Text Classification

(Natural Language Processing)

HND Thilini

hnd@ucsc.cmb.ac.lk

Machine Learning Concepts

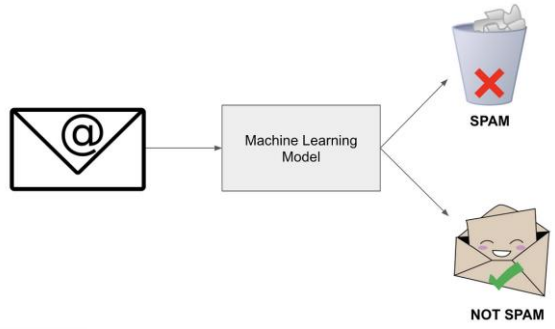
- Data preparation: Usually consists of pre-processing the data before extracting features and training
- Feature extraction: The process of extracting useful features from raw data that are used to train machine learning models
- Features: Various useful attributes of the data (examples could be age, weight, and so on for personal data)
- Training data: A set of data points used to train a model
- Testing/validation data: A set of data points on which a pre-trained model is tested and evaluated to see how well it performs
- Model: Built using a combination of data/features and a machine learning algorithm that could be supervised or unsupervised
- Accuracy: How well the model predicts something (also has other detailed evaluation metrics like precision, recall, and F1-score)

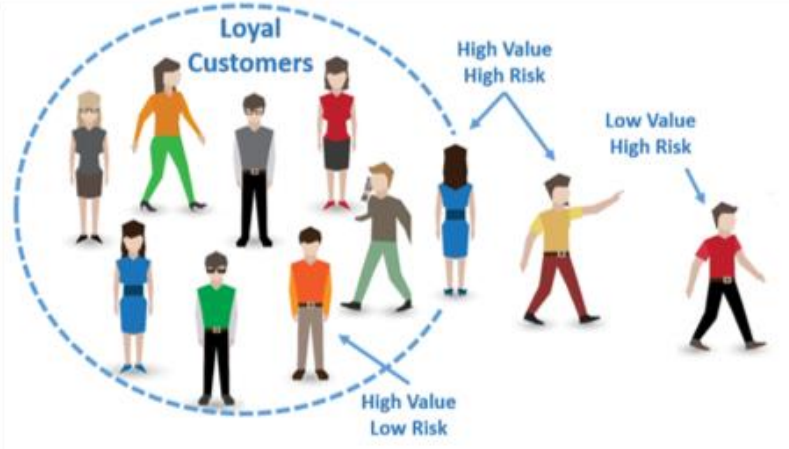
Classification: Definition

- Given a collection of examples (*training set*)
  - Each example is represented by a set of *features*, sometimes called *attributes*
  - Each example is to be given a label or class
- Find a *model* for the label as a function of the values of features.

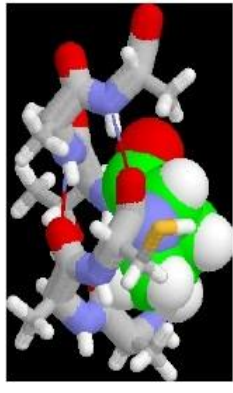
- Goal: previously unseen examples should be assigned a label as accurately as possible.
  - A *test set* is used to determine the accuracy of the model.
  - Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Examples of classification tasks





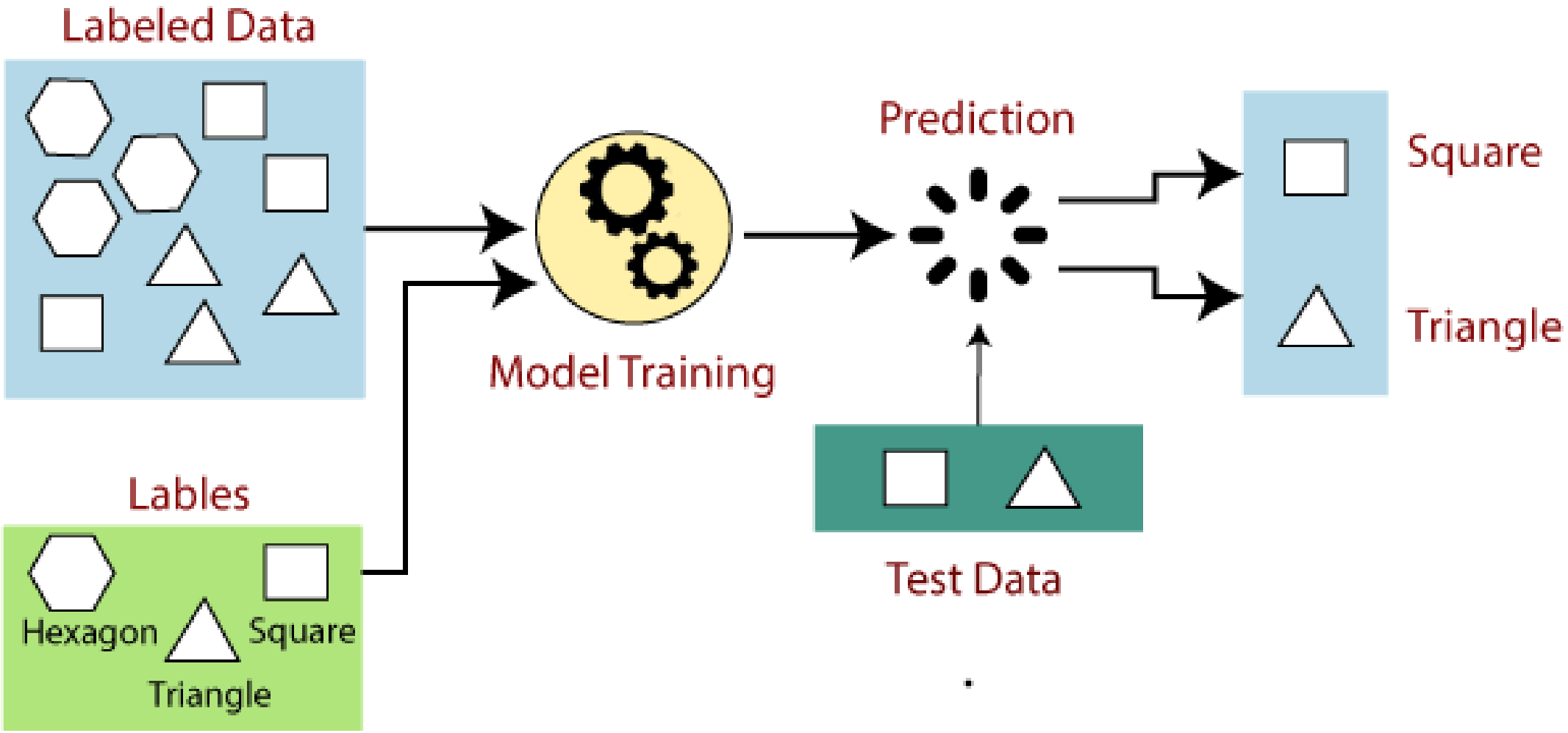
- Document Classification
- Spam Filtering
- Customer behavior/churn prediction
- Image classification
- Anomaly detection / Fraud detection
- Sentiment Analysis
- Health: predicting benign or malignant tumor cells
- …



