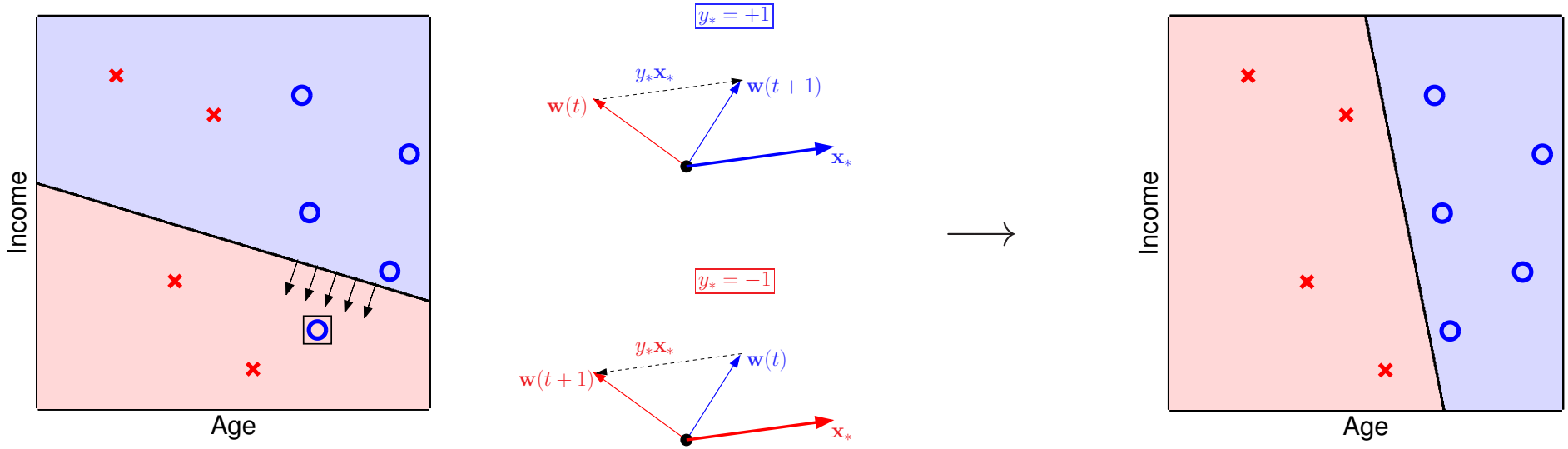
Learning From Data
Lecture 3
Is Learning Feasible?

Outside the Data
Probability to the Rescue
Learning vs. Verification
Selection Bias - A Cartoon

M. Magdon-Ismail
CSCI 4100/6100

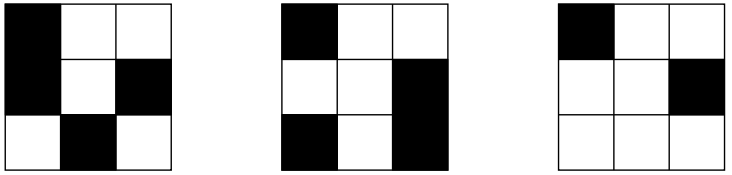
RECAP: The Perceptron Learning Algorithm



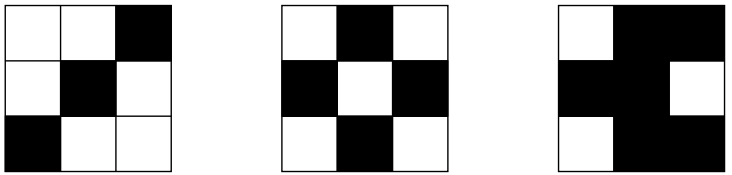
PLA finds a linear separator in finite time.

- What if data is not linearly separable?
- We want $g \approx f$
  - Separating the data amounts to "memorizing the data": $g \approx f$ only on $\mathcal{D}$.
  - $g \approx f$ means we are interested in *outside the data*.

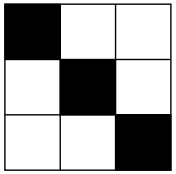
Outside the Data Set



$f = -1$

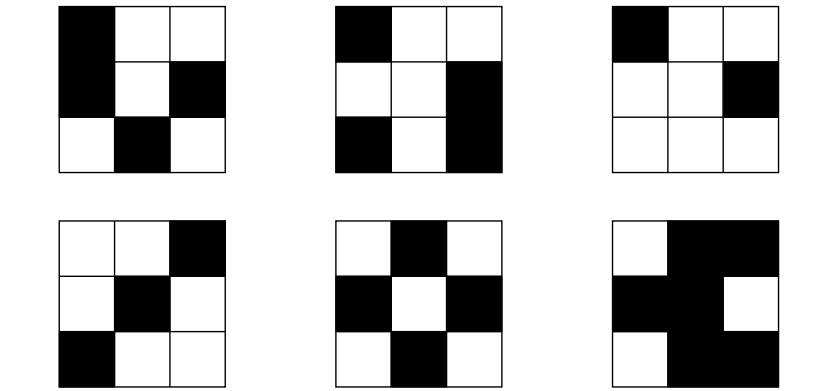


$f = +1$



$f = ?$

Outside the Data Set



$f = -1$

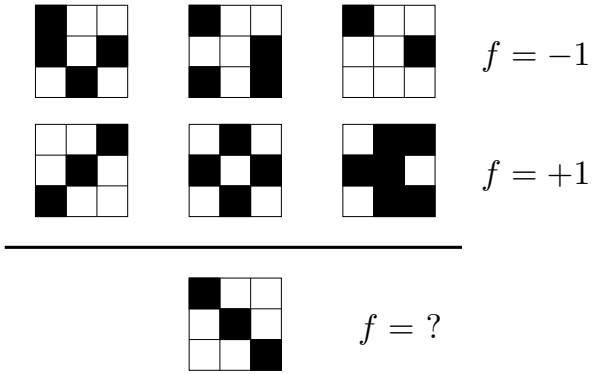
$f = +1$

$f = ?$

- Did you say $f = +1$? (f is measuring symmetry.)
- Did you say $f = -1$? (f only cares about the top left pixel.)

