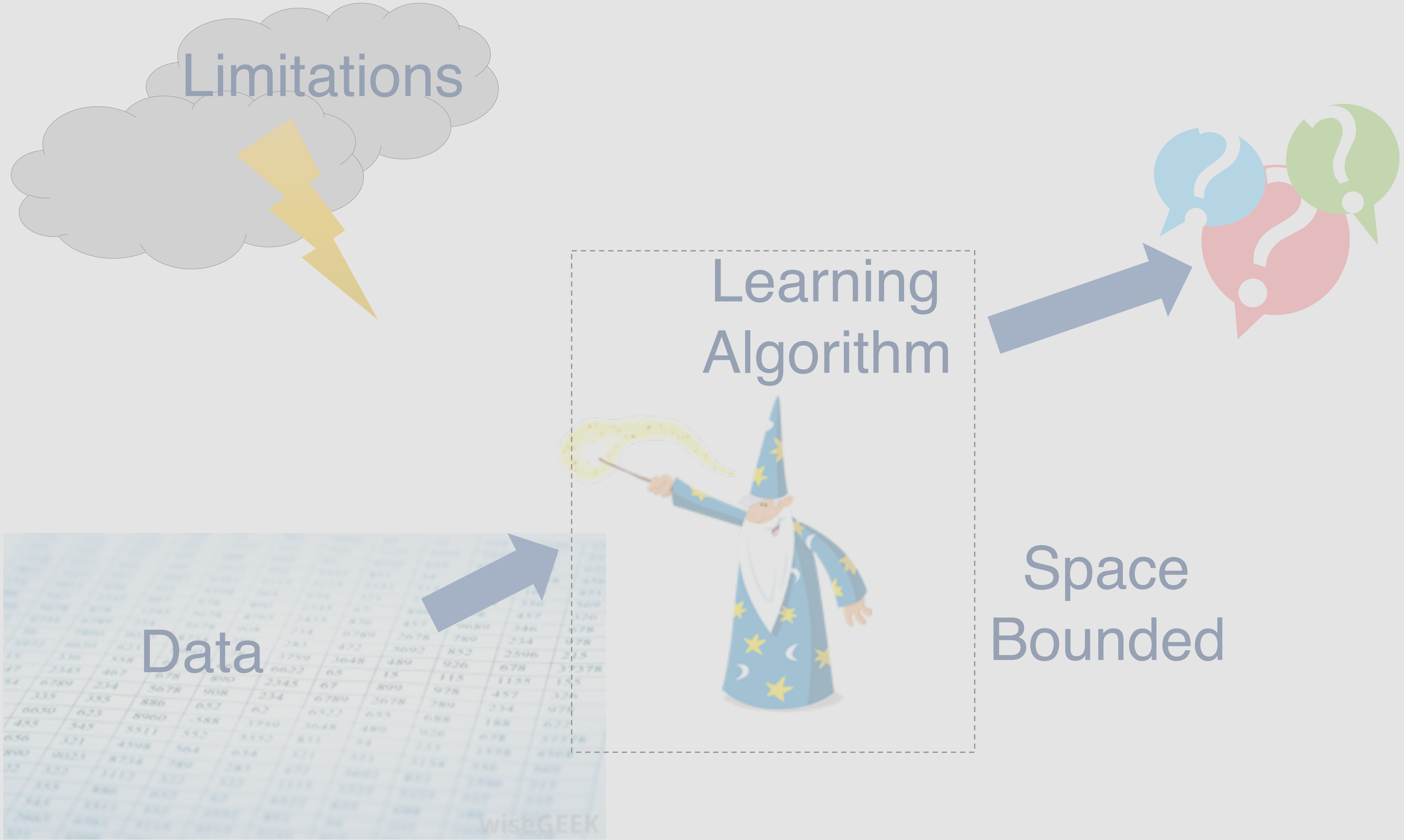
What Cannot Be Learned With Bounded Memory?

Michal Moshkovitz
The Hebrew University of Jerusalem

Joint work with Dana Moshkovitz, UT Austin


Limitations
Learning Algorithm
Space Bounded
Data
wiseGEEK

Learning with Bounded Space: Motivation

- Natural question - next step after time constraints
- We are in the middle of the big data era
- (Artificial) Neural Networks can be viewed as a bounded space algorithm
- Number of neutrons in the nervous system is bounded

Plan

- Definitions
  - What is (PAC) learning?
  - What is online learning?
  - What is bounded memory learning?
- Problem Formulation
- Main Theorem and a Surprising Conclusion

Supervised Learning: Example

labelled examples

h

PAC Learning (Valiant, 1984)

