Week 2: Language for Distribution Shifts

Feb 6, 2025

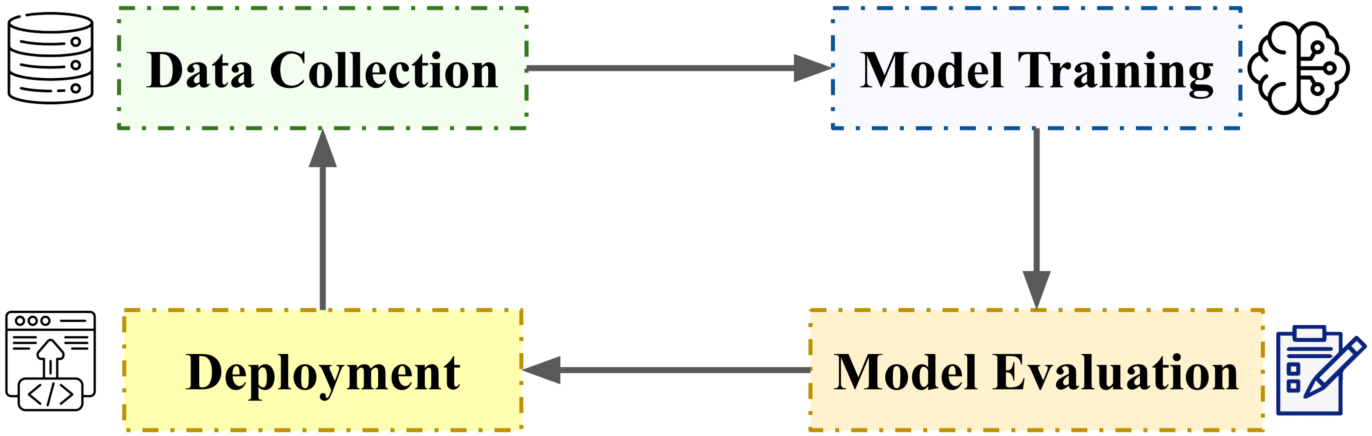
Thoughtful use of AI is challenging

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

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: Inductive modeling language for distribution shifts

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

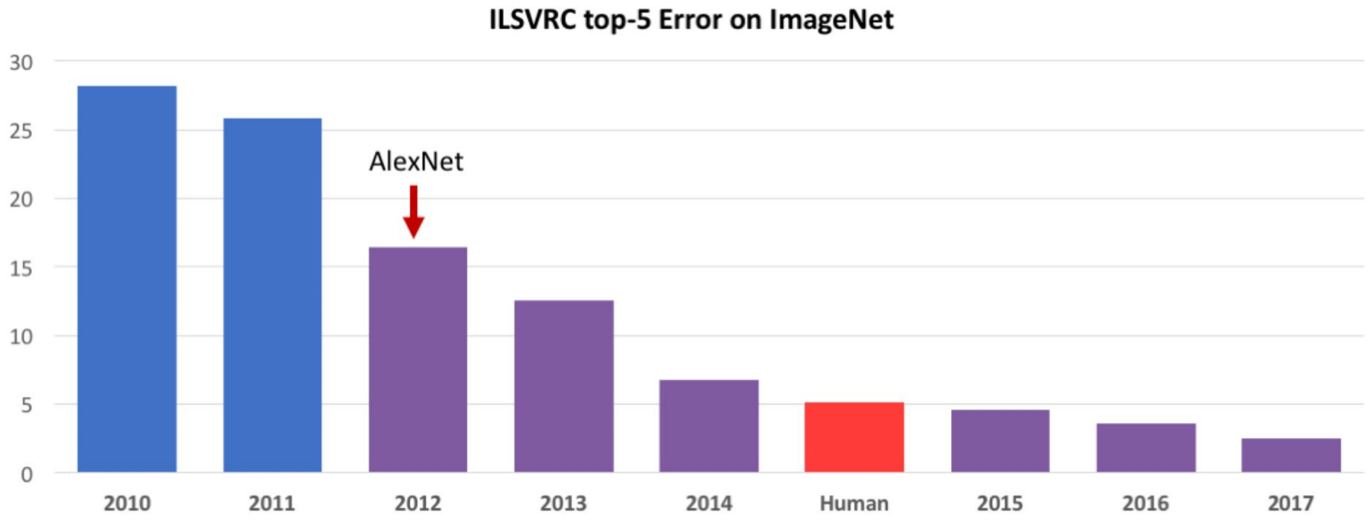
Great white shark

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]

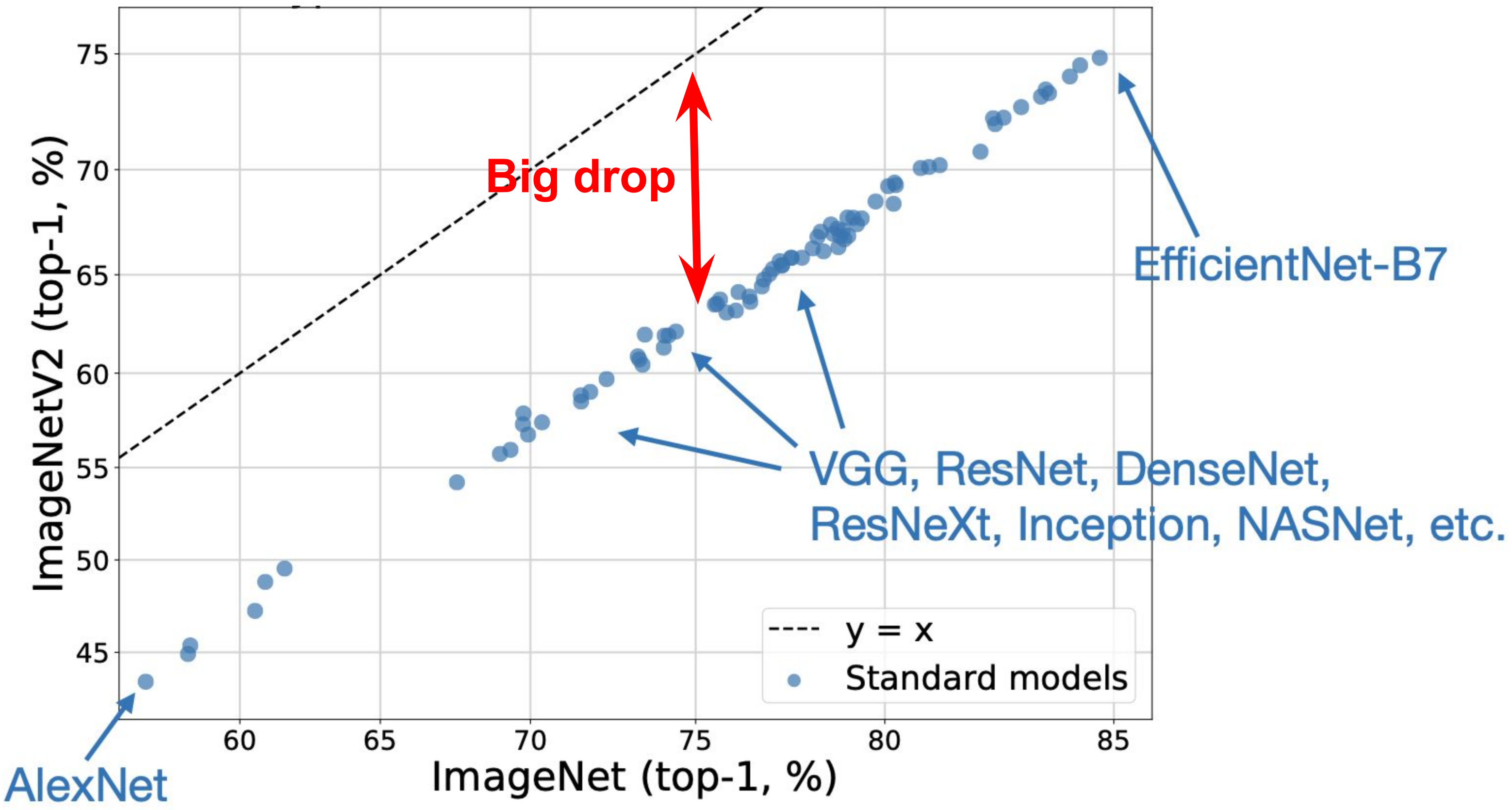
ImageNet-Sketch

[Wang, Ge, Lipton, Xing '19]

ImageNet-R

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

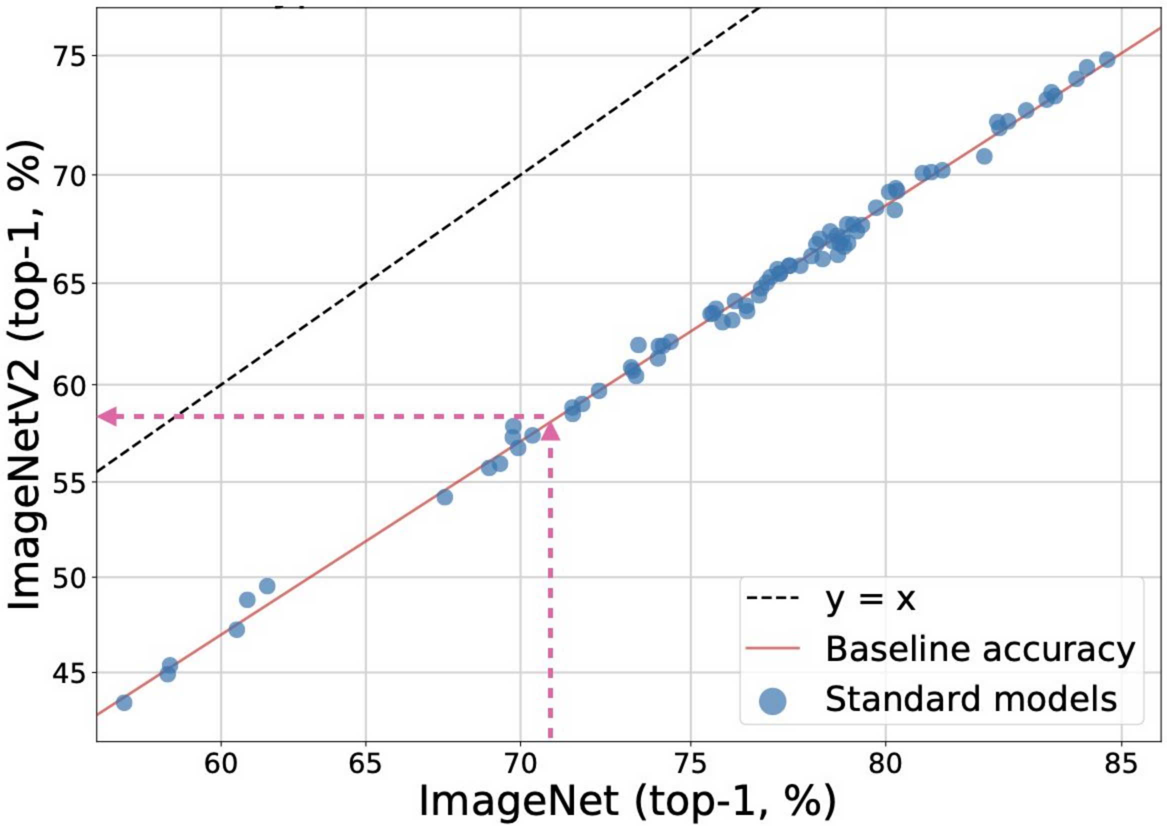
Slide credit: Ludwig Schmidt



[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.**

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

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

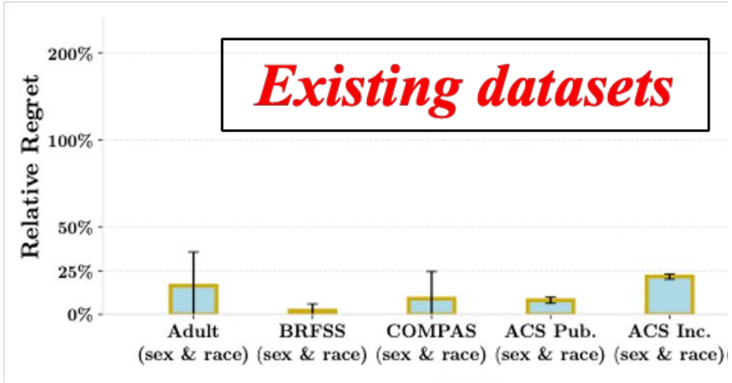
- So far: Humans are robust on all distributions. Can we get a universally good model?
- Implicitly, this view focuses on covariate shift (X-shift)
  - Traditional focus of ML
- On the other hand, we expect $Y|X$-shifts when there are unobserved factors
  - Traditional focus of causal inference
- For $Y|X$-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

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

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 $\rightarrow$ Puerto Rico	$Y\|X \gg X$
`ACS Mobility`	Mississippi $\rightarrow$ Hawaii	$Y\|X \gg X$
`Taxi`	New York City$\rightarrow$ Botogá	$Y\|X \gg X$
`ACS Pub.Cov`	Nebraska $\rightarrow$ Louisiana	$Y\|X > X$
`US Accident`	California$\rightarrow$ Oregon	$Y\|X > X$
`ACS Pub.Cov`	2010 (NY)$\rightarrow$ 2017 (NY)	$Y\|X < X$
`ACS Income`	Younger$\rightarrow$ 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.