ML approaches in Classification

- Supervised learning (**classification)**
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (includes **clustering)**
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Illustrating Classification Task using Supervised Learning



Different types of Classification Tasks in ML

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

Binary Classification

- The number of distinct categories is 2
  - Prediction of conversion (buy or not), Churn forecast (churn or not), Detection of spam email (spam or not), etc..
- One representing the normal state and the other representing the aberrant state
  - normal condition is "not spam", while the abnormal state is "spam"
  - normal condition of "cancer not identified" and an abnormal state of "cancer detected“
- Logistic Regression, Support Vector Machines, Simple Bayes, Decision Trees, etc..

Multi-Class Classification

- More than 2 categories
  - Categorization of faces, Classifying plant species, Character recognition using optical, etc..
- Instances are grouped into one of several well-known classes.

- Progressive Boosting, Choice trees, Nearest K Neighbors, Rough Forest, Simple Bayes, etc…

- Can be solved using algorithms created for binary classification
  - One-vs-One: For each pair of classes, fit a single binary classification model
  - One-vs-Rest: Fit a single binary classification model for each class versus all other classes

Multi-Label Classification

- Allow for the prediction of one or more class labels for each example
  - Photo classification - A particular photo may have multiple objects in the scene
- Solved by training a number of binary classifiers and combining them to get a multi-label result
- Multi-label Gradient Boosting, Multi-label Random Forests, Multi-label Decision Trees, etc

Imbalanced Classification

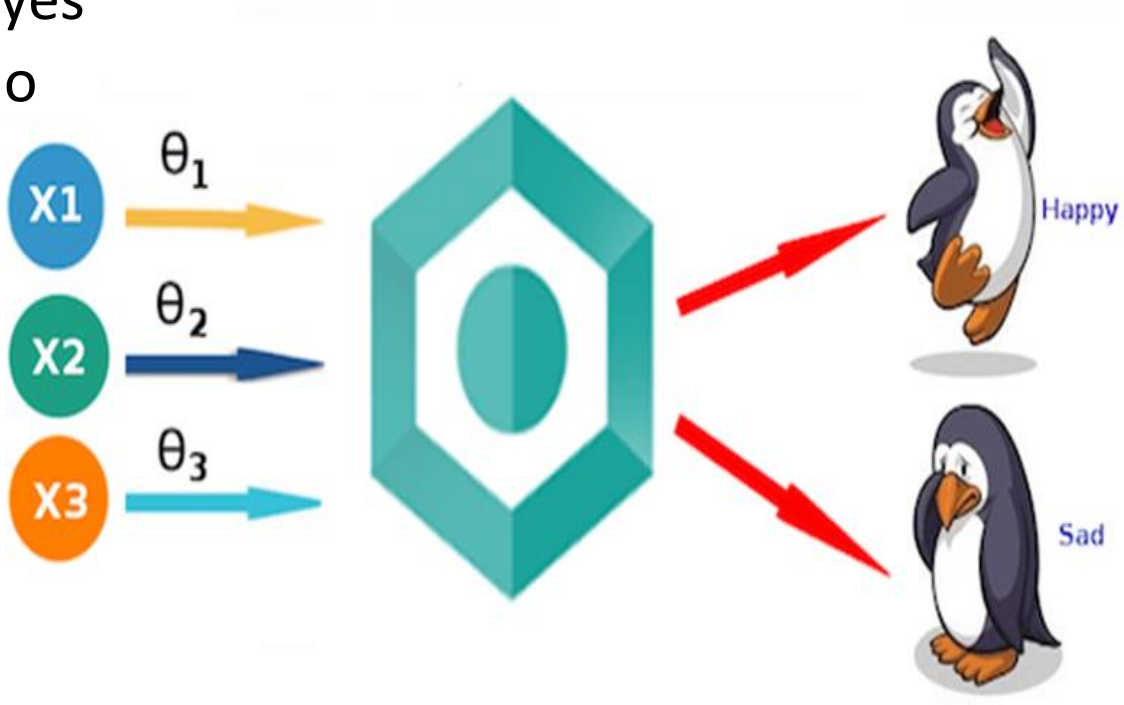
- The distribution of examples within each class is not equal.
  - A majority of the training dataset's instances belong to the normal class, while a minority belong to the abnormal class
- Clinical diagnostic procedures, Detection of outliers, Fraud investigation, etc…
- By oversampling the minority class or undersampling the majority class, specialized strategies can be employed to alter the sample composition in the training dataset.

Classification Techniques

- There are a number of different classification techniques to build a model for classification
  - Logistic Regression
  - Naïve Bayes
  - K-nearest Neighbors
  - Decision Tree based Methods
  - Random Forest algorithm
  - Support Vector Machines
  - Rule-based Methods
  - Genetic Algorithms

Logistic Regression

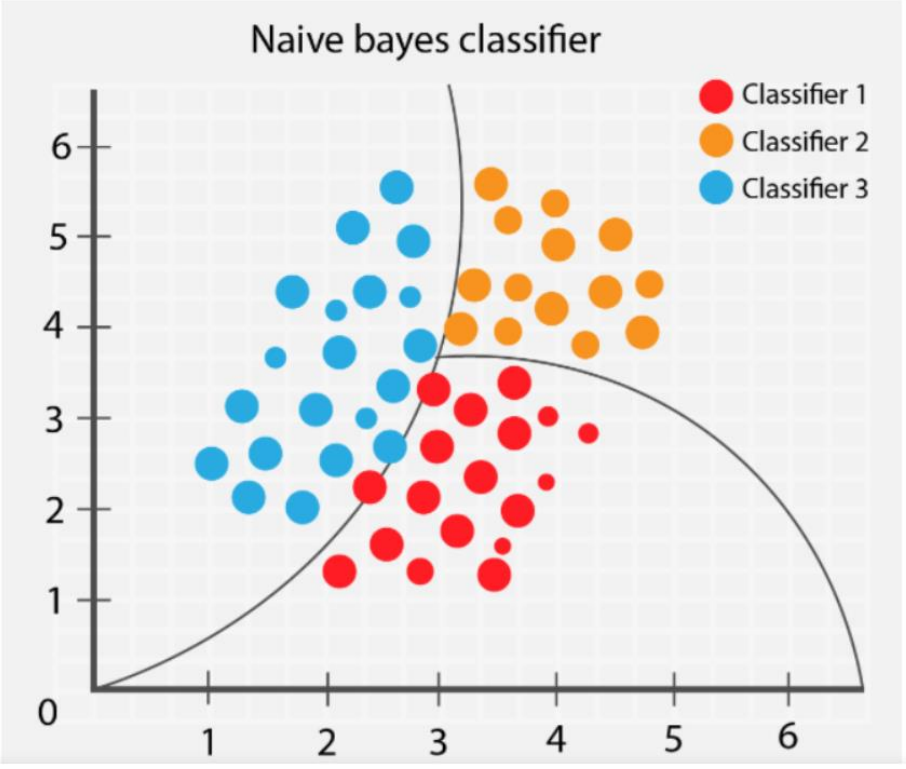
- Forecasts the likelihood of a target variable
- There will only be a choice between two classes.
- Data can be coded as either
  - representing success: 1 or yes
  - representing failure: 0 or no



