



# 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)

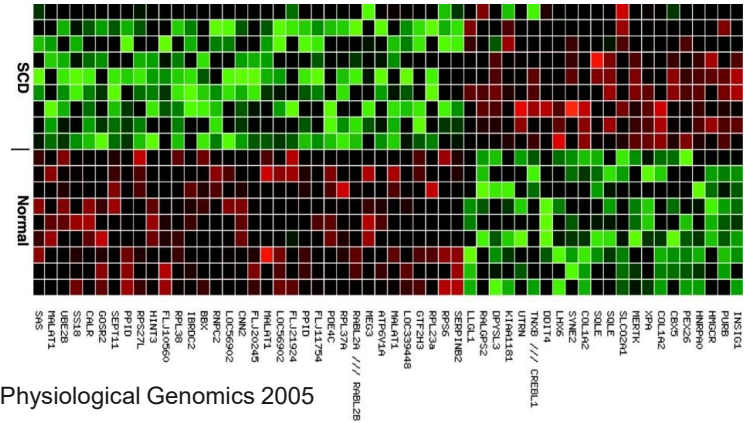
# Today's agenda



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

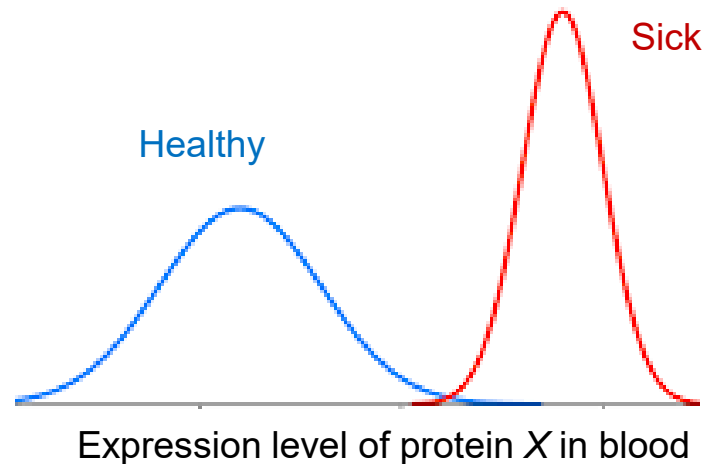


Klings et al. Physiological Genomics 2005



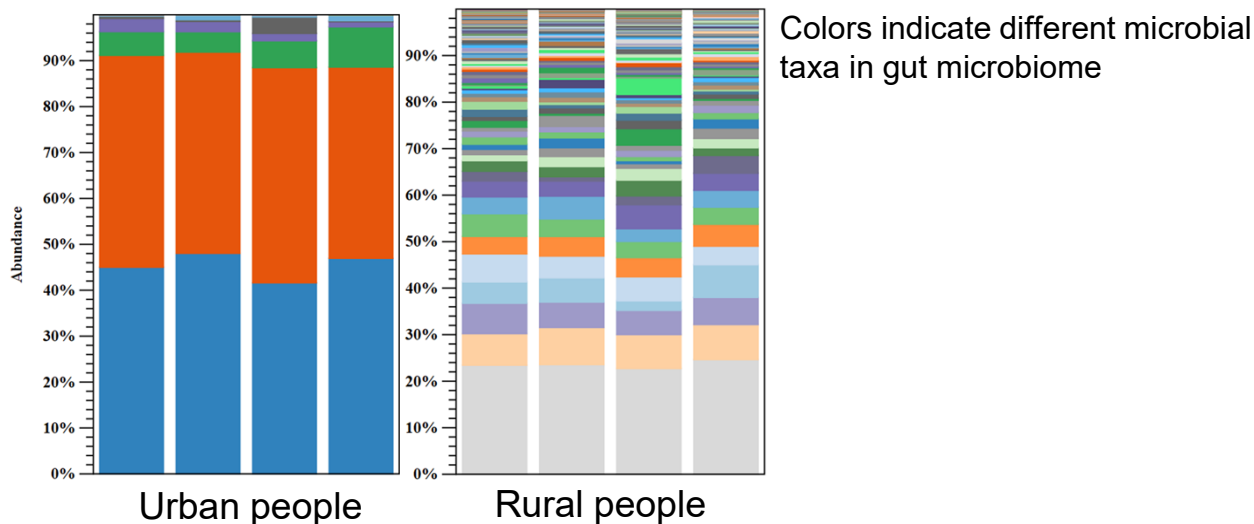
- Not just gene A is up-regulated, but **genes A is up-regulated by 2.36 folds with standard deviation of 0.18 across 12 biological replicates**
- Biology has become quantitative

# Quantitative thinking



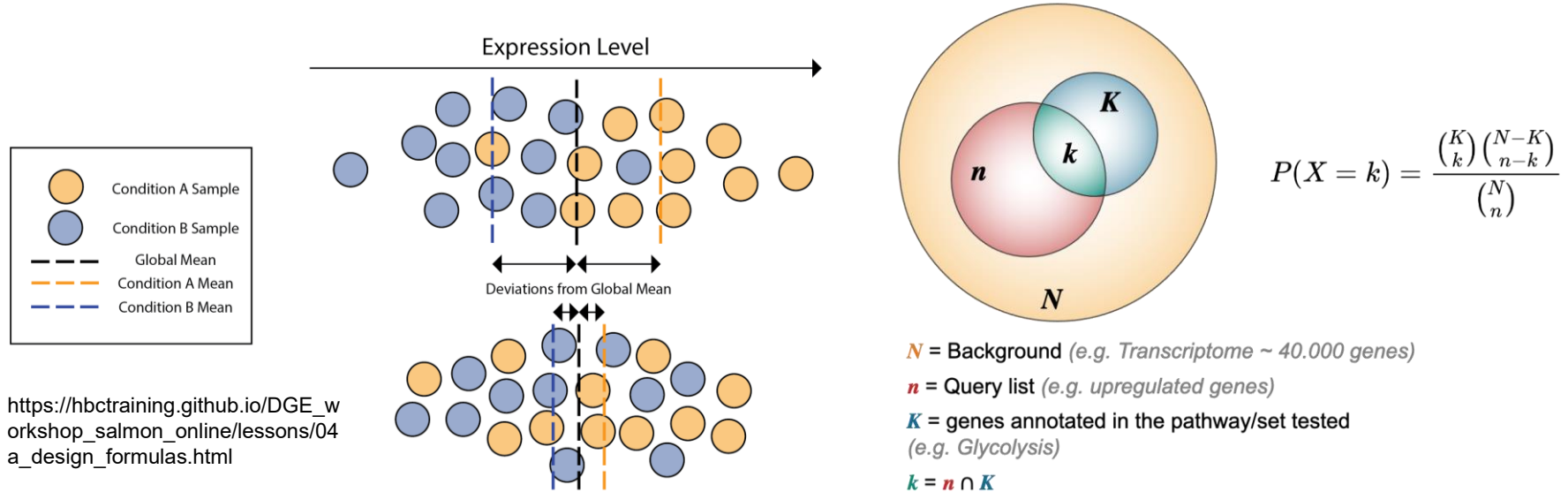
- How would you to quantify the ability of X to distinguish sick patients?
  - How about  $\text{Score}(X) = \text{Mean}_1 - \text{Mean}_2$  or  $\text{Abs}(\text{Mean}_1 - \text{Mean}_2)$  ?
  - How about  $\text{Score}(X) = \frac{\text{Mean}_1 - \text{Mean}_2}{\sqrt{\frac{1}{n}(\text{Variance}_1 +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

# Turning verbal description into mathematical formula



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

# Statistical framework provides objectivity



[https://hbctraining.github.io/DGE\\_workshop\\_salmon\\_online/lessons/04\\_a\\_design\\_formulas.html](https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/04_a_design_formulas.html)

- 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






# **Roles of algorithm in bioinformatics**



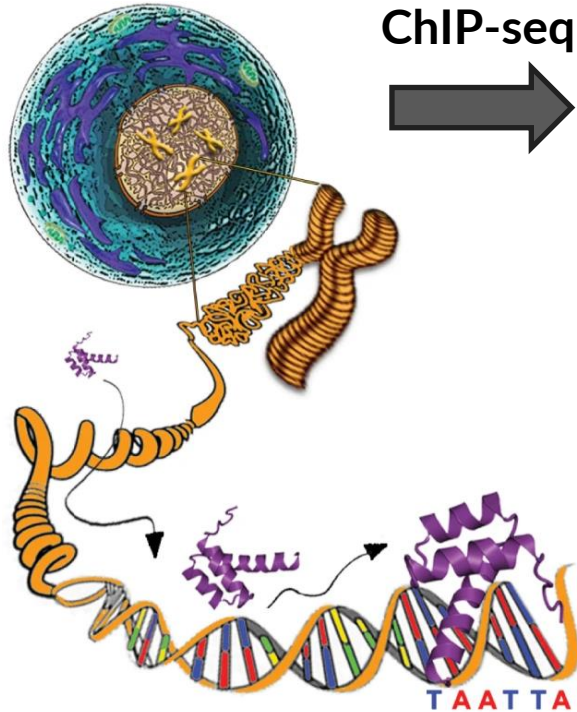
# Dynamic programming for sequence alignment

	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

# DNA-binding motif discovery

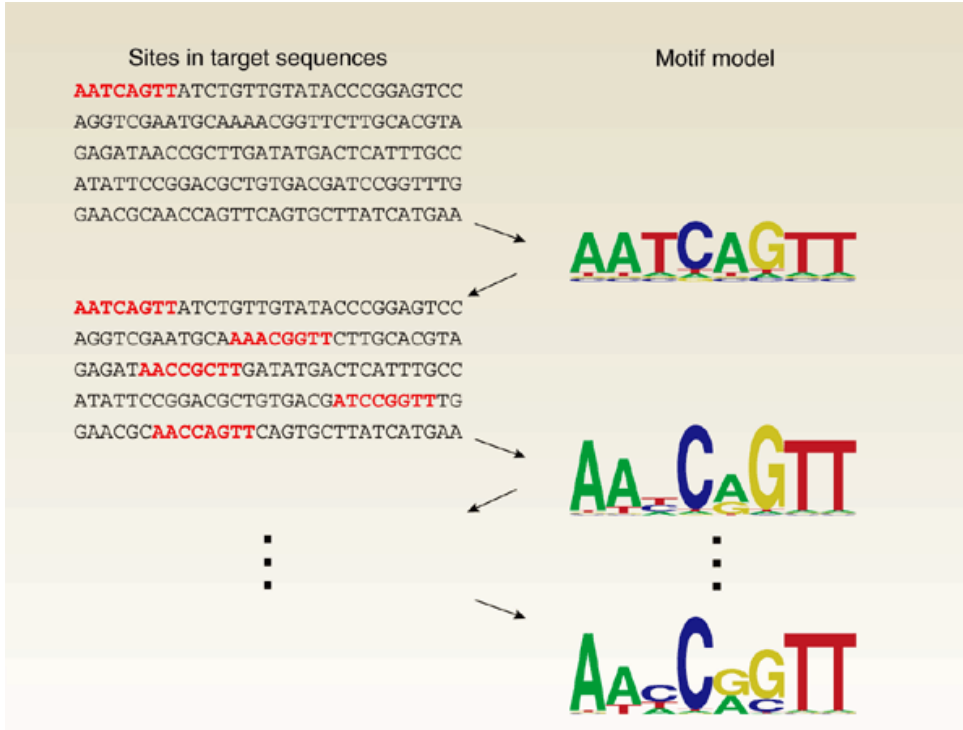


CGGGGCTATcCAGCTGGGTCGTCACATTCCCCTT  
TTTGAGGGTGCCCAATAAaggGCAACTCCAAAGCGGACAAA  
GGATGgAtCTGATGCCGTTTGACGACCTA  
AAGGAaGCAACcCCAGGAGCGCCTTTGCTGG  
AGATTATAATGTCGGTCCtTGgAACTTC  
CAACTGAGATCATGCTGCATGCcAtTTTCAAC  
TACATGATCTTTTGATGgCACTTGGATGAGGGAATGATGC

