Active Learning

Burr Settles

Machine Learning 10-701 / 15-781
Nov 14, 2012

Let’s Play 20 Questions!

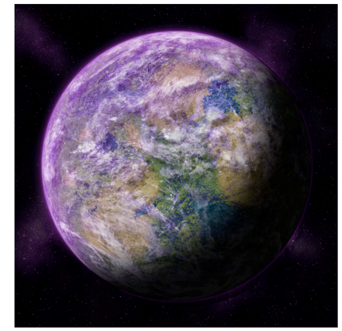
- I’m thinking of something; ask me yes/no questions to figure out what it is…

How Do We *Automate* Inquiry?

A Though Experiment

A Thought Experiment

- suppose you are on an Earth convoy sent to colonize planet Zelgon



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

people who ate the rough Zelgian fruits found them ***gross****!*

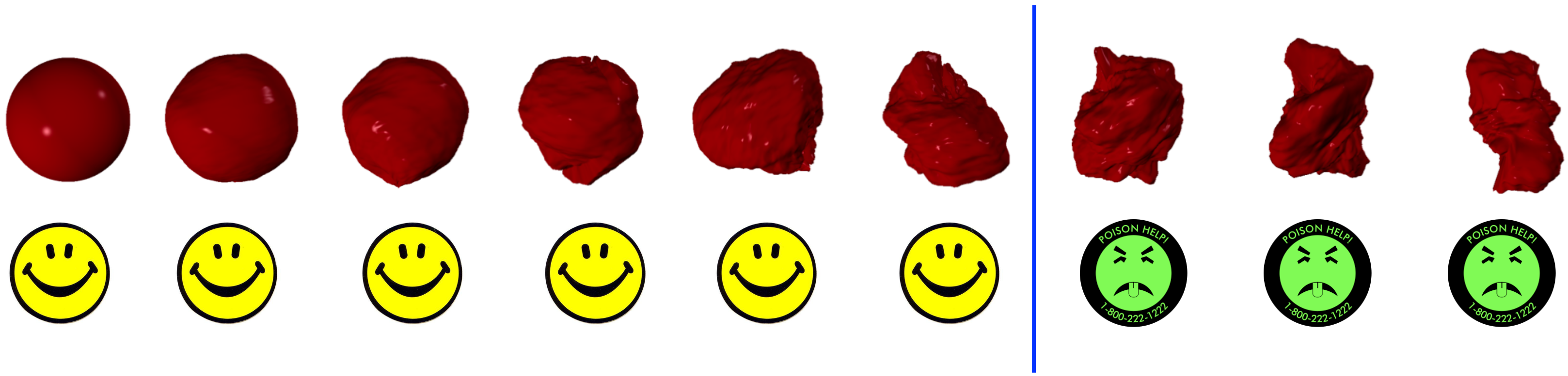
Poisonous *vs.* Yummy Alien Fruits

- there is a continuous range of round-to-rough fruit shapes on Zelgon:

you need to learn how to classify fruits as
safe or noxious

and you need to do this while risking as
little as possible (i.e., colonist health)

Supervised Learning Approach

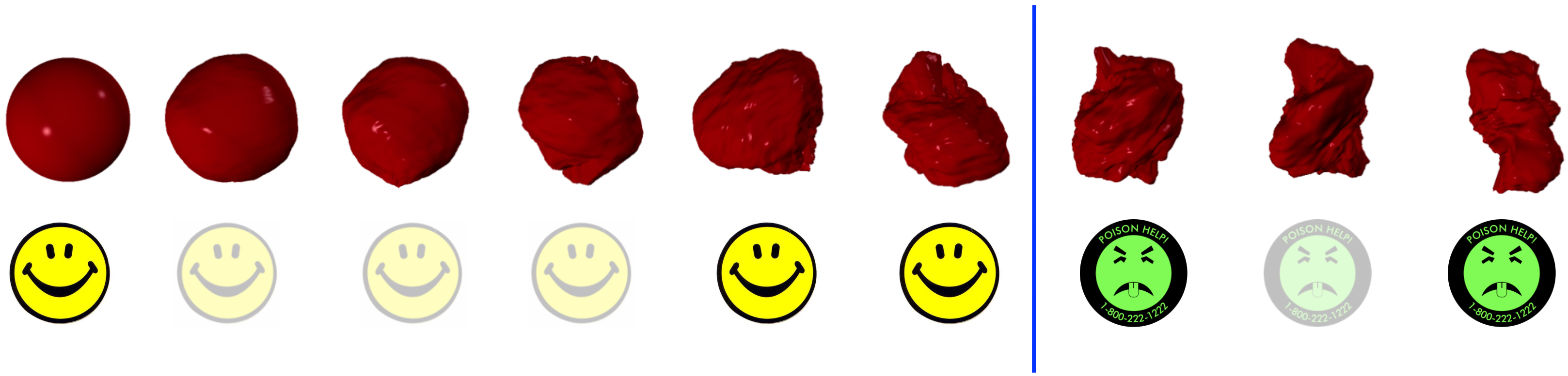


problem:

PAC theory tells us we need $O(1/\varepsilon)$ tests
to obtain an error rate of ε…

a lot of people might get sick in the process!

Can We Do Better?

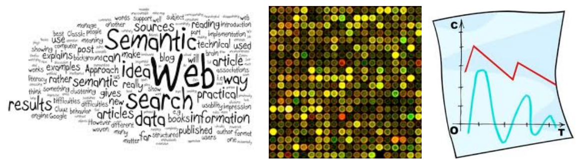


this is just a **binary search**…

requiring $O(1/\varepsilon)$ fruits (e.g., samples)
but only $O(\log_2 1/\varepsilon)$ tests (e.g., queries)

our first "active learning" algorithm!

Supervised Learning



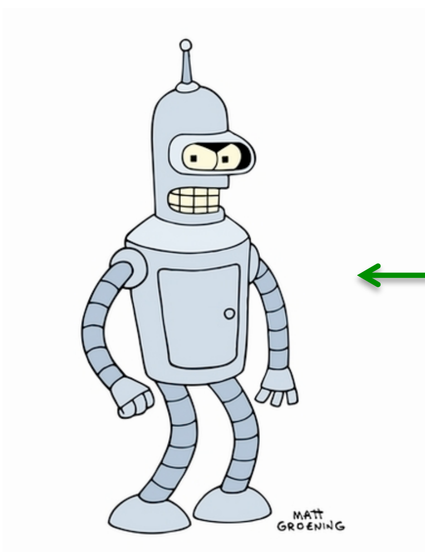
raw unlabeled data

$x_1, x_2, x_3, \ldots$

random sample

labeled training instances

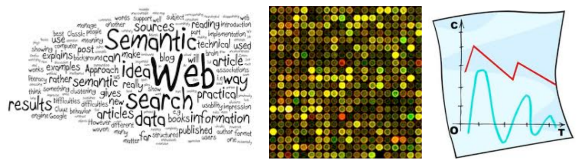
$\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \langle x_3, y_3 \rangle, \ldots$



supervised learner
induces a classifier

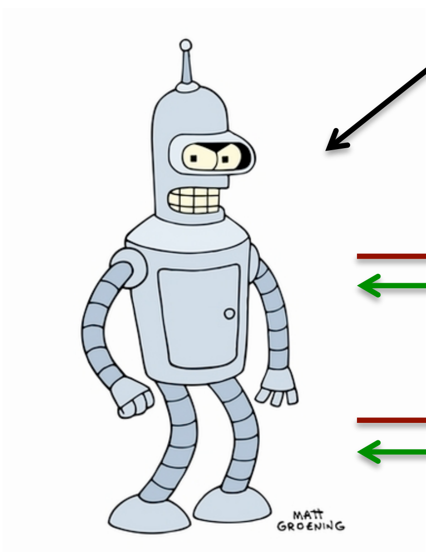
expert / oracle
analyzes experiments
to determine labels

Active Learning



raw unlabeled data
$x_1, x_2, x_3, \ldots$

inspect the
unlabeled data



request labels for selected data

$\langle x_1, ? \rangle$

$\langle x_1, y_1 \rangle$

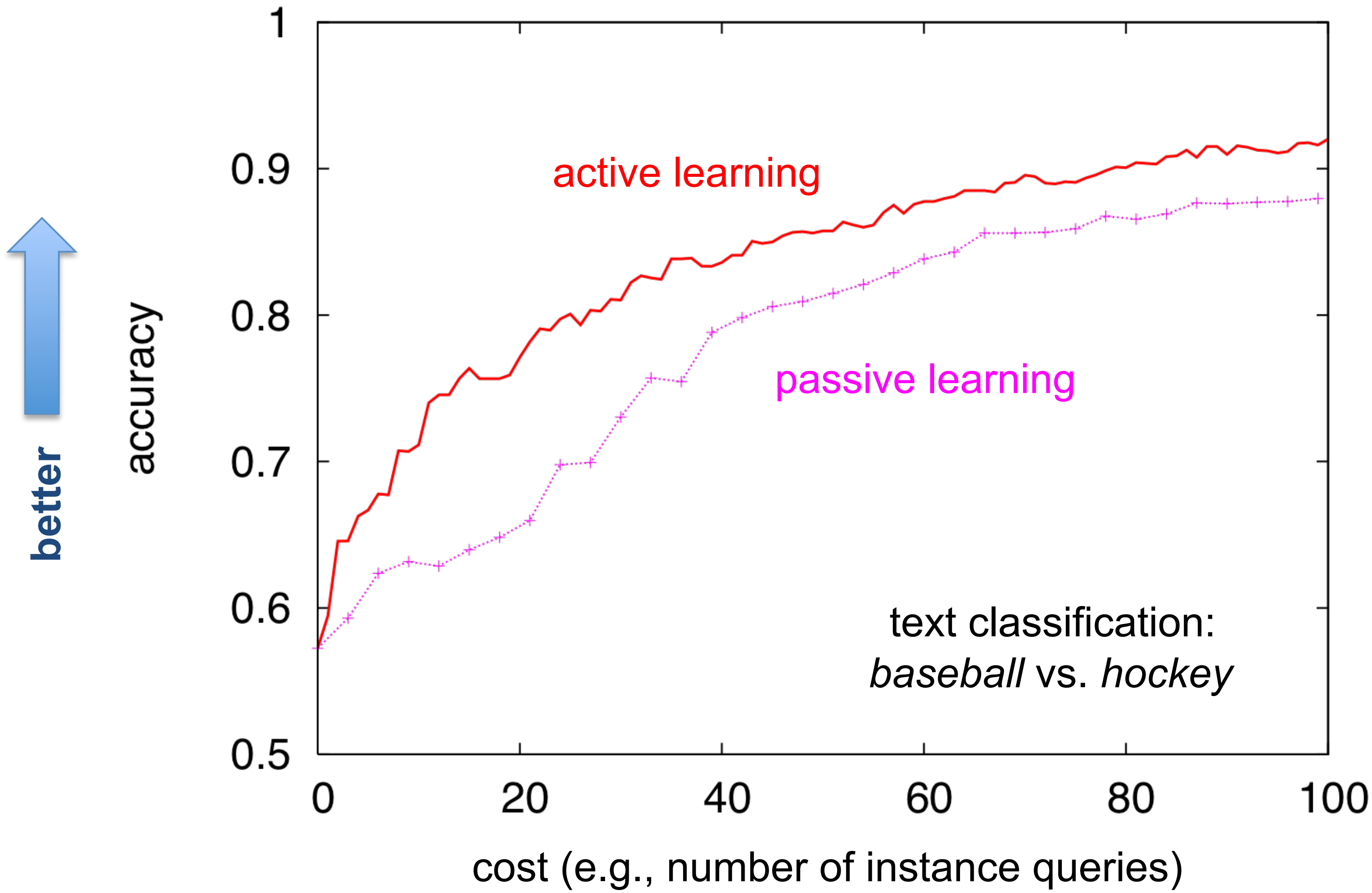
$\langle x_2, ? \rangle$

$\langle x_2, y_2 \rangle$

active learner
induces a classifier

expert / oracle
analyzes experiments
to determine labels

Learning Curves



Who Uses Active Learning?

