CS 6604 DCML
Active Learning

Instructor: Ismini Lourentzou
Assistant Professor
Computer Science, Virginia Tech
https://isminoula.github.io/

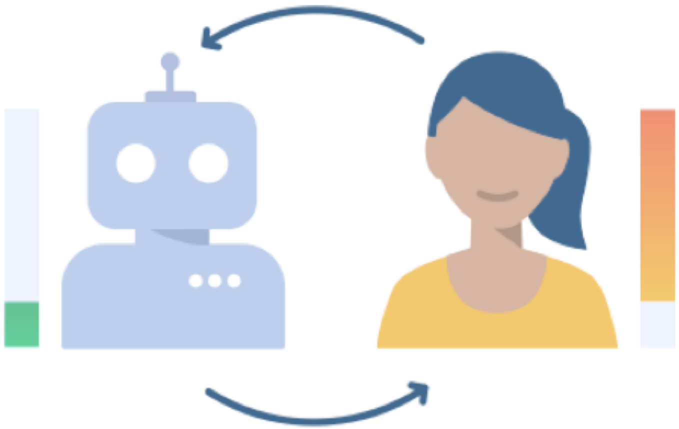
Slides 2-21 adapted from ***Robert Munro's Active Learning for NLP talk***
And Yi Zhang's slides: https://www.cs.cmu.edu/~tom/10701_sp11/recitations/Recitation_13.pdf

Human-in-the-Loop Machine Learning

How can humans and machines **work together** to solve problems?

Applicable to many domains:

- Robotics: learn from demonstrations
- Information extraction: learn from tagged entities
- Search: learn from click-logs
- …



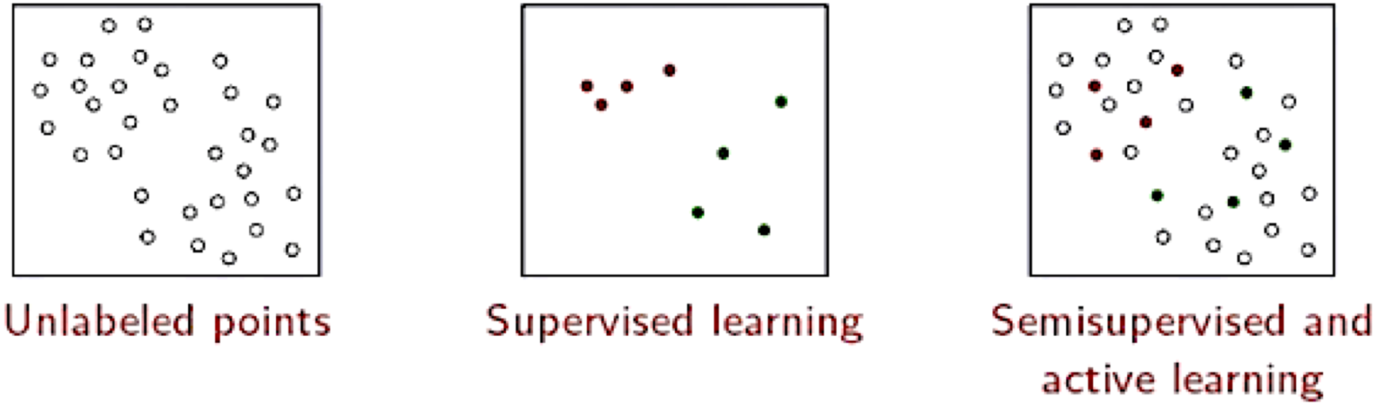
https://www.svatitech.com/images/lyout/hm.png

Exploit unlabeled data

Large amounts of unlabeled data are easily accessible for free

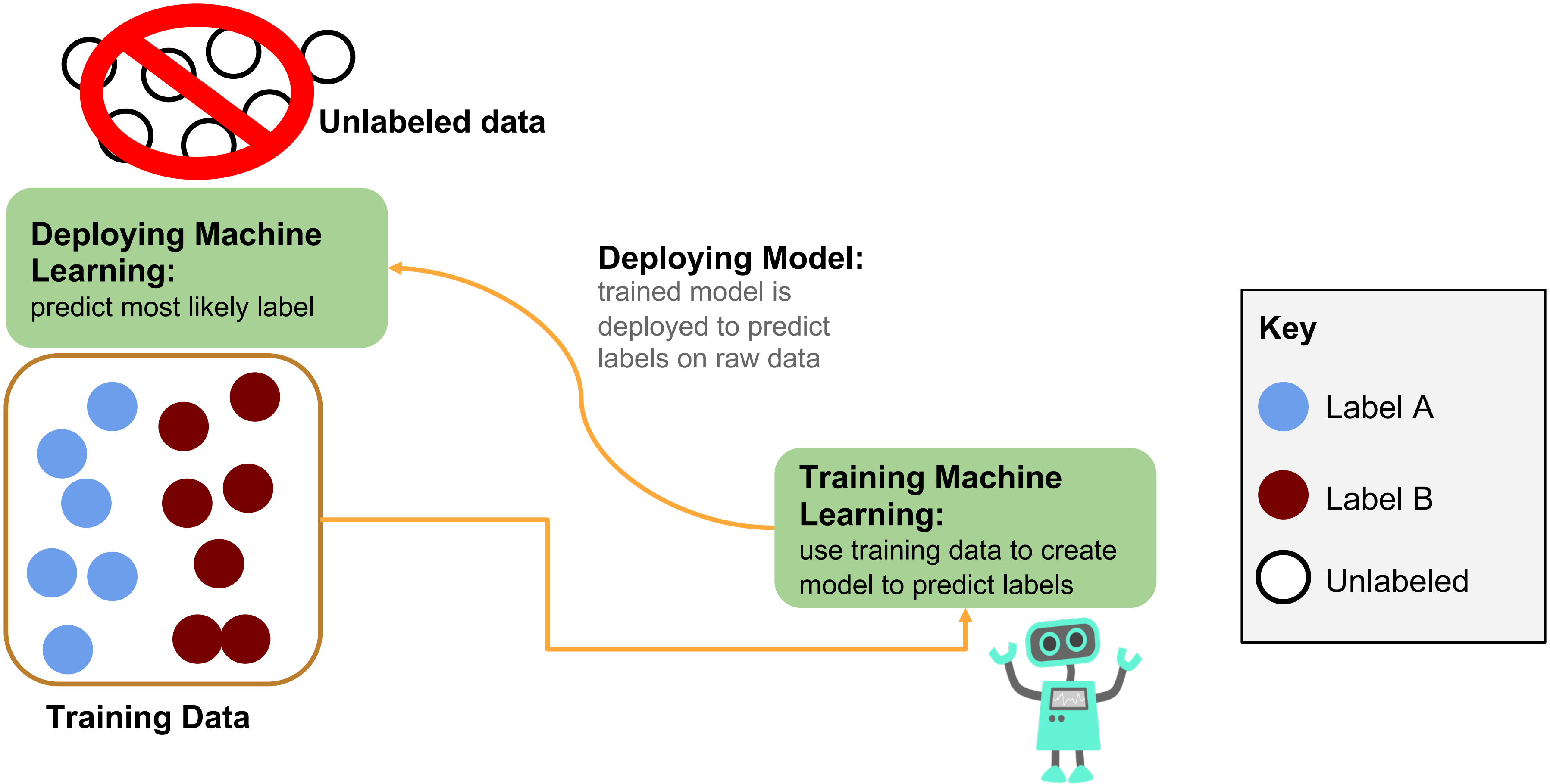
- Text, Images, Video, etc.

