Machine learning in bioinformatics

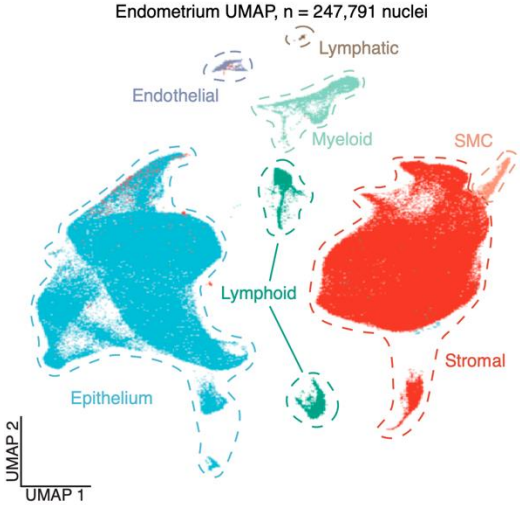
An introduction to supervised/unsupervised machine learning in transcriptomics

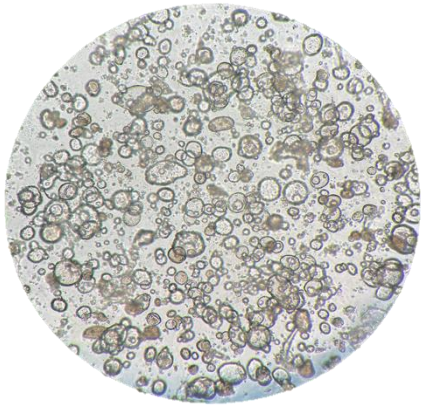
Short about me:

- Gustaw Eriksson (gustaw.eriksson@ki.se)
- B.Sc. in Molecular Biology & M.Sc. in Bioinformatics from Lund University
  → Thesis and projects at the Max Planck Institute and SciLifeLab
- PhD student in the Reproductive Endocrinology and Metabolism group led by Elisabet Stener-Victorin
- Researching the etiology of PCOS

My research

- I study cell type specific disease markers of polycystic ovary syndrome (PCOS)
  - → 10x snRNA-sequencing and analysis of endometrium, fat and muscle
  - → Cell culture, primarily endometrium epithelial organoids and stroma cells
  - → Clinical statistics, proteomics, animal experiments
  - → **Combining wet and dry lab**





Eriksson, G., Li, C., Sparovec, T.G. *et al.* Single-cell profiling of the human endometrium in polycystic ovary syndrome. *Nat Med* **31**, 1925–1938 (2025). https://doi.org/10.1038/s41591-025-03592-z

Todays schedule:

Before lunch:

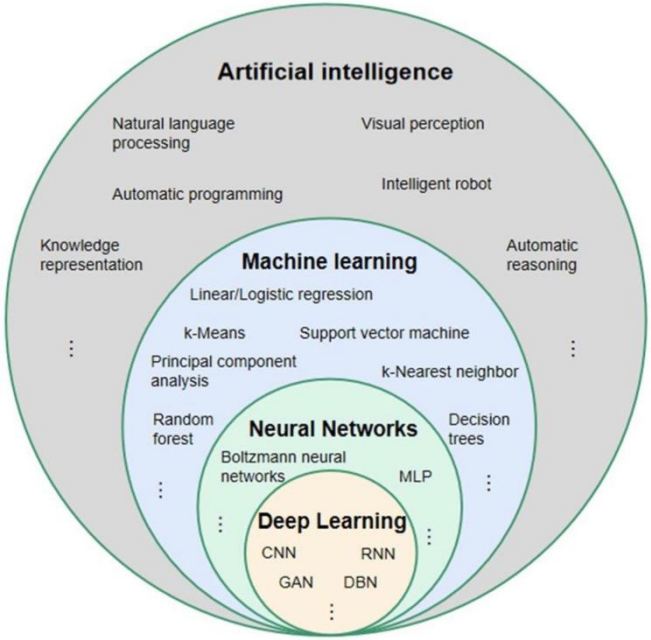
- Supervised Machine Learning (SML)
  - → Linear regression, K-nearest neighbours (KNN), Random Forest
- Unsupervised Machine Learning (UML)
  - → K-means clustering, Hierarchical clustering, Principal component analysis (PCA)

After lunch:

- Machine Learning (ML) in practice
- DESeq2 tutorial in R analysing bulk RNA-sequencing 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 and methods.
Remember that:

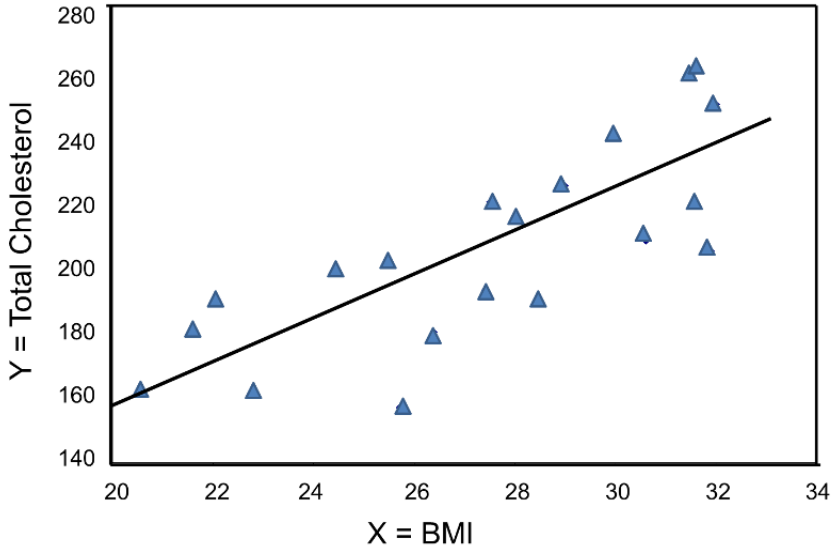
- Most important, know when to apply the tools
- Understanding the math will help you to master the tools, but it is not crucial
- Validation and quality control is vital
- **Terminology is key when communicating**

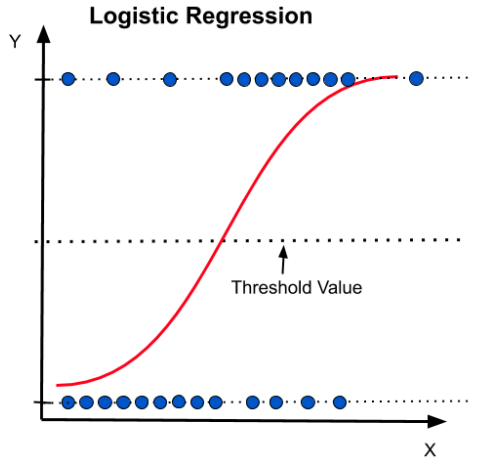
Supervised Machine Learning (SML)

- 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

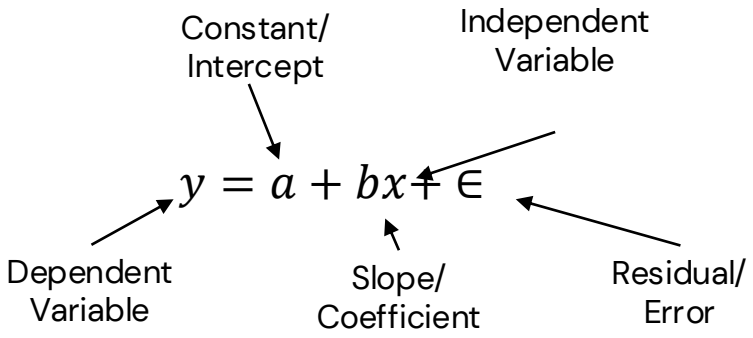
Simple Linear (and logistic) regression

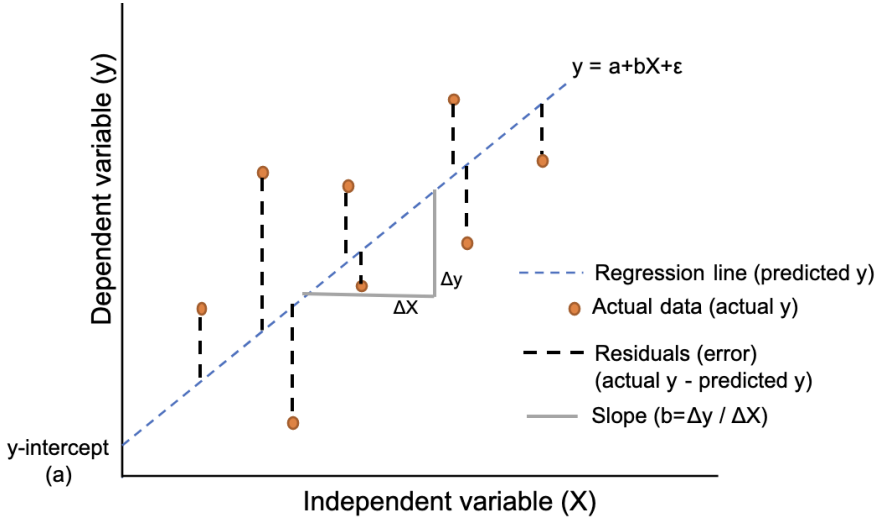
- Used to predict (forecast) the value of the dependent variable based on the independent variable
- Linear regression is applied on continuous variables, whilst logistic regression on discrete.





