Deep Active Learning: Unified and Principled Method for Query and Training

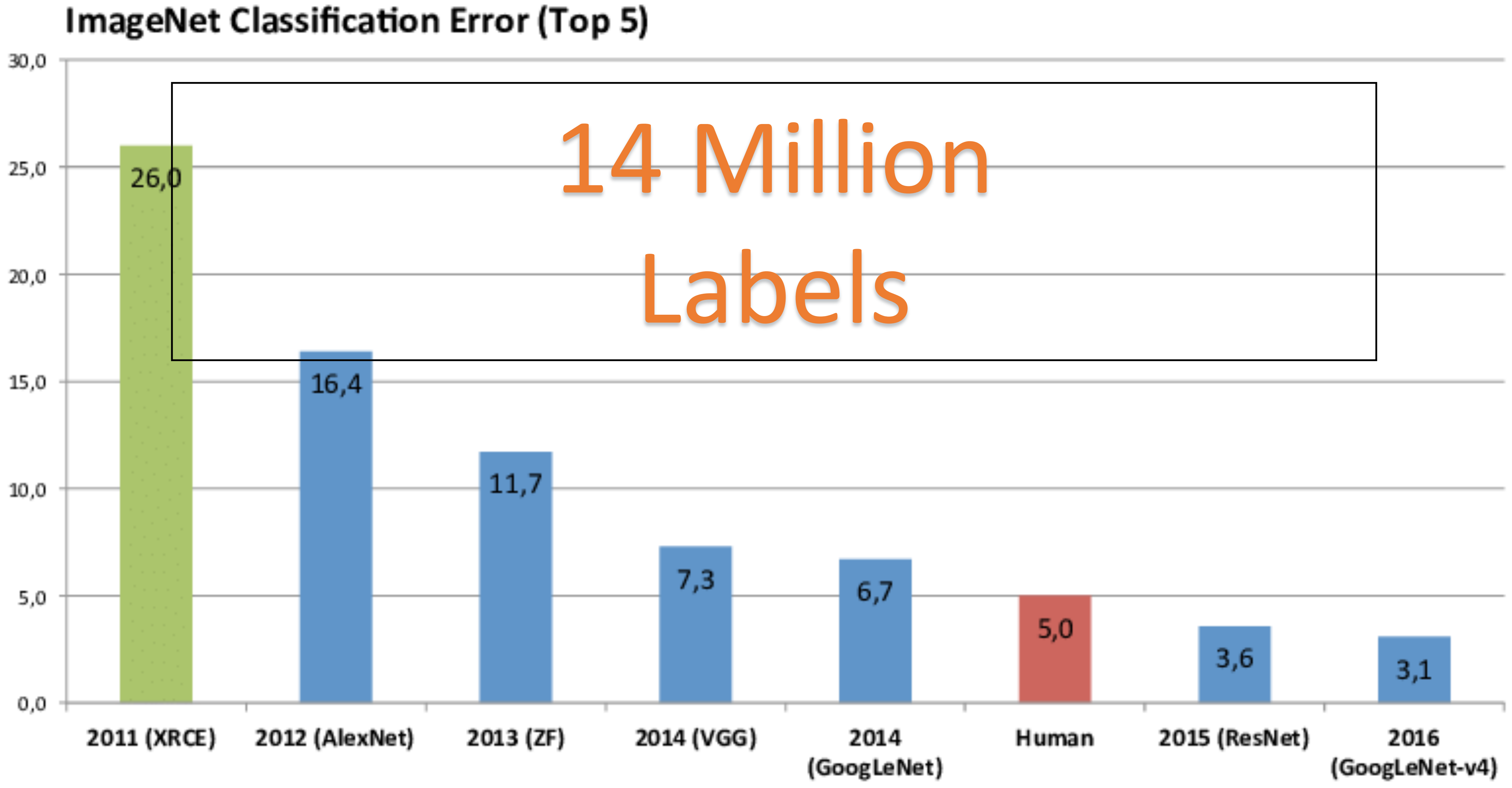
Changjian Shui^[1], Fan Zhou^[1], Christian Gagné^[1,2], Boyu Wang^[3]

