CS344: Introduction to Artificial Intelligence

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 36-37: Foundation of Machine Learning

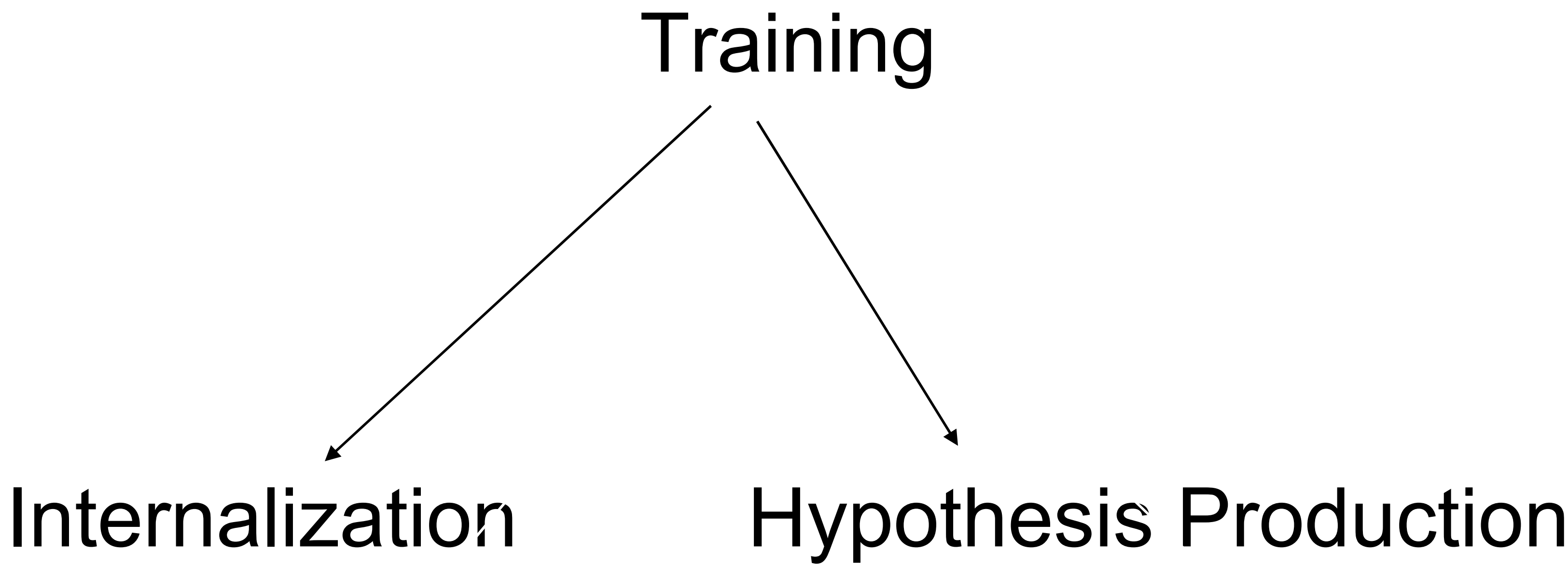
Attempt at formalizing Machine Learning

(Landmark paper by L.G.Valiant, 1984, *A Theory of Learnable,* CACM Journal)

Learning

Training (Loading)

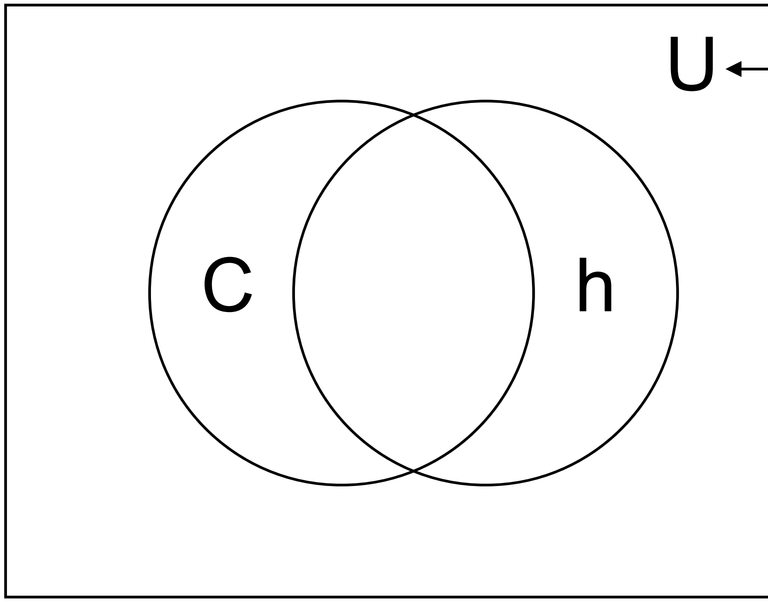
Testing (Generalization)


Training
Internalization
Hypothesis Production

Hypothesis Production

Inductive Bias

In what form is the hypothesis produced?



Universe

$C \oplus h =$ Error region

$$P(C \oplus h) <= \epsilon$$

Prob. distribution

accuracy parameter

P(X) = Prob that x is generated by the teacher – the “oracle” and is labeled

<x, +> : Positive example.

<x, -> : Negative example.

Learning Means the following Should happen:

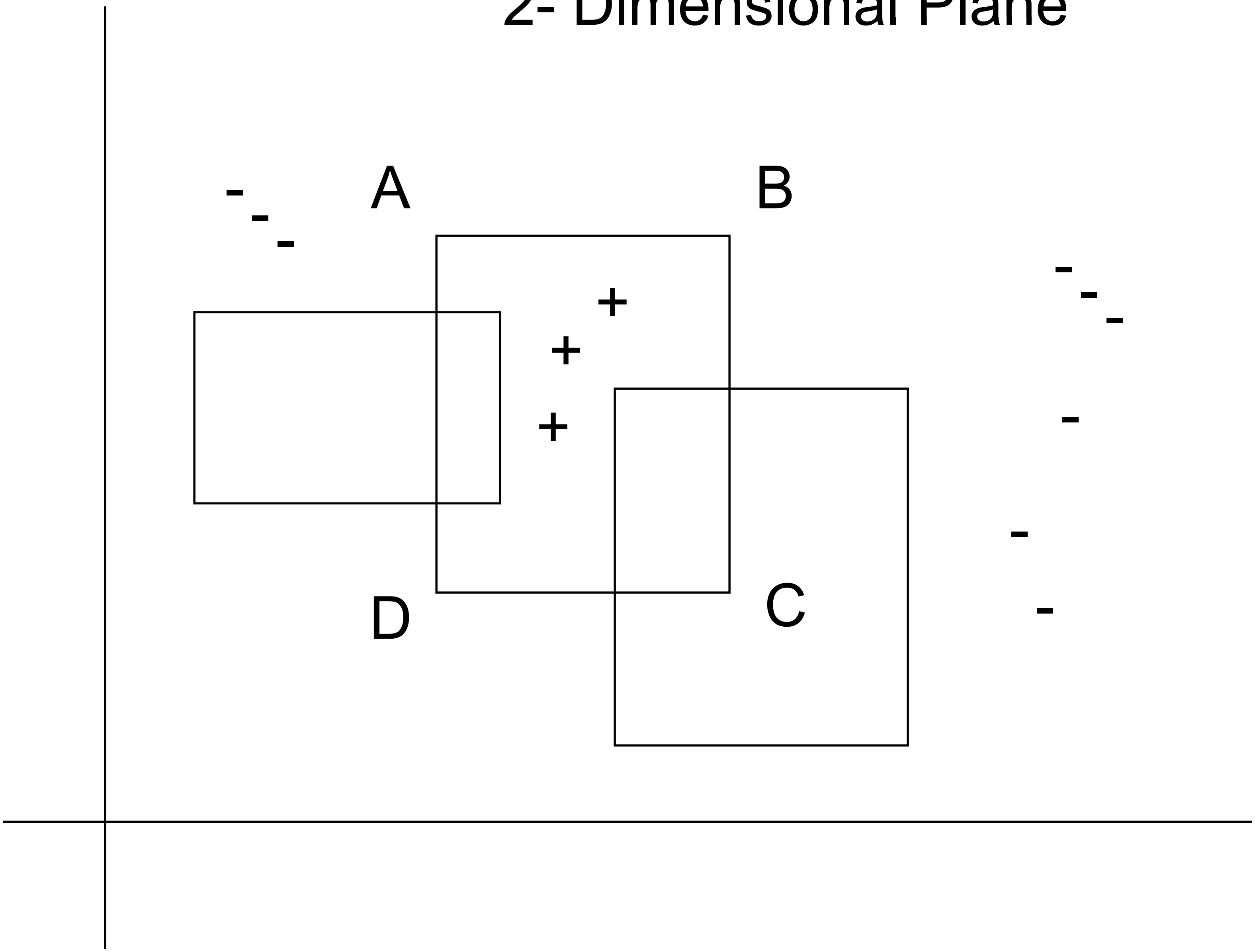
$$Pr(P(c \oplus h) <= \epsilon) >= 1 - \delta$$

PAC model of learning correct.

Probably Approximately Correct

An Example

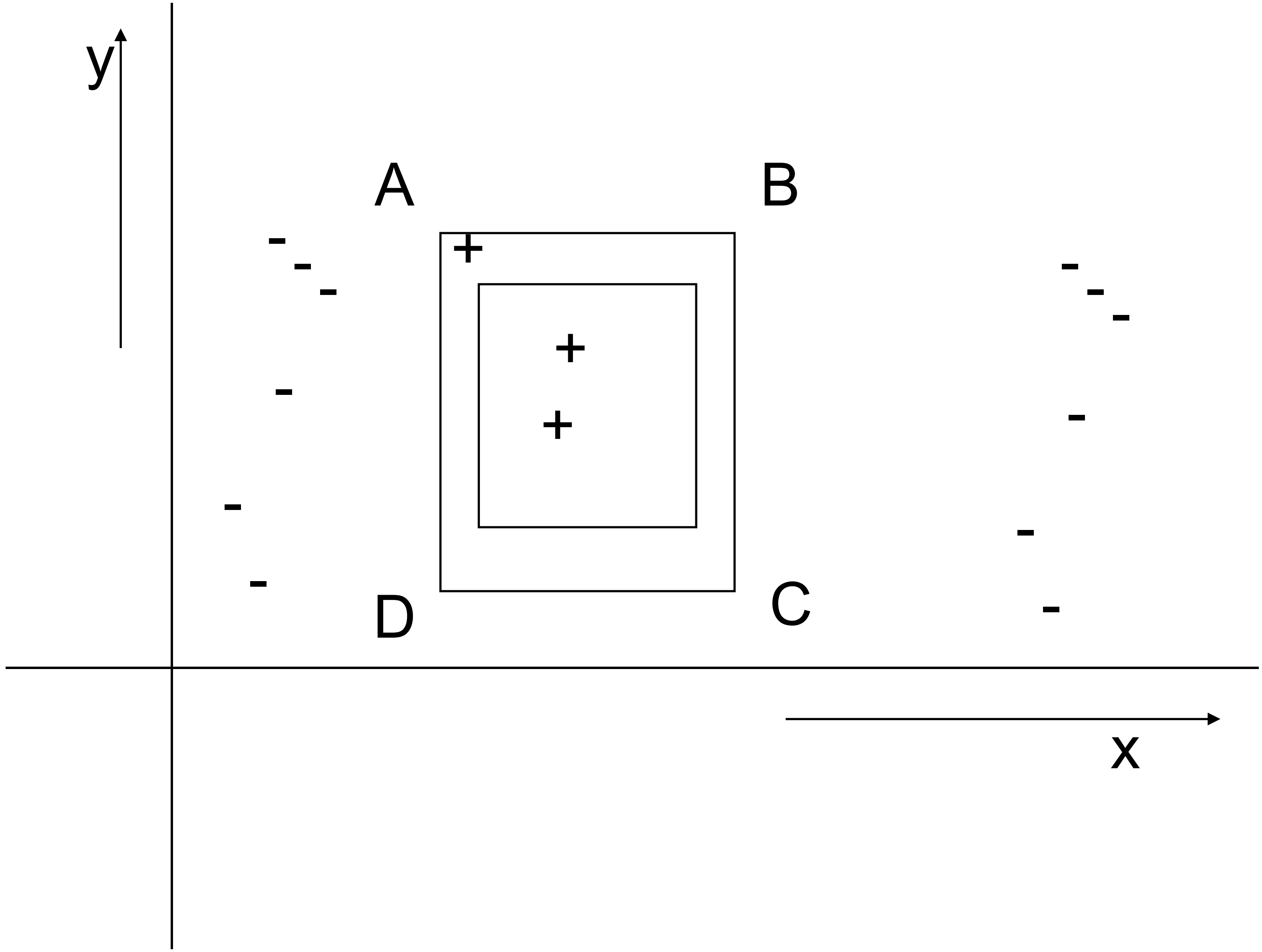
Universe:
2- Dimensional Plane



Key insights from 40 years of machine Learning Research:

1) What is it that is being learnt , and how the hypothesis should be produced ? This is a "MUST". This is called Inductive Bias .

