Making Your First Choice: To Address Cold Start Problem in Medical Active Learning

Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu,
Bihan Wen, Alan Yuille, Zongwei Zhou

MIDL
Nashville 2023

About Me

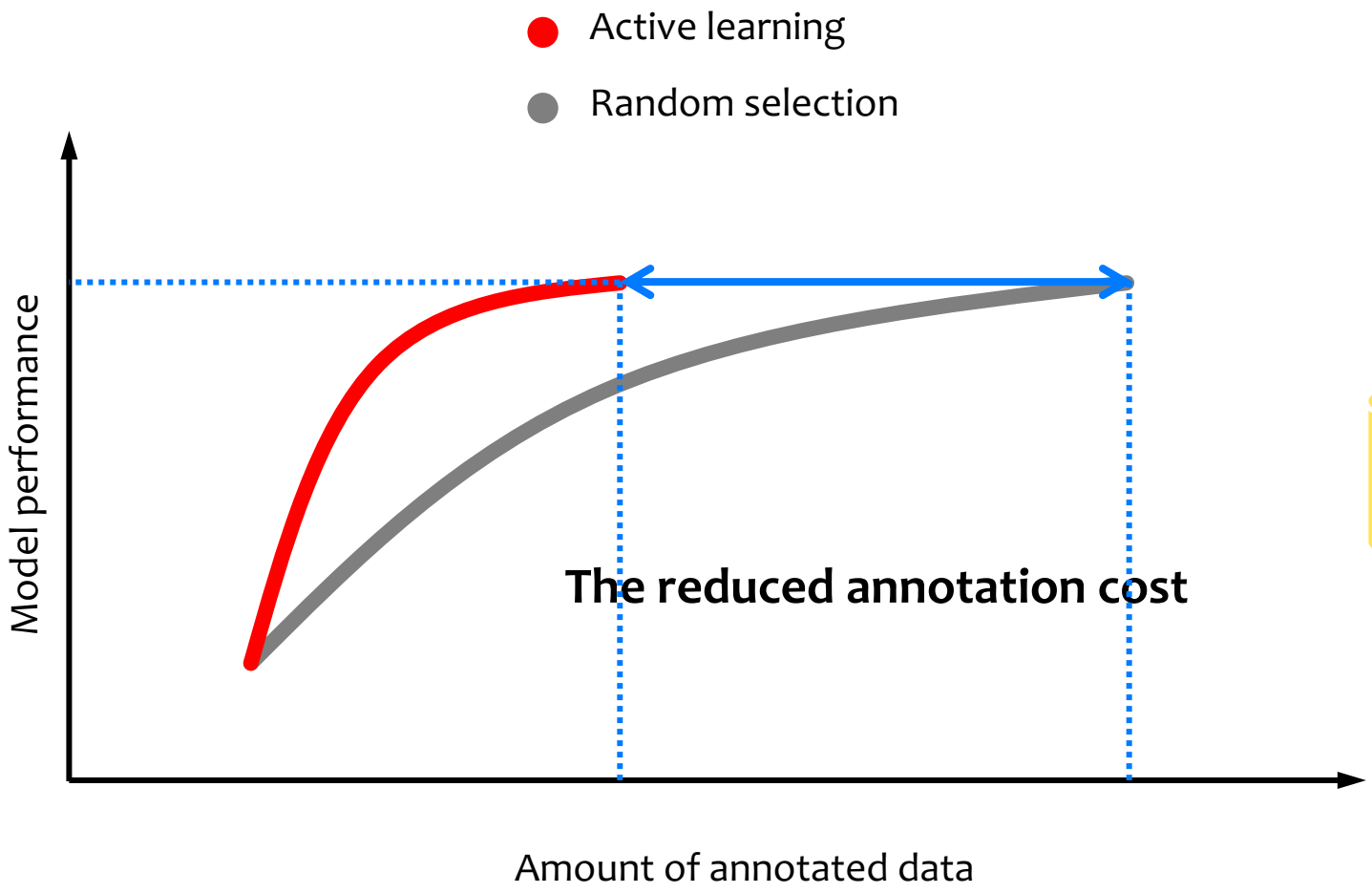
Liangyu Chen 陈亮宇
https://c-liangyu.github.io/

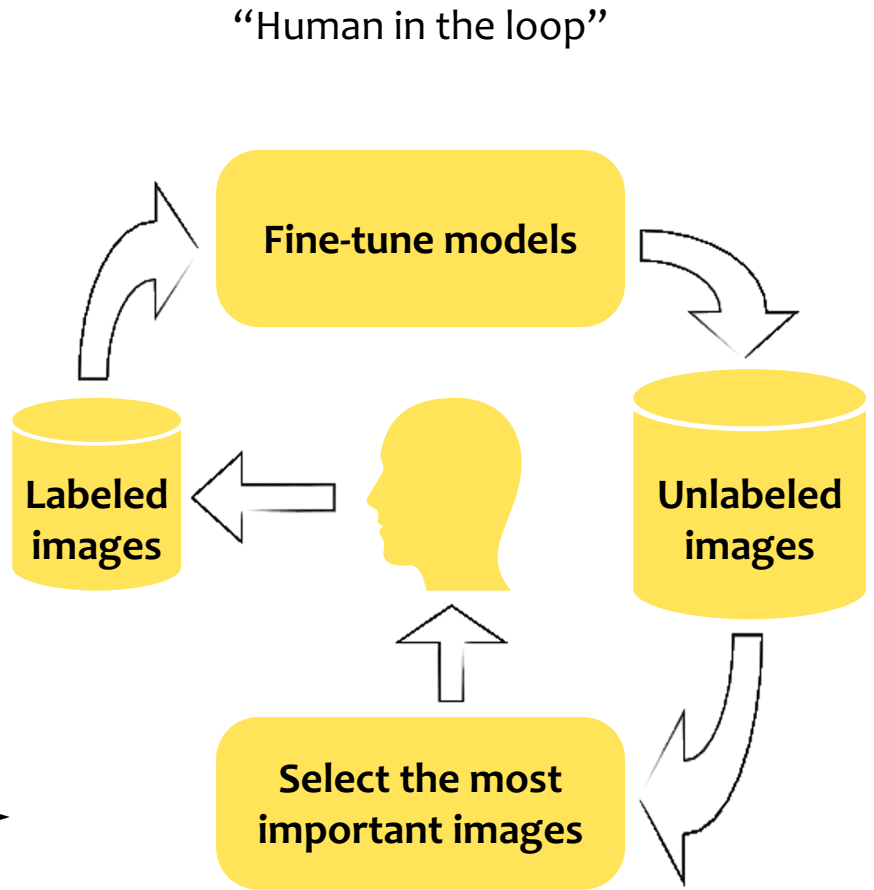
Research Engineer at Nanyang Technological University

Research:
- Generalizable and data-efficient machine learning
- Interactive or assistive AI systems

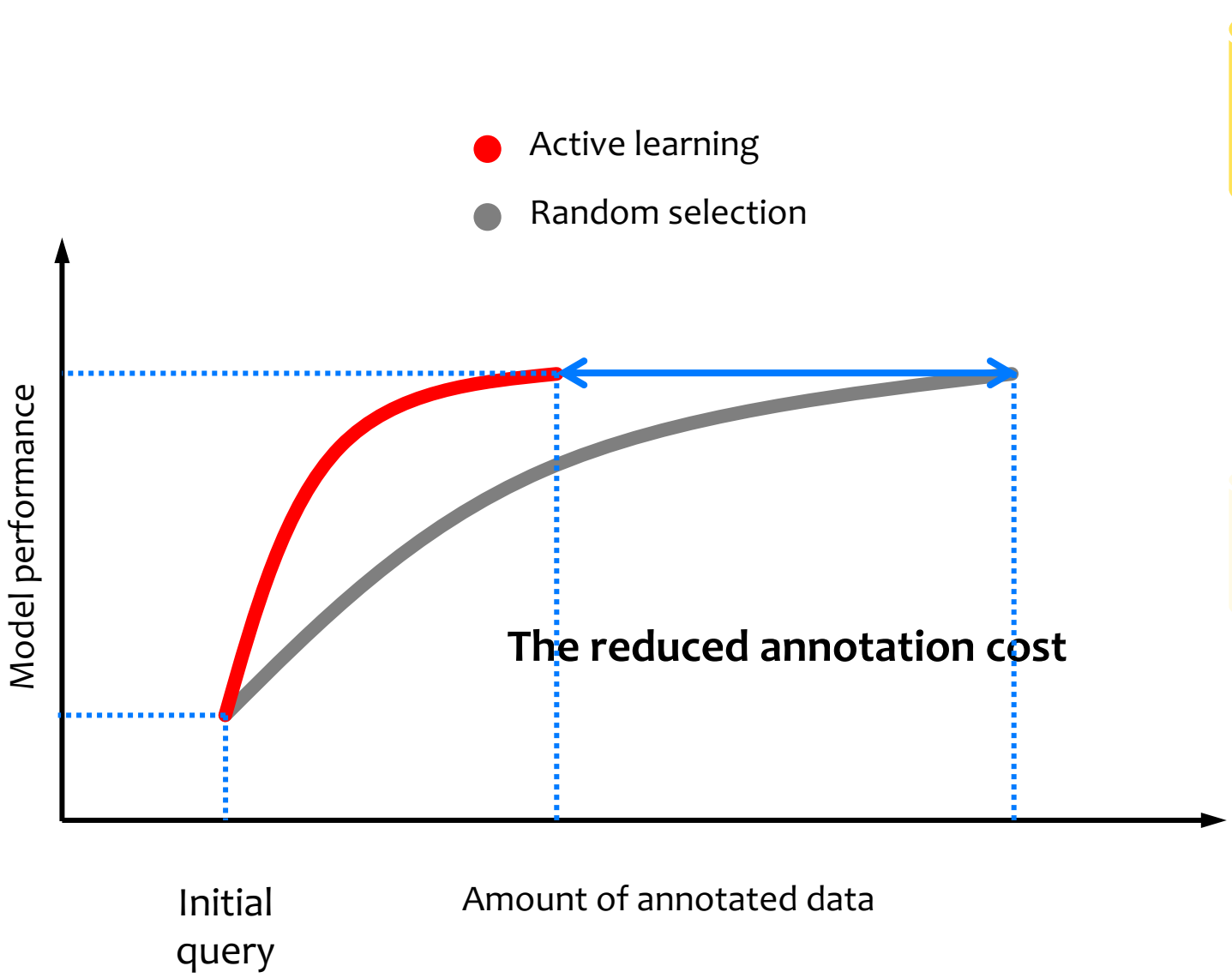
Looking for a PhD program for 2024 Fall entry.

