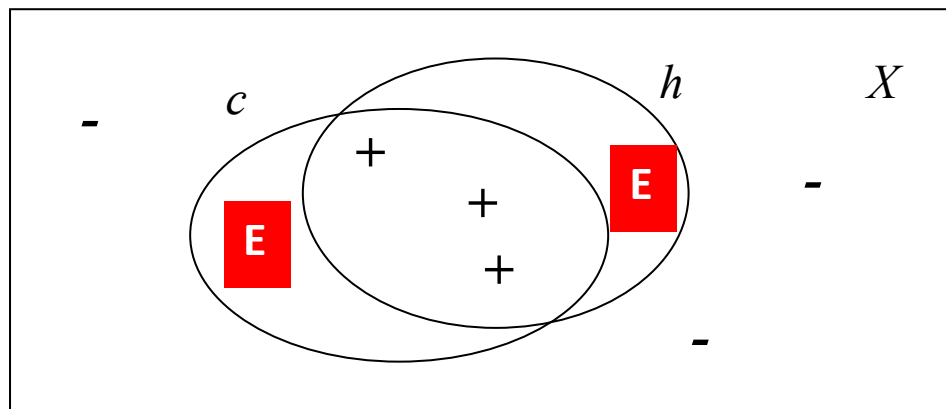


Learning Theory

Sample Complexity

Examples

Ariel Shamir
Zohar Yakhini



Outline

כנראה שכדאי לראות את ההרצאה המתאימה למצגת הזו (לשים לב שההרצאה של המצגת הזו ולא המצגת הקודמת), אם רוצה להבין טוב את הנושא: הרצאות.

