3000788 Intro to Comp Molec Biol

Week 13: Machine learning

Fall 2024

Sira Sriswasdi, PhD

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

Part 1: Supervised learning

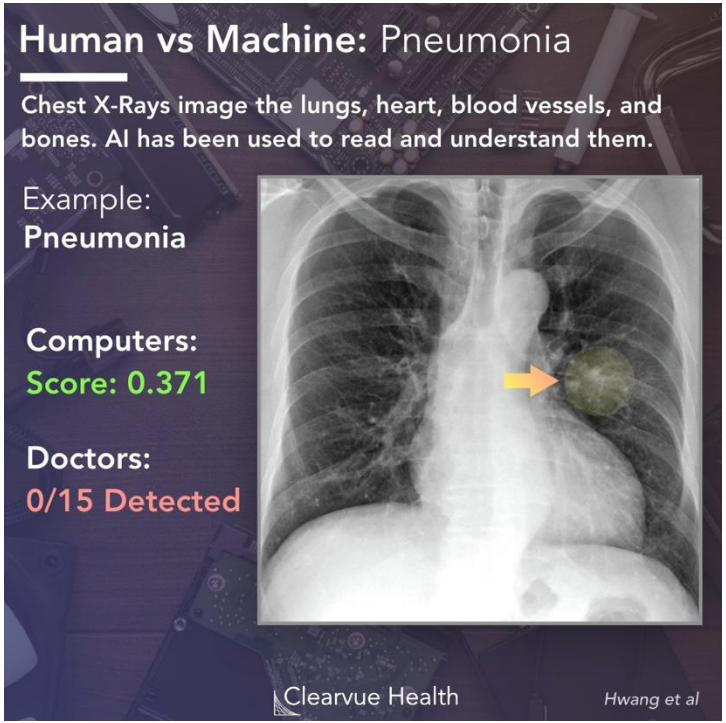
- What is machine learning (ML)?
- Differences between statistics and ML
- Supervised learning = optimizing a function that best fit the observed data
- Examples of supervised ML algorithms

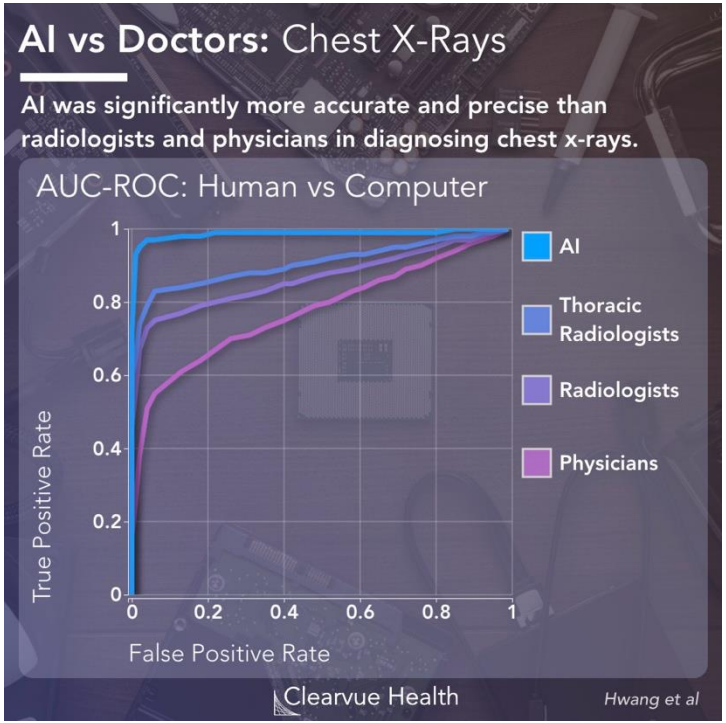
Human vs machine learning

	Human	Machine
Memorization		
Pattern recognition		
Trial and error		
Generalization		
Mechanism of learning		

Memorization (with digitization)

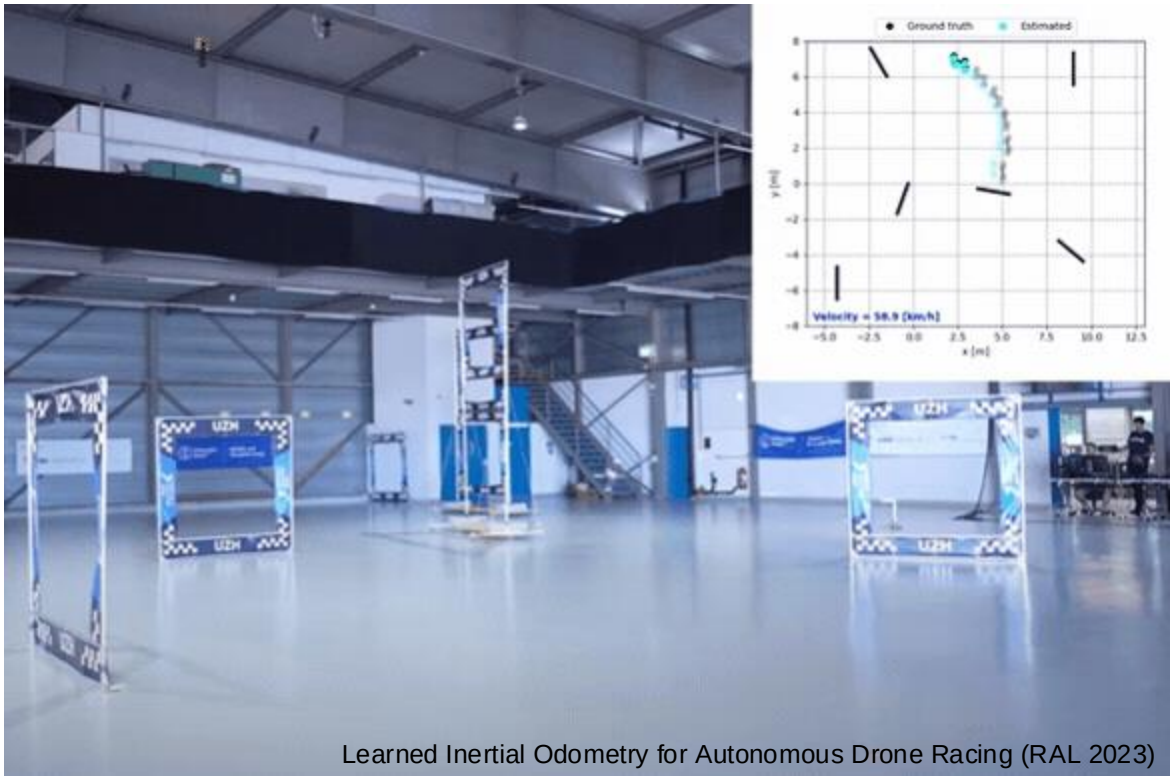
Pattern recognition (with enough data)





https://www.clearvuehealth.com/b/ai-radiology-xray/

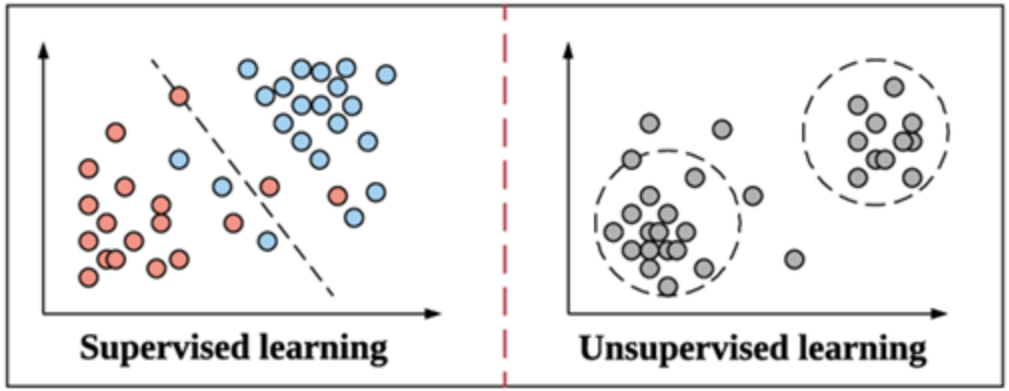
Trial and error (with reinforcement learning)



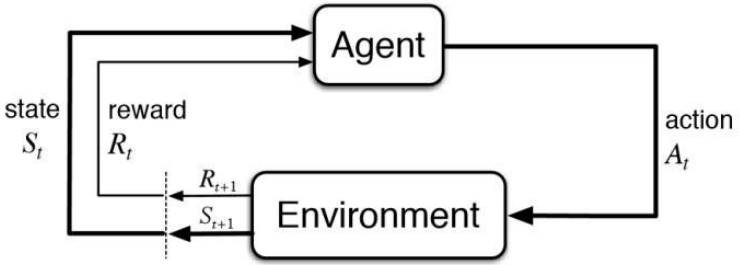
Learned Inertial Odometry for Autonomous Drone Racing (RAL 2023)

Google DeepMind's AlphaGo computer beats top player Lee Sedol for third time to sweep competition

Machine learning paradigms



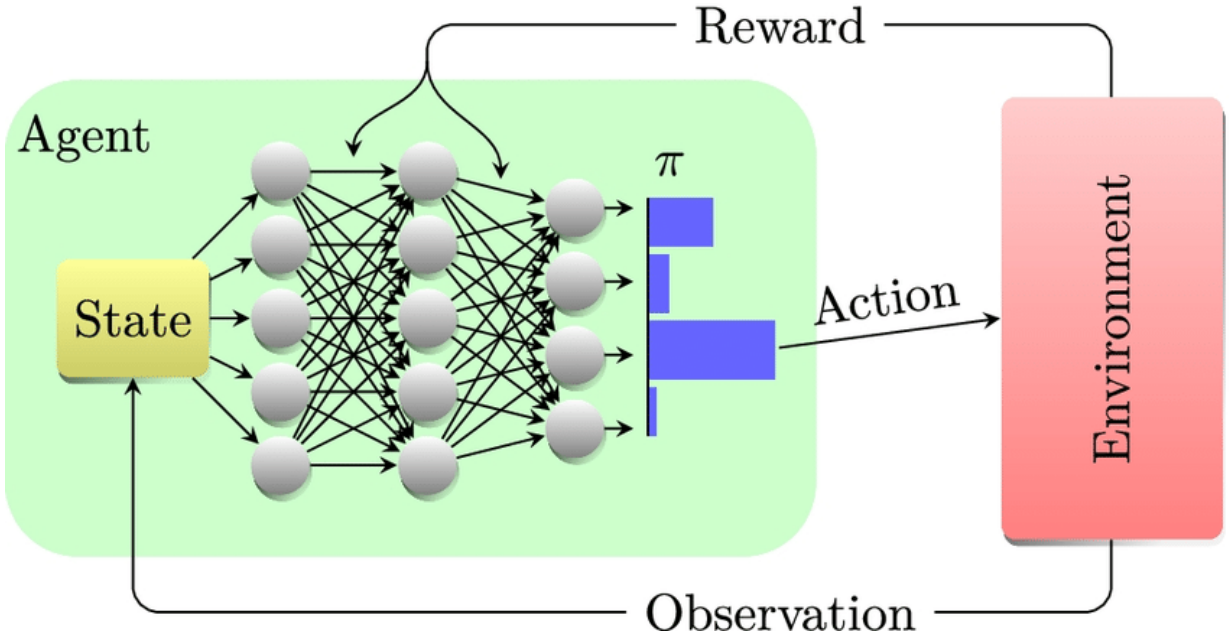
Qian, B. et al. "Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey"



Reinforcement learning

- **Supervised**: Model learns from a dataset (*x*, *y*) to predict *y* from *x*
- **Unsupervised**: Pattern recognition with no target output (only *x*)
- **Reinforcement**: Model learns by interacting and receiving feedbacks from the environment (dynamic data)

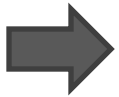
RL = dynamic supervised learning



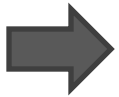
- Action can affect environment → current & future rewards

Machine learning versus human's way of thinking

Data
+
Hypotheses
Knowledge-based

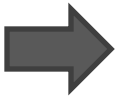


Statistics
Likelihood, goodness-of-fit

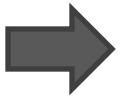


Model that best explains the data

Data
+
Algorithms



Machine Learning
Performance evaluation on unseen data points

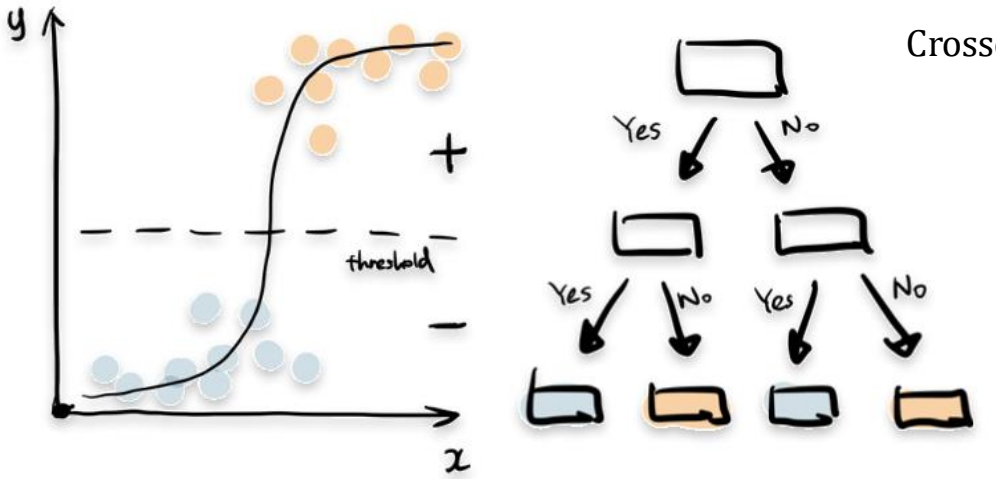


Model that best predicts new data

Supervised learning

The cores of supervised learning

Model



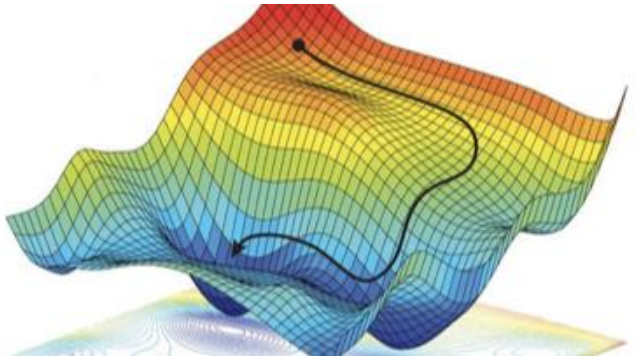
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Objective / Loss Function

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

$$\text{Crossentropy} = -\frac{1}{n}\sum_{i=1}^{n} y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i)$$

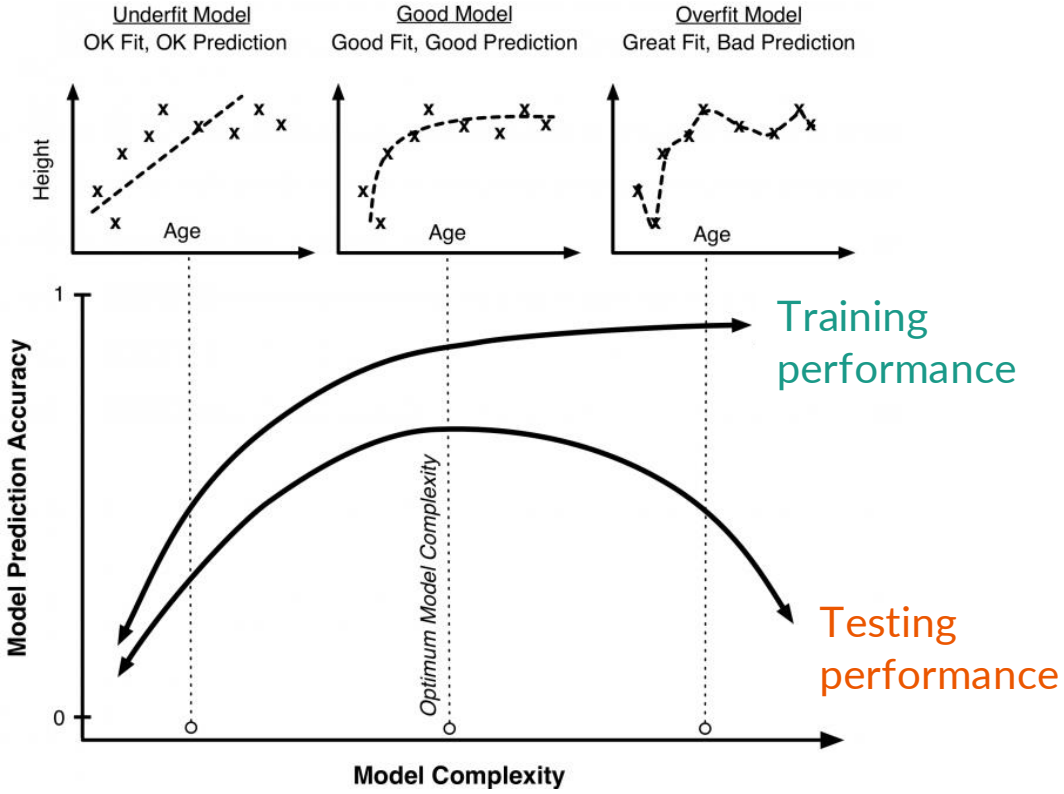
Optimization Algorithm



https://medium.com/analytics-vidhya/gradient-descent-b0dc1af33517

Supervised learning is all about control

https://en.wikipedia.org/wiki/Bull_riding



https://realkm.com/2018/04/23/optimization-and-complexity-the-cost-of-complexity-systems-thinking-modelling-series/

