



UNIVERSITÀ
DI TORINO

Dipartimento
Chimica



DataBloom
let your data flourish

Leveraging Machine Learning with SpectrApp: An Open-Source Solution for Forensic Data Analysis (W2.06.01)

Eugenio Alladio
eugenio.alladio@unito.it

Alberto Mazzoleni
alberto.mazzoleni@unito.it

Giovanni Solarino
giovanni.solarino@unito.it



THE EUROPEAN ACADEMY OF FORENSIC SCIENCE

Conference material



<https://github.com/Chemometrics-UniToto/SpectrApp-EAfs2025>



Time
WORKSHOP for

Workshop outline

- Introduction and familiarizing participants with SpectrApp
- Multivariate Data Analysis and chemometric strategies
- Graphical approaches and data visualization strategies
- Supervised classification modeling
- Supervised Regression Modeling
- Questions and answers/discussion

[https://github.com/Chemometrics-
UniTo/SpectrApp-EAFS2025](https://github.com/Chemometrics-UniTo/SpectrApp-EAFS2025)

*Time for
WORKSHOP*

Statistics – ... now we're here

BALLISTICS



How Good a Match is It? Putting Statistics into Forensic Firearms Identification

February 08, 2018

On February 14, 1929, gunmen working for Al Capone disguised themselves as police officers, entered the warehouse of a competing gang, and shot seven of their rivals dead. The St. Valentine's Day Massacre is famous not only in the annals of gangland history, but also the history of forensic science. Capone denied involvement, but an early forensic scientist named Calvin Goddard linked bullets from the crime scene to Tommy guns found at the home of one of Capone's men. Although the case never made it to trial—and Capone's involvement was never proved in a court of law—media coverage introduced millions of readers to Goddard and his strange-looking microscope.

That microscope had a split screen that allowed Goddard to compare bullets or



MEDIA CONTACT

 **Rich Press**
richard.press@nist.gov
(301) 975-0501

ORGANIZATIONS

Physical Measurement Laboratory
Sensor Science Division
Surface and Interface Metrology Group
Laboratory Programs
Special Programs Office

Statistics – ... now we're here

EAFS DUBLIN 2025



DataBloom
let your data flourish

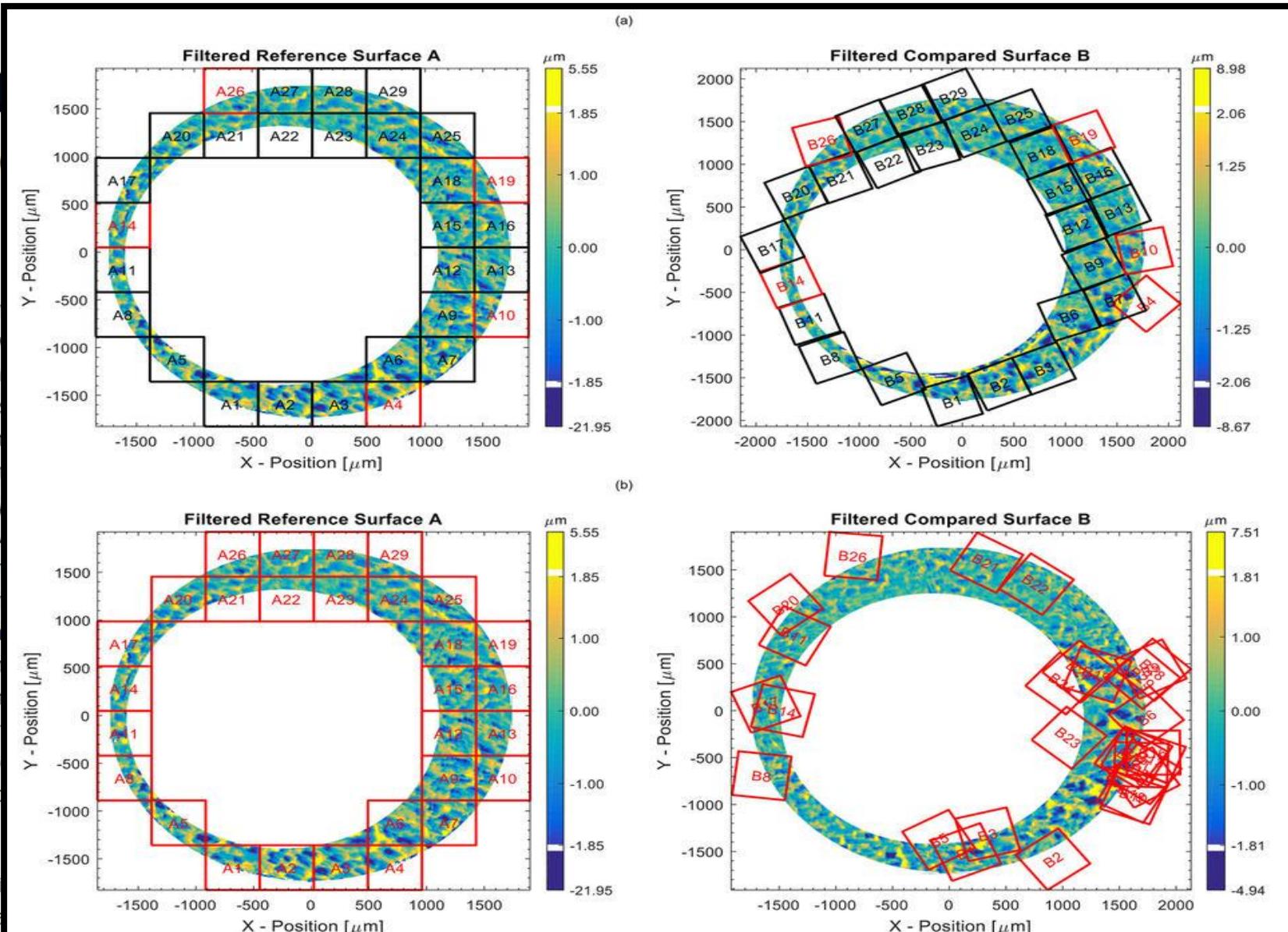
How Good Firearms I

February 08, 2018

On February 14, 1929, gunn Capone disguised themself officers, entered the wareh competing gang, and shot rivals dead. The St. Valentine is famous not only in the ar history, but also the history science. Capone denied inv early forensic scientist nam Goddard linked bullets fro to Tommy guns found at th Capone's men. Although th made it to trial—and Capon was never proved in a cour coverage introduced millio Goddard and his strange-l microscope.

That microscope had a spl allowed Goddard to compa

BALLISTICS



Statistics – ... now we're here

EAFS DUBLIN 2025



DataBloom
let your data flourish

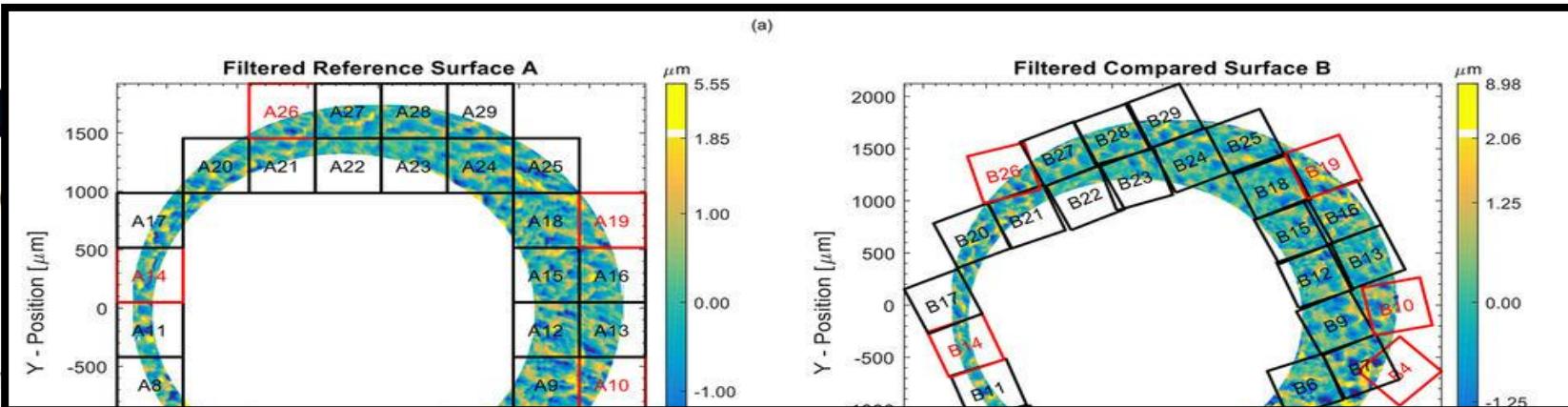
How Good Firearms I

February 08, 2018

On February 14, 1929, gunn Capone disguised themself officers, entered the wareh competing gang, and shot rivals dead. The St. Valentine is famous not only in the ar history, but also the history science. Capone denied inv early forensic scientist nam Goddard linked bullets fro to Tommy guns found at th Capone's men. Although th made it to trial—and Capon was never proved in a cour coverage introduced millio Goddard and his strange-look microscope.

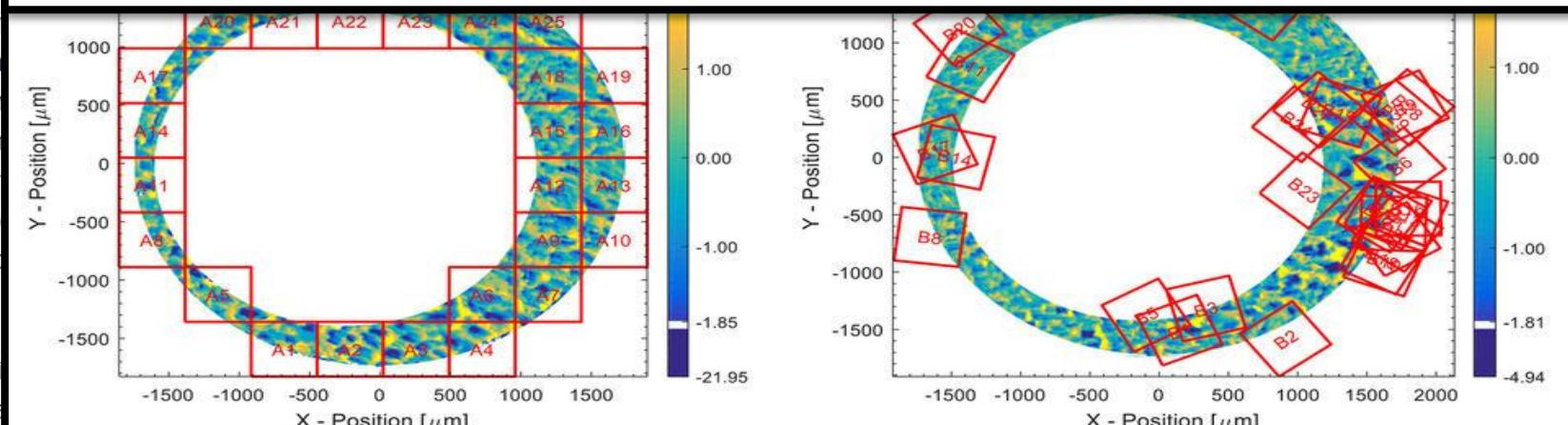
That microscope had a spl allowed Goddard to compa

BALLISTICS

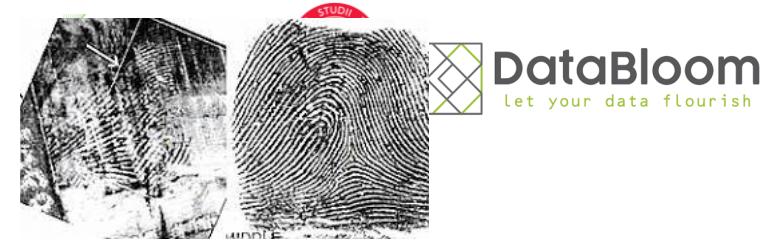


A Better Way to Testify

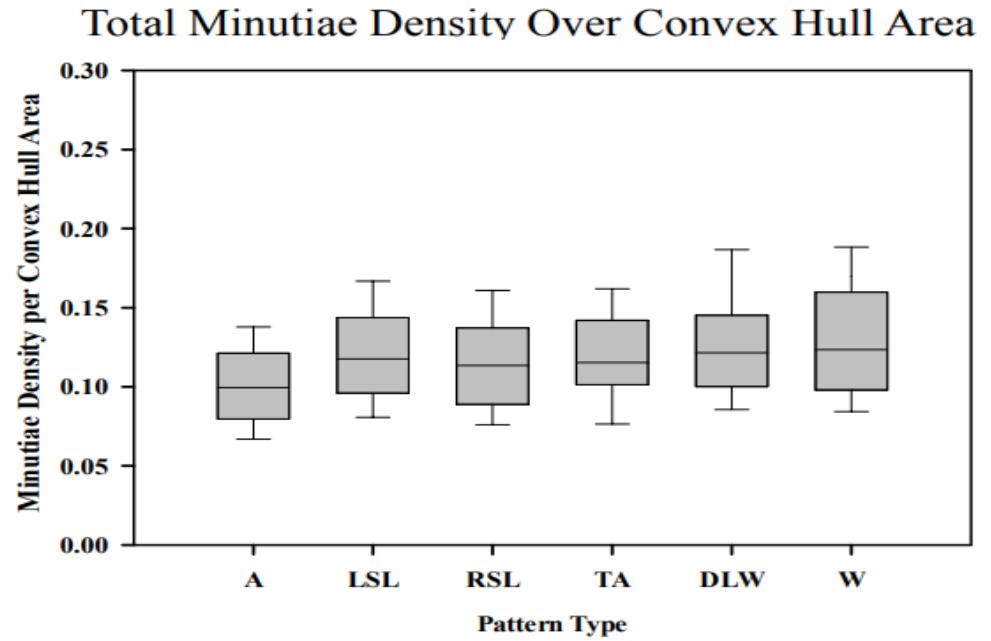
Using well-established statistical methods, the authors built a model for estimating the likelihood that random effects would cause a false positive match. Using this method, a firearms expert would be able to testify about how closely the two cartridges match based on the number of matching cells, and also the probability of a random match, similar to the way forensic experts testify about DNA.



Statistics – ... now we're here



DACTYLOSCOPY



ANOVA: Total minutiae density (no./mm²) over convex hull area separated by pattern type

	df	SSE	MSE	F value	p-value
Pattern Type	5	0.0921	0.018425	14.8	0.0000000000000417
Residuals	1194	1.4867	0.001245		

P-values: $\alpha=0.05$

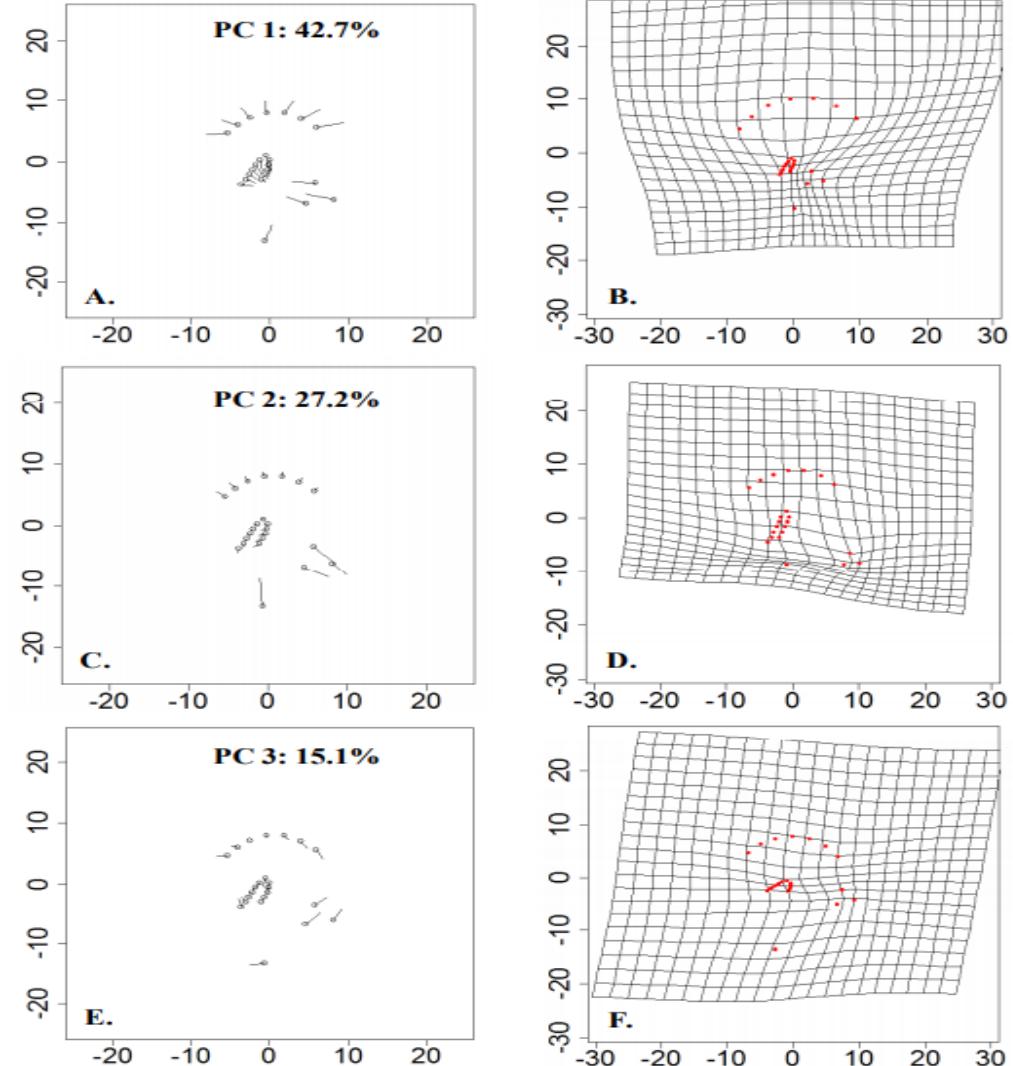
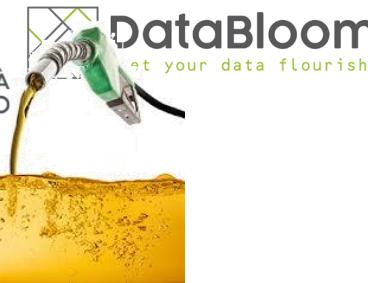


Figure 3-7. Principal component analysis and deformation modeling for left slant loops

Statistics – ... now we're here

EAFS DUBLIN
2025



CHEMISTRY

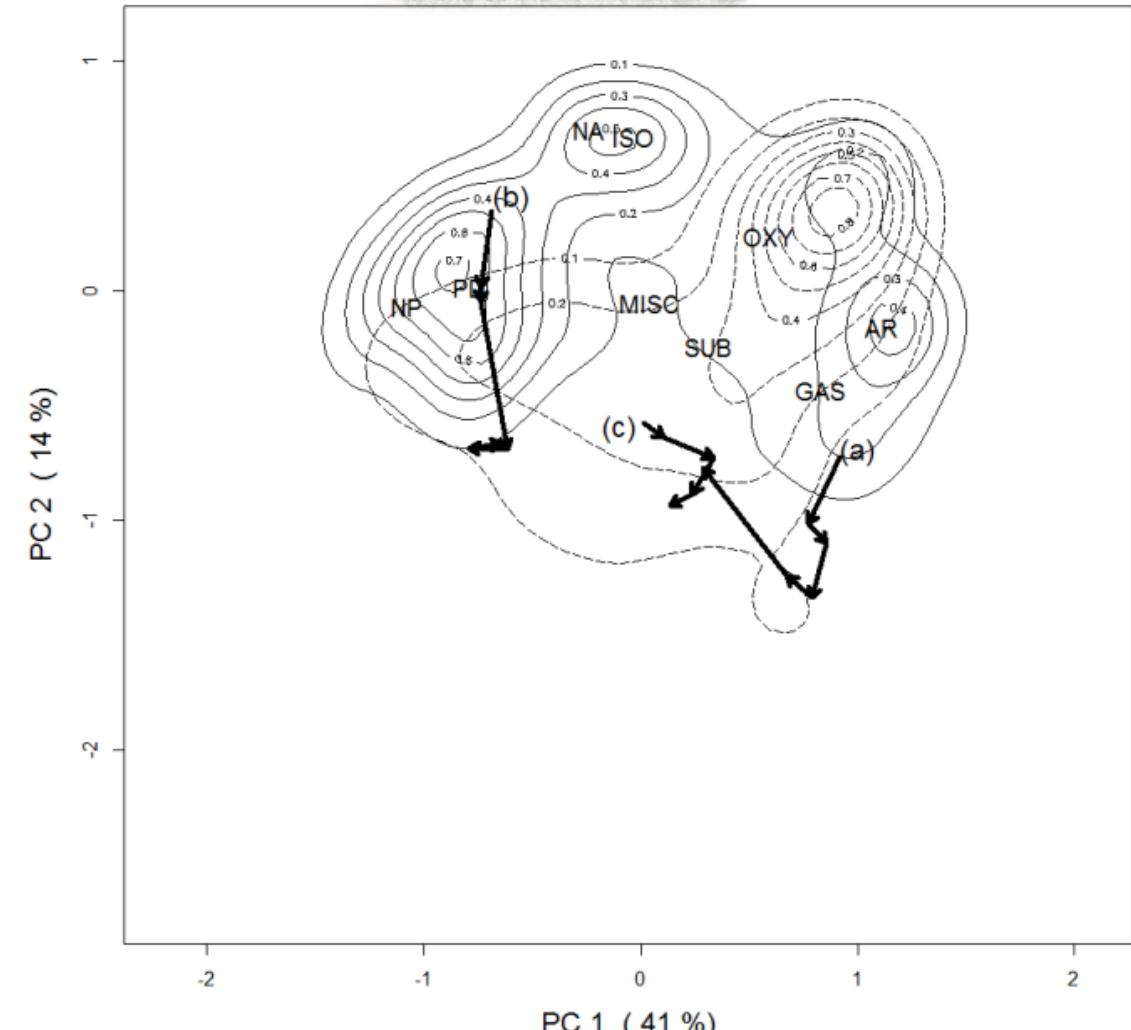
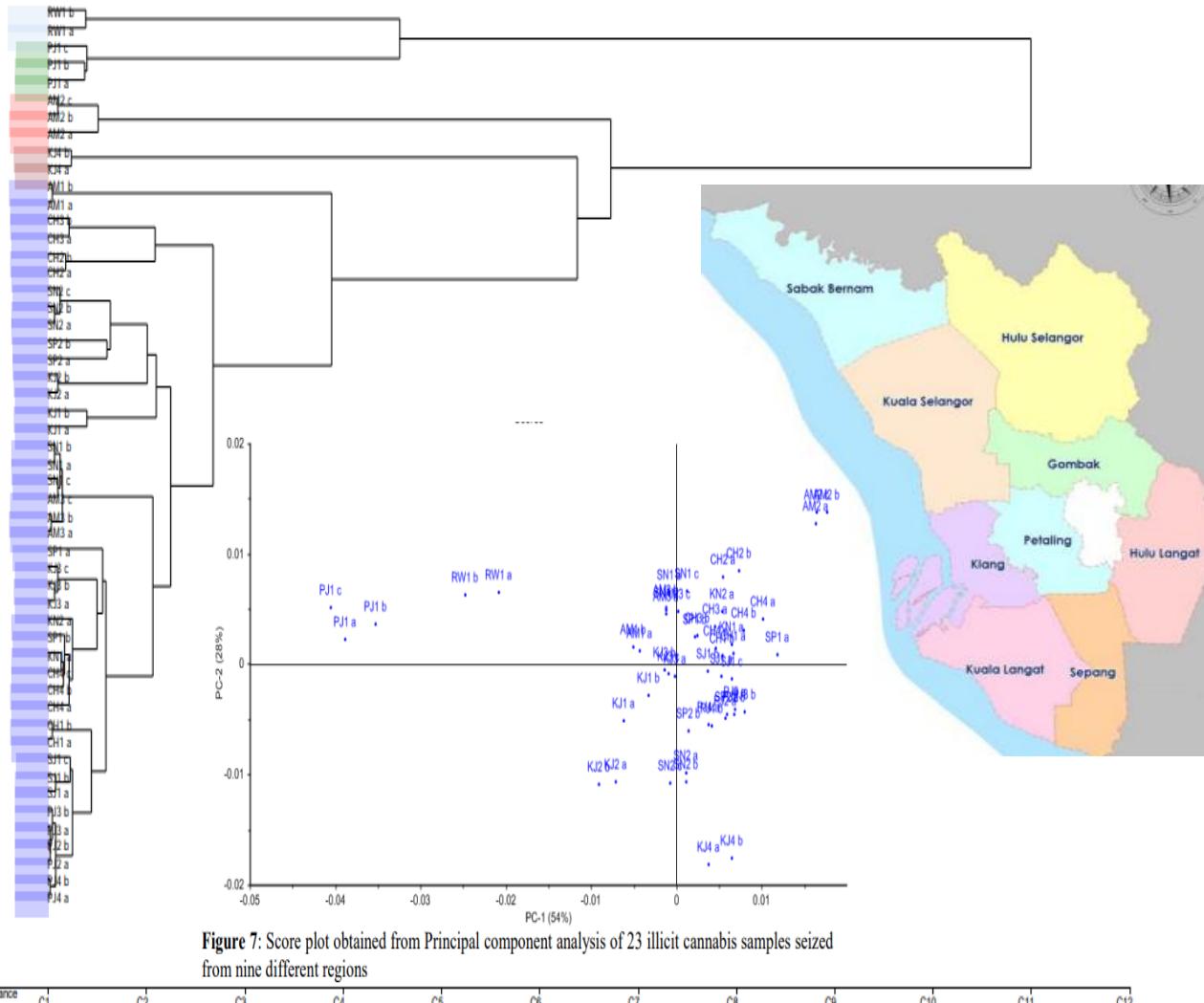
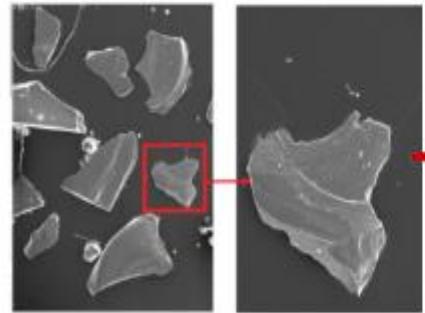


Figure 5: Dendrogram obtained from Cluster analysis of 23 illicit cannabis samples seized from nine different regions

Statistics – ... now we're here

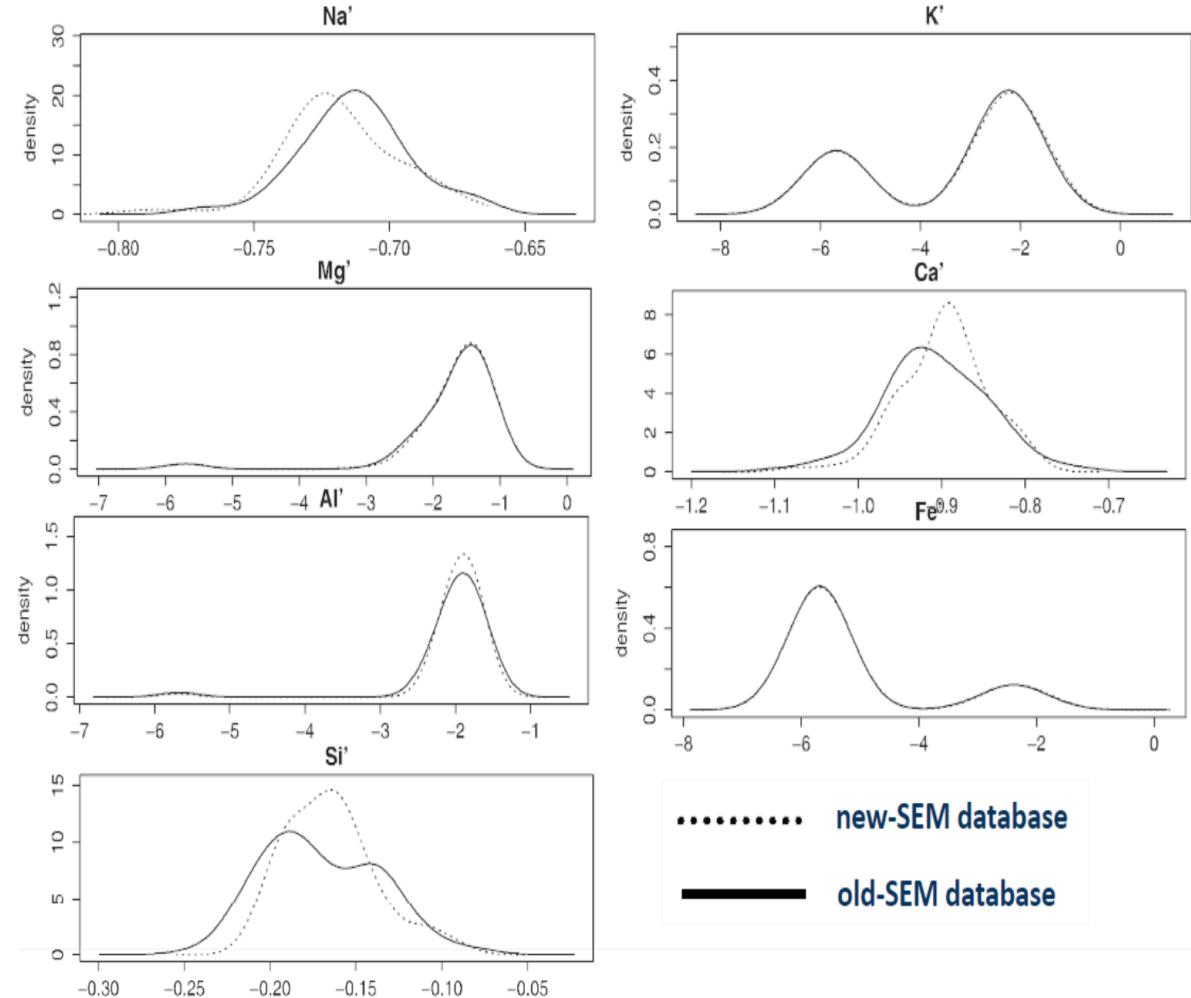
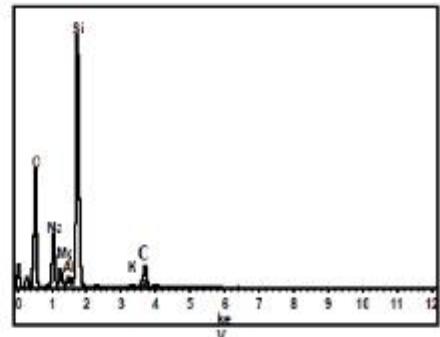
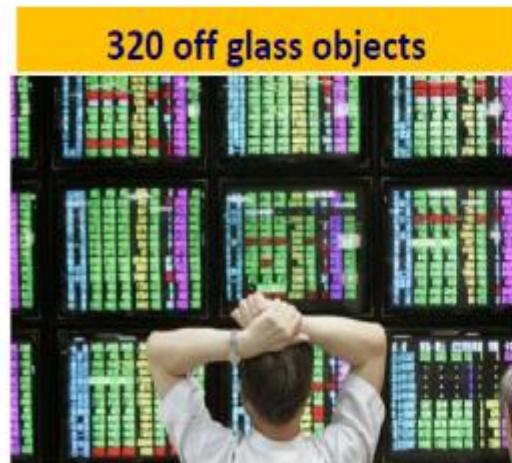
MICRO-ANALYSIS & SPECTROSCOPY



Non-Embedded procedure



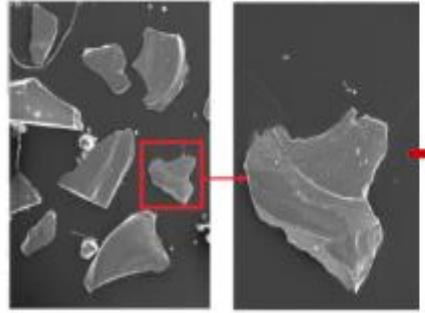
Jeol JSM-5800 SEM with a
Link ISIS 300 EDX (Oxford
Instrument Ltd.)