1. Université Laval; 2. Mila, Canada CIFAR AI Chair; 3 University of Western Ontario

Deep Learning

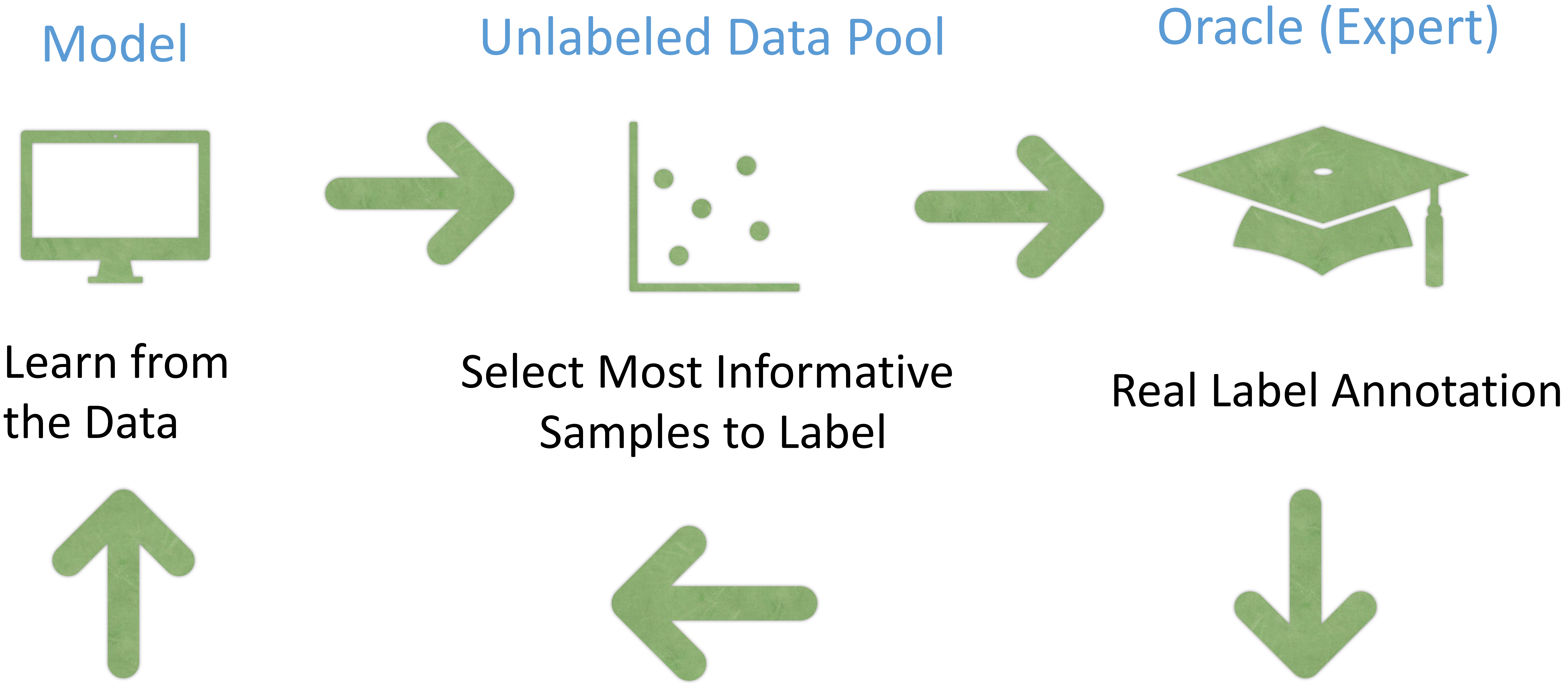
- Surpass human in many tasks
- Drawback: annotated data hungry

One solution: **Deep Active Learning**



Credit: https://devopedia.org/imagenet

Active Learning (AL)



