Stable Learning:

Finding the Common Ground between Causal Inference and Machine Learning

Peng Cui

Tsinghua University

Now AI is stepping into risk-sensitive areas

Healthcare

Law

Human

Transportation

Fintech

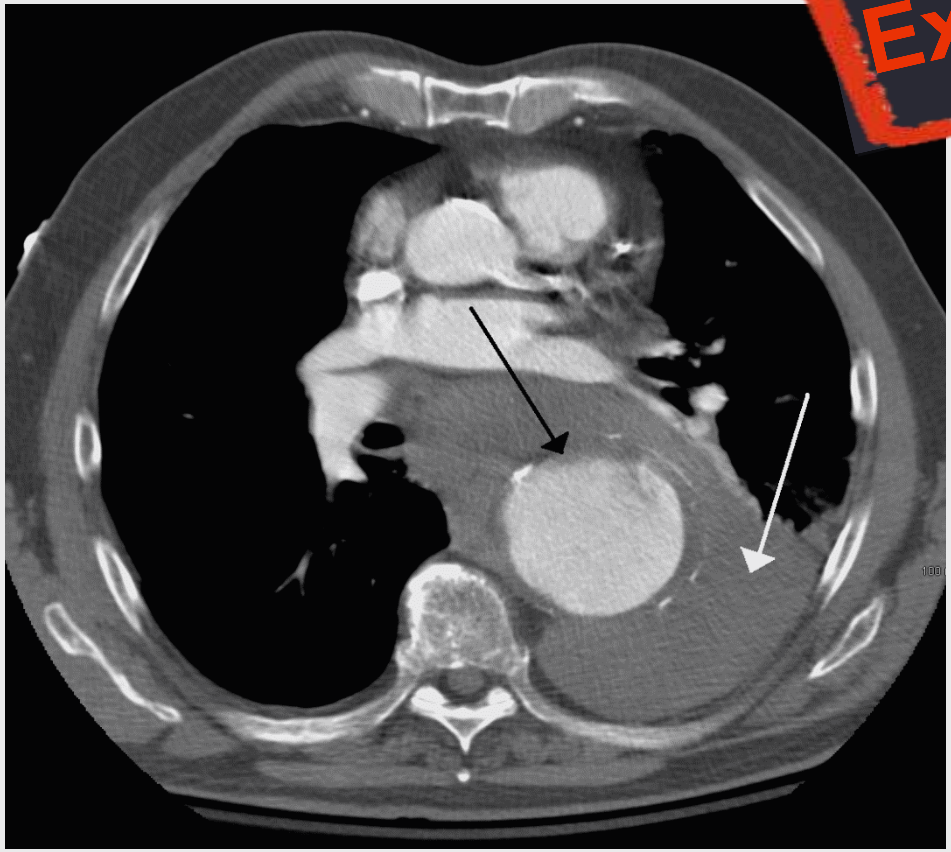
Shifting from ***Performance Driven*** to ***Risk Sensitive***

