Going Multivariate

Pietro Franceschi
Computational Biology - Research and Innovation Centre - Fondazione E. Mach

pietro.franceschi@fmach.it

Multivariate in -omics

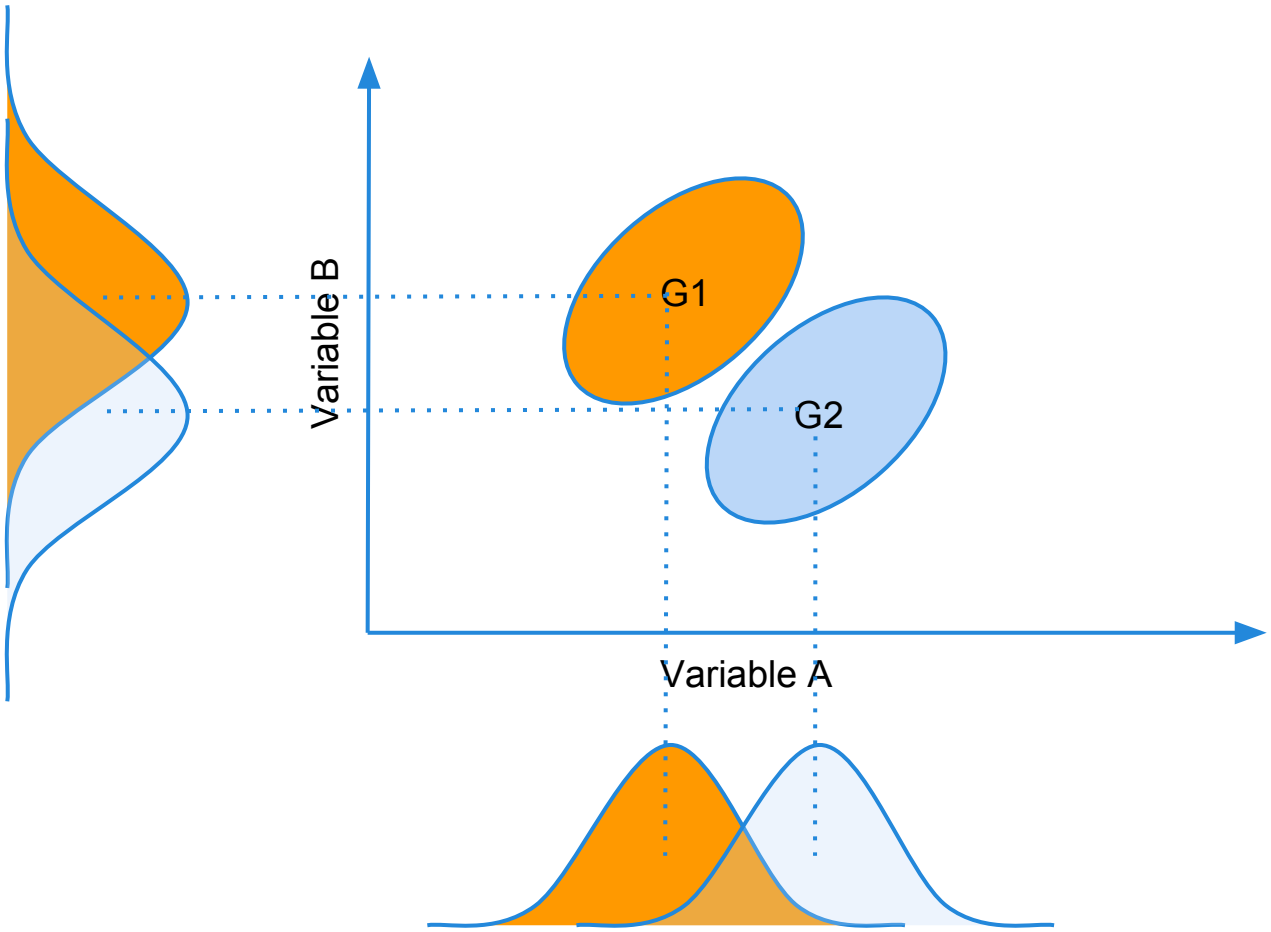
+

- Profit of variable correlation
- More fancy ;-)
- Comprehensive

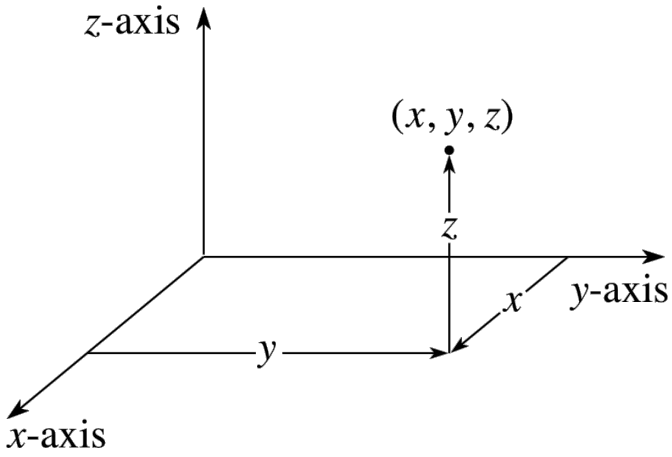
-

- Empty space
- “Random” correlations
- Often small sample size

… this is not new …



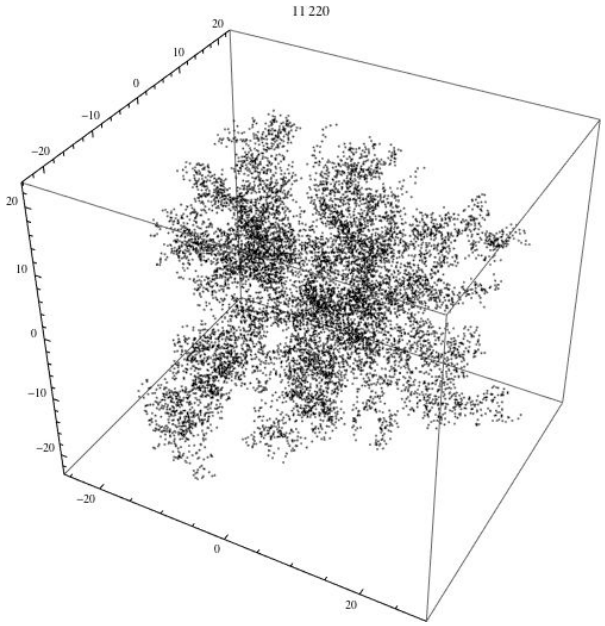
… the geometric picture



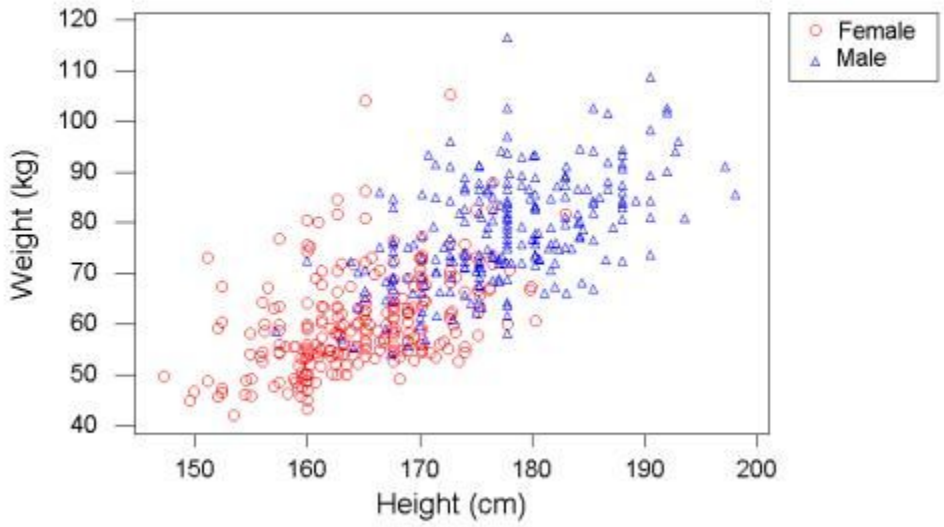
- I measure n variables per sample
- Each variable is one dimension
- Each sample is a point in the n dimensional space

… the geometric picture

- Each experiment is a cloud of points in the n dimensional space
- Due to the “relations” among the variables in the samples the cloud actually occupies a subspace of dimension smaller than n



