# Active Learning

Maria-Florina Balcan

04/16/2018

# Admin

## Projects and Project Presentations

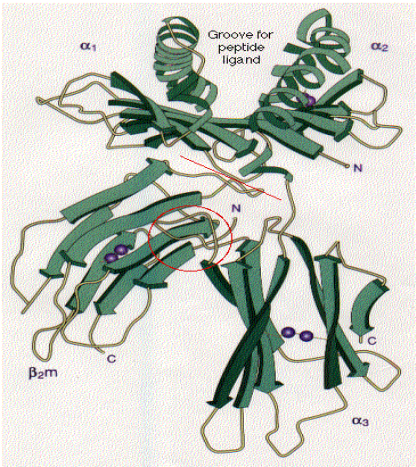
- Project presentations: April 23 and April 25.
- Presentation schedule posted on Piazza last week.
- Should be in contact with TAs to discuss progress.
- Will meet with me on April 18$^{th}$ or April 24$^{th}$.
- Final report: May 14th.

**Final: in class on May 2$^{nd}$.**

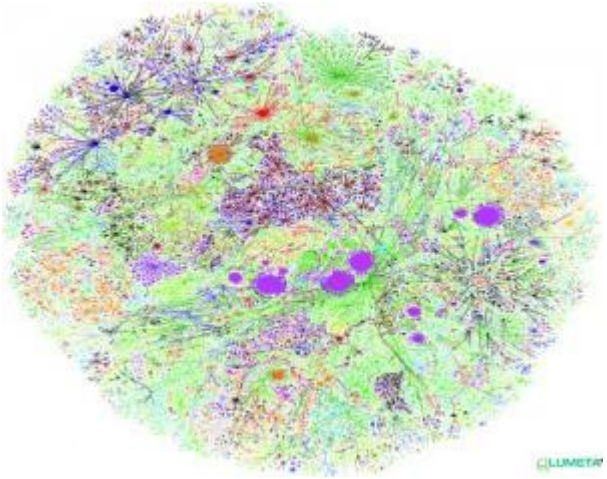
# Classic Fully Supervised Learning Paradigm Insufficient Nowadays