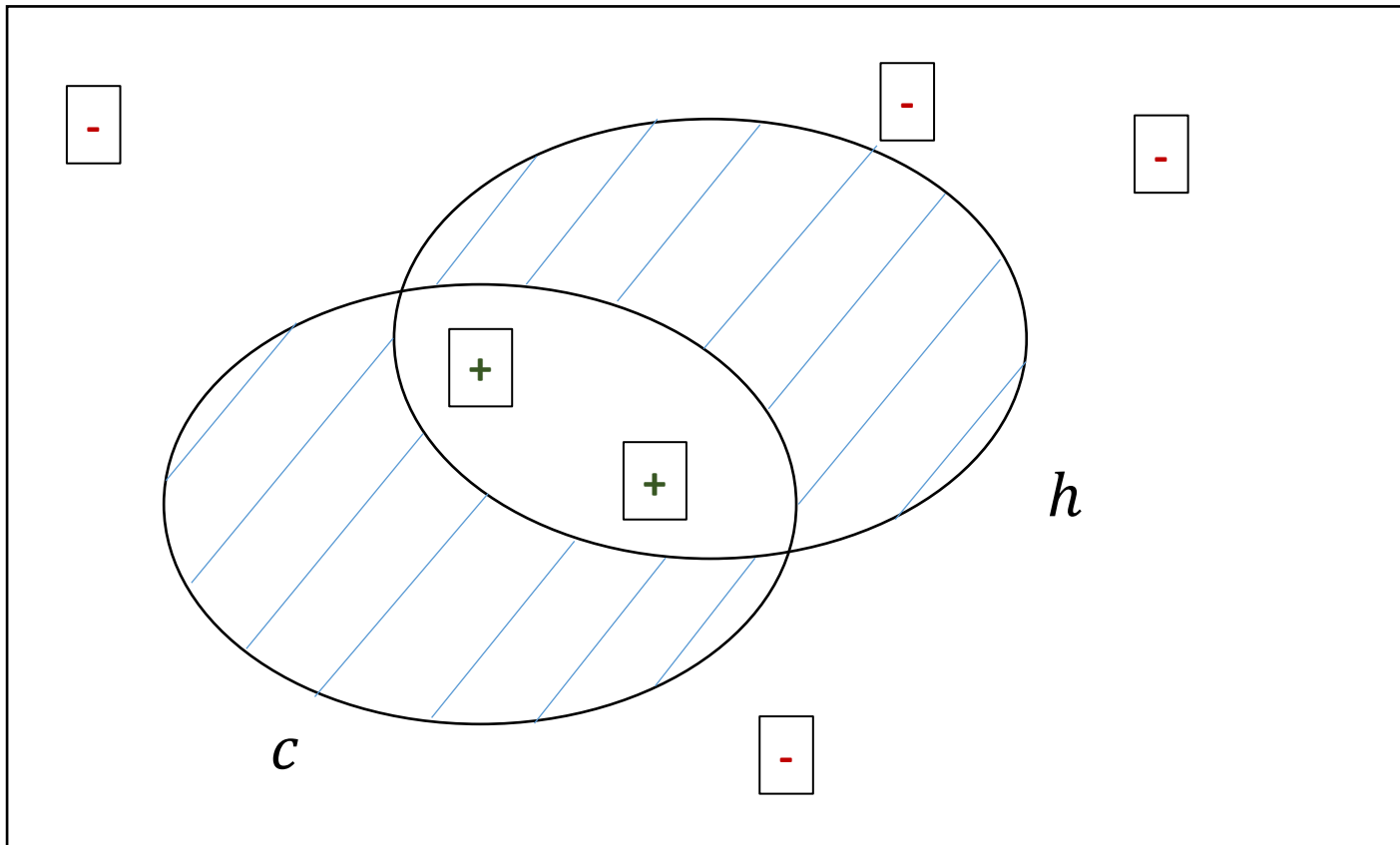
Machine Learning from Data H Sun 1730

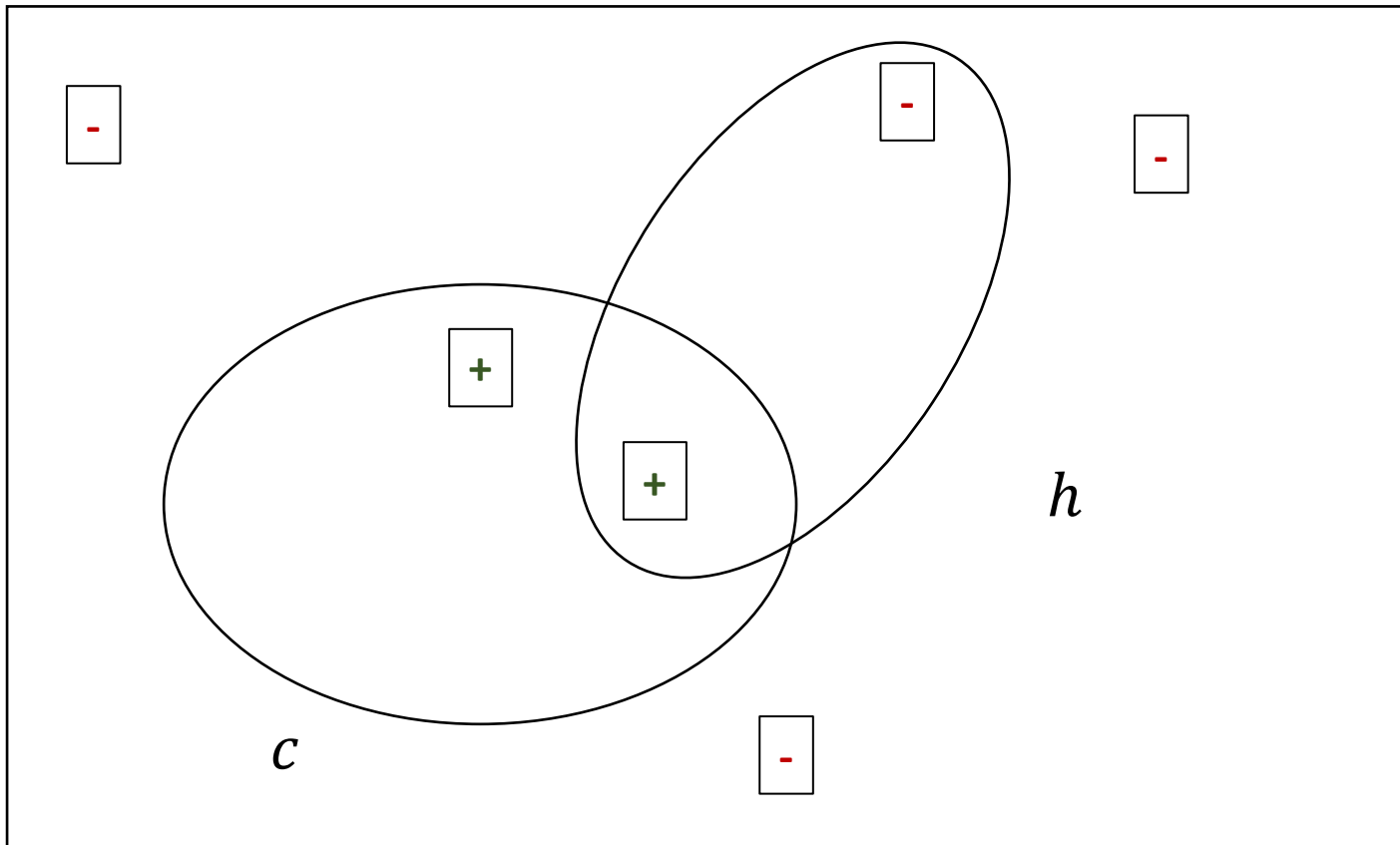
א: במצגת הקודמת למדנו concepts שהיו לינאריים, כלומר יש קו שמספיד בין הלייבלים. כאן מה שמפסיד זה ריבוע שתוחם את כל החיוביים. לכן יש למצוא נוסחה חדשה לכמות הסאמפלס הנדרשת על מנת להבטיח PAC

- The true error
- Consistent hypotheses, consistent learners
- Complexity of learning
- Sample complexity
- Bound on the sample complexity for finite H
- Examples:
 - Finite spaces of Boolean vectors
 - Circles, rectangles
- VC dimension

General setting

- Instances come from $\Omega = (X, Y, \pi)$
- The learning algorithm L takes training data $D \in \Omega^m$
- It works with some set of hypotheses, H
- It returns a hypothesis (or a model) $L(D) = h \in H$





Example from last class (disjuncts and Bool vectors)

$$m \geq \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} = \frac{1}{\varepsilon} \left(\ln |H| + \ln \frac{1}{\delta} \right)$$

- $n = 20$ attributes in the same setting
- We get $|H| = 3^{20} \sim 3.5 * 10^9$
- In this case, we can obtain 95% certainty that our hypothesis will have error $< 10\%$ when using

$$m > \frac{1}{0.1} (\ln 3.5 * 10^9 + \ln \frac{1}{0.05}) = 10(22+3) = 250 \text{ instances}$$

- Note that here $|X| = 2^{20} \approx 10^6$



When $|H|$ increases exponentially with the number of features then sample complexity increases linearly.
The required fraction of the full population decreases.

PAC Learnability

- Consider a class \mathbf{C} of possible target concepts defined over a space of instances \mathbf{X} , and a learning algorithm \mathbf{L} using hypothesis space \mathbf{H} .

הגדרה שלא באמת חייב להבין... (זוהר אמר)

- Definition

\mathbf{C} is PAC-learnable by \mathbf{L} using \mathbf{H}

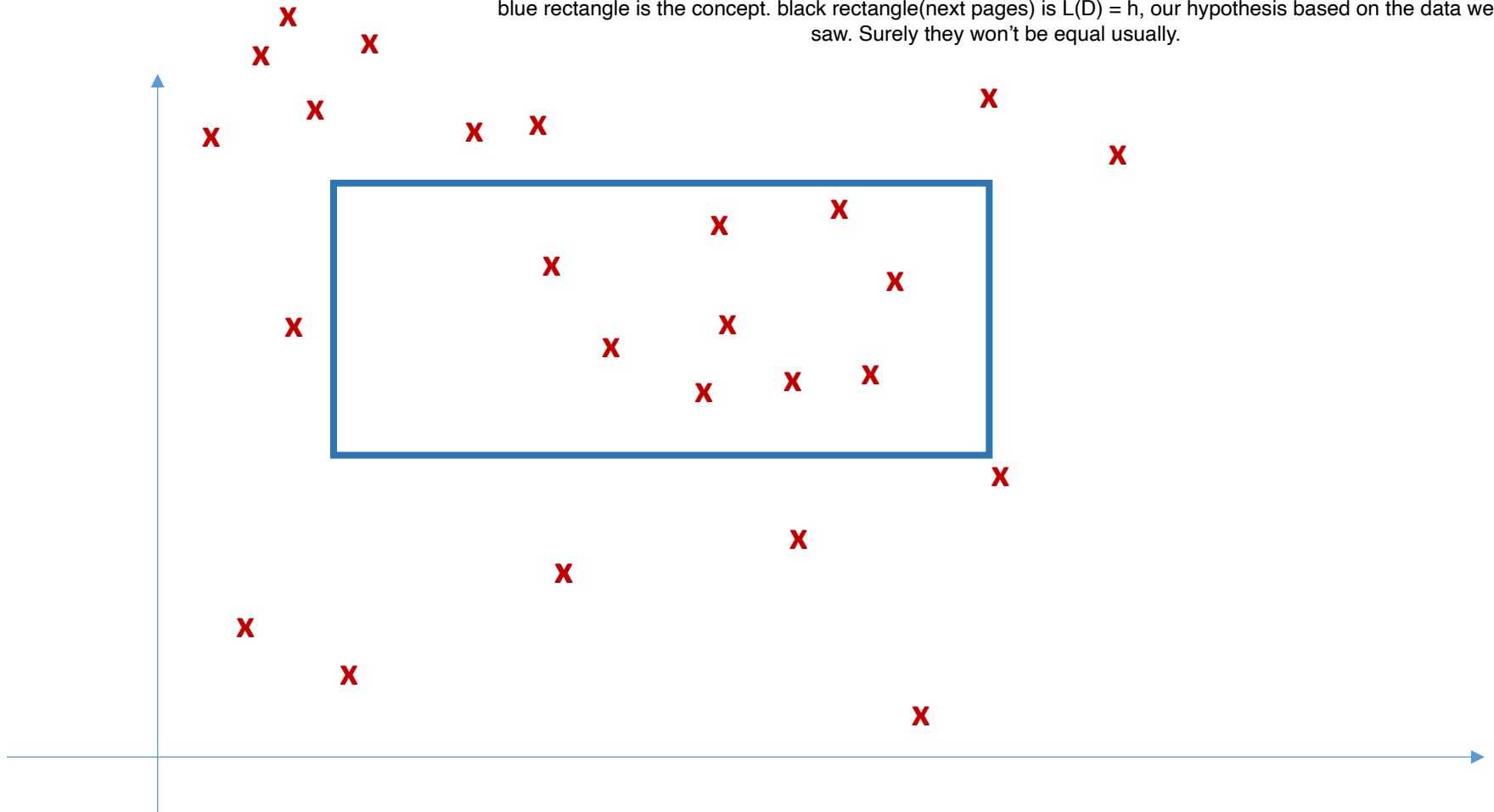
if for all $0 < \varepsilon < \frac{1}{2}$, $0 < \delta < \frac{1}{2}$, and for all $c \in \mathbf{C}$ and distributions π over \mathbf{X} , the following holds:

with data drawn independently according to π , \mathbf{L} will output, with probability at least $(1-\delta)$, a hypothesis $h \in \mathbf{H}$ such that $\text{error}_{\pi}(h) \leq \varepsilon$,

\mathbf{L} operates in time (and sample) complexity that is polynomial in $1/\varepsilon$, $1/\delta$ (and in other possible parameters).

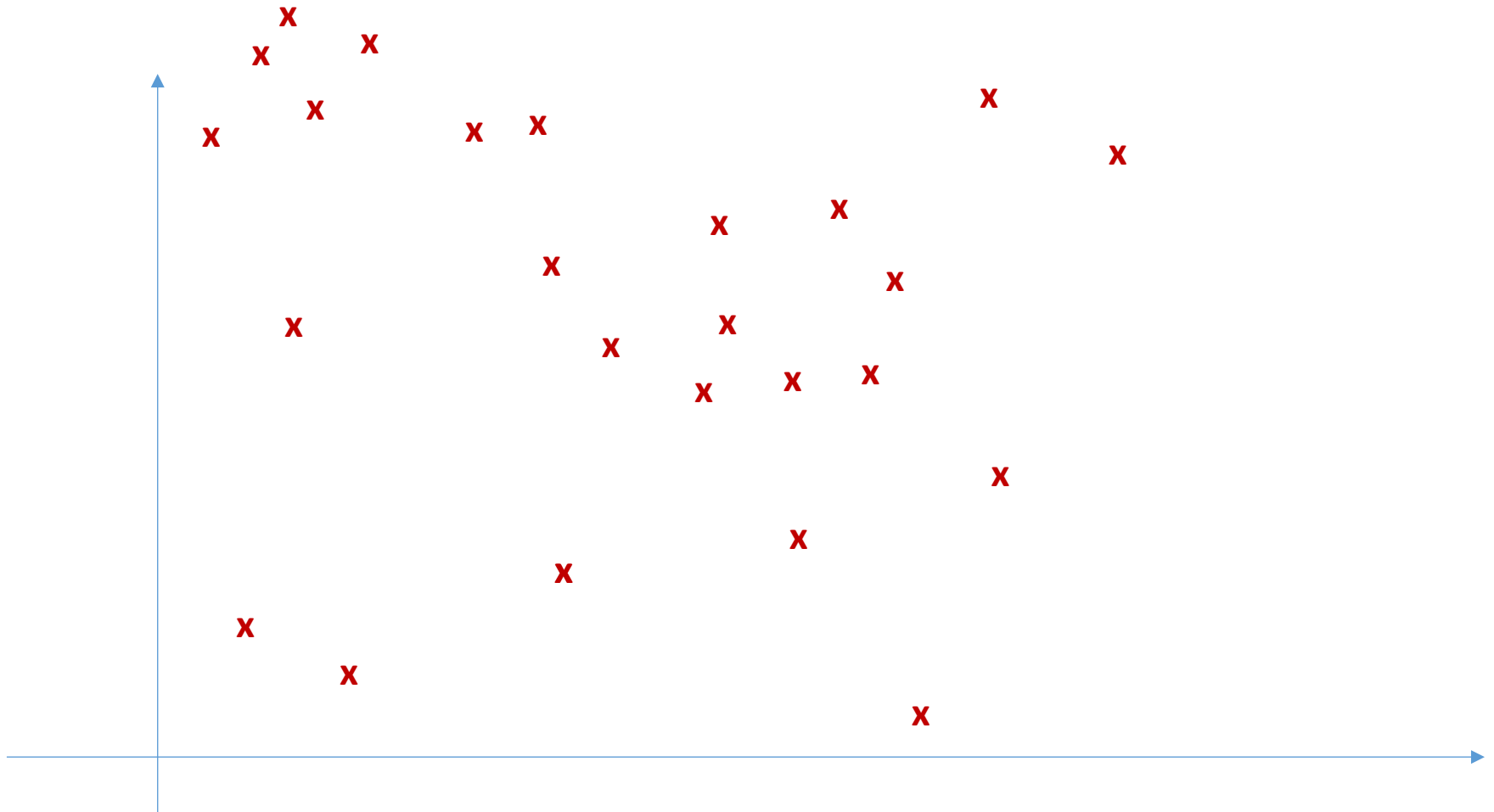
Learning axes aligned rectangles

blue rectangle is the concept. black rectangle(next pages) is $L(D) = h$, our hypothesis based on the data we saw. Surely they won't be equal usually.



