Machine Learning for Large-Scale Data Analysis and Decision Making (MATH80629A) Winter 2022

Week #2 - **Summary**

Announcement

- Class will most likely be online throughout this semester, but final exam will be in person.

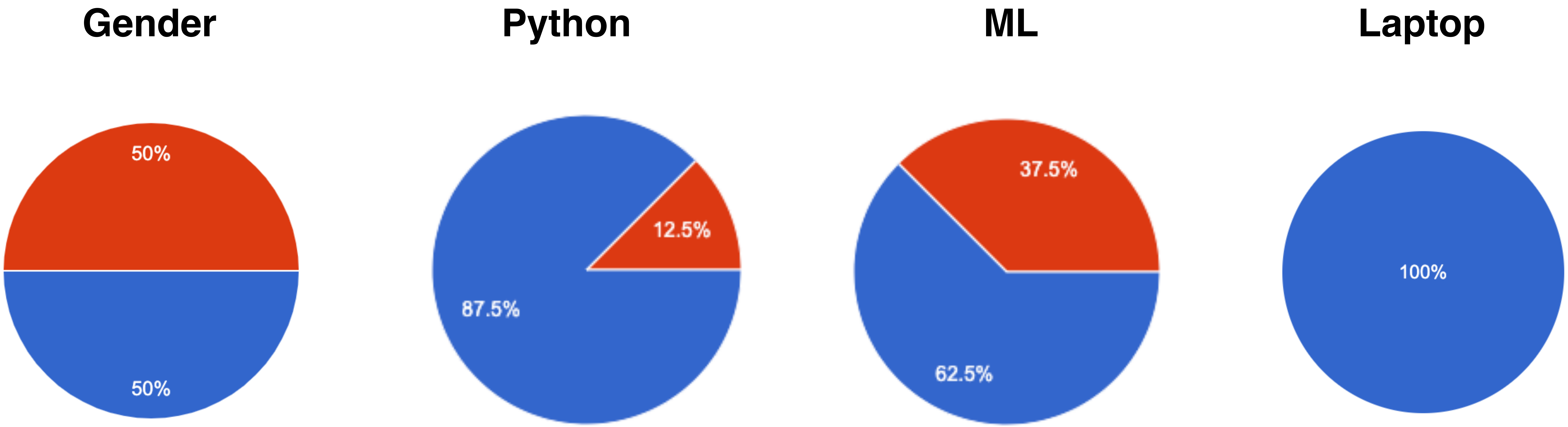
- Office hour is right after the class on Wednesdays 11:30-12:30
  dfdJoin Zoom Meeting: https://hecmontreal.zoom.us/j/81836582494?pwd=VVhvWk1rYVFLdGJzTldLZzYyc0VvQT09
  Meeting ID: 818 3658 2494
  Passcode: 379543

- Office hour (Pravish) will be on Fridays. He will announce the details on Piazza

- Student Introduction suvery, due January 26, 2022.

- Team Registration, due: January 26, 2022.

Class statistics


Gender

Python

ML

Laptop

Student Introduction Survey form **due tonight**

Today

- **First 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.

- Summary of Machine learning fundamental

- Q&A

- Hands-on session

Quiz 0

Login to your Gradescope account

Machine Learning Problem

The three components of an ML problem:

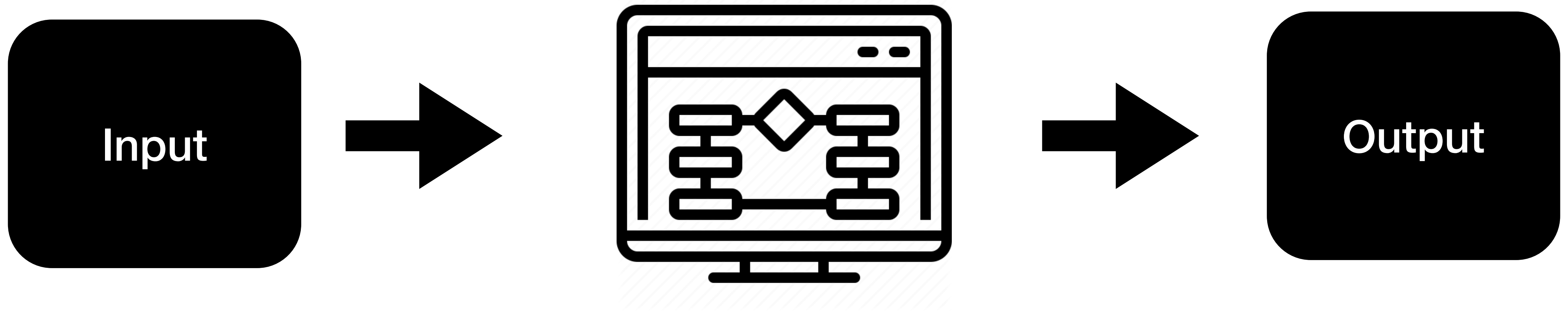
1. **Task.** What is the problem at hand?
   - Model. How are you parametrizing your solution.
2. **Performance.** How well you are doing?
3. **Experience.** What kind of data do you have access to?