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

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




Protein sequences



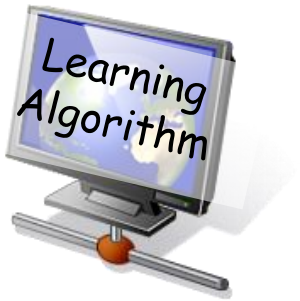
Billions of webpages

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.**

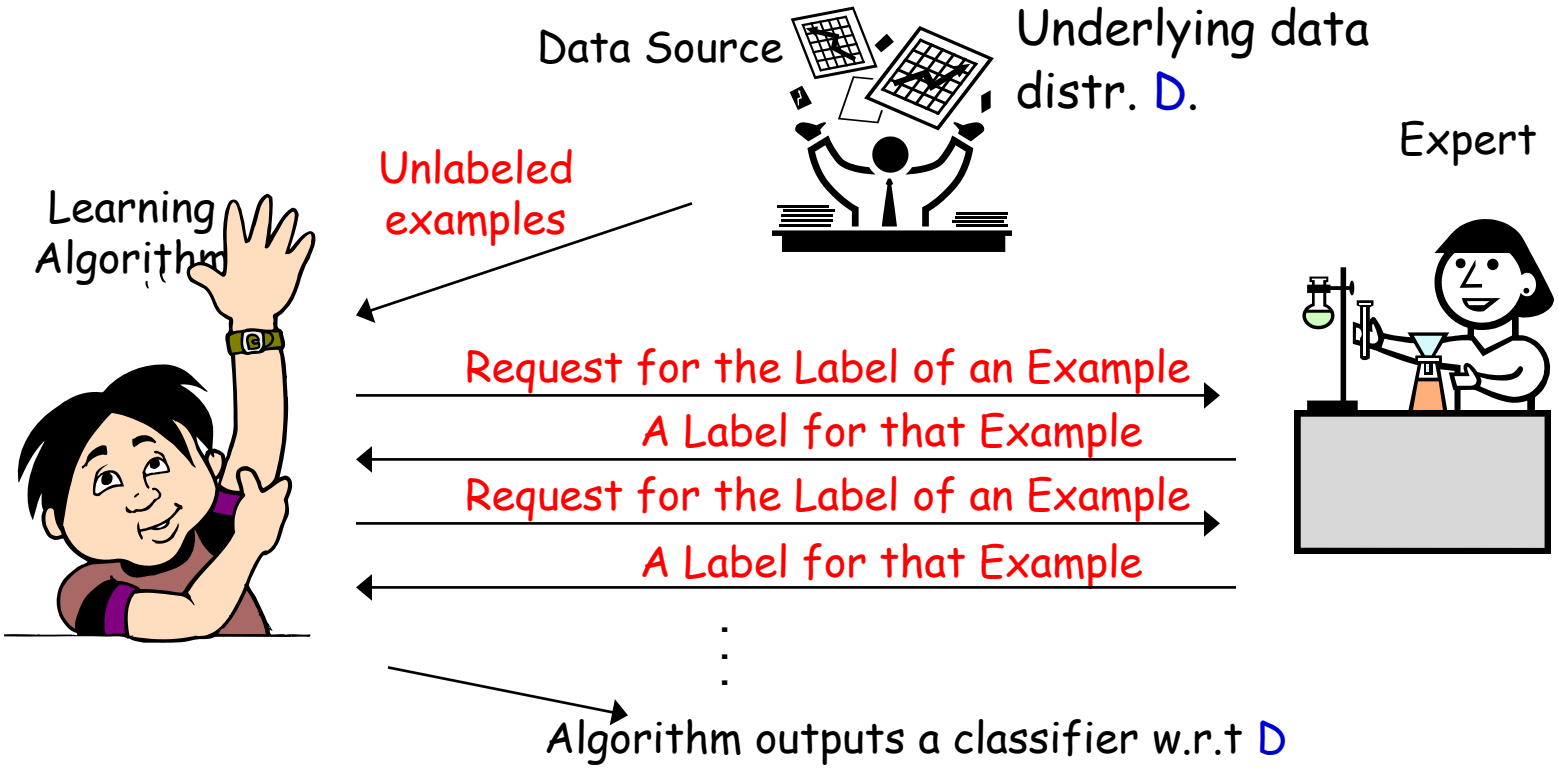



# Active Learning

<u>Additional resources:</u>

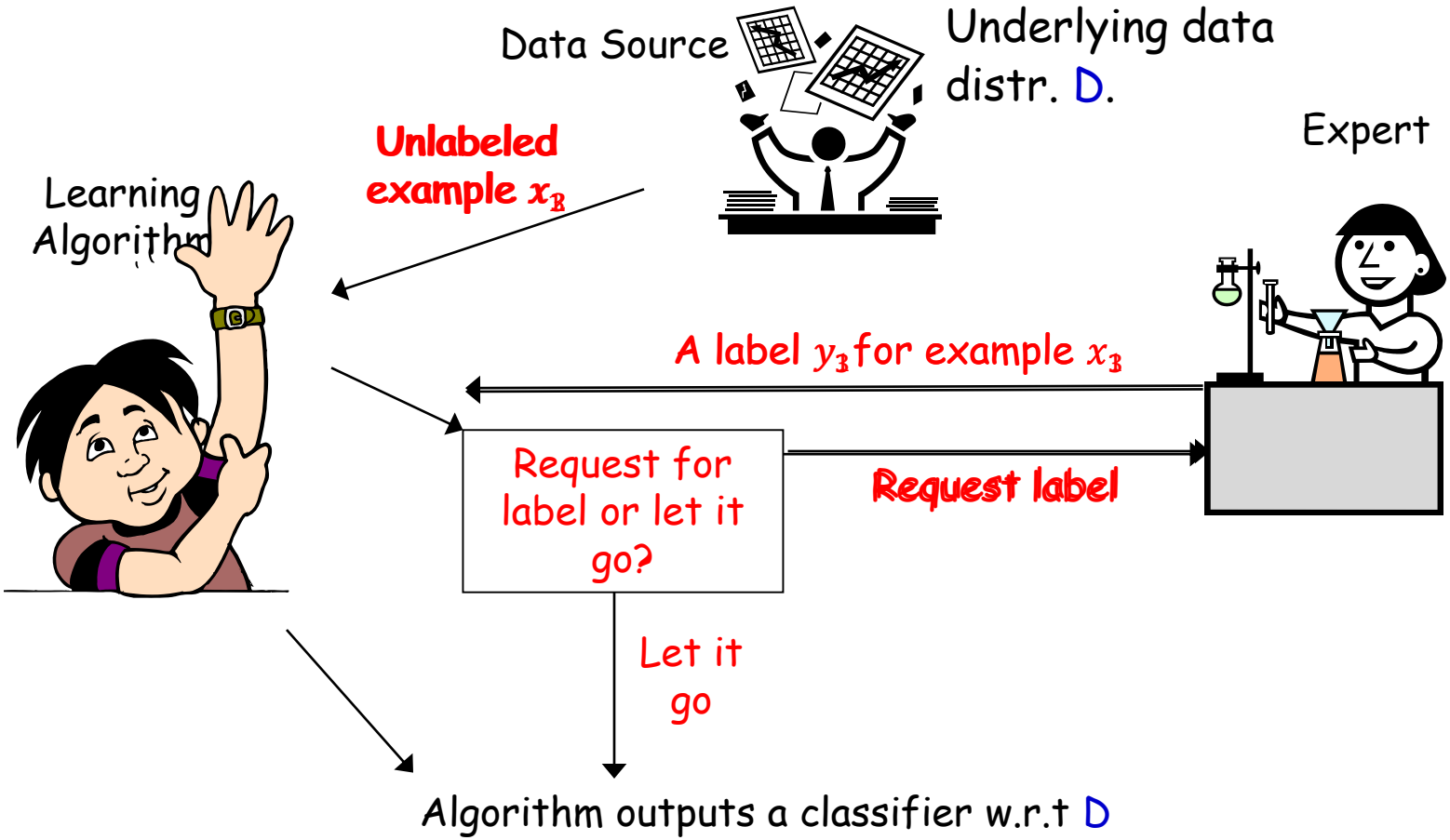
- Two faces of active learning. Sanjoy Dasgupta. 2011.
- Active Learning. Bur Settles. 2012.
