Lecture 1 – Introduction, Learning From Data

DSC 40A, Fall 2022 @ UC San Diego

Mahdi Soleymani, with help from **many others**

Agenda

1. Who are we?
2. What is this course about?
3. How will this course run?
4. How do we turn the problem of learning from data into a math problem?

Who are we?

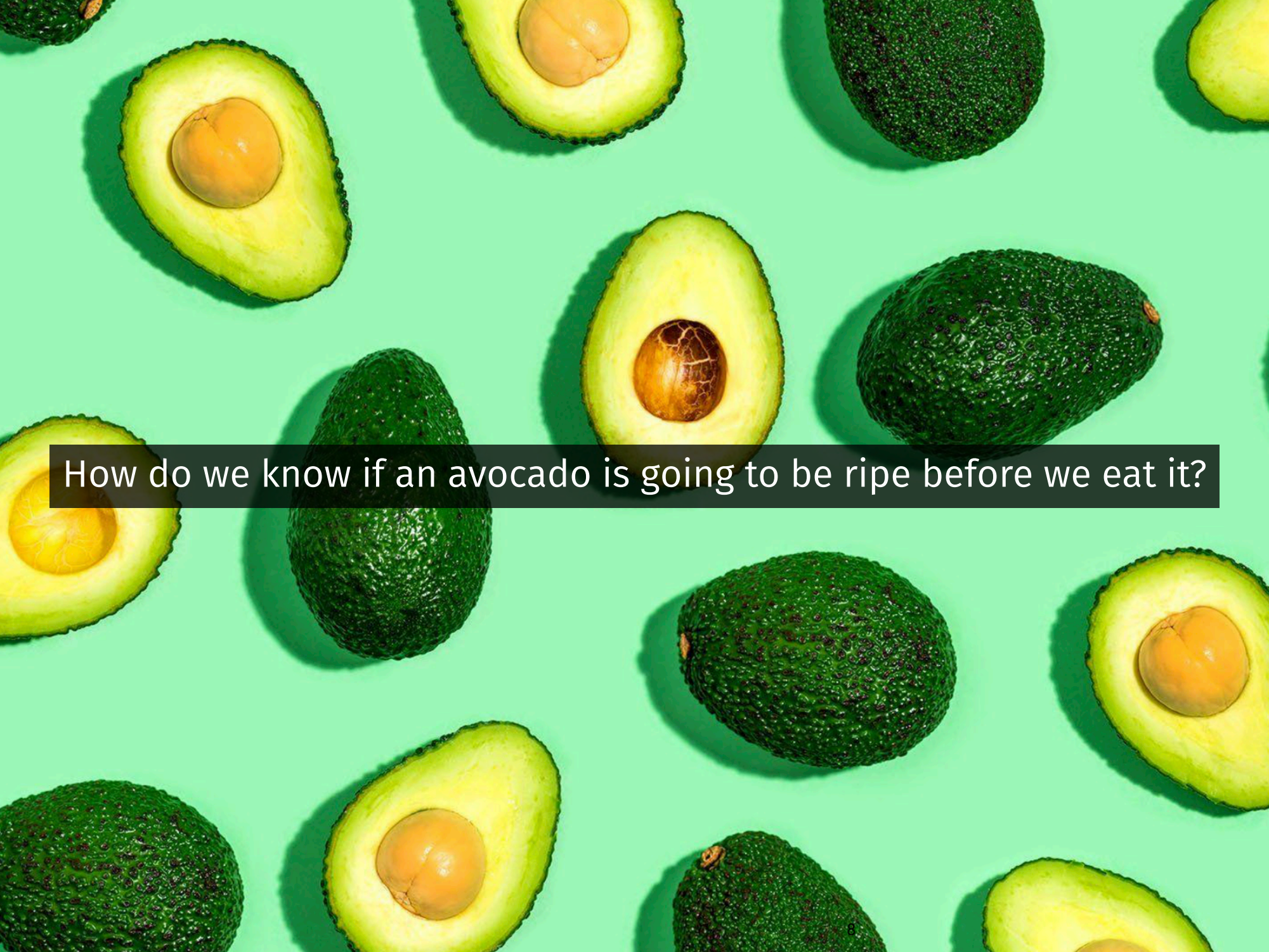
Instructor:

- Mahdi Soleymani
- Ph.D. in ECE, University of Michigan Ann Arbor.
- Research: Coding/information theory and machine learning
- Postdoctoral Scholar and Lecturer at HDSI.
- Email: msoleymani@ucsd.edu

Course staff:

- 1 TA, who will teach discussion and help run the class.
  - Pushkar Bhuse, a MS student in CSE.
- 8 tutors, who will hold OH, grade assignments, and help run the class.
  - Aryaman Sinha, Jessica Song, Karthikeya Manchala, Shiv Sakthivel, Vivian Lin, Weiyue Li, Yujia Wang, Yuxin Guo.
  - All undergrads who took DSC 40A before and did well.
- Read about them at **dsc40a.com/staff**.

What is this course about?


