Data Heterogeneity Analysis for Distribution Shifts

Tutorial at 3rd Conference on Lifelong Learning Agents
(CoLLAs 2024)

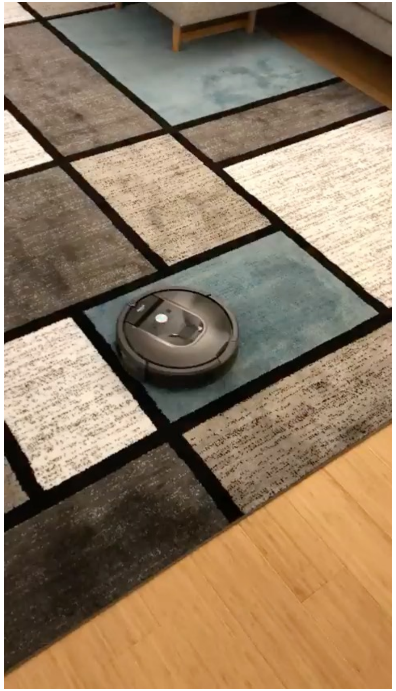
Peng Cui, Jiashuo Liu

Department of Computer Science and Technology
Tsinghua University
liujiashuo77@gmail.com

July 29th, 2024, Pisa, Italy

Based on joint work & materials with Jose Blanchet, Tiffany (Tianhui) Cai, Bo Li, Jiajin Li, Hongseok Namkoong, Tianyu Wang, Jiayun Wu

Risk of Today's AI Systems



Risk of Today's AI Systems

- AI camera ruins football game for fans after mistaking referee's bald head for ball

Risk of Today's AI Systems

- Existing AI models have extremely high bias and risk when predicting COVID-19.

nature machine intelligence
ANALYSIS
https://doi.org/10.1038/s42256-021-00307-0
Check for updates

OPEN

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

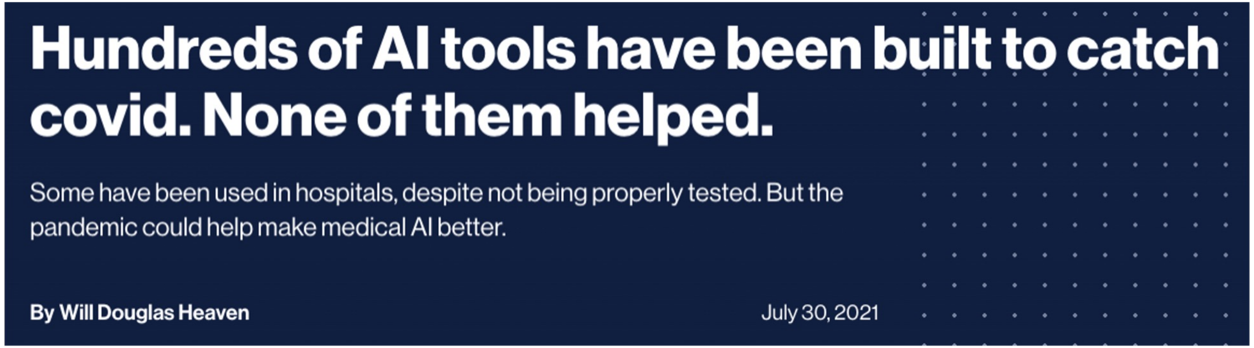
Michael Roberts[1,2]✉, Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], AIX-COVNET*, James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Elena Albu,[2] Banafsheh Arshi,[1] Vanesa Bellou,[10] Marc M J Bonten,[8,11] Darren L Dahly,[12,13] Johanna A Damen,[8,9] Thomas P A Debray,[8,14] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,15] Paula Dhiman,[4,5] Joie Ensor,[6] Shan Gao,[2] Maria C Haller,[7,16] Michael O Harhay,[17,18] Liesbet Henckaerts,[19,20] Pauline Heus,[8,9] Jeroen Hoogland,[8] Mohammed Hudda,[21] Kevin Jenniskens,[8,9] Michael Kammer,[7,22] Nina Kreuzberger,[23] Anna Lohmann,[24] Brooke Levis,[6] Kim Luijken,[24] Jie Ma,[5] Glen P Martin,[25] David J McLernon,[26] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[27,28] Chunhu Shi,[29] Nicole Skoetz,[22] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[30] René Spijker,[8,9,31] Ewout W Steyerberg,[3] Toshihiko Takada,[8,32] Ioanna Tzoulaki,[10,33] Sander M J van Kuijk,[34] Bas C T van Bussel,[1,35] Iwan C C van der Horst,[35] Kelly Reeve,[36] Florien S van Royen,[8] Jan Y Verbakel,[37,38] Christine Wallisch,[7,39,40] Jack Wilkinson,[24] Robert Wolff,[41] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

Risk of Today's AI Systems

- Correlation is no substitute for causal evidence
- COVID prediction AIs were found to be "picking up on the text font that certain hospitals used to label the scans."
- "As a result, fonts from hospitals with more serious caseloads became predictors of covid risk."



Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven July 30, 2021

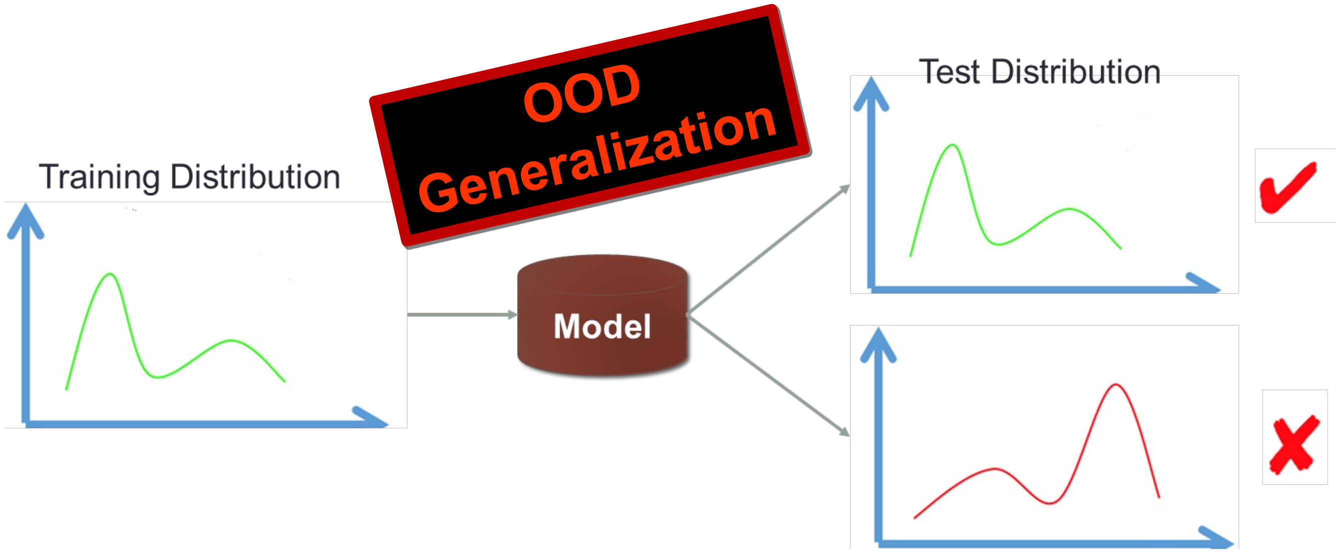
Risk of Today’s AI Systems

Owner: “Car kept jamming on the brakes thinking this was a person”

Risk of Today’s AI Systems

Risk of Today's AI Systems

Most ML methods are developed under ***i.i.d*** hypothesis



From a **DATA** Perspective

Data Problems

Distribution Shifts

Sub-population Structure

Data Corruptions

Model problems under distribution shifts

Poor generalization

Unfair to minority groups

Sensitive to perturbations

Analyze → **Solve**

Main Scope

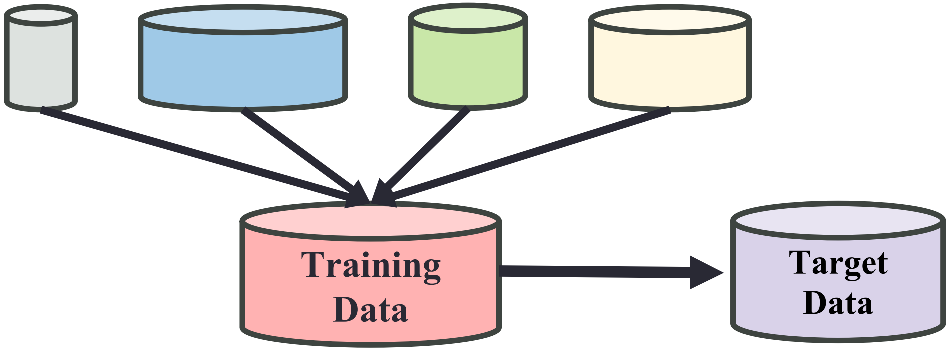
Analyze **data heterogeneity** to address the problems caused by **distribution shifts** from a **systematic** perspective

Data Heterogeneity: the complex nature of data

- sub-population structures
- hard samples, noisy samples
- different data generating processes
- different data types, sources, …

Data Heterogeneity

ML models are based on *heterogeneous* data sources



- multiple ***environments***
- different ***Y|X distributions***
- different ***data size***

Today: opportunities and challenges of heterogeneity

Main Scope

Analyze **data heterogeneity** to address the problems caused by **distribution shifts** from a **systematic** perspective

Distribution Shifts: complicated distribution shift patterns in practice

- Data corruptions
- Sub-population shifts: X-shifts vs. $Y|X$-shifts

X-shifts vs. $Y|X$-shifts

- So far: Humans are robust on all distributions. Can we get a universally good model?
- Implicitly, this view focuses on covariate shift (X-shift)
  - Traditional focus of ML
- On the other hand, we expect $Y|X$-shifts when there are unobserved factors
  - Traditional focus of causal inference
- For $Y|X$-shifts, we don't expect a single model to perform well across distributions
- Requires application-specific understanding of distributional differences

Main Scope

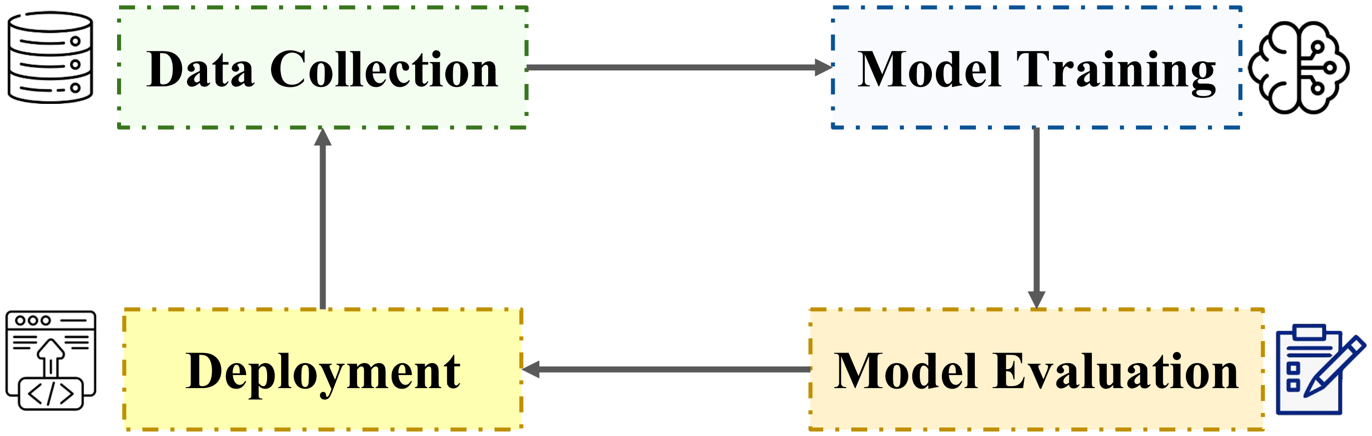
Analyze **data heterogeneity** to address the problems caused by **distribution shifts** from a **systematic** perspective

A system of view: different stages in the whole ML pipeline

- Data collection->Model training -> Model evaluation -> Deployment

A Systemic Perspective

- Building a reliable AI stack requires a holistic view



Outline

Part 1: A critical review of existing approaches

Part 2: Shift to an inductive research philosophy

Part 3: Towards heterogeneity-<u>aware</u> machine learning

Part 4: Future Directions

Outline

Part 1: A critical review of existing approaches

- Distributionally Robust Optimization
- Invariant Learning

} **make modeling assumptions**

- Pretrained "Big" Models

scale up model & data

Part 2: Shift to an inductive research philosophy

Part 3: Towards heterogeneity-aware machine learning

Part 4: Future Directions

Outline

Part 1: A critical review of existing approaches

- **Distributionally Robust Optimization**
- Invariant Learning

} **make modeling assumptions**

- Pretrained "Big" Models — **scale up model & data**

Part 2: Shift to an inductive research philosophy

Part 3: Towards heterogeneity-aware machine learning

Part 4: Future Directions

Distributionally Robust Optimization (DRO)

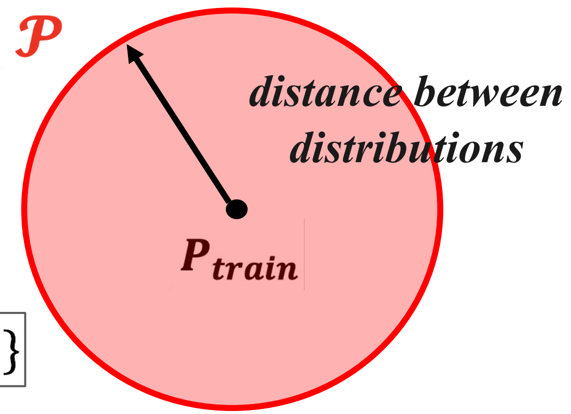
Empirical Risk Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{train}}[\ell(\theta; Z)]$$

DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$



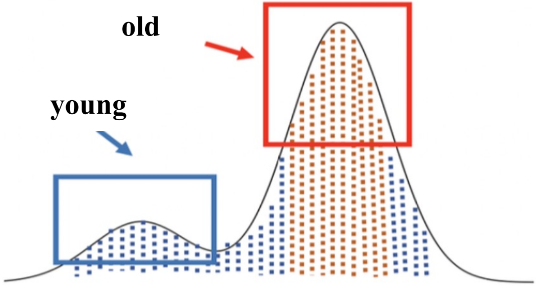
Instead of minimizing loss over training distribution, minimize loss over distributions ***near*** it

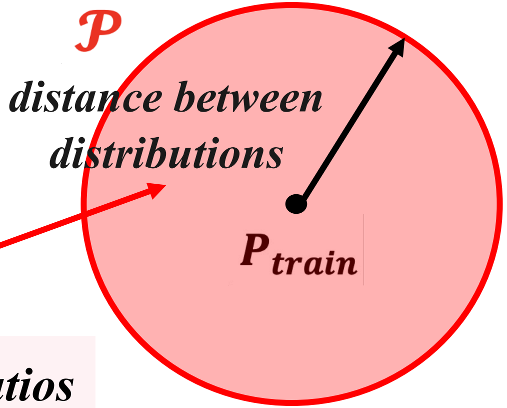
Distributionally Robust Optimization (DRO)

DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

Training distribution





Consider ***different mixture ratios*** of young and old people!

Distributionally Robust Optimization (DRO)

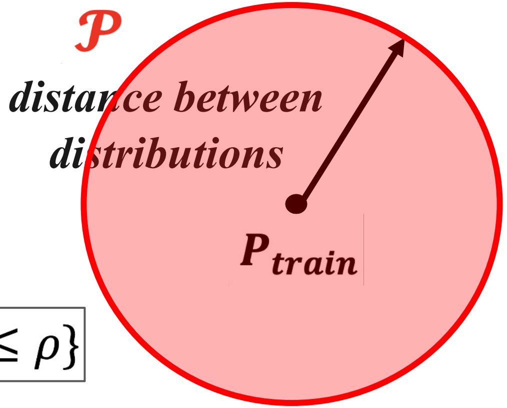
Empirical Risk Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{train}}[\ell(\theta; Z)]$$

DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

$$\mathcal{P} = \{Q : Dist(Q, P_{train}) \leq \rho\}$$



1. **Define set of distributions you care about**
2. Minimize loss on worst distribution in this set

Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

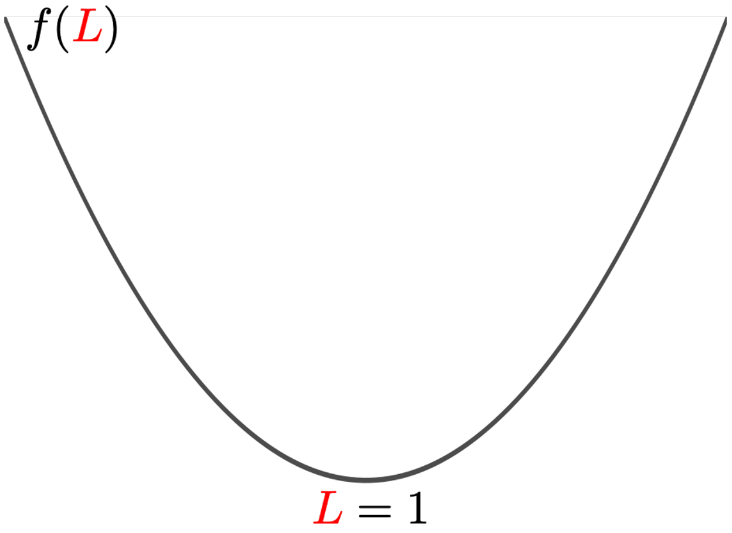
***f*-divergence:** about *densities*

If $L = \frac{dQ}{dP}$ is "near 1", then Q and P are near.

For a convex function,

$$f : \mathbb{R}_+ \to \mathbb{R} \quad \text{with } f(1) = 0,$$

$$D_f(Q \| P) := \mathbb{E}_P \left[f\left(\frac{dQ}{dP} \right) \right]$$



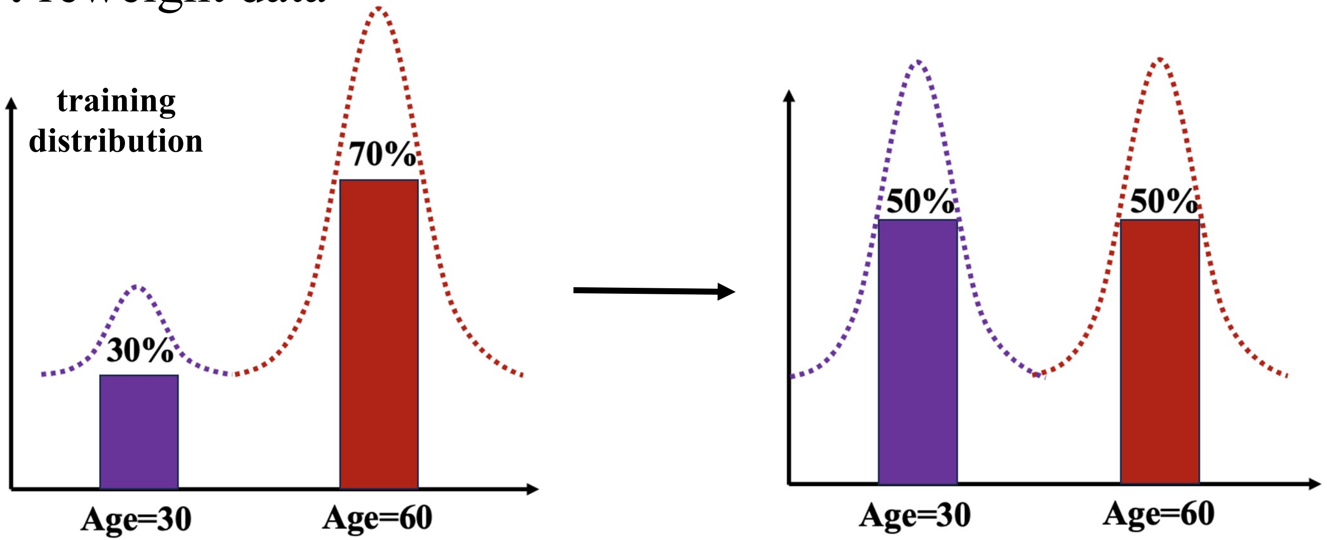
Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \le \rho\}$$

recall the objective

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

***f*-DRO:** reweight data



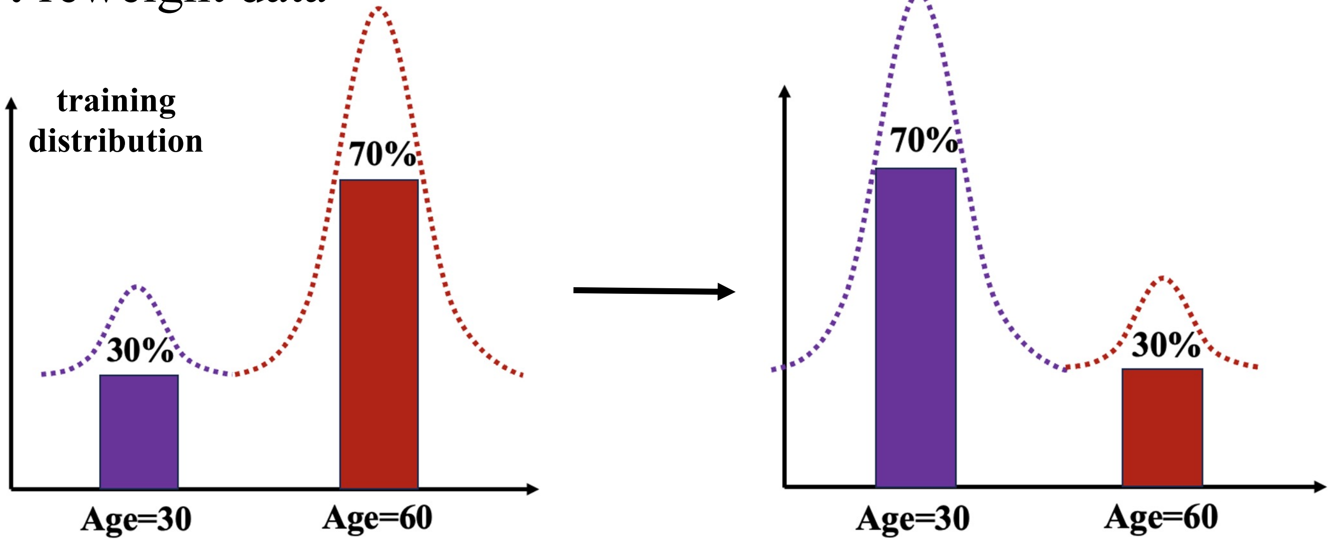
Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta\in\Theta} \sup_{Q\in\mathcal{P}} \mathbb{E}_{Z\sim Q}[\ell(\theta;Z)]$$

***f*-DRO:** reweight data



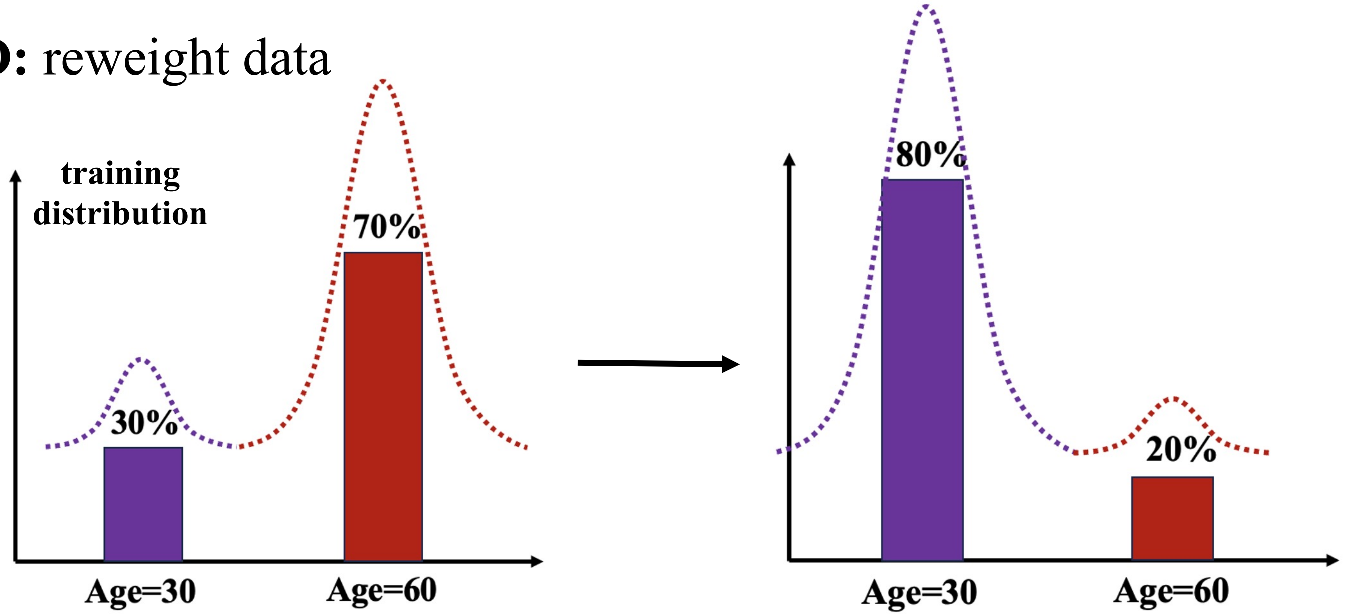
Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta\in\Theta} \sup_{Q\in\mathcal{P}} \mathbb{E}_{Z\sim Q}[\ell(\theta; Z)]$$

