3000788 Intro to Comp Molec Biol

Lecture 24: Machine learning in biology

Fall 2025

Sira Sriswasdi, PhD

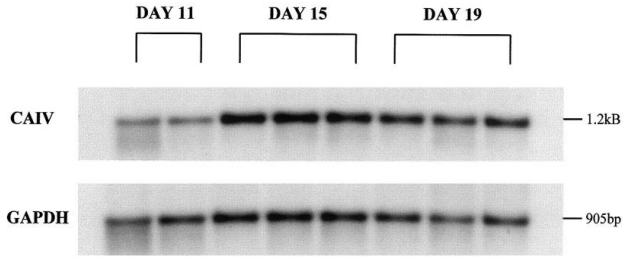
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

- Digital and data transformation of biology
- Improving bioinformatics with machine learning
- Knowledge discovery with machine learning

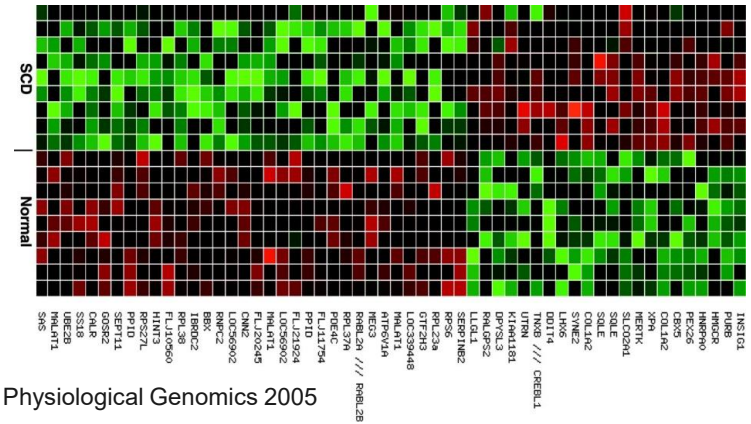
Data transformation of biology

Qualitative experiment



Rosen et al. Am J of Physiology 2001

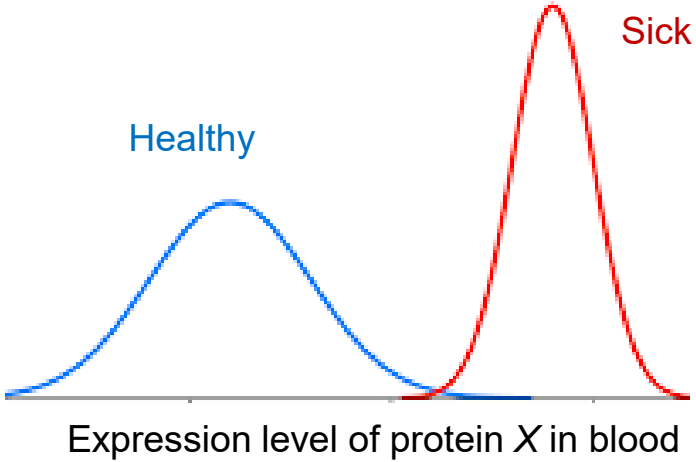
High-throughput, quantitative data



Klings et al. Physiological Genomics 2005

- Not just gene A is up-regulated, but **genes A is up-regulated by 2.36 folds with standard deviation of 0.18 across 12 biological replicates**
- Biology has become quantitative

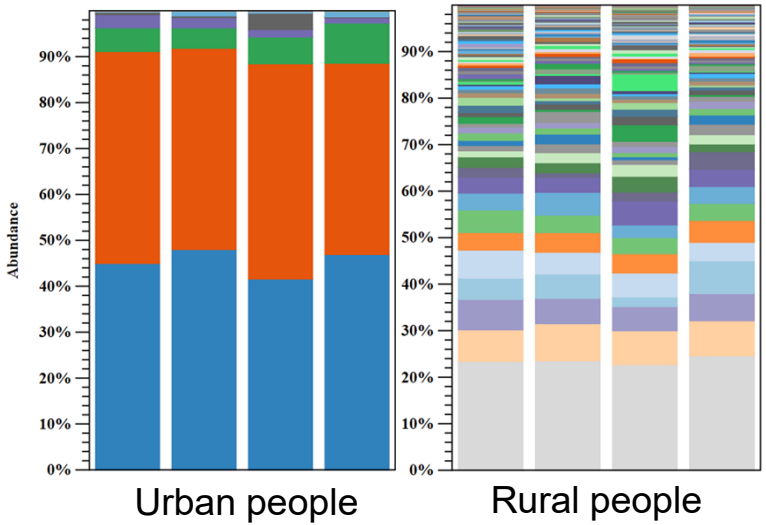
Quantitative thinking



But how much difference is high enough

- How would you to quantify the ability of *X* to distinguish sick patients?
  - How about $\text{Score}(X) = \text{Mean}_1 - \text{Mean}_2$ or $\text{Abs}(\text{Mean}_1 - \text{Mean}_2)$?
  - How about $\text{Score}(X) = \frac{\text{Mean}_1 - \text{Mean}_2}{\sqrt{\frac{1}{n}(\text{Variance}_1 + \text{Variance}_2)}}$?

Turning verbal description into mathematical formula

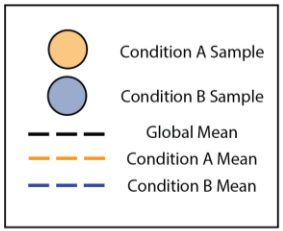


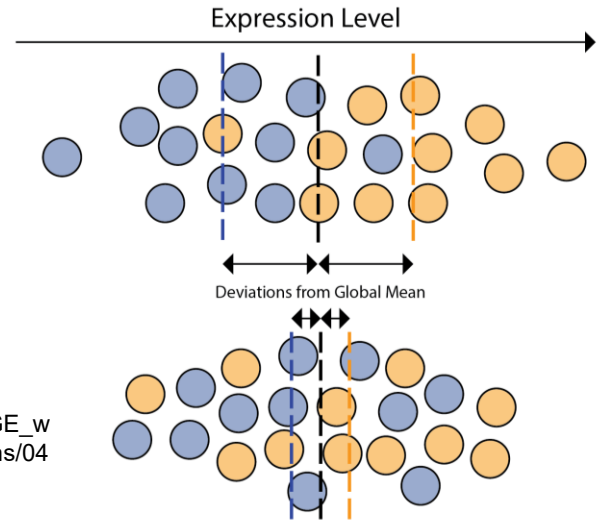
Colors indicate different microbial taxa in gut microbiome

But how much entropy is “high”

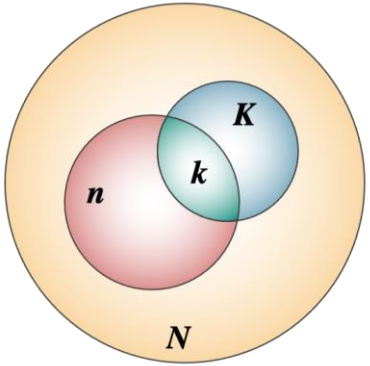
- How would you quantify the diversity of microbiome?
  - Number of different taxa = n
  - Let p_1, …, p_n be taxa frequency, how should they define diversity?
  - **Entropy** = $-p_1 \log_2(p_1) - p_2 \log_2(p_2) - \cdots - p_n \log_2(p_n)$

Statistical framework provides objectivity





https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/04a_design_formulas.html



$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

N = Background *(e.g. Transcriptome ~ 40.000 genes)*

n = Query list *(e.g. upregulated genes)*

K = genes annotated in the pathway/set tested *(e.g. Glycolysis)*

$k = n \cap K$

- Turn subjective fold-differences into objective statistical significances
- P-value, false discovery rate, etc.

Statistics alone is not enough

- Statistics help you assess the significance of an observation, after you have calculated some scores

- It doesn’t help you calculate the score itself
  - Are two protein structures similar?
  - Do two genes have similar sequences?
  - Does the drug target the immune system?

- We need algorithm

Dynamic programming for sequence alignment

	GAP	A	T	G	C	T
GAP	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
G	-4	-1	0	0	-2	-4
C	-6	-3	-2	-1	1	-1
T	-8	-5	-2	-3	-1	2

Match : 1
Mismatch : -1
GAP : -2

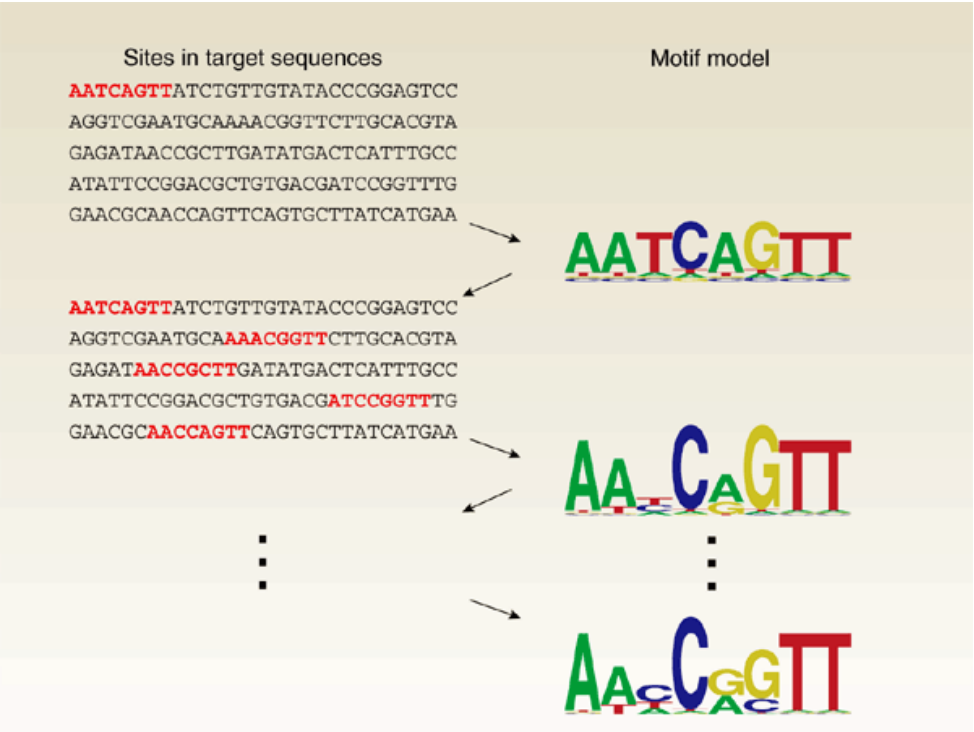
```
Seq1:  ATGCT
       |  |||
Seq2:  A-GCT
```

Use statistics to interpret if +2 is good enough

https://www.youtube.com/watch?v=ipp-pNRIp4g

Motif discovery algorithm

Use statistics to interpret the quality of the final motif



- Guess a motif (fixed length)
- Find the best match in each sequence

- Update motif nucleotide profile
- Search for (possibly better) match in each sequence

- Repeat the two steps until convergence

D'haeseleer, P. Nature Biotechnology 24:959-961 (2006)

