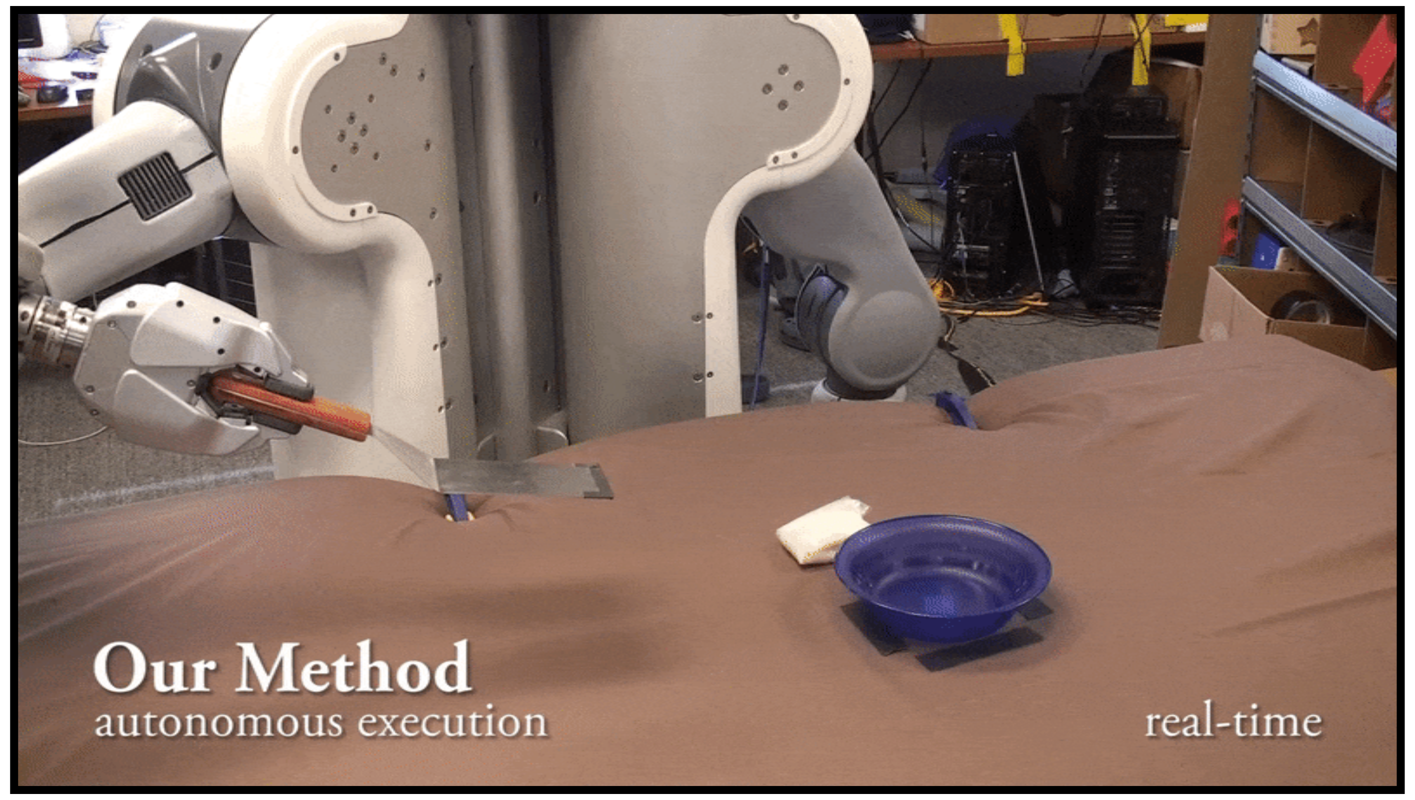
Distribution Shift as Underspecification

And What We Might Do About It

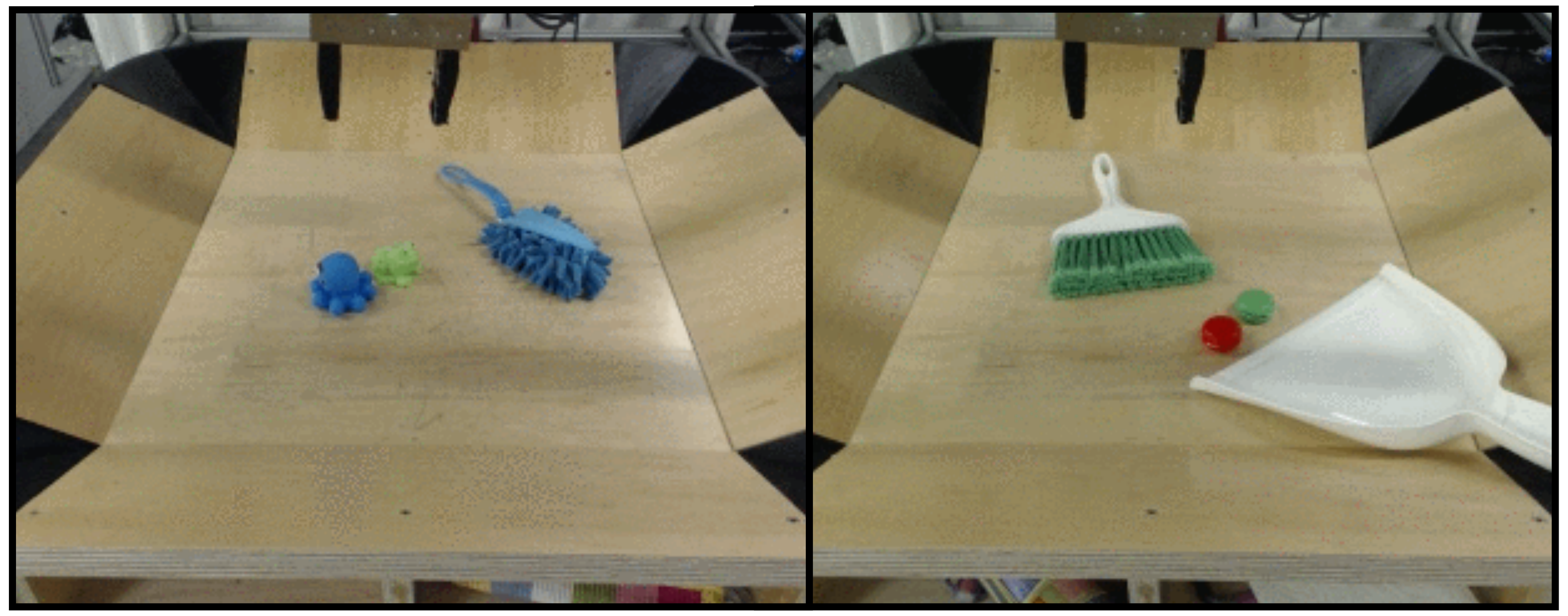
Chelsea Finn

Stanford

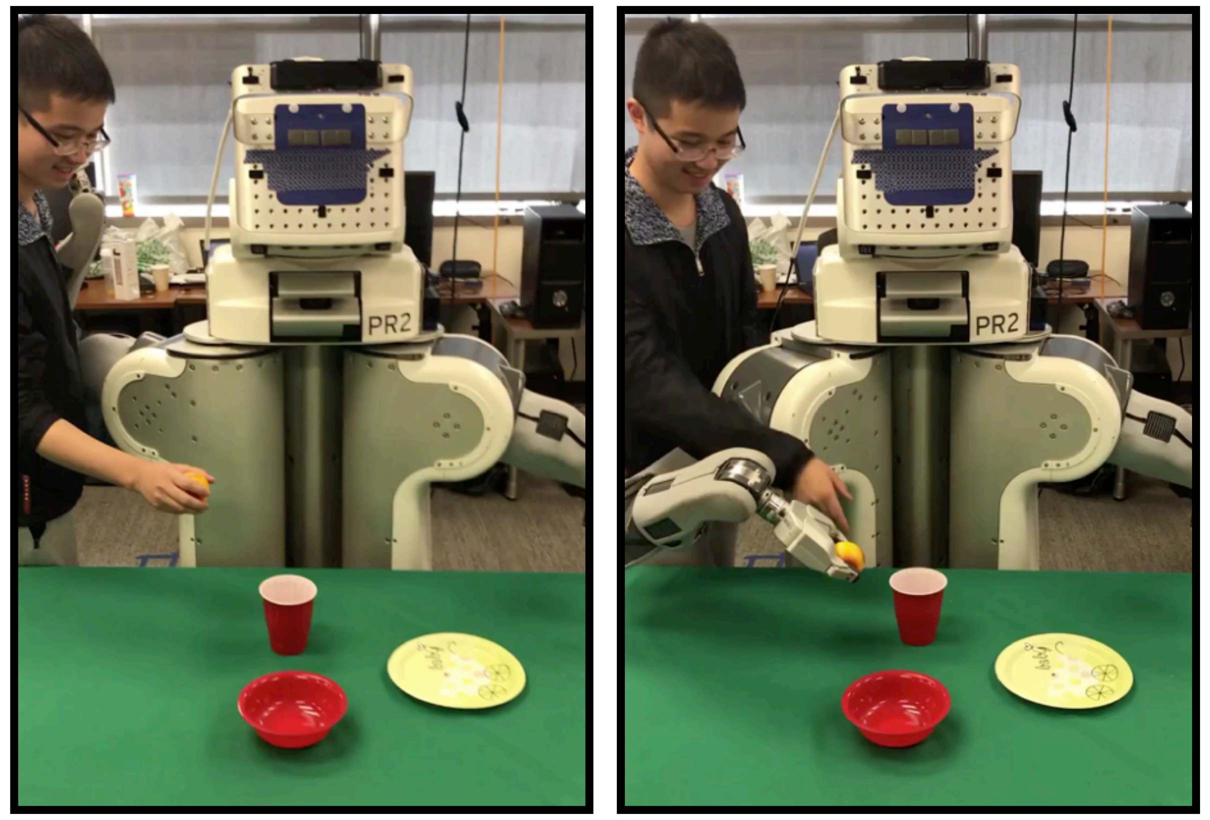
Can robots develop broadly intelligent behavior through learning & interaction?



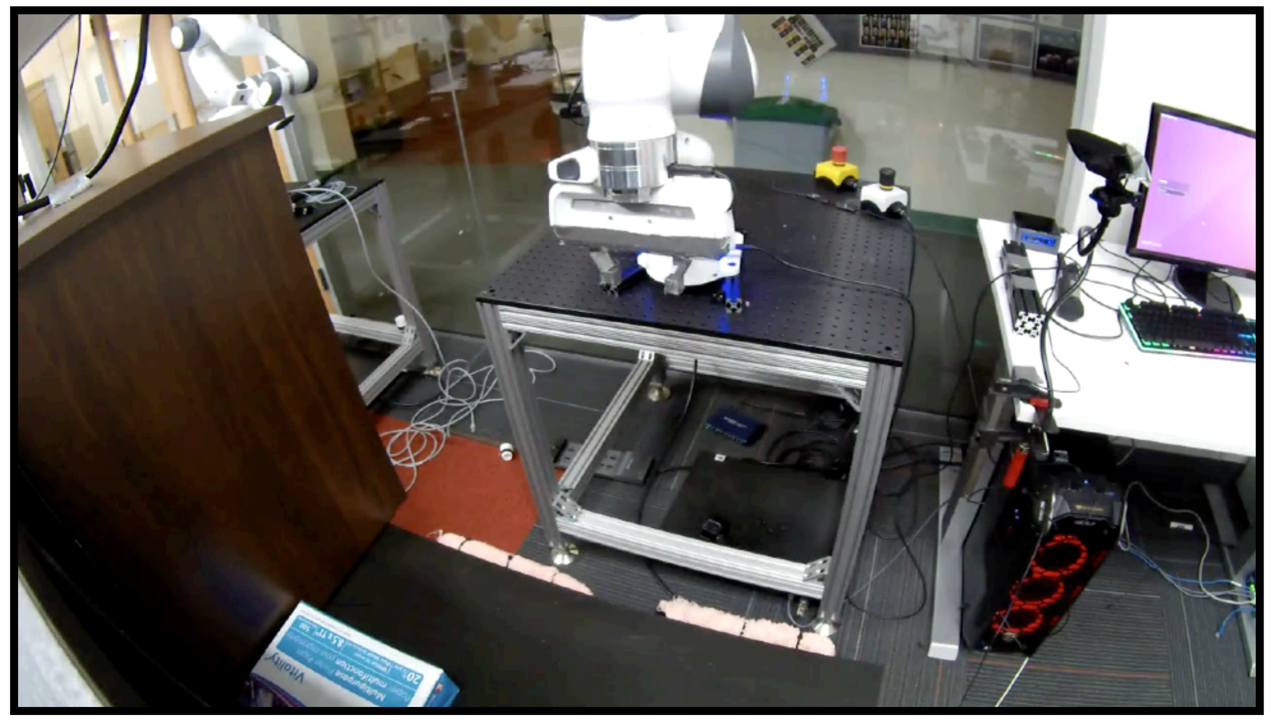
Finn, Tan, Duan, Darrell, Levine, Abbeel. ICRA '16



