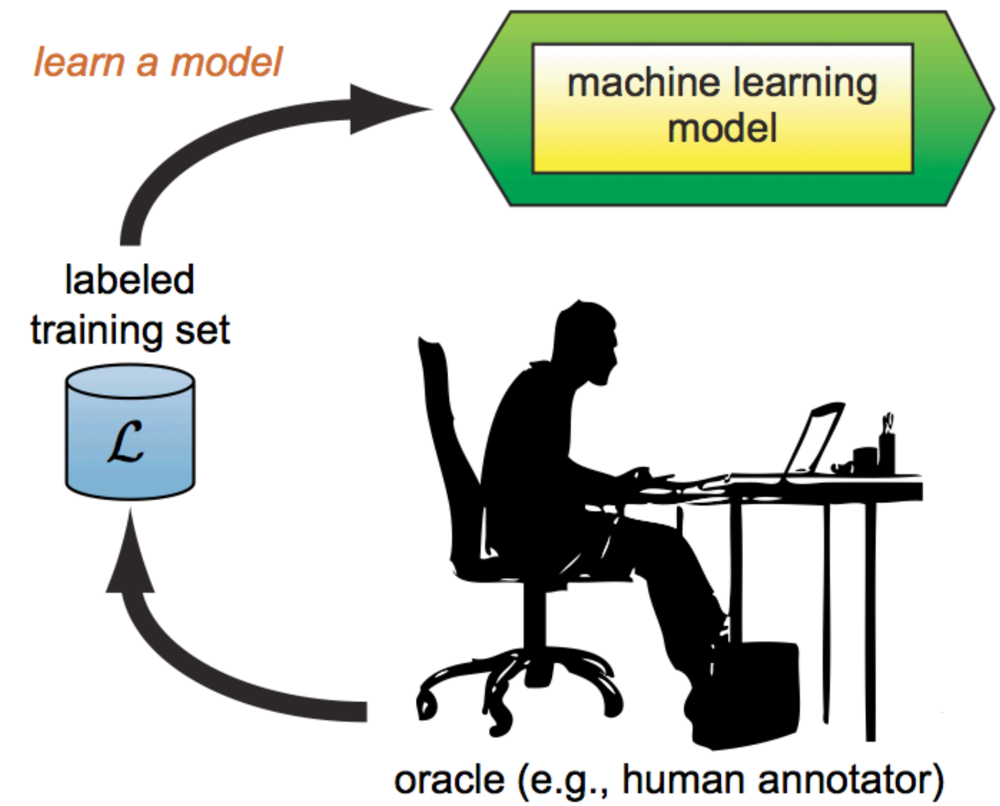
Active Learning

Machine Learning 10-601B

Batch/Passive Learning

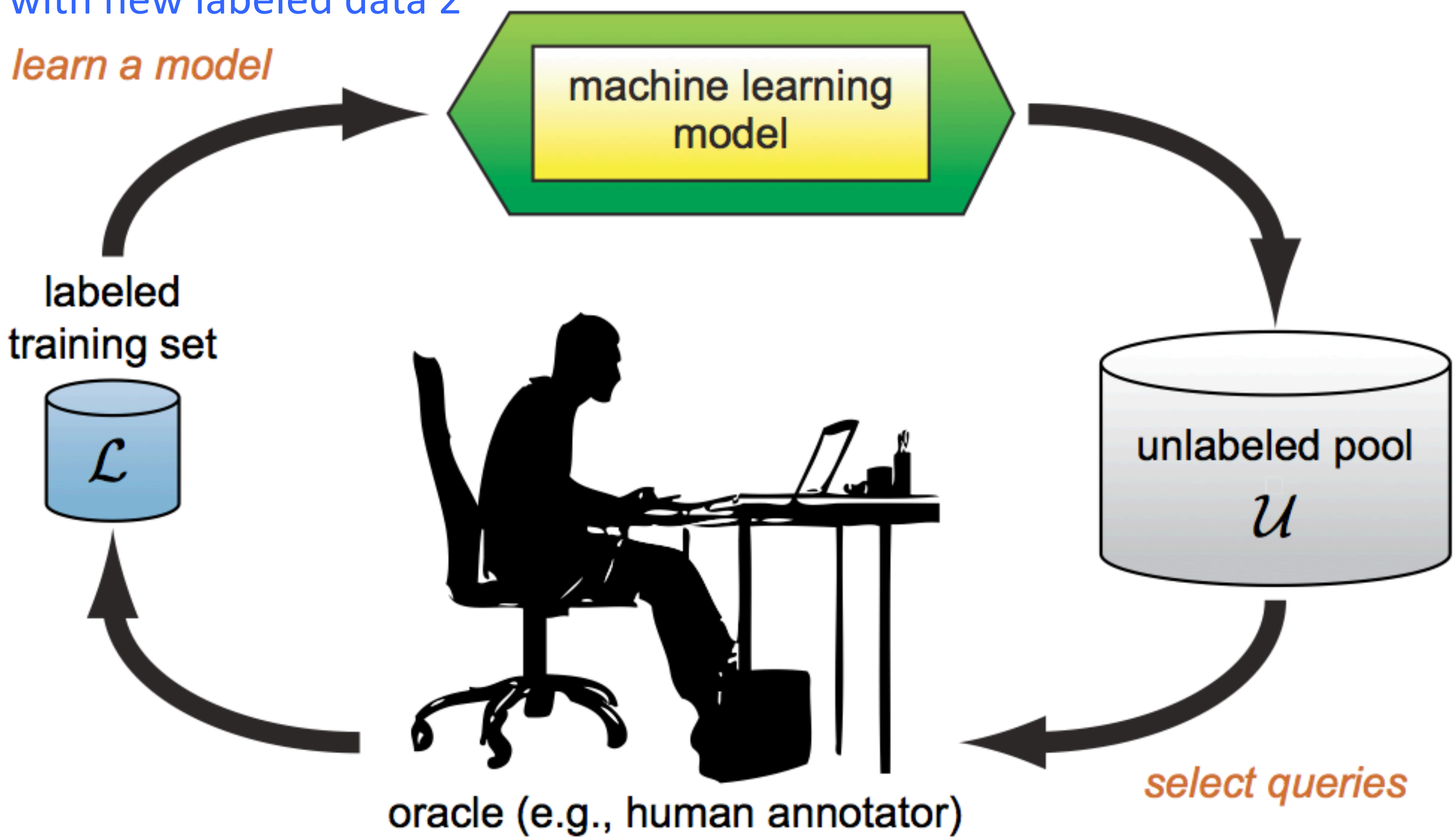
- Training data are collected at once and available to learner as a batch



