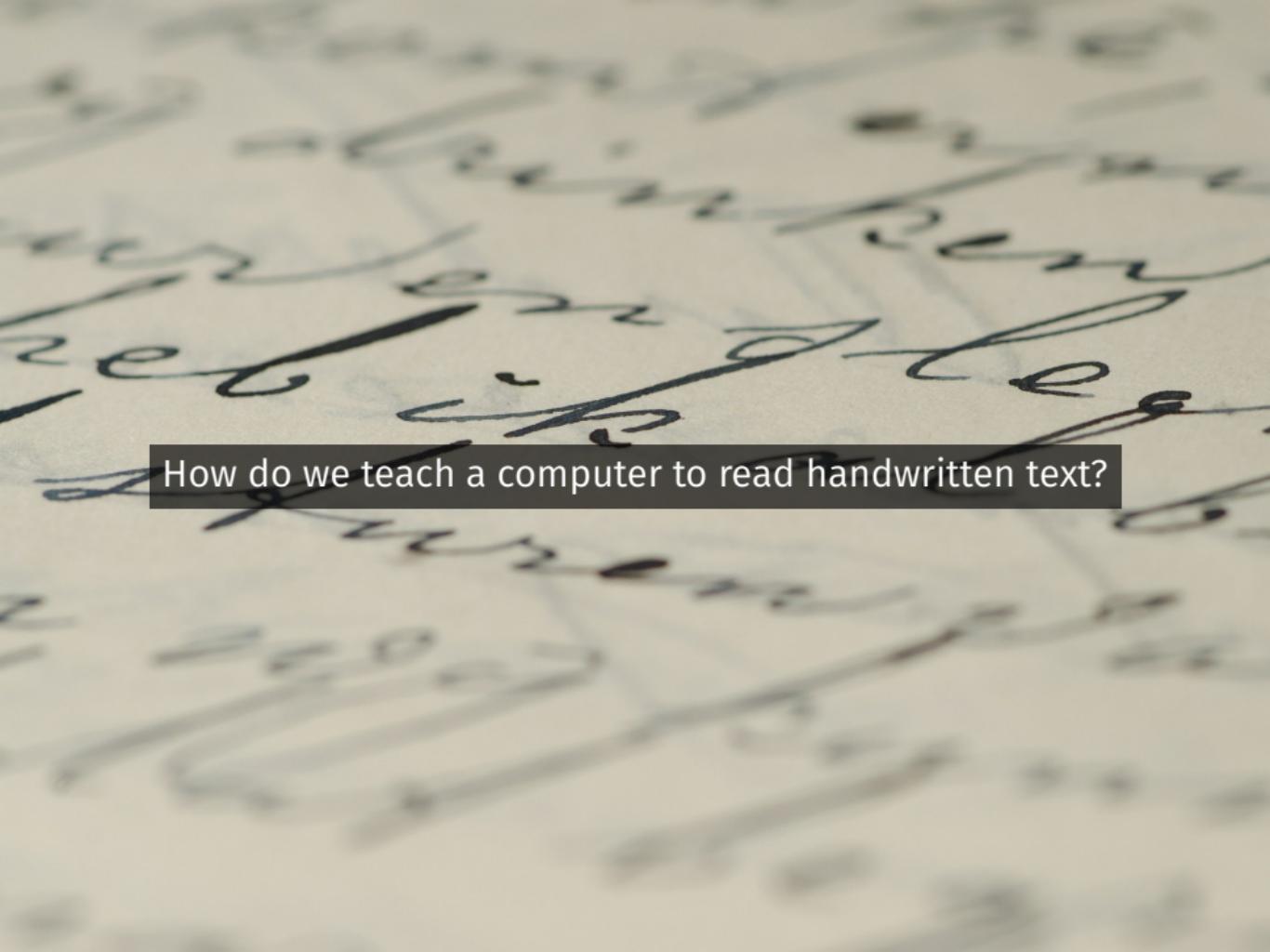


DSC 40A

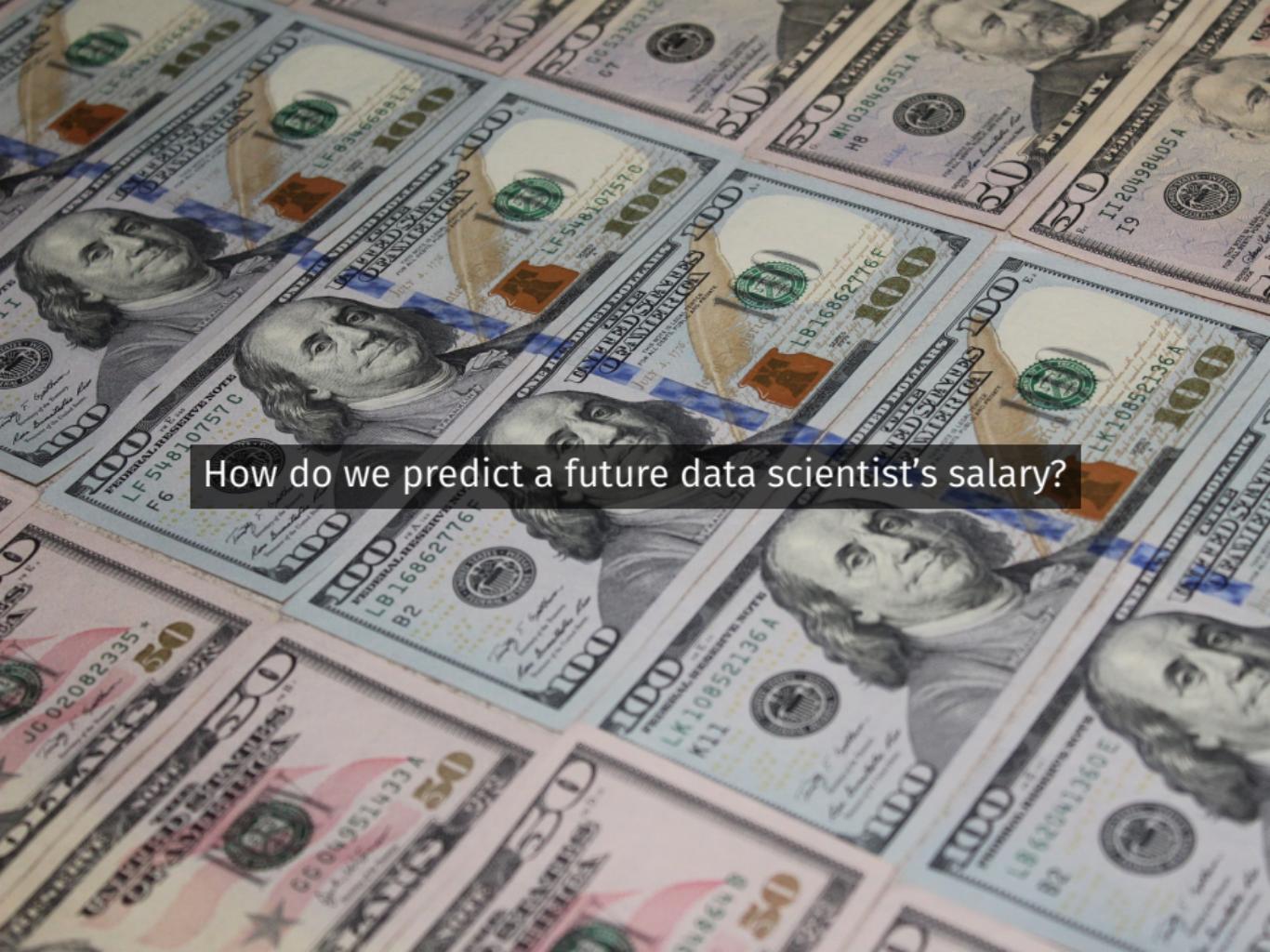
Theoretical Foundations of Data Science I

A wide-angle aerial photograph of Earth's horizon, showing a vast expanse of white and light blue clouds against the dark void of space. A thin, bright blue line marks the edge of the atmosphere where it meets the blackness of space.

How do we decide who needs the most help before a hurricane hits?

The background of the image is a close-up, slightly out-of-focus photograph of handwritten cursive text on lined paper. The text is written in black ink and appears to be a continuous stream of letters, possibly a sentence or a page from a notebook.

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

www.dsc40a.com



DSC 40A
Lecture 01
Learning via Optimization, pt I.

Lecture Format

- ▶ Lecture slides will be posted before class.
 - ▶ Suggestion: don't write everything down!
 - ▶ I'll write definitions, proofs, etc. on the slides.
-
-
-

