Machine Learning

Topic: Active Learning

Concept Learning

- Much of learning involves acquiring general concepts from specific training examples
- *Concept*: subset of objects from some space
- Concept learning: Defining a function that specifies which elements are in the concept set.

Some terms

X is the set of all possible instances

C is the set of all possible concepts c
where $c : X \rightarrow \{0,1\}$

H is the set of hypotheses considered
by a learner, $H \subseteq C$

L is the learner

D is a probability distribution over X
that generates observed instances

Concept Learning Task

GIVEN:

- Instances X
- Target function c $\rightarrow \{0,1\}$
- Hypothesis space H
- Training examples D $= \{<x_1,c(x_1)>,\ldots, <x_n,c(x_n)>\}$

FIND:

- A hypothesis h in H such that h(x)=c(x) for all x in D.

Labeling examples

Too time consuming

- Example 1: Netflix Challenge

  Concept: movies Bob would like

  Instances: 10,000 movies on netflix

  Labeling: Bob watches a movie and reports

- Example 2: Labeling phonemes

  Concept: words labeled with phonetic alphabet

  Instances: 1000 hours of talk radio recordings

  Labeling: Hire linguist to annotate each syllable

The BIG IDEA

- If we just pick the RIGHT examples to label, we can learn the concept from only a few labeled examples

  (it’s like 20 questions)

Active Learning Heuristic

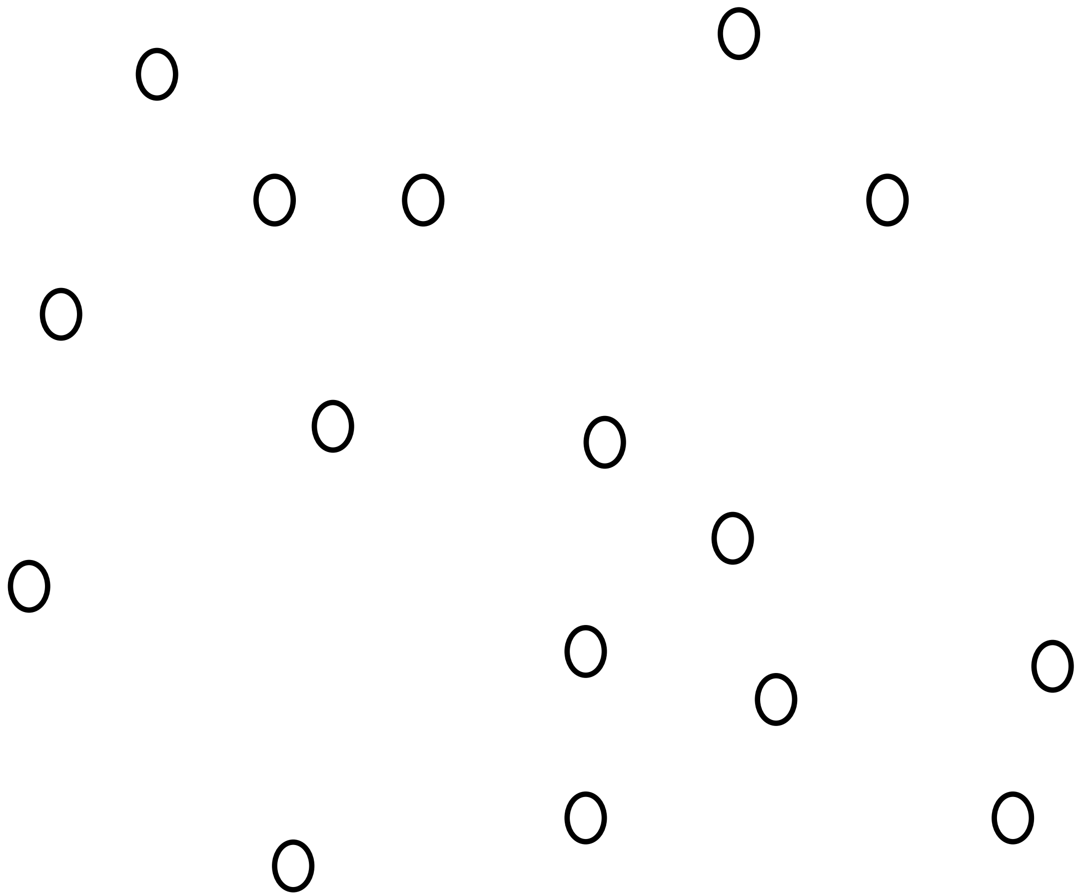
- Start with a pool of unlabeled data
- Pick a few points at random and get their labels
- Repeat the following
  1. Fit a classifier to the labels seen so far
  2. Pick the BEST unlabeled point to get a label for

     (closest to the boundary?)

