Unsupervised learning

LSSDS 2023

Mauricio Cerda, Eng. PhD

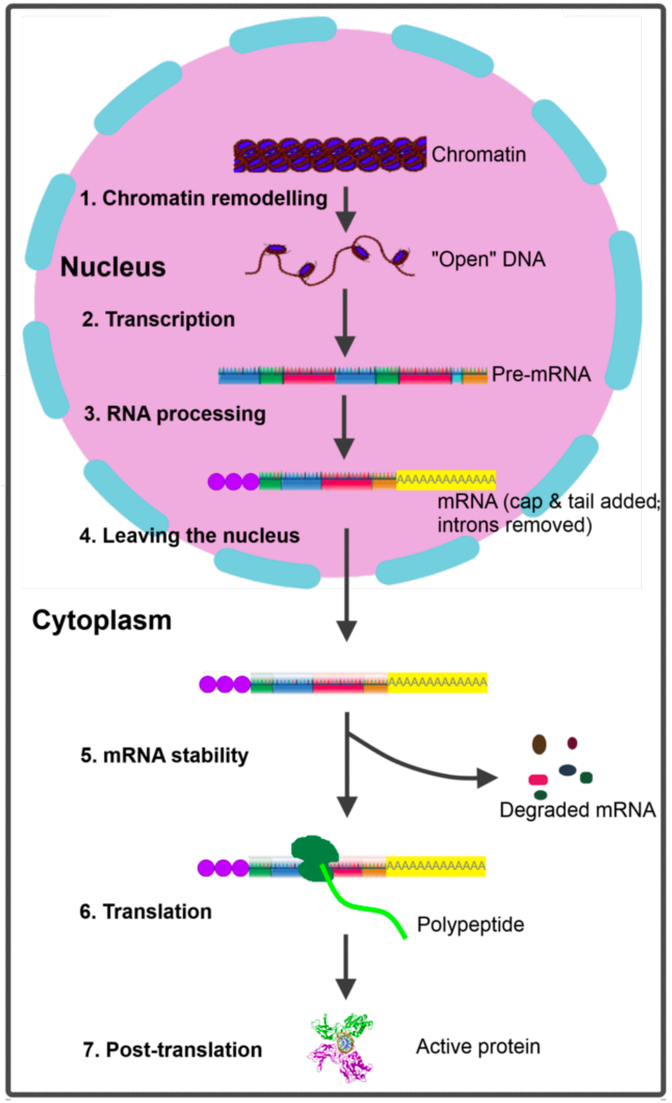
Programa de Biología Integrativa
I.C.B.M., Facultad de Medicina, Universidad de Chile

Outline

- Examples: genes expression, text topics.
- Problem definitions.
- Dimensionality reduction (PCA, t-SNE, uMAP).
- Clustering (k-means, hierarchical, DBScan).

Example: genes

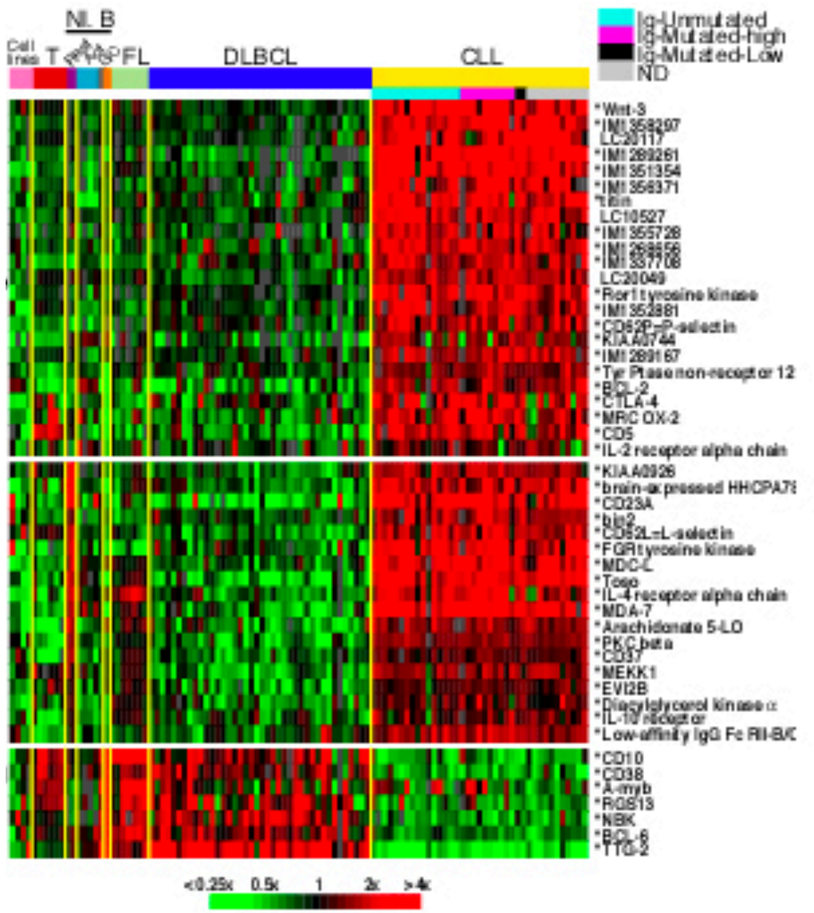
- Gens change their expression according to sample.
- This is called a gen expression profile.



Genes expression, Wikipedia.

Example: genes

- Microarray allows to observe genes expression:
  - Groups of genes can have coordinated changes: co-expressed, co-regulated.
- Group samples may have similar patterns:
  - Groups in the samples
  - Useful to detect outlayers in unexpected groups.

