

# Active Learning

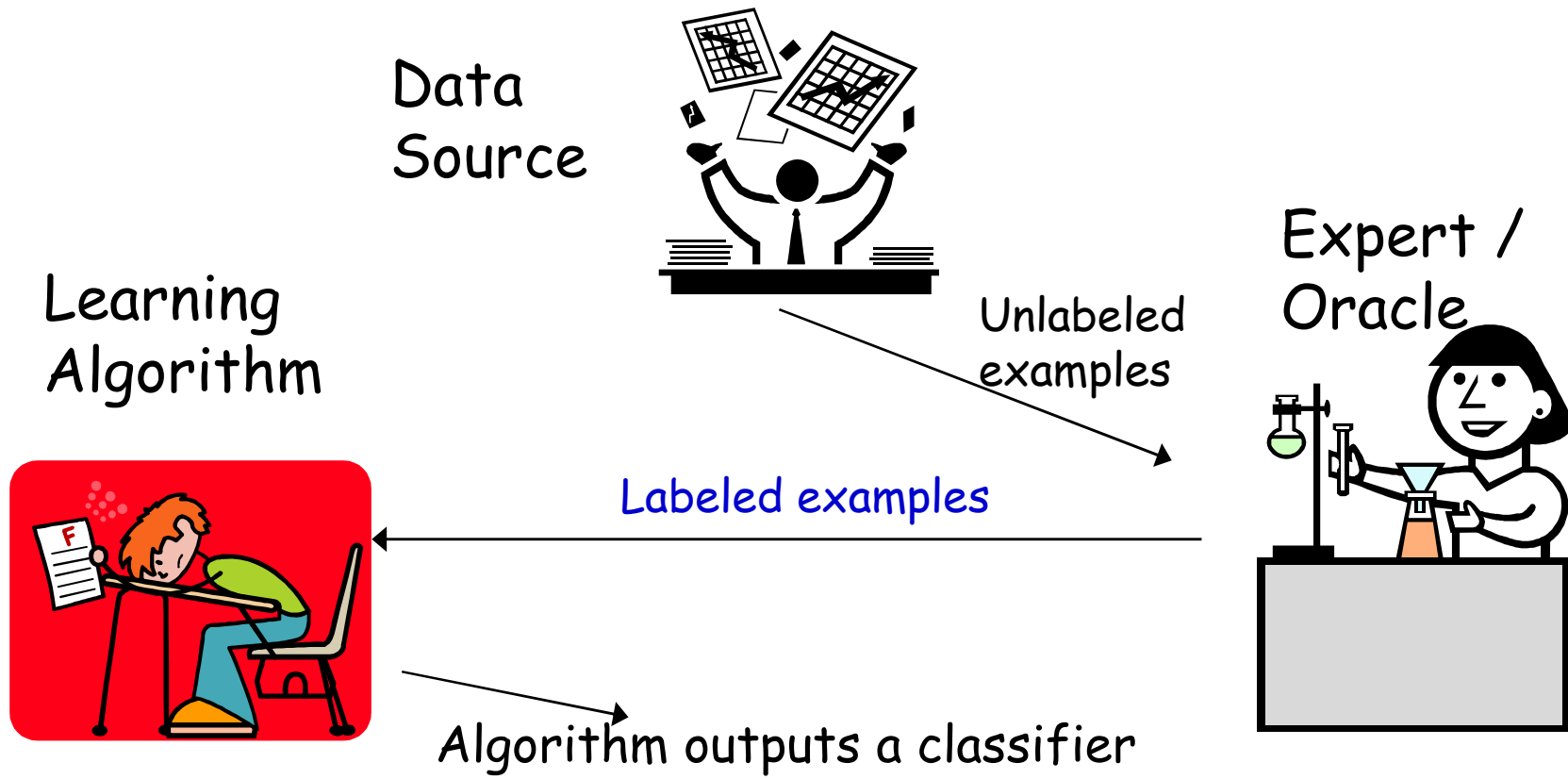
Maria-Florina Balcan

04/12/07

# Outline

- Brief Overview of Active Learning.
- Active Learning of Linear Separators.  
( $C$  - homogeneous linear separators in  $\mathbb{R}^d$ ,  
 $D$  - uniform distribution over unit sphere. )