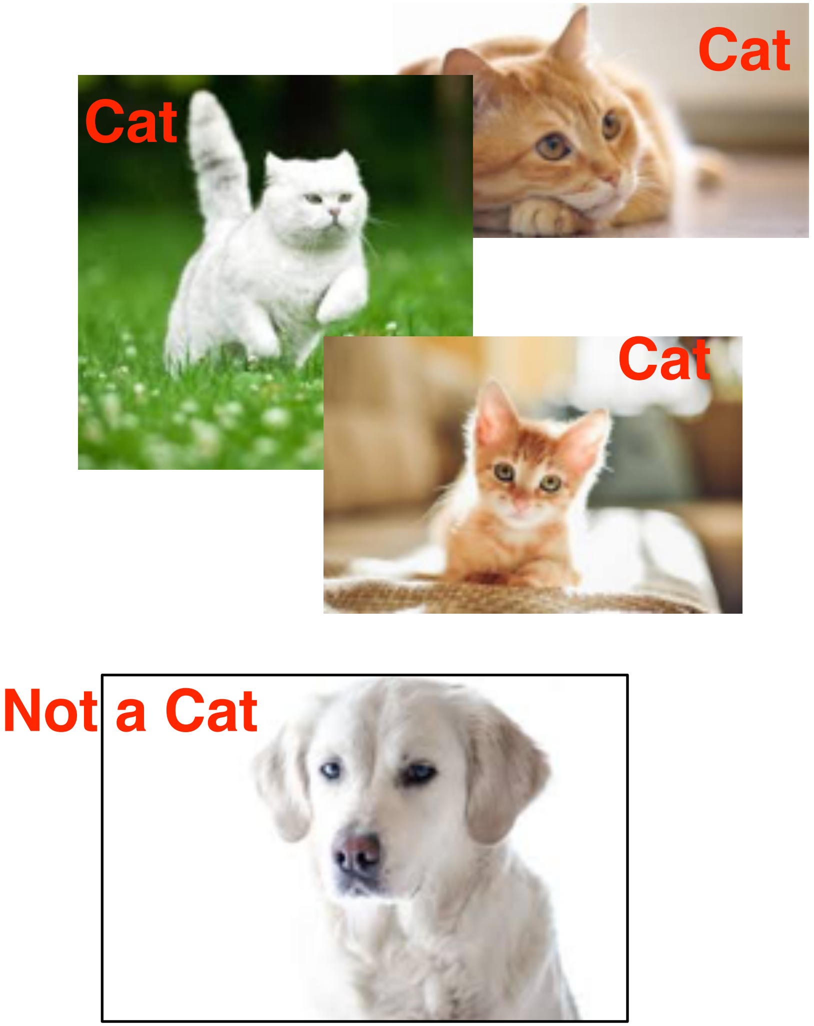
Hypothesis class H={h:X→{0,1}} is PAC-learnable if there is a learner s.t.
for any h, for any distribution over X, with probability > 0.99, the learner will come up with an approximation h' with $P_x(h'(x) \neq h(x)) < 0.01$.