But labels are **expensive**

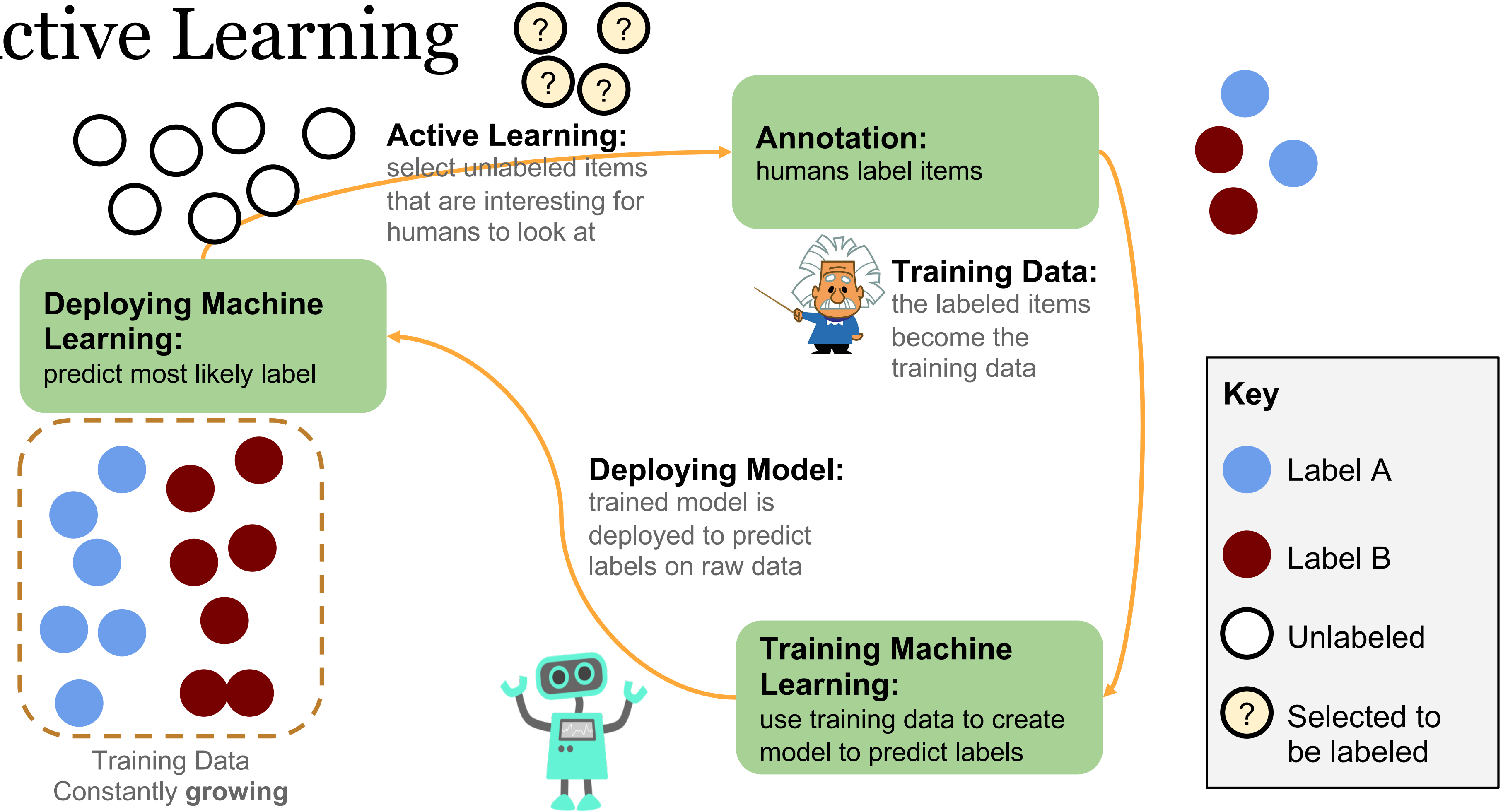


https://www.youtube.com/watch?v=FE1r7_SQq6Y

Supervised Learning



Active Learning

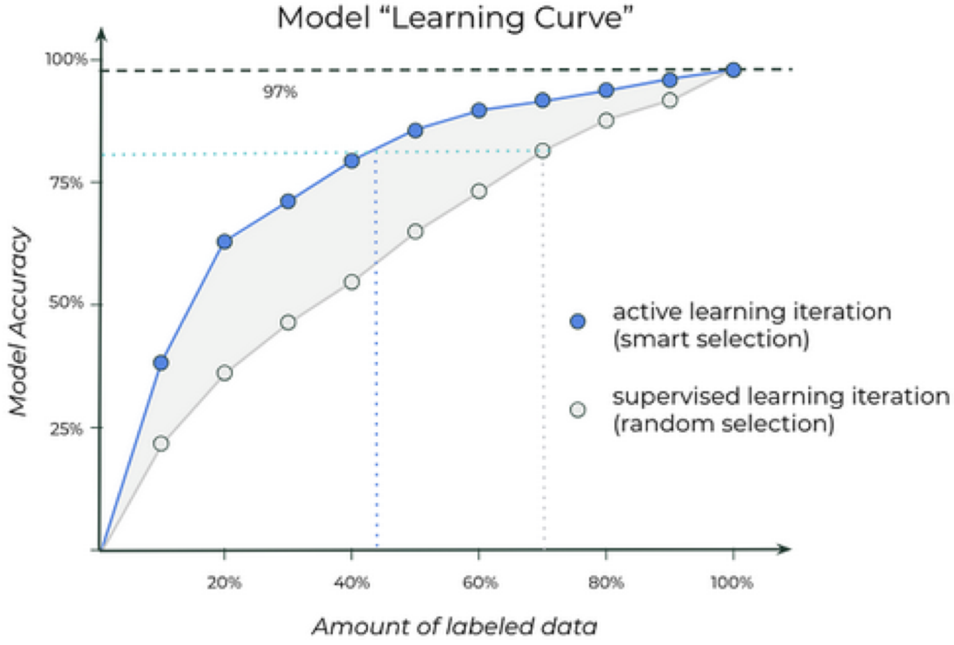


Active Learning

1. **Accuracy:** limited annotation budget (e.g., 1% of the unlabeled data)

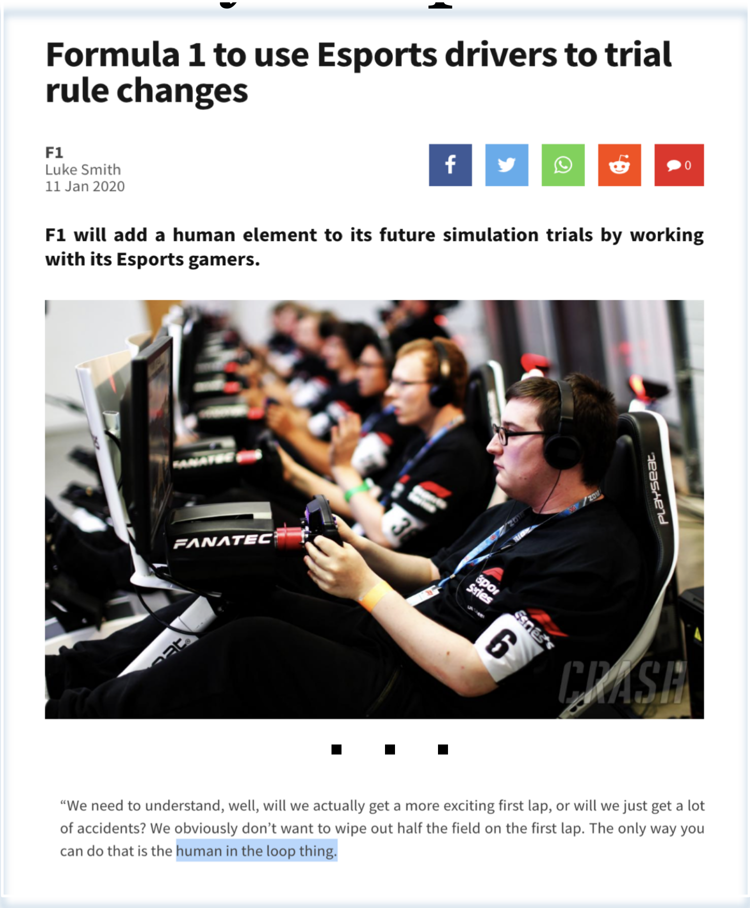
2. **Speed:** train a model to be more accurate more quickly

3. **Diversity:** random sample would be biased



https://www.kdnuggets.com/2018/10/introduction-active-learning.html

Widely adopted



Formula 1 to use Esports drivers to trial rule changes

F1
Luke Smith
11 Jan 2020

F1 will add a human element to its future simulation trials by working with its Esports gamers.

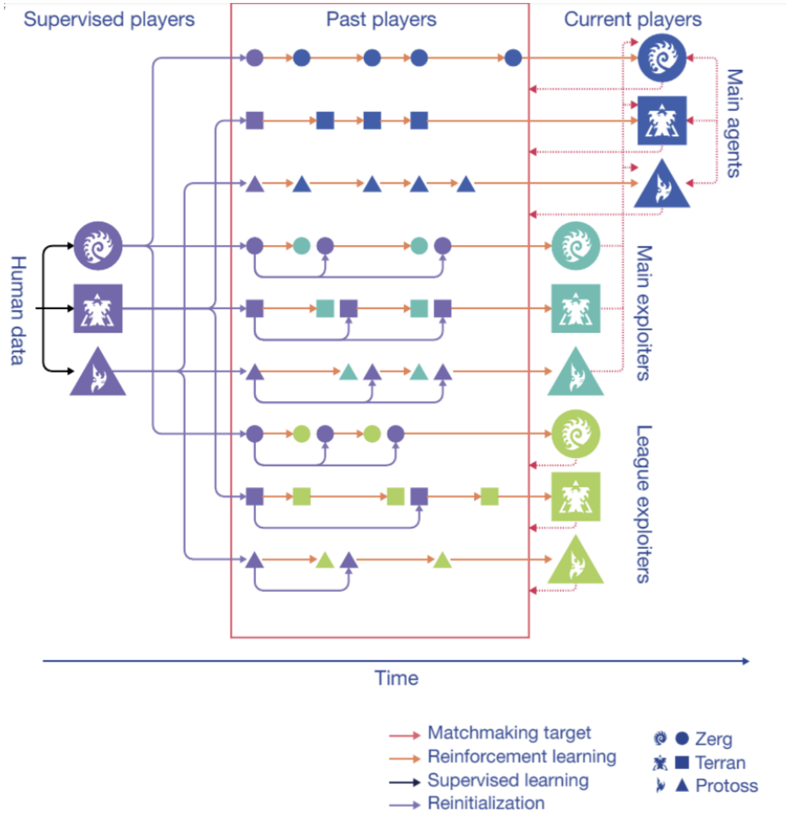
. . .

"We need to understand, well, will we actually get a more exciting first lap, or will we just get a lot of accidents? We obviously don't want to wipe out half the field on the first lap. The only way you can do that is the human in the loop thing.

- Correct bias
- Augment hard examples
- Increase model performance
- Incorporate prior domain knowledge
- Validate model
- Improve interpretability
- Safe AI

https://thumbs.gfycat.com/FavorableMeagerAfricanjacana-size_restricted.gif

Surpassing human intelligence



https://www.youtube.com/watch?v=eHipy_j29Xw

https://www.kdnuggets.com/2019/11/reinforcement-learning-methods-alphastar-outcompete-human-players-starcraft.html

AL Categories

Query Synthesis

- Construct/generate examples for labeling

Selective sampling / Stream-based

- Unlabeled data come as a stream
- For each data point, decide to label or discard

Pool-based

- Set of unlabeled data
- Decide which to label from pool
- Batch (multiple in one go) or online (one at a time)

Knowledge Quadrant for ML

	Knowns	Unknowns
Known	Confident Predictions from Model (Known Knowns) **Current Model State**	Non-Confident Predictions from Model (Known Unknowns) **Uncertainty Sampling**
Unknown	Latent Information in Related Models (Unknown Knowns) **Transfer Learning**	Gaps in Model Knowledge (Unknown Unknowns) **Diversity Sampling**

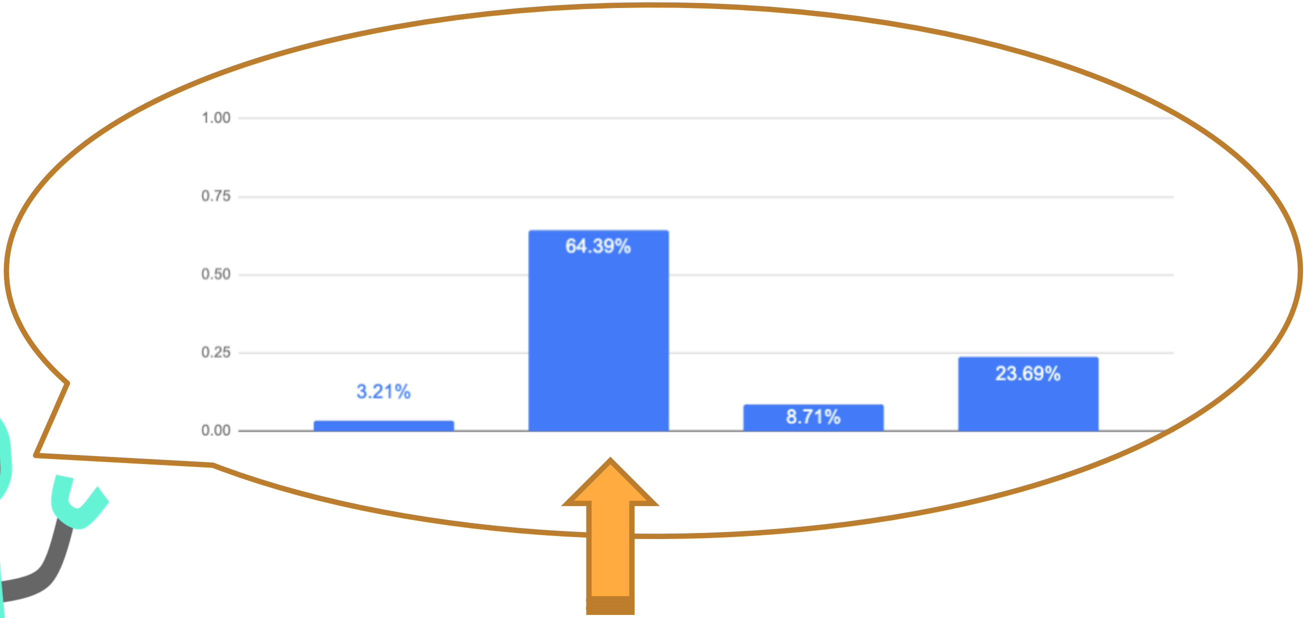
Can be addressed within your current model. Model knows that is uncertain.