- Active Learning. Balcan-Urner. Encyclopedia of Algorithms. 2015

# Batch Active Learning




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

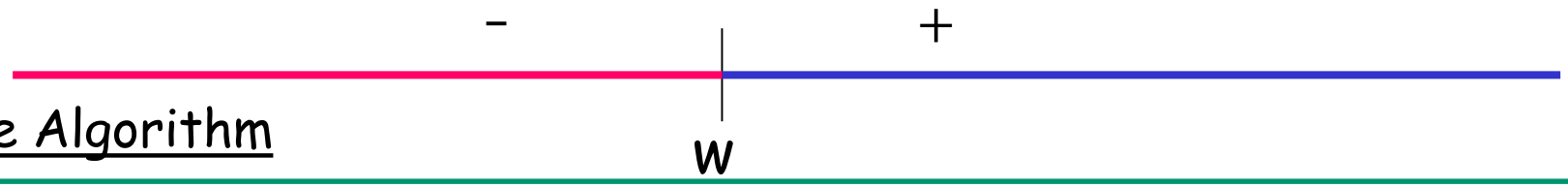
# Selective Sampling Active Learning




- 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].

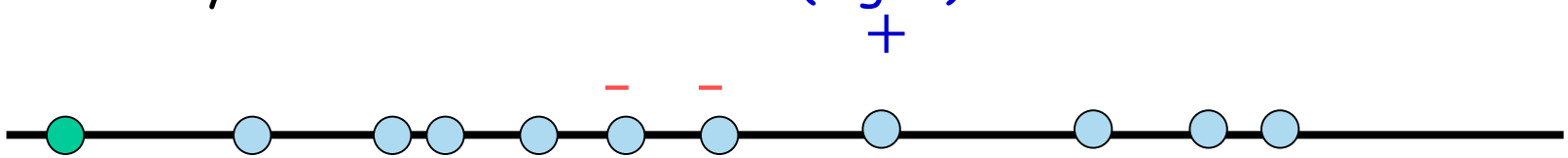
# Can adaptive querying help? [CAL92, Dasgupta04]

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




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 $\varepsilon$-accurate threshold.