Statistical control of overfitting

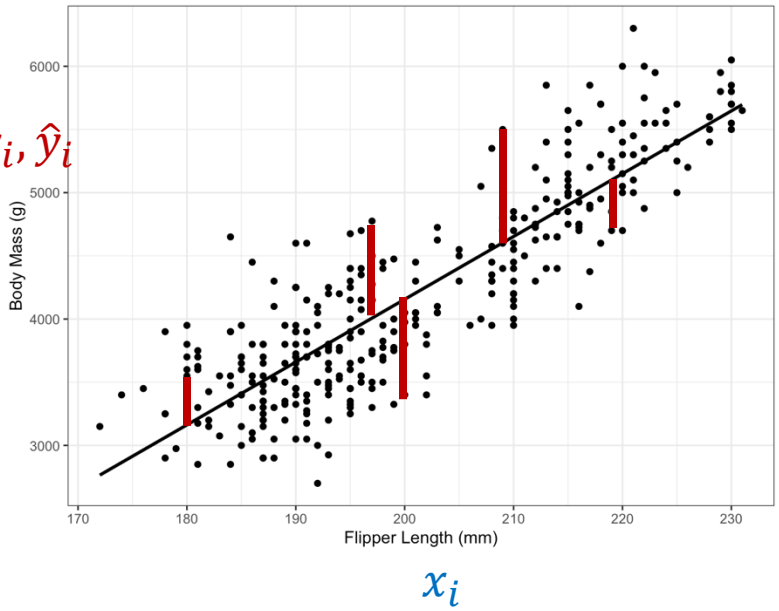
- Better model achieves higher likelihood
- Complex model has more parameters

- **Information Criterion**
  - Akaike (AIC) $= 2k - 2 \cdot \ln(\hat{L})$, where $\hat{L}$ is the likelihood
  - Bayesian (BIC) $= \ln(n)\, k - 2 \cdot \ln(\hat{L})$, where n is the sample size

- **Nested model testing**
  - Simple model has n parameters, fit the data with likelihood $\widehat{L_1}$
  - Complex model has $m > n$ parameters, fit the data with likelihood $\widehat{L_2} > \widehat{L_1}$
  - Is the improvement $\frac{\widehat{L_2}}{\widehat{L_1}}$ worth the increase in $m - n$ parameters?

Linear and logistic regression

Linear regression (Ordinary Least Square)



- Model: $\hat{y_i} = b_0 + b_1 x_i$
- Minimize MSE: $\frac{1}{n}\sum_i (y_i - [b_0 + b_1 x_i])^2$

- $\frac{\delta MSE}{\delta b_o} = -2\sum_i y_i - 2b_1 \sum_i x_i - 2nb_0$
- $\frac{\delta MSE}{\delta b_1} = -2\sum_i x_i y_i - 2b_1 \sum_i x_i^2 - 2b_0 \sum_i x_i$

- $b_0 = \frac{\sum xy \sum x - \sum x^2 \sum y}{(\sum x)^2 - n\sum x^2}$
- $b_1 = \frac{\sum y \sum x - n\sum xy}{(\sum x)^2 - n\sum x^2}$

Ordinary Least Square interpretation

- Observed value = True value + Normally-distributed noise
- **Assumption**: Noises are identical and independent across samples

- Model: $(y_i - \hat{y_i}) \sim N(0, \sigma^2)$
- Density: $\mathrm{P}(y_i - \hat{y_i} = \varepsilon_i \mid \sigma^2) \propto e^{\frac{-{\varepsilon_i}^2}{2\sigma^2}}$
- Likelihood: $\prod_i \mathrm{P}(y_i - \hat{y_i} = \varepsilon_i \mid \sigma^2) \propto e^{\frac{-\sum_i {\varepsilon_i}^2}{2\sigma^2}}$

- MSE: $\frac{1}{n}\sum_i (y_i - \hat{y_i})^2 = \frac{1}{n}\sum_i {\varepsilon_i}^2$
- Minimizing MSE is the same as maximizing likelihood

Logistic regression

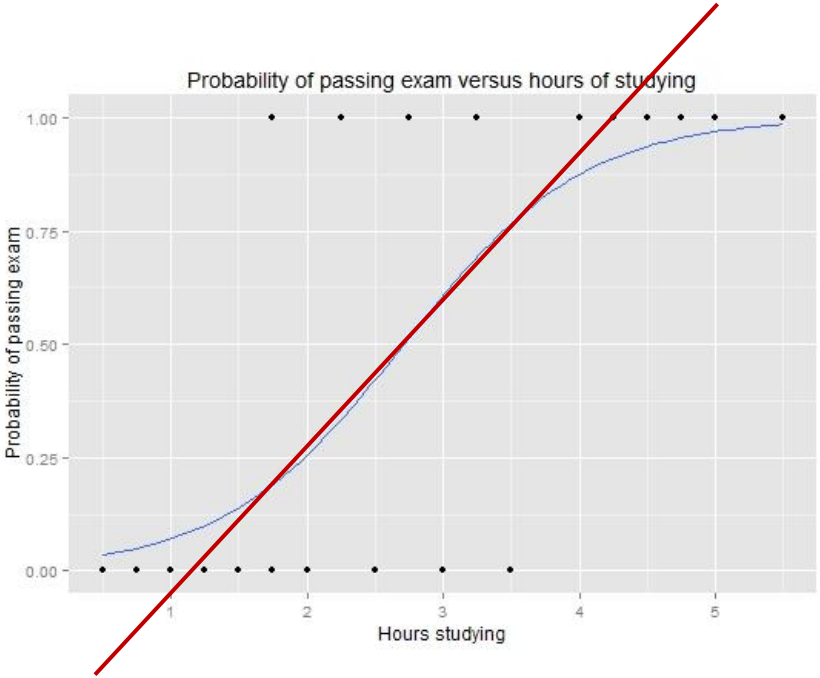


Image from Wikipedia

- Classification output = 0 or 1
- Linear regression outputs $-\infty$ to ∞

- Probability of success p
- Log odd: $\ln\left(\frac{p}{1-p}\right)$
  - $\ln\left(\frac{p}{1-p}\right) \to -\infty$ as $p \to 0$
  - $\ln\left(\frac{p}{1-p}\right) \to \infty$ as $p \to 1$

- Transform linear regression output with log odd!

Logistic regression

- Model: $\ln\left(\frac{\widehat{y_i}}{1-\widehat{y_i}}\right) = f(x_i) = b_0 + b_1 x_{i,1} + \cdots + b_n x_{i,n}$
- $\widehat{y_i} = \frac{e^{b_0+b_1 x_{i,1}+\cdots+b_n x_{i,n}}}{1+e^{b_0+b_1 x_{i,1}+\cdots+b_n x_{i,n}}}$
  - When $f(x_i) \to \infty$, $\widehat{y_i} \to 1$
  - When $f(x_i) \to -\infty$, $\widehat{y_i} \to 0$
- Can we keep using MSE as the loss function?
  - Brier score $= \frac{1}{N}\sum_i (y_i - \widehat{y_i})^2$
  - But this does not interpret logistic output as probability

Likelihood for logistic regression

- Likelihood: $\mathrm{P}(y_i \mid x_i) = \hat{y}_i^{\,y_i}(1-\hat{y}_i)^{1-y_i}$
  - y_i is either 0 or 1
  - When y_i is 0, the likelihood is $1-\hat{y}_i$
  - When y_i is 1, the likelihood is $\hat{y}_i$

- Log likelihood: $y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)$
  - This is the cross-entropy loss function!
  - Maximizing likelihood is the same as minimizing cross-entropy

Regularization of linear model

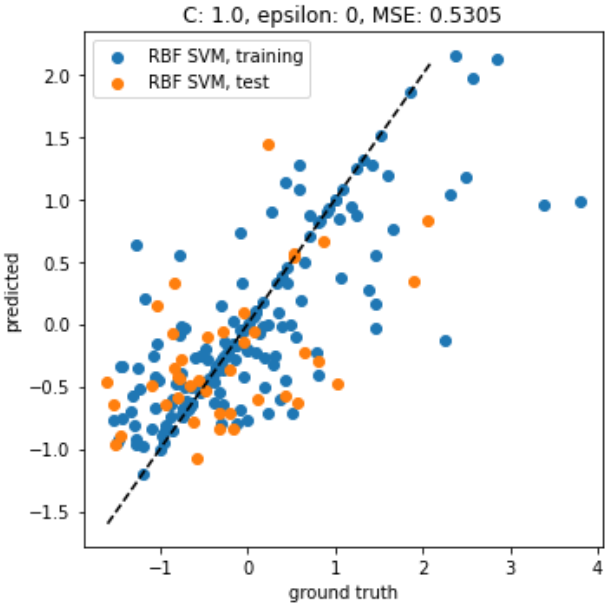
- L1 regularization (LASSO): MSE + $\alpha \sum_k |b_k|$
- L2 regularization (Ridge): MSE + $\alpha \sum_k {b_k}^2$

- α is the hyperparameter that controls the regularization strength

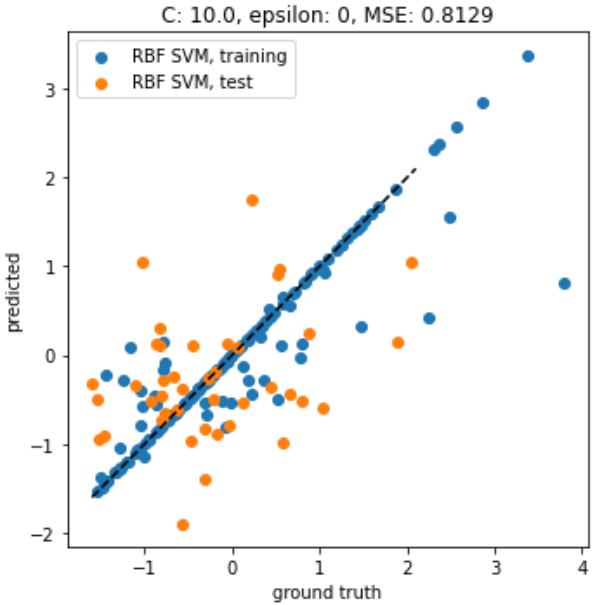
- Hyperparameter must be tuned for every dataset!

Tuning regularization strength

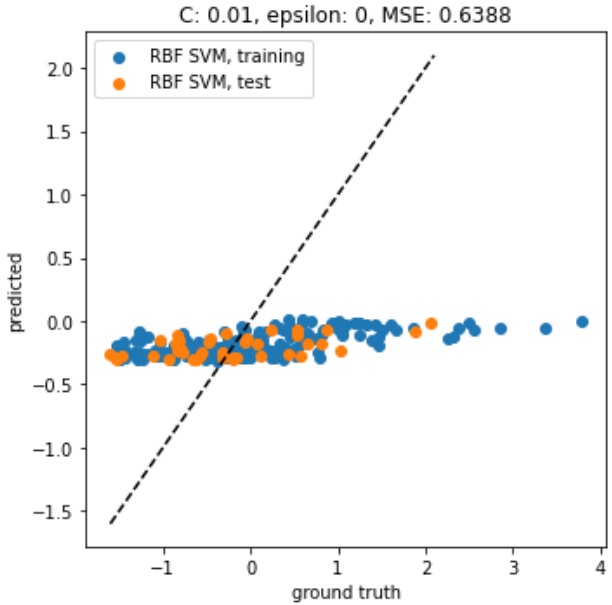
Just right



Too weak

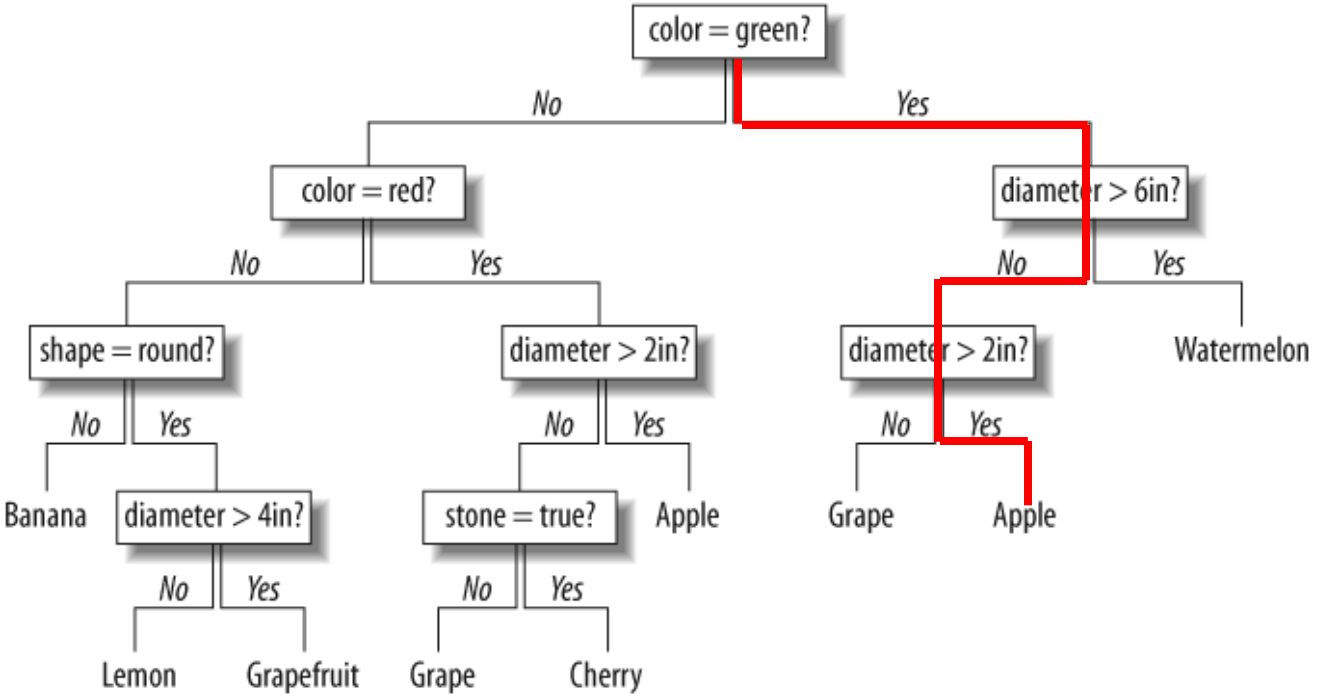


Too strong



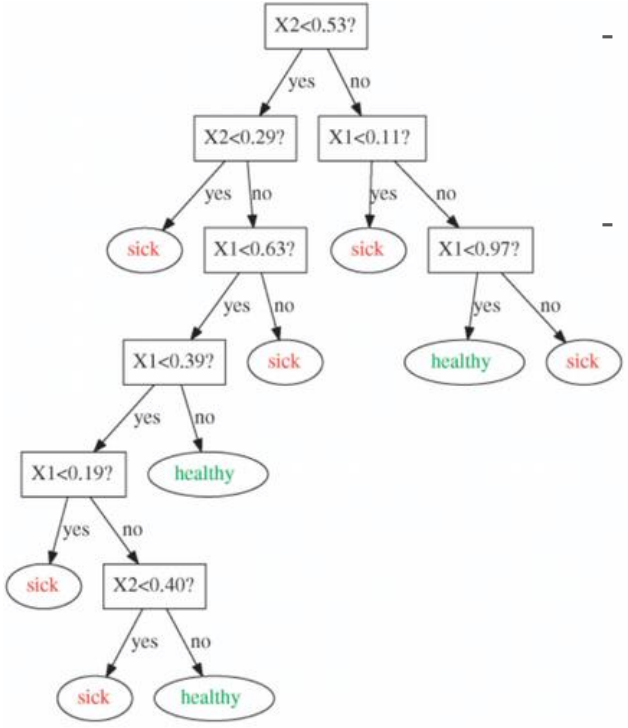
Decision tree

Decision tree

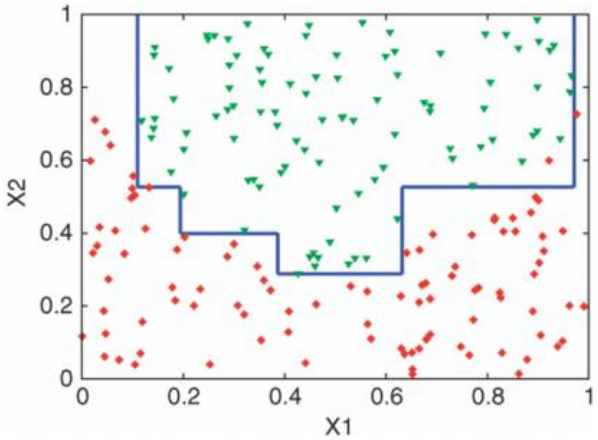


Source: Programming collective intelligence by Toby Segaran

Decision tree behaviors



- Each decision is a threshold on each feature
  - Piecewise linear
  - Parallel to an axis
- Good for criteria-based classification



