Active Learning

Yi Zhang
10-701, Machine Learning, Spring 2011
April 20th, 2011

Some pictures are from Burr and Aarti's lectures

Outline

- **Basic idea of active learning**
- Supervised, semi-supervised, and active learning
- Uncertainty sampling
- Version space reduction and query by committee
- Expected error reduction
- Other active learning methods

Why active learning?

- Learning classifiers using labeled examples

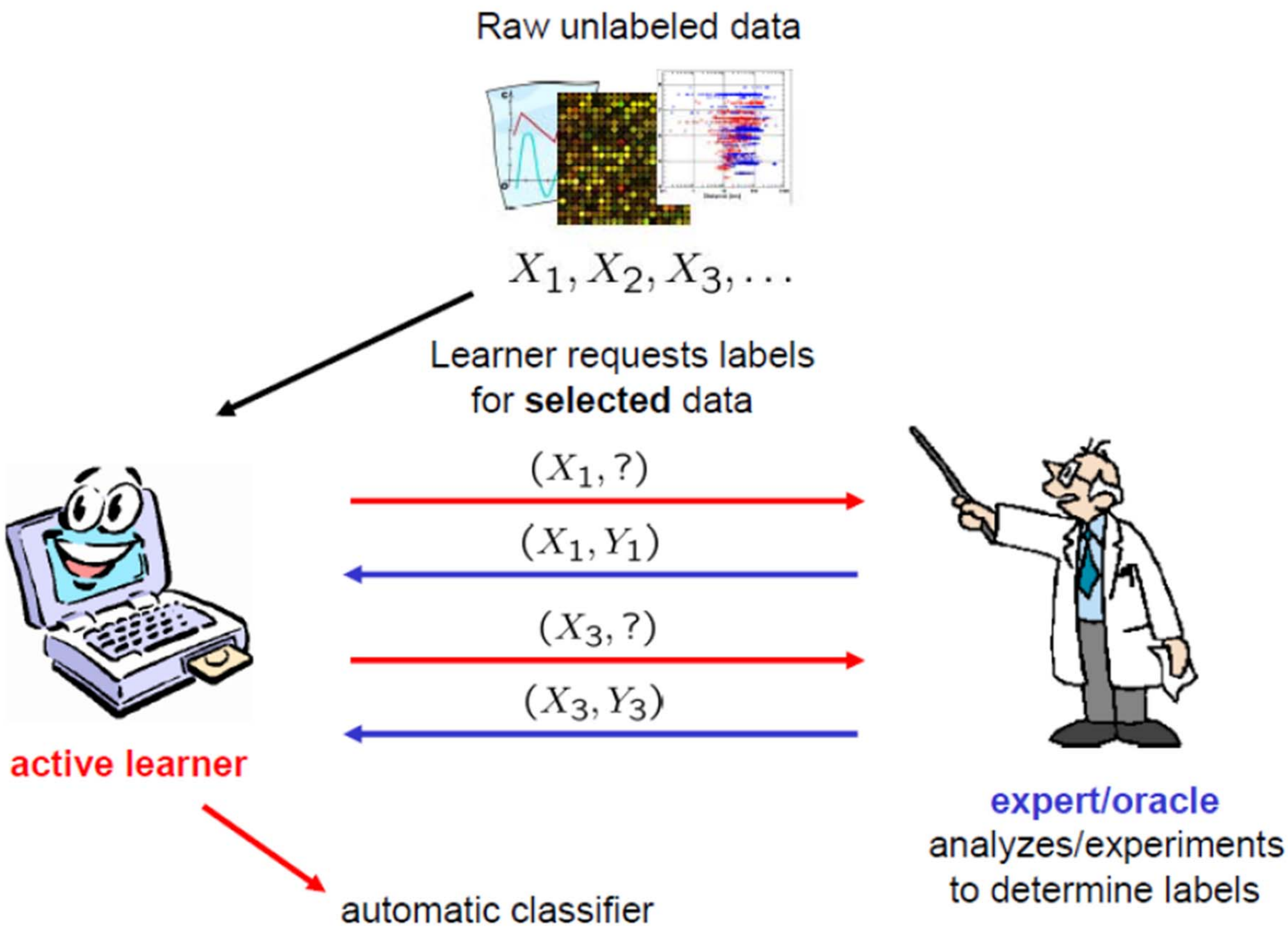
$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$$

- But obtaining labels is
  - Time consuming, e.g., document classification
  - Expensive, e.g., medical decision (need doctors)
  - Sometimes dangerous, e.g., landmine detection

Active learning

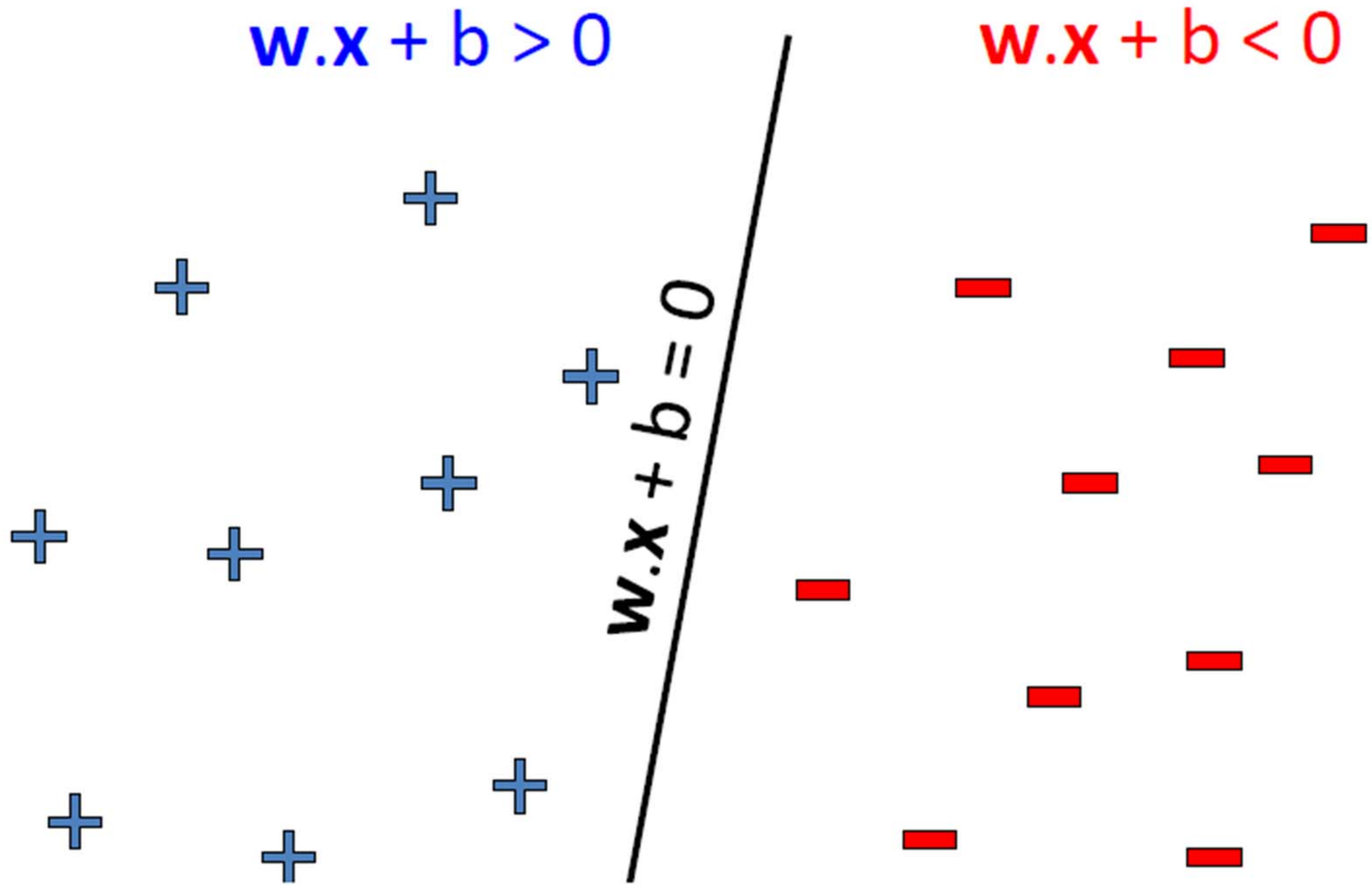
- The learner actively chooses which examples to label !
- Goal: reduce the number of labeled examples needed for learning