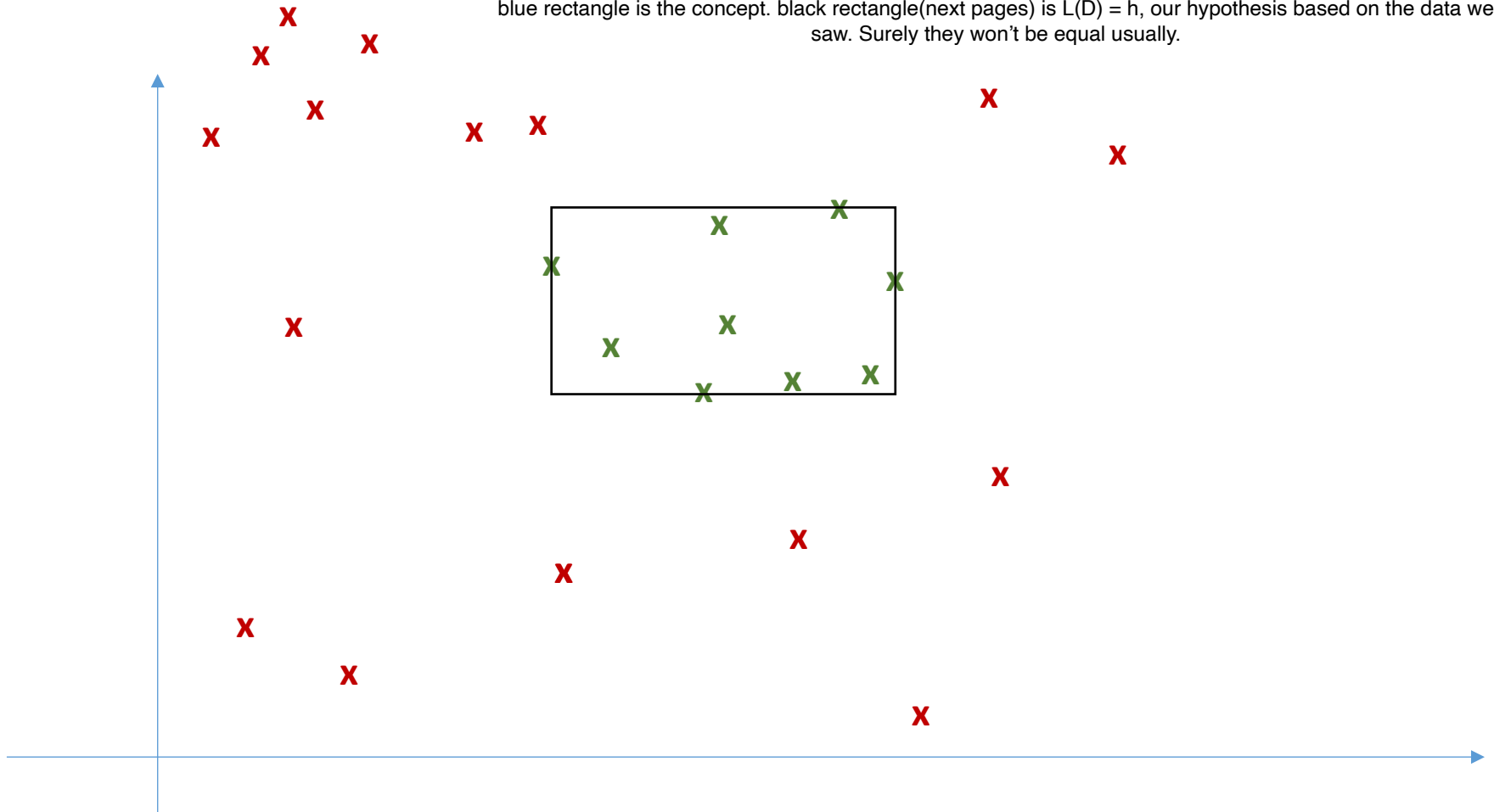
$$X = \mathbb{R}^2$$

Learning axes aligned rectangles



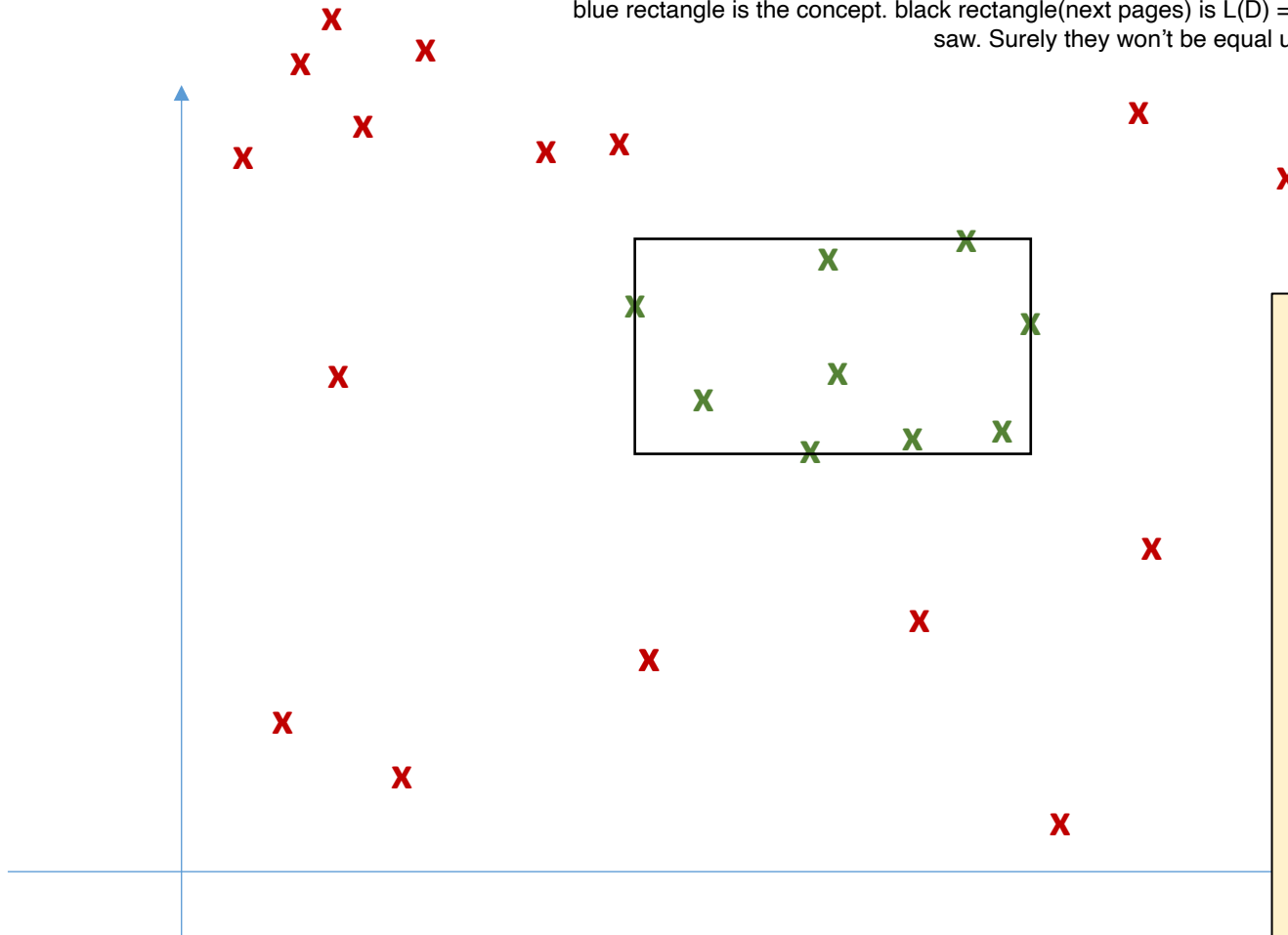
Learning axes aligned rectangles

blue rectangle is the concept. black rectangle(next pages) is $L(D) = h$, our hypothesis based on the data we saw. Surely they won't be equal usually.



Learning axes aligned rectangles

blue rectangle is the concept. black rectangle(next pages) is $L(D) = h$, our hypothesis based on the data we saw. Surely they won't be equal usually.



Consistent learner:

Find points with
max and min x,
max and min y.
Draw edges accordingly.
Linear in m.

