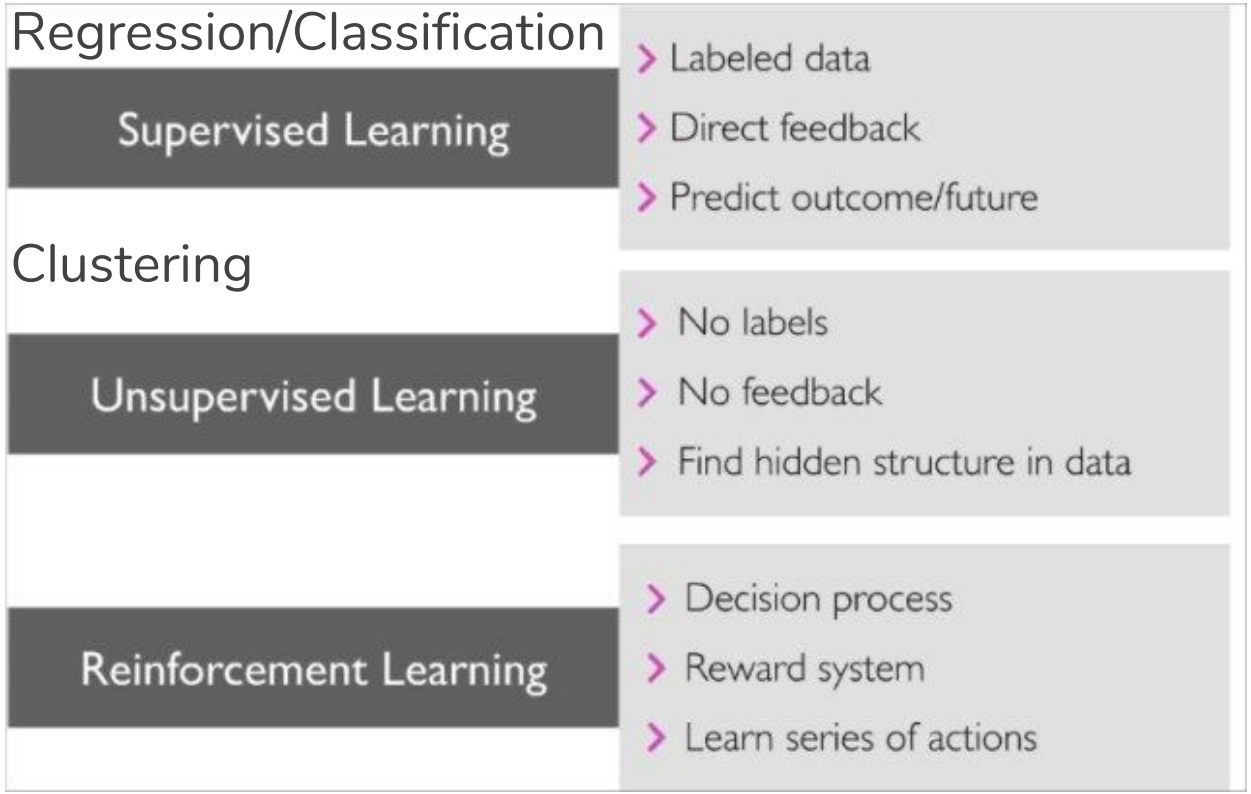
The Machine Learning and Data Science Pipeline

Agenda

1. 3 Different Classes of Machine Learning Problems
2. How are ML models trained, validated, and tested?
3. The 7 Main Steps in the Data Mining and Machine Learning Pipeline (2 DEMOs)