..... new-SEM database
— old-SEM database

Statistics – ... now we're here

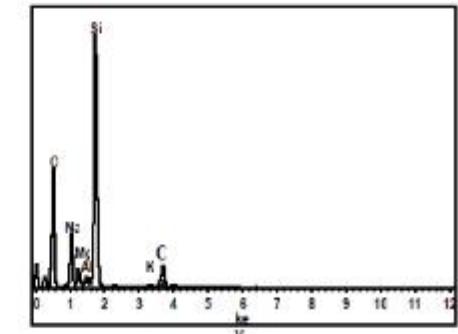
MICRO-ANALYSIS & SPECTROSCOPY



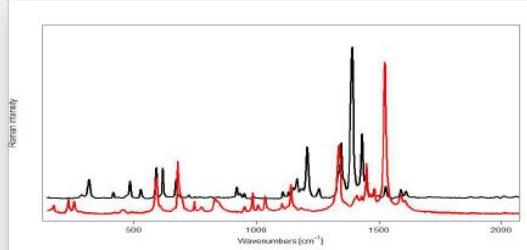
Non-Embedded procedure



Jeol JSM-5800 SEM with a
Link ISIS 300 EDX (Oxford
Instrument Ltd.)

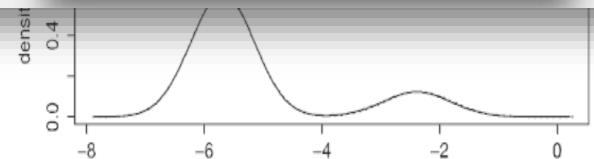
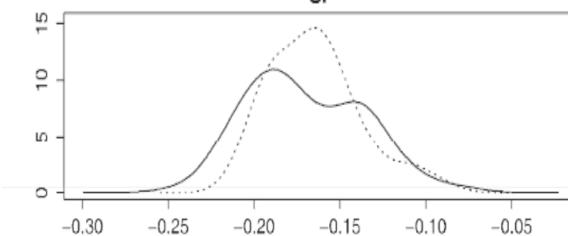
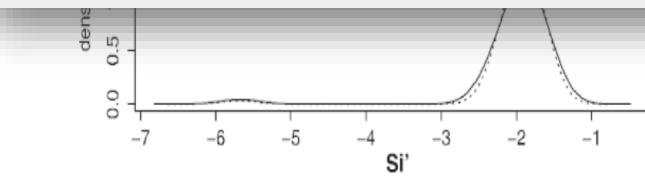
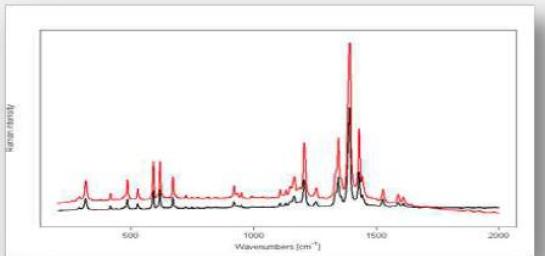


subjective



objective

$$LR = \frac{\Pr(E | H_p)}{\Pr(E | H_d)}$$



..... new-SEM database

— old-SEM database

Statistics – ... now we're here

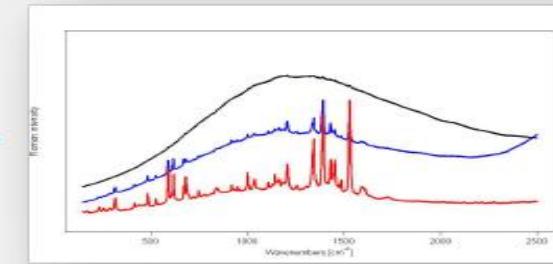
MICRO-ANALYSIS & SPECTROSCOPY

Analysis of automotive paints – Raman data

Problem with fluorescence



Photobleaching
514.5 nm



$$LR = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

Research article **LR-BA**

Received: 16 July 2014
Revised: 26 March 2015
Accepted: 27 April 2015
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jrs.4719

Application of a likelihood ratio approach in solving a comparison problem of Raman spectra recorded for blue auton

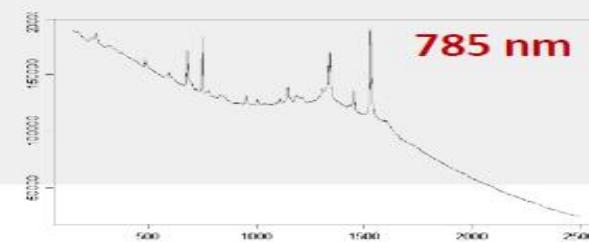
Aleksandra Michalska,^a Agnieszka Małyszko^{a,c*} and Grzegorz Zadora^{a,b,c*}

^aAnal. Bioanal. Chem. (2015) 407:3457–336
DOI: 10.1002/abc.2015-032-0338-9
^bRESEARCH PAPER

LR-DWT

Interpretation of FTIR spectra of polymers and Raman spectra of car paints by means of likelihood ratio approach supported by wavelet transform for reducing data dimensionality

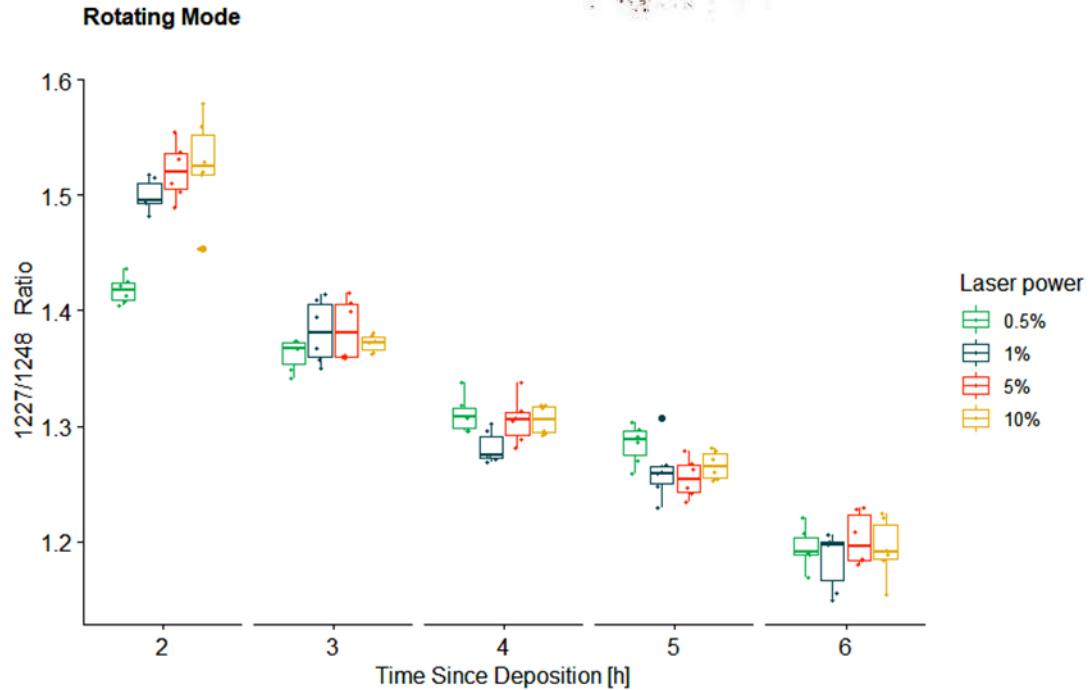
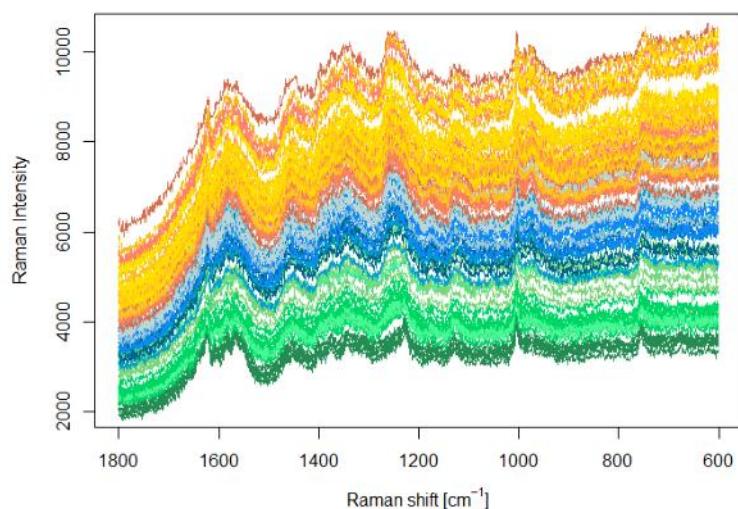
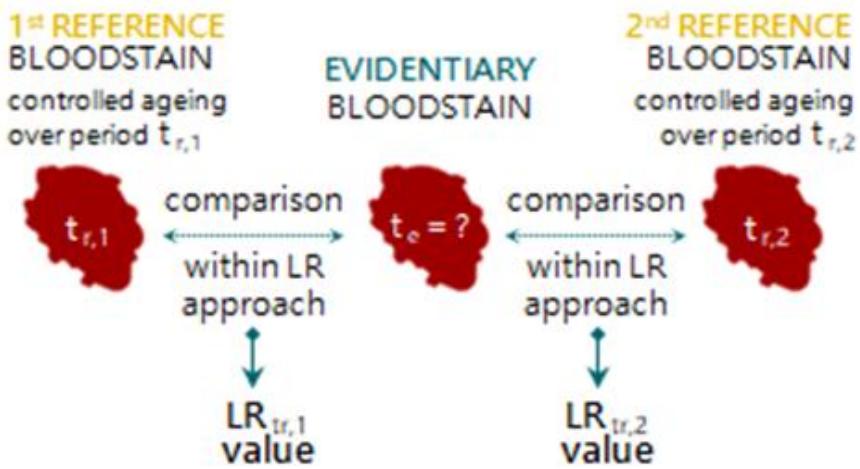
Agnieszka Martyna · Aleksandra Michalska ·
Grzegorz Zadora



Statistics – ... now we're here



BLOODSTAIN PATTERN ANALYSIS



Toward a novel framework for bloodstains dating by Raman spectroscopy: how to avoid sample photodamage and subsampling errors

Alicja Menzyk^{1*}, Alessandro Damin², Agnieszka Martyna¹, Eugenio Alladio^{2,3}, Marco Vincenti^{2,3}, Gianmario Martra², Grzegorz Zadora^{1,4}

¹ Institute of Chemistry, University of Silesia in Katowice, Szkolna 9, 40-007 Katowice, Poland

² Department of Chemistry, University of Torino, Via P. Giuria 7, 10125 Torino, Italy

³ Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Regione Gonzole 10/1, 10043 Orbassano, Torino, Italy

⁴ Institute of Forensic Research, Westerplatte 9, 31-033 Krakow, Poland

Let's talk about "chemometrics"

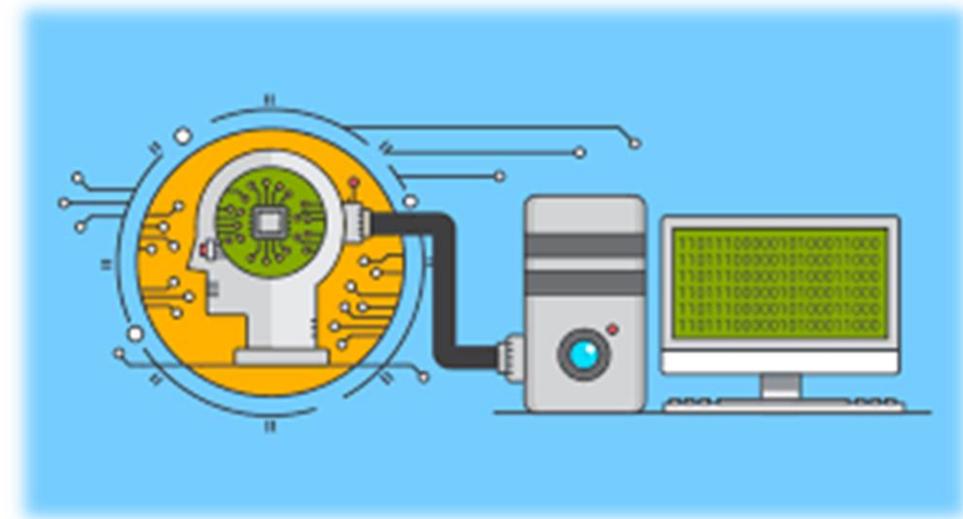


Data Mining in:
Hyperspectral and digital imaging
Chromatography
Forensic Sciences
Pharmaceutical monitoring
Food production
Etc...

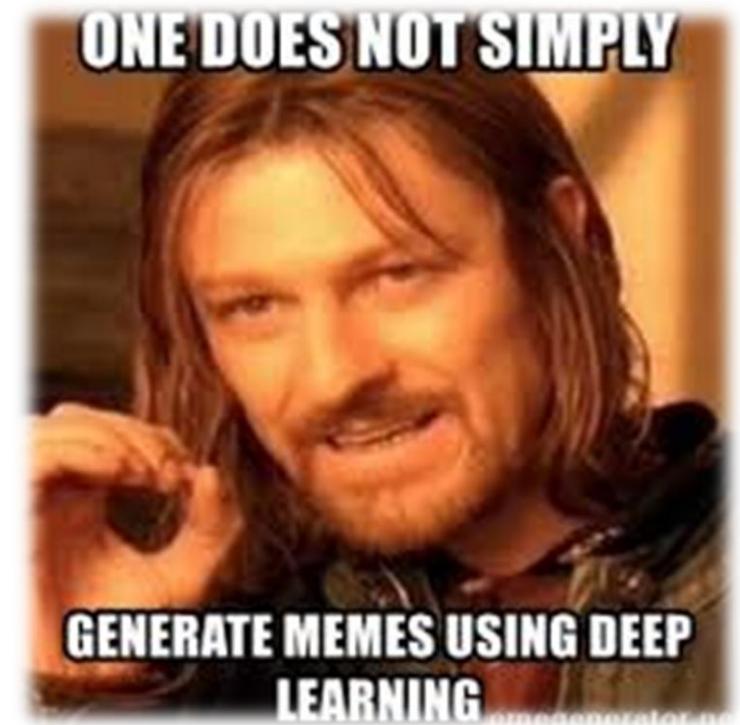


Machine Learning in:

- Hyperspectral and digital imaging
- Chromatography
- Forensic Sciences
- Pharmaceutical monitoring
- Food production
- Etc...

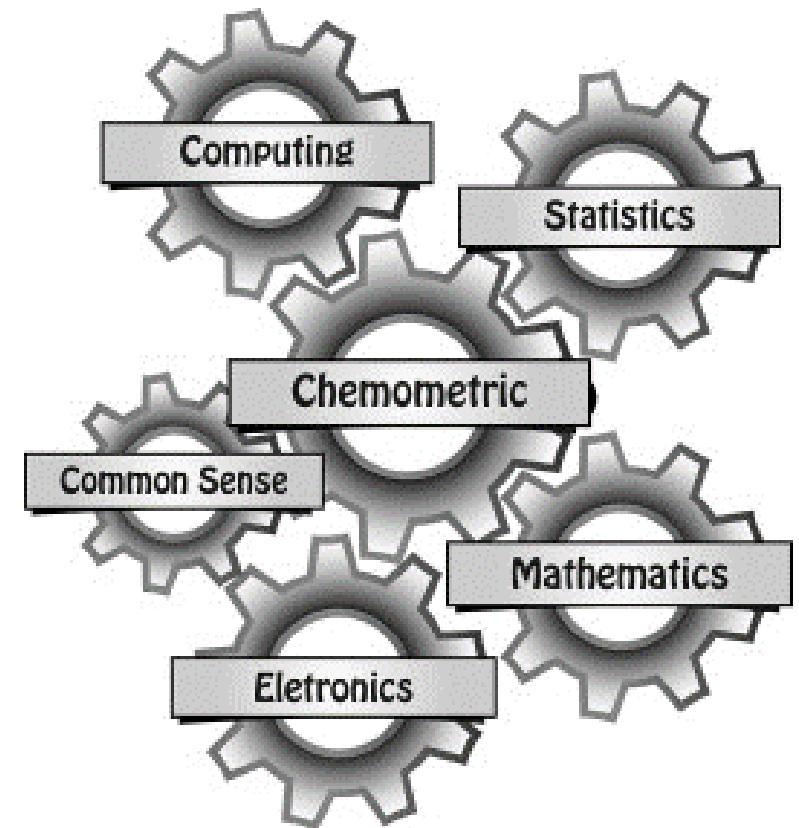


Deep Learning in:
Hyperspectral and digital imaging
Chromatography
Forensic Sciences
Pharmaceutical monitoring
Food production
Etc...



CHEMOMETRICS (♥) in:

Hyperspectral and digital imaging
Chromatography
Forensic Sciences
Pharmaceutical monitoring
Food production
Etc...



Data Mining in:
Machine learning in:
Deep learning in:
Hyperspectral and digital imaging
Hyperspectral and digital imaging
Chromatography and digital imaging
Hyperspectral and digital imaging
Chromatography
Chromatography
Forensic chromatography
Forensic Sciences
Pharmaceutical Sciences
Food production monitoring
Food production monitoring
Food production
Food production
Etc...Food production
Etc...
Etc...

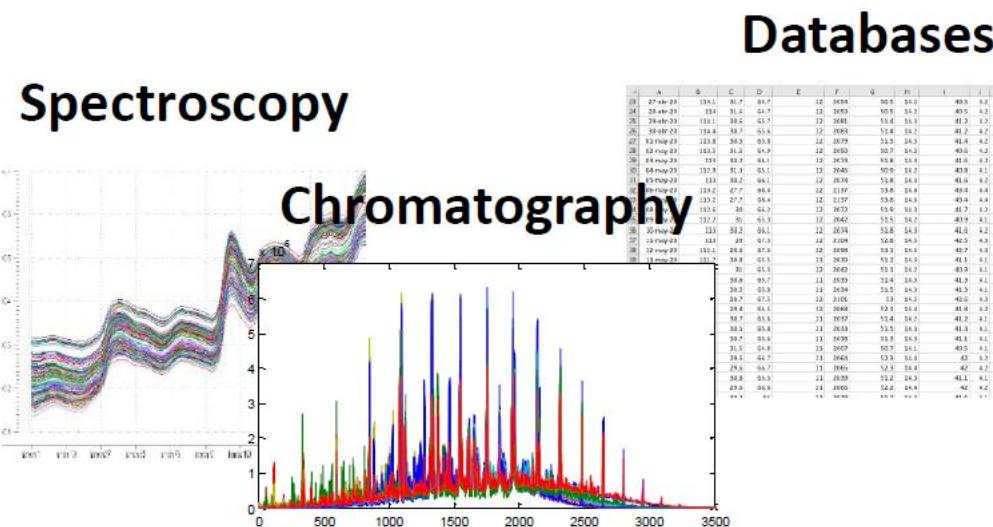
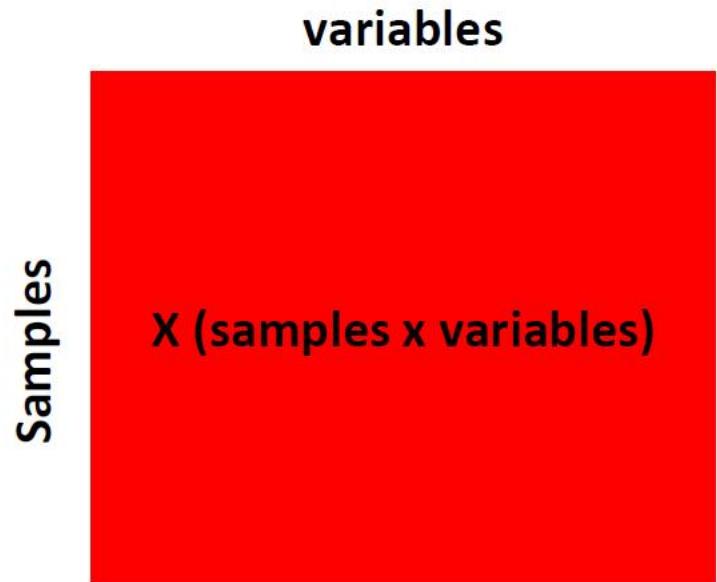


Data Mining

is a set of methods used to **extract usable information from a larger set of any raw data**. It implies analyzing data patterns in already existing data using one or more algorithms.

Premises:

- We have existing data in the shape of a matrix with samples and variables;
- The variables can be of very different nature;
- We extract any valuable information that already exist in the data. **Digging into databases**.

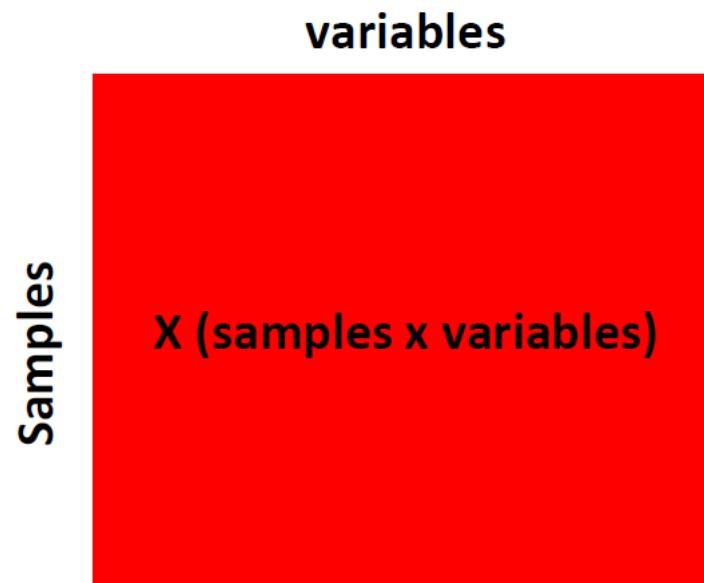


Data Mining

is a set of methods used to **extract usable information from a larger set of any raw data**. It implies analyzing data patterns in already existing data using one or more algorithms.

Premises:

- We have existing data in the shape of a matrix with samples and variables;
- The variables can be of very different nature;
- We extract any valuable information that already exist in the data. **Digging into databases**.



PATTERN RECOGNITION, CLUSTERING, UNSUPERVISED

We need to study the relationship between samples depending on the variables

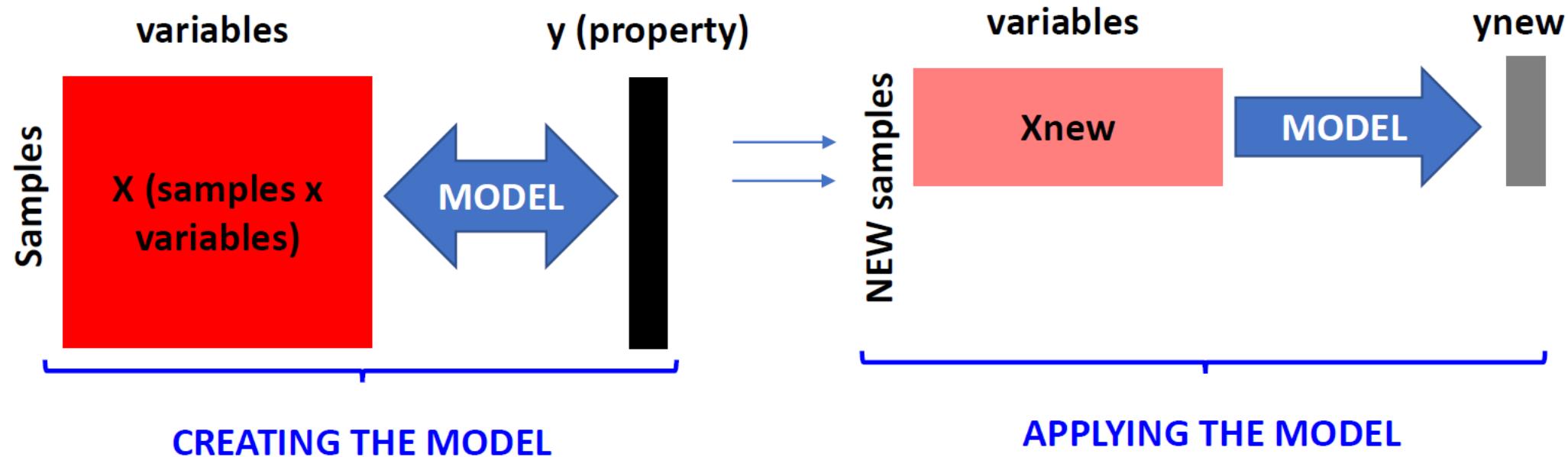
WORKHORSE method →PCA

Machine learning

Methods that parse data, **learn from that data**, and make informed decisions based on what it has learned.

Premises:

- Now the data needs to learn. **Process of training (learning)**;
- Because we will predict using the constructed model.



