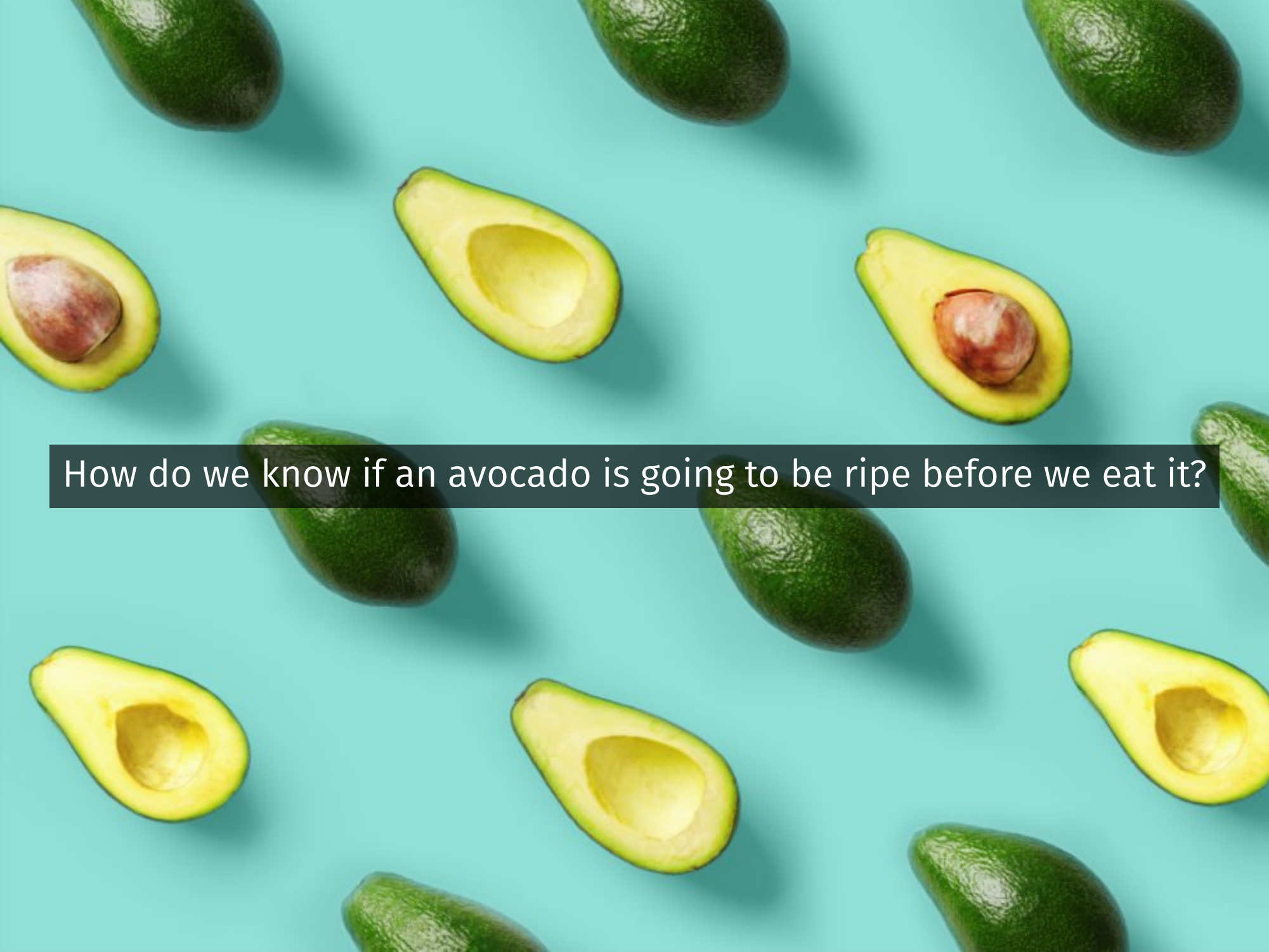
DSC 40A

Theoretical Foundations of Data Science I


How do we know if an avocado is going to be ripe before we eat it?