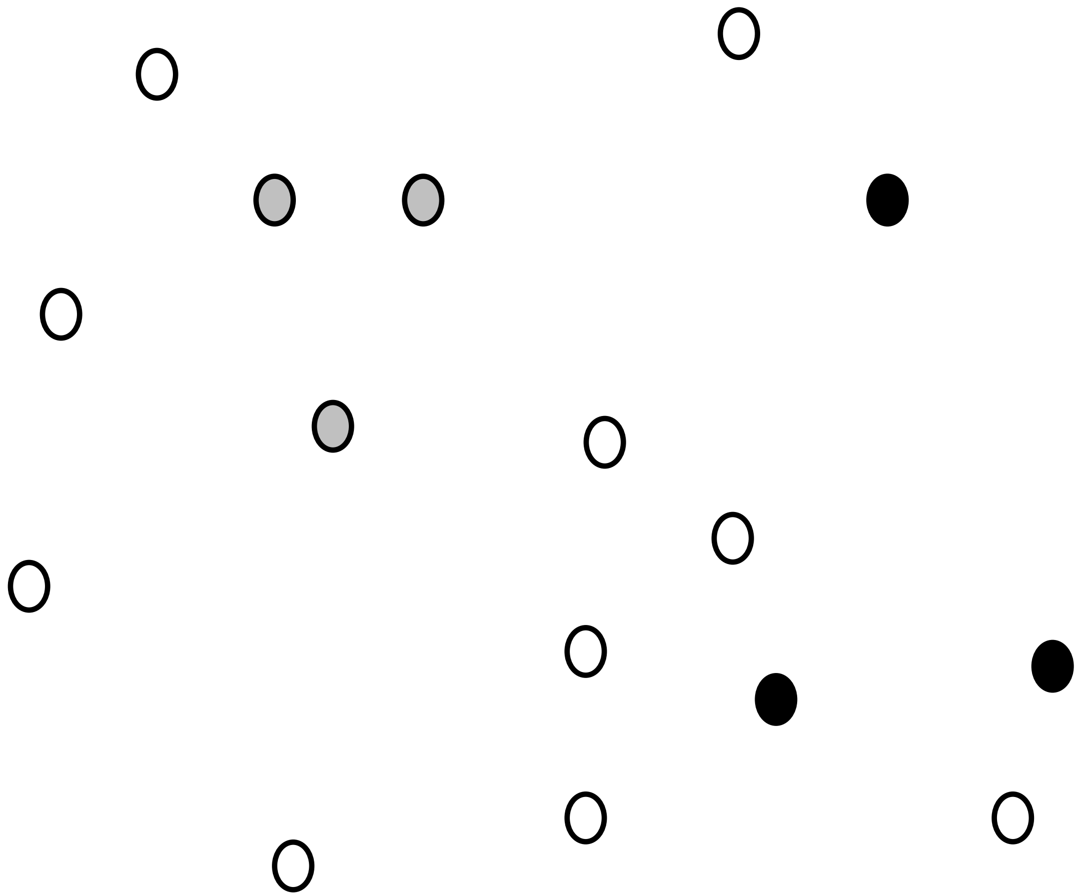
     (most uncertain?)

     (most likely to decrease overall uncertainty?)

