Modeling and Exploiting Data Heterogeneity under Distribution Shifts

Tutorial at 37th Conference on Neural Information Processing Systems (NeurIPS 2023)

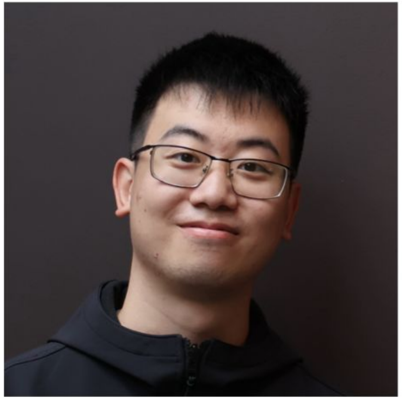
Jiashuo Liu[1], Tiffany (Tianhui) Cai[2], Peng Cui[1], Hongseok Namkoong[3]

[1] Department of Computer Science and Technology, Tsinghua University
[2] Department of Statistics, Columbia University
3 Decision, Risk, and Operations Division, Columbia Business School

liujiashuo77@gmail.com, tiffany.cai@columbia.edu
cuip@tsinghua.edu.cn, namkoong@gsb.columbia.edu

Dec 11th, 2023, New Orleans

Speakers

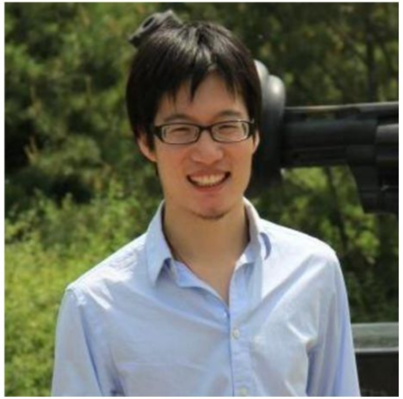


Jiashuo Liu

Tsinghua University

Tiffany (Tianhui) Cai

Columbia University



Peng Cui

Tsinghua University

Hongseok Namkoong

Columbia University

Thoughtful use of AI is challenging

AI's main value proposition: omni-present feedback generation through codification of patterns

- Recent advances are truly exciting, e.g., natural language interface to computing through LLMs
- Salient challenges remain for their reliable deployment and use
- Main value prop is also its main shortcoming: difficult to assess when said automated predictions and feedback are trustworthy

Some failures are not hard to spot

Some failures are not hard to spot

- 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.

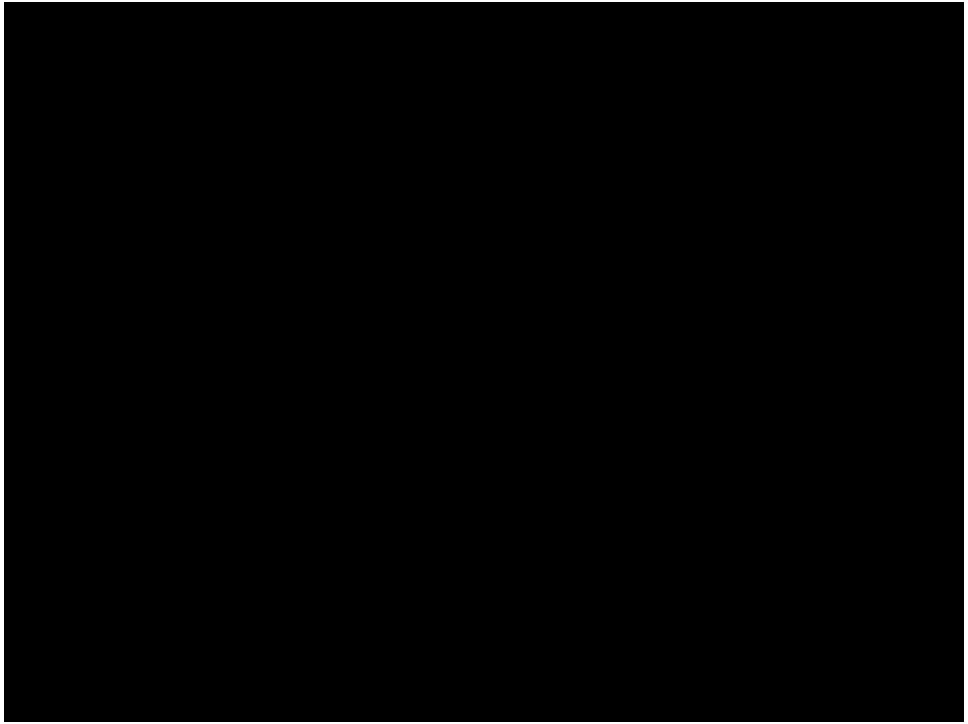
By Will Douglas Heaven

July 30, 2021

Some failures are not hard to spot

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

Some are not so easy



Federal Government Opens Safety Defect Investigation Into Tesla Autopilot Crashes

NHTSA is looking at whether the technology may be a contributing factor in multiple crashes with emergency vehicles

By Keith Barry

Published August 16, 2021 | Updated September 1, 2021

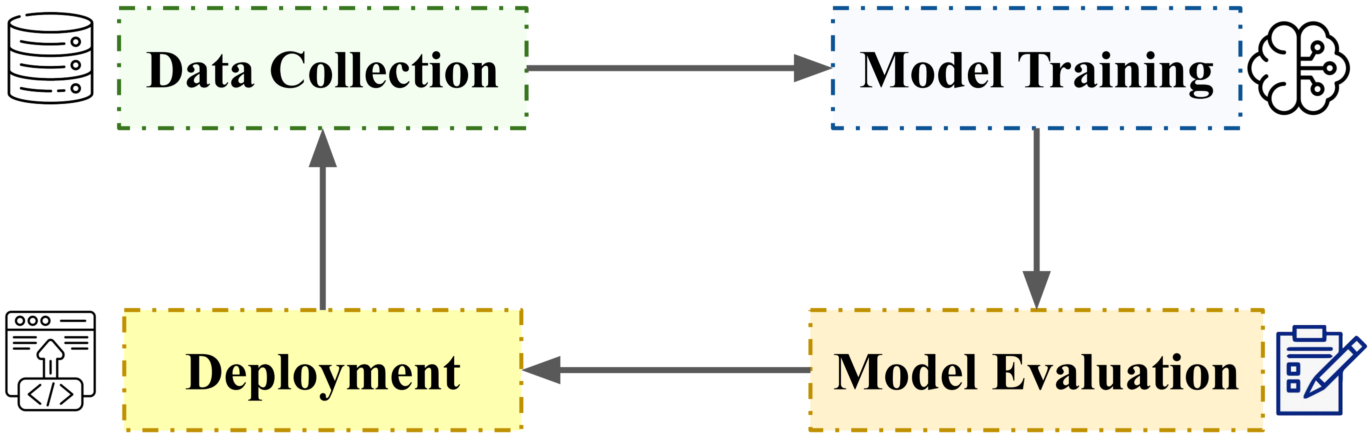
Some are not so easy

AI Camera Ruins Soccer Game For Fans After Mistaking Referee's Bald Head For Ball

Some are not so easy

System level of view of AI

- Building a reliable AI stack requires a holistic view



- Since rigorous benchmarking is the foundation of empirical progress, we begin with how we can evaluate the robustness of AI models

Outline

Part 1: Benchmarking performance under distribution shift

Part 2: A critical review of existing approaches

Part 3: Application-specific modeling of data heterogeneity

Part 4: Towards heterogeneity-<u>aware</u> machine learning

History

- Lots of research on distribution shifts and robustness in causal inference, operations research, economics, control theory, and statistics
- ML researchers like Masashi Sugiyama and Kate Saenko studied particular types of distribution shift in ‘00s, and a wave of algorithmic papers followed in ‘10s
- Most recently, exciting developments in benchmarking model robustness
    - Rigorous benchmarking is the foundation of empirical progress

ImageNet

Large **image classification** dataset: 1.2 mio training images, 1,000 image classes.

Golden retriever

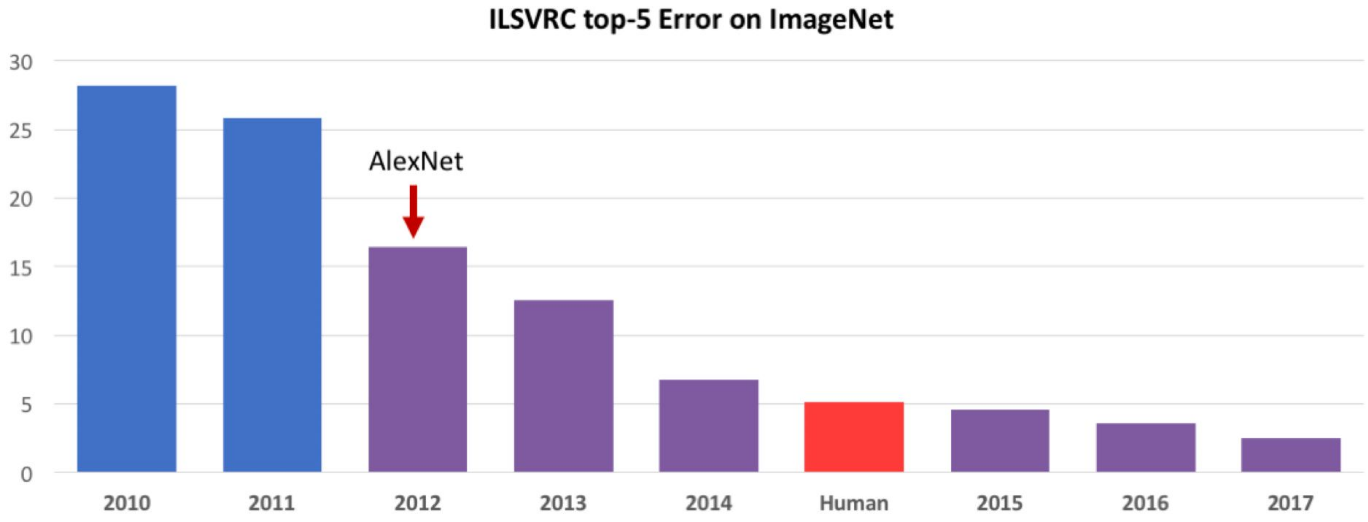
Great white shark

Minibus

Slide credit: Ludwig Schmidt

ImageNet

- Drove the bulk of empirical progress in AI for multiple years from 2010



Slide credit: Ludwig Schmidt

Robustness on ImageNet

Lots of progress on ImageNet over the past 10 years, but models are still not robust.

Evaluation: **new test sets**

ImageNetV2

[Recht, Roelofs, Schmidt, Shankar ’19]

ObjectNet

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz ’19]

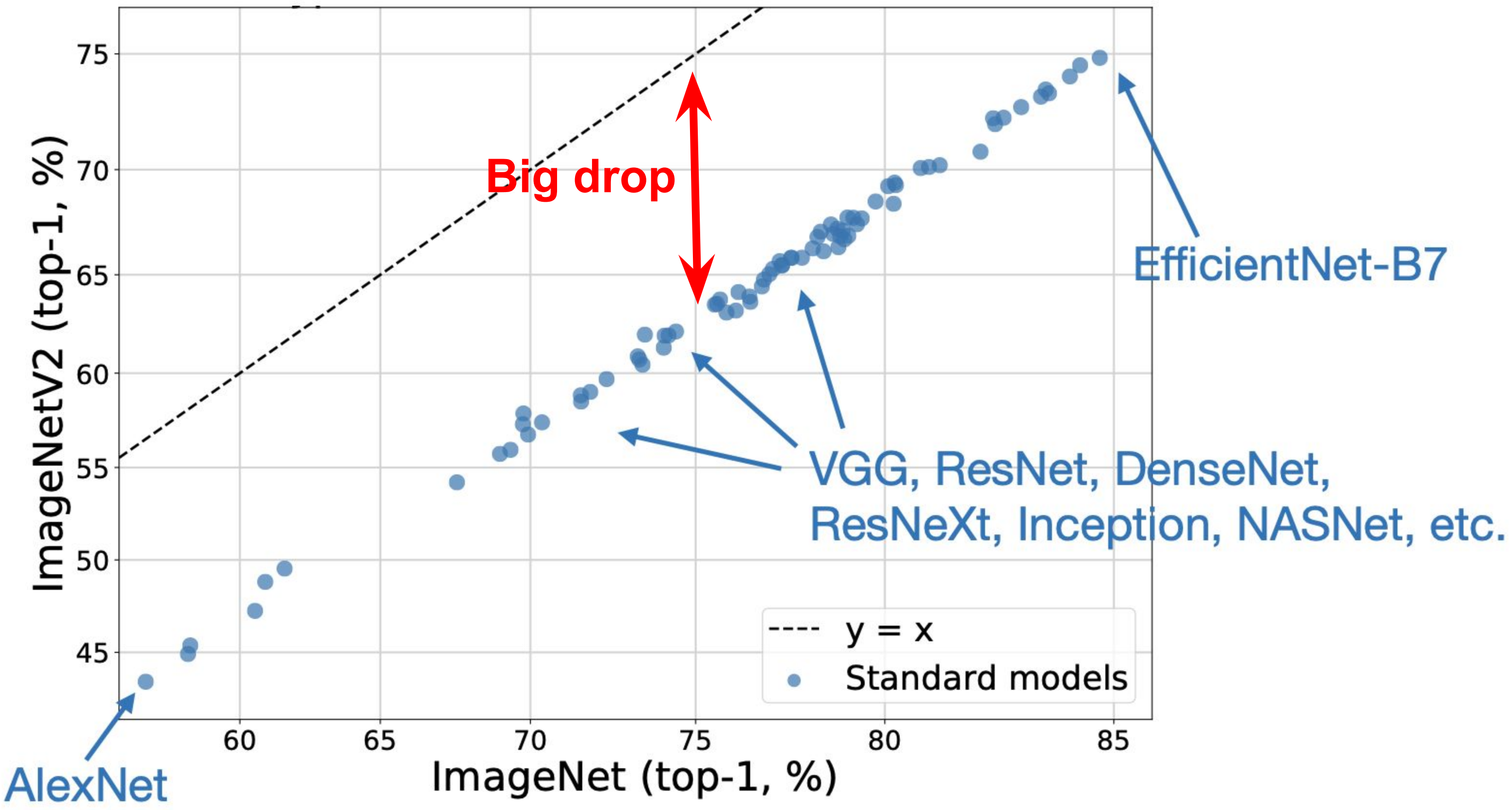
ImageNet-Sketch

[Wang, Ge, Lipton, Xing ’19]

ImageNet-R

[Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer ’20]

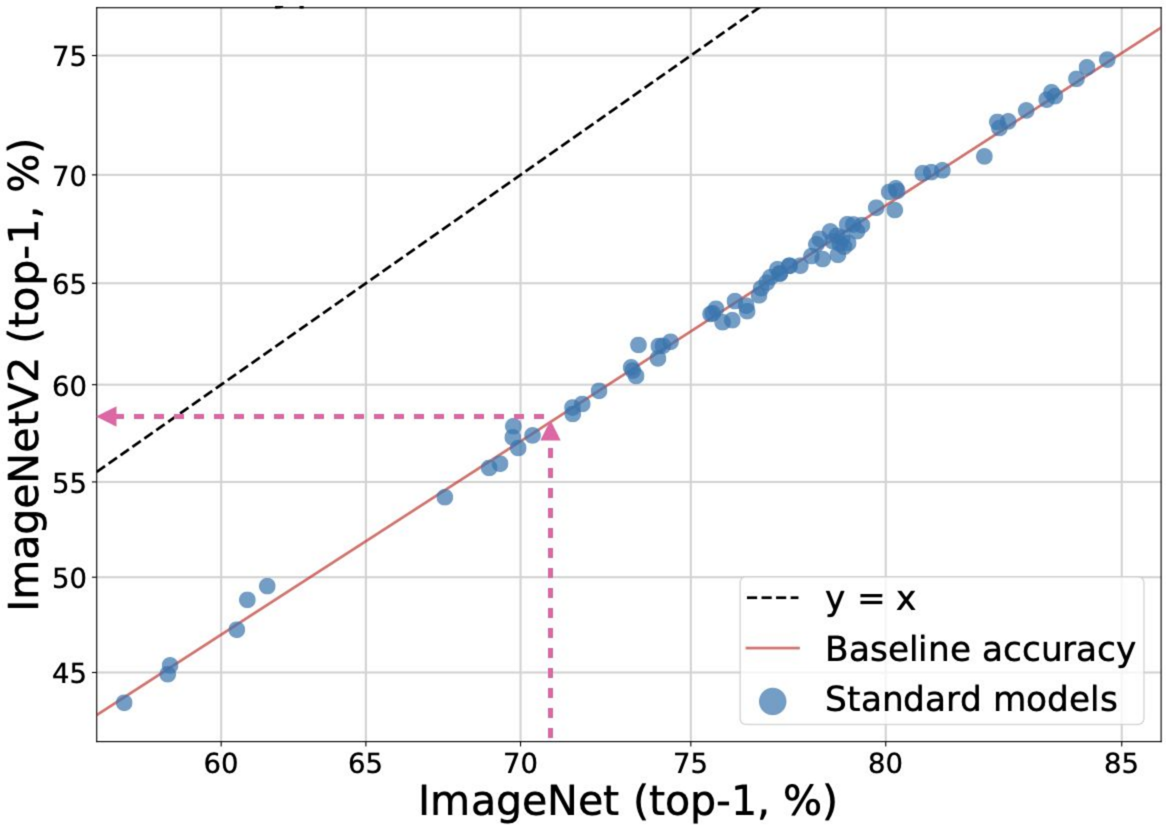
Slide credit: Ludwig Schmidt



[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

Slide credit: Ludwig Schmidt

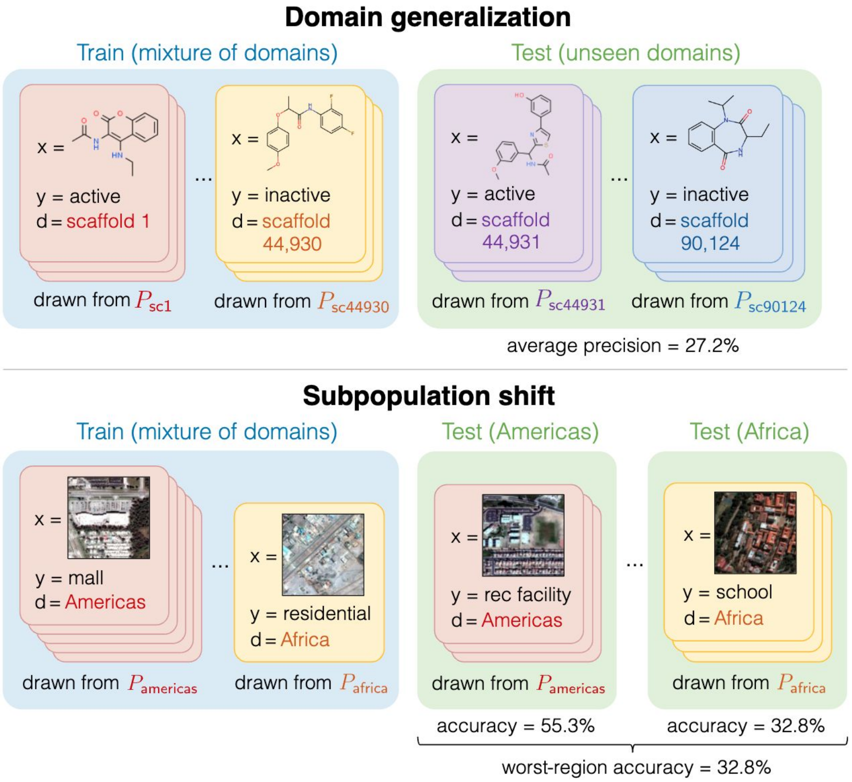
Expected out-of-distribution accuracy



In-distribution accuracy

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

Benchmarking distribution shifts



WILDS

A benchmark of in-the-wild distribution shifts spanning diverse data modalities and applications, from tumor identification to wildlife monitoring to poverty mapping.

