Toward a inductive modeling language for distribution shifts

Hongseok Namkoong
namkoong@gsb.columbia.edu

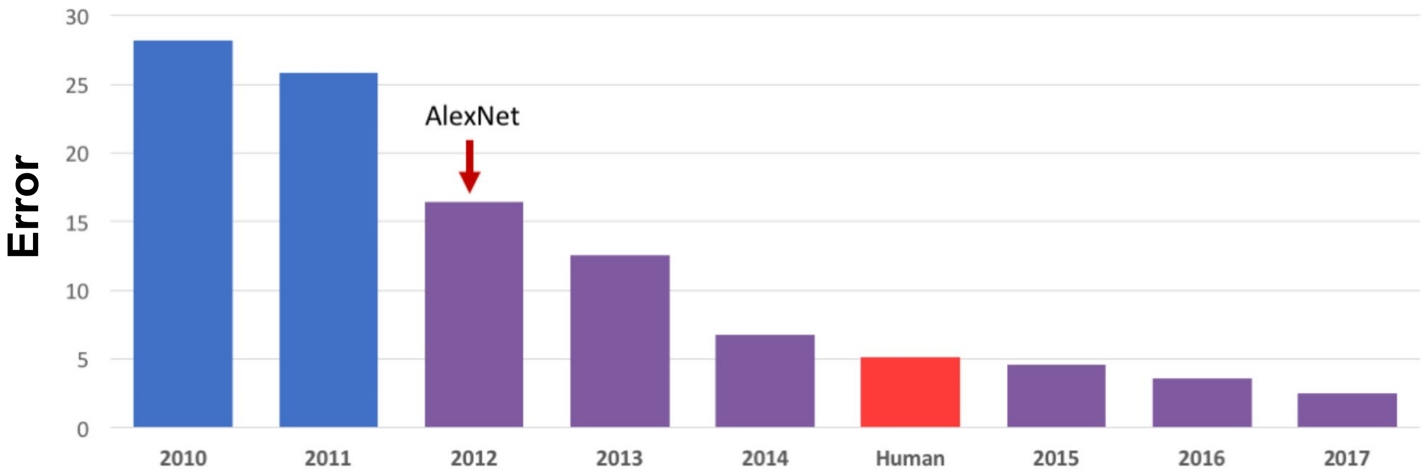
Decision, Risk, and Operations Division, Columbia Business School

Based on joint works with **Tiffany Cai**, Peng Cui, **Jiashuo Liu**, **Tianyu Wang**, Steve Yadlowsky

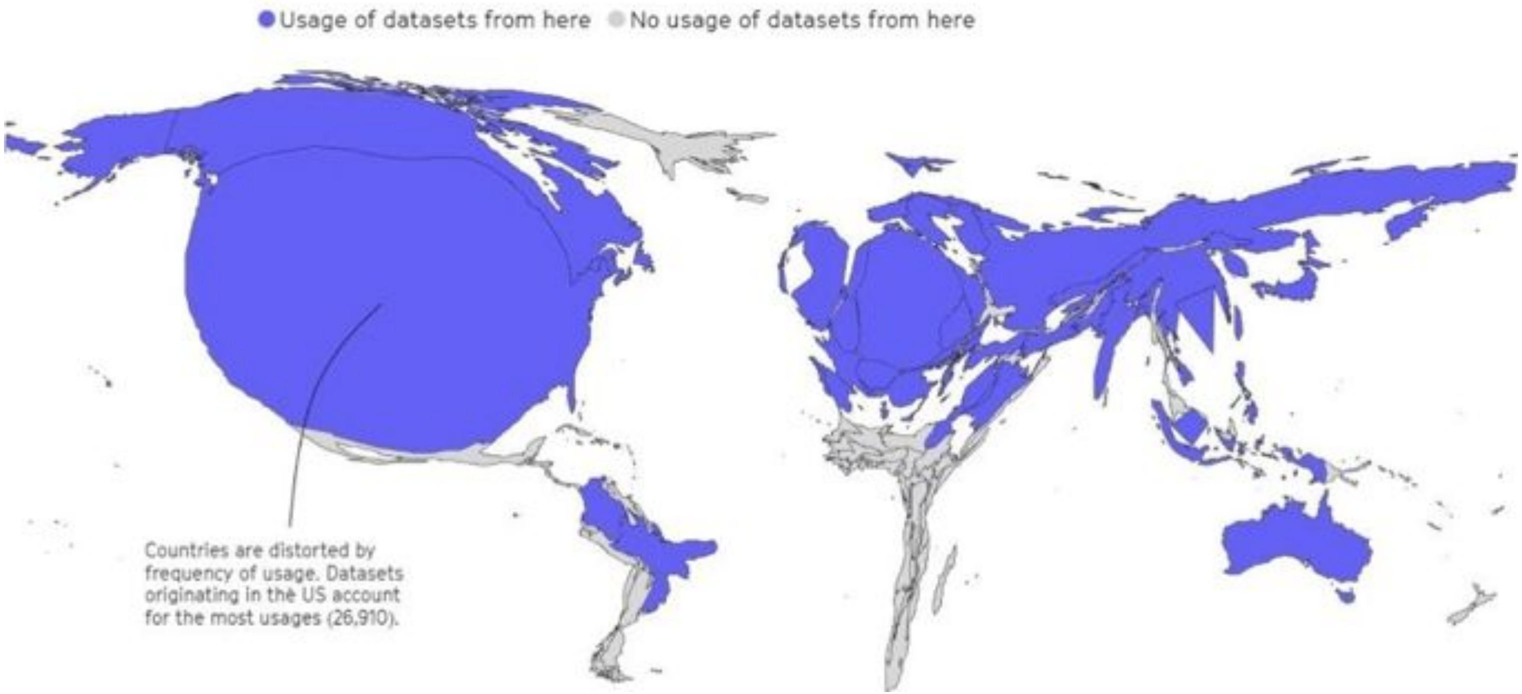
ImageNet

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

Golden retriever



AI builds on data as infrastructure



Research by: Koch, Denton, Hanna, and Foster (2021)
Visual by: The Mozilla Internet Health Report 2022

Pattern recognition will reflect existing biases

Screenshot from 2020-03-31 11-27-22.png

Technology	68%
Electronic Device	66%
Photography	62%
Mobile Phone	54%

Screenshot from 2020-03-31 11-23-45.png

Gun	88%
Photography	68%
Firearm	65%
Plant	59%

THE WHITE HOUSE MENU

OCTOBER 30, 2023

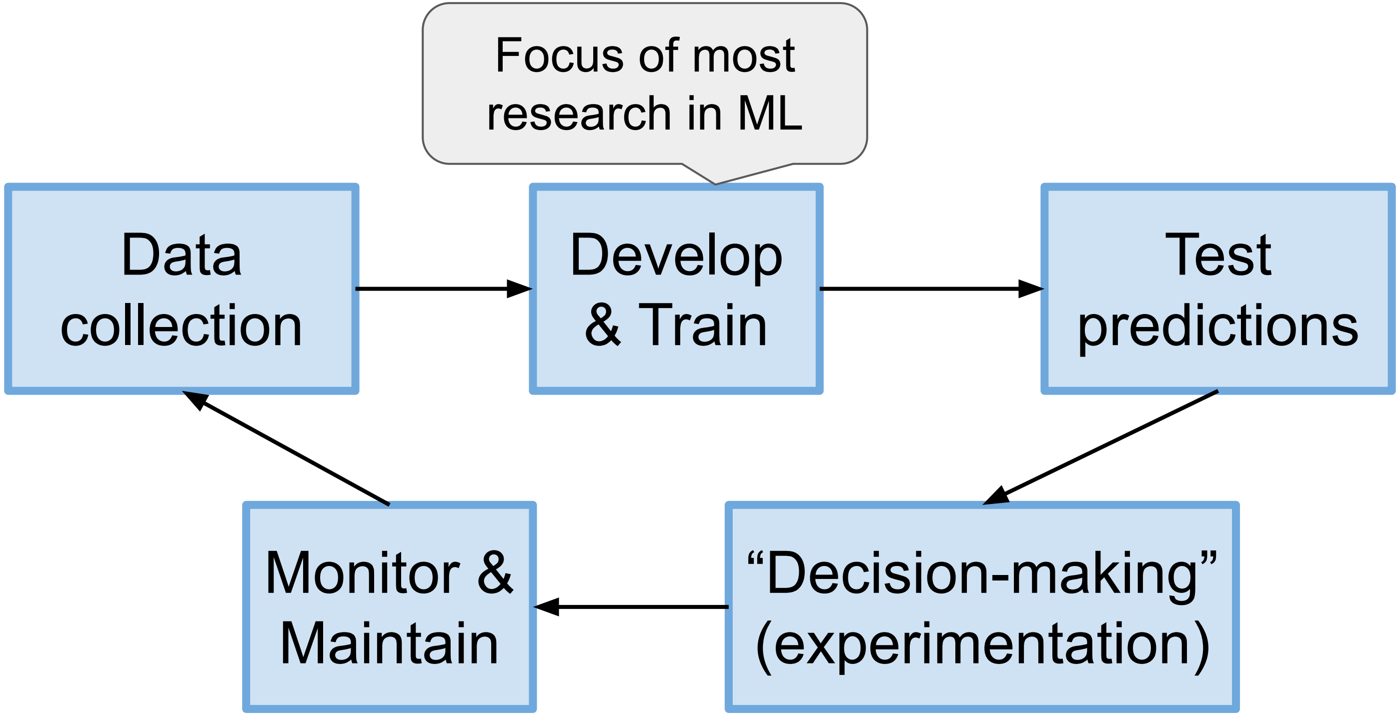
Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM ▸ PRESIDENTIAL ACTIONS

"Artificial Intelligence systems deployed irresponsibly have reproduced and intensified existing inequities, caused new types of harmful discrimination, and exacerbated online and physical harms….It is necessary to hold those developing and deploying AI accountable to standards that protect against unlawful discrimination and abuse, including in the justice system and the Federal Government."

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

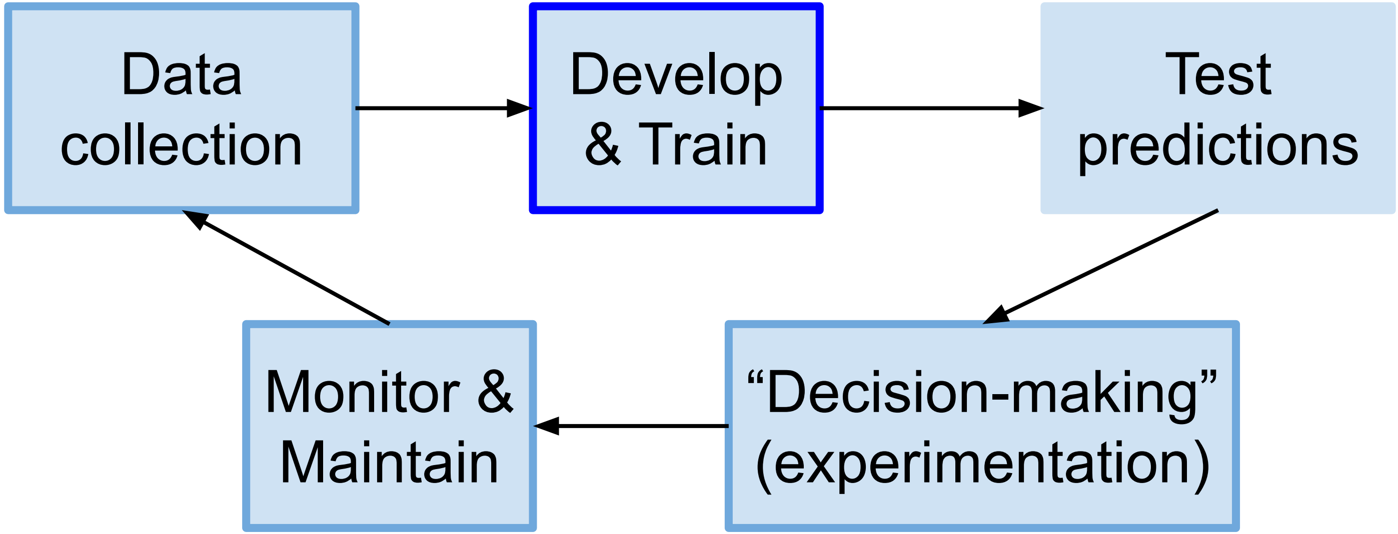
A “process” view of AI **systems** (not just a model)



Trustworthy data-driven decision-making

- Reliability is a first-order problem in AI-driven decisions
    - Standard CS ML benchmarking view breaks down
- I study AI systems with **distribution shifts** as a central concern
    - Build algorithmic + empirical foundation with a **modern ML lens**
- Main application: online platforms where AI-systems influence high-stakes decisions
    - Algorithmic hiring / sourcing, e.g., allocation of limited recruiter bandwidth across candidates at LinkedIn

A “process” view of AI systems



ML as stochastic optimization

- Standard approach: Solve average-case risk minimization

$$\text{minimize}_{f(\cdot)} \;\; \mathbb{E}_P[\ell(Y, f(X))]$$

- Distributionally robust optimization: Solve worst-case problem

$$\text{minimize}_{f(\cdot)} \;\; \max_{Q \in \mathcal{P}} \; \mathbb{E}_Q[\ell(Y, f(X))]$$

- Idea: Do well almost all the time, instead of on average!

Recent progress

- DRO can contribute to generalization, robustness, and fairness
- Intellectual foundations: training algorithms and data efficiency
- Practical impact: algorithms useful when real shifts can be modeled succinctly, e.g., fairness across demographic groups

Duchi and **N**. Learning models with uniform performance via distributionally robust optimization. Annals of Statistics, 2021.

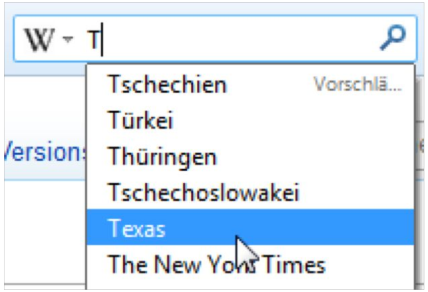
Duchi, Hashimoto, and **N**. Distributionally robust losses against mixture covariate shifts. Operations Research, 2022.

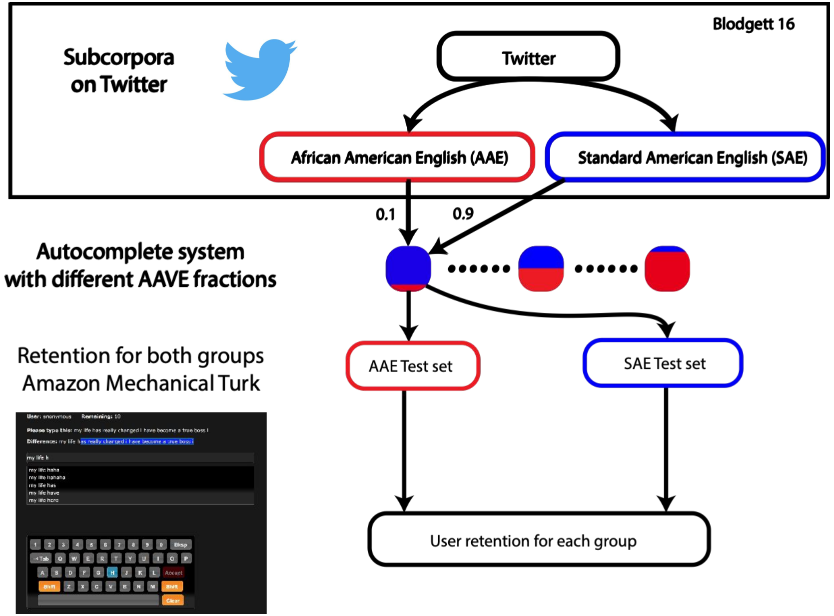
Hashimoto, Srivastava, **N**, and Liang. Fairness without demographics in repeated loss minimization. ICML, 2018. Best Paper Runner-up.

