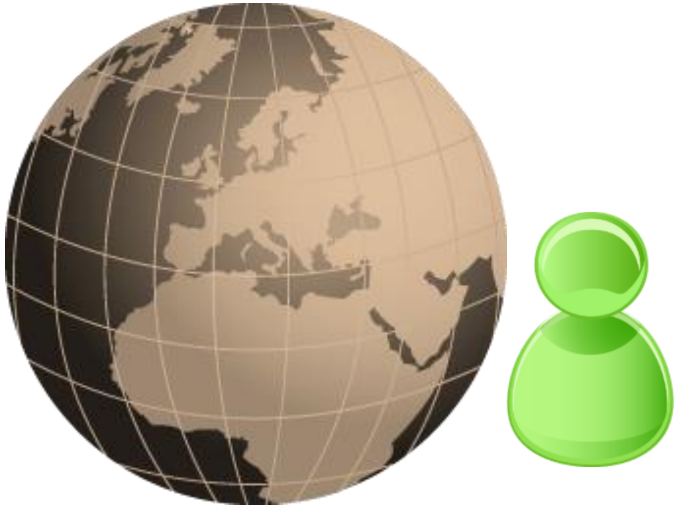
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	?
P2	0	?	0.6
P3	1	0.3	?
P4	0	?	0.1
P5	1	0.5	?
P6	0	?	0.5
P7	0	?	0.1

- 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!

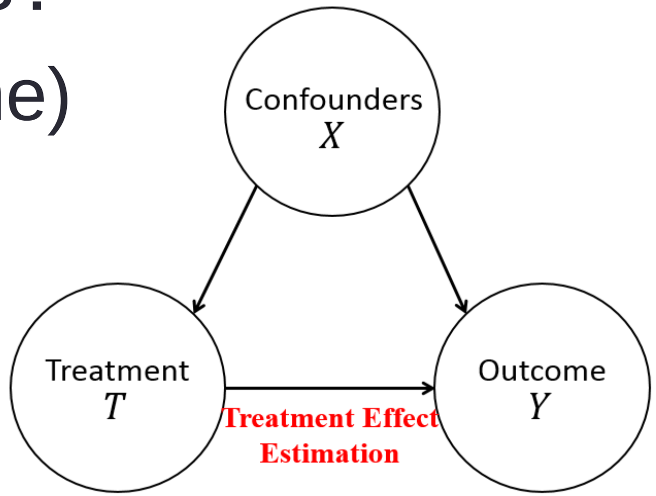
- Drawbacks … nts:
  - Cost
  - Unethical
  - Unrealistic

Causal Inference with Observational Data

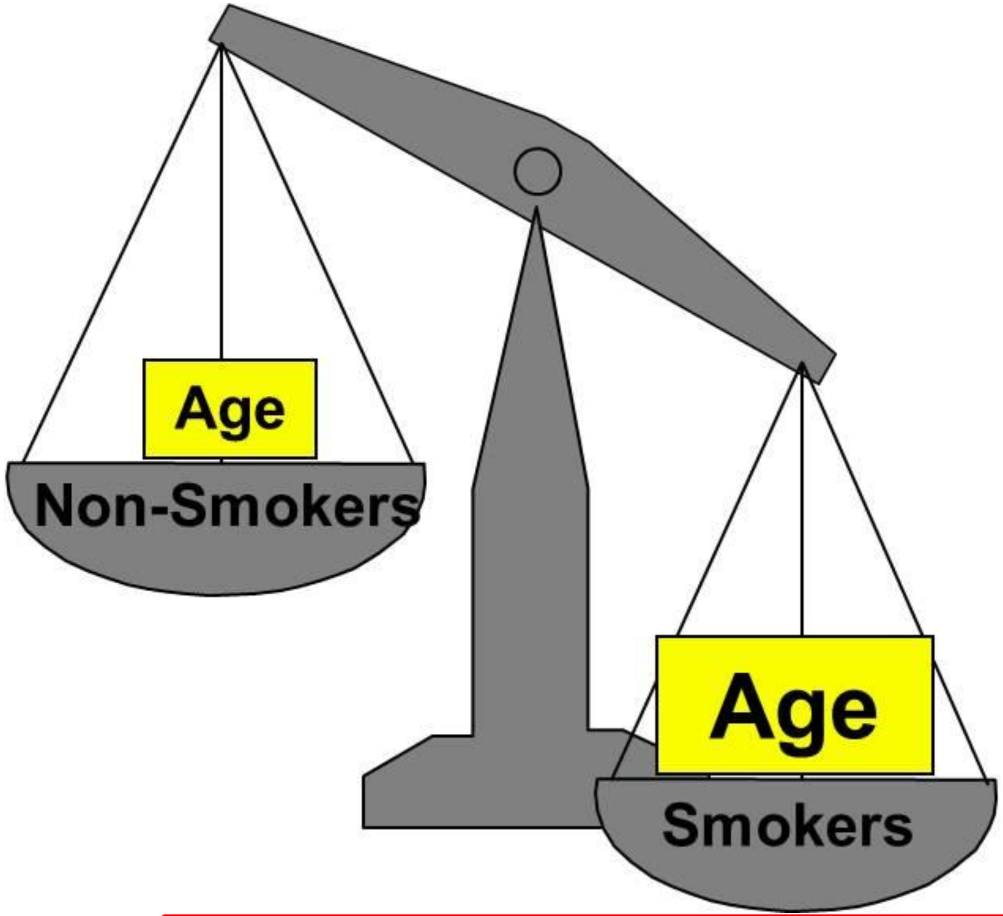
- Counterfactual Problem:

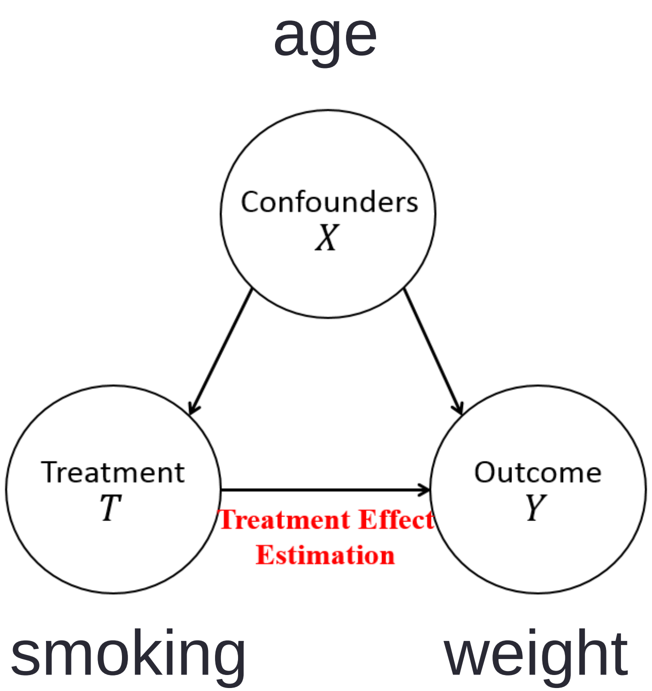
$$Y(T=1) \quad \text{or} \quad Y(T=0)$$

- 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



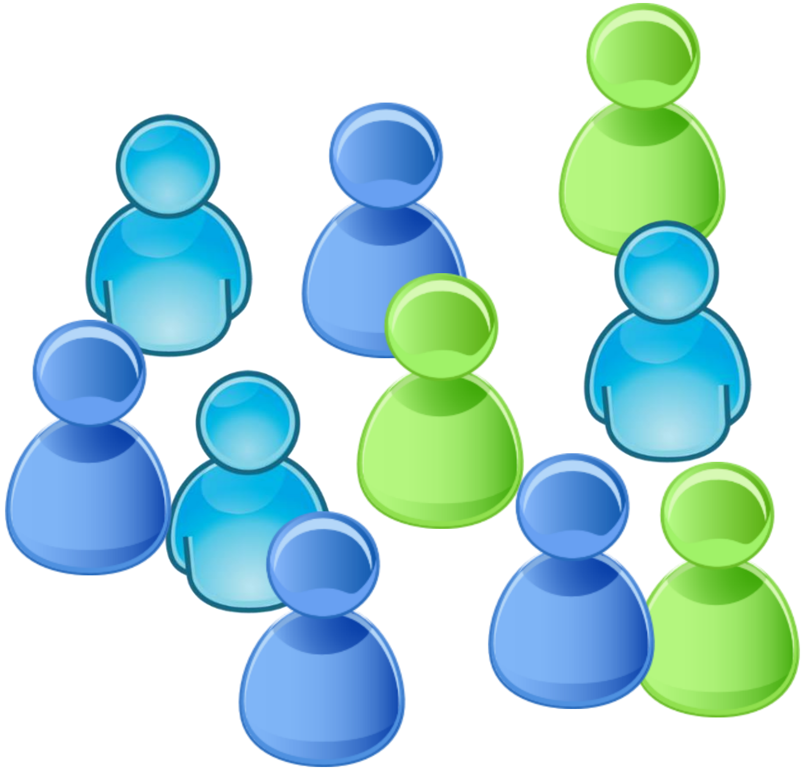


Balancing Confounders' Distribution

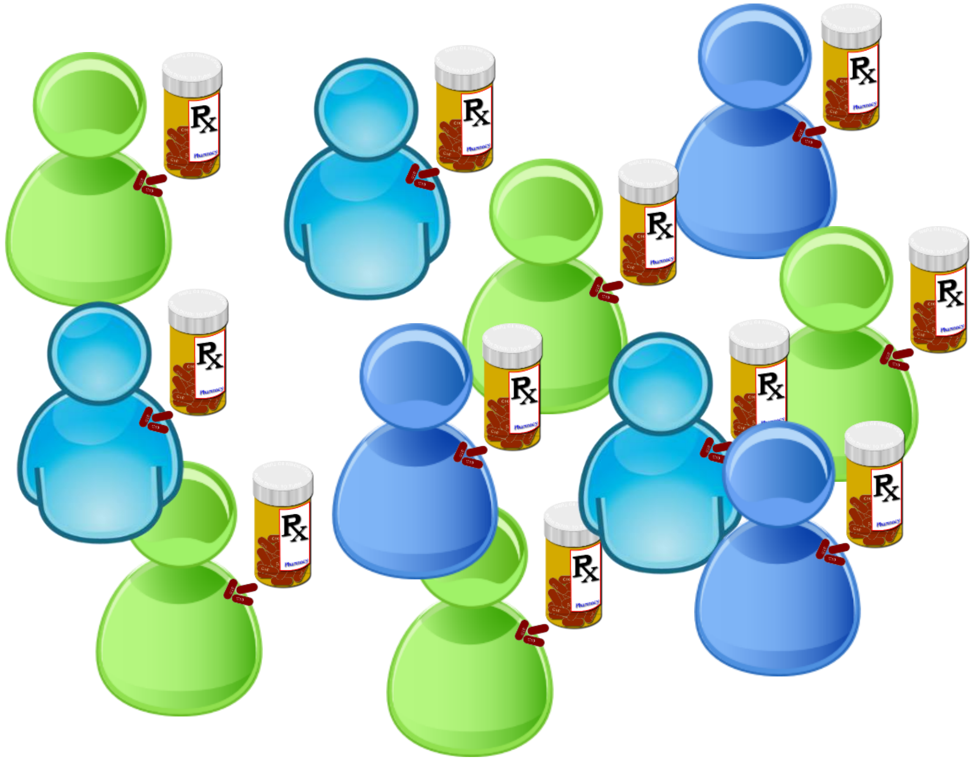
Methods for Causal Inference

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

Matching

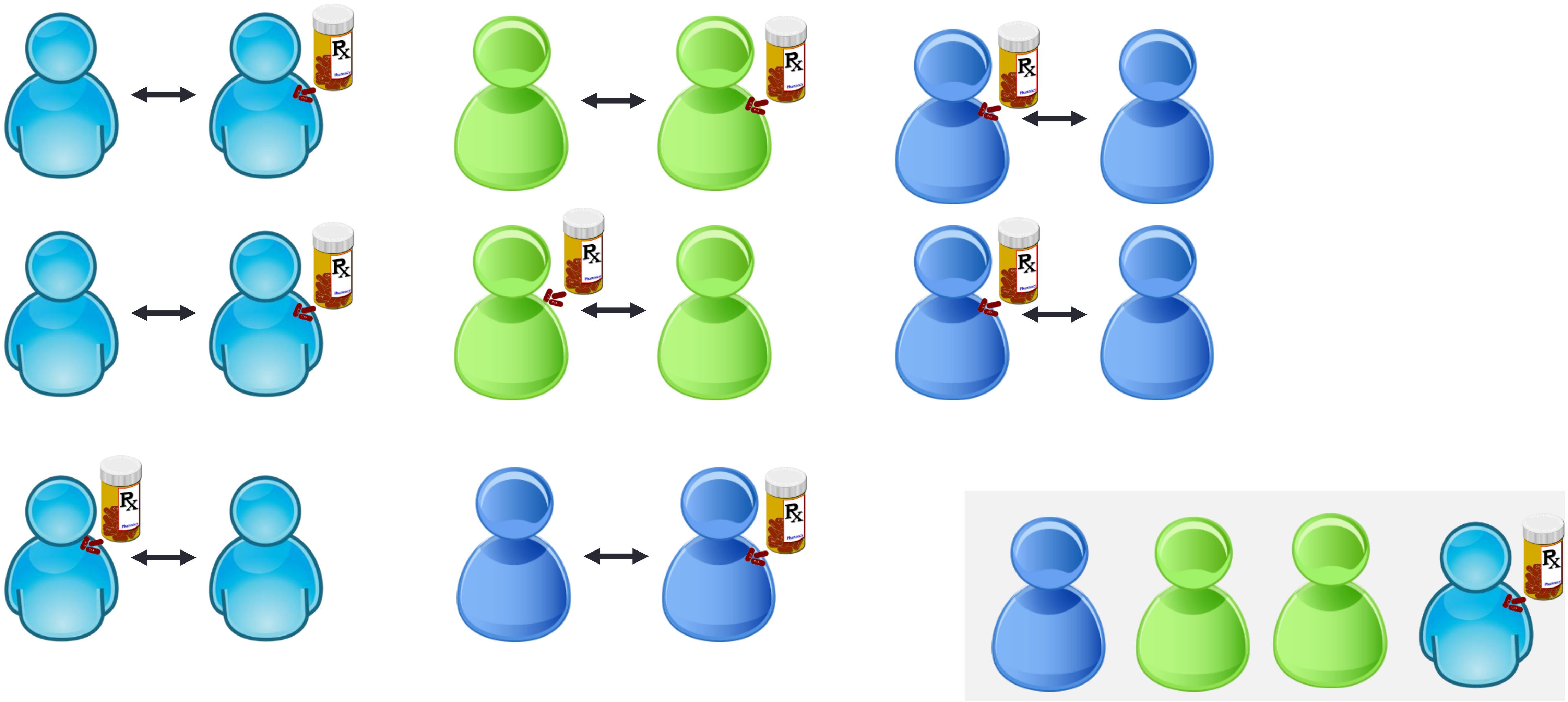


$T = 0$



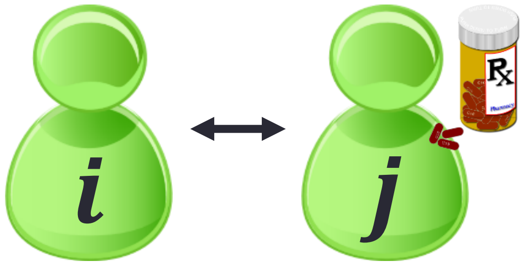
$T = 1$

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

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

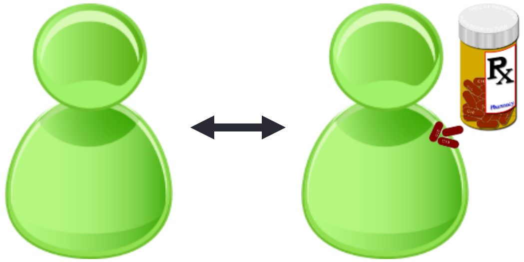
- Then, Donald Rubin shows that the propensity score is sufficient to control or summarize the information of confounders

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

- 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
- But this is a ‘hard’ solution.

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.

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 *learning*?

Outline

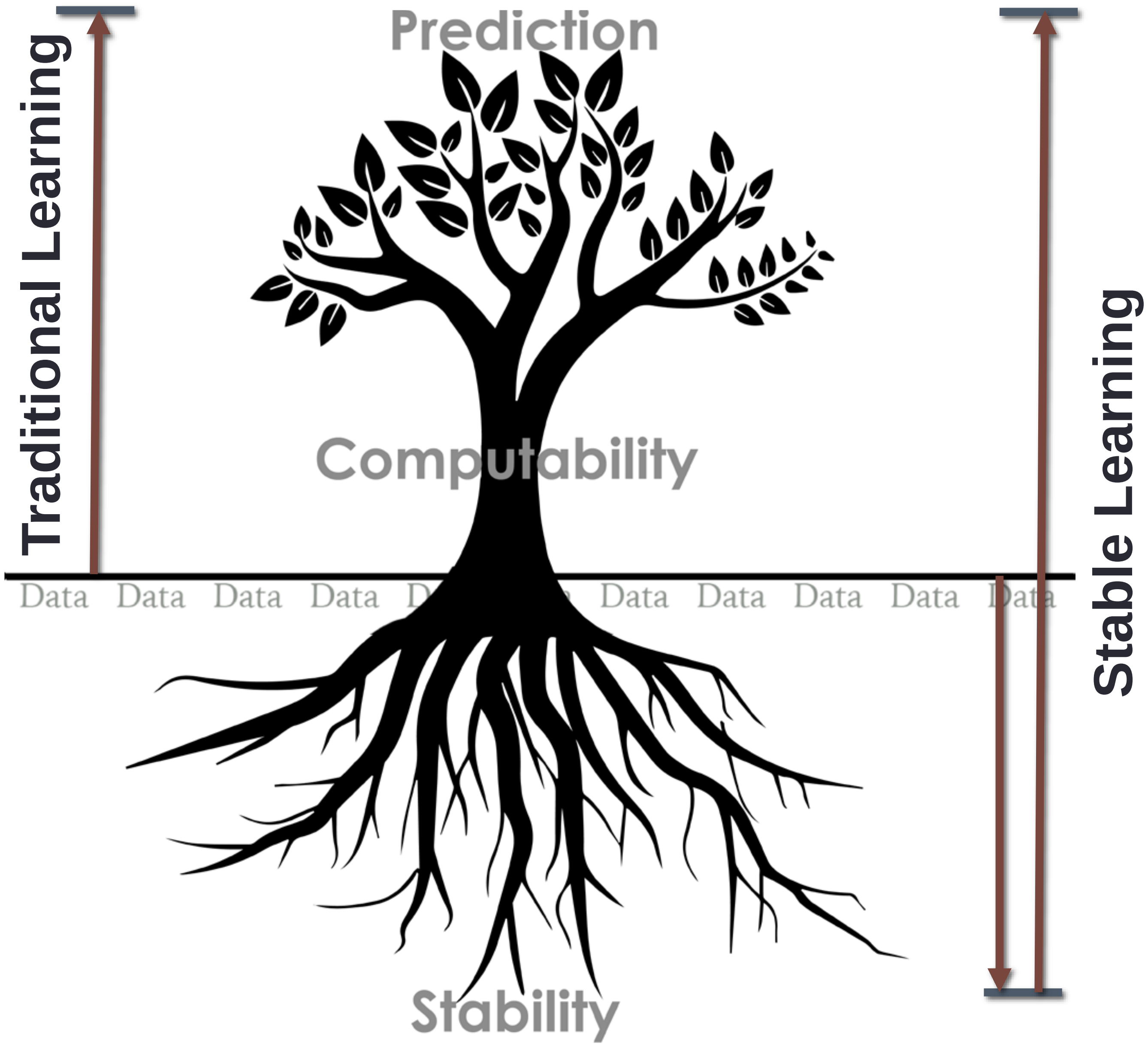
- Brief introduction to causal inference
- Stable learning and its development
- Positioning stable learning in OOD generalization
- Benchmark and dataset

Stability and Prediction

Prediction Performance

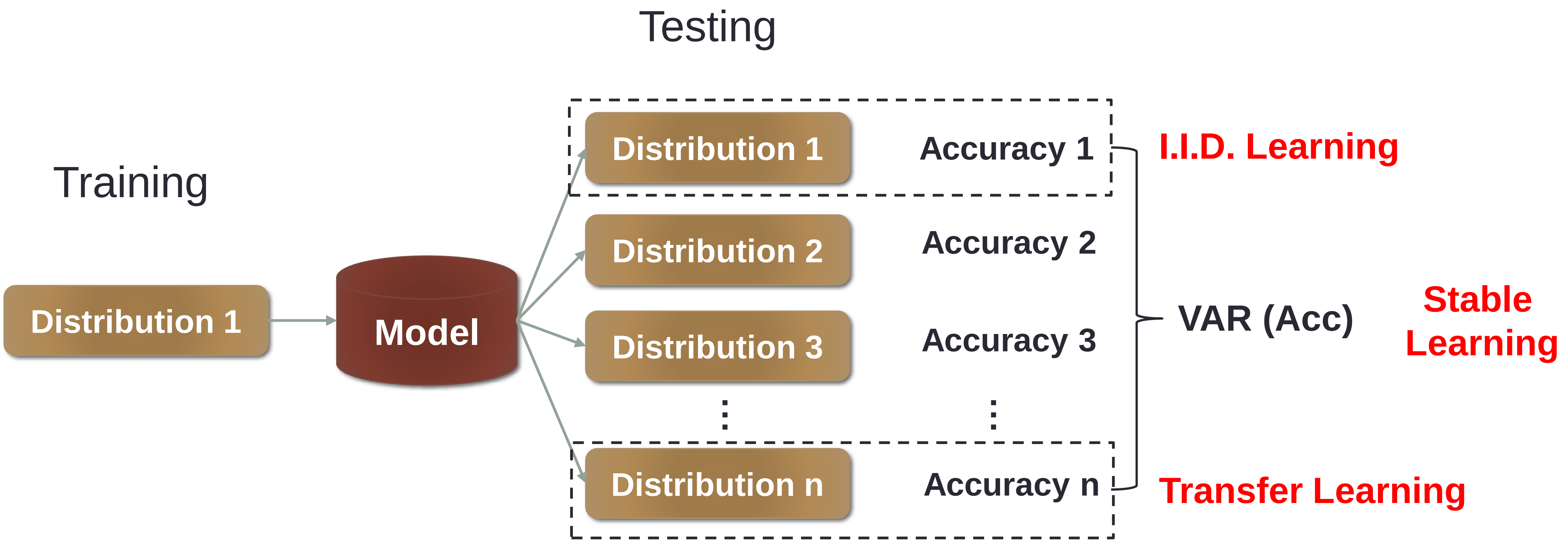
Learning Process

True Model

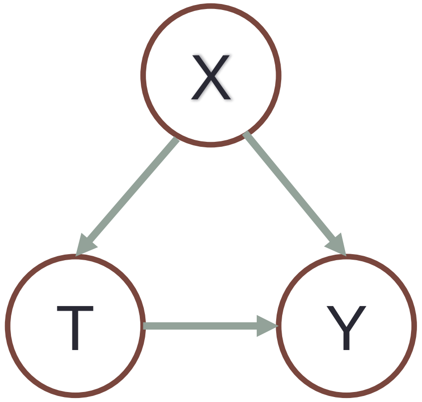


Bin Yu (2016), Three Principles of Data Science: predictability, computability, stability

Stable Learning



Revisit Directly Balancing for causal inference



Typical Causal Framework

Directly Confounder Balancing

Given a feature T

↓

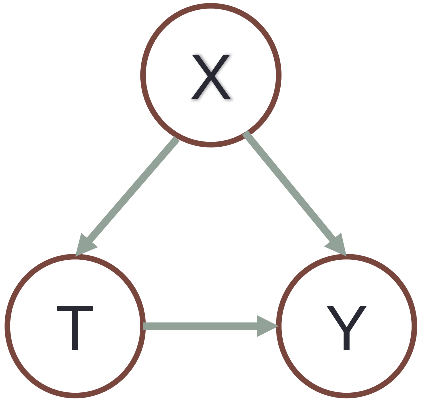
Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

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.