An Introduction to Bioinformatics Algorithm by Jones and Pevzner

- **Assumption:** The motif must be present in each DNA sequence, at least once

# Motif discovery algorithm



- 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

# Synergy between algorithm and statistics

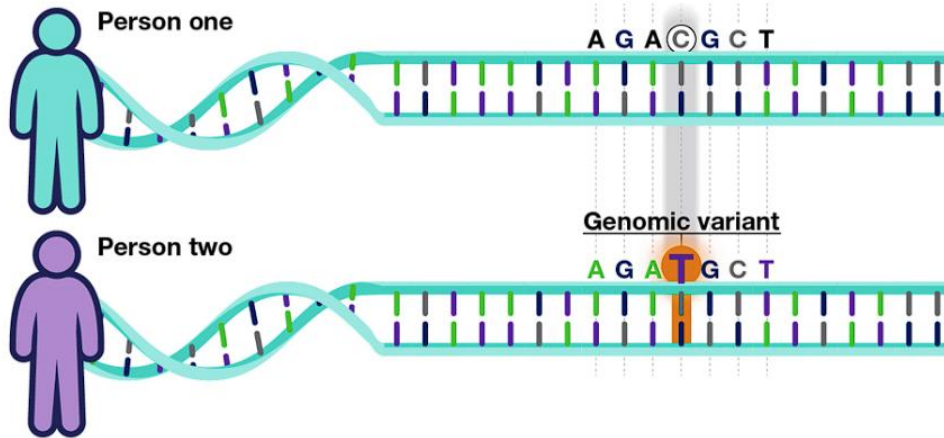


- Algorithm identifies the best possible answer in your data
  - Aligned portion of sequences
  - DNA motifs
- Statistics model the distribution of the scores and provides objective significance cutoff



# The need for machine learning

# Human judgment in bioinformatics

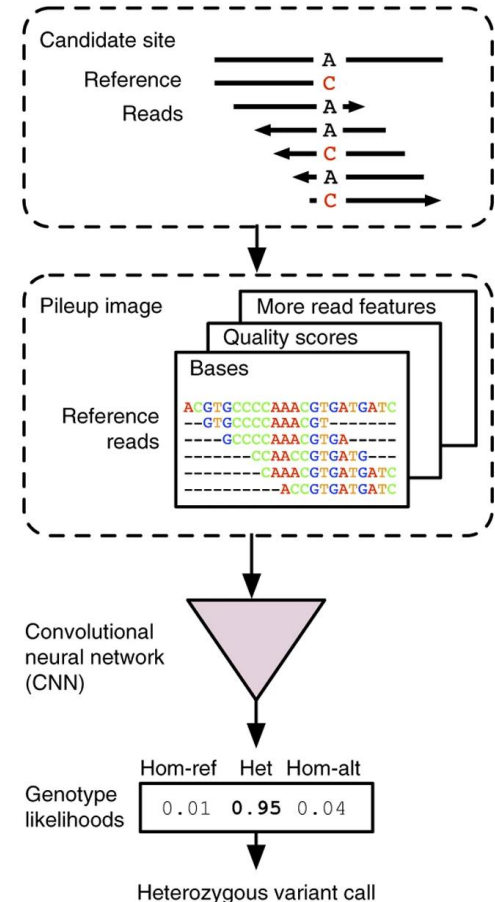


<https://storymd.com/journal/4m8ald6ipw-gene-variants-and-health/page/nrq7zt7bry-what-is-a-gene-variant-and-how-do-variants-occur>

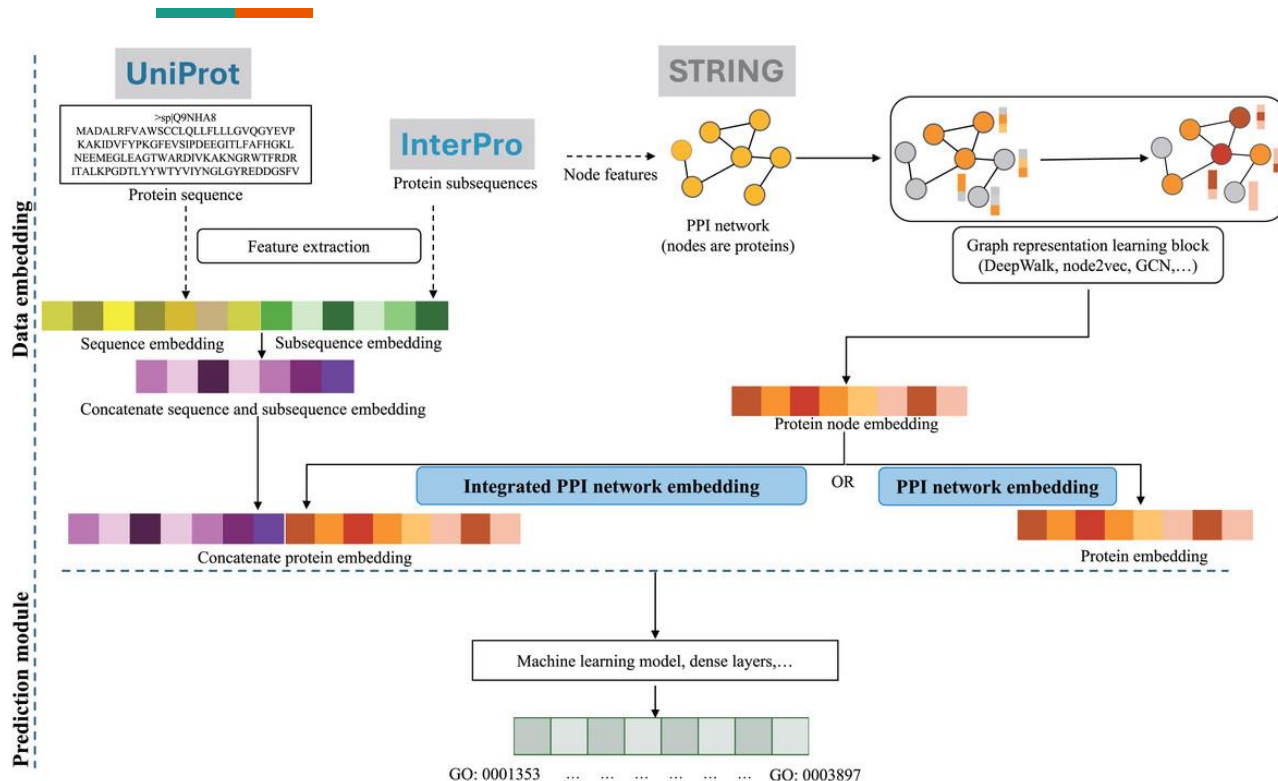
- When to focus on a mutation?
- Many subjective filters:
  - Base quality
  - Read depth
  - Allele frequency (AF)
  - Population AF
  - Coding or non-coding
- Each lab has different criteria

# 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



# Machine learning integrates data types

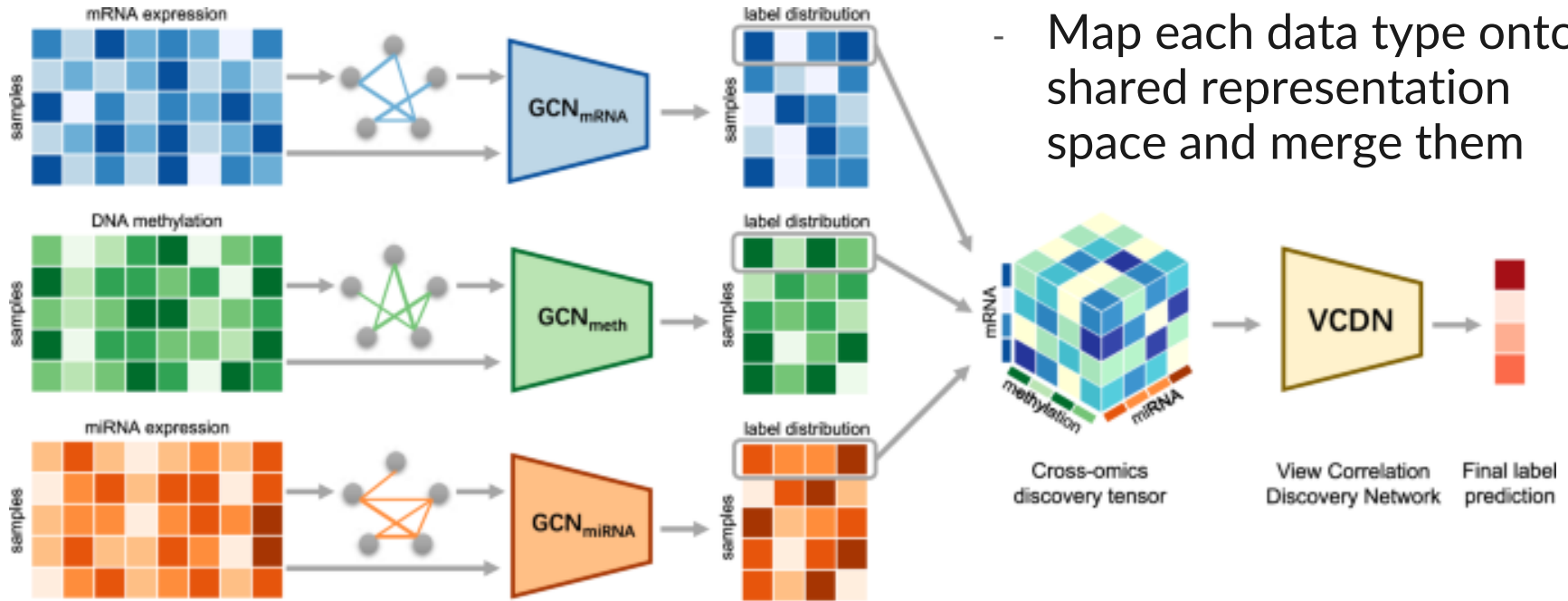


- Condition the learning to make representations from multiple raw data types computationally compatible
- Additive, concatenate, etc.