Naïve Bayes

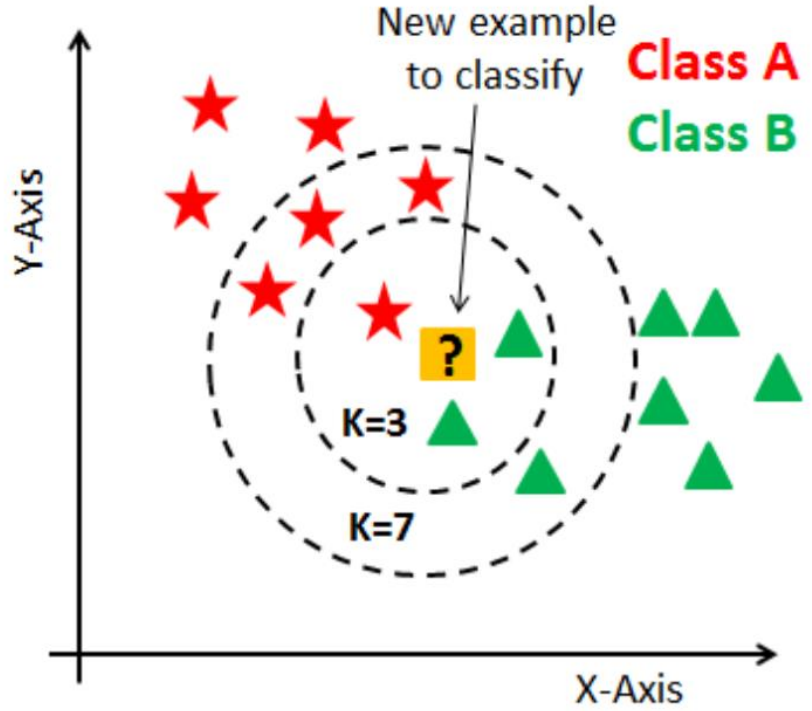
- Determines whether a data point falls into a particular category.
- It can be used to classify phrases or words in text analysis as either falling within a predetermined classification or not.

Text	Tag
“A great game”	Sports
“The election is over”	Not Sports
“What a great score”	Sports
“A clean and unforgettable game”	Sports
“The spelling bee winner was a surprise”	Not Sports



