

# Active Learning

Maria-Florina Balcan

04/19/2019

# Admin

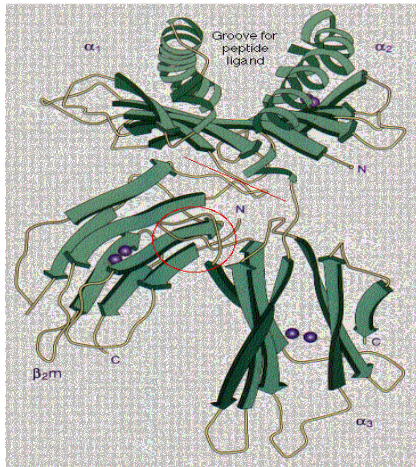
**HWK 6: due on Monday 4/22**

**Final: in class on May 3<sup>rd</sup>.**

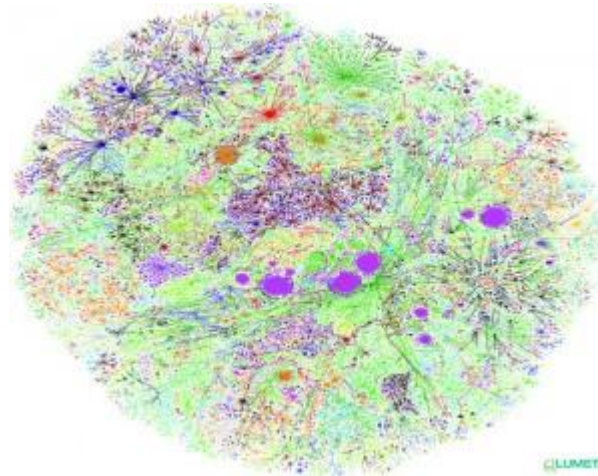
# Classic Fully Supervised Learning Paradigm Insufficient Nowadays

Modern applications: **massive amounts** of raw data.

Only **a tiny fraction** can be annotated by human experts.



Protein sequences



Billions of webpages

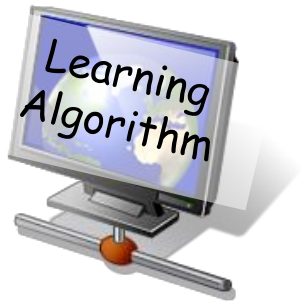


Images

# Modern ML: New Learning Approaches

Modern applications: **massive amounts** of raw data.

Active learning: techniques that best utilize data,  
**minimizing need for expert/human intervention.**

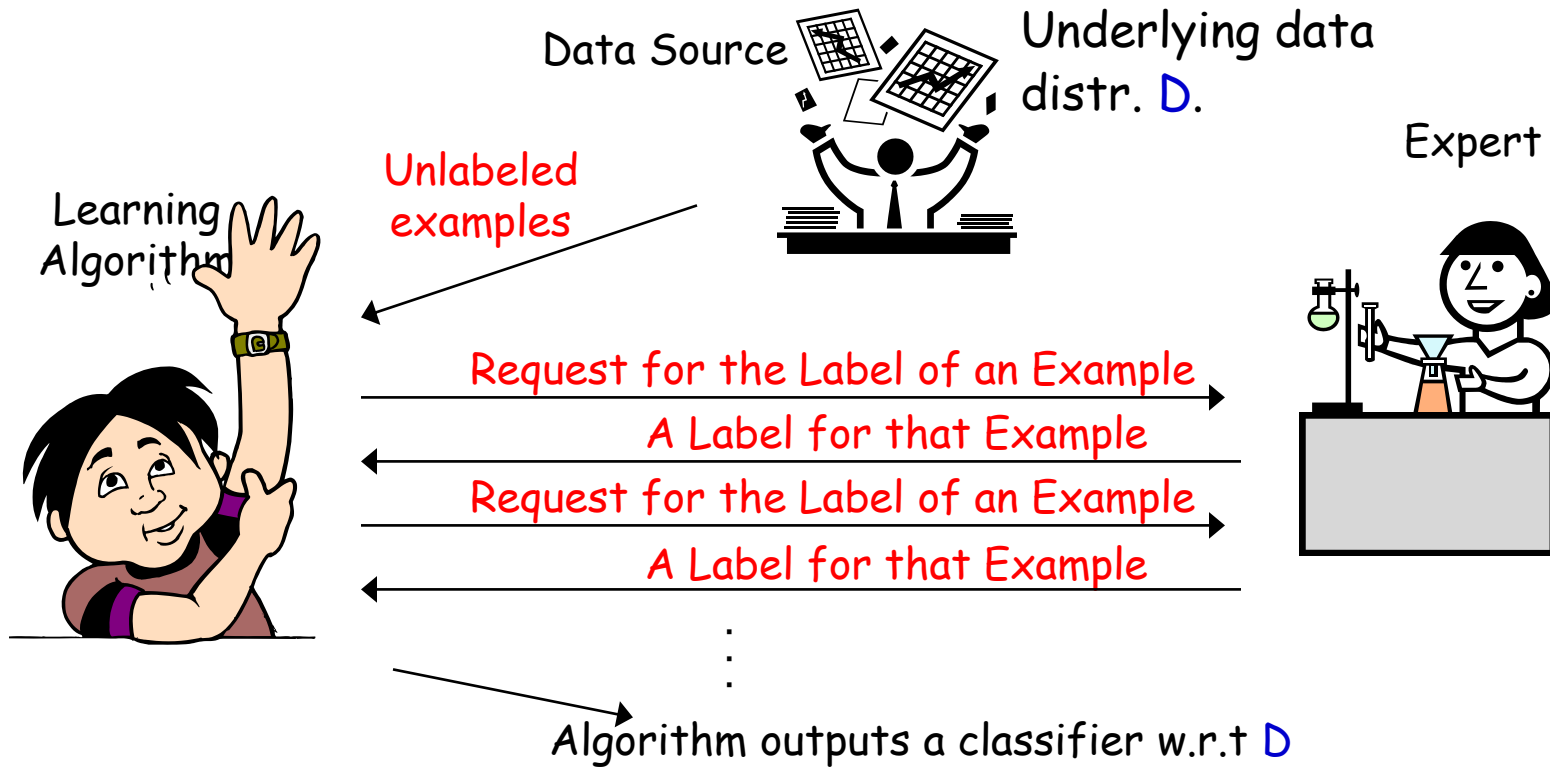


# Active Learning

## Additional resources:

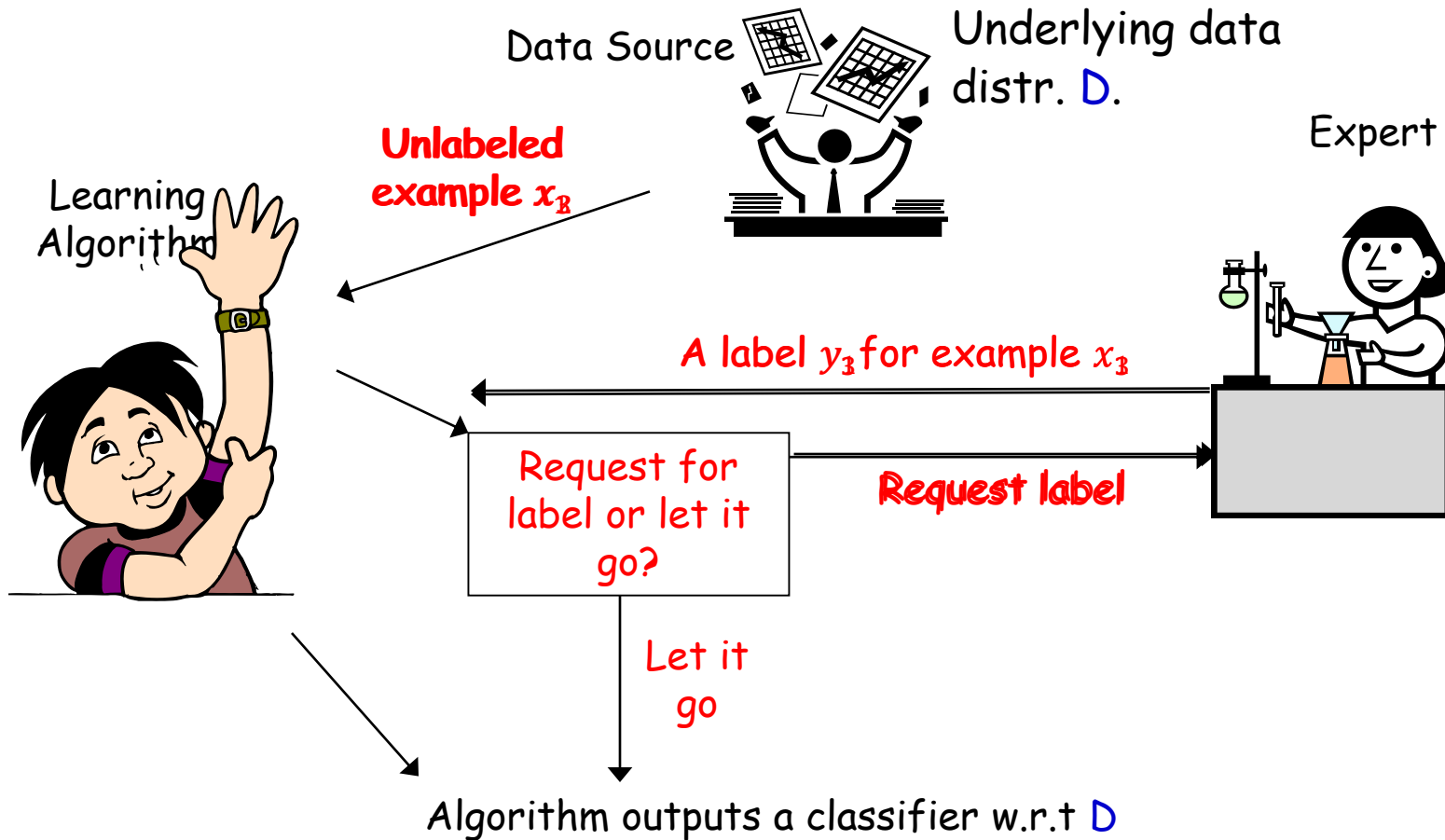
- Two faces of active learning. Sanjoy Dasgupta. 2011.
- Active Learning. Bur Settles. 2012.
- Active Learning. Balcan-Urner. Encyclopedia of Algorithms. 2015

