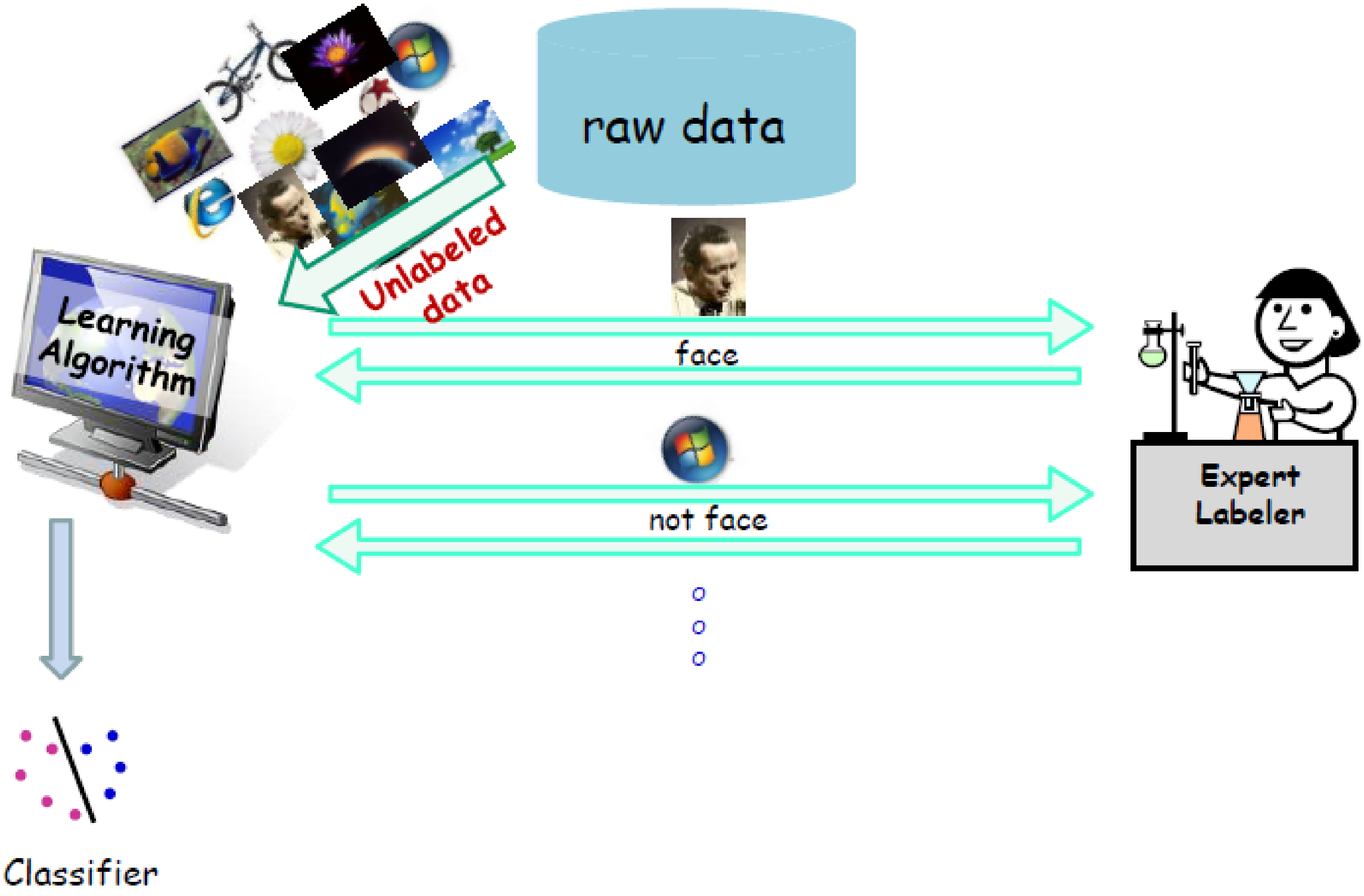
Lecture 03: Active Learning

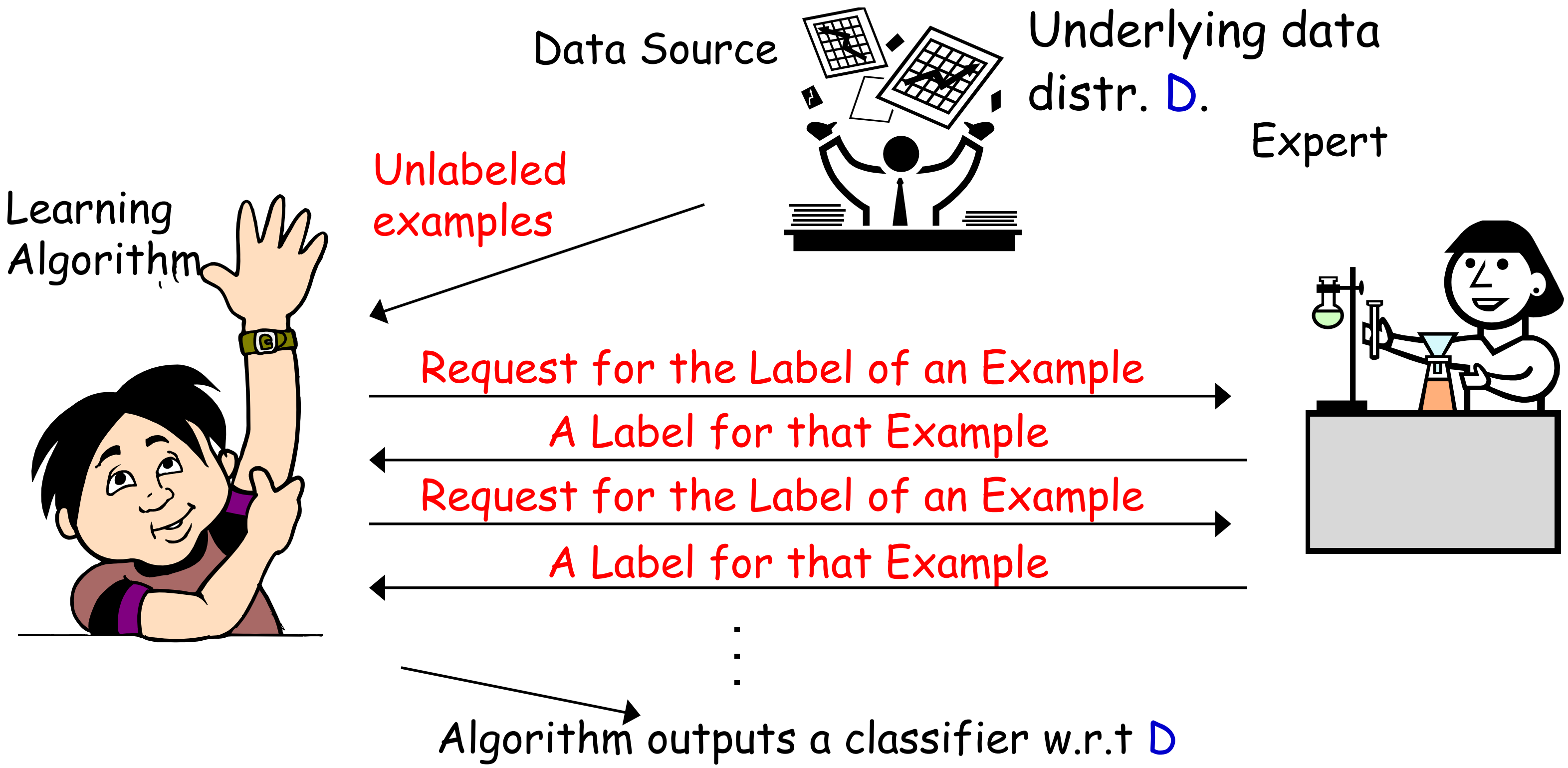
Readings:

- Survey Paper (posted)

Active Learning



Batch Active Learning



Data Source

Underlying data distr. D.

