Active Learning

Aarti Singh

Machine Learning 10-601
Dec 6, 2011

Slides Courtesy: Burr Settles, Rui Castro, Rob Nowak

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Learning from unlabeled data

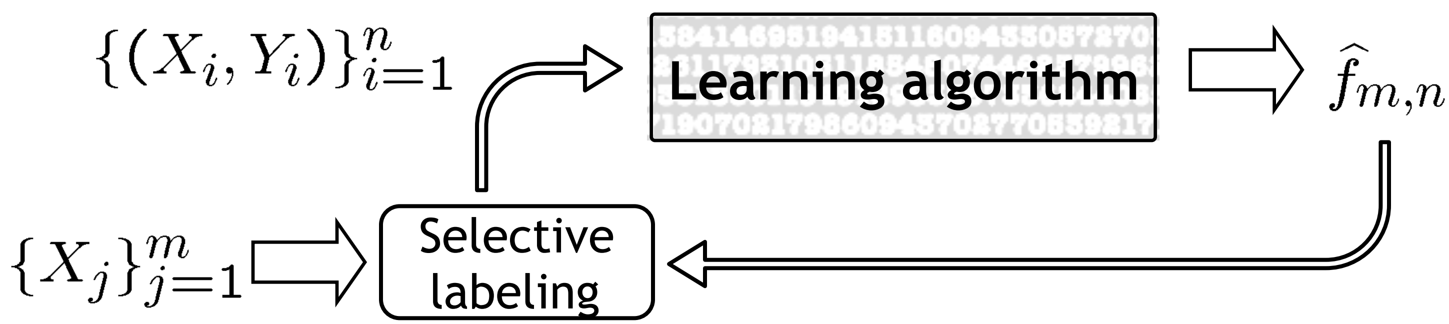
Semi-supervised learning: Design a predictor based on iid unlabeled and few ***randomly*** labeled examples.

$$\{(X_i, Y_i)\}_{i=1}^{n} \quad \{X_j\}_{j=1}^{m} \Longrightarrow \boxed{\text{Learning algorithm}} \Longrightarrow \widehat{f}_{m,n}$$

Assumption: Knowledge of marginal density can simplify prediction
e.g. similar data points have similar labels

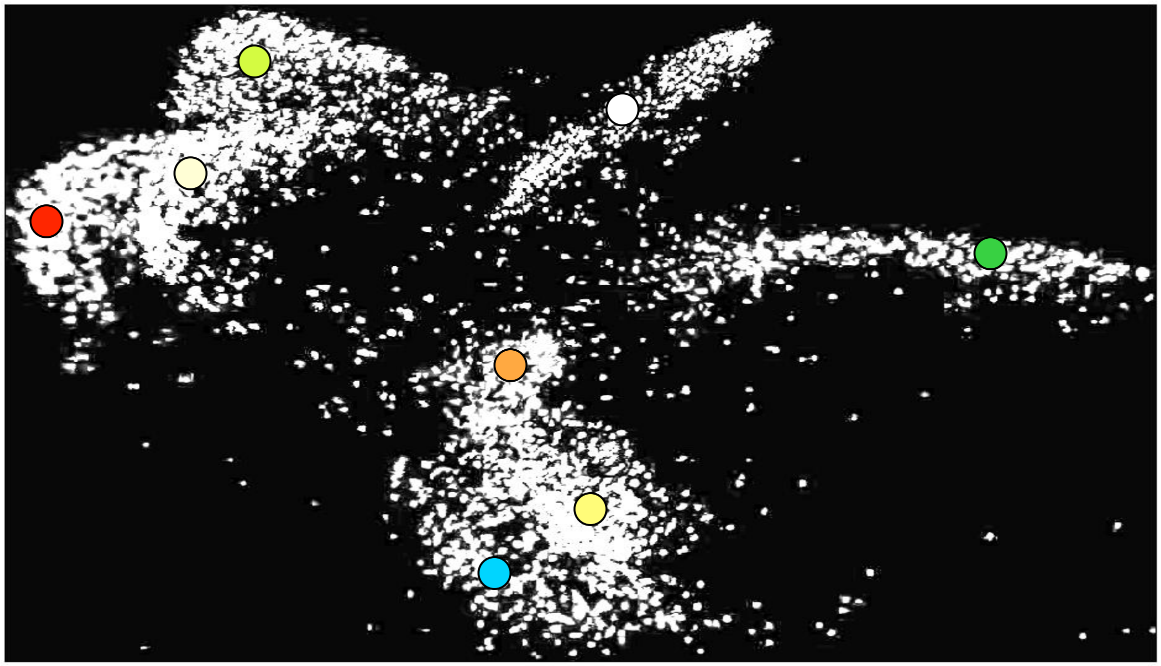
Learning from unlabeled data

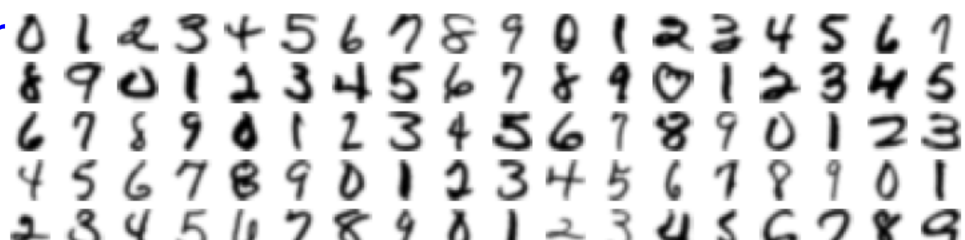
Active learning: Design a predictor based on iid unlabeled and ***selectively*** labeled examples



Assumption: Some unlabeled examples are more informative than others for prediction.

Example: Hand-written digit recognition

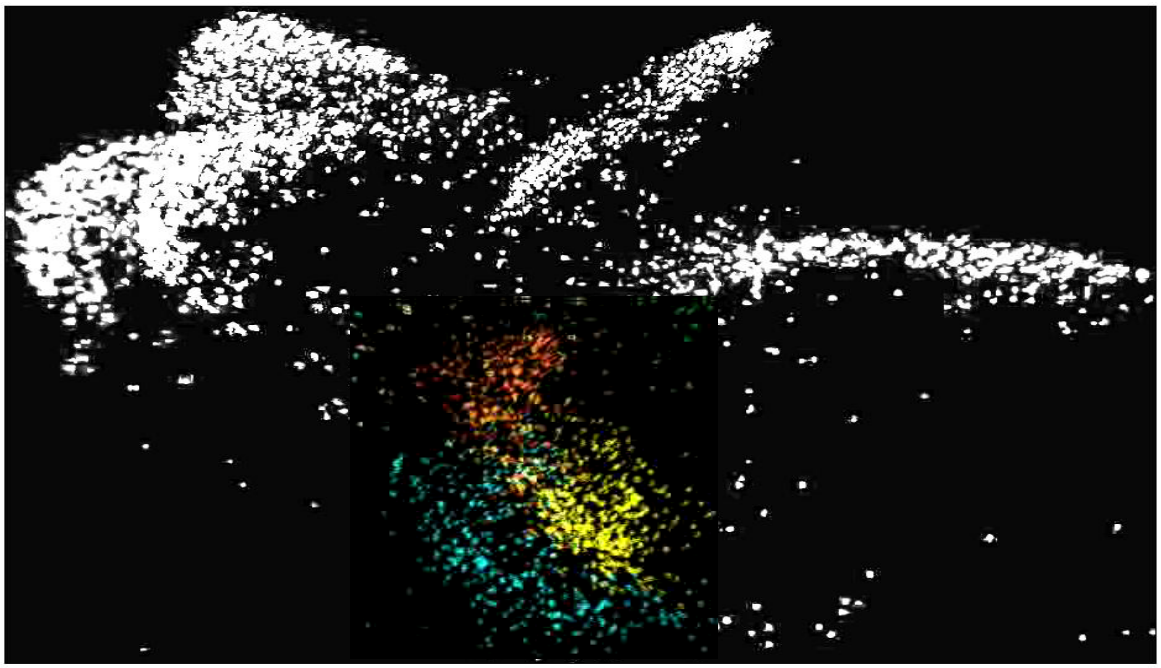




knowledge of clusters + a few labels in each is sufficient to design a good predictor – **Semi-supervised learning**