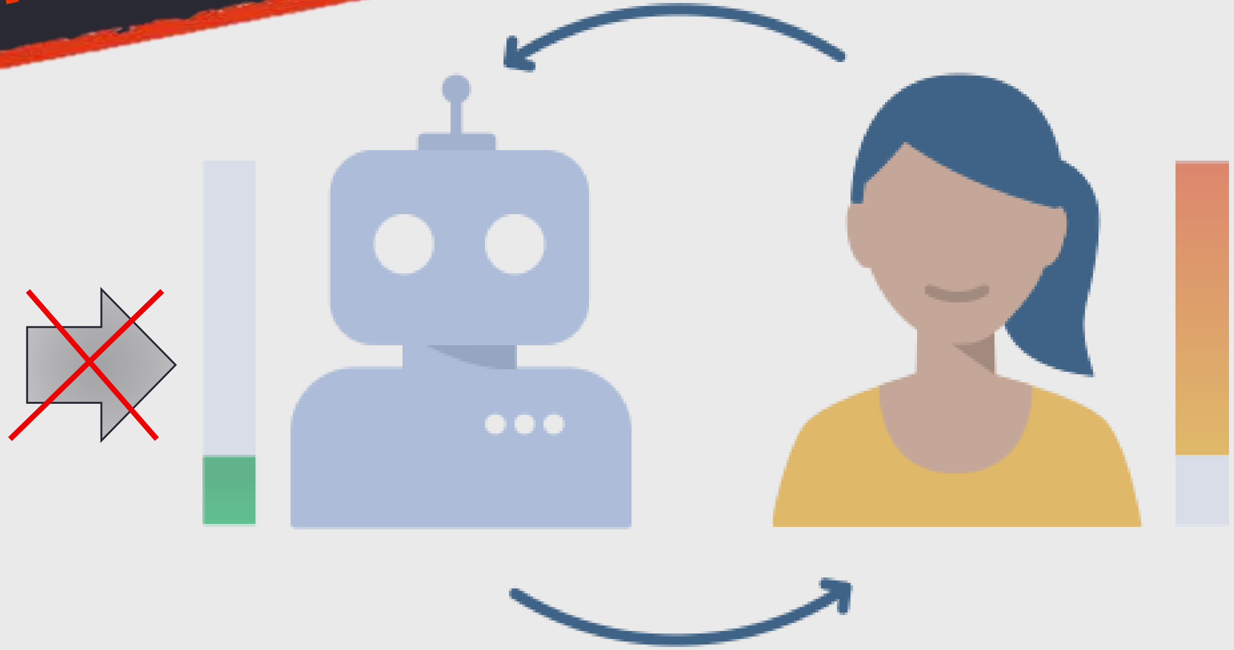
Risks of Today's AI Algorithms

Unexplainable

Explainability

man in the loop

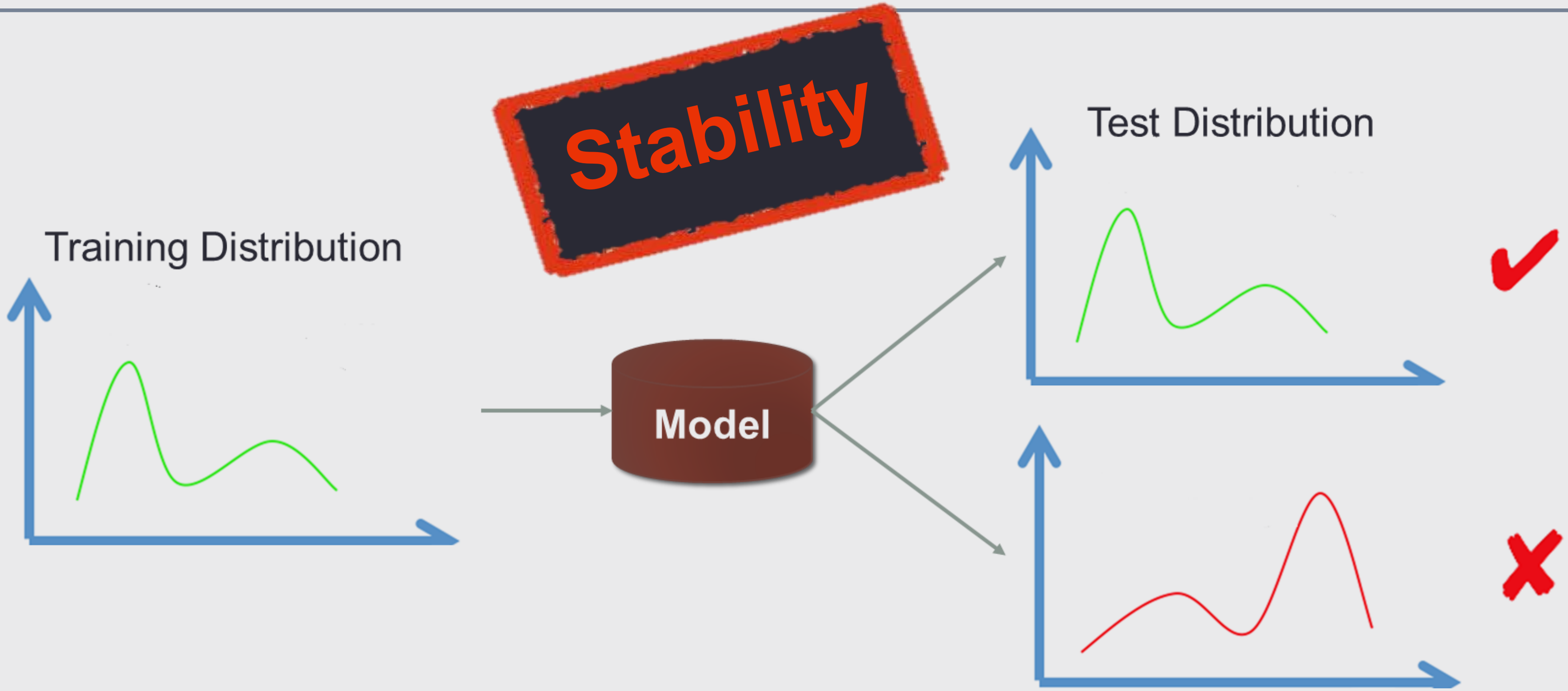




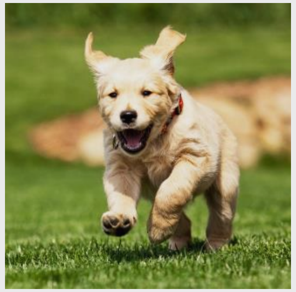
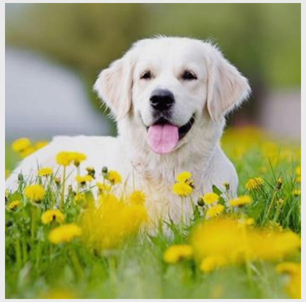
Medical Military Finance

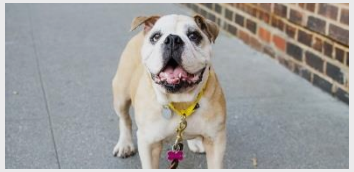
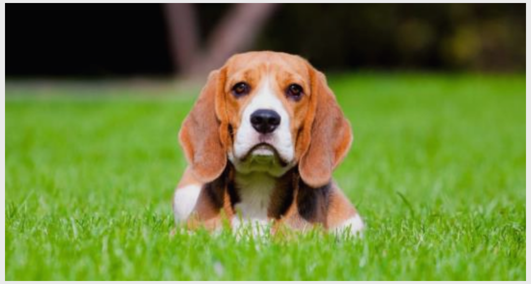
Risks of Today's AI Algorithms

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



Risks of Today’s AI Algorithms

Yes

Maybe

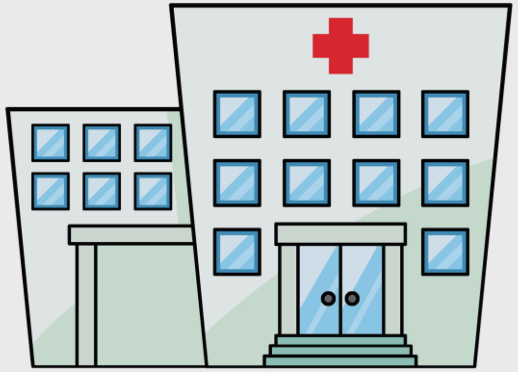
No

Risks of Today’s AI Algorithms

- Cancer survival rate prediction

Features:
- Body status
- Income
- Treatments
- Medications

Training Data

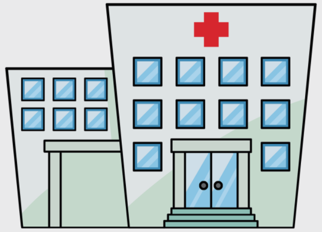


City Hospital

Higher income, higher survival rate.

Predictive Model

Testing Data



City Hospital

University Hospital

Survival rate is not so correlated with income.