The core idea of stable learning: *Sample Reweighting*



Typical Causal Framework

Analogy of A/B Testing

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)

If all variables are independent after sample reweighting,
Correlation = Causality

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 for Global Balancing

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 (CRLR)

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

Sample reweighted logistic loss

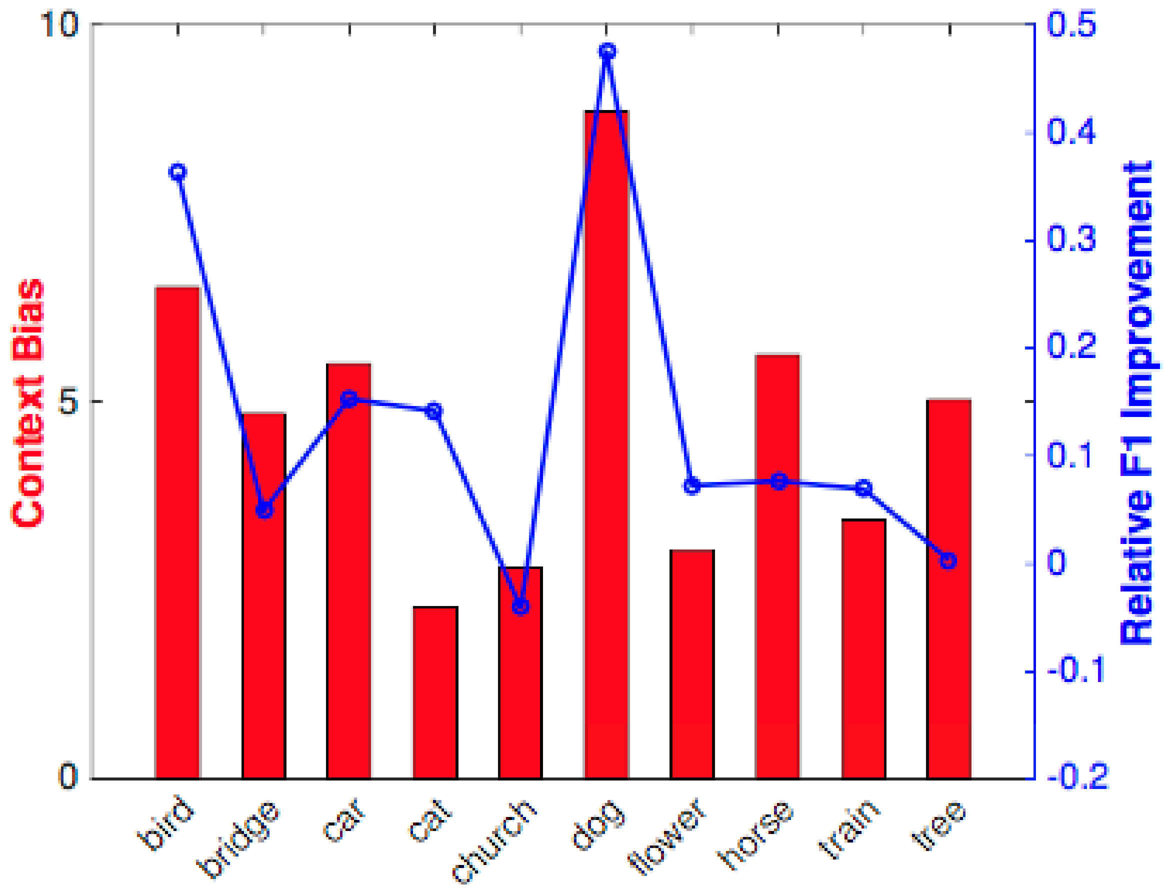
Causal Contribution

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

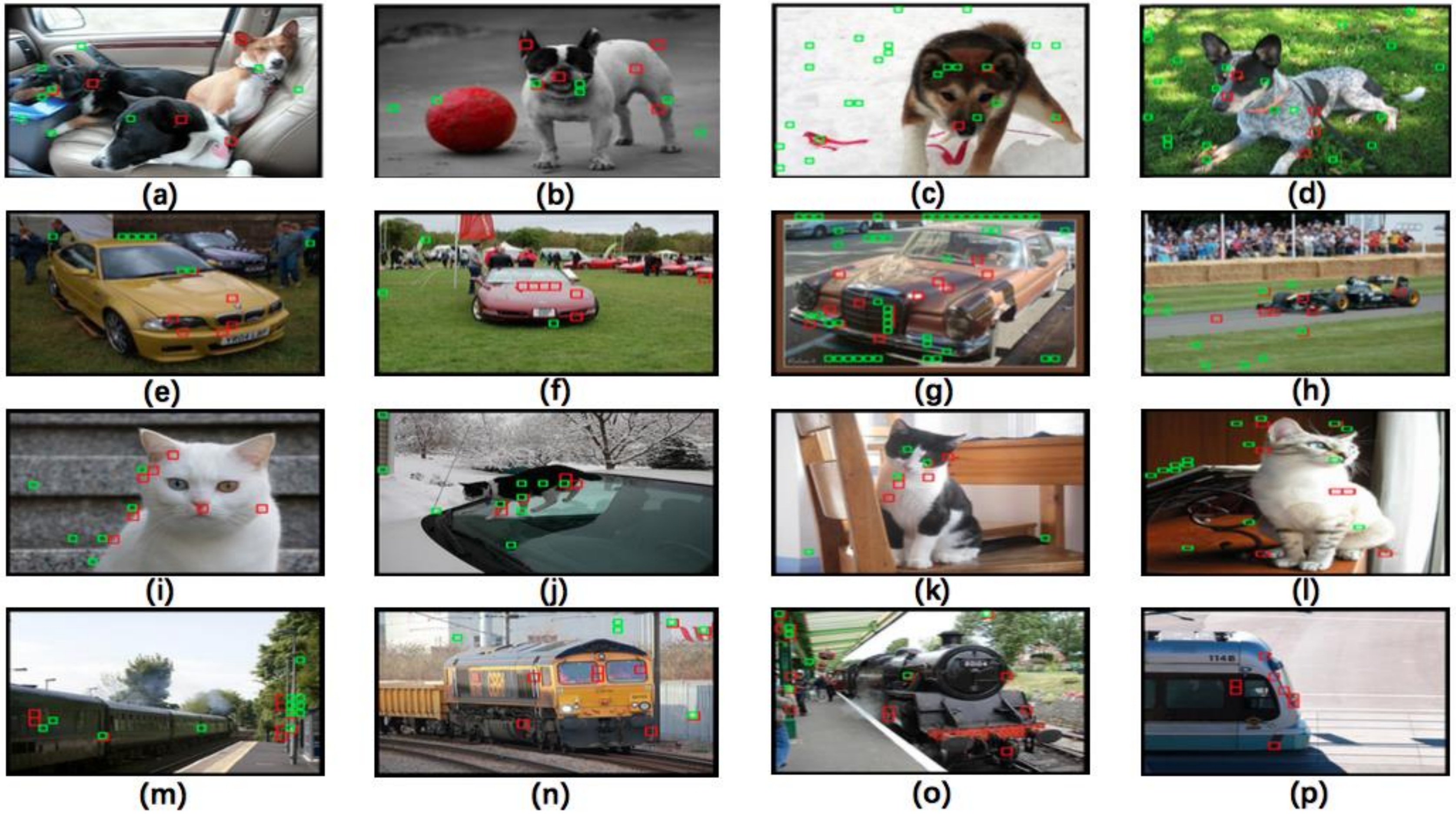
Experiment – Non-i.i.d. image classification

- 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



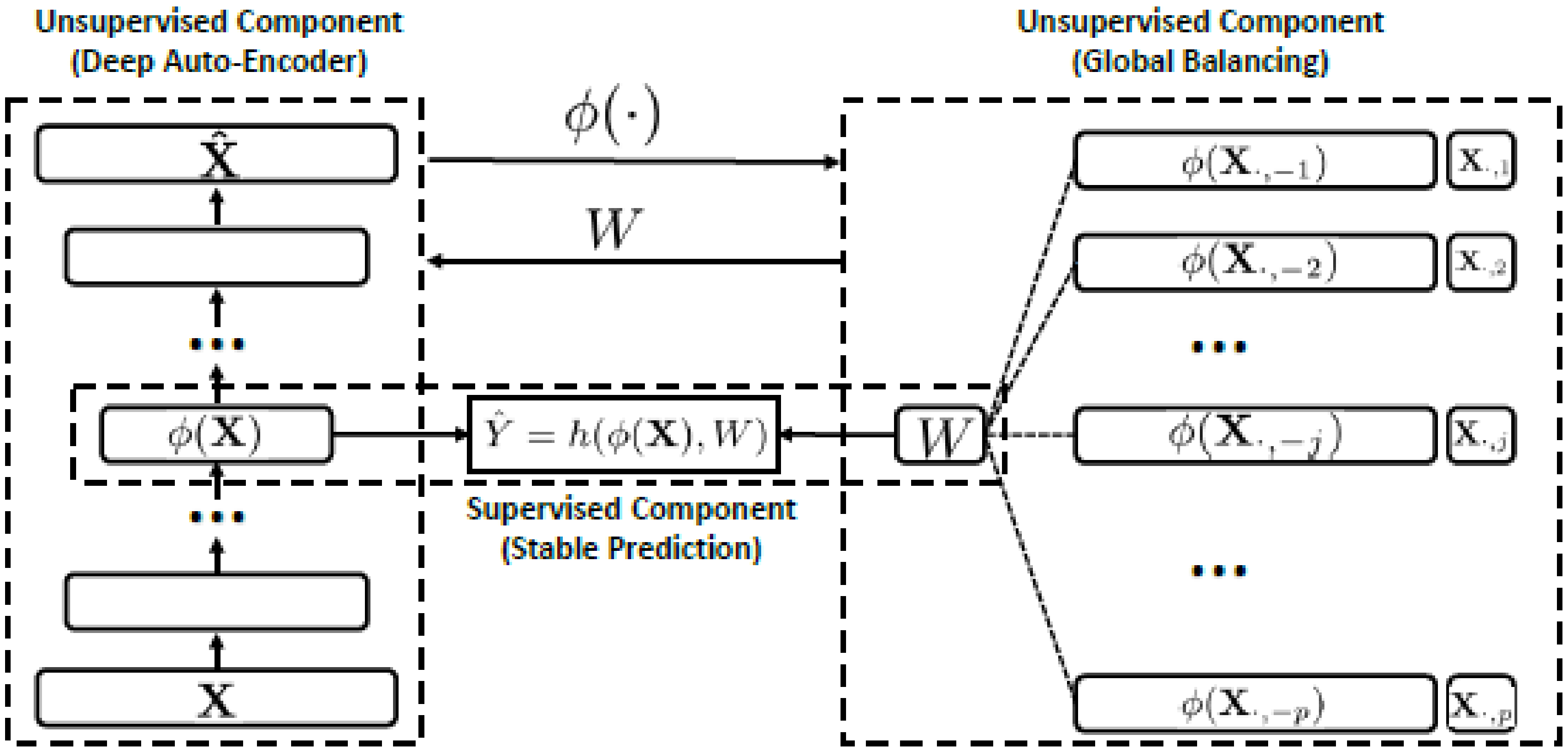
Limitations of Global Balancing

- A hidden assumption for Global Balancing to work

Assumption 2 (Overlap) *For any variable* $\mathbf{X}_{\cdot,j}$ *when setting it as the treatment variable, it has* $\forall j, 0 < P(\mathbf{X}_{\cdot,j} = 1 | \mathbf{X}_{\cdot,-j}) < 1$.

- Practical constraints
  - High dimensional features (potential treatment)
  - Sparsity of real world data
  - Possible interactions between features
  - More complex data type: categorical and continuous

From Shallow to Deep - DGBR



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

From Shallow to Deep - DGBR

- Deep Global Balancing Regression (DGBR) Algorithm

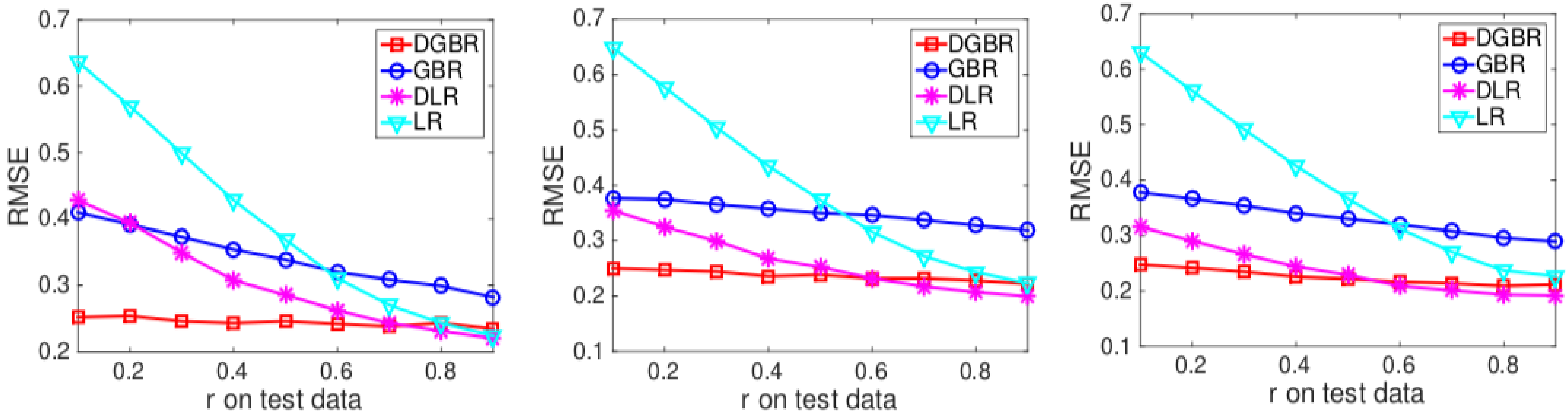
$$
\begin{aligned}
\min \quad & \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (\phi(\mathbf{X}_i)\beta))), \qquad (7)\\
s.t. \quad & \sum_{j=1}^{p} \left\| \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1,\\
& \|(W \cdot \mathbf{1}) \odot (X - \hat{X})\|_F^2 \leq \lambda_2, \quad W \geq 0, \quad \|W\|_2^2 \leq \lambda_3,\\
& \|\beta\|_2^2 \leq \lambda_4, \quad \|\beta\|_1 \leq \lambda_5, \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_6\\
& \sum_{k=1}^{K} (\|A^{(k)}\|_F^2 + \|\hat{A}^{(k)}\|_F^2) \leq \lambda_7,
\end{aligned}
$$

Deep Auto-Encoder

Global Balancing

Stable Prediction

Experiments on Synthetic Data



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

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

From Binary to Continuous Variable - DWR

Independence condition for continuous variable

For all $a, b \in \mathbb{N}, \; \mathbb{E}[\mathbf{X}_{,j}^{a}\mathbf{X}_{,k}^{b}] = \mathbb{E}[\mathbf{X}_{,j}^{a}]\mathbb{E}[\mathbf{X}_{,k}^{b}]$

Causal Regularizer for Continuous Variable

$$\min_{W} \sum_{j=1}^{p} \left\| \mathbb{E}[\mathbf{X}_{,j}^{T}\mathbf{\Sigma}_{W}\mathbf{X}_{,-j}] - \mathbb{E}[\mathbf{X}_{,j}^{T}W]\mathbb{E}[\mathbf{X}_{,-j}^{T}W] \right\|_2^2$$

Decorrelated Weighted Regression:

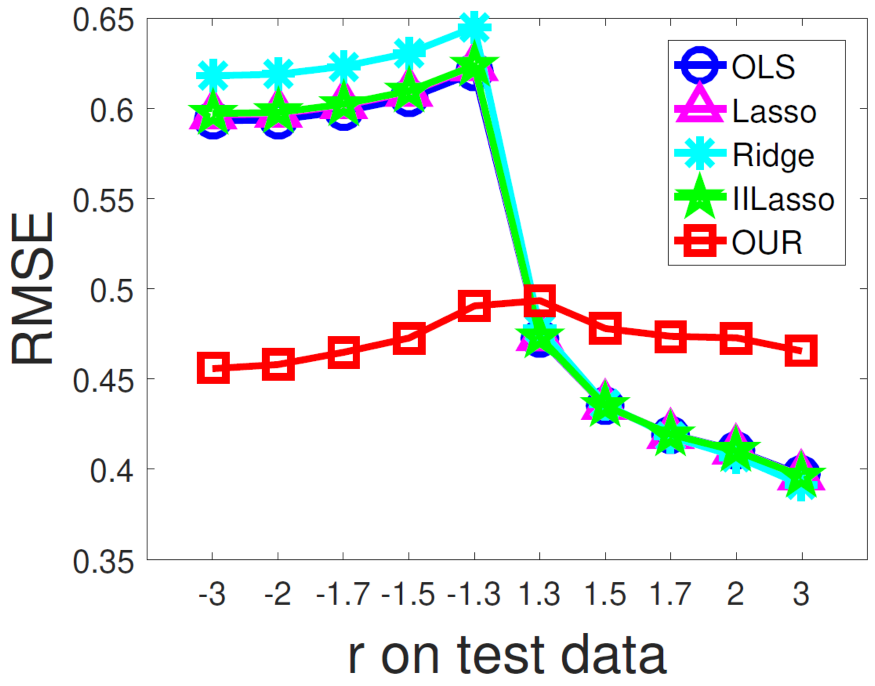
$$\min_{W,\beta} \sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2$$

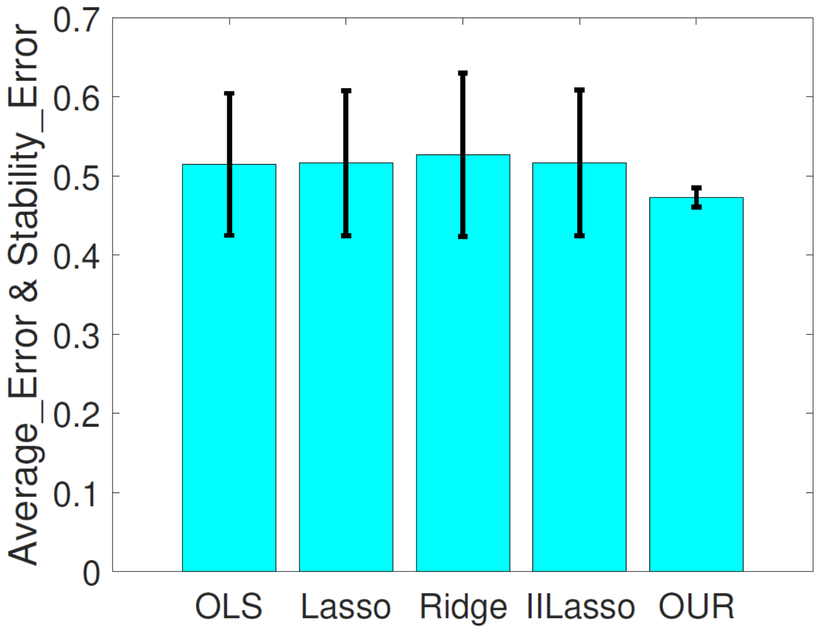
$$s.t \quad \sum_{j=1}^{p} \left\| \mathbf{X}_{,j}^{T}\mathbf{\Sigma}_{W}\mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^{T}W/n \cdot \mathbf{X}_{,-j}^{T}W/n \right\|_2^2 < \lambda_2$$

$$|\beta|_1 < \lambda_1, \; \frac{1}{n}\sum_{i=1}^{n} W_i^2 < \lambda_3,$$

$$(\frac{1}{n}\sum_{i=1}^{n} W_i - 1)^2 < \lambda_4, \;\; W \succeq 0,$$

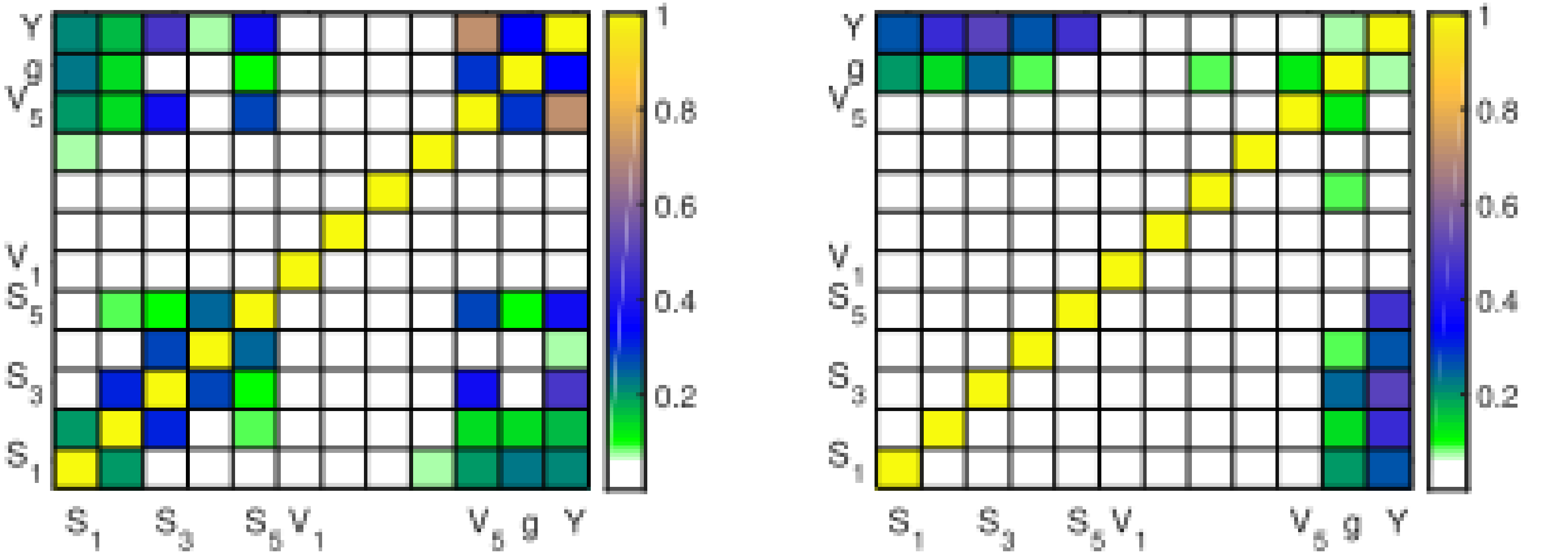
Stable Learning with ***Linear*** model





Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li. Stable Prediction with Model Misspecification and Agnostic Distribution Shift. ***AAAI***, 2020.