Miller, C. “Screening meter data: Characterization of temporal energy data from large groups of non-residential buildings”

Splitting quality

100 healthy, 100 sick

85 healthy, 5 sick

15 healthy, 95 sick

45 healthy, 51 sick

55 healthy, 49 sick

- Gini impurity: $\sum p(1-p)$
- Entropy: $-\sum p \ln(p)$
  - Minimal at p = 0 or 1 → Perfect split
  - Maximal at p = 0.5 → 50-50 split
- Search for feature and cutoff that yield lowest impurity or entropy

Control mechanisms for tree building

1. Too few samples to make a split

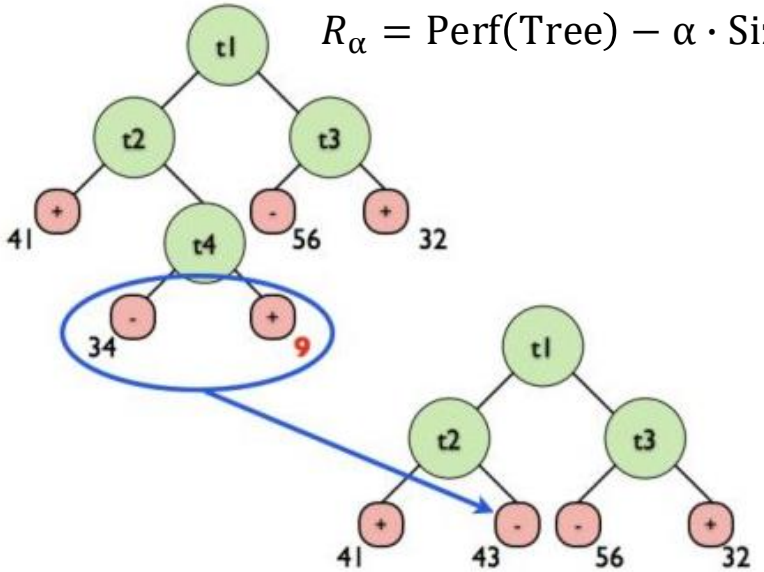
3. Impurity or entropy does not change much after the split

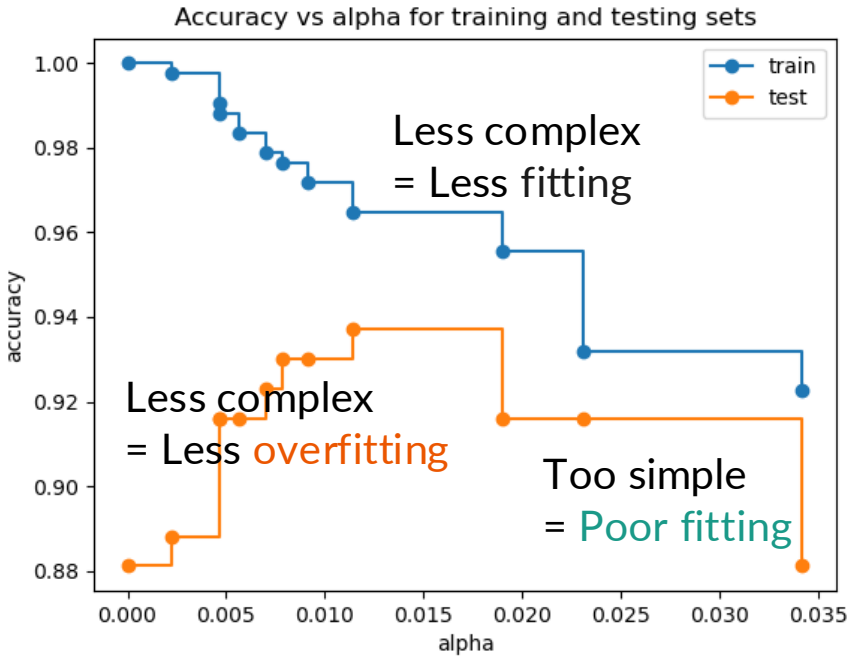
2. Too few samples on either branch

- Limit the tree size
- Limit the improvement in quality
- Limit the number of samples that support a split

Tree pruning (post-processing)

$$R_\alpha = \text{Perf}(\text{Tree}) - \alpha \cdot \text{Size}$$





Patel, N. and Upadhyay, S. “Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA” IJCA 2012

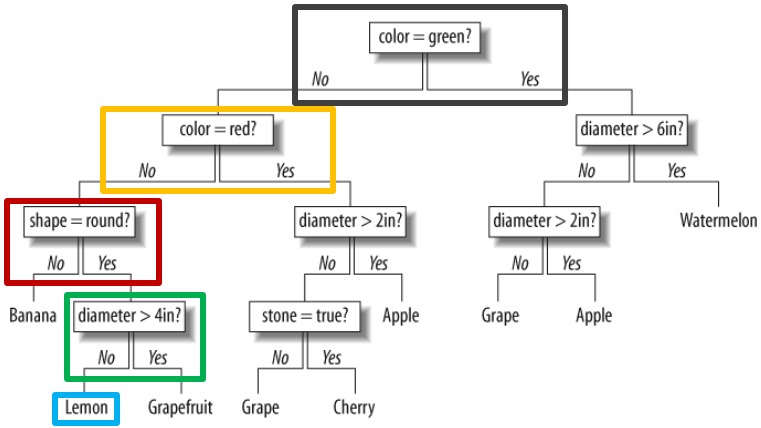
Image from scikit-learn,org

Regularization on features

- **Linear model:** $\hat{y}_i = b_0 + b_1 x_{i,1} + \cdots + b_n x_{i,n}$
  - LASSO

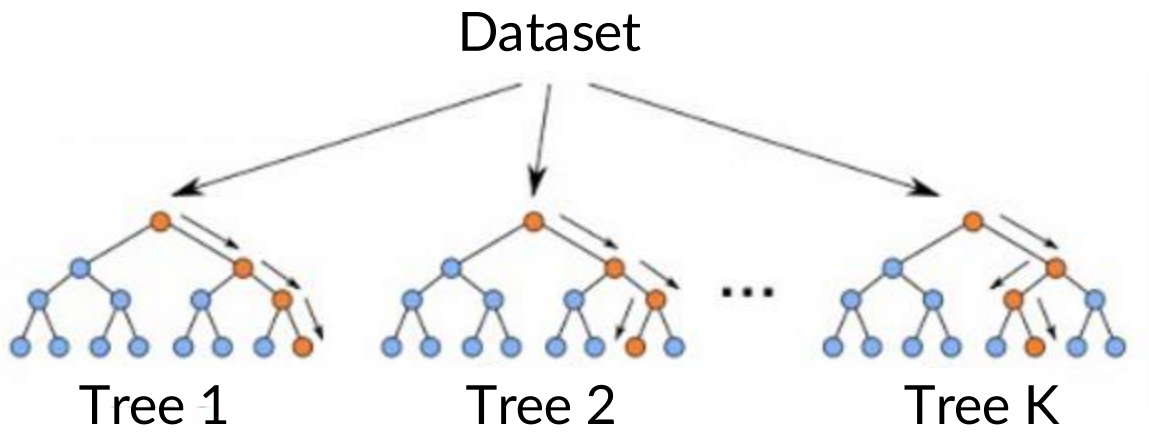
- **Tree model:**
  - Repeatedly using the same feature
  - Early decision affects the rest

- Feature bagging
  - Look at only *N* features at each step
  - Force model to use diverse features



Ensemble approaches

Bagging: Random forest



- Sample 80% of the dataset to train each decision tree
- Each tree may overfit to different part of the dataset
- But the consensus should be correct

Boosting

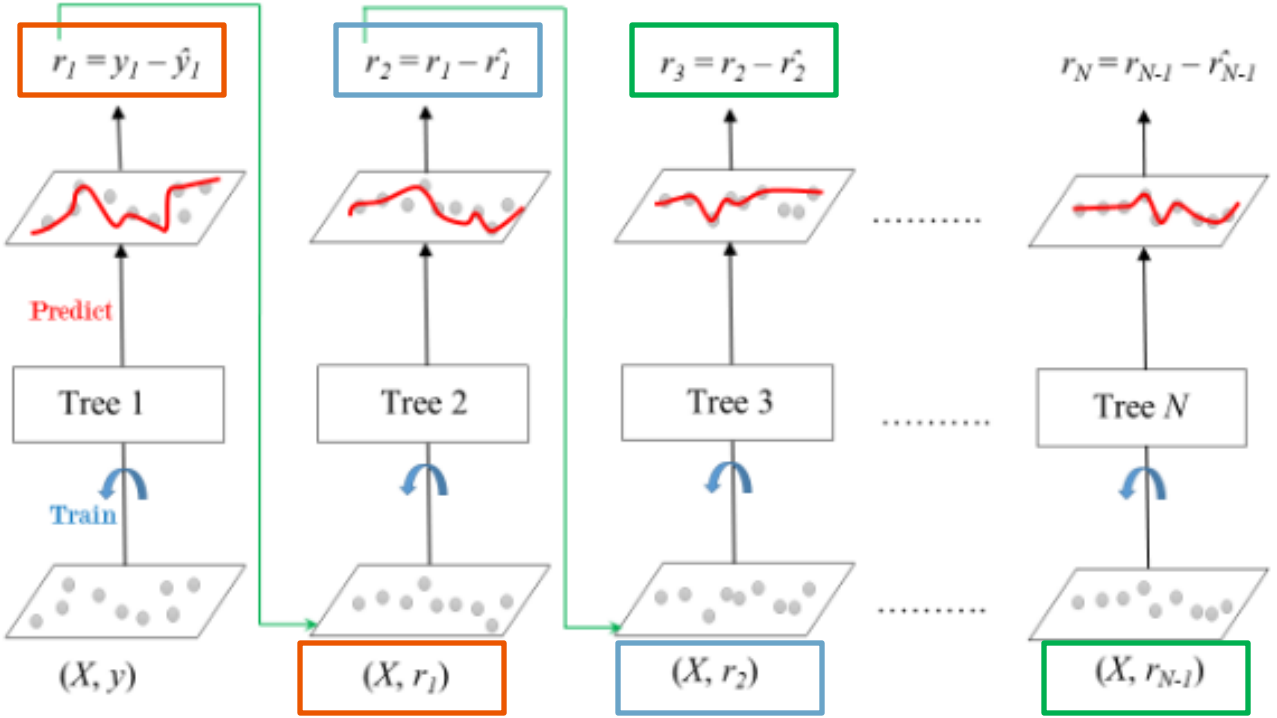
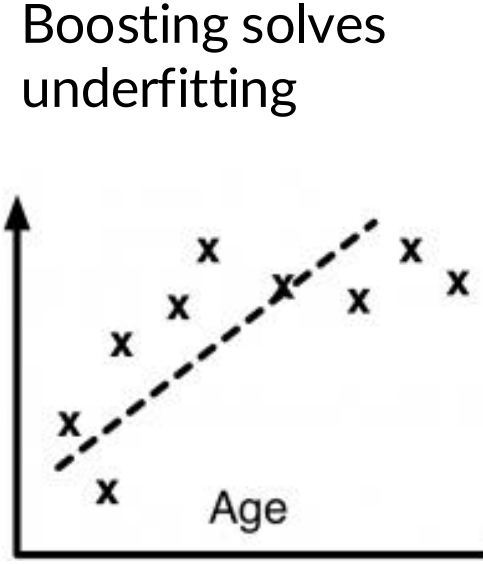
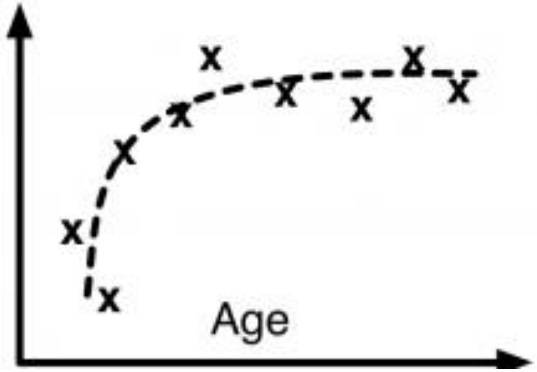


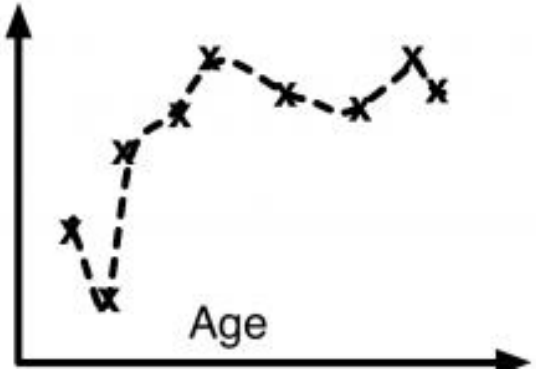
Image from geeksforgeeks.org

Impact of ensemble

Boosting solves underfitting





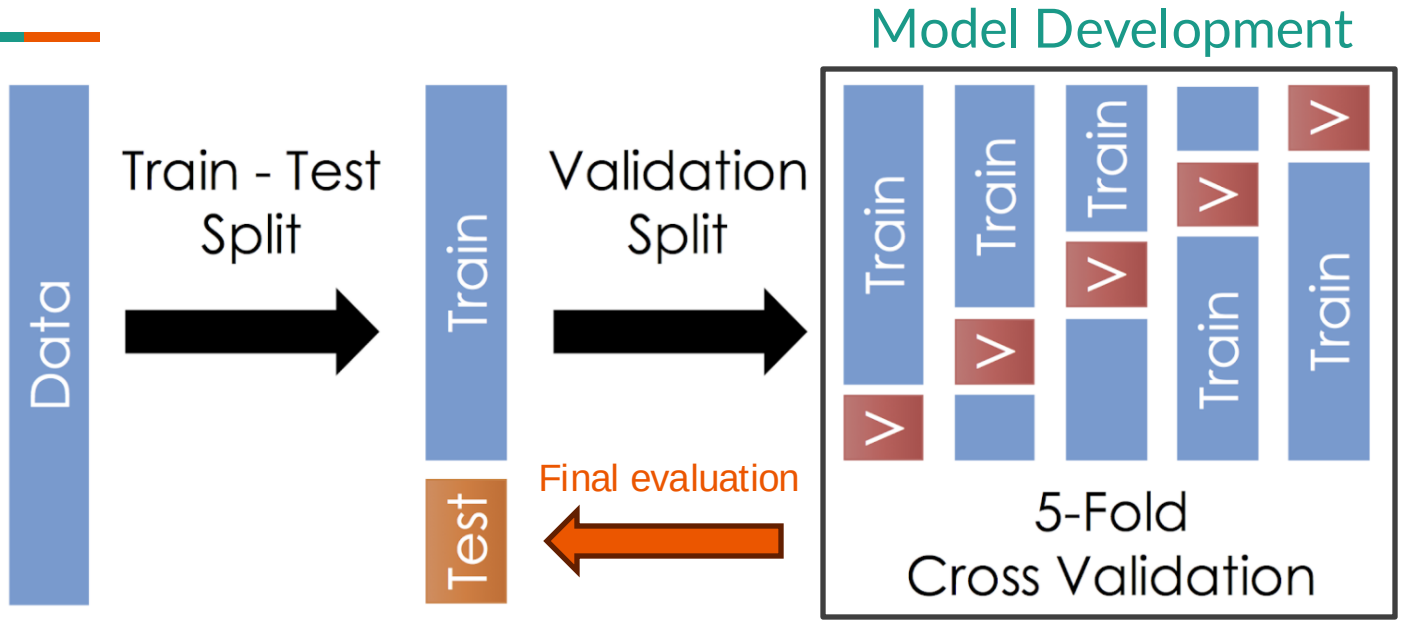


Bagging prevent overfitting

https://realkm.com/2018/04/23/optimization-and-complexity-the-cost-of-complexity-systems-thinking-modelling-series/

