

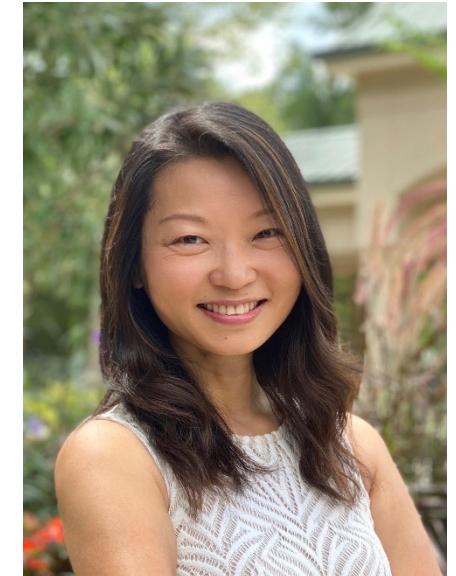
CS334 INTRODUCTION TO MACHINE LEARNING

Li Xiong

Slides adapted from Prof. Joyce Ho

CS334-1

- Lectures: MW 1-2:15pm @W303
- Instructor: Li Xiong
- Office Hours: MW 2:15-3:15PM @E412
- More information: <http://www.cs.emory.edu/~lxiong>



Li Xiong
(Professor)

Assured Information Management and Sharing (AIMS)

(<http://www.cs.emory.edu/site/aims>)



+



+



Undergraduate research opportunities available



EMORY
UNIVERSITY

Honor Thesis Students 2022-2023



CS334-2

- Lectures: TuTh 8:30-9:45am @W303
- Instructor: Joyce Ho
- Office Hours: Tu 1-2pm; Th 2-3pm @W302M
- More information: <https://joyceho.github.io/>



Joyce Ho
(Professor)

TA TEAM



Hong Kyu Lee
(2nd year PhD,
CS534TA,
CS334 TA Fall22)



Tiantian Li
(1st year PhD,
CS334 TA Fall22,
CS334 alumni)



Helen Zeng
(1st year PhD,
CS334 alumni)

Shared between CS334-1 and CS334-2
All office hours posted on Piazza

BREAKOUT: GETTING TO KNOW EACH OTHER (YOUR NEIGHBORS)

- Your name
- Share something unique about yourself
- Name (one of) the favorite machine learning applications you interact with



OVERVIEW OF MACHINE LEARNING AND LEARNING OBJECTIVES

MACHINE LEARNING: EXAMPLE



LI

write a welcome email to the students for an undergraduate machine learning class I'm teaching

GPT

Certainly! Below is a sample welcome email you can send to your students for an undergraduate machine learning class.

Subject: Welcome to the Machine Learning Class of [Semester/Year]!

Dear Students,

I am thrilled to welcome each of you to our Undergraduate Machine Learning class for the [Semester/Year]. My name is [Your Name], and I will be your instructor for this exciting journey into the world of machine learning.

In this course, we will dive into various concepts, algorithms, and tools that are pivotal in the field of machine learning. Whether you're new to the subject or have some previous experience, I'm confident that this class will offer something engaging and challenging for everyone.

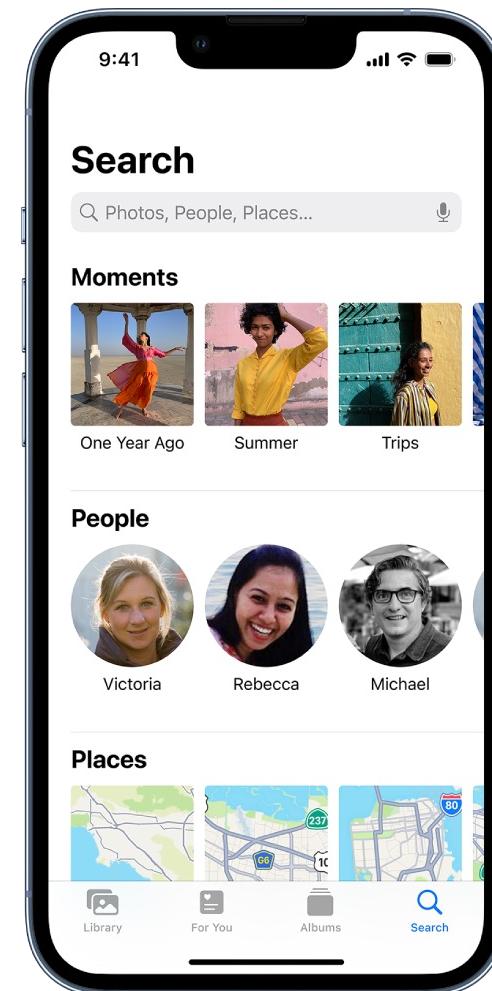
Here's a quick overview of what we'll be covering:

1. **Fundamentals of Machine Learning:** From supervised to unsupervised learning, we'll explore the core concepts that drive the field.
2. **Hands-On Projects:** You'll get the opportunity to work on real-world datasets and implement algorithms using popular programming languages like Python.
3. **Collaboration and Teamwork:** Collaborative projects will allow you to work with fellow students, fostering a sense of community and shared learning.
4. **Guest Lectures and Industry Insights:** We have some exciting guest speakers lined up who will share their real-world experiences and cutting-edge developments in the industry.

MACHINE LEARNING: EXAMPLE

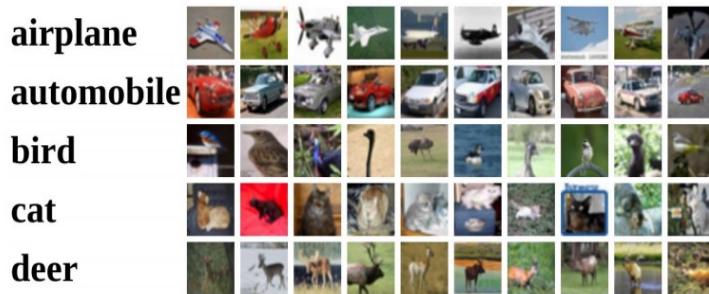


iPhone



MACHINE LEARNING: A DATA DRIVEN APPROACH

Example training set



Learning phase



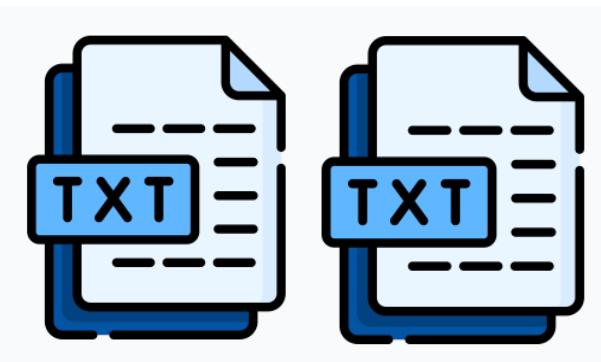
Prediction phase



```
def train(images, labels):  
    # Machine learning!  
    return model
```

```
def predict(model, test_images):  
    # Use model to predict labels  
    return test_labels
```

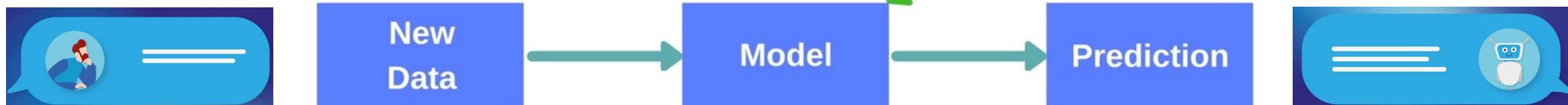
MACHINE LEARNING: A DATA DRIVEN APPROACH



Learning phase



Generation phase



WHAT IS MACHINE LEARNING?

- Arthur Samuel (1959): Machine learning is the field of study that gives the computer the ability to learn without being explicitly programmed.
- Tom Mitchell (1998): A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .



WHAT IS MACHINE LEARNING

- Kevin Murphy (2014): A set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).



“We are drowning in information and starving for knowledge.”