Expert

Unlabeled examples

Learning Algorithm

Request for the Label of an Example

A Label for that Example

Request for the Label of an Example

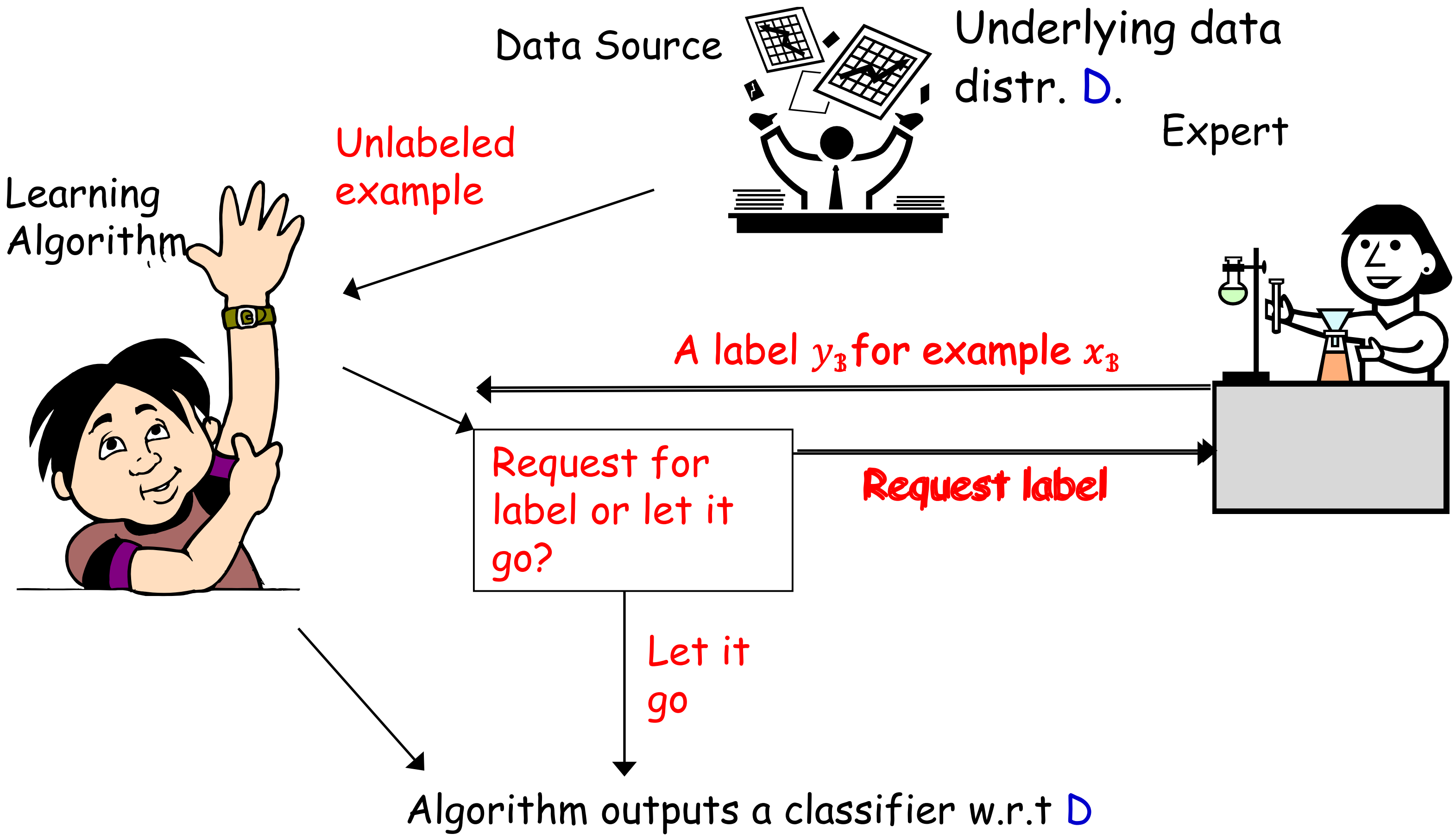
A Label for that Example

. . .

Algorithm outputs a classifier w.r.t D

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

Selective Sampling Active Learning



Data Source

Underlying data distr. D.

Expert

Unlabeled example

Learning Algorithm

A label $y_{\mathcal{B}}$ for example $x_{\mathcal{B}}$

Request for label or let it go?

Request label

Let it go

Algorithm outputs a classifier w.r.t D

- 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?

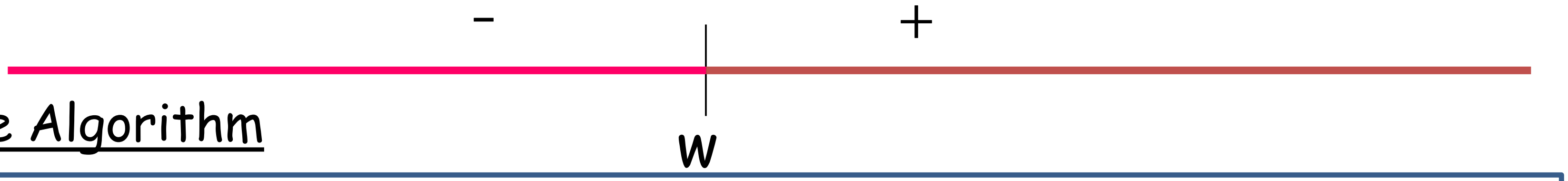
- YES! (sometimes)

- We often need far fewer labels for active learning than for passive.

- This is predicted by theory and has been observed in practice.

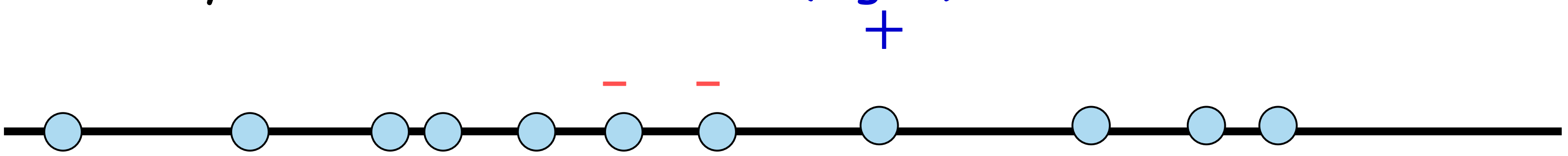
Can adaptive querying help? [CAL92, Dasgupta04]

- Threshold fns on the real line: $h_w(x) = 1(x ¸ w)$, $C = \{h_w : w\ 2\ R\}$



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 ε-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; Schohon 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)

