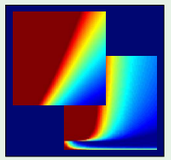
# Learning From Data

Yaser S. Abu-Mostafa
*California Institute of Technology*

## Lecture 2: **Is Learning Feasible?**


Sponsored by Caltech's Provost Office, E&AS Division, and IST • Thursday, April 5, 2012



# Feasibility of learning - Outline

- Probability to the rescue
- Connection to learning
- Connection to *real* learning
- A dilemma and a solution

# A related experiment

- Consider a 'bin' with red and green marbles.

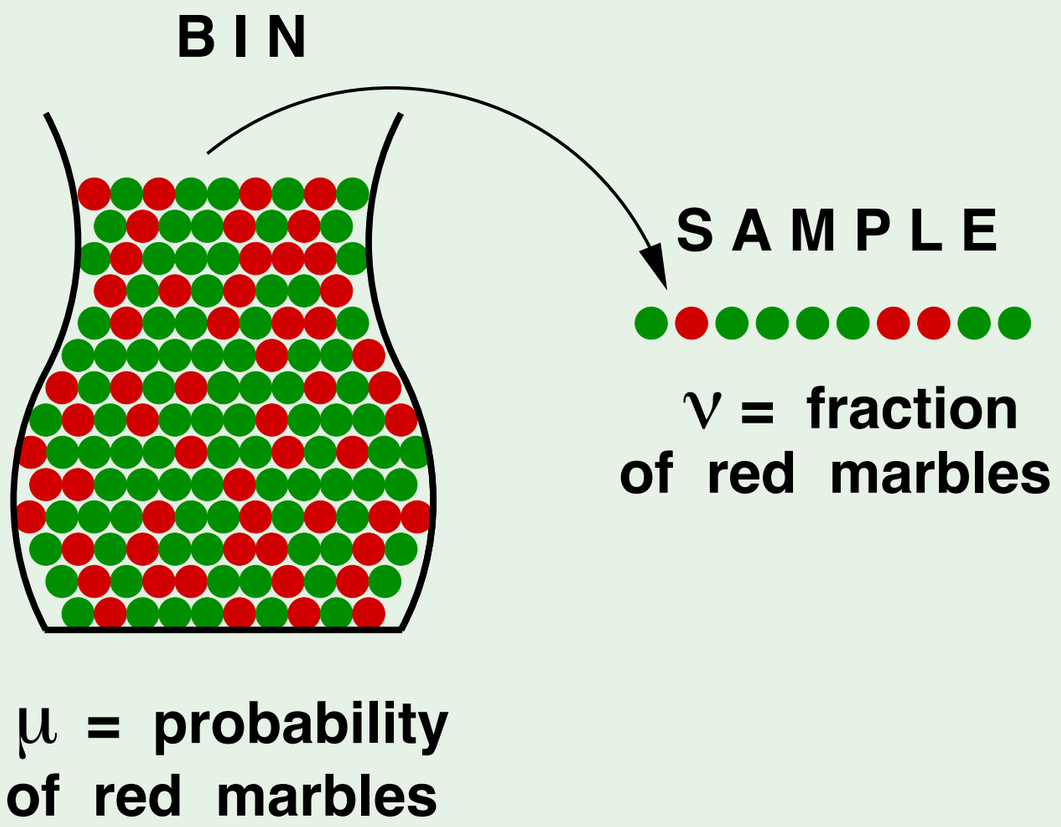
$$\mathbb{P}[\text{ picking a red marble }] = \mu$$

$$\mathbb{P}[\text{ picking a green marble }] = 1 - \mu$$

- The value of $\mu$ is unknown to us.

- We pick $N$ marbles independently.