Try a little tenderness

Australian Avocados
www.avocado.org.au

How do you know when we're ripe?

AVOCADO COLOUR & RIPENESS CHART

Colour Rating

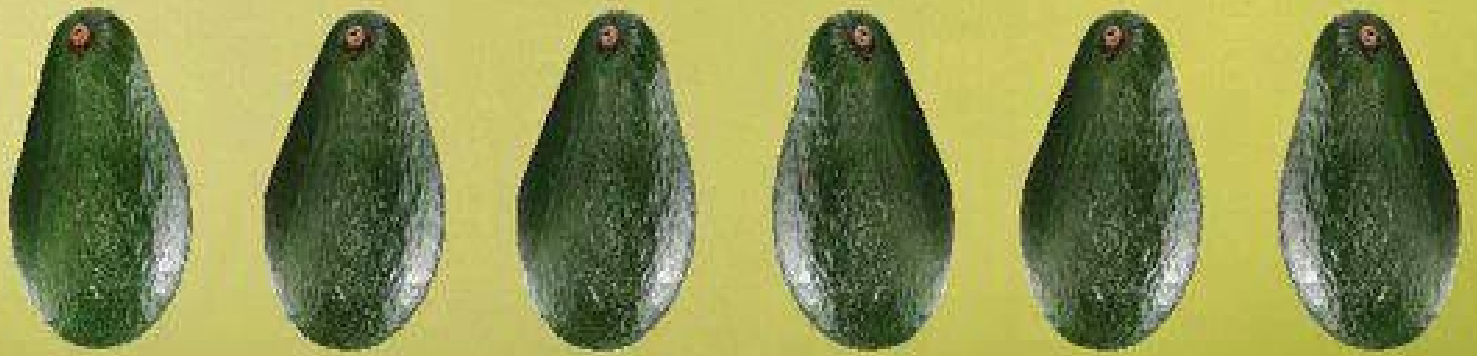
HASS
Look & Touch

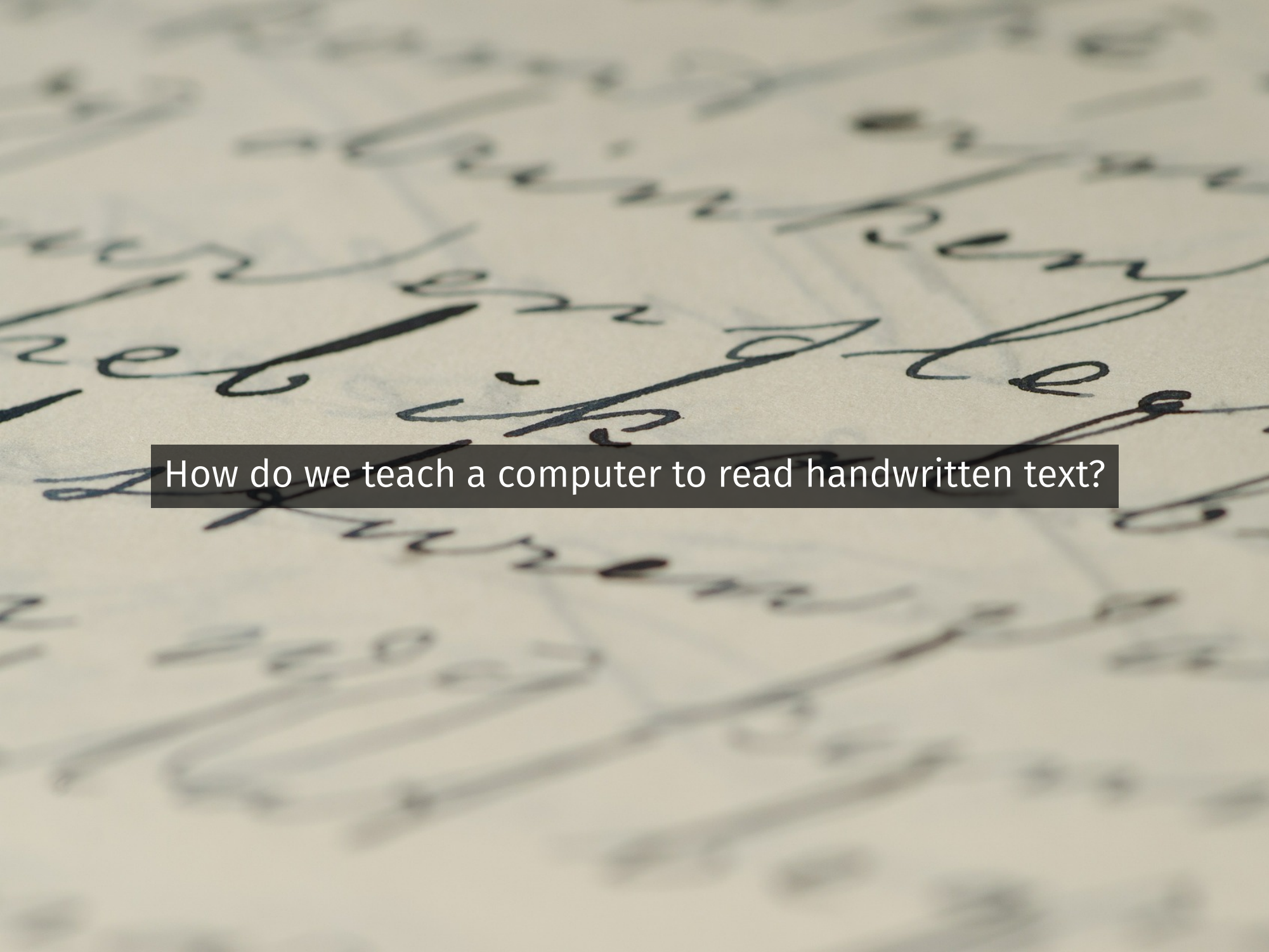
1	2	3	4	5	6
					
Hard Effogi puncture (kgf) - using 11mm tip	**Rubbery** 5kgf	**Softening** 2kgf	**Firm Ripe** 1kgf	**Medium to Soft Ripe** 0.65kgf	**Soft to Over Ripe** 0.45kgf

Firmness Rating

GREEN SKINS
Touch

(Shepard, Wurtz, Sharwil, Reed)




How do we teach a computer to read handwritten text?


How do we predict a future data scientist's salary?

...by **learning** from data.

How do we learn from data?

The fundamental approach:
1) Turn learning into a math problem.
2) Solve that problem.

After this quarter, you'll...

- understand the basic principles underlying almost every machine learning and data science method.
- be better prepared for the math in upper division: vector calculus, linear algebra, and probability.
- be able to tackle the problems mentioned at the beginning.

Theoretical Foundations of Data Science

In This Video

How do we make good predictions? What *is* a good prediction?

Recommended Reading

Course Notes: Chapter 1, Section 1


How do we predict a future data scientist's salary?

Learning from Data

- Idea: ask a few data scientists about their salary.
- StackOverflow survey.
- Five random responses:

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

Question

Given this data, how might you predict your future salary?

Some Common Approaches

- The **mean**:

$$\frac{1}{5} \times (90{,}000 + 94{,}000 + 96{,}000 + 120{,}000 + 160{,}000)$$
$$= 112{,}000$$

- The **median**:

$$90{,}000 \quad 94{,}000 \quad \underbrace{96{,}000}_{\uparrow} \quad 120{,}000 \quad 160{,}000$$

- Which is better? Are these good ways of predicting future salary?

Quantifying goodness/badness of a prediction

- The **error**: distance from prediction to the right answer.

$$\text{error} = |\text{prediction} - (\text{actual future salary})|$$

- Find prediction with smallest possible error.