Sinha$*$, **N**$*$, and Duchi. Certifiable distributional robustness with principled adversarial training. ICLR, 2018. Oral presentation.

Vignette: auto-complete service

Motivation: Autocomplete system for text





Problem: Atypical text doesn't get surfaced

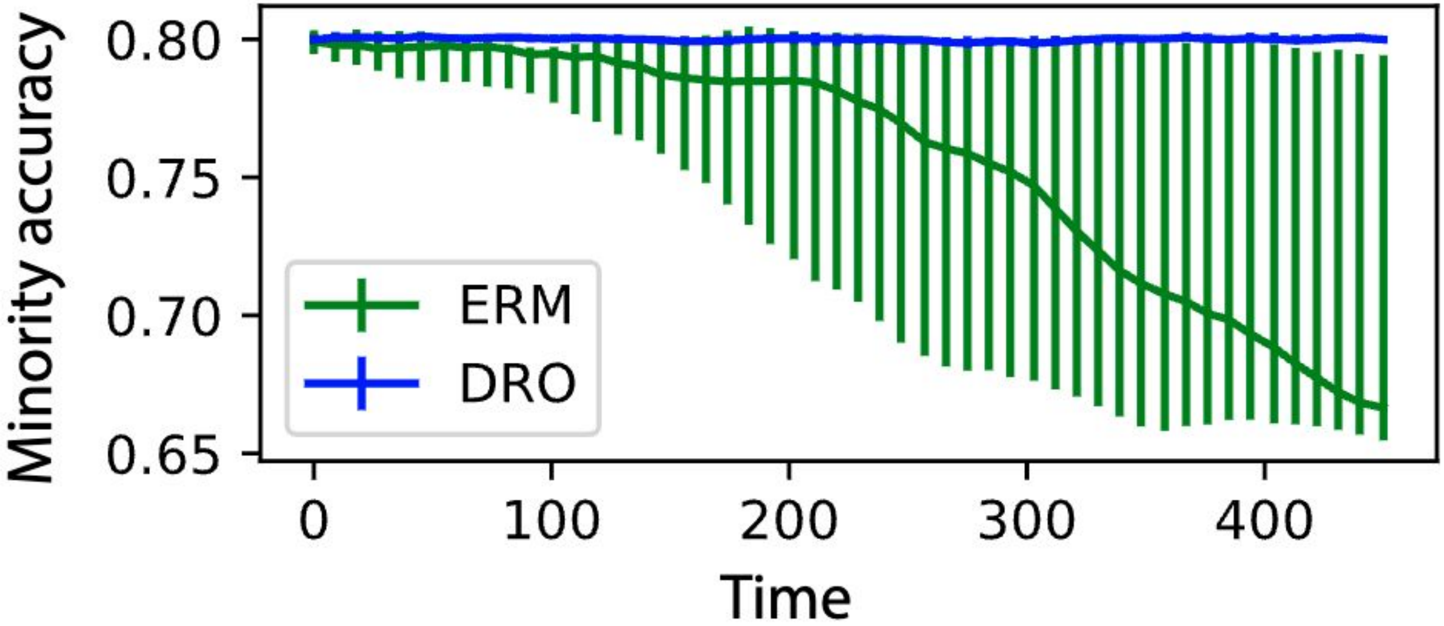
African American Vernacular (AAVE)

If u wit me den u pose to RESPECT ME

Standard American English (SAE)

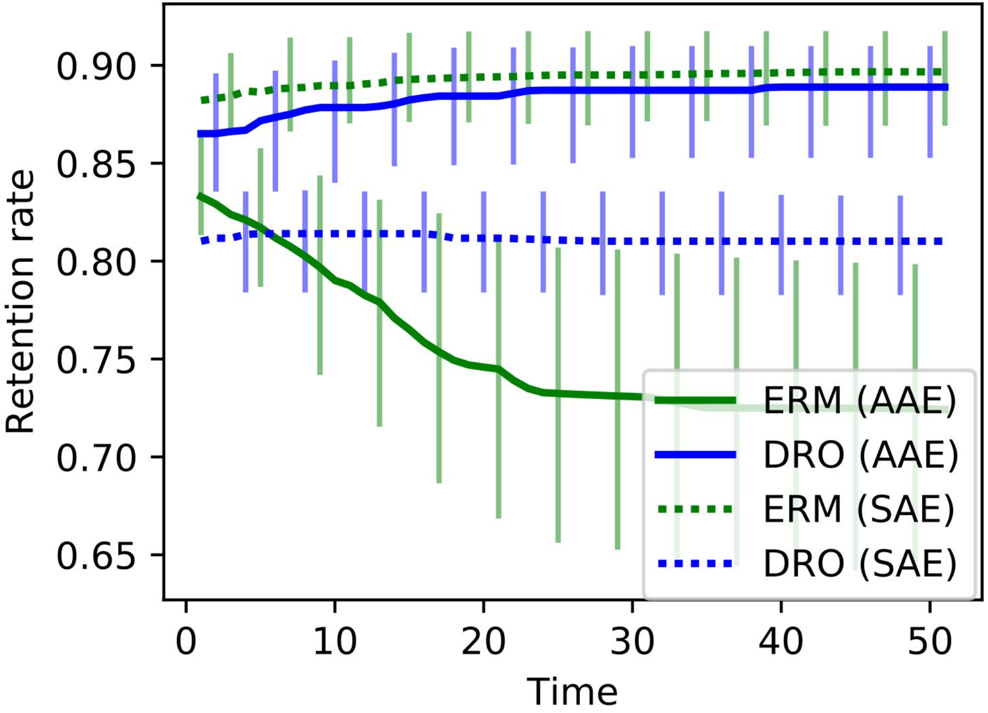
If you are with me then you are supposed to respect me.

DRO mitigates disparity amplification



Hashimoto, Srivastava, **N**, and Liang. Fairness without demographics in repeated loss minimization. ICML, 2018. Best Paper Runner-up.

DRO mitigates disparity amplification



Takeaway:

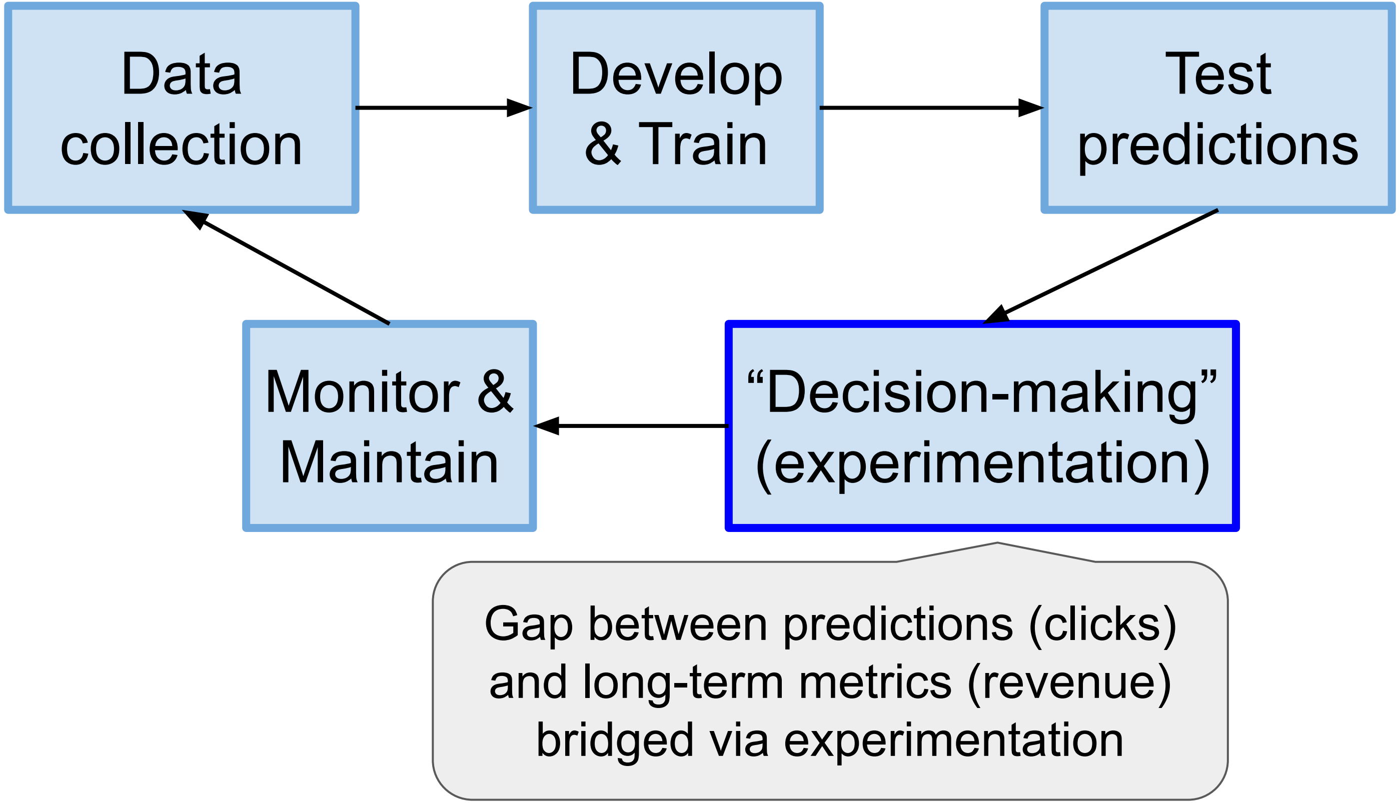
Control minority proportion

↓

Uniform performance over time

Hashimoto, Srivastava, **N**, and Liang. Fairness without demographics in repeated loss minimization. ICML, 2018. Best Paper Runner-up.

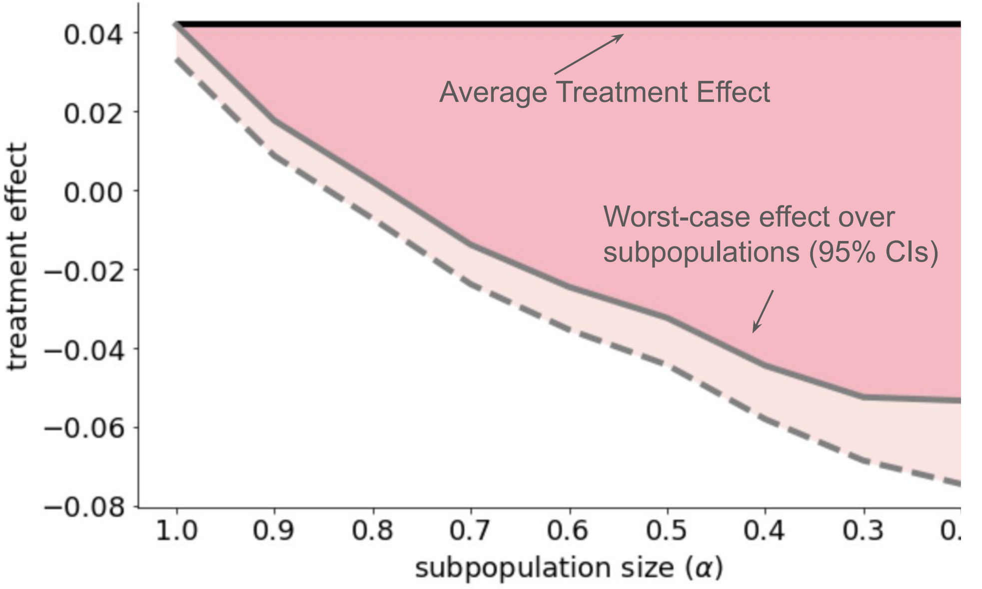
Causal inference and experimentation



Distributional robustness is a useful diagnostic

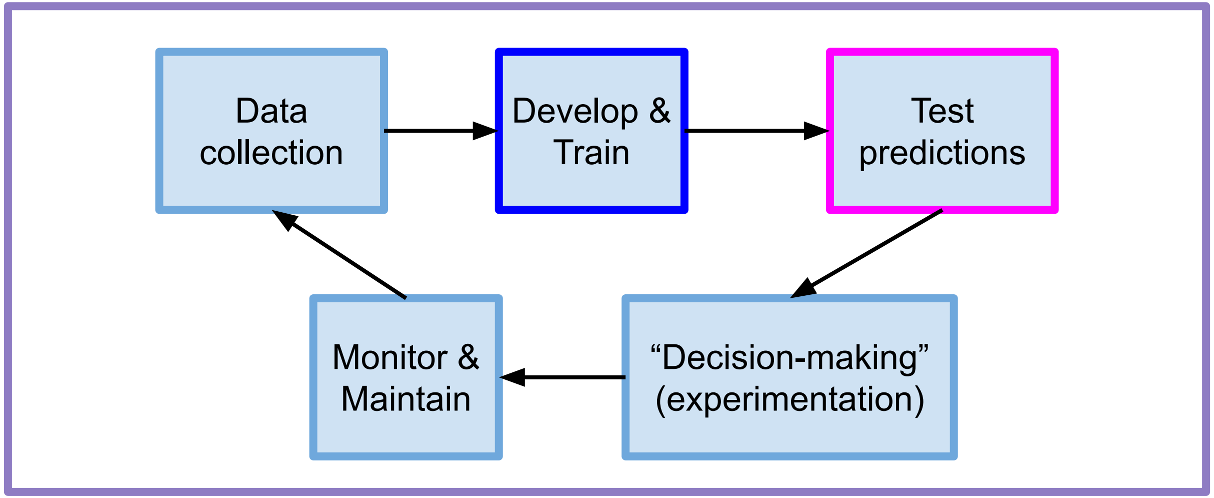
- Causal inference is fundamental to scientific decision-making
- Its reliability depends on the ability to extrapolate a study's findings
- Assess validity of findings under distribution shifts
    - Example: finding fails to hold over subpopulations comprising 80% of the study population

Effect of Medicaid enrollment on doctors' office utilization



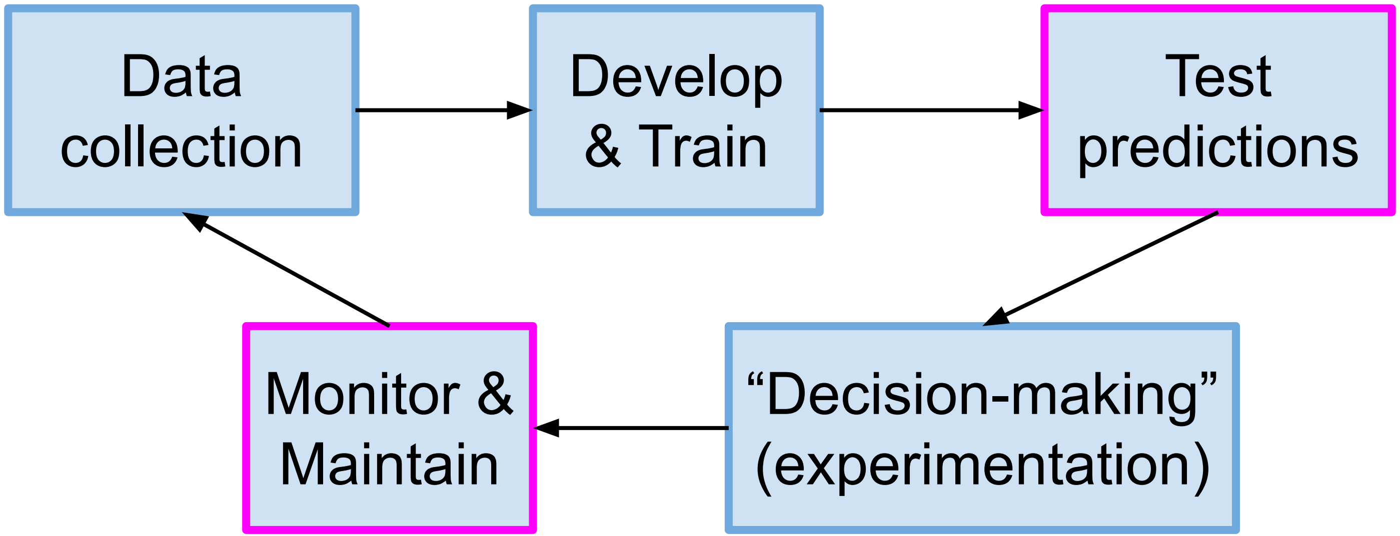
Jeong and N. Assessing external validity over worst-case subpopulations, Short version appeared at COLT2020.
YNBDT. Bounds on the conditional and average treatment effect with unobserved confounding factors. Annals of Statistics, 2022.
NKYB. Off-policy policy evaluation for sequential decisions under unobserved confounding. NeurIPS, 2020.
Boyarsky, Egami, and N.. Assessing external validity of RCTs under effect-ordering. Work in progress.
Ma, Huang, and N.. A practical minimax approach to causal inference with limited overlap. Work in progress