Simple linear regression





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

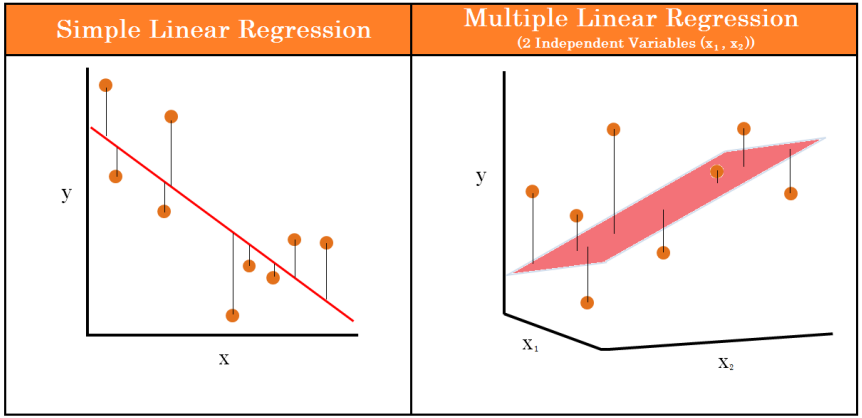
Multiple linear regression

- 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*BMI + b_2*Time exercising + b_3*Shoe size… + ∈

- But is shoe size relevant?
  - → Can be tested with a hypothesis test where H_0 is no relationship (no slope) and/or theory

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + \in$$



Multiple linear regression assumptions and Root Mean Square Error

- Parametric test based on assumptions:
  - → Linear relationship between Y and X
  - → Independent variables (X_i) are not highly correlated i.e. similar 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:

- Tells you how well your prediction works of the dependent variable (*y*)

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

n = Sample size

$\hat{y}$ = Predicted value

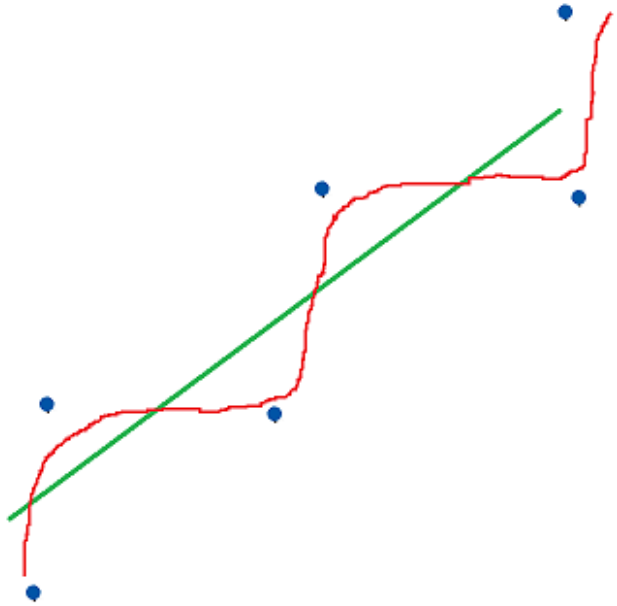
y = True value

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 → Decreasing RMSE → 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

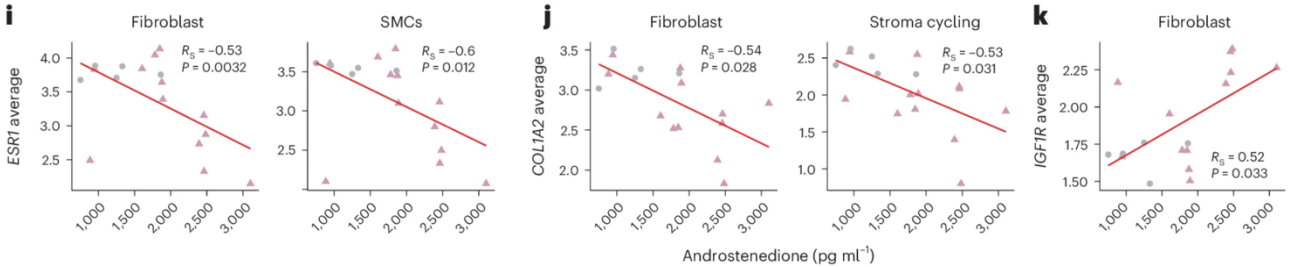
Cons:

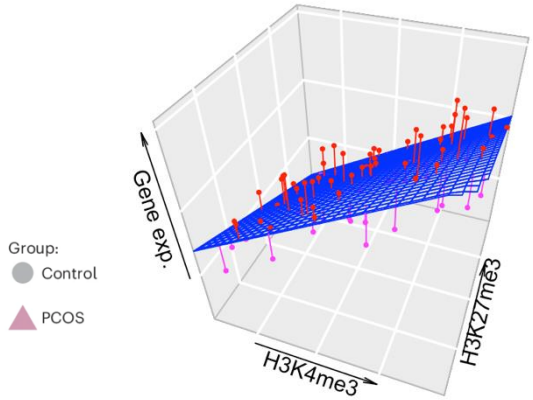
- No mixed data (continuous & discrete
- Many assumptions
- Requires complete data and no missing data

Linear/logistic regression in use

Linear regression can be used for:

- Studying gene expressions relationship to clinical variables
  - → Does a specific variable affect a gene expression levels
- Gene expression from regulatory elements
  - → Number of transcription factor binding site to predict mRNA levels
- Drug-response studies
  - → How does drug concentrations affect cell viability





Eriksson, G., Li, C., Sparovec, T.G. *et al.* Single-cell profiling of the human endometrium in polycystic ovary syndrome. *Nat Med* **31**, 1925–1938 (2025). https://doi.org/10.1038/s41591-025-03592-z

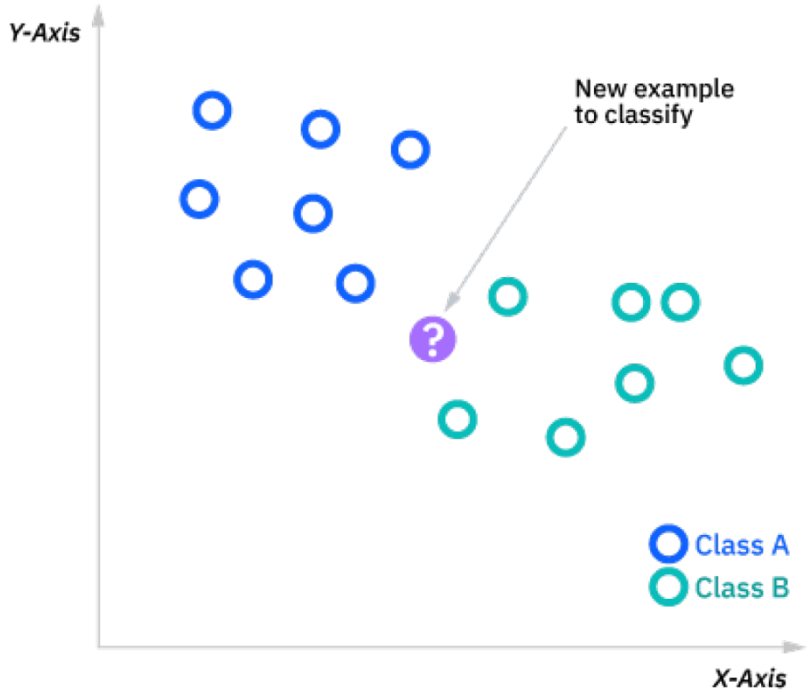
https://compgenomr.github.io/book/relationship-between-variables-linear-models-and-correlation.html

Machine Learning

Linear Regression

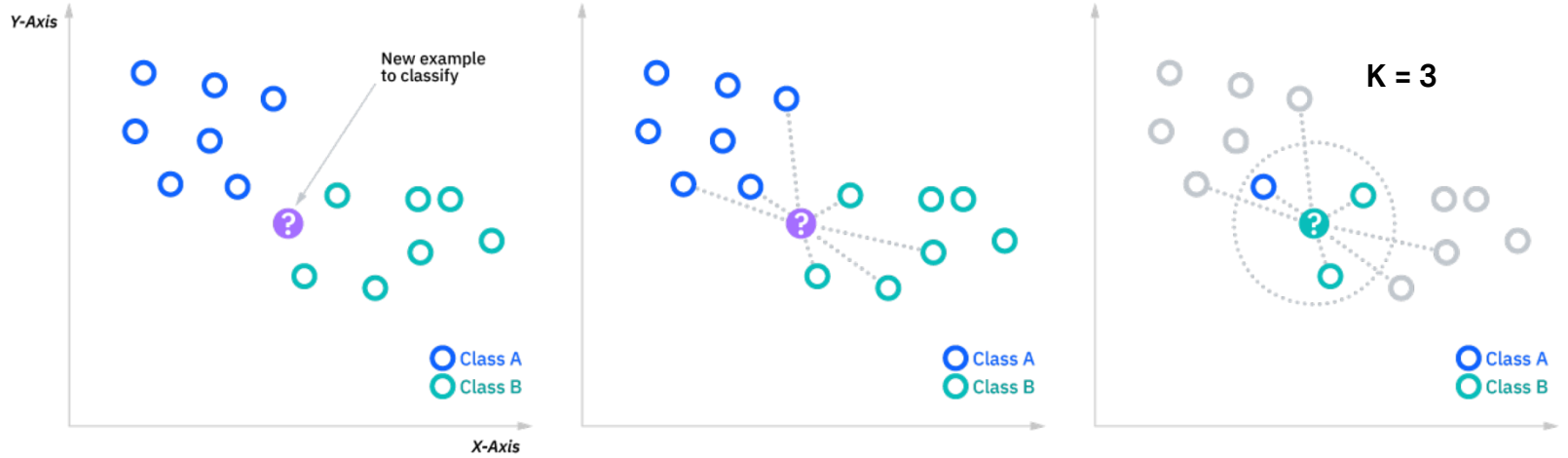
K-nearest neighbors (KNN)

- 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



IBM.com

Computing KNN by distance and K's



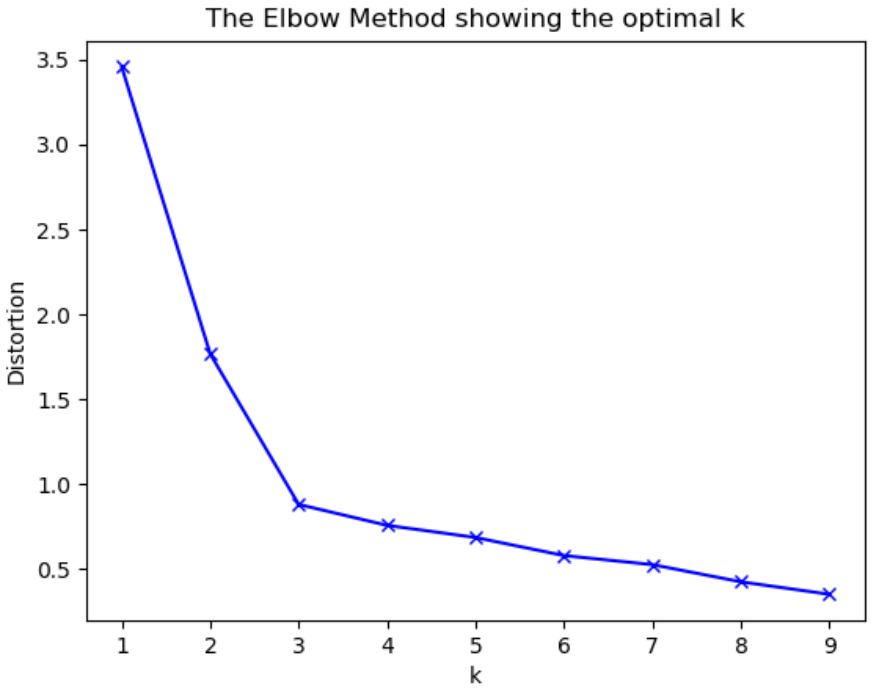
IBM.com

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

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

Elbow plot determines number of K's

- It can be challenging to set the number of K's
- 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