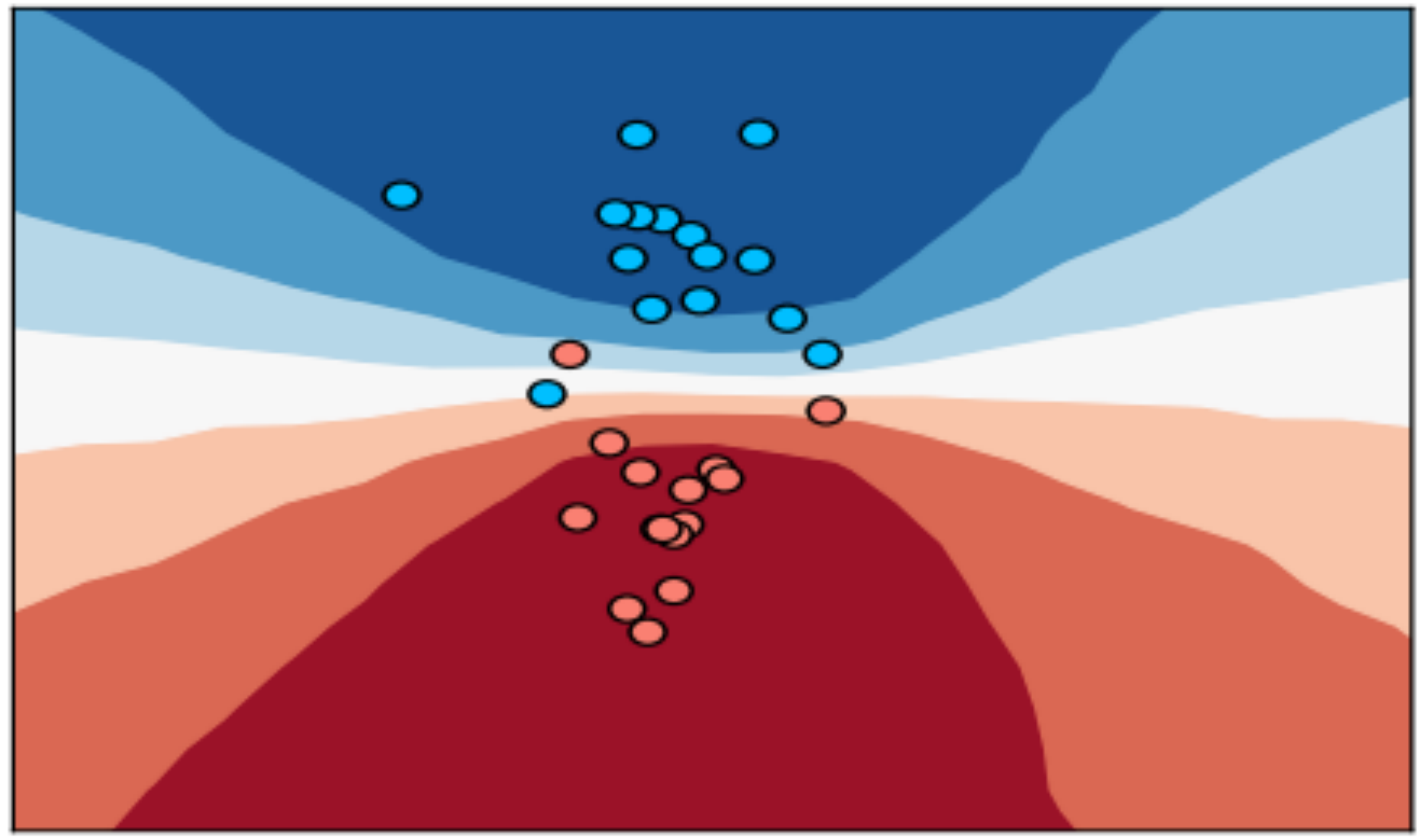
Key factors in Deep AL

- *Query* How to select the most informative samples ?

- *Training* How to train the model ?

