ACTIVE LEARNING

ROSEN
empowered by technology

WHOAMI

Hendrik Niemeyer

Theoretical Physics

Data Scientist

@hniemeye

https://github.com/hniemeyer

hniemeyer@rosen-group.com

Slides: https://github.com/rosen-group/conferences

INTRODUCING THE ROSEN GROUP

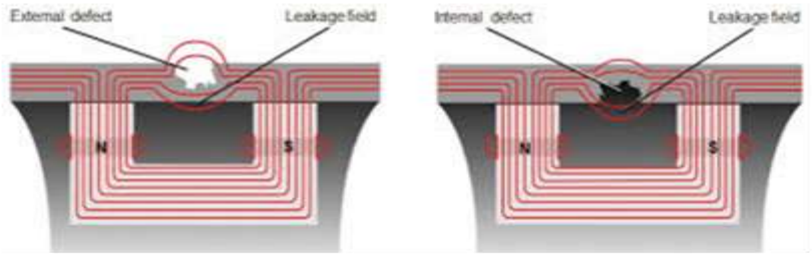
ROSEN develops and manufactures equipment, software, and methods for the inspection, diagnosis, and protection of industrial structures in a wide range of industries.

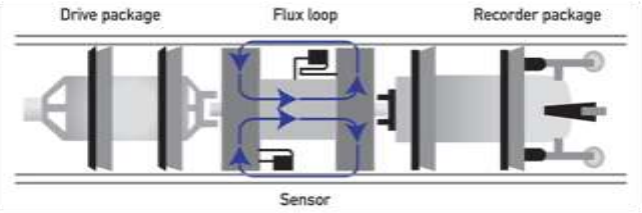
Because damage can cause serious impacts!

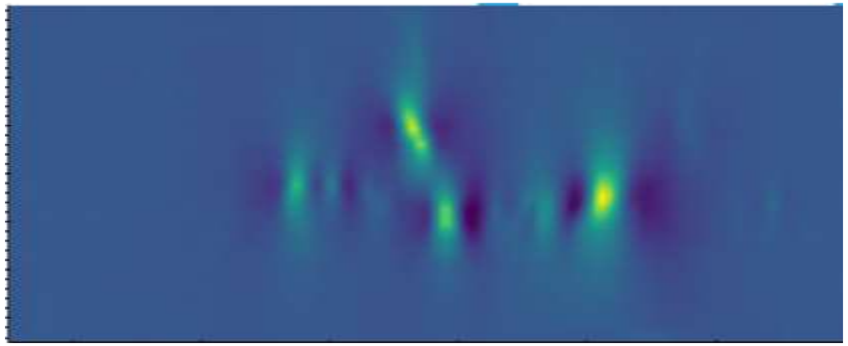
MAGNETIC FLUX LEAKAGE

- Measure volume loss in pipeline wall
- Indirect measurement principle
- Image-like data (2d array of amplitudes)

- Tasks: Detect, classify and estimate defect geometry from measured data.