De-confounding for continuous variable



(a) On raw data

(b) On the weighted data

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

- Previous logic:

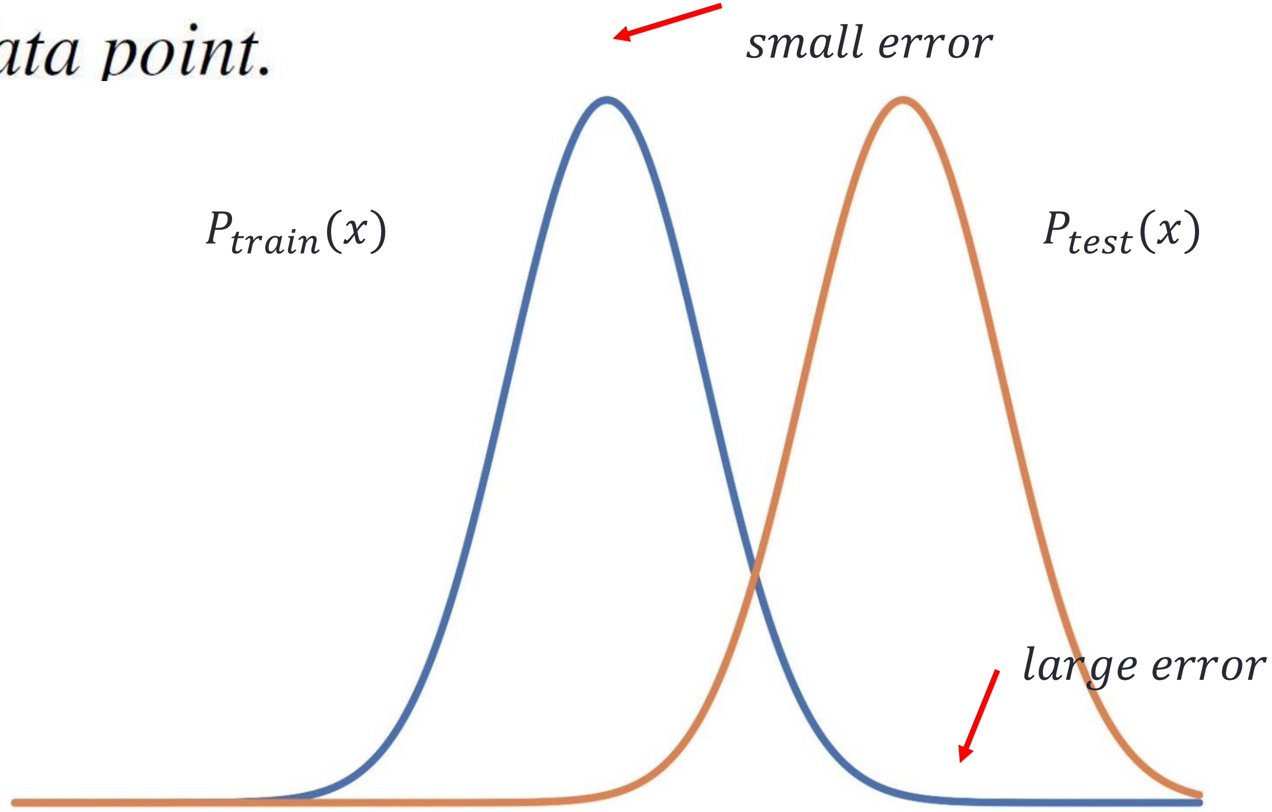
Sample Reweighting → Independent Variables → Causal Variable → Stable Prediction

- 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. AAAI, 2020.

Stable Learning of Linear Models

- Consider the linear regression with misspecification bias

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

Goes to infinity when perfect collinearity exists!

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

- By accurately estimating $\bar{\beta}$ 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 $\|\hat{\beta} - \bar{\beta}\|_2 \leq \boxed{2(\delta/\gamma) + \delta}$, 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. AAAI, 2020.

Toy Example

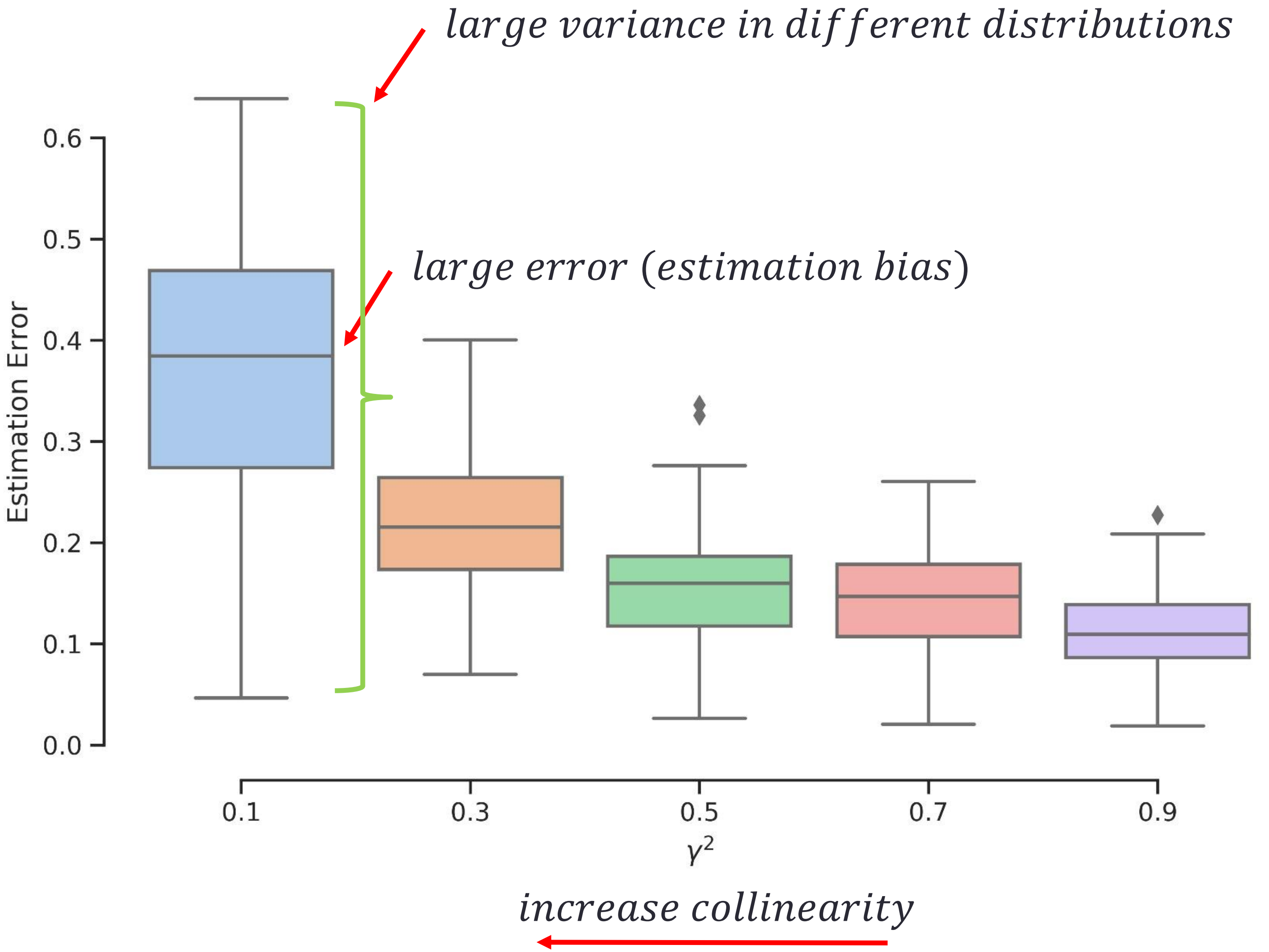
- Assume the design matrix X consists of two variables X_1, X_2, generated from a multivariate normal distribution:

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

- 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. AAAI, 2020.

Simulation Results



Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. AAAI, 2020.

Stable Learning of Sparse Linear Models

- Suppose $X=\{S,V\}$, and $Y=f(S)+\varepsilon$
- S: set of ***stable (causal) features***, i.e., eyes, ears of dog
- V: set of ***unstable (contextual) features***, i.e., grass, ground
- We assume the outcome is determined by sparse stable signals S regardless of V

Key reason of instability: Spurious correlation between V and Y

Theoretical Analysis

$$\hat{\beta}_{V_{OLS}} = \beta_V + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{V}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T g(\mathbf{S}_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{V}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{S}_i\right)\left(\beta_S - \hat{\beta}_{S_{OLS}}\right),$$

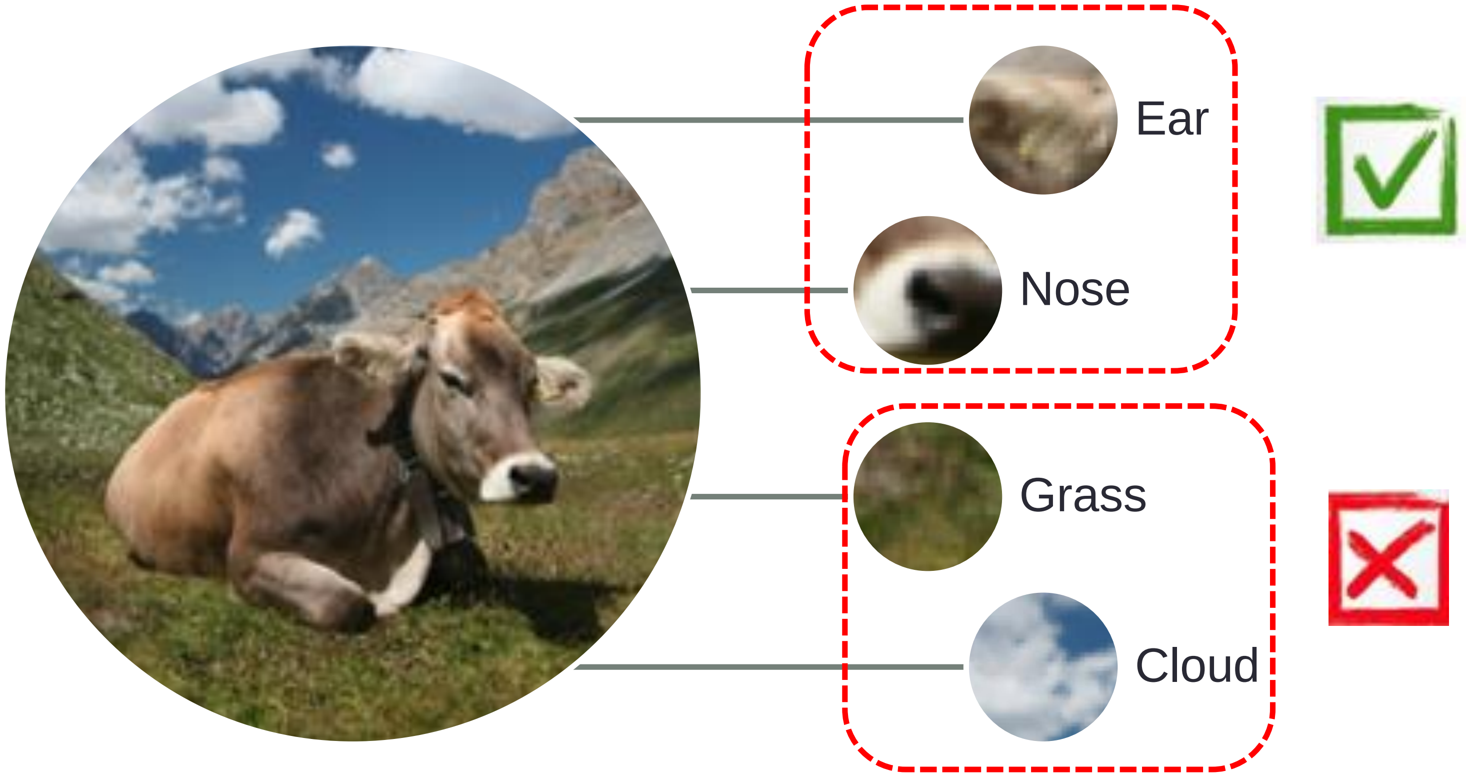
$$\hat{\beta}_{S_{OLS}} = \beta_S + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T g(\mathbf{S}_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{V}_i\right)\left(\beta_V - \hat{\beta}_{V_{OLS}}\right)$$

- The estimation error is induced by
  - Cov(S, V)
  - Cov(V, g(S))
  - Cov(S, g(S))

Spurious correlation between *V* and S may shift due to different time spans, regions and data collecting strategies, leading to unstable performance.

Our Idea – Heterogeneity & Modularity

Assumption 3. *The variables* $\mathbf{X} = \{X_1, X_2, \ldots X_p\}$ *could be partitioned into* k *distinct groups* $\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_k$. *For* $\forall i, j, i \neq j$ *and* $X_i, X_j \in \mathrm{G}_l, l \in \{1, 2, \ldots, k\}$, *we have* $P^e_{X_i X_j} = P_{X_i X_j}$.



Clustering?

Differentiated Variable Decorrelation

- Feature Partition by Stable Correlation Clustering
  - Define the dissimilarity of two variables:

$$Dis(X_i, X_j) = \sqrt{\frac{1}{M-1}\sum_{l=1}^{M}\left(Corr(X_i^l, X_j^l) - Ave\_Corr(X_i, X_j)\right)^2},$$

- Remove the correlation between variables via sample reweighting:

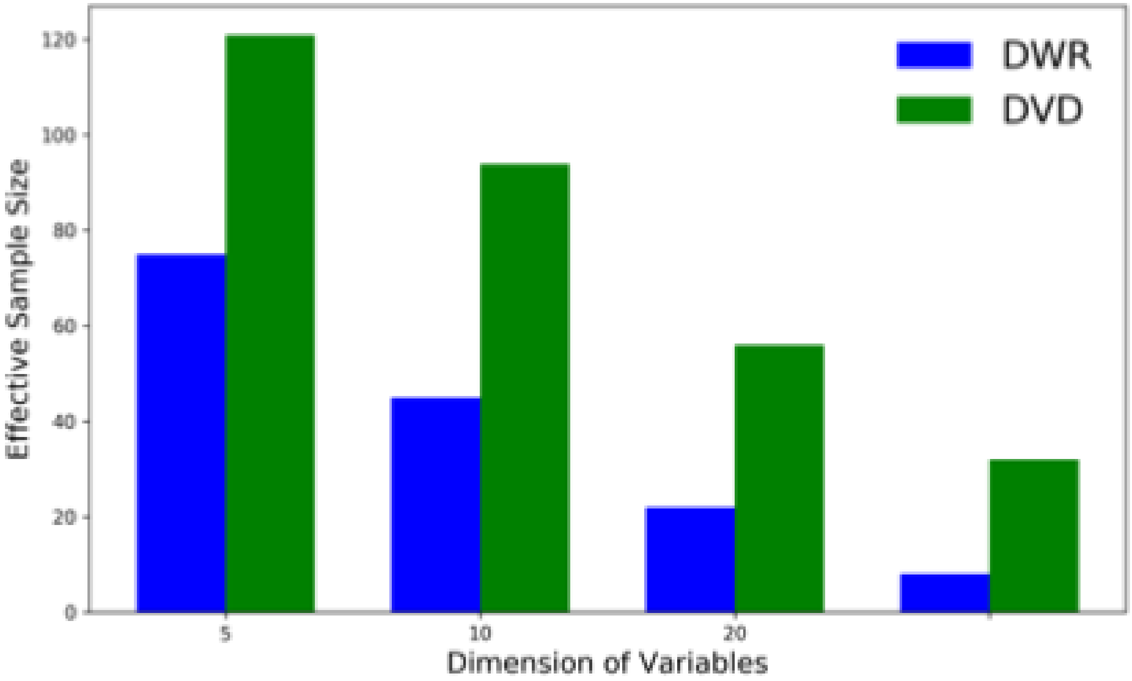
$$\min_{W} \sum_{i \neq j} \boxed{\mathbb{I}(i,j)} \left\| (\mathbf{X}_{,i}^T \Sigma_W \mathbf{X}_{,j}/n - \mathbf{X}_{,i}^T W/n \cdot \mathbf{X}_{,j}^T W/n) \right\|_2^2$$

$$\text{s.t } \frac{1}{n}\sum_{i=1}^{n} W_i^2 < \gamma_1, \quad \left(\frac{1}{n}\sum_{i=1}^{n} W_i - 1\right)^2 < \gamma_2, \quad W \geq 0$$

Zheyean Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li and Zhitang Chen. Stable Learning via Differentiated Variable Decorrelation. ***KDD***, 2020.

Experimental Results

Scenario 1: varying sample size n						
n, p_{v_b}, r	$n = 120, p_{v_b} = p * 0.2, r = 1.9$			$n = 160, p_{v_b} = p * 0.2, r = 1.9$		
Methods	β_Error	Average_Error	Stability_Error	β_Error	Average_Error	Stability_Error
OLS	1.988	0.470	0.087	1.870	0.489	0.105
Lasso	2.021	0.476	0.092	1.905	0.494	0.110
IILasso	2.035	0.475	0.094	1.920	0.498	0.113
DWR	2.012	0.545	0.099	1.991	0.502	0.076
Our	**1.892**	**0.469**	**0.040**	**1.741**	**0.489**	**0.050**
Scenario 2: varying number of unstable variables p_{v_b}						
n, p_{v_b}, r	$n = 200, p_{v_b} = p * 0.2, r = 1.9$			$n = 200, p_{v_b} = p * 0.3, r = 1.9$		
Methods	β_Error	Average_Error	Stability_Error	β_Error	Average_Error	Stability_Error
OLS	1.839	0.522	0.121	2.128	0.563	0.179
Lasso	1.876	0.529	0.129	2.176	0.571	0.186
IILasso	1.894	0.538	0.149	2.196	0.575	0.191
DWR	1.656	0.485	0.081	1.881	0.469	0.092
Our	**1.369**	**0.476**	**0.042**	**1.641**	**0.460**	**0.064**
Scenario 3: varying bias rate r on training data						
n, p_{v_b}, r	$n = 200, p_{v_b} = p * 0.2, r = 1.6$			$n = 200, p_{v_b} = p * 0.2, r = 1.8$		
Methods	β_Error	Average_Error	Stability_Error	β_Error	Average_Error	Stability_Error
OLS	1.296	**0.452**	0.064	1.780	0.510	0.117
Lasso	1.321	0.455	0.067	1.812	0.516	0.123
IILasso	1.339	0.457	0.070	1.829	0.519	0.125
DWR	**1.153**	0.457	0.033	1.262	0.458	0.035
Our	1.236	0.463	**0.021**	**1.236**	**0.450**	**0.023**



Effective Sample Size

Zheyean Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li and Zhitang Chen. Stable Learning via Differentiated Variable Decorrelation. ***KDD***, 2020.

StableNet: From Linear Models to Deep Models

Variable Decorrelation by Sample Reweighting and RFF:

- Measure and eliminate the complex non-linear dependencies among features with RFF
- The computation cost is acceptable

$$\underbrace{\boxed{\mathcal{H}_{\text{RFF}}}}_{\text{Function space of RFF}} = \{h : x \to \sqrt{2}\cos(\omega x + \phi) \mid \omega \sim N(0,1), \phi \sim \text{Uniform}(0, 2\pi)\},$$

$$\underbrace{\boxed{f^{(t+1)}}}_{\text{Representation function}}, \underbrace{\boxed{g^{(t+1)}}}_{\text{Prediction function}} = \arg\min_{f,g} \sum_{i=1}^{n} w_i^{(t)} L(g(f(\mathbf{X}_i)), y_i),$$

$$\underbrace{\boxed{\mathbf{w}^{(t+1)}}}_{\text{Sample weights}} = \arg\min_{\mathbf{w}\in\Delta_n} \sum_{1\le i<j\le m_Z} \left\| \underbrace{\boxed{\hat{\Sigma}_{\mathbf{Z}^{(t+1)}_{:,i}\mathbf{Z}^{(t+1)}_{:,j};\mathbf{w}}}}_{\text{Partial cross-covariance matrix}} \right\|_F^2$$

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. ***CVPR***, 2021

Learning sample weights globally

Optimize sample weights globally by saving and reloading all features and weights.

$$\mathbf{Z}_O = \mathrm{Concat}\left(\mathbf{Z}_{G1}, \mathbf{Z}_{G2}, \cdots, \mathbf{Z}_{Gk}, \mathbf{Z}_L\right),$$
$$\mathbf{w}_O = \mathrm{Concat}\left(\mathbf{w}_{G1}, \mathbf{w}_{G2}, \cdots, \mathbf{w}_{Gk}, \mathbf{w}_L\right)$$

Overall representation

Overall sample weights

Reloaded representation

Reloaded sample weights

$$\mathbf{Z}'_{Gi} = \alpha_i \mathbf{Z}_{Gi} + (1-\alpha_i)\mathbf{Z}_L,$$
$$\mathbf{w}'_{Gi} = \alpha_i \mathbf{w}_{Gi} + (1-\alpha_i)\mathbf{w}_L,$$

Saved representation

Saved sample weights

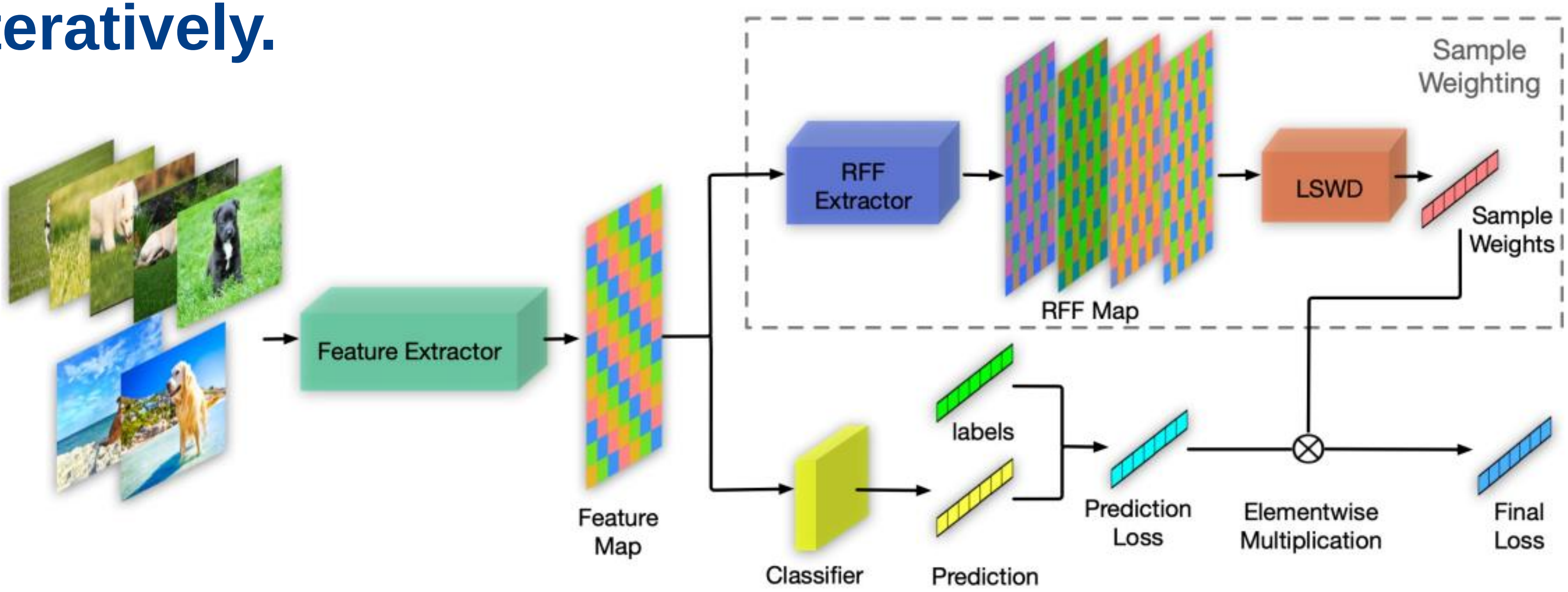
Current representation

Current sample weights

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. ***CVPR***, 2021

Learning sample weights globally

- Sample weights learning module is an independent module which can be easily assembled with current deep models.
- Sample weights and the classification model are trained iteratively.



Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. ***CVPR***, 2021

Out-Of-Distribution Generalization

- **The heterogeneity of training data is not significant nor known.**
- **The capacities of different domains can varies significantly.**

NICO dataset

Flexible OOD Generalization

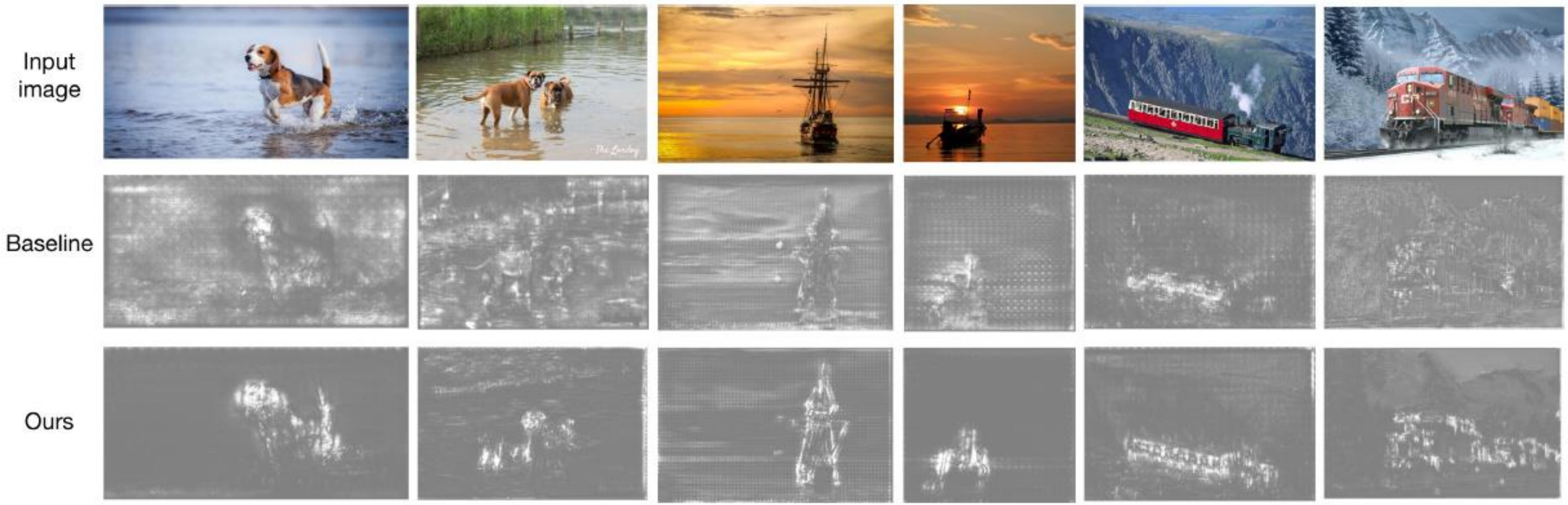
- **The domains for different categories can be different.**
- **For instance, birds can be on trees but hardly in the water while fishes are the opposite.**

	JiGen	M-ADA	DG-MMLD	RSC	ResNet-18	StableNet (ours)
PACS	40.31	30.32	42.65	39.49	39.02	**45.14**
VLCS	76.75	69.58	78.96	74.81	73.77	**79.15**
NICO	54.42	40.78	47.18	57.59	51.71	**59.76**

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. ***CVPR***, 2021

Saliency maps of StableNet and other models

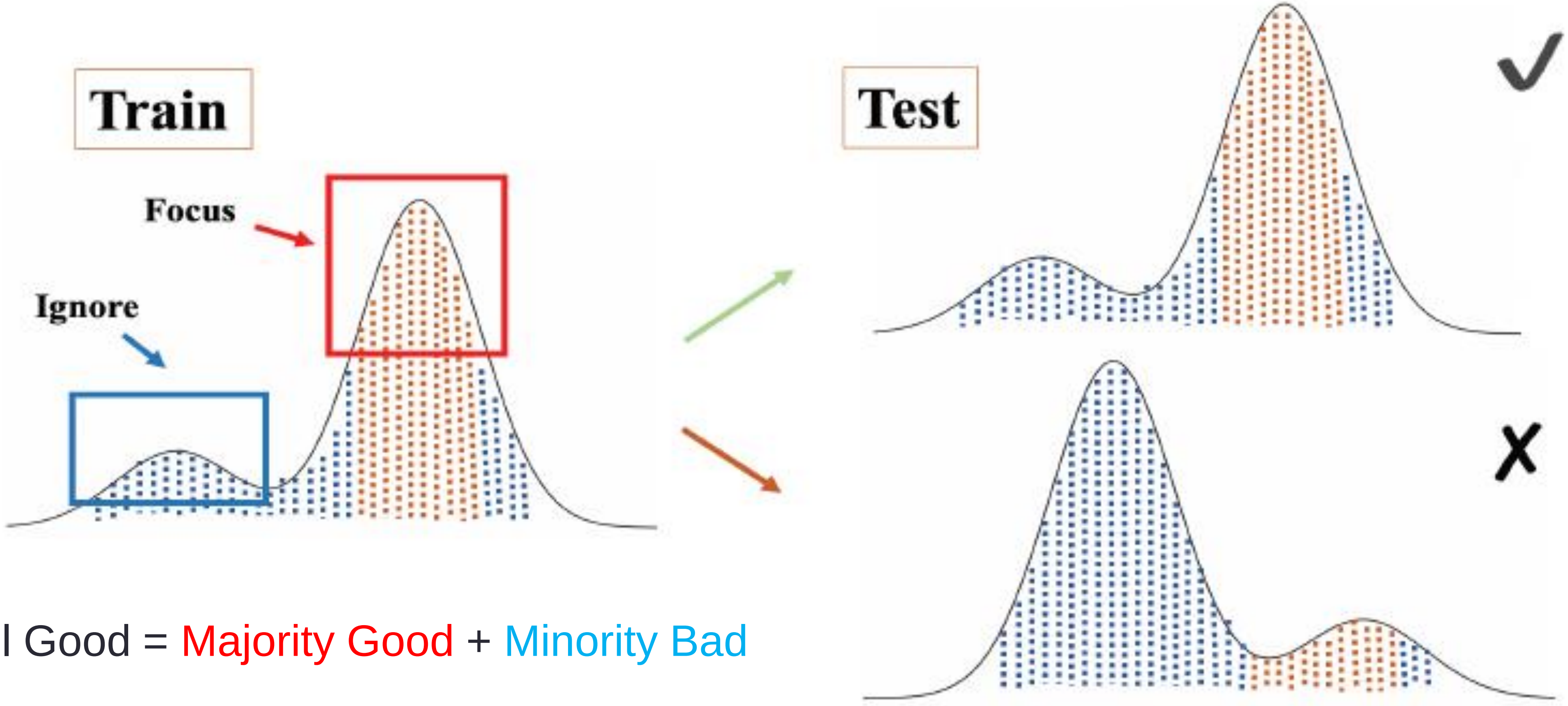
- **The visualization of the gradient of the class score function with respect to the input pixels. The brighter the pixel is, the more contribution it makes to prediction.**



Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. ***CVPR***, 2021

OOD generalization: Model v.s. ***Optimization***?

$$\theta_{ERM} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta; X_i, Y_i)$$

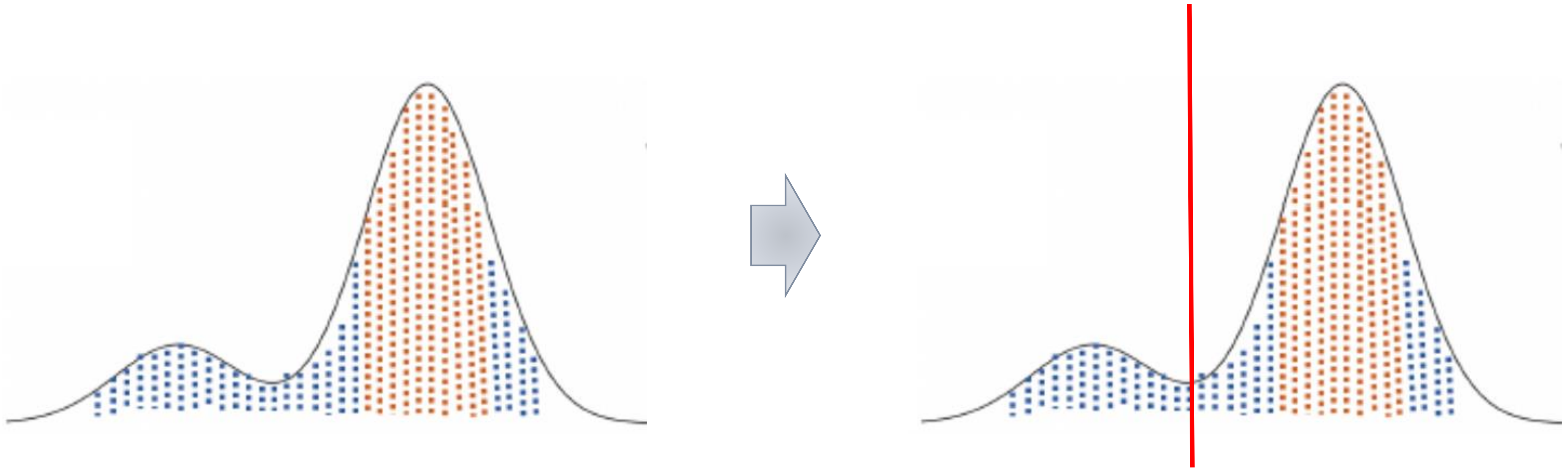


Overall Good = Majority Good + Minority Bad

Overall Good = Majority Good + Minority Good

Problem I
Uncovering Heterogeneity

Problem II
Finding Invariance



Heterogeneity → Invariance

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.

Invariance Assumption

To deal with the potential distributional shifts, one common assumption made in invariant learning is the **Invariance Assumption**.

Assumption (Invariance Assumption)

There exists random variable $\Phi^*(X)$ such that the following properties hold:

1. Invariance property: *for all $e_1, e_2 \in \mathrm{supp}(\mathcal{E})$, we have*

$$P^{e_1}(Y|\Phi^*(X)) = P^{e_2}(Y|\Phi^*(X)) \tag{4}$$

2. Sufficiency property: $Y = f(\Phi^*) + \epsilon, \ \epsilon \perp X$.

Here we make some demonstrations on the Invariance Assumption:

- The first property assumes that the relationship between $\Phi^*(X)$ and Y remains invariant across environments, which is also referred to as causal relationship.
- The second property assumes that $\Phi^*(X)$ can provide all information of the target label Y.
- $\Phi^*(X)$ is referred to as **(Causally) Invariant Predictors**.

Maximal Invariant Predictor

To obtain the invariant predictor $\Phi^*(X)$, one can seeks for the **Maximal Invariant Predictor**[1,2], which is defined as follows:

Definition (Invariance Set & Maximal Invariant Predictor)

The invariance set $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:

$$\mathcal{I}_{\mathcal{E}} = \{\Phi(X) : Y \perp \mathcal{E}|\Phi(X)\} = \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\} \tag{5}$$

where $H[\cdot]$ is the Shannon entropy of a random variable. The corresponding maximal invariant predictor (MIP) of $\mathcal{I}_{\mathcal{E}}$ is defined as:

$$S = \arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{6}$$

where $I(\cdot;\cdot)$ measures Shannon mutual information between two random variables.

Remarks:

- $\Phi^*(X)$ is MIP.
- Optimal for OOD is $\hat{Y} = \mathbb{E}[Y|\Phi^*(X)]$.
- "Find $\Phi^*(X)$" $\rightarrow$ "Find MIP"

[1] Chang, S., Zhang, Y. et al. (2020, November). Invariant rationalization.

[2] Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem ?

- The flow of Invariant Learning methods:

$$\text{Given } \mathcal{E}_{tr} \rightarrow \text{Find MIP } \Phi^*_{tr} \text{ of } \mathcal{I}_{\mathcal{E}_{tr}} \rightarrow \text{Predict using } \Phi^*_{tr} \rightarrow \text{OOD "Optimal?"}$$

- Recall the definition of MIP:

$$\arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi) \tag{7}$$

1. MIP relies on the invariance set $\mathcal{I}_{\mathcal{E}}$
2. Invariance set $\mathcal{I}_{\mathcal{E}}$ relies on the given environments $\mathcal{E}$.

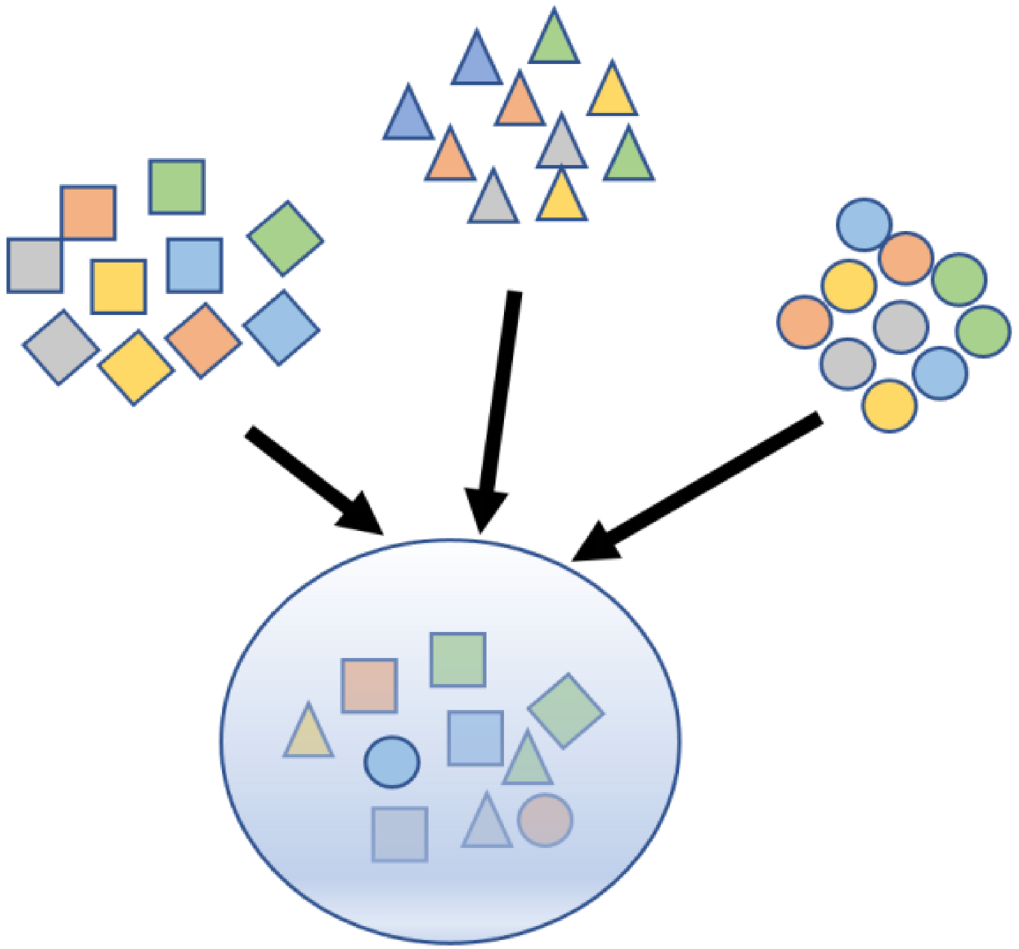
- What happens when $\mathcal{E}$ is replaced by $\mathcal{E}_{tr}$?

1. $\mathrm{supp}(\mathcal{E}_{tr}) \subset \mathrm{supp}(\mathcal{E})$
2. $\mathcal{I}_{\mathcal{E}} \subset \mathcal{I}_{\mathcal{E}_{tr}}$
3. Φ^*_{tr} NOT INVARIANT.

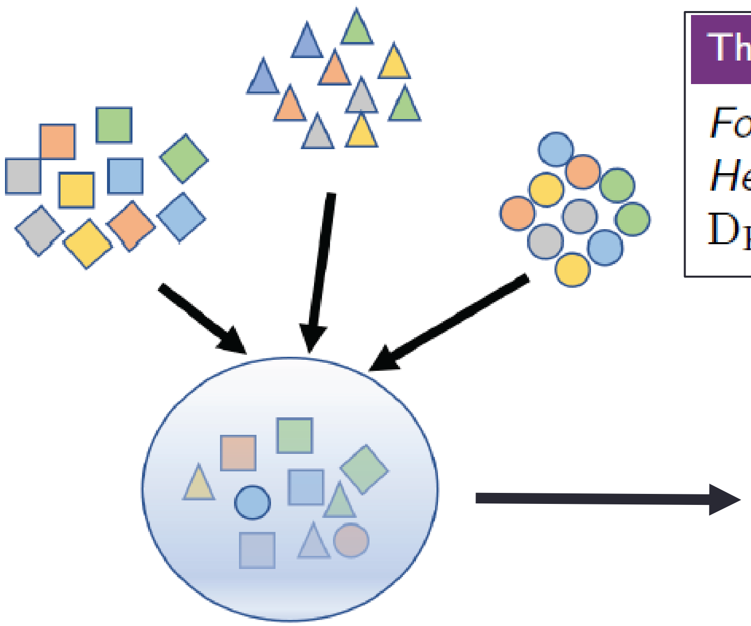
Remark: We need training environments where $\mathcal{I}_{\mathcal{E}_{tr}} \rightarrow \mathcal{I}_{\mathcal{E}}$

No Training Environments

Modern datasets are frequently assembled by merging data from multiple sources without explicit source labels, which means there are not multiple environments but only one pooled dataset.

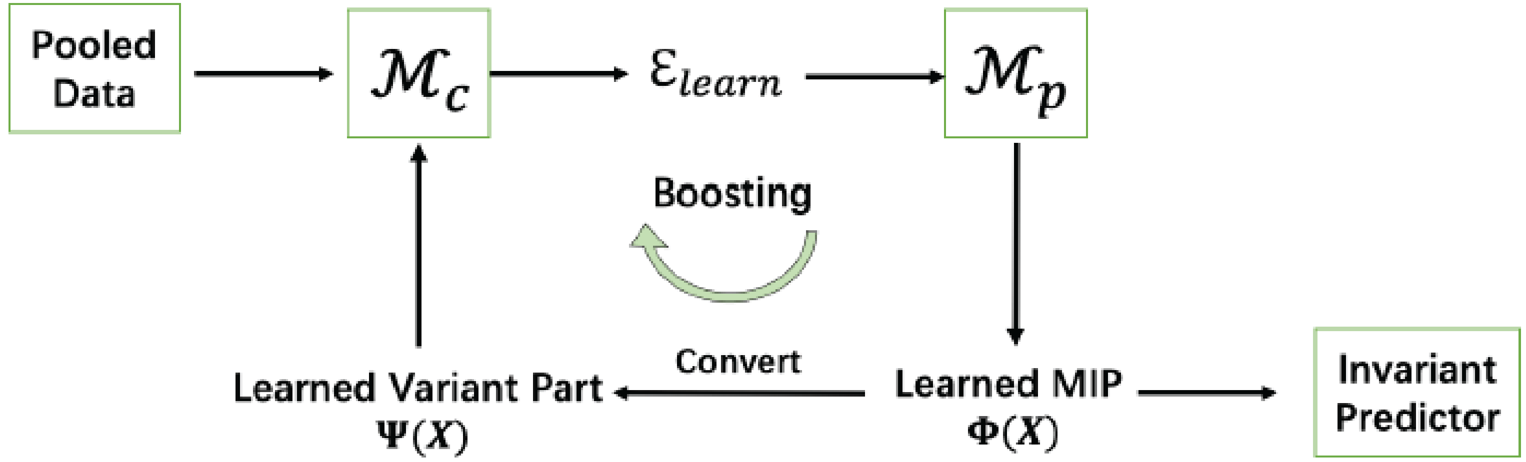


ERM → HRM (Heterogeneous Risk Minimization)



Theorem (Why using only Ψ?)

For $e_i, e_j \in \mathrm{supp}(\mathcal{E}_{tr})$, assume that $X = [\Phi^*, \Psi^*]^T$ satisfying Invariance and Heterogeneity Assumption, where Φ^* is invariant and Ψ^* variant. Then we have
$\mathrm{D_{KL}}(P^{e_i}(Y|X) \| P^{e_j}(Y|X)) \leq \mathrm{D_{KL}}(P^{e_i}(Y|\Psi^*) \| P^{e_j}(Y|\Psi^*))$



Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.

The Heterogeneity Identification Module $\mathcal{M}_c$

Recall that for $\mathcal{M}_c$,

$$\Psi(X) \rightarrow \mathcal{M}_c \rightarrow \mathcal{E}_{learn}$$

we implement it with a convex clustering method. Different from other clustering methods, we cluster the data according to the **relationship** between $\Psi(X)$ and Y.

- Assume the j-th cluster centre $P_{\Theta_j}(Y|\Psi)$ parameterized by Θ_j to be a Gaussian around $f_{\Theta_j}(\Psi)$ as $\mathcal{N}(f_{\Theta_j}(\Psi), \sigma^2)$:

$$h_j(\Psi, Y) = P_{\Theta_j}(Y|\Psi) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(Y - f_{\Theta_j}(\Psi))^2}{2\sigma^2}) \tag{8}$$

- The empirical data distribution is $\hat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_i(\Psi, Y)$
- The target is to find a distribution in $\mathcal{Q} = \{Q|Q = \sum_{j\in[K]} q_j h_j(\Psi, Y), \mathbf{q} \in \Delta_K\}$ to fit the empirical distribution best.
- The objective function of our heterogeneous clustering is:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \tag{9}$$

The Invariant Prediction Module $\mathcal{M}_p$

Recall that for $\mathcal{M}_p$,

$$\mathcal{E}_{learn} \to \mathcal{M}_p \to \Phi(X) = M \odot X$$

The algorithm involves two parts, invariant prediction and feature selection.

- For invariant prediction, we adopt the regularizer[4] as:

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) \tag{10}$$

  - Restrict the gradient across environments to be the same.
  - Only use invariant features.

- For feature selection, we adopt the continuous feature selection method that allows for continuous optimization of M:

$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M [\ell(M \odot X^e, Y^e; \theta) + \alpha \|M\|_0] \tag{11}$$

  - $\|M\|_0$ controls the number of selected features.
  - Conduct continuous optimization as [5].

---

[4] Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem ?

[5] Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates, in ICML2020

The Mutual Promotion

- Insight: We should only use Ψ^* for Heterogeneity Identification.

