Imperial College
London

Introduction to Machine Learning

Valentina Giunchiglia

v.giunchiglia20@imperial.ac.uk

What is Machine Learning?

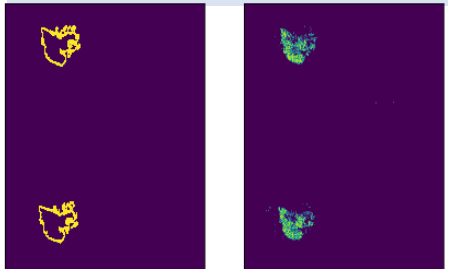
"Machine Learning: a computer is able to learn from experience without being specifically programmed"

Self-drinving cars

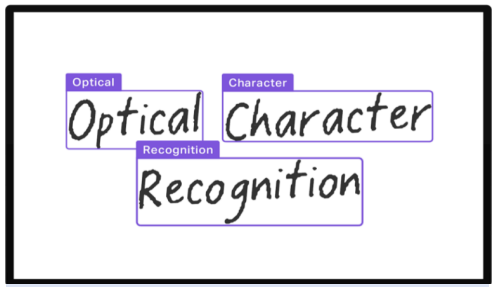
Face recognition

Content ads

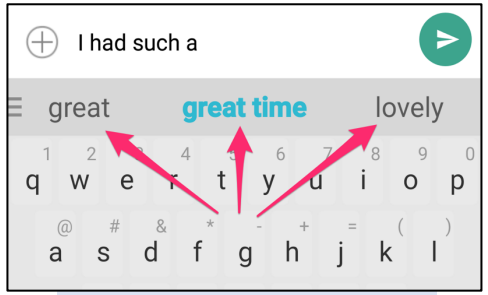
Biomedical Research



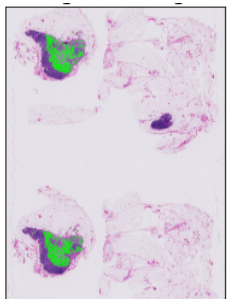
Speech recognition



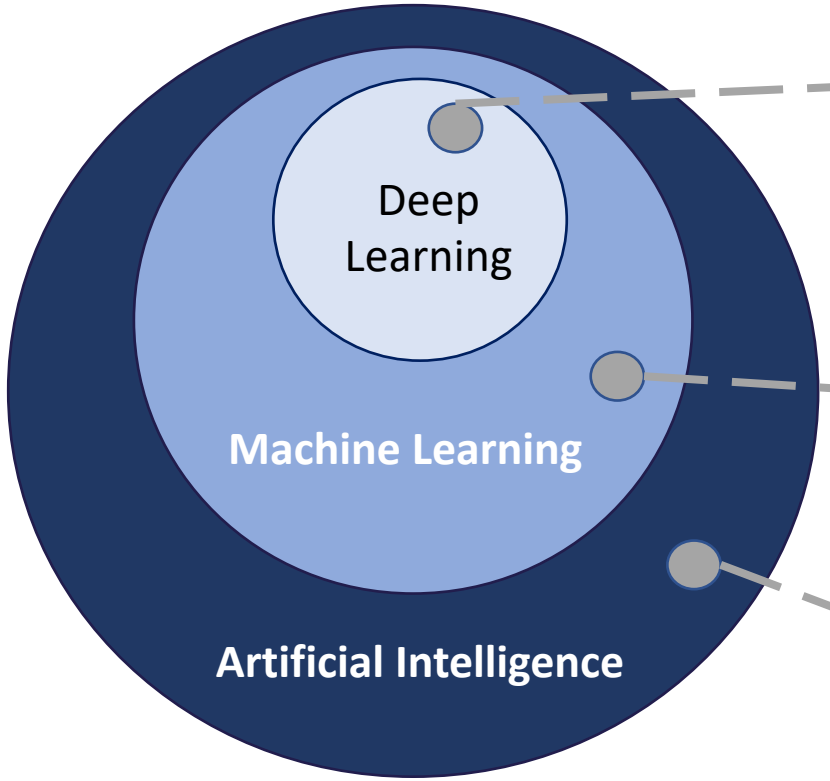
Character Recognition



Typing prediction



What is Machine Learning?



Subset of machine learning that uses artifical neural networks

Stream 1

Application of AI that allows a machine to learn and improve from experience

Today

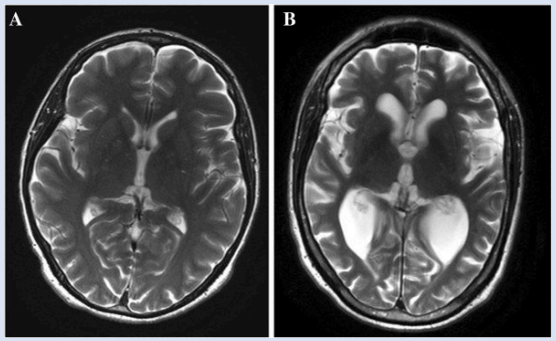
Any technique/method that alllows machines to mimic human behaviour

What is Machine Learning?

Three main **types** of Machine Learning

Supervised

ML approach that uses labeled data to learn and predict outcomes

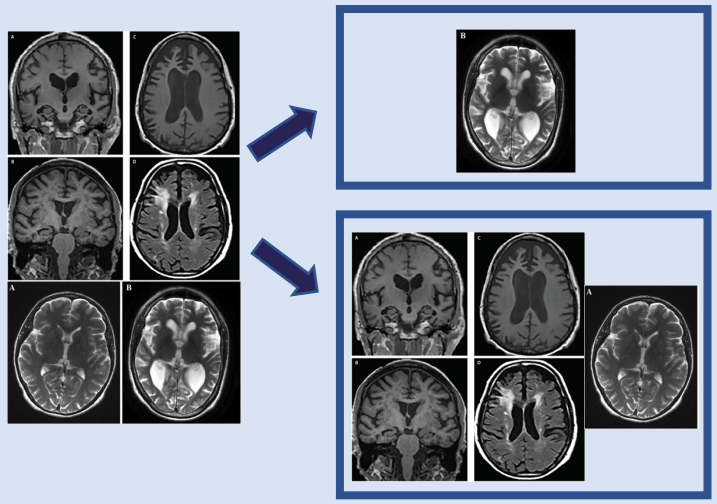


Alzheimer Healthy

Task driven

Unsupervised

ML approach that uses unlabeled data



Data driven

Reinforcement Learning

ML approach that allows an AI-driven system to learn through trial and error using feedback from its actions

Beyond the scope of this lecture

Learning from mistakes

Cornelius et al, 2013

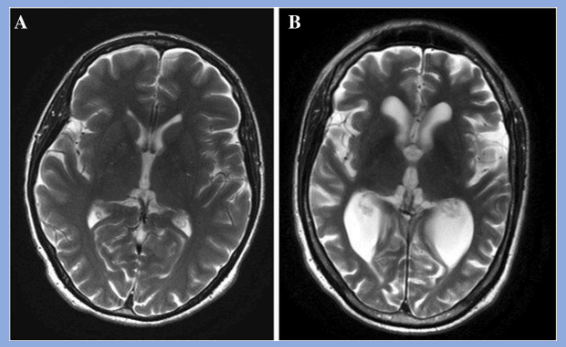
Ayers et al, 2019

What is Machine Learning?

Three main **types** of Machine Learning

Supervised

ML approach that uses labeled data to learn and predict outcomes

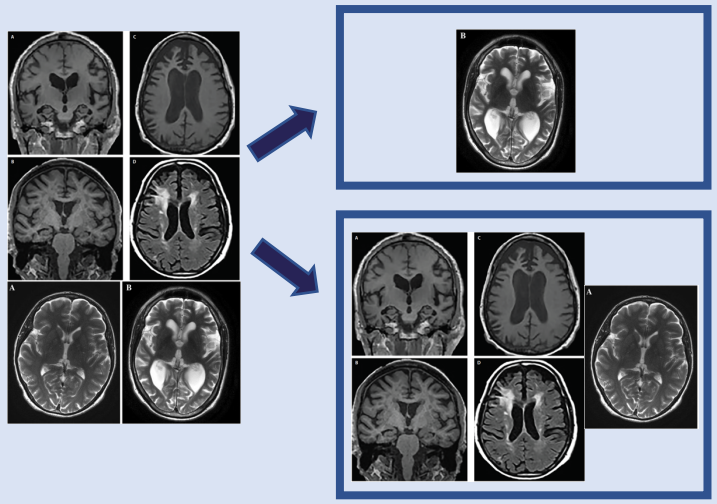


Alzheimer Healthy

Task driven

Unsupervised

ML approach that uses unlabeled data



Data driven

Reinforcement Learning

ML approach that allows an AI-driven system to learn through trial and error using feedback from its actions

Beyond the scope of this lecture

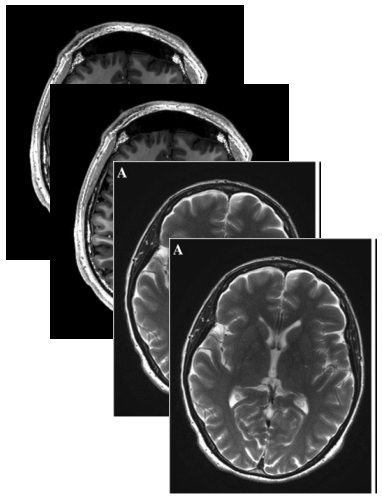
Learning from mistakes

Cornelius et al, 2013

Ayers et al, 2019

Supervised learning can be divided into **CLASSIFICATION** and REGRESSION

Input data



Healthy

Healthy

Alzheimer

Alzheimer

…

…

Data

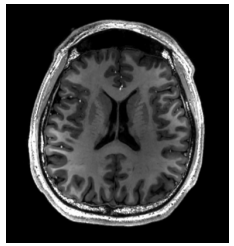
Categorical Label
(e.g. disease status)