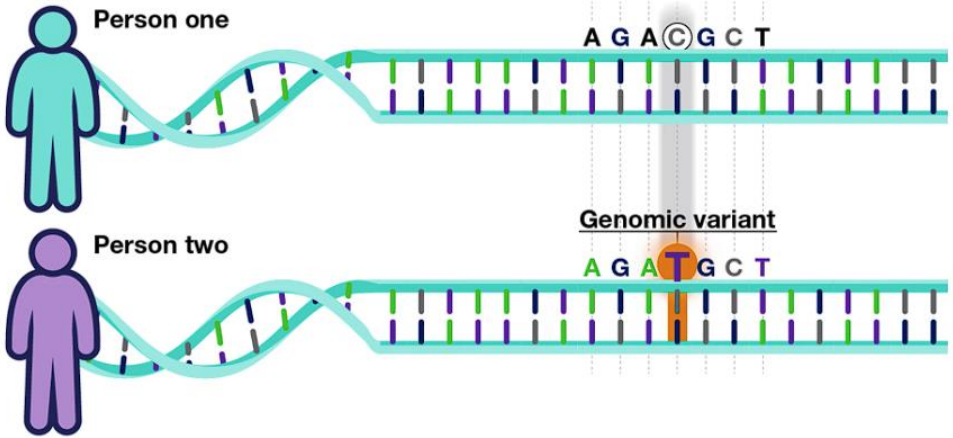
Synergy between algorithm and statistics

- Algorithm identifies the best possible answer in your data
  - Aligned portion of sequences
  - DNA motifs

- Statistics model the distribution of the scores and provides objective significance assessment of the best answer

The need for machine learning

Human judgment in bioinformatics

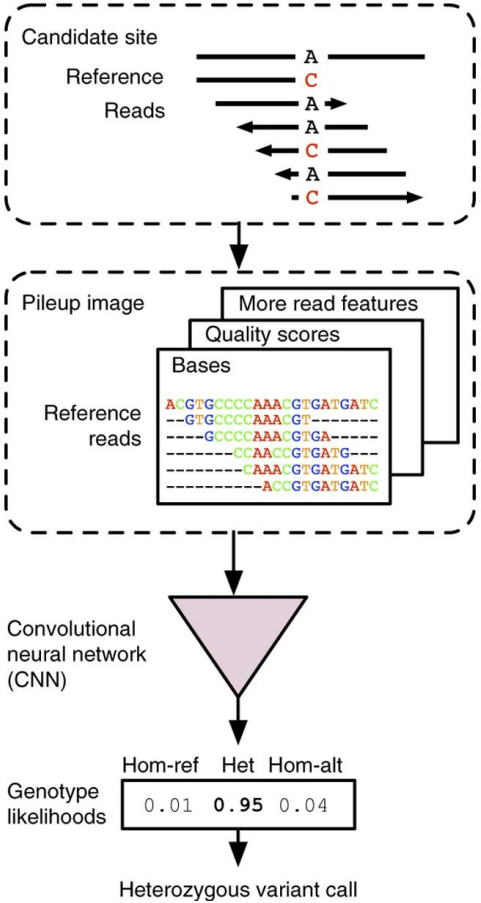


https://storymd.com/journal/4m8ald6ipw-gene-variants-and-health/page/nrq7zt7bry-what-is-a-gene-variant-and-how-do-variants-occur

- When to focus on a mutation?
- Many subjective filters:
  - Base quality
  - Read depth
  - Allele frequency (AF)
  - Population AF
  - Coding or non-coding
- Each lab has different criteria

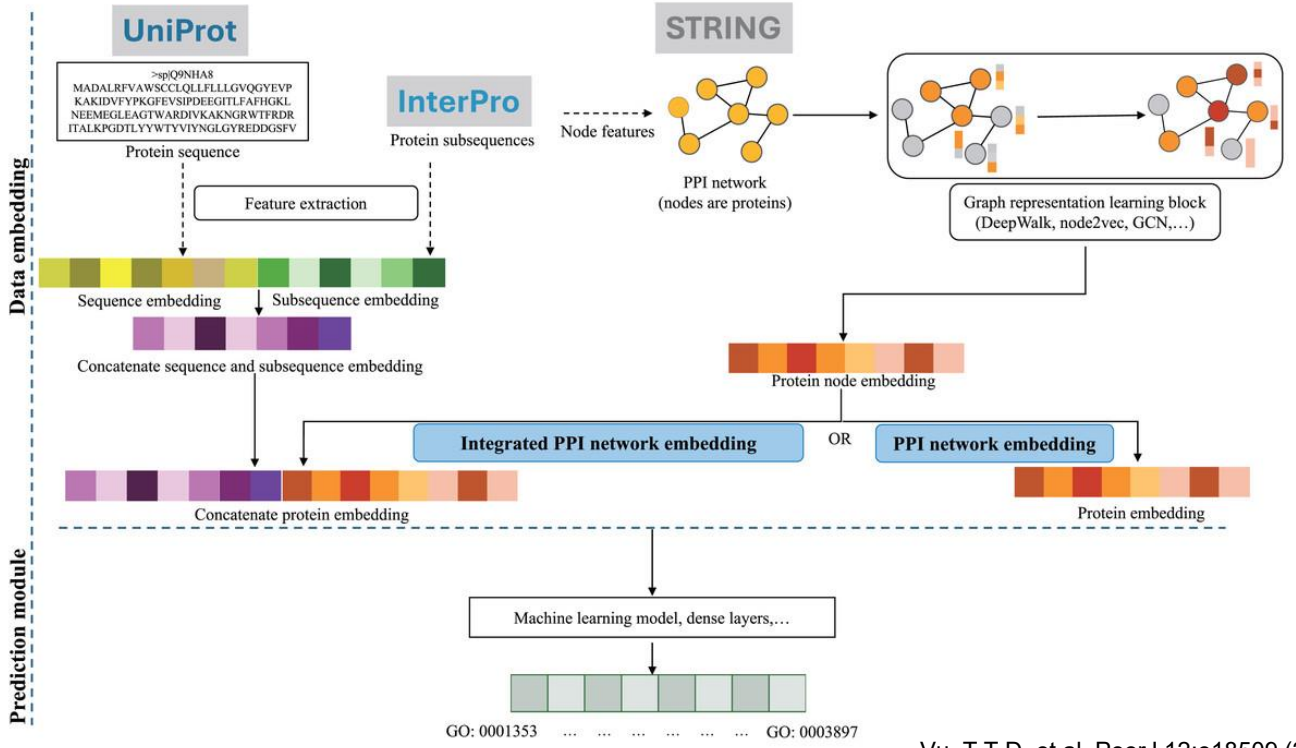
Machine learning as objective criteria

- Pain points:
  - No theoretical model for scoring variants
  - Human cannot interpret multiple scores
- What can be done?
  - Collect data from samples with known mutations
  - **Train ML model to distinguish true variants**
- **Balancing act**: Which parts to offload to ML?
  - The whole pipeline
  - Combine scores from multiple tools



Poplin, R. et al. Nature Biotech. (2018)

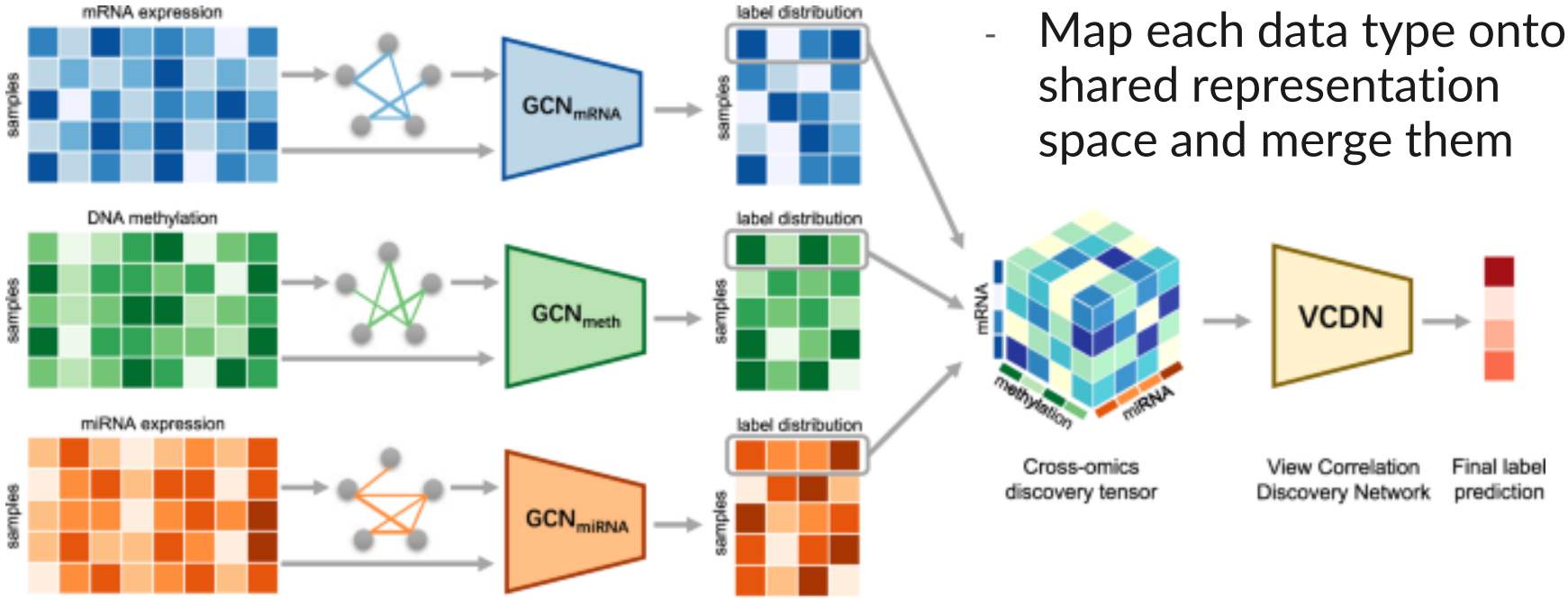
Machine learning integrates data types



- Condition the learning to **make representations from multiple raw data types computationally compatible**
- Additive, concatenate, etc.

