Learning From Data
Lecture 5
Training Versus Testing

The Two Questions of Learning
Theory of Generalization ($E_{\text{in}} \approx E_{\text{out}}$)
An Effective Number of Hypotheses
A Combinatorial Puzzle

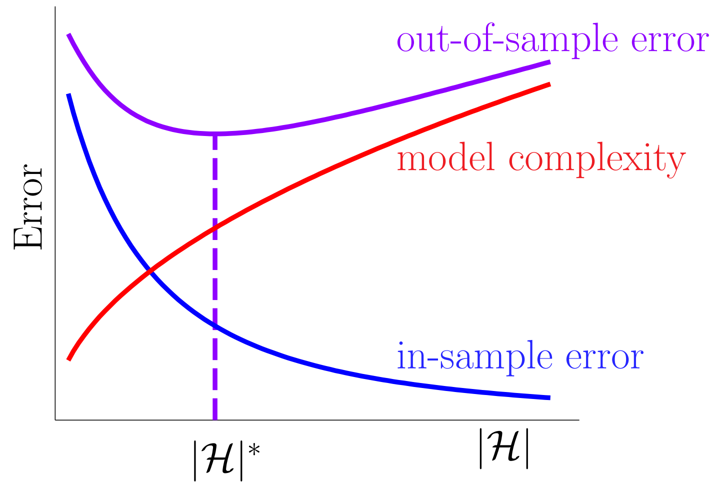
M. Magdon-Ismail
CSCI 4100/6100

RECAP: The Two Questions of Learning

1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?

The Hoeffding *generalization bound*:

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \underbrace{\sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}}_{\text{generalization error bar}}$$



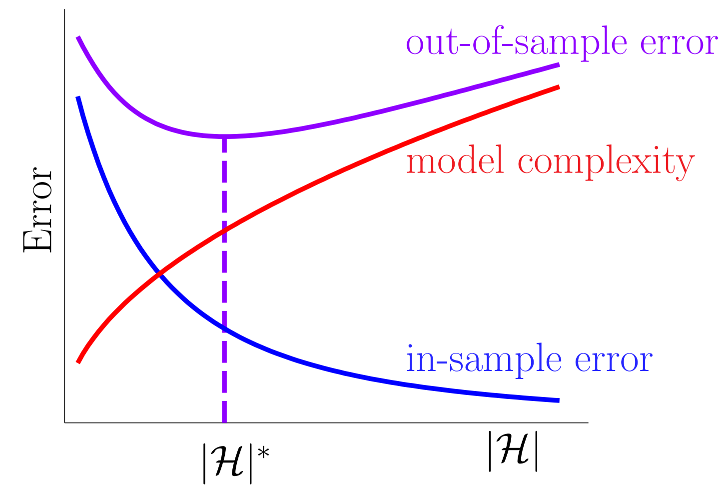
E_{in}: training (eg. the practice exam)

E_{out}: testing (eg. the real exam)

There is a tradeoff when picking $|\mathcal{H}|$.

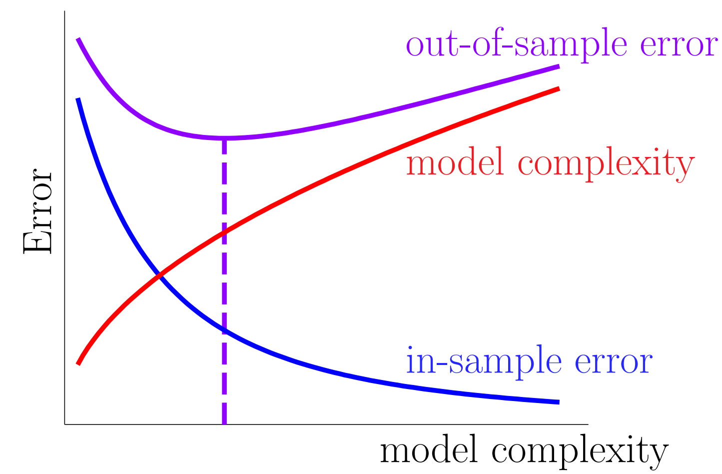
What Will The Theory of *Generalization* Achieve?