Xie, Ebert, Levine, Finn, RSS '19

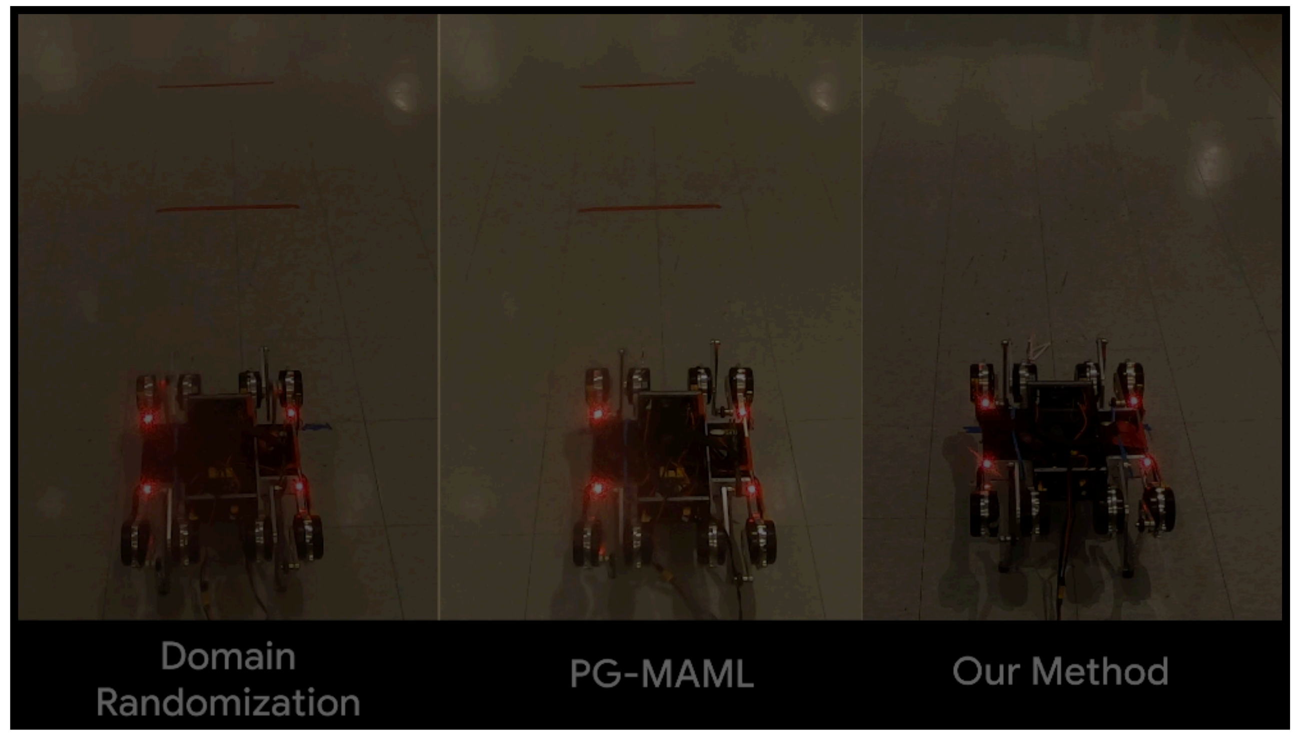


Yu*, Finn*, Xie, Dasari, Zhang, Abbeel, Levine, RSS '18



Chen*, Nam*, Nair*, Finn. ICRA '21

Nair, Rajeswaran, Kumar, Finn, Gupta. arXiv '22



Song, Yang, Choromanski, Caluwaerts, Gao, Finn, Tan. IROS '20

Machine learning works

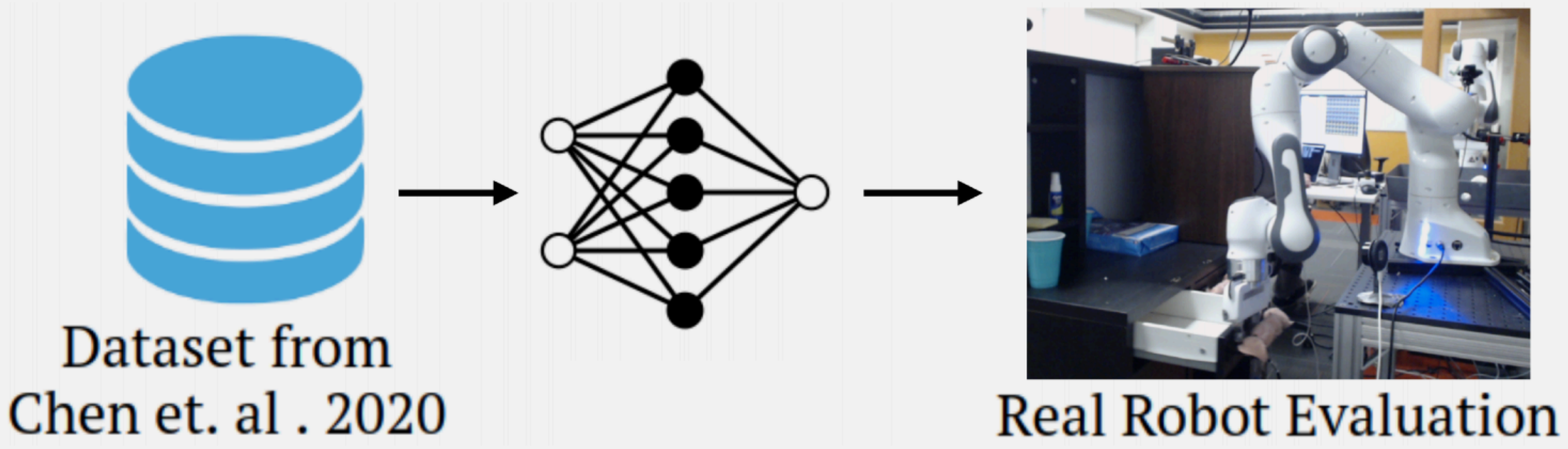
on the training data distribution

Core assumption

$$P_{\text{train}} = P_{\text{test}}$$

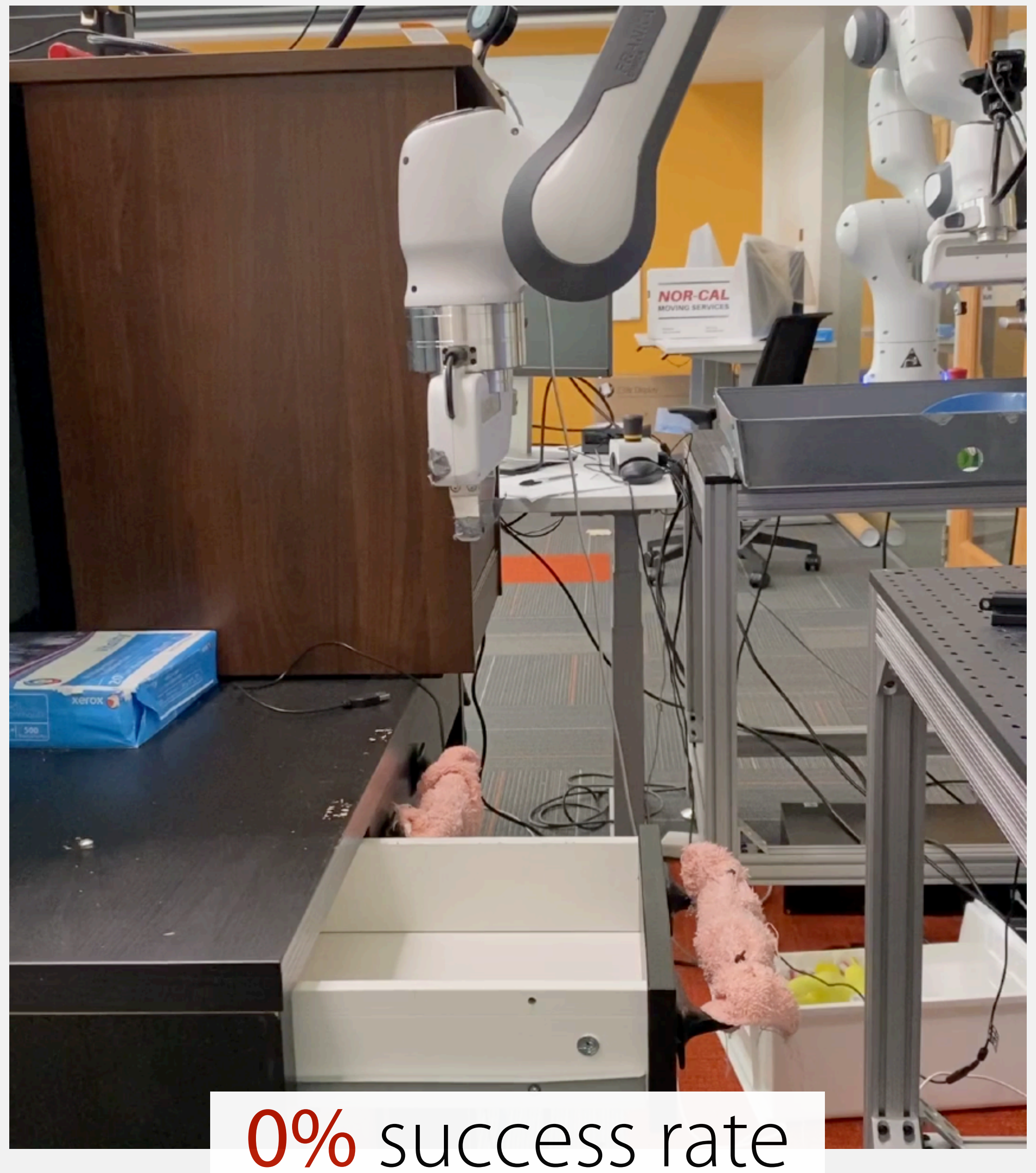
Examples of distribution shift: **offline RL** and **temporal shifts**

RL from offline datasets



Distribution shift between **policy in the dataset** and the **policy being optimized**.

If you don't account for this shift:

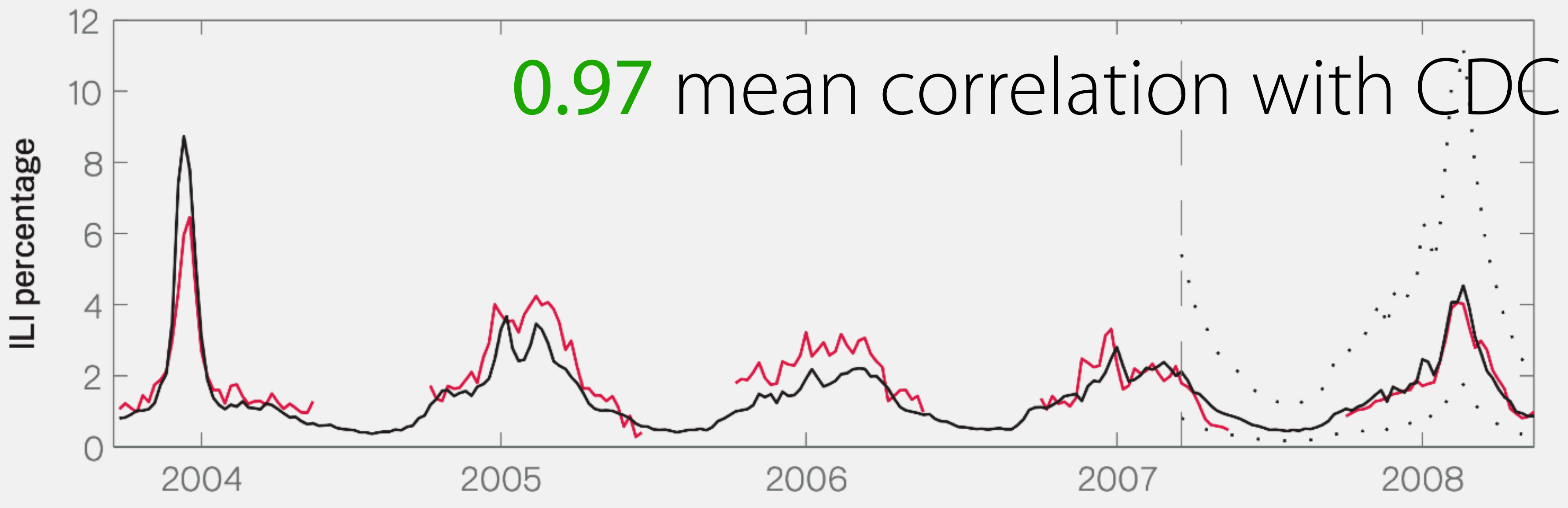


0% success rate

Shift over time

Predicting flu incidence from search queries

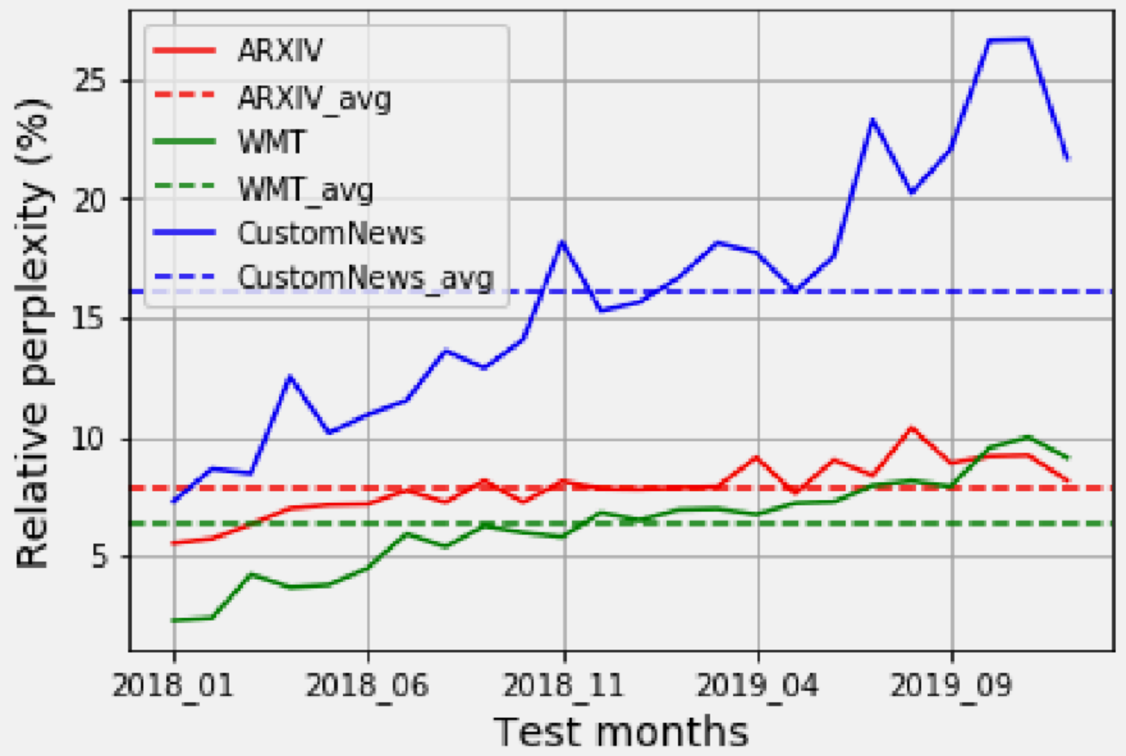
0.97 mean correlation with CDC data.



Ginsberg et al. *Detecting influenza epidemics using search engine query data*. Nature '09

Feb 2013: predicting **double** the incidence

Language model perplexity over time.



Lazaridou et al. *Pitfalls of Static Language Modeling*. '21

Examples of distribution shift: **domains** & **subpopulations**

Online content moderation (Borkan et al. 2019)

Comment: "I doubt that anyone cares whether you believe it or not" ⟶ toxic / not toxic

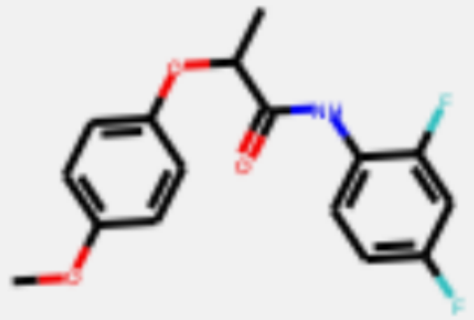
92.2% average test accuracy

Demographic	Test accuracy on non-toxic comments
Male	87.3 (0.7)
Female	89.0 (0.6)
LGBTQ	74.6 (0.5)
Christian	92.1 (0.2)
Muslim	80.9 (1.0)
Other religions	86.1 (0.1)
Black	**69.2** (1.3)
White	71.2 (1.4)

69.2% on non-toxic comments mentioning Black demographic

Molecular Property Prediction (Hu et al. 2020)

Molecule:



⟶ (0,1,1,0,0,..)

biological activity prediction

34.4% average precision on test molecules from training scaffolds

26.8% average precision on test molecules from held-out scaffolds

WILDS

WILDS has **10 datasets** with distribution shift, ranging from ecological conservation to medical imaging.

WILDS 2.0 adds unlabeled data for **8 datasets**.

Pang Wei Koh

Shiori Sagawa

Koh*, Sagawa*, Marklund, Xie, Zhang, Balsubramani, Hu, Yasunaga, Phillips, Gao, Lee, David, Stavness, Guo, Earnshaw, Haque, Beery, Leskovec, Kundaje, Pierson, Levine, Finn, Liang. **WILDS: A Benchmark of in-the-Wild Distribution Shifts**. ICML 2021.

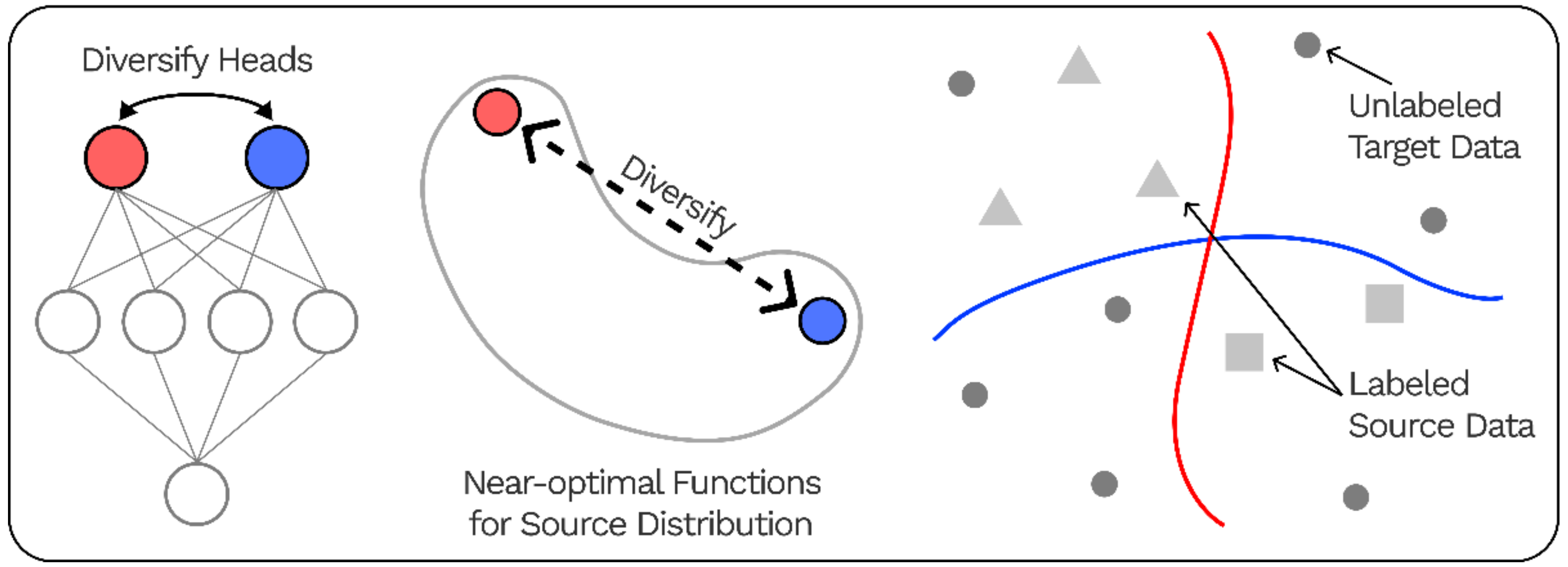
wilds.stanford.edu

Different kinds of distribution shift

Covariate shift	Change in $p(x)$	(includes domain shift, subpopulation shift)
Label shift	Change in $p(y)$	
Concept shift	Change in $p(y \mid x)$	

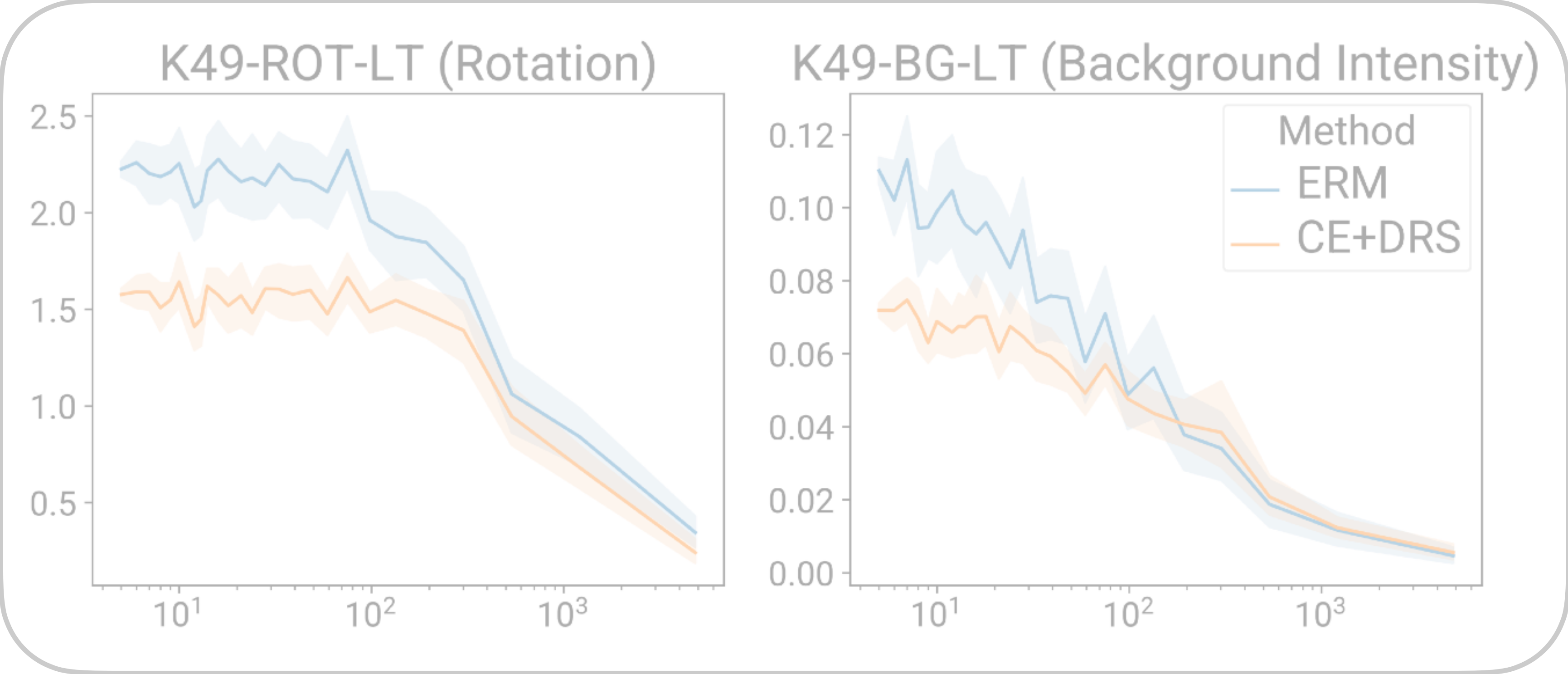
Outline

Addressing extreme covariate shift
via diverse ensembles



for supervised learning & reinforcement learning

Addressing label shift
via invariance transfer



for long-tailed image classification

A couple existing approaches for tackling covariate shift

Data rebalancing

Key idea: upweight or upsample underrepresented datapoints

- distributional robust optimization (group DRO, joint DRO)
- uniform class resampling
- learning from failure (LfF)
- just train twice (JTT)

