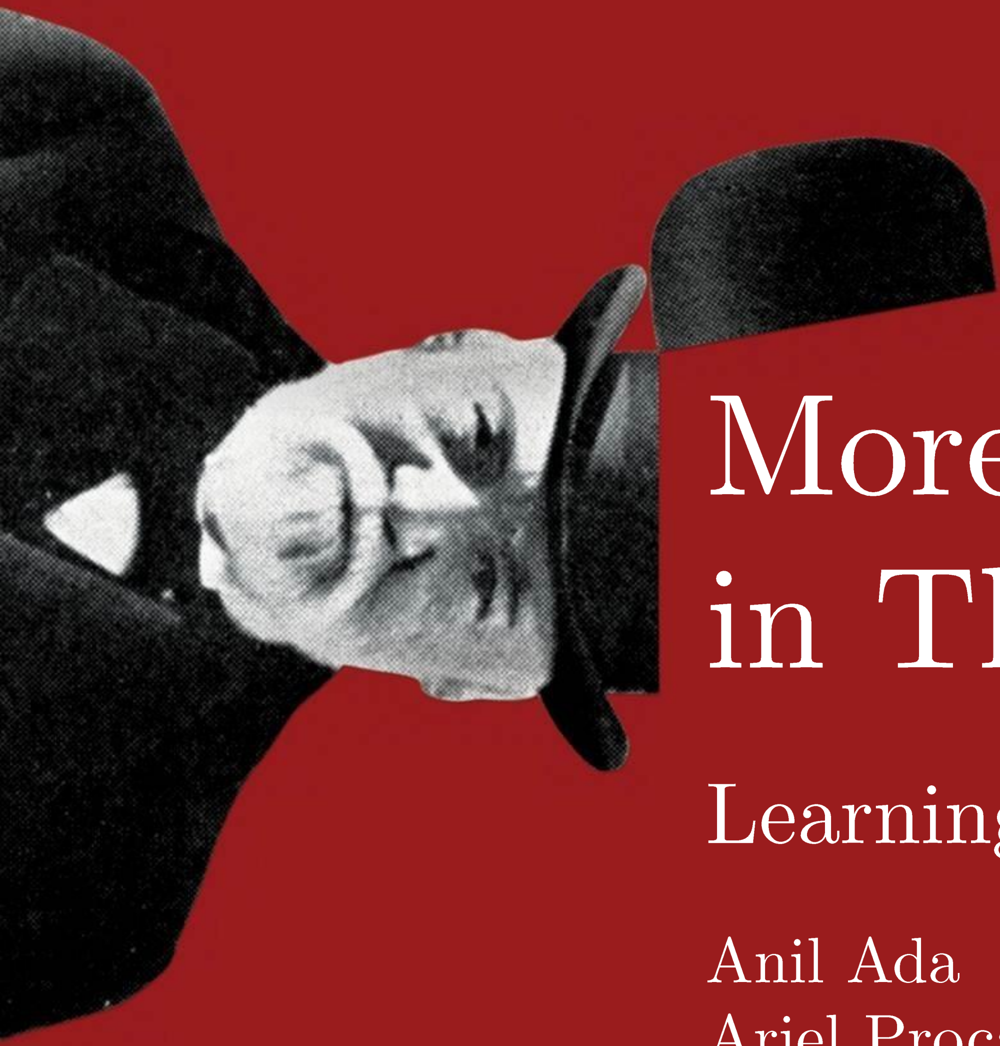


More Great Ideas in Theoretical CS

Learning Theory

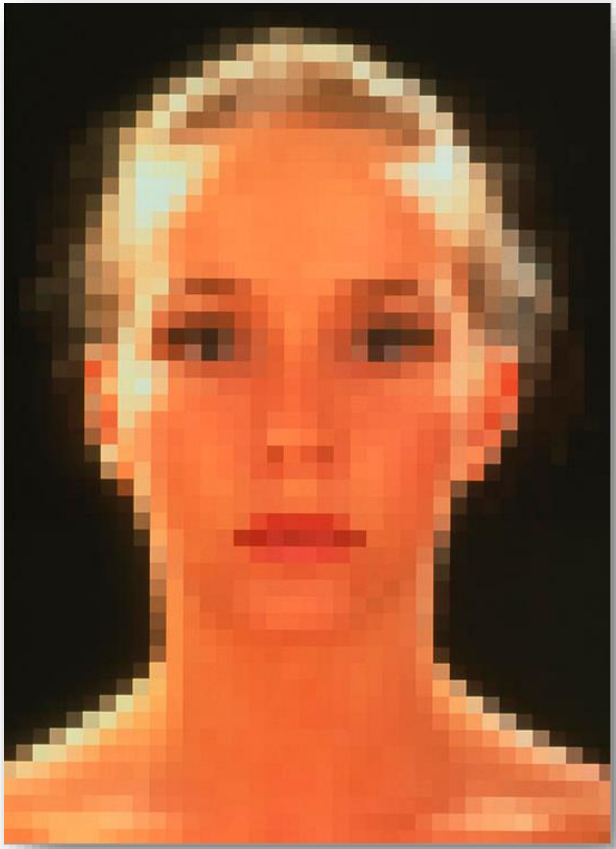
Anil Ada
Ariel Procaccia (this time)

The PAC model

- Input space X
- D distribution over X: unknown but fixed
- Learner receives a set S of m instances $x_1, \dots, x_m$, independently sampled according to D