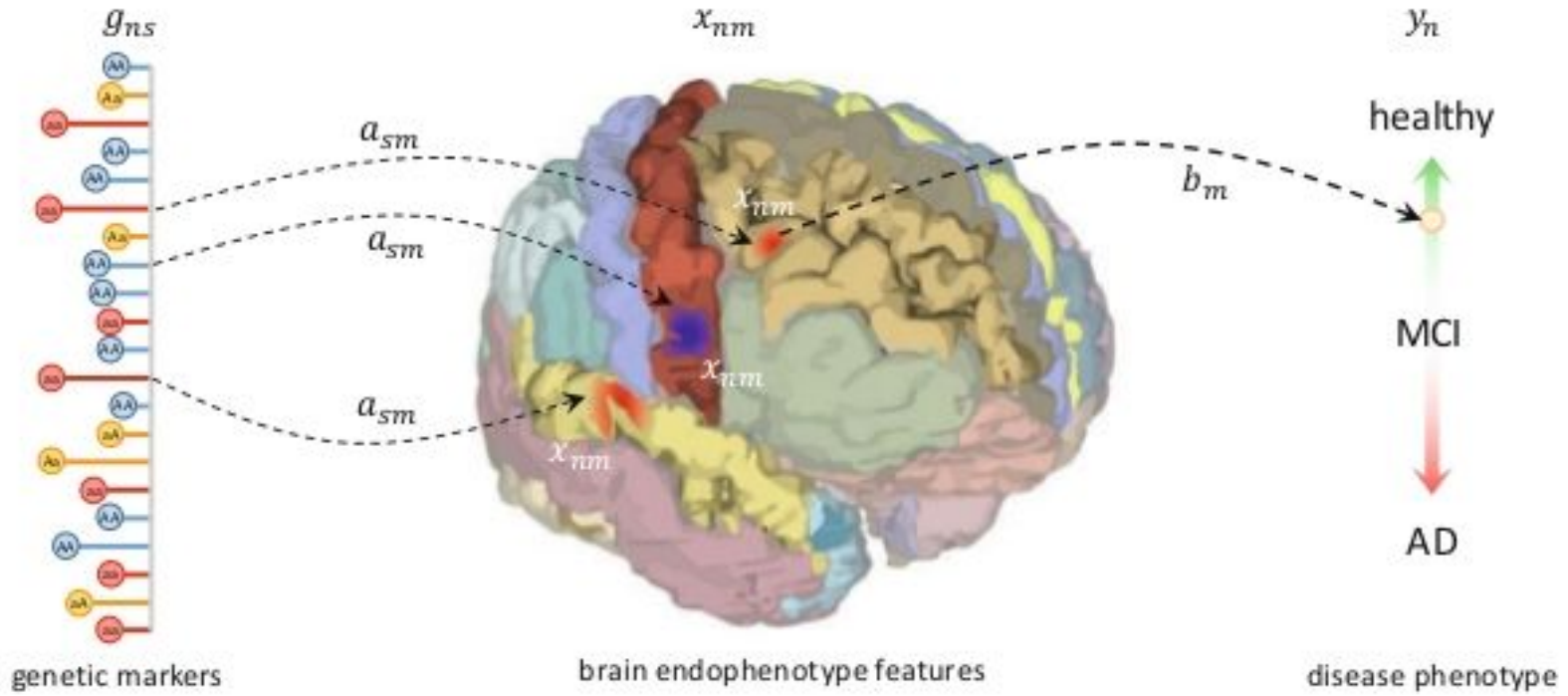
3 Different Classes of Machine Learning Problems