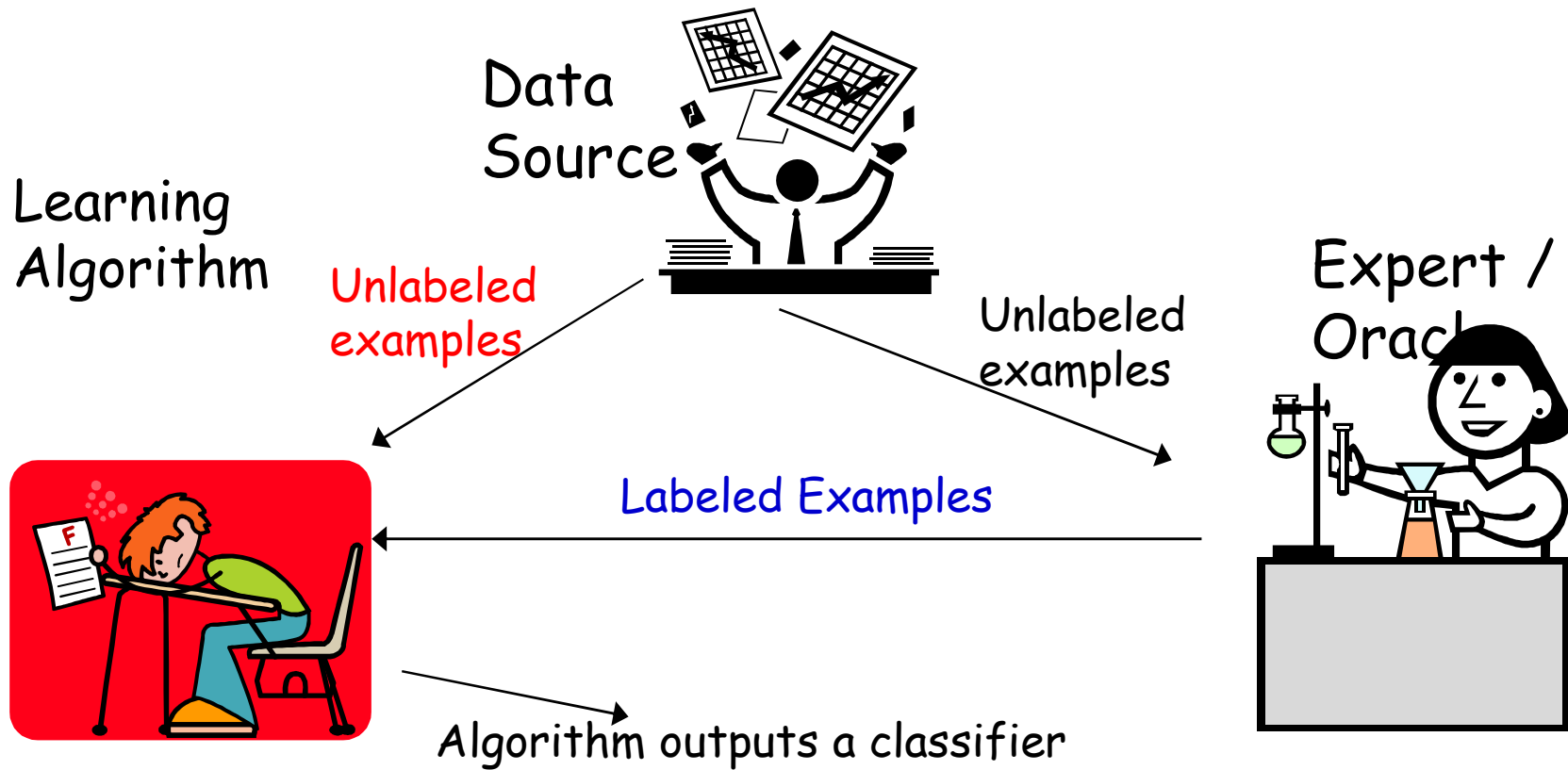
# Supervised Passive Learning



# Incorporating Unlabeled Data in the Learning process

- In many settings, unlabeled data is cheap & easy to obtain, labeled data is much more expensive.
  - Web page, document classification
  - OCR, Image classification

# Semi-Supervised Passive Learning

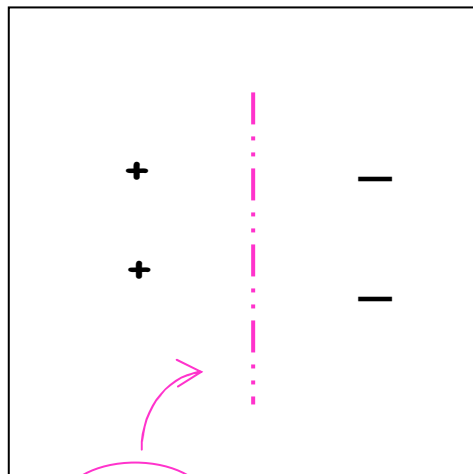


# Semi-Supervised Passive Learning

- Several methods have been developed to try to use unlabeled data to improve performance, e.g.:
  - Transductive SVM [Joachims '98]
  - Co-training [Blum & Mitchell '98], [BBY04]
  - Graph-based methods [Blum & Chawla01], [ZGL03]

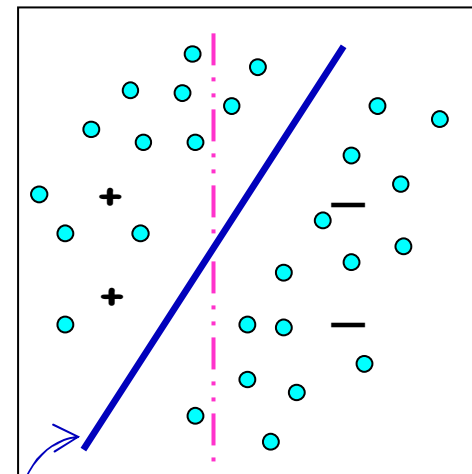
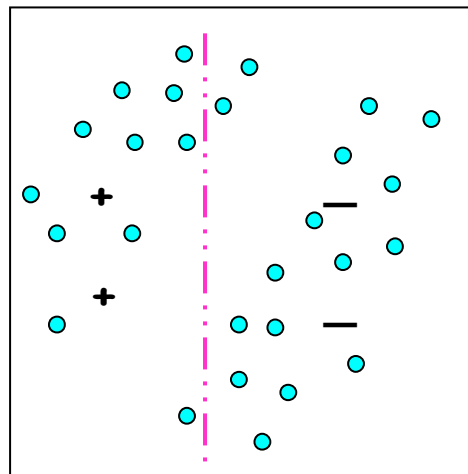
# Semi-Supervised Passive Learning

- Several methods have been developed to try to use unlabeled data to improve performance, e.g.:
  - **Transductive SVM** [Joachims '98]
  - **Co-training** [Blum & Mitchell '98], [BBY04]
  - **Graph-based methods** [Blum & Chawla01], [ZGL03]



