Active Learning

October 15, 2009

Reading the Web:
Advanced Statistical Language Processing
ML 10-709

Burr Settles

Machine Learning Department

Carnegie Mellon University

Thought Experiment

- suppose you're the leader of an Earth convoy sent to colonize planet Mars

people who ate the round Martian fruits found them *tasty!*

people who ate the spiked Martian fruits ***died!***

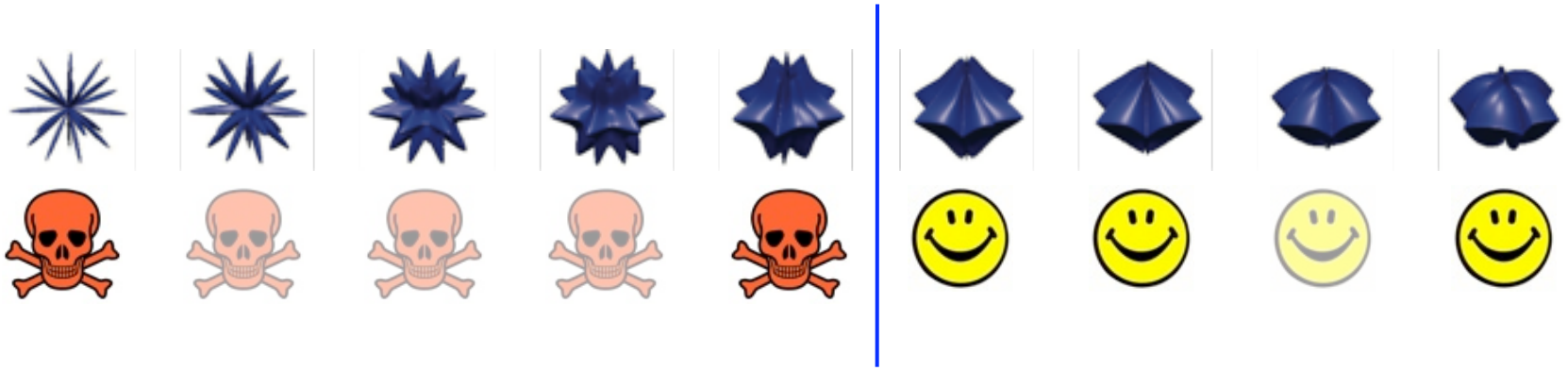
Poison vs. Yummy Fruits

- *problem*: there's a range of spiky-to-round fruit shapes on Mars:

you need to learn the "threshold" of roundness where the fruits go from poisonous to safe.

and… you need to determine this risking as **few colonists' lives** as possible!

Testing Fruit Safety…



this is just a **binary search**, so…

under the PAC model, assume we need $O(1/\varepsilon)$ i.i.d. instances to train a classifier with error ε.

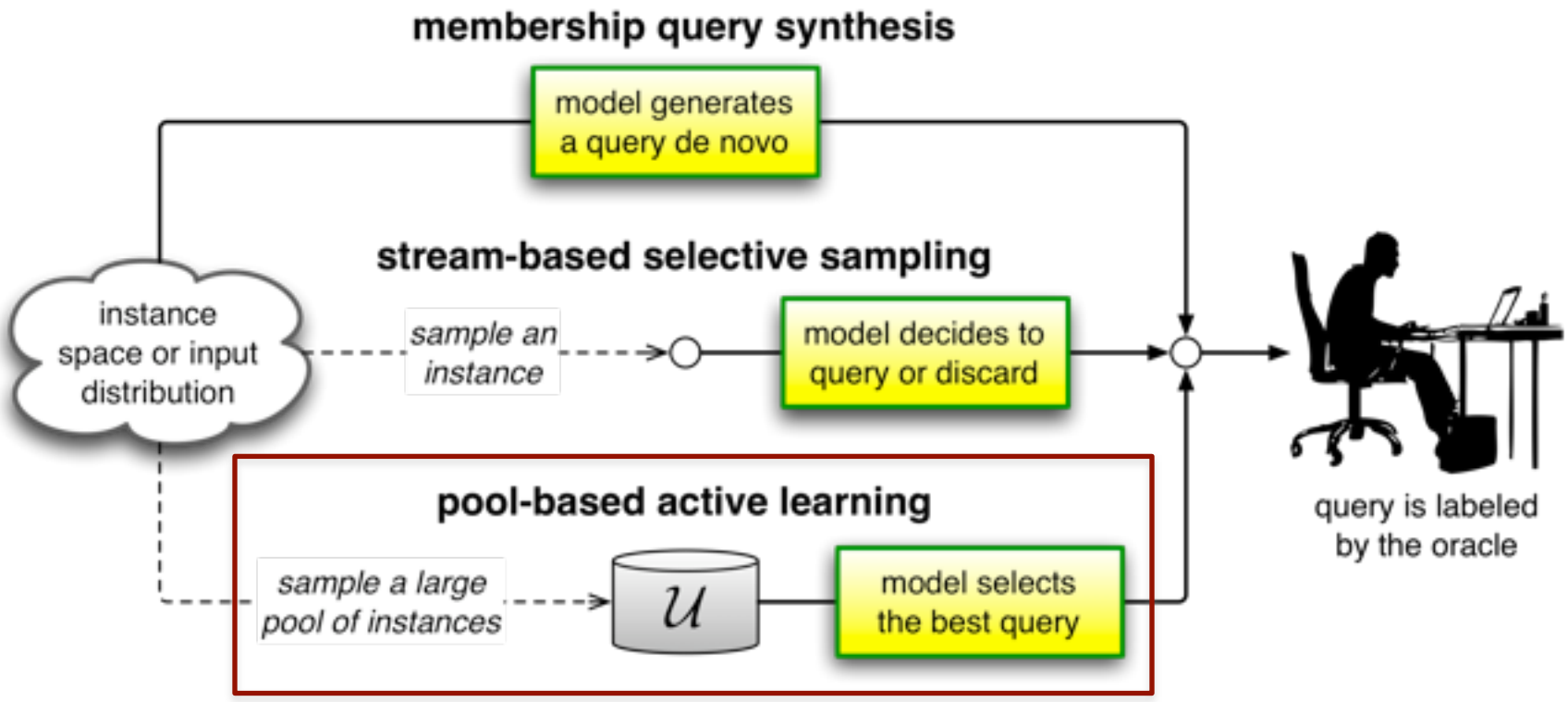
using the binary search approach, we only needed $O(\log_2 1/\varepsilon)$ instances!

Relationship to Active Learning

- **key idea:** the learner can choose training data
  - on Mars: whether a fruit was poisonous/safe
  - *in general*: the true label of some instance

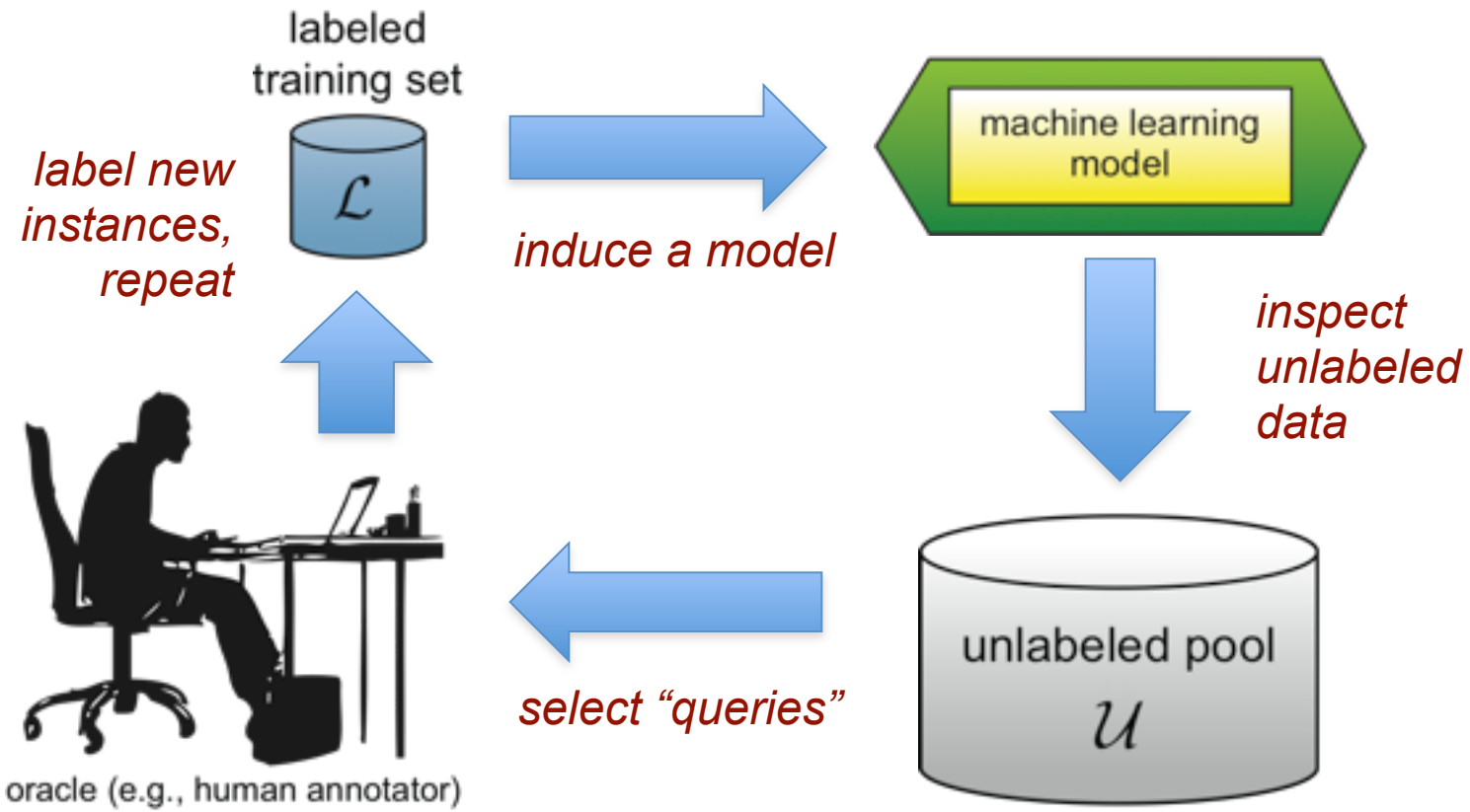
- **goal:** reduce the training costs
  - on Mars: the number of "lives at risk"
  - *in general*: the number of "queries"

