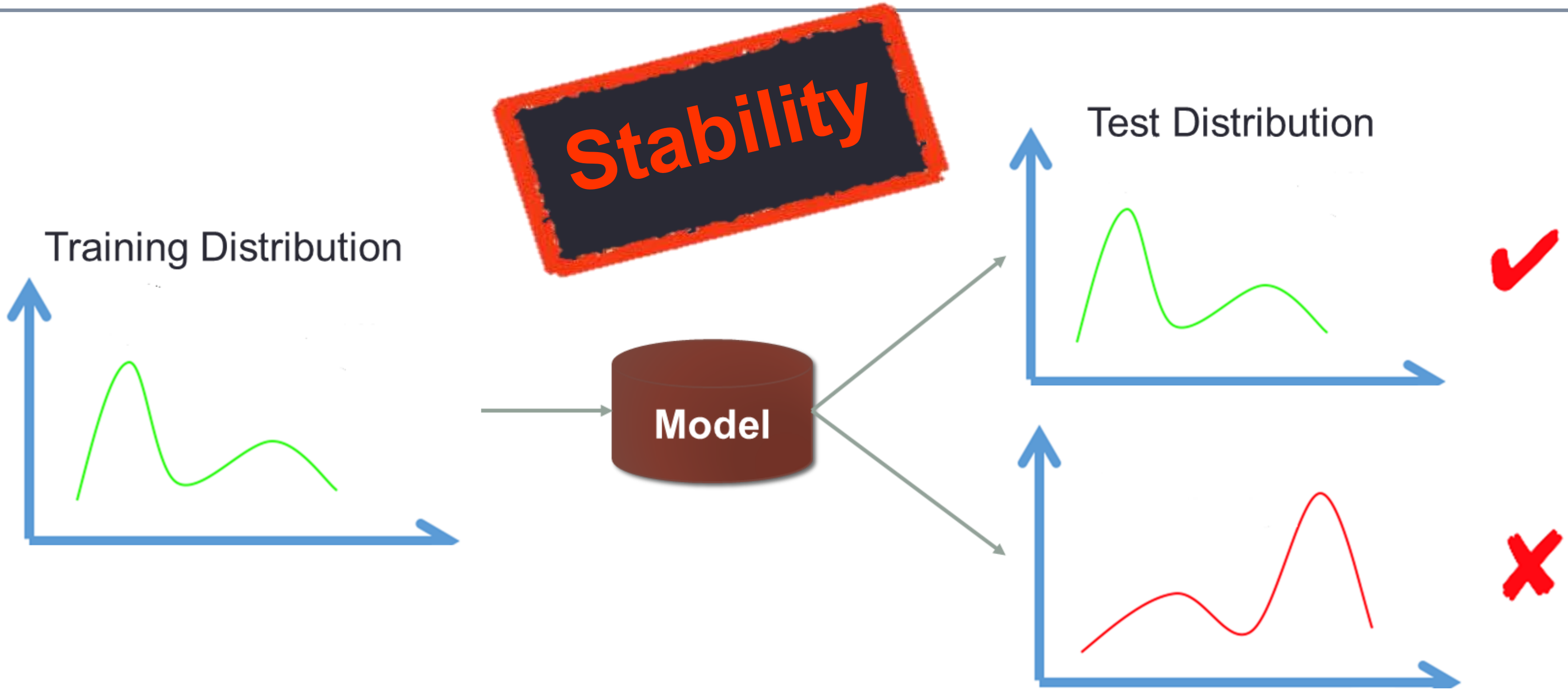
因果启发的稳定学习

崔 鹏
清华大学

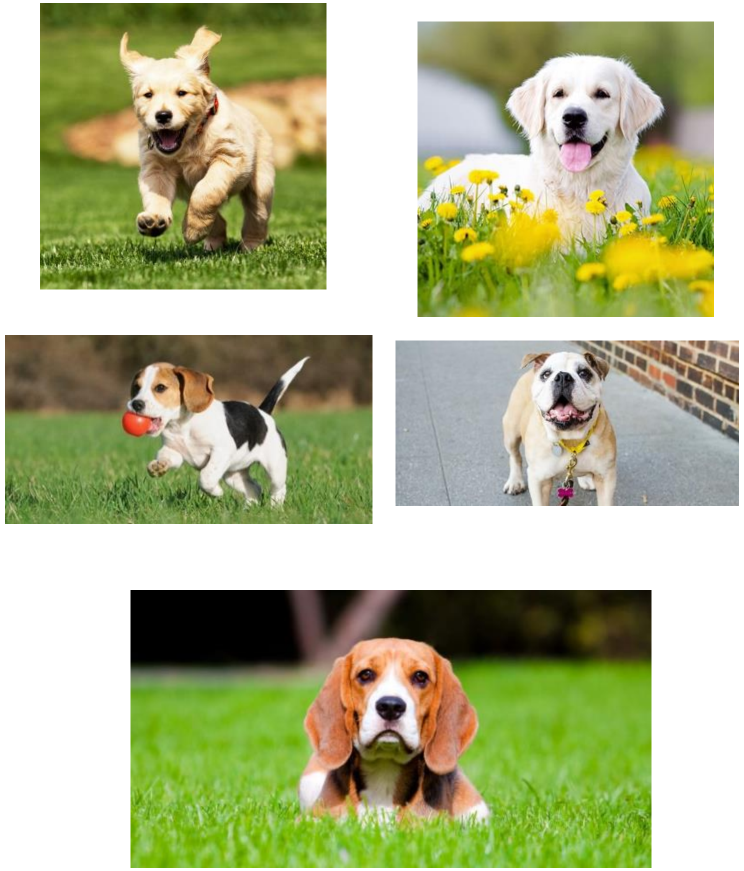
Risks of Today's AI Algorithms

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



OOD Generalization Problem

