Machine Learning from Data

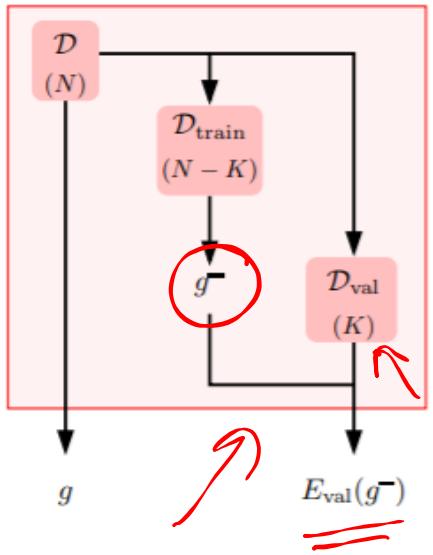
Lecture 14: Spring 2021

Today's Lecture

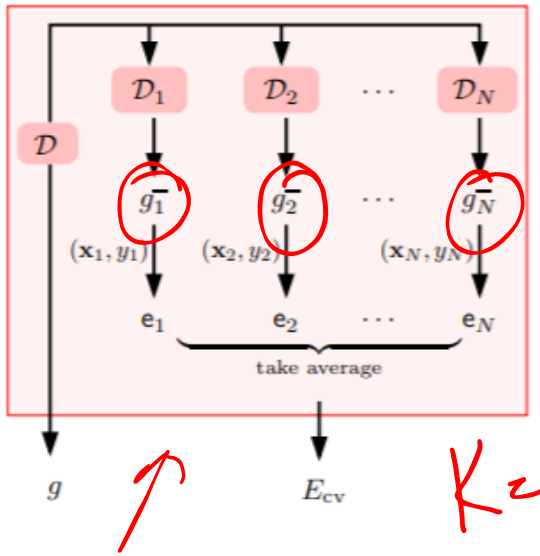
- Three Learning Principles
  - Occam's Razor
  - Sampling Bias
  - Data Snooping

Validation and Cross Validation

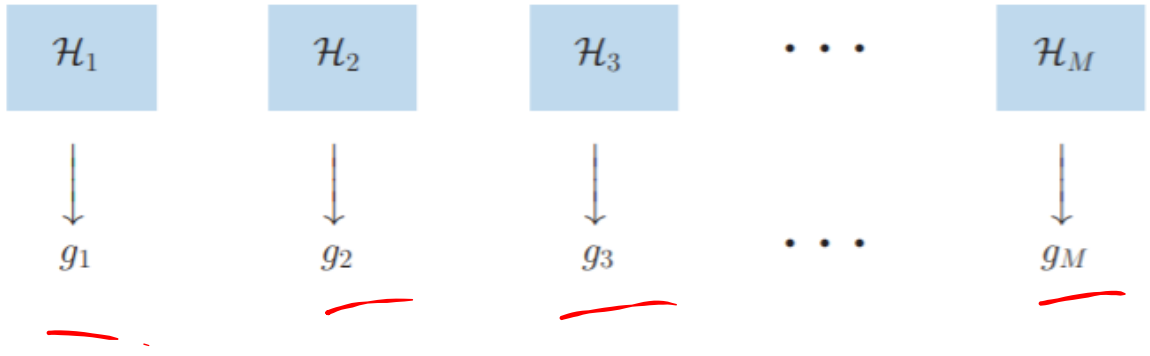
Validation



Cross Validation



Model Selection



Occam's Razor

use a '**razor**' to 'trim down'

"an explanation of the data to make it as simple as possible but no simpler."

attributed to William of Occam (14th Century) and often mistakenly to Einstein

Simpler is Better

$E_{in}(g) = 0$

E_{out}

The simplest model that fits the data is also the most plausible.

Q1. What is simpler?

Q2. How do we know?

dvc

...or, beware of using complex models to fit data

What is Simpler?

simple hypothesis h	simple hypothesis set $\mathcal{H}$
$\Omega(h)$	$\Omega(\mathcal{H})$
low order polynomial	$\mathcal{H}$ with small d_{VC}
hypothesis with small weights	small number of hypotheses
easily described hypothesis	low entropy set
⋮	⋮

The equivalence:

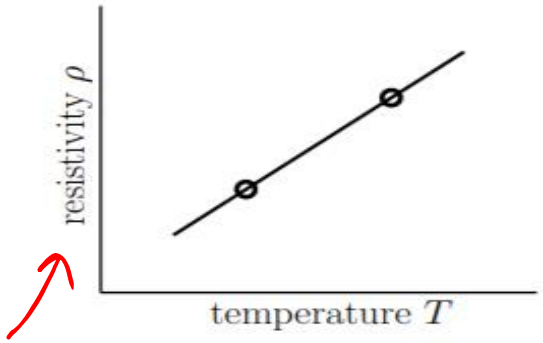
A hypothesis set with **simple** hypotheses *must be* **small**

Why is simpler better?

