

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...

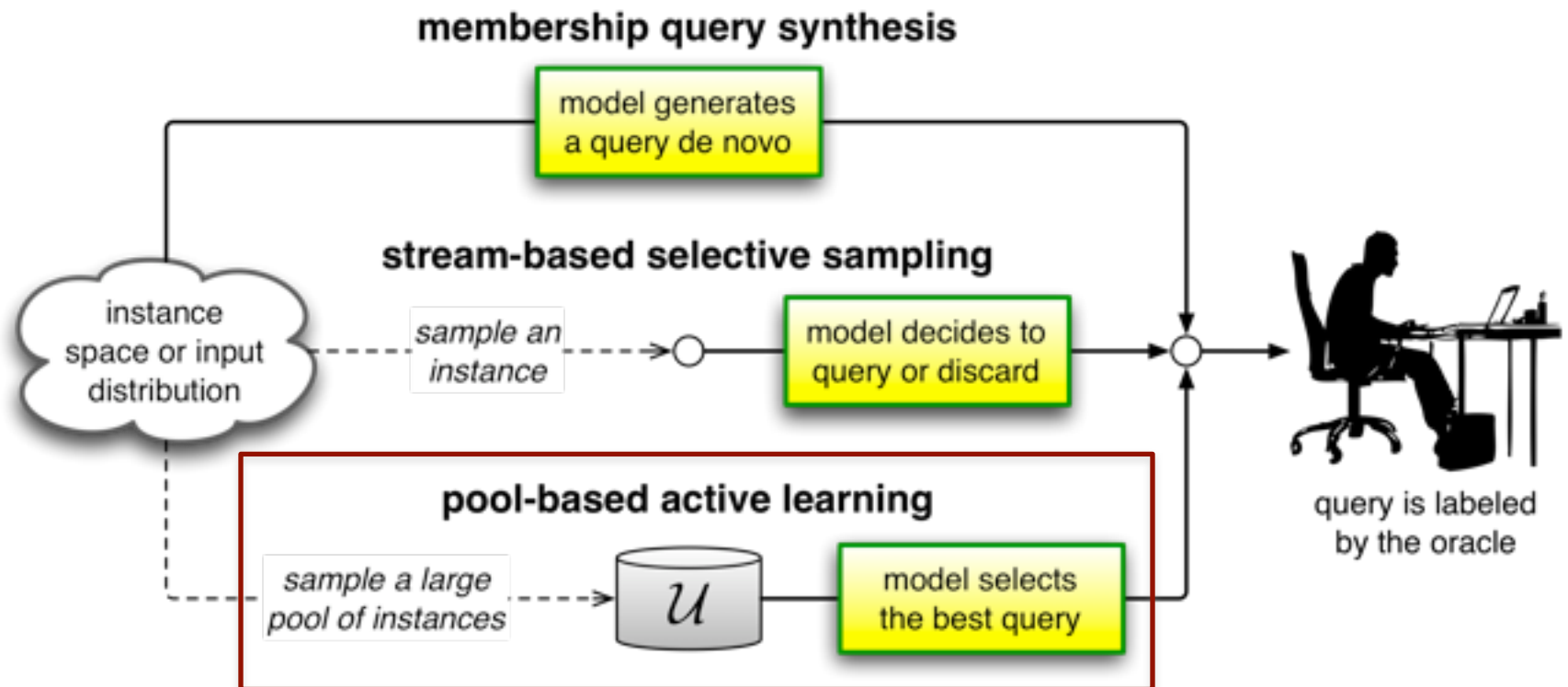
under the PAC model, assume we need $O(1/\epsilon)$ i.i.d. instances to train a classifier with error ϵ .

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

Relationship to Active Learning

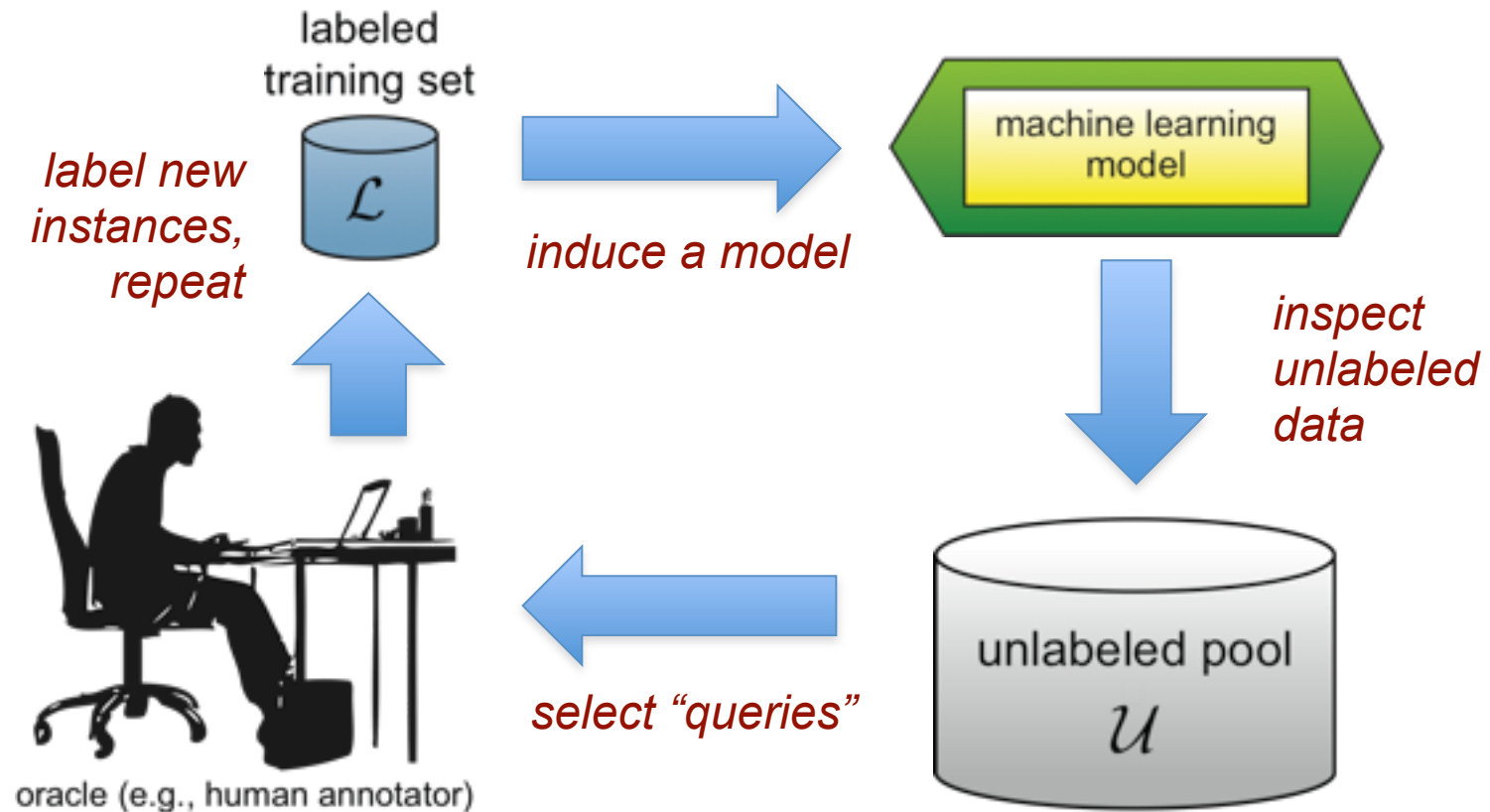
- **key idea:** the learner can choose training data
 - on Mars: whether a fruit was poisonous/safe
 - *in general*: the true label of some instance
- **goal:** reduce the training costs
 - on Mars: the number of “lives at risk”
 - *in general*: the number of “queries”