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

Try a little tenderness

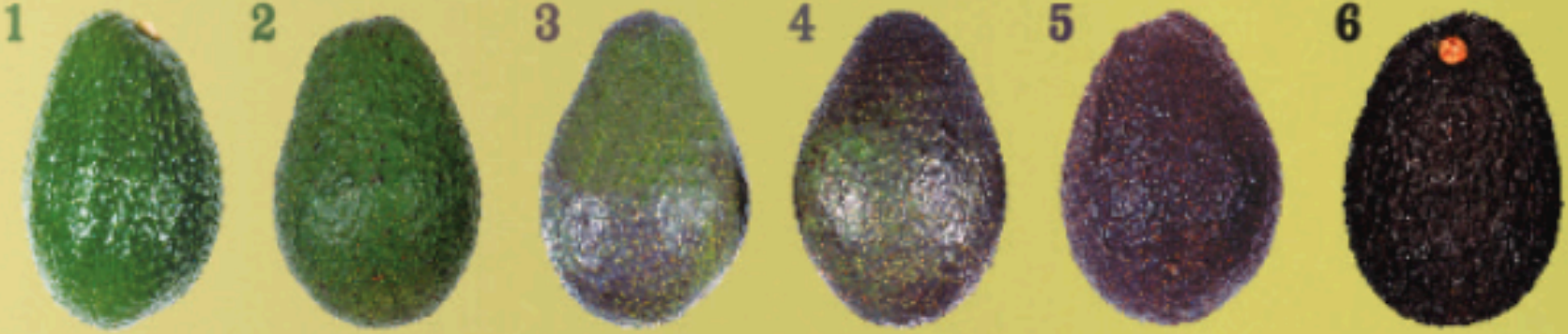
Australian Avocados
www.avocado.org.au

How do you know when we're ripe?

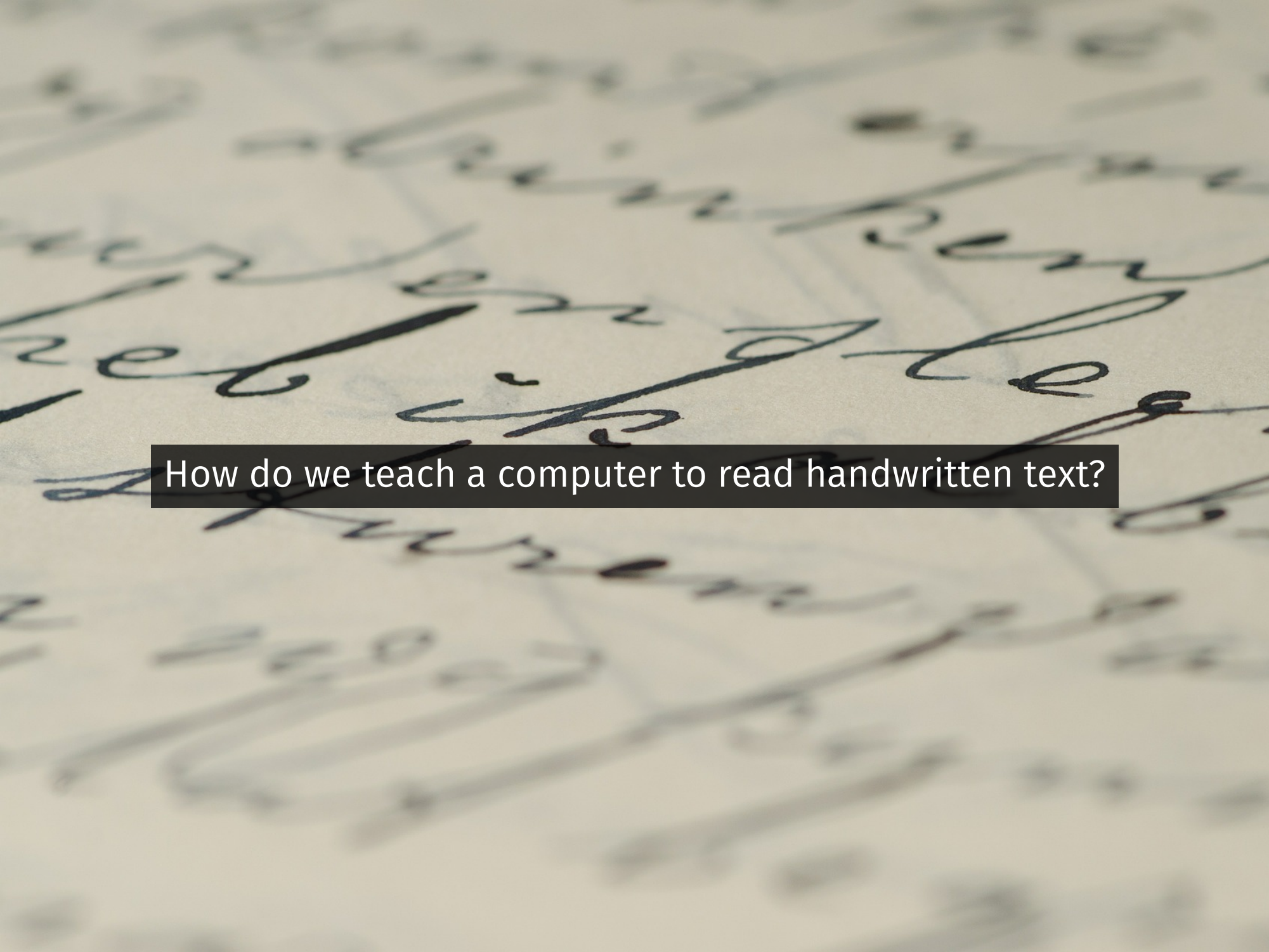
AVOCADO COLOUR & RIPENESS CHART

HASS
Look & Touch

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



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 from data into a math problem.