Industry applications



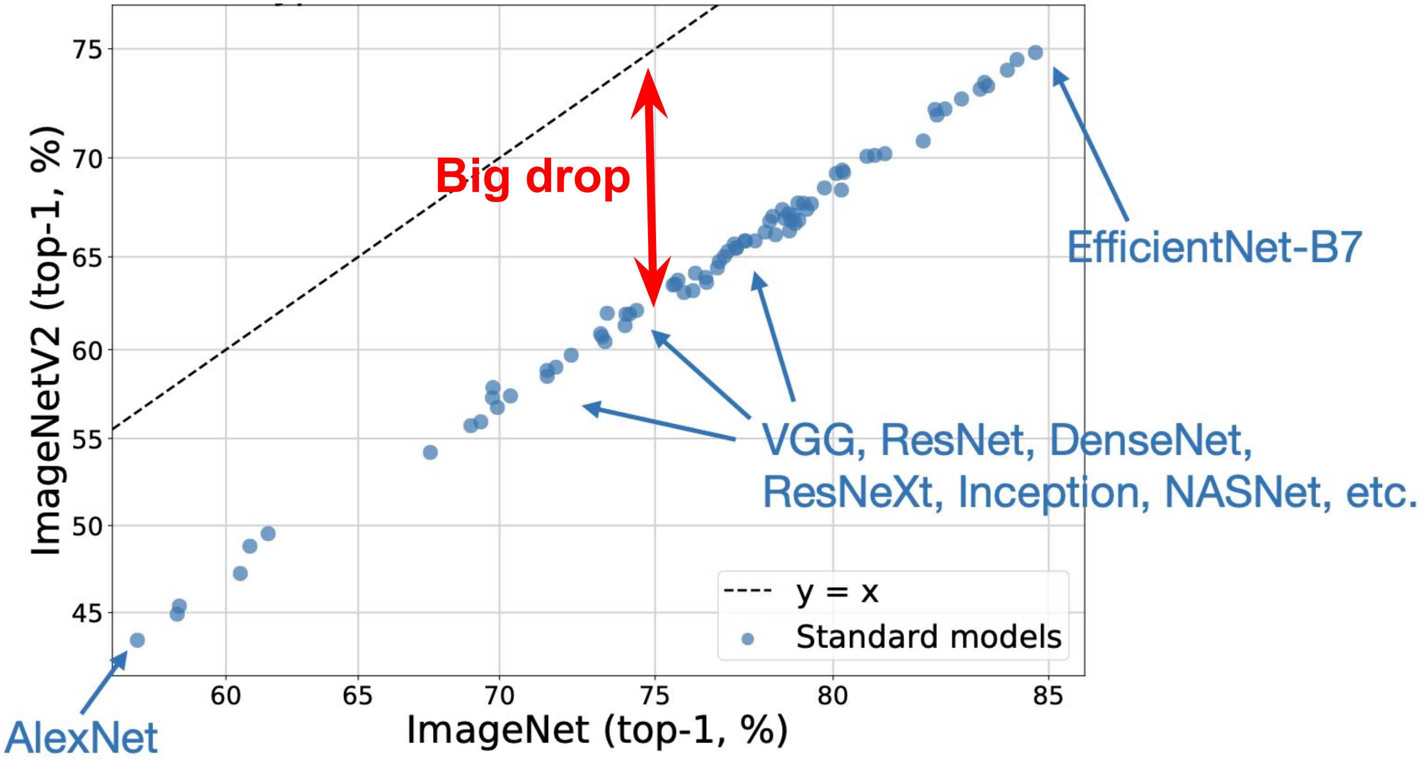
- **Engineering constraints**: Robust algos under infrastructural constraints
- **Compliance**: Disparate treatment, design best practices for "due diligence"
- **Governance**: Standardize & scale requirements at the company level

Today: Diagnostics



Understand **why** predictive performance degraded

Back to ImageNet



Slide credit: Ludwig Schmidt

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

How do we go up the red line?

- Algorithmic interventions do not provide robustness; only larger training data does—AI community focus on scaling internet data
- But cost of data collection remains a binding constraint; need to understand **which** data to collect
- Implicit assumptions in the CS benchmarking view (one-size-fits-all)
    - Building a universally robust model, just like humans!
    - Focus on covariate shift (*X*-shift), e.g., image recognition

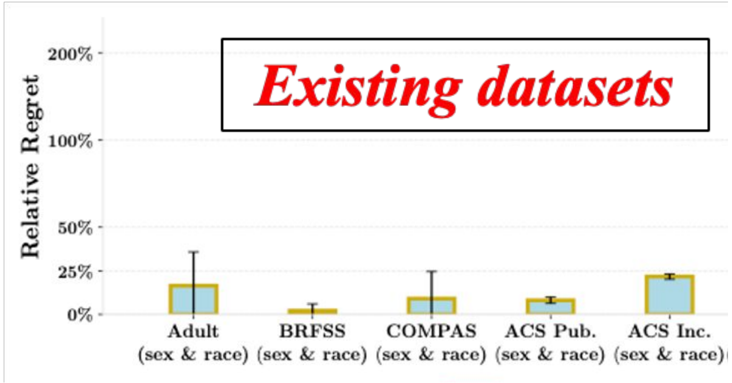
[WIKLKRGHF**N**S'22] Robust fine-tuning of zero-shot models. CVPR, 2022. Best Paper Award Finalist.
[WIGRGM**N**FCKS'22] Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. ICML, 2022.

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



Today: Design new datasets from US census data!

Even tabular benchmarks mainly focus on *X*-shifts

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

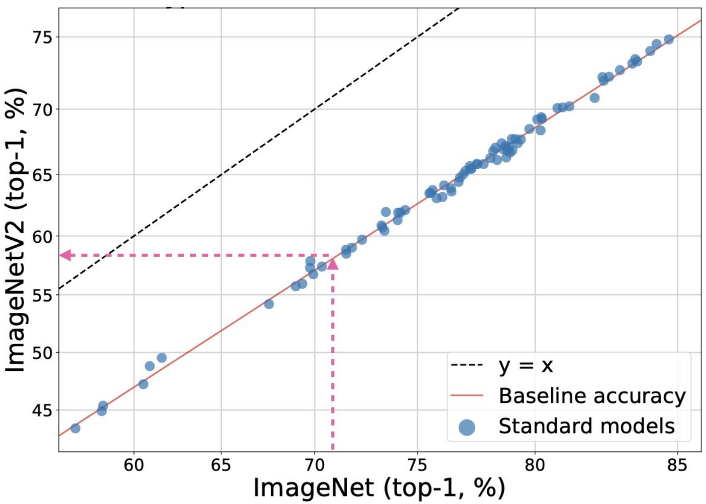
$$\frac{\mathbb{E}_Q[\ell(Y, f_P(X))]}{\min_{f\in\mathcal{F}} \mathbb{E}_Q[\ell(Y, f(X))]} - 1, \quad \text{where} \quad f_P \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_P[\ell(Y, f(X))]$$

relative regret

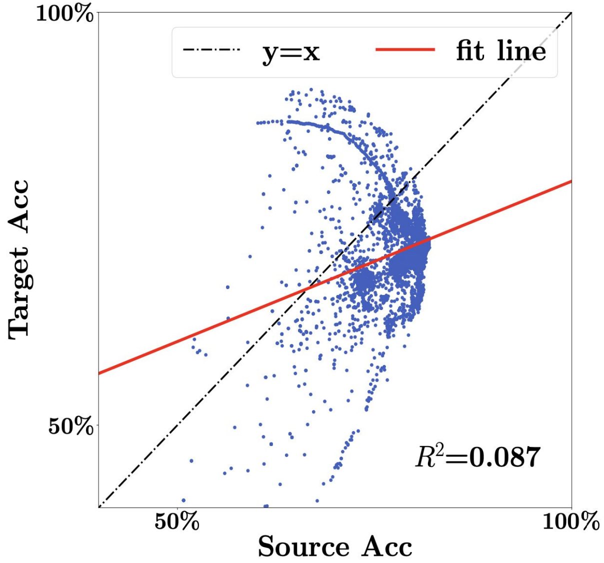
Liu, Wang, Cui, N., On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets, Short version in NeurIPS, 2023.

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

- Train & target performance correlated *only when X-shifts dominate*



ImageNet



ACS Income (CA→PR)
$Y|X$-ratio: 85.4%

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

WhyShift

arxiv

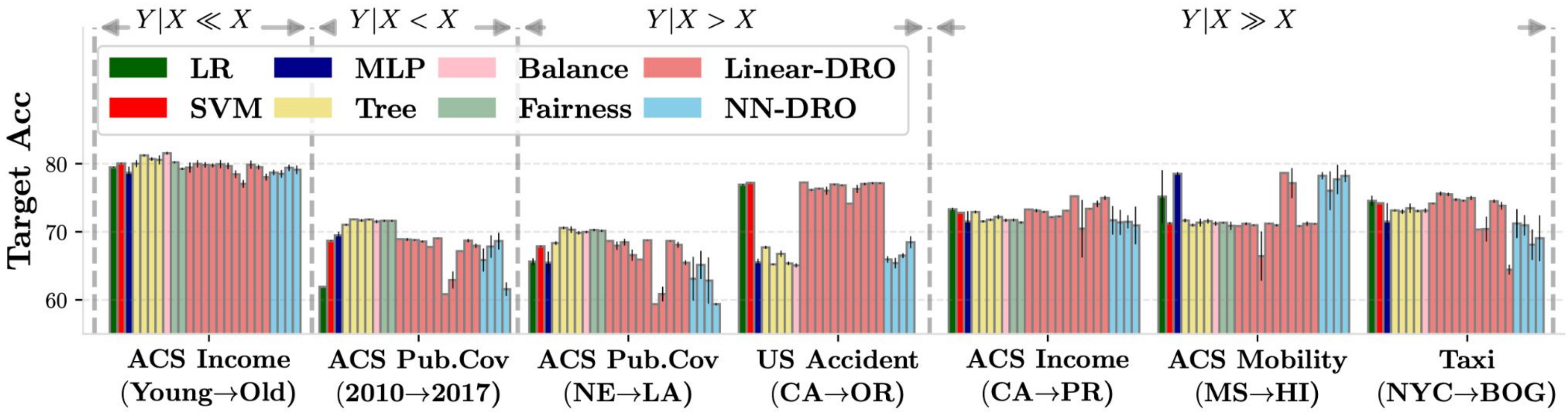
github

whyshift 0.1.3

`pip install whyshift`

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

- Out of 169 train-target pairs, 80% primarily suffer $Y|X$-shifts
- Existing algos do not show consistent robustness gains
  - They make assumptions about data distributions but do **not** check them
  - We need an understanding of **why** the distribution changed!



DRO revisited

- Distributionally robust optimization: Solve worst-case problem

$$\text{minimize}_{f(\cdot)} \max_{Q \in \mathcal{P}} \mathbb{E}_Q[\ell(Y, f(X))]$$

- Choice of ambiguity set $\mathcal{P}$ arbitrary; primarily driven by mathematical convenience and details “left to the modeler”

- Little thought given to model class $f(\cdot)$

Empirical analysis of 10,000 DRO models

- Analyze impact of algorithmic design knobs on model robustness

Task/State fixed effect

$$Acc_{i,j,s,t} = \alpha + \beta_1^\top X_{i,j,s} + \beta_2^\top D_{i,j,s,t} + \beta_3 Z_{i,j} + \beta_4^\top V_{i,j} + \boxed{\mu_i + \tau_j} + \varepsilon_{i,j,s}$$

Model Class
(Tree, Linear, MLP)

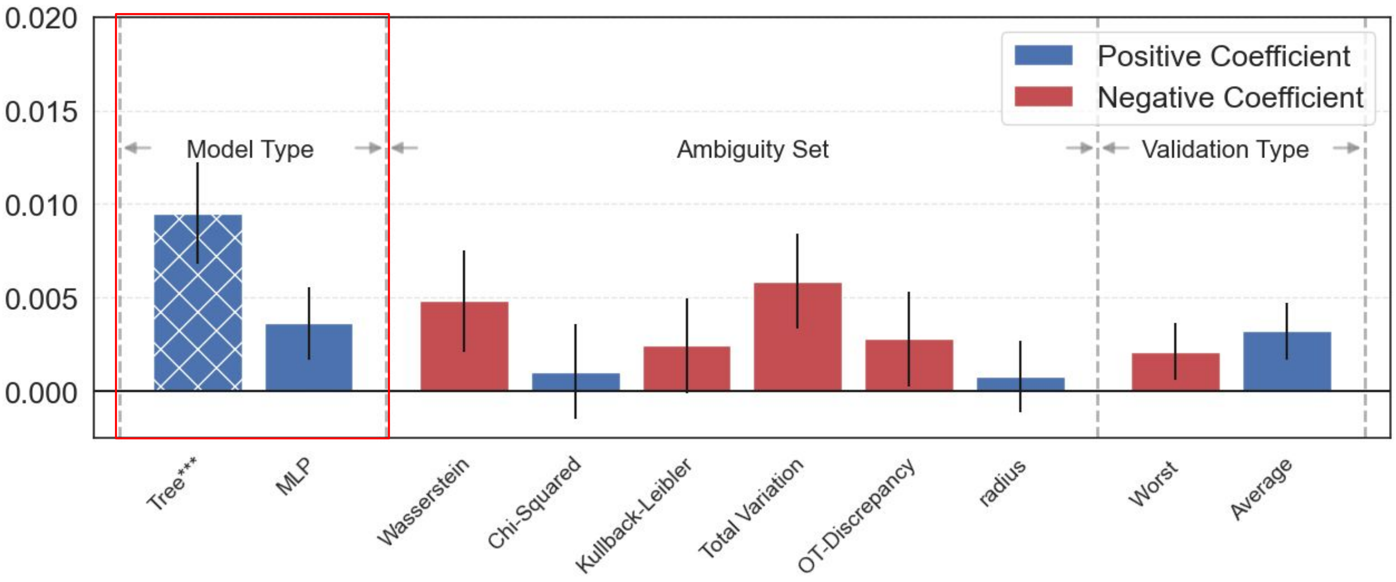
Ambiguity Set
(Distance Type, Radius)

Shift Pattern
($Y|X$-ratio)

Validation Type
(Average, Worst)