Active Learning Scenarios

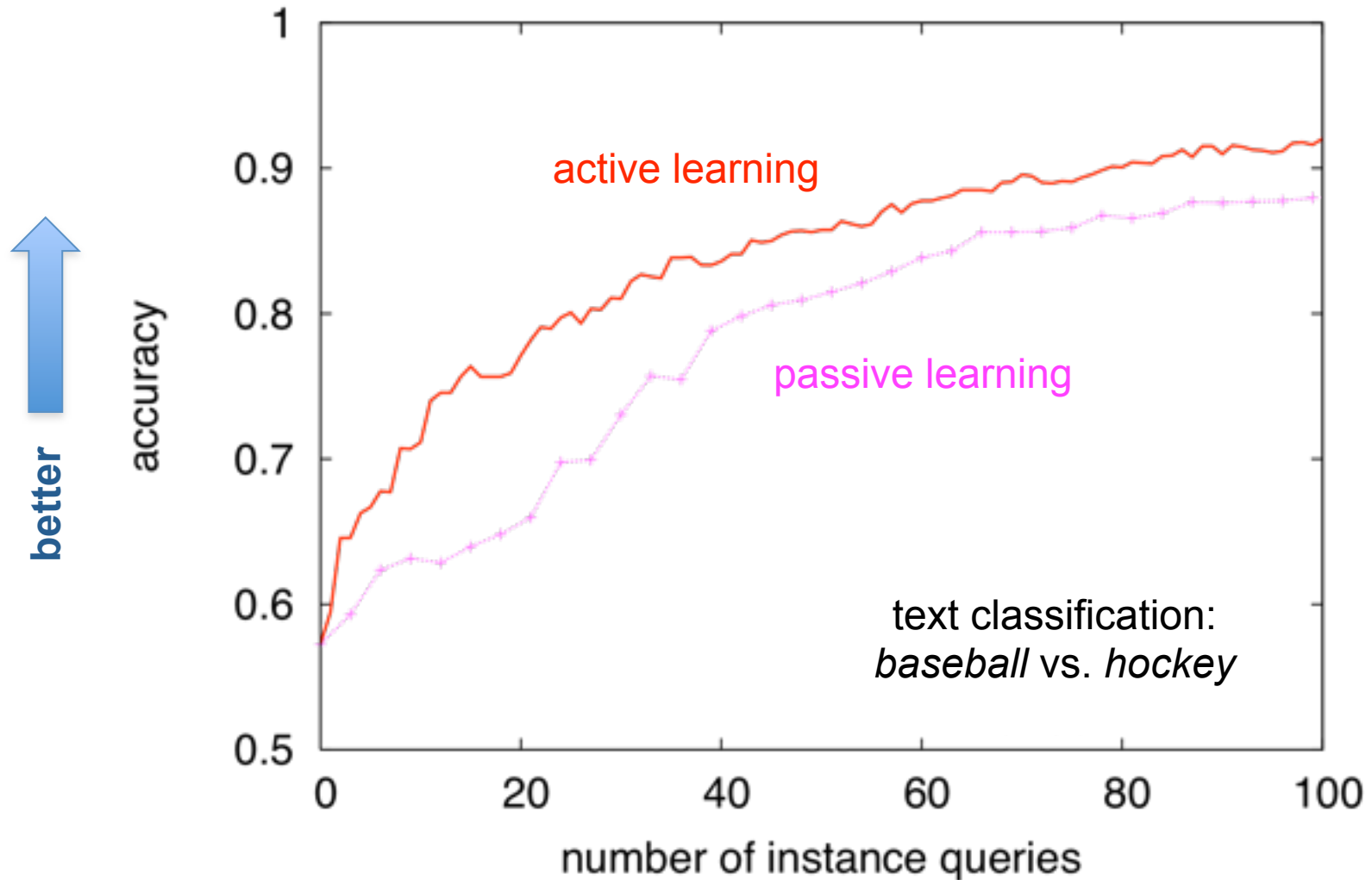


most common in NLP applications

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*



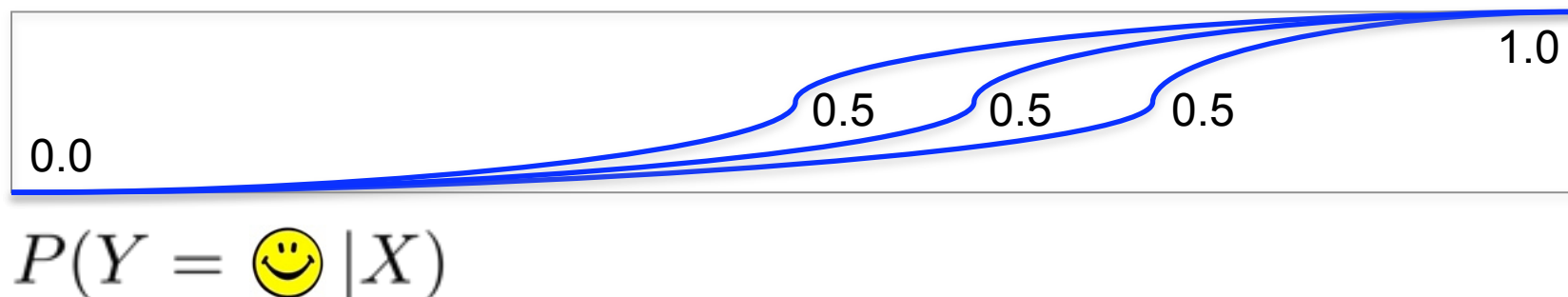
MS Outlook voicemail plug-in [Kapoor et al., IJCAI'07];
“A variety of prototypes that are in use throughout the company.” – *Eric Horvitz*



“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*

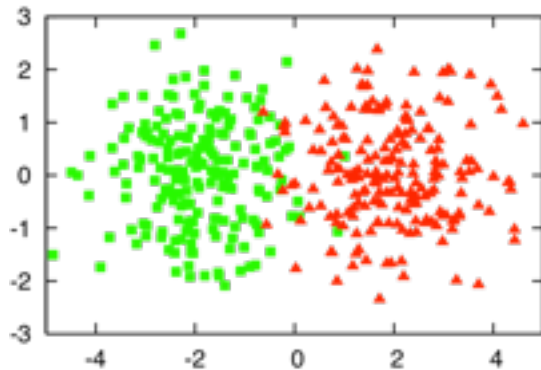
How to Select Queries?

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

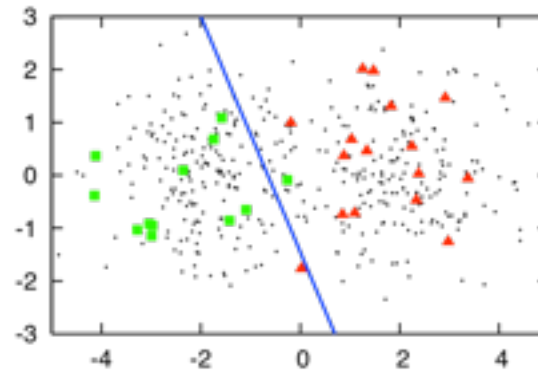


Uncertainty Sampling

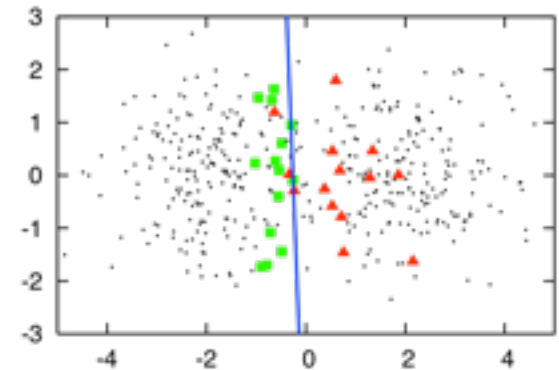
- query instances the learner is *most uncertain* about



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

Multi-Class Uncertainty Measures

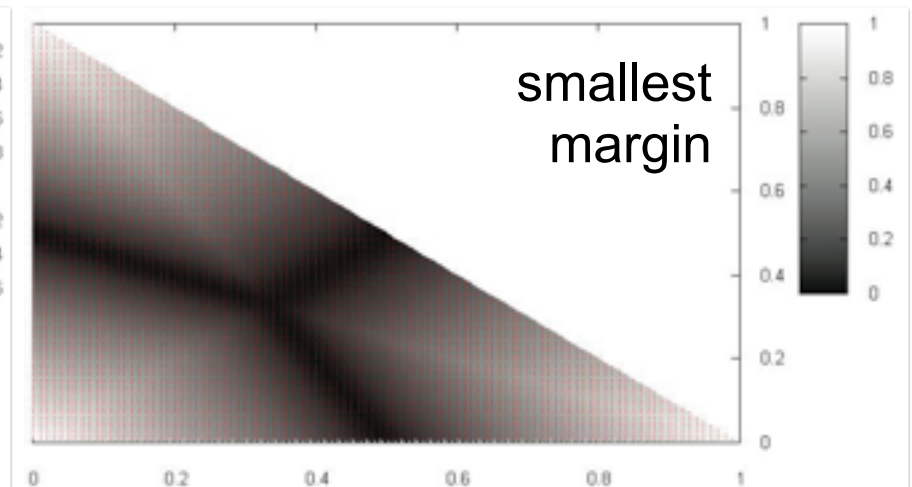
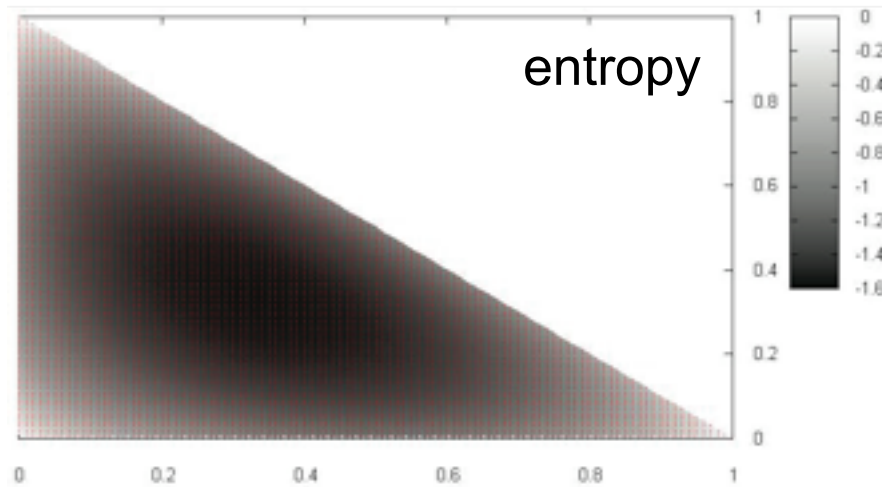
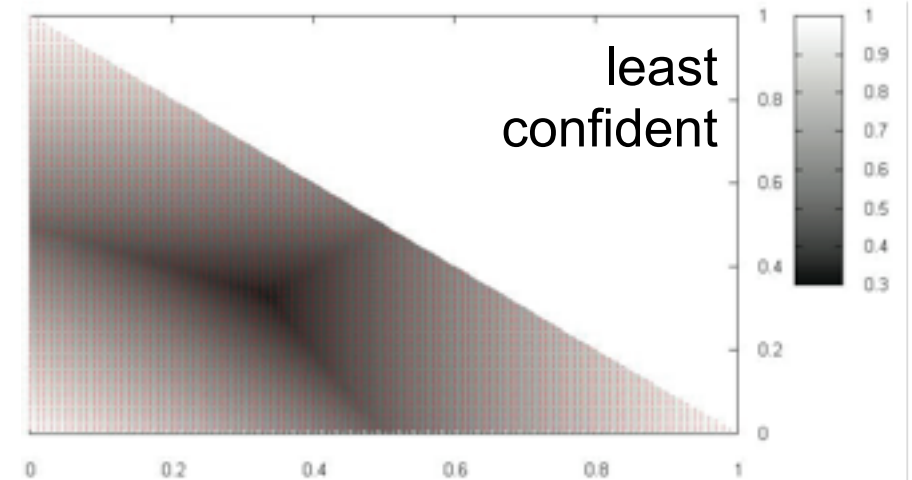


