In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness

Sang Michael Xie*, Ananya Kumar*, Robbie Jones*, Fereshte Khani,
Tengyu Ma, Percy Liang

ICLR 2021






Outline

- Robustness in remote sensing
- Empirical observations
- Theoretical insights
- In-N-Out algorithm
- Empirical results

Motivating example: remote sensing

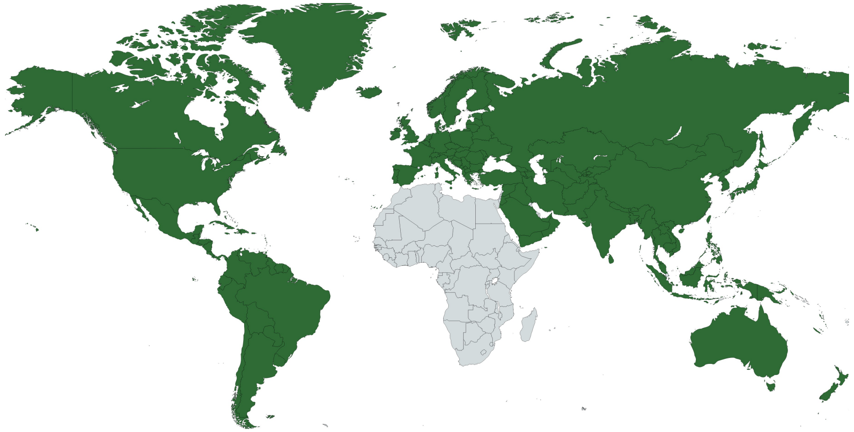
- **Task:**

x: Satellite image

y: Land cover type

{cropland, grassland, evergreen forest, ...}

- **Data:** Labels are expensive to collect/scarce in some areas, satellite imagery is everywhere



	In-distribution	Out-of-distribution (OOD)
Test accuracy (train on)	76%	**58%**

Robustness problems in remote sensing

Crop yield prediction (Wang et al. 2020):

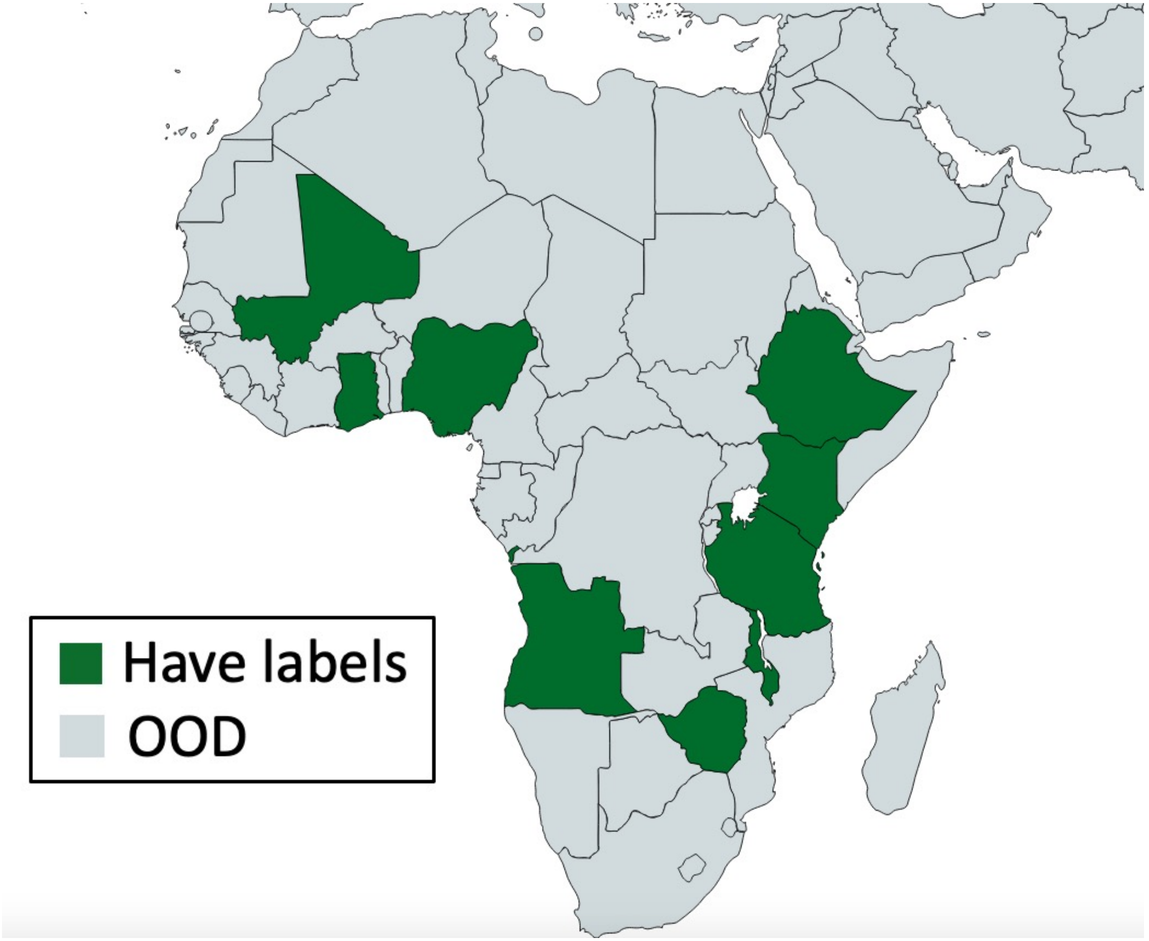
- Crop yield: how much crop will this field produce?
- Important for **improving agricultural practices** and **food security**
- Have good physics-based crop yield models if we know the crop type
- Can we predict crop type from satellites?
- Labeling crop type requires sending workers to the field -> expensive -> **scarce labeled data**

$\xrightarrow{\text{ML?}}$ {Corn, Soybean, Wheat, ...} $\longrightarrow$ Crop-specific yield model $\longrightarrow$ Crop yield

Robustness problems in remote sensing

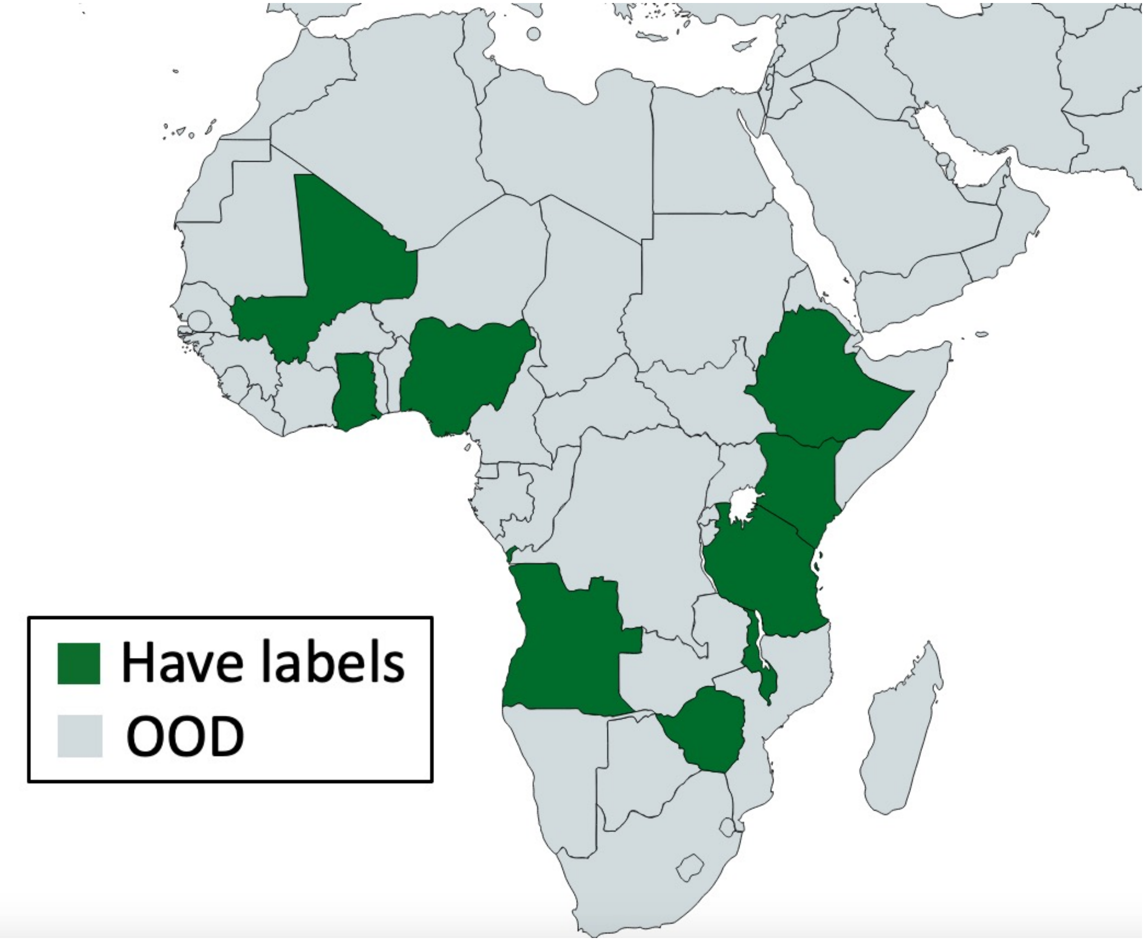
Poverty mapping:

- High-resolution poverty maps **improve policy and humanitarian decisions**
- Expensive to conduct surveys for collect labels ($400,000 to $1.5 million)
- Most African countries haven't had a survey in > 5-10 years
- Even with survey data, we have poor spatial resolution (Uganda dataset with 2,716 households)



Robustness problems in remote sensing

- Only some domains have labels - **how do we generalize globally?**
- Not possible generally without **additional structure**
- Can **unlabeled data** and **auxiliary information** from unseen domains help on out-of-distribution (OOD) examples?



Outline

- Robustness in remote sensing
- Empirical observations
- Theoretical insights
- In-N-Out algorithm
- Empirical results

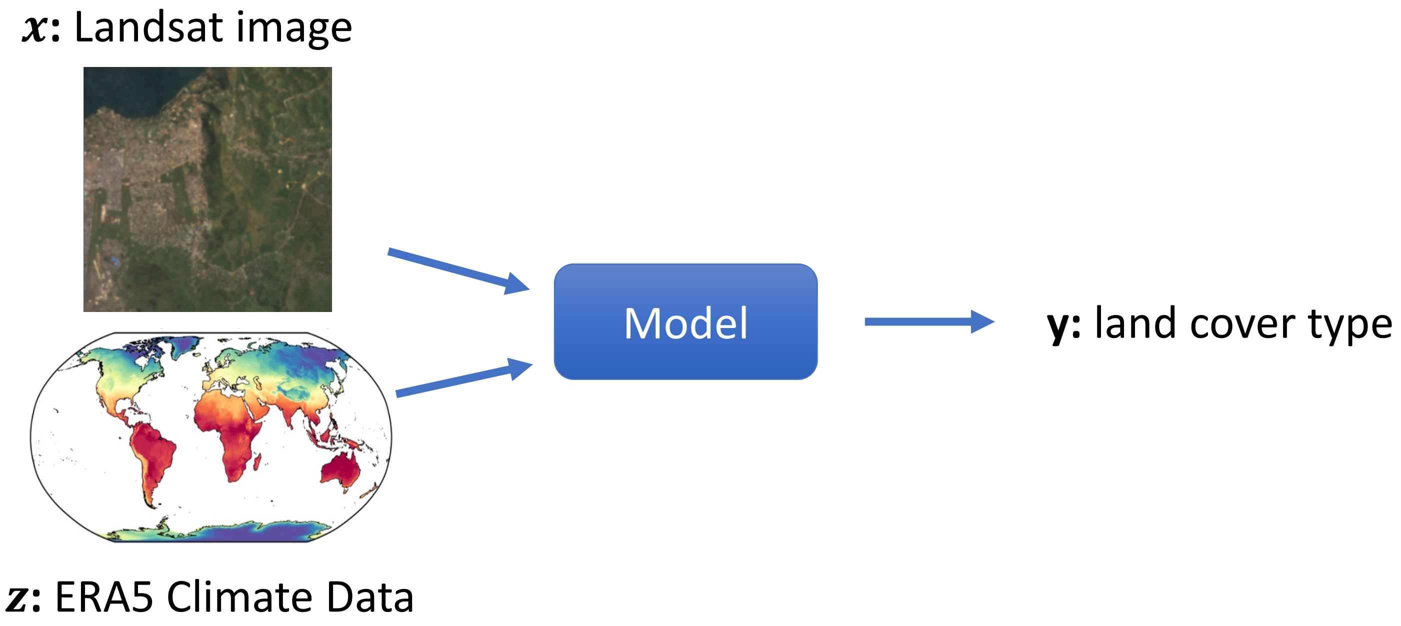
Setting

- Inputs x (satellite images), Outputs y (land cover type)
- **Auxiliary information $\boldsymbol{z}$** (climate data from other satellites)
- In-distribution (ID): few labeled (x, y, z) tuples
- Both ID and OOD: many unlabeled (x, z) pairs

How do we use unlabeled data and auxiliary information to improve OOD?

Baseline 1: Aux-inputs

- **Aux-inputs**: use z as extra input features ($x, z \rightarrow y$)



Aux-inputs can hurt OOD accuracy

Aux-inputs improves in-distribution (ID) accuracy (countries with labeled data)

ID	Cropland	Landcover
No aux	94.5	76
Aux-inputs	**95.3**	**77**

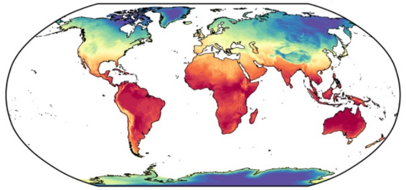
Surprisingly, **aux-inputs can hurt OOD** accuracy (unseen countries) because z can shift a lot and be misleading OOD

OOD	Cropland	Landcover
No aux	90	58
Aux-inputs	**84**	**55**

Datasets: Cropland (Wang et al. 2020) and Landcover (Rußwurm et al. 2020)