Returns $h = L(D)$ such
that

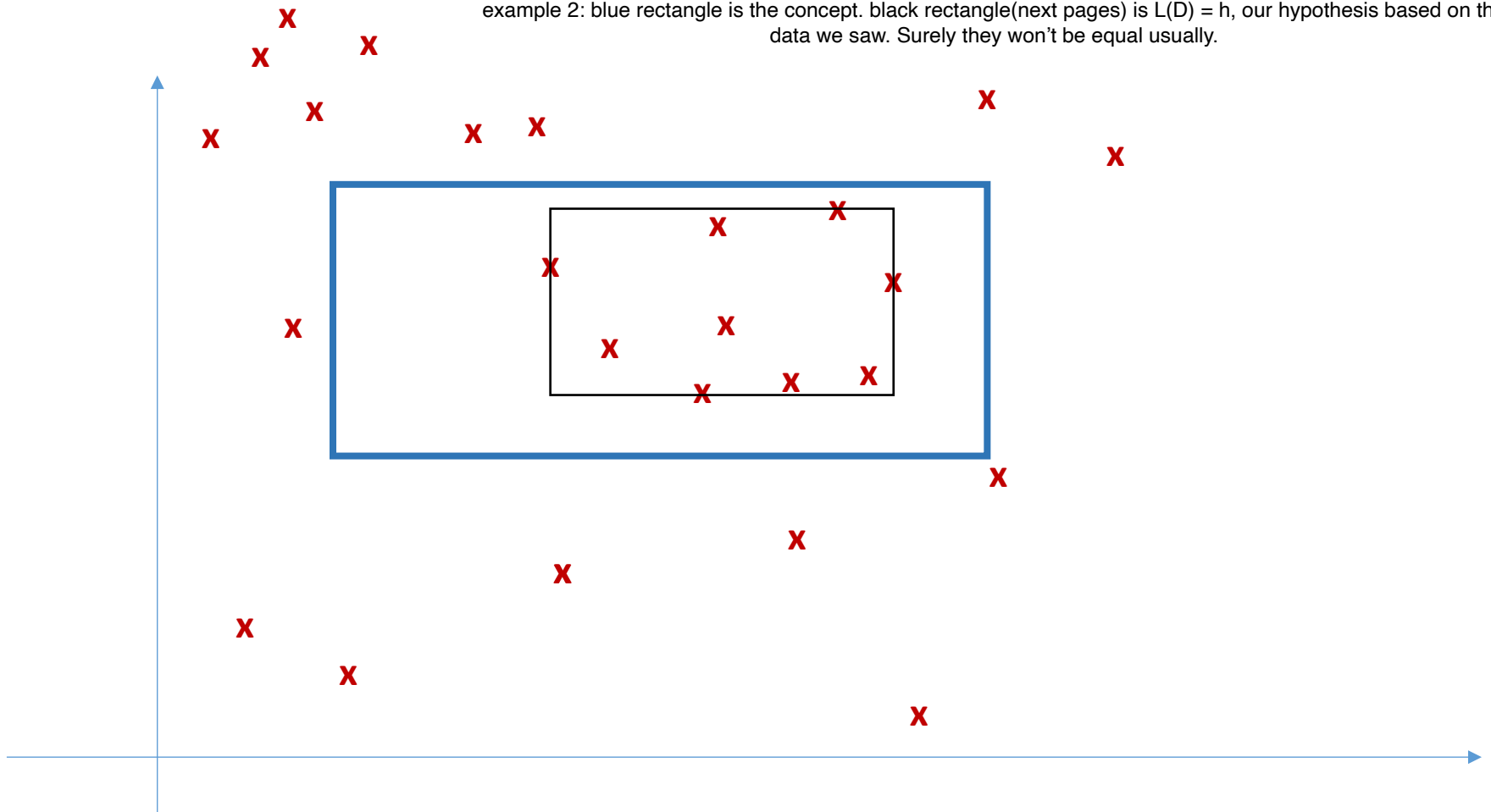
$\forall x \in X$

$h(x) = 1 \implies c(x) = 1$
but not necessarily the
opposite.

How about on D ?

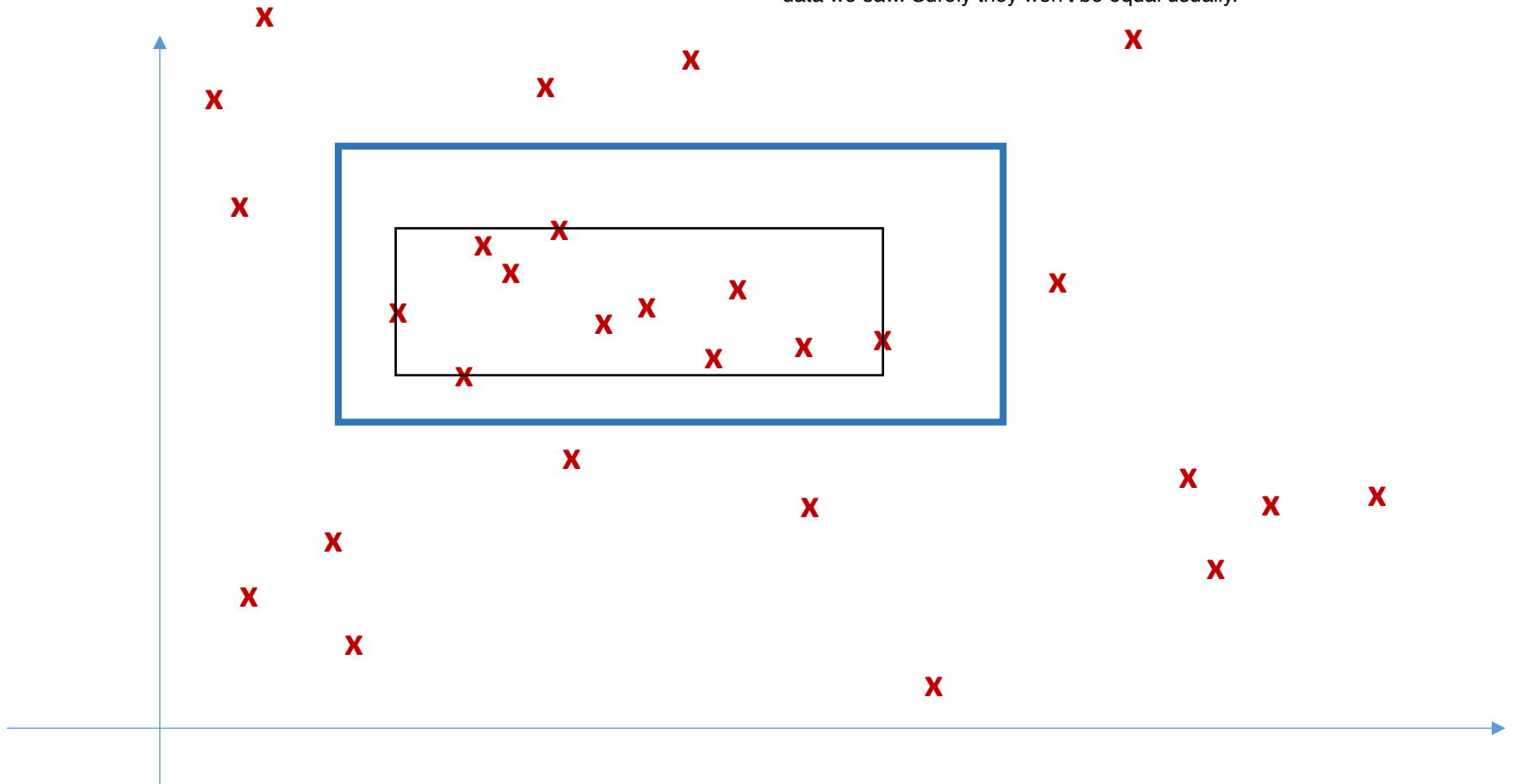
Learning axes aligned rectangles

example 2: blue rectangle is the concept. black rectangle(next pages) is $L(D) = h$, our hypothesis based on the data we saw. Surely they won't be equal usually.