# Batch Active Learning



- Learner can choose specific examples to be labeled.
- Goal: use fewer labeled examples [pick **informative** examples to be labeled].

# Selective Sampling Active Learning



- **Selective sampling AL (Online AL)**: stream of unlabeled examples, when each arrives make a decision to ask for label or not.
- **Goal**: use fewer labeled examples [pick **informative** examples to be labeled].

# What Makes a Good Active Learning Algorithm?

- Guaranteed to output a relatively good classifier for most learning problems.
- Doesn't make too many label requests.  
Hopefully a lot less than passive learning and SSL.
- Need to choose the label requests carefully, to get **informative** labels.

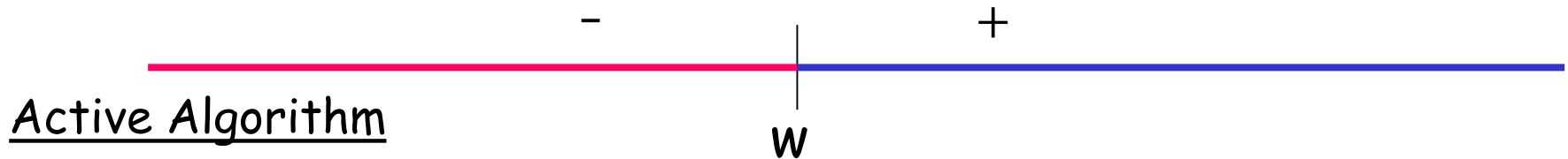


# Can adaptive querying really do better than passive/random sampling?

- YES! (sometimes)
- We often need far fewer labels for active learning than for passive.
- This is predicted by theory and has been observed in practice.

# Can adaptive querying help? [CAL92, Dasgupta04]

- Threshold fns on the real line:  $h_w(x) = 1(x \geq w)$ ,  $C = \{h_w: w \in \mathbb{R}\}$



## Active Algorithm

- Get  $N$  unlabeled examples
- How can we recover the correct labels with  $\ll N$  queries?
- Do binary search! Just need  $O(\log N)$  labels!



- Output a classifier consistent with the  $N$  inferred labels.

- $N = O(1/\epsilon)$  we are guaranteed to get a classifier of error  $\leq \epsilon$ .

Passive supervised:  $\Omega(1/\epsilon)$  labels to find an  $\epsilon$ -accurate threshold.

Active: only  $O(\log 1/\epsilon)$  labels. Exponential improvement.



# Common Technique in Practice

Uncertainty sampling in SVMs common and quite useful in practice. E.g., [Tong & Koller, ICML 2000; Jain, Vijayanarasimhan & Grauman, NIPS 2010; Schohn Cohn, ICML 2000]

## Active SVM Algorithm

- At any time during the alg., we have a “current guess”  $w_t$  of the separator: the max-margin separator of all labeled points so far.
- Request the label of the example closest to the current separator.

# Common Technique in Practice

Active SVM seems to be quite useful in practice.

[Tong & Koller, ICML 2000; Jain, Vijayanarasimhan & Grauman, NIPS 2010]

## Algorithm (batch version)

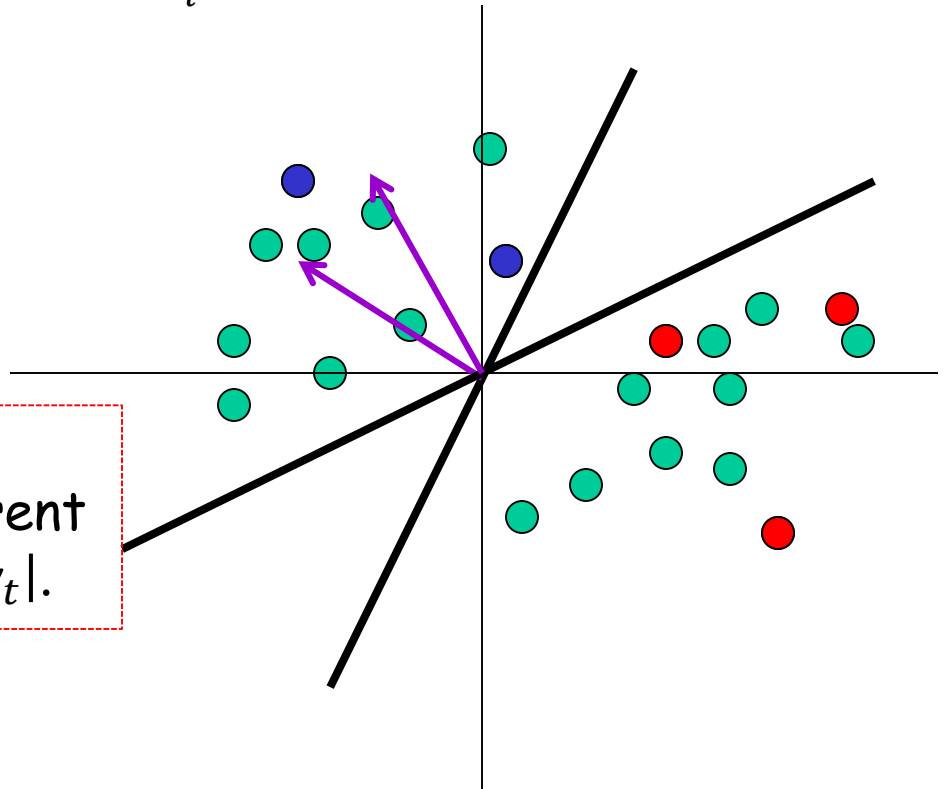
Input  $S_u = \{x_1, \dots, x_{m_u}\}$  drawn i.i.d from the underlying source  $D$

Start: query for the labels of a few random  $x_i$ s.

For  $t = 1, \dots,$

- Find  $w_t$  the max-margin separator of all labeled points so far.
- Request the label of the example closest to the current separator: minimizing  $|x_i \cdot w_t|$ .