https://wilds.stanford.edu/

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

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

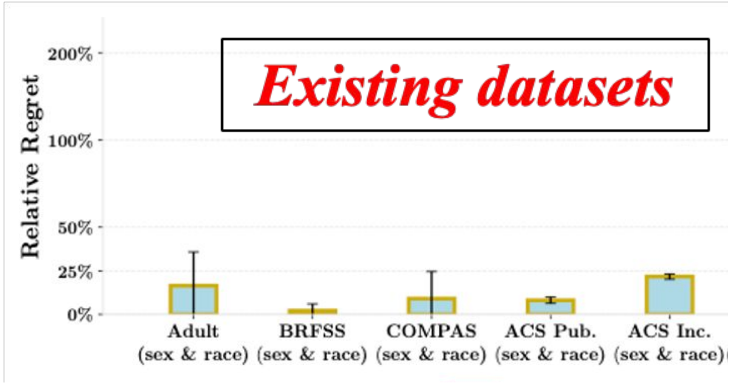
- So far: Humans are robust on all distributions. Can we get a universally good model?
- Implicitly, this view focuses on covariate shift (X-shift)
  - Traditional focus of ML
- On the other hand, we expect $Y|X$-shifts when there are unobserved factors
  - Traditional focus of causal inference
- For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

Even tabular benchmarks mainly focus on ***X*-shifts**

- Look at loss ratio of deployed model vs. best model for target

$$\frac{\mathbb{E}_Q[\ell(Y, f_P(X))]}{\min_{f\in\mathcal{F}} \mathbb{E}_Q[\ell(Y, f(X))]} - 1, \quad \text{where} \quad f_P \in \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, \mathbb{E}_P[\ell(Y, f(X))]$$

relative regret



Existing tabular benchmarks mainly focus on ***X*-shifts**

- Look at loss ratio of deployed model vs. best model for target

$$\frac{\mathbb{E}_Q[\ell(Y, f_P(X))]}{\min_{f\in\mathcal{F}} \mathbb{E}_Q[\ell(Y, f(X))]} - 1, \quad \text{where} \quad f_P \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_P[\ell(Y, f(X))]$$

relative regret

Liu, Wang, Cui, Namkoong, On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets

WhyShift

arxiv github

whyshift 0.1.3

`pip install whyshift`

- 7 spatiotemporal and demographic shifts from 5 tabular datasets

Dataset	Selected Settings	Shift Patterns
ACS Income	California → Puerto Rico	$Y\|X \gg X$
ACS Mobility	Mississippi → Hawaii	$Y\|X \gg X$
Taxi	New York City→ Botogá	$Y\|X \gg X$
ACS Pub.Cov	Nebraska → Louisiana	$Y\|X > X$
US Accident	California→ Oregon	$Y\|X > X$
ACS Pub.Cov	2010 (NY)→ 2017 (NY)	$Y\|X < X$
ACS Income	Younger→ Older	$Y\|X \ll X$

- Out of 169 source-target pairs with significant performance degradation, 80% of them are primarily attributed to $Y|X$-shifts.

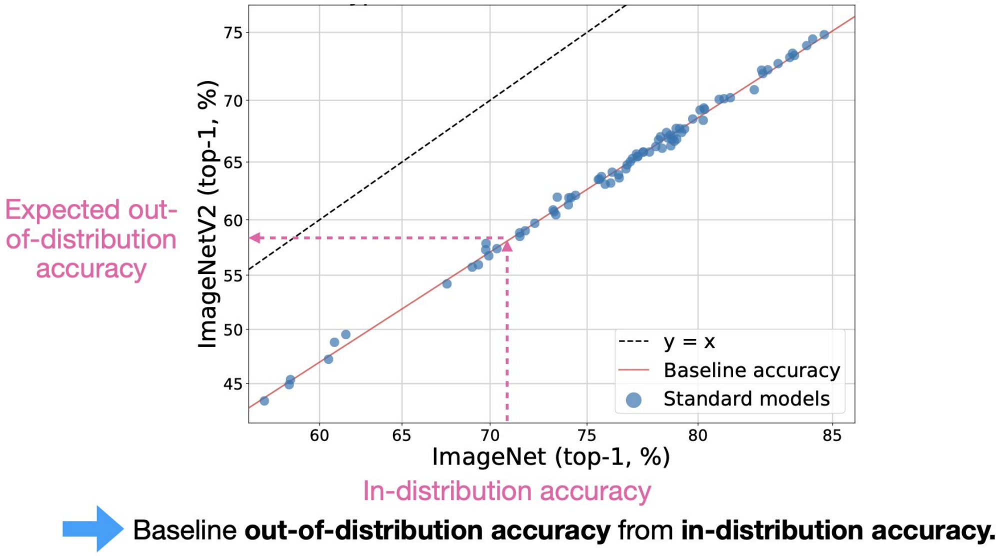
https://github.com/namkoong-lab/whyshift

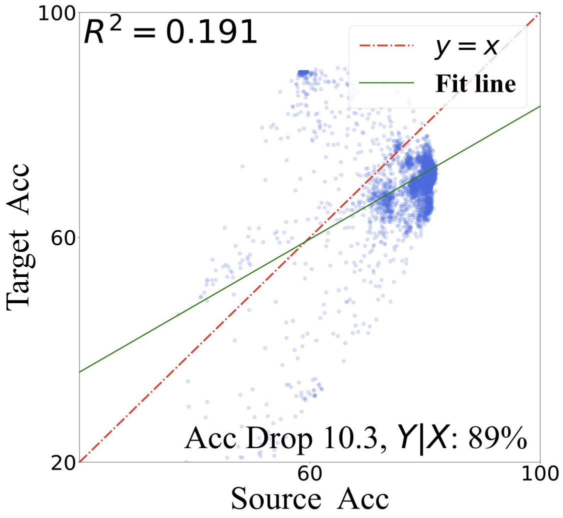
$Y|X$-shifts

- We can't just compare models based on their out-of-distribution performance
- It may not be feasible to simultaneously perform well across source and target
- We need to build an understanding of **why** the distribution changed!
- Previously observed empirical trends break if we look at $Y|X$-shifts

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

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





ACS Income (CA→ PR)

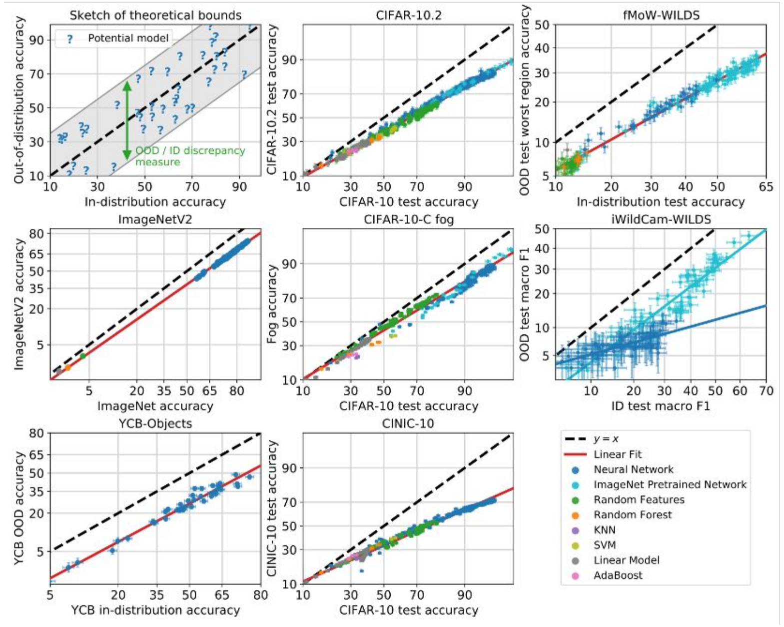
Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization.

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

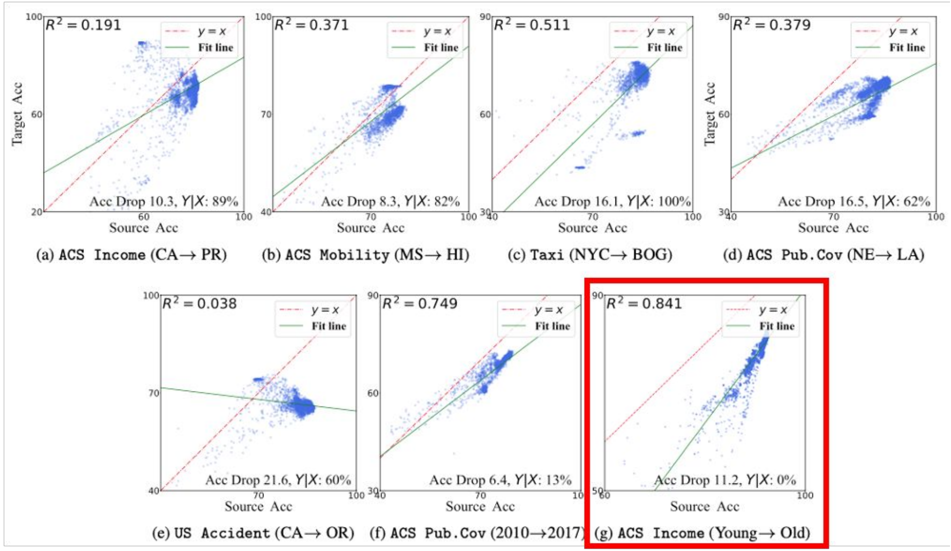
Accuracy-on-the-line **doesn't** hold under strong $\boldsymbol{Y|X}$-shifts

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

Image datasets



WHYSHIFT



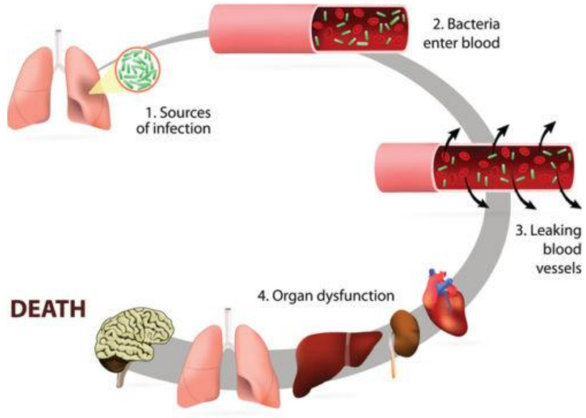
Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization.

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

Modeling: an application-driven perspective

- Measuring, understanding, and mitigating failures is nuanced
- "Modeling research" refers to building a simplified caricature of the real-world problem that we can analyze and understand
    - Not to be confused with "modeling" in the tech world
- Tremendous domain expertise is required to arrive at a concrete formulation
    - Often referred to as "institutional knowledge"
- Considered a first-order problem in disciplines like Economics, Operations Research, and Statistics. AI/ML community has long neglected this dimension.

Example: EPIC's sepsis risk scores



- More than ⅓ of deaths in US hospitals due to sepsis
- Epic Sepsis Model widely deployed as an early warning systems for sepsis in hundreds of US hospitals
- Developed based on data from 400K patients across 3 health systems from 2013-15
- Recent external validation found the model's performance to be substantially lower than vendor claims
  - Failed to identify 93% sepsis patients who did not receive timely administration of antibiotics
  - Also did not identify 67% of sepsis patients despite creating a large burden of alert fatigue

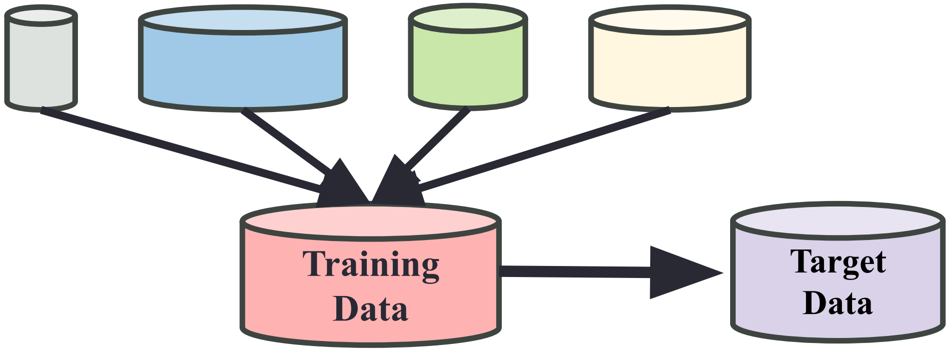
Wong et al., External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients, JAMA, 2021

Example: EPIC's sepsis risk scores

- It's common for risk scores developed on data from a particular region (North Carolina) to not generalize to other regions (New York)
- We need to better understand the level of heterogeneity that exists in data
  - How different are the patients from the two regions?
- How do we catch these failure modes?
  - More rigorous evaluation protocols
- How do we diagnose the cause of this failure?
  - Differences in age? Differences in latent factors? (e.g., genetics)
- Which interventions do we take to mitigate such failures?
  - Need better data collection mechanisms and algorithms
  - Resource constraints must be more explicitly modeled

Modeling data heterogeneity

ML models are based on *heterogeneous* data sources

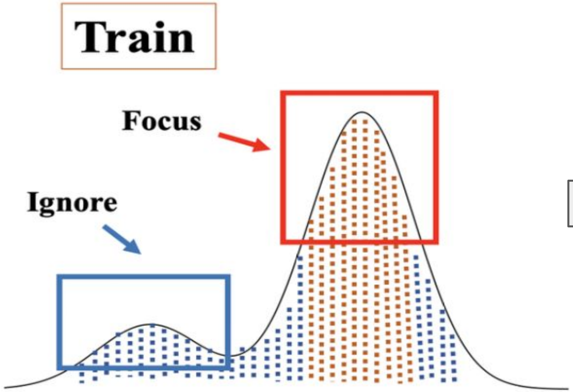


- multiple ***environments***
- different ***$Y|X$ distributions***
- different ***data size***

Today: opportunities and challenges of heterogeneity

Data Heterogeneity Matters

Ignoring heterogeneity makes models *ignore underrepresented groups*



Amazon scraps secret AI recruiting tool that showed bias against women REUTERS

Data Heterogeneity Matters

Ignoring heterogeneity makes ML algorithms *fail to generalize*

Self-Driving

common scenes

→

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

Data Heterogeneity Matters

Ignoring heterogeneity makes ML algorithms *unreliable*

Health Care


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

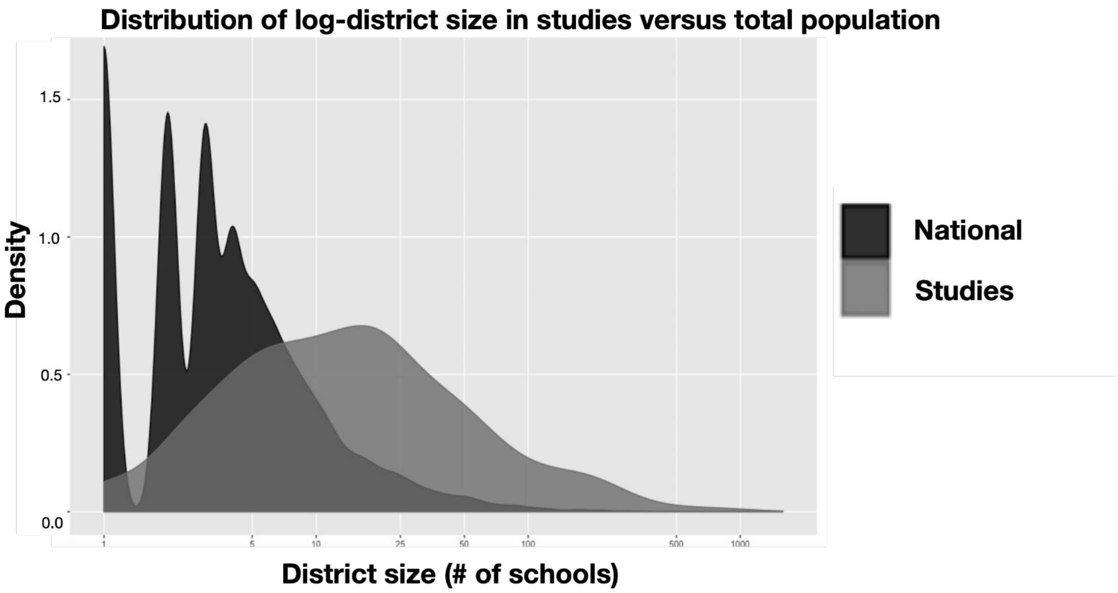
Selection bias in data leads models to focus on spurious correlations

Data Heterogeneity Matters

Ignoring heterogeneity brings *unreliable scientific discoveries*

Social Science

Even for carefully designed randomized trials, there is *large selection bias*



[Tipton et al. 2019] The convenience of large urban school districts: a study of recruitment practices in 37 randomized trials

Outline

Part 1: Benchmarking performance under distribution shift

Part 2: A critical review of existing approaches

Part 3: Application-specific modeling of data heterogeneity

Part 4: Towards heterogeneity-aware machine learning

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

Two existing approaches to distribution shift

1. Make modeling assumptions

2. Scale up data and models

Two existing approaches to distribution shift

1. **Make modeling assumptions**

2. Scale up data and models

Distributionally Robust Optimization (DRO)