Learning axes aligned rectangles

example 3: blue rectangle is the concept. black rectangle(next pages) is $L(D) = h$, our hypothesis based on the data we saw. Surely they won't be equal usually.

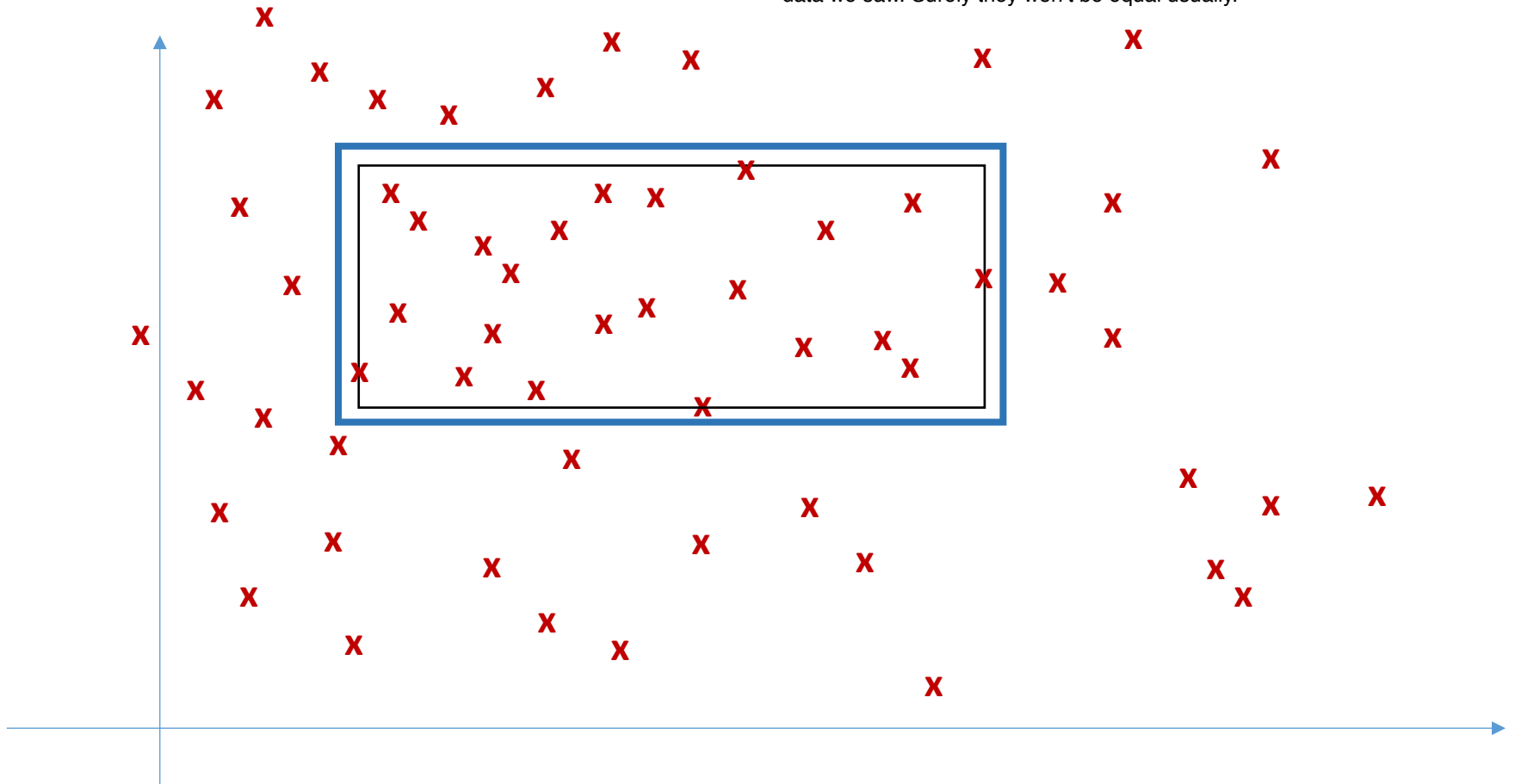


Different training dataset lead to a different output hypotheses.

$$h = L(D), D \in (X, P)^m$$

Learning axes aligned rectangles

example 4: blue rectangle is the concept. black rectangle(next pages) is $L(D) = h$, our hypothesis based on the data we saw. Surely they won't be equal usually.



Different training dataset lead to a different output hypotheses.

$$h = L(D), D \in (X, P)^m$$

A bound on sample complexity

- For every $c \in \mathcal{C}$
- We will now characterize (bound) the collection of all training datasets, D , that can conceivably lead to $h = L(D)$ with $\text{Err}(h, c) > \varepsilon$
- We will show that this collection (a subset of X^m) is contained in a union of a small number of sets (B_i , subsets of X^m) characterizable by regions that their elements (points in X^m) do not visit.

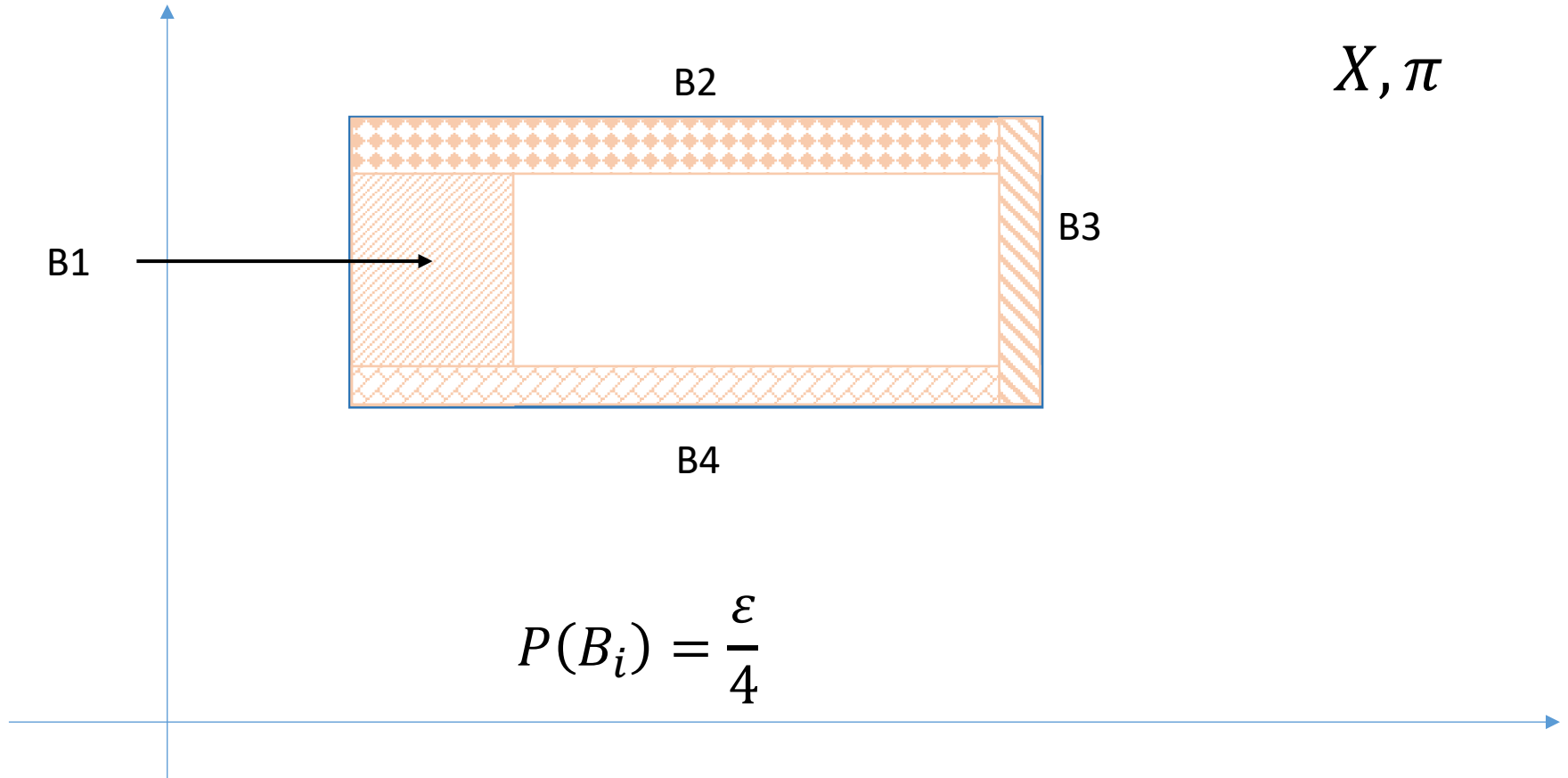
A bound on the sufficient sample complexity, cont