Model training using ML algorithm

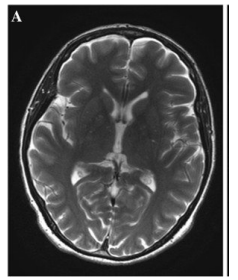
ML model

Output of ML algorithm, can be used to make predictions on unseen data

Prediction



Healthy



Alzheimer

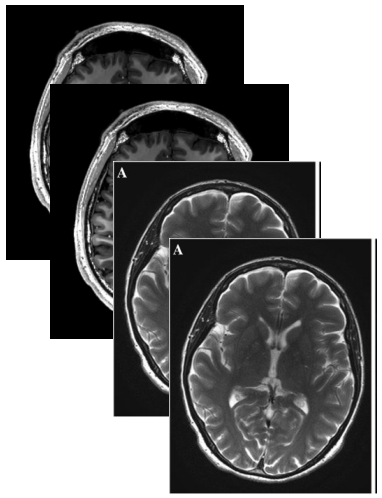
…

CLASSIFICATION: prediction of a categorical outcome label

Cornelius et al, 2013

Supervised learning can be divided into CLASSIFICATION and **REGRESSION**

Input data



70

65

67

72

…

…

Data

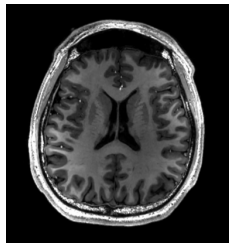
Continous Label (e.g. Age)

Model training using ML algorithm

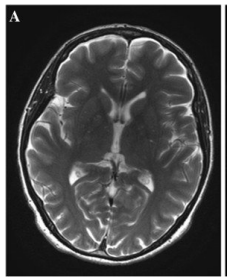
ML model

Output of ML algorithm, can be used to make predictions on unseen data

Prediction



63



74

…

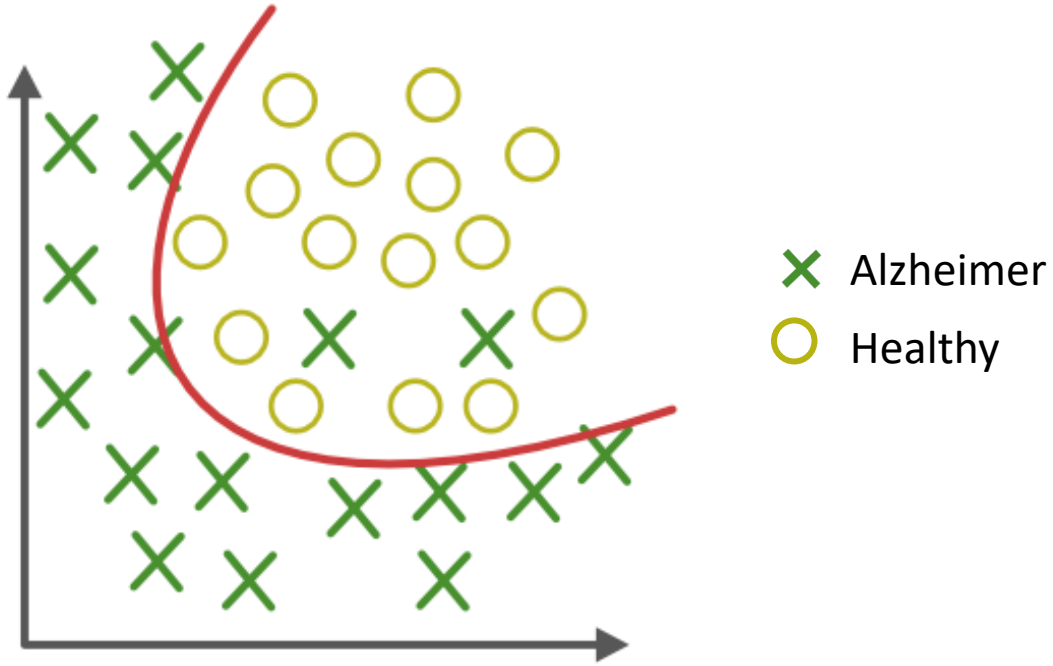
Regression: prediction of a continous outcome label

Cornelius et al, 2013

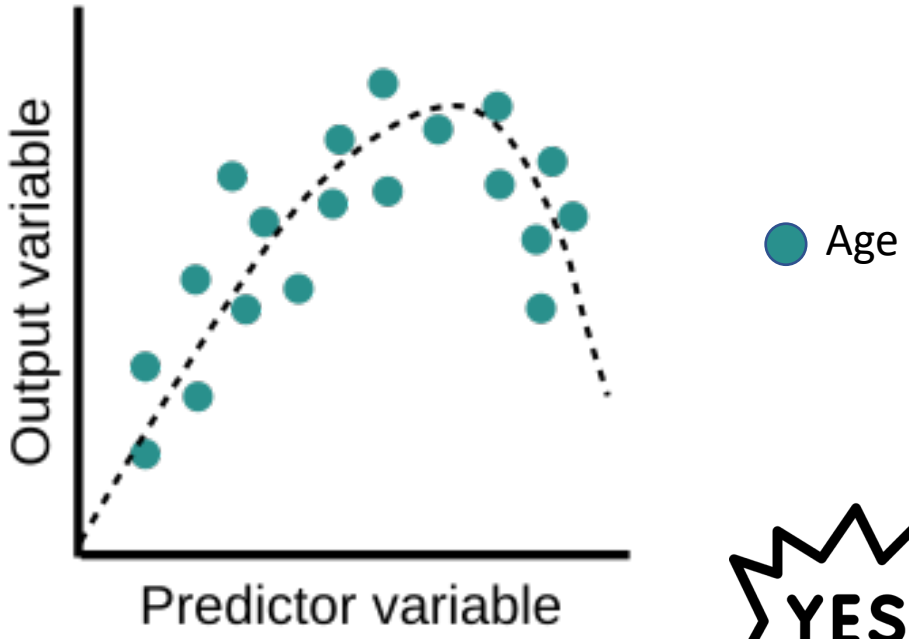
Classification vs Regression

What is the ML model trying to learn? → The ML model tries to learn what is the best line (in red) to separate the different classes (**CLASSIFICATION**) or to fit the data (**REGRESSION**)

Classification



Regression

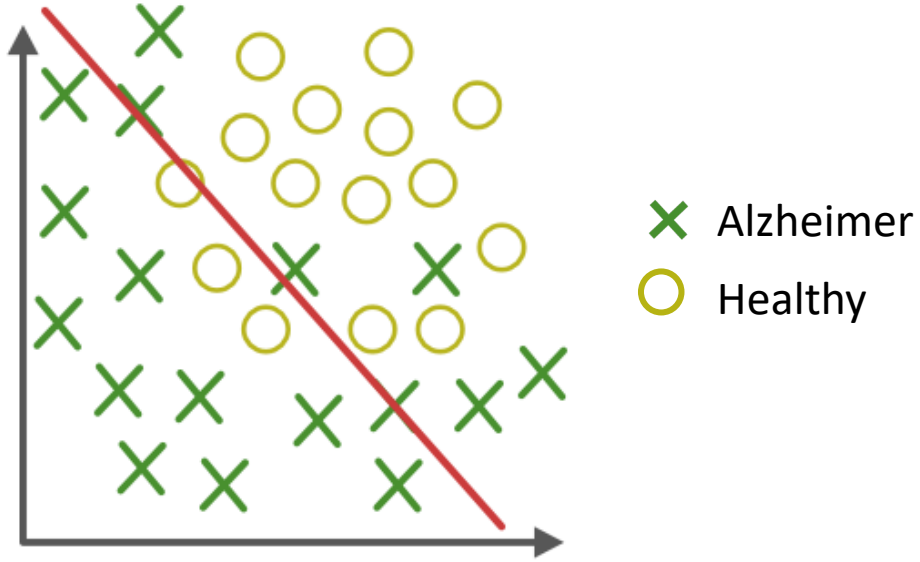


In real life there is NEVER a perfect split or fit to the data and the best line is hard to find → Can your model be wrong?

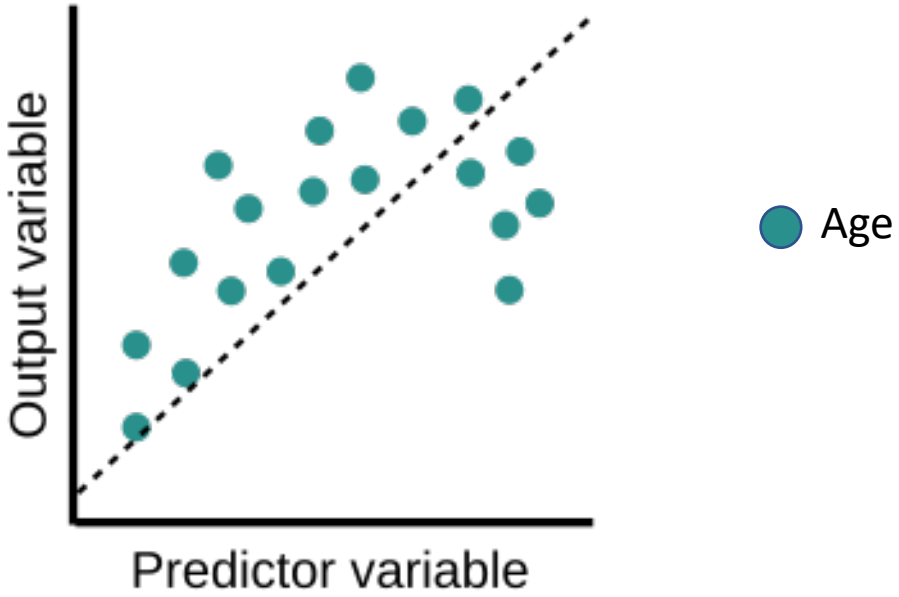
Underfitting

UNDERFITTING: the model cannot properly learn how to predict the categorical or continuos label and performs poorly both on the data on which it was trained and on unseen data

Classification



Regression



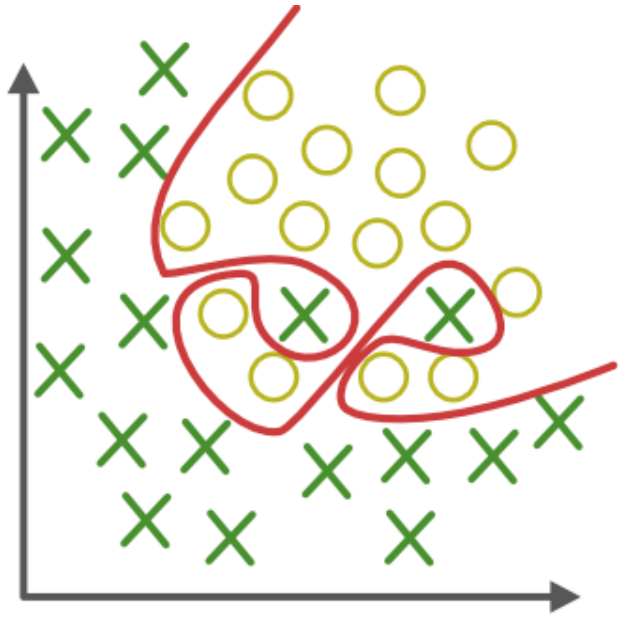
Potential **causes**:
- ✗ Not enough training samples
- ✗ Faetures provided are not enough

