Multivariate exploration of omics data

Descriptive, PCA and Clustering

Alex Sanchez-Pla

Department of Genetics, Microbiology and Statistics University of Barcelona

Version 2024-03-25

# Scheme

1. Introduction and Motivation
2. Multivariate Data Exploration
3. Reduction of dimension
4. Discovering groups in your data
5. References and Resources

2 / 45

# Introduction and motivation

3 / 45

# The right approach for Omics?

Classical Statistics

 Multiple regression

 ANOVA  Discriminant Analysis

Data tables are *long* and *not*

*very wide ("more individual than variable")*

# The nature of omics data

The omics data have a different structure than the usual one in multivariate statistics.

Measure different characteristics

 GC/MS spectra, expression, concentration...

Although they have aspects in common

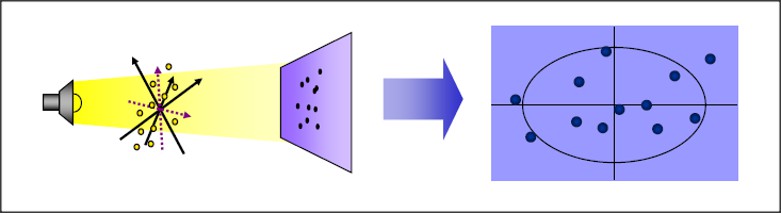


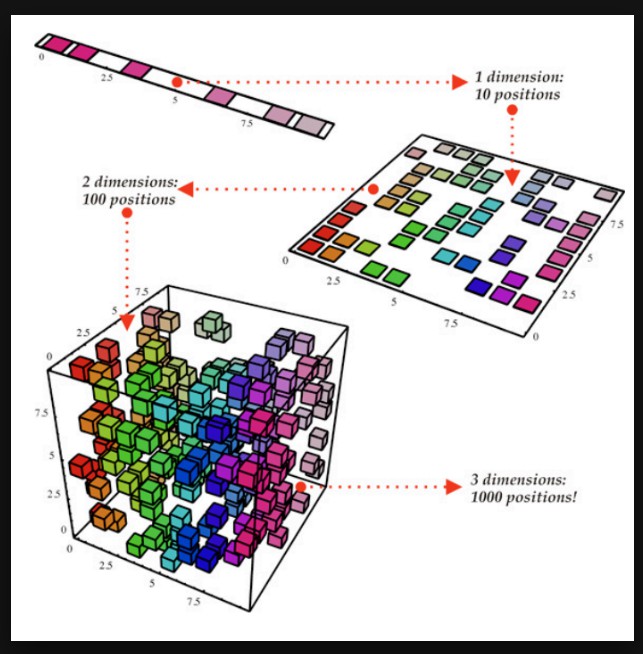
 Many variables (**K**) measured simultaneously

 Few samples (**N**) analyzed

Most of them are high-performance

# A (more) adequate approach

With omics data, it is common to apply methods called *projection* or also  *dimension reduction*

They have interesting advantages

 They examine ALL the variables together

 Prevents information loss

 Find *underlying trends*

= "variables latentes"

# What is a projection?

Variables = axes in a multidimensional space

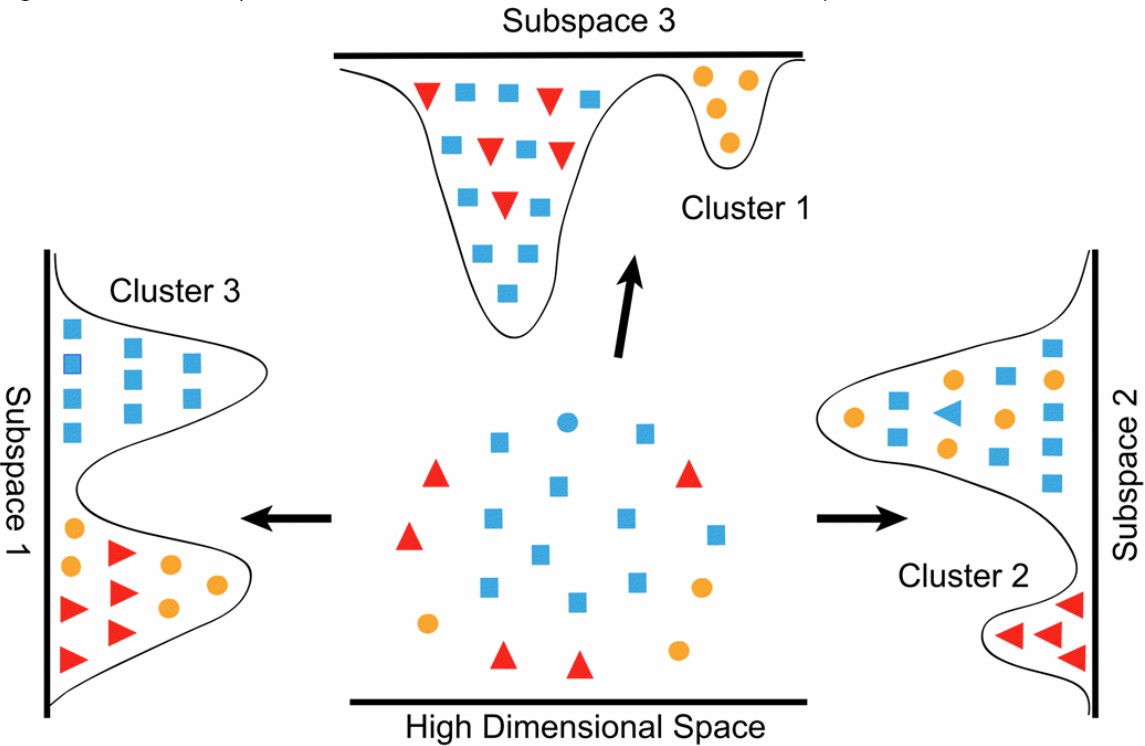
Observations = points in multidimensional space

Projecting points on (hyper)planes

*It may be possible*

 Simplify/Improve Visualization

 Highlight inherent or natural groupings of data



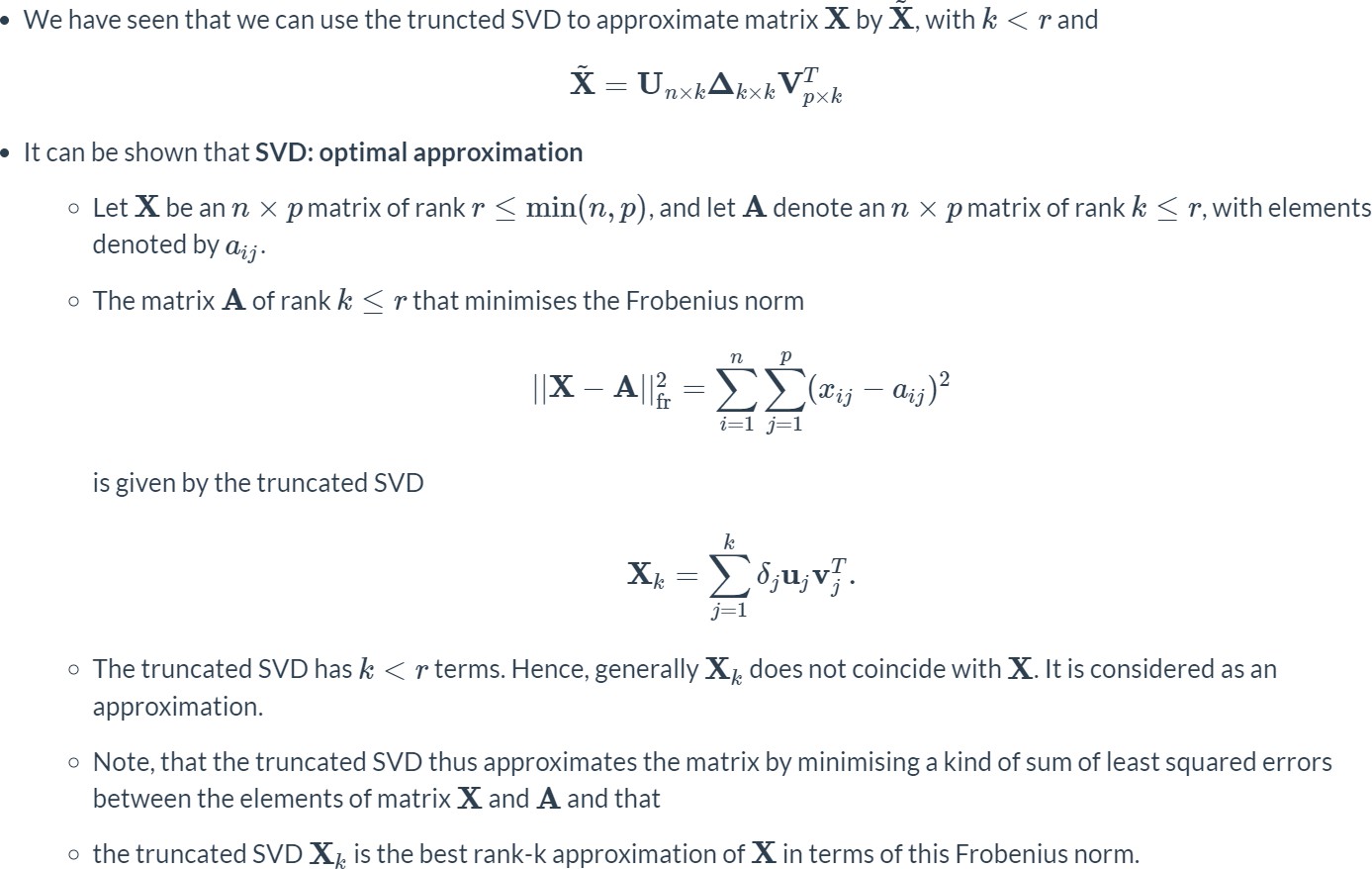
[Data Science post: What is dimensionality](https://datascience.stackexchange.com/questions/130/what-is-dimensionality-reduction-what-is-the-difference-between-feature-selecti) [reductio](https://datascience.stackexchange.com/questions/130/what-is-dimensionality-reduction-what-is-the-difference-between-feature-selecti)

 Highlight the relationship between these groups.