The Current Condition

Explainability

We cannot ***understand*** AI

Stability

We don’t ***trust*** AI

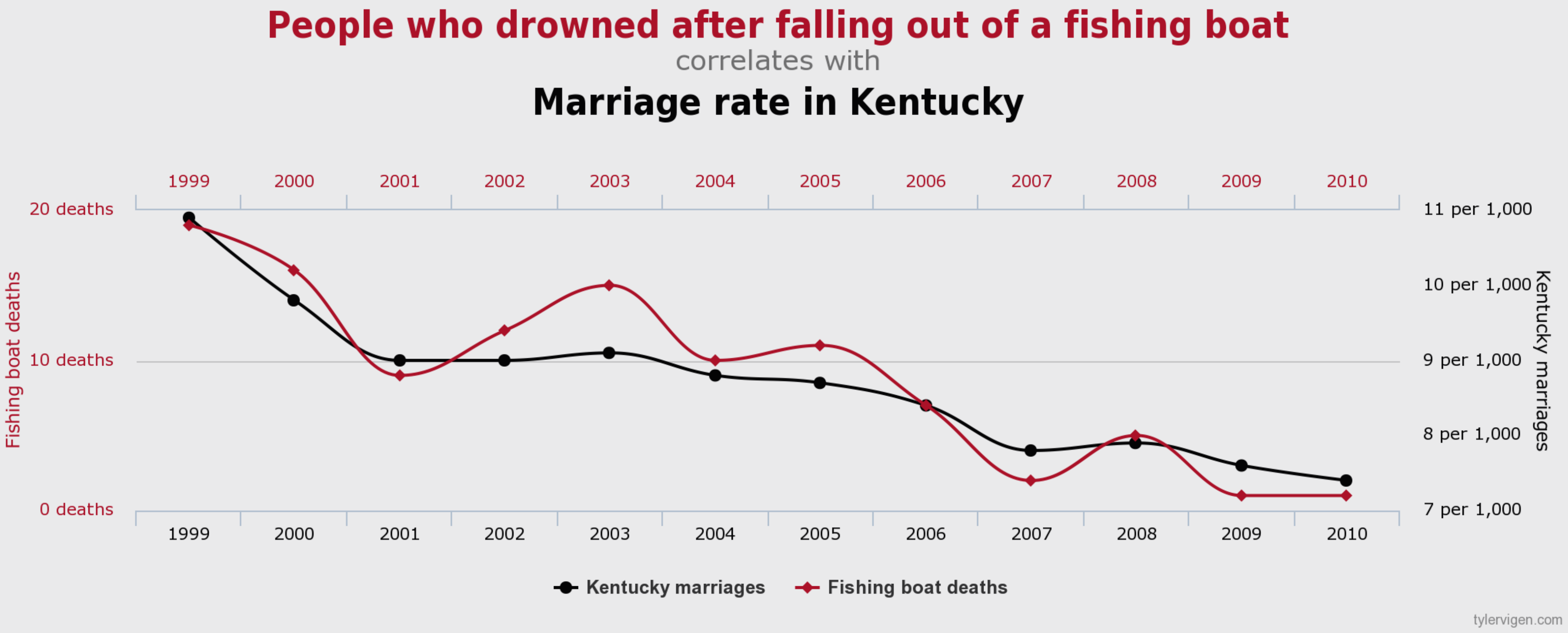
Dilemma

A plausible reason: ***Correlation***

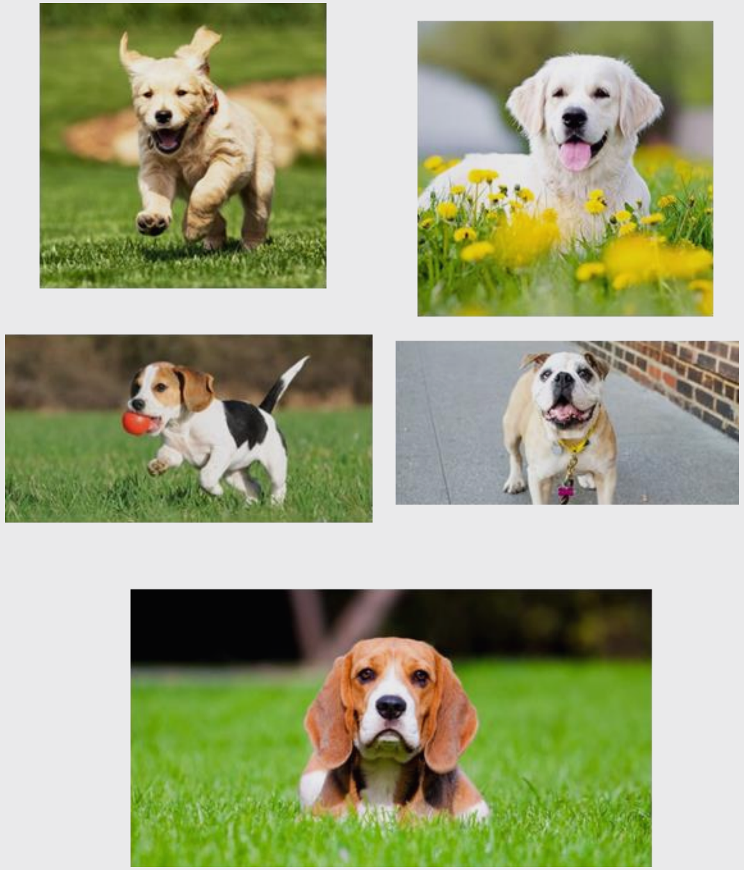
Correlation is the very basics of machine learning.

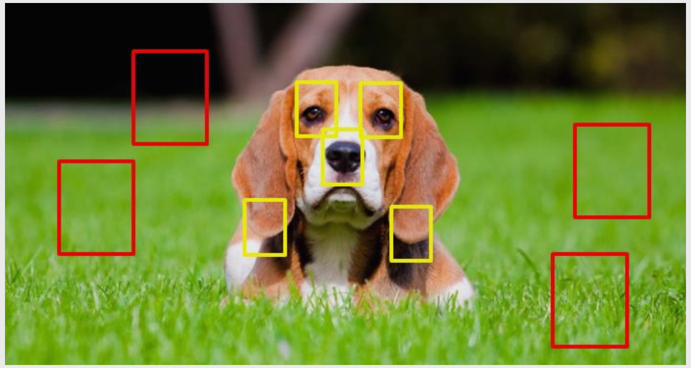
© marketoonist.com

Correlation is not explainable



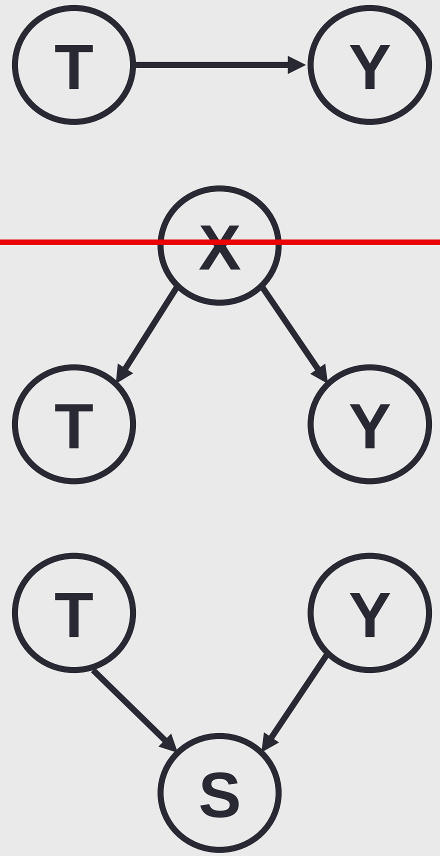
Correlation is ‘*unstable*’