Domain invariance

Key idea: learn representations that are invariant to domain

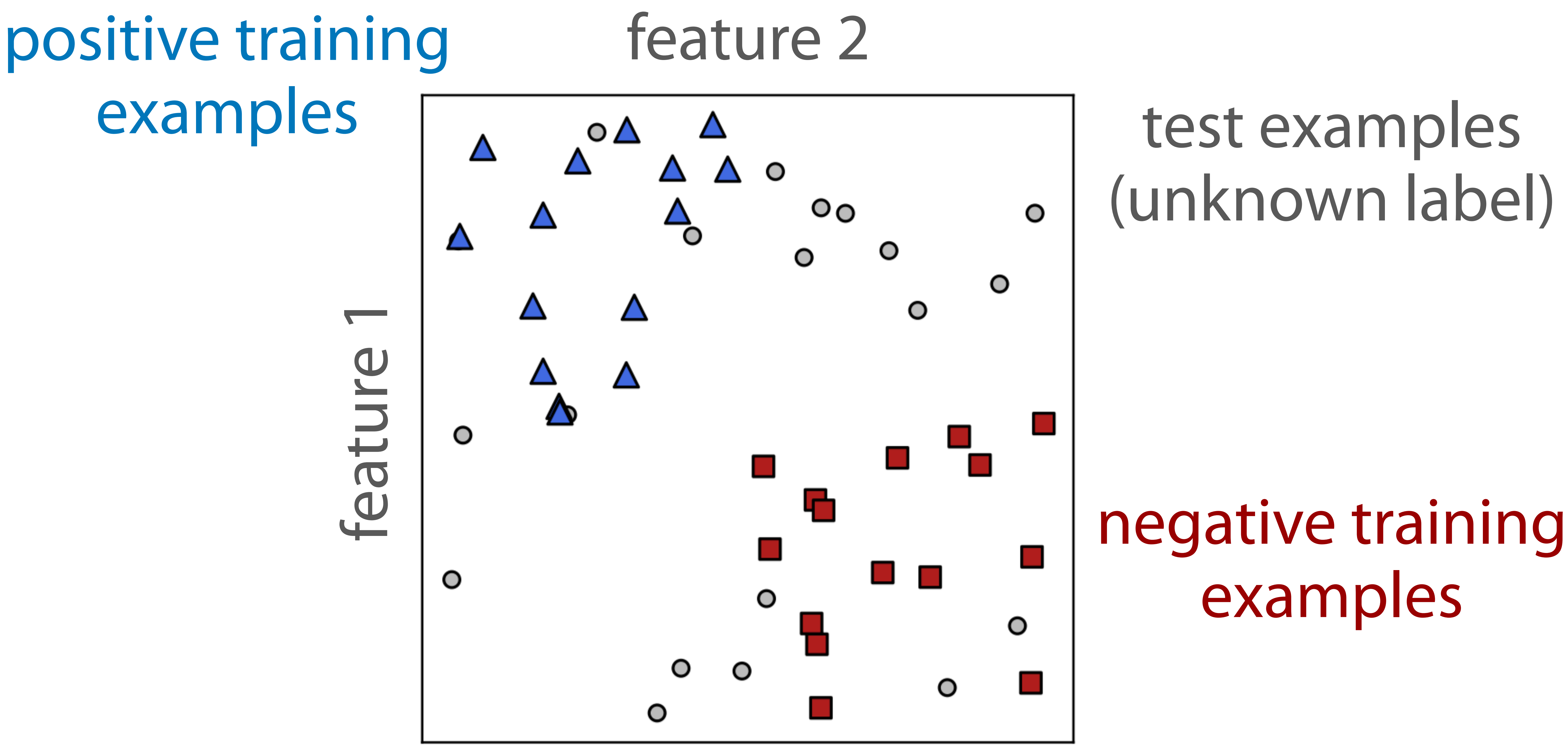
- domain adversarial neural networks & domain confusion
- invariant risk minimization (IRM)
- invariance via selective augmentation (LISA)

\+ produce models robust to spurious correlations, domain shift

- may require domain annotations
- don't address more extreme spurious correlations

Note: ALL methods for distribution shift need to go ***beyond*** standard iid assumptions!

Underspecified data - an example

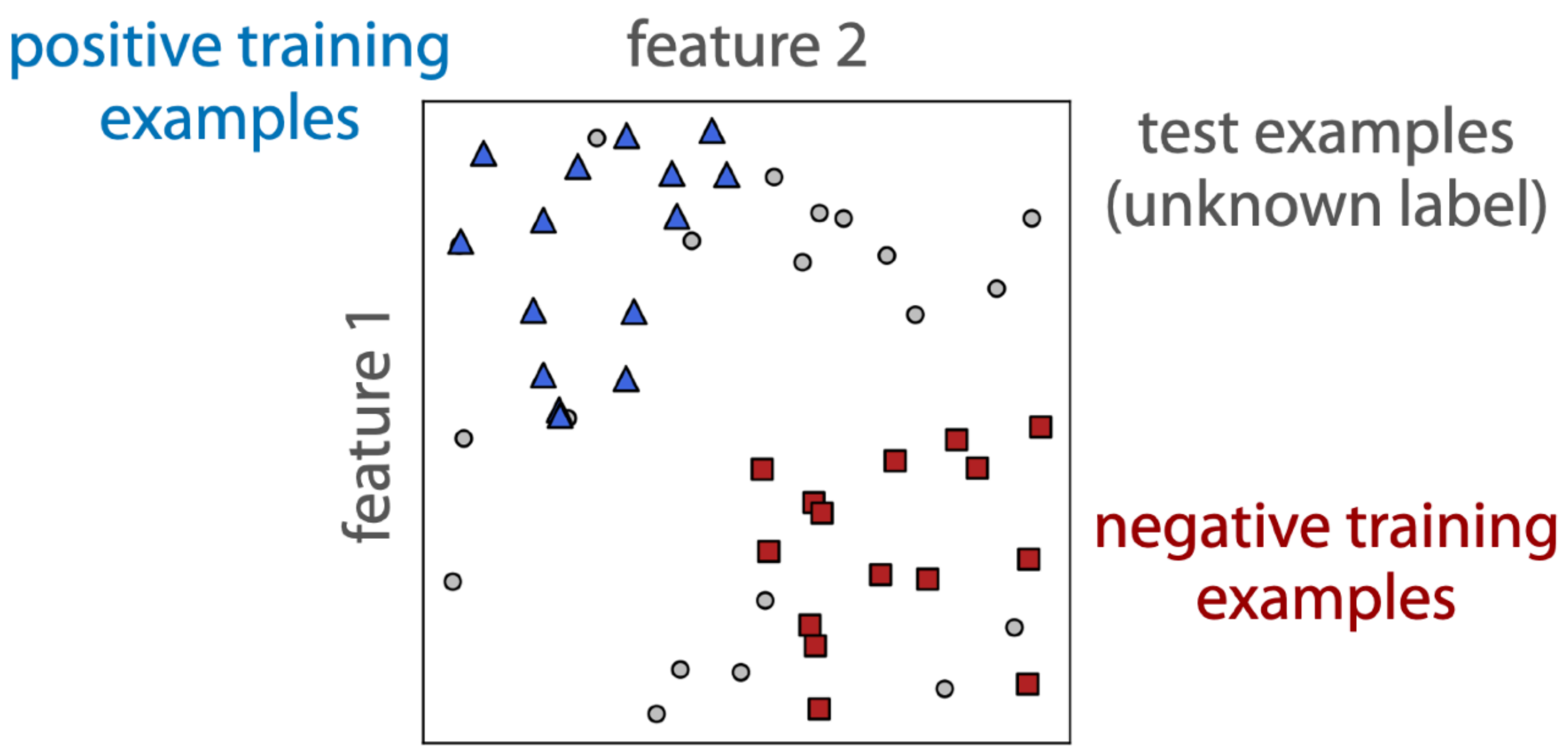


Many functions can achieve low training loss; they **can't all be correct**.

Which feature should the model use?

Underspecified *only because there is covariate shift.*

Possible Solutions



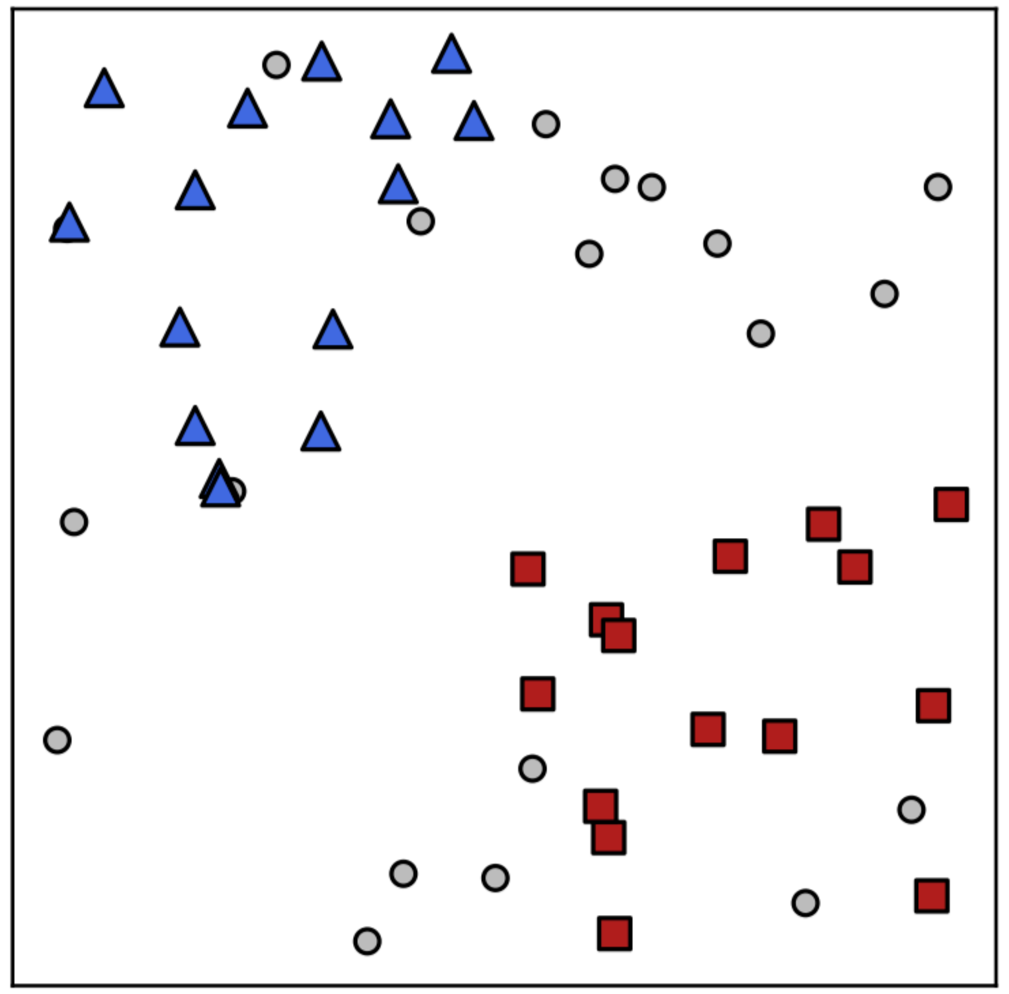
Regularize to the correct function

- requires **domain knowledge**
- requires way to **convert domain knowledge** into a **regularizer**

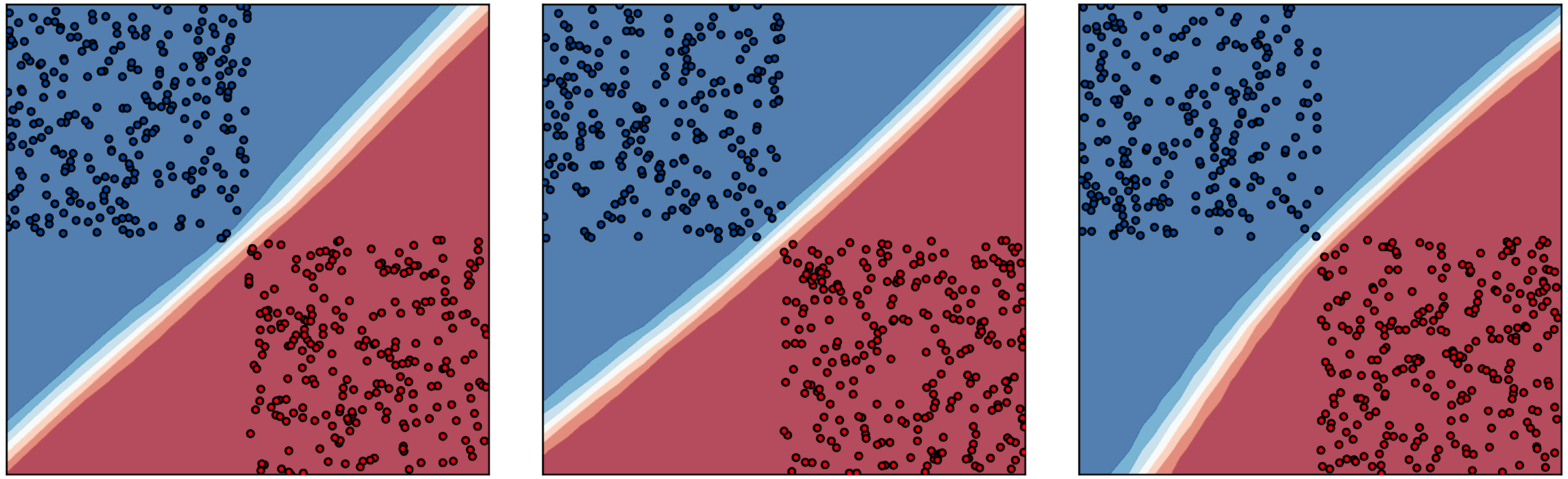
Learn Bayesian posterior over parameters

- these methods **don’t scale** to deep networks

Train an ensemble of deep networks?



Re-training with different seeds



- Vanilla ensembles show little disagreement, even in this toy dataset!
  - Can be worse in larger-scale settings: simplicity bias, texture bias etc
- Core idea: **actively diversify** on unlabeled data from test distribution

Diversify and Disambiguate (DivDis)

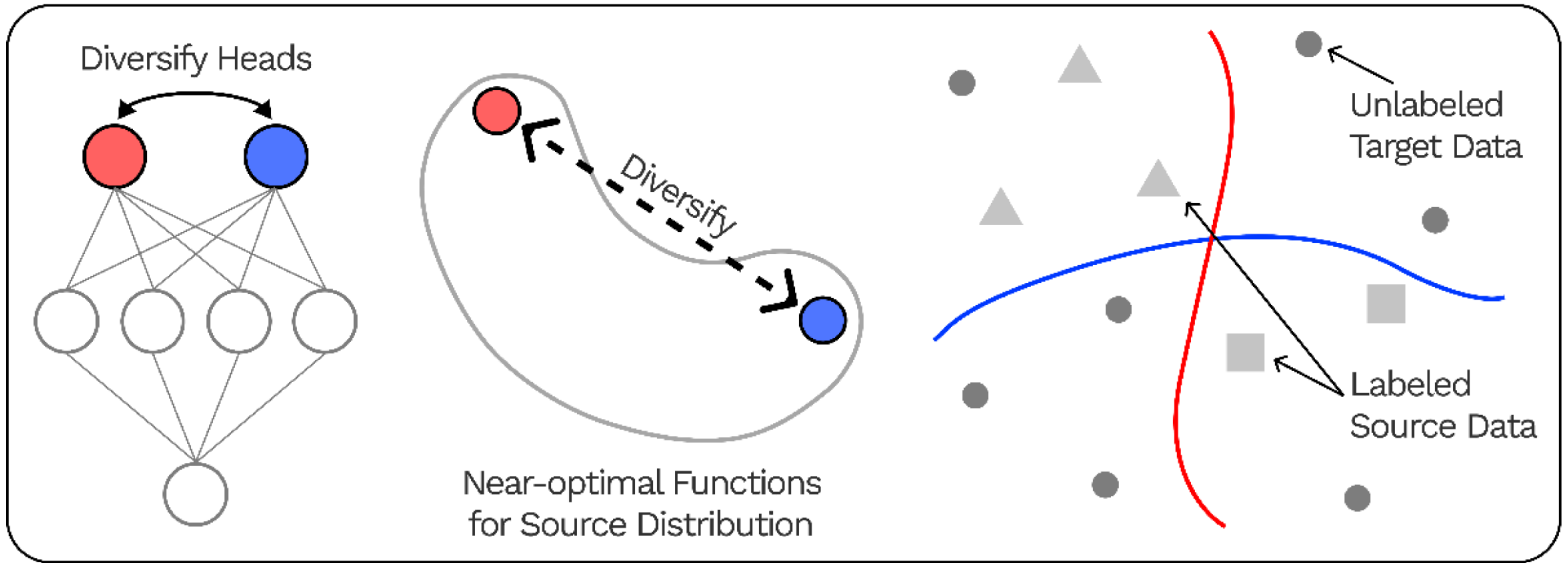
Train multiple functions
(e.g. NN with multiple heads)

Use an ensemble of NNs?