Images from here

Overfitting

OVERFITTING: the model learns the data on which it was trained *too well* and performs poorly on unseen data

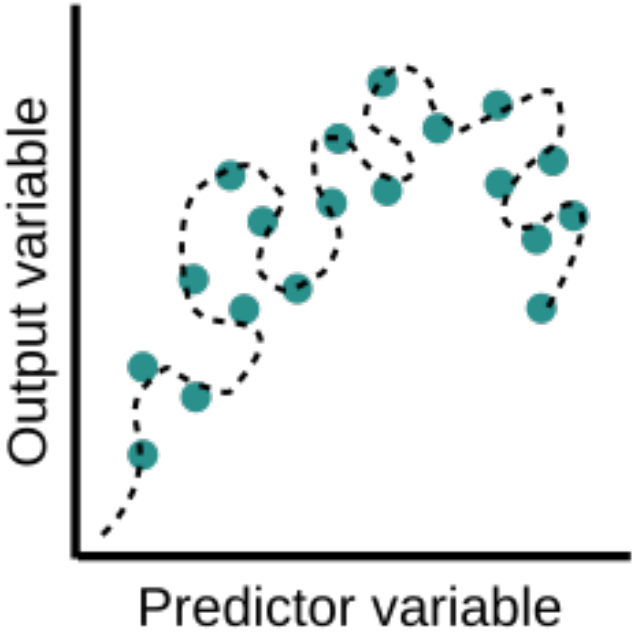
Classification



Alzheimer

Healthy

Regression

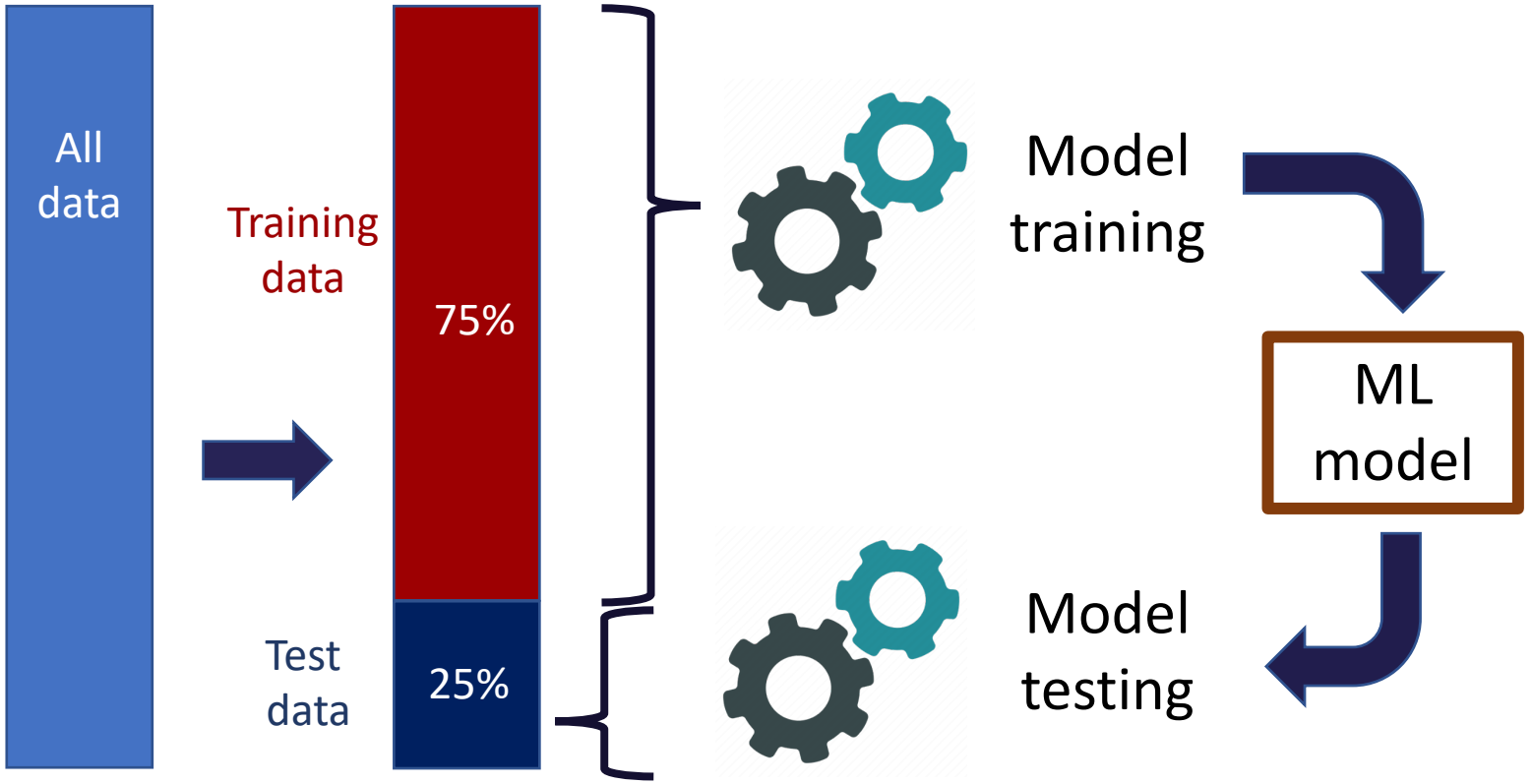


Age

Potential **causes**:
- Wrong type of model
- Too many features in the input

Is your model overfitted or underfitted?

The standard approach to test whether the model is underfitted/overfitted is to train it on a part of the data and test it on an hold-out dataset



Percentage of splitting depends on size of dataset + ML algorithm used

Predictions on unseen data

OVERFITTING

- Train data performance ↑
- Test data performance ↓

UNDERFITTING

- Train data performance ↓
- Test data performance ↓

K-Fold cross validation

! Splitting in train/test is RANDOM, but based on the splitting performance can change → Some samples are **EASIER/HARDER** to predict than others

✓ **K-Fold cross validation:** compute the train/test split K times, and calculate the performance on K different test sets

→ Final performance = Average

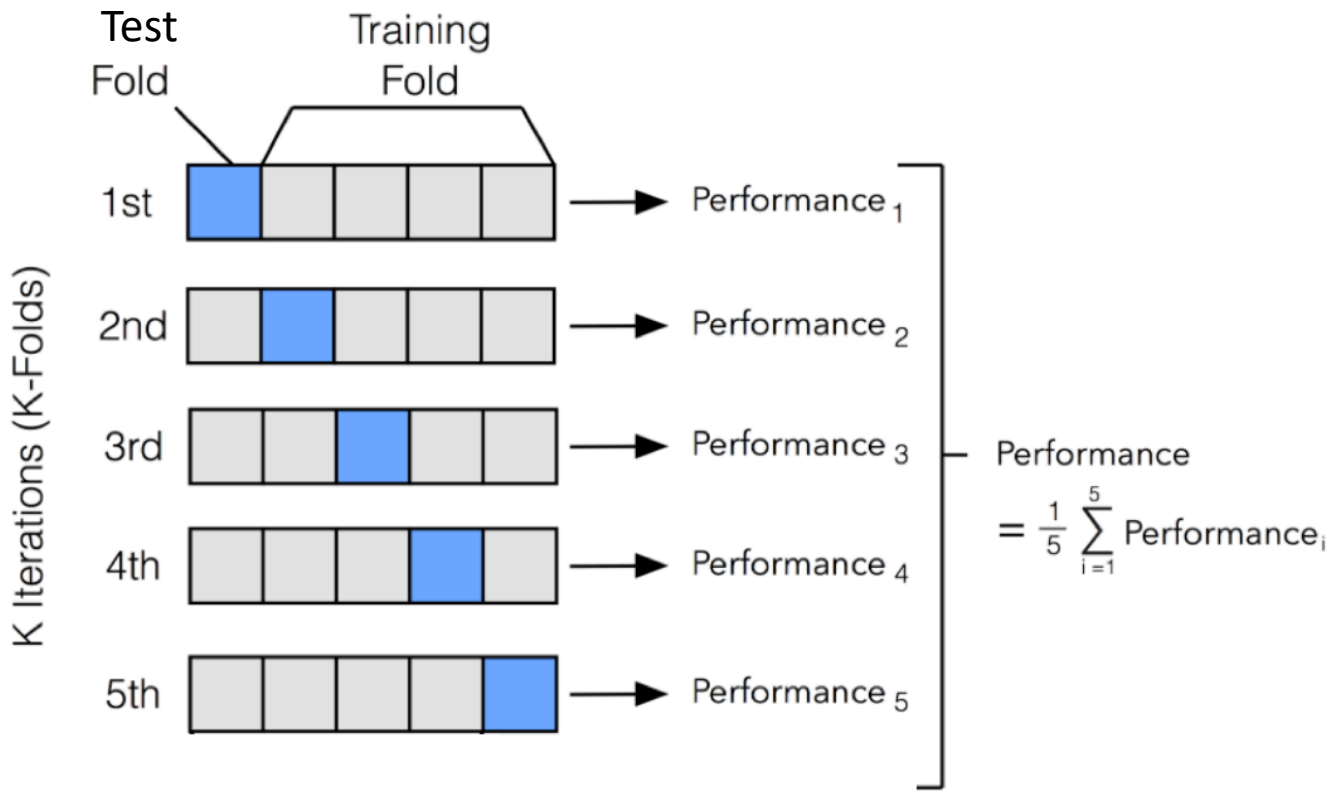


Image from here

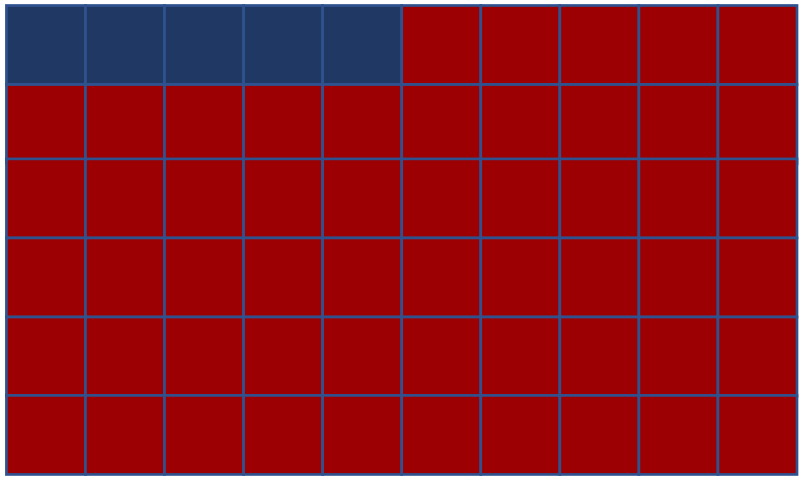
How do you measure performance?