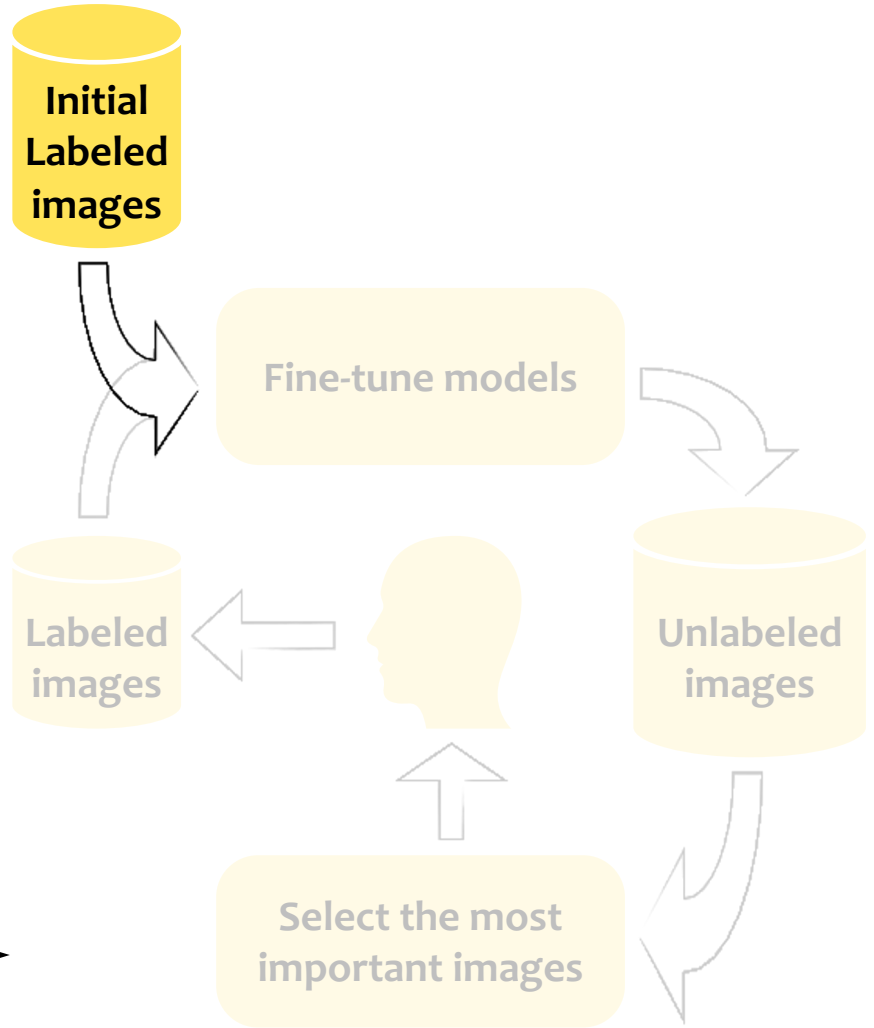
Active Learning: What, Why, How?





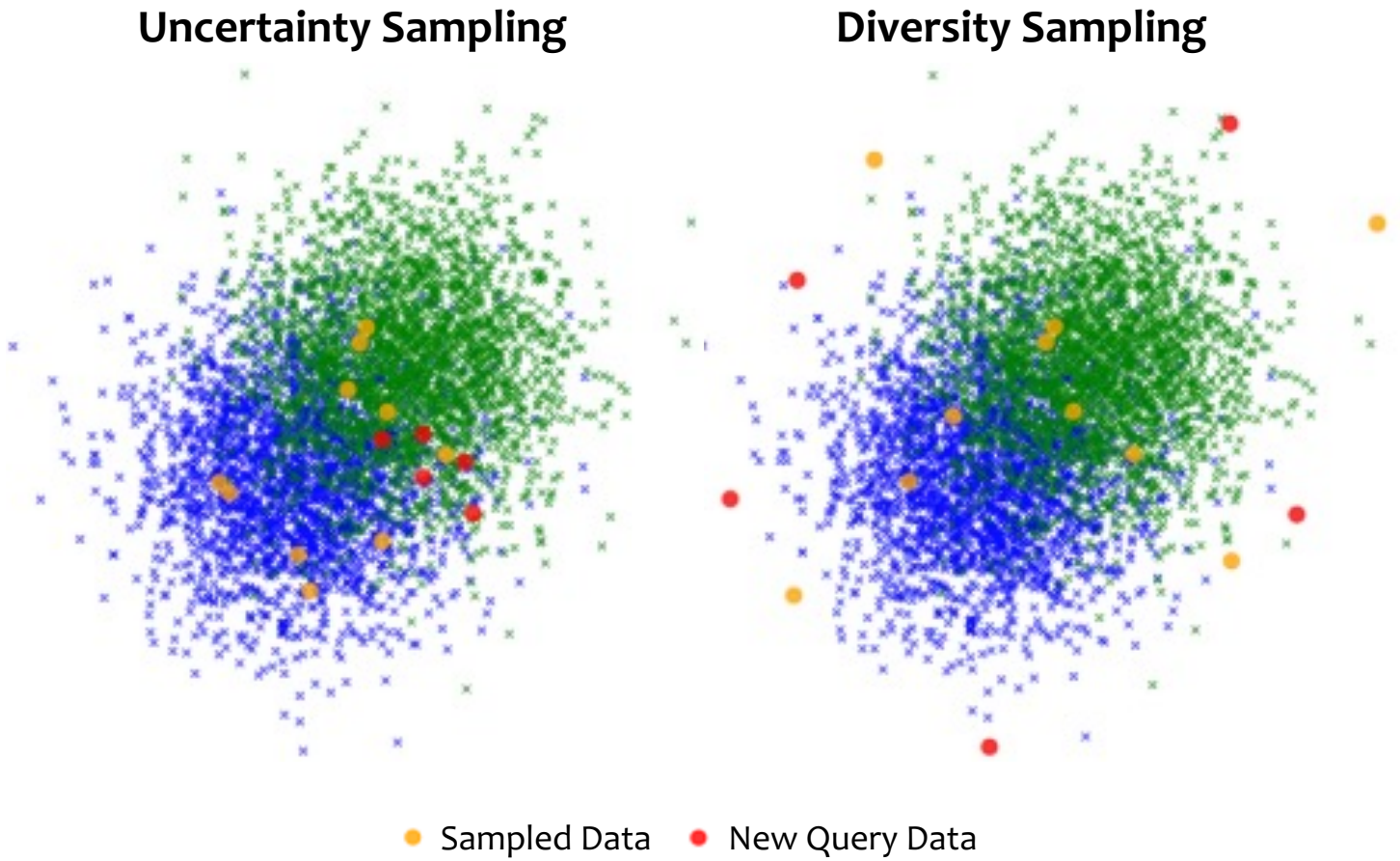
A Good Start is Half the Battle, But …





Part I: The Cold Start Problem in Active Learning

How About Active Sampling the Initial Query?