Risks of Today's AI Algorithms



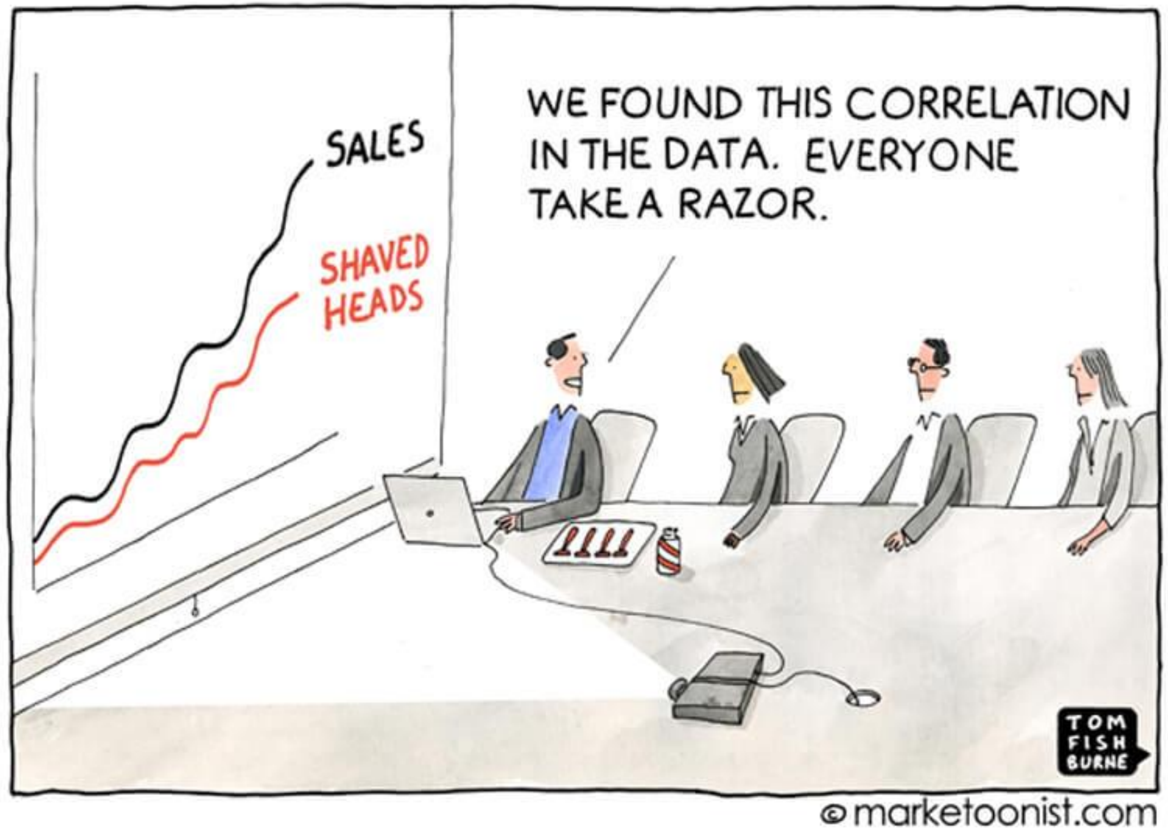
Yes

Maybe

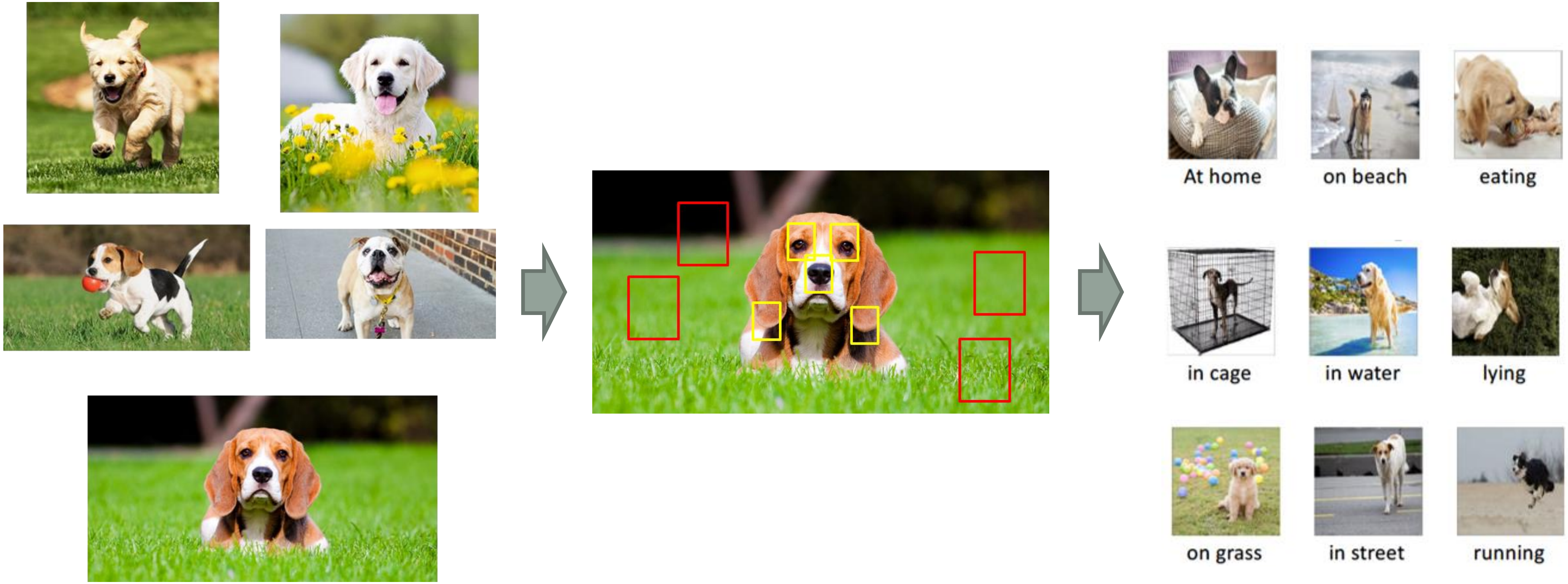
No

A plausible reason: ***Correlation***

Correlation is the very basics of machine learning.