- The fraction of red marbles in sample $= \nu$



# Does $\nu$ say anything about $\mu$?

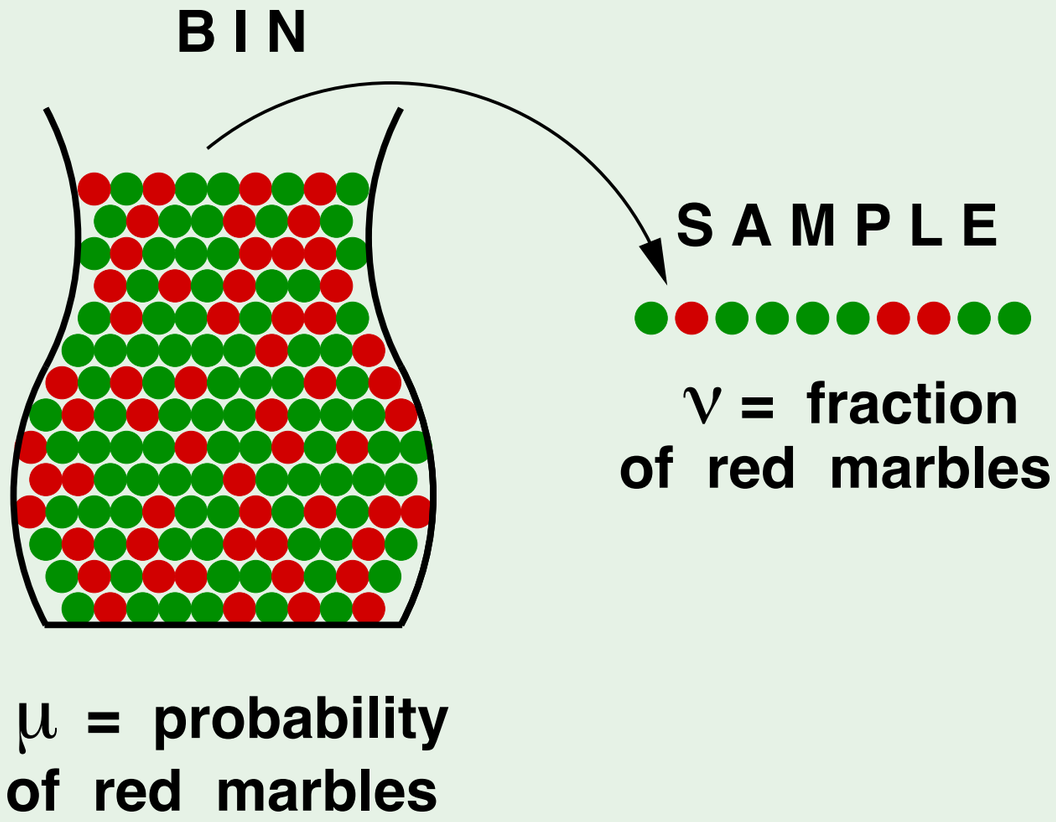
## No!

Sample can be mostly green while bin is mostly red.

## Yes!

Sample frequency $\nu$ is likely close to bin frequency $\mu$.

**possible** versus **probable**



# What does $\nu$ say about $\mu$?

In a big sample (large $N$), $\nu$ is probably close to $\mu$ (within $\epsilon$).

Formally,

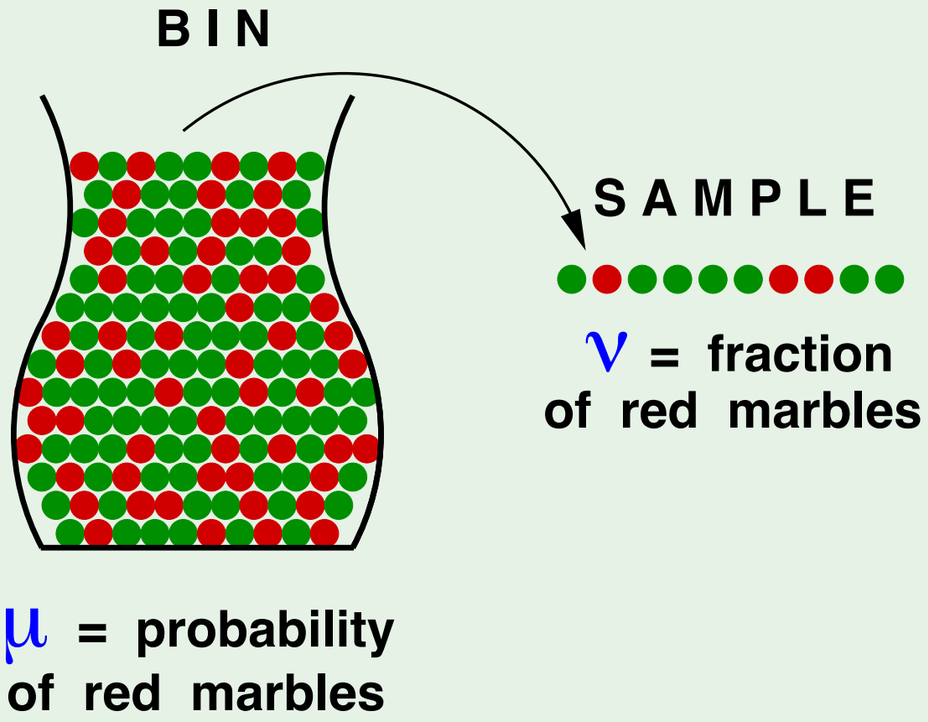
$$\mathbb{P}\left[\left|\nu - \mu\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