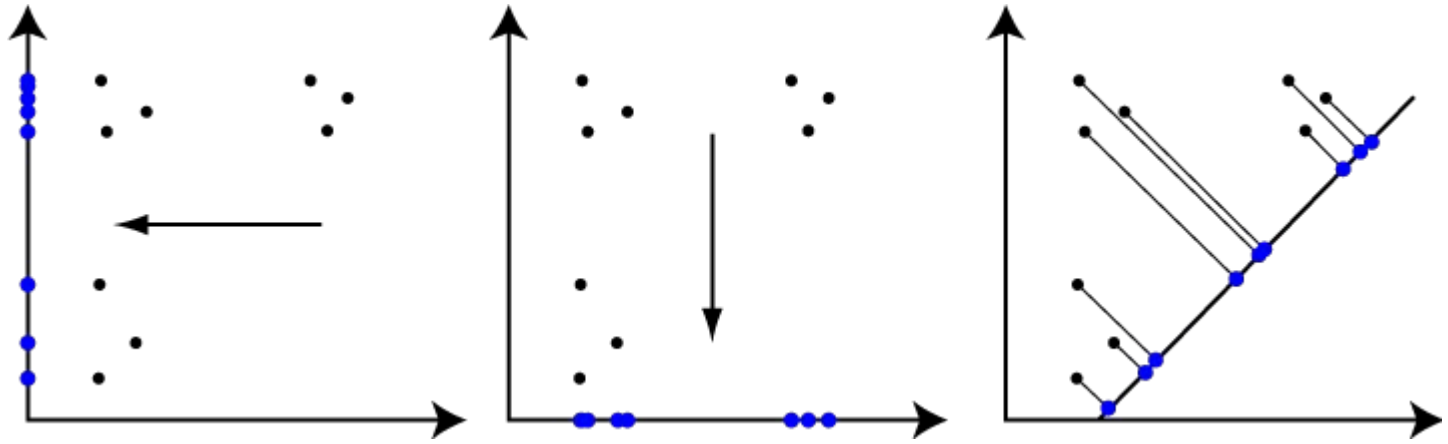
… the geometric picture



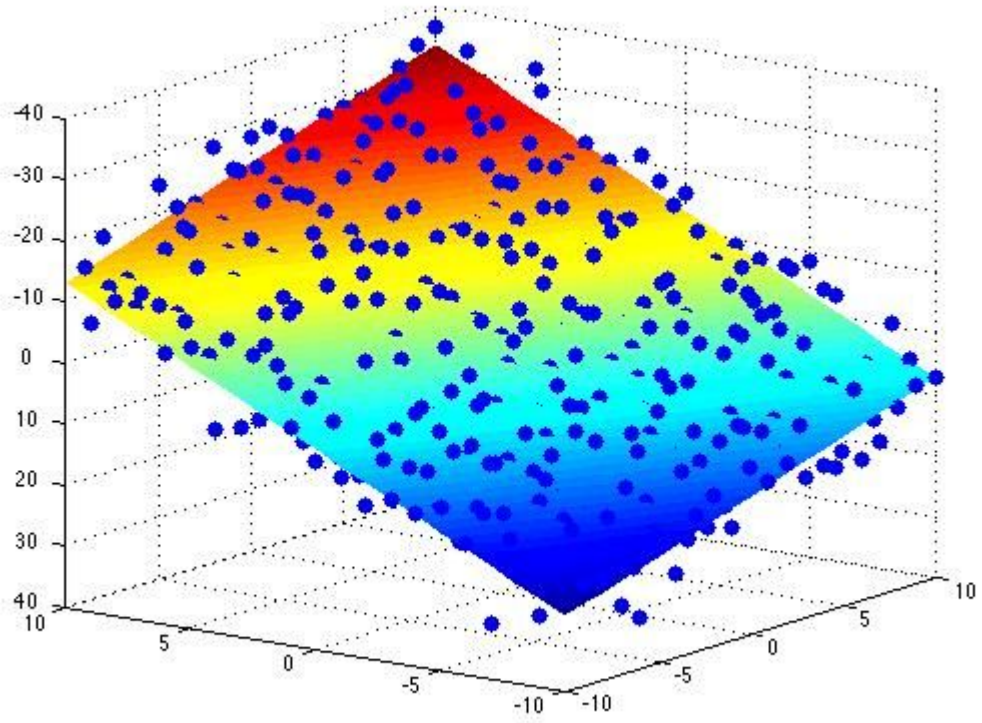
If you forget the noise, you see that this 2D plot is actually 1D … because setting one variable “blocks” the other one

Projections and visualization ...

… projections

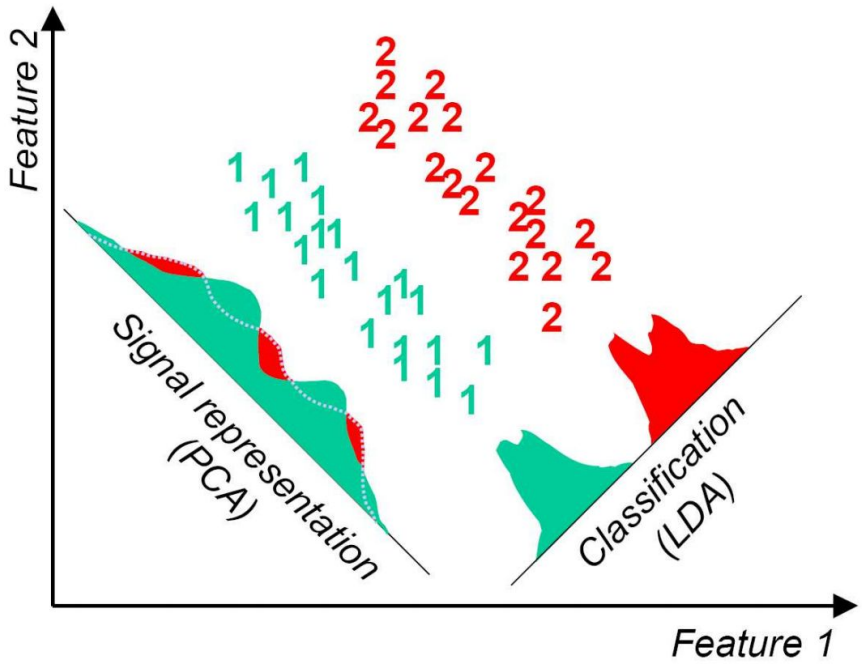


Since we cannot “see” the data in more than three dimensions, the workaround is to visualize the data projecting them in a low(er) dimensional space



This “subspace” is constructed by combining the original variables

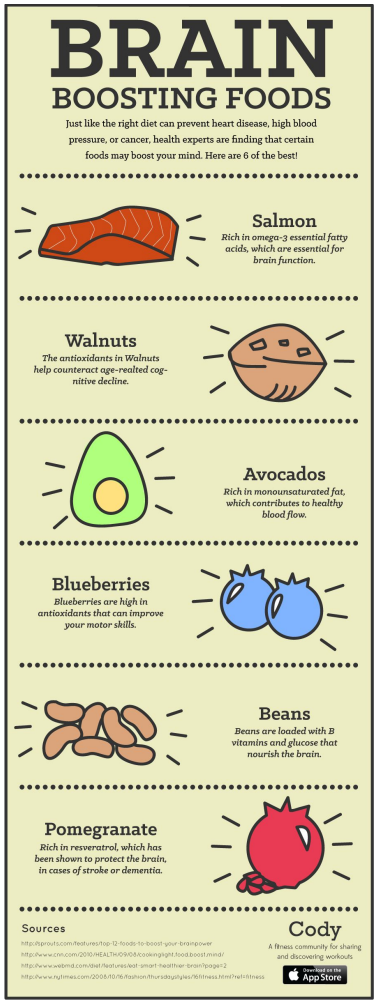
… different criteria



The “lower” dimensional space can be selected with different criteria … here one is “maximal variance”, the other “maximal separation”

… food for brain

- For PCA and LDA the (sub)space is flat
- It could be also not flat ;-)
- … but remember the balance between points and complexity …
- ...

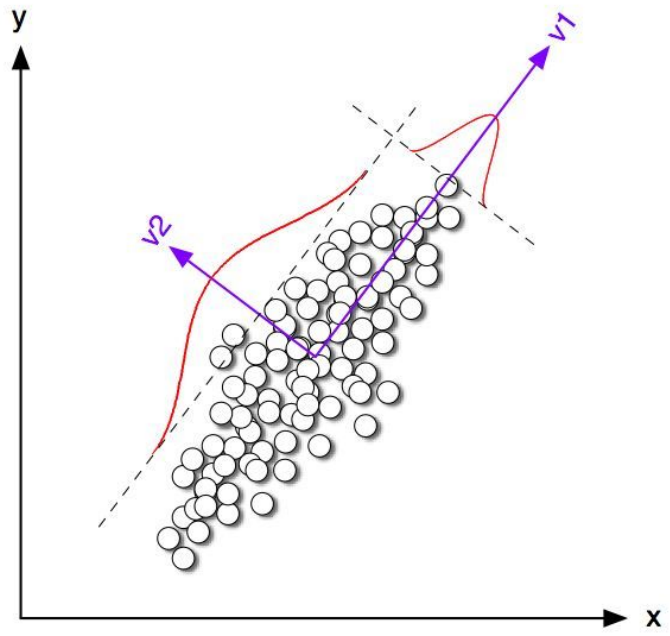


Principal Component Analysis

Maximize Variance

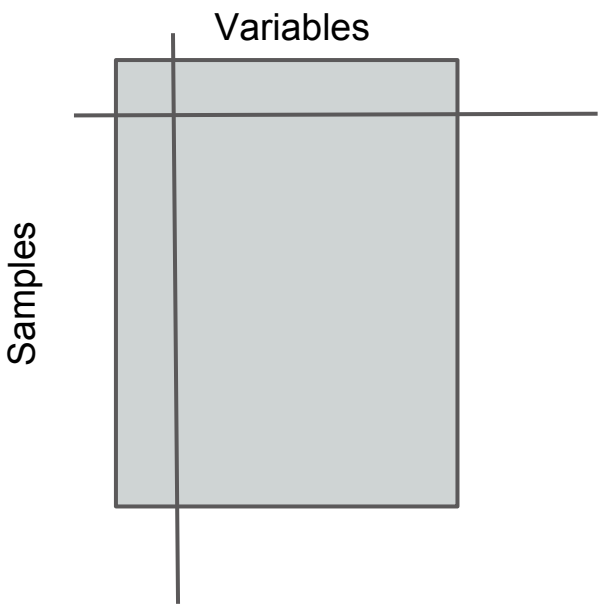
The “objective” of PCA is to identify orthogonal combinations of the original variables maximizing the explained variance of the original dataset