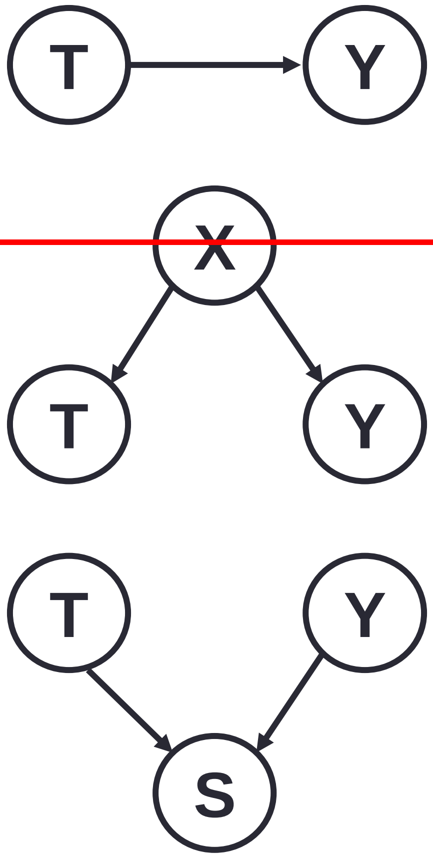


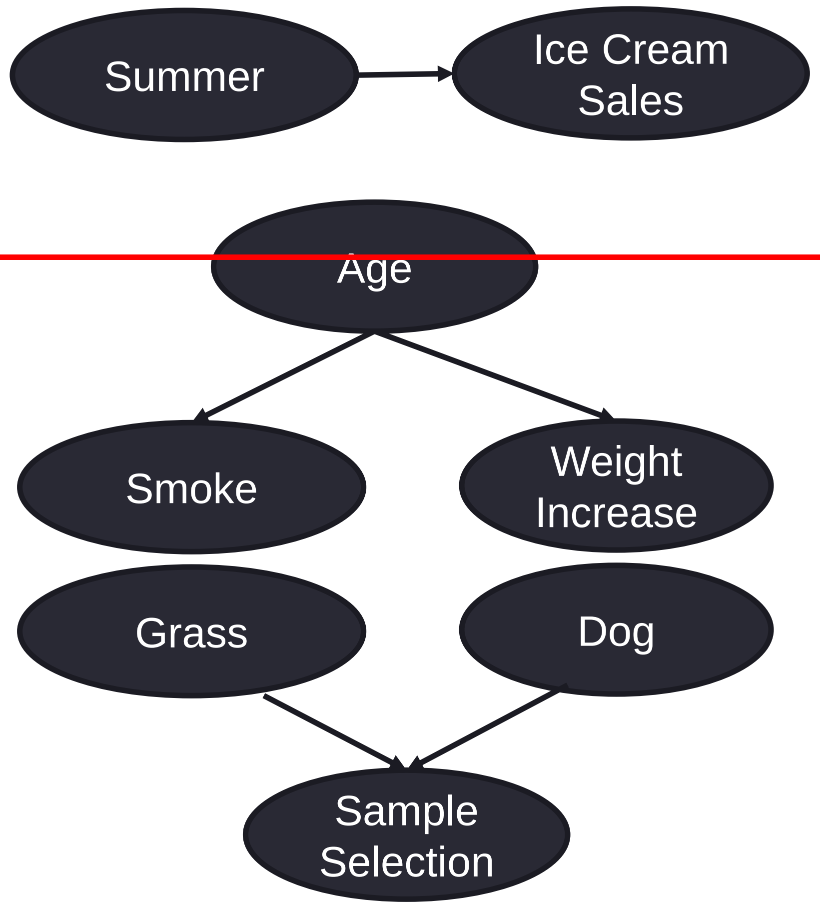
Correlation is ‘*unstable*’



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



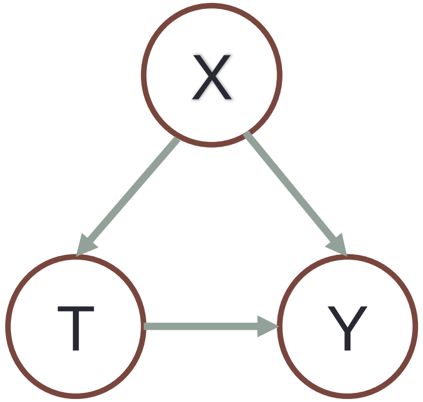


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



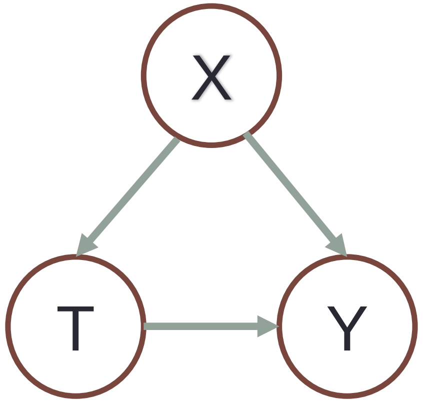
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