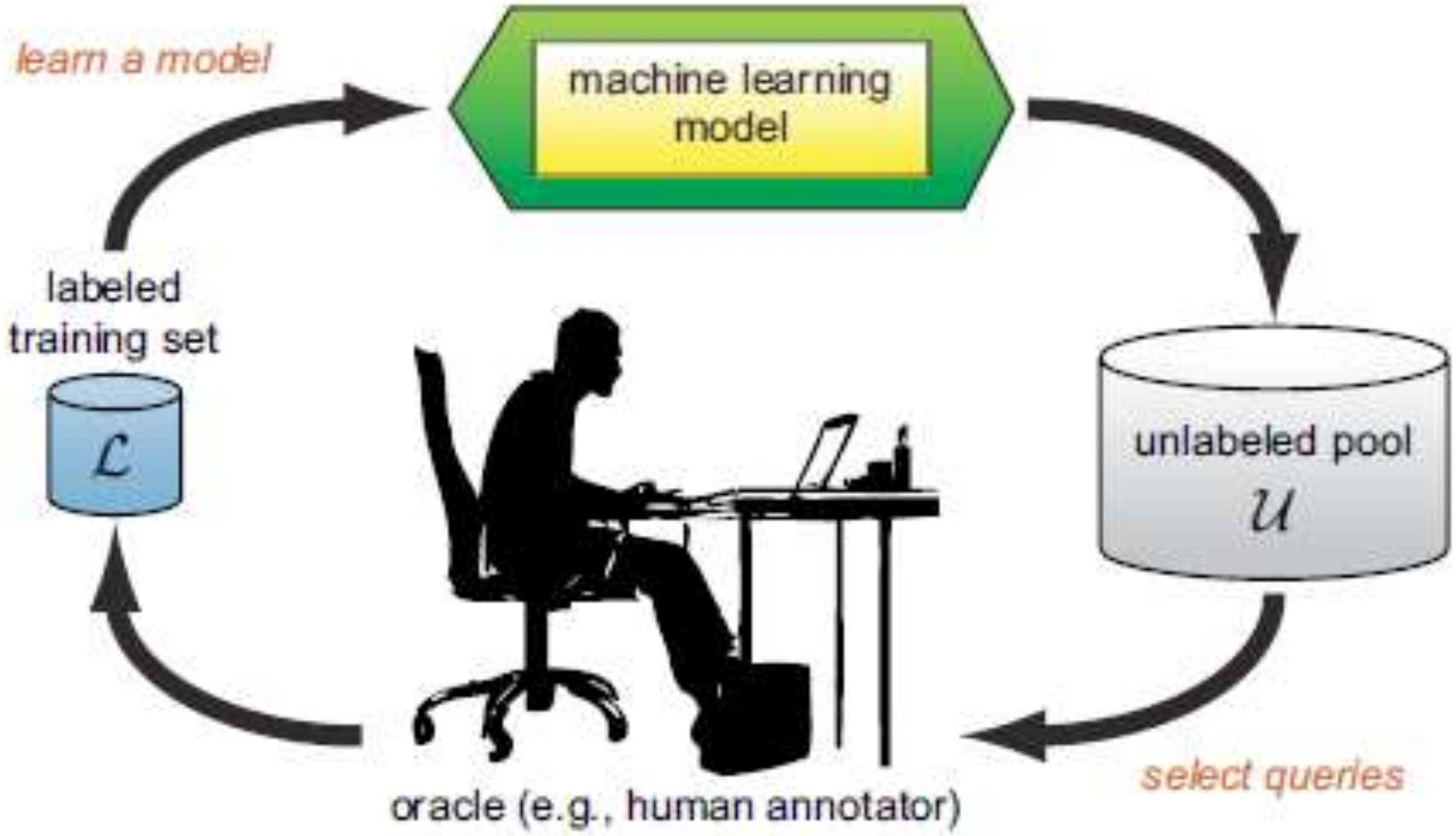




MOTIVATION FOR ACTIVE LEARNING

- Computer Vision problems
  - Large amounts of unlabeled data available
  - Human annotators can provide ground truth
  - Which instances to label?
  - Achieve high accuracy using as few labeled instances as possible

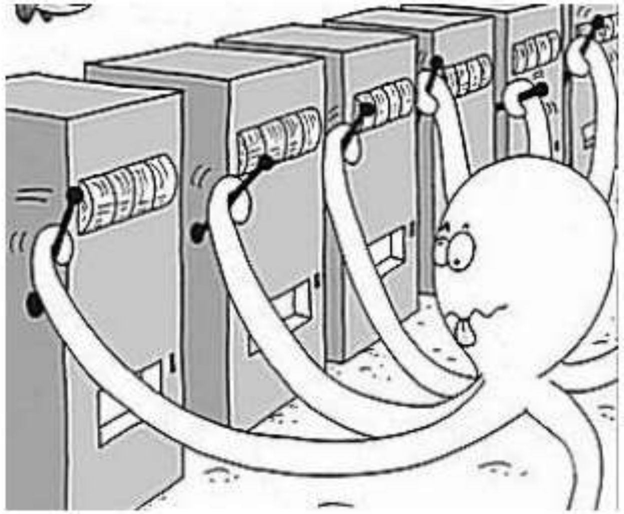
WHAT IS ACTIVE LEARNING?