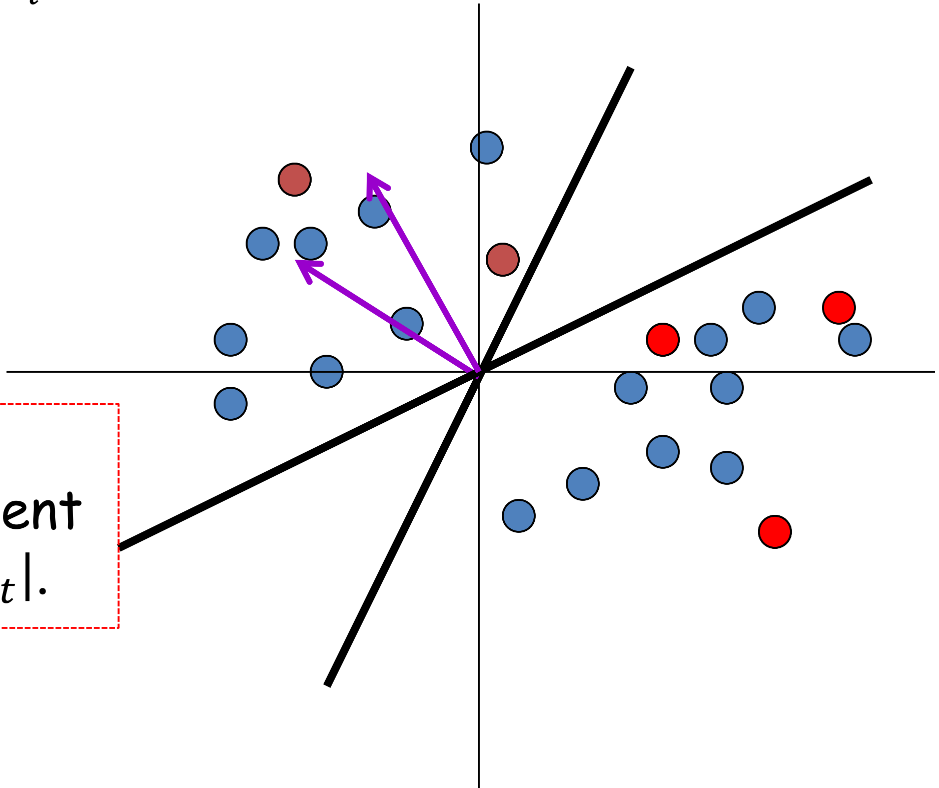
Input $S_u=\{x_1, \ldots, x_{m_u}\}$ drawn i.i.d from the underlying source D

Start: query for the labels of a few random x_is.

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)

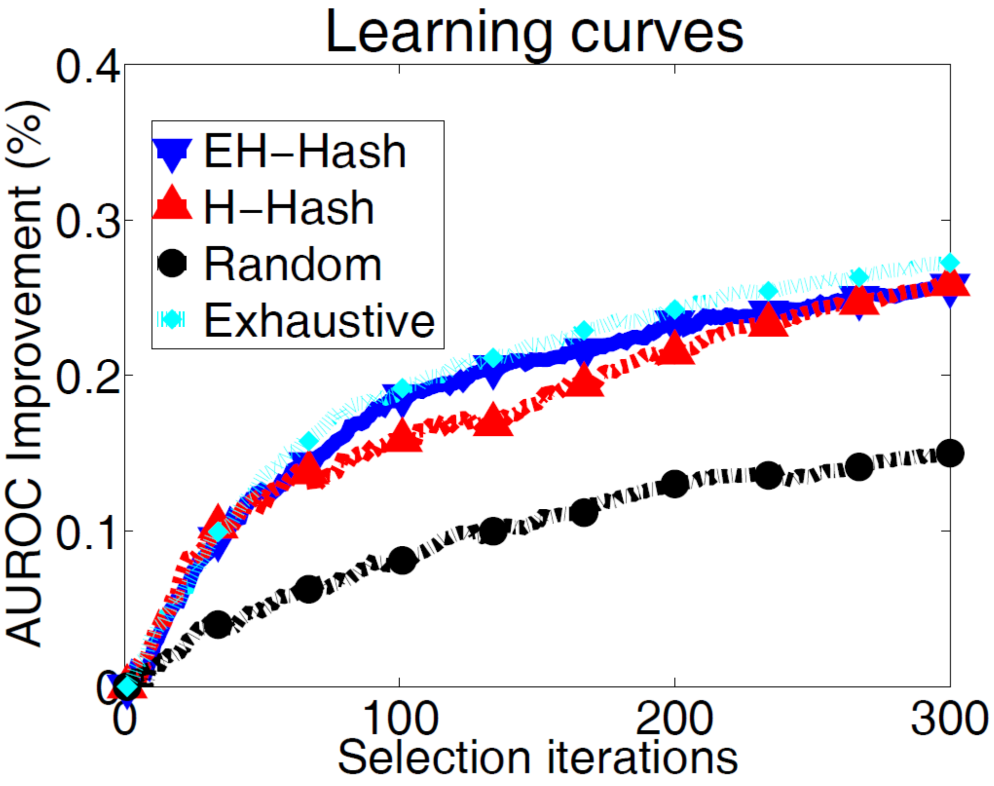


Common Technique in Practice

Active SVM seems to be quite useful in practice.

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

Newsgroups dataset (20.000 documents from 20 categories)

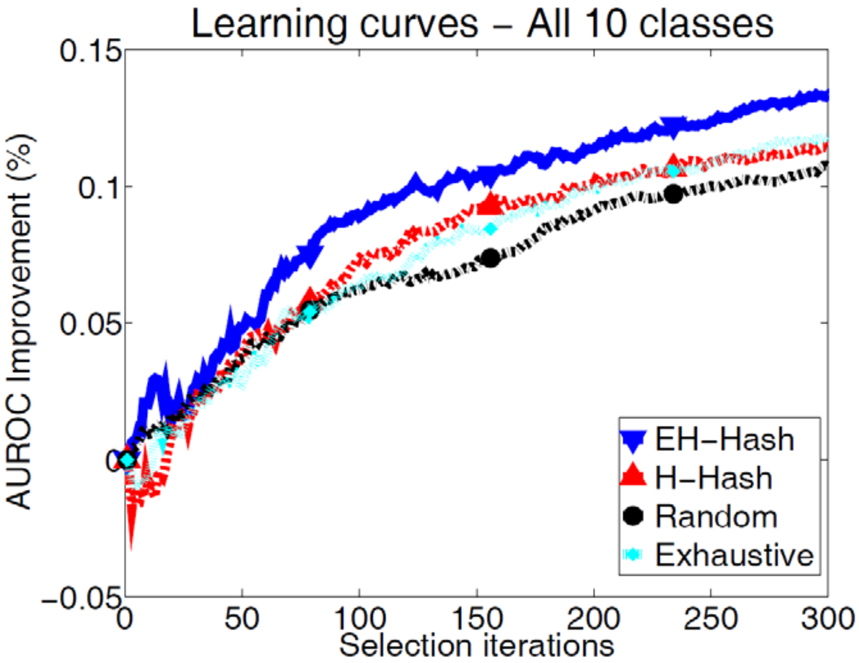


Common Technique in Practice

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

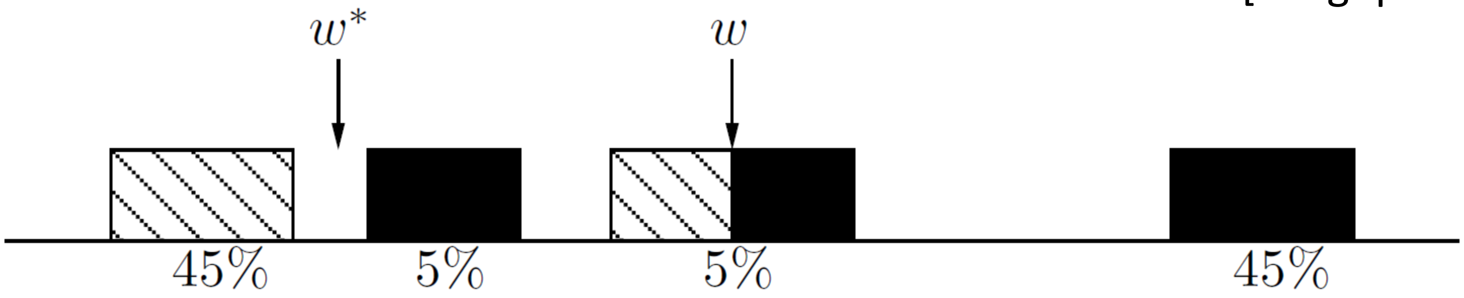
- Works sometimes….