- minimize training error
- maximize disagreement on unlabeled test data

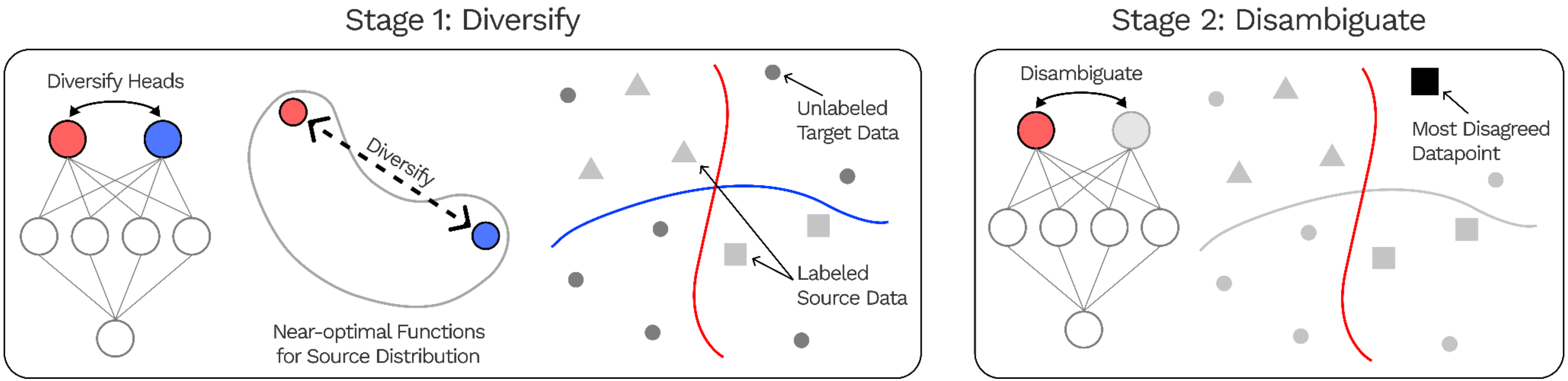
more specifically: minimize statistical dependence $\mathcal{L}_{\mathrm{MI}}(f_i, f_j) = D_{\mathrm{KL}}\left(p(\widehat{y}_i, \widehat{y}_j) \,||\, p(\widehat{y}_i) \otimes p(\widehat{y}_j)\right)$

Stage 1: Diversify



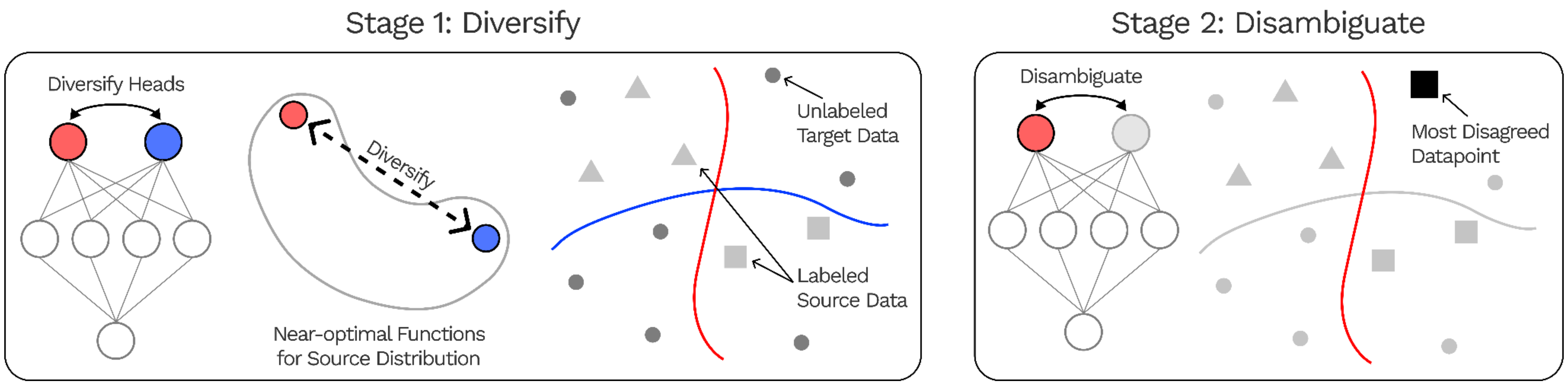
Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

Diversify and Disambiguate (DivDis)



Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

Diversify and Disambiguate (DivDis)

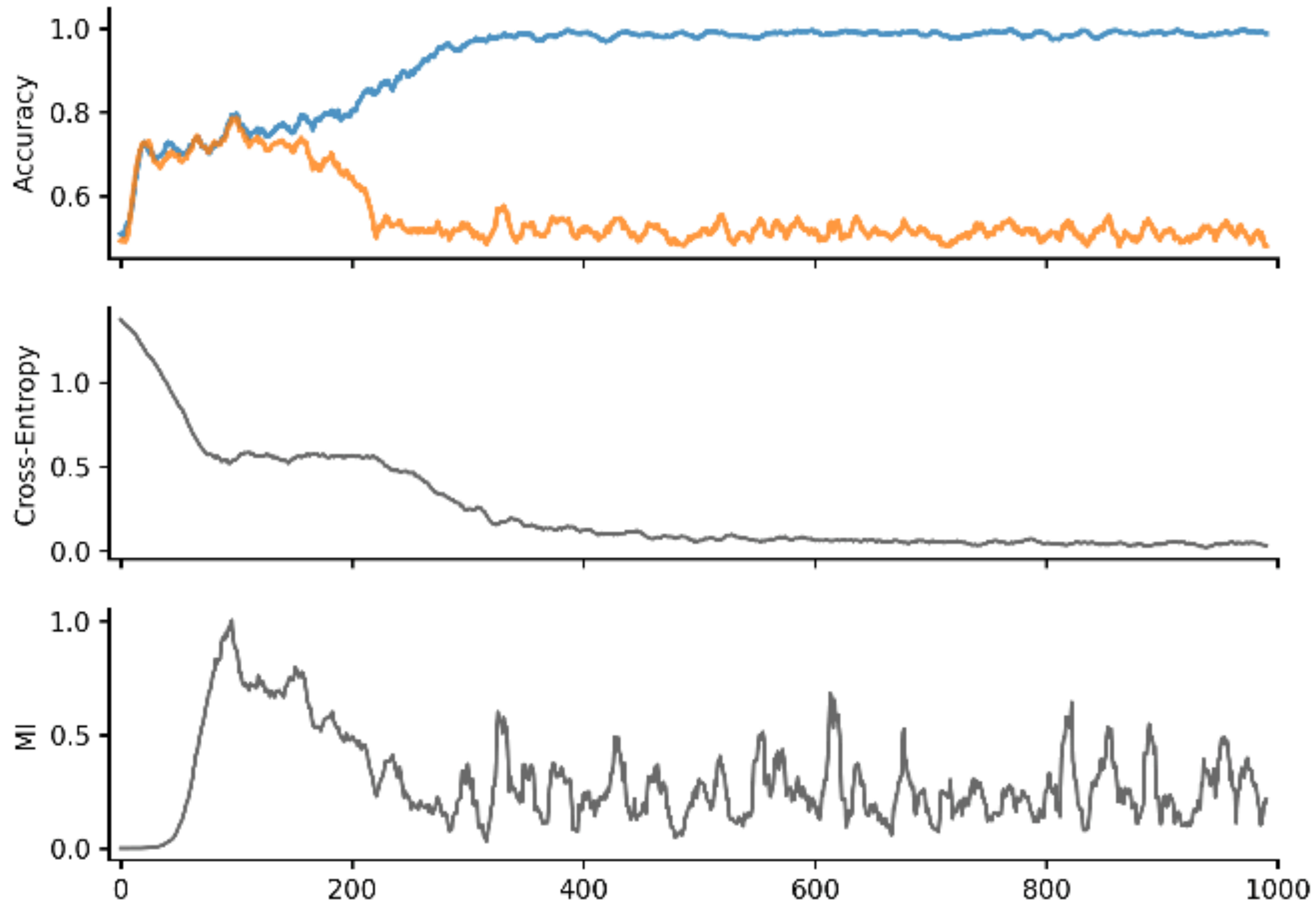


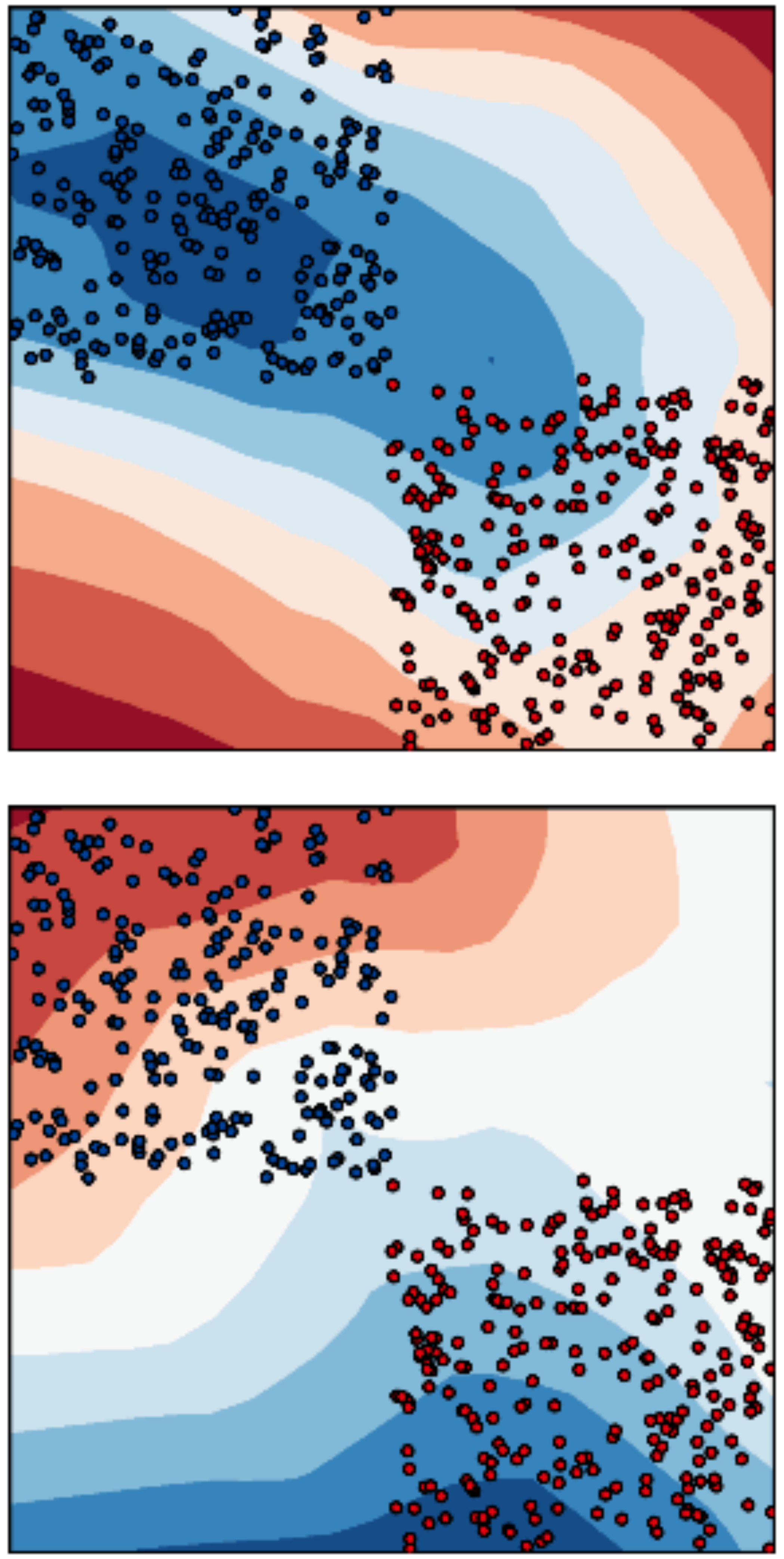
How to select the head?

A few options:

- Randomly label some test points, select most accurate head
- Query label for most disagreed points, select most accurate
- Inspect the learned functions (e.g. using interpretability methods)

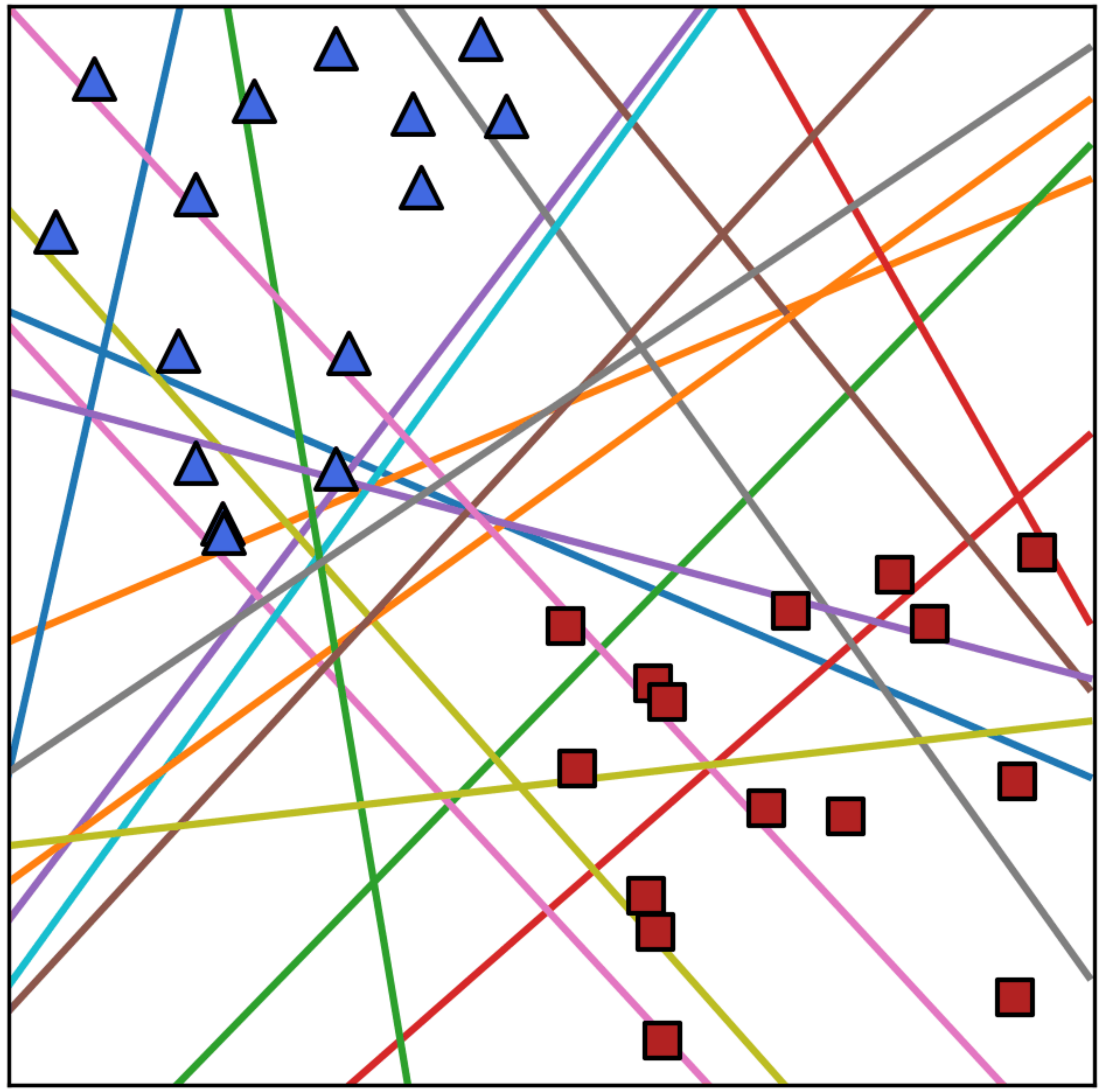
What Happens During Diversification?





Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

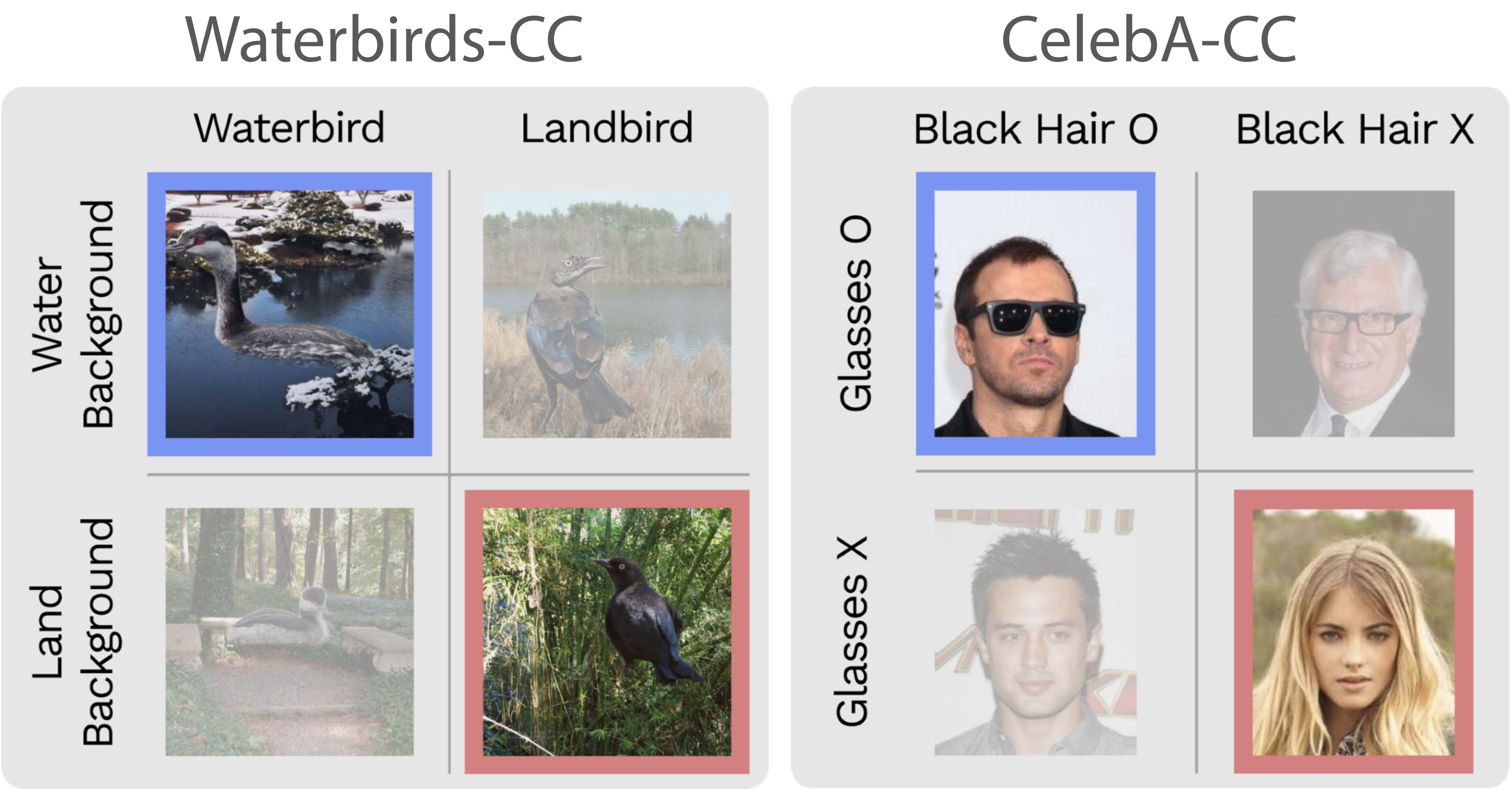
What Happens During Diversification?



The **diversified heads** cover the space of functions consistent with training data.

Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

Experiment 1: Completely Correlated Data



- design train datasets with **complete correlation** btw spurious attribute & label
- imperfect or no correlation in test data
- measure avg & worst-group accuracy
- DivDis with 2 heads, 16 active queries

Initial Comparisons:

- ERM (standard NN training)
- JTT (upweight examples w/ highest error)
- Group DRO (upweight group w/ highest error)

Note: none of these are designed to handle perfect correlation!

Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

Experiment 1: Completely Correlated Data

	Waterbirds-CC		CelebA-CC-1		CelebA-CC-2		MultiNLI-CC	
	Avg (%)	Worst (%)	Avg (%)	Worst (%)	Avg (%)	Worst (%)	Avg (%)	Worst (%)
Random guessing baseline	50.0	50.0	50.0	50.0	50.0	50.0	33.3	33.3
ERM	60.5 ± 1.6	7.0 ± 1.5	70.9 ± 2.0	57.0 ± 5.8	73.1 ± 0.9	41.1 ± 2.6	53.2 ± 1.5	22.8 ± 2.5
JTT (Liu et al., 2021)	44.6 ± 1.9	26.5 ± 1.4	71.4 ± 1.9	51.2 ± 5.4	78.7 ± 0.8	59.8 ± 1.1	80.0 ± 4.0	40.5 ± 2.3
GDRO (Sagawa et al., 2020)	55.6 ± 4.8	47.1 ± 8.9	71.6 ± 0.3	59.3 ± 2.6	71.6 ± 2.4	61.3 ± 2.3	79.1 ± 3.4	39.8 ± 1.4
DivDis w/o reg	87.2 ± 0.8	77.5 ± 4.7	91.0 ± 0.4	$\mathbf{85.9} \pm 1.0$	79.7 ± 0.4	$\mathbf{69.3} \pm 1.9$	80.3 ± 0.6	67.6 ± 4.0
DivDis	87.6 ± 1.4	$\mathbf{82.4} \pm 1.9$	90.8 ± 0.4	$\mathbf{85.6} \pm 1.1$	79.5 ± 0.2	$\mathbf{68.5} \pm 1.7$	79.9 ± 1.2	$\mathbf{71.5} \pm 2.5$

Existing methods struggle, sometimes even doing **worse than random guessing**

DivDis shows **>25% improvement** in worst-group accuracy on 3 of 4 datasets

Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

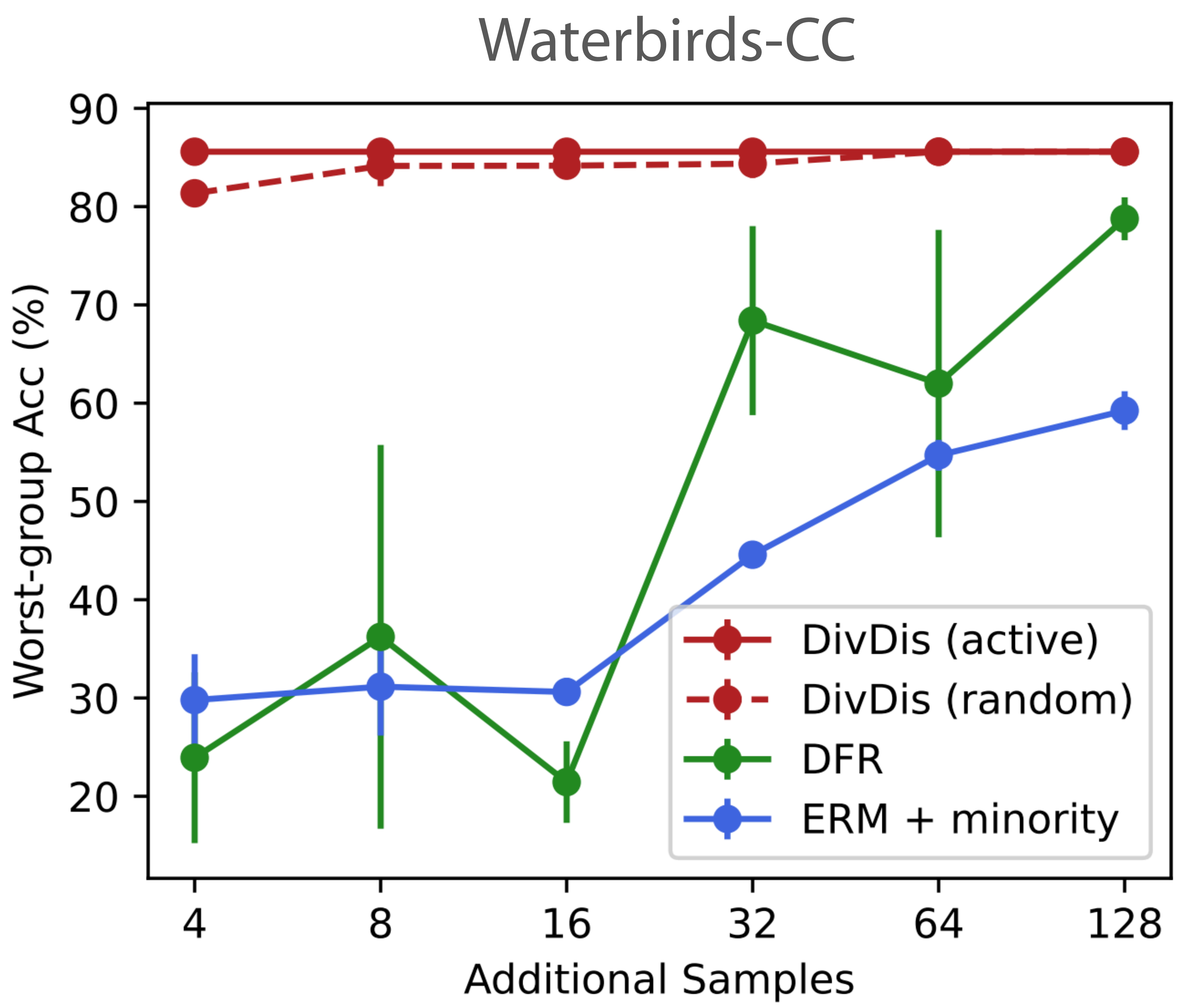
Experiment 1: Completely Correlated Data

What happens when you give a few labeled examples to ERM?

Compare to:

- **ERM+minority:** standard NN training on training data & *N* minority examples
- **DFR:** ERM + fine-tune on *N* target examples

DivDis substantially more **label efficient**, still favorable with 128 labeled target examples



Kirichenko, P., Izmailov, P., and Wilson, A. G. (2022). Last layer re-training is sufficient for robustness to spurious correlations. arXiv:2204.02937

Lee, Yao, Finn. **Diversify and Disambiguate: Learning from Underspecified Data.** arXiv '22

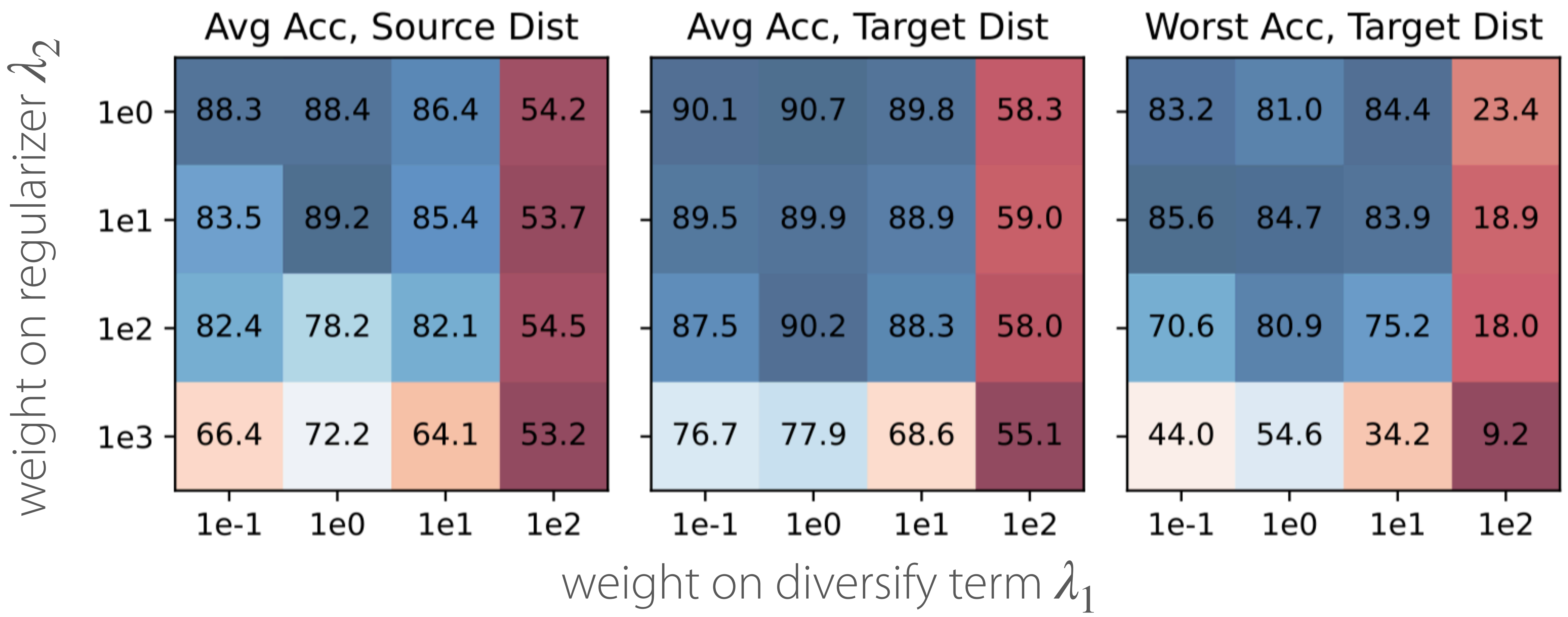
Experiment 2: Assumptions for Tuning Hyperparameters

On prior Waterbird & CelebA robustness benchmarks.

	Waterbirds worst-group test acc.		CelebA worst-group test acc.	
	Tuned w/ worst	Tuned w/ avg	Tuned w/ worst	Tuned w/ avg
CVaR DRO (Levy et al., 2020)	75.9%	62.0%	64.4%	36.1%
LfF (Nam et al., 2020)	78.0%	44.1%	77.2%	24.4%
JTT (Liu et al., 2021)	86.7%	62.5%	81.1%	40.6%
DivDis	85.6%	**81.0%**	55.0%	**55.0%**

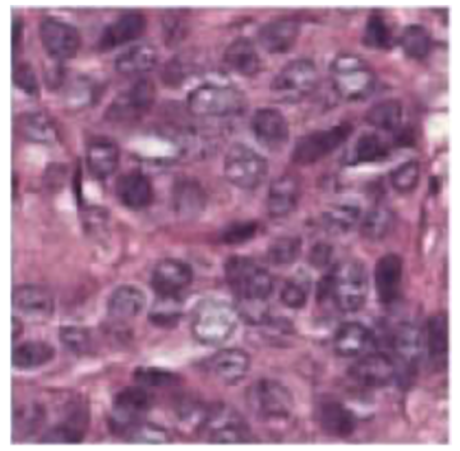
Existing methods assume access to **group labels** during hyperparameter tuning.

DivDis can be **tuned without group labels.**

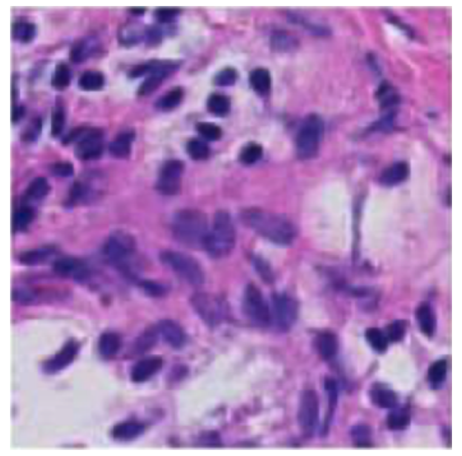


Experiment 3: Domain Shift Problems with Mild Correlations

Camelyon17-WILDS



Labeled data from **in-distribution** hospitals
(no complete correlation)



Unlabeled data from **out-of-distribution** hospitals

	Test Acc
Pseudo-Label	67.7 ± 8.2
DANN	68.4 ± 9.2
FixMatch	71.0 ± 4.9
CORAL	77.9 ± 6.6
NoisyStudent	86.7 ± 1.7
DivDis (ours)	$\mathbf{90.4} \pm 1.8$

DivDis works well on **domain shift**
(not just subpopulation shift)

DivDis compares favorably to domain adaptation methods.

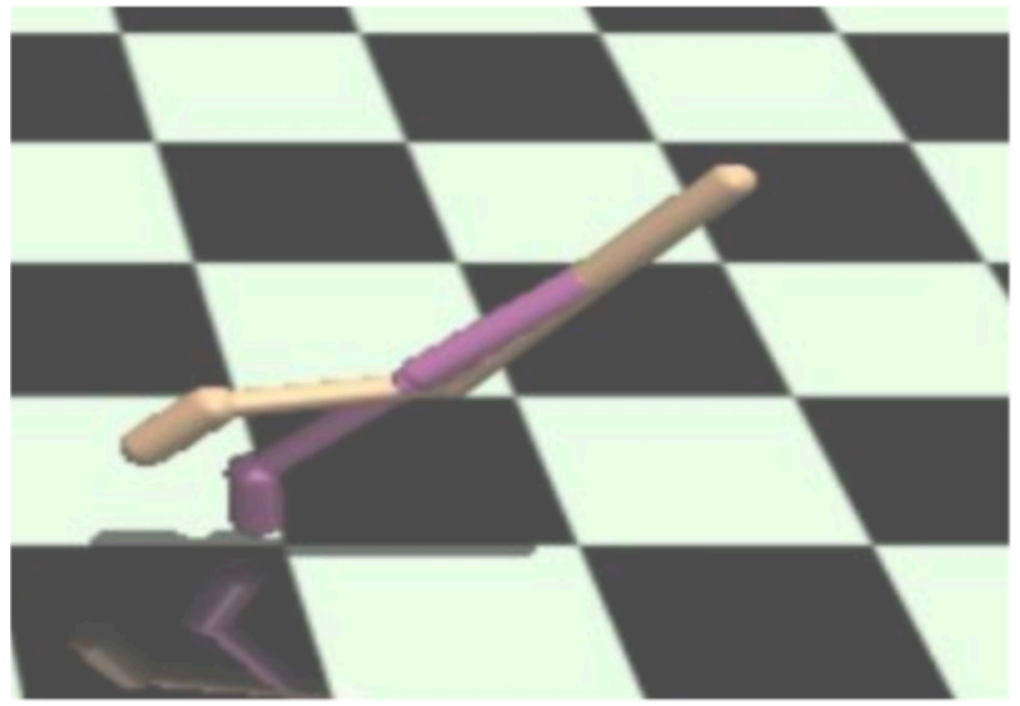
Summary of DivDis

- Tackles underspecification in data. **Existing methods fail** on data with severe underspecification through **complete correlations**.
- To deal with such highly underspecified data, we must consider **multiple hypotheses**.
- DivDis **performs well** on completely correlated data, and can be **tuned without group information**.
- Code: https://github.com/yoonholee/DivDis

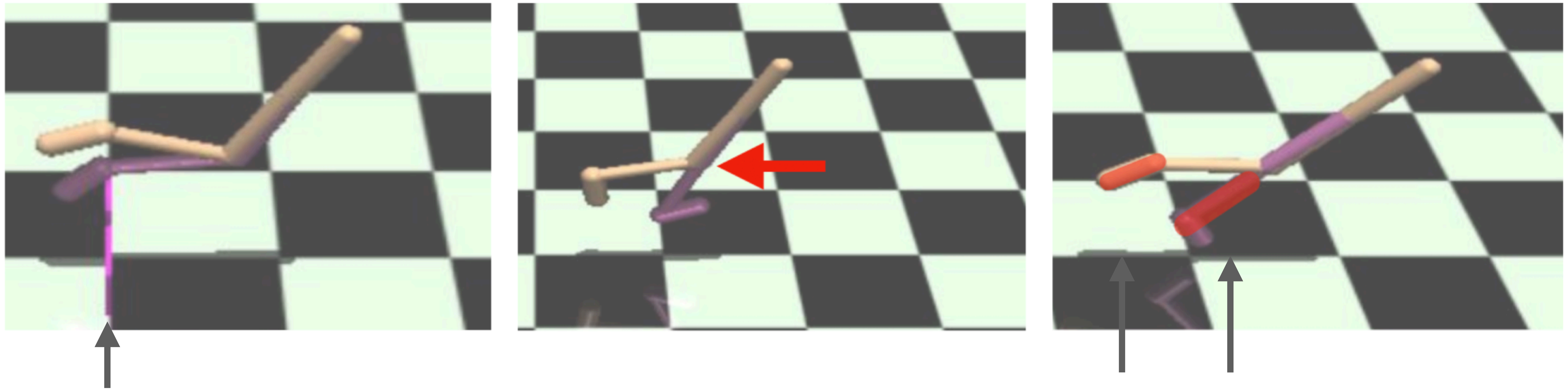
Lee, Yao, Finn. Diversify and Disambiguate: Learning from Underspecified Data. arXiv '22

Aside: Can you learn diverse ensembles of RL policies?

one training environment M_{train}



new test environments M_{test}



obstacle

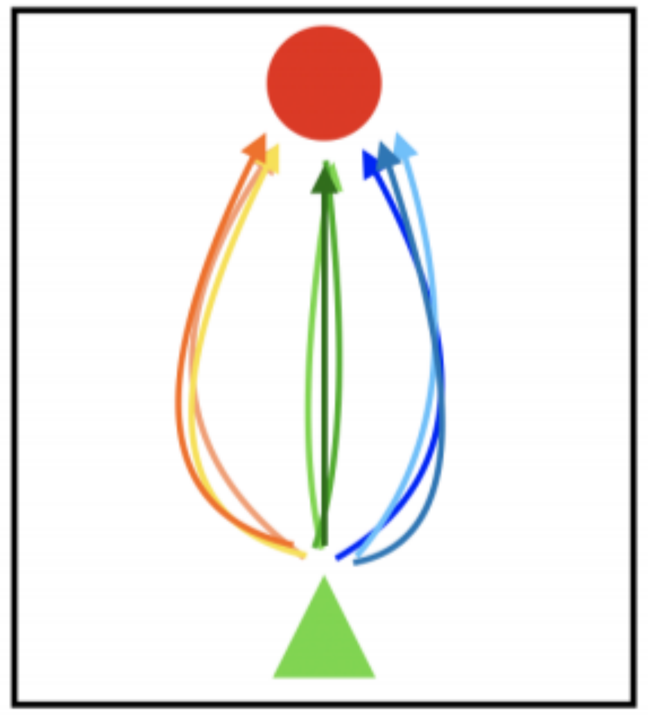
force perturbation

disabled joints

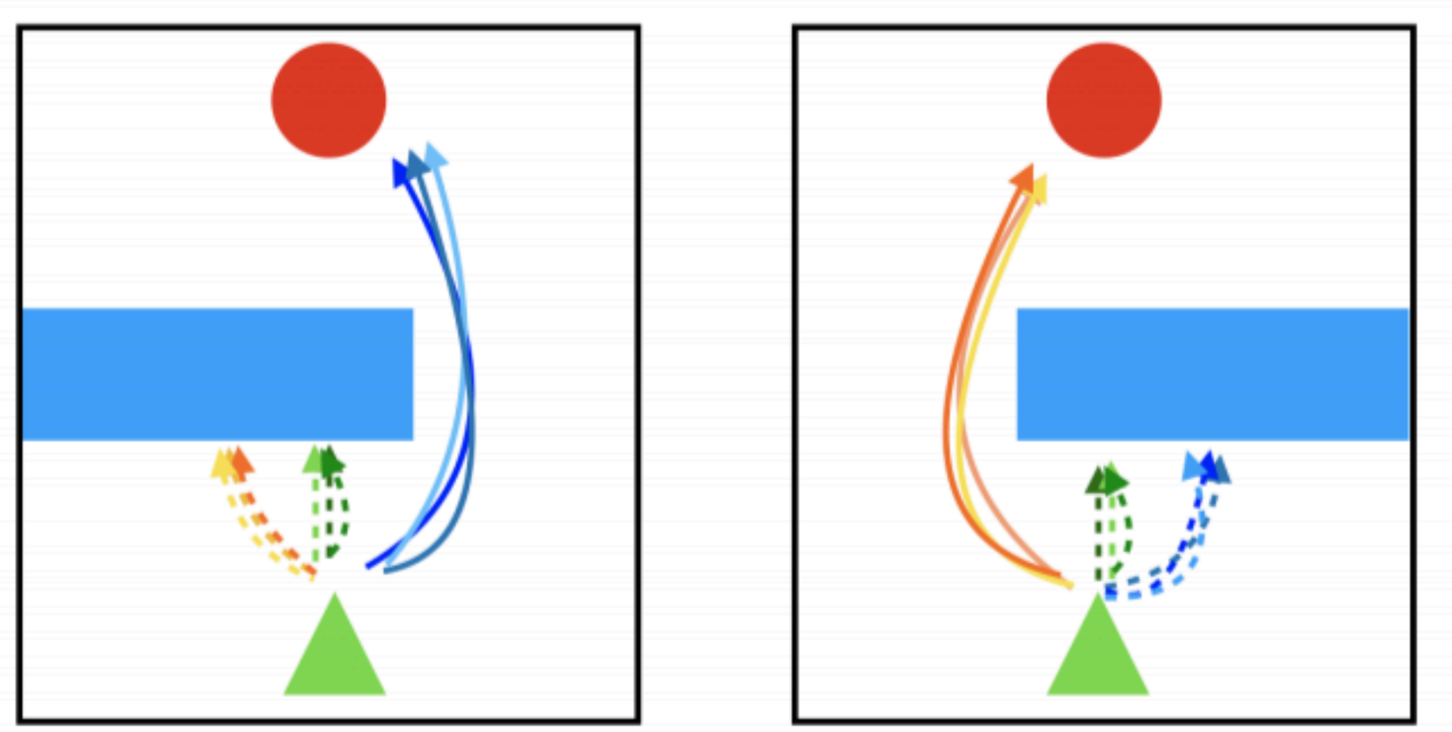
Aside: Can you learn diverse ensembles of RL policies?