IBM

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

SIEMENS

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

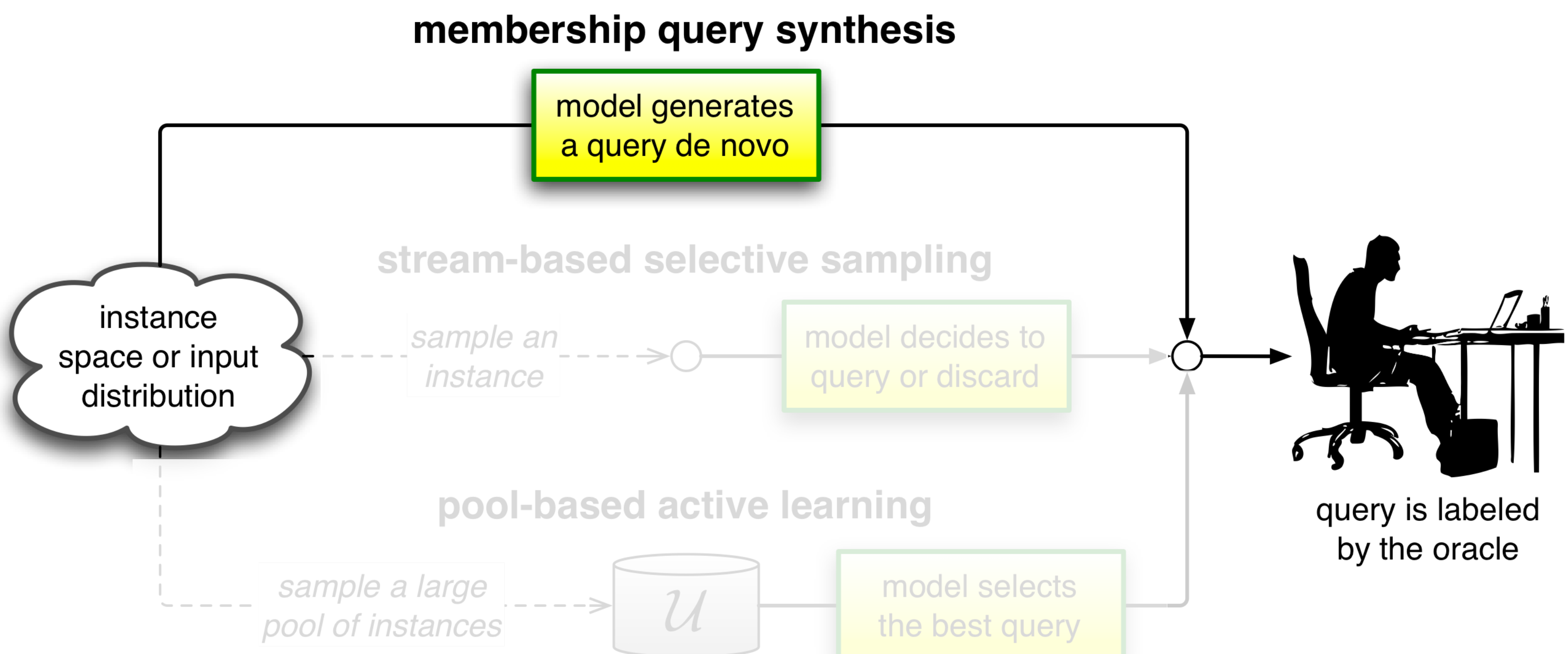
Microsoft

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

Google

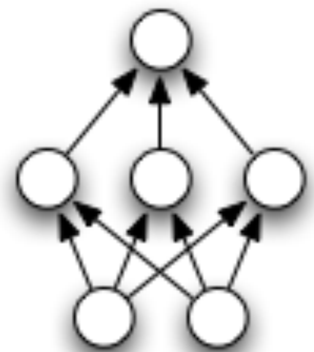
"While I can confirm that we're using active learning in earnest on many problem areas… I really can't provide any more details than that. Sorry to be so opaque!"
– *David Cohn*

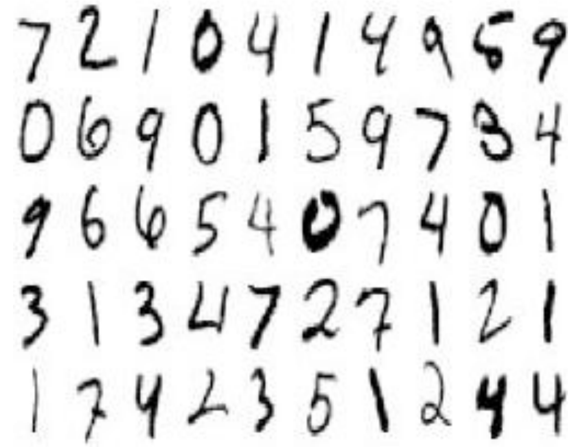
Active Learning Scenarios

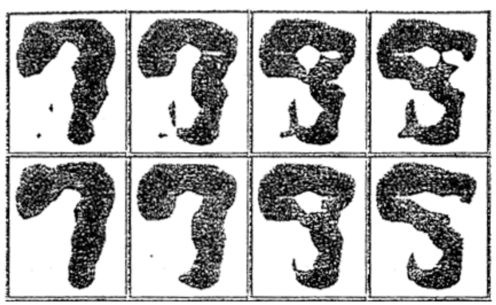


Problems with Query Synthesis

an early real-world application: neural-net queries synthesized for handwritten digits [Lang & Baum, 1992]



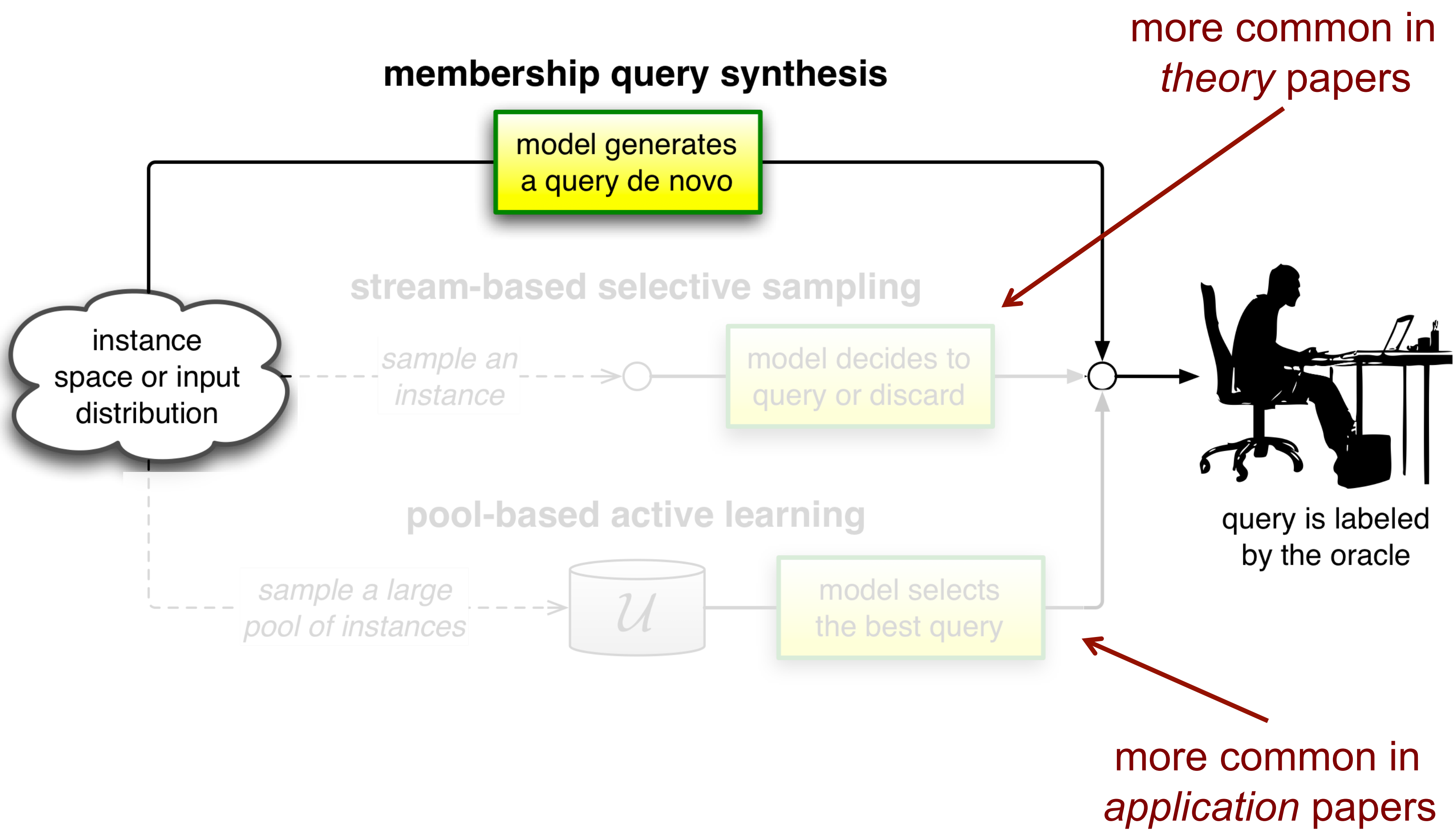




problem: humans couldn't interpret the queries!

ideally, we can ensure that the queries come from the underlying "natural" distribution

Active Learning Scenarios

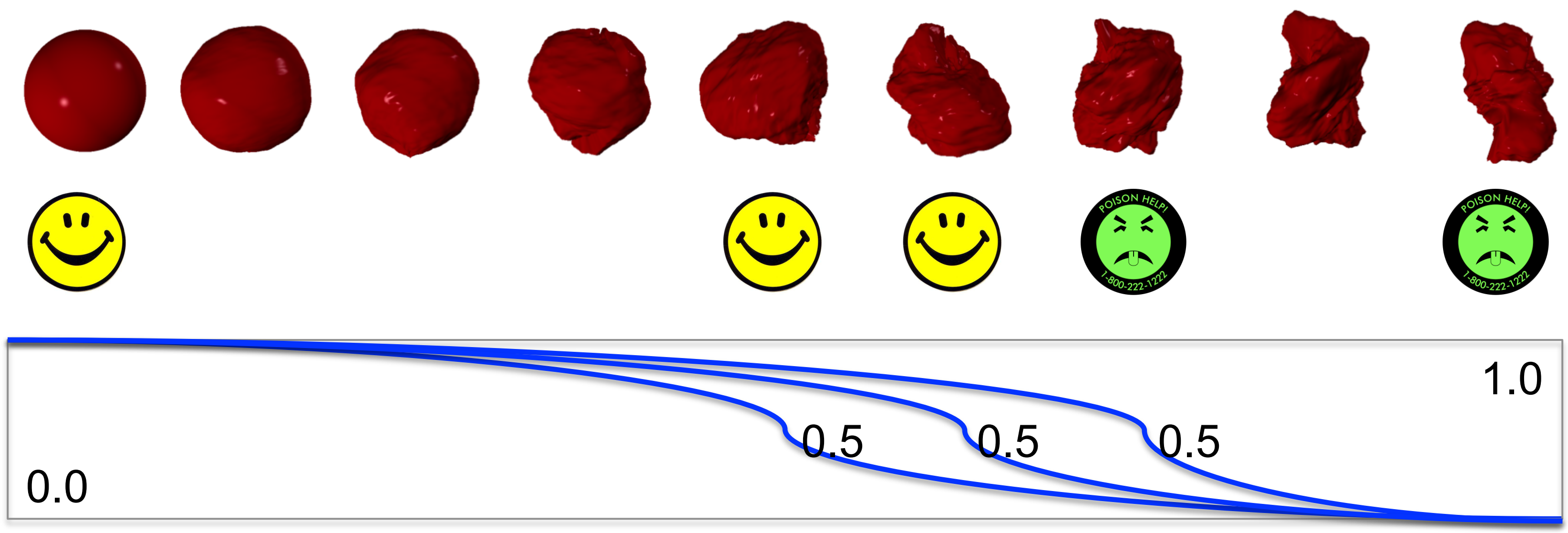


Active Learning Approaches

(1) Uncertainty Sampling

Zelgian Fruits Revisited

- let's interpret our Zelgian fruit binary search in terms of a *probabilistic* classifier:

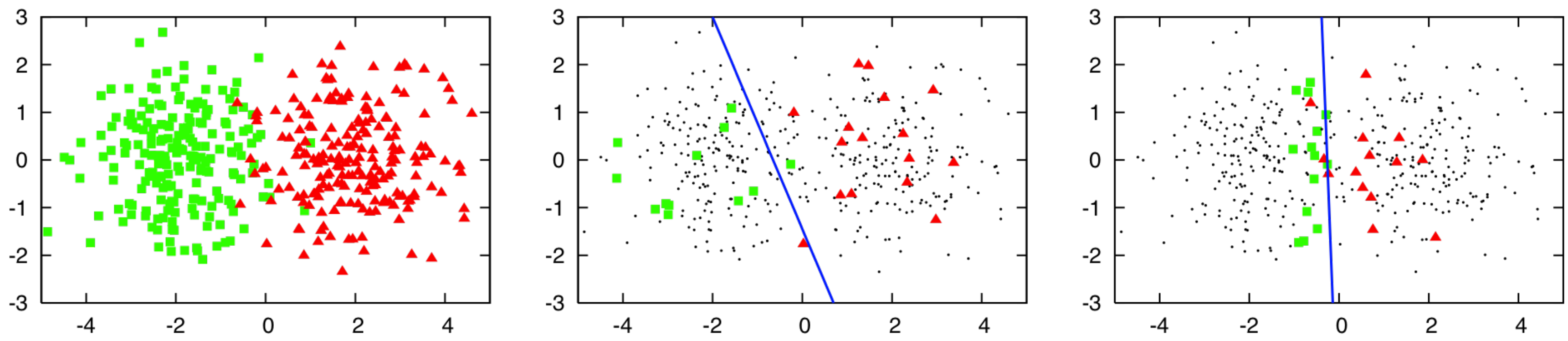


$P(Y = ☺ | X)$

[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)

uncertainty sampling
30 labeled instances (accuracy=0.9)

Common Uncertainty Measures

least confident

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

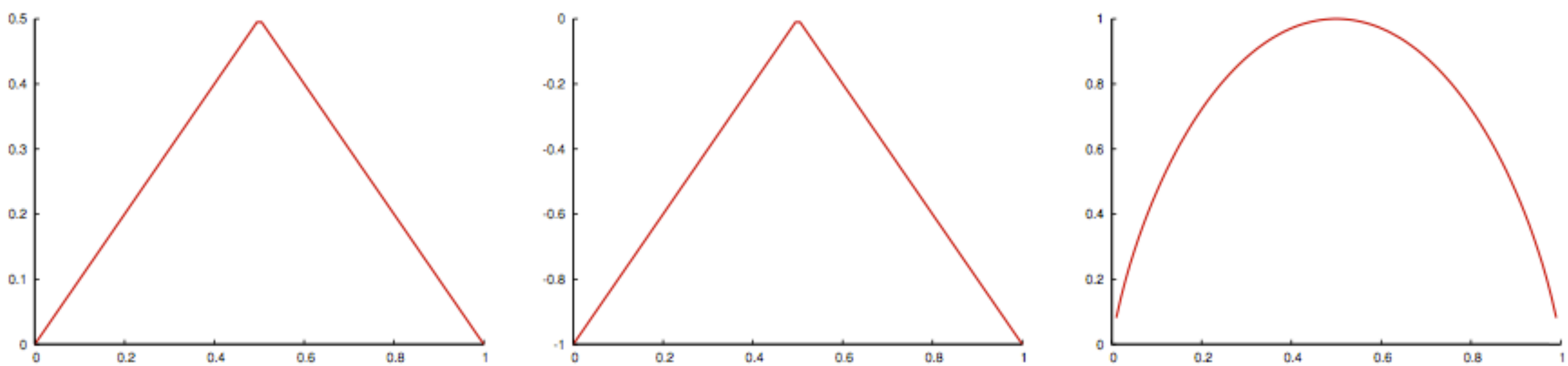
margin

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

entropy

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

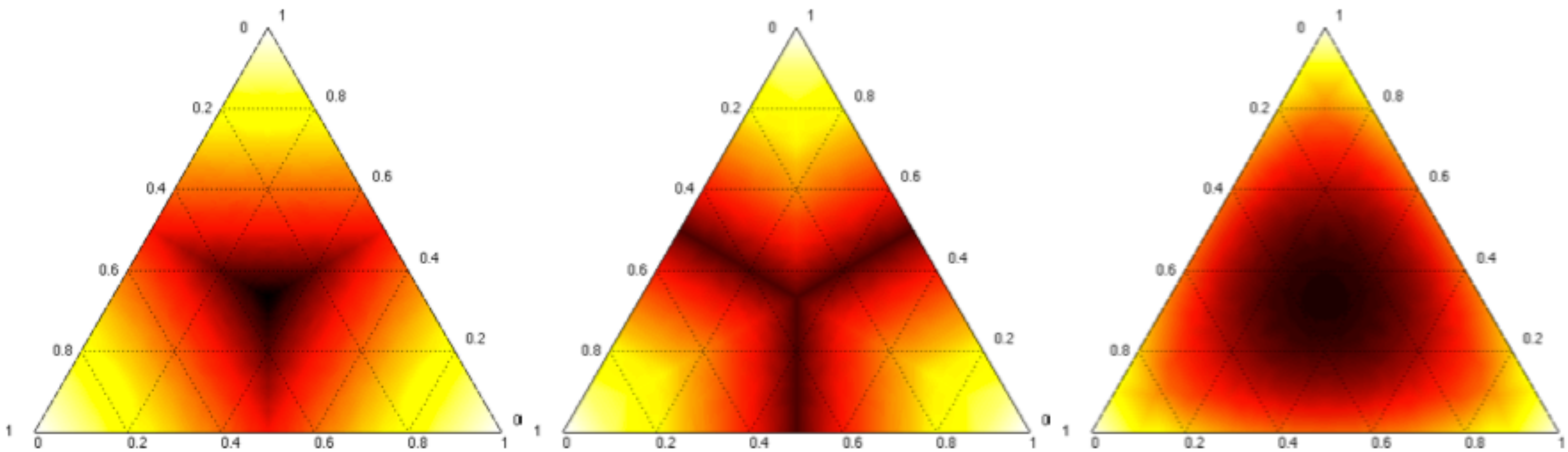
Common Uncertainty Measures



(a) least confident – binary (b) margin – binary (c) entropy – binary

note: for binary tasks, these are functionally equivalent!

Common Uncertainty Measures



(d) least confident – ternary (e) margin – ternary (f) entropy – ternary

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

note: for multi-class tasks, these are *not* equivalent!

Information-Theoretic Interpretation

- the "surprisal" $\mathcal{I}$ is a measure (in bits, nats, etc.) of the information content for outcome y of variable Y:

$$\mathcal{I}(y) = \log \frac{1}{P(y)} = -\log P(y)$$

- so this is how "informative" the oracle's label y will be
- but the learner doesn't know the oracle's answer yet! we can estimate it as an *expectation* over all possible labels:

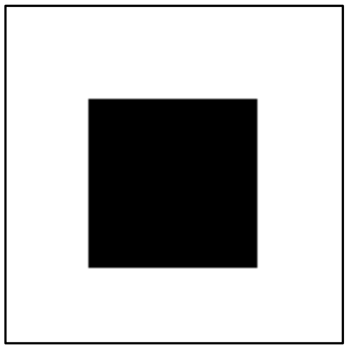
$$E_y\left[-\log P_\theta(y|x)\right] = -\sum_y P_\theta(y|x) \log P_\theta(y|x)$$

- which is **entropy**-based uncertainty sampling

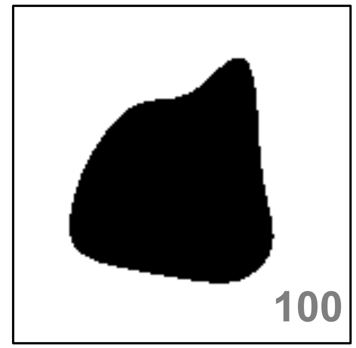
Uncertainty Sampling in Practice

- pool-based active learning:
  - evaluate each x in $\mathcal{U}$
  - rank and query the top K instances
  - retrain, repeat
- selective sampling:
  - threshold a "region of uncertainty," e.g., [0.2, 0.8]
  - observe new instances, but only query those that fall within the region
  - retrain, repeat

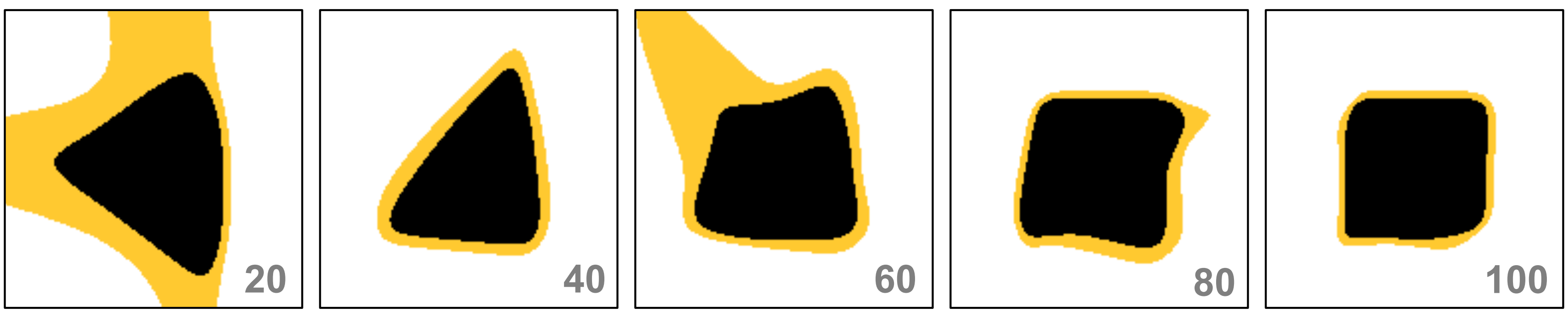
Uncertainty Sampling: Example



target function



neural net trained from
100 random pixels

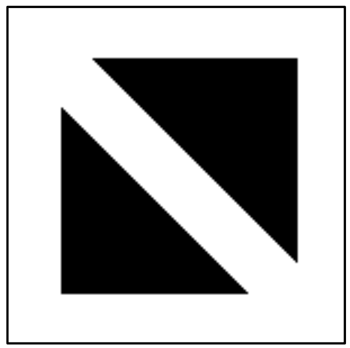


active neural net (stream-based uncertainty sampling)

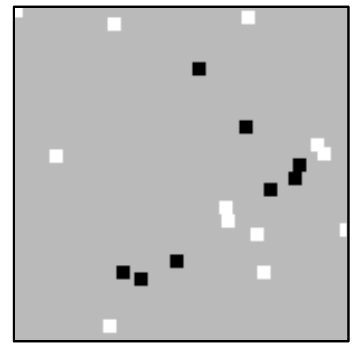
Simple and Widely-Used

- **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; Settles & 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

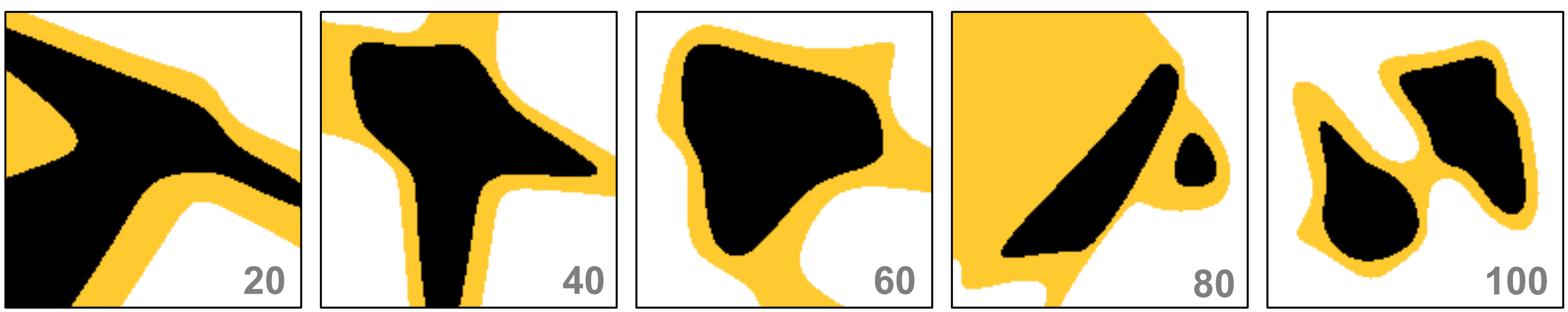
Uncertainty Sampling: Failure?!



target function



initial sample



active neural net (stream-based uncertainty sampling)

What To Do?