Active Learning

Update with new labeled data 1

Update with new labeled data 2



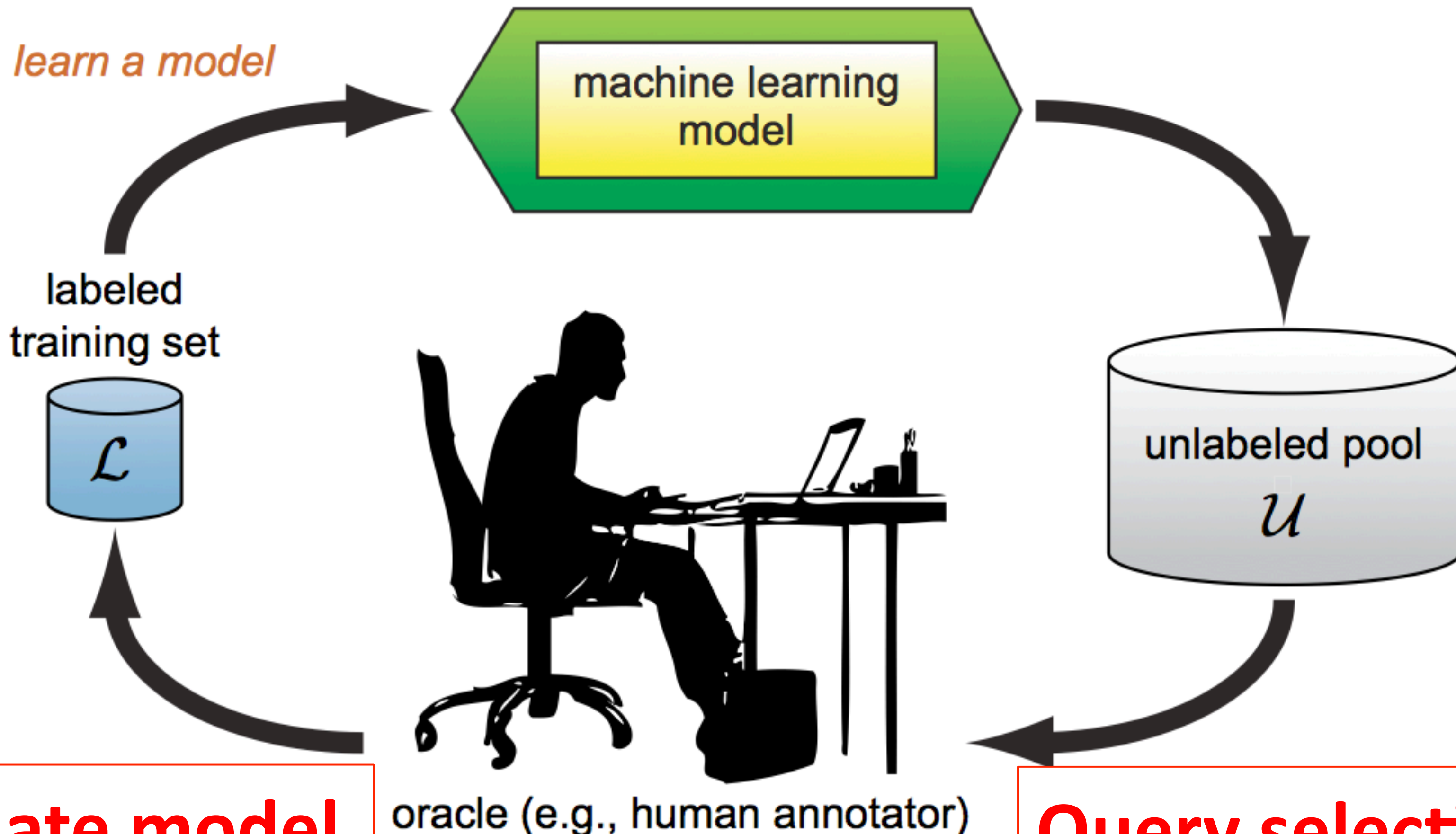
Request a new label 1

Request a new label 2

Why Active Learning?

- Want to collect best data at minimal cost
  - Collect more useful data than simply more data (quality over quantity)
  - Data collection may be expensive
    - Labeled data are more expensive and scarce than unlabeled data
      - Labeling speech data, documents, images by humans
    - Cost of time and materials for an experiment

Active Learning



Update model with new data

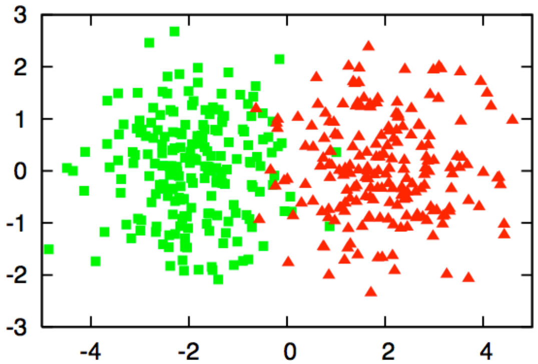
Query selection strategy

Pool Based Sampling

- Assume a small set of labeled data *L*, a large set of unlabeled data U
- Select from the pool of unlabeled data U, the most promising instances to request labels
  - Evaluate all unlabeled instances to select the best query