Active learning



Can active learning work?

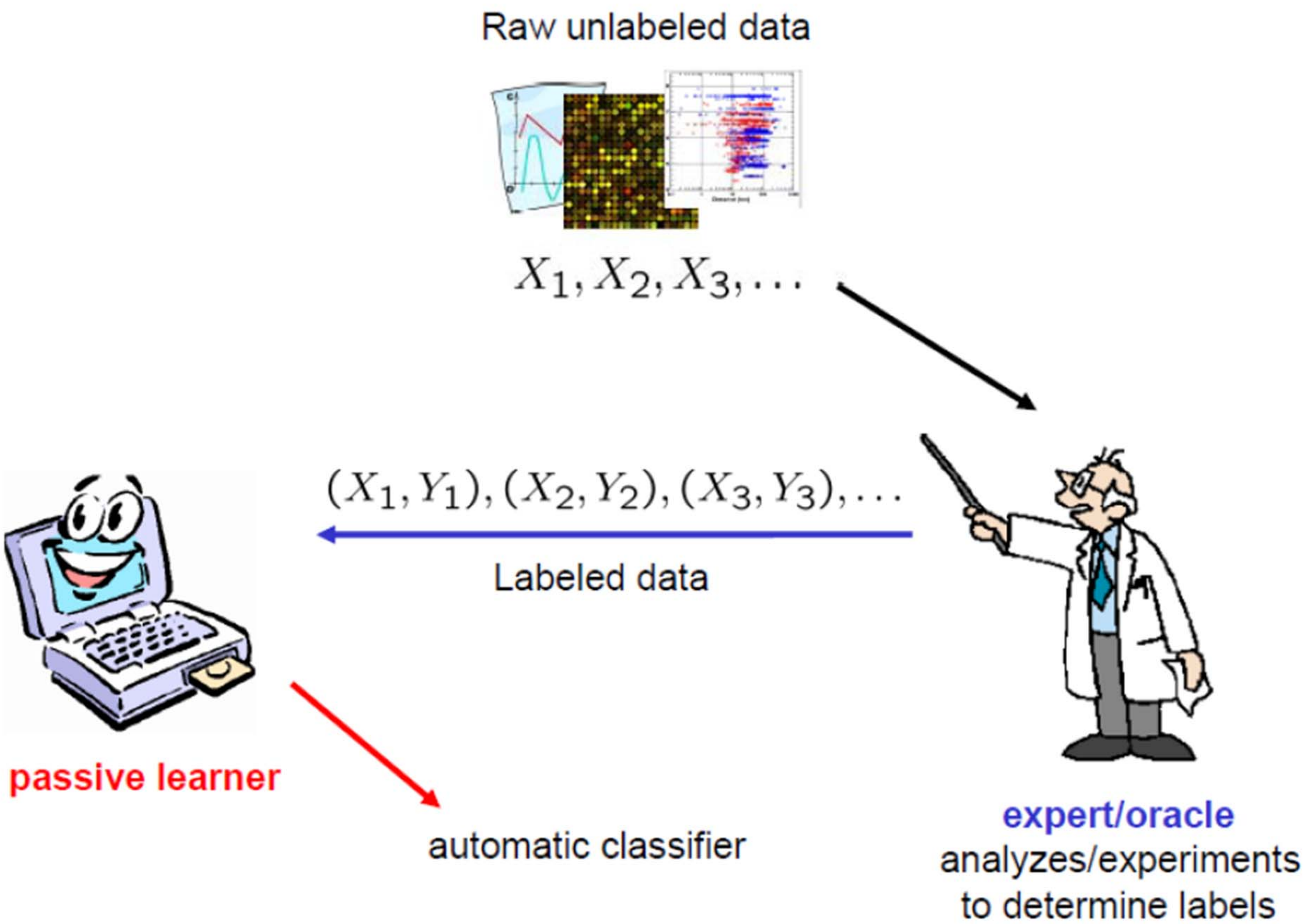
- Are all labeled examples equally important?