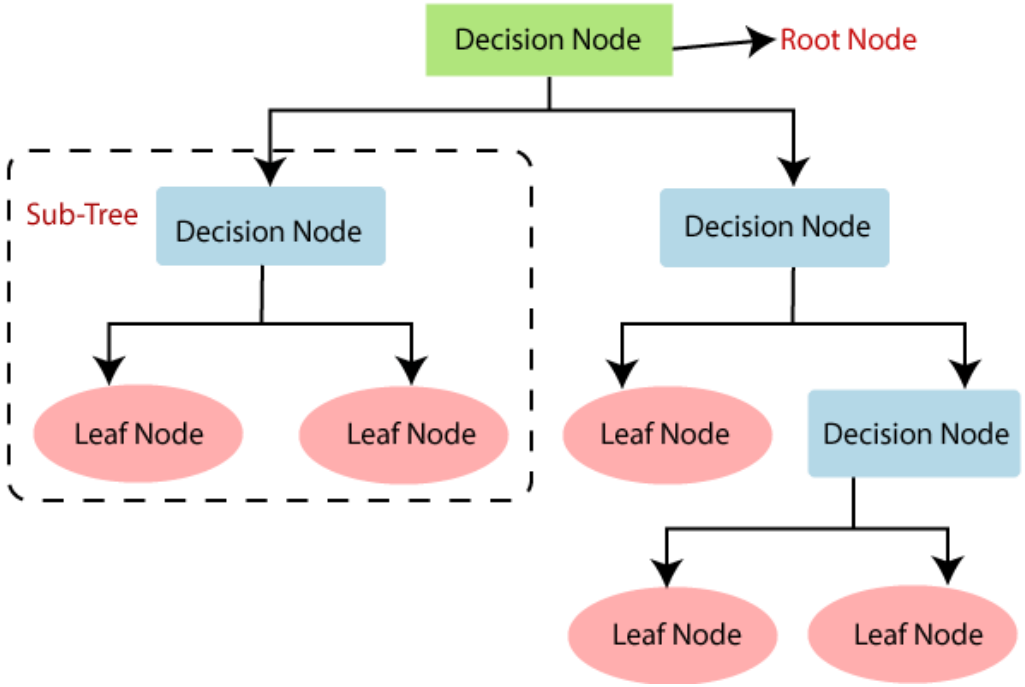
K-Nearest Neighbors

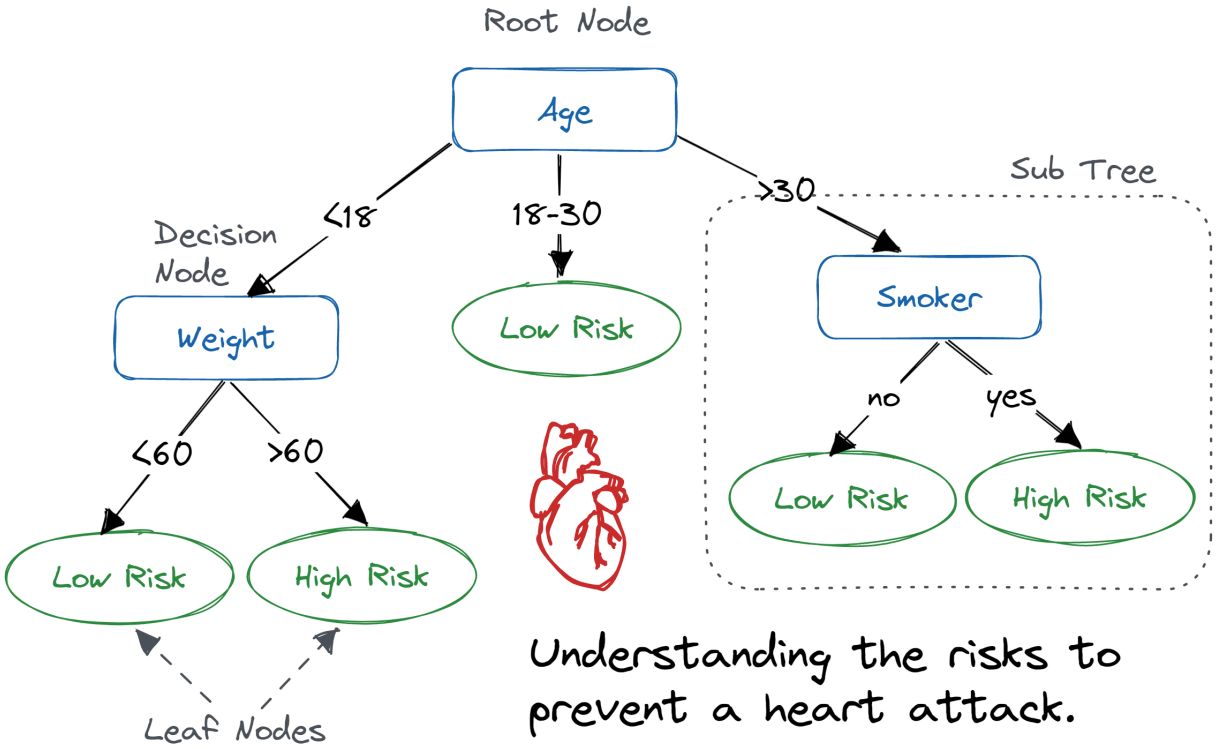
- Calculates the likelihood that a data point will join the groups based on which group the data points closest to it are a part of.
- When using k-NN for classification, you determine how to classify the data according to its nearest neighbor.



Decision Tree

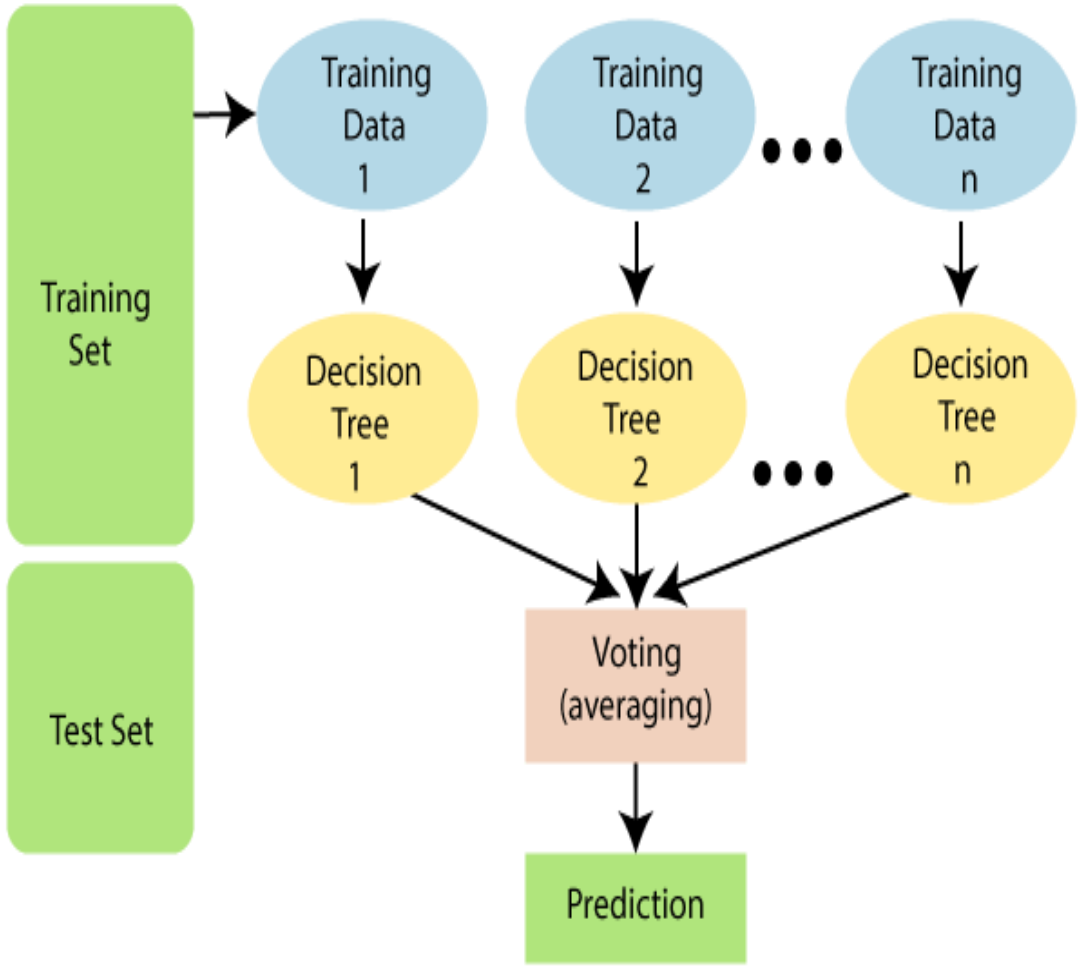
- Similar to a flow chart, it divides data points into two similar groups at a time, starting with the "tree trunk" and moving through the "branches" and "leaves" until the categories are more closely related to one another.