CLASSIFICATION

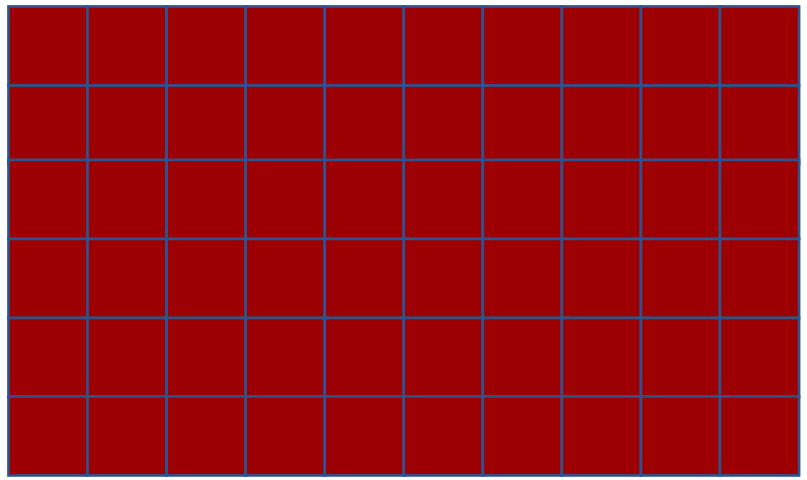
$$Accuracy = \frac{Number\ of\ corrected\ predictions}{Total\ number\ of\ predictions}$$

→ ✗ Not robust to **CLASS IMBALANCE (i.e. unequal number of samples in each class)**

Actual Labels



Predicted Labels



Accuracy:
*55/60 = **0.92***

! **High** accuracy **AND** **bad** model

Healthy

Alzheimer

How do you measure performance?

CLASSIFICATION

CONFUSION MATRIX → ✓ Robust to **CLASS IMBALANCE**

TP: 0

FP: 0

TN: 55

FN: 5

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

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

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

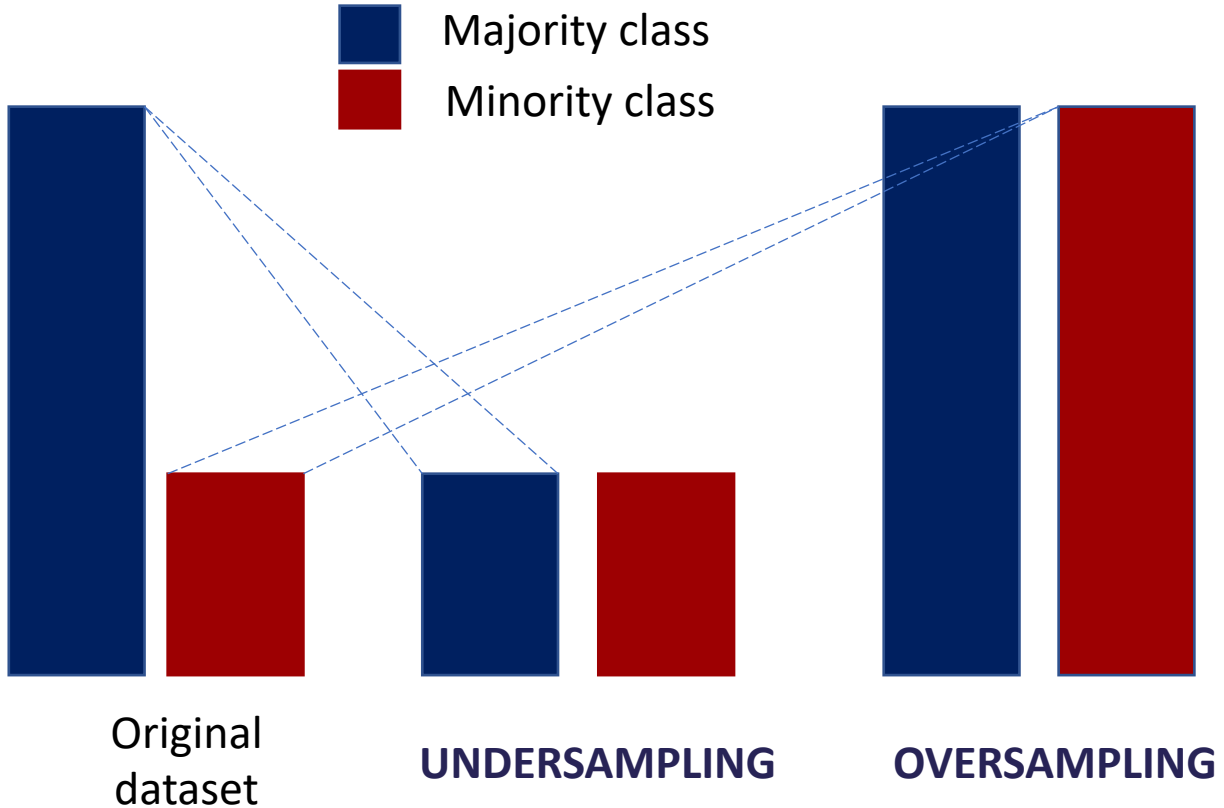
Actual Label

Predicted Label	Alzheimer	Healthy
Alzheimer	True positive	False positive
Healthy	False negative	True negative

Class imbalance

Class imbalance should be addressed DURING or BEFORE model training

Model will learn to predict everything as **MAJORITY CLASS** (i.e. class with most samples)



BEFORE

Model **PENALISATION**

Influence the training of the model so that the model makes fewer mistakes with the minority class

DURING

How do you measure performance?

REGRESSION

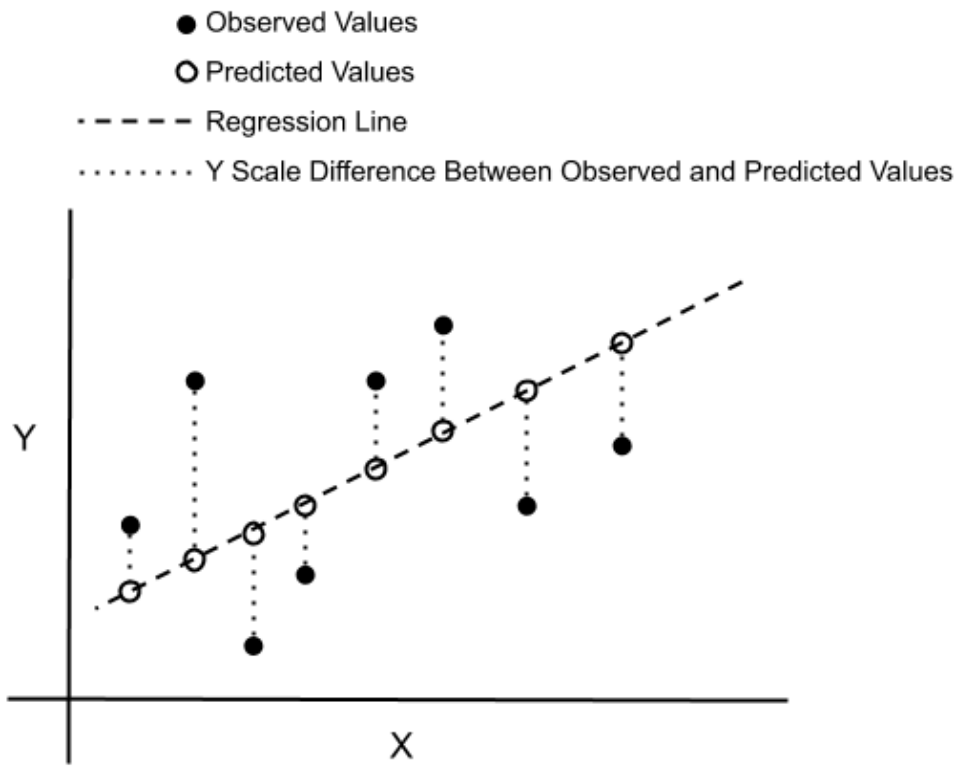


Image from here

! We cannot calculate accuracy in case of regression models

MEAN SQUARED ERROR: measures how close a fitted regression line is to the actual data

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

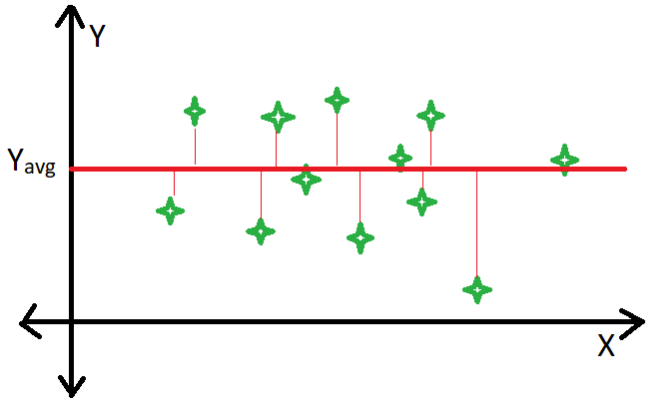
n = number of samples
Y_i = observed values
$\hat{Y}_i$ =predicted values
MSE = mean squared error

➡ Hard to interpret on its own

How do you measure performance?

REGRESSION

$$SS_{total} = \sum_{i=1}^{n} (Y_i - Y_{mean})^2$$



$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

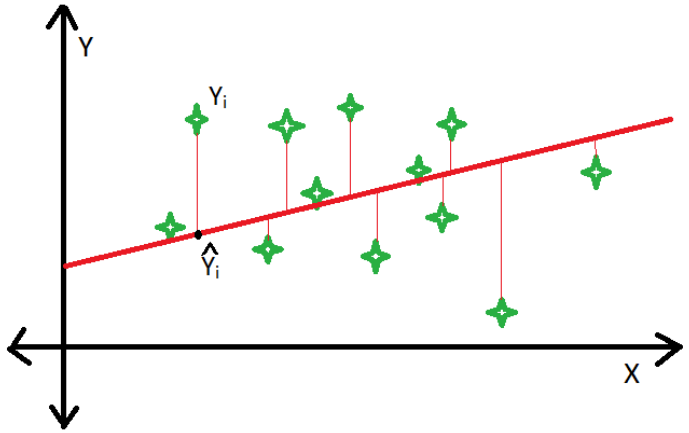


Image from here

R2 (or Coefficient of determination): the R2 represents how better your model is compared to the mean model

➡ Describes how much of the total variation in your data is explained by the independent variables

$$R2 = 1 - \frac{SS_{res}}{SS_{total}}$$

R2 = R squared
SS_{res} = residual sum of squares
SS_{tot} = total sum of squares

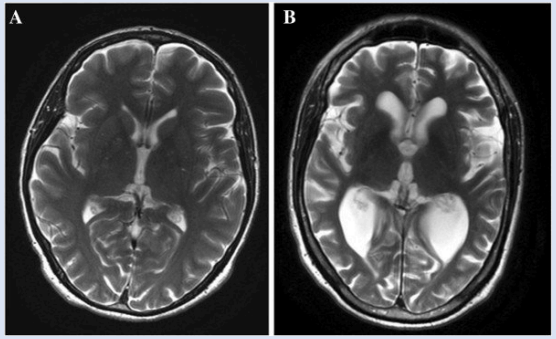
R1 = 1 ➡ Perfect fit

What is Machine Learning?