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

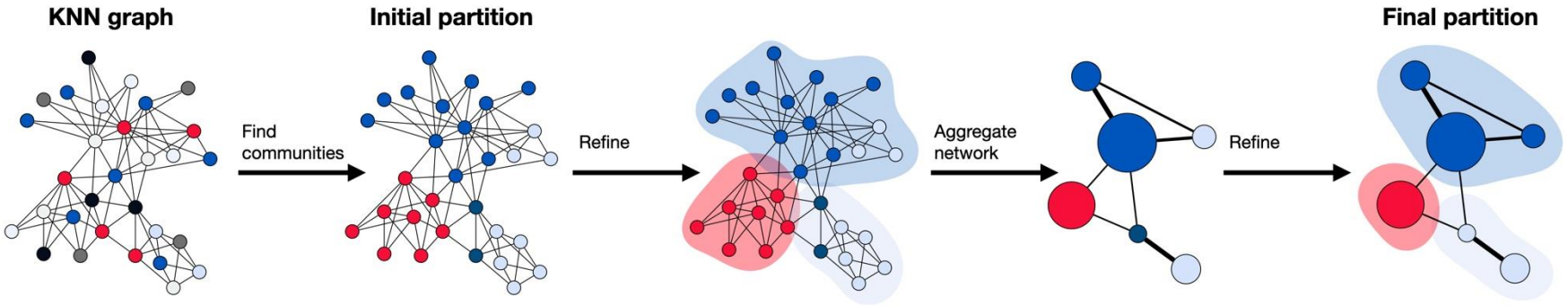
Cons:

- 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

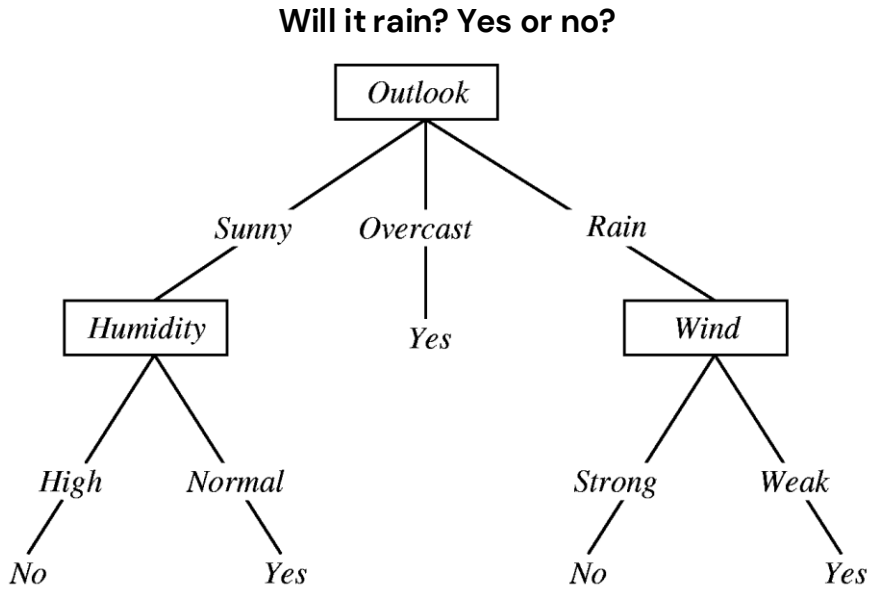
KNN in bioinformatics



- KNN is often applied in cell type classification in single-cell RNA-seq
  - → Leiden clustering uses KNN
  - → Each cluster is defined as a cell type based on cell type specific gene markers

Decision tree and random forest

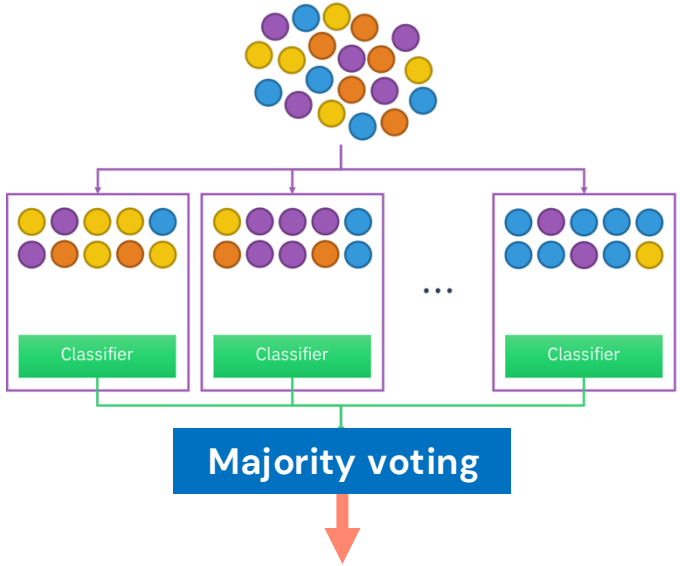
- Random forest is based on decision tree's
- Generates many decision tree's creates the random forest to classify unlabeled data
  - → A single tree is not accurate
- Can use both categorical and continuous variables



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



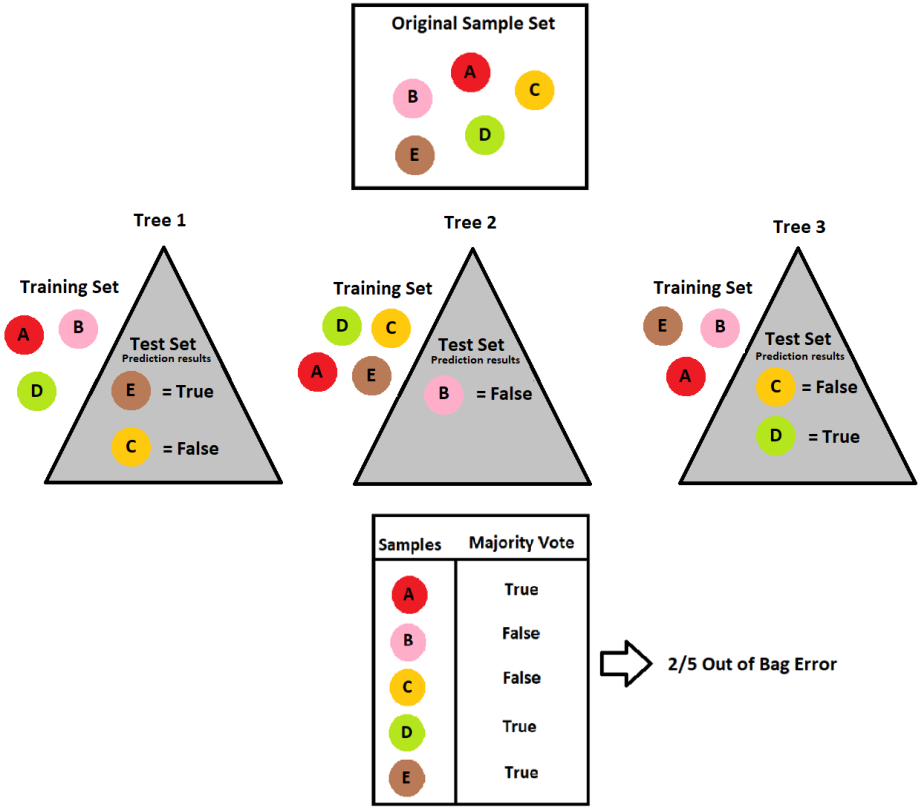
Will it rain?

Yes	No
80	20

Wikipedia.org

Random forest validation with Out-of-Bag

- The Random forest model can be validated using the Out-of-bag error (OOB)
- The decision tree is built by labelled data (training set) to predict labels of data not selected for the decision tree (test set)
- Compare the OOB predictions to the real labels to calculate the OOB error



Random forest pros and cons

Pros:

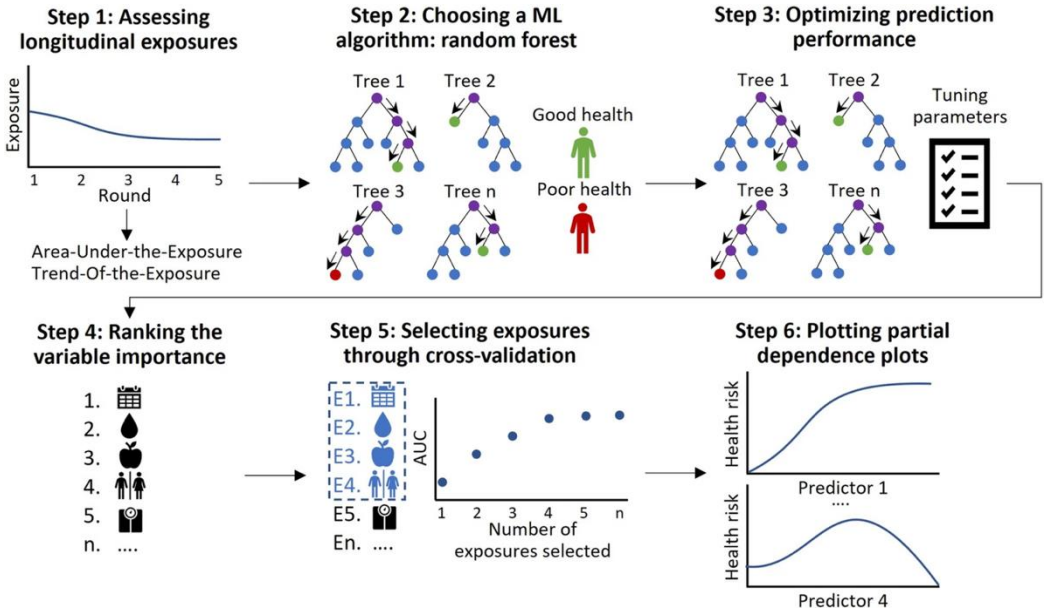
- Can be used on many types and mixes of data
- Can be applied on both classification and regression problems
- Can be applied on data with missing values
- No overfitting and curse of dimensionality

Cons:

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

Random forest examples

- Can be used for disease prediction, with input such as:
  - → Gene expression levels
  - → Clinical variables
  - → Mix of both



Loef, B., Wong, A., Janssen, N.A.H. *et al.* Using random forest to identify longitudinal predictors of health in a 30-year cohort study. *Sci Rep* **12**, 10372 (2022). https://doi.org/10.1038/s41598-022-14632-w

Unsupervised machine learning (UML)