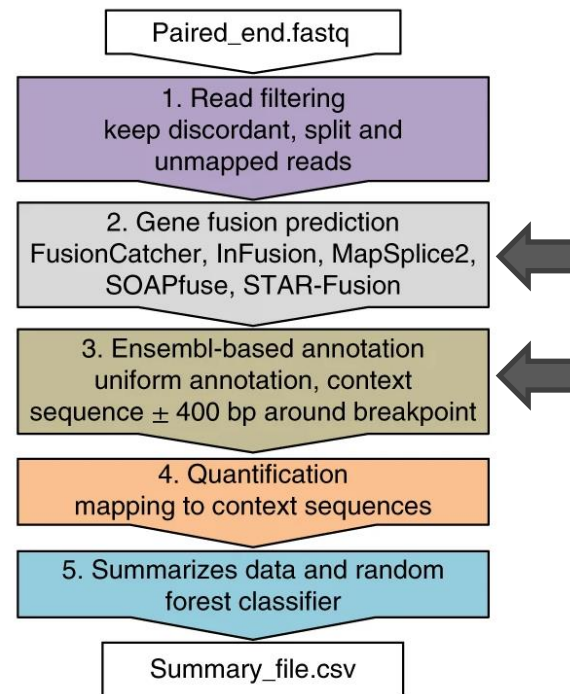
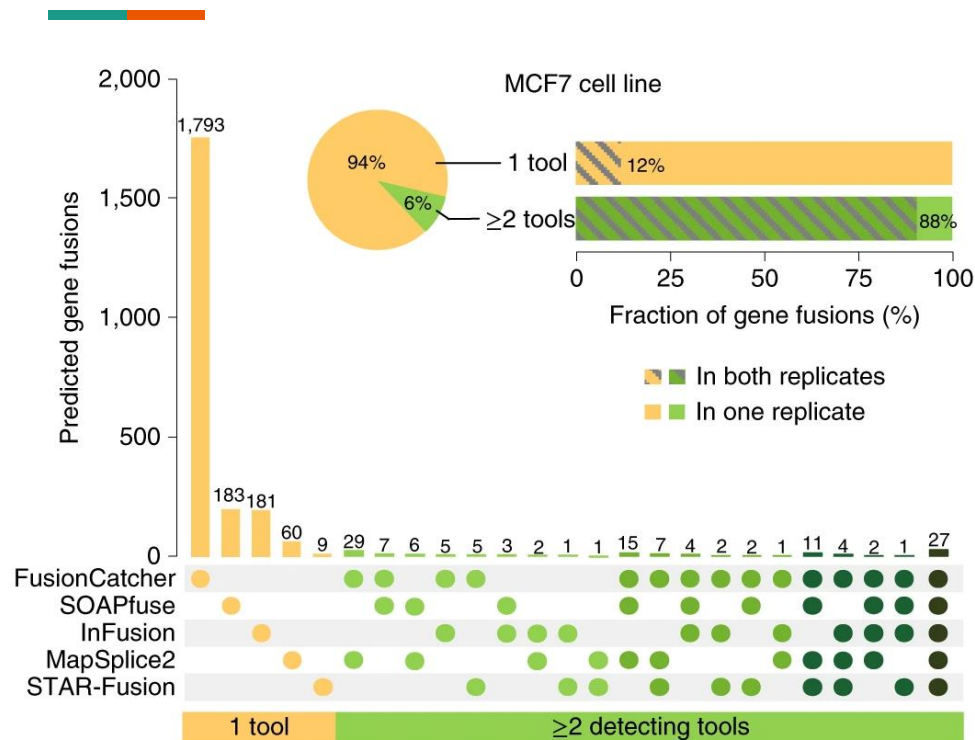
GO: 0001353 ... GO: 0003897



# Multi-omics integration with ML



# Machine learning aggregate bioinformatics tools



# Synergy between ML and bioinformatics

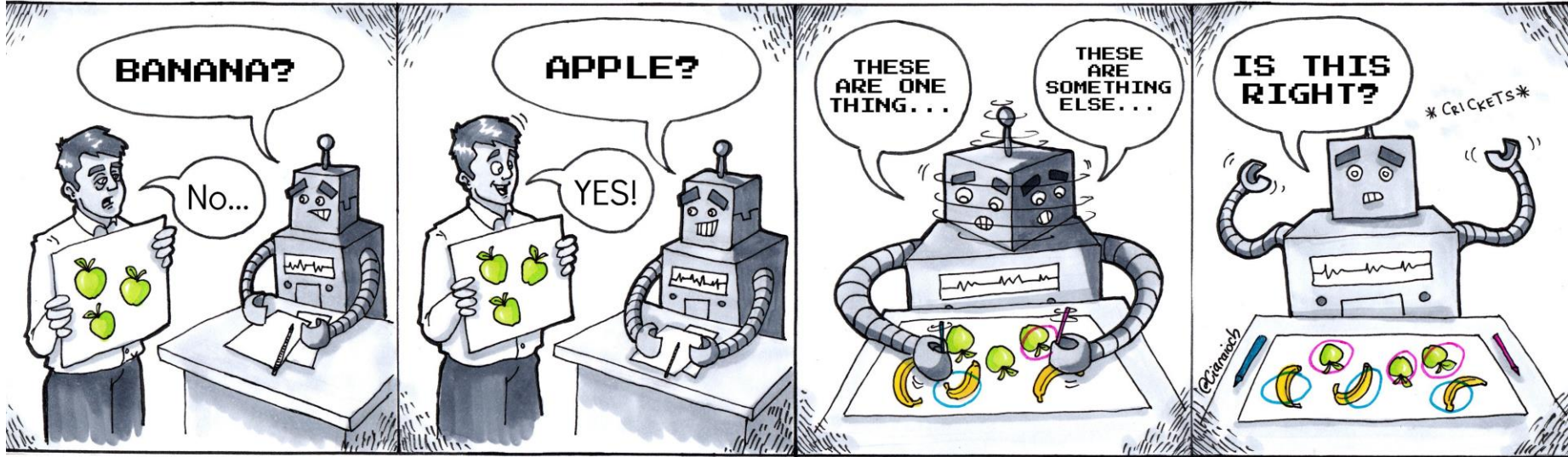


- Different bioinformatics algorithms produce different mistakes
- ML can learn to identify when to trust each algorithm from the data and confidence scores
- ML can speed up bioinformatics
  - Multiple sequence alignment: identify pairs of sequences to align first
  - Protein modeling: directly identify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



