



Karolinska
Institutet

Machine learning in bioinformatics

An introduction to supervised/unsupervised
machine learning in transcriptomics

Short about me:

- Gustaw Eriksson (gustaw.eriksson@ki.se)
- B.Sc. in Molecular Biology & M.Sc. in Bioinformatics from Lund University.
- PhD student in the Reproductive Endocrinology and Metabolism group led by Elisabet Stener-Victorin
- Researching the etiology of PCOS
 - Single-nuclei RNA-sequencing and bioinformatics
 - Endometrium organoids
 - **Wet and dry lab**

Today's schedule:

Before lunch:

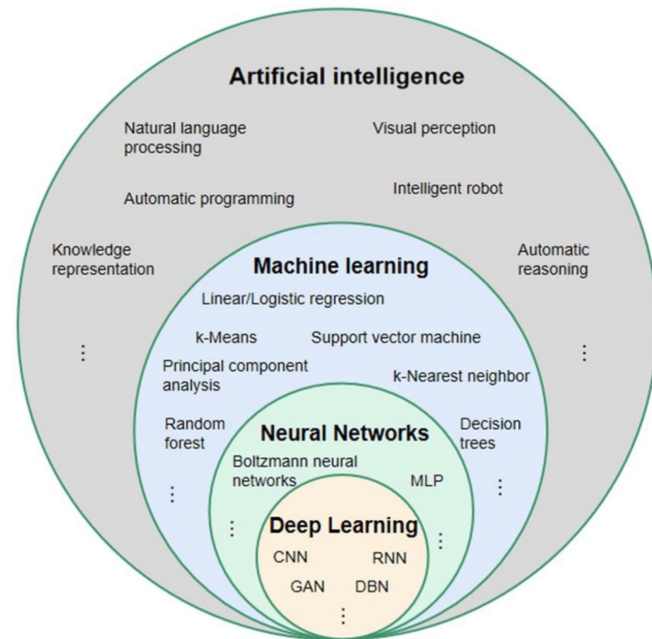
- Supervised Machine Learning (SML)
 - Linear regression, K-nearest neighbours (KNN), Random Forest
- Unsupervised Machine Learning (UML)
 - K-means clustering, Hierarchical clustering, Principal component analysis (PCA)

After lunch:

- Machine Learning (ML) in practice
- DESeq2 tutorial in R analysing bulk RNA-seq data
- Questions and discussion

AI, Machine Learning, Neural Network and Deep Learning. What's the difference?

- Machine learning (ML) is a subfield of AI, or a path to AI
 - Algorithms to learn insights and recognise patterns from data
 - Deep Learning and Neural Networks are methods of ML
 - Deep Learning structures algorithms in Neural Networks, with the aim of teaching them to take decisions



AlphaSignal.ai

ML algorithms are tools used by us and machines

Today we will go through several algorithms. Remember that:

- Most important, know when to apply the tools and when not to
- Understanding the math will help you to master the tools, but it is not crucial
- Validation and quality control is vital
- Terminology is key when communicating

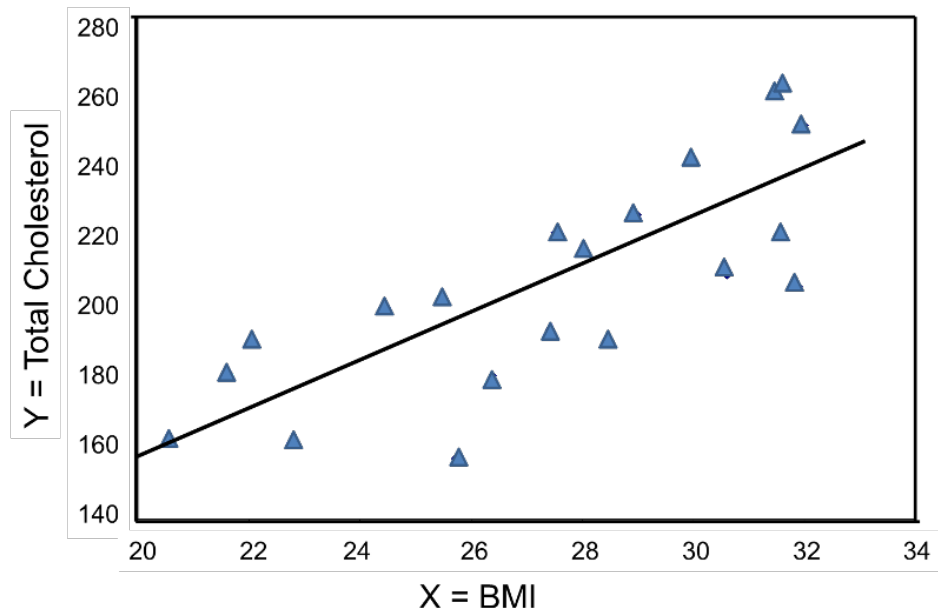


Supervised Machine Learning (SML)

- In SML, algorithms learn from **labelled data**
- **Regression** is used to understand the relationship between dependent and independent variables
- **Classification** assign test data into categories based on specific variables

Simple Linear (and logistic) regression

- Used to predict (forecast) the value of the dependent variable based on the independent variable
- Linear regression is applied on continuous variables, whilst logistic regression on discrete



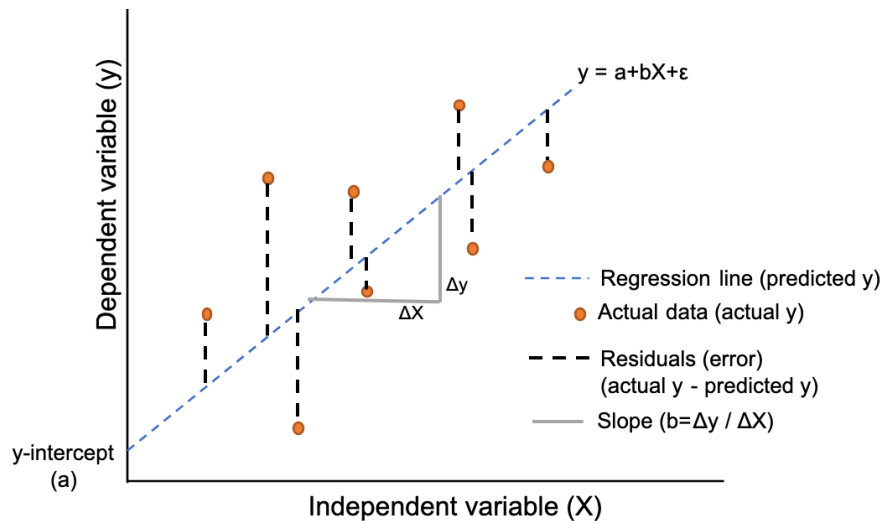
Simple linear regression

Diagram illustrating the components of the simple linear regression equation:

$$y = a + bx + \epsilon$$

Labels and arrows pointing to the equation components:

- Constant/Intercept (points to a)
- Independent Variable (points to x)
- Dependent Variable (points to y)
- Slope/Coefficient (points to b)
- Residual/Error (points to ϵ)



