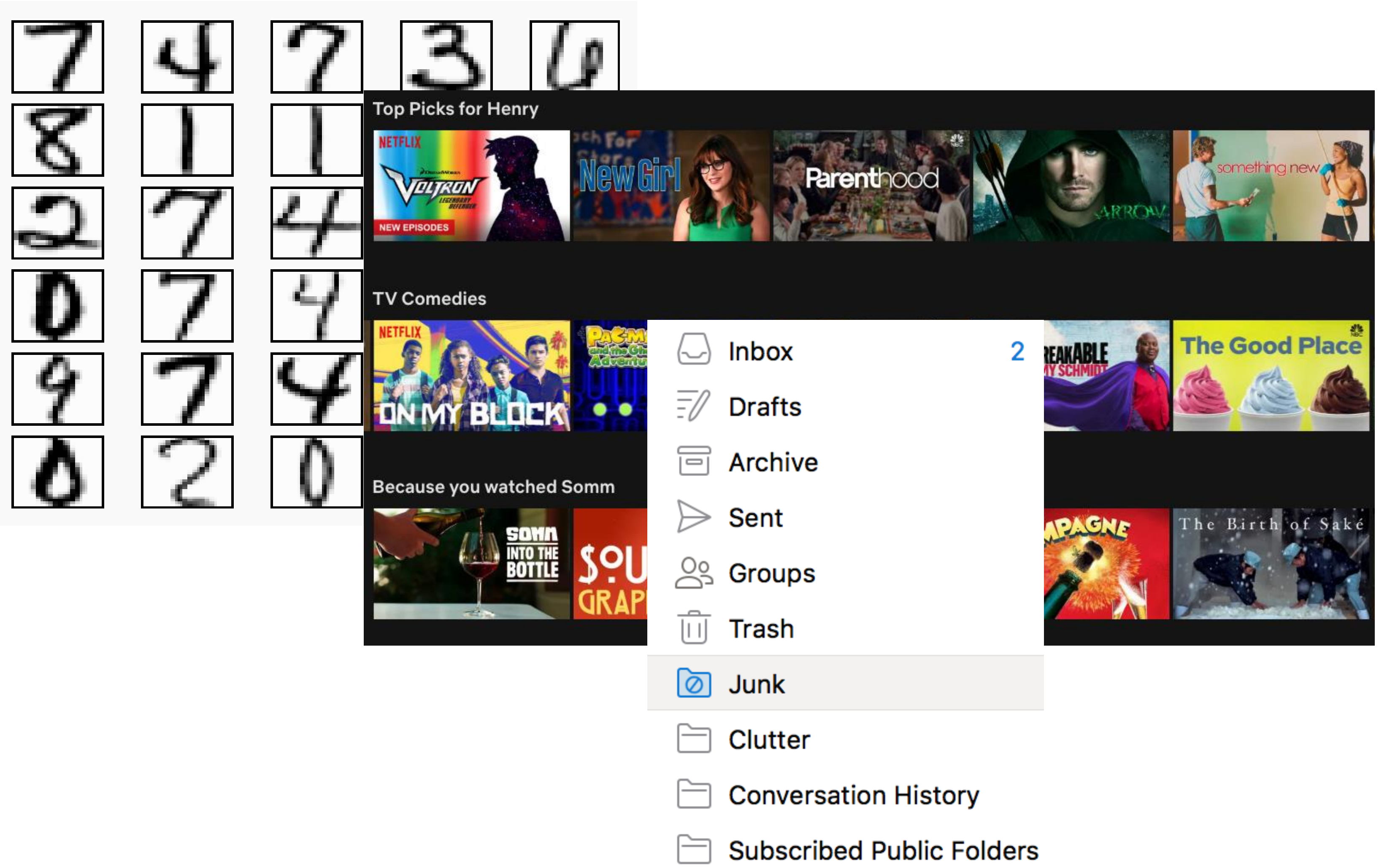
10-301/601: Introduction to Machine Learning Lecture 1 – Problem Formulation & Notation

Henry Chai

5/12/25

What is Machine Learning?

Machine Learning (A long long time ago…)

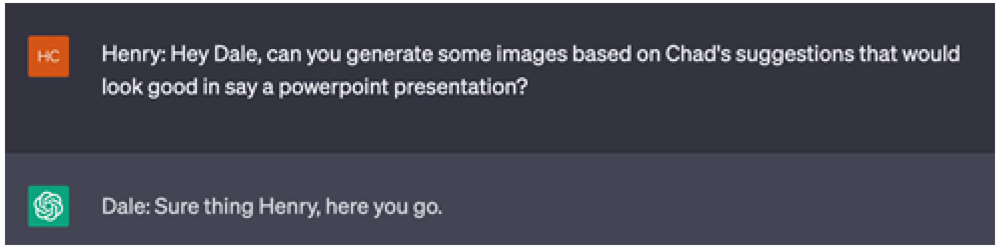


Machine Learning (A short time ago…)

Machine Learning (Now)