2) "Learning in the Vacuum" is not possible. A learner already has crucial given pieces of knowledge at its disposal.

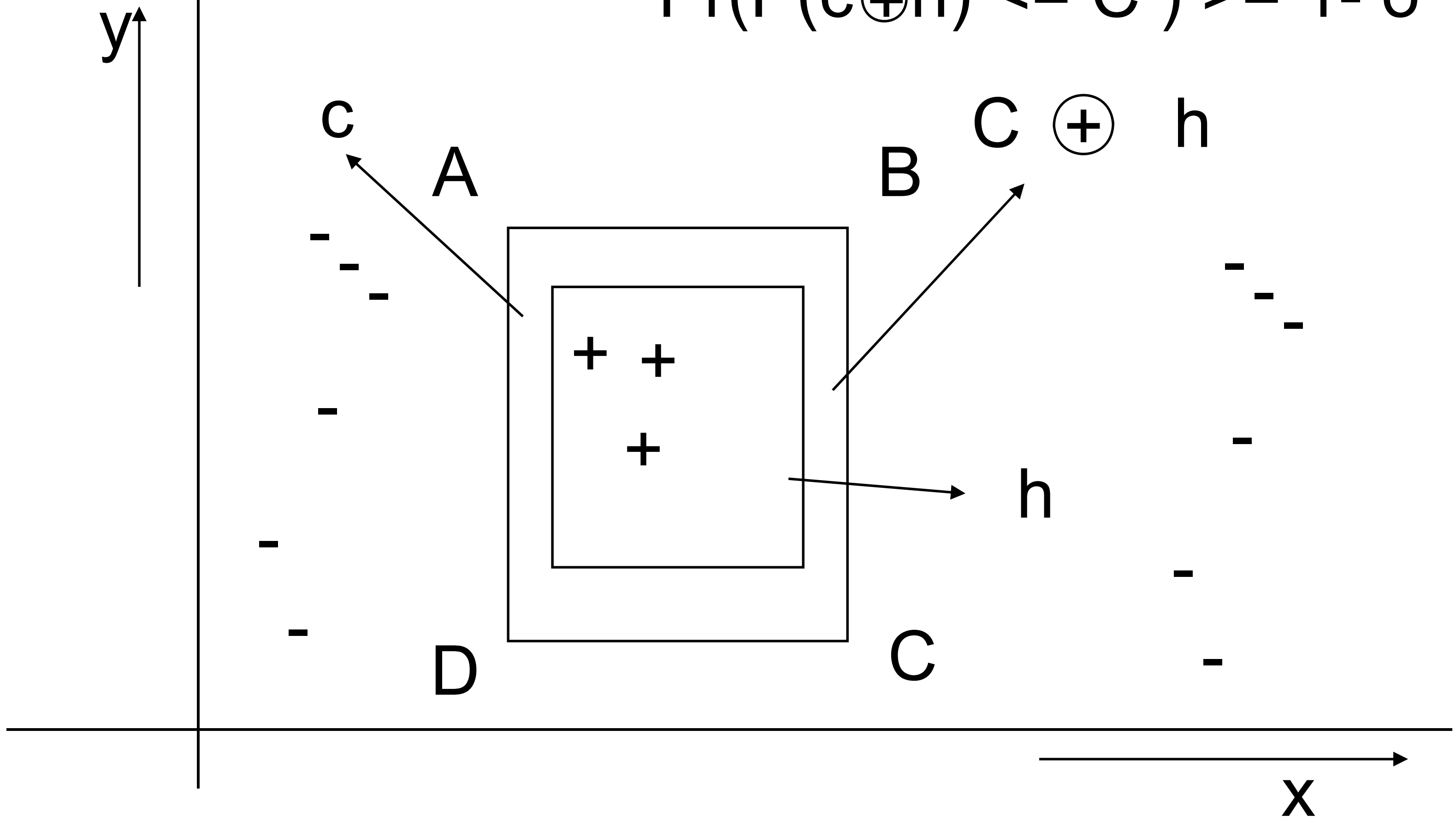

y
A
B
D
C
x

Algo:

1. Ignore –ve example.

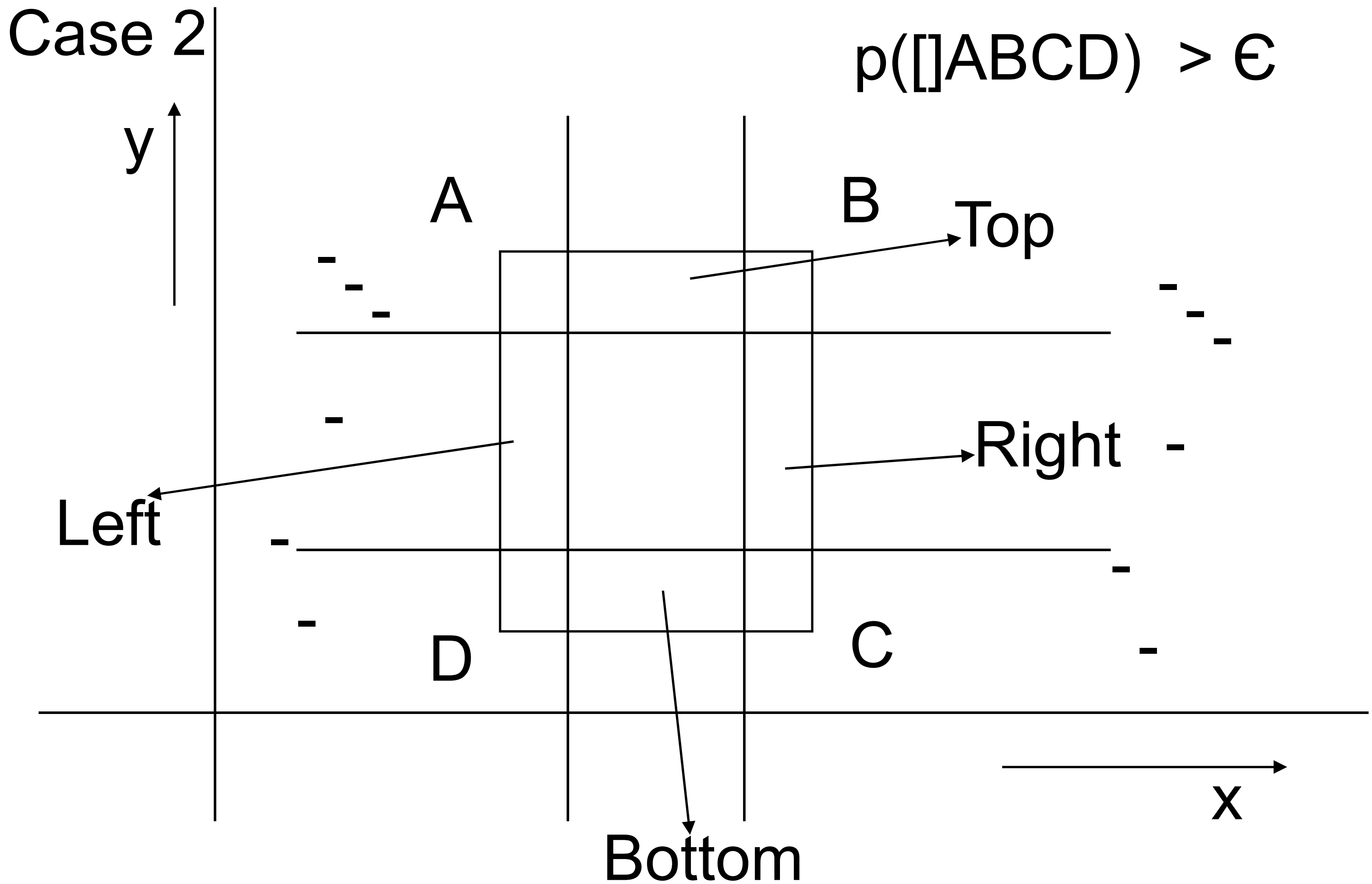
2. Find the closest fitting axis parallel rectangle for the data.