Who is correct? – we cannot *rule out either possibility.*

Outside the Data Set



- An easy visual learning problem just got very messy.

  For *every* f that fits the data and is "+1" on the new point, there is one that is "−1".

  Since f is *unknown*, it can take on any value outside the data, no matter how large the data.

- This is called **No Free Lunch (NFL)**.

  You cannot know anything *for sure* about f outside the data without making assumptions.

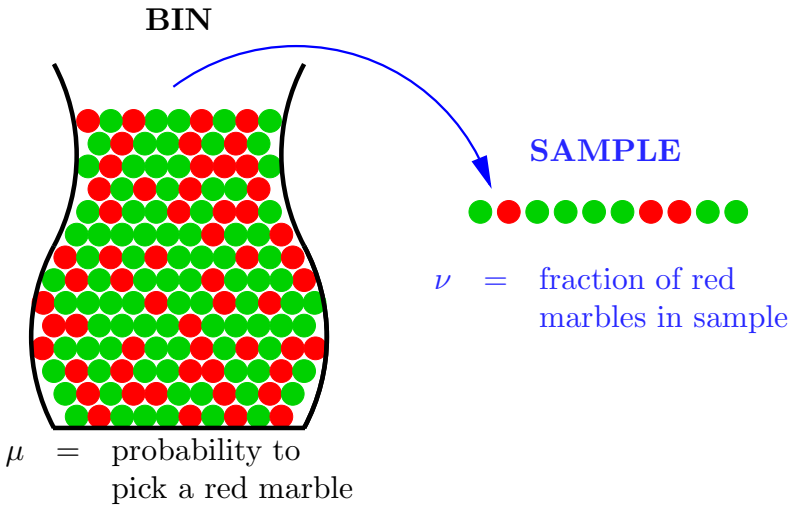
- **What now!**

  Is there *any hope* to know *anything* about f outside the data set *without* making assumptions about f?

Yes, if we are willing to give up the "for sure".

Can we infer *something* outside the data using only $\mathcal{D}$?

Population Mean from Sample Mean



The BIN Model

- Bin with red and green marbles.
- Pick a sample of N marbles *independently*.
- μ: probability to pick a red marble.
  ν: fraction of red marbles in the sample.

Sample $\longrightarrow$ the data set $\longrightarrow$ ν

BIN $\longrightarrow$ outside the data $\longrightarrow$ μ

Can we say anything about $\boldsymbol{\mu}$ **(outside the data)** after observing $\boldsymbol{\nu}$ **(the data)**?

ANSWER: No. It is *possible* for the sample to be all green marbles and the bin to be mostly red.

Then, why do we trust polling (e.g. to predict the outcome of the presidential election).

ANSWER: The bad case is *possible*, but not **probable**.

Probability to the Rescue: Hoeffding's Inequality

Hoeffding/Chernoff proved that, most of the time, ν cannot be too far from μ:

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$

$$\mathbb{P}\left[|\nu - \mu| \leq \epsilon\right] \geq 1 - 2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$

box it and memorize it 

We get to select any ϵ we want.

newsflash: $\nu \approx \mu \implies \mu \approx \nu$ ☺.
$\mu \approx \nu$ is *probably approximately correct* (PAC-learning)

Probability to the Rescue: Hoeffding's Inequality

$$\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$

$$\mathbb{P}\left[|\nu - \mu| \leq \epsilon\right] \geq 1 - 2e^{-2\epsilon^2 N}, \qquad \text{for any } \epsilon > 0.$$

box it and memorize it 

Example: $N = 1,000$; draw a sample and observe ν.

99% of the time	$\mu - 0.05 \leq \nu \leq \mu + 0.05$	$(\epsilon = 0.05)$
99.9999996% of the time	$\mu - 0.10 \leq \nu \leq \mu + 0.10$	$(\epsilon = 0.10)$

What does this mean? If I repeatedly pick a sample of size 1,000, observe ν and claim that

$$\mu \in [\nu - 0.05, \nu + 0.05], \qquad \text{(the error bar is } \pm 0.05\text{)}$$

I will be right 99% of the time. On any particular sample you **may** be wrong, but not often.

We learned *something*. From ν, we reached outside the data to μ.

How Did Probability Rescue Us?

- Key ingredient samples must be *independent*.

  If the sample is constructed in some arbitrary fashion, then indeed we cannot say anything.

  Even with independence, ν can take on arbitrary values; but some values are way more likely than others.

  This is what allows us to learn *something* – it is likely that $\nu \approx \mu$.

- The bound $2e^{-2\epsilon^2 N}$ does not depend on μ or the size of the bin

  The bin can be infinite.

  It's great that it does not depend on μ because μ is unknown; and we mean *unknown*.

- The key player in the bound $2e^{-2\epsilon^2 N}$ is N.