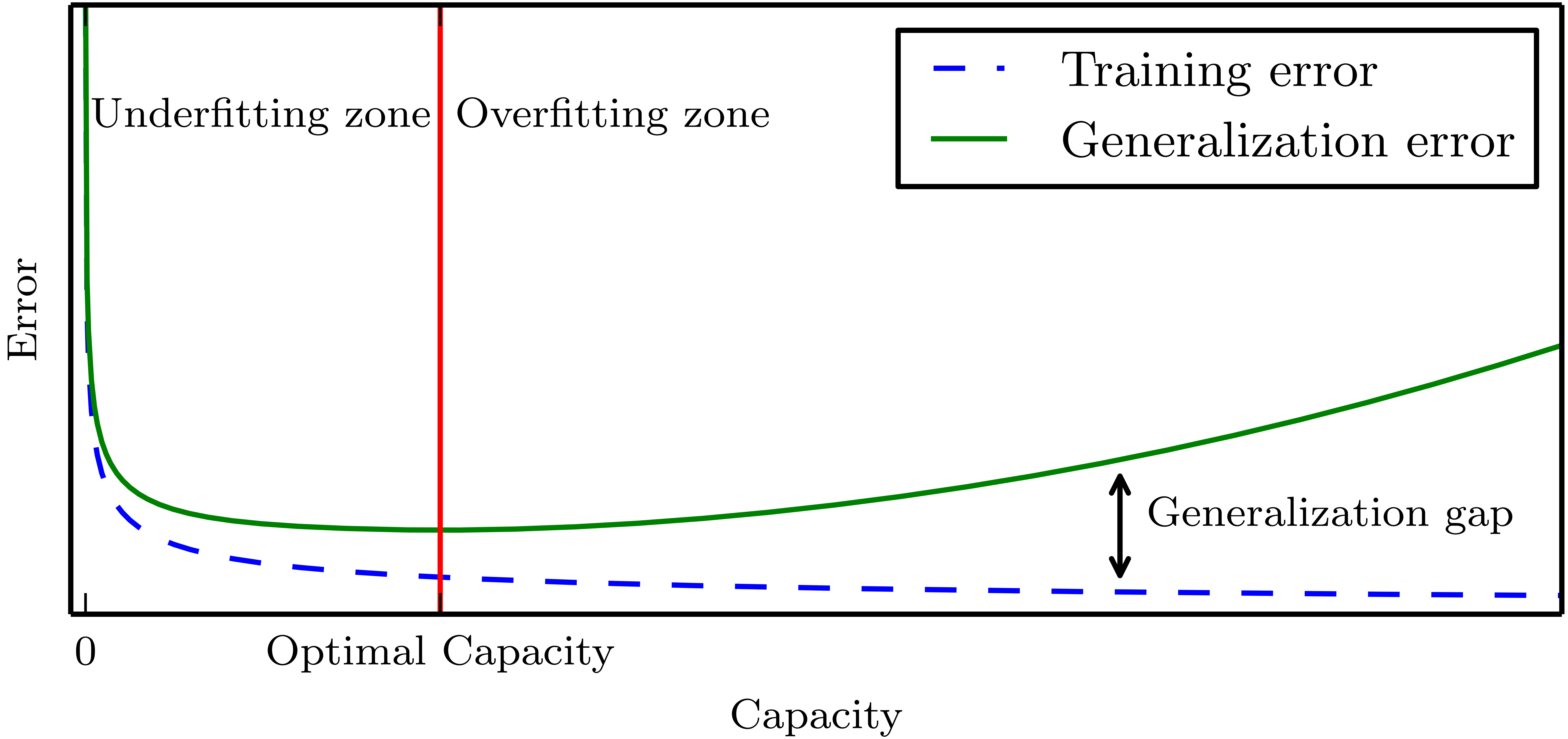
Types of Experiences

- **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.

Model Evaluation

- **Given:**
  - A performance measure
  - A train dataset
  - A model
- **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



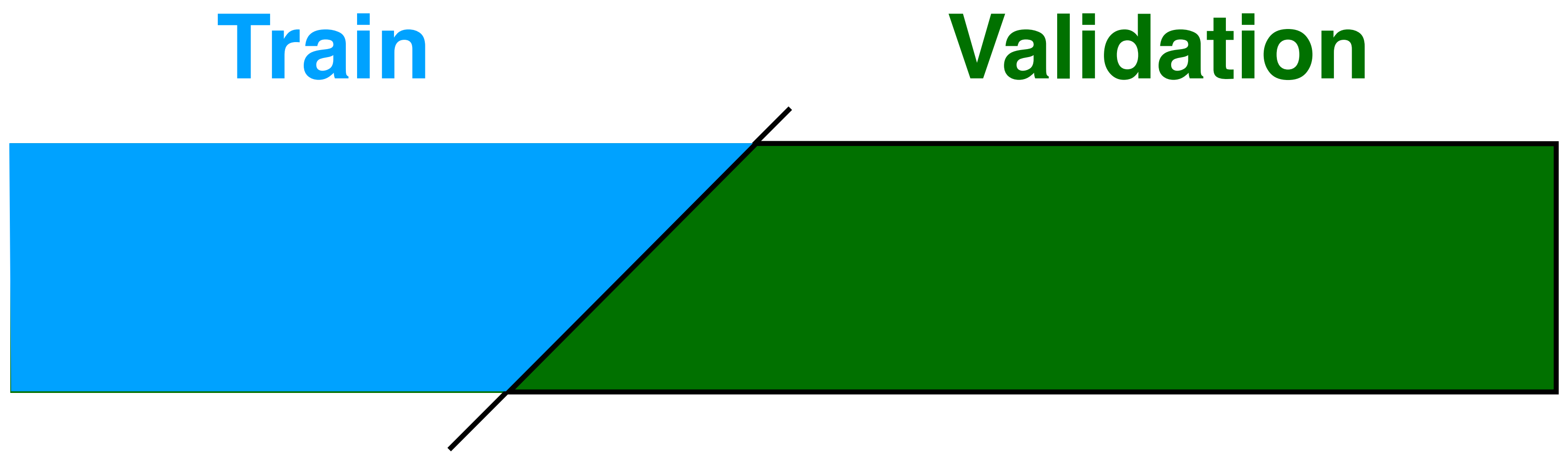

Underfitting zone
Overfitting zone
Training error
Generalization error
Generalization gap
Error
0
Optimal Capacity
Capacity