(highest uncertainty)

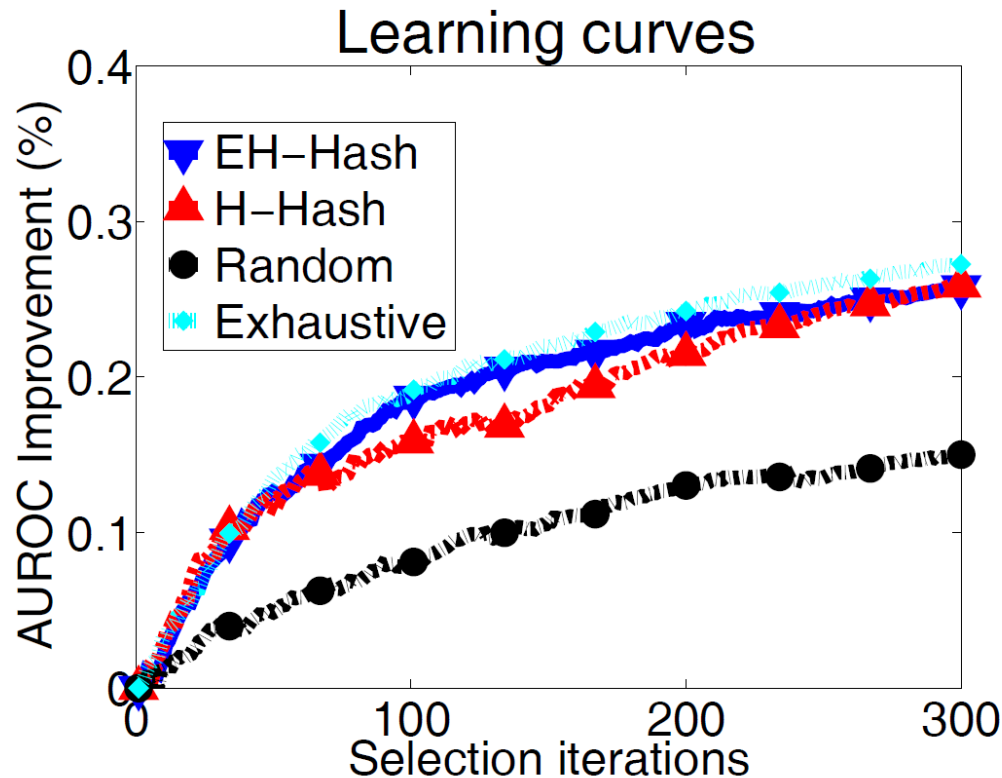


# Common Technique in Practice

Active SVM seems to be quite useful in practice.

E.g., Jain, Vijayanarasimhan & Grauman, NIPS 2010

Newsgroups dataset (20.000 documents from 20 categories)

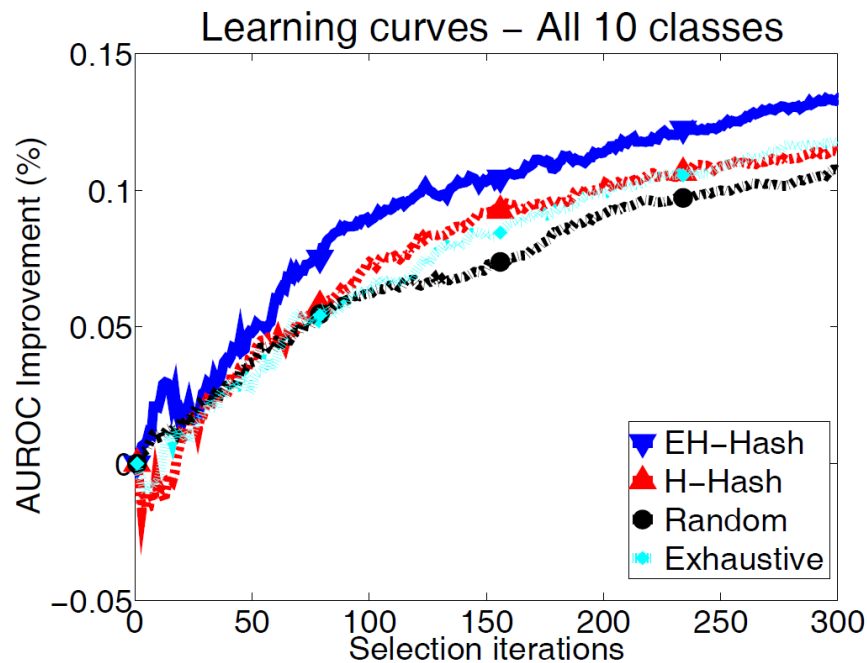


# Common Technique in Practice

Active SVM seems to be quite useful in practice.

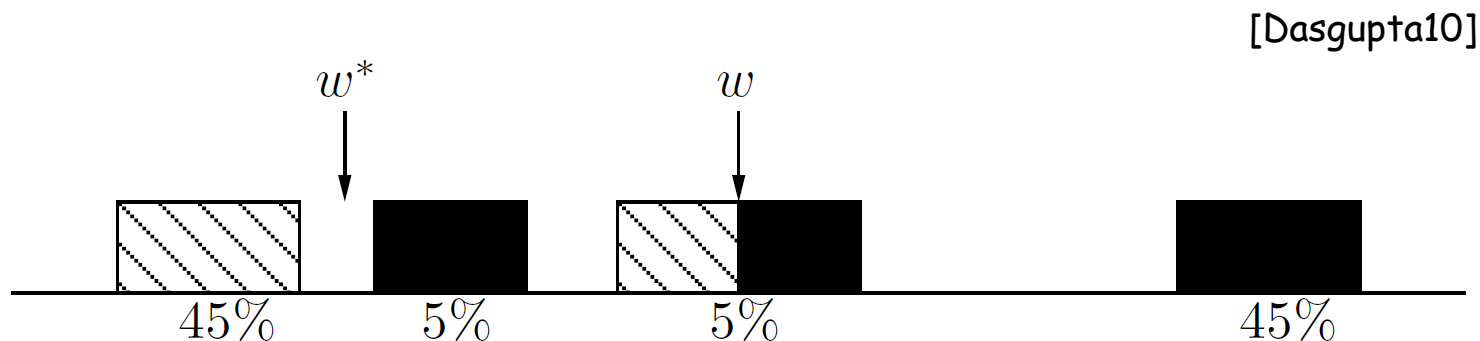
E.g., Jain, Vijayanarasimhan & Grauman, NIPS 2010

CIFAR-10 image dataset (60.000 images from 10 categories)



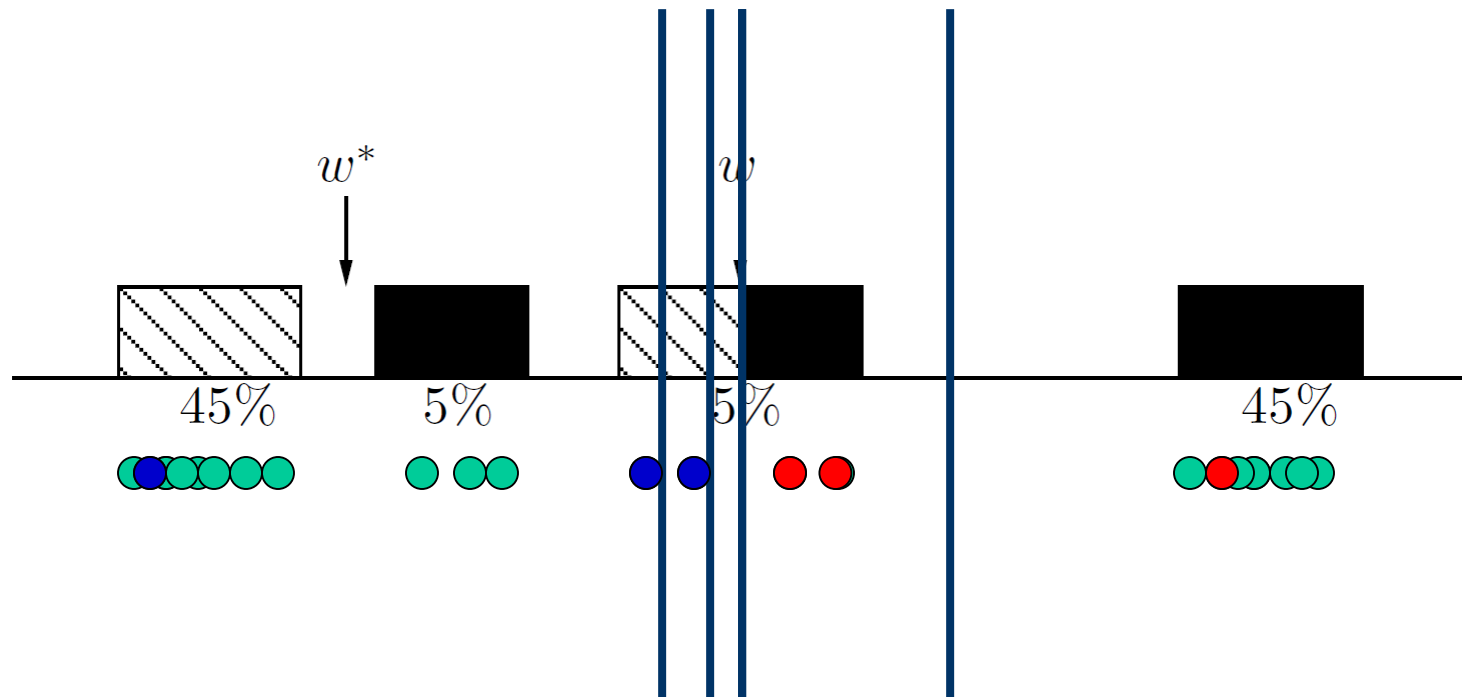
# Active SVM/Uncertainty Sampling

- Works sometimes....
- However, we need to be very very very careful!!!
  - Myopic, greedy technique can suffer from **sampling bias**.
  - A bias created because of the querying strategy; as time goes on the sample is less and less representative of the true data source.