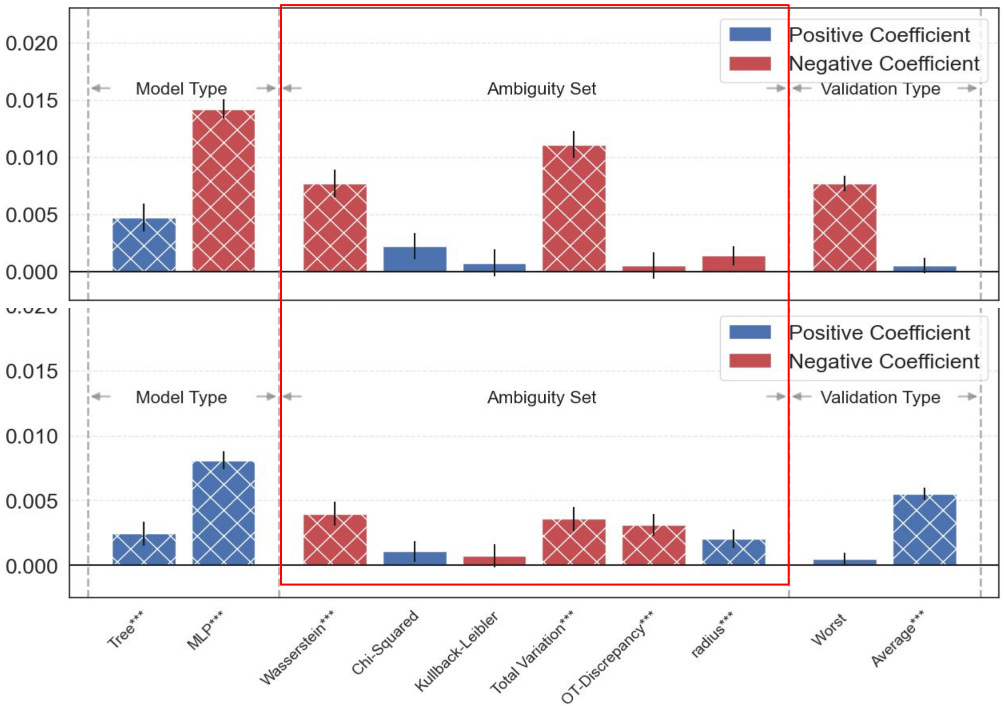
Target performance: single state

- Model class most important! Trees >>> ambiguity set
- Effect of ambiguity set inconsistent across different outcomes



Target performance: single state

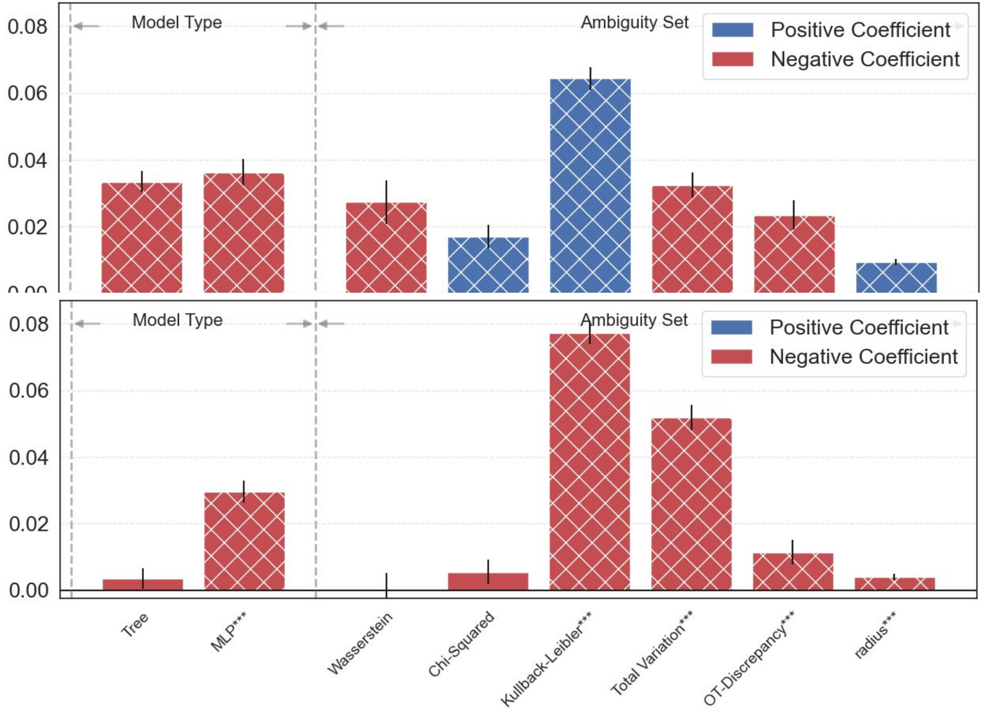
- Effect of ambiguity set inconsistent across different outcomes



Upper: Predict whether a low-income individual, not eligible for Medicare, has coverage from public health insurance.
Lower: Predict whether annual income > $50K

Target performance: **worst** state

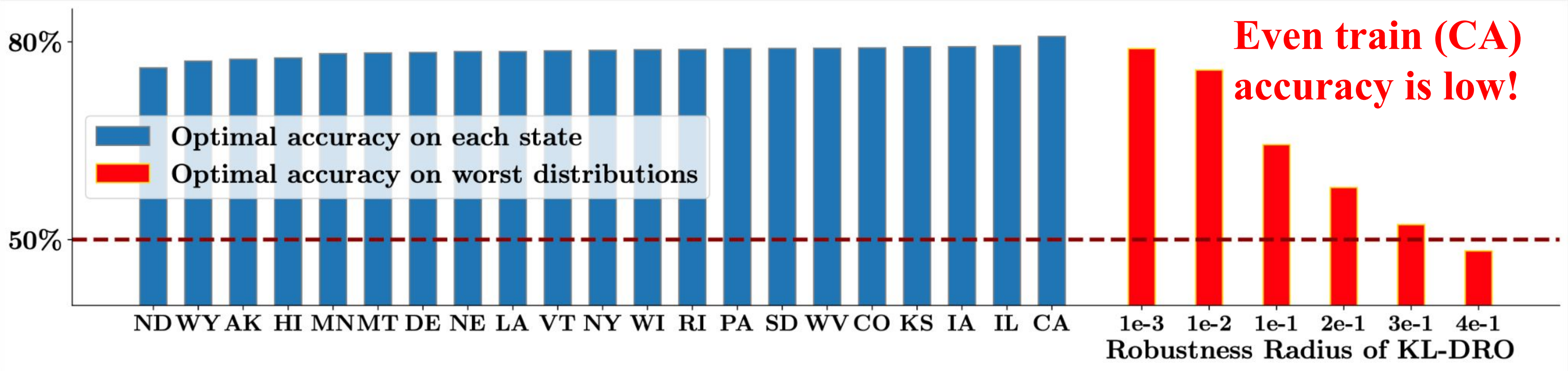
- Even for worst-state performance, DRO is unreliable



Upper: Predict whether a low-income individual, not eligible for Medicare, has coverage from public health insurance.
Lower: Predict whether annual income > $50K

Problems with deductive reasoning

Worst-case distribution does not match real targets



Blue bars: Accuracy of logistic regression models trained on each state.

Red bars: Accuracy on worst-case distribution from a DRO model trained on CA

Last week's discussion scientific methods

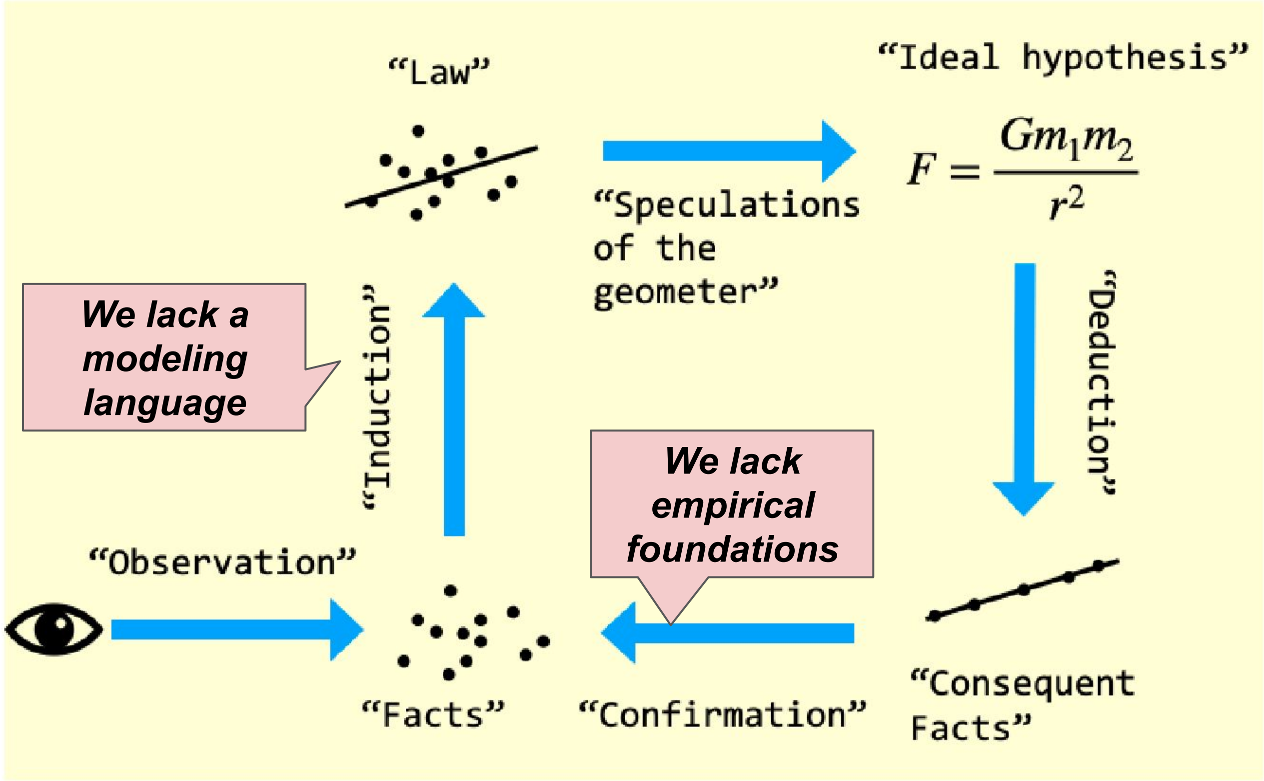
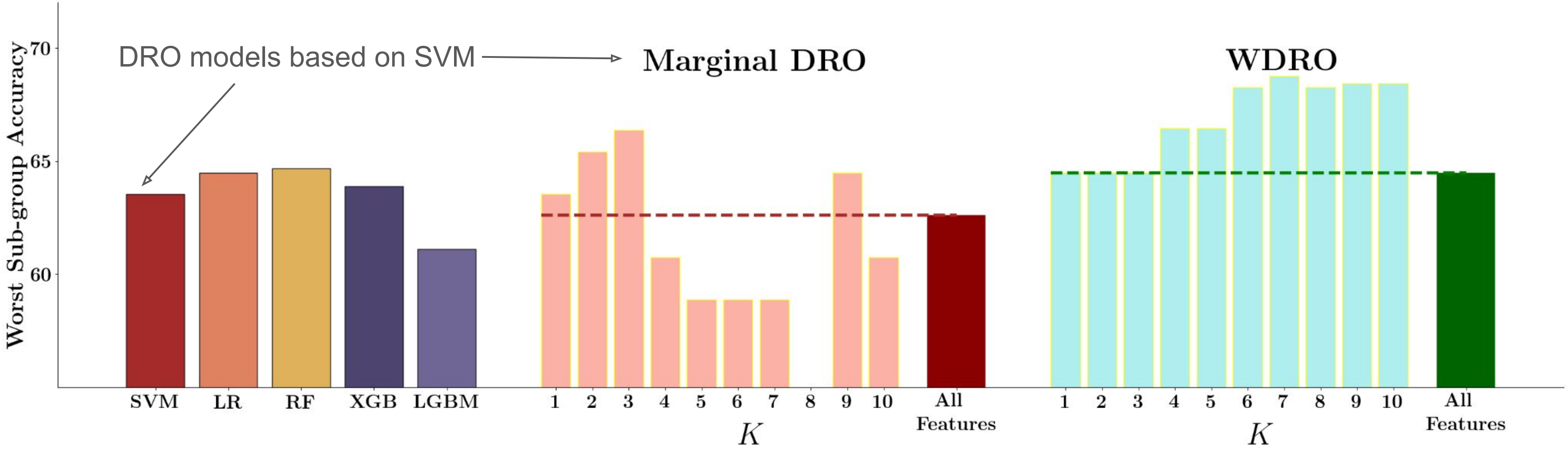


Figure from Christopher Ryan, DRO Brown Bag, April 2024

Inductive approach to ambiguity sets: *X*-shifts

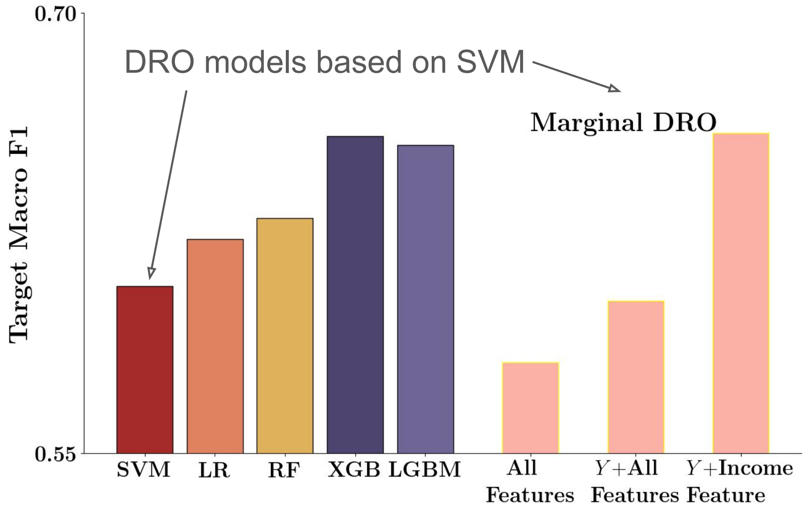
- Consider shifts induced by age groups: [20,25), [25,30), …, [75,100)
- Consider DRO methods (DH**N**'22) tailored to shifts on a subset of covariates
- Variable selection for ambiguity set: top-K with largest subgroup differences
- Performance varies a lot over variables selected



Duchi, Hashimoto, and **N**. Distributionally robust losses against mixture covariate shifts. Operations Research, 2022.