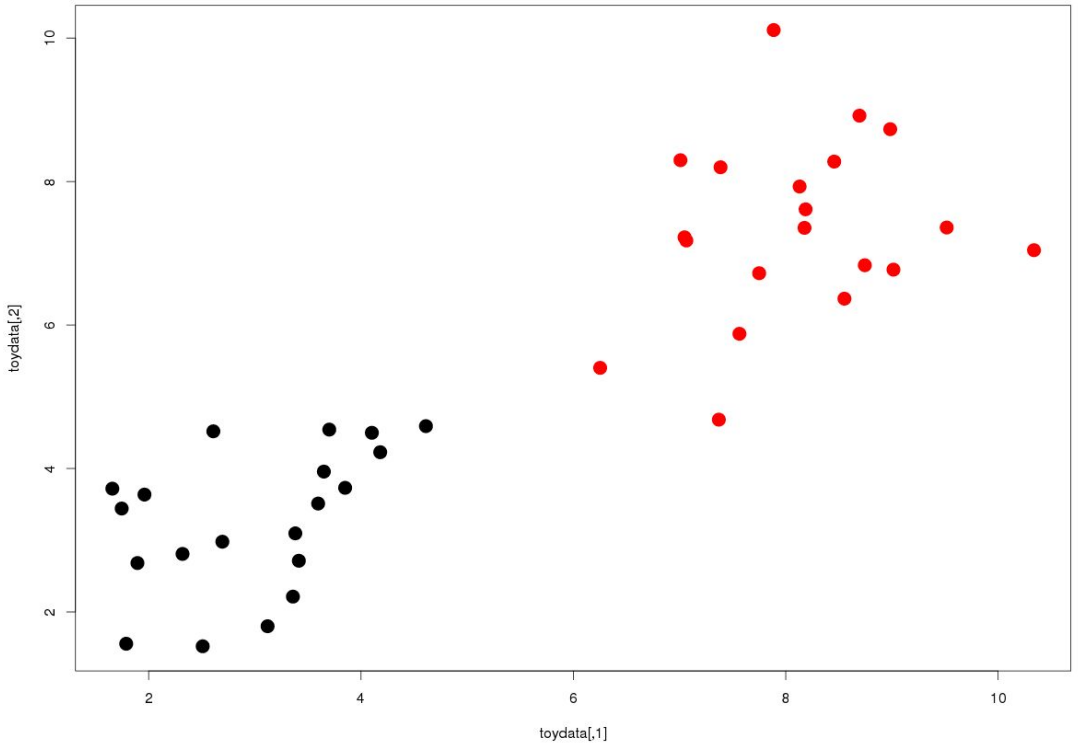
This “view” will enhance the spread of the data

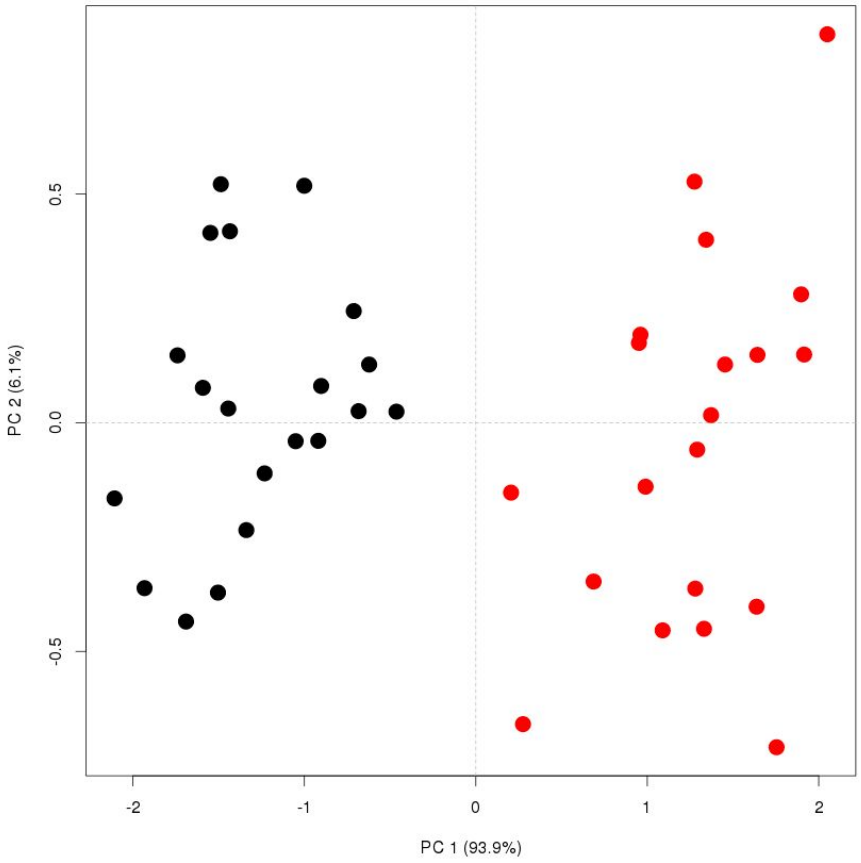


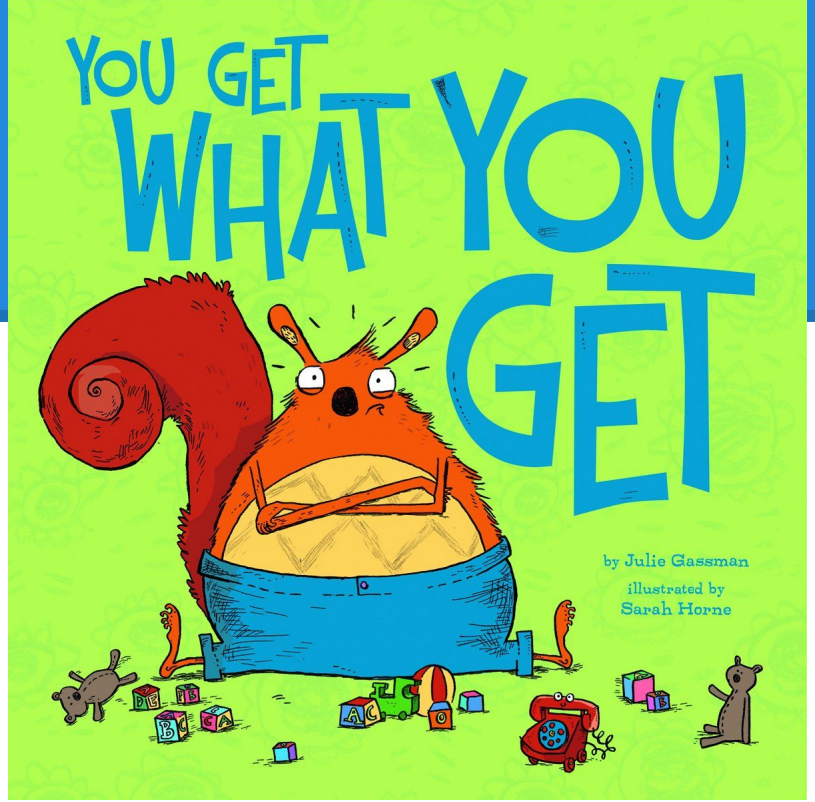
http://setosa.io/ev/principal-component-analysis/

Data Matrix





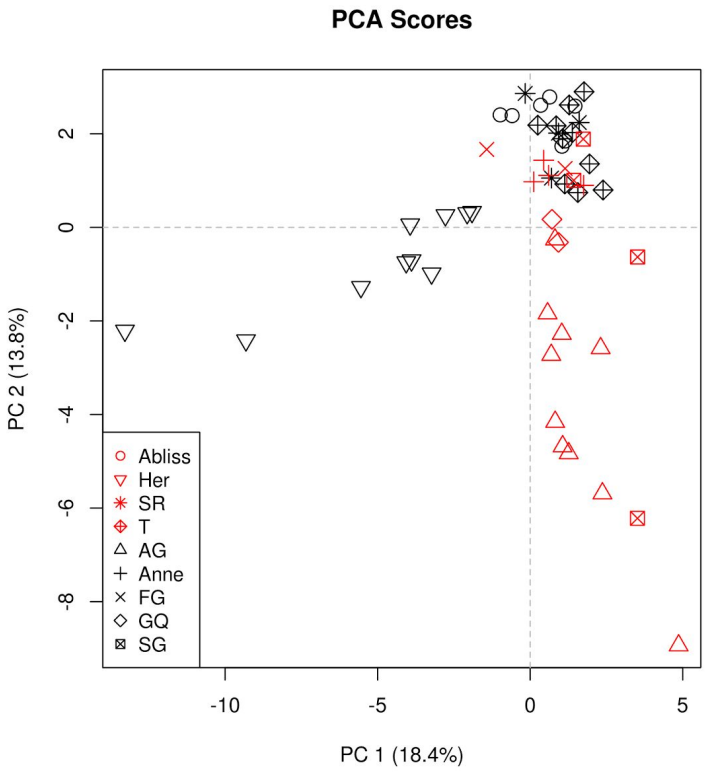

PC 2 (6.1%)
0.5
0.0
-0.5
-2
-1
0
1
2
PC 1 (93.9%)

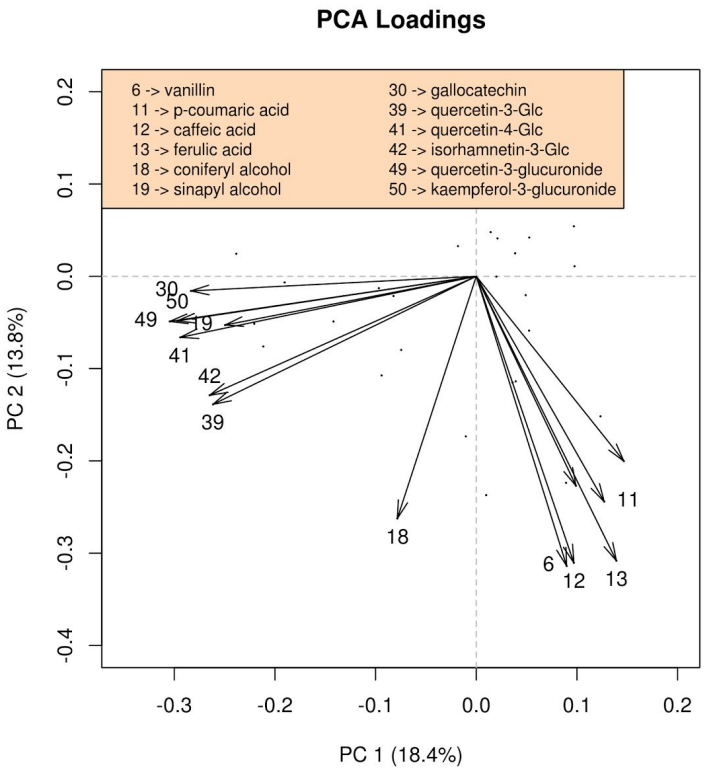


- The positions of the points in this new “space”, the SCORES (scoreplot)
- The weights of the old directions on the new ones. In other words, the rule to construct the new directions from the LOADINGS (loadingplot)