Assumption (Heterogeneity Assumption from Information Theory)

Assume the pooled training data is made up of heterogeneous data sources: $P_{tr} = \sum_{e \in \mathrm{supp}(\mathcal{E}_{tr})} w_e P^e$. *For any* $e_i, e_j \in \mathcal{E}_{tr}, e_i \neq e_j$, *we assume*

$$I^c_{i,j}(Y;\Phi^*|\Psi^*) \geq \max(I_i(Y;\Phi^*|\Psi^*), I_j(Y;\Phi^*|\Psi^*)) \quad (12)$$

where Φ^* *is invariant feature and* Ψ^* *the variant.* I_i *represents mutual information in* P^{e_i} *and* $I^c_{i,j}$ *represents the cross mutual information between* P^{e_i} *and* P^{e_j} *takes the form of* $I^c_{i,j}(Y;\Phi|\Psi) = H^c_{i,j}[Y|\Psi] - H^c_{i,j}[Y|\Phi,\Psi]$ *and* $H^c_{i,j}[Y] = -\int p^{e_i}(y)\log p^{e_j}(y)dy$.

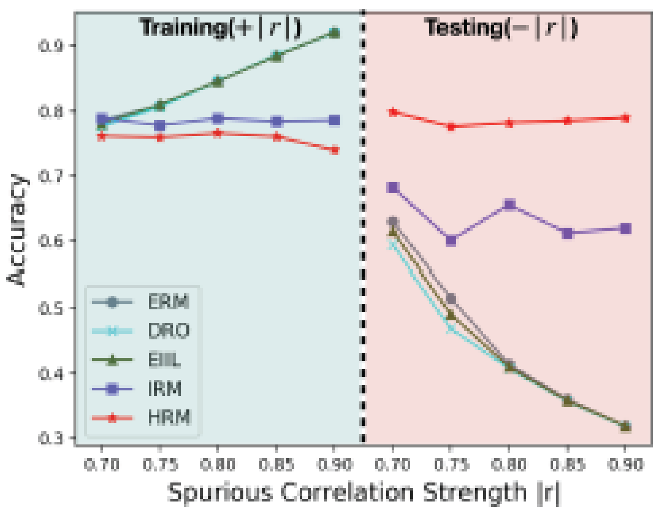
- The mutual information $I_i(Y;\Phi^*) = H_i[Y] - H_i[Y|\Phi^*]$ can be viewed as the error reduction if we use Φ^* to predict Y rather than predict by nothing.
- The cross mutual information $I^c_{i,j}(Y;\Phi^*)$ can be viewed as the error reduction if we use the predictor learned on Φ^* in environment e_j to predict in environment e_i, rather than predict by nothing.

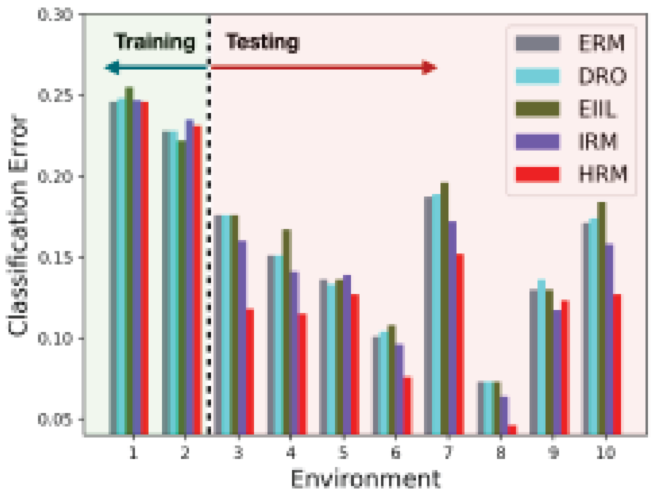
Theorem (Why using only Ψ?)

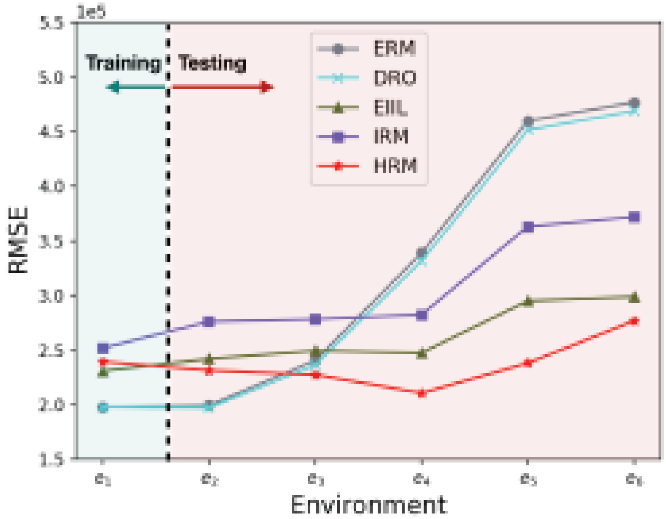
For $e_i, e_j \in \mathrm{supp}(\mathcal{E}_{tr})$, *assume that* $X = [\Phi^*, \Psi^*]^T$ *satisfying Invariance and Heterogeneity Assumption, where* Φ^* *is invariant and* Ψ^* *variant. Then we have* $\mathrm{D_{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \leq \mathrm{D_{KL}}(P^{e_i}(Y|\Psi^*)\|P^{e_j}(Y|\Psi^*))$

Results

Scenario 1: $n_\phi = 9, \ n_\psi = 1$										
e	Training environments			Testing environments						
Methods	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
ERM	0.290	0.308	0.376	0.419	0.478	0.538	0.596	0.626	0.640	0.689
DRO	0.289	0.310	0.388	0.428	0.517	0.610	0.627	0.669	0.679	0.739
EIIL	**0.075**	**0.128**	0.349	0.485	0.795	1.162	1.286	1.527	1.558	1.884
IRM(with $\mathcal{E}_{tr}$ label)	0.306	0.312	0.325	0.328	0.343	0.358	0.365	0.374	0.377	0.392
HRM^s	1.060	1.085	1.112	1.130	1.207	1.280	1.325	1.340	1.371	1.430
HRM	0.317	0.314	**0.322**	**0.318**	**0.321**	**0.317**	**0.315**	**0.315**	**0.316**	**0.320**



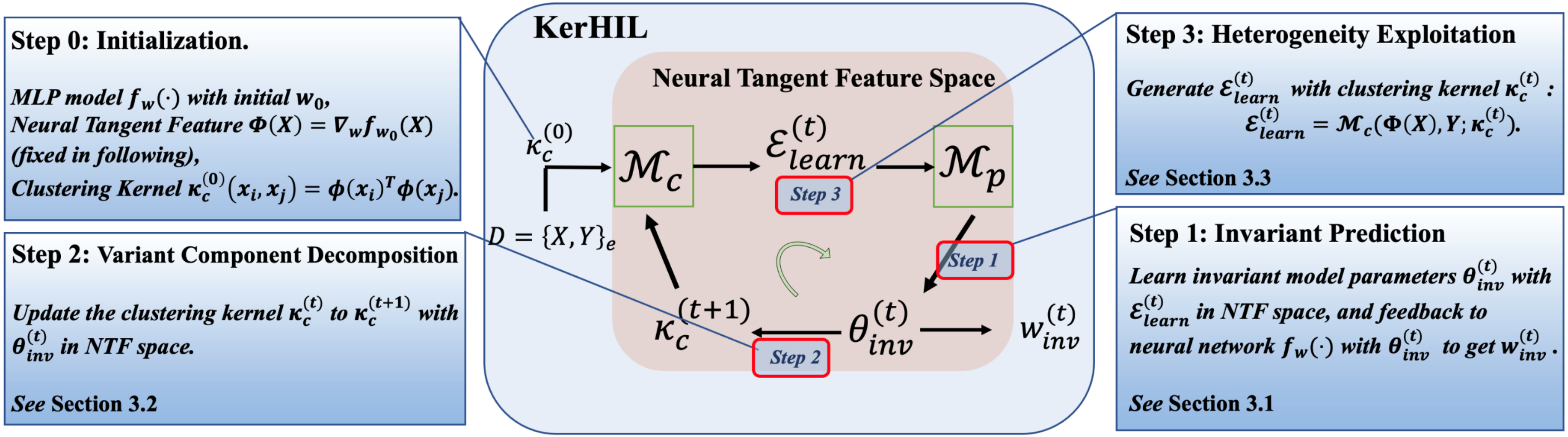




Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.

Kernelized Heterogeneous Risk Minimization

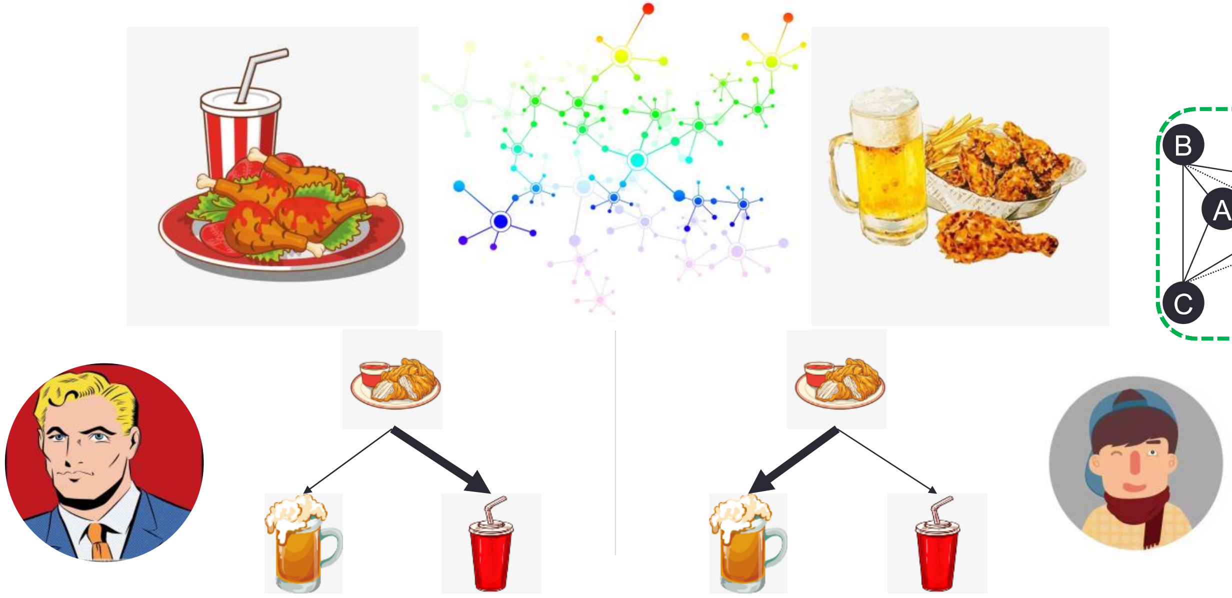
- To solve the HRM problem beyond the raw feature level.

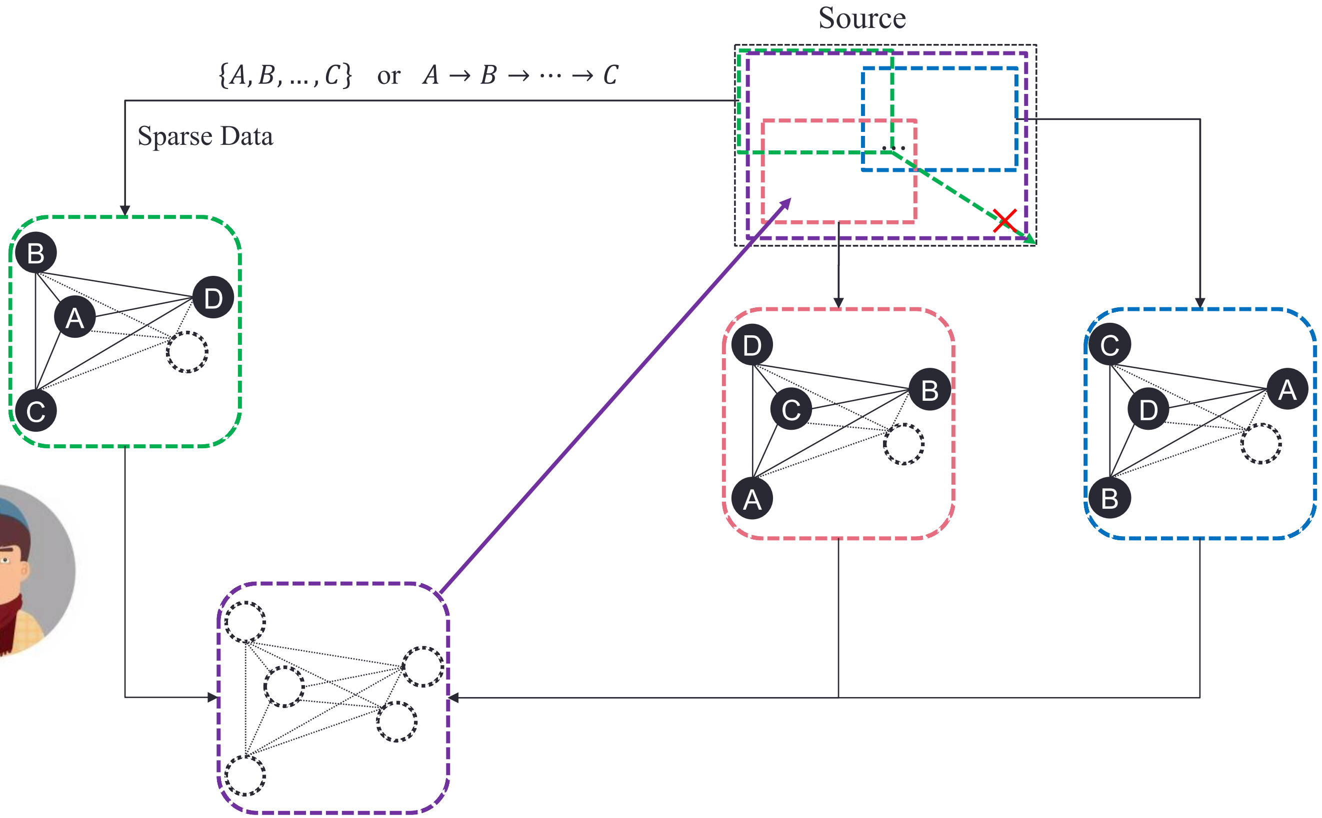


- Incorporate Neural Tangent Kernel.
- Perform the heterogeneity identification and invariant prediction in the Neural Tangent Feature Space.

Jiashuo Liu, Zheyuan Hu, Peng Cui, et al. Kernelized Heterogeneous Risk Minimization. ***NeurIPS***, 2021.

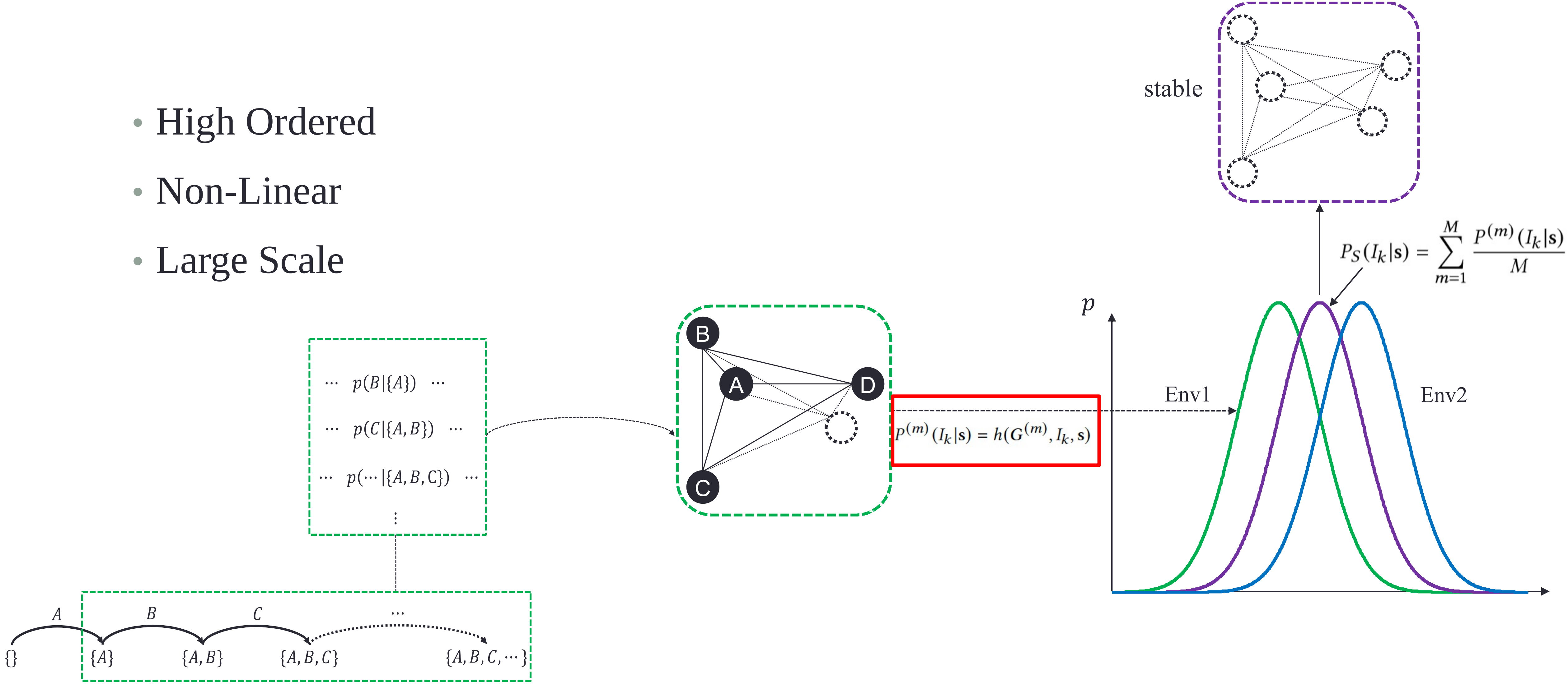
Stable Learning of Graph Structure



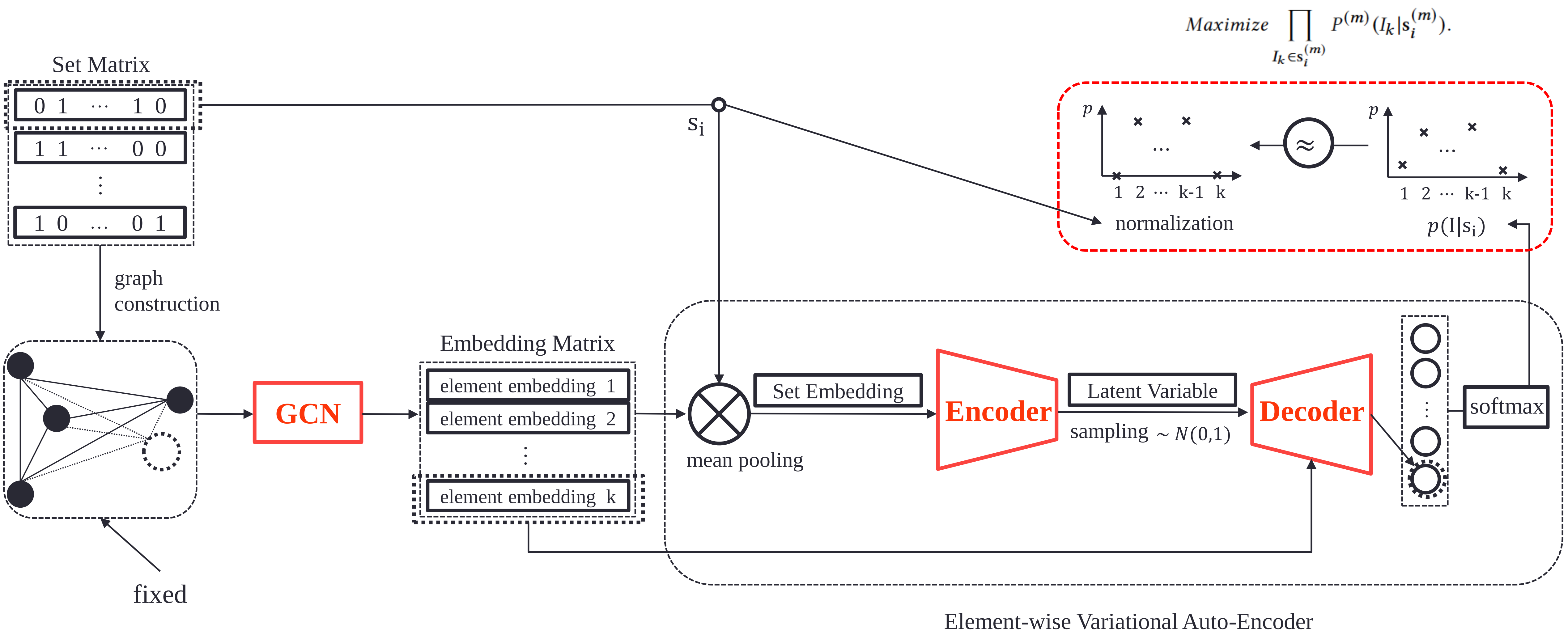


Core Idea

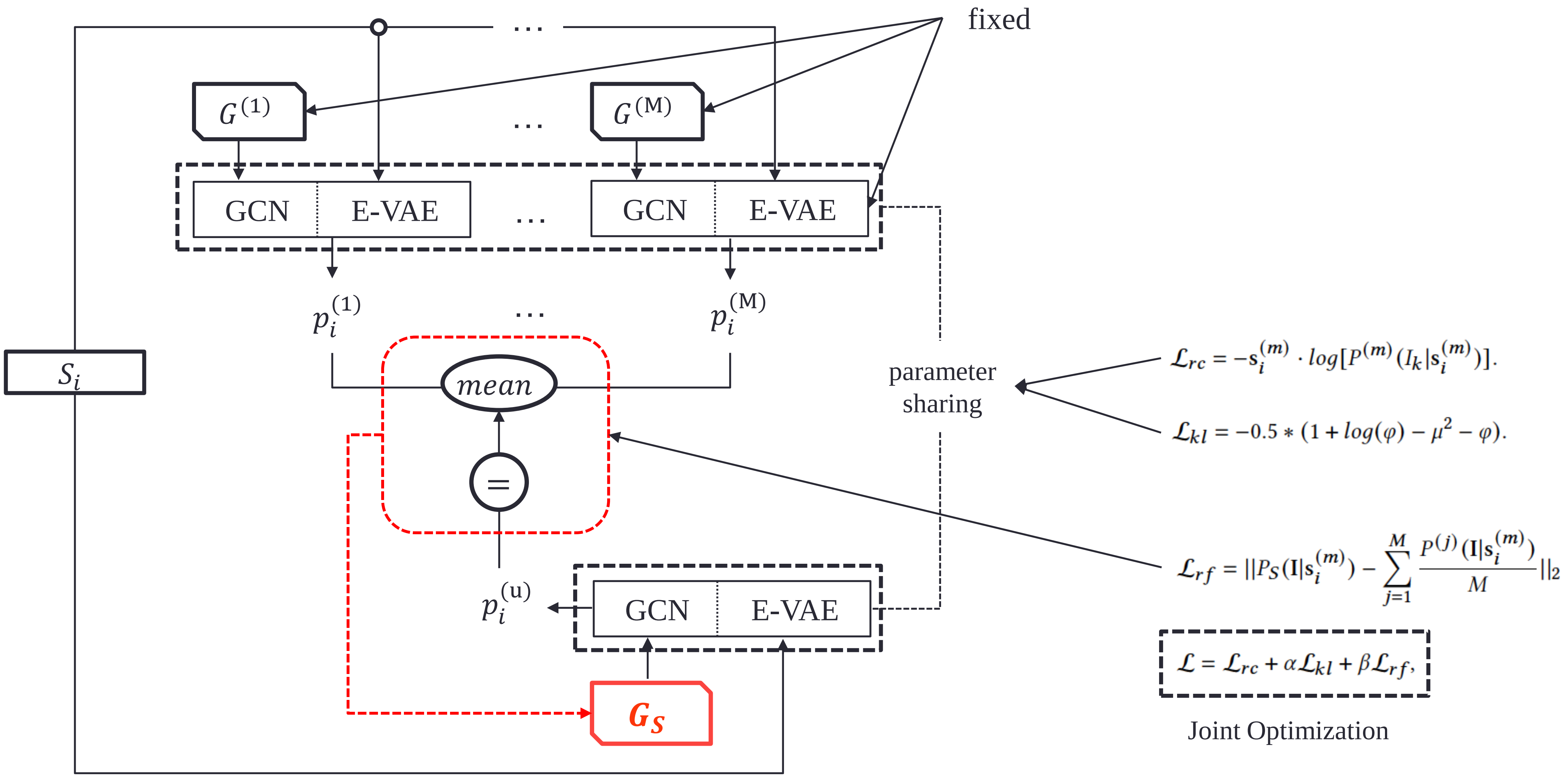
- High Ordered
- Non-Linear
- Large Scale