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2|\mathcal{H}|}{\delta}}$$



$$\downarrow$$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}}{\delta}}$$



The new bound will be applicable to *infinite* $\mathcal{H}$.

Why is $|\mathcal{H}|$ an Overkill

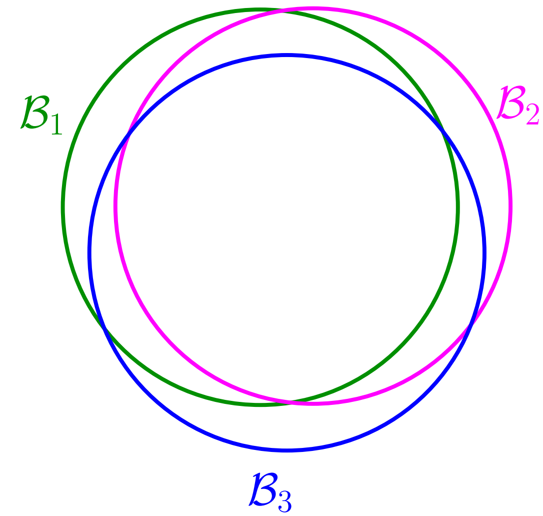
How did $|\mathcal{H}|$ come in?

$\mathcal{B}$ad events

$$\begin{aligned}\mathcal{B}_g &= \{|E_{\text{out}}(g) - E_{\text{in}}(g)| > \epsilon\}\\ \mathcal{B}_m &= \{|E_{\text{out}}(h_m) - E_{\text{in}}(h_m)| > \epsilon\}\end{aligned}$$

We do not know which g, so use a worst case union bound.

$$\mathbb{P}[\mathcal{B}_g] \le \mathbb{P}[\text{any } \mathcal{B}_m] \le \sum_{m=1}^{|\mathcal{H}|} \mathbb{P}[\mathcal{B}_m].$$



- $\mathcal{B}_m$ are events (sets of outcomes); they can overlap.
- If the $\mathcal{B}_m$ overlap, the union bound is loose.
- If many h_m are similar, the $\mathcal{B}_m$ overlap.
- There are "effectively" fewer than $|\mathcal{H}|$ hypotheses,.
- We can replace $|\mathcal{H}|$ by something smaller.

$|\mathcal{H}|$ fails to account for similarity between hypotheses.

Measuring the Diversity (Size) of $\mathcal{H}$

We need a way to measure the *diversity* of $\mathcal{H}$.

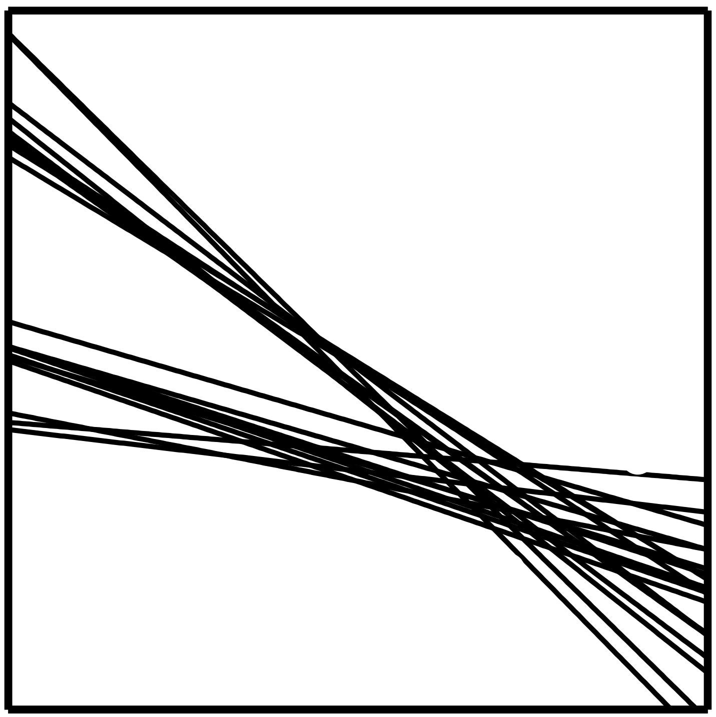
A simple idea:

Fix *any* set of N data points.