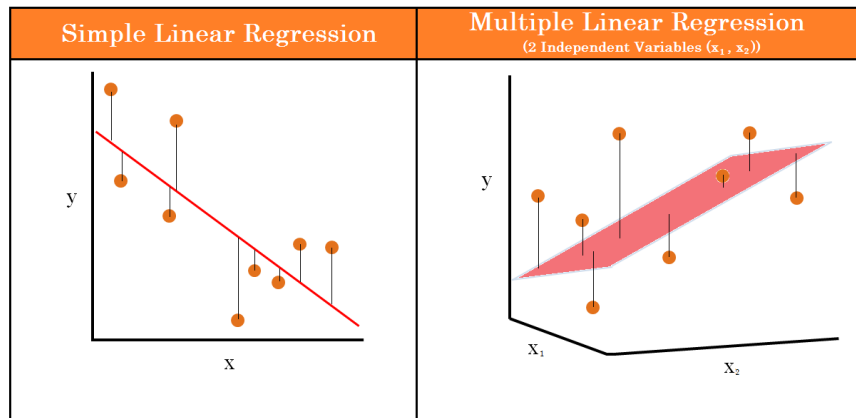
Reneshbedre.com

- Residuals can be used to validate the model by making sure that they are independent and normally distributed
- As independent variables increases, multiple linear regression is applied

Multiple linear regression

- Builds a model to describe Y in the best way using X_n
- Use independent variables to predict the dependent variable. Example:
 - Total Cholesterol = $a + b_1 \cdot \text{BMI} + b_2 \cdot \text{Time exercising} + b_3 \cdot \text{Shoe size} \dots + \epsilon$
- But is shoe size relevant?

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + \epsilon$$



Multiple linear regression assumptions and Root Mean Square Error

- Parametric test based on assumptions:
 - Linear relationship between Y and X
 - X_i are not highly correlated with each other
 - The variance of the residuals is constant
 - Independence of observations
 - Residuals are normally distributed
- Model can be tested with Root Mean Square Error (RMSE), the standard deviation of the residuals:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

n = Sample size

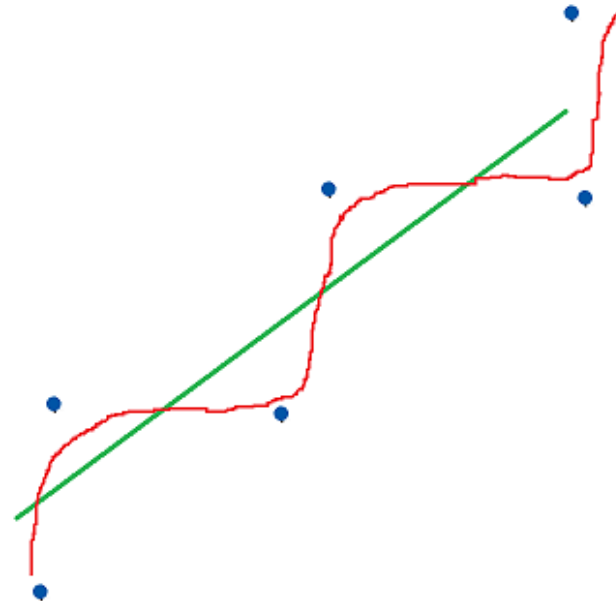
\hat{y} = Predicted value

y = True value

i = Variable i

Multiple linear regression for prediction

1. Create a random 80/20 split of the data, generating training data (80%) and test data (20%)
2. Train a regression model on the training data
3. Apply the model on the test data
4. Calculate RMSE of the training data (in-sample RMSE) and test data (out-of-sample RMSE)
 - Compare the RMSE. Indicates how well the model performs on new data.
 - More complex model \rightarrow Decreasing RMSE \rightarrow Overfitting



Linear regression models pros and cons

Pros:

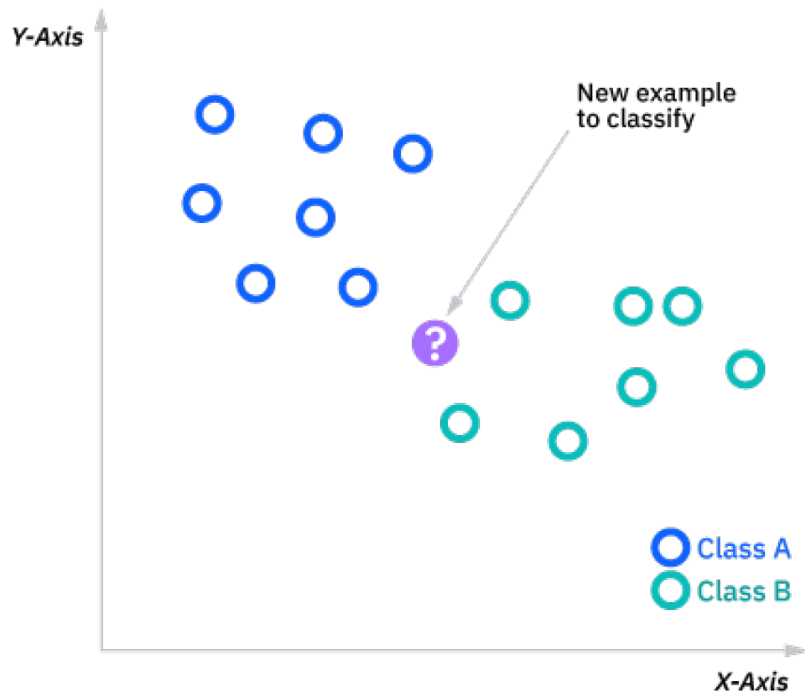
- Can be used on continuous (linear) and discrete (logistic) data
- Determine influence of independent variables on the dependent
- Identifying outliers

Cons:

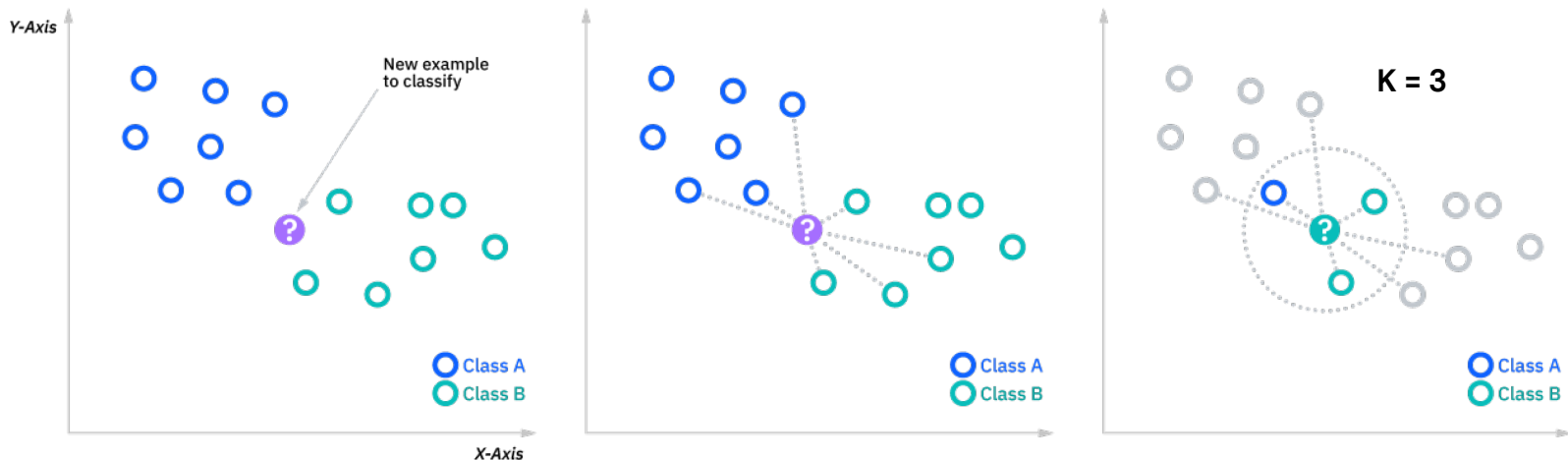
- No mixed data (continuous & discrete)
- Many assumptions
- Requires complete data and no missing data