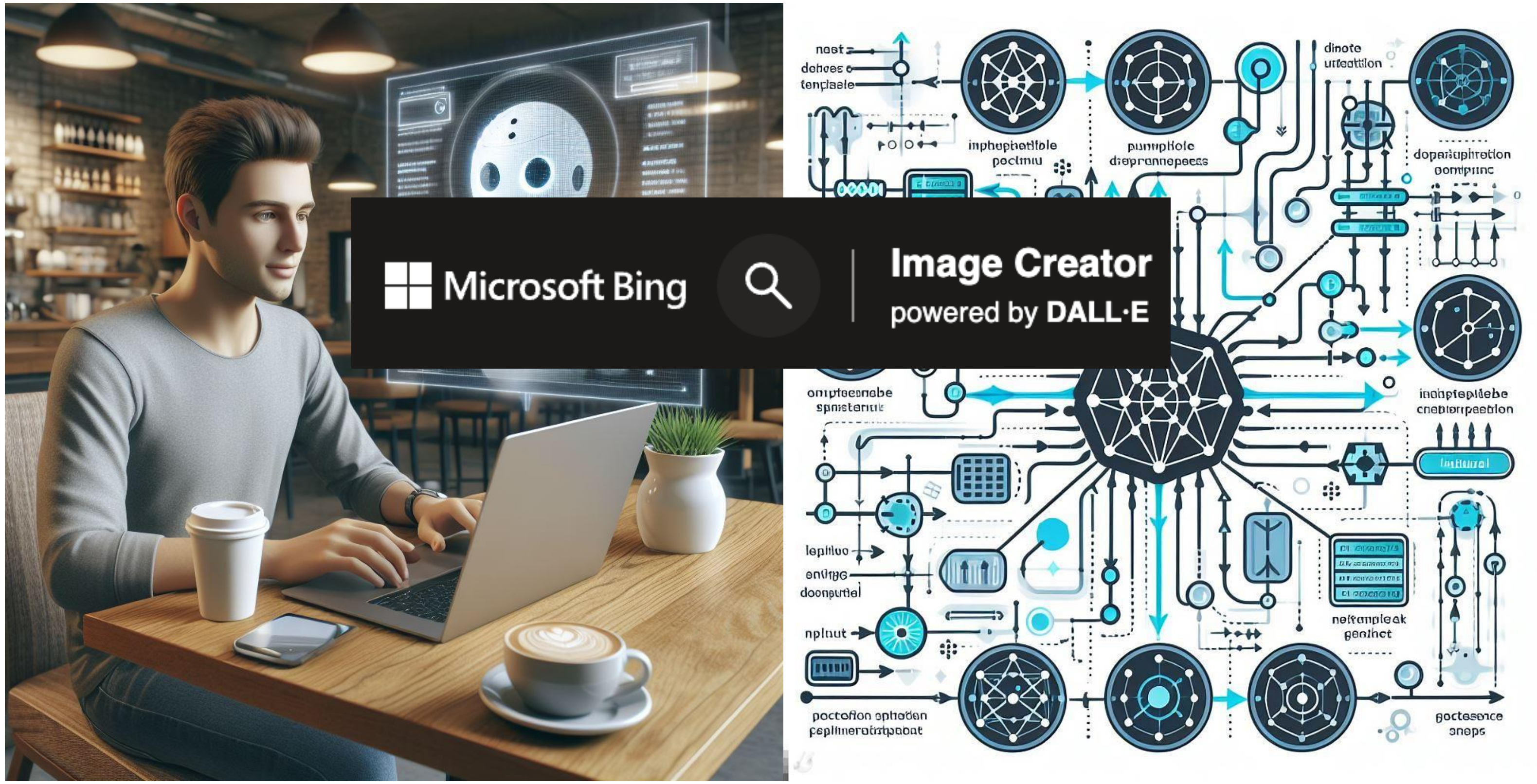
Machine Learning (Now)

Machine Learning (Now)



HC Henry: Hey Dale, can you generate some images based on Chad's suggestions that would look good in say a powerpoint presentation?

Dale: Sure thing Henry, here you go.



Source: https://www.bing.com/images/create?FORM=GERRLP

Source: https://chat.openai.com/

What is ~~Machine Learning~~ 10-301/601?

- Supervised Models
  - Decision Trees
  - KNN
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
- Unsupervised Learning
- Ensemble Methods
- Deep Learning & Generative AI
- Learning Theory
- Reinforcement Learning
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

What is Machine Learning?

Source: https://en.wikipedia.org/wiki/Panzanella

Defining a Machine Learning Task (Mitchell, 97)

- A computer program **learns** if its *performance*, *P*, at some *task*, *T*, improves with *experience*, *E*.
- Three components
  - Task, T
  - Performance metric, P
  - Experience, E