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

# 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)

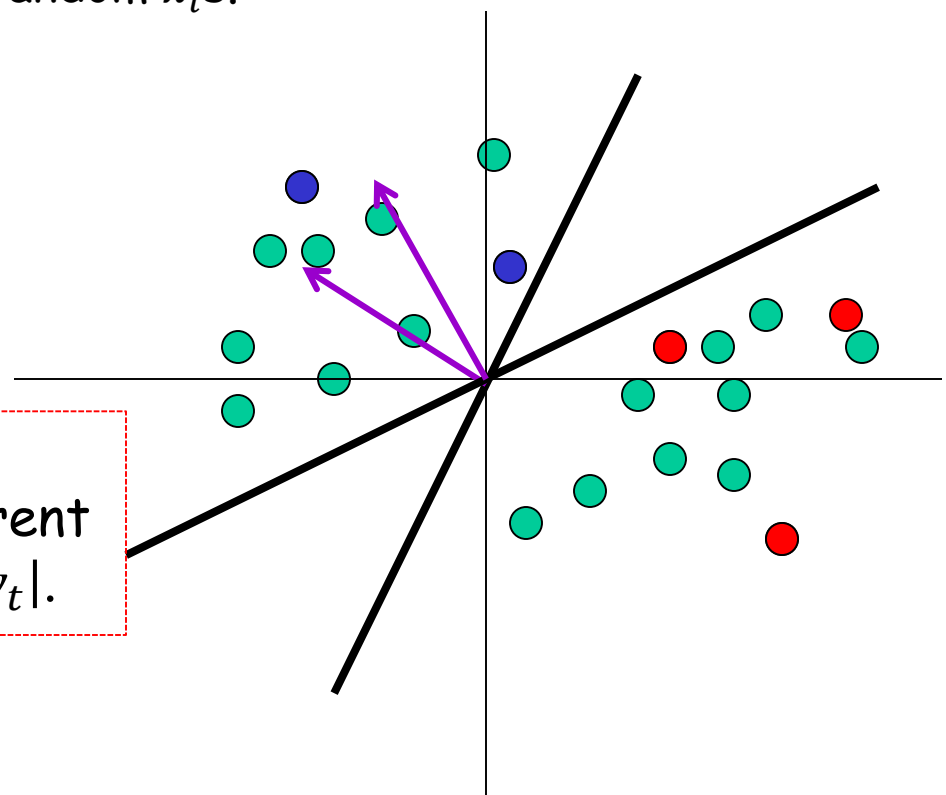
Input $S_u = \{x_1, \ldots, x_{m_u}\}$ drawn i.i.d from the underlying source D

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

**For** $t = 1, \ldots,$

- 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 $|x_i \cdot w_t|$.

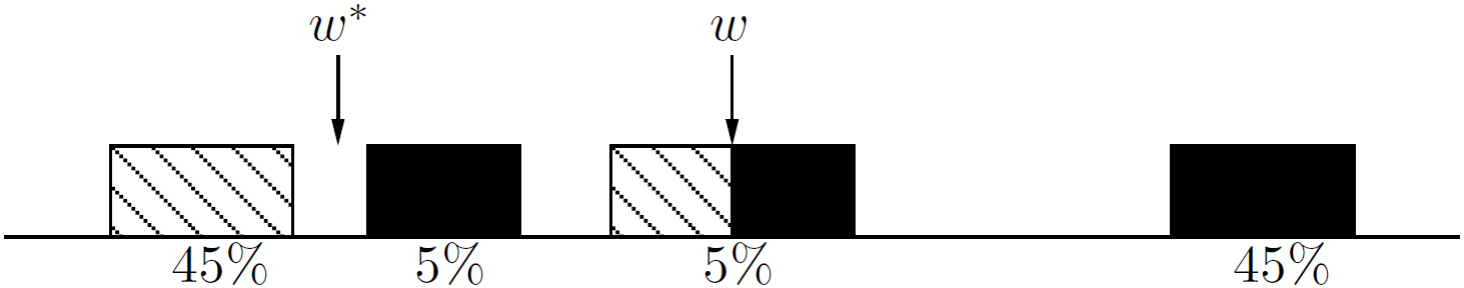
(highest uncertainty)