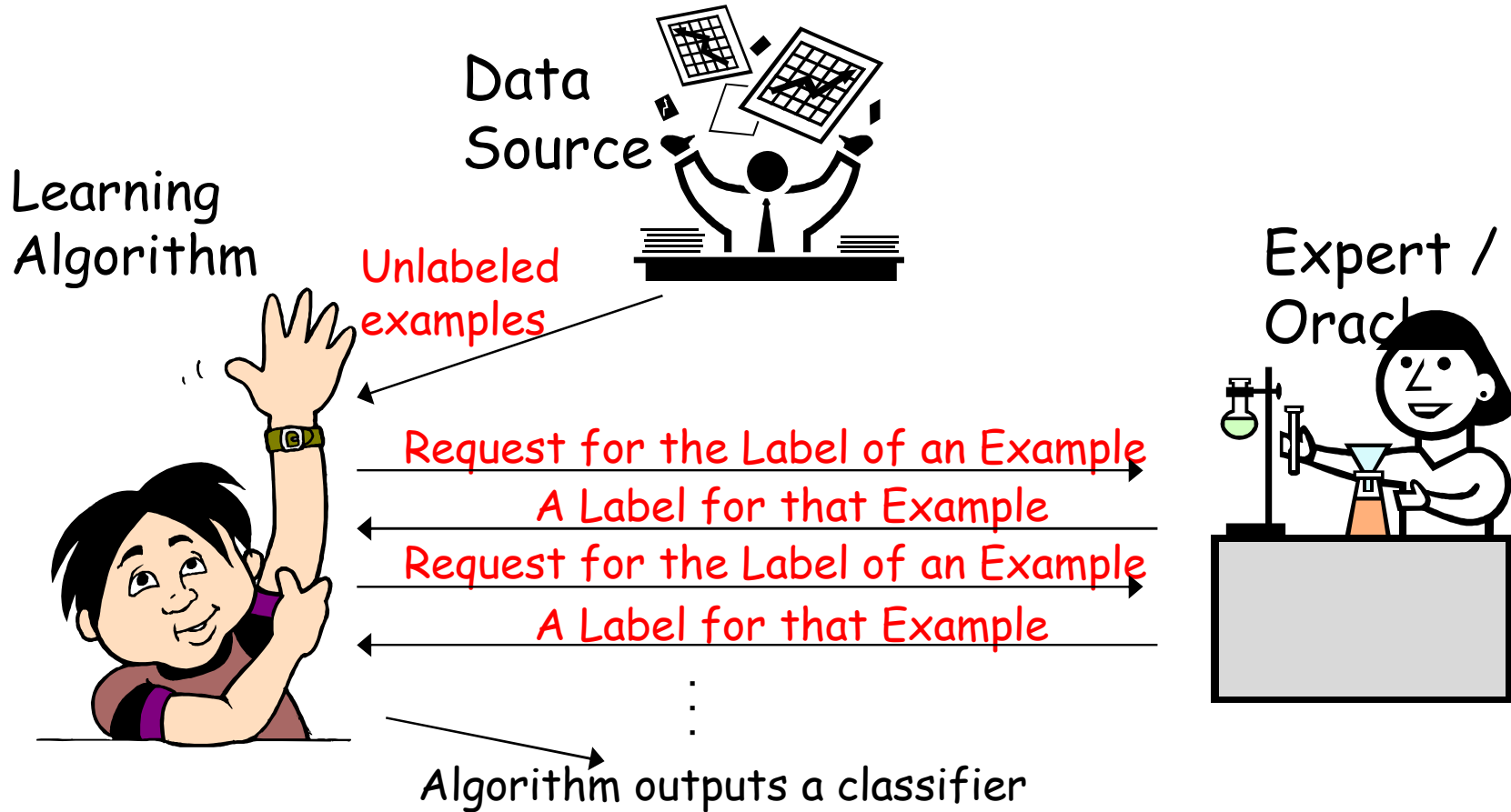
SVM

Labeled data only



Transductive SVM

# Active Learning



# What Makes a Good Active Algorithm?

- Guaranteed to output a good classifier for most learning problems.
- Doesn't make too many label requests.

Choose the label requests carefully, to get **informative** labels.

# Active Learning

- We get to see unlabeled data first, and there is a charge for every label.
- The learner has the ability to choose specific examples to be labeled:
  - The learner works harder, in order to use fewer labeled examples.
- How many labels can we save by querying adaptively?

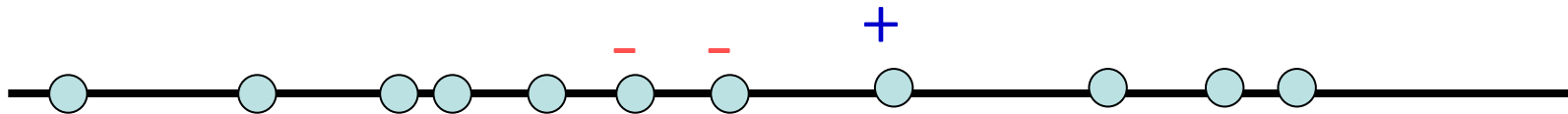
## Can adaptive querying help? [CAL92, Dasgupta04]

- Consider threshold functions on the real line:

$$h_w(x) = \begin{cases} - & x < w \\ + & x \geq w \end{cases}, \quad C = \{h_w : w \in \mathbb{R}\}$$



- Sample with  $1/\epsilon$  unlabeled examples.



- Binary search - need just  $O(\log 1/\epsilon)$  labels.