- Function class F of functions $f: X \to \{+, -\}$
- Assume target function $f_t \in F$
- Training examples $Z = \{(x_i, f_t(x_i))\}$

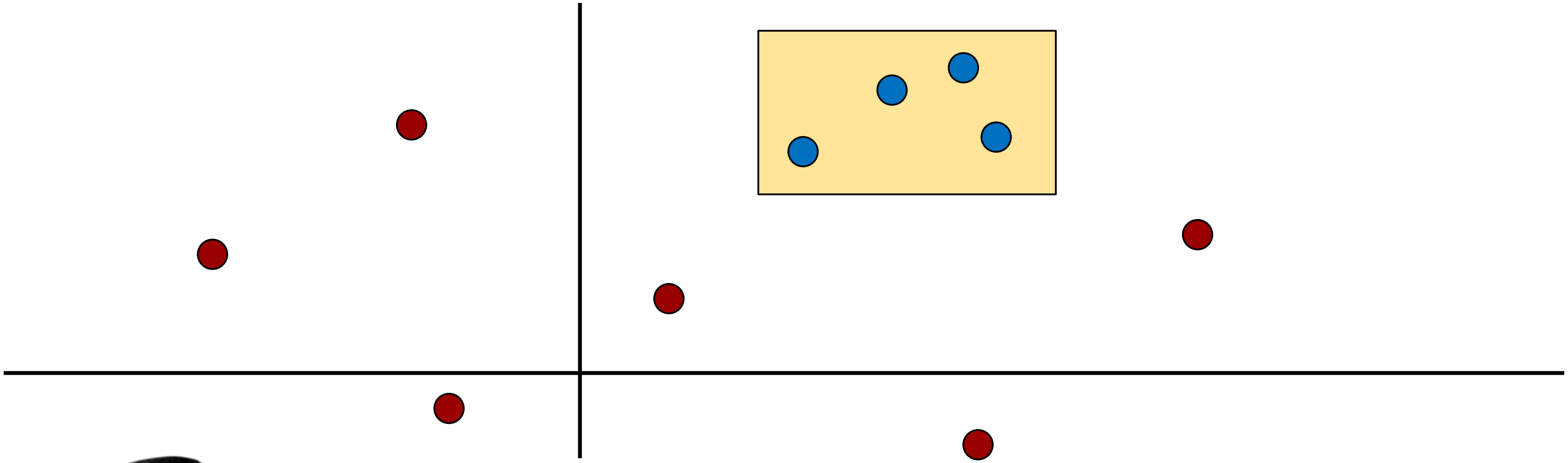
Example: Faces

- $X = \mathbb{R}^{k \times \ell}$
- Each $x \in X$ is a matrix of colors, one per pixel
- $f_t(x) = +$ iff x is a picture of a face
- Training examples: Each is a picture labeled "face" or "not face"