Active Learning Scenarios

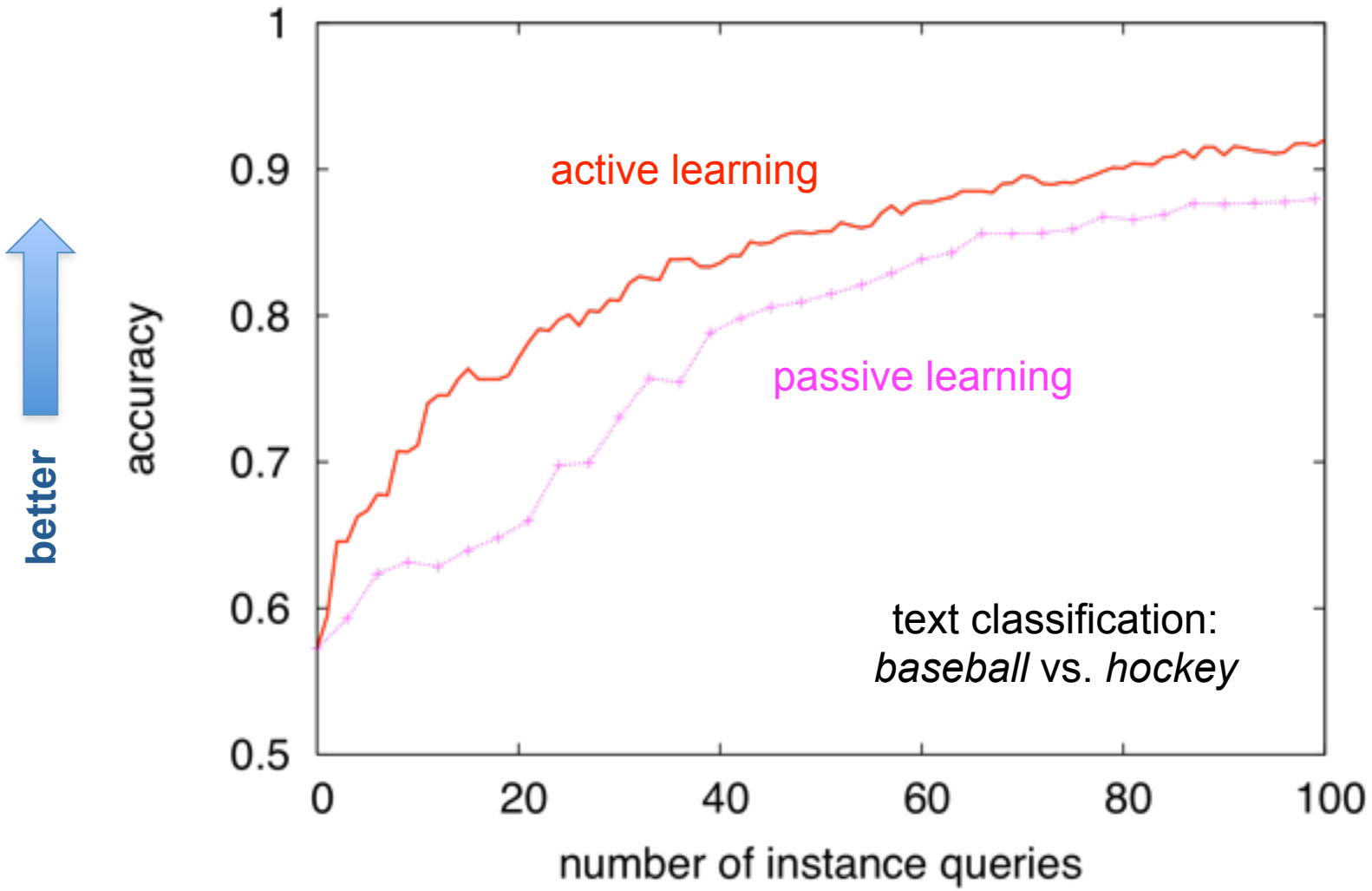


most common in NLP applications

Pool-Based Active Learning Cycle



Learning Curves



Who Uses Active Learning?

Sentiment analysis for blogs; Noisy relabeling
– *Prem Melville*

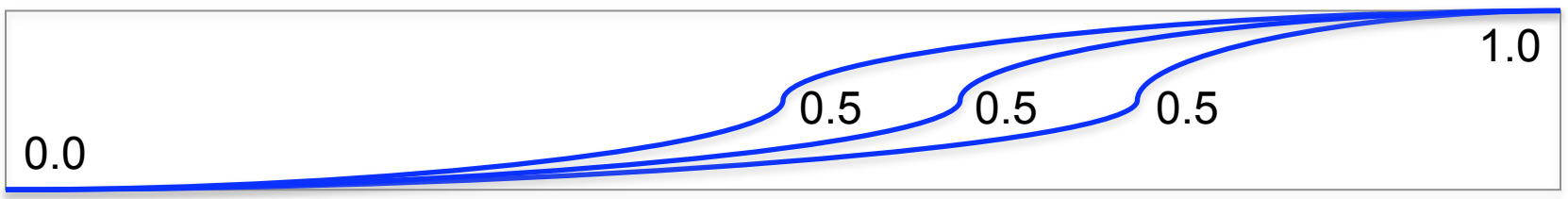
Biomedical NLP & IR; Computer-aided diagnosis
– *Balaji Krishnapuram*

MS Outlook voicemail plug-in [Kapoor et al., IJCAI'07];
"A variety of prototypes that are in use throughout the company." – *Eric Horvitz*

"While I can confirm that we're using active learning in earnest on many problem areas… I really can't provide any more details than that. Sorry to be so opaque!"
– *David Cohn*

How to Select Queries?

- let's try generalizing our binary search method using a *probabilistic* classifier:

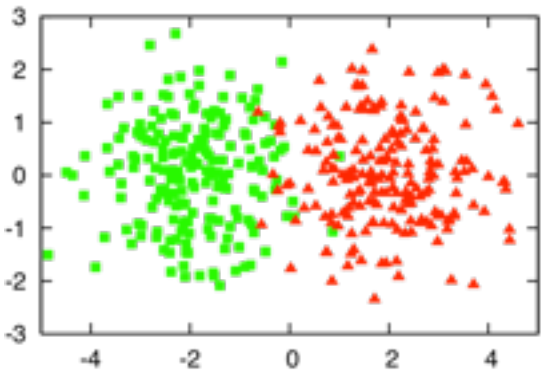


$$P(Y = \text{☺} \mid X)$$

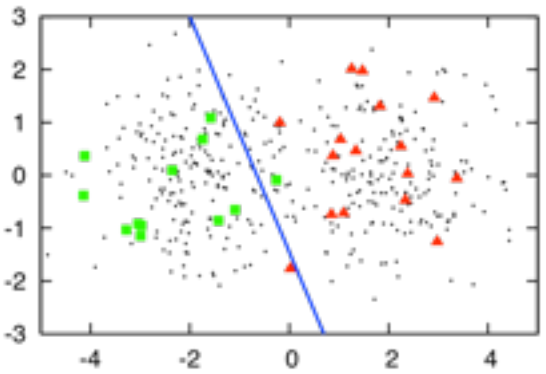
[Lewis & Gale, SIGIR'94]

Uncertainty Sampling

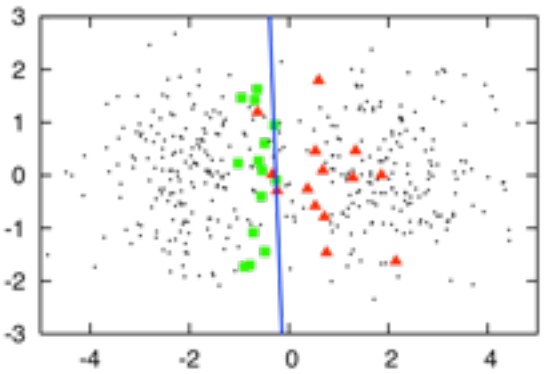
- query instances the learner is *most uncertain* about



400 instances sampled
from 2 class Gaussians



random sampling
30 labeled instances
(accuracy=0.7)



active learning
30 labeled instances
(accuracy=0.9)

Generalizing to Multi-Class Problems

entropy [Dagan & Engelson, ICML'95]

$$\phi_{ENT}(x) = -\sum_y P_\theta(y|x) \log_2 P_\theta(y|x)$$

smallest-margin [Scheffer et al., CAIDA'01]

$$\phi_M(x) = P_\theta(y_1^*|x) - P_\theta(y_2^*|x)$$

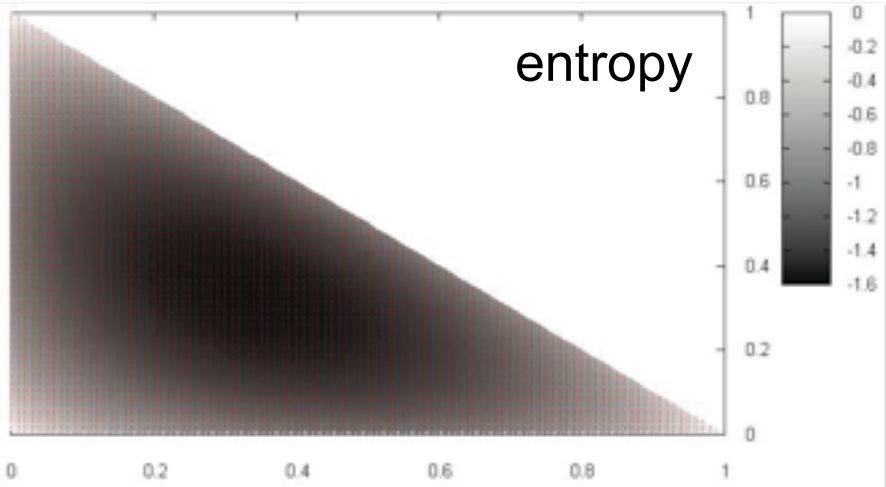
least confident [Culotta & McCallum, AAAI'05]

$$\phi_{LC}(x) = 1 - P_\theta(y^*|x)$$

note: for binary tasks, these are equivalent

[Körner & Wrobel, ECML'06]

Multi-Class Uncertainty Measures



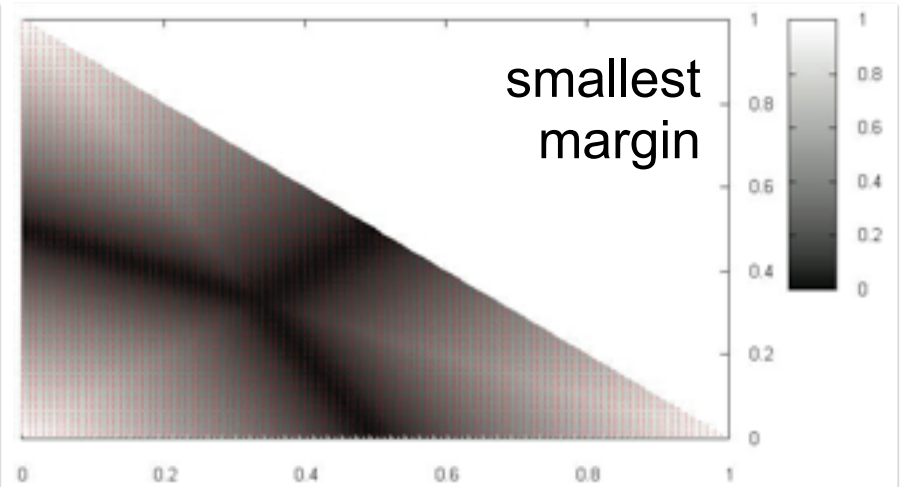
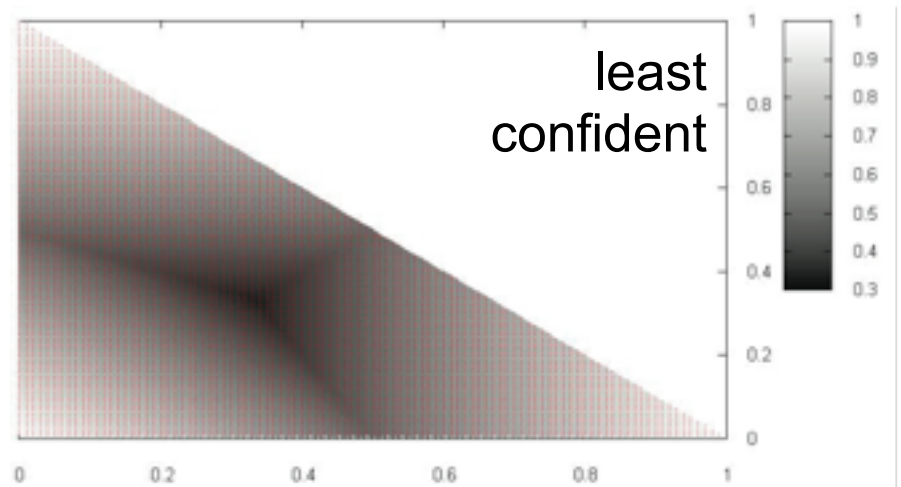


illustration of preferred (darker) posterior distributions in a 3-label classification task



[Seung et al., COLT'92]