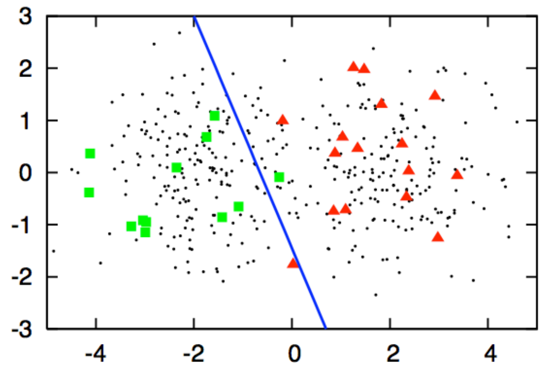
Pool Based Learning

Data space



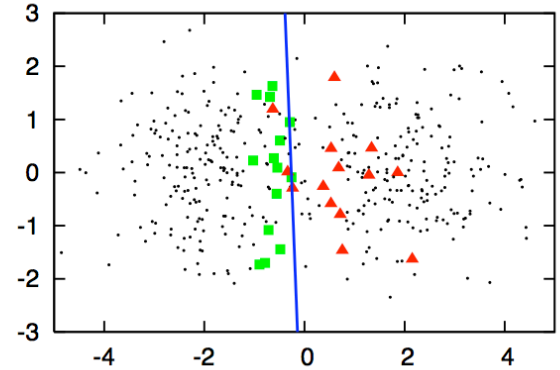
400 samples from
two class Gaussians

Batch learning



Logistic regression
trained with 30 labeled
randomly drawn
instances

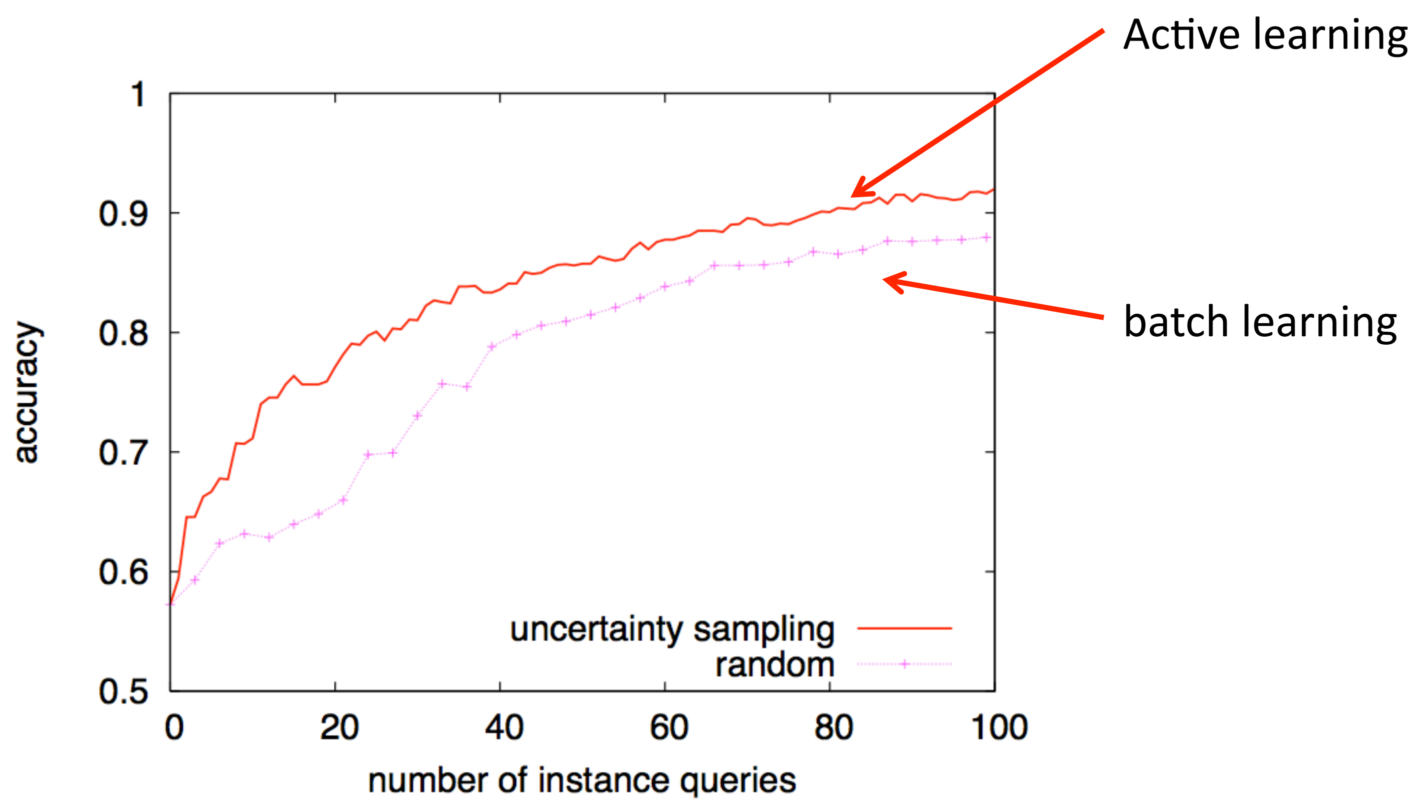
Active learning



A logistic regression
model trained with 30
actively queried
instances using
uncertainty sampling.
90% accuracy, near
Bayes optimal decision
boundary

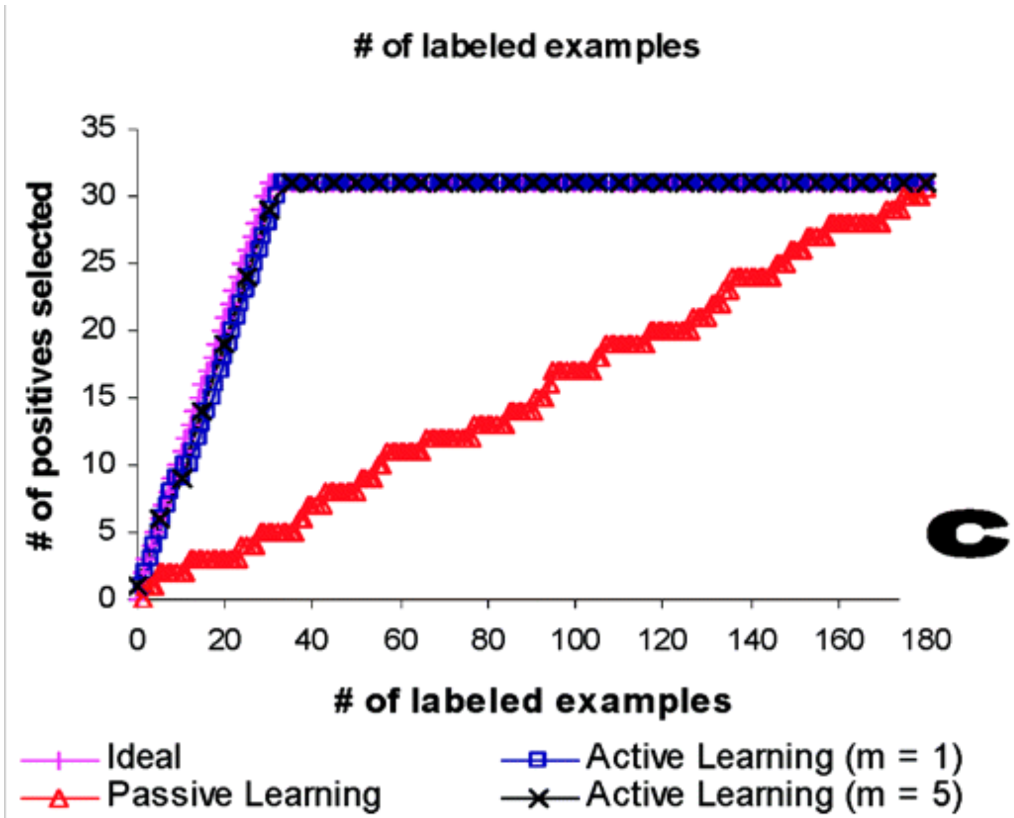
Example: Document Classification

- Logistic regression for classifying Hockey vs Baseball documents from 20 newsgroup corpus of 2000 Usenet documents