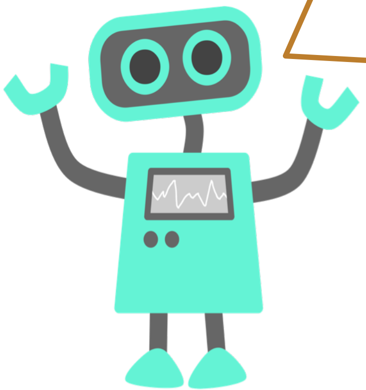
Can be addressed by the right algorithm design: how can we build accurate models that are interpretable and leverage pre-trained models?

Look outside of your model: what other pre-trained models and methods for sampling for diverse unlabeled items should you use?

Acquisition functions

Machine Learning: Classification



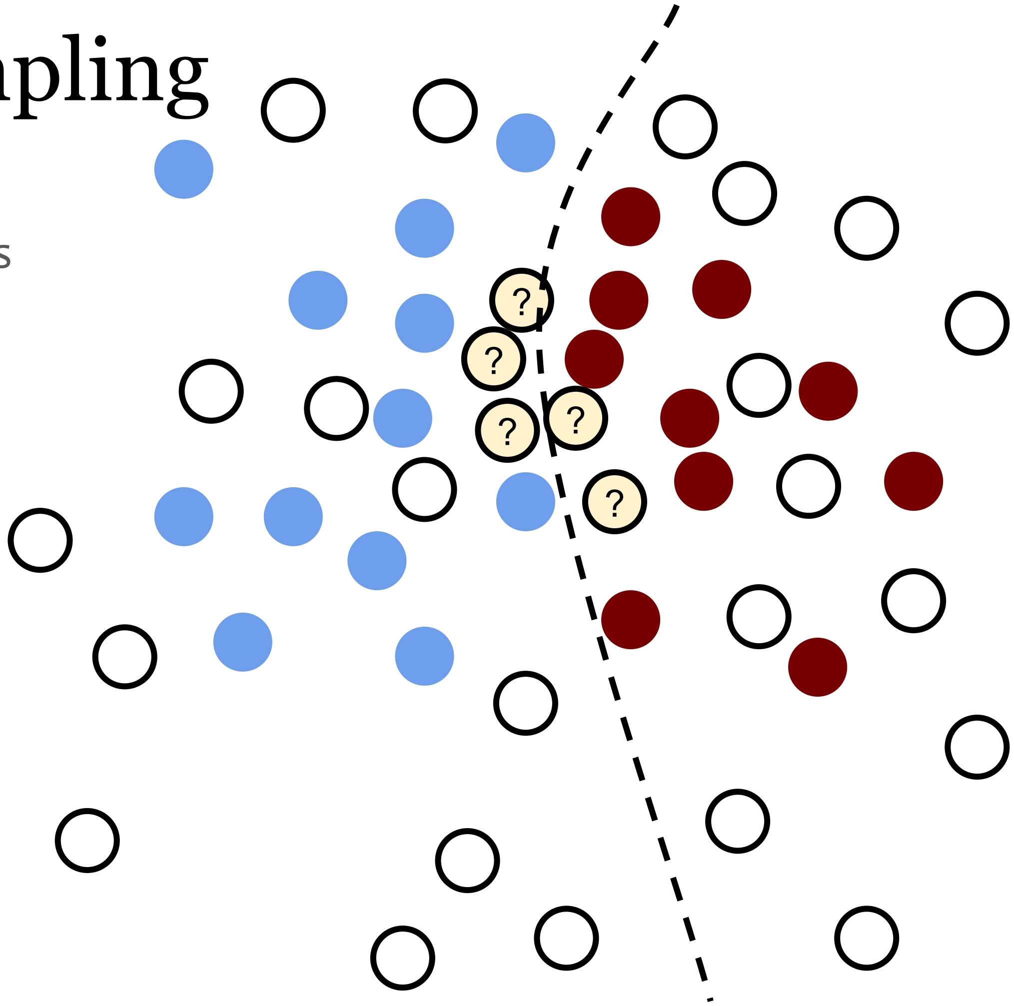


$$f(x)_{LC} = \max_{y} P_{\theta} (y^* \mid x)$$

$|y| = 4$ *(labels)*

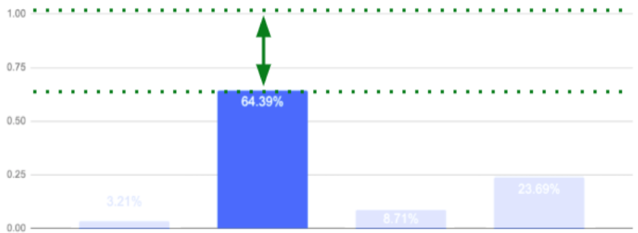
Uncertainty Sampling

Most confusing examples



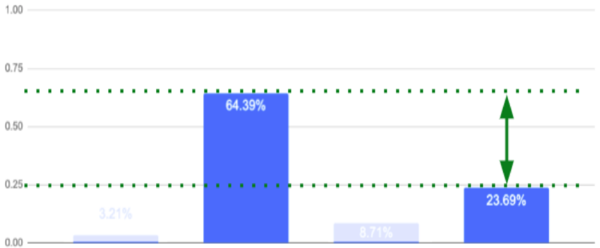
Uncertainty Sampling

Least Confident [1] : difference between most confident prediction and 100% confidence



$$f(x)_{LC} = 1 - P_\theta (y^* \mid x)$$

Margin of Confidence [2]: difference between the top two most confident predictions

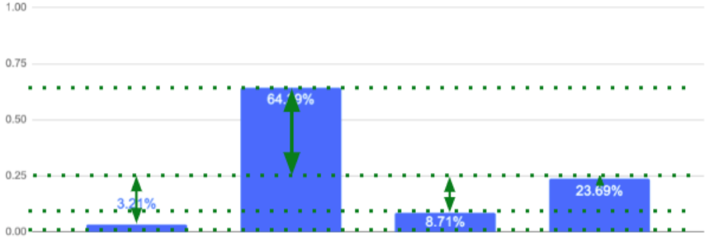


$$f(x)_{MC} = 1 - (P_\theta(y^*_1 \mid x) - P_\theta(y^*_2 \mid x))$$

[1] Culotta, Aron, and Andrew McCallum. "Reducing labeling effort for structured prediction tasks." *AAAI* 2005.
[2] Scheffer, Tobias, Christian Decomain, and Stefan Wrobel. "Active hidden markov models for information extraction." *International Symposium on Intelligent Data Analysis*. Springer, Berlin, Heidelberg, 2001.

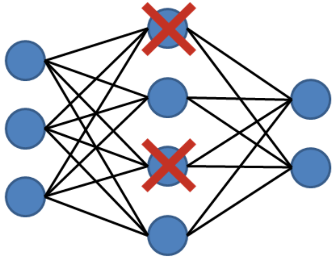
Uncertainty Sampling

Entropy [3]: difference between all predictions



$$- \sum_y P_\theta(y \mid x) \log_2 P_\theta(y \mid x))$$

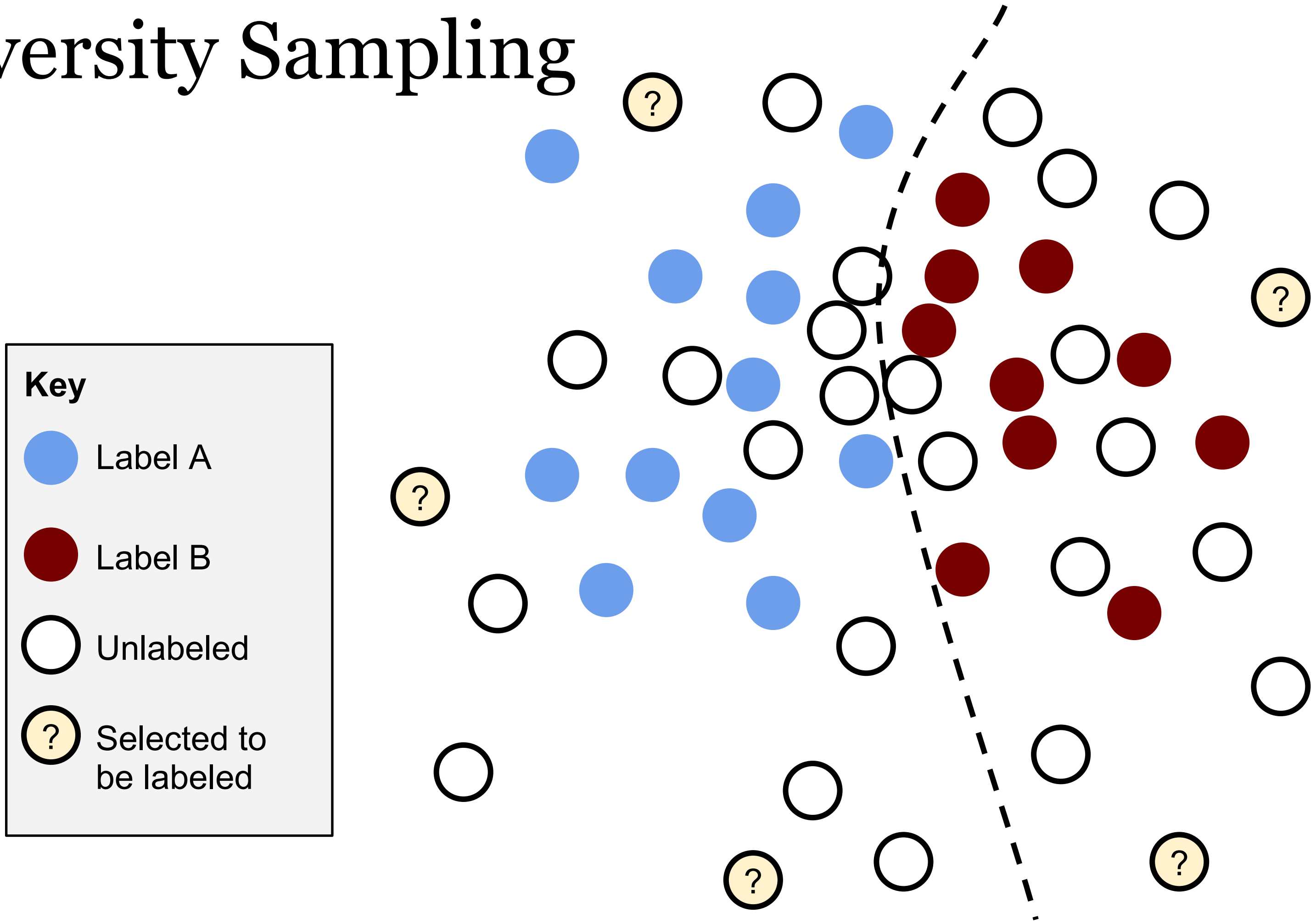
Monte Carlo Dropout [4]: random dropouts during prediction