The 3 General Classes of ML Models

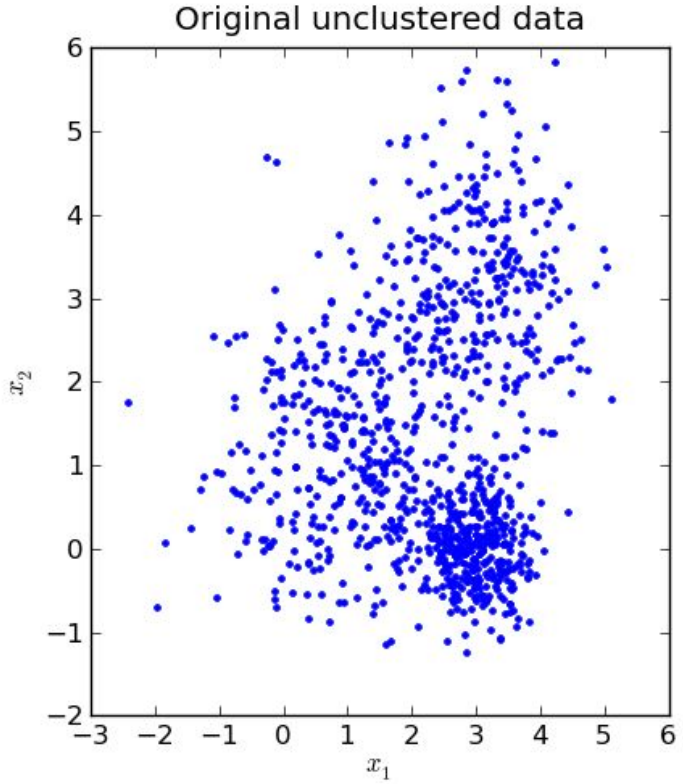


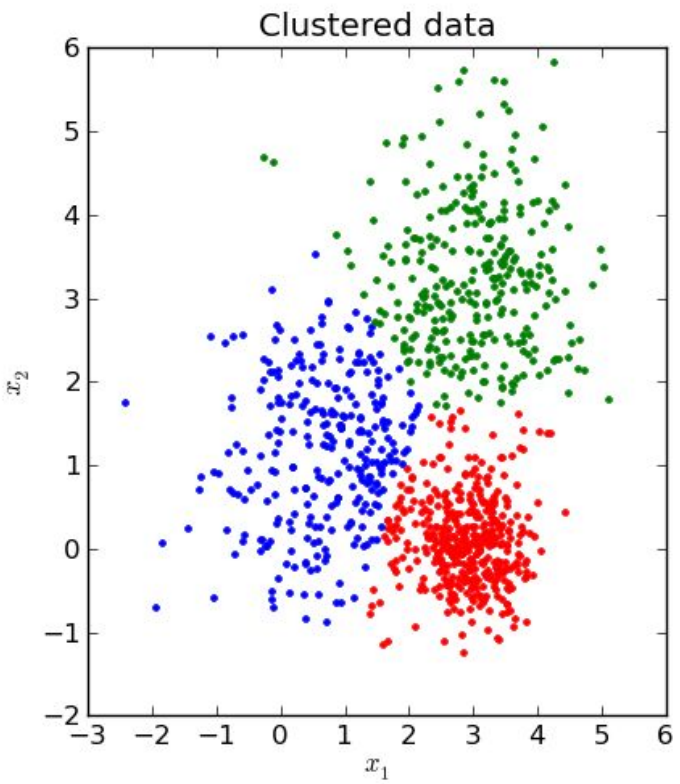
Supervised Learning (Classification)
Ex: ImageNet (10M+ images), 1000 classes

Supervised Learning (Regression)
Ex: Imaging Genetics and Genome-Wide Association Studies

