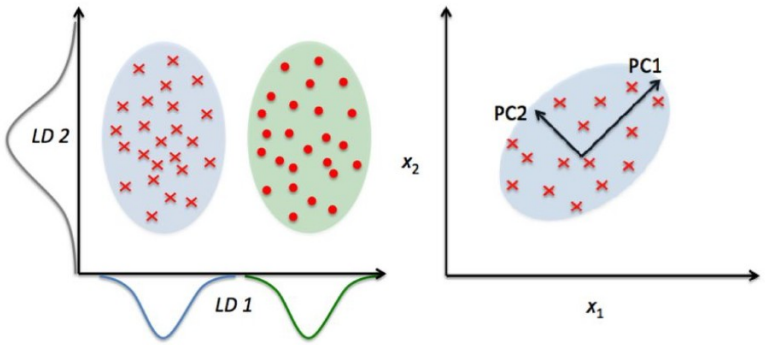
Tutorial 10

COMP90014 Algorithm for Bioinformatics

Semester 2, 2025

Why do we need dimensionality reduction?



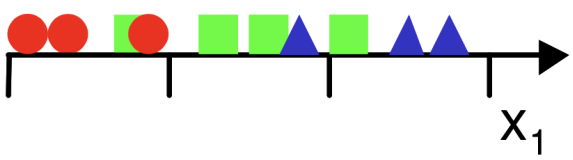
Feature selection

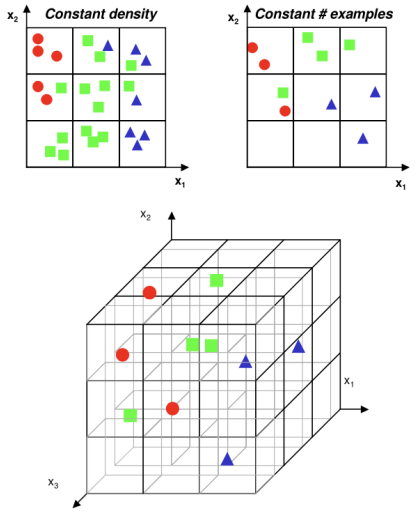
- greedy feature selection

Feature extraction

- unsupervised methods
- supervised methods
- principal component analysis (PCA)
- singular value decomposition (SVD)

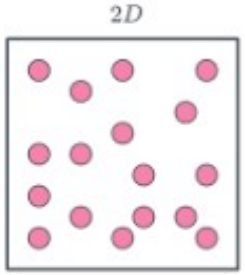
The curse of dimensionality

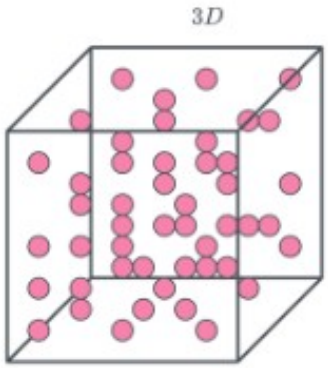


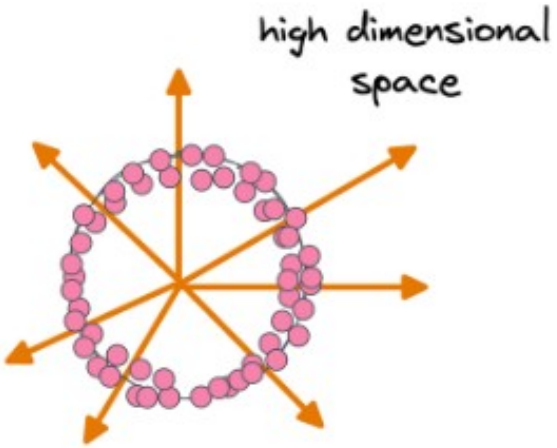


Ricardo Gutierrez-Osuna, Wright State University

- as the number of dimensions increases the data become ***sparse***
- an huge amount of data is needed to "cover" all the dimensions
  - number of data points needed **grows exponentially with the number of dimensions**

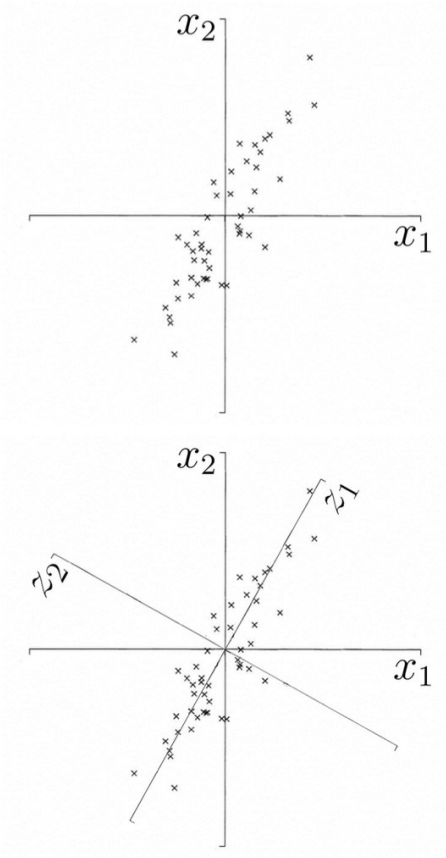






https://www.dailydoseofds.com/a-mathematical-deep-dive-into-the-curse-of-dimensionality/