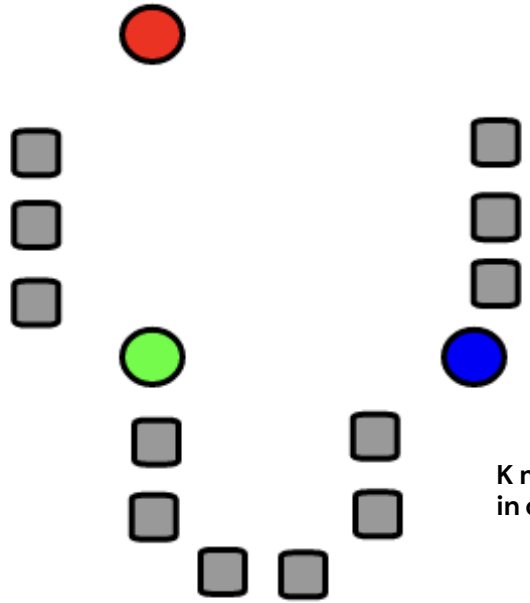
- In UML, algorithms are used to analyze and cluster **unlabelled data**
  - → Data grouping based on patterns
  - → Similarities and differences of the data

- **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

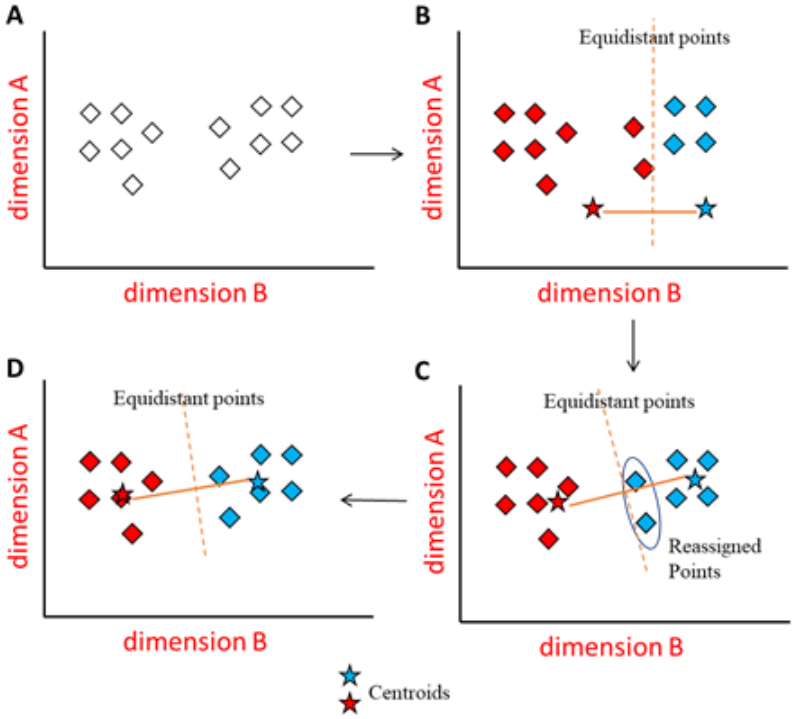


K means shown in colour

By I, Weston.pace, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2463053

K-means clustering steps

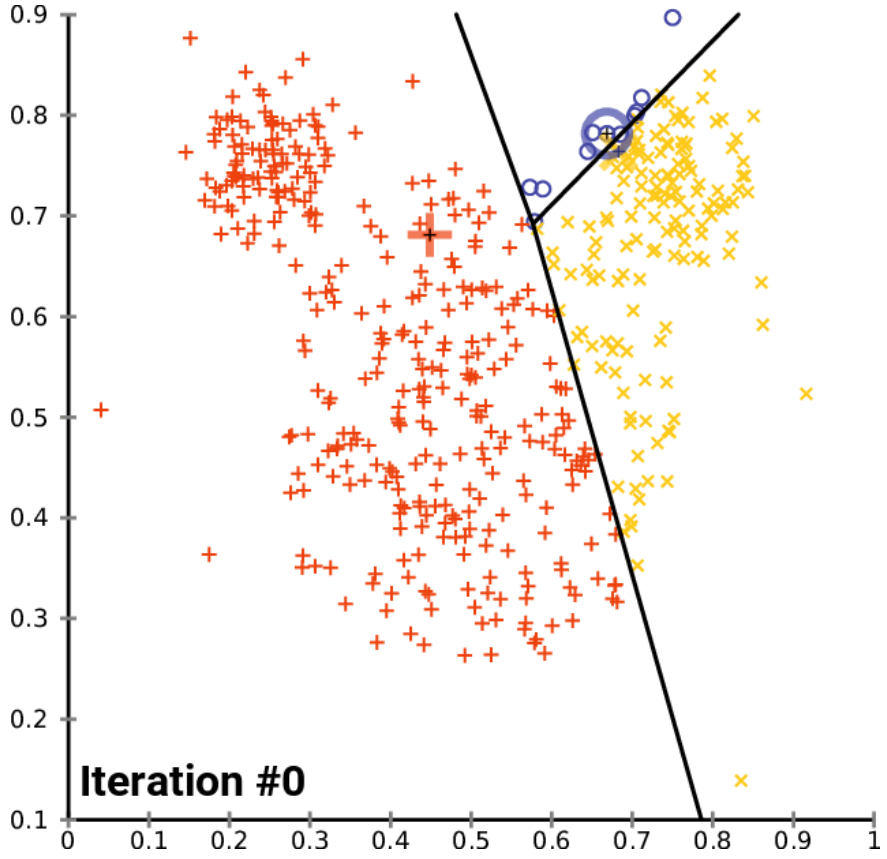
1. Set the number of K's manually
   → With Elbow plot
2. Generates K random centroids
3. Creates K clusters by assigning each data point to closest centroid
4. Calculates new centroids for each cluster
5. Reassigns points with new centroids
   → If new assignments, repeat 4
   → If no new assignments,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



By Chire - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=59409335

K-means clustering pros and cons

Pros:

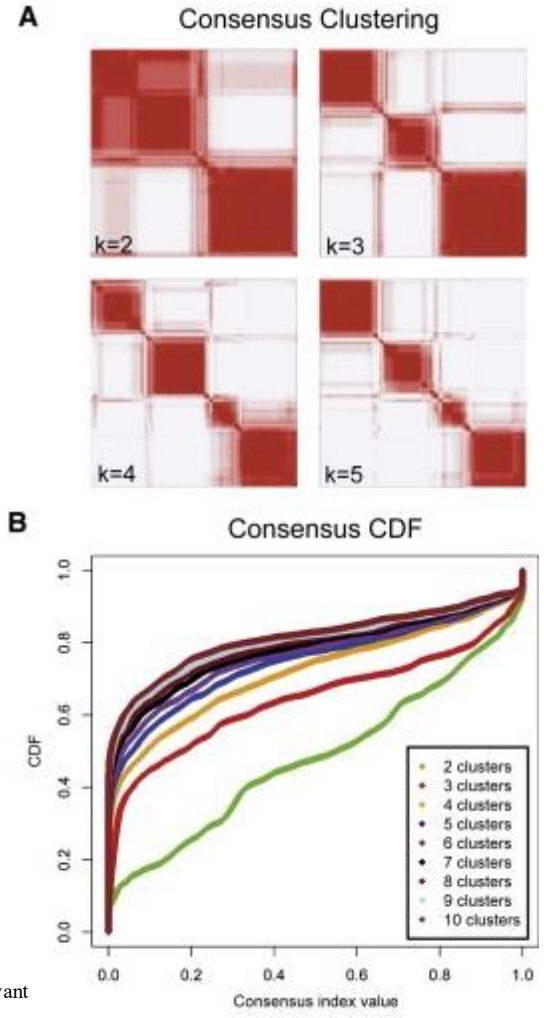
- Easy to use
- Can be used on large datasets
- Adapts to new data
- Clusters can be of different shapes and sizes

Cons:

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

K-means clustering in practice

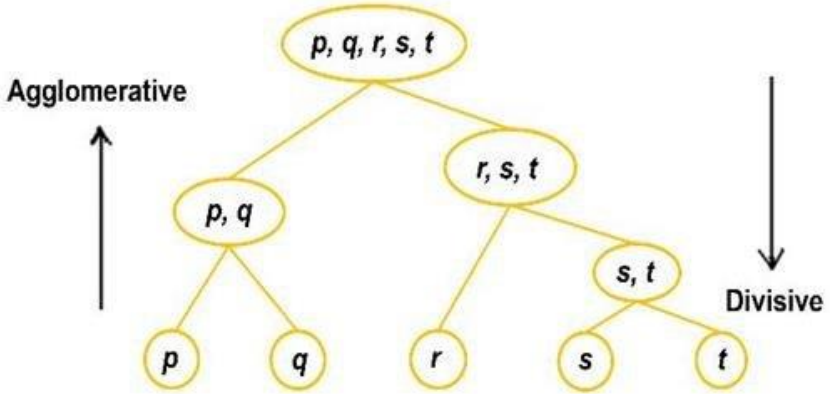
- Can be used to cluster genes with similar expression levels derived from many patient samples
  - → The clustering is used to subtype conditions
  - → In the example, glioblastoma multiforme is subtyped
  - → K-means clustering is part of a collection of clustering methods, called consensus clustering



Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98-110. doi:10.1016/j.ccr.2009.12.020

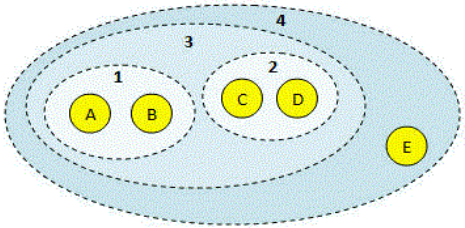
Hierarchical clustering

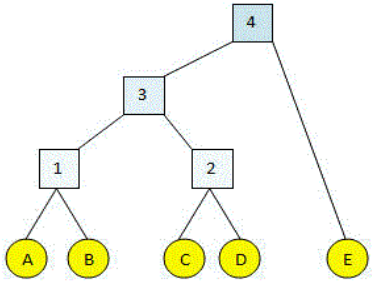
- Groups similar data points to clusters
- Defines clusters that are distinct from each other and datapoints within are similar
- Creates cluster by ordering clusters:
  - → Bottom-up (Agglomerative)
  - → Top-down (Divisive)