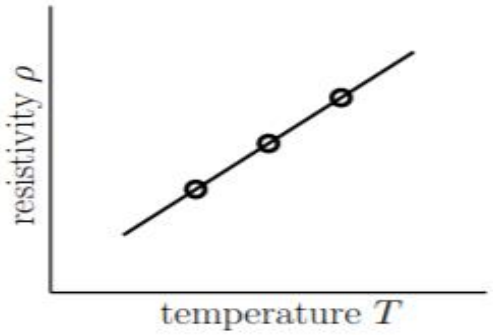
- Because you will be more surprised when you fit the data.
- When something unlikely happens, it is very significant when it happens. ✓

A Scientific Experiment

Scientist 1

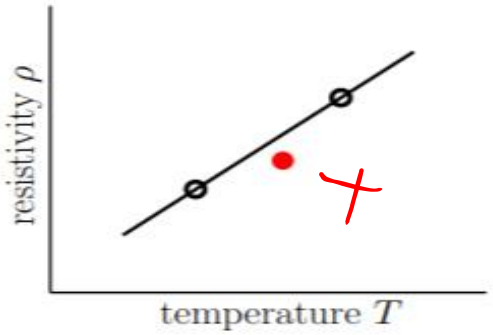


Scientist 2



Very convincing

Scientist 3



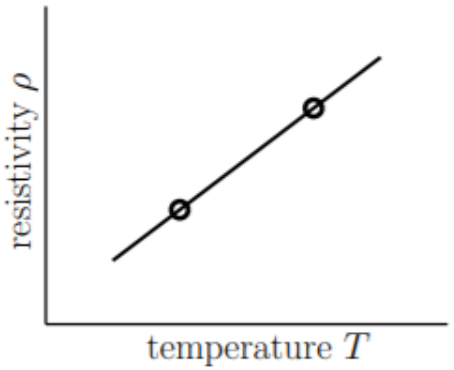
Maybe

Q Who provides most evidence for the hypothesis "ρ is linear in T"?

Axiom of Non-Falsifiability

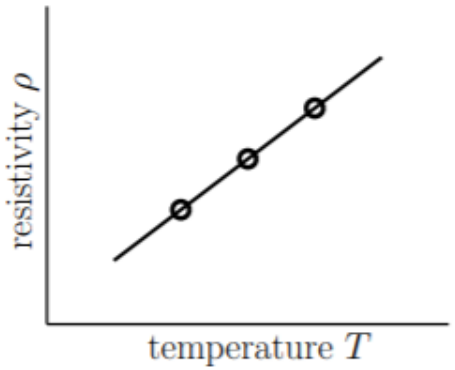
Axiom. If an experiment has no chance of falsifying a hypothesis, then the result of that experiment provides no evidence one way or the other for the hypothesis.

Scientist 1



no evidence

Scientist 2



very convincing

Falsification and $m_{\mathcal{H}}(N)$

If $\mathcal{H}$ shatters $\mathbf{x}_1, \cdots, \mathbf{x}_N$,

– Don't be surprised if you fit the data.

– Can't falsify "$\mathcal{H}$ is a good set of candidate hypotheses for f".

If $\mathcal{H}$ doesn't shatter $\mathbf{x}_1, \cdots, \mathbf{x}_N$, and the target values are uniformly distributed,

$$\mathbb{P}[\text{falsification}] \geq 1 - \frac{m_{\mathcal{H}}(N)}{2^N}.$$

A good fit is surprising with simple $\mathcal{H}$, hence significant. You can, but didn't falsify "$\mathcal{H}$ is a good set of candidate hypotheses for f"