***f*-DRO:** reweight data



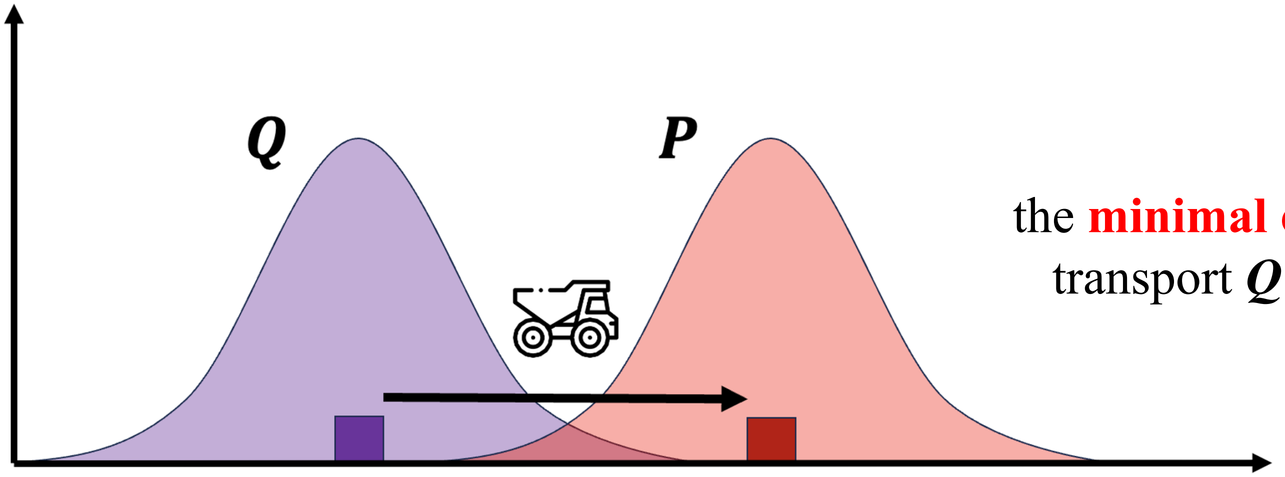
Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

Wasserstein distance**:** earth-mover's distance that considers geometry



the **minimal cost** to transport ***Q*** to ***P***

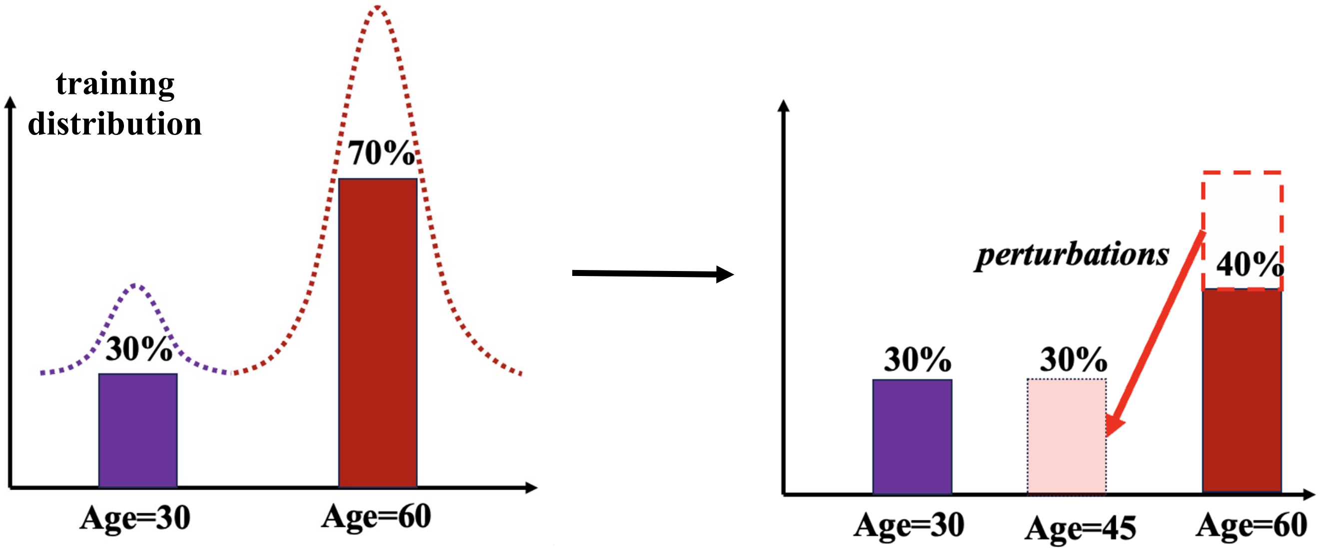
Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

Wasserstein-DRO: perturb data



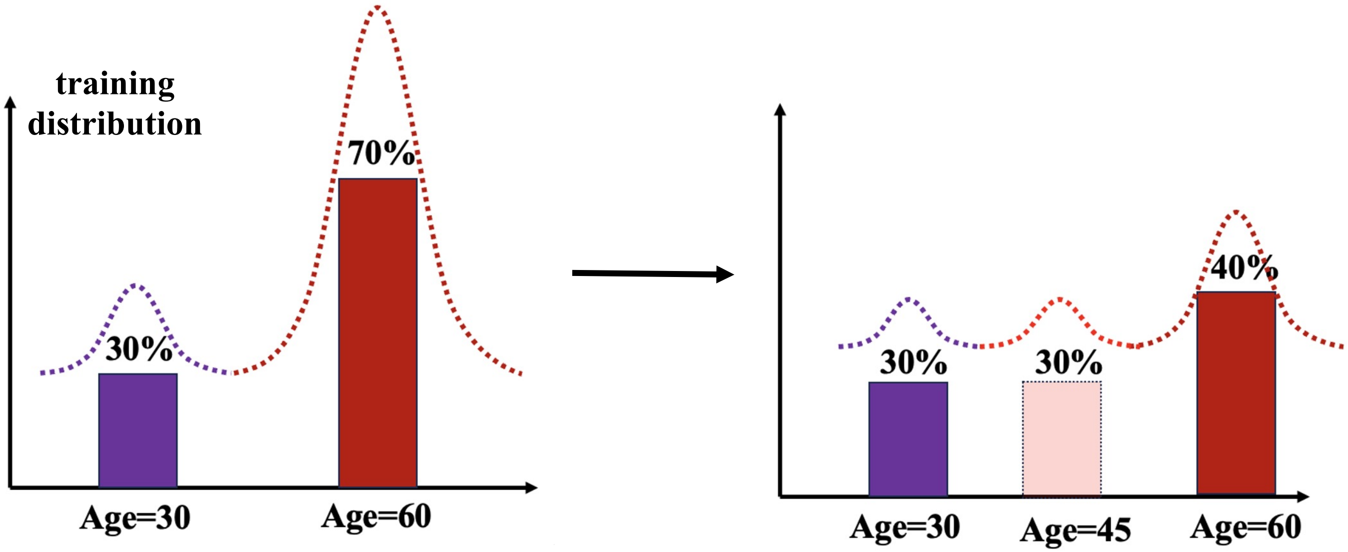
Examples: set of distributions we care about

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

Wasserstein-DRO: perturb data



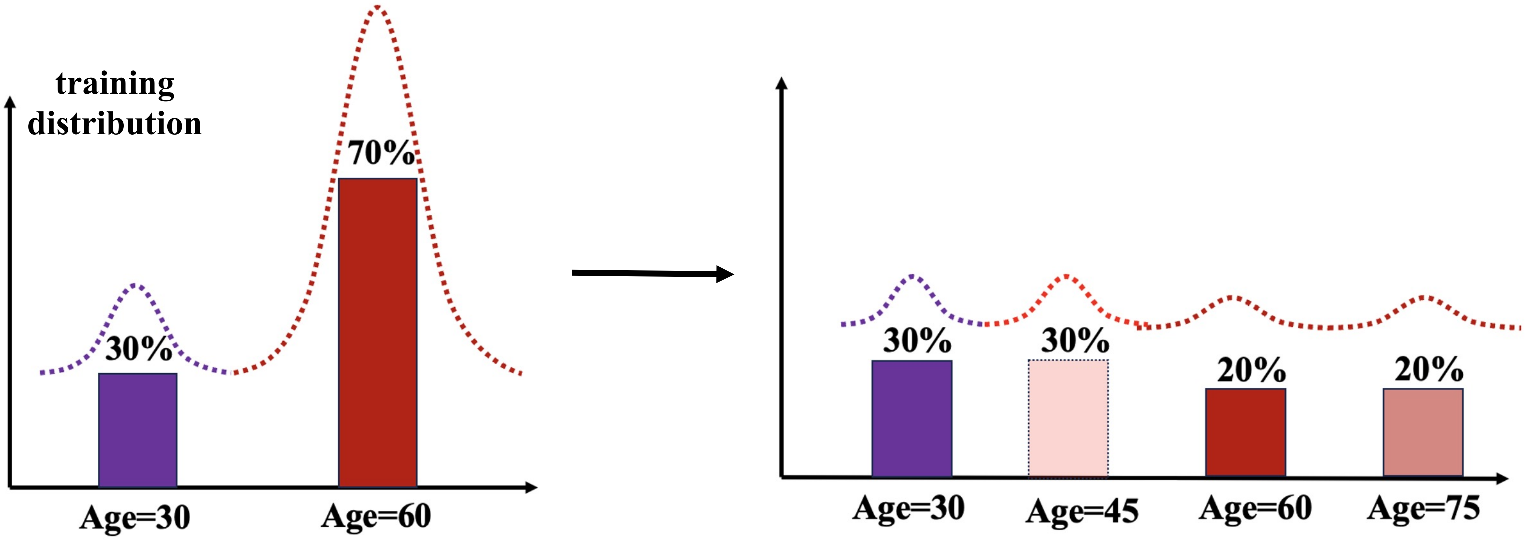
Examples: set of distributions we care about

$$\mathcal{P} = \{Q : Dist(Q, P_{train}) \leq \rho\}$$

recall the objective

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

Wasserstein-DRO: perturb data



Intuition: *f*-divergence vs Wasserstein distance

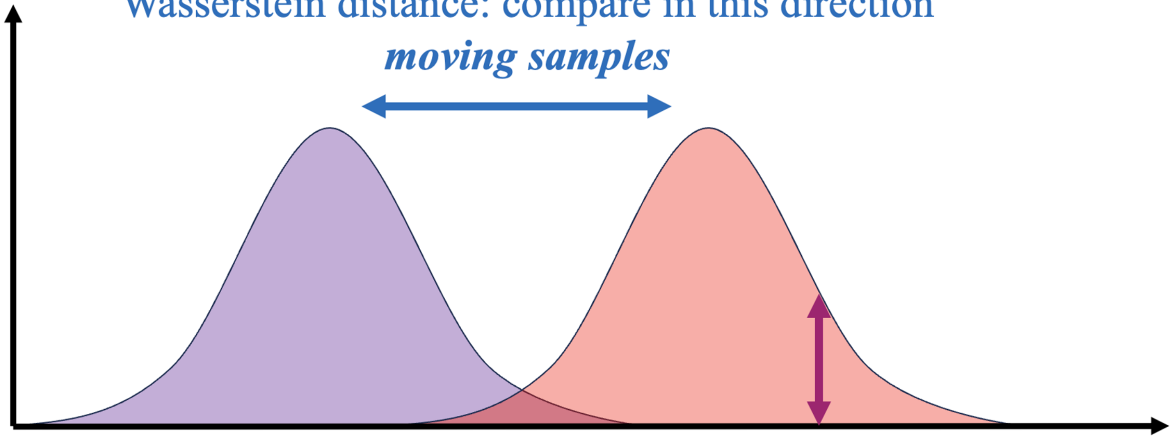
$$\mathcal{P} = \{Q : Dist(Q, P_{train}) \le \rho\}$$

recall the objective

$$\min_{\theta\in\Theta} \sup_{Q\in\mathcal{P}} \mathbb{E}_{Z\sim Q}[\ell(\theta; Z)]$$

Wasserstein distance: compare in this direction

moving samples



f-divergence: compare in this direction

comparing densities

DRO: set of distributions we care about: there are lots!

More Methods:

- Marginal DRO: only perturbs marginal distribution
- Sinkhorn DRO: adds entropy term to regularize Wasserstein distance
- Geometric DRO: uses geometric Wasserstein distance
- MMD DRO: uses MMD distance
- Holistic DRO: uses a mixture of distances
- Unified (OT) DRO: unifies Wasserstein distance and f-divergence

For more about DRO, please refer to the survey of DRO: Rahimian, H., & Mehrotra, S. (2019). Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659.

Duchi, J., Hashimoto, T., & Namkoong, H. (2023). Distributionally robust losses for latent covariate mixtures. Operations Research, 71(2), 649-664.
Wang, J., Gao, R., & Xie, Y. (2021). Sinkhorn distributionally robust optimization. arXiv preprint arXiv:2109.11926.
Liu, J., Wu, J., Li, B., & Cui, P. (2022). Distributionally robust optimization with data geometry. In NeurIPS.
Staib, M., & Jegelka, S. (2019). Distributionally robust optimization and generalization in kernel methods. In NeurIPS.
Bennouna, A., & Van Parys, B. (2022). Holistic robust data-driven decisions. arXiv preprint arXiv:2207.09560.
Blanchet, J., Kuhn, D., Li, J., & Taskesen, B. (2023). Unifying Distributionally Robust Optimization via Optimal Transport Theory. arXiv preprint arXiv:2308.05414.

DRO Package

An easy-to-use codebase for DRO

- Implement **12 typical DRO** algorithms
  - f-DRO: CVaR-DRO, KL-DRO, TV-DRO, χ^2-DRO
  - WDRO: Wasserstein DRO, Augmented WDRO, Satisficing WDRO
  - Sinkhorn-DRO
  - Holistic-DRO
  - Unified (OT)-DRO

dro 0.0.1

`pip install dro`

DRO makes a strong assumption

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q}[\ell(\theta; Z)]$$

Modeling

Carefully choose the set $\mathcal{P}$

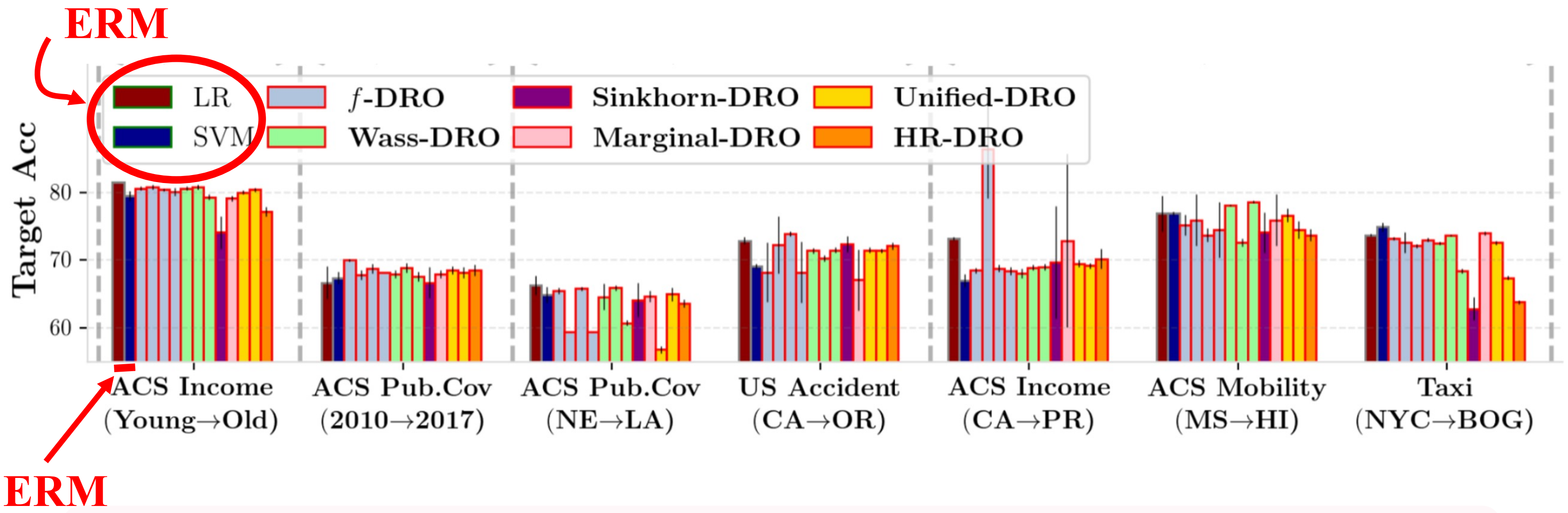
→

Goal

Do well on real distribution shifts!

Hope the worst-case distribution captures real shifts

Critical View of DRO: not better than ERM!



DRO does NOT show significant improvements over ERM!

Hard to choose this set of distributions P!!!

Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

Critical View of DRO: over-pessimism of the worst-case

Optimal *in-distribution* accuracy

$$1 - \min_{f \in \mathcal{F}} \mathbb{E}_{\widehat{P}^\star}[\ell(f(X), Y)].$$

Distribution	Source Domain	Worst-Distribution of KL-DRO		Worst-Distribution of χ^2-DRO		Worst-Distribution of TV-DRO		50 Target Domains' Quantile		
	$\epsilon = 0$	$\epsilon = 1e^{-2}$	$\epsilon = 1e^{-1}$	$\epsilon = 1e^{-1}$	$\epsilon = 5e^{-1}$	$\epsilon = 1e^{-1}$	$\epsilon = 2e^{-1}$	50%	25%	0%
LR	80.37	75.50	64.81	70.39	58.95	64.55	47.20	79.77	78.93	76.07
SVM	80.72	75.38	64.65	70.28	58.75	64.39	47.20	79.86	78.88	76.11
NN	80.26	75.55	65.57	71.08	61.13	63.66	44.65	79.81	78.52	75.08
RF	79.61	75.35	66.09	71.28	61.22	62.51	46.92	78.78	77.84	75.93
LGBM	81.74	76.18	66.76	72.23	63.02	61.85	45.01	80.51	79.47	76.43
XGB	81.29	75.84	66.31	71.92	62.73	61.45	45.47	80.13	79.13	75.08

The worst-case distribution is too conservative!

Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

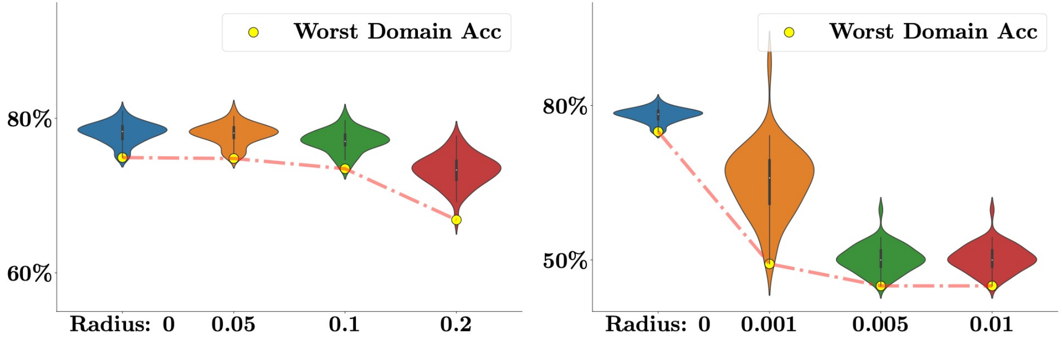
Critical View of DRO: mismatch with real target domains

Transfer accuracy from worst to target

$$\mathrm{TAcc}(\widehat{P}^\star, \widehat{Q}_t) = 1 - \mathbb{E}_{\widehat{Q}_t}[\ell(f^\star(X), Y)], \quad \text{where} \quad f^\star \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{\widehat{P}^\star}[\ell_{tr}(f(X), Y)].$$

test on the 50 target domains

model fit on the worst-case distribution



(a) ACS Income, KL-DRO

(b) ACS Income, Wasserstein-DRO

The worst-case distribution is NOT aligned with the 50 target domains!

Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

Hard to pick set of distributions; can we do better?

What if we were given a set of environments that we cared about?

Hard to pick set of distributions ***P***; can we do better?

Problem Setting:

- Train: ***Multiple*** training domains $P^1_{X,Y}, P^2_{X,Y}, \ldots, P^K_{X,Y}$
- Test: New domain $Q_{X,Y}$

Compare to DRO setting, more information about potential shifts!

Figure from Gulrajani, I., & Lopez-Paz, D. (2020, October). In Search of Lost Domain Generalization. In ICLR 2020.

Outline

Part 1: A critical review of existing approaches

- Distributionally Robust Optimization
- **Invariant Learning**
- Pretrained "Big" Models

} **make modeling assumptions** (Distributionally Robust Optimization, Invariant Learning)

scale up model & data (Pretrained "Big" Models)

Part 2: Shift to an inductive research philosophy

Part 3: Towards heterogeneity-<u>aware</u> machine learning

Part 4: Future Directions

Invariant Learning

Modeling → **Goal**

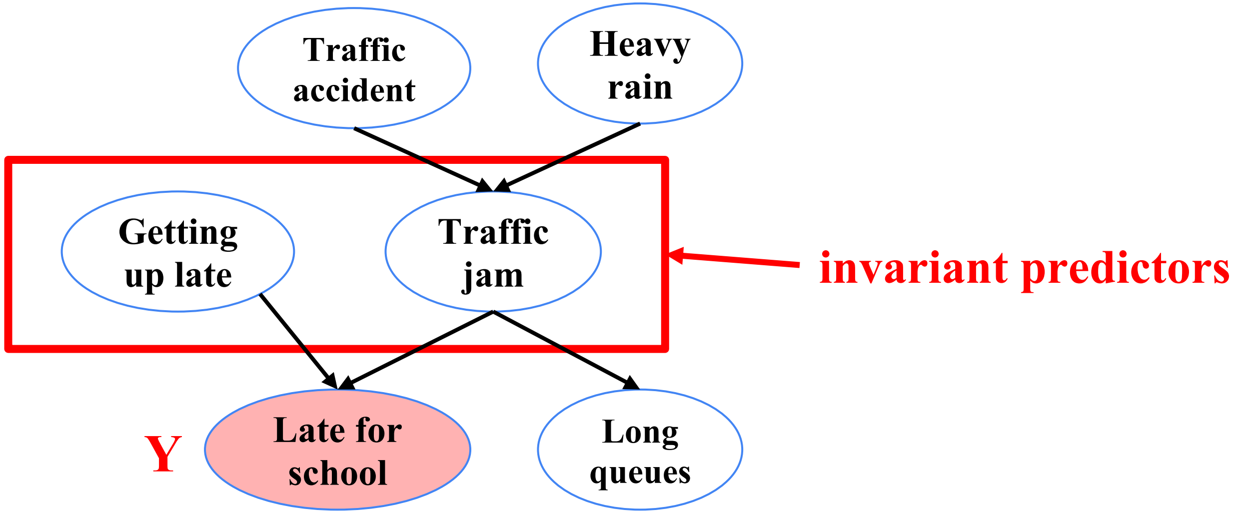
Learn an invariant mechanism across given environments

Generalize to new environments

Assume true invariant mechanism can be learned with given heterogeneous data

Invariant Learning: Invariant Causal Prediction

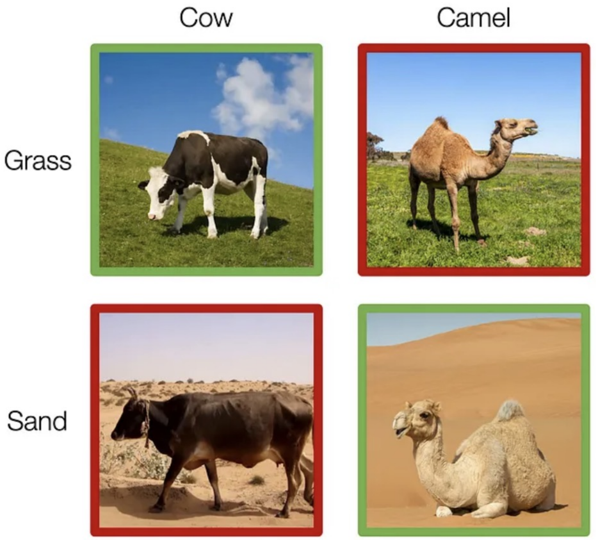
Find <u>subset</u> of covariates X with an **invariant** relationship to Y across environments!



Peters, J., Buhlmann, P., & Meinshausen, N. (2015). Causal inference using invariant prediction: identification and confidence intervals.
Figure from https://learn.saylor.org/mod/page/view.php?id=21614

Invariant Learning: Invariant Risk Minimization

Assume existence of <u>feature</u> ***Φ(X)*** such that ***Y|Φ(X)*** is **invariant** across environments. Then, learn this feature.



Task: classify between cows and camels

Use animals *Φ(X)* for prediction, rather than backgrounds!

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). <u>Invariant risk minimization</u>.
Figure from https://towardsdatascience.com/on-learning-in-the-presence-of-underrepresented-groups-8937434d3c85

Invariant Learning: Invariant Risk Minimization

Assume existence of <u>feature</u> ***Φ(X)*** such that ***Y|Φ(X)*** is **invariant** across environments. Then, learn this feature.

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi)$$

$$\text{subject to} \quad w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min}\, R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\mathrm{tr}}$$

invariance

Practical version:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(\Phi) + \lambda\cdot\|\nabla_{w|w=1.0}\, R^e(w\cdot\Phi)\|^2, \qquad \text{(IRMv1)}$$

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). <u>Invariant risk minimization</u>.

Invariance Assumption

- To deal with the potential distribution shifts, one common assumption is:

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*

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

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

- Some comments:
  - 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**.

M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor.

Maximal Invariant Predictor

- To obtain the invariant predictor $\Phi^*(X)$, we seek for:

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

$$\mathcal{I}_{\mathcal{E}} = \{\Phi(X) : Y \perp \mathcal{E} | \Phi(X)\} = \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\} \tag{6}$$

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:

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

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"

M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor.