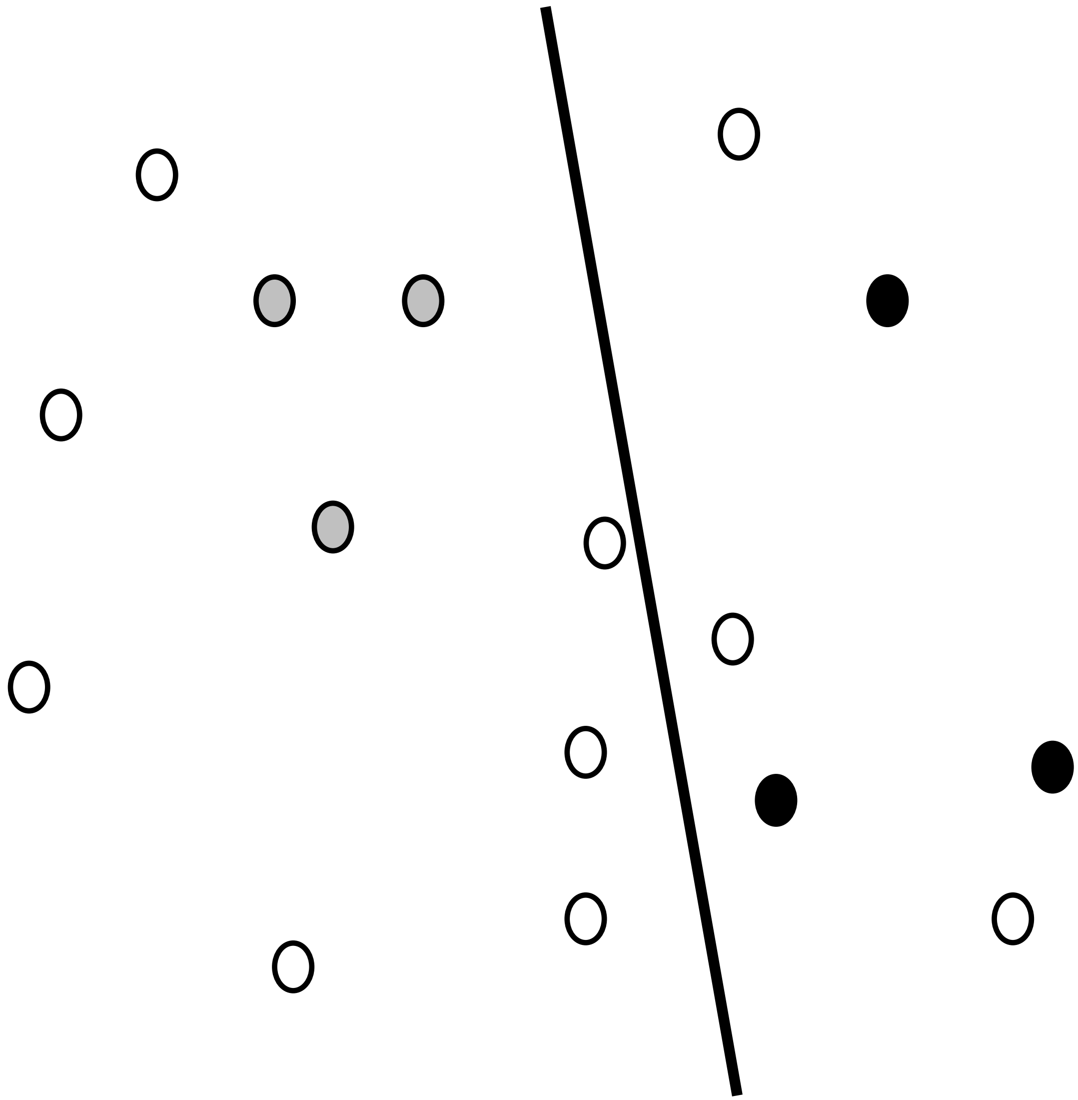
Start: Unlabeled Data



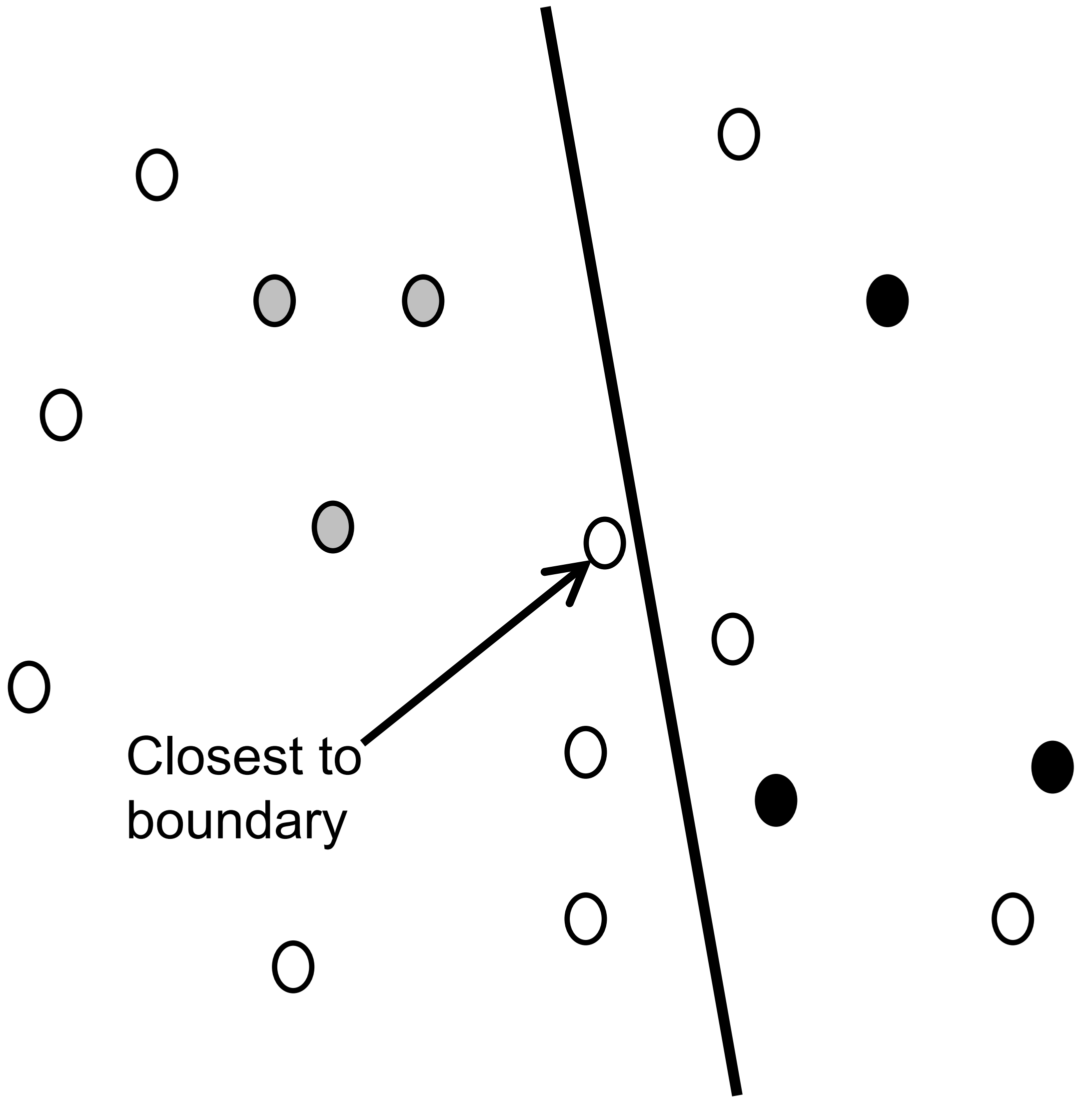
Label a Random Subset



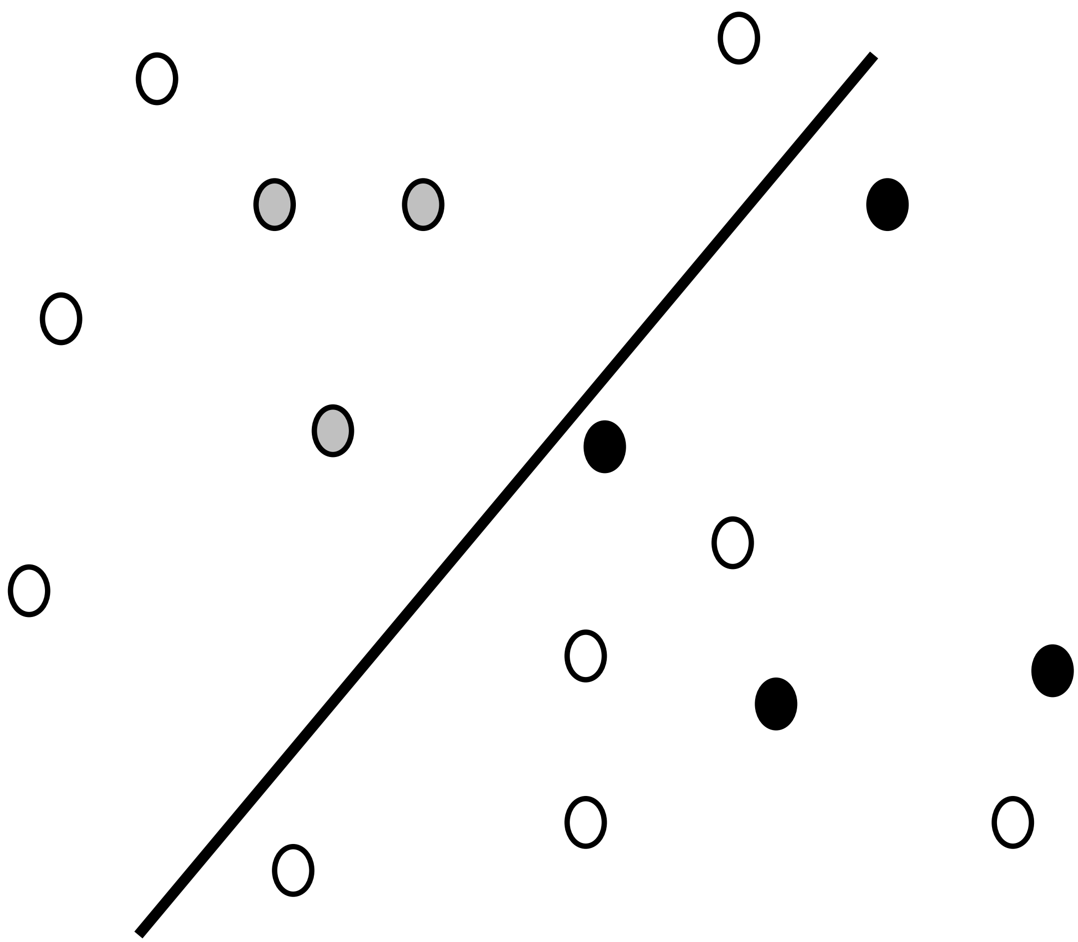