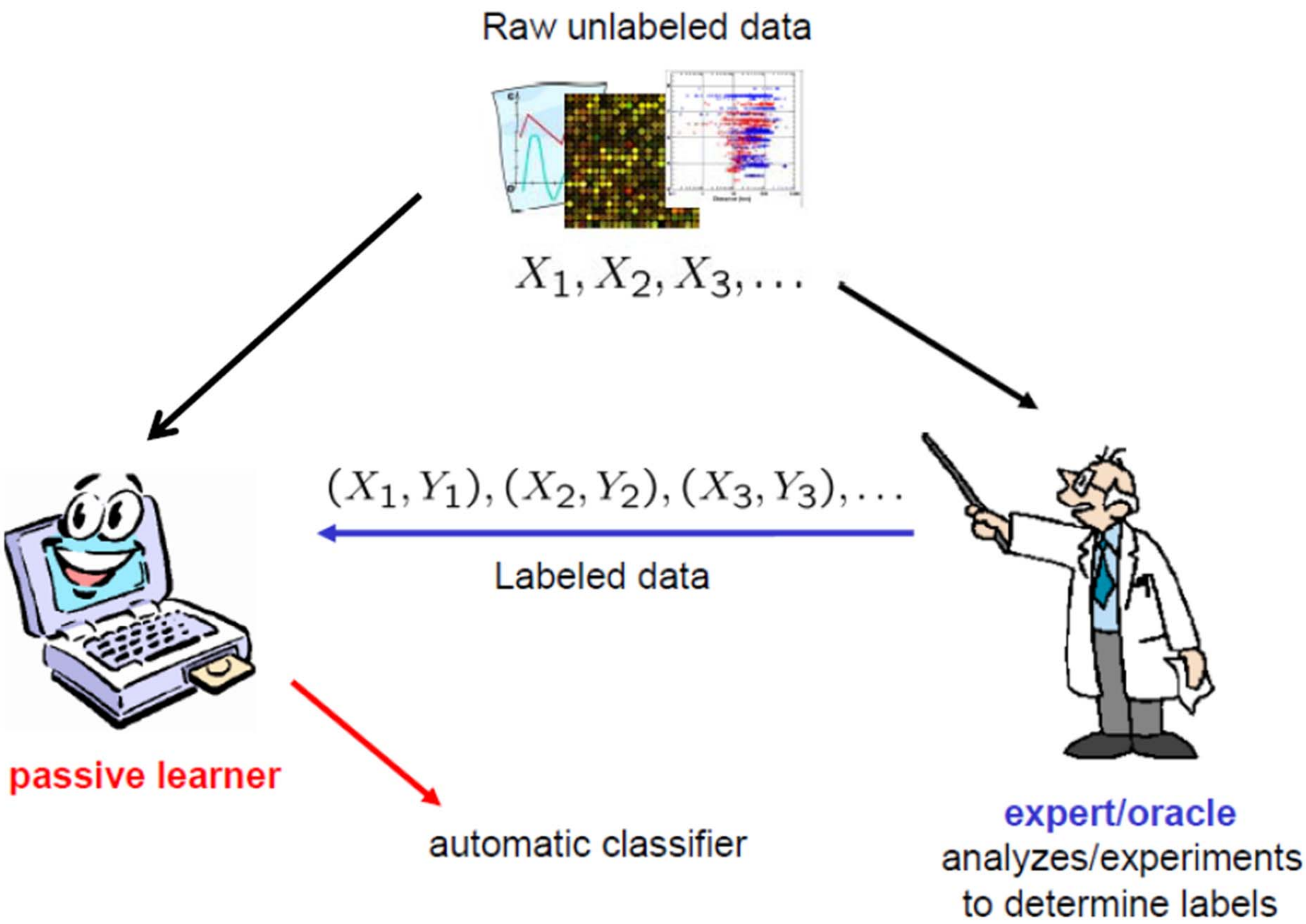
Outline

- Basic idea of active learning
- **Supervised, semi-supervised, and active learning**
- Uncertainty sampling
- Version space reduction and query by committee
- Expected error reduction
- Other active learning methods

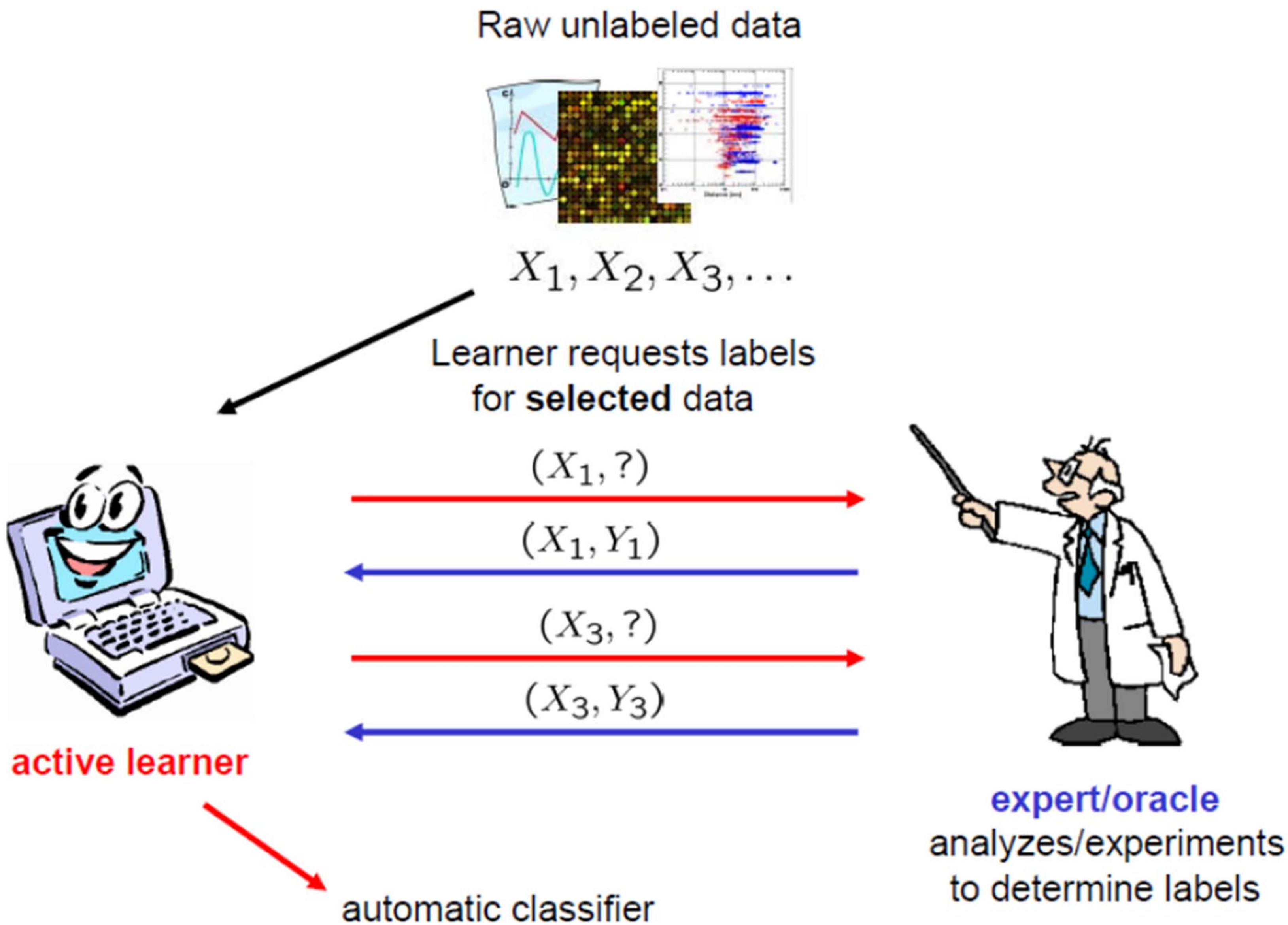
(Passive) supervised learning



Semi-supervised learning



Active learning



Active learning vs. semi-supervised learning

- The same goal:
  - Attain good learning performance (e.g., classification accuracy) without demanding too many labeled examples
- Different approaches
  - Semi-supervised learning: use unlabeled data
  - Active learning: choose labeled examples