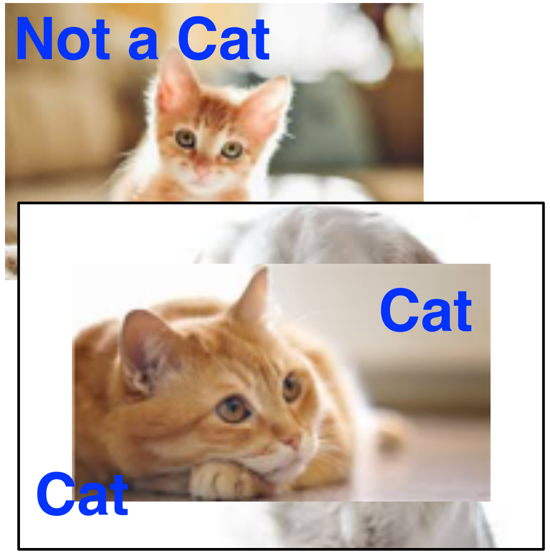
Hypothesis class
H={h:X→{0,1}}

Learner

h' close to h

Online Learning: Example





h^2

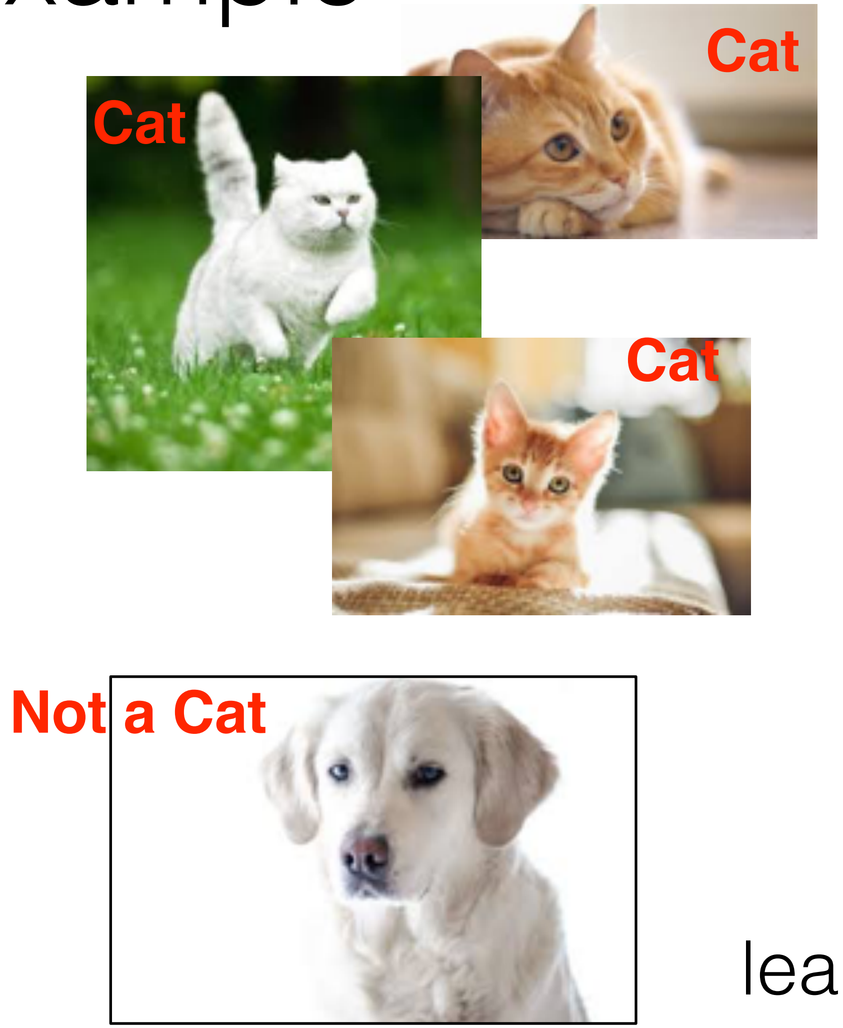
2

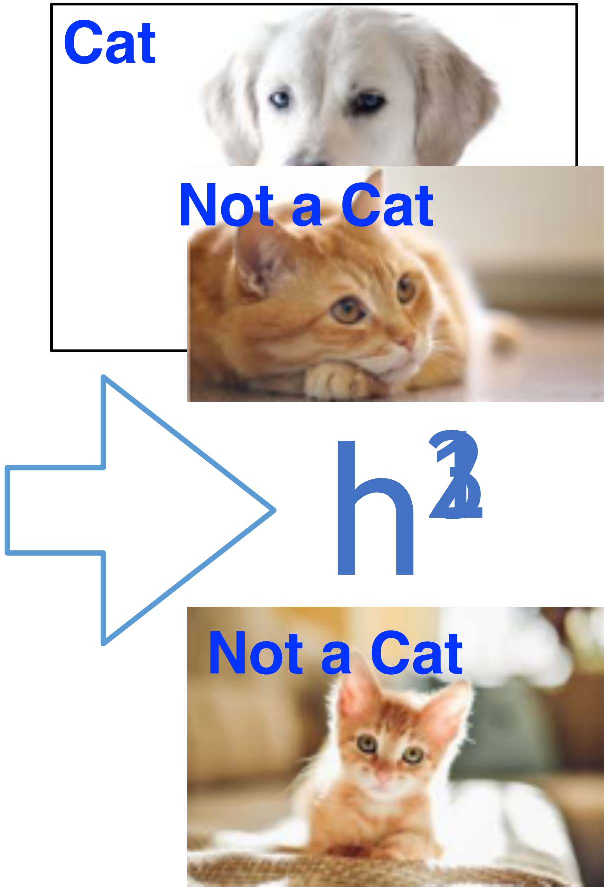
The Online-Learning Framework

- For t=1,2,…
  - An example x^t is given
  - Learner predicts label y^t
  - True label y^t is revealed