—John Naisbitt

ML IS EVERYWHERE



Facial Recognition



Spam Detection



Recommendations



Medical Diagnosis



Smart Cars



Fraud Detection

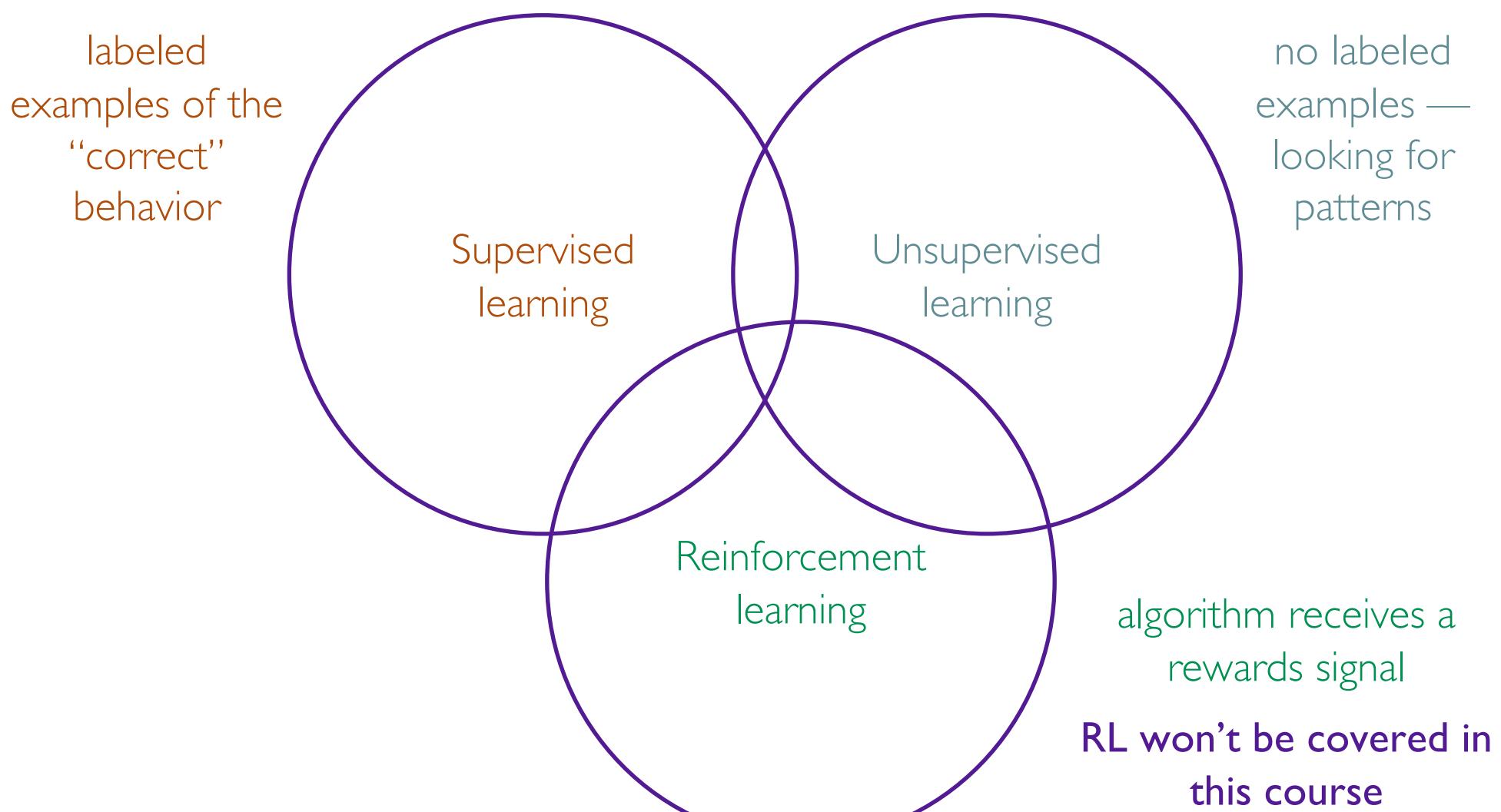


Online Search



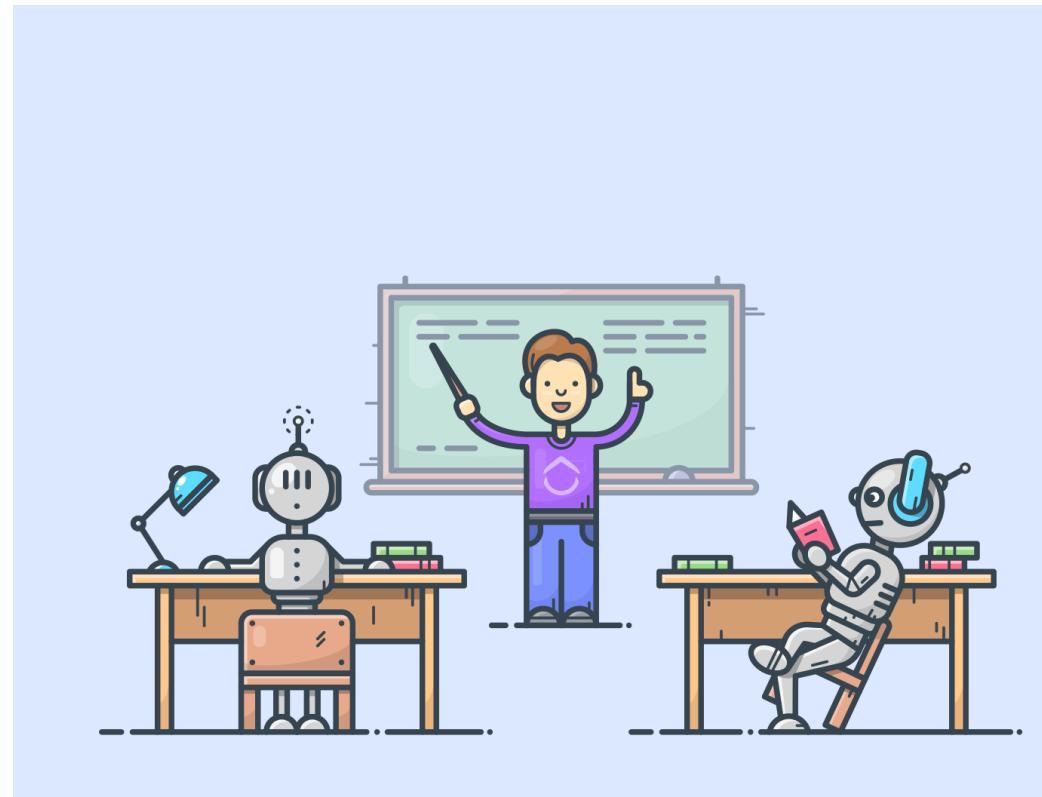
Speech

ML TAXONOMY



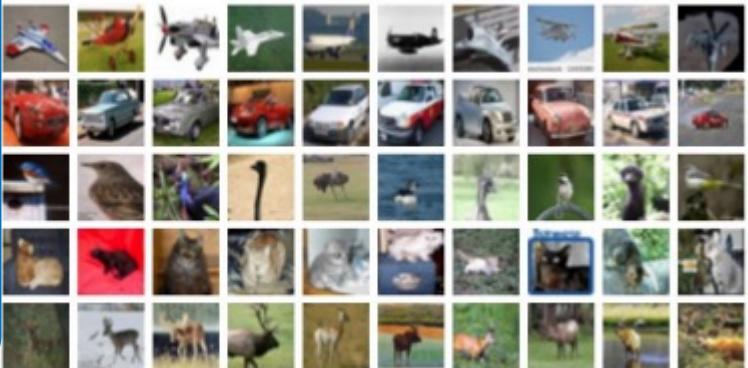
SUPERVISED VS UNSUPERVISED

Supervised



<https://clickup.com/blog/supervised-vs-unsupervised-machine-learning/>

airplane



automobile



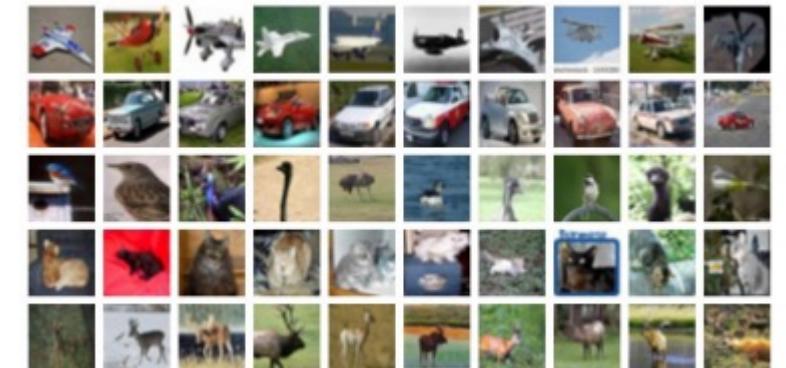
bird



cat



deer



SUPERVISED LEARNING

- Learning a mapping from input to output, given a labeled set of input-output pairs, i.e. training dataset