Defining a Machine Learning Task: Example

- Learning to approve loans/lines of credit

- Three components
  - Task, T

    Decide whether or not to extend someone a loan
  - Performance metric, P

    Minimizing the # of people who default on their loan
  - Experience, E

    Interviewing other loan officers

Defining a Machine Learning Task: Example

- Learning to approve loans/lines of credit

- Three components
  - Task, T

    Predict the probability that someone defaults
  - Performance metric, P

    Amount of interest made over 5 years
  - Experience, E

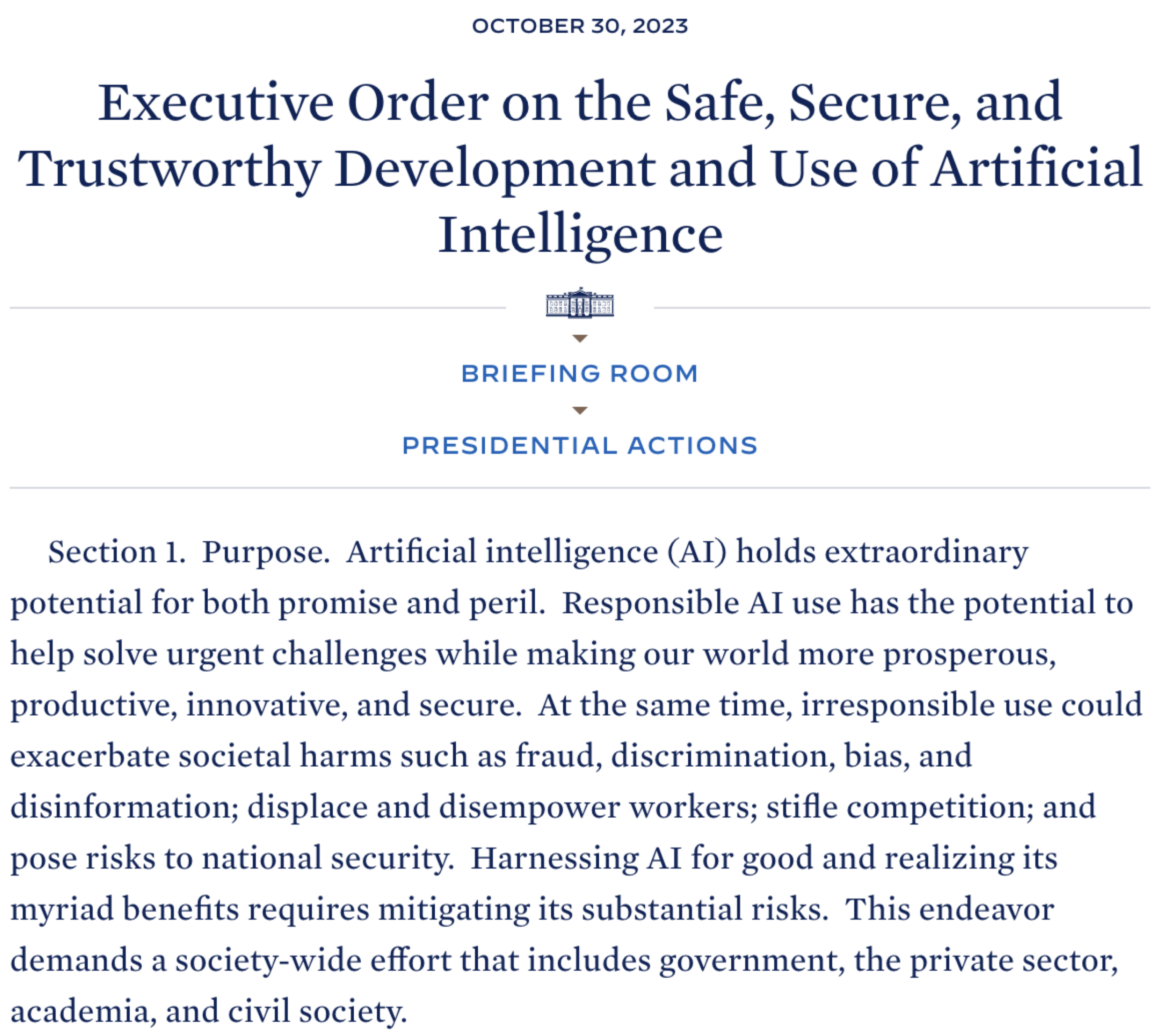
    Historical data about defaults

What is Machine Learning

- Neutral or Unbiased?

Things Machine Learning Isn’t

- Neutral or Unbiased



OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM

PRESIDENTIAL ACTIONS

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

Source: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Defining a Machine Learning Task: Example

- Learning to
- Three components
  - Task, T

    Finding optimal stuff about oxygen-sensing iridium complexes
  - Performance metric, P

    Some more chemistry stuff (linear regression)
  - Experience, E

    Lab experiments + simulations

Defining a Machine Learning Task: Example

- Learning to
- Three components
  - Task, T

    predicting complications from surgery
  - Performance metric, P

    comparing outcomes vs. predictions
  - Experience, E

(anonymized) Patient data – demographics, health data

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) binary classification task**

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) binary classification task**

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) <u>binary classification</u> task**

features

labels

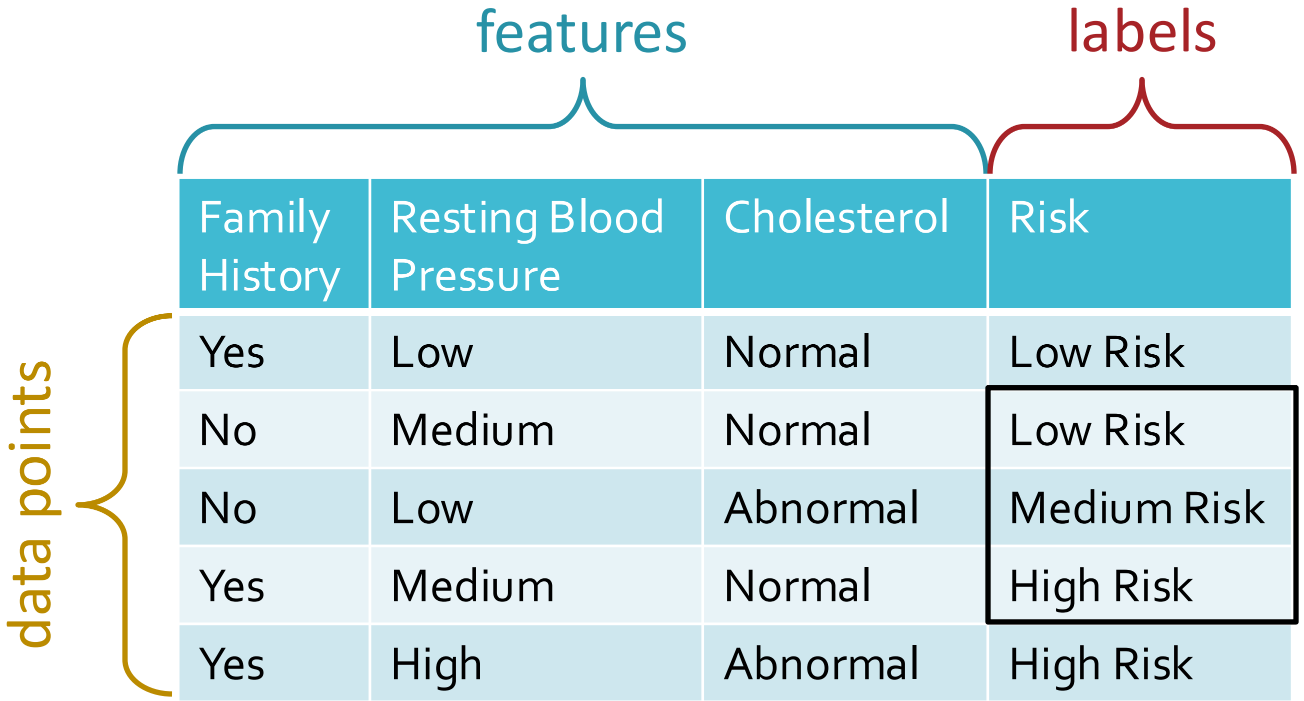
Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