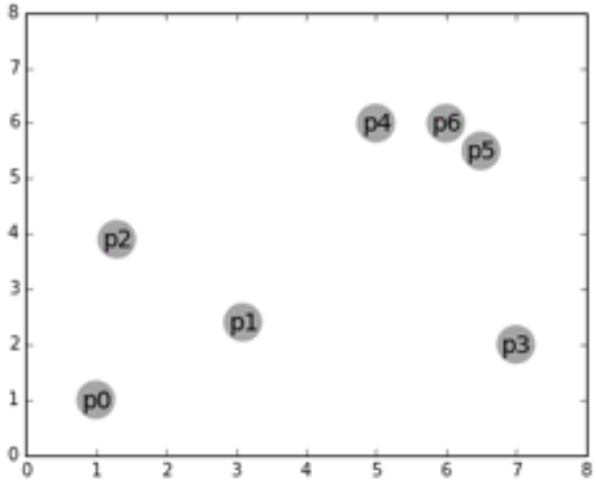
Giacoumidis E et al., "Blind Nonlinearity Equalization by Machine-Learning-Based Clustering for Single- and Multichannel Coherent Optical OFDM", *Journal of Lightwave technology*, 2017

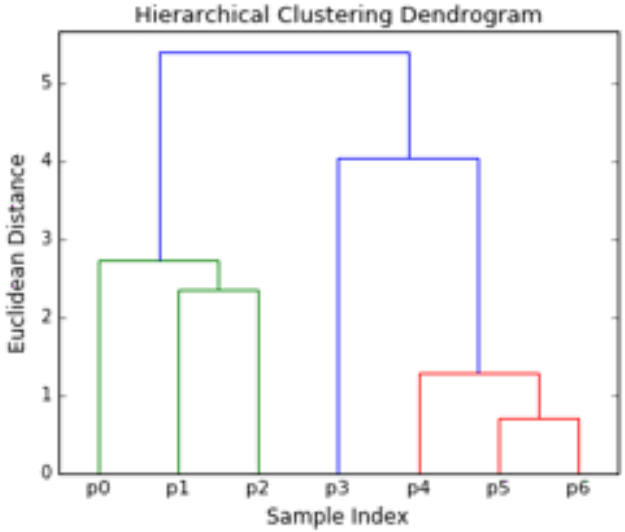
Hierarchical clustering





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





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

Hierarchical clustering pros and cons

Pros:

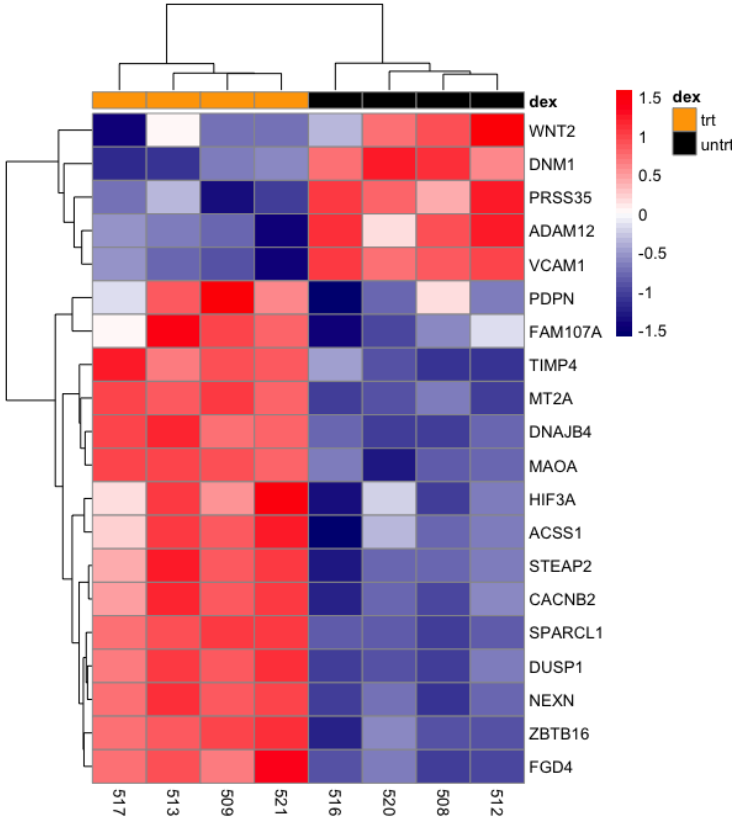
- Easy to use
- The dendrogram gives information about the data structure
- Can be used to set number of clusters

Cons:

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

Hierarchical clustering in bioinformatics

- Commonly used to cluster gene expression heatmaps
- ChIP-seq patterns
  - → Cluster genomic regions based on histone markers
- Phylogenetic trees to discover evolutionary relationships



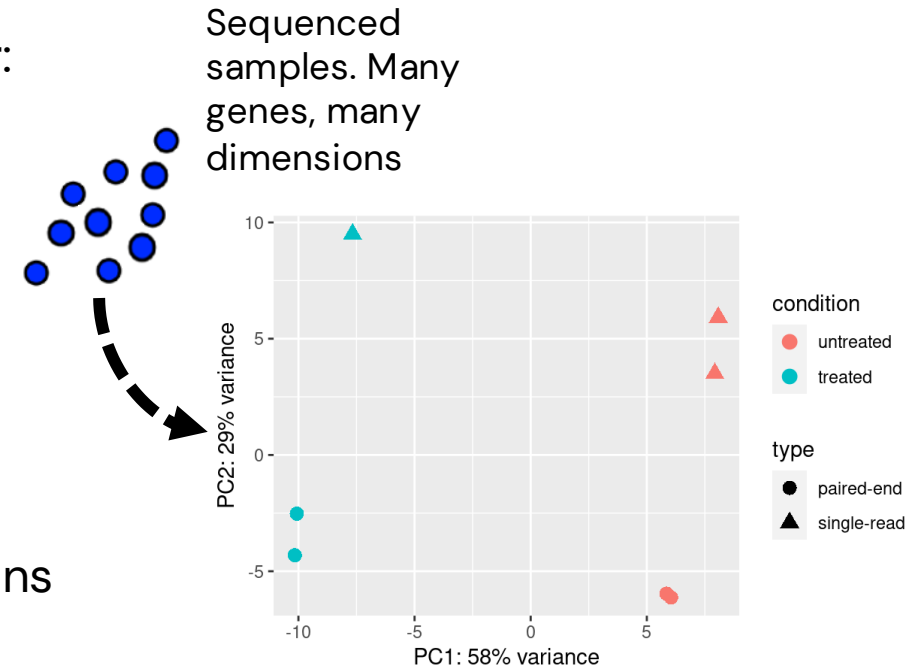
https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r/Lesson5_intro_to_ggplot/

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



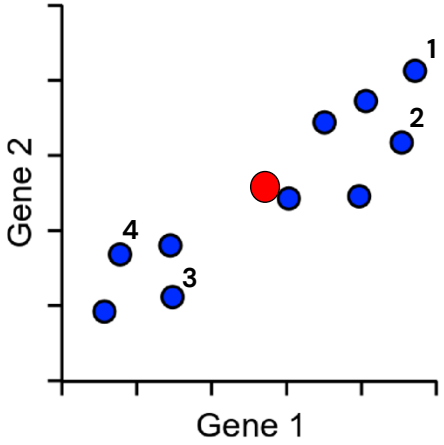
Love M et al., “Analyzing RNA-seq data with DESeq2”, *DESeq2 Vignette,* 2023