# Knowledge discovery with ML

# Machine learning paradigms



## Supervised Learning

Find accurate decision functions

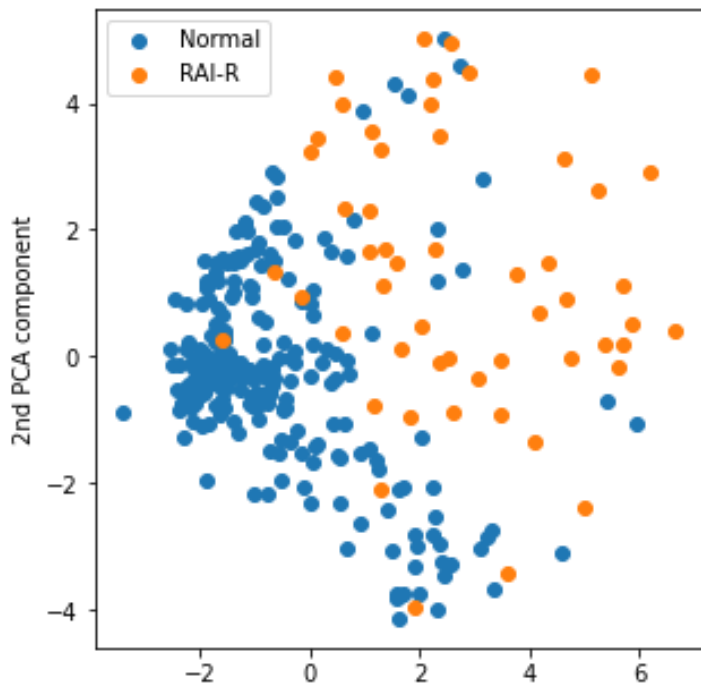
## Unsupervised Learning

Find similarities among data

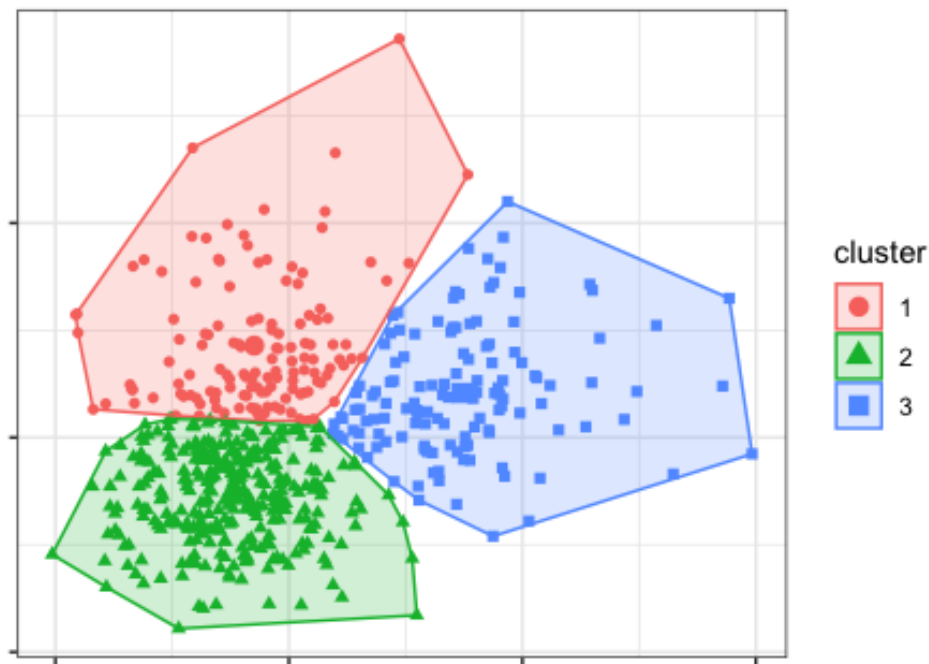
# Key unsupervised learning techniques



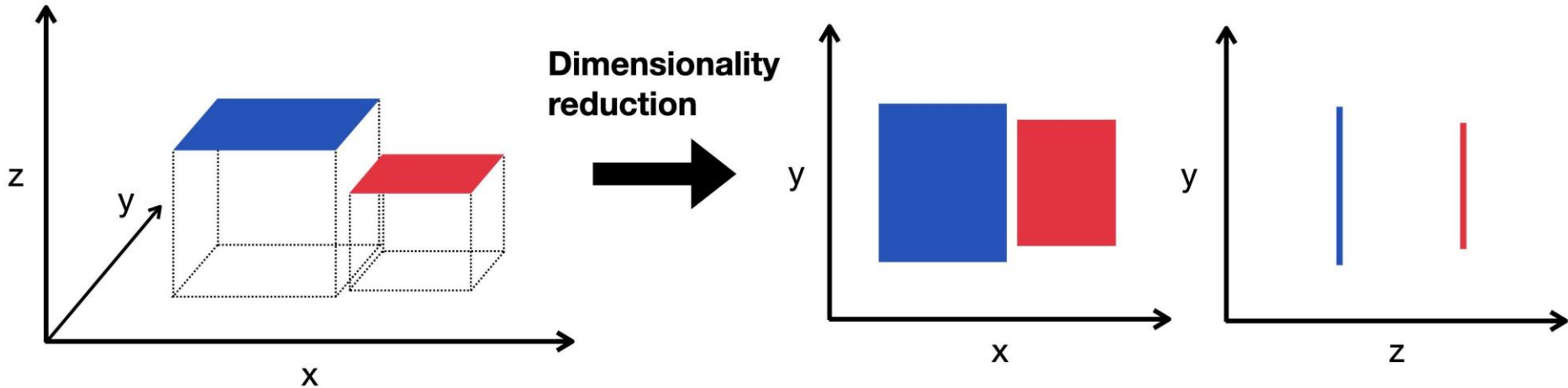
## Dimensionality Reduction



## Clustering / Anomaly Detection



# Dimensionality reduction



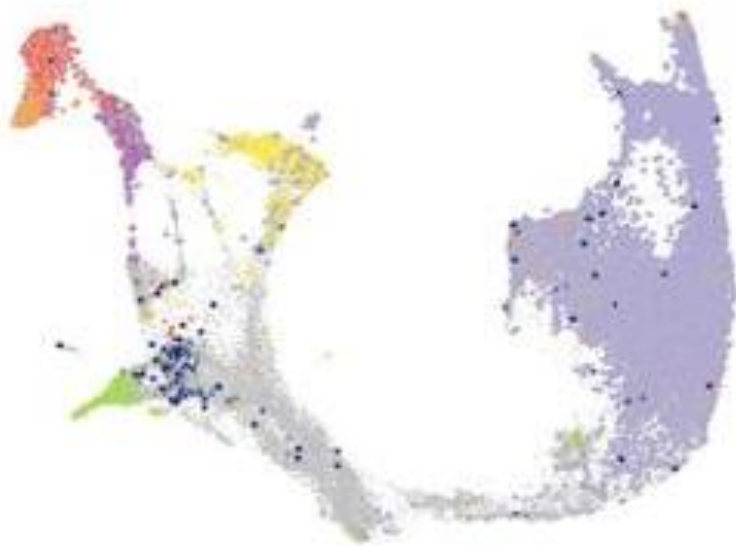
[https://www.sc-best-practices.org/preprocessing\\_visualization/dimensionality\\_reduction.html](https://www.sc-best-practices.org/preprocessing_visualization/dimensionality_reduction.html)