Simple idea:

Learn & *remember* multiple solutions to M_{train}



Adapt solution set to M_{test}



Assumption #1: ability to adapt with modest amount of data

Assumption #2: changes to the environment are local such that the optimal policy in M_{test} also does well in M_{train}

e.g., few-shot robustness to local changes in obstacles, terrains, friction, etc

Saurabh Kumar

S. Kumar, A. Kumar, Levine, Finn. *One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL*, NeurIPS '20

How to learn multiple solutions?

Learn controllable space of diverse policies that achieve return with ϵ of optimal

using latent variables

$\pi_\theta(a \mid s, z)$

constrained optimization

Train time:

$$\arg\max_{\theta} \sum_{t=1}^{T} \underbrace{I(s_t; z)}_{\mathcal{H}(s) - \mathcal{H}(s \mid z)} \quad \text{s.t.} \quad \forall z, R_{\mathcal{M}}(\pi_\theta) \geq R_{\mathcal{M}}(\pi^*_{\mathcal{M}}) - \varepsilon$$

Test time: Roll-out K policies with different z. Return $\pi_\theta(a \mid s, z_i)$ for best performing z_i.

"structured maximum entropy RL" (SMERL)

Eysenbach, Gupta, Ibarz, Levine. *DIAYN: Learning Skills without a Reward Function*, ICLR '18

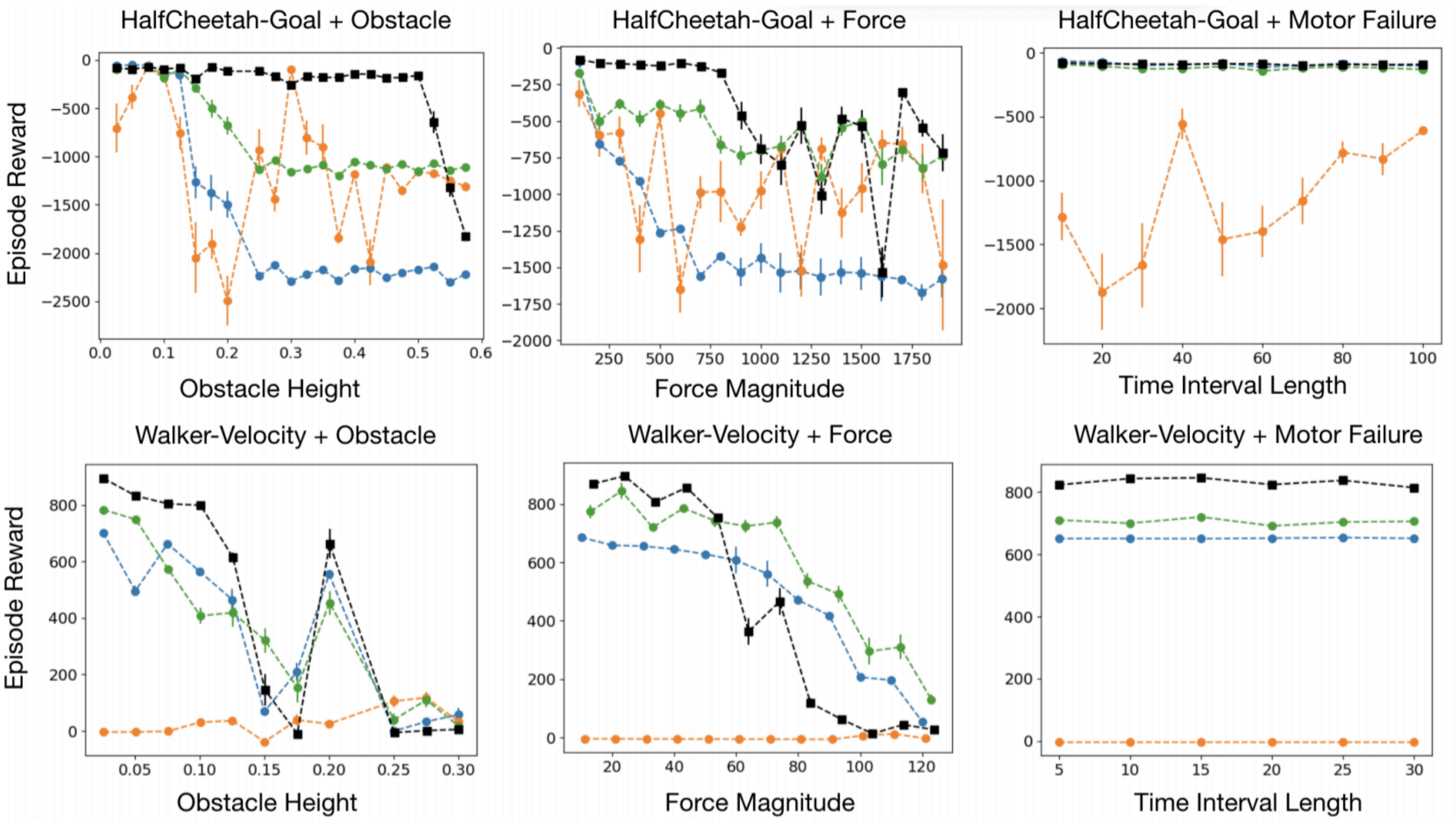
S. Kumar, A. Kumar, Levine, Finn. *One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL*, NeurIPS '20

Testing Robustness to Obstacles, Perturbations, and Motor Failures

Compare:

Measuring **5-shot** generalization.

performance

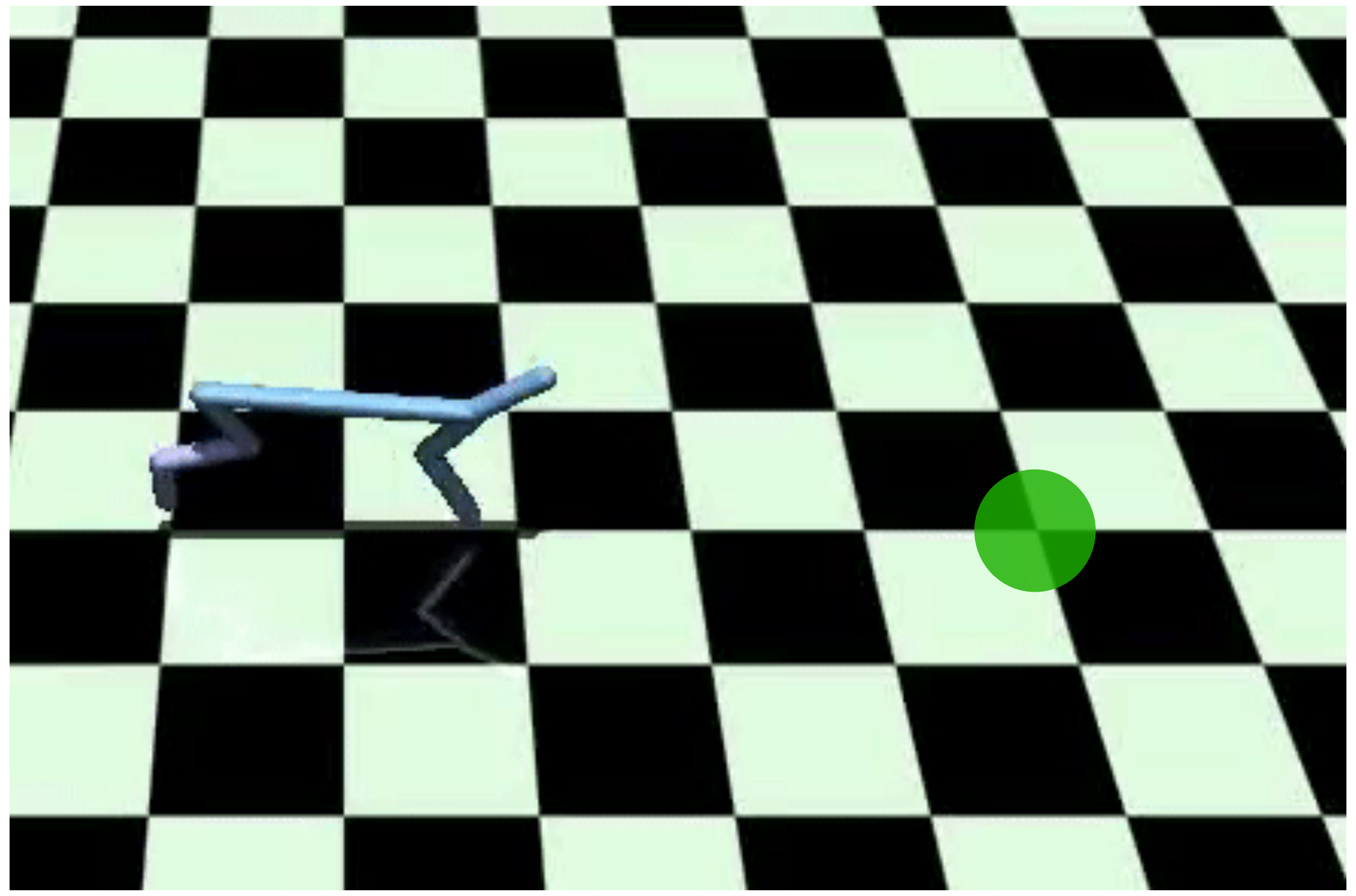


degree of environment change

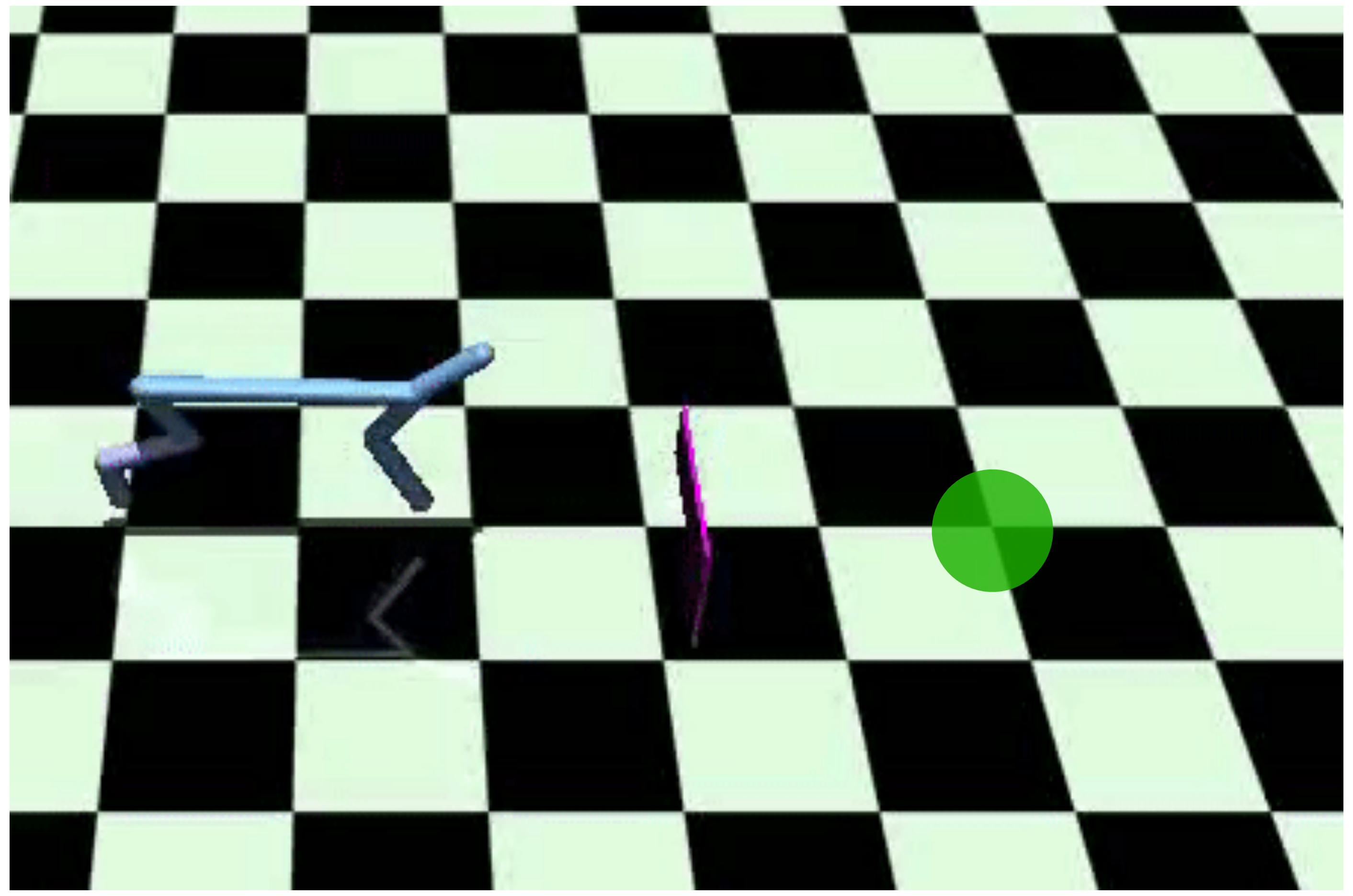
Pinto, Davidson, Sukthankar, Gupta. *Robust Adversarial Reinforcement Learning*, ICML '17

S. Kumar, A. Kumar, Levine, Finn. *One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL*, NeurIPS '20

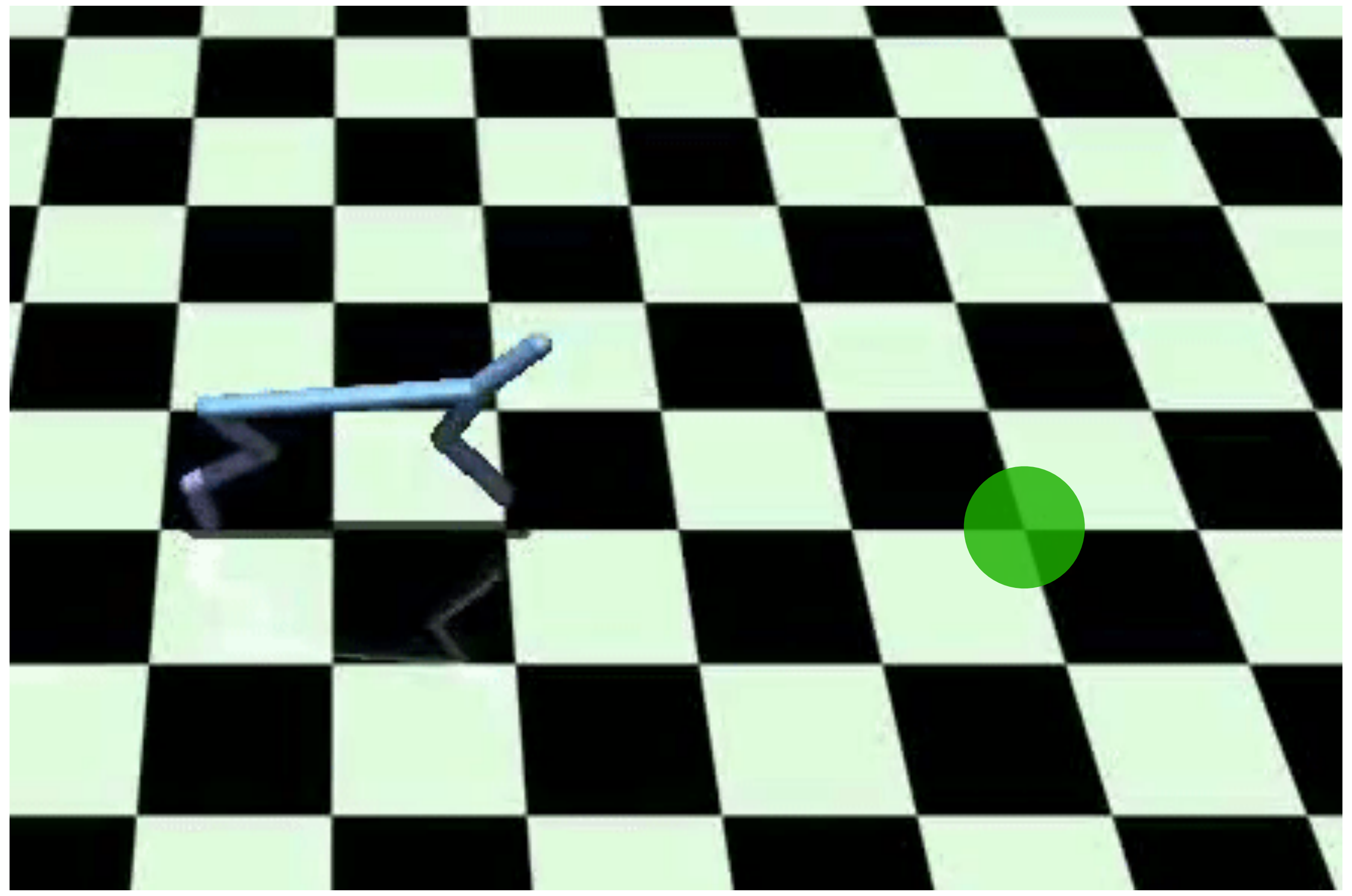
SAC policies at train time.



