# Machine Learning for Large-Scale Data Analysis and Decision Making (MATH80629A) Winter 2023

Week #2 - **Summary**


# Quiz 0

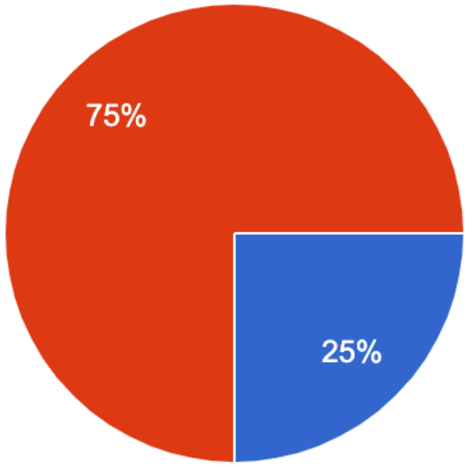
Login to your Gradescope account

# Announcement

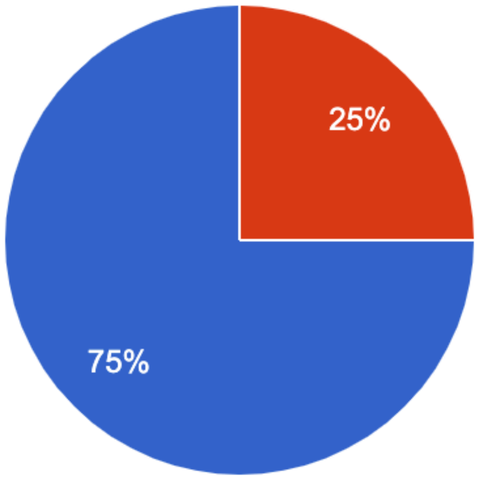
- My office hour is right after the class on **Fridays 3-4 pm**
    - Office: 4.834
- TA Office hours (William and Afaf) will be on **Wednesdays 11 am-12 pm.** (They have announced the details on Piazza.)
    - Online: https://bluejeans.com/560322885/2910

# Class statistics

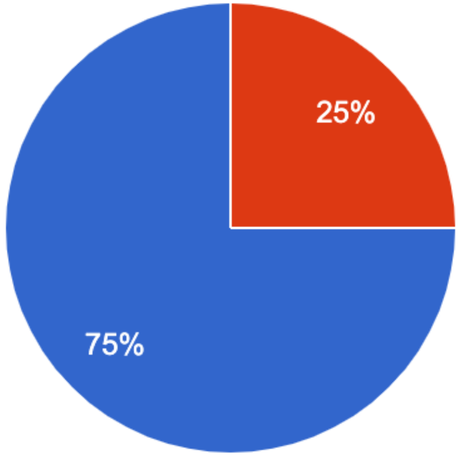
**Gender**



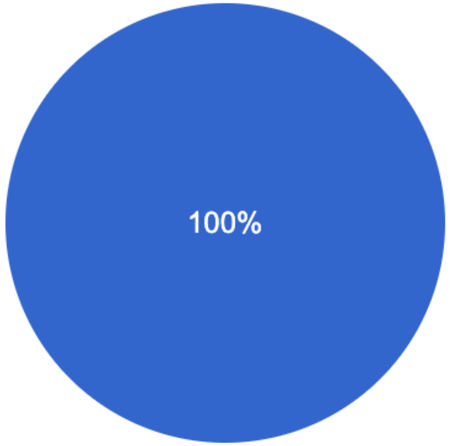

**Python**




**ML**




**Laptop**




Student Introduction Survey form **due January 20, 2023**

Team Registration form **due January 20, 2023**

# Announcement

- Two Python programming labs by TAs:
    - **Intro to Python programming**
        - **Time:** Thursday, January 18 2023, at 1pm
        - **Location**: TBA on Piazza
        - Will be recorded and the recording will be on Piazza after the lab
    - **Intro to PyTorch**
        - **Time:** Thursday, February 1 2023, at 1 pm
        - **Location**: TBA on Piazza
        - Will be recorded and the recording will be on Piazza after the lab

# Today

- **Test Quiz** on Gradescope!
- **BE PREPARED** for next week! We will have a quiz almost every week at the beginning of the class. You can check the schedule on the website.
- **Quiz 1 will be next week, based on Machine learning fundamentals & Supervised learning algorithms**
- Summary of Machine learning fundamental
- Q&A
- Hands-on session

# Machine Learning Problem

The three components of an ML problem:

1. **Task.** What is the problem at hand?