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

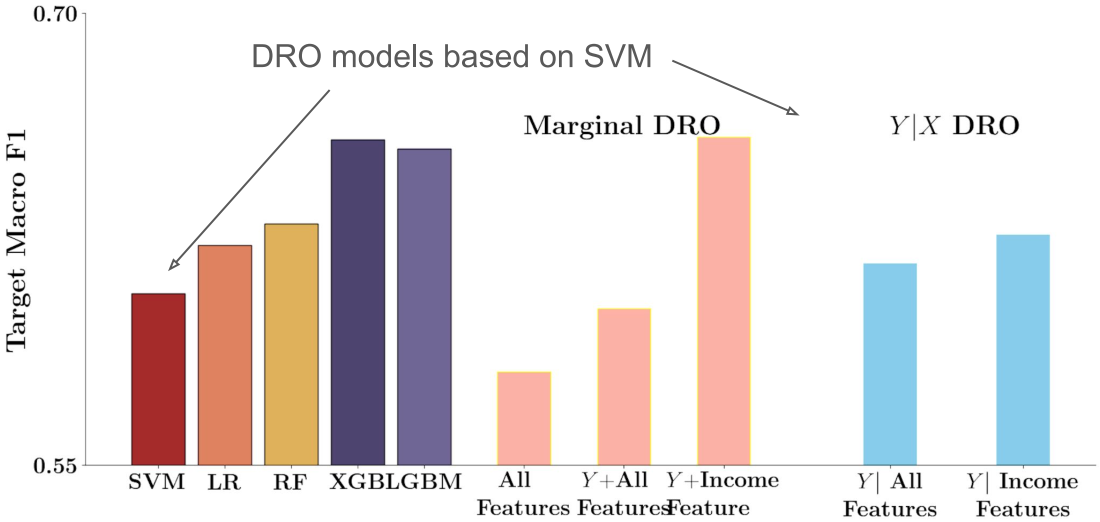
- Consider $Y|X$-shifts from NE -> LA (public coverage task)
- Consider DRO methods that consider shifts on a subset of covariates and Y
- Variable selection for ambiguity set: Y | “income” suffers the largest shift
- Performance varies a lot over variables selected



Duchi, Hashimoto, and **N**. Distributionally robust losses against mixture covariate shifts. Operations Research, 2022.

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

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

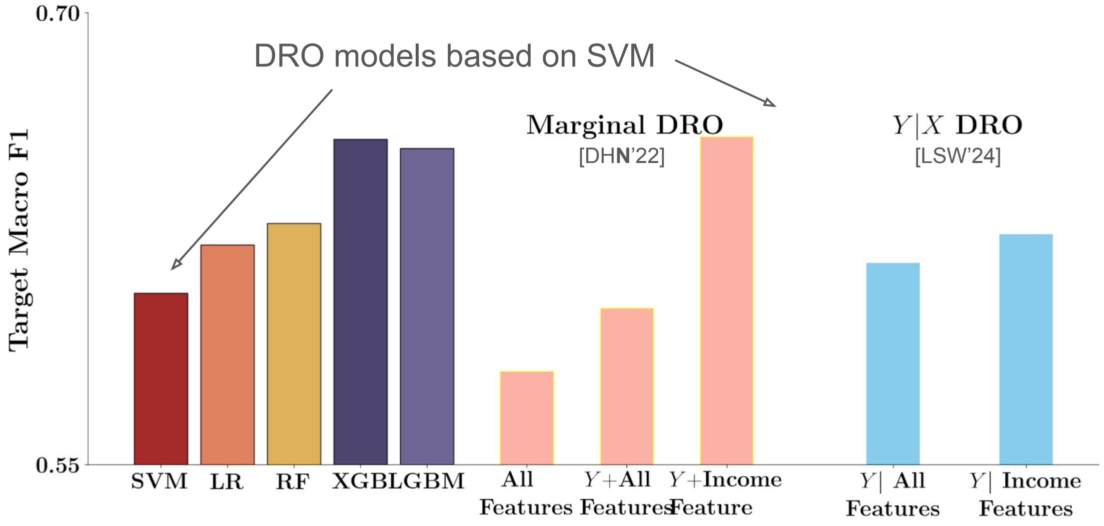


For conditional $Y|X$-DRO:
the h-net is a SVM.
the alpha-net is a two-layer MLP.

Duchi, Hashimoto, and **N**. Distributionally robust losses against mixture covariate shifts. Operations Research, 2022.

Inductive approach to ambiguity sets

- Y|X-shifts from NE -> LA; DRO over shifts on a subset of (X, Y)
- Variable selection for ambiguity set: Y | “income” suffers the largest shift
- Performance varies a lot over variables selected



Liu, Wang, Cui, **N.**, On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets, Short version in NeurIPS, 2023.
Duchi, Hashimoto, and **N**. Distributionally robust losses against mixture covariate shifts. Operations Research, 2022.

Takeaways so far

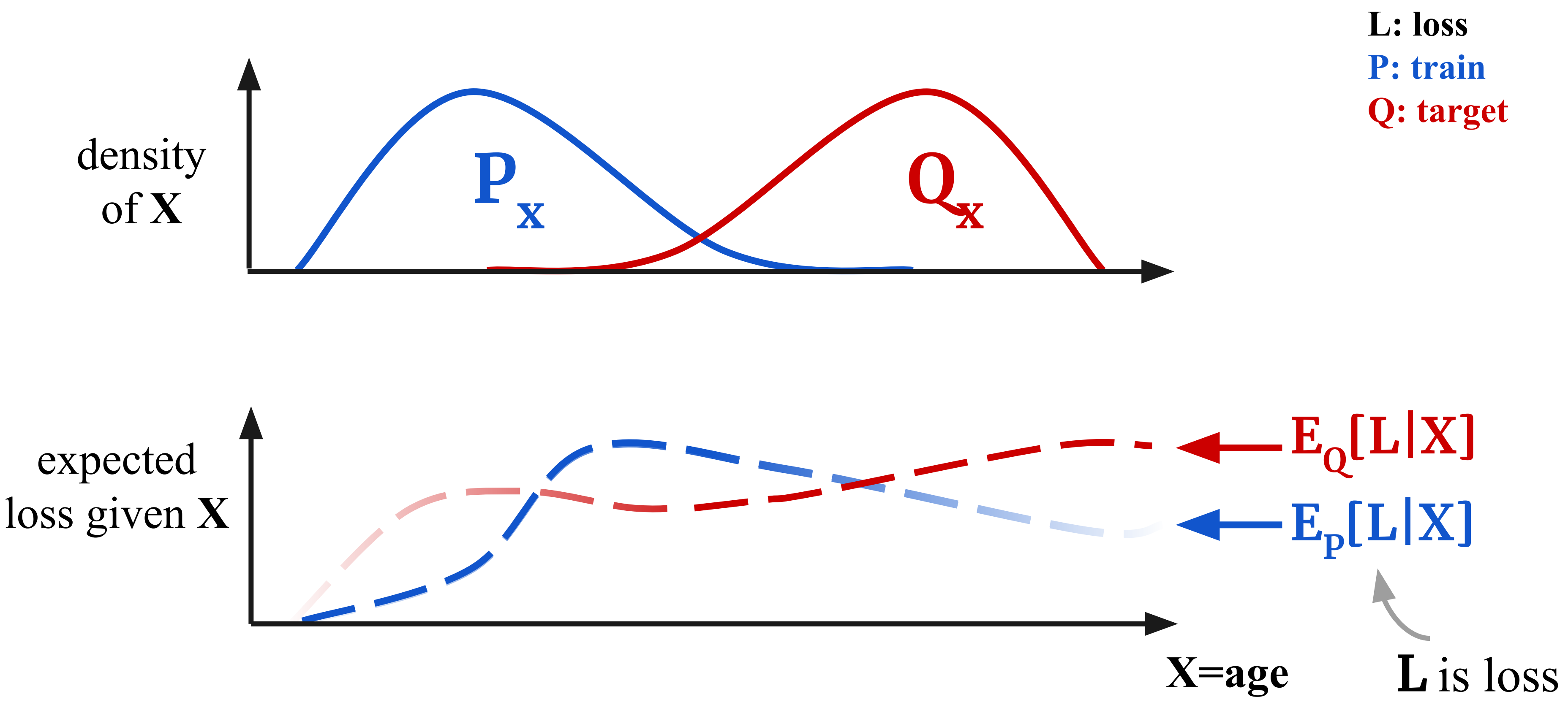
- Underlying model class (neural networks vs. tree ensembles) has first-order impact on robustness, yet frequently overlooked
- Ambiguity sets should be ***modeled.*** Move from deductive to inductive reasoning; do not optimize for math convenience
- Validation methods for hyperparameter selection matters a lot

Rest of the talk: a step toward an inductive modeling language for distribution shifts

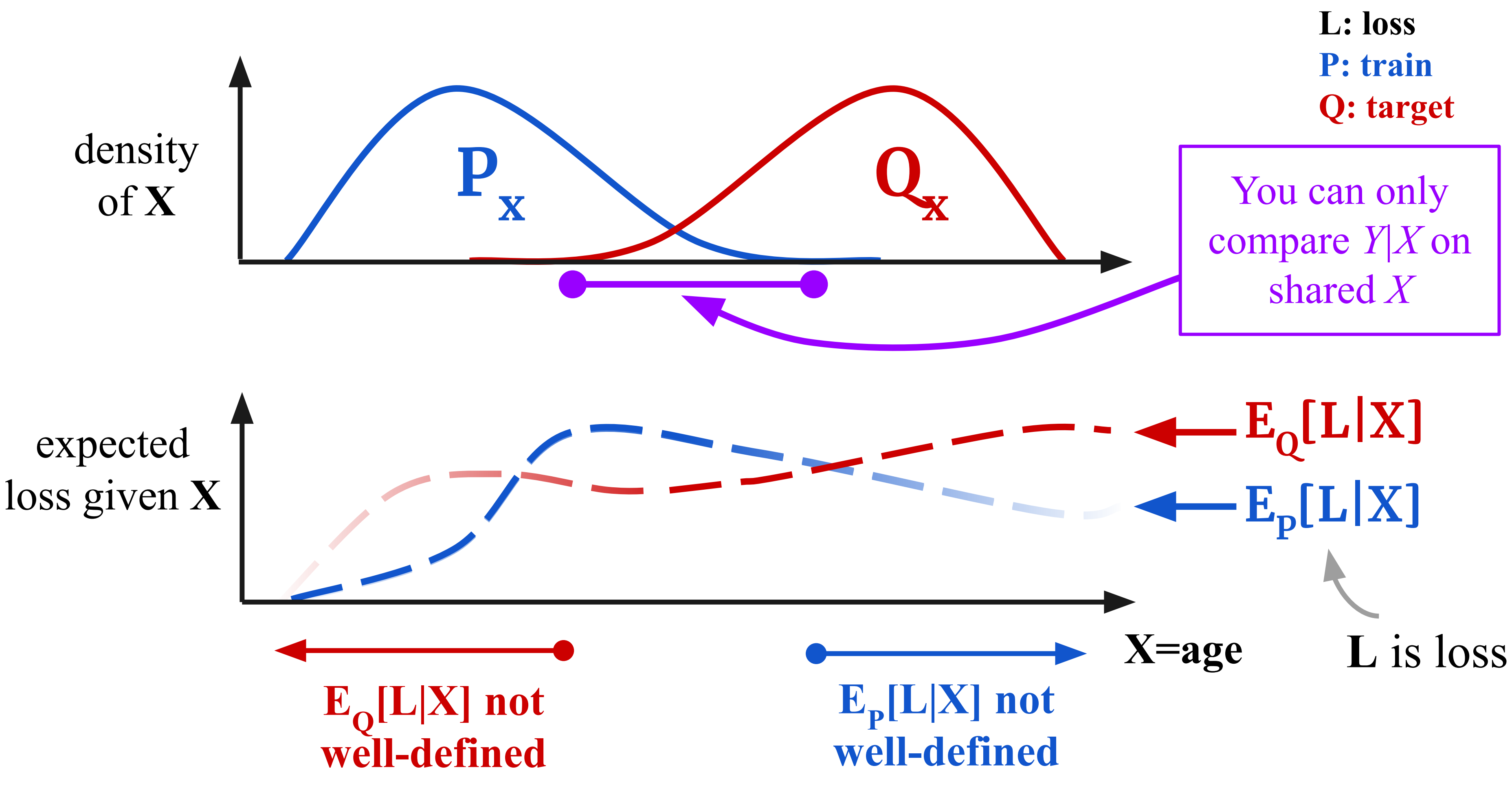
- Current ML community: out-of-distribution performance is worse than in-distribution performance,
  - i.e., **P: train** $\neq$ **Q: target**

- How do we attribute performance degradation? Not all shifts matter for model performance

- Different shifts warrant different interventions
  - Our goal today: differentiate X- vs. $Y|X$-shifts



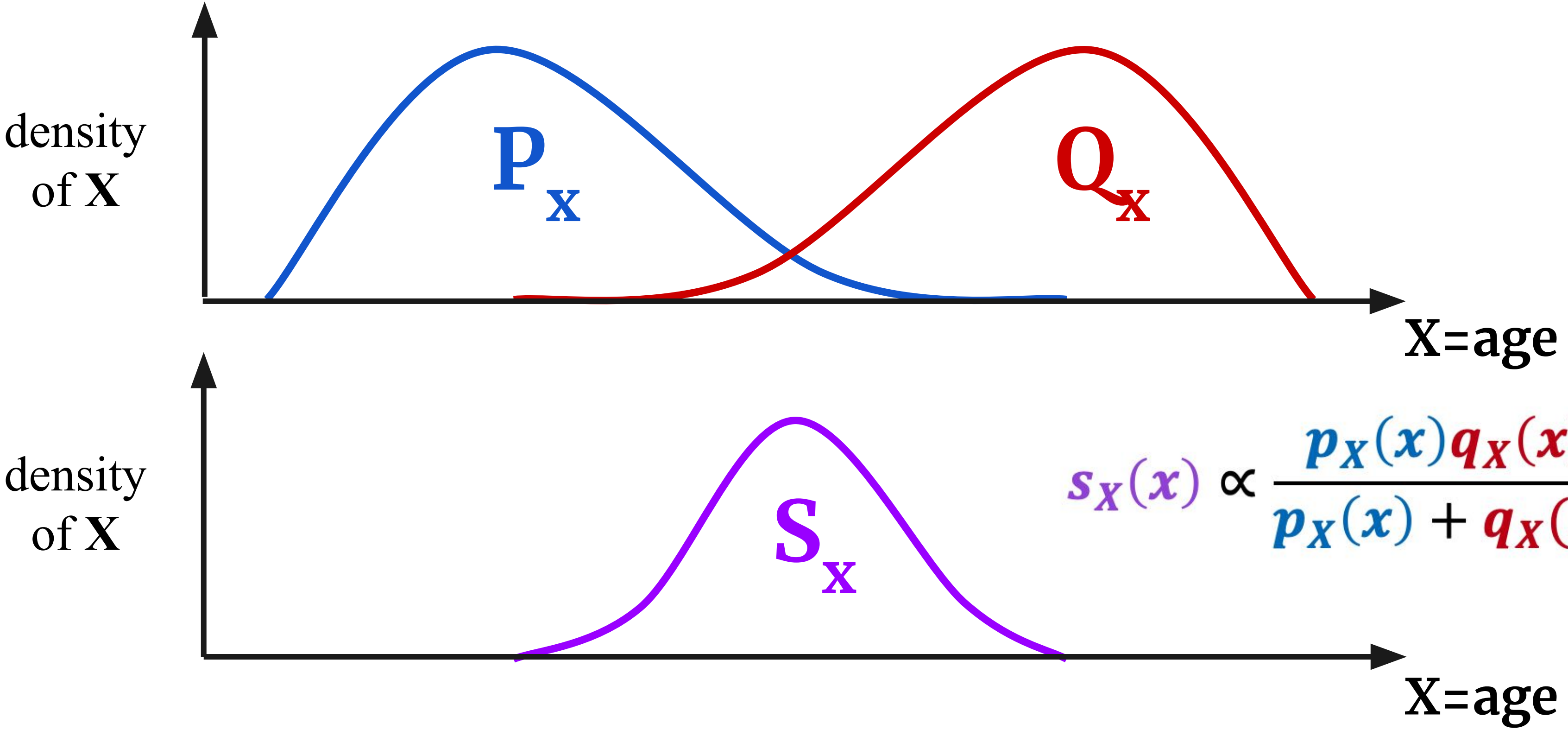
Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011