Active setting:  $O(\log 1/\epsilon)$  labels to find an  $\epsilon$ -accurate threshold.

Supervised learning needs  $O(1/\epsilon)$  labels.

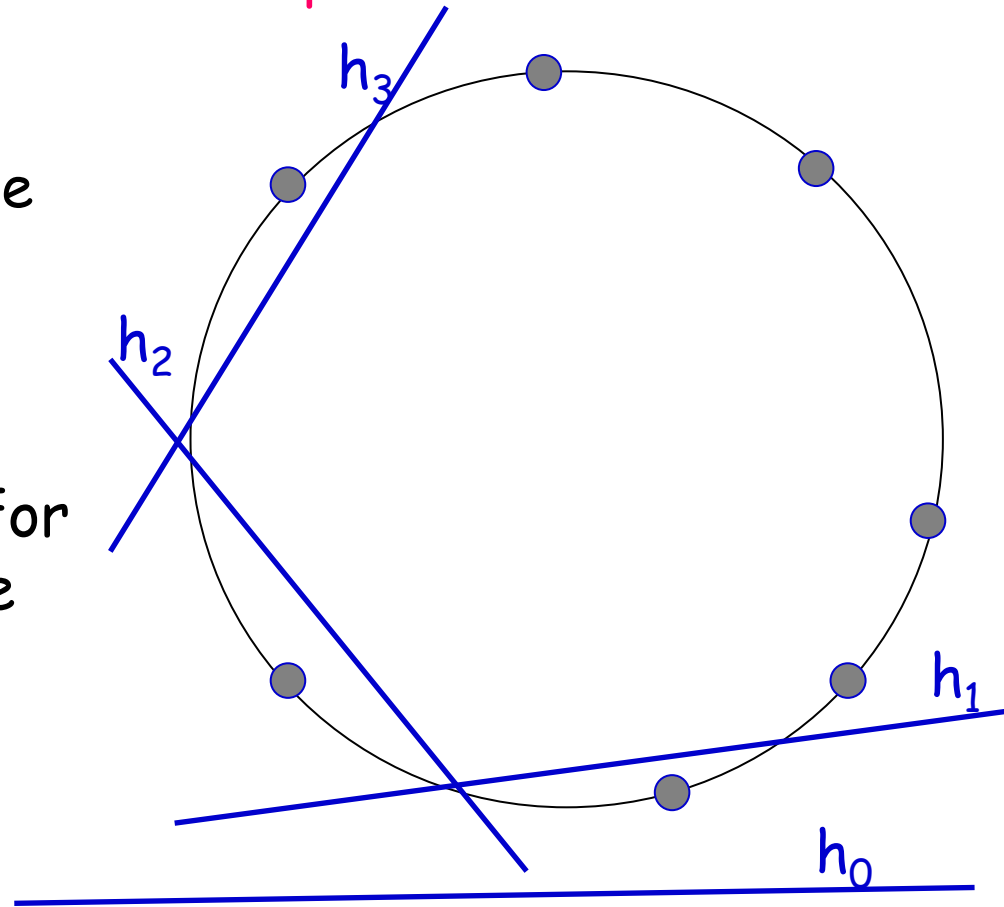
**Exponential** improvement in sample complexity  $\mathcal{J}$

# Active Learning might not help [Dasgupta04]

In general, number of queries needed depends on  $C$  and also on  $D$ .

$C = \{\text{linear separators in } \mathbb{R}^1\}$ :  
active learning reduces sample complexity substantially.

$C = \{\text{linear separators in } \mathbb{R}^2\}$ :  
there are some target hyp. for which no improvement can be achieved!  
- no matter how benign the input distr.



In this case: learning to accuracy  $\epsilon$  requires  $1/\epsilon$  labels...

# Examples where Active Learning helps

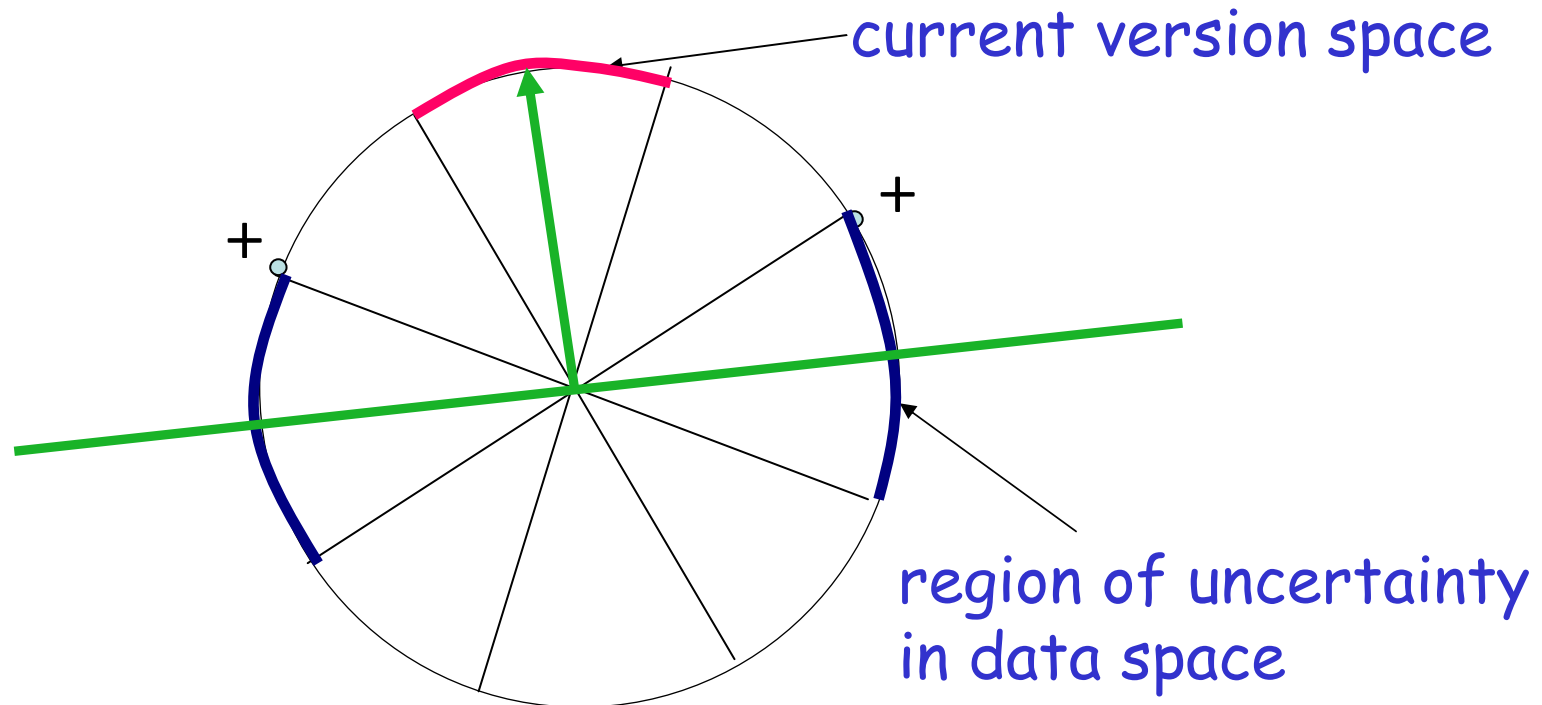
In general, **number of queries needed depends on  $C$  and also on  $D$ .**