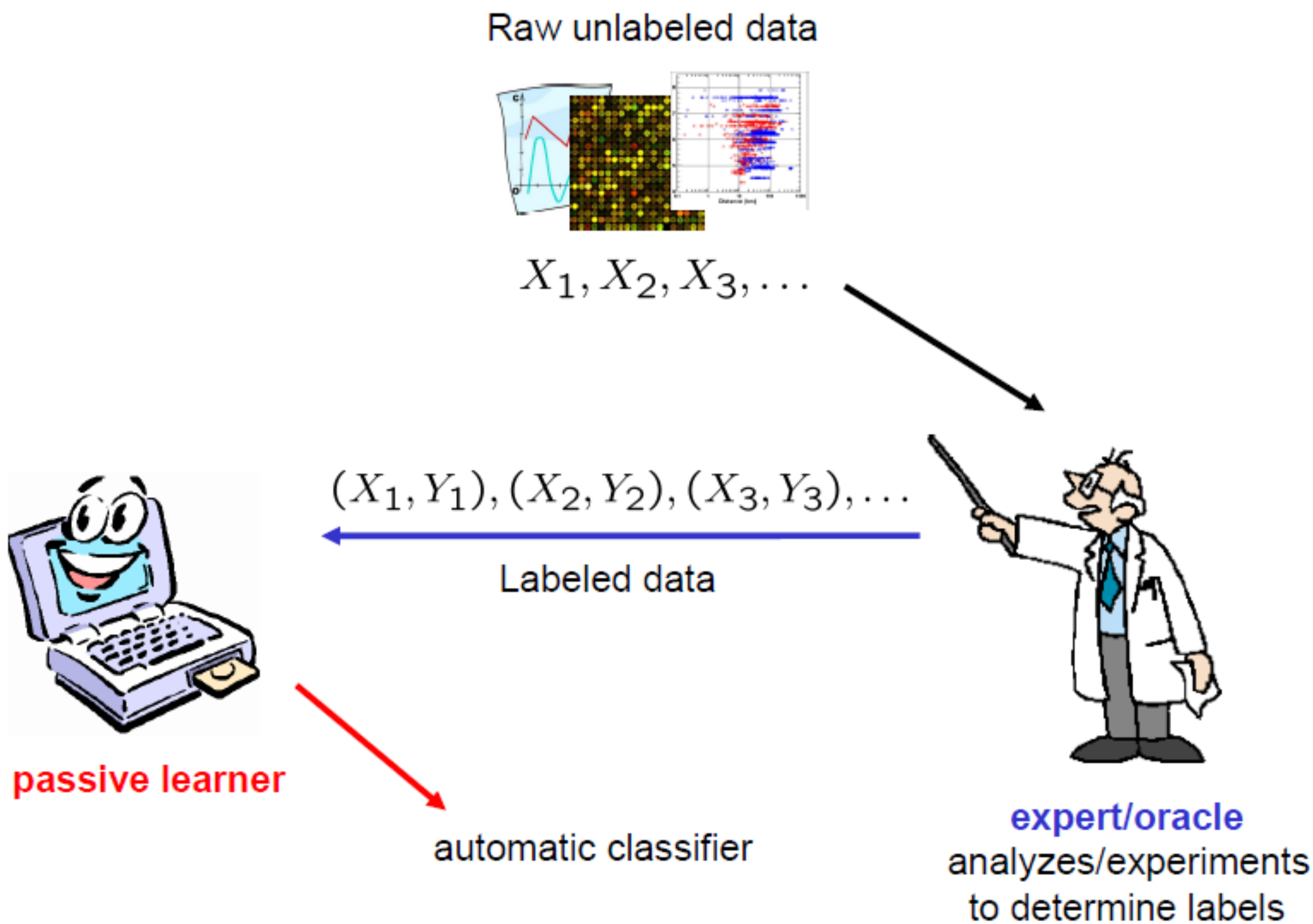
Example: Hand-written digit recognition

Not all examples are created equal

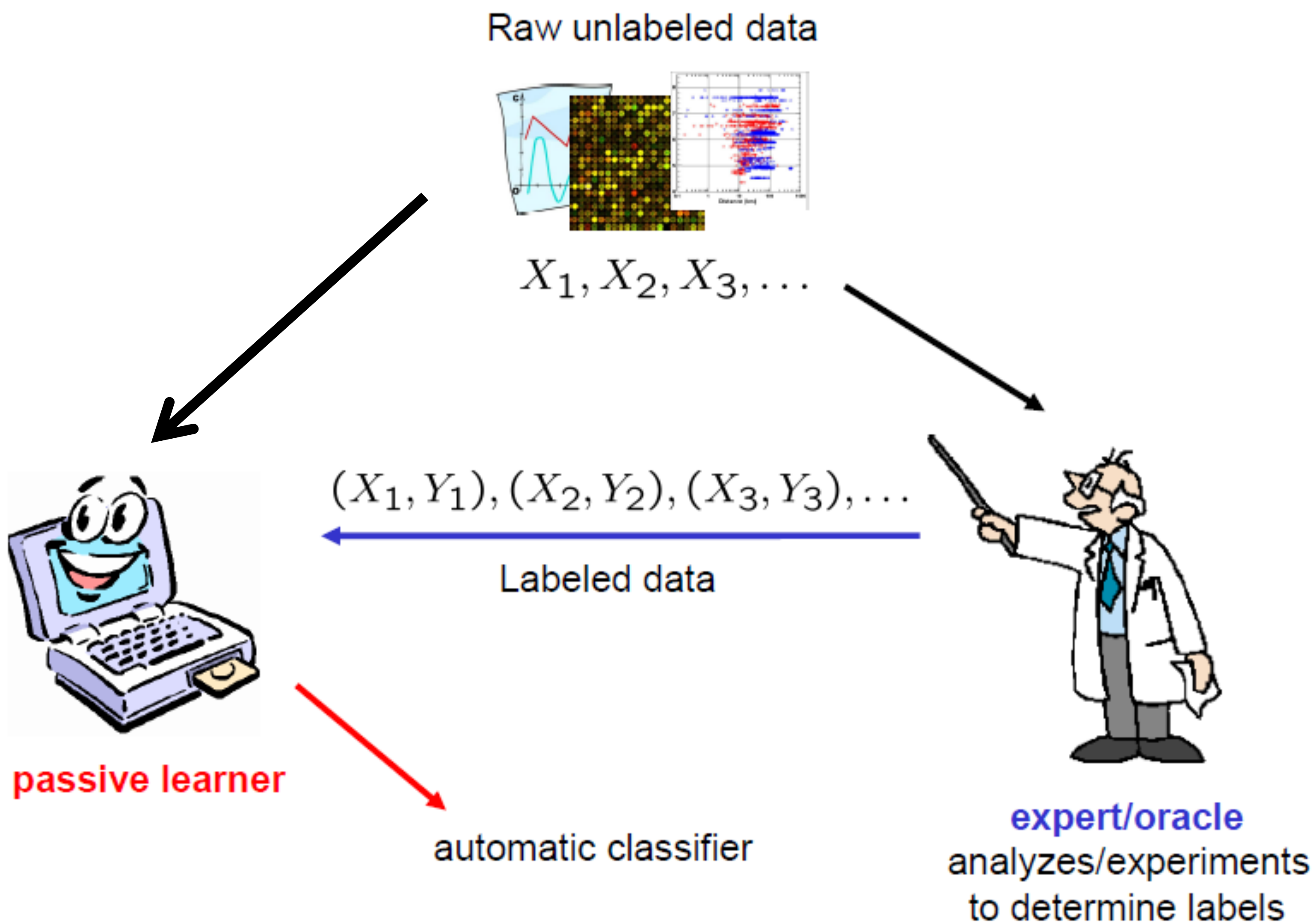


Labeled examples near “boundaries” of clusters are much more informative – **Active learning**

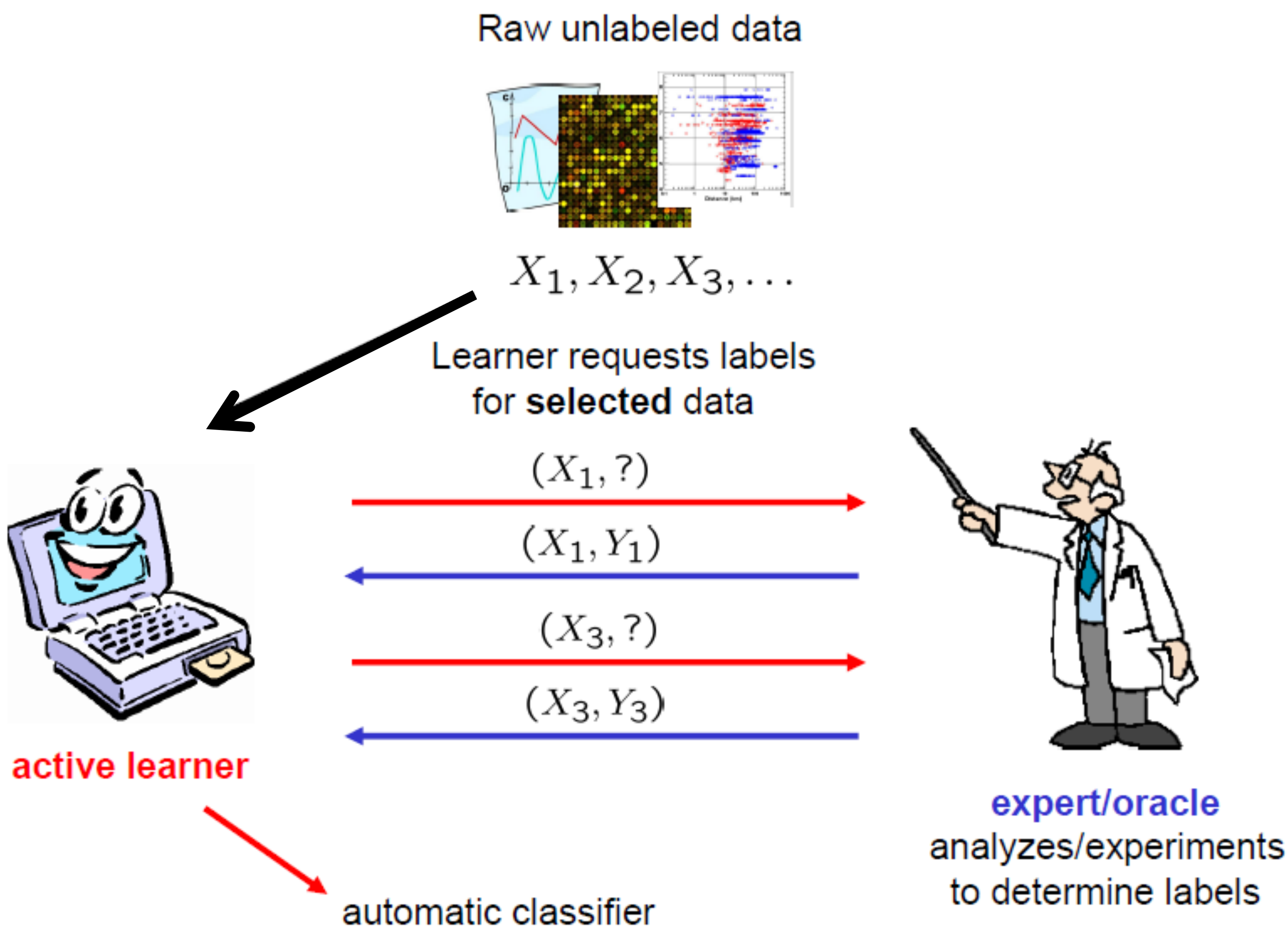
Passive Learning



Semi-supervised Learning

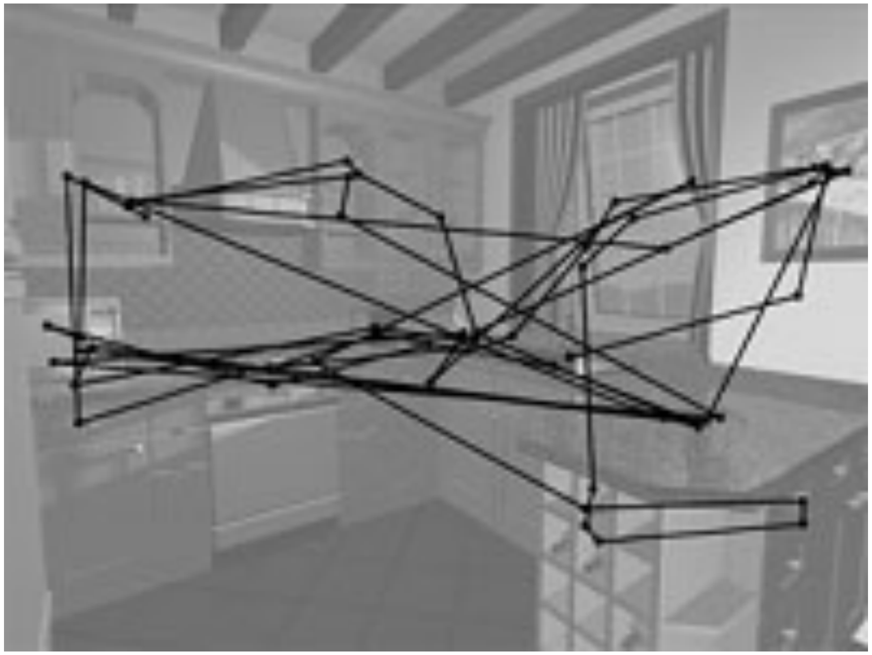


Active Learning

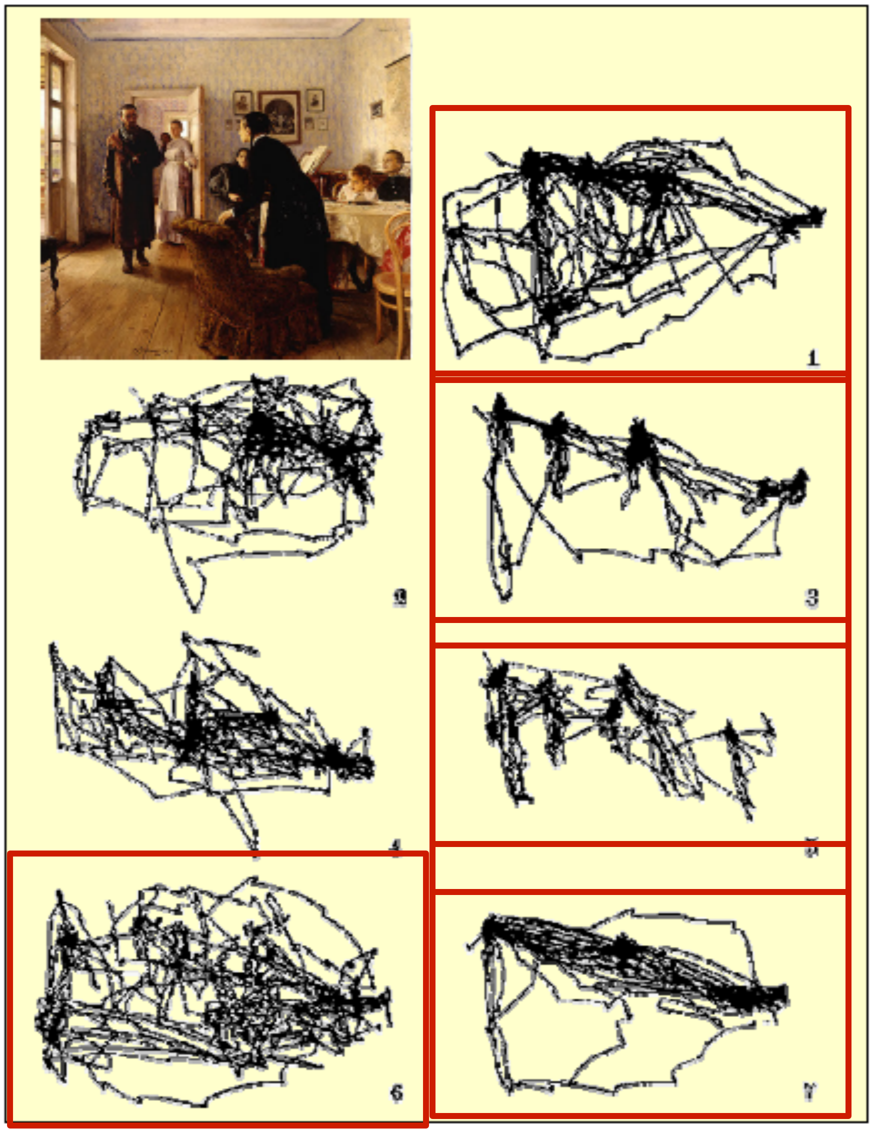


Feedback driven learning

The eyes focus on the interesting and relevant features, and do not sample all the regions in the scene in the same way.