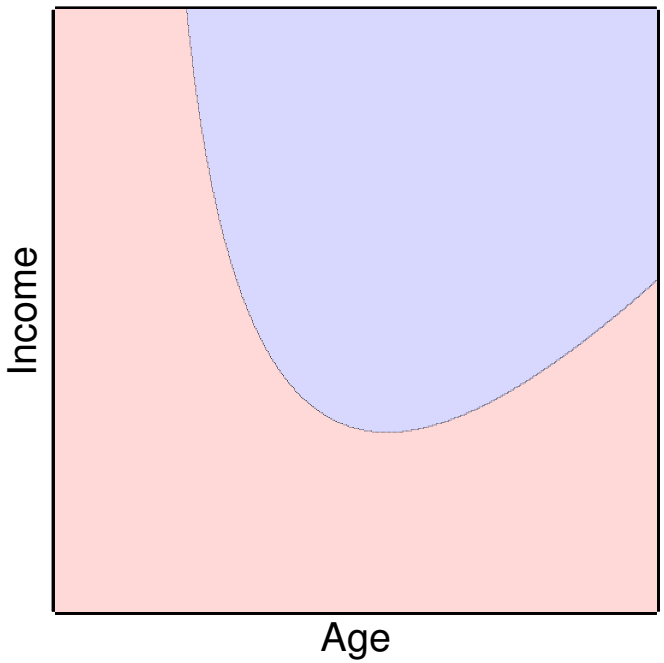
  If $N \to \infty$, $\mu \approx \nu$ with very very very ... high probabilty, *but not for sure*.

  Can you live with 10^{-100} probability of error?

We should *probably* have said *"independence to the rescue"* ☺

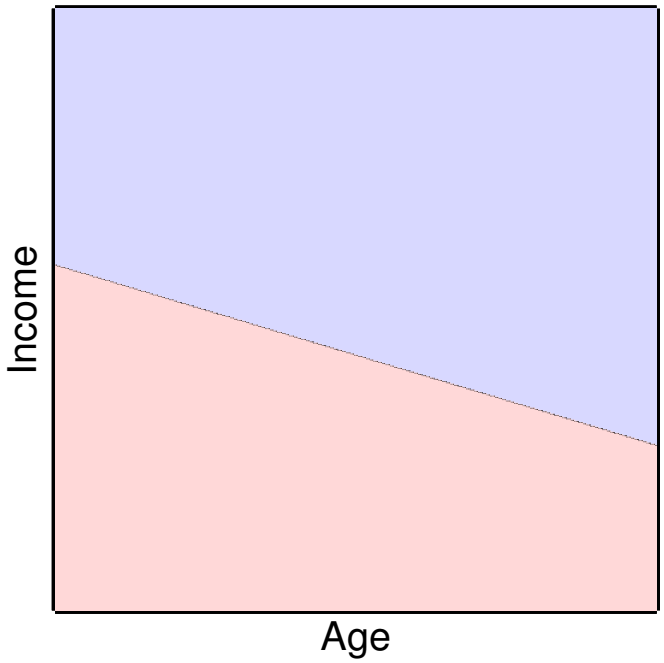
Relating the Bin to Learning

Target Function f



UNKNOWN

Fixed hypothesis h

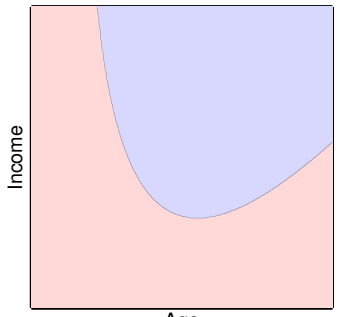


KNOWN

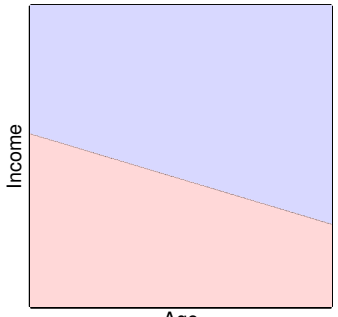
In learning, the unknown is an entire function f; in the bin it was a single number μ.

Relating the Bin to Learning - The Error Function

Target Function f



Fixed hypothesis h



UNKNOWN

green: $h(\mathbf{x}) = f(\mathbf{x})$

red: $h(\mathbf{x}) \neq f(\mathbf{x})$

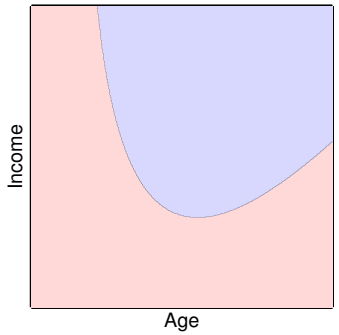
$$E(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

("size" of the red region)