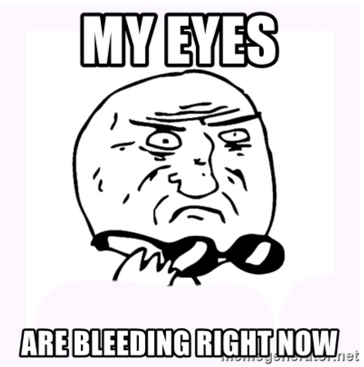


Alizadeh et al 2000 Nature

Example: text topics

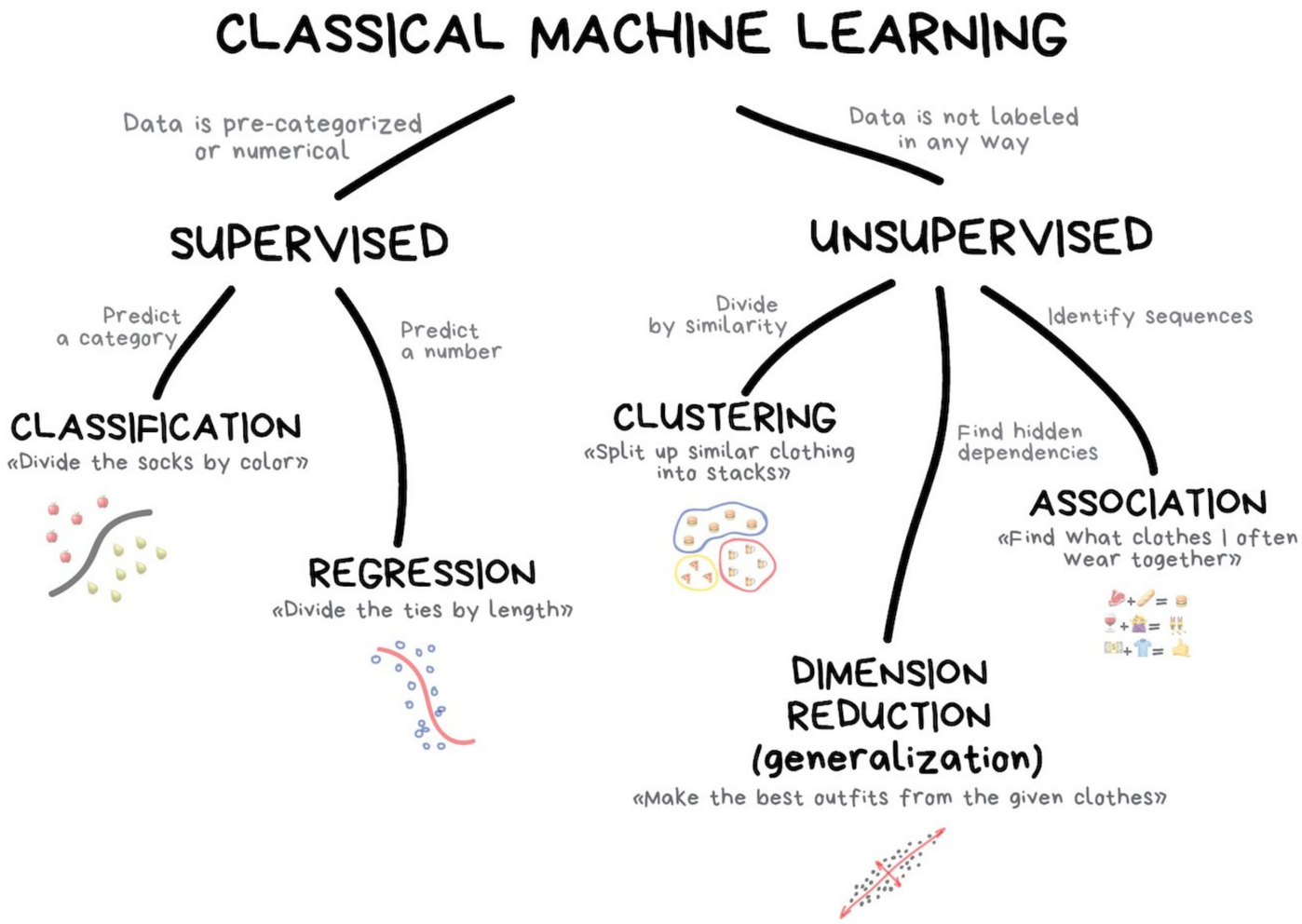
- From natural text we may need to get the “topic”, without extra information.
- Topics may be very specific or broad. And not even in proper English/ Spanish!

”el desarrollo de una propuesta cultural privada para a cojer la mas ampria bariedade de expresiones que conformas las artes de mi comunidad de mejillones es mi motibante ya que oficia es dar aconocer y el dar aconocer como se hace una empa nada de marisco a la mejillonina o hacer un cuadro deo leo una acuarela ect hacer



Tasks

- If we know the pattern (class, label), we call it a supervised problem.
- If we don't know the pattern, we call it an **unsupervised problem**.

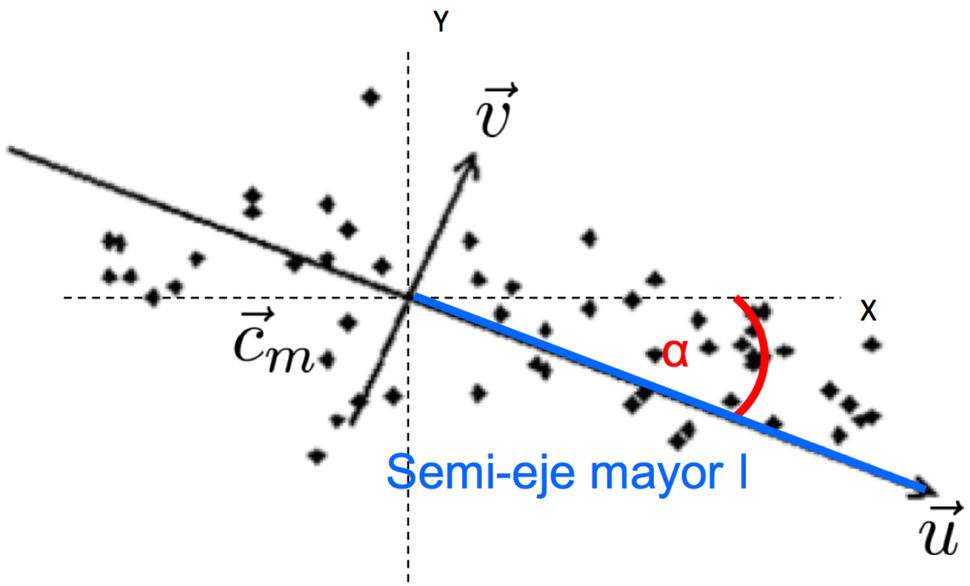


Preliminaries

- Set your environment (git, server, aws, docker)
- Load data.
- Study variables.
- Visualize.
- Understand the problem.
- Choose the right tool for your problem!