data points

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **classification task**



features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Risk
Yes	Low	Normal	Low Risk
No	Medium	Normal	Low Risk
No	Low	Abnormal	Medium Risk
Yes	Medium	Normal	High Risk
Yes	High	Abnormal	High Risk

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **<u>regression</u> task**

features

targets

data points

Family History	Resting Blood Pressure	Cholesterol	Medical Costs
Yes	Low	Normal	\$0
No	Medium	Normal	\$20
No	Low	Abnormal	\$30
Yes	Medium	Normal	\$100
Yes	High	Abnormal	\$5000

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the dataset

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Is this a "good" Classifier?

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the dataset

features

labels

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

data points

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)

training dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)
- A **test** dataset is used to evaluate a classifier's **predictions**

test dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	**Predictions**
No	Low	Normal	No	Yes
No	High	Abnormal	Yes	Yes
Yes	Medium	Abnormal	Yes	Yes

- The **error rate** is the proportion of data points where the prediction is wrong

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)
- A **test** dataset is used to evaluate a classifier's **predictions**

test dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	**Predictions**
No	Low	Normal	No	Yes
No	High	Abnormal	Yes	Yes
Yes	Medium	Abnormal	Yes	Yes

- The **test error rate** is the proportion of data points in the test dataset where the prediction is wrong ($1/3$)

A Typical (Supervised) Machine Learning Routine

- Step 1 – training
  - Input: a labelled training dataset
  - Output: a classifier
- Step 2 – testing
  - Inputs: a classifier, a test dataset
  - Output: predictions for each test data point
- Step 3 – evaluation
  - Inputs: predictions from step 2, test dataset labels
  - Output: some measure of how good the predictions are; usually (but not always) error rate

Key Takeaways

- Components of a machine learning problem
- Algorithmic bias
- Components of a labelled dataset for supervised learning
- Training vs. test datasets
- Majority vote classifier