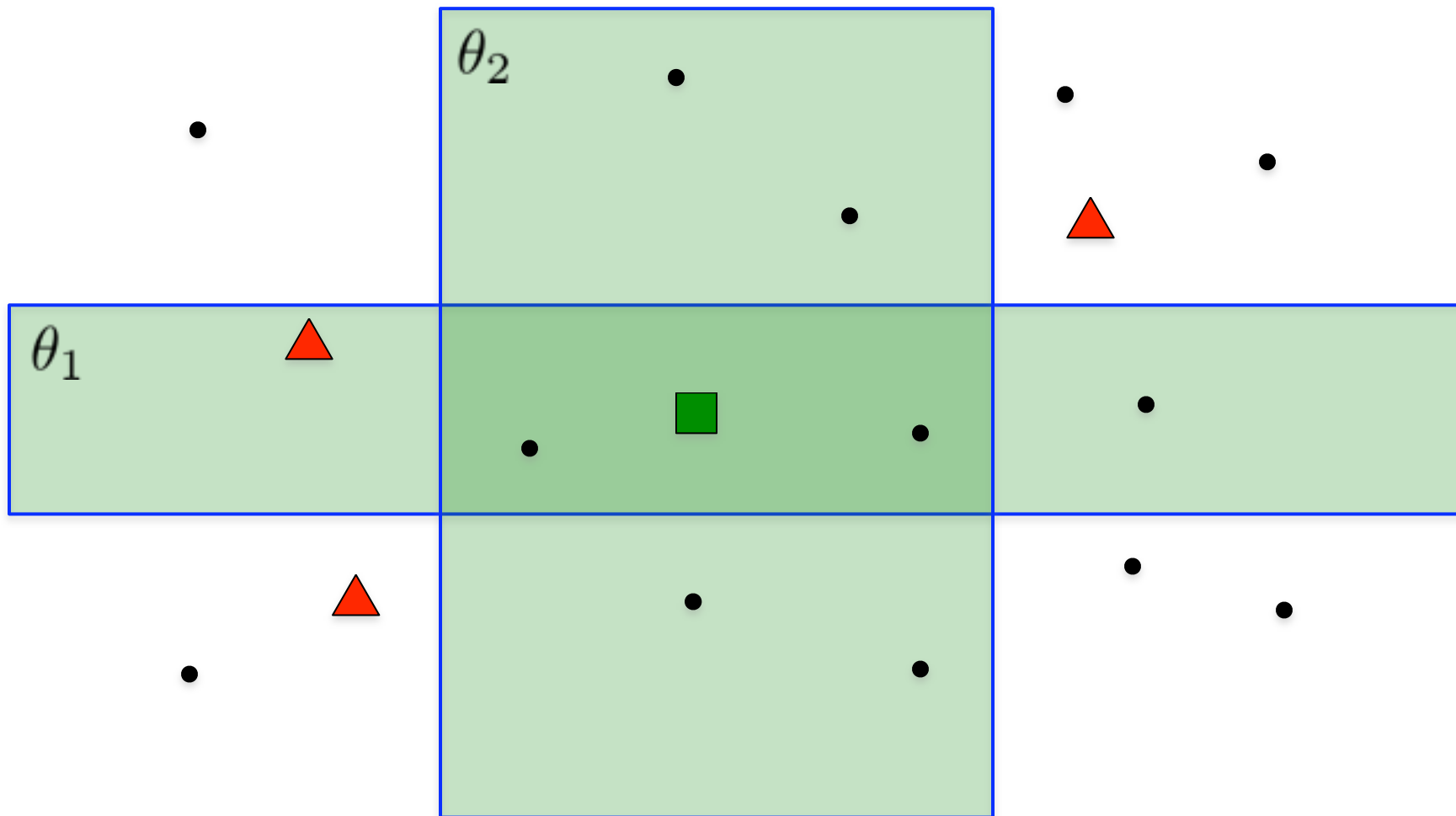
illustration of preferred (darker)
posterior distributions in a
3-label classification task



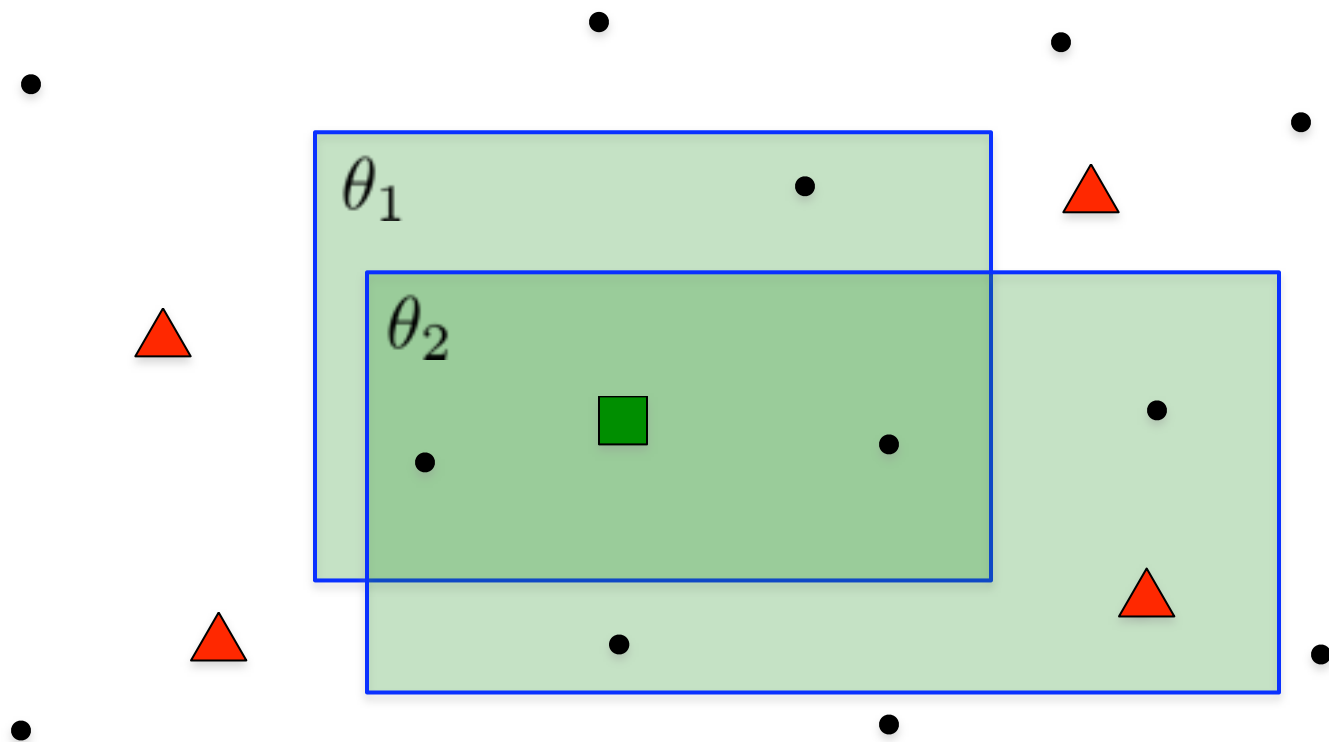
Query-By-Committee (QBC)

- train a committee $\mathcal{C} = \{\theta_1, \theta_2, \dots, \theta_C\}$ 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