This is called **Hoeffding's Inequality**.

In other words, the statement "$\mu = \nu$" is P.A.C.

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$$

- Valid for all $N$ and $\epsilon$
- Bound does not depend on $\mu$
- Tradeoff: $N$, $\epsilon$, and the bound.
- $\nu \approx \mu \implies \mu \approx \nu$ ☺



# Connection to learning

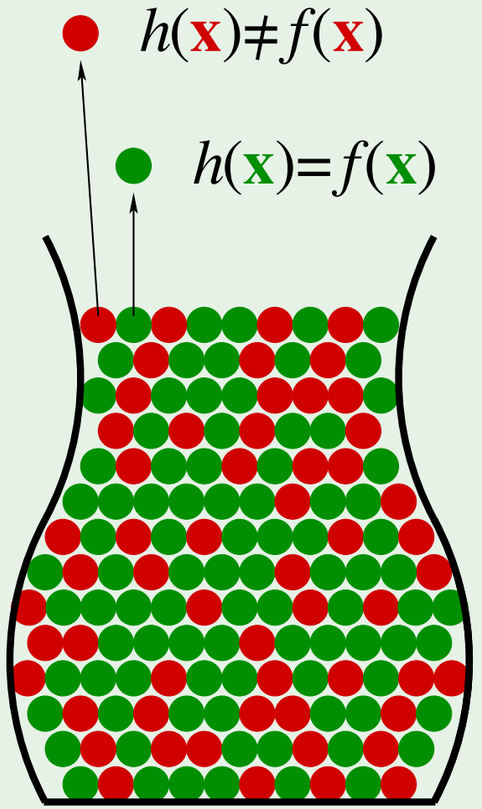
**Bin:** The unknown is a number $\mu$

**Learning:** The unknown is a function $f : \mathcal{X} \to \mathcal{Y}$

Each marble • is a point $\mathbf{x} \in \mathcal{X}$

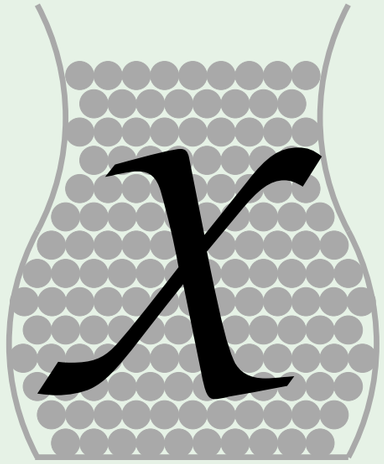
• : Hypothesis got it right $h(\mathbf{x}) = f(\mathbf{x})$