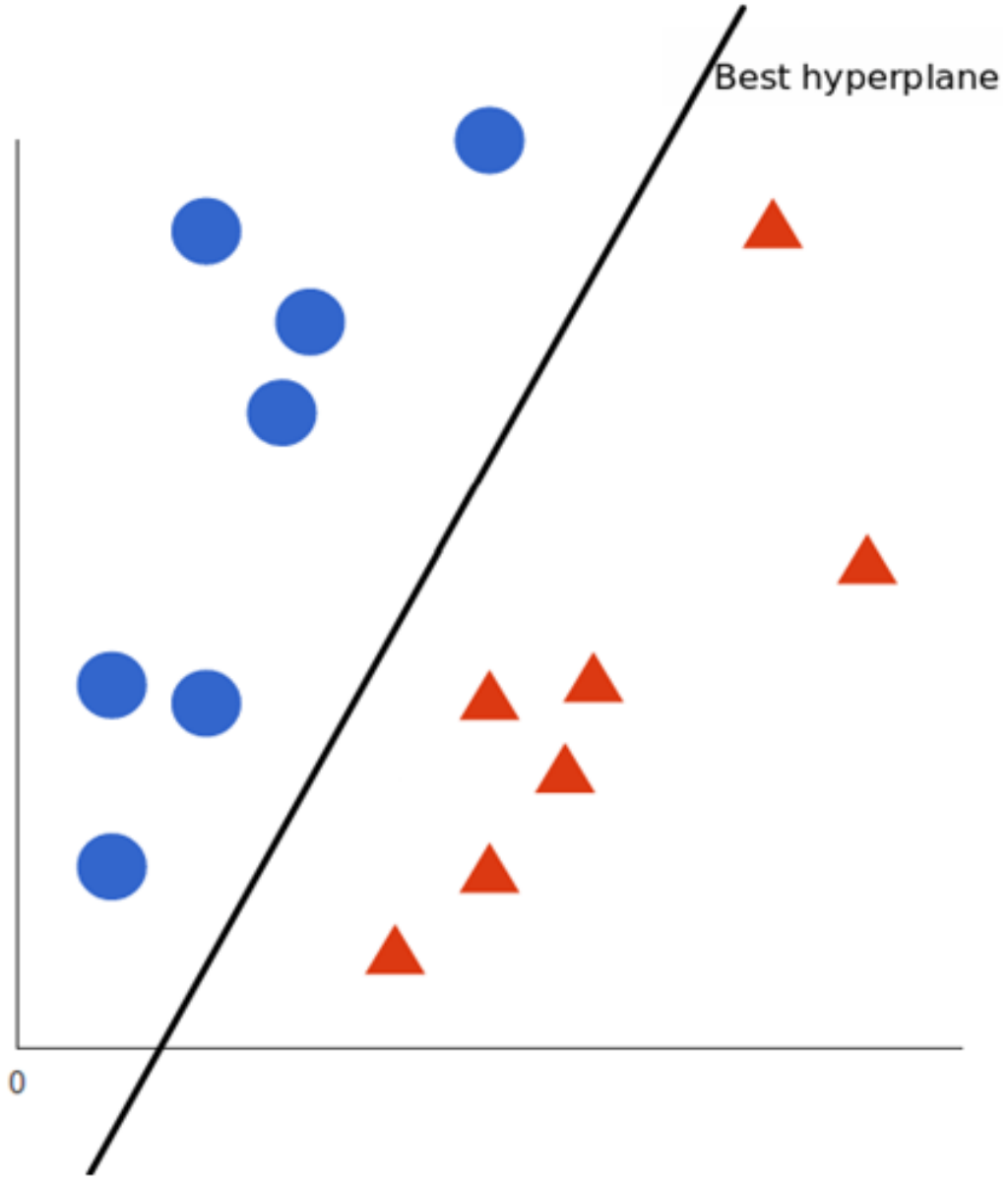
Random Forest Algorithm

- An extension of the Decision Tree algorithm
- You first create a number of decision trees using training data and then fit your new data into one of the created ‘tree’ as a ‘random forest’.
- It averages the data to connect it to the nearest tree data based on the data scale.
- These models are great for improving the decision tree’s problem of forcing data points unnecessarily within a category.



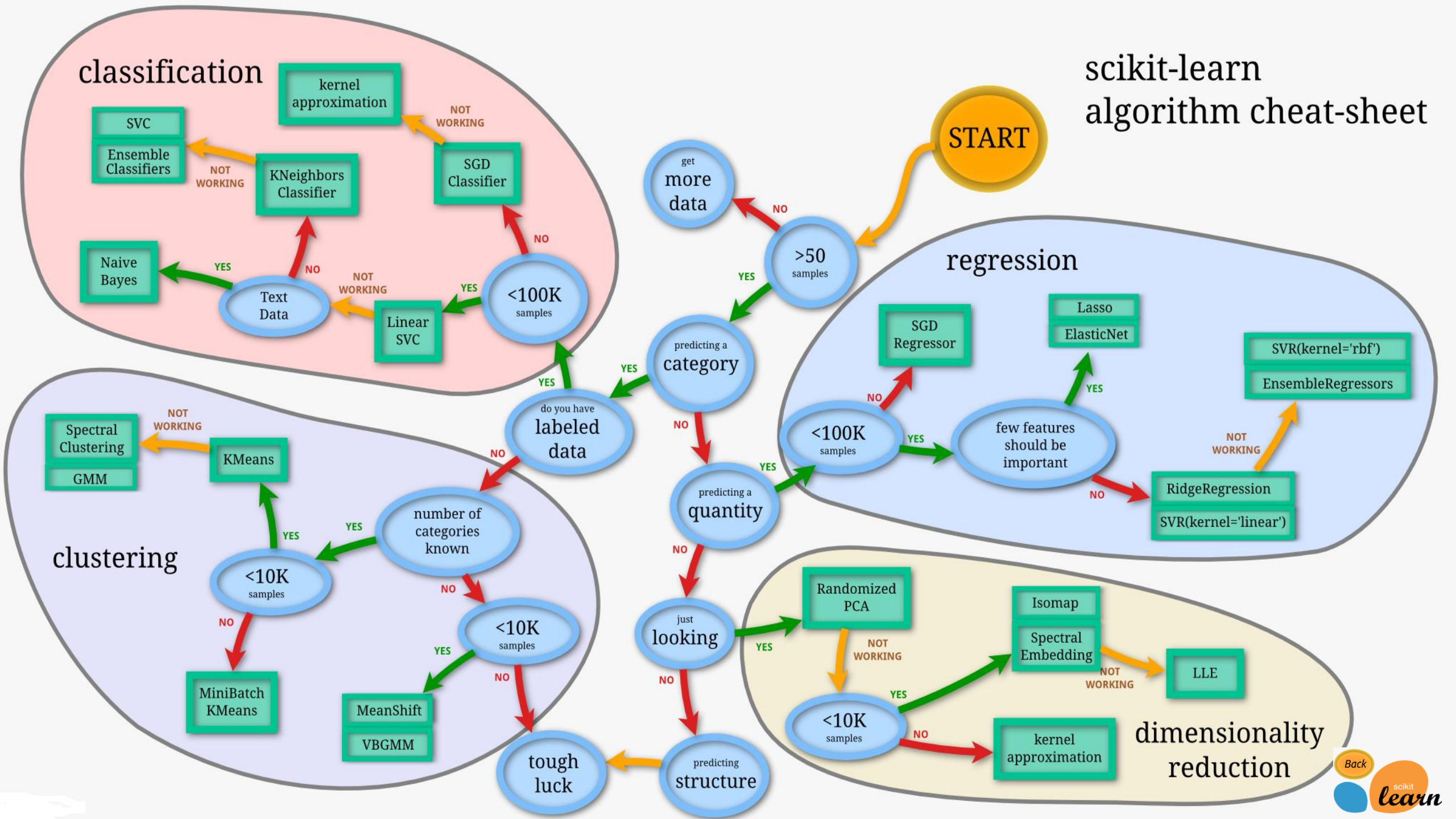
Support Vector Machine

- Work by creating a decision boundary called a "hyperplane."
- In two-dimensional space, this hyperplane is like a line that separates two sets of labeled data.
- Goal is to find the best possible decisior boundary by maximizing the margin between the two sets of labeled data.
- Reliable and can work well even with a small amount of data