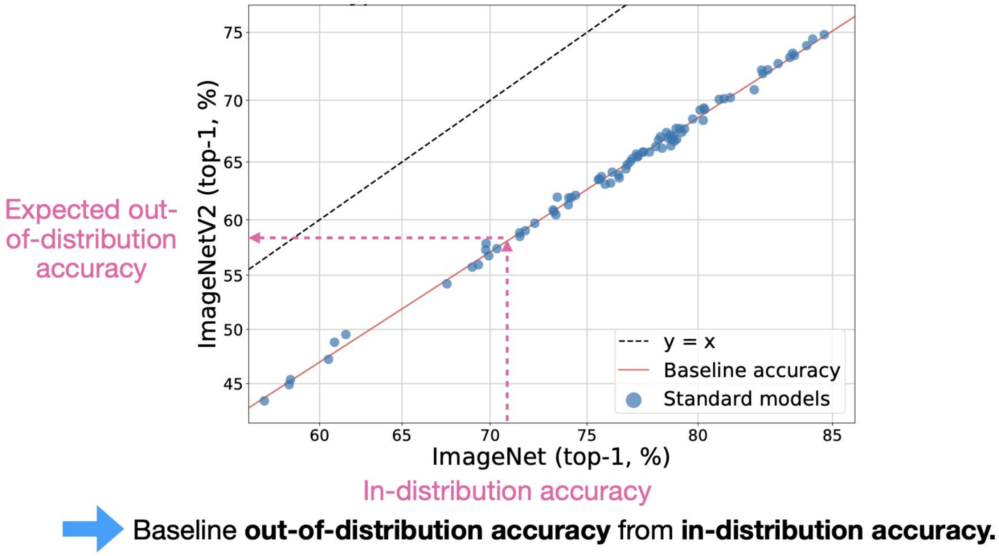
https://github.com/namkoong-lab/whyshift

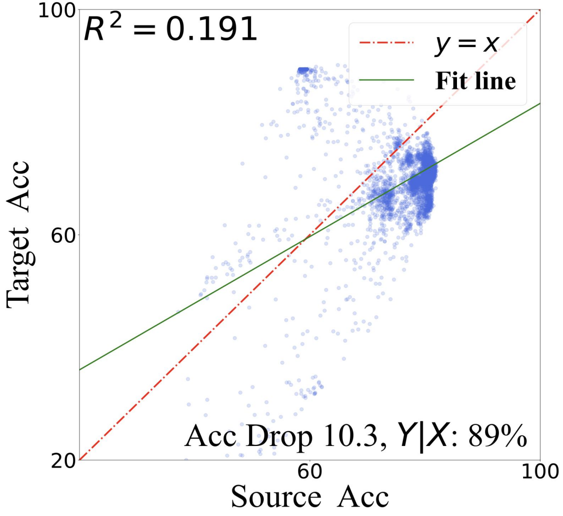
$Y|X$-shifts

- 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***





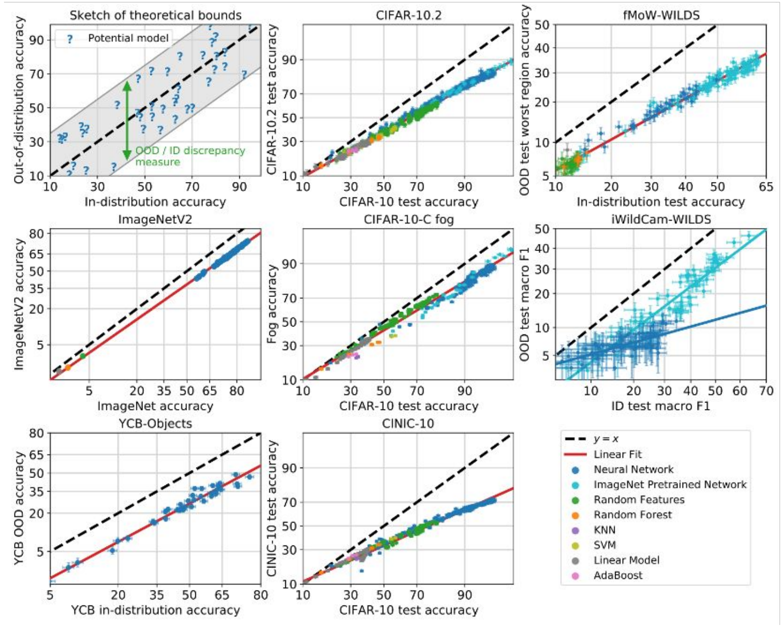
ACS Income (CA→ PR)

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

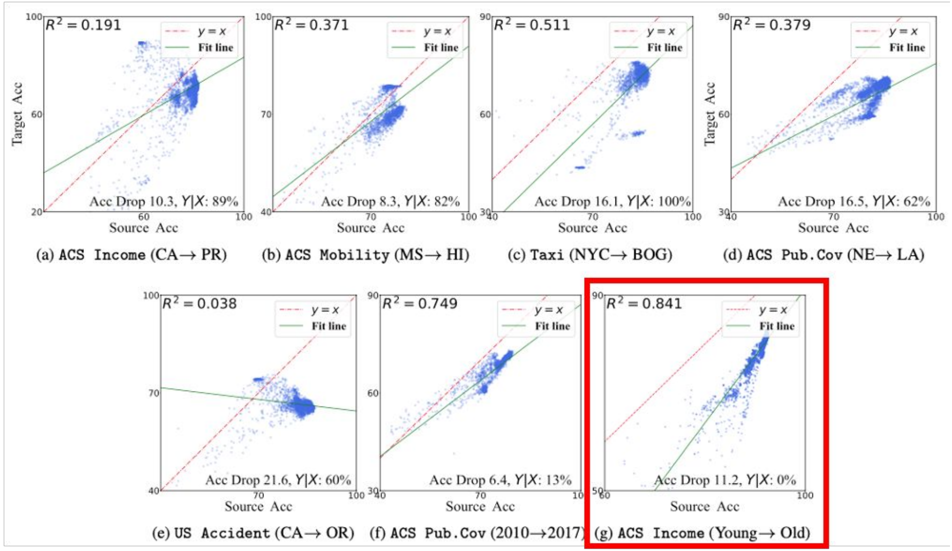
Accuracy-on-the-line **doesn't** hold under strong $\boldsymbol{Y|X}$-shifts

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

Image datasets



WHYSHIFT



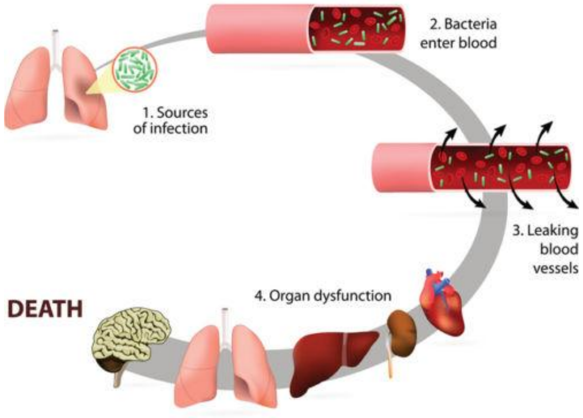
Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization.

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

Outline

Part 1: Benchmarking performance under distribution shift

Part 2: A critical review of existing approaches

Part 3: Inductive modeling language for distribution shifts

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)

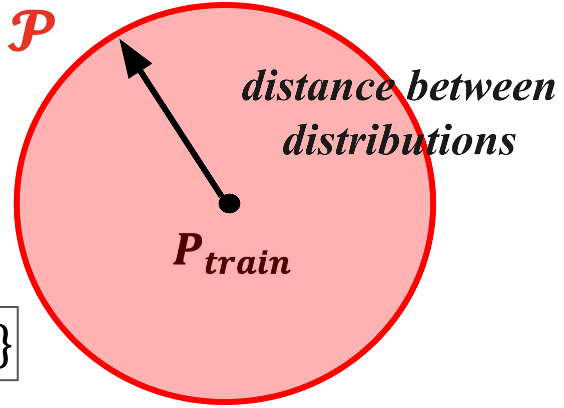
Empirical Risk Minimization

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

DRO

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

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



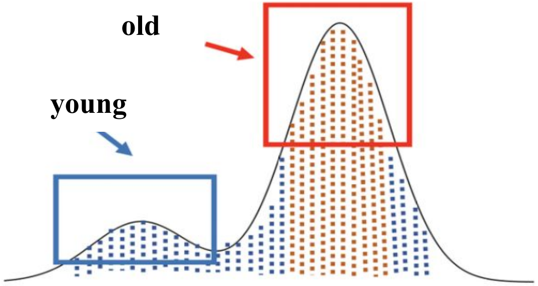
Instead of minimizing loss over training distribution, minimize loss over distributions ***near*** it

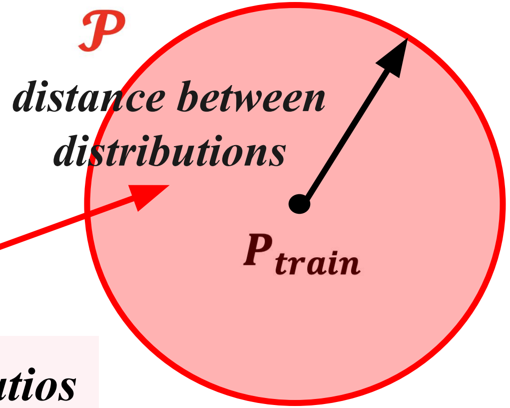
Distributionally Robust Optimization (DRO)

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)