PCA



Mario Guarracino, ICAR-CNR

Mathematical procedure:

- transform a number of (possibly correlated) variables into a (smaller) number of uncorrelated variables
- the uncorrelated variables are called ***principal components***

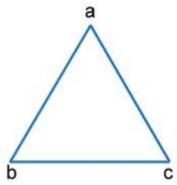
Principal components (PC):

- first PC is the projection direction that maximizes the variance of the projected data
- second PC is the projection direction that is orthogonal to the first PC and maximizes variance of the projected data

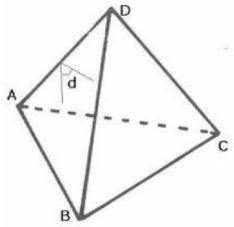
- Spatial rearrangements may reveal relationships that were hidden in higher dimension space

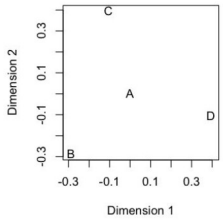
MDS

$$D(x_i, x_j) = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

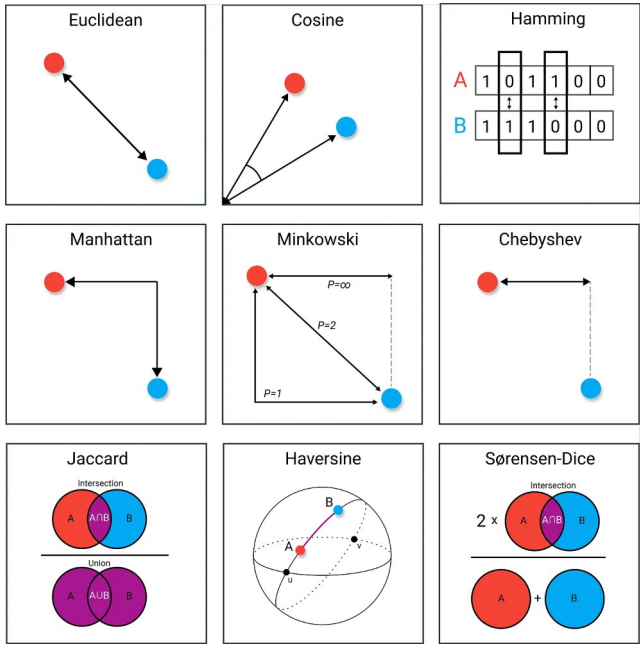


$$D(x_i, x_j) = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

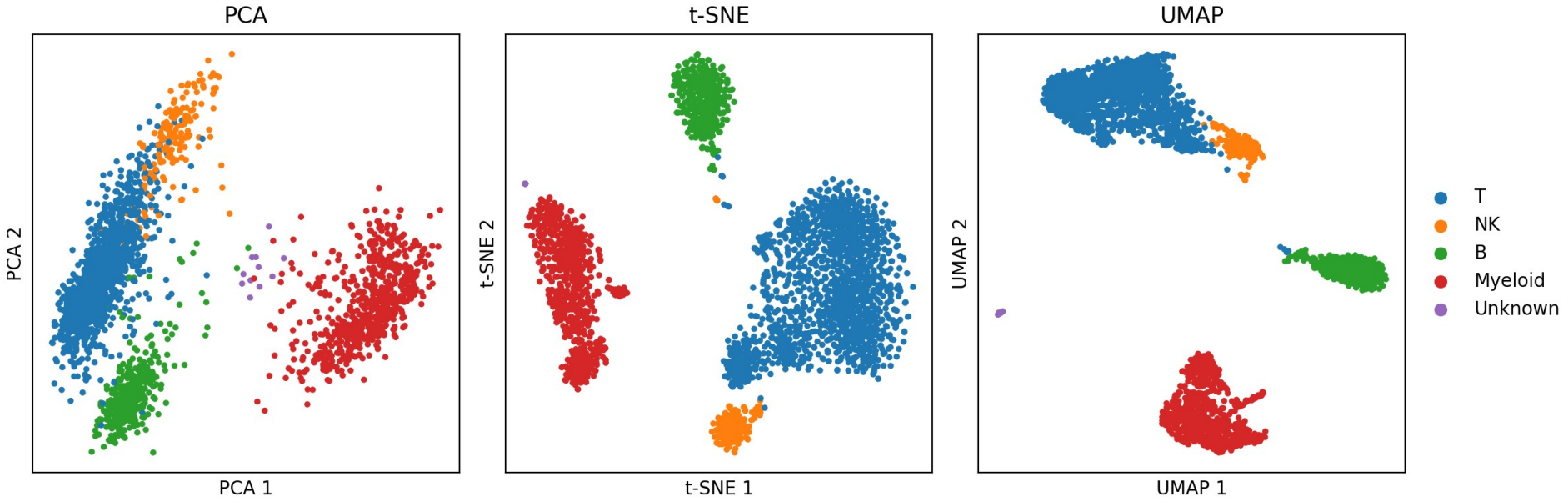




- start with a pairwise distance matrix or dissimilarity matrix
- we can represent three points that are equally-spaced in 3D **exactly in 2D**
- we can represent four points that are equally-spaced in 3D **exactly in 3D** ...
- **... but not in 2D**
- in general, we need $N - 1$ dimensions to exactly represent pairwise distances between N samples



T-SNE vs UMAP



https://pair-code.github.io/understanding-umap/

https://distill.pub/2016/misread-tsne/