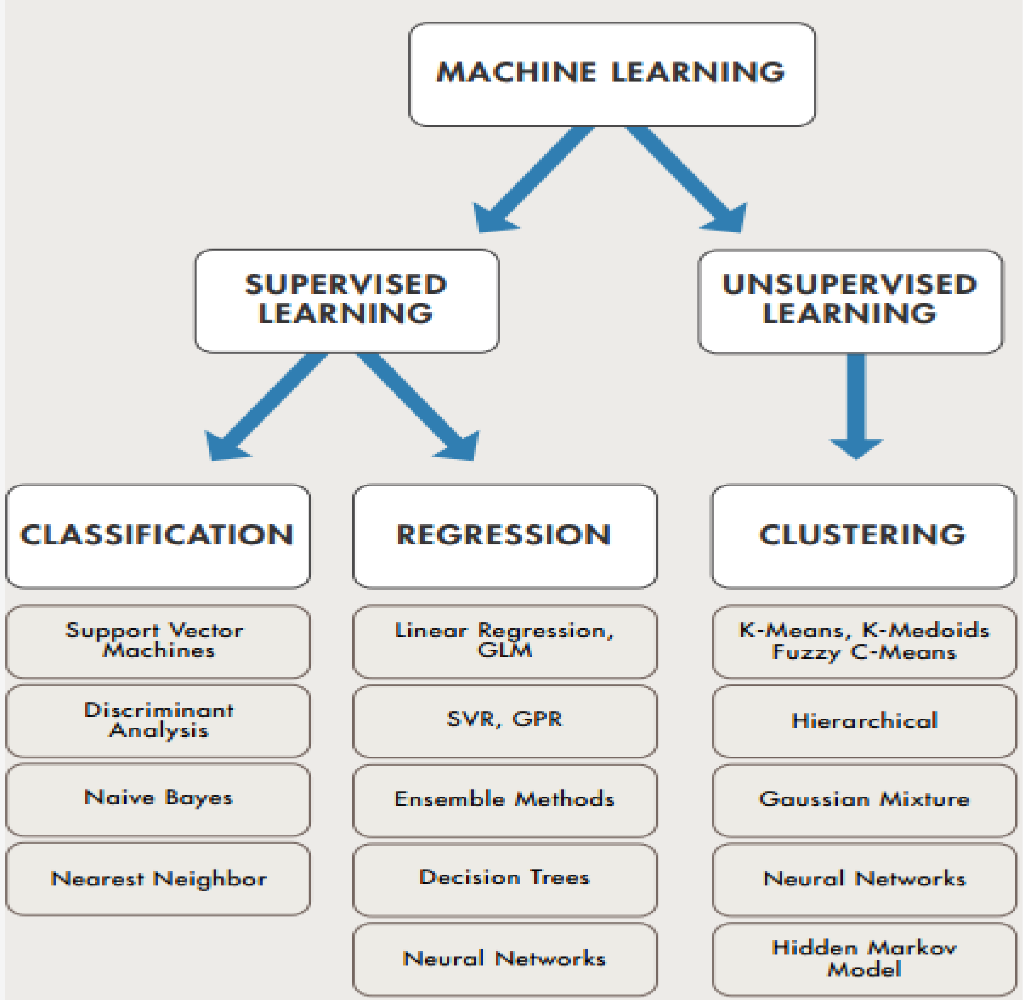
Part 2 – Common Machine Learning Algorithms

Where are we ?

Part 2 – Common Machine Learning Algorithms



MACHINE LEARNING

SUPERVISED LEARNING

UNSUPERVISED LEARNING

CLASSIFICATION

- Support Vector Machines
- Discriminant Analysis
- Naive Bayes
- Nearest Neighbor

REGRESSION

- Linear Regression, GLM
- SVR, GPR
- Ensemble Methods
- Decision Trees
- Neural Networks

CLUSTERING

- K-Means, K-Medoids Fuzzy C-Means
- Hierarchical
- Gaussian Mixture
- Neural Networks
- Hidden Markov Model