https://www.cs.cmu.edu/~hchai2/courses/10601

Logistics: Course Website

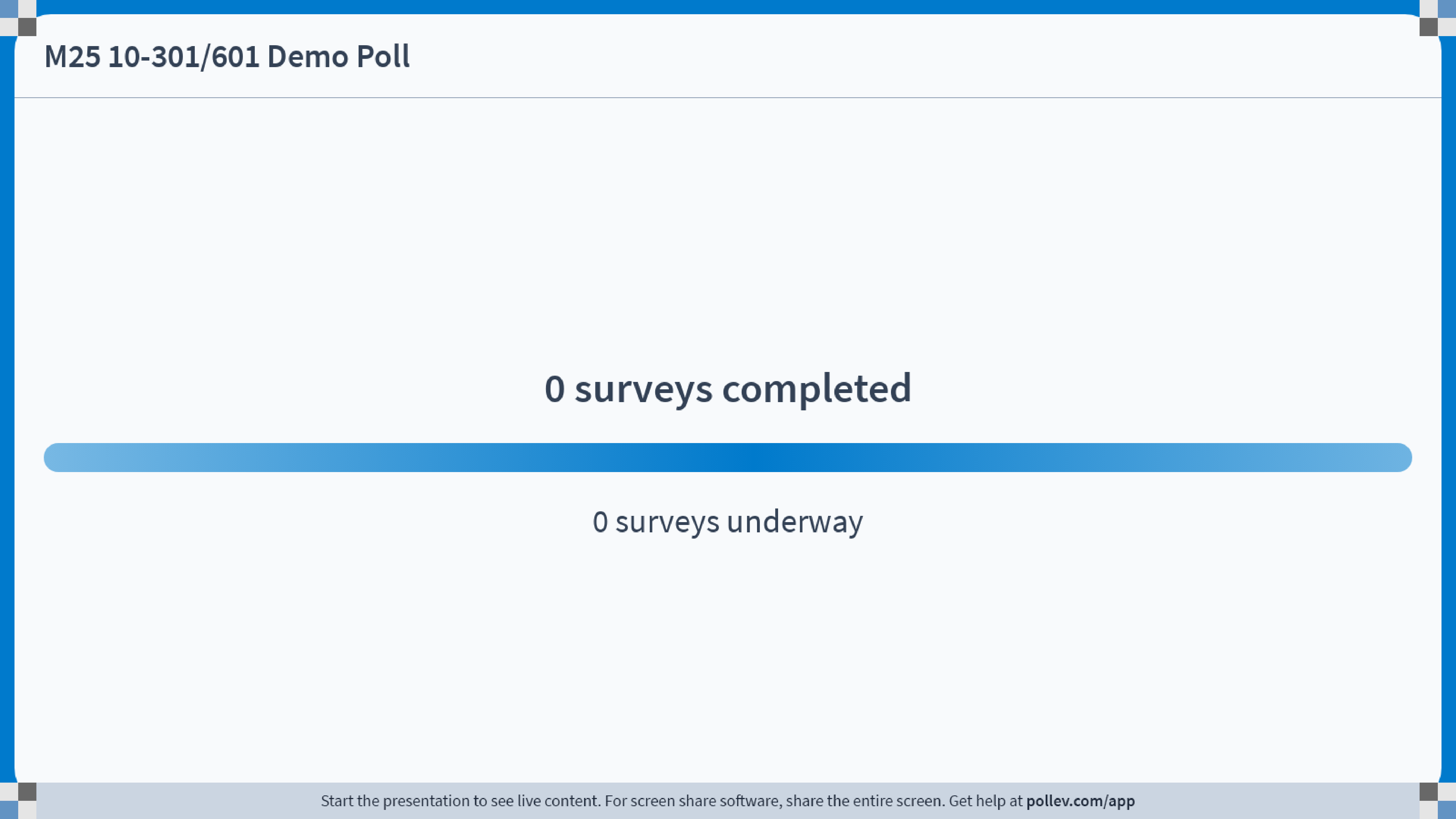
Logistics: Course Syllabus

https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

- This whole section is **required** reading https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

Logistics: Grading

- 32% = 8 homework assignments
- 24% = 4 quizzes
- 20% = midterm
- 20% = final
- 4% participation
  - 4% (full credit) for 75% or greater poll participation
  - 2% for 50%-75% poll participation
  - “Correctness” will not affect your participation grade
  - 50% credit for responses before the next lecture



M25 10-301/601 Demo Poll

0 surveys completed

0 surveys underway

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Is a hot dog a sandwich?

Yes, of course!

No, absolutely not!

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Briefly justify your response to the previous question ("Is a hot dog a sandwich?")

Join by Web

PollEv.com/301601polls

Join by QR code

Scan with your camera app

Logistics: Programming Assignments

https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

- 8 programming assignments which focus on implementing machine learning methods presented in class
  - Each will have a programming component and some written, empirical questions
  - Your answers to the written questions must be typeset in LaTeX
    - We will always provide a LaTeX starter template that you can just fill in with your answers.
  - You will submit your code and your answers to the written questions separately, both using Gradescope

Logistics: Late Policy

https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

- 8 grace days for use across all homework assignments
- Only 2 grace days may be used per homework
- Late submissions w/o grace days:
  - 1 day late = 75% multiplicative penalty
  - 2 days late = 50% multiplicative penalty
- No submissions accepted more than 2 days late

Logistics: In-class Quizzes

https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

- 4 weekly quizzes that cover the week's content
  - The goal of these regular quizzes is to keep you up to date on the material and serve as check-ins for your understanding
  - To help you prepare, we will release a "Study Guide" at the beginning of each week with practice problems
    1. You are encouraged to be working on these problems throughout the week
    2. Our TAs will go over some subset of these problems in recitations
  - **At least 50% of the points on the in-class quizzes will come from questions in the Study Guides**

Logistics: Collaboration Policy

https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

- **On study materials, you may collaborate freely, to any extent**
  - **However, you still have a duty to protect your work:** you may not post your solutions publicly/share your solutions with anyone outside of the course
- Collaboration on programming assignments is encouraged but must be documented
- **You must always write your own code/answers**
  - You may not use generative AI tools to complete the programming assignments
- Good approach to collaborating on programming assignments:
  1. Collectively sketch pseudocode on an impermanent surface, then
  2. Disperse, erase all notes and start from scratch

Logistics: Technologies

https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus

- Piazza, for course discussion: https://piazza.com/cmu/summer2025/10301601/home
- Gradescope, for submitting homework assignments: https://www.gradescope.com/courses/1030511
- Polleverywhere, for in-class participation: https://pollev.com/301601polls
- Panopto, for lecture recordings: https://scs.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=caea12f7-c2b4-48c2-b947-b2cf00e7bfee