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

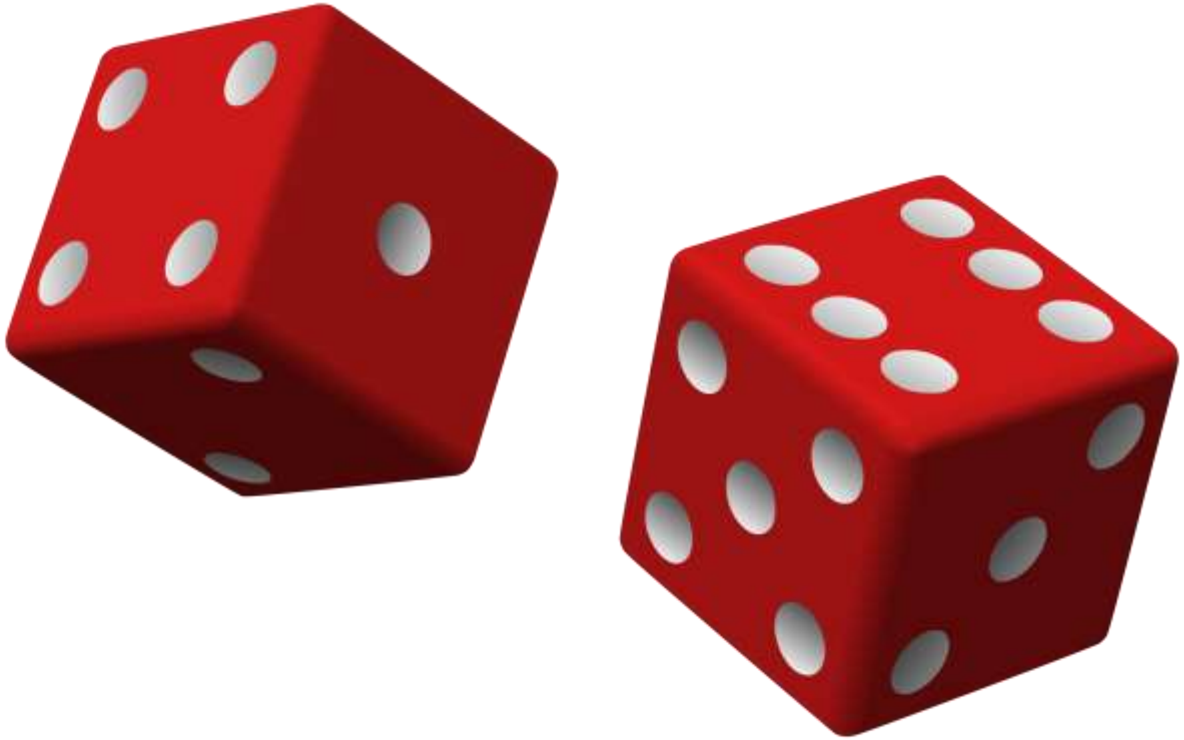
EXPLORATION VS EXPLOITATION DILEMMA

- How to design query algorithms?
  - Exploitation: make best decision based on currently available information
  - Exploration: gather more information



RANDOM QUERY

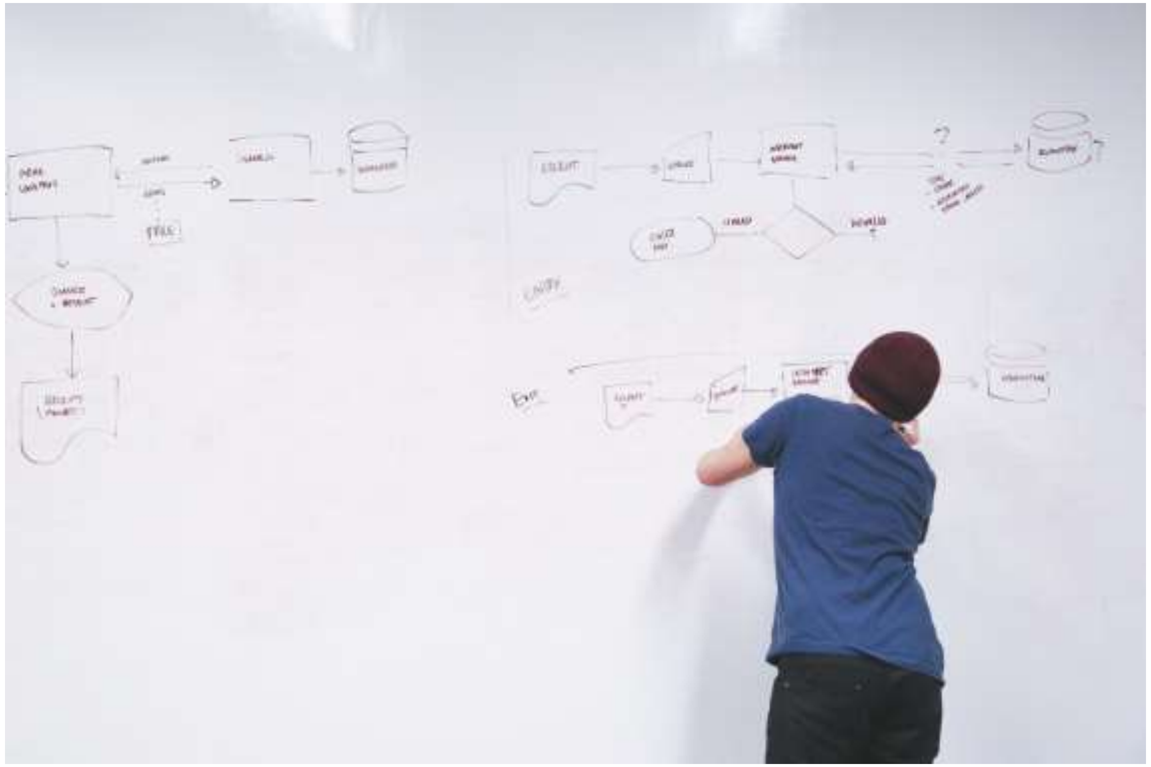
- Select instance to label randomly
- Good starting point
- Baseline algorithm : Compare other algorithms against random query