Fit a Classifier to Labeled Data



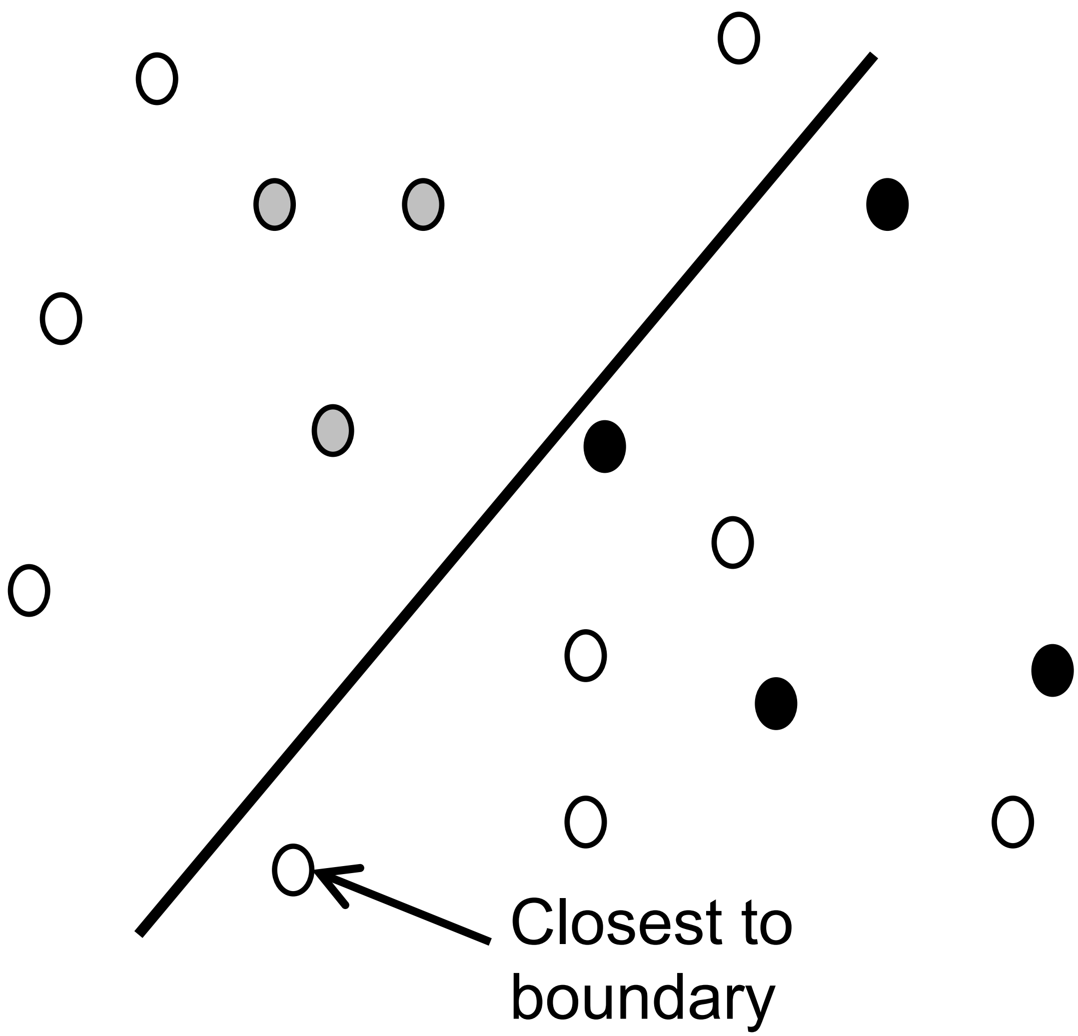
Pick the Best Next Point To Label



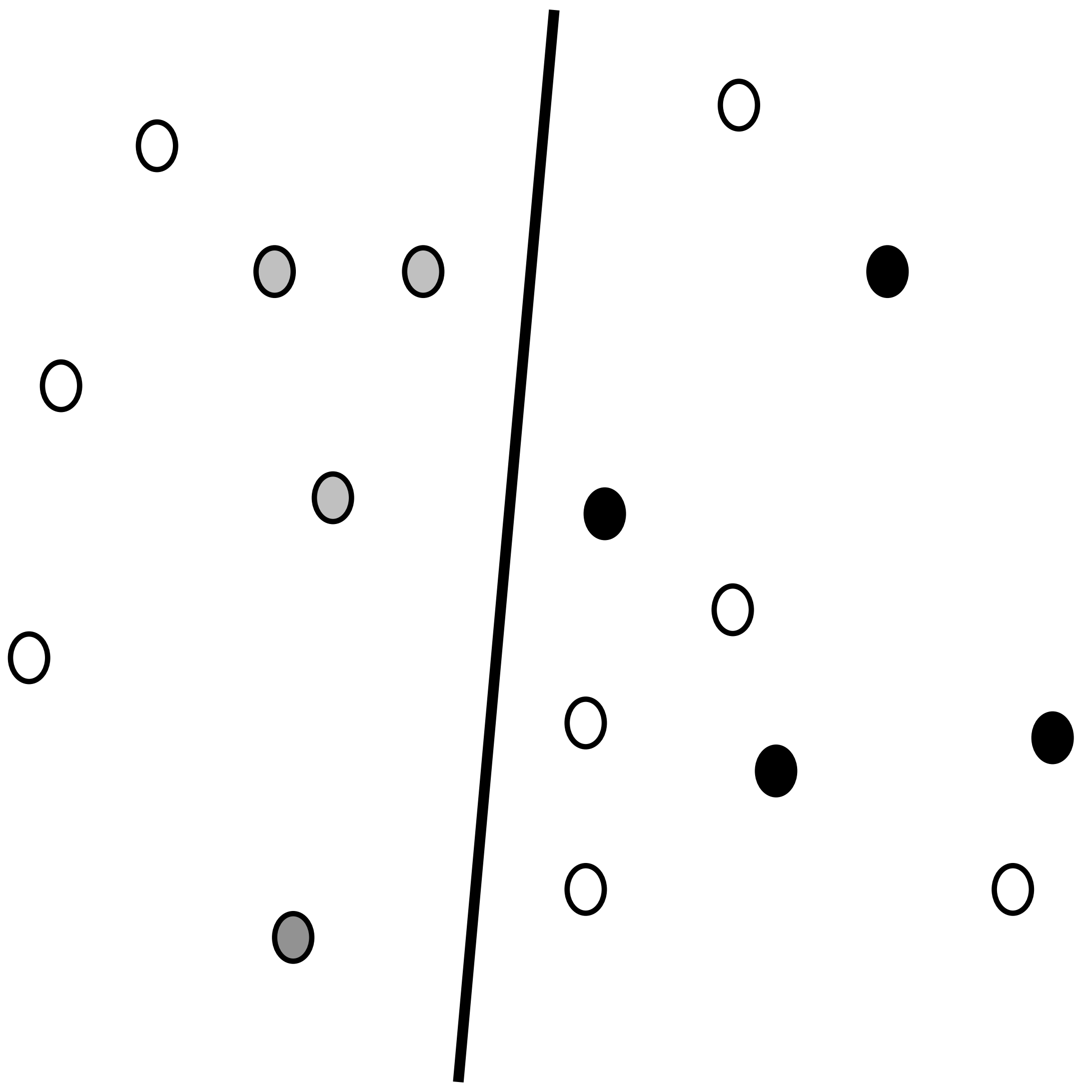
Fit a Classifier to Labeled Data