- There's a problem with this:

When making prediction, actual future salary is unknown

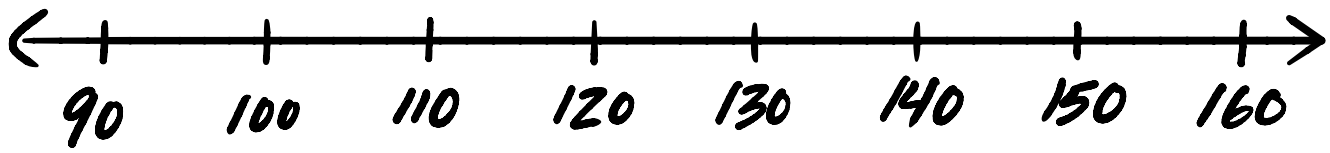
What is good/bad, intuitively?

- The data:

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

- Consider these hypotheses:

$$h_1 = 150{,}000 \qquad h_2 = 115{,}000$$



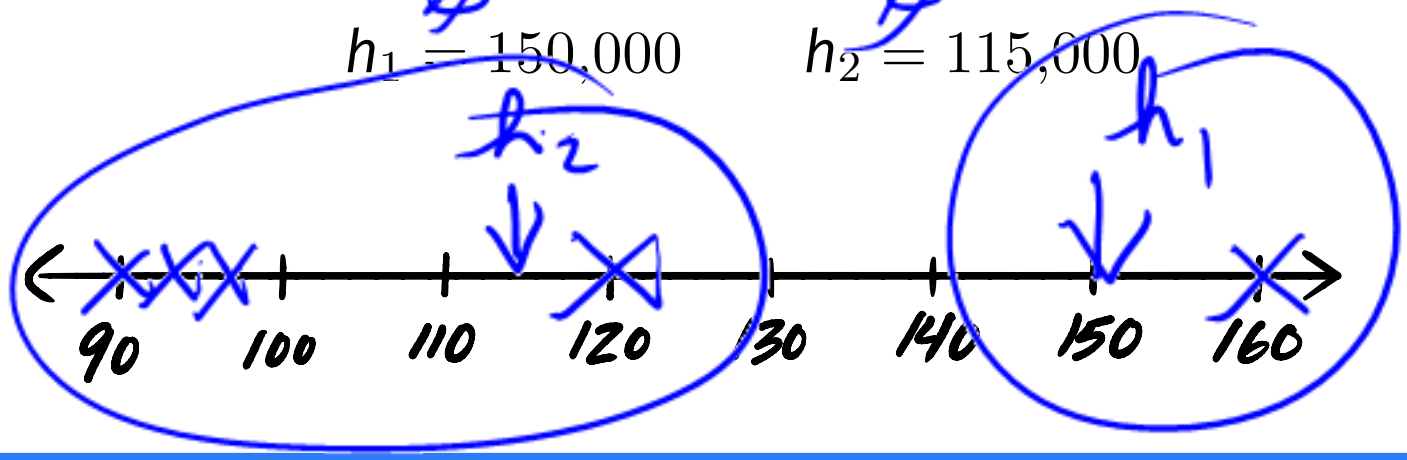
What is good/bad, intuitively?

- The data:

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

- Consider these hypotheses: = prediction

$h_1 = 150{,}000$ $h_2 = 115{,}000$



Question

Which do you think is better, h_1 or h_2? *Why*?

Quantifying our intuition

- Intuitively, a good prediction is close to the data.
- Suppose we predicted a future salary of $h_1 = 150{,}000$ *before* collecting data.

predicted

salary	error of h_1
90,000	60,000
94,000	56,000
96,000	54,000
120,000	30,000
160,000	10,000
	total error: 210,000
	mean error: 42,000

actual

high

small

Quantifying our intuition

- Now suppose we had predicted $h_2 = 115{,}000$.

salary	error of h_2
90,000	25,000
94,000	21,000
96,000	19,000
120,000	5,000
160,000	45,000
	total error: 115,000
	mean error: 23,000

small

high

Mean Errors

- Mean error on data:

$$h_1 : 42{,}000 \qquad h_2 : 23{,}000$$

- Conclusion: h_2 is the better prediction.
- In general: pick prediction with the smaller mean error.

