NBE-4070 : Basics of Biomedical Data Analysis

Stéphane Deny

Lecture 9: Recap of Important Definitions for the Oral Exam

What is a good explanation of a data analysis method for the oral exam?

Guiding principle: It's an explanation that you would have liked and understood before taking this class.

Tips:

- use an example
- draw a sketch (equations are nice to have but rarely necessary)
- make it clear what the method can and cannot do
- abstract the unnecessary details at first
- leave room for questions – ask if it is clear

Exercise 1: Explain what is:

- standard deviation (SD),
- standard error of the mean (SE) and confidence interval (CI)
- t-test and p-value
- Pearson Correlation

Exercise 2: Explain what is:

- PCA
- Linear Regression
- Fourier Transform and Low-pass filtering
- Clustering and one algorithm for clustering (K-means or hierarchical)

Exercise 3: Explain what it:

- K-nearest neighbour
- Random Forest
- t-SNE visualisation
- a deep network classifier

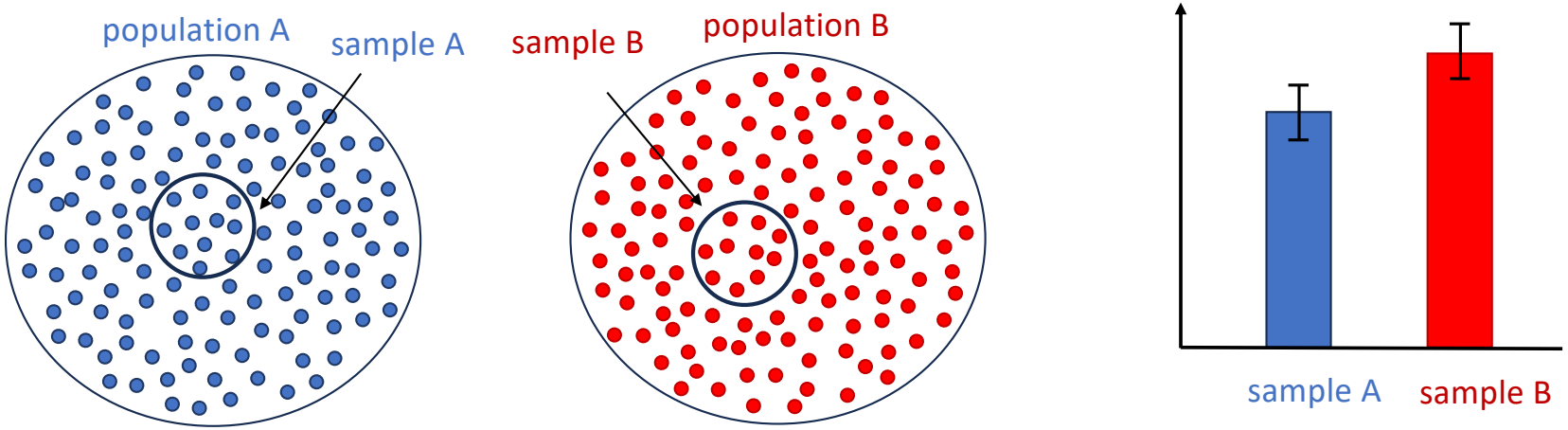
Definitions: standard deviation (SD), standard error of the mean (SE) and confidence interval (CI)

Table I. **Common error bars**

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$SD = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$
Standard error (SE)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$SE = SD/\sqrt{n}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{(n-1)} \times SE$, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times SE$.