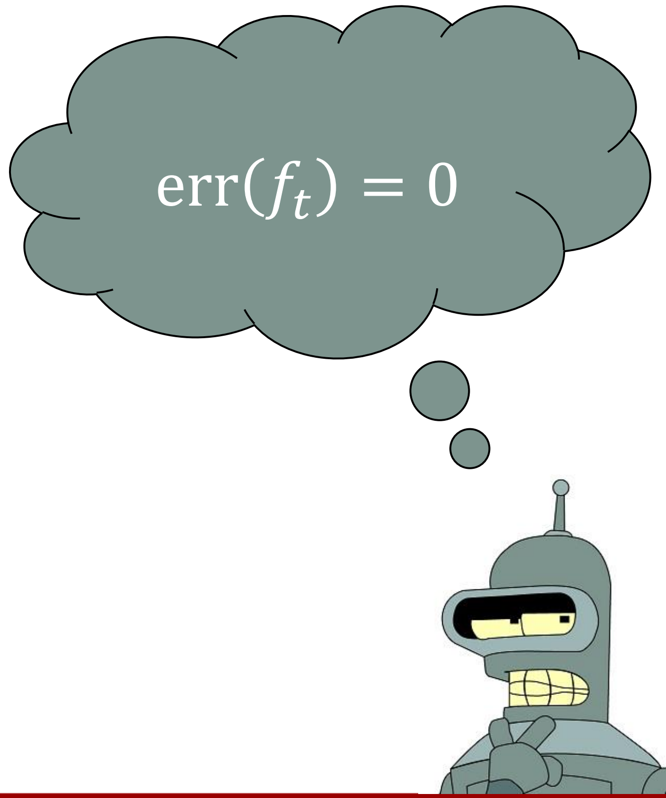
Example: rectangle learning

- $X = \mathbb{R}^2$
- F = axes-aligned rectangles
- $f(x) = +$ iff x is contained in f



The PAC model

- The error of function f is $\mathrm{err}(f) = \Pr_{x\sim D}[f_t(x) \neq f(x)]$
- Given accuracy parameter $\epsilon > 0$, would like to find function f with $\mathrm{err}(f) \leq \epsilon$
- Given confidence parameter $\delta > 0$, would like to achieve $\Pr[\mathrm{err}(f) \leq \epsilon] \geq 1 - \delta$

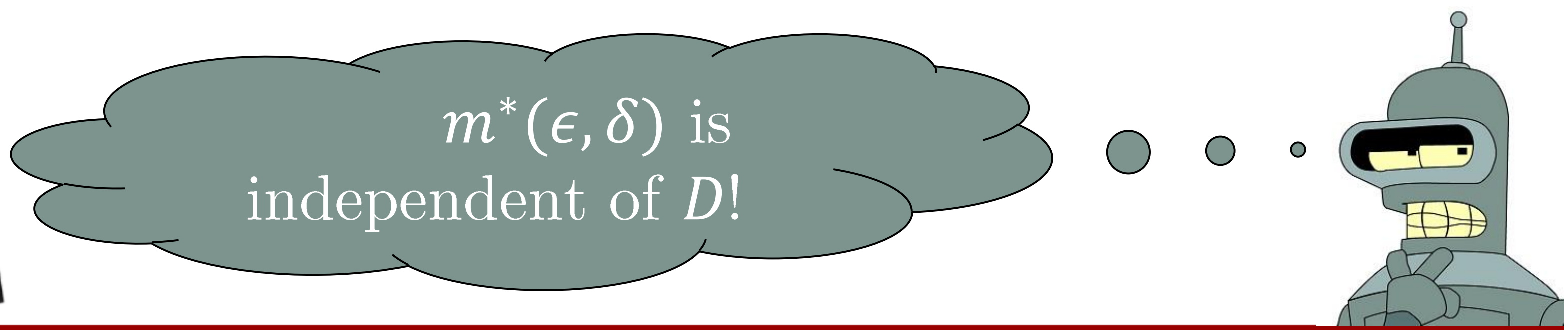


The PAC model

- A learning algorithm L is a function from training examples to F such that: for every $\epsilon, \delta > 0$ there exists $m^*(\epsilon, \delta)$ such that for every $m \geq m^*$ and every D, if m examples Z are drawn from D and $L(Z) = f$ then