# Active SVM/Uncertainty Sampling

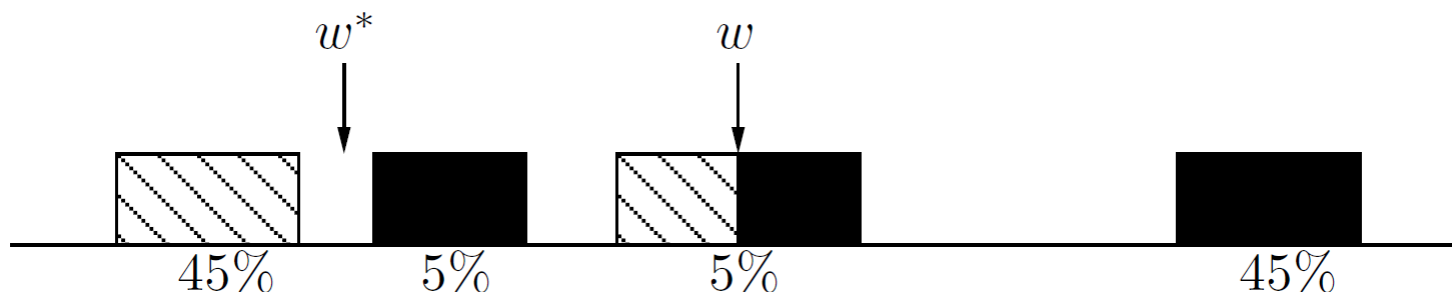
- Works sometimes....
- However, we need to be very very careful!!!





# Active SVM/Uncertainty Sampling

- Works sometimes....
- **However, we need to be very very careful!!!**
  - Myopic, greedy technique can suffer from **sampling bias**.
  - Bias created because of the querying strategy; as time goes on the sample is less and less representative of the true source.
  - Observed in practice too!!!!
- **Main tension:** want to choose informative points, but also want to guarantee that the classifier we output does well on true random examples from the underlying distribution.



# Safe Active Learning Schemes

Disagreement Based Active Learning

Hypothesis Space Search

[CAL92] [BBL06]

[Hanneke'07, DHM'07, Wang'09, Fridman'09, Kolt10, BHW'08, BHLZ'10, H'10, Ailon'12, ...]

# Version Spaces

- $X$  - feature/instance space; distr.  $D$  over  $X$ ;  $c^*$  target fnc
- Fix hypothesis space  $H$ .

**Definition (Mitchell'82)** Assume realizable case:  $c^* \in H$ .

Given a set of labeled examples  $(x_1, y_1), \dots, (x_{m_1}, y_{m_1}), y_i = c^*(x_i)$

**Version space of  $H$ :** part of  $H$  consistent with labels so far.

I.e.,  $h \in VS(H)$  iff  $h(x_i) = c^*(x_i) \forall i \in \{1, \dots, m_1\}$ .

# Version Spaces

- $X$  - feature/instance space; distr.  $D$  over  $X$ ;  $c^*$  target fnc
- Fix hypothesis space  $H$ .

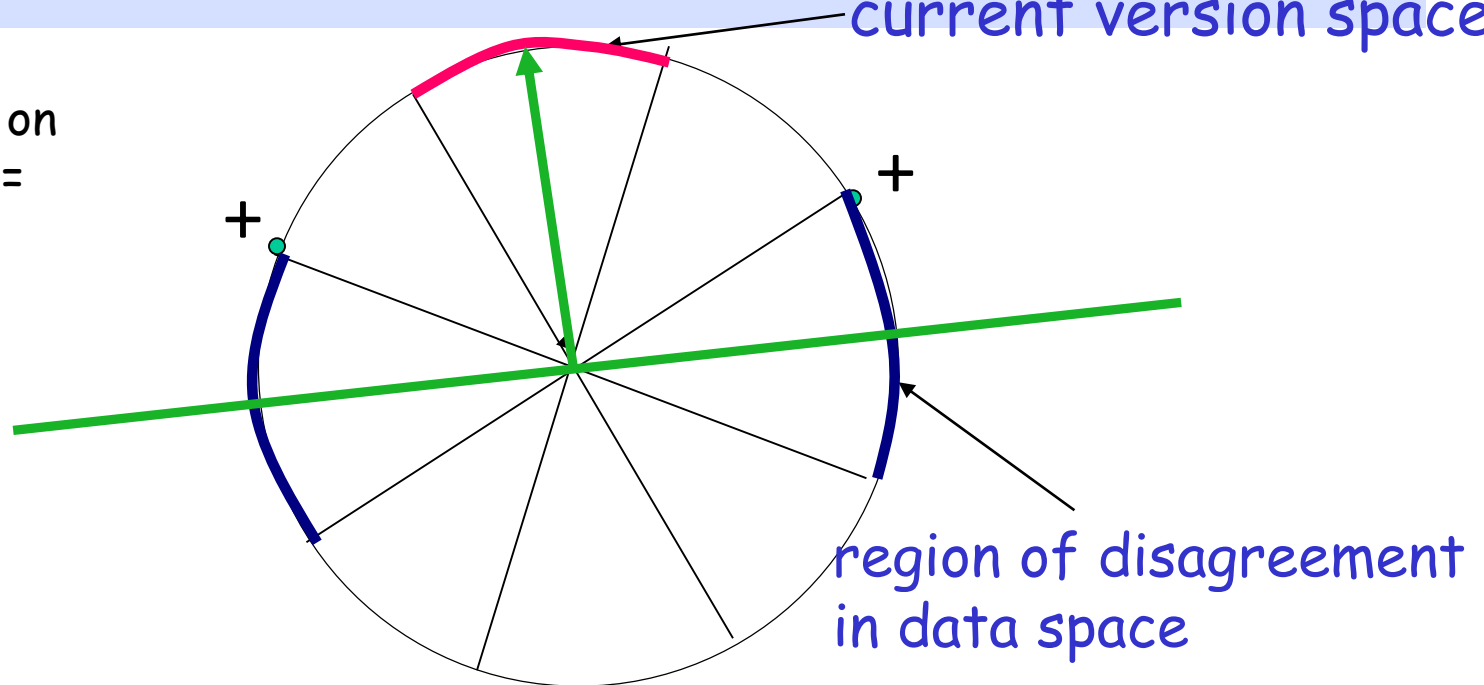
**Definition (Mitchell'82)** Assume realizable case:  $c^* \in H$ .

Given a set of labeled examples  $(x_1, y_1), \dots, (x_{m_1}, y_{m_1}), y_i = c^*(x_i)$

**Version space of  $H$ :** part of  $H$  consistent with labels so far.

current version space

E.g.: data lies on circle in  $\mathbb{R}^2$ ,  $H$  = homogeneous linear seps.