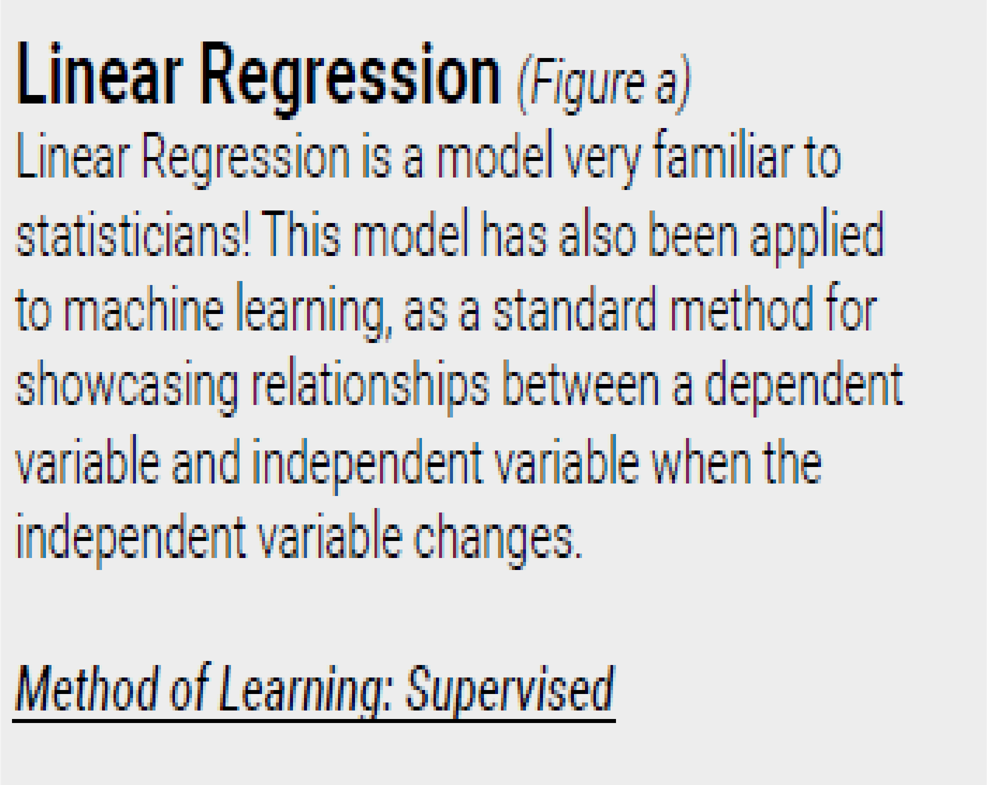
Part 2 – Common Machine Learning Algorithms

- We will not cover every one of the machine learning methodologies in details in this course.
- Instead we will go over the high level intuition of many machine learning problems and its solution in this lecture. We will focus in more details in some of the most common techniques in the next few weeks.
- So, you will at least understand what many of the terminology and have a solid foundation in some of the most common algorithms

Common Data Mining Problems and Solutions

Supervised Learning (Predictive ability based on past data)	Classification – Machine Learning	Decision Trees K-Nearest Neighbors Support Vector Machine
		Neural Networks
	Classification - Statistics	Regression; Dimension Reduction
Unsupervised Learning (Exploratory analysis to discover patterns)	Clustering Analysis	K-Means
	Association Rules	Apriori