DETERMINISTIC QUERY

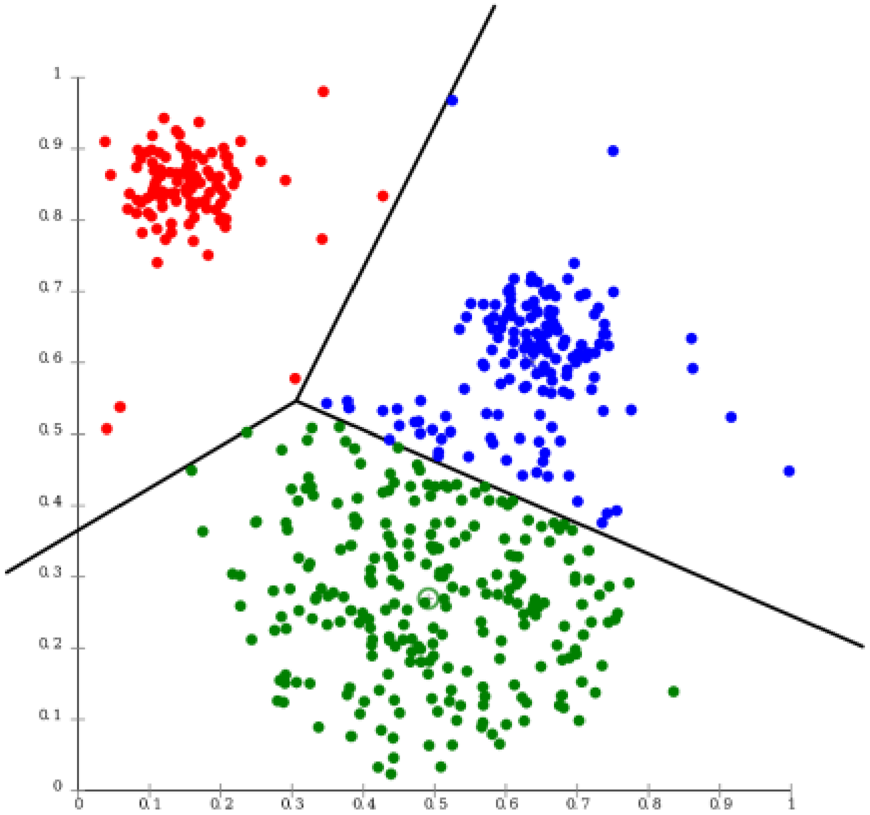
ROSEN empowered by technology

- Use knowledge about features and data to select instances
- Good starting point
- Class discovery



CLUSTER HOMOGENEITY

- Cluster label and unlabeled data
- Are cluster labeled inhomogeneously (by annotator or by classifier)?
- Heuristic algorithm
- Good for exploration



UNCERTAINTY SAMPLING

- Most popular algorithm
- Query the instance which the classifier is most uncertain how to label
- Least confident prediction in terms of prediction probability