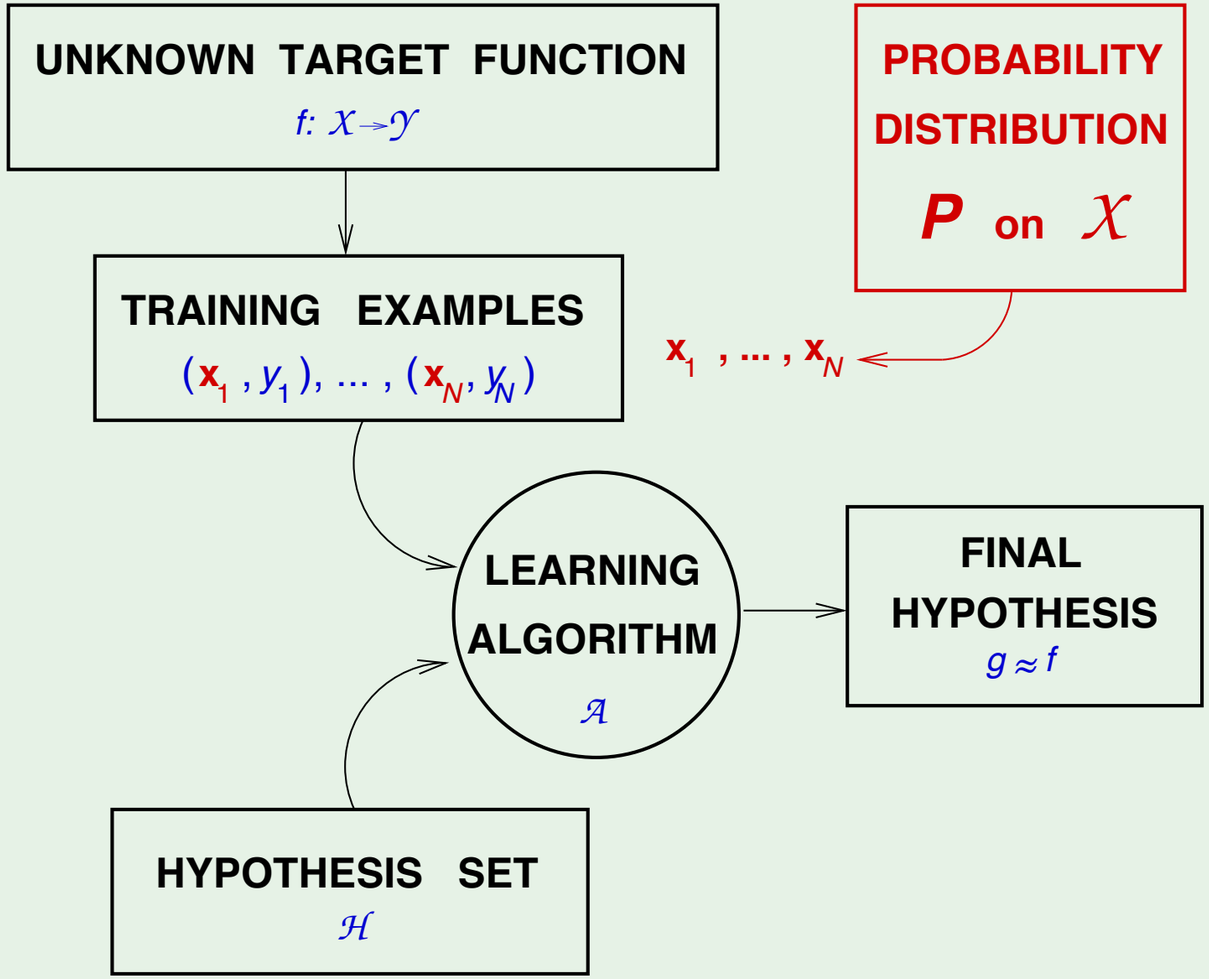
• : Hypothesis got it wrong $h(\mathbf{x}) \neq f(\mathbf{x})$



# Back to the learning diagram

The bin analogy:



# Are we done?

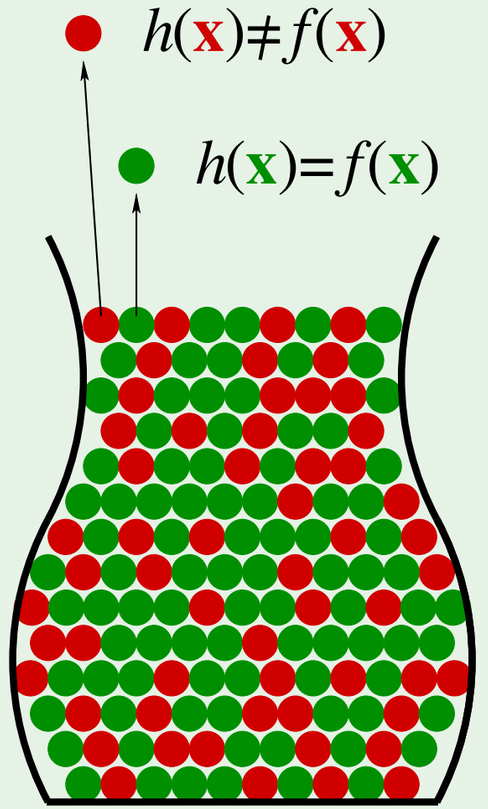
Not so fast! $h$ is fixed.