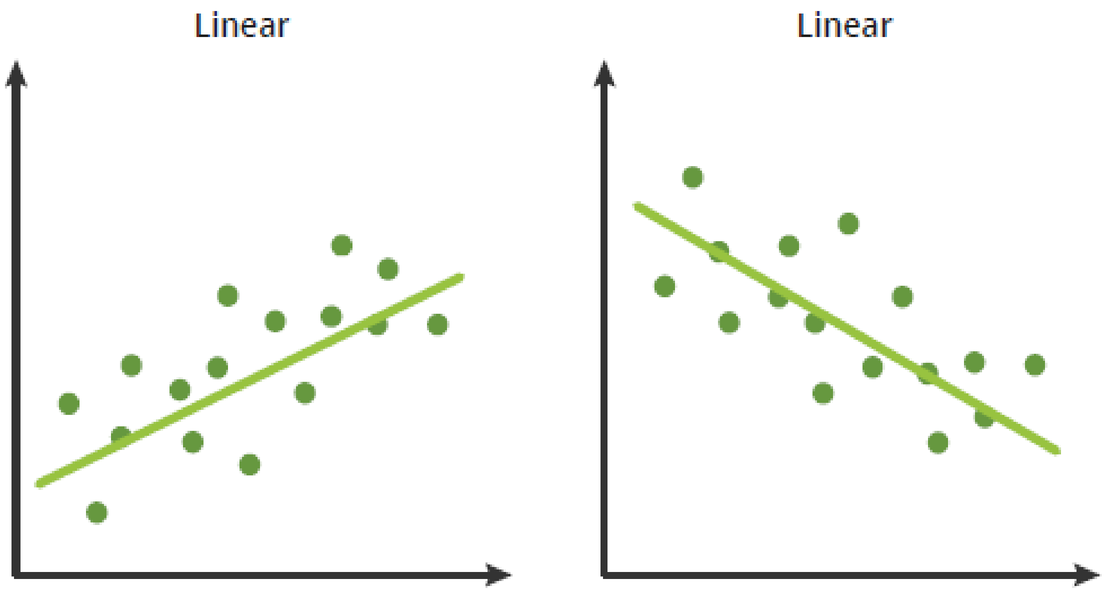
Machine Learning Overview



Linear Regression *(Figure a)*

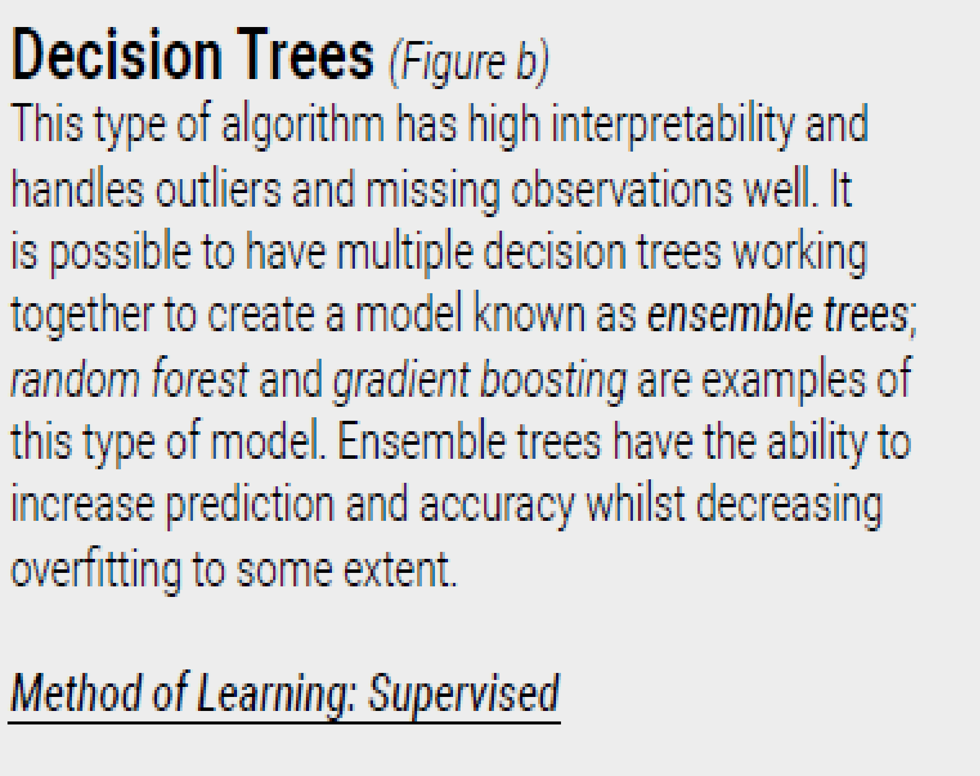
Linear Regression is a model very familiar to statisticians! This model has also been applied to machine learning, as a standard method for showcasing relationships between a dependent variable and independent variable when the independent variable changes.