Machine learning

Methods that parse data, **learn from that data**, and make informed decisions based on what it has learned.

Premises:

- Now the data needs to learn. **Process of training (learning)**;
- Because we will predict using the constructed model.

REGRESSION AND CLASSIFICATION, SUPERVISED

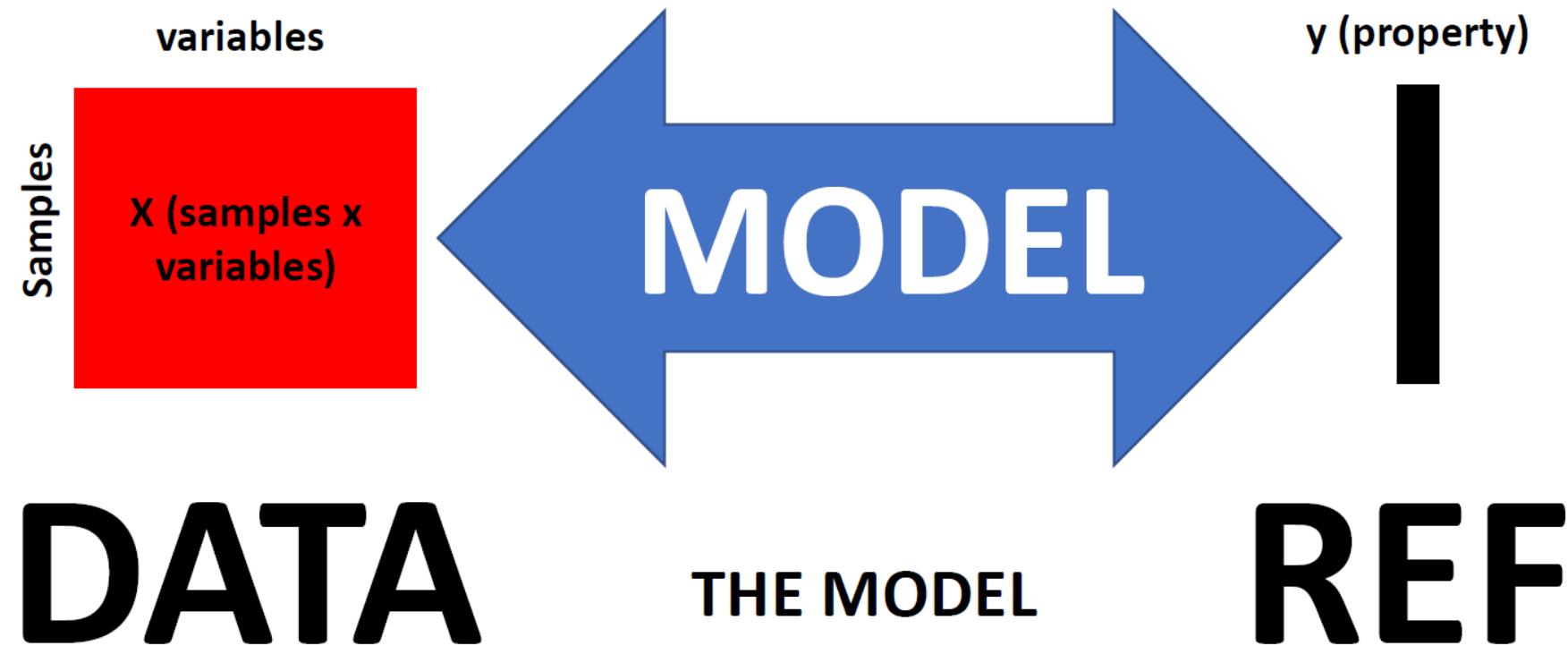
We need to study and optimize the relationship
between X and Y samples depending on the
variables

WORKHORSE method → PLS, LDA, ...

Machine learning

Methods that parse data, **learn from that data**, and make informed decisions based on what it has learned.

Coming back to the definition: Where is the part of Learning and taking decisions???



Machine learning

**Learning procedure:
FIRST, creating the model:**

High quality data:

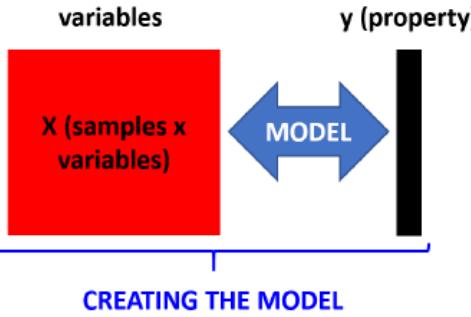
- Reliable information linked to the property we want to measure;
- Comprehensive databases and good instruments;
- Many of them (**How many? As many as you need**).

High quality reference values:

- Standardized protocols approved by authorities and with replicates.

The proper model:

- Studying the structure of the data and the nature of the **relationship with Y**;
- Remember, **the simpler, the better... ALWAYS**;
- Remember to **VALIDATE**;



Deep learning

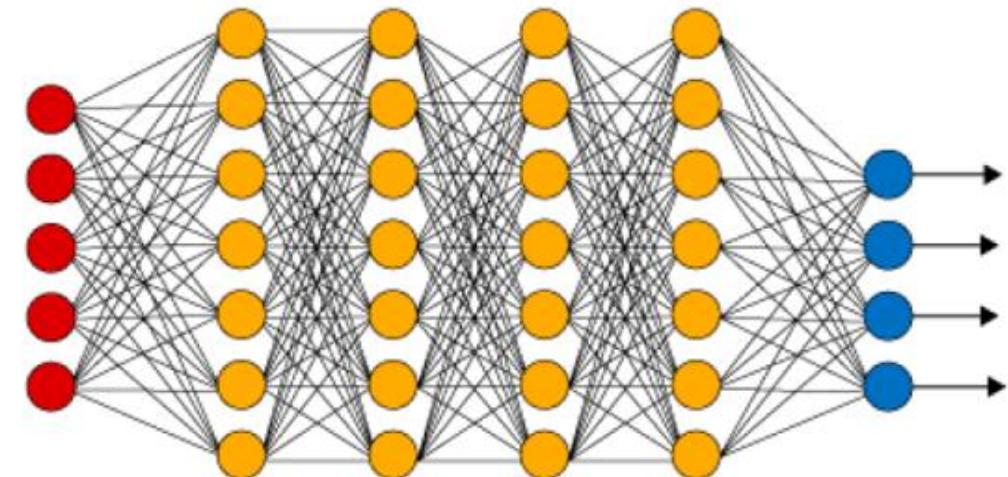
Methods that structure algorithms in layers to create an "artificial neural network" that can learn and make intelligent decisions on its own.

Premises:

- More than one problem/model to solve at the same time. Normally neural networks with MORE than one hidden layer.
- To make decisions on its own??????

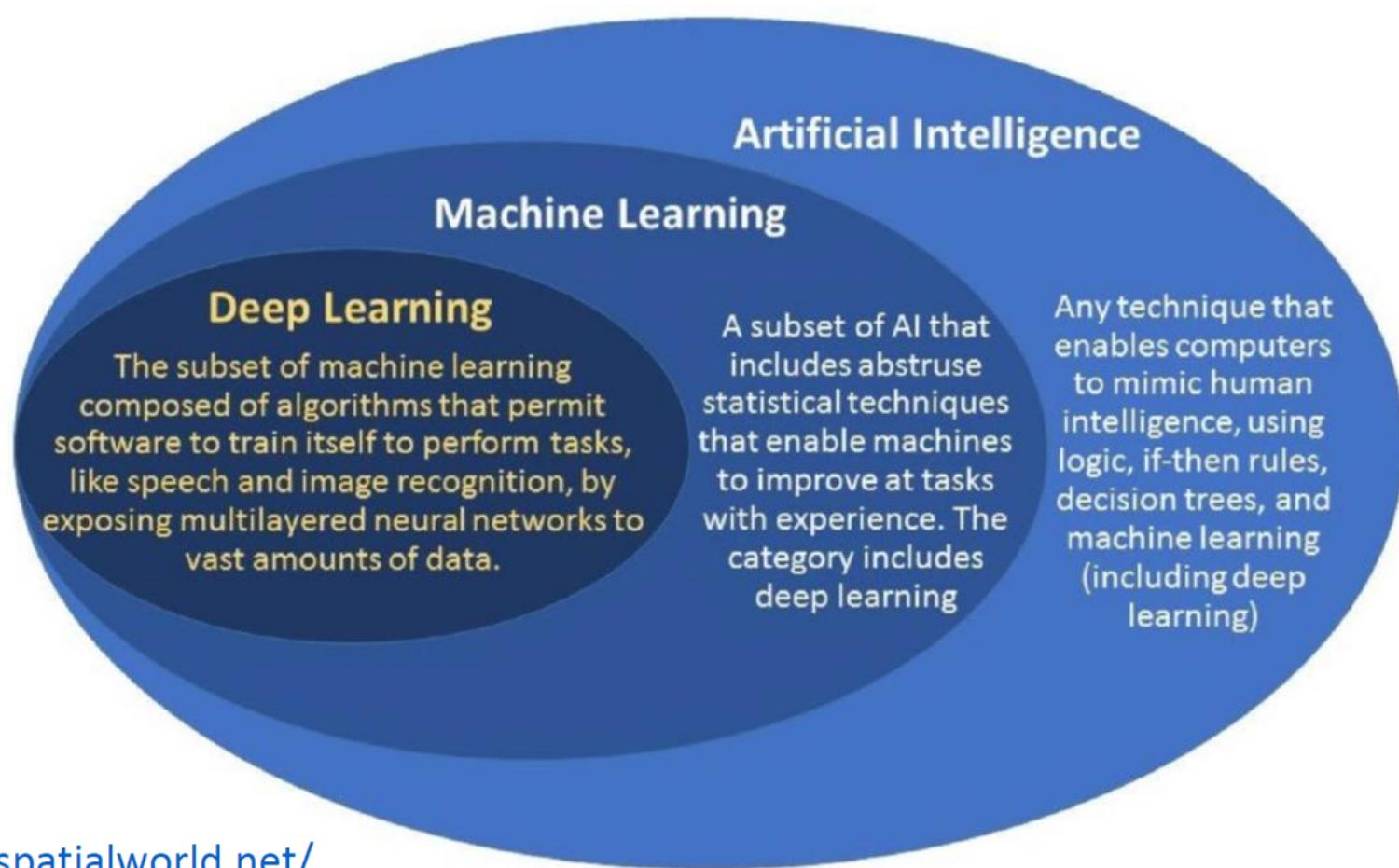
Machines Learn because
WE teach them

We give them access to different databases so they can extract the information



Artificial Intelligence

Providing an example (out of many)

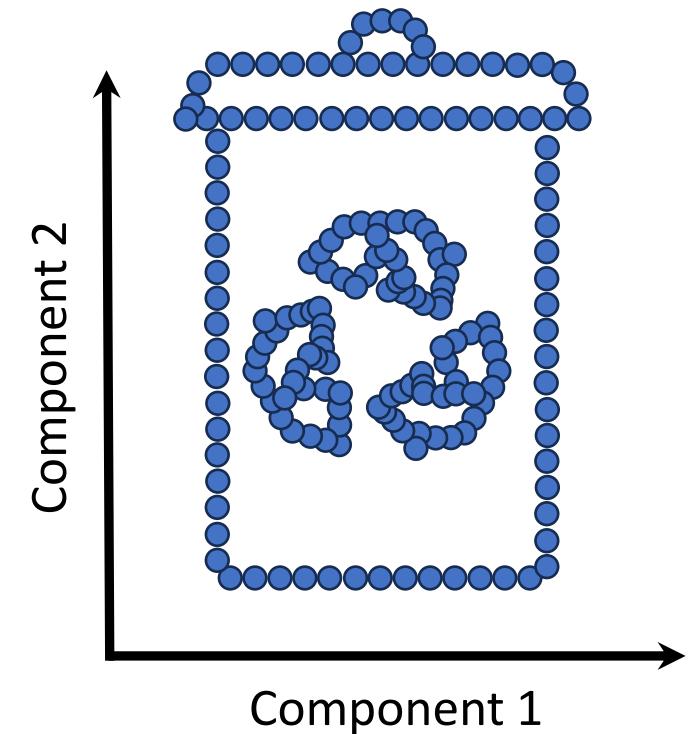
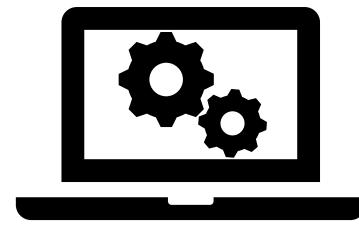


<https://www.geospatialworld.net/>

But remember...

Garbage in, garbage out

Record ID	RowID	Variable_1	Variable_2	Variable_3
Record 1	1	12.34	aa	12
Record 2	2	34	aa	33
Record 3	3	44	cc	22
Record 4	4	-433	ff	44
Record 5	5		N/A	66
Record 6	6	43	dd	33
Record 7	7	34	N/A	66
Record 8	8	6743	df	22
Record 9	9	3	fg	
Record 10	10	4	dd	88
Record 11	11	3	g	55
Record 12	12	N/A	dd	aa



A little bit oh history...

NEWS

History of Chemometrics

Chemometrics was developed in the 1960s. It extracts information from chemical systems by using methods such as multivariate statistics, applied mathematics, and computer science, to address problems in chemistry, biochemistry, medicine, biology, and chemical engineering. Hence, its area took off with the advent of scientific computing, especially with the development of computerized laboratories.

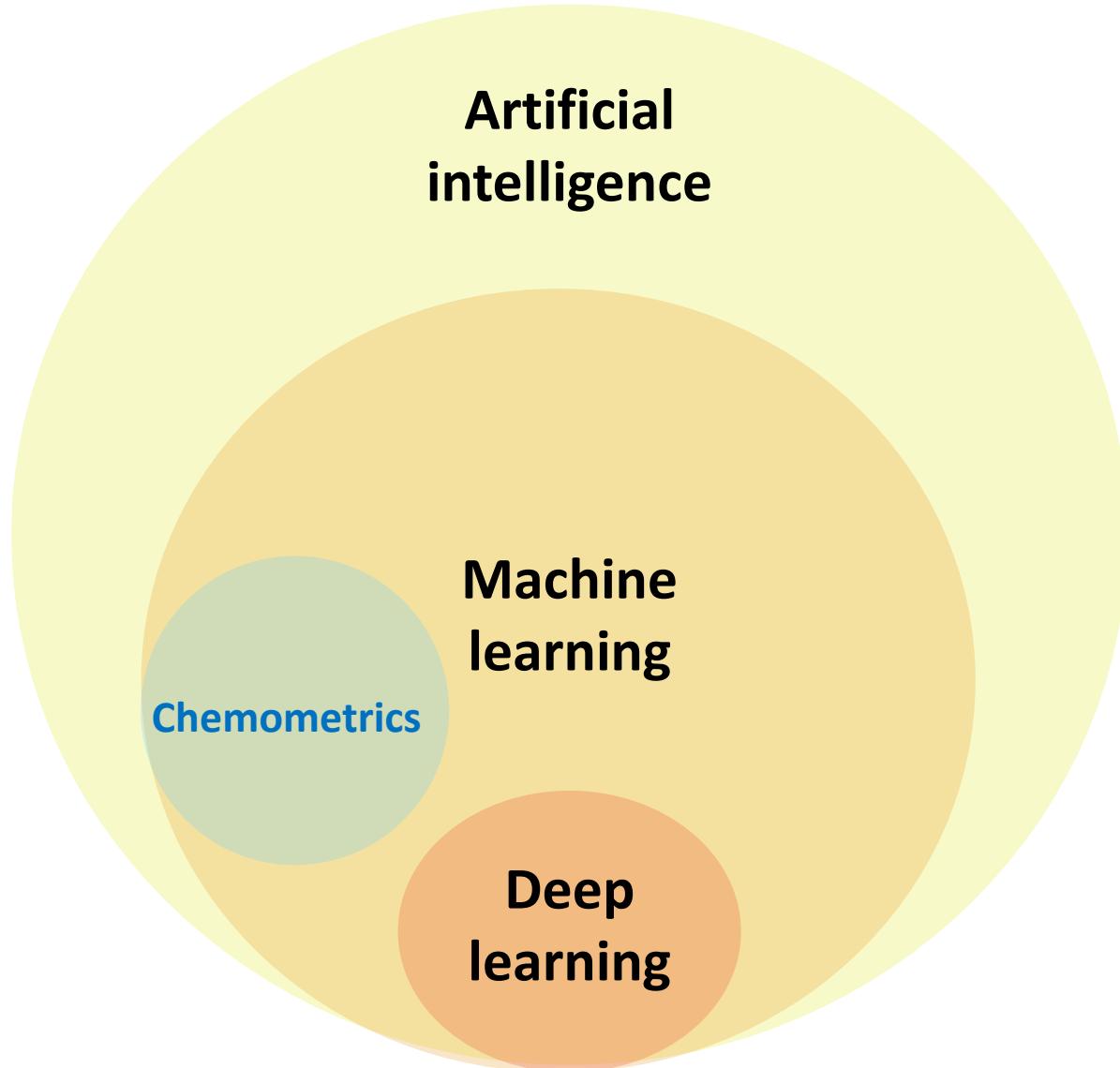
Svante Wold, Umeå Universitet, Sweden, invented the word chemometrics for a grant application in late 1971. In 1974, together with Bruce Kowalski, University of Washington, Seattle, USA, he created the International Chemometrics Society (ICS). The first paper with the word chemometrics in it was published by Wold in 1972. Remarkably, it is only cited seven times according to the Web of Science. Richard G. Brereton, University of Bristol, UK, sees Wold and Kowalski clearly amongst the important pioneers, but they named an existing discipline that had already been seeded in the mid-1960s.

In the 1980s, the first dedicated journals *Chemometrics and Intelligent Laboratory Systems* and *Journal of Chemometrics*, the first book with chemometrics in the title, several ACS symposia, the first book series (Research Studies Press), the first dedicated software (ARTHUR, SIMCA, and UNSCRAMBLER), and the first workshops appeared. A meeting held in Cosenza, Italy, in 1983 was probably the first major attempt to get together a diverse international range of scientists working in chemometrics.

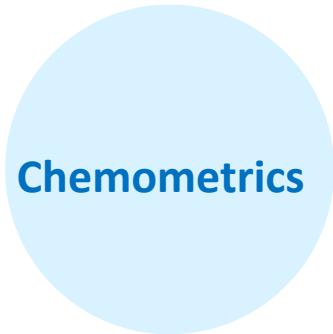
[https://www.chemistryviews.org/details/news/6330601/History_of_Chemometrics/#:~:text=Chemometrics%20was%20developed%20in%20the%201960s.&text=Svante%20Wold%2C%20Ume%C3%A5%20Universitet%2C%20Sweden,International%20Chemometrics%20Society%20\(ICS\).](https://www.chemistryviews.org/details/news/6330601/History_of_Chemometrics/#:~:text=Chemometrics%20was%20developed%20in%20the%201960s.&text=Svante%20Wold%2C%20Ume%C3%A5%20Universitet%2C%20Sweden,International%20Chemometrics%20Society%20(ICS).)



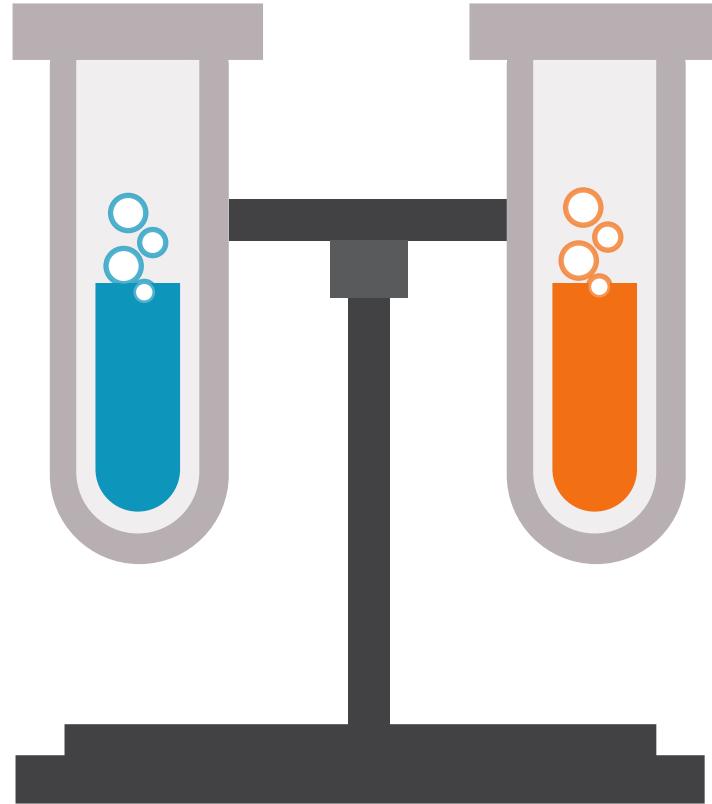
A gentle introduction



A gentle introduction



Multivariate
Data Analysis
(MDA)



Design of
Experiment
(DoE)

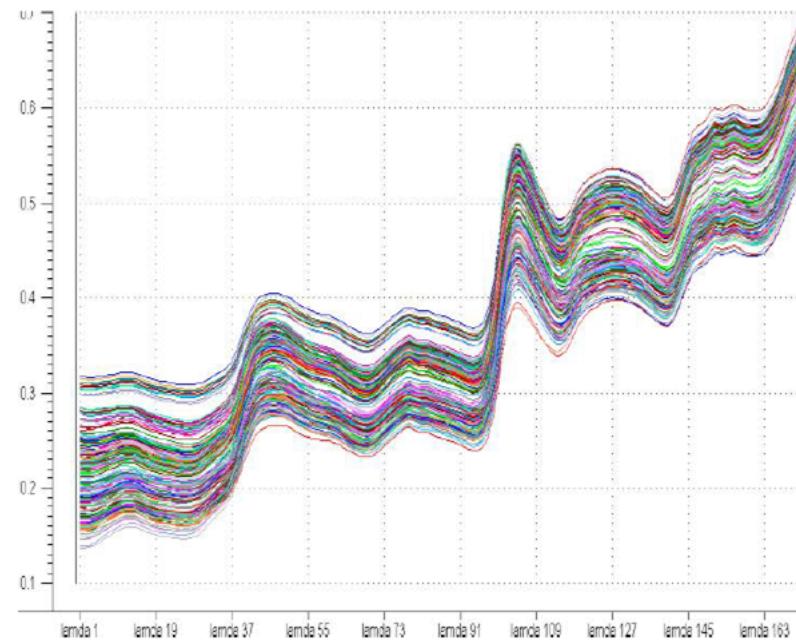
A gentle introduction



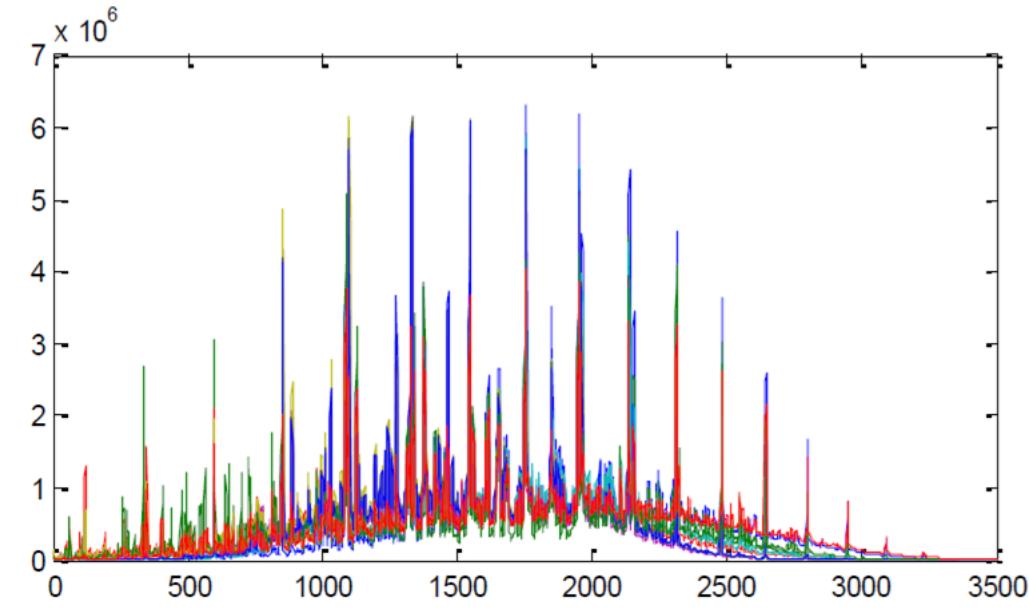
How is the information in modern Science?

- Large amount of data for each sample;
- Large amount of samples;
- Results must be handled in short time;

Spectroscopy

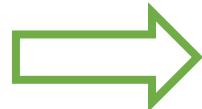


Chromatography

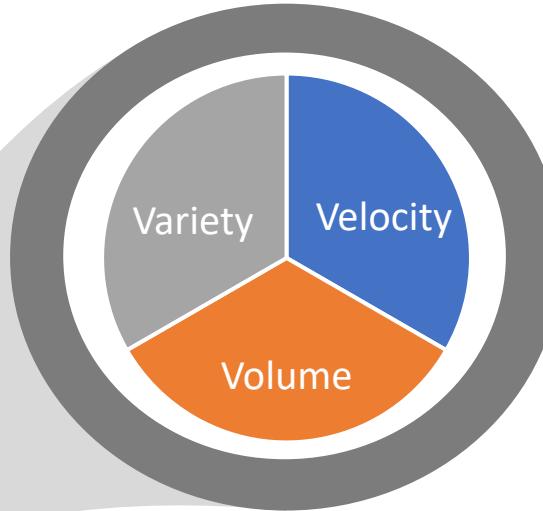


How is the information in modern Science?

3V-data



Continuous technological improvements

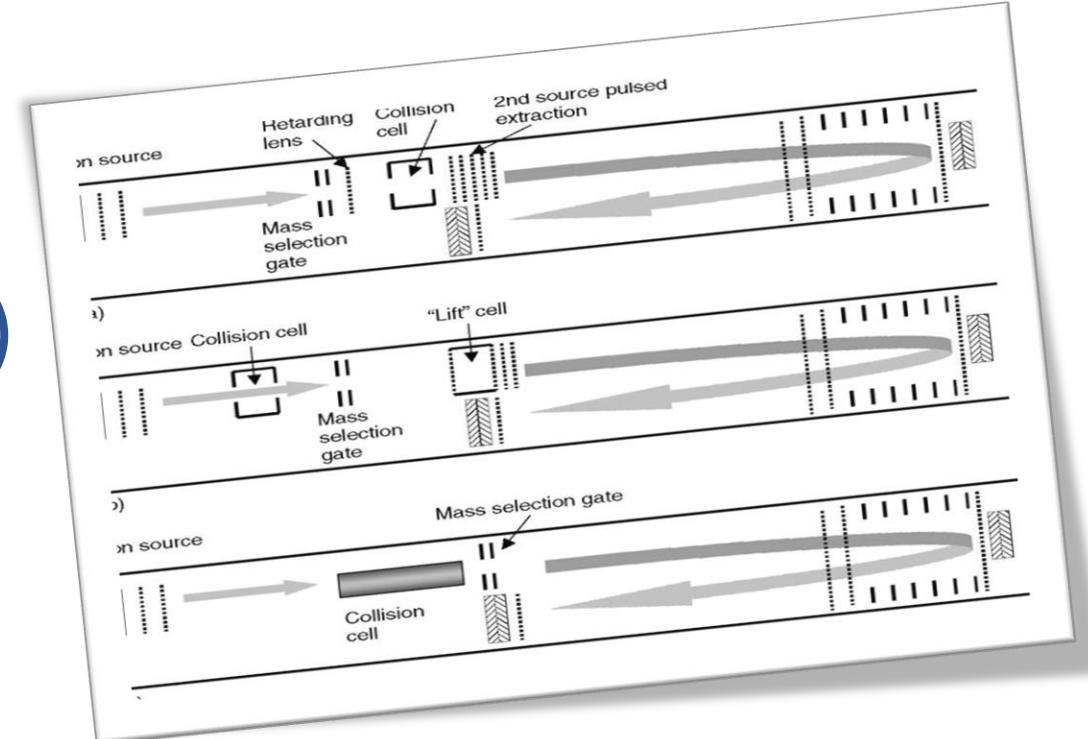
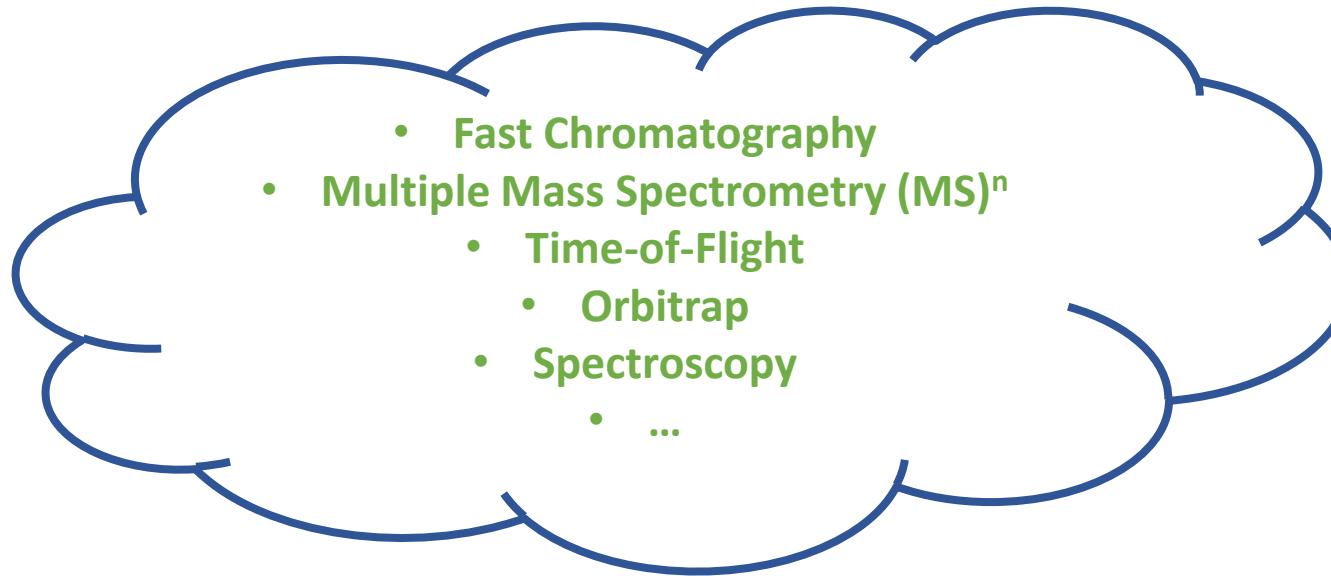


How is the information in modern Science?