- 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

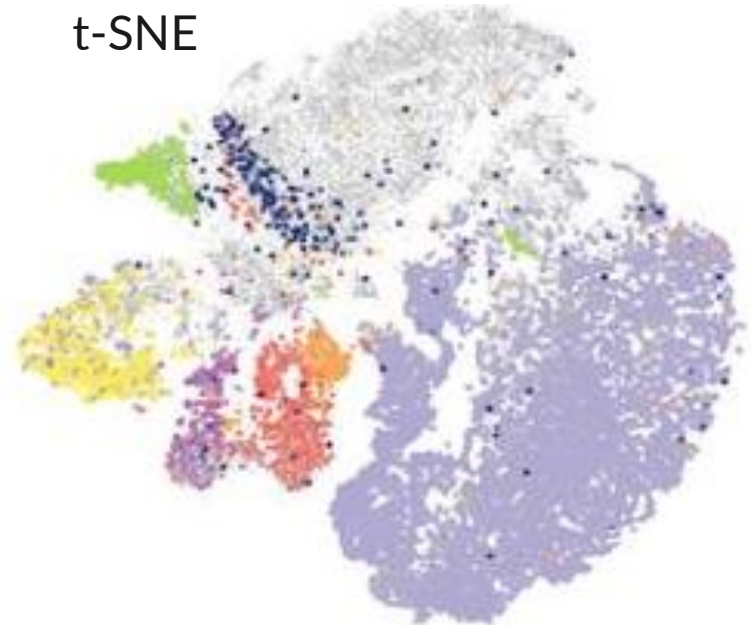
# Visualization of single-cell data



UMAP



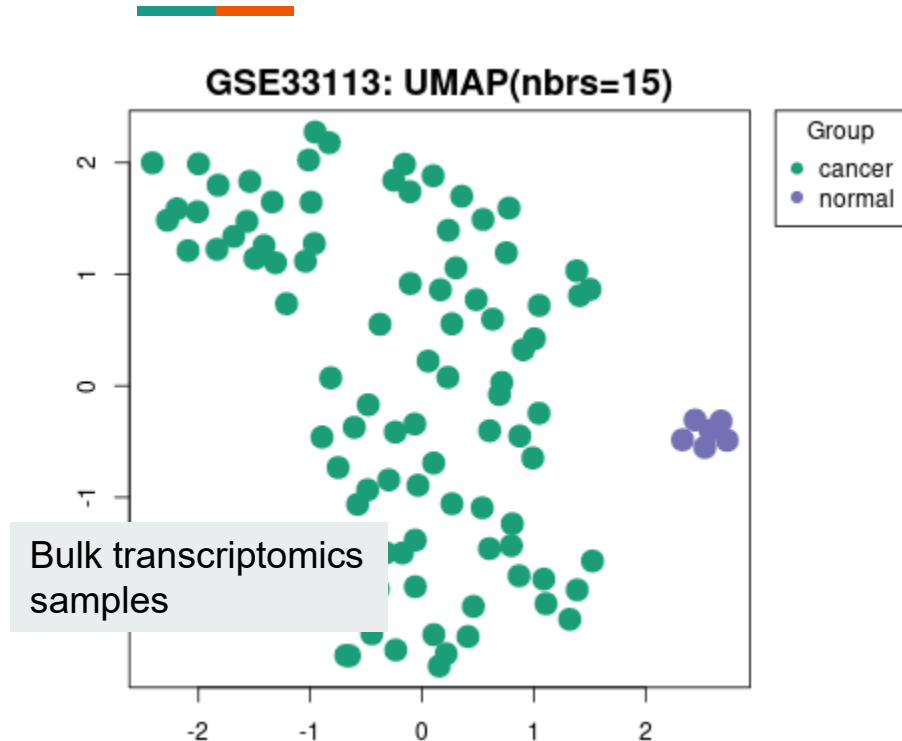
t-SNE



● MPP ● Macrophage ● Neutrophil ● Erythrocyte ● B cell ● T cell ● NK cell

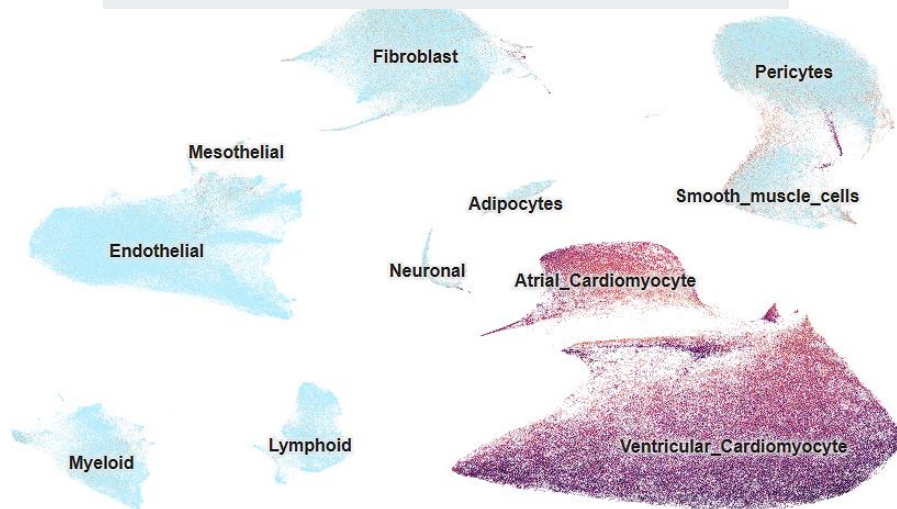


# Visualization helps generate/validate hypothesis



PCA plot from GEO2R

## Expression of CTNNA3 in cardiac cells



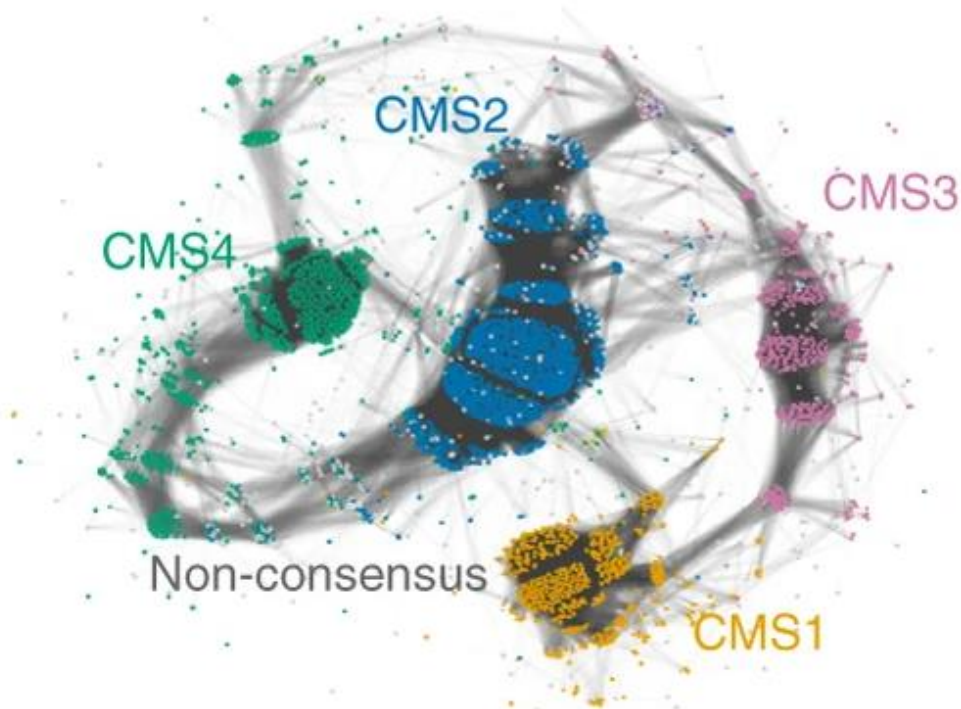
UCSC Cell Browser

# The heart of clustering

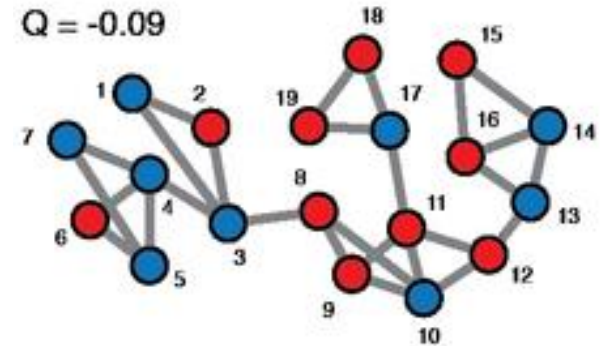
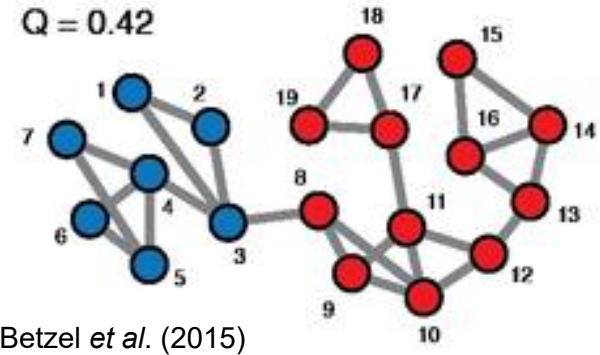


- **Goal:** Group **similar** data point together
- How to define **similarity**?
  - **Distance:** Between two data points
  - **Linkage:** Between groups of data points
- How many clusters is appropriate?
  - **Within-cluster (small) versus between-cluster (large) distance**

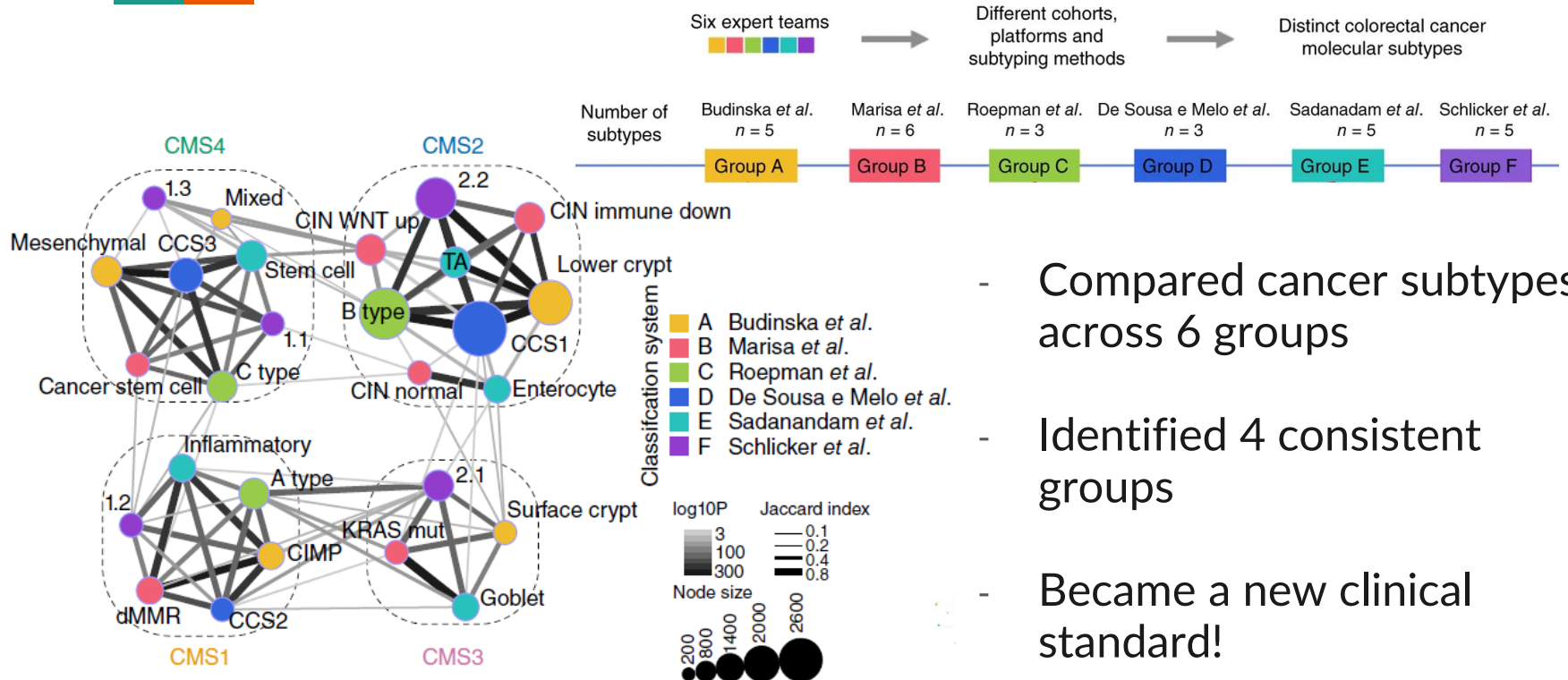
# Network clustering with modularity score



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

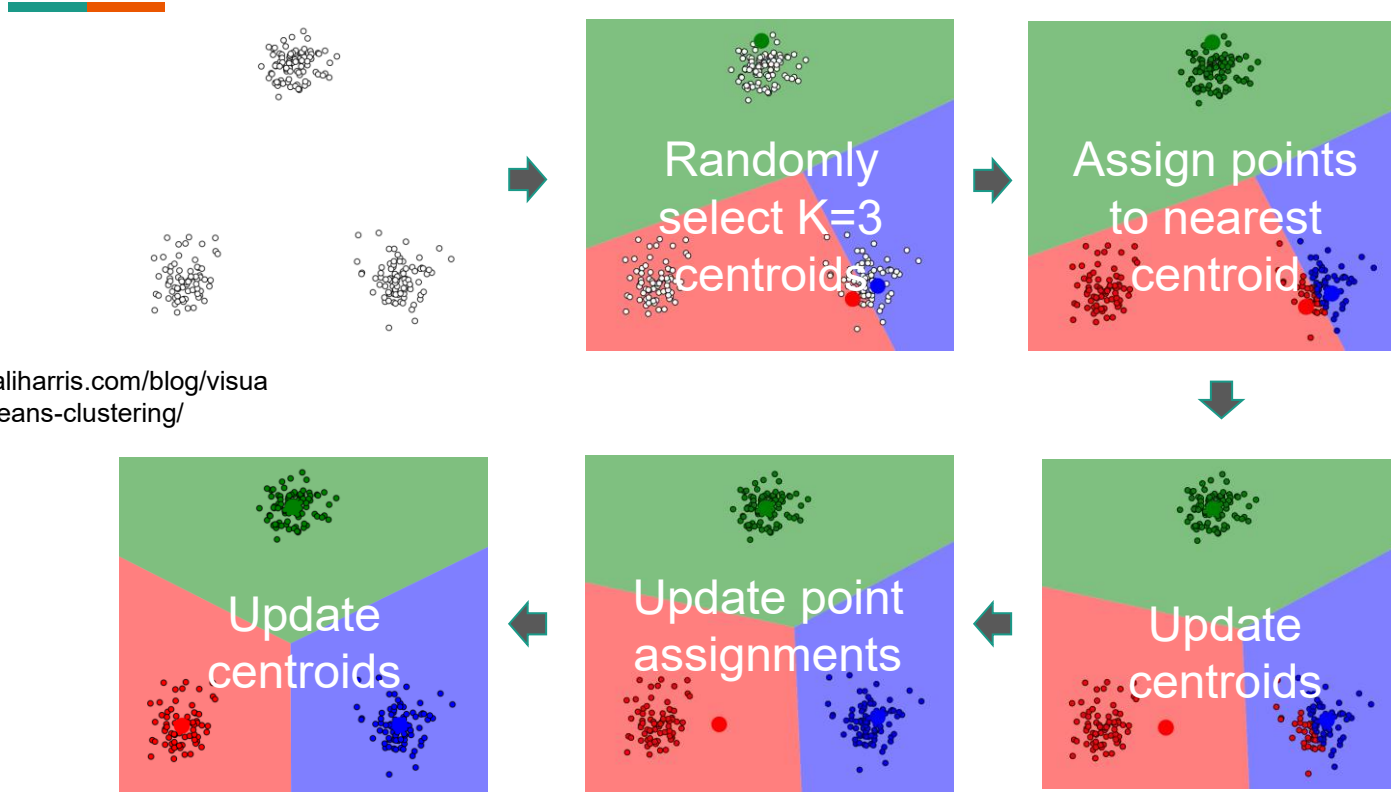


# Cancer consensus subtype discovery



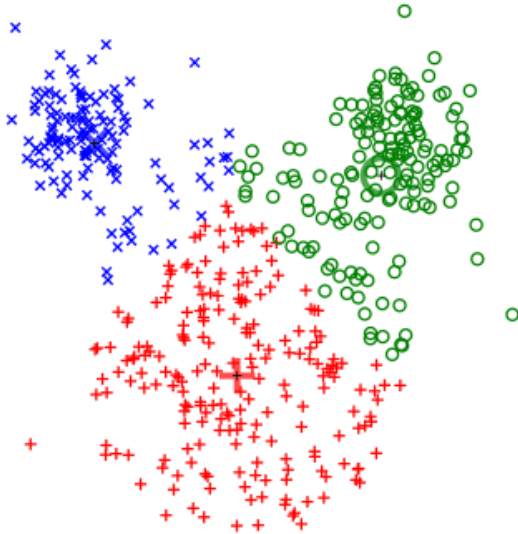
- Compared cancer subtypes across 6 groups
- Identified 4 consistent groups
- Became a new clinical standard!

