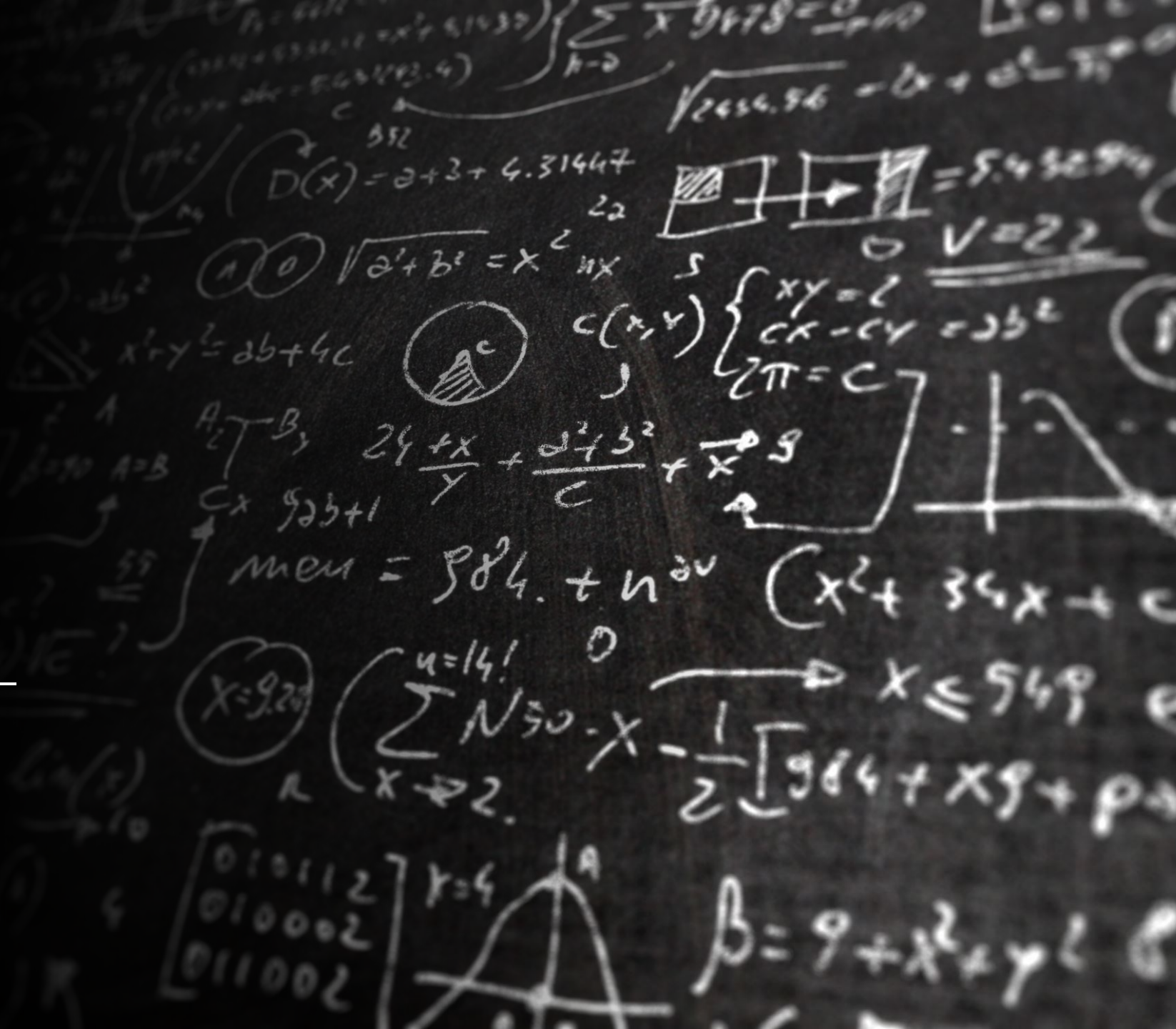
Mentor Session scikit-learn

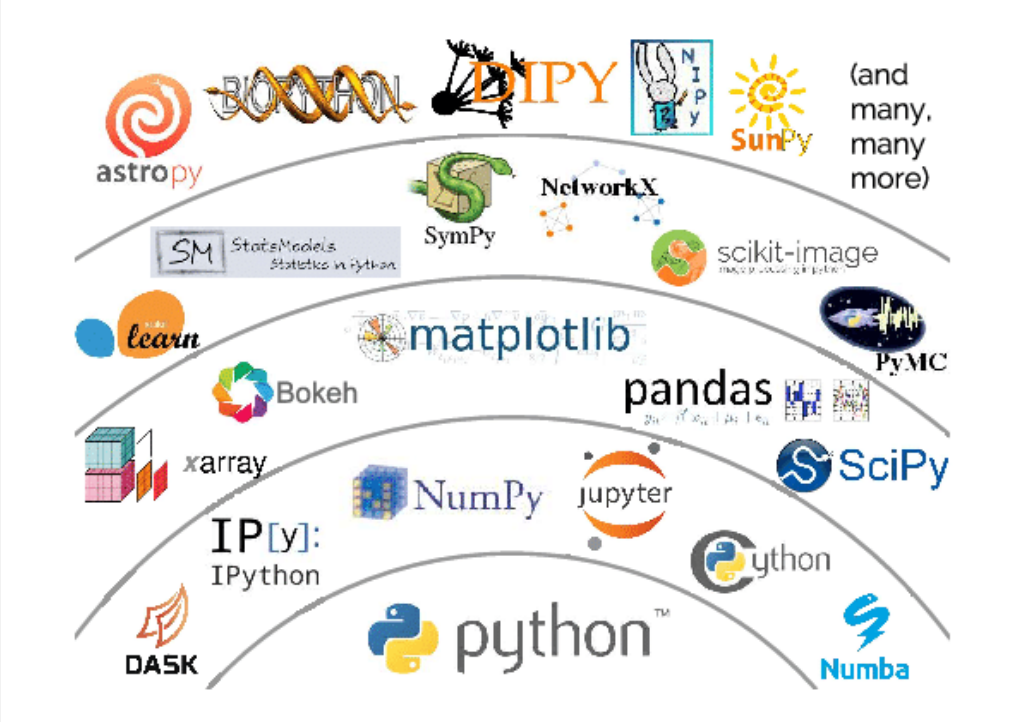
Machine Learning & Data Science II

Yingjie(Chelsea) Wang



Python Ecosystem for Machine Learning

- SciPy is an ecosystem of Python libraries for mathematics, science and engineering. It is an add-on to Python that you will need for machine learning. The SciPy ecosystem is comprised of the following core modules relevant to machine learning:
  - **NumPy**: A foundation for SciPy that allows you to efficiently work with data in arrays.
  - **Matplotlib**: Allows you to create 2D charts and plots from data.
  - **Pandas**: Tools and data structures to organize and analyze your data.
- The **scikit-learn** library is how you can develop and practice machine learning in Python. It is built upon and requires the SciPy ecosystem. The name scikit suggests that it is a SciPy plug-in or toolkit.
  - The focus of the library is machine learning algorithms for classification, regression, clustering and more.
  - It also provides tools for related tasks such as pre-processing data, modelling, evaluating models and tuning parameters.