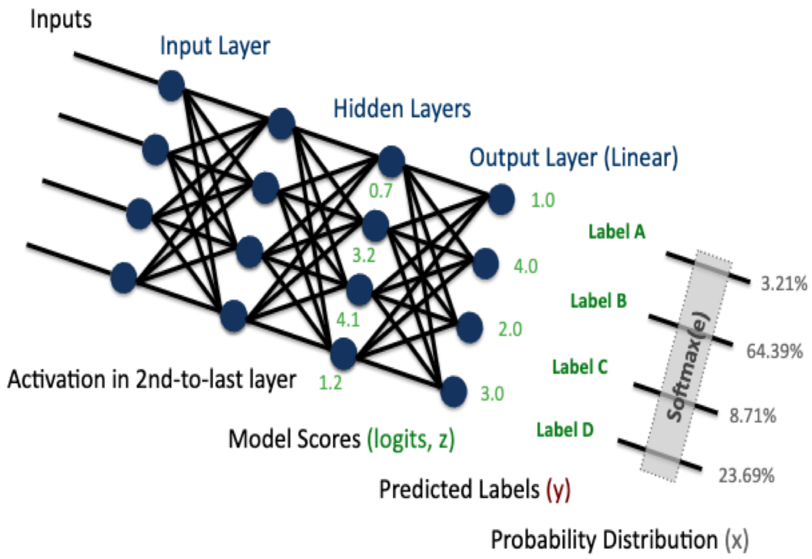
[3] Dagan, Ido, and Sean P. Engelson. "Committee-based sampling for training probabilistic classifiers." *Machine Learning Proceedings 1995*. Morgan Kaufmann, 1995. 150-157.

[4] Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian active learning with image data." JMLR 2017

Diversity Sampling



Diversity Sampling



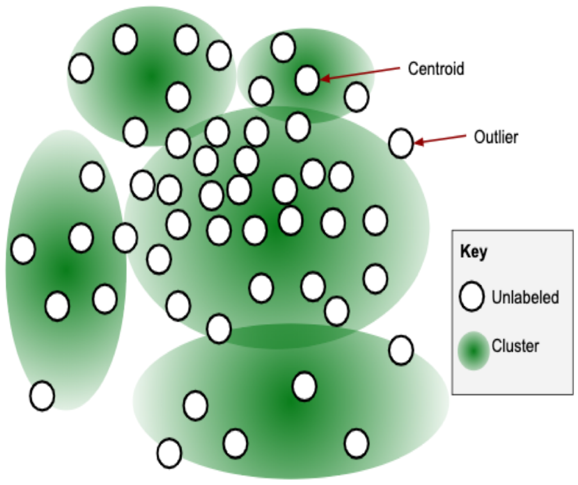
Model-based Outliers:

Create a model for all samples.

Predict outliers as those having large deviations from the established profiles.

- Sampling for low activation in logits and hidden layers.

Diversity Sampling



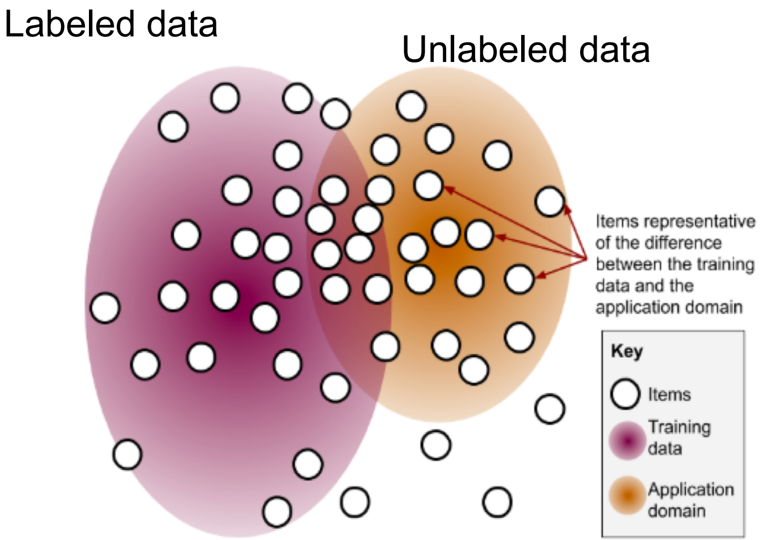
Cluster-based Sampling:

Uses unsupervised learning to pre-segment the data.

Model-agnostic; just use similarity in feature space.

Ensures that you are sampling data from all the meaningful density regions.

Diversity Sampling



Representative Sampling: finding items most representative of the target domain.

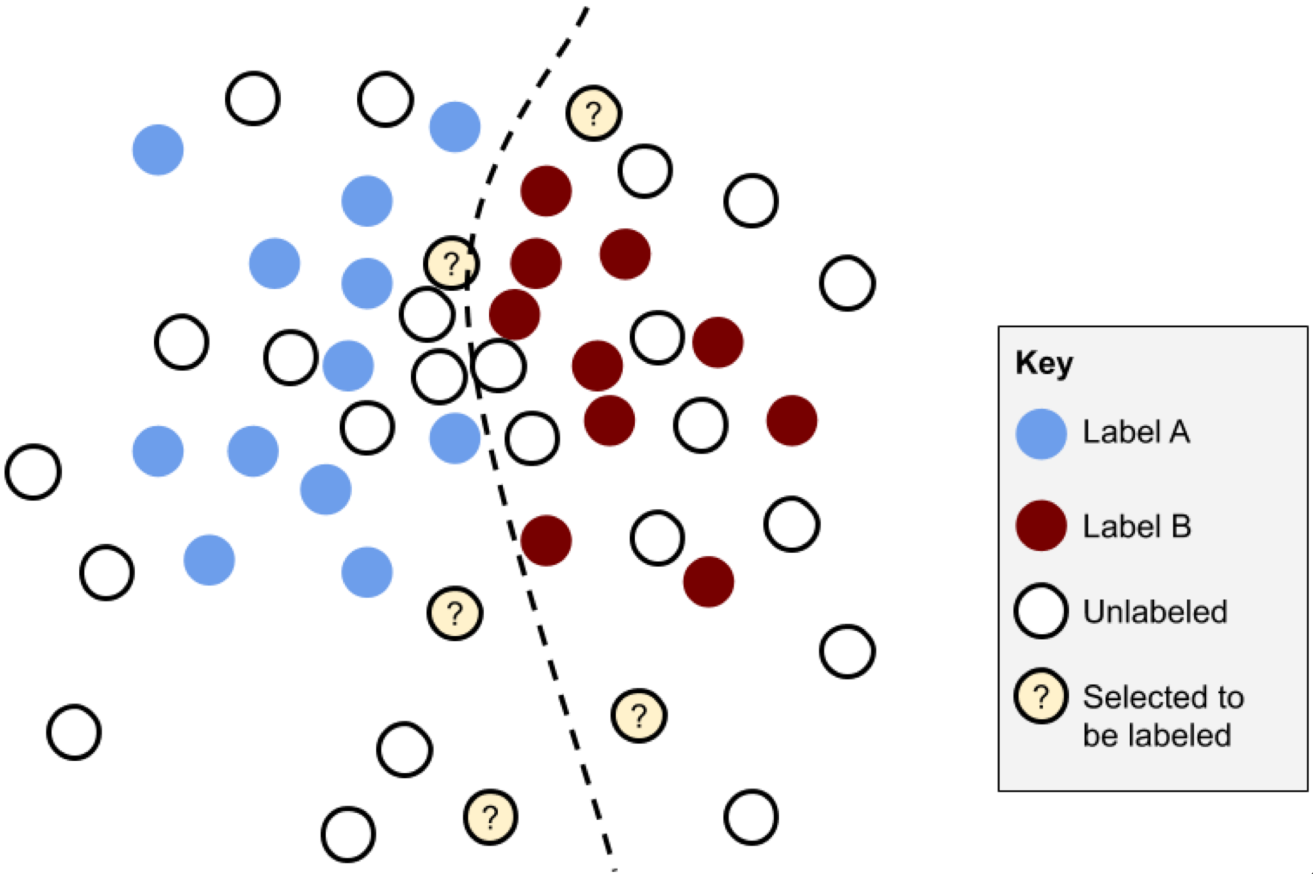
Select example that looks most like your unlabeled data relative to labeled data.

Looking for what you currently do not have a label for that is representative of the unlabeled data.