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

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

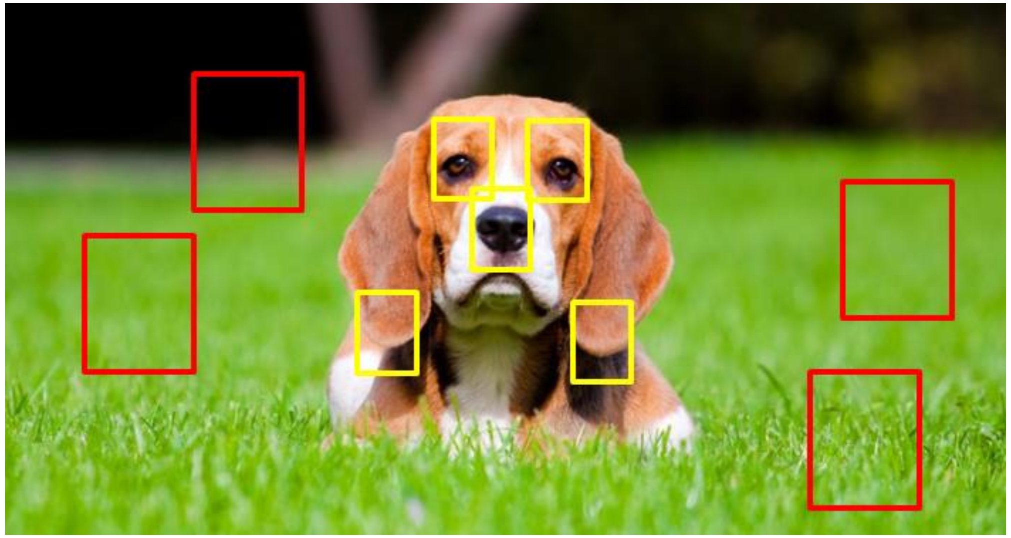
Causal Framework



T： grass
X： dog nose
Y： label

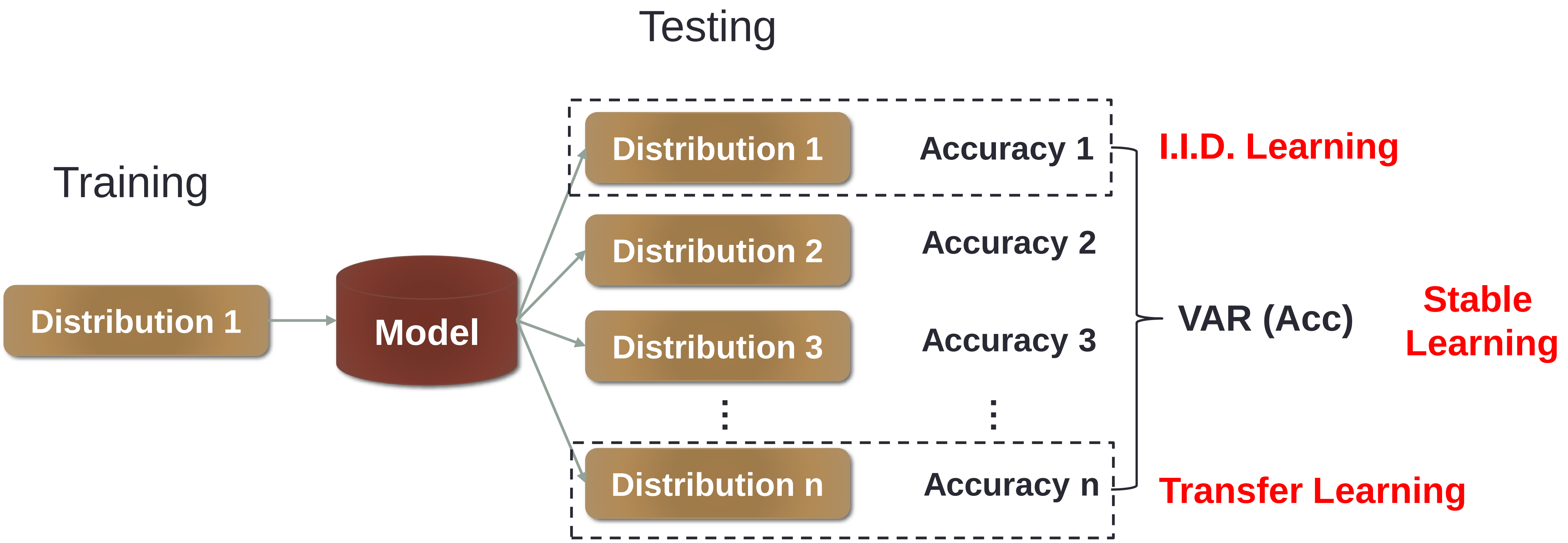
Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation



More *Explainable* and More *Stable*

Stable Learning



NICO - A Benchmark for OOD Generalization

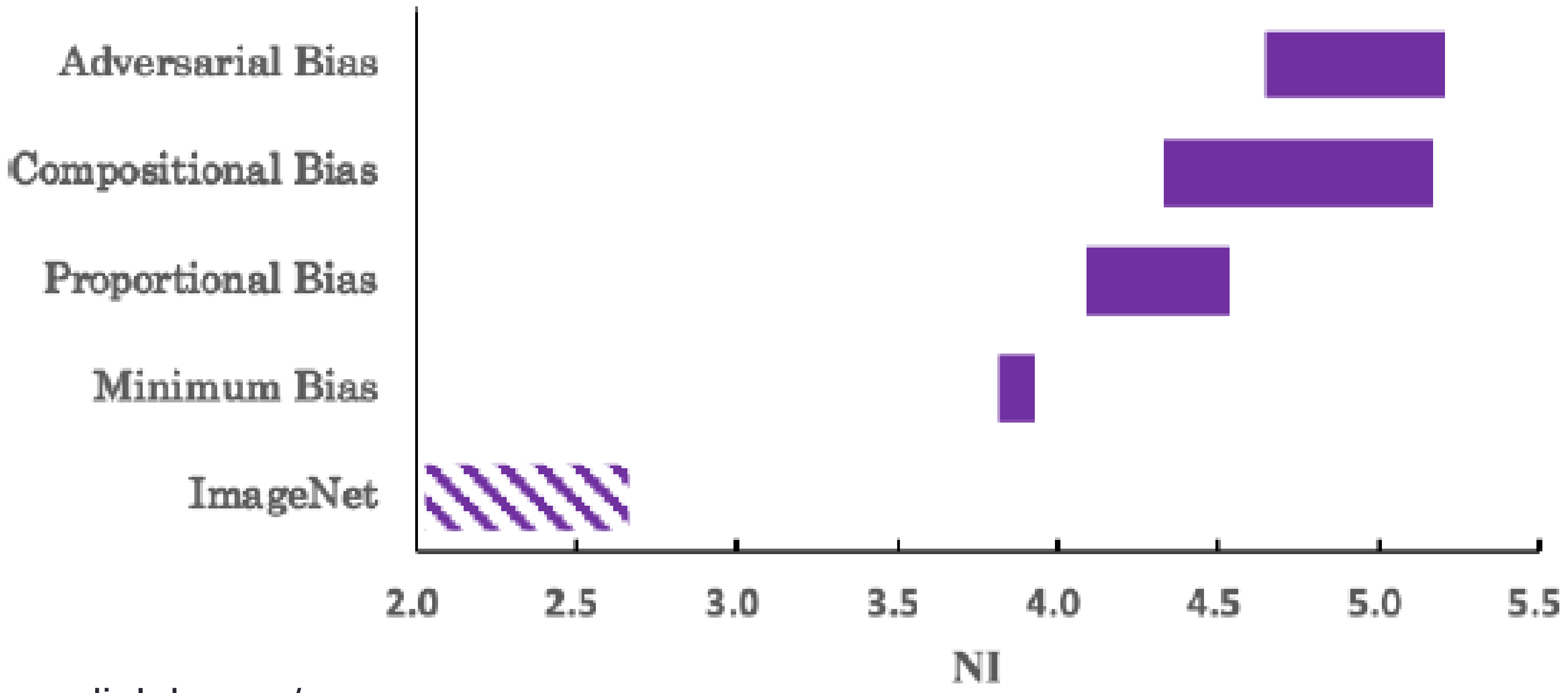
- Data size of each class in NICO
  - Sample size: thousands for each class
  - Each superclass: 10,000 images
  - Sufficient for some basic neural networks (CNN)

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

Yue He, Zheyan Shen, Peng Cui. Towards Non-I.I.D. Image Classification: A Dataset and Baselines. ***Pattern Recognition***, 2020.