- **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.

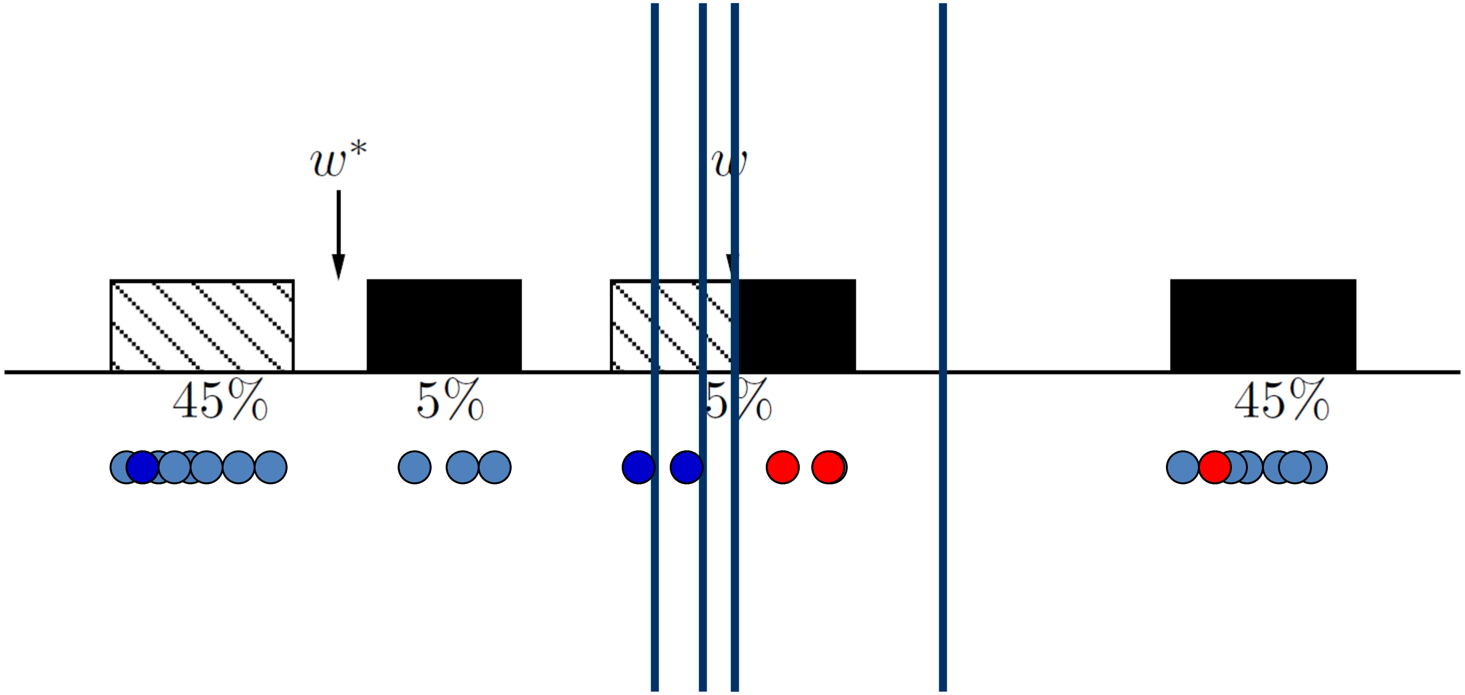
[Dasgupta10]



Active SVM/Uncertainty Sampling

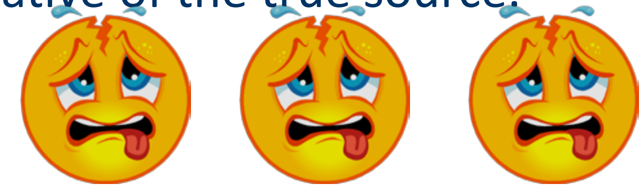
- Works sometimes….

- **However, we need to be very very careful!!!**

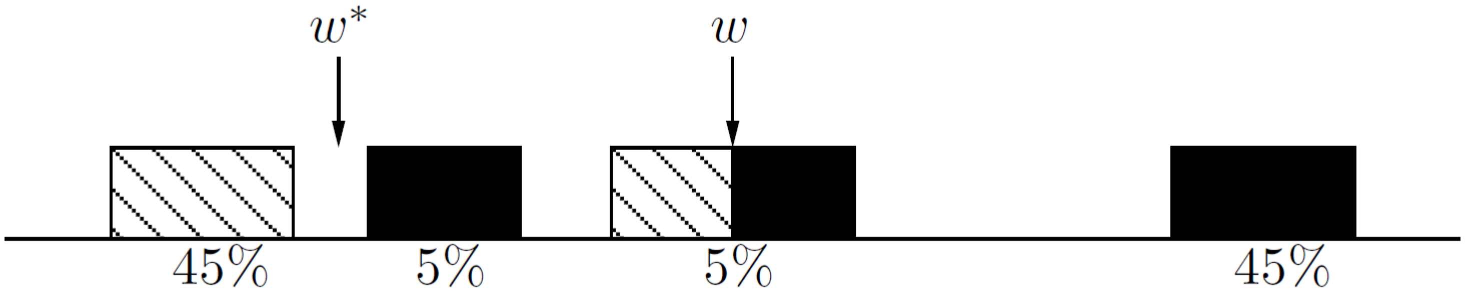


Active SVM/Uncertainty Sampling

- 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 $(x_1, y_1), \ldots, (x_{m_l}, y_{m_l})$, $y_i = c^*(x_i)$