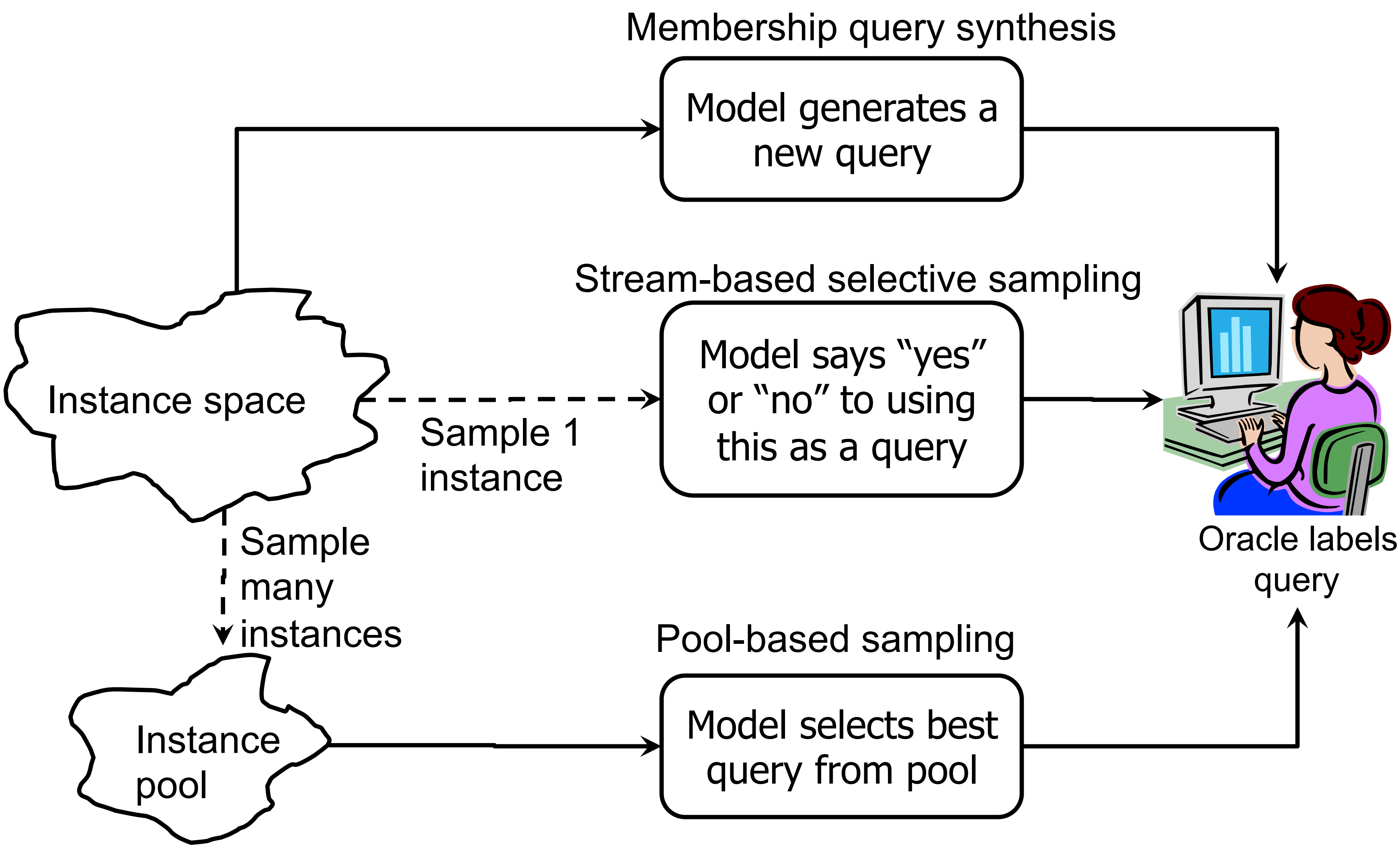
Pick the Best Next Point To Label



Fit a Classifier to Labeled Data



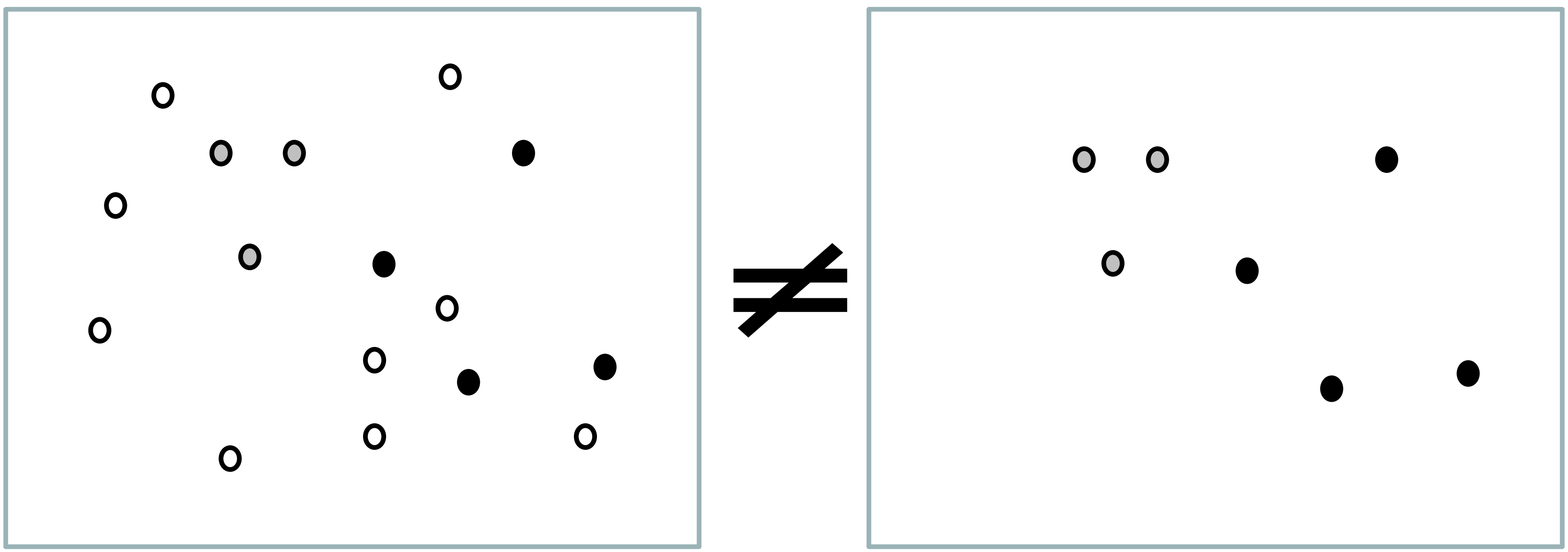
3 Approaches to Querying



Biased Sampling

- The labeled points may not be representative of the underlying distribution

- This can increase error in the limit (as number of labeled examples goes to infinity) (Schutze et al 03)