Feedback driven learning



Seven records of eye movements by the same subject. Each record lasted 3 minutes. 1) Free examination. Before subsequent recordings, the subject was asked to: 2) estimate the material circumstances of the family; 3) give the ages of the people; 4) surmise what the family had been doing before the arrival of the "unexpected visitor;" 5) remember the clothes worn by the people; 6) remember the position of the people and objects in the room; 7) estimate how long the "unexpected visitor" had been away from the family (from Yarbus 1967).

The Twenty questions game

"Does the person have blue eyes ?"

"Is the person wearing a hat ?"

Focus on most informative questions

"Active Learning" works very well in simple conditions

Thought Experiment

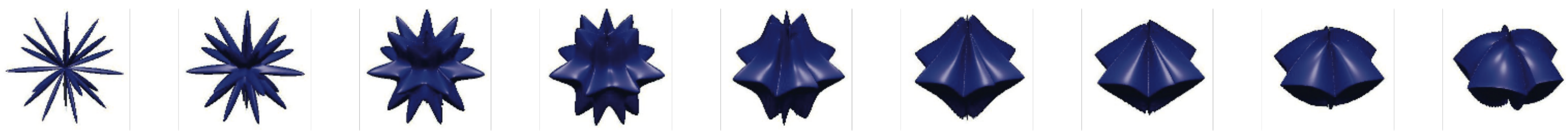
- suppose you're the leader of an Earth convoy sent to colonize planet Mars

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 bisection search**

Your first active learning algorithm!

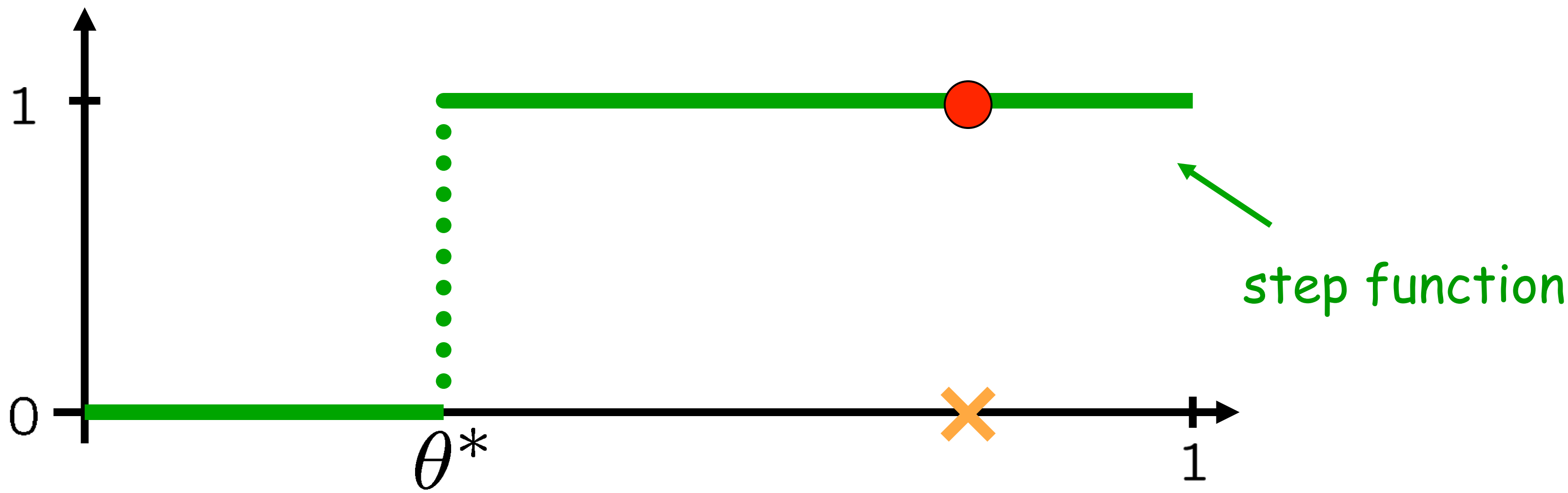
Active Learning

- **key idea:** the learner can choose training data on the fly
  - 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"

Learning a change-point

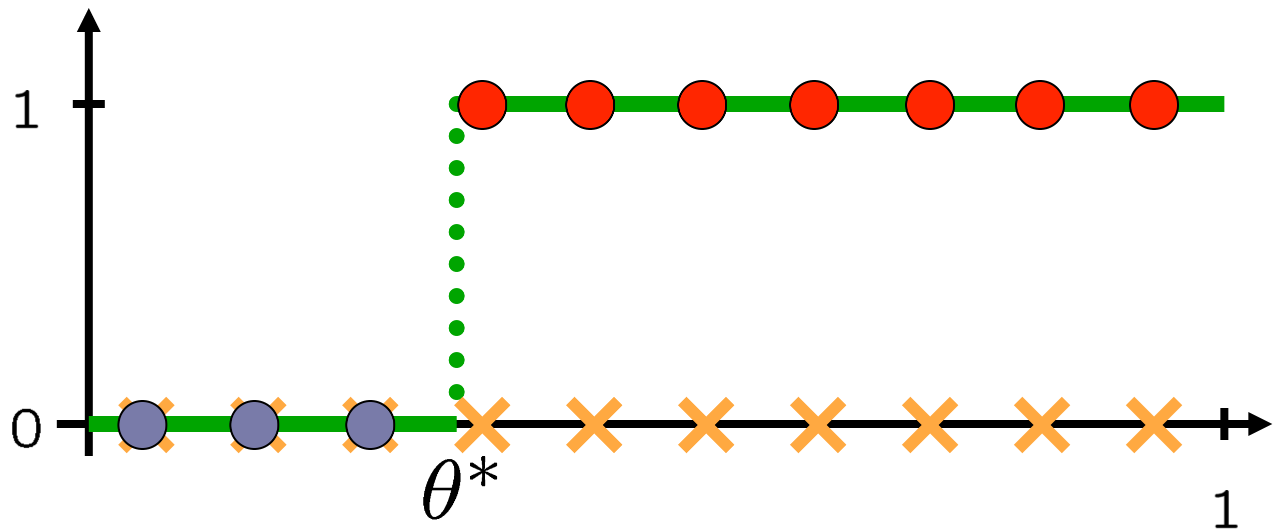
Locate a change-point or threshold (poisonous/yummy fruit,
contamination boundary)



Goal: Given a budget of n samples, learn threshold θ^* as accurately as possible

Passive Learning

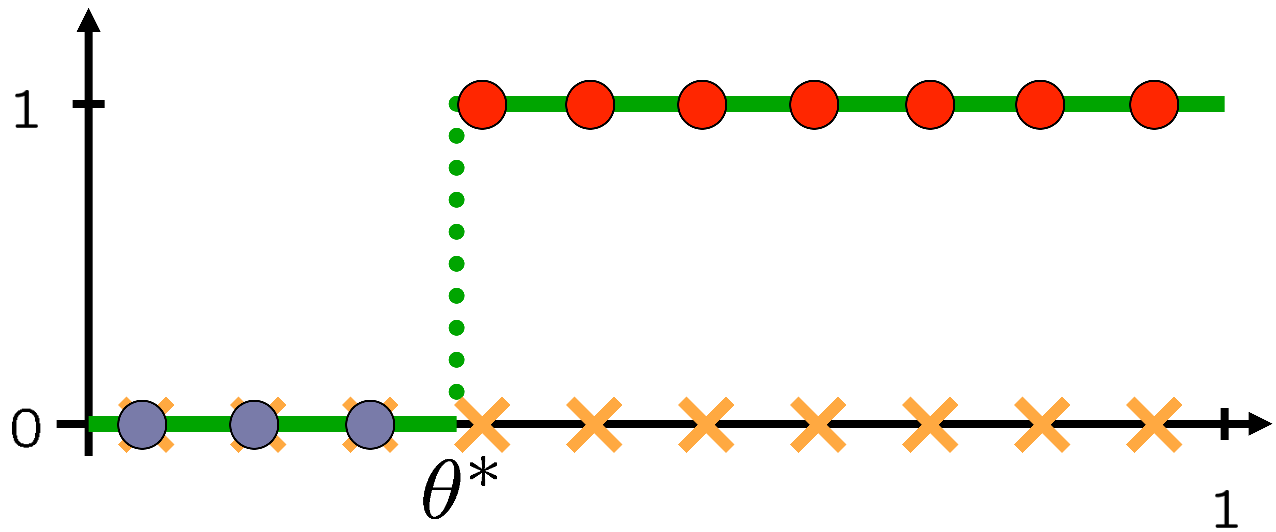
Sample locations must be chosen before any observations are made