Three main **types** of Machine Learning

Supervised

ML approach that uses labeled data to learn and predict outcomes

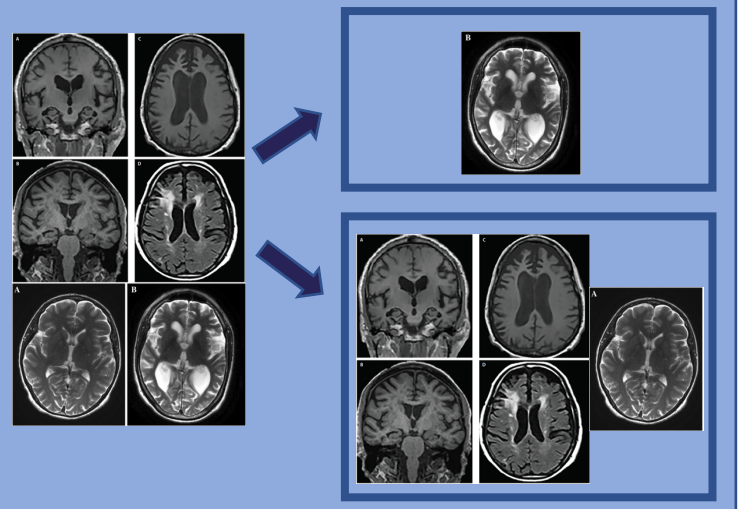


Alzheimer Healthy

Task driven

Unsupervised

ML approach that uses unlabeled data to discover patterns



Data driven

Reinforcement Learning

ML approach that allows an AI-driven system to learn through trial and error using feedback from its actions

Beyond the scope of this lecture

Learning from mistakes

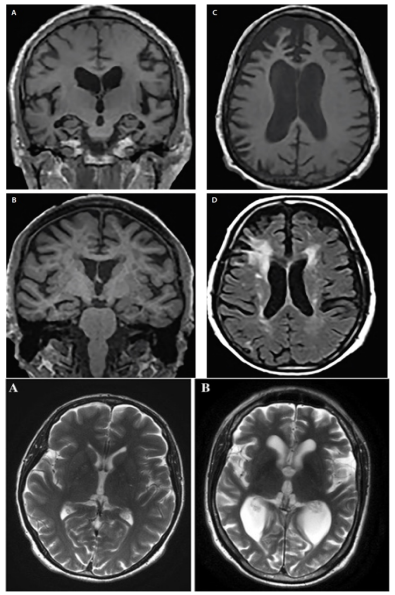
Cornelius et al, 2013

Ayers et al, 2019

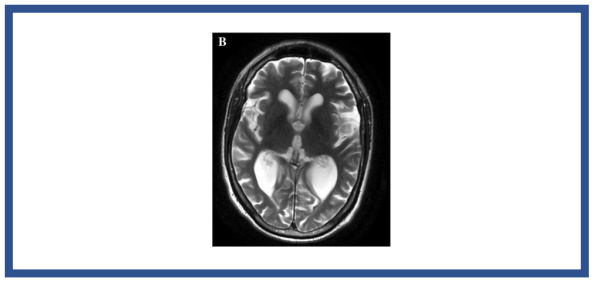
Unsupervised Learning

Unsupervised learning is commonly used for **CLUSTERING** and DIMENSIONALITY REDUCTION

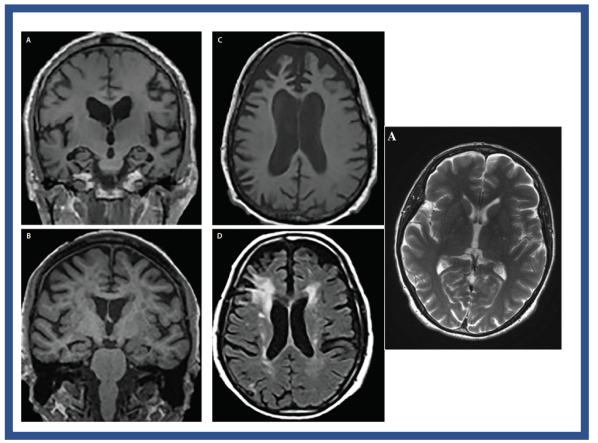
Input data



ML algorithm



Cluster 1 (Healthy?)



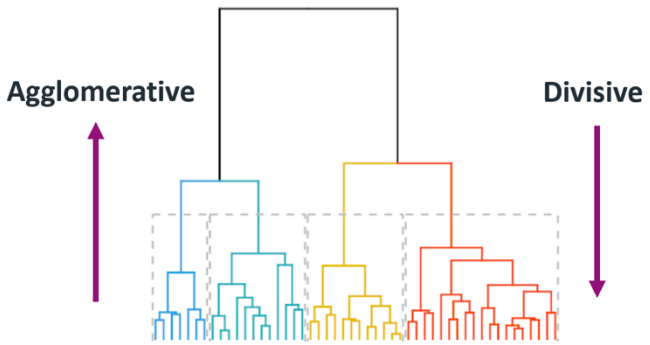
Cluster 2 (Dementia?)

Clustering: grouping of unlabeled data – **similar** data / data points are grouped together

Unsupervised Learning: types of clustering

CONNECTIVITY-BASED

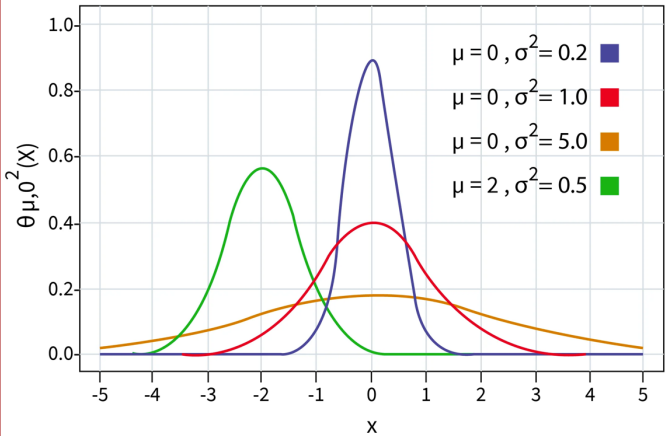
Clusters are created by building a hierarchical tree-type structure



HIERARCHICAL CLUSTERING

DISTRIBUTION-BASED

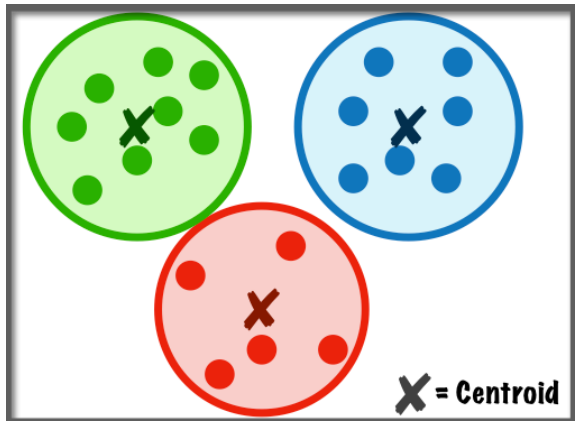
Data points are grouped together based on their probability of belonging to the same probability distribution



GAUSSIAN MIXTURE MODELS

CENTROID-BASED

Similairty is defined based on the distance between the data points and the centroid/center of the cluster



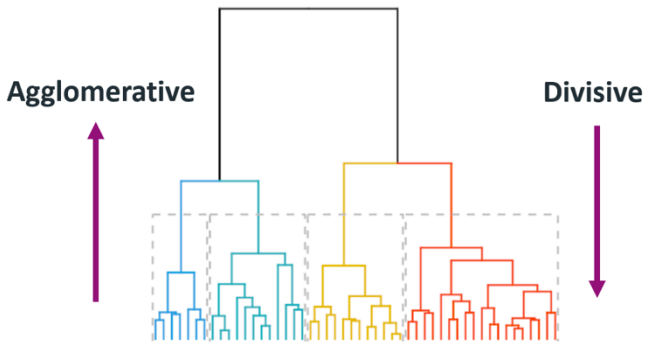
K-MEANS CLUSTERING

This list includes only some of the most common types, but not all of them!