Learning From Data in General

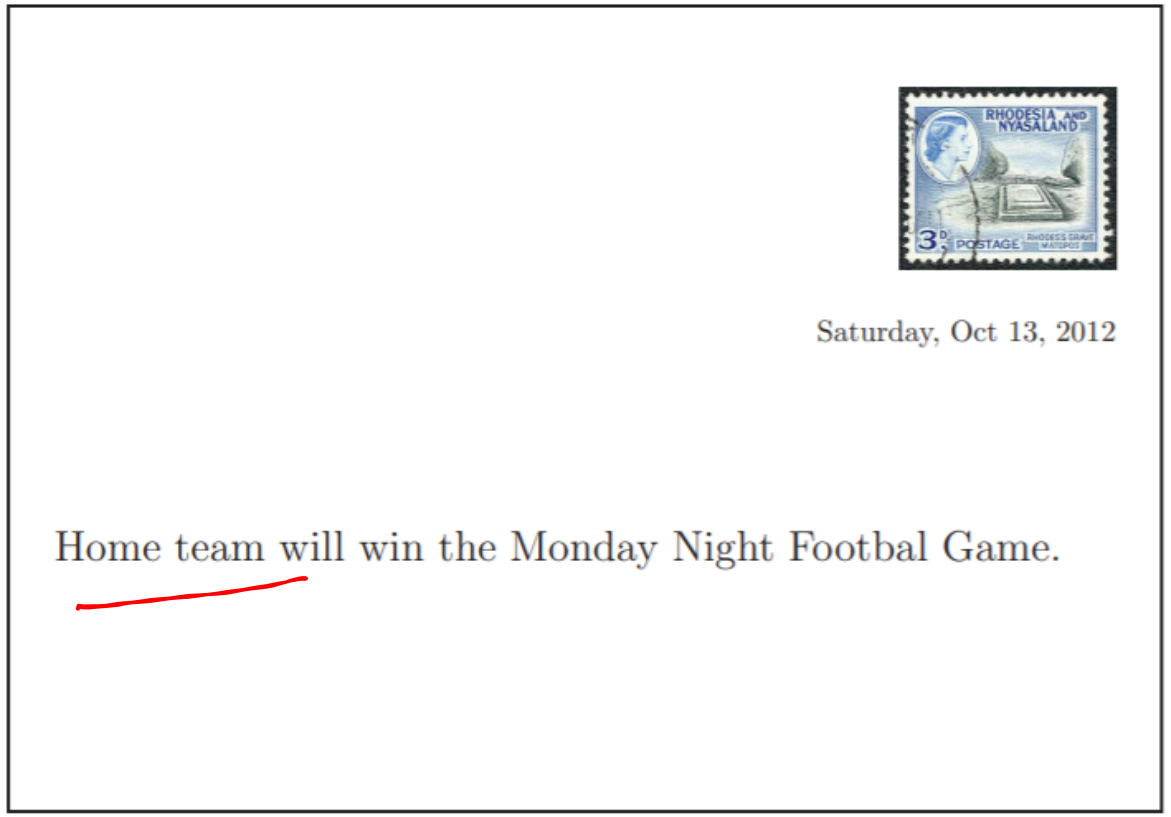
$E_{in} \rightarrow$ small

- We may opt for a simpler fit than possible.
- Imperfect fit with a simpler mode is better than a perfect fit with a more complex model.