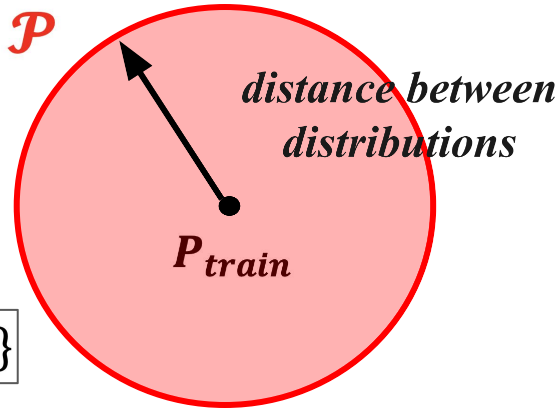
Empirical Risk Minimization

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

DRO

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

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



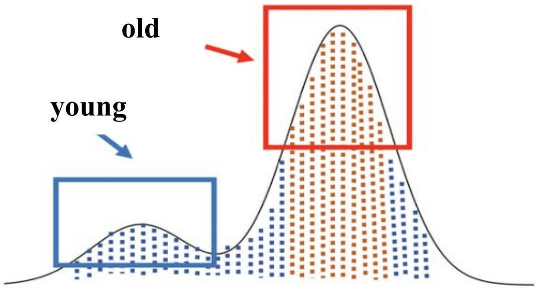
Instead of minimizing loss over training distribution, minimize loss over distributions ***near*** it

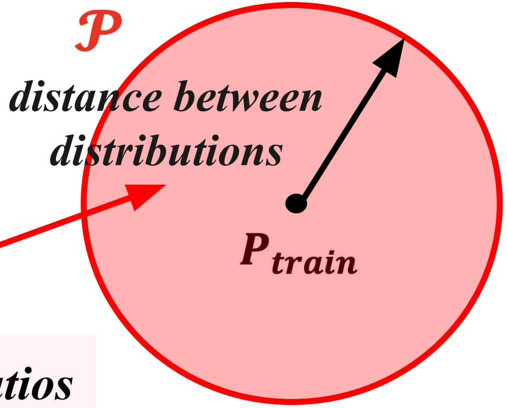
Distributionally Robust Optimization (DRO)

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)

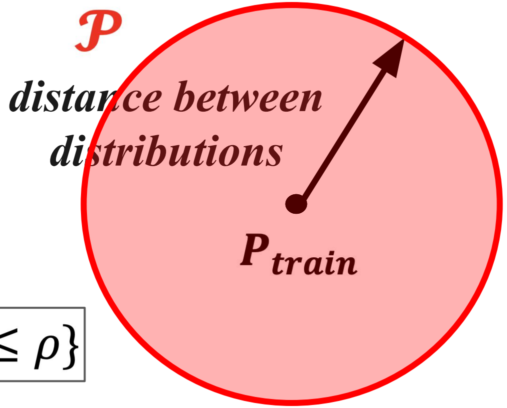
Empirical Risk Minimization

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

DRO

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

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



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

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

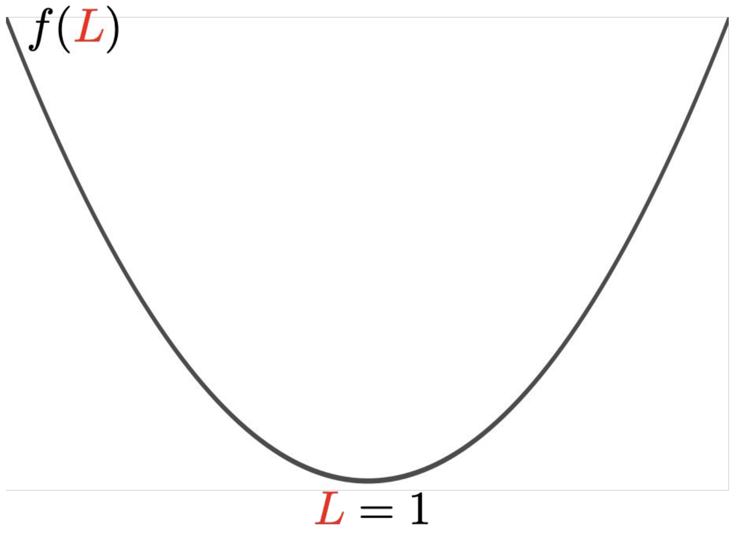
***f*-divergence:** about *densities*

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

For a convex function,

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

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



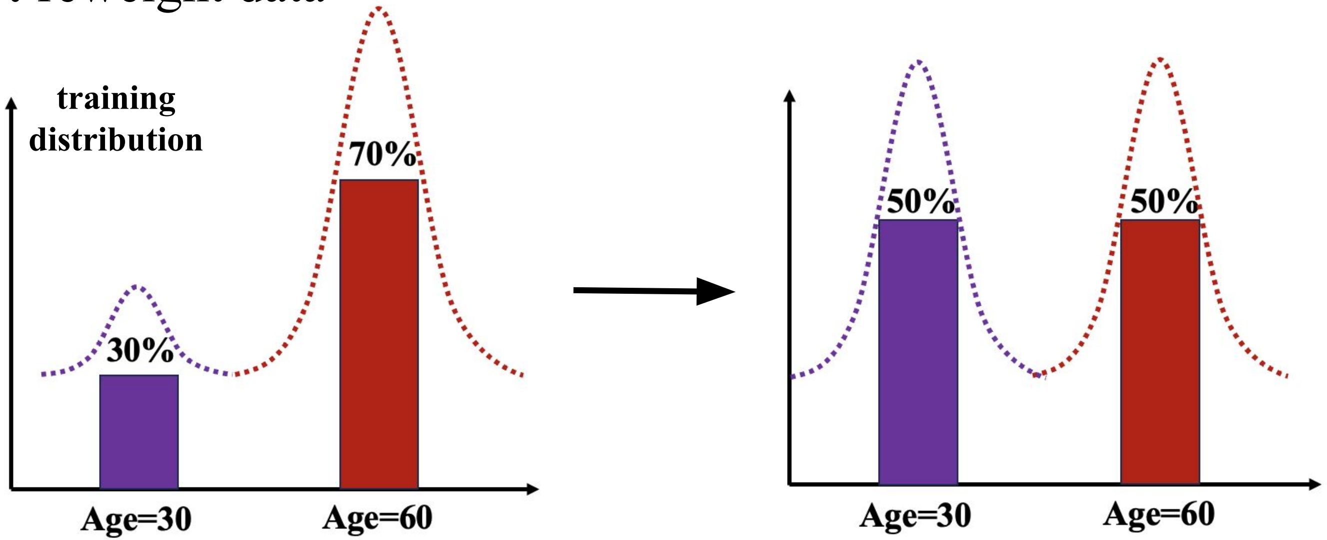
Examples: set of distributions we care about

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

recall the objective

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

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



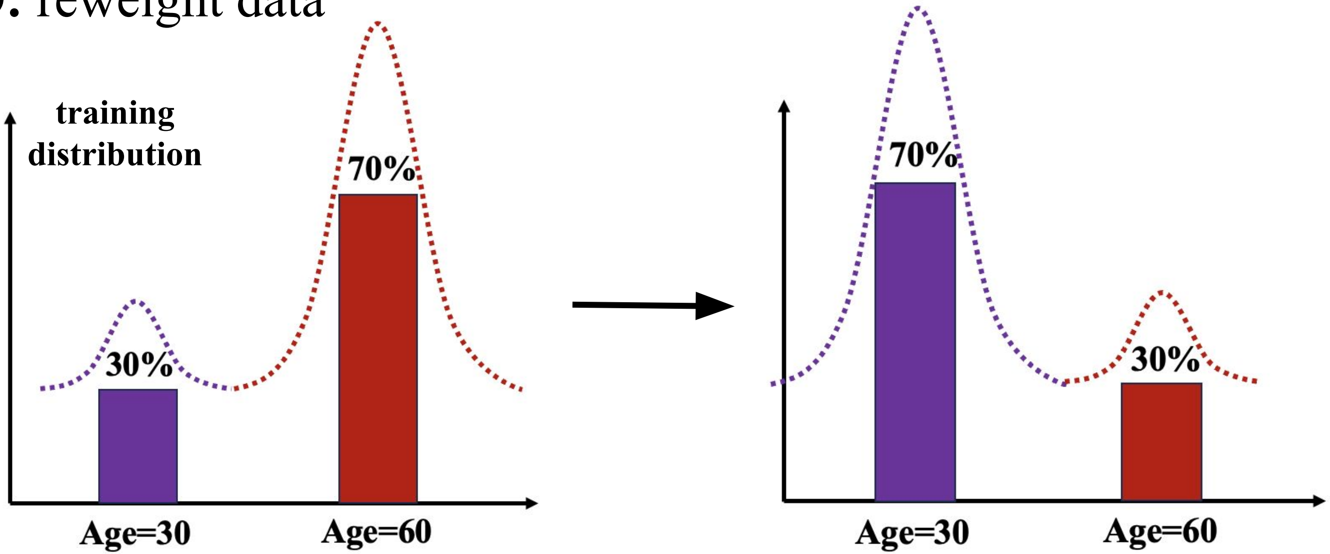
Examples: set of distributions we care about

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

recall the objective

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

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



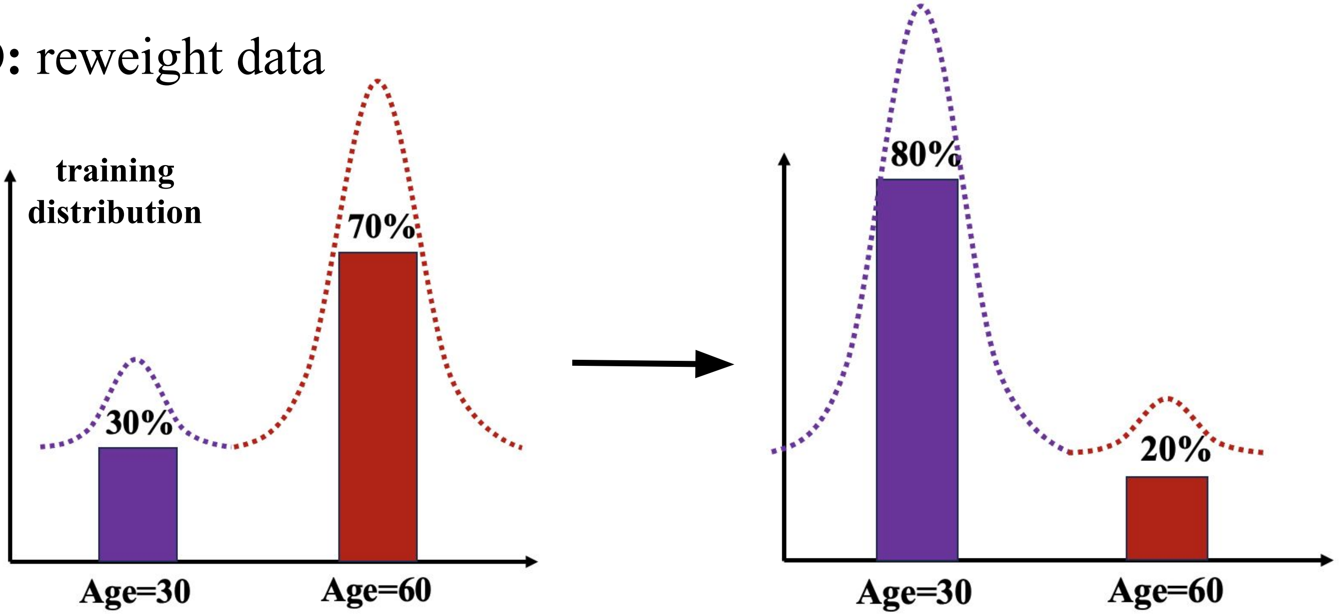
Examples: set of distributions we care about

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

recall the objective

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

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



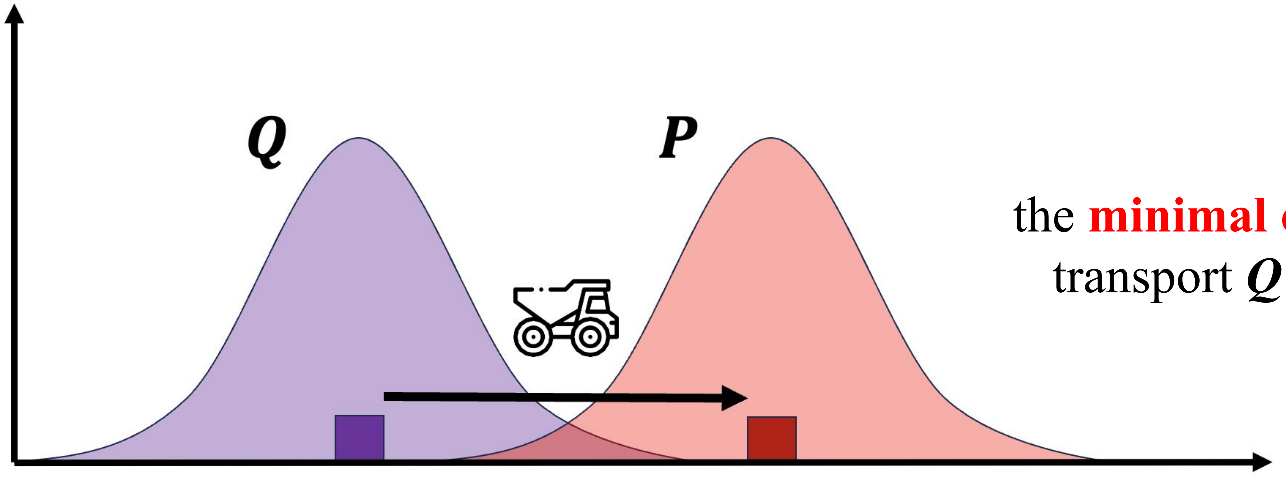
Examples: set of distributions we care about

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

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***

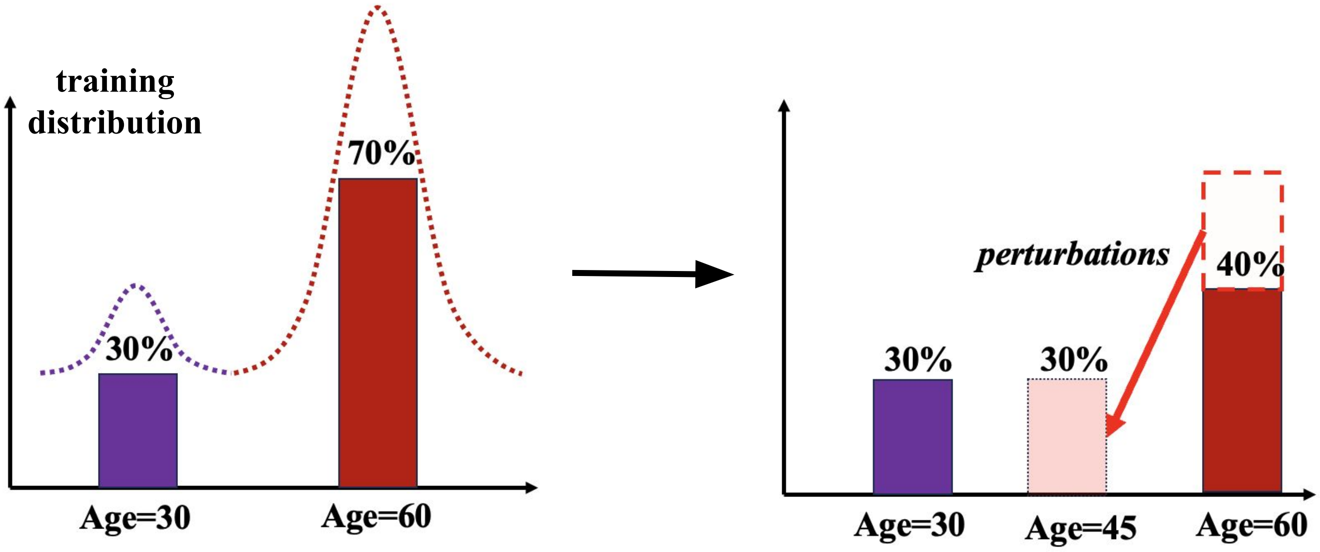
Examples: set of distributions we care about

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

recall the objective

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

Wasserstein-DRO: perturb data



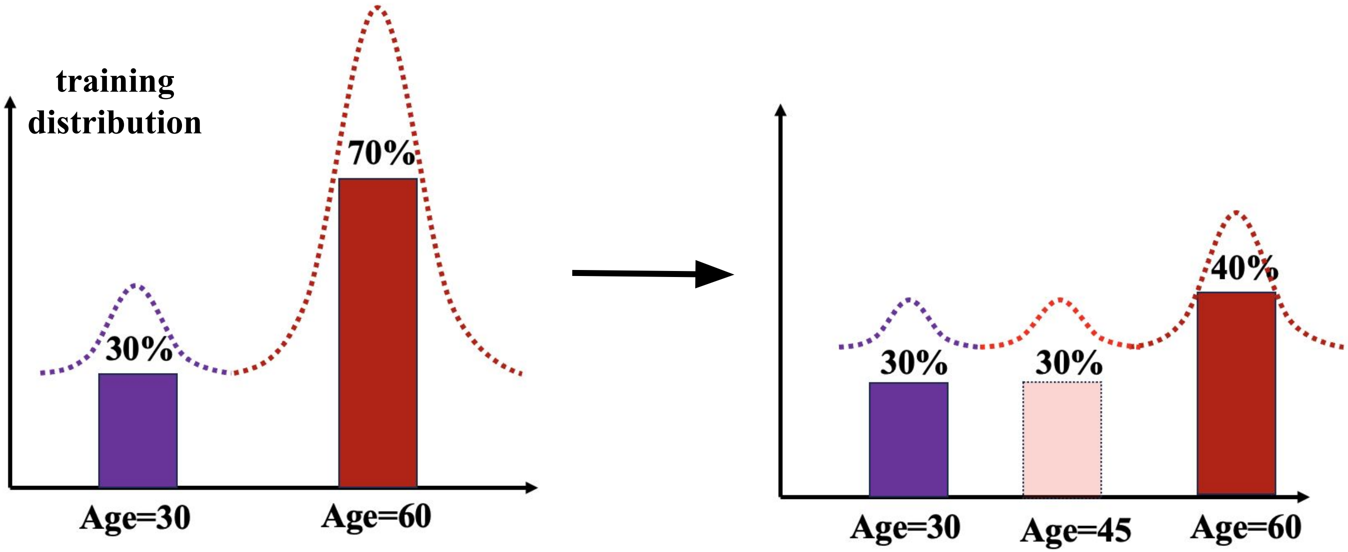
Examples: set of distributions we care about