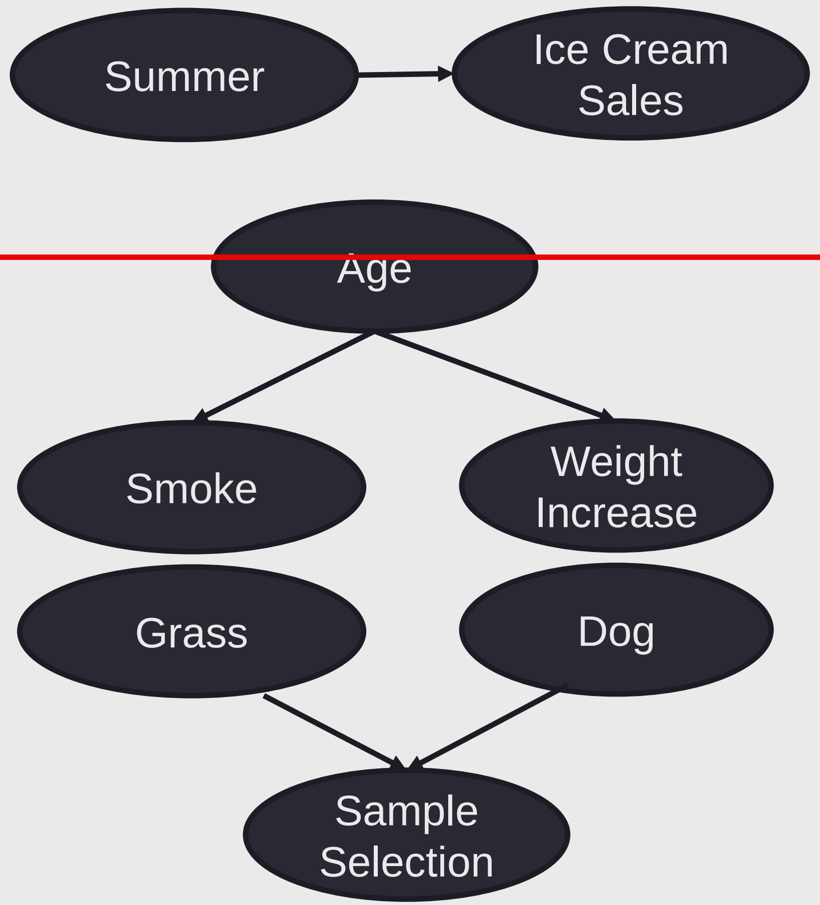




It’s not the fault of ***correlation***, but the way we use it

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



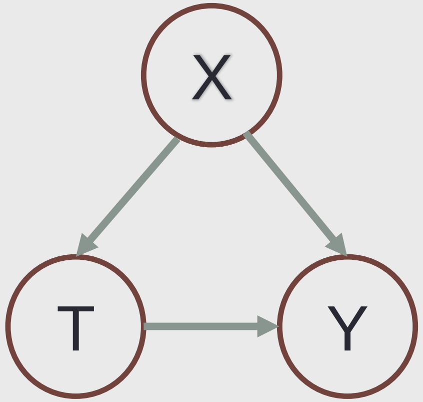


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



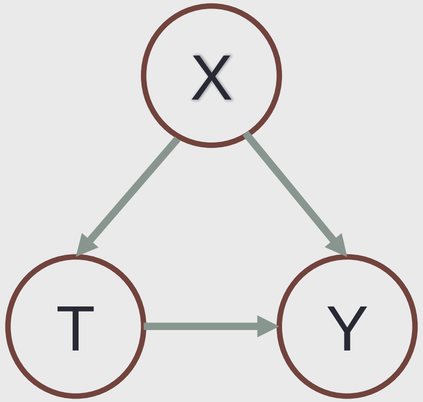
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the "interventionist" interpretation of causality.

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

The *benefits* of bringing causality into learning

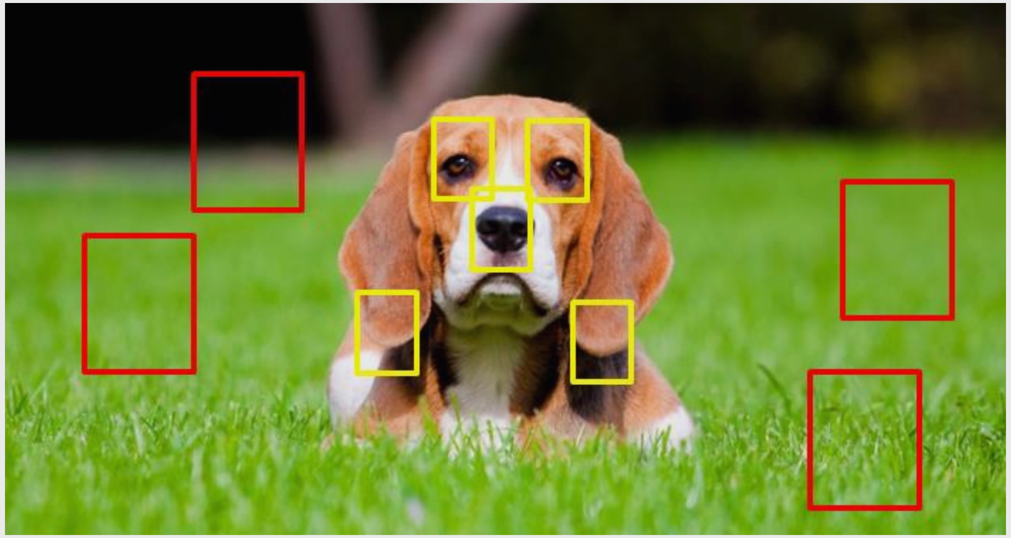
Causal Framework



T： grass
X： dog nose
Y： label

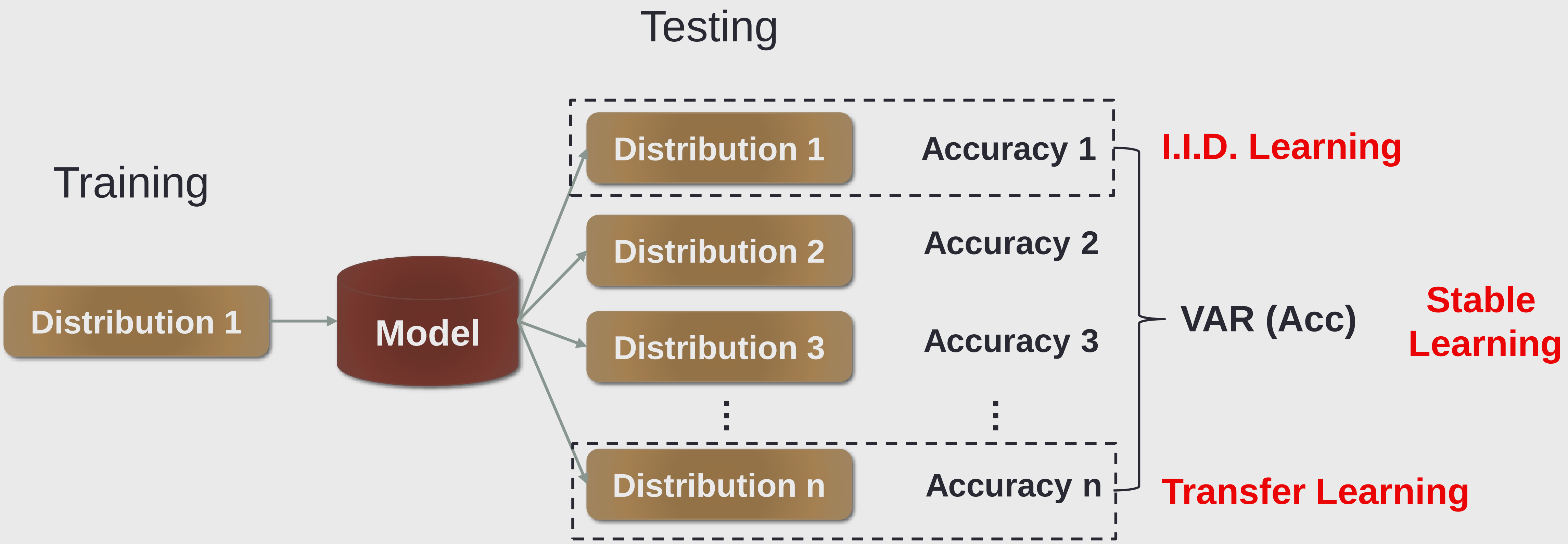
Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation

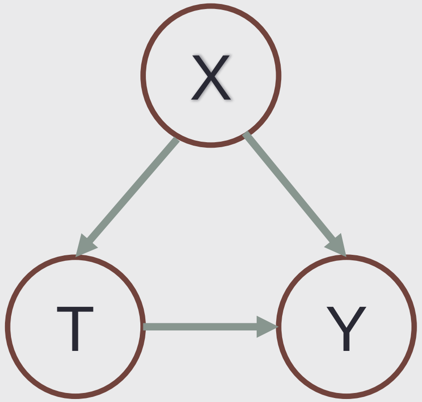


More ***Explainable*** and More ***Stable***

Stable Learning



Revisit Directly Balancing for causal inference



Typical Causal Framework

Directly Confounder Balancing

Given a feature T

↓

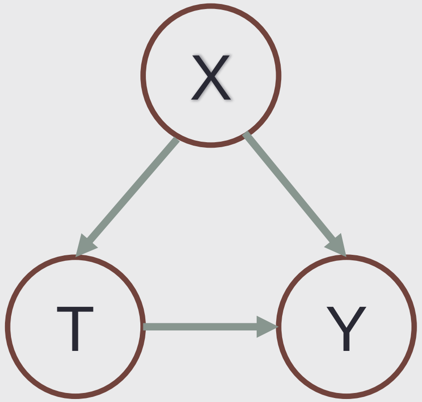
Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

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.

Global Balancing: making all variables independent



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^*.

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

b

$\downarrow$

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, et al. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Causal Regularizer

Set feature *j* as treatment variable

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

All features excluding treatment *j*

Sample Weights

Indicator of treatment status

Zheyan Shen, et al. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM***, *2018*.

Causally Regularized Logistic Regression

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

Sample reweighted logistic loss

Causal Contribution

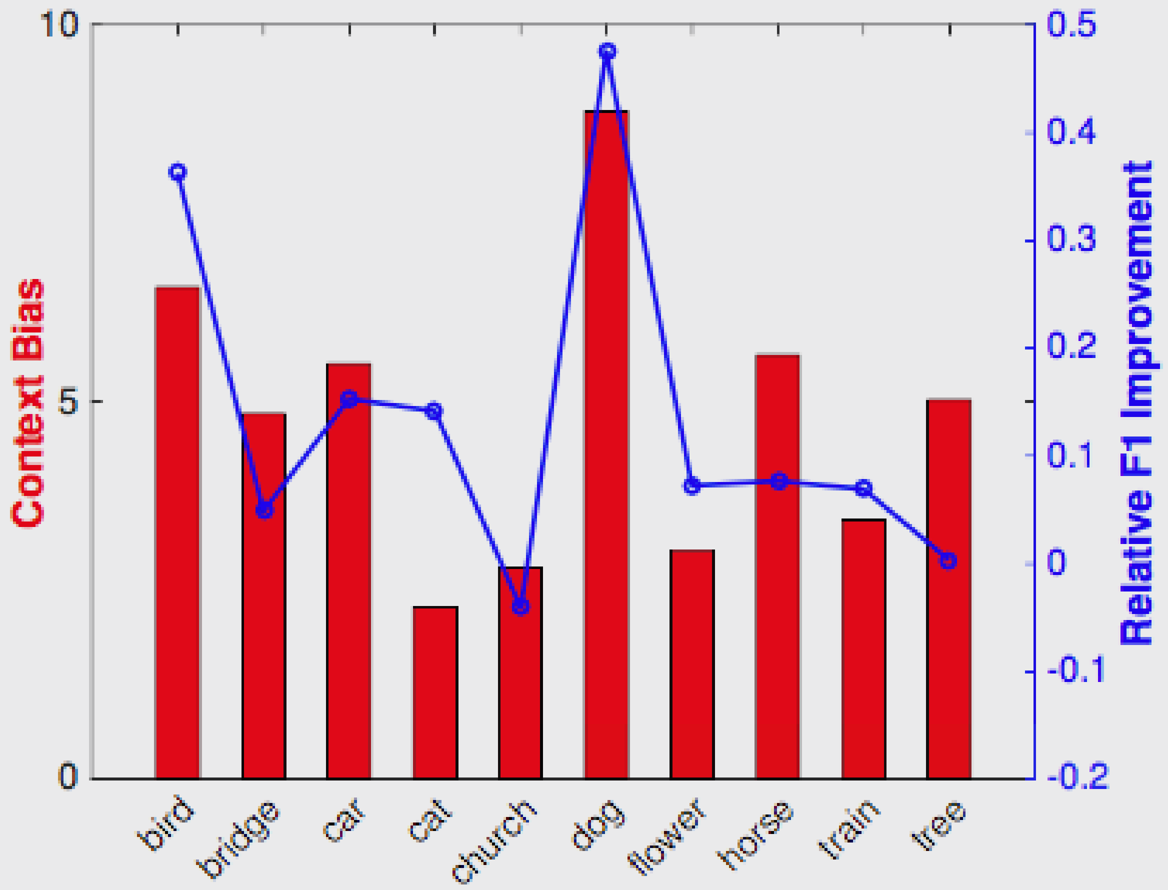
Zheyan Shen, et al. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM**, 2018*.