For this $h$, $\nu$ generalizes to $\mu$.

**'verification'** of $h$, not **learning**

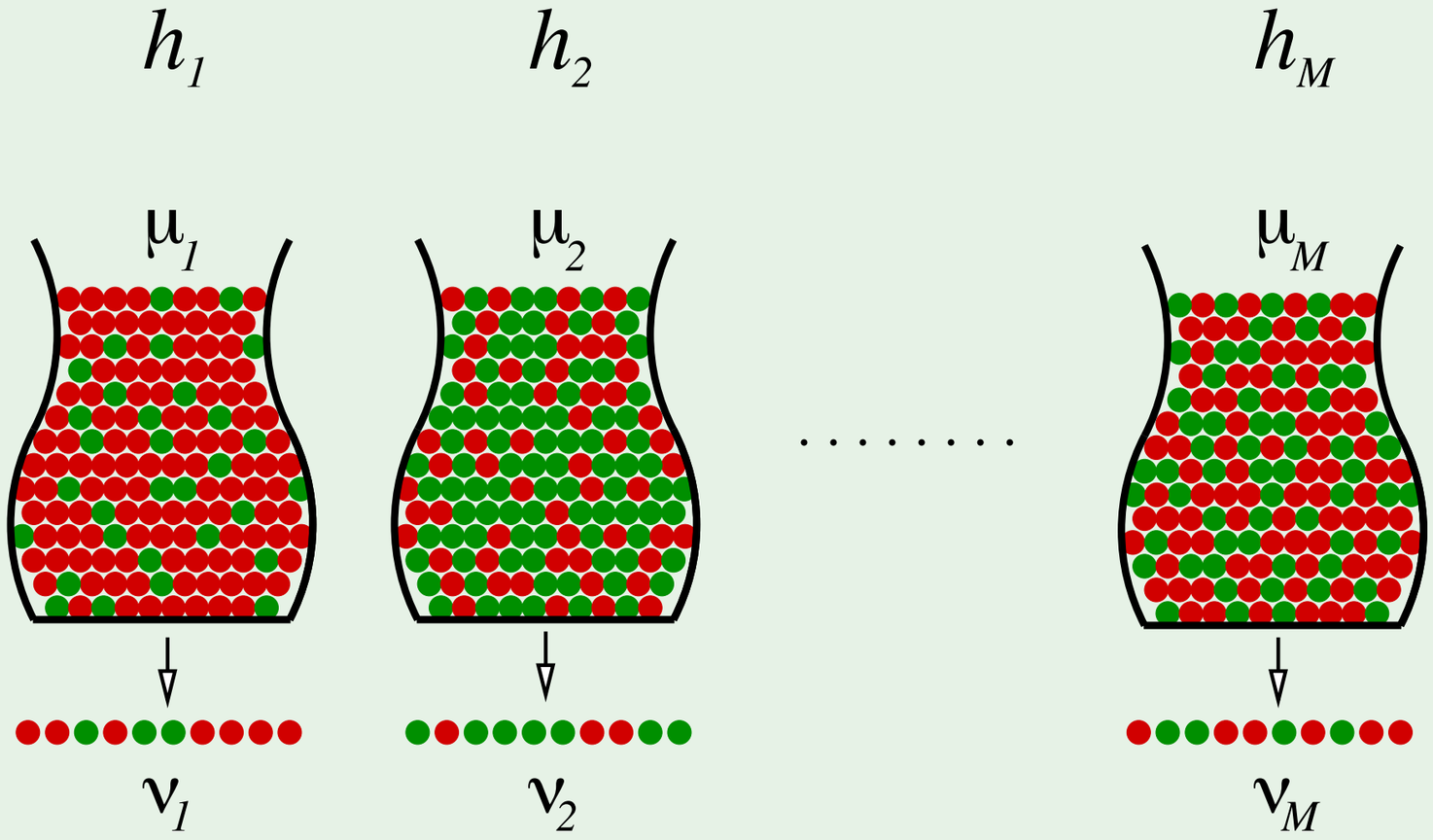
No guarantee $\nu$ will be small.