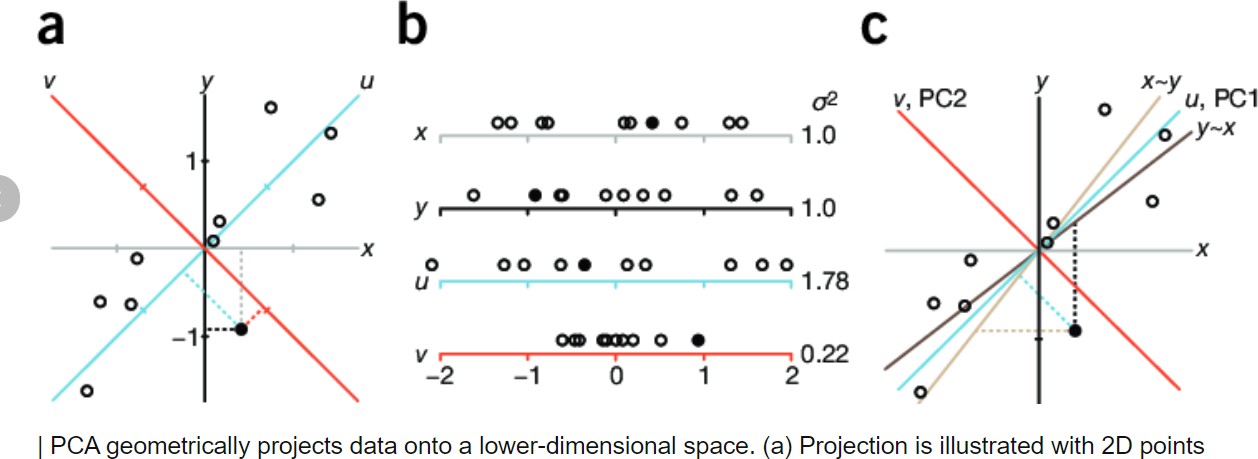
 Detect issues such as batch effects or outliers

7 / 45

# Geometric and algebraic approach

Projections can always be considered from an algebraic or geometric point of view

 Algebraically: The information from the observations is summarized in some new (latent) variables

 Geometrically: The set of points in a dimensional space K (K = number of variables) is *approximated* by a (hyper)plane, that is, a space of smaller dimension where the points are projected.

# Content organization

The initial exploratory analysis

 Description and possible transformations

Dimension projection/reduction or matrix decomposition methods.

 We will focus on Principal Component Analysis (PCA), but we will mention other approaches

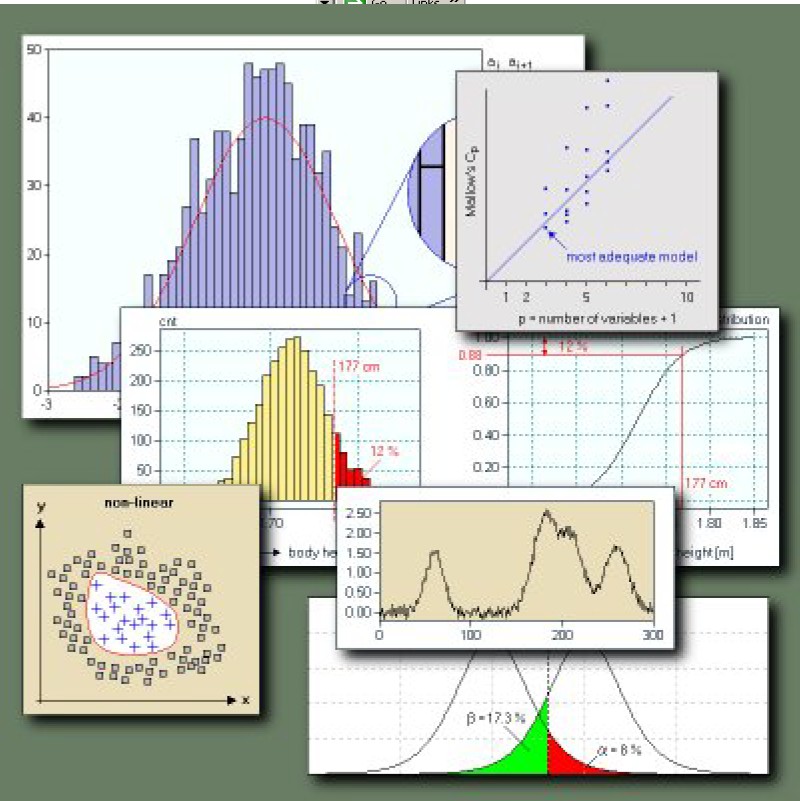
Methods of searching for patterns or classes

 Cluster Analysis and Agglomerative Methods (KMeans/PAM)

# Multivariate Data Exploration

10 / 45

# I. The description of the data

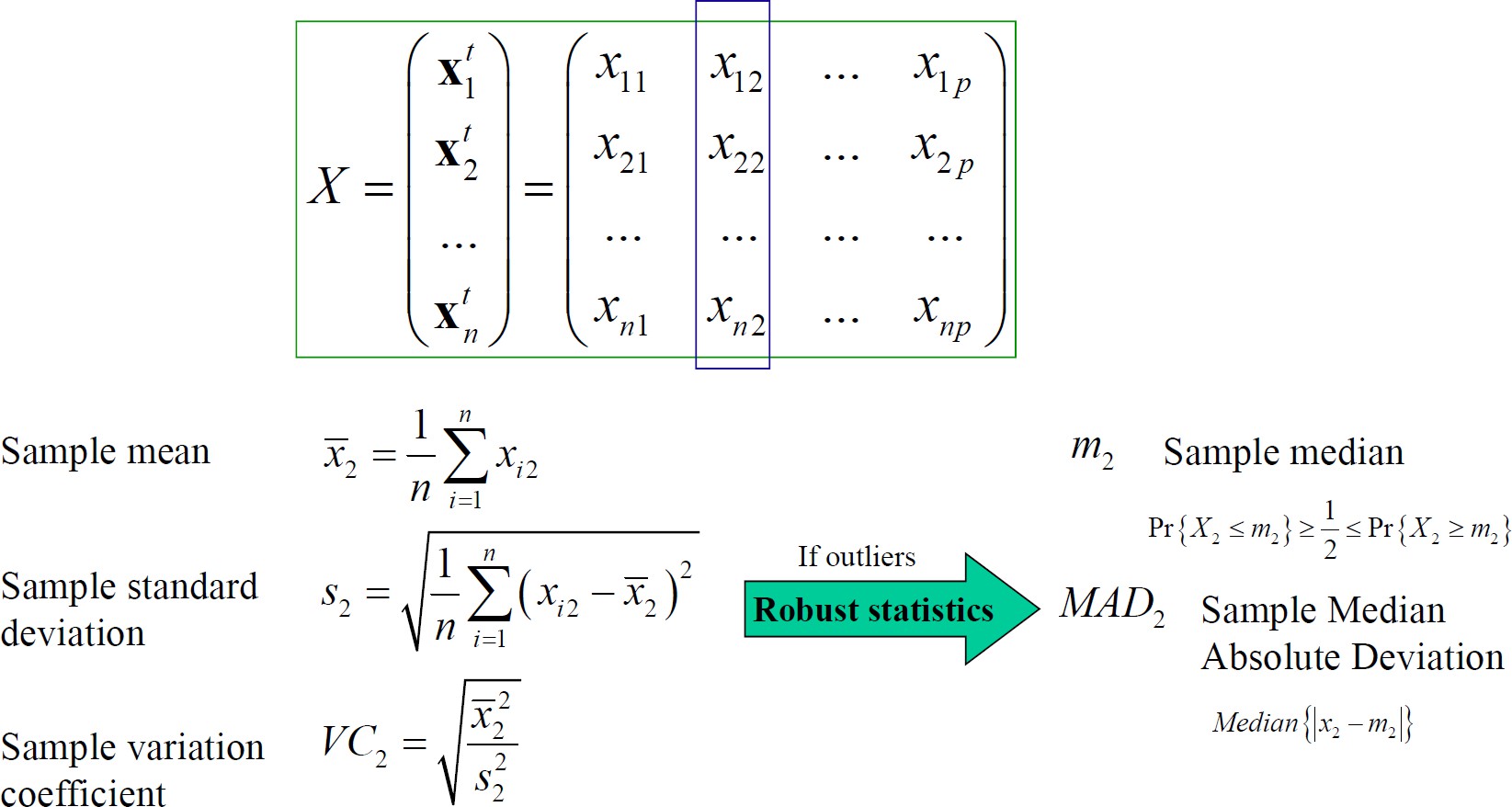
We will begin any statistical analysis by looking at the data

 Which/How many variables,  Which/How many samples  Missing values?