Vu, T.T.D. et al. PeerJ 12:e18509 (2024)

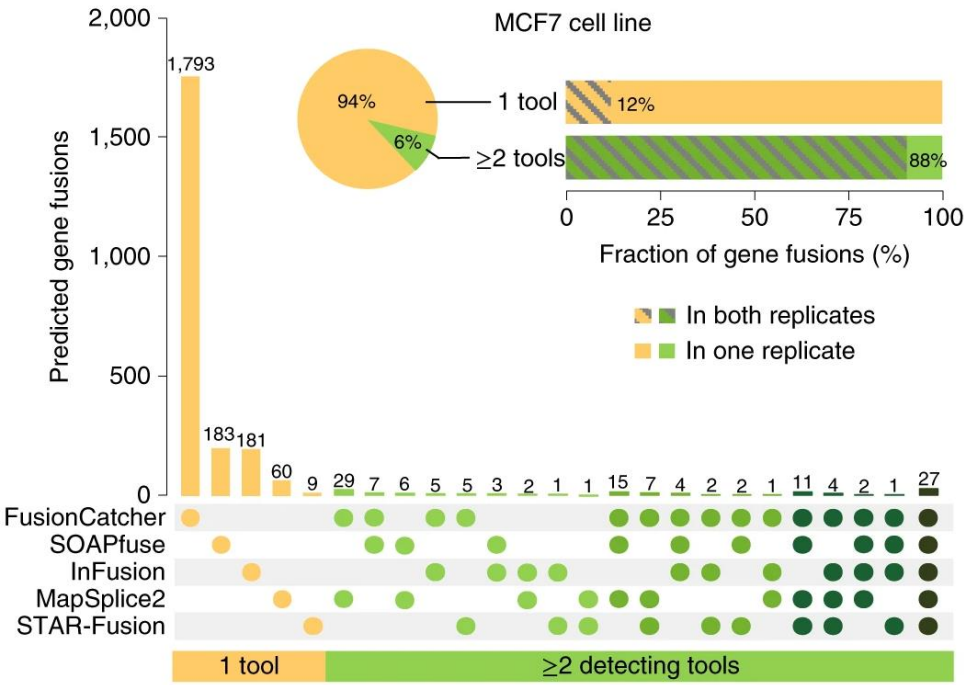
Multi-omics integration with ML

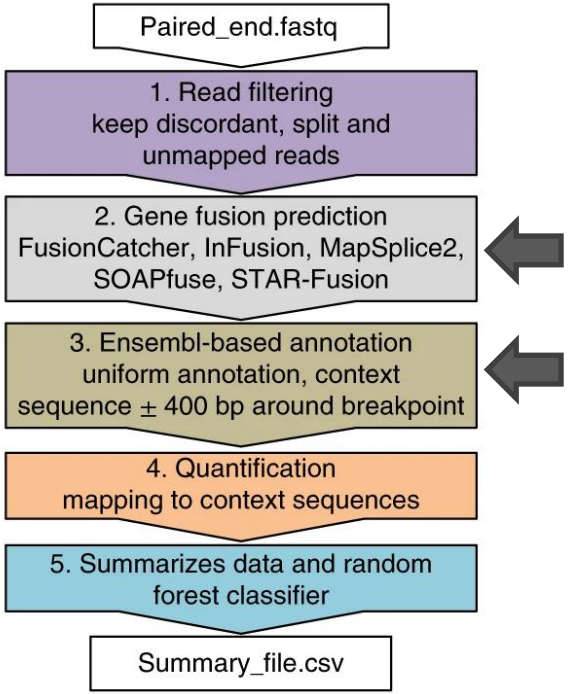


- Map each data type onto a shared representation space and merge them

Wang, T. et al. Nature Communications 2021

Machine learning aggregate bioinformatics tools





Weber, D. et al. Nature Biotechnology 40:1276-1284 (2022)

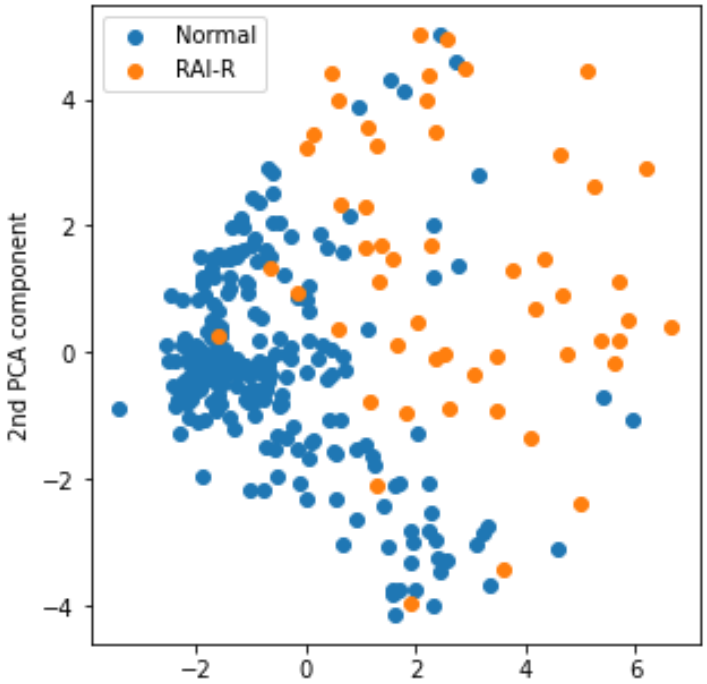
Synergy between ML and bioinformatics

- Different bioinformatics algorithms produce different mistakes
- ML can learn to identify when to trust each algorithm from the data and confidence scores
- ML can speed up bioinformatics
  - Multiple sequence alignment: identify pairs of sequences to align first
  - Protein modeling: directly identify structure models

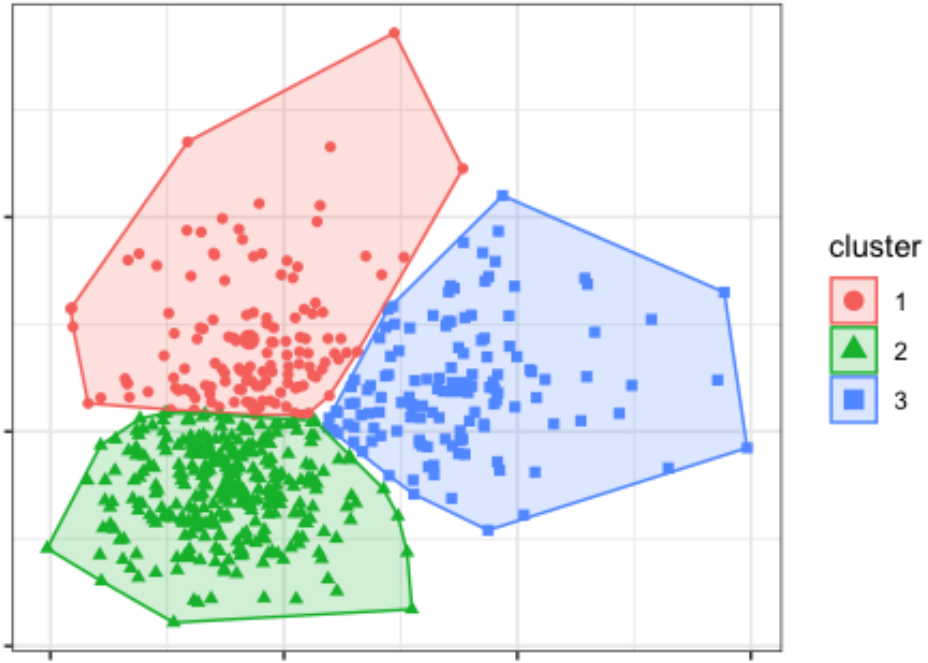
Knowledge discovery with unsupervised ML

Key unsupervised learning techniques

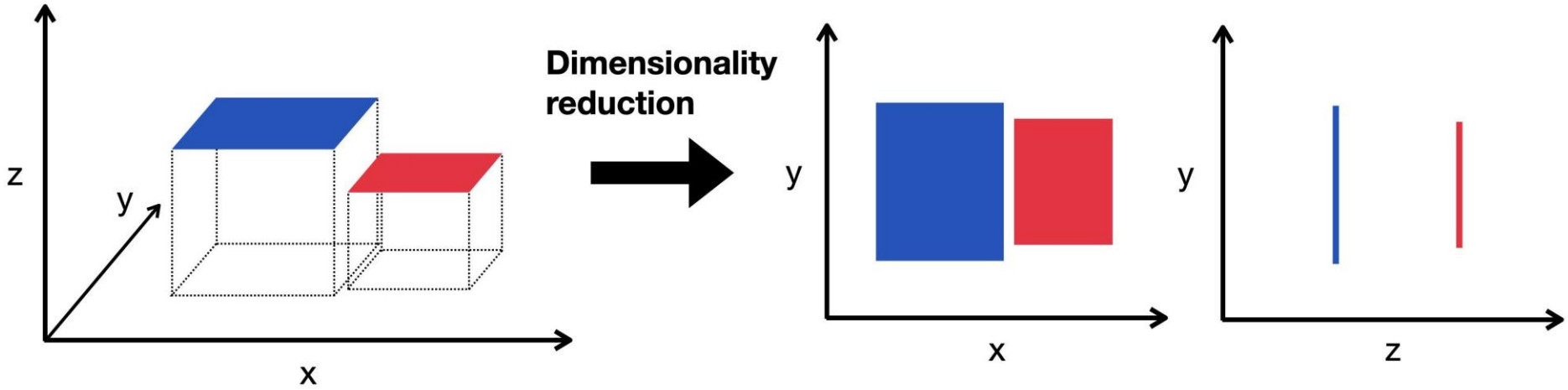
Dimensionality Reduction



Clustering / Anomaly Detection



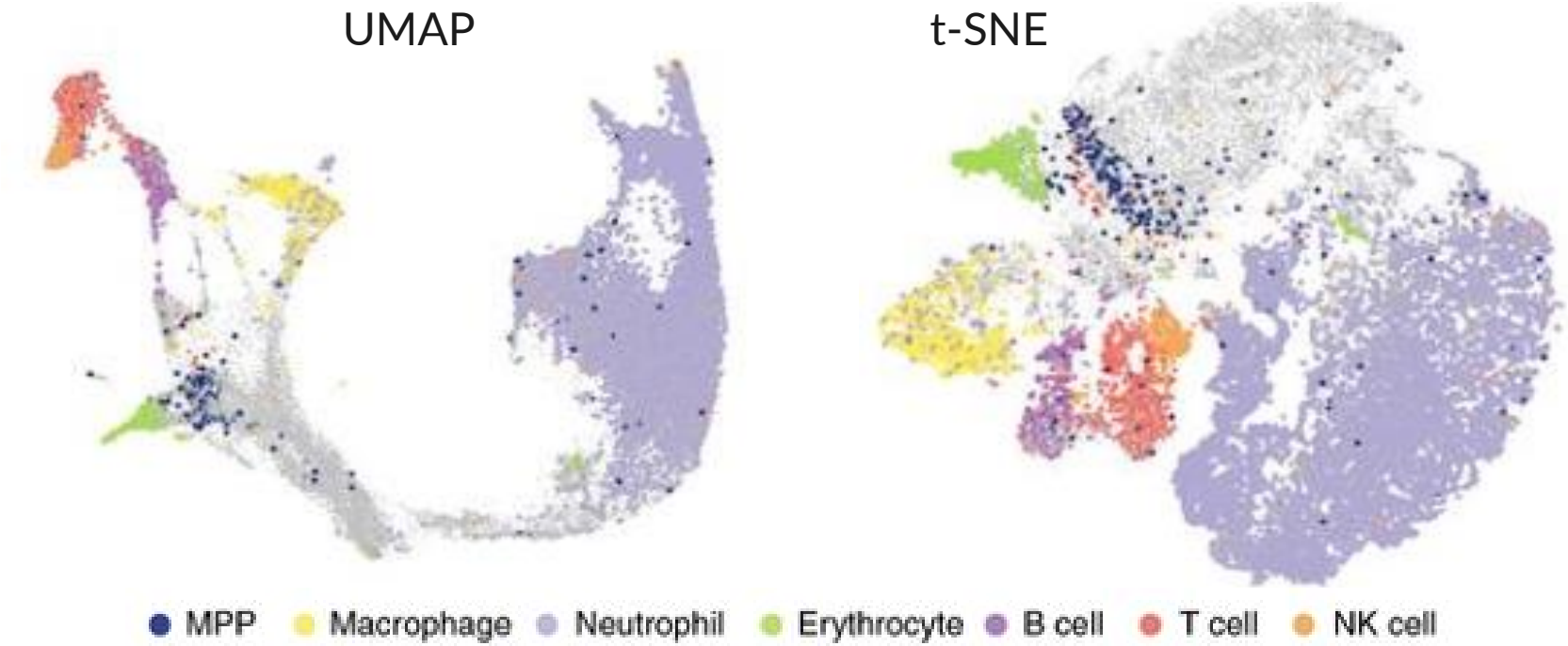
Dimensionality reduction



https://www.sc-best-practices.org/preprocessing_visualization/dimensionality_reduction.html