<u>Method of Learning: Supervised</u>



(Figure a)

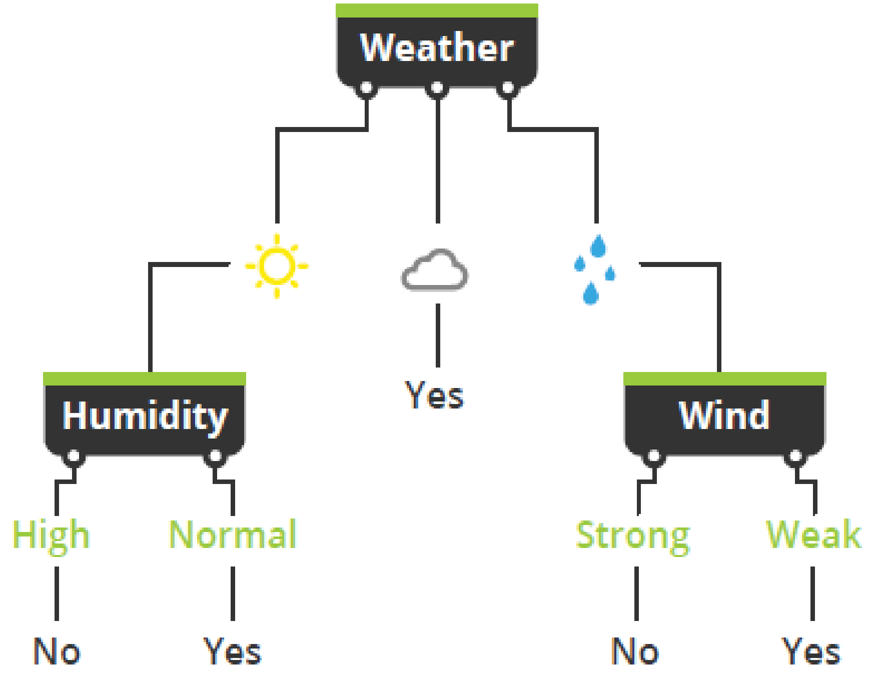
Machine Learning Overview



Decision Trees *(Figure b)*

This type of algorithm has high interpretability and handles outliers and missing observations well. It is possible to have multiple decision trees working together to create a model known as *ensemble trees*; *random forest* and *gradient boosting* are examples of this type of model. Ensemble trees have the ability to increase prediction and accuracy whilst decreasing overfitting to some extent.

Method of Learning: Supervised

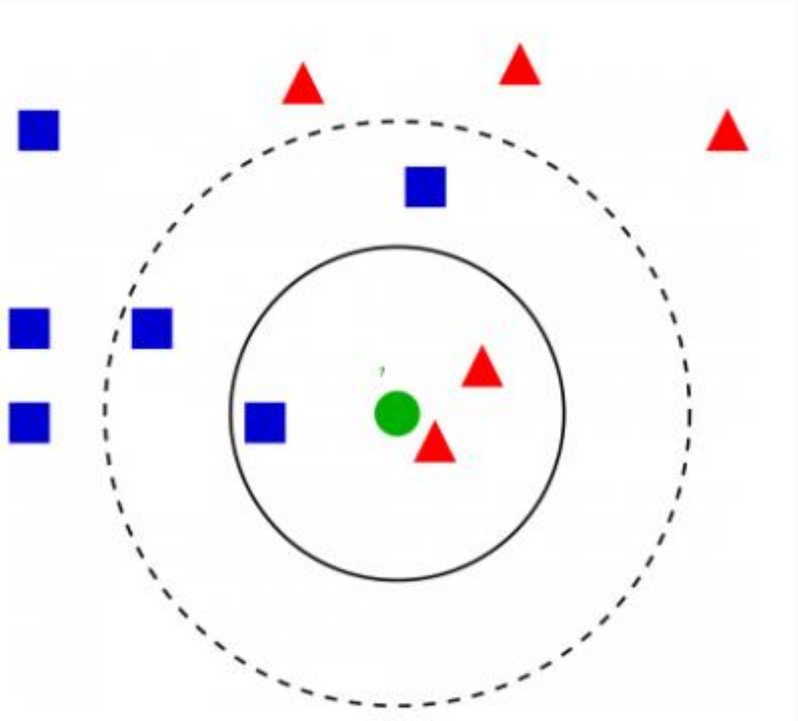


(Figure b)