- $C = \{\text{linear separators in } \mathbb{R}^1\}$ : active learning reduces sample complexity **substantially no matter what is the input distribution.**
- $C$  - **homogeneous** linear separators in  $\mathbb{R}^d$ ,  $D$  - **uniform distribution** over unit sphere:
  - need only  **$d \log 1/\epsilon$**  labels to find a hypothesis with error rate  $< \epsilon$ .
    - Dasgupta, Kalai, Monteleoni, COLT 2005
    - Freund et al., '97.
    - Balcan-Broder-Zhang, COLT 07

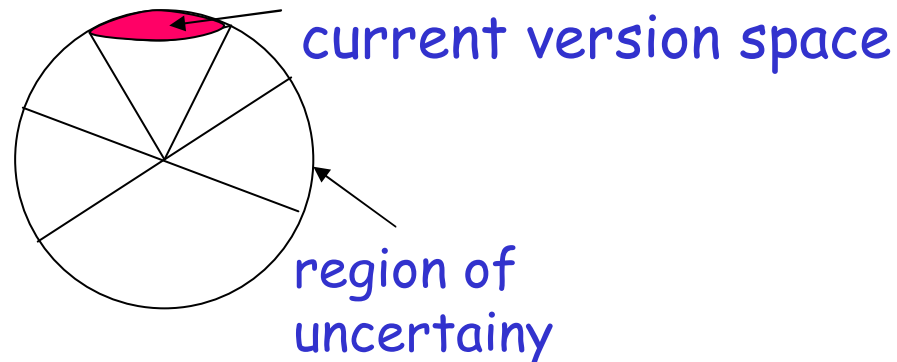
Note:  **$d (1/\epsilon)$**  needed in the **passive** learning setting. (Phil Long)

## Region of uncertainty [CAL92]

- Current **version space**: part of  $C$  consistent with labels so far.
- "**Region of uncertainty**" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)
- Example: data lies on circle in  $\mathbb{R}^2$  and hypotheses are homogeneous linear separators.