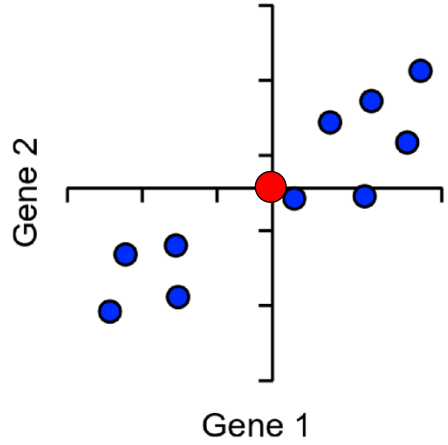
Gene expression data format

	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
Gene *n*	X	X	X	X

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



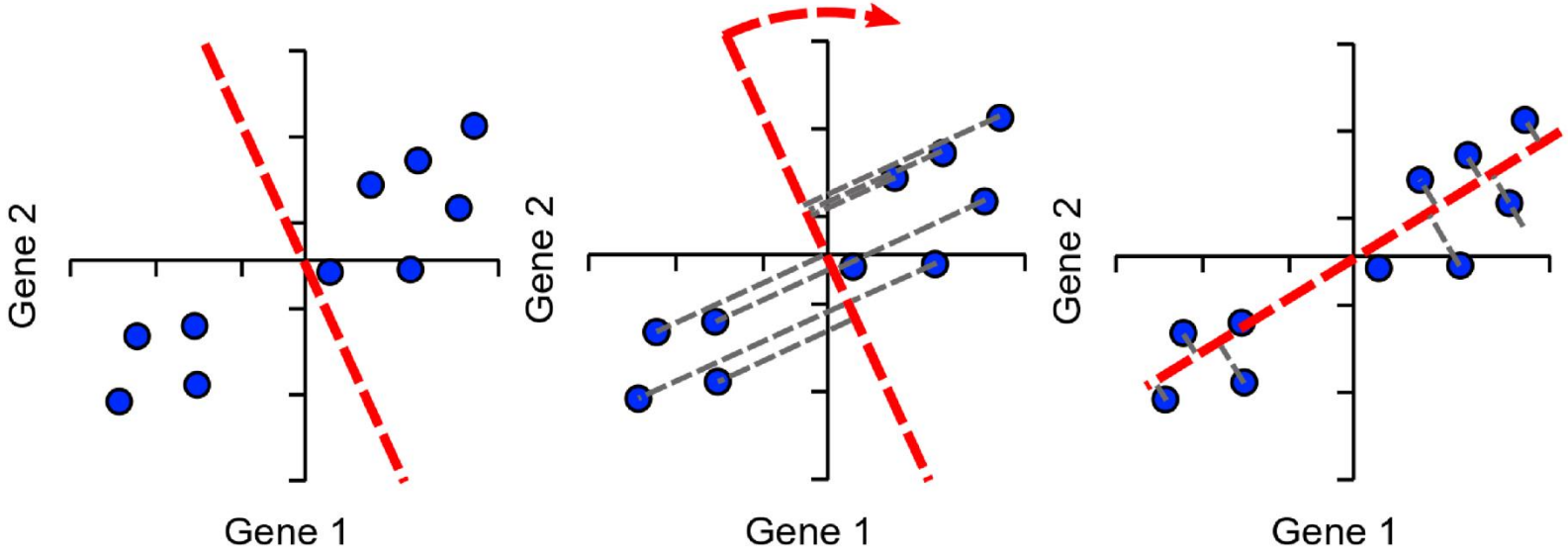


1. Plot the data. Gene 1 & 2 is higher in sample 1 & 2...

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



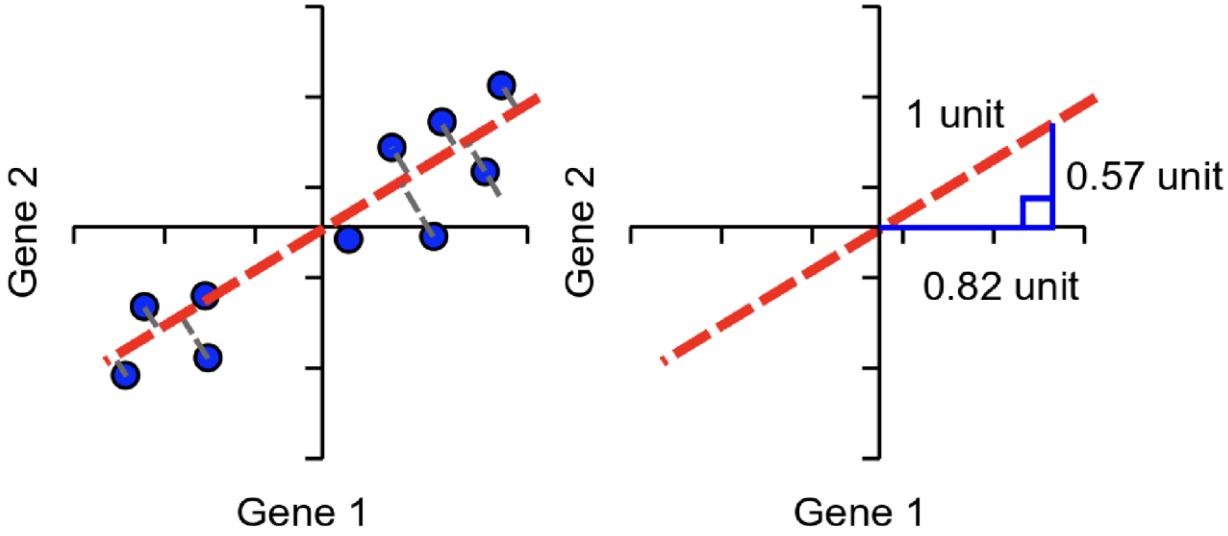
Find the line, through the origin, with the best fit. The best fit is defined by PCA projecting the distance of the point to the line and minimizing it.

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

Calculating the eigenvector

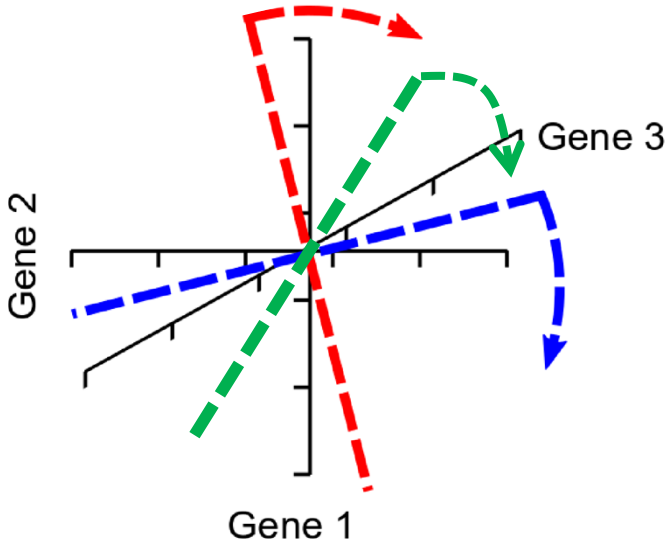
The eigenvectors are calculated.

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

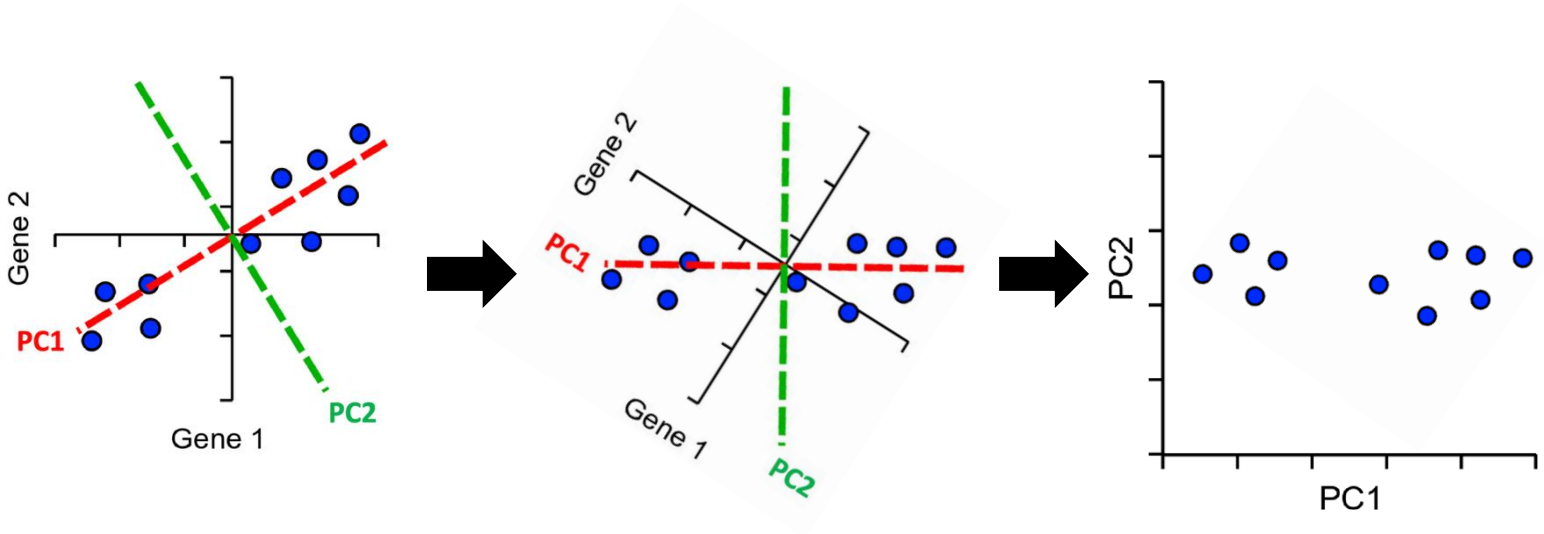


Multi-dimensions and PC *n*

- PC2 is perpendicular (vertical) 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. P2 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 alternatives to PCA, that sometimes reveal cluster structure better than PCA

PCA in real life...

- The 2-gene example illustrates the mechanism
  - → It shows how variance is captures in PC
  - → Easy to visualise

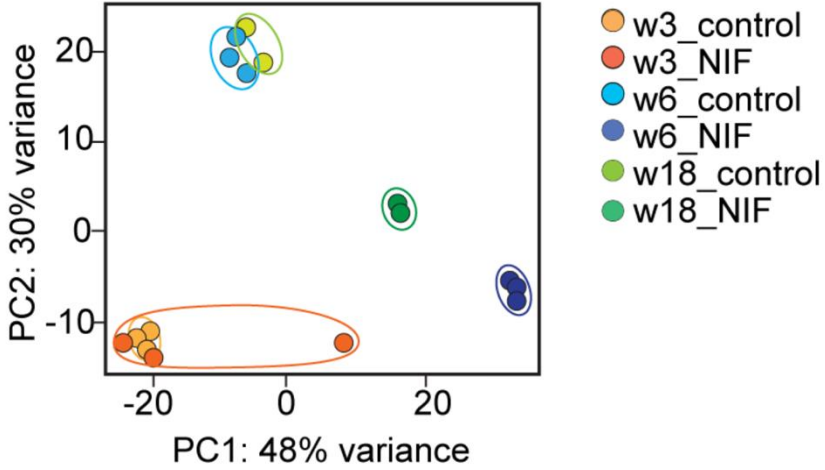
- When applied on real RNA-seq data
  - → ≈20,000 genes or more if including non-coding RNA
  - → Every gene contribute to every PC, but with different loadings (weights)

- Loadings are eigenvectors of the covariance matrix
  - → How much to the gene expression relate to each other
  - → Metadata such as condition, treatment, batch, etch is used to interpretation

PCA, main ingredient in quality control

- PCA is performed in almost all high throughput –omics methods
  - → Transcriptomics, proteomics, lipidomics, metabolomics…
  - → The PCA is generated from the expression levels of the -omics

- How well to replicates cluster? Do conditions separate? Is there a batch effect?
  - → These questions can be investigated with PCA



Schmidt-Christensen A, Eriksson G, Laprade WM, et al. Structure-function analysis of time-resolved immunological phases in metabolic dysfunction-associated fatty liver disease (MASH) comparing the NIF mouse model to human MASH. *Sci Rep*. 2024;14(1):23014. Published 2024 Oct 3. doi:10.1038/s41598-024-73150-z

If interested in more PCA and machine learning…

StatQuest has really good tutorials on machine learning and statistics: https://www.youtube.com/watch?v=FgakZw6K1QQ

Machine learning in bioinformatics

An introduction to supervised/unsupervised machine learning in transcriptomics

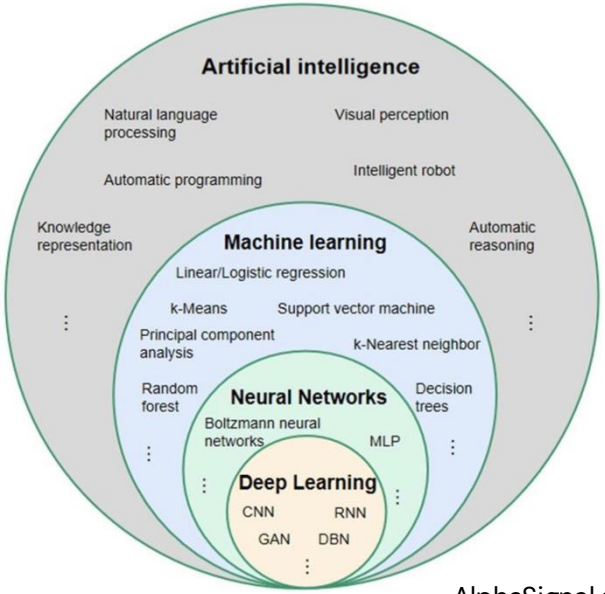
AI, deep learning and machine learning

Today we have learned about classical machine learning, but how does AI and deep learning fit in?