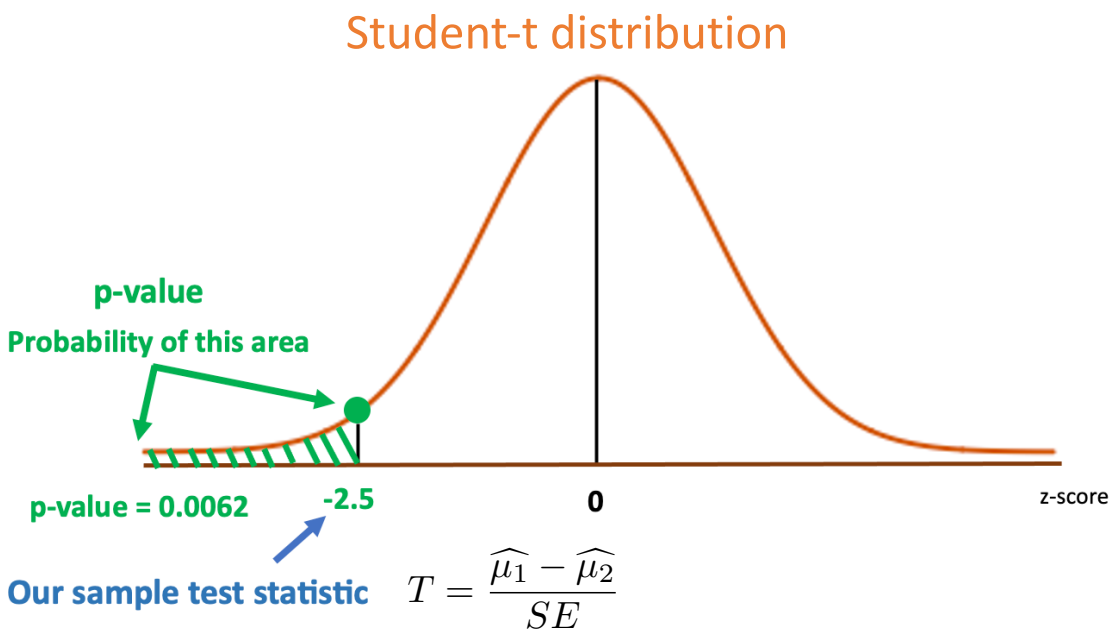
Definition: t-test

- A *t-test* is a statistical test to compare the means of two populations from their samples.



Definition: p-value

- The *p-value* is the probability that the difference in means observed would occur by chance.
  example: A p-value of 0.05 implies a chance of 95% that the difference observed is statistically significant.



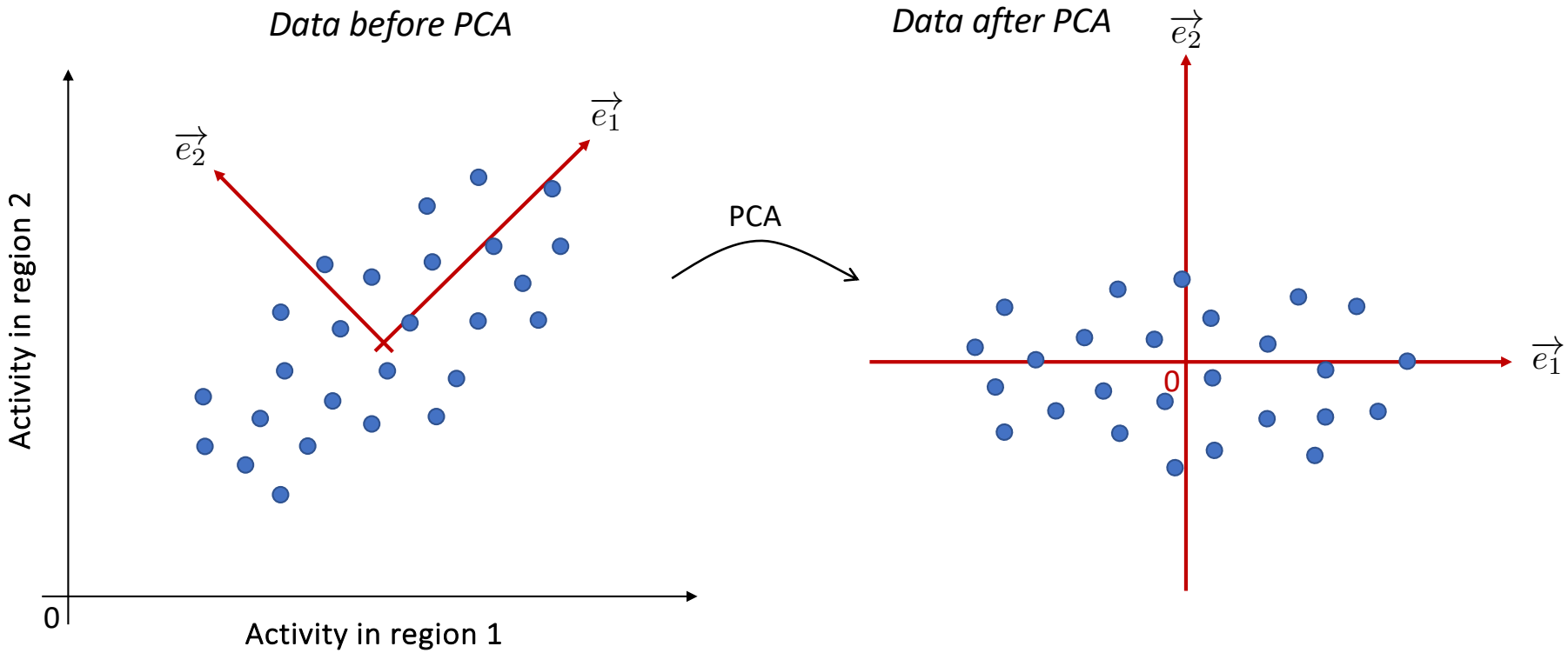
Definition: Pearson correlation

- Given two paired variables $X = \{x_1, x_2, ..., x_N\}$
  $Y = \{y_1, y_2, ..., y_N\}$
- Their *(Pearson) correlation coefficient* is given by

$$R = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{(\sum_{i=1}^{N}(x_i - \mu_X)^2)}\sqrt{(\sum_{i=1}^{N}(y_i - \mu_Y)^2)}}$$