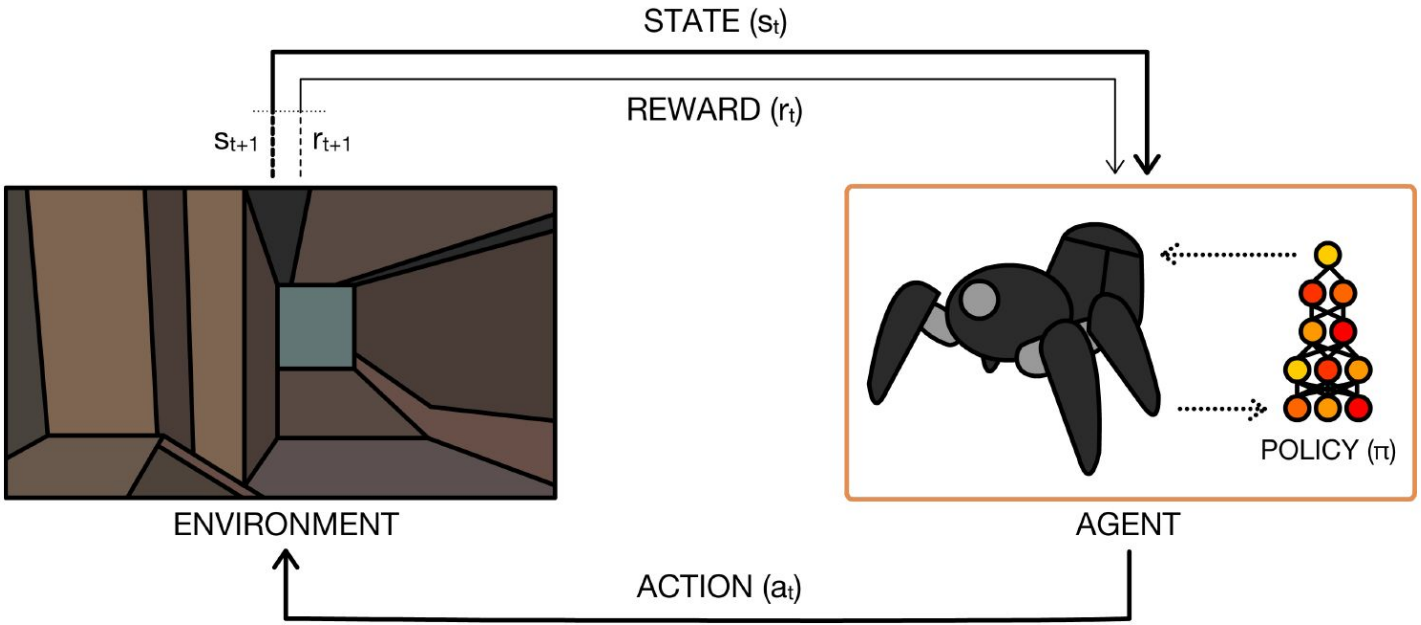


Unsupervised Learning (Clustering)





Reinforcement Learning (Training autonomous agents to behave optimally in complex environments)



The Initial Development, Training, Validation, and Testing of Machine Learning Models

Typical ML Pipeline Goes Like This:

1. Acquire the data
2. Prepare & Visualize Data
3. Choose a Model
4. Train a Model on the Training Set
5. Evaluate Model Performance
6. Tune Model Hyperparameters
7. Prediction!