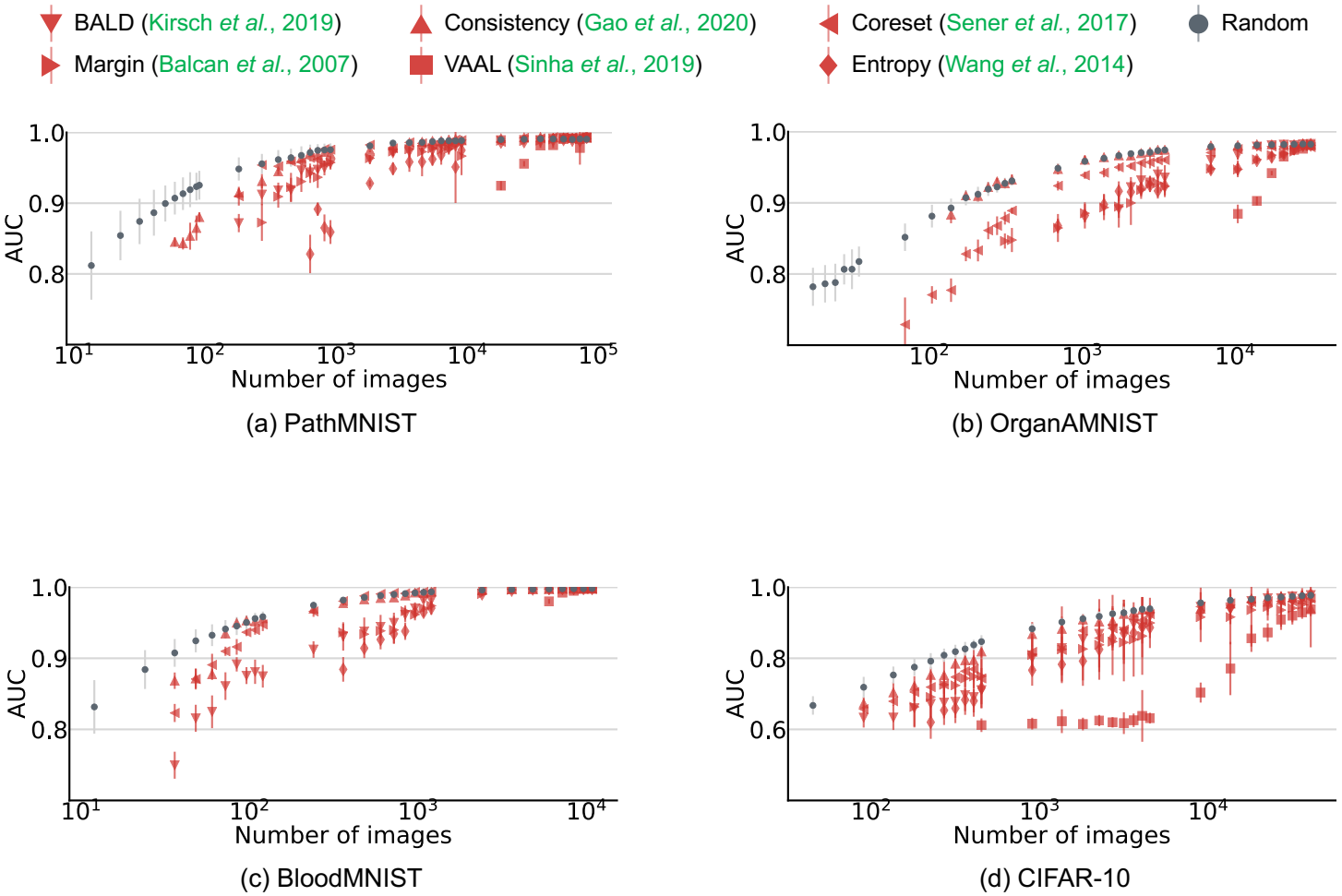


Uncertainty:
ISAL (ICCV, 2021)
Consistency (ECCV, 2020)
Learning loss (CVPR, 2020)
Margin (COLT, 2007)
Entropy (Machine Learning, 1994)
…

Diversity:
CDAL (ECCV, 2020)
Coreset (ICLR, 2018)
Pre-clustering (ICML, 2004)
…

Hybrid:
BADGE (ICLR, 2020)
VAAL (ICCV, 2019)
BALD (NeurIPS, 2019)
…

Current Active Samplings Underperform Random Sampling (I)



Uncertainty:
ISAL (ICCV, 2021)
Consistency (ECCV, 2020)
Learning loss (CVPR, 2020)
Margin (COLT, 2007)
Entropy (Machine Learning, 1994)
…

Diversity:
CDAL (ECCV, 2020)
Coreset (ICLR, 2018)
Pre-clustering (ICML, 2004)
…

Hybrid:
BADGE (ICLR, 2020)
VAAL (ICCV, 2019)
BALD (NeurIPS, 2019)
…

Current Active Samplings Underperform Random Sampling (II)

"Experimental results could not conclusively prove that intelligently sampled initial pools are better for AL than random initial pools in the long run."

Chandra et al. 2020, On Initial Pools for Deep Active Learning

1. Chandra, Akshay L et al. "On Initial Pools for Deep Active Learning." ArXiv abs/2011.14696 (2020): n. Pag.

Current Active Samplings Underperform Random Sampling (II)

"Experimental results could not conclusively prove that intelligently sampled initial pools are better for AL than random initial pools in the long run."

Chandra et al. 2020, On Initial Pools for Deep Active Learning

"We show that although the state-of-the-art active learning methods work well given a large budget of data labeling, a simple K-means clustering algorithm can outperform them on low budgets."

Pourahmad et al. 2020, A Simple Baseline for Low-Budget Active Learning

1. Chandra, Akshay L et al. "On Initial Pools for Deep Active Learning." ArXiv abs/2011.14696 (2020): n. Pag.
2. Pourahmadi, Kossar et al. "A Simple Baseline for Low-Budget Active Learning." ArXiv abs/2110.12033 (2021): n. Pag.

Current Active Samplings Underperform Random Sampling (II)

"Experimental results could not conclusively prove that intelligently sampled initial pools are better for AL than random initial pools in the long run."

Chandra et al. 2020, On Initial Pools for Deep Active Learning

"We show that although the state-of-the-art active learning methods work well given a large budget of data labeling, a simple K-means clustering algorithm can outperform them on low budgets."

Pourahmad et al. 2020, A Simple Baseline for Low-Budget Active Learning

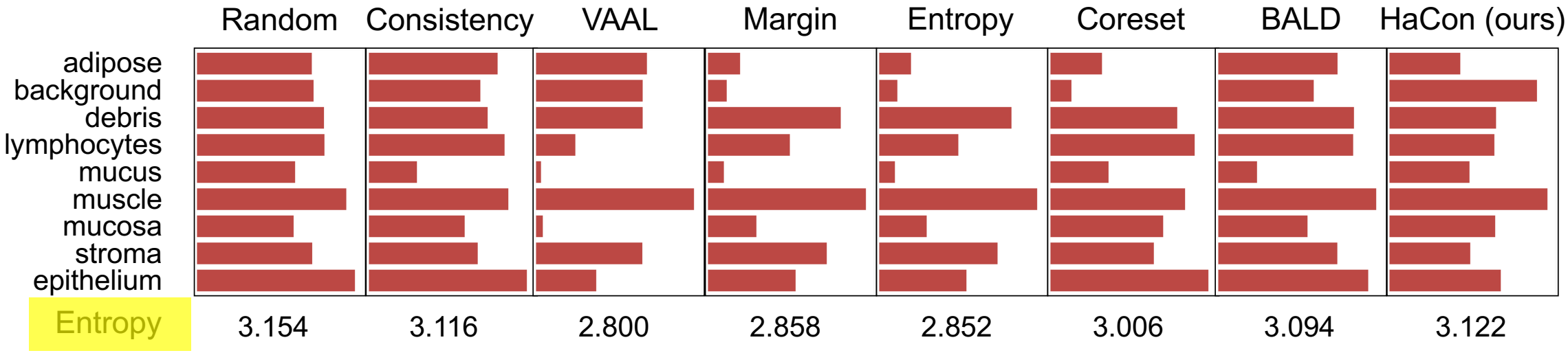
"One key observation is that the benefit or the drawback of using another method than random sampling to construct the initial set disappears rapidly within the next few AL cycles."

Lang et al. 2021, Best Practices in Pool-based Active Learning for Image Classification

1. Chandra, Akshay L et al. "On Initial Pools for Deep Active Learning." ArXiv abs/2011.14696 (2020): n. Pag.
2. Pourahmadi, Kossar et al. "A Simple Baseline for Low-Budget Active Learning." ArXiv abs/2110.12033 (2021): n. Pag.
3. Lang et al 2021. "Best Practices in Pool-based Active Learning for Image Classification." ICLR 2022 submission.

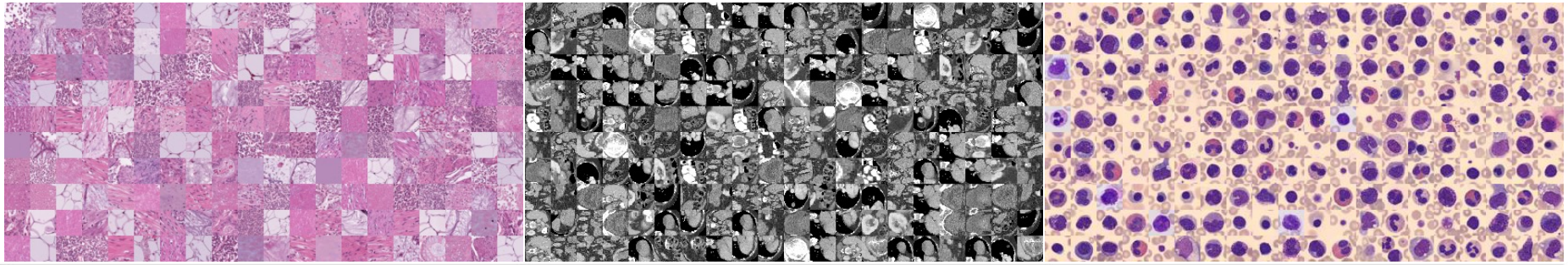
Part II: Causes of Cold Start Problem & Our Solution

Inter-class Factor: Enforcing Label Diversity to Reduce Biased Query



Higher Entropy = More Balanced Query

Our Benchmark: Uniform vs. Non-uniform



	Unif.	PathMNIST 0.5% (499)	PathMNIST 1% (899)	OrganAMNIST 0.5% (172)	OrganAMNIST 1% (345)	BloodMNIST 0.5% (59)	BloodMNIST 1% (119)
Random	✓	96.8±0.6	97.6±0.6	91.1±0.9	93.3±0.4	94.7±0.7	96.5±0.4
	✗	96.4±1.3	97.6±0.9	90.7±1.1	93.1±0.7	93.2±1.5	95.8±0.7
Consistency	✓	96.4±0.1	97.9±0.1	92.3±0.5	92.8±1.0	92.9±0.9	95.9±0.5
	✗	96.2±0.0	97.6±0.0	91.0±0.3	94.0±0.6	87.9±0.2	95.5±0.5
Margin	✓	97.9±0.2	96.0±0.4	81.8±1.2	85.8±1.4	89.7±1.9	94.7±0.7
	✗	91.0±2.3	96.0±0.3	-	85.9±0.7	-	-
Entropy	✓	93.2±1.6	95.2±0.2	79.1±2.3	86.7±0.8	85.9±0.5	91.8±1.0
	✗	-	87.5±0.1	-	-	-	-
Coreset	✓	95.0±2.2	94.8±2.5	85.6±0.4	89.9±0.5	88.5±0.6	94.1±1.1
	✗	95.6±0.7	97.5±0.2	83.8±0.6	88.5±0.4	87.3±1.6	94.0±1.2
BALD	✓	95.8±0.2	97.0±0.1	87.2±0.3	89.2±0.3	89.9±0.8	92.7±0.7
	✗	92.0±2.3	95.3±1.0	-	-	83.3±2.2	93.5±1.3

Inter-class Factor: Enforcing Label Diversity to Reduce Biased Query

Label uniformity: How uniform the sampled labels are in the query

Inter-class Factor: Enforcing Label Diversity to Reduce Biased Query

Label uniformity: How uniform the sampled labels are in the query

We hypothesize *label uniformity* is a good indicator.