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

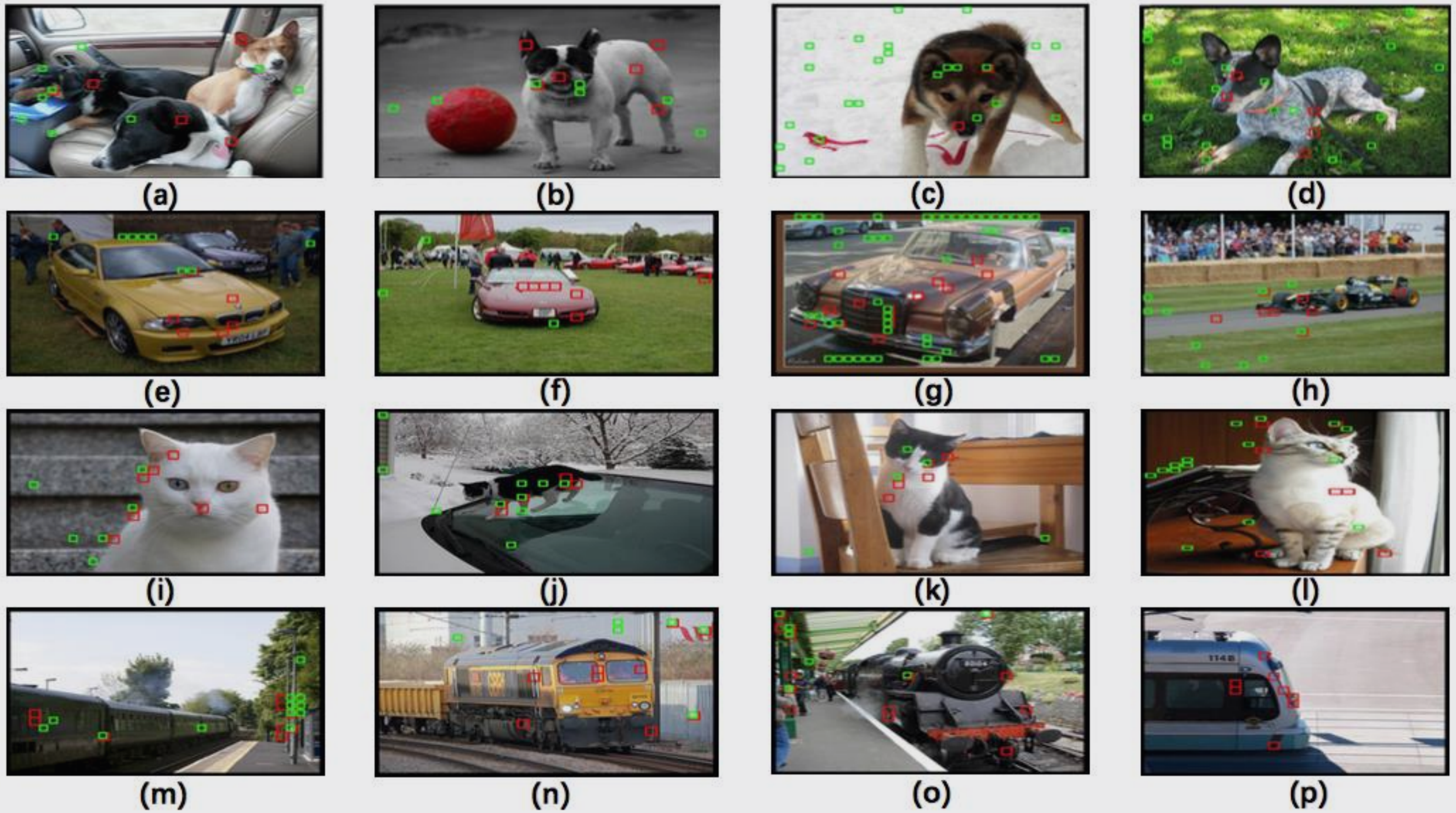
Animal	Data Size	*Vehicle*	Data Size
Bear	1609	Airplane	930
Bird	1590	Bicycle	1639
Cat	1479	Boat	2156
Cow	1192	Bus	1009
Dog	1624	Car	1026
Elephant	1178	Helicopter	1351
Horse	1258	Motorcycle	1542
Monkey	1117	Train	750
Rat	846	Truck	1000
Sheep	918		

- 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)
- Samples with contexts in NICO

Experimental Result - insights



Experimental Result - insights



Stable Learning with ***Continuous*** Variables

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:

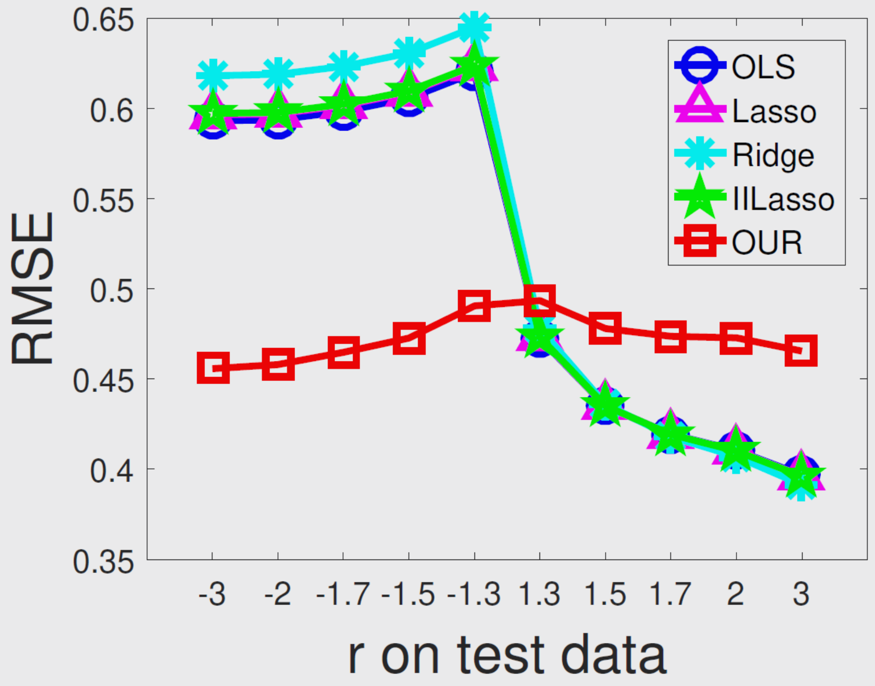
$$\min_{W,\beta} \sum_{i=1}^{n} \boxed{W_i} \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \qquad (12)$$