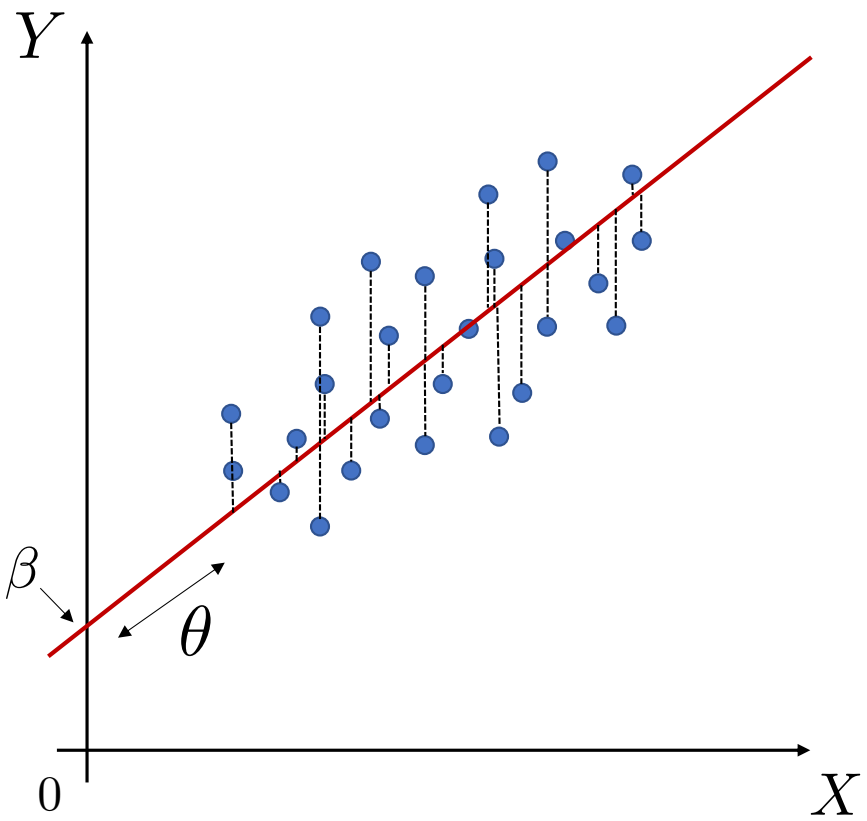
- It is a normalized version of the covariance, such that the result always has a value between −1 and 1. There exists tests to decide whether the correlation is statistically significant or not.

PCA is a change of basis “aligned with the data”



Formal: PCA is an orthornomal change of basis such that the covariance of the data in that new basis is diagonal.

Linear regression : definition



Linear regression predicts the value of an output variable Y based on the value of an input variable X, assuming a linear relationship between X and Y:

$$\widehat{y_i} = \theta * x_i + \beta$$

where x_i is the **input** variable for sample *i*

y_i is the **output** variable for sample *i*

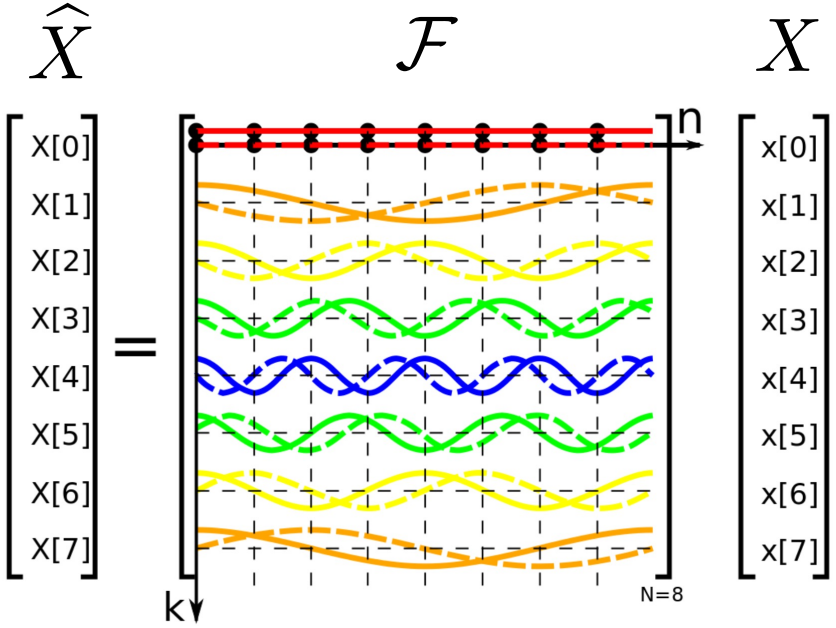
$\widehat{y_i}$ is the **prediction** for sample *i*