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$$

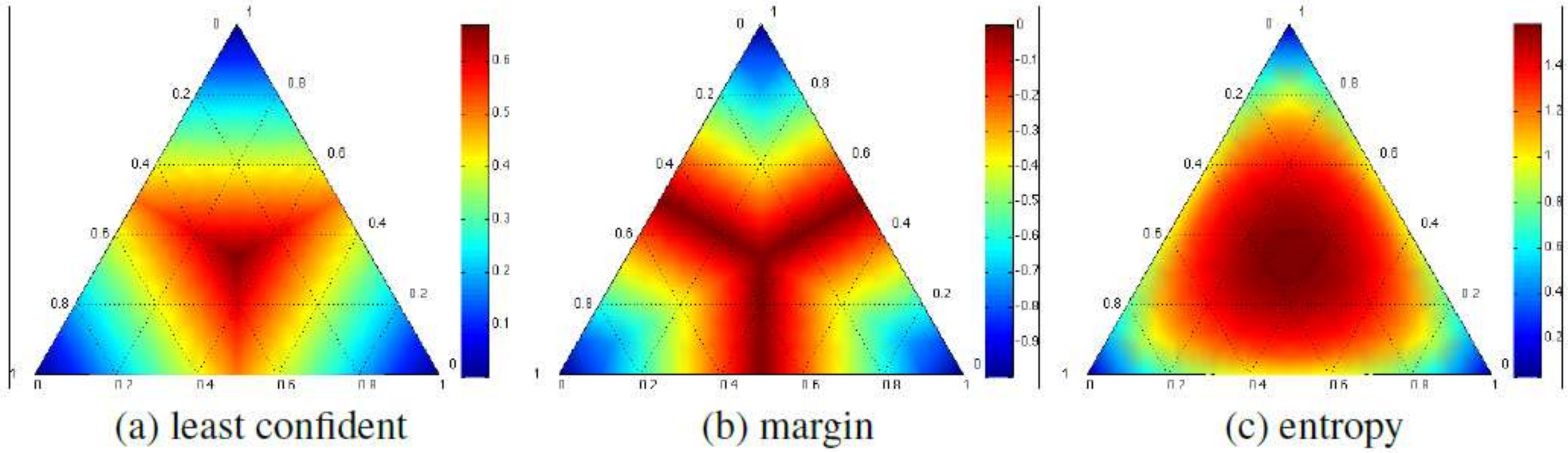
$$\hat{y} = \operatorname{argmax}_y P_\theta(y|x)$$

- Make use of margin between most probable classes or entropy of class prediction probabilities to capture the whole distribution

$$x^*_M = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

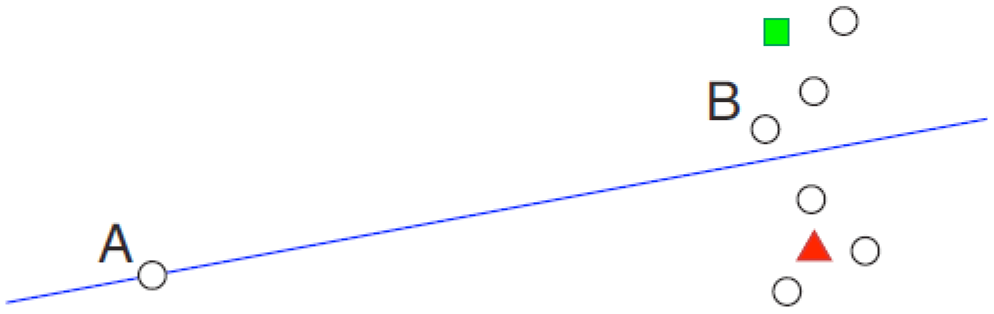
$$x^*_H = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