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

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

Version Spaces

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

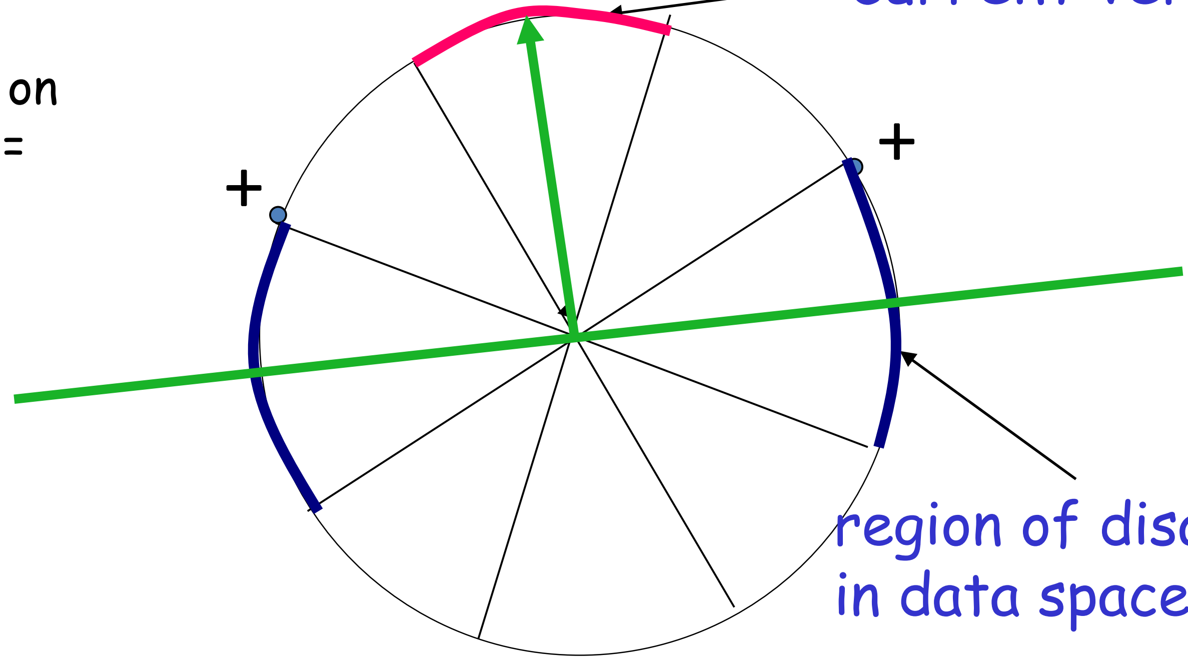
Definition (Mitchell'82) Assume realizable case: $c^* \in H$.

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

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

E.g.,: data lies on circle in R^2, H = homogeneous linear seps.

current version space



region of disagreement in data space

Version Spaces. Region of Disagreement

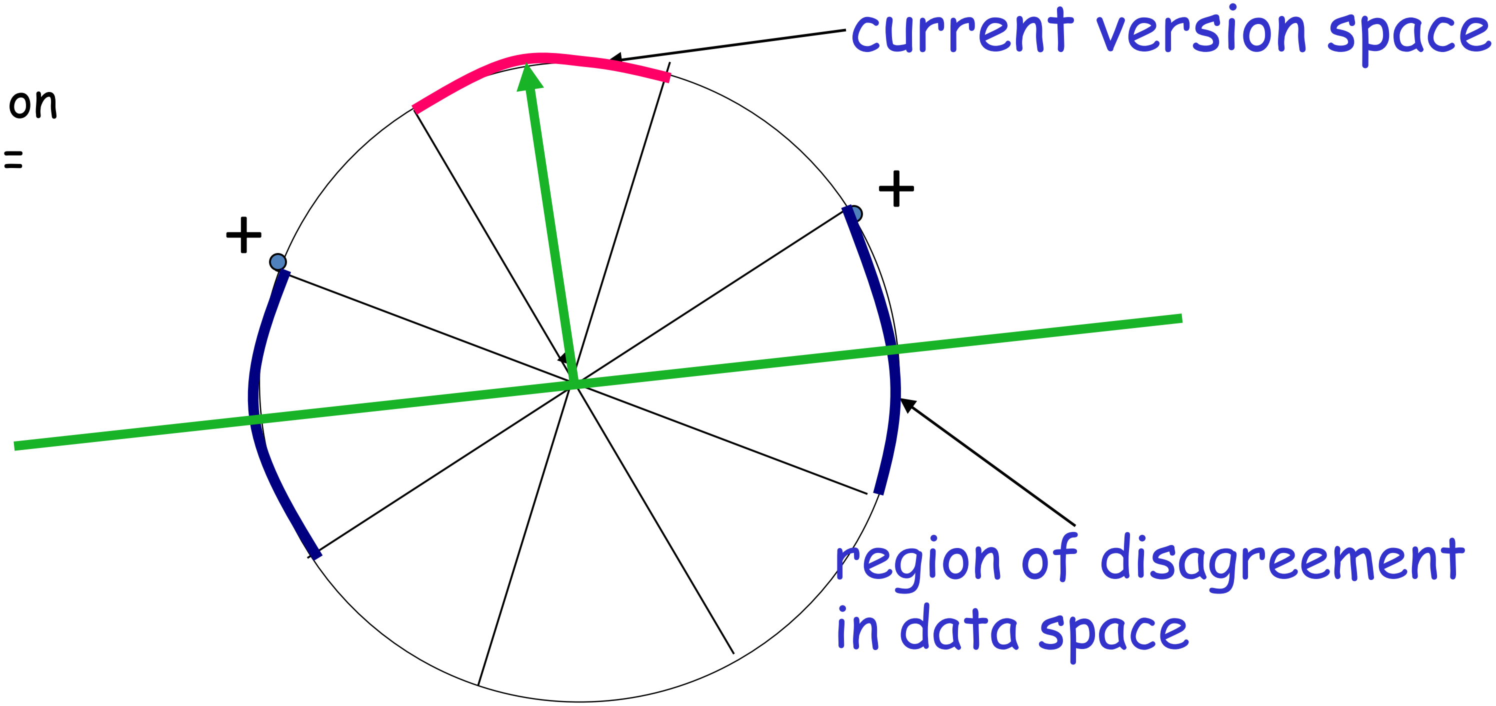
Definition (CAL'92)

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

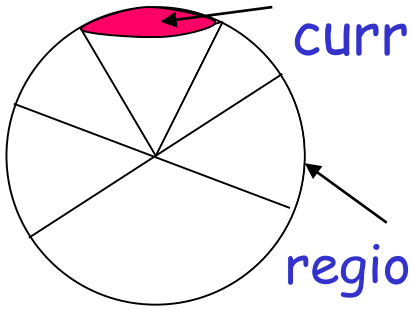
Region of disagreement = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

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

E.g.,: data lies on circle in R^2, H = homogeneous linear seps.



Disagreement Based Active Learning [CAL92]



current version space

region of uncertainy

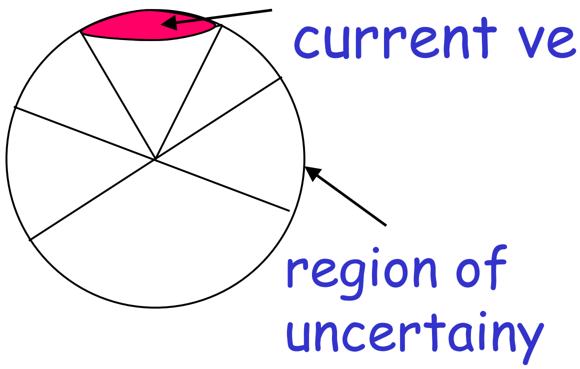
Algorithm:

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_is.

Let H_1 be the current version space.