K-nearest neighbors (KNN)

- Non-parametric algorithm *i.e.* no strong assumptions
- Often used for classification, predicting the group of a data point
- Applies majority voting based on:
 - Distance metrics
 - Number of K's



Computing KNN by distance and K's



IBM.com

1. Calculate the distances, usually with Euclidean distance
2. Find the nearest neighbours by ranking the distances
3. Majority vote on the predicted class label based on the K nearest neighbours

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

KNN pros and cons

Pros:

- It is easy to implement
- No need to train a model
- Versatile, distance algorithms can handle different types of data

Cons:

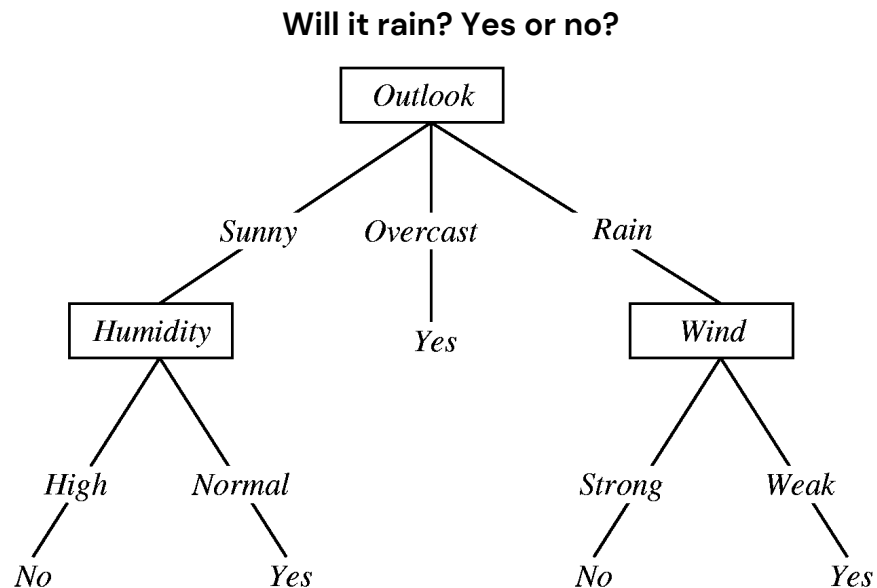
- Data should be of the same scale which can be difficult with large datasets
- Setting the K can be challenging

Tips:

- Test different K's
- K should be odd numbers to avoid any draws

Decision tree and random forest

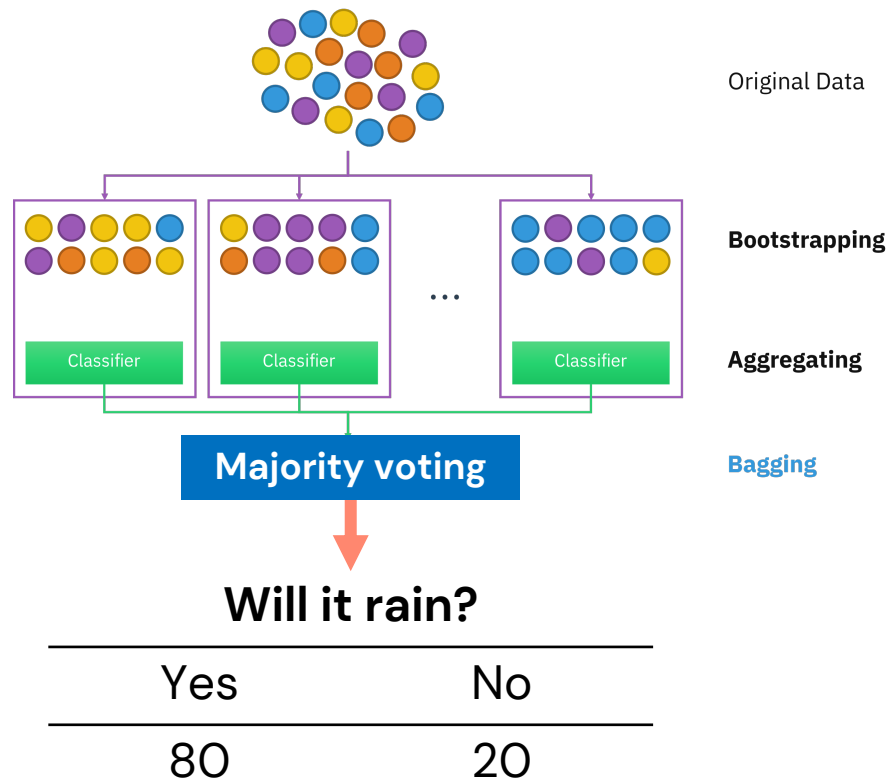
- Random forest is based on decision tree's
- Generates many decision tree's creates the random forest to classify unlabeled data
 - A single tree is not accurate
- Can use both categorical and continuous variables



Towardsdatascience.com

Random forest

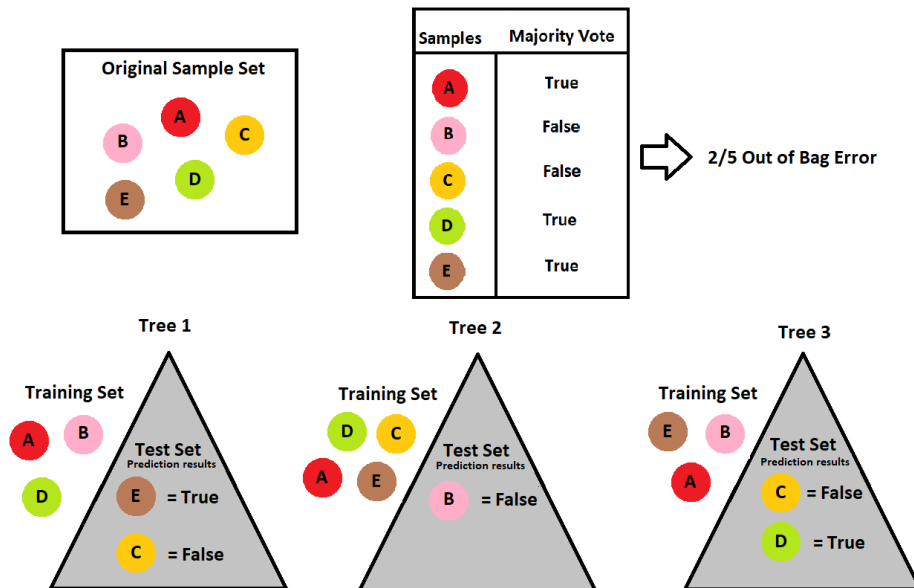
1. Create a bootstrapped dataset that is the same size of the original
 - Randomly selected data, where duplicates are allowed
2. Create a decision tree using the bootstrapped data using a random subset of variables
3. Repeat 1 and 2 multiple times
4. Impute your unlabeled data and let the random forests' many classifiers label
5. Majority vote classifies the unlabeled data