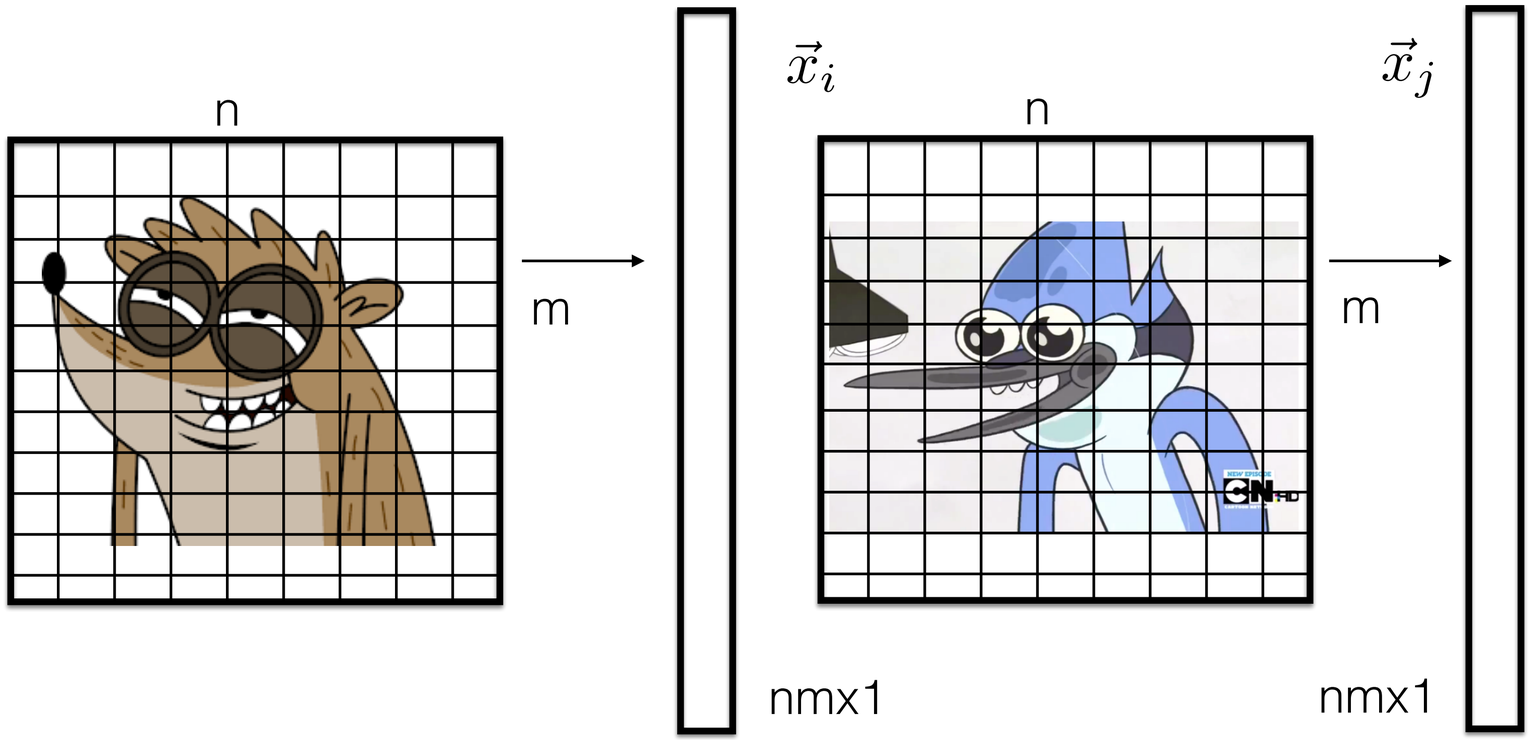
Dimensionality reduction

- If we have many variables (N), can we reduce the problem into less variables (n)?
- To visualize, to understand what variables are more relevant.
- Look for the direction with more variance.



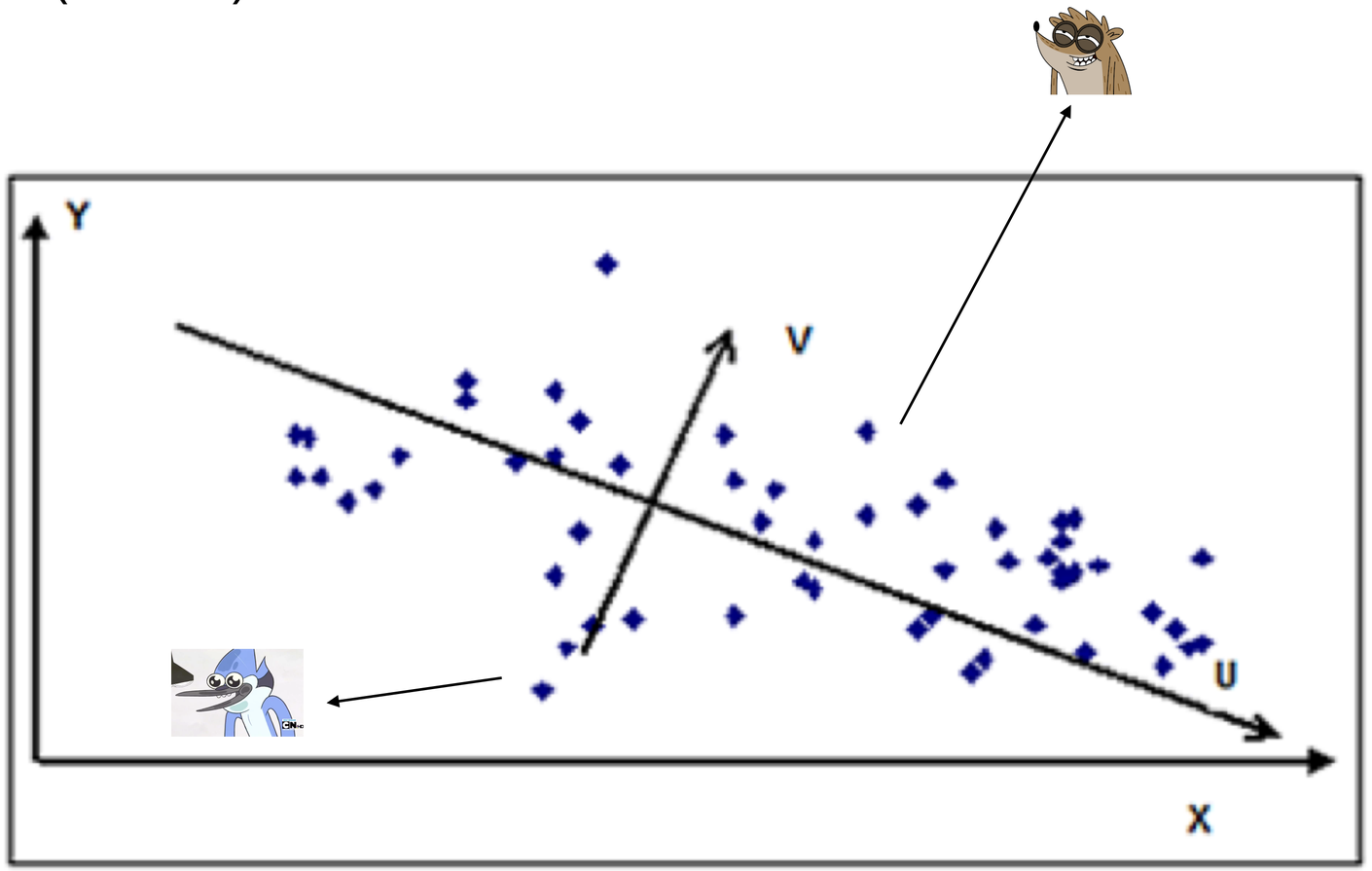
Dimensionality reduction: high dimension

- For images, we can build a point per image.