$Pr(P(c \oplus h) <= Є\ ) >= 1- \delta$



Case 1: If P([]ABCD) < Є

than the Algo is PAC.



P(Top) = P(Bottom) = P(Right) = P(Left) = Є /4

Let # of examples = m.

- Probability that a point comes from top = $\epsilon/4$

- Probability that none of the m example come from top = $(1 - \epsilon/4)^m$

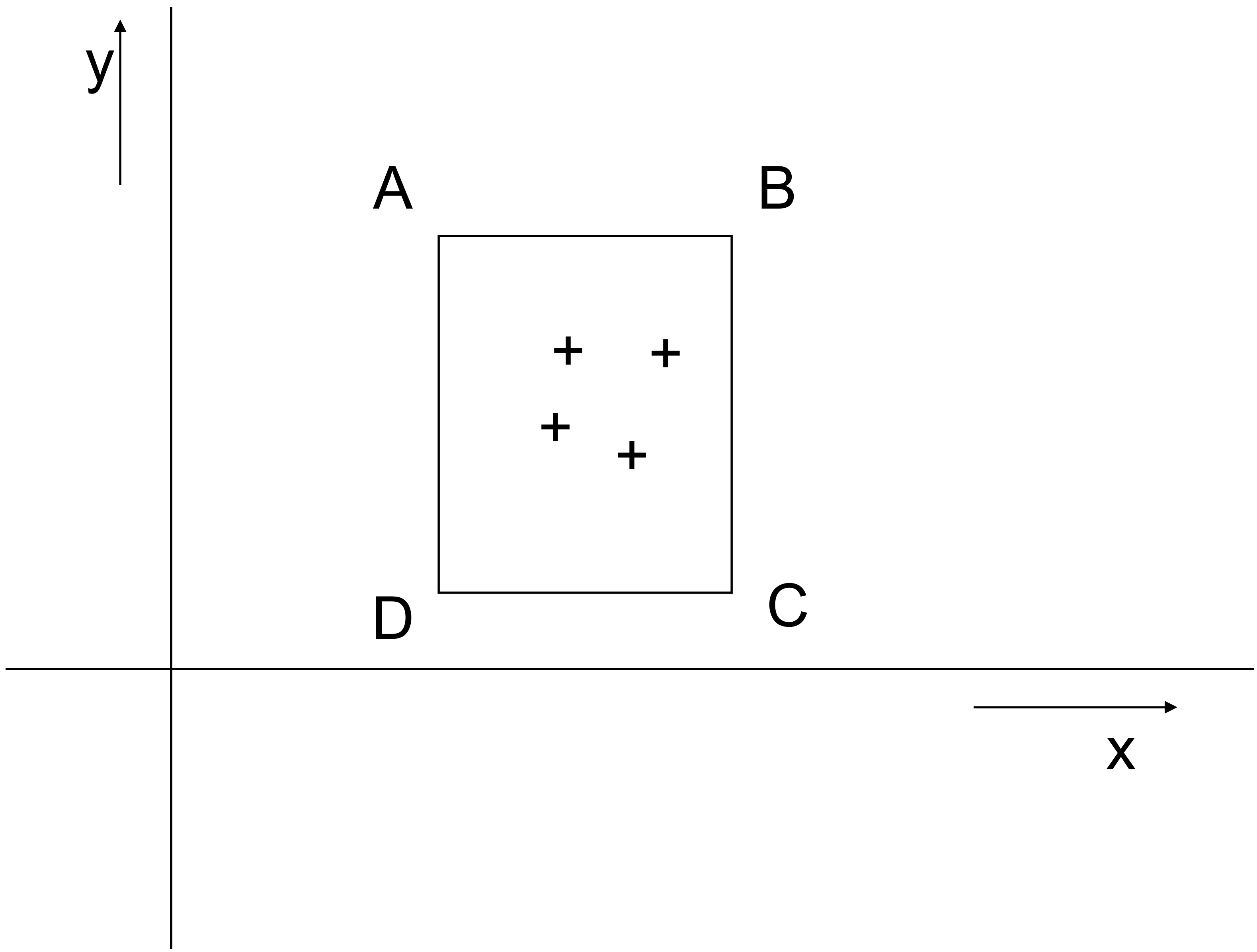
Probability that none of m examples come from one of top/bottom/left/right = $4(1 - \epsilon/4)^m$

Probability that at least one example will come from the 4 regions = $1- 4(1 - \epsilon/4)^m$