Wikipedia.org

Random forest validation with Out-of-Bag

- The Random forest model can be validated using the Out-of-bag error
- The Random forest is used to predict labels of data not selected for the bootstrapped data (test set)



By Sauxeris – Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=97324306>

Random forest pros and cons

Pros:

- Can be used on many types and mixes of data
- Can be applied on both classification and regression problems
- Can be applied on data with missing values
- No overfitting and curse of dimensionality

Cons:

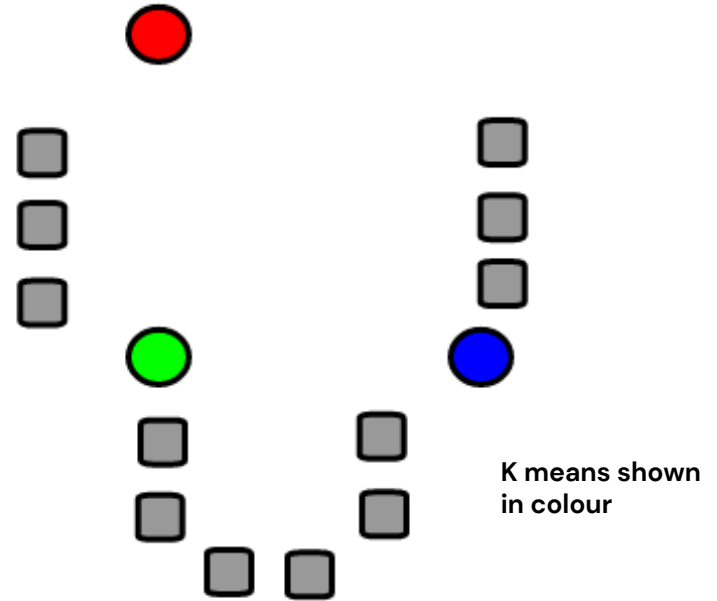
- Very complex and you can't follow the decision of the tree
- Training the model takes time and computing power

Unsupervised machine learning (UML)

- In UML, algorithms are used to analyze and cluster **unlabelled data**
 - Data grouping based on patterns
 - Similarities and differences of the data
- **Clustering** is applied on raw data and groups it based on similarities and differences between the structure and/or patterns of the data
- **Dimensionality reduction** can be applied to reduce complexity of data whilst preserving the structure to reduce "noise" and overfitting ML algorithms.

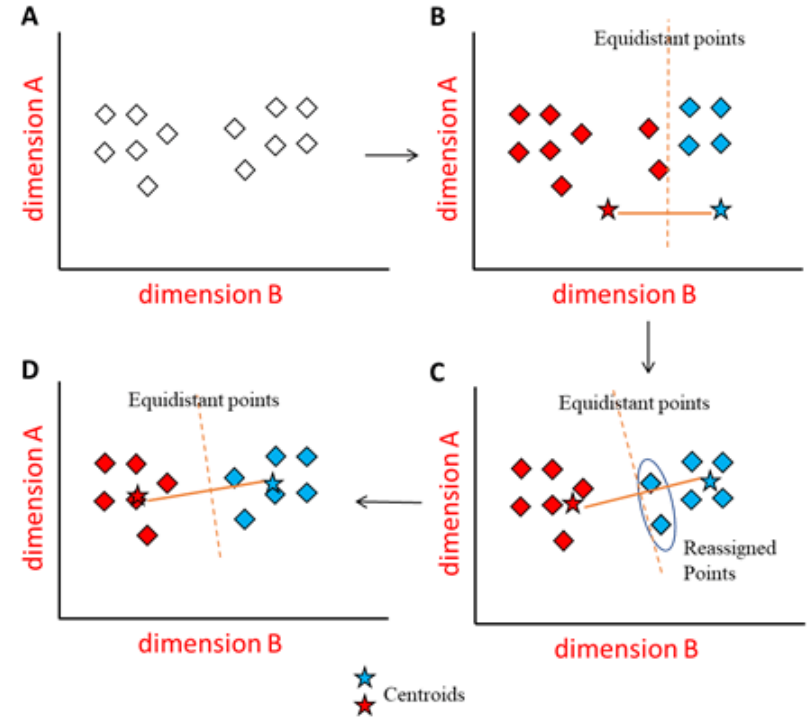
K-means clustering

- Not to be confused with KNN
- Groups similar datapoints in clusters
- K is the number of cluster and means generated



K-means clustering steps

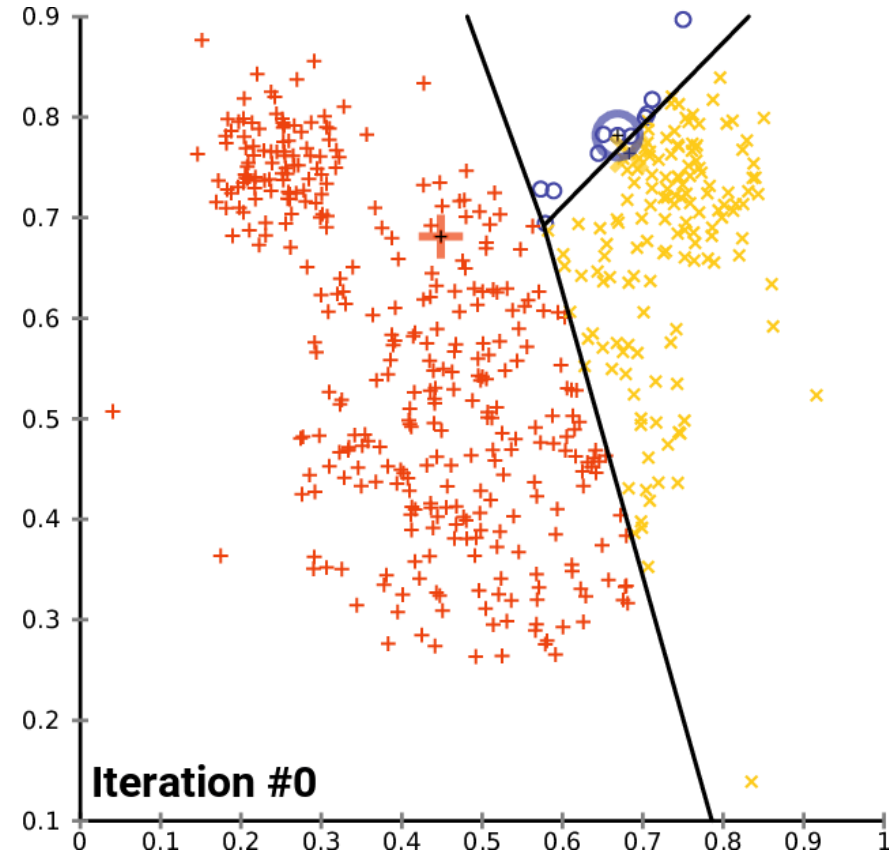
1. Set the number of K's
 - With Elbow plot
2. Generates K random centroids
3. Creates K clusters by assigning each data point to closest centroid
4. Calculates new centroids for each cluster
5. Reassigns points with new centroids
 - If new assignments, repeat 4
 - If no new assignments, terminate algorithm