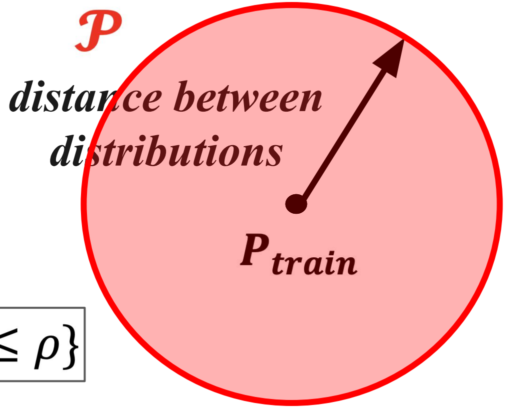
Empirical Risk Minimization

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

DRO

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

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



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

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

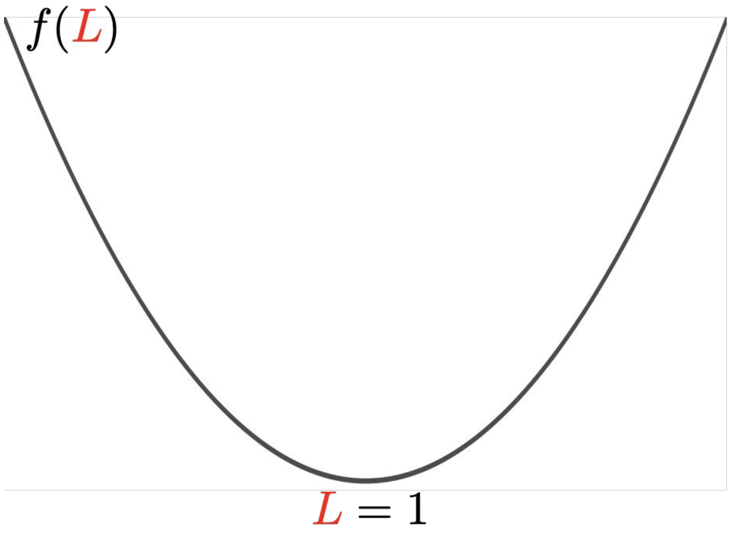
***f*-divergence:** about *densities*

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

For a convex function,

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

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



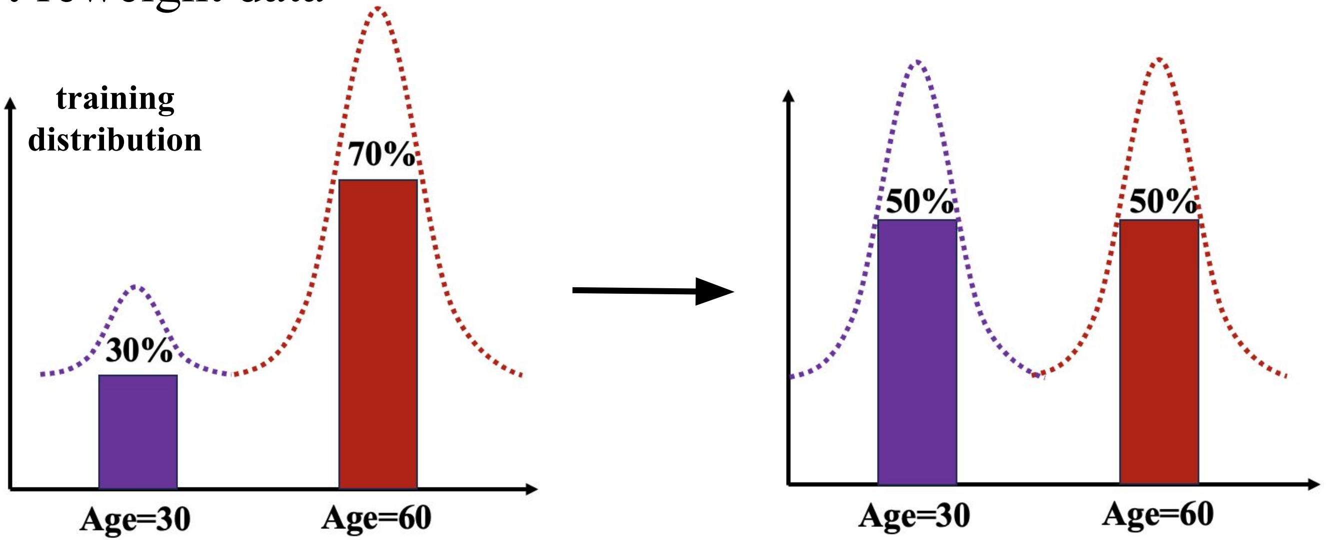
Examples: set of distributions we care about

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

recall the objective

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

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



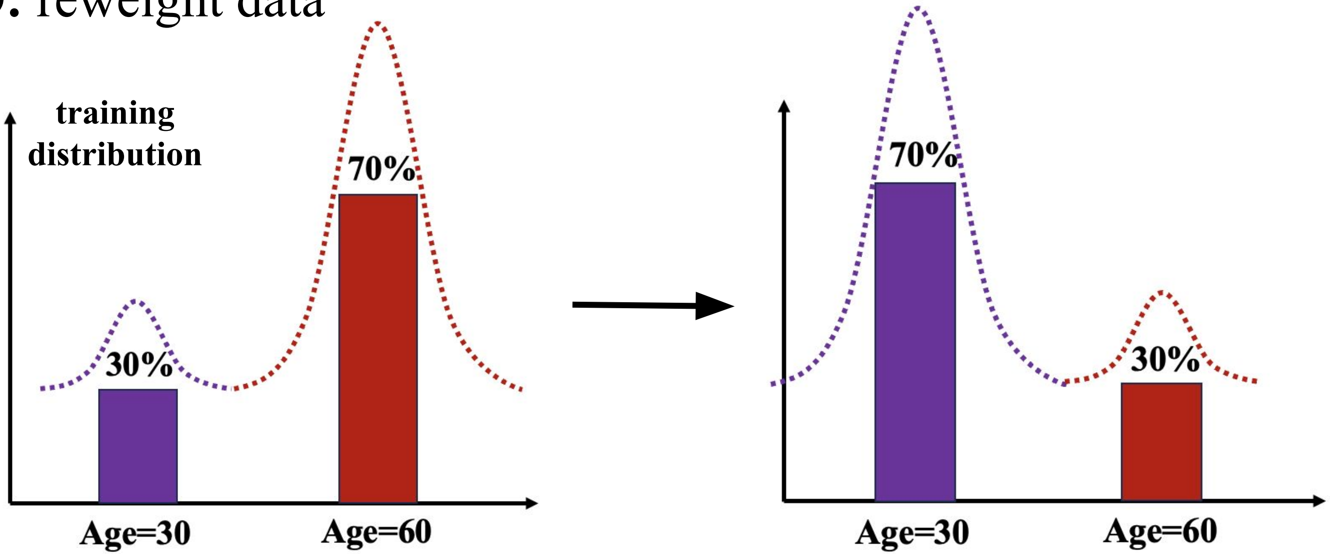
Examples: set of distributions we care about

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

recall the objective

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

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



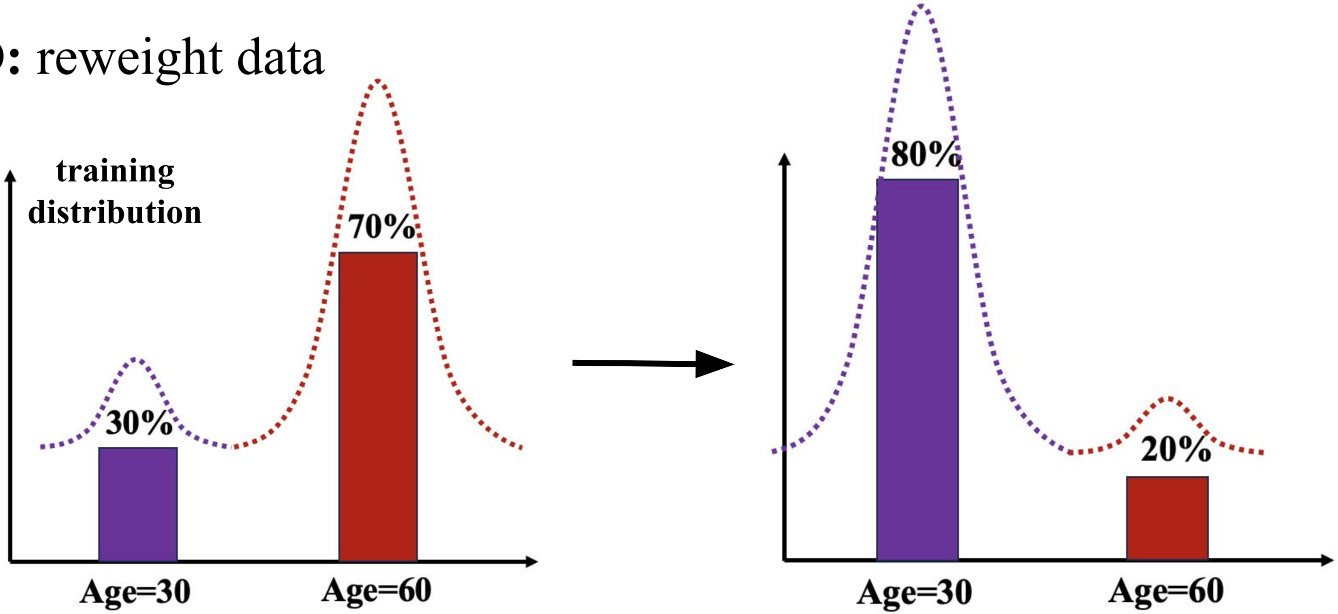
Examples: set of distributions we care about

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

recall the objective

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

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



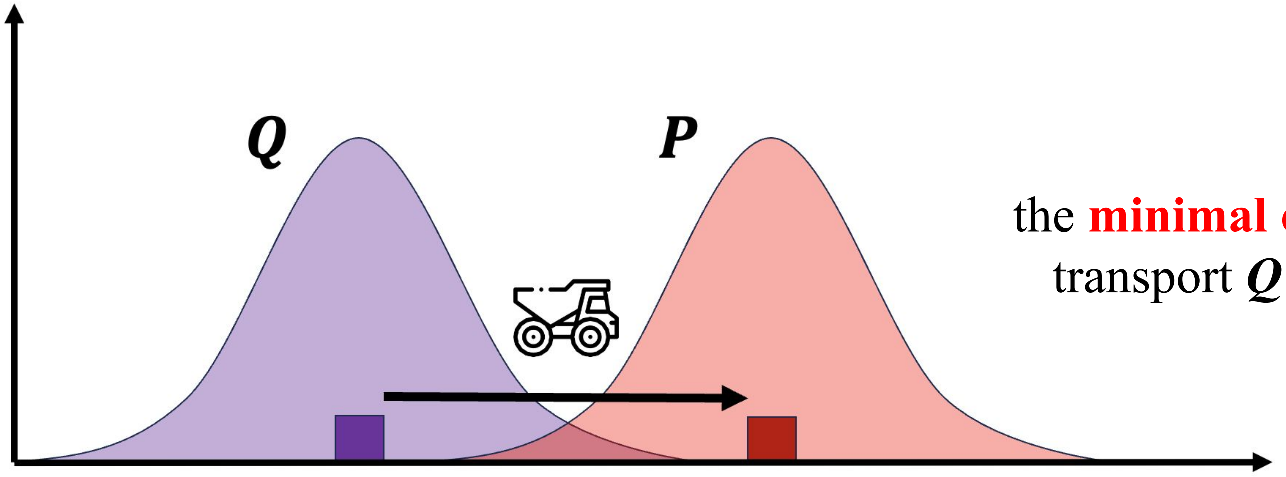
Examples: set of distributions we care about

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

recall the objective

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

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



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

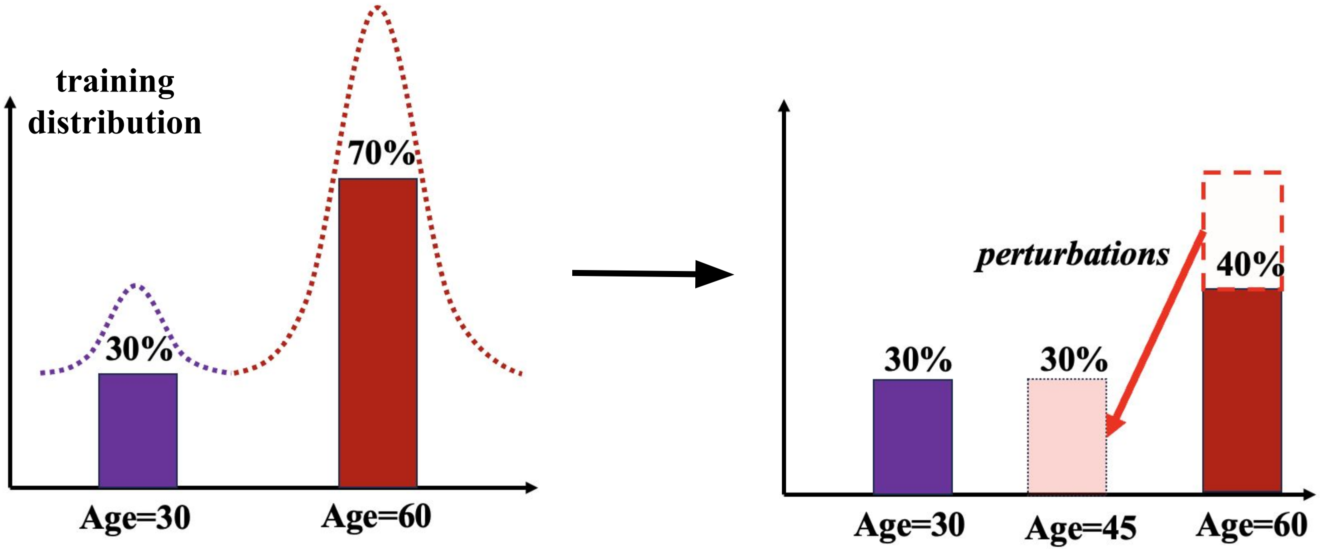
Examples: set of distributions we care about