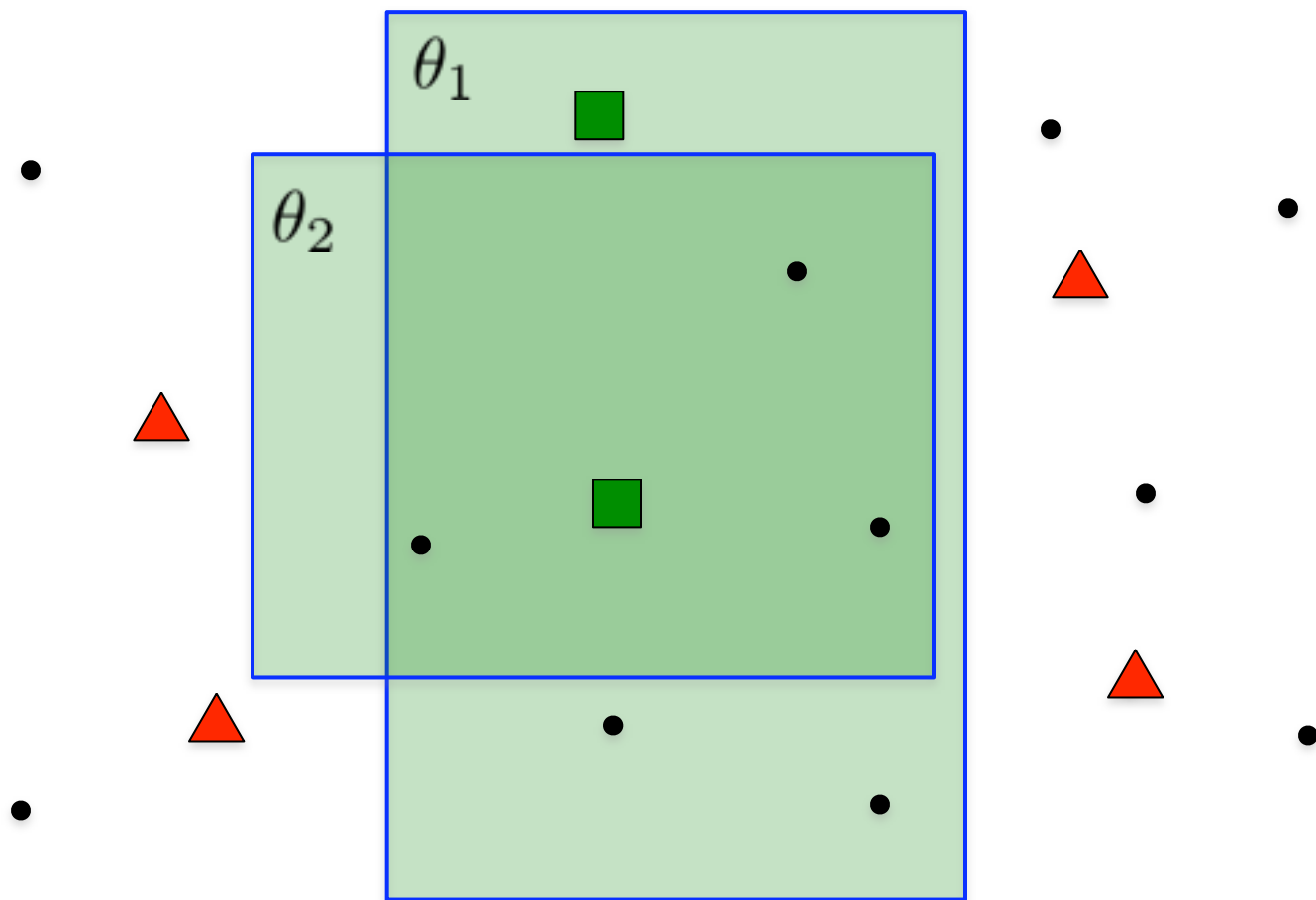
QBC Example



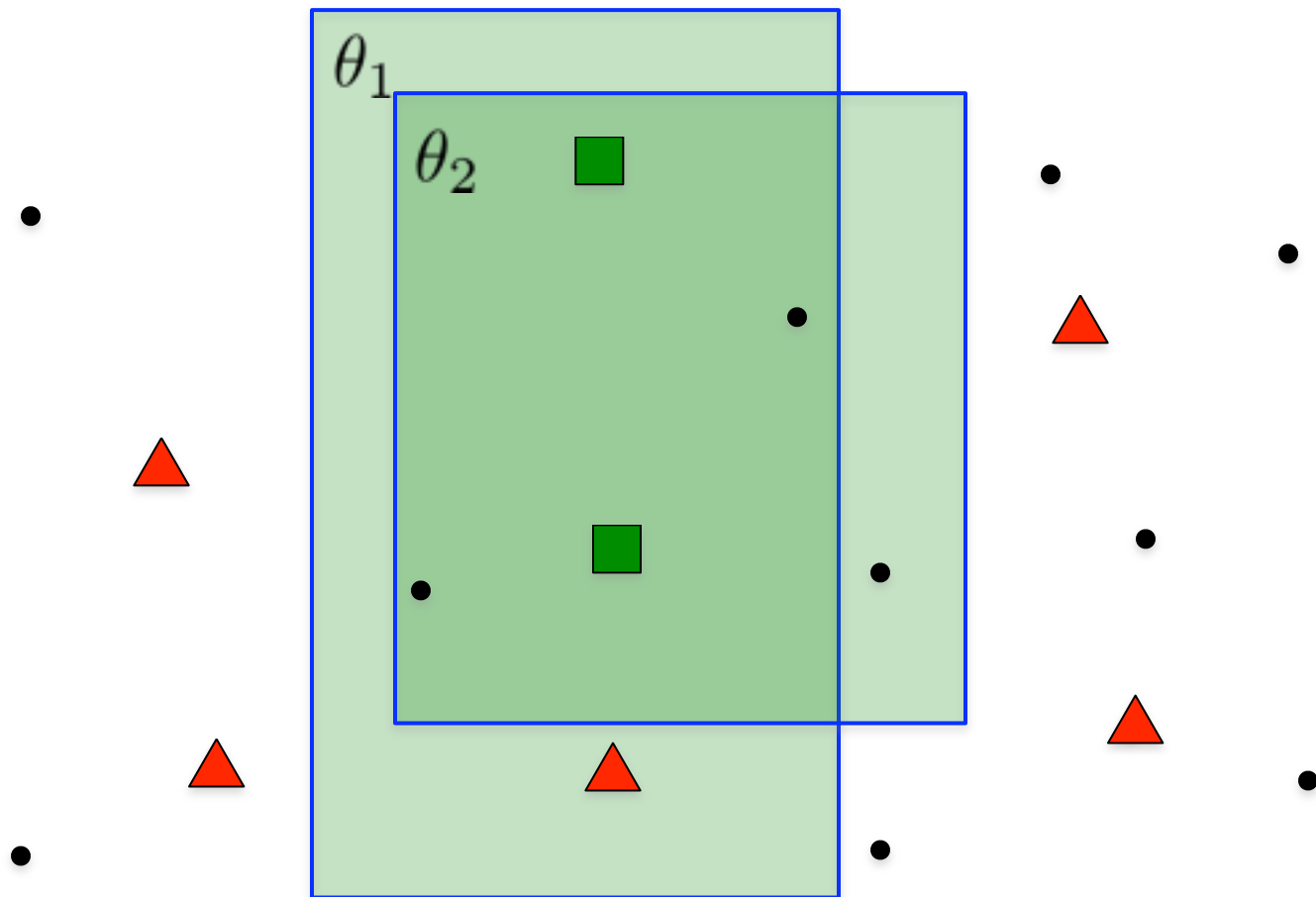
QBC Example



QBC Example



QBC Example



QBC: Design Decisions

- how to build a committee:
 - “sample” models from $P(\theta|\mathcal{L})$
 - [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

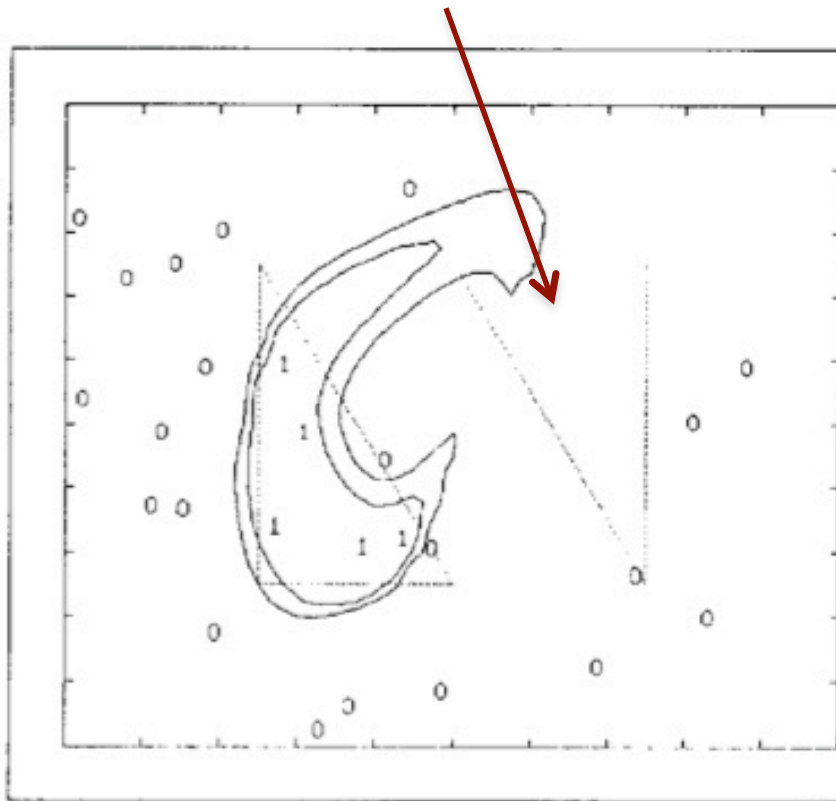
- QBC is a more *general* strategy, incorporating uncertainty over both:
 - instance label
 - model hypothesis

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

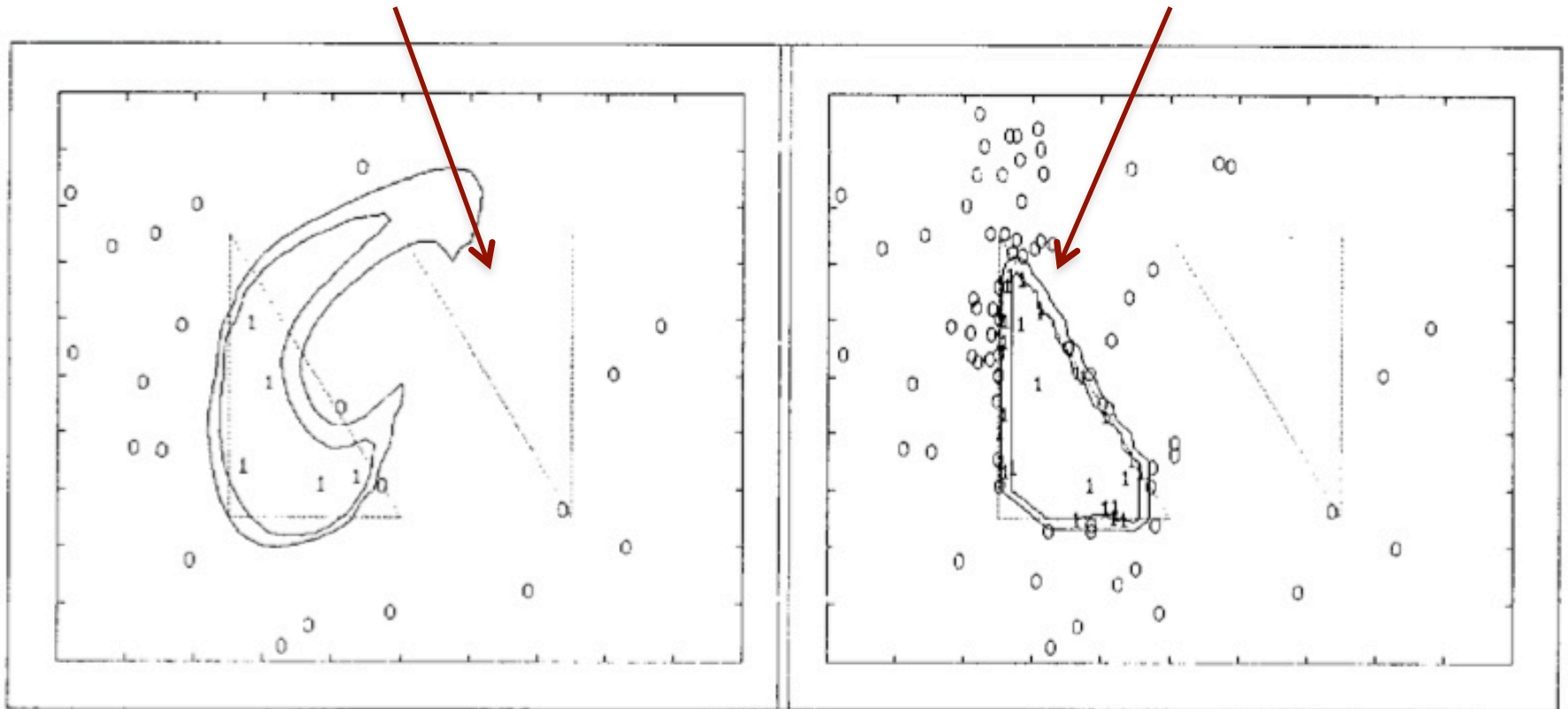
- theoretical guarantees...
 - QBC: $O(\log_2 d/\epsilon)$ query complexity [Seung et al., ML'97]
 - uncertainty sampling: none

Pathological Case for Uncertainty

initial random sample fails
to hit the right triangle

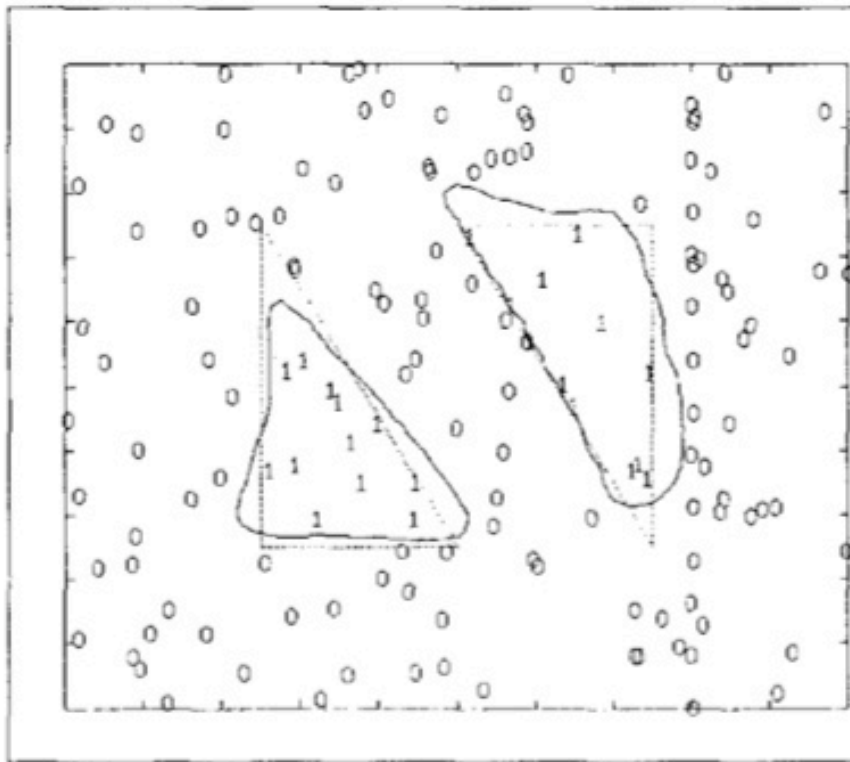


uncertainty sampling only
queries the left side!