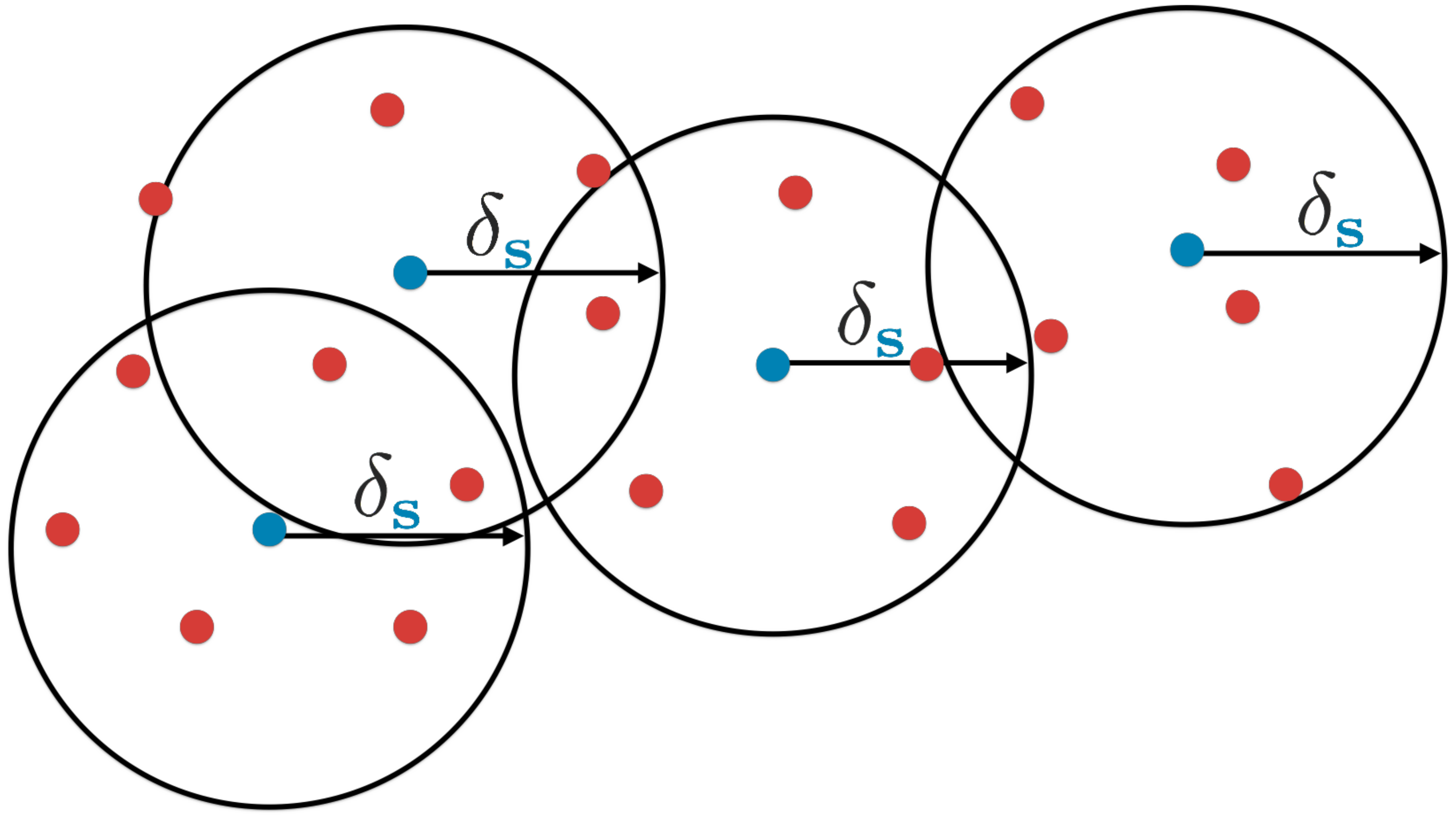
Query Strategy

Uncertainty



- Least Confident Samples
- Sampling Bias

Diversity



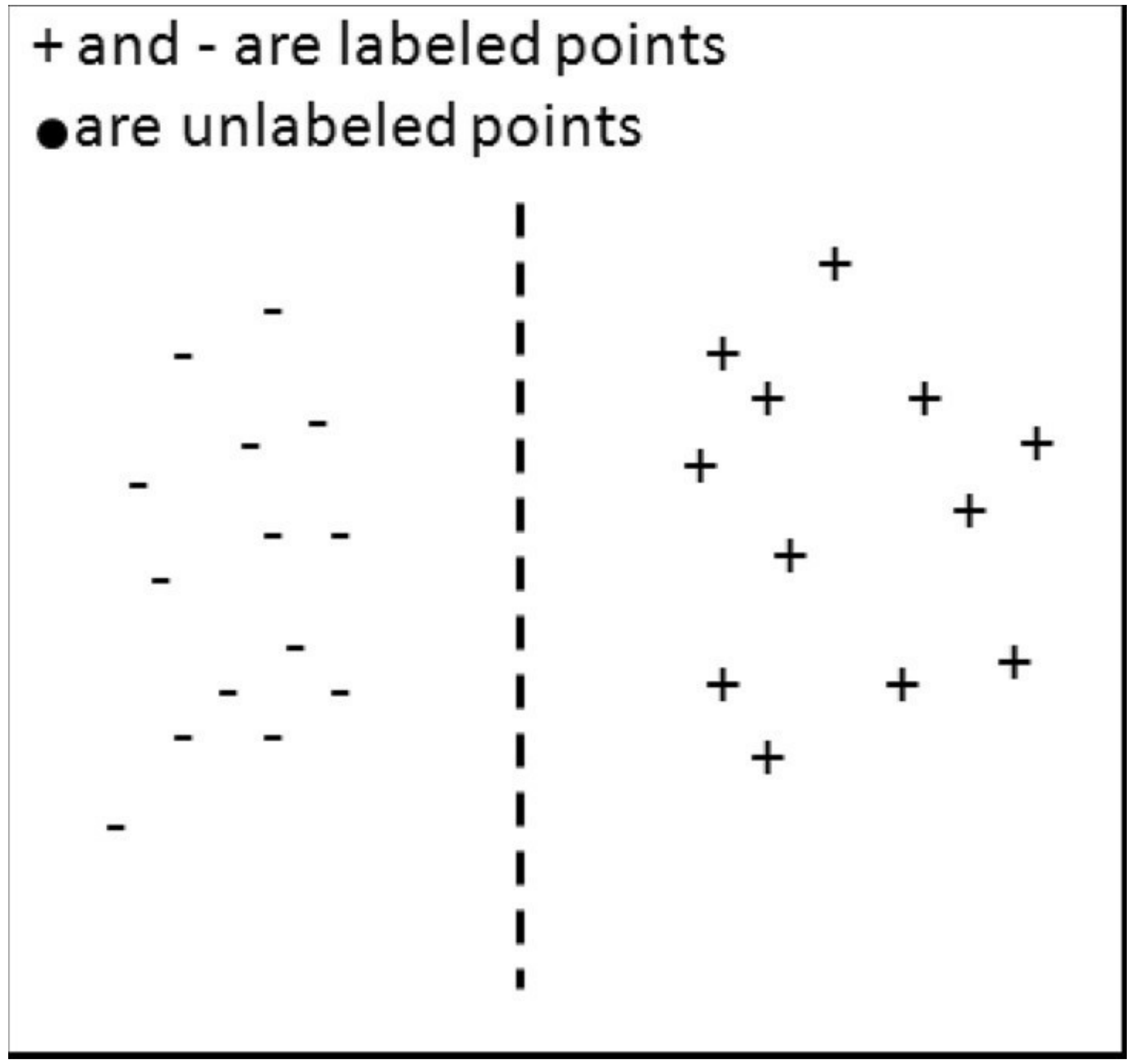
- Representative Samples
- Computationally Inefficient

Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." ICLR, 2018

Tim Pearce, et.al. Uncertainty in Neural Networks: Approximately Bayesian Ensembling Arxiv.1810.05546