$$|\widehat{\theta}_n - \theta^*| \sim \frac{1}{n}$$

Passive Learning

Sample locations must be chosen before any observations are made

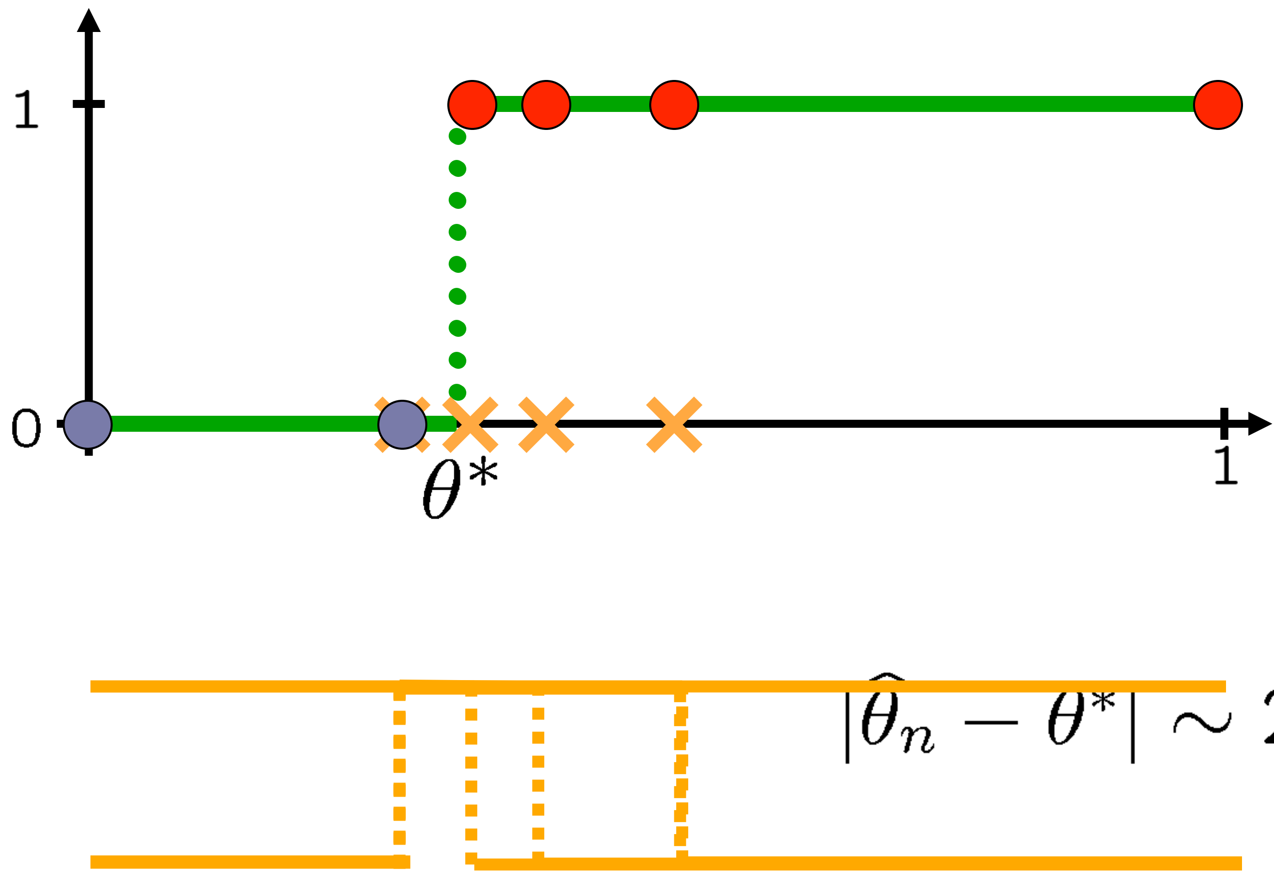


$$|\widehat{\theta}_n - \theta^*| \sim \frac{1}{n}$$

Too many wasted samples. Learning is limited by sampling resolution

Active Learning

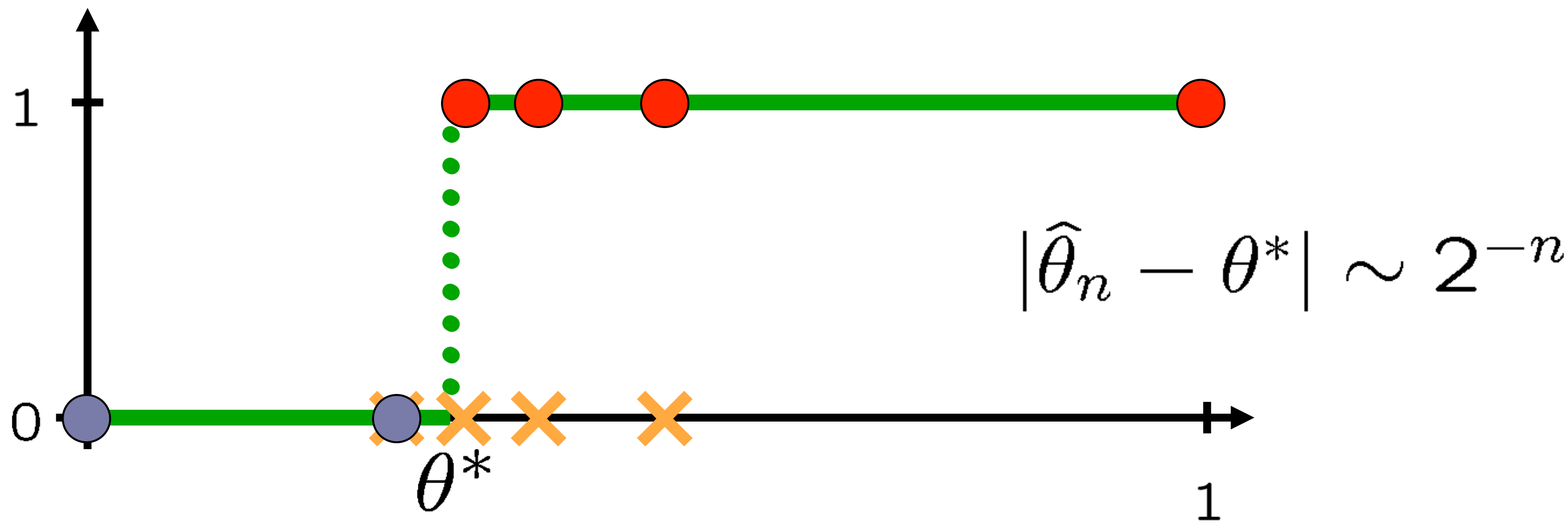
Sample locations are chosen based on previous observations



$$|\widehat{\theta}_n - \theta^*| \sim 2^{-n}$$

Active Learning

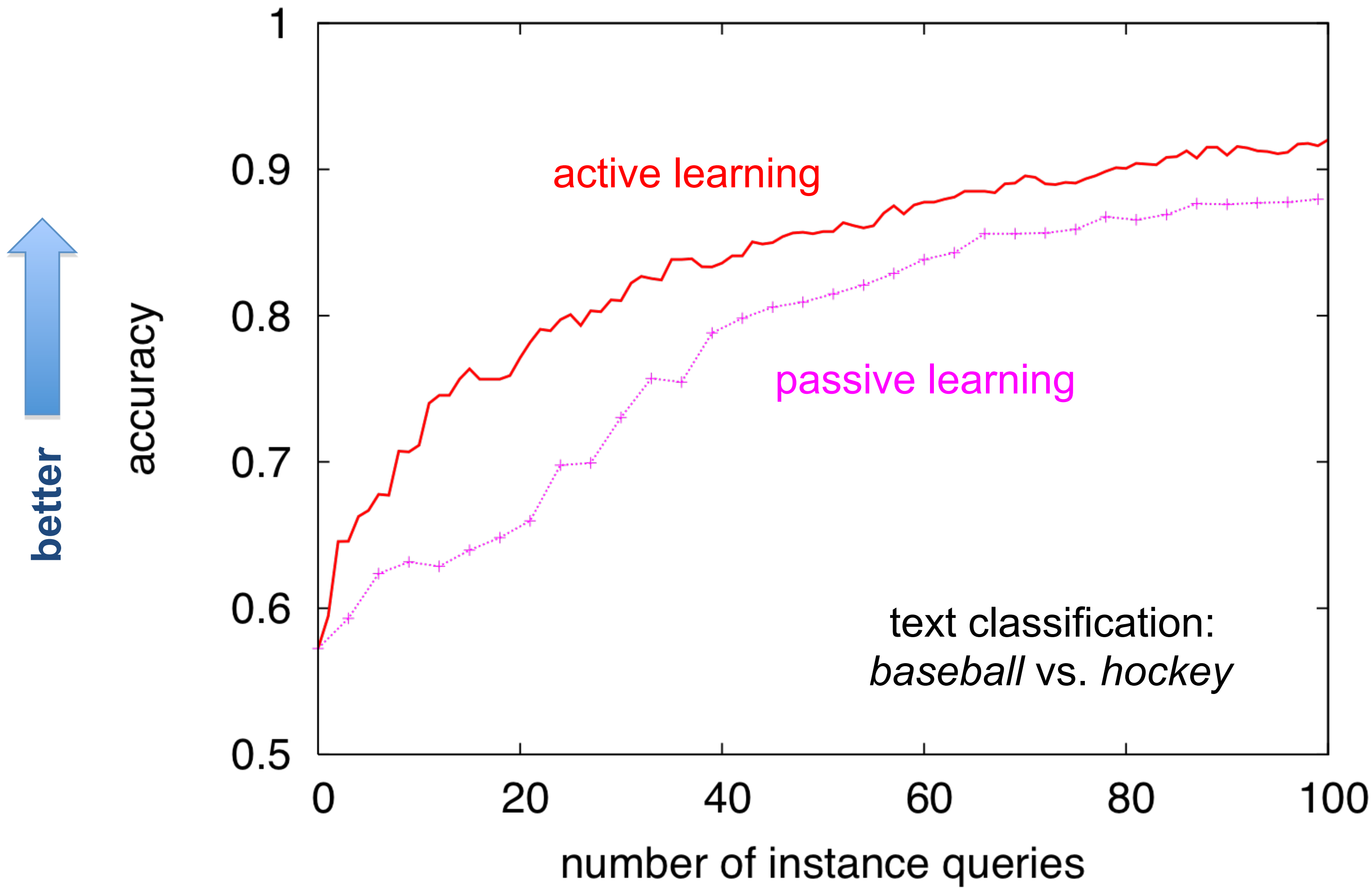
Sample locations are chosen based on previous observations



The error decays much faster than in the passive scenario. No wasted samples… Exponential improvement!

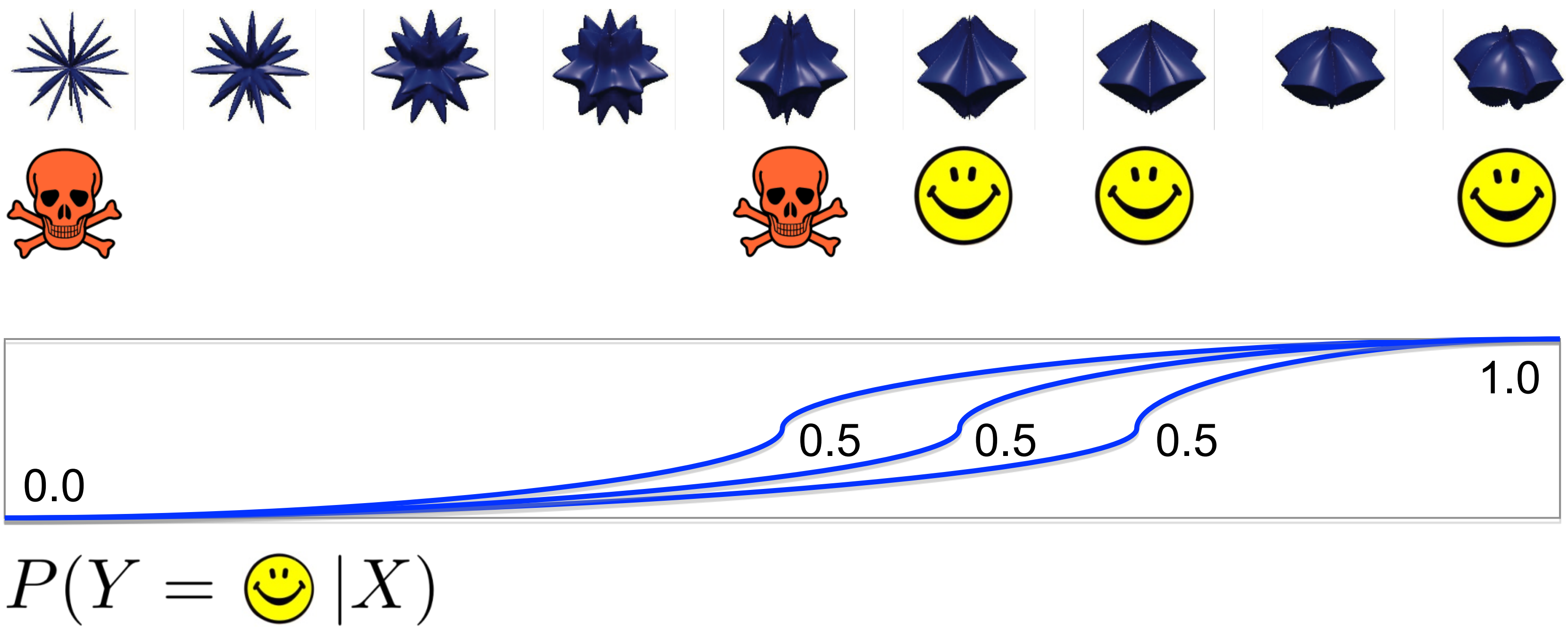
Works even when labels are noisy … though improvement depends on amount of noise