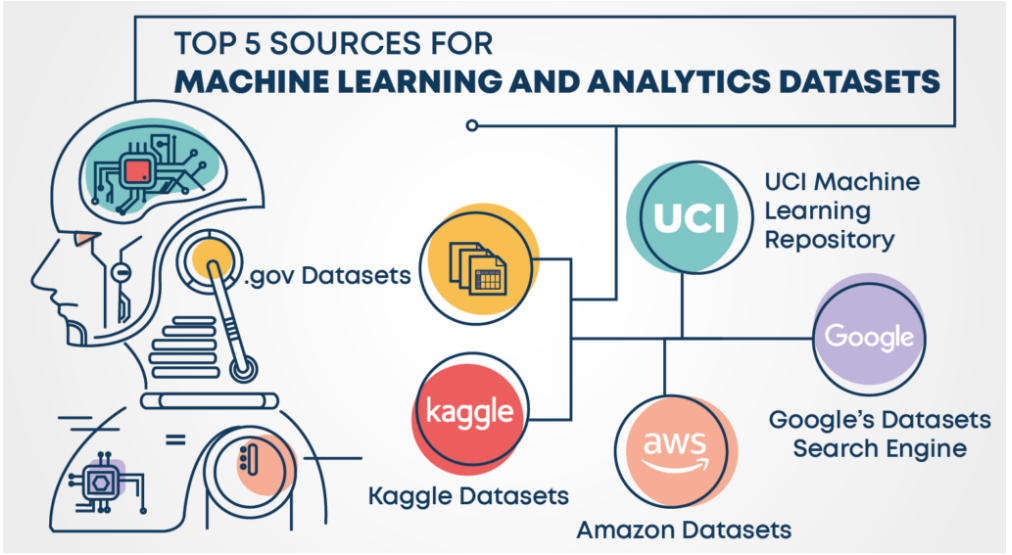


Data Science Workflow

- Data Preparation
  - Data Collection Sources
  - Importing and merging DataFrames in Pandas
- Exploratory Data Analysis (EDA)
  - Descriptive Statistics
  - Creating correlation heatmaps
  - Principal components analysis (PCA) with scikit-learn
  - Feature Engineering
- Modelling
  - Doing ML classification prediction with scikit-learn
  - Doing ML Regression prediction with scikit-learn
- Modelling Evaluation with scikit-learn

Data Preparation

- Data Collection
  - Download Existing Data (Kaggle, Google Cloud Public Datasets, etc.)
  - Data Scraping (Beautiful Soup, urllib, etc.)
  - API (application programming interface)
  - Create your own data (survey, observations, etc.)
- Data Preprocessing
  - Data Cleaning (missing rate, outliers, data format, etc.)
  - Read in Pandas DataFrame (https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html)