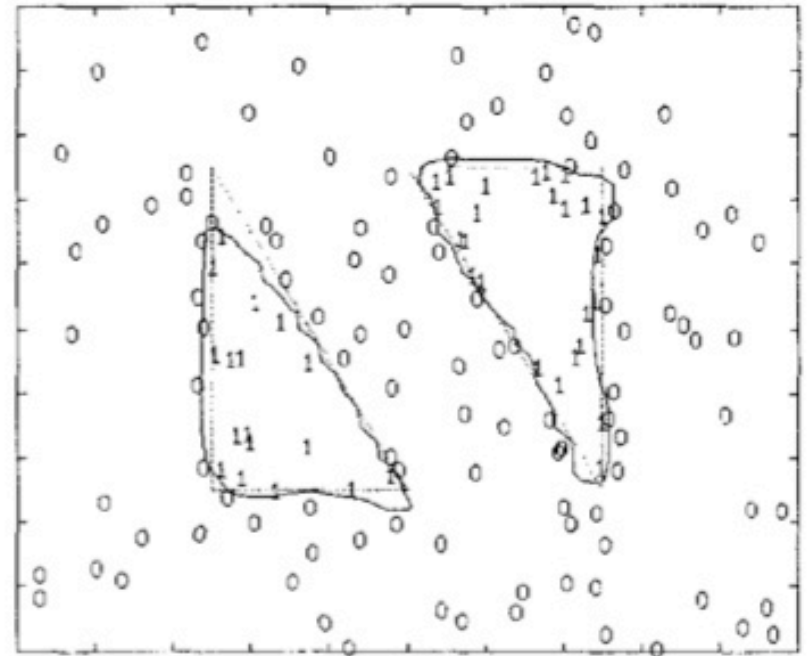


Version-Space Sampling Instead

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)

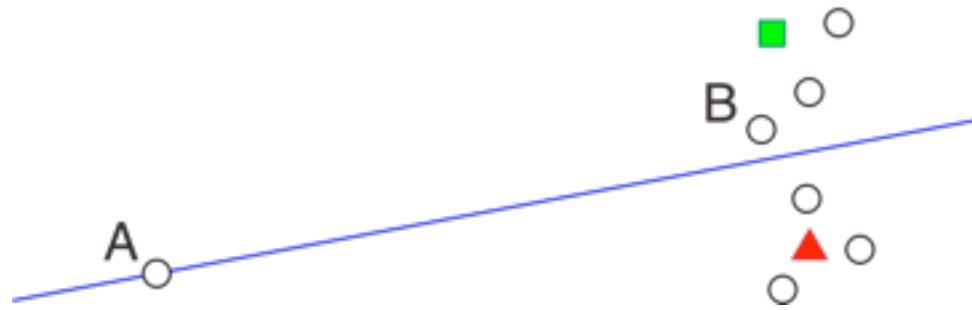
use ensembles to rapidly
reduce the version space



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]

$$\phi_{ID}(x) = \underbrace{H_{\theta}(Y|x)}_{\text{“base” informativeness}} \times \underbrace{\left(\frac{1}{U} \sum_{u \in \mathcal{U}} \text{sim}(x, u) \right)^{\beta}}_{\text{density term}}$$

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

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

Solution 2: Estimated Error Reduction

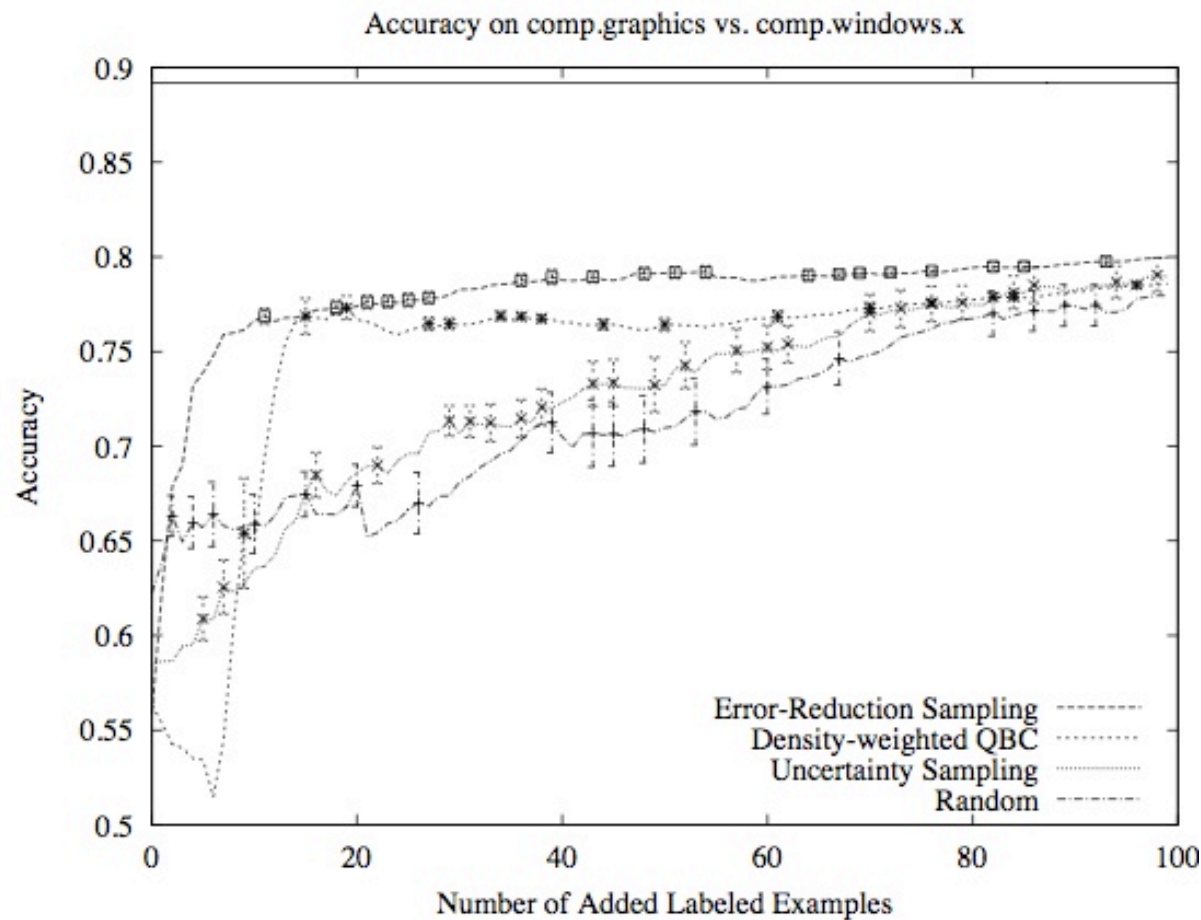
- minimize the risk $R(x)$ of a query candidate
 - _ expected uncertainty over $\mathcal{U}$ if x is added to $\mathcal{L}$

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

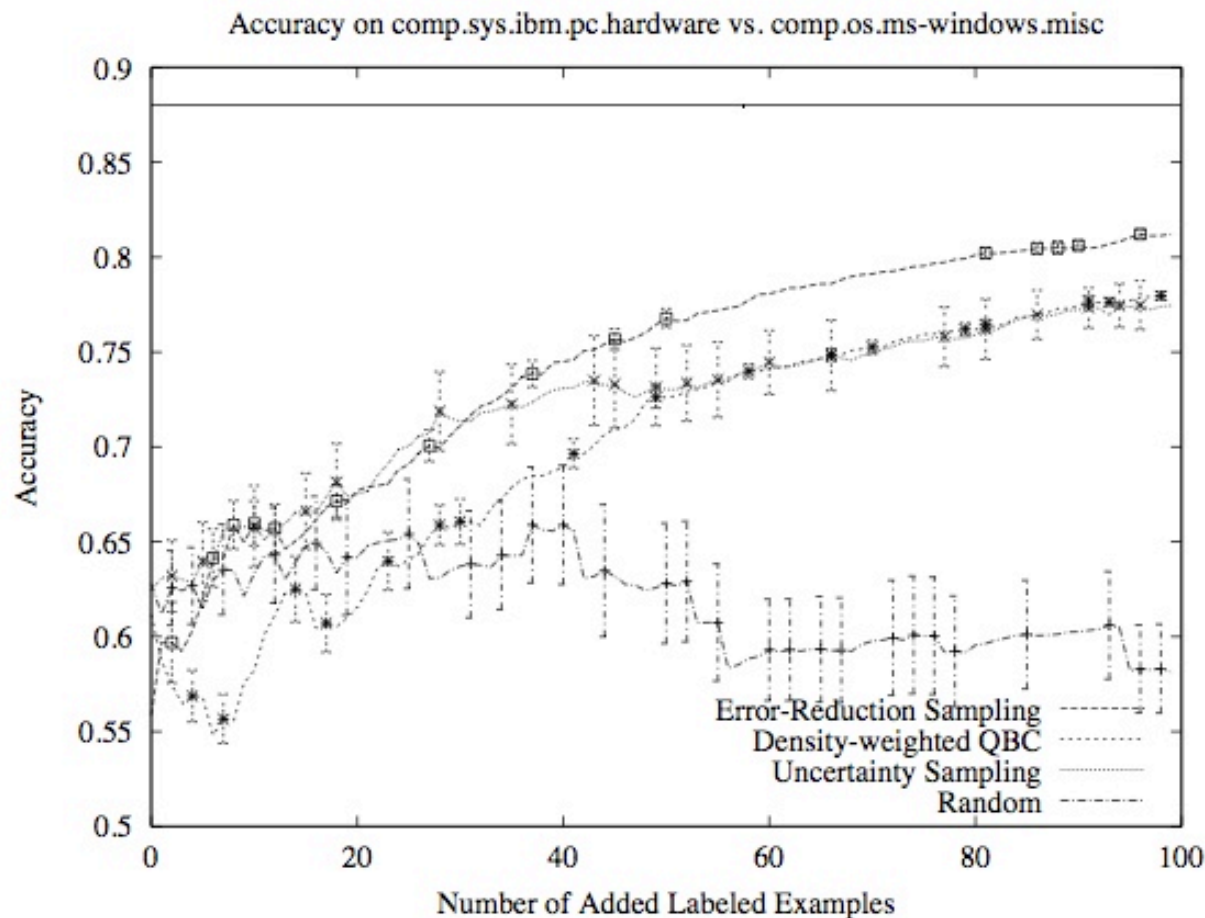
Diagram illustrating the components of the risk function $R(x)$:

- expectation over possible labelings of x** : Points to the E_y term.
- sum over unlabeled instances**: Points to the $\sum_{u \in \mathcal{U}}$ term.
- uncertainty of u after retraining with x** : Points to the $H_{\theta + \langle x, y \rangle} (Y | u)$ term.