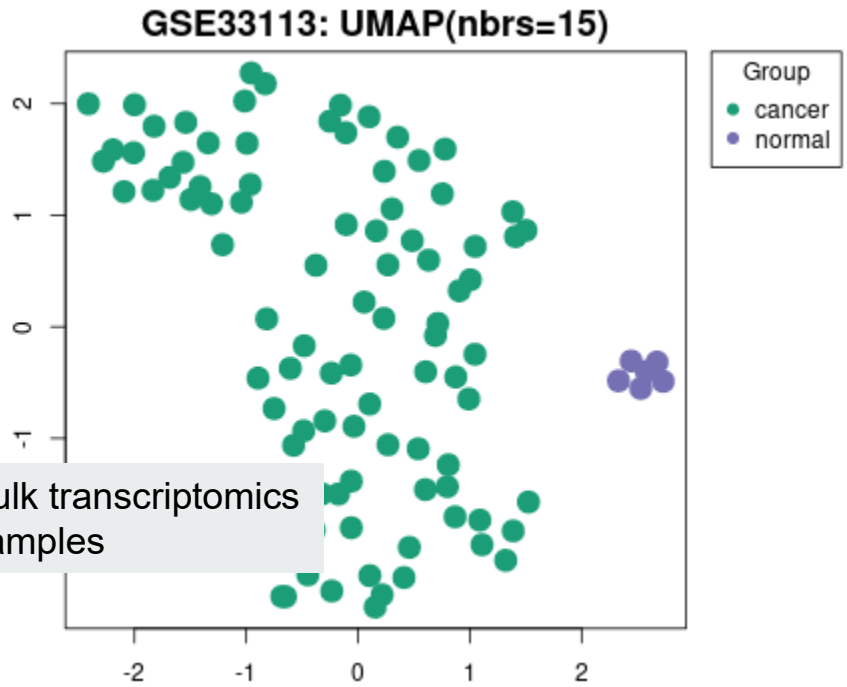
- Reduce dimension (number of features) while maintaining information
- Patient with similar symptoms also exhibit similar lab tests or have similar demographics or similar medical history

Visualization of single-cell data



Becht, E. et al. Nature Biotechnology 37:38-44 (2019)

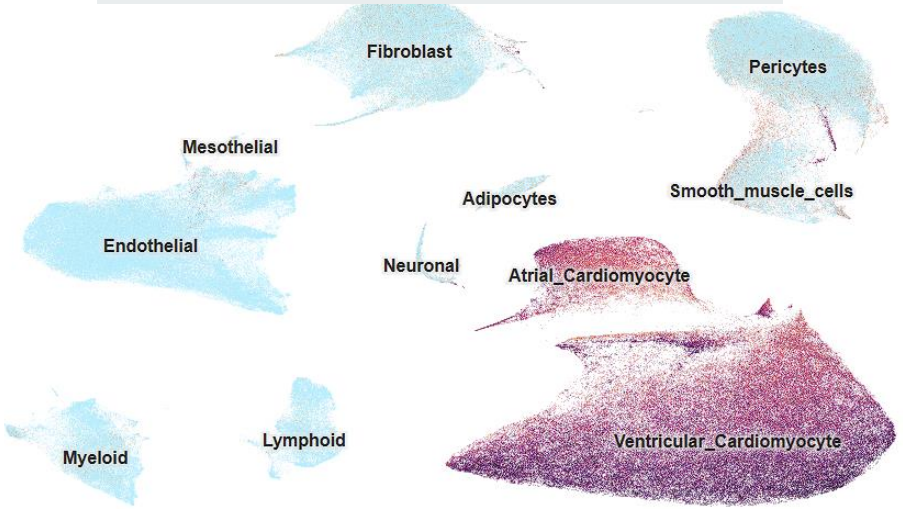
Visualization helps generate/validate hypothesis



Bulk transcriptomics samples

PCA plot from GEO2R

Expression of CTNNA3 in cardiac cells



UCSC Cell Browser

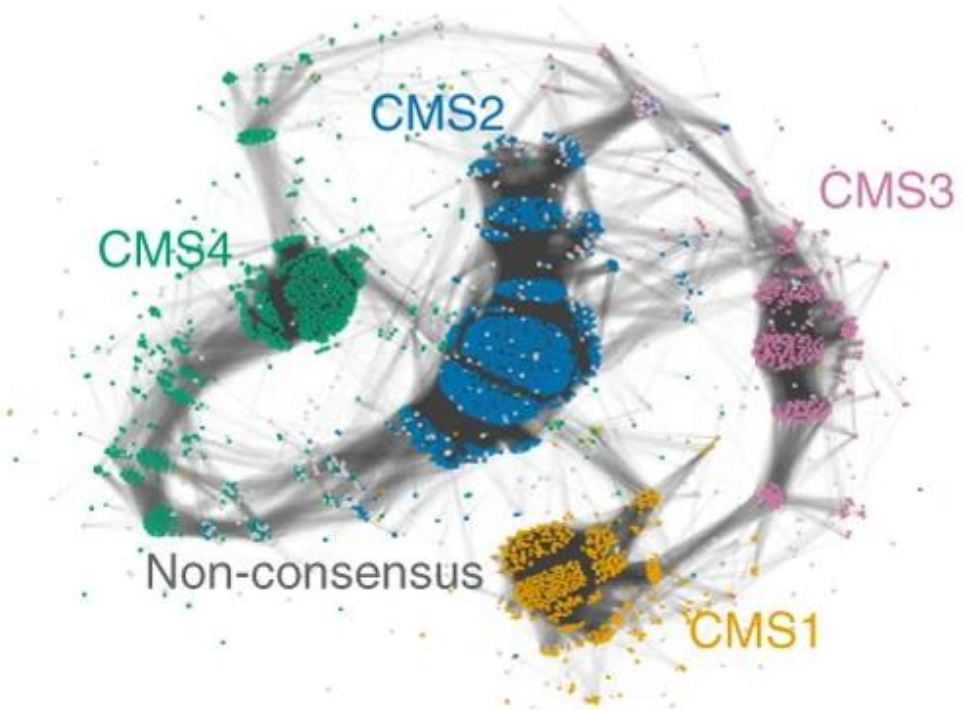
The heart of clustering

- **Goal**: Group similar data point together

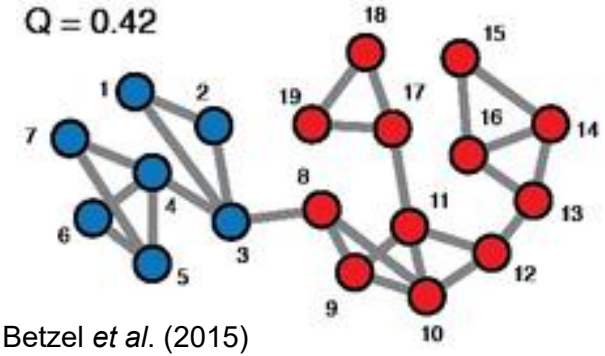
- How to define similarity?
  - Distance: Between two data points
  - Linkage: Between groups of data points

- How many clusters is appropriate?
  - Within-cluster (small) versus between-cluster (large) distance

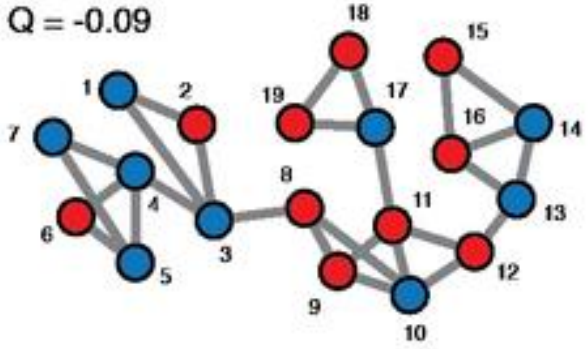
Network clustering with modularity score



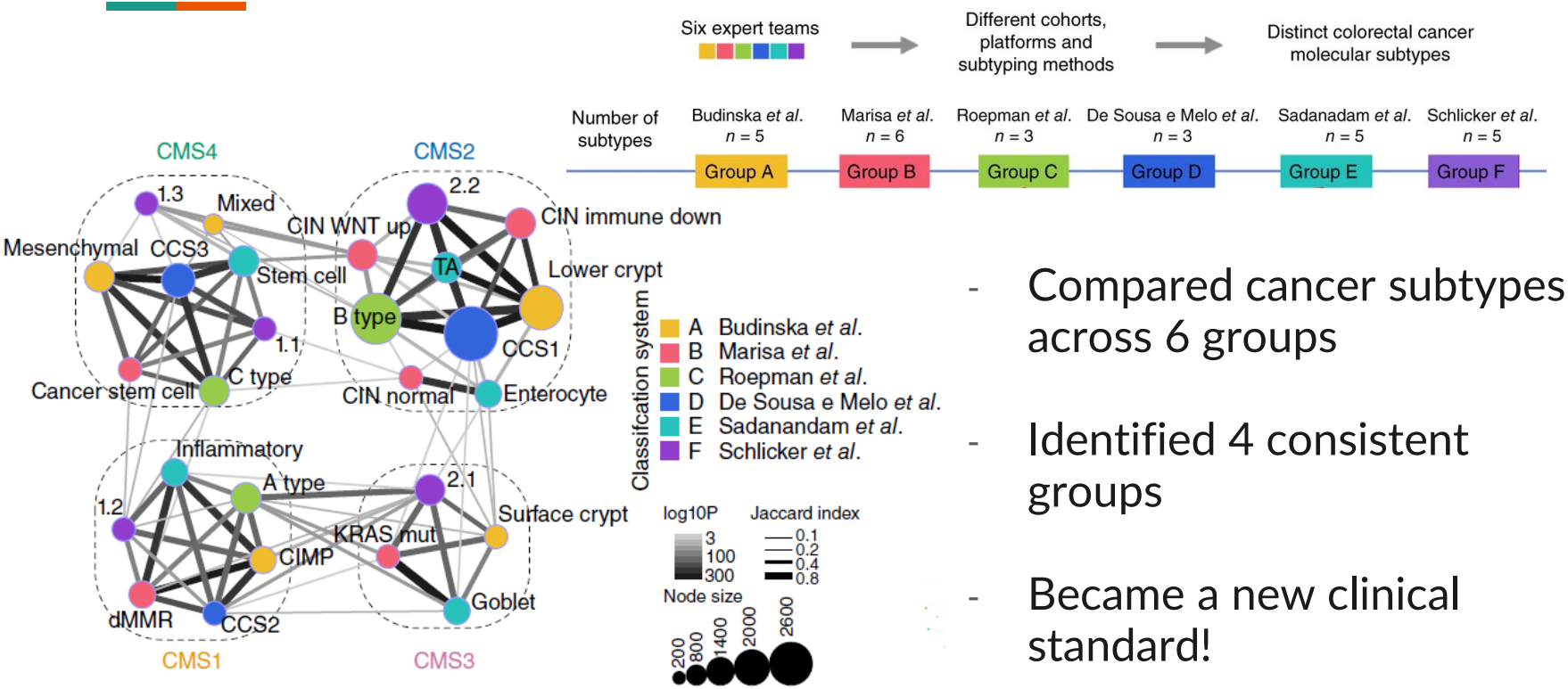
Guinney, J. et al. Nature Medicine 21:1350-1356 (2015)



Betzel *et al*. (2015)