Aux-inputs can hurt OOD accuracy: Intuition

- Climate info can help predict land cover type



→ **y:** land cover type

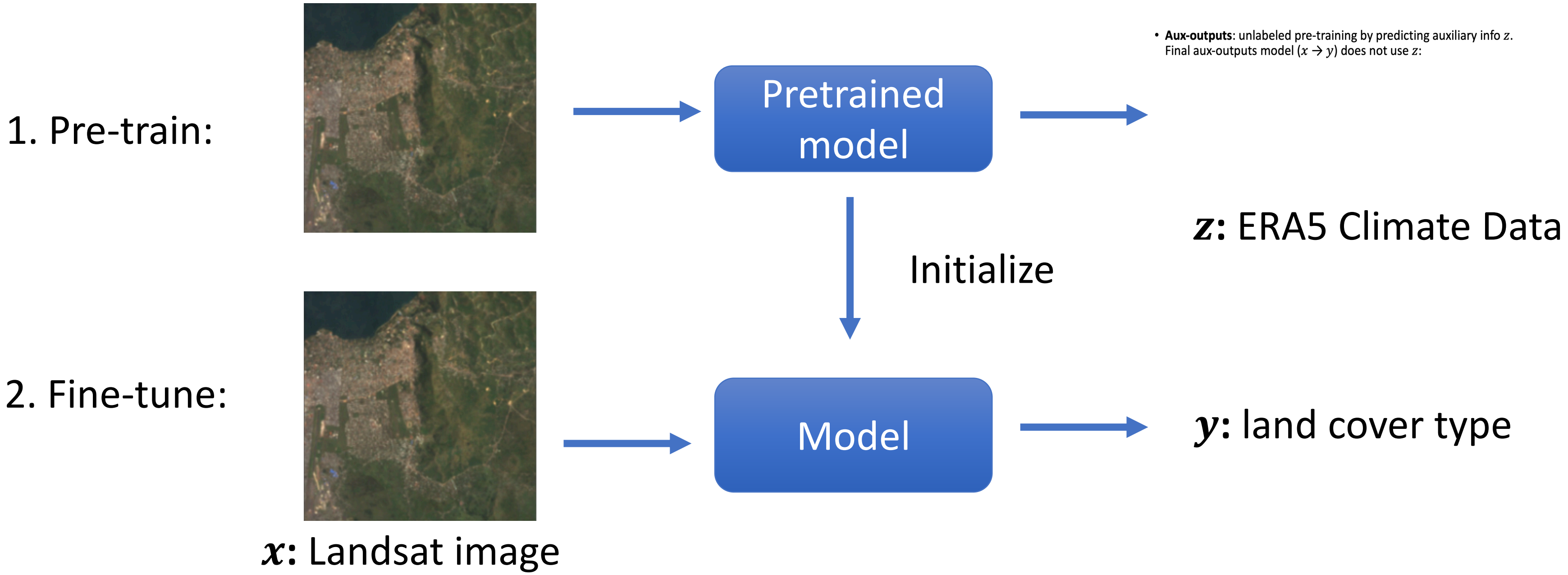
$\boldsymbol{z}$: ERA5 Climate Data

- However, climate info is noisy and changes with spatial location
  - More possible spurious correlations
  - May need to extrapolate on unseen climates in OOD data

Auxiliary info may introduce additional spurious correlations

Baseline 2: Aux-outputs

- **Aux-outputs**: unlabeled pre-training by predicting auxiliary info z. Final aux-outputs model ($x \to y$) does not use z:



Aux-outputs improves OOD

But **ID accuracy is not as good as aux-inputs,** since it doesn't use extra info in z

ID	Cropland	Landcover
No aux	94.5	76
Aux-inputs	**95.3**	**77**
Aux-outputs	**95.1**	**73**

Aux-outputs improves OOD accuracy by using unlabeled data and **using z to extract useful features only**

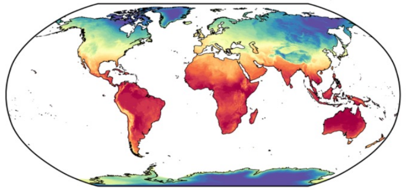
OOD	Cropland	Landcover
No aux	90	58
Aux-inputs	**84**	**55**
Aux-outputs	**92**	**61**

Aux-outputs improves OOD: Intuition

- Model must learn useful land features to predict climate

x: Landsat image

trees
water
grass
sand



z: ERA5 Climate Data

- Since climate data can be noisy, we learn these features on a large unlabeled dataset

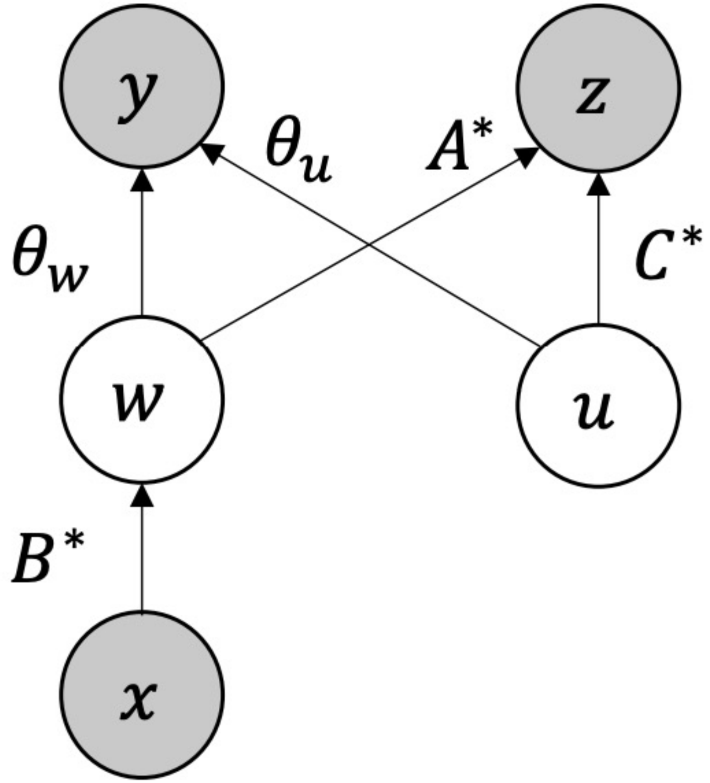
Predicting auxiliary info on unlabeled data extracts useful features

Outline

- Robustness in remote sensing
- Empirical observations
- Theoretical insights
- In-N-Out algorithm
- Empirical results

Multi-task linear regression setting

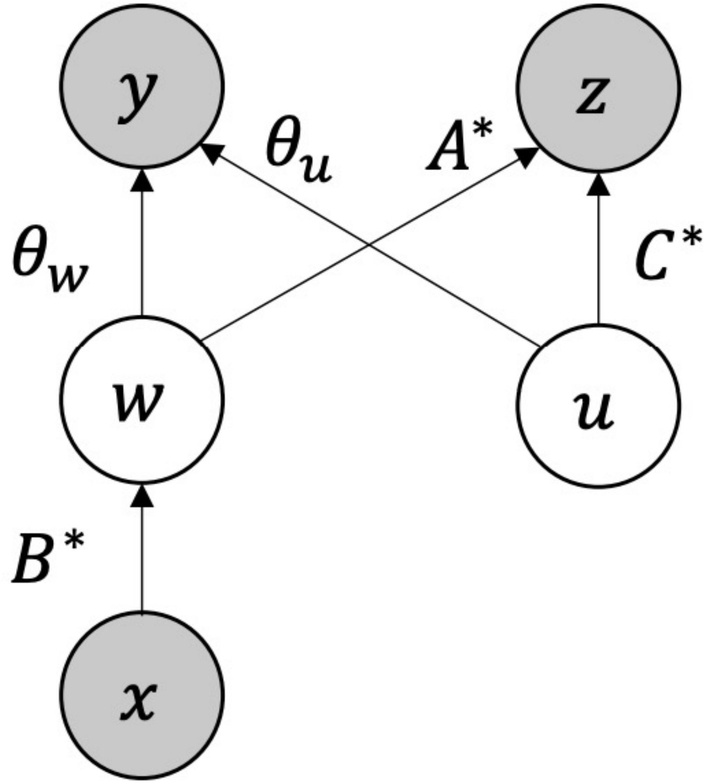
- Inputs $x \in \mathbb{R}^d$
- Targets $y \in \mathbb{R}$ with noise $N(0, \sigma^2)$
- Auxiliary info $z \in \mathbb{R}^T$
- Latent features $w \in \mathbb{R}^k$ with $k \leq d$
- Latent noise $u \in \mathbb{R}^m$
- x, u can shift OOD



All arrows describe a linear relation with true parameter labeled on the arrow

Multi-task linear regression setting

- Well-specified linear regression setting
$$y = \theta_w^\top w + \theta_u^\top u + \epsilon$$
$$z = A^* w + C^* u$$
- **Baseline:** learn $\hat{\theta}^\top x$
- **Aux-inputs:** learn $\hat{\theta}_x^\top x + \hat{\theta}_z^\top z$
- **Aux-outputs:**
  - Pretrain: learn $\hat{z} = \hat{A}\hat{B}x$ to learn feature space $\widehat{w} = \hat{B}x$
  - Fine-tune: learn $\hat{y} = \hat{\theta}_w^\top \widehat{w}$