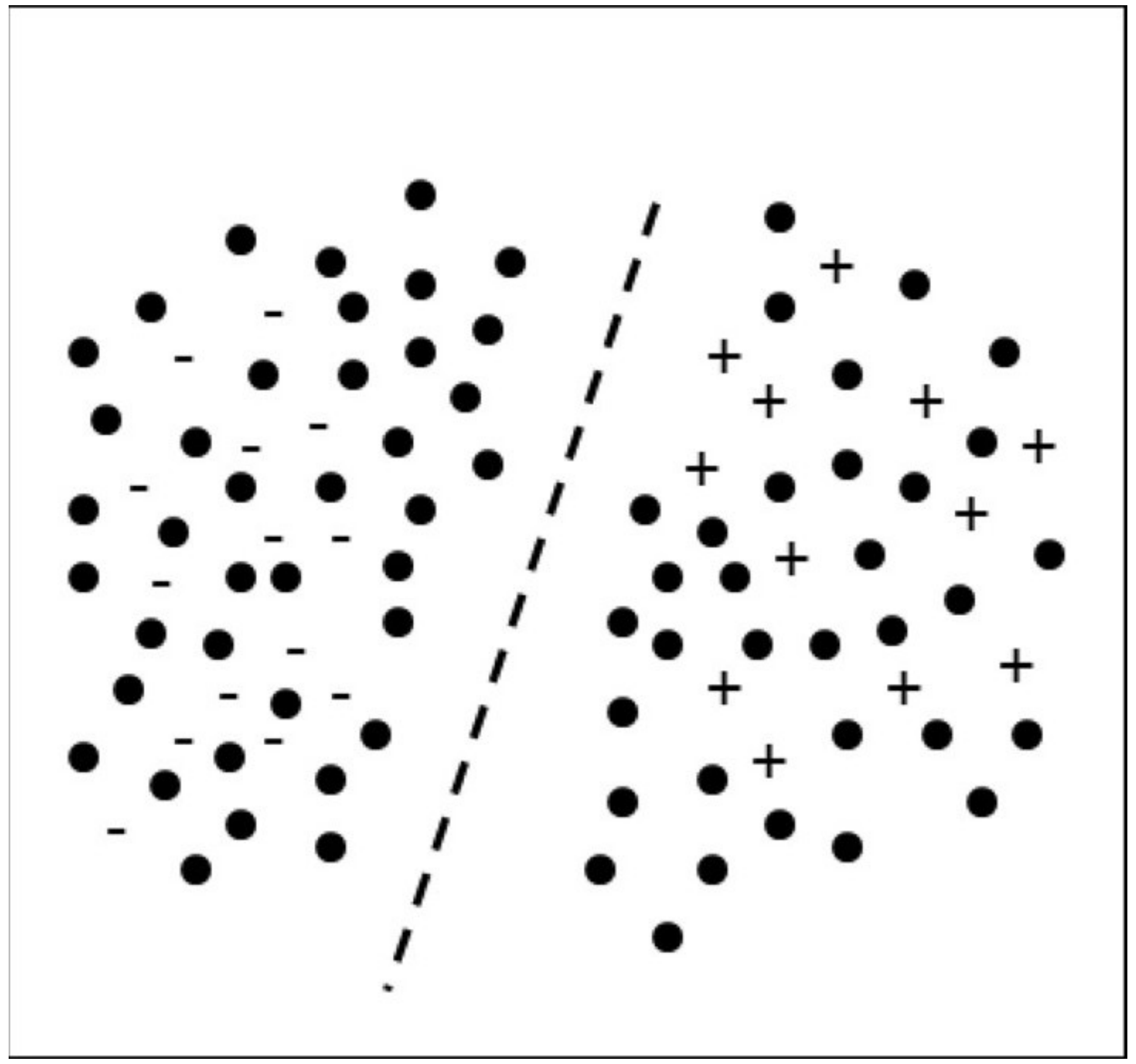
Training Strategy in Deep AL

Labelled



- Simple, Efficient

Labelled + Unlabeled



- Better representation

Mohammad Peikari, et.al. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification, Scientific Reports, 2018

Summary of Our Work

- Query from the unlabeled data:

  Explicit strategy with Uncertain + Diverse criteria.

- Train the model:

  Training Strategy on the labeled + unlabeled dataset.

- Our approach is principled, unified and query efficient.

Active Learning as Distribution Matching

Data distribution D(x), Query distribution Q(x)

$$R_D(h) \leq R_Q(h) + C_0 W_1(D, Q) + C_1$$

Risk of D

Risk of query Q

Distribution Similarity
Wasserstein distance

Wasserstein distance captures the diverse property

Practical Loss in Deep Learning

$$R_Q(h) + \mu W_1(D, Q)$$

$$\min_{\hat{B},h} \hat{R}_{\hat{L}\cup\hat{B}}(h) + \mu \boxed{W_1(\hat{L}\cup\hat{U}, \hat{L}\cup\hat{B})}$$

Estimation via dual term (introducing citric function)