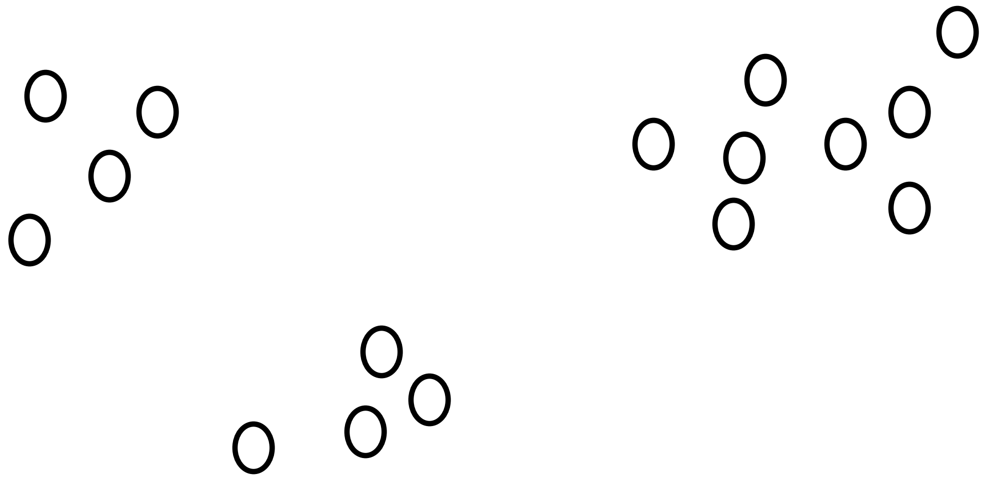
Two Rationales for Active Learning

Rationale 1: We can exploit cluster structure in data

Rationale 2: We can efficiently search through the hypothesis space

Exploiting structure in data

If the data looked like this…



…then we might just need 3 labeled points

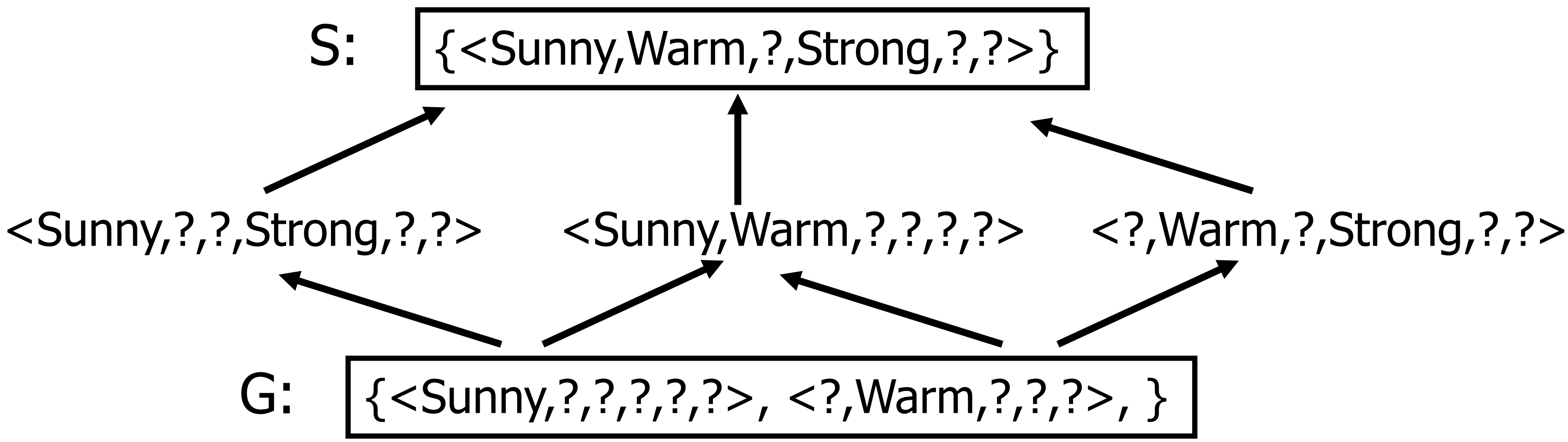
Issues:

- Structure may not be so clearly defined
- Structure exists at many levels of granularity
- Clusters may not be all one label

Efficient Hypothesis Search

If each query cuts the version space in 2, we may need only log(|H|) to get a perfect hypothesis.

Which example should we label?



x_5 = <Sunny Warm Normal Strong Cool Change> + 6/0

x_6 = <Rainy Cold Normal Light Warm Same> - 0/6

x_7 = <Sunny Warm Normal Light Warm Same> ? 3/3

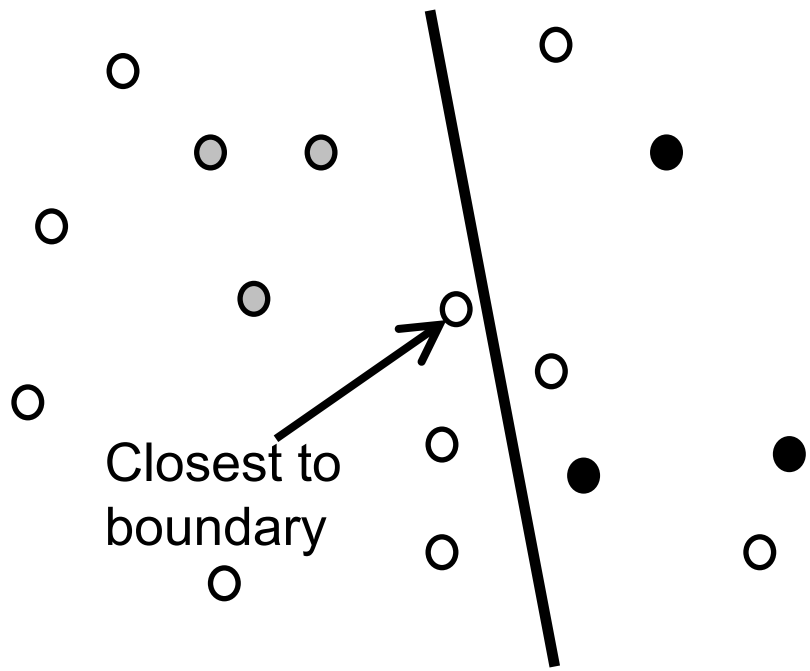
x_8 = <Sunny Cold Normal Strong Warm Same> ? 2/4

Questions

- Do there always exist queries that will cut off a good portion of the version space?
- If so, how can these queries be found?
- What happens in the nonseparable case?