Machine Learning Algorithms

- Popular machine learning algorithms:
  - https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/
- Scikit-learn arsenal of algorithms:
  - http://scikit-learn.org/stable/
- Supervised learning with scikit-learn:
  - http://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
tough luck
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
NO
predicting structure
dimensionality reduction
Back
scikit learn

Classifier Evaluation Metrics

- Accuracy
- Confusion Matrix
- Precision Recall
- F1-Score
- Specificity
- Receiver Operating Characteristics Curve (ROC)
- Area Under Curve (AUC)
- …

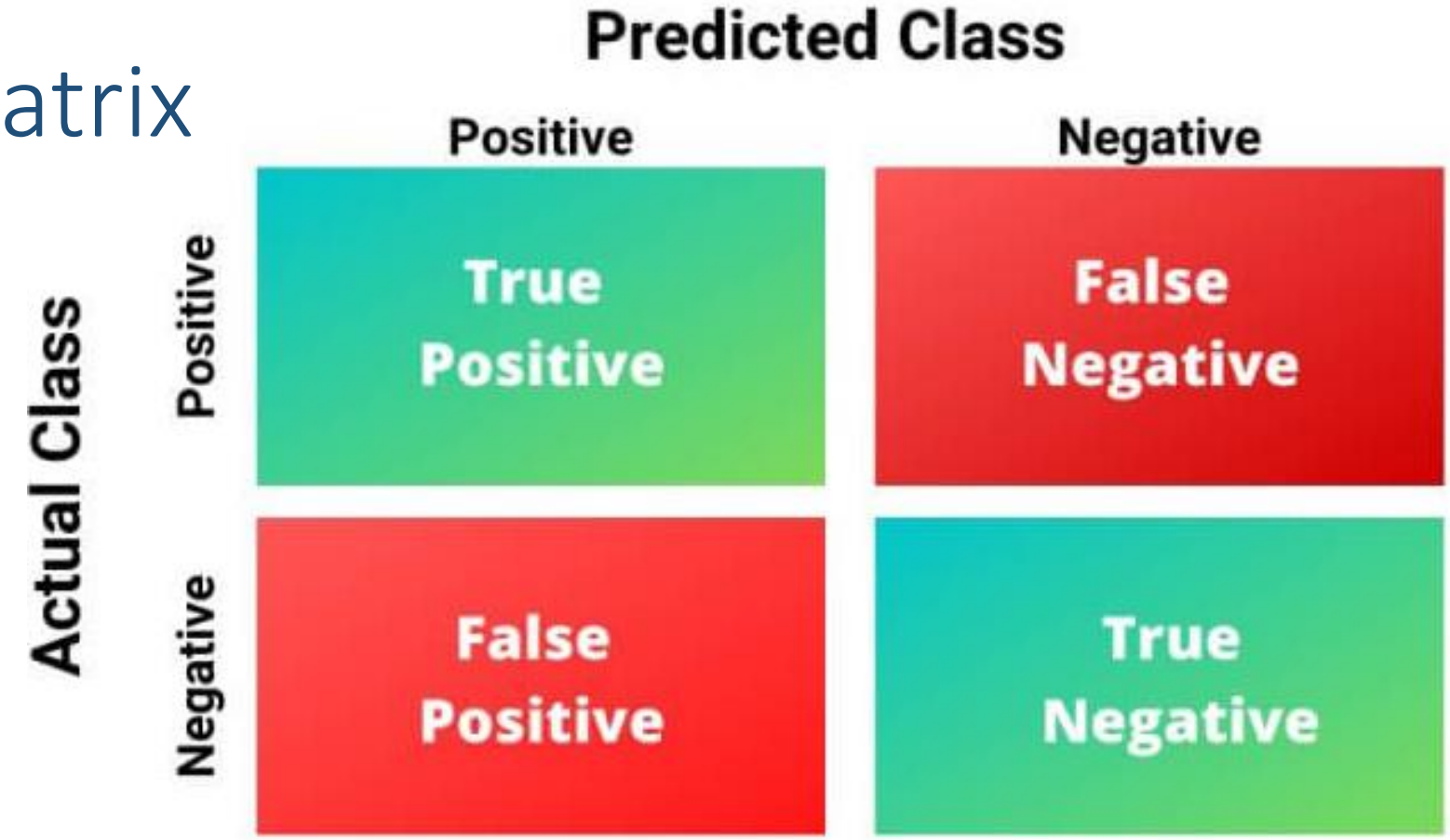
Accuracy

- Straight forward calculation

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

- Limitations:
  - Works well with a balanced data set but not with an unbalanced data set.

Confusion Matrix



- True Positive (TP): Predicted positive and it's true
- True Negative (TN): Predicted negative and it's true
- False Positive (FP): Predicted positive and it's false (Type I error)
- False Negative (FN)- We predicted negative and it's false (Type II error)

Precision & Recall

- Precision - how many of the correctly predicted cases actually turned out to be positive

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall - how many of the actual positive cases we were able to predict correctly with our model

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score

- Gives a combined idea about Precision and Recall metrics.
- It is maximum when Precision is equal to Recall.