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

recall the objective

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

Wasserstein-DRO: perturb data



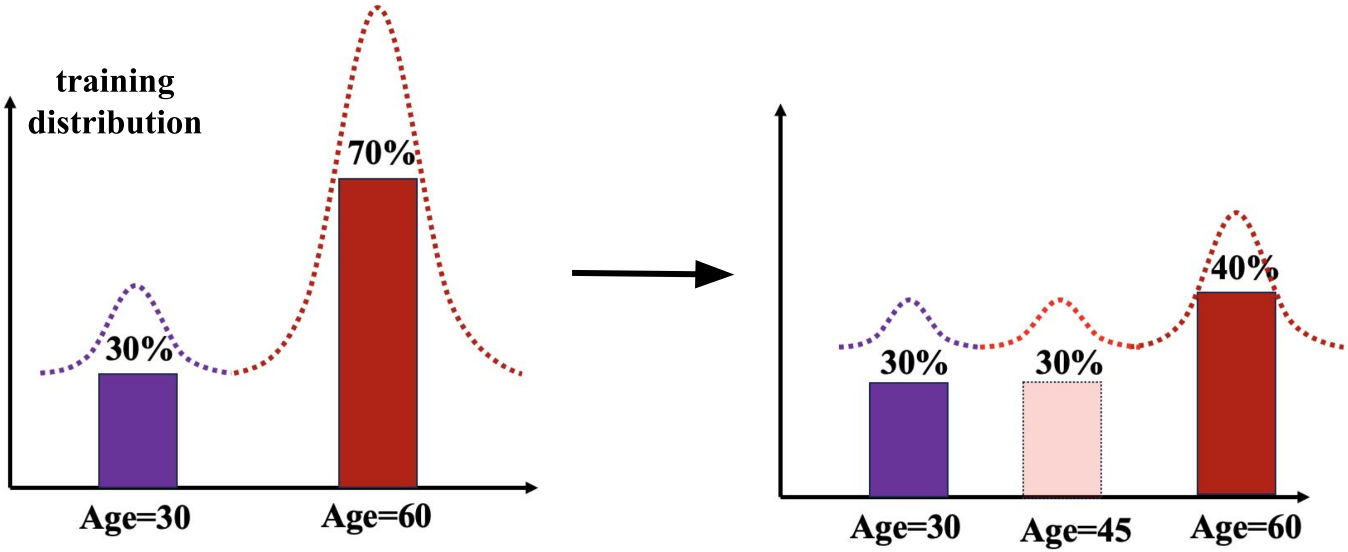
Examples: set of distributions we care about

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

recall the objective

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

Wasserstein-DRO: perturb data



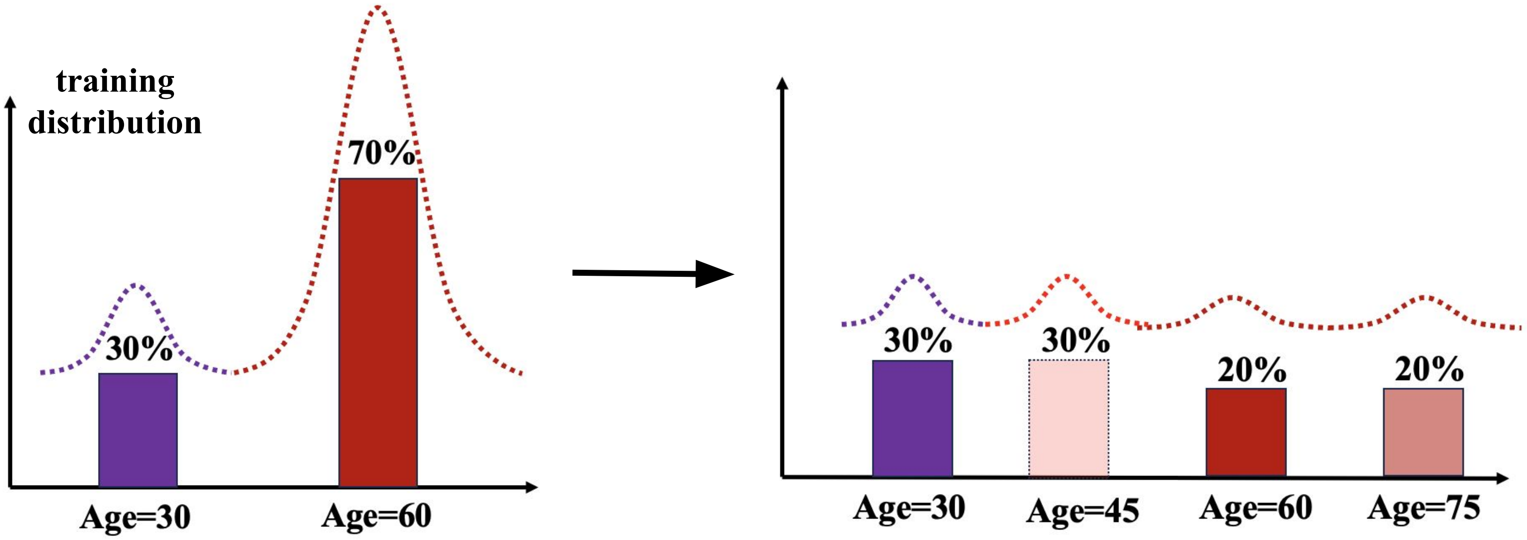
Examples: set of distributions we care about

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

recall the objective

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

Wasserstein-DRO: perturb data

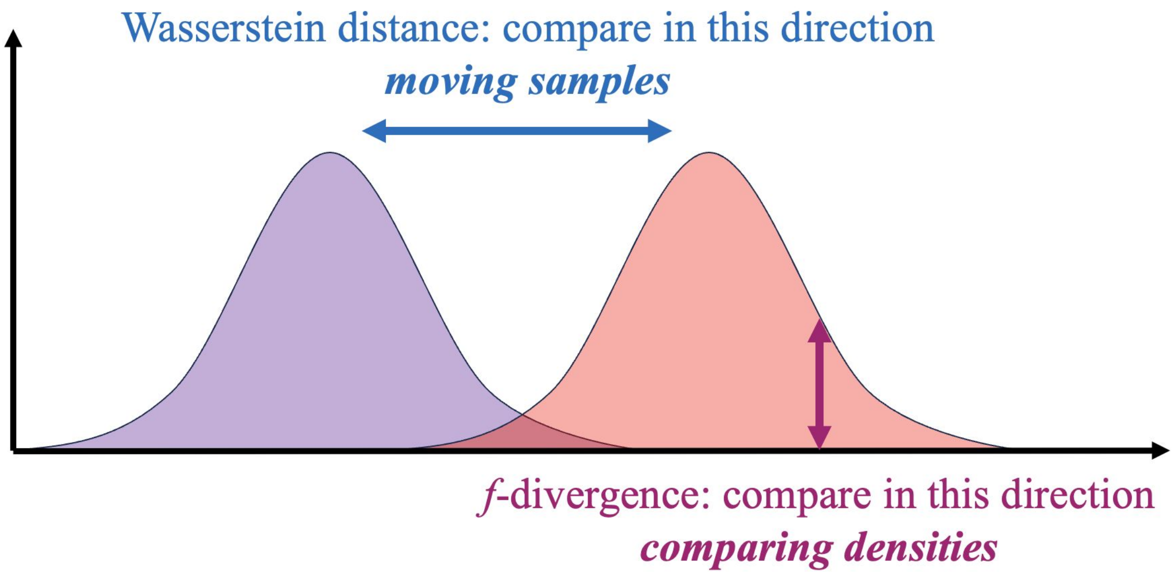


Intuition: *f*-divergence vs Wasserstein distance

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

recall the objective

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



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

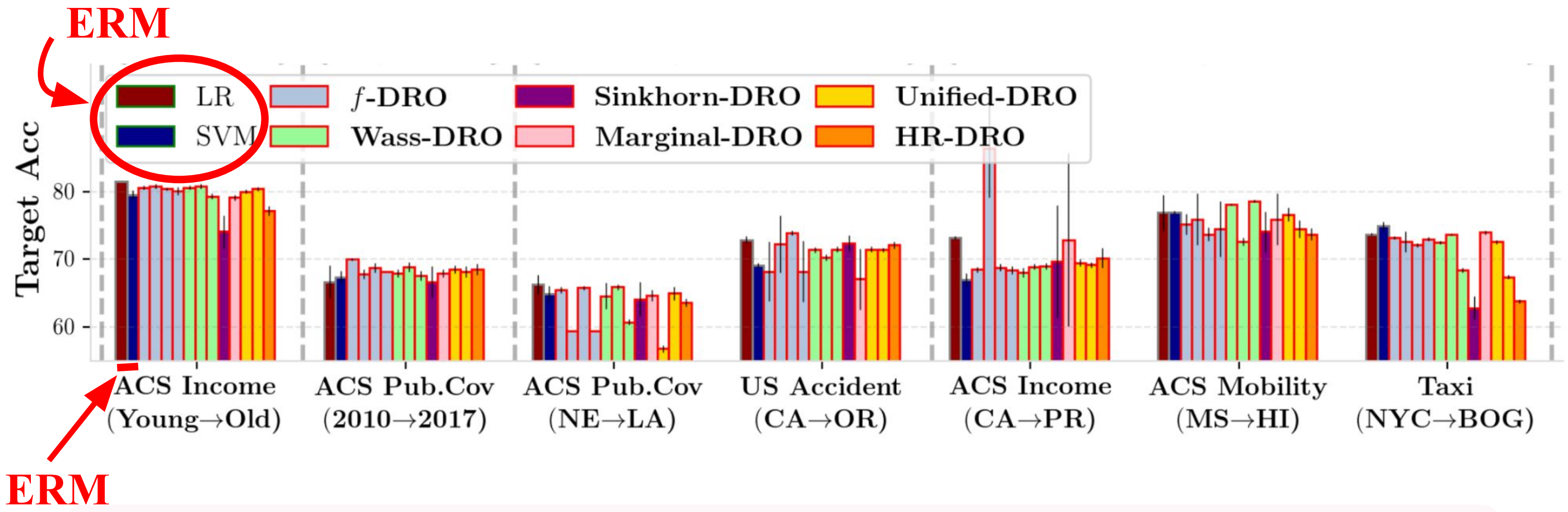
Carefully choose the set $\mathcal{P}$

Goal

Do well on real distribution shifts!

Hope the worst-case distribution captures real shifts

Critical View of DRO: not better than ERM!



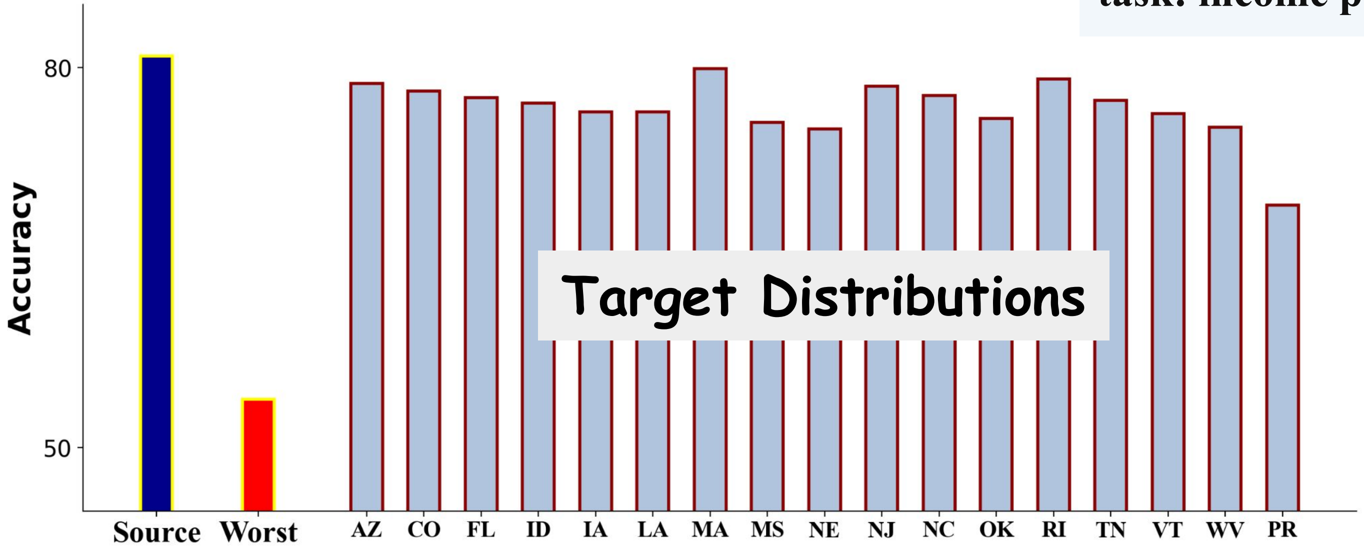
DRO does NOT show significant improvements over ERM!