You have a metabolomics experiment where you measure 60 metabolites on 30 urine samples:

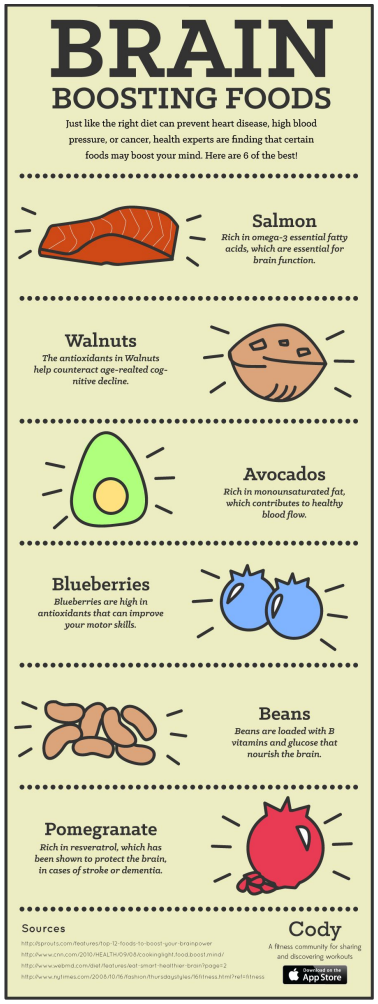
- How many points do I have?
- How big is my multidimensional space?
- How many points do I have in a scoreplot?
- How many points in a loadingplot?

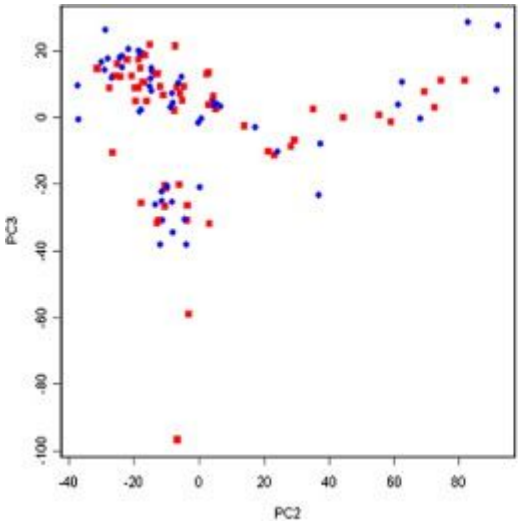

PCA Scores
PC 2 (13.8%)
PC 1 (18.4%)
2
0
-2
-4
-6
-8
-10
-5
0
5
Abliss
Her
SR
T
AG
Anne
FG
GQ
SG


PCA Loadings
6 -> vanillin
11 -> p-coumaric acid
12 -> caffeic acid
13 -> ferulic acid
18 -> coniferyl alcohol
19 -> sinapyl alcohol
30 -> gallocatechin
39 -> quercetin-3-Glc
41 -> quercetin-4-Glc
42 -> isorhamnetin-3-Glc
49 -> quercetin-3-glucuronide
50 -> kaempferol-3-glucuronide
PC 2 (13.8%)
PC 1 (18.4%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
30
50
49
19
41
42
39
18
6
12
13
11

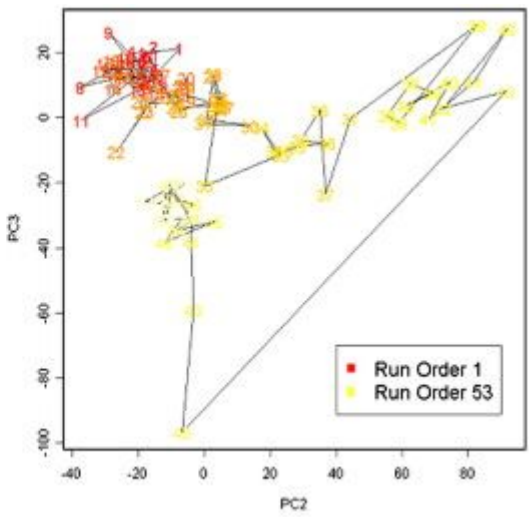
… food for brain

- PCA is useful to reduce the number of relevant variables
- PCA show the big “structure” of my data and this can help in interpretation
- PCA will change if you add points !!!
- PCA is sensitive to the data scaling
- The loadings are not always easy to interpret

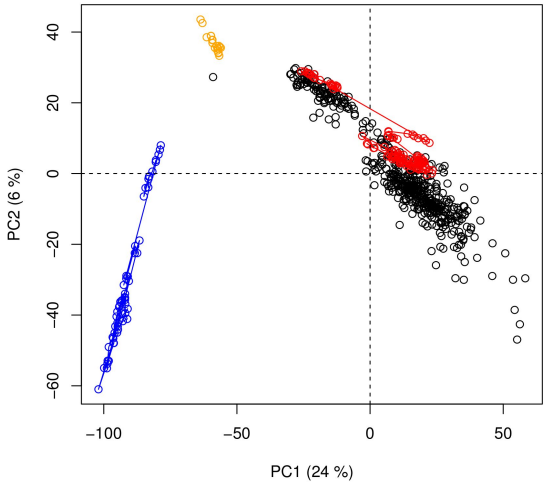


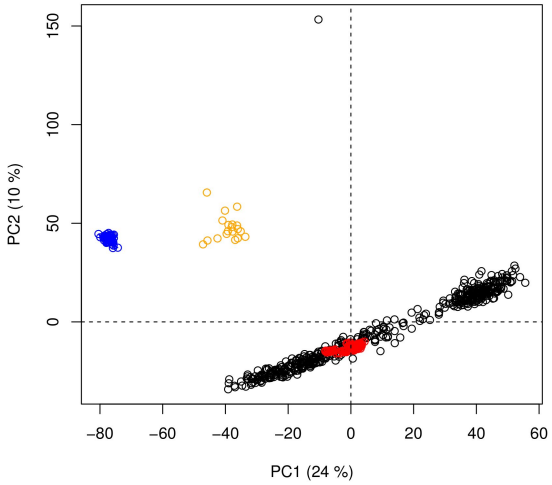


PCA



PCA on Norm. data





Cluster Analysis

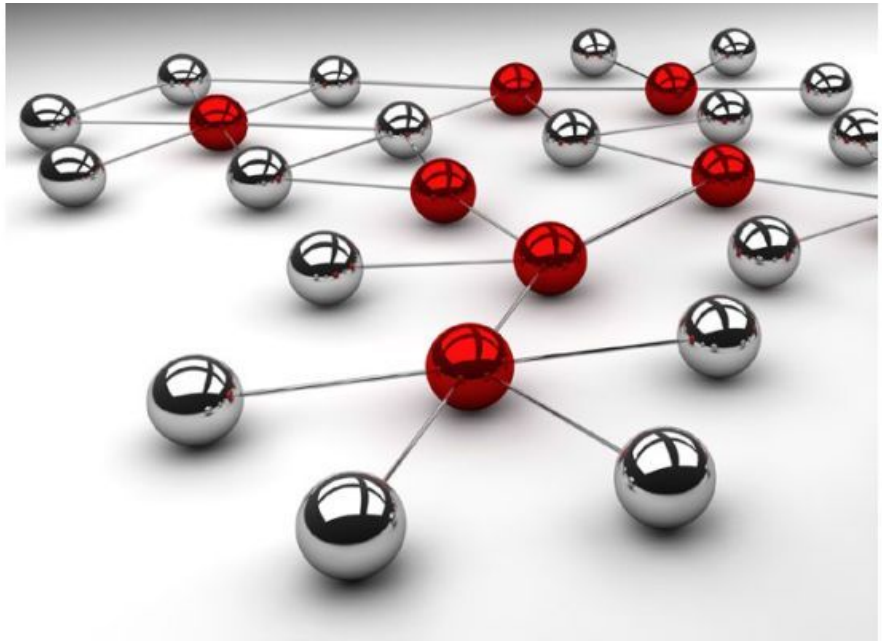
… clustering

Instead of looking for specific projection of the data cloud, an alternative is to look for “groups” (clusters) of samples in the n dimensional space

Clusters ...