"K-means strategies outperform non-uniform methods and are on par with uniform."

Pourahmad et al. 2020, A Simple Baseline for Low-Budget Active Learning

1. Pourahmadi, Kossar et al. "A Simple Baseline for Low-Budget Active Learning." ArXiv abs/2110.12033 (2021): n. Pag.

Inter-class Factor: Enforcing Label Diversity to Reduce Biased Query

Label uniformity: How uniform the sampled labels are in the query

We hypothesize *label uniformity* is a good indicator.

"K-means strategies outperform non-uniform methods and are on par with uniform."

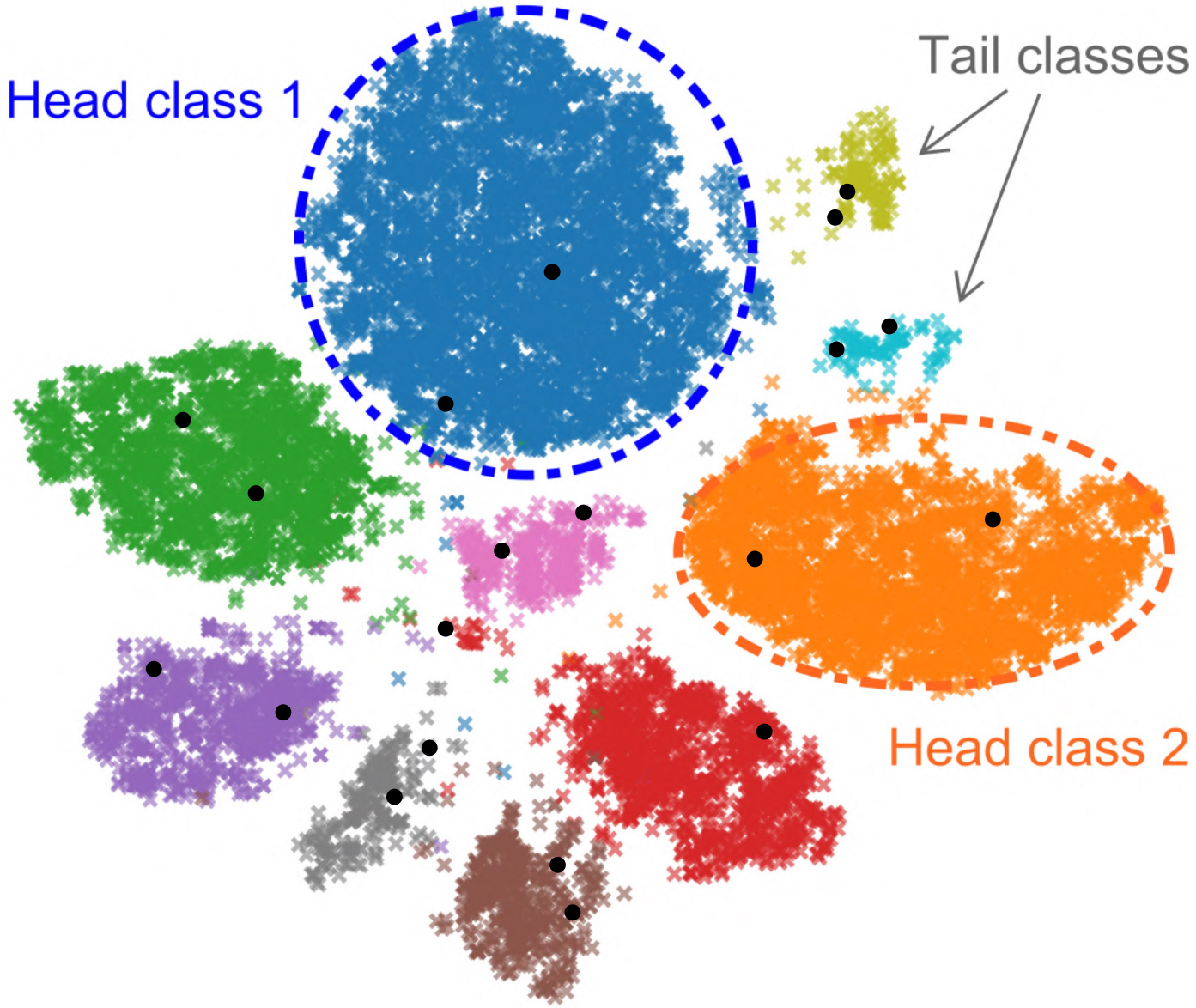
Pourahmad et al. 2020, A Simple Baseline for Low-Budget Active Learning

Label uniformity is not easy to achieve because

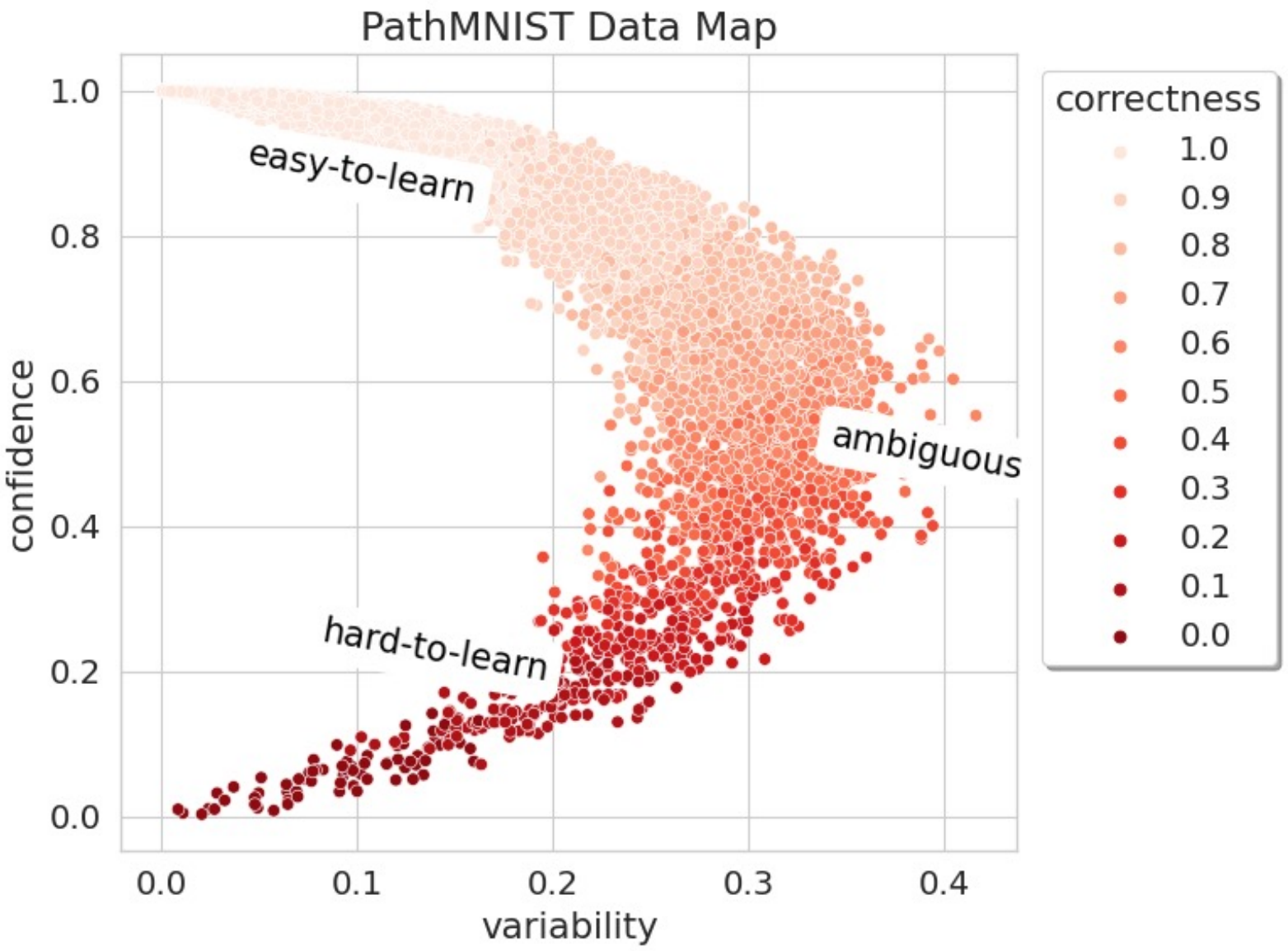
- Active learning tends to bias towards certain classes
- Random sampling ≠ uniform sampling in unbalanced datasets

1. Pourahmadi, Kossar et al. "A Simple Baseline for Low-Budget Active Learning." ArXiv abs/2110.12033 (2021): n. Pag.

Following Uniform Distribution, Which Data to Sample?



A Primer on Data Maps



Confidence

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}\left(y_i^* | x_i\right)$$

Mean of GT prediction of the model over # of epochs (E).

Variability

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\theta^{(e)}}\left(y_i^* | x_i\right) - \hat{\mu}_i\right)^2}{E}}$$

Standard deviation of the confidence over E.