Hard to choose this set of distributions P!!!

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

Summary

- Overall *philosophy* to algo development is sensible
  - But empirically current methods 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

Two existing approaches to distribution shift

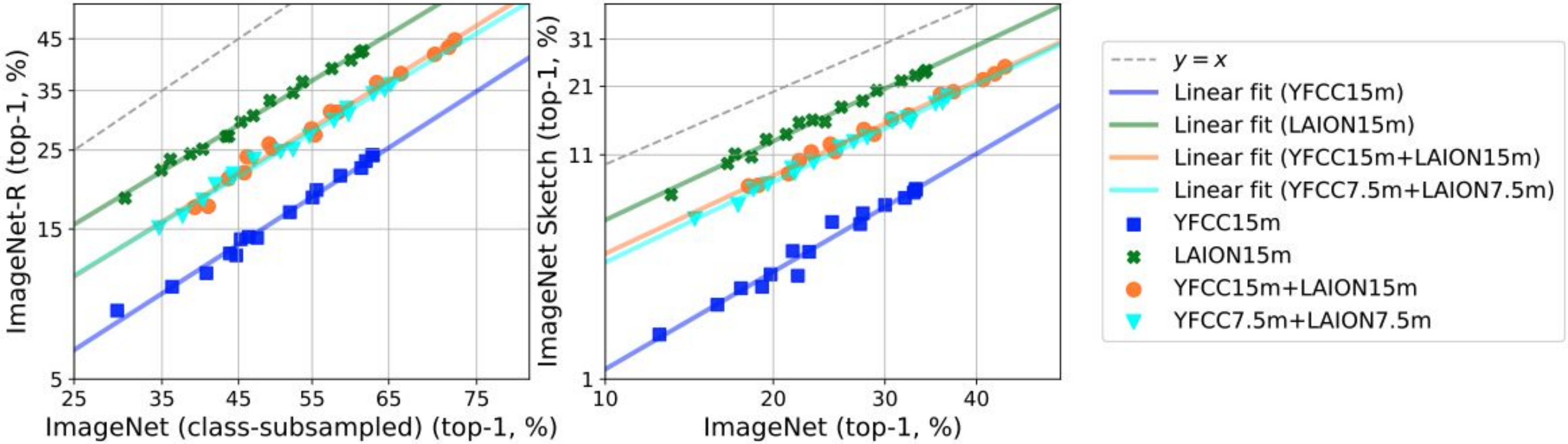
1. Make modeling assumptions

2. **Scale up data and models**

Just adding more data ≠ better

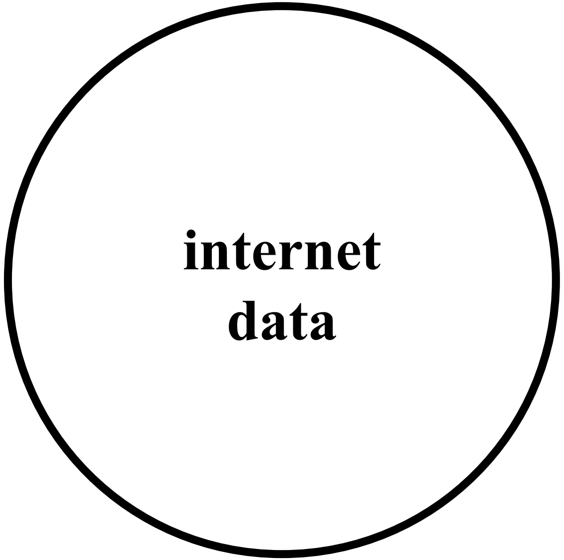
Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP

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

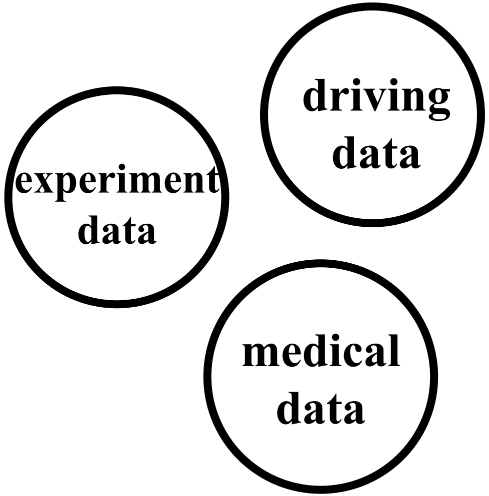


Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP
Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, Ludwig Schmidt

Sometimes you need (costly) specialized data!



$ cheap!



$$$ expensive!

Many important applications!

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

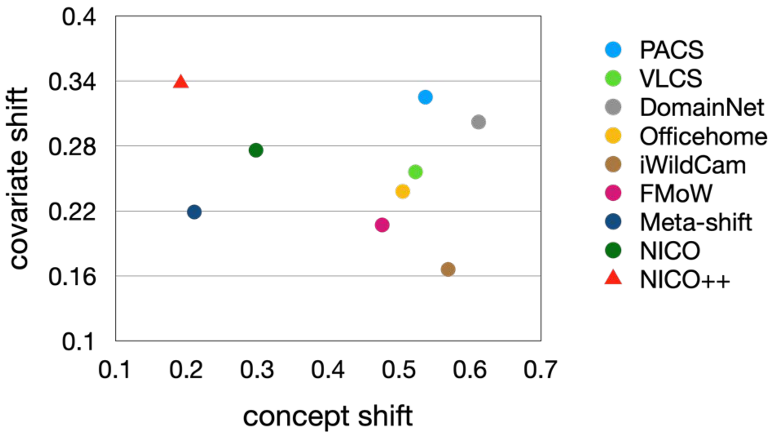
Part 3: Inductive modeling language for distribution shifts

Distribution shifts are complicated in real applications

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

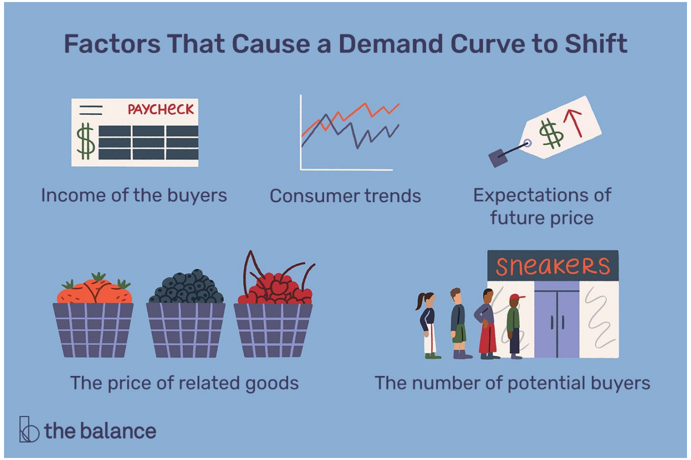


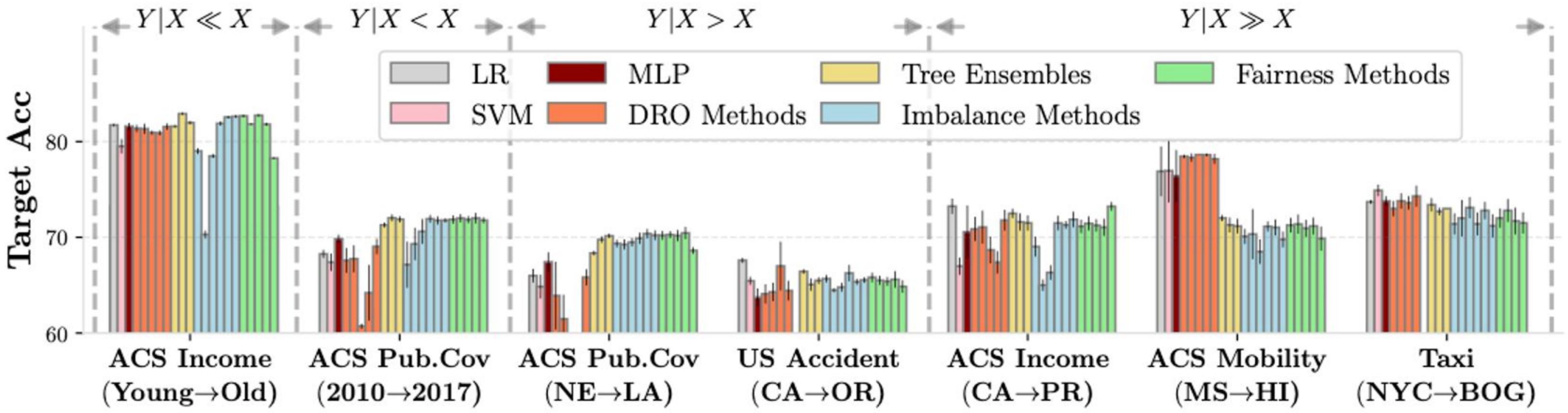
PHOTO: TH BALANCE / MADDY PRICE

Average rent for a 1-bedroom

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

~~One size fits all~~

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



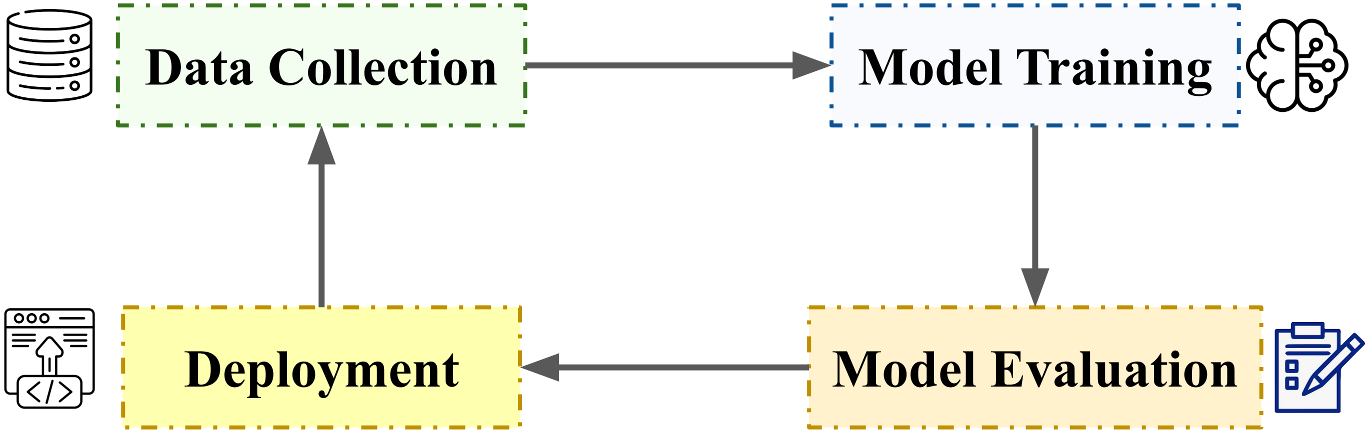
https://github.com/namkoong-lab/whyshift

A different philosophy

- **Model: Application specific** v.s. one model fits all
  - Given an application, first understand its real distribution shift pattern characterized by heterogeneity, and then derive realistic assumptions accordingly for the subsequent modeling process
- **Data: Concerted data collection** v.s. more the better
  - 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.