 Get simple statistics and summary charts

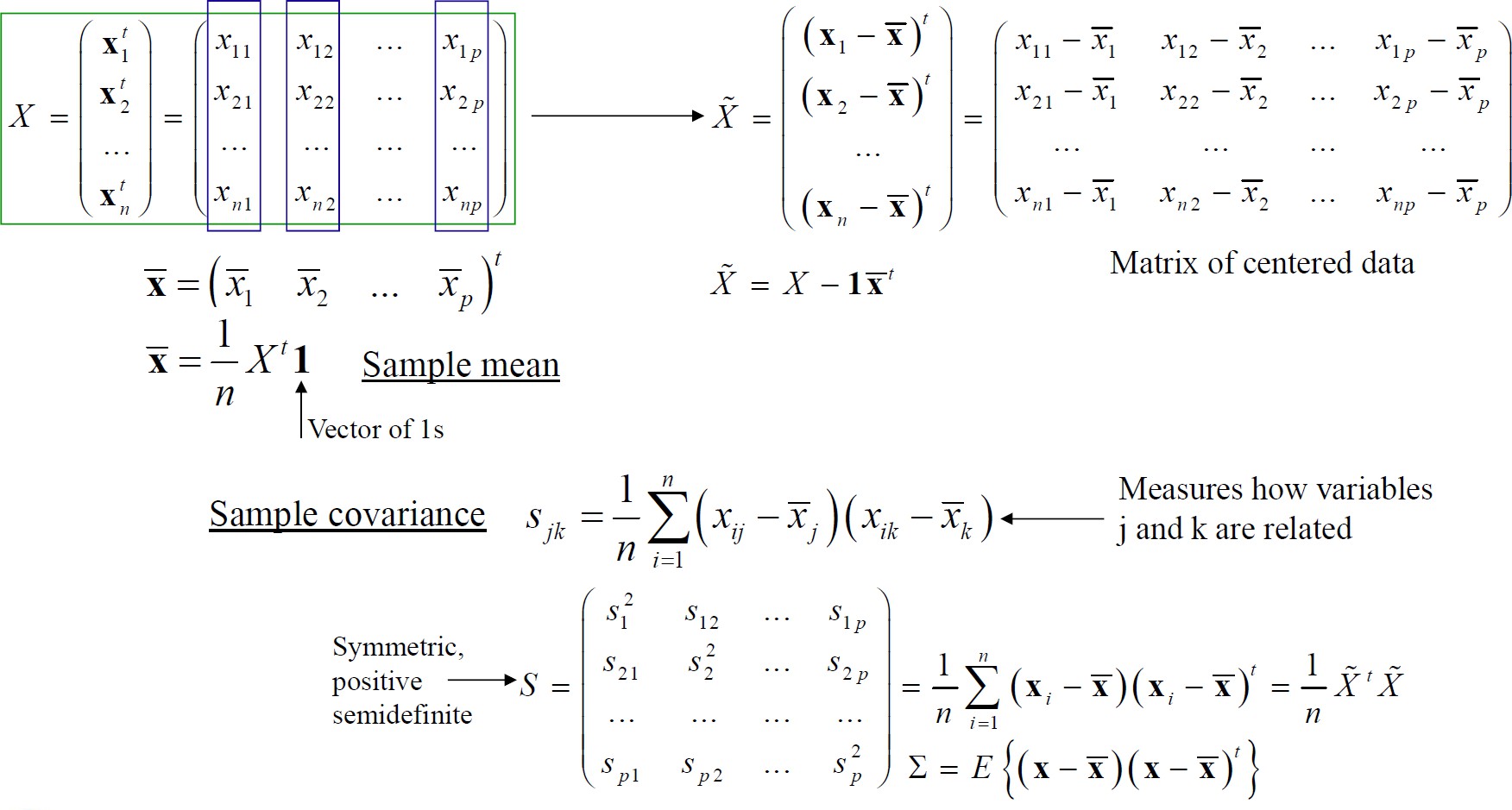
We can calculate the usual descriptive statistics, starting with univariate measures and progressing to bivariate or multivariate statistics

# Univariate numerical abstracts



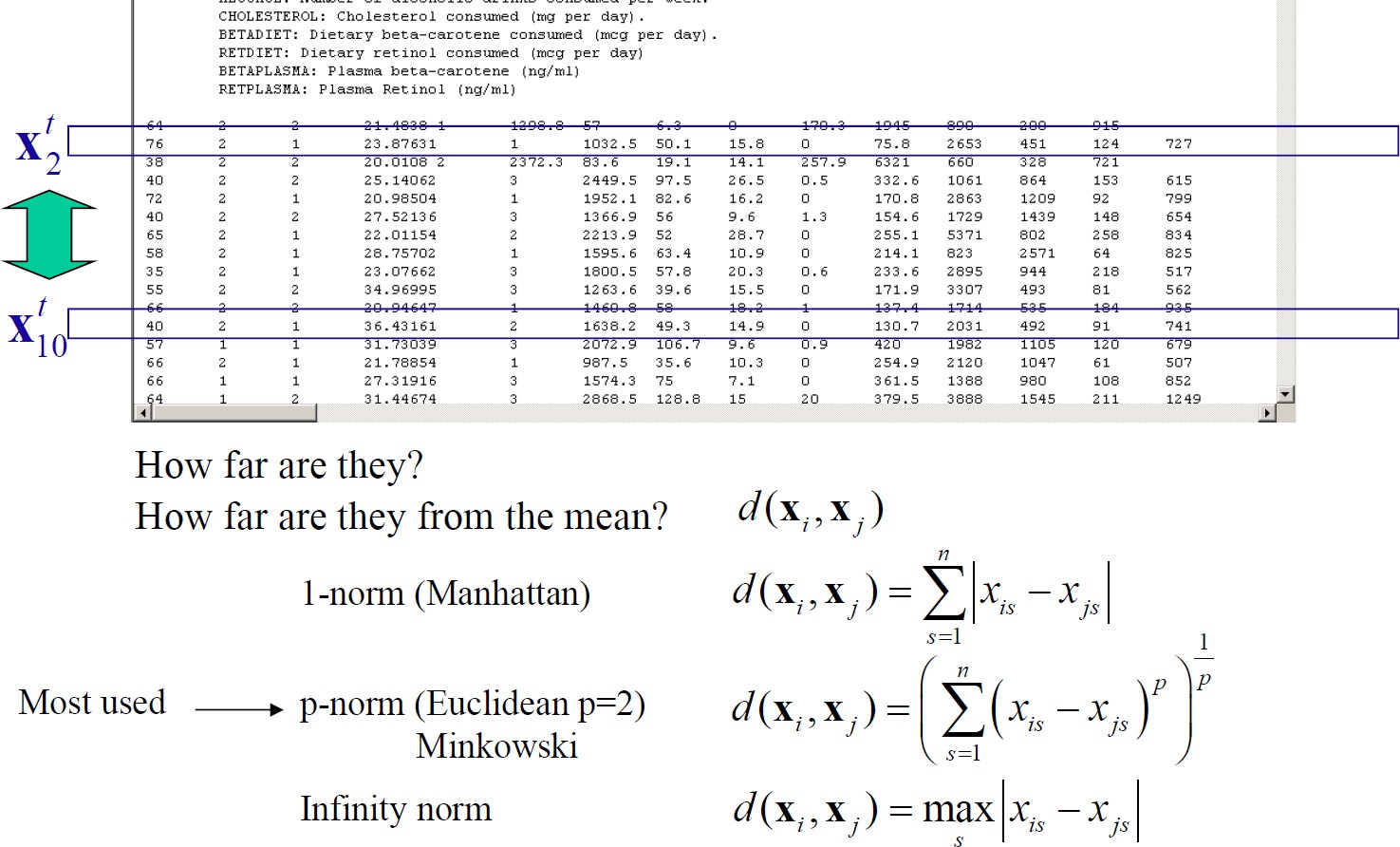
[Source: Multivariate Data Analysis Course. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Bivariate numerical abstracts



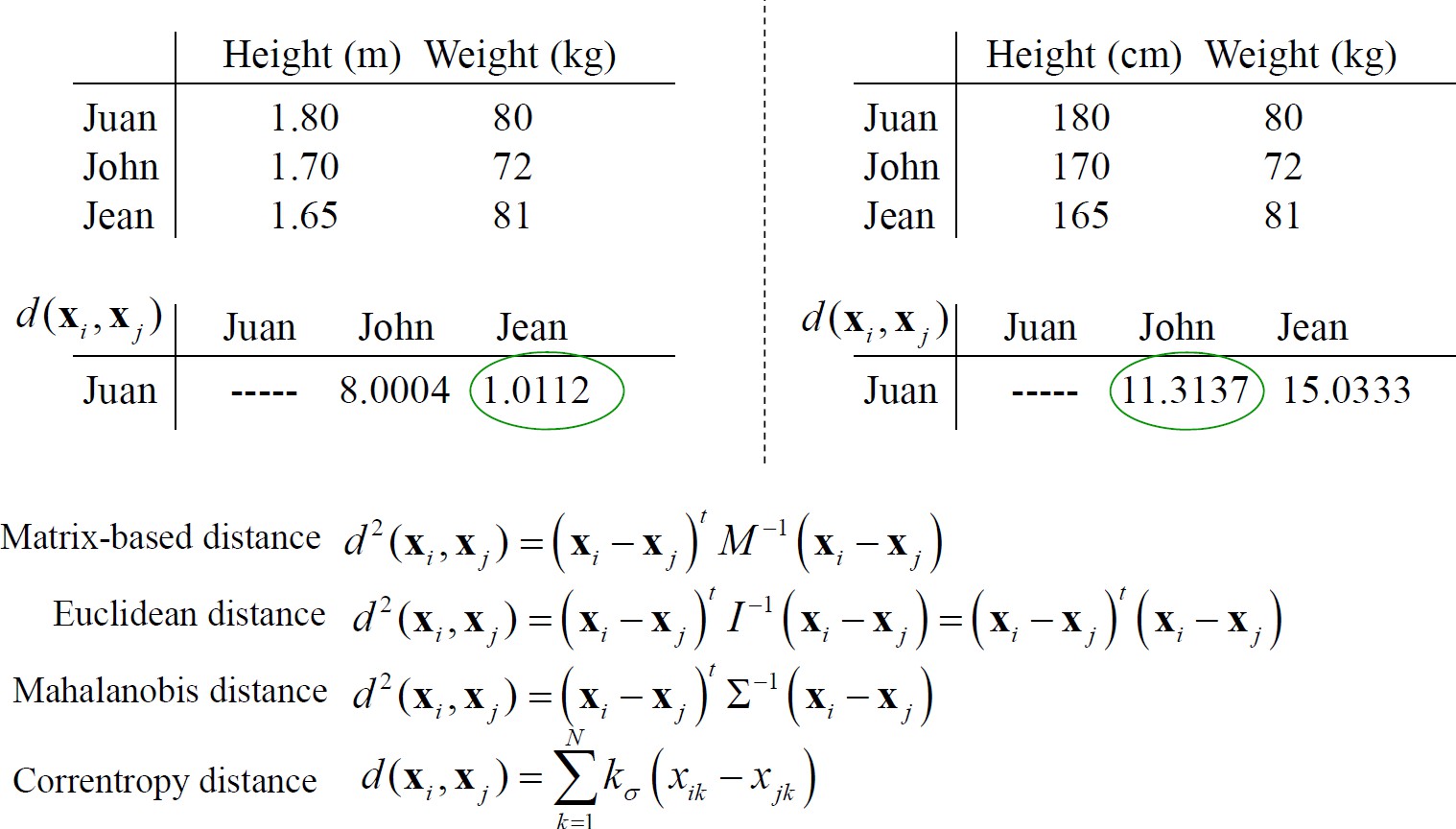
[Source: Multivariate Data Analysis Course. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Multivariate distances (1)