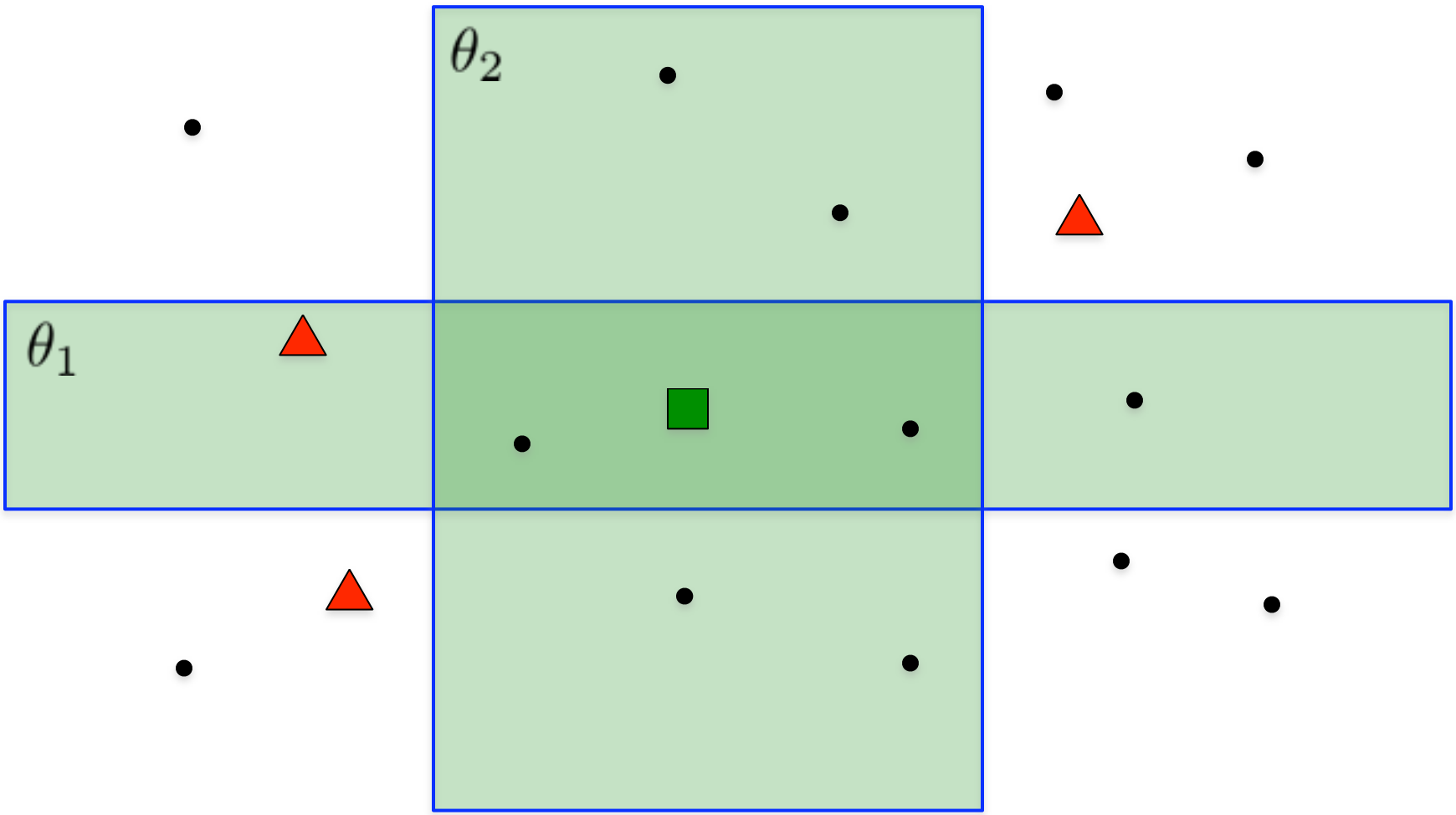
Query-By-Committee (QBC)

- train a committee $\mathcal{C} = \{\theta_1, \theta_2, \ldots, \theta_C\}$ of classifiers on the labeled data in $\mathcal{L}$

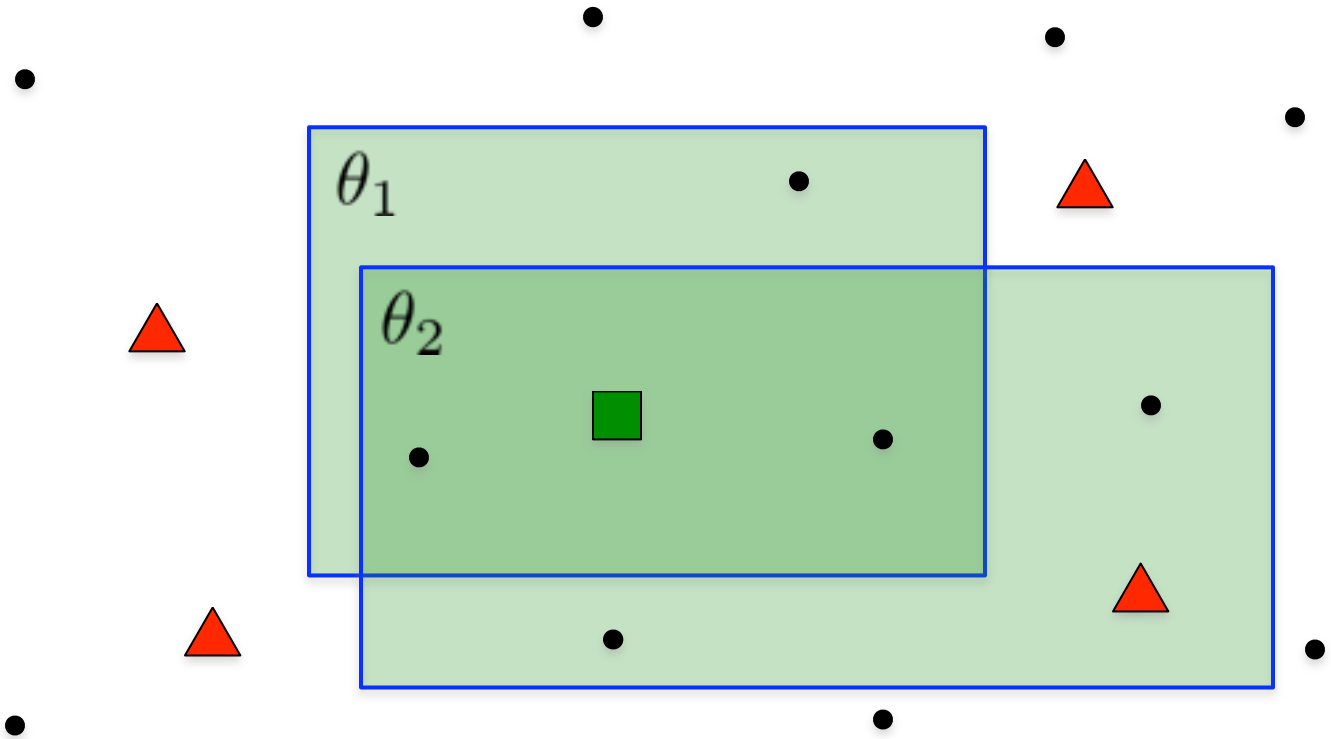
- query instances in $\mathcal{U}$ for which the committee is in most *disagreement*

- **key idea:** reduce the model *version space*
  - expedites search for a model during training