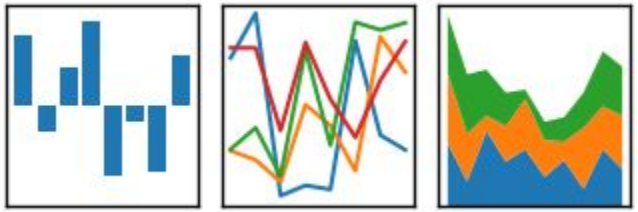
STEP ONE: Acquire the Data

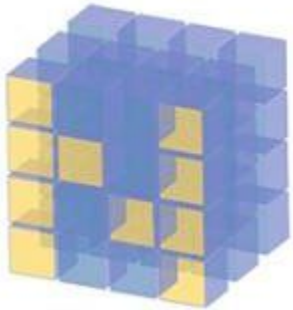
BERKELEY LAB

STEP TWO: Prepare and Visualize Data

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$





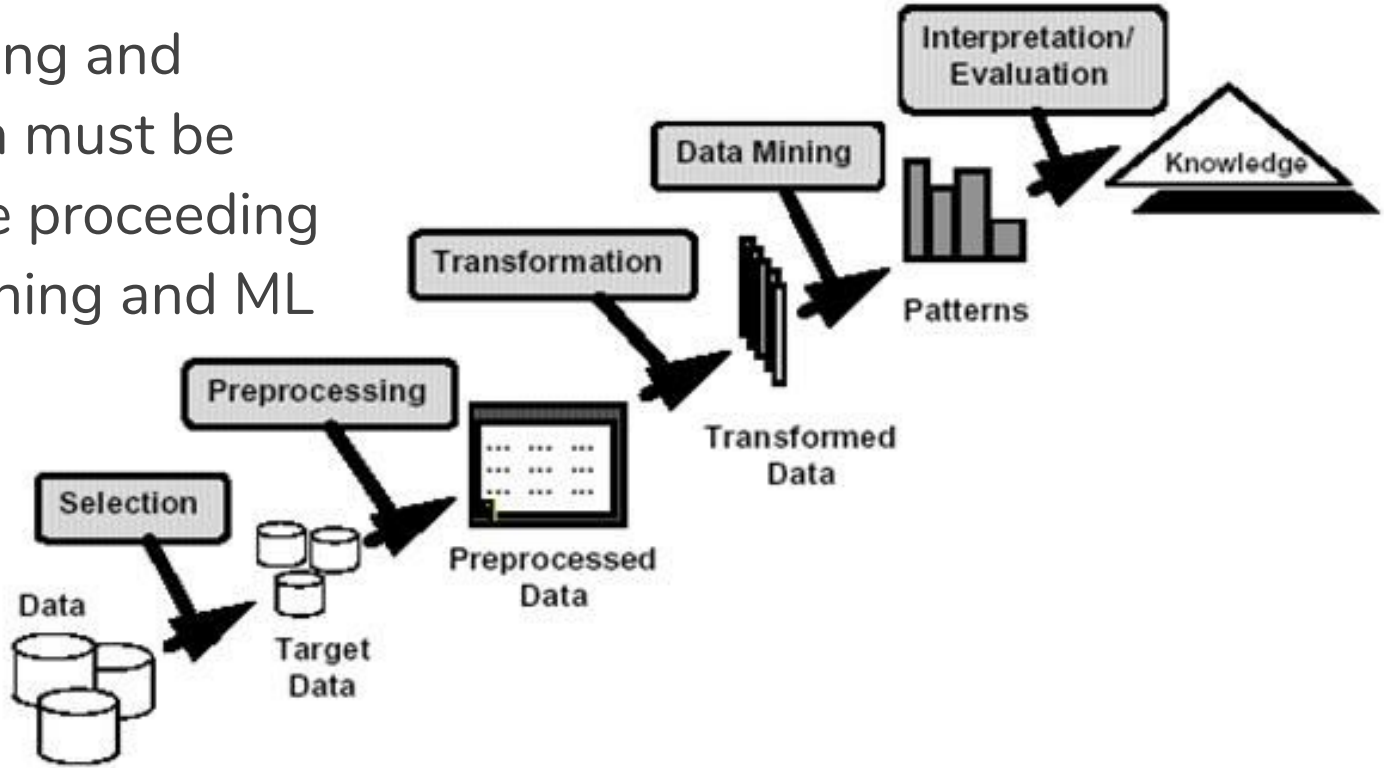
NumPy

scikits learn

machine learning in Python

Importance of Feature Selection and Data Preprocessing in ML problems

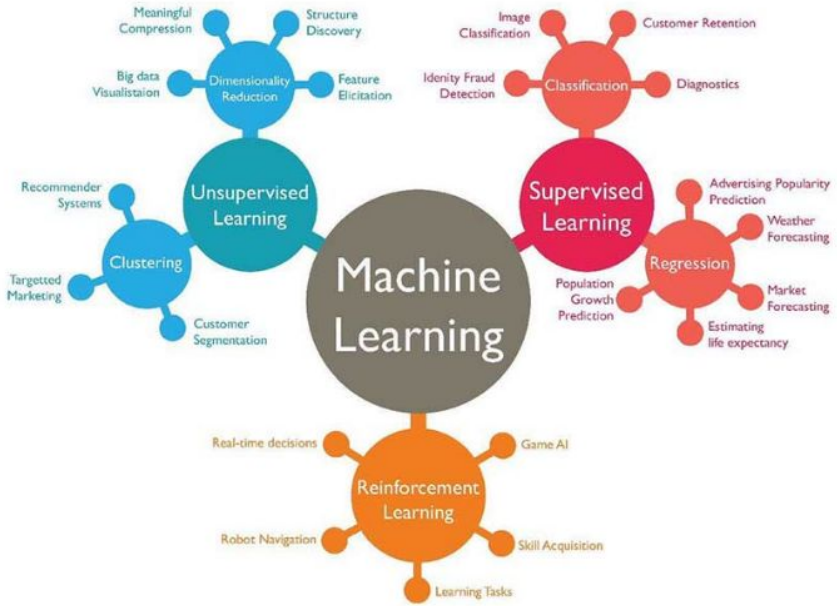
Data Preprocessing and Feature Selection must be performed before proceeding to actual data mining and ML analyses.