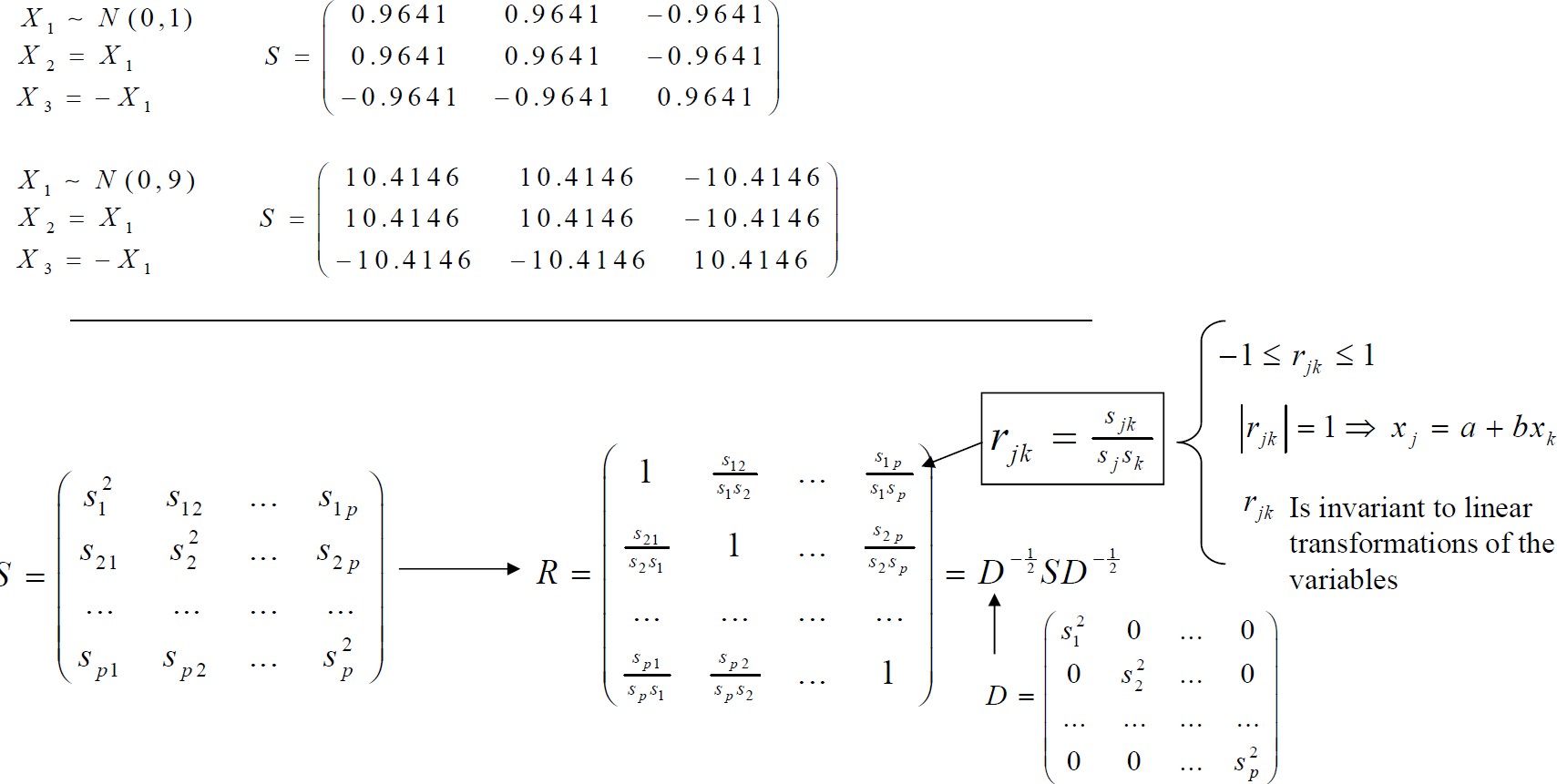
[Source: Multivariate Data Analysis Course. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Multivariate distances (2)



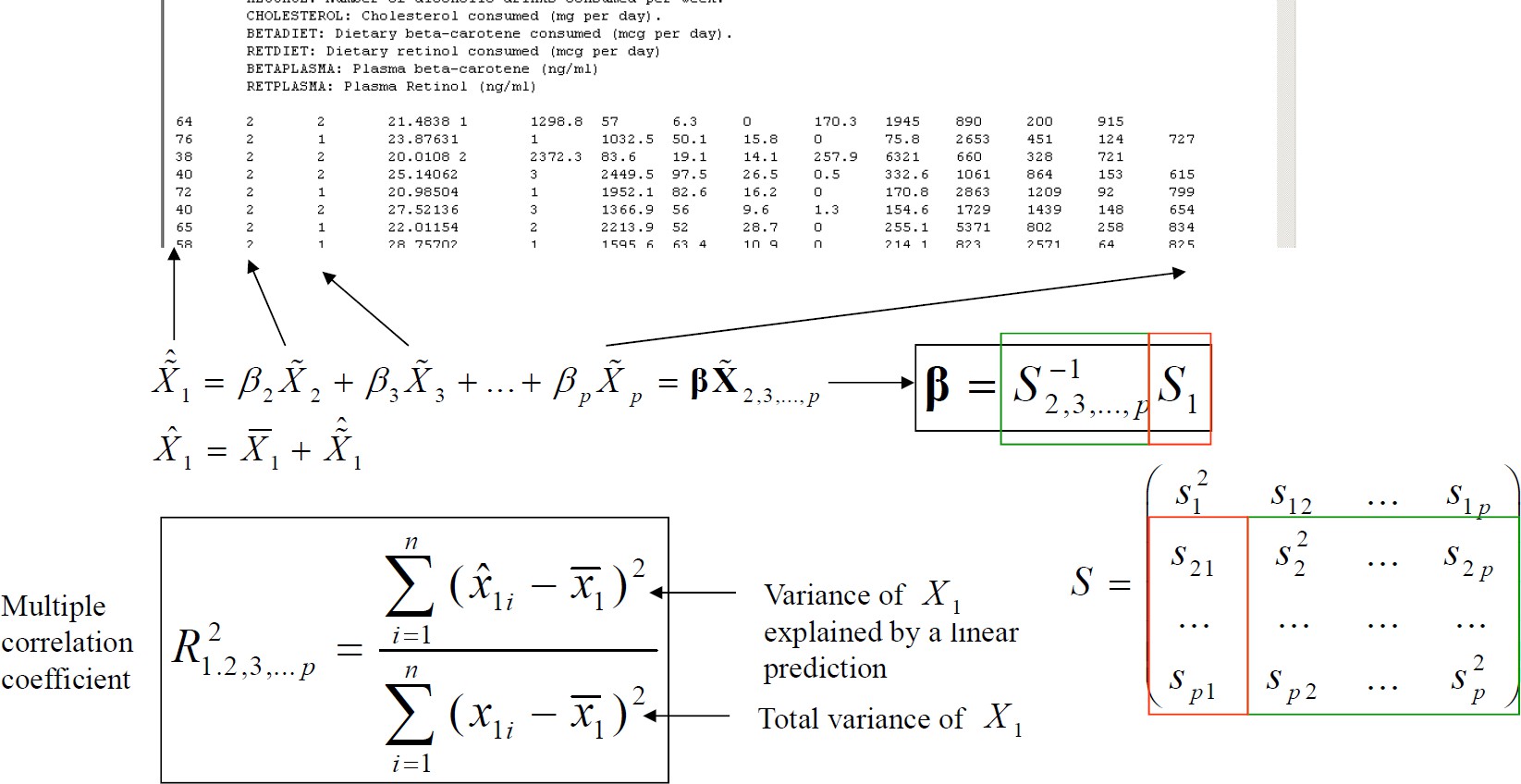
[Source: Multivariate Data Analysis Course. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Dependence bivariant