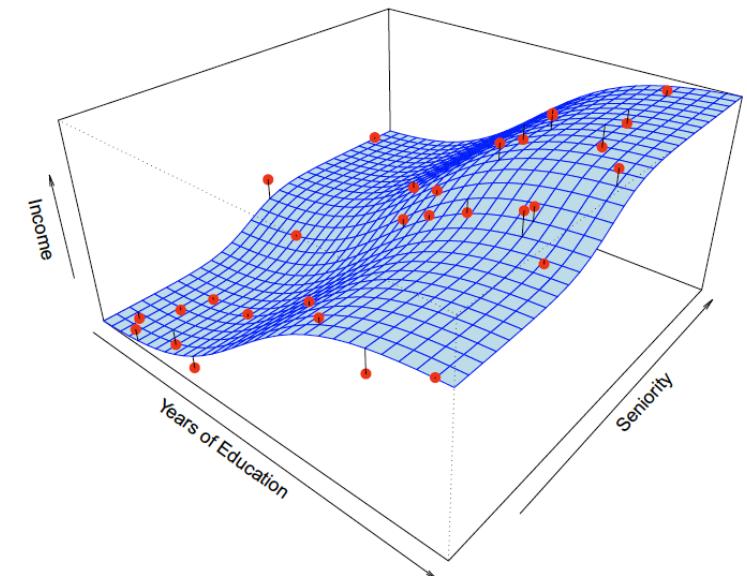
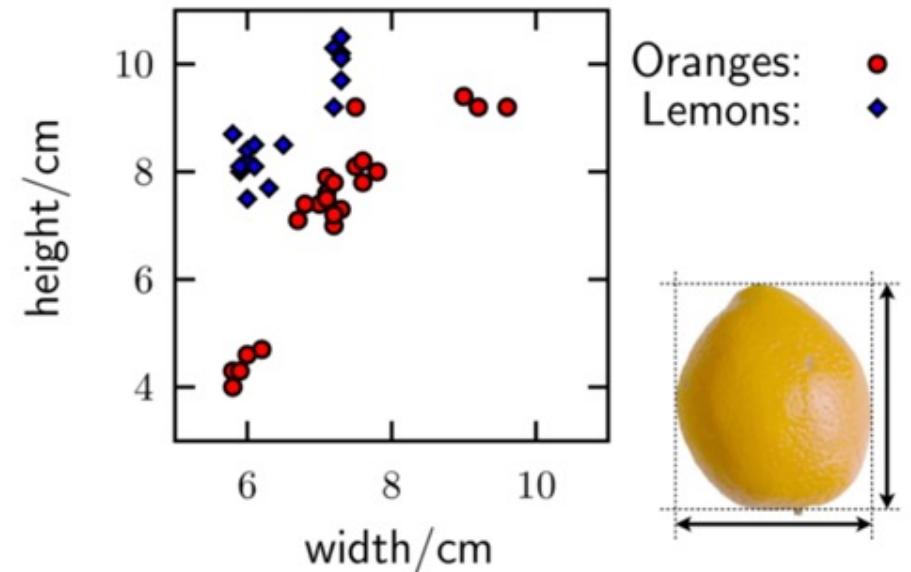
D

$$\{(\mathbf{x}_i, y_i)\}, i = 1, \dots, N$$

- Each input instance represents an object/sample as a p -dimensional vector of features
- Classification**: output is categorical (e.g. orange, lemon)

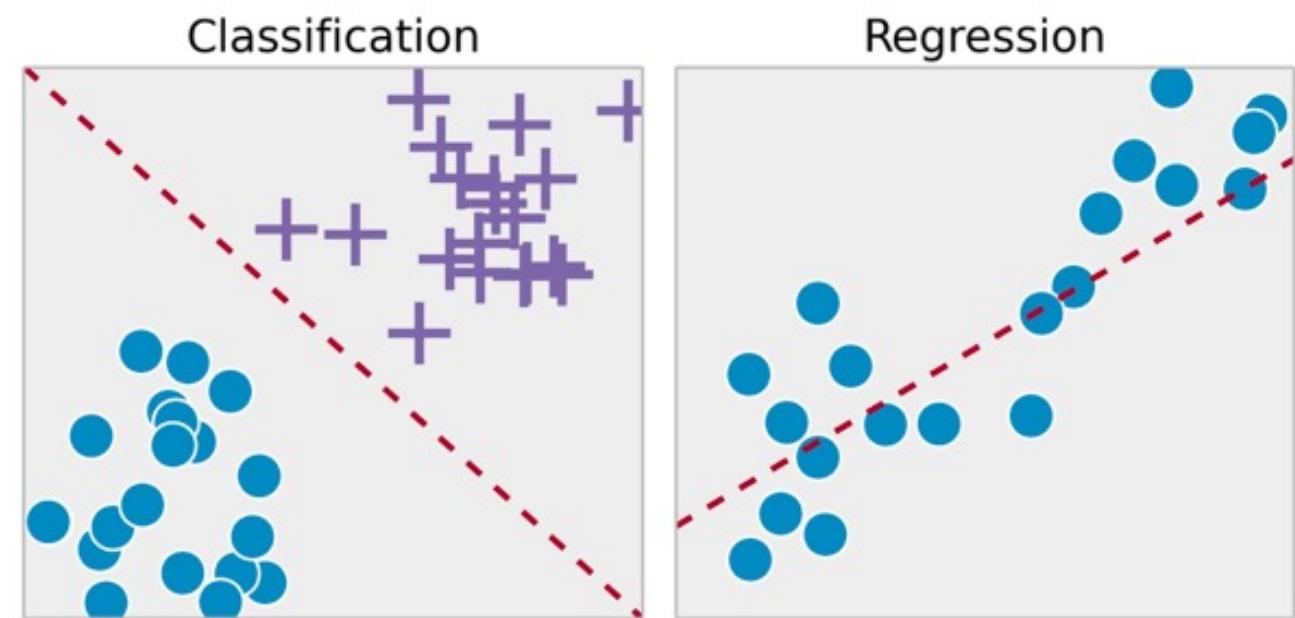
$$y_i \in \{1, \dots, C\}$$

- Binary vs. multiclass classification
- Regression**: output is real-valued



SUPERVISED LEARNING: PREDICTION TASKS

- **Classification:** Predicting qualitative targets (values in a finite set)
- **Regression:** Predicting quantitative responses (continuous valued, natural ordering)