Idea similar to Domain Adaptation

Density Weighted Uncertainty Sampling

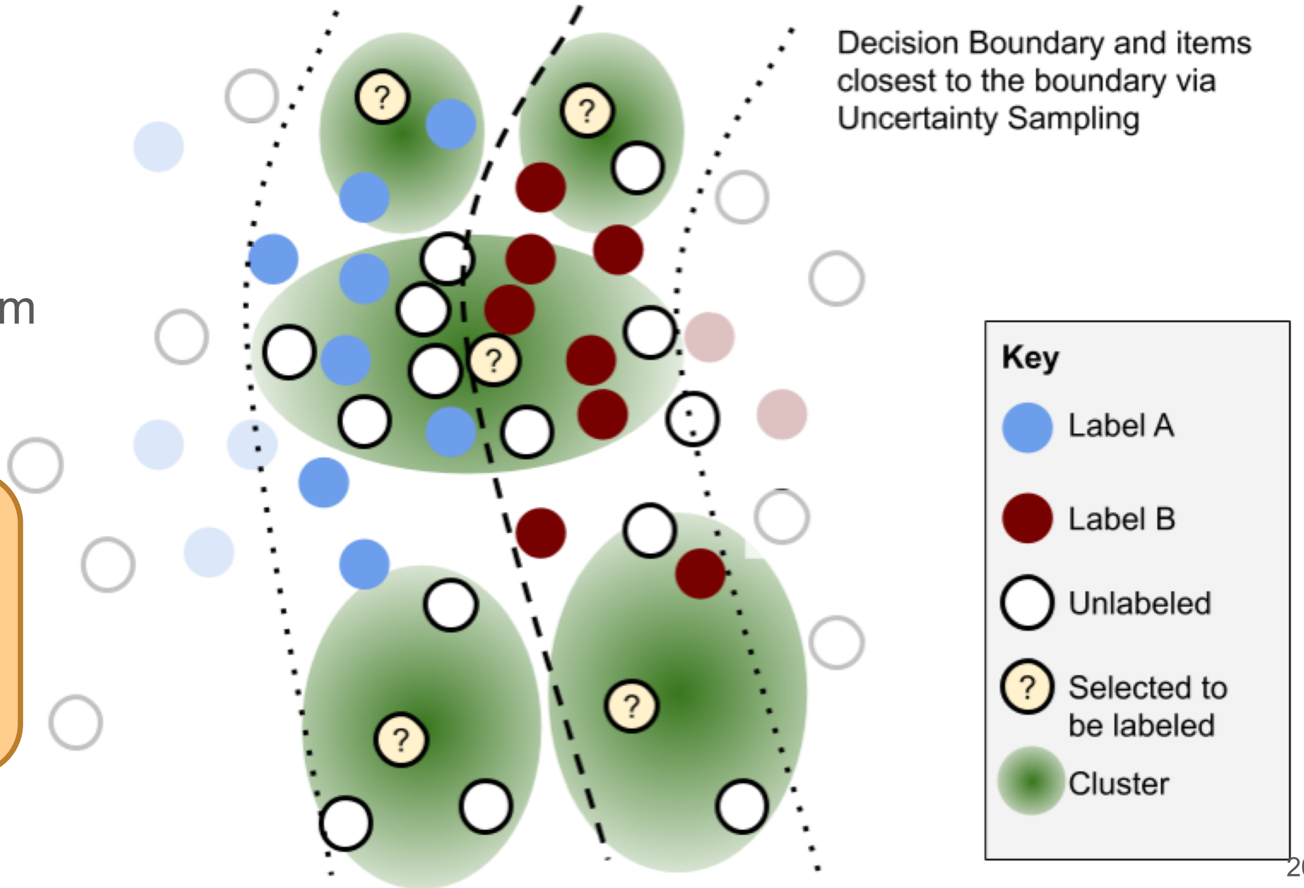
- 😐 Uncertainty will select examples that look too much like each other.
- 😐 Diversity will pick examples too far from the decision boundary.
- ✓ Sampling items that are both uncertain and diverse.



Density Weighted Uncertainty Sampling

Uncertain + Diverse = Sampling items that are near the decision boundary but distant from each other

1. Discard items far from decision boundary
2. Apply clustering
3. Sample items from each cluster (centroid)



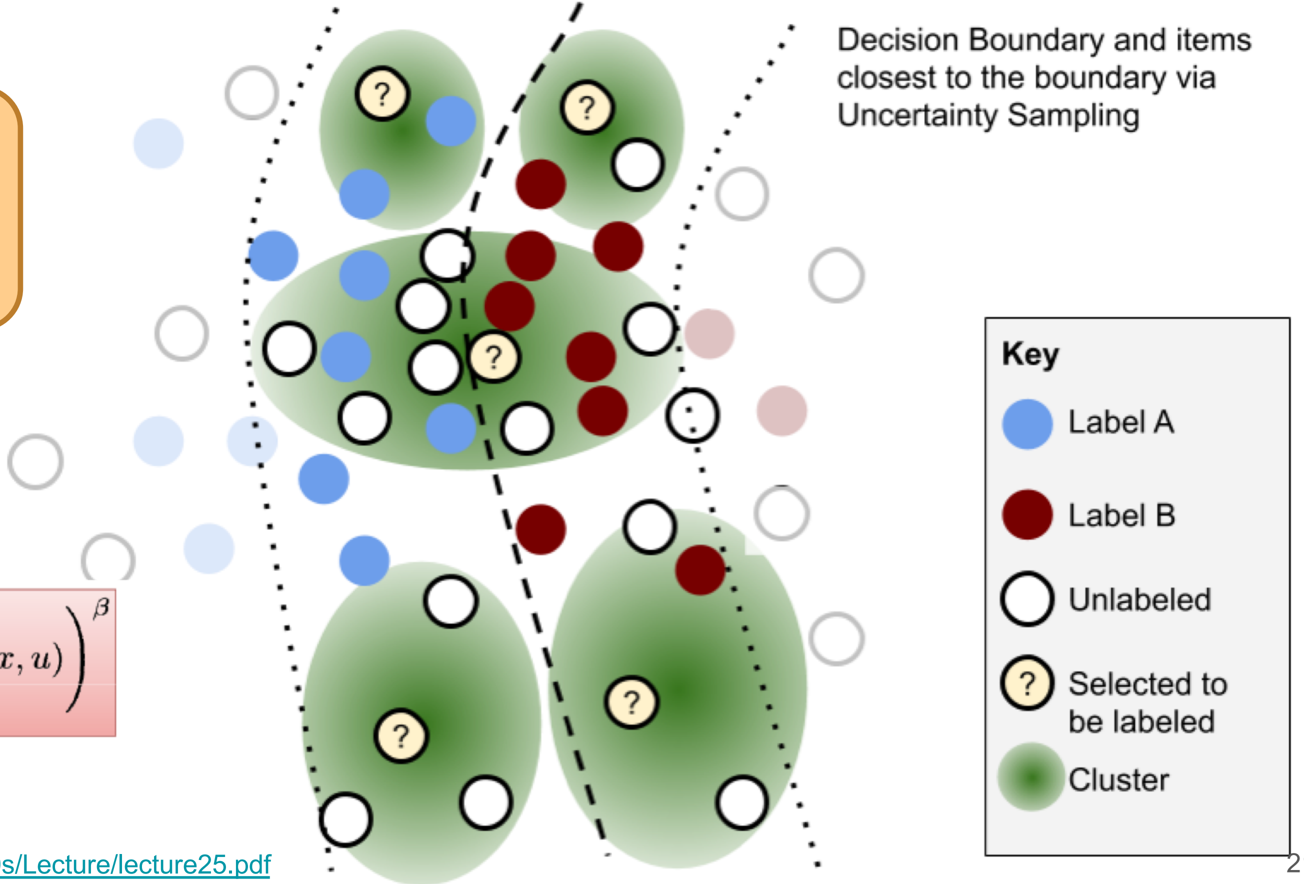
Density Weighted Uncertainty Sampling

Weight the uncertainty of an instance by its density w.r.t. the pool

$$\phi_{ID}(x) = \boxed{\phi(x)} \times \boxed{\left(\frac{1}{U}\sum_{u\in\mathcal{U}} \mathrm{sim}(x,u)\right)^{\beta}}$$

"base" informativeness

density term



https://www.cs.cmu.edu/~epxing/Class/10701-10s/Lecture/lecture25.pdf

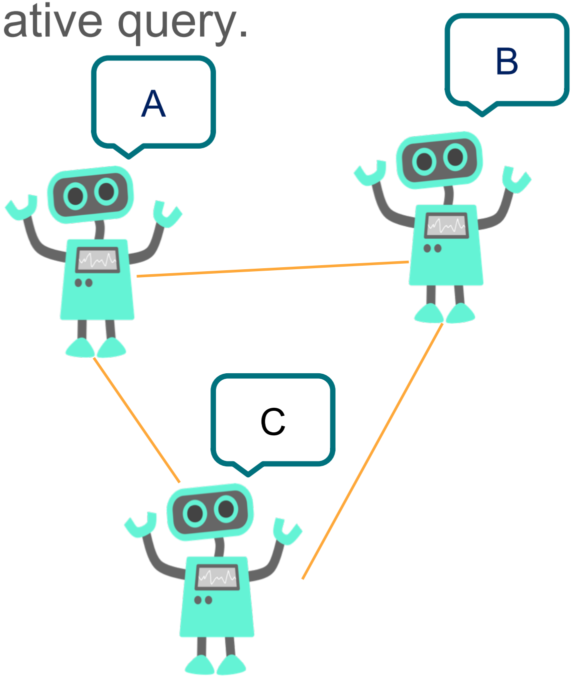
Query-by-committee

Committee of models trained on same labelled set.
Each member is allowed to vote on labelling of query candidates.
Instance about which they most **disagree** is most informative query.

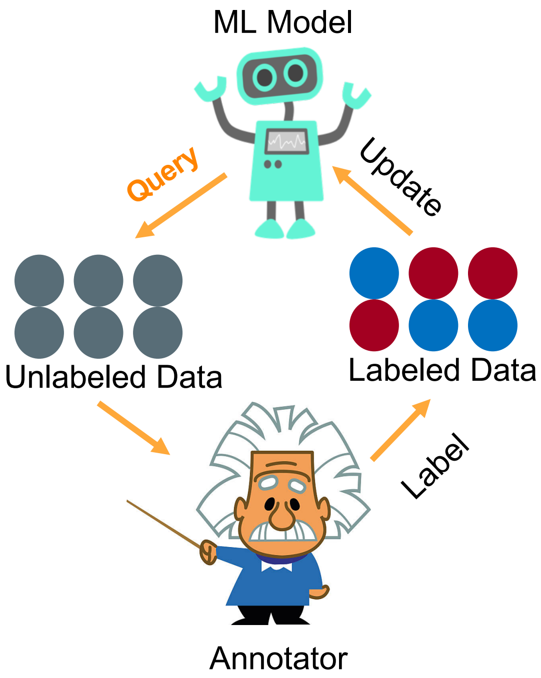
Vote entropy

$$\underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

Selects instance for which entropy of the vote distribution is the largest.



Active learning "on-the-fly"



<u>So many acquisition functions</u>

rs: random

us: uncertainty

ds: diversity

dwus: density weighted uncertainty ….

Which heuristic is best?