Text Classification Examples



Text Classification Examples



Relationship to Uncertainty Sampling

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

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

Annotations:

- uncertainty *before* query (points to $\sum_{u \in \mathcal{U}} H_{\theta}(Y|u)$)
- risk term (points to $E_y \left[H_{\theta + \langle x, y \rangle}(Y|u) \right]$)
- assume x is representative of $\mathcal{U}$ (points to the approximation $\approx H_{\theta}(Y|x)$)
- assume this evaluates to zero (points to the expectation term in the second line)

...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]

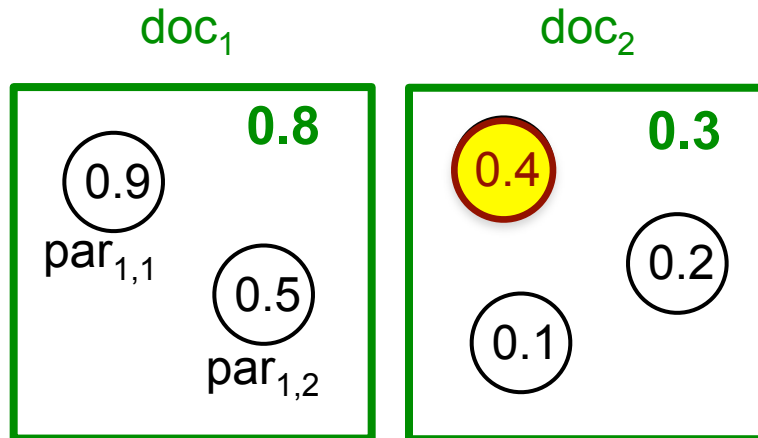
MI Active Learning

- 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

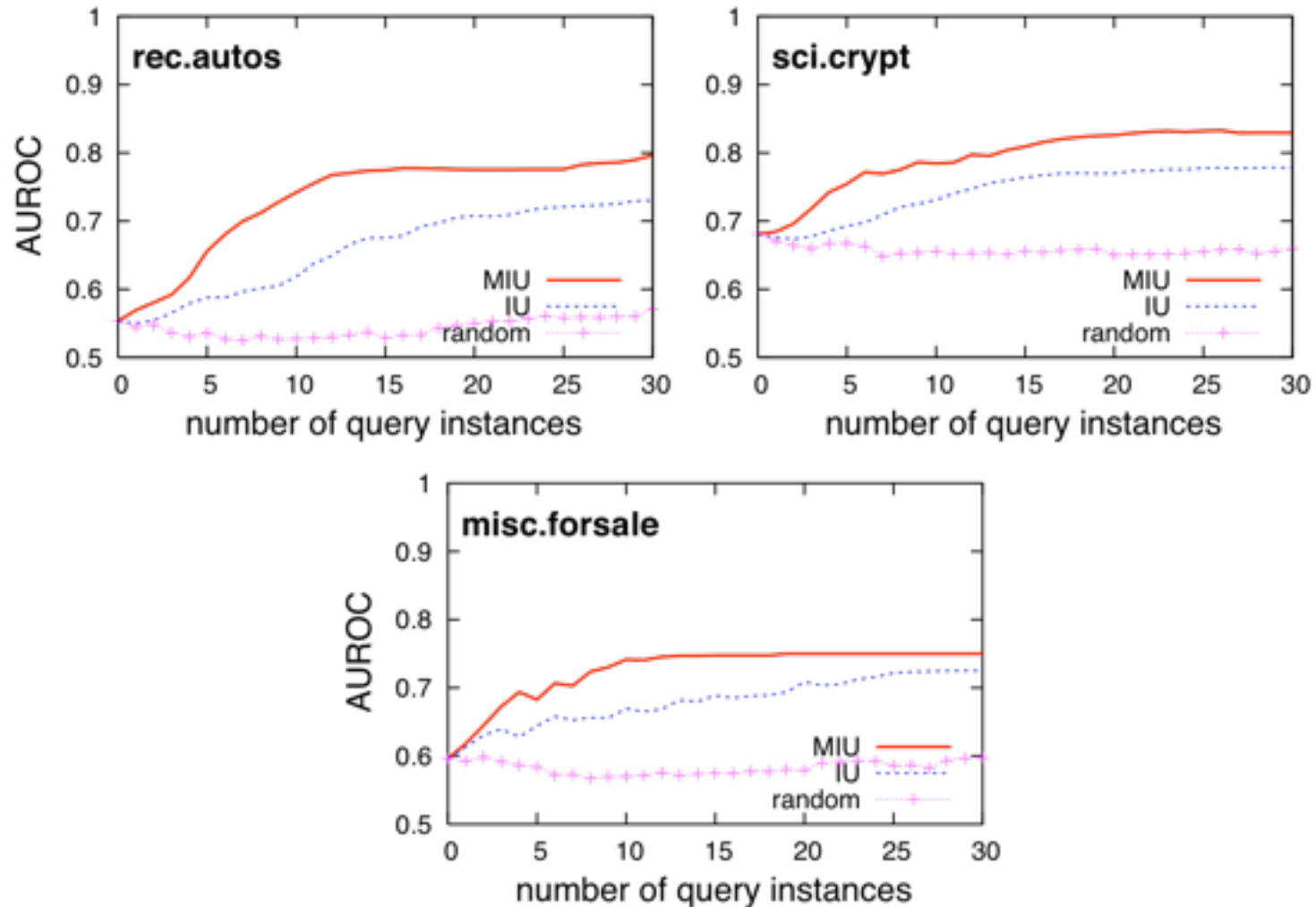
MI Uncertainty (MIU)

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

$$\phi_{MIU}(x_n) = \underbrace{H_{\theta}(Y|x_n)}_{\text{“base” uncertainty}} \underbrace{\left(\frac{\partial o}{\partial o_n}\right)^{\beta}}_{\text{“relevance” term}}$$



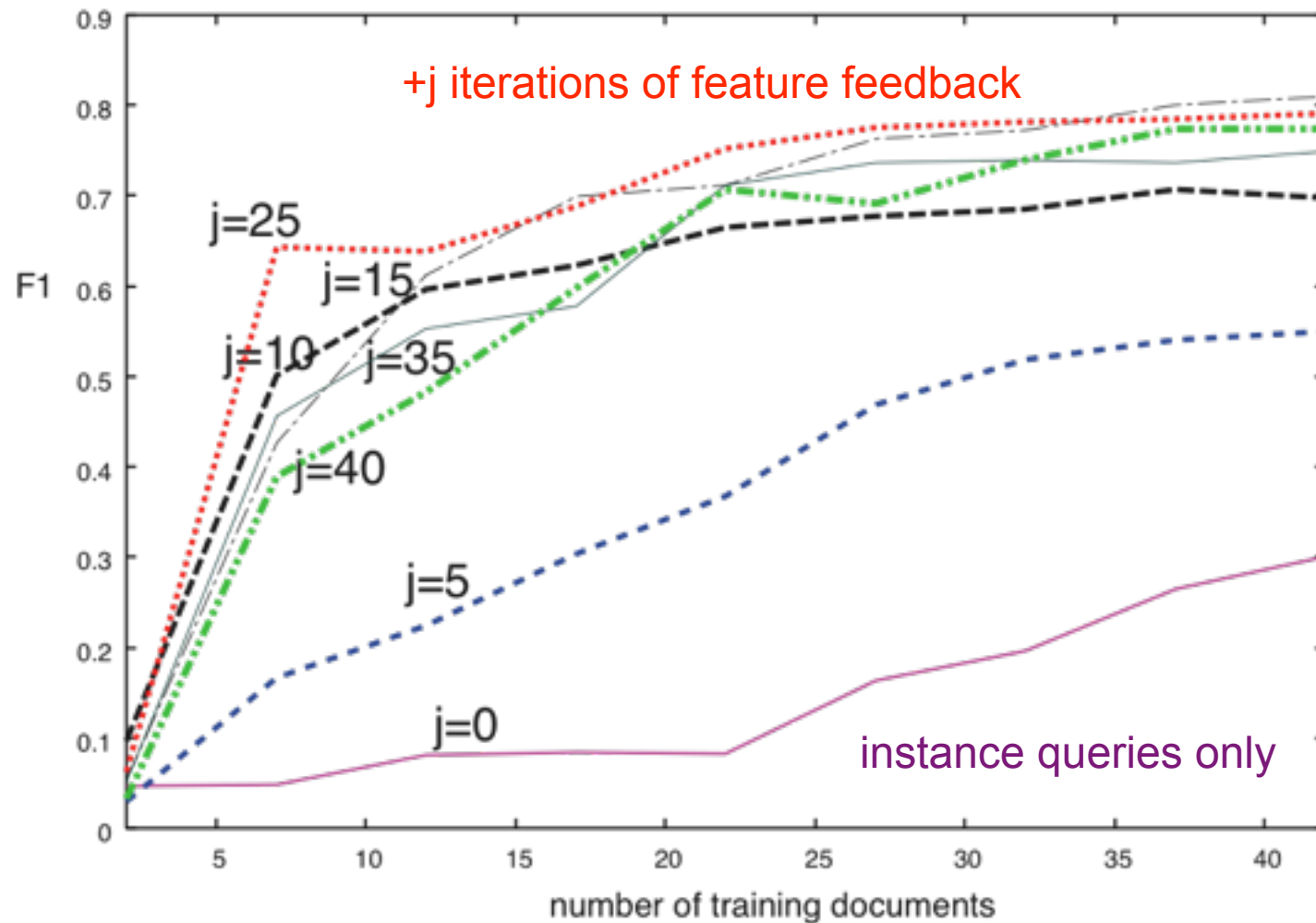
MI Active Learning Results



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?”

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

Lg 1 Bdrm No long-term lease

http://madiso

NAACL... The Ne... NAACL... http...bib Lg 1 ...

\$490 / 1br - Lg 1 Bdrm No long-term lease (Poynette)

[best of craigslist](#)

Reply to: hous-hnxwu-1198961460@craigslist.org [Errors when replying to ads?]
Date: 2009-05-31, 10:08PM CDT

Large one bedroom apartment in Poynette, only 20 mintues to east side of Madison. \$490 /month includes heat, water, and appliances. Large back yard for gardening or grilling. No long-term lease. Available now. Non-smoking building. 608-576-1734.