Dimensionality reduction: PCA

- To maximize variance is to make covariance matrix diagonal (PCA).



Dimensionality reduction: PCA formulation

$$\mathbf{X} = [\vec{x}_1 ... \vec{x}_i ... \vec{x}_k]$$

$$\mu_i = E(\vec{x}_{:,i})$$

$$\Sigma_{ij} = cov(\vec{x}_{:,i}, \vec{x}_{:,j}) = E[(\vec{x}_{:,i} - \mu_i)(\vec{x}_{:,j} - \mu_j)]$$

$$\mathbf{\Sigma} = \begin{bmatrix} E[(\vec{x}_{:,1} - \mu_1)(\vec{x}_{:,1} - \mu_1)] & E[(\vec{x}_{:,1} - \mu_1)(\vec{x}_{:,2} - \mu_2)] & \cdots & E[(\vec{x}_{:,1} - \mu_1)(\vec{x}_{:,k} - \mu_k)] \\ \vdots & \ddots & \vdots & \\ E[(\vec{x}_{:,k} - \mu_k)(\vec{x}_{:,1} - \mu_1)] & E[(\vec{x}_{:,k} - \mu_k)(\vec{x}_{:,2} - \mu_2)] & \cdots & E[(\vec{x}_{:,k} - \mu_k)(\vec{x}_{:,k} - \mu_k)] \end{bmatrix}$$

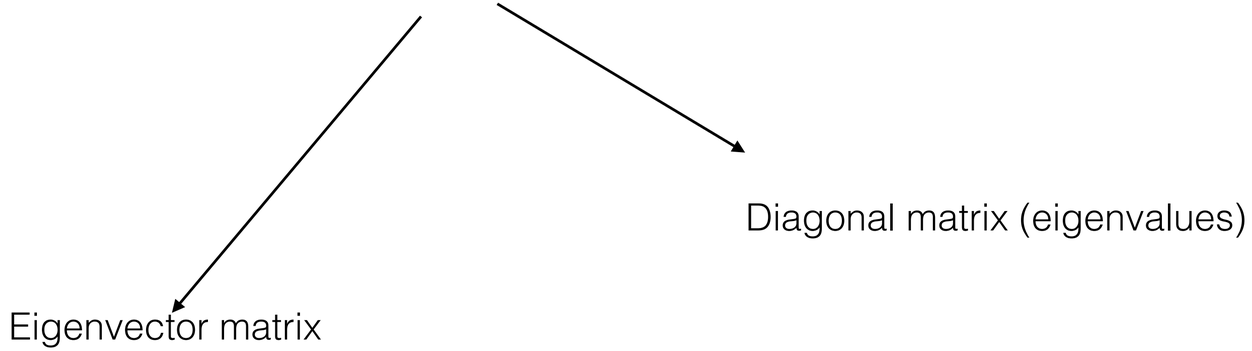
Dimensionality reduction: PCA formulation

- If $\mu_i = 0$ (centered data)

$$\mathbf{\Sigma} = \frac{1}{k}\mathbf{X}^T\mathbf{X}$$

- We can solve if we diagonalize, e.g. SVD

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathbf{T}$$



Dimensionality reduction: PCA as descriptor

- Example, eigenfaces.

Let’s take k different pictures of the same subject (x_k).



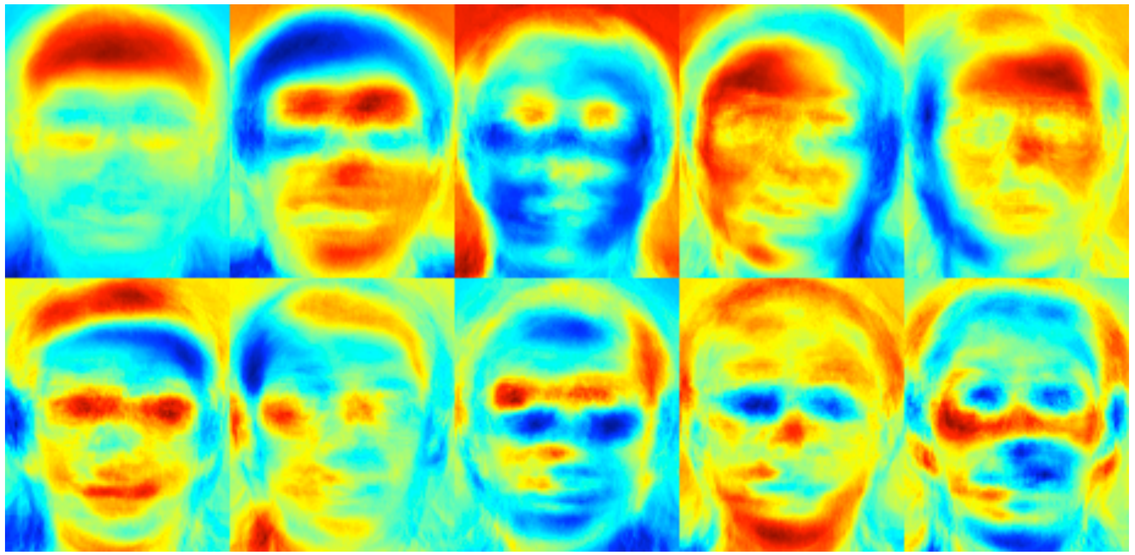
…


1 2 k

Dimensionality reduction: PCA as descriptor

- Example, *eigenfaces*.