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

recall the objective

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

Wasserstein-DRO: perturb data



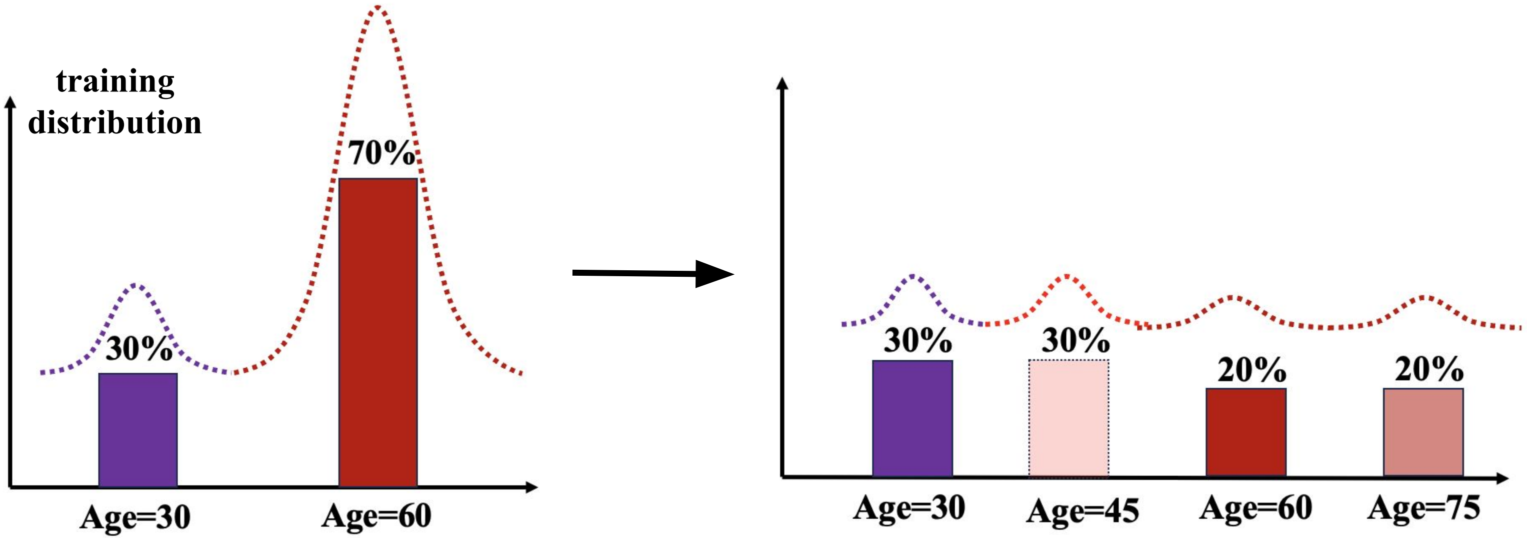
Examples: set of distributions we care about

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

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

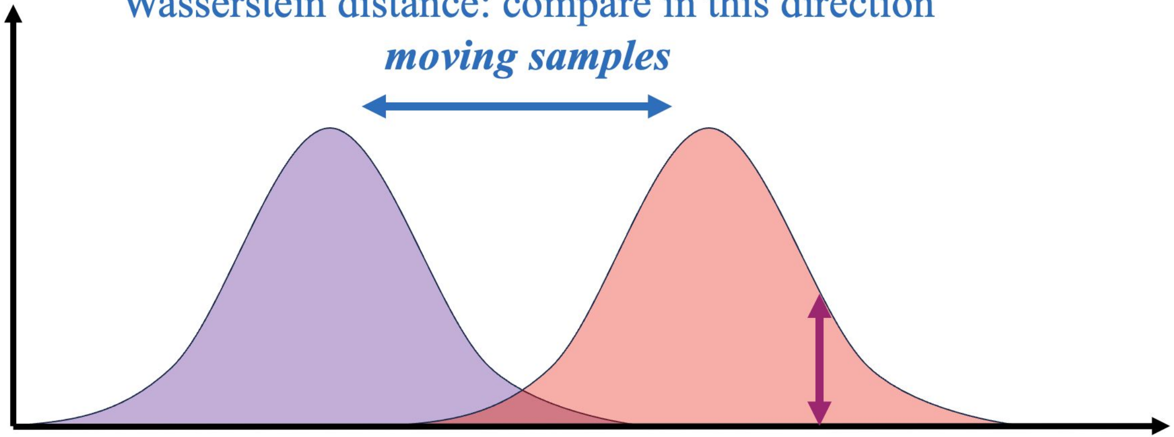
$$\mathcal{P} = \{Q : Dist(Q, P_{train}) \leq \rho\}$$

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

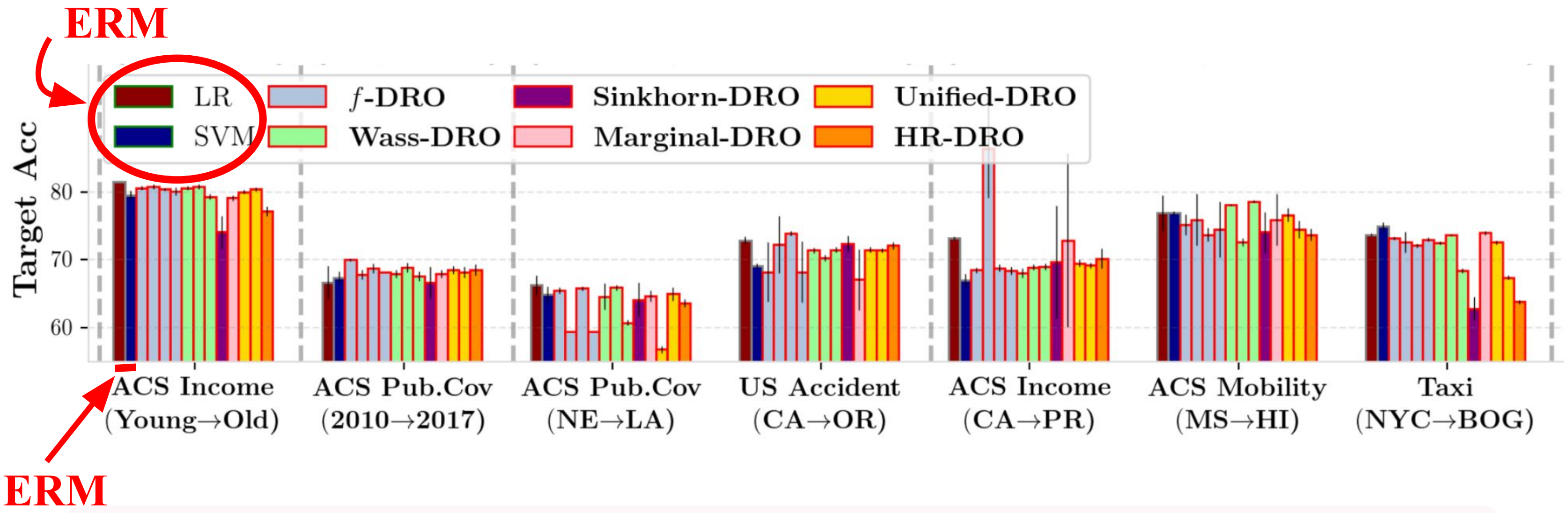
Goal

Carefully choose the set $\mathcal{P}$

Do well on real distribution shifts!

Hope the worst-case distribution captures real shifts

Critical View of DRO: not better than ERM!



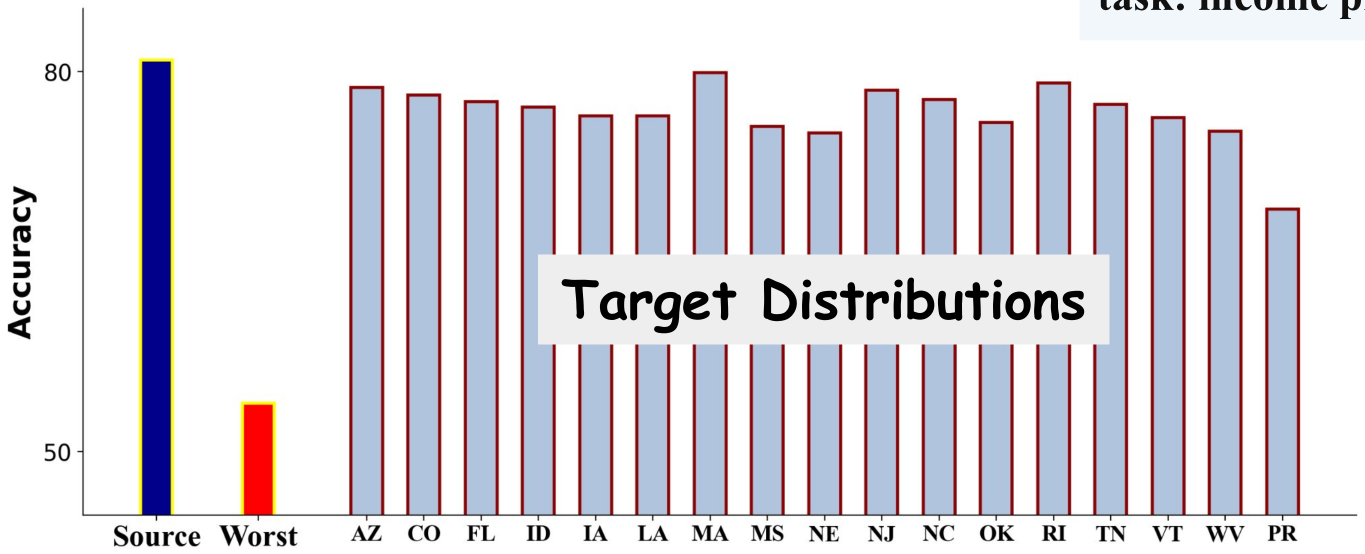
DRO does NOT show significant improvements over ERM!

Hard to choose this set of distributions P!!!

Liu, J., Wang, T., Cui, P., & Namkoong, H. (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*.

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

task: income prediction



χ^2-DRO: the worst-case distribution is too conservative!

Hard to pick set of distributions $\boldsymbol{P}$; can we do better?

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

Hard to pick set of distributions $\boldsymbol{P}$; 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.

Invariant Learning

Modeling → **Goal**

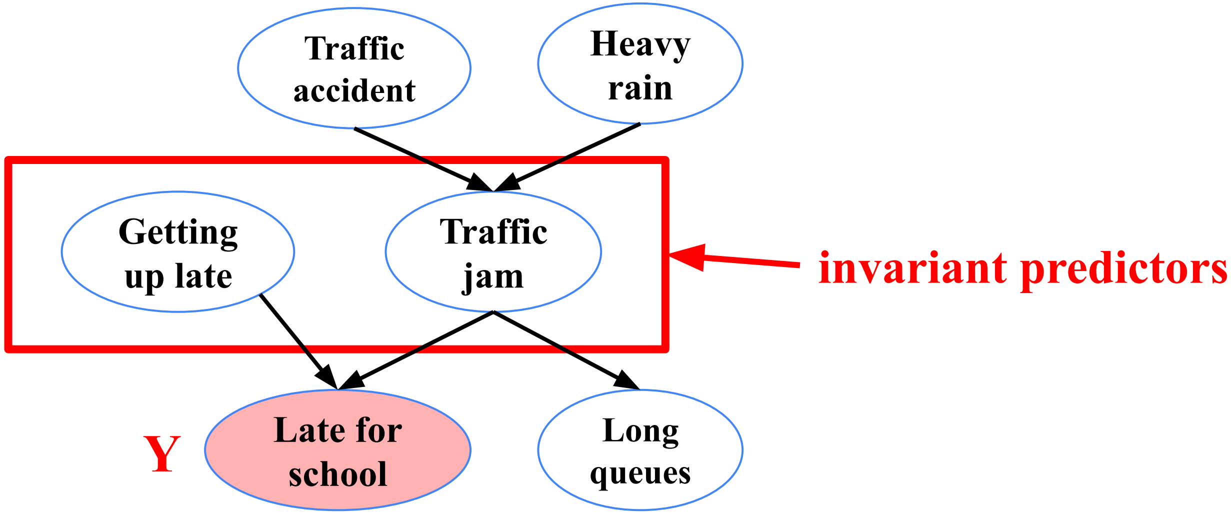
Learn an invariant mechanism across given environments

Generalize to new environments

Assume true invariant mechanism can be learned with given heterogeneous data

Invariant Learning: Invariant Causal Prediction

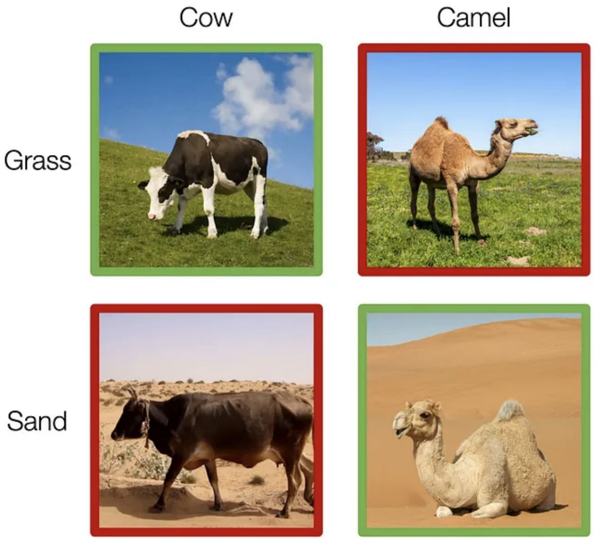
Find <u>subset</u> of covariates X with an **invariant** relationship to Y across environments!



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

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

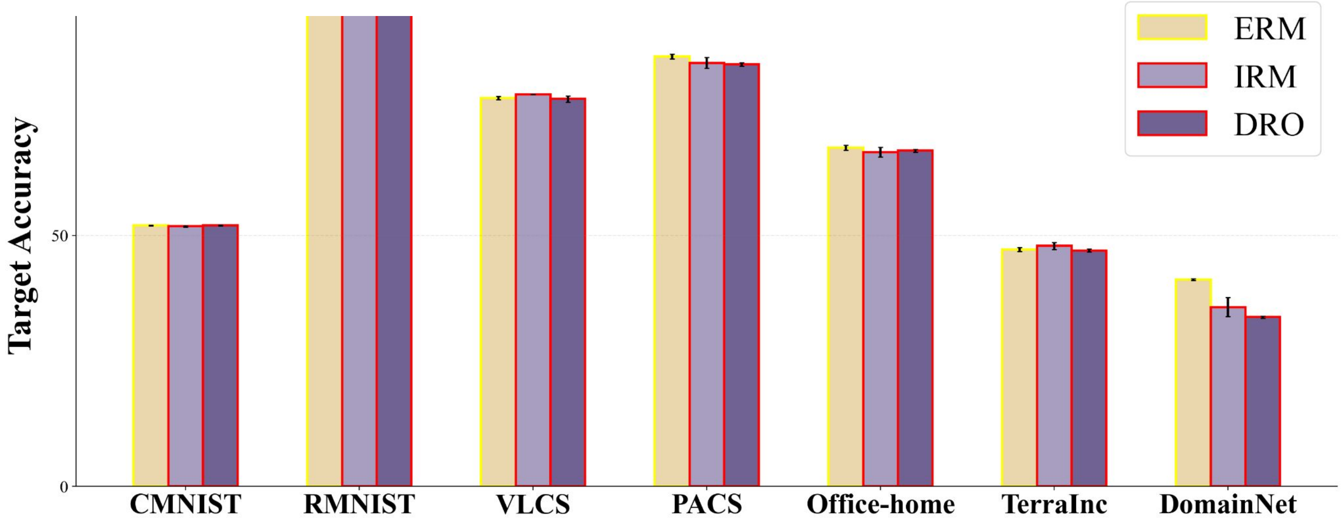
Do these assumptions work in practice?

NO! Domain generalization methods do not beat ERM!

IN SEARCH OF LOST DOMAIN GENERALIZATION

Ishaan Gulrajani*
Stanford University
igul222@gmail.com

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



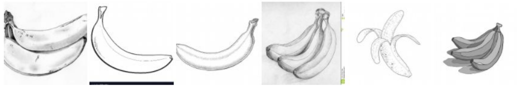
Plot generated from Table 4 from Gulrajani, I., & Lopez-Paz, D. (2020, October). In Search of Lost Domain Generalization. In International Conference on Learning Representations.