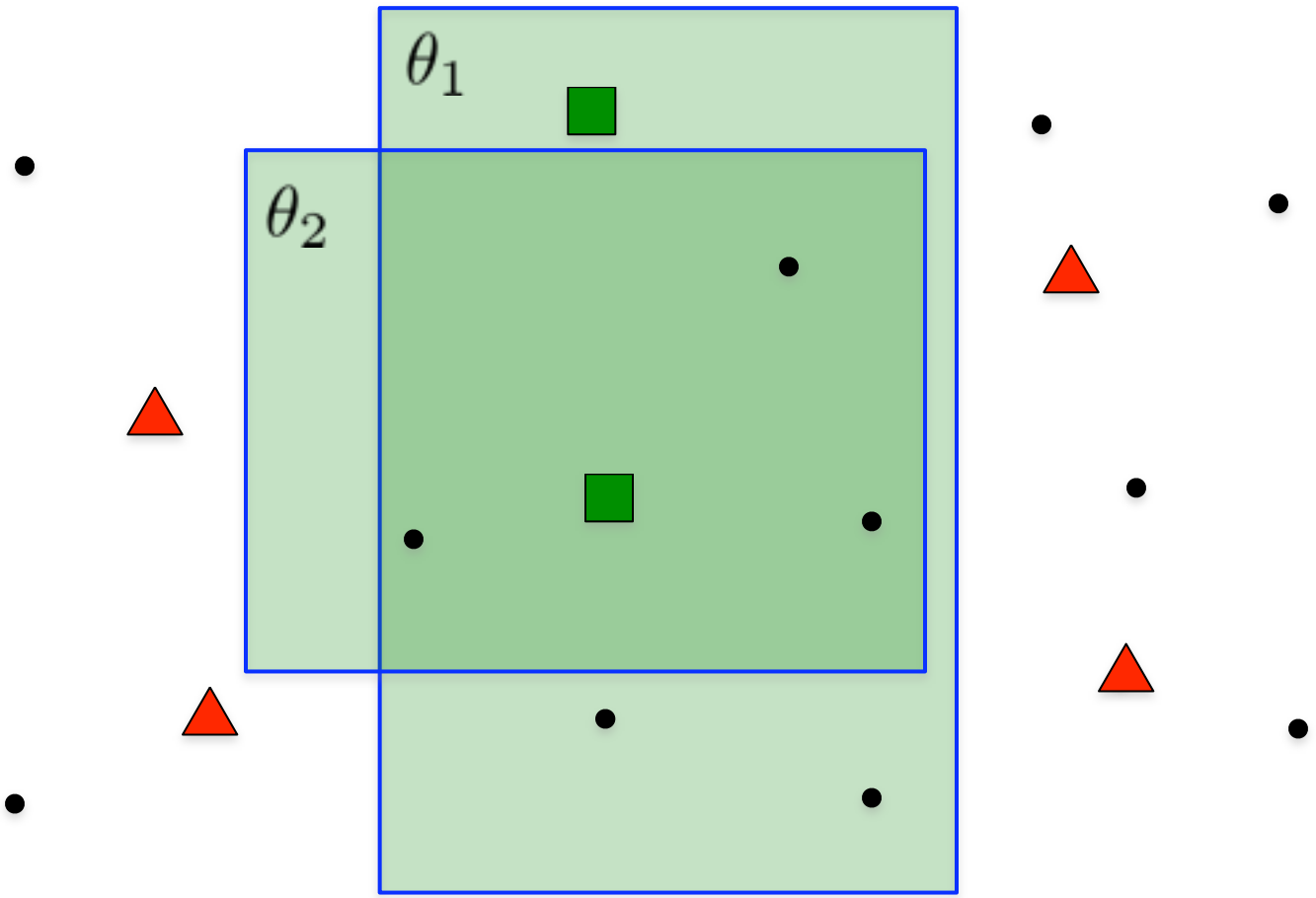
QBC Example



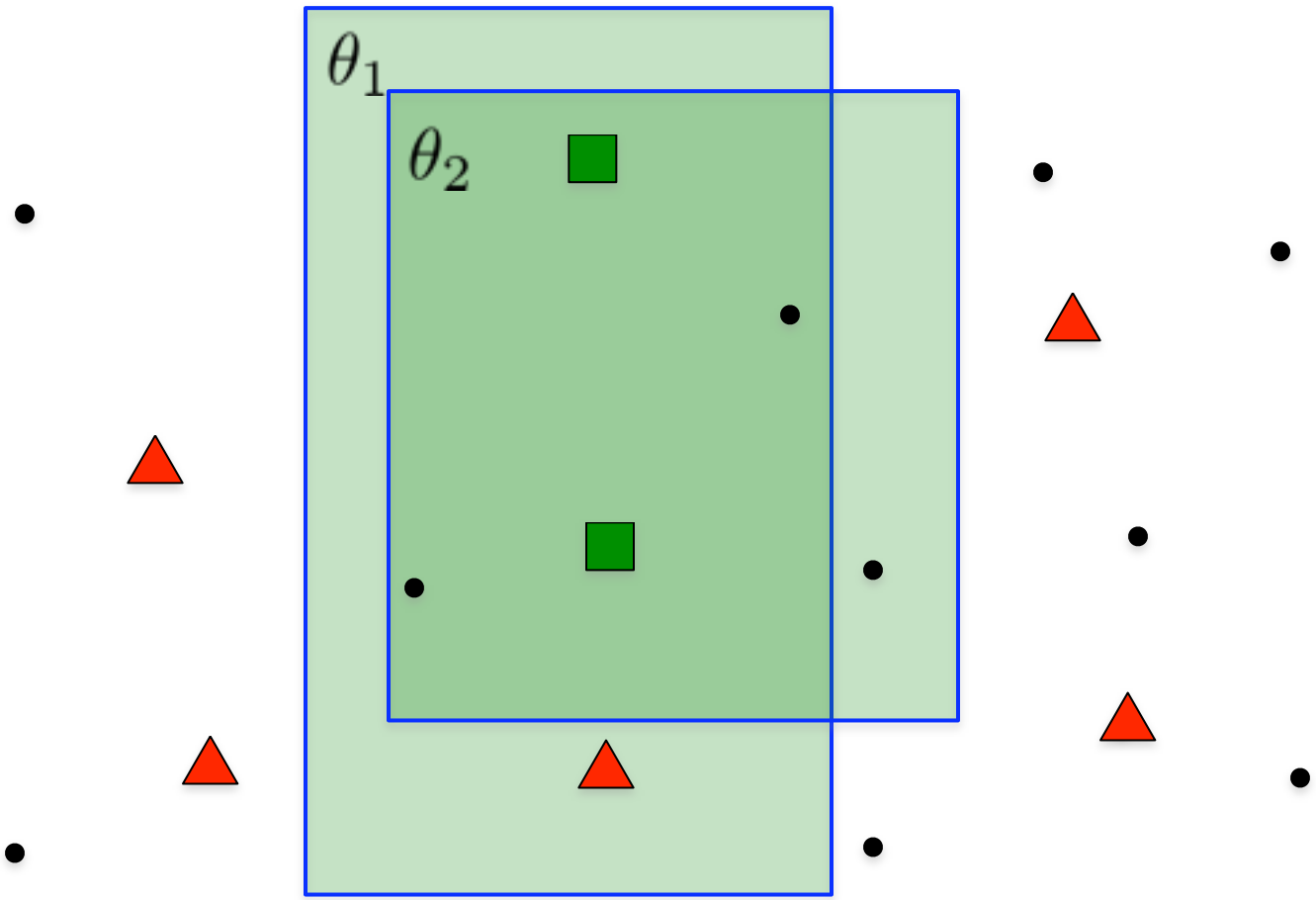
QBC Example



QBC Example



QBC Example



QBC: Design Decisions

- how to build a committee:
  - "sample" models from $\mathrm{P}(\theta|\mathcal{L})$
    - [Dagan & Engelson, ICML'95; McCallum & Nigam, ICML'98]
  - standard ensembles (e.g., bagging, boosting)
    - [Abe & Mamitsuka, ICML'98]
- how to measure disagreement:
  - "XOR" committee classifications
  - view vote distribution as probabilities, use uncertainty measures (e.g., entropy)

Uncertainty vs. QBC

- QBC is a more *general* strategy, incorporating uncertainty over both:
  - instance label
  - model hypothesis

$$\phi_{QBC}(x) = -\sum_{y} \frac{\sum_{\theta \in \mathcal{C}} V_\theta(y|x)}{|\mathcal{C}|} \log_2 \frac{\sum_{\theta \in \mathcal{C}} V_\theta(y|x)}{|\mathcal{C}|}$$

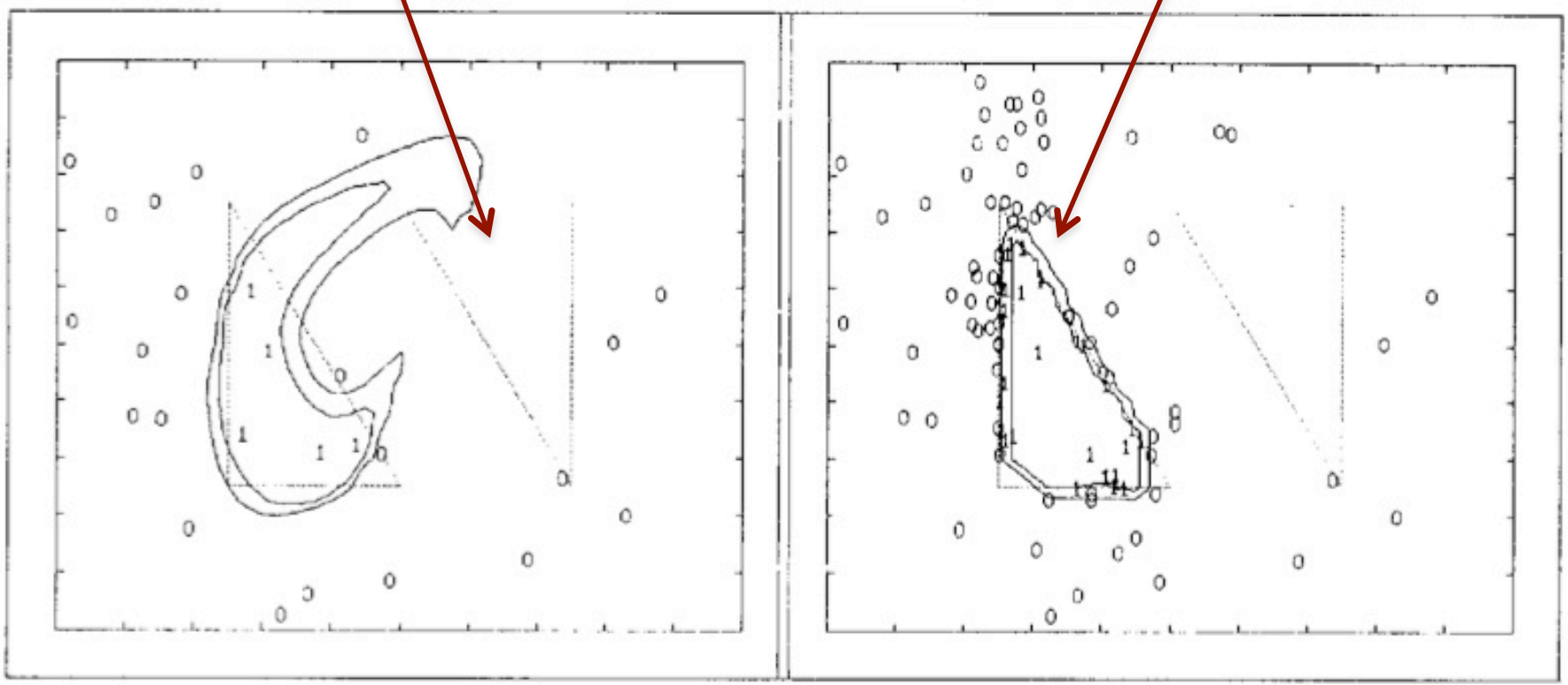
- theoretical guarantees…
  - QBC: $\mathrm{O}(\log_2 \mathrm{d}/\varepsilon)$ query complexity [Seung et al., ML'97]
  - uncertainty sampling: none

[Cohn et al., ML'94]

Pathological Case for Uncertainty

initial random sample fails to hit the right triangle

uncertainty sampling only queries the left side!

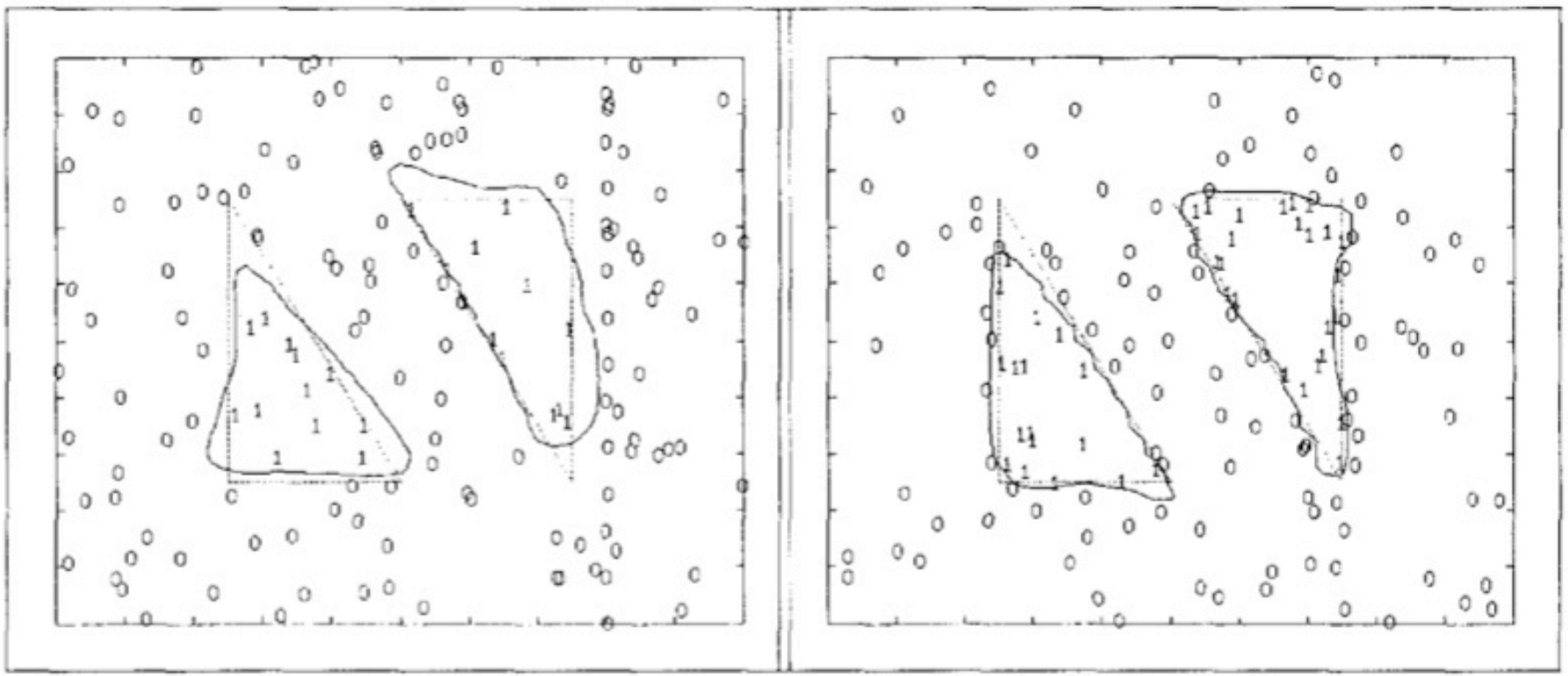


[Cohn et al., ML'94]

Version-Space Sampling Instead

150 random samples

150 active queries (QBC variant)



Active vs. Semi-Supervised

- both try to attack the same problem: making the most of unlabeled data $\mathcal{U}$

- each attacks from a different direction:
  - **semi-supervised** learning ***exploits*** what the model thinks it knows about unlabeled data
  - **active** learning ***explores*** the unknown aspects of the unlabeled data

Active vs. Semi-Supervised

uncertainty sampling
query instances the model
is least confident about

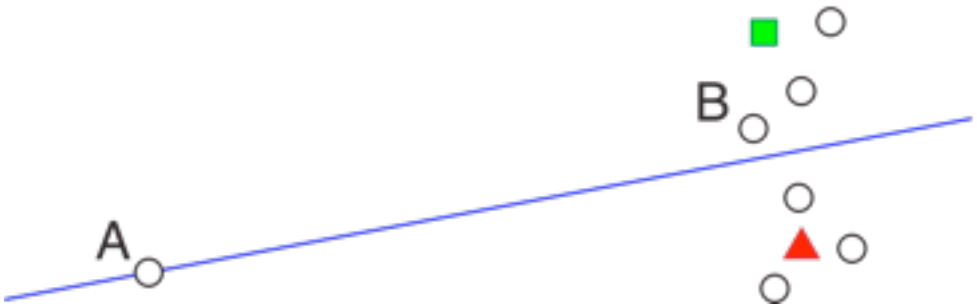
self-training
expectation-maximization (EM)
entropy regularization (ER)
propagate confident labelings
among unlabeled data

query-by-committee (QBC)
use ensembles to rapidly
reduce the version space

co-training
multi-view learning
use ensembles with multiple views
to constrain the version space

Problem: Outliers

- an instance may be uncertain or controversial (for QBC) simply because it's an *outlier*



- querying outliers is not likely to help us reduce error on more typical data

Solution 1: Density Weighting

- weight the uncertainty ("informativeness") of an instance by its density w.r.t. the pool $\mathcal{U}$

[Settles & Craven, EMNLP'08]

$$\phi_{ID}(x) = H_\theta(Y|x) \times \left(\frac{1}{U} \sum_{u \in \mathcal{U}} \text{sim}(x, u) \right)^\beta$$

"base" informativeness

density term

- use $\mathcal{U}$ to estimate $\mathrm{P}(x)$ and avoid outliers

[McCallum & Nigam, ICML'98; Nguyen & Smeulders, ICML'04; Xu et al., ECIR'07]