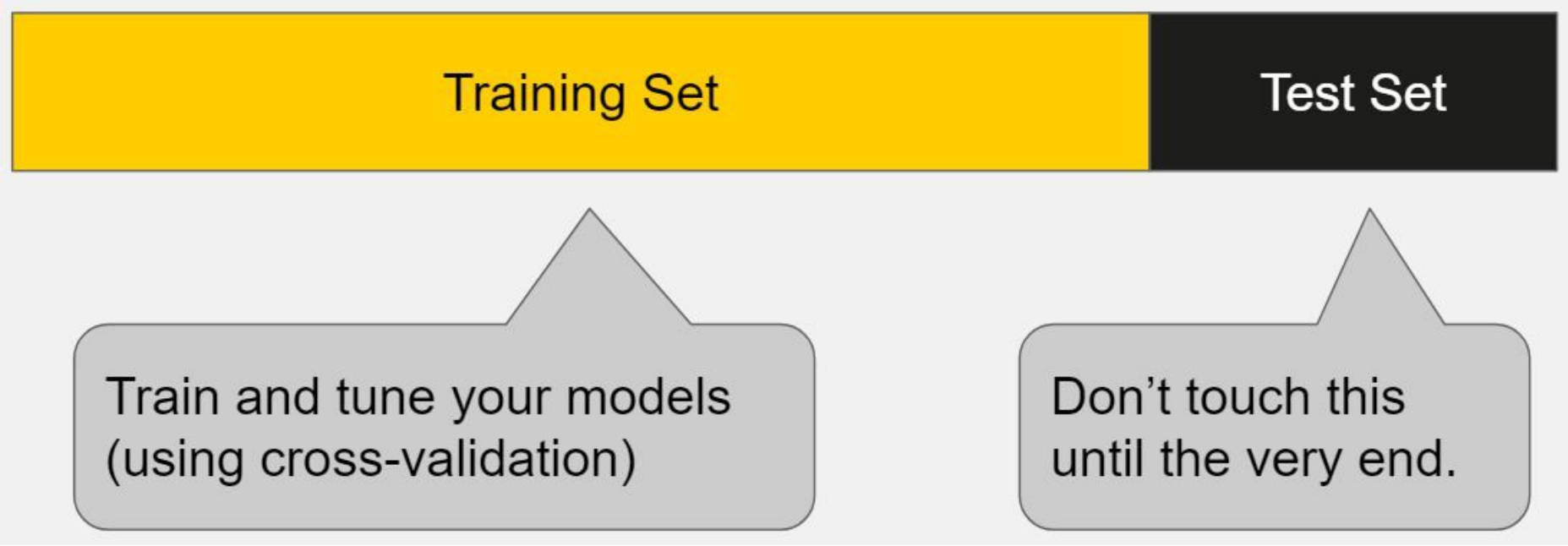
Live Demo: Understanding, Cleaning, and Visualizing Climate Change Data from Lawrence Berkeley National Laboratory

STEPs THREE and FOUR: Choose A Model (3) and Train It! (4)