Invariant Learning

More literature

S. Chang, et al. Invariant rationalization. In ICML, 2020.
M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor.
K. Ahuja, et al. Invariant risk minimization games. In ICML, 2020.
E. Rosenfeld, et al. The risks of invariant risk minimization.In ICLR, 2020.
D. Krueger, et al. Out-of-distribution generalization via risk extrapolation (rex). In ICML, 2021.
D. Mahajan, et al. Domain generalization using causal matching. In ICML, 2021.
P. Kamath, et al. Does invariant risk minimization capture invariance? In AISTATS, 2021.
B. Li, et al. Invariant information bottleneck for domain generalization. In AAAI, 2022.
H. Wang, et al. Provable domain generalization via invariant-feature subspace recovery. In ICML, 2022.
J. Fan, et al. Environment invariant linear least squares, 2023.

……

Methods and assumptions

	Distributionally Robust Optimization	Invariant Learning
Heterogeneity	Pre-defined set of distributions near training distribution	Pre-defined set of environments
Assumptions	Worst-case distribution guarantees generalization	Learn true invariant mechanism

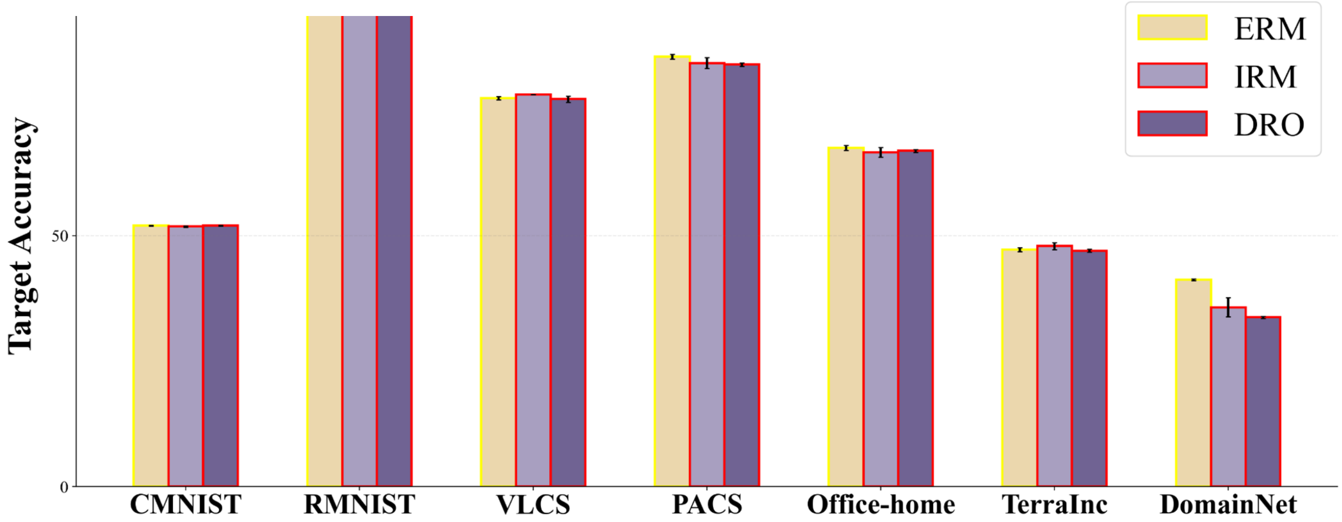
Do these assumptions work in practice?

Not Really! IRM does not beat ERM on Image Datasets!

In Search of Lost Domain Generalization

Ishaan Gulrajani*
Stanford University
igul222@gmail.com

David Lopez-Paz
Facebook AI Research
dlp@fb.com



Plot generated from Table 4 from Gulrajani, I., & Lopez-Paz, D. (2020, October). In Search of Lost Domain Generalization. In International Conference on Learning Representations.

Outline

Part 1: A critical review of existing approaches

- Distributionally Robust Optimization
- Invariant Learning
- **Pretrained "Big" Models**

} **make modeling assumptions** (Distributionally Robust Optimization, Invariant Learning)

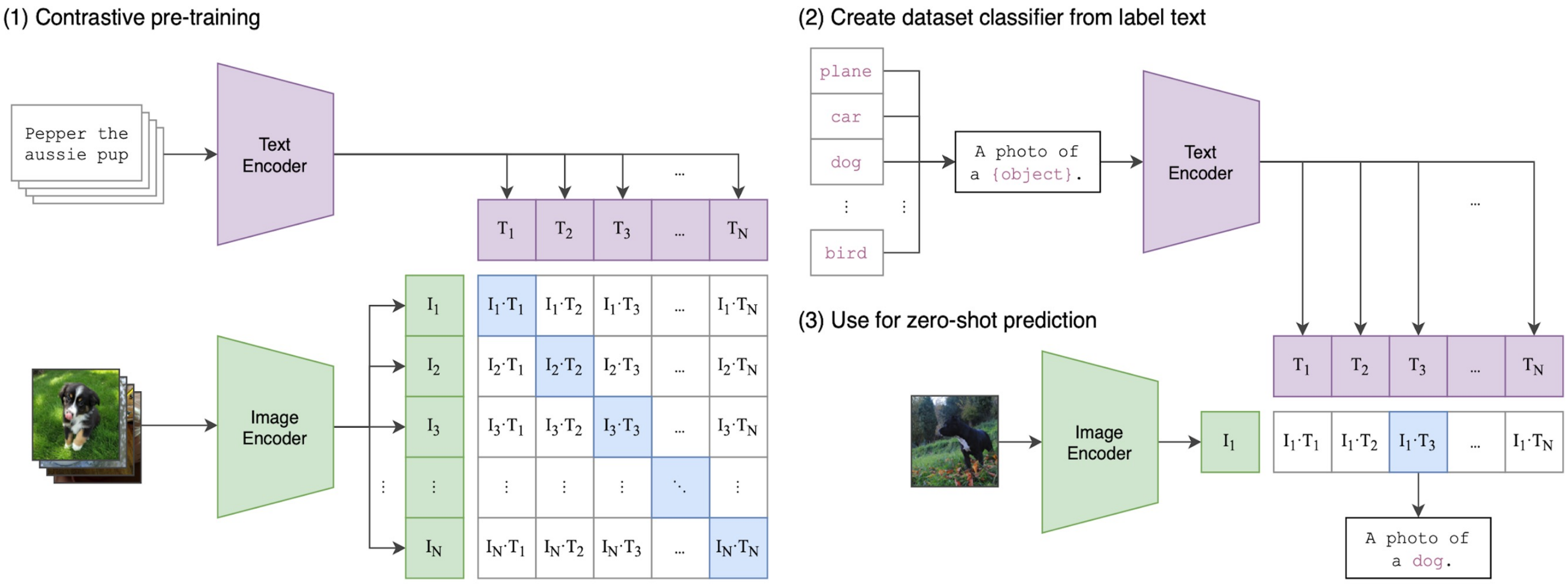
scale up model & data (Pretrained "Big" Models)

Part 2: Shift to an inductive research philosophy

Part 3: Towards heterogeneity-aware machine learning

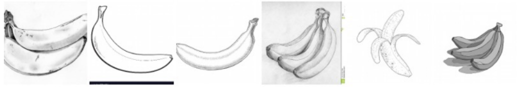
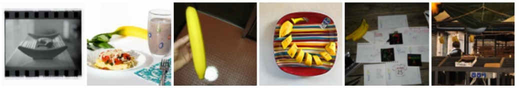
Part 4: Future Directions

CLIP: learn relationship between images and captions



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. Learning transferable visual models from natural language supervision. ICML, 2021.

“Big” Models: CLIP is robust to natural distribution shifts!

DATASET	IMAGENET RESNET101	CLIP VIT-L	Effective robustness
 ImageNet	76.2%	76.2%	
 ImageNet V2	64.3%	70.1%	+6%
 ImageNet Rendition	37.7%	88.9%	+51%
 ObjectNet	32.6%	72.3%	+40%
 ImageNet Sketch	25.2%	60.2%	+35%
 ImageNet Adversarial	2.7%	77.1%	+74%

Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever

Learning Transferable Visual Models From Natural Language Supervision (2021)

CLIP: scale up data

Supervised ImageNet training data	CLIP training data
● ~1M (image, label) pairs ● Data from one source ● Needs labelers	● ~400M (image, caption) pairs ● Data from all over the internet; more diverse ● No need for labelers; there is lots of (image, caption) data across the internet

Where are gains coming from? Data!

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)

Alex Fang[†] Gabriel Ilharco[†] Mitchell Wortsman[†] Yuhao Wan[†]

Vaishaal Shankar[◇] Achal Dave[◇] Ludwig Schmidt[†◦]

Abstract

Contrastively trained language-image models such as CLIP, ALIGN, and BASIC have demonstrated unprecedented robustness to multiple challenging natural distribution shifts. Since these language-image models differ from previous training approaches in several ways, an important question is what causes the large robustness gains. We answer this question via a systematic experimental investigation. Concretely, we study five different possible causes for the robustness gains: (i) the training set size, (ii) the training distribution, (iii) language supervision at training time, (iv) language supervision at test time, and (v) the contrastive loss function. Our experiments show that the more diverse training distribution is the main cause for the robustness gains, with the other factors contributing little to no robustness. Beyond our experimental results, we also introduce ImageNet-Captions, a version of ImageNet with original text annotations from Flickr, to enable further controlled experiments of language-image training.

~~Language supervision~~

Training distribution

~~Training set size~~

~~Loss function~~

~~Test-time prompting~~

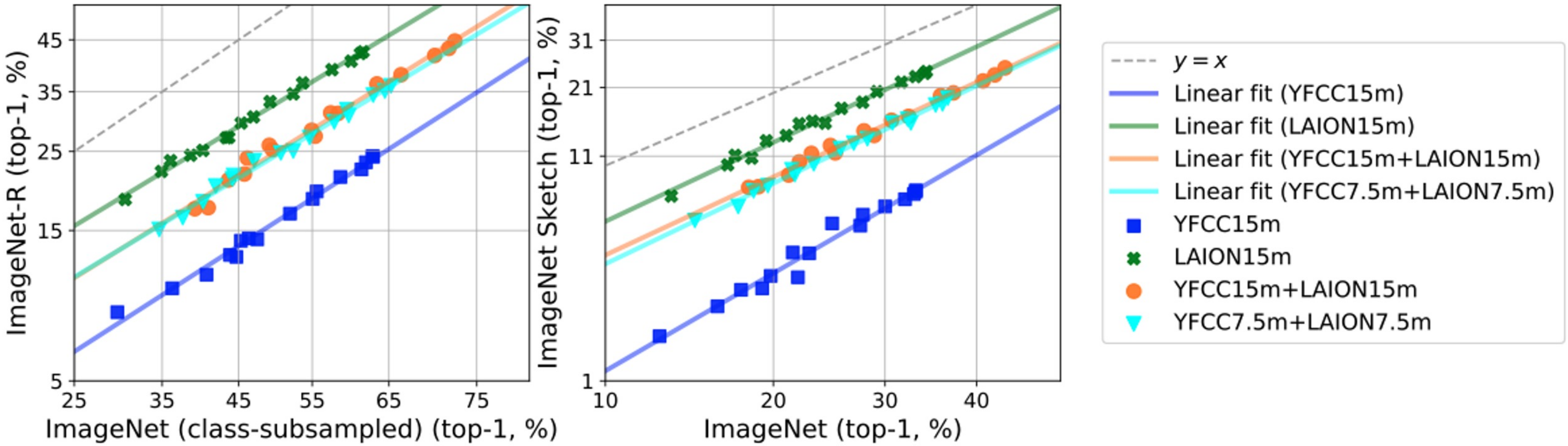
~~Model architecture~~

Is generalization under distribution shifts solved?

Just adding more data ≠ better

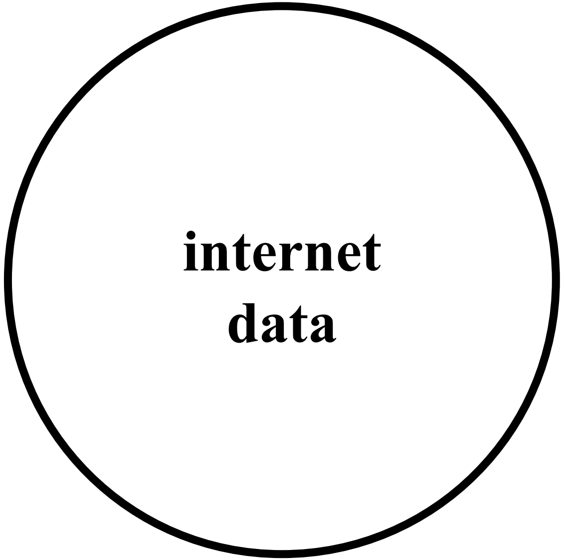
Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP

Thao Nguyen[1] Gabriel Ilharco[1] Mitchell Wortsman[1]
Sewoong Oh[1] Ludwig Schmidt[1,2]

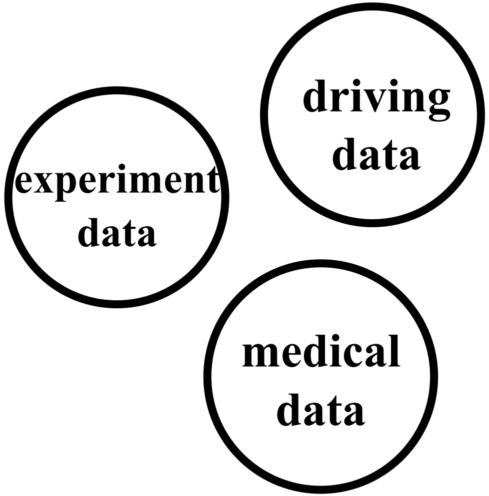


Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP
Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, Ludwig Schmidt

Sometimes you need (costly) specialized data!



$ cheap!



Many important applications!

$$$ expensive!

Not only in terms of dollars! E.g. time to perform an experiment

Two existing approaches to distribution shift

1. Make **modeling assumptions**

2. **Scale up data** and models

Strengths	Limitations
Clear assumptions about distribution shift	Current methods do not consistently provide robustness to many real distribution shifts
Works well to improve robustness to many real distribution shifts	Relevant, application-specific data can be costly to acquire

Two existing approaches to distribution shift

1. Make **modeling assumptions**

2. **Scale up data** and models

Can we do better?

Strengths	**Limitations**
Clear assumptions about distribution shift	Current methods do not consistently provide robustness to many real distribution shifts
Works well to improve robustness to many real distribution shifts	Relevant, application-specific data can be costly to acquire

Can we do better?

Don't just do this!

1. Make **modeling assumptions**

2. **Scale up data** and models

Instead, do this!

Understand the application

First understand your application and your data, and then make appropriate modeling assumptions!

Understand where you need data

Especially when data is costly, first identify what data is most helpful to collect!

Takeaways

- Empirically current methods (e.g. DRO, invariant learning) do **not** provide large gains.
- These methods make assumptions about the relationship between data distributions, but do **not** check them.
- We must model **real distributions shifts** rather than **hypothetical** ones, in an application-specific manner.
- For large pretrained models, we also need a better understanding of data distribution.
- In response, we propose carefully **understanding** the real distribution shift patterns in each application.

Outline

Part 1: A critical review of existing approaches

Part 2: Shift to an inductive research philosophy

- Inductive vs. Deductive
- Motivated examples
- The need for an inductive way

Part 3: Towards heterogeneity-aware machine learning

Part 4: Future Directions

Inductive vs. Deductive

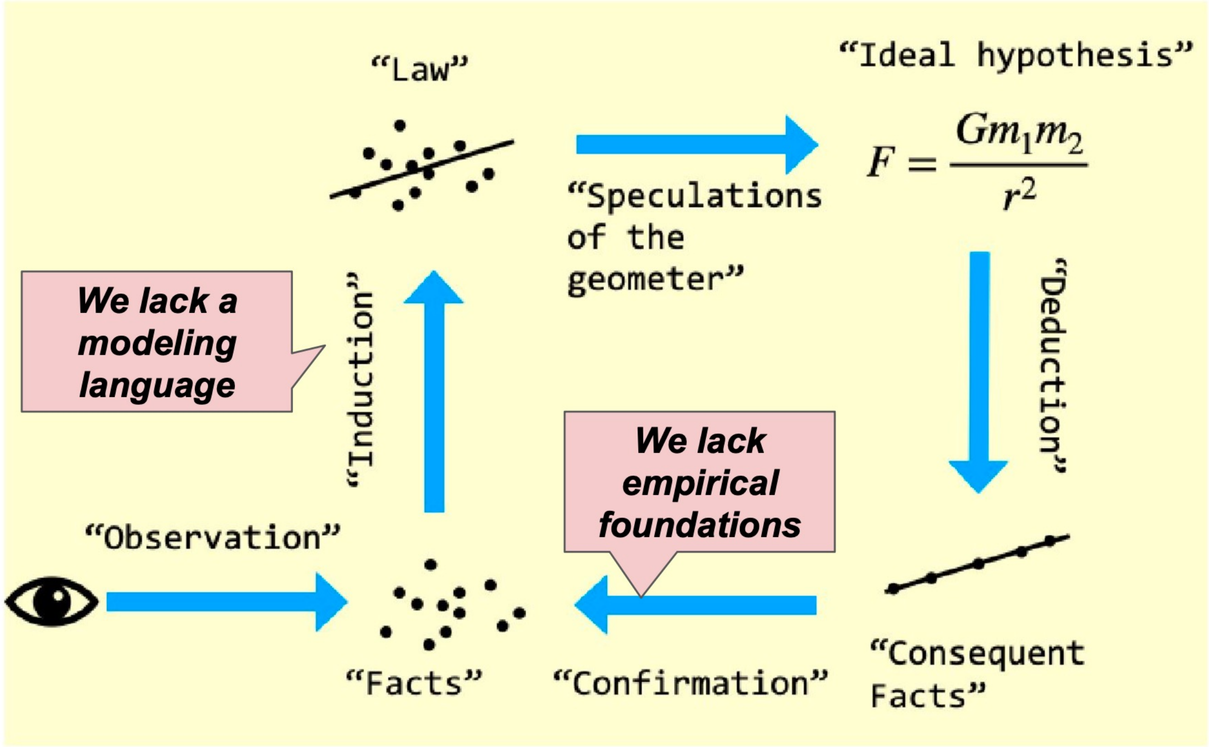
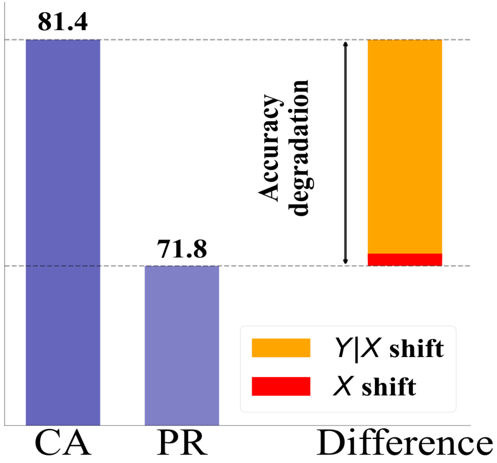


Figure from Christopher Ryan, Hong DRO Brown Bag, Columbia

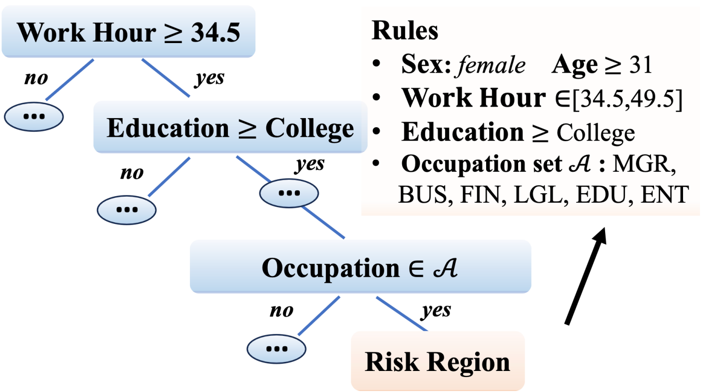
Motivated Example

Income prediction (source: CA, target: PR)

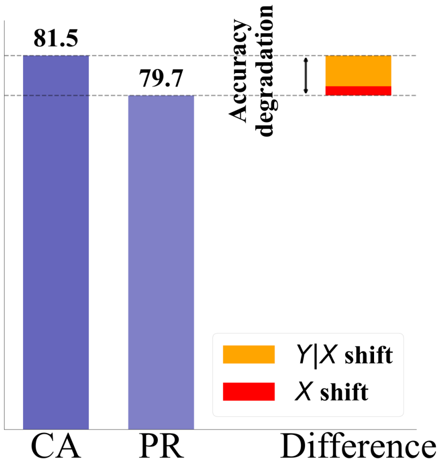
Inductive way!



Performance drops!



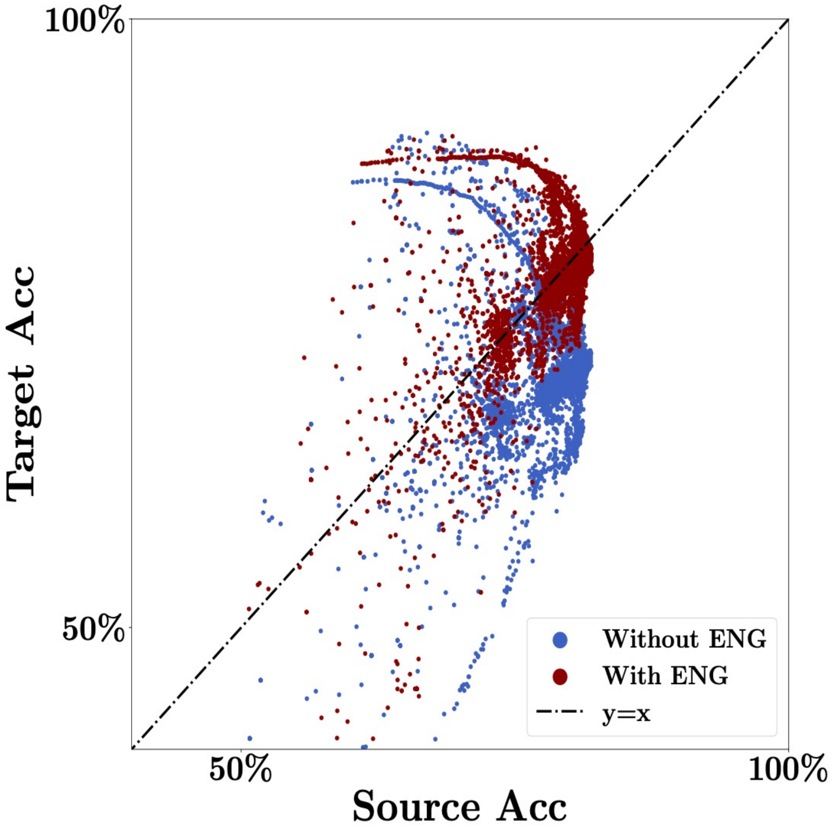
Identify covariate regions



Simply intervention works a lot!

Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

Motivated Example



Not only for one method but for ALL methods!

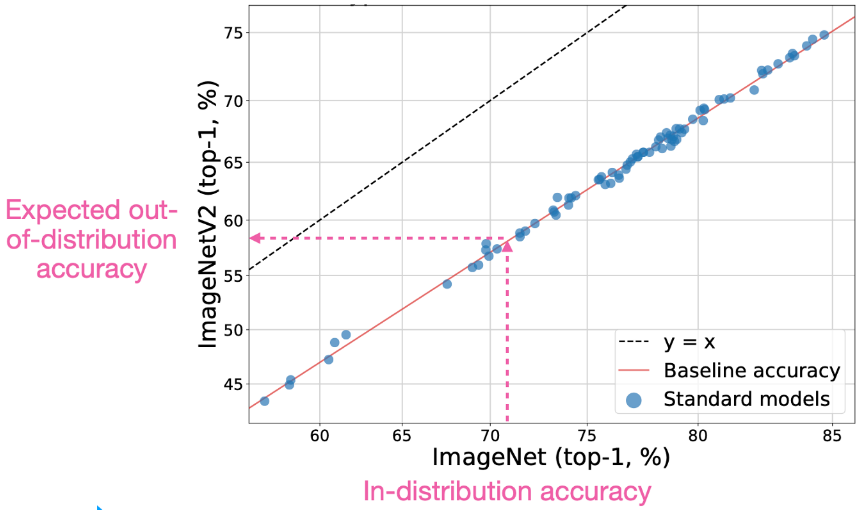
Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

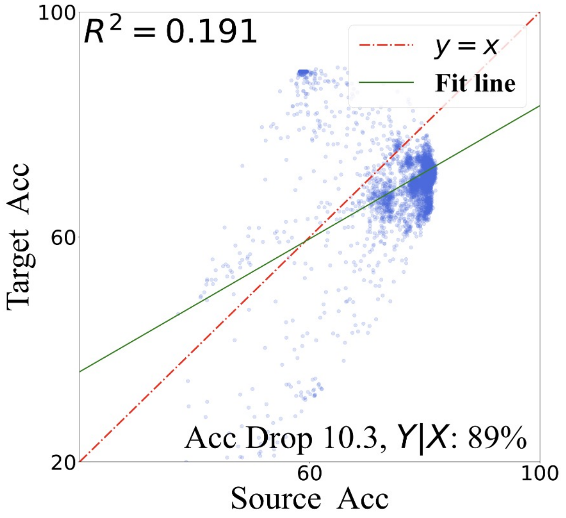
The need for Induction

- If not, we may have **FALSE** empirical discoveries!