- Goal: minimize number of mistakes

Online Learning with Bounded Memory: Example



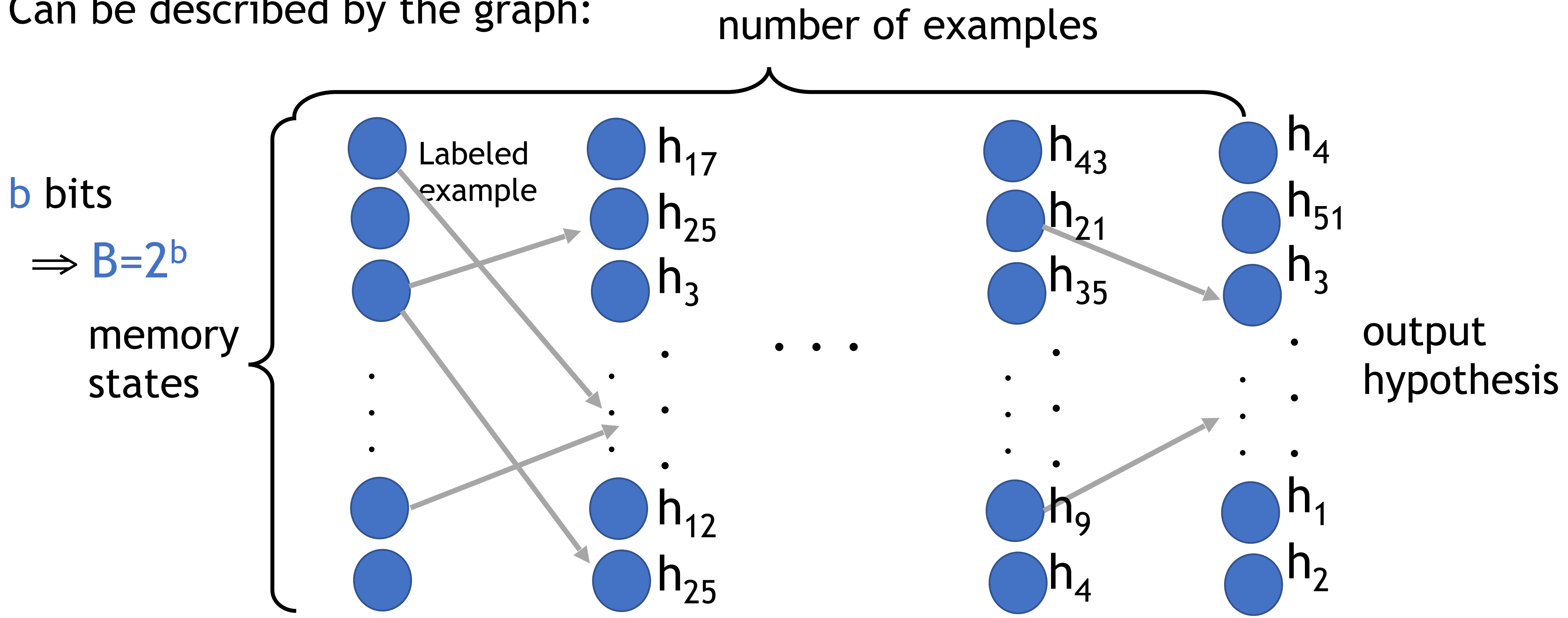


learning with bounded memory is hard

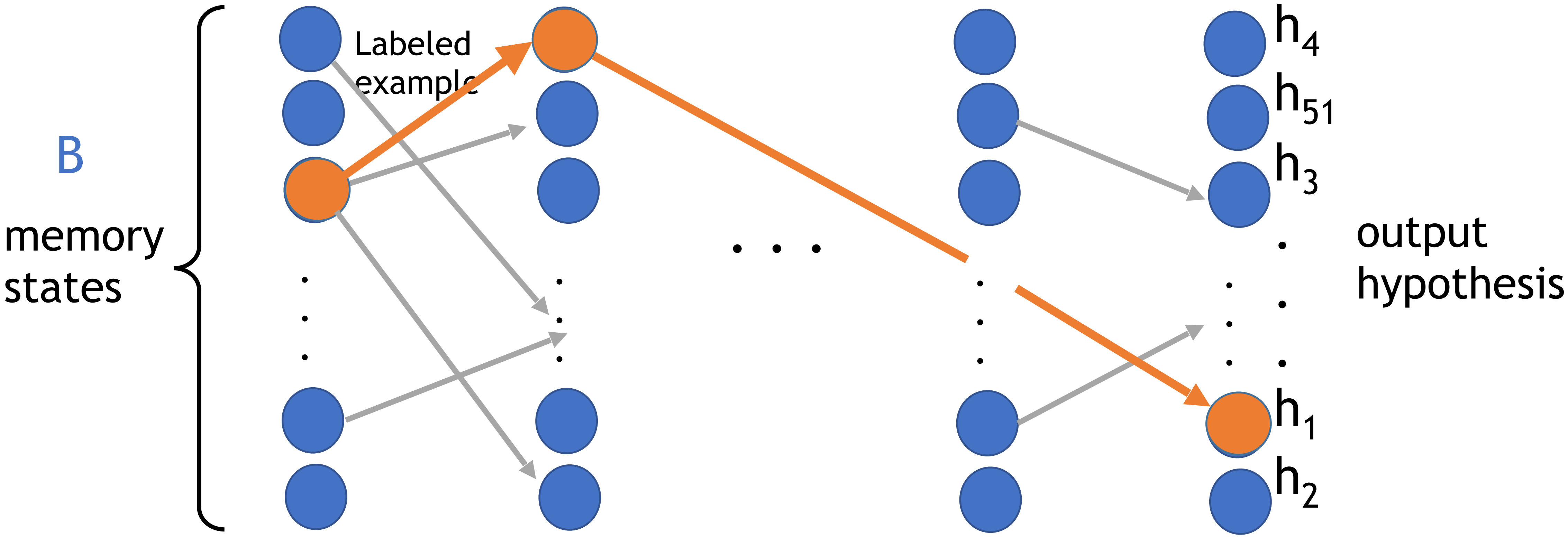
What Cannot be Learned with Bounded Memory?

Learning with Bounded Memory

Fix an algorithm A.
Can be described by the graph:



Learning with Bounded Memory



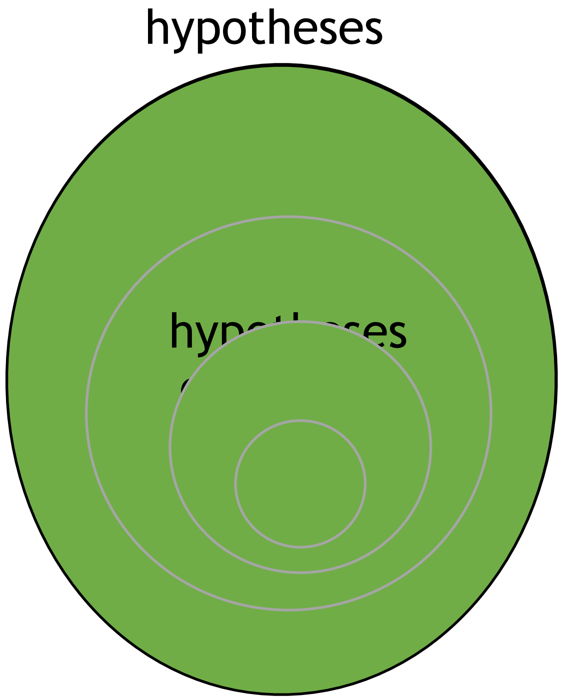
Unbounded Learner

Initially, all h in H are possible.

Repeat: Given example (x,b), rule out h' where $h'(x) \neq b$.

Claim: With high probability, after seeing $O(\log|H|)$ random examples any h' not ruled out satisfies $P_z(h'(z)=h(z))>0.99$.

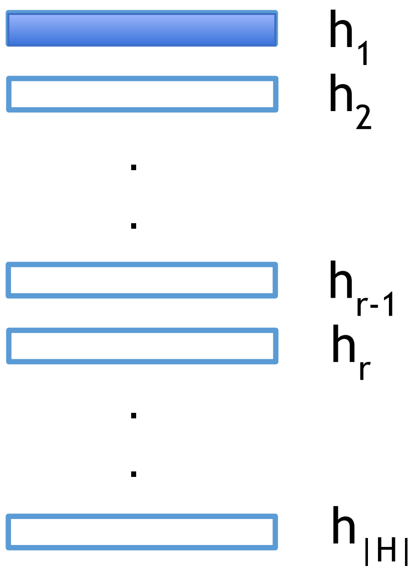
But this requires a lot of memory! Must store received examples in memory, which requires $\min\{\Theta(\log|H|\log|X|), |H|\}$ memory bits.



Bounded Learner

Initially, h’ is the first function in H.