- We will then estimate the probability (π^m) of each such set of datasets as well as that of their union.
- From here we will infer an upper bound on sufficient sample complexity, as a function of ε and δ , of the form

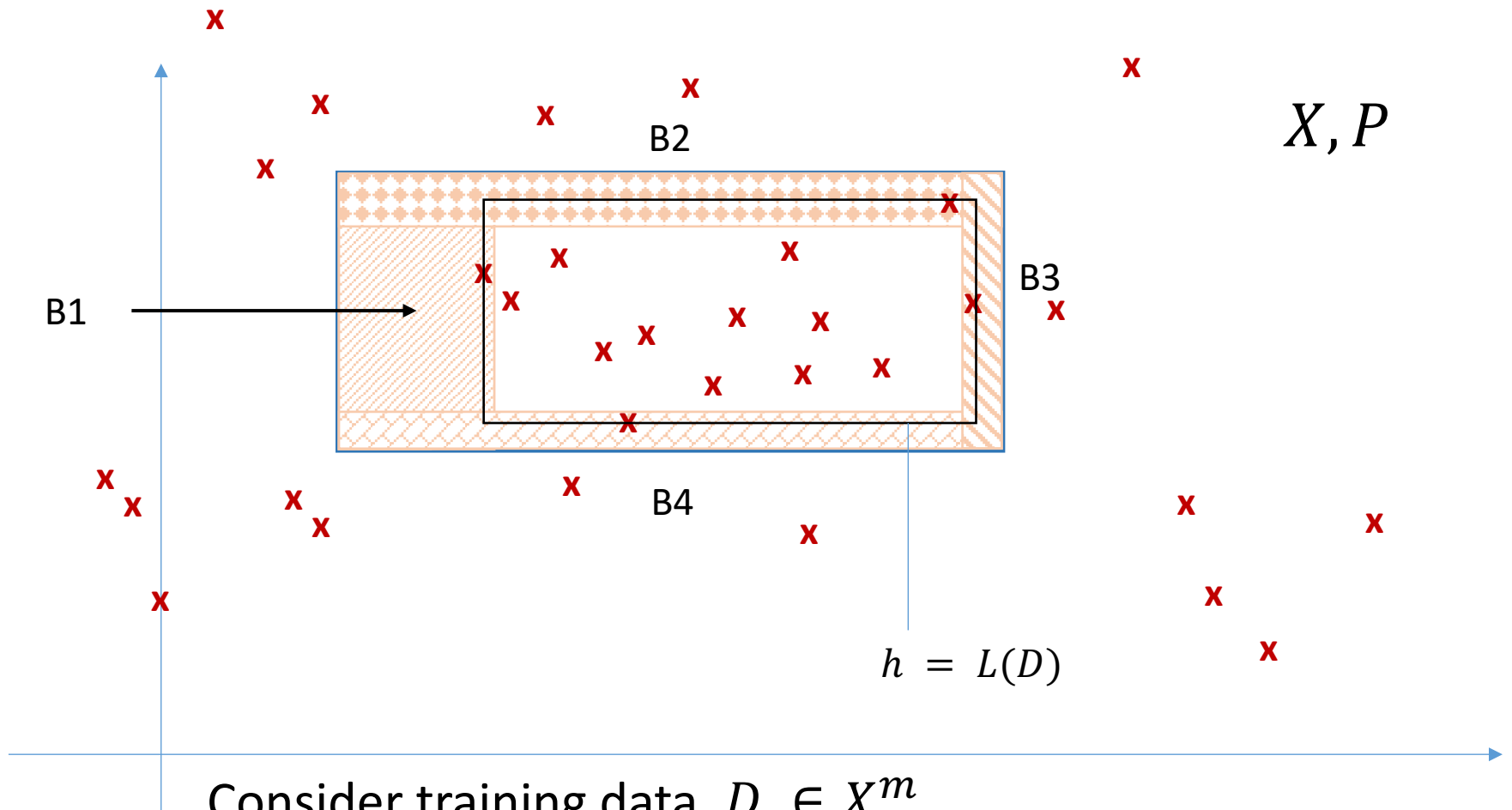
$$m(\varepsilon, \delta) \in O(f(\varepsilon), g(\delta))$$

Learning axes aligned rectangles

Machine Learning from Data H Sun 1730: הרצאה: אם רוצה להבין את זה, דקה 1:10 בערך (שעה ועשר דקות)