Accuracy-on-the-line **doesn't** hold under strong $\boldsymbol{Y|X}$-shifts

- Source and target performances correlated ***only when X-shifts dominate***





Baseline **out-of-distribution accuracy** from **in-distribution accuracy.**

ACS Income (CA→ PR)

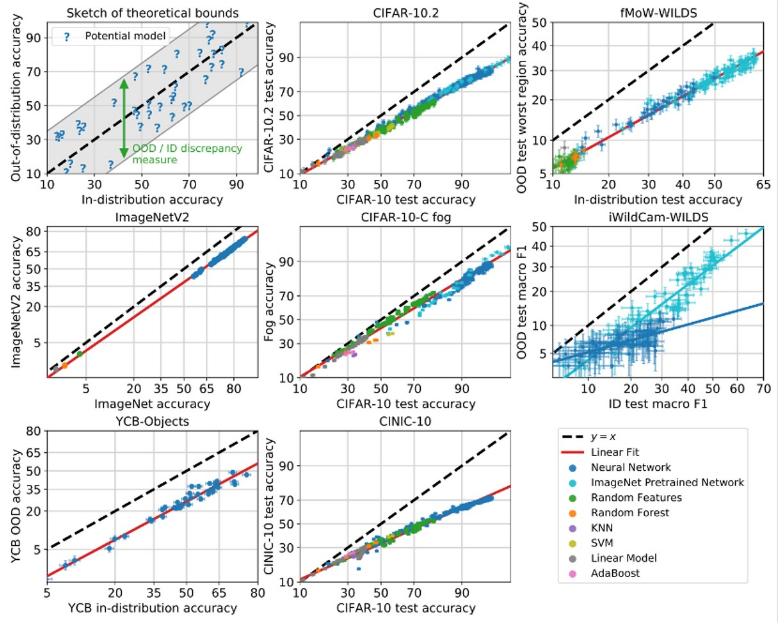
Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization.

On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets

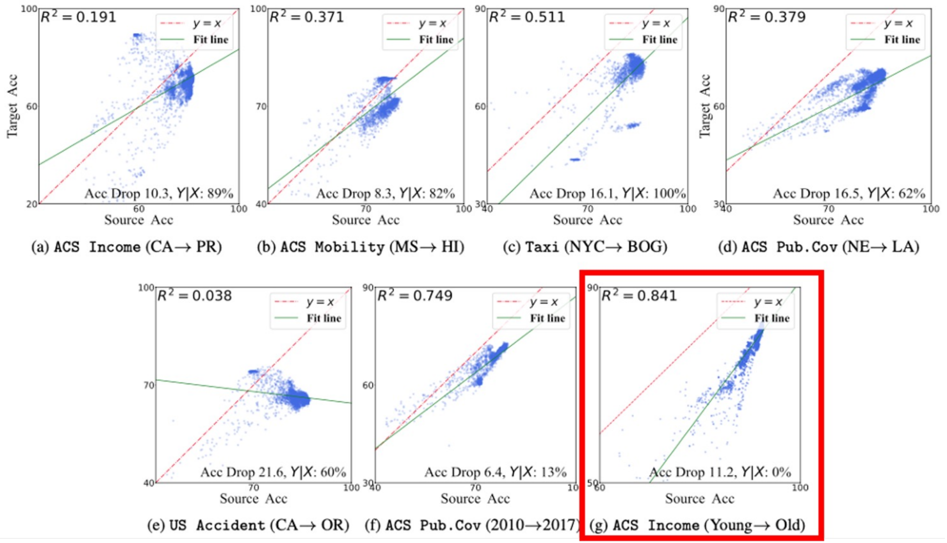
Accuracy-on-the-line **doesn't** hold under strong $Y|X$-shifts

- Source and target performances correlated ***only when X-shifts dominate***

Image datasets



WHYSHIFT



Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization.

On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets

The need for Induction

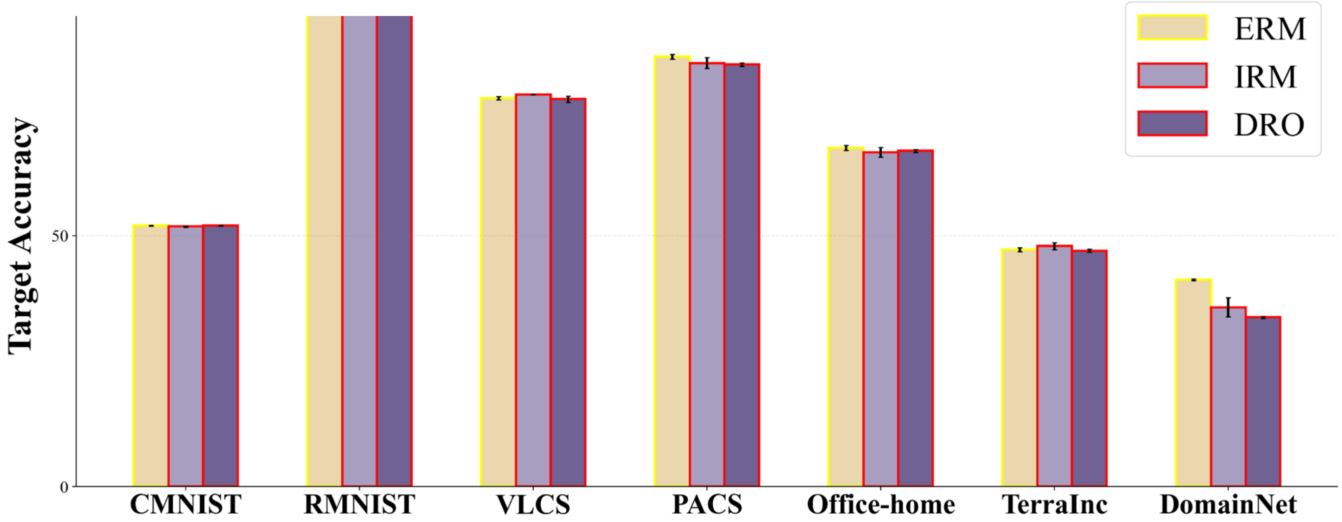
- If not, we may have **FALSE** empirical discoveries!
- If not, the empirical value of methods tailored for distribution shifts is **LIMITED**.

Recall: DRO & IRM don't outperform ERM on image data

In Search of Lost Domain Generalization

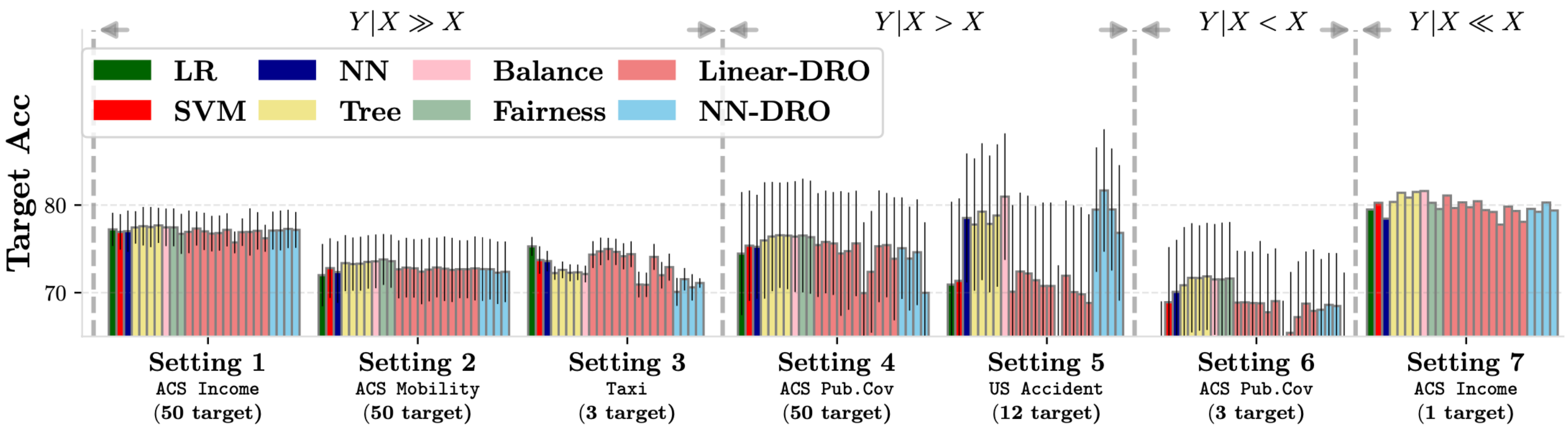
Ishaan Gulrajani*
Stanford University
igul222@gmail.com

David Lopez-Paz
Facebook AI Research
dlp@fb.com



Plot generated from Table 4 from Gulrajani, I., & Lopez-Paz, D. (2020, October). In Search of Lost Domain Generalization. In International Conference on Learning Representations.

Also: DRO doesn't outperform ERM on tabular data



Typical DRO methods do not significantly outperform traditional ERM or tree-based methods!

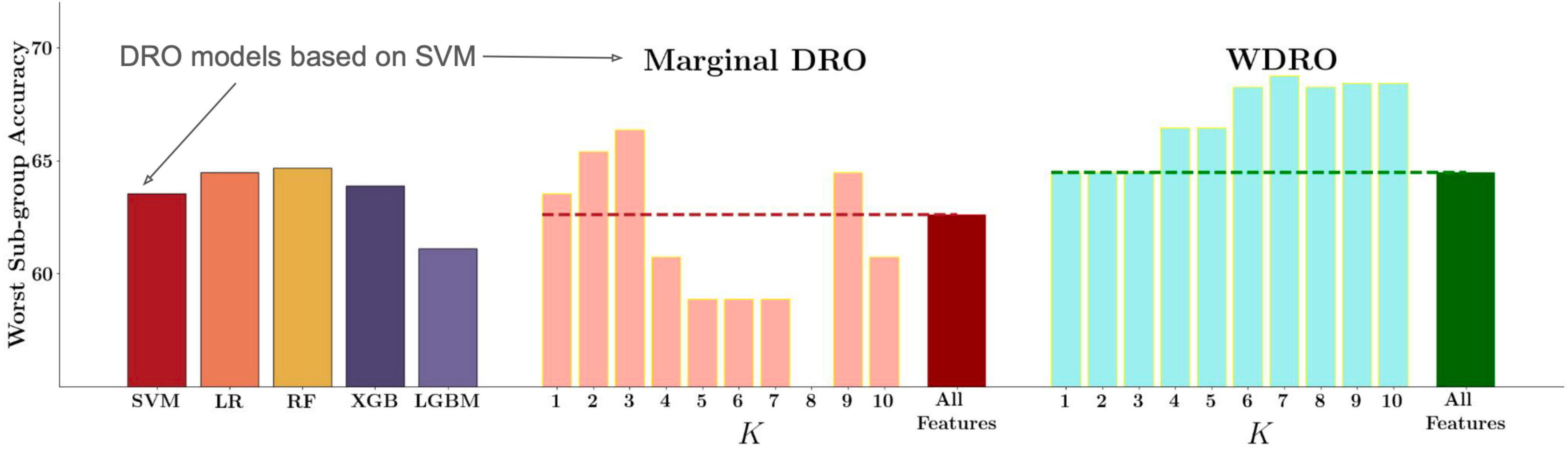
Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

The need for Induction

- If not, we may have **FALSE** empirical discoveries!
- If not, the empirical value of methods tailored for distribution shifts is **LIMITED**.
- If so, we can design/select **TARGETED** methods!

Inductive approach to ambiguity sets: *X*-shifts

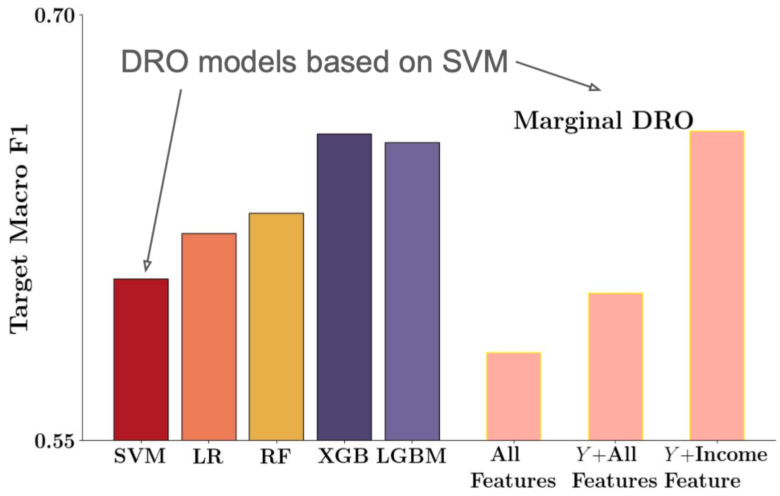
- Consider shifts induced by age groups: [20,25), [25,30), …, [75,100)
- Consider DRO methods (DH**N**'22) tailored to shifts on a subset of covariates
- Variable selection for ambiguity set: top-K with largest subgroup differences
- Performance varies a lot over variables selected



Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

Inductive approach to ambiguity sets: $Y|X$-shifts

- Consider $Y|X$-shifts from NE -> LA (public coverage task)
- Consider DRO methods that consider shifts on a subset of covariates and Y
- Variable selection for ambiguity set: Y | “income” suffers the largest shift
- Performance varies a lot over variables selected



Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

The need for Induction

- If not, we may have **FALSE** empirical discoveries!
- If not, the empirical value of methods tailored for distribution shifts is **LIMITED**.
- If so, we can design/select **TARGETED** methods!
- If so, we can obtain better improvements!

Analyze **data heterogeneity** to address the problems caused by **distribution shifts** from a **systematic** perspective

Outline

Part 1: A critical review of existing approaches

Part 2: Shift to an inductive research philosophy

Part 3: Towards heterogeneity-aware machine learning

- Tools to analyze data heterogeneity
- Model training
- Model evaluation & Improvement

Part 4: Future Directions

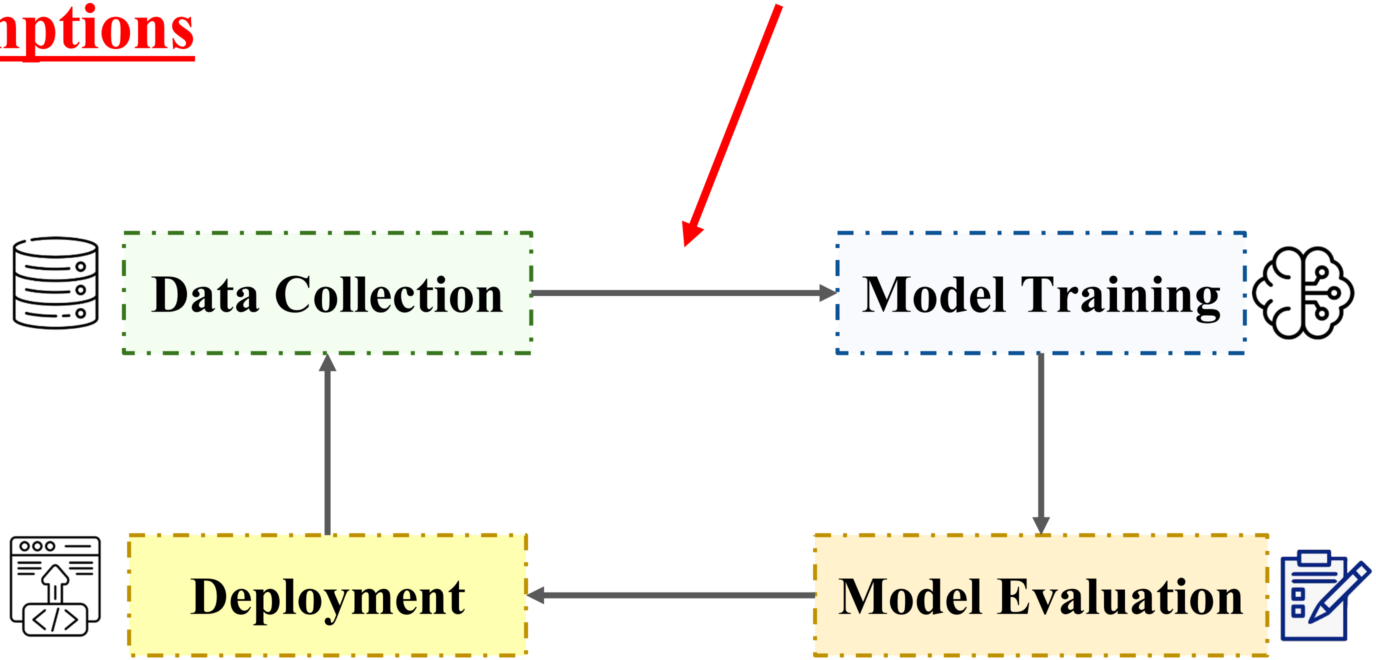
Recap: Terminology

- "Distribution shift" refers to mismatch between training distribution P and target distribution Q
- "Distributional robustness" refers to model performance **not** becoming worse even when Q is different from P
- "Heterogeneity" refers to the diverse mixture of distributions that generated the data, including both training and target

Recap: What’s left?

- How to measure the data heterogeneity?
- How to analyze the distribution shift patterns?

Stage 1: Analyze heterogeneity **before making modeling assumptions**



Perspective 1: It's important to understand if your data has heterogeneous subpopulations

After collecting data, we **need** to know

Does the training data contain ***sub-populations*** with ***different $Y|X$*** ?

Then we might want to model them separately!

In contrast, invariance methods assume the same $X \rightarrow Y$ across the entire population. This assumption can be inappropriate.

Discover heterogeneous subpopulations: **predictive heterogeneity**

Divide the dataset into subpopulations with different $Y|X$ by maximizing additional usable information gain

Definition

$$\sup_{\mathcal{E} \text{ is a split}} \mathbb{I}_{\mathcal{V}}(Y;X|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(Y;X)$$

→ *mutual information with* ***model constraints***

optimization algorithm

finite sample bounds

Xu, Y., Zhao, S., Song, J., Stewart, R., & Ermon, S. (2019, September). A Theory of Usable Information under Computational Constraints. In International Conference on Learning Representations.
Liu, J., Wu, J., Pi, R., Xu, R., Zhang, X., Li, B., & Cui, P. (2022, September). Measure the Predictive Heterogeneity. In The Eleventh International Conference on Learning Representations.

Preliminary: Mutual Information

Mutual Information

$$\mathbb{I}(X;Y) = H(Y) - H(Y|X)$$

- $H(Y)$: the entropy of Y
  - measuring the uncertainty of Y

the "**hardness**" of the original prediction task

- $H(Y|X)$: the conditional entropy of Y given X
  - measuring the uncertainty of Y after having access to some features X

- $\mathbb{I}(X;Y)$: how much information X can provide to **reduce the uncertainty of Y**

Liu, J., Wu, J., Pi, R., Xu, R., Zhang, X., Li, B., & Cui, P. (2022, September). Measure the Predictive Heterogeneity. In The Eleventh International Conference on Learning Representations.

Predictive Heterogeneity

$$\sup_{\mathcal{E}\ is\ a\ split} \mathbb{I}_{\mathcal{V}}(Y;X|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(Y;X) \quad \longleftrightarrow \quad \sup_{\mathcal{E}\ is\ a\ split} \mathbb{I}_{\mathcal{V}}(Y;X|\mathcal{E})$$

Equivalent

$$\mathbb{I}_{\mathcal{V}}(Y;X|\mathcal{E}) = \sum_{e\in\mathcal{E}} P(e)\mathbb{I}_{\mathcal{V}}(Y;X|\mathcal{E}=e)$$

$$= \sum_{e\in\mathcal{E}} P(e)\left(\boxed{H_{\mathcal{V}}(Y|e)} - H_{\mathcal{V}}(Y|X,e)\right)$$

the "hardness" of the prediction task in environment *e*

Liu, J., Wu, J., Pi, R., Xu, R., Zhang, X., Li, B., & Cui, P. (2022, September). Measure the Predictive Heterogeneity. In The Eleventh International Conference on Learning Representations.

Algorithm

the "hardness" of the prediction task in environment e

- **Objective Function:**

$$\min_{W \in \mathcal{W}_K} \mathcal{R}_{\mathcal{V}}(W, \theta_1^*(W), \ldots, \theta_K^*(W)) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} \ell_{\mathcal{V}}(f_{\theta_j^*}(x_i), y_i) + U_{\mathcal{V}}(W, Y_N) \right\},$$

$$\text{s.t.} \quad \theta_j^*(W) \in \arg\min_{\theta} \left\{ \mathcal{L}_{\mathcal{V}}(W, \theta) = \sum_{i=1}^{N} w_{ij} \ell_{\mathcal{V}}(f_\theta(x_i), y_i) \right\}, \quad \text{for } j = 1, \ldots, K,$$

- **Penalties reflect the difficulty of each 'sub-task'**
  - regression:

$$U_{\mathcal{V}_1}(W, Y_N) = \mathrm{Var}_{j \in [K]}(\overline{Y_N^j}) = \sum_{j=1}^{K} \left(\sum_{i=1}^{N} w_{ij} y_i \right)^2 \frac{1}{N \sum_{i=1}^{N} w_{ij}} - \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right)^2$$

  - classification:

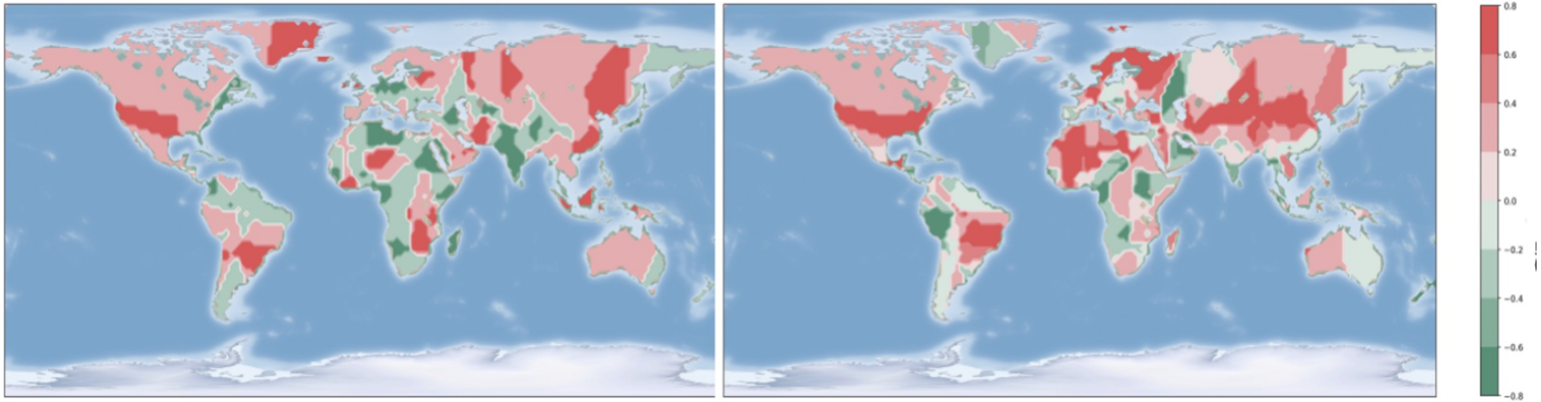
$$U_{\mathcal{V}_2}(W, Y_N) = -\sum_{j=1}^{K} \frac{1}{N} (\sum_{i=1}^{N} w_{ij}) H(Y_N^j),$$

Liu, J., Wu, J., Pi, R., Xu, R., Zhang, X., Li, B., & Cui, P. (2022, September). Measure the Predictive Heterogeneity. In The Eleventh International Conference on Learning Representations.

Example: predictive heterogeneity

Application in Agriculture

Task: predict ***crop yields*** from ***climate features***



true division of two crop types (rice vs wheat)

learned two sub-populations

probability of crop type / sub-population

learned sub-populations correspond to ***different crop types; model separately!***

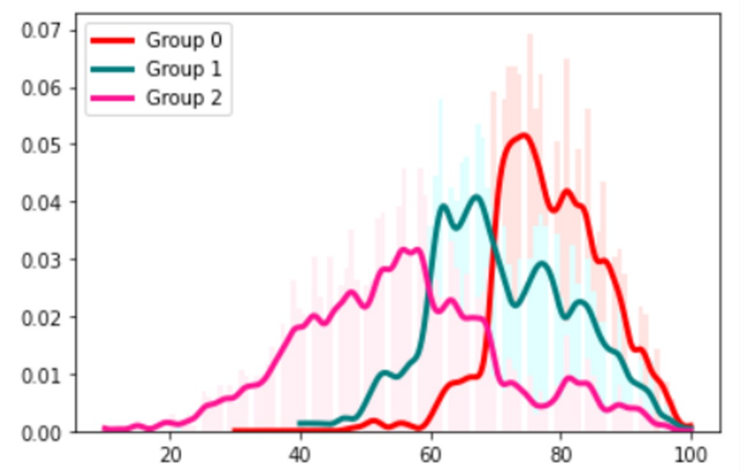
Liu, J., Wu, J., Pi, R., Xu, R., Zhang, X., Li, B., & Cui, P. (2022, September). Measure the Predictive Heterogeneity. In The Eleventh International Conference on Learning Representations.

Example: predictive heterogeneity

Task: predict ***mortality*** from ***symptom*** and ***underlying disease*** for people with COVID-19

Application in COVID-19

Age distributions of learned sub-populations



Top 4 Features:

Group 0: SPO2 Diabetes Renal Neurologic

Group 1: Diabetes SPO2 Neurologic Cardiovascular

Group 2: Fever Cough Renal Vomiting/Diarrhea

Serious covid symptoms!