Querying Batch $\hat{B}$

Labelled Data $\hat{L}$

Unlabeled Data $\hat{U}$

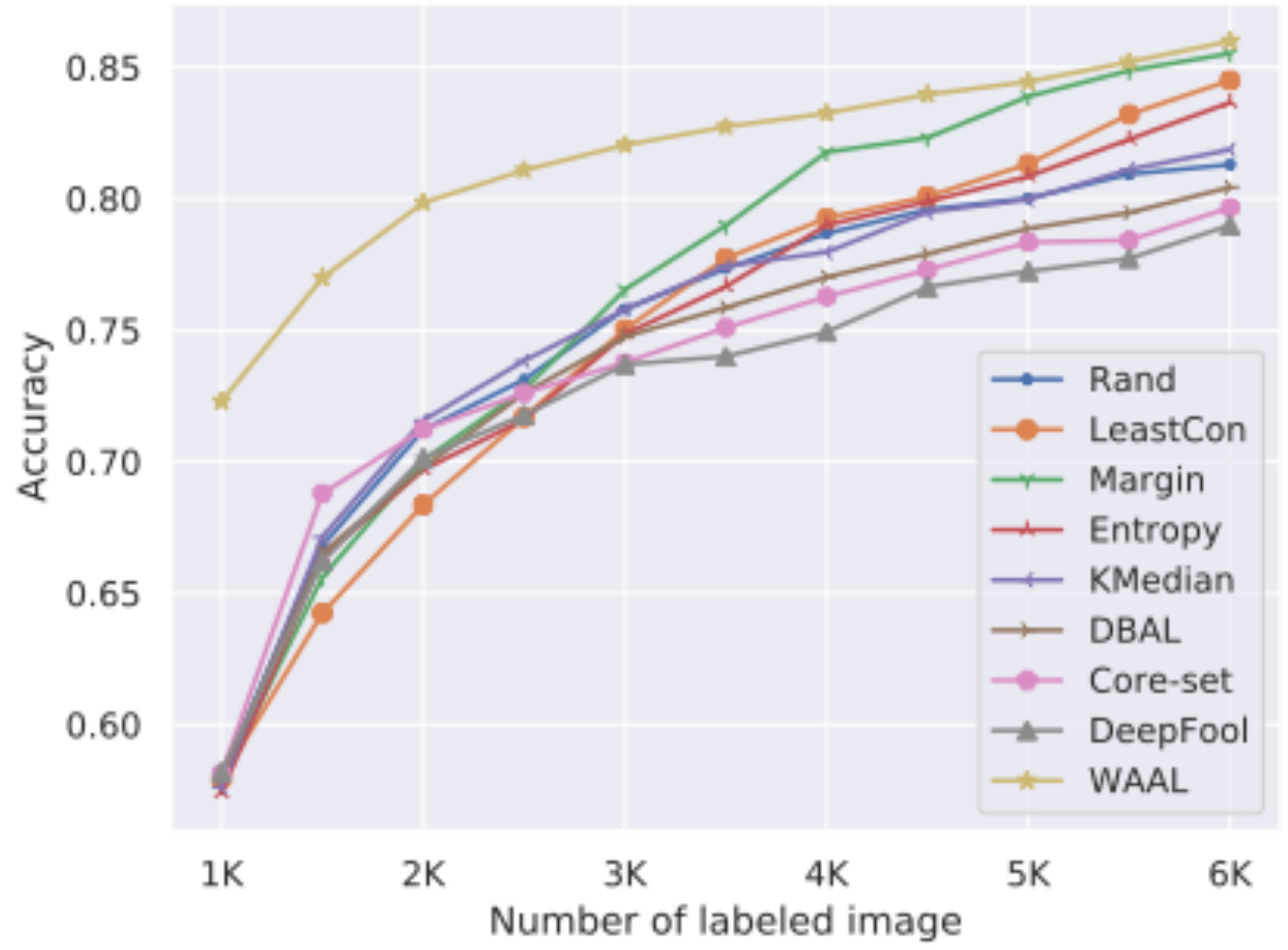
Loss Decomposition in Deep AL

Hypothesis h, feature learning function f, critic function g

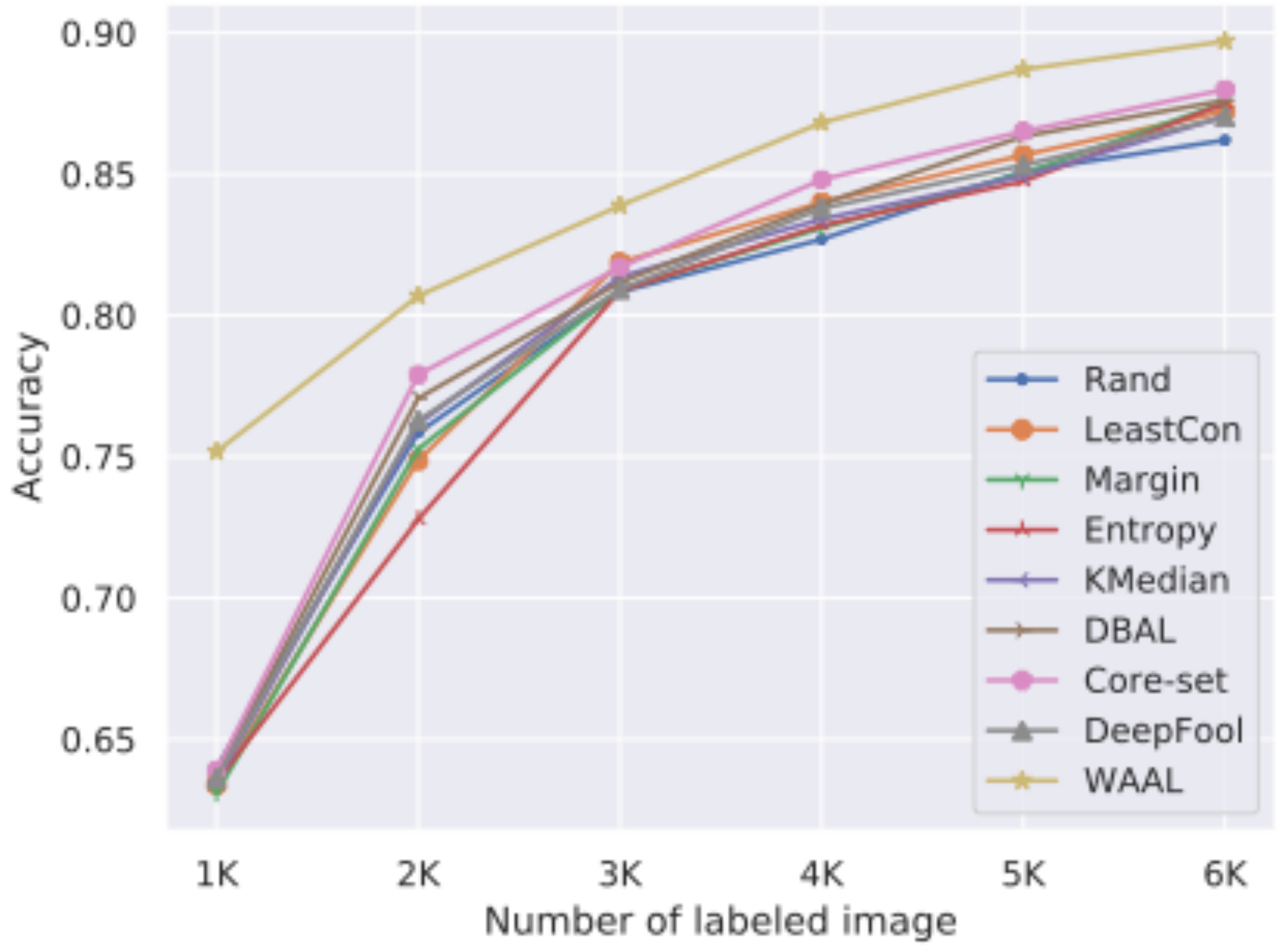
- **Training Stage**
  - Label Prediction Loss. $min_{f,h} \quad \hat{R}_{\hat{L}}(h, f)$
  - Leverage unlabeled data. $min_f \; max_g \; C_0 \sum_{x\sim\hat{U}} g\big(f(x)\big) - C_1 \sum_{x\sim\hat{L}} g\big(f(x)\big)$
- **Query Stage**
  - Query Batch $min_{\hat{B}} \; \underbrace{C_2 \hat{R}_{\hat{B}}(h, f)}_{\text{Upper Bound: Uncertainty}} - \underbrace{C_3 \sum_{x\sim\hat{B}} g\big(f(x)\big)}_{\text{Diversity}}$