[Roy & McCallum, ICML'01; Zhu et al., ICML-WS'03]

Solution 2: Estimated Error Reduction

- minimize the risk $R(x)$ of a query candidate
  - _ expected uncertainty over $\mathcal{U}$ if x is added to $\mathcal{L}$

expectation over possible labelings of x

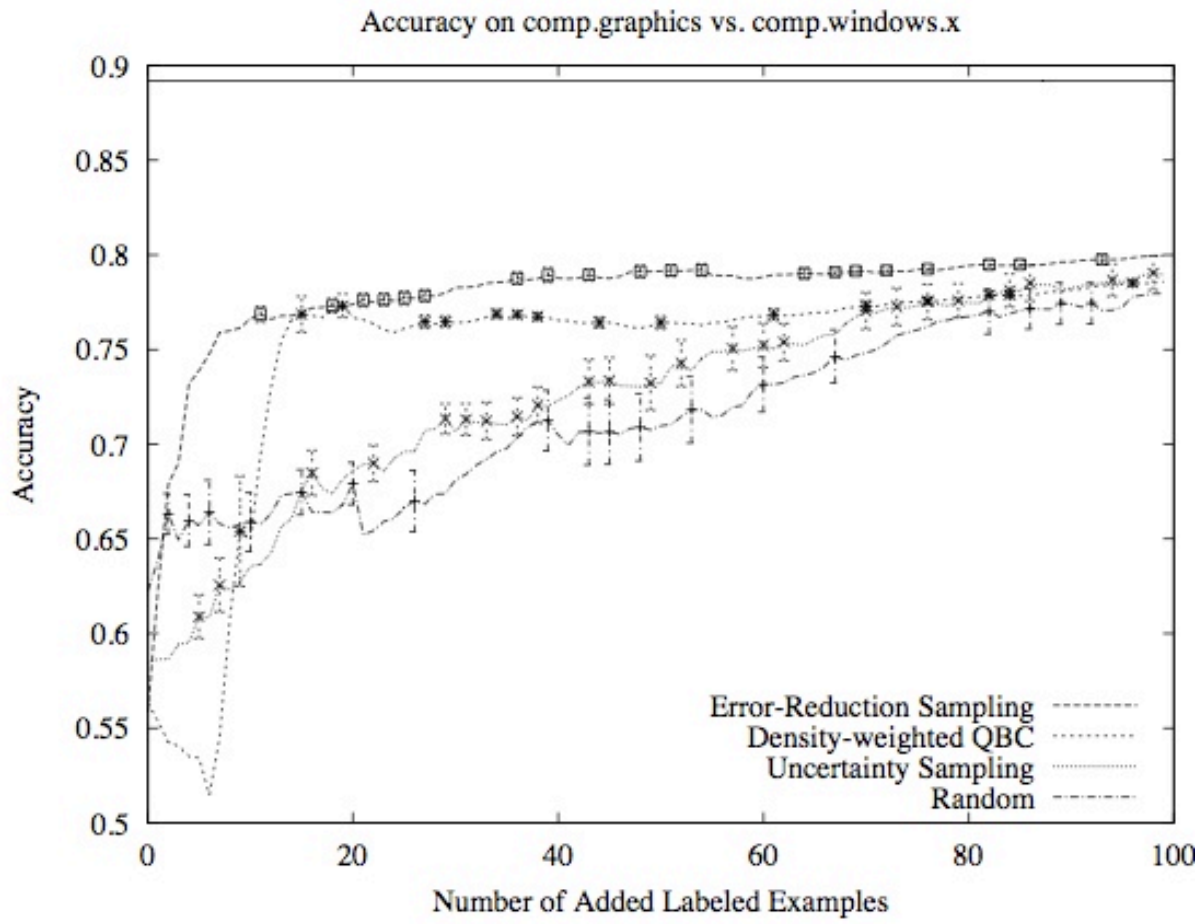
$$R(x) = \sum_{u \in \mathcal{U}} E_y \left[H_{\theta + \langle x, y \rangle} (Y|u) \right]$$

sum over unlabeled instances

uncertainty of u after retraining with x

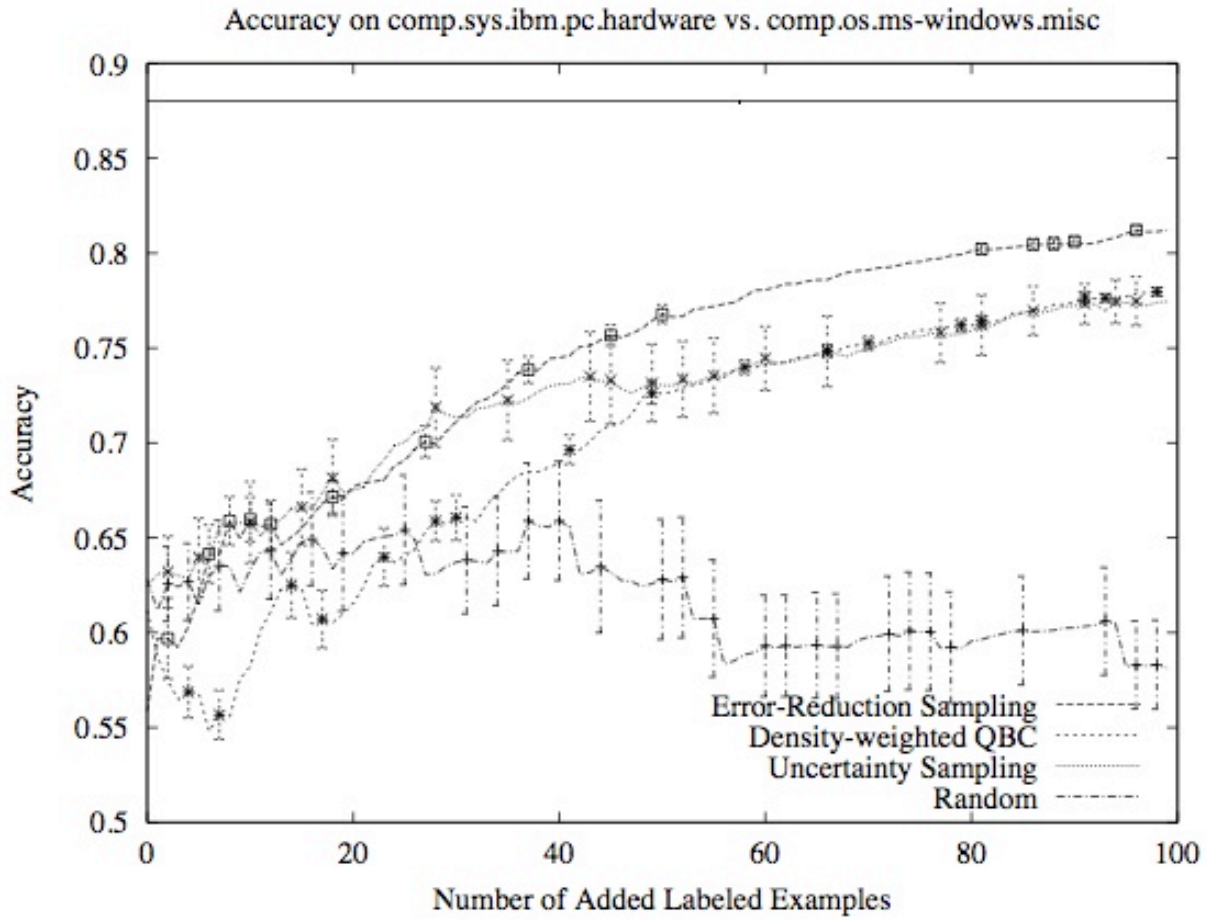
[Roy & McCallum, ICML'01]

Text Classification Examples



[Roy & McCallum, ICML'01]

Text Classification Examples



Relationship to Uncertainty Sampling

- a different perspective: aim to maximize the *information gain* over $\mathcal{U}$

uncertainty *before* query

risk term

$$\phi_{IG}(x) = \sum_{u \in \mathcal{U}} H_\theta(Y|u) - E_y\left[H_{\theta+\langle x,y\rangle}(Y|u)\right]$$

assume x is representative of $\mathcal{U}$

$$\approx H_\theta(Y|x) - E_y\left[H_{\theta+\langle x,y\rangle}(Y|x)\right]$$

assume this evaluates to zero

$$\approx H_\theta(Y|x)$$

…reduces to uncertainty sampling!

“Error Reduction” Scoresheet

- pros:
  - more principled query strategy
  - can be model-agnostic
    - literature examples: naïve Bayes, LR, GP, SVM
- cons:
  - too expensive for most model classes
    - some solutions: subsample $\mathcal{U}$; use approximate training
  - intractable for structured outputs

Alternative Query Types

- so far, we assumed queries are *instances*
  - e.g., for document classification the learner queries *documents*

- can the learner do better by asking different types of questions?
  - *multiple-instance* active learning
  - *feature* active learning

Multiple-Instance (MI) Learning

[TREC Genomics Track 2004]

bag: document = { instances: paragraphs }

- *multiple instance* (MI) learning is one approach to problems like this [Dietterich et al., 1997]

[Andrews et al., NIPS'03; Ray & Craven, ICML'05]

MI Active Learning

- traditional MI learning
  - high ambiguity vs. low cost
- in some MI domains (e.g., text classification), labels can be obtained at the instance level
  - low ambiguity vs. high cost
- MI active learning
  - obtain low-cost bag labels, selectively query instances
  - reduce ambiguity *and* overall labeling cost

[Settles, Craven, & Ray NIPS'07]

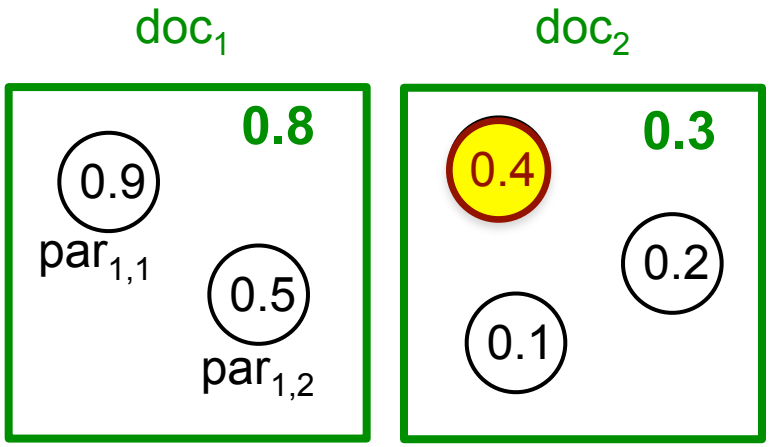
MI Uncertainty (MIU)

- weight the uncertainty of an instance by its "relevance" to the bag-level output

$$\phi_{MIU}(x_n) = H_\theta(Y|x_n) \left(\frac{\partial o}{\partial o_n}\right)^\beta$$