$$\Pr[\text{err}(f) \leq \epsilon] \geq 1 - \delta$$

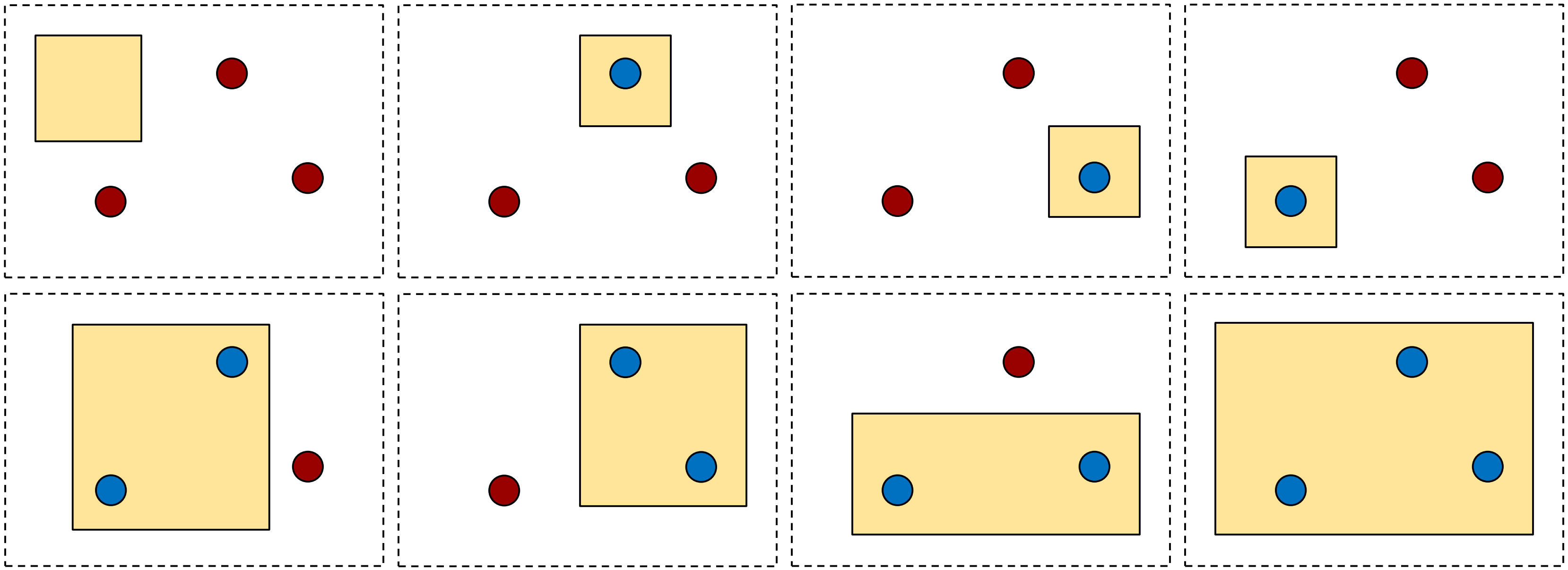
- F is learnable if there is a learning algorithm for F



VC DIMENSION

- We would like to obtain a general connection between learnability and combinatorial properties of the function class
- Let $S = \{x_1, \dots, x_m\}$
- $\Pi_F(S) = \left\{\left(f(x_1), \dots, f(x_m)\right) : f \in F\right\}$

VC DIMENSION



$$\Pi_F(S) = \{(-,-,-),(-,+,-),(-,-,+),(+,-,-),\\ (+,+,-),(-,+,+),(+,-,+),(+,+,+)\}$$

VC DIMENSION

- X = real line
- F = intervals; points inside interval are labeled by $+$, outside by $-$
- Poll 1: what is $|\Pi_F(S)|$ for $S =$ 
  1. 1
  2. 2
  3. 3
  4. 4