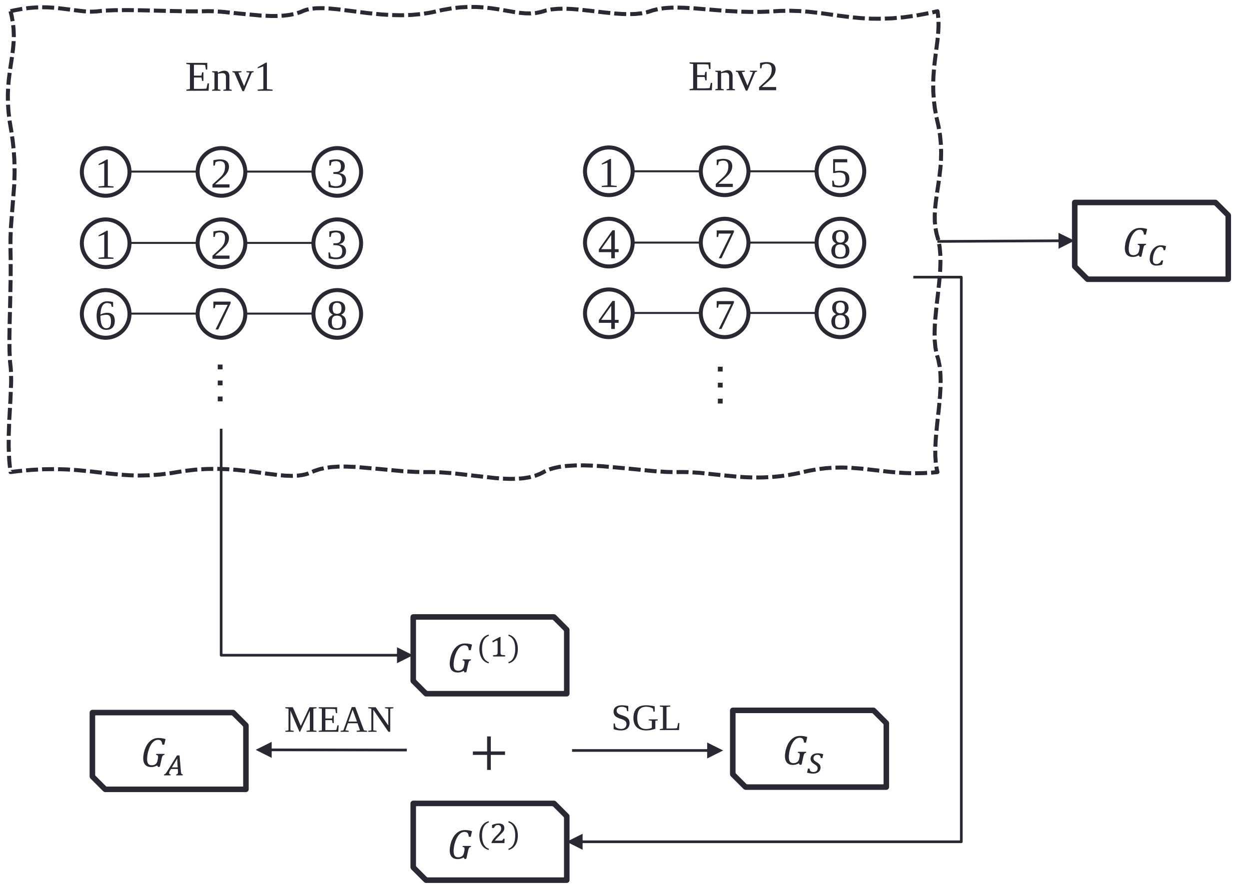
Algorithm: Graph Based Set Generation in Single Environment



Algorithm: Stable Graph Learning from Multiple Environment



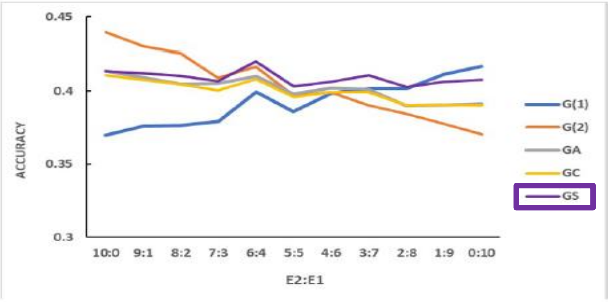
Experiment: Simulation Data



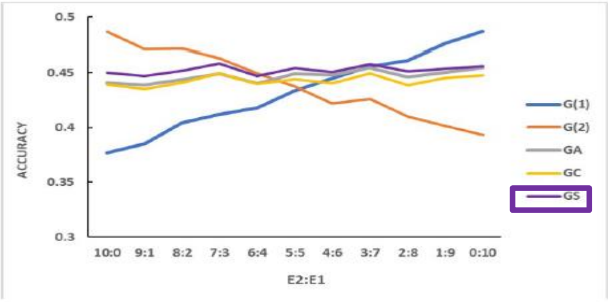
MEAN of ACCURACY				
	$q_0 = 1$			
	$q_1 = 2, q_2 = 4$	$q_1 = 2, q_2 = 6$	$q_1 = 3, q_2 = 6$	$q_1 = 3, q_2 = 9$
$G^{(1)}$	39.24%	43.23%	44.03%	45.18%
$G^{(2)}$	40.36%	43.93%	44.29%	45.34%
G_A	40.14%	44.69%	43.94%	47.62%
G_C	39.98%	44.28%	44.38%	46.96%
G_S	**40.91%**	**45.27%**	**45.02%**	**48.38%**
	$q_0 = 3$			
	$q_1 = 2, q_2 = 4$	$q_1 = 2, q_2 = 6$	$q_1 = 3, q_2 = 6$	$q_1 = 3, q_2 = 9$
$G^{(1)}$	39.58%	40.31%	43.70%	44.97%
$G^{(2)}$	39.43%	40.89%	43.67%	43.78%
G_A	38.40%	39.92%	44.02%	46.64%
G_C	38.68%	40.34%	43.23%	46.68%
G_S	**39.90%**	**41.45%**	**44.99%**	**48.79%**

11 testing datasets: mixing of Env1 and Env2 (10:0 to 0:10)

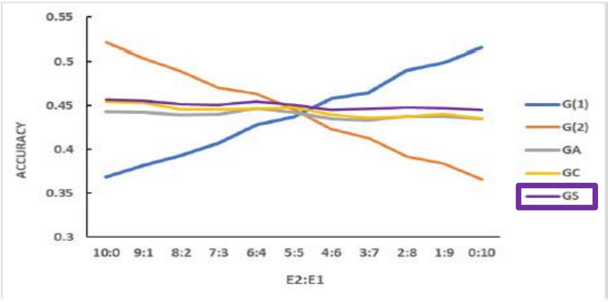
Experiment: Simulation Data



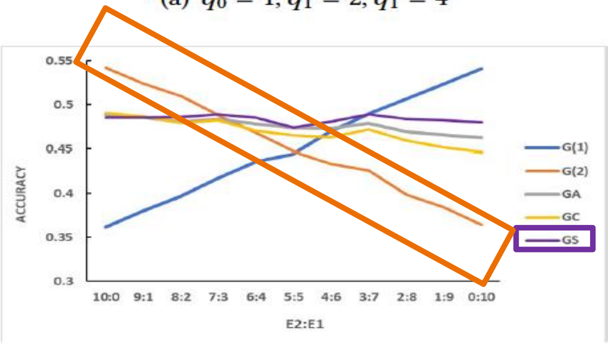
(a) $q_0 = 1, q_1 = 2, q_1 = 4$



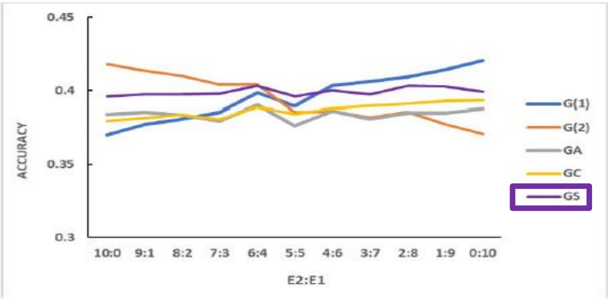
(b) $q_0 = 1, q_1 = 2, q_2 = 6$



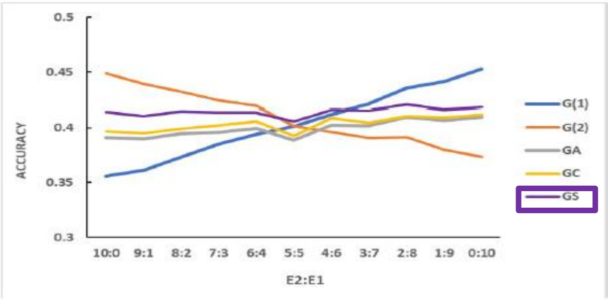
(c) $q_0 = 1, q_1 = 3, q_2 = 6$



(d) $q_0 = 1, q_1 = 3, q_2 = 9$

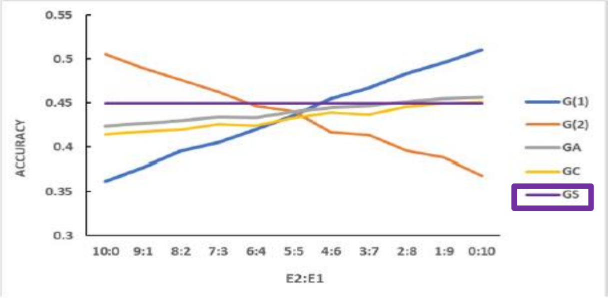


(e) $q_0 = 3, q_1 = 2, q_1 = 4$

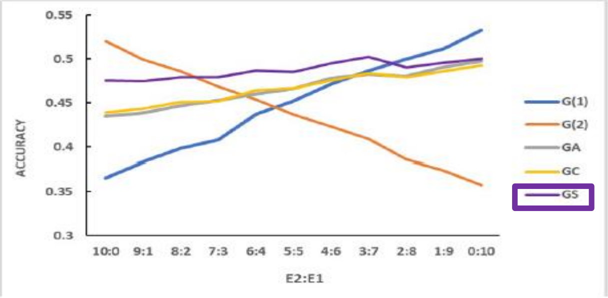


(f) $q_0 = 3, q_1 = 2, q_2 = 6$

Stability Improvement



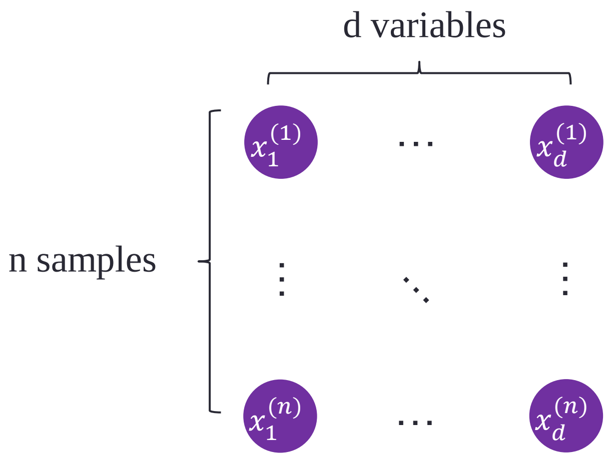
(g) $q_0 = 3, q_1 = 3, q_2 = 6$



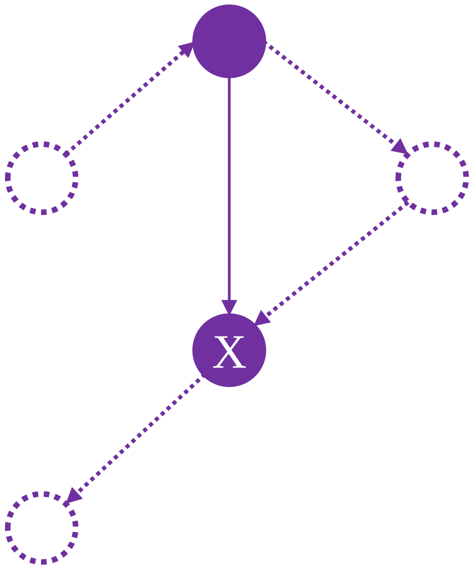
(h) $q_0 = 3, q_1 = 3, q_2 = 9$

Causal Graph--- stable graph structure

- Causal Discovery Problem



Observation Data



Causal Graph

- Functional Causal Models (FCMs)
  - Additional Noise Model

$$\mathrm{x} = f_{\mathrm{x}}\big(P_G(\mathrm{x})\big) + \epsilon_{\mathrm{x}}$$

exogenous noise

  - Linear Model

$$X = WX + \epsilon$$

Continuous Optimization for Structure Learning

- DAG Constraint

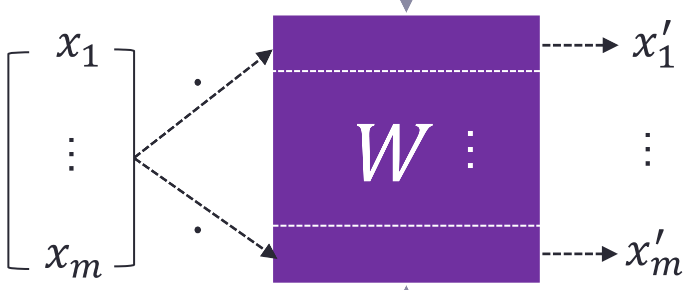
Theorem 1. *A matrix $W \in \mathbb{R}^{d \times d}$ is a DAG if and only if*

$$h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - d = 0, \tag{5}$$

where $\circ$ is the Hadamard product and e^A is the matrix exponential of A. Moreover, $h(W)$ has a simple gradient

$$\nabla h(W) = \left(e^{W \circ W}\right)^T \circ 2W, \tag{6}$$

and satisfies all of the desiderata (a)-(d).



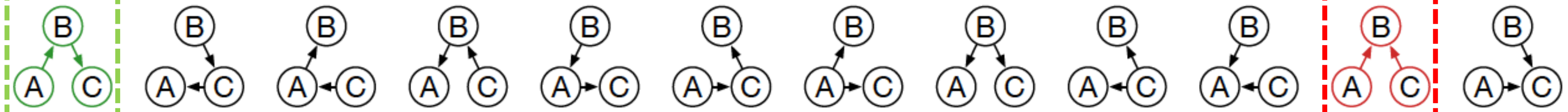
Reconstruction Constraint

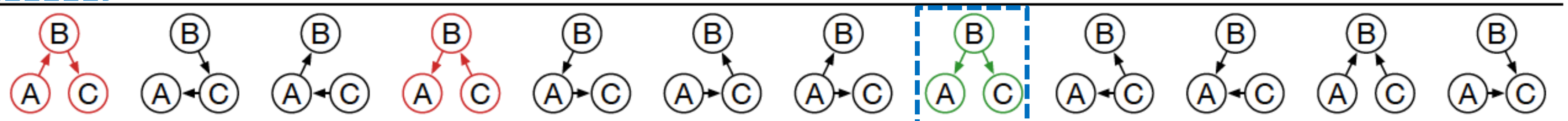
$$\min \sum_{i=1}^{m} E(Rx_i{}^2) \qquad Rx_i = x_i - x'_i$$

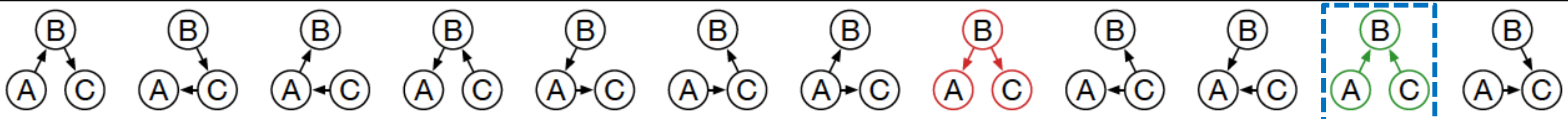
Sparsity Constraint l_1 *or* l_2

Inconsistency between Reconstruction and Causality

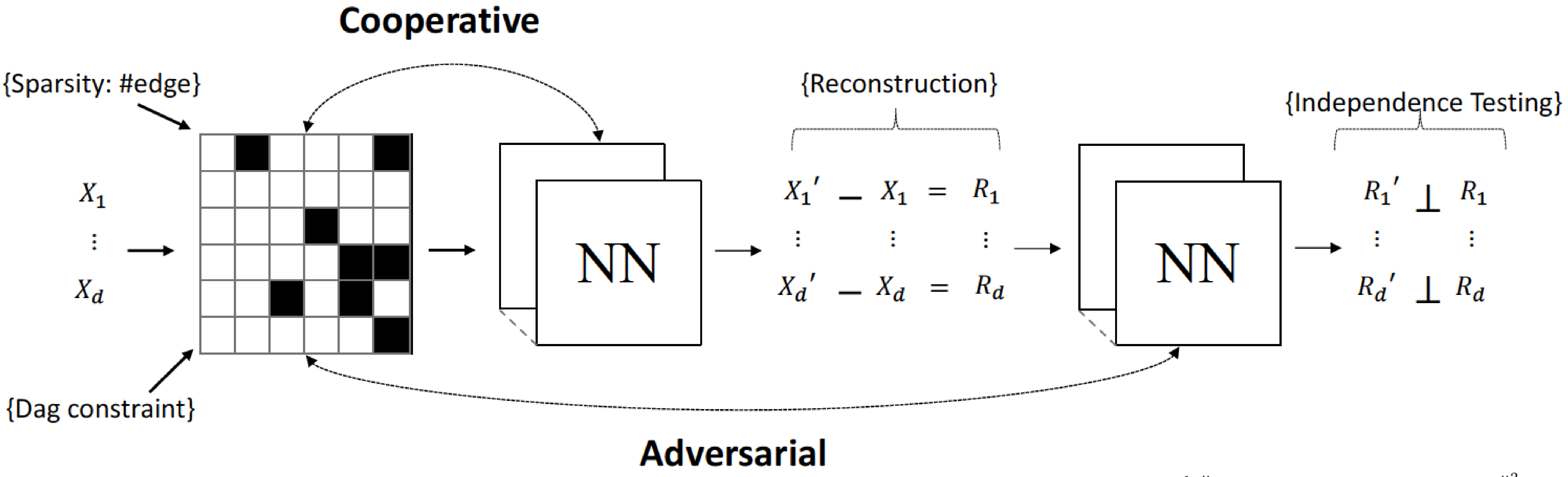
Table 1: Chain example: $A = \epsilon_A(\sim \mathcal{N}(0,1))$, $B = A + \epsilon_B(\sim \mathcal{N}(0,4))$, $C = B/5 + \epsilon_C(\sim \mathcal{N}(0,1))$. **Fork example:** $B = \epsilon_B(\sim \mathcal{U}(-2,2))$, $A = B/2 + \epsilon_A(\sim \mathcal{U}(-1,1))$, $C = B/2 + \epsilon_C(\sim \mathcal{U}(-1,1))$. **Collider example:** $A = \epsilon_A(\sim \mathcal{N}(0,1))$, $C = \epsilon_C(\sim \mathcal{N}(0,1))$, $B = A/3 + C/3 + \epsilon_B(\sim \mathcal{N}(0,1/9))$. **The graph in green lines denotes the ground truth, but the red one is the false structure learned by traditional differential FCMs (owing to the minimal reconstruction loss). Independence regularization can help to identify the true graph.**

Predicted Graph (Chain)												
Reconstruction Loss	6.00	6.97	6.17	6.17	6.96	6.33	6.16	6.80	6.33	7.00	5.65	7.00
Mutually Independent?	✓			✓				✓				

Predicted Graph (Fork)												
Reconstruction Loss	1.67	2.17	1.83	1.67	2.17	1.83	1.83	2.00	1.83	2.33	1.78	2.33
Mutually Independent?								✓				

Predicted Graph (Collider)												
Reconstruction Loss	1.89	2.00	2.22	1.89	2.00	2.22	2.22	1.67	2.22	1.83	2.11	1.83
Mutually Independent?											✓	

Algorithm---Differentiable Adversarial Causal Discovery



$$\min \quad \mathcal{L} = \mathcal{L}_{\text{rec}}(G, \mathbf{X}, \theta) + \alpha \mathcal{L}_{\text{DAG}}(G) + \beta \mathcal{L}_{\text{sparse}}(G) + \gamma \mathcal{L}_{\text{M}}(X - f(X, \theta), \phi).$$

$$\max_{\phi} \mathcal{L}_{\text{M}}(R, \phi) = \sum_{i=1}^{d} \left\| \frac{\text{Cov}[\text{MLP}(R_{-i}, \phi_i), R_i]}{\sqrt{\text{Var}[\text{MLP}(R_{-i}, \phi_i)]} \cdot \sqrt{\text{Var}[R_i]}} \right\|_2^2 .$$