# 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

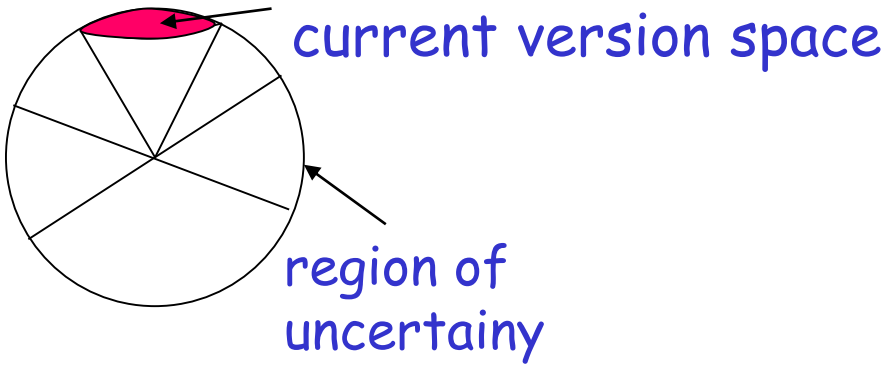
## Disagreement Based Active Learning

## Hypothesis Space Search

[CAL92] [BBL06]

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

# 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.

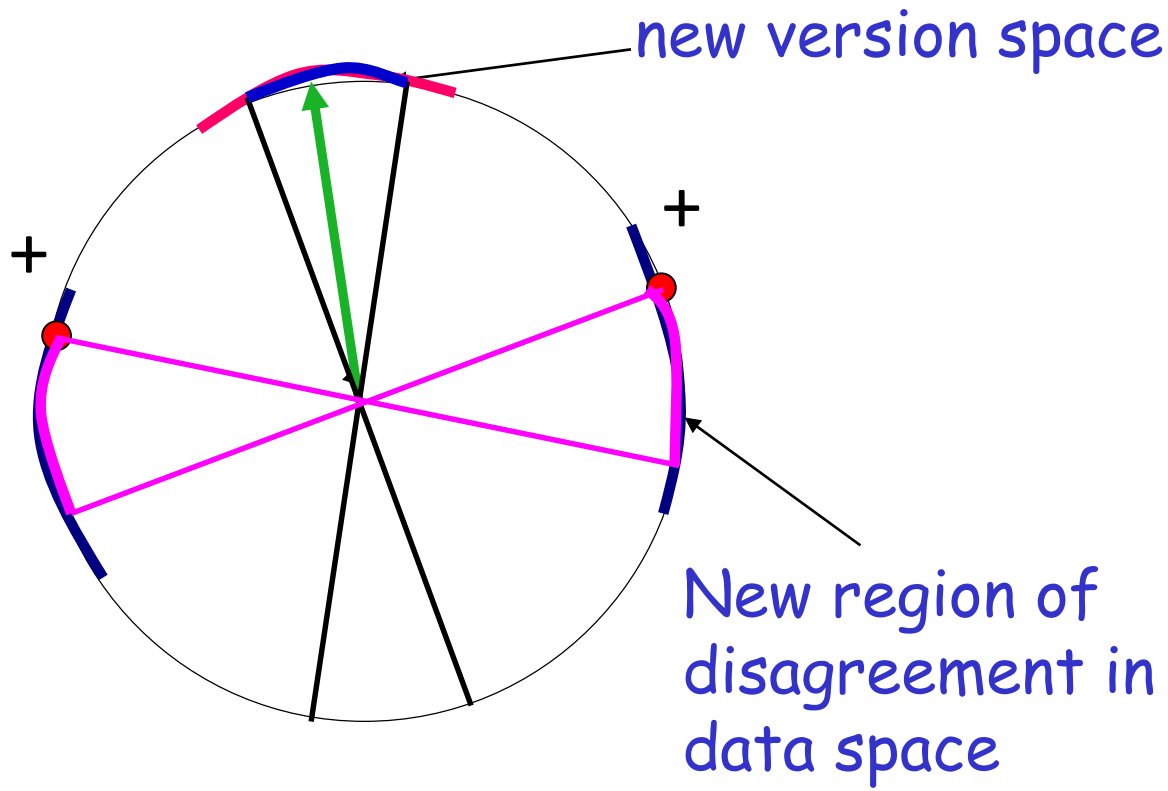
**For** $t = 1, \ldots,$

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

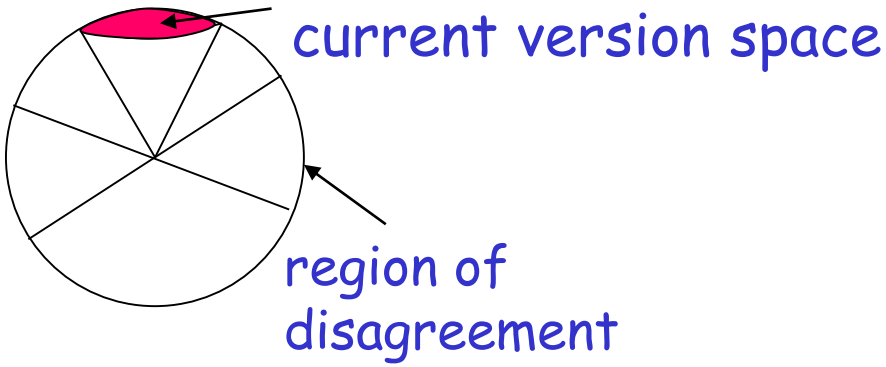
Let $H_{t+1}$ be the new version space.

# 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)



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




**Algorithm:**

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

Let $H_1 = H$.

**For** $t = 1, \ldots,$

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

# What You Should Know

- Active learning could be really helpful, could provide exponential improvements in label complexity (both theoretically and practically)!

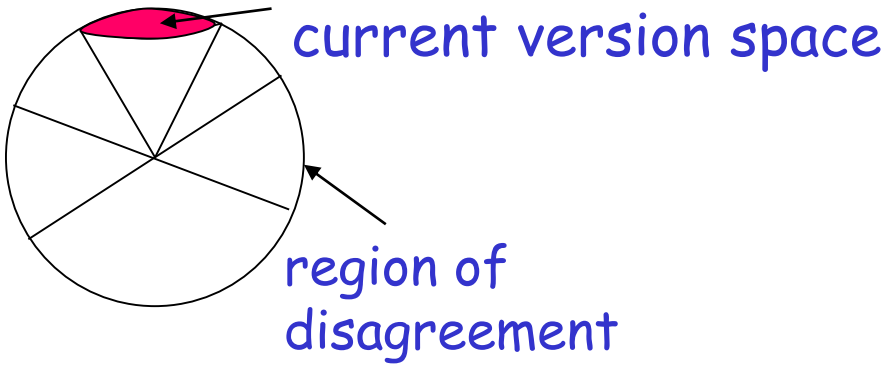