Oxford Protein Informatics group, <https://www.blopig.com/blog/2020/07/k-means-clustering-made-simple/>

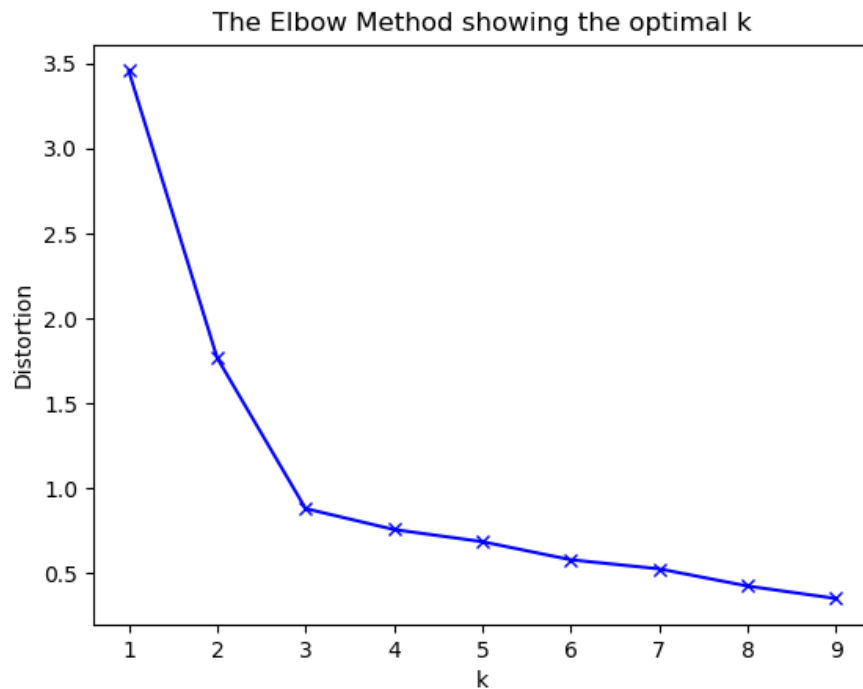
K-means clustering steps

1. Set the number of K's
 - With Elbow plot
2. Generates K random centroids
3. Creates K clusters by assigning each data point to closest centroid
4. Calculates new centroids for each cluster
5. Reassigns points with new centroids
 - If new assignments, repeat 4
 - If no new assignments, terminate algorithm



Elbow plot determines number of K's

- First step of K-means clustering is to set the K
- The Elbow method is common
- Distortions is the sum of squared distances of data points from cluster centers
 - Decreases as K increases.
 - 0 when K = number of points



Geekforgeeks.org

K-means clustering pros and cons

Pros:

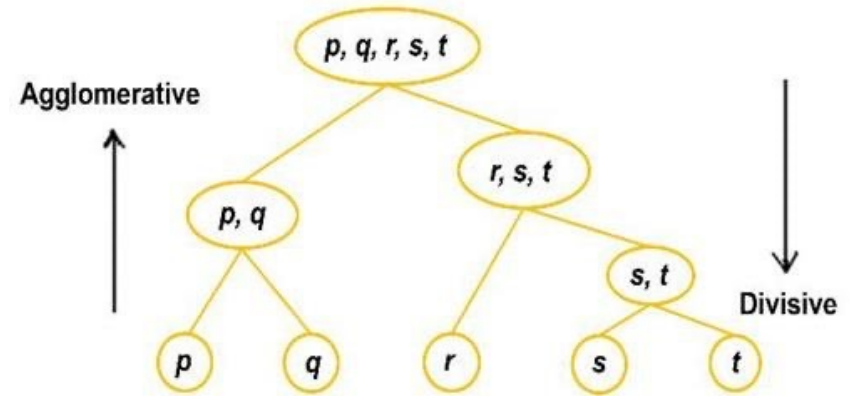
- Easy to use
- Can be used on large datasets
- Adapts to new data
- Clusters can be of different shapes and sizes

Cons:

- Sensitive to outliers
- Choosing K is manual labour and sometimes tough

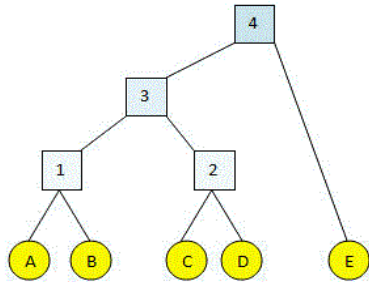
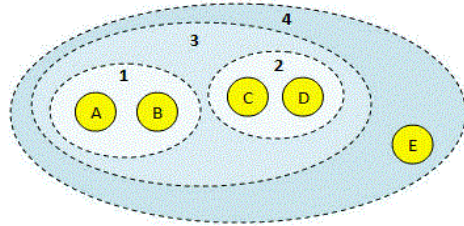
Hierarchical clustering

- Groups similar data points to clusters
- Defines clusters that are distinct from each other and datapoints within are similar
- Creates cluster by ordering clusters:
 - Bottom-up (Agglomerative)
 - Top-down (Divisive)