Example: Gene expression and Cancer classification

- Active learning for SVM takes 31 points to achieve same accuracy as passive/batch learning with 174



Liu 2004

Selecting Instances for Labeling

- Challenges in active learning: Query strategy!
  - how to evaluate the informativeness of samples to select the most informative samples for labeling
    - Uncertainty sampling
    - Query by committee
    - Expected model changes

Uncertainty Sampling: Least Confident Sample

- Select the instance with the least confident prediction by the current probabilistic classifier $P_\theta(y|x)$

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

where $\hat{y} = \mathrm{argmax}_y\, P_\theta(y|x)$ is the predicted class label by the current estimate of the classifier

- For two-class classification, this selects samples with class probabilities near 0.5

- Does not extend well to multi-class classification

Uncertainty Sampling: Entropy

- Use entropy as a measure of uncertainty in prediction to select query

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

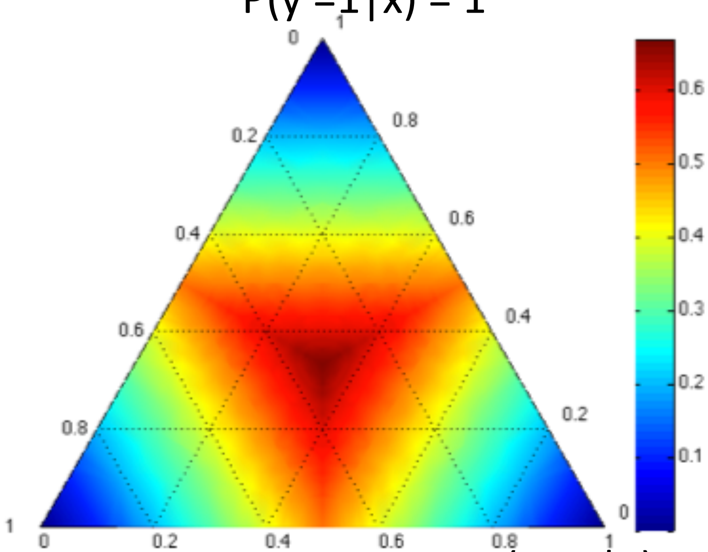
  the summation is over all possible class labels

- Select an instance with the highest uncertainty measured by entropy

Least Confident vs Entropy

- The simplex of P(y|x) for 3 class classification
  - The middle of the simplex: the largest uncertainty
  - Corners of the simplex: the lowest uncertainty