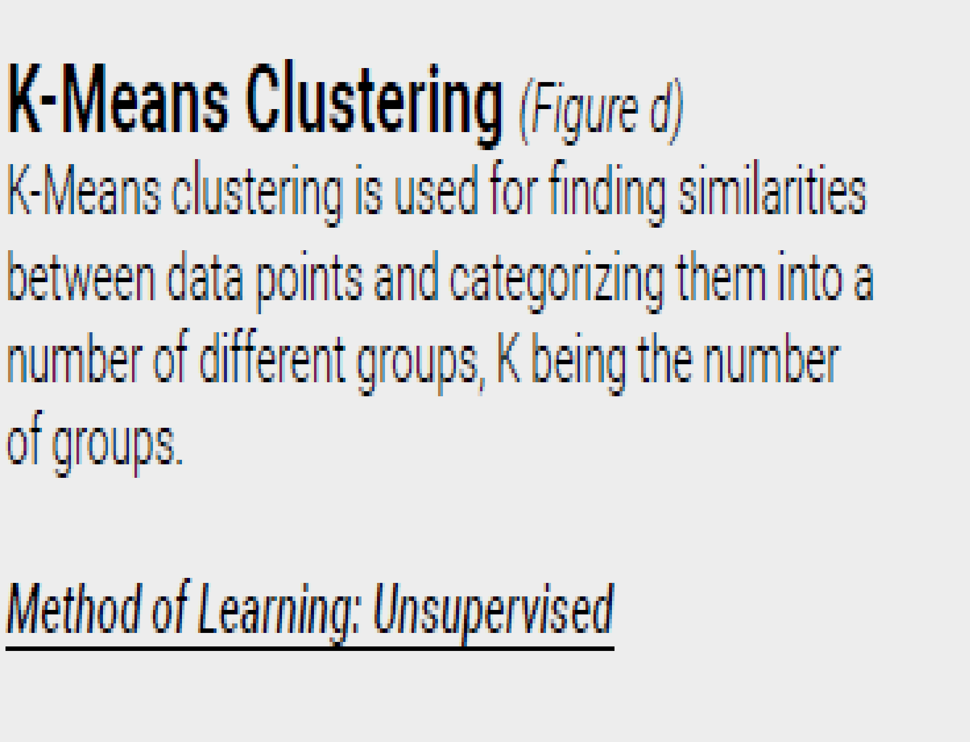
Machine Learning Overview

K-Nearest Neighbors

is one of the most popular machine learning algorithms in solving supervised classification problem. The idea is based on computing a similarity metrics among data points. To classify a new observation is to identify the nearest K number of neighbors based on the similarity metrics, the prediction label will be the most common class label among all the K neighbors.



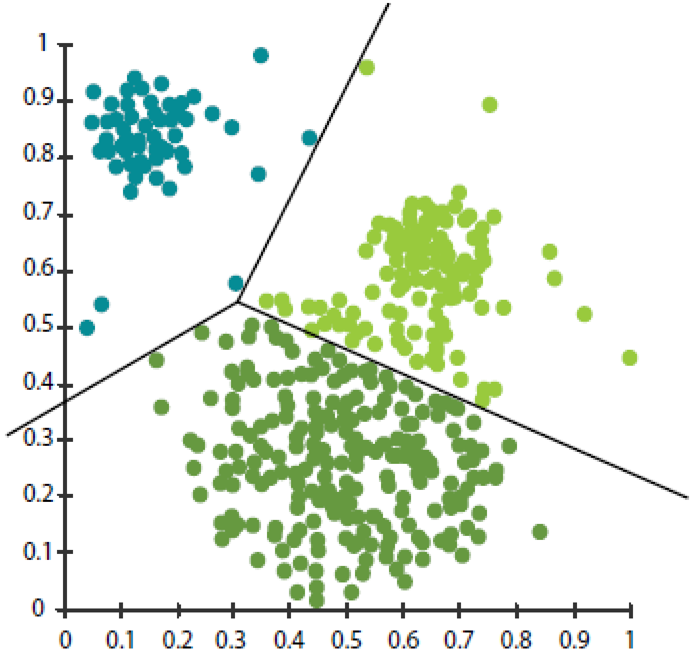
Machine Learning Overview



K-Means Clustering *(Figure d)*

K-Means clustering is used for finding similarities between data points and categorizing them into a number of different groups, K being the number of groups.