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



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

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

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

Performance on the training distribution

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

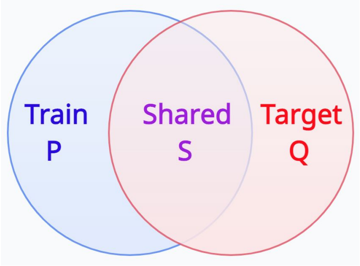
Performance on the target distribution

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

Decompose change in performance

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



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

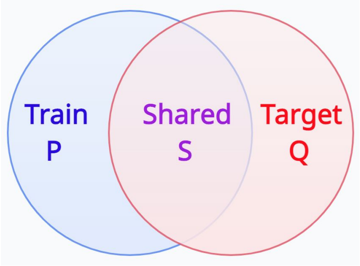
Diagnosis:
S has more X's that are harder to predict than **P**

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

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

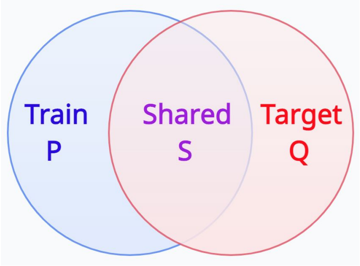
Potential interventions:

Re-collect data
or modify covariates

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:

$$\xrightarrow{X\text{ shift}} \qquad \overset{Y \mid X}{\downarrow}\ \text{shift}$$

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

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

Estimation

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

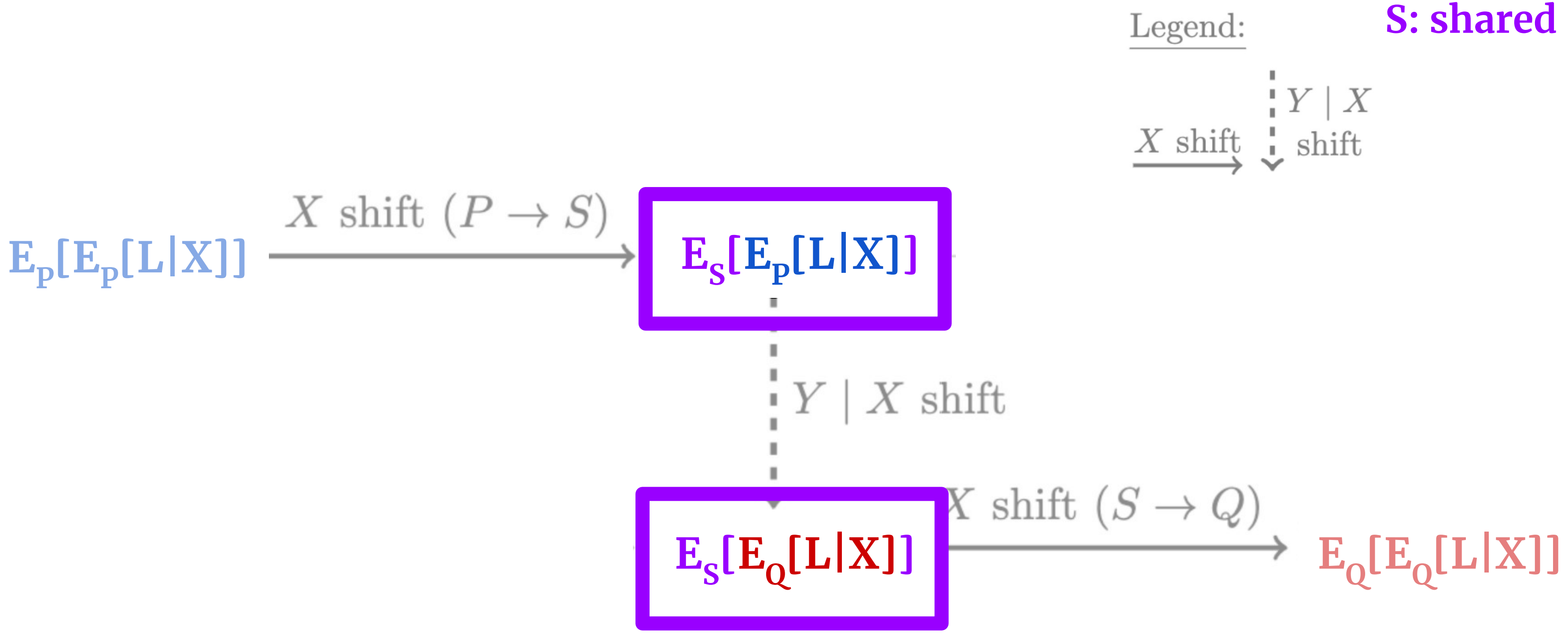
Legend:

$\xrightarrow{X \text{ shift}}$ $\quad$ $\downarrow$ $Y \mid X$ shift

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

Estimation

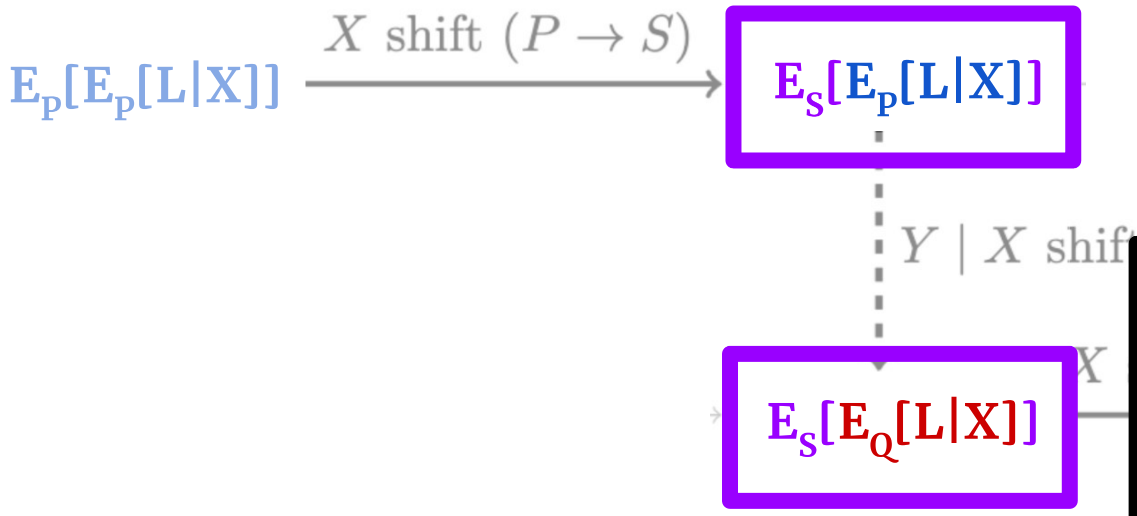
L: loss
P: train
Q: target
S: shared



How do you take expectations over S???

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

Legend:
$\xrightarrow{X \text{ shift}}$
$Y \mid X$ shift



Importance weighting!

Importance weights look like classifier probabilities

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

Reweight samples from P and Q into S using importance weighting.
The importance weights are

$$\frac{dS_X}{dP_X}(x) \propto \frac{q(x)}{p(x)+q(x)} \quad \text{and} \quad \frac{dS_X}{dQ_X}(x) \propto \frac{p(x)}{p(x)+q(x)}$$

Importance weights look like **classifier probabilities** of X being from P vs Q

Method

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

1. Train domain classifier to classify X as coming from P vs Q
2. Reweight losses from P and Q into S using class probabilities

Shared S inputs are those that can't be confidently classified as **P** vs **Q**

Confidence intervals

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

[Theorem: asymptotics] For a nonparametric classifier / reweighting that is asymptotically accurate, our estimator for $\theta_Q = E_S[E_Q[L|X]]$ is asymptotically normal

$$\sqrt{n}(\hat{\theta}_Q - \theta_Q) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\psi_Q(W)))$$

and we can estimate $\mathrm{Var}(\psi_Q(W))$ using plug-ins to calculate confidence intervals.

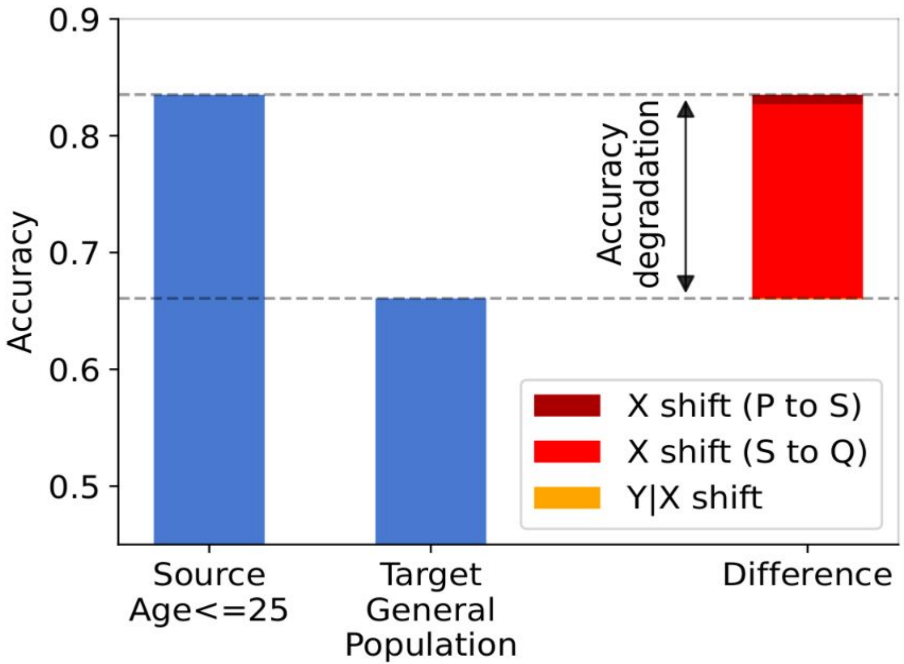
[Theorem: semiparametric efficiency] Our estimator gives the tightest possible confidence interval, achieving the lowest possible (asymptotic) variance

Cai, **N.**, and Yadlowsky, Diagnosing Model Performance Under Distribution Shift, 2023

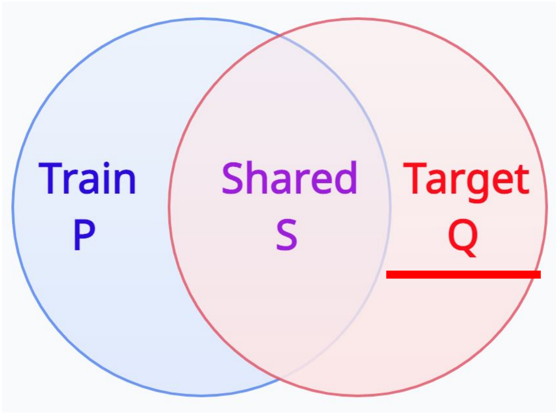
Employment prediction case study

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

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



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

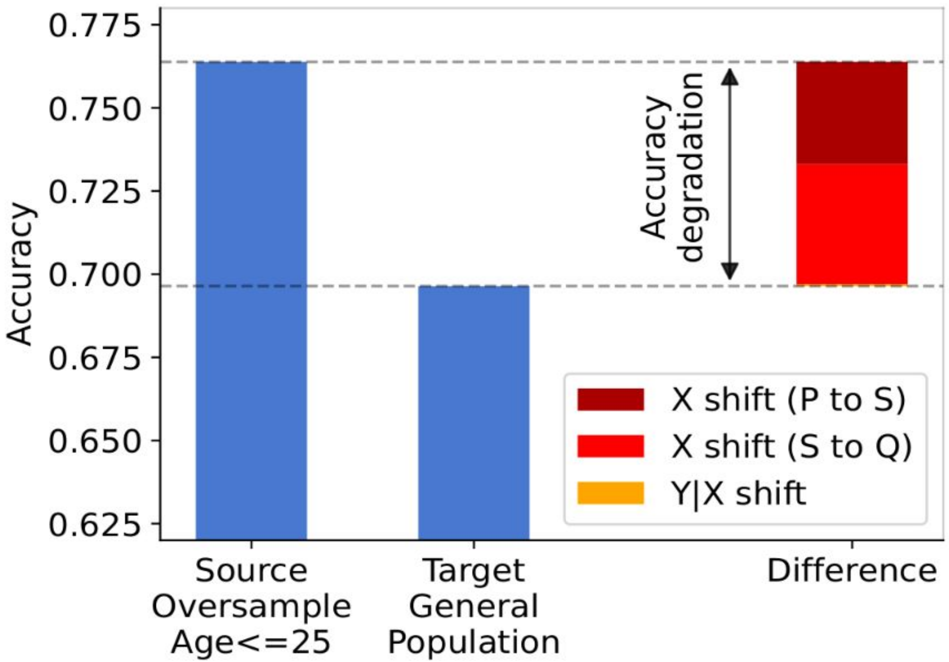


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

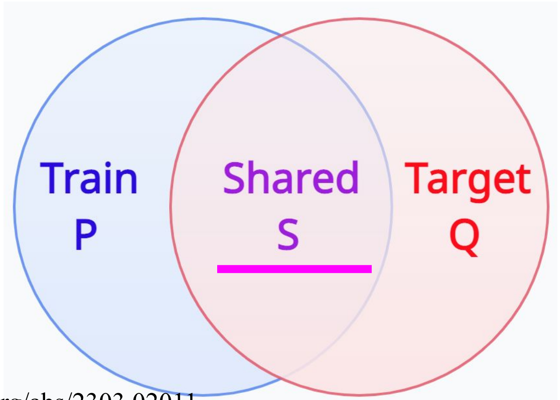
Employment prediction case study

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** ⟶ **S**), suggesting domain adaptation may be effective

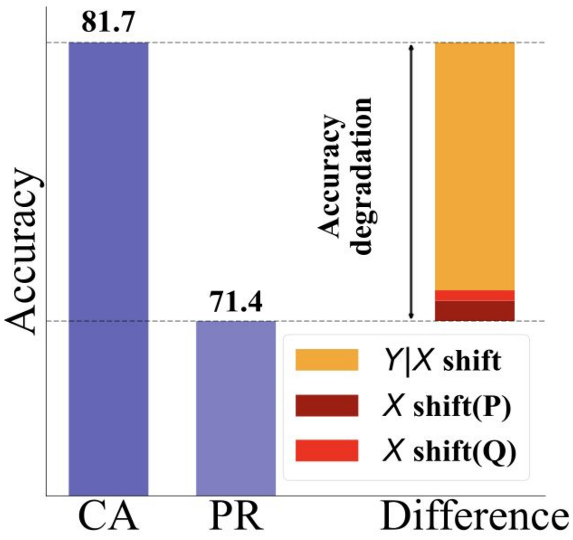


Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

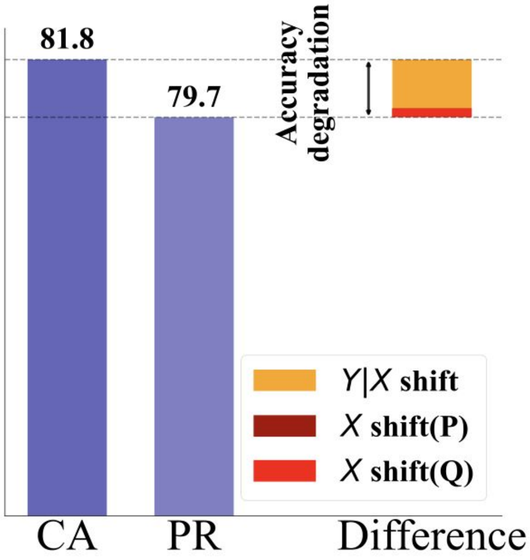
Better data can be effective

[$Y|X$ shift] **P:** California (CA), **Q:** Puerto Rico (PR)