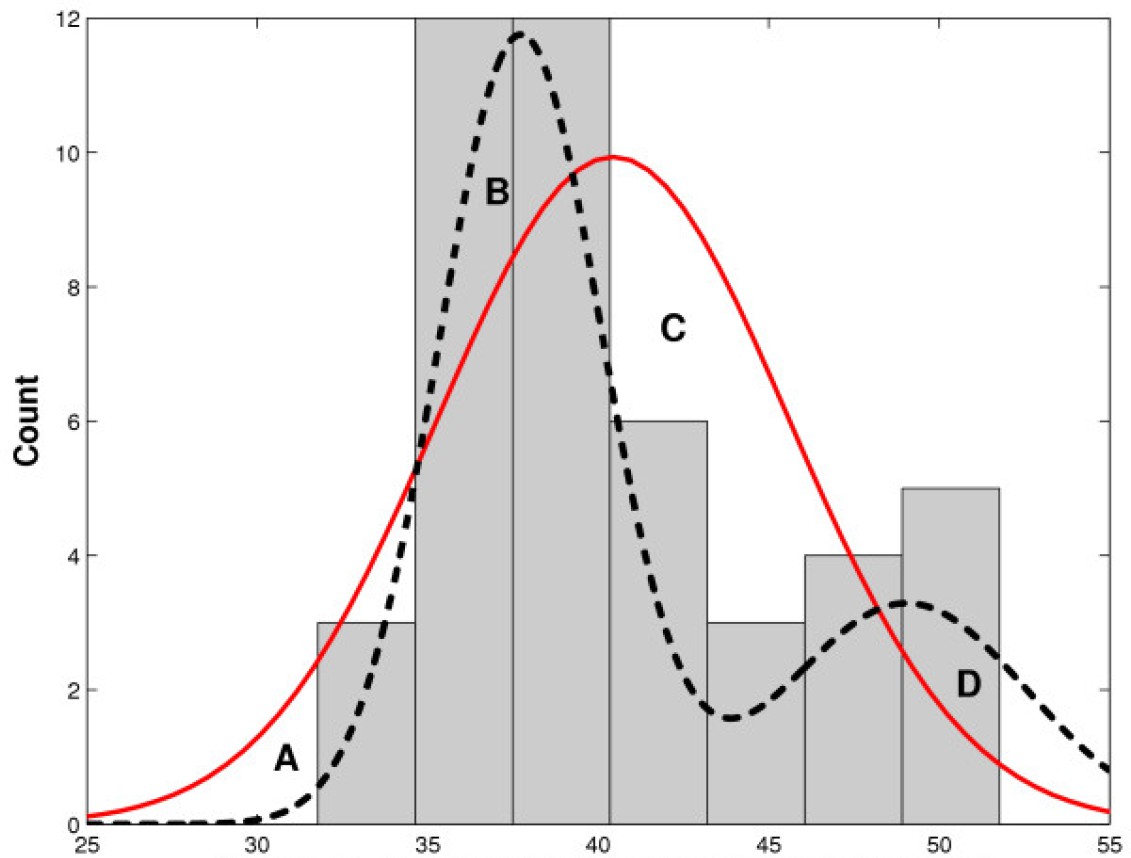
[Source: Course on Multivariate Data Analysis. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Dependence Multivariant



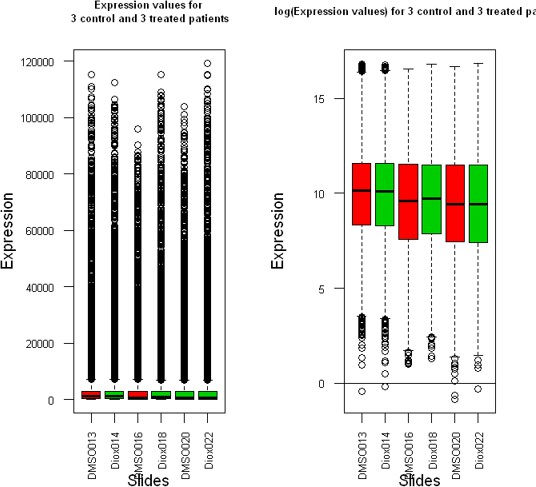
[Source: Multivariate Data Analysis Course. Carlos O. Sanchez](http://i2pc.es/coss/Docencia/ADAM/Notes/MultivariateAnalysisSlides.pdf)

# Data visualization

It is possible to visualize the variables one by one

 Histogram, Boxplts, Violin plots, Bar Charts

But the usual thing is to start with graphics in 2 or 3 dimensions

 Multiple Boxplot, Scatterplots, MosaicplotsGraphics are usually 2D or 3D

 Heatmaps

are very popular

18 / 45

# Visualizing Multiple Variables

To work with three or more variables, there are different alternatives

 Obviously, one is to start with the above and represent the variables by identifying their relationship or study it in pairs

 Another alternative is to work in small dimensions, i.e. to represent for example 2 to 2 or 3 to 3 some of the first components.

This is the approach discussed in the next section, devoted to dimension reduction.

19 / 45

# Data Exploration Example

This link [(Link to the workflow)](https://bioconductor.org/packages/release/bioc/vignettes/POMA/inst/doc/POMA-demo.html) contains a complete workflow of metabolomics data using Bioconductor's POMA package.



20 /

 The first steps contain aspects such as data exploration, transformations, and detection and imputation of missing values.

 Note the use of

one

SummarizedExperiment

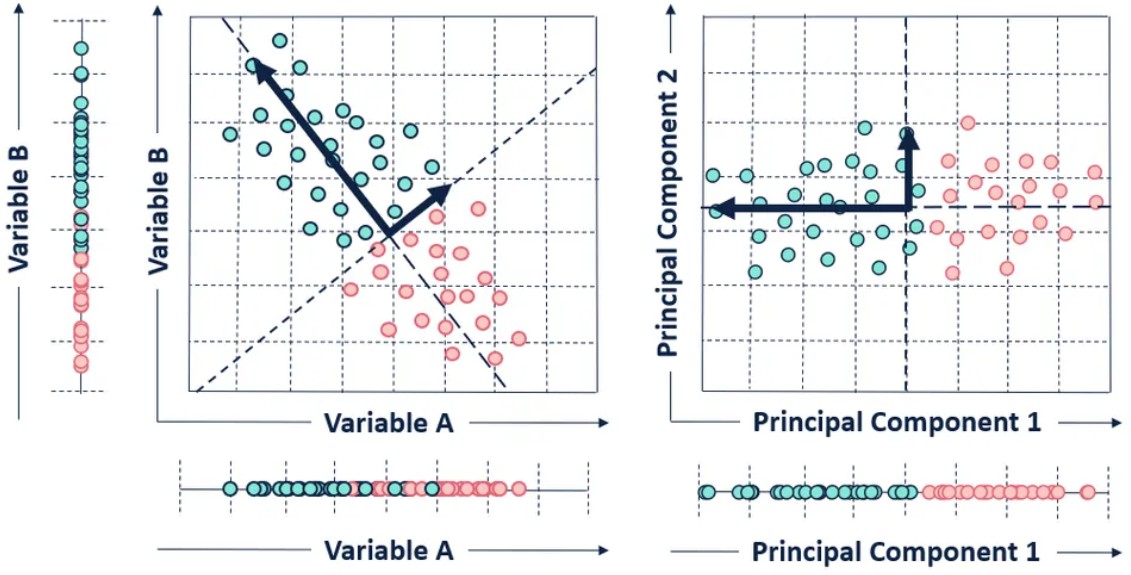
generalization of the ExpressionSet class, which was originally intended for microarrays.

45

# Reduction of dimension

21 / 45

# Principal Component Analysis

Given a ***KxN data***  matrix containing ***K*** variables (*probably* correlated measurements) in ***N*** samples (objects/individuals...)

The PCA carries out a transformation of the variables into ***K*** new components that

 They reflect the different sources of data variability, but

 are not correlated, i.e., *each component represents a different source of variability,*

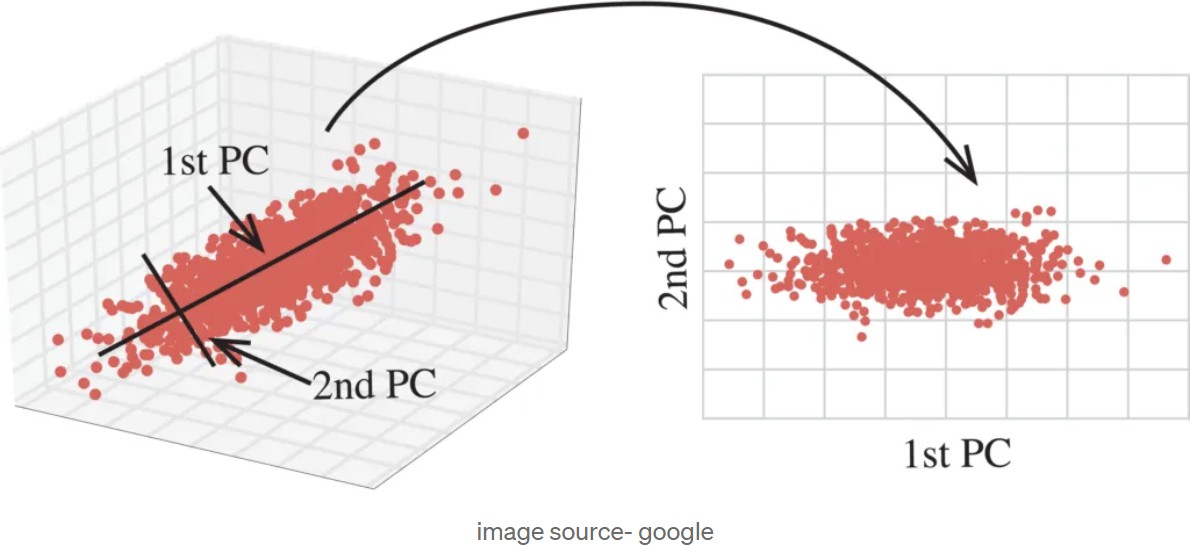
[Fountain](https://towardsdatascience.com/tidying-up-with-pca-an-introduction-to-principal-components-analysis-f876599af383)

# Designed to improve

These new components are constructed in such a way that *they have a decreasing explanatory capacity*: each component explains more than the next (less than the previous one).