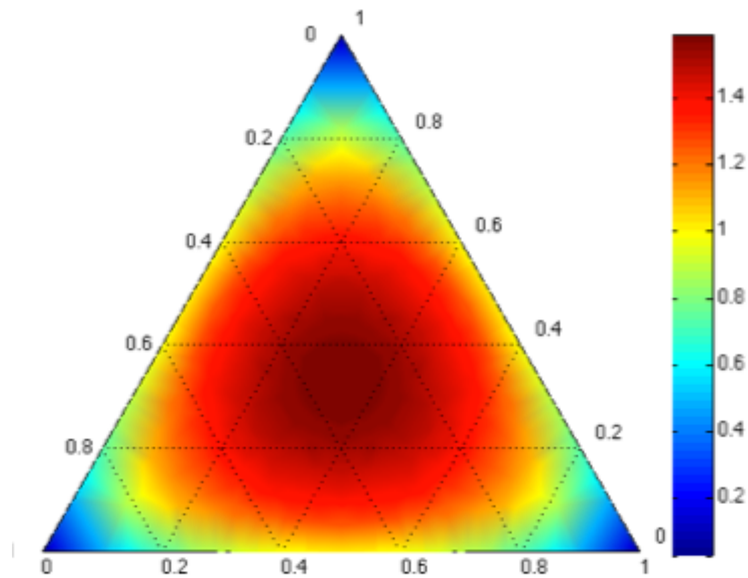
P(y =1|x) = 1



P(y =2|x) = 1

P(y =3|x) = 1

Least confident



Entropy

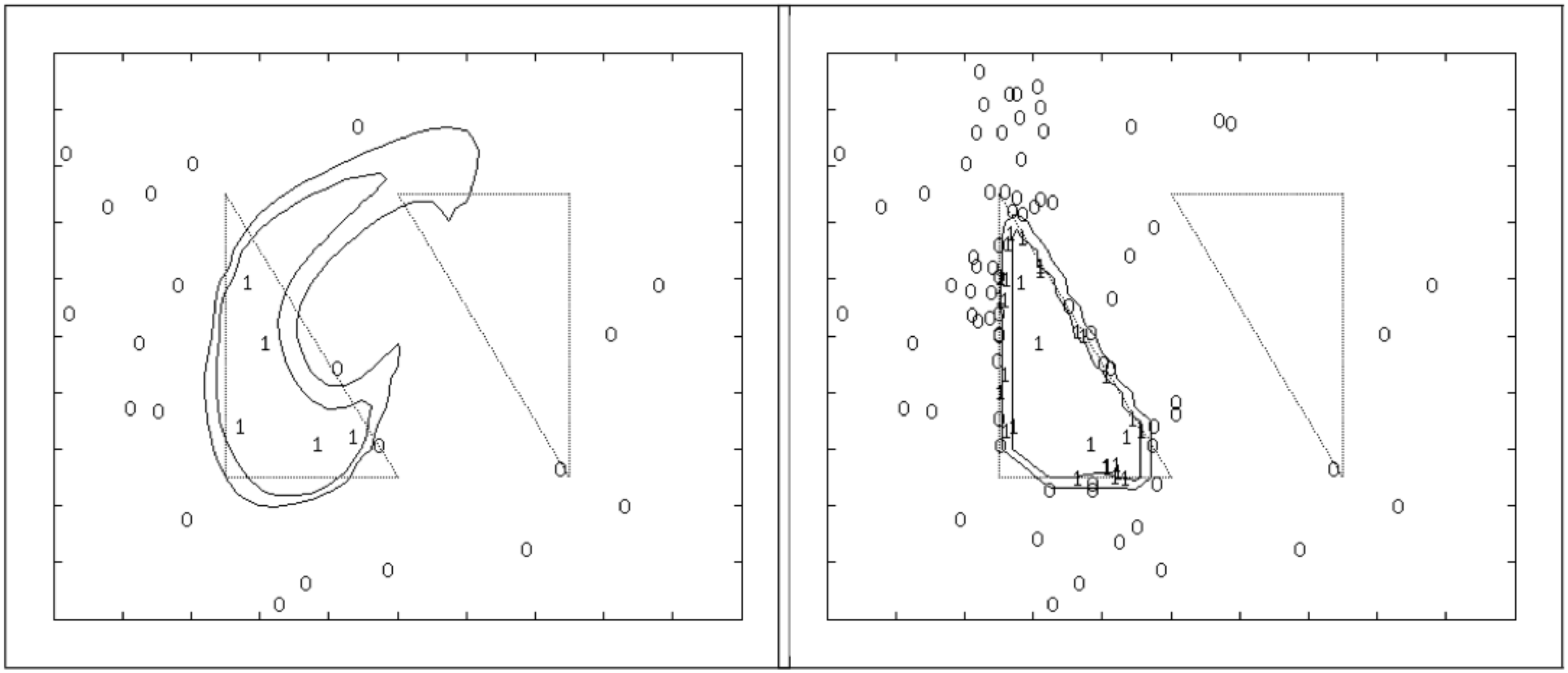
Simple and Widely Used

- text classification
  - Lewis & Gale ICML'94;
- POS tagging
  - Dagan & Engelson, ICML'95; Ringger et al., ACL'07
- disambiguation
  - Fujii et al., CL'98;
- parsing
  - Hwa, CL' 04;
- information extraction
  - Scheffer et al., CAIDA'01; Se0les & Craven, EMNLP'08
- word segmentation
  - Sassano, ACL'02
- speech recognition
  - Tur et al., SC'05
- transliteration
  - Kuo et al., ACL'06
- translation
  - Haffari et al., NAACL'09

Problems with Uncertainty Sampling

Initial random sample misses the right triangle

Neural net uncertainty sampling only queries the left side



Cohn et al., ML 1994

Problems with Uncertainty Sampling

- Plain uncertainty sampling only uses the confidence of a single classifier
  - Sometimes called a point estimate for parametric models
  - This classifier can become overly confident about instances it really knows nothing about!

- Instead let's consider a different notion of uncertainty, about the classifier itself

Query by Committee

- Maintain a committee of classifiers $\mathcal{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$, all of which were trained on labeled data L **Uncertainty among the classifiers**

- Let the committee vote for the labels of unlabeled data

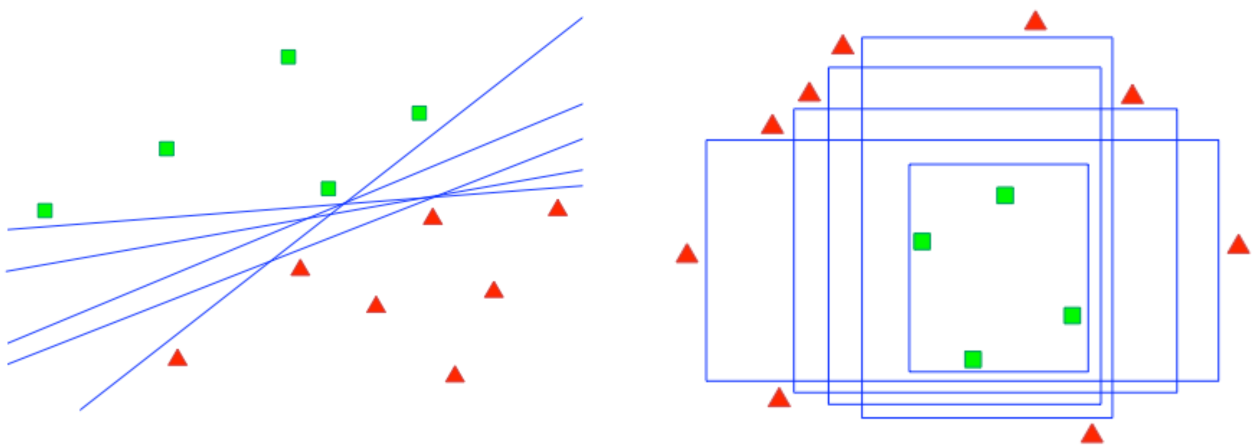
- Select the samples on which the committee disagrees the most
  - Vote entropy: C is # of classifiers in the committee, $V(y_i)$ is the votes from

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

Query by Committee

- Committee consists of classifiers in the same version space (all classifiers consistent with the training data)

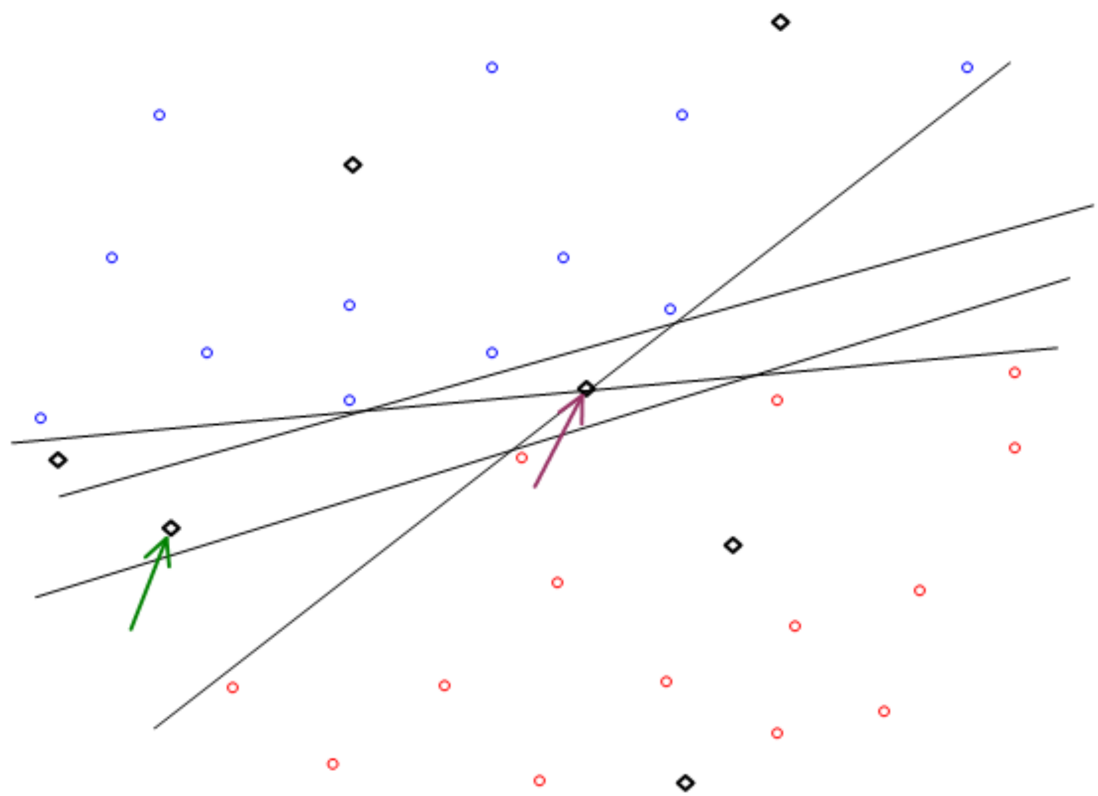
- By selecting the samples that the committee disagrees on, we are trying to reduce the version space