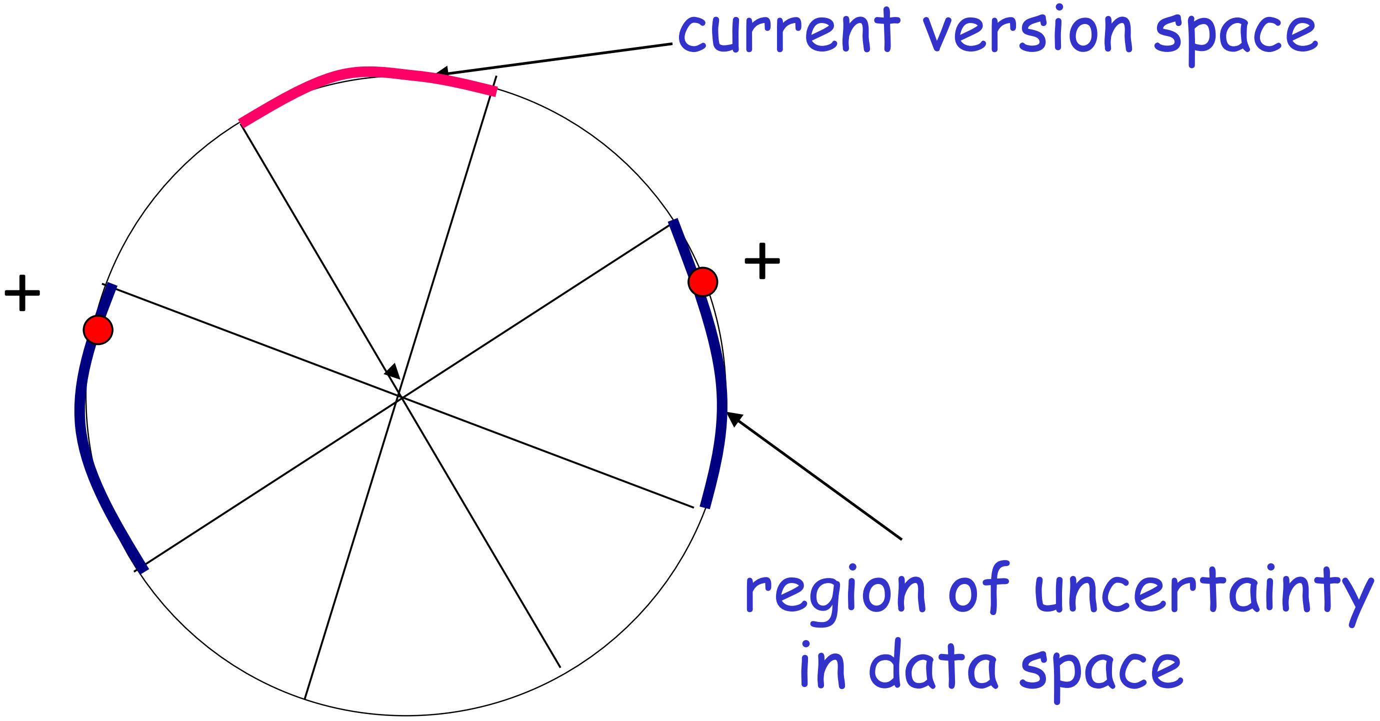
For $\boldsymbol{t = 1}$**, ….,**

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

Let $\mathrm{H}_{\mathrm{t+1}}$ be the new version space.

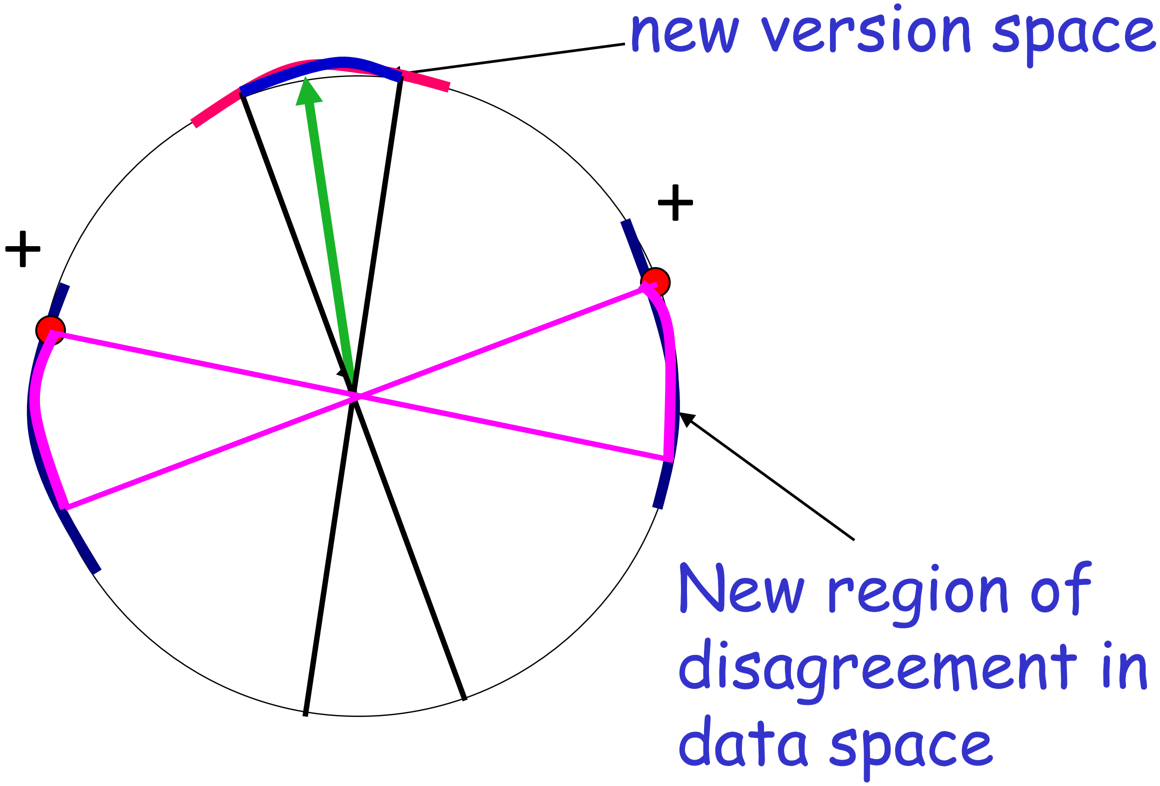
Region of uncertainty [CAL92]

- 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)



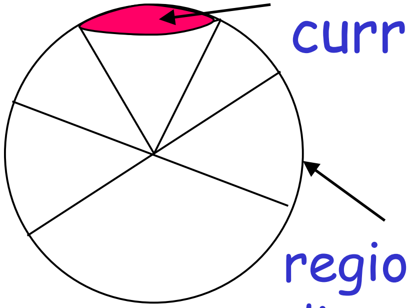
Region of uncertainty [CAL92]

- 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]



current version space

region of disagreement

Algorithm:

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

Let $H_1 = H.$

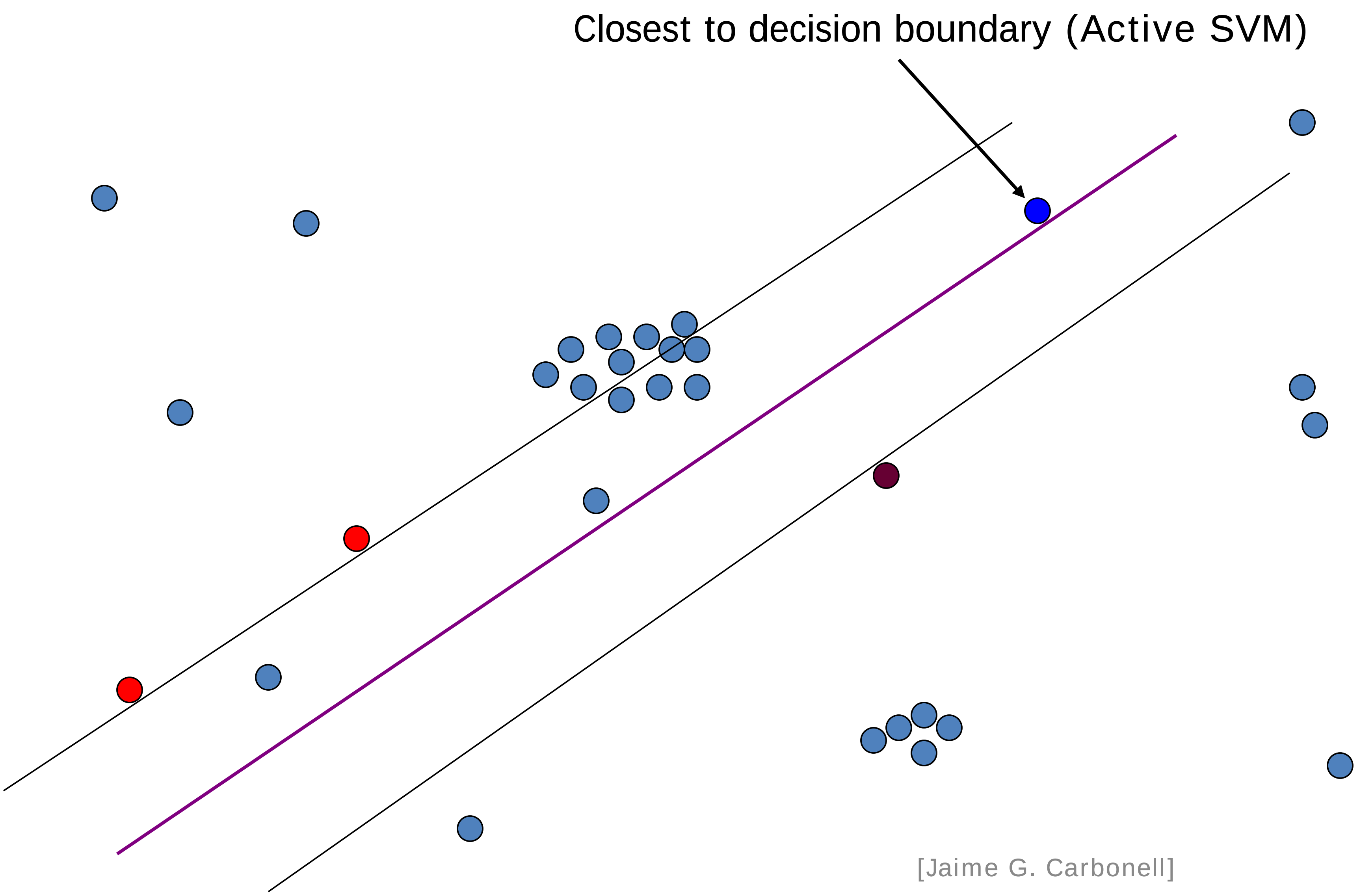
For $\boldsymbol{t = 1, \ldots.,}$

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

Other Interesting AL Techniques/ Heuristics used in Practice

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

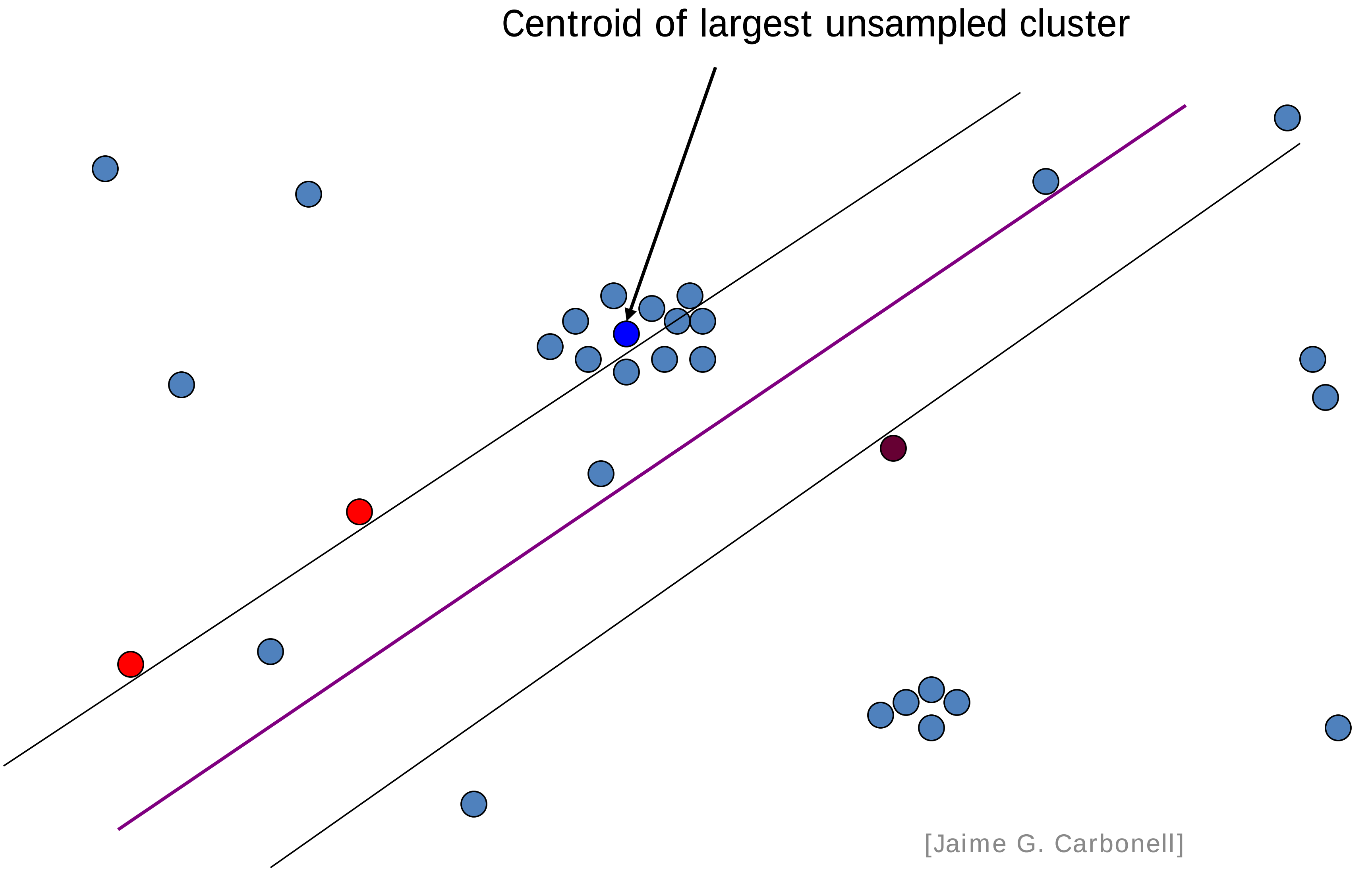
Uncertainty Sampling



Closest to decision boundary (Active SVM)