Two existing approaches to distribution shift

1. Make modeling assumptions

2. **Scale up data and models**

OpenAI’s 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%

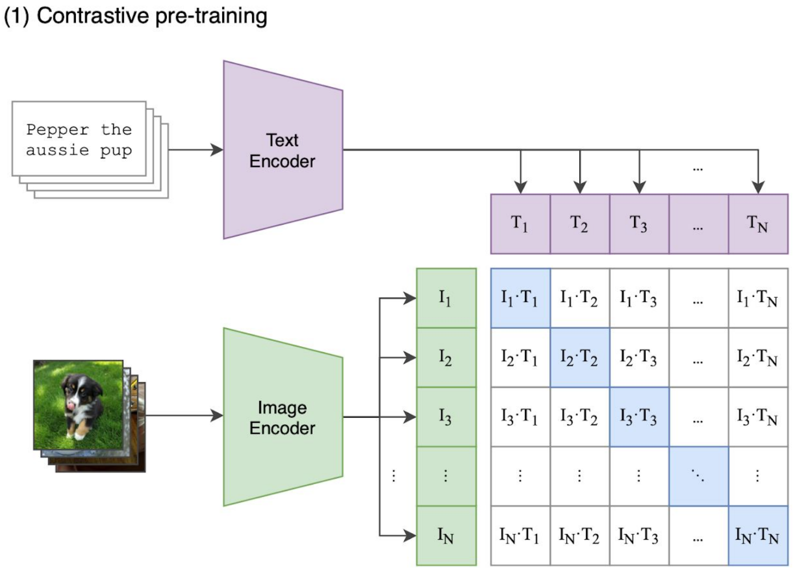
Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger, Sutskever

Learning Transferable Visual Models From Natural Language Supervision (2021)

CLIP: scale up data

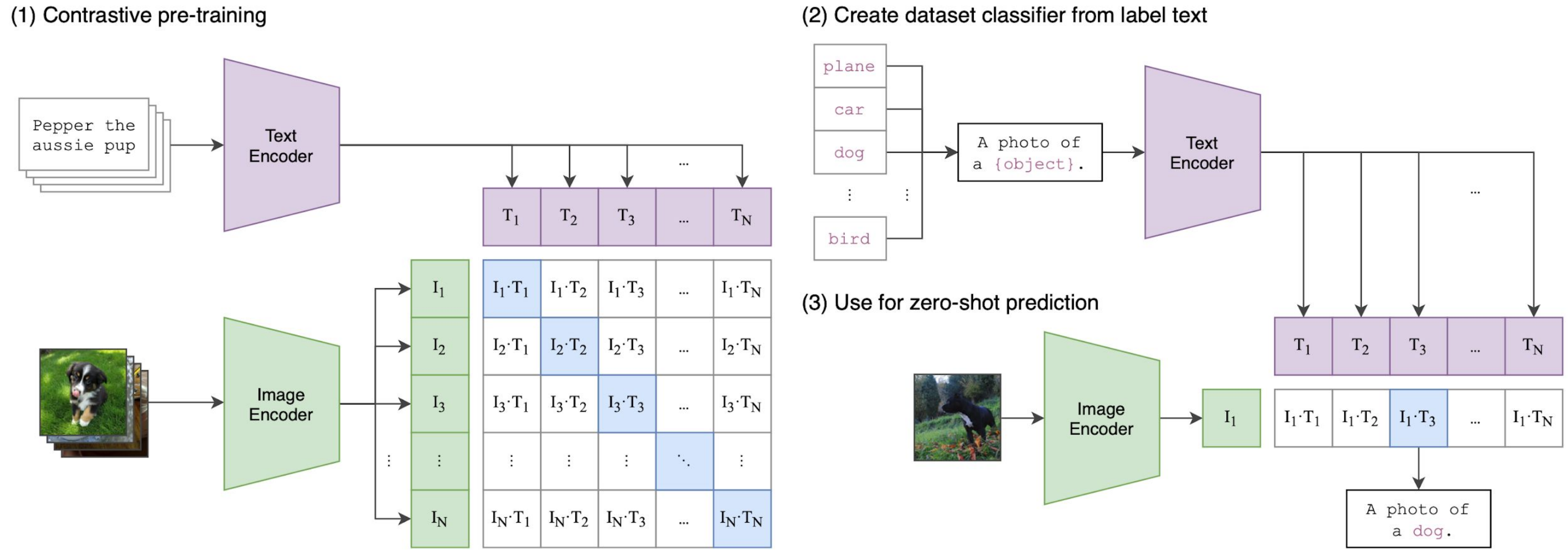
Supervised ImageNet training data	CLIP training data
● ~1M (image, label) pairs ● Data from one source ● Needs labelers	● ~400M (image, caption) pairs ● Data from all over the internet; more diverse ● No need for labelers; there is lots of (image, caption) data across the internet

CLIP: learn relationship between images and captions



1. Learn embeddings for images and text so that embeddings for images and text that correspond are similar, and embeddings for images and text that don't are different

CLIP: learn relationship between images and captions



1. Learn embeddings for images and text so that embeddings for images and text that correspond are similar, and embeddings for images and text that don’t are different
2. To make a zero-shot classifier: for each image embedding, find the closest class label (caption) embedding

→ enables using a huge dataset of (image, caption) pairs

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~~

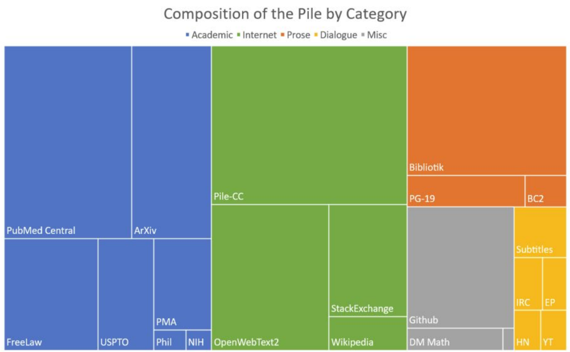
~~Model architecture~~

Scale up data for LLMs, too

→ bigger, more diverse datasets → better LLMs →

Common Crawl

The Pile



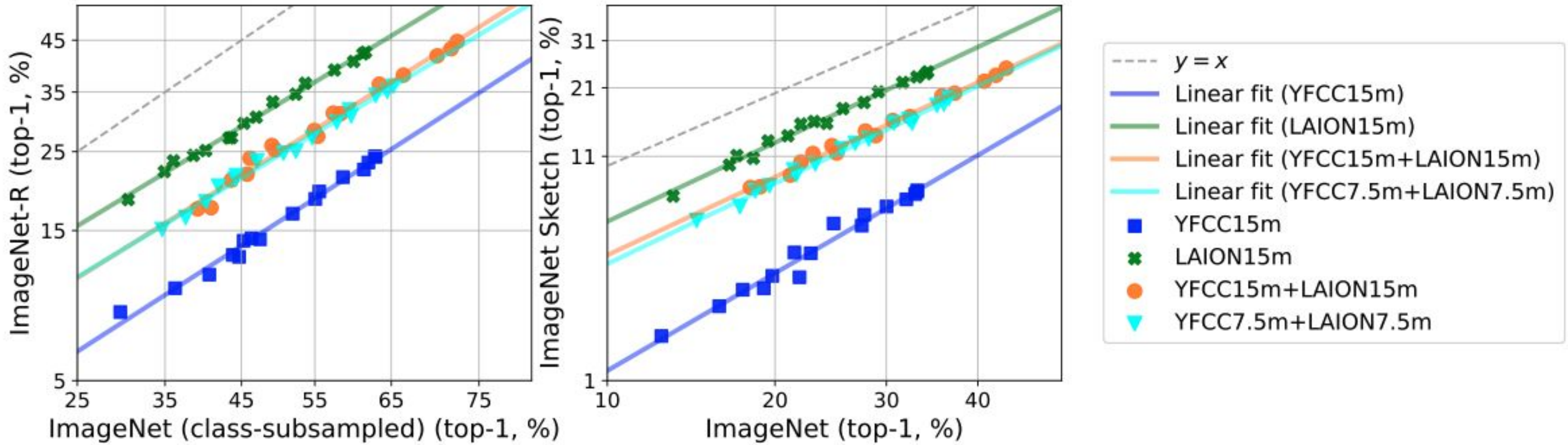
Red Pajama

Have we solved domain generalization?

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

Which training data do we use?

DataComp

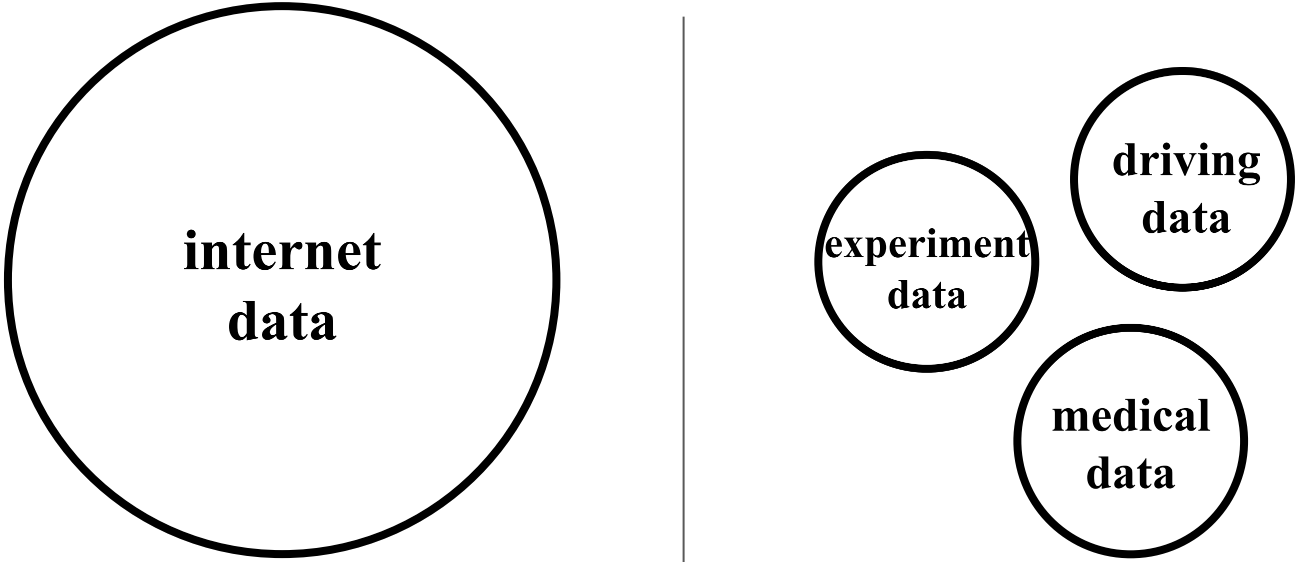
datacomp.ai

Welcome to DataComp, the machine learning benchmark where the models are fixed and the challenge is to find the best possible data!

DATACOMP:
In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2], Gabriel Ilharco*[1], Alex Fang*[1], Jonathan Hayase[1], Georgios Smyrnis[5], Thao Nguyen[1], Ryan Marten[7,9], Mitchell Wortsman[1], Dhruba Ghosh[1], Jieyu Zhang[1], Eyal Orgad[3], Rahim Entezari[10], Giannis Daras[5], Sarah Pratt[1], Vivek Ramanujan[1], Yonatan Bitton[11], Kalyani Marathe[1], Stephen Mussmann[1], Richard Vencu[6], Mehdi Cherti[6,8], Ranjay Krishna[1], Pang Wei Koh[1,12], Olga Saukh[10], Alexander Ratner[1,13], Shuran Song[2], Hannaneh Hajishirzi[1,7], Ali Farhadi[1], Romain Beaumont[6], Sewoong Oh[1], Alex Dimakis[5], Jenia Jitsev[6,8], Yair Carmon[3], Vaishaal Shankar[4], Ludwig Schmidt[1,6,7]

Sometimes you need (costly) specialized data!



Many important applications!

$ cheap!

$$$ 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!

Outline

Part 1: Benchmarking performance under distribution shift

Part 2: A critical review of existing approaches

Part 3: Application-specific modeling of data heterogeneity

Part 4: Towards heterogeneity-<u>aware</u> machine learning

Alarm and Proposition

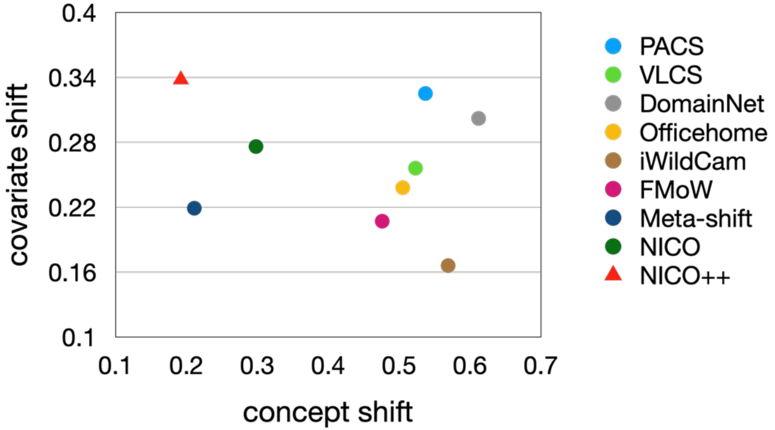
- Empirically, current algorithmic robustness methods (e.g. DRO, invariant learning) do **not** improve domain generalization.
- These methods usually make assumptions about the relationship between data distributions, but do **not** check them.
- In theory, **no** model can generalize to arbitrarily shifted distributions.
- A more realistic goal of studying OOD generalization (or distribution shifts) is to deal with **real** rather than **hypothetical** distribution shifts.
- In response, we propose carefully **understanding** the real distribution shift patterns in each application.

Distribution shifts are complicated in real applications

- Different **types**
  - different $\boldsymbol{X}$ distributions
    - examples: demographic shifts, minority groups
  - different $\boldsymbol{Y} \mid \boldsymbol{X}$ distributions
    - examples: different user preferences over time

Distribution shifts are complicated in real applications

- Different ***Applications***
  - For **image data**: ***X***-shifts are more common
  - A sample will not have different labels in training and testing, as X include complete information for predicting Y



Xingxuan Zhang, et al. **NICO++: Towards Better Benchmarks for Out-of-Distribution Generalization**. *CVPR*, 2023.

Distribution shifts are complicated in real applications

- Different ***Applications***
  - For **tabular data**: both X-shift and $Y|X$-shift exists
  - A sample may have different labels in training and testing when X can not provide complete information for predicting Y, due to missing variables

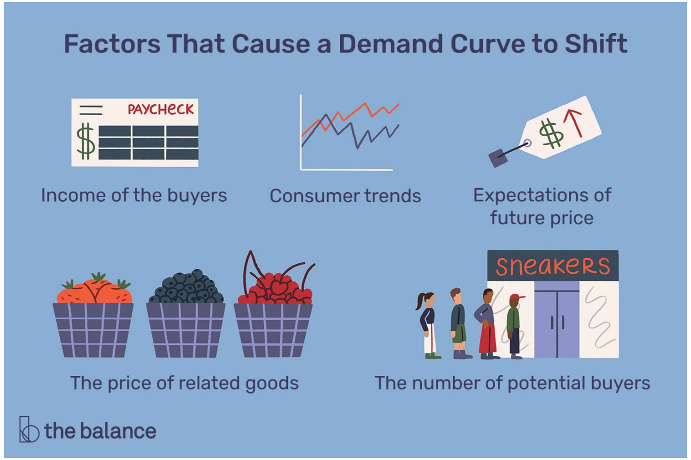


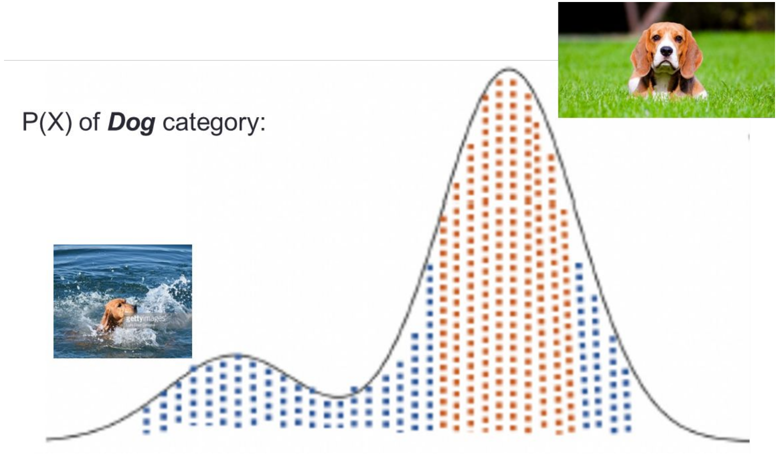
PHOTO: TH BALANCE / MADDY PRICE

Average rent for a 1-bedroom

Manhattan	Pittsburgh
$3,075	**$1,050**

Heterogeneity: a language for characterizing distribution shifts

- Modeling heterogeneity is an art of pursuing the tradeoff between commonality and differences



X Heterogeneity

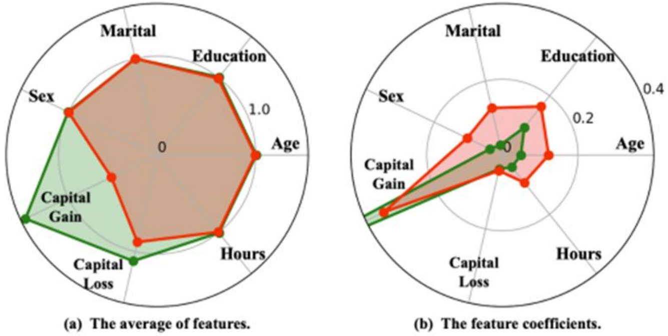
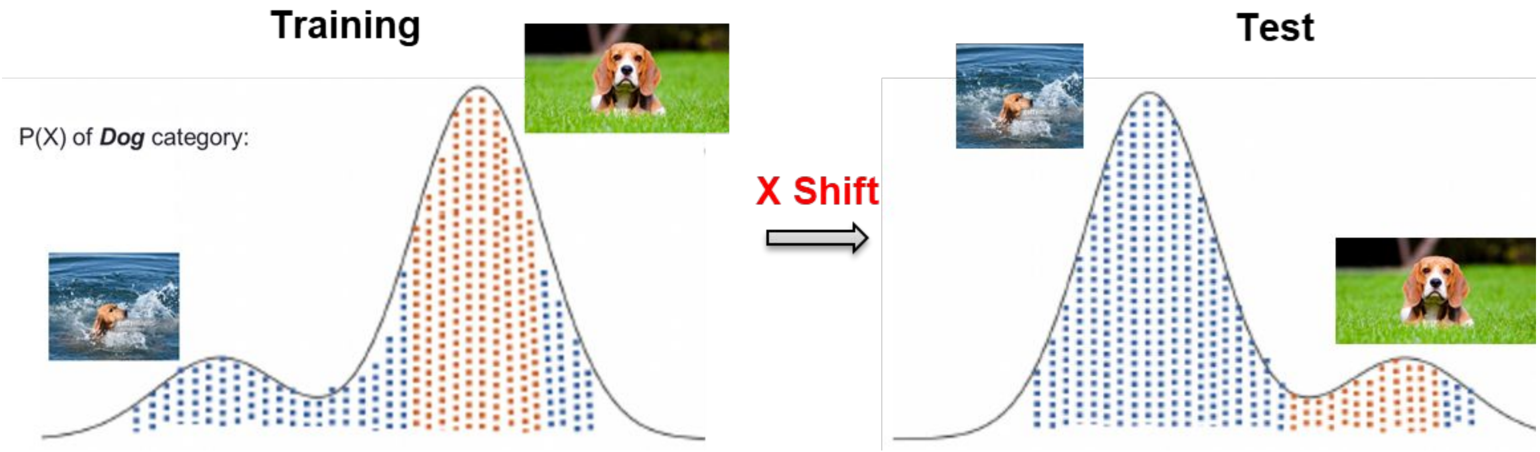


Figure 2: Results on the Adults data. Here we show the average of features and the feature coefficients of the two learned sub-populations.