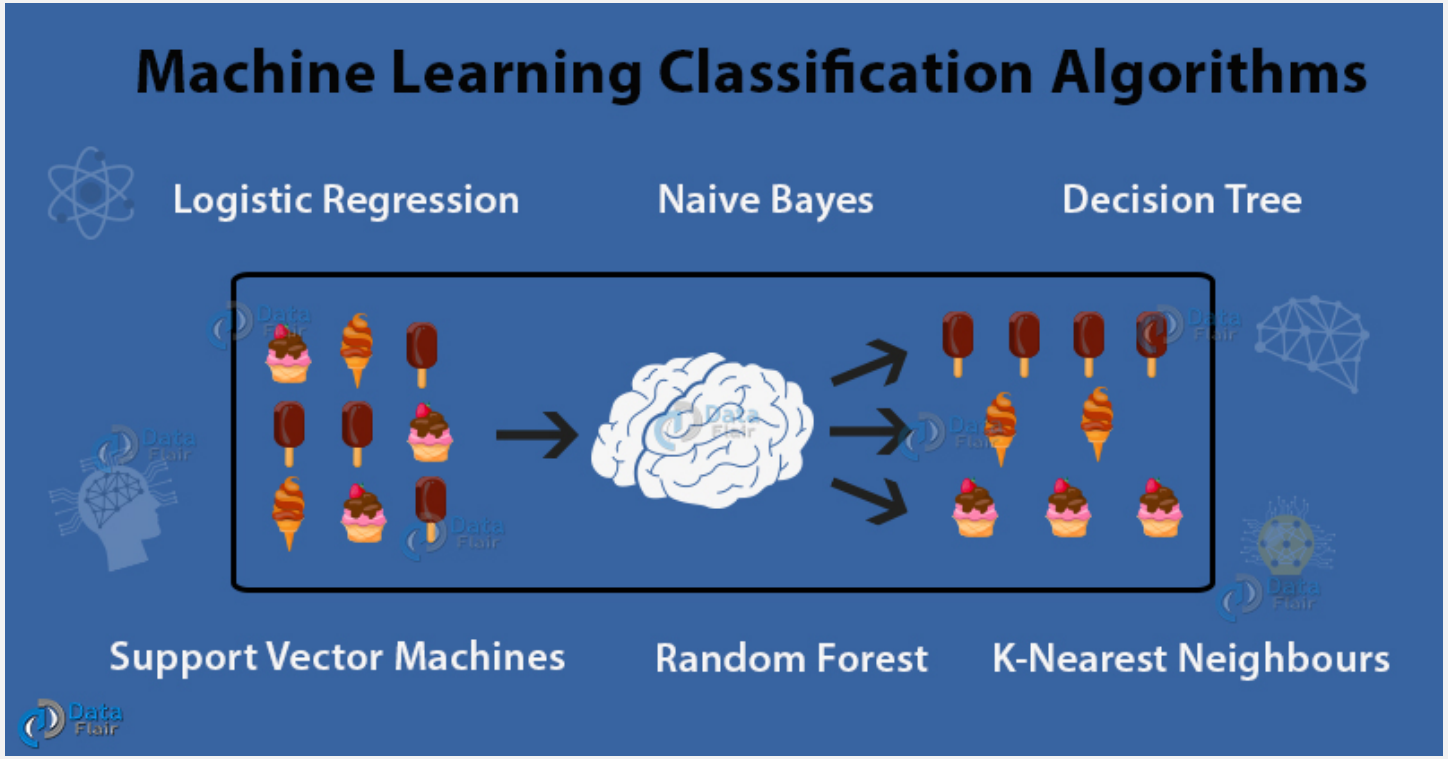


Exploratory Data Analysis

- Descriptive Statistics
  - Dimensions of Your Data, Data Types, Class Distribution, Data Summary, Correlations, Skewness, Outliers, etc.
- Correlation heatmaps
  - Correlation matrix to show which variable is having a high or low correlation in respect to another variable.
- Principal components analysis (PCA) with scikit-learn
  - Unsupervised method for dimensionality reduction
  - Resulting in a lower-dimensional projection of the data that preserves the maximal data variance.
  - Usage: Data Visualization and Speeding ML algorithm
- Feature Engineering (optional)