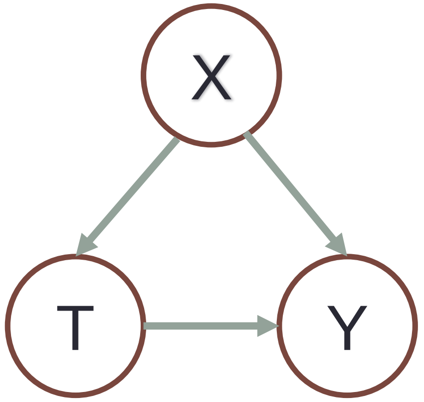
NICO - A Benchmark for OOD Generalization



http://nico.thumedialab.com/

Yue He, Zheyan Shen, Peng Cui. Towards Non-I.I.D. Image Classification: A Dataset and Baselines. ***Pattern Recognition***, 2020.

Revisit Directly Balancing for causal inference



Typical Causal Framework

Directly Confounder Balancing

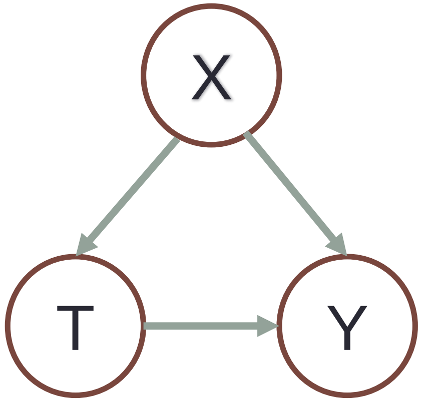
Given a feature T

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.

Stable Learning with ***Linear*** model

Variable Decorrelation by Sample Reweighting:

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

Decorrelated Weighted Regression:

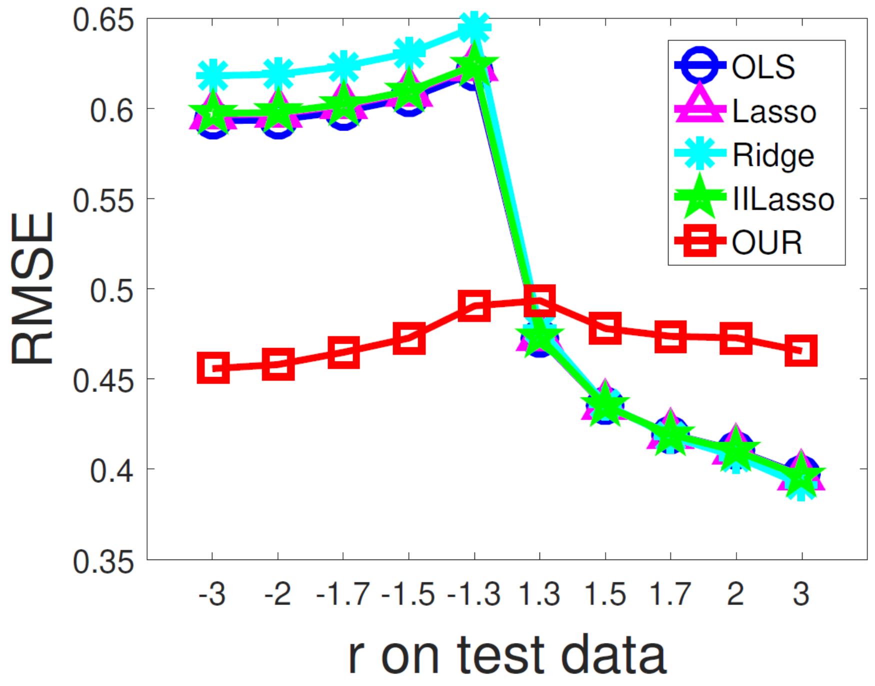
$$\min_{W,\beta} \sum_{i=1}^{n} \boxed{W_i} \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \qquad (12)$$

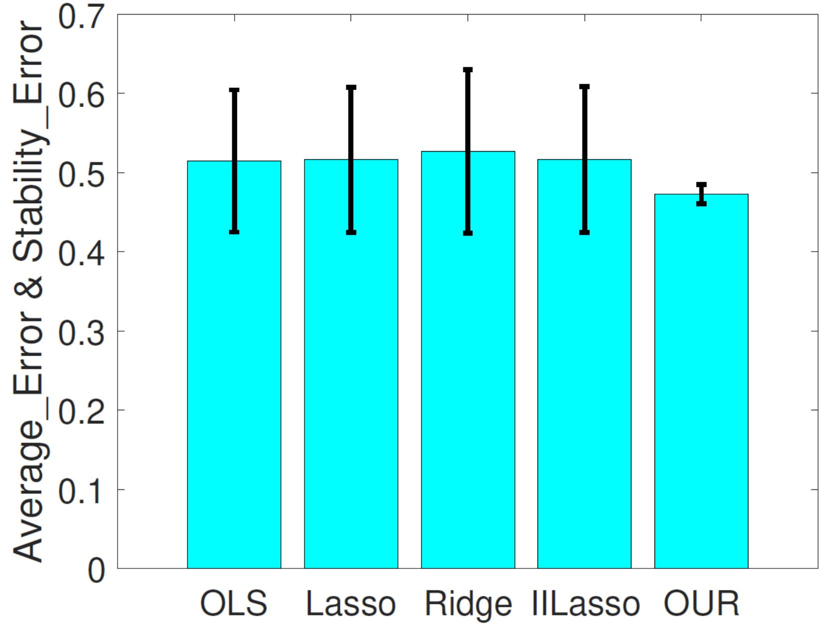
$$s.t \quad \boxed{\sum_{j=1}^{p} \left\| \mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n \right\|_2^2 < \lambda_2}$$

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

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

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.