UNCERTAINTY SAMPLING



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

SHORTCOMINGS OF UNCERTAINTY SAMPLING

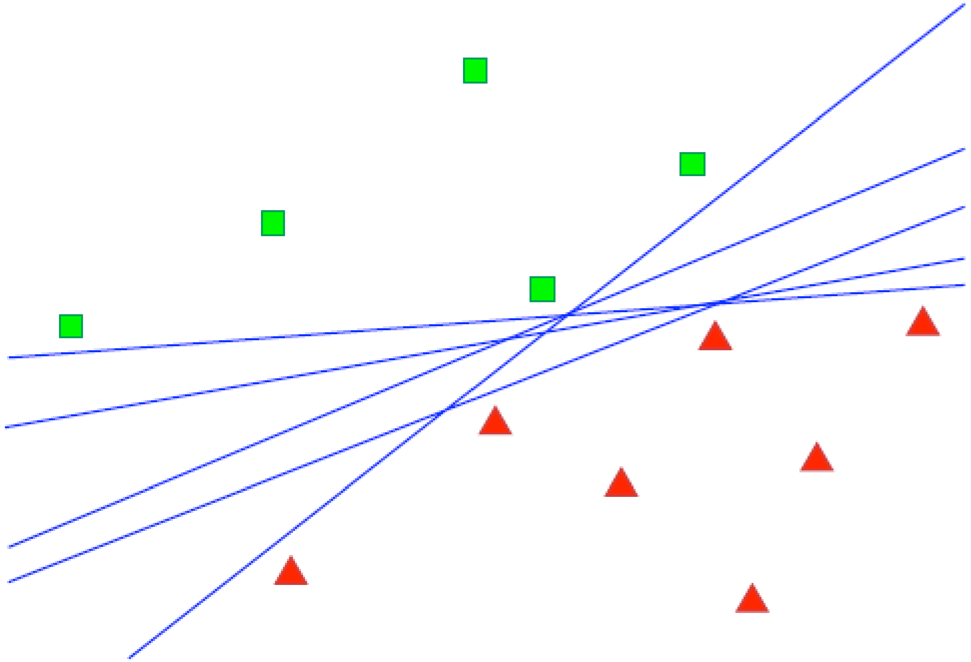


Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

QUERY-BY-COMMITTEE

- Train a set of different classifiers
- Select the instance which the classifiers most degree about
- “Query controversial regions”
- Bagging can be used to construct the committee

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

EXPECTED MODEL CHANGE

- Select the instance which implies the greatest change of the current classifier if its label would be known
- Possible strategy: Which instance has greatest impact on the training gradient
- True labels are unknown so expectation value is used
- Can be computationally expensive

$$x^*_{EGL} = \underset{x}{\operatorname{argmax}} \sum_i P_\theta(y_i|x) \left\| \nabla \ell_\theta(\mathcal{L} \cup \langle x, y_i \rangle) \right\|$$

EXPECTED ERROR REDUCTION

- Select instance which is likely to reduce classification error on unlabeled instances
- Since true labels are unknown expectation values of the loss function for the unlabeled instances is used.
- Very high computational costs
- Sum can be approximated with Monte Carlo sampling to reduce the sum over all unlabeled instances

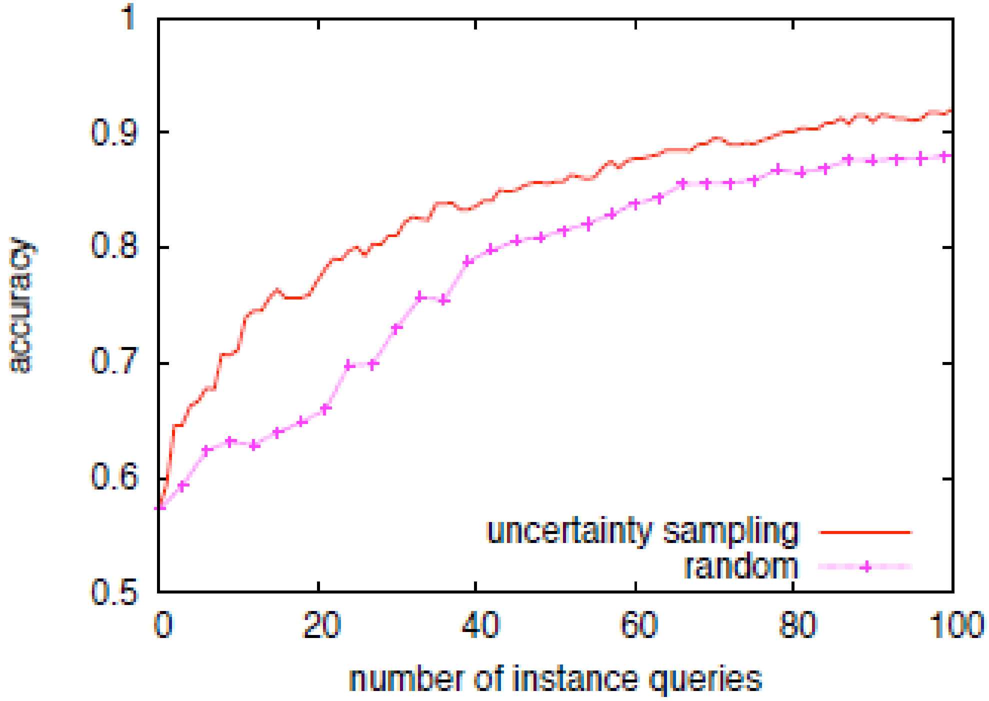
$$x^*_{\log} = \underset{x}{\operatorname{argmin}} \sum_i P_\theta(y_i|x) \left(- \sum_{u=1}^{U} \sum_j P_{\theta+\langle x,y_i\rangle}(y_j|x^{(u)}) \log P_{\theta+\langle x,y_i\rangle}(y_j|x^{(u)}) \right)$$