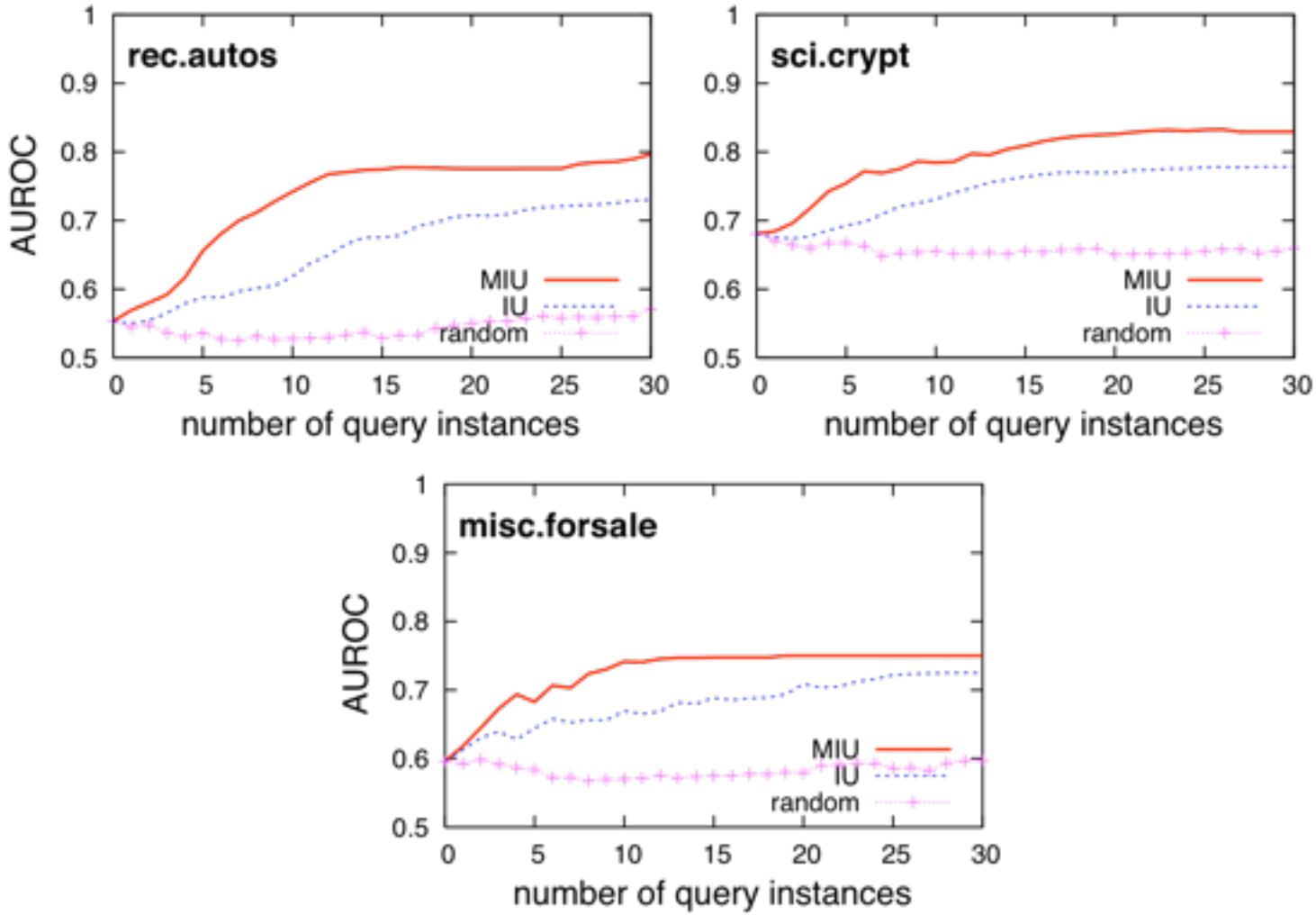
"base" uncertainty

"relevance" term



[Settles, Craven, & Ray NIPS'07]

MI Active Learning Results



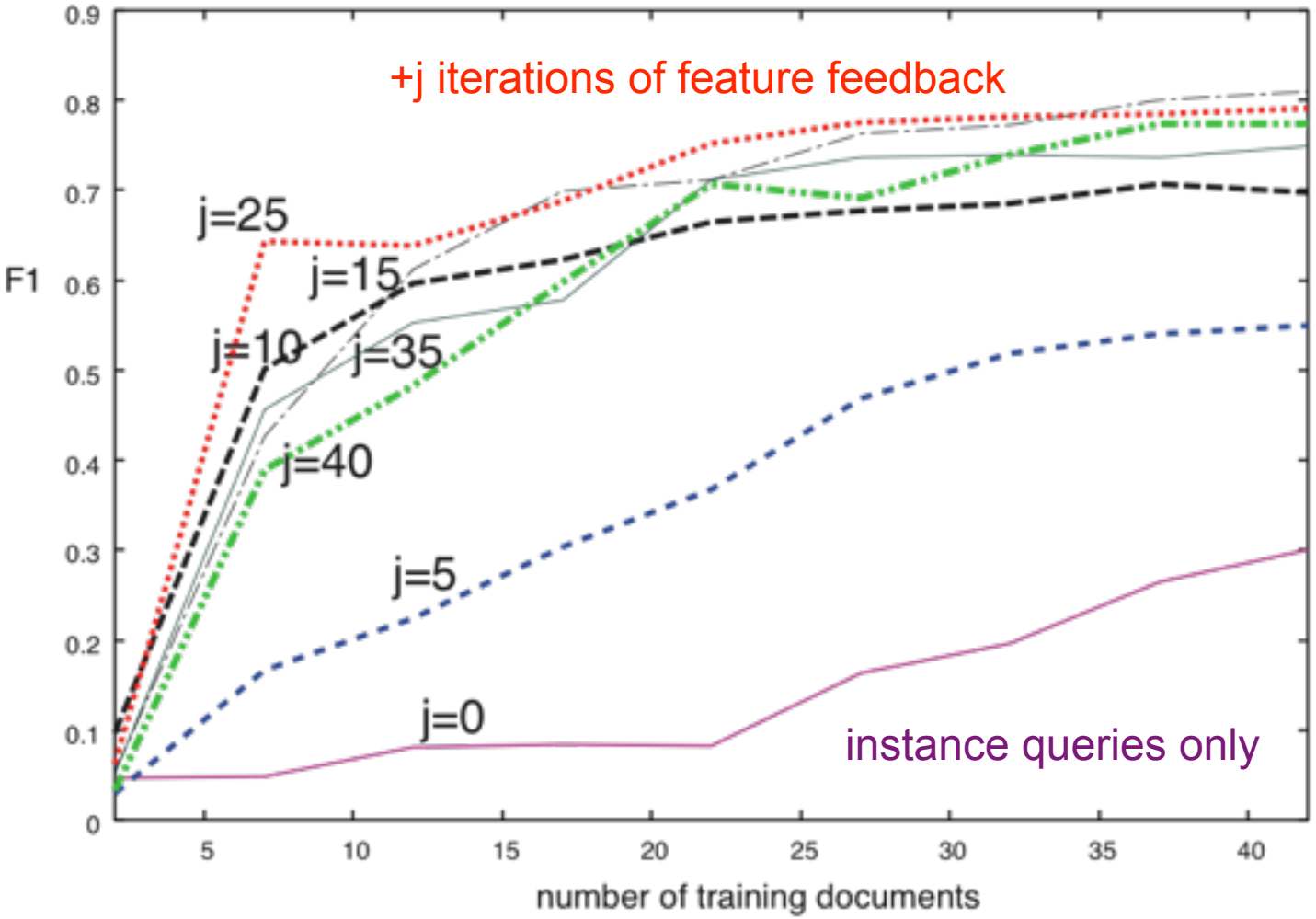
Feature Active Learning

- in NLP tasks, we can often intuitively label *features*
  - the feature word "*puck*" indicates the class **hockey**
  - the feature word "*strike*" indicates the class **baseball**

- **tandem learning** exploits this by asking both instance-label and feature-relevance queries
  [Raghavan et al., JMLR'06]
  - e.g., "is *puck* an important discriminative feature?"

[Raghavan et al., JMLR'06]

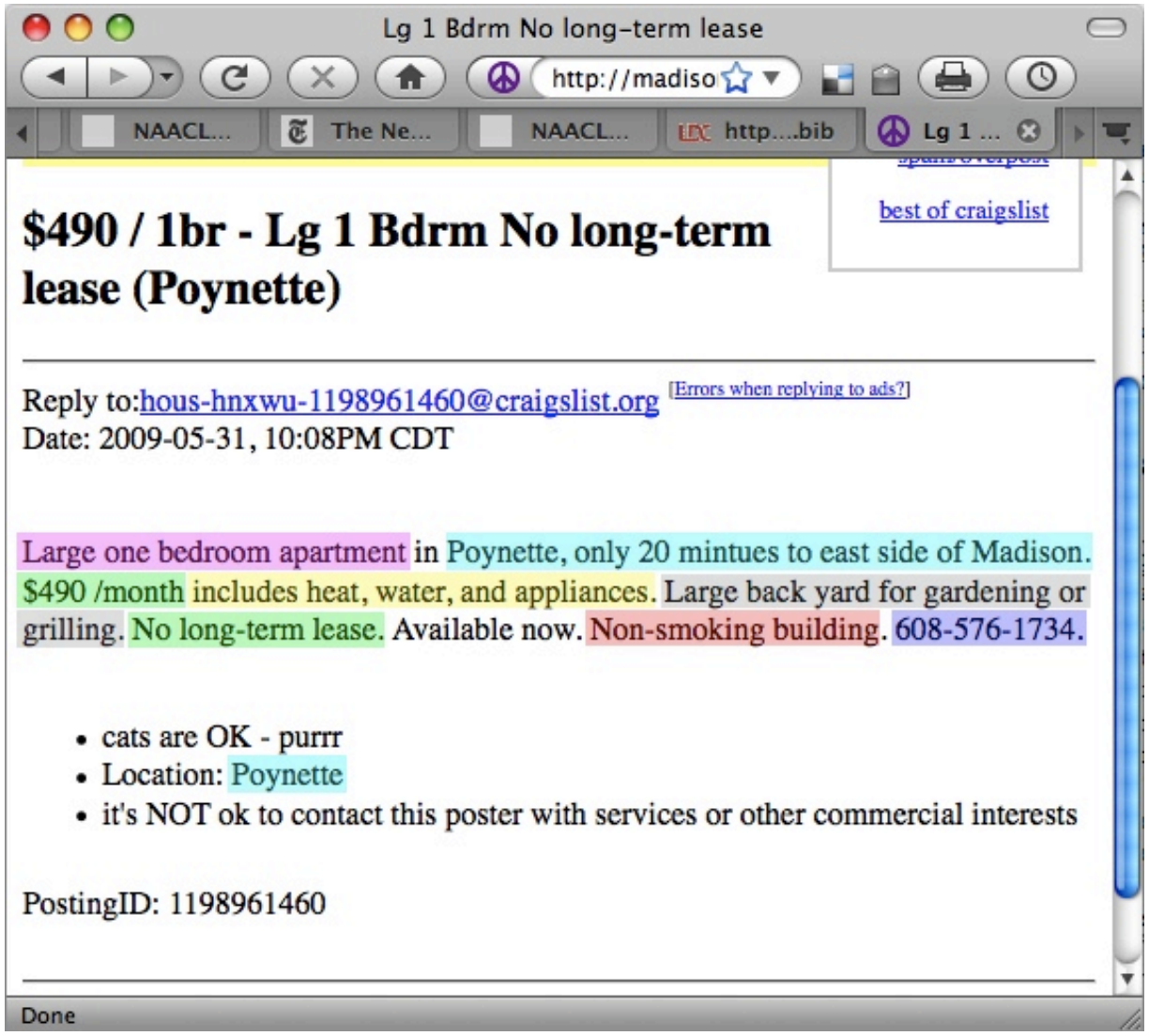
Tandem Learning: Text Classification



Feature Labeling

- recent, alternative forms of "supervision" combine "feature labels" with $\mathcal{U}$ for semi-supervised learning
  - prototype-driven learning [Haghighi & Klein, NAACL'06]
  - generalized expectation (GE) criteria [Mann & McCallum ACL'08; Druck et al., SIGIR'08]
- can we *actively* solicit these feature labels?