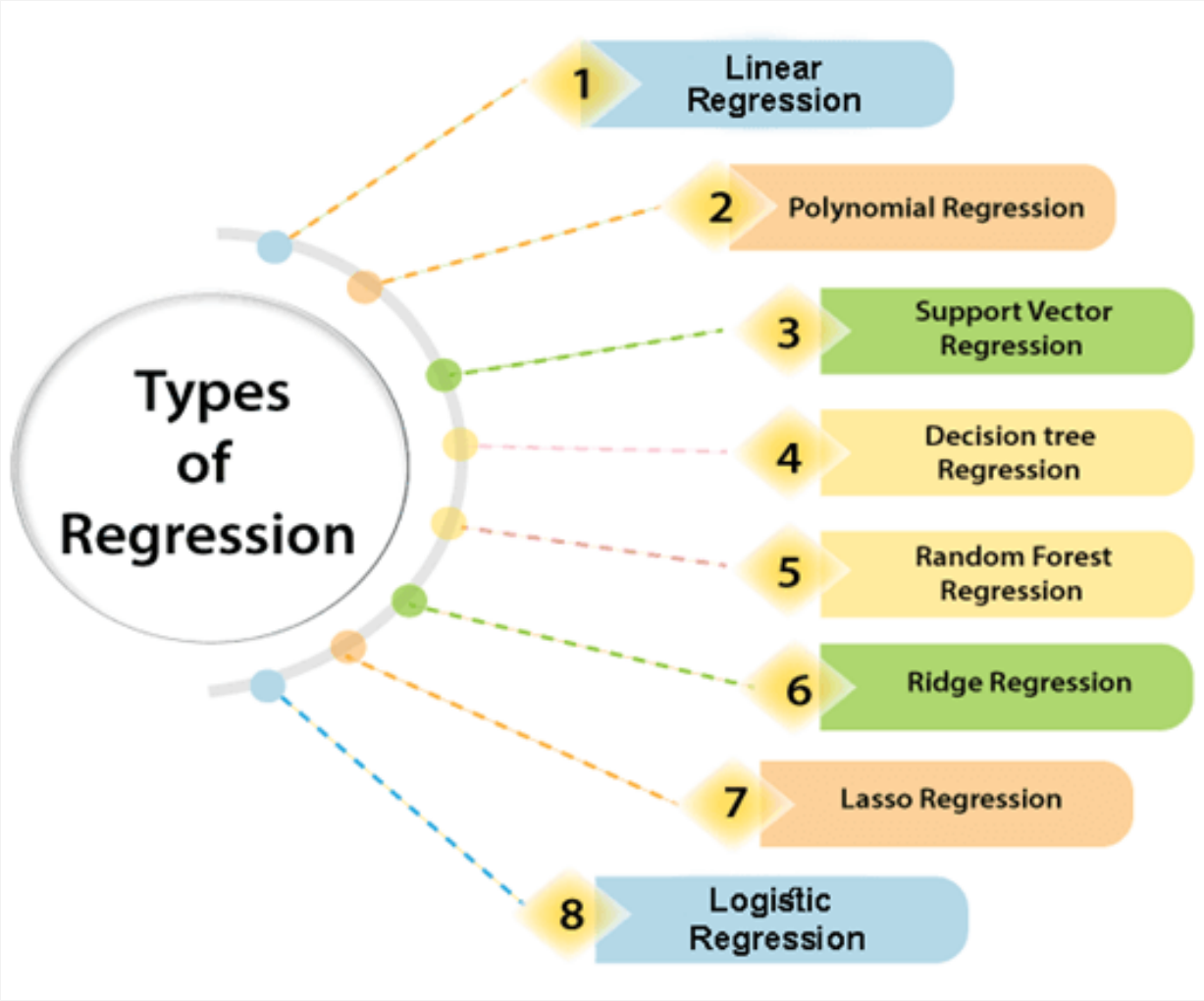
Classification Algorithms

- Linear machine learning algorithms:
  - Logistic Regression
- Nonlinear machine learning algorithms:
  - Decision tree
  - Random forest
  - Naive Bayes
  - Support Vector Machines
  - k-Nearest neighbors



Regression Algorithms

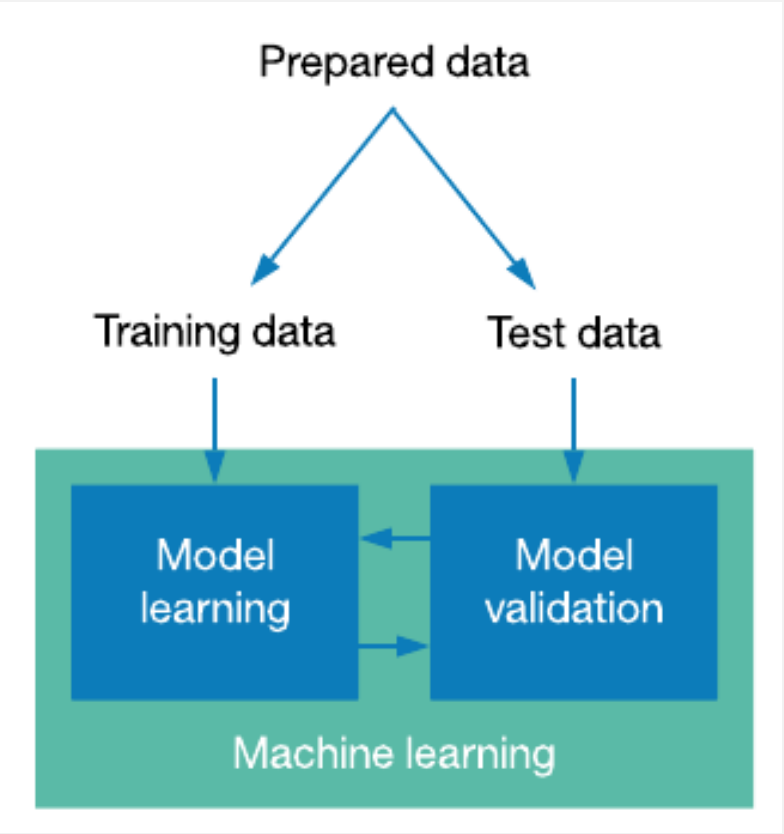
- Linear machine learning algorithms:
  - Linear Regression
  - Ridge Regression (L2-norm)
  - LASSO Linear Regression (L1-norm)
  - Elastic Net Regression (ridge + lasso)
- Nonlinear machine learning algorithms:
  - Decision tree Regressor
  - Bagging Regressor
  - Random forest Regressor
  - Support Vector Regressor
  - k-Nearest neighbors Regressor