If $\mathcal{H}$ is diverse it should be able to implement all functions

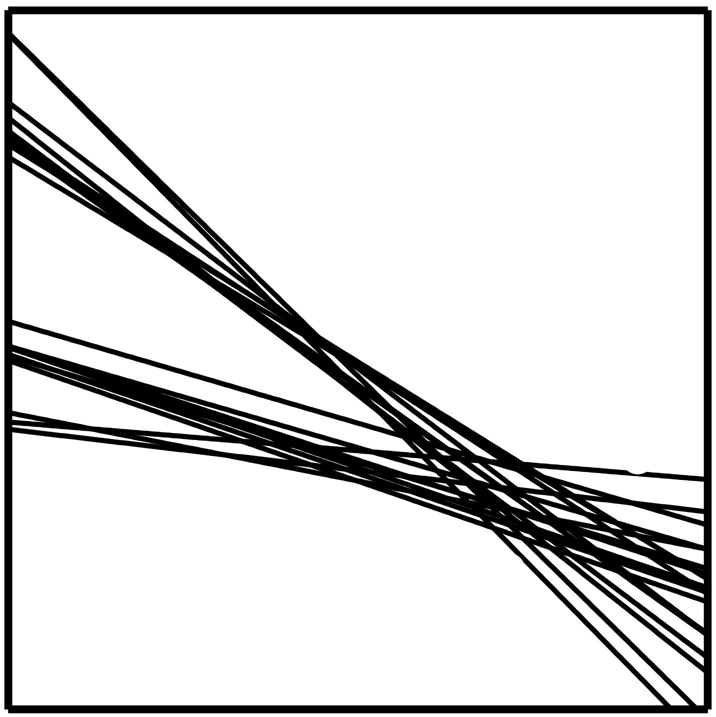
. . . on these N points.