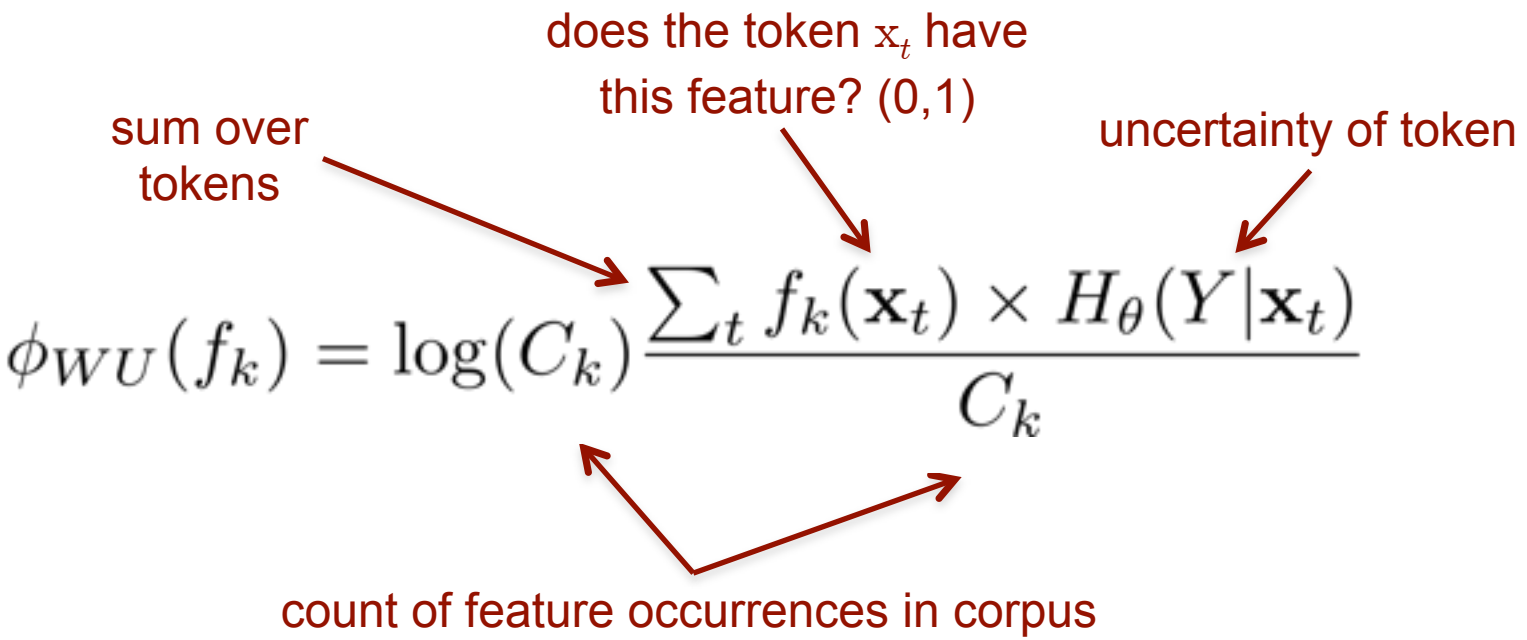
Example: Information Extraction



feature	label
[PHONE]	contact
lease	rent
bedroom	size
large	size /
water	utilities
east	neighborhood
non-smoking	restrictions

[Druck et. al, EMNLP'09]

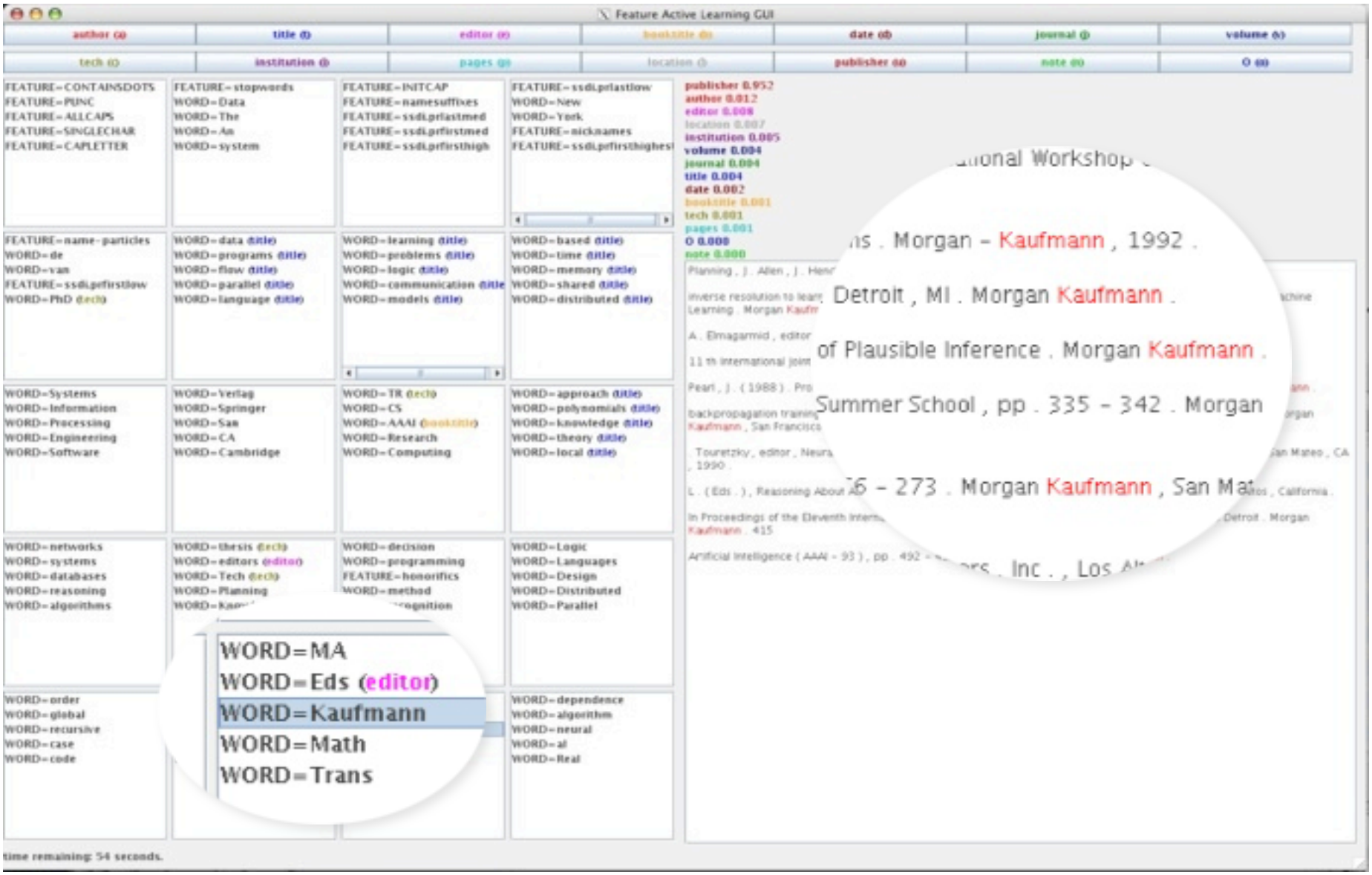
Weighted Uncertainty for Features



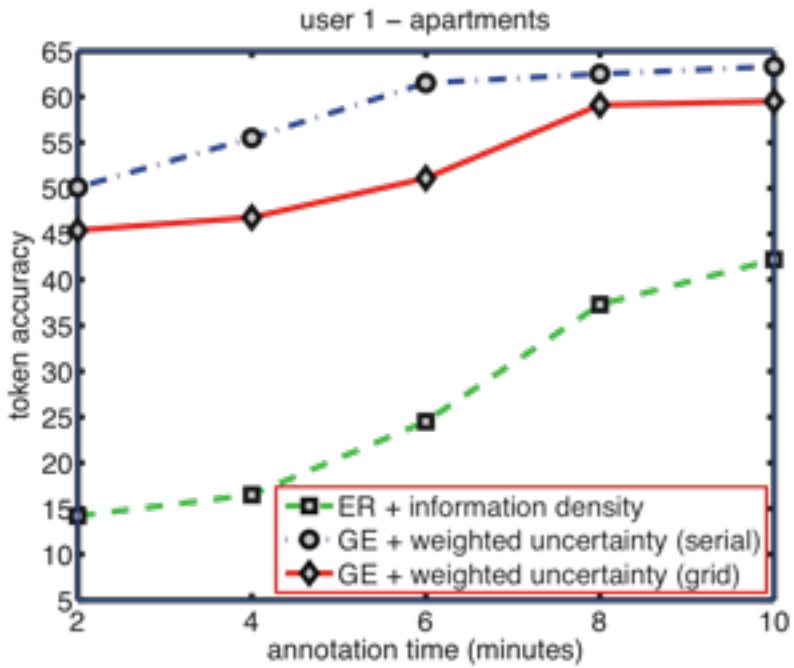
a form of density-weighted uncertainty sampling

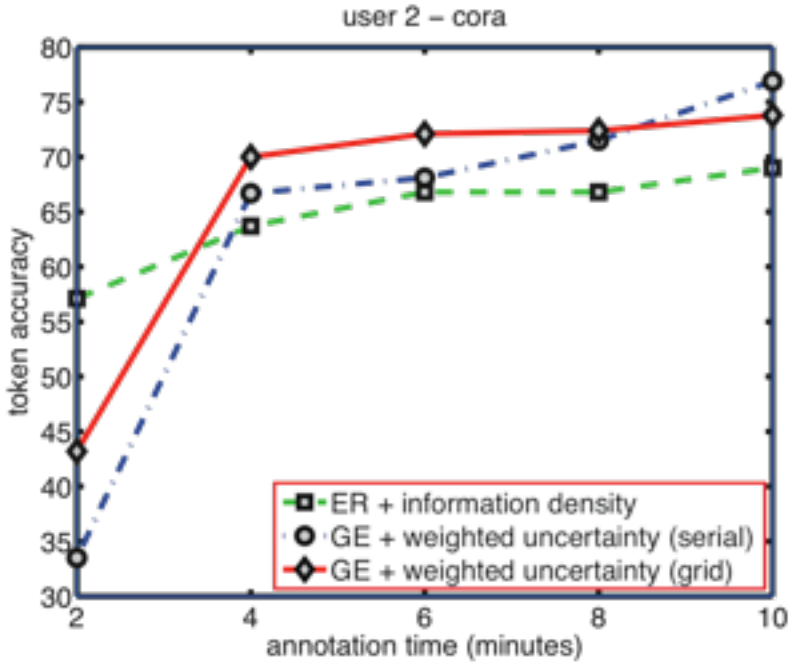
[Druck et. al, EMNLP'09]

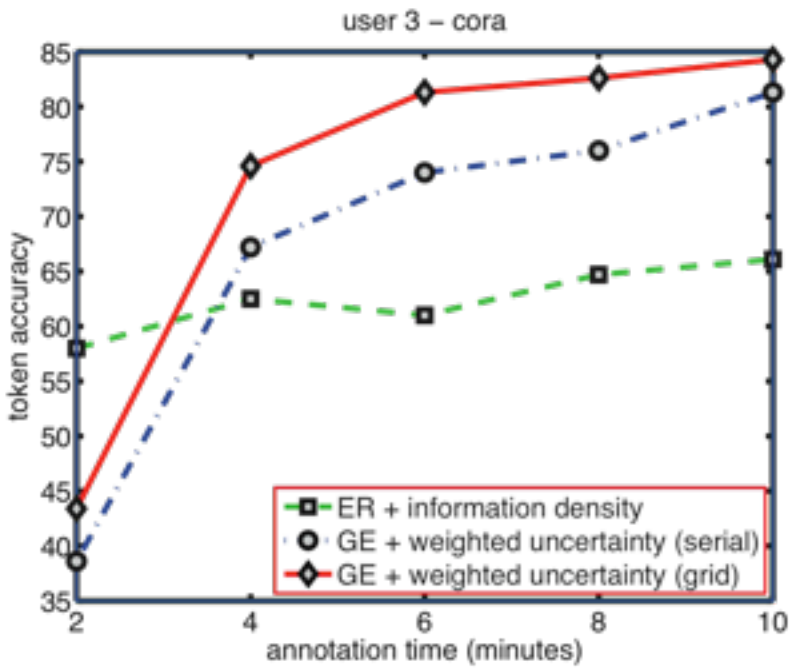
"Grid" Labeling Interface



User Experiments







five 2-minute labeling sessions
with real human annotators

[Druck et. al, EMNLP'09]

Real-World Annotation Costs

- so far, we've assumed that queries are equally expensive to label
  - for many tasks, labeling "costs" vary