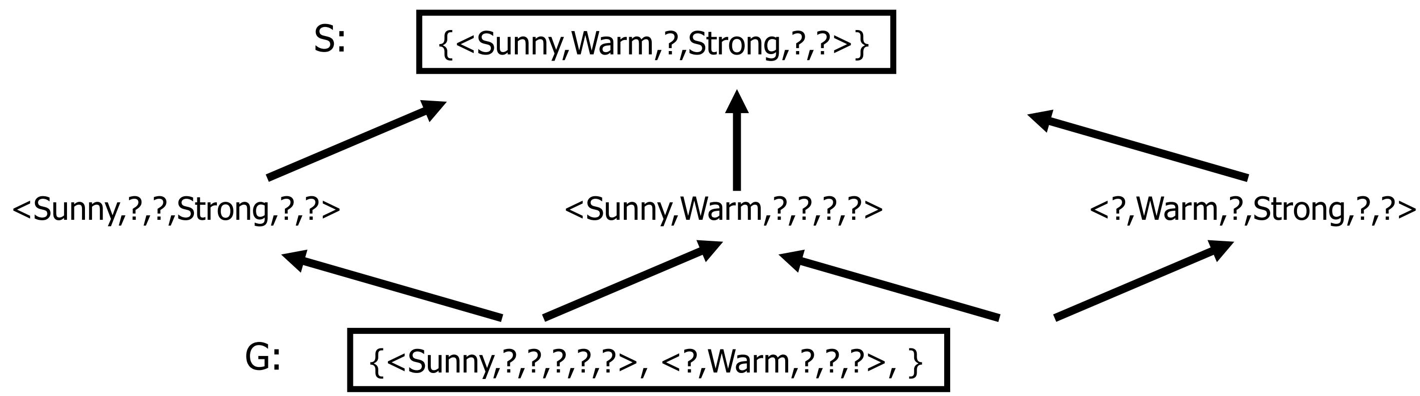
Query Selection Strategies

- Uncertainty Sampling
  - A single model
  - Query the instances we are least certain how to label
    (e.g. closet to the decision boundary)



Query Selection Strategies

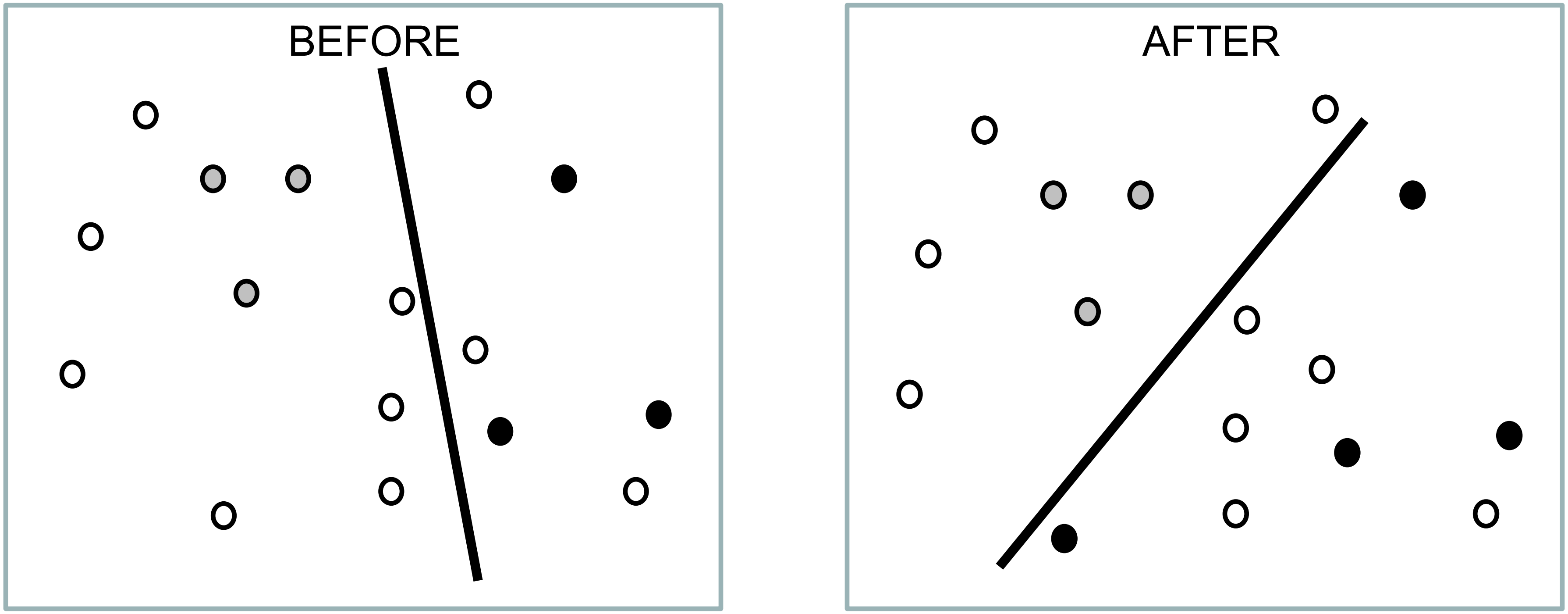
- Query by Committee (QBC)
  - Maintain a version space of hypotheses
  - Pick the instances generating the most disagreement among hypotheses



x_5 = <Sunny Warm Normal Strong Cool Change>	+ 6/0
x_6 = <Rainy Cold Normal Light Warm Same>	- 0/6
x_7 = <Sunny Warm Normal Light Warm Same>	? 3/3
x_8 = <Sunny Cold Normal Strong Warm Same>	? 2/4

Query Selection Strategies

- Expected Model Change
  - A single model
  - Pick the unlabeled instance that would cause the greatest change to the model, if we knew its label



Query Selection Strategies

- Expected Error Reduction
  - A single probabilistic model
  - Query the instances that would most reduce error.
  - **most computationally expensive query framework**
    - we have to estimate given all possible labelings for each new instance

Density Weighting Selections

Pick instances that are both “informative” and “representative”

“informative” = score highly on one of the query evaluation measures discussed earlier

“representative” = inhabit dense regions of the input space

Example Density Weighting