Learning axes aligned rectangles

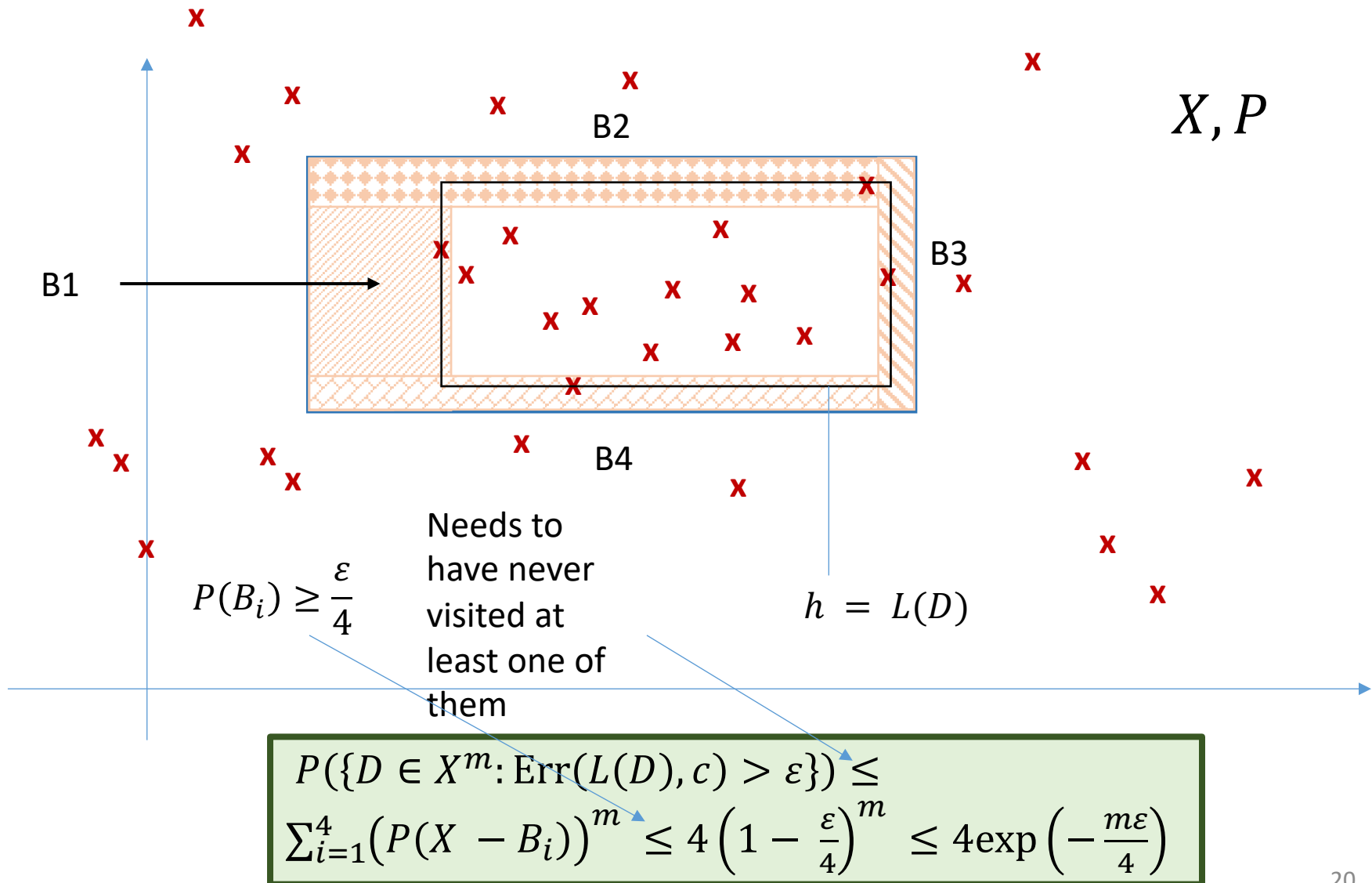


Consider training data, $D \in X^m$.

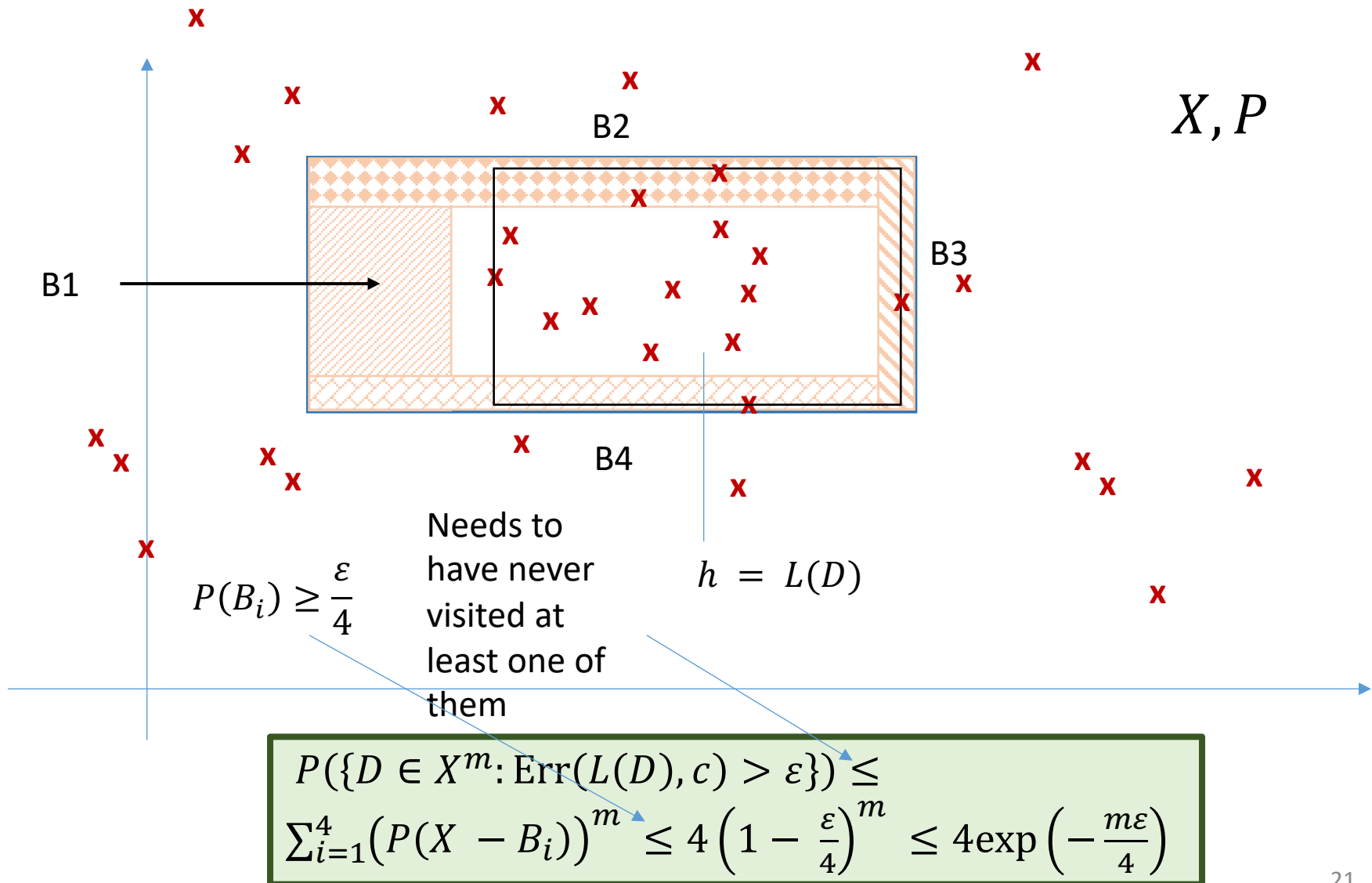
Assume that D visits each one of the 4 sets B_i defined above.

What can we say about $\text{Err}(h, c)$? $(P(B_i) \leq \frac{\epsilon}{4})$

Not visiting one of the regions B_i



Learning axes aligned rectangles



Sufficient sample size

$$m(\varepsilon, \delta) = \frac{4}{\varepsilon} \cdot \ln \frac{4}{\delta}$$

Summary

- Sample complexity
- Consistent learners for finite hypotheses spaces
- Directly calculating bounds on the sample complexity of consistent learners
- Concepts in \mathbb{R}^n - use the geometry!
- VC dimension - a more general formula
- In the recitation
 - More sample complexity examples
 - VC examples