Wild Scenario

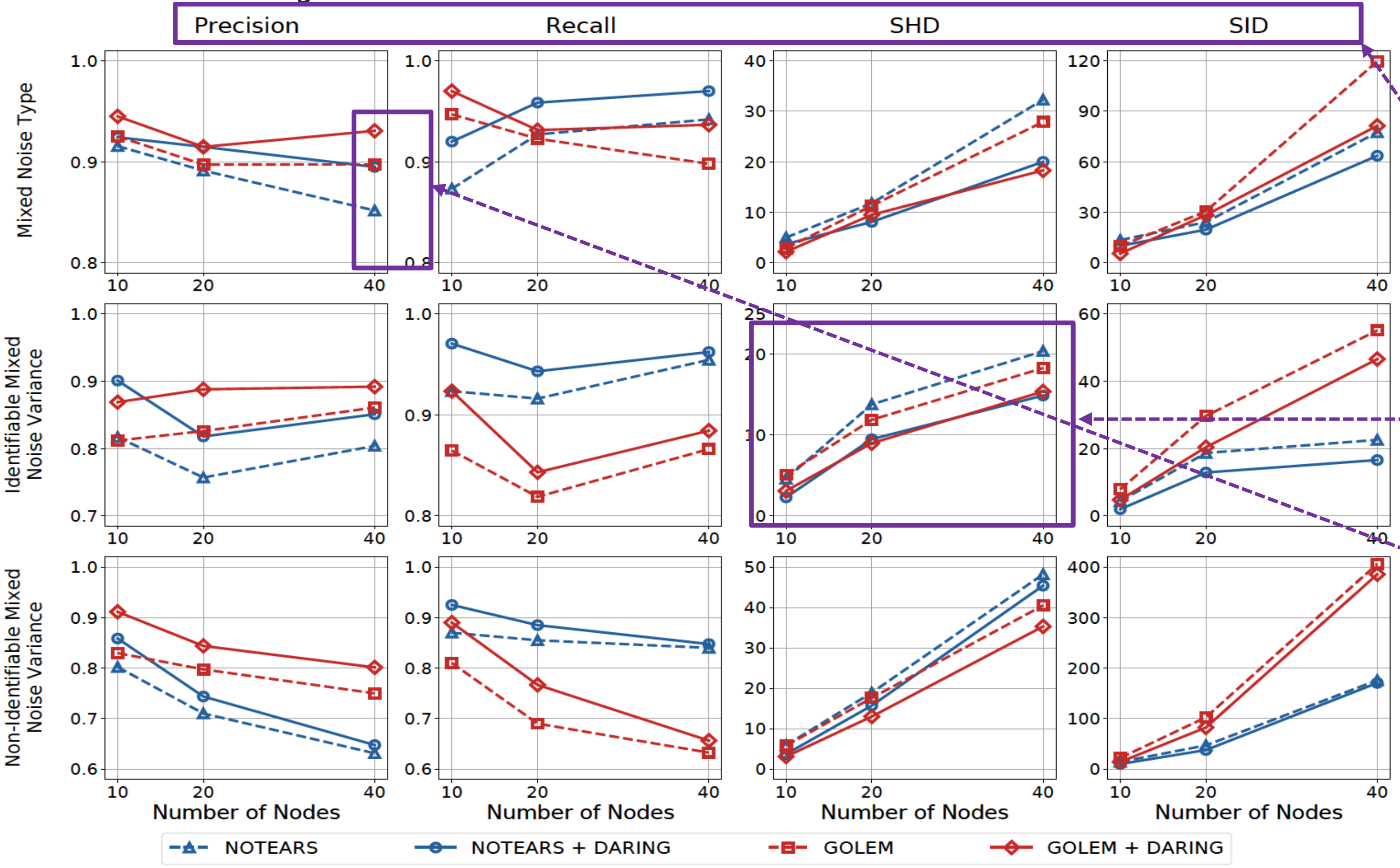
exogenous noise

Small and Equal Variance & Single Type √	Large and Unequal Variance & Single Type √
Small and Equal Variance & Various Type √	**Large and Unequal Variance** **& Various Type** √

Realistic

Simulation Experiment

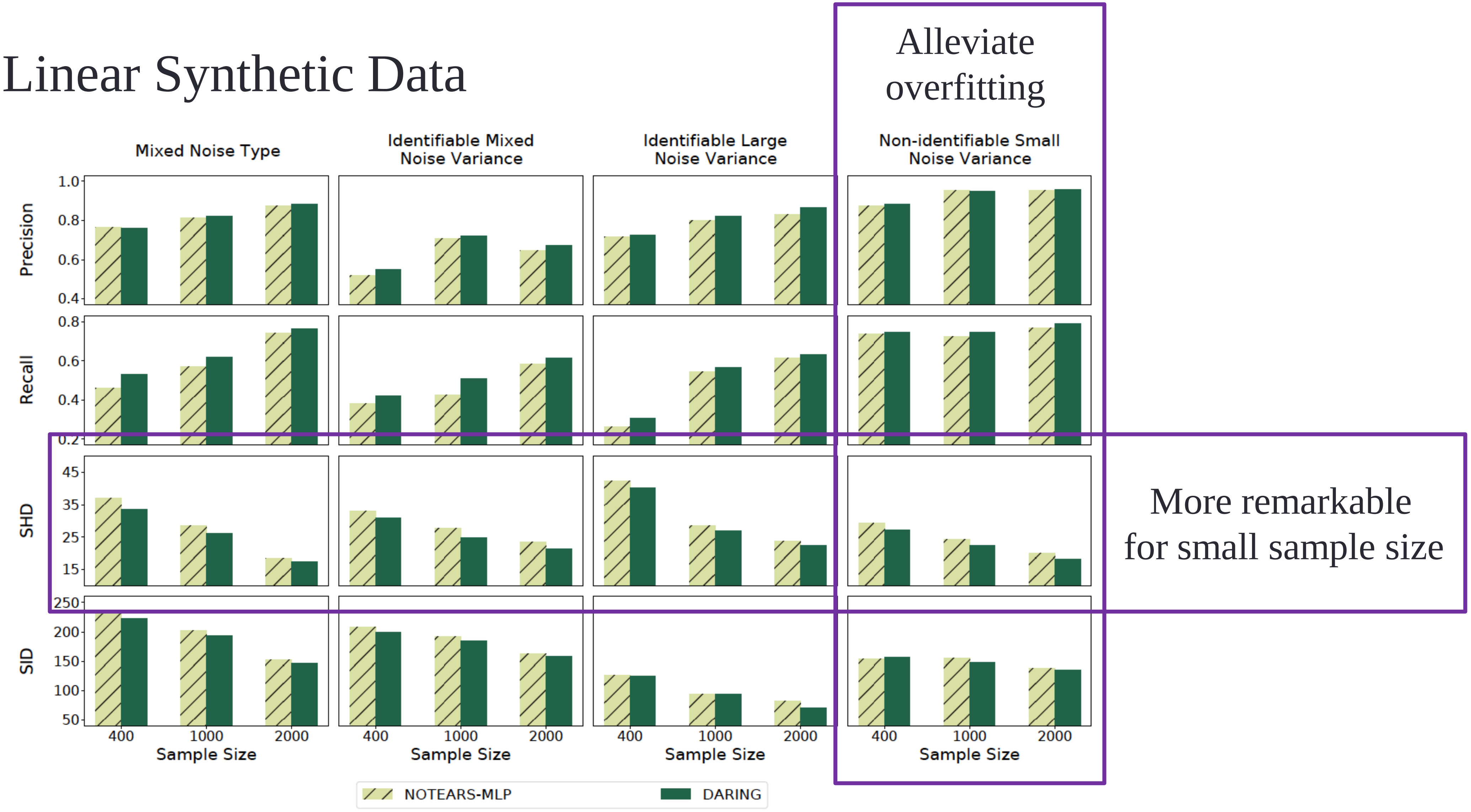
- Linear Synthetic Data



- Promote and achieve best performance for all metrics
  - Global optimization
- More remarkable for large scale
  - Availability
- Make up gap of baseline models
  - Robustness

Simulation Experiment

- Non-Linear Synthetic Data



Outline

- Brief introduction to causal inference
- Stable learning and its development
- Positioning stable learning in OOD generalization
- Benchmark and dataset

Problem Definition

Problem 1 (Supervised Learning). *Given a set of n training samples of the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ which are drawn from training distribution* $\boxed{P_{tr}(X, Y)}$*, a supervised learning problem is to find an optimal model f_θ^* which can generalize best on data drawn from test distribution* $\boxed{P_{te}(X, Y)}$*:*

$$f_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{X,Y \sim P_{te}}[\ell(f_\theta(X), Y)]$$

Key Question: $\boldsymbol{P_{tr}(X, Y) \neq P_{te}(X, Y)}$

Categorization of OOD Methods

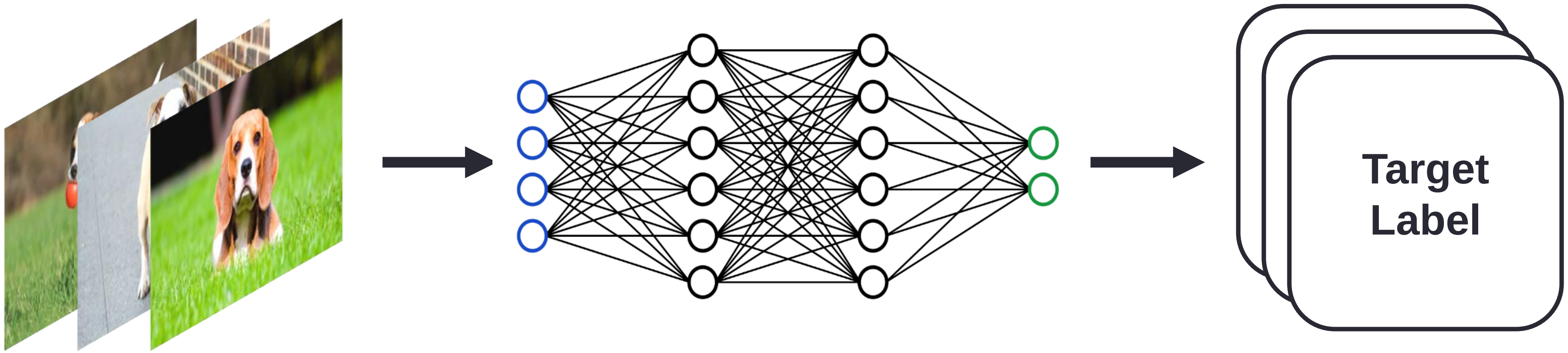
$$f_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{X,Y \sim P_{te}}[\ell(f_\theta(\boxed{X}), Y)]$$

Unsupervised Representation Learning

Categorization of OOD Methods

Stable Learning

$$f_\theta^* = \arg\min_{f_\theta} \mathbb{E}_{X,Y\sim P_{te}}[\ell(\boxed{f_\theta}(X), Y)]$$



Supervised Model Learning

Categorization of OOD Methods

Stable Learning

$$f_\theta^* = \boxed{\arg\min_{f_\theta} \mathbb{E}_{X,Y\sim P_{te}}[\ell}(f_\theta(X), Y)]$$

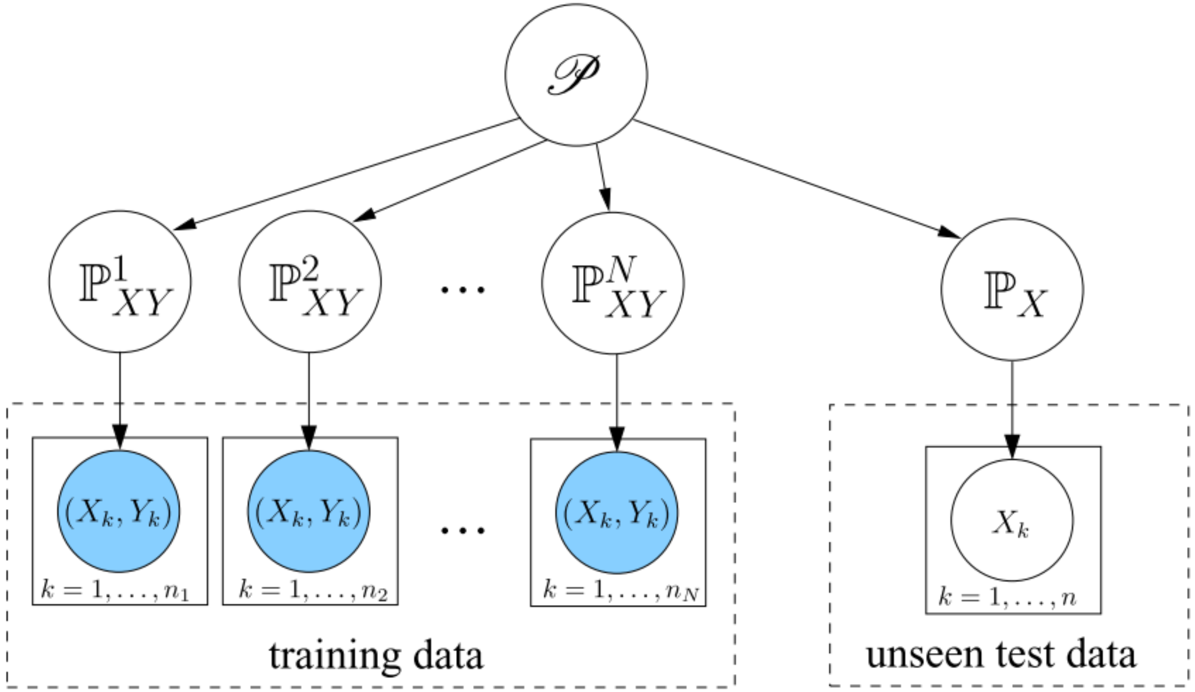
⇩

Optimization

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***
  - May help for robustness

Domain Generalization



- 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 $1-\delta$

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

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

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*

Distributionally Robust Optimization

- Problem Definition:

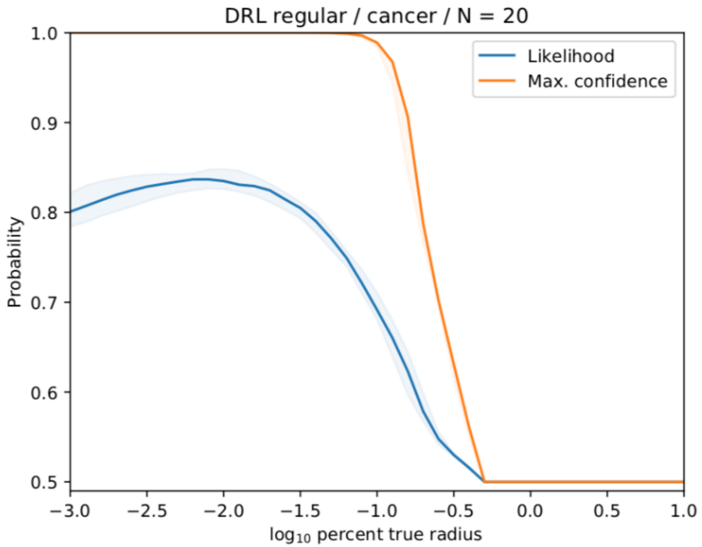
$$\underset{\theta \in \Theta}{minimize} \; \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$$

where $\mathcal{P}$ is a class of distributions around the data-generating distribution P_0

- Idea: if class $\mathcal{P}$ contains all distributions under shift-interventions or do-interventions, then causal parameter θ_{causal} is the distributionally robust parameter.

Over Pessimism Problem

- DRO has the over-pessimism problem.
  - When the radius of the $\mathcal{P}$ is large, the distribution set includes many unrealistic/useless cases, which will make the learned model refuse to make a decision in order to guarantee such a overwhelmingly-considered robustness.



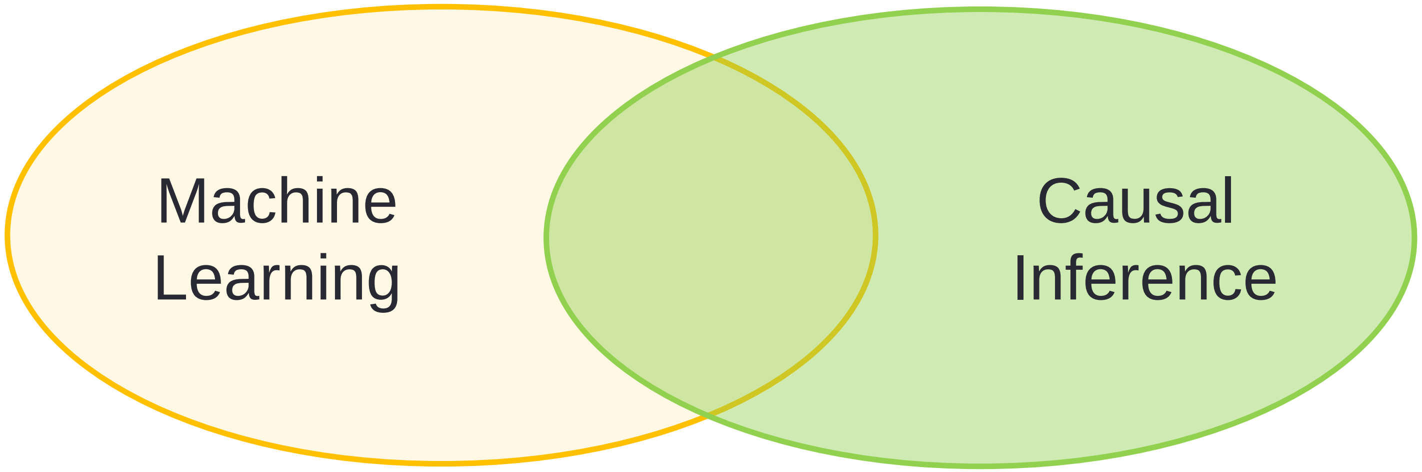
Assign equal probability to each class !

  - When the radius of the distribution set is too small, the distribution set may not contain the possible test distributions, resulting in an inability to guarantee the expected robustness.

Image from Frogner, C.; Claici, S.; Chien, E.; and Solomon, J. 2019. Incorporating Unlabeled Data into Distributionally Robust Learning.

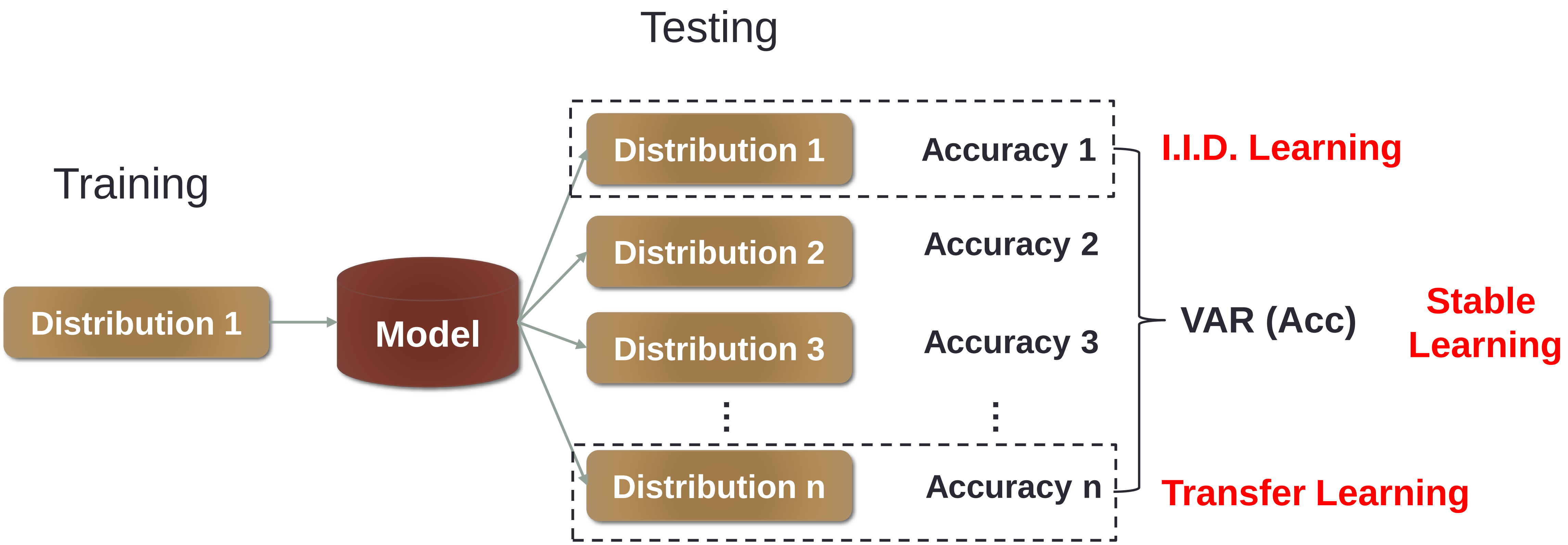
Stable Learning

- Finding the common ground between causal inference and machine learning



Stable Learning

- One training distribution, multiple testing distributions



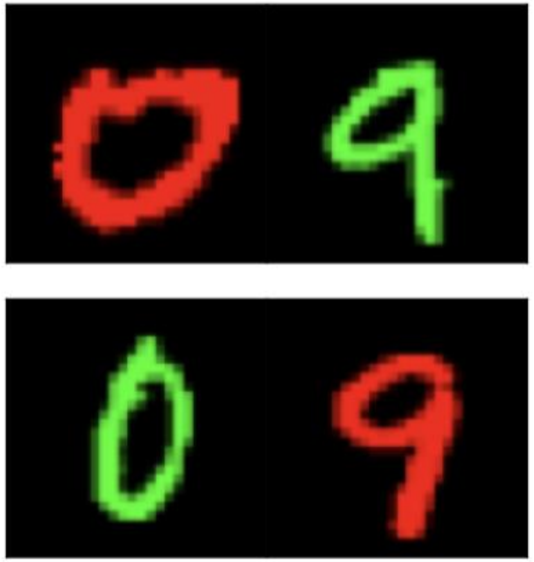
Outline

- Brief introduction to causal inference
- Stable learning and its development
- Positioning stable learning in OOD generalization
- Benchmark and dataset

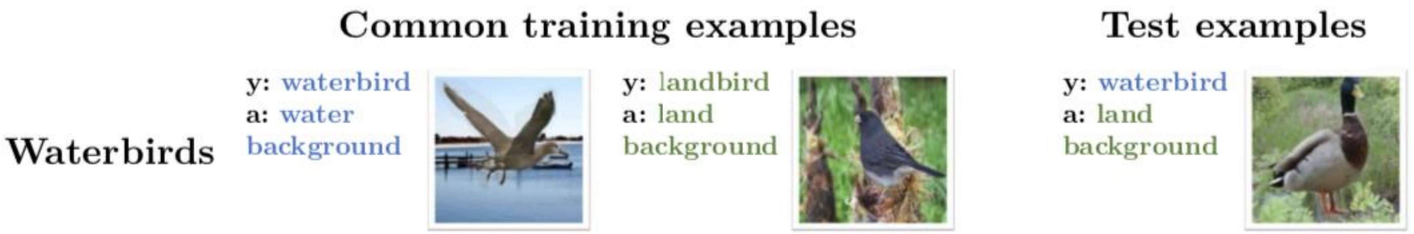
Image Dataset —— Synthetic Transformation