Unsupervised Learning: types of clustering

CONNECTIVITY-BASED

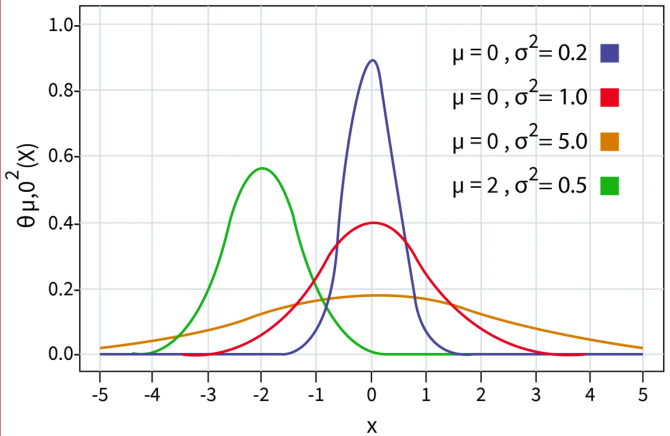
Clusters are created by building a hierarchical tree-type structure



HIERARCHICAL CLUSTERING

DISTRIBUTION-BASED

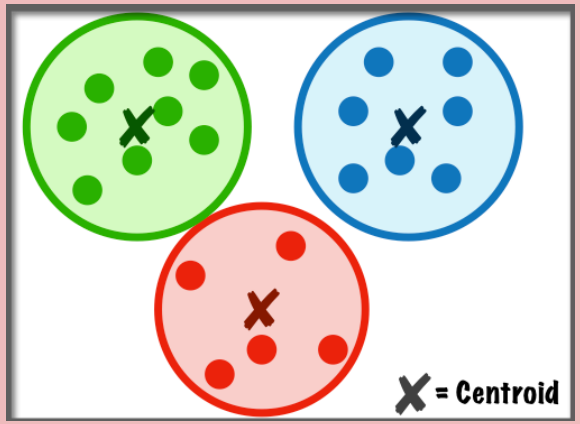
Data points are grouped together based on their probability of belonging to the same probability distribution



GAUSSIAN MIXTURE MODELS

CENTROID-BASED

Similairty is defined based on the distance between the data points and the centroid/center of the cluster



K-MEANS CLUSTERING

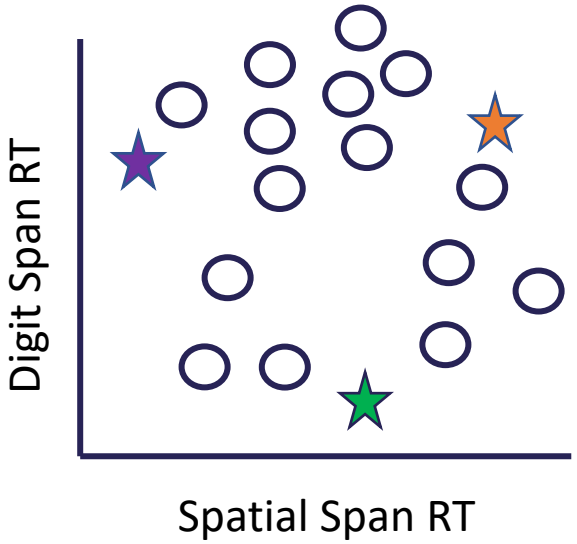
This list includes only some of the most common types, but not all of them!

Centroid-based clustering: k-means

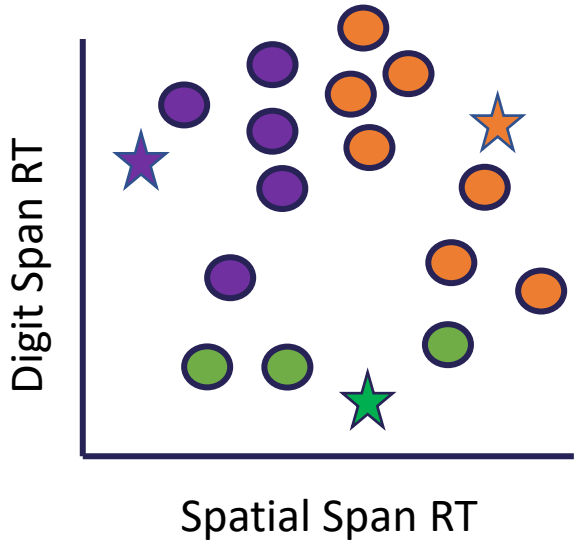
K-means clustering aims to minimise the total intra-cluster variation (or within cluster sum of squares)

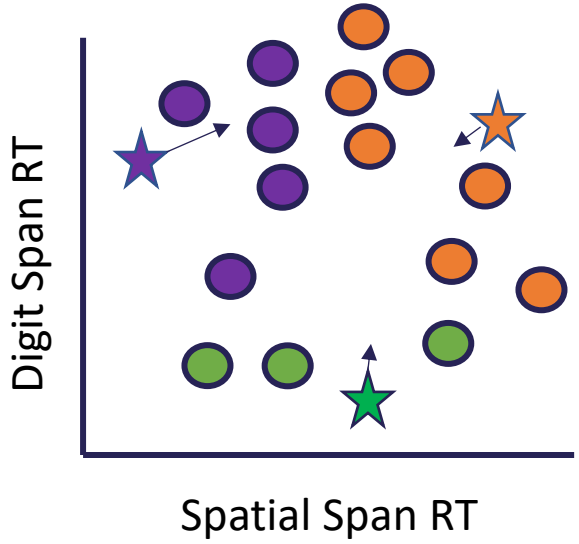
! The number of clusters is pre-defined

Initial centroids



Iteration 1



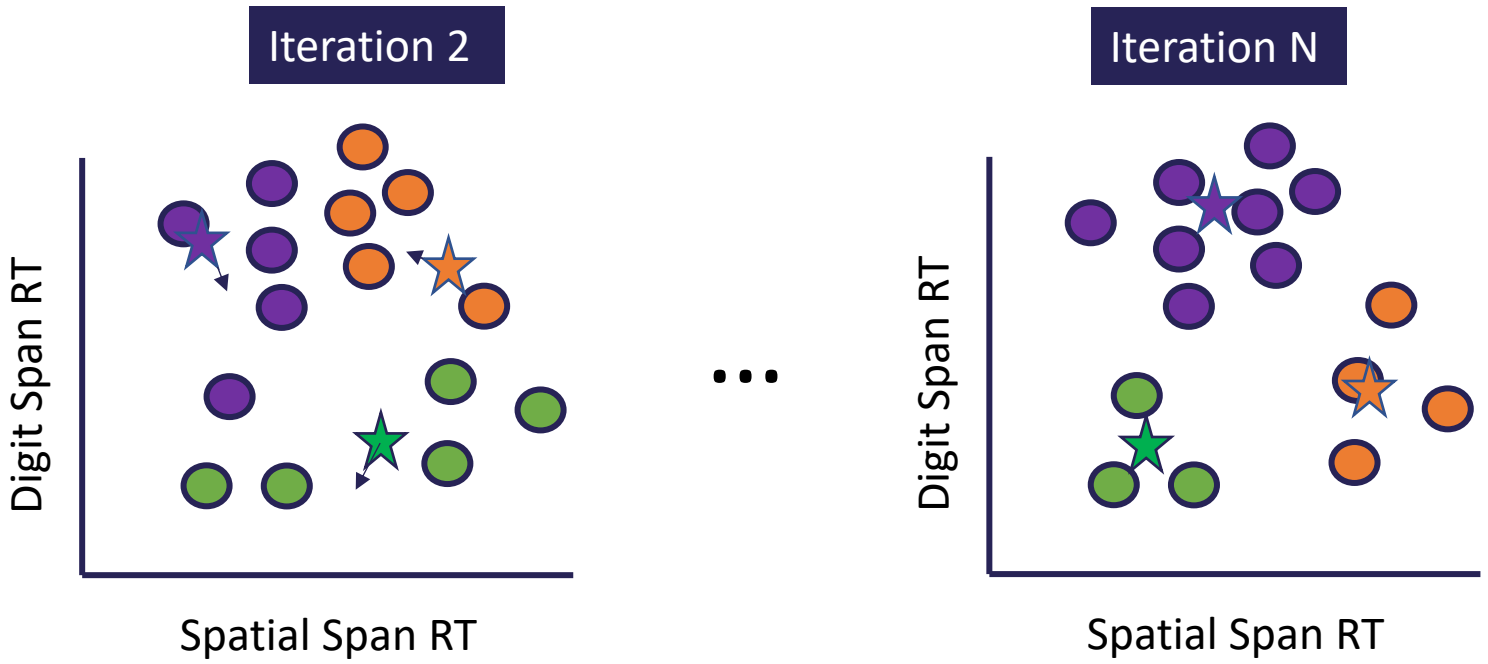


Centroids are randomly selected → Data points are grouped with the cluster of the closest centroid → New centroids are located at the cluster means

Centroid-based clustering: k-means

K-means clustering aims to minimise the total intra-cluster variation (or within cluster sum of squares)

The number of clusters is pre-defined

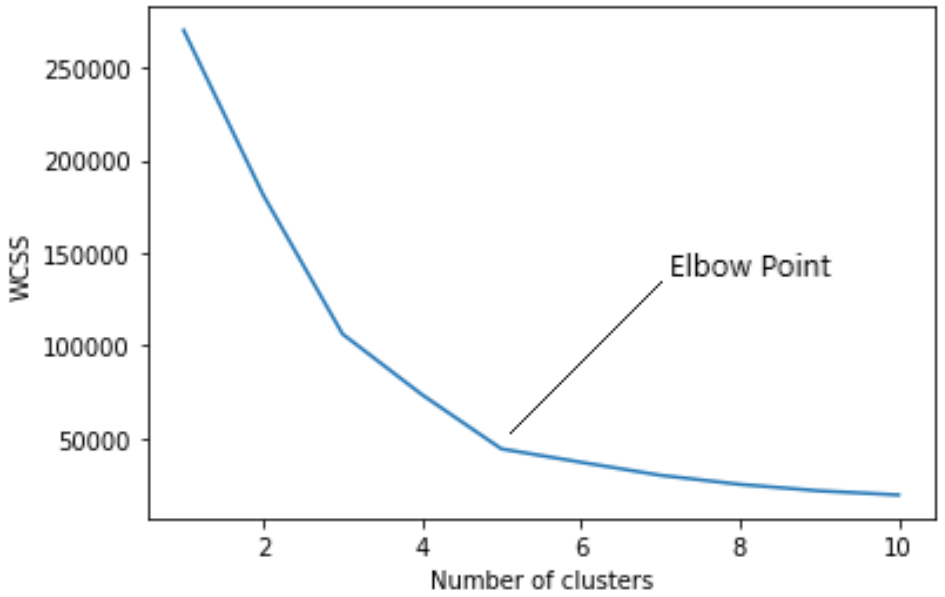


How do select the optimal number of clusters?