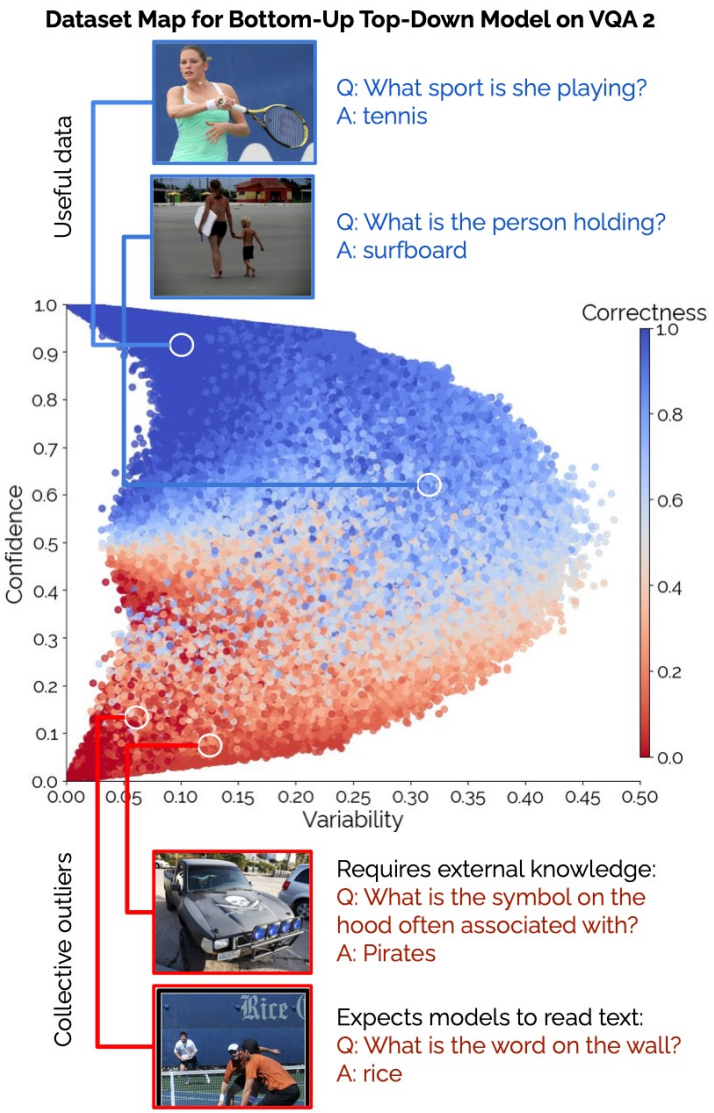
Correctness

$$\hat{\phi}_i = \frac{1}{E}\sum_{e=1}^{E} 1\left(\hat{y}_i = y_i^* | x_i\right)$$

Probability of correct prediction over E.

y_i^* = prediction on GT class

Data Map for Active Learning (II)

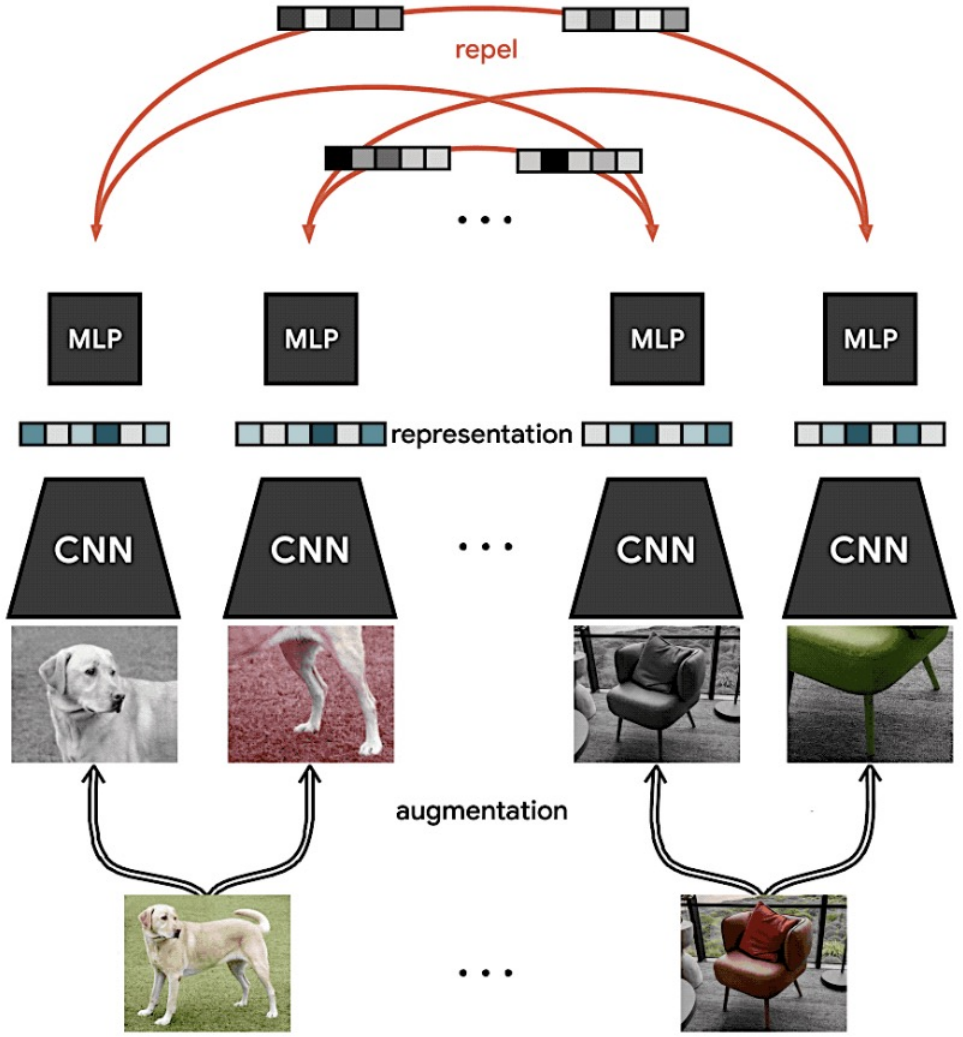


Collective outliers harm active learning performance!

- Contemporary AL methods sample outliers.
- Such outliers are usually difficult to learn.

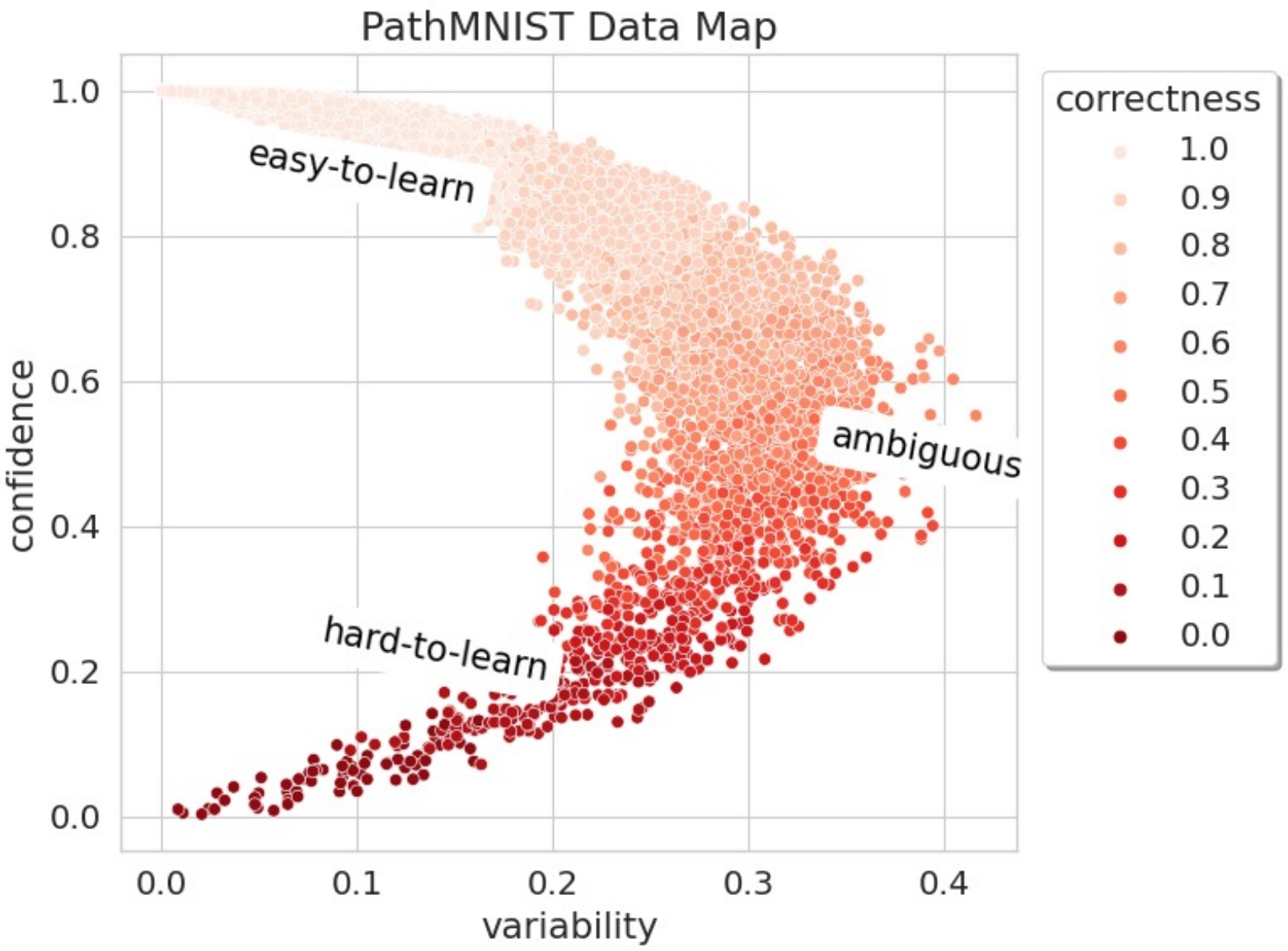
1. Karamcheti, Siddharth et al. "Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering." ArXiv abs/2107.02331 (2021): n. pag.
2. Swayamdipta, Swabha et al. "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics." EMNLP (2020).