We need to **choose** from multiple $h$'s.



# Multiple bins

Generalizing the bin model to more than one hypothesis:



# Notation for learning

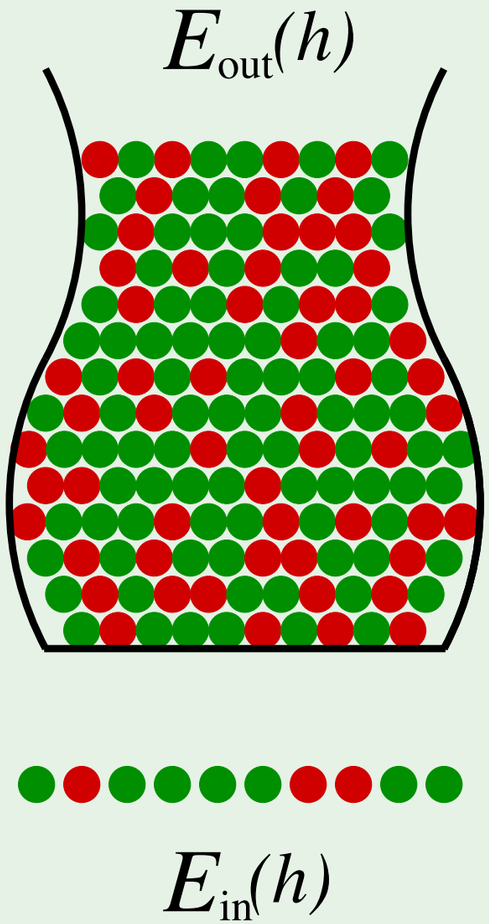
Both $\mu$ and $\nu$ depend on which hypothesis $h$