Practical Learning Curves



Probabilistic Binary Bisection

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



[Lewis & Gale, SIGIR'94]

Uncertainty Sampling

- query instances the learner is *most uncertain* about

$$x^*_{LC} = \text{arg}\max_x \; 1 - P_\theta(y^*|x), \qquad \text{where } y^* = \text{arg}\max\nolimits_y \; P_\theta(y|x),$$

Using logistic regression

Generalizing to Multi-Class Problems

least confident [Culotta & McCallum, AAAI'05]

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

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

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

entropy [Dagan & Engelson, ICML'95]

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

note: for binary tasks, these are equivalent

[Seung et al., COLT'92]

Query-By-Committee (QBC)

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

- query instances in U for which the committee is in most *disagreement*

- **key idea:** reduce the model *version space* (set of hypotheses which are consistent with training examples)
  - expedites search for a model during training

Version Space Examples

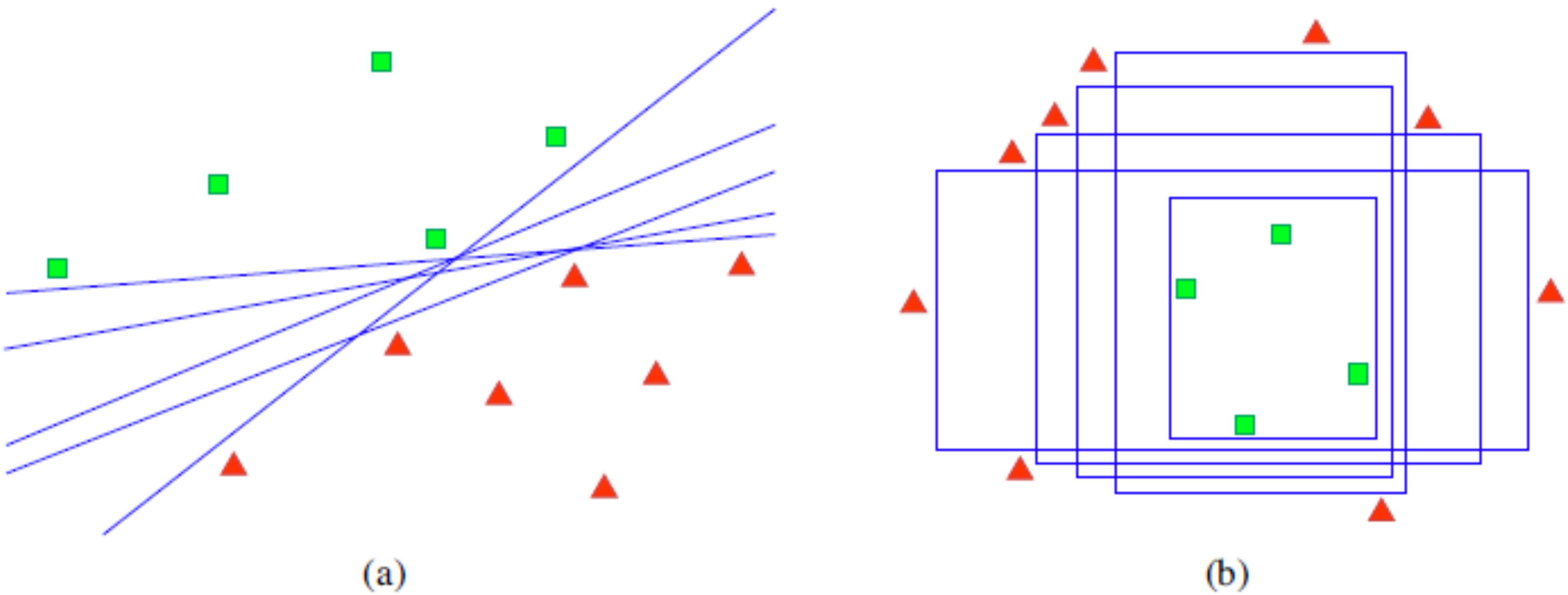
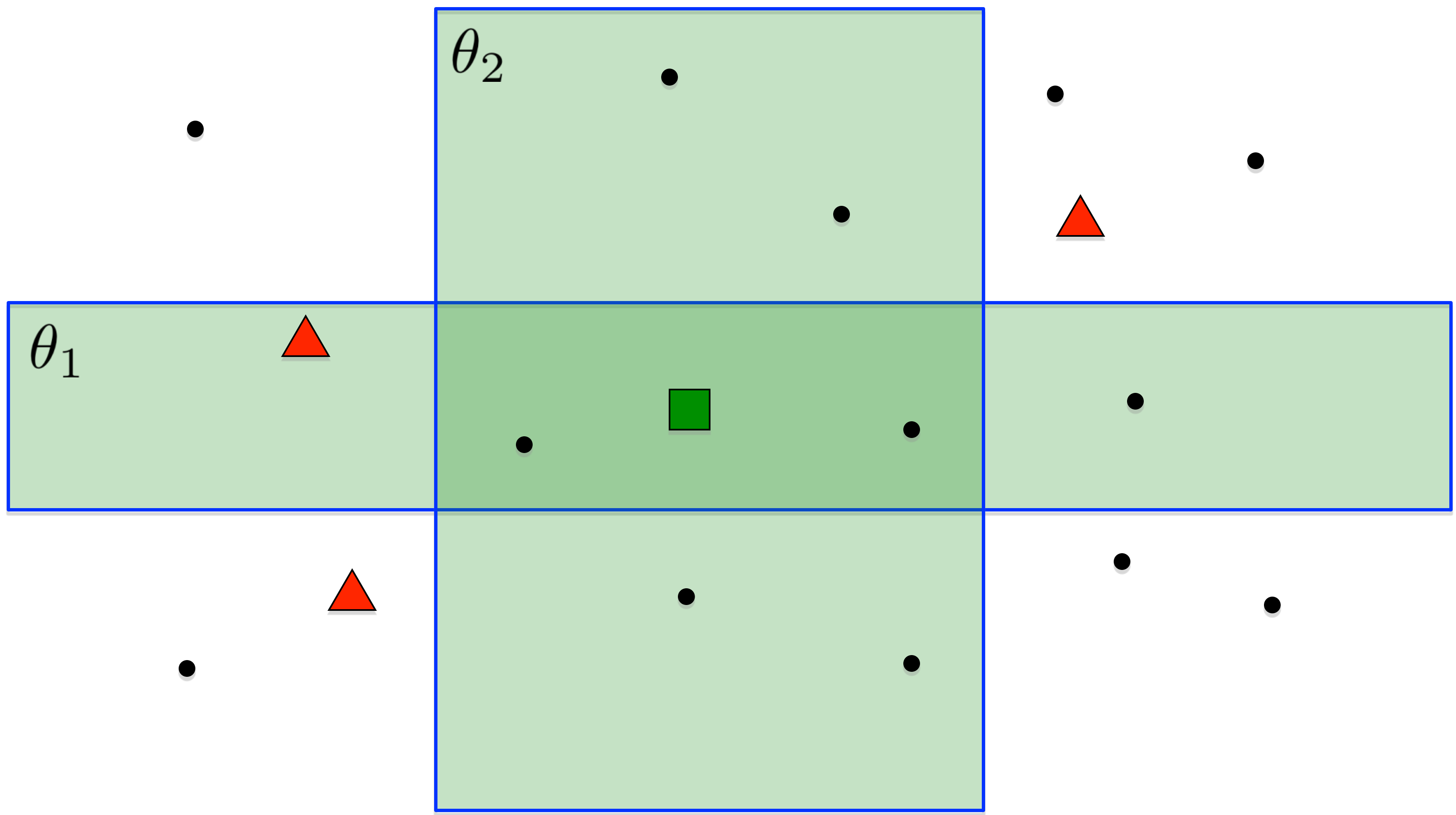
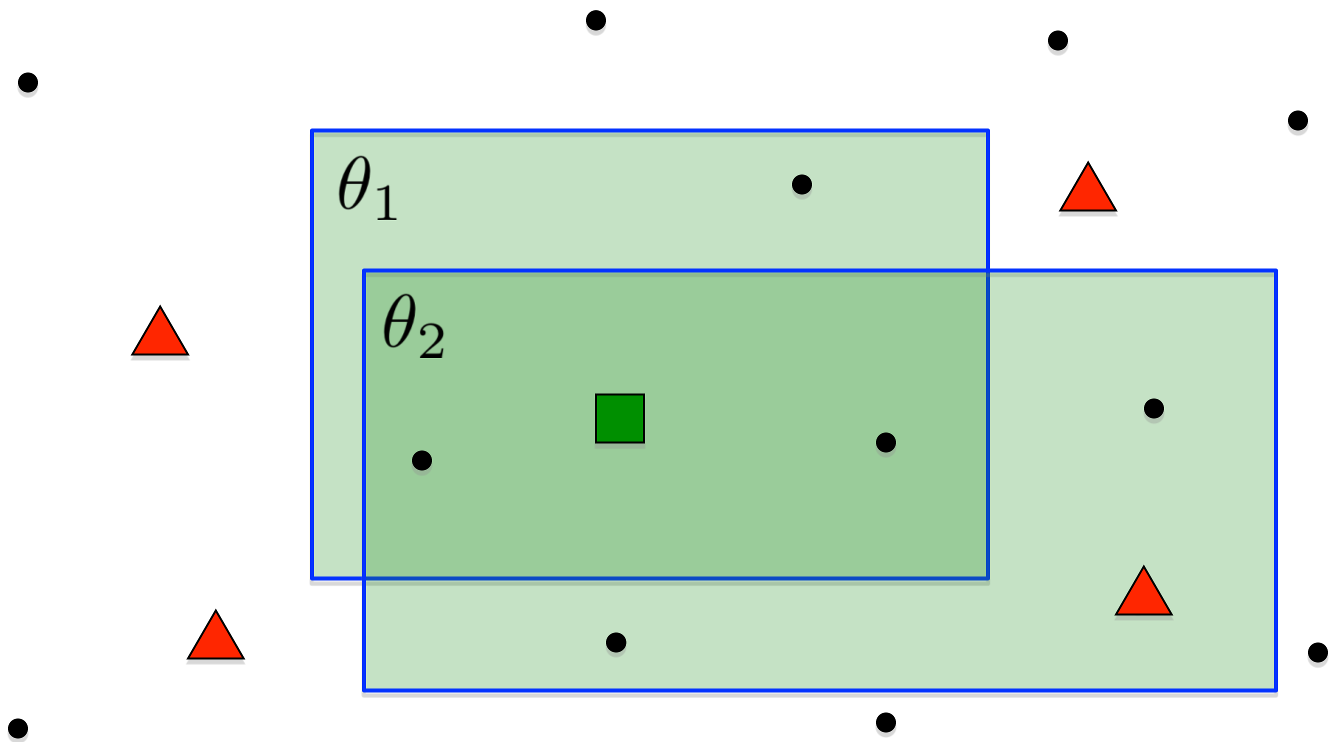


Figure 6: Version space examples for (a) linear and (b) axis-parallel box classifiers. All hypotheses are consistent with the labeled training data in $\mathcal{L}$ (as indicated by shaded polygons), but each represents a different model in the version space.