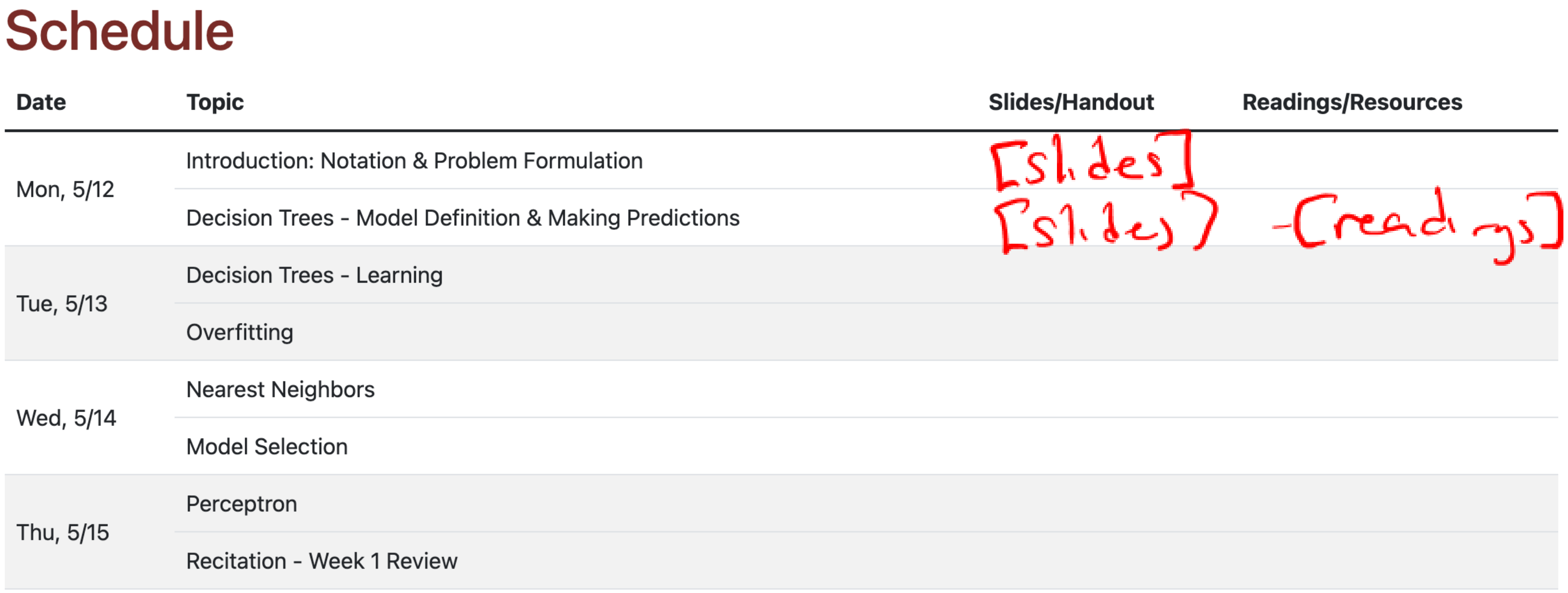
Logistics: Weekly Schedule

https://www.cs.cmu.edu/~hchai2/courses/10601/#Schedule

Lecture	Monday 9:30 – 10:30	
	Monday 11 – 12	
	Tuesday 9:30 – 10:30	
	Tuesday 11 – 12	
	Wednesday 9:30 – 10:30	
	Wednesday 11 – 12	
	Thursday 9:30 – 10:30	
Recitation	Thursday 11 – 12	
Quiz	Friday 11 – 12	
HW1	Released Tuesday	Due Friday
HW2	Released Friday	Due Tuesday

https://www.cs.cmu.edu/~hchai2/courses/10601/#Schedule

Logistics: Lecture Schedule

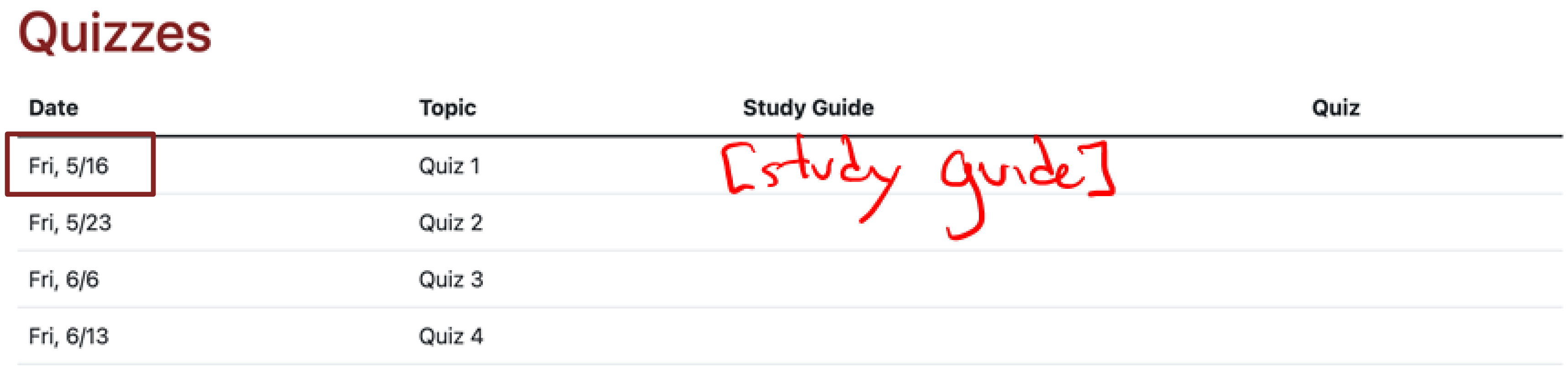


Schedule

Date	Topic	Slides/Handout	Readings/Resources
Mon, 5/12	Introduction: Notation & Problem Formulation	[slides]	
	Decision Trees - Model Definition & Making Predictions	[slides]	-[readings]
Tue, 5/13	Decision Trees - Learning		
	Overfitting		
Wed, 5/14	Nearest Neighbors		
	Model Selection		
Thu, 5/15	Perceptron		
	Recitation - Week 1 Review		