Regularization

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

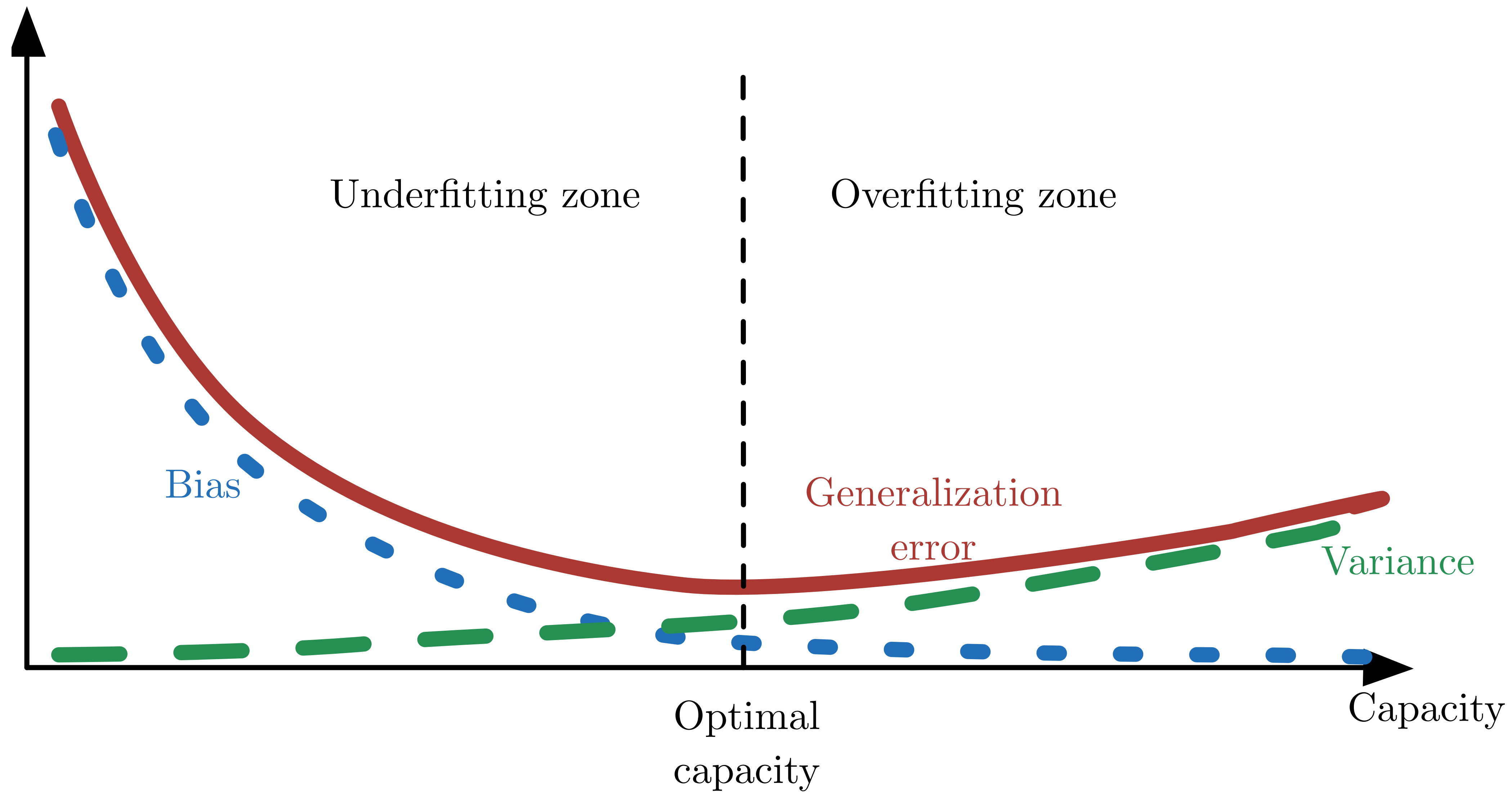
Validation set

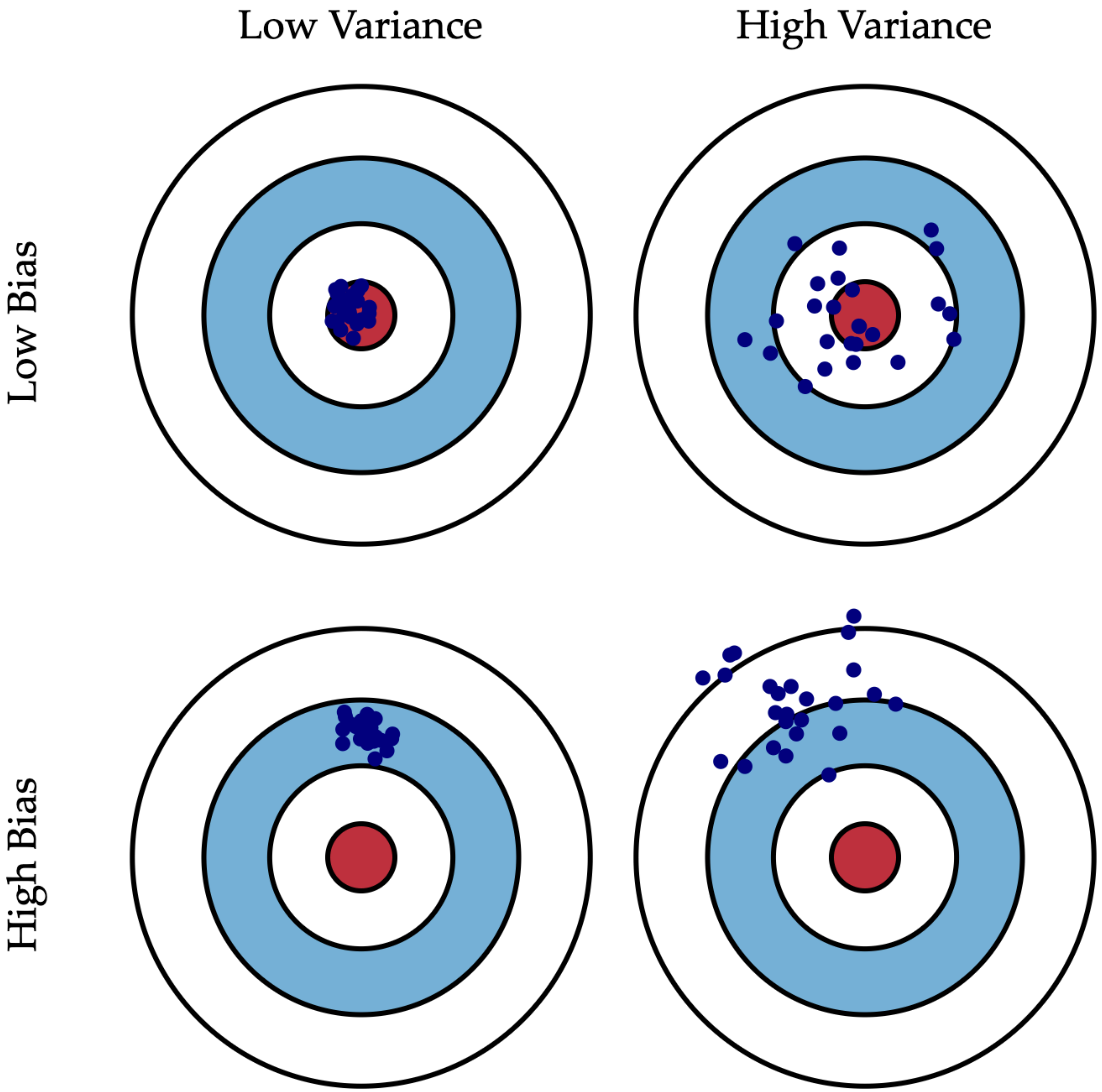
- How do we choose the right model and set its hyper parameters (e.g. λ)?
  - **Use a validation set**
    - **Split the original data into two:**
      1. Train set
      2. Validation set
         - Proxy to the test set
  - **Train different models/hyper-parameter settings on the train set**
  - **Pick 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


Underfitting zone
Overfitting zone
Bias
Generalization error
Variance
Optimal capacity
Capacity


Low Variance
High Variance
Low Bias
High Bias