ERM: SPO2 Renal Neurologic Diabetes

learned sub-populations correspond to ***different causes of death***

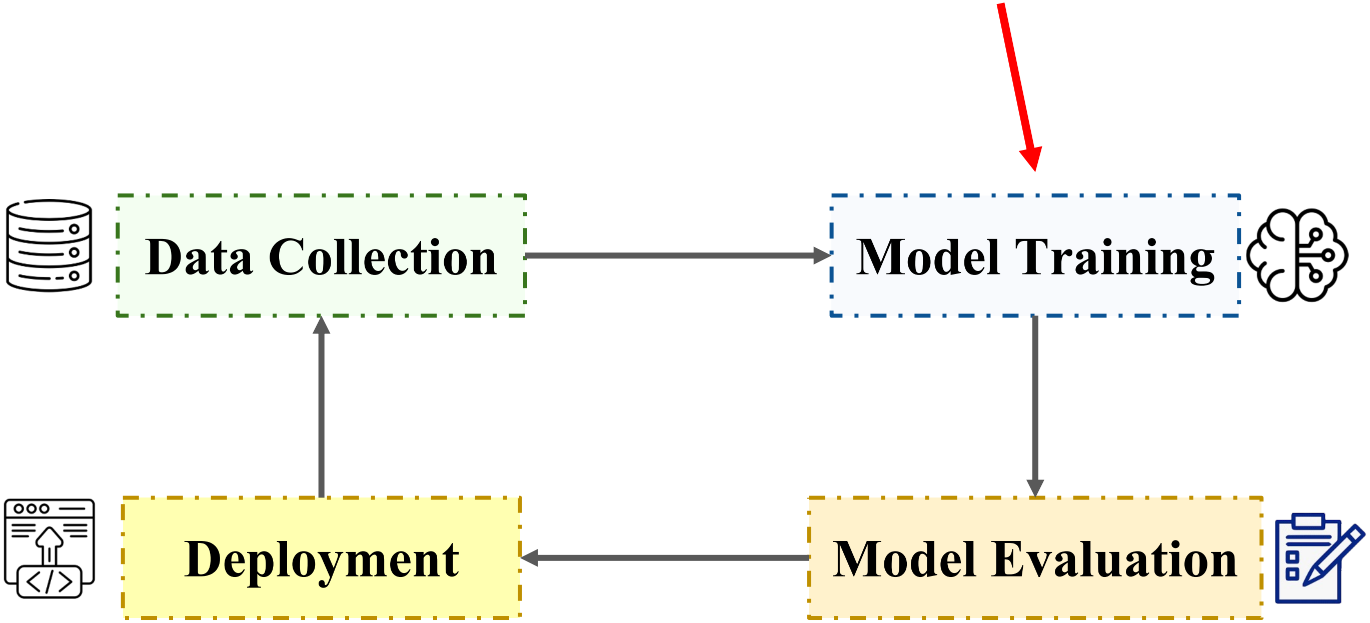
Liu, J., Wu, J., Pi, R., Xu, R., Zhang, X., Li, B., & Cui, P. (2022, September). Measure the Predictive Heterogeneity. In The Eleventh International Conference on Learning Representations.

Discovering heterogeneous subpopulations: where to go next?

- Limitations of this method: need more efficient ways to discover heterogeneous subpopulations
  - Scale up to larger tasks and models

- Next goal: *Understanding* heterogeneous subpopulations
  - Why do subpopulations have the $Y|X$ shifts that they have?
    - E.g .unobserved confounders, different generating process
    - How do these causes affect how we should model them?

Stage 2: Analyze heterogeneity **during model training**



Example 1: For invariant learning

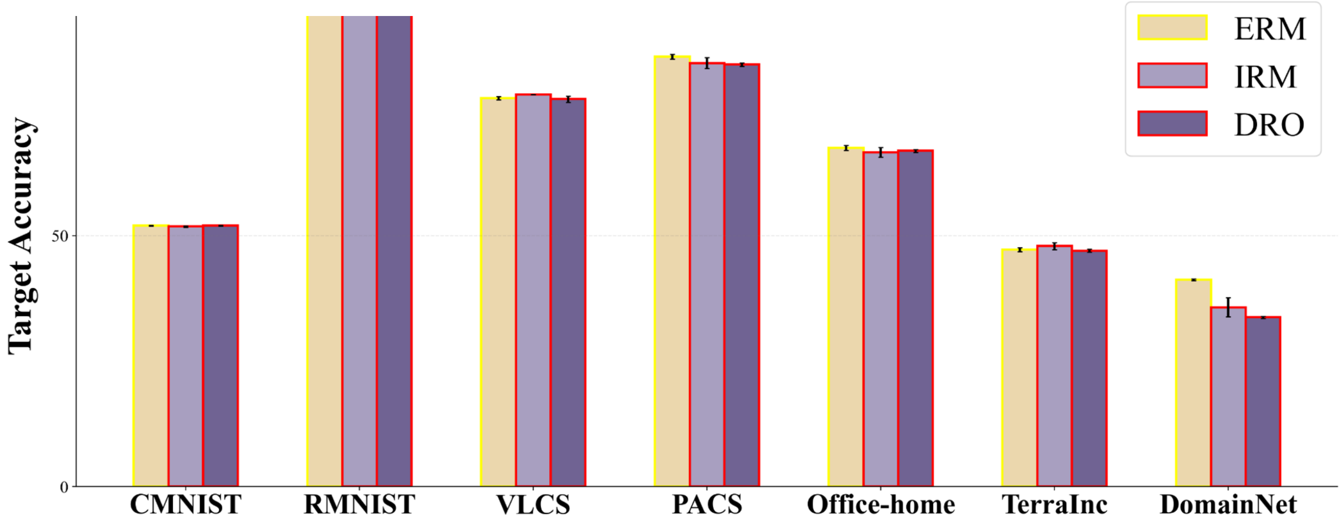
Example 2: For DRO

Recall: IRM doesn't outperform ERM on image data

IN SEARCH OF LOST DOMAIN GENERALIZATION

Ishaan Gulrajani*
Stanford University
igul222@gmail.com

David Lopez-Paz
Facebook AI Research
dlp@fb.com



Plot generated from Table 4 from Gulrajani, I., & Lopez-Paz, D. (2020, October). In Search of Lost Domain Generalization. In International Conference on Learning Representations.

Quality of Training Environments

- Invariance set

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

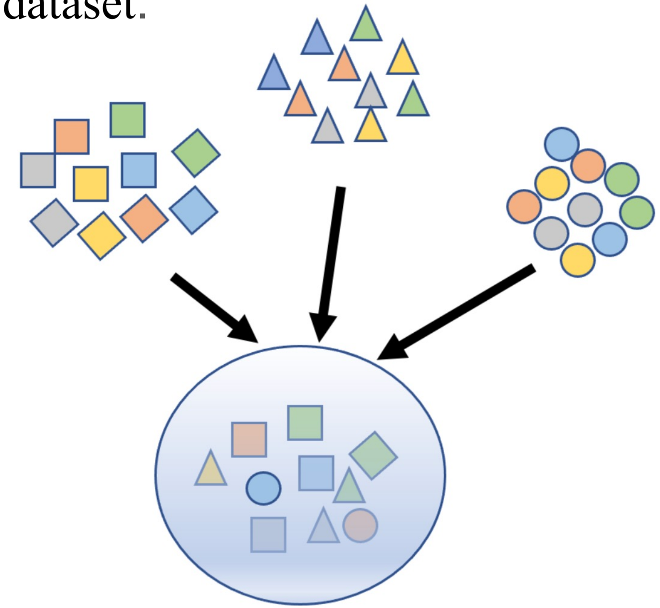
$$\mathcal{I}_{\mathcal{E}} = \{\Phi(X) : Y \perp \mathcal{E} | \Phi(X)\} = \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}]\}$$

- What happens when $\mathcal{E}$ is replaced by $\mathcal{E}_{tr}$?
  - **$\mathrm{supp}(\mathcal{E}_{tr}) \subset \mathrm{supp}(\mathcal{E})$**
  - **$\mathcal{I}_{\mathcal{E}} \subset \mathcal{I}_{\mathcal{E}_{tr}}$**
  - $\Phi^*(X)$ **NOT** invariant!

M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor.

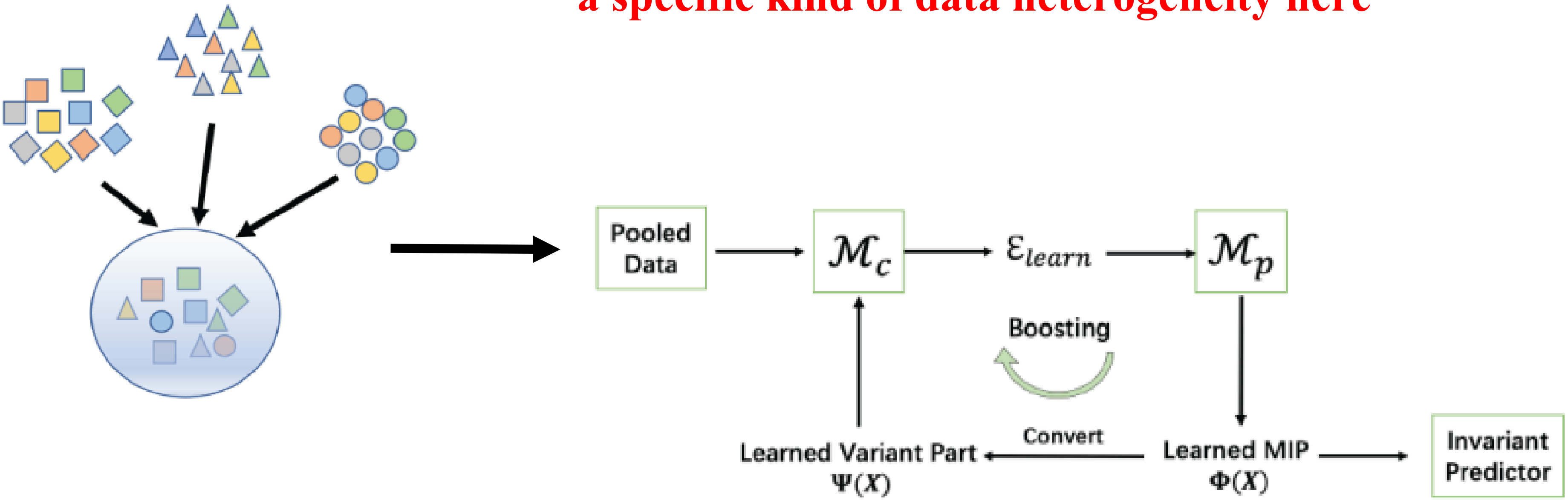
No Training Environments!

- 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.



Perspective 2: Explore **heterogeneous environments** during training

a specific kind of data heterogeneity here



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

Heterogeneity Identification Module

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

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 $\mathcal{N}(f_{\Theta_j}(\Psi), \sigma^2)$:

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

- 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:

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

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

Invariant Prediction Module

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

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

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

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

  - 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:

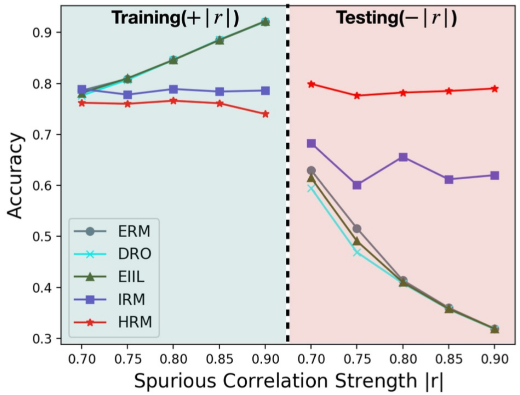
$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M[\ell(M \odot X^e, Y^e; \theta) + \alpha \|M\|_0] \quad (11)$$

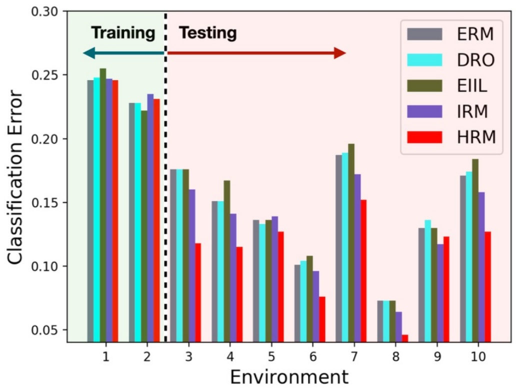
  - $\|M\|_0$ controls the number of selected features.

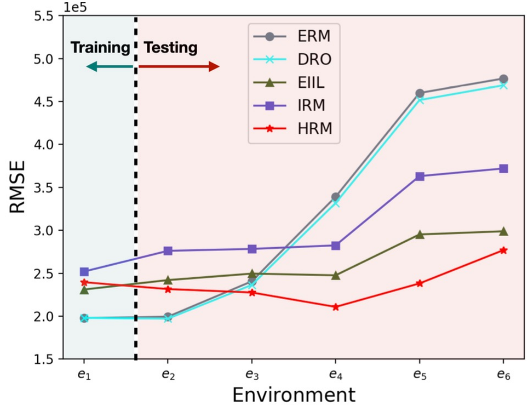
Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem ?

Performance

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

Example: heterogeneous risk minimization

- The two modules can boost each other
- The target accuracy is consistent with the heterogeneity of learned sub-populations

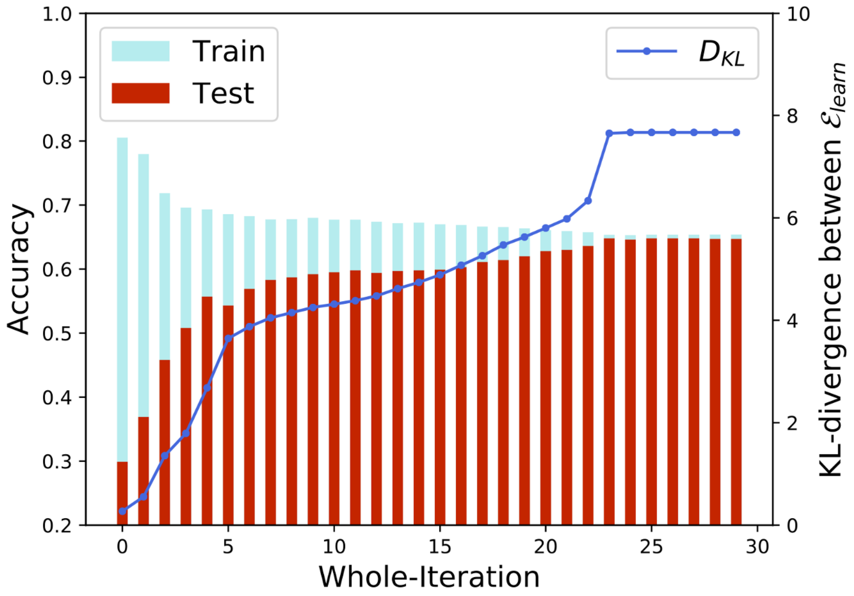


Figure from Figure 2 in Liu, J., Hu, Z., Cui, P., Li, B., & Shen, Z. Kernelized heterogeneous risk minimization. In *NeurIPS*, 2021.

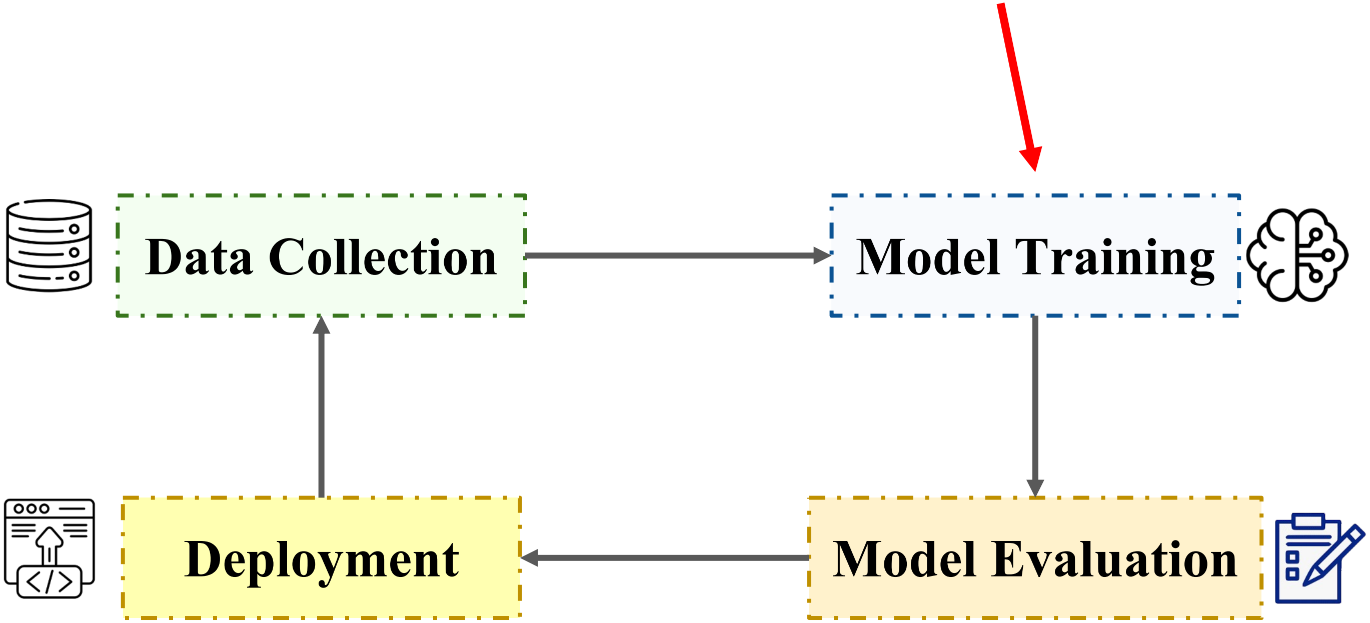
Example: heterogeneous risk minimization

Follow-up works on various tasks

- In recommendation:
  - **InvPref**
    Wang, Z. et al. Invariant preference learning for general debiasing in recommendation. In KDD.
  - **InvRL**
    Du, X. et al. Invariant Representation Learning for Multimedia Recommendation. In MM.

- On graph data:
  - **EERM**
    Wu, Q. et al. Handling Distribution Shifts on Graphs: An Invariance Perspective. In ICLR.
  - **LECI**
    Gui, S. et al. Joint Learning of Label and Environment Causal Independence for Graph Out-of-Distribution Generalization. In NeurIPS.
  - **GALA**
    Chen, Y. et al. Does Invariant Graph Learning via Environment Augmentation Learn Invariance?. In NeurIPS.

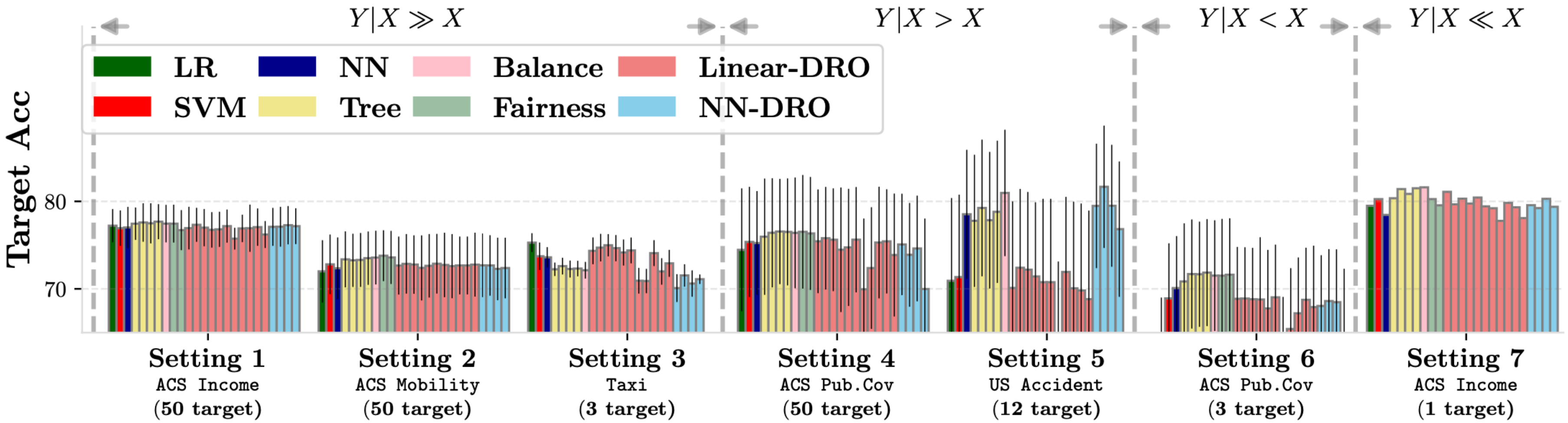
Stage 2: Analyze heterogeneity **during model training**



Example 1: For invariant learning

Example 2: For DRO

Recall: DRO doesn't outperform ERM on tabular data



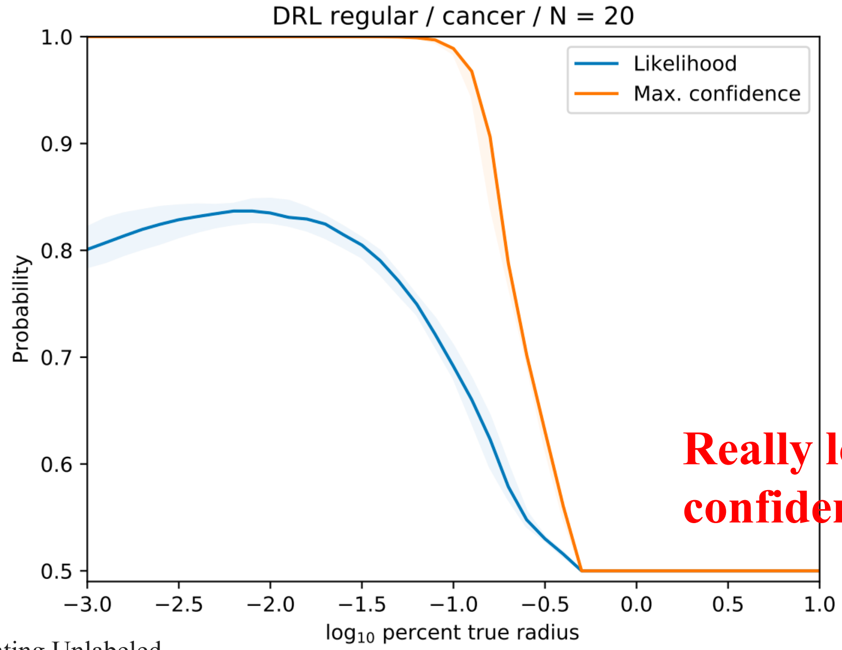
Typical DRO methods do not significantly outperform traditional ERM or tree-based methods!

Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

Recall: Over-pessimism problem of DRO

- When the uncertainty set is overwhelmingly large, the learned model predicts with low confidence.

$$\min_\theta \sup_{P:Dist(P,P_{tr})\le\epsilon} \mathbb{E}_P[\ell(\theta;X,Y)]$$

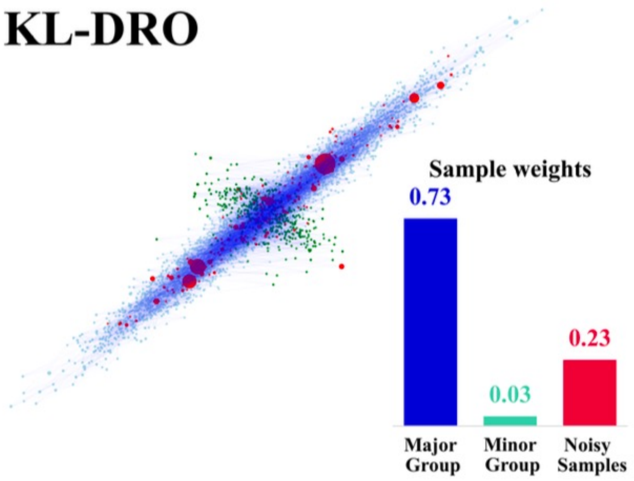


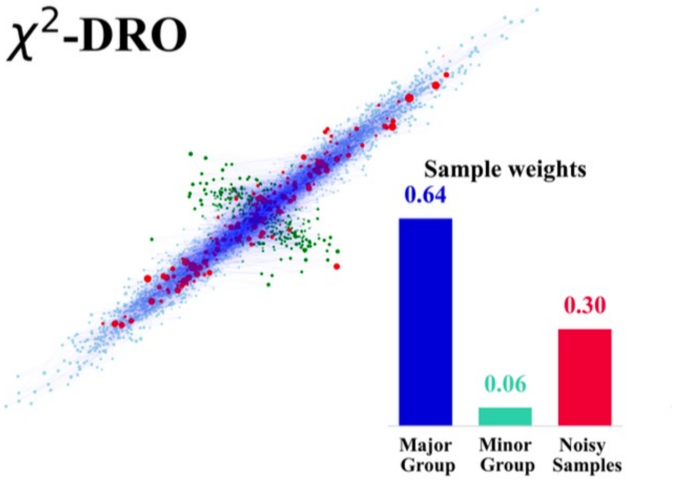
Really low confidence!

Figure from Frogner, C., Claici, S., Chien, E., & Solomon, J. (2021). Incorporating Unlabeled Data into Distributionally Robust Learning. *Journal of Machine Learning Research*, *22*(56), 1-46.

Perspective 3: Avoid noisy samples in DRO

another specific kind of data heterogeneity here





DRO methods focus too much on noisy samples!