Method of Learning: Unsupervised



(Figure d)

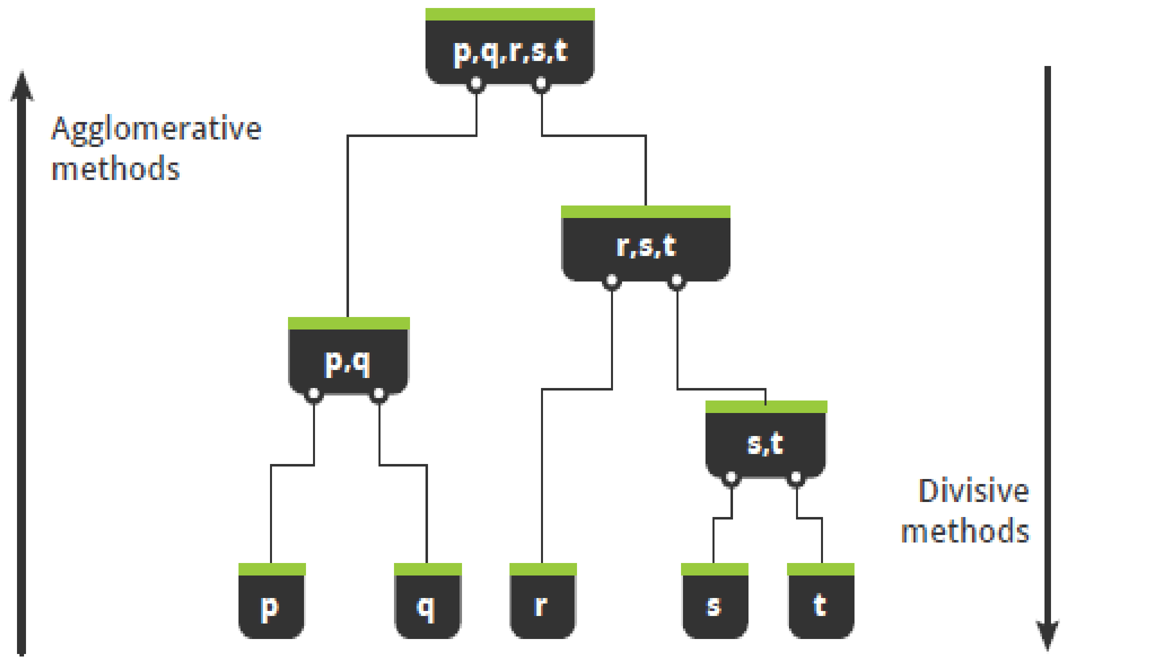
Machine Learning Overview

Hierarchical Clustering *(Figure e)*

Hierarchical clustering creates a known number of overlapping clusters of different sizes along a hierarchical tree to form a classification system. This type of clustering can be achieved through various methods with the most common methods being *agglomerative* and *divisive*.