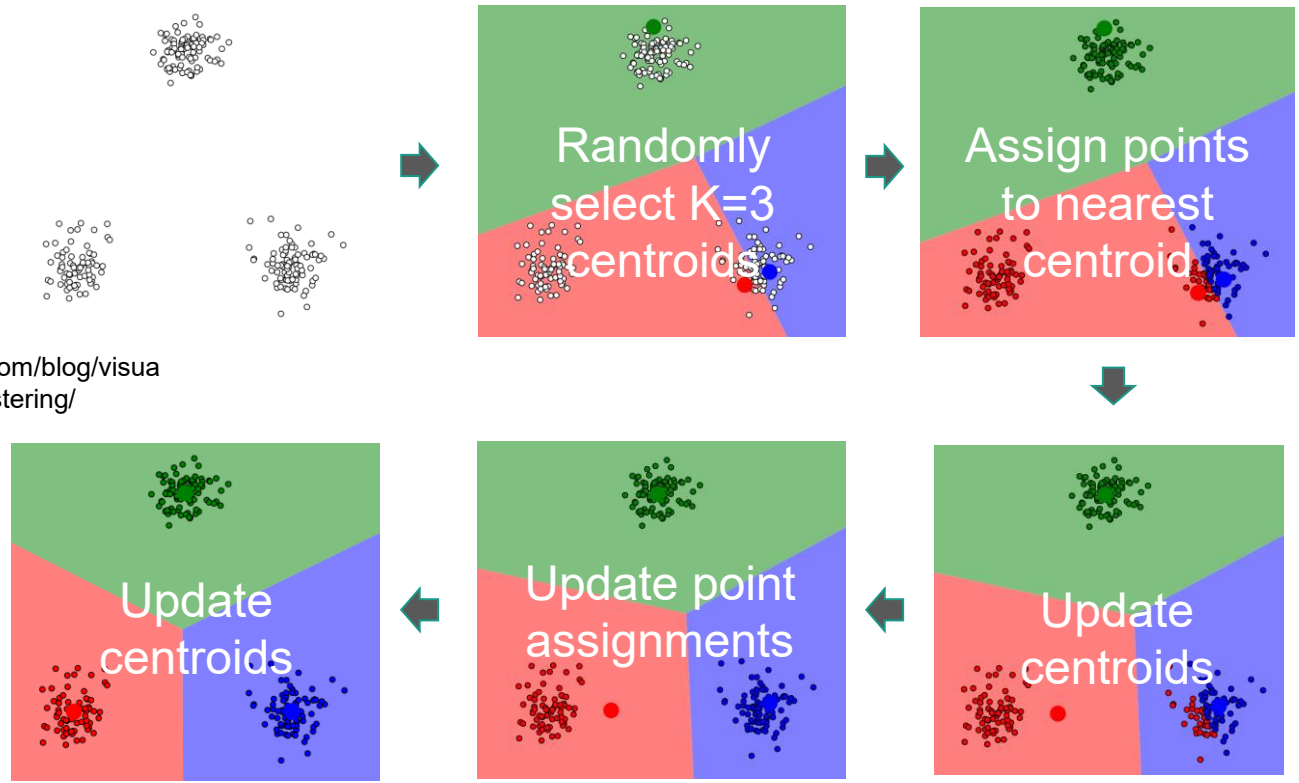
Cancer consensus subtype discovery



- Compared cancer subtypes across 6 groups
- Identified 4 consistent groups
- Became a new clinical standard!

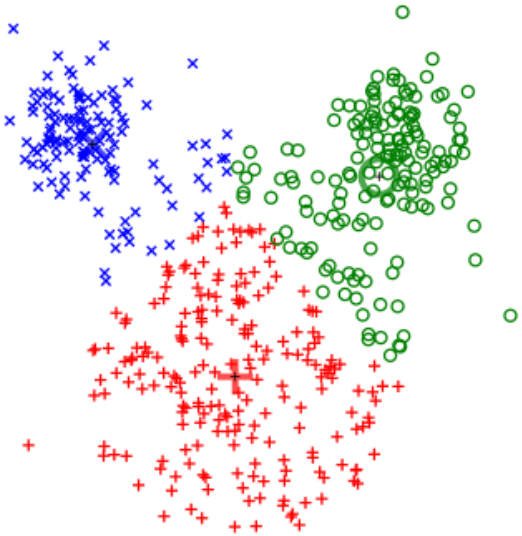
Guinney *et al*. Nat Medicine 21: 1350-1356 (2015)

k-mean: radius-based



www.naftaliharris.com/blog/visualizing-k-means-clustering/

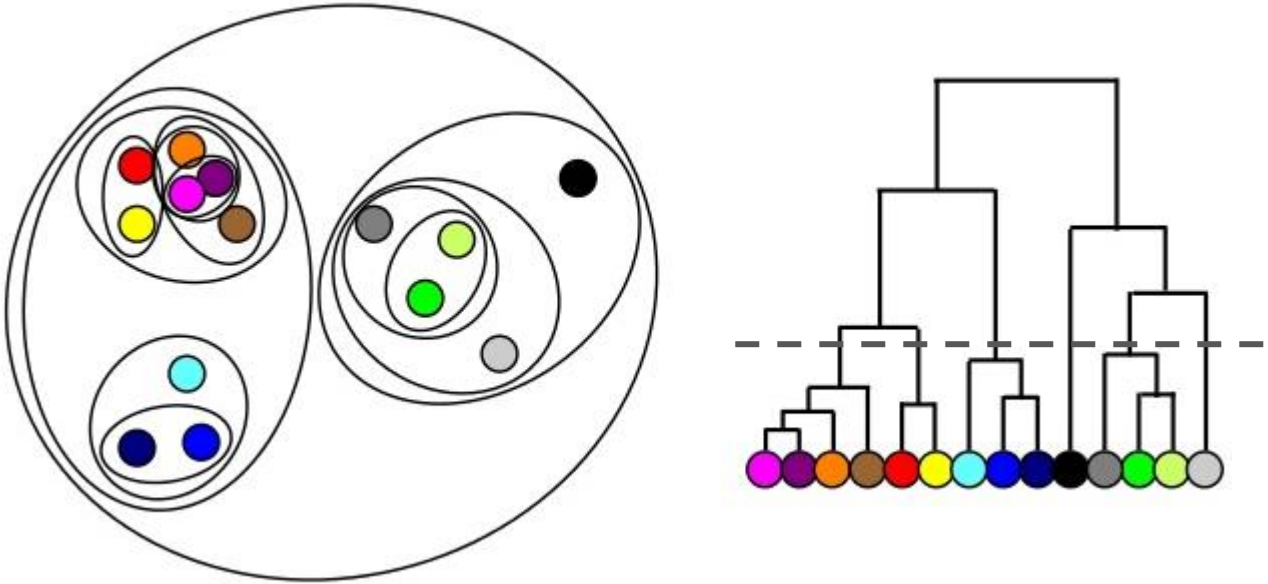
Limitation of *k*-mean



- Assume Euclidean distance
- Assume that clusters are of equal radius
- The initial guess of the locations of *k* means can affect the final clusters
  - Repeat multiple times

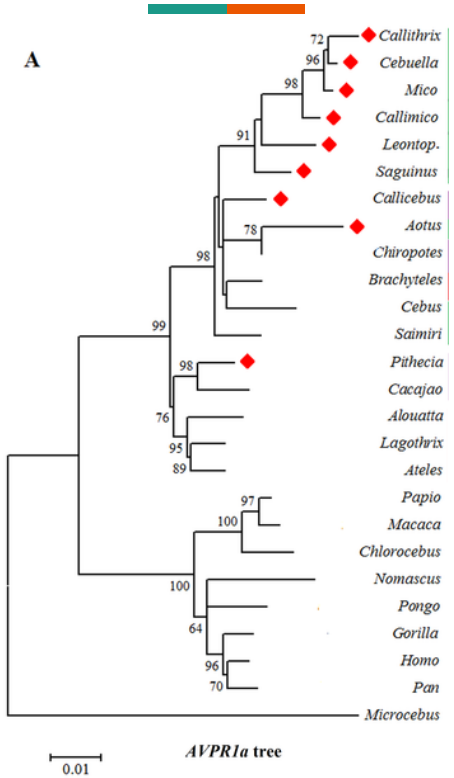
https://en.wikipedia.org/wiki/K-means_clustering

Agglomerative/Hierarchical: neighbor-based

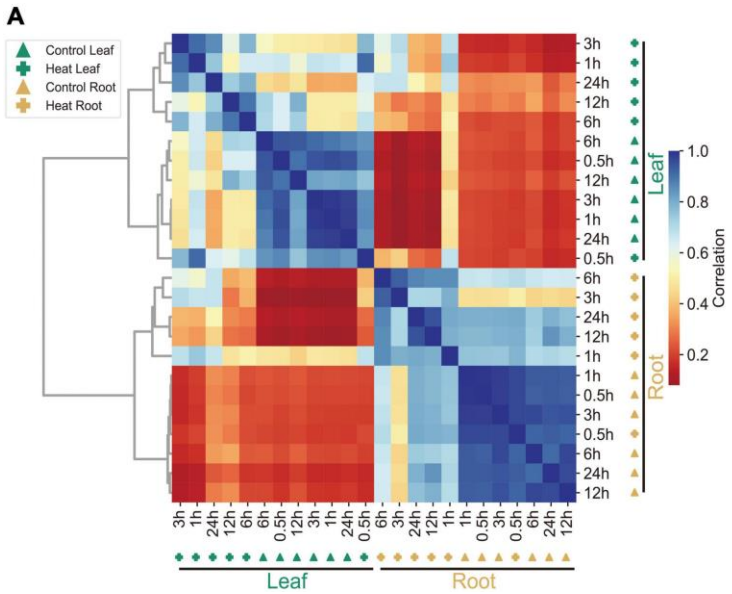


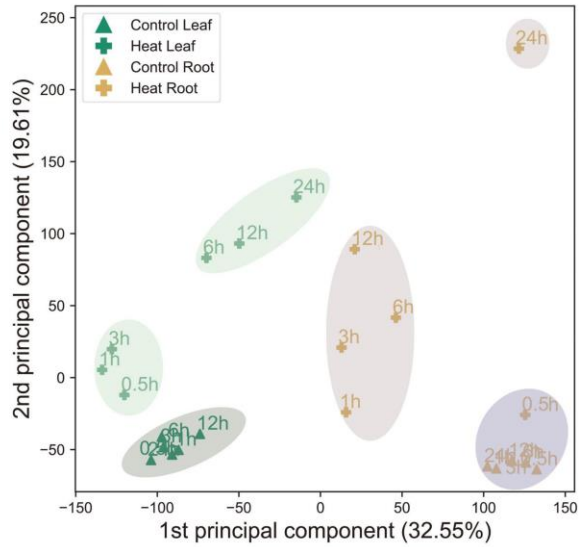
Source: www.slideshare.net/ElenaSgis/data-preprocessing-and-unsupervised-learning-methods-in-bioinformatics