- uncertainty sampling only uses the confidence of *one single* classifier
  - e.g., a "point estimate" for parametric models
  - this classifier can become overly confident about instances is really knows nothing about!

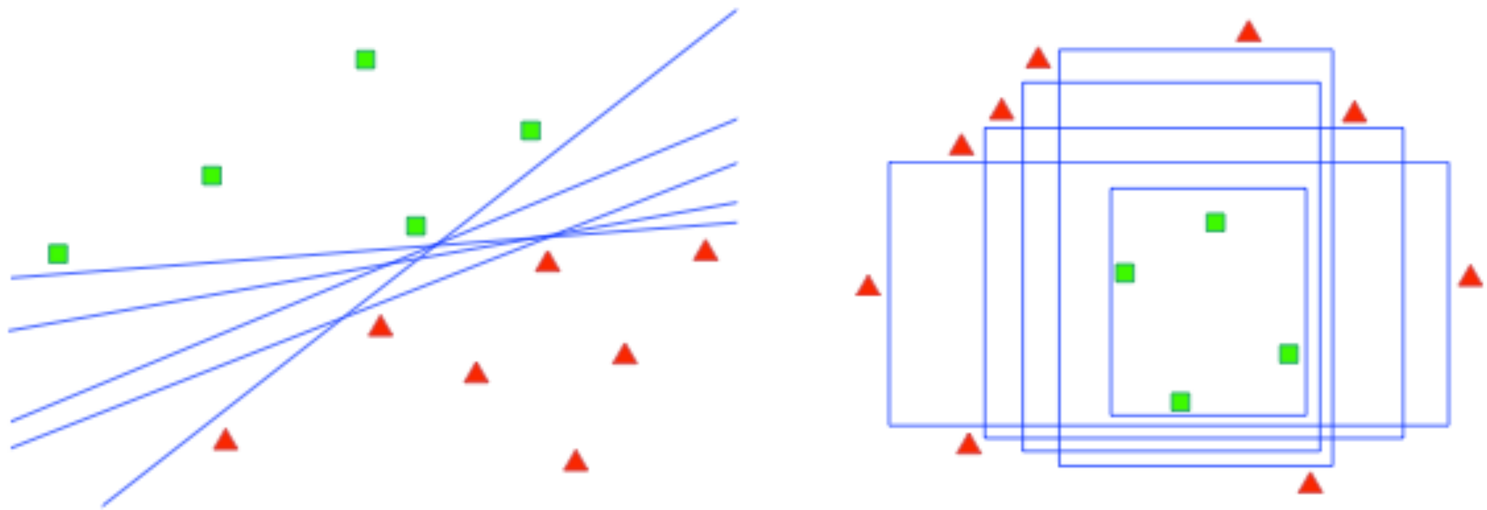
- instead, let's consider a different notion of "uncertainty"… about the *classifier itself*

Active Learning Approaches

(2) Hypothesis Space Search

Remember Version Spaces?

- the set of all classifiers that are consistent with the labeled training data

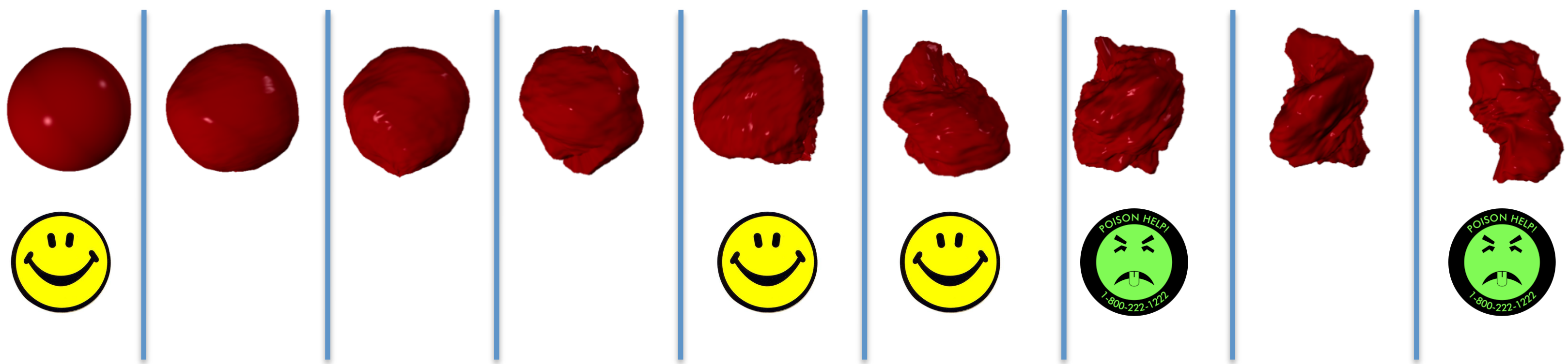


(a) (b)

- the larger the version space $\mathcal{V}$, the less likely each possible classifier is… we want queries to *reduce* $|\mathcal{V}|$

Alien Fruits Revisited

- let's try interpreting our binary search in terms of a version space search:



possible classifiers (thresholds): **1**

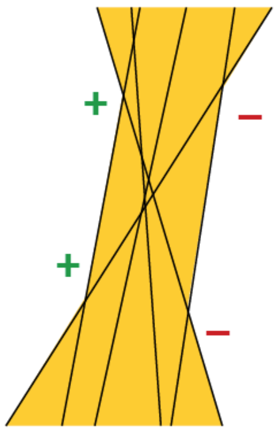
Version Space Search

- in general, the version space $\mathcal{V}$ may be too large to enumerate, or to measure the size $|\mathcal{V}|$ through analytical trickery

- **observation**: for the Zelgian fruits example, uncertainty sampling and version space search gave us *the same queries*!

- how far can uncertainty sampling get us?

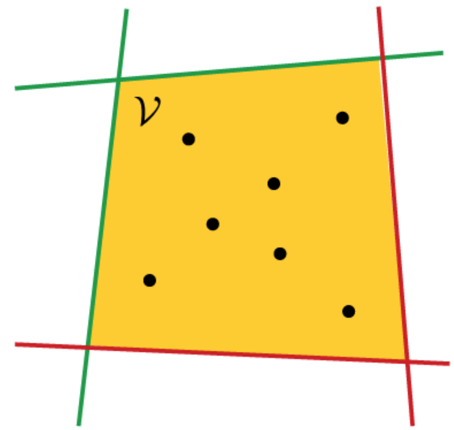
Version Spaces for SVMs



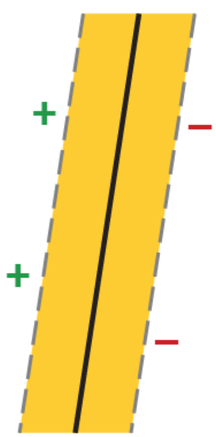
$\mathcal{F}$ (feature space)

"version space duality"
(Vapnik, 1998)

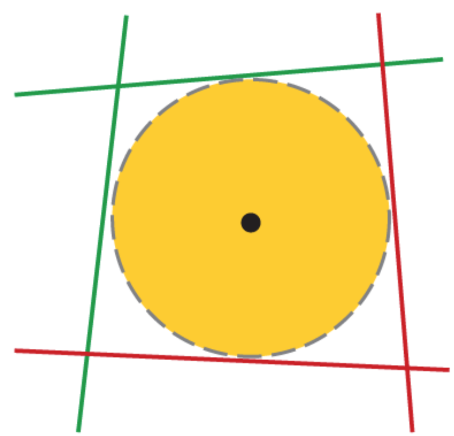
points in $\mathcal{F}$ correspond
to hyperplanes in $\mathcal{H}$
and *vice versa*



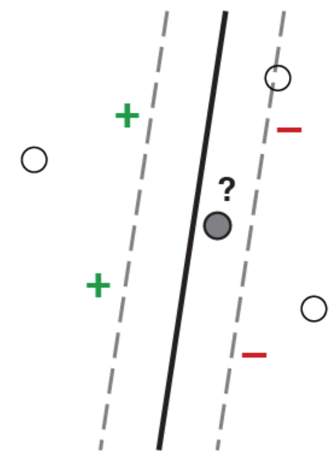
$\mathcal{H}$ (hypothesis space)



SVM with largest margin
is the center of the largest
hypersphere in $\mathcal{V}$

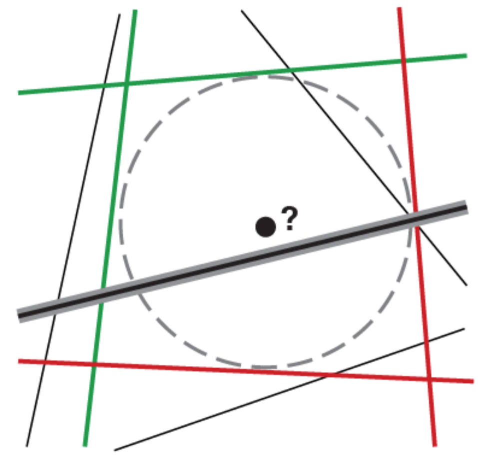


Bisecting the SVM Version Space



$\mathcal{F}$ (feature space)

points closest to the class boundary in $\mathcal{F}$ correspond to planes that bisect $\mathcal{V}$ in $\mathcal{H}$



$\mathcal{H}$ (hypothesis space)

- hence, uncertainty sampling is a special case of version space search for SVMs (and other so-called "max-margin" classifiers)

[Cohn et al., ML'94]

Query By Disagreement (QBD)

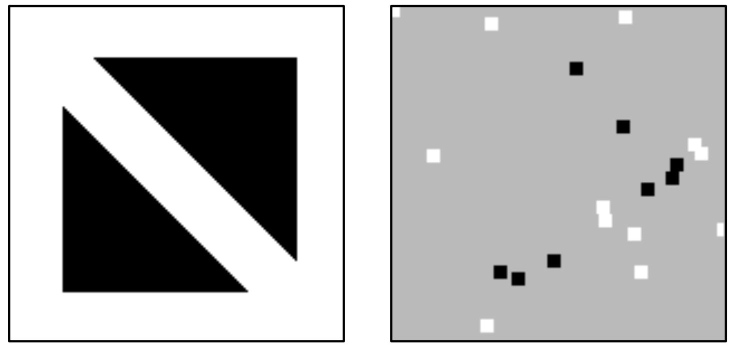
- in general, uncertainty doesn't cut it

- **idea**: we wish to quickly *eliminate bad hypotheses*; train two classifiers G and S which represent the two "extremes" of the version space

- if these two models disagree, the instances falls within the "region of uncertainty"