Understanding heterogeneity throughout the modeling process

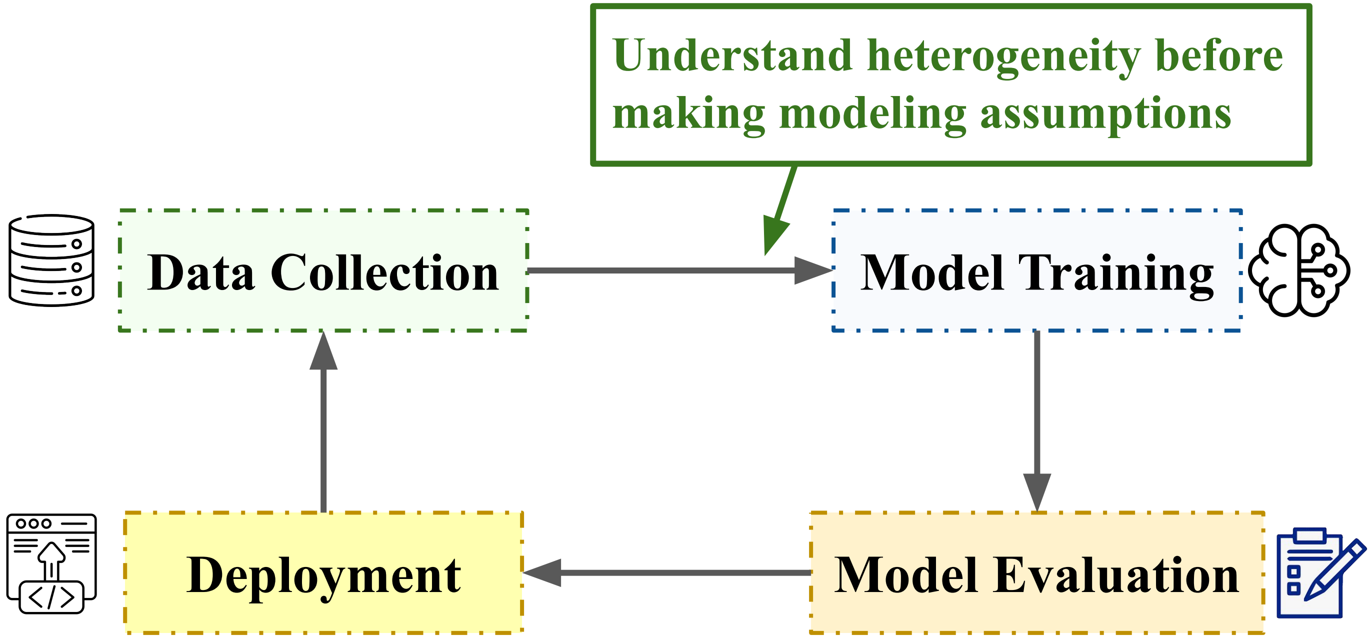
We discuss how understanding heterogeneity can be important throughout the modeling process



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



Understand heterogeneous subpopulations

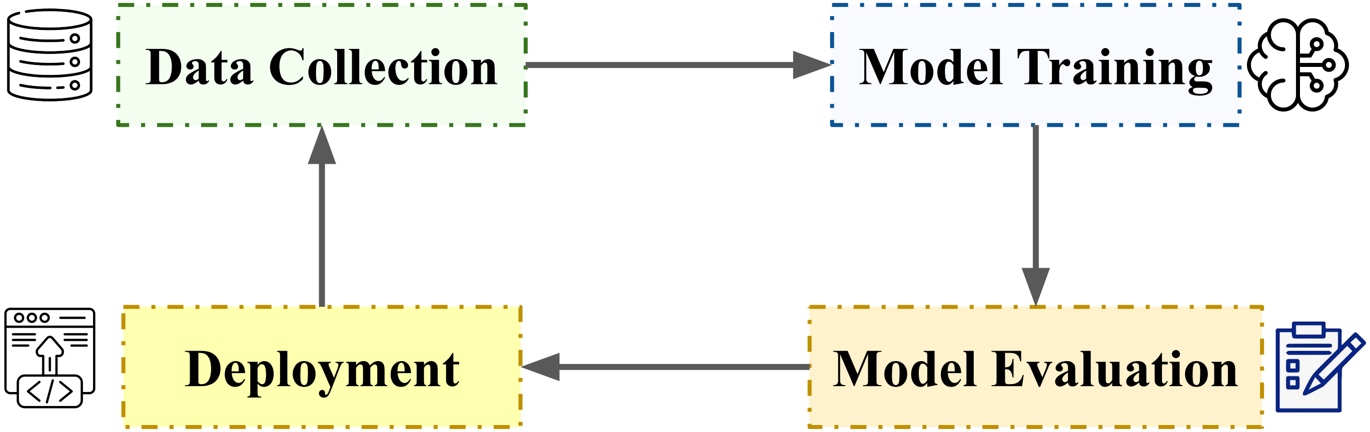
After collecting data, we **need** to know

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.

Understanding heterogeneity throughout the modeling process



Understand important subsets of training data

Understand where your model performs poorly

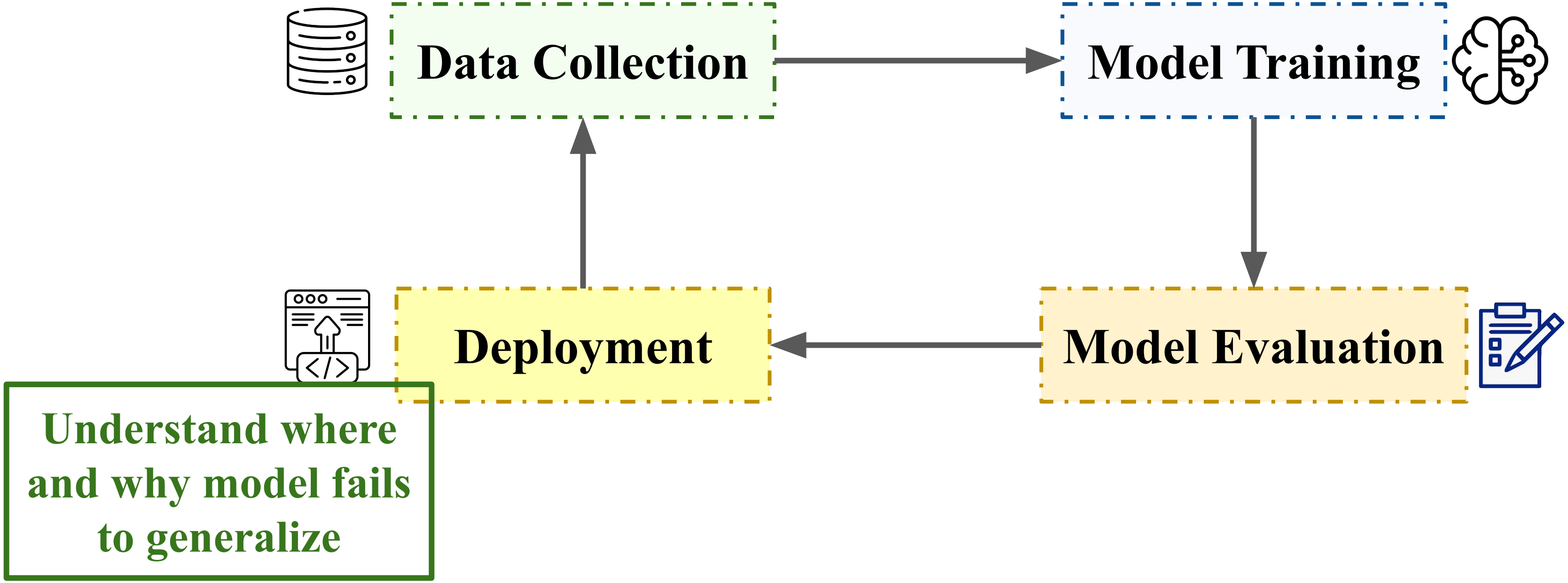
After training a model, we **need** to know

On what 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

Understanding heterogeneity throughout the modeling process



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!

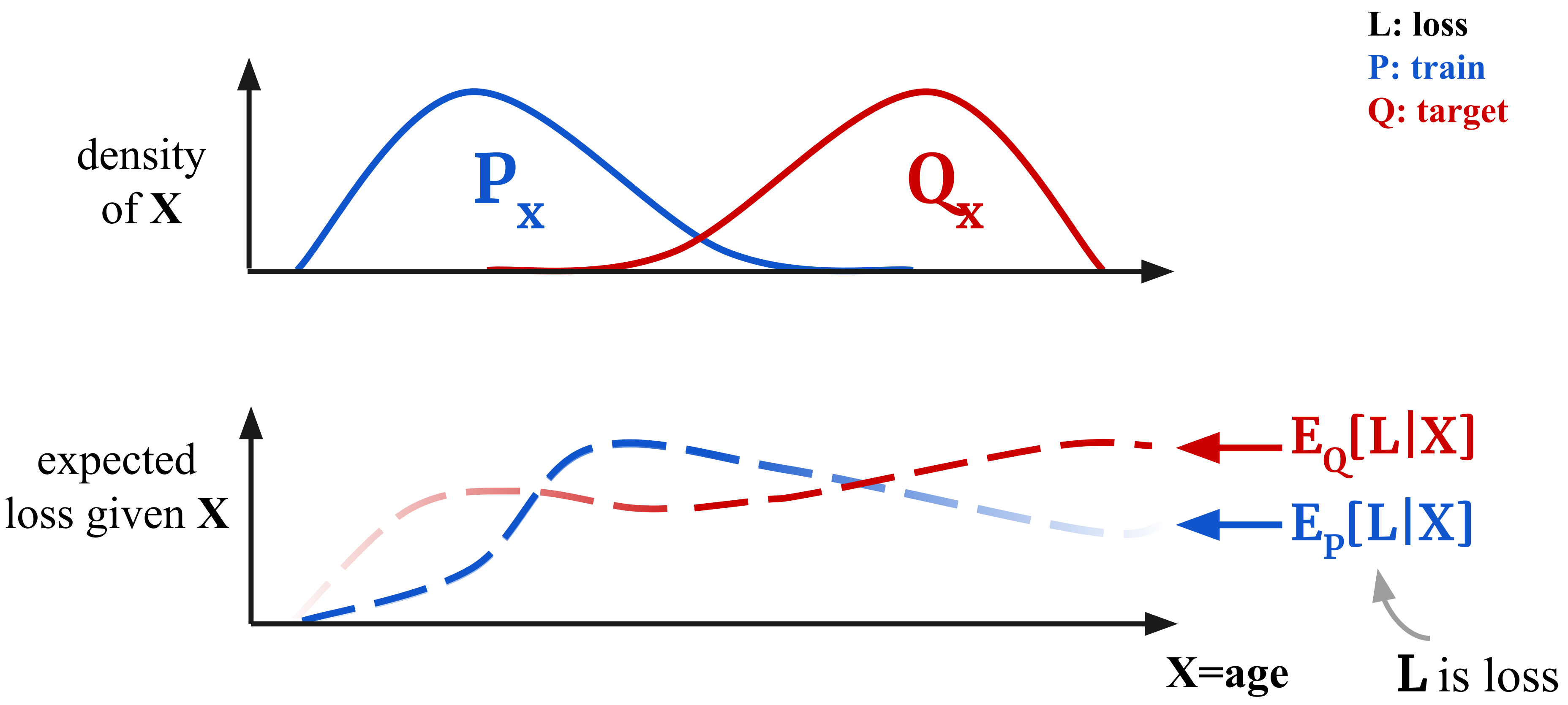
Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