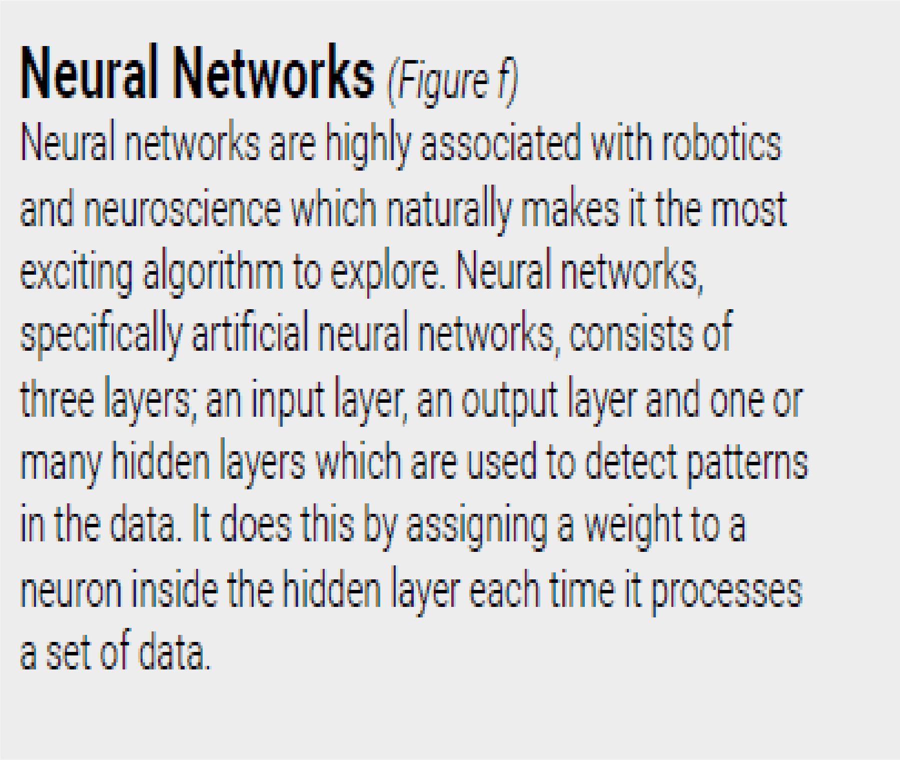
The agglomerative approach is a bottom-up method which consists of all objects starting within their own respective clusters. These clusters of objects are then joined together by taking the two most similar clusters and merging them. Conversely, divisive clustering takes a top-down approach where all the objects start in the same cluster and are then divided into two separate clusters through an algorithmic process similar to K-Means. The splitting process is repeated until the desired number of clusters is achieved.

Method of Learning: Unsupervised



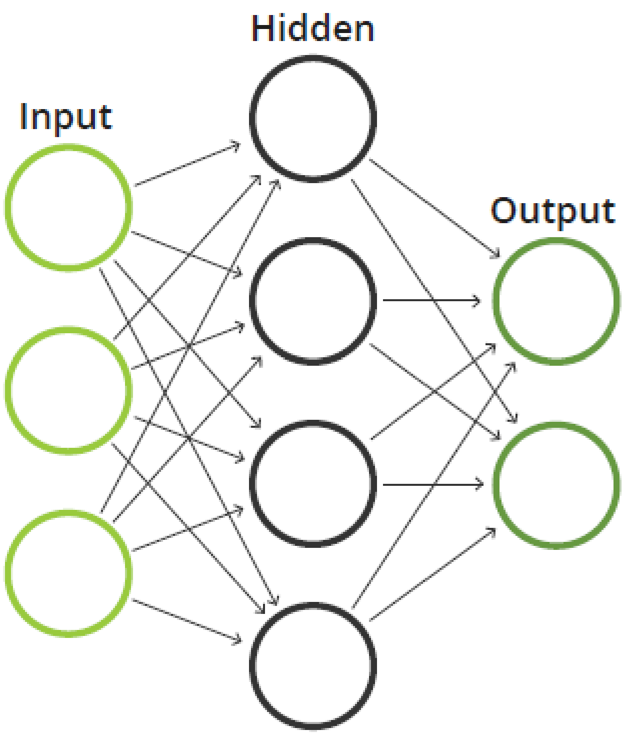
(Figure e)

Machine Learning Overview



Neural Networks *(Figure f)*

Neural networks are highly associated with robotics and neuroscience which naturally makes it the most exciting algorithm to explore. Neural networks, specifically artificial neural networks, consists of three layers; an input layer, an output layer and one or many hidden layers which are used to detect patterns in the data. It does this by assigning a weight to a neuron inside the hidden layer each time it processes a set of data.



(Figure f)

More Advanced Techniques (which we will not cover in this course)

- Ensembled Techniques
  - Random Forest
  - AdaBoost
- Reinforcement Learning
  - Q-Learning
- Deep Learning
- Survival Analysis
  - Cox Proportional Hazard Model
- Natural Language Processing

I will take you up to the 19th century, but it will still be quite an achievement