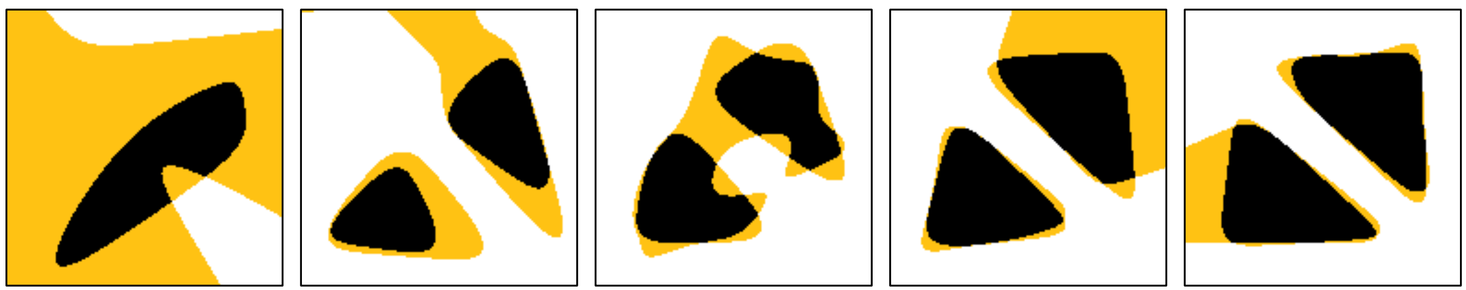
Neural Network Triangles Revisited

target function

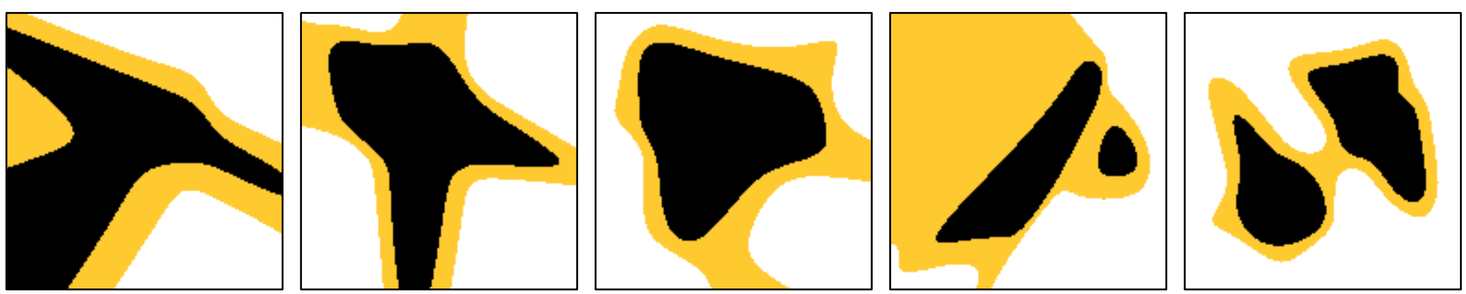


initial sample

QBD:



uncertainty sampling:



[Seung et al., COLT'92]

Query By Committee (QBC)

- simpler, more general approach

- train a committee of classifiers $\mathcal{C}$
  - no need to maintain G and S
  - committee can be any size

- query instances for which committee members disagree

QBC in Practice

- selective sampling:
  - train a committee $\mathcal{C}$
  - observe new instances, but only query those for which there is disacreement (or a lot of disagreement)
  - retrain, repeat
- pool-based active learning:
  - train a committee $\mathcal{C}$
  - measure disagreement for each x in $\mathcal{U}$
  - rank and query the top K instances
  - retrain, repeat

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., boosting, bagging)
    - [Abe & Mamitsuka, ICML'98]

- how to measure disagreement (many):
  - "XOR" committee classifications
  - view vote distributions as probabilities, use uncertainty measures…

QBC Disagreement Measures

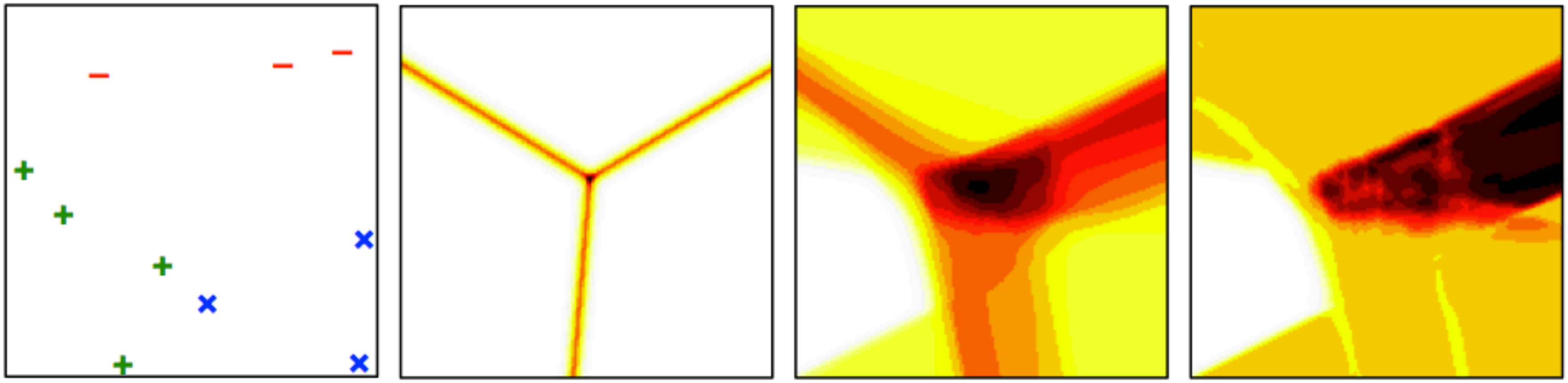
- “soft” vote entropy:

$$x^*_{SVE} = \underset{x}{\operatorname{argmax}} - \sum_y P_{\mathcal{C}}(y|x) \log P_{\mathcal{C}}(y|x)$$

- average Kullback-Liebler (KL) divergence:

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} KL(\ P_\theta(Y|x) \parallel P_{\mathcal{C}}(Y|x)\ )$$

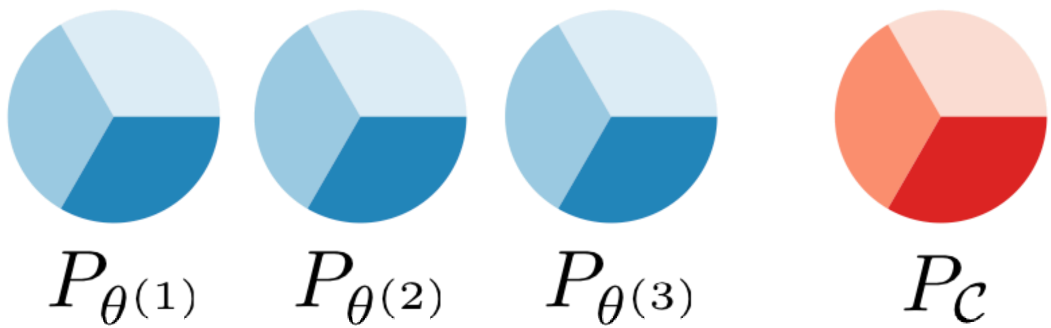
QBC Disagreement Measures



(a) training set (b) entropy (c) soft vote entropy (d) KL divergence

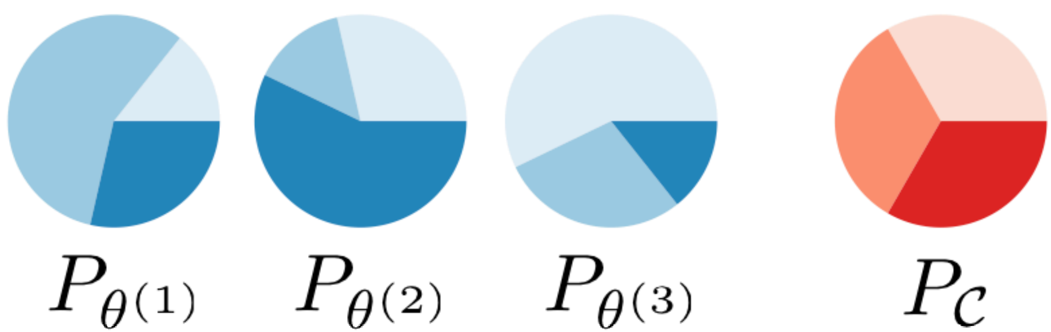
heatmaps illustrating query heuristics for a 3-label classification task using multinomial logistic regression (e.g., a MaxEnt model)

QBC Disagreement Measures



*un*certain hypotheses; but in agreement

SVE cannot tell either of these apart



confident hypotheses; but in *dis*agreement

KL divergence will query this

Information-Theoretic Interpretation

- we want to query the instance whose label contains maximal mutual information about the version space: $I(Y; \mathcal{V})$
- consider the identity:

$$\begin{aligned} I(Y;\mathcal{V}) &= H(\mathcal{V}) - H(\mathcal{V}|Y) \\ &= H(\mathcal{V}) - \mathbb{E}_Y\left[H(\mathcal{V}|y)\right] \end{aligned}$$

- this justifies querying instances which will reduce $|\mathcal{V}| \approx H(\mathcal{V})$ in expectation

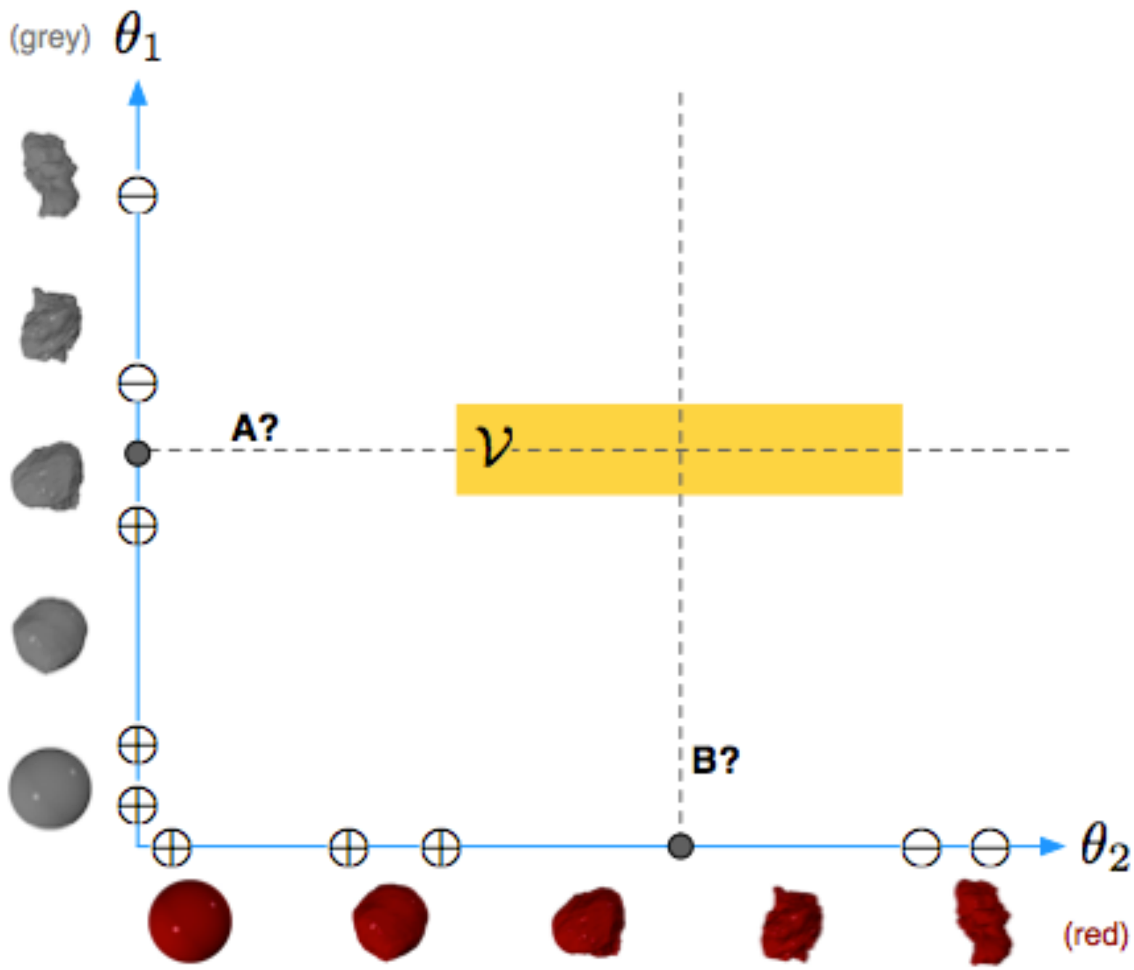
Information-Theoretic Interpretation

- an alternate, equivalent identity:

$$\begin{aligned} I(Y;\mathcal{V}) &= KL(\, P(Y,\mathcal{V}) \parallel P(Y)P(\mathcal{V}) \,) \\ &= \mathbb{E}_{\theta\in\mathcal{V}}\Big[\, KL(\, P_\theta(Y) \parallel P(Y) \,) \,\Big] \end{aligned}$$

- which, under a few simple assumptions, reduces to the KL-divergence heuristic for QBC

Limitations of Version Space Search



imagine Zelgon has both grey and red fruits, with different thresholds?

there are two queries **A** and **B** both bisect $\mathcal{V}$

which query will result is the lowest *classification error*?