- cats are OK - purrr
- Location: Poynette
- it's NOT ok to contact this poster with services or other commercial interests

PostingID: 1198961460

Done

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

Weighted Uncertainty for Features

Diagram illustrating the formula for weighted uncertainty for features, $\phi_{WU}(f_k)$, with annotations:

- sum over tokens**: points to the summation $\sum_t$.
- does the token x_t have this feature? (0,1)**: points to the feature function $f_k(\mathbf{x}_t)$.
- uncertainty of token**: points to the conditional entropy $H_\theta(Y|\mathbf{x}_t)$.
- count of feature occurrences in corpus**: points to the denominator C_k .

$$\phi_{WU}(f_k) = \log(C_k) \frac{\sum_t f_k(\mathbf{x}_t) \times H_\theta(Y|\mathbf{x}_t)}{C_k}$$

a form of density-weighted uncertainty sampling

“Grid” Labeling Interface

Feature Active Learning GUI

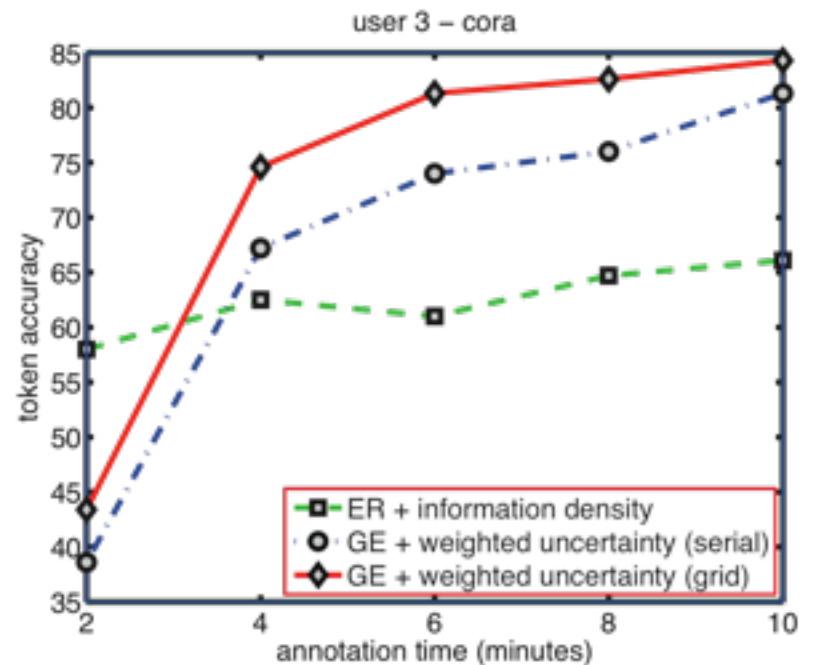
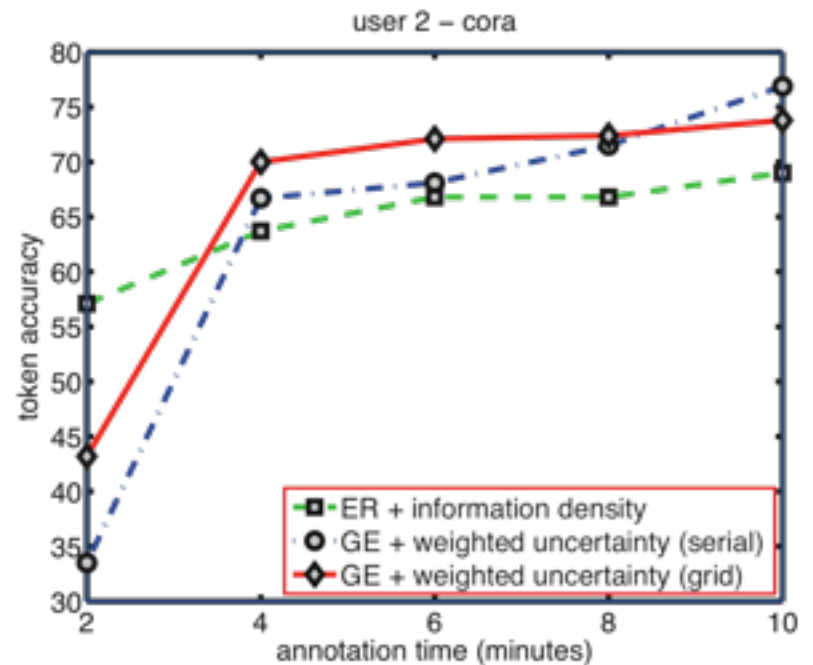
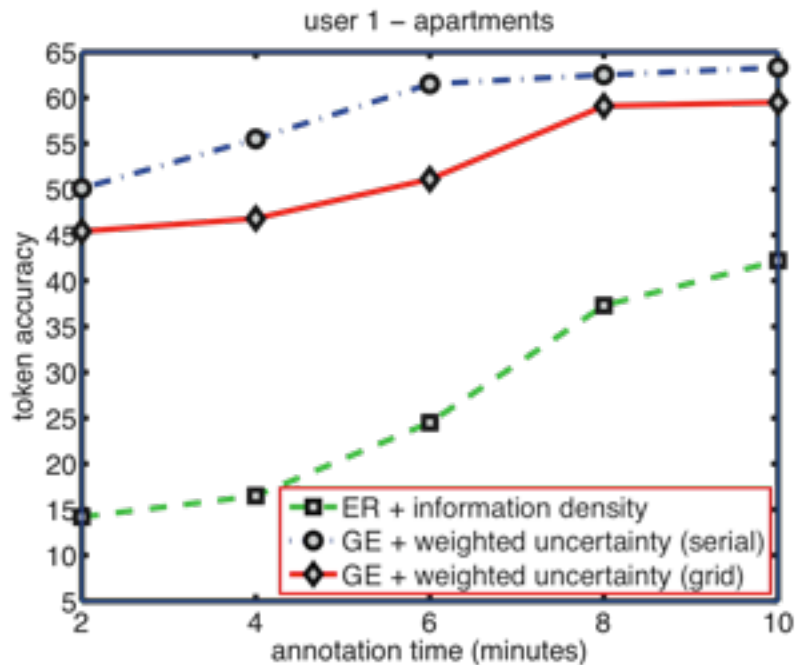
author (0)	title (0)	editor (0)	booktitle (0)	date (0)	journal (0)	volume (0)
tech (0)	institution (0)	pages (0)	location (0)	publisher (0)	note (0)	O (0)
FEATURE=CONTAINS.DOTS FEATURE=PUNC FEATURE=ALL.CAPS FEATURE=SINGLECHAR FEATURE=CAPLETTER	FEATURE=stopwords WORD=Data WORD=The WORD=An WORD=system	FEATURE=INITCAP FEATURE=namesuffixes FEATURE=ssdLprlastmed FEATURE=ssdLprfirstmed FEATURE=ssdLprfirsthigh	FEATURE=ssdLprlastlow WORD=New WORD=York FEATURE=nicknames FEATURE=ssdLprfirsthighes	publisher 0.952 author 0.912 editor 0.908 location 0.887 institution 0.885 volume 0.894 journal 0.894 title 0.884 date 0.882 booktitle 0.881 tech 0.881 pages 0.881 O 0.880 note 0.880		
FEATURE=name-particles WORD=de WORD=van FEATURE=ssdLprfirstlow WORD=PhD (0.88)	WORD=data (title) WORD=programs (title) WORD=flow (title) WORD=parallel (title) WORD=language (title)	WORD=learning (title) WORD=problems (title) WORD=logis (title) WORD=communication (title) WORD=models (title)	WORD=based (title) WORD=time (title) WORD=memory (title) WORD=shared (title) WORD=distributed (title)			
WORD=Systems WORD=Information WORD=Processing WORD=Engineering WORD=Software	WORD=Verlag WORD=Springer WORD=San WORD=CA WORD=Cambridge	WORD=TR (0.88) WORD=CS WORD=AAAI (booktitle) WORD=Research WORD=Computing	WORD=approach (title) WORD=polynomials (title) WORD=knowledge (title) WORD=theory (title) WORD=local (title)			
WORD=networks WORD=systems WORD=databases WORD=reasoning WORD=algorithms	WORD=thesis (0.88) WORD=editors (0.88) WORD=Tech (0.88) WORD=Planning WORD=Kaufmann	WORD=decision WORD=programming FEATURE=honorifics WORD=method WORD=cognition	WORD=Logic WORD=Languages WORD=Design WORD=Distributed WORD=Parallel			
WORD=order WORD=global WORD=recursive WORD=case WORD=code			WORD=dependence WORD=algorithm WORD=neural WORD=al WORD=Real			

time remaining: 54 seconds.

11th International Joint Summer School, pp. 335 - 342. Morgan Kaufmann, San Mateo, California.

Artificial Intelligence (AAAI - 93), pp. 492 - 500. Morgan Kaufmann, Inc., Los Alamitos, CA.

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

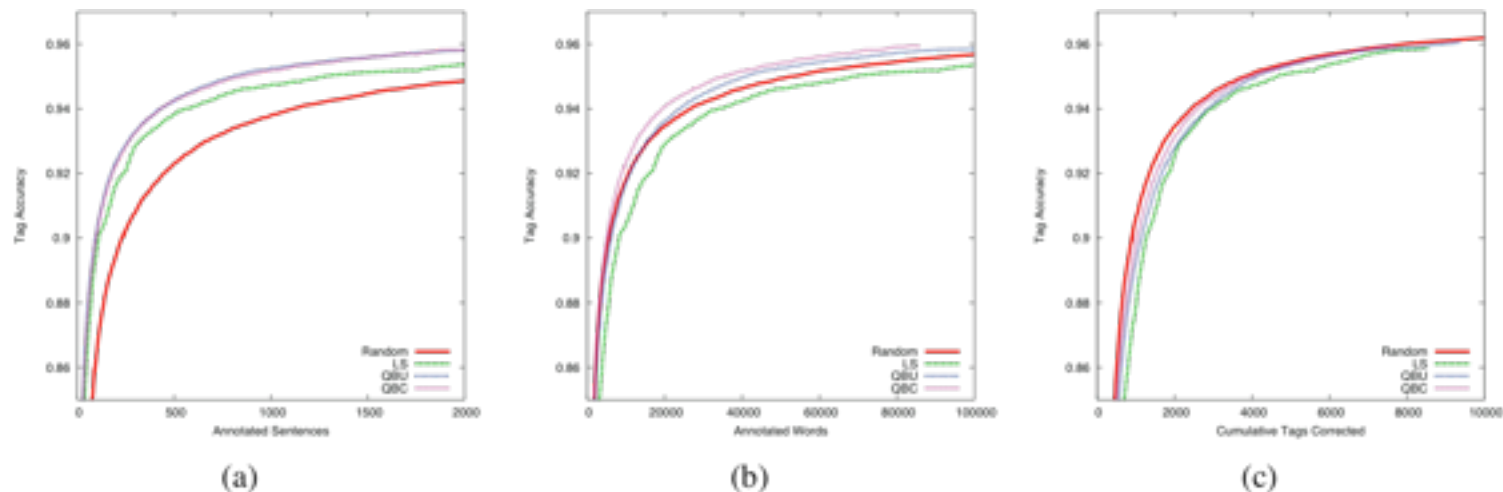
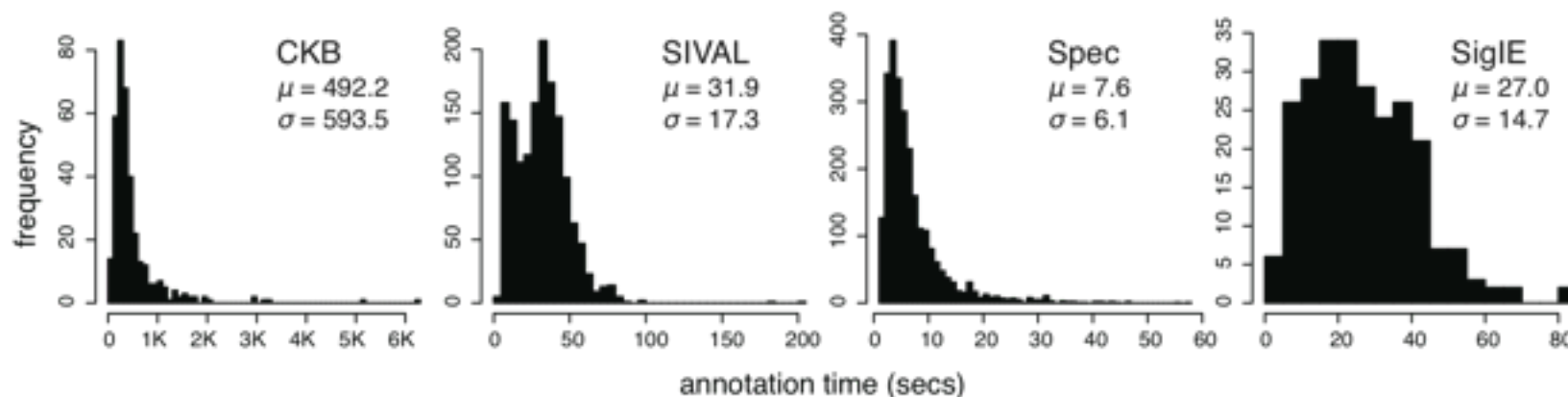