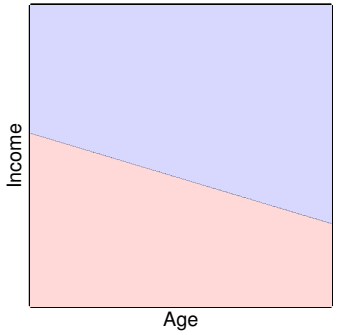
$P(\mathbf{x})$

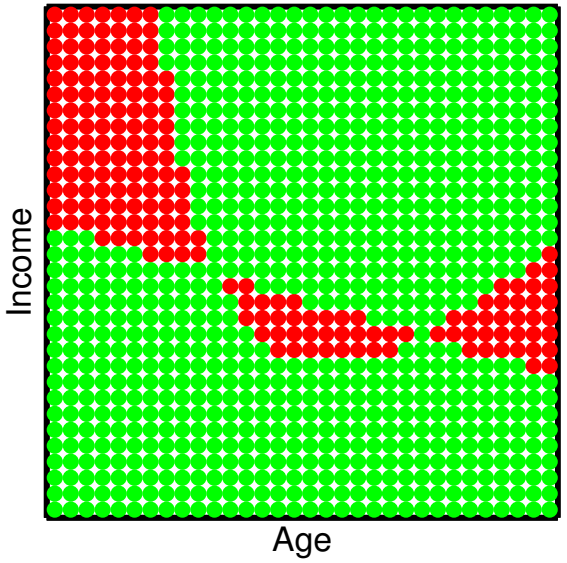
Relating the Bin to Learning - The Error Function

Target Function f



Fixed a hypothesis h





green "marble": $h(\mathbf{x}) = f(\mathbf{x})$

red "marble": $h(\mathbf{x}) \neq f(\mathbf{x})$

BIN: $\mathcal{X}$

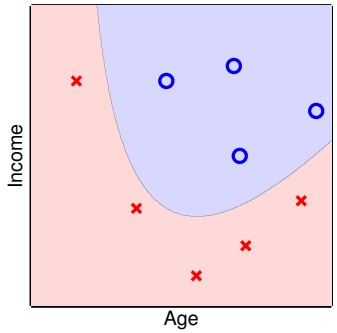
$$E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

out-of-sample

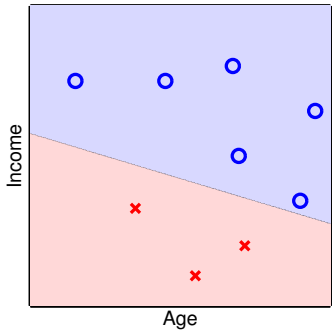
UNKNOWN

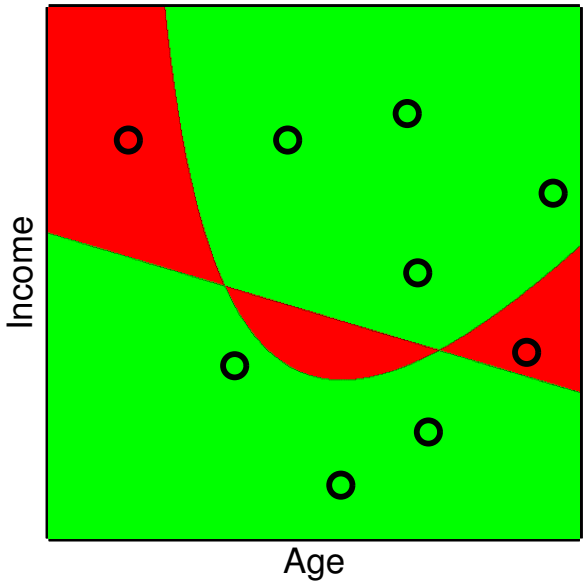
Relating the Bin to Learning - the Data

Target Function f

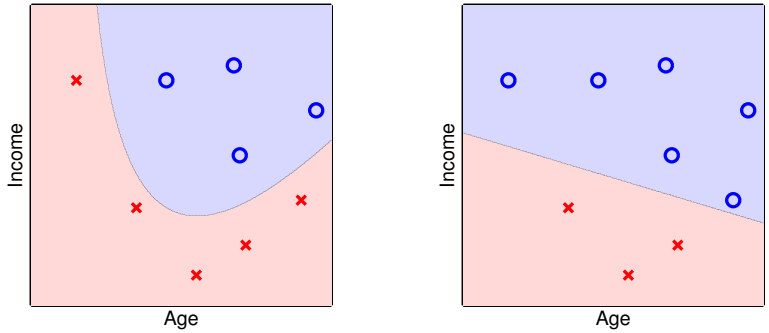


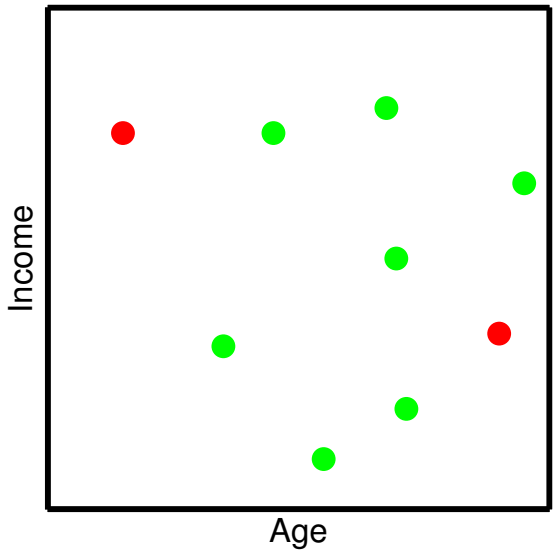
Fixed a hypothesis h





Relating the Bin to Learning - the Data





KNOWN!