Empirical Validations

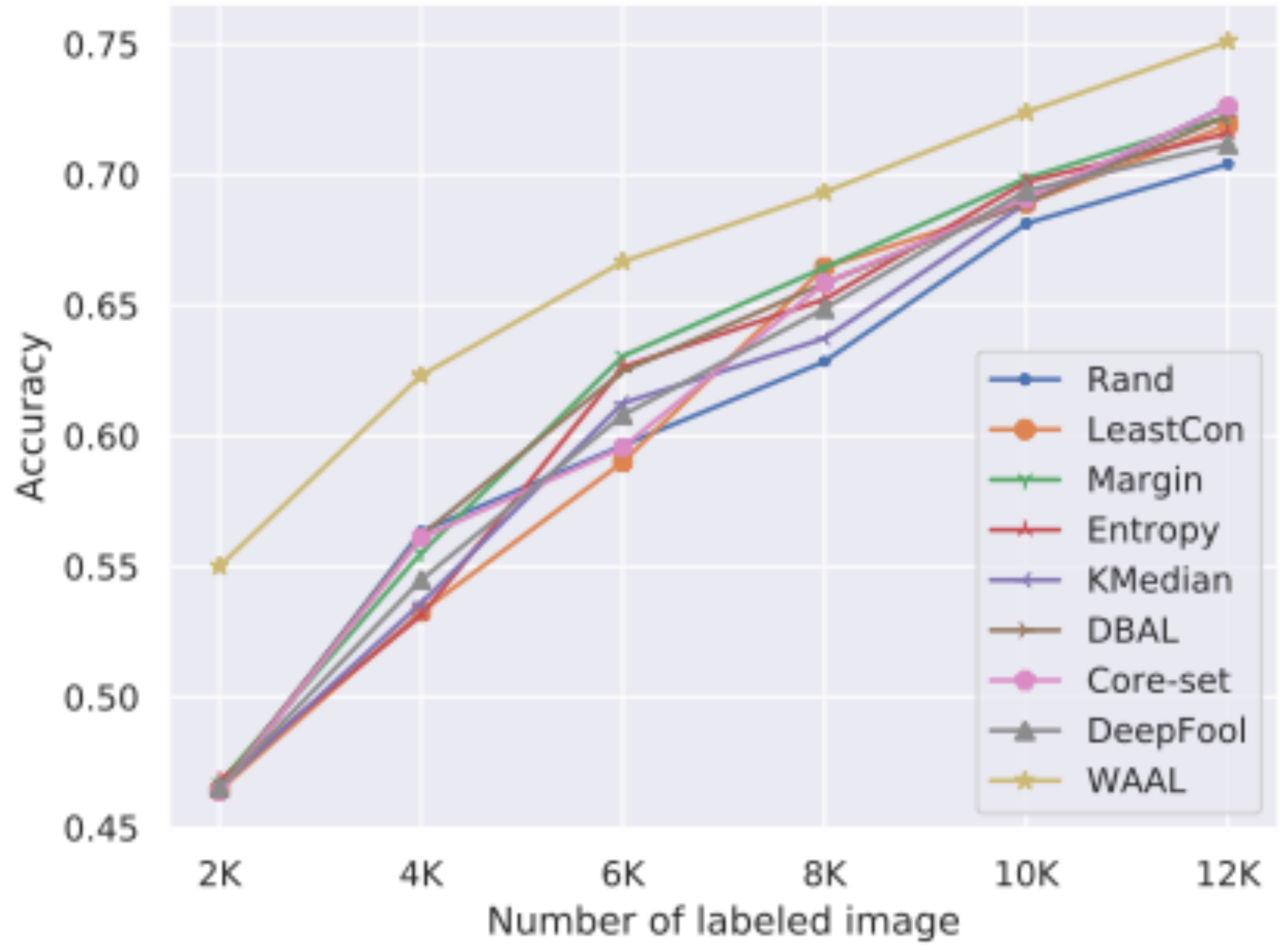
Accuracy



Fashion-MNIST



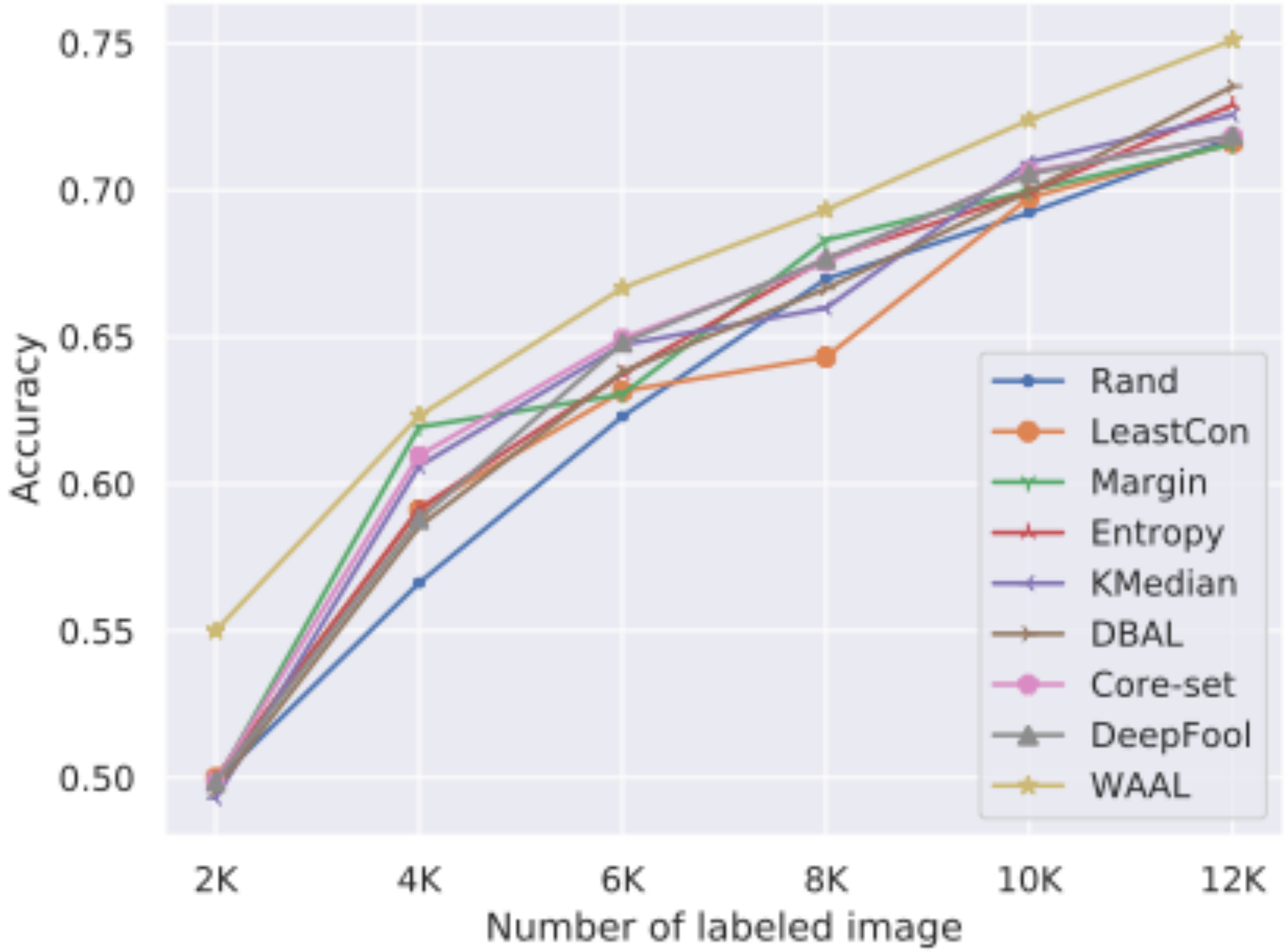
SVHN



CIFAR-10

- WAAL (Wasserstein Adversarial Active Learning)
- Role of unlabeled data: Significant improvement in the initialization

Wasserstein vs. GANs loss



CIFAR-10

- All baselines are leveraged with unlabeled samples via GANs loss.
- Consistently better!

Query time

Method	LeastCon	Margin	Entropy	K-Median	DBAL	Core-set	DeepFool	WAAL
Time	0.94	0.95	0.95	33.98	9.25	45.88	124.46	1

Table 1: Relative Average querying time, assuming the query time of WAAL as the unit.

Conclusion

- WAAL is a simple and efficient approach for deep AL
- Paper: https://arxiv.org/abs/1911.09162
- Code : https://github.com/cjshui/WAAL