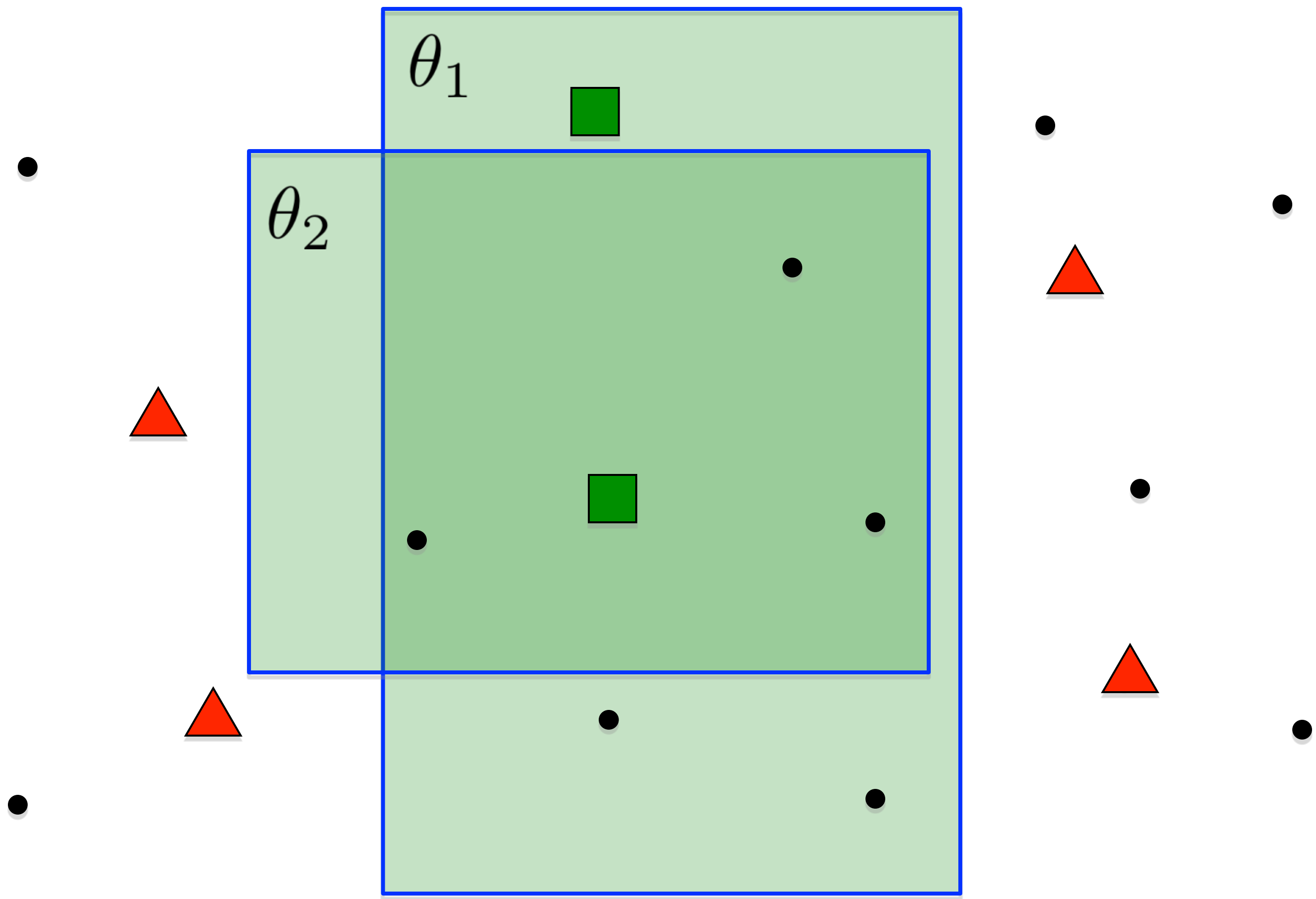
QBC Example



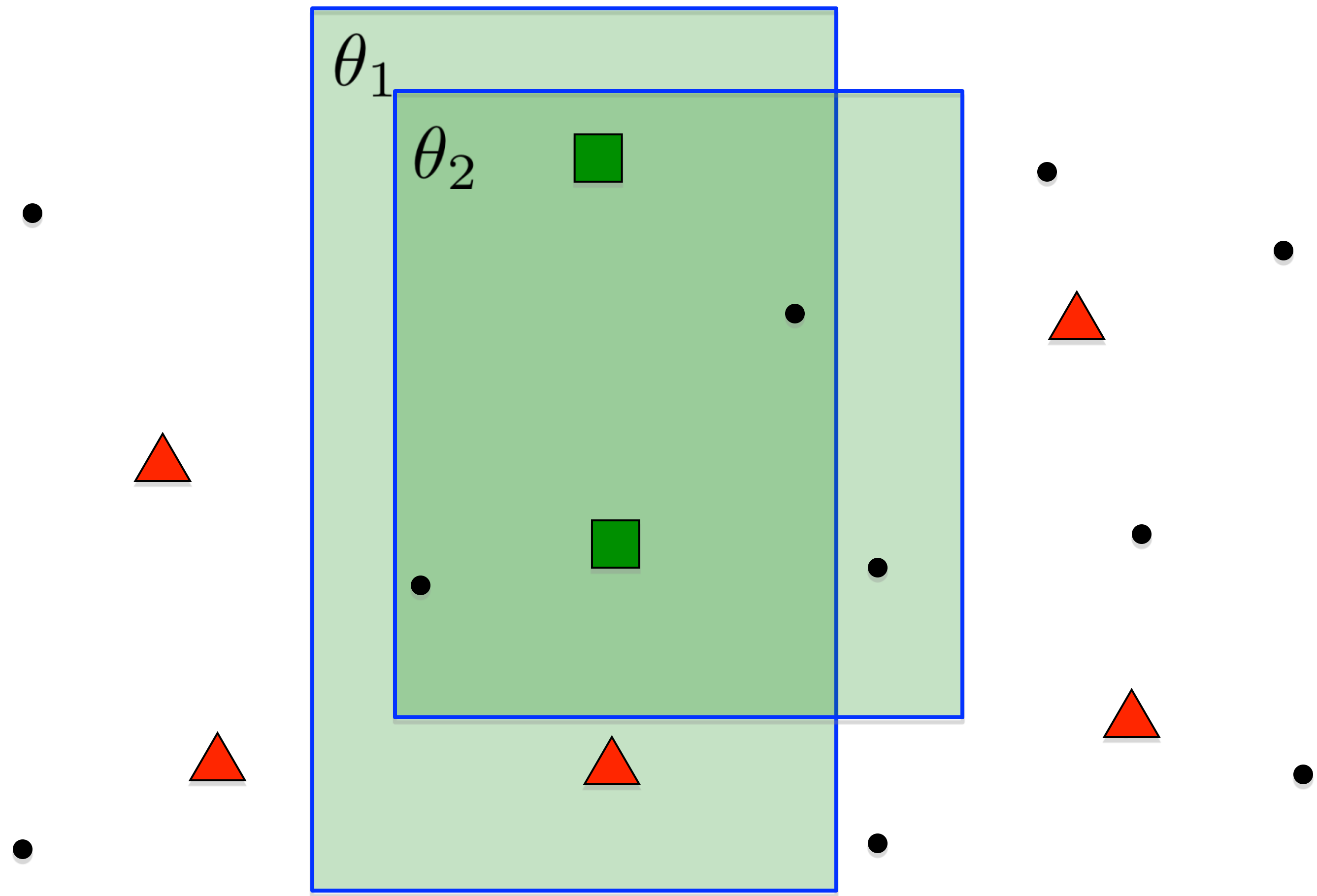
QBC Example



QBC Example



QBC Example



QBC Guarantees

- theoretical guarantees… [Freund et al.,'97]

d – VC dimension of committee classifiers

Under some mild conditions, the QBC algorithm achieves a prediction accuracy of ε and w.h.p.

# unlabeled examples generated	$O(d/\varepsilon)$
# labels queried	$O(\log_2 d/\varepsilon)$

Exponential improvement!

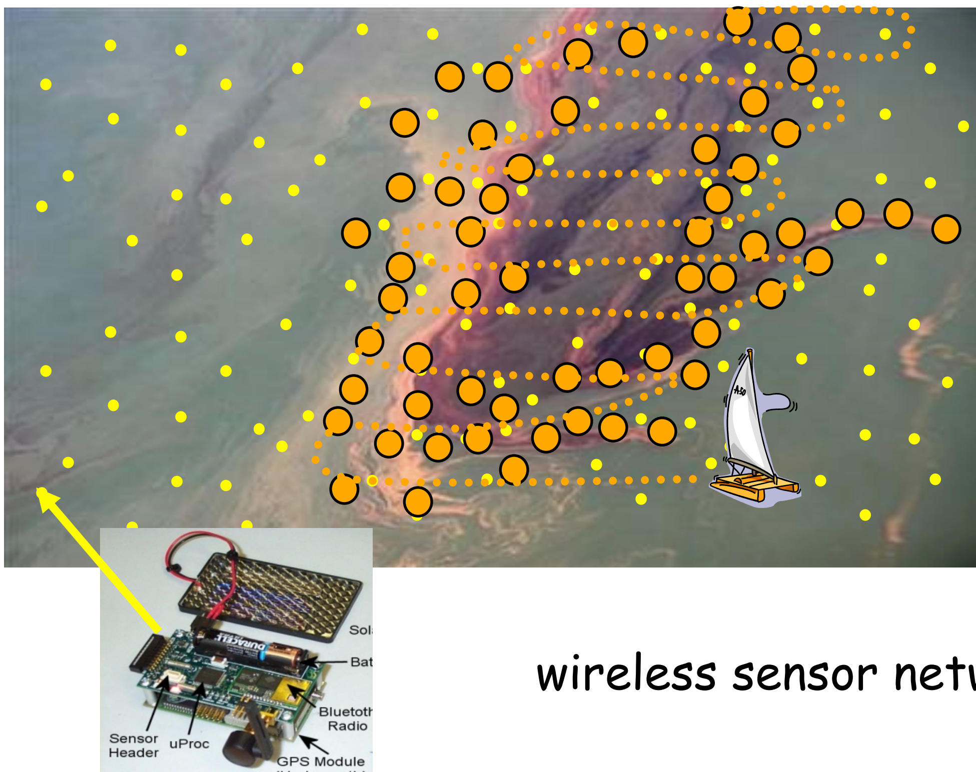
QBC: Design Decisions

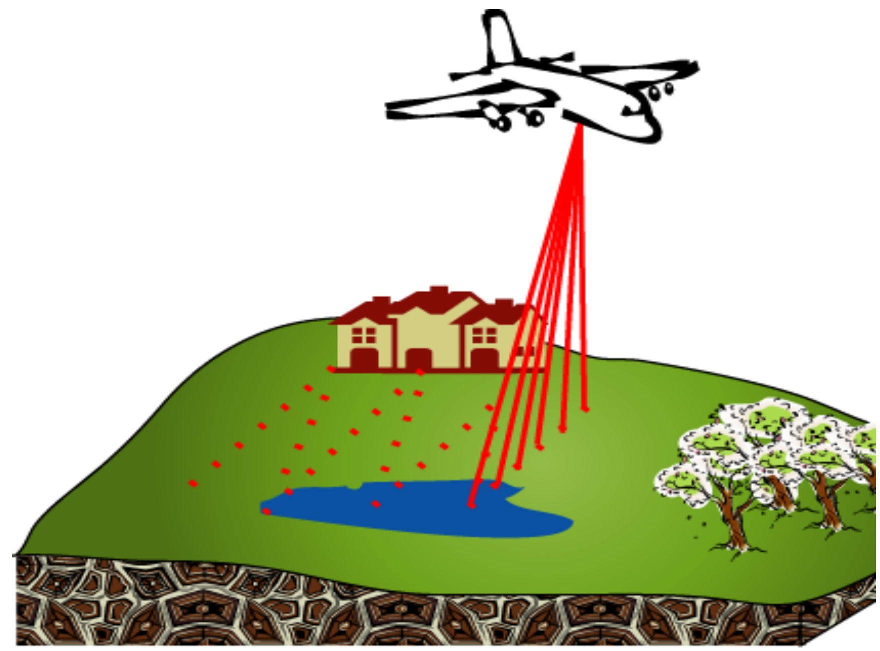
- how to build a committee:
  - "sample" models from $P(\theta|L)$
    - [Dagan & Engelson, ICML'95; McCallum & Nigam, ICML'98]
  - standard ensembles (e.g., bagging, boosting)
    - [Abe & Mamitsuka, ICML'98]