# Region of uncertainty [CAL92]

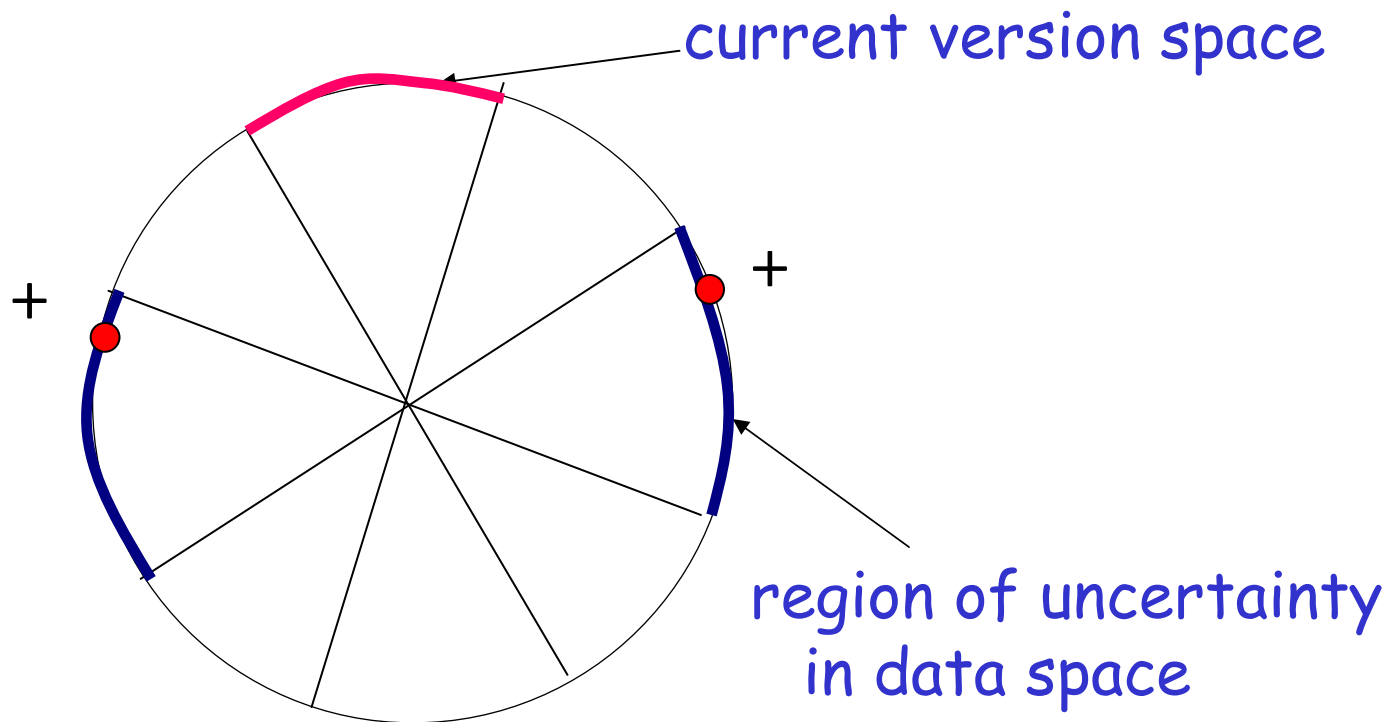


## Algorithm:

Pick a few points at random from the current region of uncertainty and query their labels.

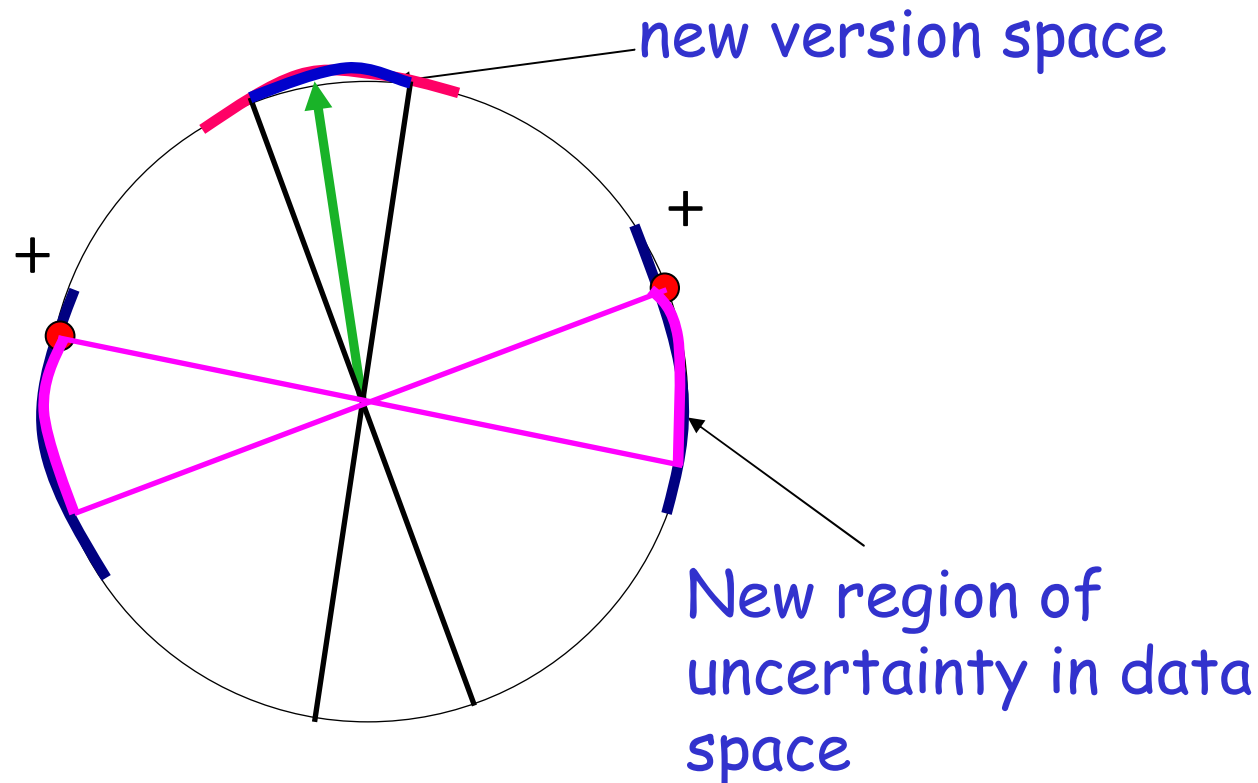
## Region of uncertainty [CAL92]

- Current **version space**: part of  $C$  consistent with labels so far.
- "**Region of uncertainty**" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)



## Region of uncertainty [CAL92]

- Current **version space**: part of  $C$  consistent with labels so far.
- "**Region of uncertainty**" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)



## Region of uncertainty [CAL92], Guarantees

**Algorithm:** Pick a few points at random from the current region of uncertainty and query their labels.

[Balcan, Beygelzimer, Langford, ICML'06]

Analyze a version of this alg. which is **robust to noise**.

- **C**- linear separators on the line, low noise, exponential improvement.
- **C** - homogeneous linear separators in  $\mathbb{R}^d$ , **D** -uniform **distribution** over unit sphere.
  - low noise, need only  $d^2 \log 1/\epsilon$  labels to find a hypothesis with error rate  $< \epsilon$ .
  - realizable case,  $d^{3/2} \log 1/\epsilon$  labels.
    - supervised --  $d/\epsilon$  labels.