Fast electronics

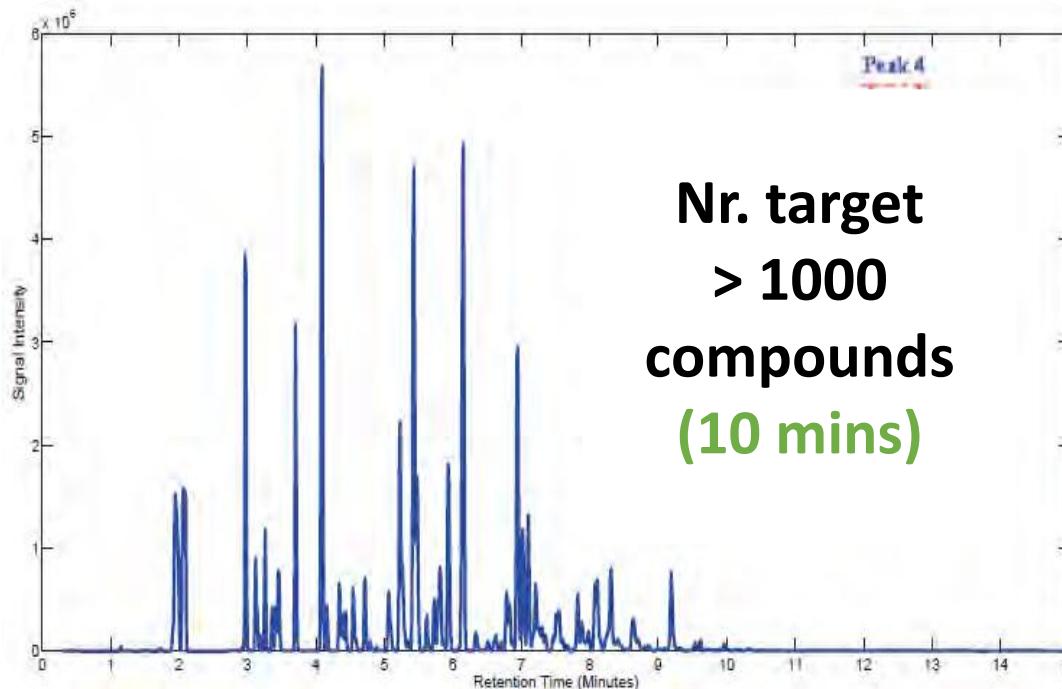


Multi-target analytical techniques



How is the information in modern Science?

Number of compounds



Big Data Storage



How is the information in modern Science?

Features/Variables

2D

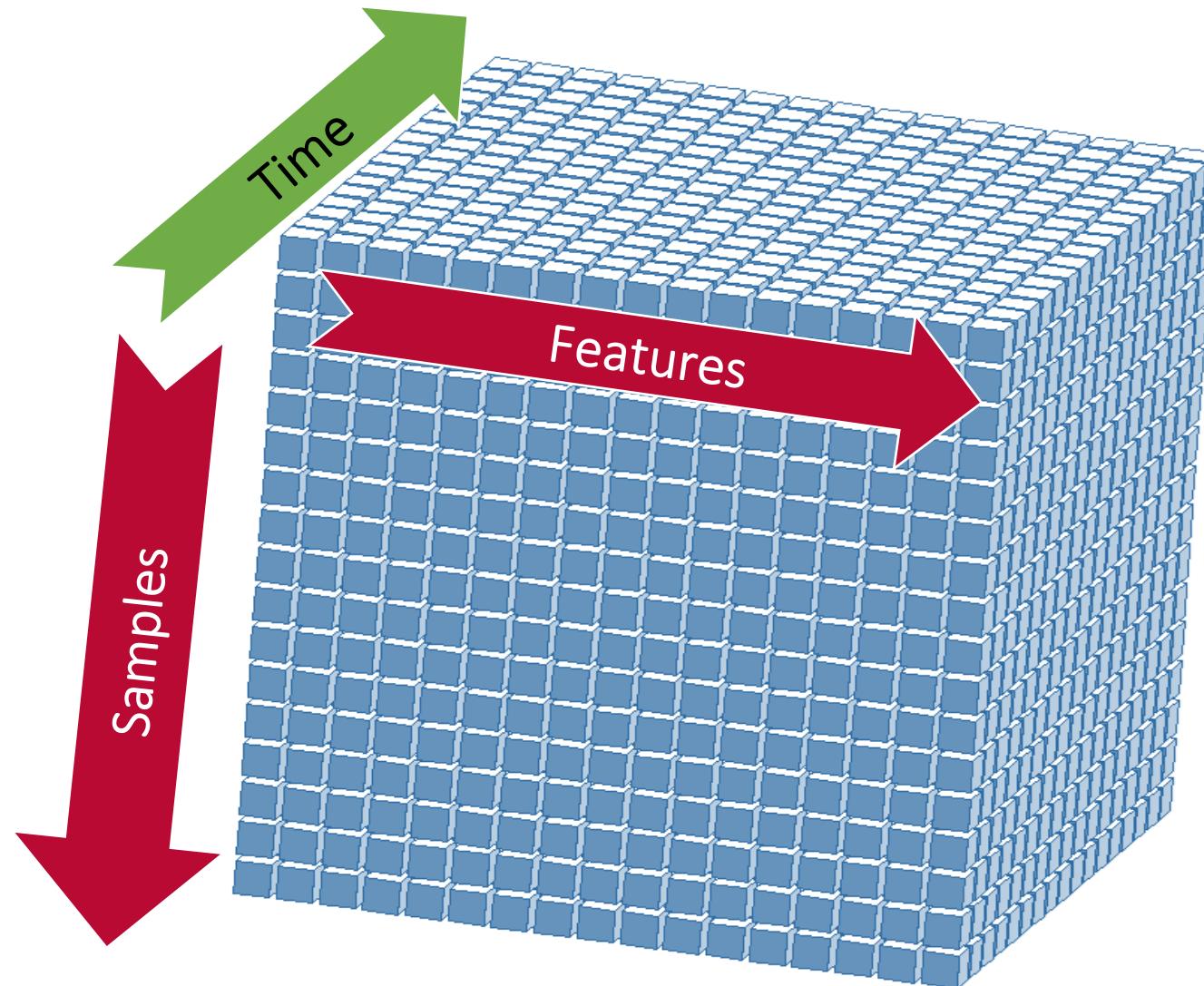
Samples/Instances

Viewer
Relation: final

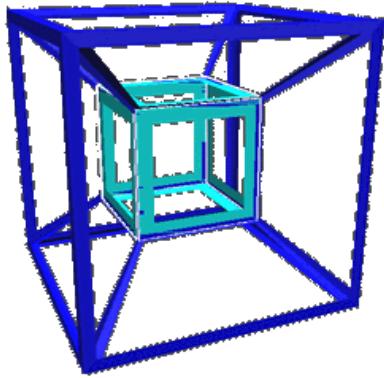
No.	Rating Numeric	Survey Numeric	Prize Numeric	Punishment Numeric	Aspen Numeric	Snowmass Numeric	Breckenridge Numeric	Keystone Numeric	ABasin Numeric	Loveland Nominal	CrestedButte Nominal	Vail Numeric	Silverton Numeric	WinterPark Numeric	Mary Jane Numeric	Eldora Numeric
1	0.675	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	0.9	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0
4	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0
5	0.6	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
6	0.95	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
7	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
8	0.8	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
10	0.85	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0
11	0.94	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
12	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
13	0.8	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
14	1.0	20.0	10.0	30.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
15	0.95	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0
16	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
17	0.85	20.0	10.0	30.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0
18	0.9	20.0	10.0	30.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0
19	1.0	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
20	0.675	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.575	20.0	1.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
22	0.925	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
23	0.9	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
24	0.6	20.0	10.0	30.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0

Undo OK Cancel

How is the information in modern Science?



How is the information in modern Science?



0 dimensions:

POINT



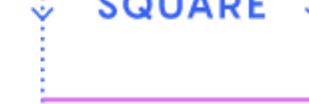
1 dimension:

LINE SEGMENT



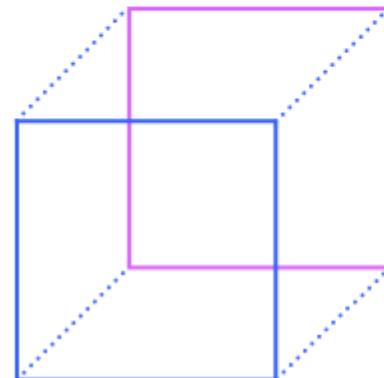
2 dimensions:

SQUARE



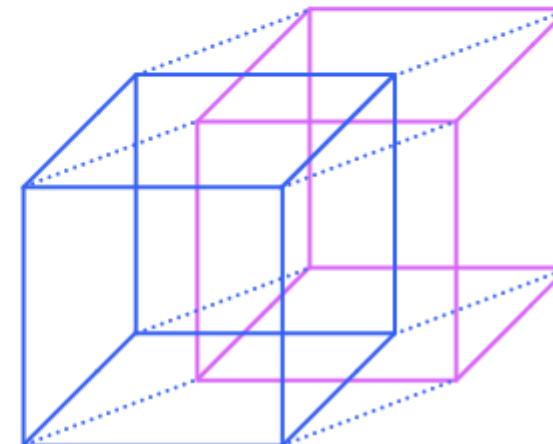
3 dimensions:

CUBE



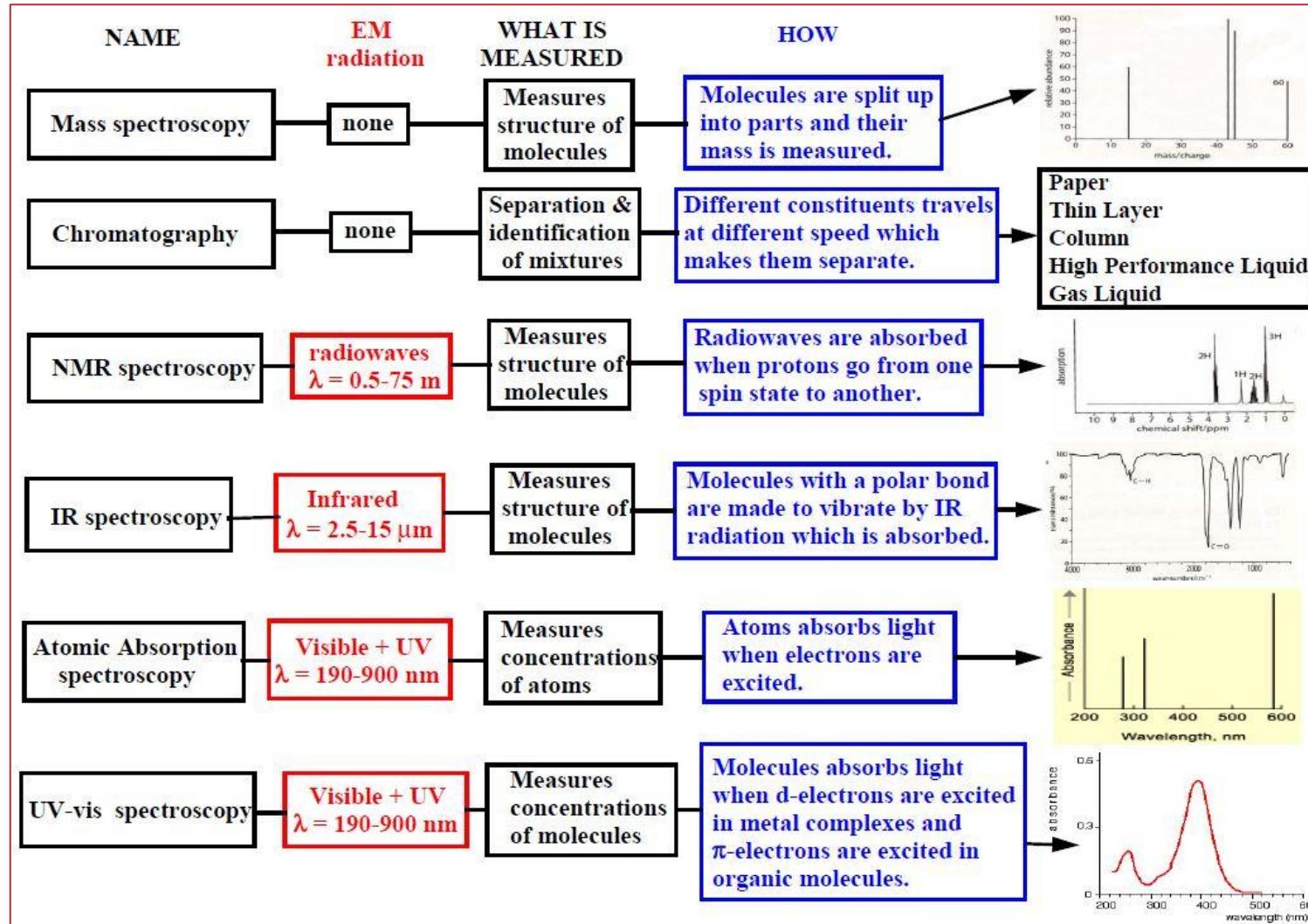
4 dimensions:

TESSERACT



- 1. Samples;**
- 2. Features;**
- 3. Time;**
- 4. Location.**

A spectroscopic focus



- Simultaneous evaluation of several data and techniques;
- Possibility to "combine" the results from different analytical techniques → DATA FUSION.

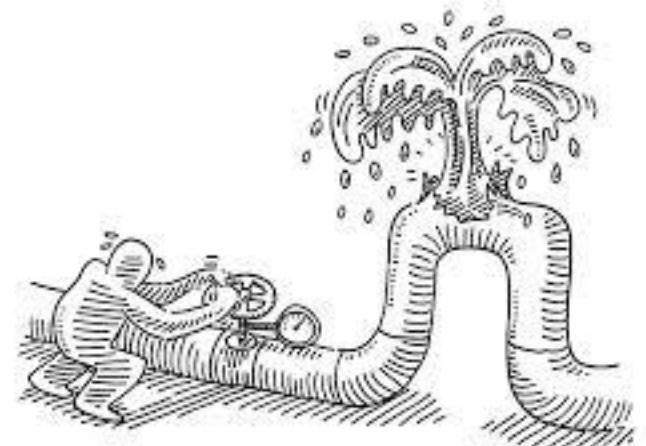
Finally, we can say that today we will do Data Mining, Machine Learning and a bit of Deep Learning (when needed) in chemistry...
So, we do...

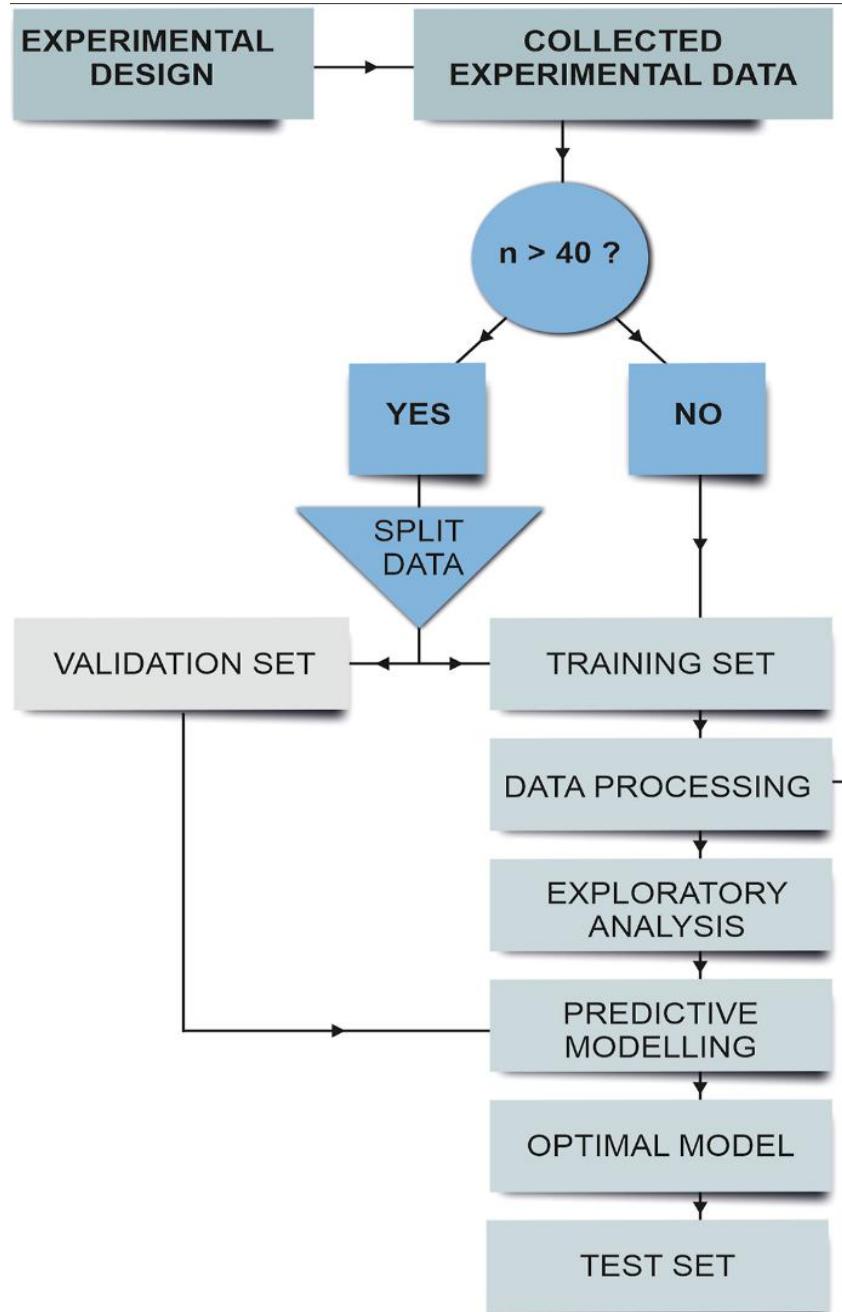
CHEMOMETRICS

But let us tell you a secret: we don't care how you want to call it!

- Obtain your data according to what you want to answer
- UNDERSTAND your data!
- Construct BIG and COMPREHENSIVE databases
- Feed your databases with more GOOD data
- Construct your models. And start from the simplest approach. Thinking simple at the beginning can help you to understand the relationship between your data, your model and the response that you want to obtain. **THE SIMPLER, THE BETTER!!!**
- **AND REMEMBER TO VALIDATE!**

The chemometric pipeline





TrAC Trends in Analytical Chemistry

Volume 135, February 2021, 116157



A guide to good practice in chemometric methods for vibrational spectroscopy, electrochemistry, and hyphenated mass spectrometry

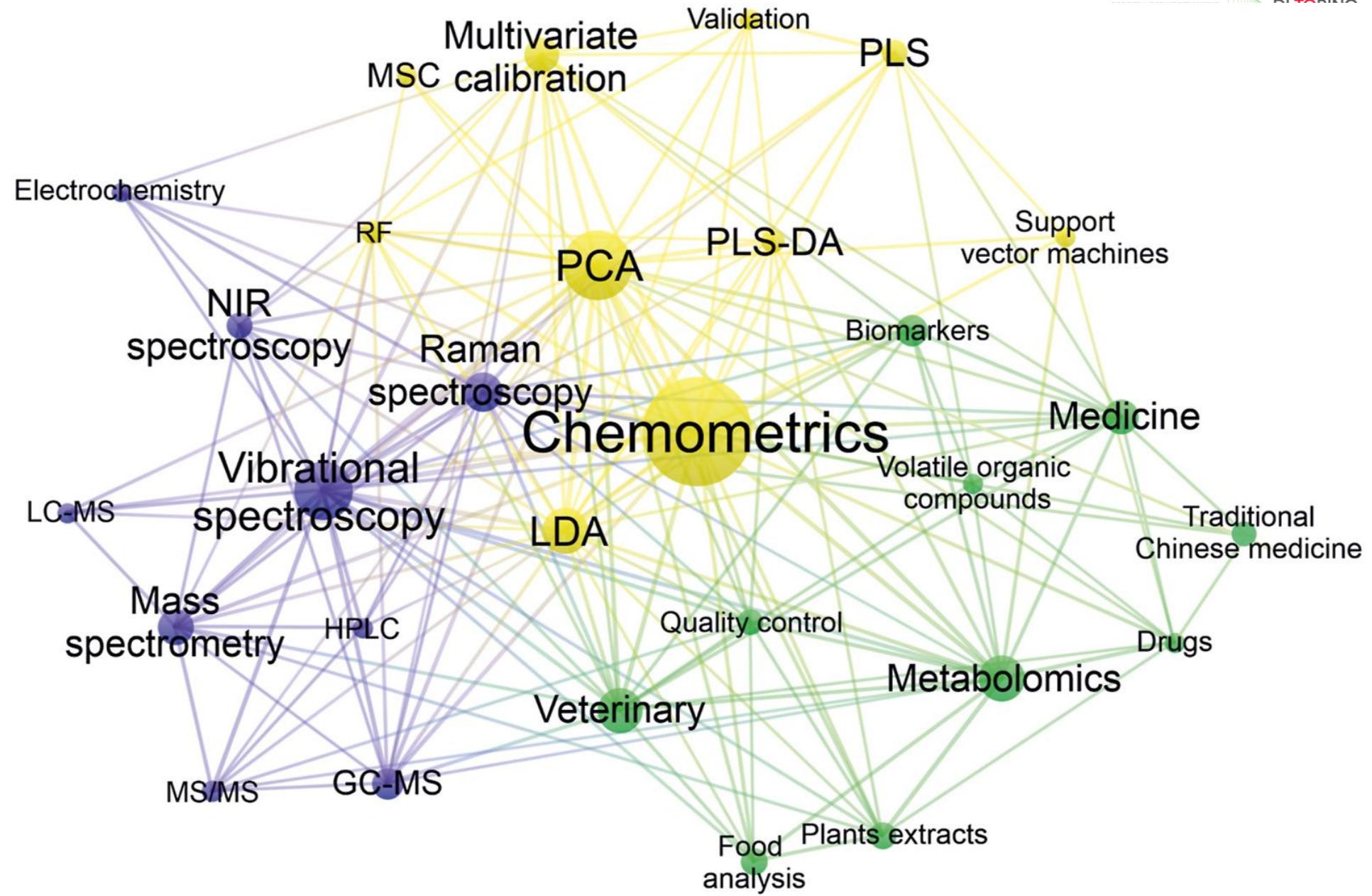
[Manuel David Peris-Díaz](#) , [Artur Kręzel](#)

Show more

+ Add to Mendeley Share Cite

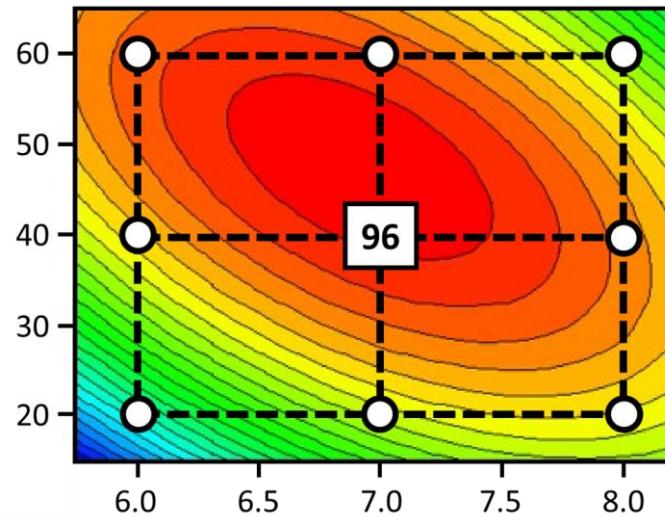
<https://doi.org/10.1016/j.trac.2020.116157>