Evaluate the Performance of ML Algorithms with Resampling

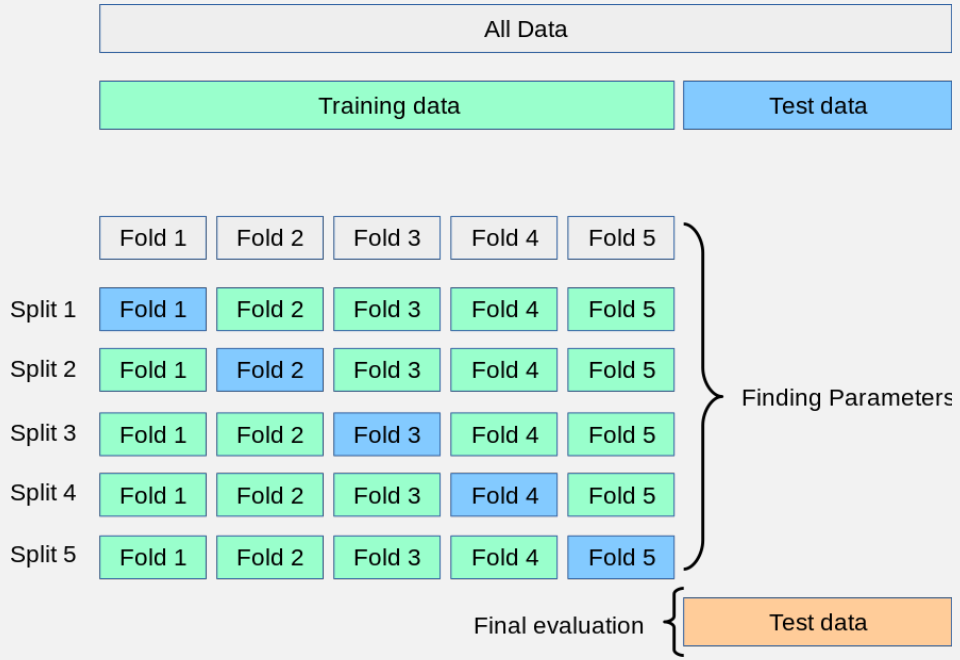
Four different techniques to split up our training dataset and create useful estimates of performance for our machine learning algorithms:

- Train and Test Sets (Most Common)
  - Split data into 2 parts (e.g. 70% vs 30%)
  - This algorithm evaluation technique is **very fast**. It is ideal for large datasets (millions of records) where there is strong evidence that both splits of the data are representative of the underlying problem.
  - A downside of this technique is that it can have a high variance. This means that differences in the training and test dataset can result in meaningful differences in the estimate of accuracy.
- k-fold Cross Validation
- Leave One Out Cross Validation
- Repeated Random Test-Train Splits



Evaluate the Performance of ML Algorithms with Resampling

- k-fold Cross Validation
  - Split the dataset into k-parts. Each split of the data is called a fold.
  - The algorithm is trained on k – 1 folds with one held back and tested on the held back fold. This is repeated so that each fold of the dataset is given a chance to be the held back test set. After running cross validation you end up with k different performance scores that you can summarize using a mean and a standard deviation.
- Leave One Out Cross Validation
  - Configure cross validation so that the size of the fold is 1 (k is set to the number of observations in the dataset).
  - Don’t Use LOOCV: Large datasets or costly models to fit.
  - Use LOOCV: Small datasets or when estimated model performance is critical.
- Repeated Random Test-Train Splits
  - Create a random split of the data like the train/test split but repeat the process of splitting and evaluation of the algorithm multiple times, like cross validation.
  - A downside is that repetitions may include much of the same data in the train or the test split from run to run, introducing redundancy into the evaluation.



Machine Learning Algorithm Performance Metrics

Supervised Learning

Classification Metrics

- Classification Accuracy
  - Ratio of number of correct predictions to the total number of input samples.
- Logarithmic Loss
  - Works by penalizing the false classifications
- Confusion Matrix
  - Matrix as output and describes the complete performance of the model. (TP, TN, FP, FN)
- Area under Curve
  - x-axis: False Positive Rate; y-axis: True Positive Rate
- F1 Score
  - Mean between precision and recall.

Regression

- Mean Absolute Error
- Mean Squared Error
- R-squared

Unsupervised learning

Clustering

- Adjusted Rand Index
  - The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.
  - The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical (up to a permutation).
- Silhouette Score
  - The Silhouette Score attempts to describe how similar a datapoint is to other datapoints in its cluster, relative to datapoints not in its cluster (this is aggregated over all datapoints to get the score for an overall clustering).
  - It is bounded between -1 and 1. Closer to -1 suggests incorrect clustering, while closer to +1 shows that each cluster is very dense.

NLP Metrics

- Text Generation: Perplexity
- Translation: BLEU

Automate Machine Learning Workflows with Pipelines

- Python scikit-learn provides a Pipeline utility to help automate machine learning workflows.
- Pipelines work by allowing for a linear sequence of data transforms to be chained together culminating in a modeling process that can be evaluated.
  - Data Preparation and Modeling Pipeline
    - Standardize the data
  - Feature Extraction and Modeling Pipeline
    - Feature Extraction with Principal Component Analysis
    - Feature Extraction with Statistical Selection
    - Feature Union
- Last Step: hyperparameter optimization (Tuning)
  - Grid Search Parameter Tuning
  - Random Search Parameter Tuning