# Margin Based Active-Learning Algorithm

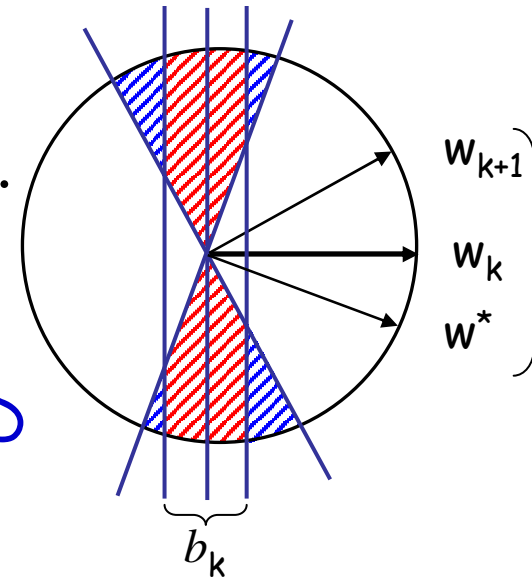
[Balcan-Broder-Zhang, COLT 07]

Use  $O(d)$  examples to find  $w_1$  of error  $1/8$ .

iterate  $k=2, \dots, \log(1/\epsilon)$

- rejection sample  $m_k$  samples  $x$  from  $D$  satisfying  $|w_{k-1}^T \cdot x| \leq b_k$ ;
- label them;
- find  $w_k \in B(w_{k-1}, 1/2^k)$  consistent with all these examples.

end iterate



# BBZ'07, Proof Idea

iterate  $k=2, \dots, \log(1/\epsilon)$

Rejection sample  $m_k$  samples  $x$  from  $D$

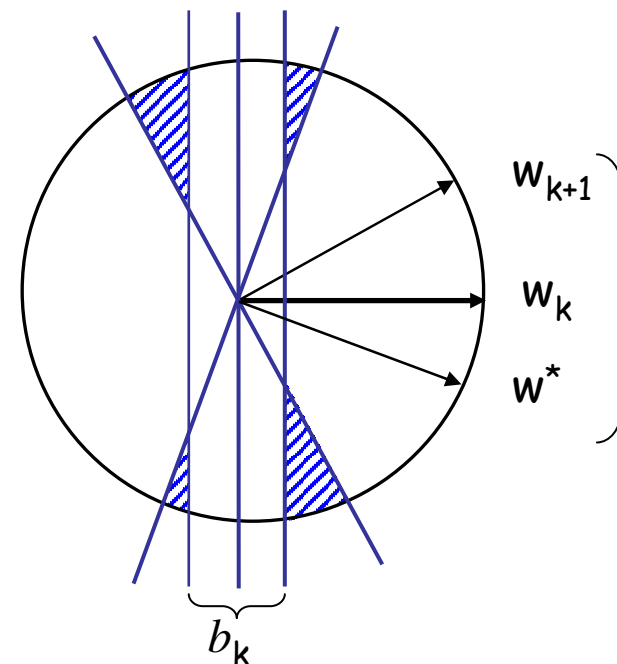
satisfying  $|w_{k-1}^\top \cdot x| \leq b_k$ ;

ask for labels and find  $w_k \in B(w_{k-1}, 1/2^k)$

consistent with all these examples.

end iterate

Assume  $w_k$  has error  $\leq \alpha$ . We are done if  $\exists b_k$  s.t.  $w_{k+1}$  has error  $\leq \alpha/2$  and only need  $O(d \log(1/\epsilon))$  labels in round  $k$ .



$$\text{err}(w_{k+1}) = \Pr(w_{k+1} \text{ errs on } x, |w_k \cdot x| \geq b_k)$$

# BBZ'07, Proof Idea

iterate  $k=2, \dots, \log(1/\varepsilon)$

Rejection sample  $m_k$  samples  $x$  from  $D$

satisfying  $|w_{k-1}^\top \cdot x| \leq b_k$ ;

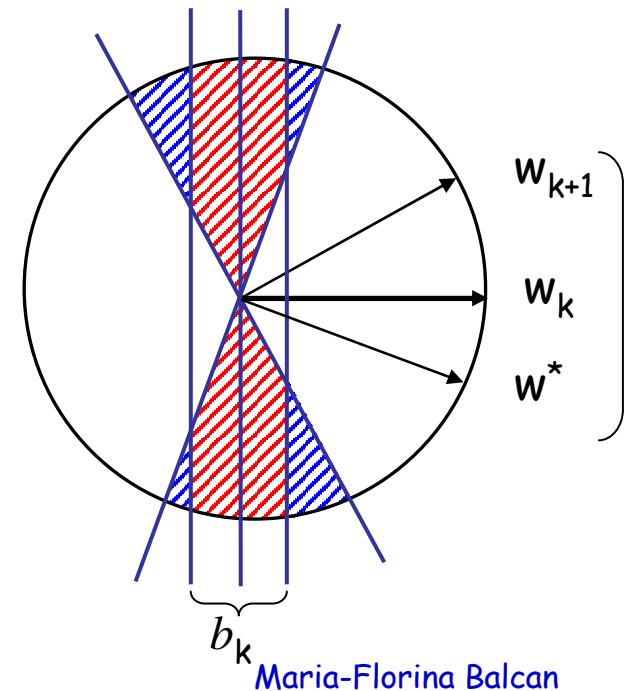
ask for labels and find  $w_k \in B(w_{k-1}, 1/2^k)$

consistent with all these examples.