$\nu$ is '**in sample**' denoted by $E_{\text{in}}(h)$

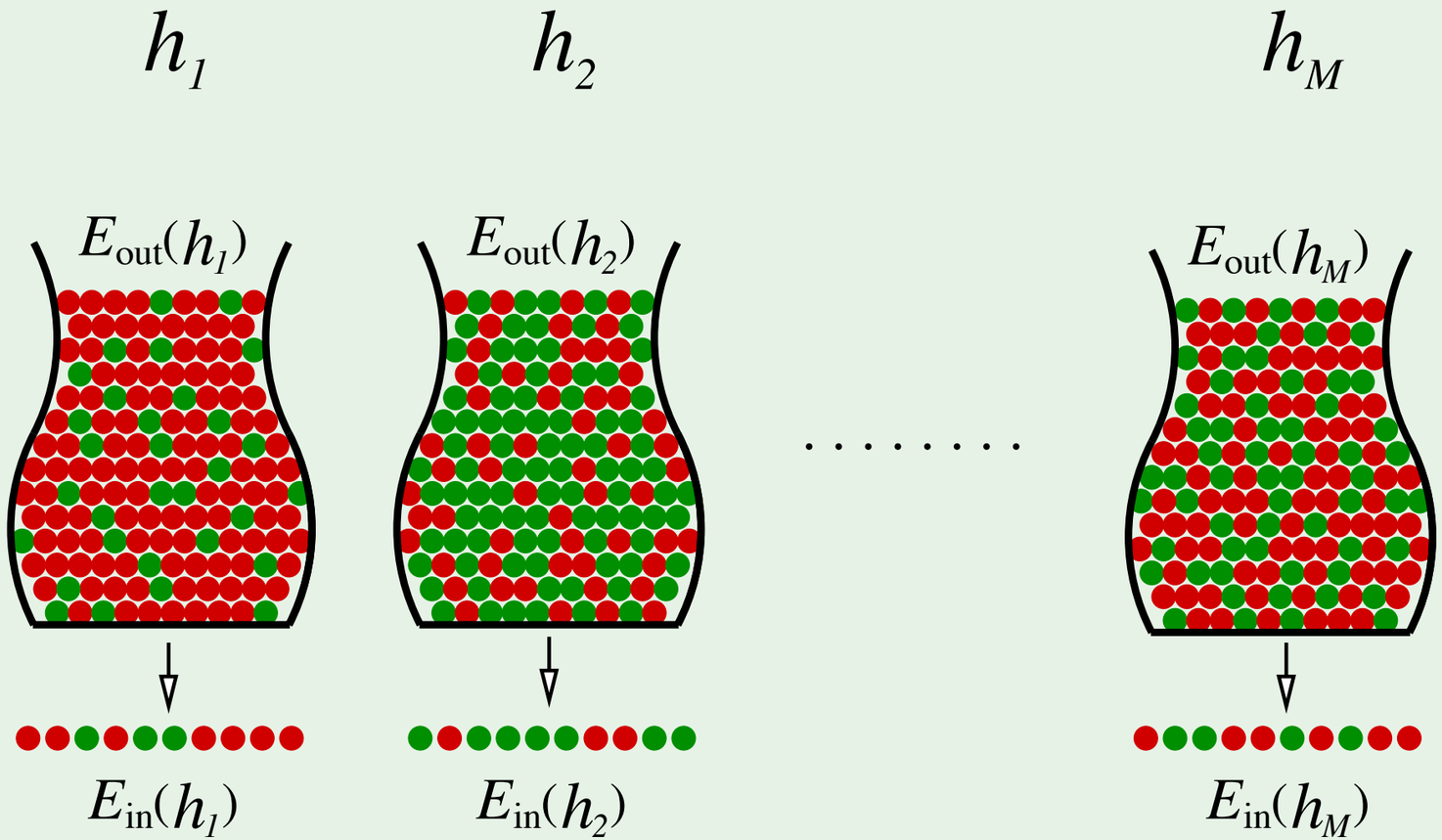
$\mu$ is '**out of sample**' denoted by $E_{\text{out}}(h)$

The Hoeffding inequality becomes:

$$\mathbb{P}\left[\ |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\ \right] \ \leq\ 2e^{-2\epsilon^2 N}$$



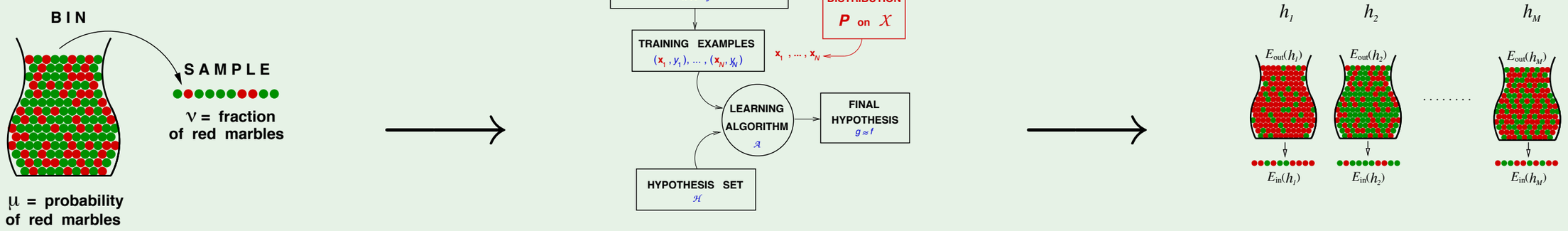
# Notation with multiple bins



# Are we done already? ☺

Not so fast!! Hoeffding doesn't apply to multiple bins.