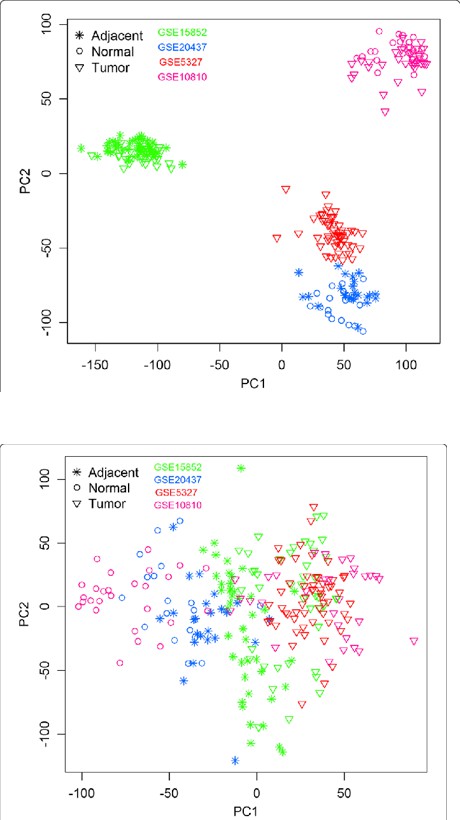
 This *can allow you* to keep the first components and discard the rest.

 That is, it is possible to use the values of the first principal components to obtain a representation of the data *in a reduced dimension*, i.e. smaller than the original



[In this example](https://medium.com/alliedoffsets/unveiling-the-power-of-pca-turbocharge-your-data-science-with-dimensionality-reduction-fa9d1fc6e510) we go from three to two dimensions with almost no loss of representation quality

# Revealing latent information

In general, it is considered that these transformations

 they are used to reveal the dominant patterns and main trends in the data

 while allowing natural groupings to be visualized

or erroneous such as those due to batch effect.

The image shows the graph of the first two components before (top) and after (bottom) of adjusting to remove a batch effect

# How does PCA work?

Let's assume a 2xN data matrix of two correlated variables.

Because the data is correlated, it is difficult to separate each source of variability

If K were much taller, it would be even harder.

To obtain the new bases, the mathematical problem of optimization is posed, consisting *of searching, sequentially, for a transformation of the original matrix in which the resulting vectors are orthogonal two to two*, that is,

 The new variables (new coordinate axes) are independent of each other.

 In this case, the matrix of variances will be diagonal, since the covariance of independent variables is zero.

# Calculating PCs

Under general conditions, it is seen that this will be obtained *by diagonalizing* the variance-covariance matrix of the centered data, a calculation that results in

 A vector of eigenvalues (VAPs) proportional to the variance of each variable in the new coordinates

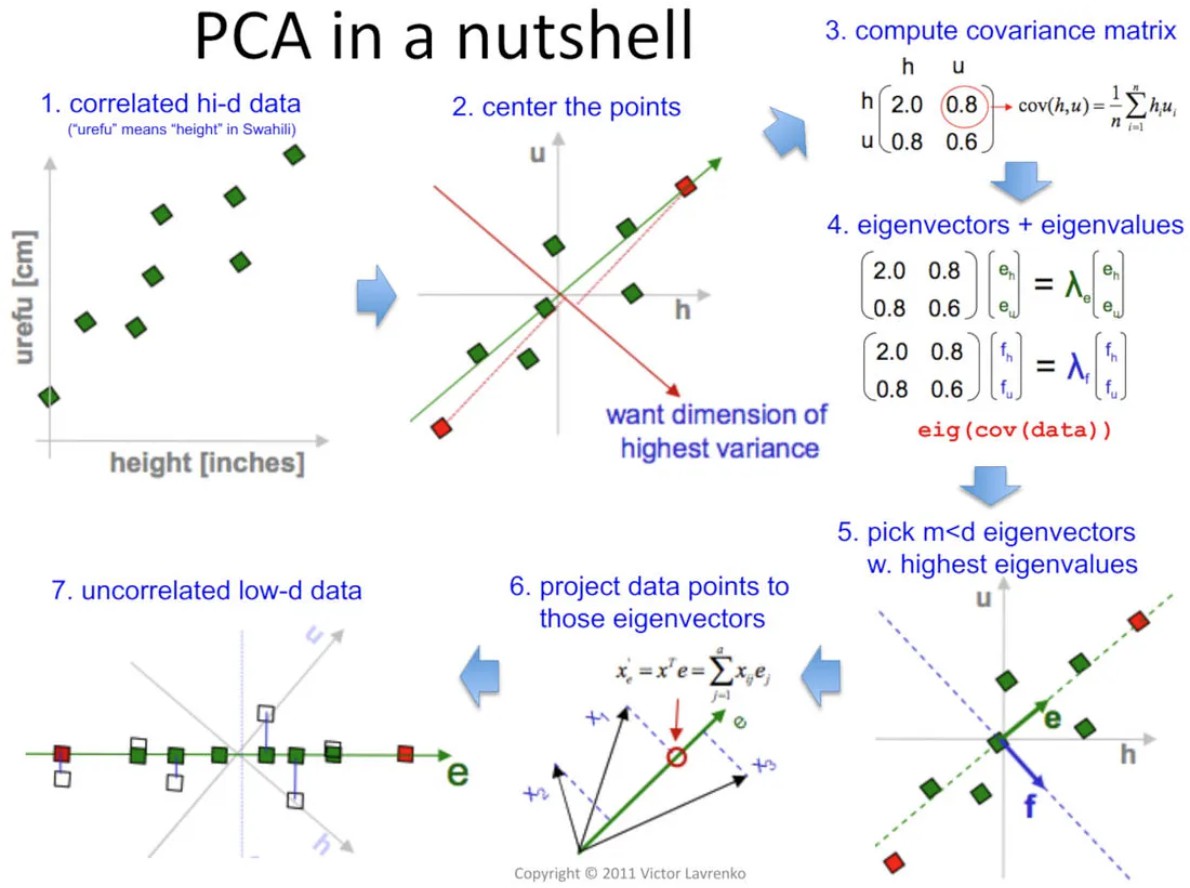
 A matrix of eigenvectors (VEPs) that are, precisely, the scores or

*scores* of the observations in these new coordinates.

Without going into the mathematical details, note that when transforming the matrix of variances diagonally, the covariances of the new variables become 0, that is, independent variables are obtained.

In addition, the eigenvalues (the diagonal of the new matrix) are ordered from highest to greatest, which corresponds to the idea that each one has a greater variability than the next, and therefore explains a greater percentage than the following ones.

# Obtaining the PCs



Example with R

The two chapters of the open text "[Applied Multivariate Statistics with R](https://uw.pressbooks.pub/appliedmultivariatestatistics/)" present, in very affordable way.

 [The diagonalization of an array ("EigenAnalysis")](https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/eigenanalysis/)

 [How this applies in the PCA](https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/pca/)

An example can also be seen in the first case of the compilation [Cases and examples](https://aspteaching.github.io/AMVCasos/) [of Multivariate Analysis with R](https://aspteaching.github.io/AMVCasos/)

 [PCA-1 Example: Searching for Latent Factors in Ecological Data](https://aspteaching.github.io/AMVCasos/#ejemplo-pca-1-b%C3%BAsqueda-de-factores-latentes-en-datos-ecol%C3%B3gicos1)

# Interpreting PCs

The first main component is

 A linear combination (CL) of the original variables

 That goes in the direction of greater variability in the data  Explains the *maximum amount of variation in the data*

The 2nd and successive CPs is also a CL of all the original variables, although with other coefficients calculated in such a way that:

 It aligns with the *next direction of greatest variability*, ***orthogonally*** to the previous PCs.

 Explain the *maximum amount of variation remaining*

The interpretation of the new components may be clear, corresponding to real biological dimensions or, sometimes, not at all.

 To facilitate this, programs that calculate the GWP also calculate the correlation between the new components and the original variables.

# Representation in the PCA space

Scores are the new coordinates in the orthogonal system defined by the PCs and have been obtained for each PC to explain the *maximum amount of* ***variation remaining*** *in the data*

This means that it is not necessary to use all PCs to visualize the data in this new coordinate system, so a visualization can be made in a lower dimension.]

 Often, taking the first PCs will explain a high percentage of variability.

 Usually only the first 2 or 3 are taken, but

It should always be verified that ***these few are enough***

 The graph called "scree plot" together with a criterion of "change of slope" or elbow, are usually useful.