end iterate

Assume  $w_k$  has error  $\leq \alpha$ . We are done if  $\exists b_k$  s.t.  $w_{k+1}$  has error  $\leq \alpha/2$  and only need  $O(d \log(1/\varepsilon))$  labels in round  $k$ .

$$\text{err}(w_{k+1}) = \Pr(w_{k+1} \text{ errs on } x, |w_k \cdot x| \geq b_k) + \\ \Pr(w_{k+1} \text{ errs on } x, |w_k \cdot x| \leq b_k)$$



# BBZ'07, Proof Idea

iterate  $k=2, \dots, \log(1/\epsilon)$

Rejection sample  $m_k$  samples  $x$  from  $D$

satisfying  $|w_{k-1}^\top \cdot x| \leq b_k$ ;

ask for labels and find  $w_k \in B(w_{k-1}, 1/2^k)$

consistent with all these examples.

end iterate

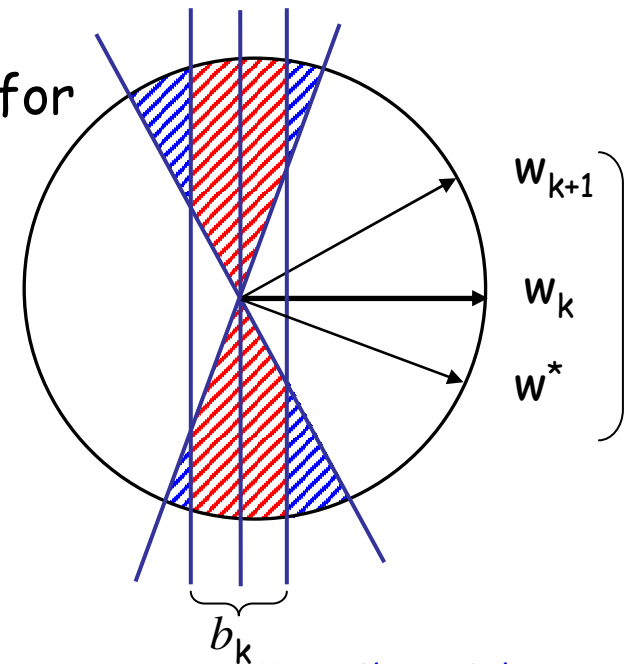
Assume  $w_k$  has error  $\leq \alpha$ . We are done if  $\exists b_k$  s.t.  $w_{k+1}$  has error  $\leq \alpha/2$  and only need  $O(d \log(1/\epsilon))$  labels in round  $k$ .

**Key Point**

Under the uniform distr. assumption for

$$b_k = O\left(\frac{\alpha \log \frac{1}{\alpha}}{\sqrt{d}}\right) \text{ we have } \leq \alpha/4$$

$$\text{err}(w_{k+1}) = \underbrace{\Pr(w_{k+1} \text{ errs on } x, |w_k \cdot x| \geq b_k)} + \Pr(w_{k+1} \text{ errs on } x, |w_k \cdot x| \leq b_k)$$



Maria-Florina Balcan

# BBZ'07, Proof Idea

**Key Point**

Under the uniform distr. assumption for

$$b_k = O\left(\frac{\alpha \log \frac{1}{\alpha}}{\sqrt{d}}\right) \text{ we have}$$

$$\leq \alpha/4$$

$$\text{err}(w_{k+1}) = \underbrace{\Pr(w_{k+1} \text{ errs on } x, |w_k \cdot x| \geq b_k)} +$$

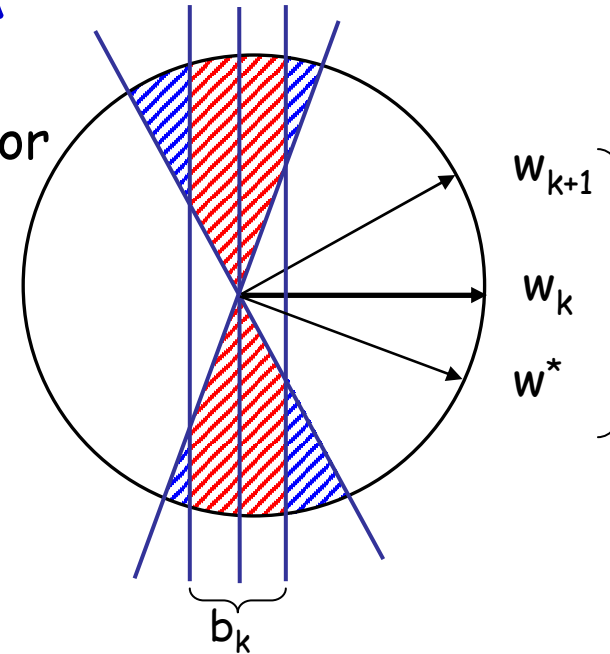
$$\Pr(w_{k+1} \text{ errs on } x \mid |w_k \cdot x| \leq b_k) \Pr(|w_k \cdot x| \leq b_k)$$

**Key Point**

$$\Pr(|w_k \cdot x| \leq b_k) \leq C\alpha \log \frac{1}{\alpha}$$

So, it's enough to ensure that

$$\Pr(w_{k+1} \text{ errs on } x \mid |w_k \cdot x| \leq b_k) \leq \frac{1}{2C \log \frac{1}{\alpha}}$$



We can do so by only using  $O(d \log(1/\epsilon))$  labels in round  $k$ .

## BBZ'07 : Extensions

- A robust version - add a testing step.
- Deals with certain types of noise (e.g. bounded noise), a more general class of distributions.