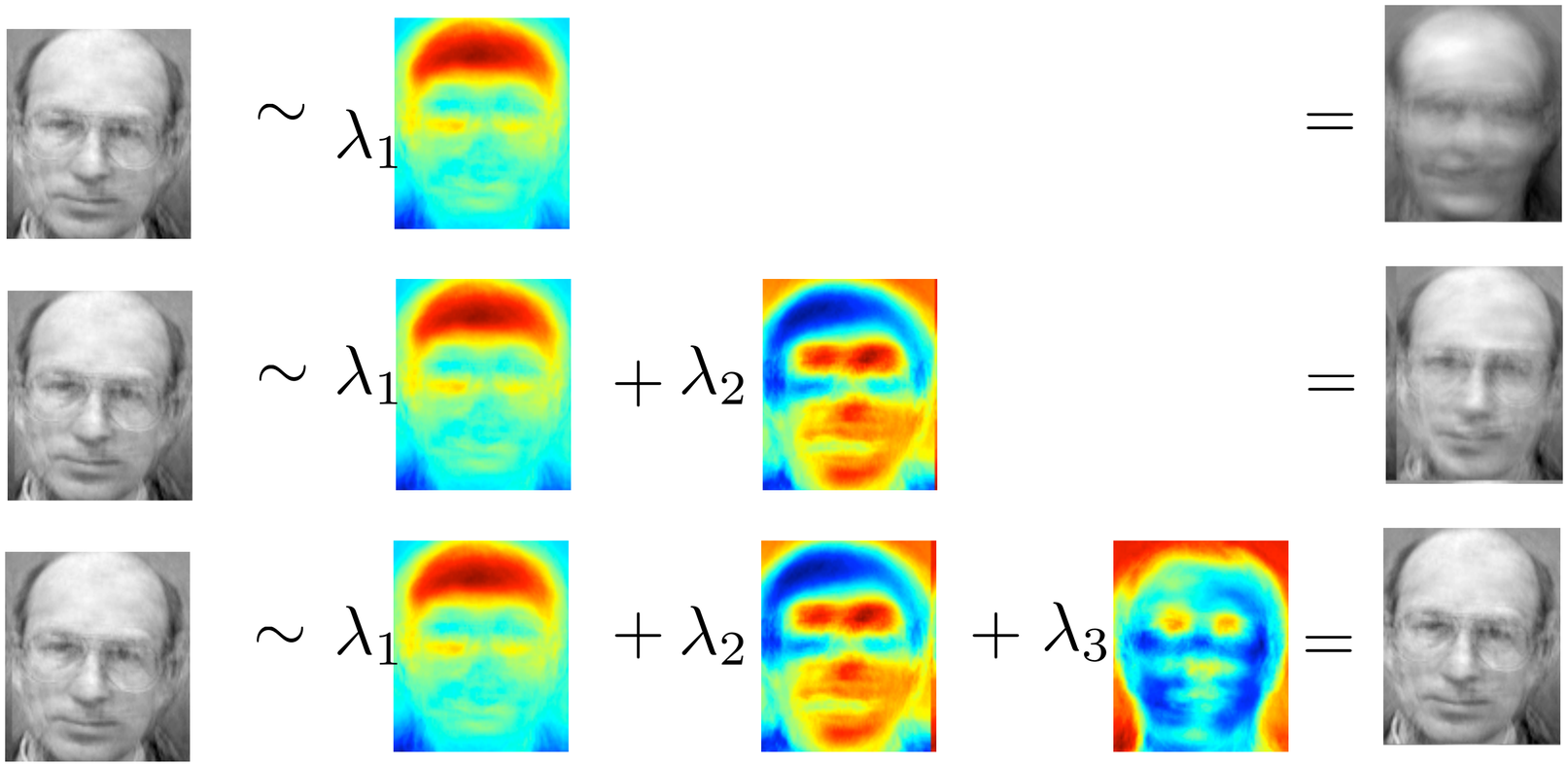
  How the first 10 eigenvectors look like?



Dimensionality reduction: PCA as descriptor

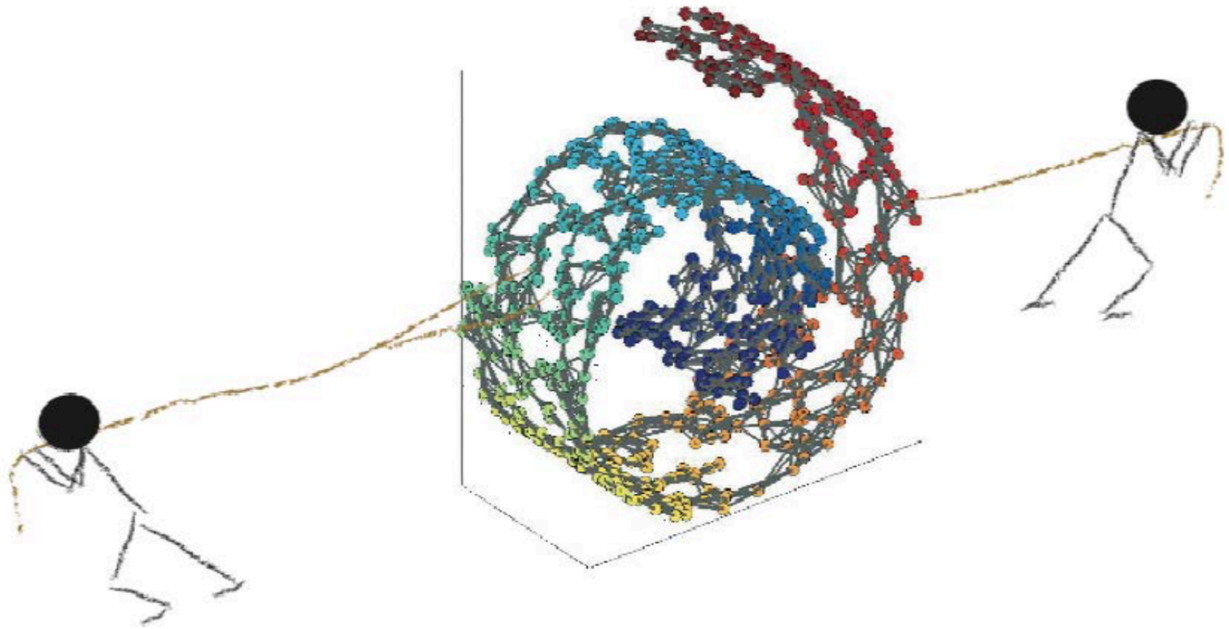
- Example, eigenfaces.

We can simplify the problem, or get features.



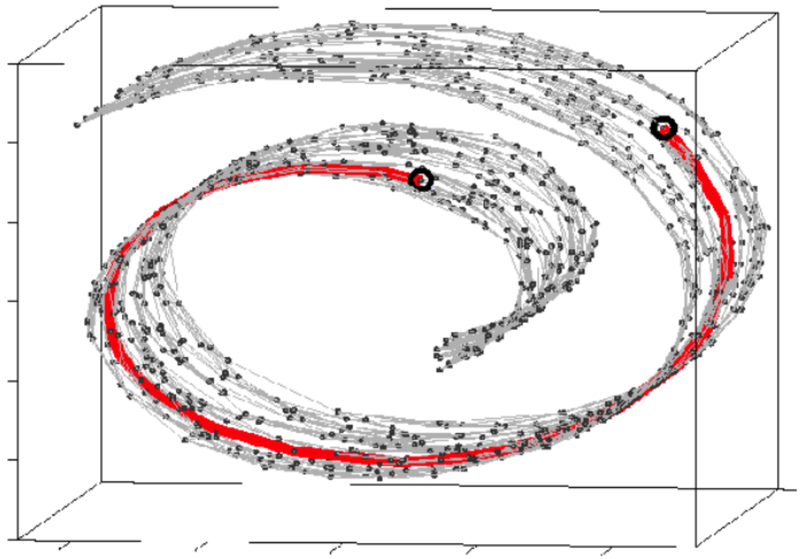
Dimensionality reduction: linear methods

- PCA is a linear transformation (to start).
- Complex data distribution cannot be simplified by PCA.