Clustering techniques require a way to assess the similarity between the samples/objects

In many cases this similarity/dissimilarity is evaluated using a measure of distance

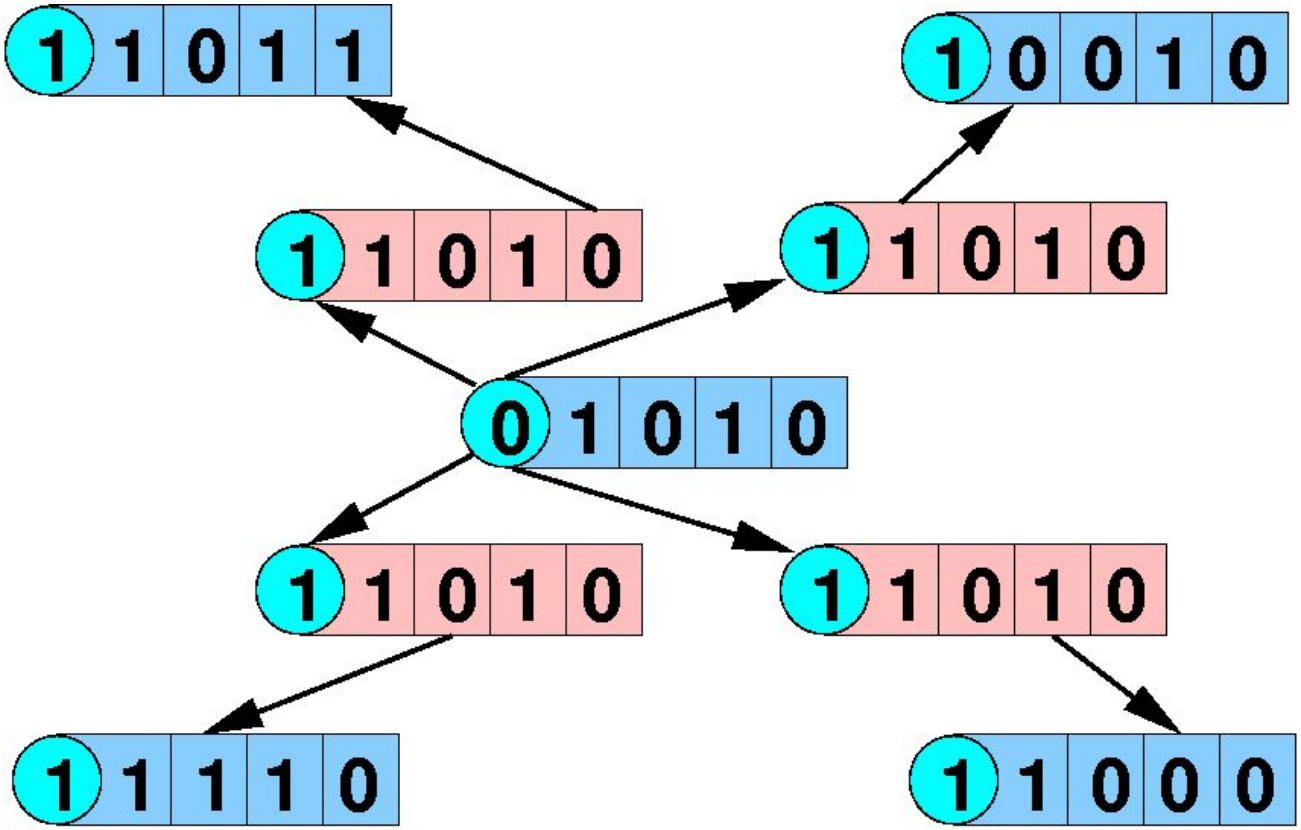


… my keyword

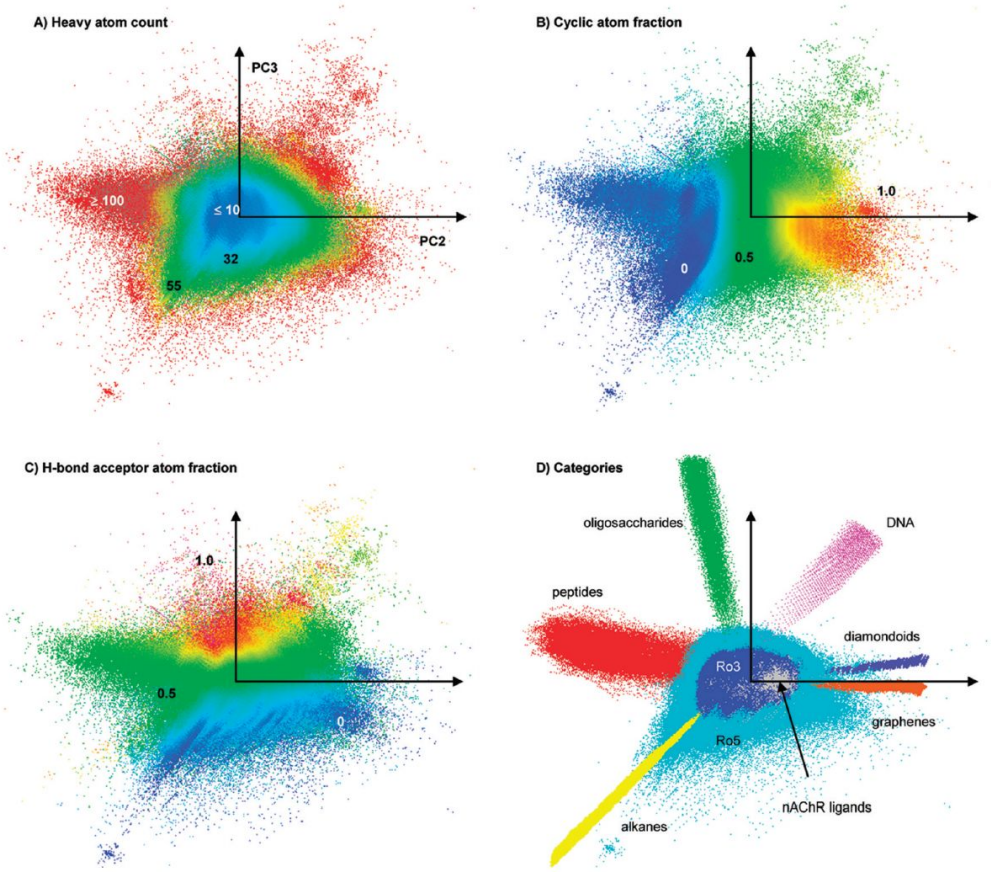
- Scaling
- Distance
- Algorithm
- ...

The risk is to play around until you see what you want to see ...

Hamming Distance



Chemical Similarity

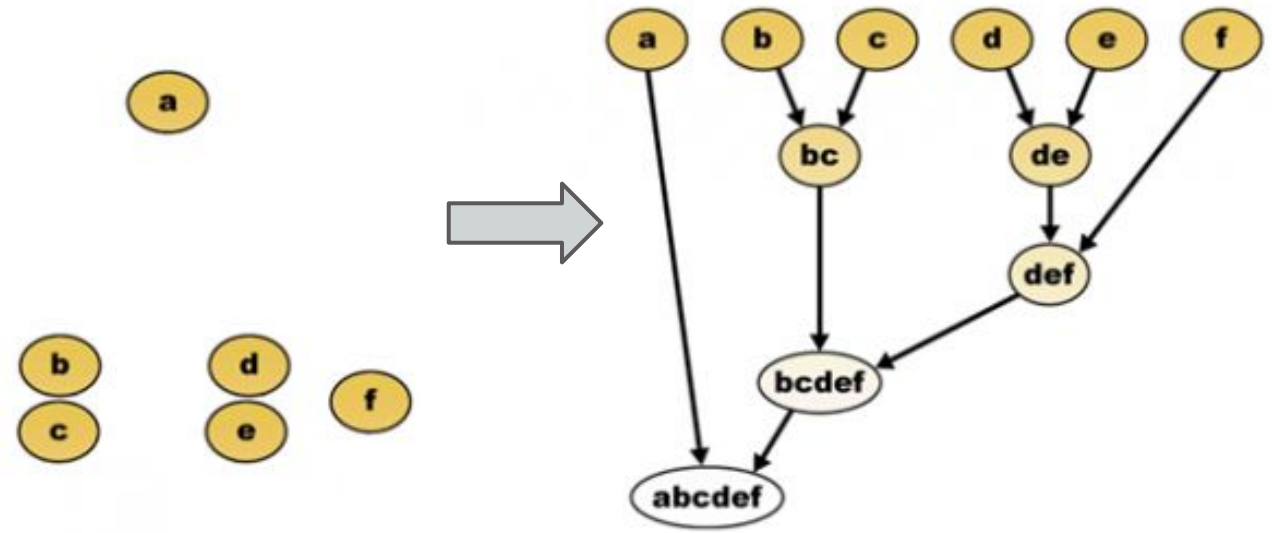


What are “reasonable” measures of similarity in the following omics?

- Metabolomics
- Genomics
- ...

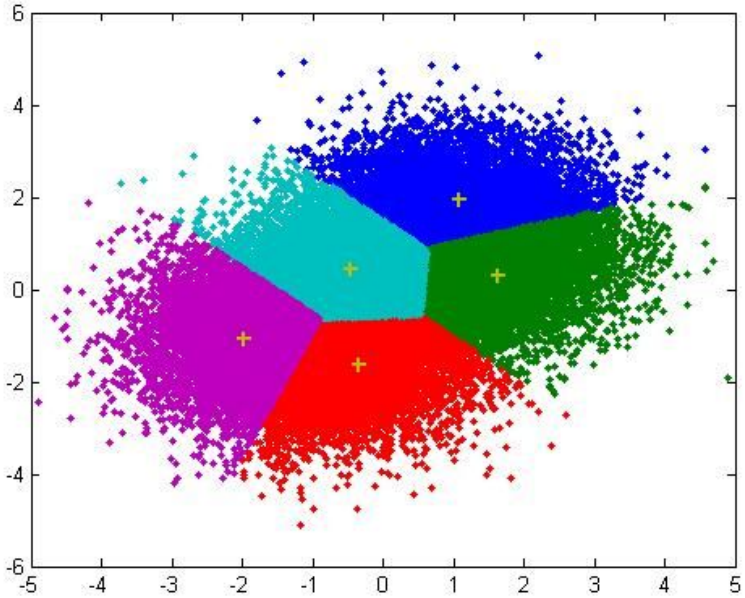
Hierarchical Clustering

Initially each element constitutes a cluster on its own. Sequentially, the clusters are combined until all the elements belong to one big cluster