Examples of agglomerative/hierarchical clustering



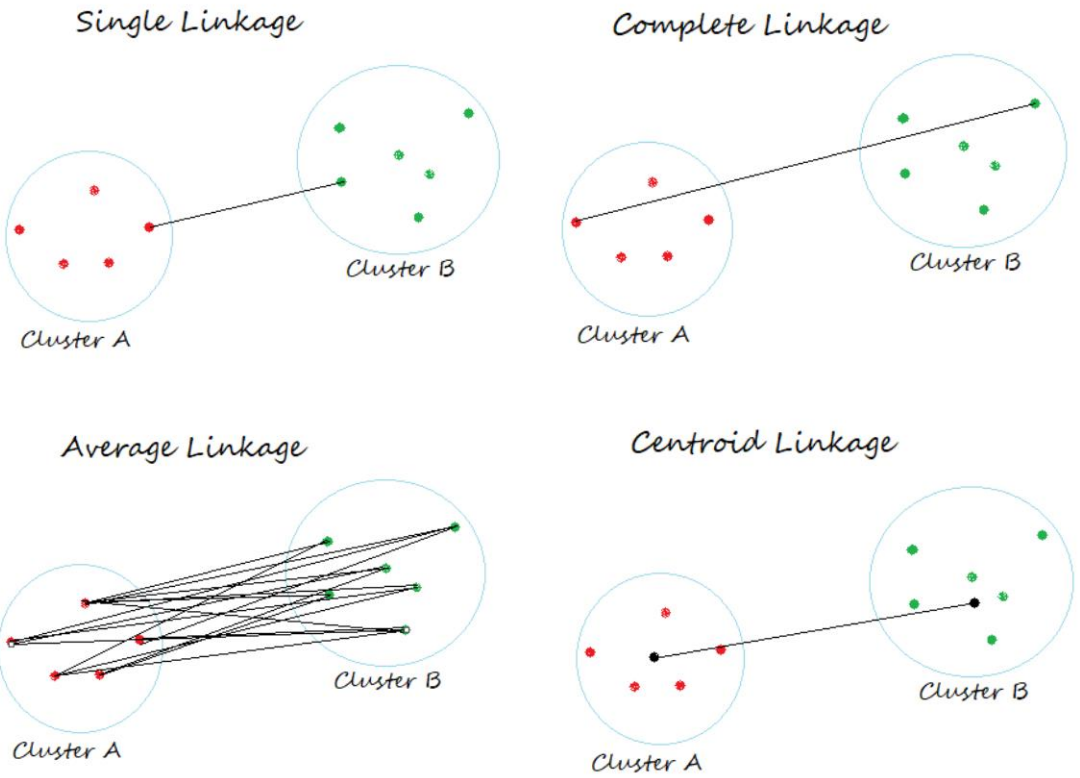
Ren, D. et al. PLoS ONE 9:e222638 (2014)





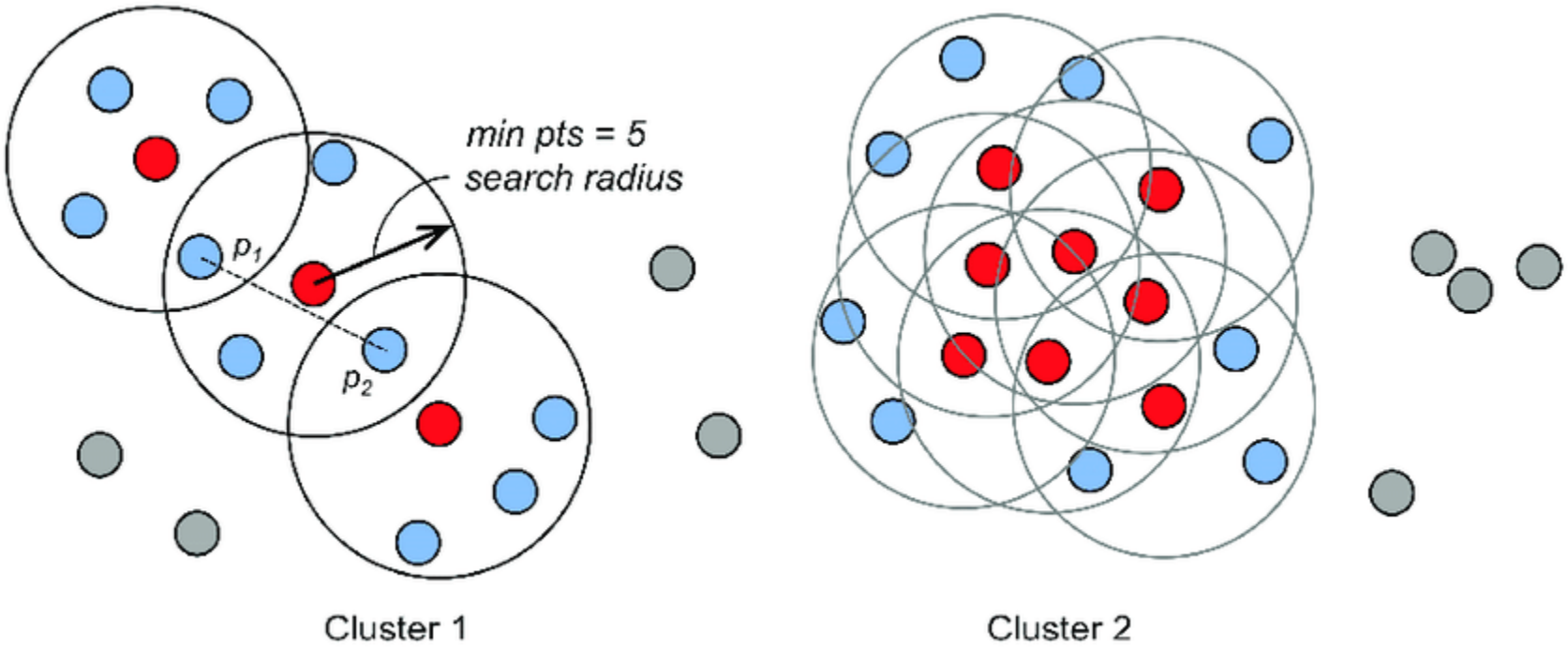
Tang, B. et al. Frontiers in Plant Sciences 13:946475 (2022)

Linkage = distance metric for groups of data points



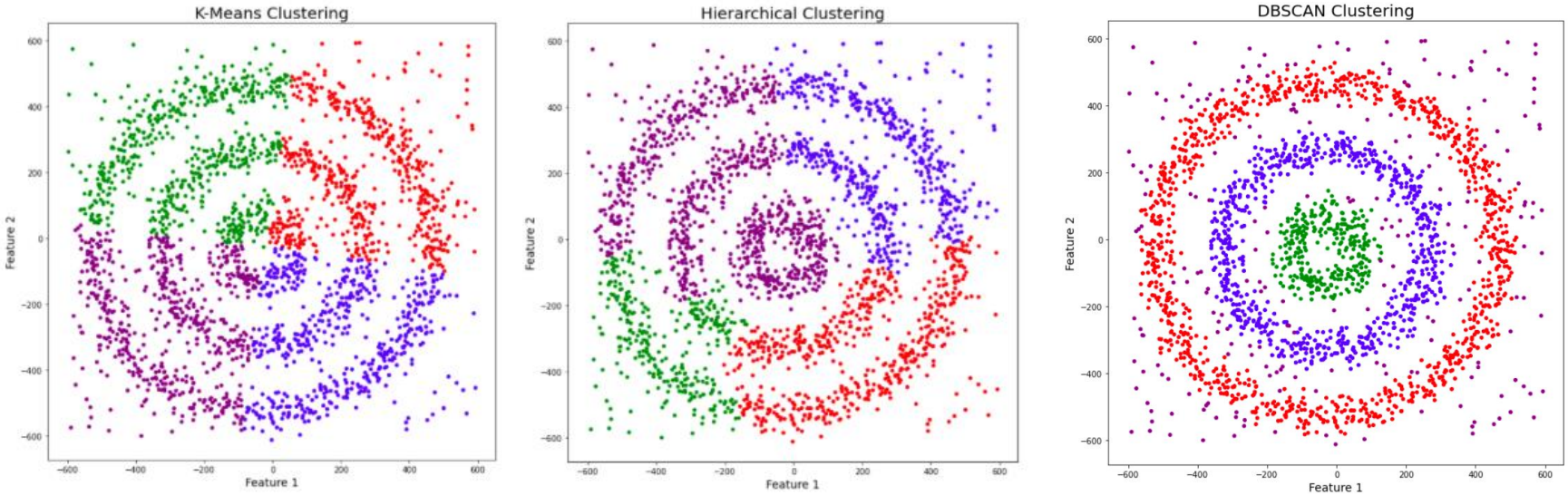
https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/

DBSCAN: density- and connectivity-based



Difrancesco, P.-M. Remote Sensing 12:1885 (2020)

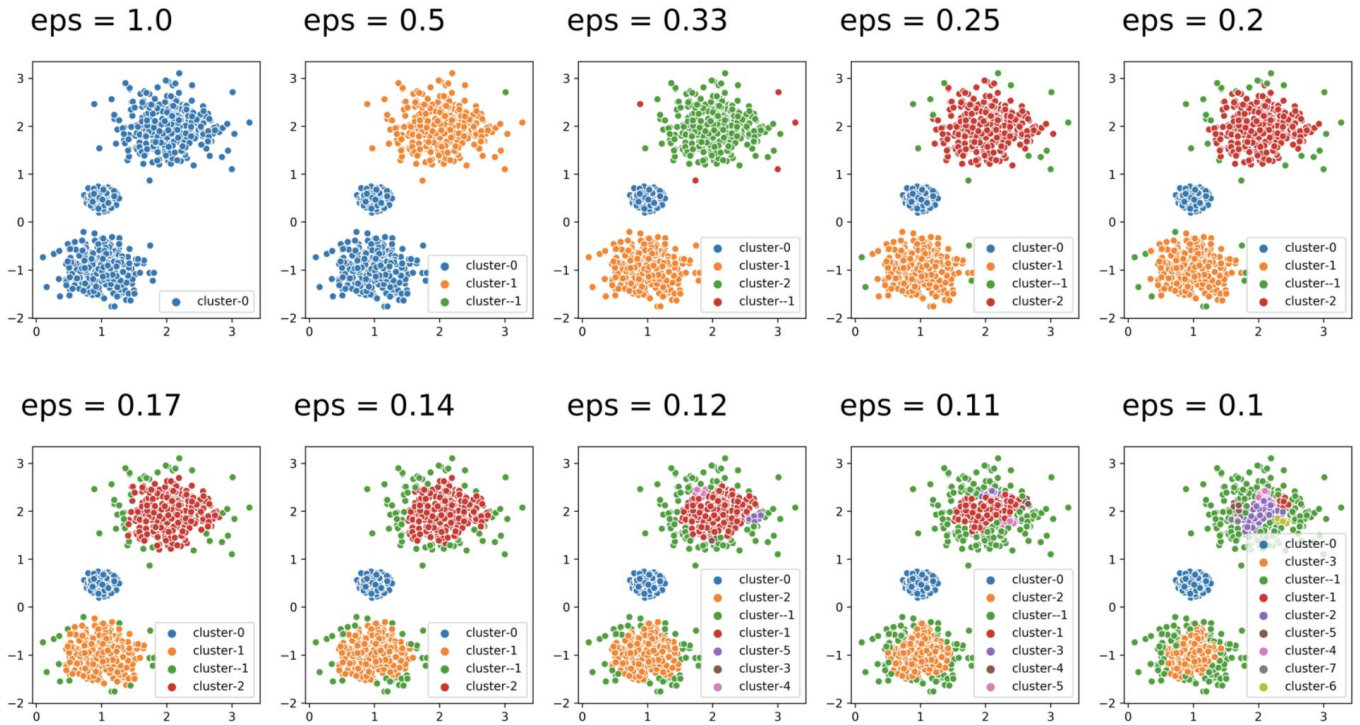
Complex, non-circular clusters



https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

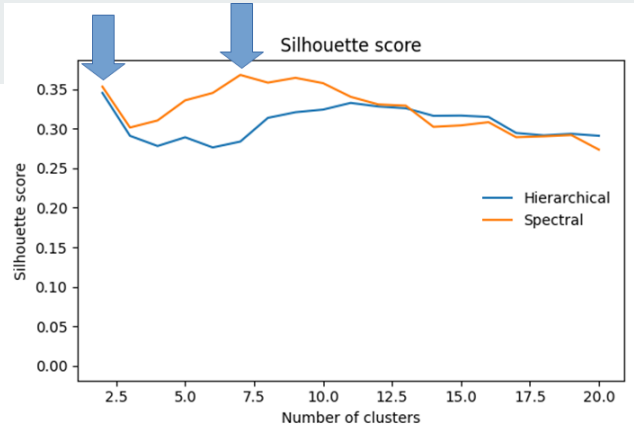
- Distance-based techniques assume that data are spread in all directions