Model evaluation and explanation

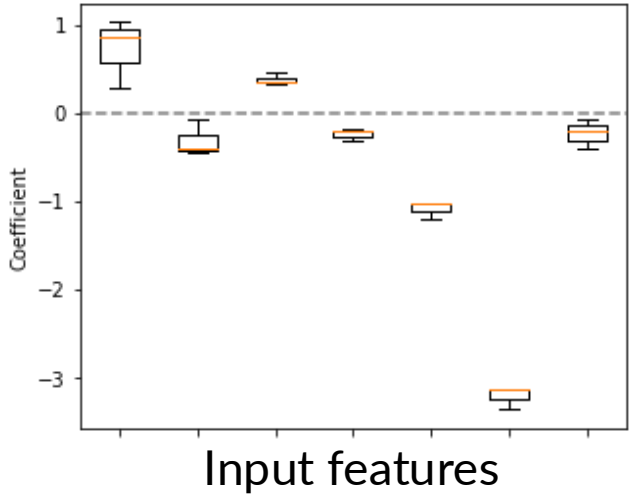
Train-Val-Test

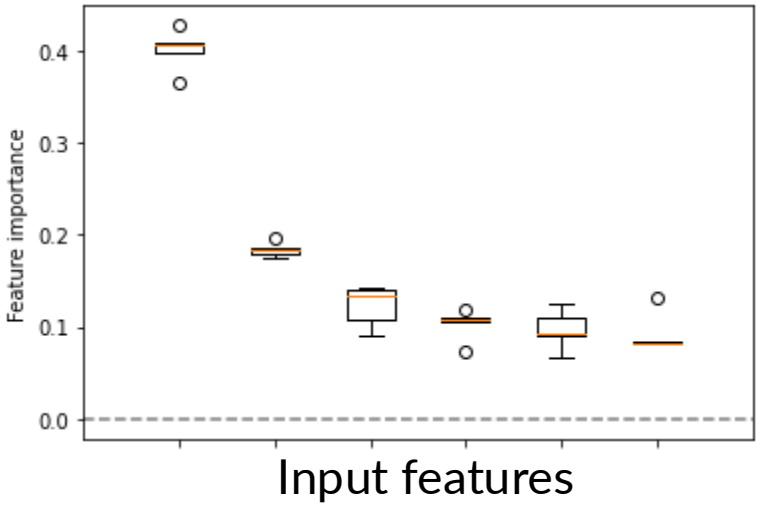


Source: medium.com

- **Training** data determines the best coefficients / weights
- **Validation** data determine the best hyperparameters
- **Test** data determine performance on new datasets

Feature importance

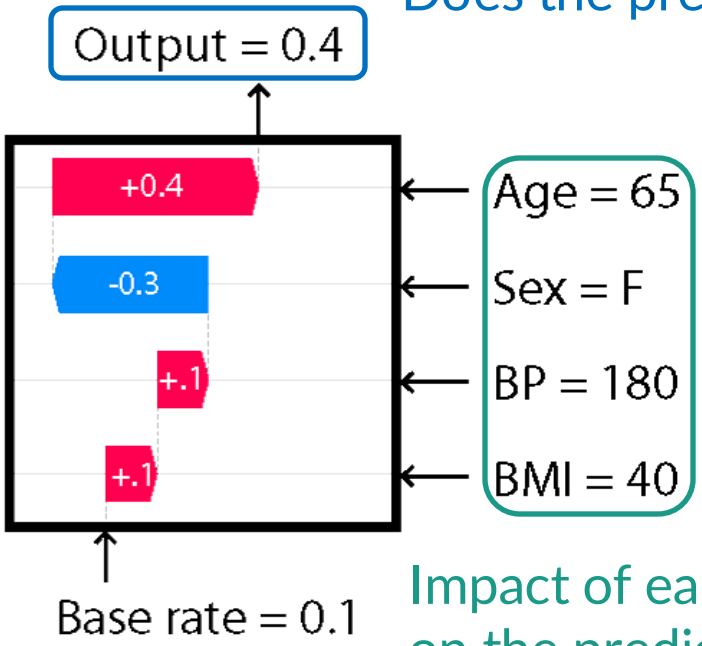




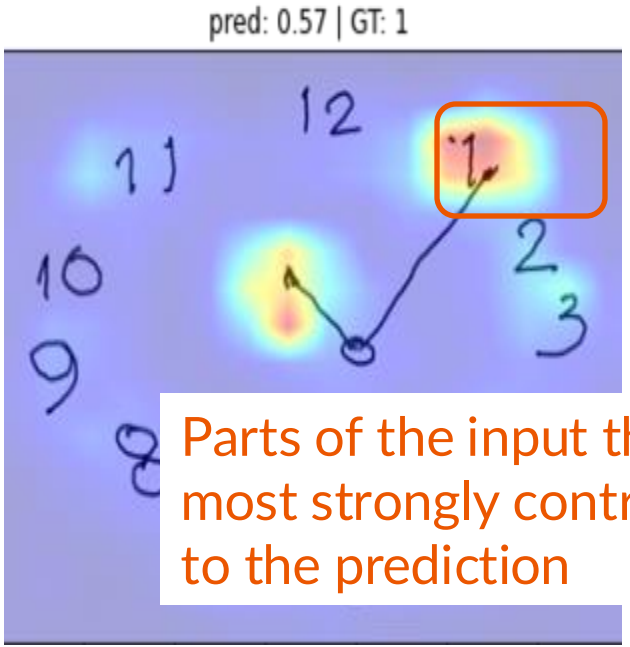
- Coefficients of linear, logistic, and SVM models
- Average improvement in impurity or entropy in tree models
- Model-level explanation

Explainability

Does the predicted confidence match your expectation?



Impact of each input on the prediction



Parts of the input that most strongly contribute to the prediction

Part 1: Unsupervised learning

- Even without target output, the machine can learn from the data
- Similarity (homology) is the key!
  - **Example**: BLAST lets you identify protein families
- Clustering
- Dimensionality reduction

Similarity

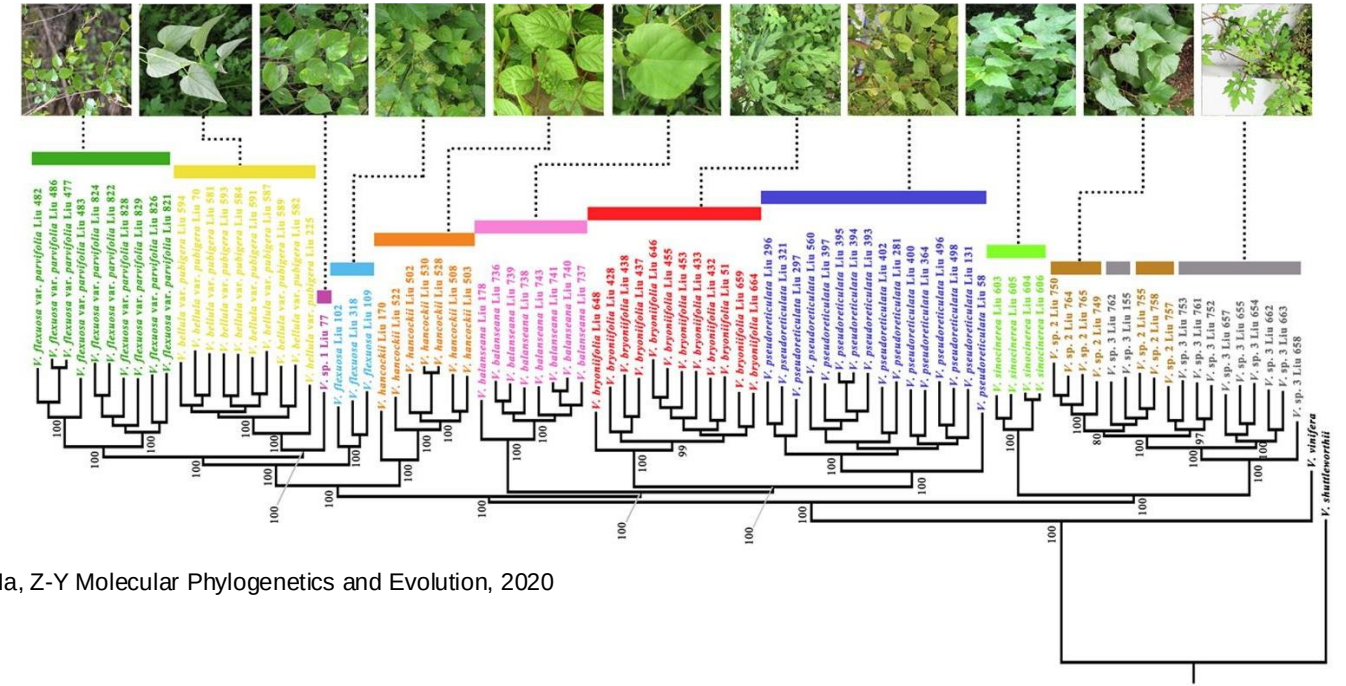
A proxy for causal relationship

- **Same disease** → Similar phenotypes across patients
- **Same biological function** → Similar amino acid sequences across proteins
- **Spiciness from red chili** → Similar red colors across dishes
- **Related evolution** → Similar genomic sequences across species

Inference through similarity

- Patients with similar phenotypes → **Same disease**?
- Proteins with similar amino acid sequences → **Same function**?
- Dishes with red colors → **Spicy**?
- Species with similar genomic sequences → **Evolutionary relation**?

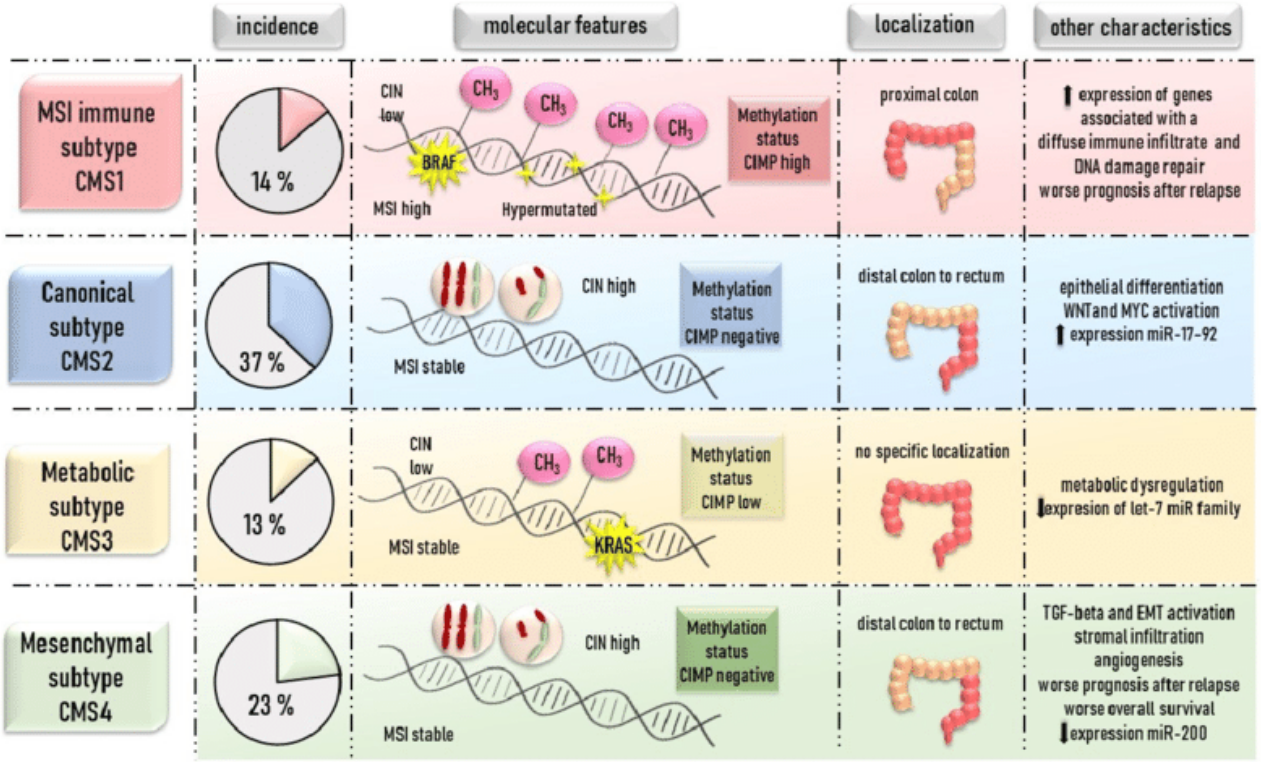
Phylogenetics: Clustering of similar genomes



Ma, Z-Y Molecular Phylogenetics and Evolution, 2020

- Plants in the same group possess similar morphology

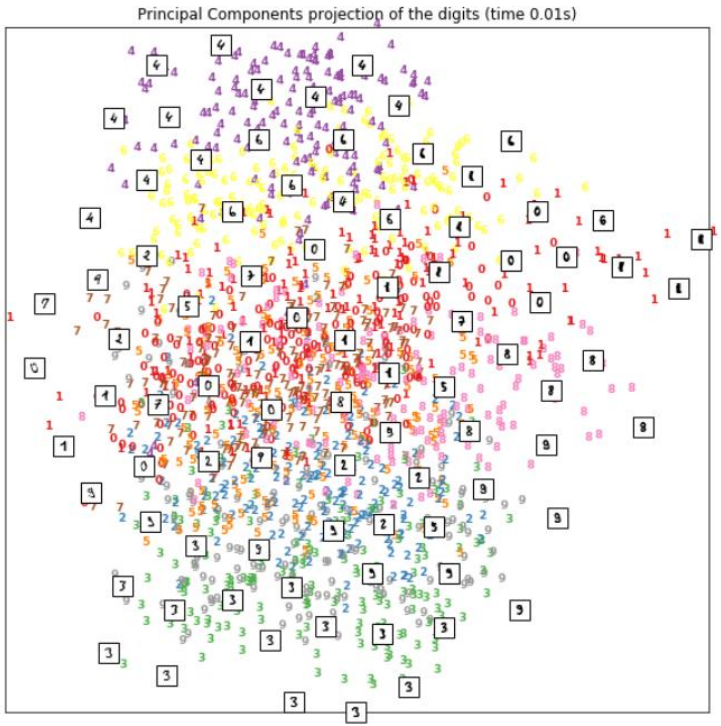
Cancer subtyping: Clustering of molecular signatures