Contrastive Self-supervised Learning



1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G.E. (2020). A Simple Framework for Contrastive Learning of Visual Representations. ArXiv, abs/2002.05709.
2. Chen, X., Fan, H., Girshick, R.B., & He, K. (2020). Improved Baselines with Momentum Contrastive Learning. ArXiv, abs/2003.04297.

Self-supervised Pseudo Labels and Data Map



Confidence

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}\left(\boldsymbol{y_i^*}|x_i\right)$$

Mean of GT prediction of the model over # of epochs (E).

Variability

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\theta^{(e)}}\left(\boldsymbol{y_i^*}|x_i\right) - \hat{\mu}_i\right)^2}{E}}$$

Standard deviation of the confidence over E.

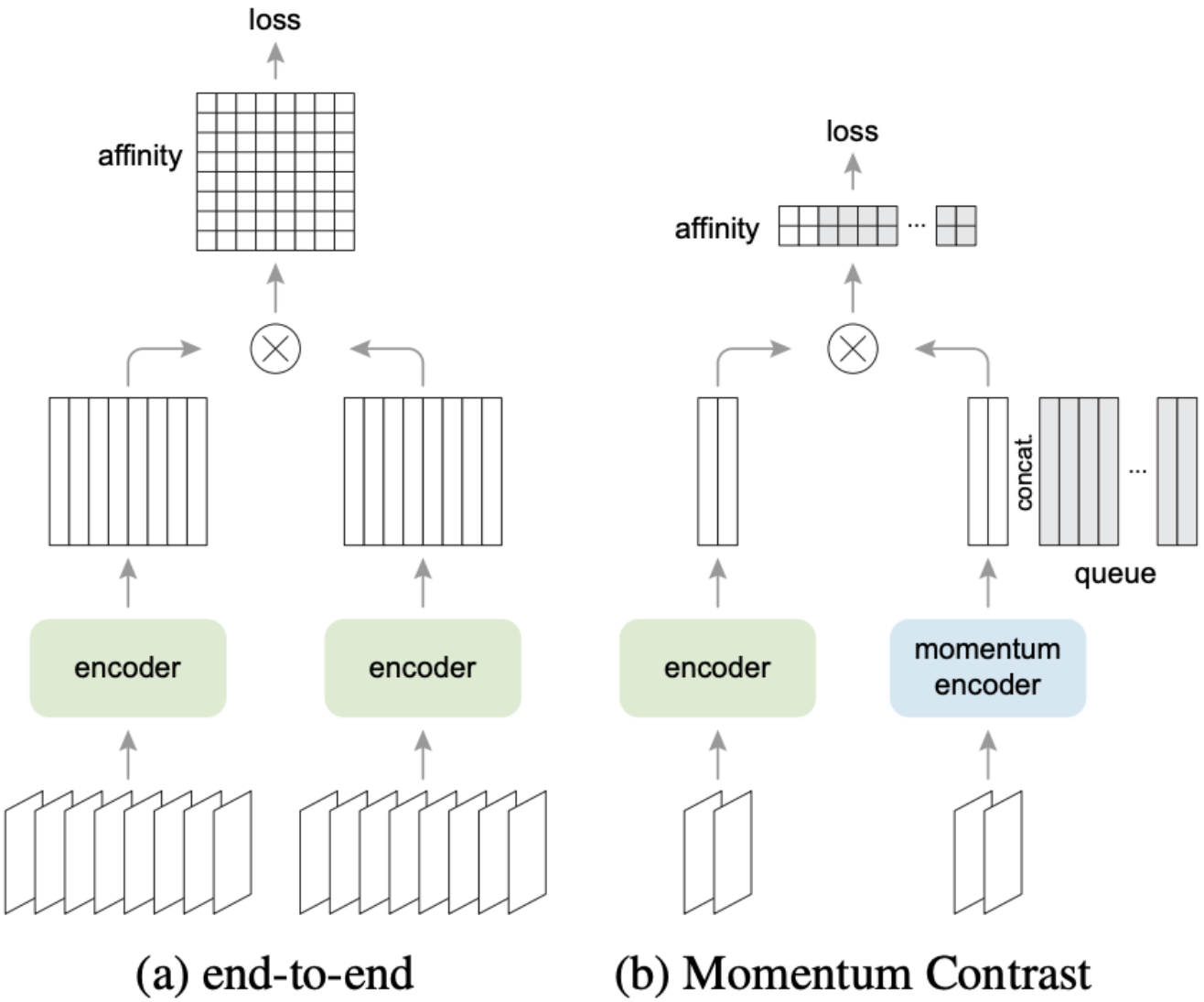
Correctness

$$\hat{\phi}_i = \frac{1}{E}\sum_{e=1}^{E} 1\left(\hat{y}_i = \boldsymbol{y_i^*}|x_i\right)$$

Probability of correct prediction over E.

$\boldsymbol{y_i^*}$ = prediction on pseudo class of self-supervised pretext task

Self-supervised Pseudo Labels and Data Map



Confidence

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}\left(\boldsymbol{y_i^*}|x_i\right)$$

Mean of GT prediction of the model over # of epochs (E).

Variability

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\theta^{(e)}}\left(\boldsymbol{y_i^*}|x_i\right) - \hat{\mu}_i\right)^2}{E}}$$

Standard deviation of the confidence over E.

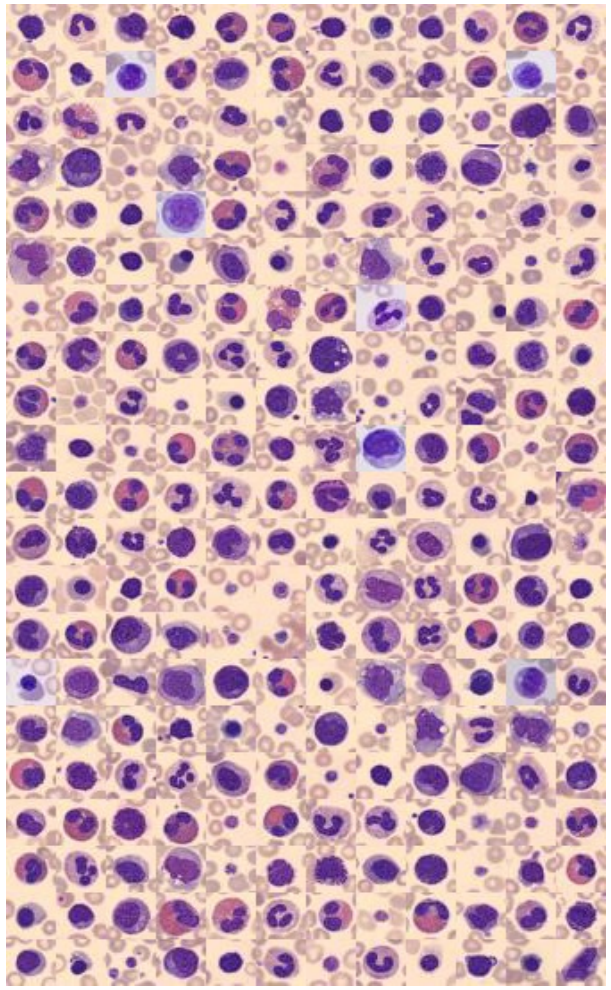
Correctness

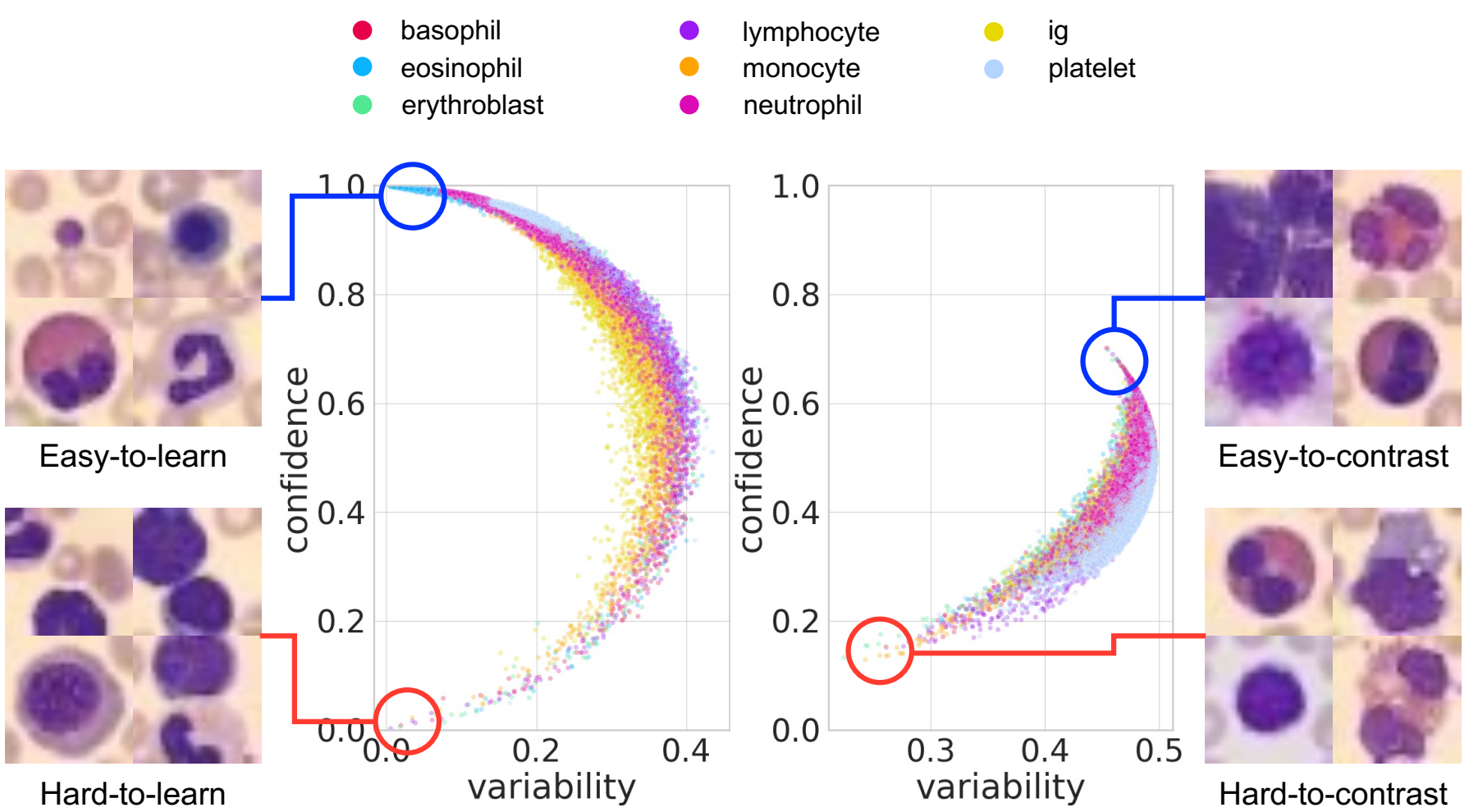
$$\hat{\phi}_i = \frac{1}{E}\sum_{e=1}^{E} 1\left(\hat{y}_i = \boldsymbol{y_i^*}|x_i\right)$$