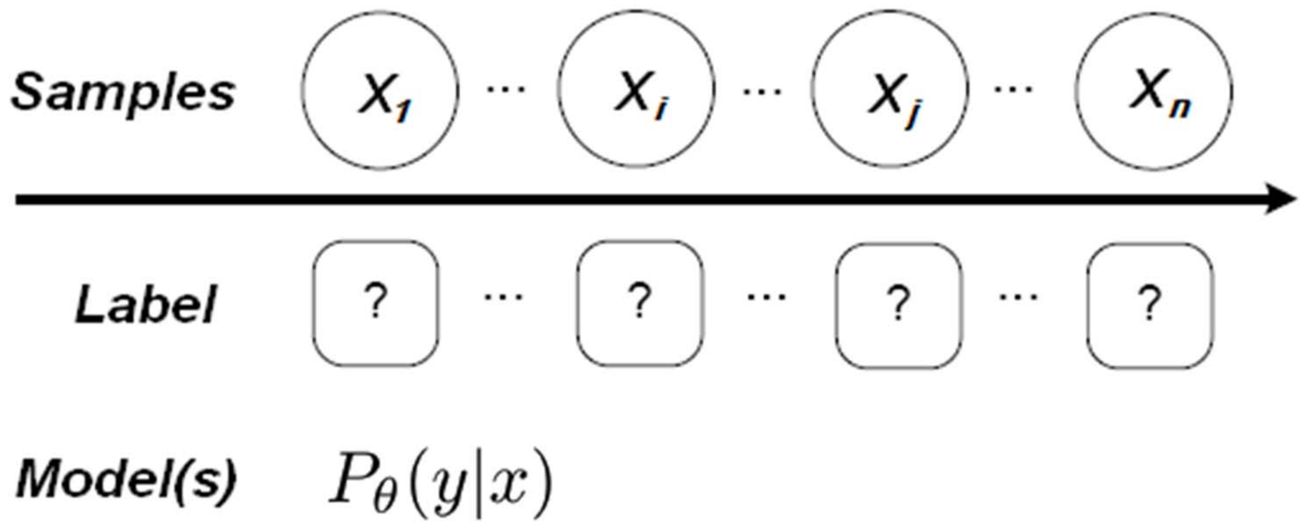
Outline

- Basic idea of active learning
- Supervised, semi-supervised, and active learning
- **Uncertainty sampling**
- Version space reduction and query by committee
- Expected error reduction
- Other active learning methods

Three active learning scenarios

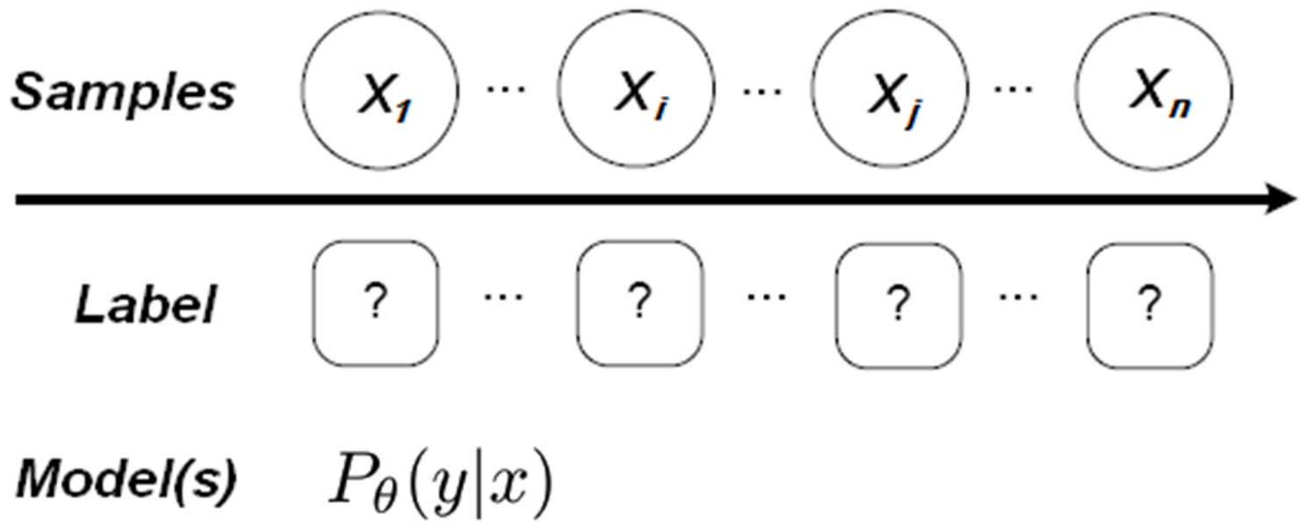
- Query synthesis
  - learner constructs examples for labeling
- Selective sampling
  - Unlabeled data come as a stream
  - For each arrived point, learner decides to query or discard
- Pool-based active learning **(*)**
  - Given a pool of unlabeled data
  - Learner chooses from the pool for labeling