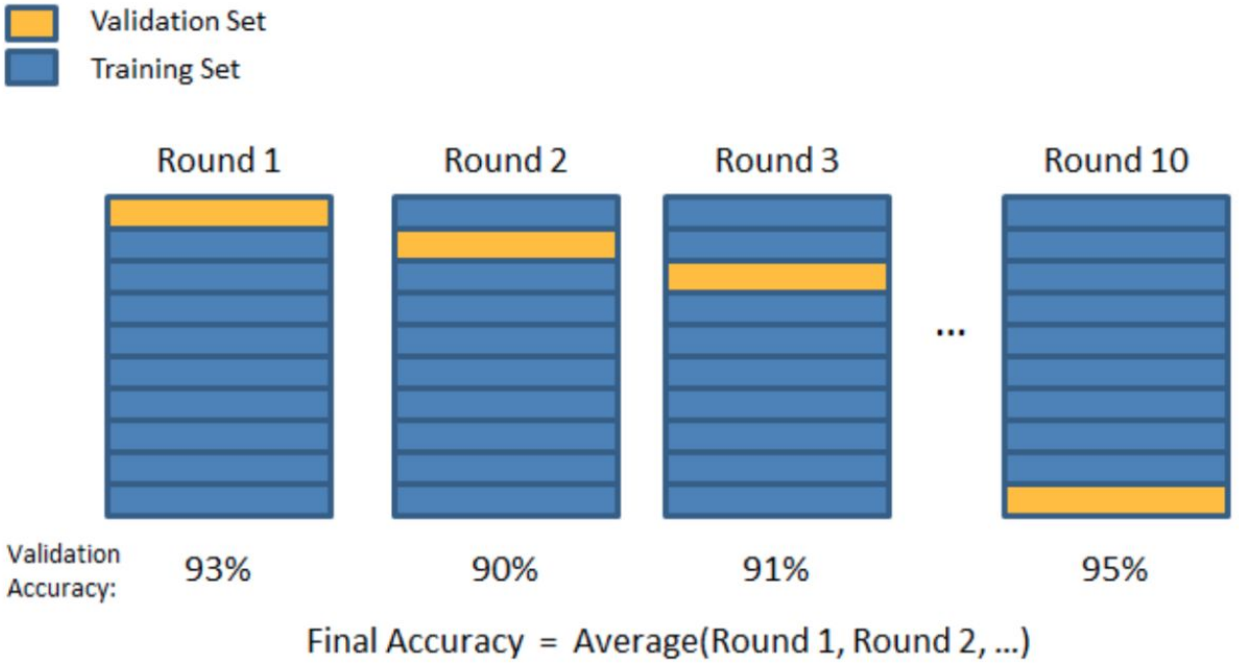
We’re going to spend the whole semester covering best practices for choosing the best ML models.



Dealing with Datasets in Machine Learning



STEP FIVE: Evaluate the Model's Performance in the Validation Step (K-fold Cross-Validation)



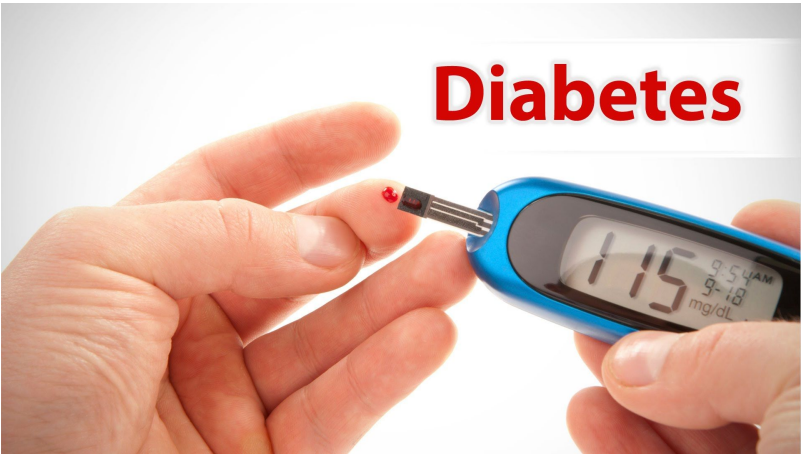
STEP SIX: Tune Model Hyperparameters Via K-Fold Cross-Validation

What happens if our cross-validation accuracy is rather low, and we would like to improve it?

Choose different **Hyperparameters** to our model!

While the ML model finds optimal values of the parameters to minimize some type of loss function, the ML practitioner gets to choose values for the hyperparameters.

The Diabetes Dataset: How to Tune a Model's Hyperparameters to achieve Best Possible Performance using K-fold Cross Validation



Let's use Age, Gender, BMI, BP, and a couple of Blood Serum Measurements to predict a quantitative metric of diabetes progression in one year using a form of regularized regression called the LASSO.

One Hyperparameter (ƛ) that we choose

LASSO (Least Absolute Shrinkage and Selection Operator) Regularization for Linear Regression Models

$$Cost\ (W) = RSS(W) + \ \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i \ - \sum_{j=0}^{M} w_j \ x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

Live Demo: Using 20-fold Cross Validation with a LASSO Regularized Regression to select optimal hyperparameter for Diabetes Prediction Problem

STEP SEVEN: Characterize Your Model's Performance on a Test Dataset and Predict using Your Model!