Partitional Clustering

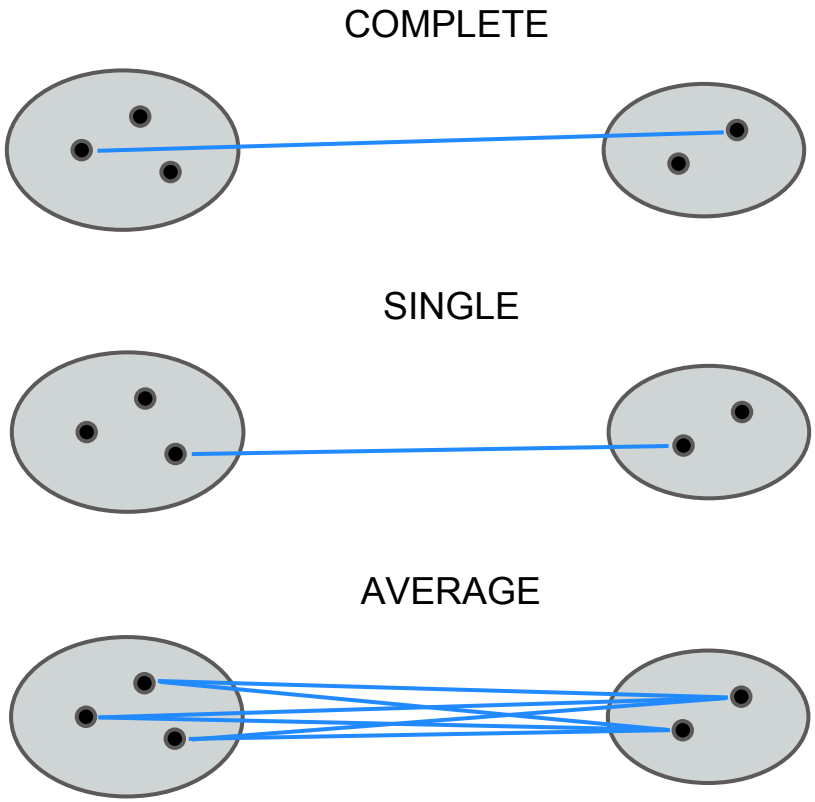
The complete set of elements is divided into n groups putting together the more similar elements



Hierarchical Clustering: ingredients

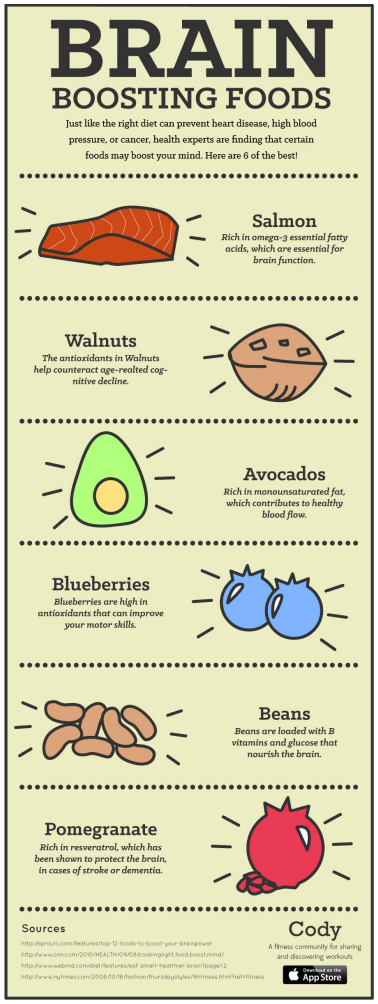
- A dataset
- A measure of distance
- A strategy to calculate the distance between clusters (linkage)
- … a good computer

Distance between among clusters: linkage



… food for brain

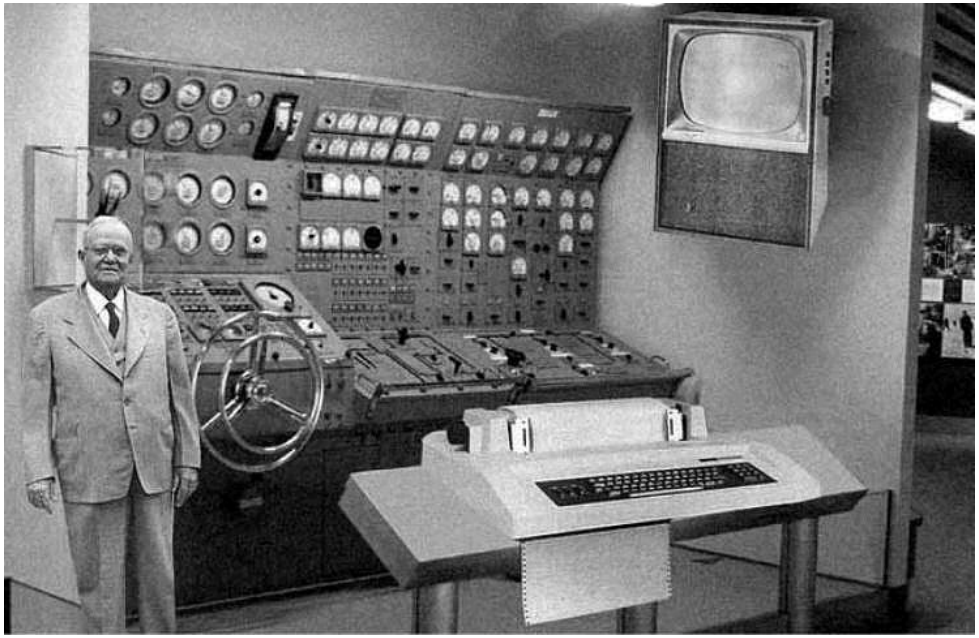
- “Single” and “complete” linkages can give distorted representation of the dataset in presence of outliers
- The “best” linkage can be selected knowing the characteristics of the dataset under study ...



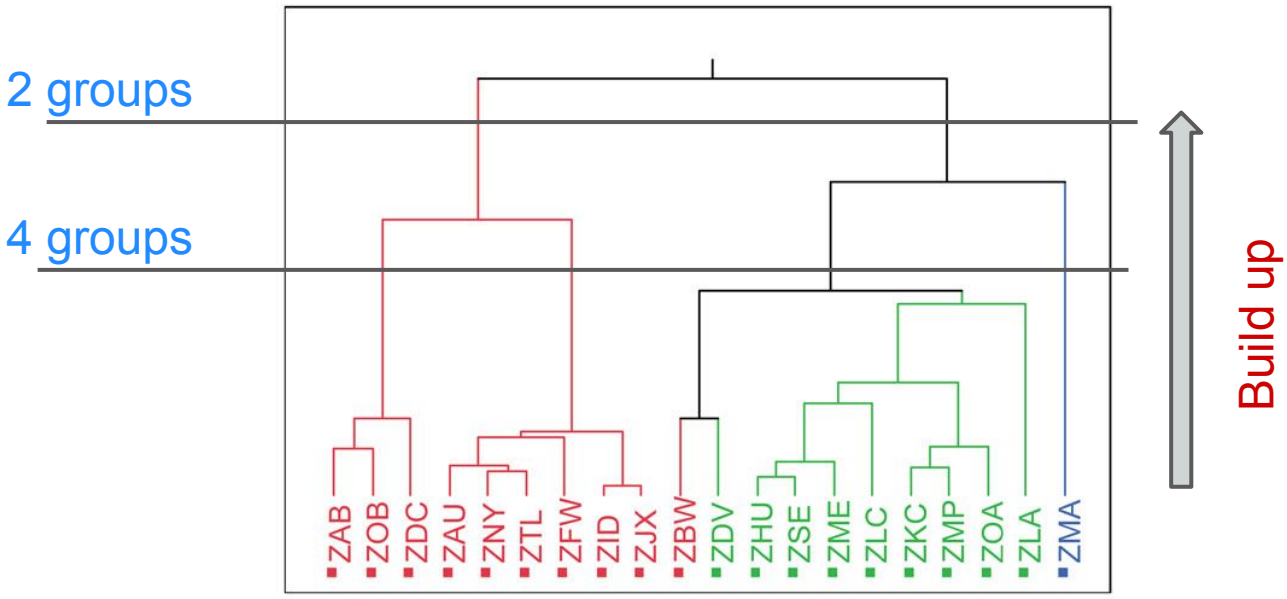
Un good computer ...

You need to calculate the distance between all the elements

It scales with the square of the number of samples



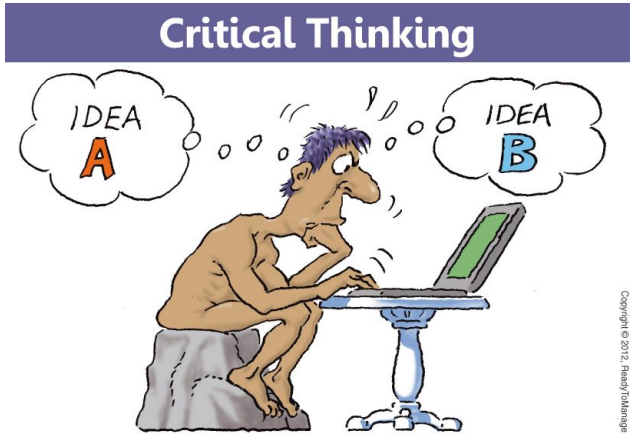
Dendrogram



Where should I cut it ?

General thoughts

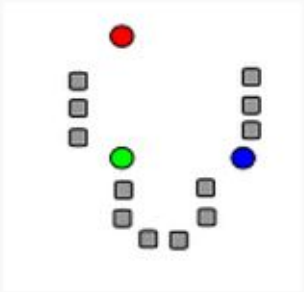
- You need to calculate the distance among all the elements.
- Once done, you can cut the tree wherever you want.
- If you read it from the top, if two elements are split, they will be in different groups until the end ...
- With big datasets it becomes quite computationally demanding ...