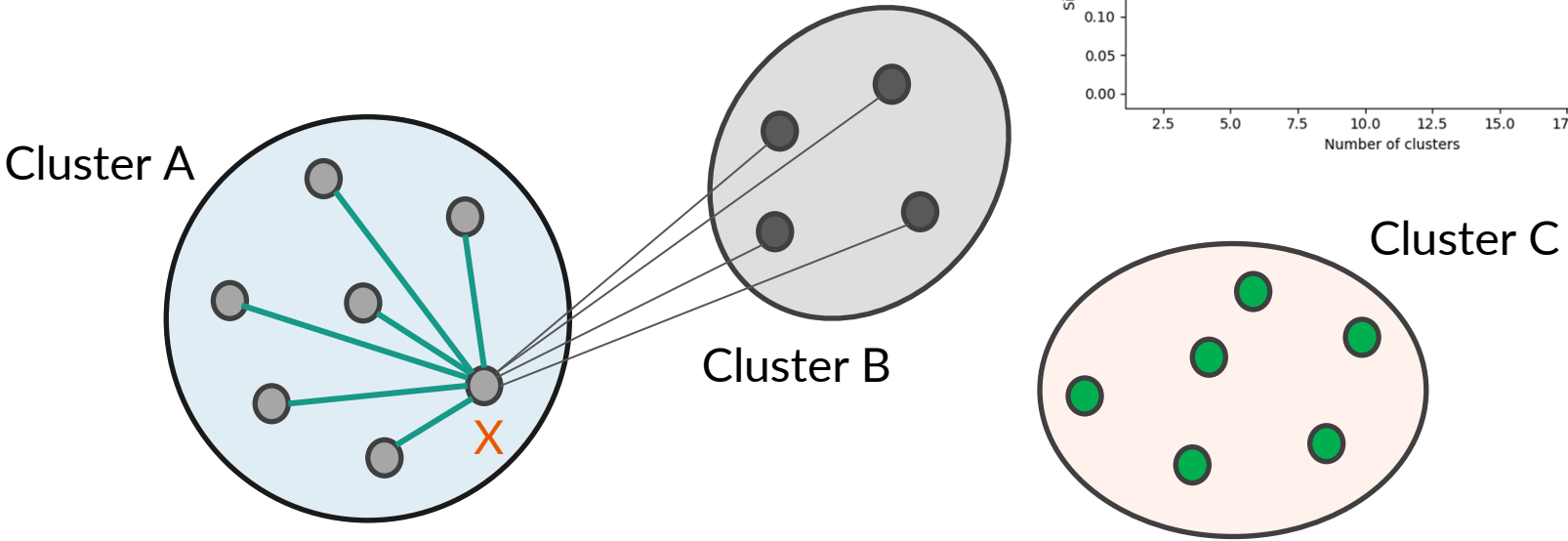
Simultaneous detection of clusters and outliers



https://towardsdatascience.com/

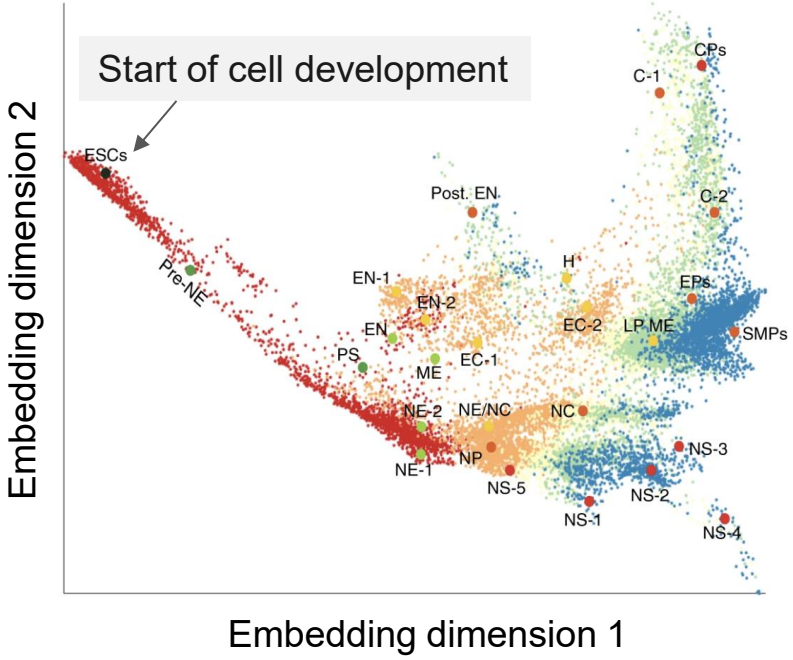
Cluster selection: Silhouette score

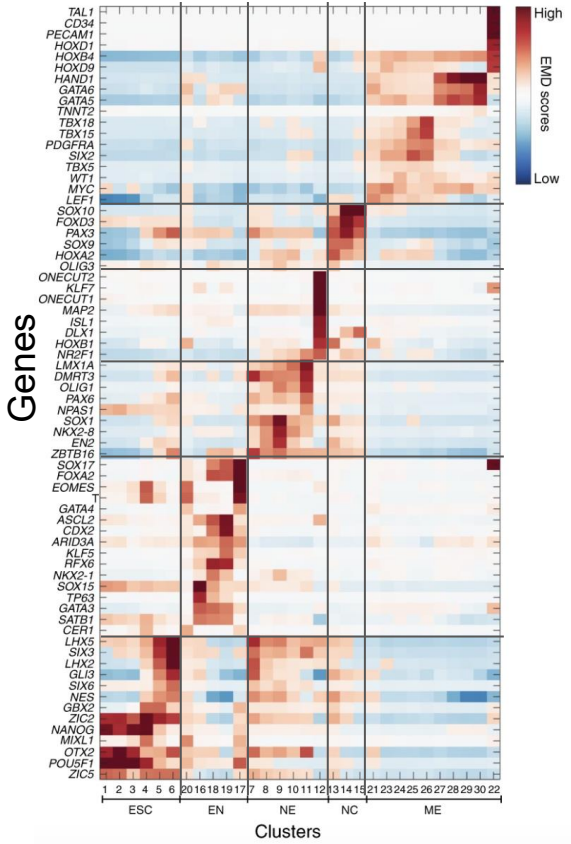




- Compare distances from X to other members of cluster A versus distances from X to members of cluster B (the closest cluster from A)

Inference of cell development markers

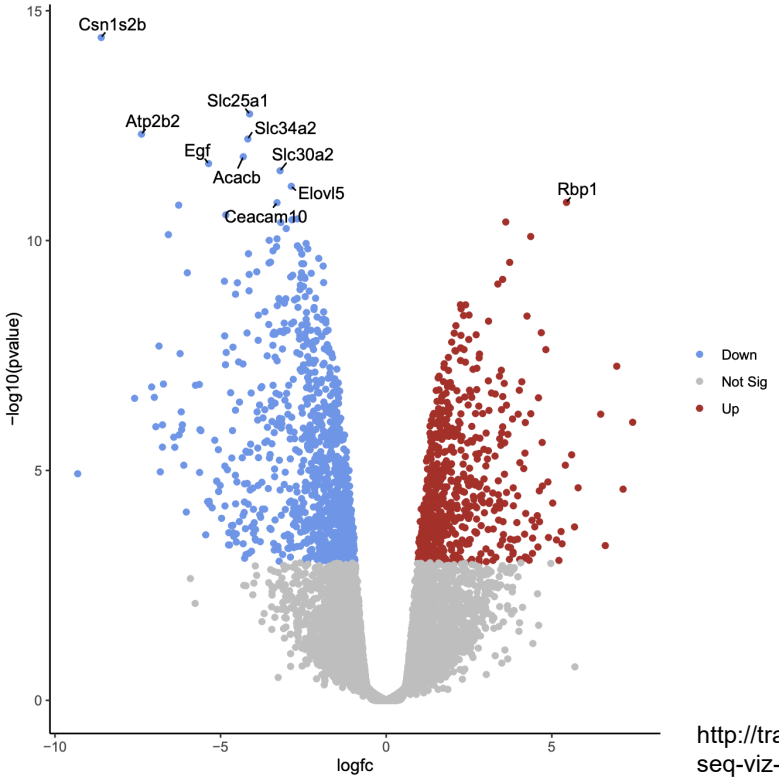




Source: Moon et al. Nature Biotechnology 37:1482-92 (2019)

Knowledge discovery with supervised ML

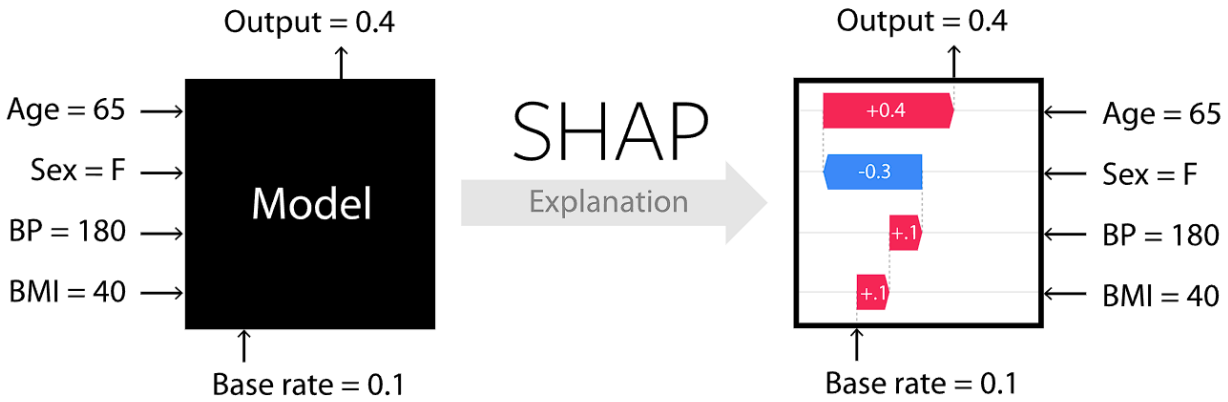
Univariate feature selection



- Most statistical and bioinformatics approaches are univariate
- Searching for panel of factors or markers is much more challenging
- **Some genes are important only in combination with others**