β is the **intercept** of the linear fit

θ is the **slope** of the linear fit

Discrete Fourier Transform (DFT) : definition

DFT is a change of basis, which re-expresses an input sequence X into a new basis of sine and cosine functions (i.e. *frequency domain*):



The real part (cosine wave) is denoted by a solid line, and the imaginary part (sine wave) by a dashed line.

where $\hat{X}$ is the Fourier representation

$\mathcal{F}$ is the Fourier transform

X is the input sequence

Low-pass, high-pass and band-pass filters

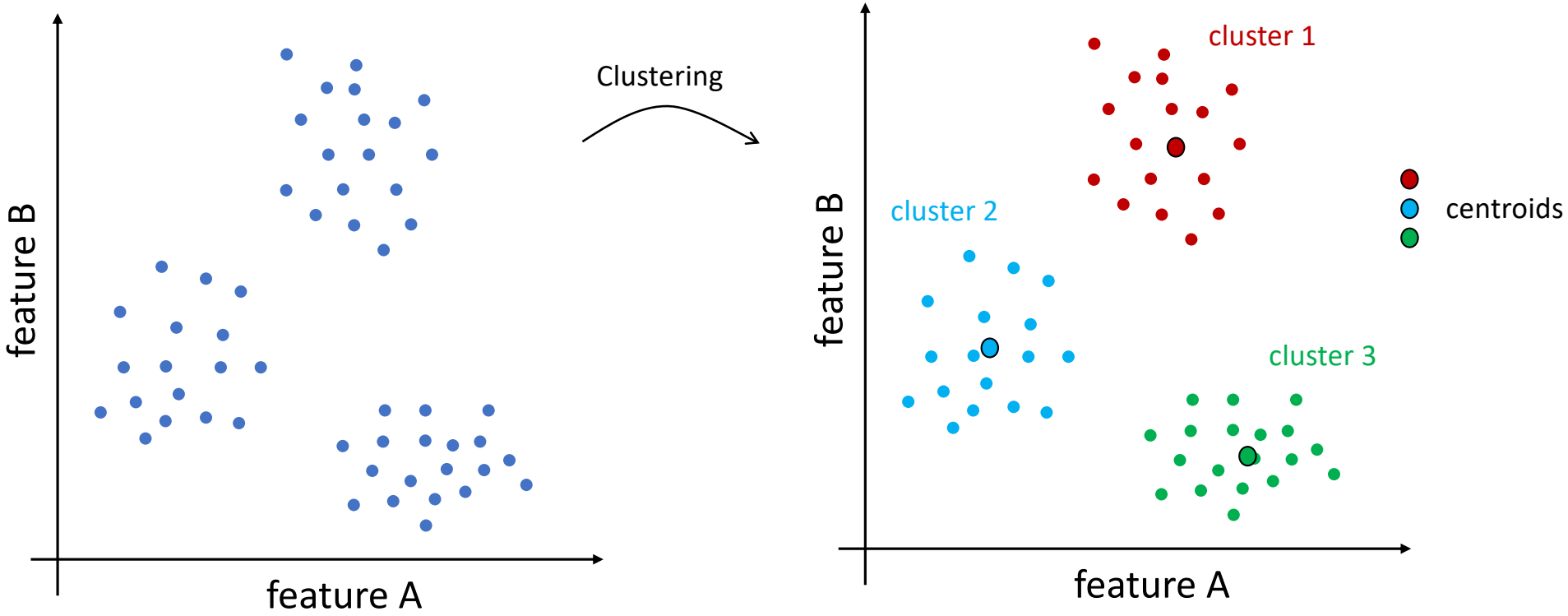
- A low-pass filter is a filter that passes signals with a frequency lower than a selected cutoff frequency and attenuates signals with frequencies higher than the cutoff frequency.

- A high-pass filter is a filter that passes signals with a frequency higher than a certain cutoff frequency and attenuates signals with frequencies lower than the cutoff frequency.

- A band-pass filter is the combination of a high-pass and a low-pass filter

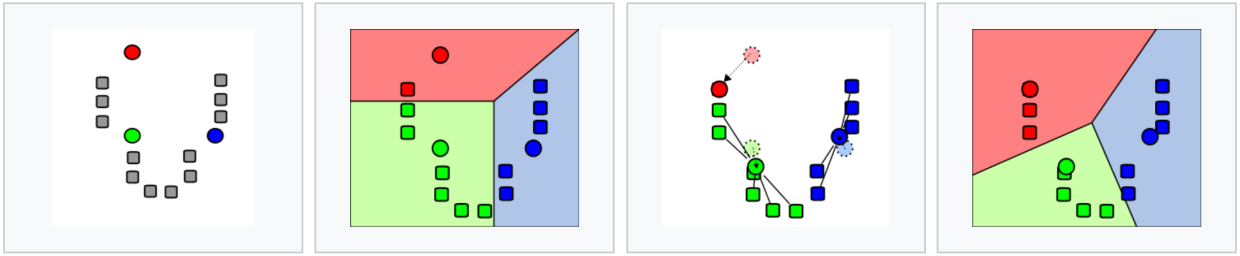
Definition of clustering

Clustering is grouping a collection of data points into subsets or “clusters”, such that the points within each cluster are closer to one another than points assigned to different clusters.