https://www.cs.cmu.edu/~hchai2/courses/10601/#Quizzes

Logistics: Quiz Schedule

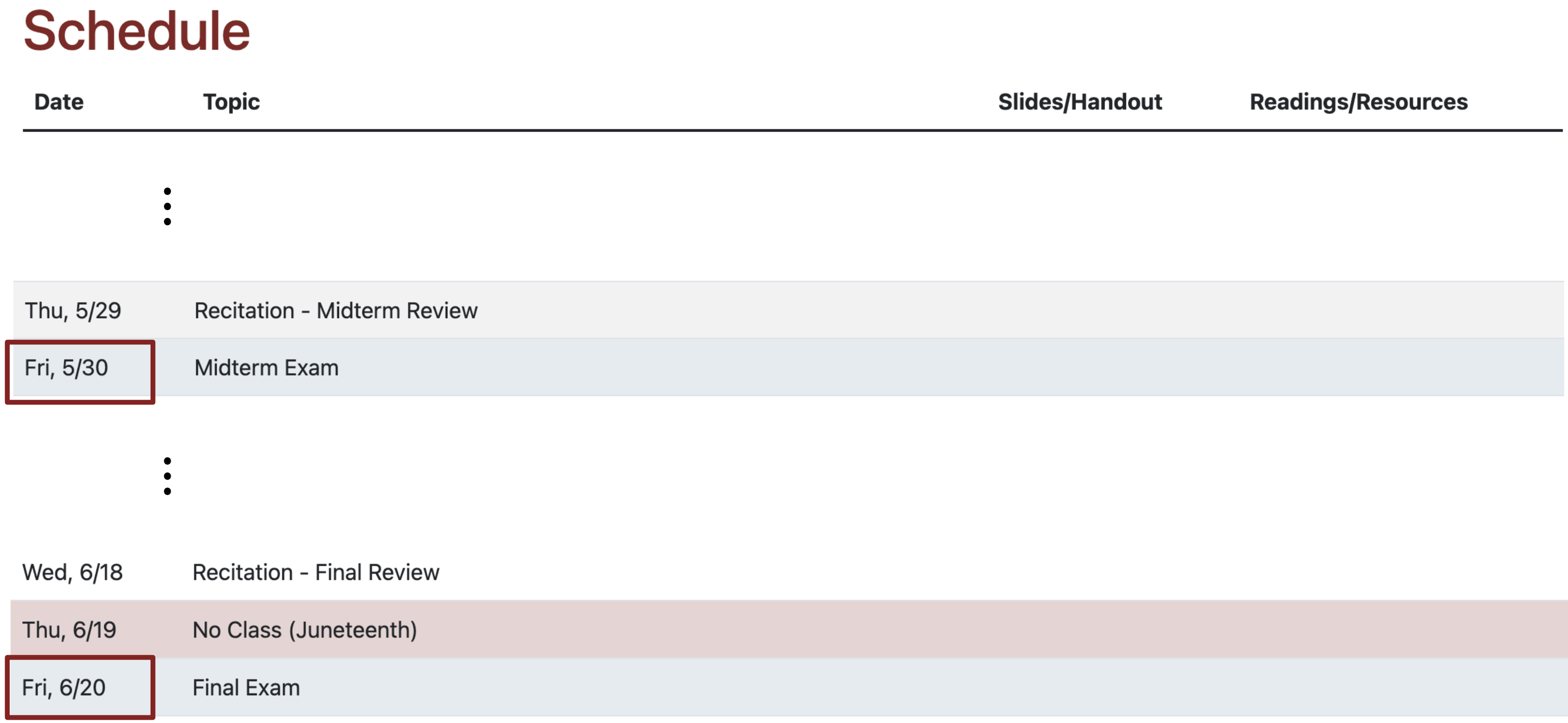


Quizzes

Date	Topic	Study Guide	Quiz
Fri, 5/16	Quiz 1	[study guide]	
Fri, 5/23	Quiz 2		
Fri, 6/6	Quiz 3		
Fri, 6/13	Quiz 4		

https://www.cs.cmu.edu/~hchai2/courses/10601/#Schedule

Logistics: Exam Schedule

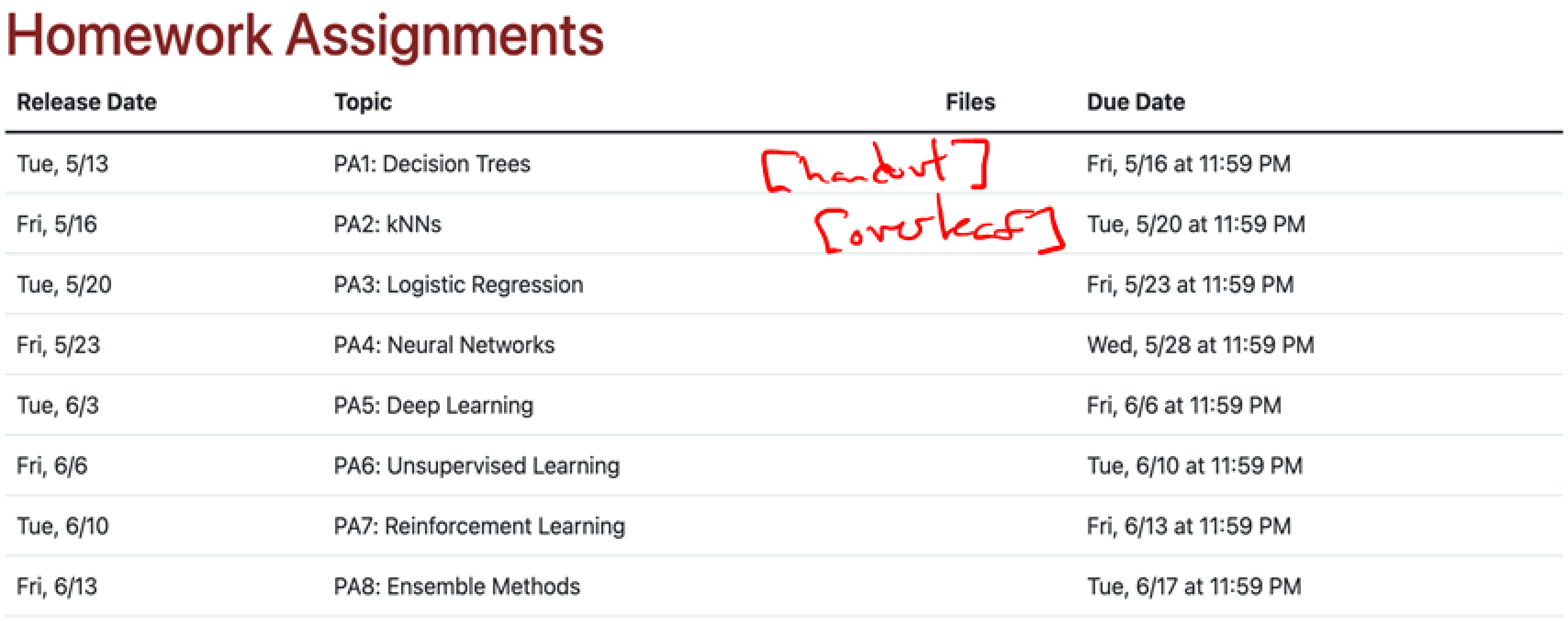


Schedule

Date	Topic	Slides/Handout	Readings/Resources
	⋮		
Thu, 5/29	Recitation - Midterm Review		
Fri, 5/30	Midterm Exam		
	⋮		
Wed, 6/18	Recitation - Final Review		
Thu, 6/19	No Class (Juneteenth)		
Fri, 6/20	Final Exam		