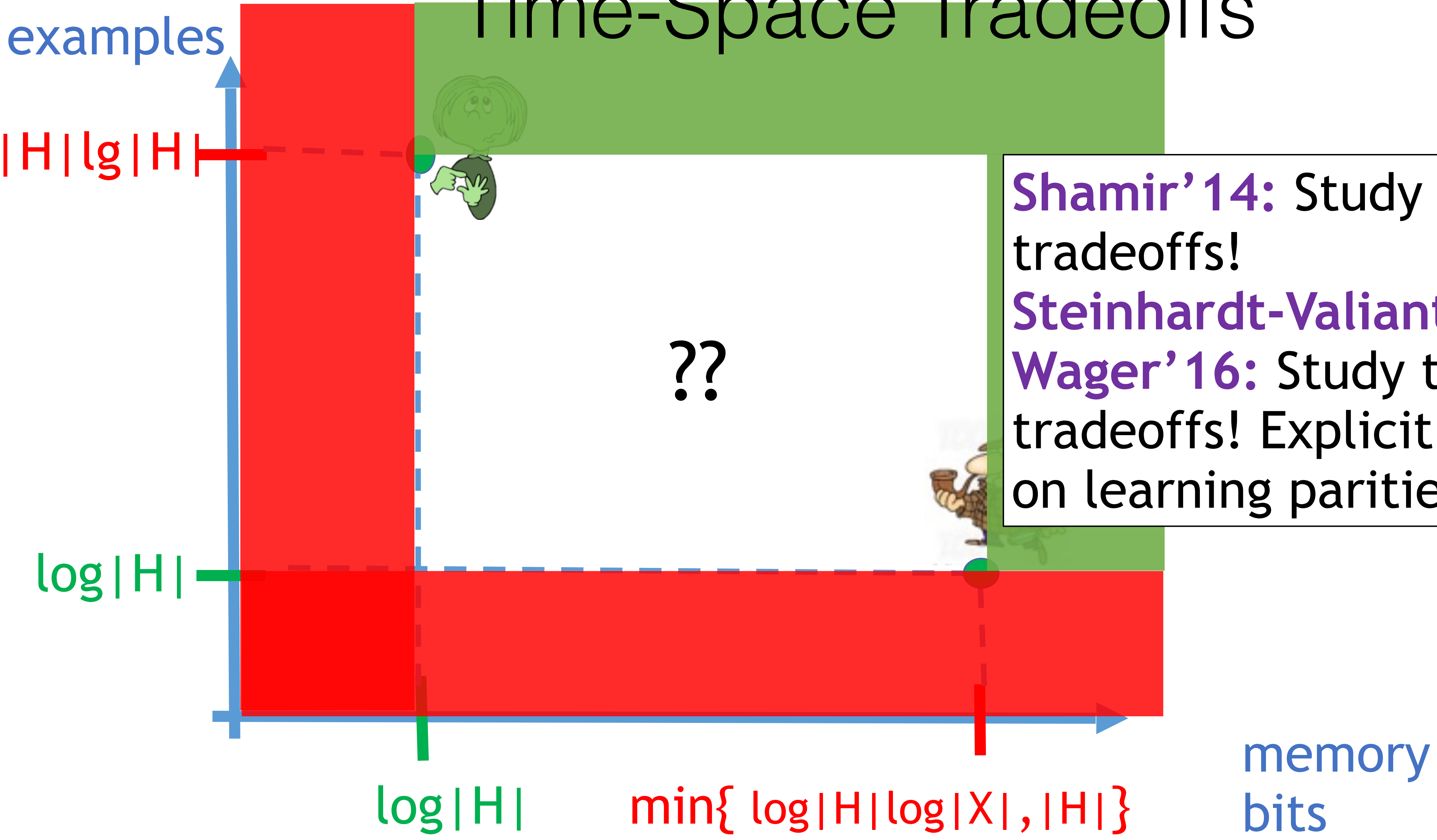
Repeat: Given example (x,b), if $h'(x) \neq b$, then let h’ be the next function in H.



Claim: With high probability, after seeing $O(|H| \log|H|)$ random examples, h’ satisfies $P_z(h'(z)=h(z))>0.99$.

This requires only a minimal number of $\log|H|$ memory bits.