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

Perspective 3: Avoid noisy samples in DRO

Example 1 (Weighted Least Square): Consider the data generation process as $Y = kX + \xi$, where $X, Y \in \mathbb{R}$ and random noise ξ satisfies $\xi \perp X$, $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2]$ (*abbr.* σ^2) is finite. Assume that the training dataset X_D consists of clean samples $\{x_c^{(i)}, y_c^{(i)}\}_{i\in[N_c]}$ and noisy samples $\{x_o^{(i)}, y_o^{(i)}\}_{i\in[N_o]}$ with $\sigma_c^2 < \sigma_o^2$. Consider the weighted least-square model $f(X) = \theta X$. Denote the sample weight of a clean sample $(x_c^{(i)}, y_c^{(i)})$ as $w_c^{(i)} \in \mathbb{R}_+, i \in [N_c]$, and the sample weight of a noisy sample $(x_o^{(i)}, y_o^{(i)})$ as $w_o^{(i)} \in \mathbb{R}_+, i \in [N_o]$ with $\sum_{i\in[N_c]} w_c^{(i)} + \sum_{i\in[N_o]} w_o^{(i)} = 1$. The variance of the estimator $\hat{\theta}$ is given by:

$$\mathrm{Var}[\hat{\theta}|X_D] = \frac{\sum_{i=1}^{N_c}(w_c^{(i)})^2(x_c^{(i)})^2\sigma_c^2 + \sum_{i=1}^{N_o}(w_o^{(i)})^2(x_o^{(i)})^2\sigma_o^2}{\left[\sum_{i=1}^{N_c} w_c^{(i)}(x_c^{(i)})^2 + \sum_{i=1}^{N_o} w_o^{(i)}(x_o^{(i)})^2\right]^2}, \tag{2.3}$$

DRO methods focus too much on noisy samples!

The parameter estimation will be quite random!

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

Data geometry matters

- Main Idea: data geometric information should be leveraged
  - High dimensional data lie on low dimensional manifolds
  - Noisy samples are mainly some **isolated** points
  - Hard samples (or minority group samples) are **continuous** within a neighborhood

Liu, J., Wu, J., Li, B., & Cui, P. (2022). Distributionally robust optimization with data geometry. NeurIPS, 2022.
Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

How to leverage?

- A geometry-aware distance metric: **Geometric Wasserstein Distance**

$$\mathcal{P}(G_0) = \{(p_i)_{i=1}^n \in \mathbb{R}^n | \sum_i p_i = 1, p_i \geq 0, i \in V\}$$

Definition 3.1 (Discrete Geometric Wasserstein Distance $\mathcal{GW}_{G_0}(\cdot,\cdot)$ [4]). *Given a finite graph G_0, for any pair of distributions $p^0, p^1 \in \mathscr{P}_o(G_0)$, define the Geometric Wasserstein Distance:*

$$\mathcal{GW}^2_{G_0}(p^0, p^1) := \inf_v \left\{ \int_0^1 \frac{1}{2} \sum_{(i,j)\in E} \kappa_{ij}(p) v_{ij}^2 dt : \frac{dp}{dt} + div_{G_0}(pv) = 0, p(0) = p^0, p(1) = p^1 \right\}, \quad (2)$$

where $v \in \mathbb{R}^{n\times n}$ denotes the velocity field on G_0, p is a continuously differentiable curve $p(t)$: $[0,1] \to \mathscr{P}_o(G_0)$, and $\kappa_{ij}(p)$ is a pre-defined interpolation function between p_i and p_j.

the support of distributions is restricted to the graph nodes

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

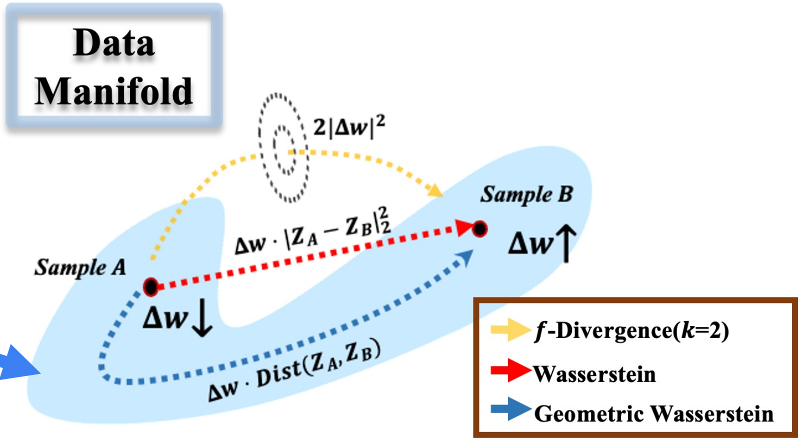
How to leverage?

Definition 3.1 (Discrete Geometric Wasserstein Distance $\mathcal{GW}_{G_0}(\cdot,\cdot)$ [4]). *Given a finite graph G_0, for any pair of distributions $p^0, p^1 \in \mathscr{P}_o(G_0)$, define the Geometric Wasserstein Distance:*

$$\mathcal{GW}^2_{G_0}(p^0,p^1) := \inf_v \left\{ \int_0^1 \frac{1}{2} \sum_{(i,j)\in E} \kappa_{ij}(p) v_{ij}^2 dt \;\; \frac{dp}{dt} + div_{G_0}(pv) = 0, p(0) = p^0, p(1) = p^1 \right\}, \quad (2)$$

where $v \in \mathbb{R}^{n\times n}$ denotes the velocity field on G_0, p is a continuously differentiable curve $p(t) : [0,1] \to \mathscr{P}_o(G_0)$, and $\kappa_{ij}(p)$ is a pre-defined interpolation function

The density transfers smoothly along the data manifold.



Liu, J., Wu, J., Li, B., & Cui, P. (2022). Distributionally robust optimization with data geometry. NeurIPS, 2022.
Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

How to leverage?

- A geometry-aware distance metric: **Geometric Wasserstein Distance**
- Geometry-Aware **calibration terms**

$$\min_{\theta\in\Theta} \underbrace{\sup_{\mathbf{q}:\mathcal{GW}^2_{G_N}(\hat{P}_X,\mathbf{q})\leq\rho}}_{\text{Geometric Wasserstein set}} \left\{\mathcal{R}_N(\theta,\mathbf{q}) := \sum_{i=1}^{N} q_i\ell(f_\theta(x_i),y_i) - \underbrace{\frac{\alpha}{2}\cdot\sum_{(i,j)\in E} w_{ij}q_iq_j(\ell_i-\ell_j)^2}_{\text{Calibration Term I}} - \underbrace{\beta\cdot\sum_{i=1}^{N} q_i\log q_i}_{\text{Calibration Term II}}\right\}$$

Graph total variation: penalize noisy samples

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

How to leverage?

- A geometry-aware distance metric: **Geometric Wasserstein Distance**
- Geometry-Aware **calibration terms**

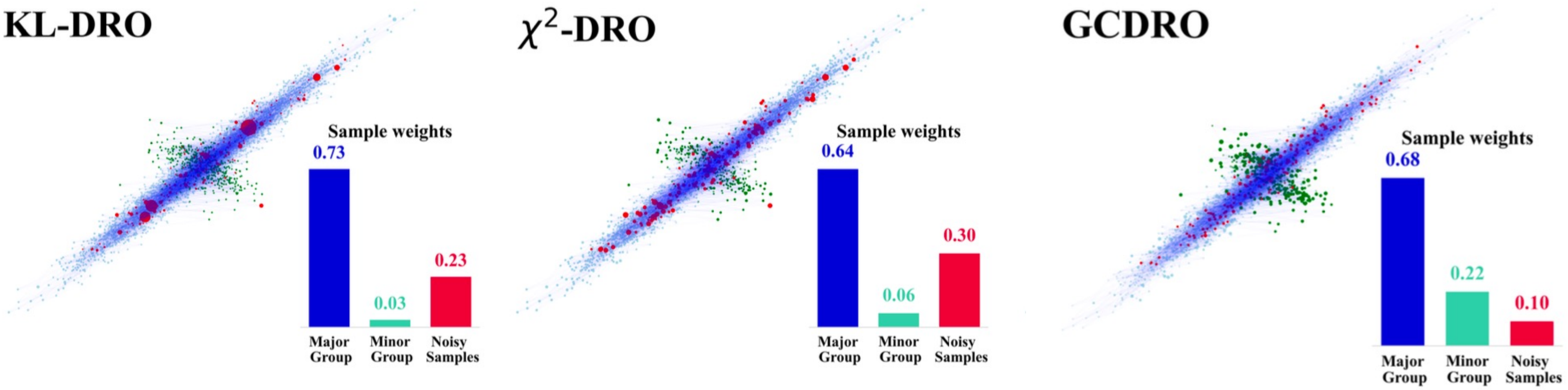
$$\min_{\theta\in\Theta} \underbrace{\sup_{\mathbf{q}:\mathcal{GW}^2_{G_N}(\hat{P}_X,\mathbf{q})\le\rho}}_{\text{Geometric Wasserstein set}} \left\{\mathcal{R}_N(\theta,\mathbf{q}) := \sum_{i=1}^{N} q_i \ell(f_\theta(x_i),y_i) - \underbrace{\frac{\alpha}{2}\cdot\sum_{(i,j)\in E} w_{ij}q_iq_j(\ell_i-\ell_j)^2}_{\text{Calibration Term I}} - \underbrace{\beta\cdot\sum_{i=1}^{N} q_i\log q_i}_{\text{Calibration Term II}}\right\}$$

Gradient of sample weights:

$$\frac{dq_i}{dt} = \sum_{(i,j)\in E} w_{ij}\xi_{ij}\Big(\mathbf{q},\ \ell_i-\ell_j+\beta(\log q_j-\log q_i)+\alpha\big(\sum_{h\in N(j)}(\ell_h-\ell_j)^2 w_{jh}q_h - \sum_{h\in N(i)}(\ell_h-\ell_i)^2 w_{ih}q_h\big)\Big)$$

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

Results



lower the sample weights on noisy samples

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

Side product: free energy implications

- Our objective

$$\mathcal{R}_N(\theta, \mathbf{q}) := \sum_{i=1}^{N} q_i \ell(f_\theta(x_i), y_i) - \underbrace{\frac{\alpha}{2} \cdot \sum_{(i,j)\in E} w_{ij} q_i q_j (\ell_i - \ell_j)^2}_{\text{Calibration Term I}} - \underbrace{\beta \cdot \sum_{i=1}^{N} q_i \log q_i}_{\text{Calibration Term II}} \}$$

- Free energy function

$$\mathcal{E}(\mathbf{q}) = \underbrace{\mathbf{q}^\top K \mathbf{q}}_{\text{Interaction Energy}} + \underbrace{\mathbf{q}^\top V}_{\text{Potential Energy}} - \underbrace{\beta \sum_{i=1}^{N} (-q_i \log q_i)}_{\text{Temperature}\times\text{Entropy}} = -\mathcal{R}_N(\theta, \mathbf{q})$$

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

Side product: a free energy understanding of DRO

Method	Energy Type			Specific Formulation			
	Interaction	Potential	Entropy	K	V	$H[\mathbf{q}]$	$\mathscr{P}$
KL-DRO	✗	✓	✓	-	$-\vec{\ell}$	$H[\mathbf{q}]$	Δ_N
χ^2-DRO	✓	✓	✗	λI	$-\vec{\ell}$	-	Δ_N
MMD-DRO	✓	✓	✗	Kernel Gram Matrix K	$-\vec{\ell} - \frac{2\lambda}{N} K^\top \mathbf{1}$	-	Δ_N
Marginal χ^2-DRO	✗	✓	✗	-	$-(\vec{\ell} - \eta)_+$	-	Δ_N with Hölder continuity
GDRO	✗	✓	✓	-	$-\vec{\ell}$	$H[\mathbf{q}]$	Geometric Wasserstein Set
GCDRO	✓	✓	✓	Interaction Matrix K	$-\vec{\ell}$	$H[\mathbf{q}]$	Geometric Wasserstein Set

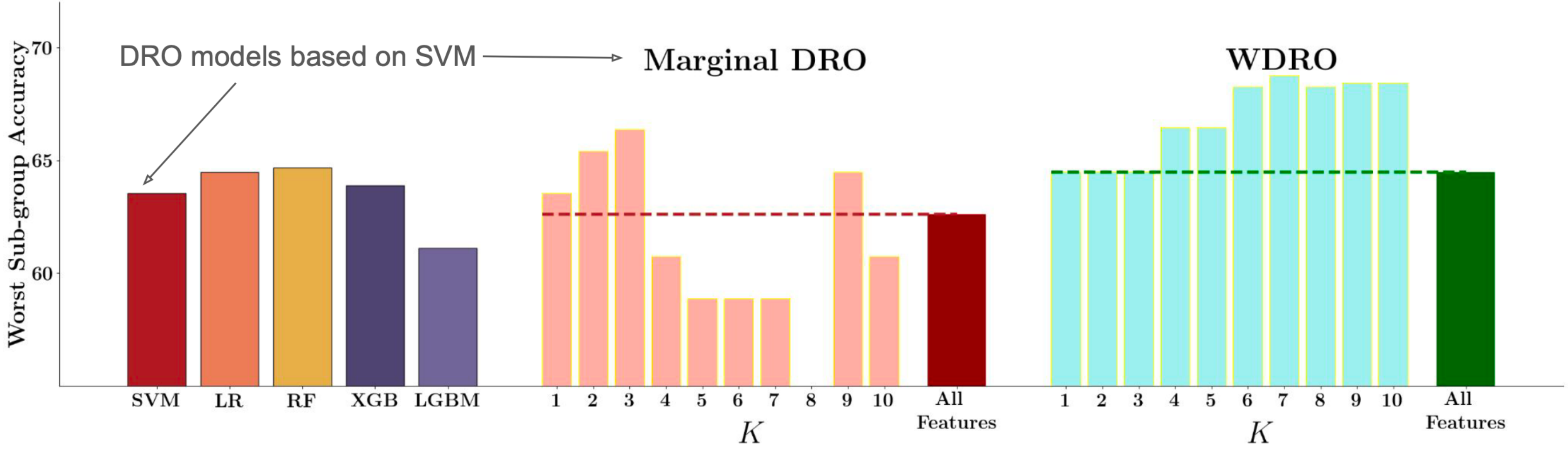
Liu, J., Wu, J., Wang, T., Zou, H., Li, B., & Cui, P. Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications. ICML, 2024.

Perspective 4: DRO tailored for specific shifts

another specific kind of data heterogeneity here

Perspective 4: DRO tailored for specific shifts

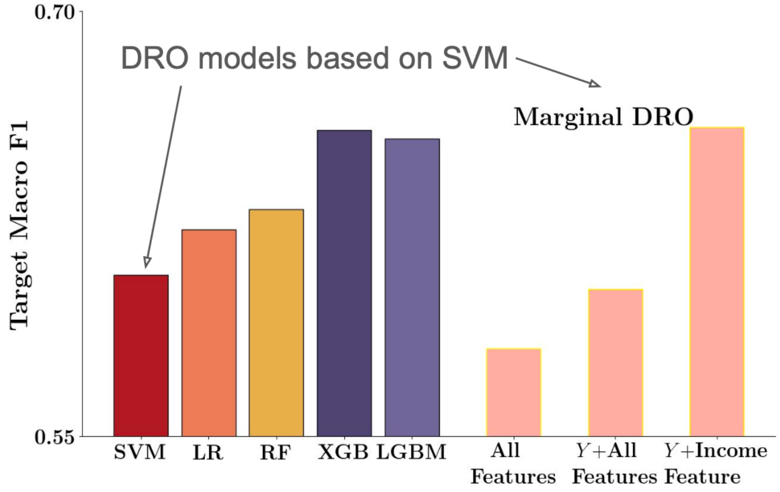
- Consider shifts induced by age groups: [20,25), [25,30), …, [75,100)
- Consider DRO methods (DH**N**'22) tailored to shifts on a subset of covariates
- Variable selection for ambiguity set: top-K with largest subgroup differences
- Performance varies a lot over variables selected



Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

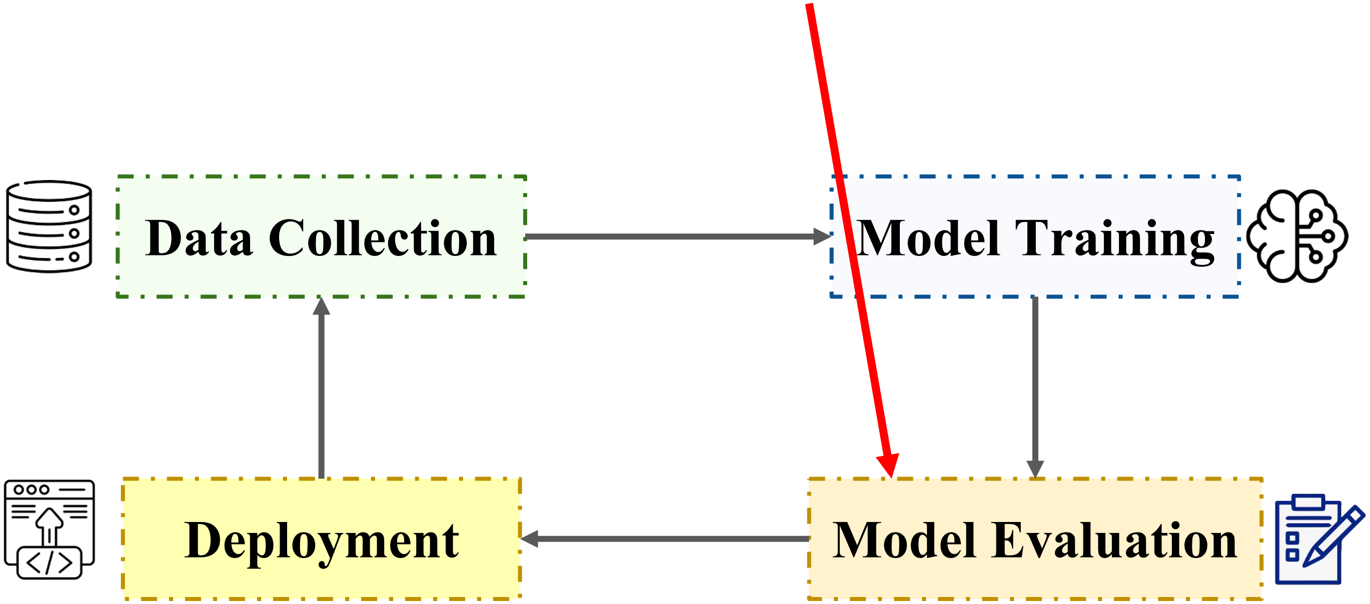
Perspective 4: DRO tailored for specific shifts

- Consider *Y*|*X*-shifts from NE -> LA (public coverage task)
- Consider DRO methods that consider shifts on a subset of covariates and *Y*
- Variable selection for ambiguity set: Y | “income” suffers the largest shift
- Performance varies a lot over variables selected



Liu, J., Wang, T., Cui, P., & Namkoong, H. On the Need of a Modeling Language for Distribution Shifts: Illustrations on Tabular Datasets.

Stage 3: Analyze heterogeneity **in evaluation**



Example 1: Error slice discovery

Example 2: Stability Evaluation

Perspective 5: it's important to understand where a model performs poorly

After training a model, we **need** to know

On what training data does the model perform **POORLY**?

If we understand this, we can

- do efficient data re-collection
- do model patching/re-training
- not use the model on certain regions

Example: Slice discovery in training distribution

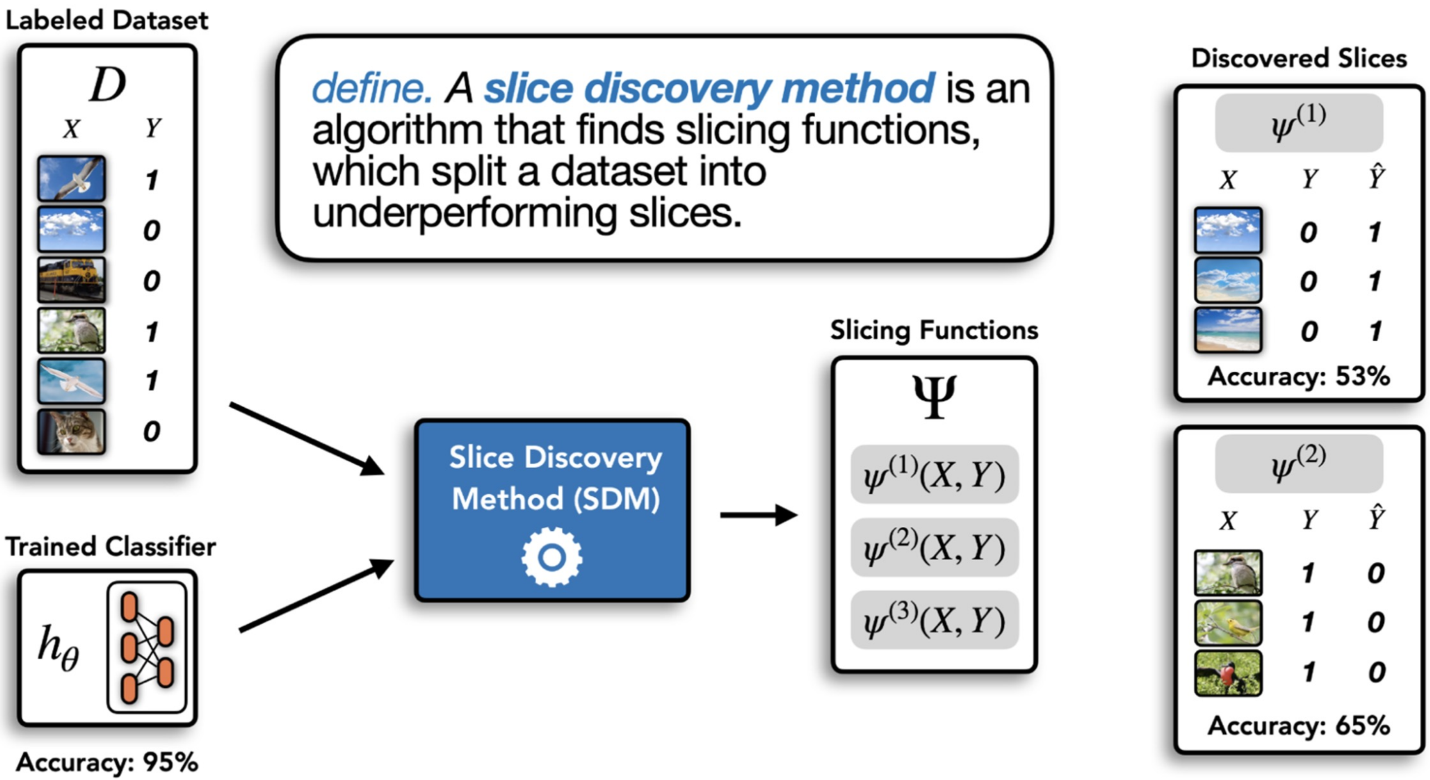


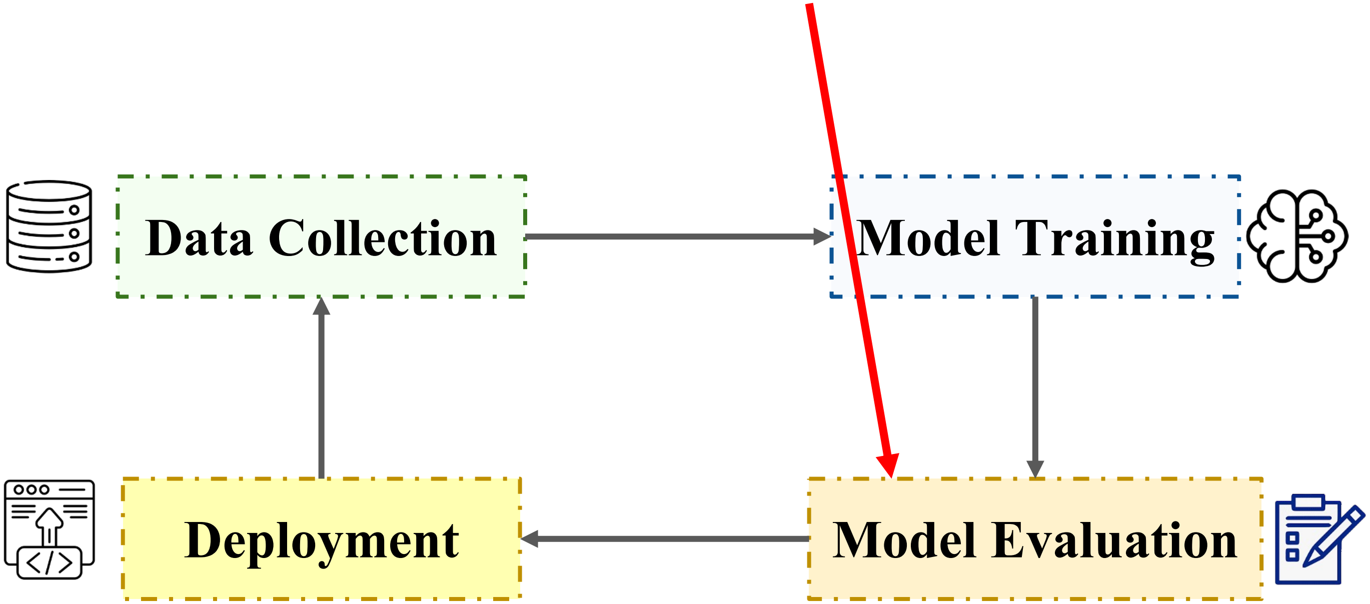
Figure from Eyuboglu, S.,et al. http://ai.stanford.edu/blog/domino/

Example: slice discovery in training distribution

More literature on **cross-modal diagnosis**

Eyuboglu, S.,et al. Domino: Discovering Systematic Errors with Cross-Modal Embeddings. In ICLR Gao, I., et al. Adaptive testing of computer vision models. In ICCV.
Metzen, J. H., et al. Identification of Systematic Errors of Image Classifiers on Rare Subgroups.
Jain, S., et al. Distilling model failures as directions in latent space.
Wiles, O., et al. Discovering Bugs in Vision Models using Off-the-shelf Image Generation and Captioning. In NeurIPS ML Safety Workshop.
Mozannar, H., et al. Effective Human-AI Teams via Learned Natural Language Rules and Onboarding. In NeurIPS

Stage 3: Analyze heterogeneity **in evaluation**



Example 1: Error slice discovery

Example 2: Stability Evaluation

Perspective 6: beyond accuracy, evaluate stability

What kind of data distribution is the model most sensitive to?