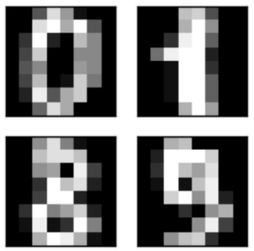
Dimensionality reduction: non-linear methods

- PCA use euclidean distance, but it is geodesic distance what is relevant.
- Multiple algorithms approximate “local” (ϵ) geodesic distance (ISOMap).

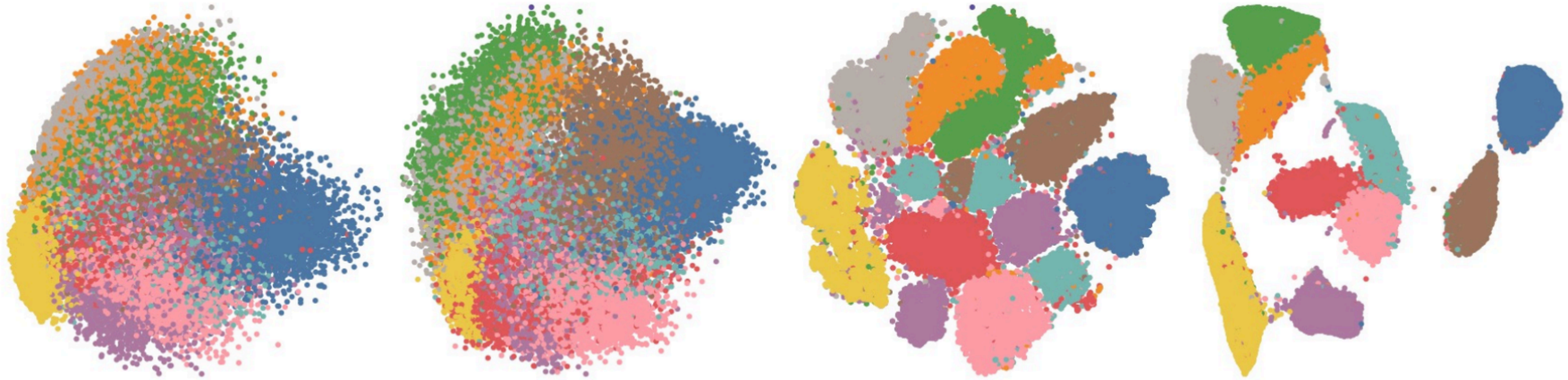


Dimensionality reduction: other methods

- Dimensionality reduction techniques (PCA, ICA, ISOMap, MDS, t-SNE) will help to explore data, build features, select variables.
- Other non-linear methods look for the inner data structure, such as uMAP.



MNIST



PCA *ISOMAP* *t-SNE* *uMAP*

https://arxiv.org/pdf/2205.00420.pdf

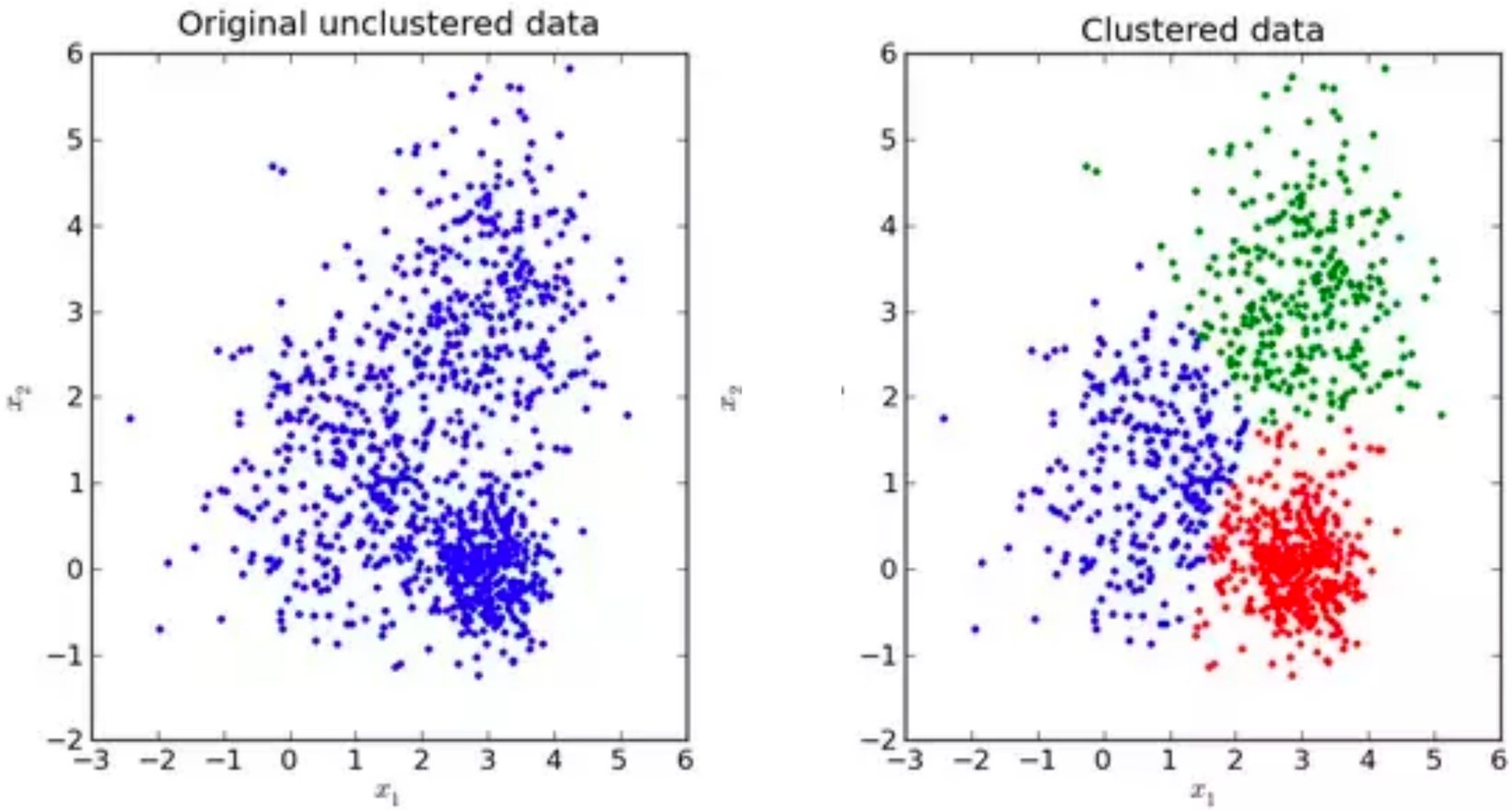
Clustering

- If we have…
  - A dataset
  - A mesure of similarity (metric)
- We want a **partition** where…
  - Same group -> more similar
  - Different group -> more different
  - That make “sense”, “interesting”