Training

Test



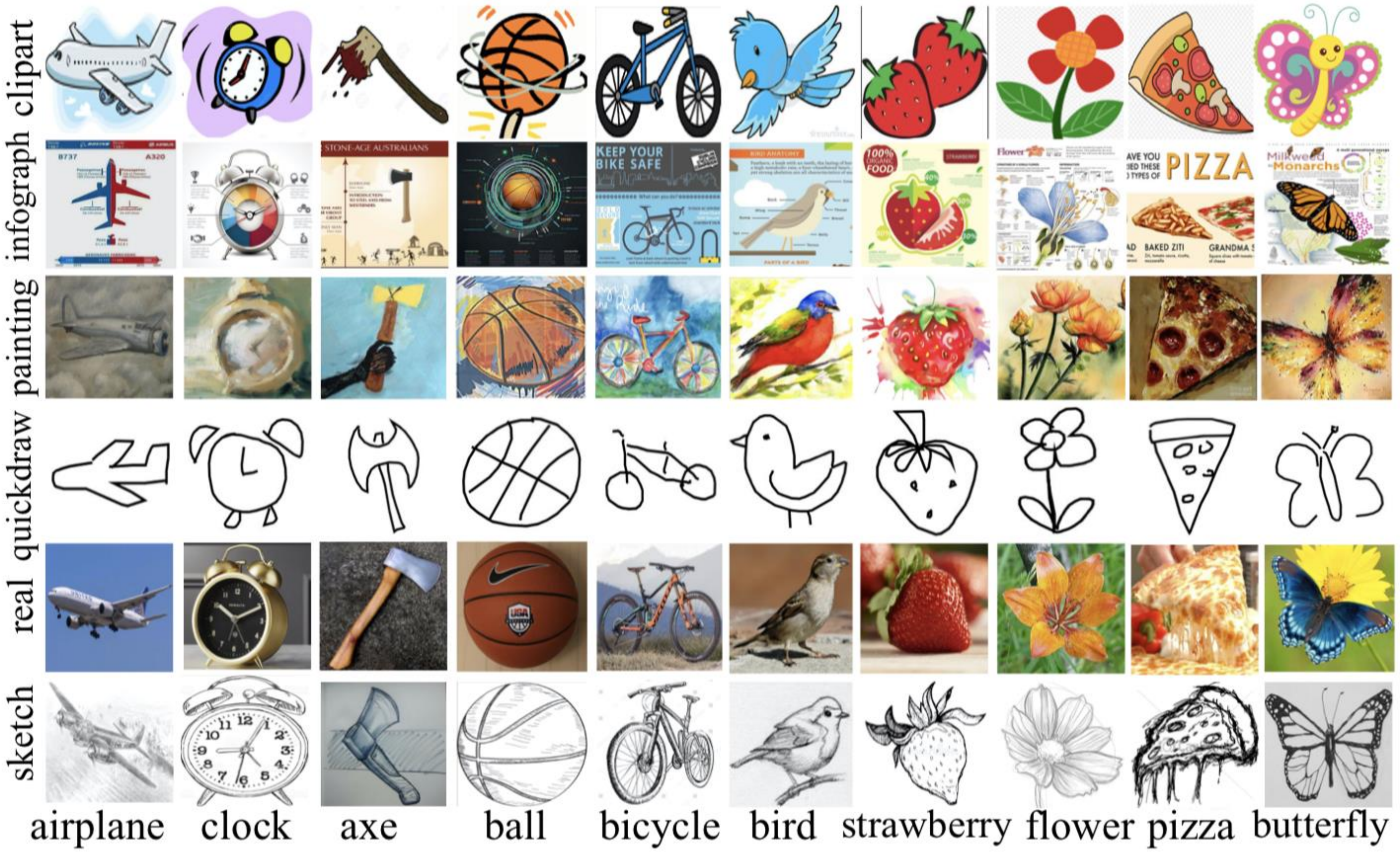
Colored MNIST[1]



Waterbirds[2]

[1] Ye, N., Li, K., Hong, L., Bai, H., Chen, Y., Zhou, F., & Li, Z. (2021). OoD-Bench: Benchmarking and Understanding Out-of-Distribution Generalization Datasets and Algorithms. *arXiv preprint arXiv:2106.03721*.
[2] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Image Dataset —— Multi-Style

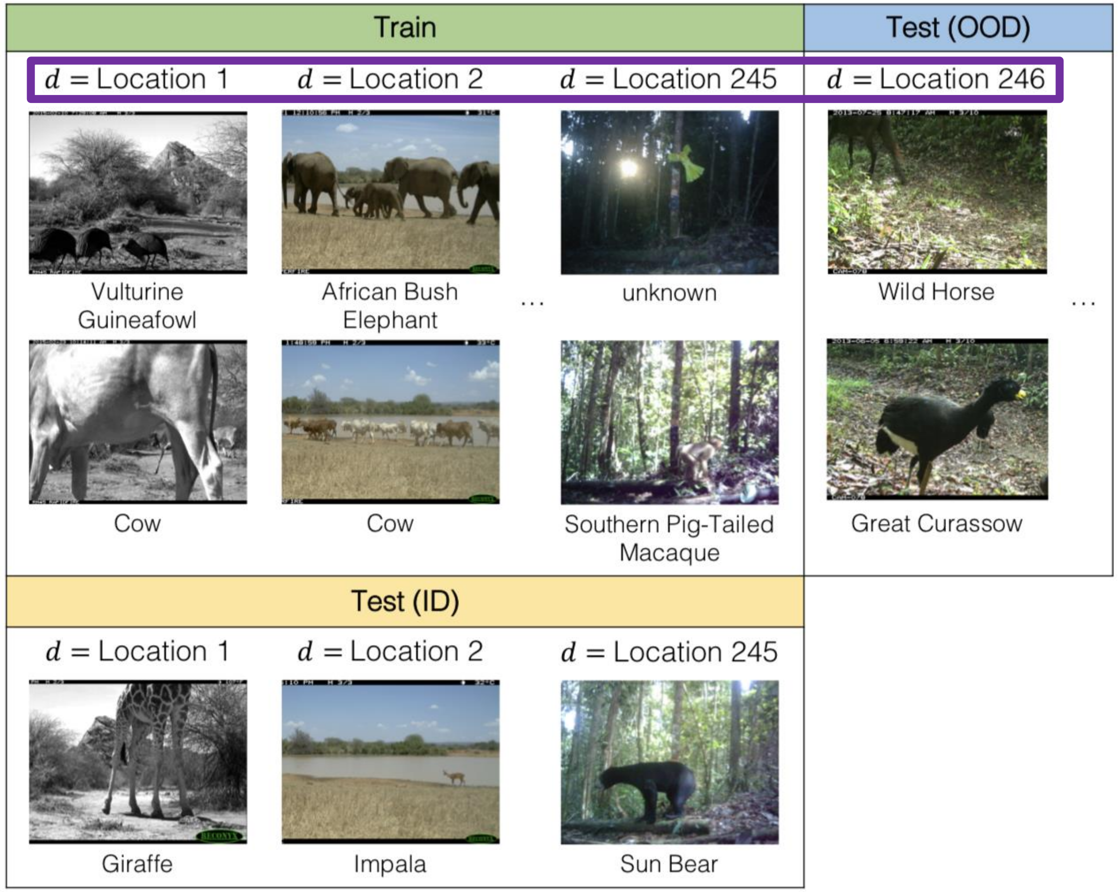


DomainNet

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1406-1415).

Image Dataset —— Fixed Wild Data

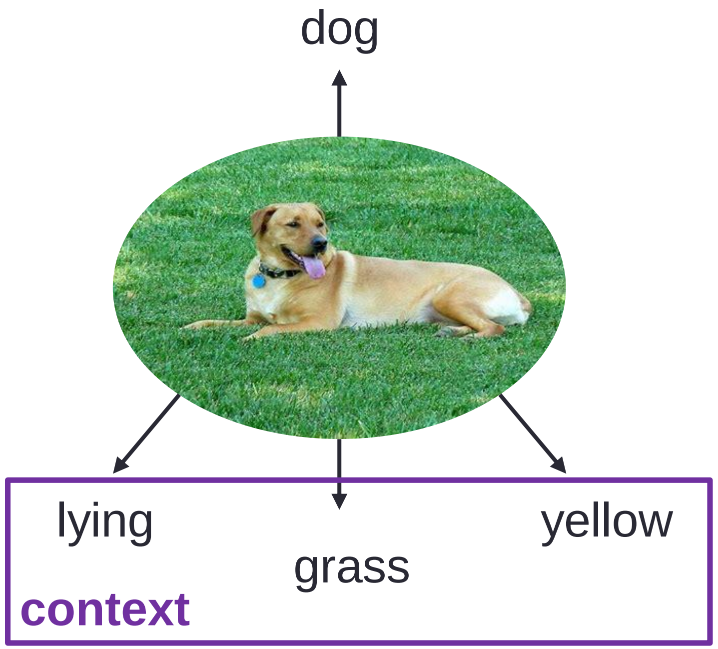
camera at

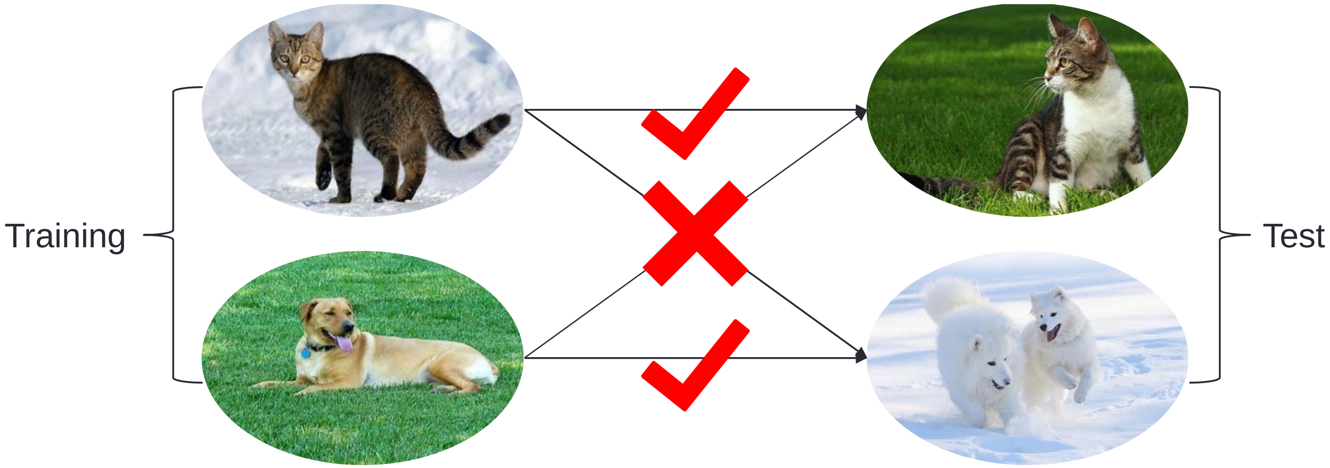


iWildCam[1]

[1] Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* (pp. 5637-5664). PMLR.

Image Dataset —— Controllable Wild Data



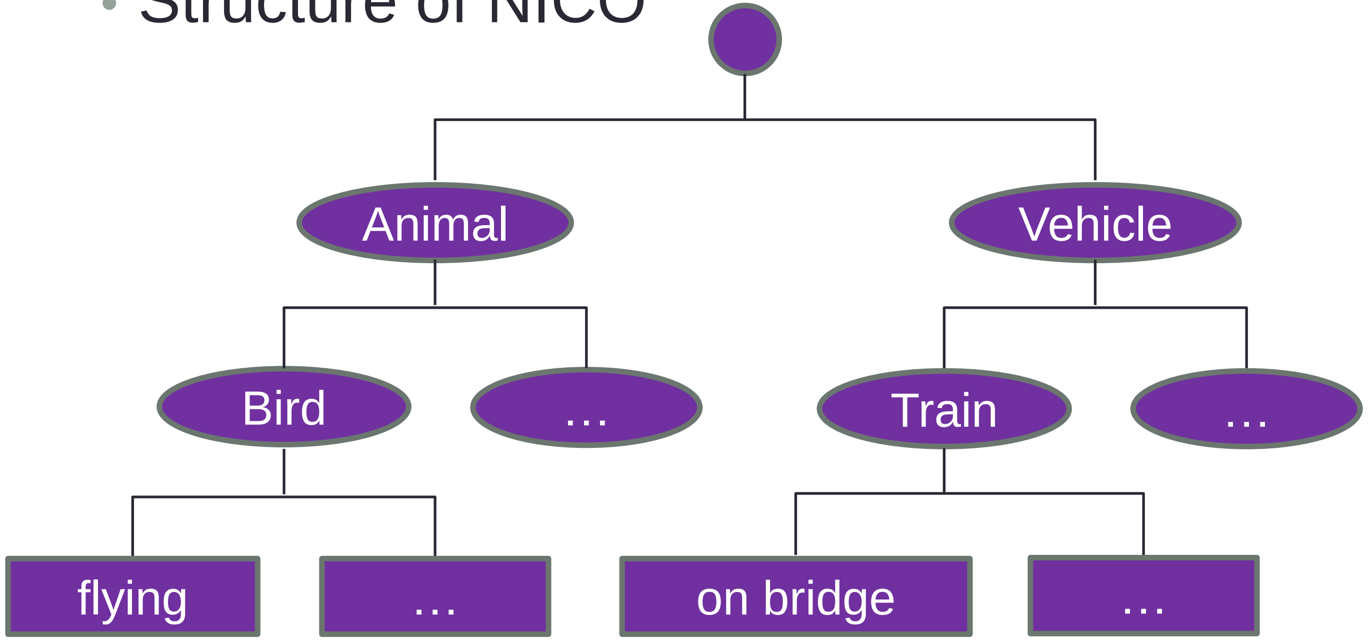


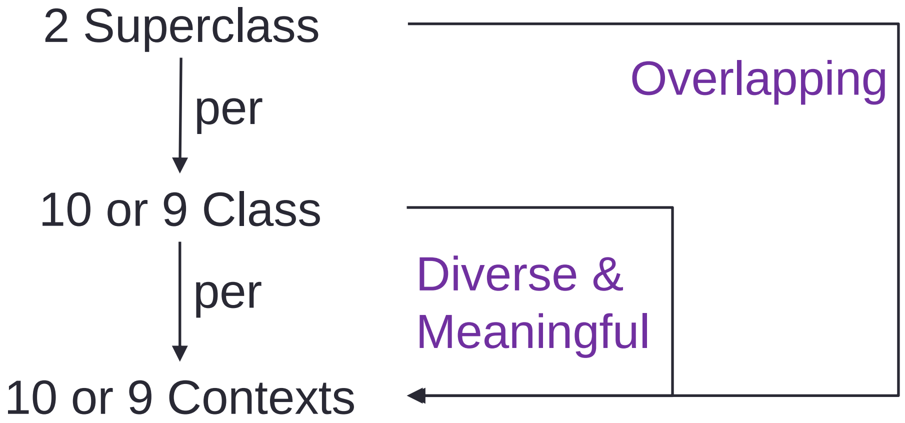
NICO[1] (Non-I.I.D. Image Dataset with Contexts)

[1] He, Y., Shen, Z., & Cui, P. (2021). Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, *110*, 107383.

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

- Contextual labels (Contexts)
  - the attributes or actions of a category
    - e.g. white bear, double decker
  - the background or scene of a category
    - e.g. cat on snow, airplane in sunrise
- Structure of NICO





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

- Data size of each class in NICO

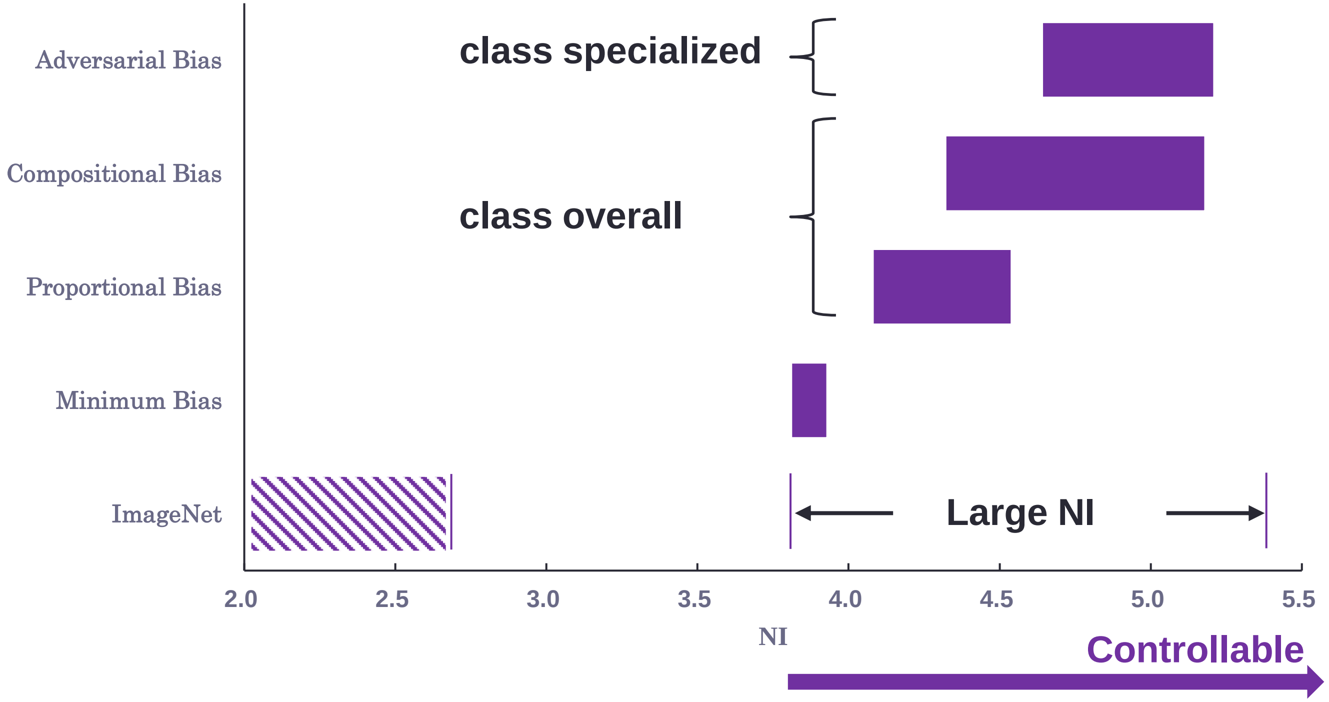
Animal	Data Size	*Vehicle*	Data Size
Bear	1609	Airplane	930
Bird	1590	Bicycle	1639
Cat	1479	Boat	2156
Cow	1192	Bus	1009
Dog	1624	Car	1026
Elephant	1178	Helicopter	1351
Horse	1258	Motorcycle	1542
Monkey	1117	Train	750
Rat	846	Truck	1000
Sheep	918		

- Samples with contexts in NICO

[1] He, Y., Shen, Z., & Cui, P. (2021). Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, *110*, 107383.

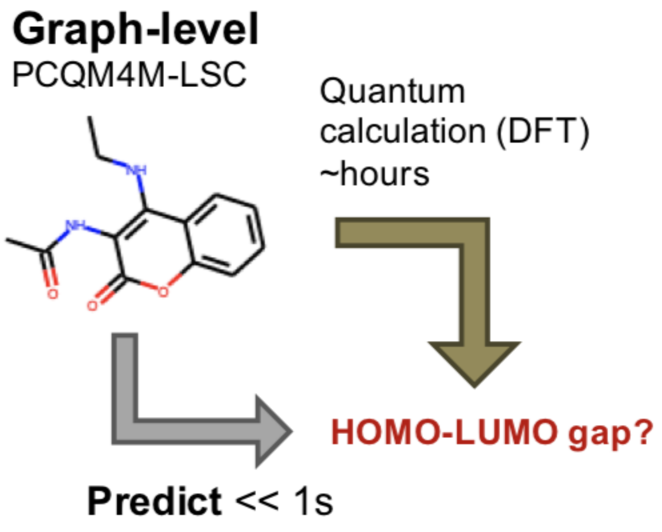
NICO——Non-I.I.D. Image Dataset with Contexts

- Range of average NI over Animal superclass for different settings supported in NICO.



[1] He, Y., Shen, Z., & Cui, P. (2021). Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, *110*, 107383.

Other Data Type



(c) PCQM4M−LSC

Graph Data (OGB-LSC[1])

	Reviewer ID (**d**)	Review Text (**x**)	Stars (**y**)
Train	Reviewer 1	They are decent shoes. Material quality is good but the color fades very quickly. Not as black in person as shown.	5
		Super easy to put together. Very well built.	5
	Reviewer 2	This works well and was easy to install. The only thing I don't like is that it tilts forward a little bit and I can't figure out how to stop it.	4
		Perfect for the trail camera	5
		…	
	Reviewer 10,000	I am disappointed in the quality of these. They have significantly deteriorated in just a few uses. I am going to stick with using foil.	1
		Very sturdy especially at this price point. I have a memory foam mattress on it with nothing underneath and the slats perform well.	5
Test	Reviewer 10,001	Solidly built plug in. I have had 4 devices plugged in and all charge just fine.	5
		Works perfectly on the wall to hang our wreath without having to do any permanent damage.	5
		…	

Text Data (Amazon Review[2])

[1] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., ... & Leskovec, J. (2020). Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*.

[2] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

OOD Evaluation Metric

Average Accuracy

$$\overline{Acc} = \frac{1}{K}\sum_{k=1}^{K} acc_k$$

performance in k_{th} environment

Standard Deviation (STD)

$$ACC_{std} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(acc_k - \overline{Acc})^2}$$

Worst-Case Accuracy

$$ACC_{worst} = \min_{k\in[K]} acc_k$$

Conclusions

- Stable Learning: finding the common ground between causal inference and machine learning
  - StableNet demonstrates its capacity and power in CNN networks
- Rethink the risk minimization framework
  - HRM: heterogeneity + invariance

Conclusions

- ***Explainability***, ***Stability***, ***Fairness***, ***Verifiability*** problems are becoming more critical
- They are not independent!
- Stable Learning: finding the common ground between causal inference and machine learning
  - Theoretical problems
  - Sample efficiency problems
  - Application problems

A survey on OOD generalization

Towards Out-Of-Distribution Generalization: A Survey

Zheyan Shen*, Jiashuo Liu*, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, Peng Cui†, *Senior Member, IEEE*

Abstract—Classic machine learning methods are built on the $i.i.d.$ assumption that training and testing data are independent and identically distributed. However, in real scenarios, the $i.i.d.$ assumption can hardly be satisfied, rendering the sharp drop of classic machine learning algorithms' performances under distributional shifts, which indicates the significance of investigating the Out-of-Distribution generalization problem. Out-of-Distribution (OOD) generalization problem addresses the challenging setting where the testing distribution is unknown and different from the training. This paper serves as the first effort to systematically and comprehensively discuss the OOD generalization problem, from the definition, methodology, evaluation to the implications and future directions. Firstly, we provide the formal definition of the OOD generalization problem. Secondly, existing methods are categorized into three parts based on their positions in the whole learning pipeline, namely unsupervised representation learning, supervised model learning and optimization, and typical methods for each category are discussed in detail. We then demonstrate the theoretical connections of different categories, and introduce the commonly used datasets and evaluation metrics. Finally, we summarize the whole literature and raise some future directions for OOD generalization problem. The summary of OOD generalization methods reviewed in this survey can be found at http://out-of-distribution-generalization.com.

Index Terms—Out-of-Distribution Generalization, Causal Inference, Invariant Learning, Stable Learning, Representation Learning, Distributionally Robust Optimization

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, Peng Cui. Towards Out-Of-Distribution Generalization: A Survey. *arxiv*, 2021.
http://out-of-distribution-generalization.com/

Reference

- Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Kernelized Heterogeneous Risk Minimization, ***NeurIPS***, 2021.
- Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, Peng Cui. Towards Out-Of-Distribution Generalization: A Survey. arxiv, 2021.
- Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.
- Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, Zheyan Shen. Deep Stable Learning for Out-Of-Distribution Generalization. ***CVPR***, 2021
- Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, Yishi Lin. Stable Adversarial Learning under Distributional Shifts. ***AAAI***, 2021.
- Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, Yue He. Counterfactual Prediction for Bundle Treatments. ***NeurIPS***, 2020.
- Zheyean Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li and Zhitang Chen. Stable Learning via Differentiated Variable Decorrelation. ***KDD***, 2020.
- Yue He, Zheyan Shen, Peng Cui. Towards Non-I.I.D. Image Classification: A Dataset and Baselines. ***Pattern Recognition***, 2020.
- Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning via Sample Reweighting. ***AAAI***, 2020.
- Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, Bo Li. Stable Prediction with Model Misspecification and Agnostic Distribution Shift. ***AAAI***, 2020.
- Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.
- Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM Multimedia***, 2018.

Thanks!

Peng Cui

cuip@tsinghua.edu.cn

http://pengcui.thumedialab.com