VC dimension

- Poll 2: what is $|\Pi_F(S)|$ for $S =$ 

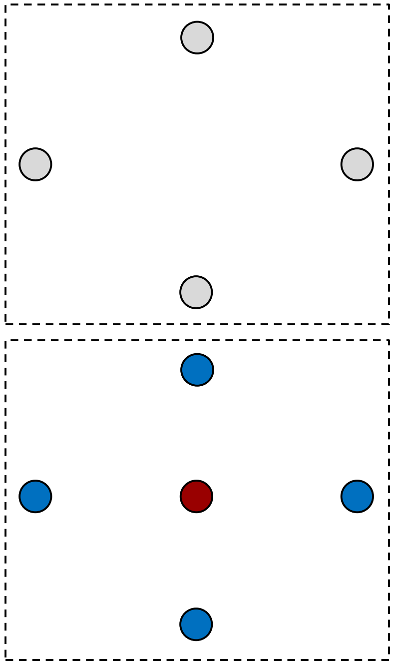
  1. 5
  2. 6
  3. 7
  4. 8

VC DIMENSION

- S is shattered by F if $|\Pi_F(S)| = 2^{|S|}$
- The VC dimension of F is the cardinality of the largest set that is shattered by F

EXAMPLE: RECTANGLES

- There is an example of four points that can be shattered
- For any choice of five points, one is "internal"
- A rectangle cannot label outer points by 1 and inner point by 0
- VC dimension is 4

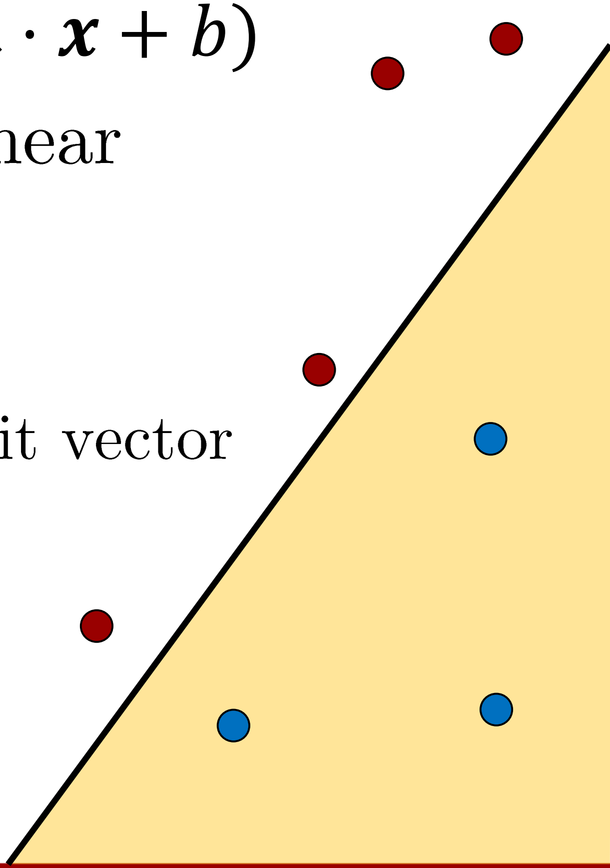


VC DIMENSION

- Poll 3: X = real line, F = intervals, what is VC-dim(F)?
  1. 1
  2. 2
  3. 3
  4. None of the above
- Poll 4: X = real line, F = unions of intervals, what is VC-dim(F)?
  1. 2
  2. 3
  3. 4
  4. None of the above

Example: Linear separators

- $X = \mathbb{R}^d$
- A linear separator is $f(\boldsymbol{x}) = \text{sgn}(\boldsymbol{a} \cdot \boldsymbol{x} + b)$
- Theorem: The VC dimension of linear separators is $d + 1$
- Proof (lower bound):
    - $\boldsymbol{e}^i = (0, \ldots, 0,1,0, \ldots, 0)$ is the i-th unit vector
    - $S = \{\mathbf{0}\} \cup \{\boldsymbol{e}^i : i = 1, \ldots, d\}$
    - Given $y^0, \ldots, y^d \in \{-1,1\}$, set $\boldsymbol{a} = (y^1, \ldots, y^d), b = y^0/2$ ■



SAMPLE COMPLEXITY

- If for any k there is a sample of size k that can be shattered by F, we say that VC-dim(F)$= \infty$
- Theorem: A function class F with VC-dim(F)$=$ ∞ is not PAC learnable
- Theorem: Let F with VC-dim(F)$= d$. Let L be an algorithm that produces an $f \in F$ that is consistent with the given samples S. Then L is a learning algorithm for F with sample complexity

$$m^*(\epsilon, \delta) = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right)$$