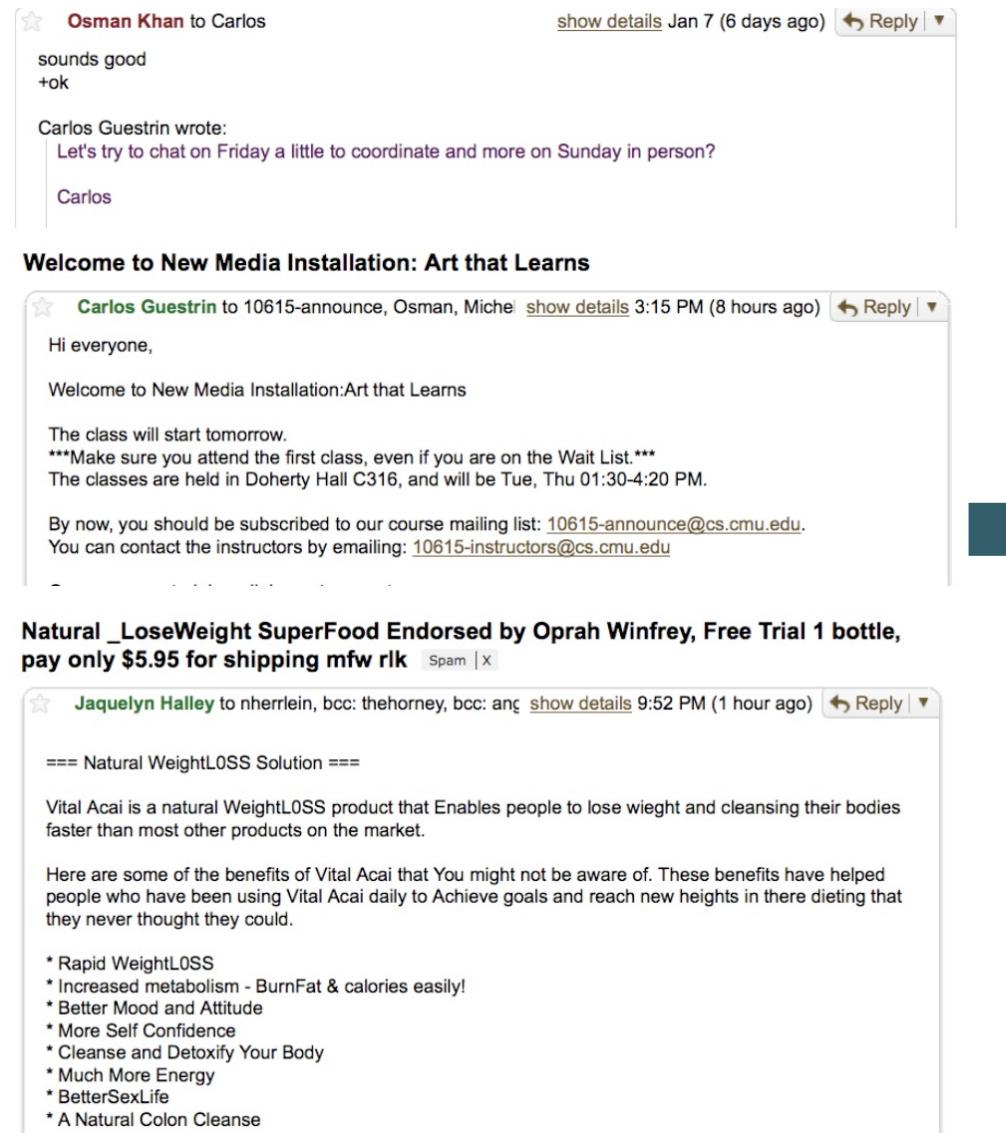
ML VOCABULARY

- Example or Instance: A single data point
- Features (predictors, (independent) variables): Properties of the data
- Target (response, dependent variable): Variable to predict (category or value)
- Label: Target value for a single data point



sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

CLASSIFICATION: SPAM FILTERING



The image displays three examples of email messages, each with a 'Reply' button and a 'spam' link:

- Top Message:** Osman Khan to Carlos. Content: sounds good +ok. Carlos Guestrin wrote: Let's try to chat on Friday a little to coordinate and more on Sunday in person? Carlos. This is labeled as 'spam'.
- Middle Message:** Carlos Guestrin to 10615-announce, Osman, Michel. Content: Hi everyone, Welcome to New Media Installation:Art that Learns. The class will start tomorrow. ***Make sure you attend the first class, even if you are on the Wait List.*** The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM. By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu. You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu. This is labeled as 'not spam'.
- Bottom Message:** Jaquelyn Halley to nherlein, bcc: thehorney, bcc: anç. Content: === Natural WeightLOSS Solution === Vital Acai is a natural WeightLOSS product that Enables people to lose weight and cleansing their bodies faster than most other products on the market. Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in their dieting that they never thought they could.
 - * Rapid WeightLOSS
 - * Increased metabolism - BurnFat & calories easily!
 - * Better Mood and Attitude
 - * More Self Confidence
 - * Cleanse and Detoxify Your Body
 - * Much More Energy
 - * BetterSexLife
 - * A Natural Colon CleanseThis is labeled as 'spam'.

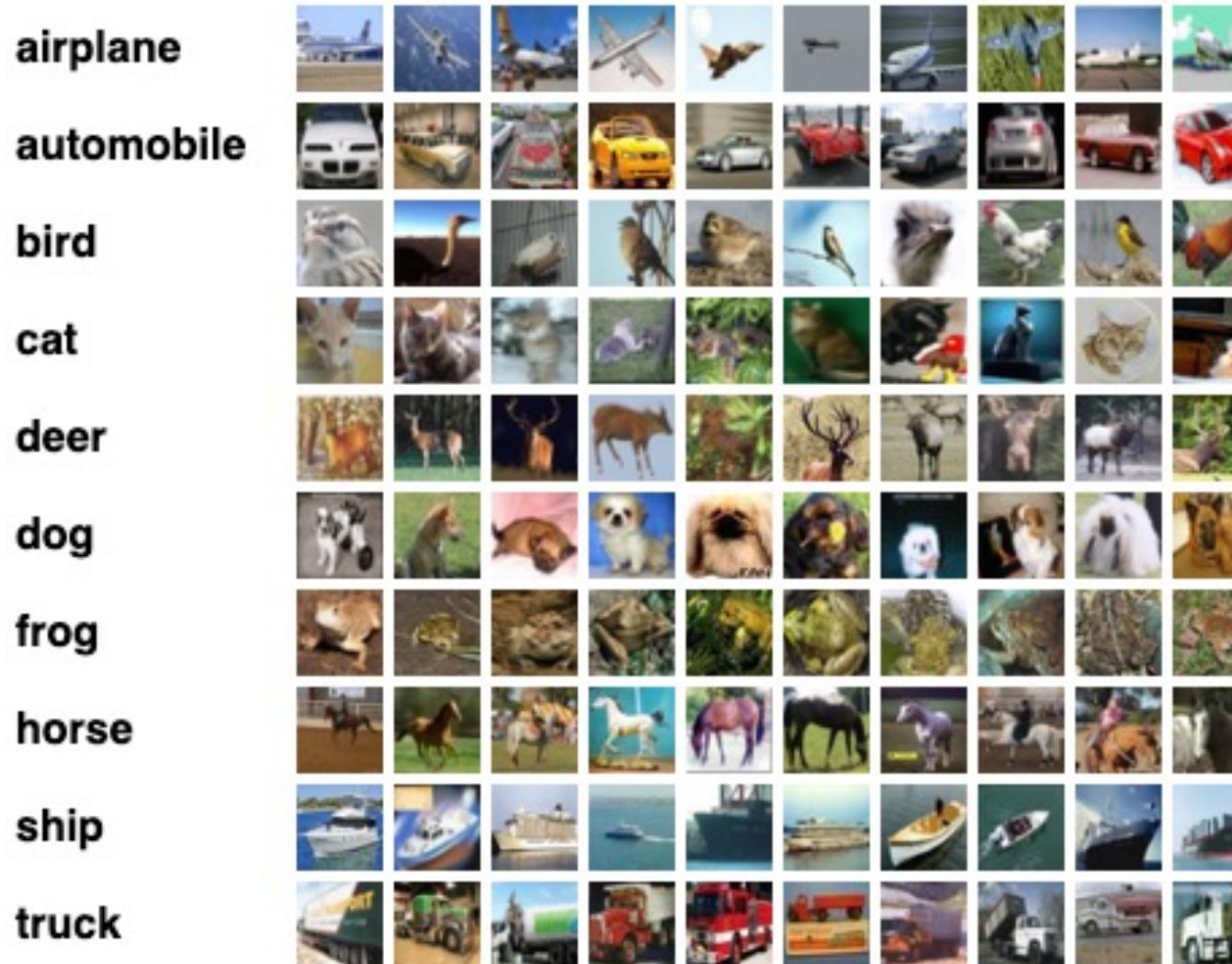


spam
vs
not spam

CLASSIFICATION: DIGIT RECOGNITION



CLASSIFICATION: IMAGE CLASSIFICATION



REGRESSION: PREDICTING STOCK PRICE

Market Summary > Tesla Inc

889.36 USD

+ Follow

+180.87 (25.53%) ↑ past year

Closed: Aug 23, 7:59 PM EDT • Disclaimer

After hours 889.60 +0.24 (0.027%)