Pool-based active learning



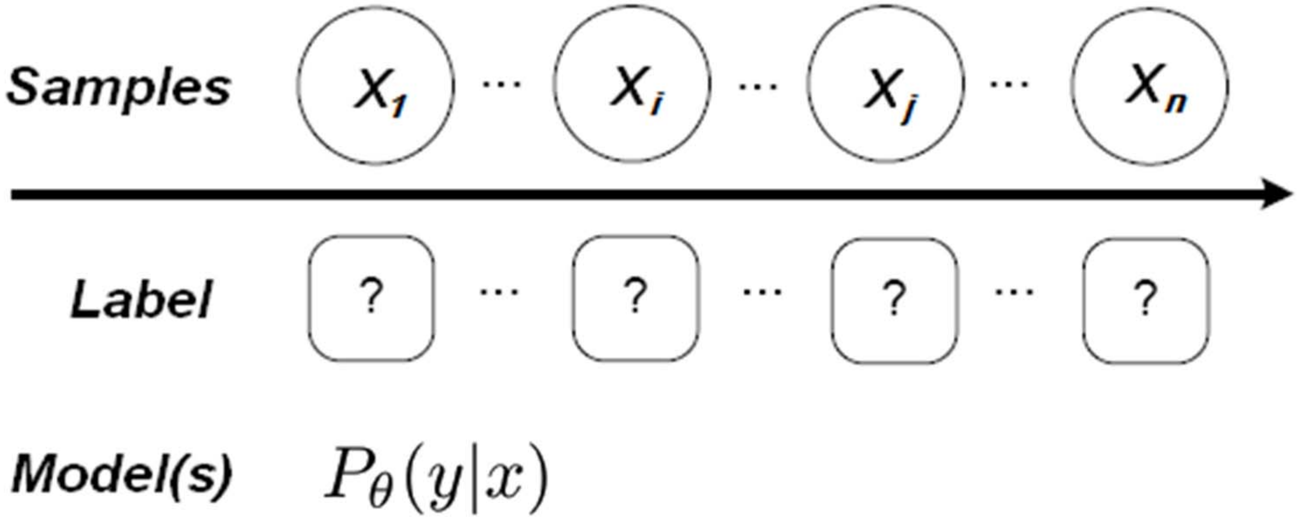
- A pool of unlabeled samples
- A learned model (or a set of models)
- Choose: which sample to label next?

Uncertainty sampling



- Uncertainty sampling
  - Query the sample x that the learner is most uncertain about
  - How to measure "uncertainty"?

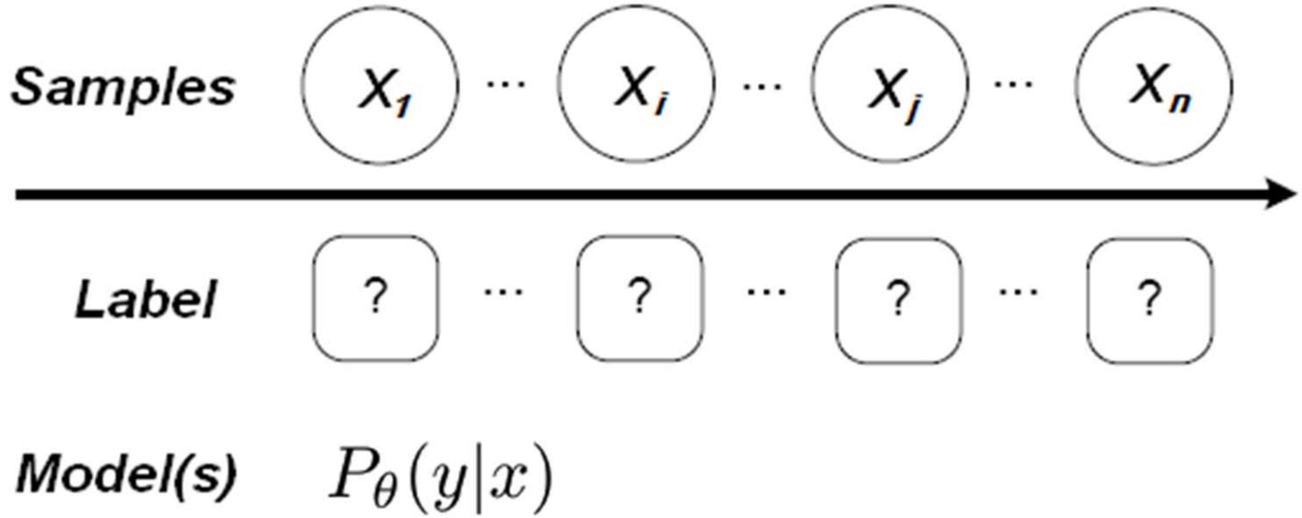
Uncertainty sampling



- Uncertainty sampling
  - Maximum entropy [Dagan & Engelson, ICML'95]

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

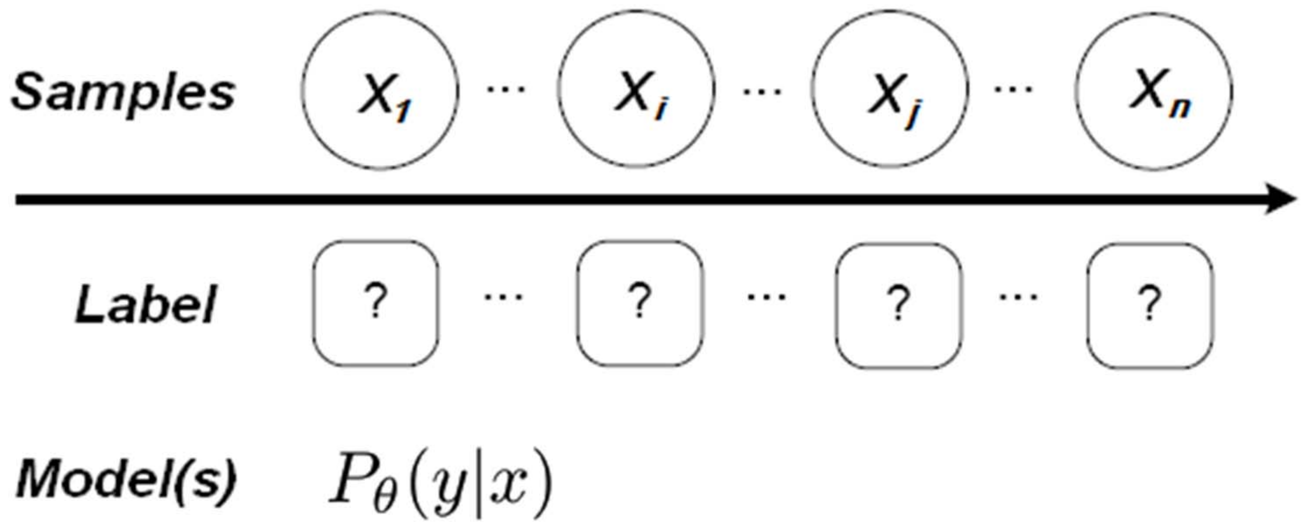
Uncertainty sampling



- Uncertainty sampling
  - Smallest margin (between most likely and second most likely labels) [Scheffer et al., CAIDA'01]

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

Uncertainty sampling



- Uncertainty sampling
  - Least confidence [Culotta & McCallum, AAAI'05]