Figure 1: QBU, LS, QBC, and the random baseline plotted in terms of accuracy versus various cost functions: (a) number of sentences annotated; (b) number of words annotated; and (c) number of tags corrected.

[Haertel et al., ACL'08]

Annotation Time As Cost

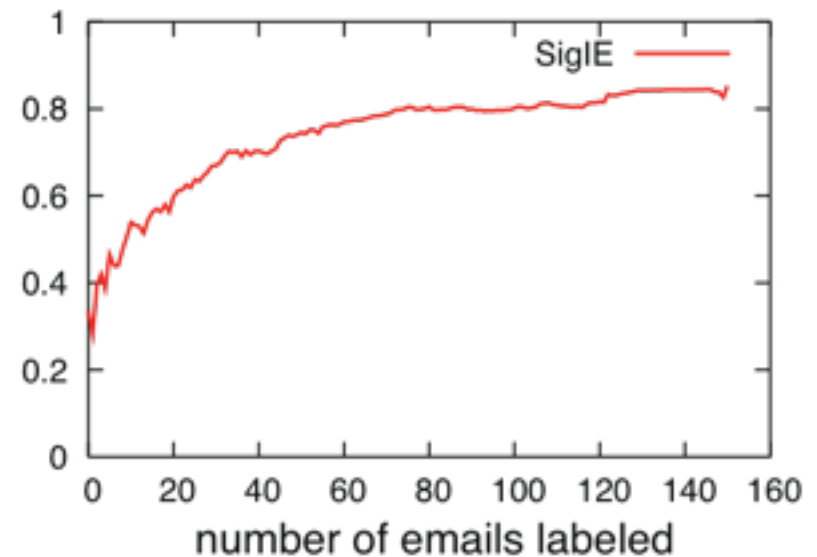
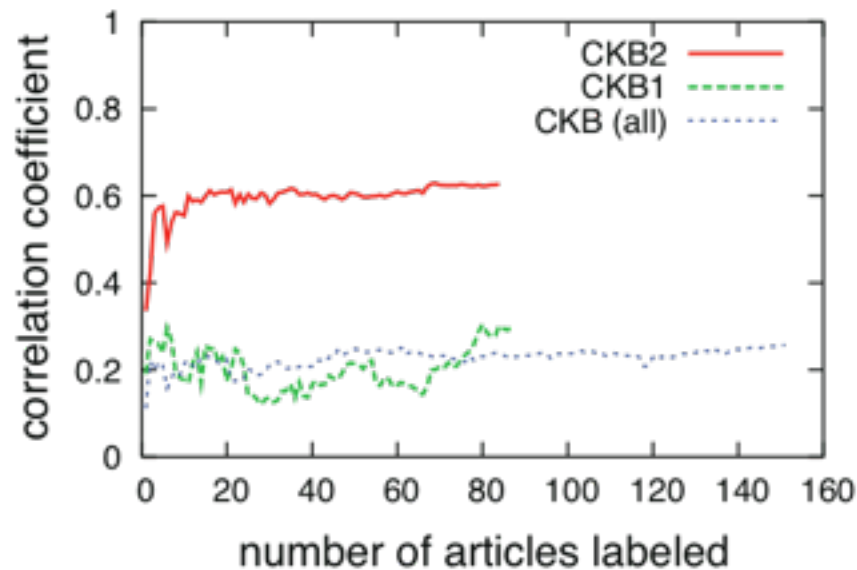
- do annotation times vary among instances?



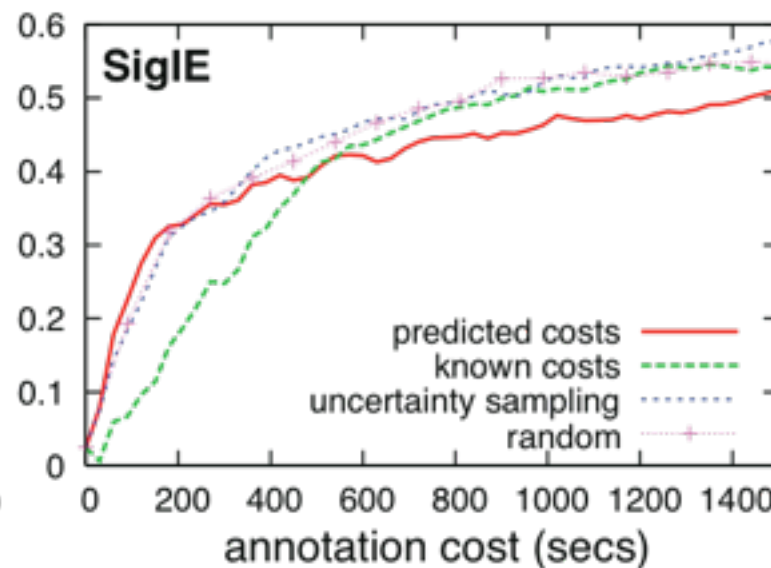
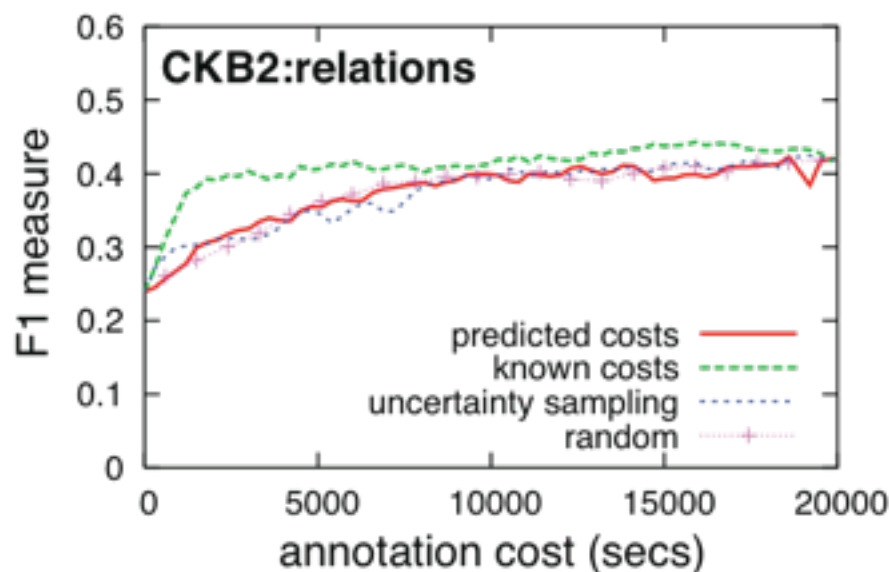
- where does this variance come from?
 - sometimes annotator-dependent
 - stochastic effects

Can Labeling Times be Predicted?

cost predictor: regression model using meta-features



Can Predicted Times Improve AL?



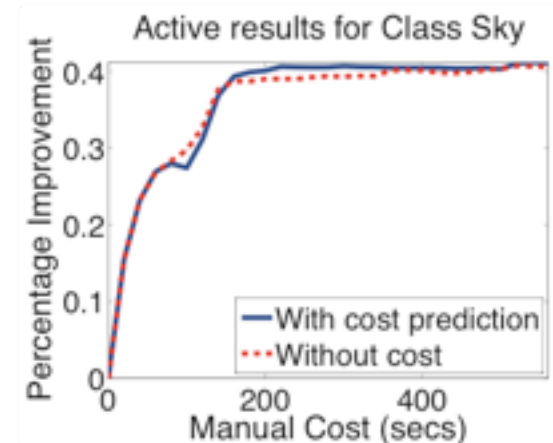
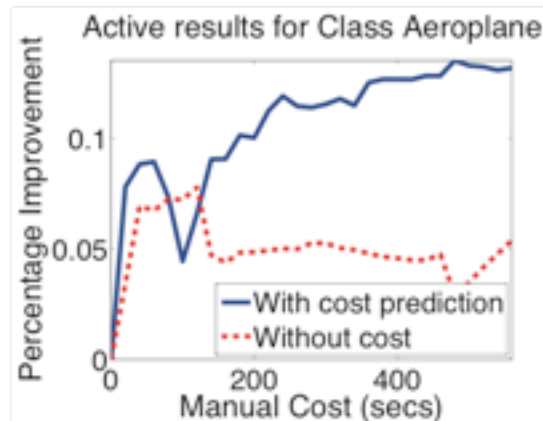
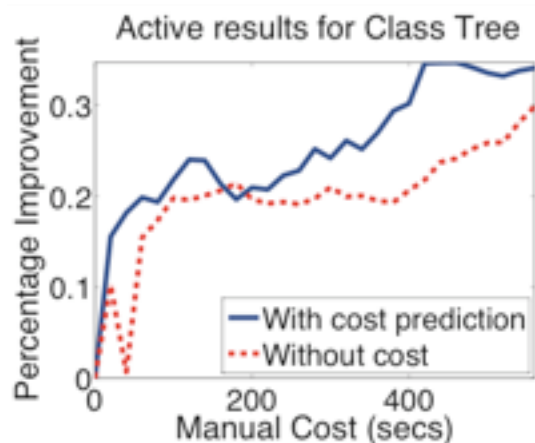
$$\phi(x) = \frac{H_{\theta}(Y|x)}{\hat{\$}(x)}$$

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 [Baldrige & 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?