1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max



Open

874.36

Mkt cap

928.93B

52-wk high

1,243.49

High

896.48

P/E ratio

106.75

52-wk low

620.57

Low

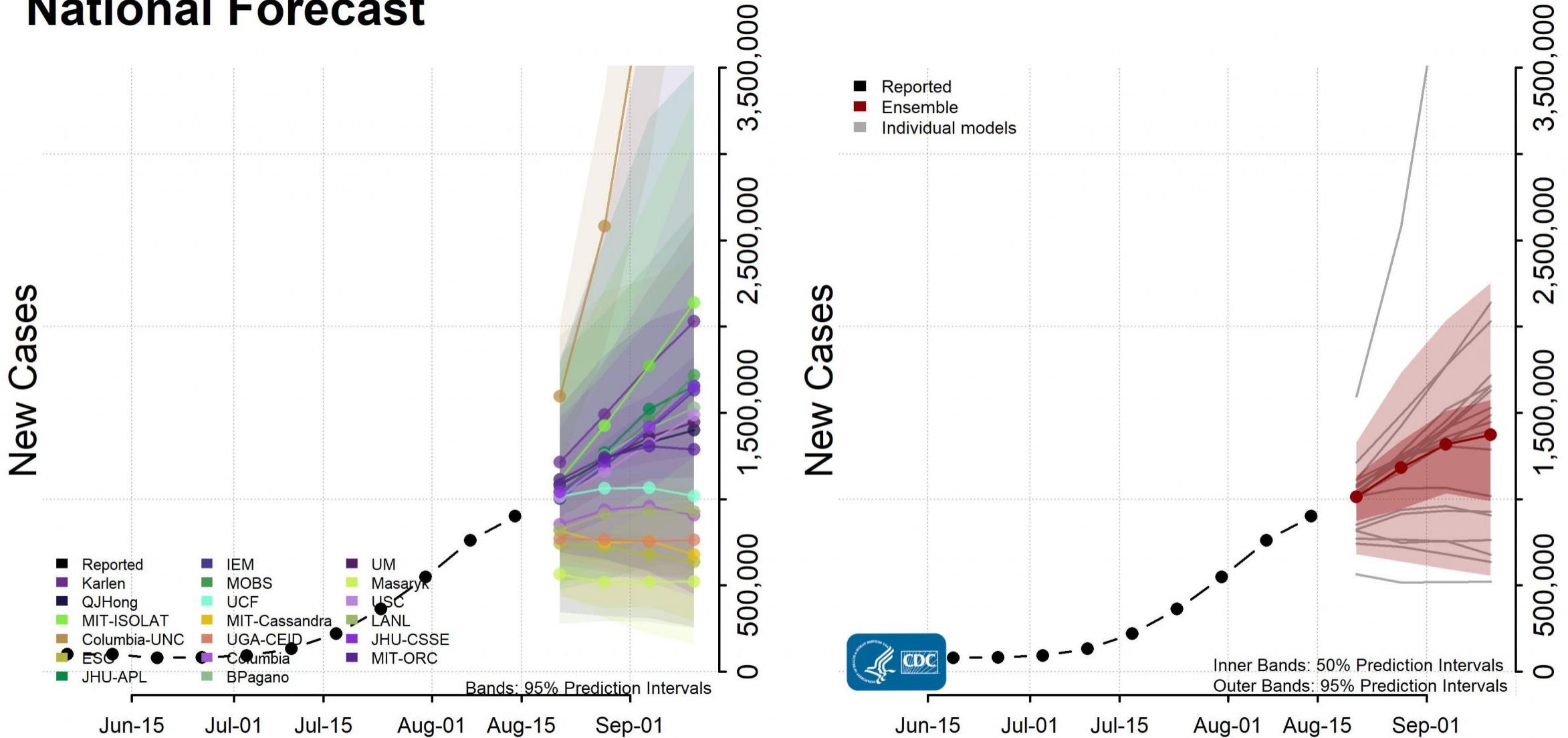
863.77

Div yield

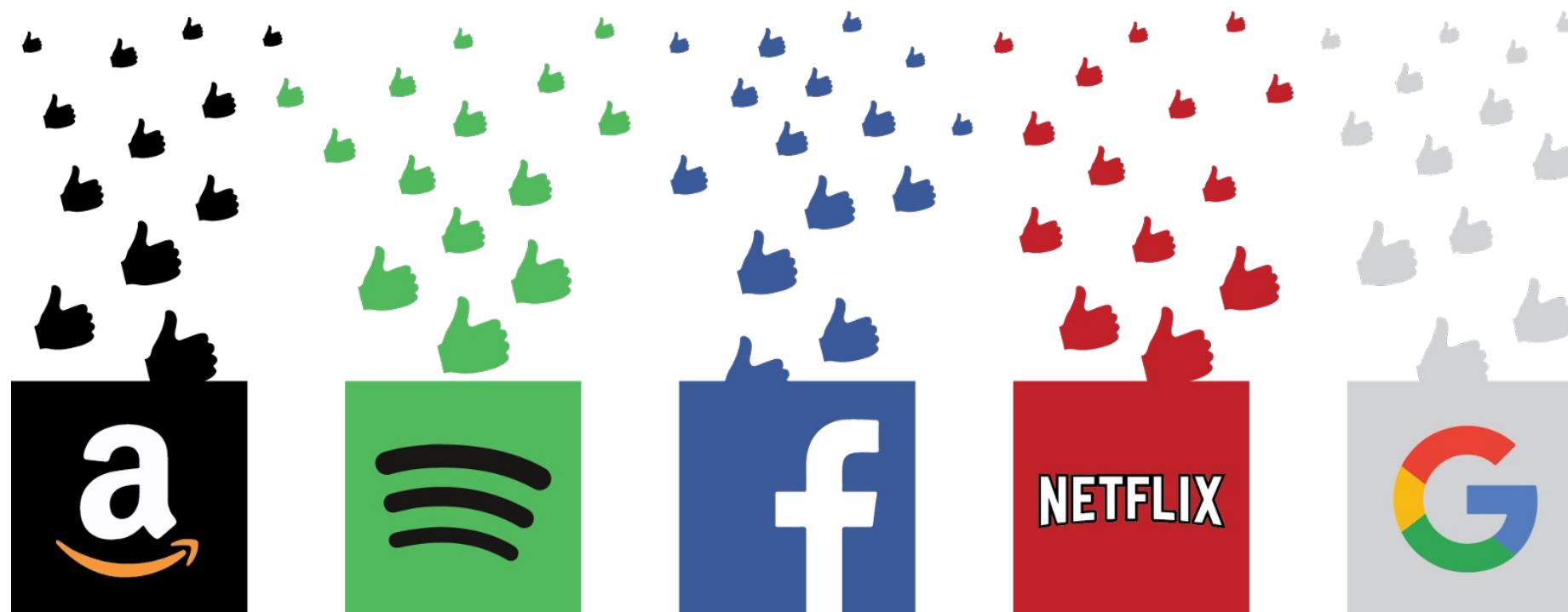
-

COVID-19 FORECASTS: CASES

National Forecast



CLASSIFICATION/REGRESSION: RECOMMENDATION SYSTEMS



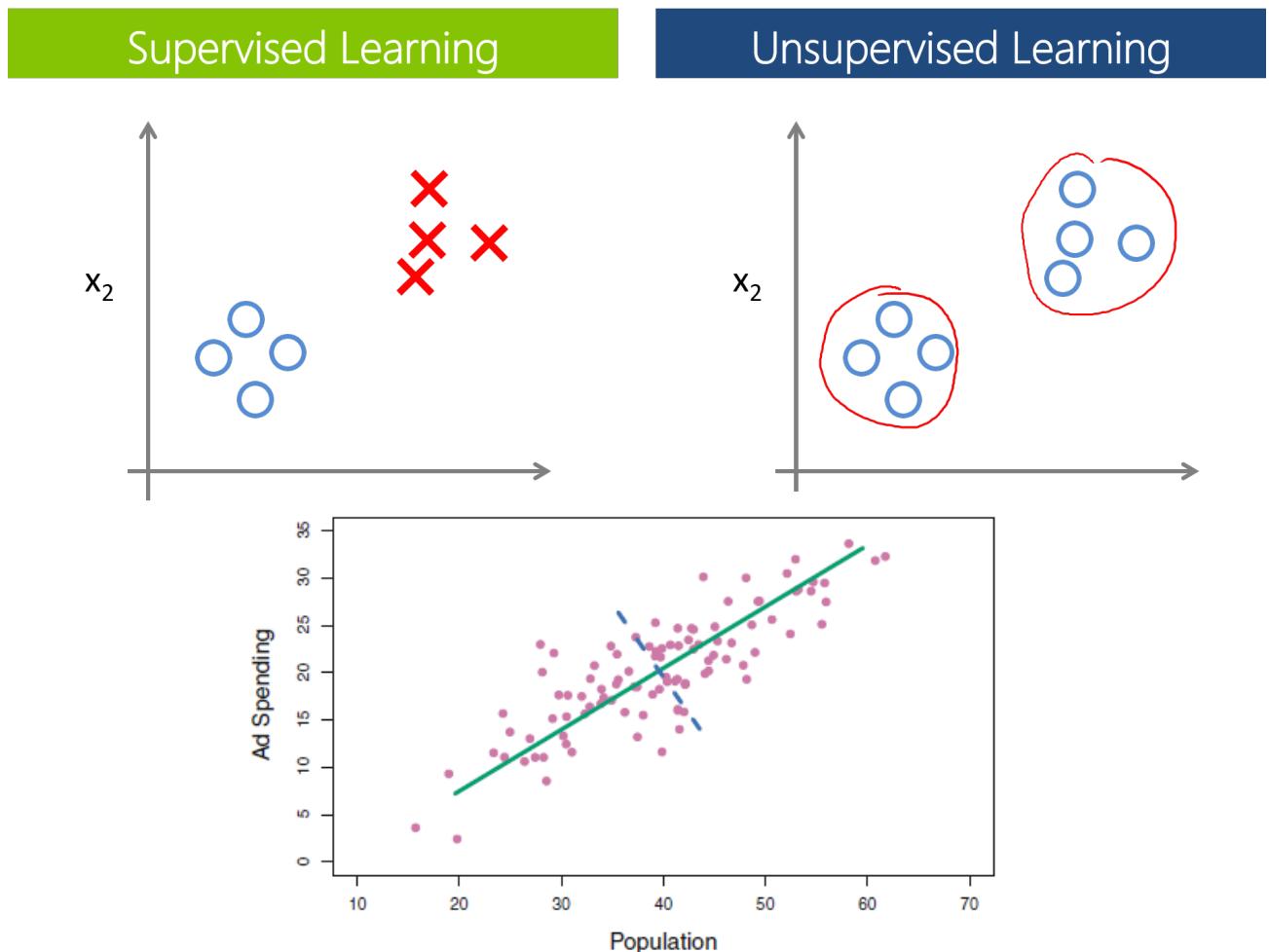
<https://www.districtdatalabs.com/the-algorithm-issue-1-recommender-system-intro>

UNSUPERVISED LEARNING

- Finding interesting patterns given inputs

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$$

- Applications: clustering; visualization; anomaly detection; dimension reduction; association analysis



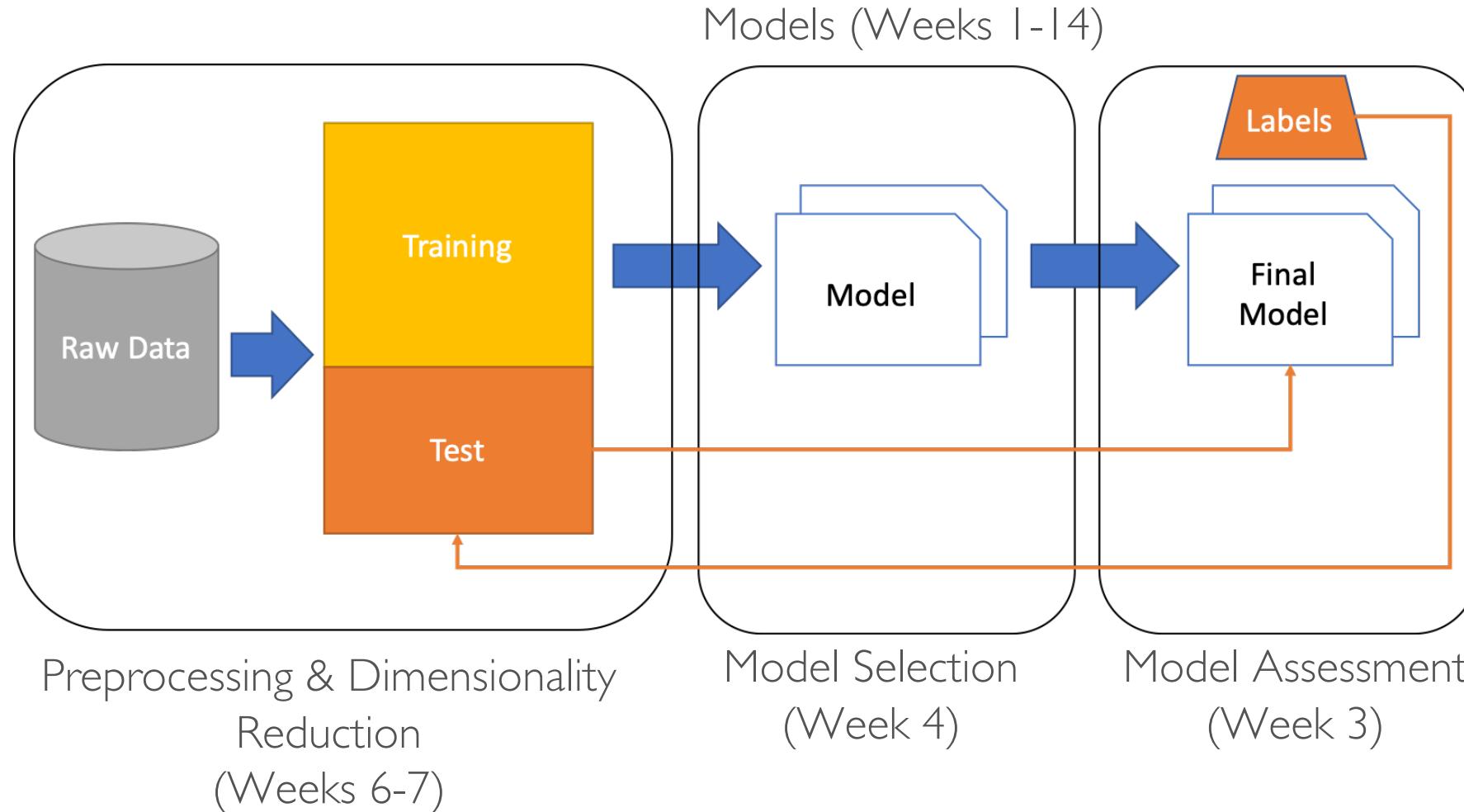
LEARNING OBJECTIVES

- Understand the basic building blocks and general principles that underlie machine learning algorithms;