Active Learning Approaches
(3) Using the Data Distribution

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

Expected Error Reduction

- minimize the expected 1/0 loss of a query x

$$x^*_{ER} = \underset{x}{\operatorname{argmin}}\ \mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}[y \neq \hat{y}]\right]$$

$$= \underset{x}{\operatorname{argmin}} \sum_{y} P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} 1 - p_{\theta^+}(\hat{y}|x')\right]$$

expectation over possible labelings of x

sum over unlabeled instances

0/1 error of x'
after retraining with x

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

Expected Error Reduction

- minimize the expected *log loss* of a query x

$$
\begin{aligned}
x^*_{LL} &= \underset{x}{\text{argmin}}\ \mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}\left[-\log p_{\theta^+}(y|x')\right]\right] \\
&= \underset{x}{\text{argmin}} \sum_{y} P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} -\sum_{y'} p_{\theta^+}(y'|x')\log p_{\theta^+}(y'|x')\right] \\
&= \underset{x}{\text{argmin}} \sum_{y} P_\theta(y|x) \sum_{x' \in \mathcal{U}} H_{\theta^+}(Y|x'),
\end{aligned}
$$

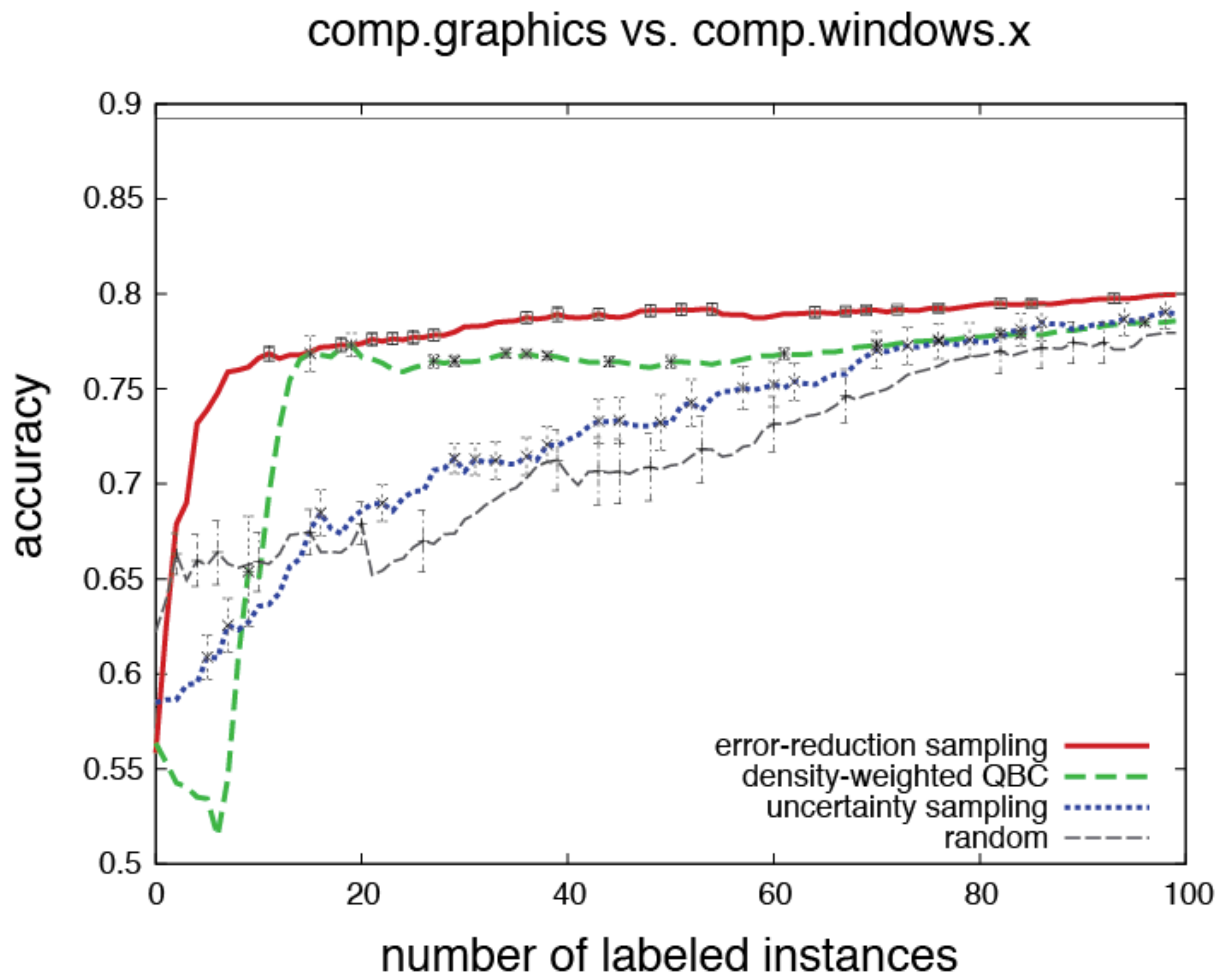
expectation over labelings of x

sum over unlabeled instances

entropy of x'
after retraining with x

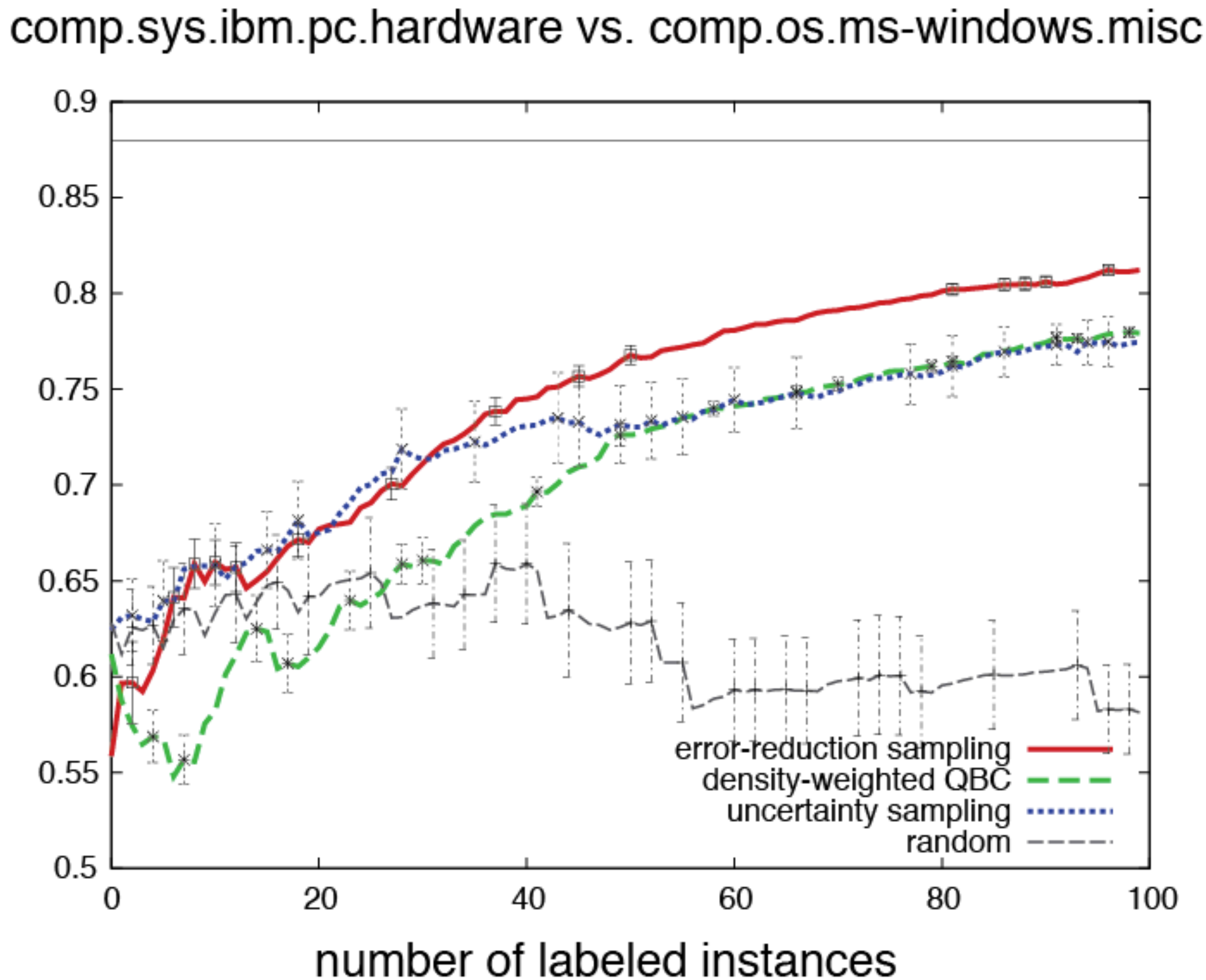
[Roy & McCallum, ICML'01]

Text Classification Examples



[Roy & McCallum, ICML'01]

Text Classification Examples



Information-Theoretic Interpretation

- aim to maximize the *information gain* over $\mathcal{U}$

uncertainty *before* query

expected loss

$$x^* = \operatorname*{argmax}_x \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big] \Big)$$

distribute the sum

$$= \operatorname*{argmax}_x \sum_{x' \in \mathcal{U}} H_\theta(Y|x') - \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big]$$

drop this constant term

$$= \operatorname*{argmin}_x \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big].$$

Poor Scalability

- expected error reduction tries to directly optimize the loss of interest, but…
- quickly becomes intractible
  - logistic regression requires $O(ULG)$ time
  - MaxEnt would require $O(M^2ULG)$ time

Approximation: Density-Weighting

- assume that the information gained per unlabeled instance x' is proportional to its similarity to the query x:

$$x^* = \underset{x}{\operatorname{argmax}} \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x}\big[H_{\theta+}(Y|x')\big]\Big)$$

$$\approx \underset{x}{\operatorname{argmax}} \sum_{x' \in \mathcal{U}} \Big(\underbrace{\mathrm{sim}(x,x')}_{\substack{\text{density term} \\ \text{(i.e., similarity)}}} \times \underbrace{H_\theta(Y|x)}_{\substack{\text{"base" utility} \\ \text{measure}}} \Big).$$

Active Learning++

Beyond Instance Queries

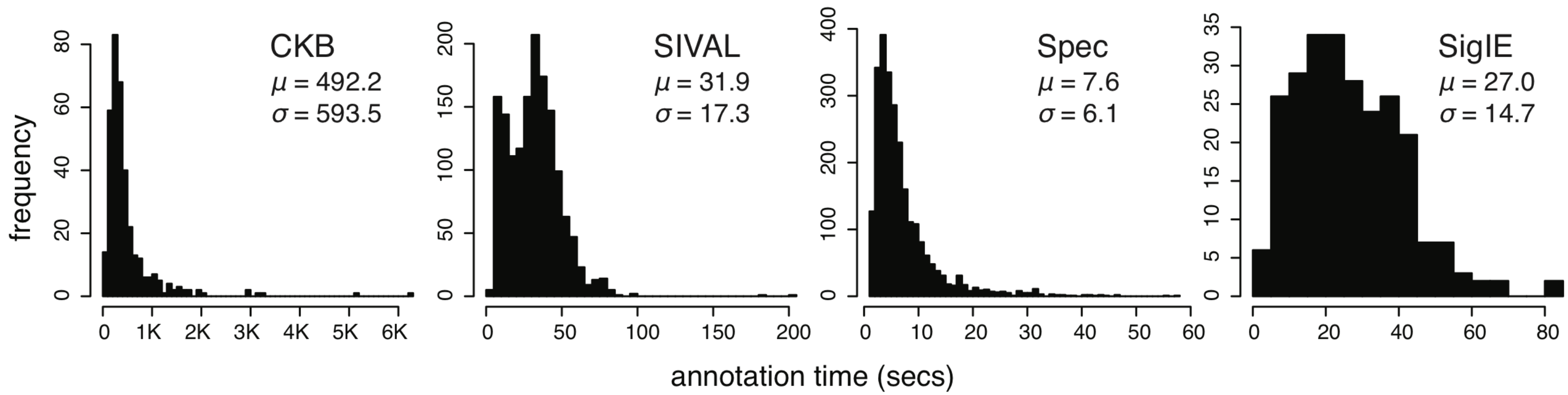
Beyond Instance Queries

- most research in active learning has been based on a few simple assumptions:
  - "cost" is proportional to training set size
  - queries must be unlabeled instances
  - there is only a single classifier to train