- Common heuristics (e.g., those based on uncertainty sampling). Need to be very careful due to sampling bias.

- 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:**

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

Let $H_1 = H$.

**For** $t = 1, \ldots,$

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

# Formal General Guarantees for Agnostic AL

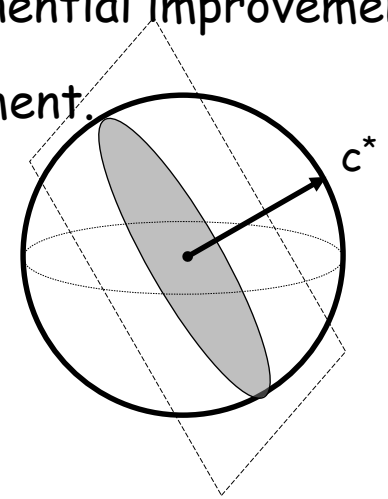
$A^2$ the first algorithm which is robust to noise.

[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 $R^d$, $D$ - uniform, low noise, only $d^2 \log(1/\varepsilon)$ 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

<u>Guarantees for $A^2$</u> [Hanneke'07]:

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

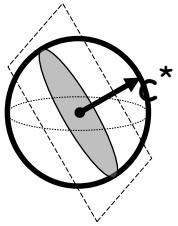
**Theorem**

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

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

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

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, …]