2. Solve that problem.

Course overview

Part 1: Learning from Data (Week 0-5)

- Summary statistics and loss functions; mean absolute error and mean squared error.
- Linear regression (incl. linear algebra).
- Clustering.

Part 2: Probability (Week 6-10)

- Set theory and combinatorics; probability fundamentals.
- Conditional probability and independence.
- Naïve Bayes (mix of both parts of the class).

Learning objectives

After this quarter, you'll...

- understand the basic principles underlying the mainstream machine learning and data science algorithms.
- 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.

How will this course run?

Technology

- The course website, **dsc40a.com**, is where all content (lectures, **readings**, homeworks, discussions) will be posted. It also contains a calendar of office hours (with Zoom links).
- **Campuswire** is where all announcements will be sent, and where all student-staff and student-student communication will occur. **Ask questions here!**
- **Gradescope** is where all assignments are submitted and all grades live.
- **Zoom** will be used for virtual office hours and discussion.

Lectures

- M/W/F 4:00-4:50PM, Pepper Canyon Hall (PCYNH). No attendance required; recordings posted.
- Content in the first few weeks will closely follow readings.
- Lecture slides will be posted before class.
- I'll write definitions, proofs, etc. on the slides.

_____

_____

Discussion

- **Discussion**: Monday 5:00-5:50 and 6:00-6:50.
  - Come to work on problems in small groups (”groupwork”) of 2-4.
  - Attendance is highly recommended but not required, however you **must** work on the groupwork problems in a group (whether that’s in discussion or on your own time).
- Groupwork problems must be submitted to Gradescope by **Monday at 11:59pm**.
  - Only one group member should submit; they should add the rest of the group to the assignment on Gradescope.

Assessments and exams

- **Homeworks**: Released weekly, and usually due **Friday at 2:00pm** on Gradescope. Worth 40% of your grade.
- **Groupworks**: Due **Monday at 11:59pm**. Worth 10% of your grade.
- **Midterm Exam**: TBD, In-person. Worth 20% of your grade.*
- **Final Exam**: In-person 12/03, 7 PM-9:59PM. Worth 30% of your grade.*

Leniency

We have some leniency built into our grading scheme:

- **Slip days**: 5. Can only be used on homework. Can only use one per homework.
- **Drops**: We will drop your lowest homework and groupwork.

Support

- **Office Hours (starting next week)**: held throughout the week, but concentrated near deadlines. Calendar on course website will be updated with times by the weekend.
  - Some staff OH are remote via Zoom. See Calendar for Zoom links. Others are in-person in the CSE Basement. Put yourself on the queue at autograder.ucsd.edu ("The Autograder").

- **Campuswire**: Use it! We're here to help you.
  - Do not post answers.

  - Do not DM TA and tutors.

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


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

Learning from data

- Idea: ask a few data scientists about their salary.
  - StackOverflow does this annually.

- 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?

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 the goodness/badness of a prediction

- We want a metric that tells us if a prediction is good or bad.
- One idea: compute the **absolute error**, which is the distance from our prediction to the right answer.

  absolute error = |(actual future salary) – prediction|

- Then, our goal becomes to **find the prediction with the smallest possible absolute error**.
- There's a problem with this:

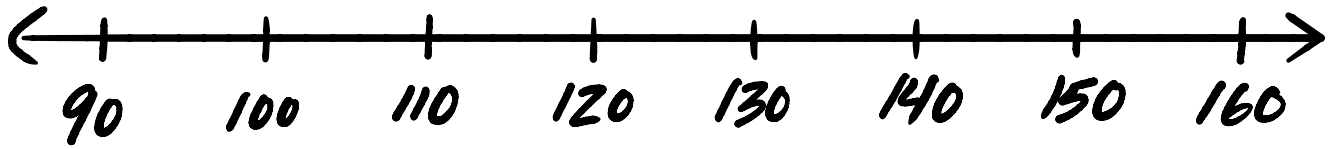
What is good/bad, intuitively?

- The data:

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

- Consider these hypotheses:

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



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	absolute error of h_1
90,000	60,000
94,000	56,000
96,000	54,000
120,000	30,000
160,000	10,000
	sum of absolute errors: 210,000
	mean absolute error: 42,000

Quantifying our intuition

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

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

Mean absolute error (MAE)

- Mean absolute 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 absolute error.