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

Interpretation from Statistical Learning perspective

- 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, et al. Stable Learning via Sample Reweighting. AAAI, 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

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

Function space of RFF

$$f^{(t+1)}, g^{(t+1)} = \arg\min_{f,g} \sum_{i=1}^{n} w_i^{(t)} L(g(f(\mathbf{X}_i)), y_i),$$

$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\Delta_n} \sum_{1\le i<j\le m_Z} \left\| \hat{\Sigma}_{\mathbf{Z}_{:,i}^{(t+1)}\mathbf{Z}_{:,j}^{(t+1)};\mathbf{w}} \right\|_F^2$$

Representation function

Prediction function

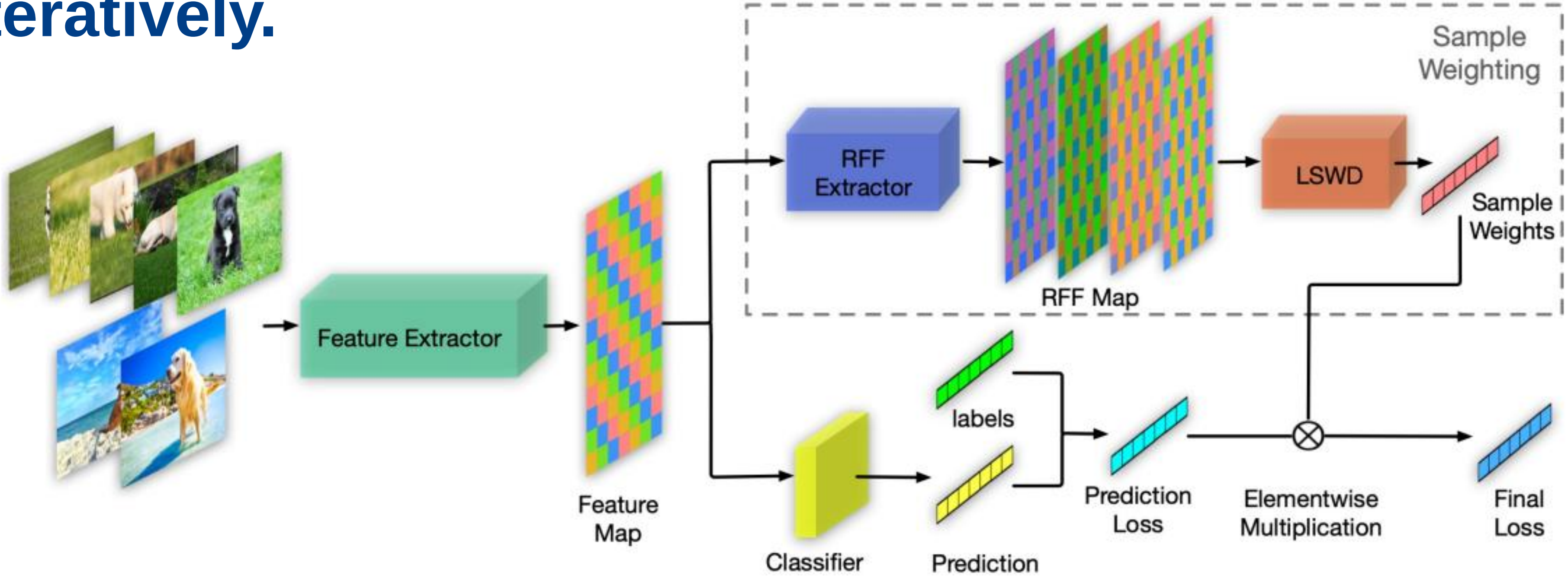
Sample weights

Partial cross-covariance matrix

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

Flexible OOD Generalization

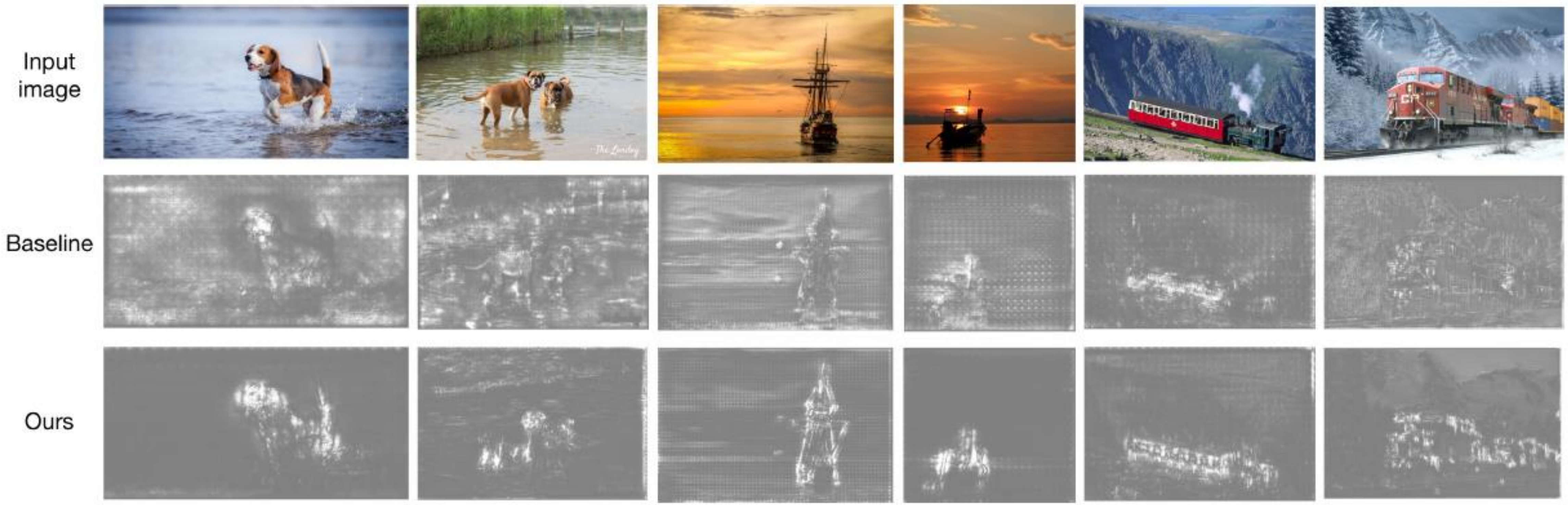
- The domains for different categories can be different.
- For instance, birds can be on trees but hardly in the water while fishes are the opposite.