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)}$$

- Higher the F1-Score, the better the model will be

Specificity

- How many of the actual negative cases we were able to predict correctly with our model

$$\text{Specificity} = \frac{TN}{TN + FP}$$

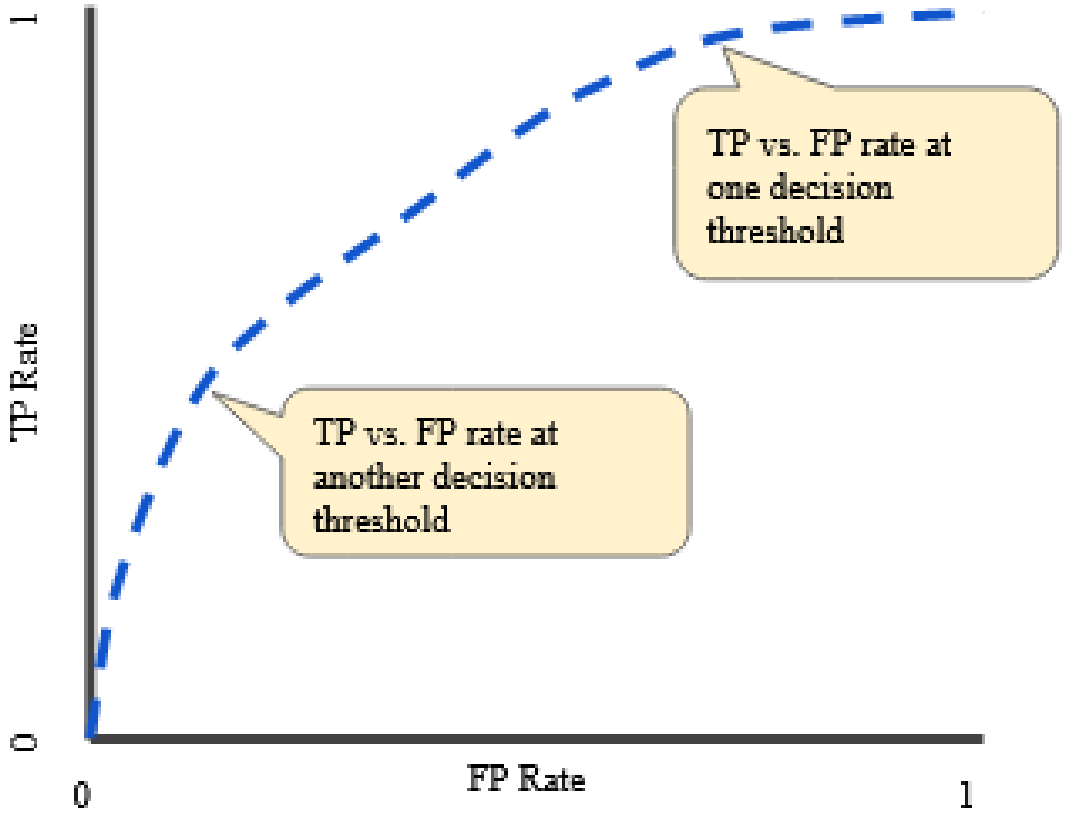
Receiver Operating Characteristics Curve (ROC)

- A graph showing the performance of a classification model at all classification thresholds
- Plots two parameters:
  - True Positive Rate (Recall)

$$TPR = \frac{TP}{TP + FN}$$

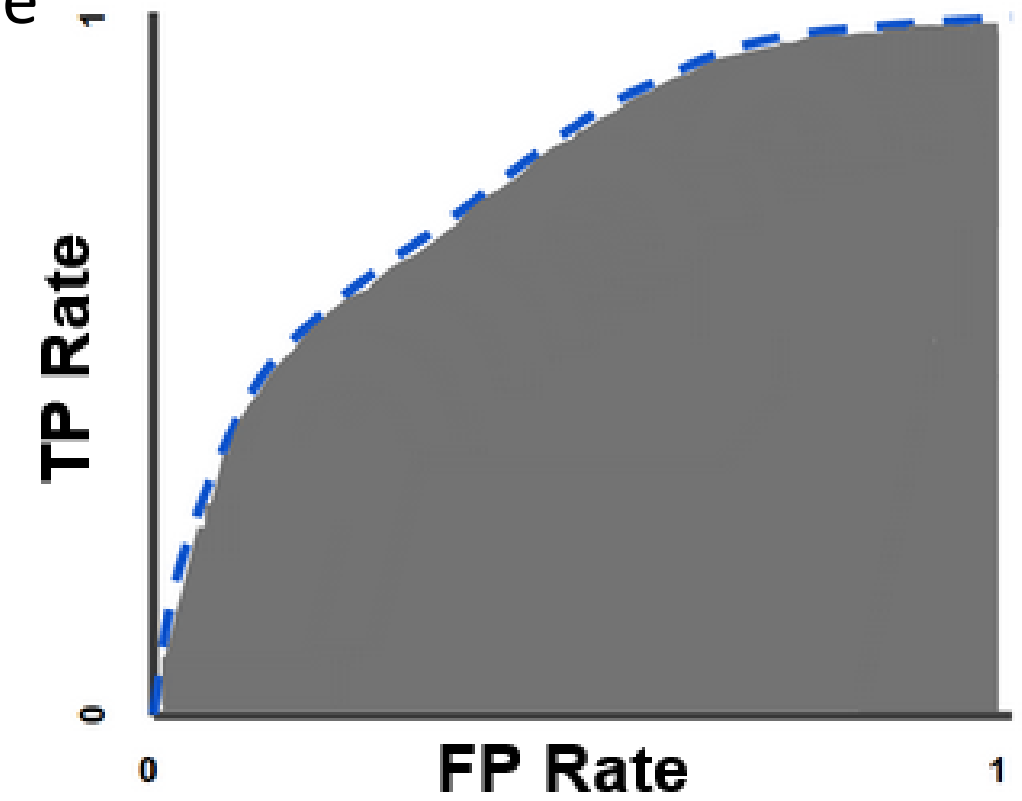
  - False Positive Rate (1-Specificity)