Each of the classifiers is consistent with the training data

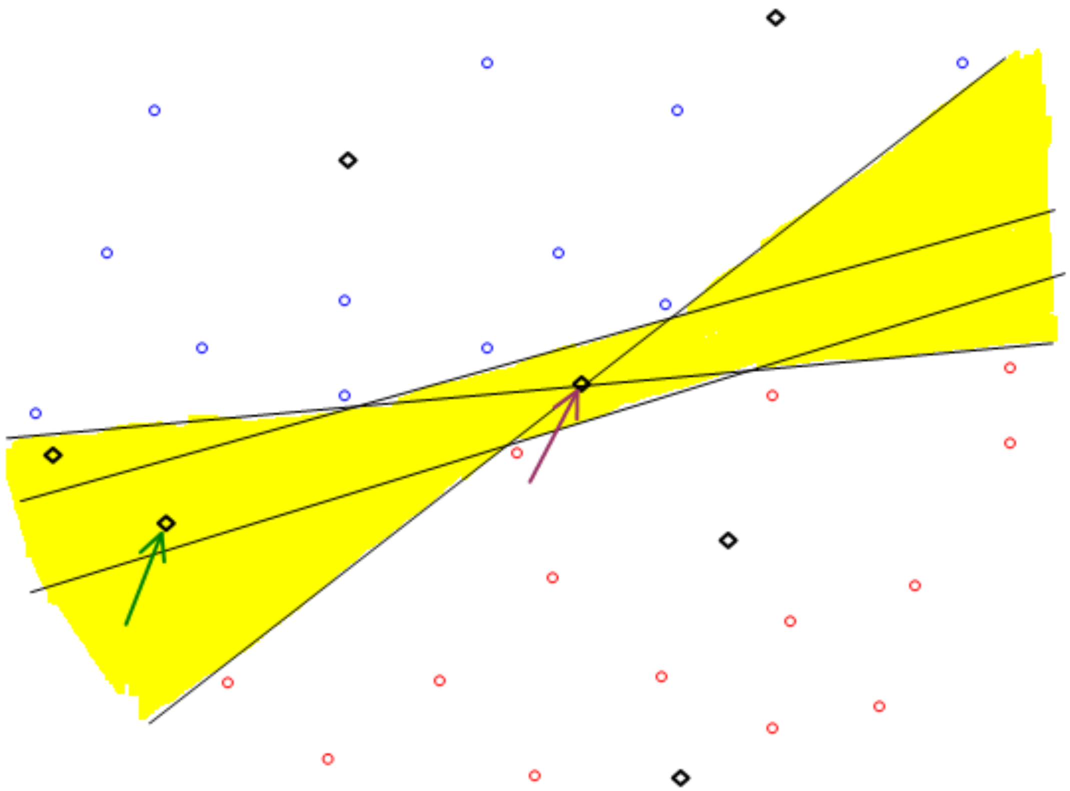
Query by Committee

- Which unlabelled point should you choose?