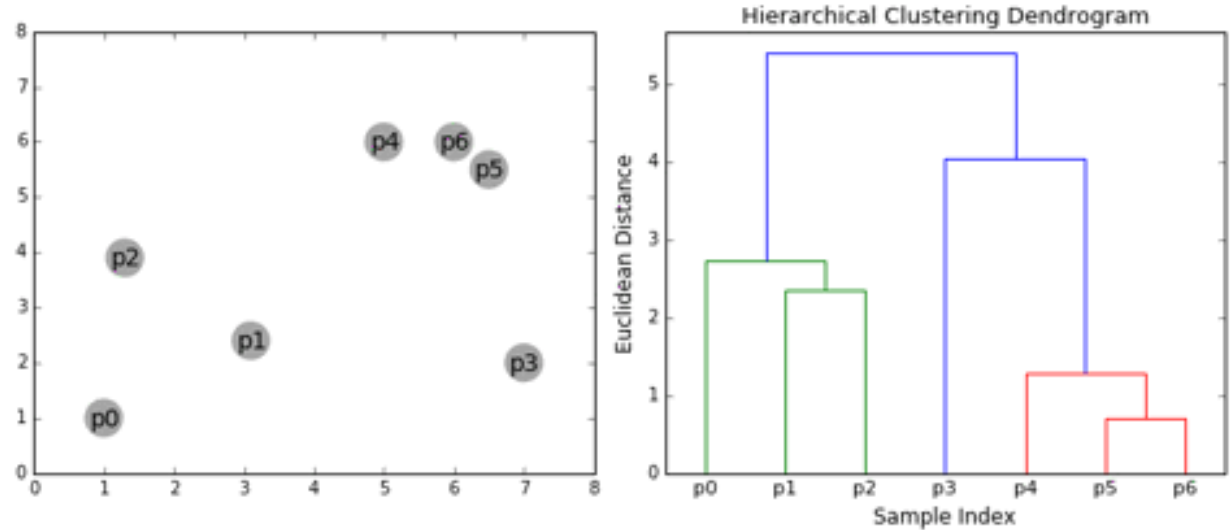


Giacoumidis E et al., "Blind Nonlinearity Equalization by Machine-Learning-Based Clustering for Single- and Multichannel Coherent Optical OFDM", *Journal of Lightwave technology*, 2017

Hierarchical clustering



<https://drive5.com/usearch/manual/agg.html>



- The length of the branch in the dendrogram show how similar the data points are.
 - Long branch = dissimilar, short branch = similar

Hierarchical clustering pros and cons

Pros:

- Easy to use
- The dendrogram gives information about the data structure
- Can be used to set number of clusters

Cons:

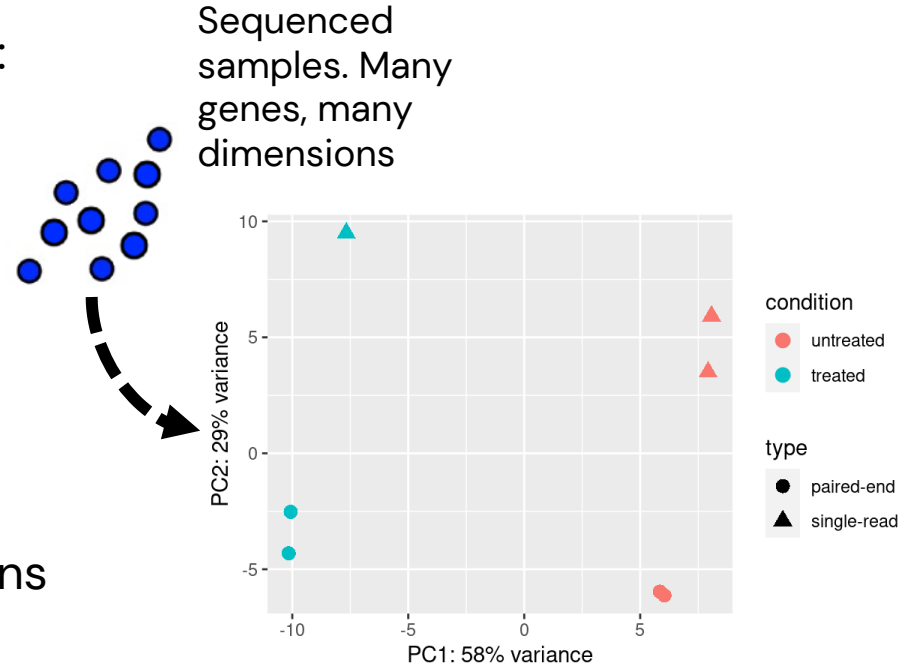
- Sensitive to outliers
- Does not work well with missing data or mixed data
- In complex data, difficult to determine number of relevant clusters

Principal component analysis (PCA)

Common and versatile method used for:

- Analysing the structure of data features
- Pre-processing for other ML algorithms
- Visualisation

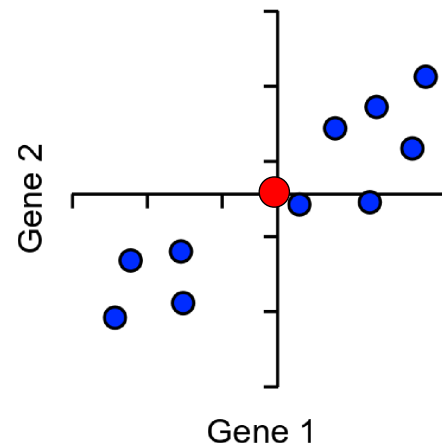
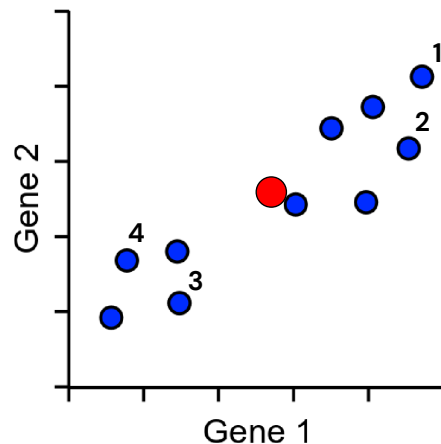
Summarises large multi-dimensional datasets to smaller number of dimensions (ideally 2) that can be visualised



Love M et al., "Analyzing RNA-seq data with DESeq2", *DESeq2 Vignette*, 2023

Initializing the PCA

	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	10	11	3	2
Gene 2	15	14	2	5
Gene 3	10	7	8	9