K-means clustering algorithm

K-means clustering starts with guesses for the 'K' cluster centers. Then it alternates the following steps until convergence:
1) for each data point, the closest cluster center (in Euclidean distance) is identified;
2) each cluster center is replaced by the coordinate-wise average of all data points that are closest to it (i.e., center of mass).



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

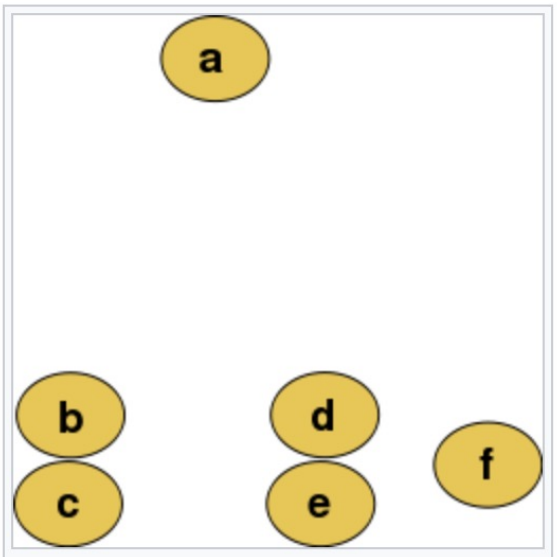
4. Steps 2 and 3 are repeated until convergence has been reached.

source: https://en.wikipedia.org/wiki/K-means_clustering

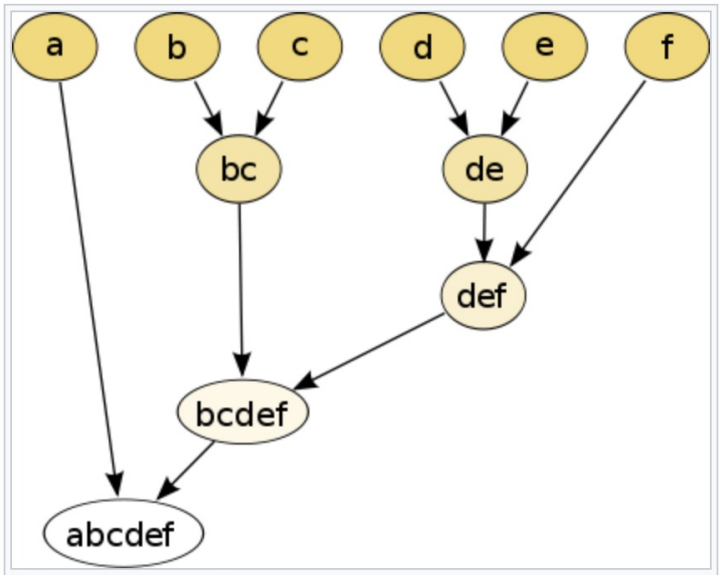
Hierarchical clustering algorithm

Hierarchical clustering recursively merges a selected pair of clusters into a single cluster. This produces a grouping at the next higher level with one less cluster. The pair chosen for merging consists of the two groups with the smallest intergroup dissimilarity.