Probability of correct prediction over E.

$\boldsymbol{y_i^*}$ = prediction on pseudo class of self-supervised pretext task

Intra-class Factor: Selecting Typical Data to Avoid Outlier Query



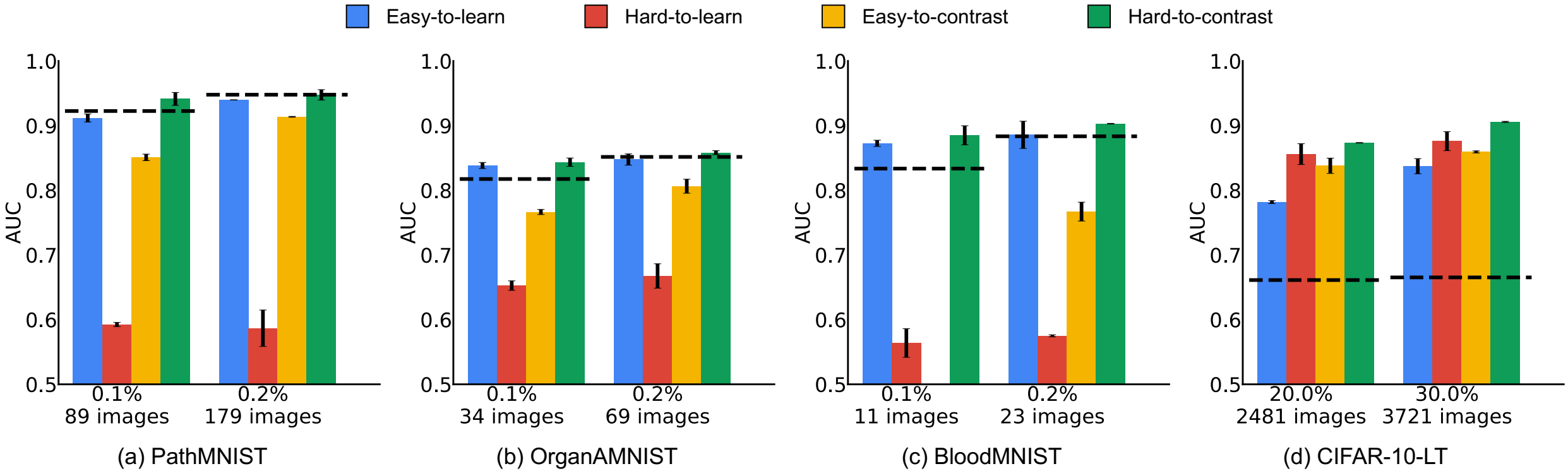


(a) Overall distribution

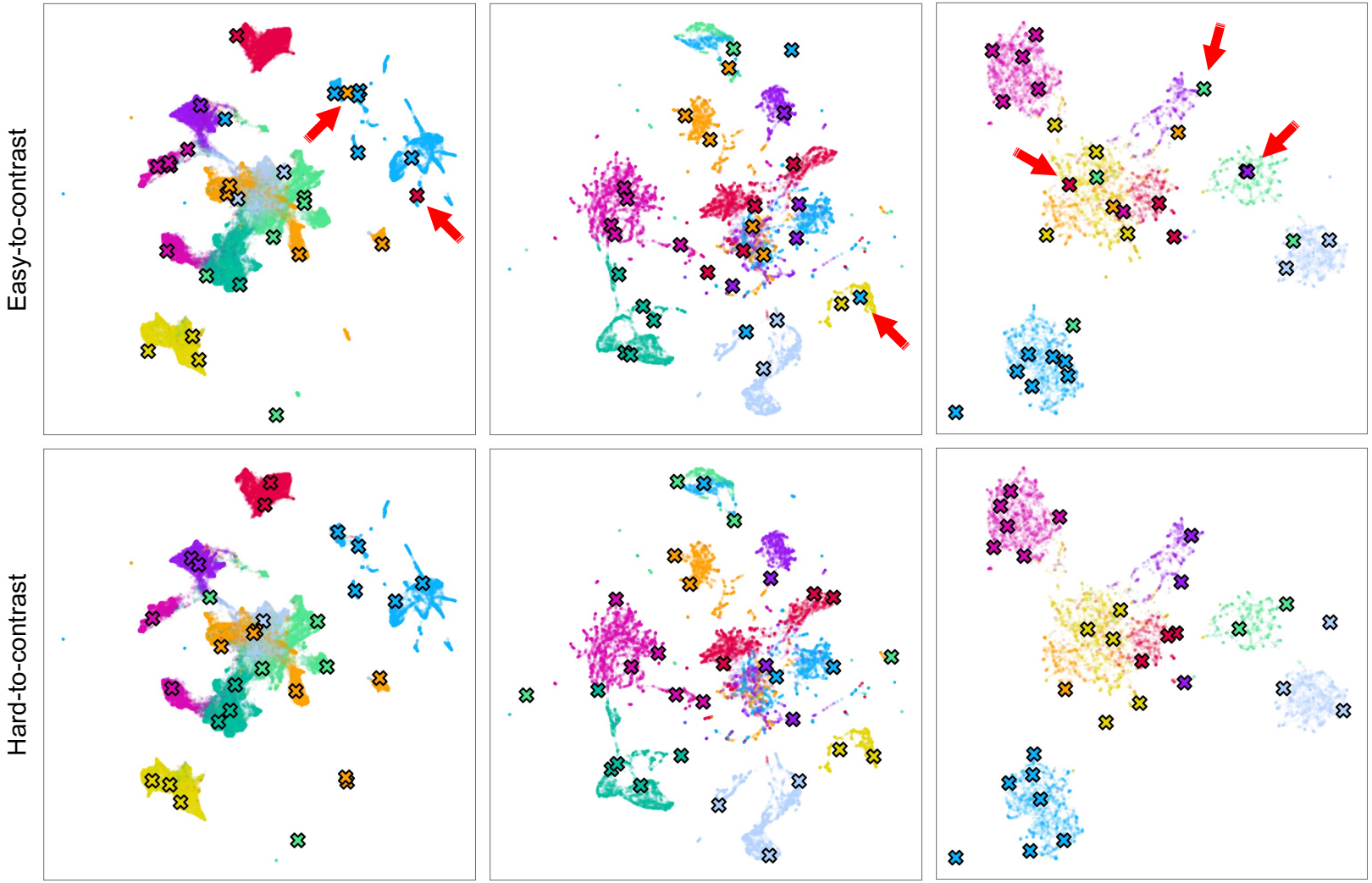
(b) Cartography by ground truth

(c) Cartography by pseudo label

Intra-class Factor: Selecting Typical Data to Avoid Outlier Query



Pseudo-labels Select Typical Data to Avoid Outlier Query



Label Uniformity & Hard-to-Contrast Data

We hypothesize that the level of

(i) Label uniformity [inter-class factor]

(ii) Hard-to-Contrast (typical) data [intra-class factor]

are the underlying indicators of the data importance.

Label Uniformity & Hard-to-Contrast Data

We hypothesize that the level of

(i) Label uniformity [inter-class factor]

(ii) Hard-to-Contrast (typical) data [intra-class factor]

are the underlying indicators of the data importance.