[Settles et al., 2008]

1. Real Annotation Costs

empirical study of *time* as labeling cost for four data sets:



information extraction | computer vision | subjectivity analysis | information extraction

[Results supported by Aurora et al., ALNLP'09; Vijayanarasimhan & Grauman, CVPR'09]

Strategies for Variable Annotation Costs

- use the current trained model assist with automatic pre-annotation
  - some successes [Baldridge & Osbourne '04; Culotta & McCallum '05; Baldridge & Palmer '09; Felt et al. '12]
- train a regression cost model in parallel (i.e., to predict time or $$) and incorporate that into the query selection heuristic
  - mixed results [Settles et al. '08; Haertel et al. '08; Tomanek and Hahn '10]

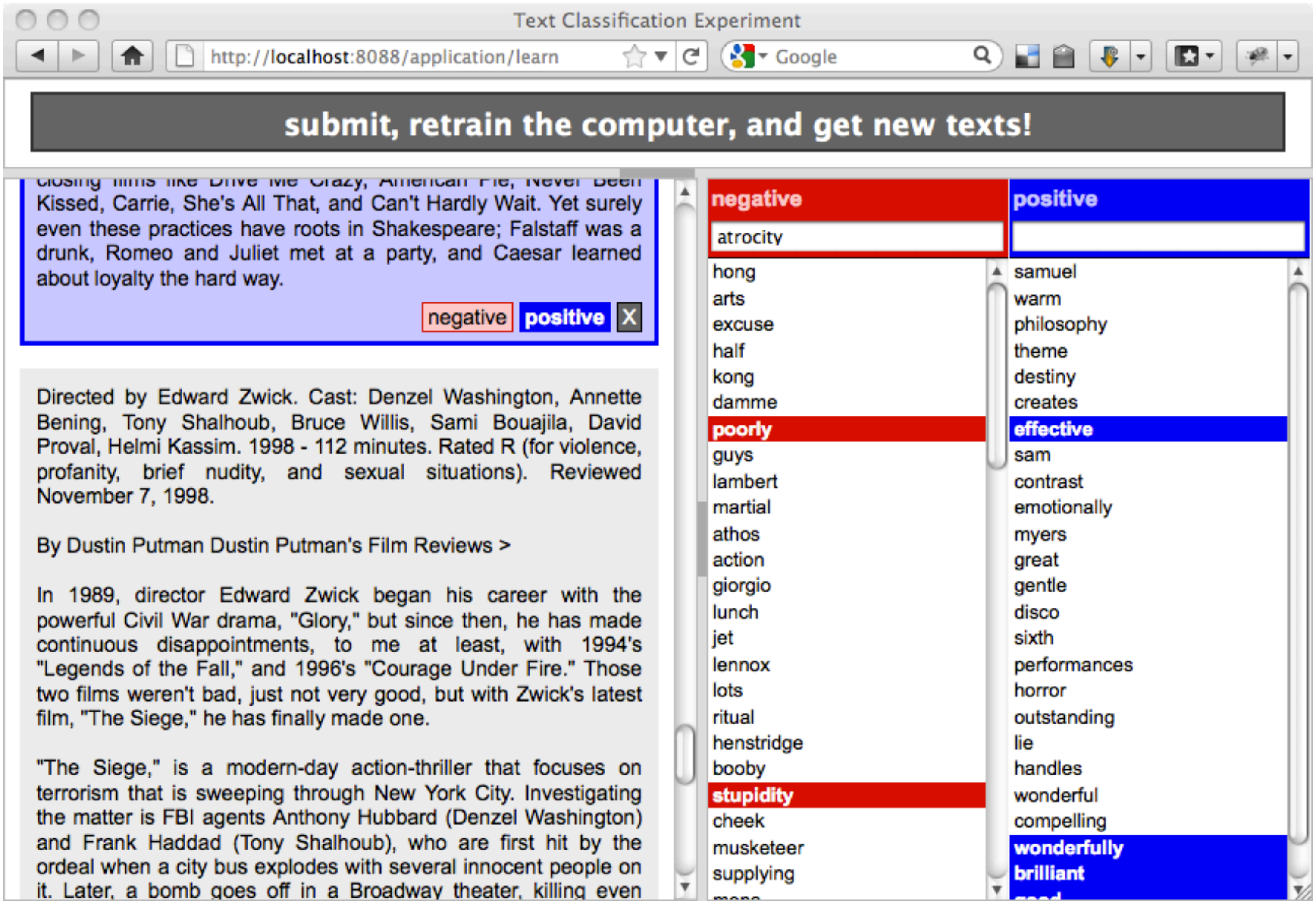
2. New Query Types

- in many NLP applications, “features” are discrete variables with semantic meaning:
  - words
  - affixes
  - capitalization
  - other orthographic patterns
- what if active learning systems could ask about “feature labels,” too?

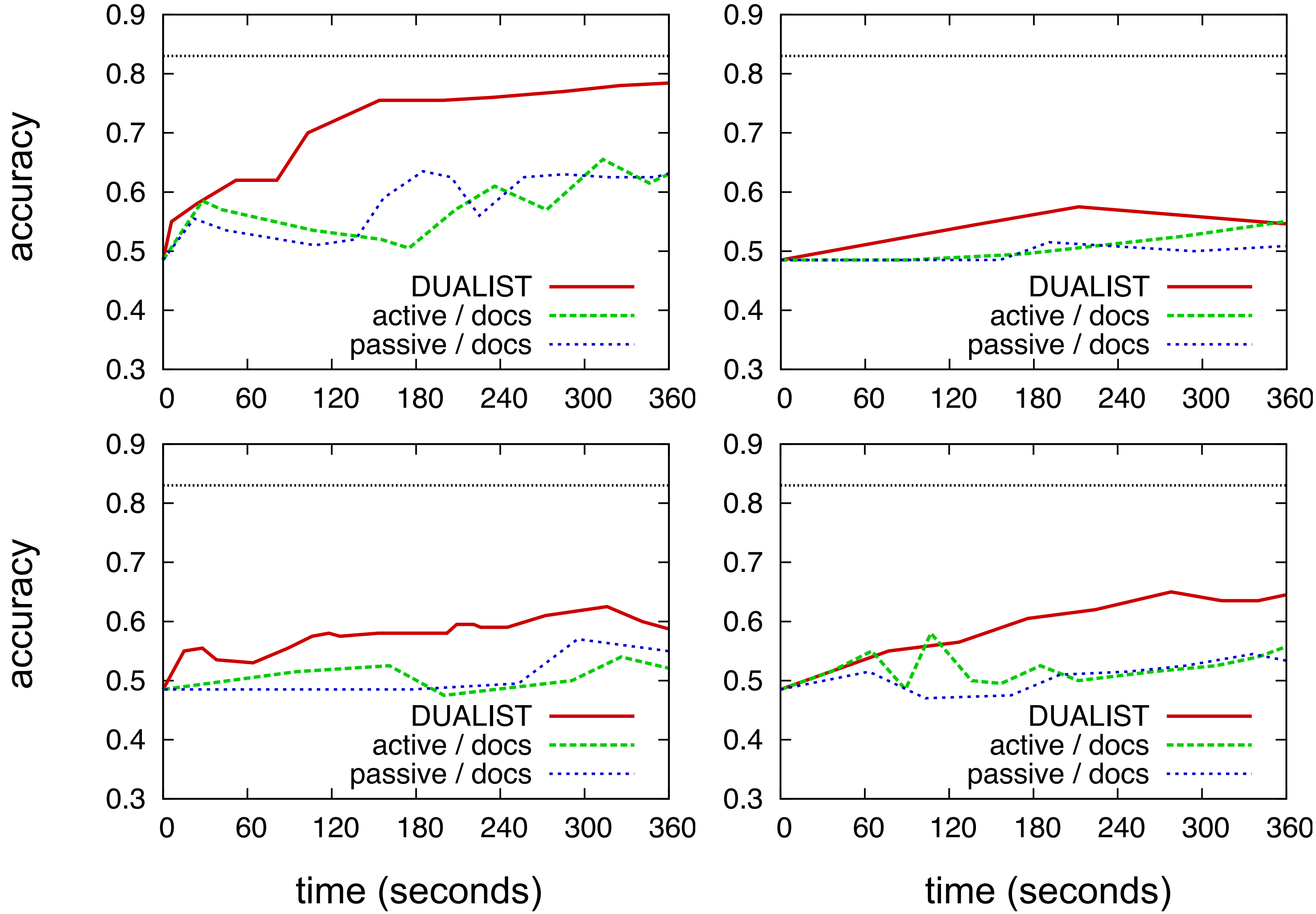
[Druck et al., EMNLP’09; Settles, EMNLP’11]

[Settles, EMNLP'11]