Raw data



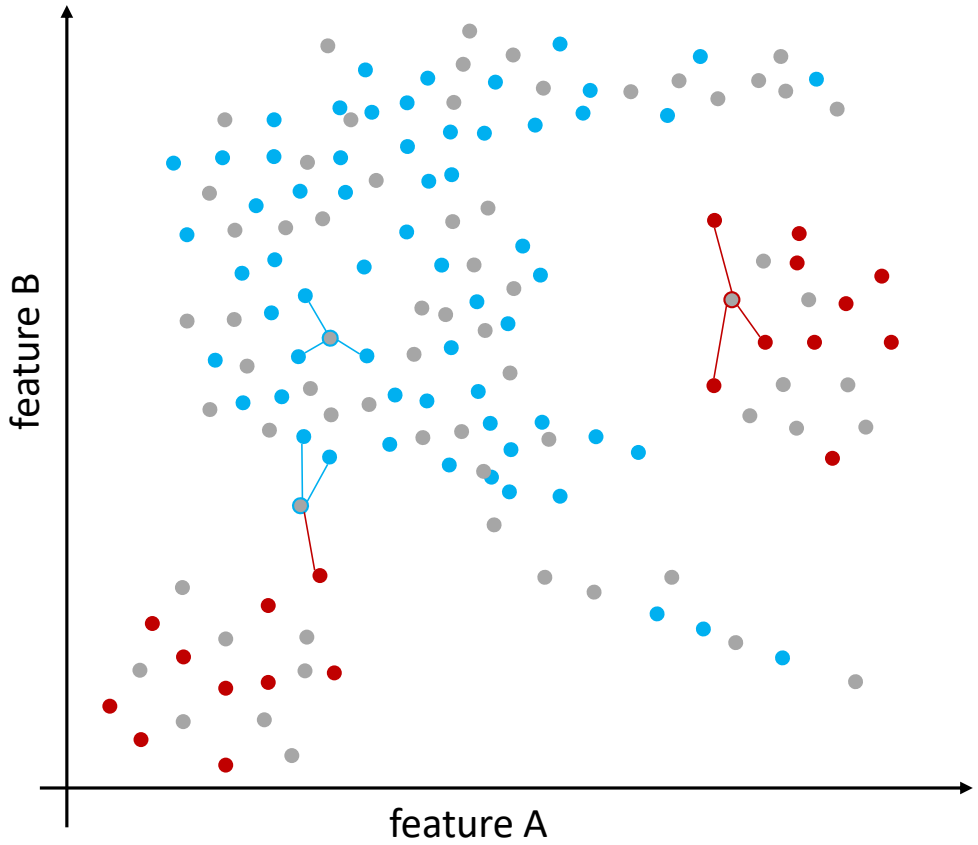
Hierarchical clusters



source: https://en.wikipedia.org/wiki/Hierarchical_clustering

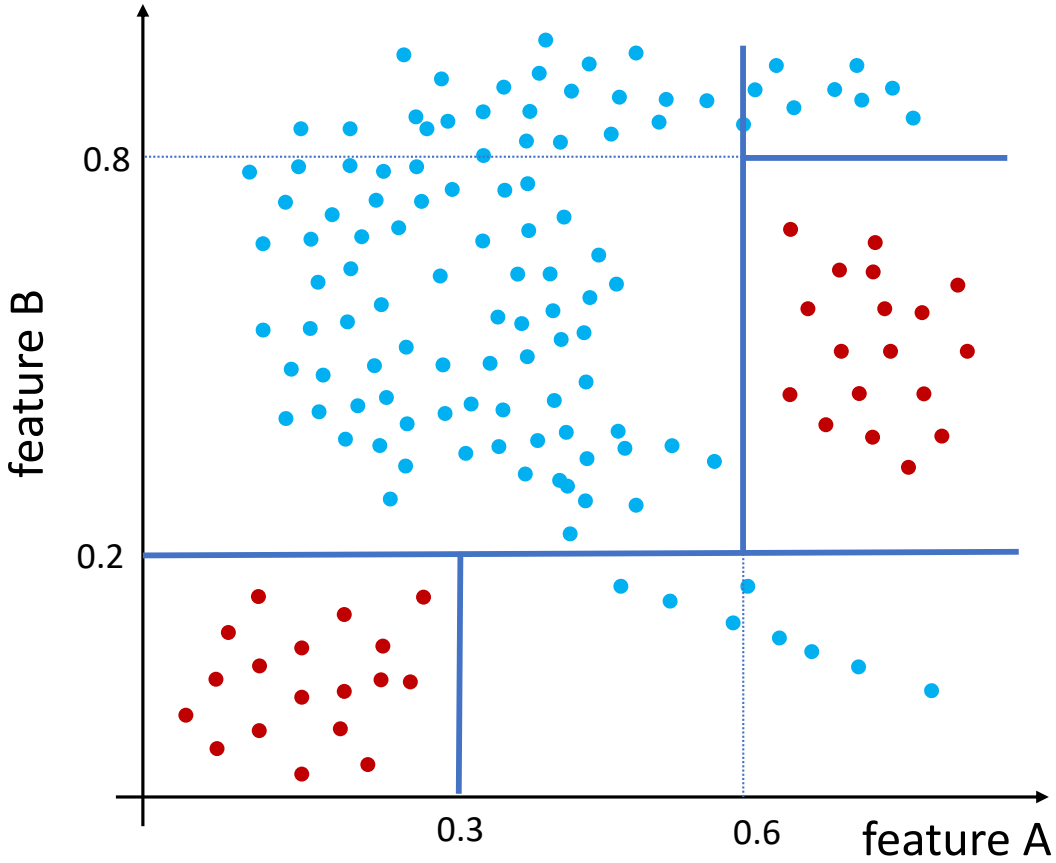
K-nearest neighbor algorithm: definition