We are making an assumption...

- We're assuming that future salaries will look like present salaries.

- That a prediction that was good in the past will be good in the future.

Question

Is this a good assumption?

Which is better: the mean or median?

- Recall:

h_3 h_4

$$\text{mean} = 112{,}000 \qquad \text{median} = 96{,}000$$

- We can calculate the average error of each:

$$\text{mean} : 22{,}400 \qquad \text{median} : 19{,}200$$

- The median is the best prediction so far!

- But is there an even better prediction?

Finding the best prediction?

- Any (non-negative) number is a valid prediction.
- Goal: out of all predictions, find the prediction h^* with the smallest mean error.
- This is an **optimization problem**.

among $h \in \mathbb{R}$, find the one, h^*, that minimizes mean error

A Formula for the Mean Error

- We have data:

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

- Suppose our prediction is h.

- The **mean error** of our prediction is:

$$R(h) = \frac{1}{5}\Big(|90{,}000 - h| + |94{,}000 - h| + |96{,}000 - h| + |120{,}000 - h| + |160{,}000 - h|\Big)$$

function of h

A Formula for the Mean Error

- We have a function for computing the mean error of **any** possible prediction.

$$\begin{aligned} R(\textbf{150,000}) = \frac{1}{5}\Big(&|90{,}000 - \textbf{150,000}| + |94{,}000 - \textbf{150,000}| \\ &+ |96{,}000 - \textbf{150,000}| + |120{,}000 - \textbf{150,000}| \\ &+ |160{,}000 - \textbf{150,000}|\Big) \\ &= \textbf{42,000} \end{aligned}$$

A Formula for the Mean Error

- We have a function for computing the mean error of **any** possible prediction.

$$\begin{aligned} R(\textbf{115,000}) = \frac{1}{5}\Big(&|90{,}000 - \textbf{115,000}| + |94{,}000 - \textbf{115,000}| \\ &+ |96{,}000 - \textbf{115,000}| + |120{,}000 - \textbf{115,000}| \\ &+ |160{,}000 - \textbf{115,000}|\Big) \\ &= \textbf{23,000} \end{aligned}$$

A Formula for the Mean Error

- We have a function for computing the mean error of **any** possible prediction.

$$\begin{aligned} R(\pi) = \frac{1}{5}\Big(&|90{,}000 - \pi| + |94{,}000 - \pi| \\ &+ |96{,}000 - \pi| + |120{,}000 - \pi| \\ &+ |160{,}000 - \pi| \Big) \\ &= \mathbf{111{,}996.8584\ldots} \end{aligned}$$

A Formula for the Mean Error

- We have a function for computing the mean error of **any** possible prediction.

$$\begin{aligned} R(\pi) = \frac{1}{5}\Big(&|90{,}000 - \pi| + |94{,}000 - \pi| \\ &+ |96{,}000 - \pi| + |120{,}000 - \pi| \\ &+ |160{,}000 - \pi| \Big) \\ &= 111{,}996.8584\ldots \end{aligned}$$

Question

Without doing any calculations, which is correct?

A) $R(50) < R(100)$
B) $R(50) = R(100)$
C) $R(50) > R(100)$

A *General* Formula for the Mean Error

- Suppose we collect n salaries, $y_1, y_2, \ldots, y_n$.
- The mean error of the prediction h is:

$$R(h) = \frac{1}{n}\left(|y_1 - h| + |y_2 - h| + \ldots + |y_n - h|\right)$$

- Or, using **summation notation**:

$$R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$$

$|h - y_i|$

The Best Prediction

- We want the best prediction, h^*.
- The smaller $R(h)$, the better h.
- Goal: find h that minimizes $R(h)$.

Summary

- We started with the learning problem:

  Given salary data, predict your future salary.

- We turned it into this problem:

  Find a prediction h^* which has smallest mean error on the data.

- We have turned the problem of learning into a specific type of math problem: an **optimization problem**.

- **Next time:** We solve this math problem.