   - Model. How are you parametrizing your solution.

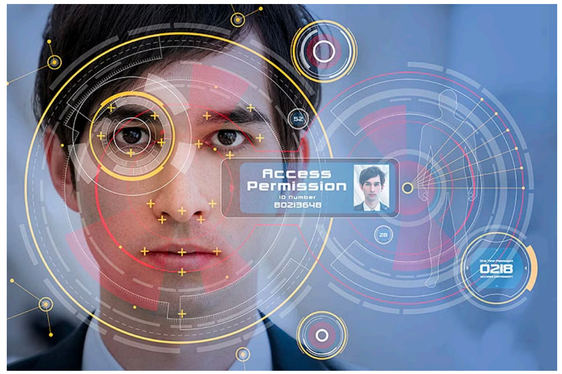
2. **Performance.** How well you are doing?

3. **Experience.** What kind of data do you have access to?

# Types of Experiences

- **Supervised {(x,y)}.** e.g., regression, classification. f: X -> Y
- **Unsupervised {(x)}**. e.g., clustering, dim. reduction, density estimation
- **Reinforcement learning**. Agent takes actions in an environment.

# Applications

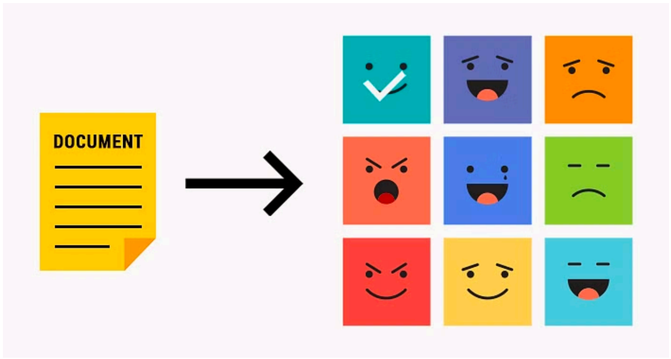



Face recognition

Fraud detection

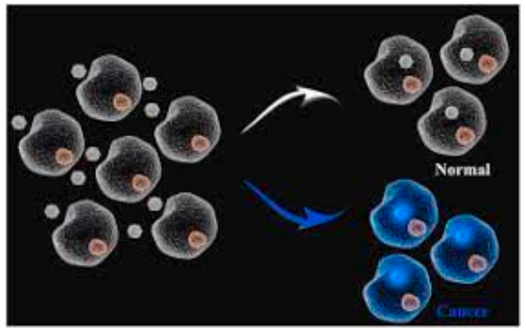

Weather forecast

Autonomous vehicle




Sentiment Analysis
(Emotion Detection)