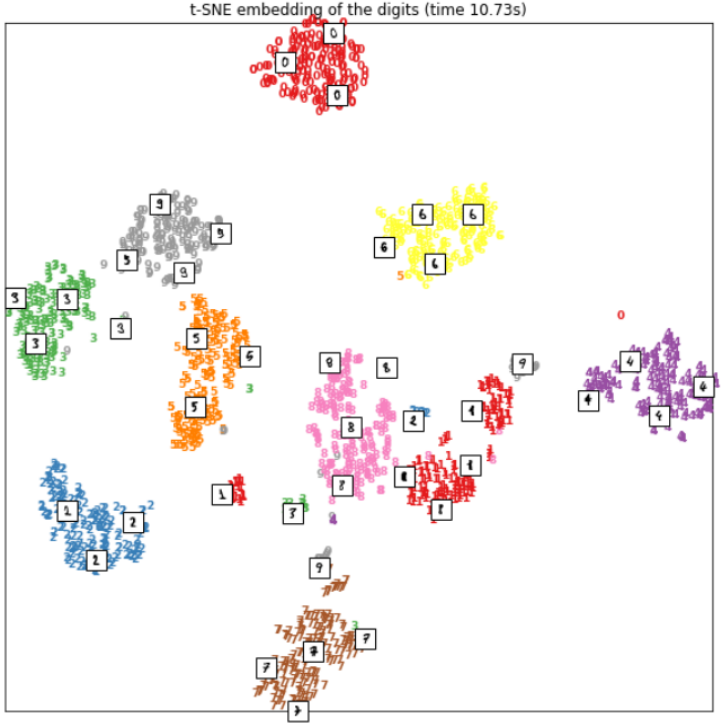


Distinct prognoses and treatment choices

Cervena, K. Cells, 2020

Visualization guided by similarity

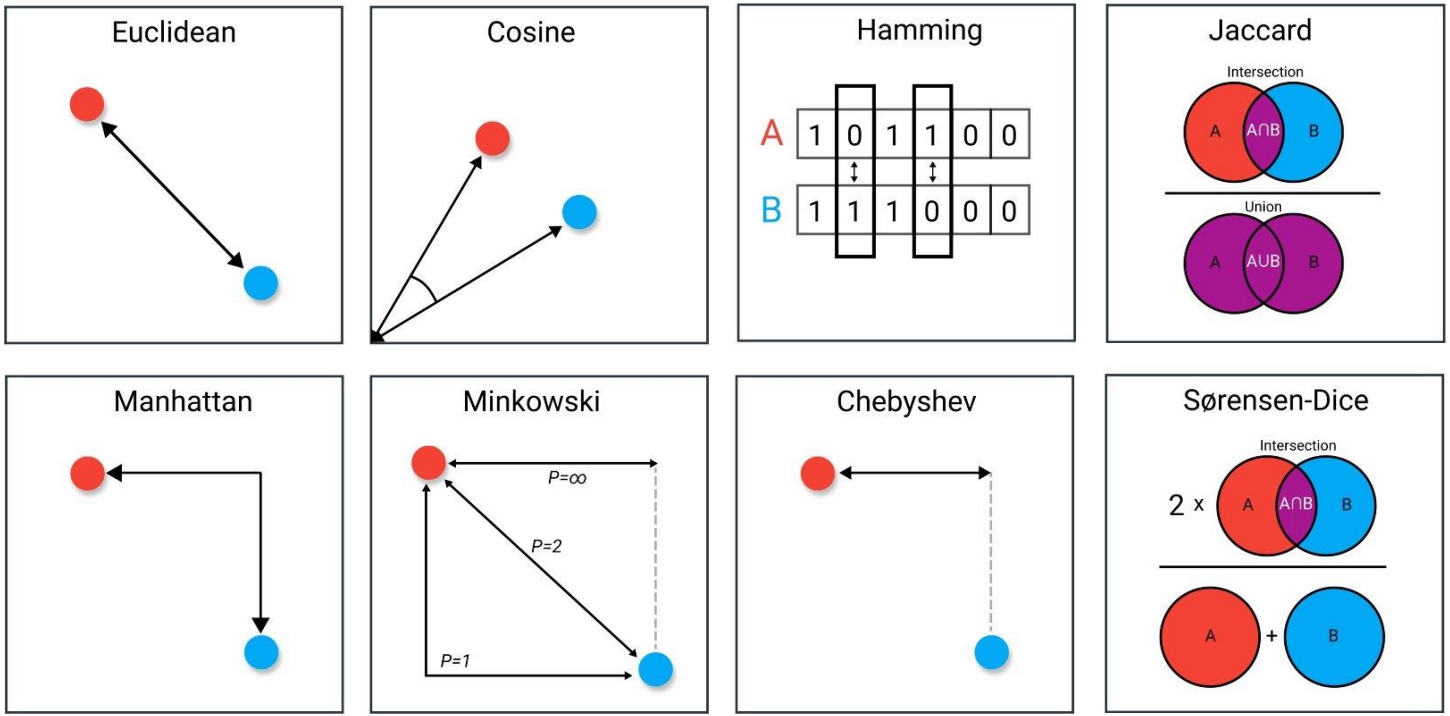




https://medium.com/ai-innovation/interactive-data-visualization-cd79927efc71

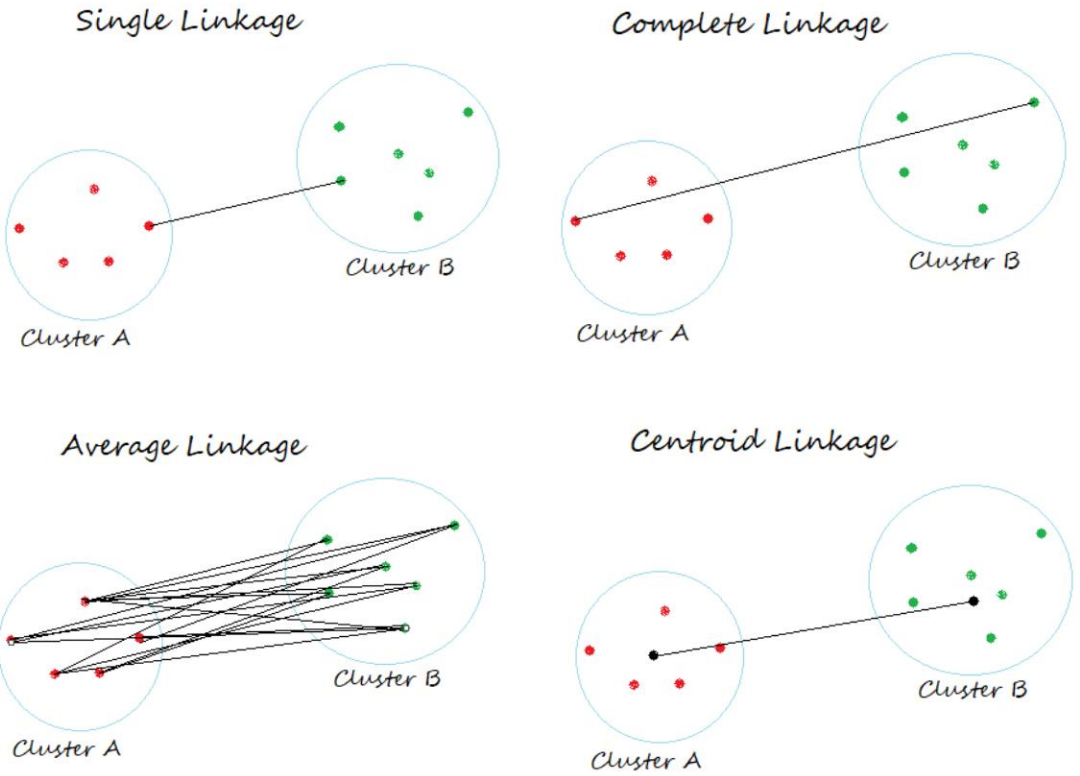
Definitions for similarity

Mathematical definitions for similarity



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Similarity between groups



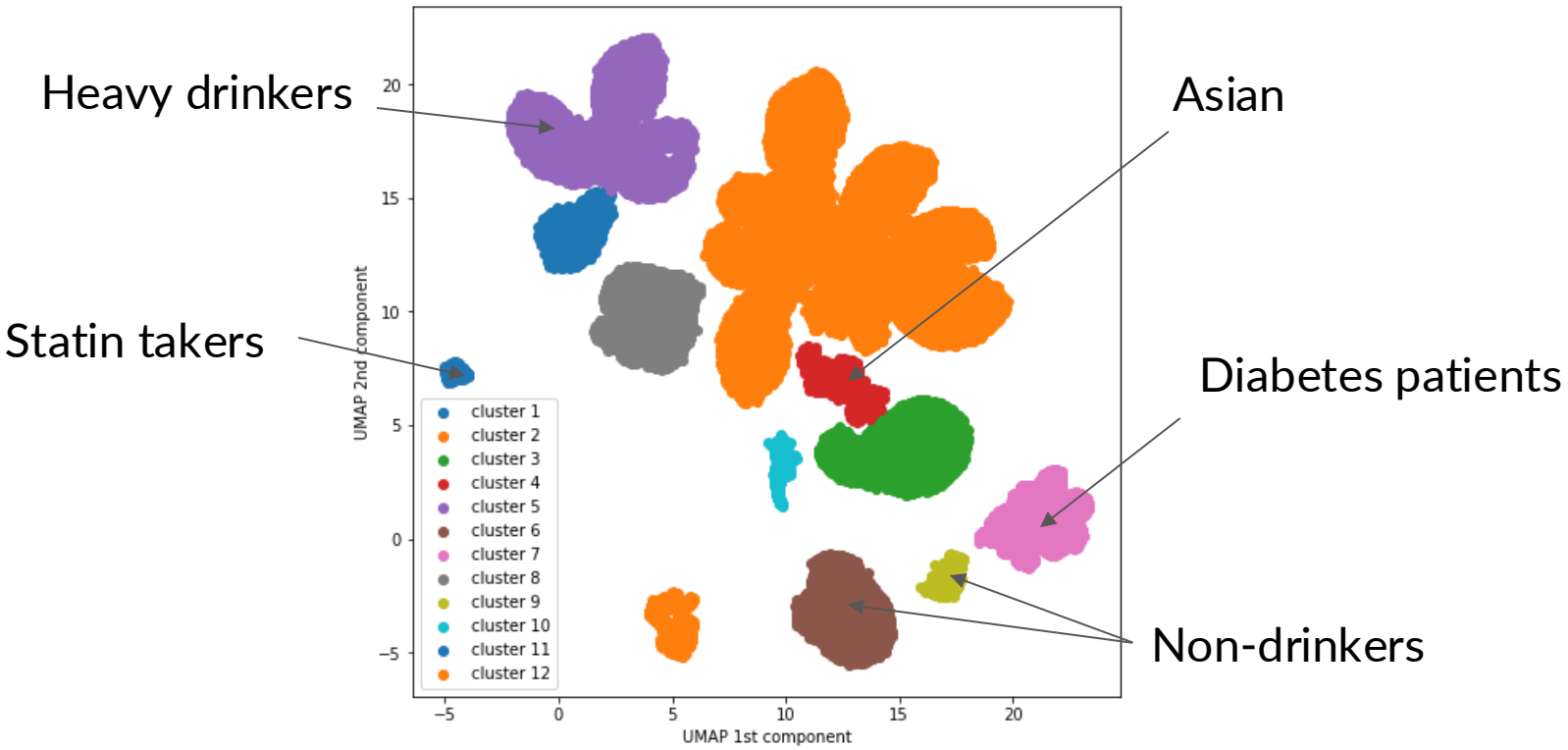
https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/

Similarity on different aspects

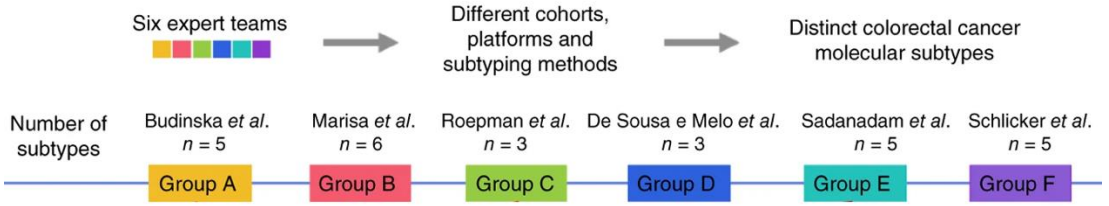
- Red **color** → Spiciness
- Oily **look** → High calorie
- Similar **ingredients** → Similar cuisine

Why unsupervised learning?

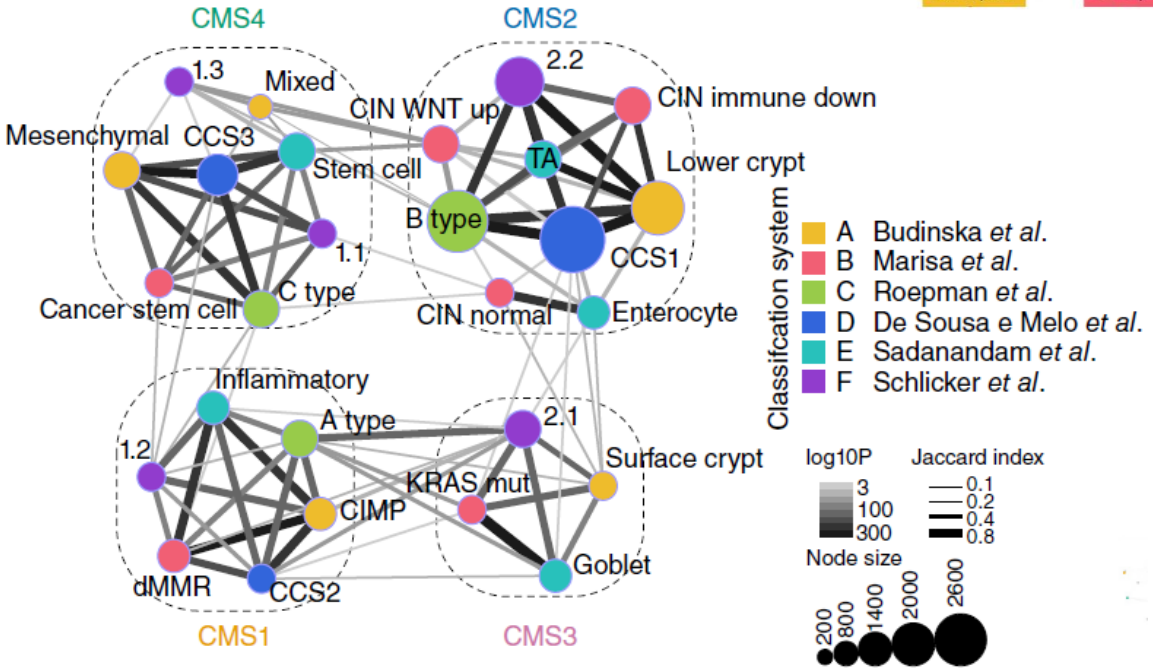
It is unbiased and can work without prior knowledge



Chance to discover



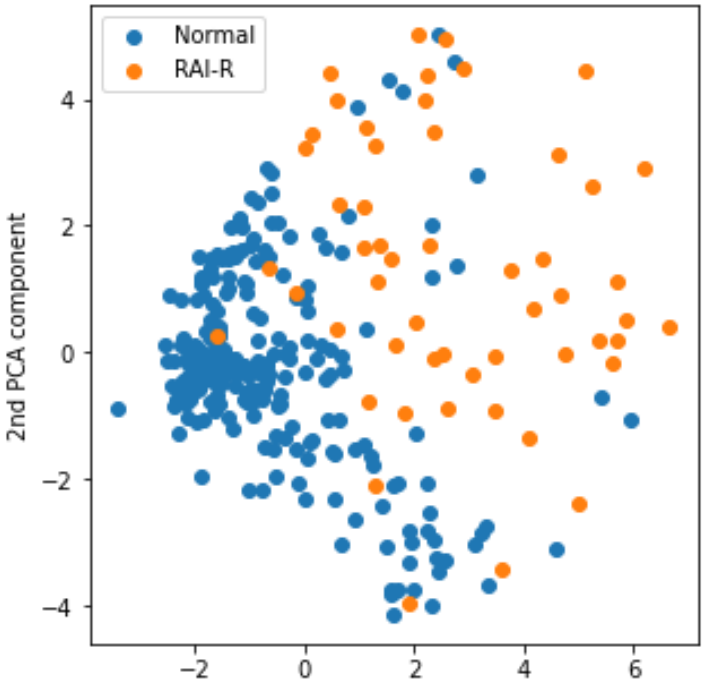
Guinney *et al*. Nat Medicine 21: 1350-1356 (2015)



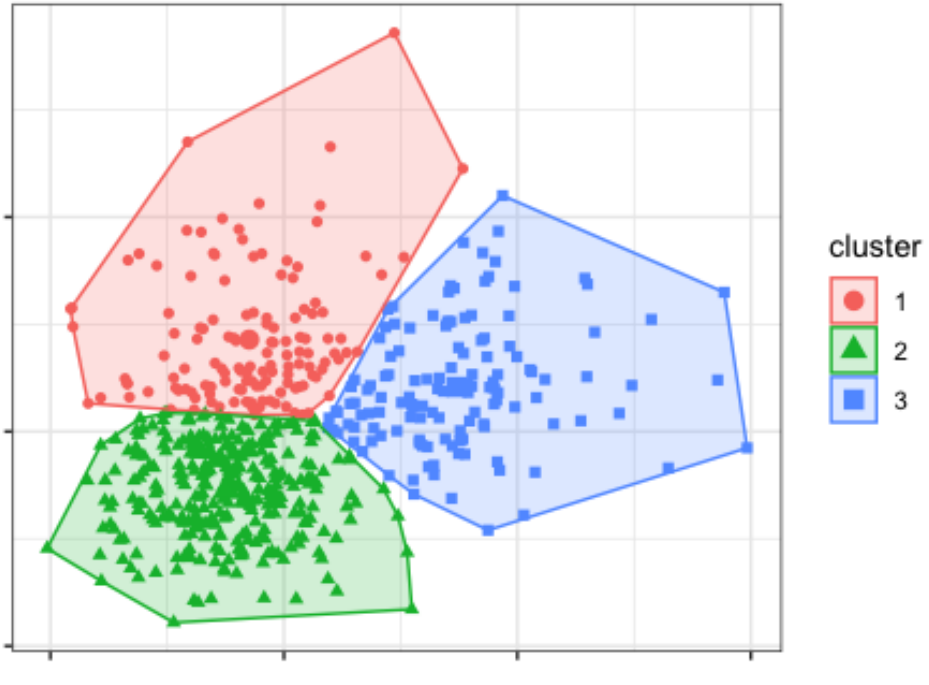
- Compare 6 cancer subtyping systems
- Found 4 consensus groups
- Develop new classifier