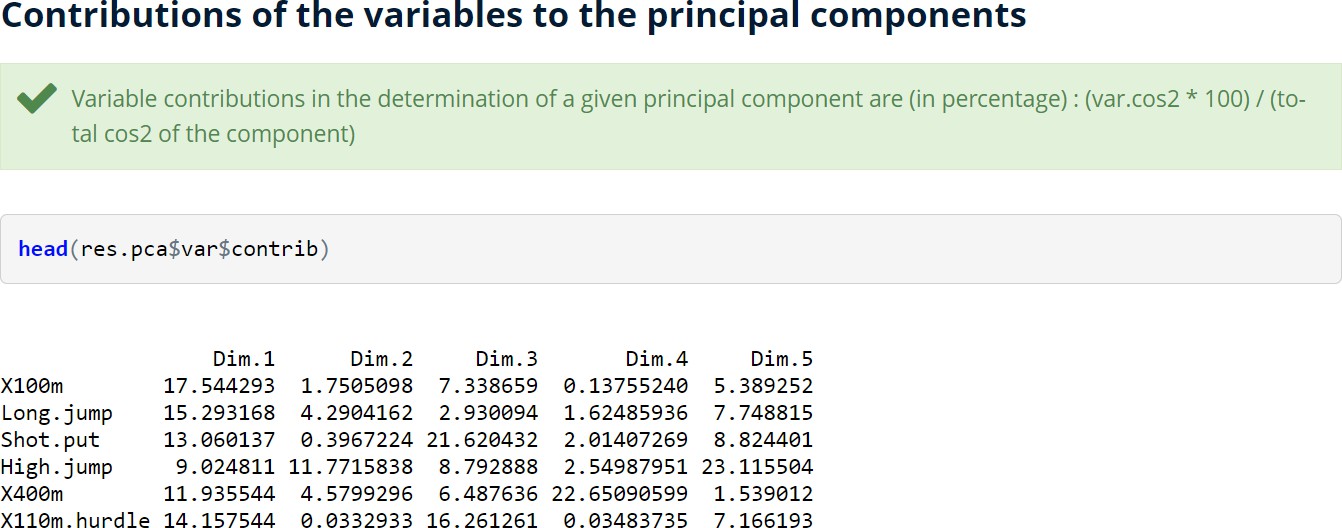
# Interpreting PCs

PCs can be interpreted by looking at which of the original variables contribute most to their variability

 The more a variable correlates with a CP, the greater its influence.

The magnitude of the contribution of each variable to a PC is its *load on that PC*

 Loads are the *cosines of the angle between the variables and the PCs*

31 / 45

# The PCA in Brief

The PCA performs a transformation on a new set of orthogonal coordinates that are called principal coordinates.

These new coordinates (PCs)

 explain the observed variability, with decreasing capacity (the greater that of the 1st CP, the smallest, that of the last CP).

 And, because they are orthogonal, each PC explains different aspects of the data.

PCA can be seen as a method of unraveling underlying trends or patterns in data

 The concept of *latent variable* is related to it

The PCA can provide an overview of the data to be revealed

 Dominant variables  Trends

 Patterns such as outliers, groups, clusters

32 / 45

# Discovering groups in your data

33 / 45

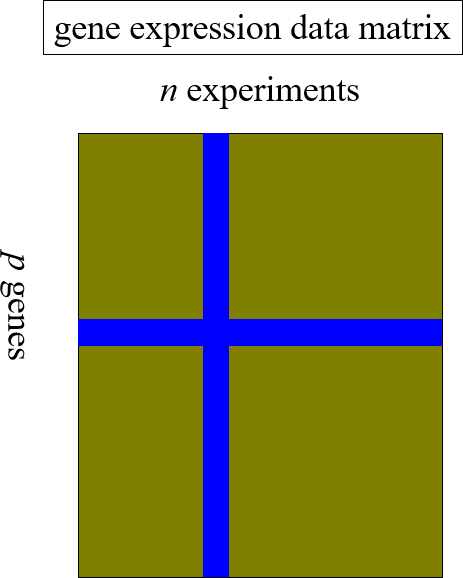
# Let's remember: the omics data

In an omics data experiment, once the data have been preprocessed, and it is common to have them summarized in an omics data matrix, for example of gene expression where

 Rows represent **genes or more generically *features***

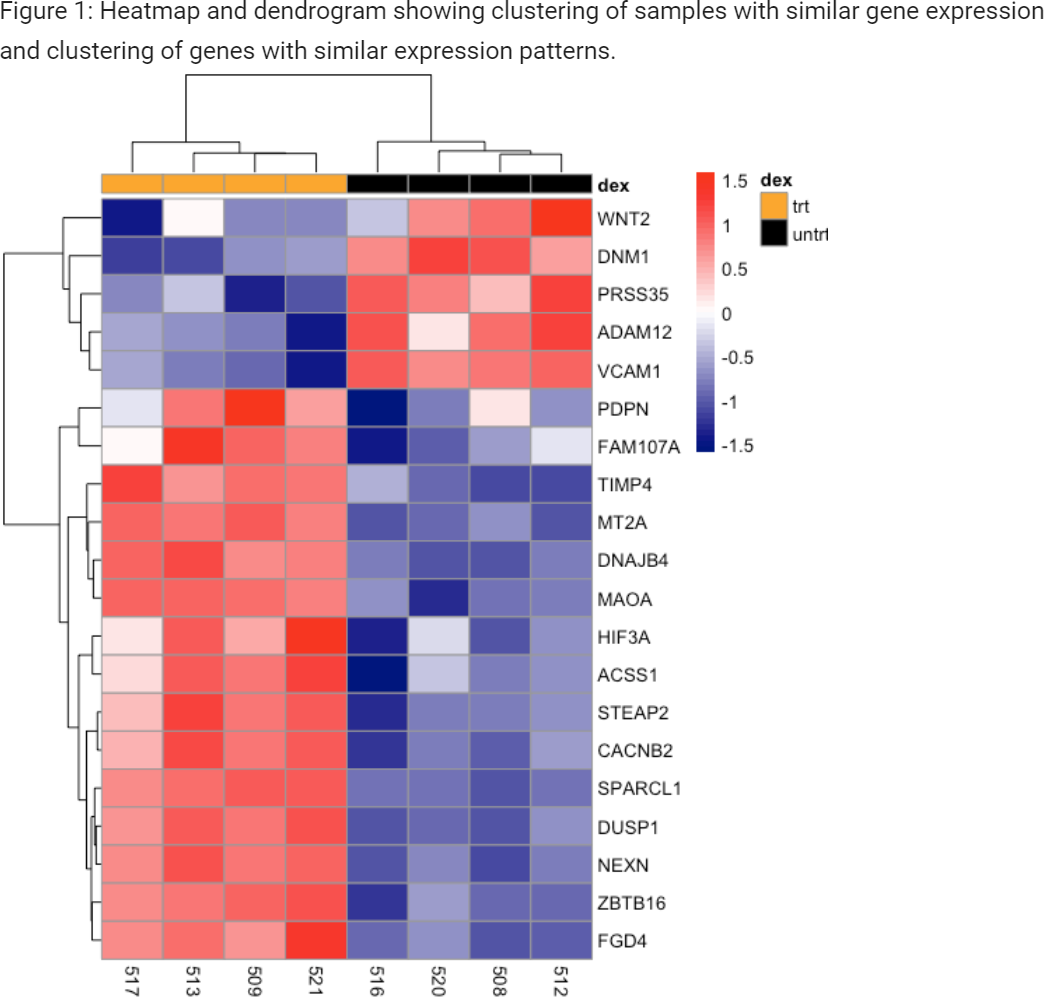
 Columns represent **individuals, samples, or observations**

The observations may or may not correspond to different experimental conditions (e.g. treatment/control) or to different groups of samples (e.g. male/female).



34 / 45

# Patterns in the data

When we talk about finding patterns in data, we can refer to identifying

 **Groups of related variables** that show *coordinated changes* between conditions.

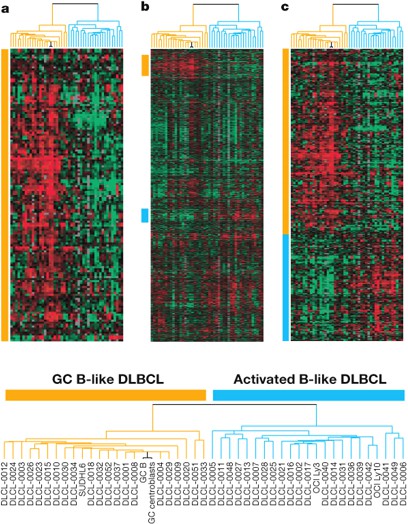
 This may suggest co-expression, co-regulation, etc

 **Sample groups** showing similar expression patterns

This may suggest the existence of previously unidentified (sub-)classes among the samples

35 / 45

# Benefits of grouping

It is common for the clustering process to lead to interpretable results.

It is especially useful

 To identify patterns in time or space.

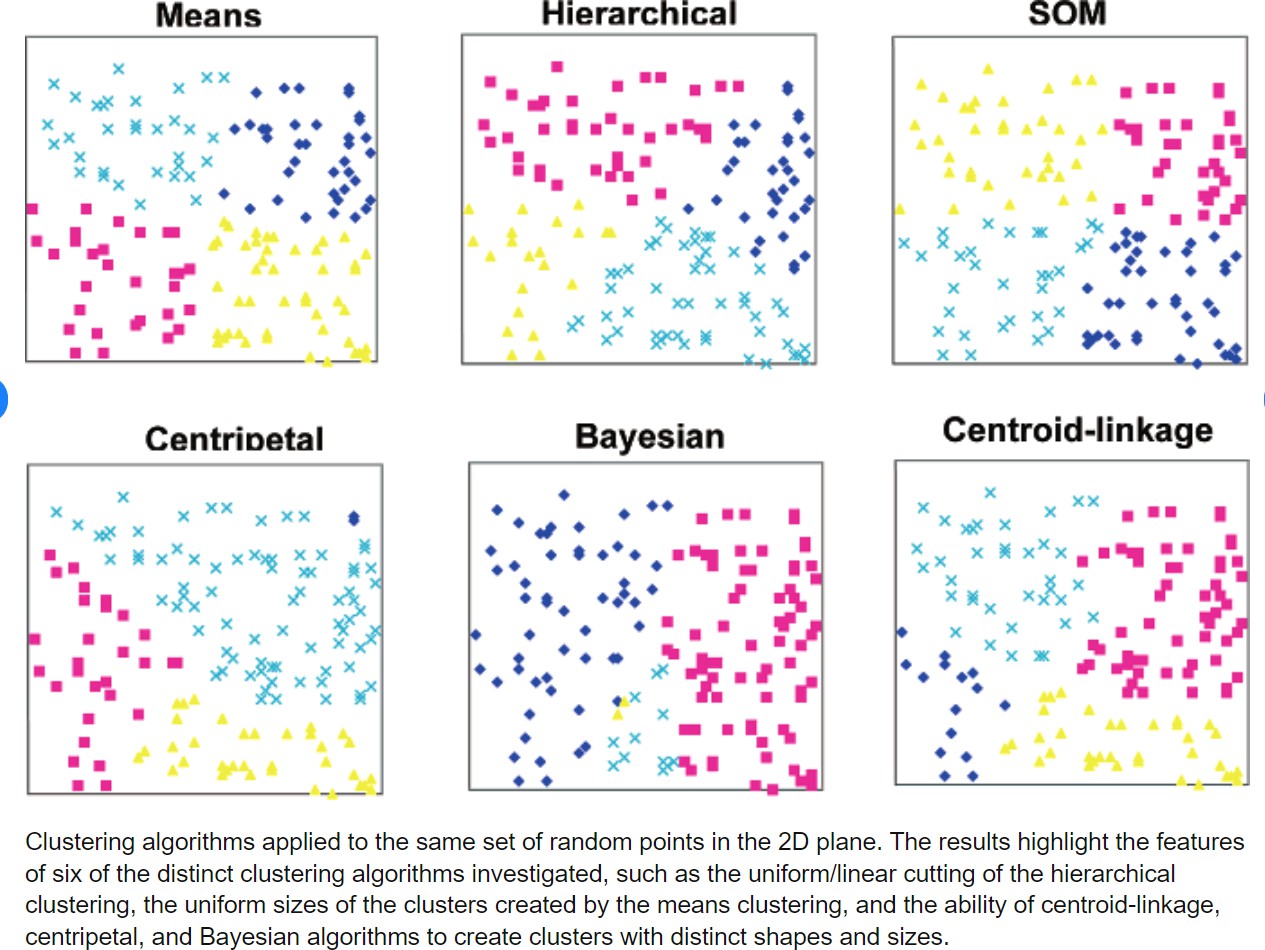
 When looking for new subclasses of cell samples (tumors, etc.).