Exercise 5.2

Football Oracle

•

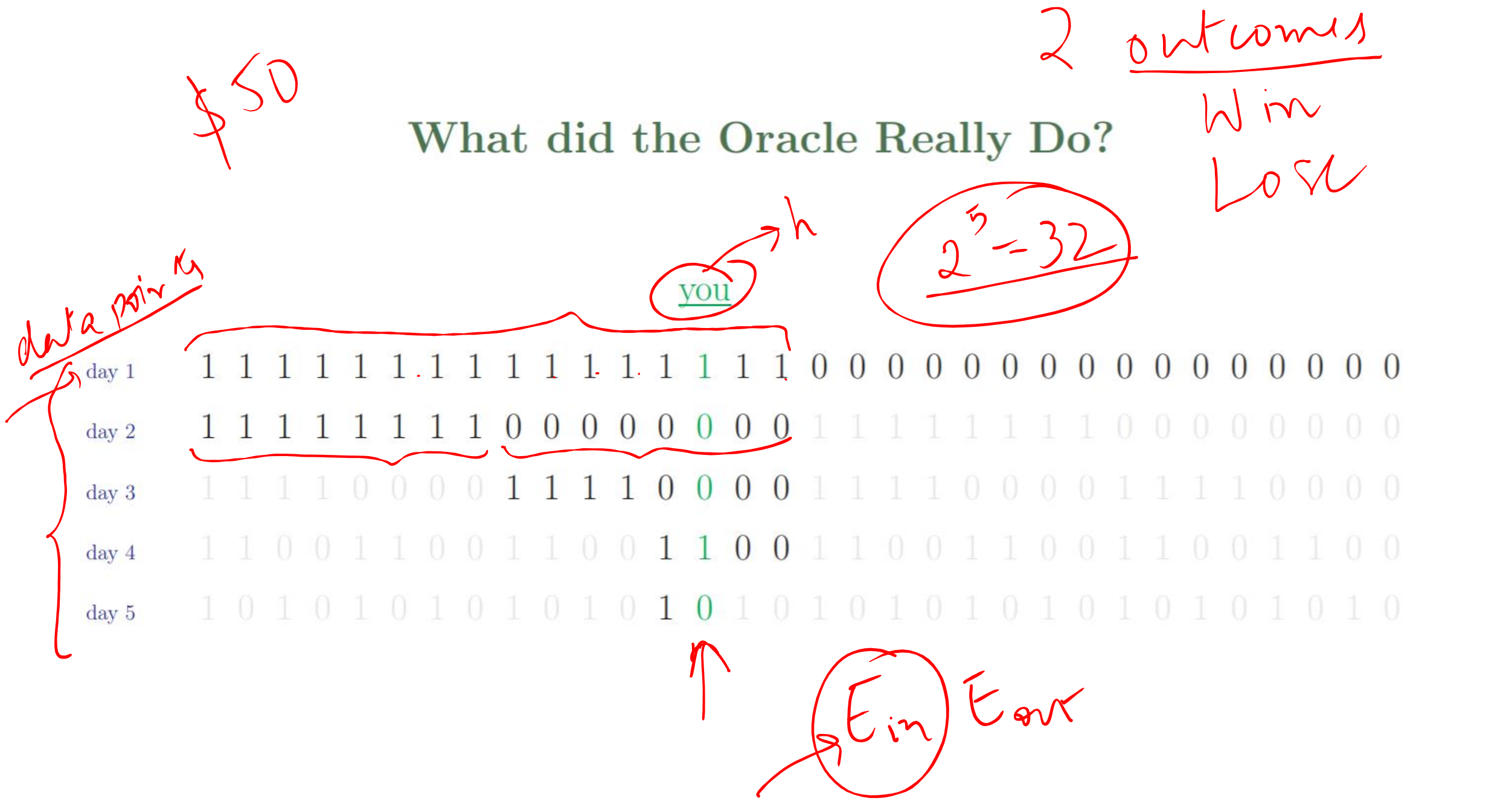


Monday1
Monday2

Prediction
↓
Outcome

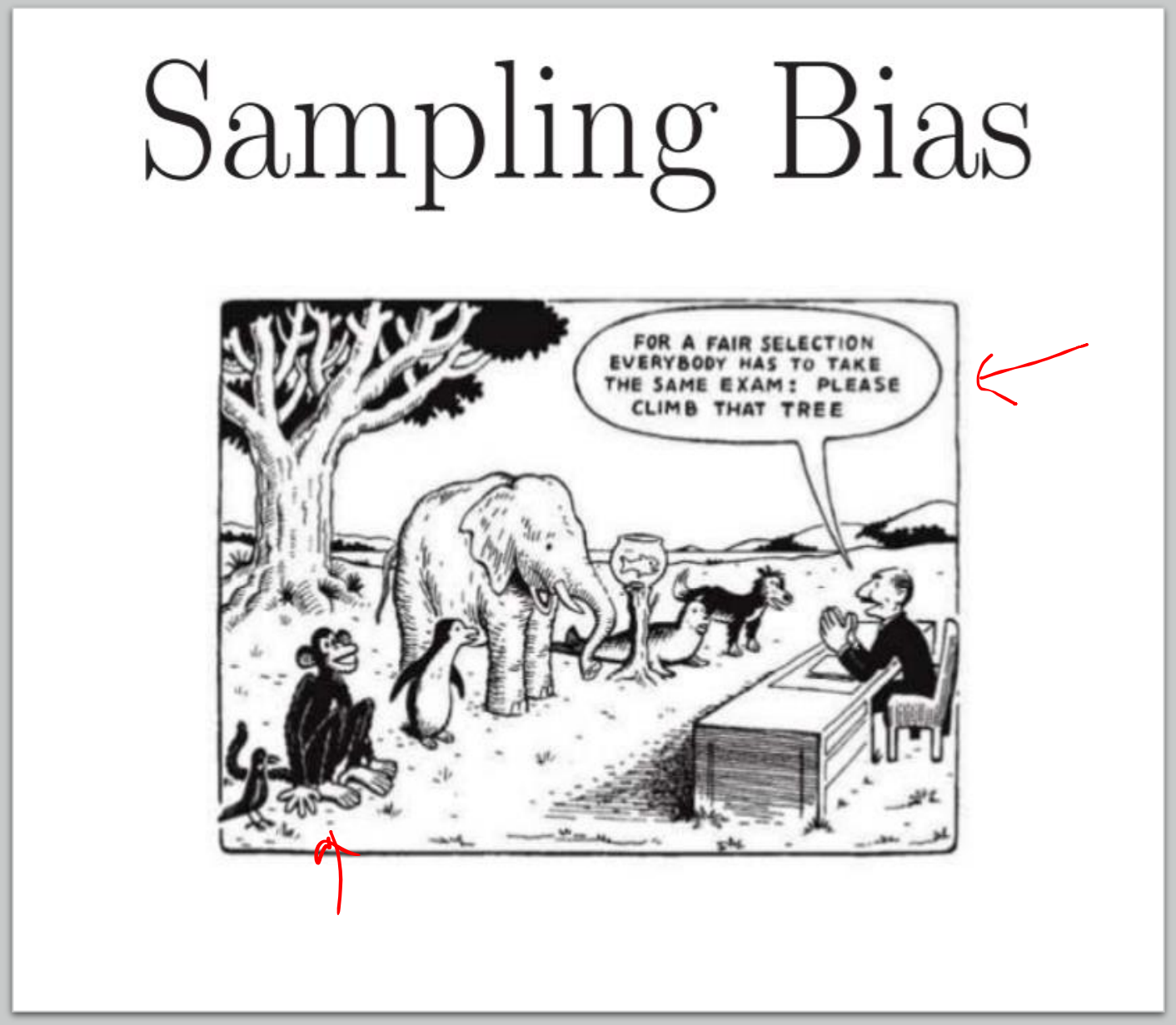
6.th week ⟶ $50!