Key domains in unsupervised learning

Dimensionality Reduction

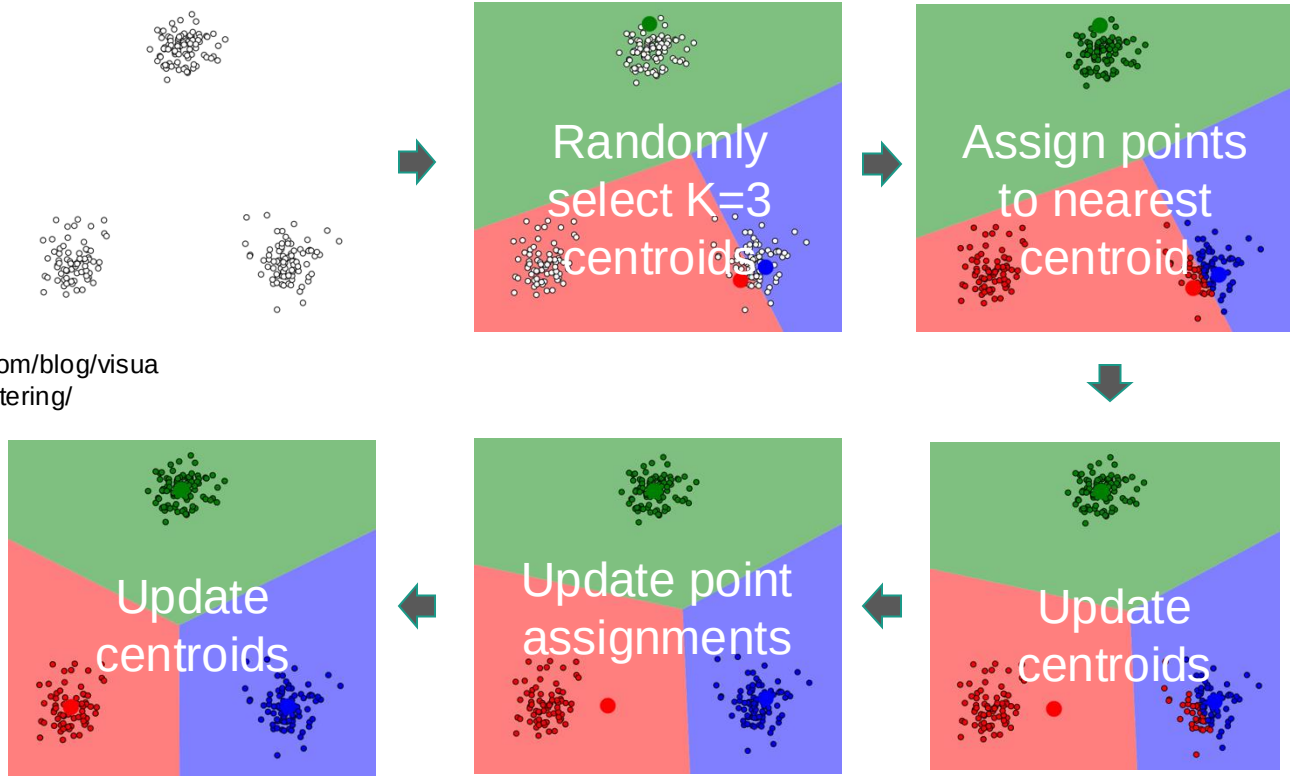


Clustering & Anomaly Detection



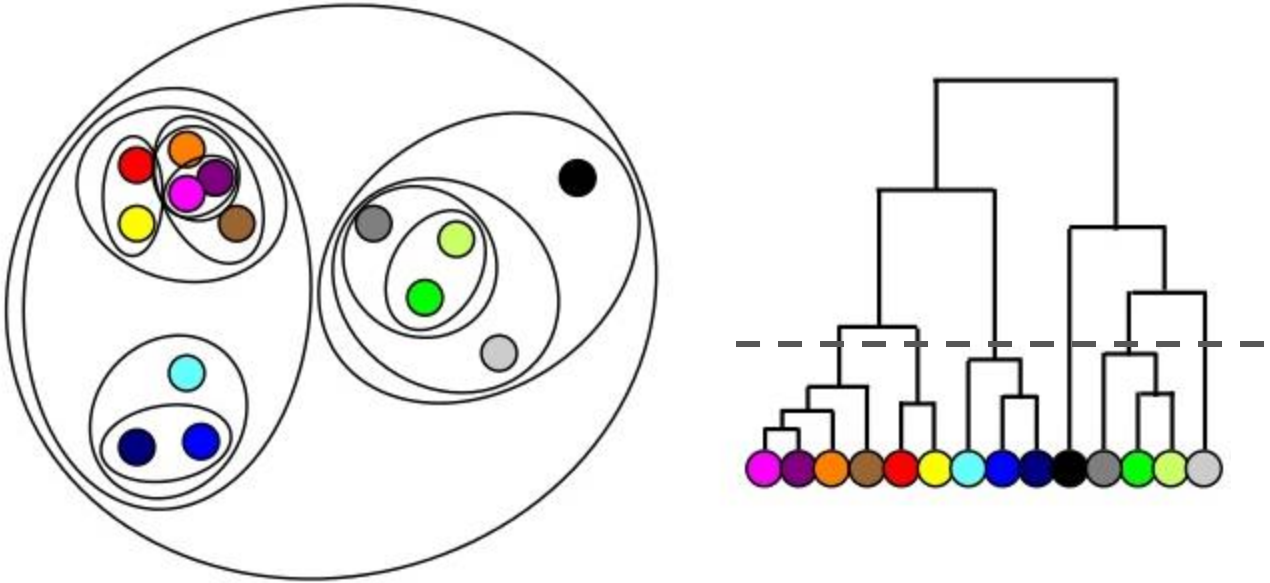
Clustering (and anomaly detection)

k-mean clustering



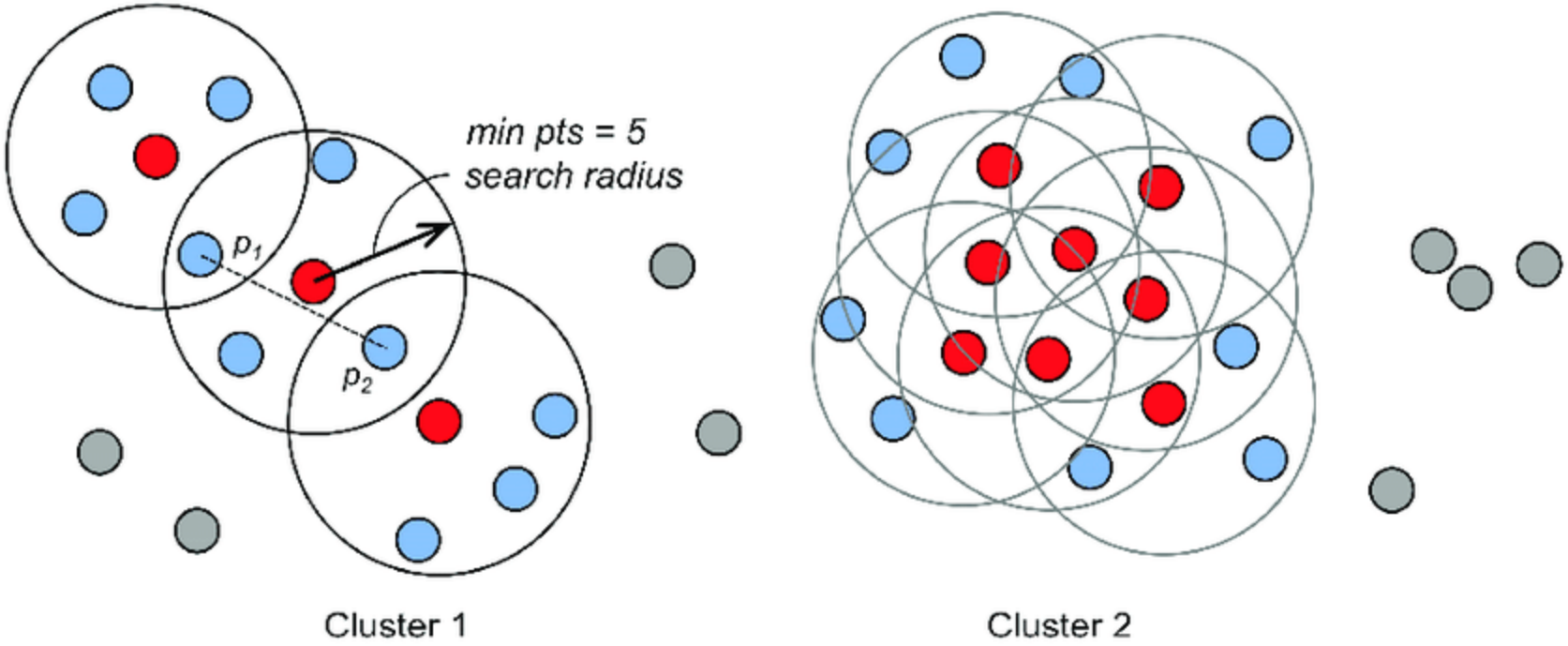
www.naftaliharris.com/blog/visualizing-k-means-clustering/

Agglomerative / Hierarchical clustering



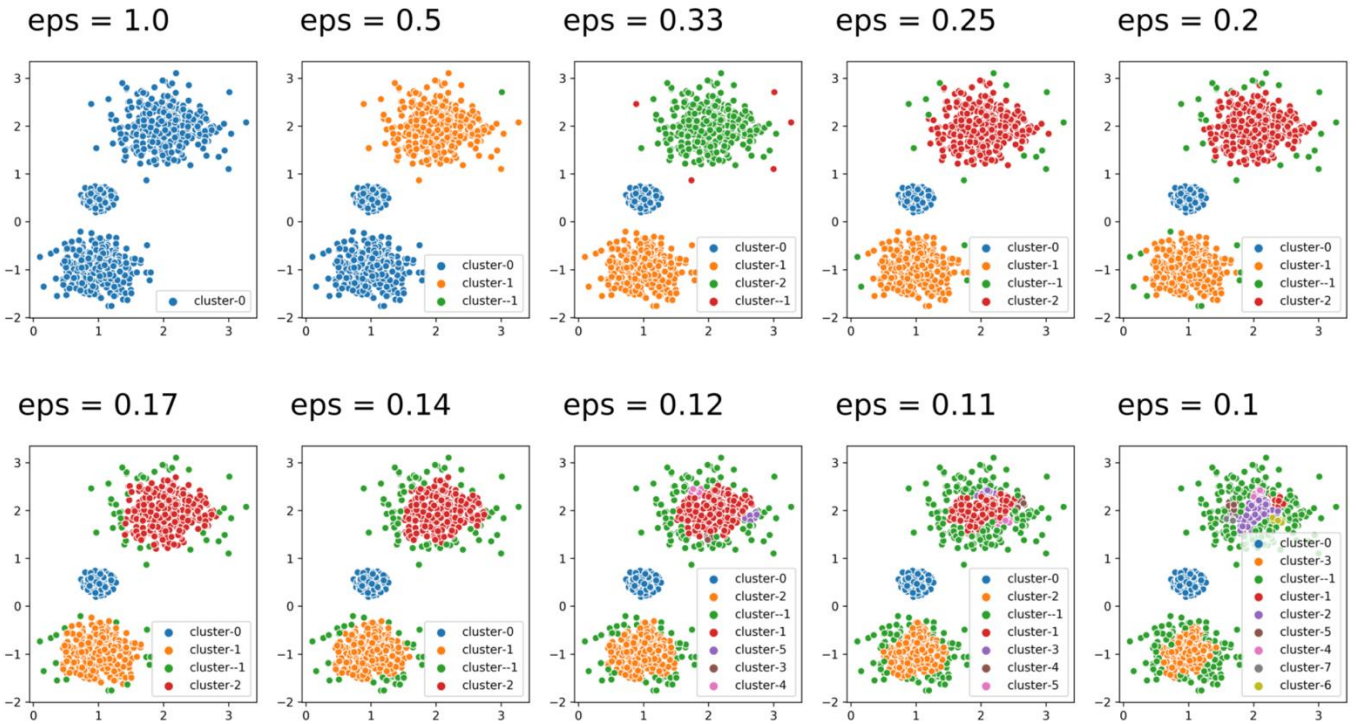
Source: www.slideshare.net/ElenaSgis/data-preprocessing-and-unsupervised-learning-methods-in-bioinformatics

DBSCAN: A density-based technique



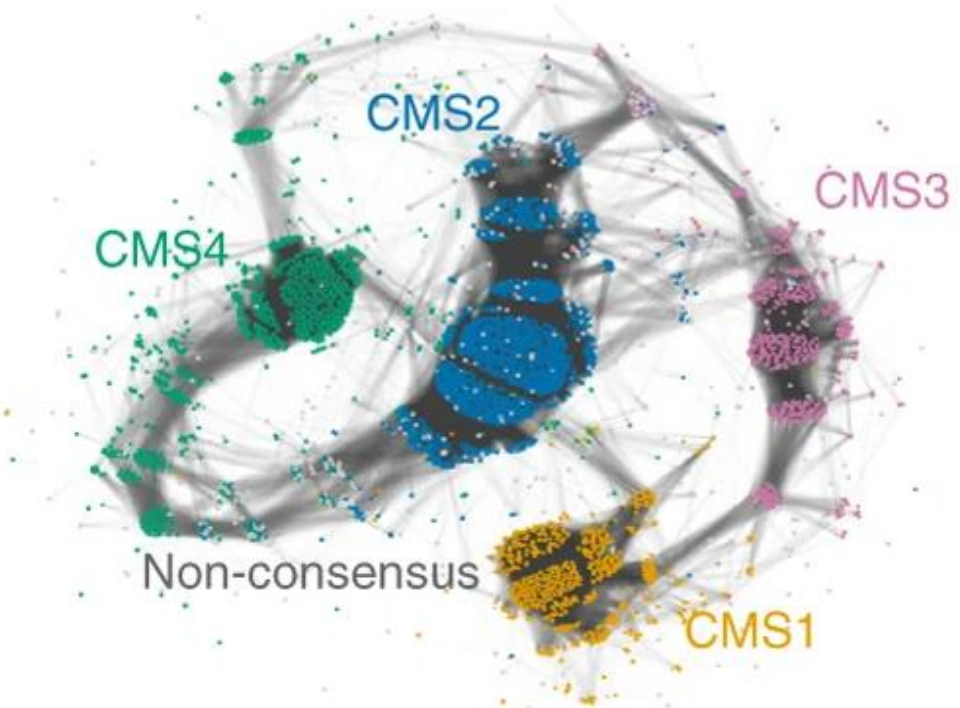
Difrancesco, P.-M. Remote Sensing 12:1885 (2020)

Simultaneous detection of clusters and outliers

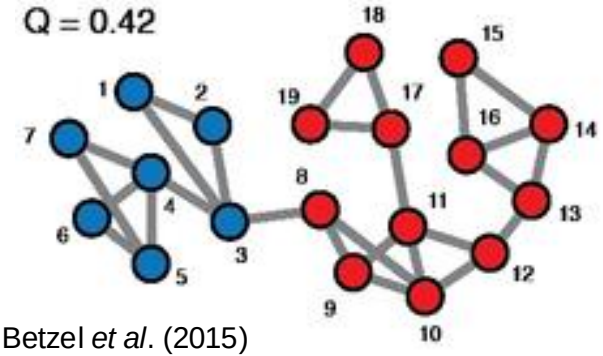


https://towardsdatascience.com/

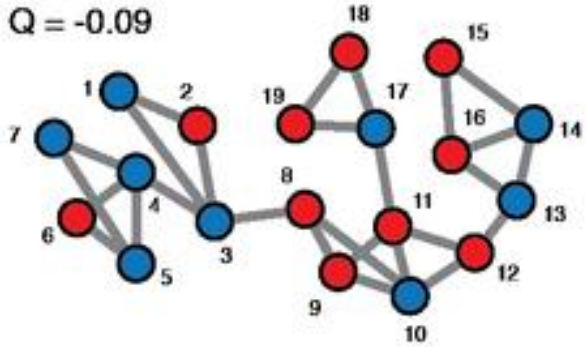
Network clustering with modularity score



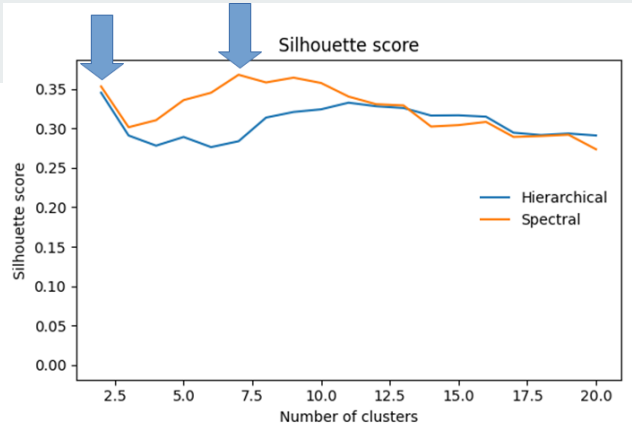
Guinney, J. et al. Nature Medicine 21:1350-1356 (2015)