- Understand and **implement** specific, widely used machine learning algorithms;
- Learn methodology and tools to **apply** machine learning algorithms to real data and evaluate their effectiveness and performance.

COURSE OUTLINE

- Algorithms for **supervised learning**: nearest neighbors, decision trees, linear regression, logistic regression, neural networks, naïve bayes, ~~support vector machines~~, ensembles, boosting, **deep learning**
- Algorithms for **unsupervised learning**: principal component analysis, ~~clustering~~
- Model assessment and model selection
- New learning paradigms and emerging topics

ML WORKFLOW / ROADMAP



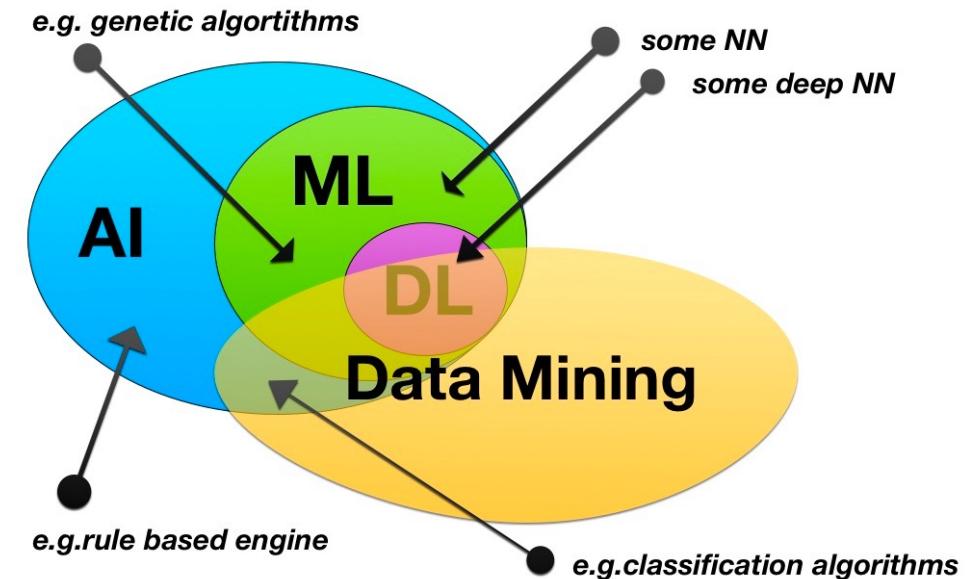
MACHINE LEARNING: BASICS

- Prerequisites
 - Basic programming and computing principles (CS 170 + 171)
 - Data structures and algorithms (CS 224 + 253)
 - Linear algebra (Math 221)
 - Python programming (workshop upcoming!)