# Version Spaces. Region of Disagreement

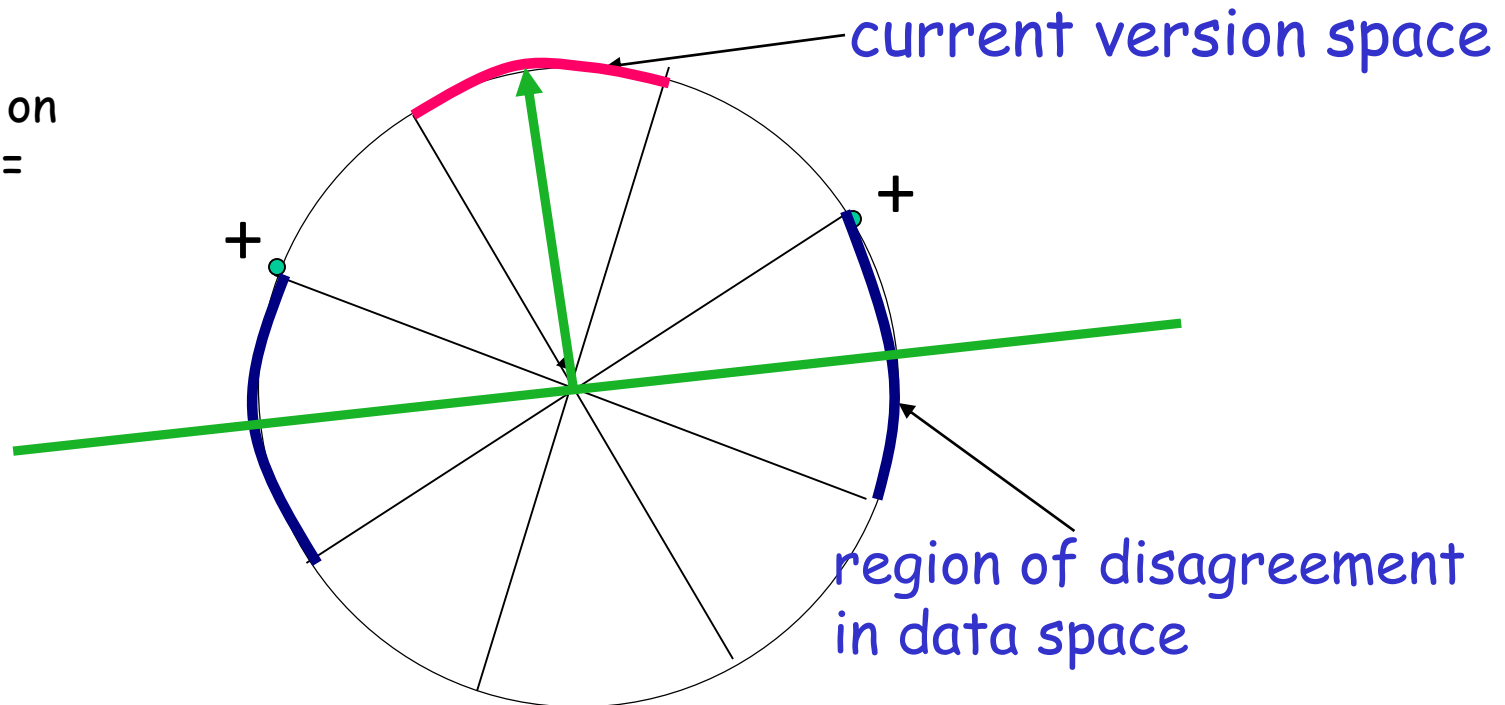
## Definition (CAL'92)

**Version space:** part of  $H$  consistent with labels so far.

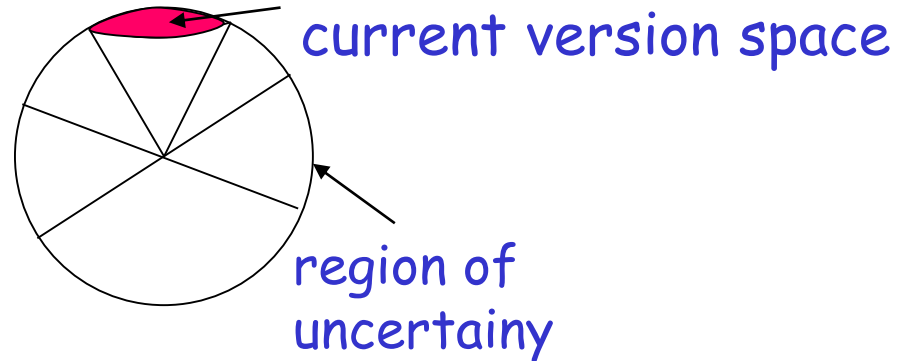
**Region of disagreement** = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

$x \in X, x \in \text{DIS}(\text{VS}(H))$  iff  $\exists h_1, h_2 \in \text{VS}(H), h_1(x) \neq h_2(x)$

E.g.: data lies on circle in  $\mathbb{R}^2$ ,  $H$  = homogeneous linear sep.



# Disagreement Based Active Learning [CAL92]



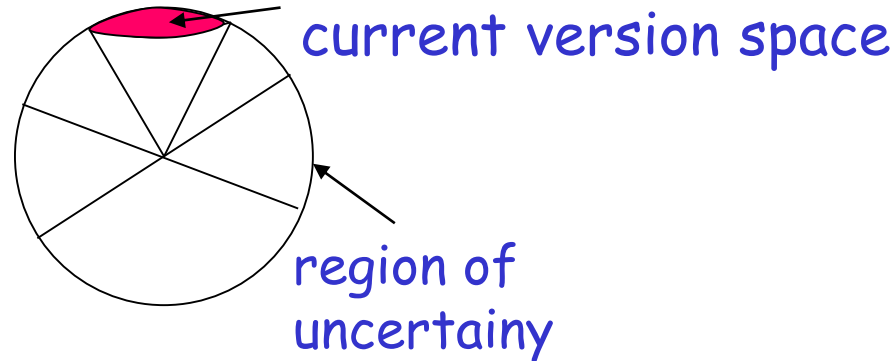
## Algorithm:

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

Stop when region of uncertainty is small.

**Note:** it is active since we do not waste labels by querying in regions of space we are certain about the labels.

# Disagreement Based Active Learning [CAL92]



## Algorithm:

Query for the labels of a few random  $x_i$ s.

Let  $H_1$  be the current version space.

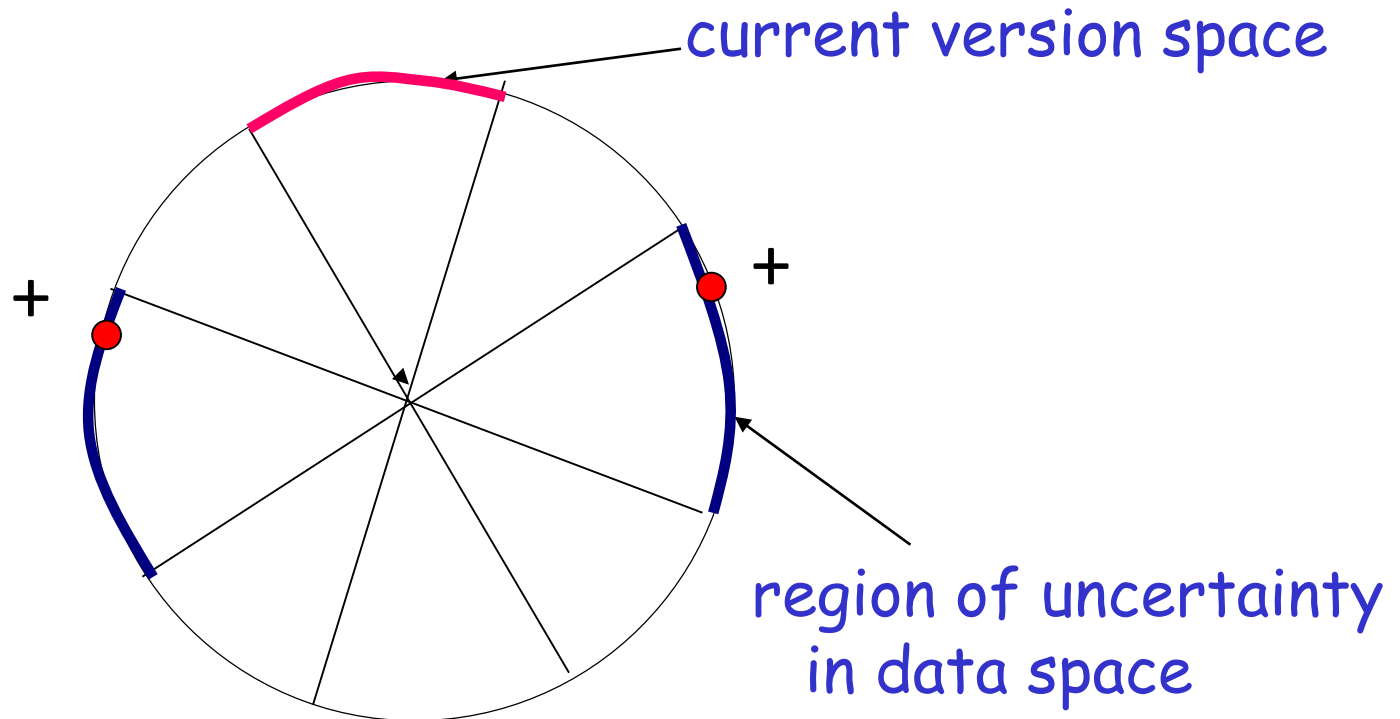
For  $t = 1, \dots,$

Pick a few points at random from the current region of disagreement  $\text{DIS}(H_t)$  and query their labels.

Let  $H_{t+1}$  be the new version space.