The iterations are continued until convergence

The data points don't change cluster anymore

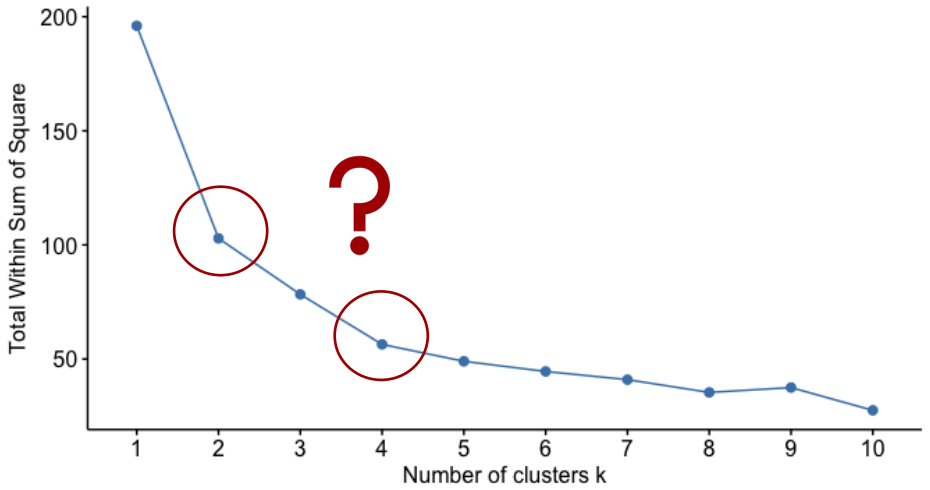
k-means: elbow method



WCSS = within-cluster sum of squares

Elbow point = point in the curve where the knee is located –> WCSS falls suddenly

The elbow is very subjective and often hard to identify

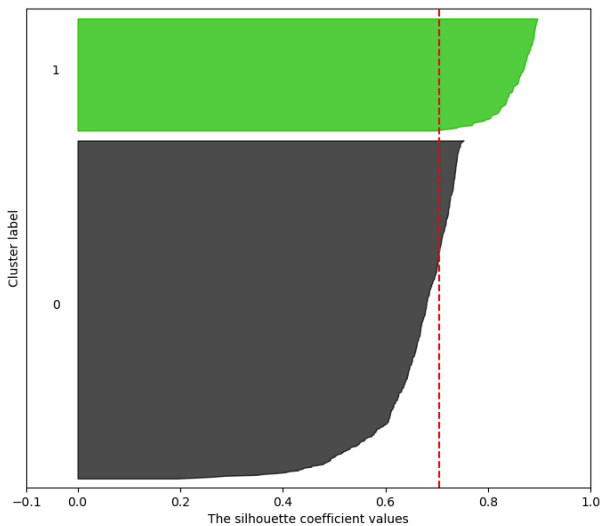


k-means: Silhouette method

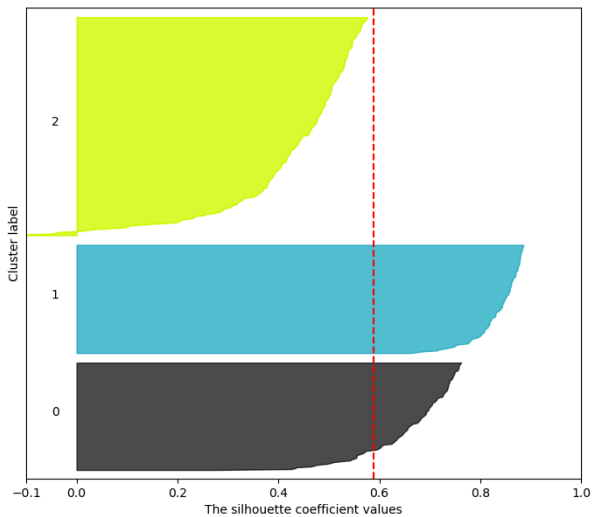
Silhouette score: measure of how close each data point in one cluster is to the data points of the neighboring cluster

- S(i) = 1 Point i is far from neighboring cluster
- S(i) = 0 Point i is at the border
- S(i) = -1 Point i is in the wrong cluster

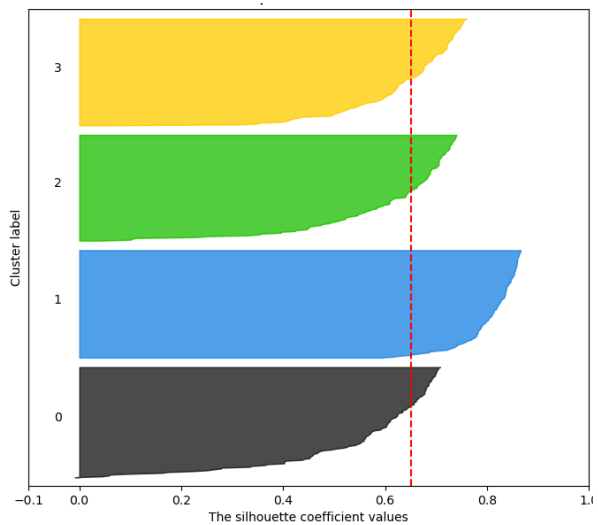
S(average) = 0.70



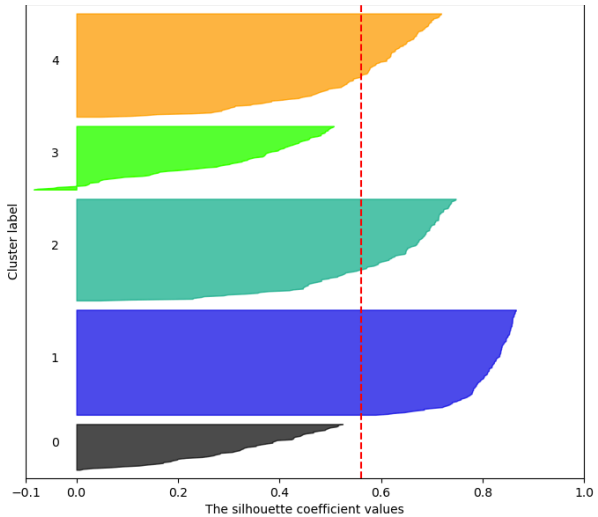
S(average) = 0.59



S(average) = 0.65



S(average) = 0.56



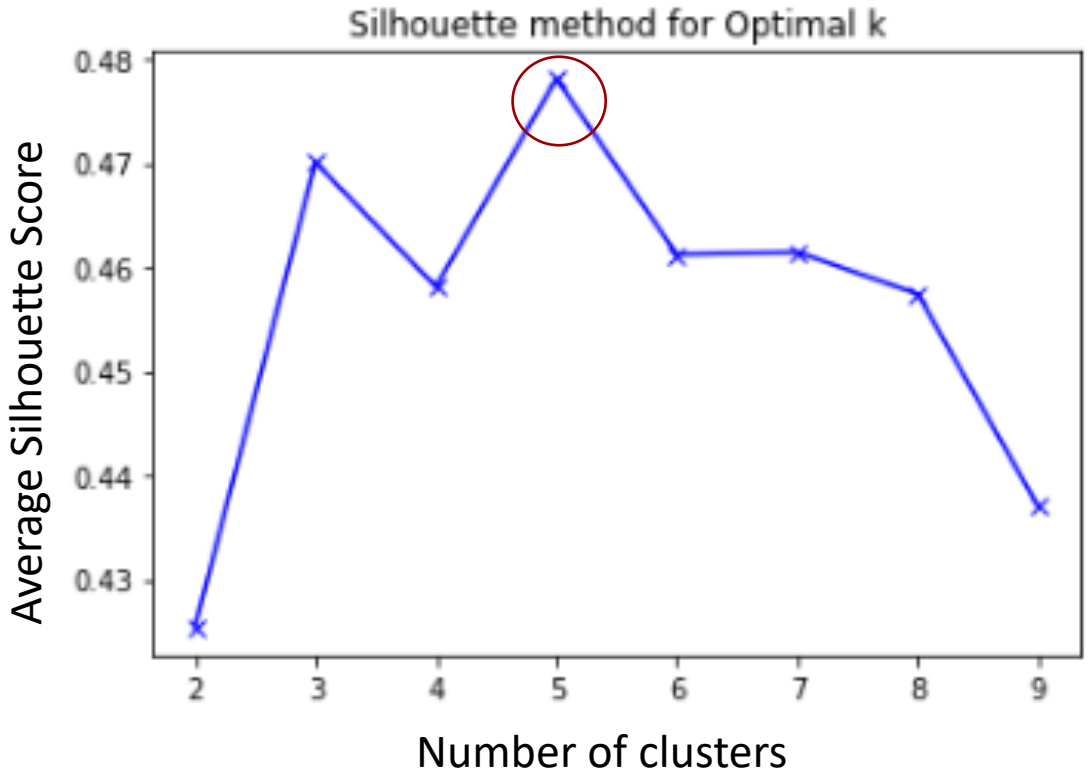
Silhouette scores

Figures from sklearn tutorial

k-means: Silhouette method

Silhouette score: measure of how close each data point in one cluster is to the data points of the neighboring cluster

S(i) = 1	Point i is far from neighboring cluster
S(i) = 0	Point i is at the border
S(i) = -1	Point i is in the wrong cluster



The optimal K has the highest average silhouette score

The silhouette score is more objective!

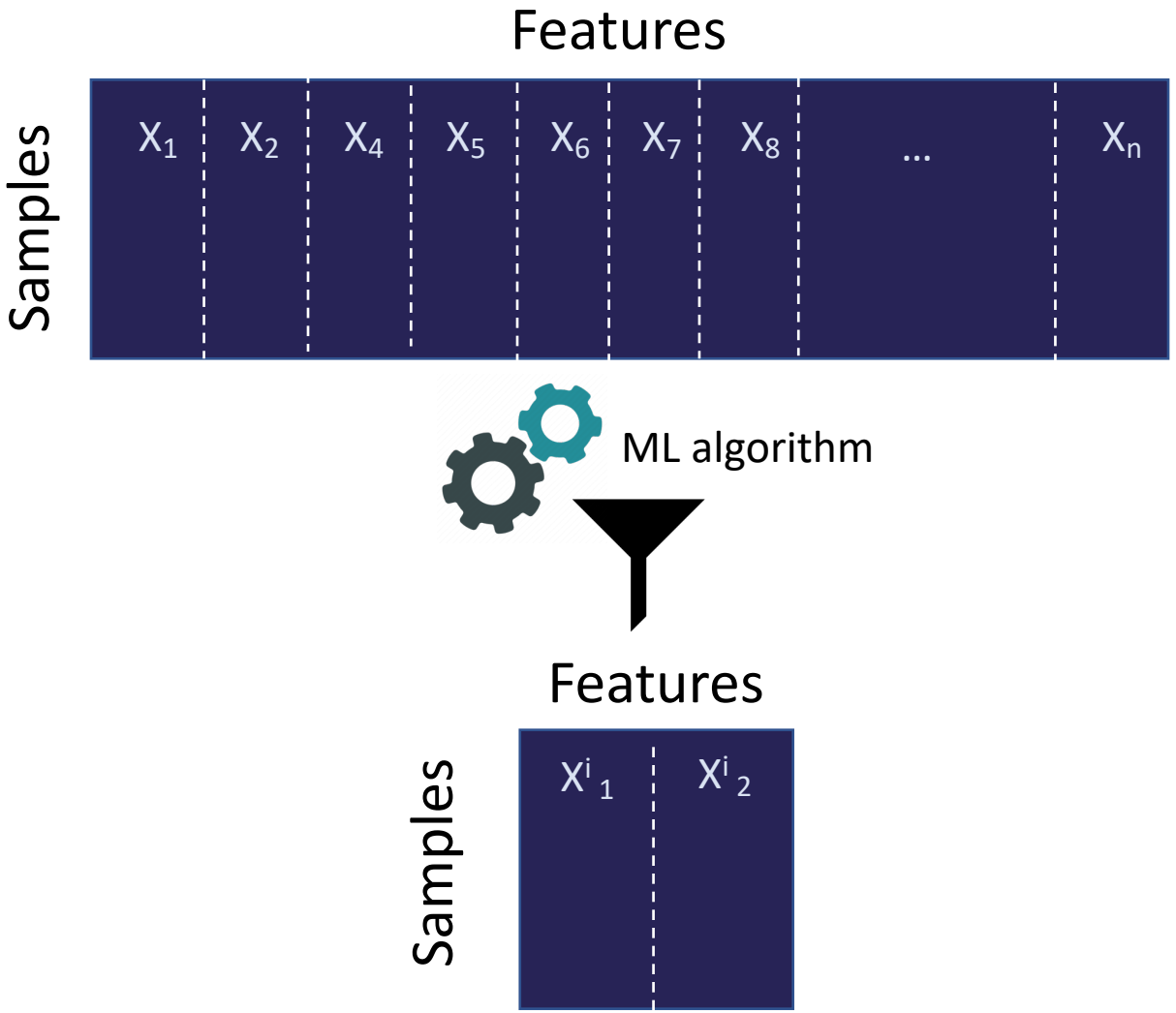
K-means vs k-modes

What if the data are categorical?