No language features



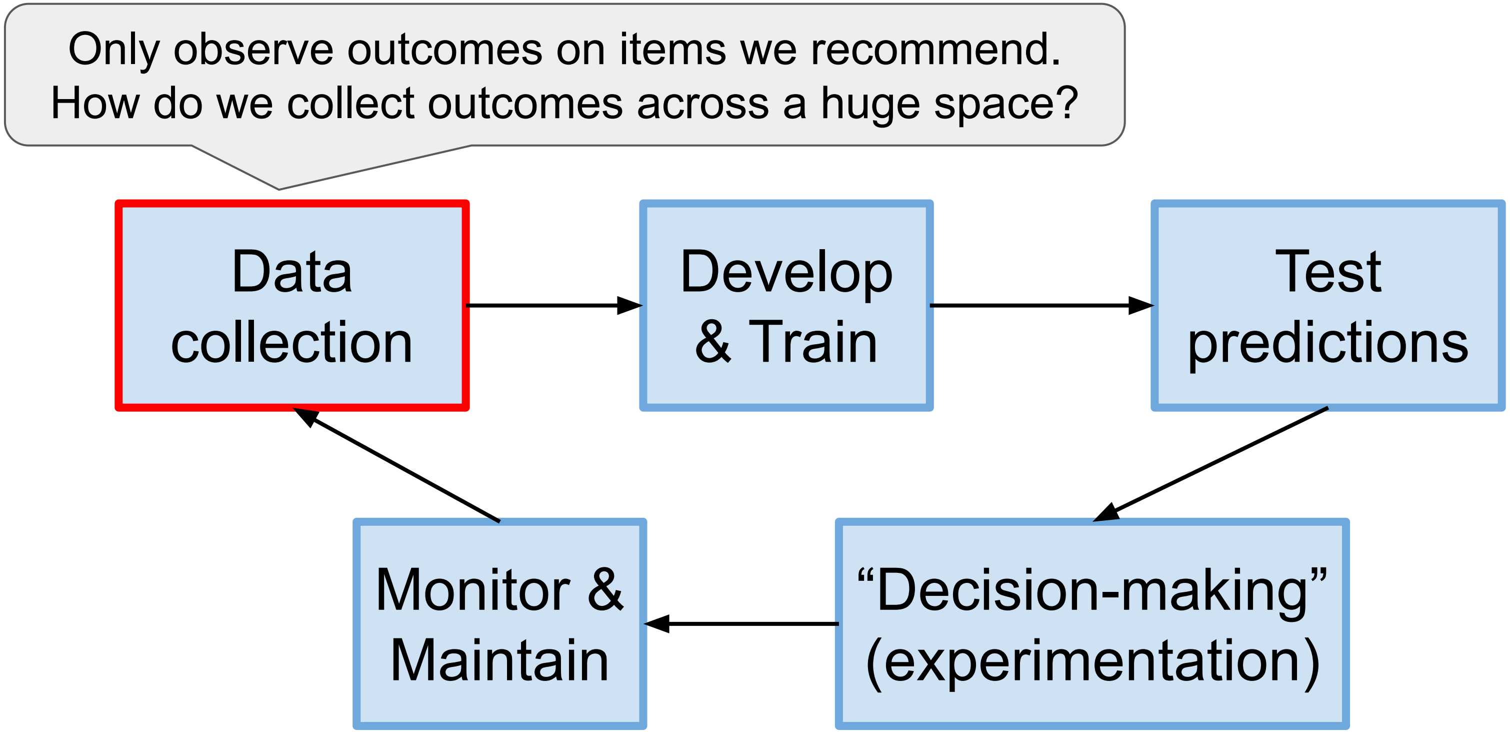
With language features



CA model does not use language.

$Y|X$ shift because of missing covariate: language affects outcome → better performance in PR

A methodological bottleneck: uncertainty



Zhang, Cai, N. and Russo. Posterior Sampling via Autoregressive Generation. Work in progress.

Distribution Shift Decomposition (DISDE)

- Diagnostic for understanding why performance dropped in terms of X vs $Y|X$ shift
- Can help articulate modeling assumptions + data collection

We need a modeling language for a data-centric view of AI

- Develop modeling tools in an **application-specific** manner!
- Top of mind: resolving methodological bottlenecks in uncertainty quantification

Cai, **N.**, and Yadlowsky, Diagnosing Model Performance Under Distribution Shift,
Major revision in Operations Research, Conference version appeared in Foundations of Responsible Computing 2022, https://github.com/namkoong-lab/disde
Liu, Wang, Cui, and **N.**, On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets, Conference version in NeurIPS 2023, https://github.com/namkoong-lab/whyshift

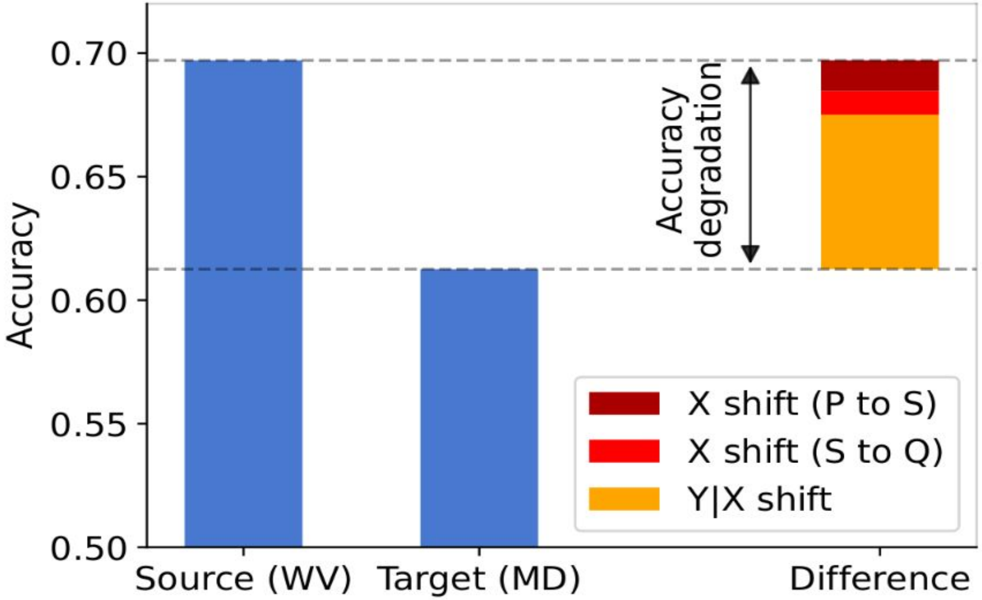
What's next?

- Industrial applications
  - Governance: Scale minimal requirements at the company level
  - Compliance: Design best practices for "due diligence" in responsible AI
  - Engineering constraints: Design algorithms under infrastructural constraints
- Methodological bottlenecks: uncertainty quantification, objective and actions defined on different timescales
- Top of mind: Measurement and mitigation in shifting AI paradigms

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.

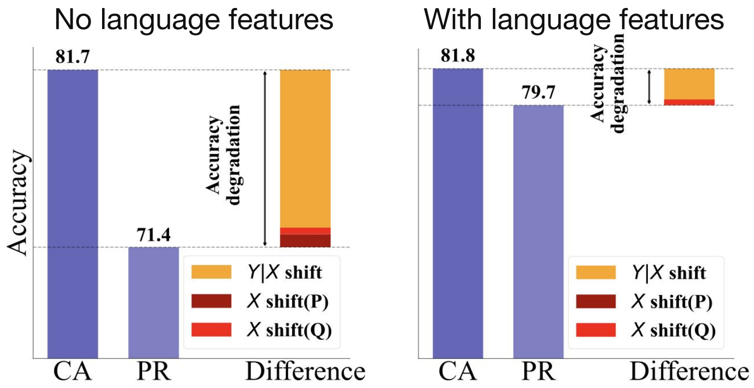
$Y|X$ shift because of missing covariate: education affects employment

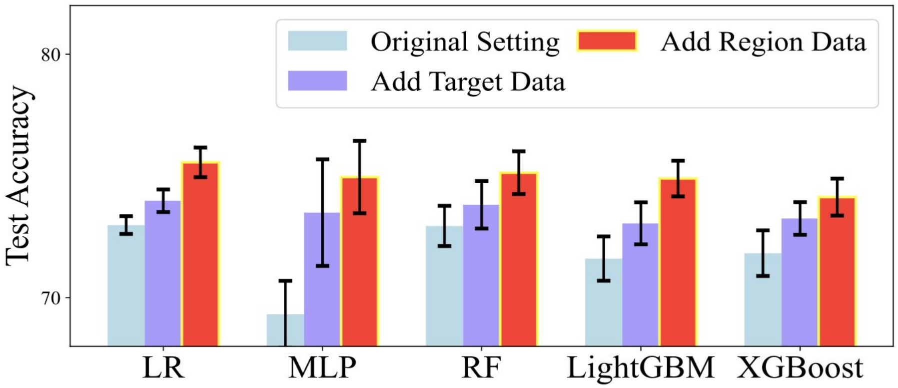
Diagnosing Model Performance Under Distribution Shift https://github.com/namkoong-lab/disde https://arxiv.org/abs/2303.02011

Better data can be more effective than better algorithms!

[Y|X shift] **P:** California (CA), **Q:** Puerto Rico (PR)

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





collecting better features

collecting better target data

Appendix: Variables in Linear Analysis

•

Type	Name	Definition
Model Class $X_{i,j,s}$	Tree	A dummy variable that takes value one if the base learner of the model configuration is tree-structure
	MLP	A dummy variable that takes value one if the base learner of the model configuration is MLP
Ambiguity Set $D_{i,j,s,t}$	Wasserstein	A dummy variable that takes value one if the model configuration belongs to DRO and uses Wasserstein-type metric
	Chi-squared	A dummy variable that takes value one if the model configuration belongs to DRO and uses χ^2-divergence metric
	Kullback-Leibler	A dummy variable that takes value one if the model configuration belongs to DRO and uses KL-divergence metric
	Total Variation	A dummy variable that takes value one if the model configuration belongs to DRO and uses TV-distance metric
	OT-Discrepancy	A dummy variable that takes value one if the model configuration belongs to DRO and uses the optimal transport-discrepancy with conditional moment constraints
	Radius	The rescaled ambiguity size if the model configuration belongs to DRO and equal to zero if the model configuration does not belong to DRO
Shift Pattern $Z_{i,j}$	$Y\|X$-ratio	The $Y\|X$-shift percentage calculated by `DISDE` from the source domain to the target domain
Validation Type $V_{i,j}$	Worst	A dummy variable that takes value one if the best configuration is obtained through the largest accuracy from the worst target domain
	Average	A dummy variable that takes value one if the best configuration is obtained through the largest accuracy from the average-case target domain

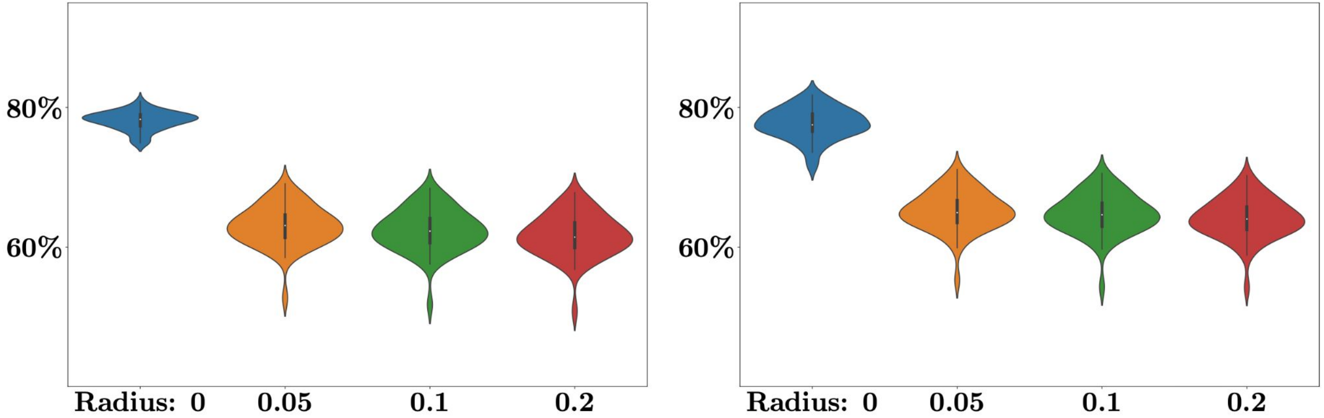
Appendix: Configurations

- Algorithms evaluated in our empirical study:

1. *Basic learners*: Logistic Regression (LR), SVM, fully-connected neural networks (MLP) with standard ERM optimization;
2. *Tree-based learners*: Random Forest (RF) [8], GBM [26], LightGBM [19], XGBoost [9];
3. *Imbalanced learning algorithms*: SUBY, RWY, SUBG, RWG [17], which reweight or subsample data to balance the samples of different classes (Y) or different demographic groups (G);
4. *Fairness-enhancing algorithms*: In-processing methods [4] with demographic parity, equal opportunity, and error parity as constraints, and post-processing methods [15] with exponential and threshold controls;
5. *Linear-DRO algorithms*: Distributionally robust optimization (DRO) methods based on linear SVM using different uncertainty sets, including CVaR-DRO [28], χ^2-DRO [12], TV-DRO [18], KL-DRO [16], Wasserstein-DRO [6], Augmented Wasserstein-DRO [30], Satisificing Wasserstein-DRO [22], Sinkhorn-DRO [31], Holistic-DRO [5], Unified-DRO (with L_2-norm) [7], and Unified-DRO (with L_{inf}-norm) [7];
6. *NN-DRO algorithms*: DRO methods based on MLP using different uncertainty sets, including CVaR-DRO (NN) and χ^2-DRO (NN) with fast implementation [21], CVaR-DORO (NN) and χ^2-DORO (NN) that are designed for outlier robustness [34].