# Region of uncertainty [CAL92]

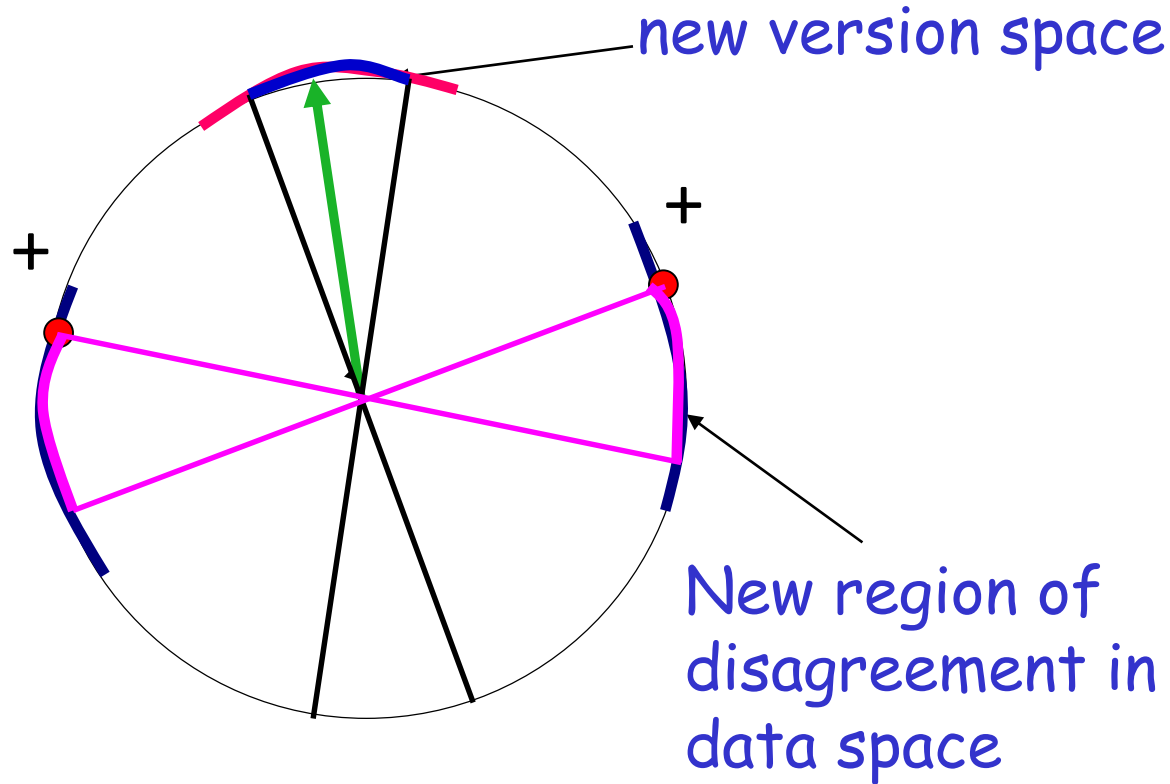
- Current **version space**: part of  $C$  consistent with labels so far.
- "**Region of uncertainty**" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)





## Region of uncertainty [CAL92]

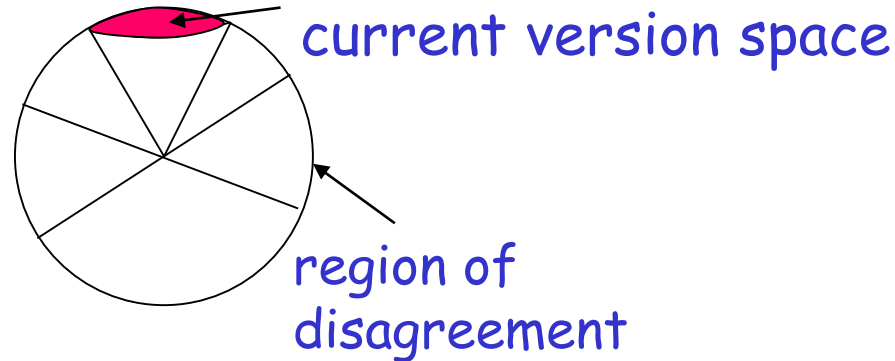
- Current **version space**: part of  $C$  consistent with labels so far.
- "**Region of uncertainty**" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)





How about the agnostic case  
where the target might not  
belong the  $H$ ?

# $A^2$ Agnostic Active Learner [BBL'06]



## Algorithm:

Let  $H_1 = H$ .

For  $t = 1, \dots,$

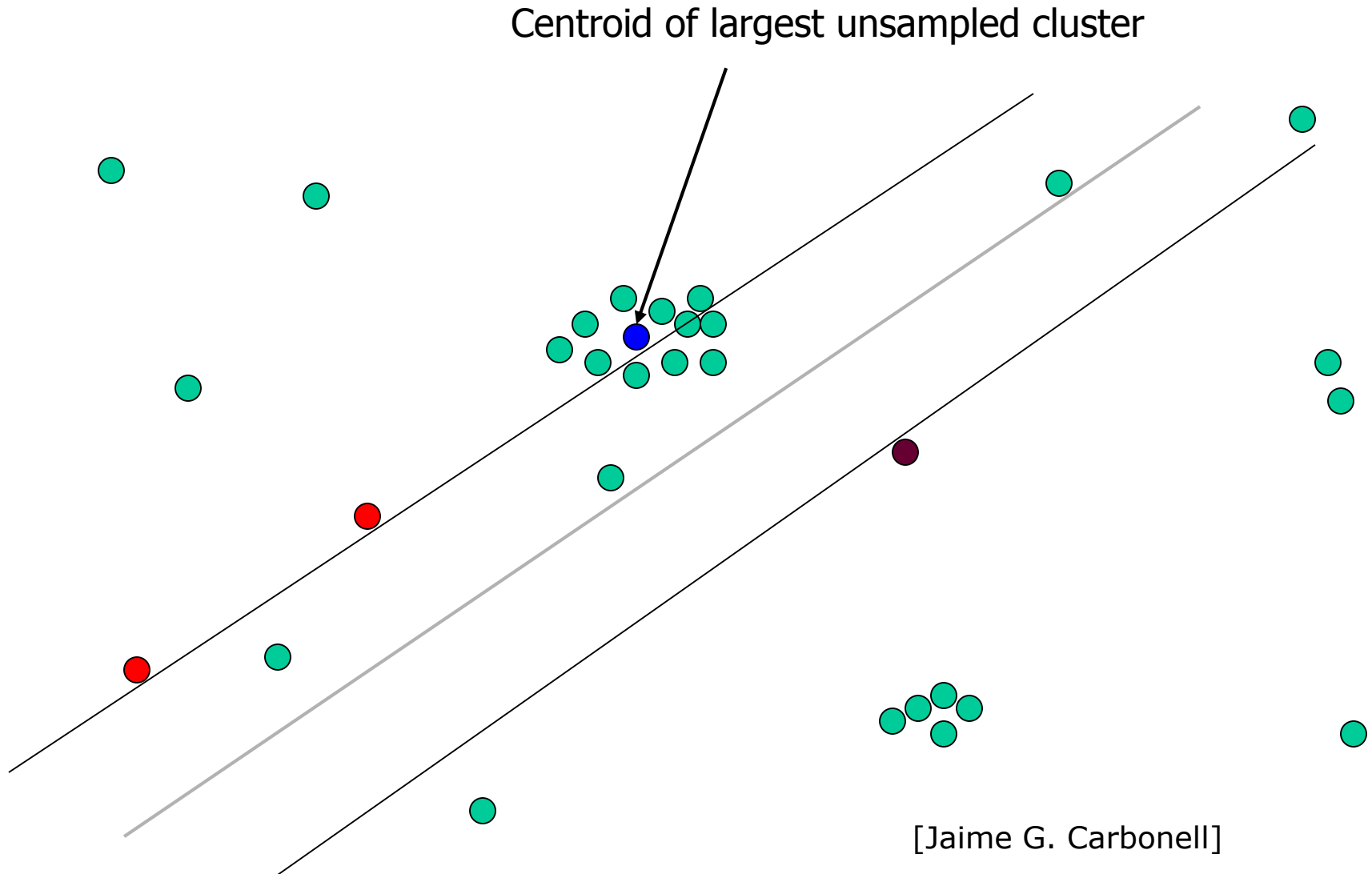
- Pick a few points at random from the current region of disagreement  $\text{DIS}(H_t)$  and query their labels.
- Throw out hypothesis if you are statistically confident they are suboptimal.

Careful use of generalization bounds;  
Avoid the sampling bias!!!!