Clustering reinforces the signal when averages are taken within groups of genes (Eisen).

[Fountain](https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r/Lesson5_intro_to_ggplot/)

36 / 45

Disadvantages of the *clustering*

It is an exploratory technique

Significance is usually not available .

Different approaches often produce different groupings

Hard to decide which is the best (or the "real")

It is difficult to avoid the temptation to select the grouping that best suits our hypotheses.

Any dataset can be grouped into clusters

It is difficult to decide whether the grouping is real or random.

37 / 45

# The Components of the Problem

Faced with any problem of this type, we must make different decisions

 What measure of similarity or dissimilarity?

 Which grouping algorithm to use?

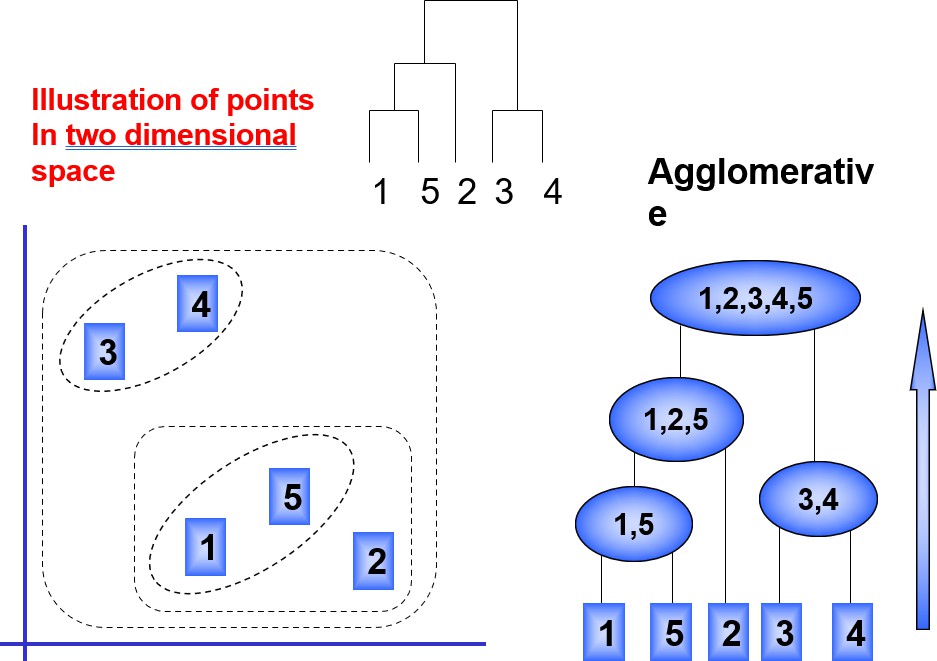
We often start with basic algorithms for And move on to more elaborate ones

 How will the number of groups be decided?

 Which individuals and variables to use

It is usually interesting to filter to eliminate redundant and uninformative variables, as well as defective data (outliers or with a lot of missings)

# Cluster hierarchical

Hierarchical grouping methods produce a tree or dendrogram.

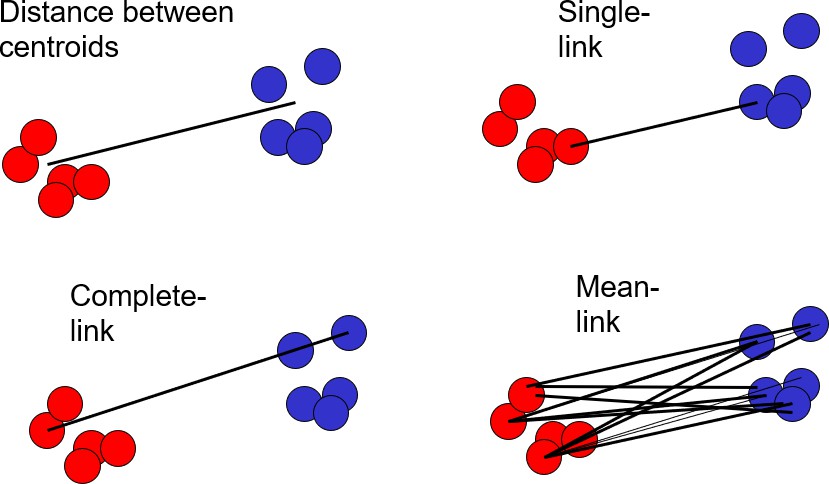
They avoid specifying how many clusters are appropriate by providing a partition for each *k* obtained from cutting the tree at some level.

The tree can be built in two different ways

 from bottom to top: agglomerated grouping;

 From top to bottom: dividing grouping.

# Distances for hierarchical grouping

They are based on the pairwise distances between cluster members

 **Average link**: average of the differences between peers

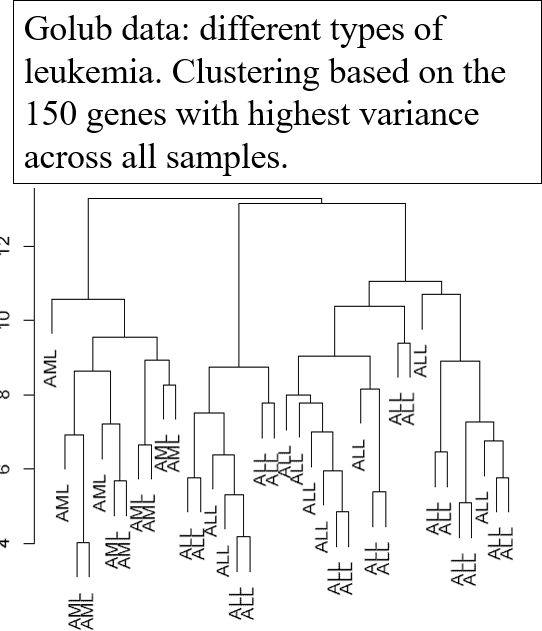
 **Single link**: minimum of dissimilarities in pairs.

 **Complete link**: Maximum dissimilarities per pair.

 **Distance Between Centroids**

The distance method chosen influences the results.

# Example of hierarchical grouping

This example illustrates some characteristics of hierarchical clustering

 The height of a node in the dendrogram represents the distance of the two secondary clusters.

 There is loss of information: n objects have n(n-1)/2 distances in pairs, the tree has n-1 internal nodes.

 The order of the leaves is not defined solely by the dendrogram: 2n-2 possible options.

# Summary: Clustering methods

They provide powerful and flexible methods for discovering groups in data and therefore

They are a powerful resource for exploratory data analysis.

The nature of the problem they solve makes them prone to ambiguous outcomes (what do they discover if there are no classes to discover?)

Not to mention that changes in methods and parameters are accompanied by different results.

The conclusion is the usual one in these matters: We must use these methods carefully and always with a good understanding of how they operate, validating whenever possible the results they provide us.

# References and Resources

43 / 45

# External References and Resources

[STHDA Statistical Tools for High Throughput Data Analysis](http://www.sthda.com/english/)

 This website contains a multitude of articles on statistics and machine learning grouped by topic. In fact, the author uses it to promote his courses and books, but most of the materials are good and free

 Specifically, we highlight two series of articles on the topics that concern us. To introduce yourself to the subject, the first chapters of each series are enough, but all of them are relevant to your training.

[Cluster Analysis in R. A practical guide](https://www.datanovia.com/en/blog/cluster-analysis-in-r-practical-guide/)

[Principal components methods in R](http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/)

# References and own resources

[Cases and examples of multivariate analysis with R](https://aspteaching.github.io/AMVCasos/)

Notes on Multivariate Analysis for the undergraduate subject "Design of Experiments and Data Analysis" of the Faculty of Biology (UB)

 Principal Component Analysis

 Cluster analysis

Multivariate analysis materials for the Multivariate Analysis subject of the UOC-UB Master's Degree

 Summaries of Linear Algebra