Get rights and content



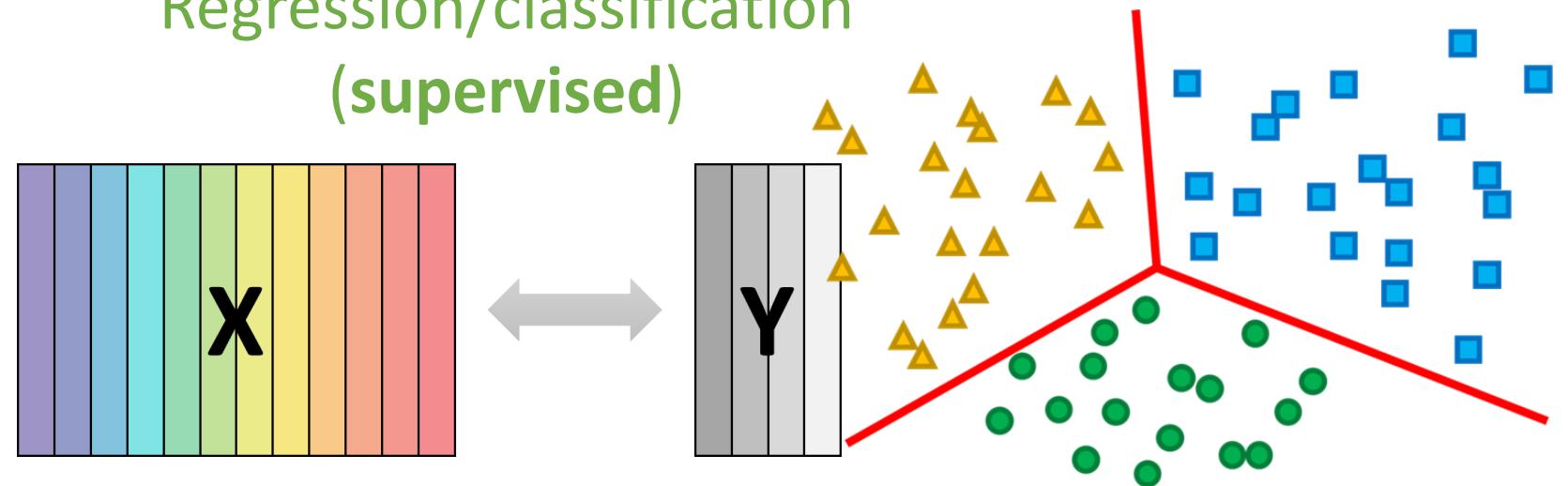


Design of Experiment

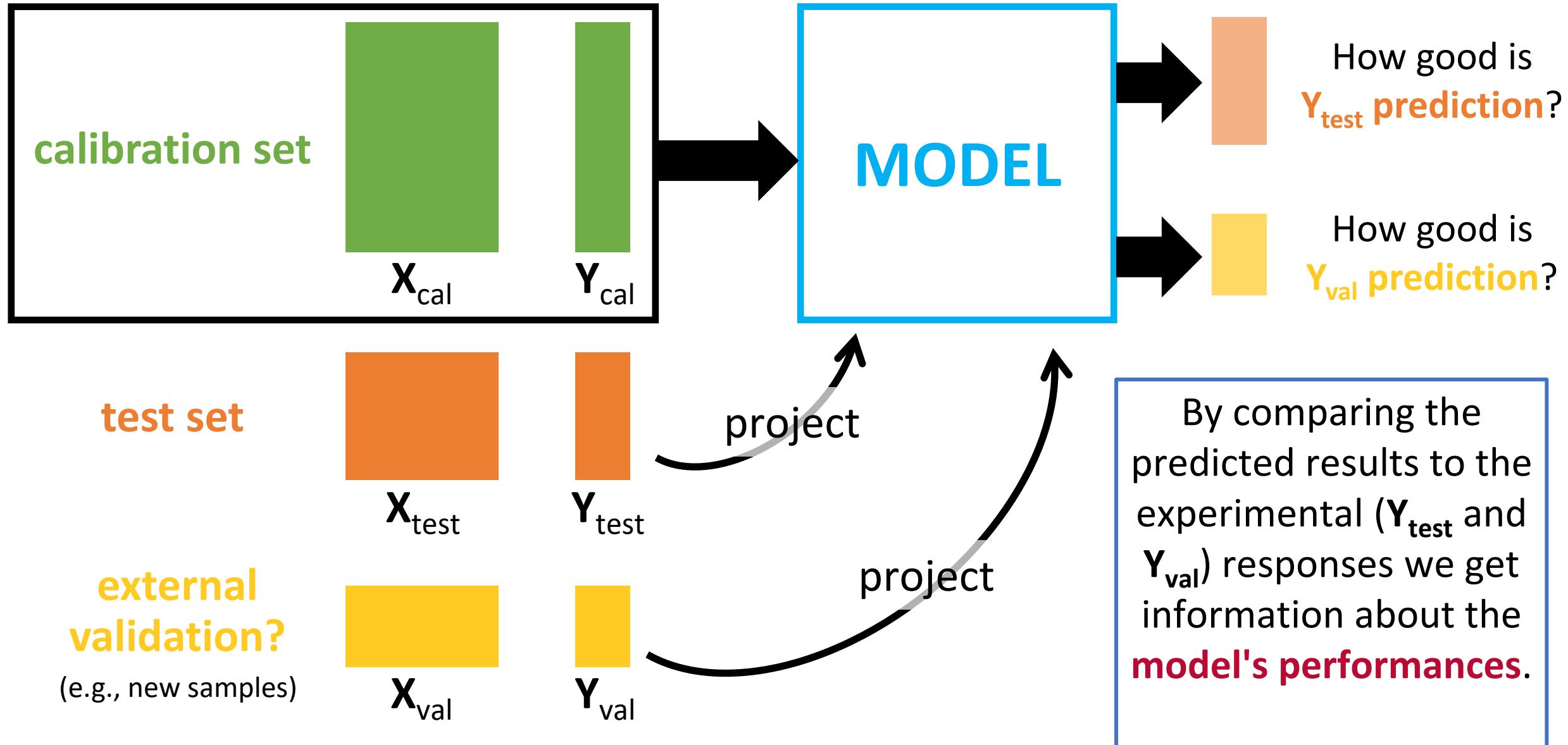


Exploratory analysis
(unsupervised)

Regression/classification
(supervised)

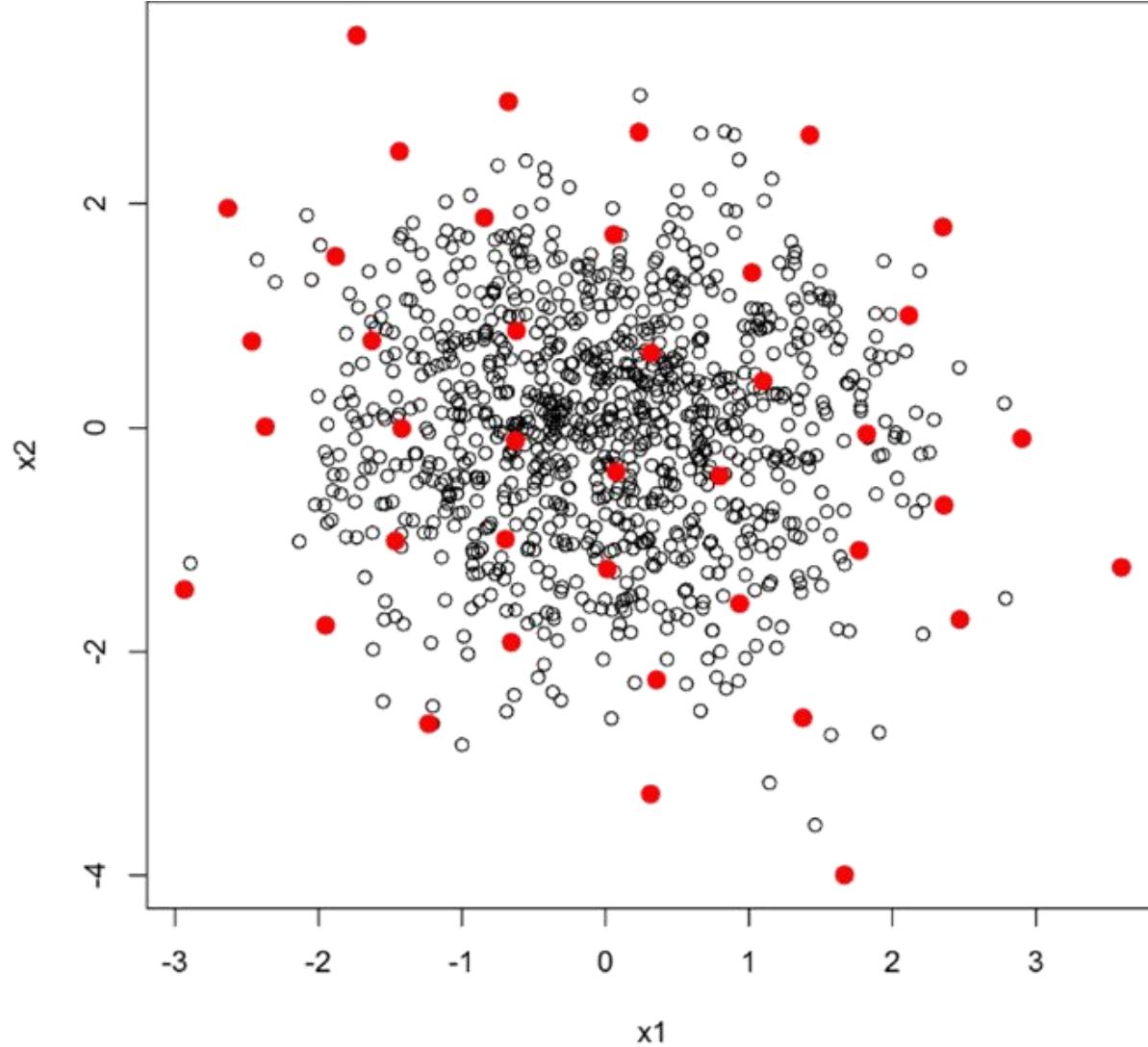


Validation



Sampling

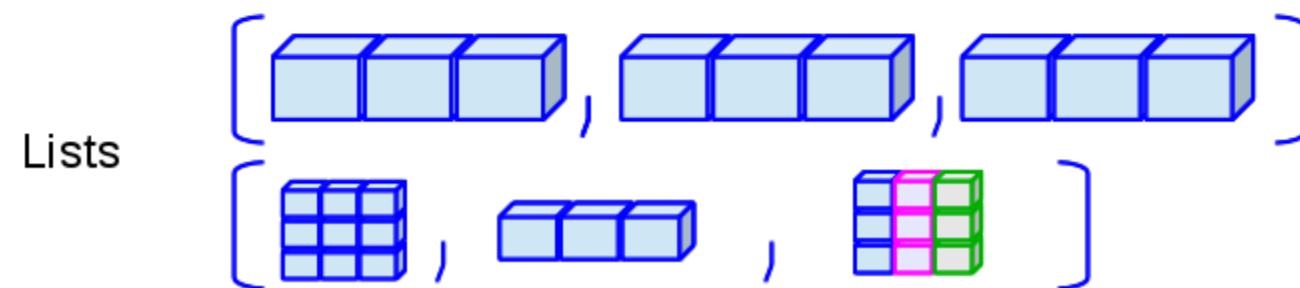
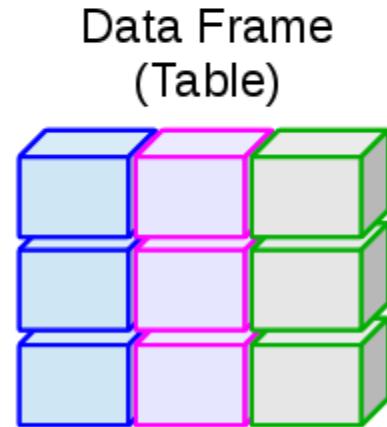
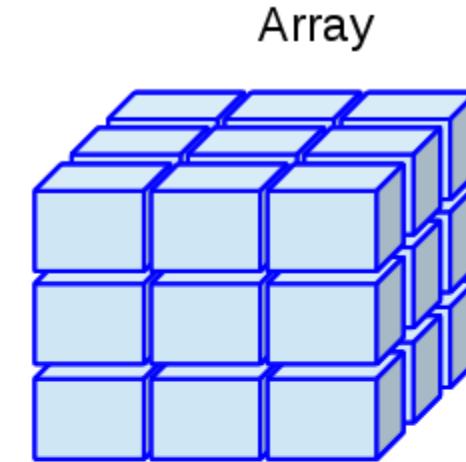
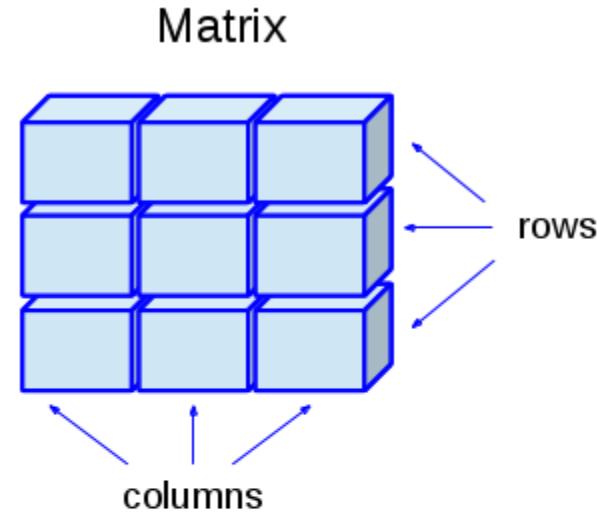
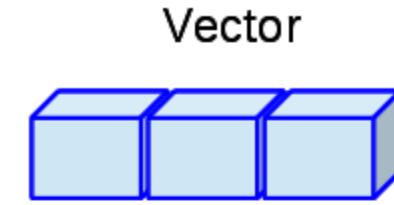
Kennard & Stone
algorithm



but... data first!



Data structure



Data structure

Usually **data** (\neq information) are reported in TABLEs:

MATRIX

each **rows** contains an **OBJECT** (sample, individual, ...) described by a (**vector**) or mode **VARIABLES (features, attributes, ...)**, one for each column of the matrix.

MISSING DATA
NA
(Not Available)

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Sample 1	1.81	0.18	30.18	0.19	199.90
Sample 2	2.15	0.21	30.22	0.23	240.28
Sample 3	2.14		30.22	0.23	240.39
Sample 4	2.26		30.23	0.24	254.12
Sample 5	1.99	0.20	30.20	0.21	218.49
Sample 6	1.96		30.20	0.20	217.39
Sample 7	2.02	0.20	30.20		222.90
Sample 8	1.87	0.19	30.19	0.19	204.85
Sample 9	2.18		30.22	0.23	243.71
Sample 10	2.15	0.22	30.21	0.22	237.53

NaN
(Not a Number)

Data structure

Usually **data** (\neq information) are reported in TABLEs:

MATRIX

each **rows** contains an **OBJECT** (sample, individual, ...) described by a (**vector**) or mode **VARIABLES (features, attributes, ...)**, one for each column of the matrix.

data
1.5
6.3
3.2
9.1
2.9

Data in increasing order	Relative ranks
1.5	1
2.9	2
3.2	3
6.3	4
9.1	5

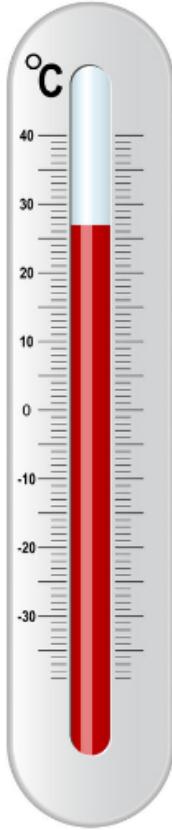
Rank: position when sorted in increasing order.

Loss of information (drawback) but there are no assumptions over the distribution
→ non-parametric

Data type

CONTINUOUS

They describe quantities that can be measured and read according to a **scale**.

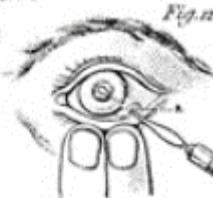


DISCRETE (CATEGORICAL)

They describe categories that can be counted on the basis of **nominal** classes or can be **sorted**.

Eye colour:

- Brown
- Blue
- Green
- Grey



Gender:

- Male
- Female



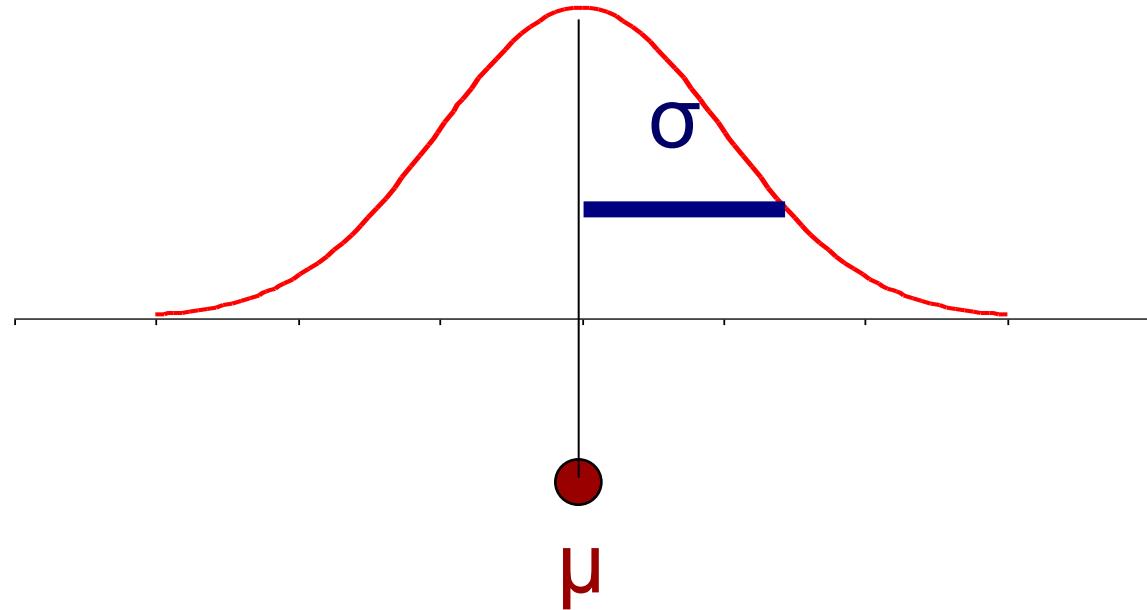
Politics:

- Labour
- Conservative
- Liberal



The normal (Gaussian) distribution

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

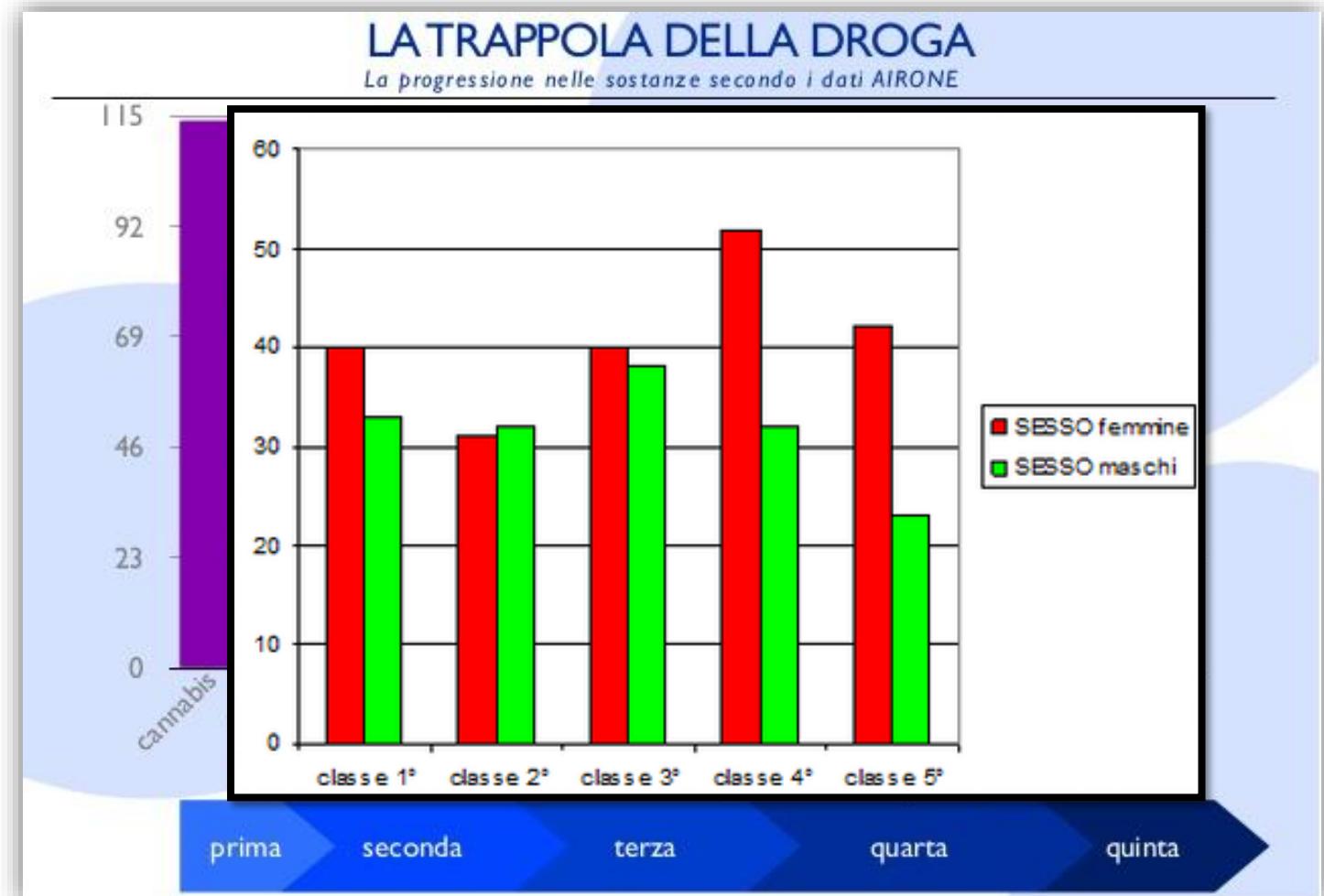


Frequency vs probability - histograms

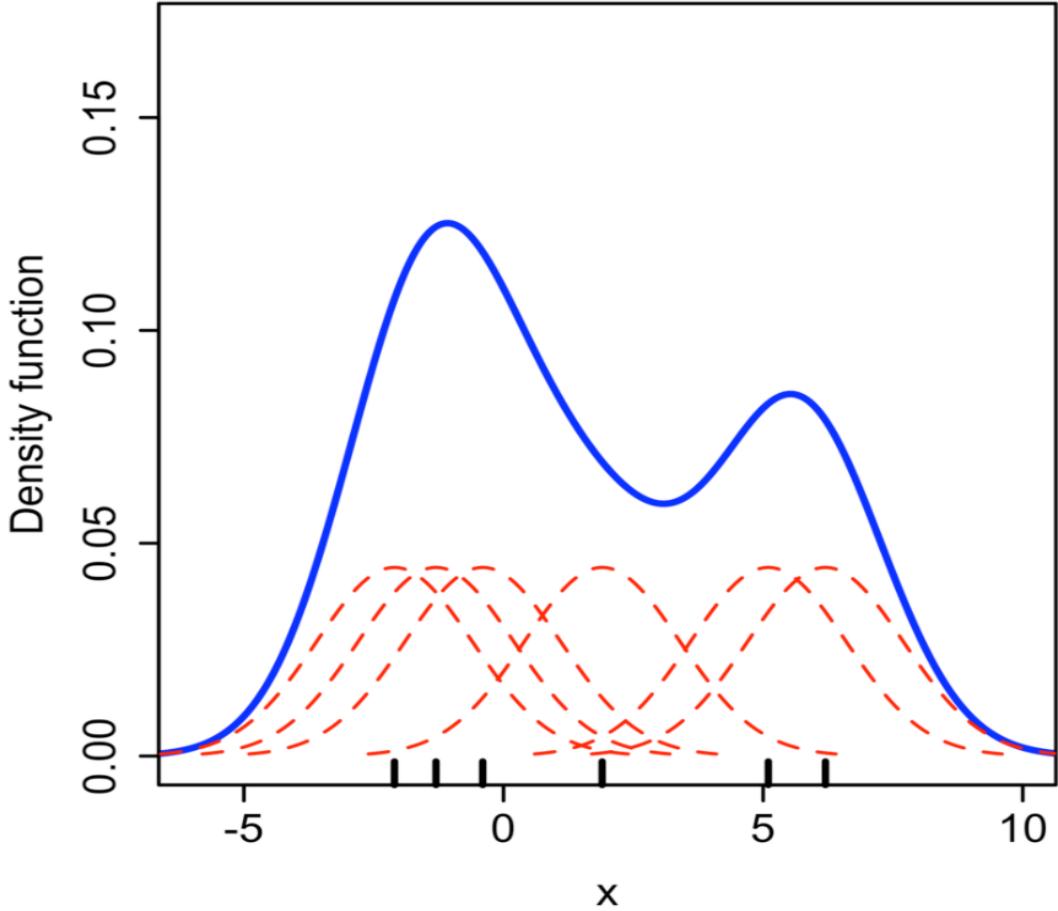
Designed by Karl Pearson and illustrated on November 18, 1891.

Defined as "time-diagram" in its lecture entitled "Maps and Chartograms".

Y axis can show frequency (%)

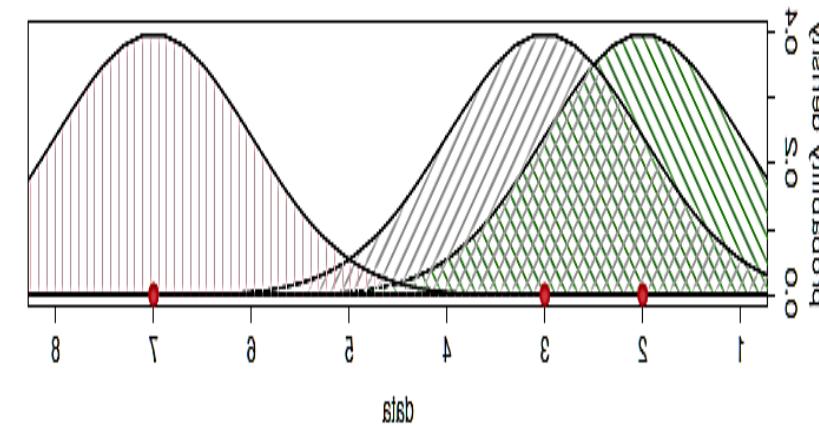


Frequency vs probability – Kernel Density Estimation (KDE)



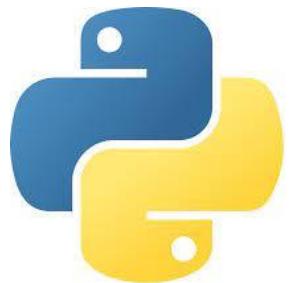
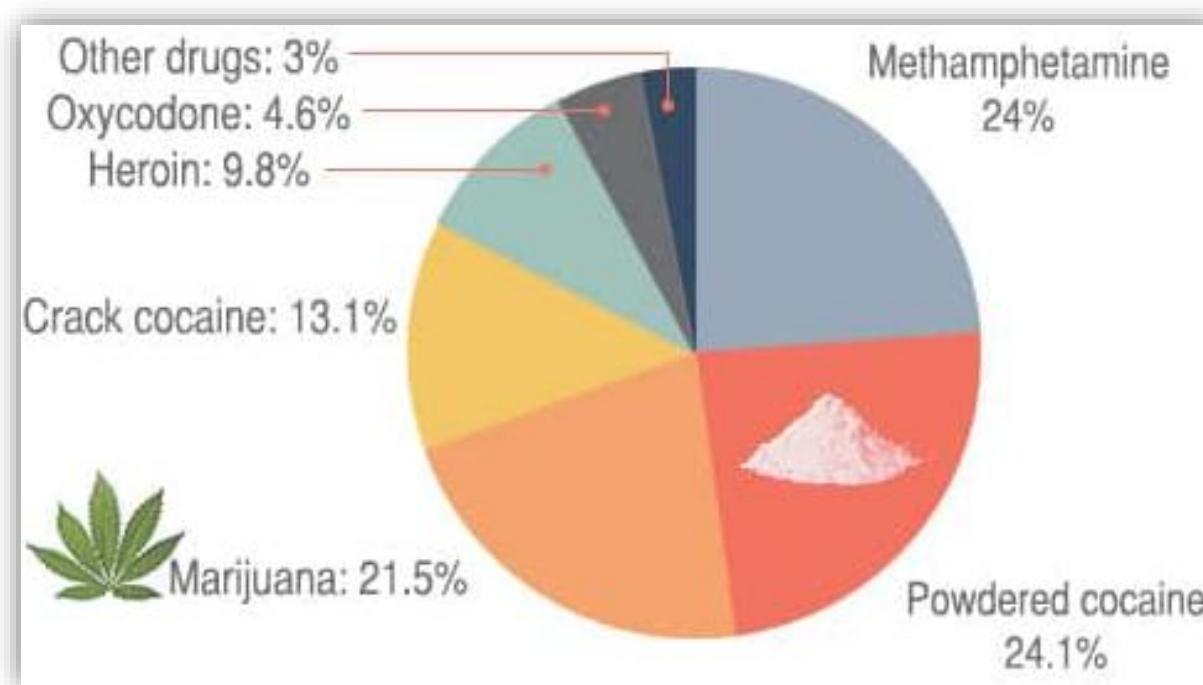
$$f(x_i, h) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right)$$

where K is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the bandwidth



Frequency vs probability – pie-chart

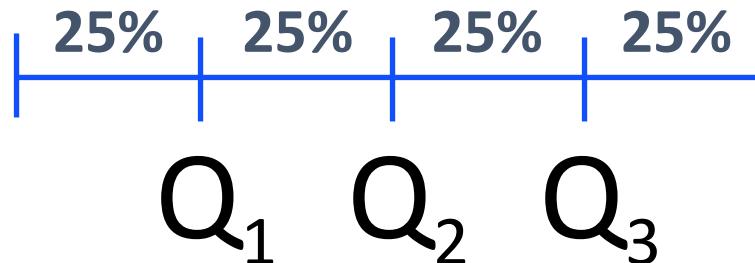
Also known as [aerograms](#), it involves making proportions starting from the values measured for each class/sample under examination and, once their relative percentage frequency is obtained, converting it into degrees (100% corresponds to 360 °).