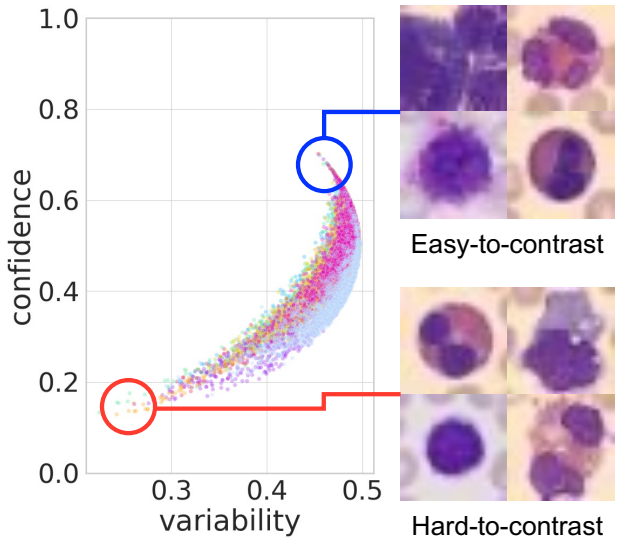
- *Label uniformity*: Higher category coverage is associated with higher performance
- *Typical data*: Typical data are useful for model training

Part III: Towards Effective Initial Query in Active Learning

Our Solution: HaCon

Contrastive pre-training

Data Map by pseudo-labels



Cluster image features

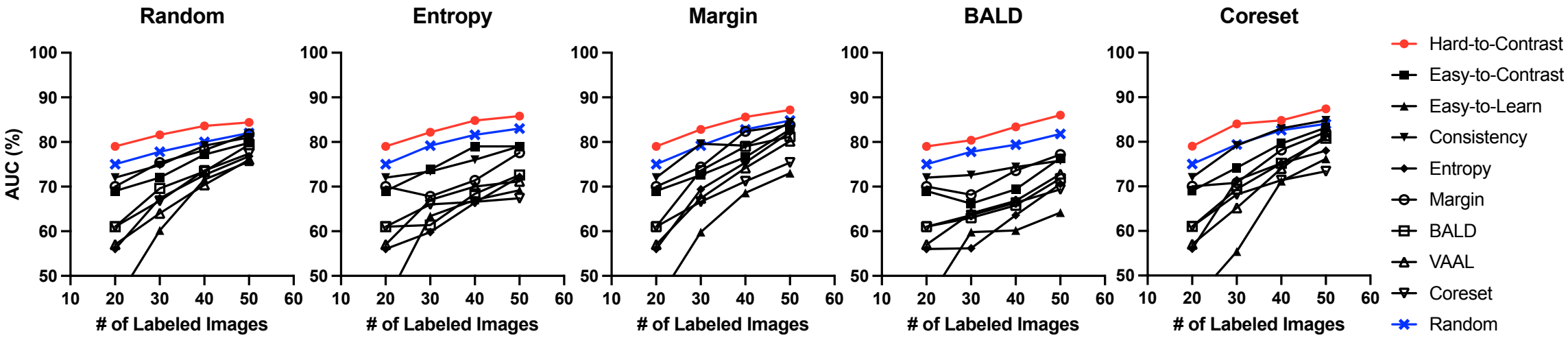
Query the most typical images among each cluster

Inter-class factor: Label Uniformity

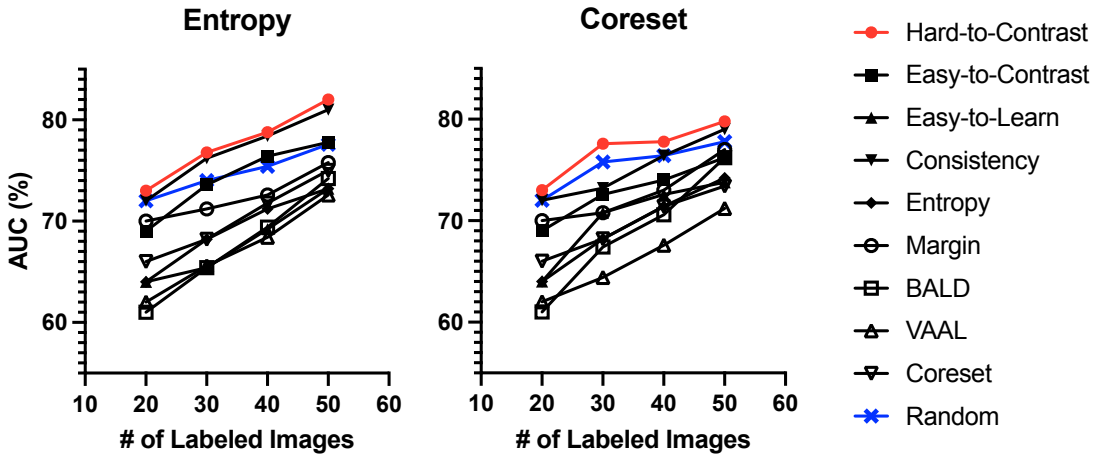
Intra-class factor: Hard-to-Contrast Data

Active Learning Performance

OrganAMNIST, CT scan, image size 28*28



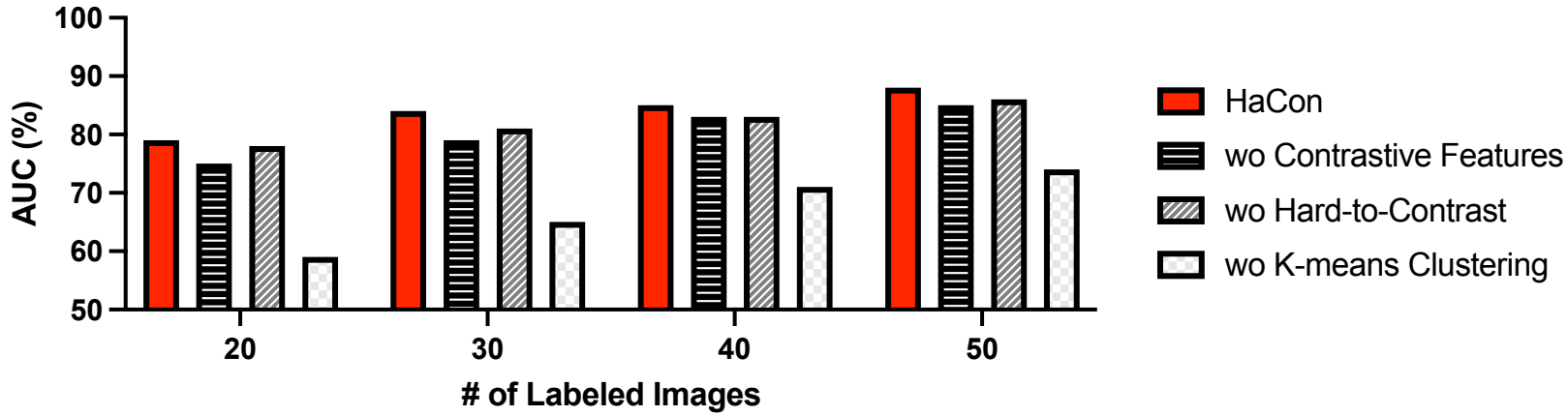
HAM10000, dermatoscopy, image size 512*512



- HaCon > Random selection
- HaCon > Contemporary AL methods

Better initial query leads to better AL!

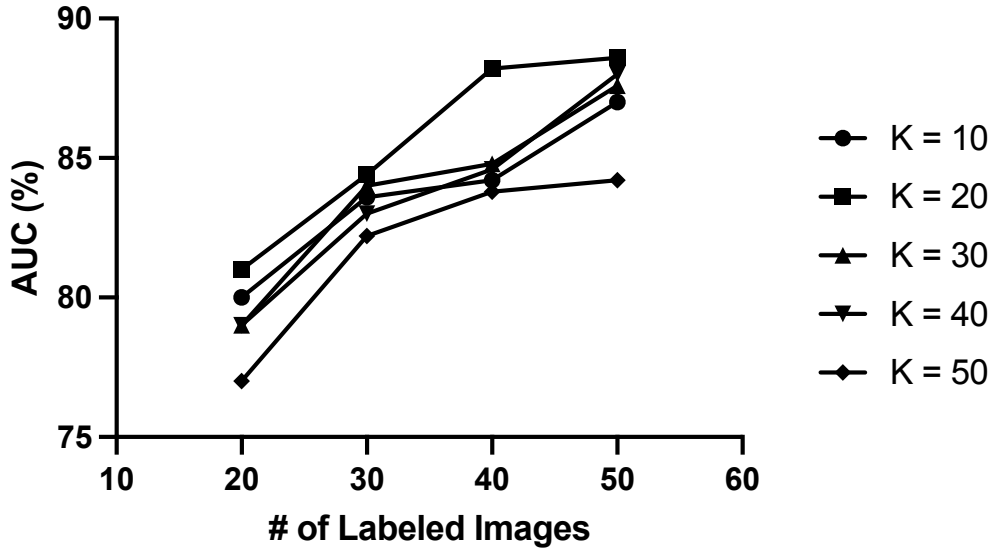
Ablation Study: HaCon Components



- 1) Contrastive learning, 2) K-means clustering, and 3) Hard-to-contrast criterion all contribute to HaCon performance.

- *Diversity matters*: K-means clustering is the most critical component.

Ablation Study: Number of Clusters (*K*)



- *Overclustering*: The number of clusters (*K*) should be larger than the number of classes to ensure coverage of all classes.

- *Optimal K*: size of the initial query

Take-home Messages

1. Illustrate the cold start problem in vision active learning.

Take-home Messages

1. Illustrate the cold start problem in vision active learning.

2. Discover that
   - biased query [inter-class factor]
   - outlier query [intra-class factor]

   are the underlying causes of the cold start problem.

Take-home Messages

1. Illustrate the cold start problem in vision active learning.

2. Discover that
   - biased query [inter-class factor]
   - outlier query [intra-class factor]

   are the underlying causes of the cold start problem.

3. Investigate the role of hard-to-contrast data in the cold start problem.

Take-home Messages

1. Illustrate the cold start problem in vision active learning.

2. Discover that
   - biased query [inter-class factor]
   - outlier query [intra-class factor]

   are the underlying causes of the cold start problem.

3. Investigate the role of hard-to-contrast data in the cold start problem.

4. Extend data map to self-supervised learning.

This work serves as a simple yet strong baseline to sample the initial query for the "human-in-the-loop" active learning procedure.

“The secret of getting ahead is getting started.”

— Mark Twain