- how to measure disagreement:
  - "XOR" committee classifications
  - view vote distribution as probabilities,
    use uncertainty measures (e.g., entropy)

Batch-based active learning

Active sensing

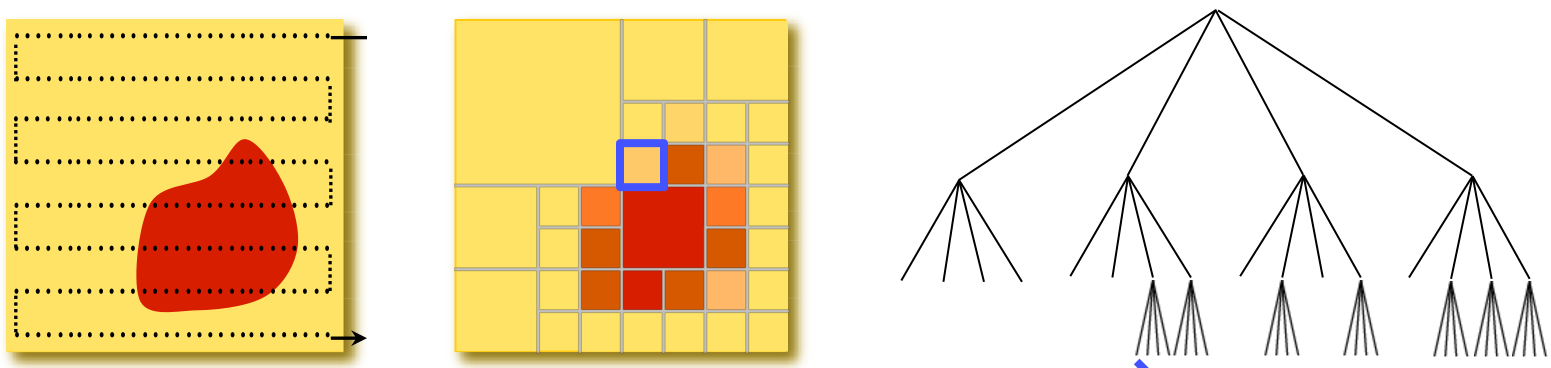




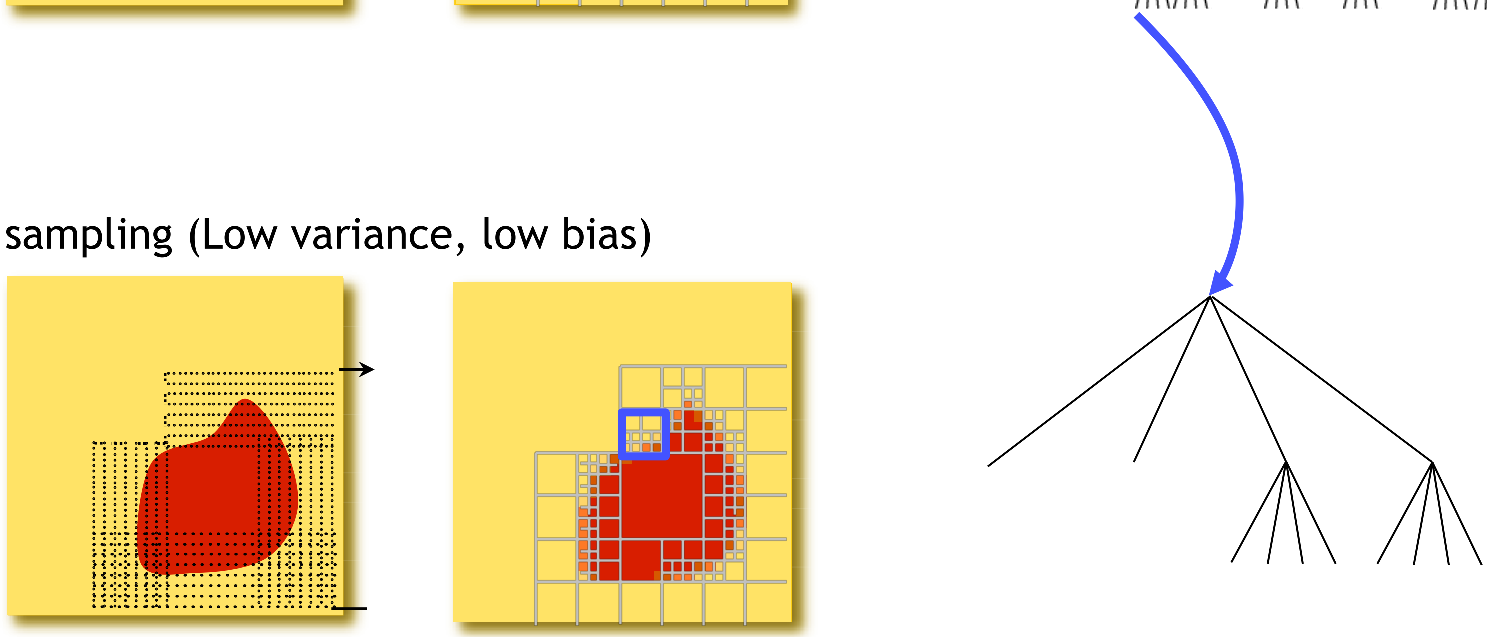
wireless sensor networks/mobile sensing

Batch-based active learning

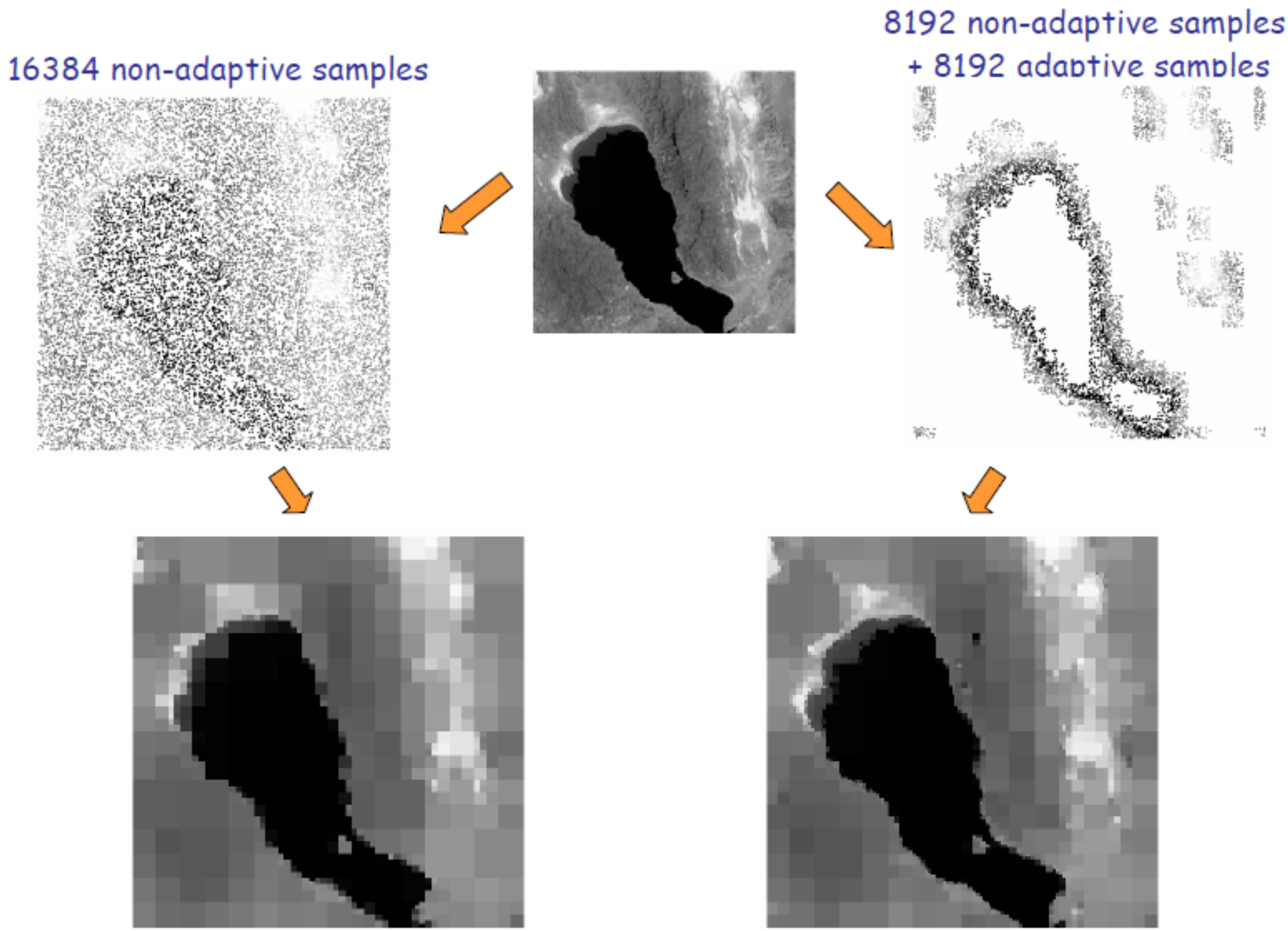
Coarse sampling (Low variance, bias limited)



Refine sampling (Low variance, low bias)



Active Learning for Terrain Mapping

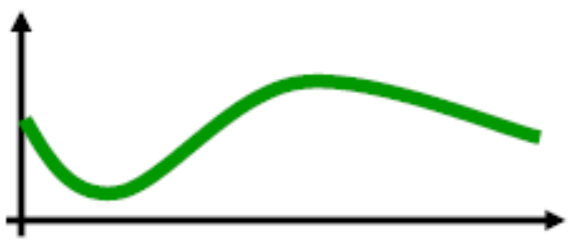


When does active learning work?

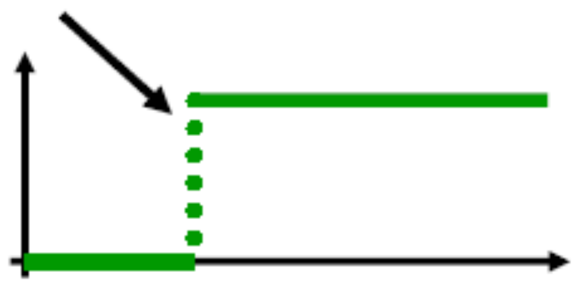
1-D

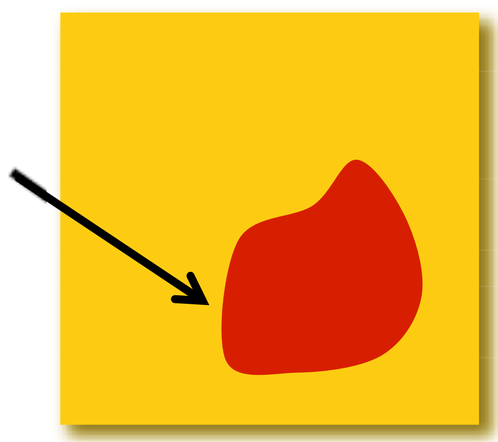
2-D

[Castro et al.,'05]



Passive = Active





Passive

$$\epsilon \sim n^{-\frac{1}{d}}$$

Active

$$\epsilon \sim n^{-\frac{1}{d-1}}$$

Active learning is useful if complexity of target function is localized – labels of some data points are more informative than others.