This fact must have probability greater than or equal to 1- δ

$$1-4\,(1 - \epsilon/4)^m > 1-\delta$$

$$\text{or}\quad 4(1 - \epsilon/4)^m < \delta$$



y
A
B
D
C
x

$$(1 - \epsilon/4)^m < e^{(-\epsilon m/4)}$$

We must have

$$4\, e^{(-\epsilon m/4)} < \delta$$

Or $m > (4/\epsilon) \ln(4/\delta)$

Lets say we want 10% error with 90% confidence

$$M > ((4/0.1) \ln (4/0.1))$$

Which is nearly equal to 200

Criticism against PAC learning

1. The model produces too many –ve results.

2. The Constrain of arbitrary probability distribution is too restrictive.

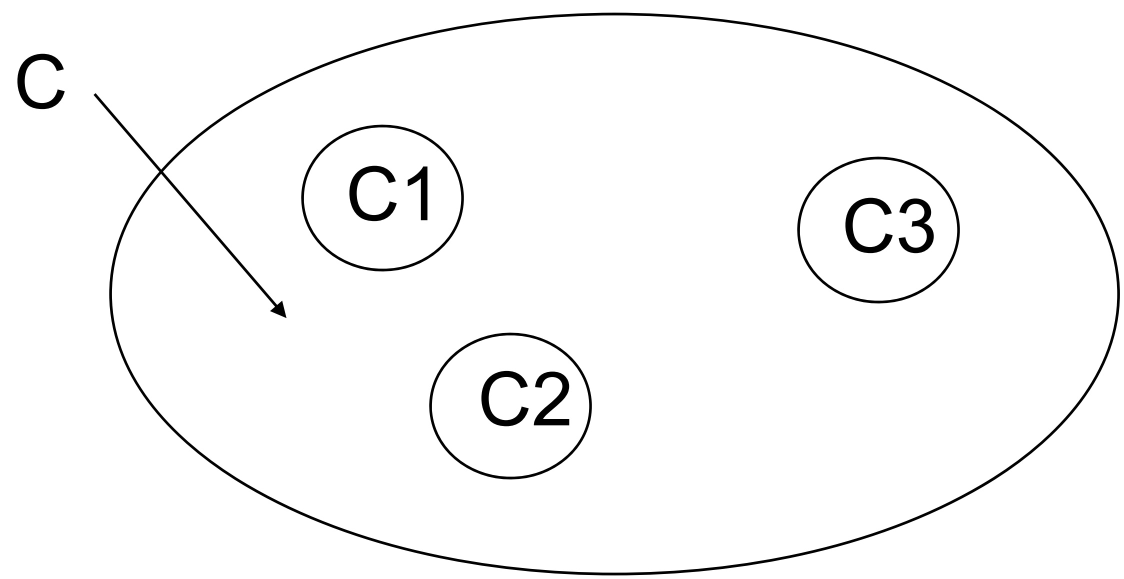
In spite of –ve results, so much learning takes place around us.

VC-dimension

Gives a necessary and sufficient condition for PAC learnability.

Def:-

Let C be a concept class, i.e., it has members c1,c2,c3,…… as concepts in it.



Let S be a subset of U (universe).

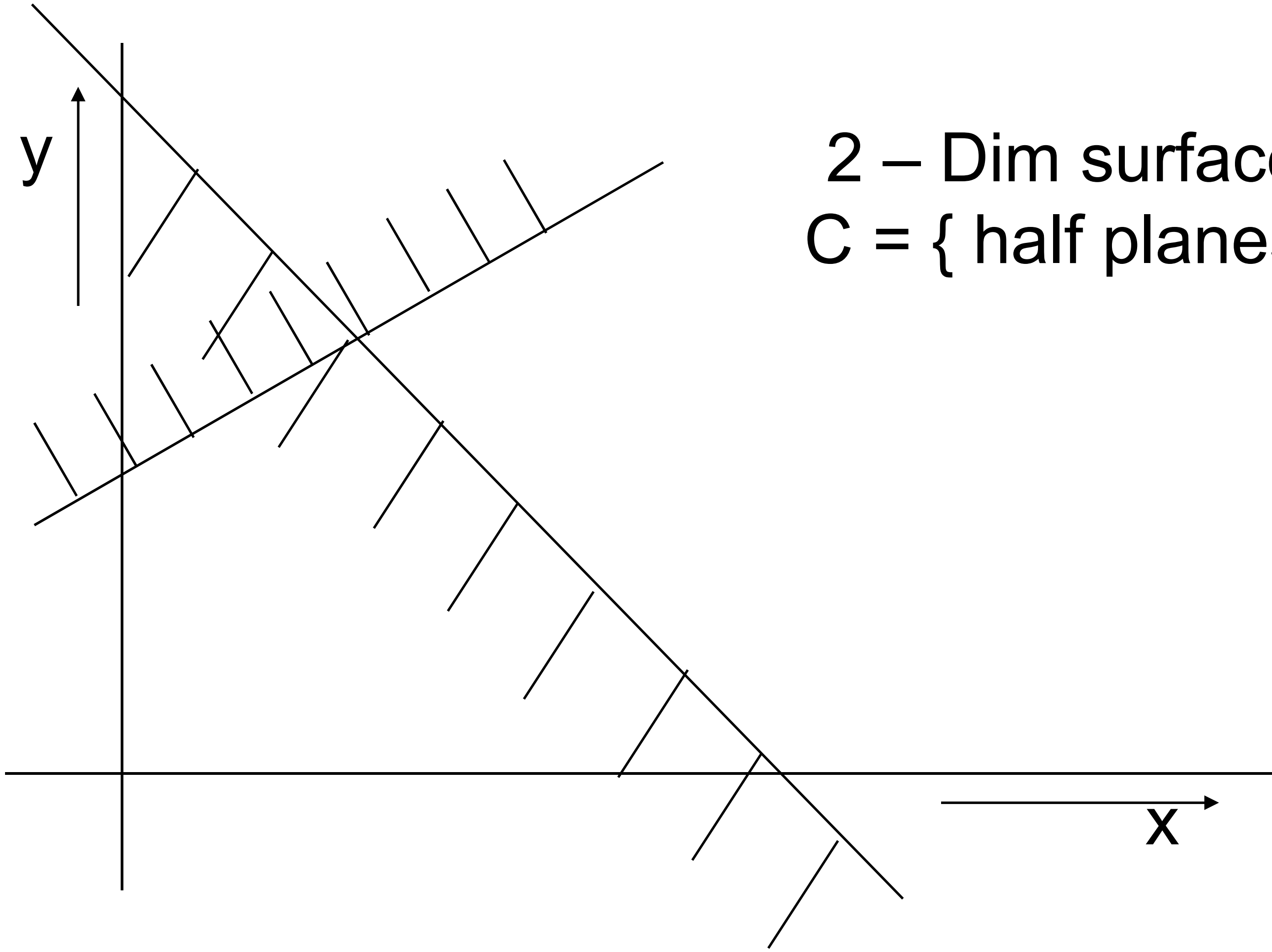
Now if all the subsets of S can be produced by intersecting with C_i^s, then we say C shatters S.