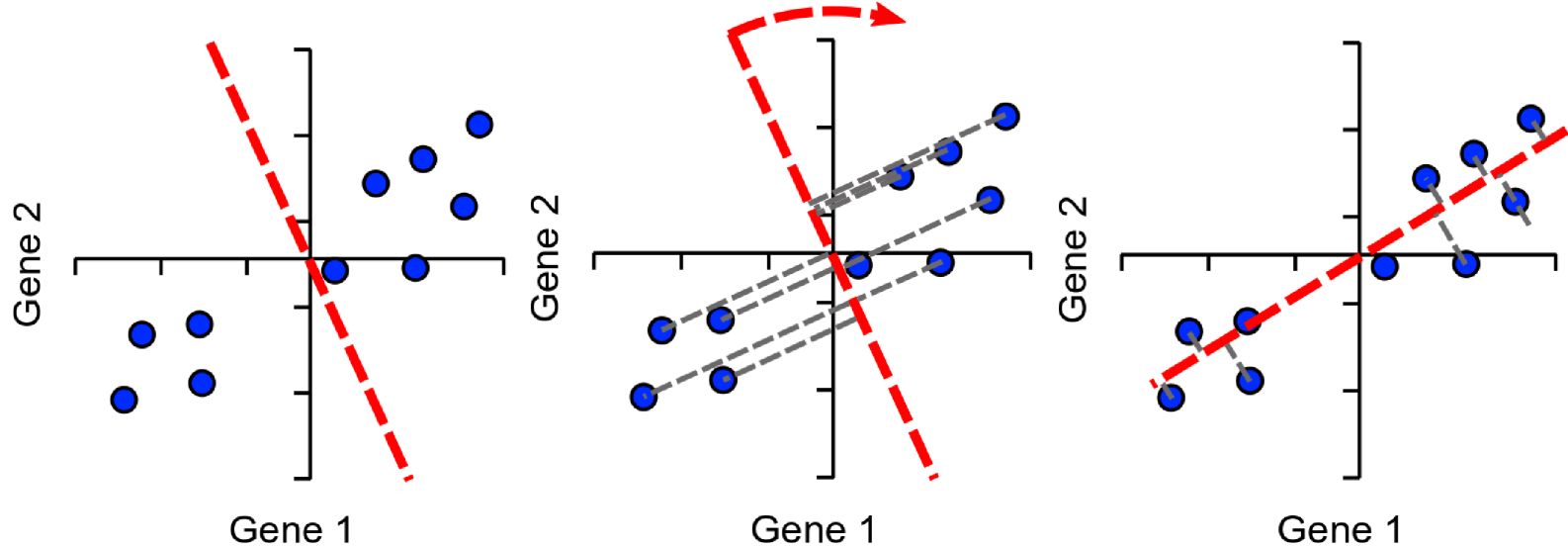


1. Plot the data. Gene 1 & 2 is higher in sample 1 & 2...

2. Calculate the average of gene 1 and 2 (and n) to find the **center** of the data.

3. Center the data at the origin (0,0)

Initializing the PCA and PC1



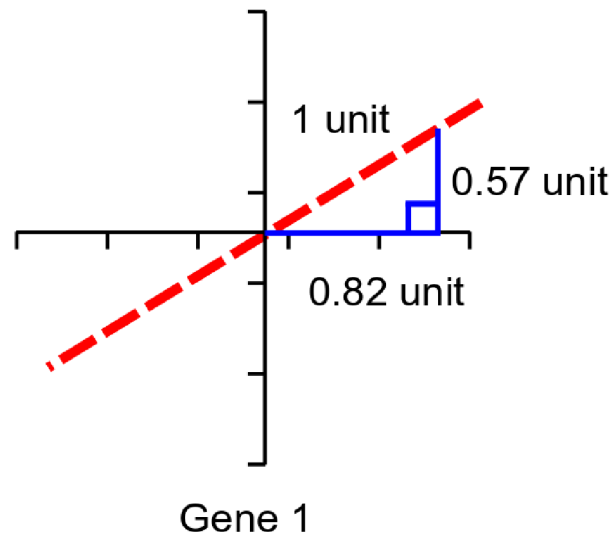
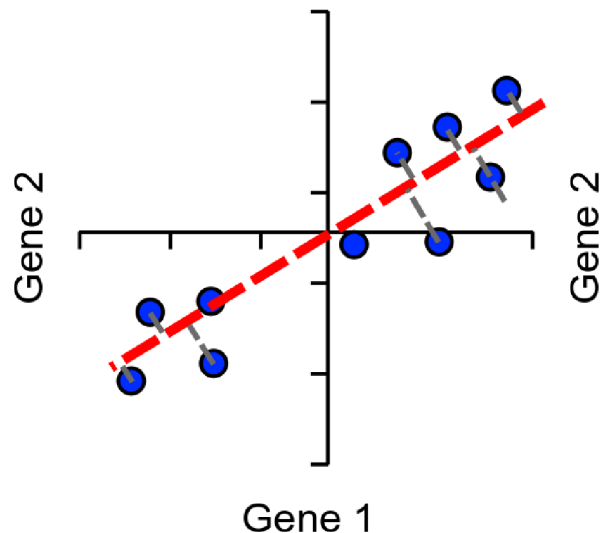
Find the line, through the origin, with the best fit. The best fit is defined by PCA projecting the distance of the point to the line and minimizing it.

The line is called **Principal Component 1 (PC1)**

Calculating the eigenvector

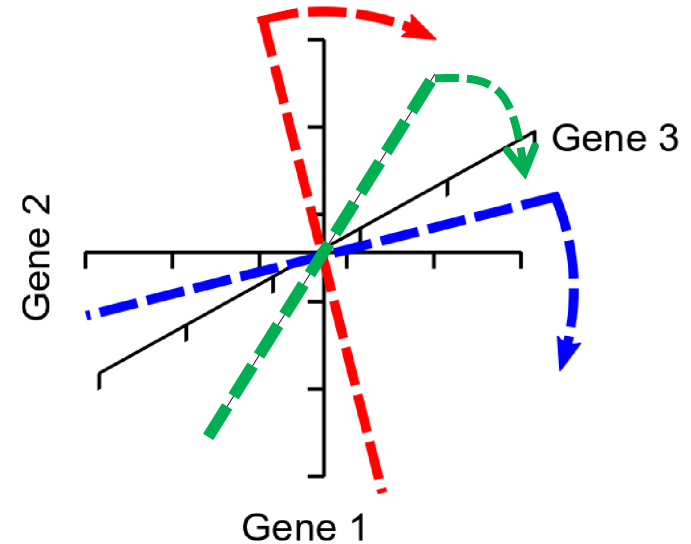
The eigenvectors are calculated.

Higher loading indicated more influence on the PC *i.e.* Gene 1 (0.82) influence more than Gene 2 (0.57).

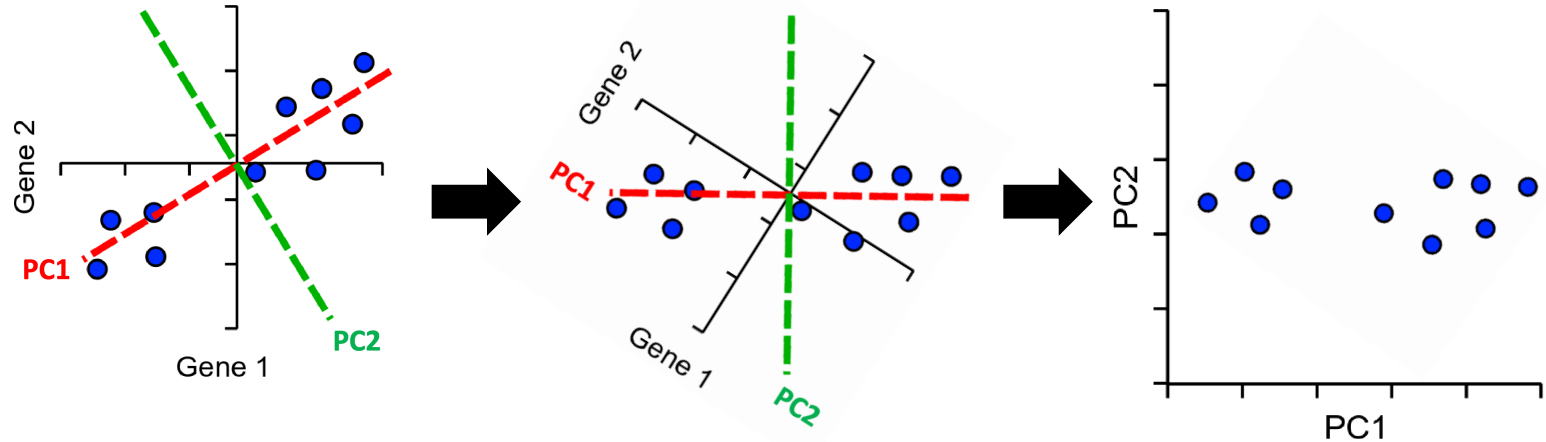


Multi-dimensions and PC n

- PC2 is perpendicular to PC1. PC3 is perpendicular to PC1 and PC2 etc.
- PCs are the same number as genes
- PC1 explains most of the variance in the data. PC2 the second most etc.
- Projection in 2D, so two PC's are projected



Generating the PCA plot



- The datapoints are projected onto PC.
- Hopefully, we see some clustering...