	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

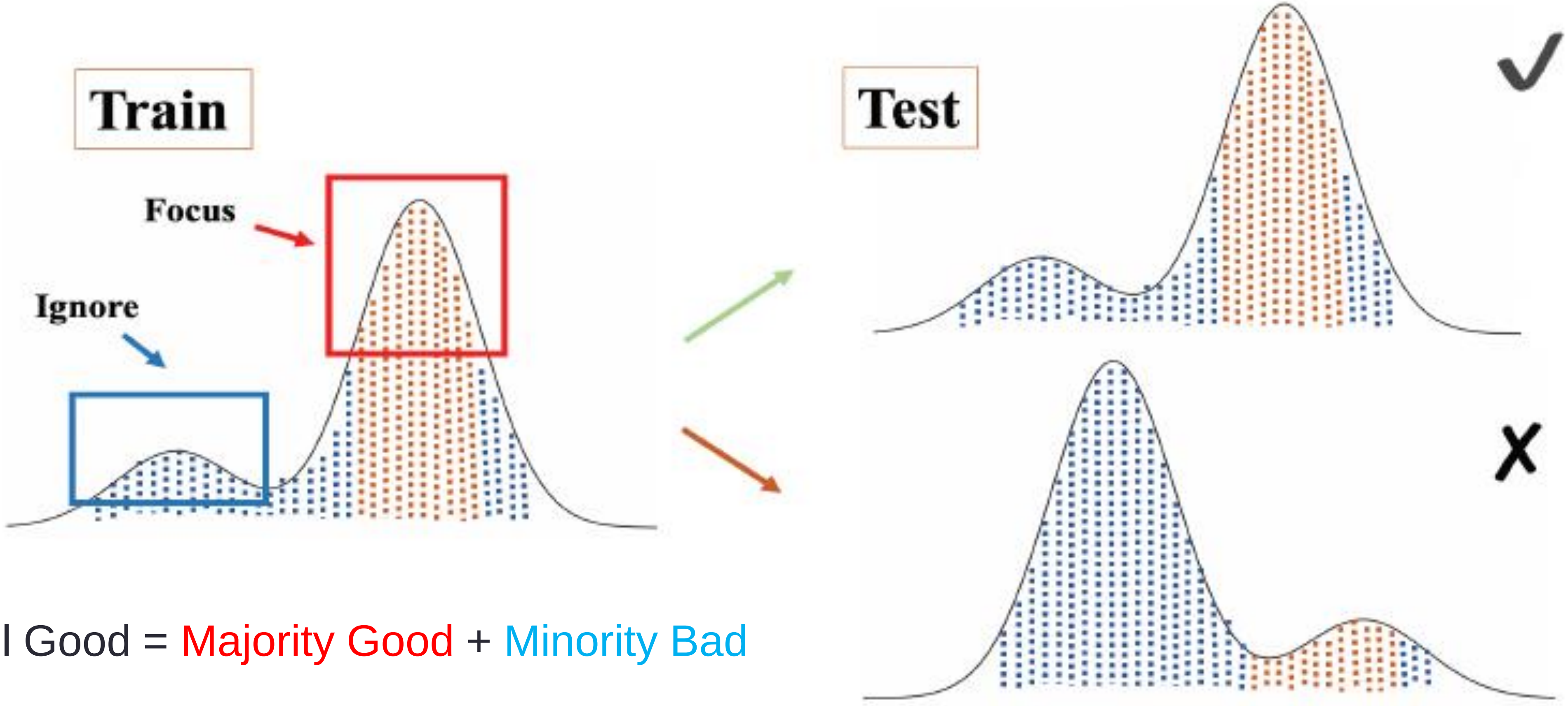
- **The visualization of the gradient of the class score function with respect to the input pixels. The brighter the pixel is, the more contribution it makes to prediction.**



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***?