## Other Interesting AL Techniques used in Practice

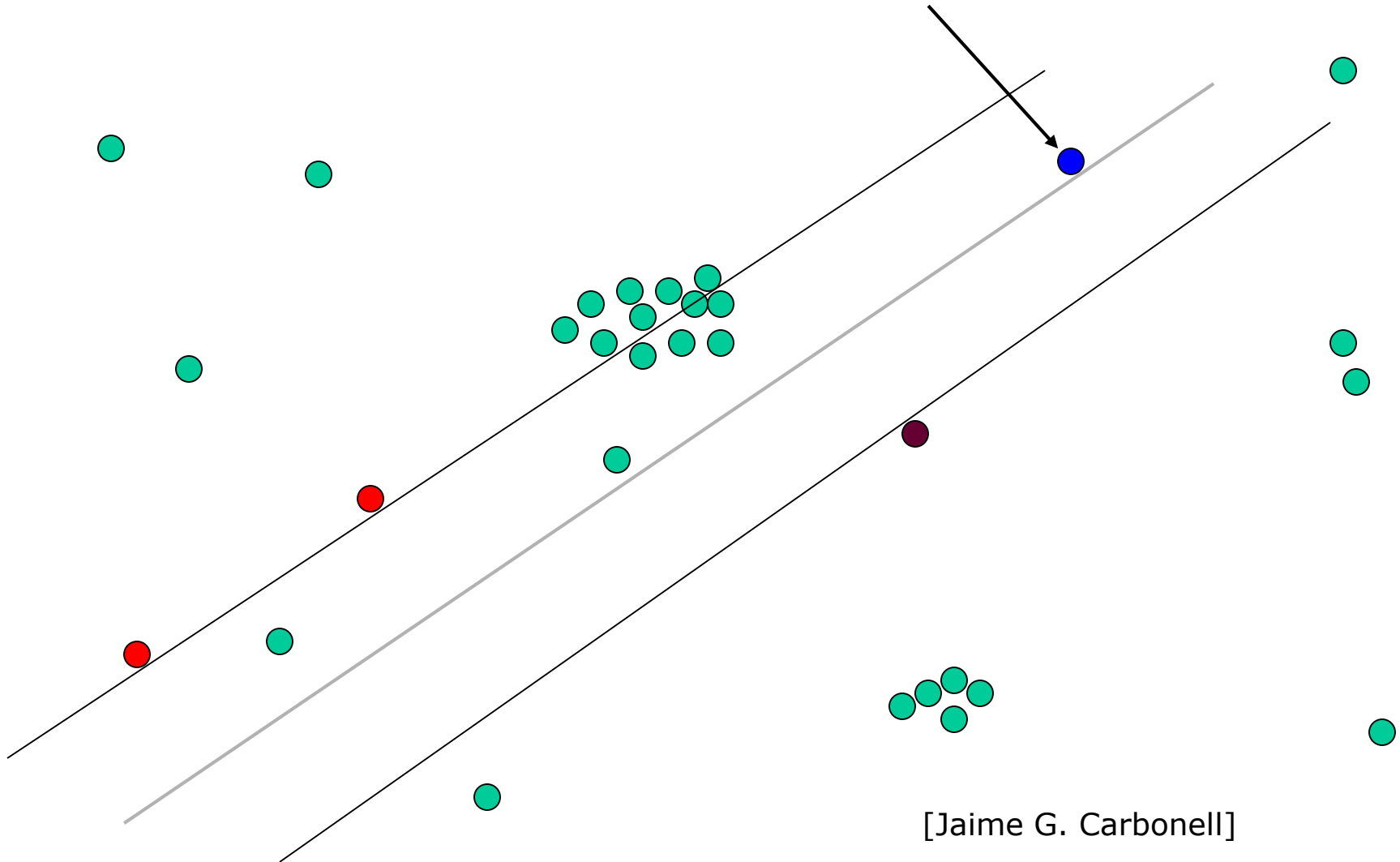
Interesting open question to analyze  
under what conditions they are successful.

# Density-Based Sampling



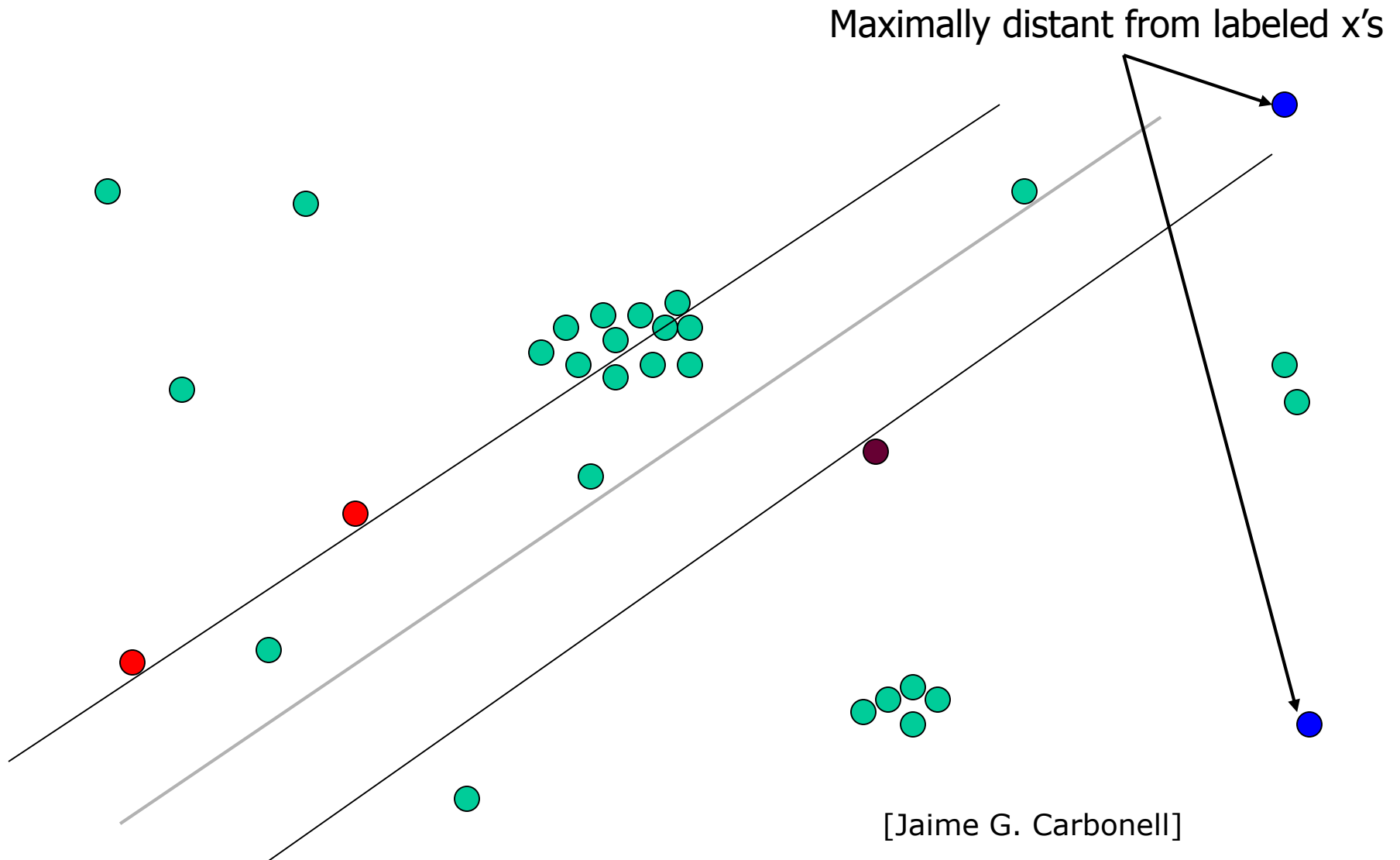
# Uncertainty Sampling

Closest to decision boundary (Active SVM)