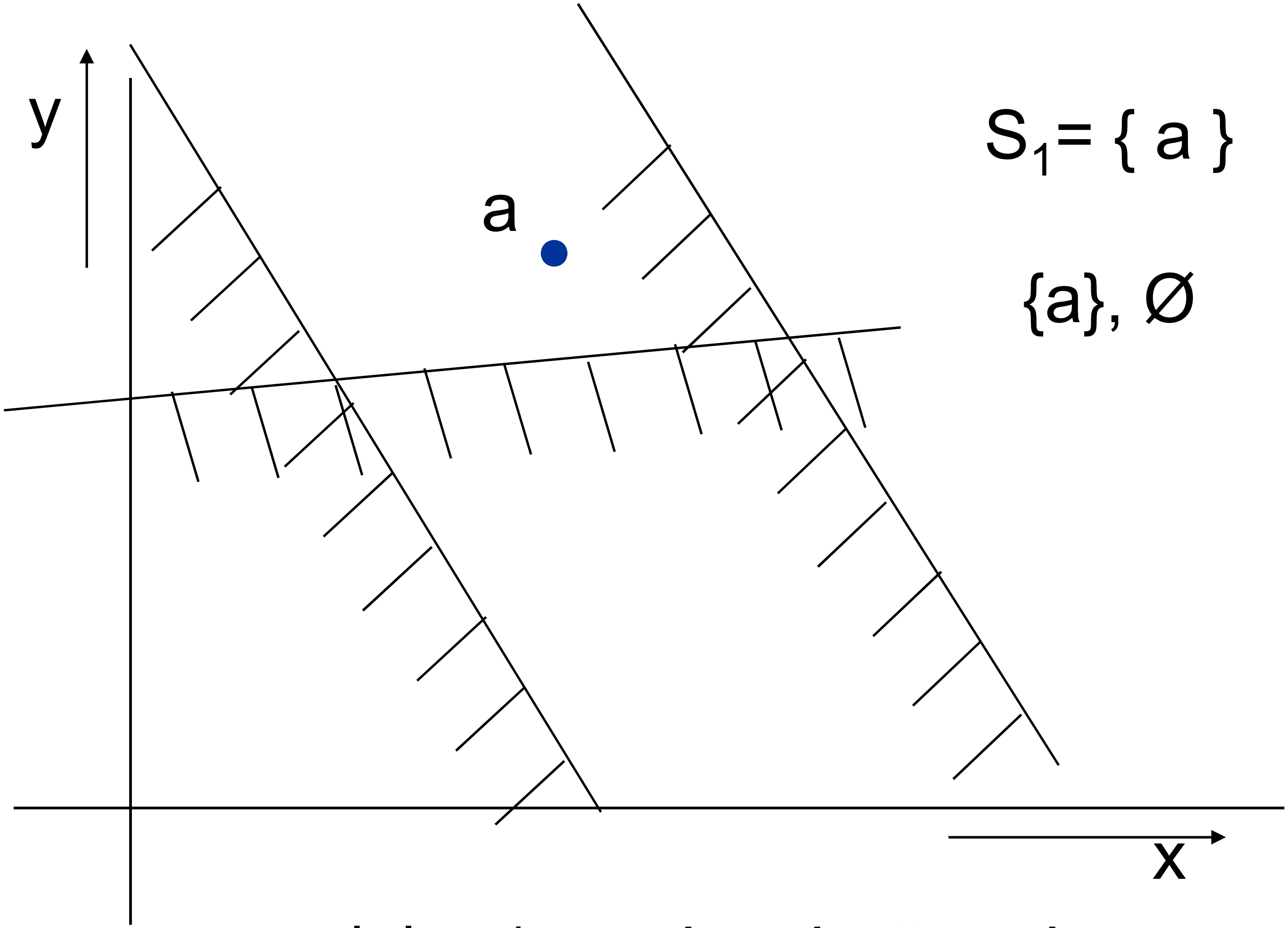
The highest cardinality set S that can be shattered gives the VC-dimension of C.

VC-dim(C)= |S|

VC-dim: Vapnik-Cherronenkis dimension.