# $k$ -mean: radius-based



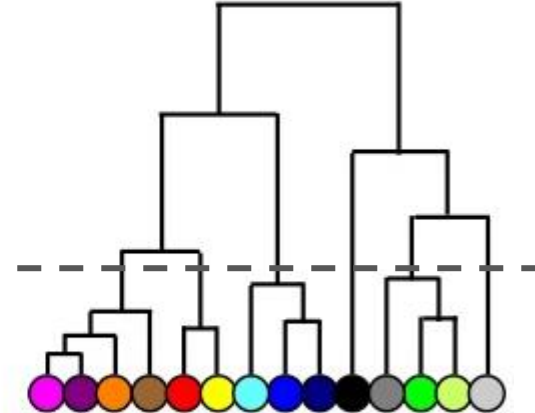
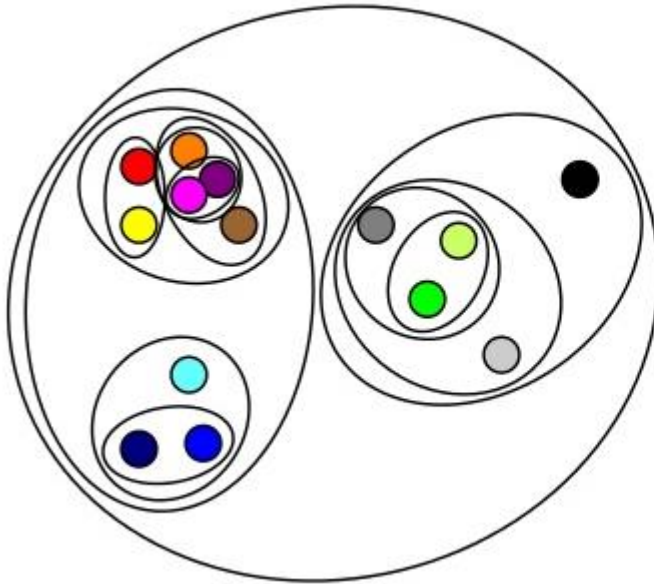
[www.naftaliharris.com/blog/visualizing-k-means-clustering/](http://www.naftaliharris.com/blog/visualizing-k-means-clustering/)

# 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

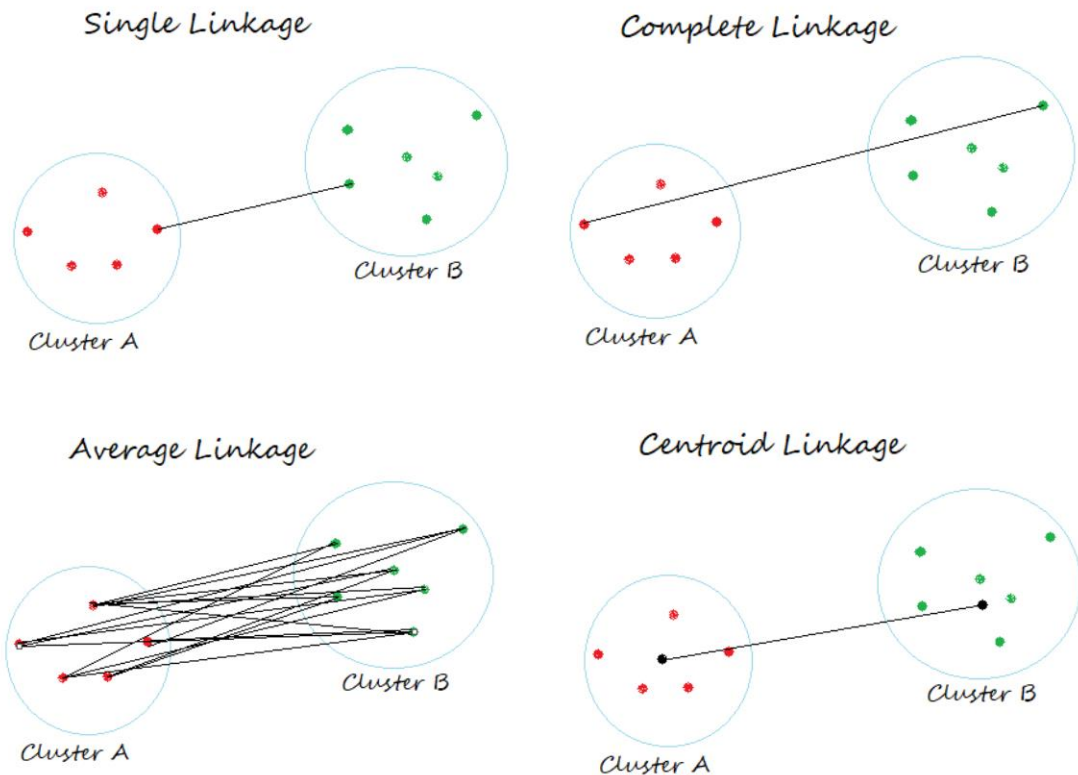
# Agglomerative/Hierarchical: neighbor-based



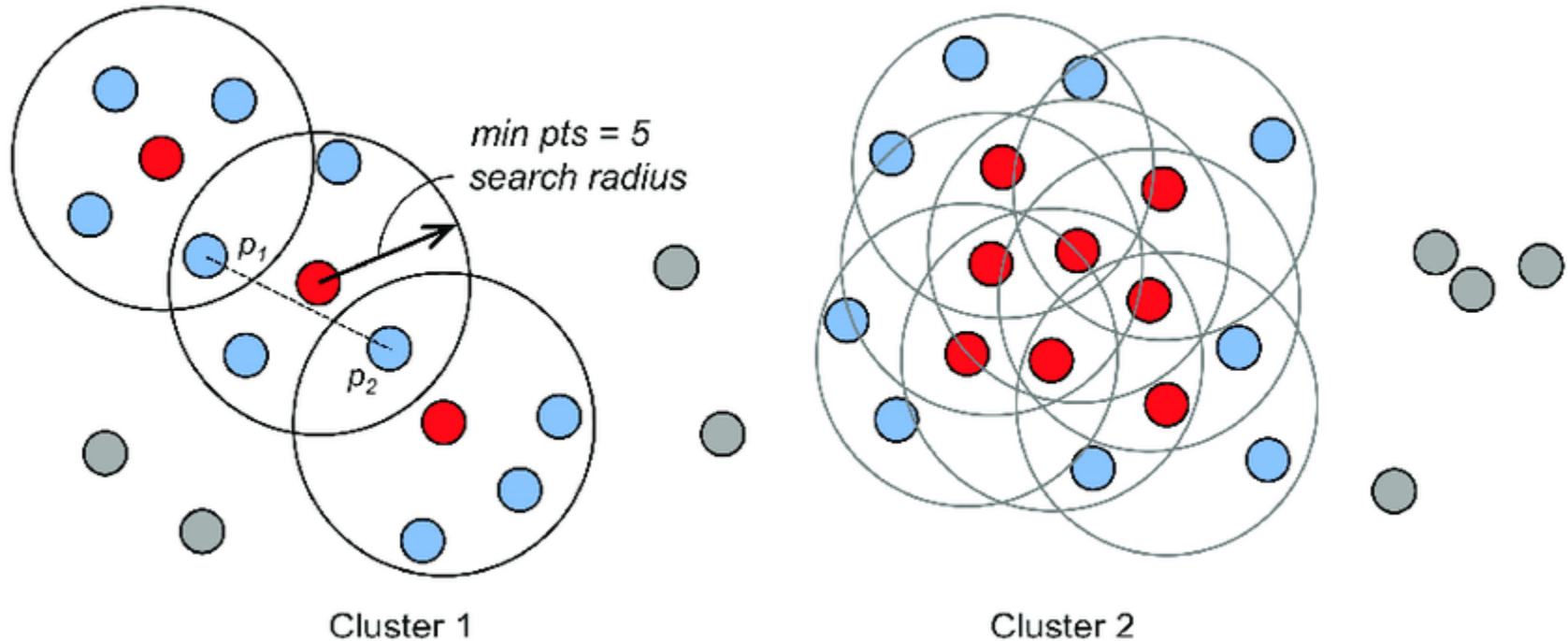




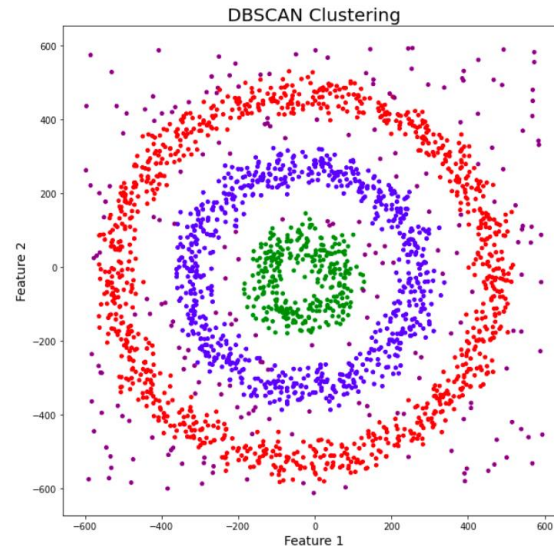
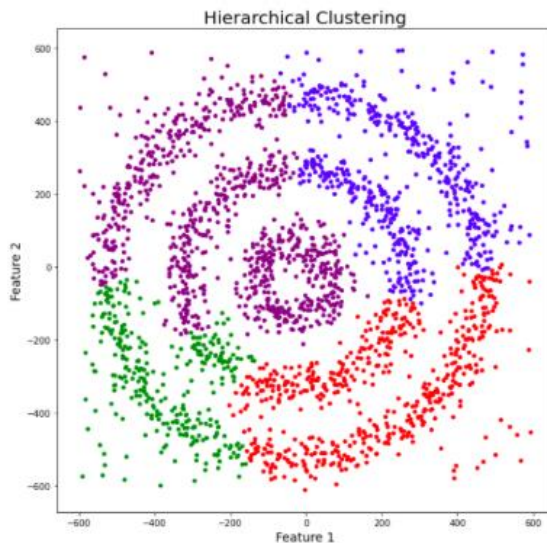
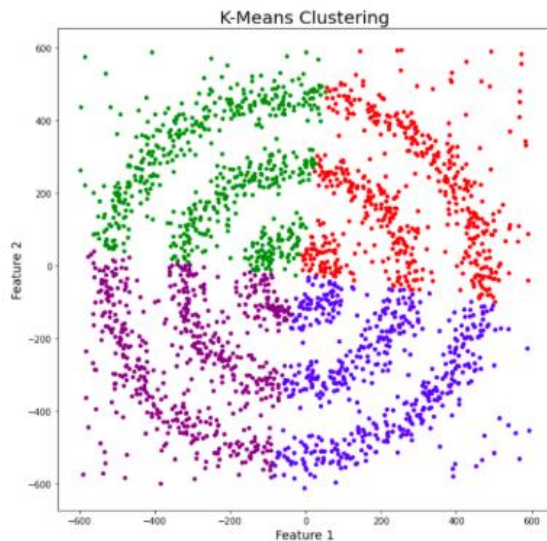
# Linkage = distance metric for groups of data points