K-MODES CLUSTERING

Centroids are random points	⟷	Centroids are random data points among those available
Measure of **distance**	⟷	Measure of **dissimilarity** (categories in common with the centroids)
At each iteration, the centroids are redefined based on the **mean**	⟷	At each iteration, the centroids are redefined based on the **mode**

Unsupervised learning is commonly used for CLUSTERING and **DIMENSIONALITY REDUCTION**



WHY?

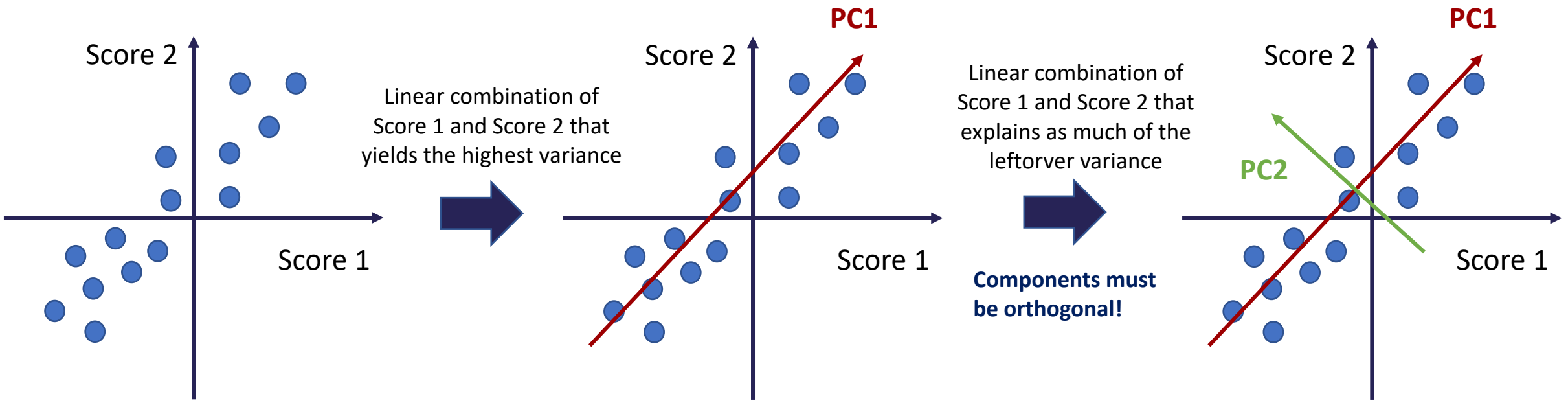
1. Visualisation
2. Decrease model complexity and prevent overfitting
3. Feature selection

! Feature selection differs from DR because it simply selects a subset of features without changing them

Dimensionality Reduction: transformation of data from high to low dimensional space

Principal Component Analysis

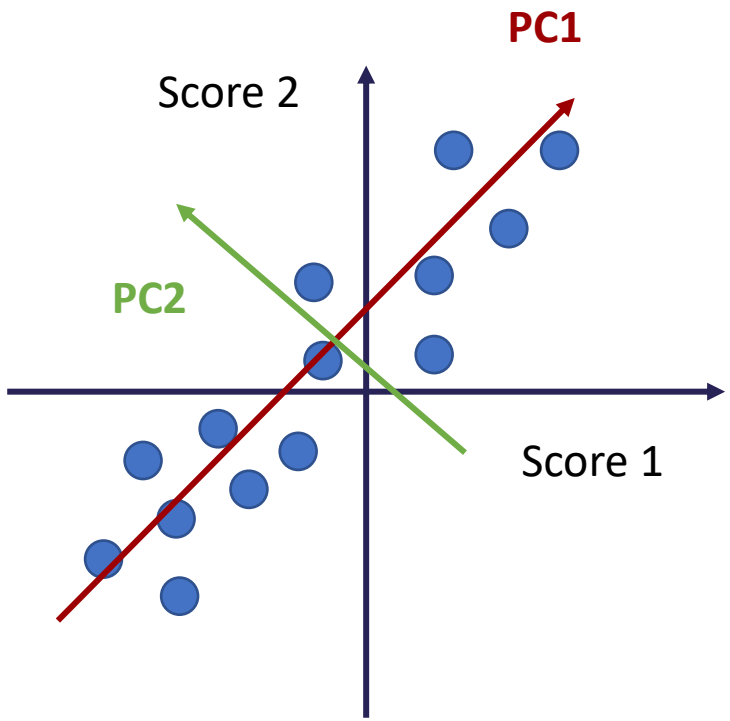
Principal Component Analysis (PCA) is a common unsupervised ML technique for dimensionality reduction that transforms a dataset by creating new dimensions that try to explain as much variance in the data as possible.



Let's imagine data with only 2 features (e.g. scores in 2 cognitive tests)

PC1 and PC2 are the 2 new lower dimensional features → You can have as many components as features

How do you interpret the principal components?



NOT STRAIGHTFORWARD

Pincipal components are **linear combinations** of the original features

$$PC1 = \boxed{1.5} * Score1 + \boxed{2} * Score2$$

Loadings

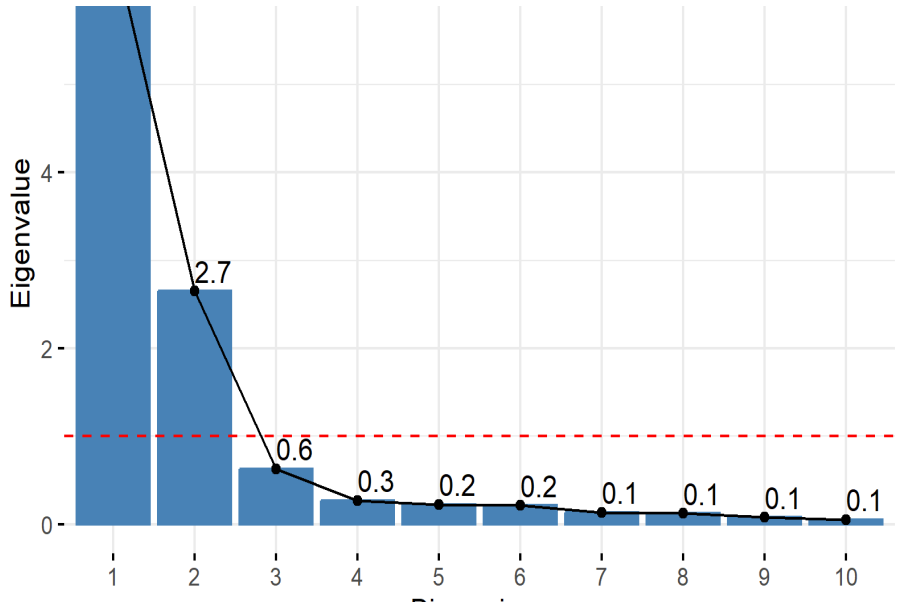
Features with higher loadings contributed the most to define the PC

➡ Can be potentially used for **FEATURE SELECTION**

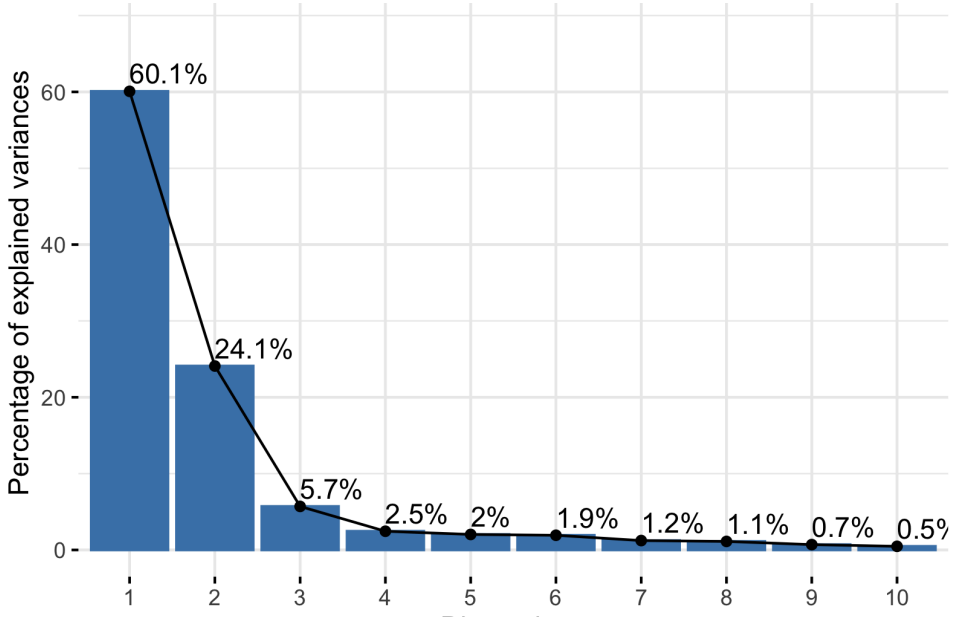
Principal Component Analysis

How do you select the optimal number of components?

Kaiser's criterion



Elbow method



Imperial College London