WHAT THIS COURSE IS NOT

- Deep learning (**new course coming!**)
- Large language models (e.g., chatGPT)
- Computer vision
- Learning all the popular machine learning models at a high-level only (intro to AI)



<https://softwareengineering.stackexchange.com/questions/366996/distinction-between-ai-ml-neural-networks-deep-learning-and-data-mining>



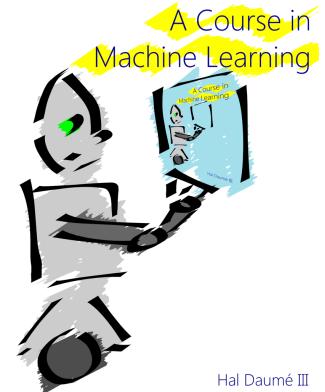
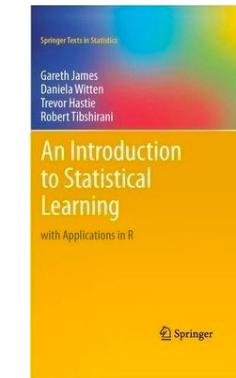
COURSE LOGISTICS

COURSE COMMUNICATION

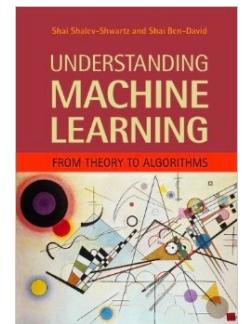
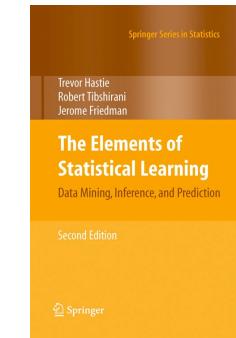
- **Canvas** (section-specific syllabus, announcements, schedule, lecture slides, grades):
<https://canvas.emory.edu/courses/120597>
- **Piazza** (shared announcements, homework handouts, solutions, Q/A):
<http://piazza.com/emory/fall2023/cs334>
- **Gradescope** (homework submission/feedback, **sync'ed with Canvas**):
<https://www.gradescope.com/courses/569901> (entry code: **ZZN4JW**)

TEXTBOOK(S)

- Required (R): An Introduction to Statistical Learning with Applications in Python, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani & Jonathon Taylor
- Supplemental (S1): A Course in Machine Learning, by Hal Daumé III
- Supplemental (S2): A First Encounter with Machine Learning, by Max Welling
- Supplemental (S3): The Elements of Statistical Learning: Data Mining, Inference, and Prediction, by Trevor Hastie, Robert Tibshirani & Jerome Friedman
- Supplemental (S4): Understanding Machine Learning: From Theory to Algorithms



Hal Daumé III



PDF is freely available online!

COURSE EVALUATION

- 5 Homeworks (50%)
 - Roughly 80% programming, 20% conceptual
- Midterm Exam (25%) **No final exam!**
- Focus on ML theory / conceptual understanding
- Project (25%)

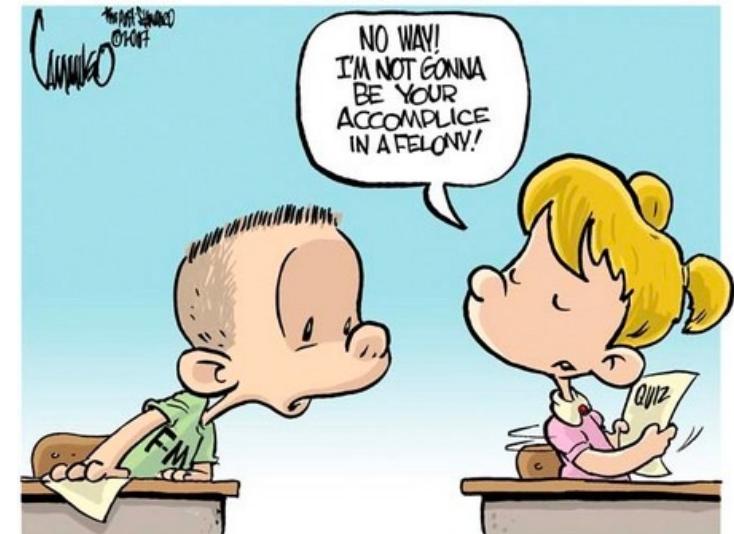
Grades will only be curved (up) so the class mean falls at least in a B range

LATE HOMEWORK POLICY

- 6 flexible “24-hour late days” to be distributed across 5 assignments
- No more than **3** late days per assignment
- Tracking based on submission date
- **Zero credit** if you have NO MORE LATE DAYS
- Additional extension requires documentation from OUE

HONOR CODE

- Write-up solutions on your own (independently)
- Attach a README file that contains the honor code and any source of collaboration (classmates or online resources)

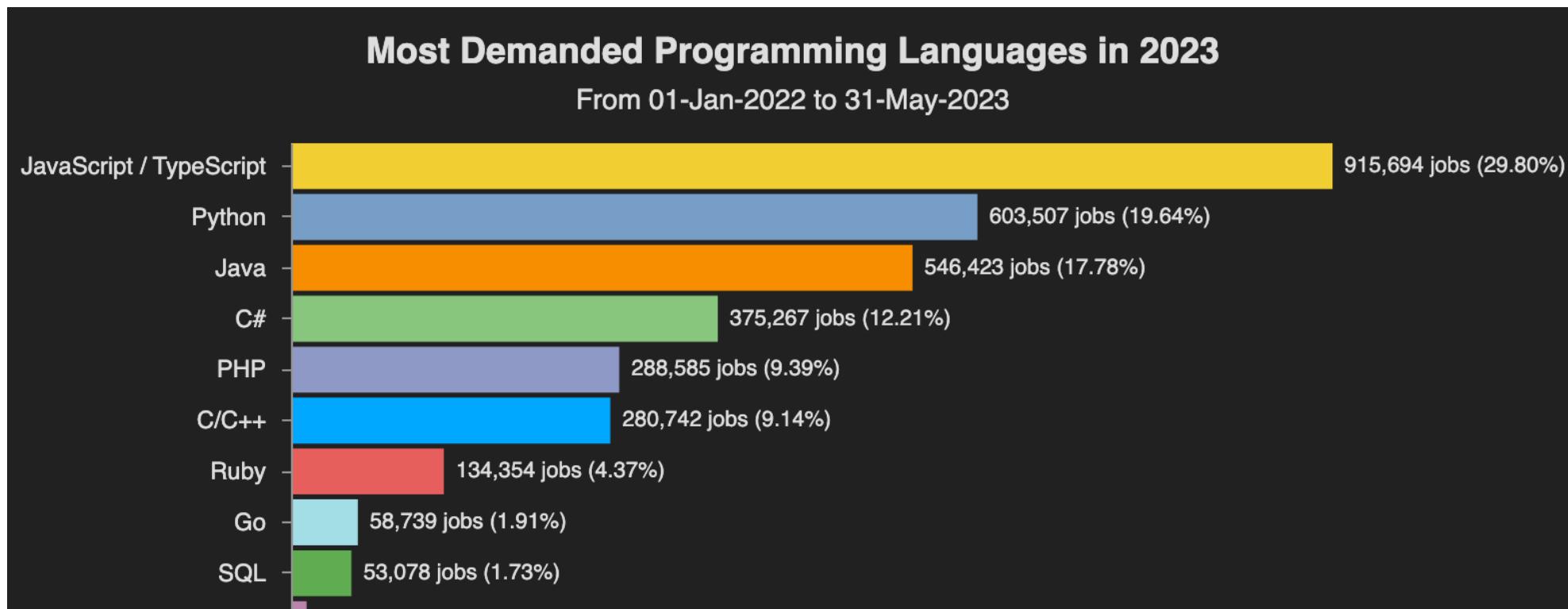


`/* THIS CODE IS MY OWN WORK, IT WAS WRITTEN WITHOUT
CONSULTING CODE WRITTEN BY OTHER STUDENTS OR LARGE
LANGUAGE MODELS LIKE CHATGPT. Your_Name_Here */`