Two ways of generating distribution shifts:

- ***Data corruptions***: changes in the distribution support (i.e., observed data samples).
- ***Sub-population shifts***: perturbation on the probability density or mass function while keeping the same support.

Preliminary

Definition (OT discrepancy with moment constraints)

If $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\mathcal{W} \subseteq \mathbb{R}_+$ are convex and closed sets, $c : (\mathcal{Z} \times \mathcal{W})^2 \to \mathbb{R}_+$ is a lower semicontinuous function, and $\mathbb{Q}, \mathbb{P} \in \mathcal{P}(\mathcal{Z} \times \mathcal{W})$, then the OT discrepancy with moment constraints induced by c, $\mathbb{Q}$ and $\mathbb{P}$ is the function $\mathbb{M}_c : \mathcal{P}(\mathcal{Z} \times \mathcal{W})^2 \to \mathbb{R}_+$ defined through

$$\mathbb{M}_c(\mathbb{Q}, \mathbb{P}) = \begin{cases} \inf & \mathbb{E}_\pi[c((Z, W), (\hat{Z}, \hat{W}))] \\ \text{s.t.} & \pi \in \mathcal{P}((\mathcal{Z} \times \mathcal{W})^2) \\ & \pi_{(Z,W)} = \mathbb{Q}, \ \pi_{(\hat{Z},\hat{W})} = \mathbb{P} \\ & \mathbb{E}_\pi[W] = 1 \quad \pi\text{-a.s,} \end{cases}$$

where $\pi_{(Z,W)}$ and $\pi_{(\hat{Z},\hat{W})}$ are the marginal distributions of (Z, W) and $(\hat{Z}, \hat{W})$ under π.

We choose the cost function as:

$$c((z, w), (\hat{z}, \hat{w})) = \underbrace{\theta_1 \cdot w \cdot (\|x - \hat{x}\|_2^2 + \infty \cdot |y - \hat{y}|)}_{\text{differences between samples}} + \underbrace{\theta_2 \cdot (\phi(w) - \phi(\hat{w}))_+}_{\text{differences in probability mass}}.$$

Blanchet, J., Kuhn, D., Li, J., & Taskesen, B. (2023). Unifying distributionally robust optimization via optimal transport theory
Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Perspective 6: beyond accuracy, evaluate stability

Given a learning model f_β and the distribution $\mathbb{P}_0 \in \mathcal{P}(\mathcal{Z})$, we formally introduce the **OT-based stability evaluation criterion** as

$$\mathfrak{R}(\beta, r) = \begin{cases} \inf\limits_{\mathbb{Q} \in \mathcal{P}(\mathcal{Z} \times \mathcal{W})} & \mathbb{M}_c(\mathbb{Q}, \hat{\mathbb{P}}) \\ \text{s.t.} & \mathbb{E}_{\mathbb{Q}}[W \cdot \ell(\beta, Z)] \geq r. \end{cases} \qquad \text{(P)}$$

Some notations:

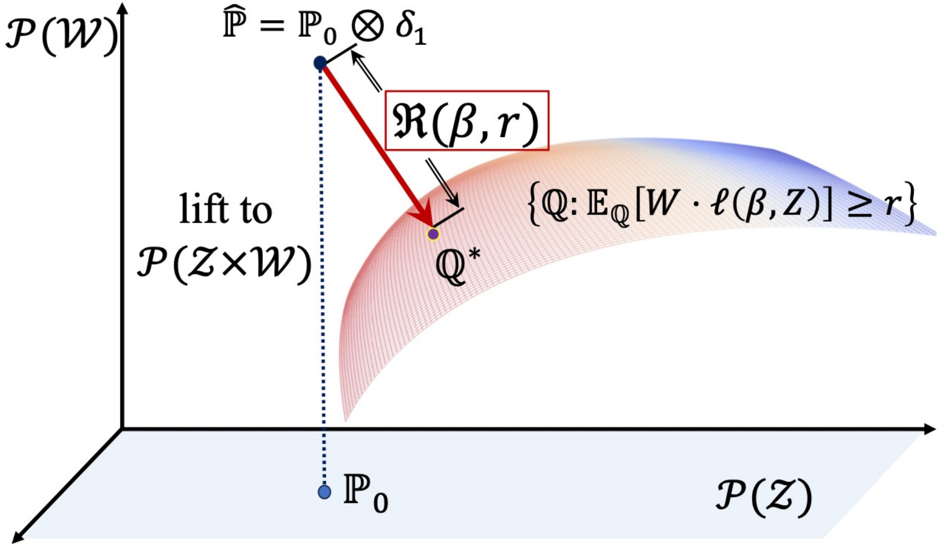
- $\hat{\mathbb{P}}$: The reference measure selected as $\mathbb{P}_0 \otimes \delta_1$, with δ_1 denoting the Dirac delta function.
- $\ell(\beta, z)$: The prediction risk of model f_β on sample z.
- $r > 0$: the *pre-defined risk threshold*.

Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Larger $\mathfrak{R}(\beta, r) \Rightarrow$ More Stable

Perspective 6: beyond accuracy, evaluate stability

Projection distance to the distribution set where the model performance falls below a specific threshold



Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Perspective 6: beyond accuracy, evaluate stability

Theorem (Dual reformulations)

Suppose that $\mathcal{W} = \mathbb{R}_+$. *(i) If* $\phi(t) = t\log t - t + 1$, *then the dual problem* (D) *admits:*

$$\sup_{h \geq 0} hr - \theta_2 \log \mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right)\right]; \quad (1)$$

(ii) If $\phi(t) = (t-1)^2$, *then the dual problem* (D) *admits:*

$$\sup_{h \geq 0, \alpha \in \mathbb{R}} hr + \alpha + \theta_2 - \theta_2 \mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\ell_{h,\theta_1}(\hat{Z}) + \alpha}{2\theta_2} + 1\right)_+^2\right], \quad (2)$$

where the d*-transform of* $h \cdot \ell(\beta, \cdot)$ *with the step size* θ_1 *is defined as*

$$\ell_{h,\theta_1}(\hat{z}) := \max_{z \in \mathcal{Z}} h \cdot \ell(\beta, z) - \theta_1 \cdot d(z, \hat{z}).$$

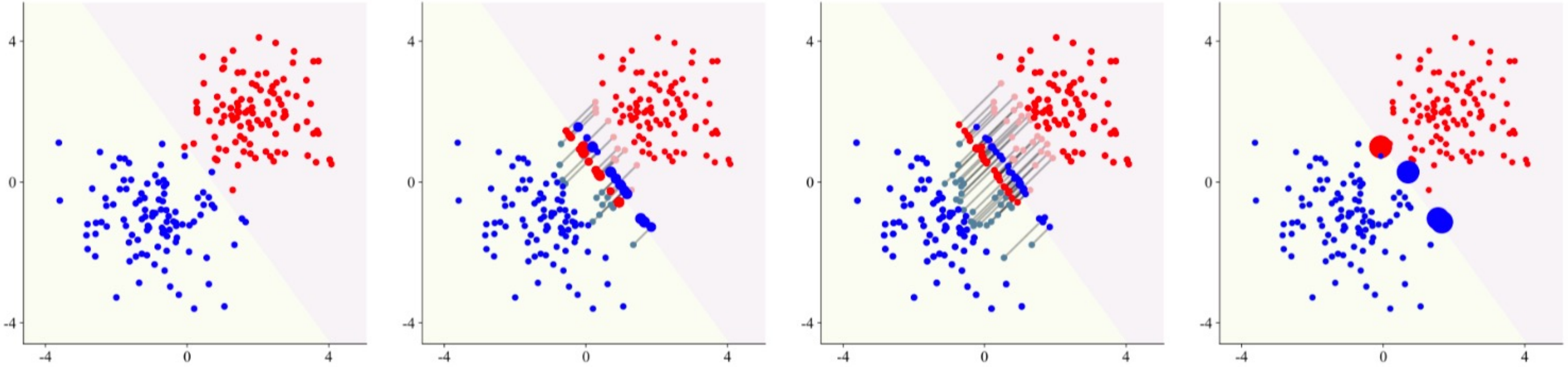
sample reweighting

data corruption

Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Visualization on toy examples

Visualize the most sensitive distribution $\mathbb{Q}^\star$:



(a) Original Dataset (b) $\theta_1 = 1.0, \theta_2 = 0.25$ (c) $\theta_1 = 0.2, \theta_2 = +\infty$ (d) $\theta_1 = +\infty, \theta_2 = 0.2$

Figure 2: Visualizations on toy examples with *0/1* loss function under different θ_1, θ_2. The original prediction error rate is 1%, and the error rate threshold r is set to 30%. The size of each point is proportional to its sample weight in $\mathbb{Q}^\star$

Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Model stability analysis

Task: Predict individual's income based on personal features.

Method under evaluation:

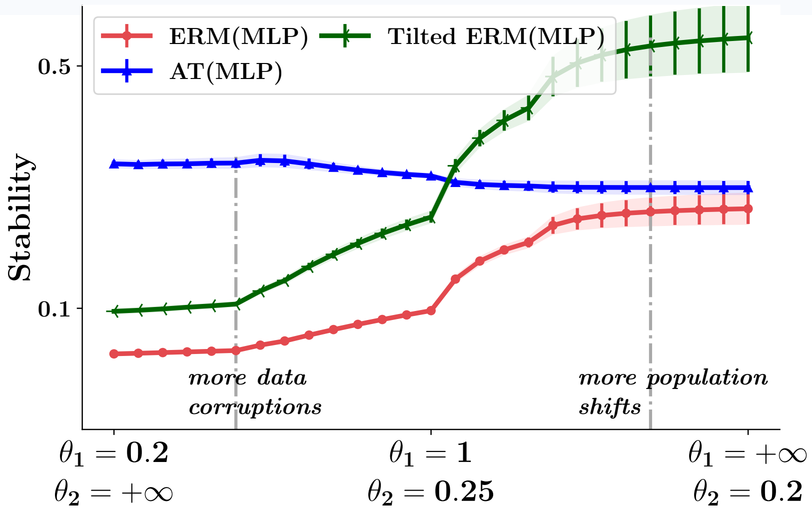
- Empirical Risk Minimization (ERM)
- Adversarial Training (AT): designed for robustness to data corruptions
- Tilted ERM: designed for robustness to sub-population shifts

Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Model stability analysis

A method designed for one class of data perturbation may not be robust against another:

- AT is not stable under sub-population shifts.
- Tilted ERM is not stable under data corruptions.



Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Feature stability analysis

Feature Stability

- perturbing on which feature will cause model's performance drop
- providing more fine-grained diagnosis for a prediction model

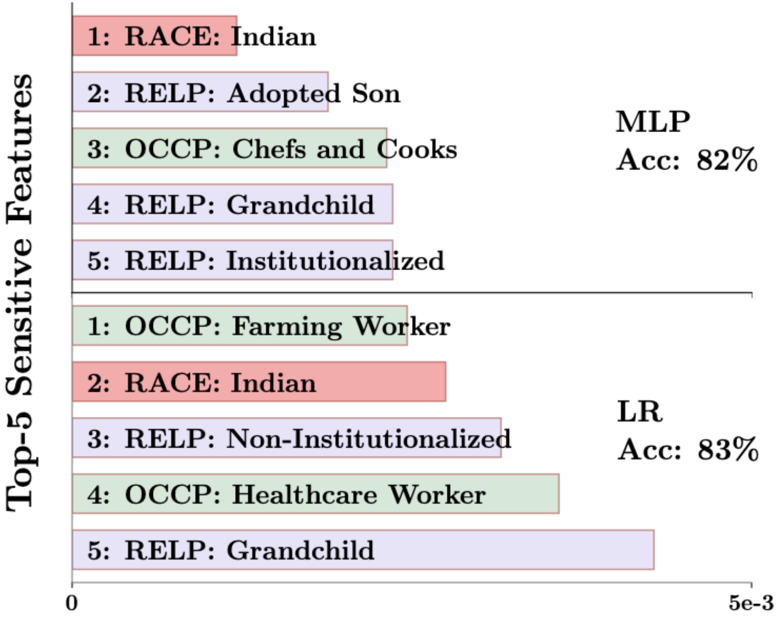
For i-th feature, choose the cost function as:

$$c((z,w),(\hat{z},\hat{w})) = \theta_1 \cdot w \cdot \underbrace{(\|z_{(i)} - \hat{z}_{(i)}\|_2^2 + \infty \cdot \|z_{(,-i)} - \hat{z}_{(,-i)}\|_2^2)}_{\text{only allow perturbations on } i\text{-th feature}} + \theta_2 \cdot (\phi(w) - \phi(\hat{w}))_+ .$$

Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Feature stability analysis

Task: predict individual's income based on personal features
Dataset: ACS Income

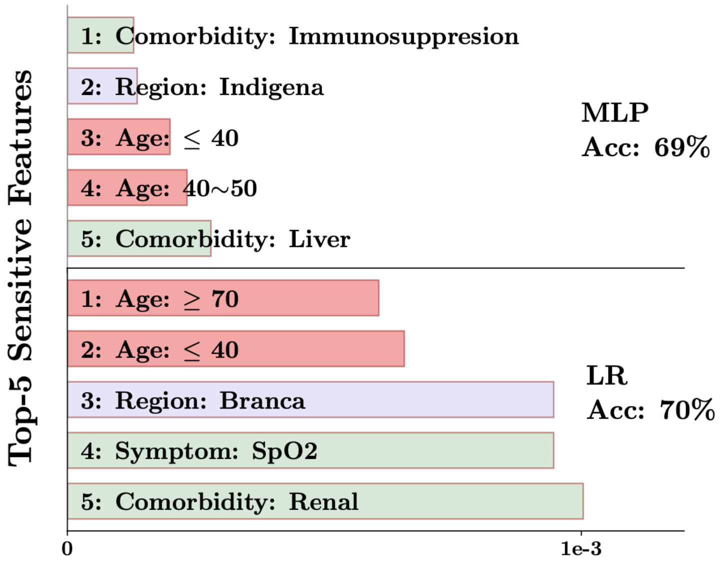


Focus too much on "American Indian" feature

Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Feature stability analysis

Task: predict mortality caused by COVID-19



"Age" matters a lot

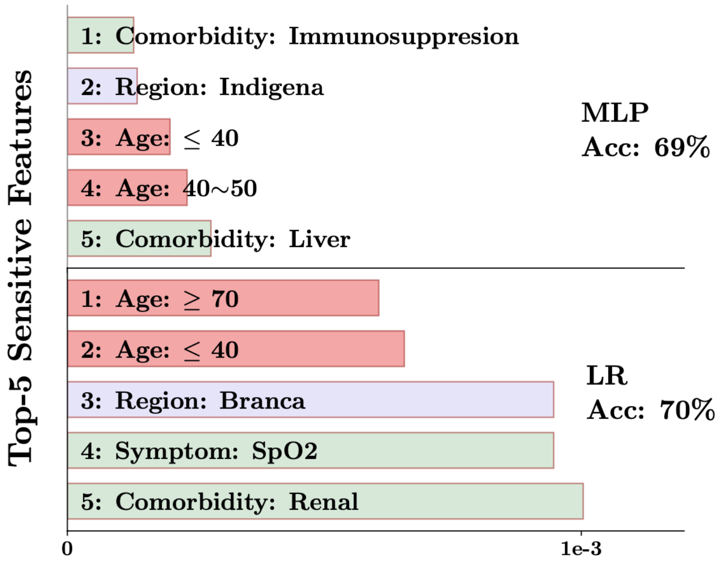
Does the model perform well **over all age groups**?

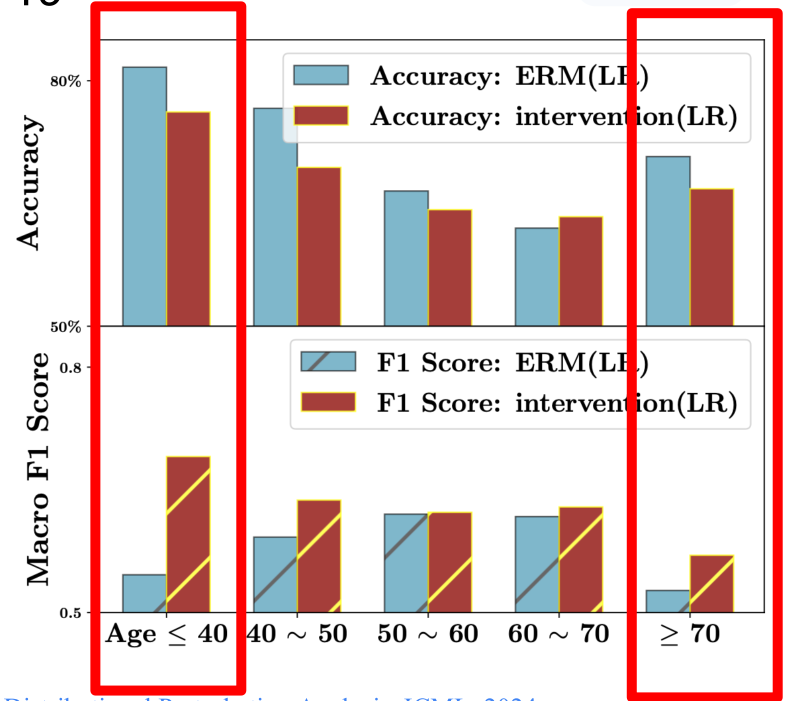
Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Feature stability analysis

NO!

Task: predict mortality caused by COVID-19





"Age" matters a lot

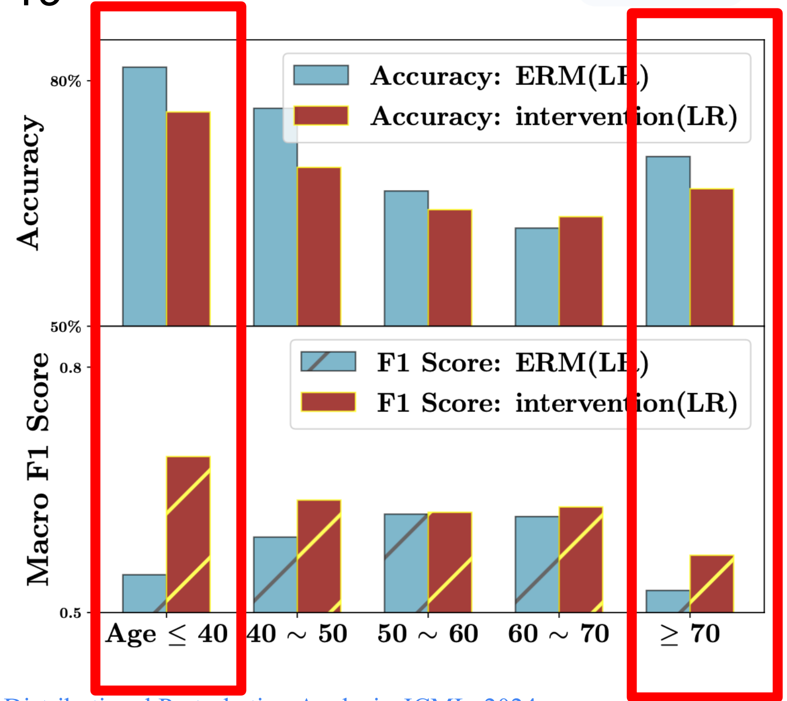
Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Feature stability analysis

Task: predict mortality caused by COVID-19

- For Age<40 and Age>70, the accuracy is high, but **Macro-F1 score is too low**
- It simply predicts based on Age!

NO!



Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

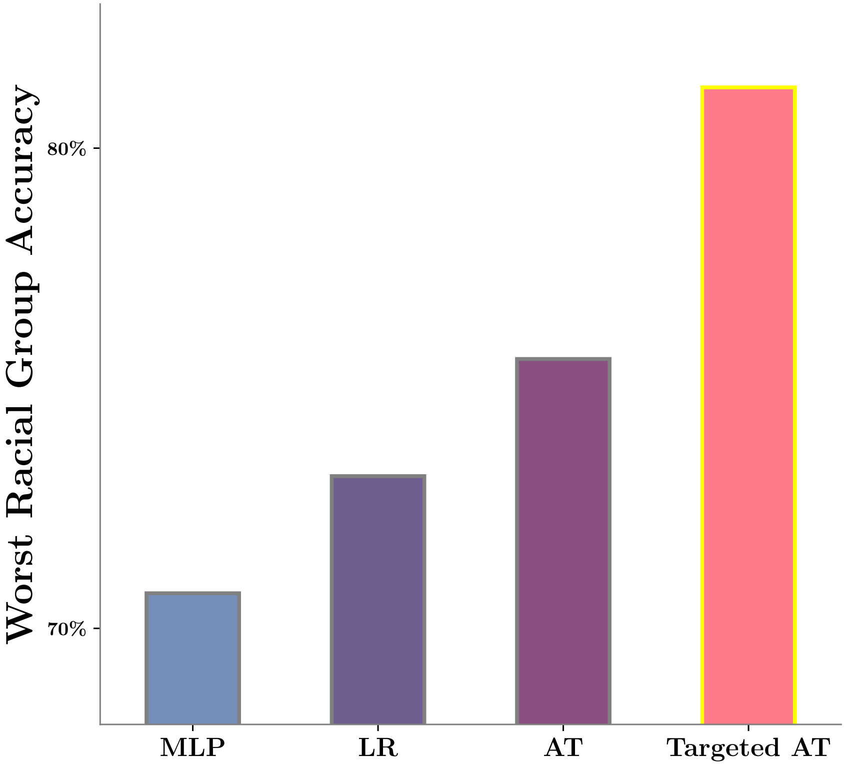
Targeted algorithmic intervention

Insight: Feature stability can motivate refined algorithmic intervention.

- for AT, only perturb the identified sensitive racial feature "American Indian"
- significantly increase the worst racial group accuracy

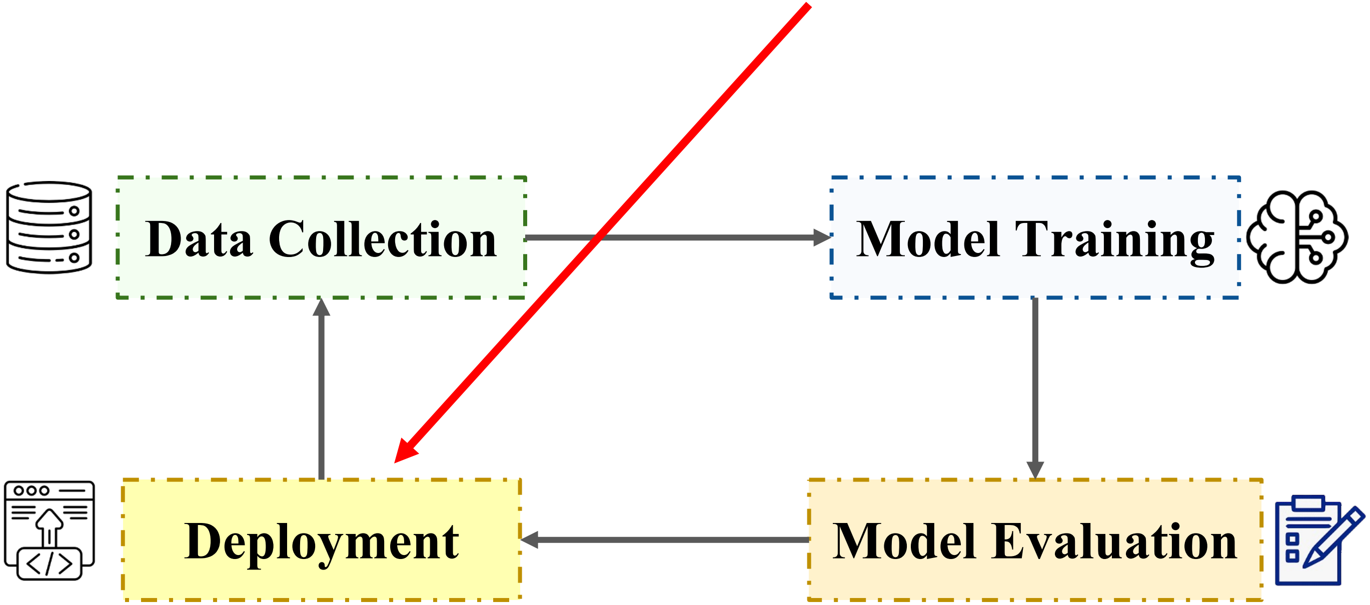
Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Targeted algorithmic intervention



Blanchet, J., Cui, P., Li, J., & Liu, J. Stability Evaluation via Distributional Perturbation Analysis. ICML, 2024.

Stage 4: Analyze heterogeneity **after deployment**



Example 1: $Y|X$-shifts vs. X-shifts

Example 2: Covariate region analysis

Perspective 7: it's important to understand **why** your model performs poorly *across a distribution shift*

Train $\rightarrow$ Target e.g. deployment

$P \longrightarrow Q$

Different interventions for different shifts!

1. Algorithm #1: domain adaptation
2. Algorithm #2: DRO
3. Algorithm #3: invariant learning
4. ….
5. Collect more data from target
6. Collect more features

These make modeling assumptions. Do they apply?