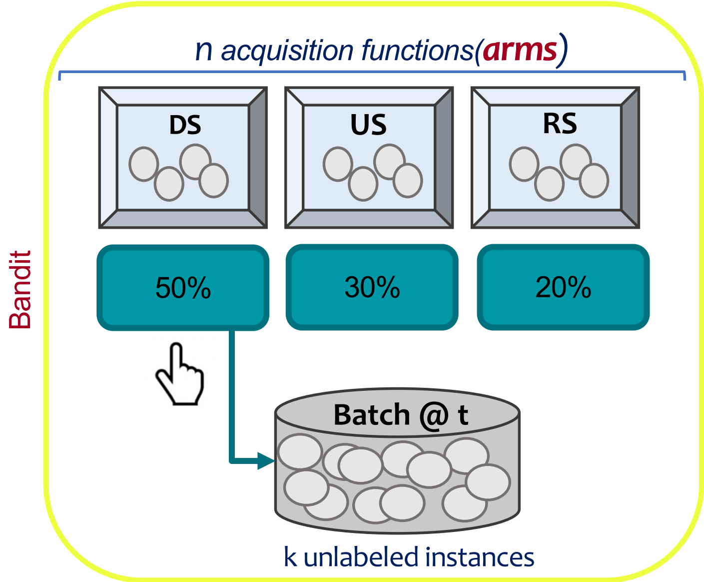
Find the most **efficient** way to **query** unlabeled data.

Train ML model with the minimal amount of human supervision.

Dynamically decide on a task-driven basis

Decide which technique to use on-the-fly

Given no a-priori knowledge on which active learning strategy will work best



Active learning by learning [1]

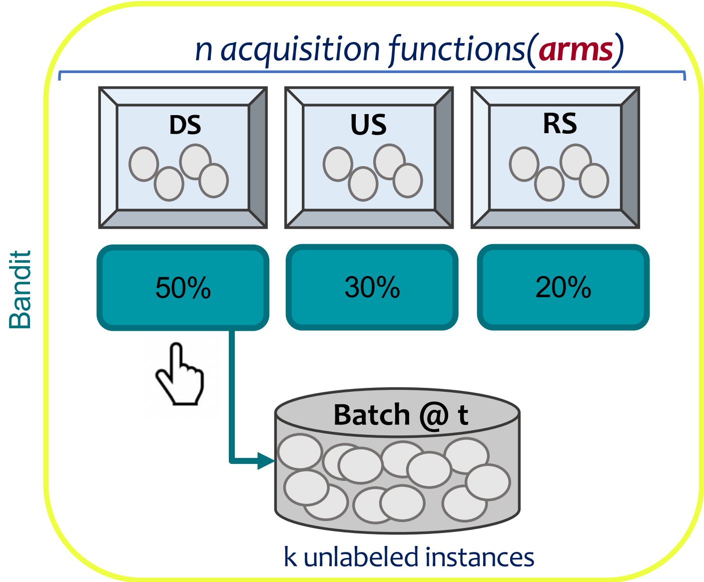
- Over-exploit (Winner takes all prematurely)
  - Ensure sufficient **exploration**
- Keep all players (Do we need the worst?)
  - Eliminate **unpromising** candidates

[1] Hsu Wei-Ning and Hsuan-Tien Lin. "Active learning by learning" *AAAI* 2015

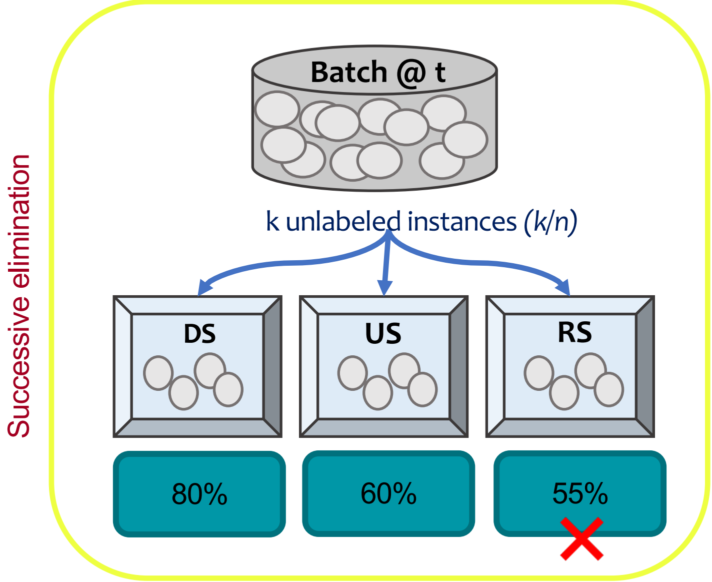
Dynamically decide on a task-driven basis

Decide which technique to use on-the-fly

Given no a-priori knowledge on which active learning strategy will work best



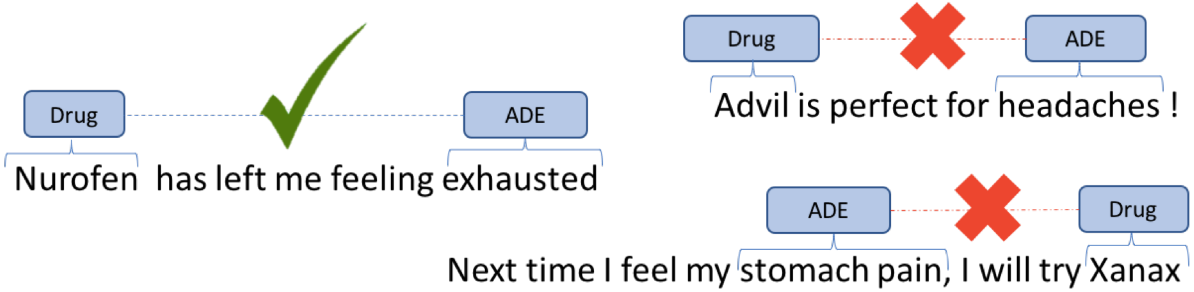
Active learning by learning



Prune: Dynamic Budget Allocation

Experimental Setup

Causal relationships between Drugs and Adverse Drug Events (ADEs)



<u>6 active learning strategies:</u>

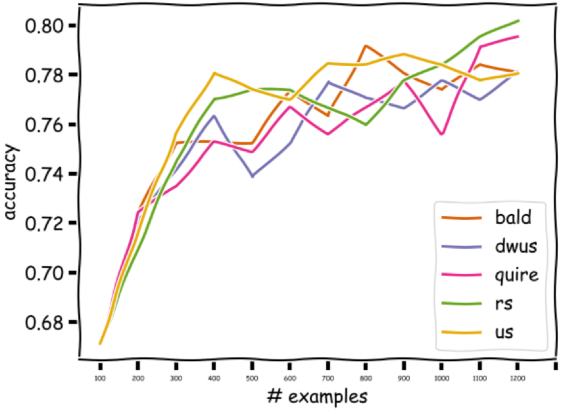
rs random
us uncertainty
ds diversity
dwus density weighted uncertainty
quire informative + representative
bald bayesian AL by disagreement

ALBL: Active learning by learning

Prune: Eliminate least efficient learner

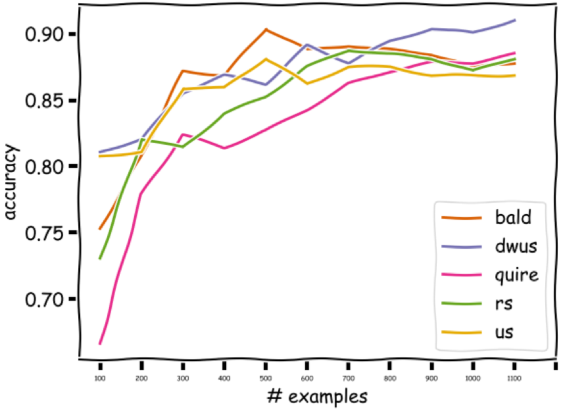
No solution fits all

Best Method: **random**



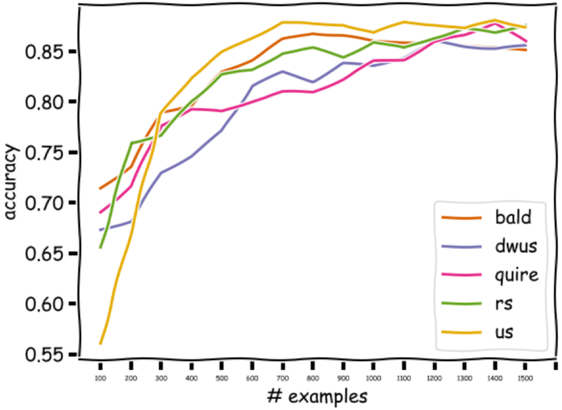
CausalADEs

Best Method: **dwus**



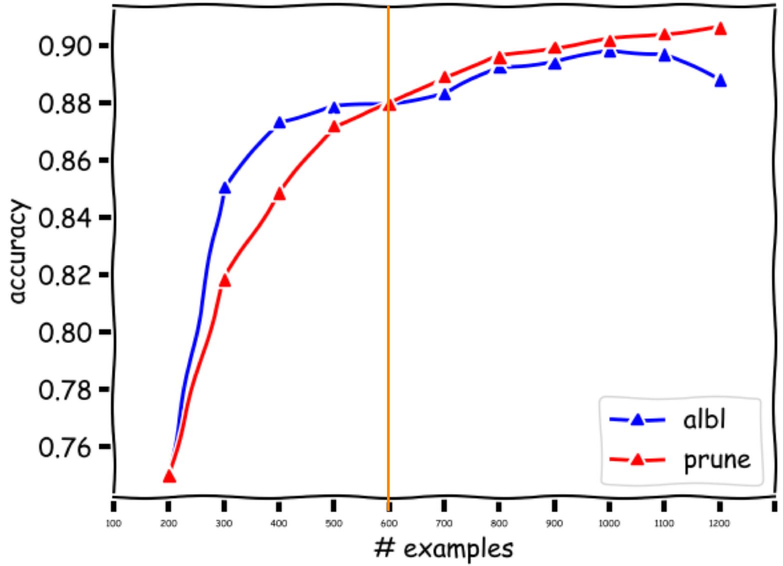
Instrument - Agency

Best Method: **us**

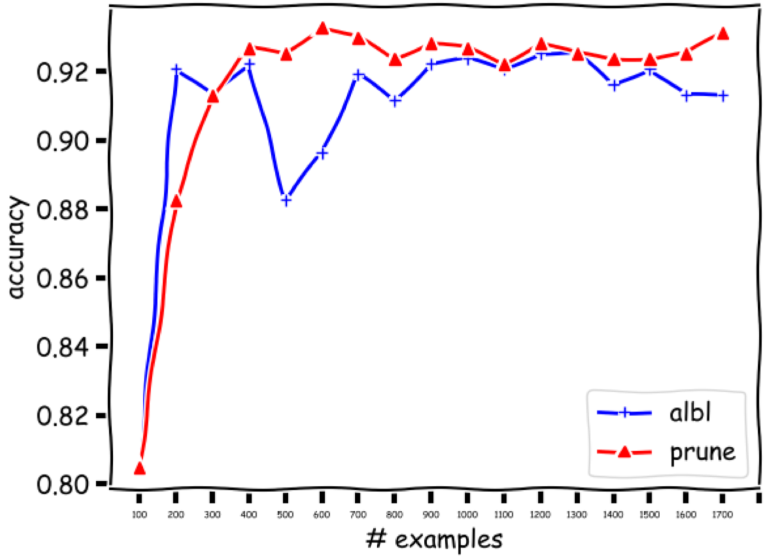


Product - Producer

Prune vs albl: same human effort → better label quality



Average on 10 relations



Cause - Effect

Batch Active Learning

Batch active learning: select a **batch** of instances at each iteration
Trade-off between **efficiency** and **performance**
Large batches result in…