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

Uncertainty sampling

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

Uncertainty sampling

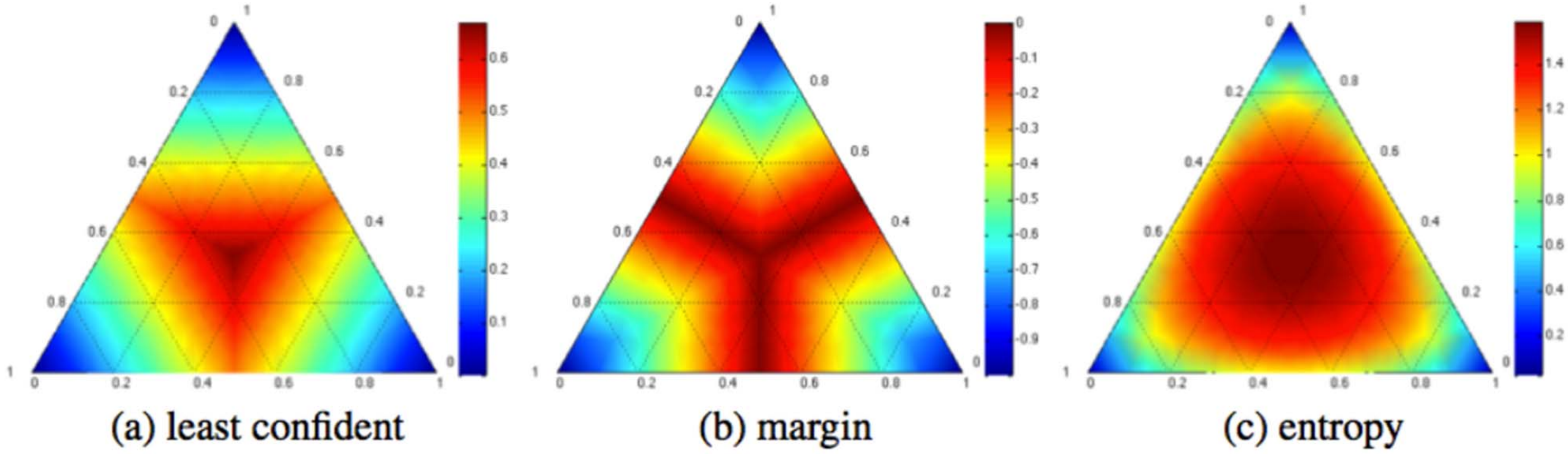


illustration of preferred (dark red) posterior distributions in a 3-label classification task

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

Outline

- Basic idea of active learning
- Supervised, semi-supervised, and active learning
- Uncertainty sampling
- **Version space reduction and query by committee**
- Expected error reduction
- Other active learning methods

Version space

- The set of classifiers that are consistent with labeled examples

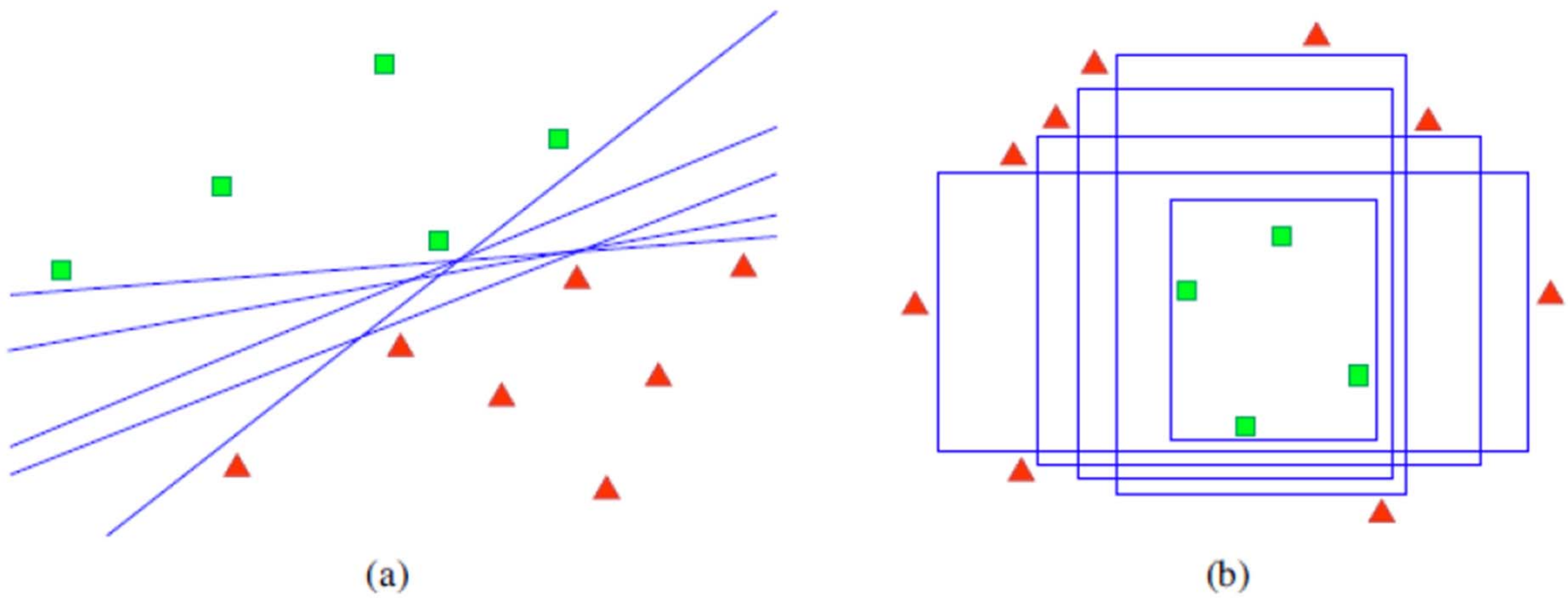
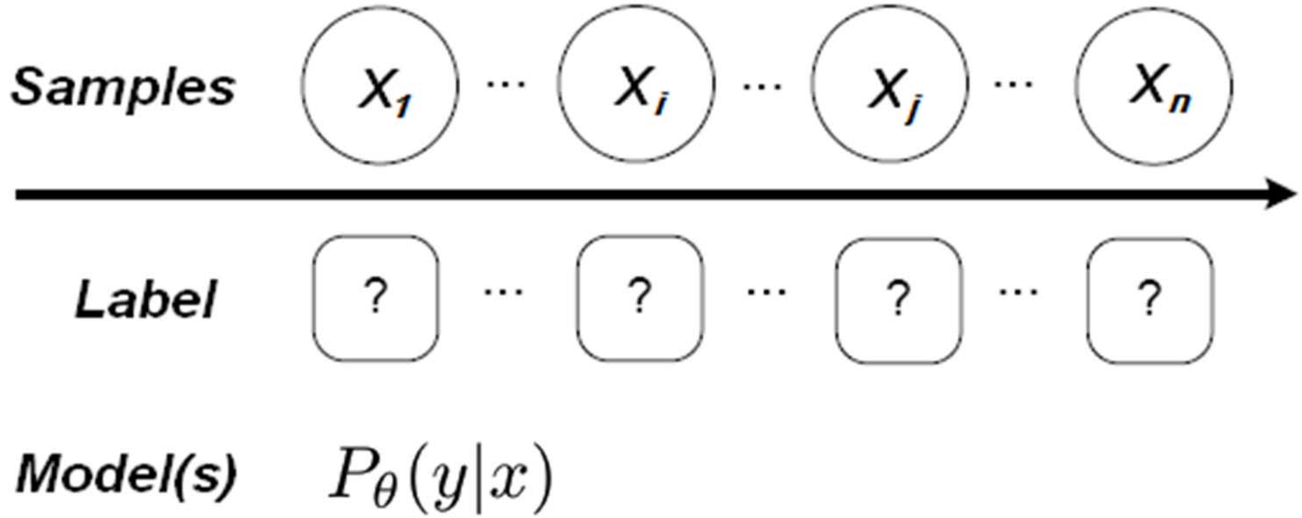