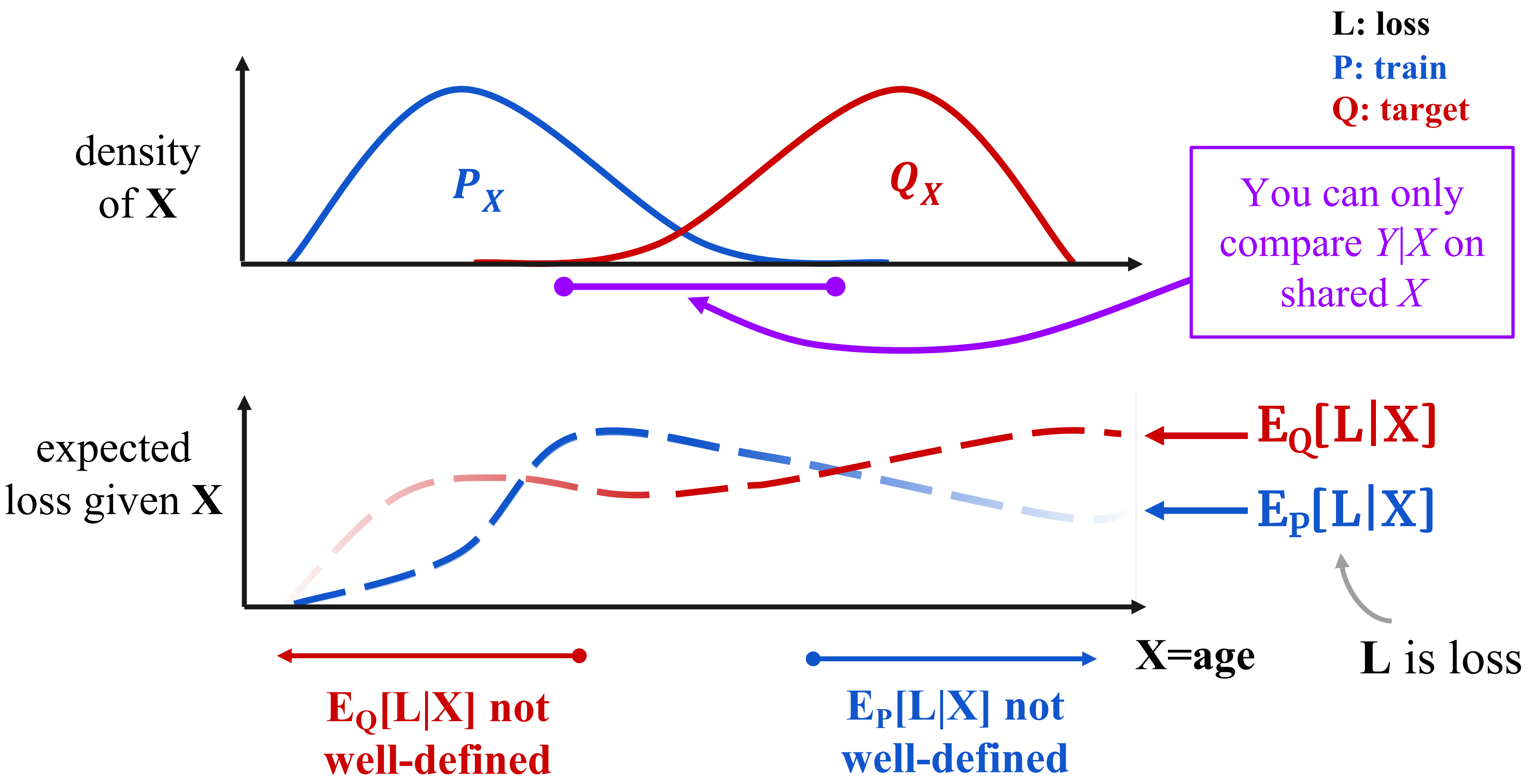
Understand distribution shift to determine next steps!

Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

Attribute change in performance to distribution shifts

***X*-shifts**	***Y*|*X*-shifts**
changes in sampling, population shifts, minority groups	changes in labeling or mechanism, poorly chosen X

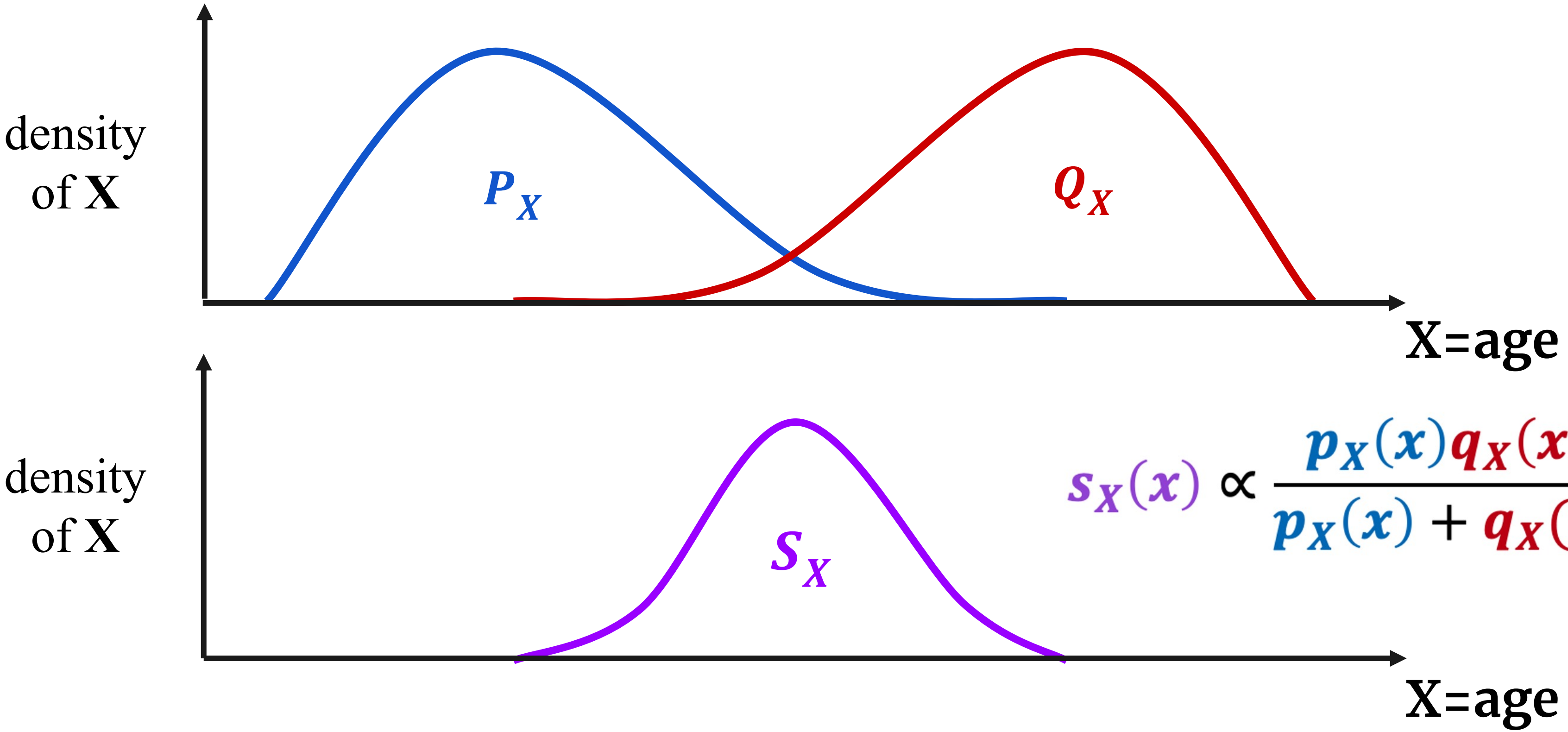
- Real distribution shifts involve a combination of both shifts
- ***Attribute*** change in model performance to shifts: not all shifts matter



Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

L: loss
P: train
Q: target
S: shared

Define Shared Distribution

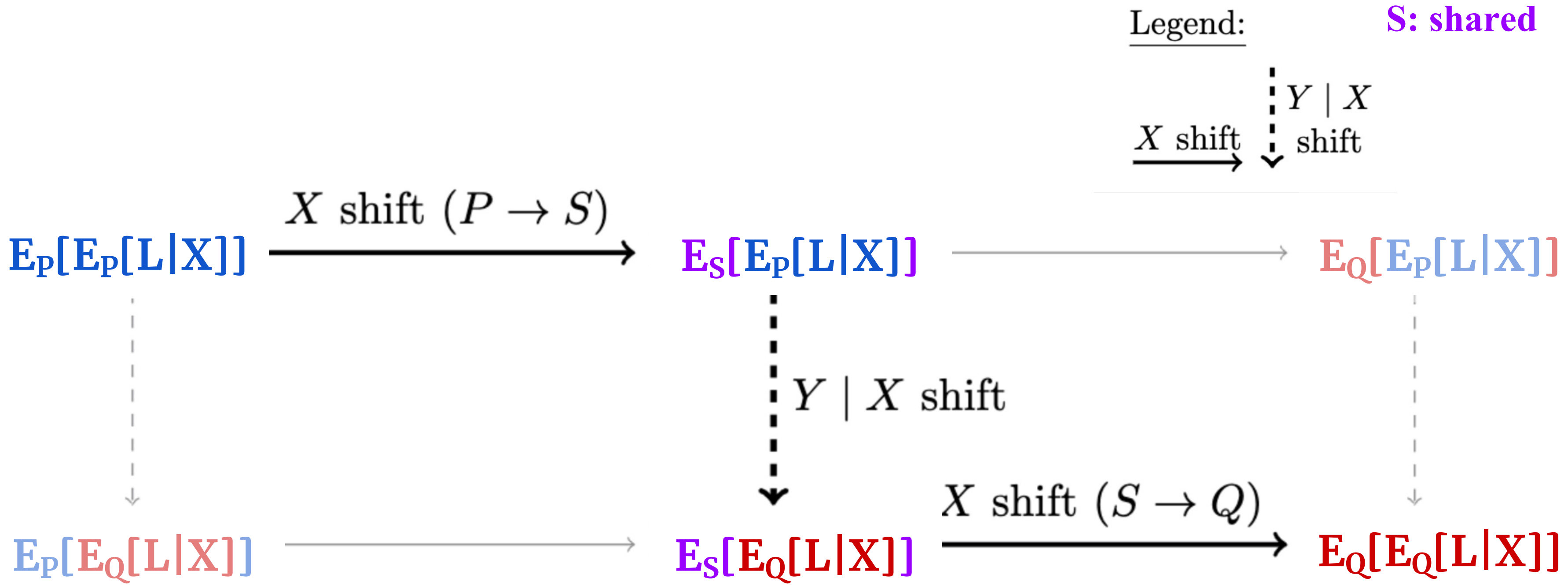


$$s_X(x) \propto \frac{p_X(x)q_X(x)}{p_X(x) + q_X(x)}$$

Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

Decompose change in performance

L: loss
P: train
Q: target
S: shared



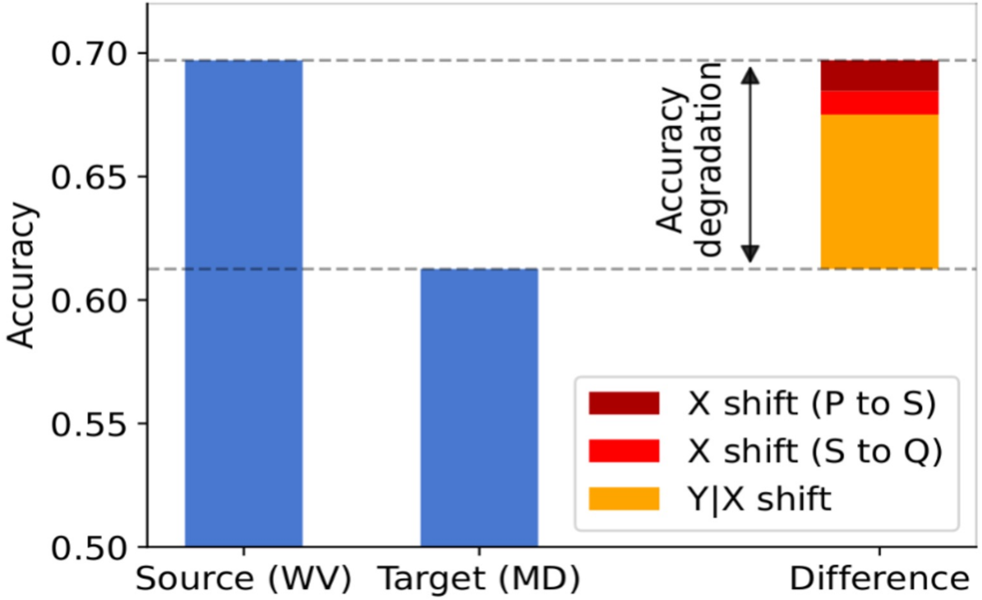
shared *X* distribution

Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

Employment prediction case study

L: loss
P: train
Q: target
S: shared

[$Y|X$ shift] P: West Virginia, Q: Maryland



WV model does not use education.

$Y|X$ shift because of missing covariate: education affects employment

Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

For reference: other diagnostic tools

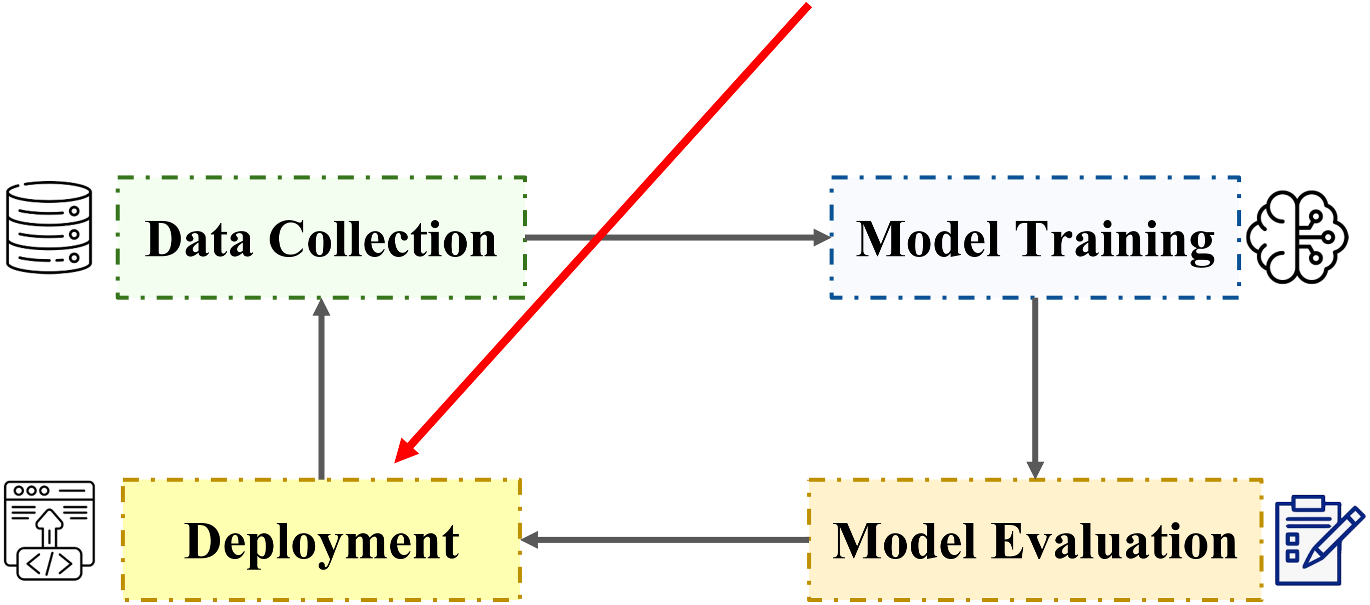
Haoran Zhang, Harvineet Singh, Marzyeh Ghassemi, Shalmali Joshi. "Why did the Model Fail?": Attributing Model Performance Changes to Distribution Shifts (2022)

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, Peng Cui. NICO++: Towards Better Benchmarking for Domain Generalization (2022)

Adarsh Subbaswamy, Roy Adams, Suchi Saria. Evaluating Model Robustness and Stability to Dataset Shift (2021)

Finale Doshi-Velez, Been Kim. Towards A Rigorous Science of Interpretable Machine Learning (2017)

Stage 4: Analyze heterogeneity **after deployment**



Example 1: $Y|X$-shifts vs. X-shifts

Example 2: Covariate region analysis

Perspective 8: it's important to understand where you have $Y|X$ shifts

When model performance drops after deployment, we **need** to know

Where does the model performance drop because of $Y|X$ shift?

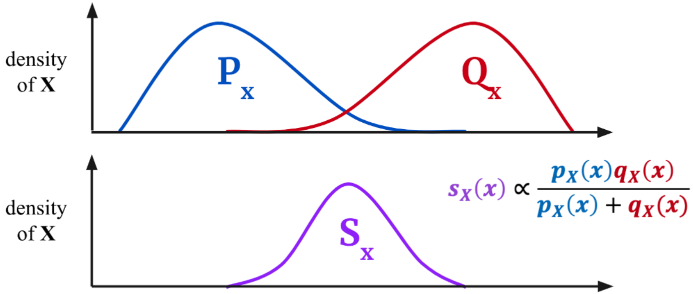
If we understand this, then we can collect data better.

Identify covariate regions with $Y|X$-shifts

How to Better Understand $Y|X$-Shifts?

Find Covariate Regions with Strong $Y|X$-Shifts!

1. Construct shared distribution from training and target
2. Model Y separately on each of training and target: f_p, f_q
3. Model difference in Y between train and target $|f_p(x) - f_q(x)|$ on shared distribution using interpretable tree-based model

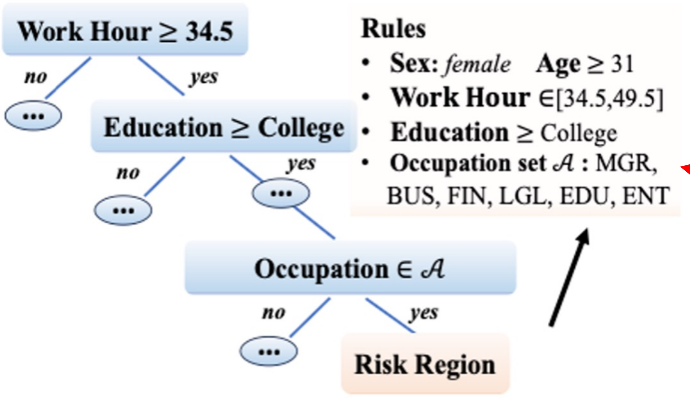


Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Identify covariate regions with $Y|X$-shifts

Task: Income Prediction
Shift: CA -> PR

Tabular Data



(c) Region with $Y|X$-shifts (XGBoost)

Y|X shift region consists of occupations that require language

Official languages are ***different*** in CA and PR!

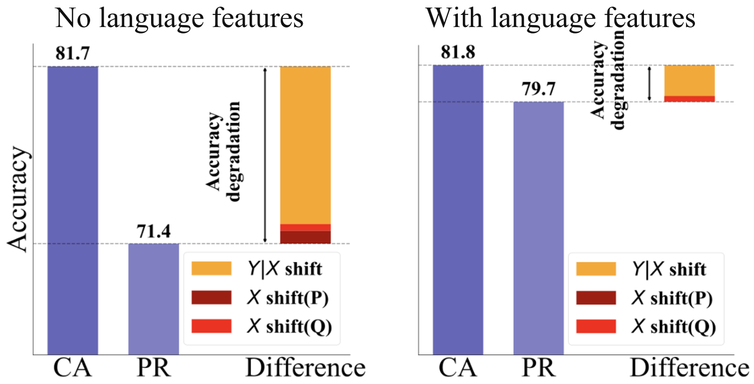
Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

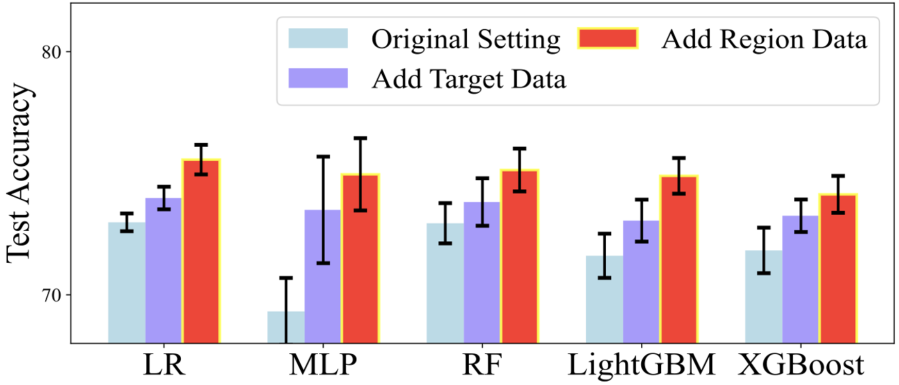
Tool 4: Identify Regions with $Y|X$-Shifts

Good data may be **more effective**!

Task: Income Prediction
Shift: CA -> PR

Include language features when training on CA → better performance in PR





collecting better features

collecting better target data

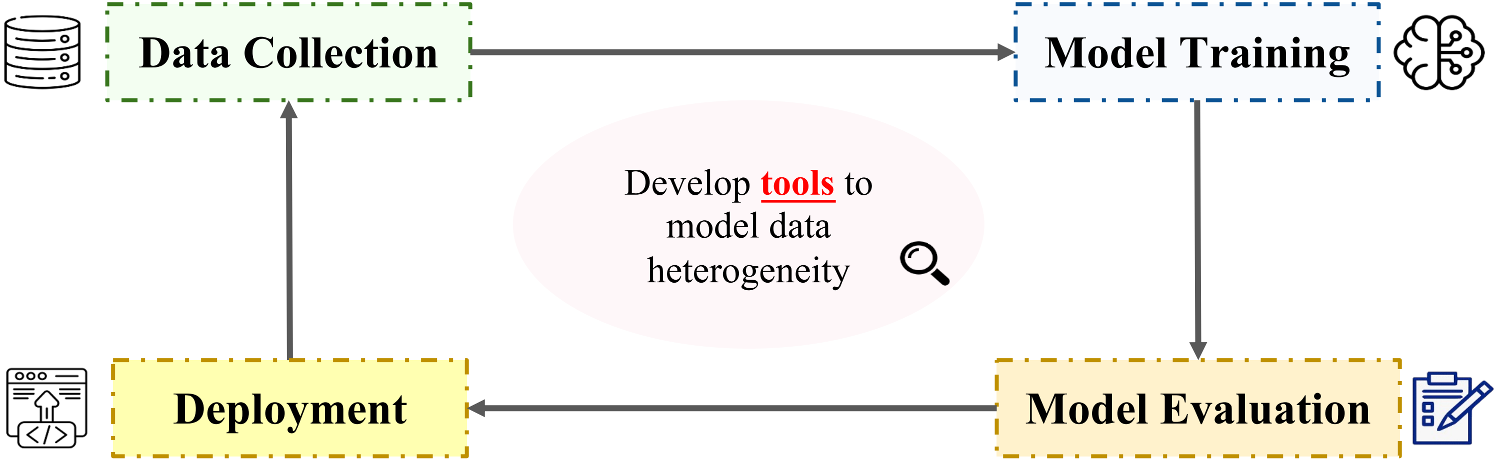
Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Recap

- Heterogeneity is really important!
- Two existing approaches to domain generalization
  - Make modeling assumptions: principled, but do the assumptions hold?
  - Scaling up data: effective for internet-scale data, but for many problems data is costly
- Heterogeneity-aware approach:
  - Develop and use tools to understand heterogeneity in your setting.
  - Then, use this understanding throughout the entire modeling process.

Future directions

- We need a system-level view; “industrial engineering” for AI
  - Design better workflows

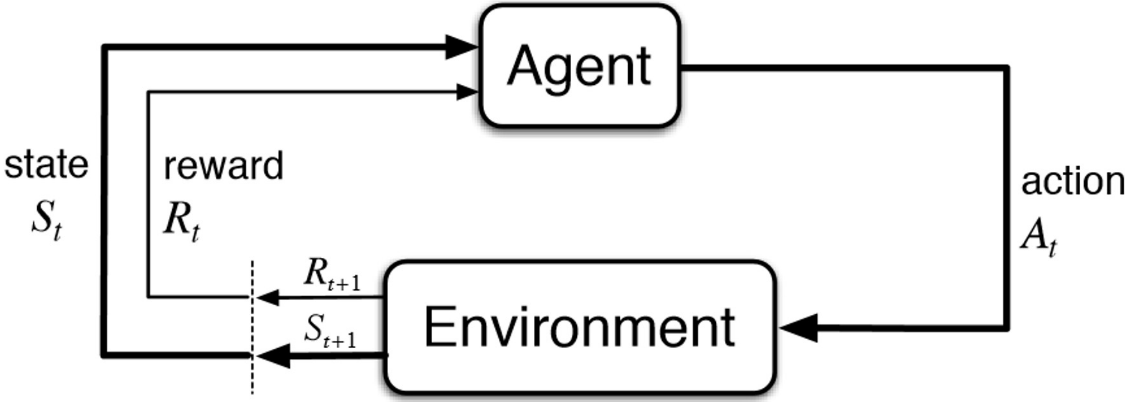


Future directions

- We must build models that know what it doesn’t know
- Recognize unforeseen heterogeneity at test time
- Connections to uncertainty quantification
    - Bayesian ML, conformal prediction etc
    - Requires explicitly modeling unobserved factors

Future directions

- Based on this uncertainty, agents must decide how to actively collect data to reduce this uncertainty
- Connections to reinforcement learning and active learning



Future directions

- We need a system-level view; “industrial engineering” for AI
  - Design better workflows
- We must build models that know what it doesn’t know
  - We only collect outcomes on actions (observations) we take (measure)
- Based on this uncertainty, agents must decide how to actively collect data to reduce this uncertainty
- Overall, exciting research space with many open problems!