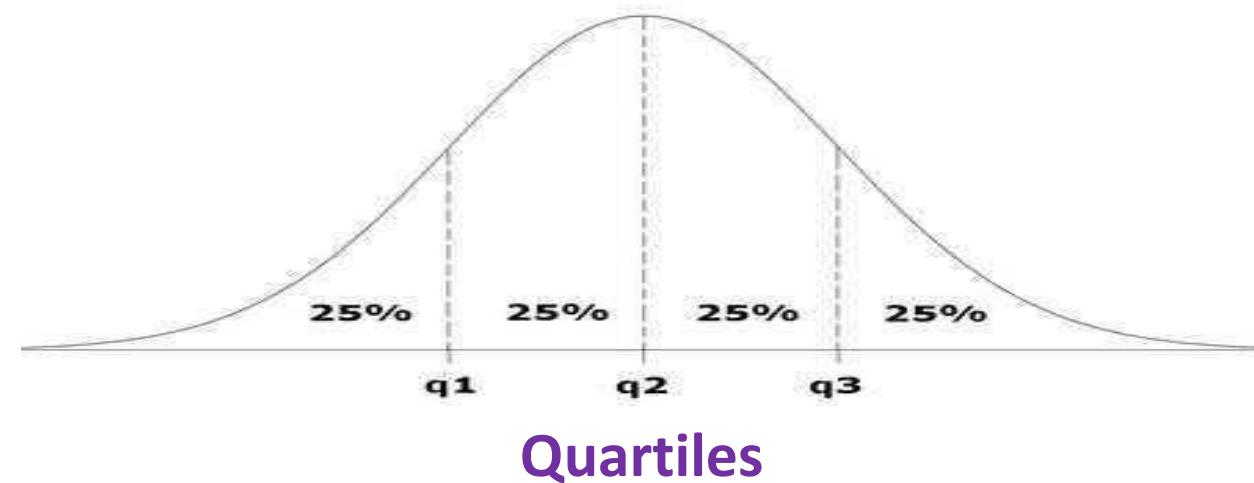
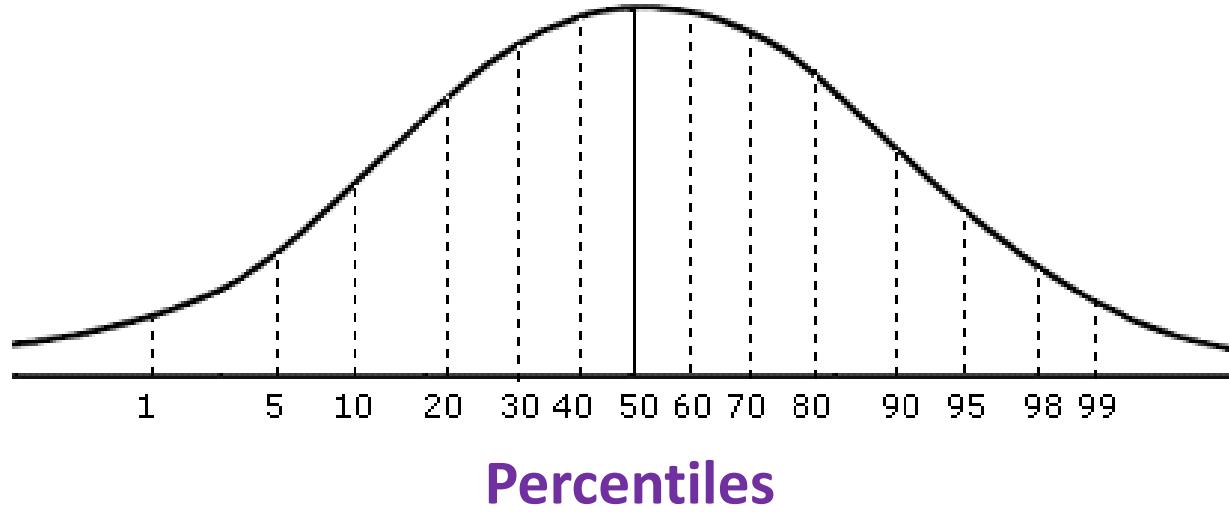
Quartiles and percentiles

Q_1 , Q_2 , Q_3

divide the **ranked scores** into 4 equal parts

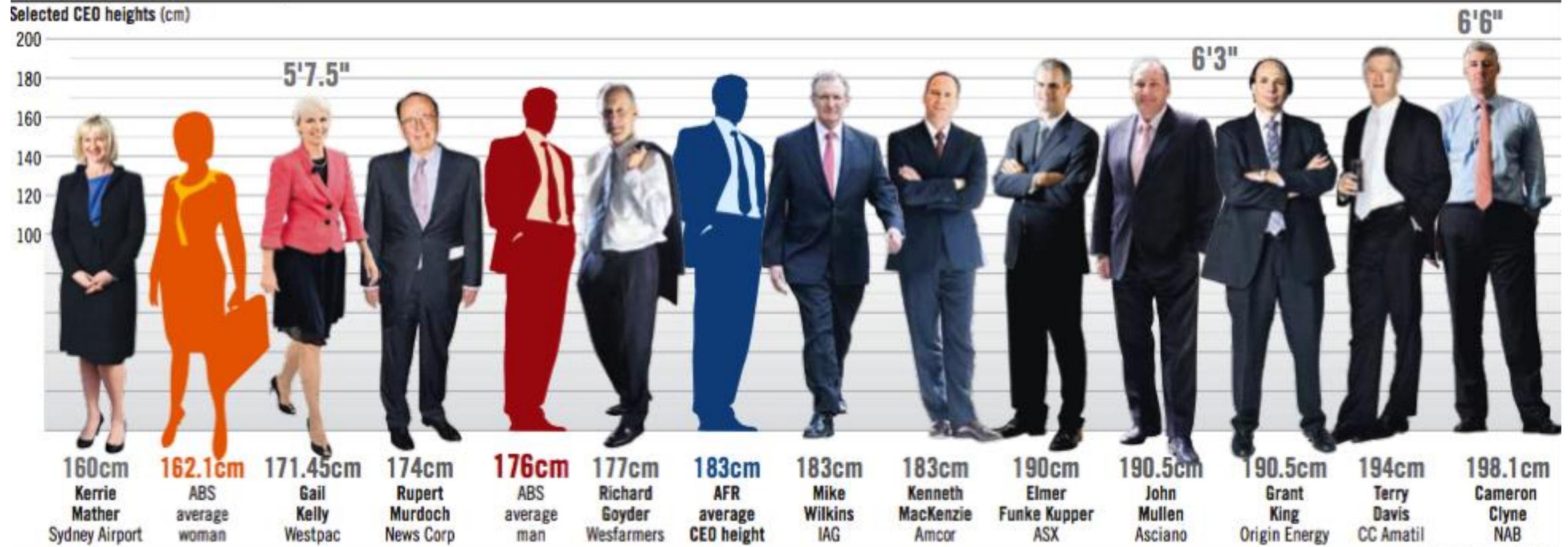


Quartiles and percentiles



Quartiles and percentiles

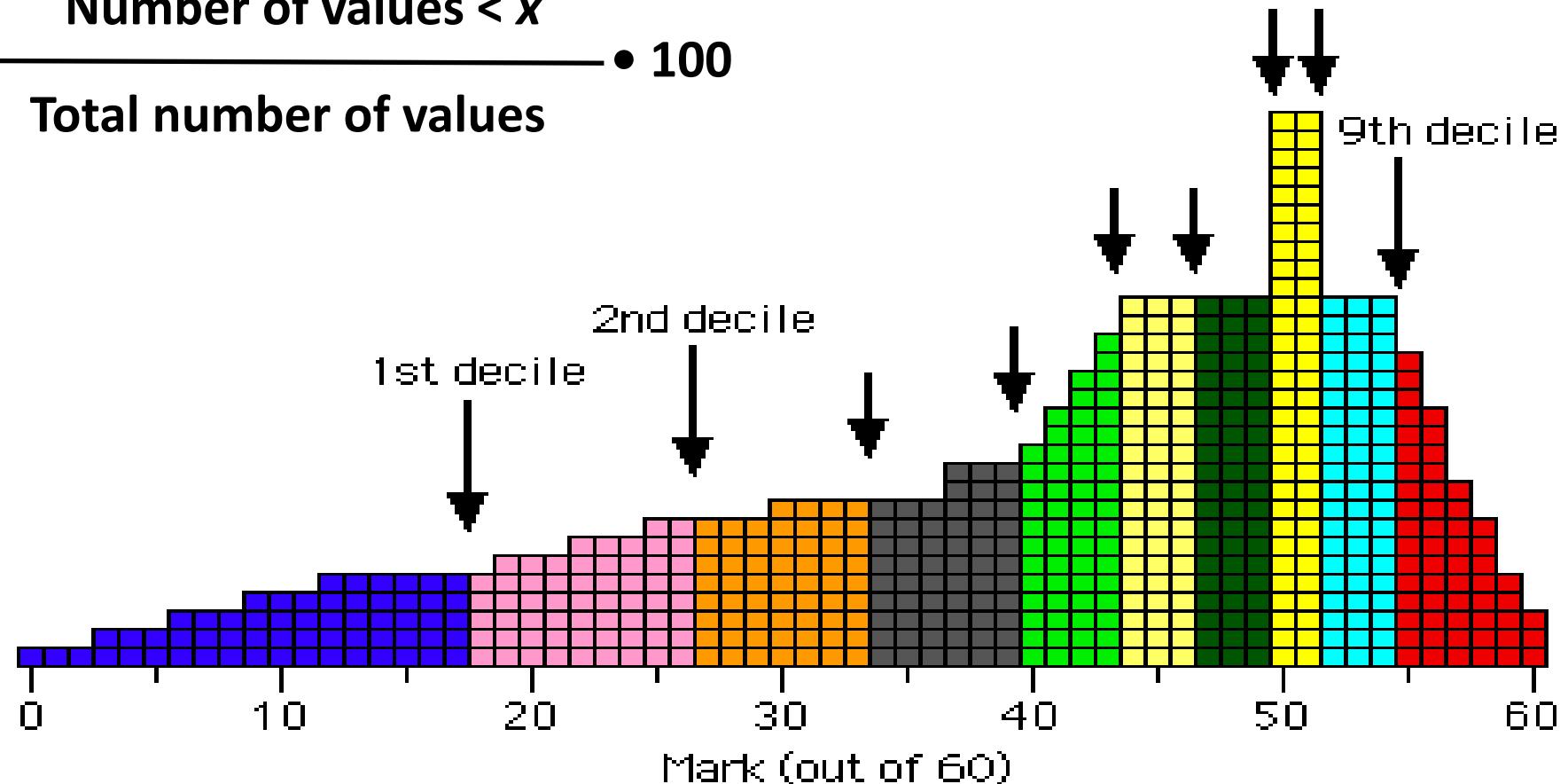
How do you measure up?



Quartiles and percentiles

99 percentiles (k^{th} percentile)

$$\text{Percentile of a value } x = \frac{\text{Number of values} < x}{\text{Total number of values}} \cdot 100$$



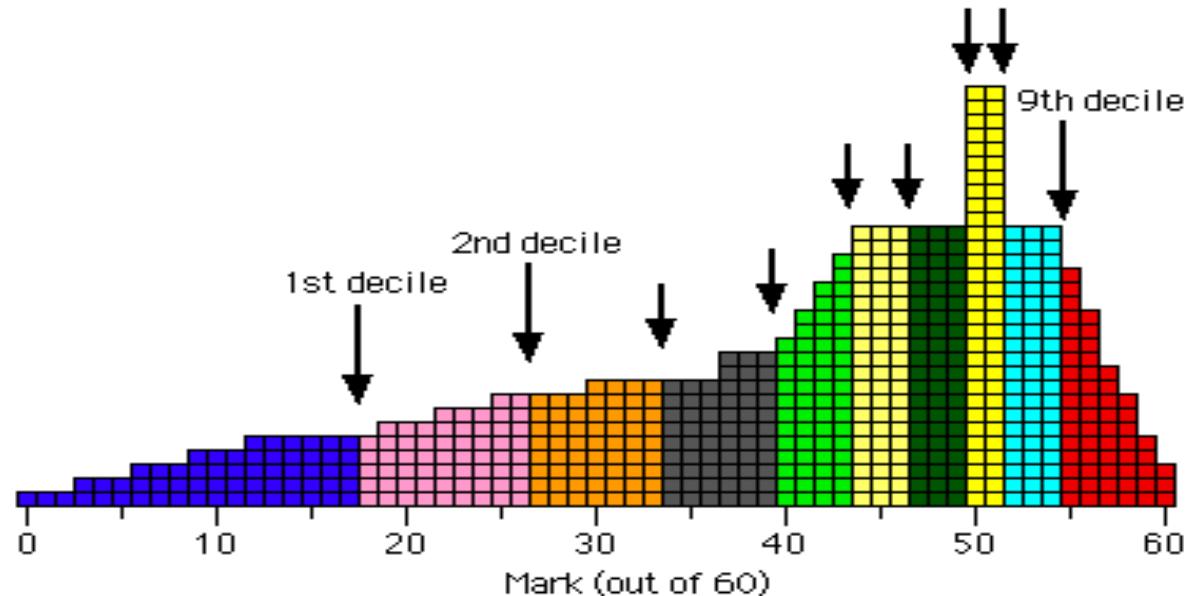
Boxplots

Interquartile Range: $Q_3 - Q_1$

Semi-interquartile Range: $\frac{Q_3 - Q_1}{2}$

Midquartile Range: $\frac{Q_1 + Q_3}{2}$

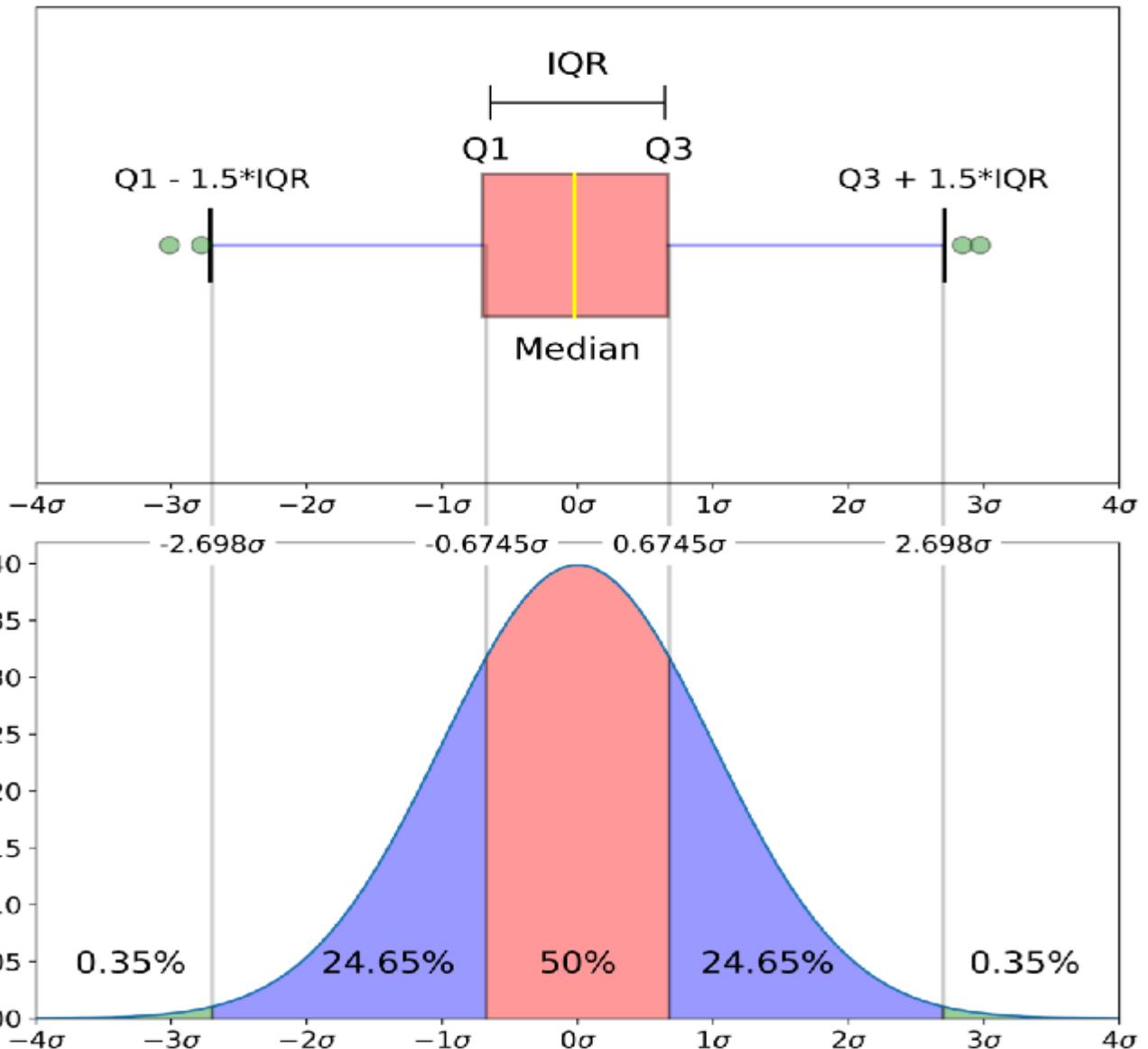
10–90 Percentile Range: $P_{90} - P_{10}$



Boxplots

Optimal for:

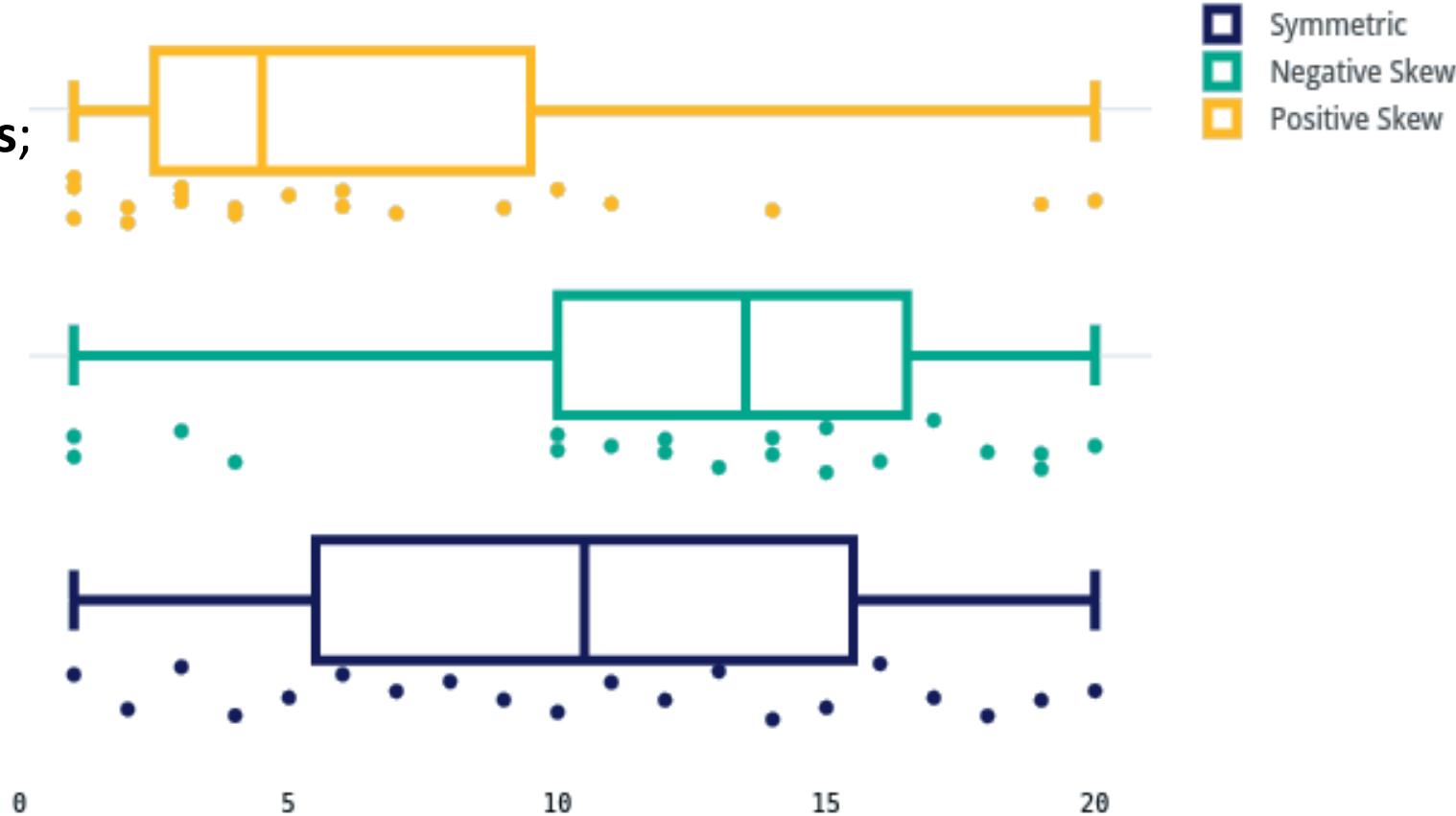
- Identify **anomalous data (outliers)**;



Boxplots

Optimal for:

- Identify **anomalous data (outliers)**;
- Representing one or more **populations**;

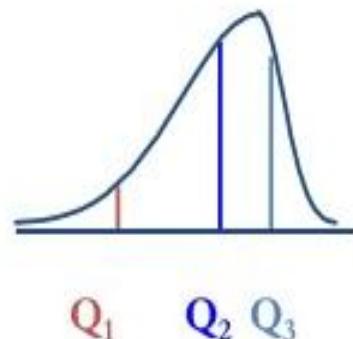


Boxplots

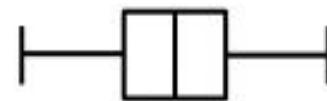
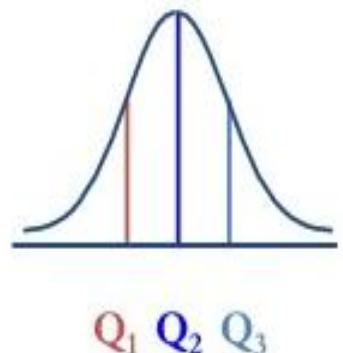
Optimal for:

- Identify **anomalous data (outliers)**;
- Representing one or more **populations**;
- Quickly evaluate **normality**;

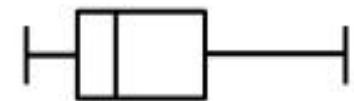
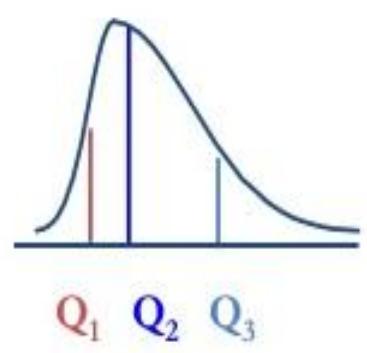
Left-Skewed



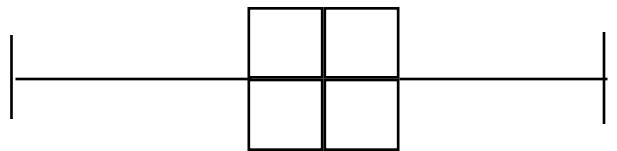
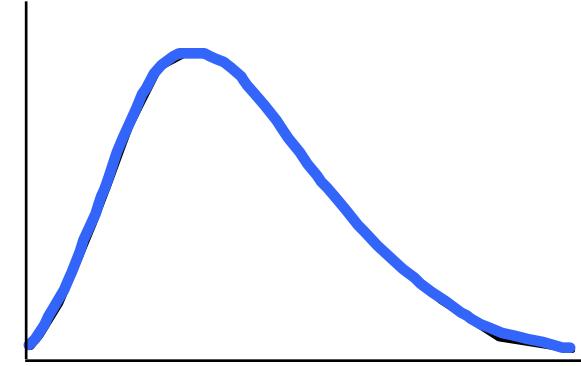
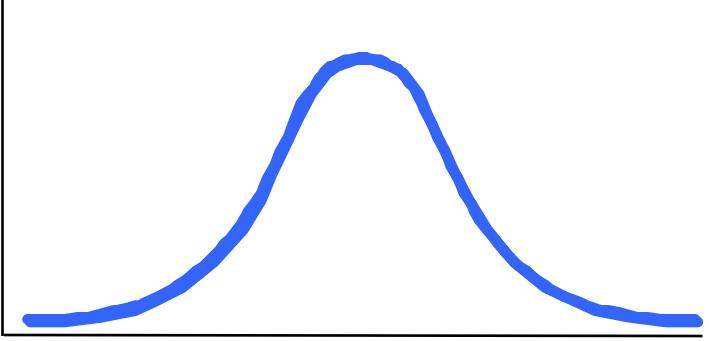
Symmetric



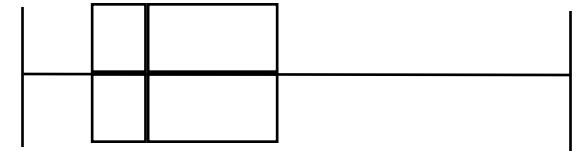
Right-Skewed



Boxplots



Normal (Gaussian)

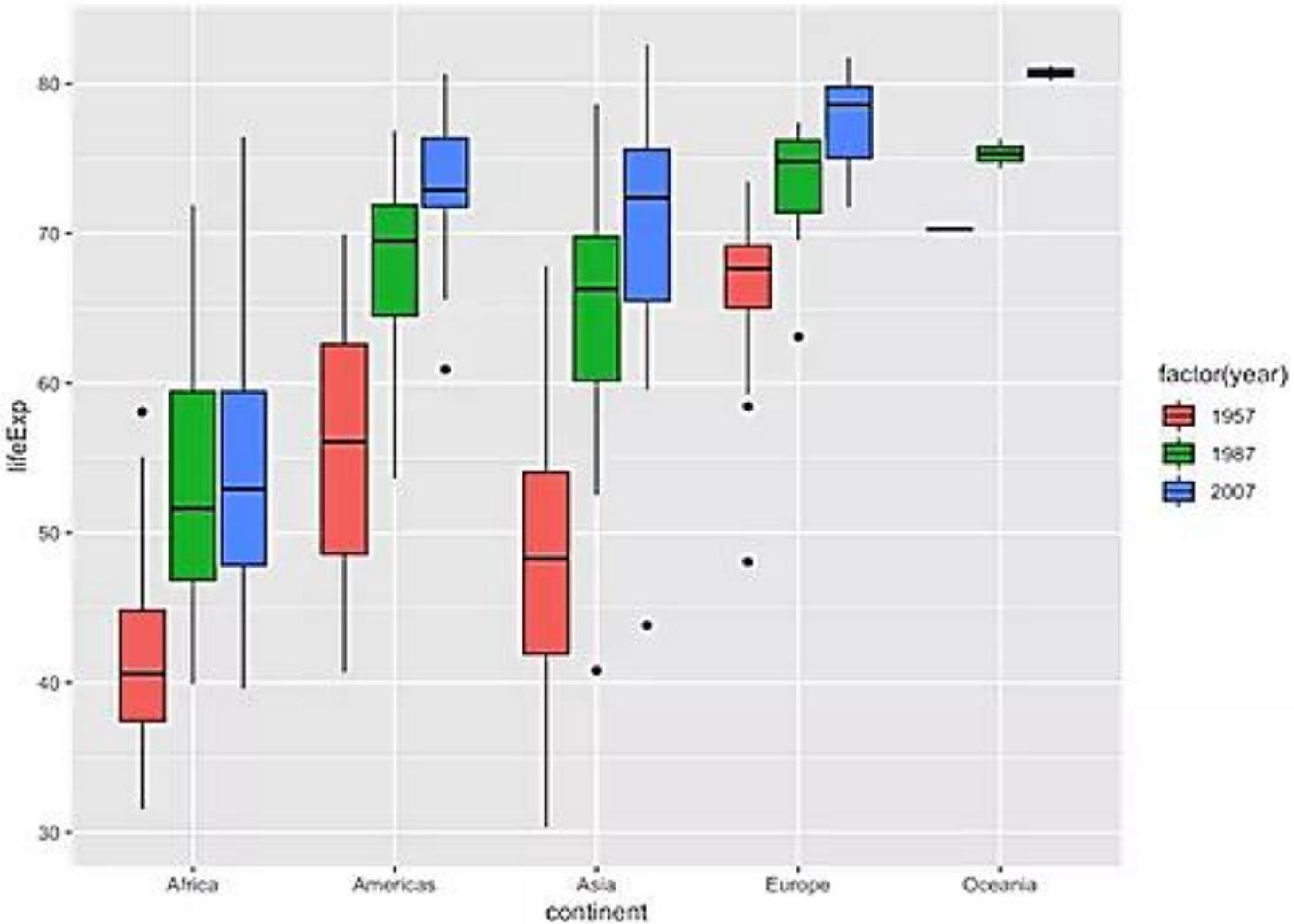
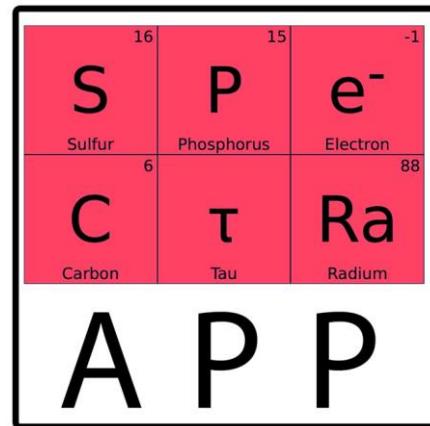


Skewed

Boxplots

Optimal for:

- Identify **anomalous data (outliers)**;
- Representing one or more **populations**;
- Quickly evaluate **normality**;
- Comparing one or more **groups**.



but... data pre-processing
first!