# DBSCAN: density- and connectivity-based



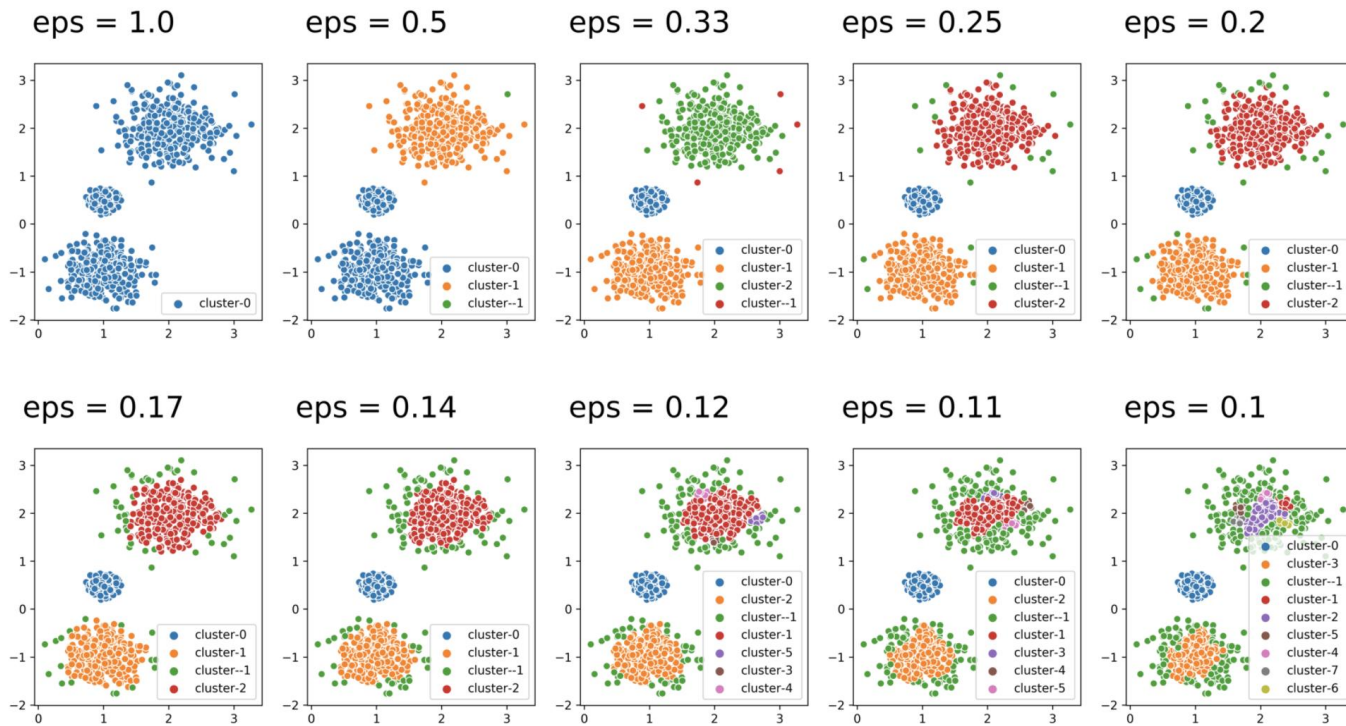
# Complex, non-circular clusters



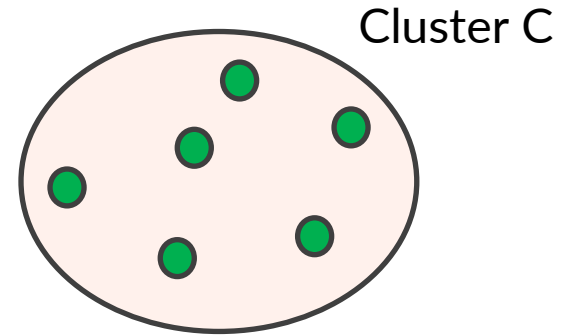
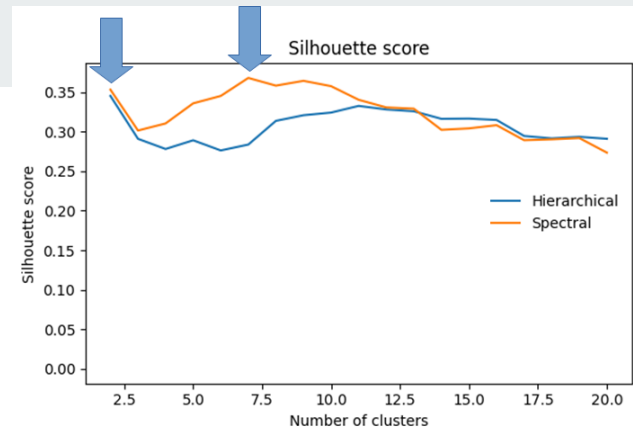
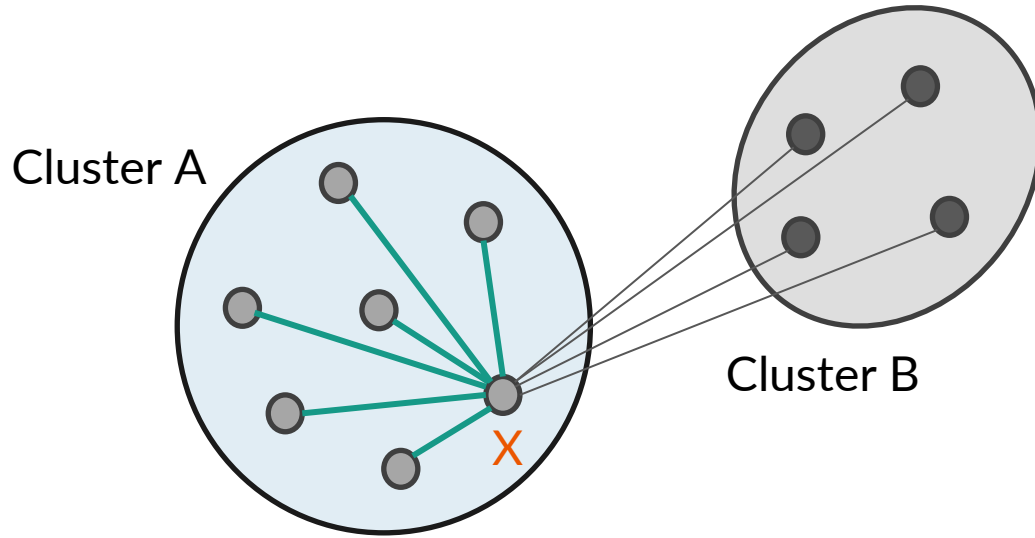
<https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/>