Y|X Heterogeneity

Using heterogeneity to characterize distribution shifts

- **Two cases**
  - The shift is induced by the proportion of heterogeneity components
    - Do NOT need target distribution
    - Divide and conquer, or learning invariance from heterogeneity



Using heterogeneity to characterize distribution shifts

- **Two cases**
  - The shift is induced by the proportion of heterogeneity components
  - The shift goes beyond the heterogeneity identified in training distribution
    - Need the target distribution
    - Diagnose the shift region, and collect more data or features accordingly

A different philosophy

- **Application specific** v.s. One model fits all (***Model-centric View***)
  - Given an application, first understand its real distribution shift pattern characterized by heterogeneity, and then derive realistic assumptions accordingly for the subsequent modeling process
- **Less is more** v.s. The more the better (***Data-centric view***)
  - Distribution shift problem can be regarded as a problem of data representativeness w.r.t. X or Y|X which CANNOT be solved by collecting MORE data, but need to collect the **RIGHT** data.

Outline

Part 1: Benchmarking performance under distribution shift

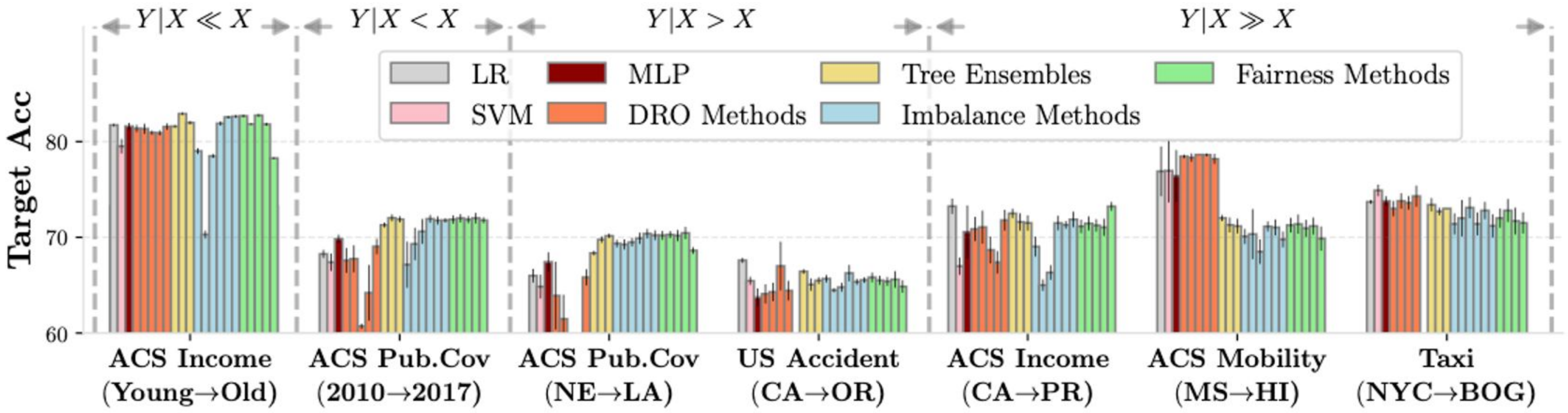
Part 2: A critical review of existing approaches

Part 3: Application-specific modeling of data heterogeneity

Part 4: Towards heterogeneity-<u>aware</u> machine learning

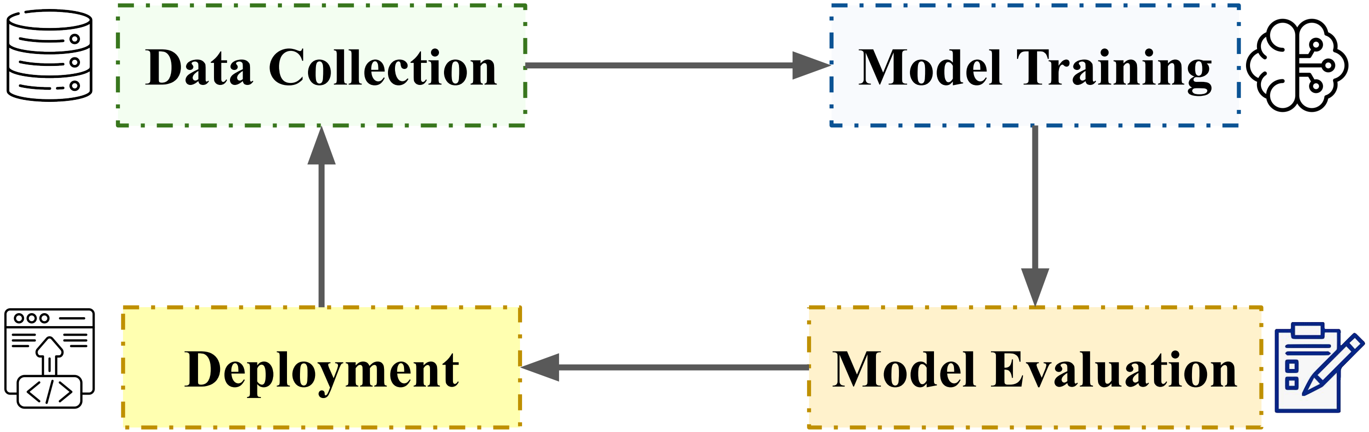
~~One size fits all~~

- Algorithms **don't** exhibit consistent rankings over different shifts
- Algos **sensitive** to configurations: rankings vary across 7 different settings



Understanding heterogeneity throughout the modeling process

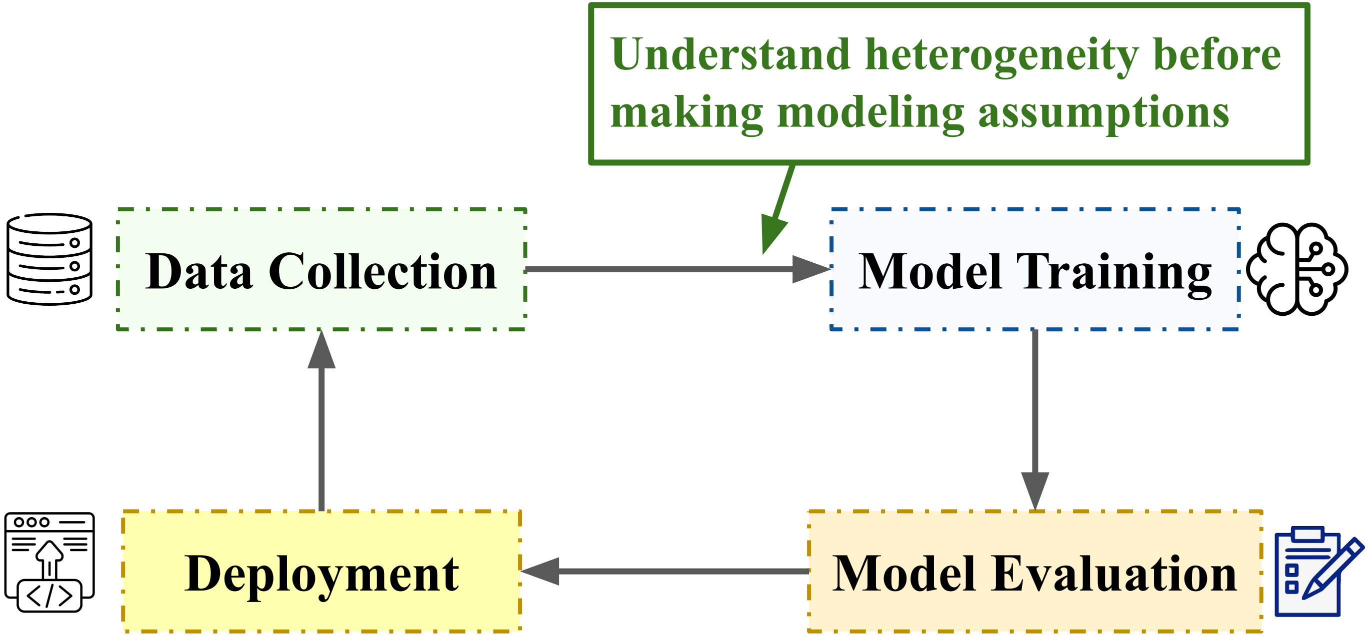
We discuss how understanding heterogeneity can be important throughout the modeling process



Perspective 0: Data as infrastructure

- Data is the infrastructure that all AI models build on
  - Big set up cost
- What are the main resource constraints?
  - Time, money, human & social capital
- Inclusion-exclusion criteria: Who in the data? Who’s **not** in the data?
  - Data depends on the social conditions under which it’s collected
  - See CVPR 2020 tutorial by Timnit Gebru and Emily Denton
- Cross-pollination needed with best practices experimental design
  - Long line of work on a thoughtful design process for experiments
  - For example, see Beth Tipton’s 2020 OCI talk
- Rigorous documentation: Datasheets (Gebru et al. 2018, Mitchell et al. 2019)

Understanding heterogeneity throughout the modeling process



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 $X \longrightarrow Y$ across the entire population. This assumption can be inappropriate.

Example: discover heterogeneous subpopulations: **predictive heterogeneity**

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

Definition

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

mutual information with ***model constraints***

optimization algorithm

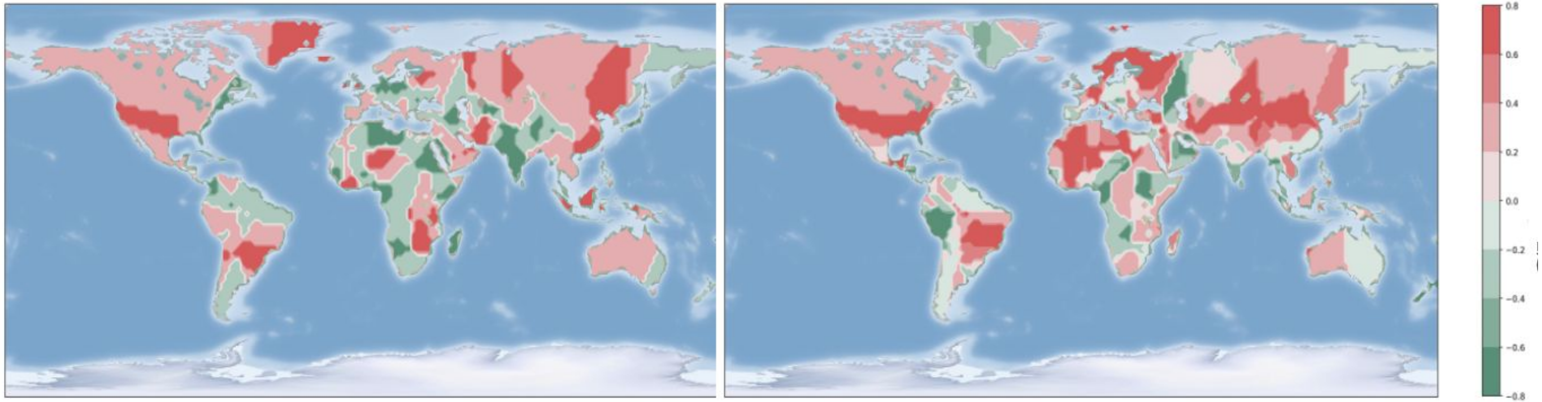
finite sample bounds

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.

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

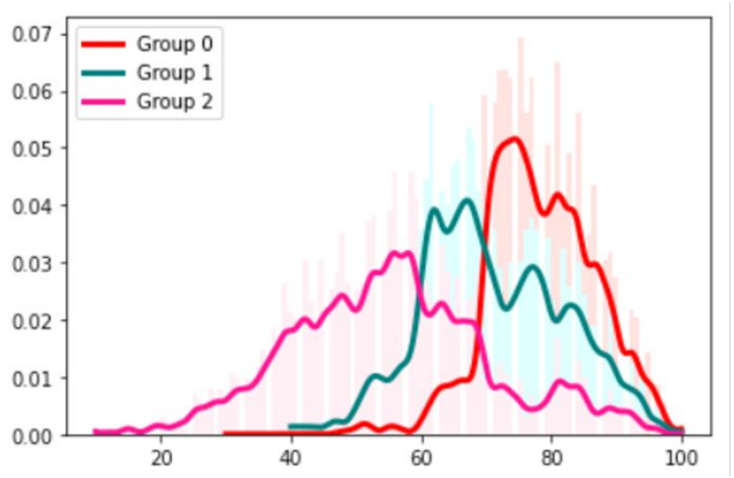
learned sub-populations correspond to ***different crop types; model separately!***

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!

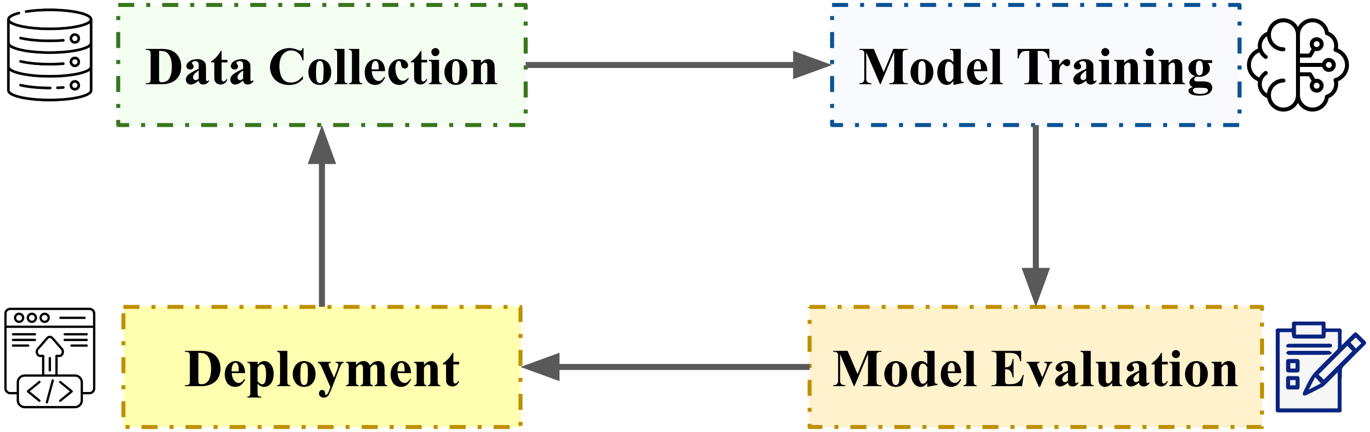
ERM: SPO2 Renal Neurologic Diabetes

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

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?

Understanding heterogeneity throughout the modeling process



Understand important subsets of training data

Perspective 2: 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

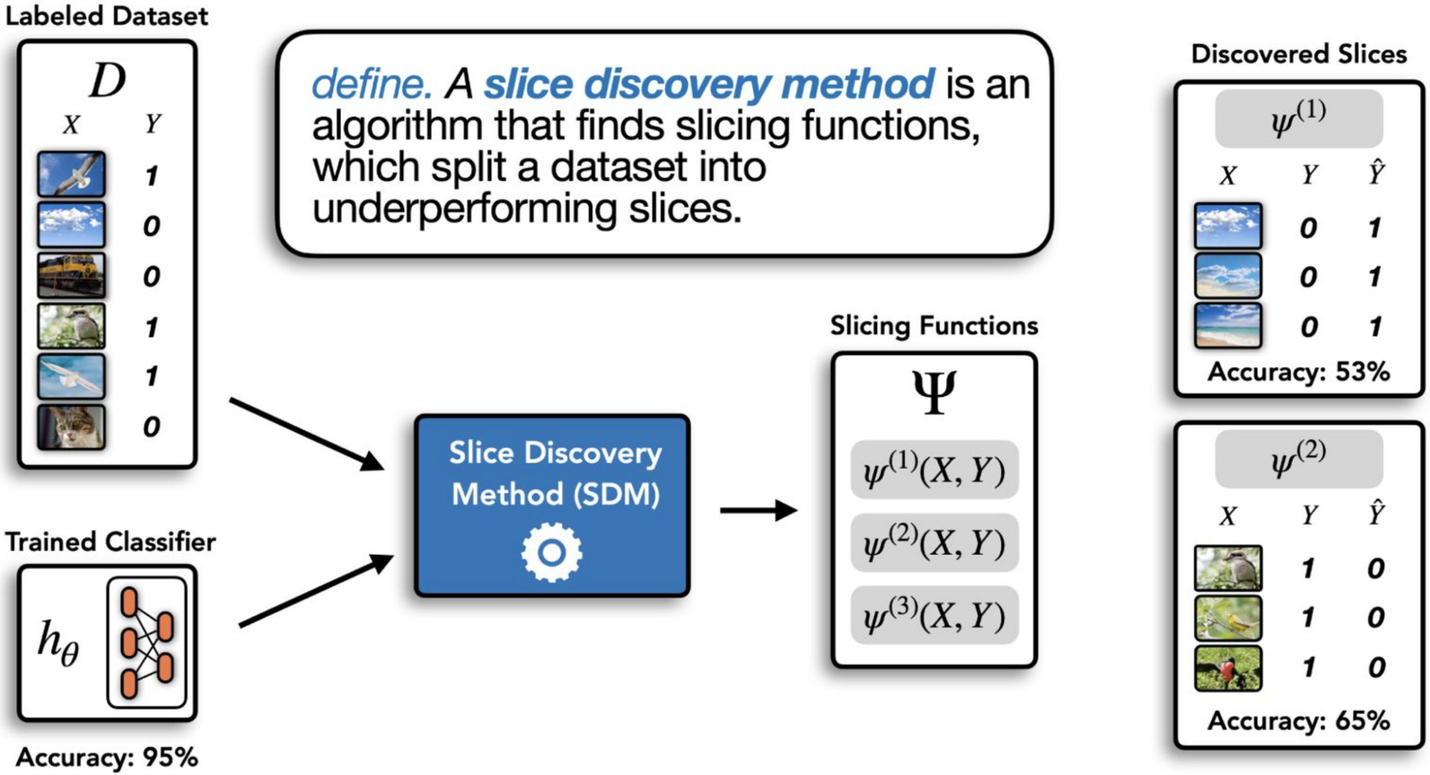


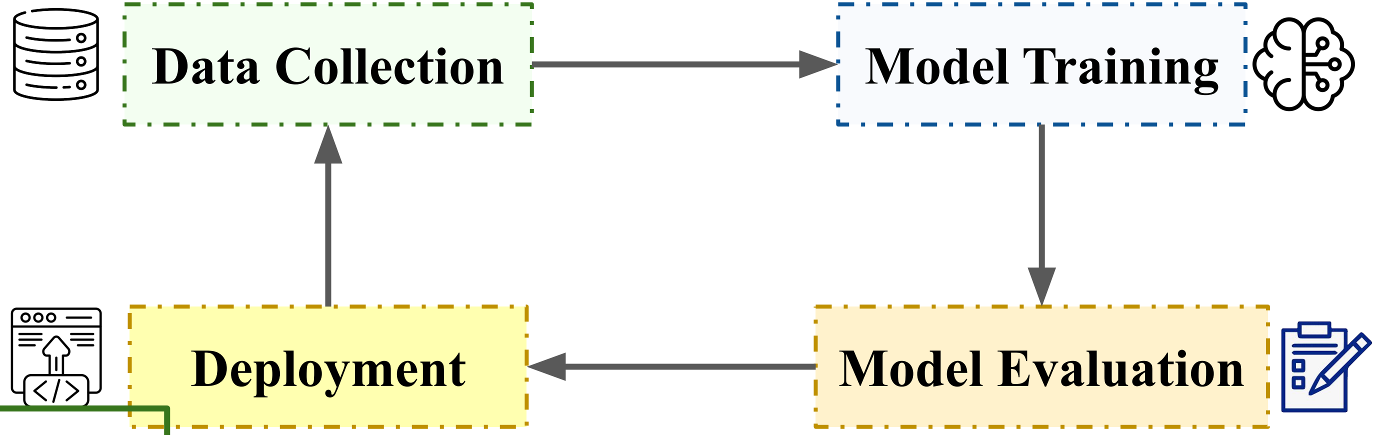
Figure from Eyuboglu, S.,et al. http://ai.stanford.edu/blog/domino/

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

Understanding heterogeneity throughout the modeling process



Understand where and why model fails to generalize

Perspective 3: it's important to understand **why** your model performs poorly *across a distribution shift*

Train → Target e.g. deployment

P → **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!