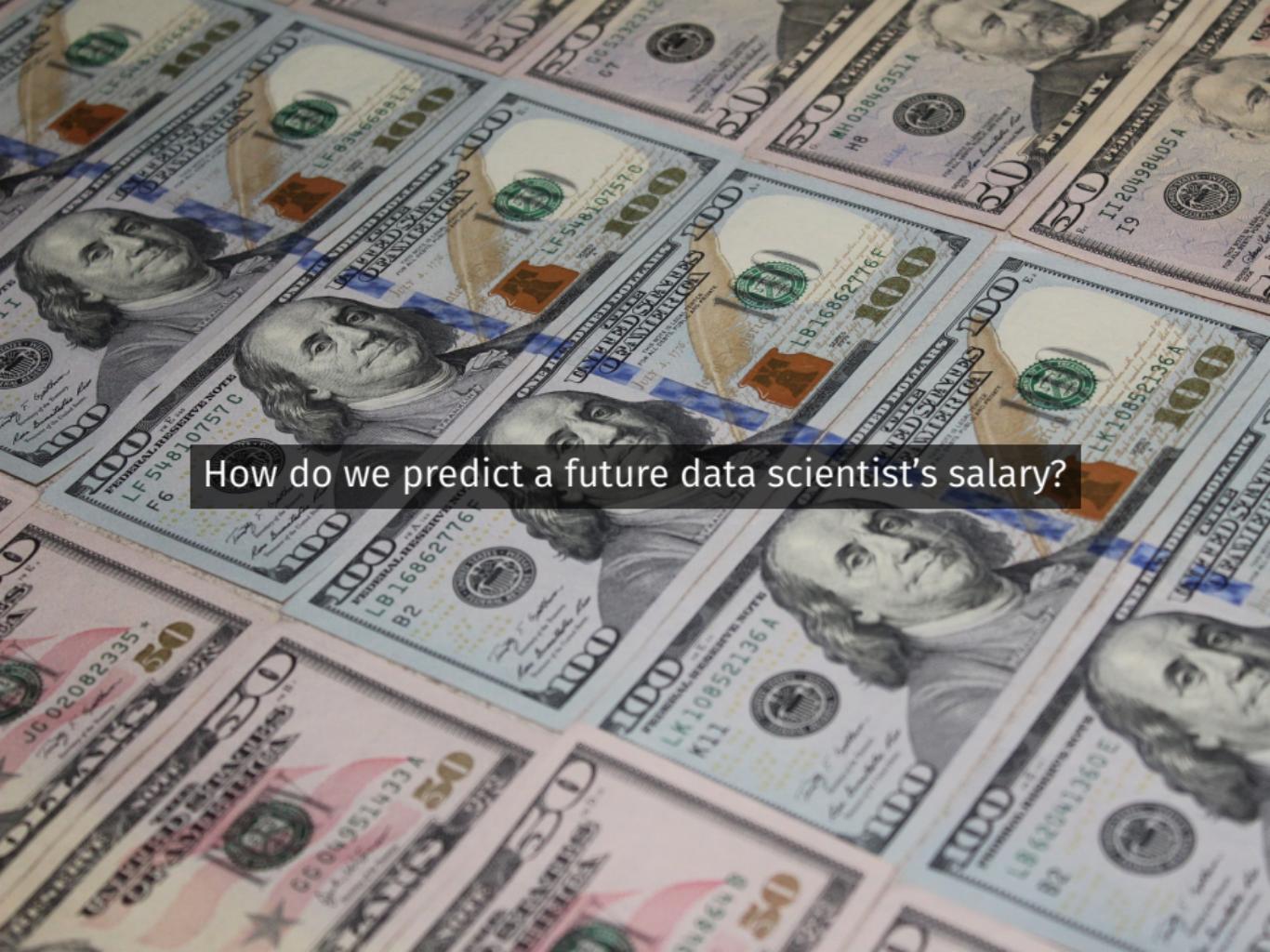
- ▶ Value of lecture: **interaction** and **discussion**.

Today's Question

How do we turn the problem of learning into a math problem?

Recommended Reading

Chapter 01, Section 01



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 94,000 96,000 120,000 160,000

Discussion Question

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

The Mean and the Median

- ▶ The **mean**:

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

- ▶ The **median**:

90,000 94,000 96,000 120,000 160,000
 ↑

- ▶ 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:
-

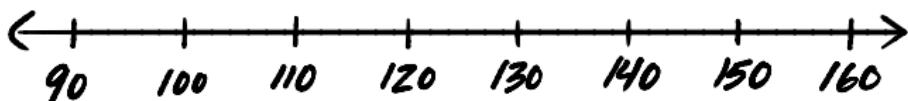
What is good/bad, intuitively?

- ▶ The data:

90,000 94,000 96,000 120,000 160,000

- ▶ Consider these hypotheses:

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



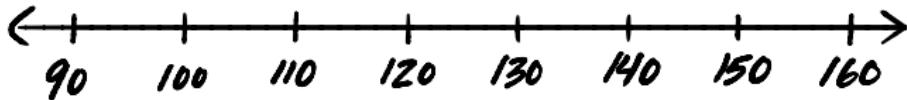
What is good/bad, intuitively?

- ▶ The data:

90,000 94,000 96,000 120,000 160,000

- ▶ Consider these hypotheses:

$$h_1 = 150,000 \quad h_2 = 115,000$$



Discussion 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.

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

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

Mean Errors

- ▶ Mean error on data:

$$h_1 : 42,000 \quad h_2 : 23,000$$

- ▶ 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.

Discussion Question

Is this a good assumption?

Which better: the mean or median?

- ▶ Recall:

mean = 112,000 median = 96,000

- ▶ We can calculate the average error of each:

mean : 22,400 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**.

Status Update

- ▶ 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**.

What's Left?

- ▶ We need to solve this math problem.
- ▶ Next time: math.