WHICH ALGORITHM SHALL I USE?

- Algorithm for cold start when no classifier is available (random query, deterministic query)
- Balance exploration and exploitation
- Combine different selection strategies
- Test test test

TESTING QUERY ALGORITHM

- Active learning can be simulated on a fully labeled data set
- Human annotator is simulated by giving the correct label
- Compare performance over number of labeled instances against random query



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

CONVERGENCE

- When to stop labeling?
  - Costs
  - Human annotator notices convergence
  - Stopping criterion

T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (TSNE)

ROSEN
empowered by technology

- Learn low-dimensional embedding of feature vector by minimizing KL between similarities in both spaces
- Different distribution in low-dimensional space to compute similarities
- “Tends to cluster points by their classes”
- Van der Maaten, JMLR 2008

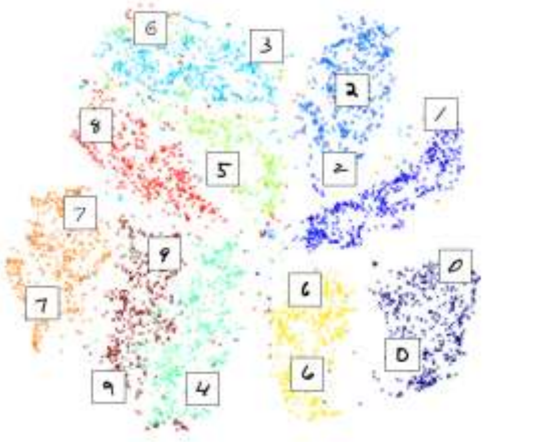
$$p_{j|i} = \frac{\exp\left(-d(\boldsymbol{x}_i, \boldsymbol{x}_j)/(2\sigma_i^2)\right)}{\sum_{i \neq k} \exp\left(-d(\boldsymbol{x}_i, \boldsymbol{x}_k)/(2\sigma_i^2)\right)}, \quad p_{i|i} = 0,$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

$$q_{ij} = \frac{(1 + ||\boldsymbol{y}_i - \boldsymbol{y}_j)||^2)^{-1}}{\sum_{k \neq l}(1 + ||\boldsymbol{y}_k - \boldsymbol{y}_l)||^2)^{-1}},$$

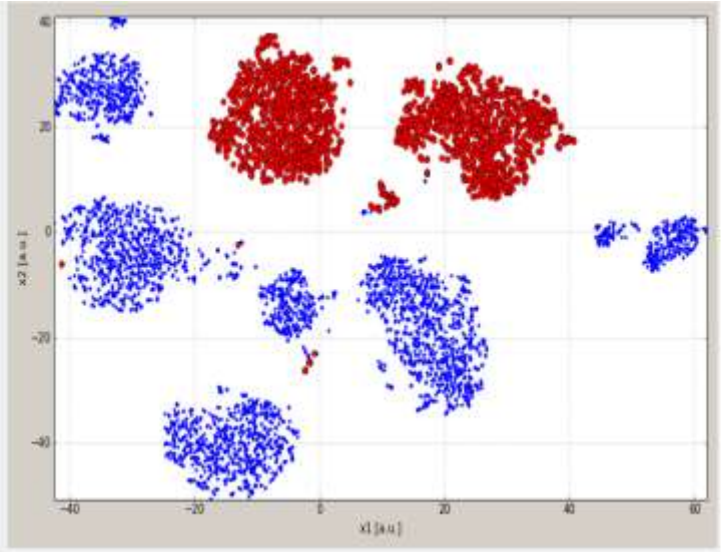
$$KL(P|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