$$FPR = \frac{FP}{FP + TN}$$

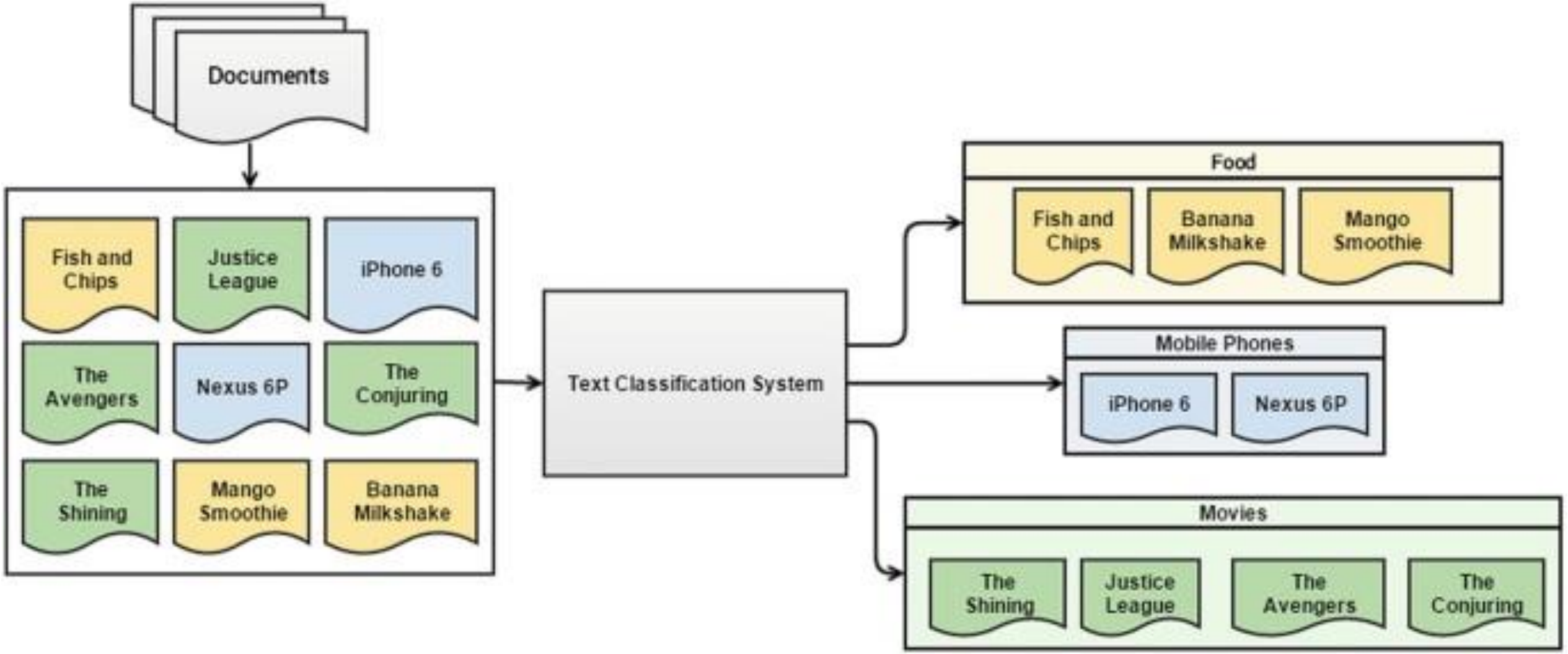


Area Under Curve (AUC)

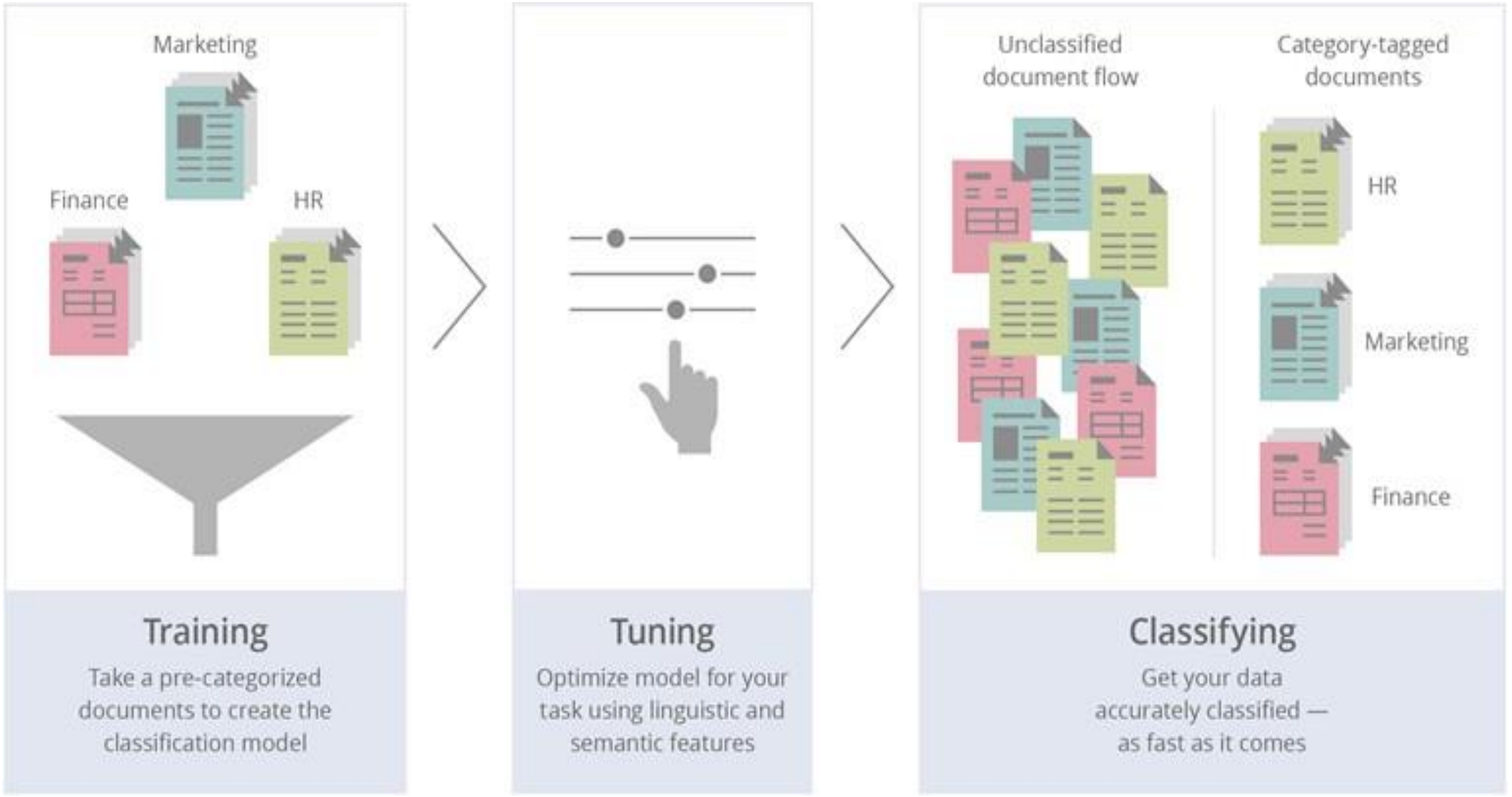
- Area under the ROC Curve
- probability that the model ranks a random positive example more highly than a random negative example



Text Classification



Supervised Classification of Text



Text Classification Pipeline

- Prepare and separate training and testing datasets
- Text normalization (expanding, removing special chars, words, lemmatizing)
- Feature extraction (vector space model)
- Model training (with training dataset)
- Model tuning (with validation dataset)
- Model prediction (with test dataset)
- Model evaluation (metrics)

Text Classification Pipeline