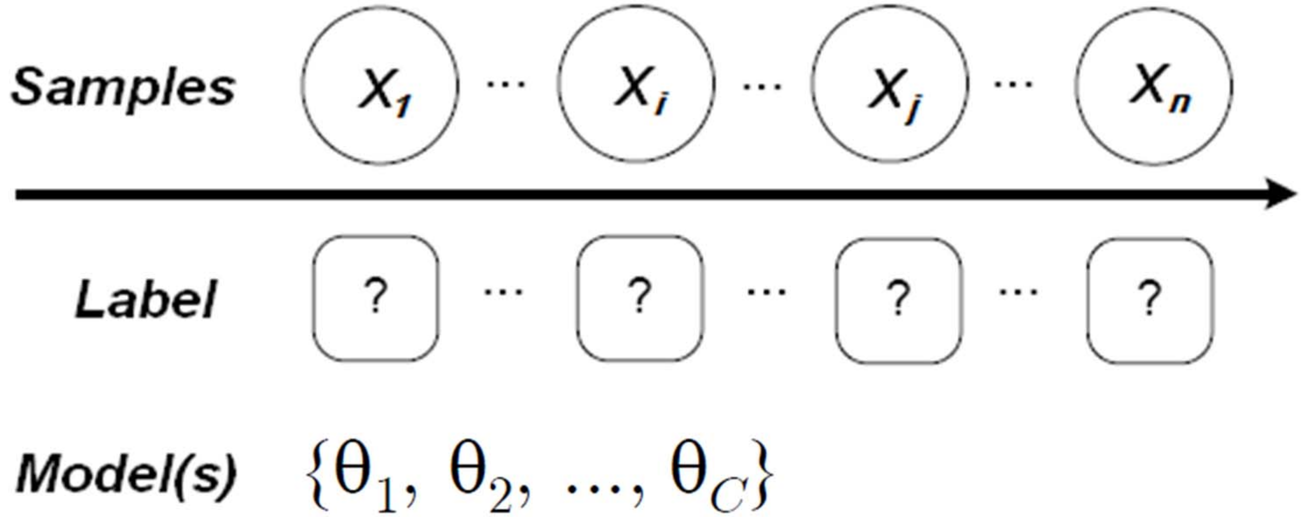
Figure 6: Version space examples for (a) linear and (b) axis-parallel box classifiers.

Recall: uncertainty sampling



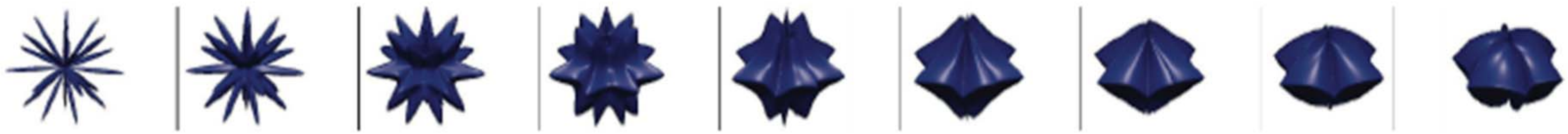
- Uncertainty sampling

Version space reduction



- Version space reduction
  - Query the sample x that reduces the version most
    - "Expected" reduction of version space
    - "Worst case" reduction of version space
    - "Best-case" reduction of version space

A simplifying example



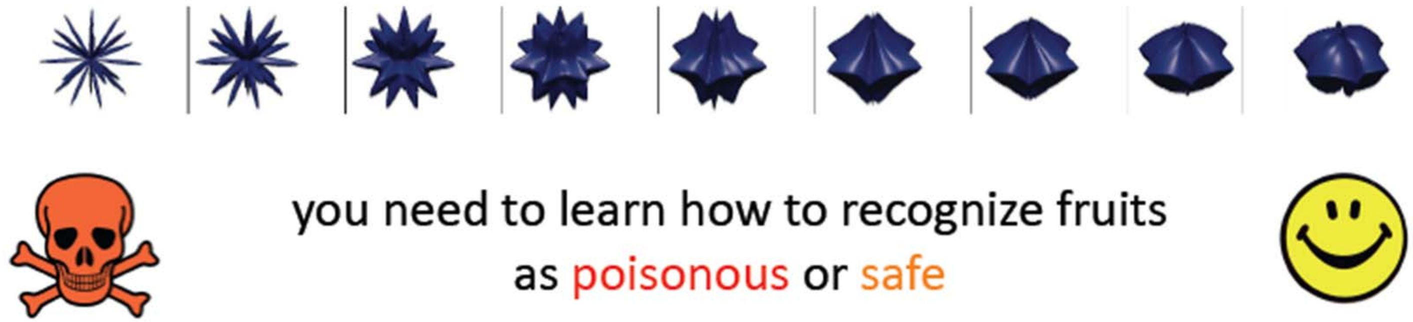
you need to learn how to recognize fruits
as poisonous or safe

- How to query? Binary search !