Pre-processing



THE IMPORTANCE
OF BEING EARLY

OSCAR WILDE



pre-processed

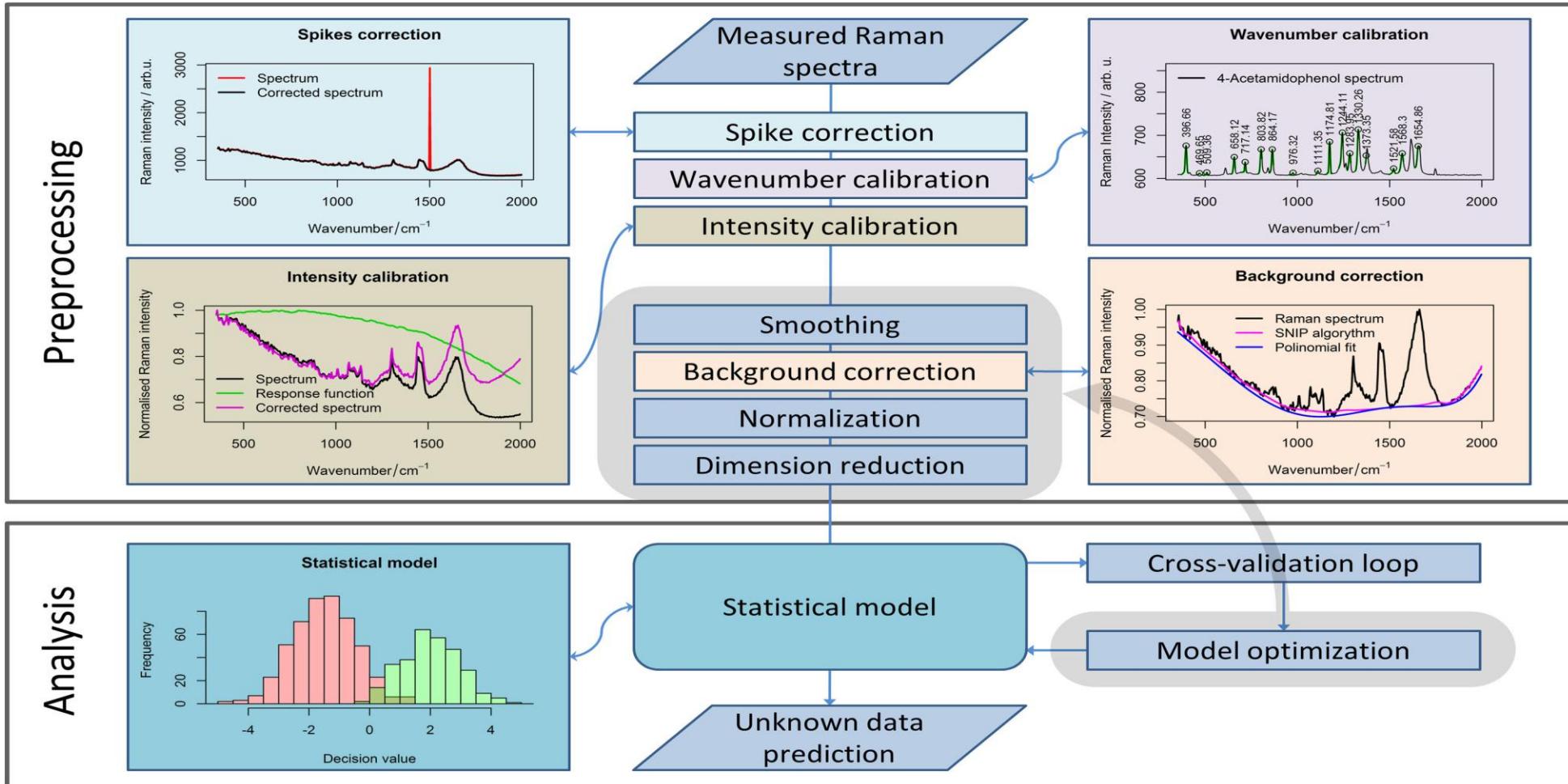


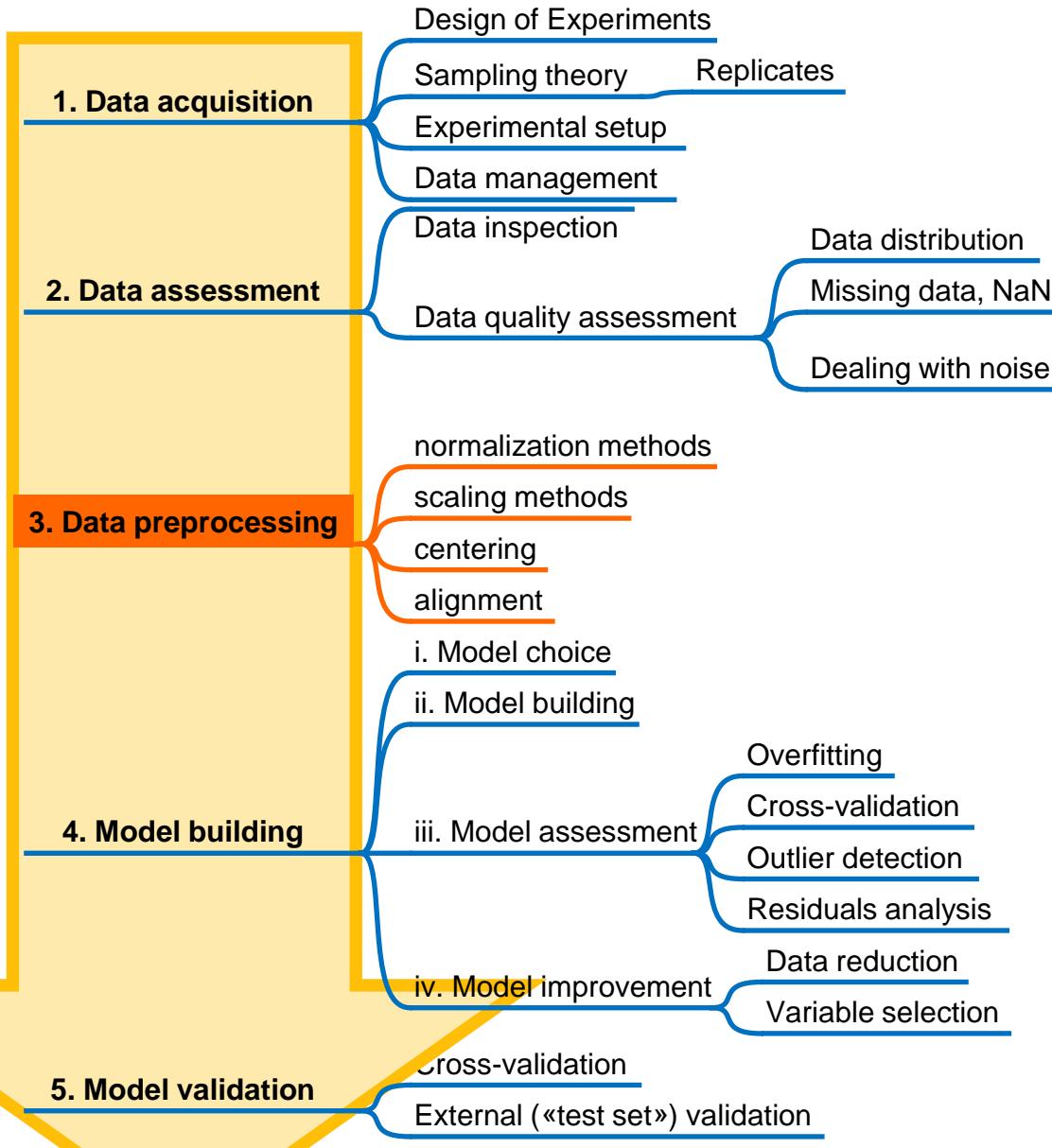
Why pre-processing?

ML models

data pre-processing

Pre-processing





Preprocessing: “Generic term for methods to go from raw instrumental data to clean data for data processing.”
Goodacre R. et al. (2007)

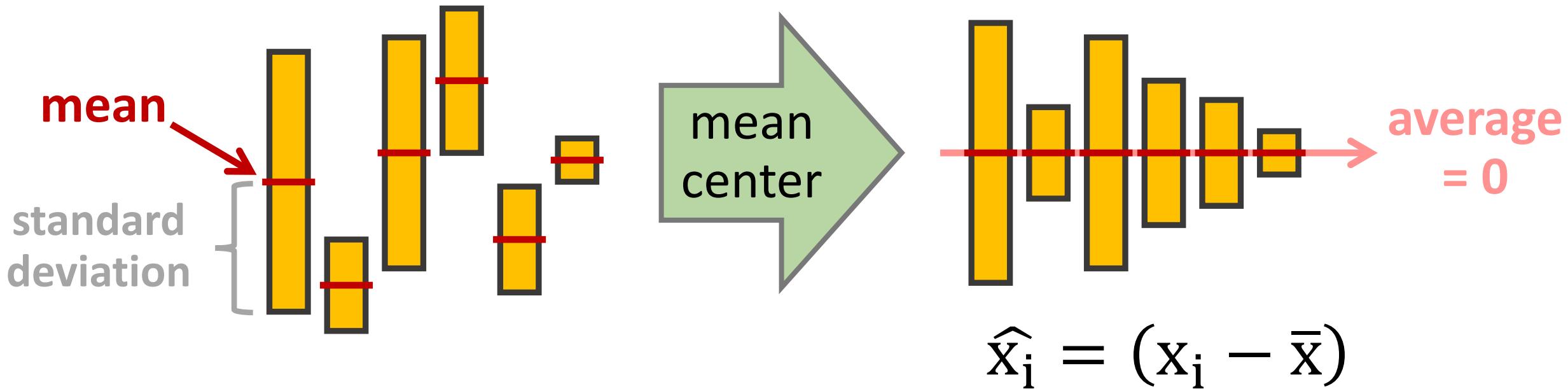
Why do we need to preprocess our data?

- To “clean” the data by removing
 - noise
 - offsets
 - unwanted effects/defects
- Correct for batch effects or instrumental drifts, ...

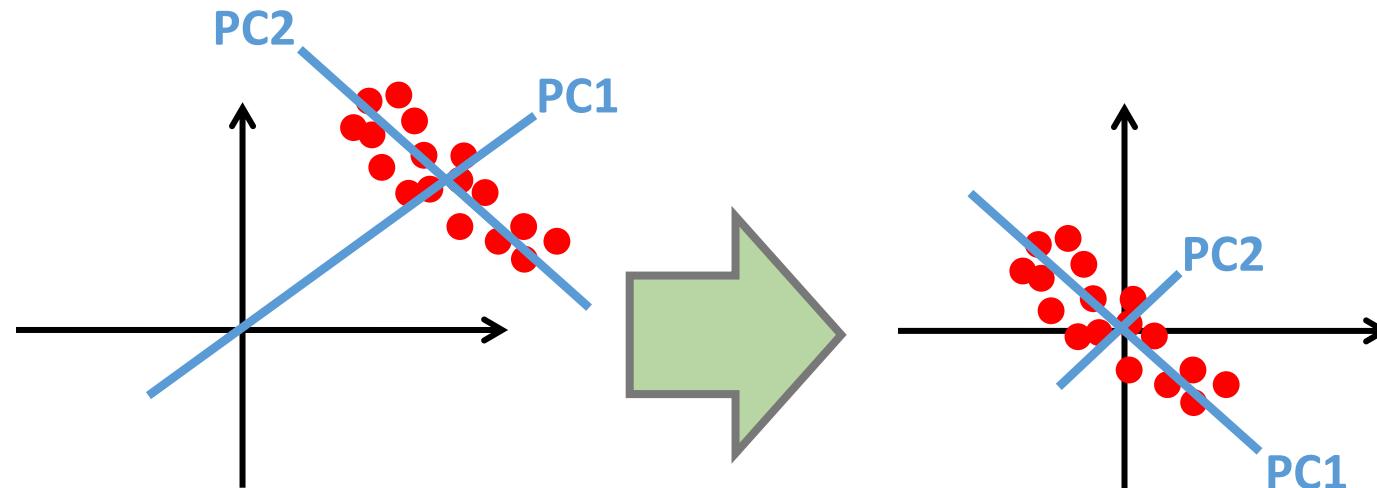
Column-wise pre-processing

Mean centering

For each value the mean of its column is subtracted.



Effect on PCA modelling:



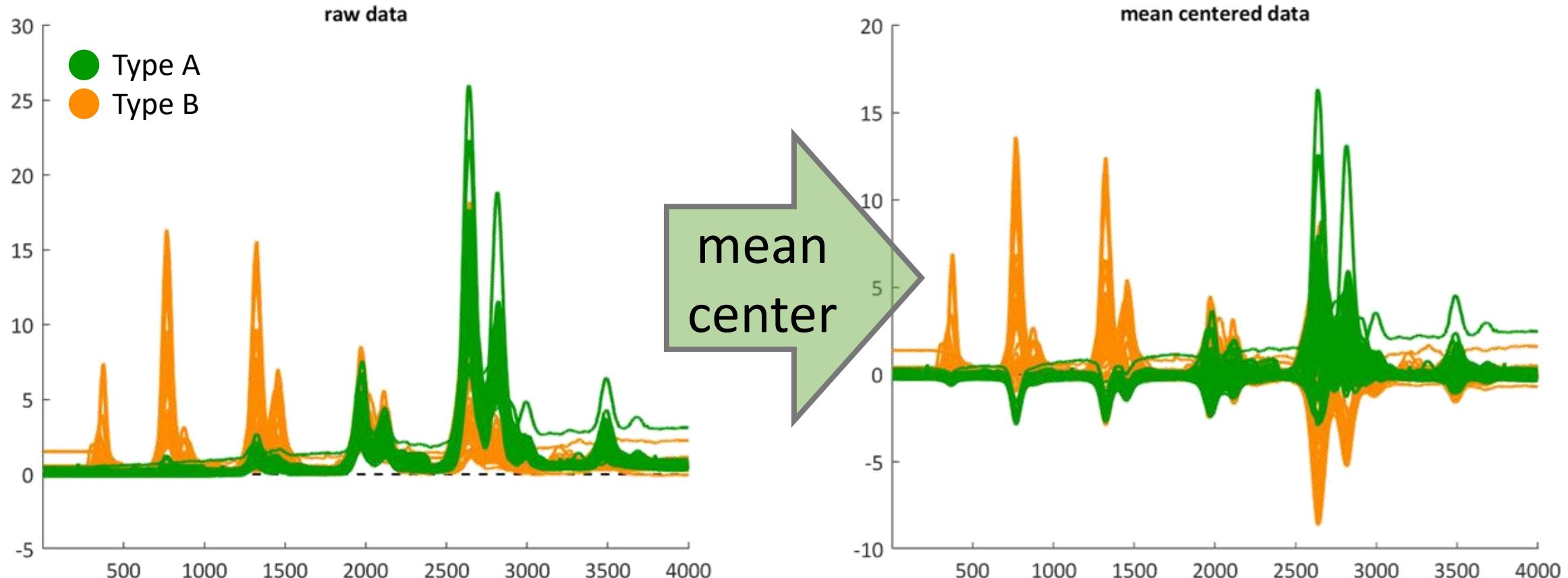
Mean centering

Continuous data: 73 samples \times 15 physical–chemical properties

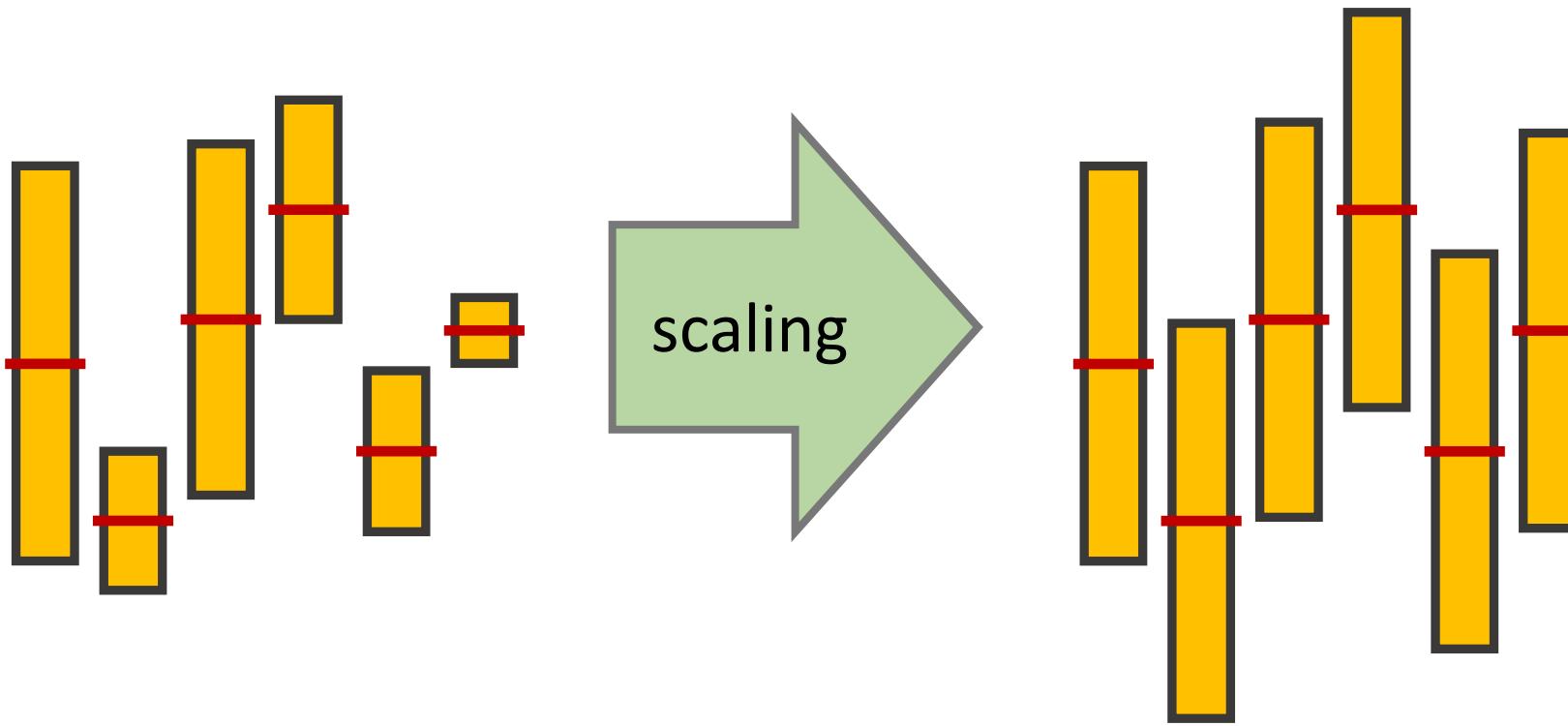


Scaling

Continuous data: 115 samples \times 4001 retention time values



Scaling

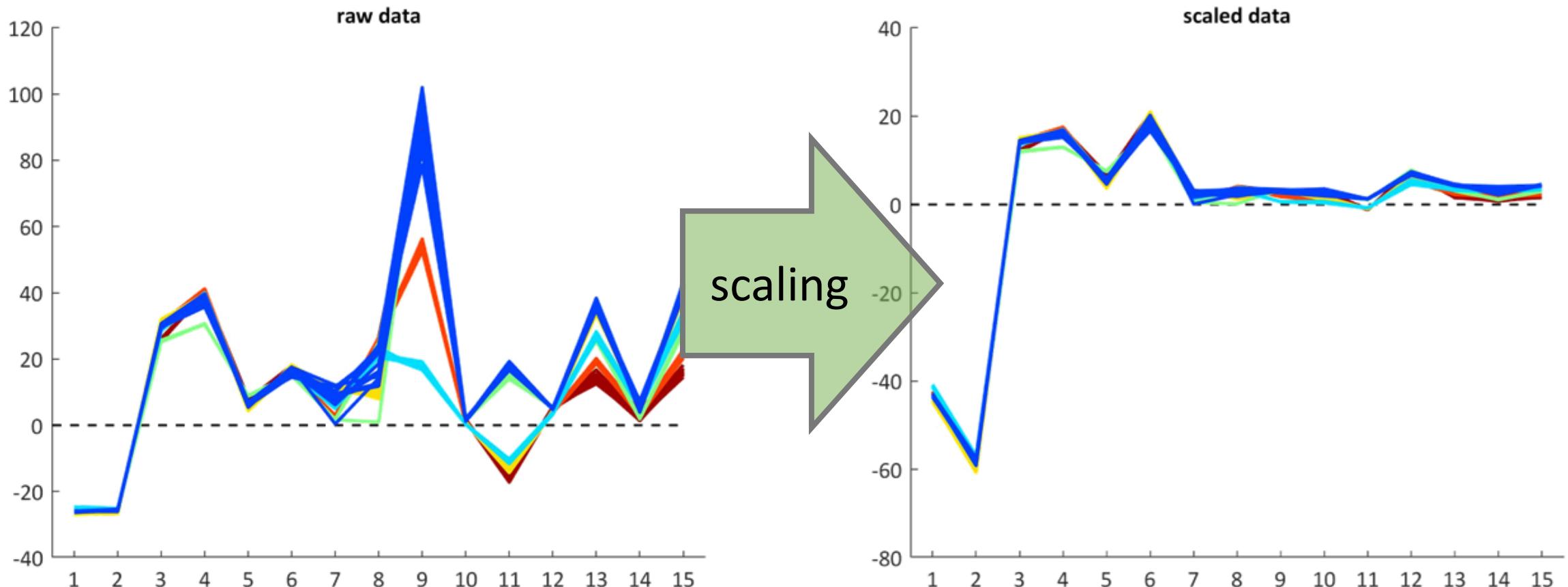


$$\hat{x}_i = \frac{x_i}{S_{x_i}}$$

standard deviation
of the column

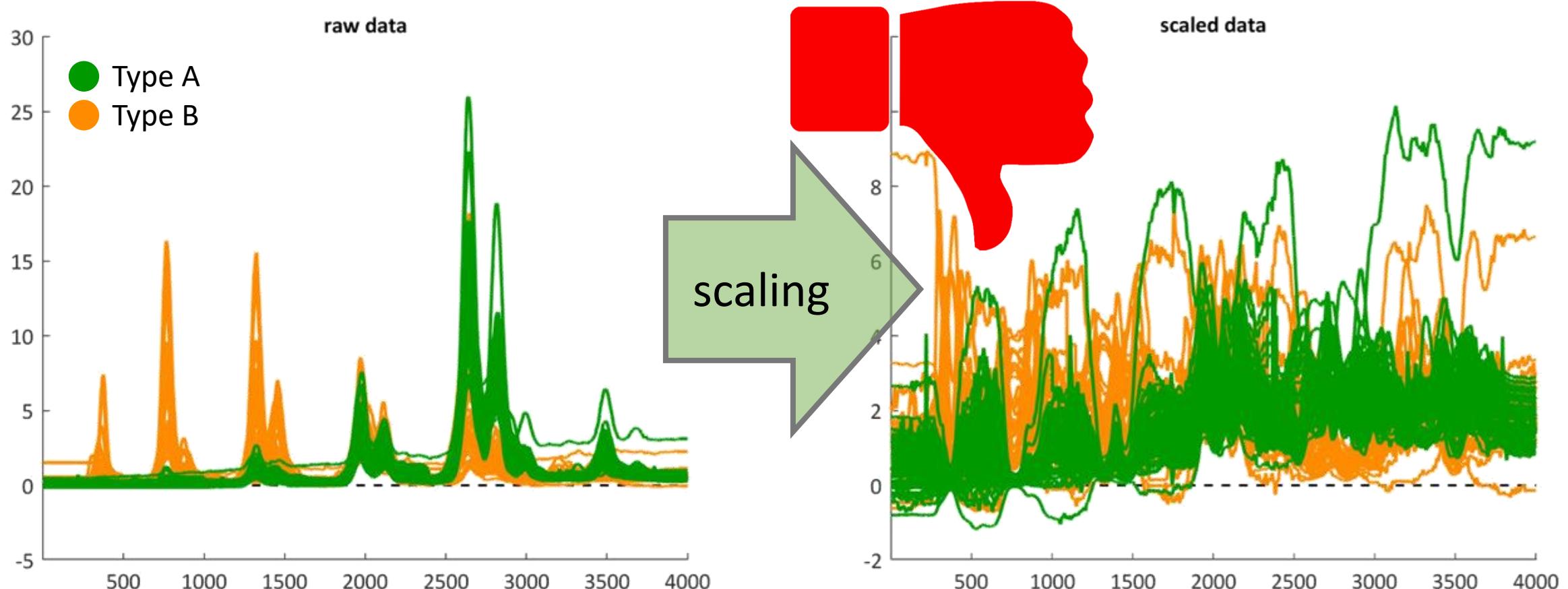
Scaling

Continuous data: 73 samples \times 15 physical–chemical properties



Scaling

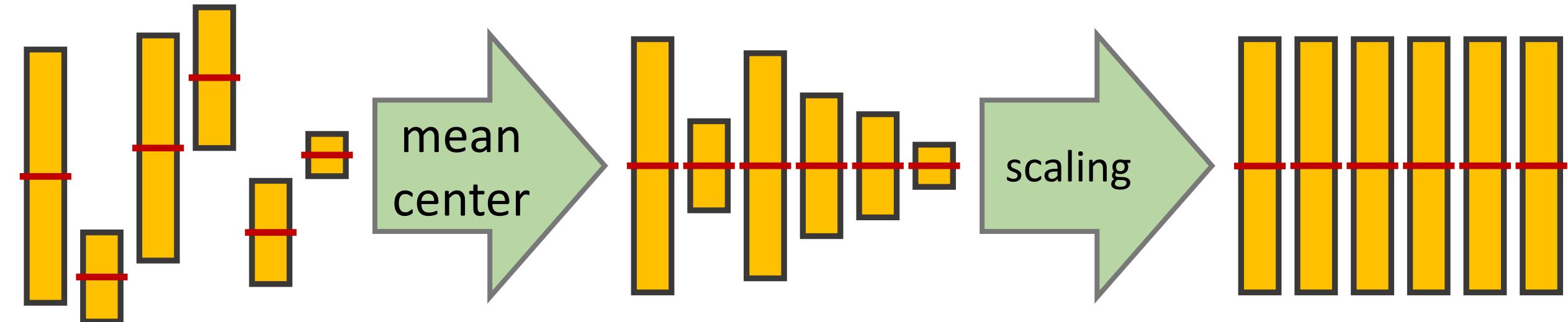
Continuous data: 115 samples \times 4001 retention time values



Autoscaling

Also called “unit variance scaling”.