A Data Set Reveals the True Colors of an $\mathcal{H}$

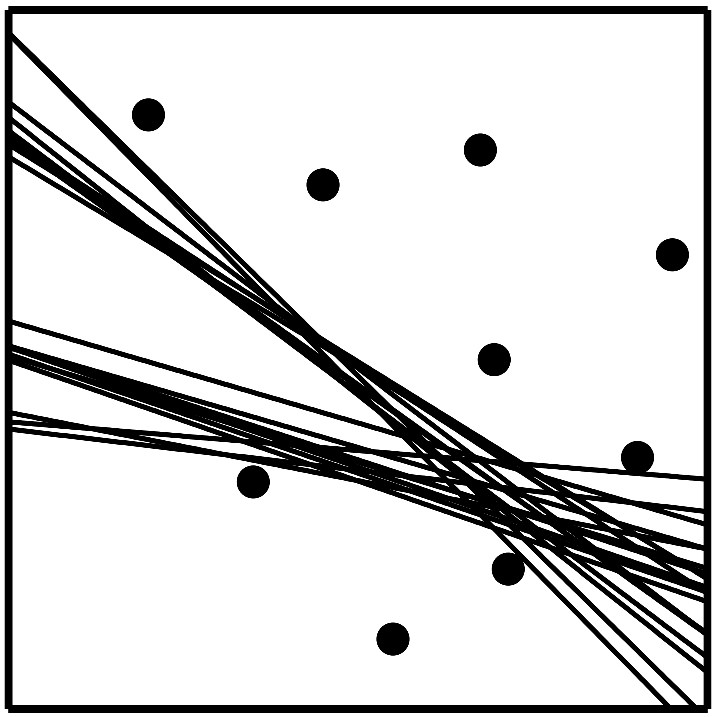


$\mathcal{H}$

A Data Set Reveals the True Colors of an $\mathcal{H}$

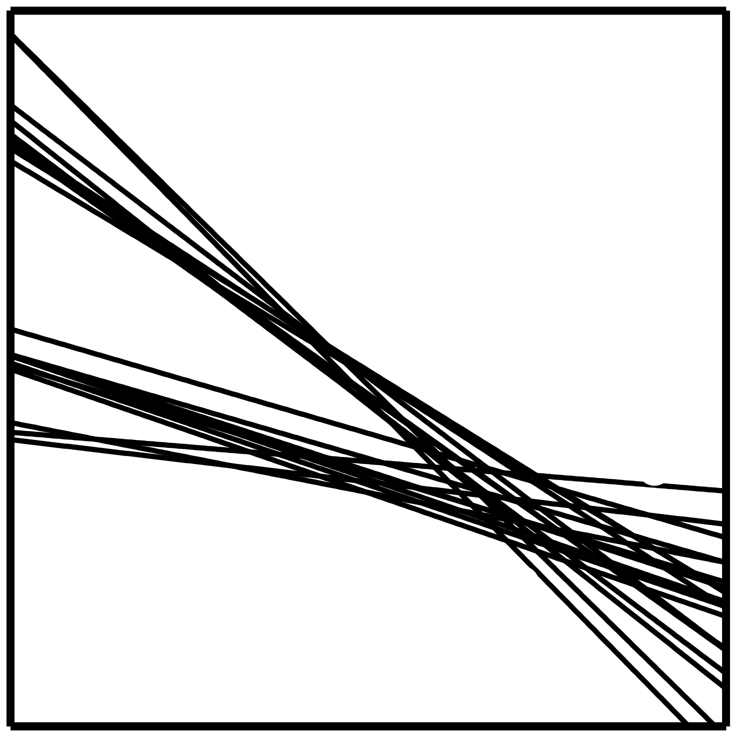


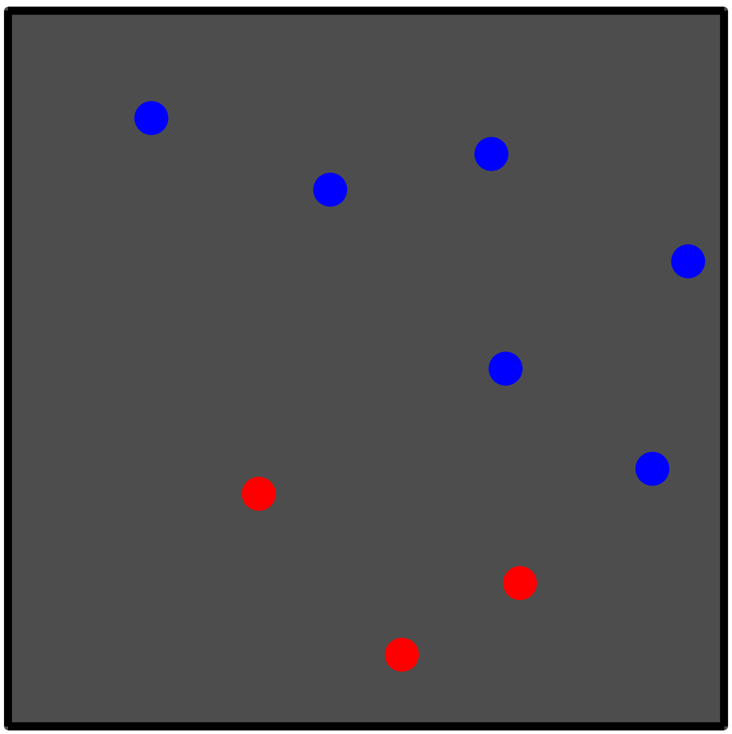
$\mathcal{H}$



$\mathcal{H}$ through the eyes of the $\mathcal{D}$

A Data Set Reveals the True Colors of an $\mathcal{H}$





From the point of view of $\mathcal{D}$, the entire $\mathcal{H}$ is just one *dichotomy*.

An Effective Number of Hypotheses

If $\mathcal{H}$ is diverse it should be able to implement many dichotomys.

$|\mathcal{H}|$ only captures the maximum possible diversity of $\mathcal{H}$.

Consider an $h \in \mathcal{H}$, and a data set $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

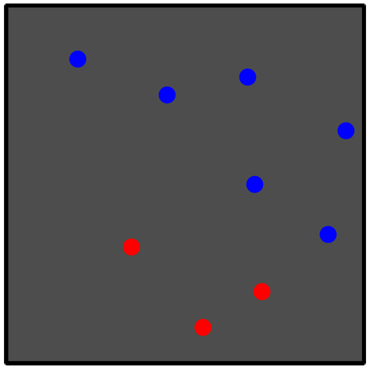
h gives us an N-tuple of ± 1's:

$(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_N))$.

A *dichotomy* of the inputs.

If $\mathcal{H}$ is diverse, we get many different dichotomies.