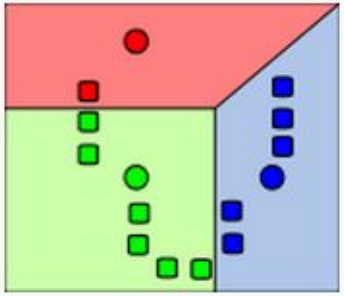
K-means clustering: ingredients

- A dataset
- An euclidean distance
- The number of groups!
- … a reasonable computer

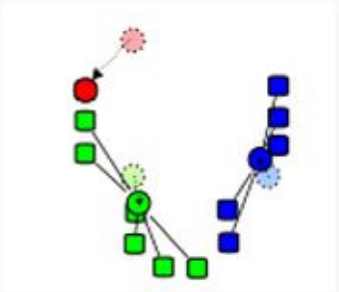
K-means algorithm



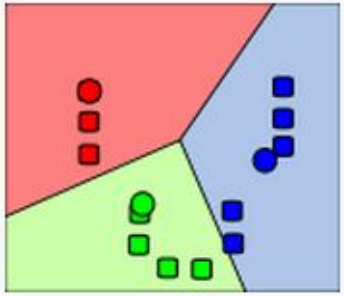
1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3) The centroid of each of the *k* clusters becomes the new mean.



4) Steps 2 and 3 are repeated until convergence has been reached.

Minimizes ...

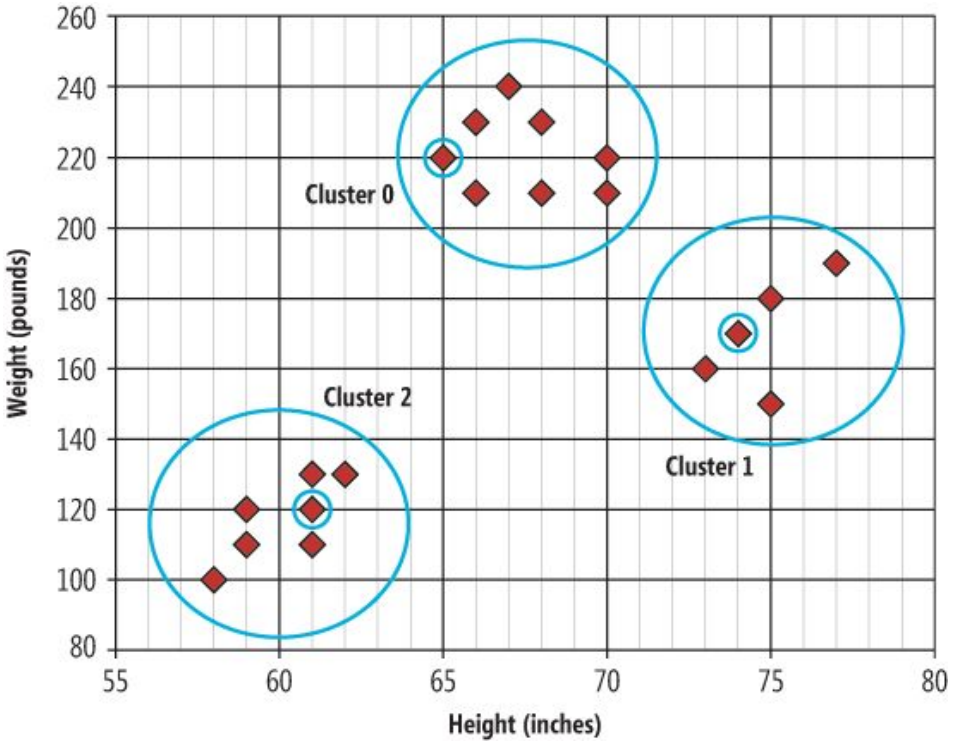
$$\sum_{i=1}^{k} \sum_{p \epsilon C_i} \|p - \mu_i\|^2$$

Distance from the center

For each cluster

For each element in the cluster

k-medoids

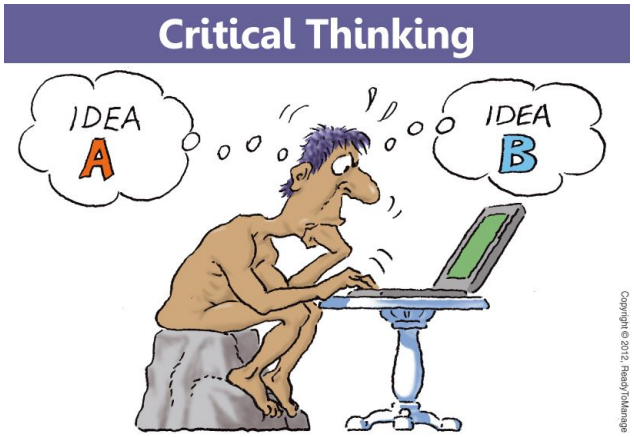

Cluster 0
Cluster 1
Cluster 2
Weight (pounds)
Height (inches)
260
240
220
200
180
160
140
120
100
80
55
60
65
70
75
80

A reasonable computer ...

You need to calculate the distance between the centers and all the elements of the dataset

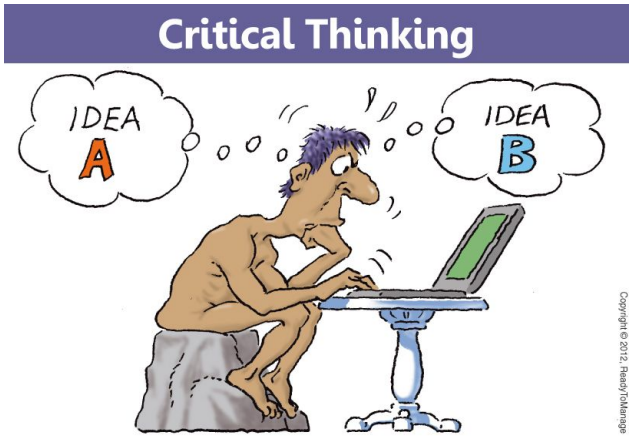
General thoughts (1)

- You do not need need to calculate the distance among all the elements.
- If you change your mind you should re-calculate everything.
- It is fast also with big dataset …
- If you change the starting points the class membership could change.



General thoughts (2)

- You should decide the number of clusters ...
- …

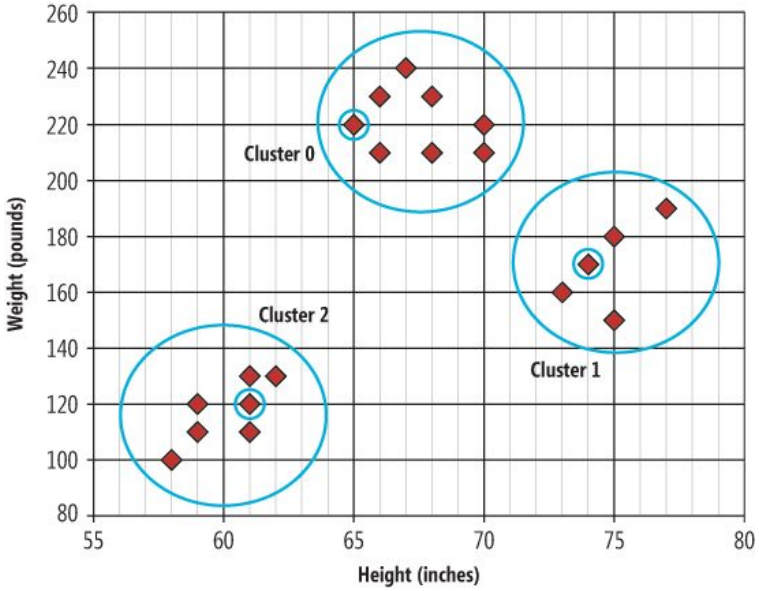


GUESS how MANY?

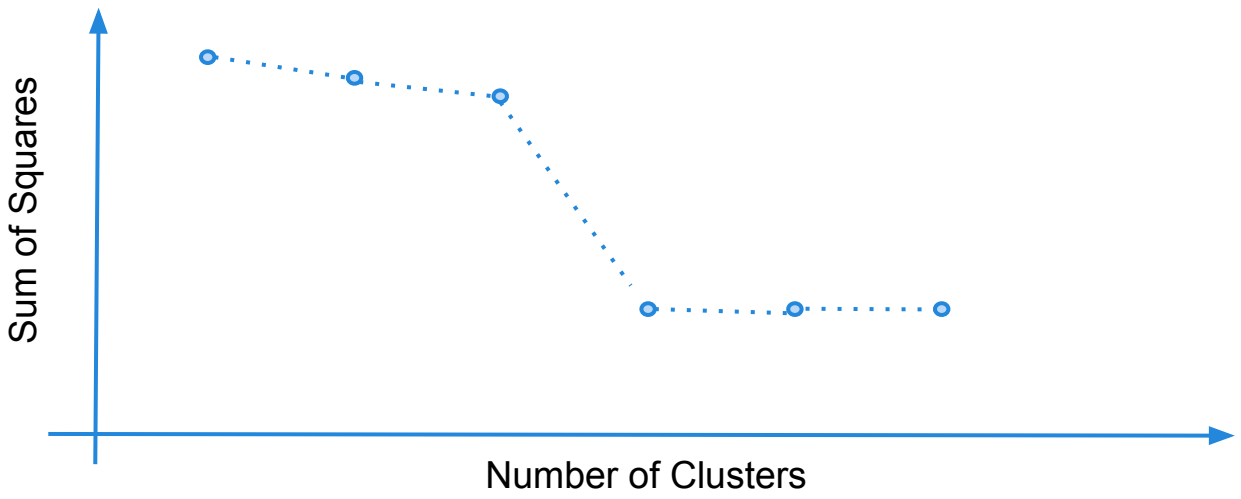
Clusters ...

Choosing the number of clusters

… an intelligent idea is to look to the "within-cluster" sum of squares as a function of the number of clusters ...