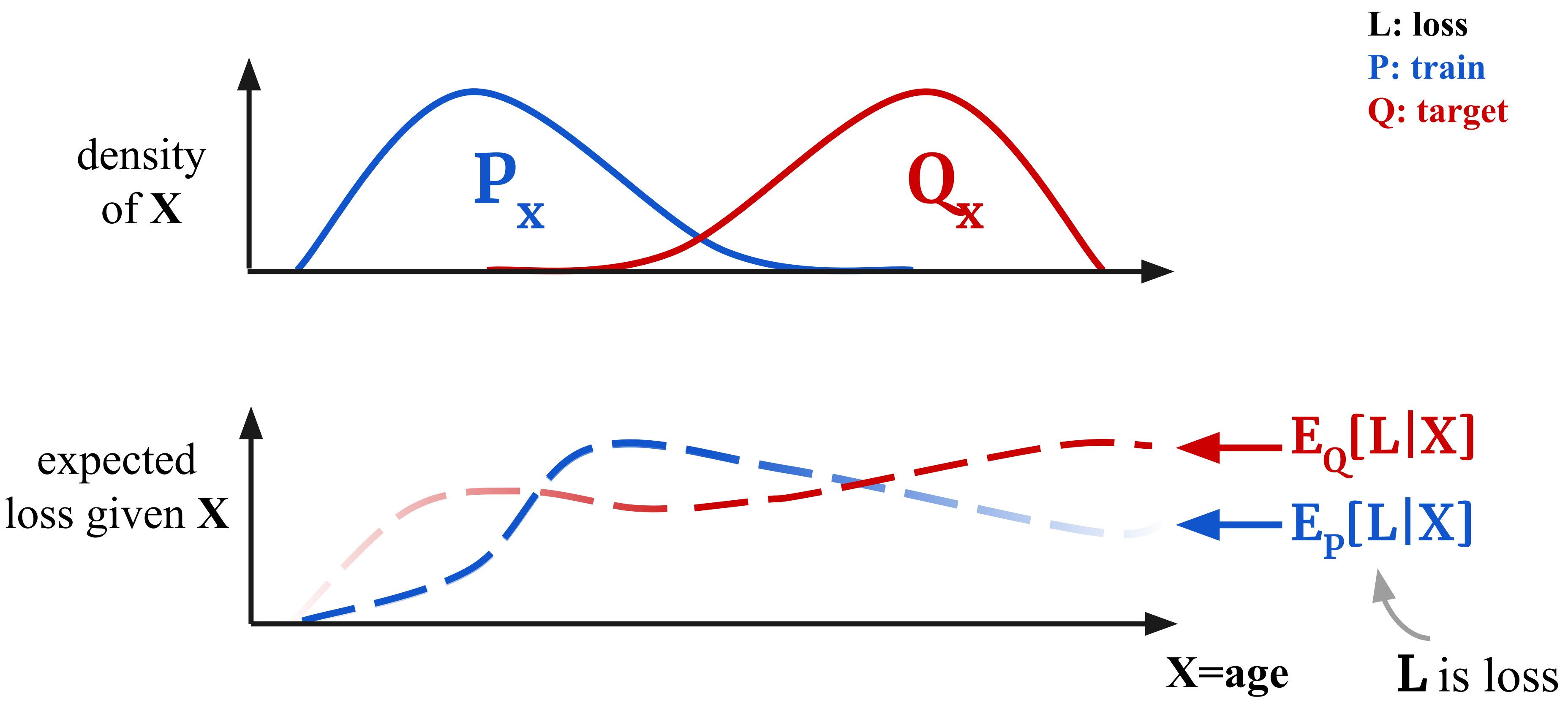
Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

Attribute change in performance to distribution shifts

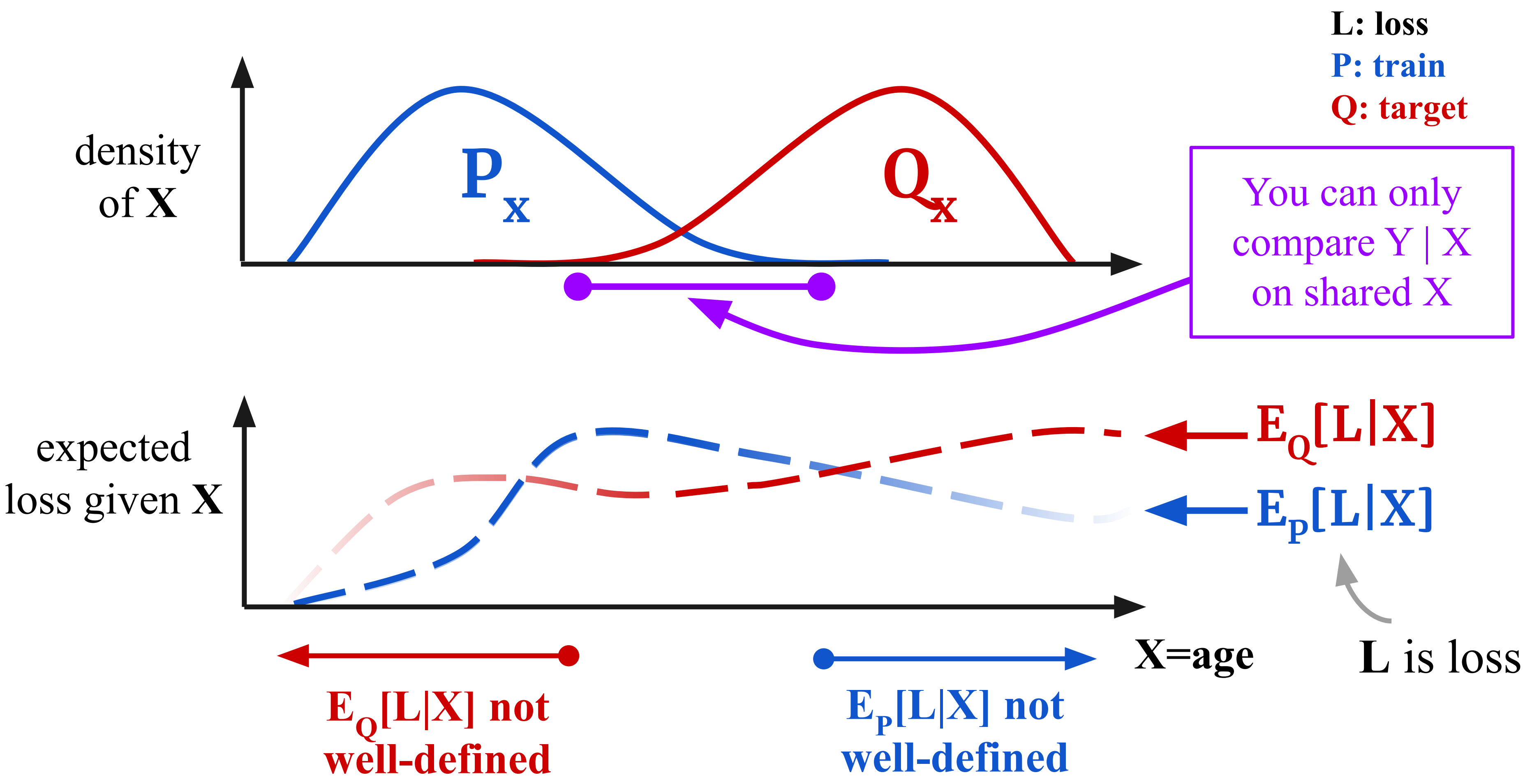
X shifts	**Y | X shifts**
changes in sampling, population shifts, subpopulations	changes in labeling or mechanism, poorly chosen X

- 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



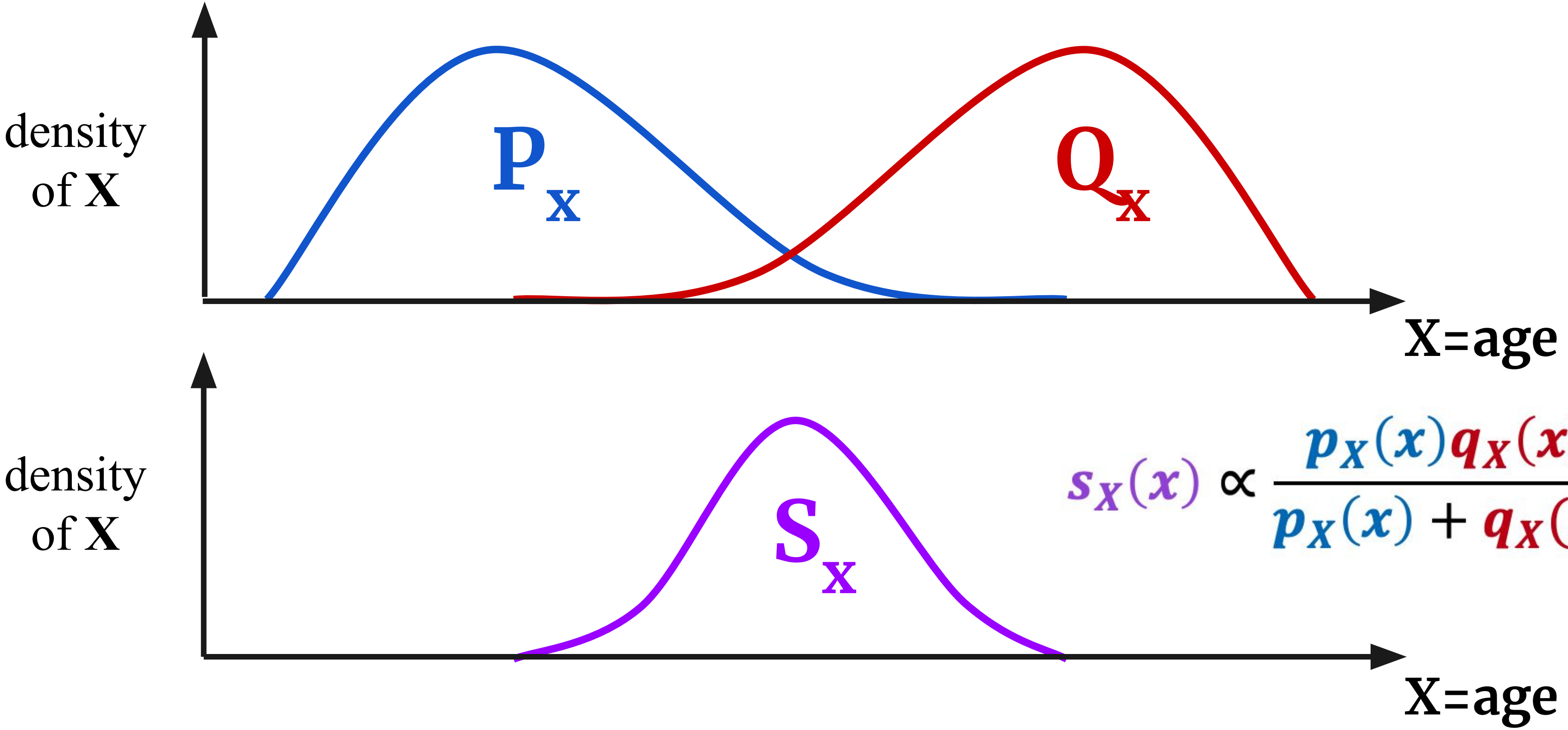
Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011



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

Define **Shared Distribution**

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



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

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

Attribute change in performance to distribution shifts

$$E_P[E_P[L|X]] \qquad\qquad E_Q[E_Q[L|X]]$$

Performance on the training distribution

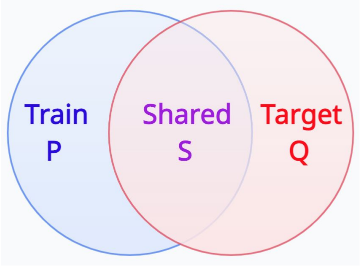
Performance on the target distribution

Decompose into X-shift vs. Y|X-shift

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

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



$$\mathbf{E_P[E_P[L|X]]} \xrightarrow{X \text{ shift } (P \to S)} \mathbf{E_S[E_P[L|X]]}$$

Diagnosis:

S has more X's that are harder to predict than P

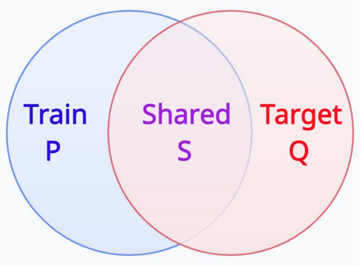
Potential interventions:

Use domain adaptation, e.g. importance weighting

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

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



Diagnosis:

$Y \mid X$ moves farther from predicted model

Potential interventions:

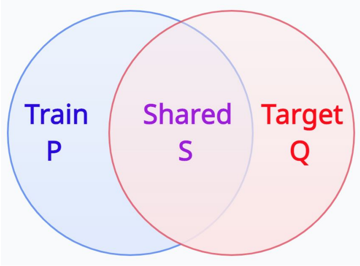
Re-collect data
or modify covariates

$$\mathbf{E_S[E_P[L|X]]} \xrightarrow{Y \mid X \text{ shift}} \mathbf{E_S[E_Q[L|X]]}$$

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

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



Diagnosis:

Q has "new" X's that are harder to predict than S

Potential interventions:

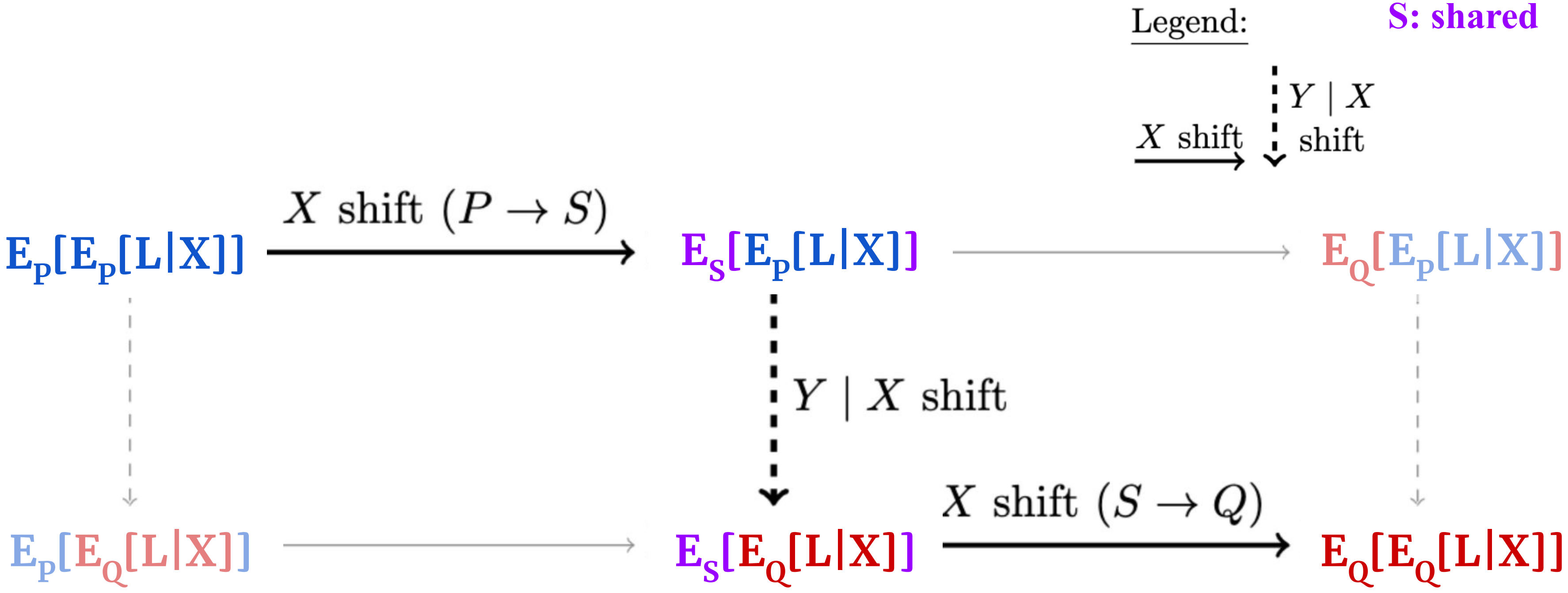
Collect + label more data on "new" examples

$$\mathbf{E_S[E_Q[L|X]]} \xrightarrow{X \text{ shift } (S \to Q)} \mathbf{E_Q[E_Q[L|X]]}$$