https://www.cs.cmu.edu/~hchai2/courses/10601/#Homeworks

Logistics: Homework Assignments

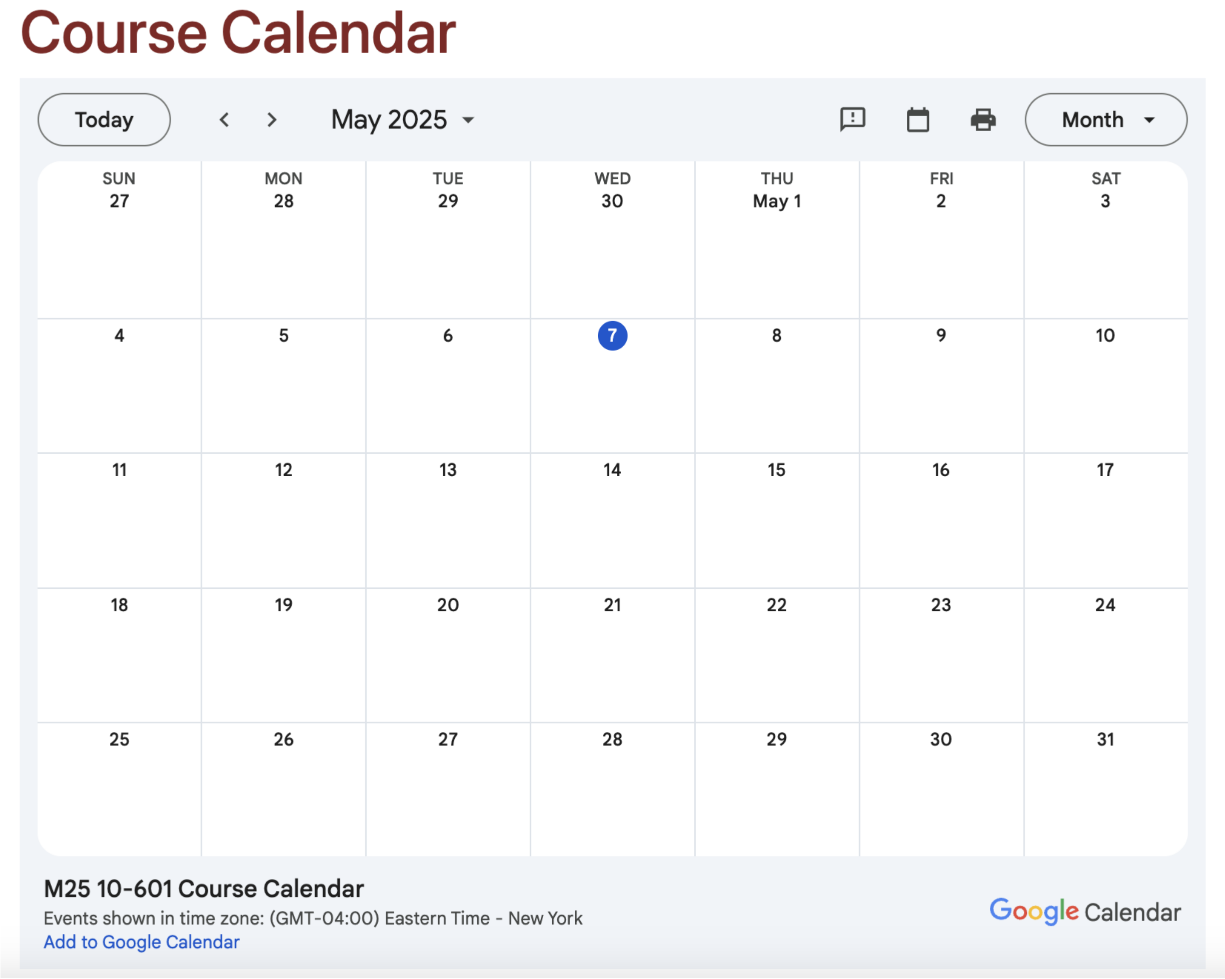


Homework Assignments

Release Date	Topic	Files	Due Date
Tue, 5/13	PA1: Decision Trees	[handout]	Fri, 5/16 at 11:59 PM
Fri, 5/16	PA2: kNNs	[overleaf]	Tue, 5/20 at 11:59 PM
Tue, 5/20	PA3: Logistic Regression		Fri, 5/23 at 11:59 PM
Fri, 5/23	PA4: Neural Networks		Wed, 5/28 at 11:59 PM
Tue, 6/3	PA5: Deep Learning		Fri, 6/6 at 11:59 PM
Fri, 6/6	PA6: Unsupervised Learning		Tue, 6/10 at 11:59 PM
Tue, 6/10	PA7: Reinforcement Learning		Fri, 6/13 at 11:59 PM
Fri, 6/13	PA8: Ensemble Methods		Tue, 6/17 at 11:59 PM

Logistics: Course Calendar

https://www.cs.cmu.edu/~hchai2/courses/10601/#Calendar



Course Calendar

SUN	MON	TUE	WED	THU	FRI	SAT
27	28	29	30	May 1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

M25 10-601 Course Calendar

Events shown in time zone: (GMT-04:00) Eastern Time - New York

Add to Google Calendar