MNIST dataset – Two-dimensional embedding of 70,000 handwritten digits with t-SNE



STRATEGIES FOR MEASURING CONVERGENCE

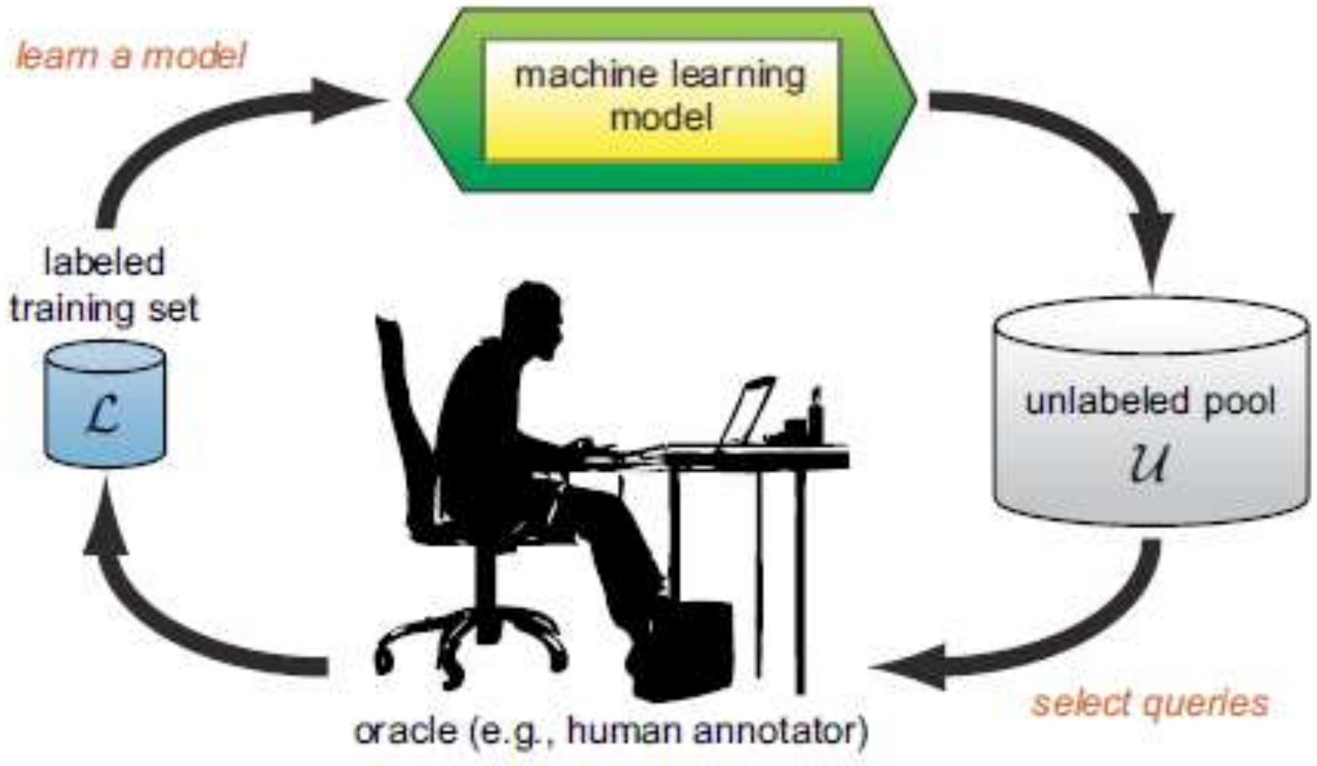
- Color TSNE visualization by classifier decisions and manual labels for the human annotator
- Track the changes of the classifier predictions on unlabeled data
- Cluster consistency



CONCLUSION

- Active learning can help you if you have a huge amount of unlabeled data and humans can provide ground truth
- There is no best query algorithm
- There is no best way to measure convergence

- Problems to consider:
  - Noisy human annotators
  - Class discovery
  - Training speed of ml algorithms



Source: Settles, Burr (2010), "Active Learning Literature Survey", Computer Sciences Technical Report 1648. University of Wisconsin–Madison

THANK YOU FOR JOINING THIS PRESENTATION.

ROSEN
empowered by technology