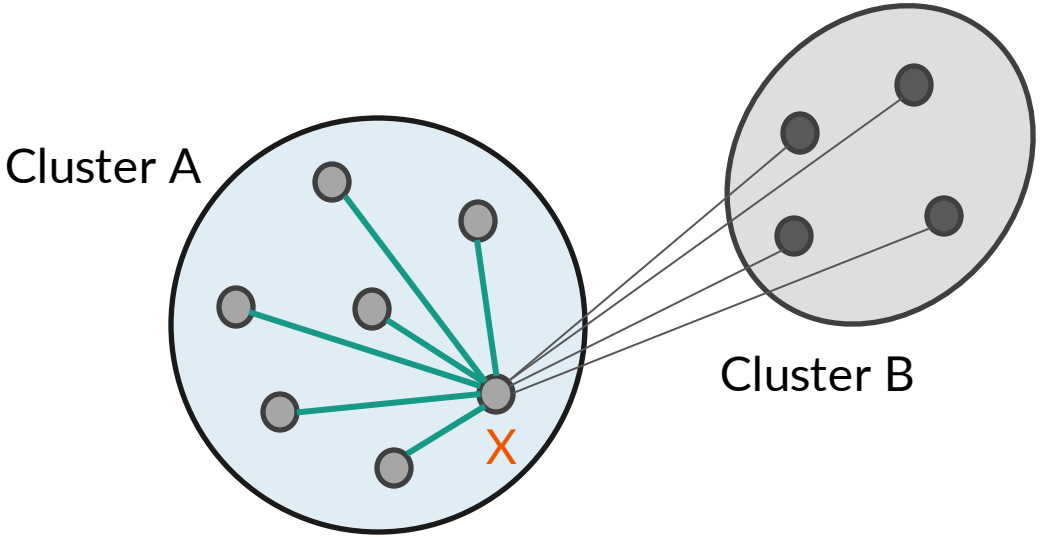


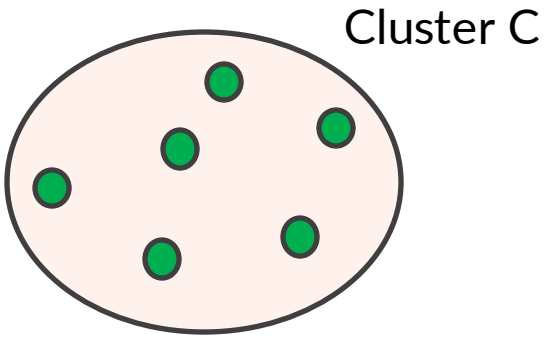
Betzel *et al*. (2015)



Cluster selection: Silhouette score







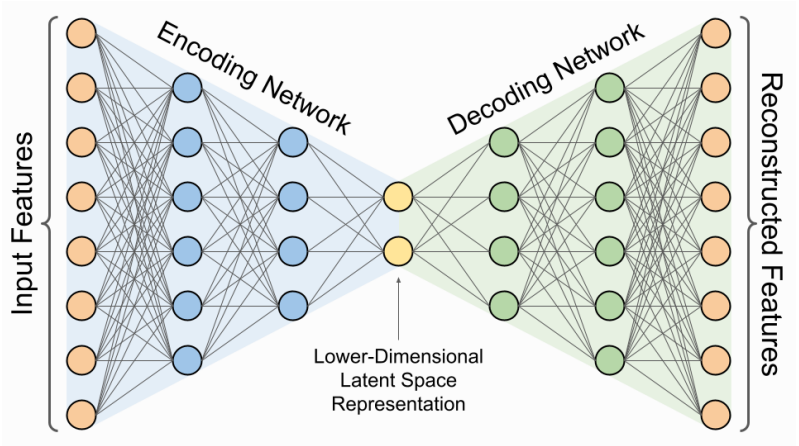
- Compare distances from X to other members of cluster A versus distances from X to members of cluster B (the closest cluster from A)

Dimensionality reduction

What is the dimension of a dataset?

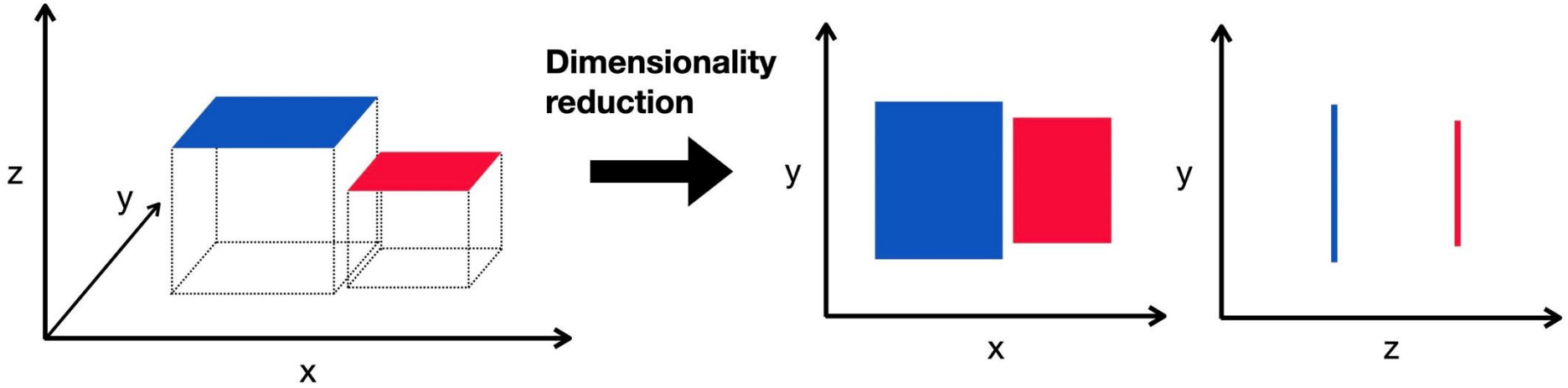
	Feature 1	Feature 2	Feature 3	Feature 4	…
Sample 1					
…					

- Number of features?
- Number of non-redundant features?
- **The minimal size of latent vectors from which the original data can be accurately reproduced**



https://www.assemblyai.com/blog/introduction-to-variational-autoencoders-using-keras/

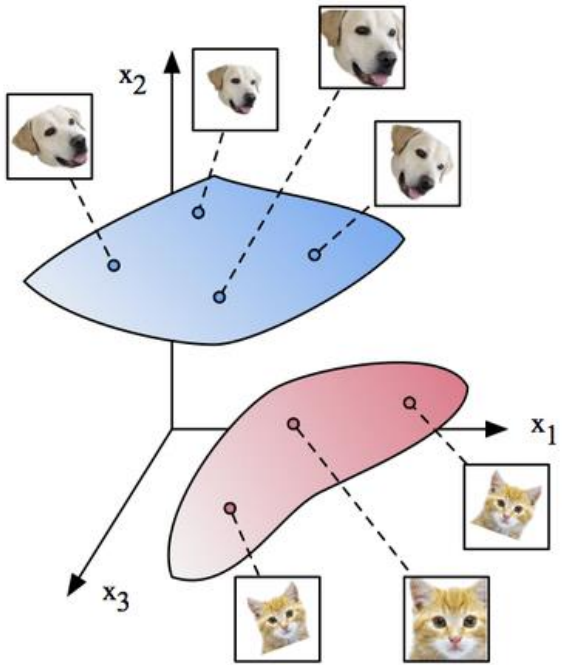
Dimensionality reduction



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

- Reduce the number of features while retaining information content

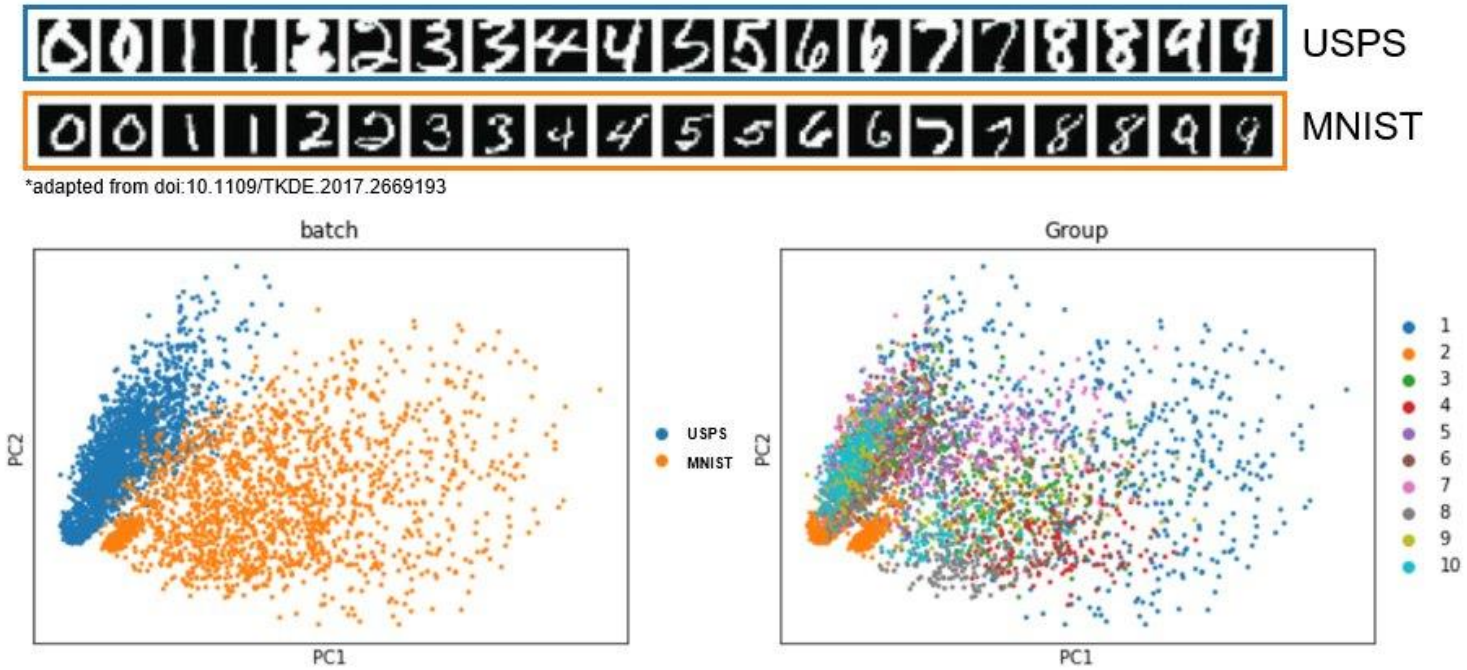
Manifold hypothesis



Chung, S. et al. “Classification and Geometry of General Perceptual Manifolds”

- Within a complex dataset with large number of features
- We can often find **low-dimensional manifolds**, each containing data points coming from the same class
- This lets us identify **a low-dimensional embedding** of the original features that can be used for unsupervised and supervised learning

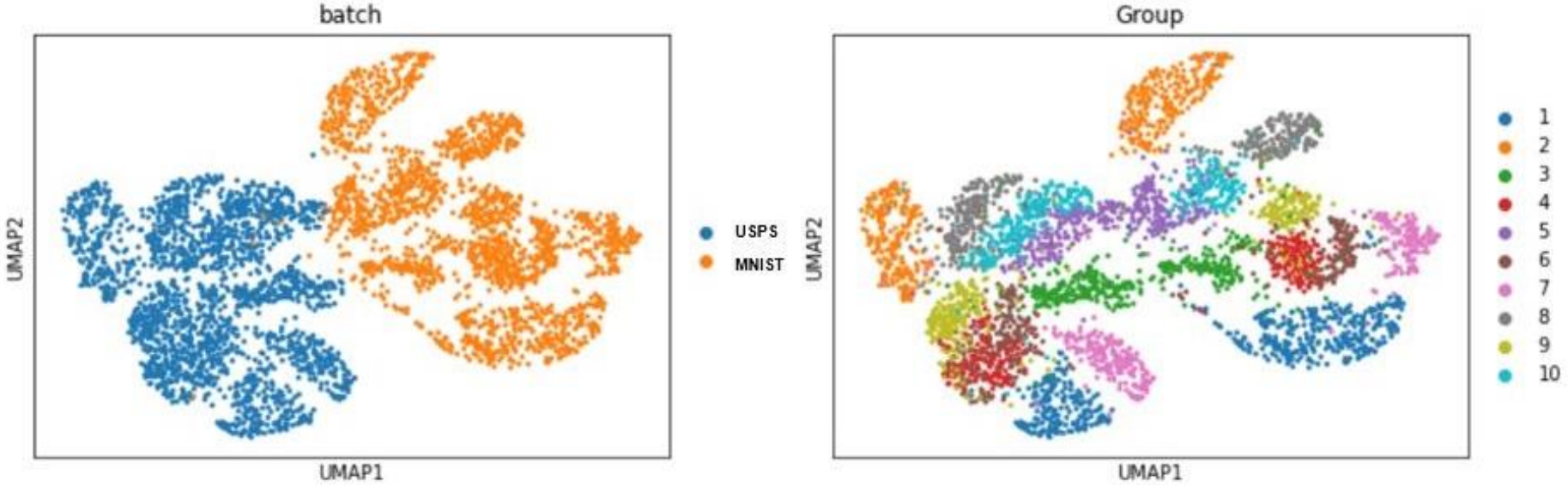
Example: Principal Component Analysis



*adapted from doi:10.1109/TKDE.2017.2669193

https://twitter.com/lkmklsmn/status/1436357177887895555

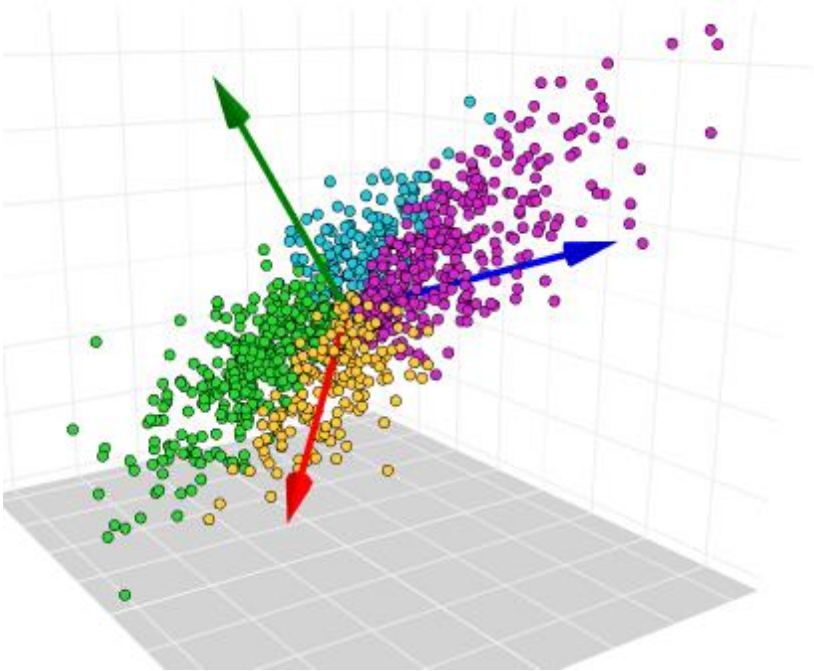
Example: Uniform Manifold Approximation and Projection



https://twitter.com/lkmklsmn/status/1436357177887895555

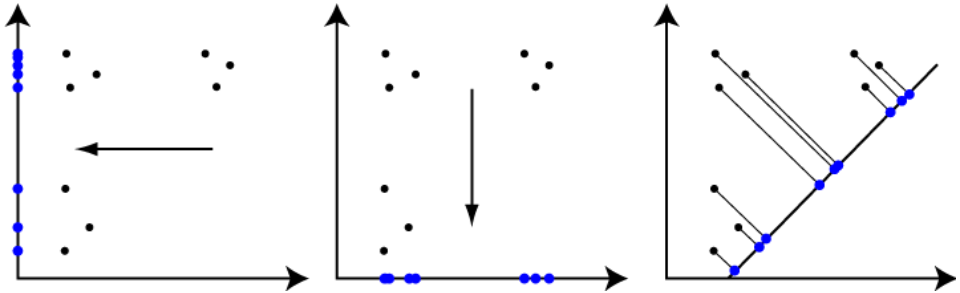
- UMAP embedding can distinguish both the data sources and digits