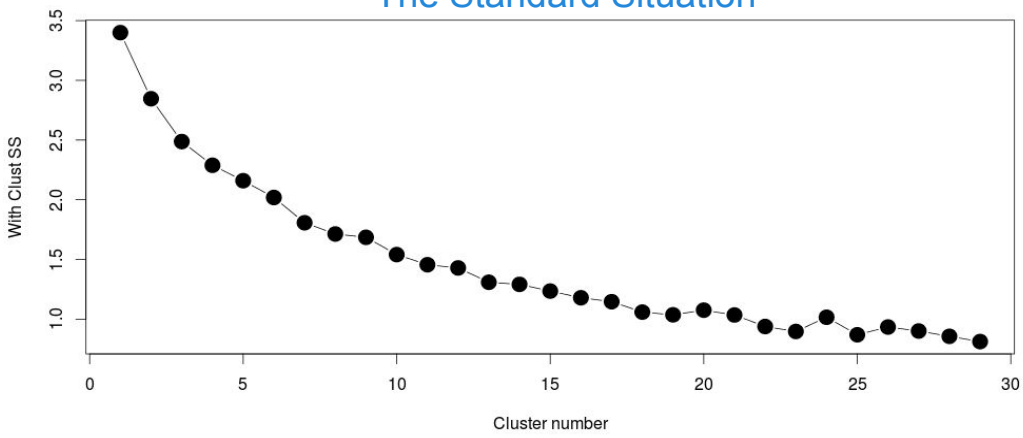


Choosing the number of clusters

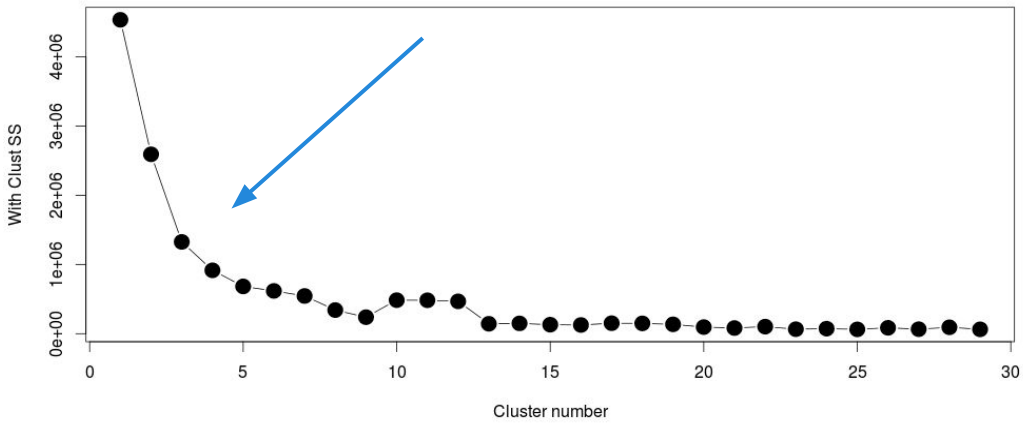


Real Life

The Standard Situation

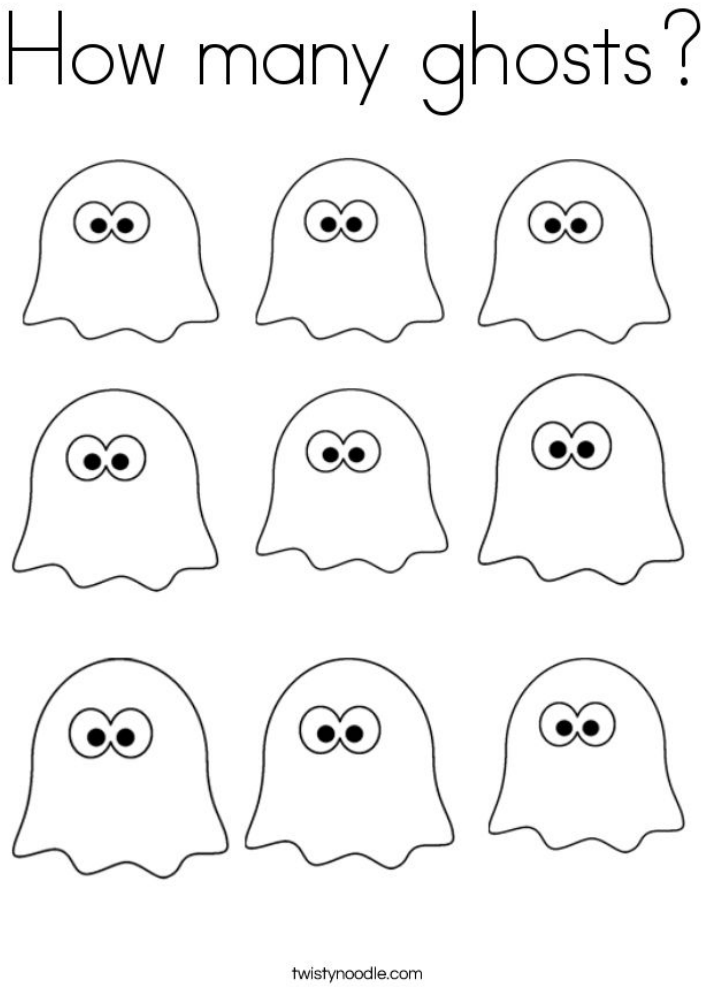


Nicely clustered dataset

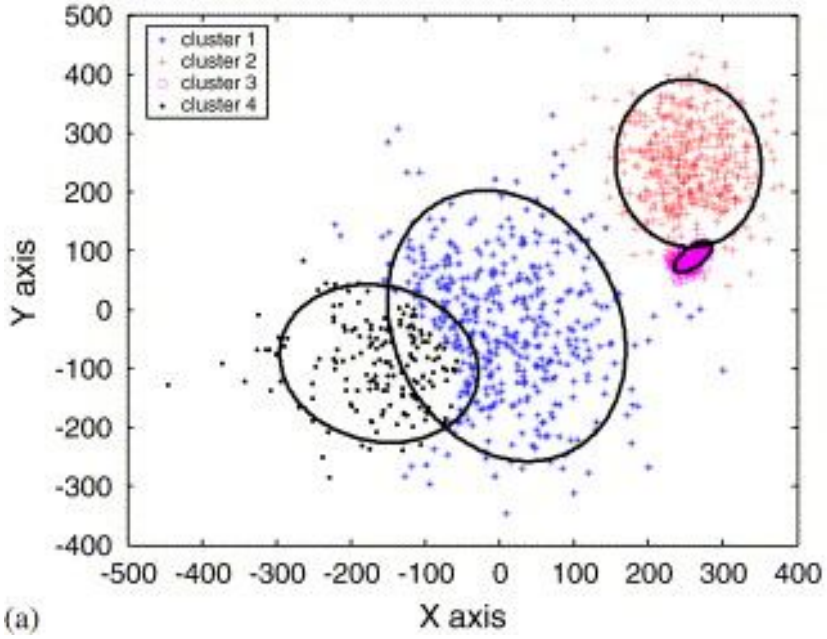


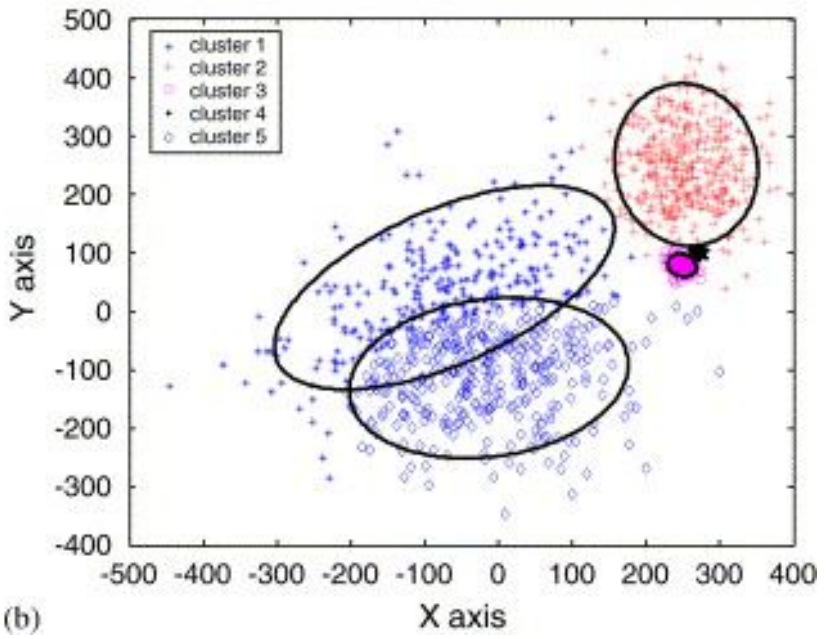
Again on the number of clusters ...

- In segmentation problems this parameter is given
- External knowledge (biological) can help
- Gap statistics (Tibshirani 2001)
- ...



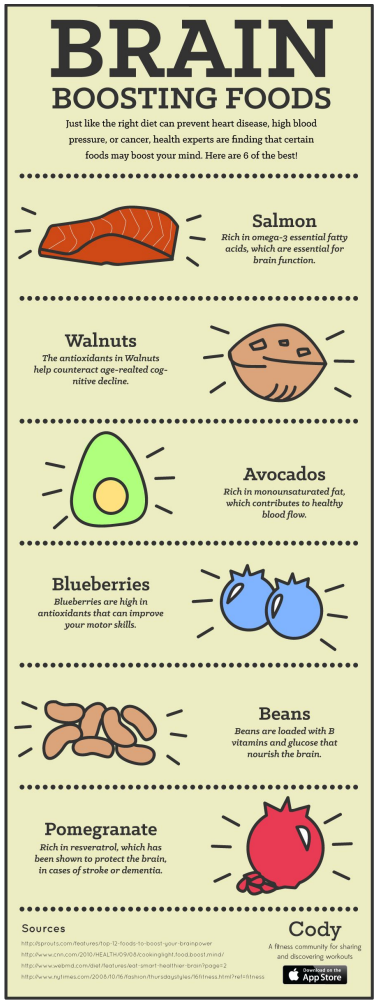
… only the starting point …





… flexibility

- Use it to get a feeling on your data
- In this way is good to play with it
- Do not overinterpret … ;-)
- Remember that 20 points in a 20k dimensional space looks really alone …
-




#1
LETS GO
LIVE!!!