If $\mathcal{H}$ contains similar functions, we only get a few dichotomies.



dichotomy

The growth function quantifies this.

The Growth Function $m_{\mathcal{H}}(N)$

Define the the restriction of $\mathcal{H}$ to the inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$:

$$\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_N)) \mid h \in \mathcal{H}\}$$

(set of dichotomies induced by $\mathcal{H}$)

The Growth Function $m_{\mathcal{H}}(N)$

The largest set of dichotomies induced by $\mathcal{H}$:

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \ldots, \mathbf{x}_N} |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)|.$$

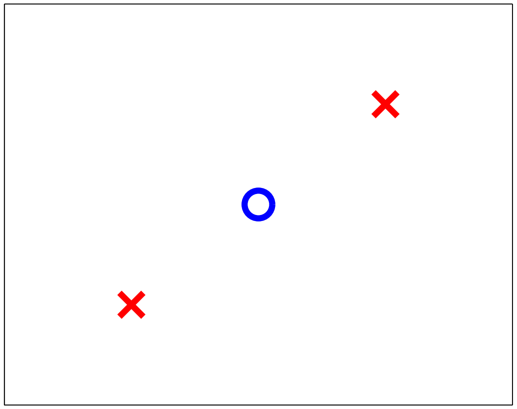
$m_{\mathcal{H}}(N) \le 2^N$.

Can we replace $|\mathcal{H}|$ by $m_{\mathcal{H}}$, an effective number of hypotheses?

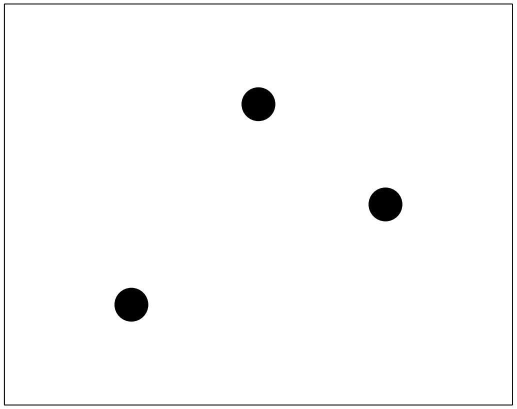
- Replacing $|\mathcal{H}|$ with 2^N is no help in the bound. (why?)
- We want $m_{\mathcal{H}}(N) \le \text{poly(N)}$ to get a useful error bar.

$\left(\text{the error bar is } \sqrt{\frac{1}{2N} \ln \frac{2|\mathcal{H}|}{\delta}}\right)$

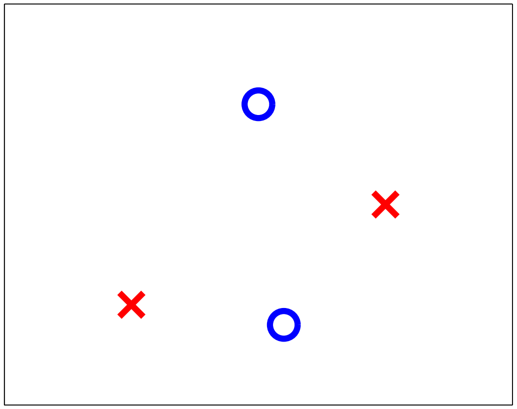
Example: 2-D Perceptron Model



Cannot implement



Can implement all 8



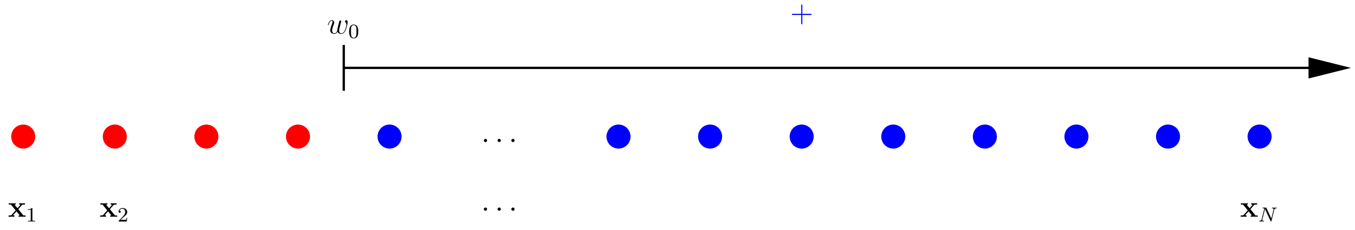
Can implement at most 14

$$m_{\mathcal{H}}(3) \;=\; 8 = 2^3.$$

$$m_{\mathcal{H}}(4) \;=\; 14 < 2^4.$$

What is $m_{\mathcal{H}}(5)$?