What did the Oracle Really Do?



Sampling Bias

- If the data is sampled in a biased way, learning will produce a similarly biased outcome.



November 3rd 1948, Dewey Defeats Truman

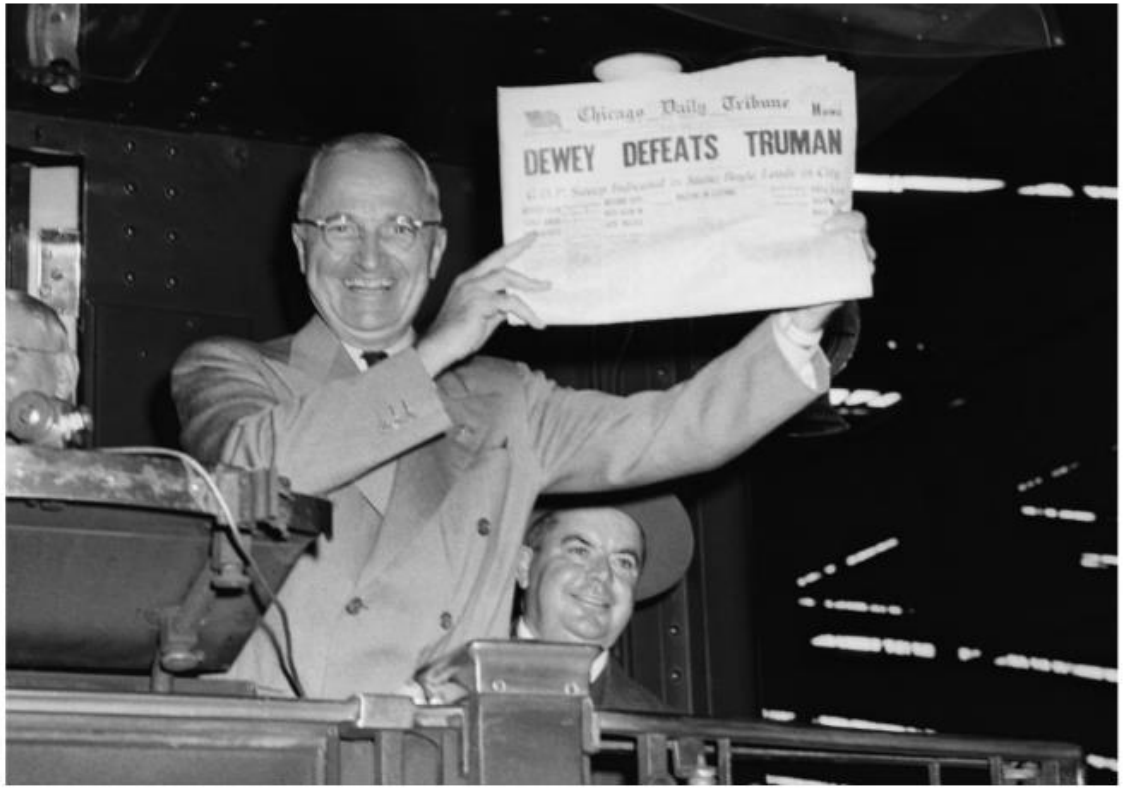
Tribune wanted to show off its latest technology

could go earlier to press.