All arrows describe a linear relation with true parameter labeled on the arrow

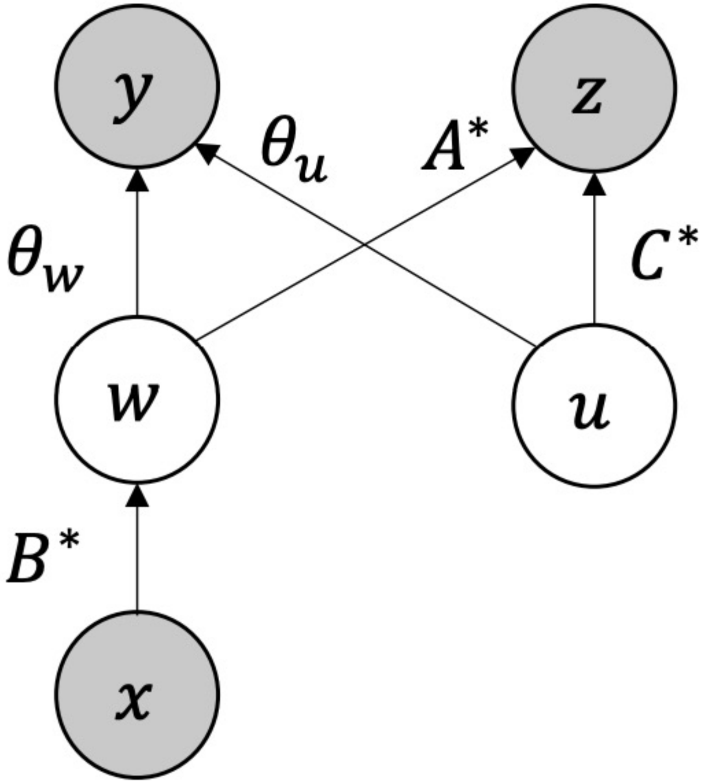
Aux-inputs helps ID, but can hurt OOD

ID

- Access to auxiliary z recovers unobserved u: *aux-inputs better than baseline by improving Bayes-opt error*

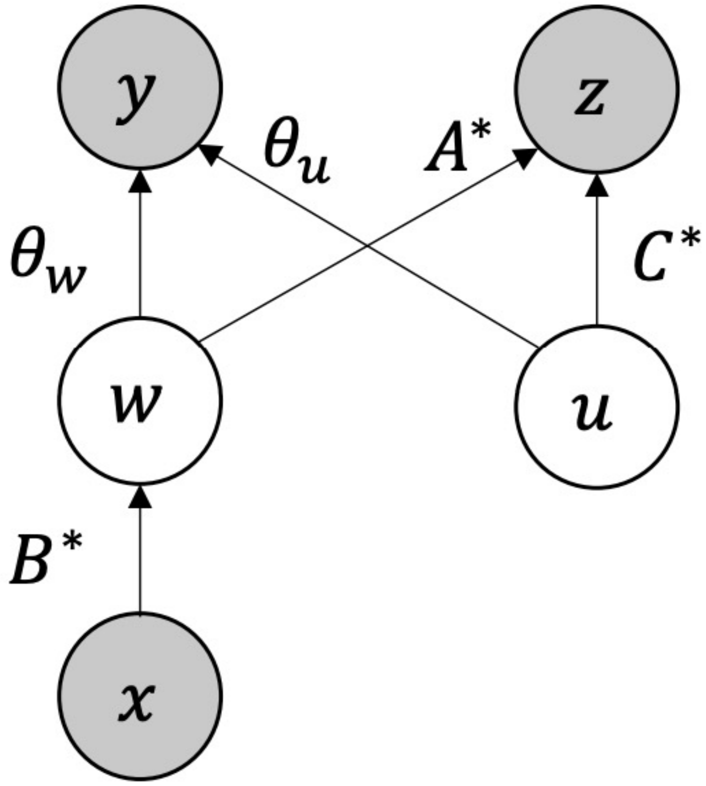
OOD

- Latent noise u can shift OOD making z non-robust – *always exists some shift where aux-inputs worse than using no auxiliary info*



Aux-outputs improves OOD robustness

- Pre-training to predict z learns latent features w, reducing to a $w \to y$ problem (lower dimensional)
- **ID:** expected excess risk is $\frac{d\sigma^2}{n} \Rightarrow$ aux-outputs trivially improves with $k \leq d$
- **OOD:** worst-case risk depends on data conditioning, and w can have worse conditioning



However, we prove that pre-training improves expected risk on **arbitrary covariate shifts!**

Self-training for further gains

- Self-training: use a teacher model to pseudo-label unlabeled data
- Suppose aux-inputs generates accurate pseudolabels on ID points (formally, irreducible noise σ^2 is small)
- On unlabeled ID data:
  - Aux-inputs model better than baseline
  - Pseudolabels $x, z \to \hat{y}$ are accurate
  - Increases number of effective labeled examples

We prove that self-training improves OOD error even more over aux-outputs as $\boldsymbol{\sigma^2 \to 0}$ (as pseudolabels are more accurate)

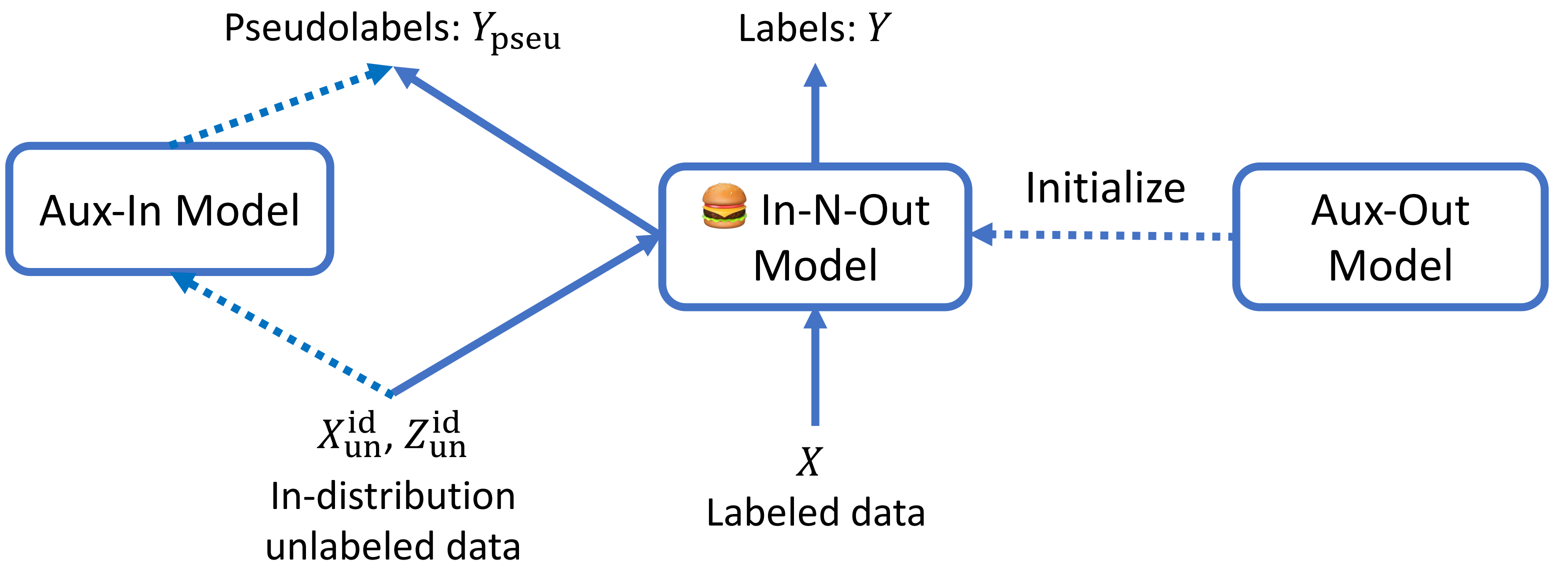
Outline

- Robustness in remote sensing
- Empirical observations
- Theoretical insights
- In-N-Out algorithm
- Empirical results