Clustering: why

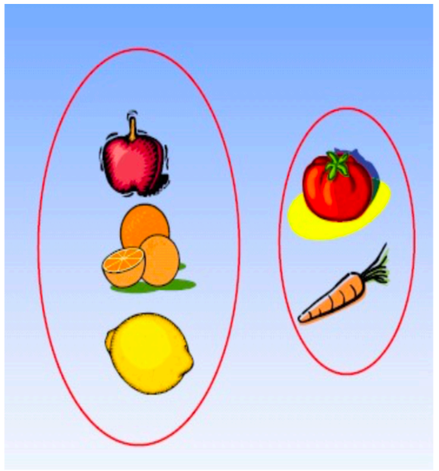
- To discover patterns.
  - To find a “natural” groups, where no classes are known.
  - To build hierarchies of similarities.
- To summarice information.
  - To define “templates”, representatives of a larger example set.

Clustering: not easy!

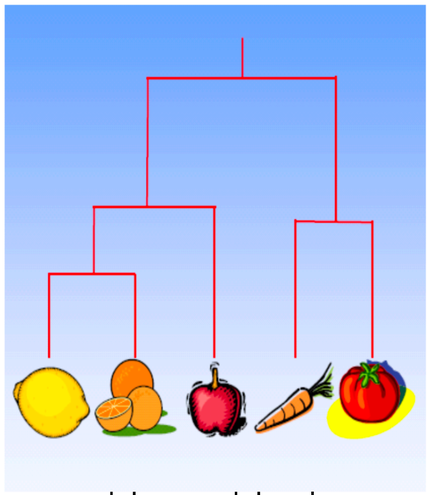


Clustering: divisive, hierarchical

- Divisive:
  - Flat clustering, space partitions.
  - Meaningful partition with a fixed set of parts.
- Hierarchical:
  - Set of partitions, where we can extract parts.



divisive

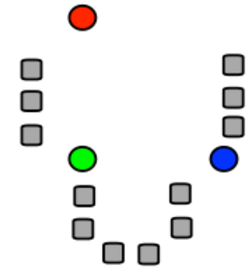


hierarchical

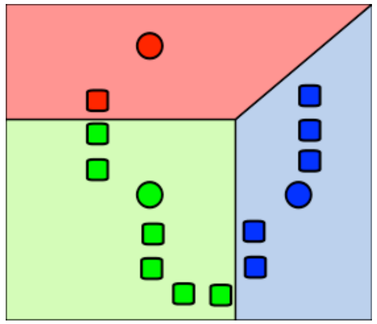
[divisive clustering 1954: JSTOR]

Clustering: divisive (k-means)

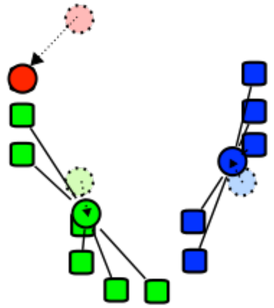
- We start by guessing we have k groups.
- Wee will look for the best k groups.
- Example k=3 (***k-means***)