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

# Simultaneous detection of clusters and outliers

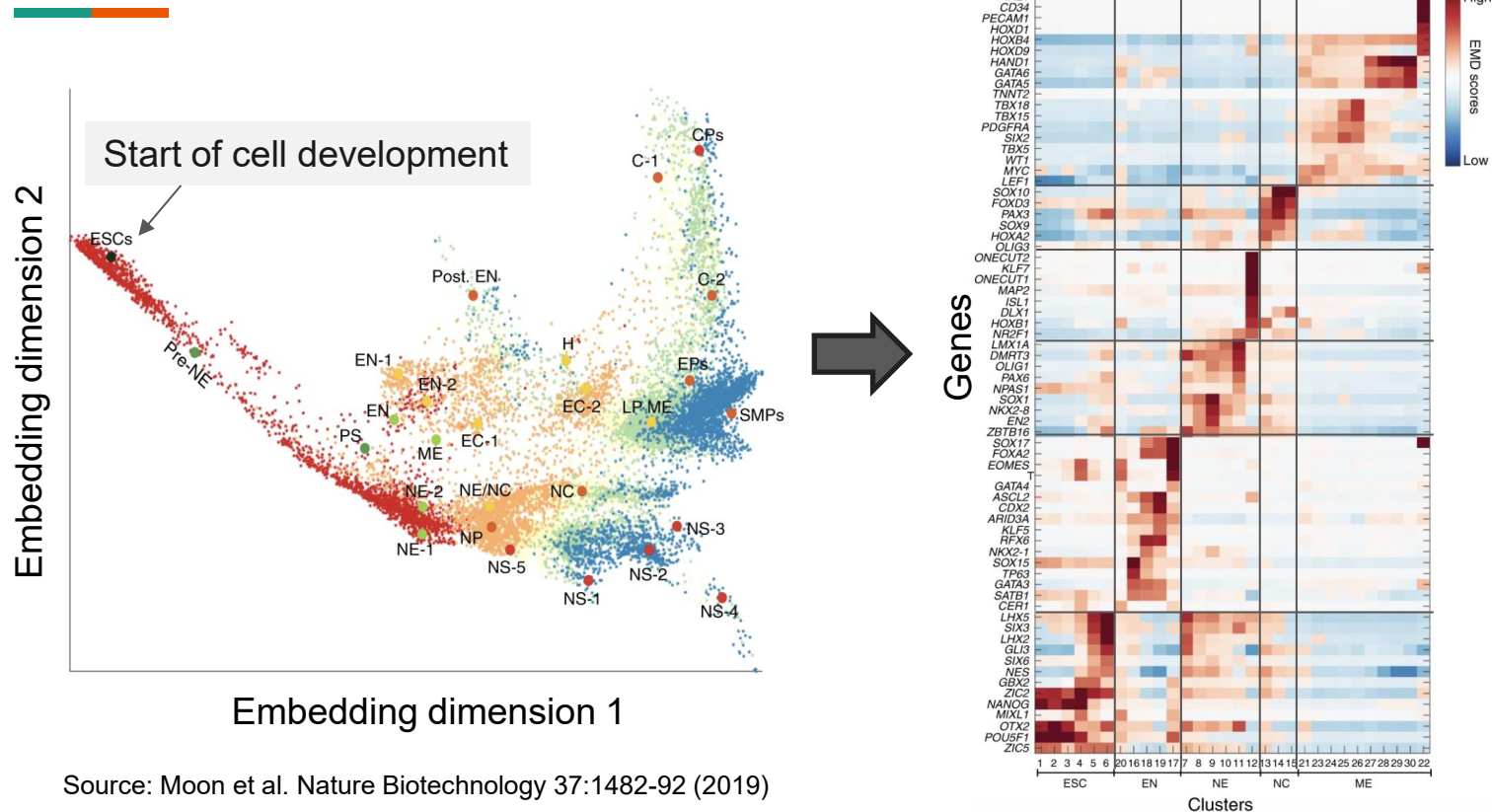


# Cluster selection: Silhouette score



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

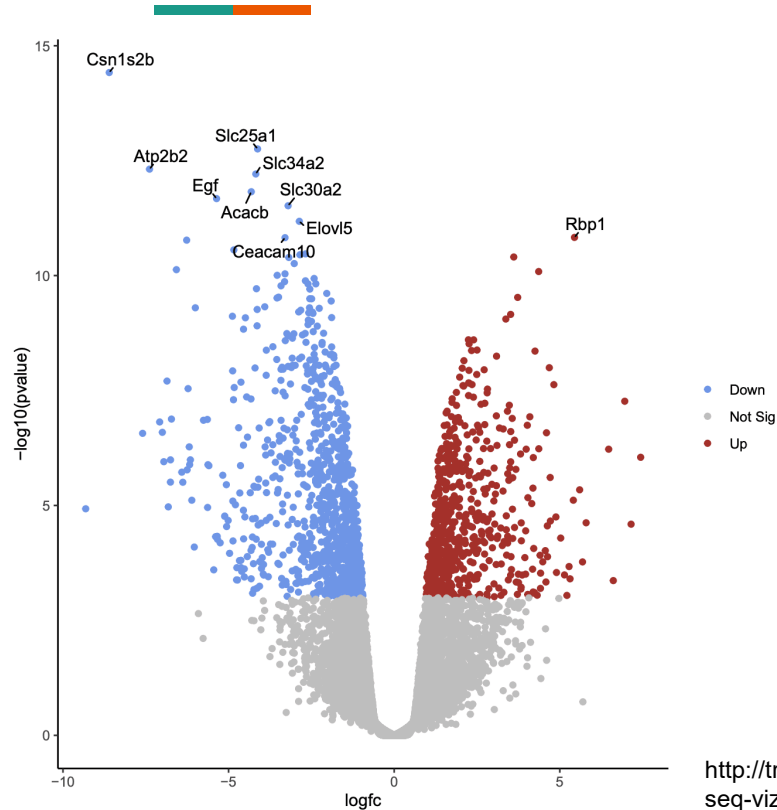
# Inference of cell development markers





# Feature selection with ML

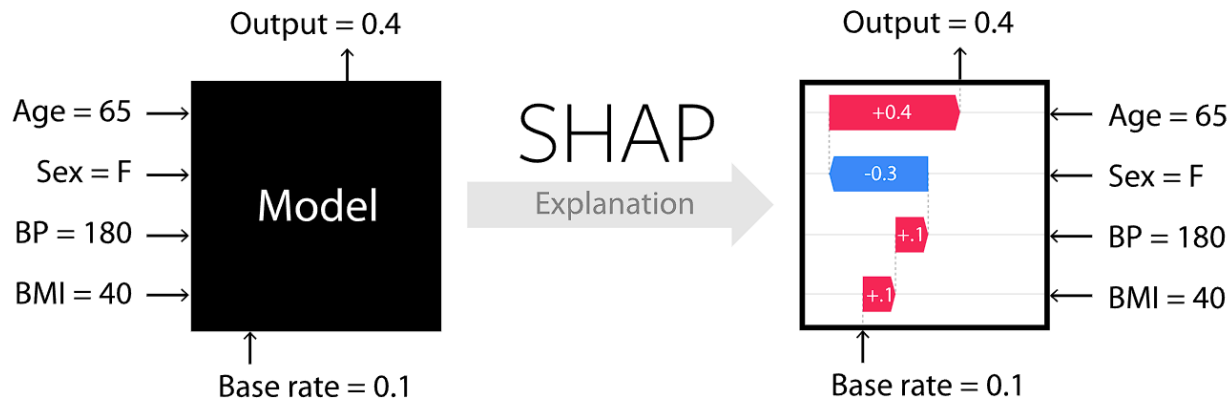
# Univariate feature selection



- 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

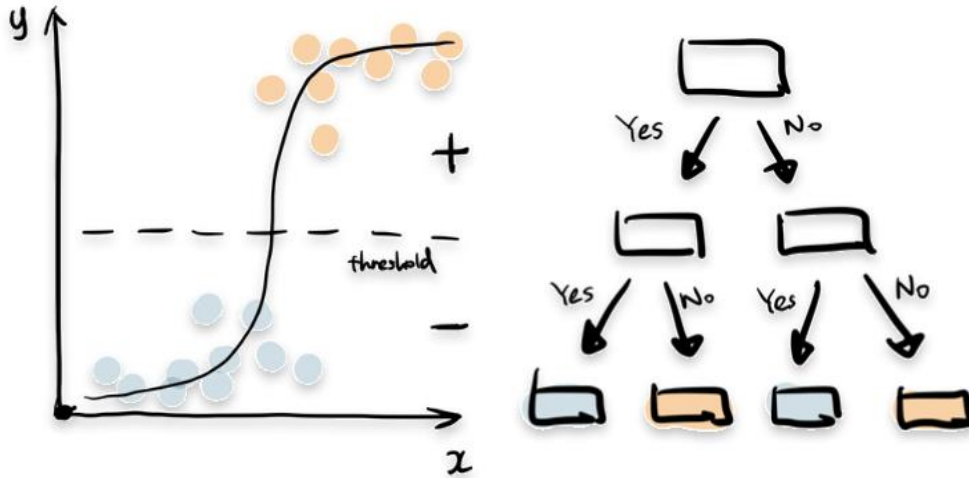


# Multivariate feature selection with explainable ML



- 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)

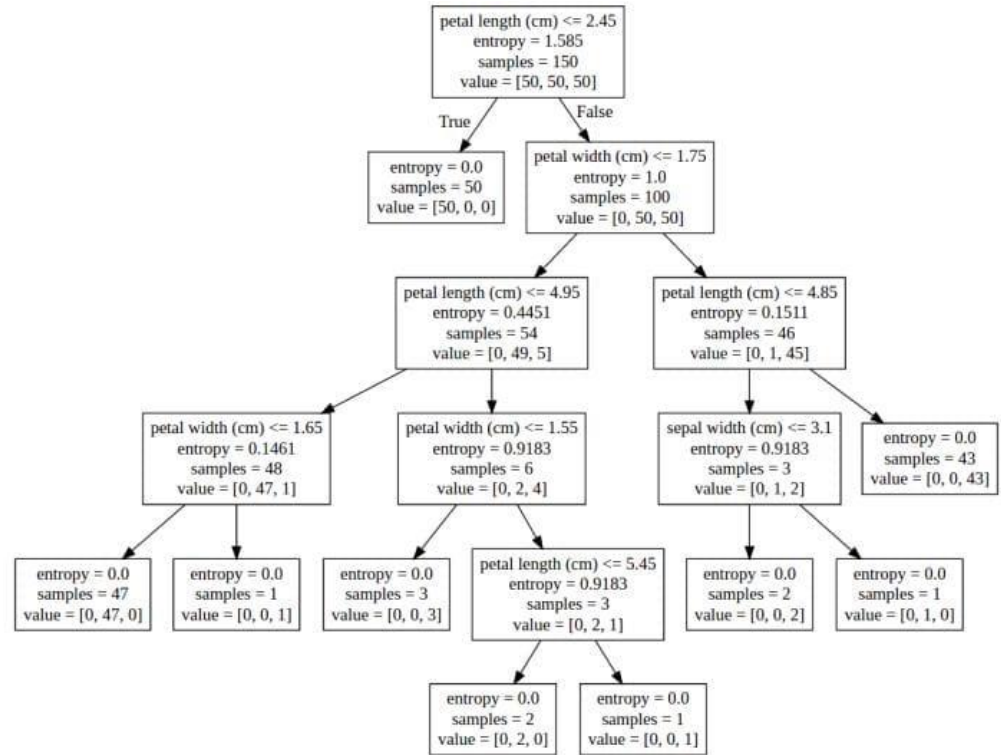
# Inherently explainable models: Linear and tree



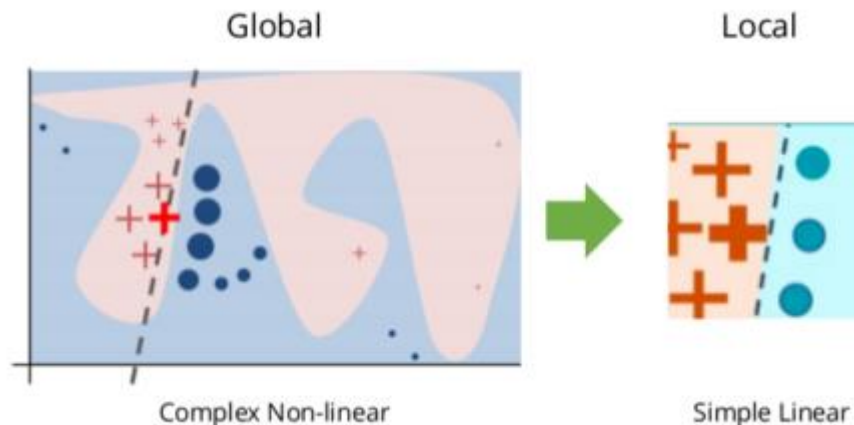
- Model decisions are immediately understandable
  - Examine coefficients
  - Trace the decision in a tree
- We can use these models to approximate a more complex model (around a data point)

# 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
  - $\sum p \cdot \log(p)$
  - $\sum p(1 - p)$



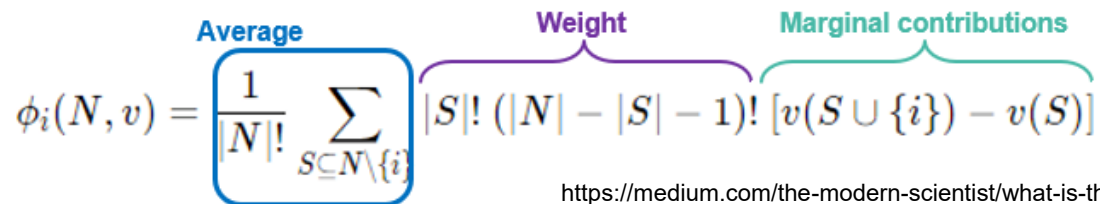
# LIME: Local Interpretable Model-Agnostic Explanation



<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



The diagram shows the Shapley value formula with three annotations: 'Average' in blue above the fraction, 'Weight' in purple above the factorial terms, and 'Marginal contributions' in green above the difference in value terms. The formula is:

$$\phi_i(N, v) = \frac{1}{|N|!} \sum_{S \subseteq N \setminus \{i\}} |S|! (|N| - |S| - 1)! [v(S \cup \{i\}) - v(S)]$$

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

- **Dropout:** remove one or more features from the data and measure the changes in model performance and behaviors
- Derive from Game Theory
  - Attribute contribution of a feature,  $i$ , on the final prediction
  - Average change when adding/dropping the feature

# Summary




- **Hypothesis driven:** Literature review



Validation  
Idea refinement

- **Data driven:**

- Unsupervised learning
  - Visualization
  - Clustering



Targets for feature  
selection and prediction

- Supervised learning
  - Feature selection
  - Predicted outcomes



Validation and inspection

# Any question?



- See you next time