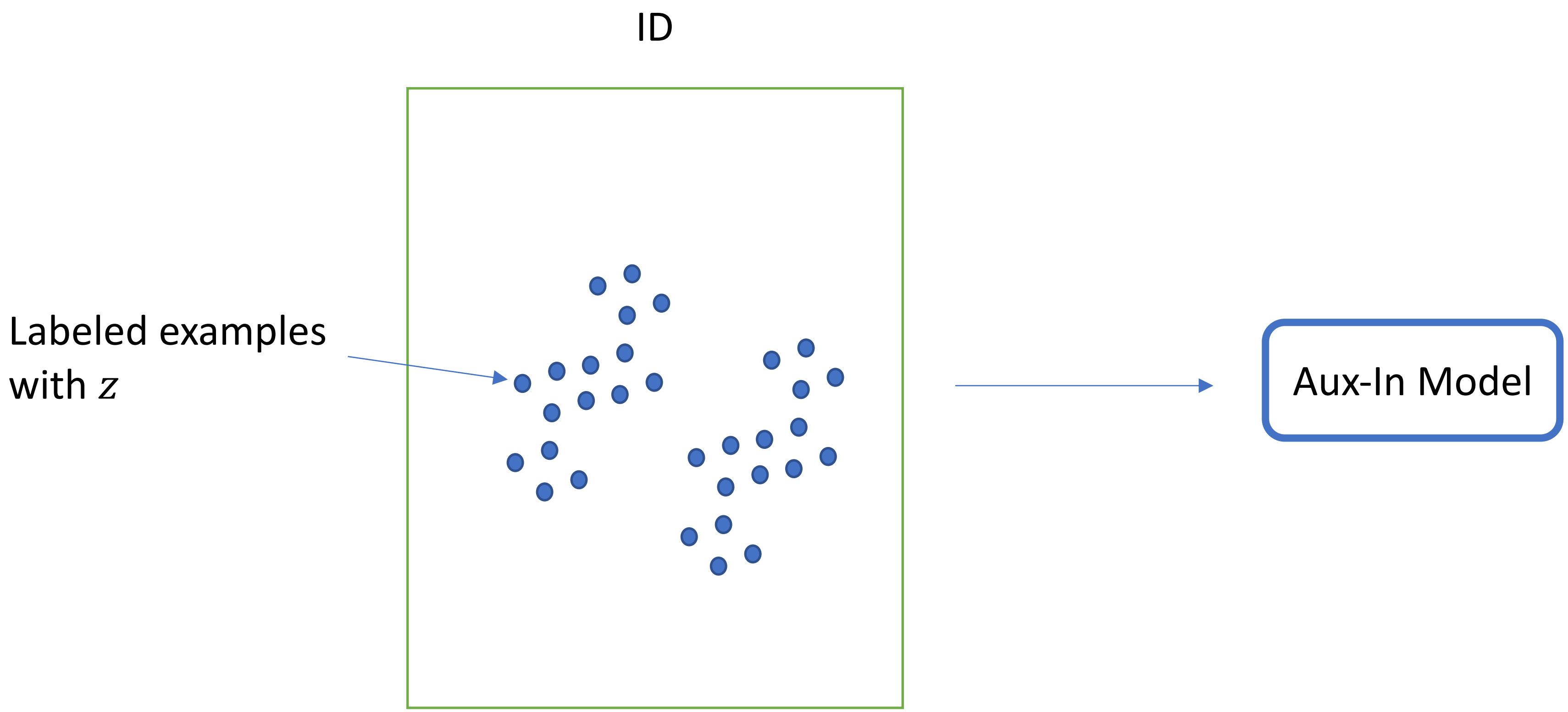
In-N-Out: best of both worlds

Aux-inputs good ID, aux-outputs good OOD, combine using self-training

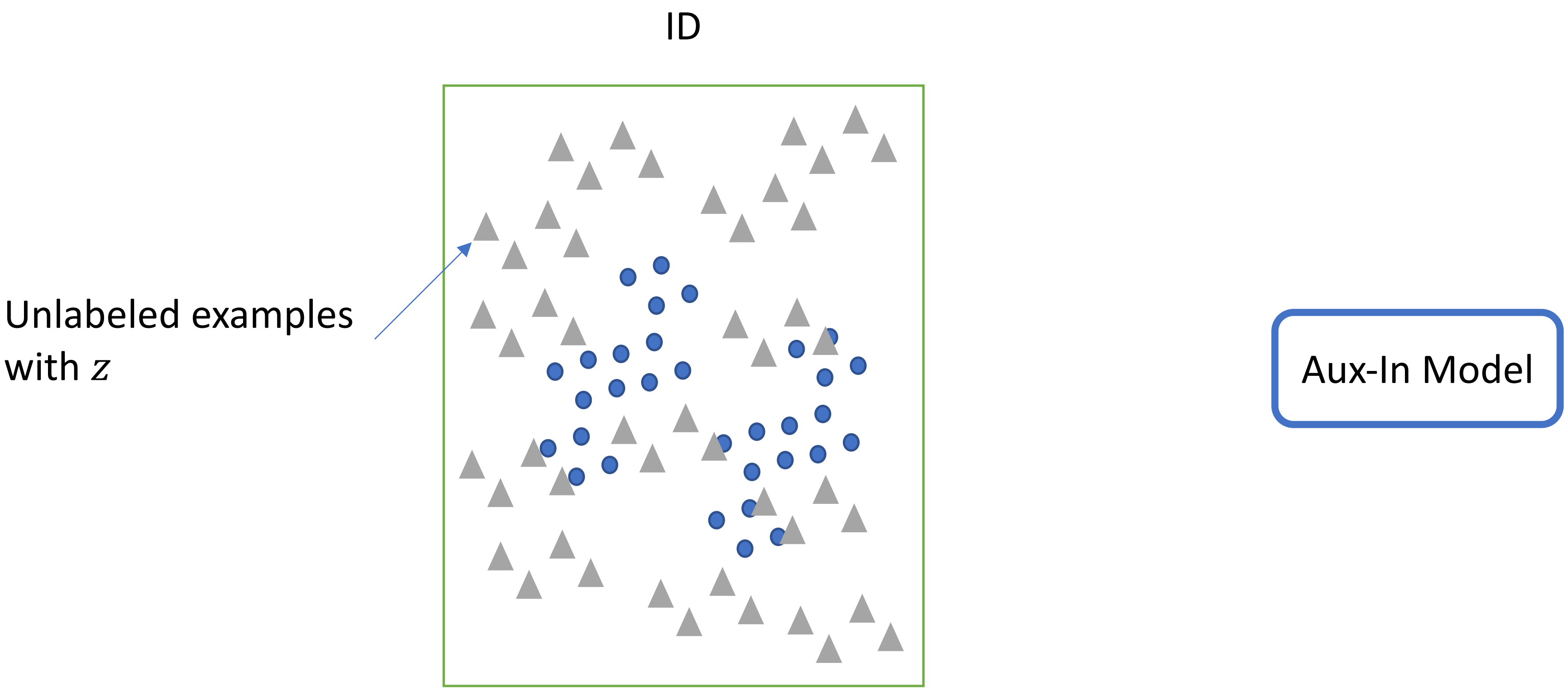
1. Use aux-inputs model to pseudolabel unlabeled ID data
2. Initialize In-N-Out model from aux-outputs model (pre-training)
3. Fine-tune In-N-Out model with original labels and pseudolabels (self-training)