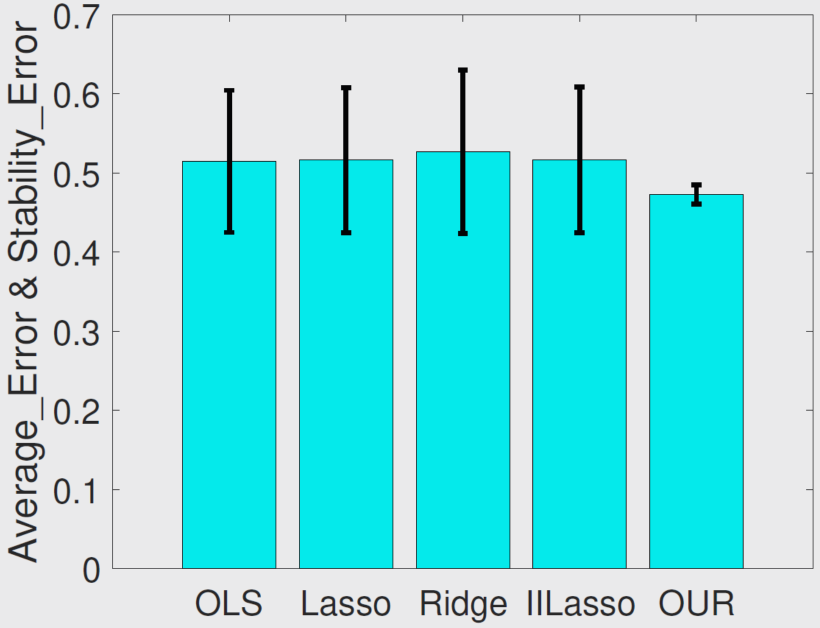
$$s.t \quad \boxed{\sum_{j=1}^{p} \left\| \mathbf{X}_{,j}^{T} \mathbf{\Sigma}_{W} \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^{T} W/n \cdot \mathbf{X}_{,-j}^{T} W/n \right\|_{2}^{2} < \lambda_2}$$

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

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

Stable Learning with ***Continuous*** Variables





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

Stable Learning with ***Differentiated*** Variables

- More detailed analysis:

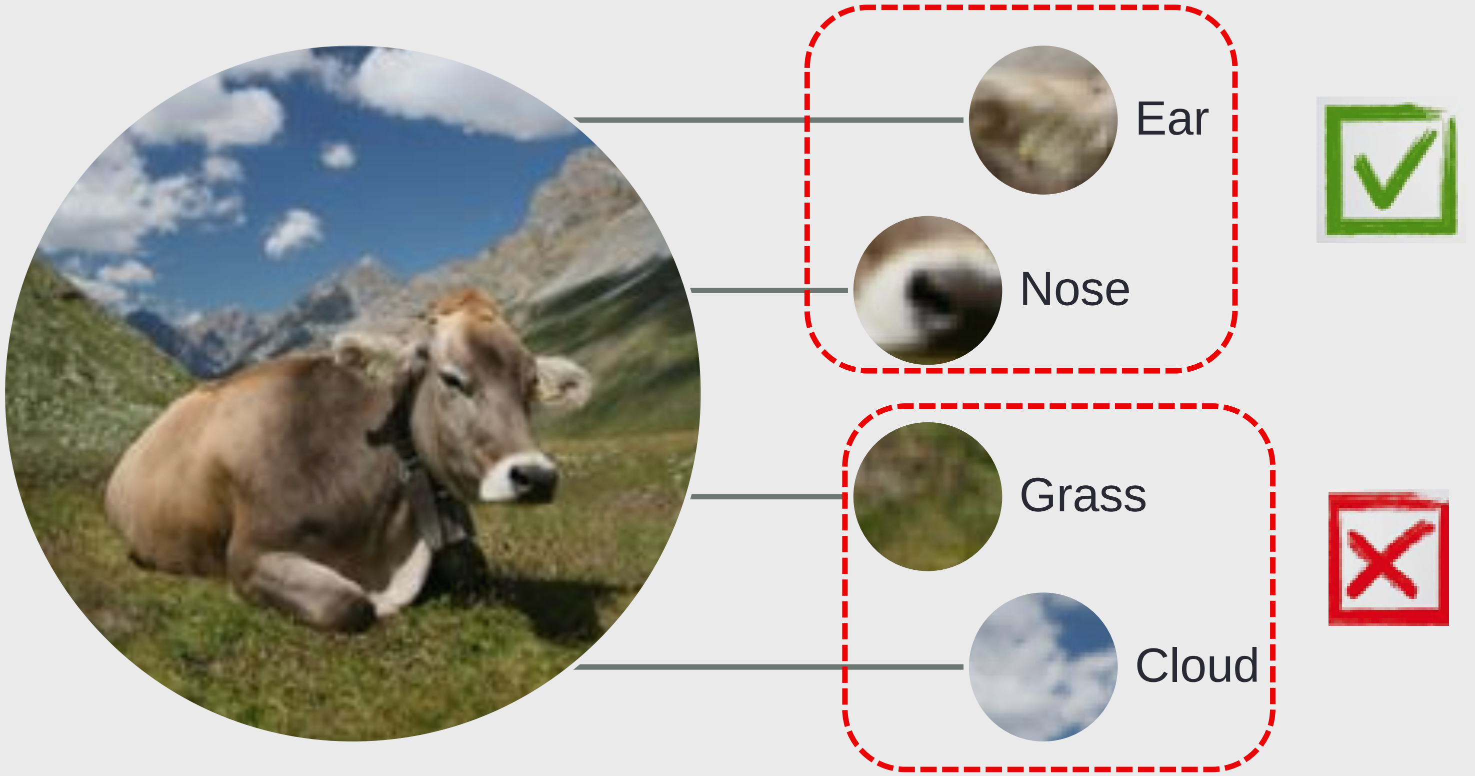
$$\hat{\beta}_{V_{OLS}} = \beta_V + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{V}_i\right)^{-1}\boxed{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T g(\mathbf{S}_i)\right)} + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{V}_i\right)^{-1}\boxed{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{S}_i\right)}\left(\beta_S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

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

- We can focus on only the spurious part of correlation
- But how?
- Leveraging the abundant sources of unlabeled data!

Stable Learning with ***Differentiated*** Variables

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



Clustering?

Stable Learning with ***Differentiated*** Variables

- 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:

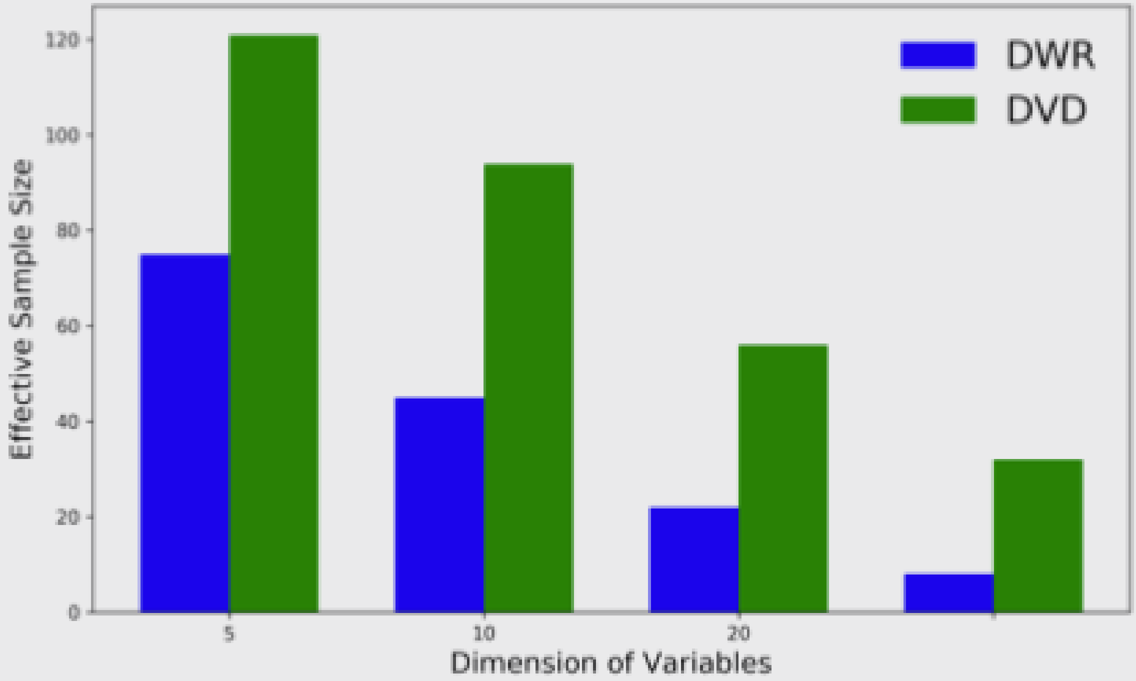
$$\min_{W} \sum_{i \neq j} \boxed{\mathbb{I}(i,j)} \left\| (\mathbf{X}_{,i}^T \Sigma_W \mathbf{X}_{,j}/n - \mathbf{X}_{,i}^T W/n \cdot \mathbf{X}_{,j}^T W/n) \right\|_2^2$$

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

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

Stable Learning with ***Differentiated*** Variables

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.

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.

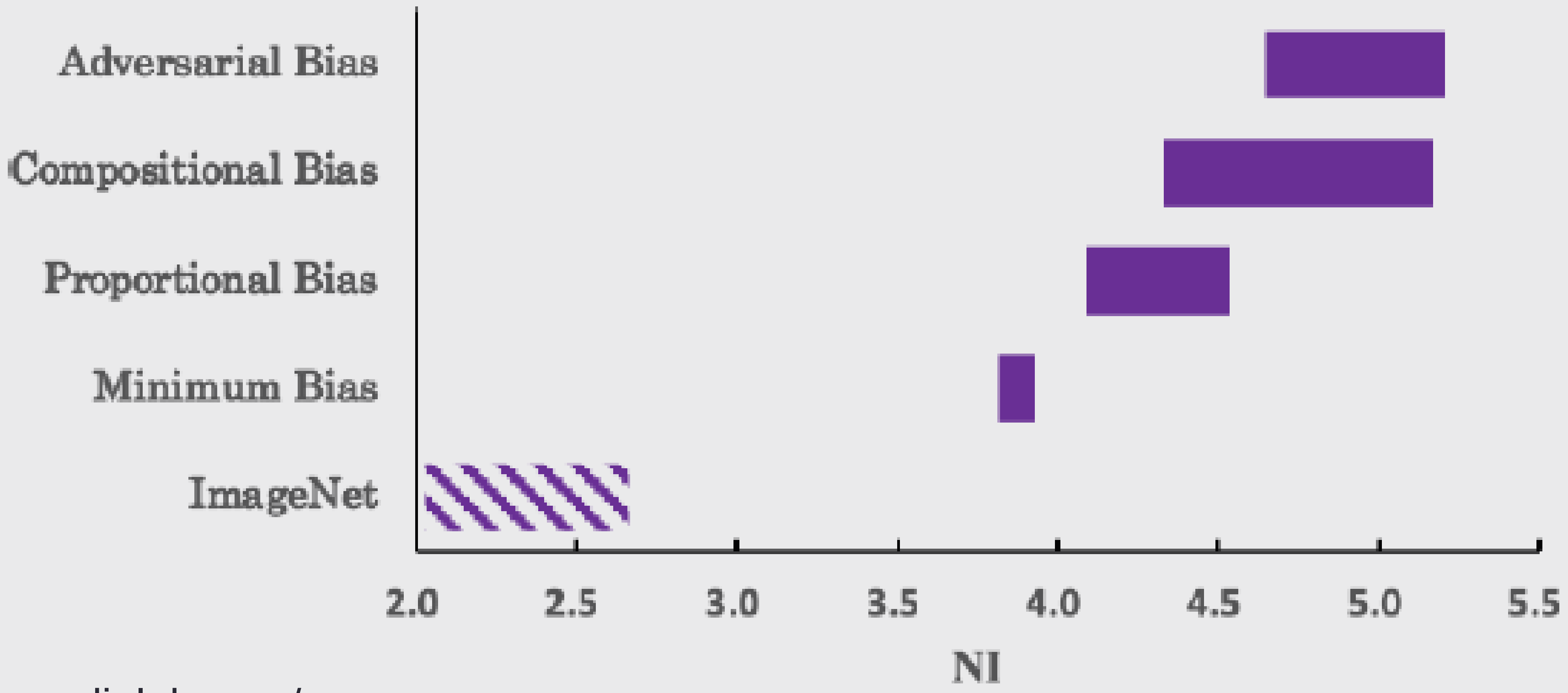
Zheyan Shen, et al. Stable Learning via Sample Reweighting. AAAI, 2020.

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

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		

- 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)
- Samples with contexts in NICO

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



http://nico.thumedialab.com/

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

Conclusions

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

Know What, but don’t know Why

知其 然 ，但不知其 所以然

Correlation Causality

Stable Learning: Try to promote the convergence of causal inference and machine learning.

Reference

- 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, Peng Cui, Jianxin Ma, Zou Hao, Xiaowei Wang, Hongxia Yang and Philip S. Yu. Learning Stable Graphs from Multiple Environments with Selection Bias. ***KDD***, 2020.
- Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen and Wei Cui. Algorithmic Decision Making with Conditional Fairness. ***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.
- Hao Zou, Kun Kuang, Boqi Chen, Peng Cui, Peixuan Chen. Focused Context Balancing for Robust Offline Policy Evaluation. ***KDD***, 2019.
- Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, Wenwu Zhu. Disentangled Graph Convolutional Networks. ***ICML***, 2019.
- 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*.
- Kun Kuang, Peng Cui, Bo Li, Shiqiang Yang. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. ***KDD***, 2017.
- Kun Kuang, Peng Cui, Bo Li, Shiqiang Yang. Treatment Effect Estimation with Data-Driven Variable Decomposition. ***AAAI***, 2017.

Thanks!

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