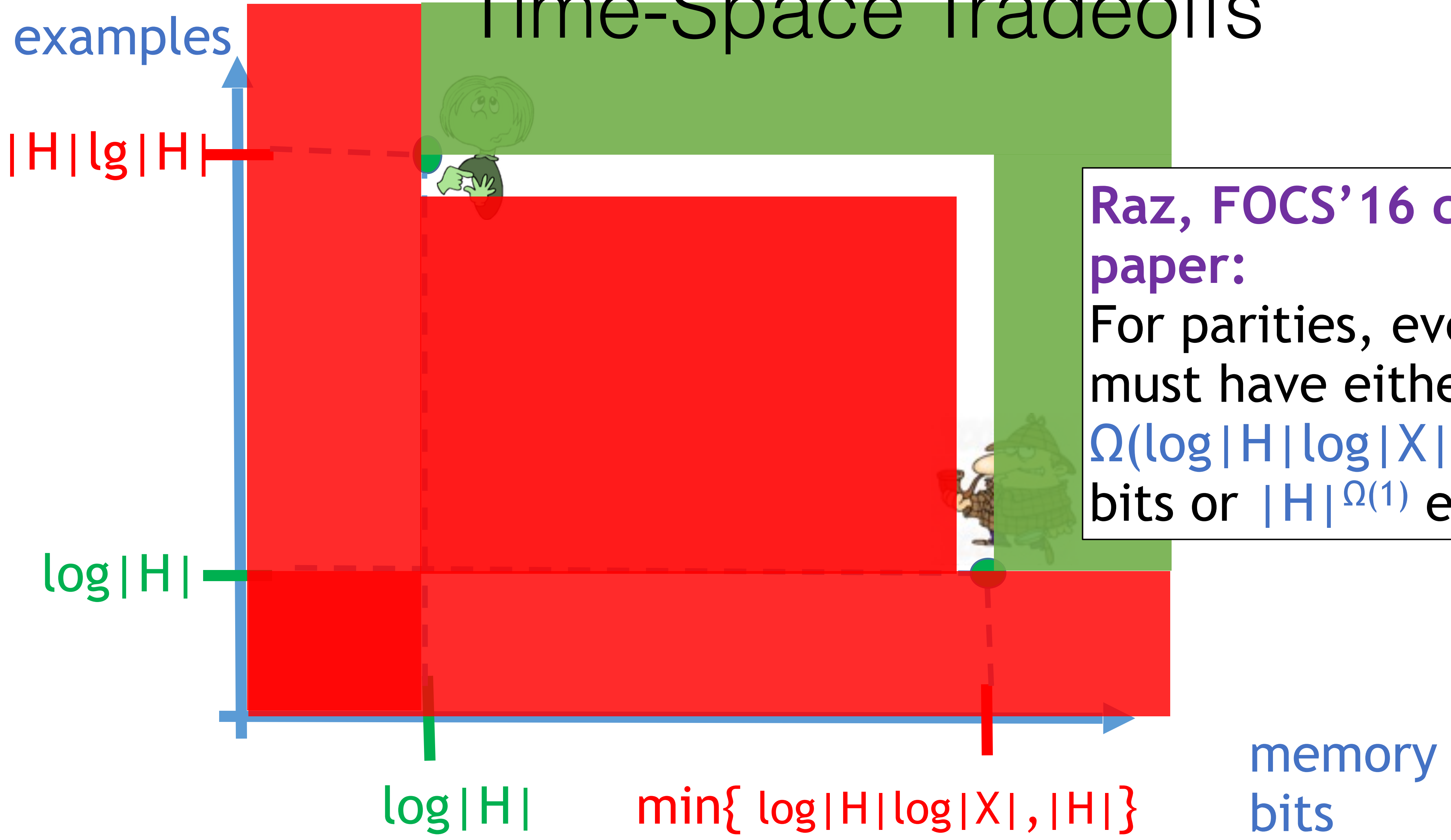
Time-Space Tradeoffs



Shamir'14: Study time-space tradeoffs!

Steinhardt-Valiant-Wager'16: Study time-space tradeoffs! Explicit conjecture on learning parities.

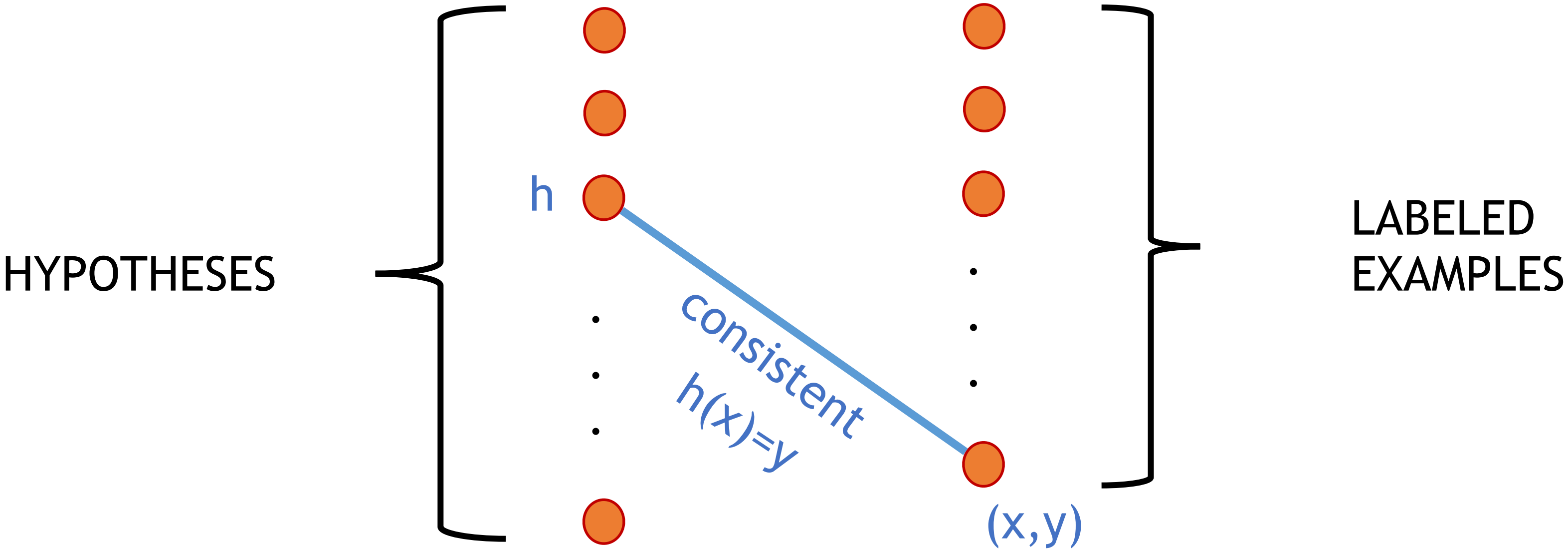
Time-Space Tradeoffs



Raz, FOCS'16 co-best paper:
For parities, every learner must have either $\Omega(\log|H|\log|X|)$ memory bits or $|H|^{\Omega(1)}$ examples.

How Can We Prove Lower Bounds For General Hypotheses Classes?

Hypotheses Graph



Mixing

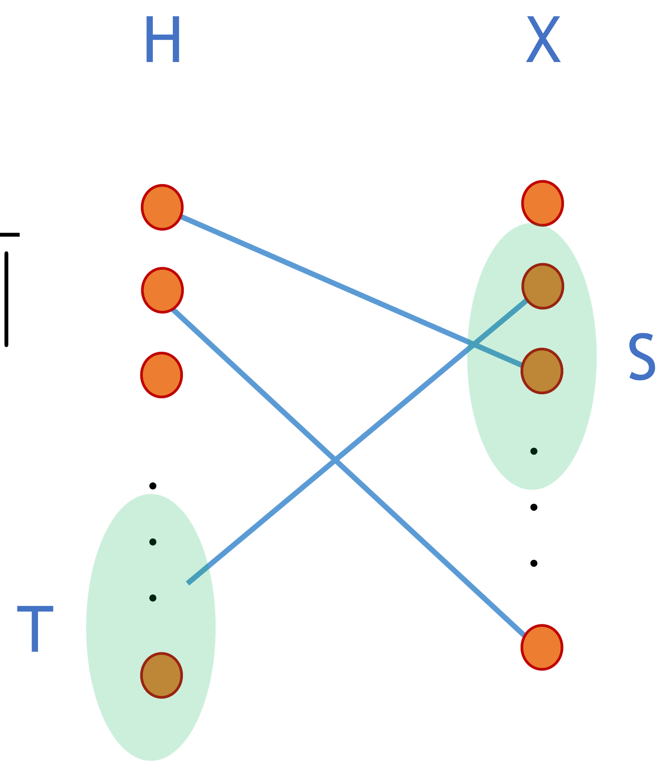
$$p = \frac{E(H,X)}{|H||X|}$$

D-Mixing: For every set S of hypotheses, every set T of labeled examples,

$$\big|E(S,T) - p|S||T|\big| \leq D\sqrt{|S||T|}$$

number of edges between S and T

expected number of edges between S and T



Examples:

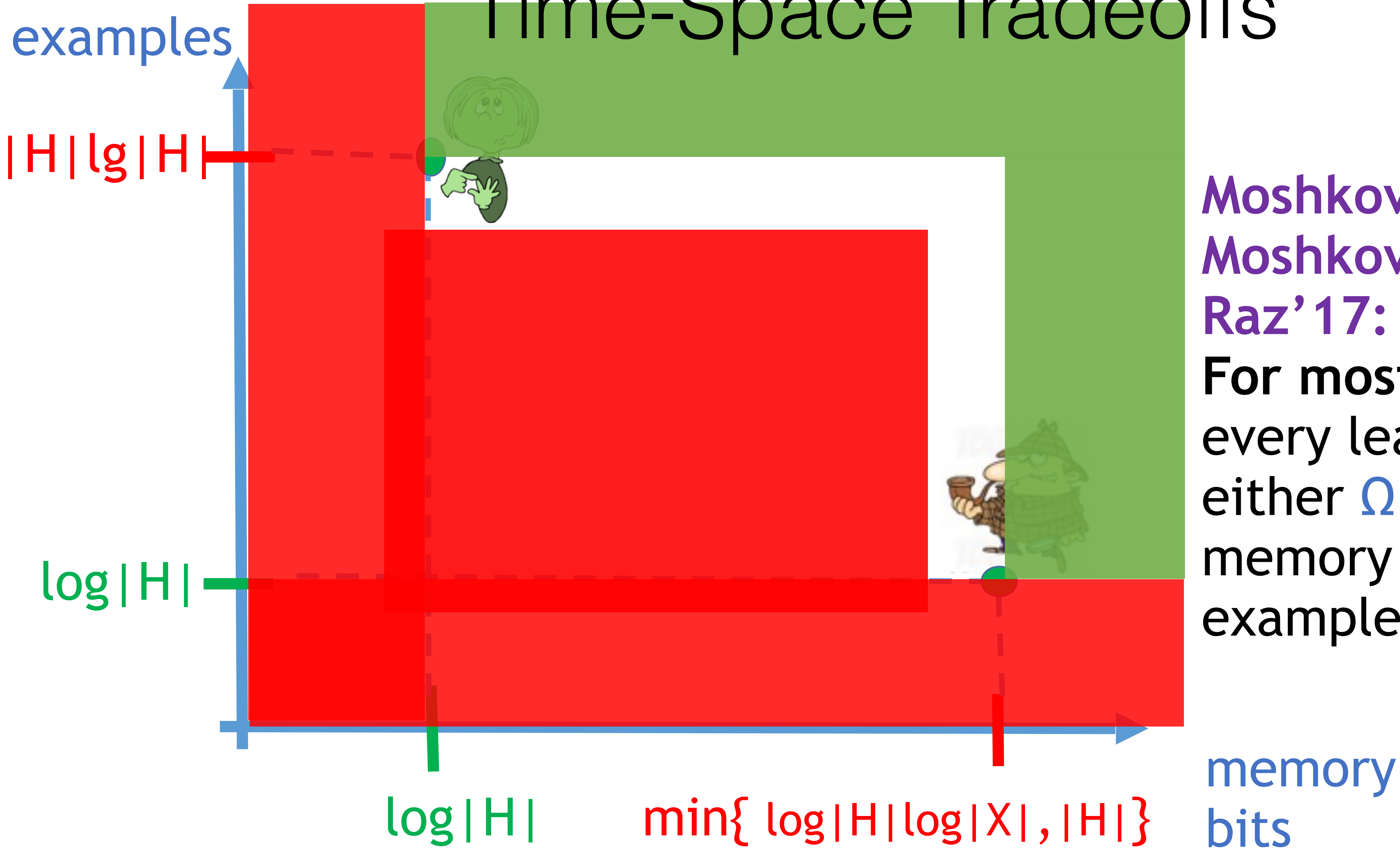
- Parities are O(√|X|)-mixing
- Almost surely G(n,m,0.5) is O(√n)-mixing (for n>m)

Our Theorem: From Mixing To Lower Bounds

Main Thm: If the hypotheses graph is $O(\sqrt{|X|})$-mixing, then either $B=\Omega(\log^2|H|)$ memory bits or $|H|^{\Omega(1)}$ examples are needed to learn.

- A pseudorandomness **sufficient condition** for unlearnability with bounded memory.
- A new combinatorial **framework** for proving lower bounds on space bounded learning.

Time-Space Tradeoffs



Moshkovitz-Moshkovitz'17a+b; Raz'17:

For most classes H, every learner must have either $\Omega(\log^2|H|)$ memory bits or $|H|^{\Omega(1)}$ examples.

The Low Certainty Framework

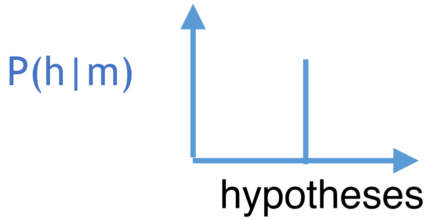
Certainty

We'll pick the underlying hypothesis at random.