Worst-case Distribution Analysis

- Misalignment between worst-case distributions and target distributions
  - when we use the worst-case distribution of KL-DRO to train tree-based methods, their target accuracies **even drop a lot**



(a) ACS Income, LightGBM (b) ACS Income, XGB

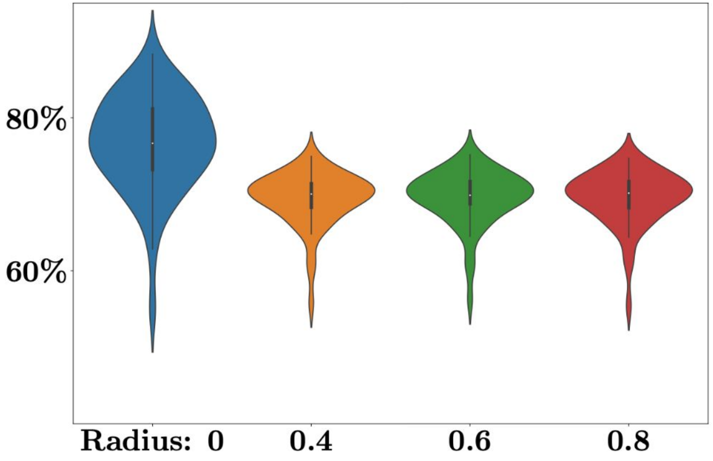
Worst-case Distribution Analysis

- Recall that KL-DRO improves the worst target performance on ACS Pub.Cov

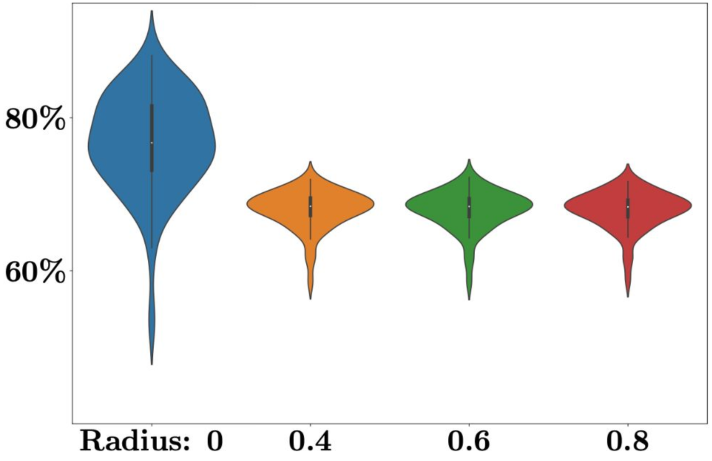
		Dependent variable: Accuracy					
		Setup 1: one-to-one			Setup 2: one-to-worst		
Variable Name		All	Pubcov	Income	All	Pubcov	Income
Model Class	Tree	.0095***	.0047***	.0024***	−.0113***	−.0335***	−.0035
		(.0027)	(.0012)	(.0009)	(.0023)	(.0032)	(.0031)
	MLP	.0036*	−.0142***	.0081***	−.0355***	−.0363***	−.0296***
		(.0019)	(.0009)	(.0007)	(.0024)	(.0040)	(.0033)
Ambiguity Set	Wasserstein	−.0048*	−.0077***	−.0040***	−.0469***	−.0274***	.0002
		(.0027)	(.0012)	(.0009)	(.0035)	(.0066)	(.0050)
	Chi-squared	.0011	.0022*	.0011	−.0015	.0170***	−.0054
		(.0025)	(.0011)	(.0008)	(.0026)	(.0035)	(.0036)
	Kullback-Leibler	−.0024	.0008	−.0008	−.0062**	.0643***	−.0773***
		(.0025)	(.0012)	(.0009)	(.0025)	(.0034)	(.0035)

Worst-case Distribution Analysis

- But still conservative!
  - We train LightGBM and XGBoost models on the worst-case distribution of KL-DRO
  - The worst-case performance over 50 target states improves
  - But the overall target performances drop a lot!



(c) ACS Pub.Cov, LightGBM



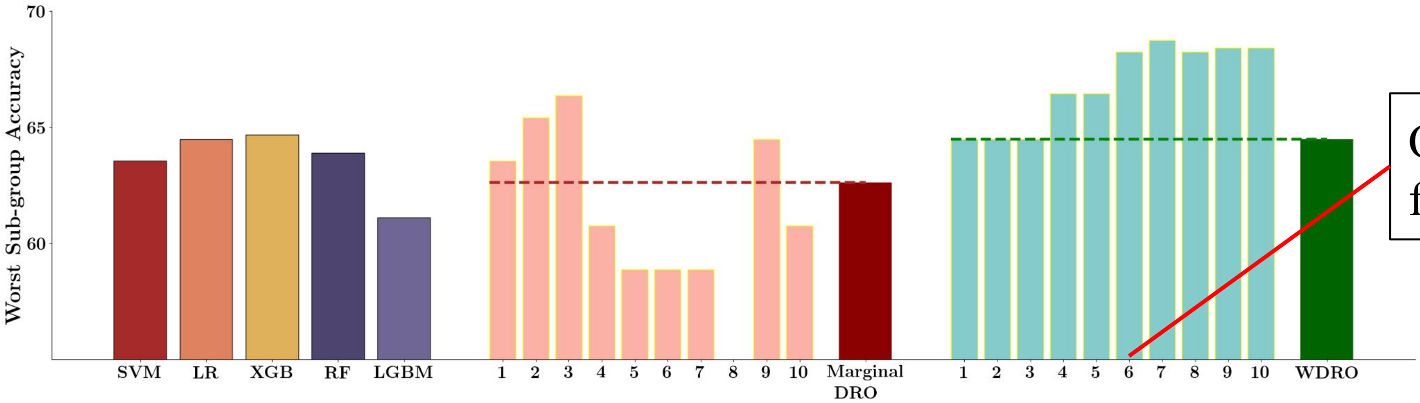
(d) ACS Pub.Cov, XGB

Algorithmic Intervention: design better ambiguity sets?

Task: income prediction
Source: Age < 25
Target: Age ≥ 25

Case study on covariate shifts:

- for Marginal-DRO and Wasserstein DRO
- only perturb the covariates whose distributions shifte a lot among age groups
  - pick the Top-shifted covariates
- measure the worst sub-group accuracy (age groups: [20,25), [25,30), …, [75,100))



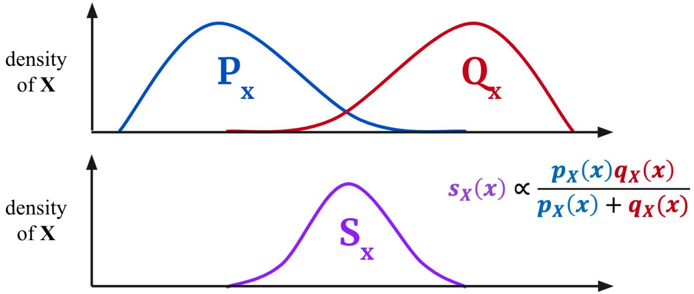
Only perturb top-i features at time i

Non-Algorithmic Intervention: collect better features/data?

- Region Analysis on $Y|X$-shift

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

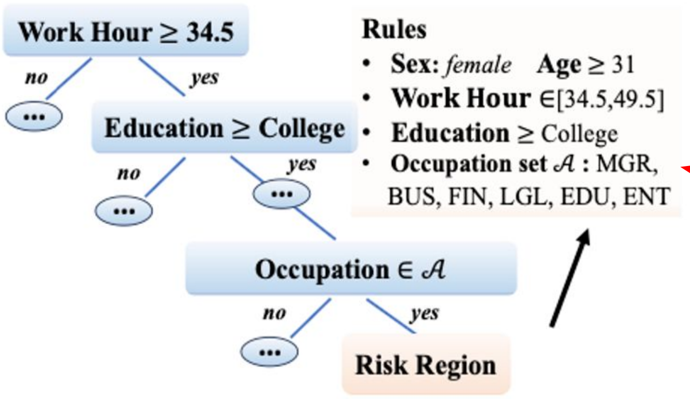


Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023, November). On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Non-Algorithmic Intervention: collect better features/data?

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!

WhyShift

arxiv

github

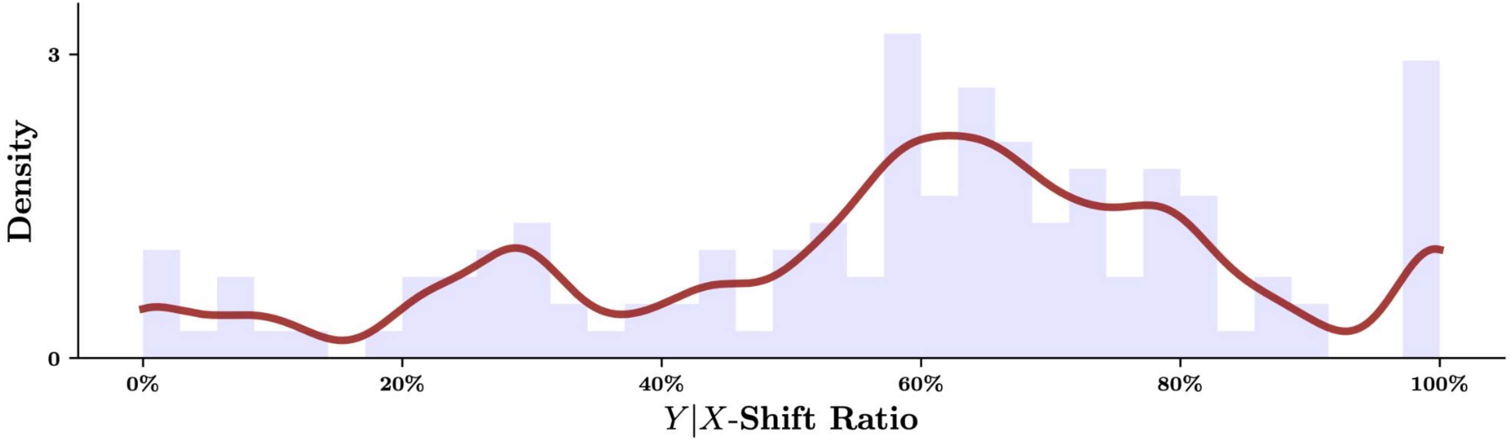
whyshift 0.1.3

```
pip install whyshift
```

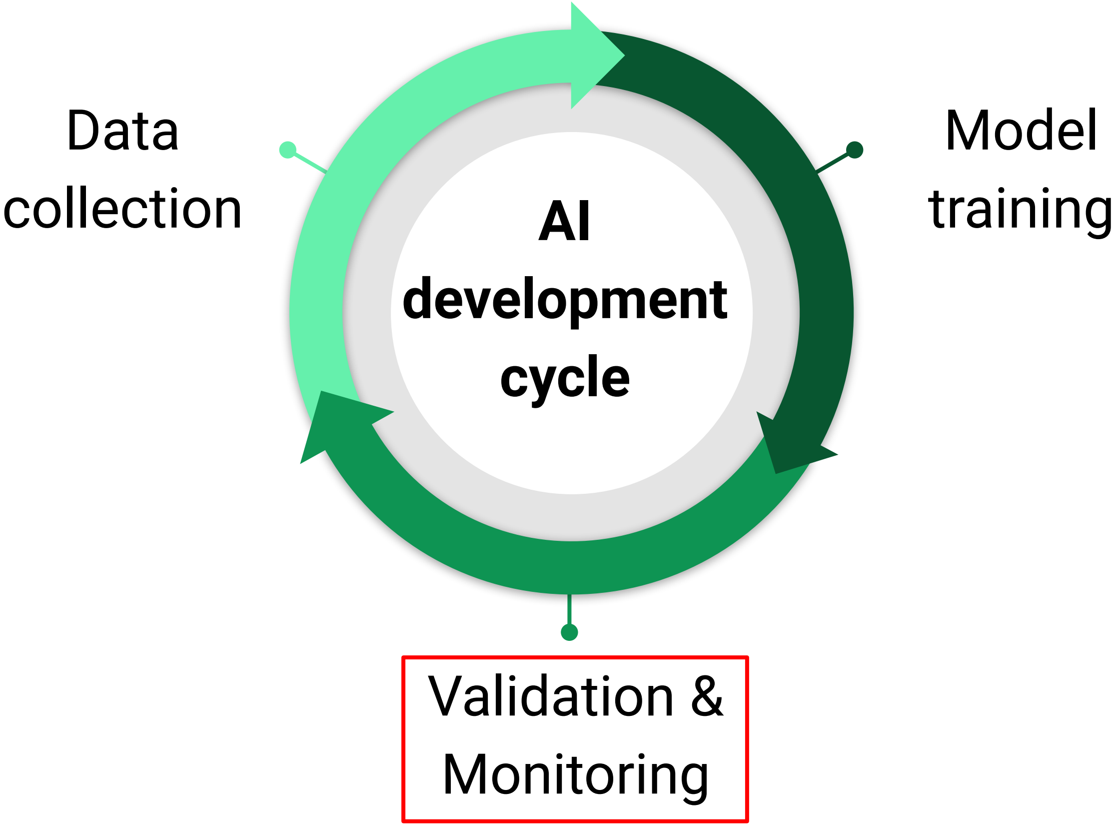
https://github.com/namkoong-lab/whyshift

- Initial conjecture: Y|X-shifts are more prominent than X-shifts in practice

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



AI pipeline


Data
collection
Model
training
AI
development
cycle
Validation &
Monitoring

Decompose change in performance

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

Legend:

$$\xrightarrow{X \text{ shift}} \qquad \begin{tikzcd} {} \arrow[d, dotted, "Y \mid X \text{ shift}"] \\ {} \end{tikzcd}$$

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

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

Estimation

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

Legend:

$\xrightarrow{X \text{ shift}}$ $\quad$ $Y \mid X$ shift (dashed downward arrow)

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

Estimation

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

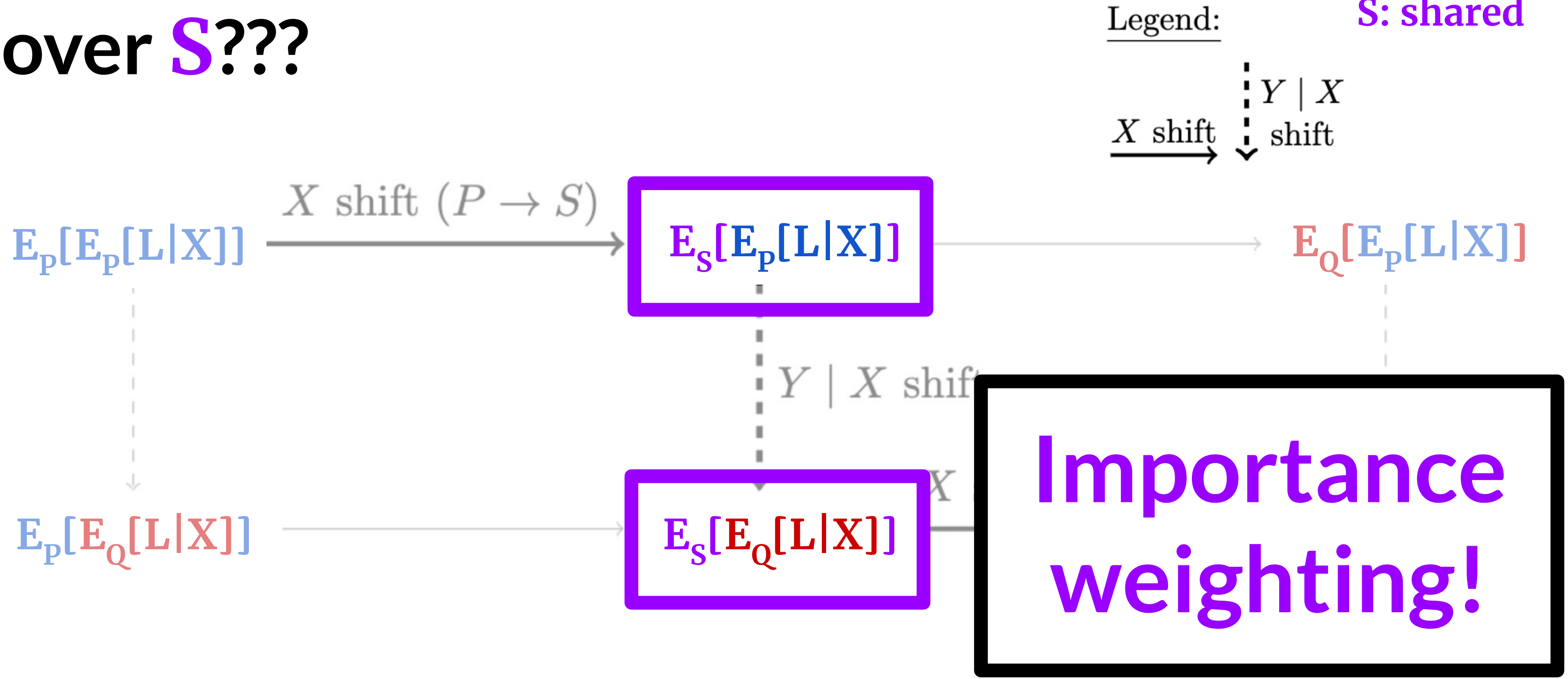
Legend:

$$\xrightarrow{X\ \text{shift}} \qquad \Big\downarrow Y \mid X \text{ shift}$$

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

How do you take expectations over S???

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



More description of datasets and shifts. Outcomes etc.

In spirit, describe the dro that actually works

Expand on why can’t we just do regular ml benchmarking on distribution shifts