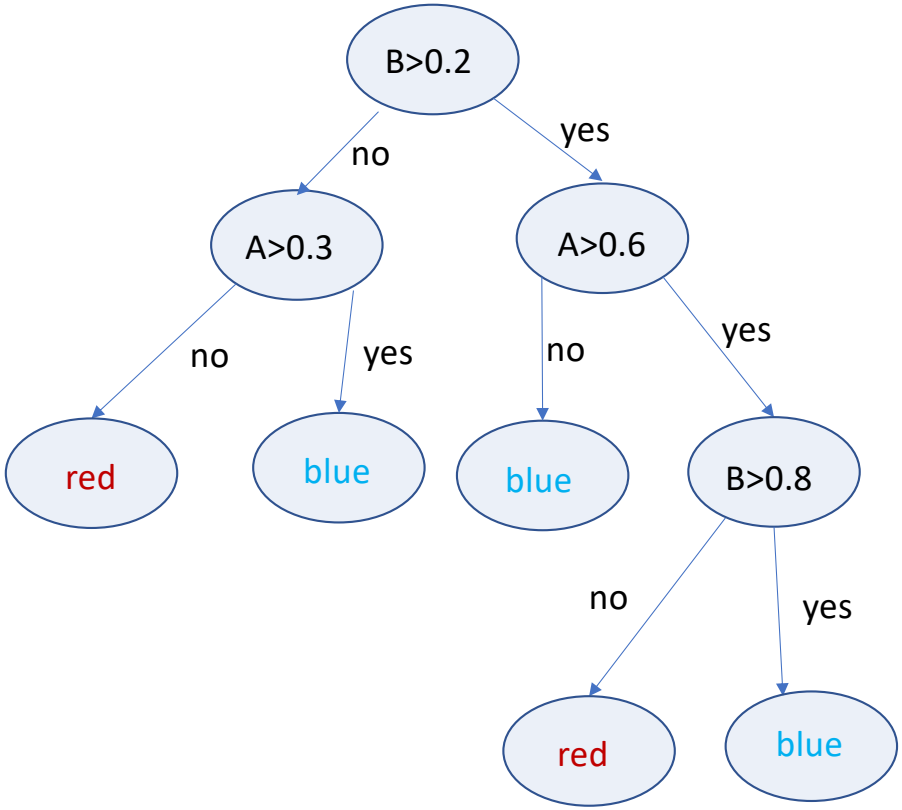
A data point is classified by a vote of its neighbors, with the point being assigned to the class most common among its ‘k’ nearest neighbors (‘k’ is a positive integer, typically small).



Decision tree: definition

A decision tree is built by splitting a dataset into subsets recursively. Each split operates on one feature of the dataset, and tries to maximize the separation of classes. The recursion is completed when the subset of points at any given node all have the same value.