Attribute change in performance to distribution shifts

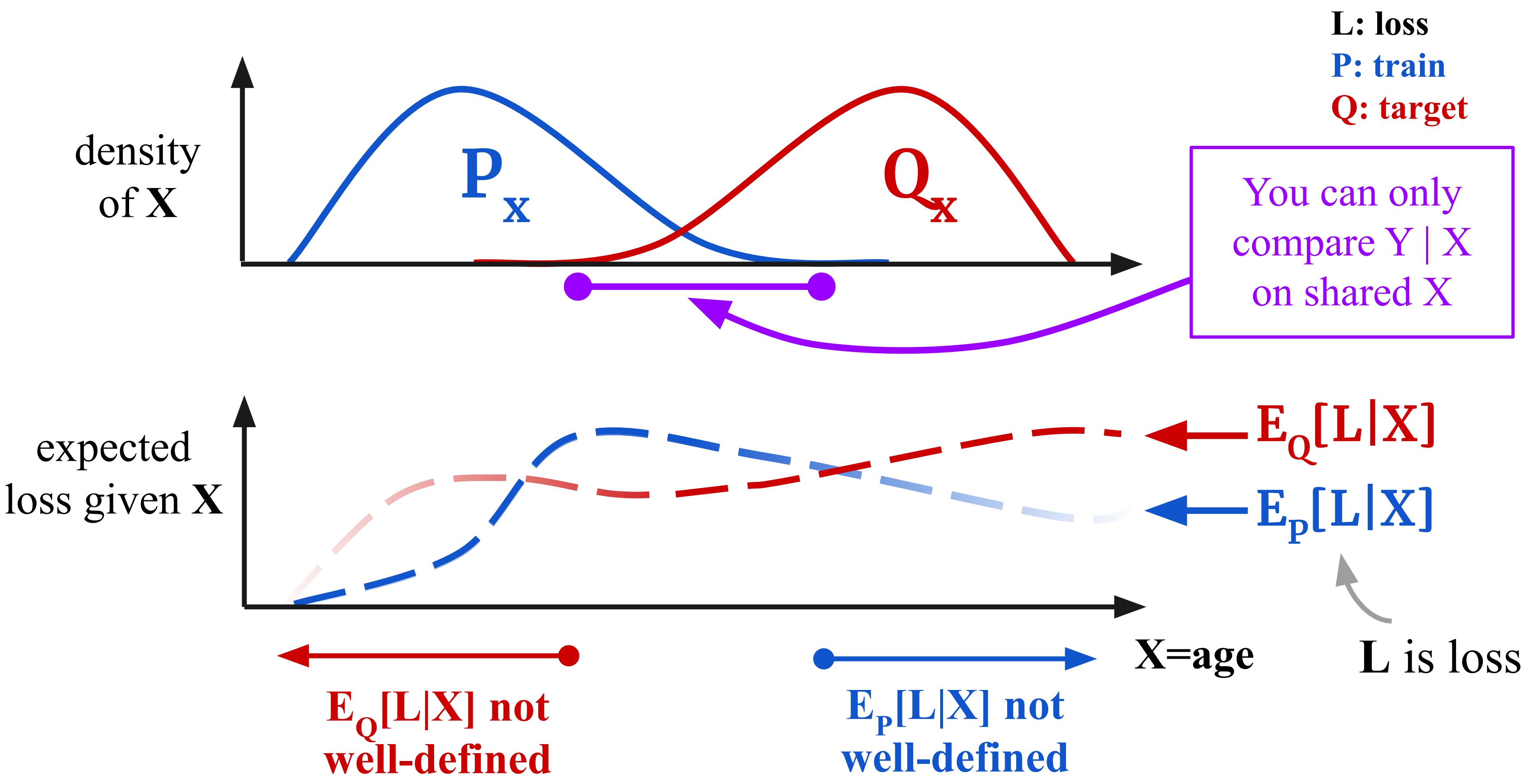
X shifts	**Y | X shifts**
changes in sampling, population shifts, minority groups	changes in labeling or mechanism, poorly chosen X

- 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



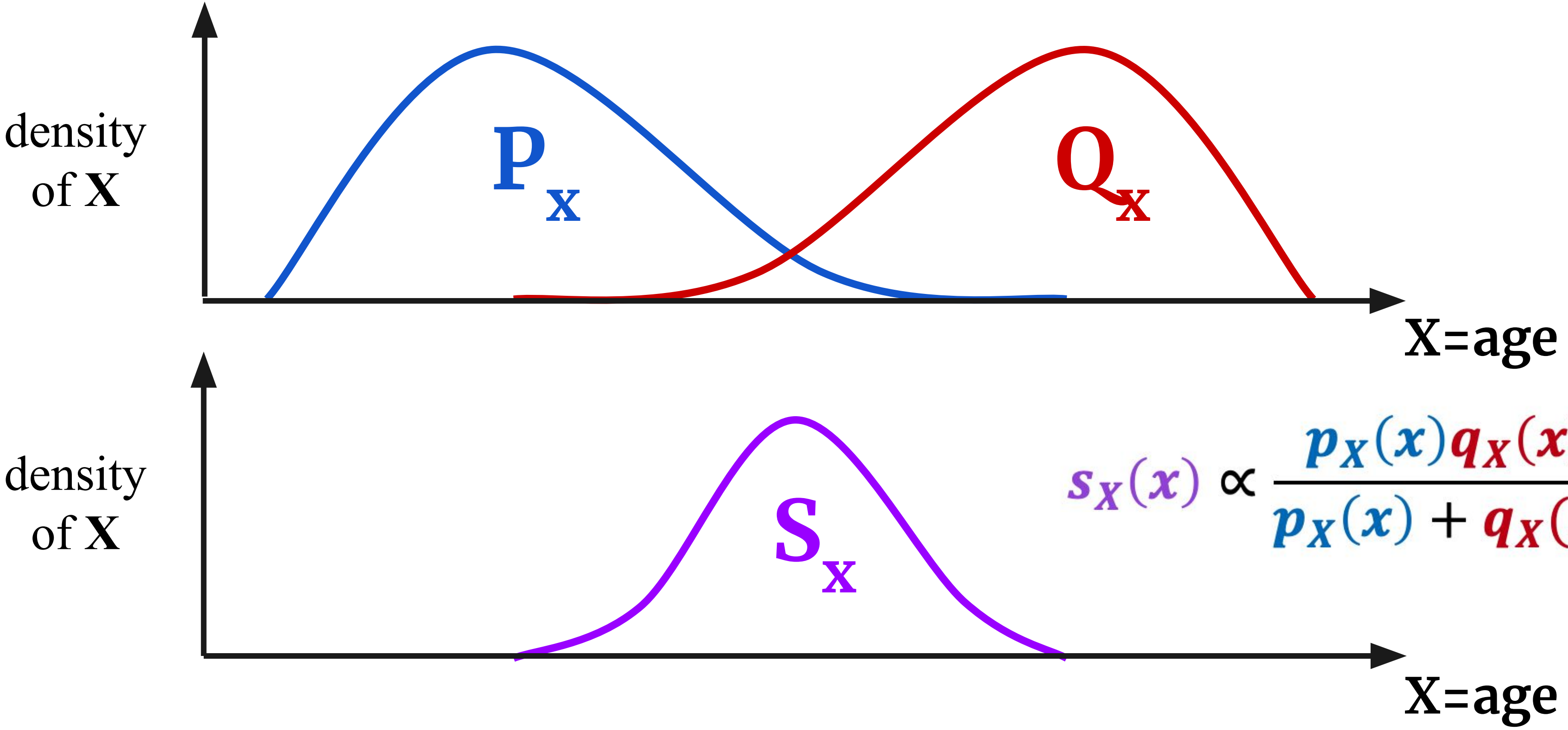
Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011



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



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

Decompose change in performance

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

$$\mathbb{E}_P[\mathbb{E}_P[L|X]]$$

Performance on the training distribution

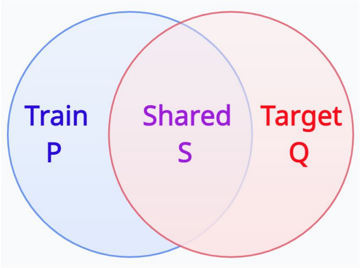
$$\mathbb{E}_Q[\mathbb{E}_Q[L|X]]$$

Performance on the target distribution

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

Decompose change in performance

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



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

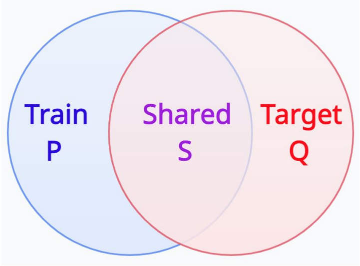
Diagnosis:
S has more X's that are harder to predict than P

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

Decompose change in performance

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



Diagnosis:

Y | X moves farther from predicted model

Potential interventions:

Re-collect data
or modify covariates

$$\mathbf{E_S[E_P[L|X]]}$$

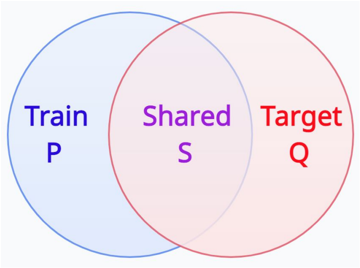
$$\Big\downarrow \; Y \mid X \text{ shift}$$

$$\mathbf{E_S[E_Q[L|X]]}$$

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

Decompose change in performance

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

Decompose change in performance

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

Legend:

$$
\begin{tikzcd}
\cdot \arrow[r, "X\text{ shift}"] & \cdot \arrow[d, dashed, "Y \mid X\text{ shift}"] \\
 & \cdot
\end{tikzcd}
$$

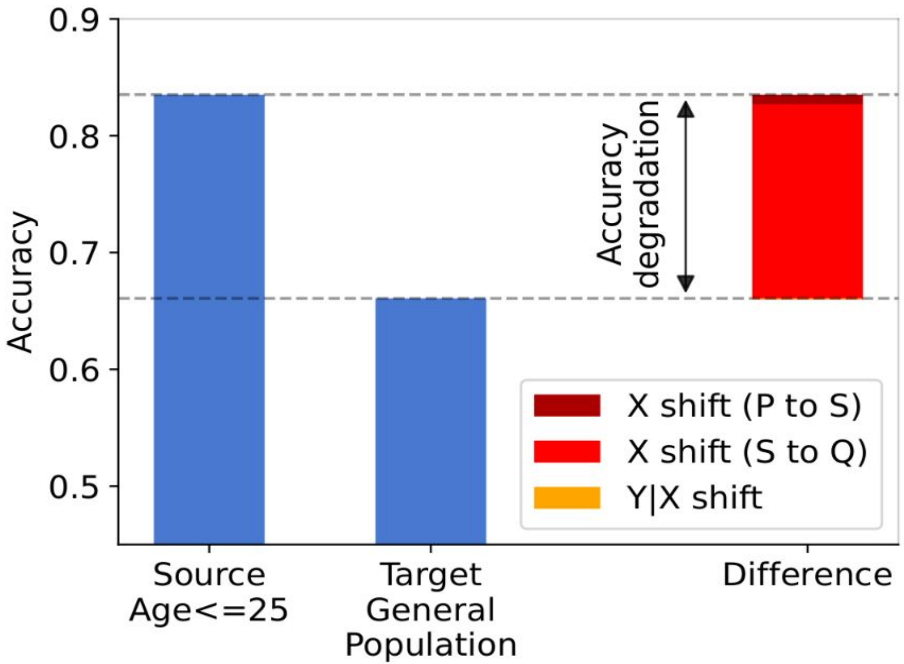
$$
\begin{tikzcd}
\mathbf{E_P[E_P[L|X]]} \arrow[r, "X\text{ shift }(P \to S)"] \arrow[d, dashed] & \mathbf{E_S[E_P[L|X]]} \arrow[r] \arrow[d, dashed, "Y \mid X\text{ shift}"] & \mathbf{E_Q[E_P[L|X]]} \arrow[d, dashed] \\
\mathbf{E_P[E_Q[L|X]]} \arrow[r] & \mathbf{E_S[E_Q[L|X]]} \arrow[r, "X\text{ shift }(S \to Q)"] & \mathbf{E_Q[E_Q[L|X]]}
\end{tikzcd}
$$