green data: $h(\mathbf{x}_n) = f(\mathbf{x}_n)$

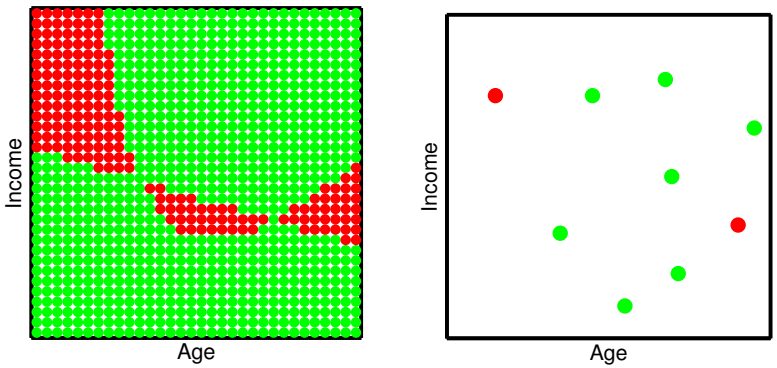
red data: $h(\mathbf{x}_n) \neq f(\mathbf{x}_n)$

$E_{\text{in}}(h)$ = fraction of red data

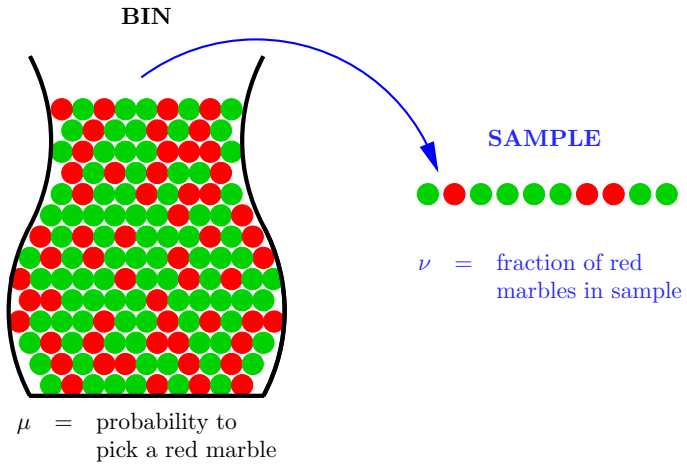
in-sample

misclassified

Relating the Bin to Learning



Unknown f and $P(\mathbf{x})$, **fixed** h



Learning

input space $\mathcal{X}$

$\mathbf{x}$ for which $h(\mathbf{x}) = f(\mathbf{x})$

$\mathbf{x}$ for which $h(\mathbf{x}) \neq f(\mathbf{x})$

$P(\mathbf{x})$

data set $\mathcal{D}$

Out-of-sample Error: $E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})]$

In-sample Error: $E_{\text{in}}(h) = \frac{1}{N} \sum_{n=1}^{N} [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$

Bin Model

Bin

green marble

red marble

randomly picking a marble

sample of N marbles

μ = probability of picking a red marble

ν = fraction of red marbles in the sample

Hoeffding says that $E_{\text{in}}(h) \approx E_{\text{out}}(h)$

$$\mathbb{P}\left[|\boldsymbol{E}_{\text{in}}(\boldsymbol{h}) - \boldsymbol{E}_{\text{out}}(\boldsymbol{h})| > \boldsymbol{\epsilon}\right] \leq \boldsymbol{2e^{-2\epsilon^2 N}}, \qquad \textbf{for any } \boldsymbol{\epsilon > 0}.$$

$$\mathbb{P}\left[|\boldsymbol{E}_{\text{in}}(\boldsymbol{h}) - \boldsymbol{E}_{\text{out}}(\boldsymbol{h})| \leq \boldsymbol{\epsilon}\right] \geq \boldsymbol{1 - 2e^{-2\epsilon^2 N}}, \qquad \textbf{for any } \boldsymbol{\epsilon > 0}.$$

E_{in} is **random, but known**; E_{out} **fixed, but unknown**.

- If $E_{\text{in}} \approx 0 \implies E_{\text{out}} \approx 0$ (with high probability), i.e. $\mathbb{P}_{\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \approx 0$;
  We have learned something about the *entire* f: $f \approx h$ over $\mathcal{X}$ (outside $\mathcal{D}$)

- If $E_{\text{in}} \gg 0$, we're out of luck.
  But, we have still learned something about the entire f: $f \not\approx h$; it is not very useful though.

Questions:

Suppose that $E_{\text{in}} \approx 1$, have we learned something about the entire f that *is* useful?

What is the worst E_{in} for inferring about f?

That's Verification, not Real Learning

The entire previous argument assumed a **FIXED** h and then came the data.

- *Given* $h \in \mathcal{H}$, a sample can **verify** whether or not it is good (w.r.t. f):

  if E_{in} is small, h is good, with high confidence.

  if E_{in} is large, h is bad with high confidence.

  We have no control over E_{in}. It is what it is.

- In learning, you actually try to **fit** the data, as with the perceptron model