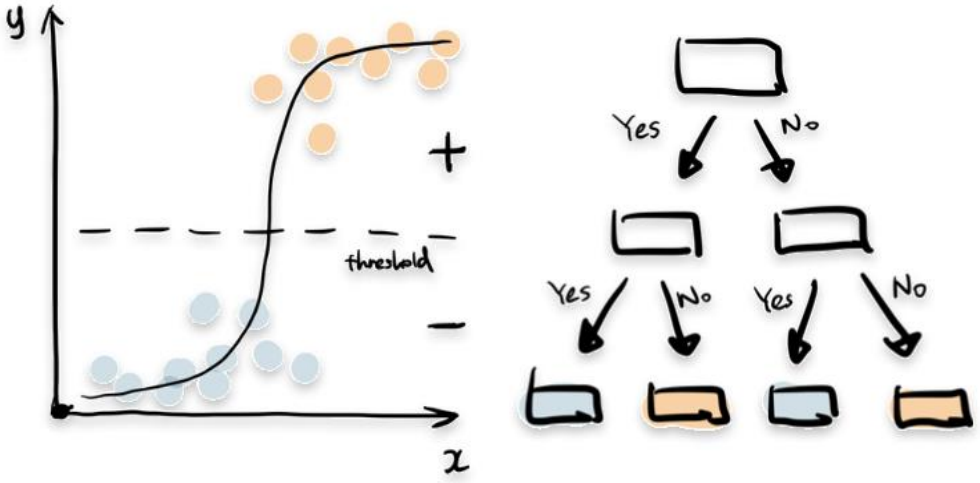
http://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html

Multivariate feature selection with explainable ML



- Black box model does not provide knowledge
- **Feature selection**: Remove unimportant features
- **Explainability**: Quantify feature contribution to the model's behavior
  - Model-level (performance) or sample-level (output)

Inherently explainable models: Linear and tree



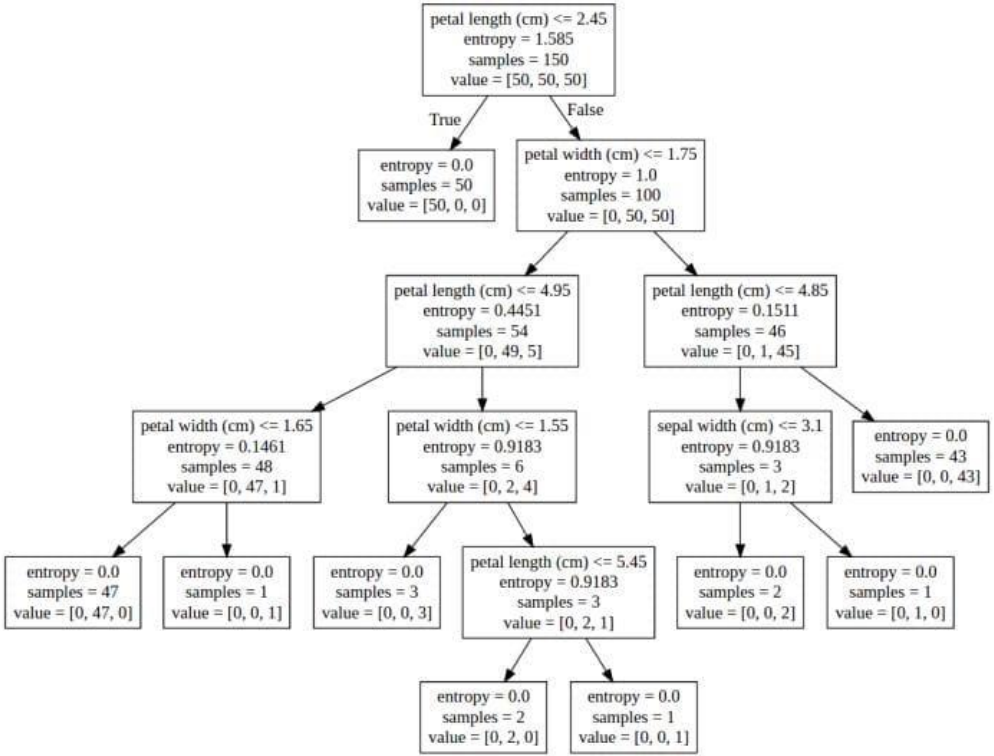
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

- Model decisions are immediately understandable
  - Examine coefficients
  - Trace the decision in a tree
- We can use these models to approximate a more complex model (around a data point)

Interpretation of tree models

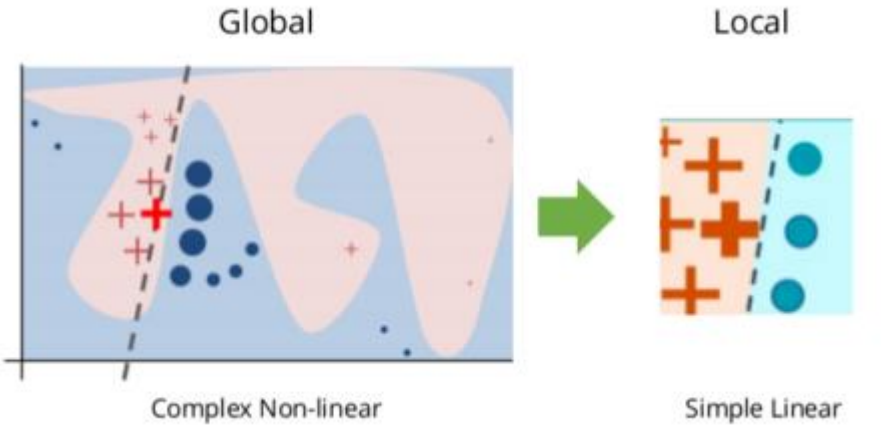
- **Measures of importance**
  - How often is a feature used?
  - When used, how good can it separate the data groups?
  - How many samples were involved?

- Entropy / Gini impurity scores
  - $\sum p \cdot \log(p)$
  - $\sum p(1-p)$



https://www.kdnuggets.com/2017/05/simplifying-decision-tree-interpretation-decision-rules-python.html

LIME: Local Interpretable Model-Agnostic Explanation



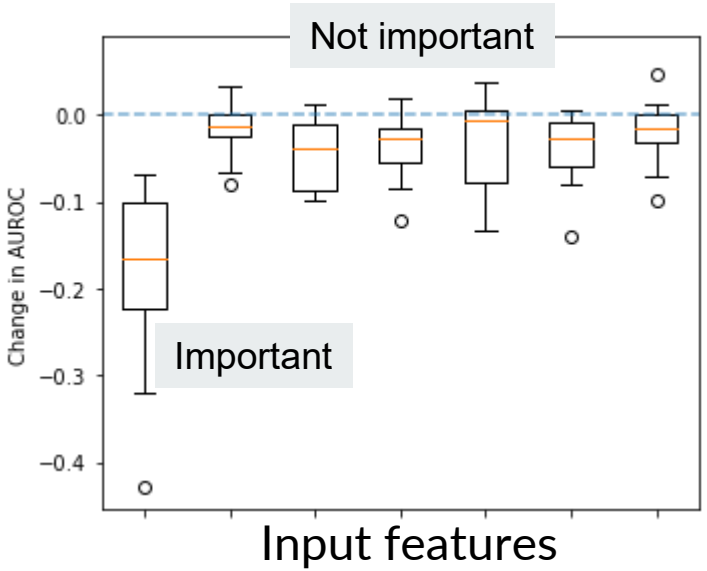
https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

- Focus on the decision boundary surrounding a data point of interest
- Approximate the original model with an explainable model (e.g., linear) by fitting on (input, output) surrounding a data point

Shapley value and dropout technique

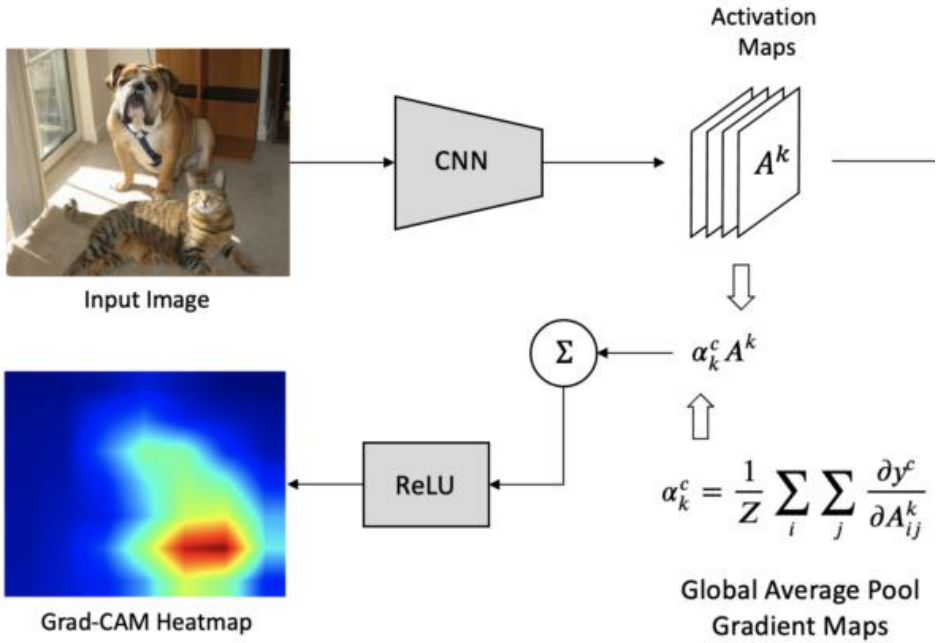
$$\phi_i(N,v) = \underbrace{\frac{1}{|N|!}\sum_{S\subseteq N\setminus\{i\}}}_{\text{Average}} \underbrace{|S|!\,(|N|-|S|-1)!}_{\text{Weight}} \underbrace{[v(S\cup\{i\})-v(S)]}_{\text{Marginal contributions}}$$

https://medium.com/the-modern-scientist/what-is-the-shapley-value-8ca624274d5a



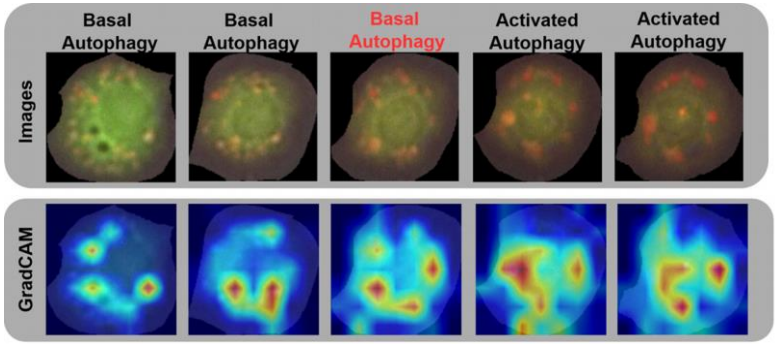
- Principle from game theory in economics
- **Dropout**: remove one or more features from the data and measure the changes in model performance and behaviors

Explanation for image data



https://learnopencv.com/intro-to-gradcam/

Presacan, O. et al. PLoS ONE 20:e0331045 (2025)



- Trace back to where the signal that contributed to the prediction came from
- Compare with knowledge

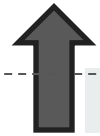
Summary

- **Hypothesis driven**: Literature review

Validation
Idea refinement

- **Data driven**:
  - Unsupervised learning
    - Visualization
    - Clustering

Targets for feature selection and prediction



Validation and inspection

  - Supervised learning
    - Feature selection
    - Predicted outcomes

Any question?

- See you next time