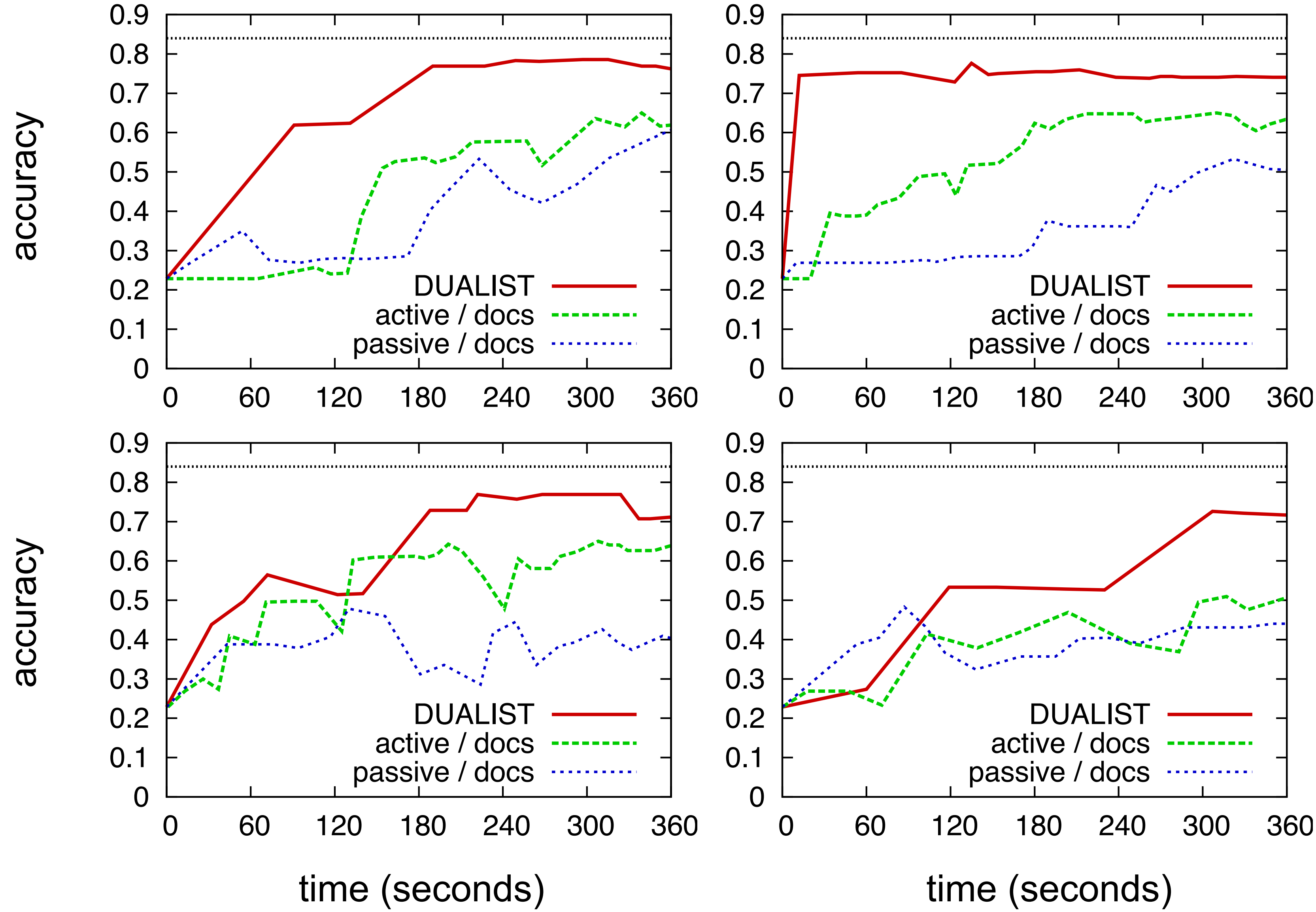
DUALIST



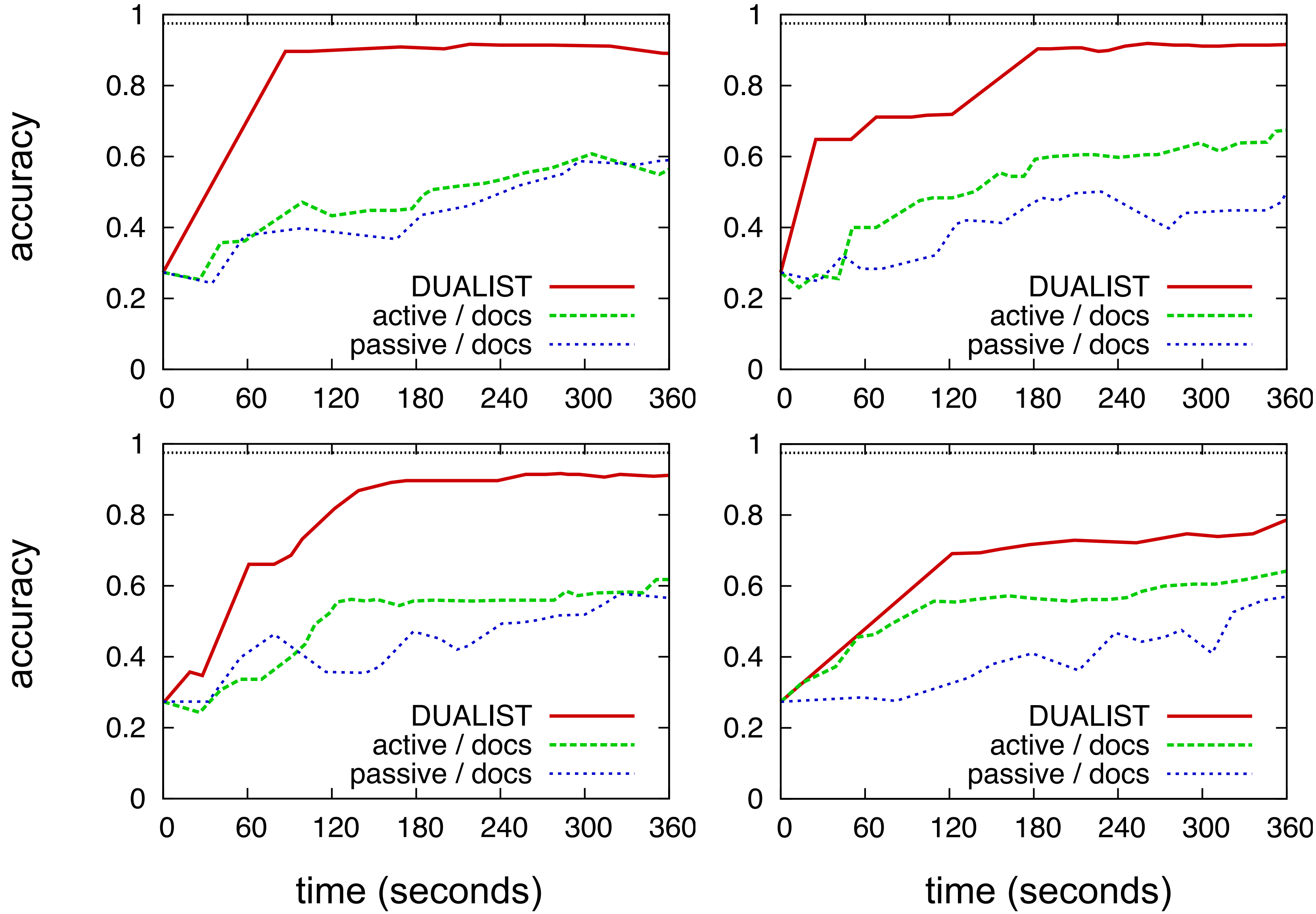
Results: Movie Reviews



Results: WebKB



Results: Science



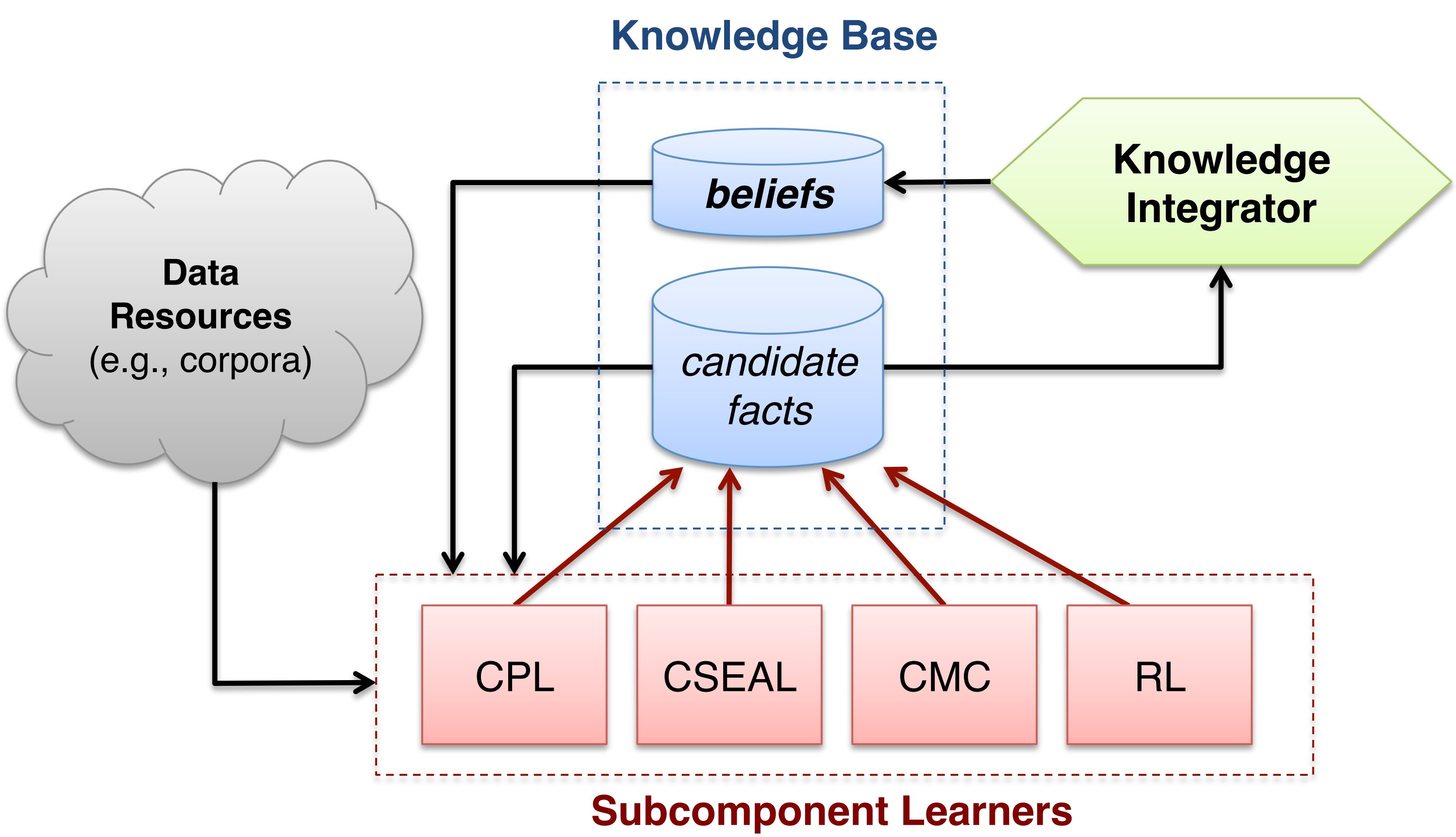
3. Multi-Task, Multi-View Active Learning

- CMU's NELL (Never Ending Language Learner)
- **given:** an ontology (schema), access to the Web, and a few seed examples per predicate, and periodic access to humans
- **task:** run 24x7 each day, populating a knowledge base with new facts
  - learning to read and reading to learn …

[Carlson et al., AAAI'10]

NELL's Architecture

- multiple tasks/views constrain each other, helping to prevent concept drift ("checks and balances")
- to date: >1.5 million beliefs at 80% precision



One View: CPL
(contextual patterns)

Predicate	Pattern
emotion	hearts full of X
beverage	cup of aromatic X
newspaper	op-ed page of X
teamPlaysInLeague	X ranks second in Y
bookAuthor	Y classic X

Another View: CMC
(orthographic features)

Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=state	1.992
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

Gender Issues

@cmunell
NELL

I think "sarah palin" is a #Male (http://bit.ly/dz11Wc)

3 Nov via NELLbot

Retweeted by _stephie_c and 71 others

I proudly voted for _
_ is still the governor
_ is the Republican nominee
_ signed the legislation
_ signed this bill

impeachment proceedings of _
_ 's inaugural
_ signs bill
endorsed _
vice presidential candidates like _

- these CPL patterns are generally correlated with ***male***s across the Web

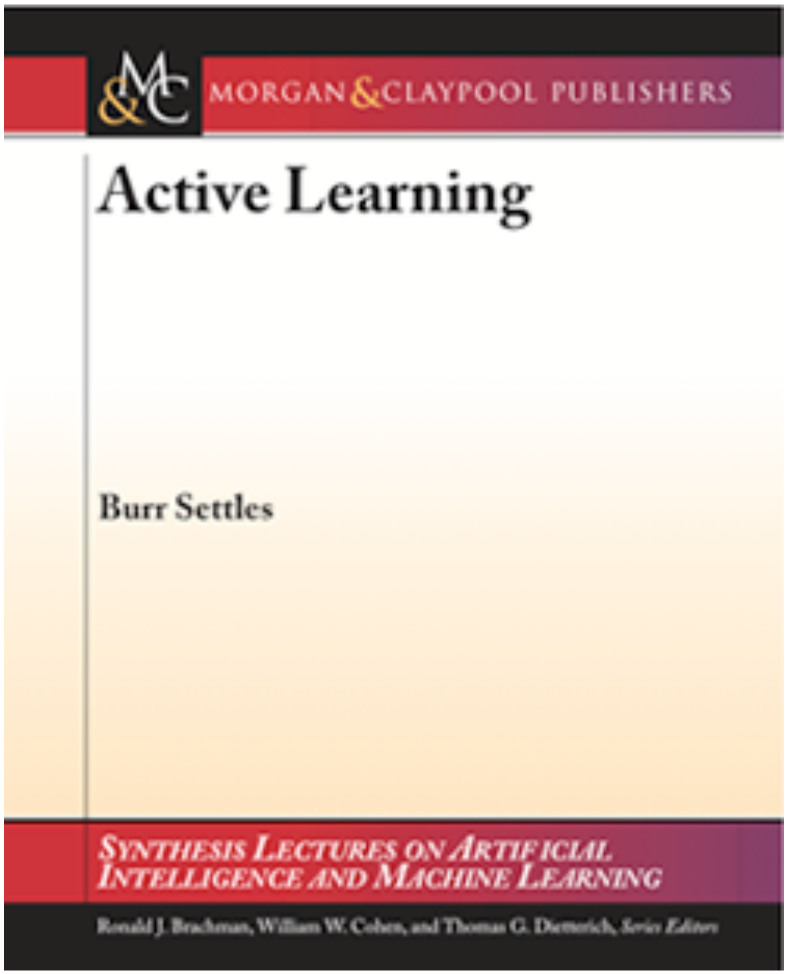
- even though CMC learned that "Sarah" is a ***female*** name, these patterns initially overwhelmed all other evidence, and NELL predicted ***male***

- these days, NELL uses multi-task/view active learning algorithms to identify beliefs with "conflicting" evidence, and query them

Interesting Open Issues

- better cost-sensitive approaches
- “crowdsourced” labels (noisy oracles)
- batch active learning (many queries at once)
- HCI / user interface issues
- data reusability

For Further Reading…



new book published by Morgan & Claypool

free to download from the CMU campus network

active-learning.net