Cyberbullying detection

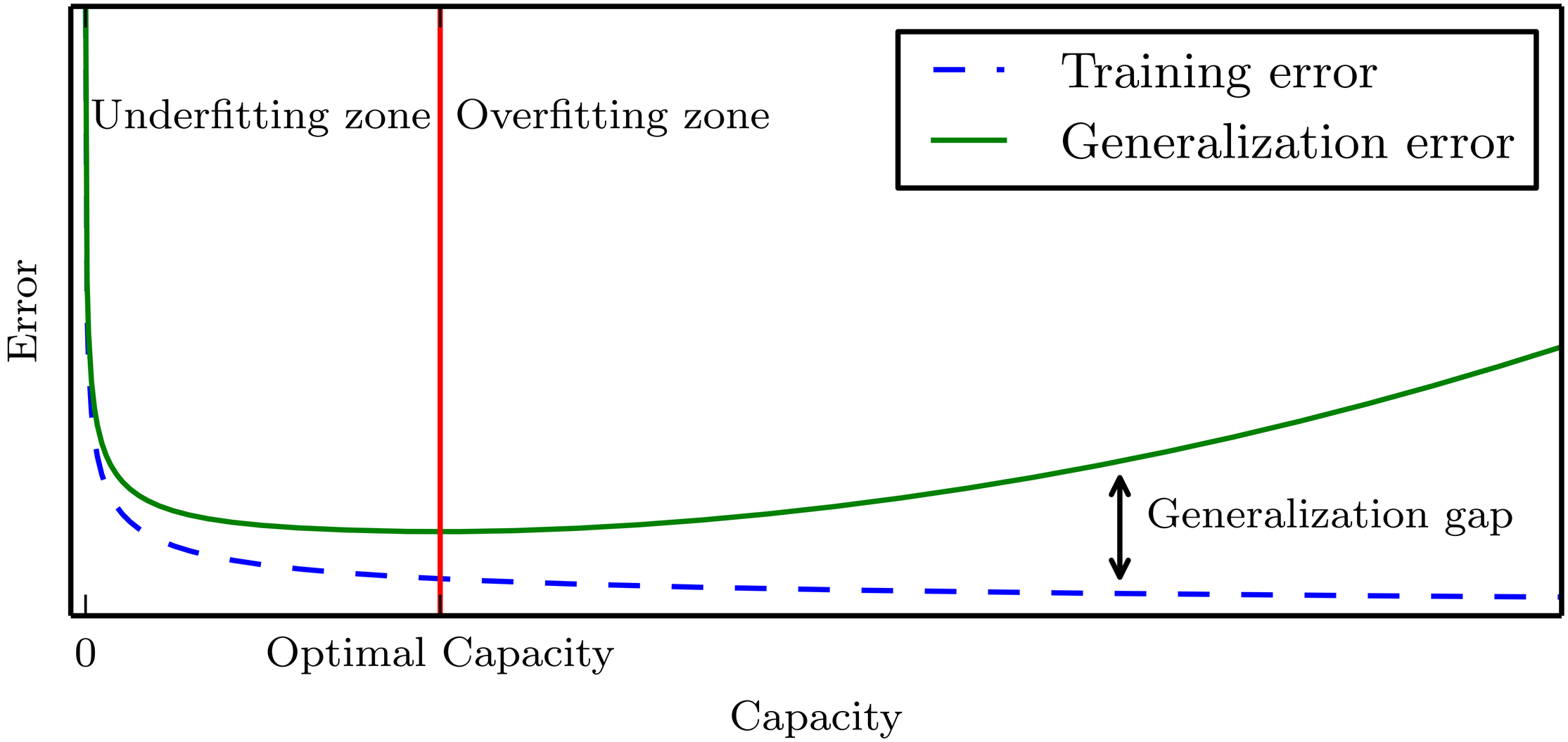



Cancer cell detection

Market Segmentation

# Model Evaluation

- **Given:**
  - A performance measure, e.g., MSE
  - A train dataset
  - A model, e.g., Linear regression
- **Can calculate:**
  - Train error: used to learn (to train).
  - Train error cannot be used to evaluate your model
  - Must use a separate dataset for evaluation

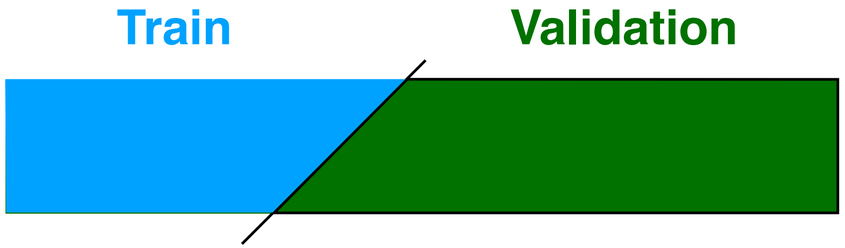

Underfitting zone
Overfitting zone
Training error
Generalization error
Error
Generalization gap
0
Optimal Capacity
Capacity

# Regularization

- Can be thought of as way to limit a model's capacity
- $\text{Loss} := \text{MSE}^{\text{train}} + \underbrace{\lambda \mathbf{w}^\top \mathbf{w}}_{||W||_2}$

# Validation set

- How do we choose the right model and set its hyper parameters (e.g. $\lambda$)?
  - **Use a validation set**
    - **Split the original data into two:**
      1. Train set
      2. Validation set
         - Proxy to the test set




  - **Train different models/hyper-parameter settings on the train set**
  - **Pick the best according to their performance on the validation set**

# Bias / Variance

- The goal is to hit the bull's eye
- Each blue dot represents the "performance" of a fixed model on different data from the same distribution