Telephone poll on how people voted

statisticians had done their thing and were confident.

Imagine Their Surprise When . . .



Telephone poll?

1978

Out of sample.

Sampling Bias in Learning

If the data is sampled in a biased way, learning will produce a similarly biased outcome.

. . . or, make sure the training and test distributions are the same.

You cannot draw a sample from one bin and make claims about *another* bin

Puzzle - Credit Analysis

Selection bias

Applied → denied → system

- Determine credit given salary, debt, years in residence, ...
- Banks have lots of data
  - customer information: salary, debt, etc.
  - whether or not **who?** defaulted on their credit.

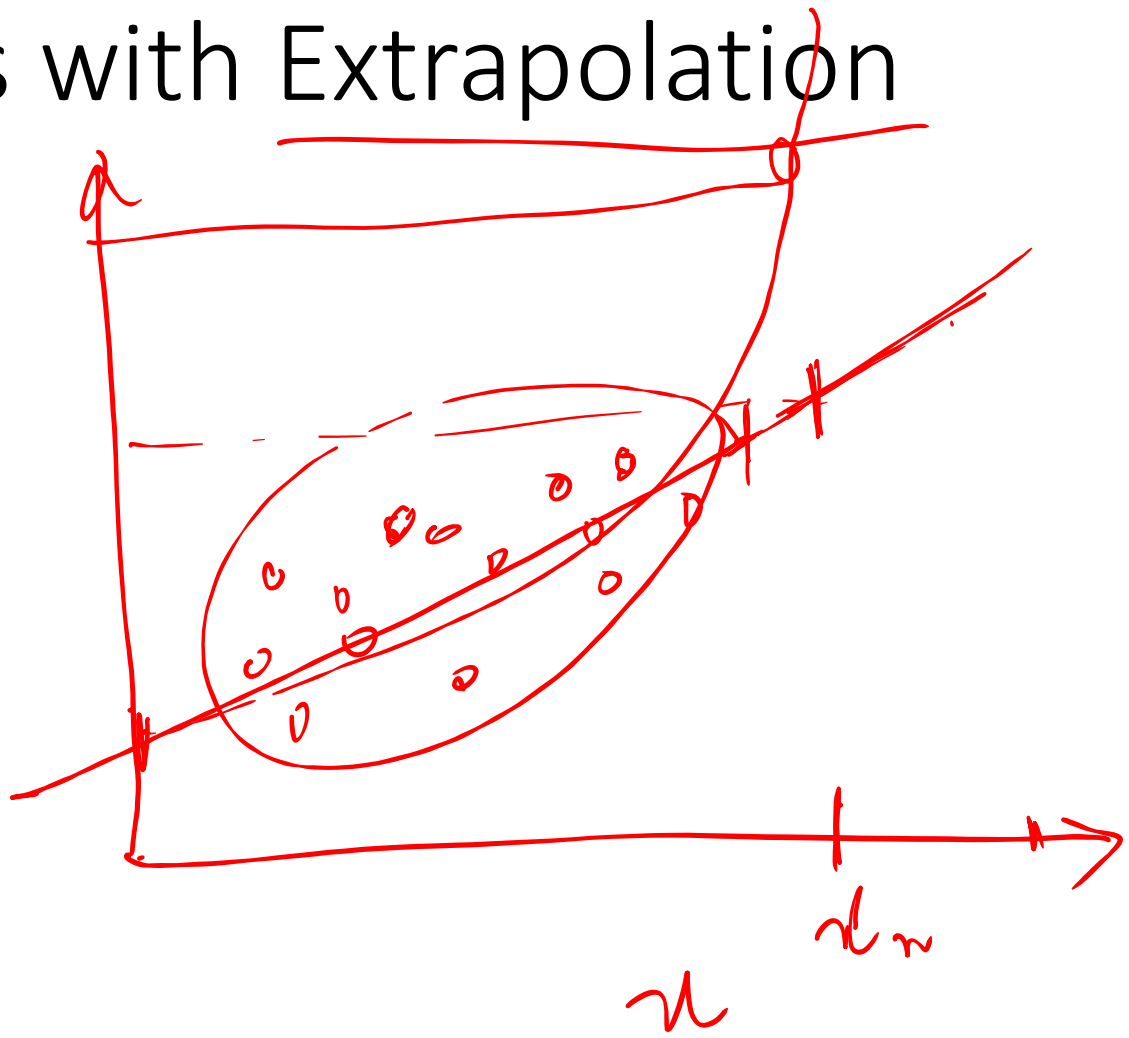
age	32 years
gender	male
salary	40,000
debt	26,000
years in job	1 year
years at home	3 years
...	...