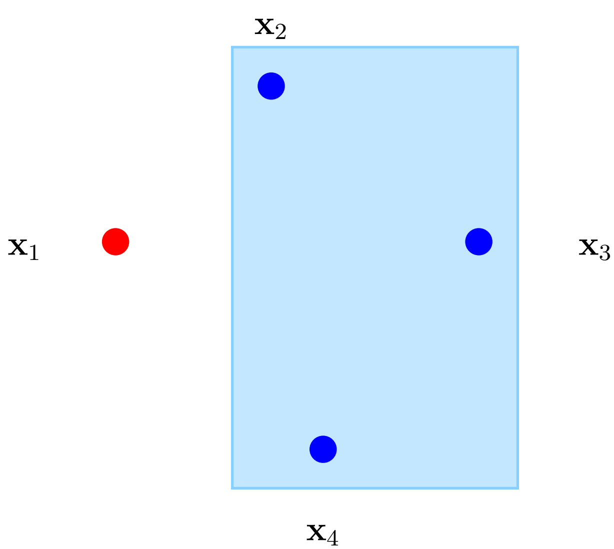
Example: 1-D Positive Ray Model



- $h(x) = sign(x - w_0)$
- Consider N points.
- There are $N+1$ dichotomies depending on where you put w_0.
- $m_{\mathcal{H}}(N) = N + 1$.

Example: Positive Rectangles in 2-D

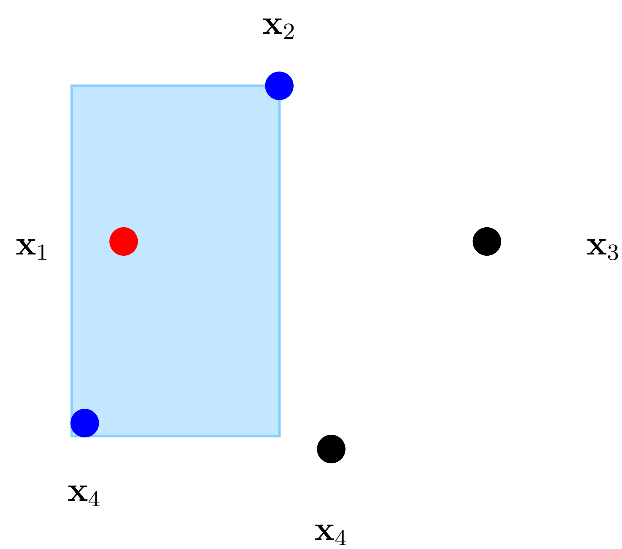
$\underline{N = 4}$



$\mathcal{H}$ implements all dichotomies

$$m_{\mathcal{H}}(4) = 2^4$$

$\underline{N = 5}$



some point will be inside a rectangle defined by others

$$m_{\mathcal{H}}(5) < 2^5$$

We have not computed $m_{\mathcal{H}}(5)$ – not impossible, but tricky.

Example Growth Functions

	N					
	1	2	3	4	5	$\cdots$
2-D perceptron	2	4	8	14	$\cdots$	
1-D pos. ray	2	3	4	5	$\cdots$	
2-D pos. rectangles	2	4	8	16	$< 2^5$	$\cdots$

- $m_{\mathcal{H}}(N)$ drops below 2^N – there is hope for the generalization bound.
- A break point is any n for which $m_{\mathcal{H}}(n) < 2^n$.

A Combinatorial Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
○	●	●

A set of dichotomys

A Combinatorial Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
○	●	●

Two points are *shattered*

A Combinatorial Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
●	○	○

No pair of points is shattered

A Combinatorial Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
●	○	○

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
○	○	○	○
○	○	○	●
	⋮		

4 dichotomies is max.

If $N = 4$ how many possible dichotomys with no 2 points shattered?