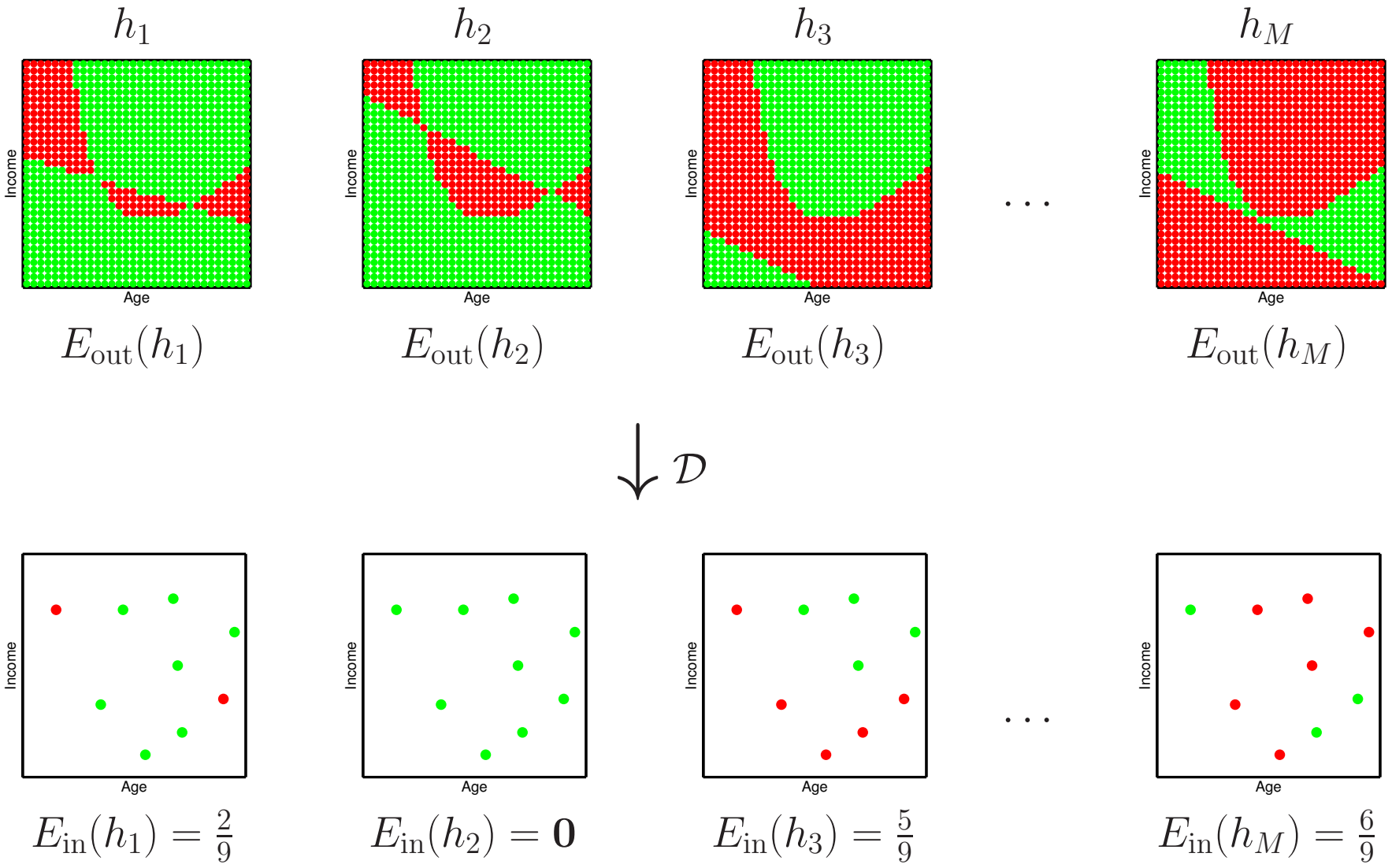
  g results from searching an entire hypothesis set $\mathcal{H}$ for a hypothesis with small E_{in}.

Verification	**Real Learning**
Fixed single hypothesis h	Fixed *hypothesis set* $\mathcal{H}$
h to be certified	g to be certified
h does not depend on $\mathcal{D}$	g results after searching $\mathcal{H}$ to fit $\mathcal{D}$
No control over E_{in}	Pick best E_{in}

Verification: we can say *something* outside the data about h?

Learning: can we say *something* outside the data about g?

Real Learning – Finite Learning Model



Pick the hypothesis with minimum E_{in}; will E_{out} be small?

Selecting the Best Coin

1. Everyone take out a coin.

2. Each of you toss your coin 5 times and count the number of heads.

3. Who got the smallest number of heads (probably 0)?

4. Can I have that coin please?

Is this a Freak Coin?

Do we expect $\mathbb{P}[\text{HEADS}] \approx 0$?

Let's toss this coin (this coin has never come up heads).

HEADS: you give me **$2;**
TAILS: I give you **$1.**

Who wants this bet?

(we're gonna play this game 100 times)

Selection Bias

Coin tossing example:

- If we toss one coin and get no HEADS, its very surprising. $\mathbb{P} = \frac{1}{2^N}$

  We expect it is biased: $\mathbb{P}[\text{heads}] \approx 0$.

- Tossing 70 coins, and *find one* with no heads. Is it surprising? $\mathbb{P} = 1 - \left(1 - \frac{1}{2^N}\right)^{70}$

  Do we expect $\mathbb{P}[\text{heads}] \approx 0$ for the selected coin?

  Similar to the "birthday problem": among 30 people, two will likely share the same birthday.

- This is called *selection bias*.

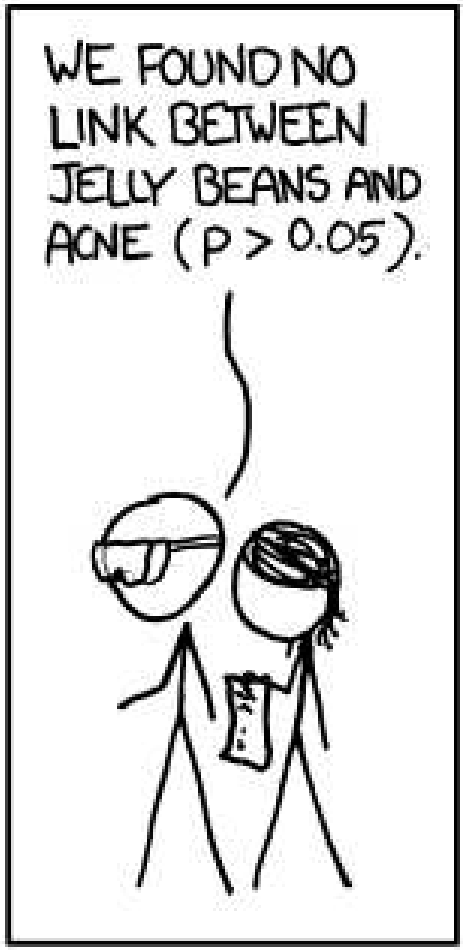
  Selection bias is a very serious trap. For example iterated medical screening.