- Deep learning is a different branch of ML using layered architecture
- Large Language Models (LLMs) used in GPTs, are specialised deep learning models for language
- Both are applied on large scale biological datasets

For bioinformatics, classical ML remains essential as the data is:

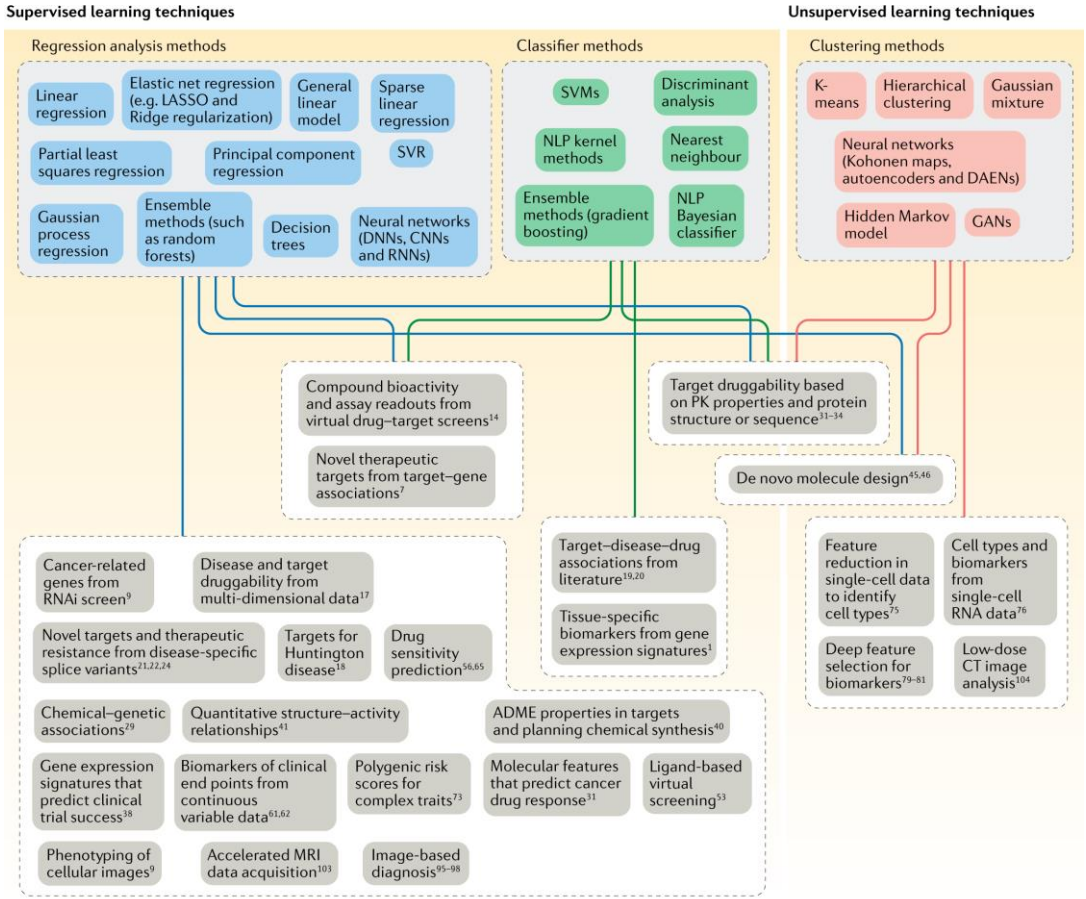
- High dimensional
- Small sample sizes
- Require biological interpretation



AlphaSignal.ai

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



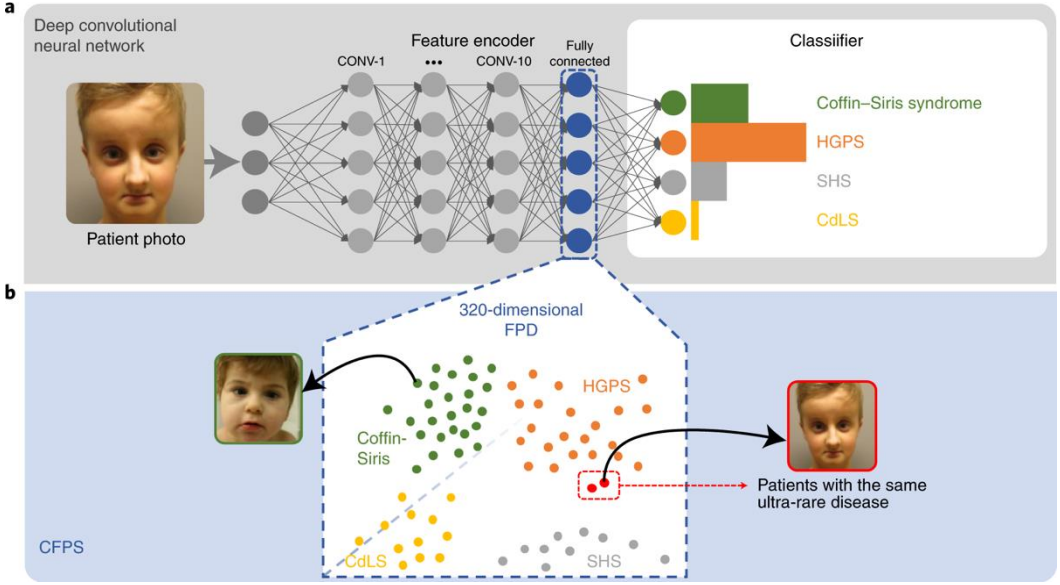
Vamathevan, J., Clark, D., Czodrowski, P. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18**, 463–477 (2019). https://doi.org/10.1038/s41573-019-0024-5

GestaltMatcher and Face2Gene, machine learning nested deep learning neural networks

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

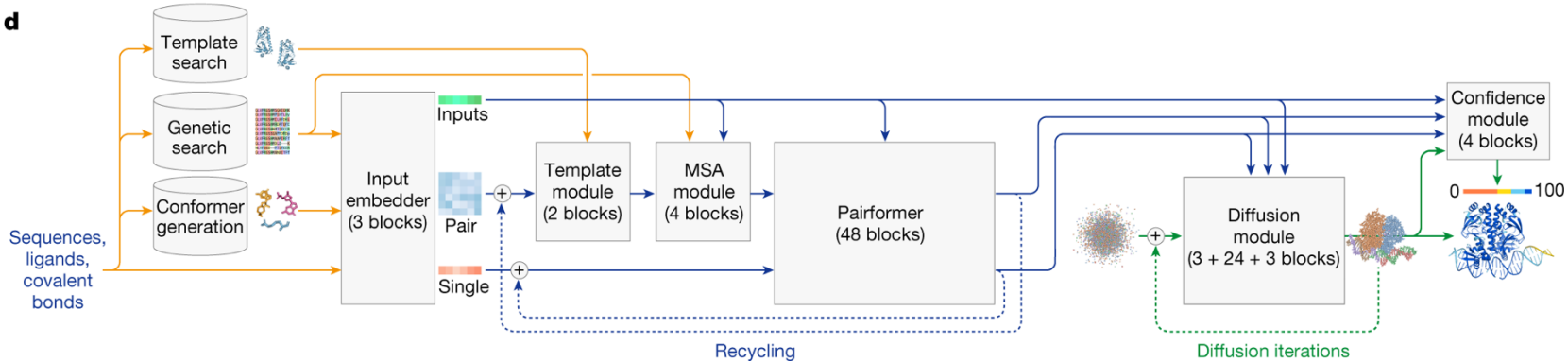
Here, KNN is nested into a Deep Neural Network.

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

AlphaFold for protein structure prediction

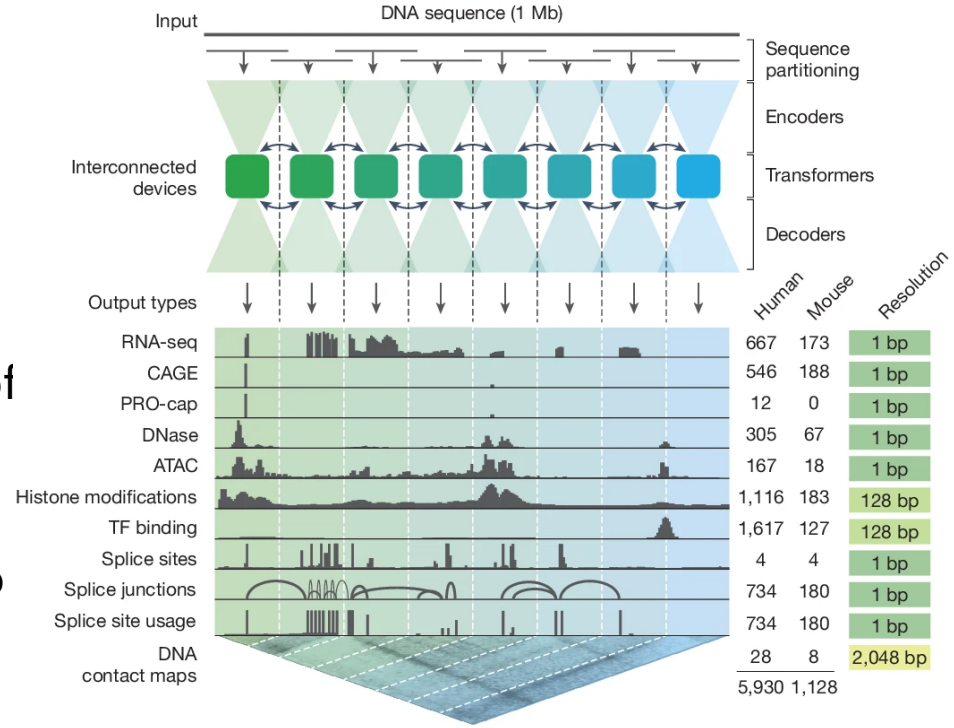


- AlphaFold is based on deep learning including templates and evolutionary conserved sequences
- Performs bootstrapping ranking its predicted models
- Works well for data of complex data

Abramson J, *et al.*, Nature (2024)

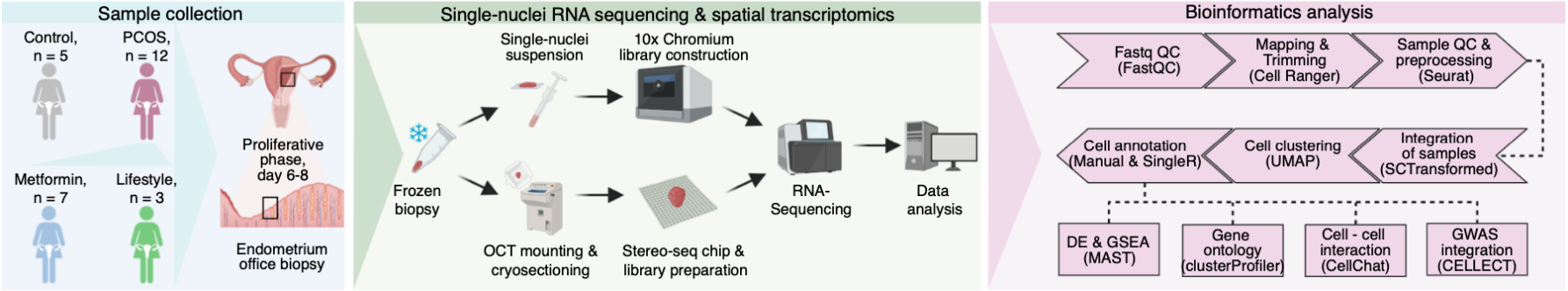
AlphaGenome, predicting functional genomic measurements from a DNA sequence