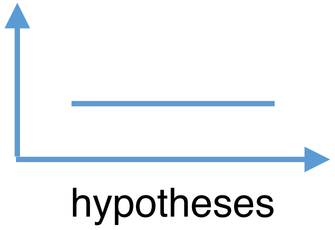
The *certainty* of a memory state m at time t:

- $P(h|m)$ - probability h correct given the algorithm in state m

high certainty:



low certainty:



- $cer^t(m)=\Sigma_h P(h|m)^2$.

The *average certainty* at time t is $cer^t=E_m[cer(m)]$.

Proof Outline

1. **Initially:** $cer^1 = O(1/|H|)$.
2. **Eventually:** Learning at time $T \Rightarrow cer^T \geq \Omega(1)$.
3. **We'll show:** For any learner with B memory states, for every time t, after removal of few hypotheses and examples of low total probability, $cer^{t+1} \leq (1+1/|H|^{\Theta(1)})cer^t$.

As a result: either B memory states or $|H|^{\Omega(1)}$ examples are needed.

The Intuition (The paper is 51 pages)

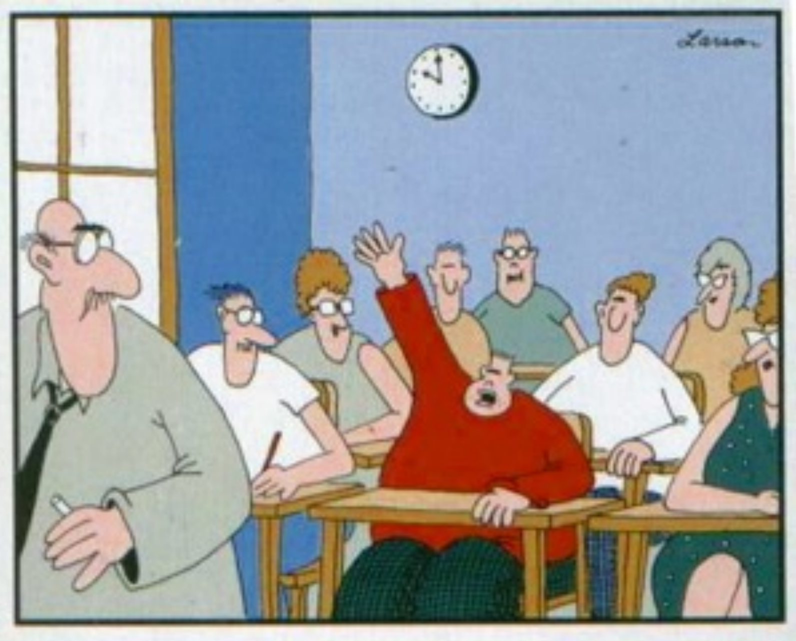
Given an example there are two things the learner can do:

(1) **Heavy step:**

- **Remember only a little about the example.** For mixing H, gain almost no information about h.

(2) **Many Step:**

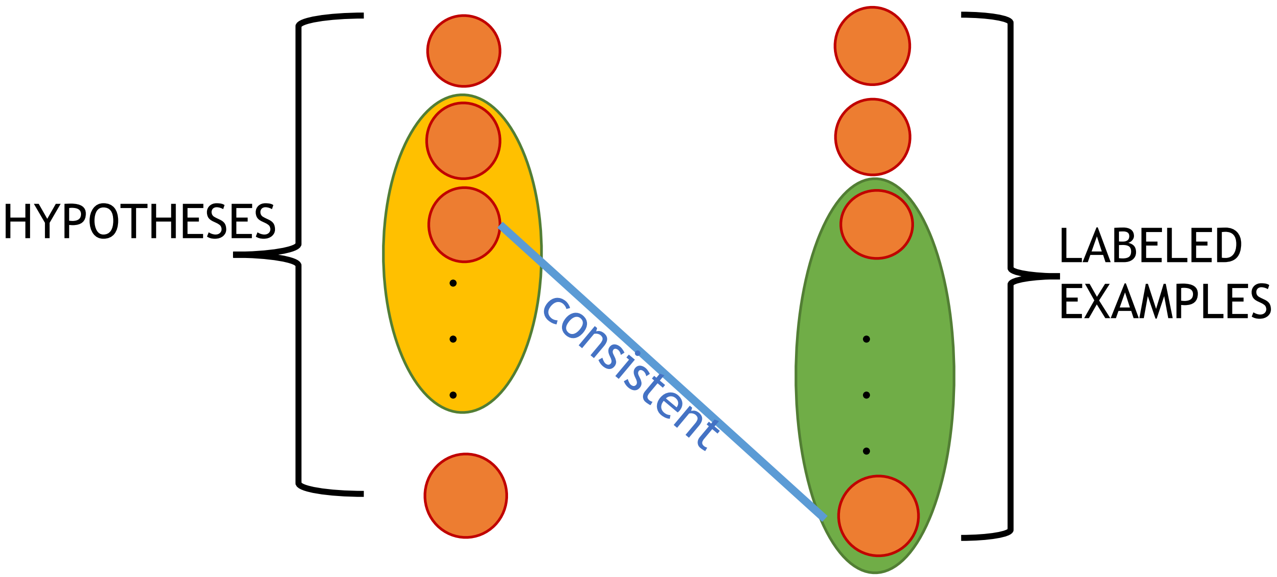
- **Remember the example in detail**, but then have to erase previous memories.



"Mr. Osborne, may I be excused? My brain is full."

Forget Some Lose All Principle

Fix a mixing hypotheses graph. Suppose that one picks uniformly at random a hypothesis h and an example x. Suppose that one stores a short string s about (x,h(x)). Then h|s is close to uniform.



Surprising Conclusion

Most problems cannot be learned with bounded memory …

Luckily, real-problems are not mixing
and can be learned with bounded memory

Thank you!