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

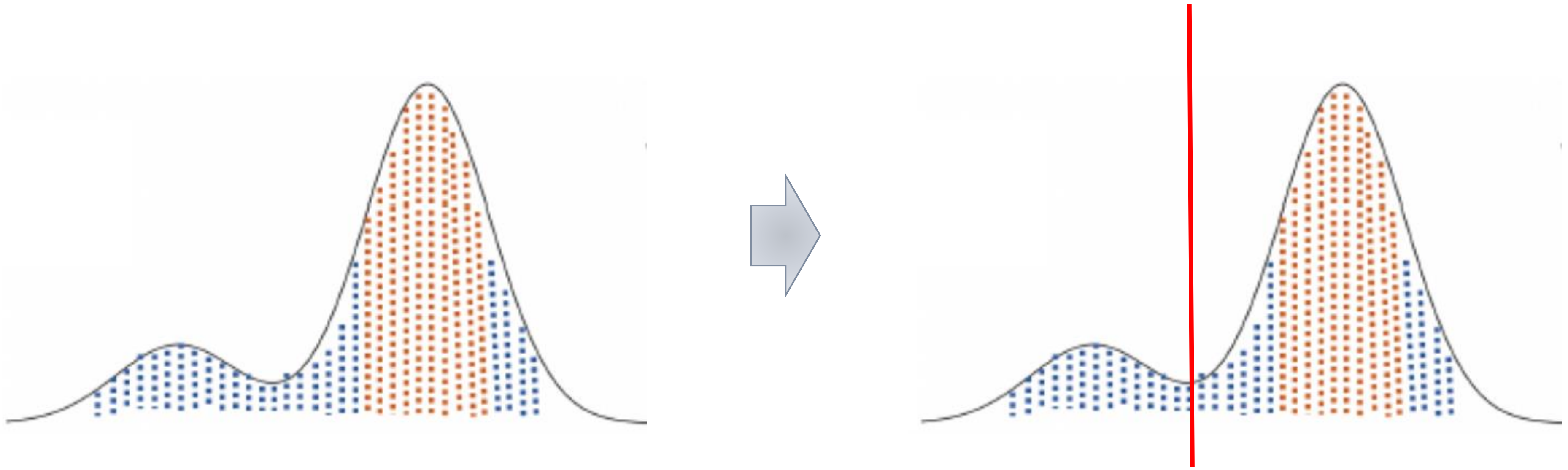


Overall Good = Majority Good + Minority Bad

Overall Good = Majority Good + Minority Good

Problem I
Uncovering Heterogeneity

Problem II
Finding Invariance



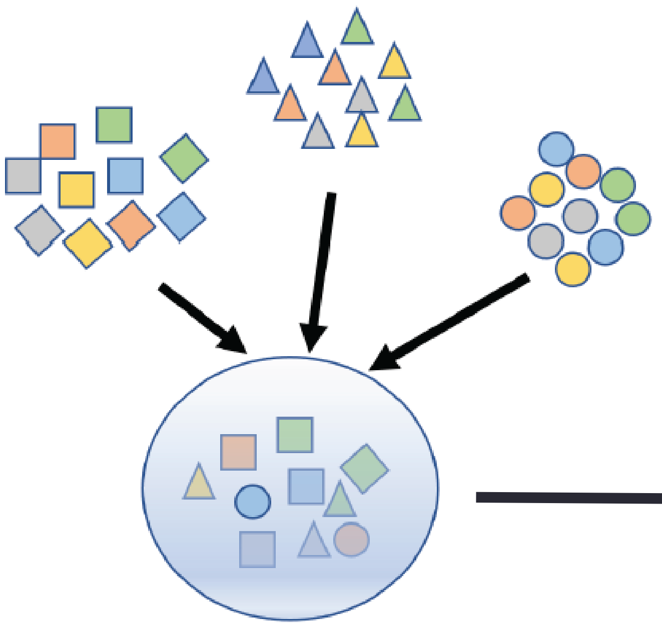
Heterogeneity → Invariance

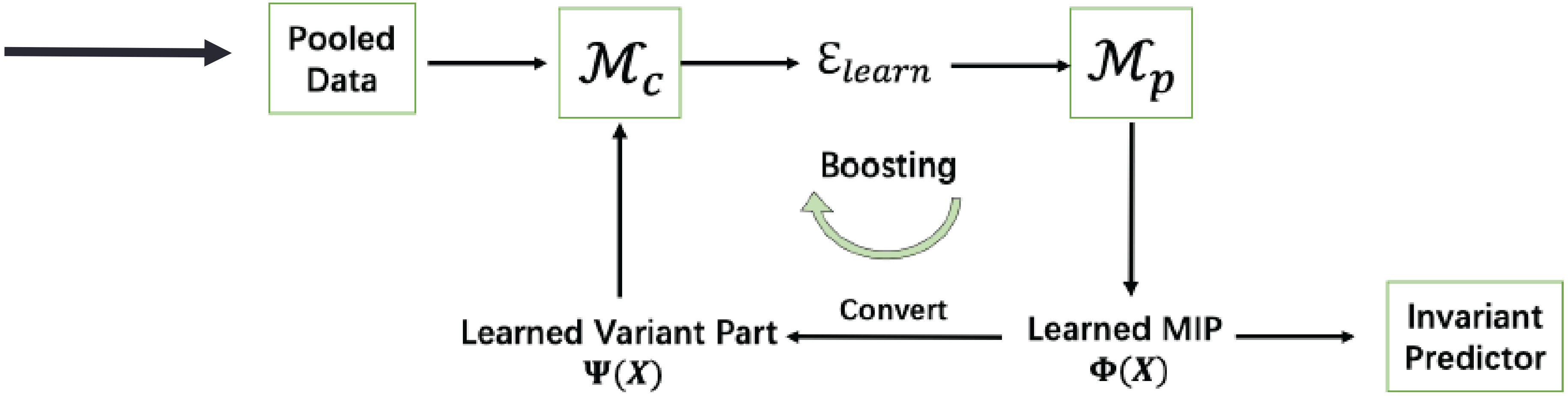
Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.

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*
$$\mathrm{D}_{\mathrm{KL}}(P^{e_i}(Y|X)\|P^{e_j}(Y|X)) \leq \mathrm{D}_{\mathrm{KL}}(P^{e_i}(Y|\Psi^*)\|P^{e_j}(Y|\Psi^*))$$

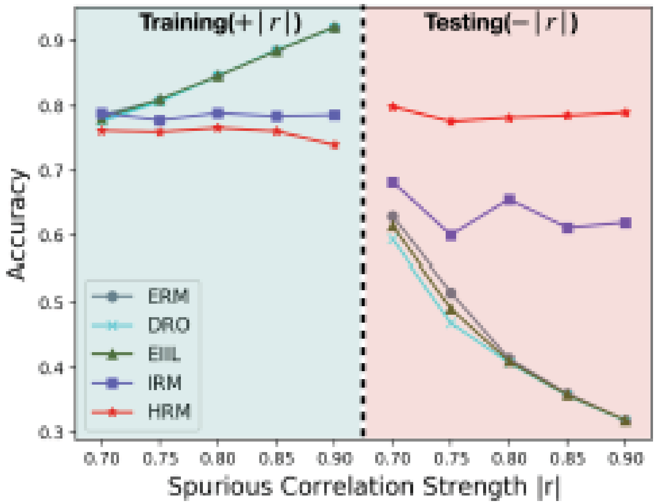


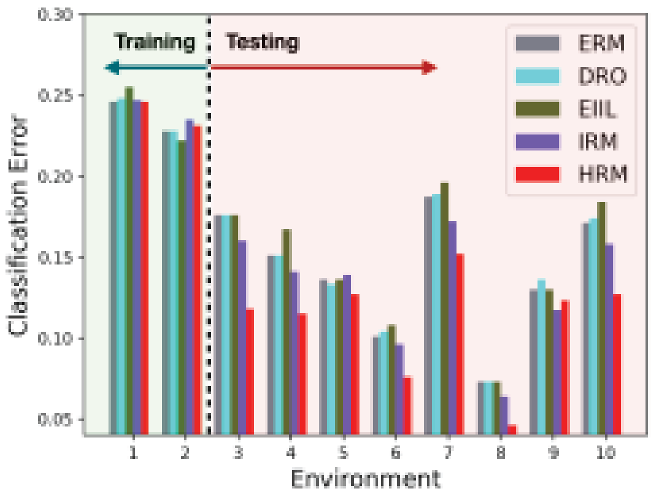


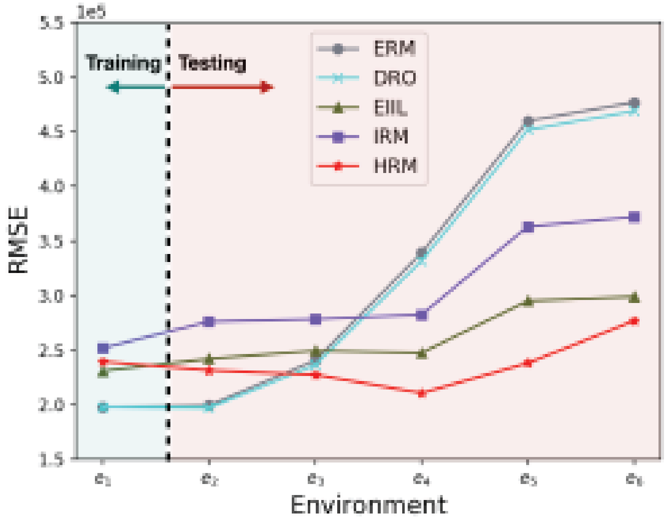
Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ***ICML***, 2021.

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.

Conclusions

- Why can't the current AI generalize well to unknown environments?

Know <u>What</u>, but don't know <u>Why</u>

知其 <u>然</u>，但不知其 <u>所以然</u>

Correlation Causality

Stable Learning: Finding the common ground between causal inference and machine learning.

Reference

- 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