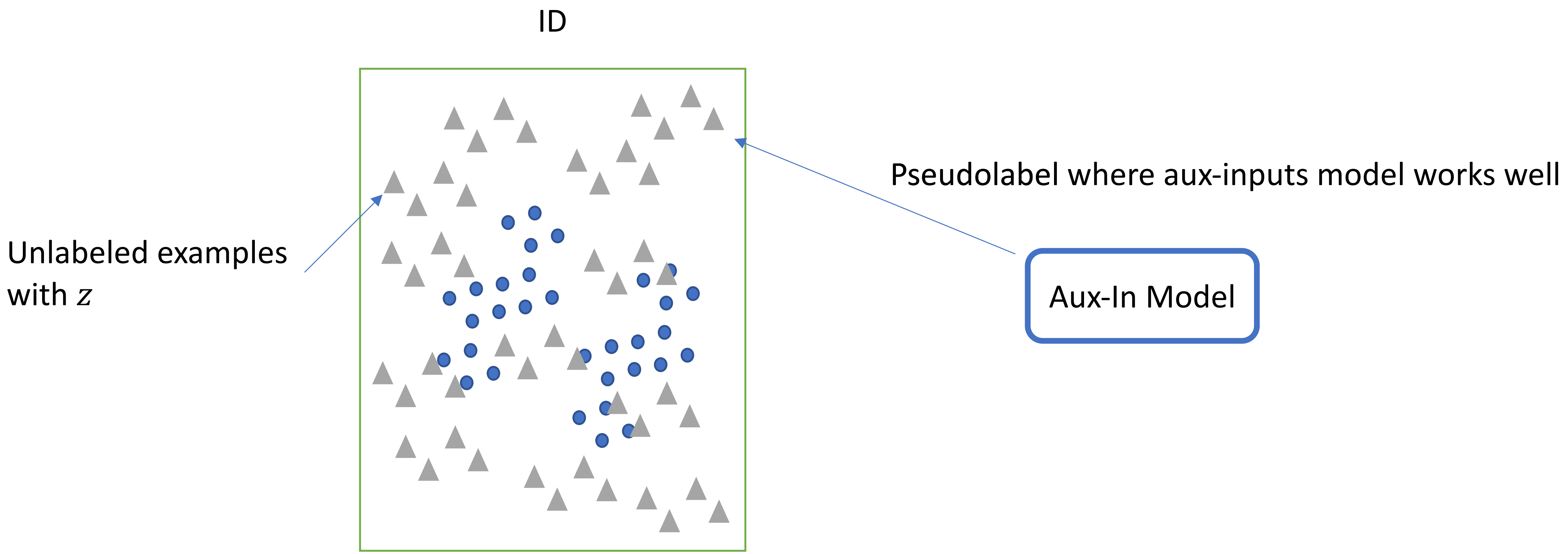
In-N-Out: Illustrated



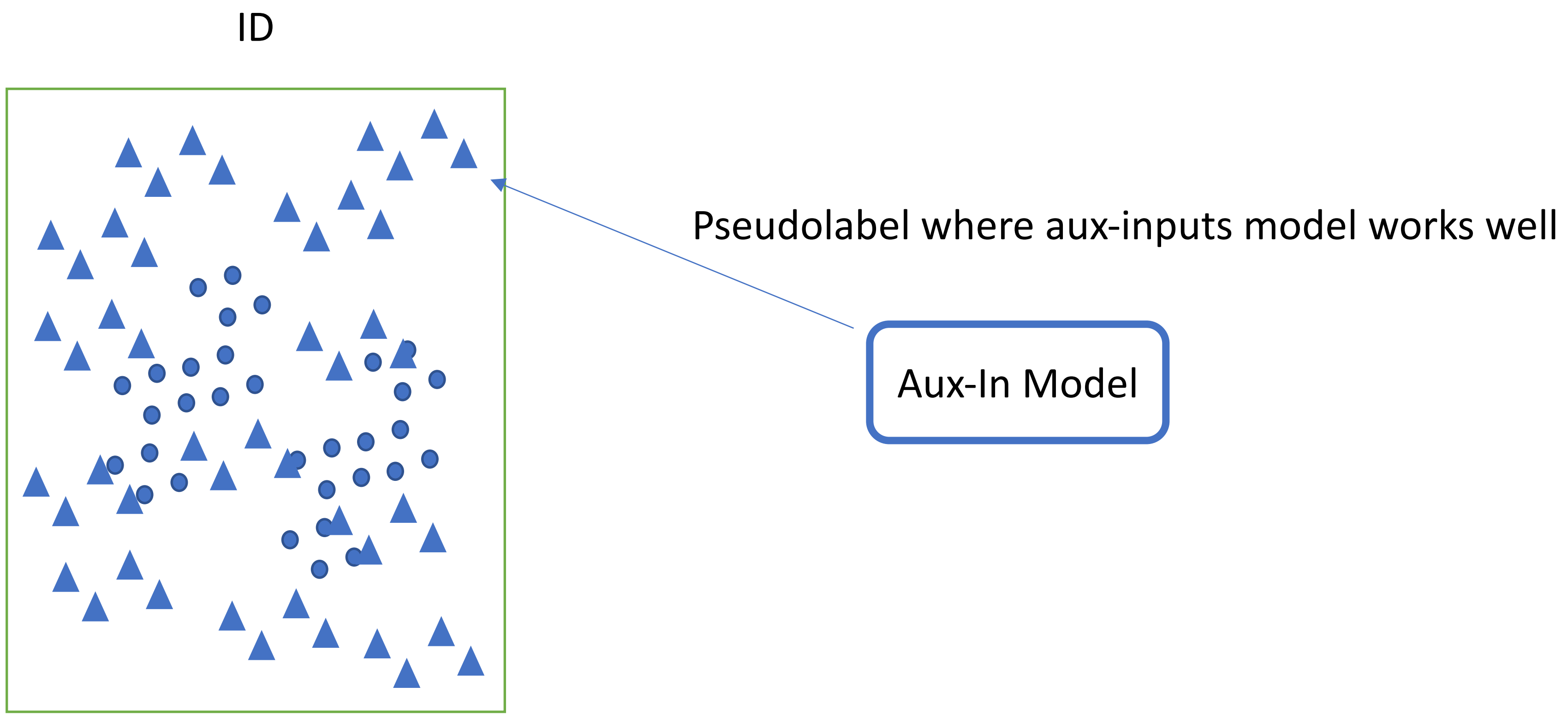
In-N-Out: Illustrated



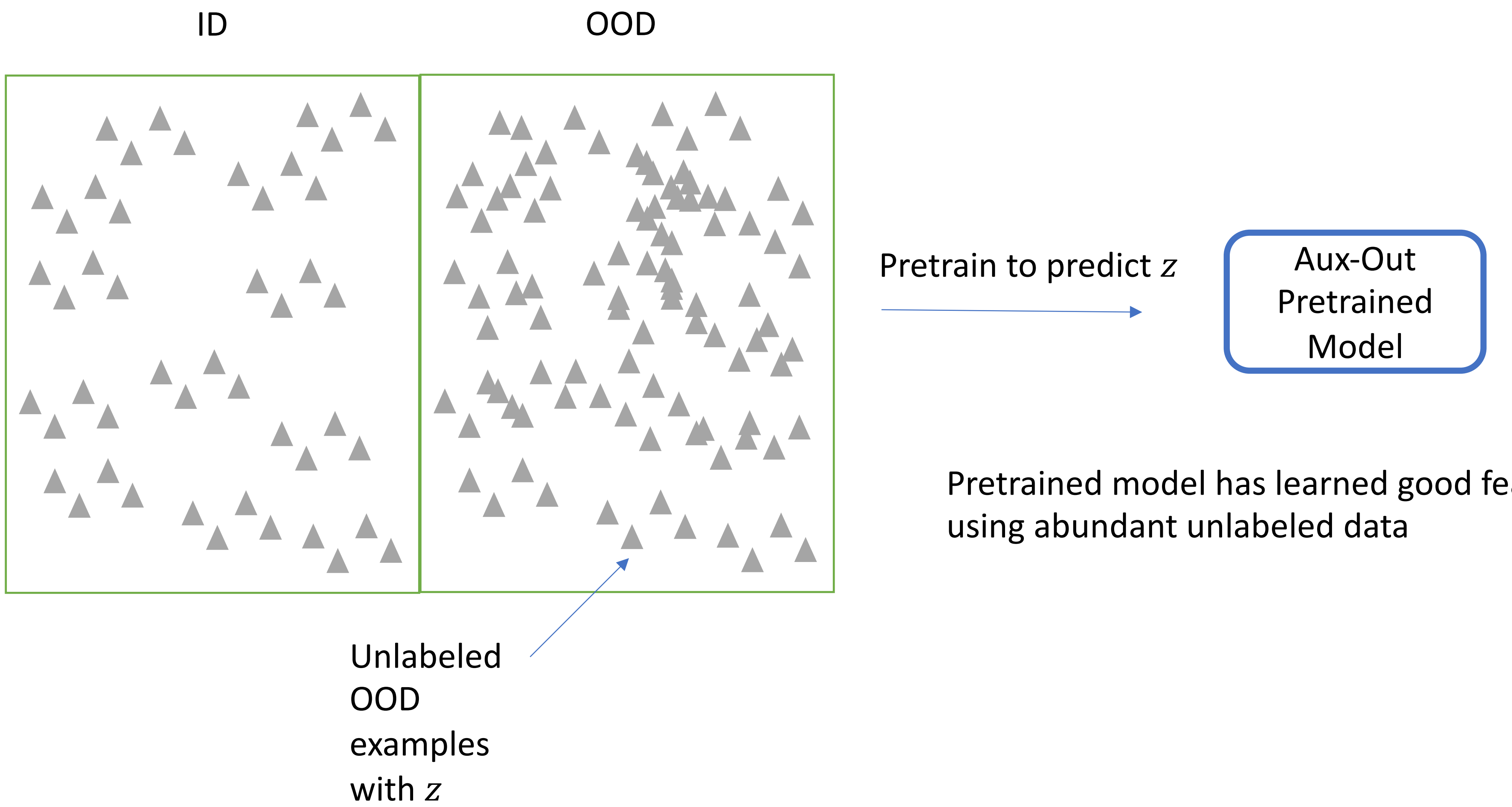
In-N-Out: Illustrated



In-N-Out: Illustrated



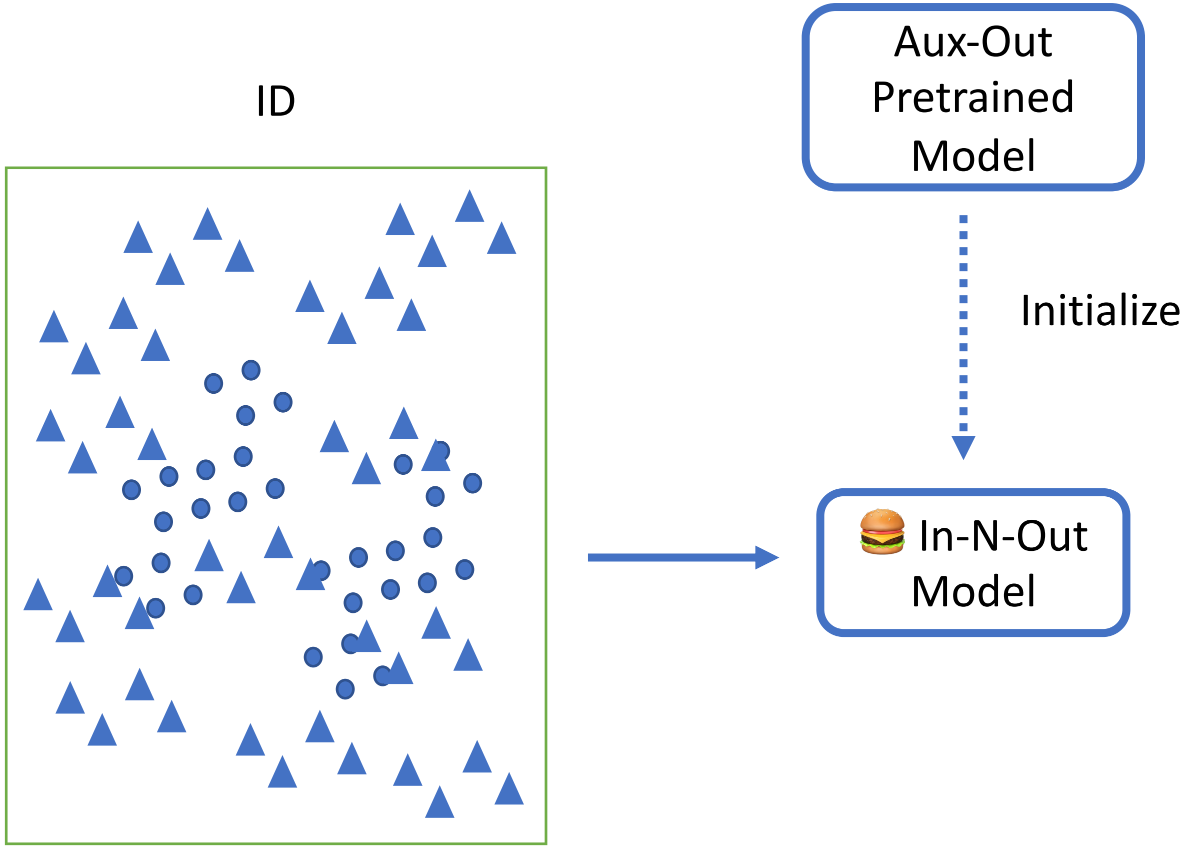
In-N-Out: Illustrated



Pretrained model has learned good features using abundant unlabeled data

In-N-Out: Illustrated

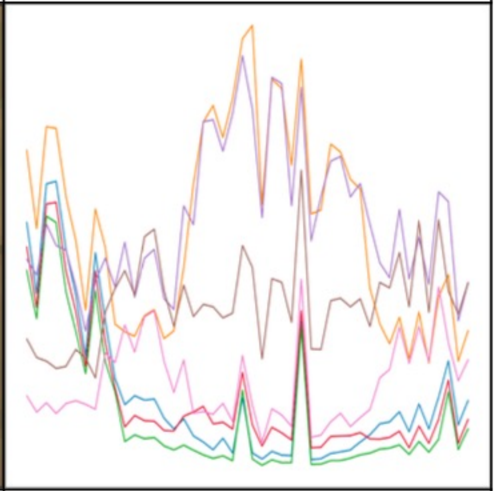
Pseudolabeled and original examples provide a larger dataset for fine-tuning



Outline

- Robustness in remote sensing
- Empirical observations
- Theoretical insights
- In-N-Out algorithm
- Empirical results

Datasets

	CelebA	Cropland	Landcover
Visualization (x)			
Aux Info (z)	7 binary attributes	Vegetation, Lat/Lon	Meteorological Data
Target (y)	Male/female?	Cropland/not cropland?	Land cover class
ID-Split	People without hats	IA, MN, IL	Outside Africa
OOD-Split	People with hats	IN, KY	Africa

Empirical results