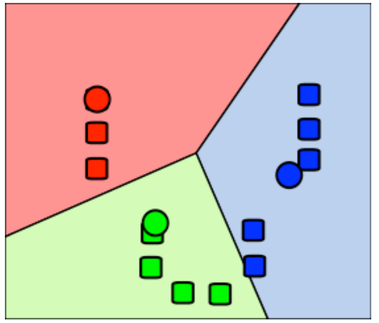
Random centroids



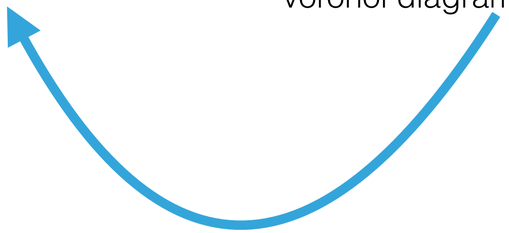
clusters assignation +
voronoi diagram



centroids re-computation

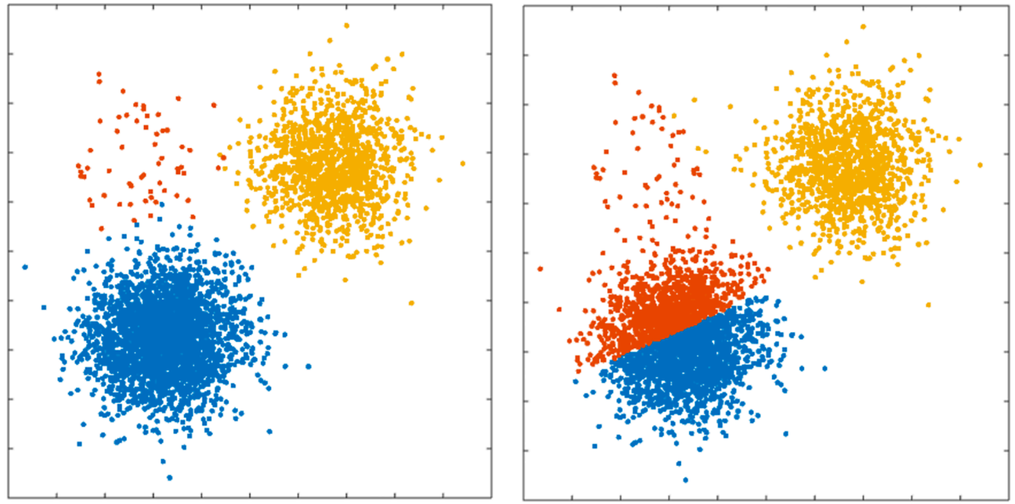


cluster assignation +
voronoi diagram



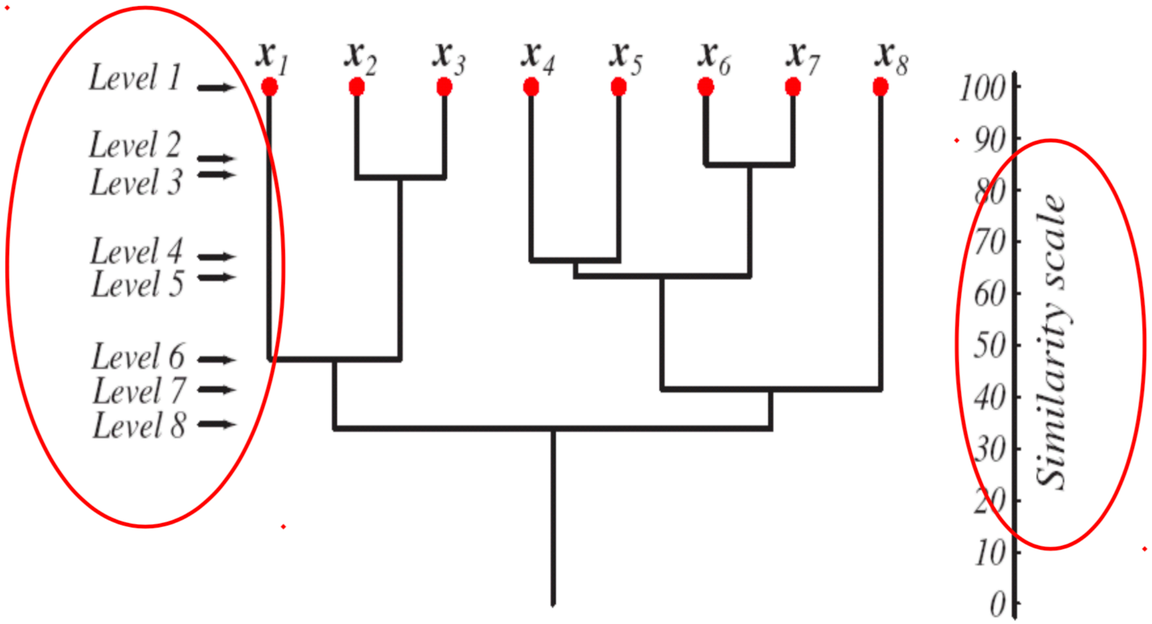
Clustering: divisive (k-means)

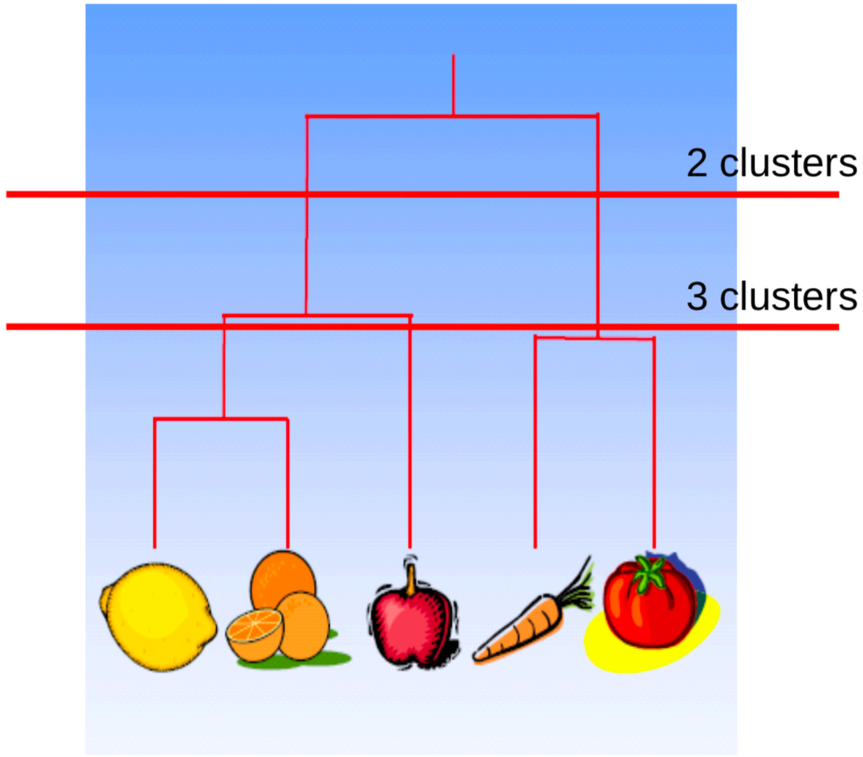
- Advantages:
  - Easy to interpret.
  - Useful to find patterns (groups).
- Disadvantages:
  - Exploratory (statistics?).
  - Variability (seed)
  - It will always give you clusters (!).



Clustering: hierarchical

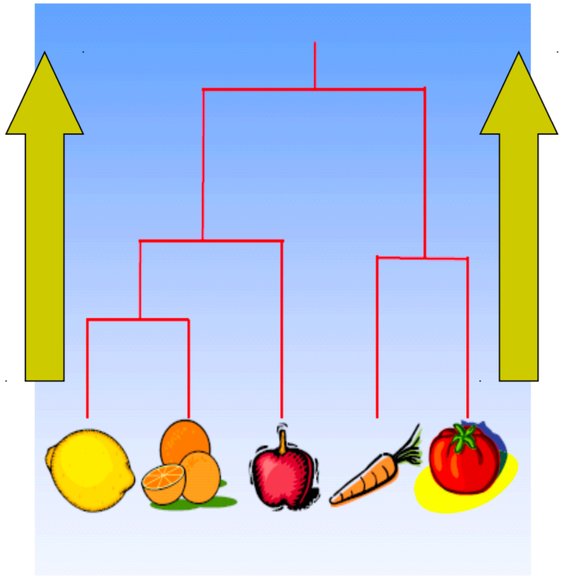
- To build a nested set of partitions.
- At each level, grouped data are more similar than to the rest.

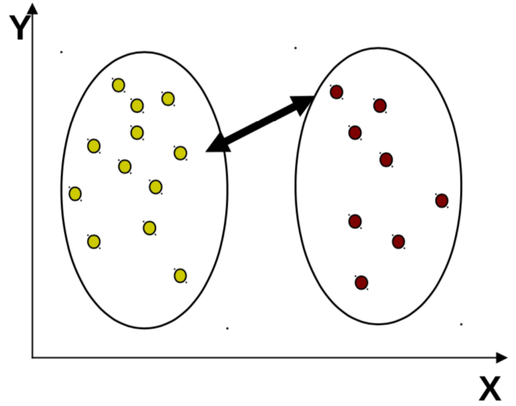




Clustering: hierarchical (bottom up)

- Starts from data points.
- Merge more similar clusters at each step.
- Ends when all points are in one cluster.
- Multiple similarity measures:
  - Single, complete, mean linkage.
  - Ward (minimize cluster inner distance)





Clustering: microarray

- En una matriz de expresión de genes
  - Filas -> genes
  - Columnas -> medidas de diferentes condiciones
- Resumir los datos
  - El valor en cada posición en la matriz representa el nivel de expresión (absoluto).