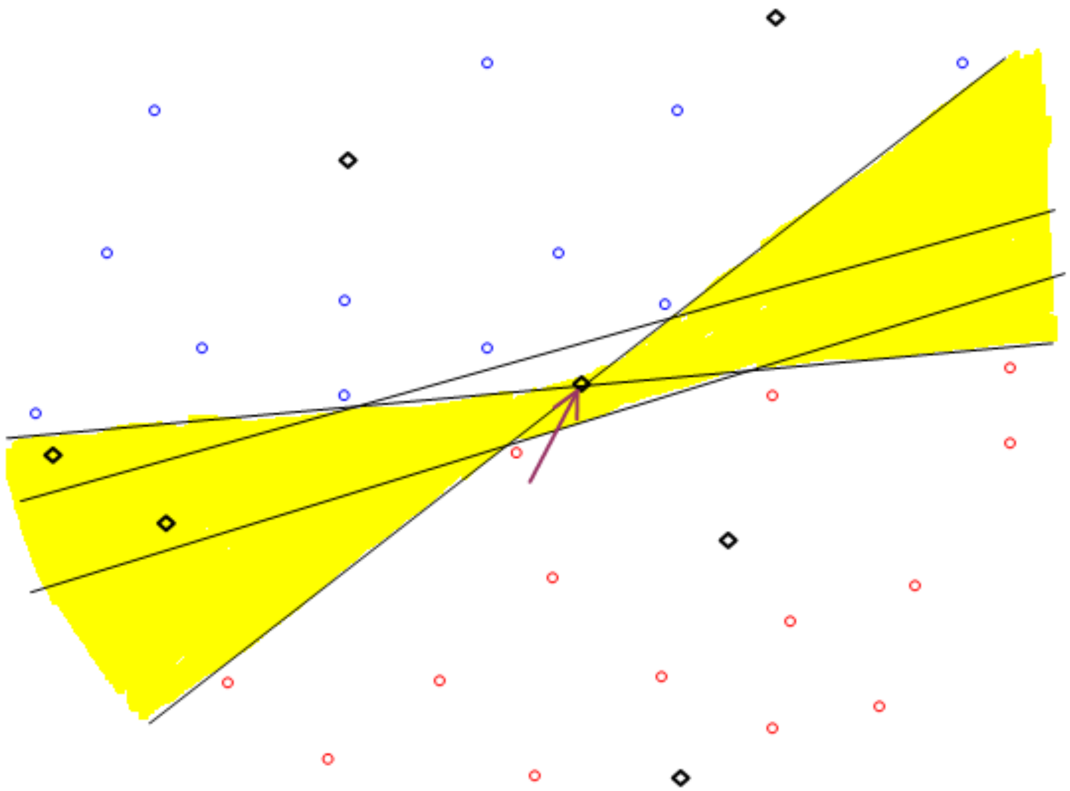
Query by Committee

- Yellow = valid hypotheses



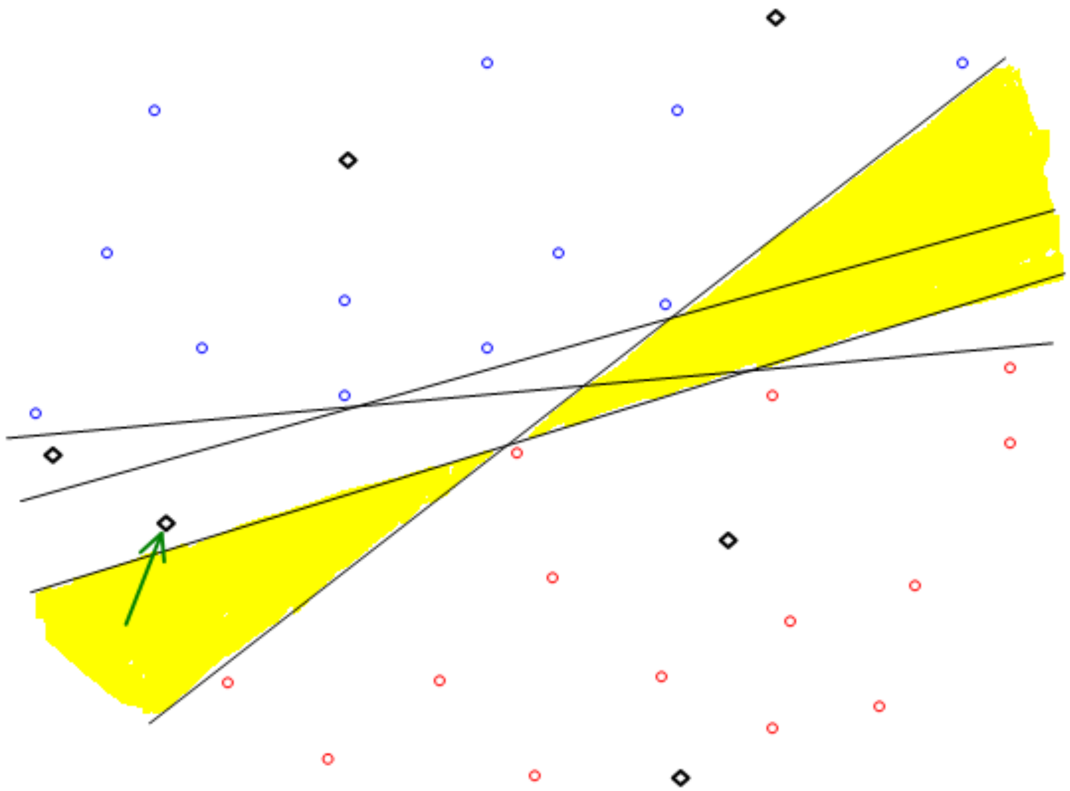
Query by Committee

- Point on max-margin hyperplane does not reduce the number of valid hypotheses by much



Query by Committee

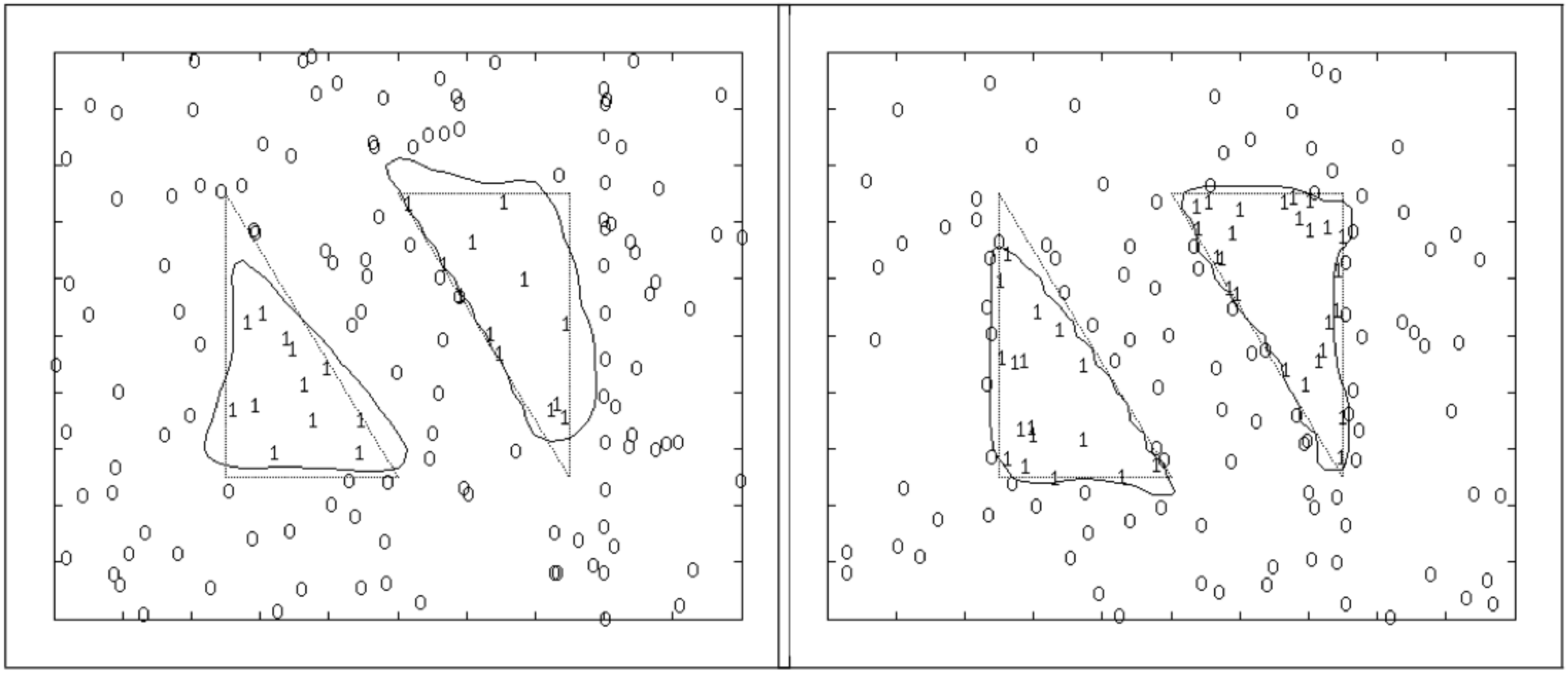
- Queries an example based on the degree of disagreement between committee of classifiers



How to Form a Committee

- Sample models from the posterior distribution of the parameter θ, $P(\theta|L)$
- Standard ensemble methods (bagging, boosting etc.)