CODE MUST BE IN PYTHON

- Rich data science/ML libraries
- Homeworks will be (partially) autograded

<https://www.devjobsscanner.com/blog/top-8-most-demanded-programming-languages>



PYTHON WORKSHOP FOR ML/DATA SCIENCE

1. Environment Setup + Basic python
 - a. Data structures (dictionaries, tuple, set, etc.)
 - b. Useful syntaxes (Lambda function, Decorators, etc.)
 - c. Basic Numpy
 - d. Basic Matplotlib
2. General ML pipeline
 - a. Data loading, cleaning, train/test split
 - b. Preprocessing (Normalization, etc.)
 - c. Selecting the right model
 - d. Model training
 - e. Model testing (Choose the right evaluation metric)
 - f. Visualizing results
3. LAB (Get your hands dirty)
 - a. Basic ML
 - b. Debugging



Hong Kyu Lee
(2nd year PhD,
CS334/534 TA)



Chase Fensore
(2nd year PhD,
CS534TA)

PROJECT

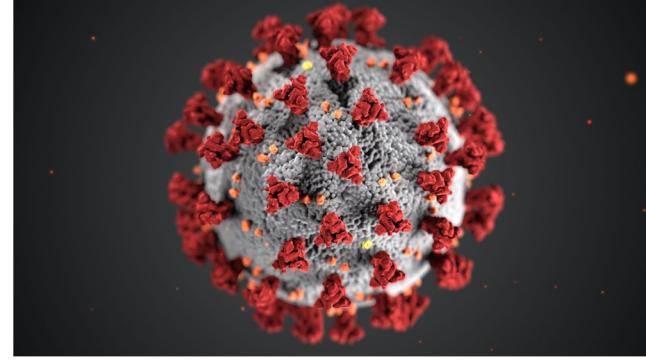
- Work in groups of 2-3
- Goal: Apply ML to real-word tasks or prepare you for ML/AI research
- Can be any topic you choose
- Emphasis on large, publicly available datasets (e.g., Kaggle, KDD Cup, etc.)

PROJECT: DETAILS

- Written proposal: 1-2 pages of problem, dataset, what you plan to do
- Spotlight: ~2 min short overview of your project (similar to a lightning talk / elevator pitch)
- Presentation: ~6-10 minute conference-like talk about your project
- Report: 6-12 pages (1-column) written report
- Best projects voted by class (bragging rights!)



A MACHINE LEARNING APPROACH TO PREDICTING COVID-19 RISK FROM DEMOGRAPHIC FACTORS AND SOCIAL DETERMINANTS



Alex Welsh and Bhargav Annigeri

Datasets:

- Kaggle Uncover COVID-19 Challenge
 - Contains county level data on cases, deaths, testing & hospitalization
- NYU Langone Health Demographic Data
 - Tracks social, economic, and racial factors for major cities in each state
- Plan is to combine data

Pre-processing:

- Feature selection will likely be an involved process
 - Choosing social determinants and factors to hone in on
- Combine county level COVID data into some metric of severity/risk for our output
- Will need to partition data into train and test
- Look for confounding features with Pearson correlation and remove them

Model Evaluation:

- Since our model will likely solve a regression problem, and in accordance with previous research, we will use RMSE and R^2 to evaluate the accuracy of our model.
- Use Lasso regularization to help prevent overfitting and improve model accuracy.

DATASET

FROM KAGGLE

160k+ songs on Spotify

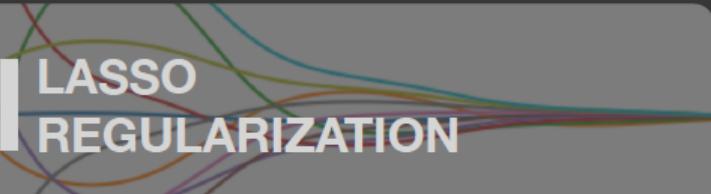
4 categorical features

15 numerical features

Target: numerical scaled popularity



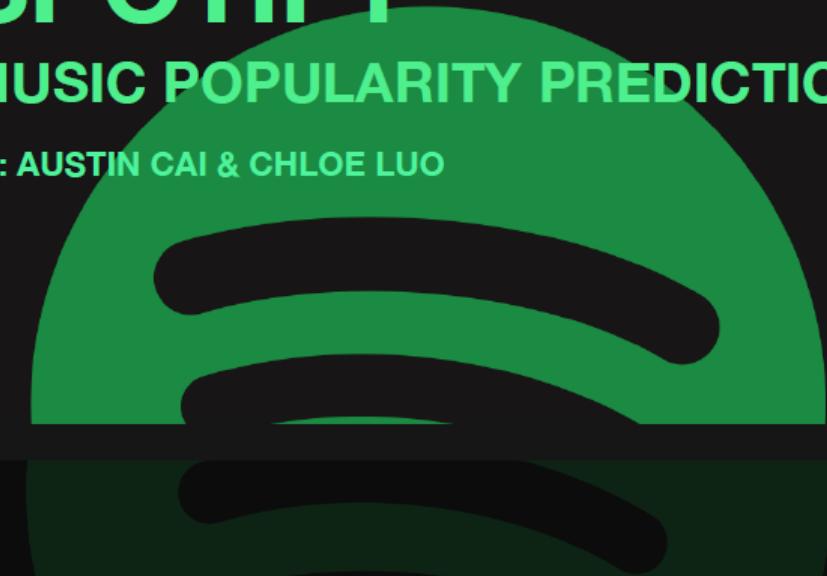
FEATURE SELECTION & PREPROCESS



What does it take
to make a **POPULAR** hit song?

SPOTIFY MUSIC POPULARITY PREDICTION

by: AUSTIN CAI & CHLOE LUO



We are music lovers and we want to apply our knowledge to something we love, so we decided to use a machine learning model to predict songs' popularity with its tempo, key, danceability, instrumentalness, and etc.

EVALUATION

K-FOLD
CROSS
VALIDATION

ACCURACY

We will use k-fold to validate our model

Also we will check the accuracy of learning model with the actual popularity value

LR with SGD

- Linear Regression is better for predicting numerical results
- SGD will be applied since we have a large set of data

MODEL

PREDICTING PRICE POINT FOR DIFFERENT WINES

Question: Can we predict the price of wine based on its features?

Why: Wine wholesalers would be interested in this to better set their prices

Dataset

- 13 features and 150k entries
 - Features including: Points (rating), Wine Description, Country, Province, Region, Winery, Designation, Price (Which will be used as a label)
- Data provided by the website "Wine Enthusiast". Test and Train are not split beforehand.

Methods

- Feature extraction and selection
 - Translate wine description into one-hot encoding of of most frequently used descriptive terms
 - Use Pearson correlation for Feature Selection
- Use Linear regression and Neural Network models to predict Price point while tuning hyperparameters with K folds for optimized performance
- Check relative performance of models using accuracy

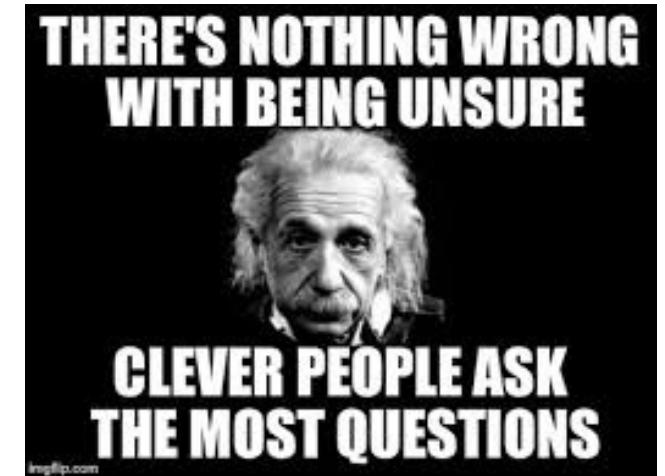


What has been done?
What problems have been found?

- Predicting Score
- Data Correlation
- Data visualization (France produces the most expensive wines)
- Special characters in dataset

CLASS EXPECTATION

- No attendance required
- Class participation encouraged (Q/A, breakout discussion)
- Piazza discussions, Q/A
- Start homeworks and projects early, and get help early





QUESTIONS?