- In-N-Out improves over all baselines on both ID and OOD (bold are within error bars)

ID	CelebA	Cropland	Landcover
No aux	91	95	76
Aux-inputs	92	**95**	77
Aux-outputs	**94**	95	73
In-N-Out	**94**	**96**	**77**

OOD	CelebA	Cropland	Landcover
No aux	73	90	58
Aux-inputs	77	84	55
Aux-outputs	78	92	61
In-N-Out	**80**	**92**	**63**

Model comparisons

	ID	OOD
Aux-inputs (use z as input feature) • More potential spurious correlations	✓	✗
Aux-outputs (use z as pre-training output) • Learn better features for robustness	✗	✓
In-N-Out (use z as input and output) • Use spurious correlations for robustness	✓	✓

Ablations (only pre-training or self-training)

- In-N-Out improves over only self-training or only pretraining (aux-outputs) on both ID and OOD accuracy

ID	CelebA	Cropland	Landcover
In-N-Out (no pretrain)	93.8	94.9	76.5
Aux-outputs	**94.0**	95.1	72.5
In-N-Out	**93.8**	**95.5**	**77.1**

OOD	CelebA	Cropland	Landcover
In-N-Out (no pretrain)	78.5	91.2	59.2
Aux-outputs	77.7	91.6	61.0
In-N-Out	**80.4**	**92.2**	**62.6**