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

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

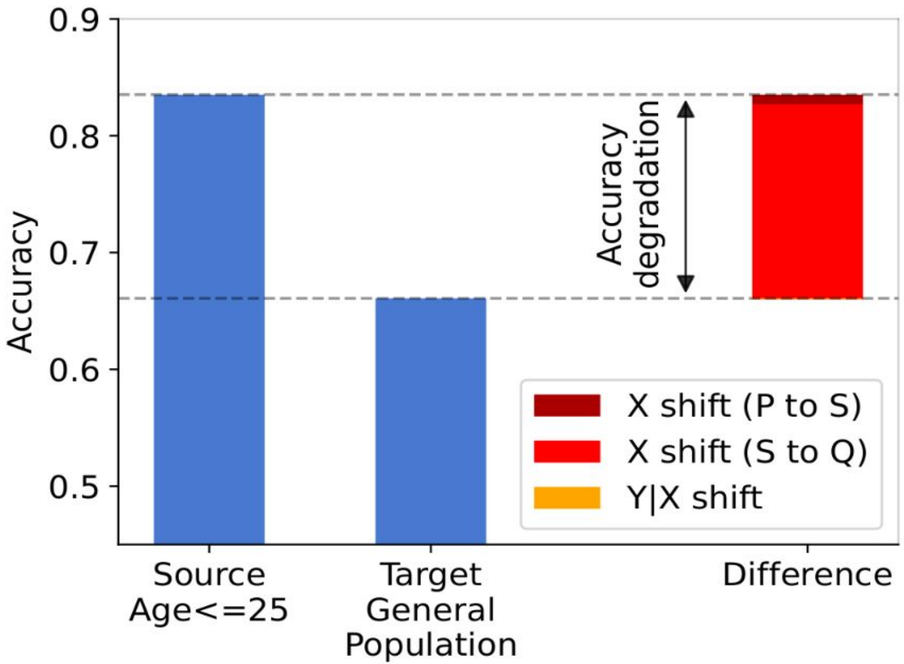


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

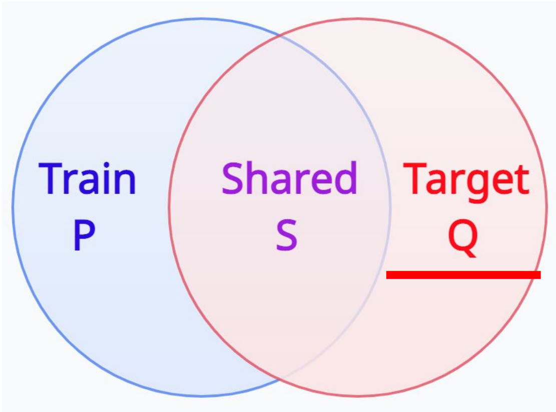
L: loss
P: train
Q: target
S: shared

Employment prediction case study

[X shift] P: only age ≤25, Q: general population



Performance attributed to X shift (S⟶Q), meaning "new examples" such as older people

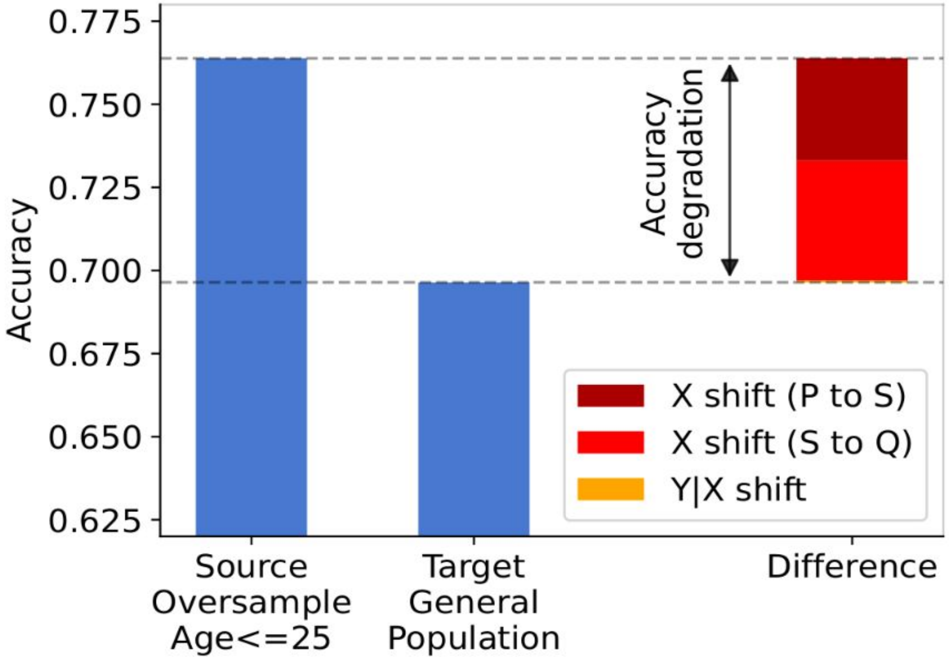


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

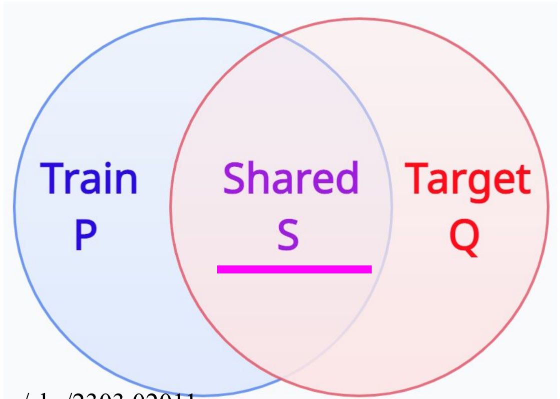
Employment prediction case study

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

[X shift] P: age ≤25 overrepresented, Q: evenly-sampled population



Substantial portion attributed to X shift ($P \longrightarrow S$), suggesting domain adaptation may be effective

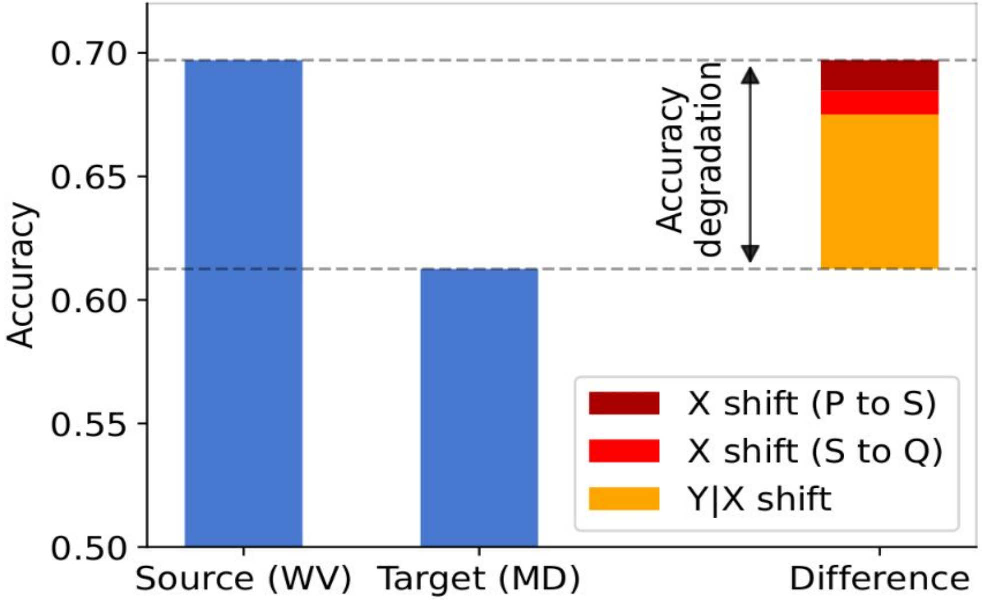


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

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

Employment prediction case study

[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

Recap

- Diagnostic for understanding **why** performance dropped, in terms of X vs Y|X shift
- Diagnostic can be used to help decide on modeling assumptions + data collection

Where to go next?

- Limitations of this diagnostic
  - Shared space not easy to understand / interpret in high dimensions
- Lots of unanswered questions!
  - We’re only diagnosing between X vs Y|X shift! This is a bare minimum.
  - In practical settings, need more fine-grained actionable insights

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)

Perspective 4: 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?

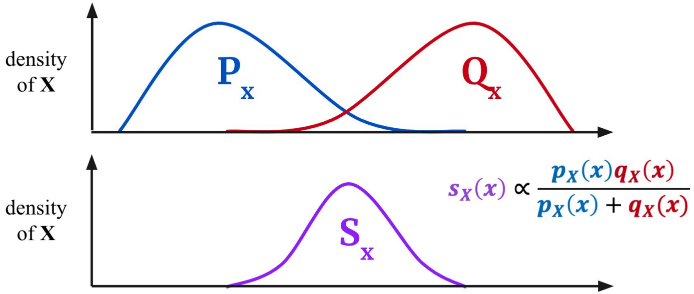
If we understand this, then we can collect data better.

Example: Identify 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

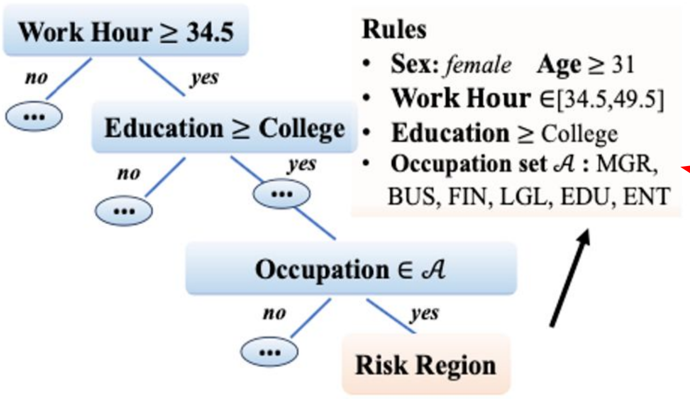


Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Tabular Data

Task: Income Prediction
Shift: CA -> PR



(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!

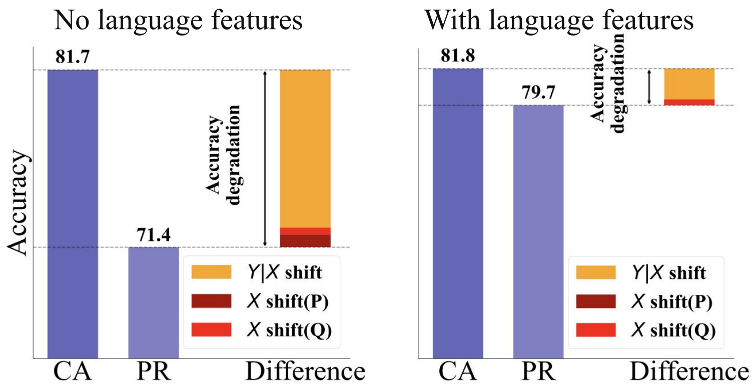
Figure from Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

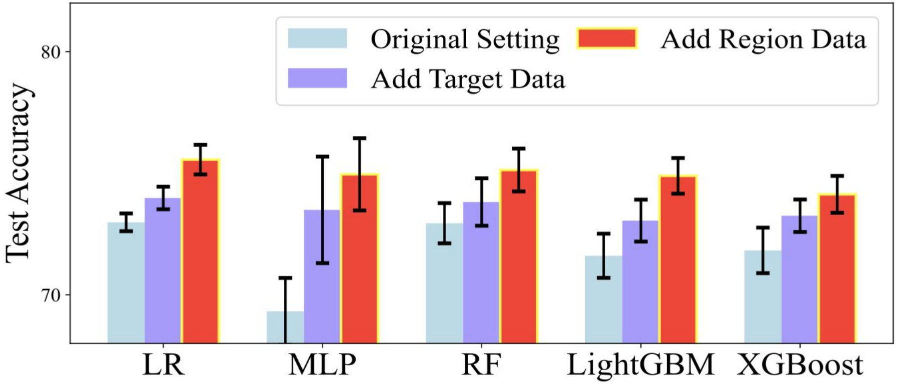
Good data may be **more effective**!

Task: Income Prediction
Shift: CA -> PR

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



collecting better features



collecting better target data

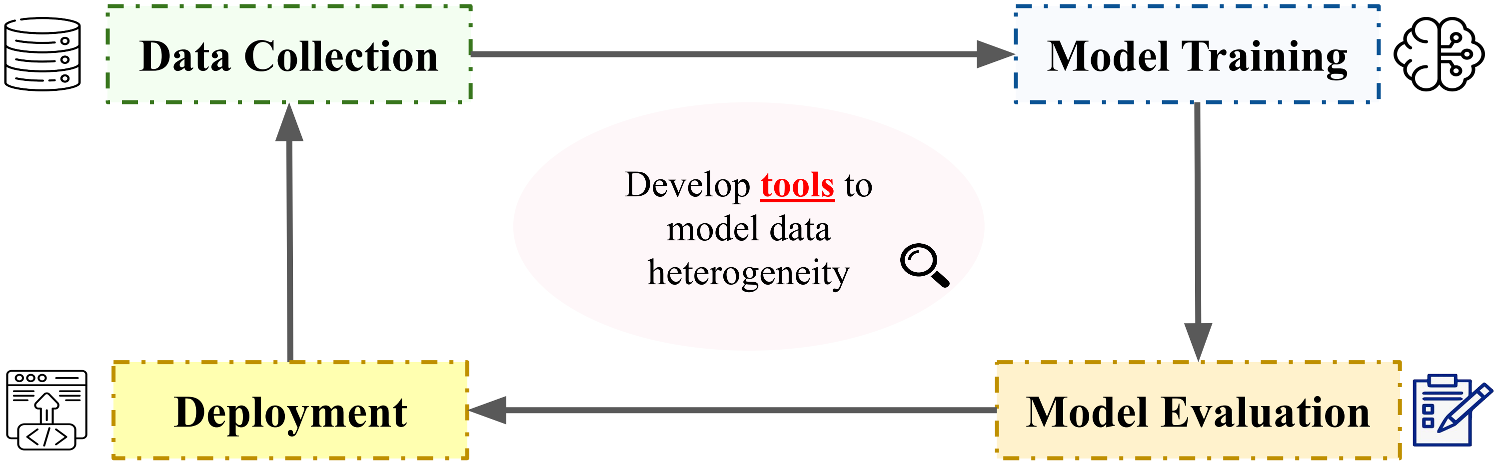
Figure from Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

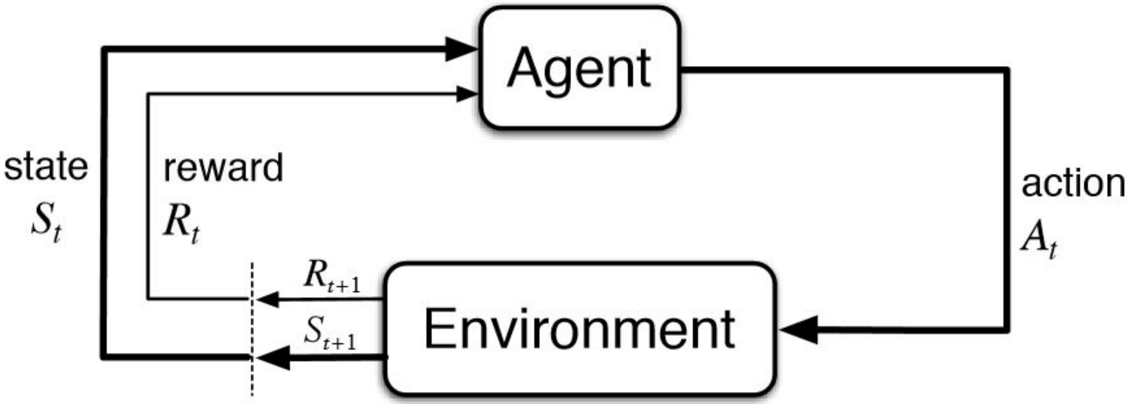


Future directions

- 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

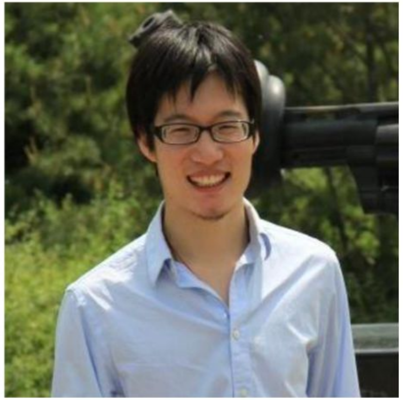


Future directions

- 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!

Contact

tutorial website



Peng Cui

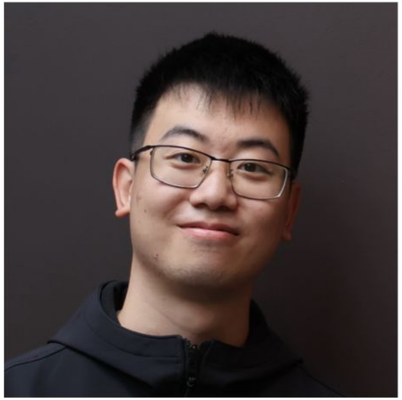
Tsinghua University

cuip@tsinghua.edu.cn

Hongseok Namkoong

Columbia University

namkoong@gsb.columbia.edu



Jiashuo Liu

Tsinghua University

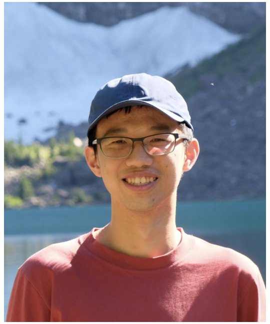
liujiashuo77@gmail.com

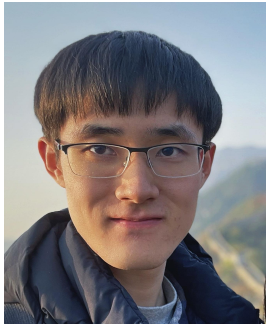
Tiffany (Tianhui) Cai

Columbia University

tiffany.cai@columbia.edu

Many thanks to





Tianyu Wang
Columbia University
tw2837@columbia.edu

Jiayun Wu
Tsinghua University
jiayun.wu.work@gmail.com

Zimu Wang
Tsinghua University
14317593@qq.com

Steve Yadlowsky
Google Deepmind

Welcome our Panelists!

[Shalmali Joshi](#)

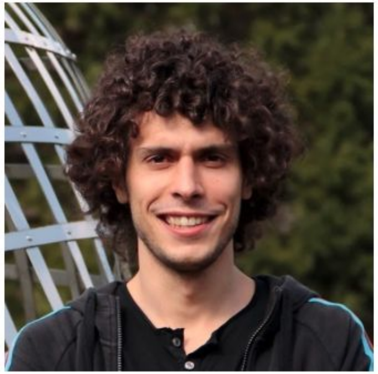
Columbia University

[Aditi Raghunathan](#)

Carnegie Mellon University

[Sara Beery](#)

MIT



[Dominik Rothenhäusler](#)

Stanford University