A simplifying example

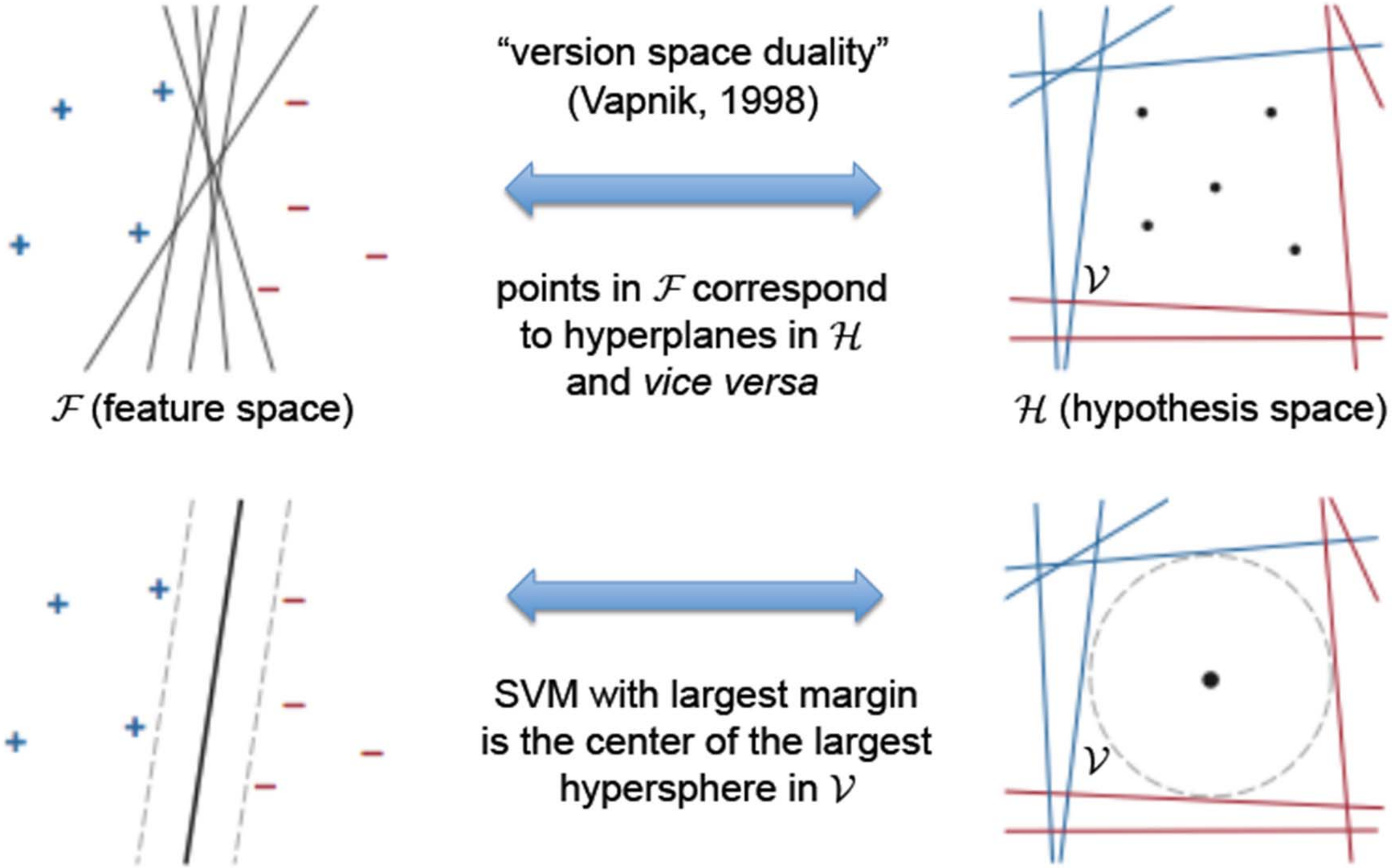
- Why binary search ?

  - Recall: version space

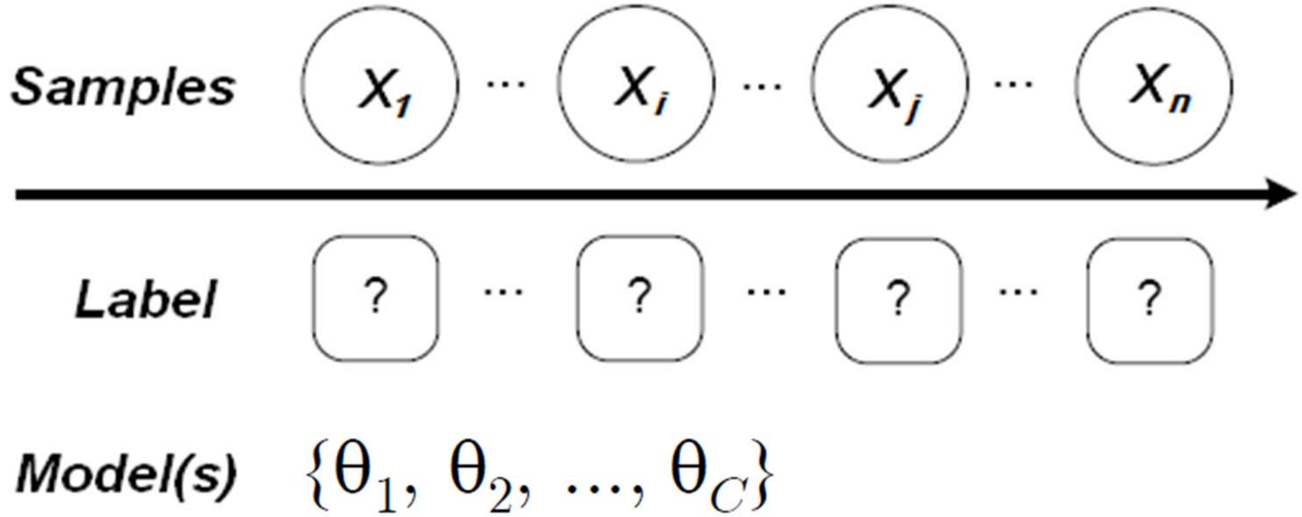


  - Maximize the “worst-case” reduction of version space

Version space reduction for SVMs



Query by committee



- Query by committee
  - Keep a committee of classifiers $\{\theta_1, \theta_2, \ldots, \theta_C\}$
  - Query the instance that the committee members disagree

Query by committee

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

Query by committee

- Query by committee
  - Keep a committee of classifiers
  - Query the instance that the committee members disagree
- QBC as version space reduction
  - Committee is an approximation to the version space
- QBC as uncertainty sampling
  - Use committee members to measure the uncertainty

Outline

- Basic idea of active learning
- Supervised, semi-supervised, and active learning
- Uncertainty sampling
- Version space reduction and query by committee
- **Expected error reduction**
- Other active learning methods

Expected error (uncertainty) 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

Outline

- Basic idea of active learning
- Supervised, semi-supervised, and active learning
- Uncertainty sampling
- Version space reduction and query by committee
- Expected error reduction
- **Other active learning methods**

Other active learning methods

- Active learning is an active research area ☺
- Cost sensitive active learning
  - Labeling costs of examples differ
- Batch mode active learning
  - Query multiple instances at once
- Multi-task active learning
  - Query labels for multiple learning tasks
- See survey of Burr Settles ! [Settles, 2008]