Query by Committee



Learned from 150 random samples

Learned from 150 samples selected by query-by-committee method

Expected Model Change

- Select the instance that would induce the greatest change in the model

- Can be applied to any models that involves gradients during training, whereas uncertainty sampling can be applied mostly for probabilistic models

Expected Model Change

- $\nabla \ell_\theta(\mathcal{L})$: gradient of the model given the current estimate of tne parameter

- $\nabla \ell_\theta(\mathcal{L} \cup \langle x, y \rangle)$: Gradient of the model after seeing the query x and the label y

- Since we do not know the label y, we take the expectation with respect to y and select the sample for labeling as

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

- $\|\nabla \ell_\theta(\mathcal{L})\|$ is near zero after training with L, so we approximate

$$\nabla \ell_\theta(\mathcal{L} \cup \langle x, y_i \rangle) \approx \nabla \ell_\theta(\langle x, y_i \rangle)$$

Active vs Semi-supervised Learning

- both try to attack the same problem: making the most of unlabeled data U

Uncertainty sampling

query instances the model is least confident about

Expectation-maximization

Propagate confident labelings among unlabeled data

Query by committee

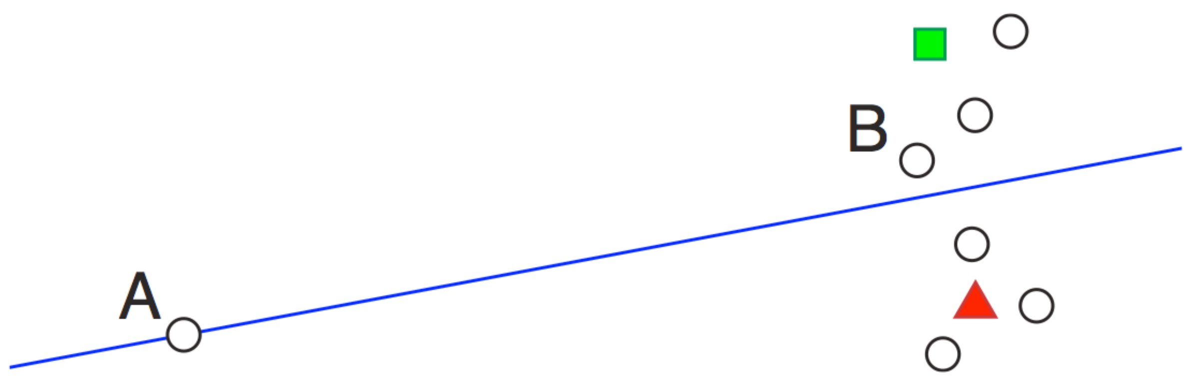
use ensembles to rapidly reduce the version space

Co-training

Use ensembles with multiple views to constrain the version space w.r.t. unlabeled data

Issues with Outlier

- A sample may be selected for labeling simply because it is an outlier



  - Data A is an outlier
  - Data B is more likely to improve the classifier if labeled

Handling Outlier Issues

- Density-weighted sampling
  - Takes into account the underlying distribution in x
  - Informative instance x is the representative sample from the full sample space

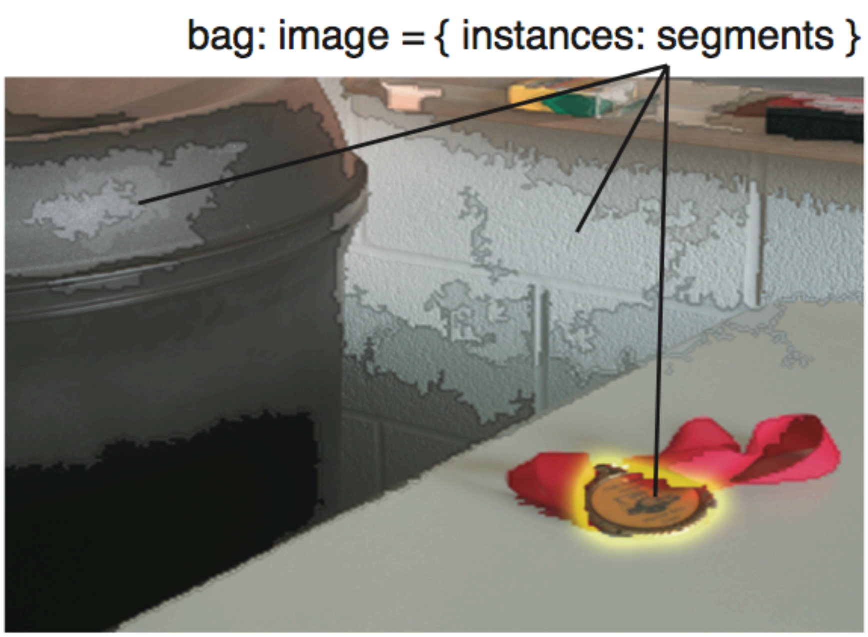
$$x^*_{ID} = \underset{x}{\operatorname{argmax}}\, \phi_A(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} \operatorname{sim}(x, x^{(u)}) \right)^{\beta}$$

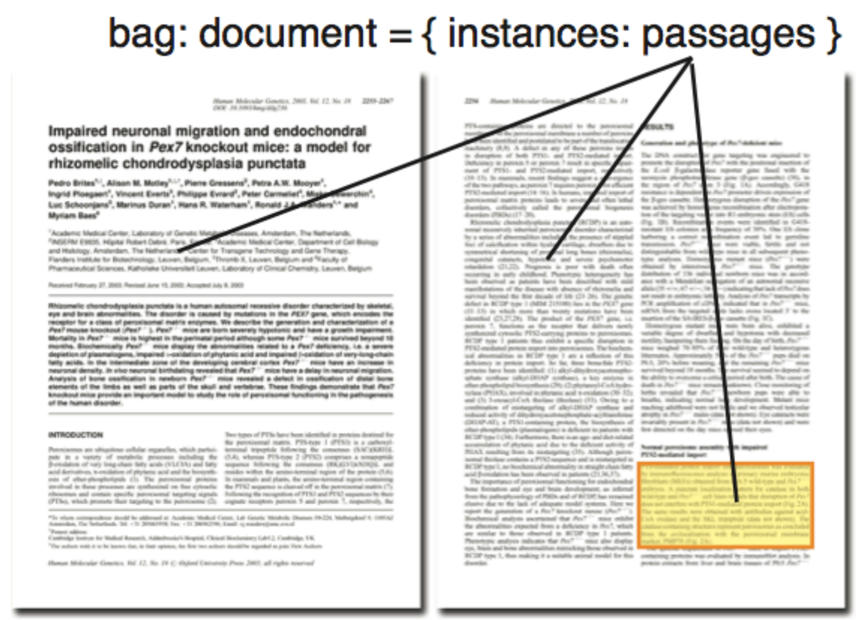
Informativeness measure from the query strategy

- Average similarity to other instances in the input distribution using unlabeled data U
- β: user-determined weight for the amount of outlier control

More Applications of Active Learning

- Bag-of-words for document classification
- bag-of-segments for image classification
- Request labelings for instances in a "bag"





Summary

- Active learning vs passive learning
- Query strategies
  - Uncertainty sampling
  - Query by committee method