PCA pros and cons

Pros:

- Can remove noise (correlated features)
- Improve ML algorithms by removing noise
 - Reduces overfitting
- Visualisation

Cons:

- PCA turns independent variables to PC's which can be hard to interpretate
- Requires standardised data and therefore does not work well on mixed data

tSNE and UMAP are advancements of PCA, projecting the data better making clustering easier

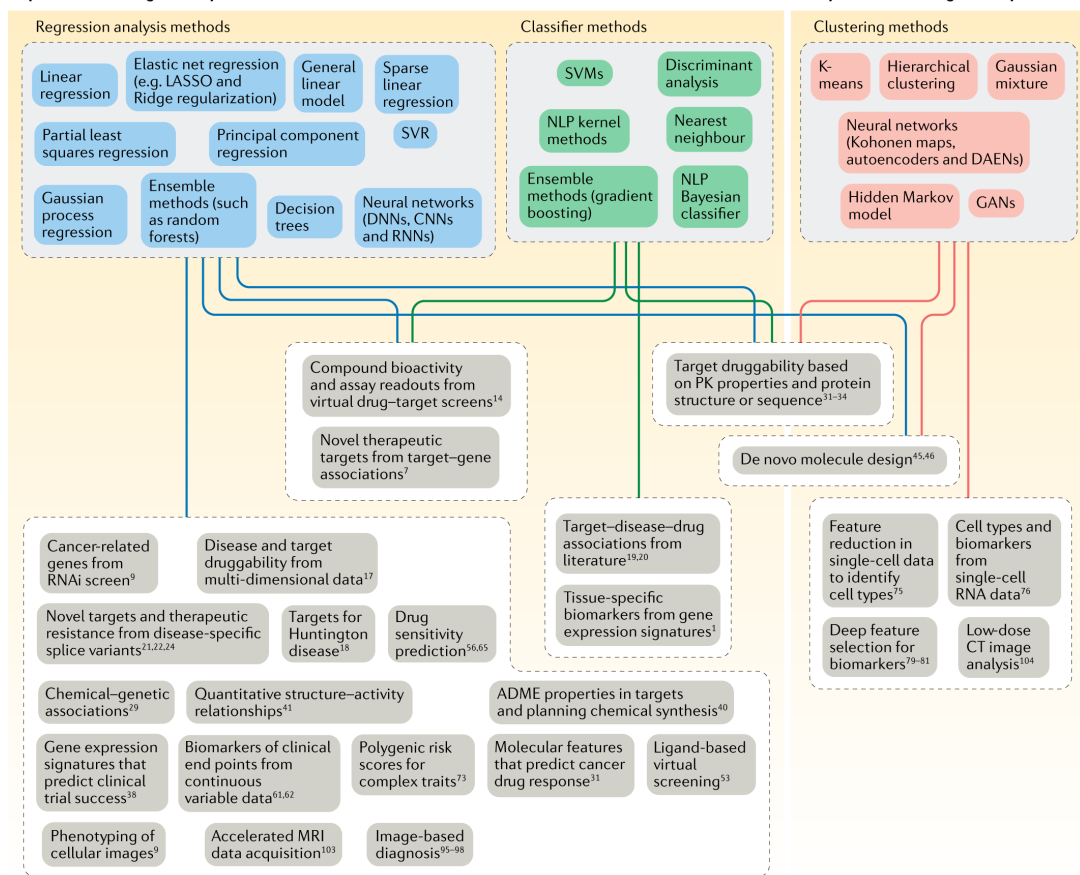


**Karolinska
Institutet**

ML in medicine and pharmacology

- ML algorithms are used together
- Nested in networks or parts of pipelines
- Used as tools, from a ML toolbox
- Important to know when and why to use it

Supervised learning techniques

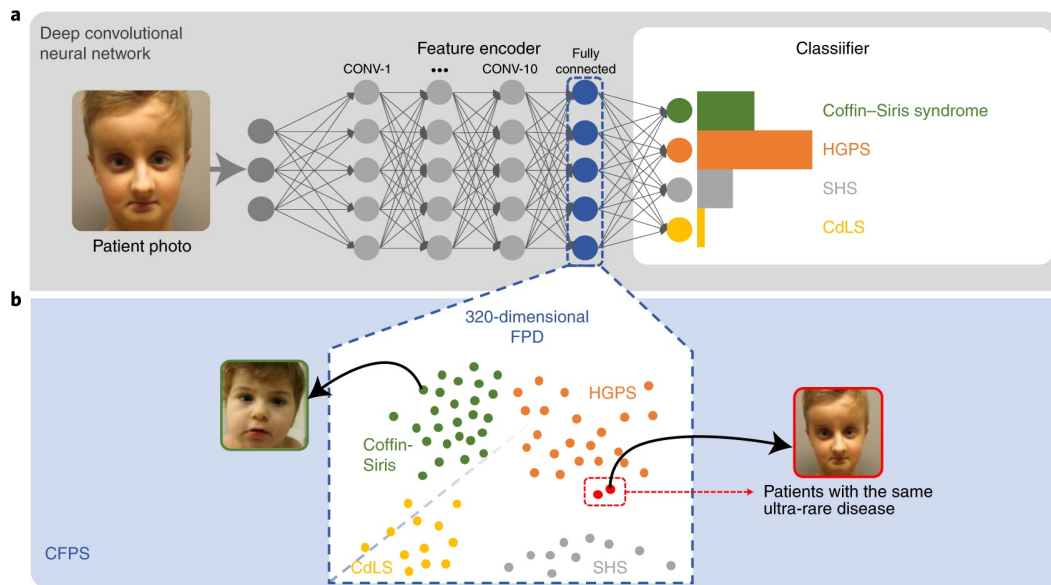


GestaltMatcher and Face2Gene

Supervised classifiers are often used in image analysis, for example when diagnosing rare diseases.

Here, KNN is nested into a Deep Neural Network.

Datapoints in the KNN is other phenotype patients



Hsieh, TC., Bar-Haim, A., Moosa, S. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat Genet* **54**, 349–357 (2022). <https://doi.org/10.1038/s41588-021-01010-x>

DESeq2, a brief introduction

- Most used method in analysing bulk RNA-seq data
- Other methods are limma and edgeR. Common aim is to find differentially expressed genes (proteins, lipids etc.)
- Great vignette and good start when going into bioinformatics:
<http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>

Love et al. *Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8



METHOD

Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love^{1,2,3}, Wolfgang Huber² and Simon Anders^{2*}



**Karolinska
Institutet**