Variance is information



https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

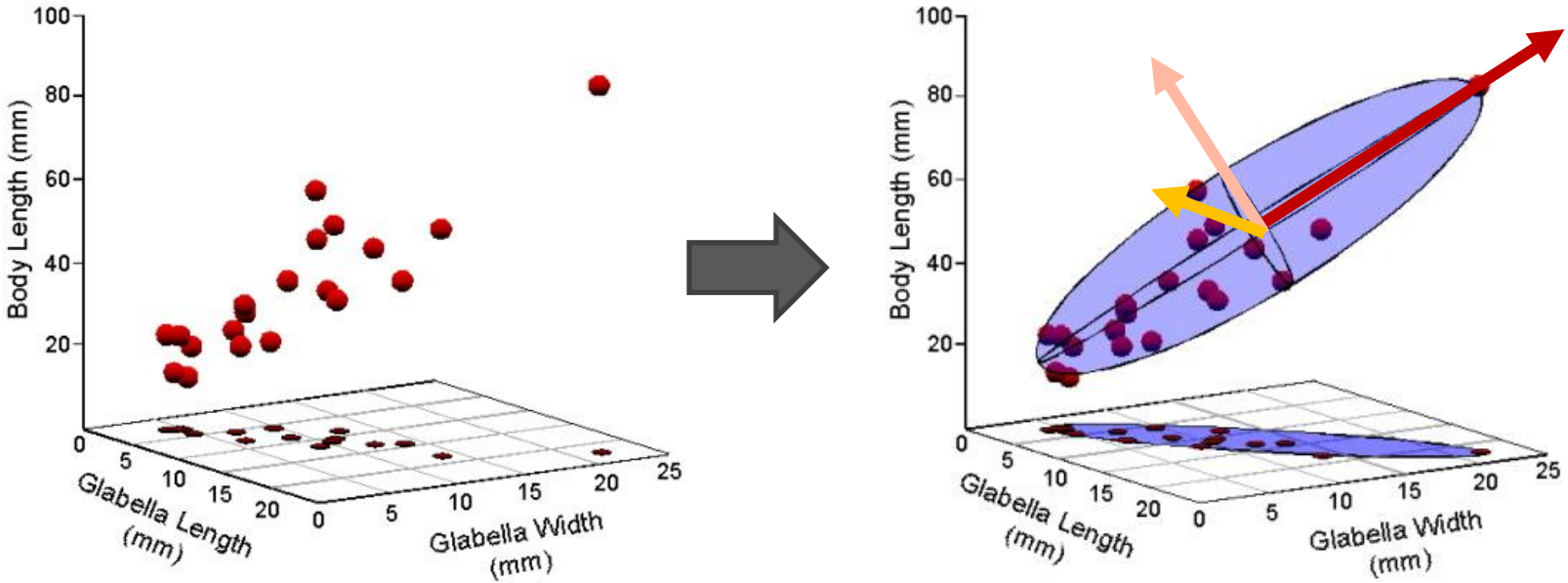
Projection



https://shapeofdata.wordpress.com/2013/04/16/visualization-and-projection/

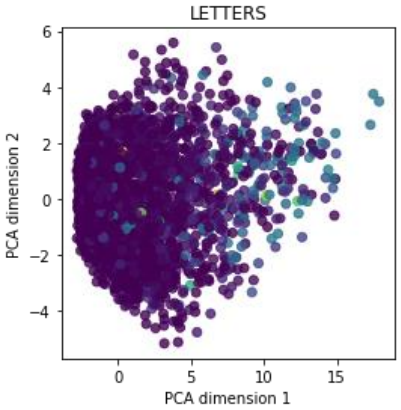
- High variances = more power to distinguish groups of data points

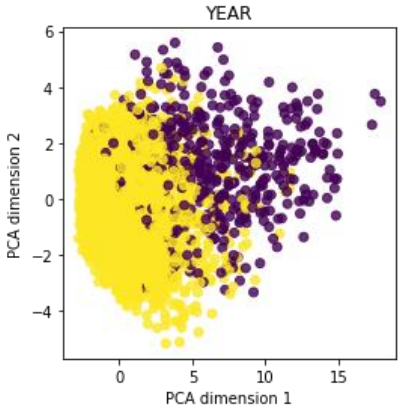
PCA prioritizes directions with high variances

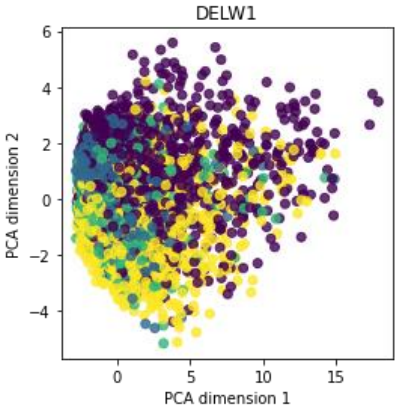


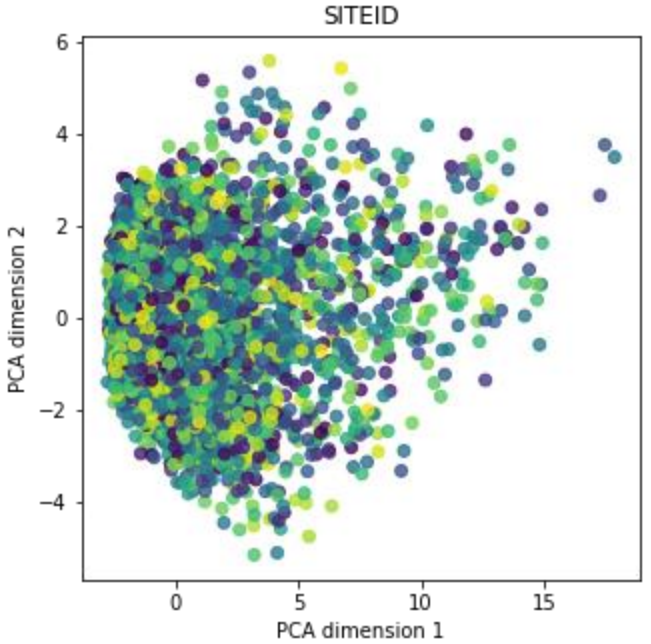
Source: the paleontological association

PCA of ADNI's MoCA test scores



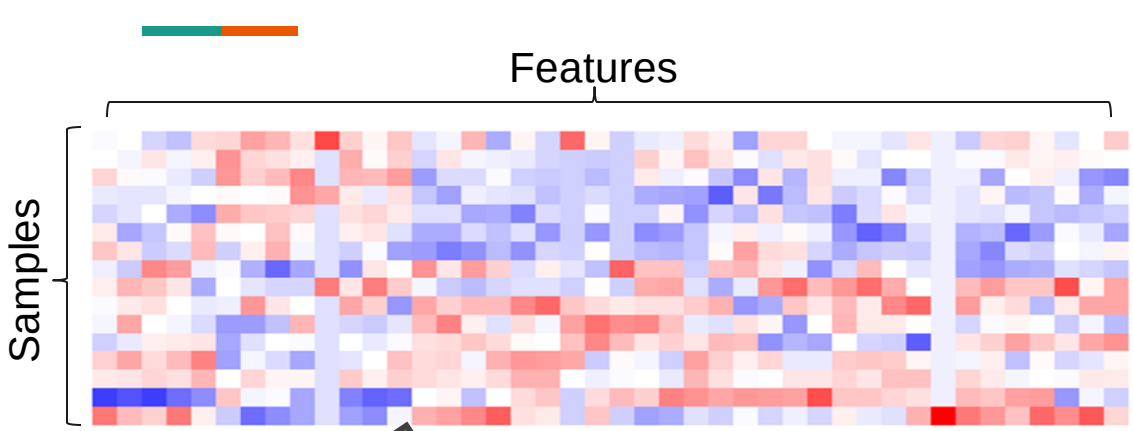


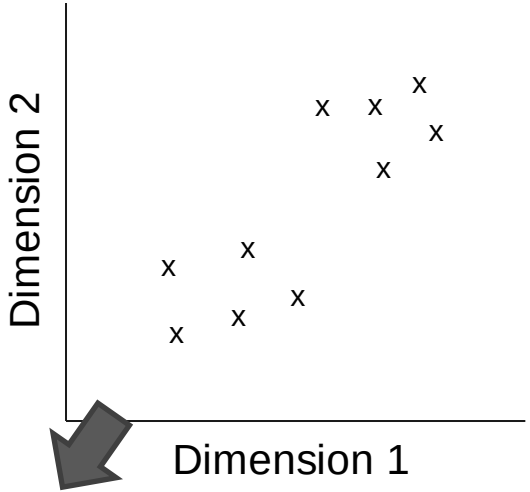




- LETTERS, YEAR, and DELW1 scores cluster in different direction
- SITEID exhibits no pattern

A general framework for dimensionality reduction





User-defined distance

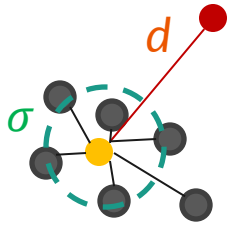
	x_1	x_2	x_3	x_4
x_1	0	3	3	3
x_2	3	0	1	1
x_3	3	1	0	1
x_4	3	1	1	0

Correlation

	x_1	x_2	x_3	x_4
x_1	0	9	9	9
x_2	9	0	1	1
x_3	9	1	0	1
x_4	9	1	1	0

Euclidean distance

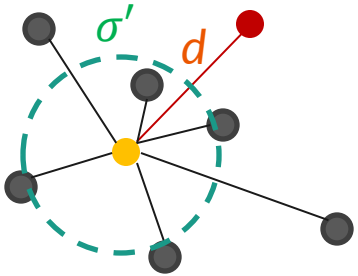
Probability of being a neighbor



score(o | o) = probability that o would pick o as neighbor under a **normal distribution** center at o

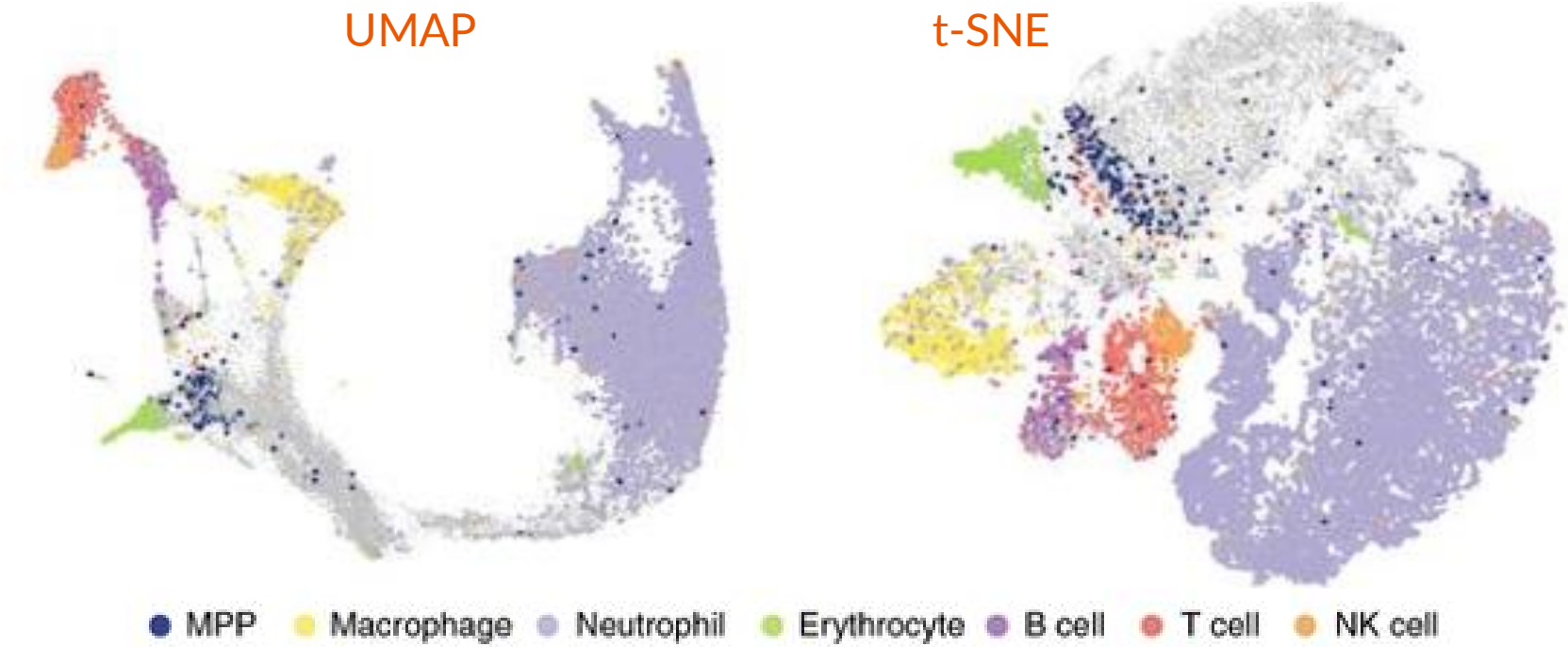
$$= \frac{e^{-\frac{d^2}{2\sigma^2}}/\sigma}{\sum e^{-\frac{(\text{dist}(o,o))^2}{2\sigma^2}}/\sigma}$$

o = other data points



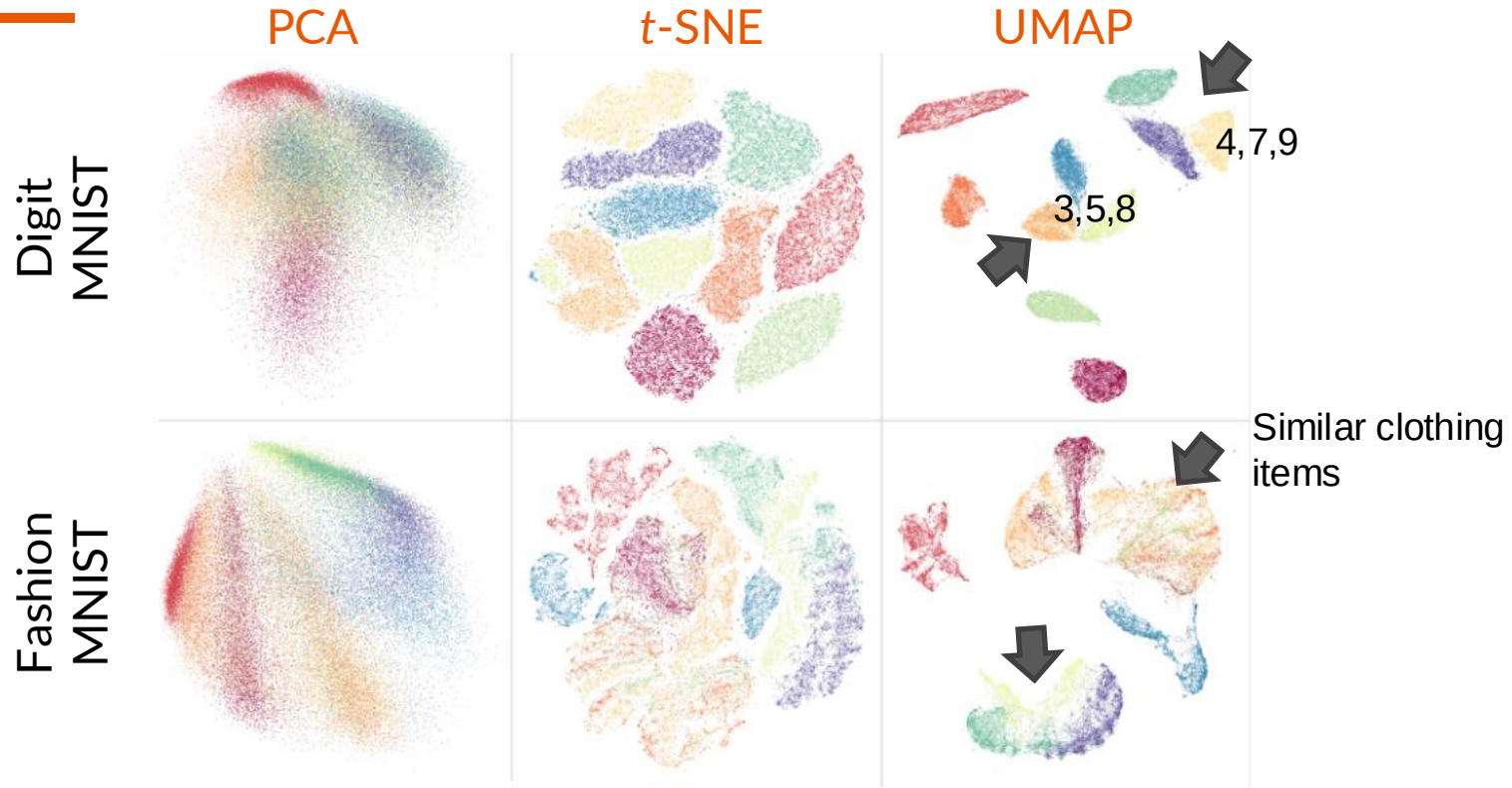
- Same distance d normalized against density σ and distances to other nearby data points o

t-SNE vs UMAP on single-cell gene expression data



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

UMAP is better at capturing inter-group relationships



McInnes, L., Healy, J. and Melville, J. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”

Any question?