Approve for credit?

predict → action

only data on approved customers

Issues with Extrapolation

- 

Not representative of the data

Data Snooping

Choice → final g

If a data set has affected any step in the learning process, it cannot be fully trusted in assessing the outcome.

Not trustworthy

. . . or, estimate performance with a *completely* uncontaminated test set

. . . and, choose $\mathcal{H}$ **before** looking at the data

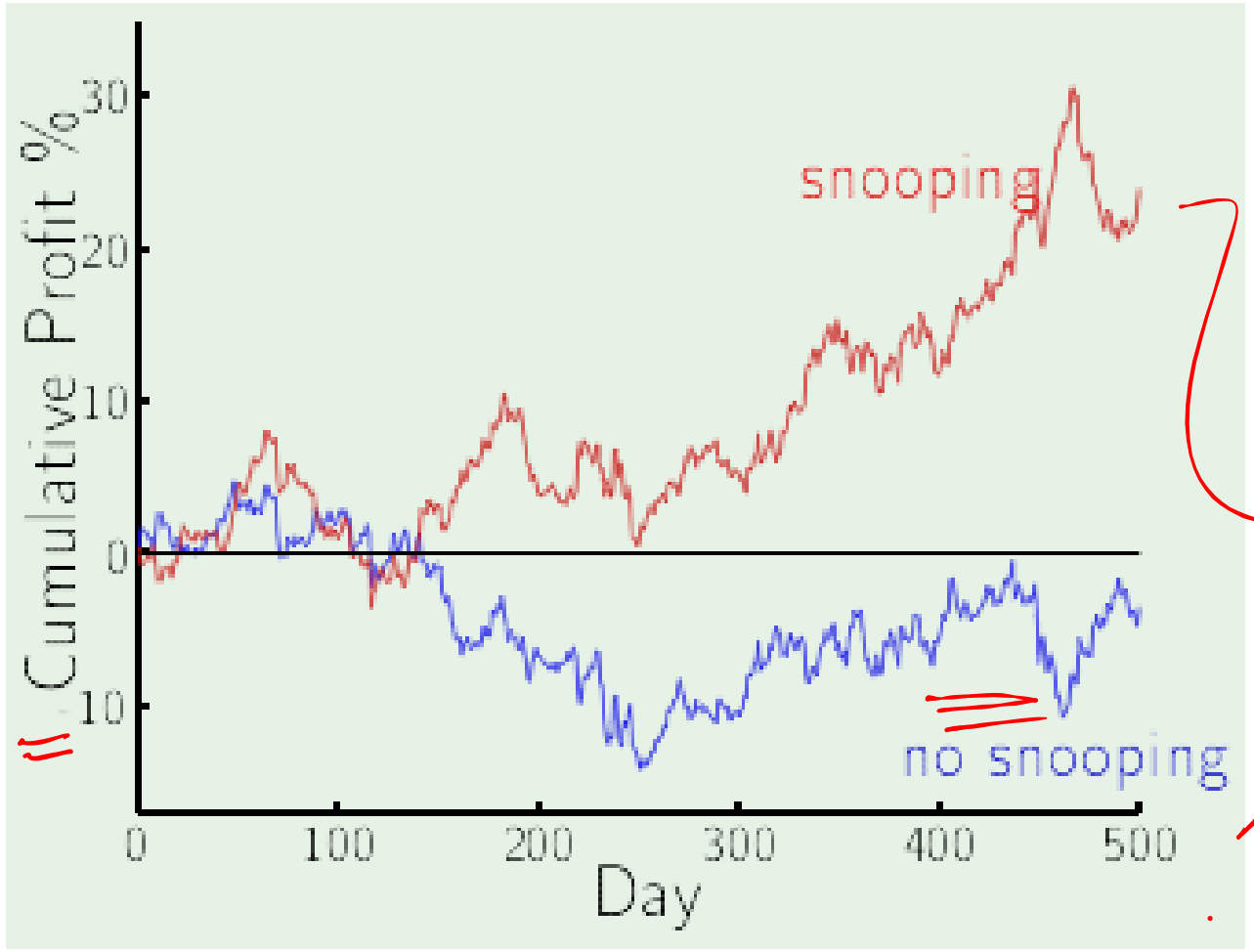
5.3

Training → 75%

25% → Test

50%

Normalized



50 years

Data Snooping is a Subtle Happy Hell

- The data looks linear, so I will use a linear model, and it worked.