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

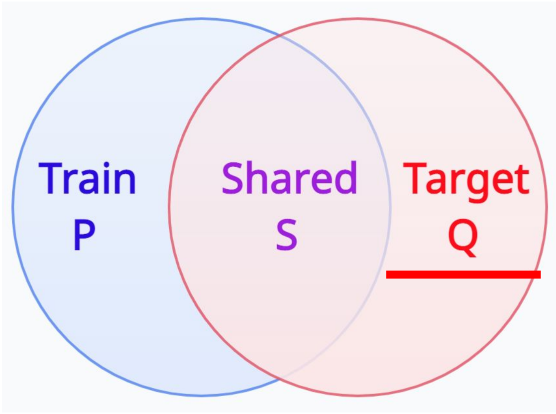
Employment prediction case study

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

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



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

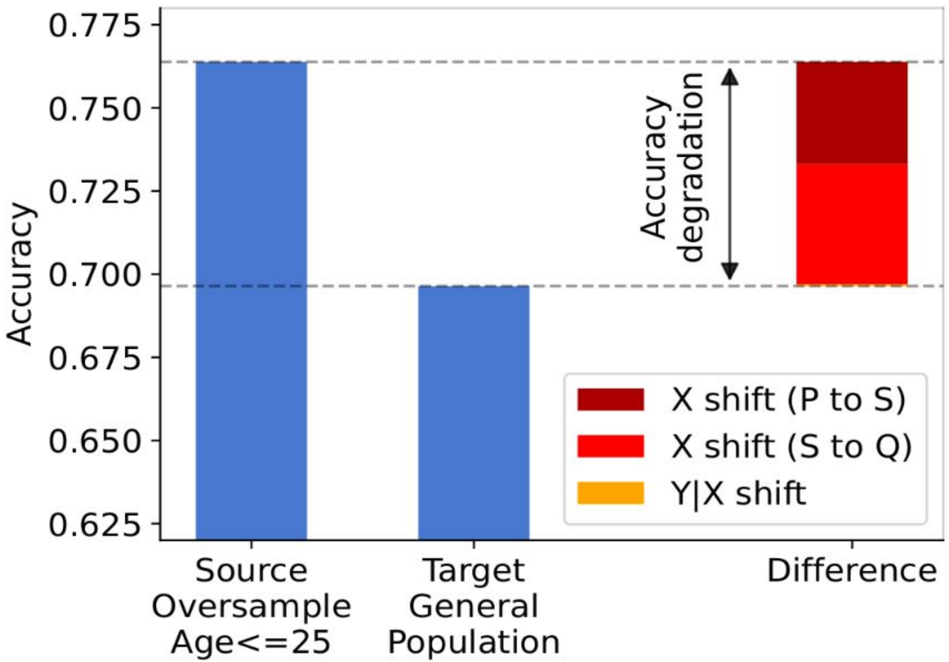


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

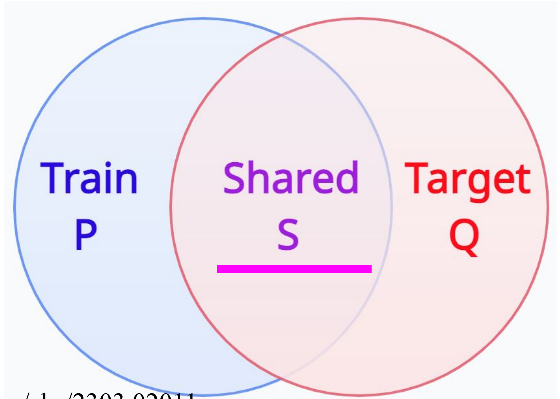
L: loss
P: train
Q: target
S: shared

Employment prediction case study

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



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

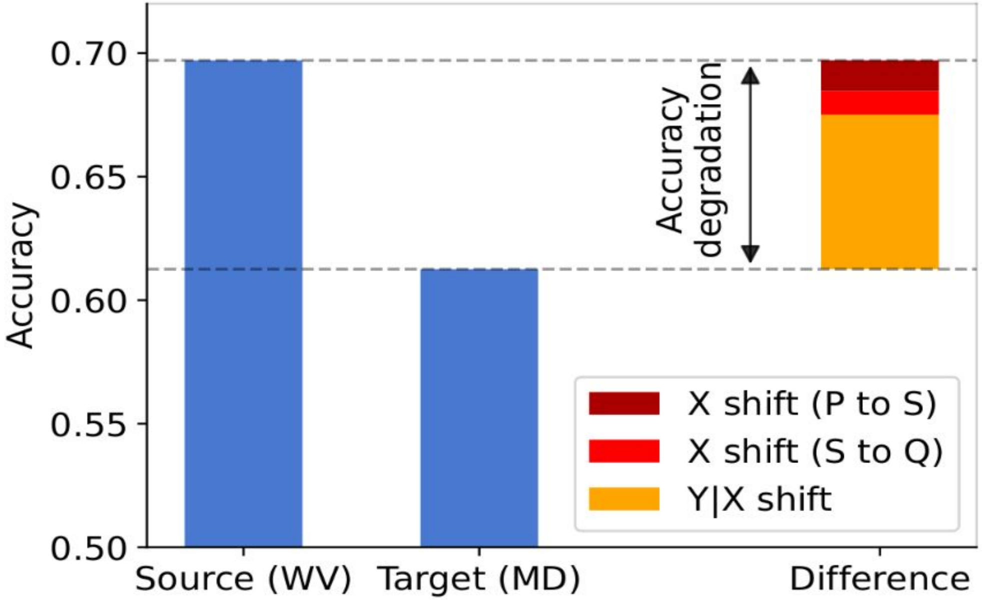


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

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)

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?

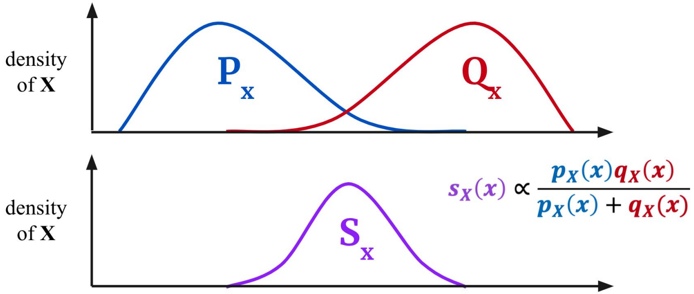
If we understand this, then we can collect
data better.

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

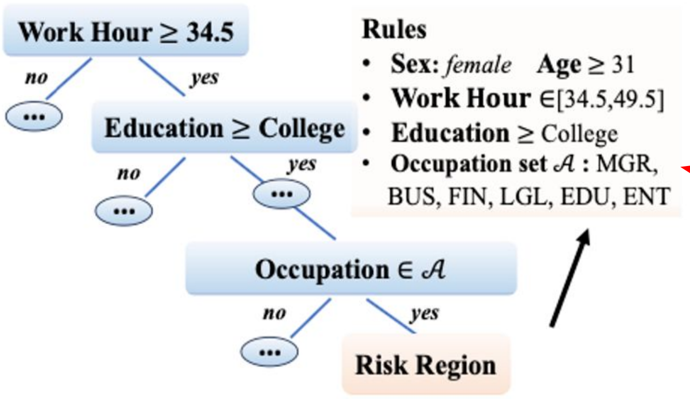


Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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!

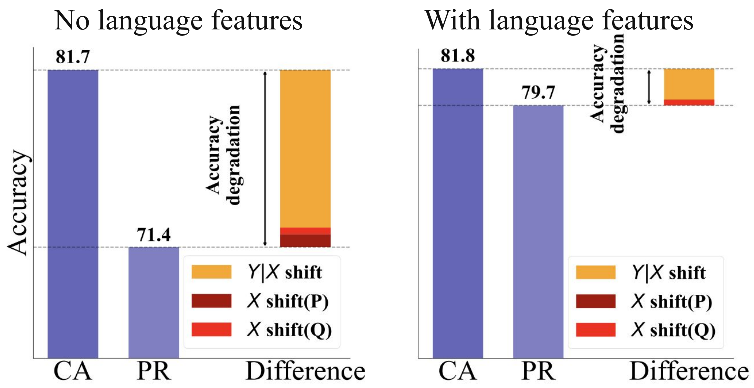
Figure from Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

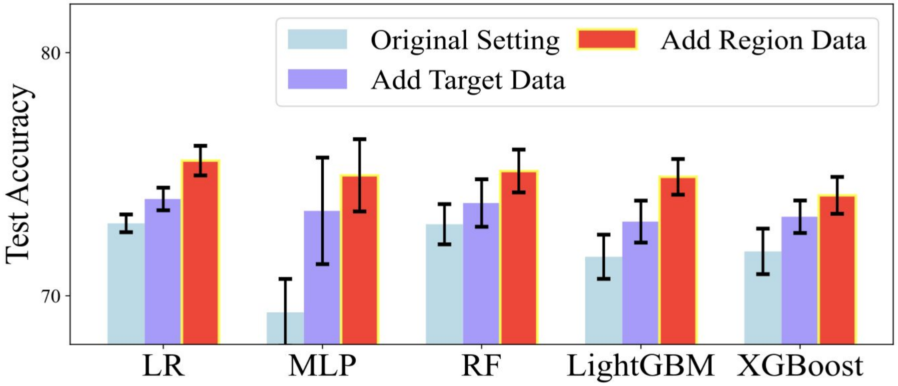
Identify Regions with $Y|X$-Shifts

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

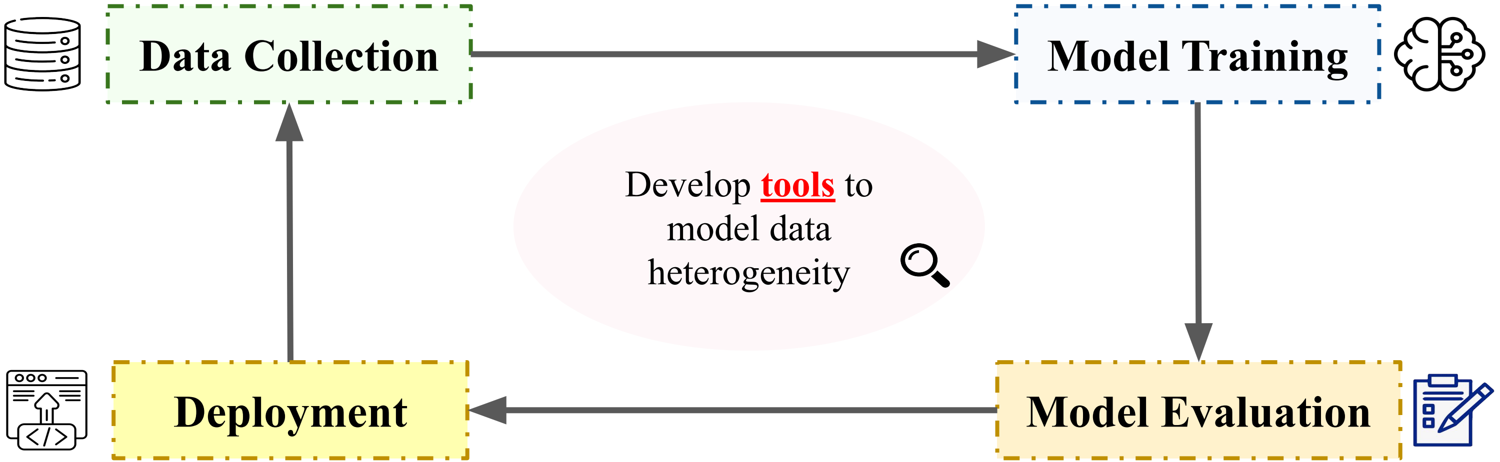
Figure from Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

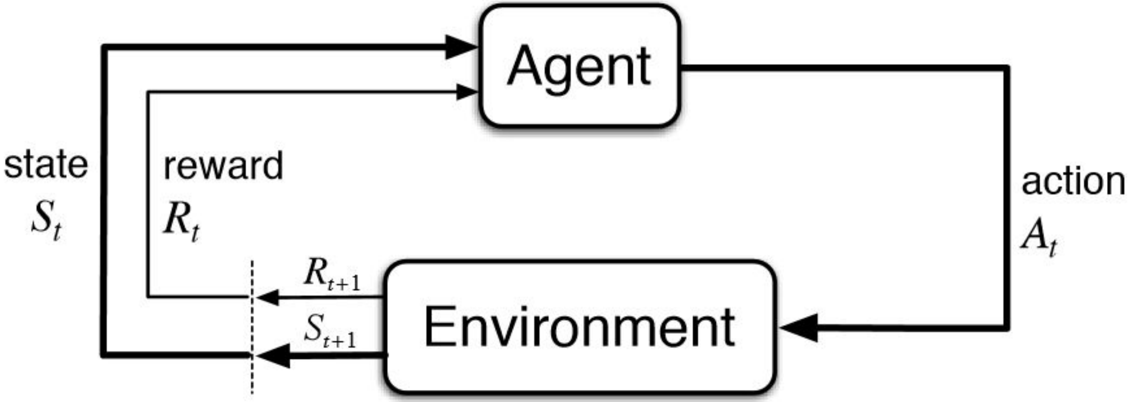


Future directions

- 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!