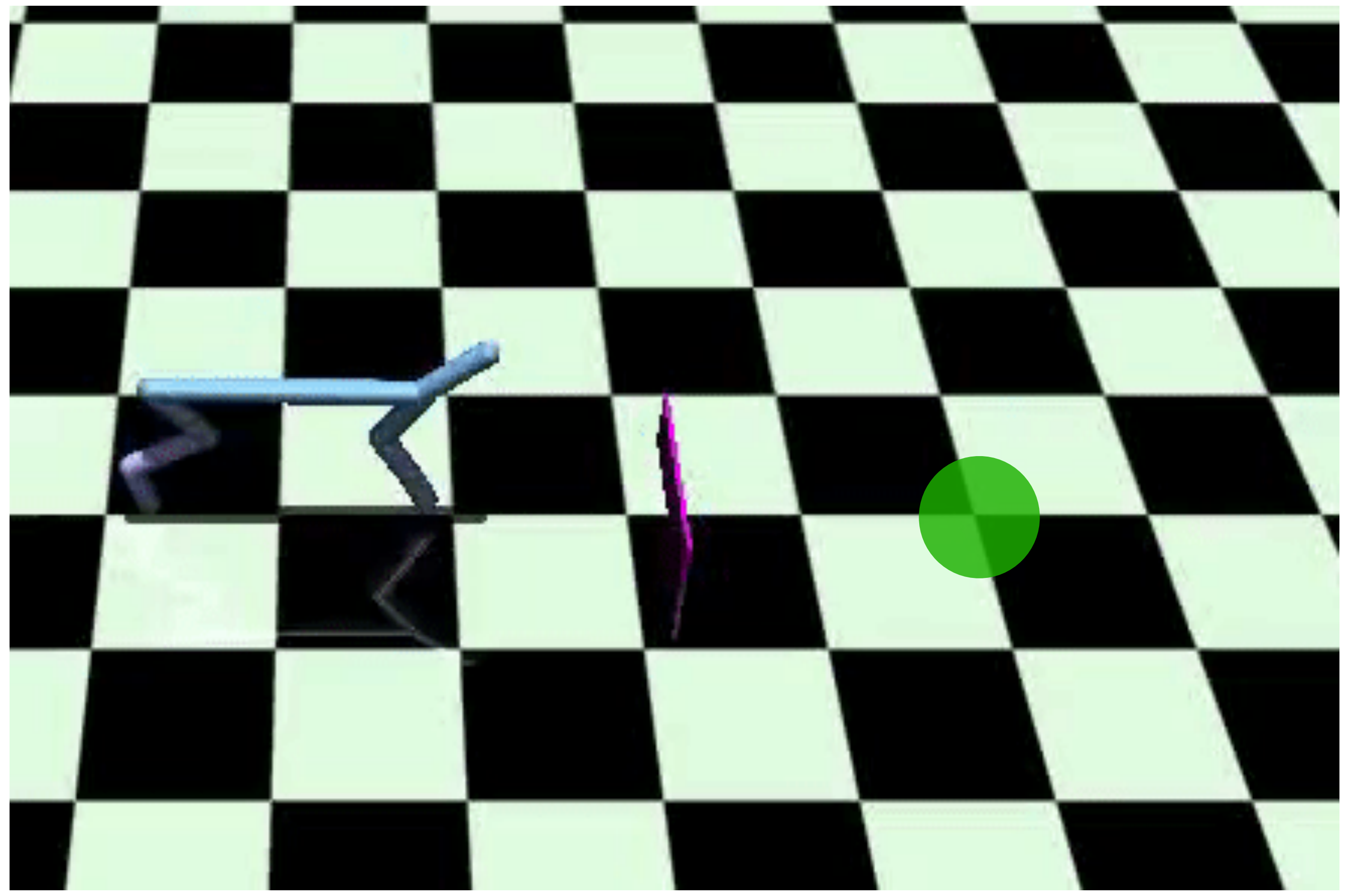
Best SAC policy at test time.



SMERL policies at train time.



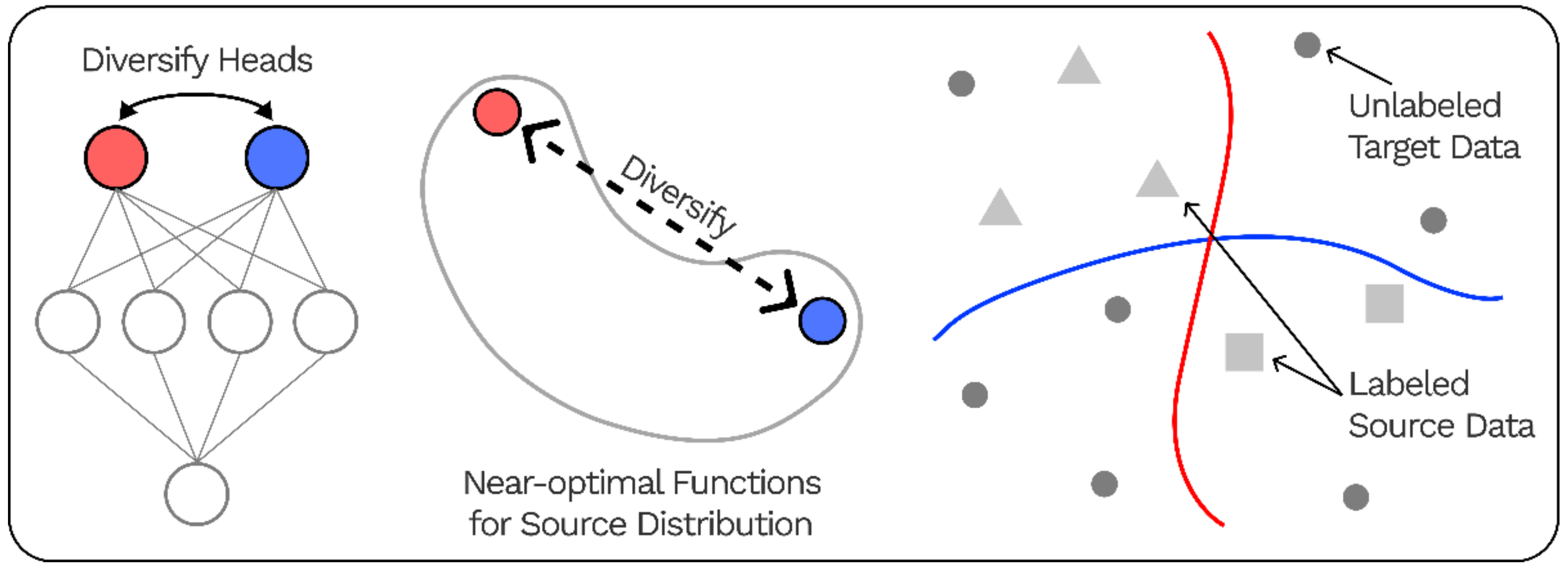
Best **SMERL** policy at test time.



S. Kumar, A. Kumar, Levine, Finn. *One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL*, NeurIPS '20

Outline

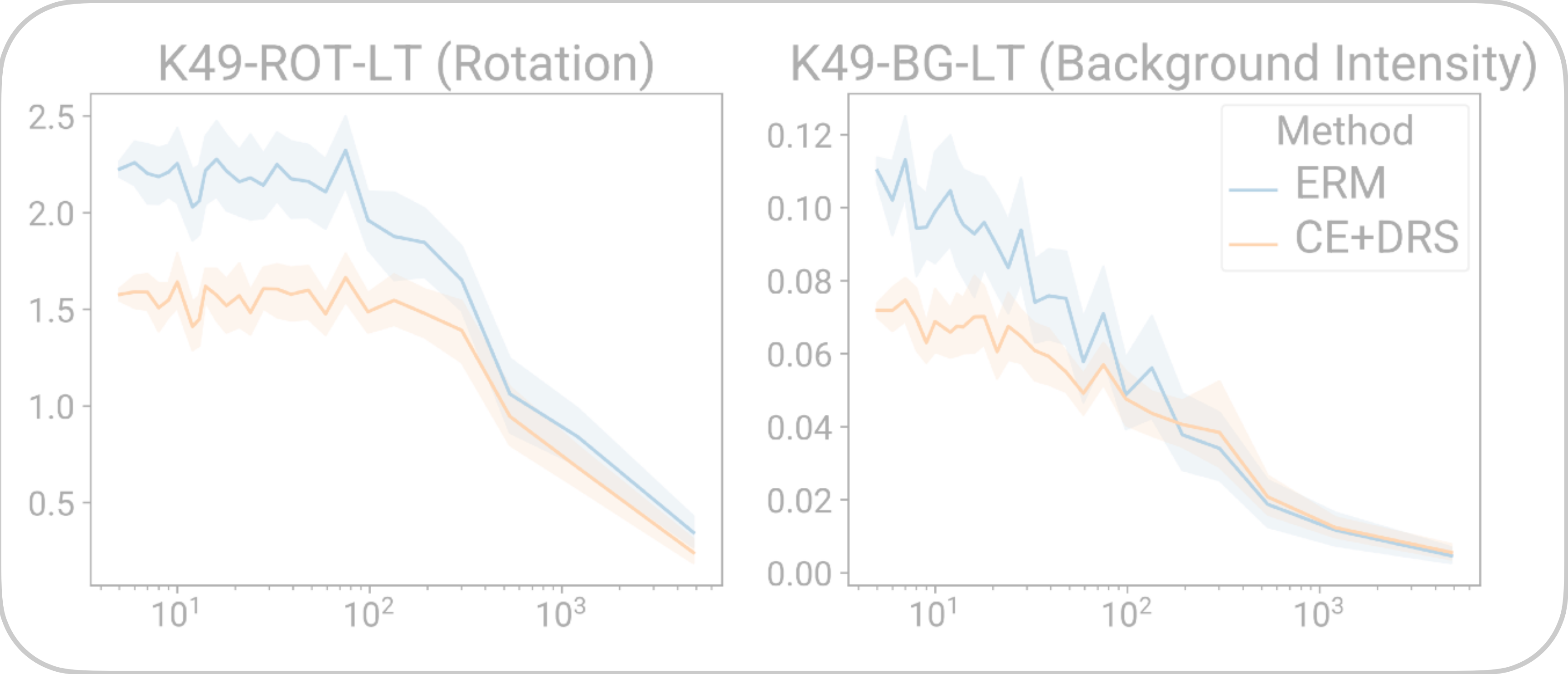
Addressing extreme covariate shift
via diverse ensembles



for supervised learning & reinforcement learning

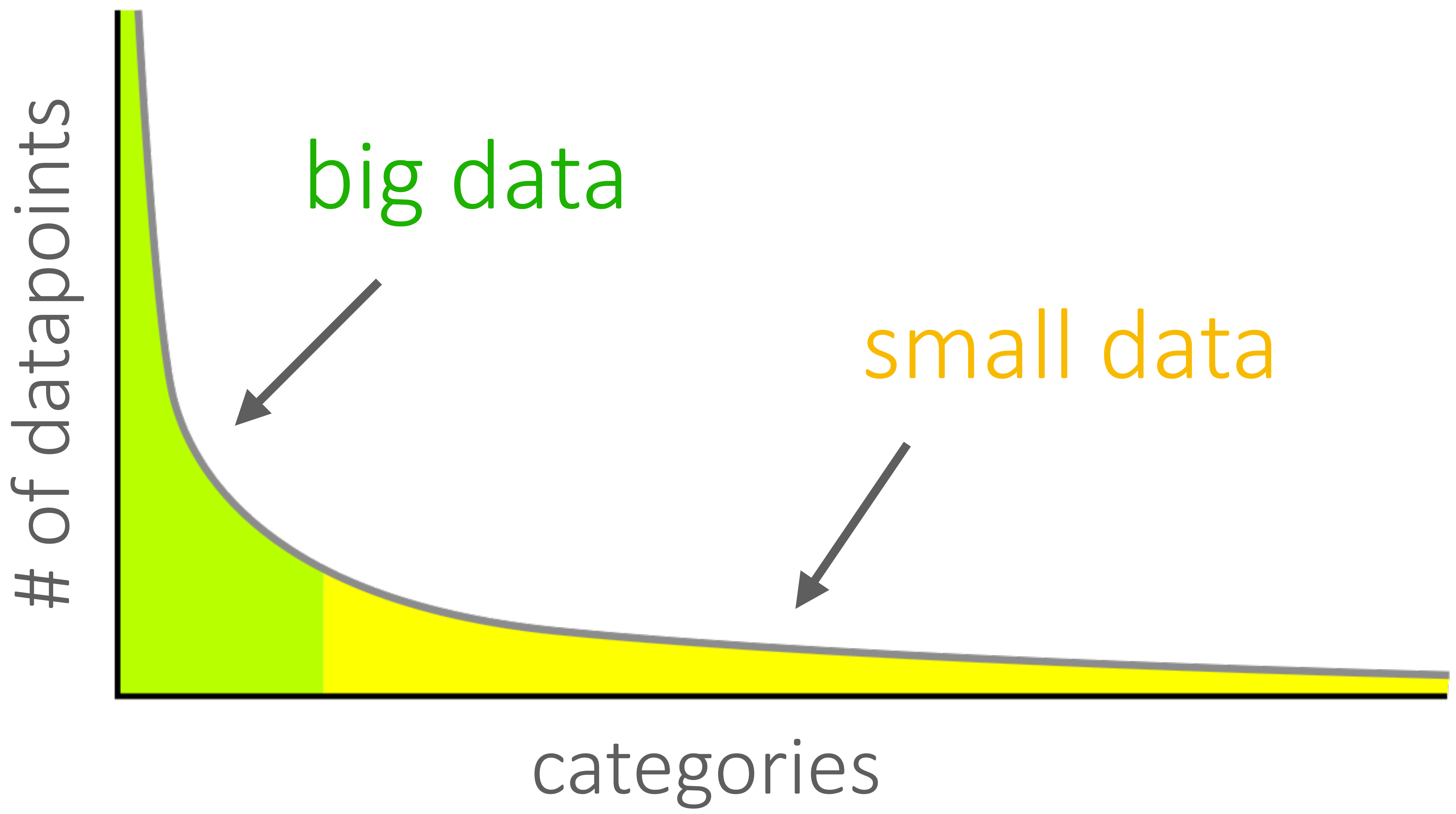
Takeaway: Learning diverse classifiers & policies enables fast adaptation to OOD situations

Addressing label shift
via invariance transfer



for image classification

What if your data has a long tail?

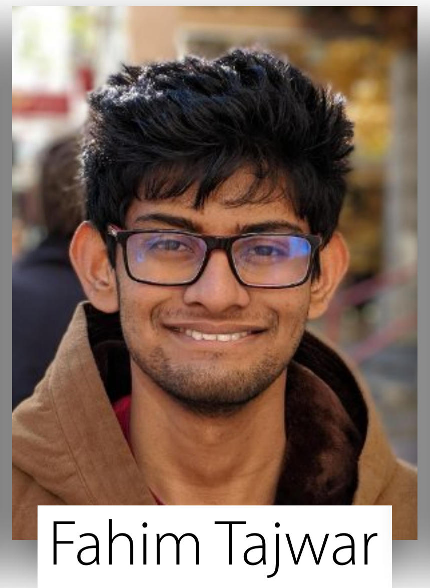


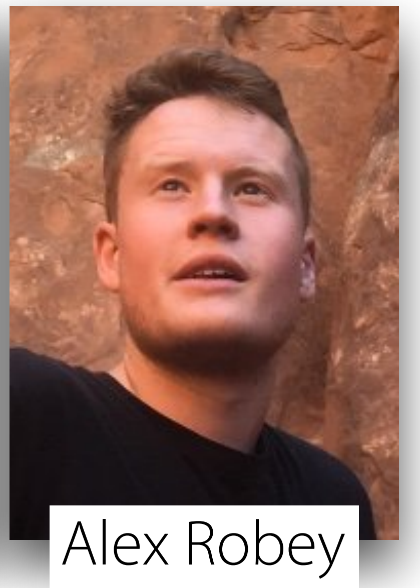
Why do deep networks fail on the tail?

Hypothesis

The model fails to transfer **class-agnostic invariances** from the head classes to the tail classes

—> if true, would lead to poor generalization on the tail.





Zhou, Tajwar, Robey, Knowles, Pappas, Hasani, Finn. **Do Deep Networks Transfer Invariances Across Classes**. ICLR '21.

Hypothesis

The model fails to transfer **class-agnostic invariances** from the head classes to the tail classes

Empirically testing this hypothesis:

- Create **synthetic long-tailed dataset** with **invariance to transformation T**
- Train models and evaluate their invariance to T.

T: Background shading

T: Image dilation/erosion

T: Rotation

based on Kuzushiji-49 (K49) dataset

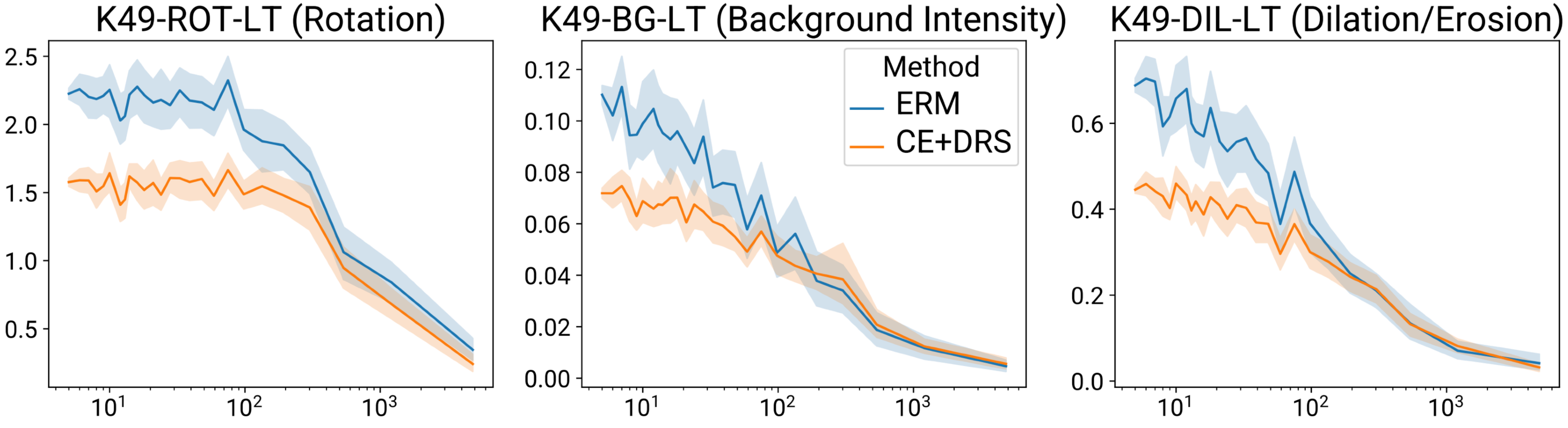
Zhou, Tajwar, Robey, Knowles, Pappas, Hasani, Finn. **Do Deep Networks Transfer Invariances Across Classes**. ICLR '21.

Hypothesis

The model fails to transfer **class-agnostic invariances** from the head classes to the tail classes

Measure invariance to T w.r.t. class size.