# What?



# Coin analogy

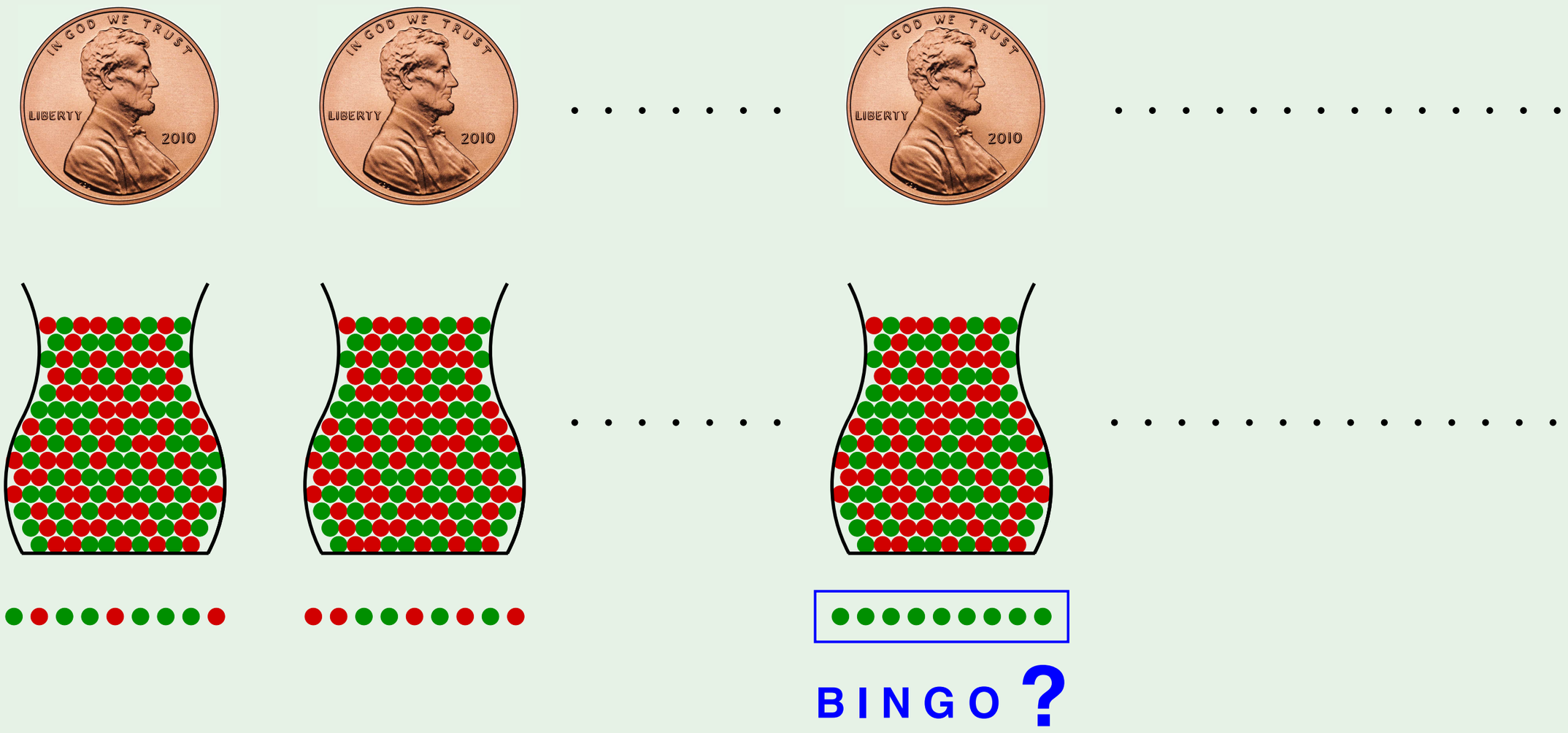
**Question:** If you toss a fair coin 10 times, what is the probability that you will get 10 heads?

**Answer:** $\approx 0.1\%$

**Question:** If you toss 1000 fair coins 10 times each, what is the probability that some coin will get 10 heads?

**Answer:** $\approx 63\%$

# From coins to learning



# A simple solution

$$\begin{aligned}
\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq\ \mathbb{P}[\ & |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\
& \textbf{or}\ |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\
& \dots \\
& \textbf{or}\ |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon\ ] \\
\leq\ & \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\right]
\end{aligned}$$

# The final verdict

$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq \ \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\right]$$

$$\leq \ \sum_{m=1}^{M} 2e^{-2\epsilon^2 N}$$

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$