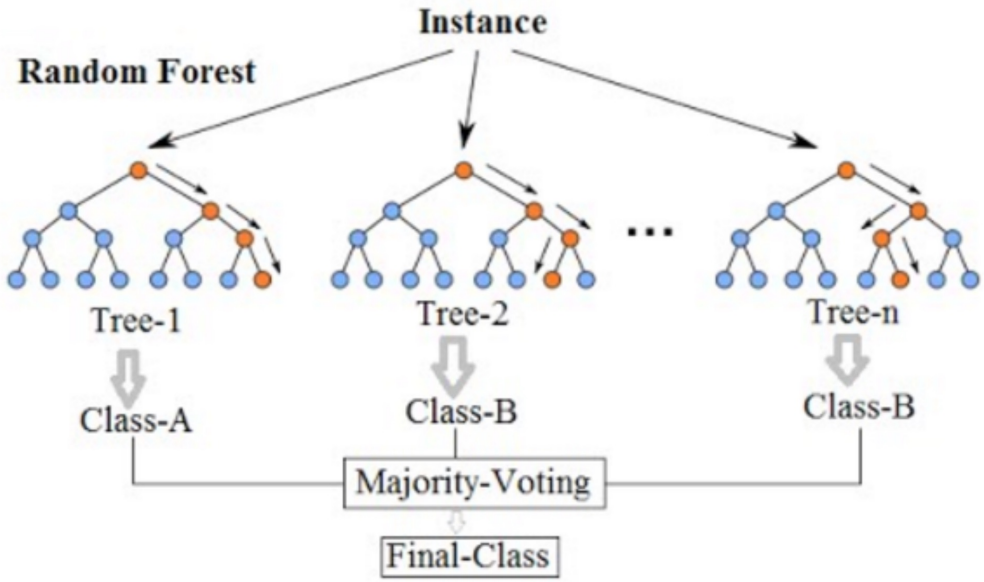




source: https://en.wikipedia.org/wiki/Decision_tree_learning

Random Forest: definition

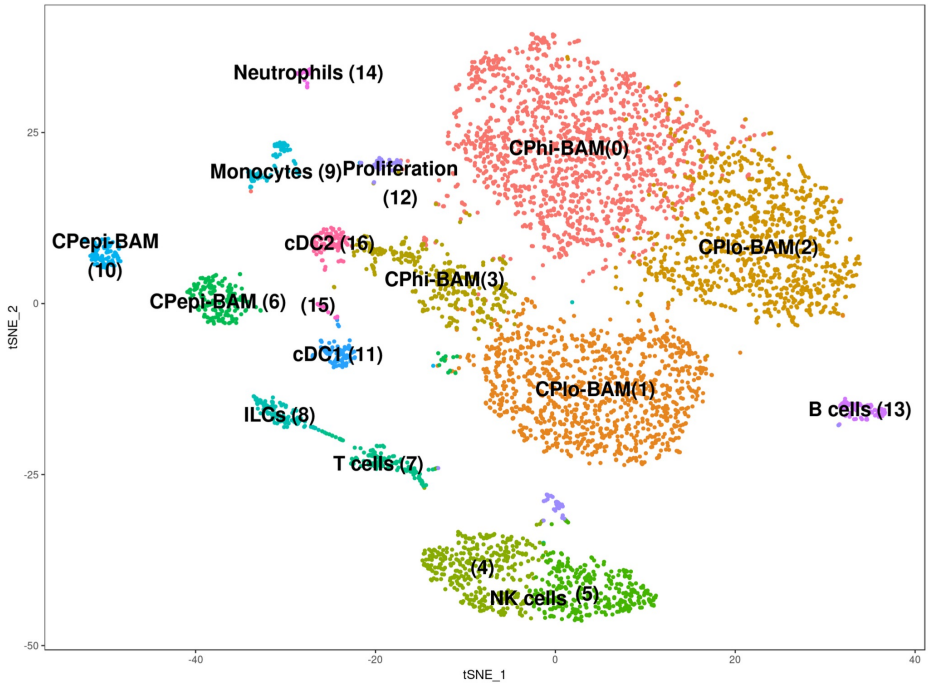
Decision trees are prone to overfitting. A random forest operates by constructing a multitude of decision trees at training time. The output of the random forest is the class selected by most trees. Random forests generally outperform decision trees. Random forest belongs to the class of “ensemble learning” algorithms.



source: https://en.wikipedia.org/wiki/Random_forest

t-SNE data visualization: definition

t-distributed stochastic neighbor embedding (t-SNE) is a method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.



Single-cell RNA sequencing datasets of brain immune cells, projected into 2 dimensions using t-SNE

source: https://www.brainimmuneatlas.org/index.php

Deep networks: definition

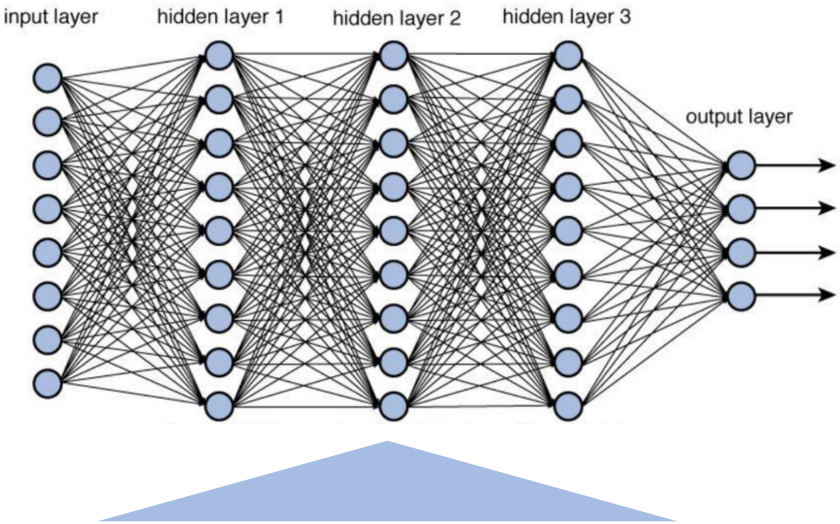
A deep neural network is an artificial neural network with multiple layers between the input and output layers. The network takes as an input a data sample ‘x’ and predicts as its output the data label ‘y’:

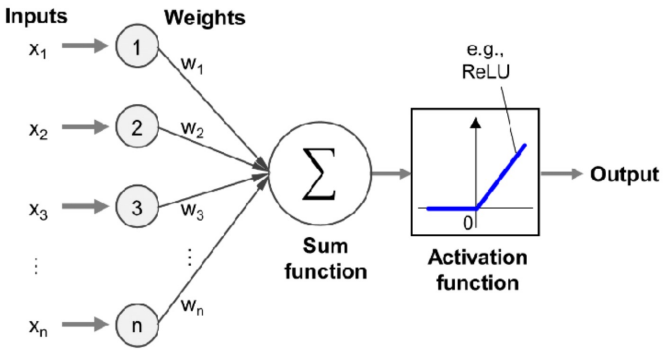
The network output is given by:

$$\widehat{y} = net(x)$$

where net() is composed of layers defined by:

$$a_i^l = ReLU\left[\sum_j w_{ij} a_j^{l-1}\right]$$





Function perfomed by one neuron