Active vs. Semi-Supervised

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

uncertainty sampling
query instances the model is least confident about

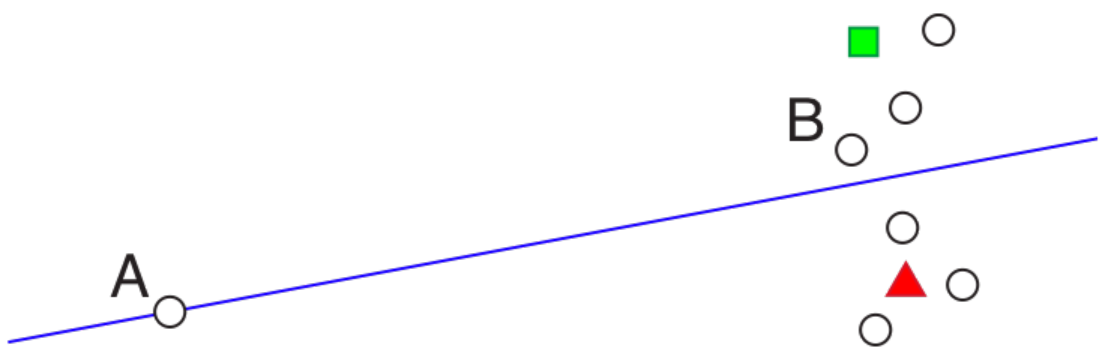
Generative model expectation-maximization (EM)
propagate confident labelings among unlabeled data

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 U [Settles & Craven, EMNLP'08]

$$\phi_{ID}(x) = \boxed{\phi(x)} \times \boxed{\left(\frac{1}{U} \sum_{u \in \mathcal{U}} \mathrm{sim}(x, u) \right)^{\beta}}$$

"base" informativeness

density term

- use U to estimate $P(x)$ 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 U if x is added to L

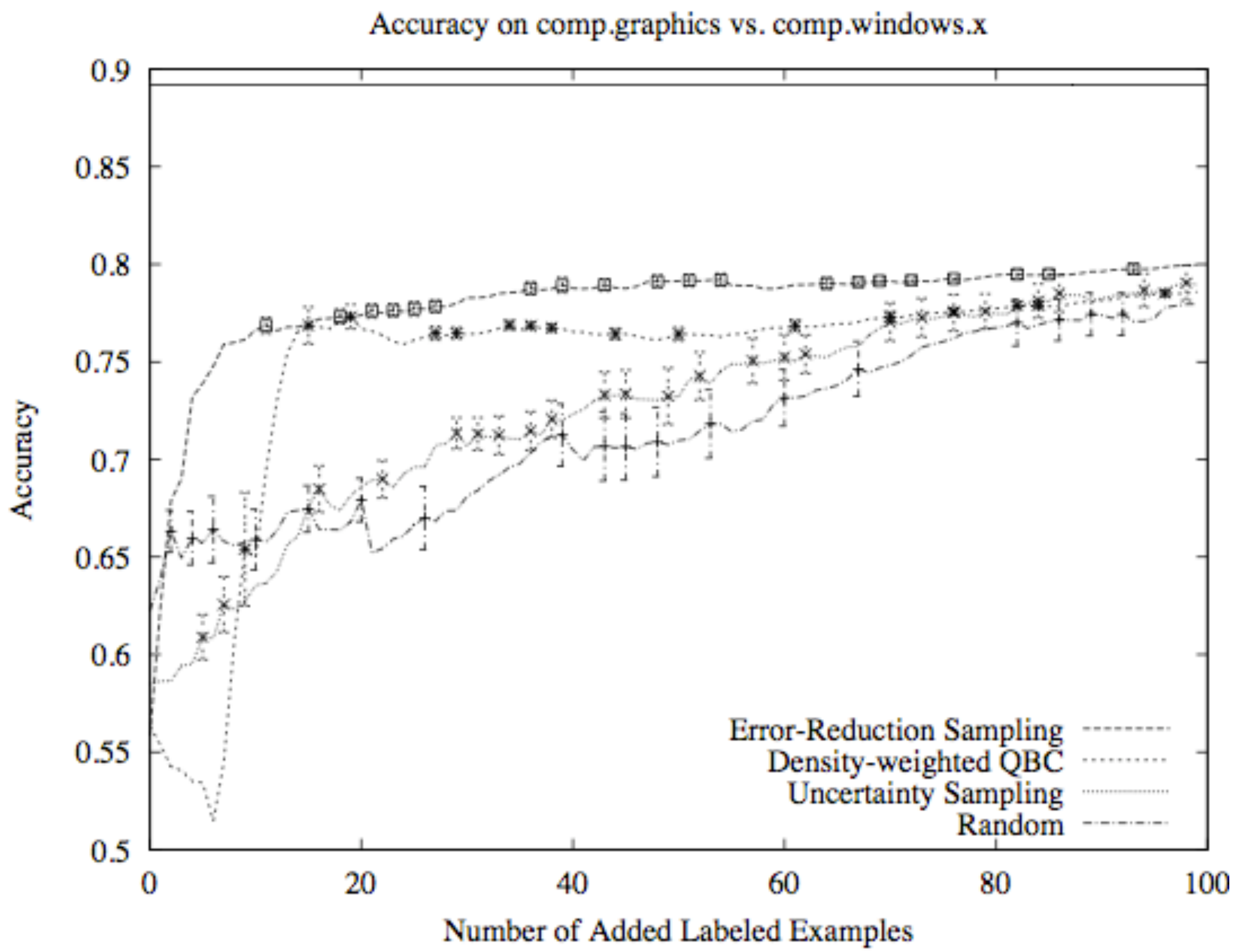
expectation over
possible labelings of x

$$R(x) = \sum_{u \in \mathcal{U}} E_y \left[\underbrace{1 - P_{\theta + <x,y>}(y^* | u)}_{\text{uncertainty of } u \text{ after retraining with } x} \right]$$

sum over
unlabeled instances

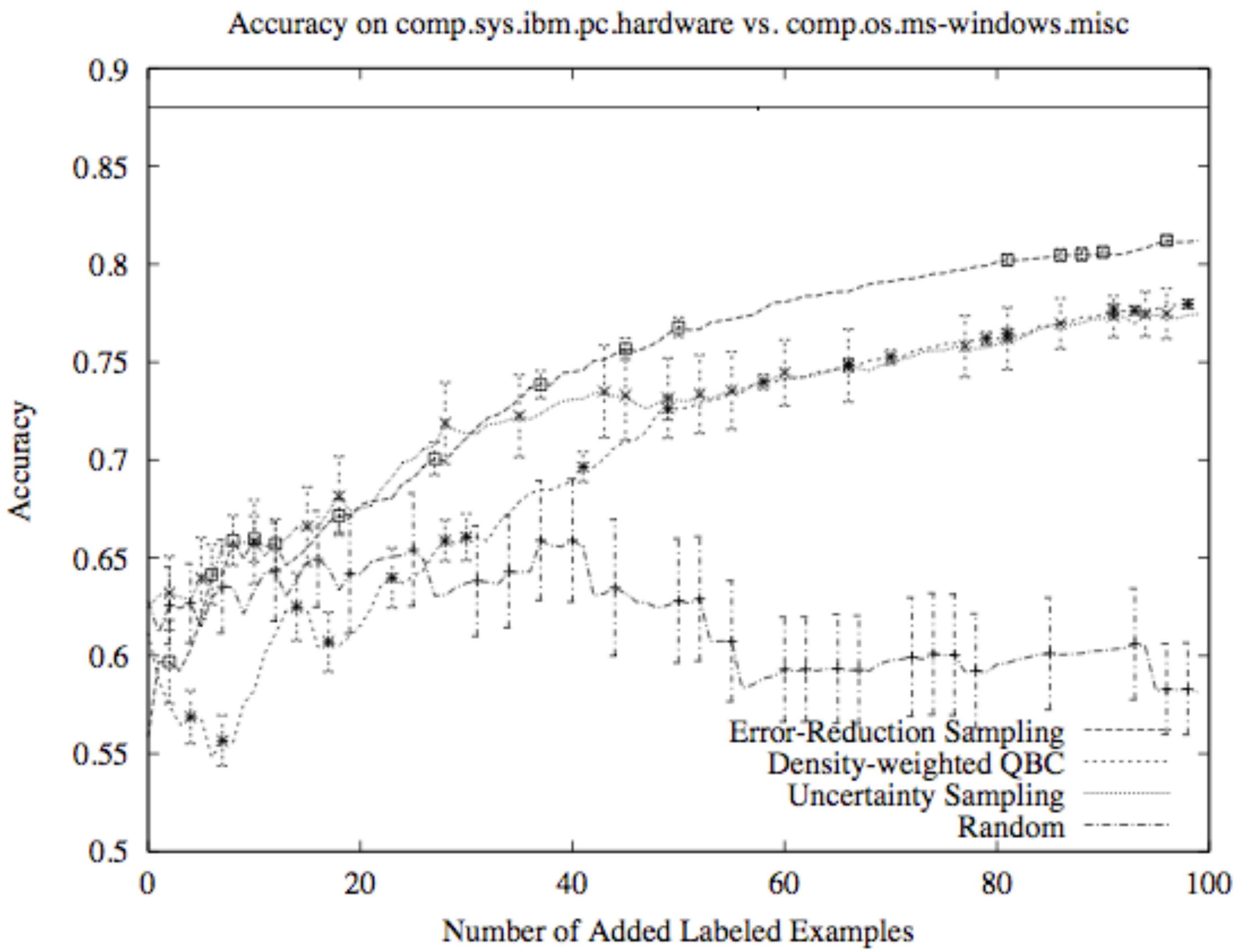
[Roy & McCallum, ICML'01]

Text Classification Examples



[Roy & McCallum, ICML'01]

Text Classification Examples



Active Learning Scenarios

Query synthesis: construct desired query/questions

Stream-based selective sampling: unlabeled data presented in a stream, decide whether or not to query its label

Pool-based active learning: given a pool of unlabeled data, select one and query its label

Alternate Settings

So far we focused on querying labels for unlabeled data.

Other query types:

Active feature acquisition – deciding whether or not to obtain a particular feature, e.g. features such as gene expressions might be correlated.

Multiple Instance active learning - one label for a bag of instances, e.g. label for a document (bag of instances) but can query passages (instance) – coarse-scale labels are cheaper

Other settings:

Cost-sensitive active learning – some labels may be more expensive than others, e.g. collecting patient vitals vs. complex and expensive medical procedures for diagnosis.

Multi-task active learning – if each label provides information for multiple tasks, which instances should be queried so as to be maximally informative across all tasks, e.g. an image can be labeled as art/photo, nature/man-made objects, contains a face or not.

Active Learning Summary

- Binary bisection
- Uncertainty sampling
- Query-by-committee
- Density Weighting
- Estimated Error Reduction

- Extensions – Active Feature acquisition, Multiple-instance active learning, Cost-sensitive active learning, Multi-task active learning

Active learning is a powerful tool if complexity of target function is localized – labels of some data points are more informative than others.