  If the data were different and didn't look linear, would you do something different?

- Try linear, it fails; try circles it works.

  If you torture the data enough, it will confess.

- Try linear, it works; so I don't need to try circles.

  Would you have tried circles if the data were different?

- Read papers, see what others did on the data. Modify and improve on that.

  If the data were different, would that modify what others did and hence what you did?

  the data snooping can happen all at once or sequentially by different people

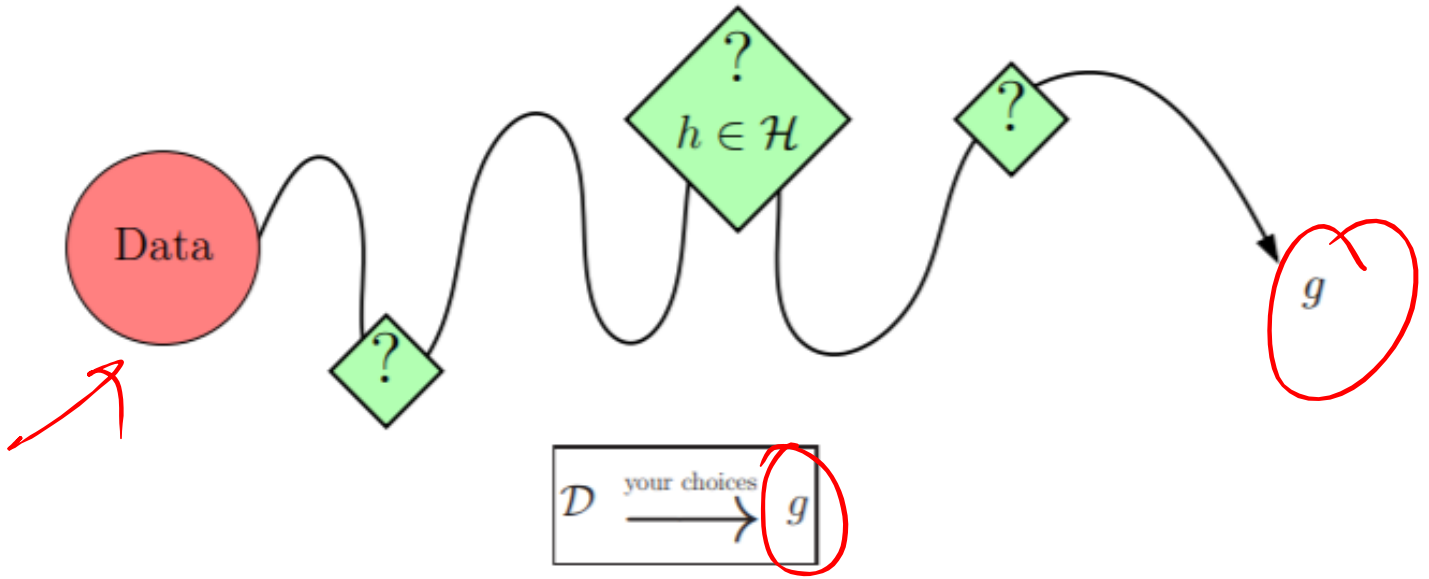
- Input normalization: normalize the data, now set aside the test set.

  Since the test set was involved in the normalization, wouldn't your g change if the test set changed?

Account for Data Snooping

Ask yourself: "If the *data* were different, could/would I have done something different?"

if yes, then there is data snooping.



You must account for ***every choice*** influenced by $\mathcal{D}$.

We know how to account for the choice of g from $\mathcal{H}$.

Three Learning Principles

- **Occam's Razor**: pick a model carefully ✓
  Simpler $\mathcal{H}$ is better.

- **Sampling Bias**: generate the data carefuly ✓
  Make sure you train and test from the same bin.

- **Data Snooping**: handle the data carefully ✓
  Account for all choices the data influenced. Choose $\mathcal{H}$ *before* you see the data.

Thanks!