- 1 megabase DNA can be used to predict and output multiple modalities
- Can be applied in clinical data, recapitulate mechanisms of genes of interest
- Similar deep learning architecture to those of GPT's (Generative AI)



Avsec, Ž., Latysheva, N., Cheng, J. *et al.* Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026). https://doi.org/10.1038/s41586-025-10014-0

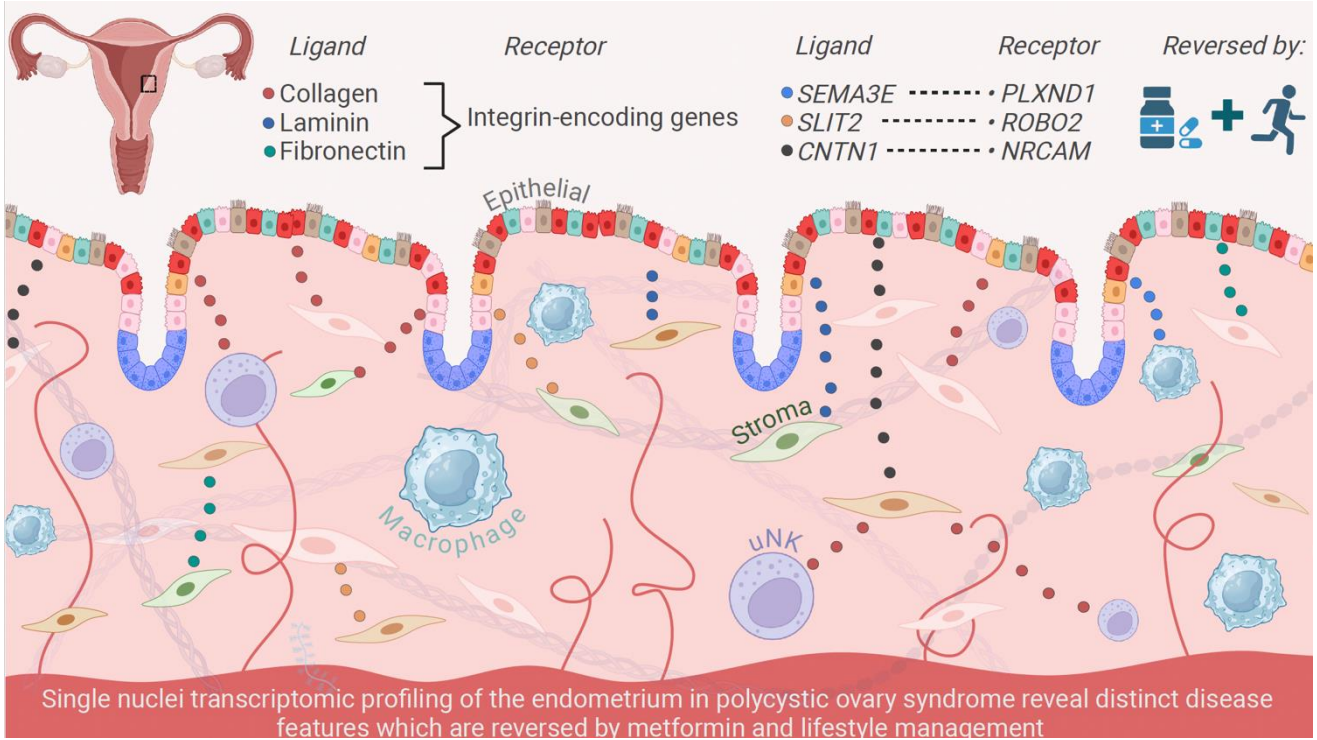
Single-Cell Profiling of human tissue in Polycystic Ovary Syndrome



- We collect endometrium, fat and muscle biopsies from women with and without PCOS
- New samples after 16 weeks of lifestyle management and metformin treatment
- Our group performs snRNA-seq, spatial RNA-seq and proteomics analysis

Eriksson, G., Li, C., Sparovec, T.G. *et al.* Single-cell profiling of the human endometrium in polycystic ovary syndrome. *Nat Med* **31**, 1925–1938 (2025). https://doi.org/10.1038/s41591-025-03592-z

Single-cell profiling of the human endometrium in polycystic ovary syndrome



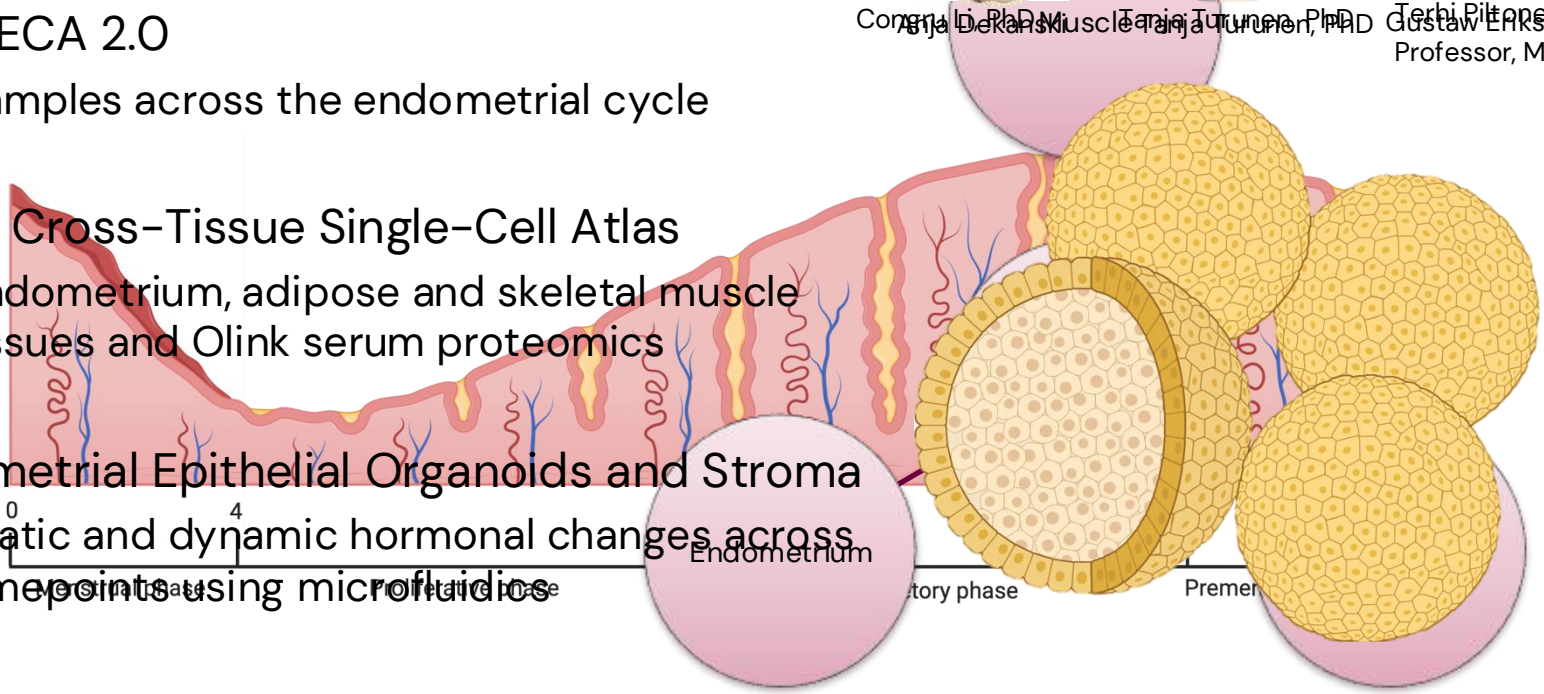
SCAN ME

Eriksson, G., Li, C., Sparovec, T.G. *et al.* Single-cell profiling of the human endometrium in polycystic ovary syndrome. *Nat Med* **31**, 1925–1938 (2025). https://doi.org/10.1038/s41591-025-03592-z

Current projects

Congru Li, PhD
Tanja Turunen, PhD
Terhi Piltonen
Professor, MD

- The PECA 2.0
  - → Samples across the endometrial cycle
- PCOS Cross-Tissue Single-Cell Atlas
  - → Endometrium, adipose and skeletal muscle tissues and Olink serum proteomics
- Endometrial Epithelial Organoids and Stroma
  - → Static and dynamic hormonal changes across timepoints using microfluidics

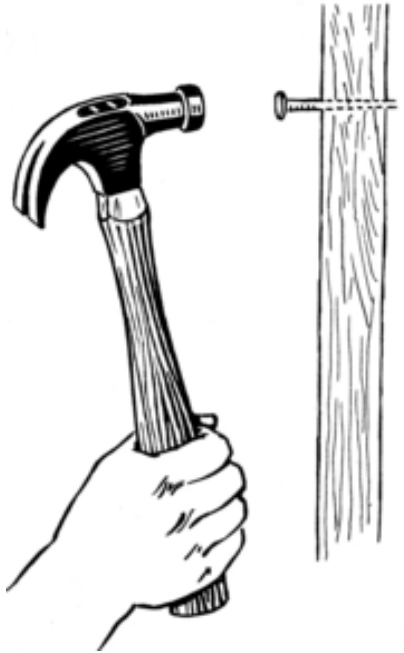


DESeq2 analysis of bulk RNA-seq

A short demonstration and workshop

DESeq2 tutorial

The different methods we have learned today are tools.
Let's apply the tools to answer our research questions



DESeq2 tutorial

- Data transformation with variance stabilizing transfortion (VST)
  - → Make the data suitable for PCA and clustering
- PCA
  - → Quality control, visualise sample relationships and check for batch effects
- Hierarchical clustering
  - → On gene expression heatmaps to identify clustering of genes and samples
- Statistical testing with DESeq2 for differentially expressed genes
  - → Based on generalised linear models
- Visualisation of the results

DESeq2, a brief introduction

- Most used method in analysing bulk RNA-sequencing data
- Other methods are limma and edgeR. Commom 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

DESeq2 tutorial

Please go to:
https://github.com/GustawEriksson/Introduction_Machine_Learning_in_Bioinformatics

If interested in our groups on-going bioinformatic projects:
https://github.com/ReproductiveEndocrinologyMetabolism

Karolinska
Institutet