Invariance to T (lower is better)



# of examples per class

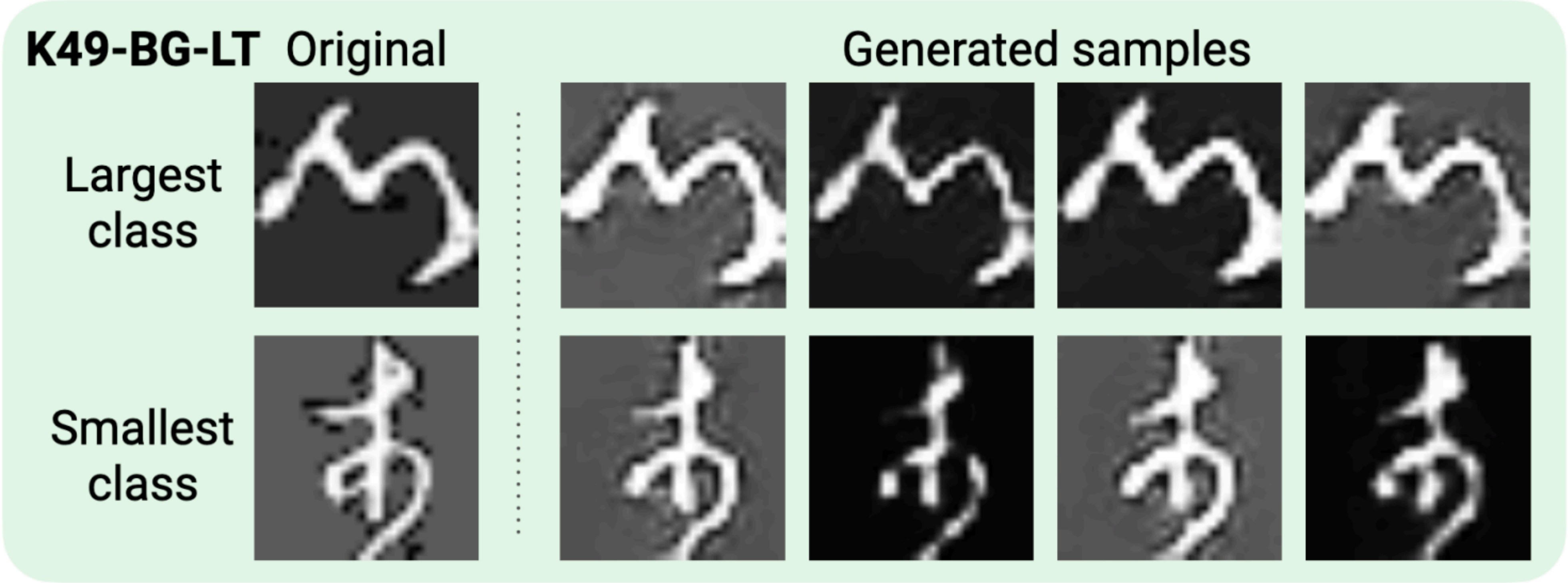
Takeaway: Evidence suggests that invariances are **not** transferred across classes.

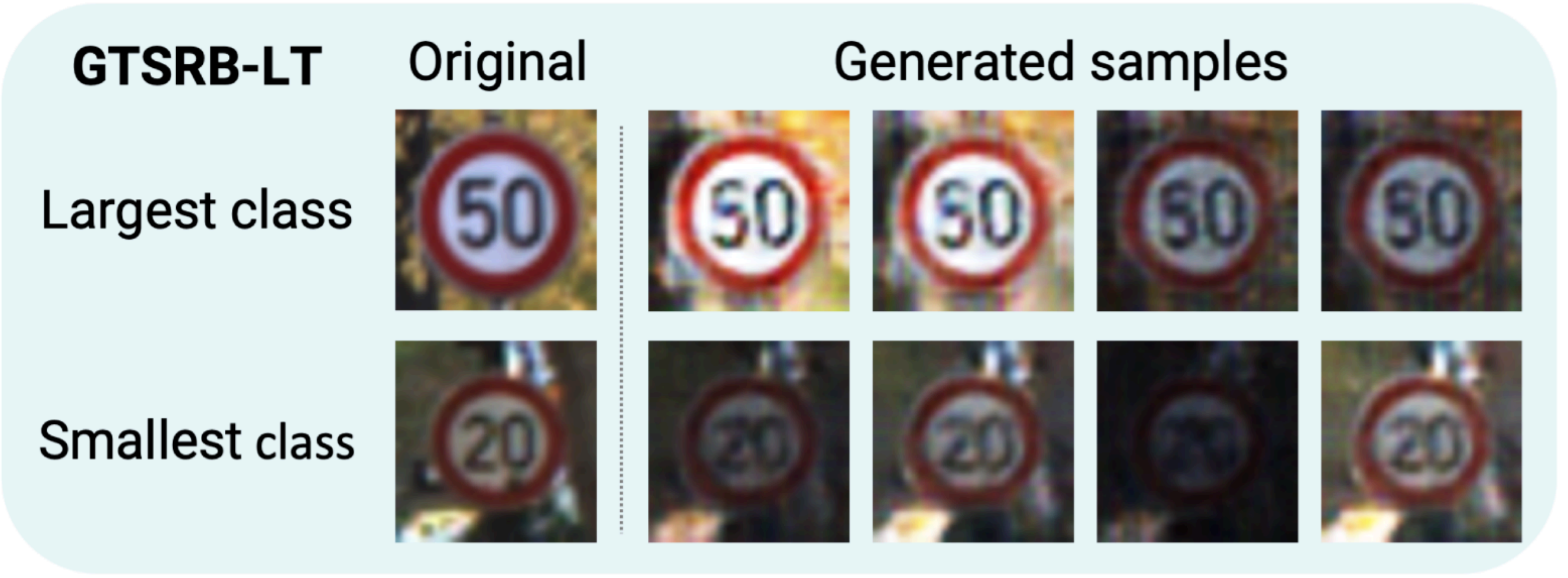
Zhou, Tajwar, Robey, Knowles, Pappas, Hassani, Finn. **Do Deep Networks Transfer Invariances Across Classes**. ICLR '21.

Can we encourage the model to transfer invariances across classes?

Generative invariance transfer:

1. Train a conditional generative model to estimate class-preserving transformations.(1)
2. Use the model to augment small classes.(2)





(1)Related works, which use paired transformation data:

Robey et al. Model-Based Robust Deep Learning. 2020

Wong & Kolter. Learning Perturbation Sets for Robust Deep Learning. 2020

(2)Related augmentation works:

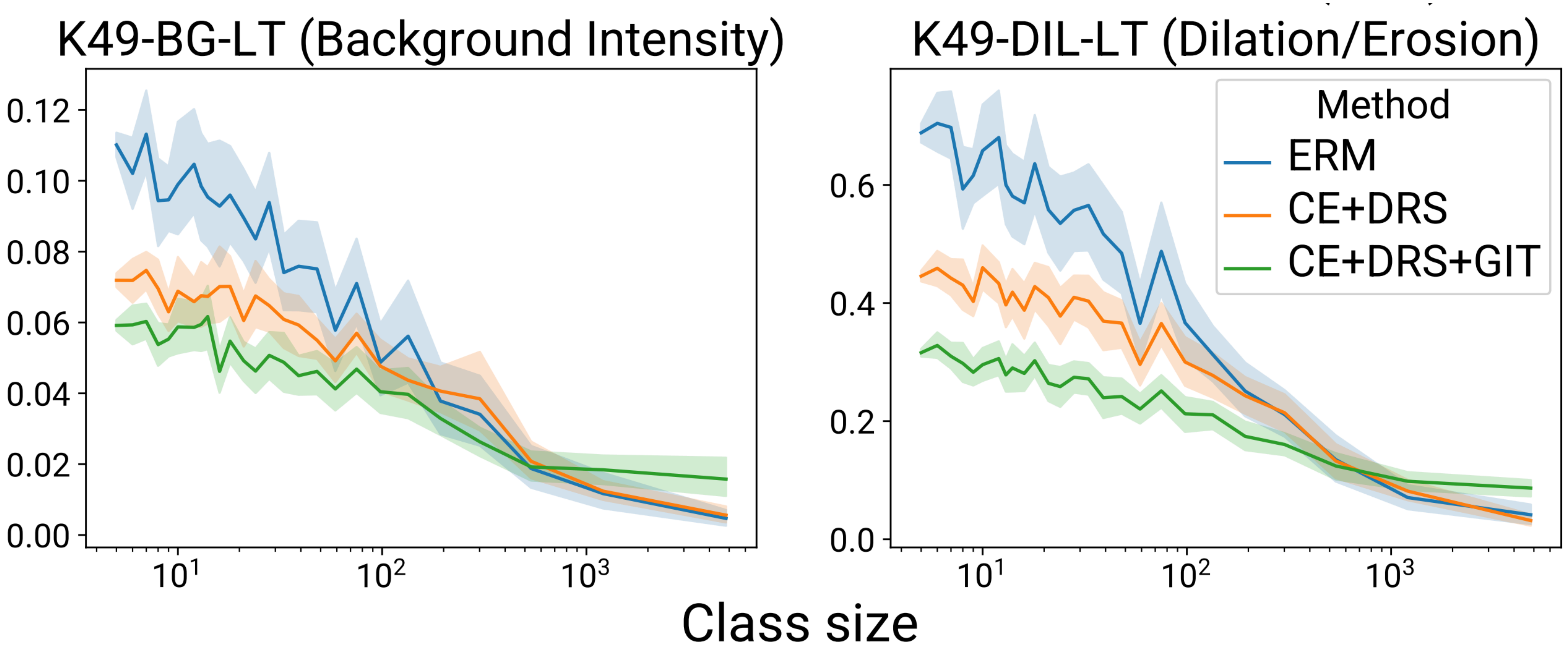
Antoniou et al. Data Augmentation GAN. 2017

Mariani et al. Data Augmentation with Balancing GAN. 2018

Zhou, Tajwar, Robey, Knowles, Pappas, Hasani, Finn. **Do Deep Networks Transfer Invariances Across Classes.** ICLR '21.

Does GIT improve invariance on small classes?

Invariance to T
(lower is better)



Yes! It also worsens invariance on well-represented classes, likely since generative model is imperfect.

—> Only apply augmentation to small classes

Zhou, Tajwar, Robey, Knowles, Pappas, Hasani, Finn. **Do Deep Networks Transfer Invariances Across Classes.** ICLR '21.

Do these improvements translate into better balanced accuracy?

Baseline	Strategy	Dataset	
		K49-BG-LT	K49-DIL-LT
ERM		42.29 ± 1.46	39.49 ± 1.47
CE+DRS		42.21 ± 1.36	39.48 ± 1.37
	+GIT	$\mathbf{49.99 \pm 1.25}$	$\mathbf{49.18 \pm 1.23}$
LDAM+DRS		54.08 ± 1.21	50.44 ± 1.24
	+GIT	$\mathbf{58.86 \pm 1.11}$	$\mathbf{56.76 \pm 1.11}$

4-10% improvement on K49

		Dataset		
Baseline	Strategy	GTSRB-LT	CIFAR-10 LT	CIFAR-100 LT
ERM		68.88 ± 1.75	70.74 ± 0.13	38.69 ± 0.32
CE + DRS		64.45 ± 1.15	74.28 ± 0.56	40.97 ± 0.40
	+GIT	$\mathbf{75.19 \pm 0.50}$	$\mathbf{77.25 \pm 0.18}$	$\mathbf{42.73 \pm 0.27}$
Focal + DRS		65.68 ± 2.09	73.51 ± 0.50	40.77 ± 0.21
	+GIT	$\mathbf{71.29 \pm 0.73}$	$\mathbf{76.87 \pm 0.14}$	$\mathbf{41.25 \pm 0.26}$
LDAM + DRS		77.25 ± 1.29	76.73 ± 0.74	43.21 ± 0.31
	+GIT	$\mathbf{81.39 \pm 0.98}$	$\mathbf{78.76 \pm 0.19}$	$\mathbf{44.35 \pm 0.2}$[illegible]

1-10% improvement

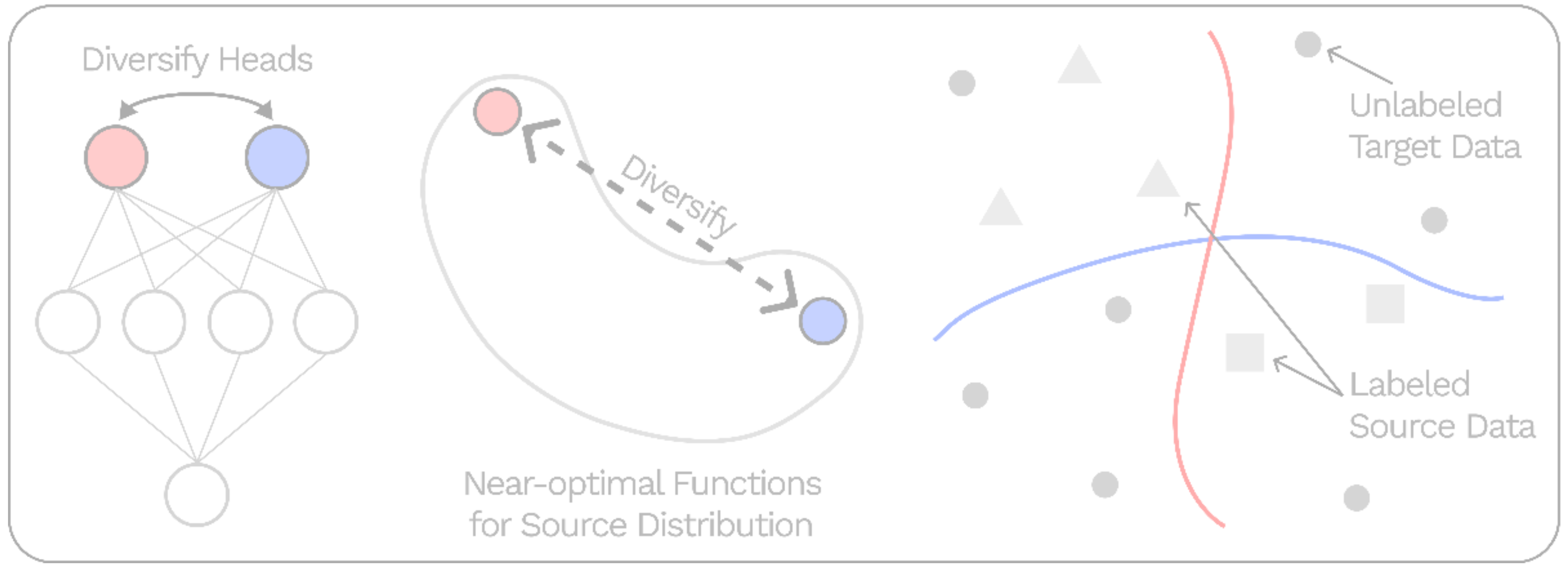
on GTSRB-LT, CIFAR-LT

Takeaway: Explicitly transferring invariances can significantly improve balanced accuracy.

Zhou, Tajwar, Robey, Knowles, Pappas, Hasani, Finn. **Do Deep Networks Transfer Invariances Across Classes**. ICLR '21.

Outline

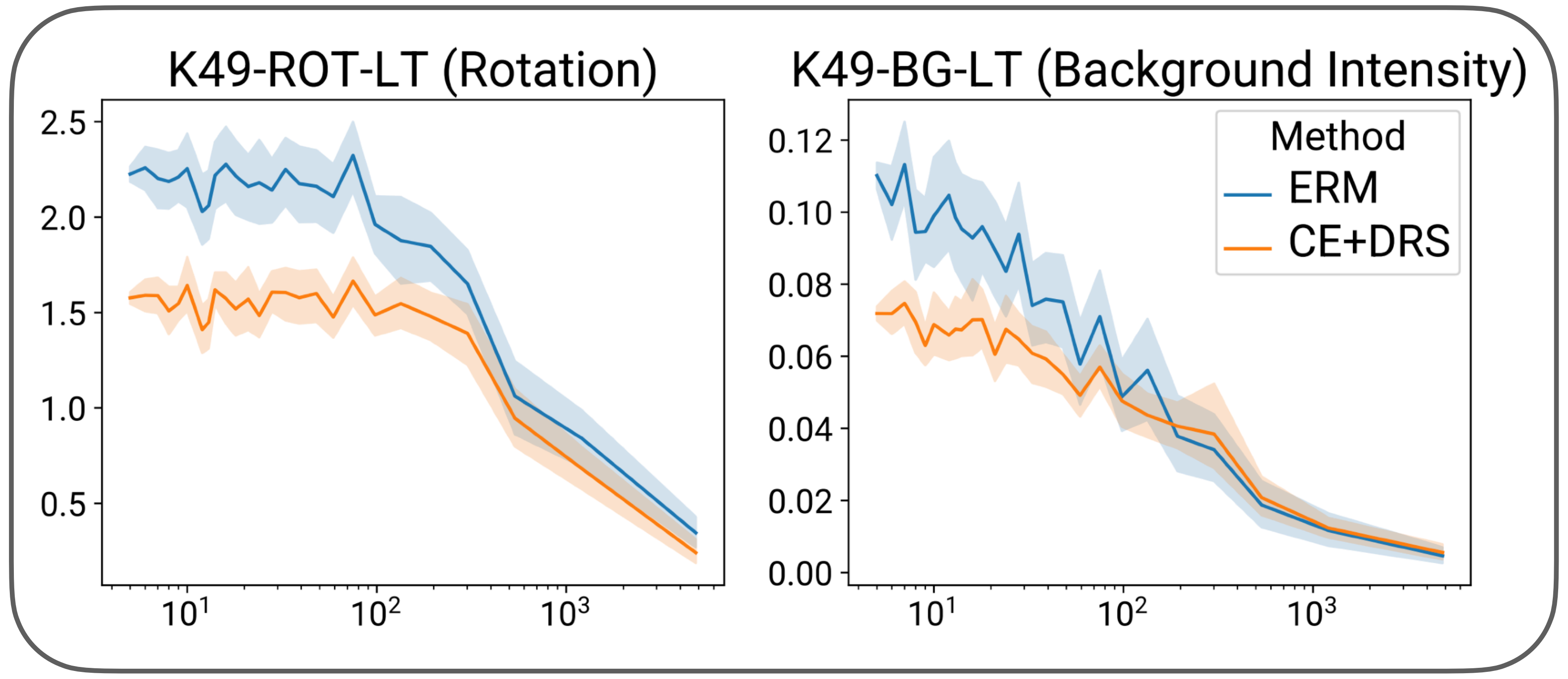
Addressing extreme covariate shift
via diverse ensembles



for supervised learning & reinforcement learning

Takeaway: Learning diverse classifiers & policies enables fast adaptation to OOD situations

Addressing label shift
via invariance transfer



for image classification

Takeaway: Invariances do not transfer across classes. Transferring them can help with label shift

Working on distribution shift?

WILDS

Benchmark with distribution shifts
arising in real-world applications.

wilds.stanford.edu

Questions?