- Less frequent model updates
- Increased prediction error

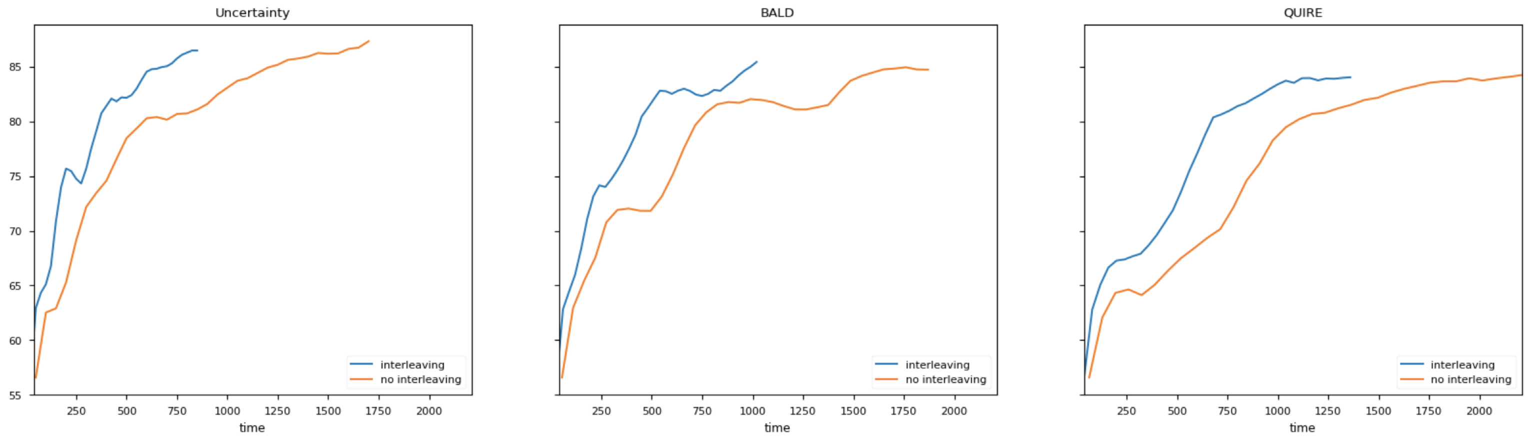
Computing the next "batch" & loading it into the UI for the SME to score takes time
Annotation time is the largest cost in a HuML system

- In an ideal world SMEs would be scoring constantly.
- Over 80% of the time the user is waiting!

Interleaving to reduce waiting time

Keep **last unlabeled batch** for future scoring

- Use $B_0 \ldots B_{n-2}$ batches to produce next batch B_n
- User scores batch Bn−1 while system ranks the next batch B_n



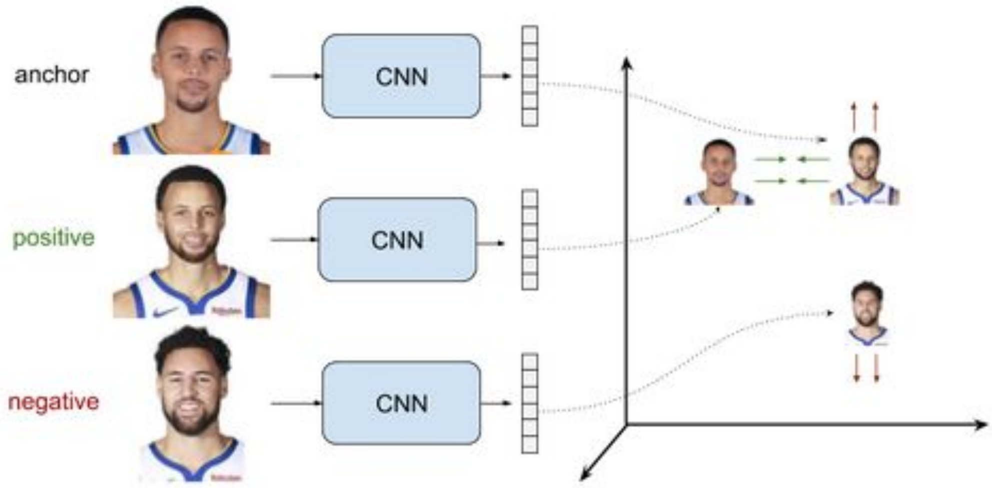
✓ **Continuous** human work

✓ Comparable performance, in **≈ 50% less training time**

Active (Ordinal) embedding

Methods that seek to recover a **distance metric** without underlying features

Learn a representation of the data points that **preserves the distance information** provided in the similarity **triplets**/pairs

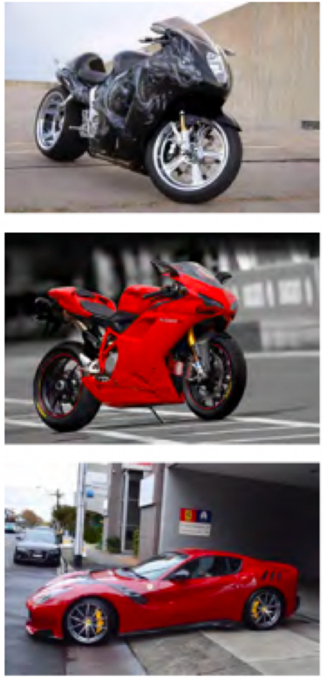


https://gombru.github.io/2019/04/03/ranking_loss/

Pairwise vs. Relative Comparisons

Are A and B similar?

Constraints:
Cannot link
Must link



Is A closer to B or C?

Constraints:
d(A,B) ? d(B,C)

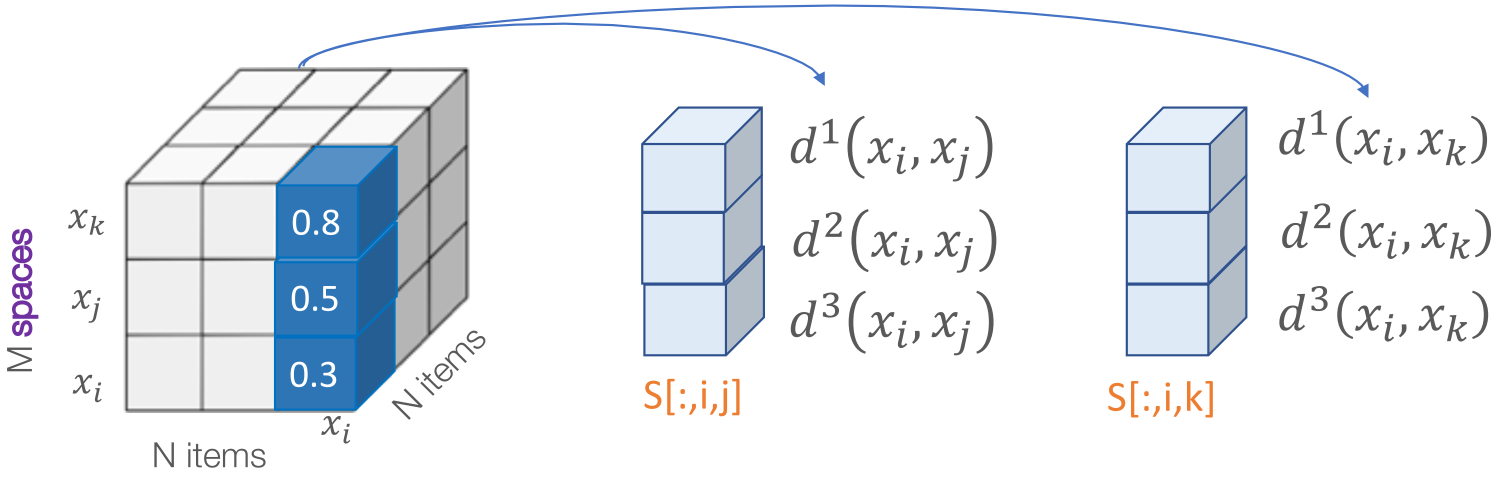
Zeng, Qiong, et al. "Bundle Optimization for Multi-aspect Embedding." *arXiv:1703.09928* (2017)

Problem Definition (multiple distances)

Set of objects $\mathcal{X} = \{x_1, x_2, \ldots, x_N\} = \{\text{cat, dog, mango, } \ldots\}$

$S \in \Re^{M\,x\,N\,x\,N}$ distances matrices, i.e. S_{mij}: $d^m(x_i, x_j)$ distance of x_i, x_j in space m

- Each space defines a different similarity between objects, i.e. “multiple views”



Problem Definition (triplets)

Triplet query (i, j, k): Is x_i more similar to x_j or x_k?

- Specifies relative distance in at least one of the spaces, i.e. $d^m(x_i, x_j) < d^m(x_i, x_k)$

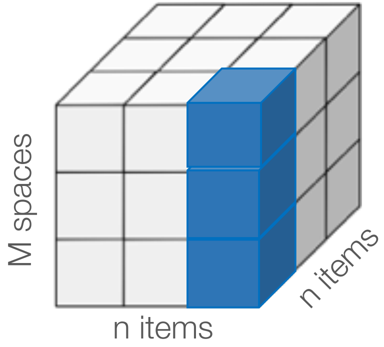
Is $\boxed{\text{cat} \; x_i}$ closer to $\boxed{\text{dog} \; x_j}$ or $\boxed{\text{mango} \; x_k}$?

$$\Downarrow$$

$$d(\text{cat}, \text{dog}) < d(\text{cat}, \text{mango})$$

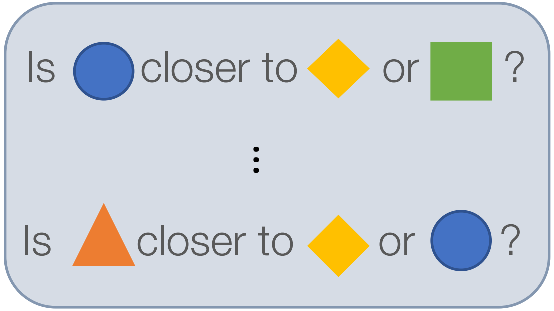
Goal: Find space S^* that best matches the user's view, adaptively by ranking spaces S^m and with as few queries as possible

Which queries?