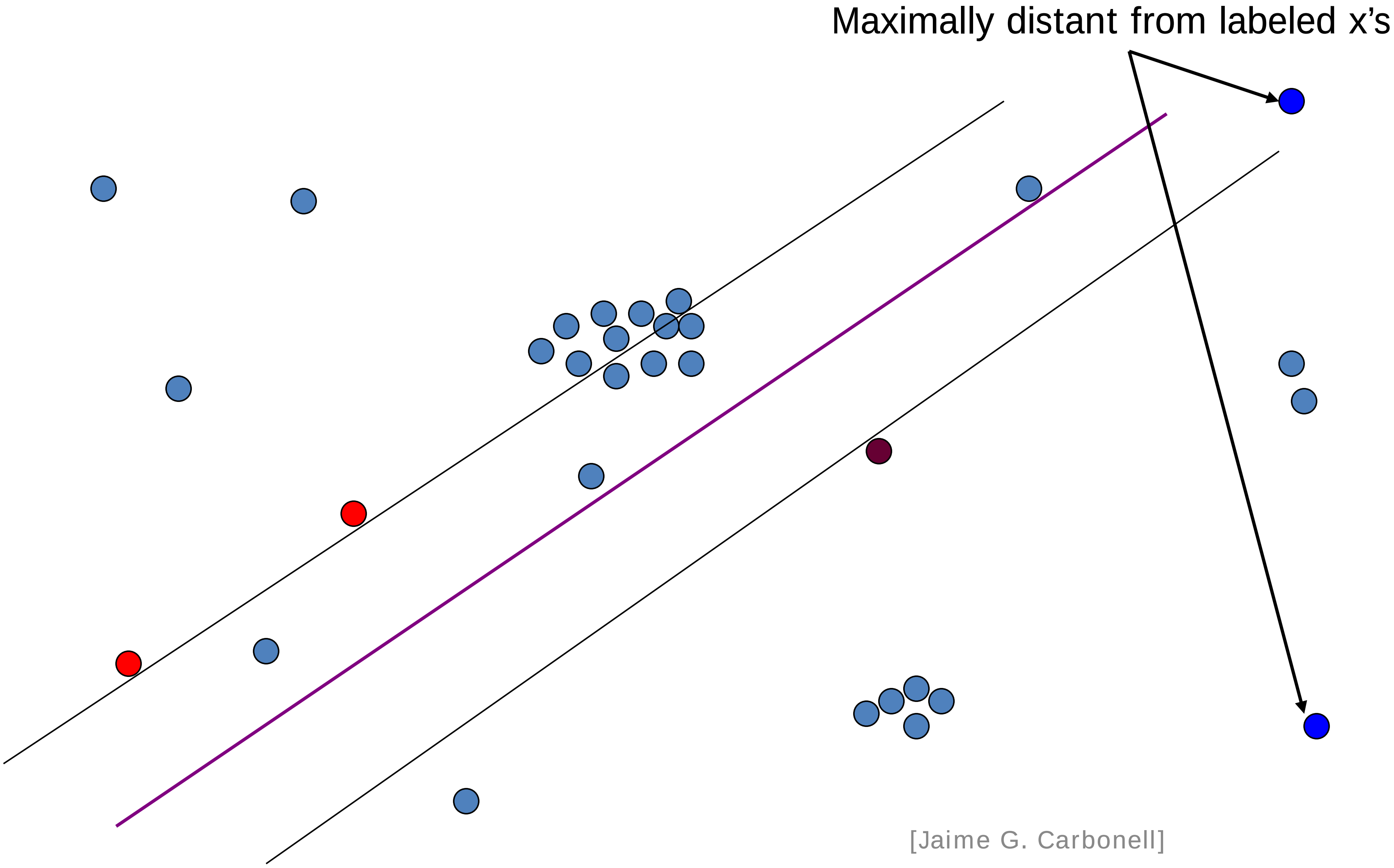
[Jaime G. Carbonell]

Density-Based Sampling



[Jaime G. Carbonell]

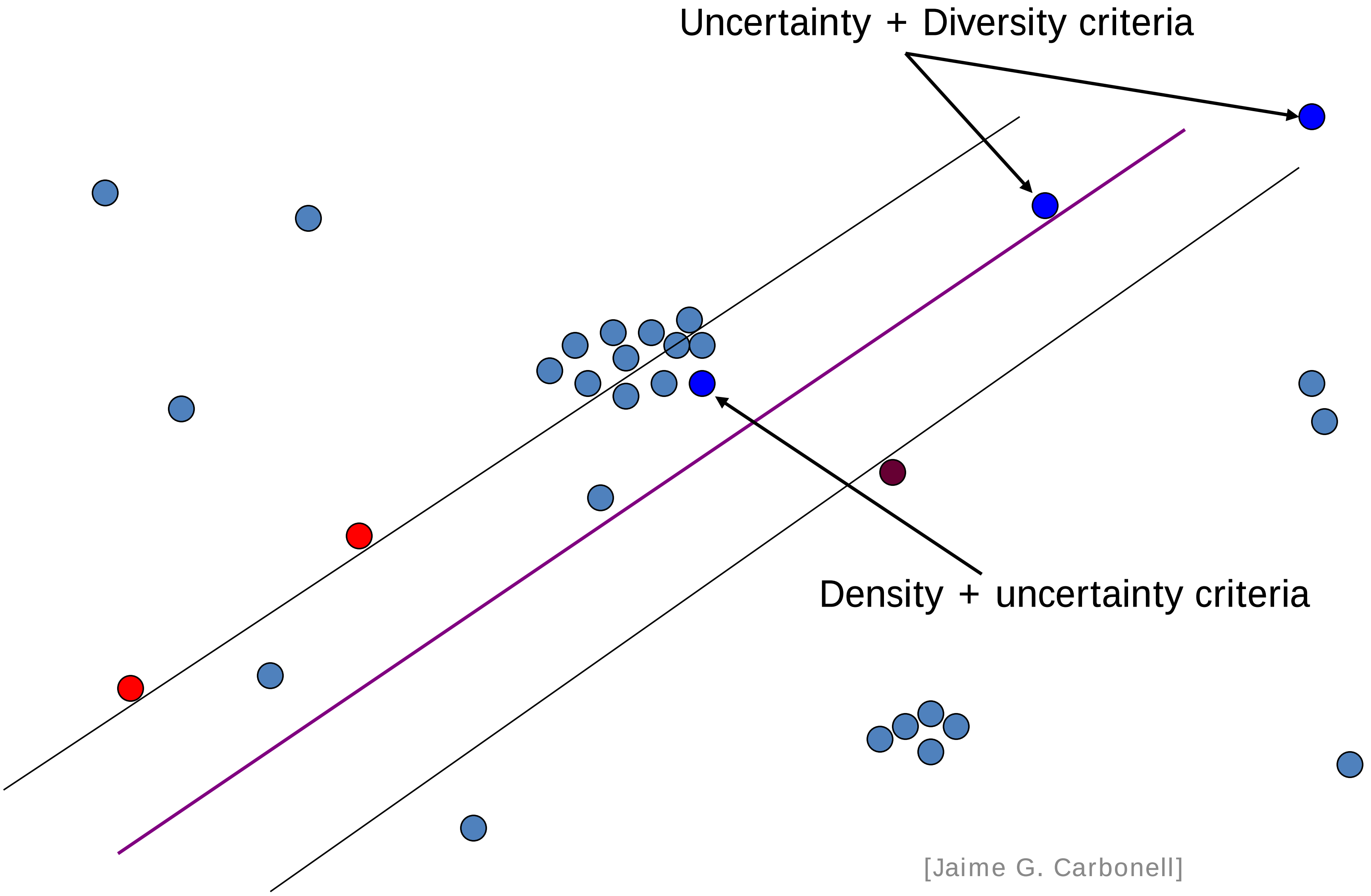
Maximal Diversity Sampling



Maximally distant from labeled x’s

[Jaime G. Carbonell]

Ensemble-Based Possibilities



Uncertainty + Diversity criteria

Density + uncertainty criteria

[Jaime G. Carbonell]