If we *select* an $h \in \mathcal{H}$ with smallest E_{in}, can we expect E_{out} to be small?

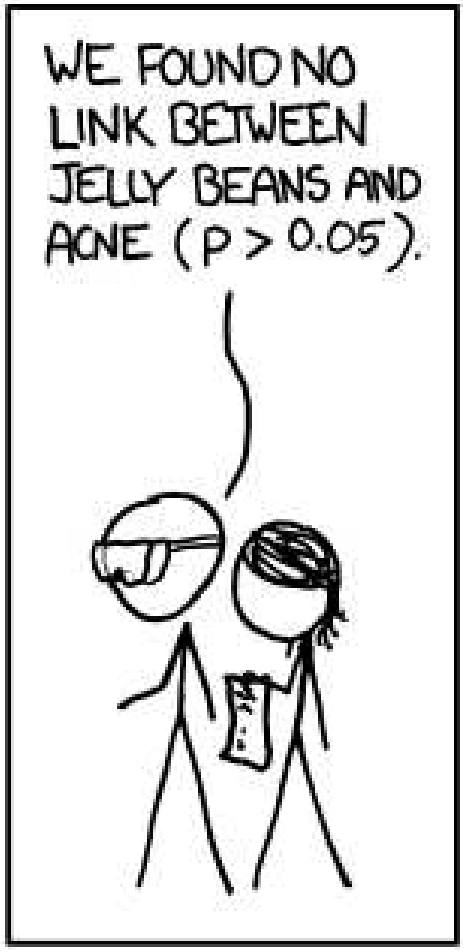
Search Causes Selection Bias

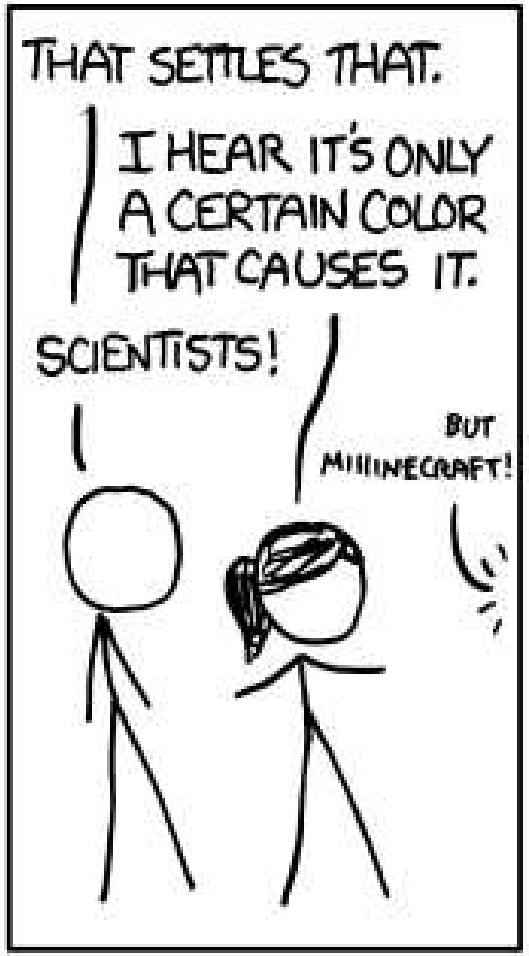
Jelly Beans Cause Acne?

Jelly Beans Cause Acne?