Input: $\Re^{M \; x \; N \; x \; N}$

similarity matrix



Input: $\mathcal{T} = \{(i,j,k)\}$

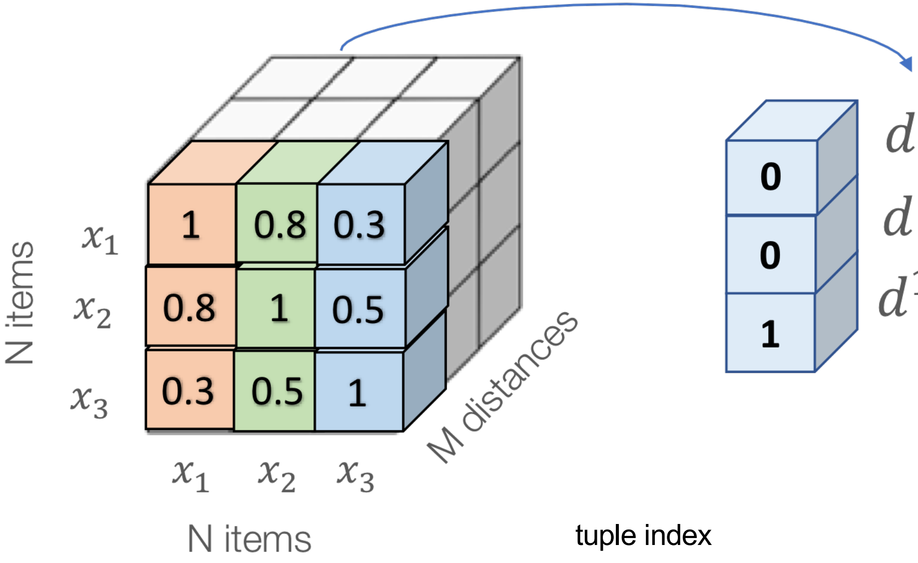
with N items → set of all queries = $\mathbf{N}\binom{N-1}{2}$

$$\delta^{(\mathcal{Q})}(x_i; x_j) < \delta^{(\mathcal{Q})}(x_i; x_k)$$

Ordinal Oracle

$\mathcal{Q}$ $\quad \delta^{(\mathcal{Q})} : \mathcal{X} \times \mathcal{X} \times \mathcal{X} \to \{0, 1\}$

Which queries? (Ordinal Vectors)



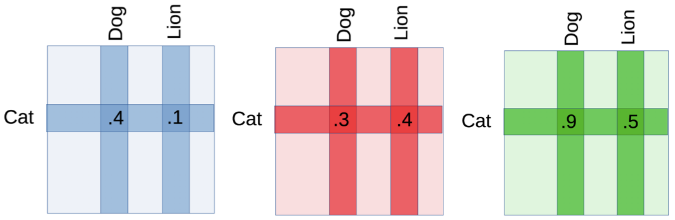
$$d^1(x_1, x_2) > d^1(x_1, x_3)$$
$$d^1(x_2, x_1) > d^1(x_2, x_3)$$
$$d^1(x_3, x_1) < d^1(x_3, x_2)$$

tuple index

Distance matrix

$$[O_m]_\tau = \begin{cases} 1, & \text{if } (x_i, x_j, x_k)_\tau \\ 0, & \text{otherwise} \end{cases}$$

where $[\cdot]_\tau$ is an entry identifier (index).



D(Cat,Dog) > D(Cat,Lion) D(Cat,Dog) < D(Cat,Lion) D(Cat,Dog) > D(Cat,Lion)

Is D(Cat,Dog) > D(Cat,Lion)?

(1, 0, 1)

Perfect setting → elimination

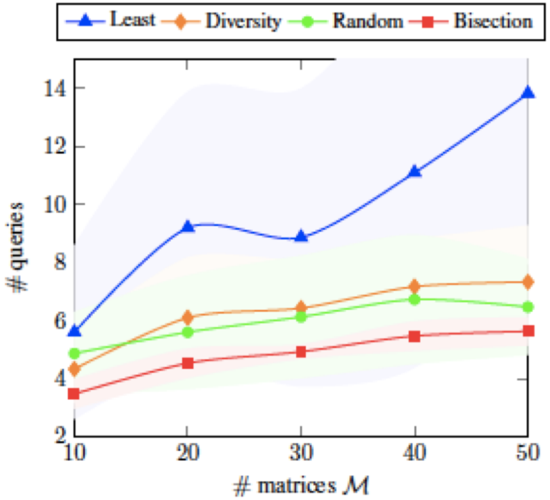
✓ At least one of the distance matrices satisfies all oracle-provided triplets

✓ Oracle is deterministic and replies correctly

Select the triplet that maximally divides the space

logM ordinal queries

PROBLEM SOLVED!



Least satisfied triplet
Most **diverse** triplet
Random triplet
Bisection search

Approximate setting

✓ Oracle's view not perfectly aligned with any of the given distance matrices
✓ Noisy oracle that replies incorrectly

How many queries do we need in order to find a good approximate choice of distance?

Allow bypassing low-confidence/ambiguous questions?

Reading Roles (presenting)

Two (2) students will team up
Pairings can change in each class

- **Presenter:** Create the main presentation
  - Motivation, problem definition, method, experimental findings

- **Archaeologist:** Previous and subsequent work report
  - Older paper that substantially influenced current paper
  - Newer paper citing current paper

- **Industry Expert:** Propose new application or company product based on paper
  - Discuss positive and negative impact of this application.
  - Convince your industry boss that it's worth investing time and money to implement this paper.
  - With arguments particularly applicable to the chosen industry market.

- **Hacker:** Implement a small part of the paper
  - On a small dataset or toy problem or any other simplified version of the paper.
  - Share a Jupyter Notebook with code
  - DO NOT simply download and run an existing implementation
  - You can use existing implementations for "backbone" code (build model, load data, train, etc.)

Reading Roles (presenting)

Designed for solo work, 1 student

- **Reviewer:** Complete review of the paper.
  - Follow NeurIPS review questions 1-6 under "Review Content"
  - Assign Overall score (question 9) + Confidence score (question 10)

- **Researcher**: Propose follow-up project that has become possible due to the existence and success of the current paper

- **Ethicist**: You are an ethical impact assessor from 2021 (or even 2051). What has been the impact (good or bad) of this paper on the economy, society, and/or the environment?

- *Any roles too difficult?*
- *Prefer solo or 2 students together?*
- *Any roles preferred to be made optional?*
- *Keep same subtopic twice for a group or each group present for each subtopic?*

Diversity & Inclusion discussion

Ethicist: You are an ethical impact assessor from 2021 (or even 2051). What has been the impact (good or bad) of this paper on the economy, society, and/or the environment?

Bring a relevant blog post, news article or paper about diversity & inclusion and bias in ML/NLP/CV/DS to discuss in class

Black in AI

A place for **sharing** ideas, **fostering** collaborations and **discussing** initiatives to increase the presence of **Black people** in the field of **Artificial Intelligence**.

blackinai.org/

LXAI

WiML

Women in Machine Learning

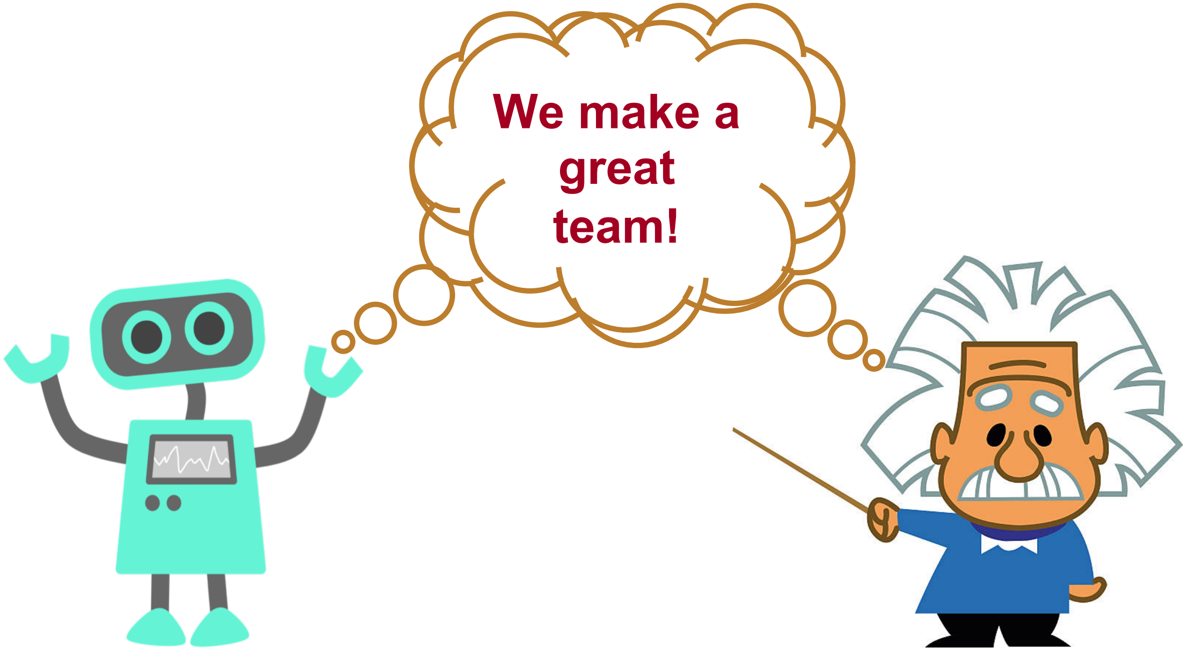
{DIS}ABILITY IN AI

Next Class

Tuesday, 01/26/2021	Semi-supervised: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks
Thursday, 01/28/2021	Active Learning: Learning How to Active Learn by Dreaming

Groups & students can suggest papers
Post your suggestions on Piazza

Thank you! Questions?



References:

- Settles, Burr. *Active learning literature survey*. University of Wisconsin-Madison Computer Sciences, 2009.
- Lewis, David D. and Catlett, Jason. ”Heterogeneous uncertainty sampling for supervised learning”, *Machine learning proceedings* 1994
- Brinker, Klaus. "Incorporating diversity in active learning with support vector machines”, ICML 2003
- Huang, Sheng-Jun, Rong Jin, and Zhi-Hua Zhou. "Active learning by querying informative and representative examples”, *NeurIPs* 2010
- Nguyen, Hieu T., and Arnold Smeulders. "Active learning using pre-clustering”, *ICML* 2004
- Gal, Yarin, Riashat Islam, Zoubin Ghahramani. "Deep bayesian active learning with image data." JMLR 2017
- Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight uncertainty in neural networks", *ICML* 2015
- Houlsby, Neil, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. "Bayesian active learning for classification and preference learning" arXiv 2011
- Hsu Wei-Ning and Hsuan-Tien Lin. “Active learning by learning” *AAAI* 2015