2 – Dim surface
C = { half planes}

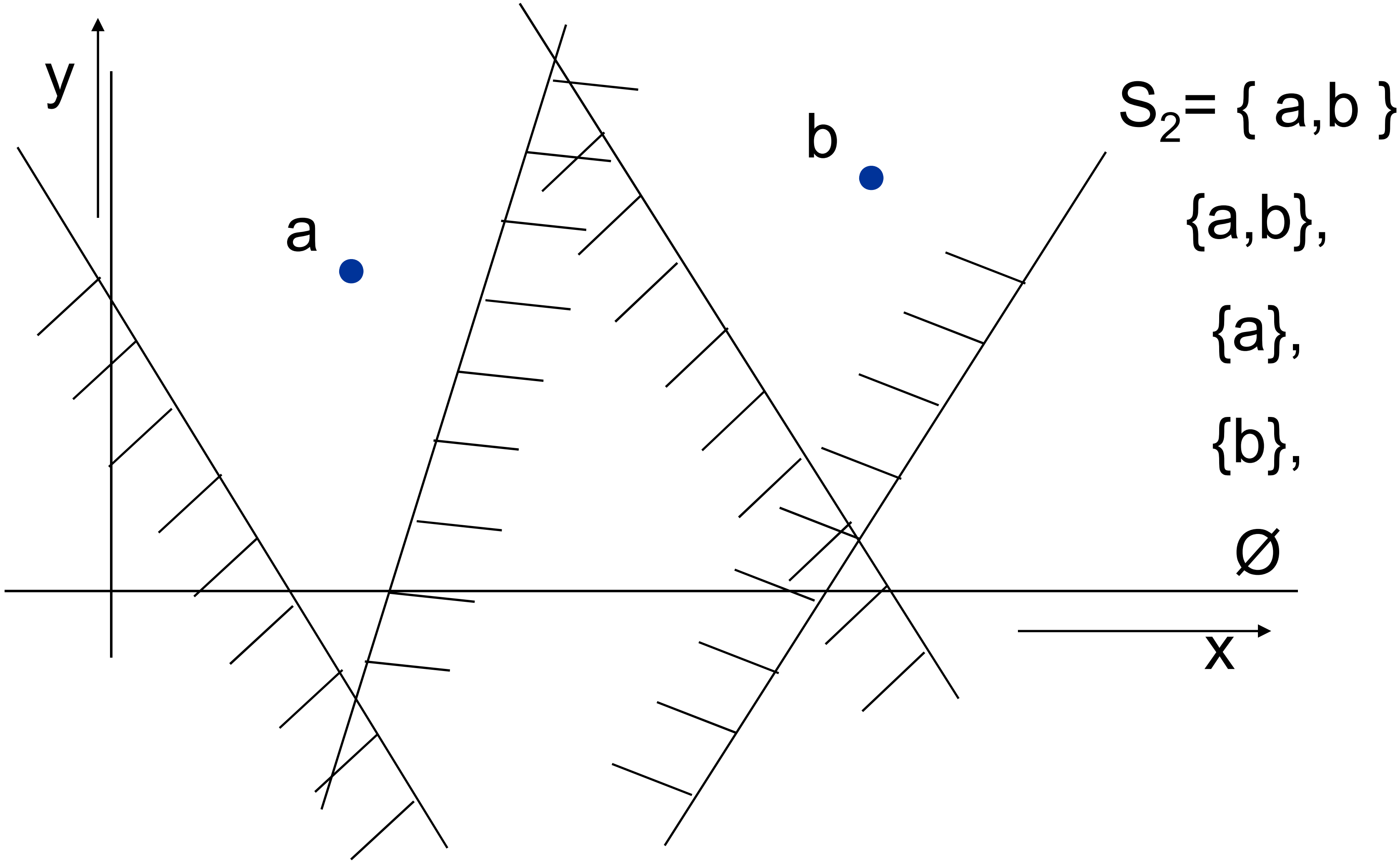




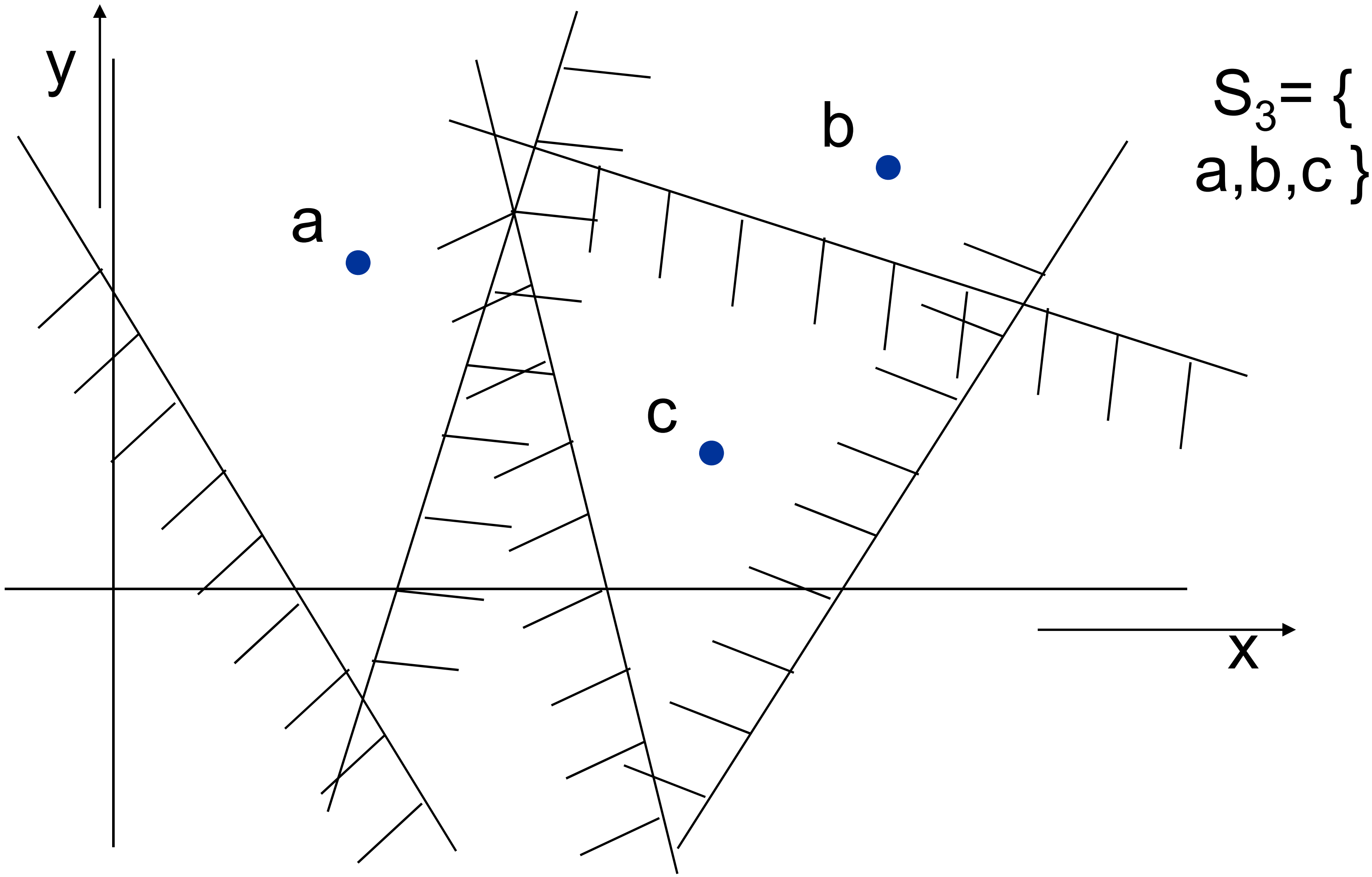
$S_1 = \{ a \}$

$\{a\}, \emptyset$

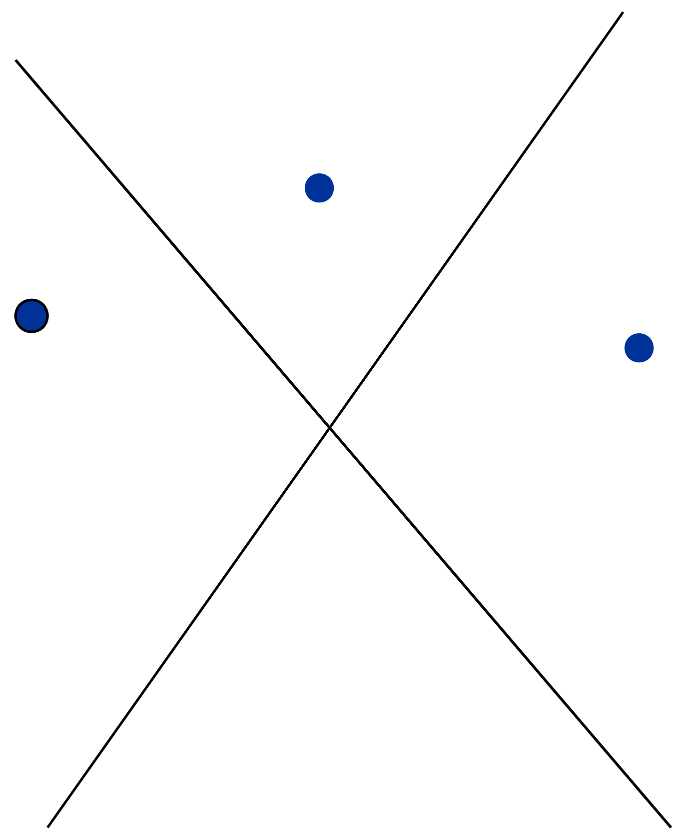
|s| = 1 can be shattered



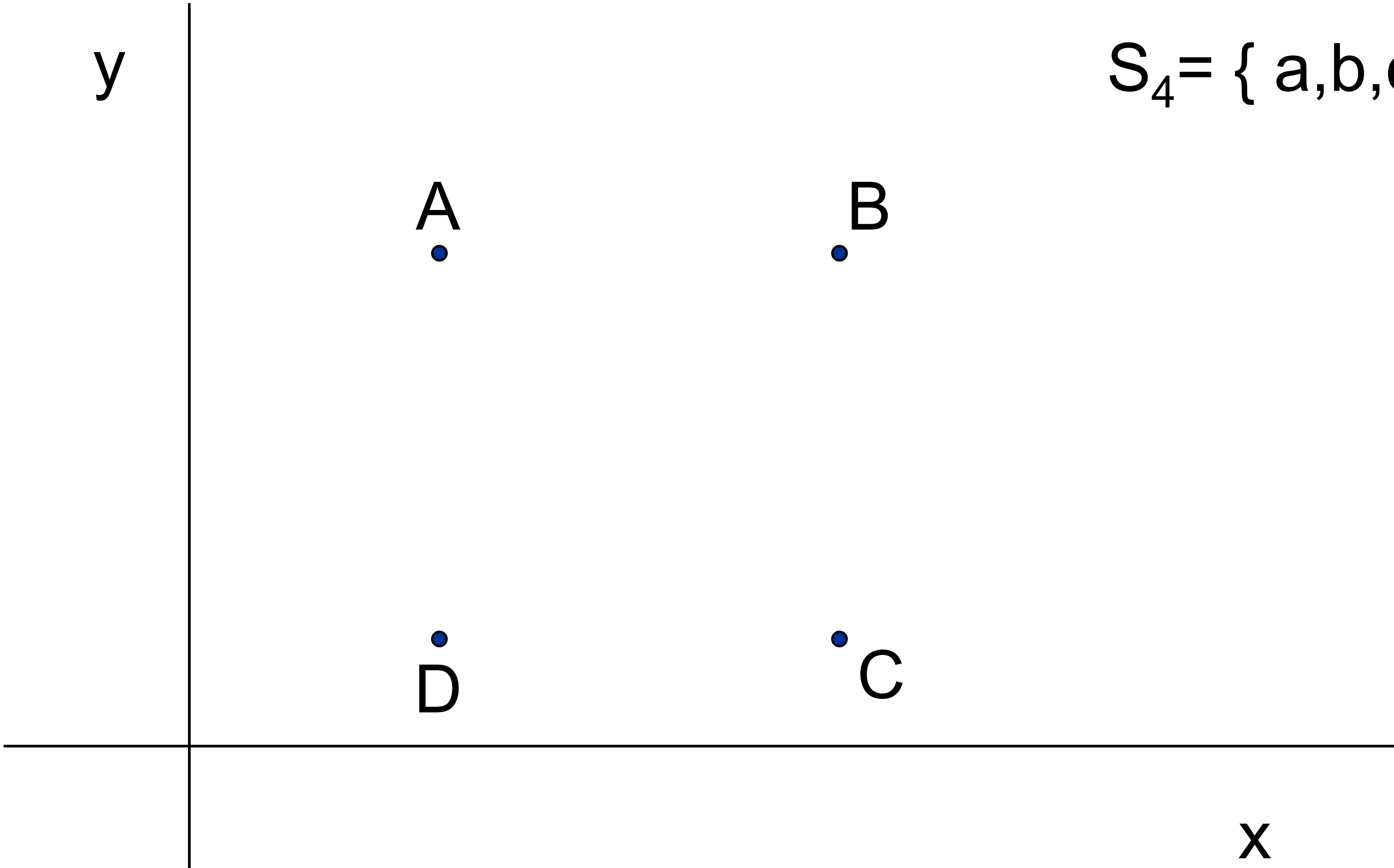
|s| = 2 can be shattered



|s| = 3 can be shattered



$S_4 = \{ a,b,c,d \}$



$|s| = 4$ cannot be shattered

Fundamental Theorem of PAC learning *(Ehrenfeuct et. al, 1989)*

- A Concept Class *C* is learnable for all probability distributions and all concepts in *C* if and only if the VC dimension of *C* is finite
- If the VC dimension of *C* is *d*, then…(next page)

Fundamental theorem (contd)

(a) for $0<\varepsilon<1$ and the sample size at least

$$\max[(4/\varepsilon)\log(2/\delta),\ (8d/\varepsilon)\log(13/\varepsilon)]$$

any consistent function $A{:}S_c \rightarrow C$ is a learning function for C

(b) for $0<\varepsilon<1/2$ and sample size less than

$$\max[((1-\varepsilon)/\ \varepsilon)\ln(1/\ \delta),\ d(1-2(\varepsilon(1-\ \delta)+\ \delta))]$$

No function $A{:}S_c \rightarrow H$, for any hypothesis space is a learning function for C.

Book

1. Computational Learning Theory, M. H. G. Anthony, N. Biggs, Cambridge Tracts in Theoretical Computer Science, 1997.

Paper's

1. A theory of the learnable, Valiant, LG (1984), Communications of the ACM 27(11):1134 -1142.

2. Learnability and the VC-dimension, A Blumer, A Ehrenfeucht, D Haussler, M Warmuth - Journal of the ACM, 1989.