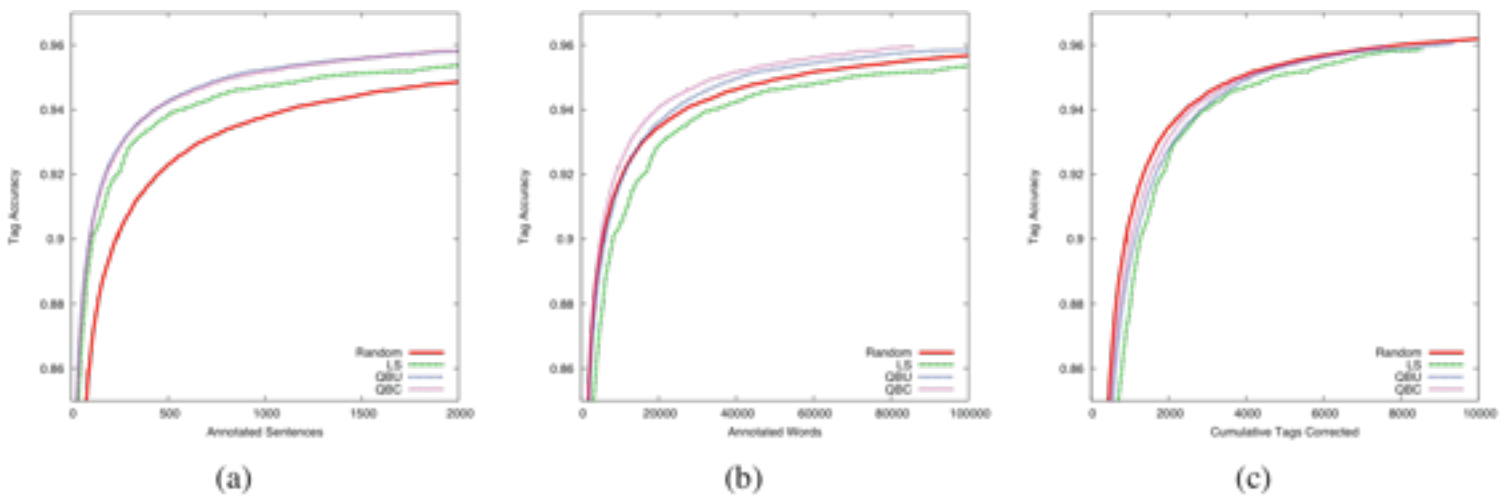


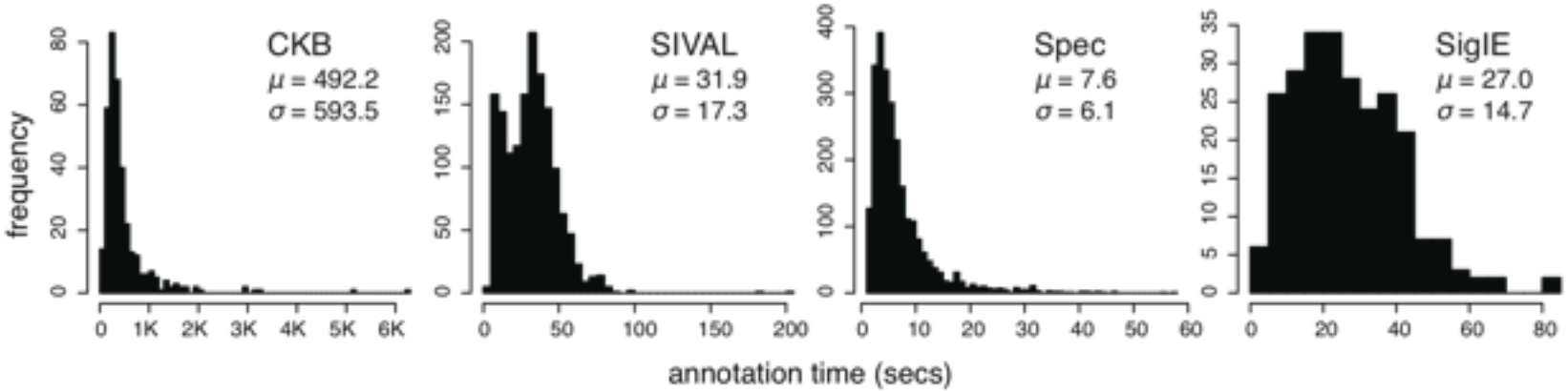
Figure 1: QBU, LS, QBC, and the random baseline plotted in terms of accuracy versus various cost functions: (a) number of sentences annotated; (b) number of words annotated; and (c) number of tags corrected.

[Haertel et al., ACL'08]

[Settles et. al, NIPS'08]

Annotation Time As Cost

- do annotation times vary among instances?

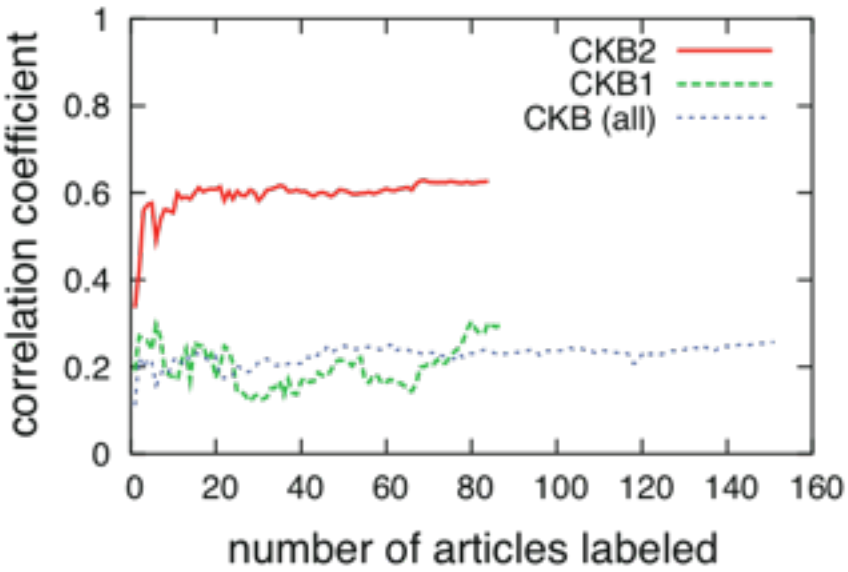


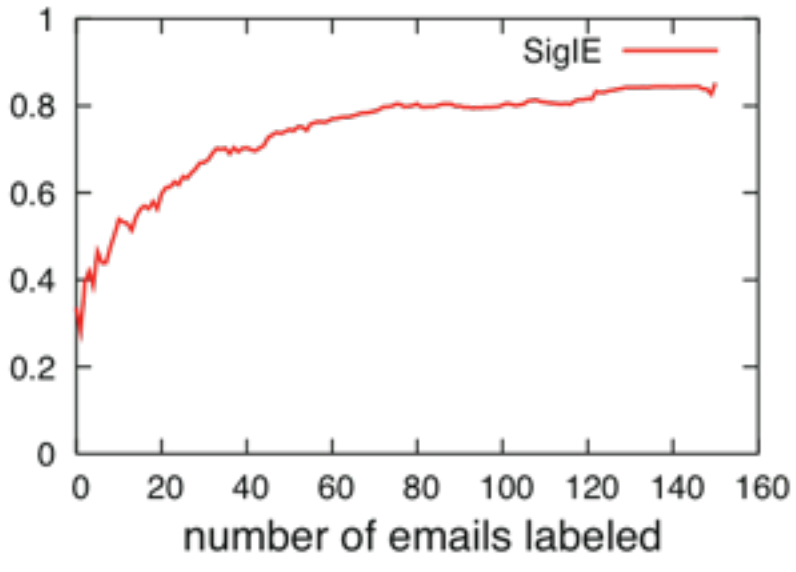
- where does this variance come from?
  - sometimes annotator-dependent
  - stochastic effects

[Settles et. al, NIPS'08]

Can Labeling Times be Predicted?

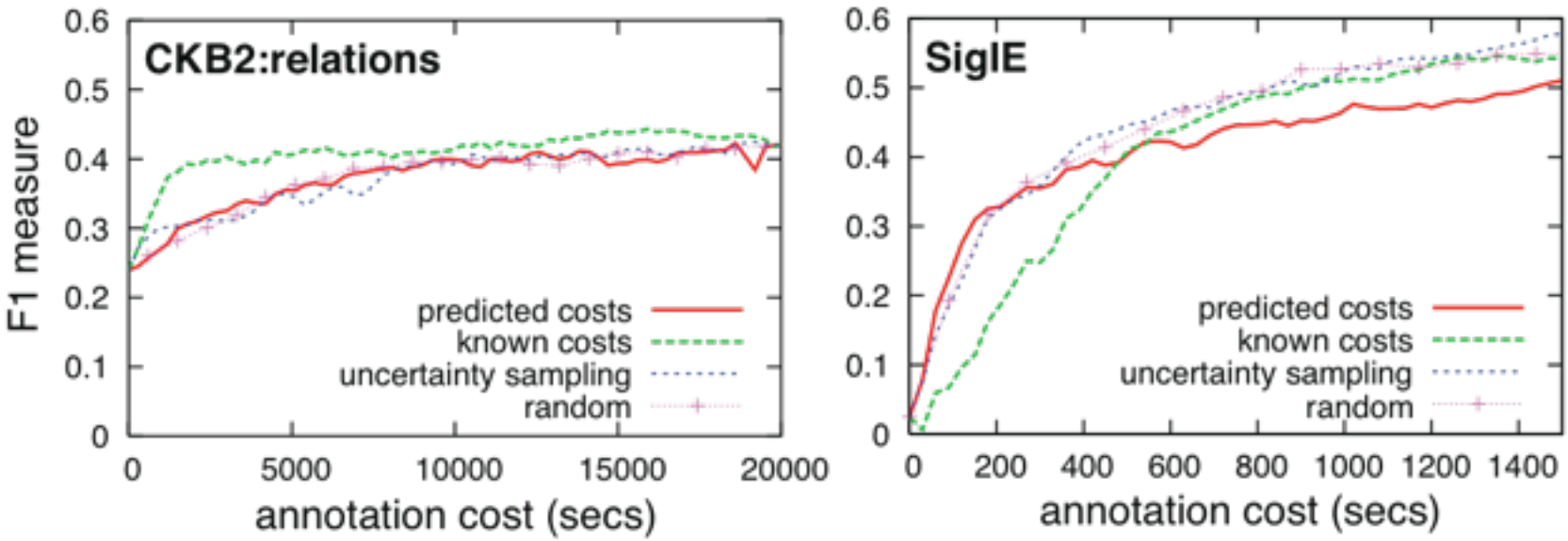
cost predictor: regression model using meta-features





[Settles et. al, NIPS'08]

Can Predicted Times Improve AL?



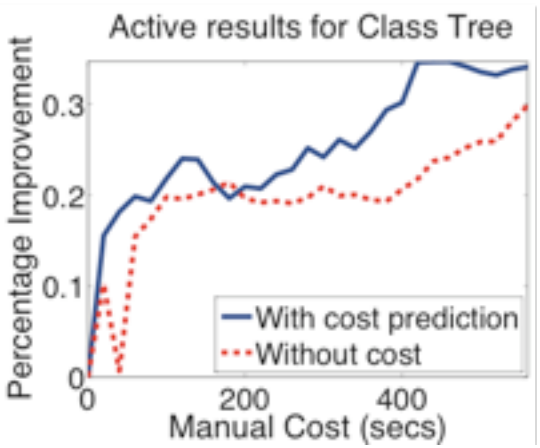
$$\phi(x) = \frac{H_\theta(Y|x)}{\hat{\$}(x)}$$

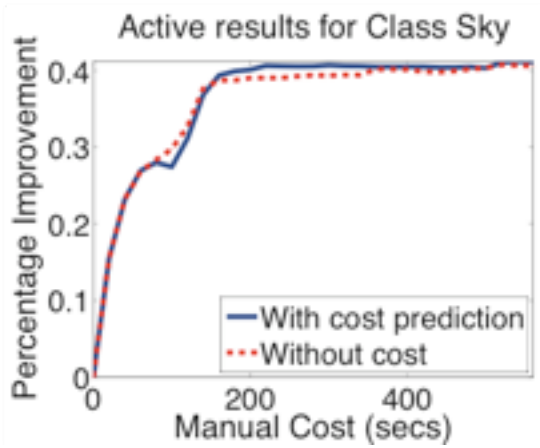
several other negative/ambiguous results in NLP domains

[Aurora et al., ALNLP'09; Tomanek et al.]

But All Is Not Lost!

- some positive results using predicted costs have been obtained in the vision community





[Vijayanarasimhan & Grauman, CVPR'09]

Other Interesting Issues

- many non-expert annotators [Sheng et al., KDD'08]
- user interface issues [Culotta et al., AI'06]
- data reusability [Baldridge & Osbourne, EMNLP'04]
- batch mode active learning [Hoi et al., ICML'06]
- multi-task active learning [Reichart et al., ACL'08]

HW3 Discussion

1. What are some active learning **opportunities** in the context of a NELL system?

- How are these opportunities **similar** to or **different** from problem settings in previous work on active learning? Think about the kinds of queries the learner can ask and how the labels are obtained.

- Identify some **practical issues** for active learning in a never-ending learning system like NELL. How might we address these issues?