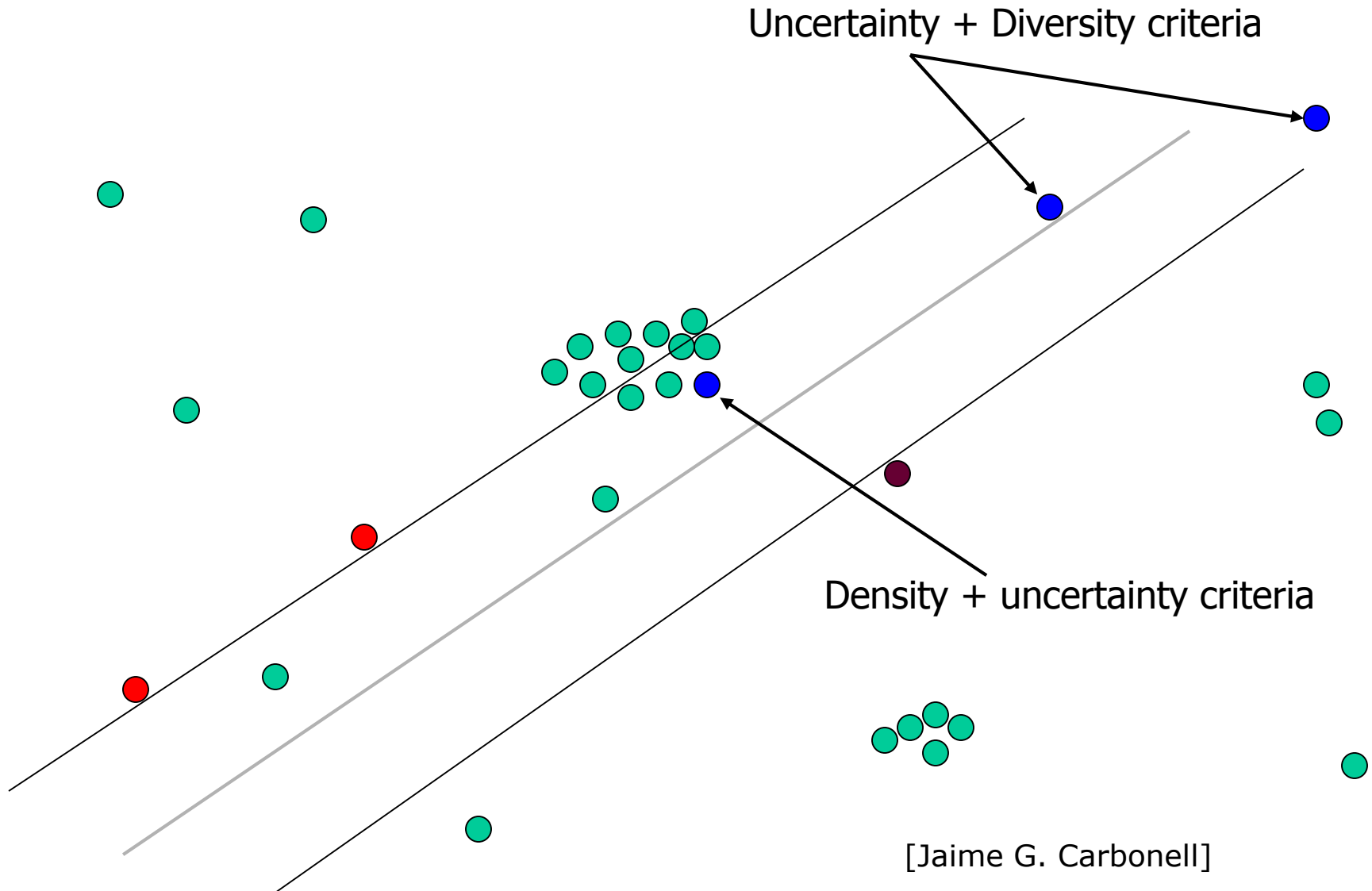


[Jaime G. Carbonell]

# Maximal Diversity Sampling



# Ensemble-Based Possibilities



[Jaime G. Carbonell]



# What You Should Know

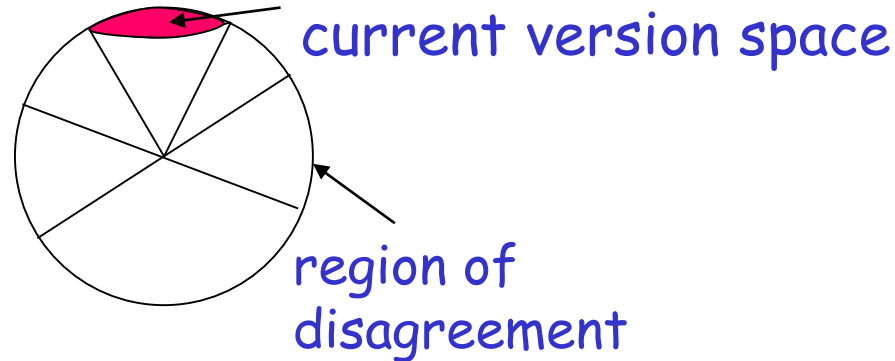
- Active learning could be really helpful, could provide exponential improvements in label complexity (both theoretically and practically)!
- Common heuristics (e.g., those based on uncertainty sampling). Need to be very careful due to sampling bias.
- Safe Disagreement Based Active Learning Schemes.
  - Understand how they operate precisely in the realizable case (noise free scenarios).



Advanced additional (not  
required material)

Disagreement based algorithms:  
How about the agnostic case  
where the target might not  
belong the  $H$ ?

# $A^2$ Agnostic Active Learner [BBL'06]



## Algorithm:

Let  $H_1 = H$ .

For  $t = 1, \dots,$

- Pick a few points at random from the current region of disagreement  $\text{DIS}(H_t)$  and query their labels.
- Throw out hypothesis if you are statistically confident they are suboptimal.

Careful use of generalization bounds;  
Avoid the sampling bias!!!!

# Formal General Guarantees for Agnostic AL

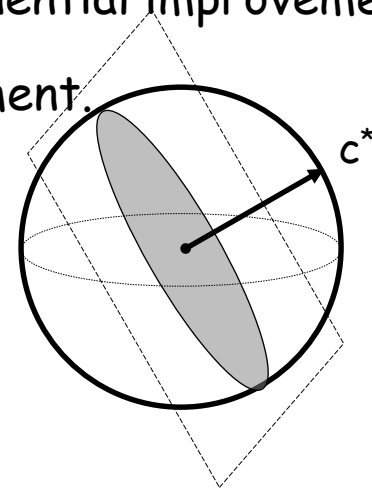
$A^2$  the first algorithm which is robust to noise.

[Balcan, Beygelzimer, Langford, ICML'06] [Balcan, Beygelzimer, Langford, JCSS'08]

"Region of disagreement" style: Pick a few points at random from the current region of disagreement, query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

Guarantees for  $A^2$  [BBL'06,'08]:

- It is **safe** (never worse than passive learning) & exponential improvements.
- $C$  - thresholds, low noise, exponential improvement.
- $C$  - homogeneous linear separators in  $\mathbb{R}^d$ ,  
 $D$  - uniform, low noise, only  $d^2 \log(1/\epsilon)$  labels.



A lot of subsequent work.

[Hanneke'07, DHM'07, Wang'09, Fridman'09, Kolt10, BHW'08, BHLZ'10, H'10, Ailon'12, ...]

# General guarantees for $A^2$ Agnostic Active Learner

"Disagreement based": Pick a few points at random from the current region of uncertainty, query their labels, throw out hypothesis if you are statistically confident they are suboptimal. [BBL'06]

How quickly the region of disagreement collapses as we get closer and closer to optimal classifier

Guarantees for  $A^2$  [Hanneke'07]:

Disagreement coefficient  $\theta_{c^*} = \sup_{r \geq \eta + \epsilon} \frac{\Pr(DIS(B(c^*, r)))}{r}$

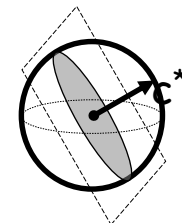
Theorem

$$m = \left(1 + \frac{\eta^2}{\epsilon^2}\right) VCdim(C) \theta_{c^*}^2 \log\left(\frac{1}{\epsilon}\right)$$

labels are sufficient s.t. with prob.  $\geq 1 - \delta$  output  $h$  with  $err(h) \leq \eta + \epsilon$ .

Realizable case:  $m = VCdim(C) \theta_{c^*} \log\left(\frac{1}{\epsilon}\right)$

Linear Separators, uniform distr.:  $\theta_{c^*} = \sqrt{d}$



# Disagreement Based Active Learning

"Disagreement based " algos: query points from current region of disagreement, throw out hypotheses when statistically confident they are suboptimal.

- Generic (any class), adversarial label noise.
- Computationally efficient for classes of small VC-dimension

Still, could be suboptimal in label complex & computationally inefficient in general.

Lots of subsequent work trying to make is more efficient computationally and more aggressive too: [Hanneke07, DasguptaHsuMontleoni'07, Wang'09 , Fridman'09, Koltchinskii10, BHW'08, BeygelzimerHsuLangfordZhang'10, Hsu'10, Ailon'12, ...]