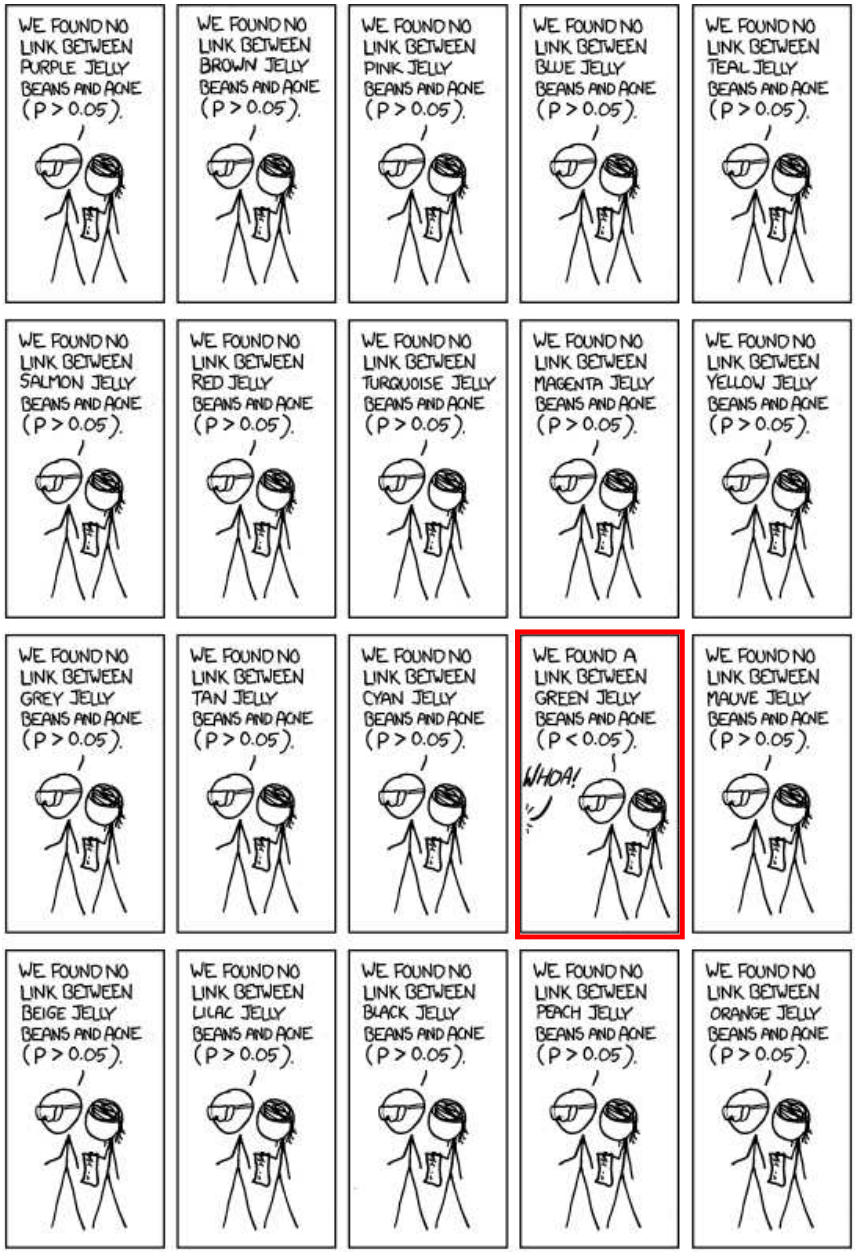


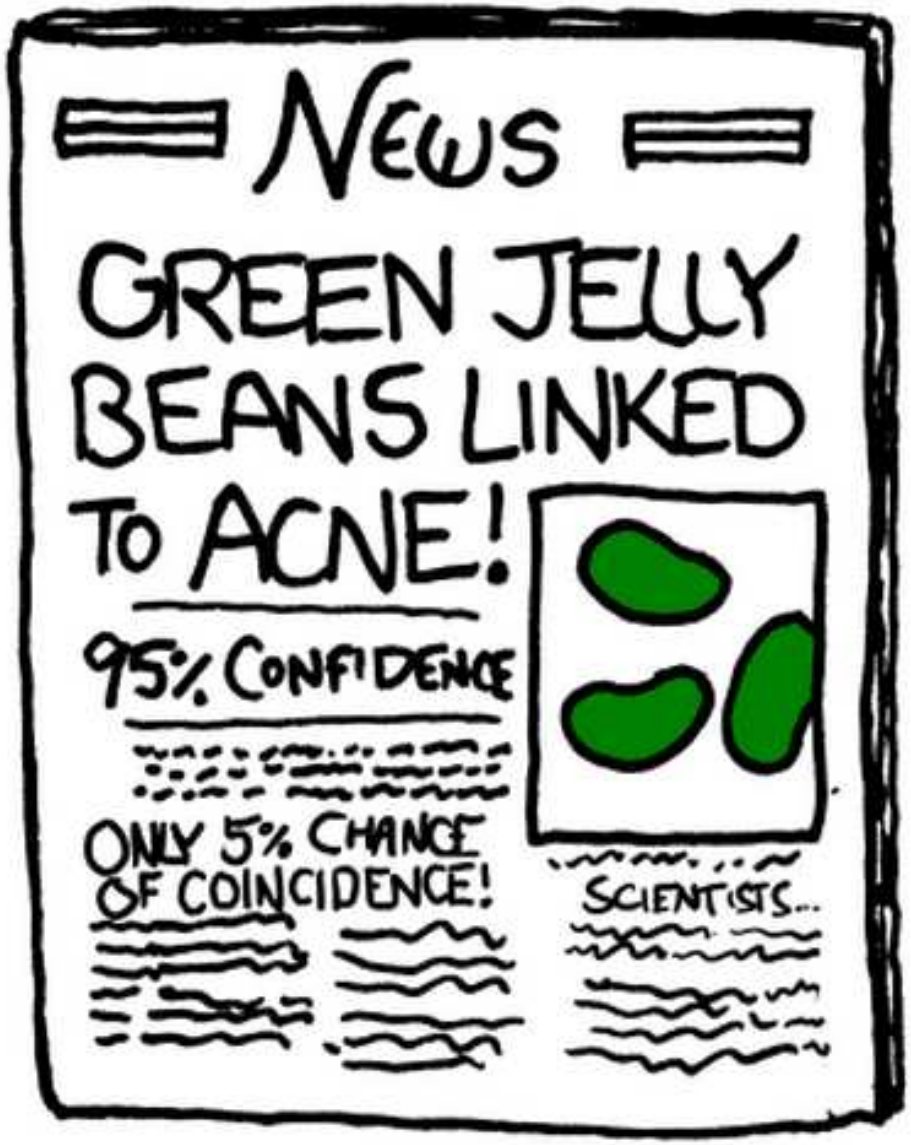
Jelly Beans Cause Acne?





Jelly Beans Cause Acne?




NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...