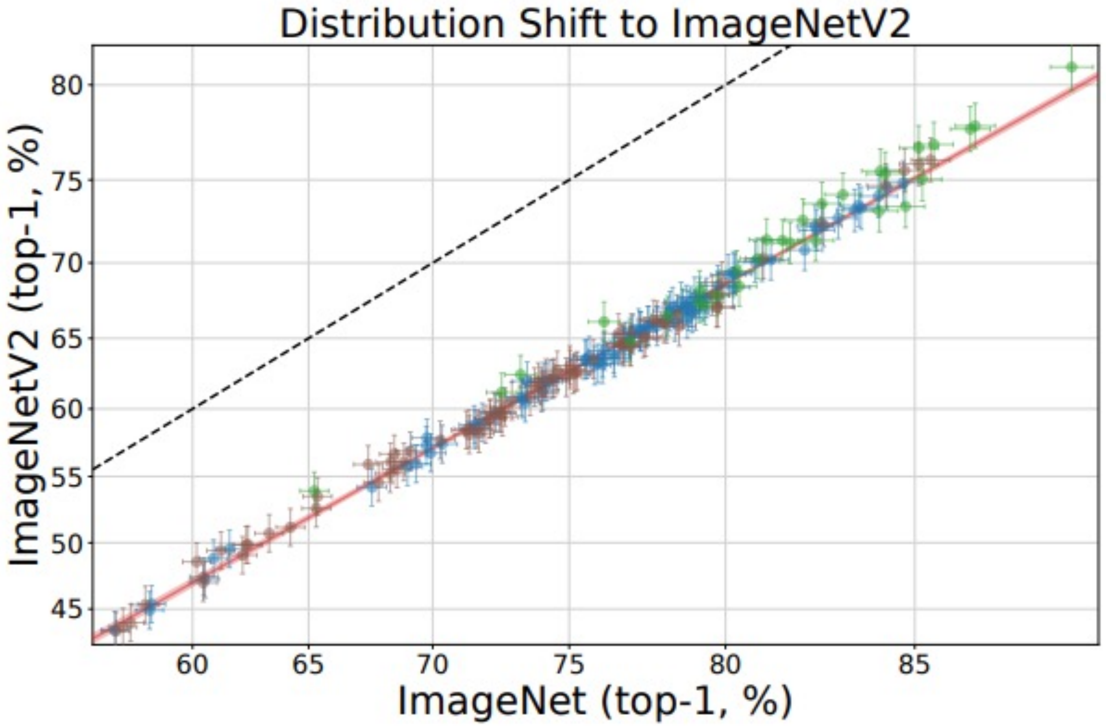
OOD unlabeled data is important for OOD

- Pre-training on ID unlabeled data vs OOD unlabeled data
- We standardized unlabeled data size (much smaller data than previous tables) and compare on Landcover

Unlabeled data used	ID Acc	OOD Acc
Only ID	69.7	57.7
Only OOD	69.9	**59.3**
Both	**70.1**	**59.8**

Reversing the ID-OOD correlation

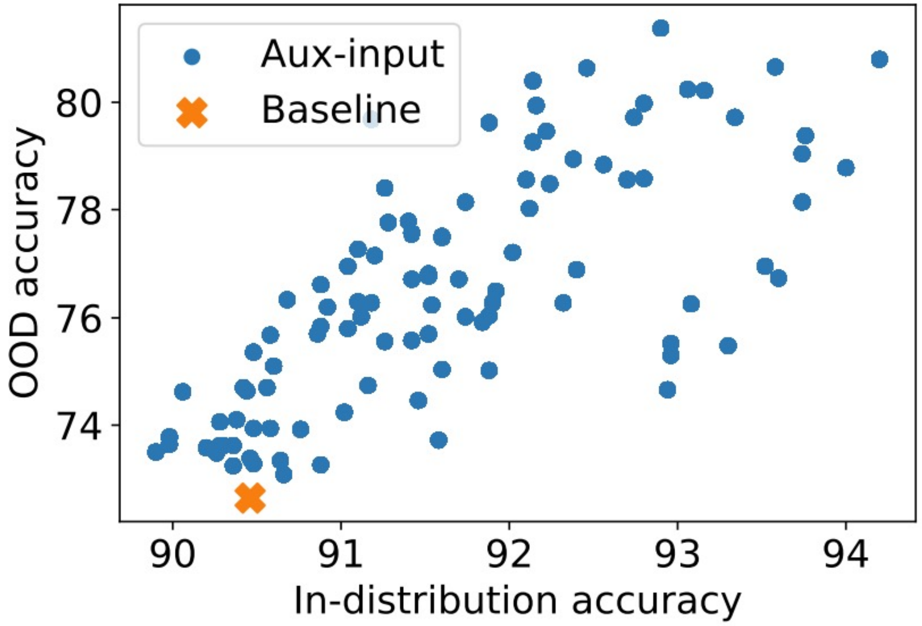
- Numerous works (Taori et al. 2020, Recht et al. 2020, Miller et al. 2021) show that ID accuracy correlates with OOD accuracy on curated benchmark datasets
- Perhaps we just have to improve ID accuracy to improve OOD?
- However, we showed on real-world data, adding features (aux-inputs) can buck the trend
- Can this phenomenon happen on curated benchmarks?



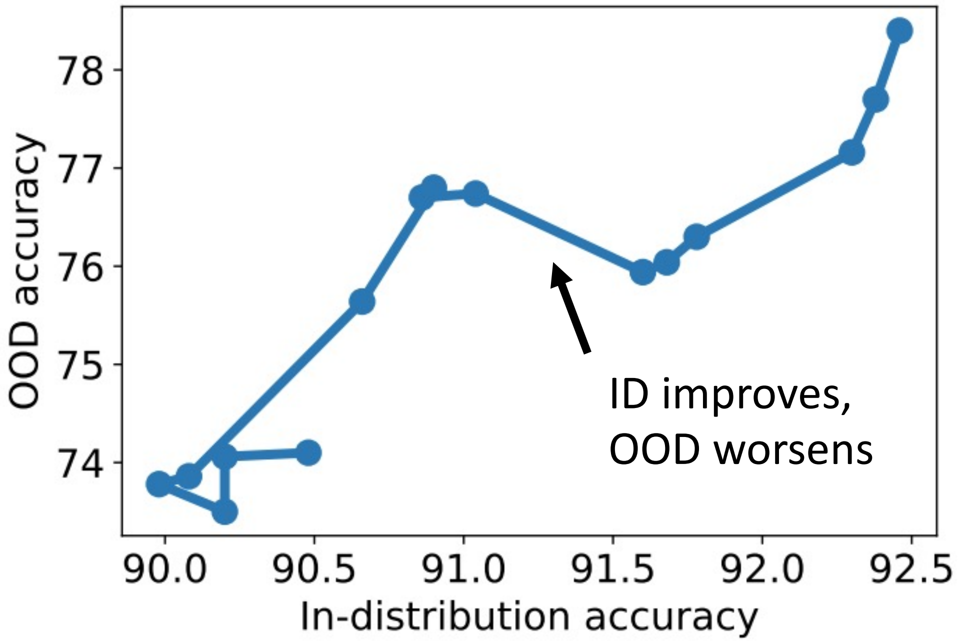
Taori et al. 2020

Reversing the ID-OOD correlation

On CelebA, adding features usually leads to better ID and OOD accuracy



Consider adding a sequence of features one-by-one. We find that almost any sequence has instances where ID-OOD accuracy are anti-correlated



Takeaways

- Real-world tasks require OOD generalization
- Adding features as inputs improves ID accuracy, but can hurt OOD
- Pre-training to predict features as outputs usually improves OOD accuracy
- In-N-Out combines these with pre-training and self-training to give gains both ID and OOD
- OOD unlabeled data is important for OOD benefits

Thanks!

Acknowledgements: Sherrie Wang, Andreas Schlueter, Daniel Levy, Albert Gu, Shyamal Buch, Pang Wei Koh, Shiori Sagawa

This work was supported by an Open Philanthropy Project Award, an NSF Frontier Award as part of the Center for Trustworthy Machine Learning (CTML). SMX was supported by an NDSEG Fellowship. AK was supported by a Stanford Graduate Fellowship. TM was partially supported by the Google Faculty Award, JD.com, Stanford Data Science Initiative, and the Stanford Artificial Intelligence Laboratory.