To remove **differences in measurements units!**

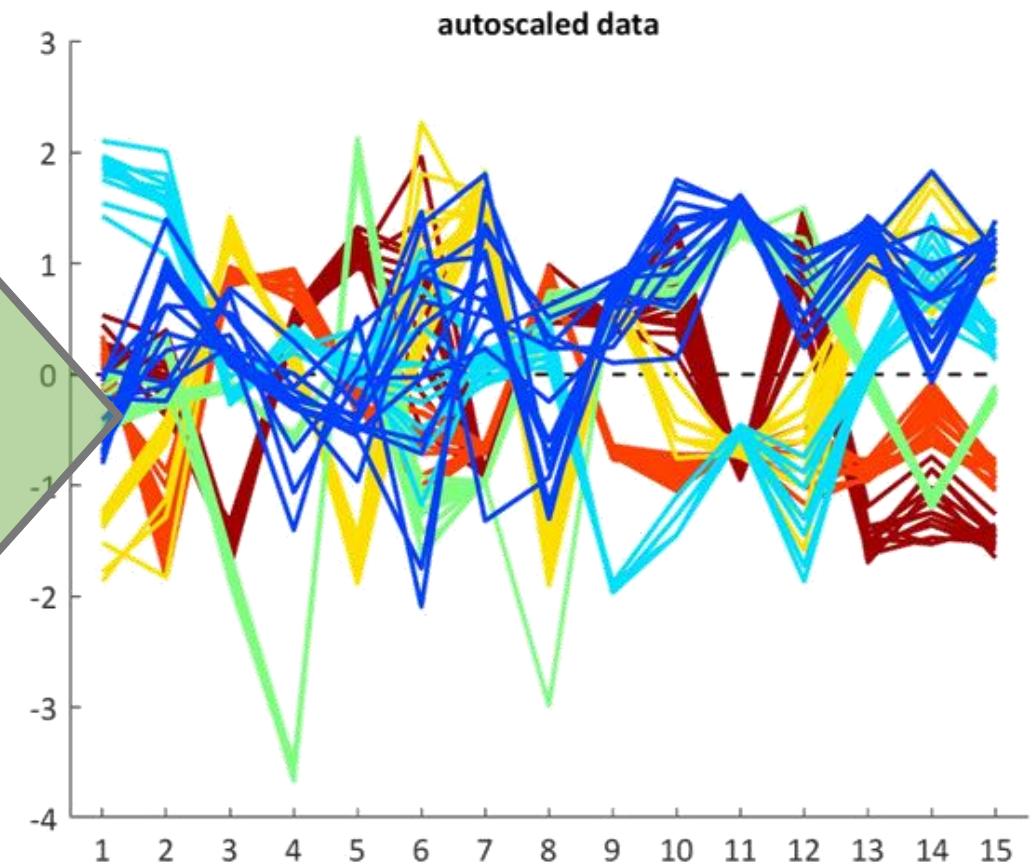
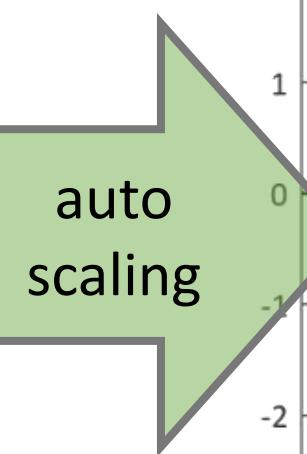
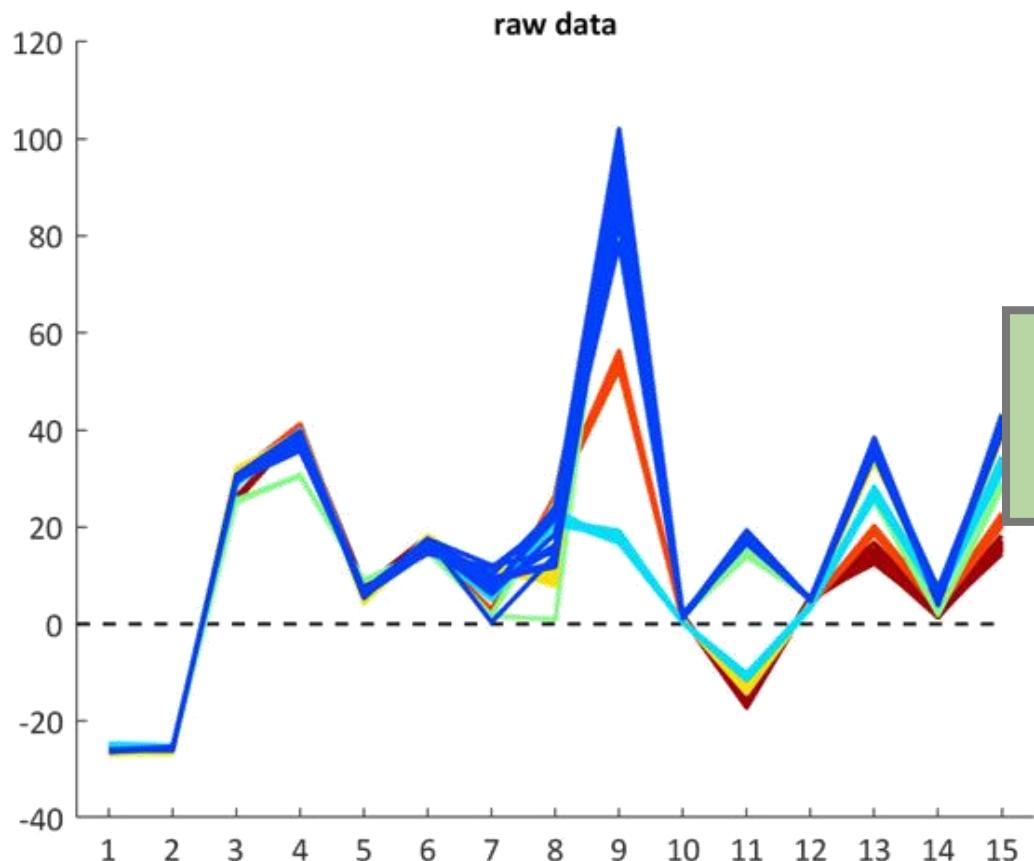


$$\hat{x}_i = (x_i - \bar{x})$$

$$\hat{x}_i = \frac{x_i}{S_{x_i}}$$

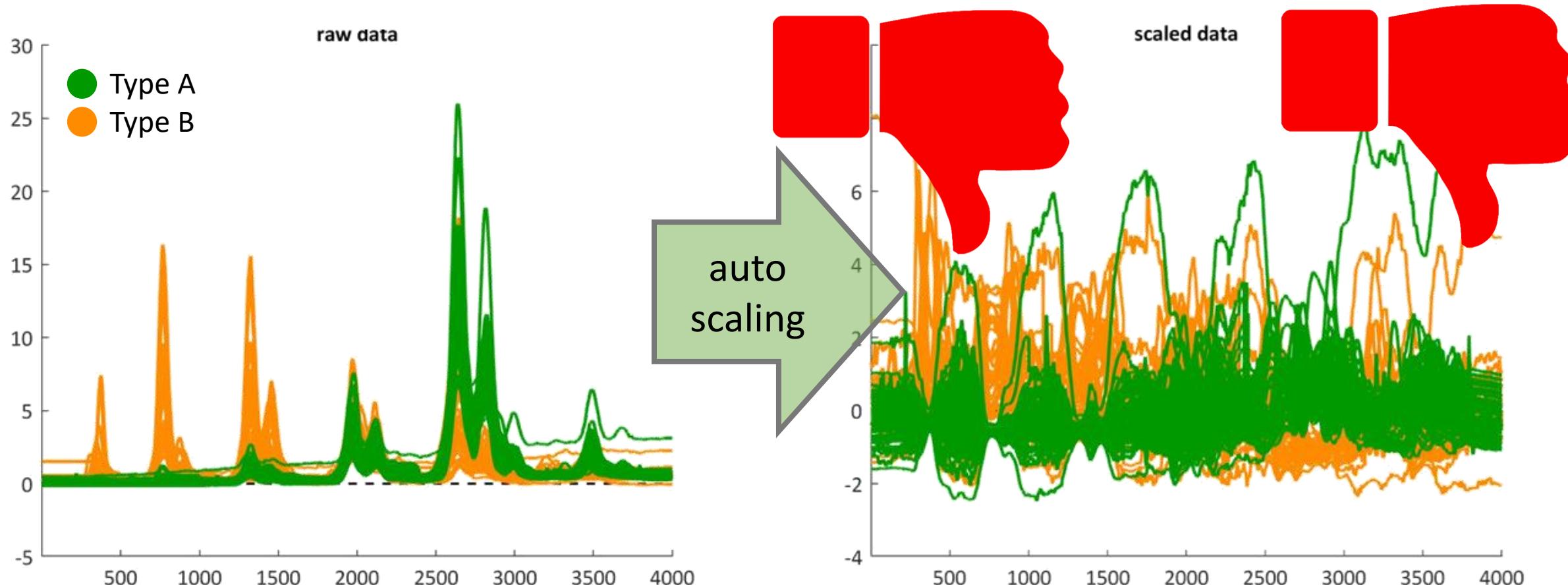
Autoscaling

Continuous data: 73 samples \times 15 physical–chemical properties

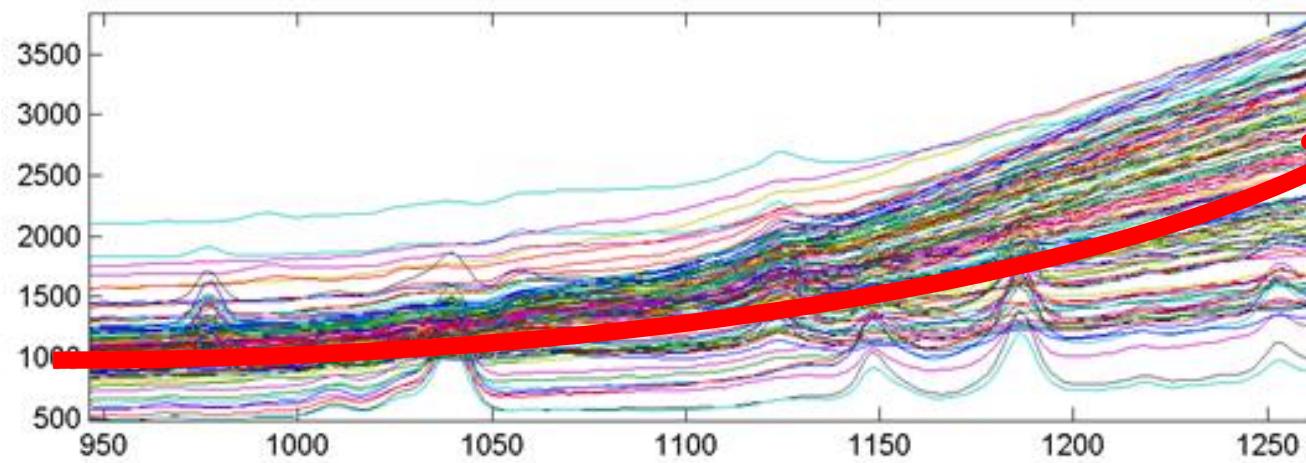


Autoscaling

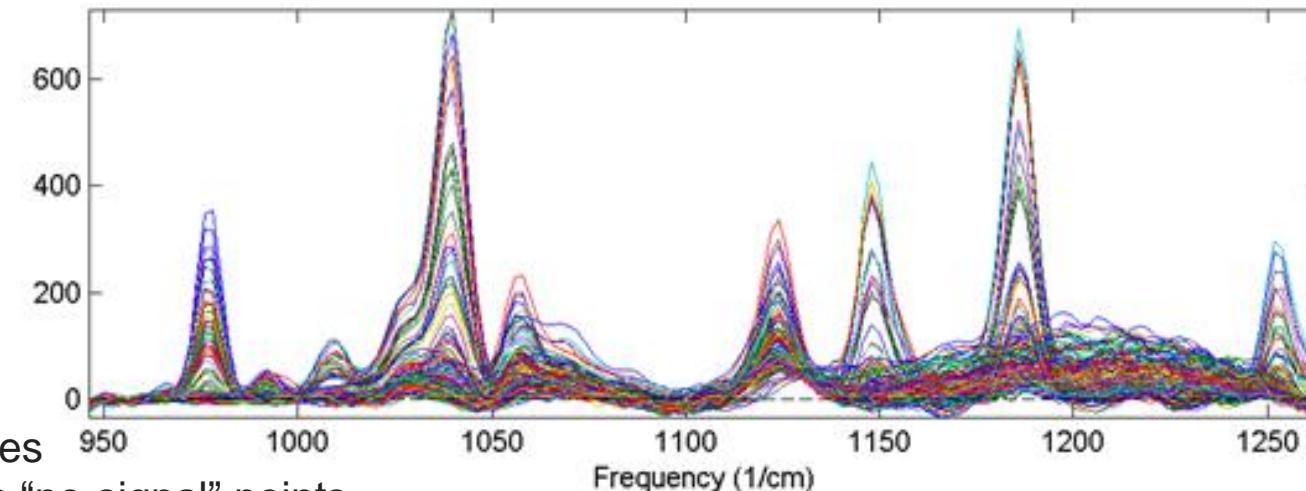
Continuous data: 115 oil samples \times 4001 retention time values



Baseling removal



curved
baseline



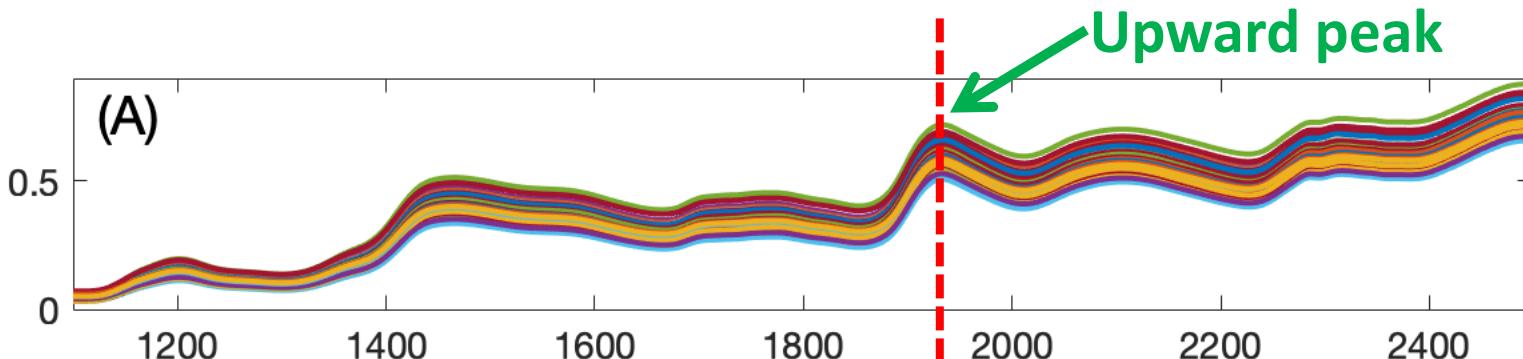
peaks can
“emerge”

Some methods:

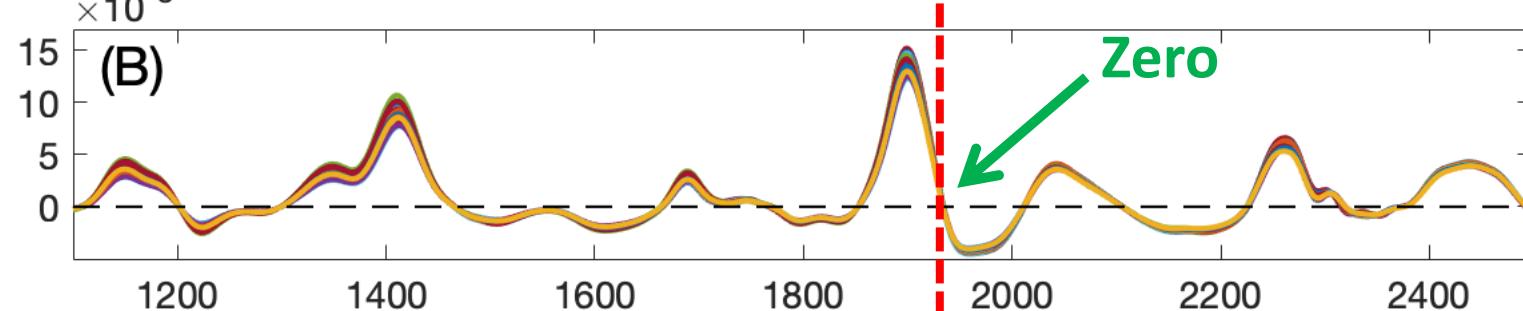
- Detrending
- Weighted Least Squares
- Estimation from known “no-signal” points

Derivatives (Savitzky-Golay method)

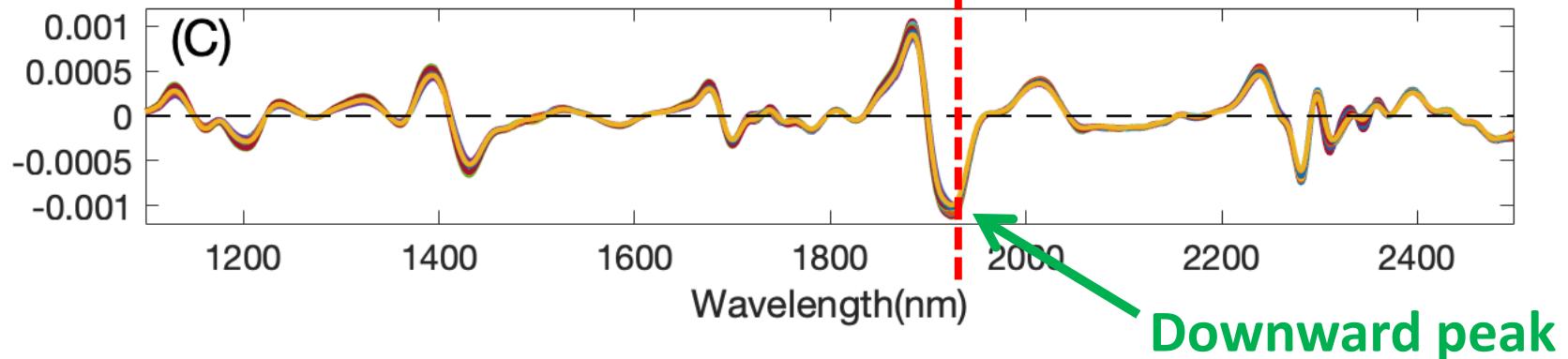
raw spectra (NIR)



1st derivative



2nd derivative



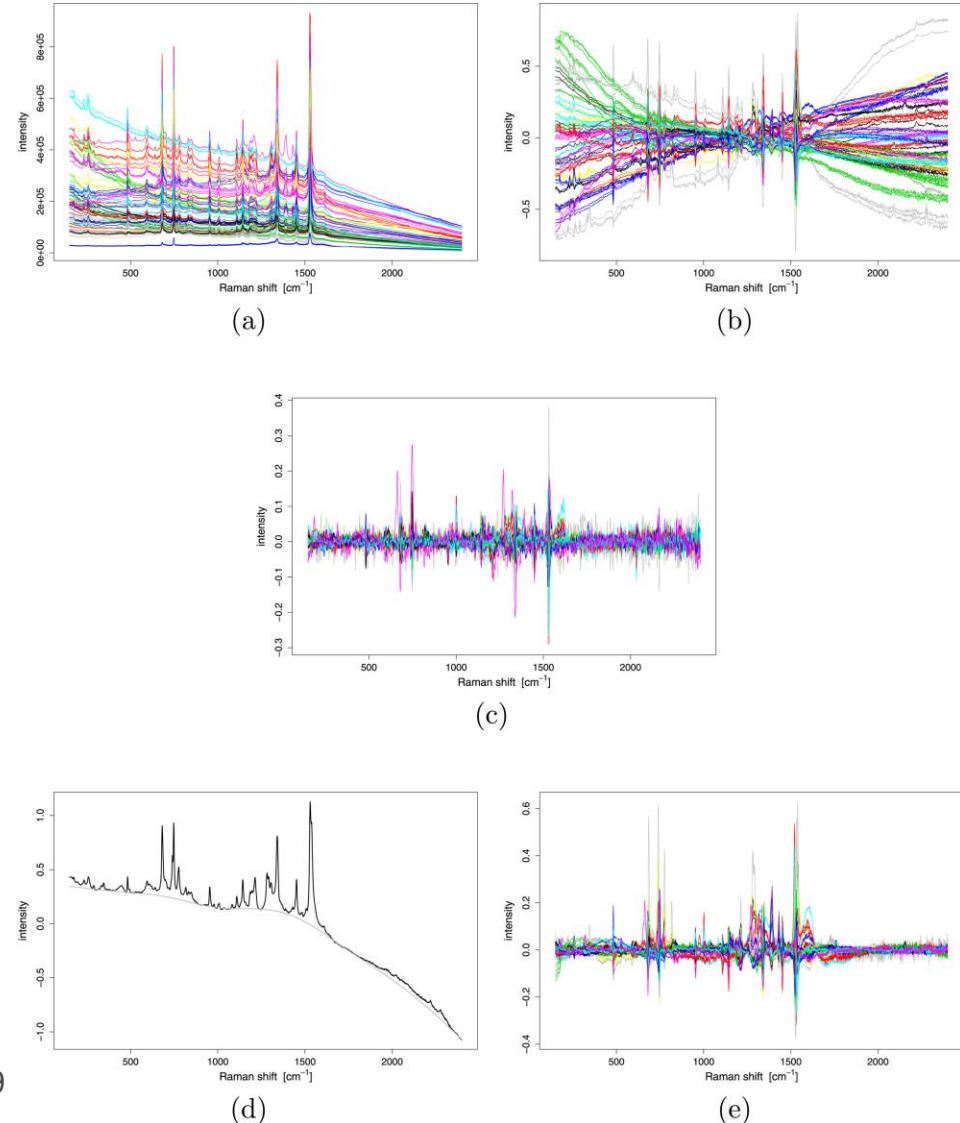
row

Row-wise pre-processing

Pre-processing in spectroscopy

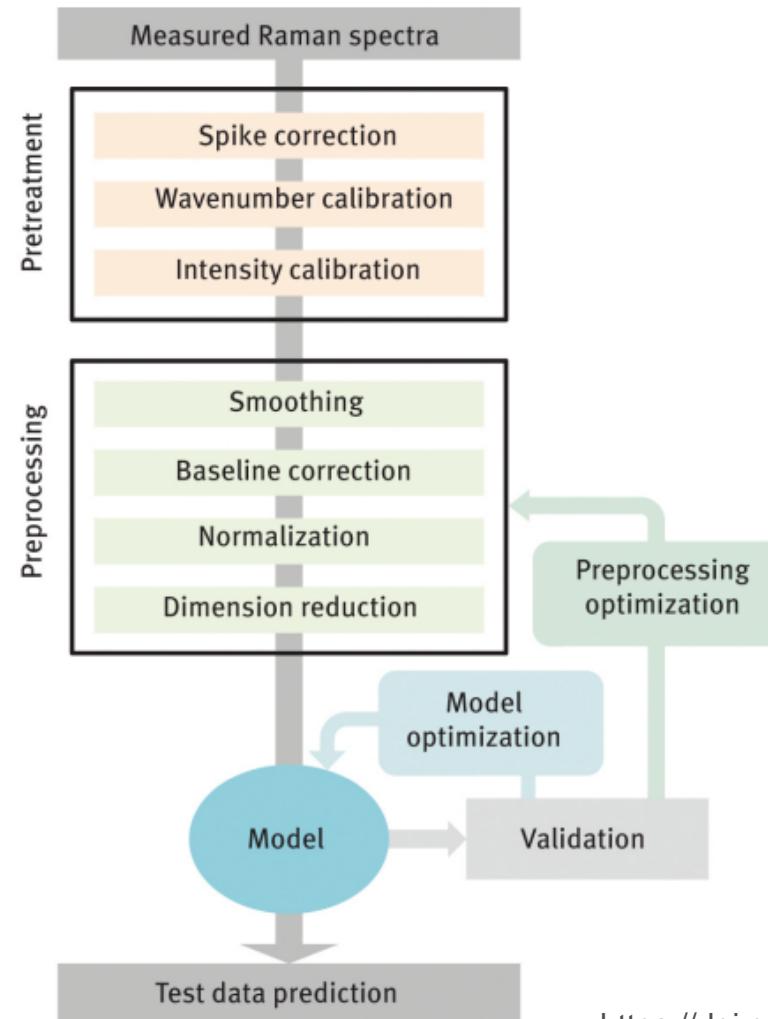
Pre-processing of spectra:

- Spike removal;
- Intensity and wavenumber calibration;
- Smoothing;
- Background correction;
- Normalization;
- and more...



Pre-processing in spectroscopy

A further distinction:



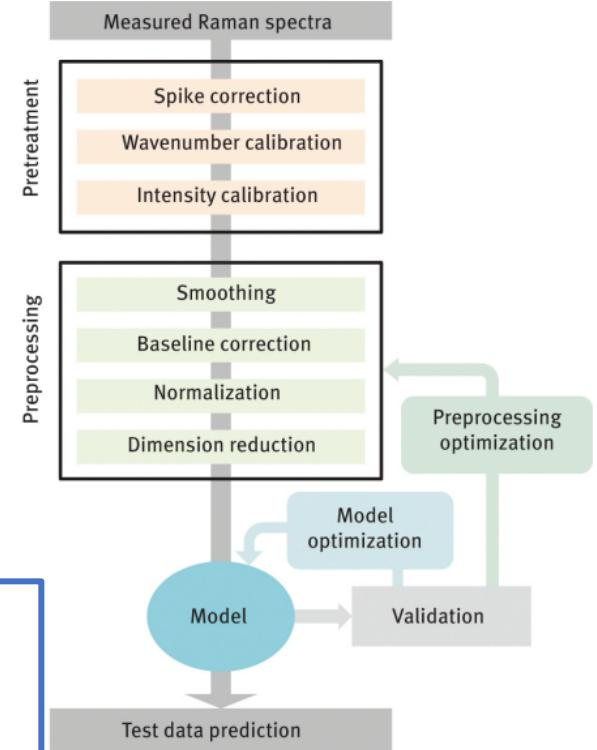
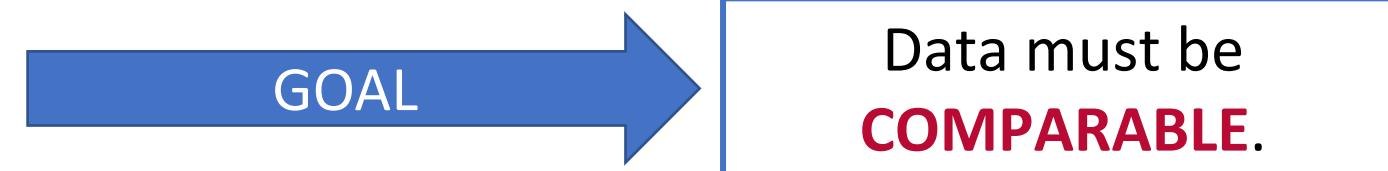
<https://doi.org/10.1515/psr-2017-0043>

Pre-processing in spectroscopy

Why?

Reduce or eliminate irrelevant, random and systematic variations in the data:

- signal intensity variations (laser intensity);
- low signal-to-noise ratio;
- high background (fluorescence);
- spikes;
- etc.



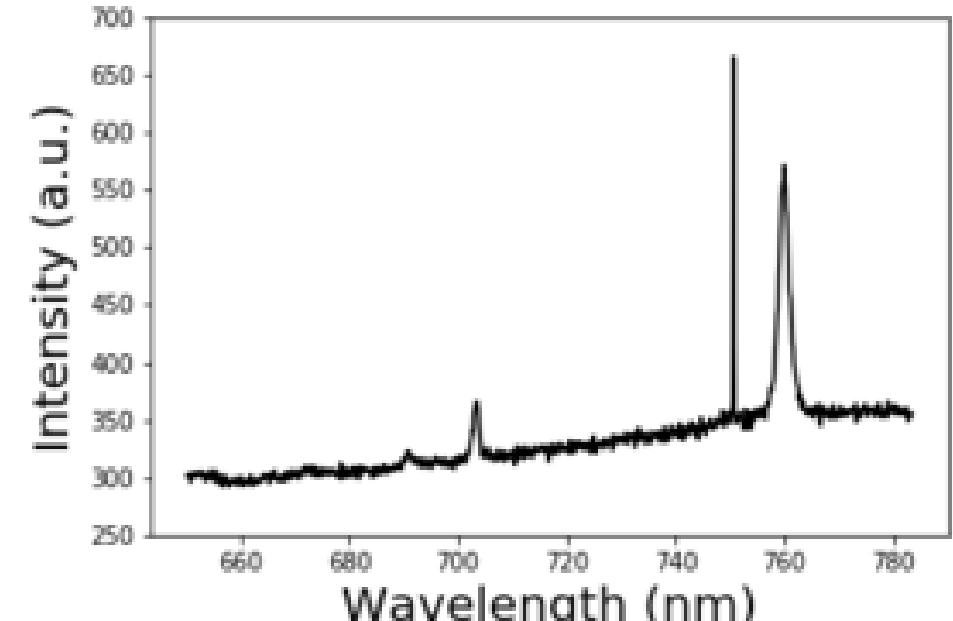
Spike removal

Spikes:

- single and random events;
- positive peaks with narrow bandwidth

How to identify?

- Spike band width much smaller than Raman band width;
- spatially adjacent spectra are similar.

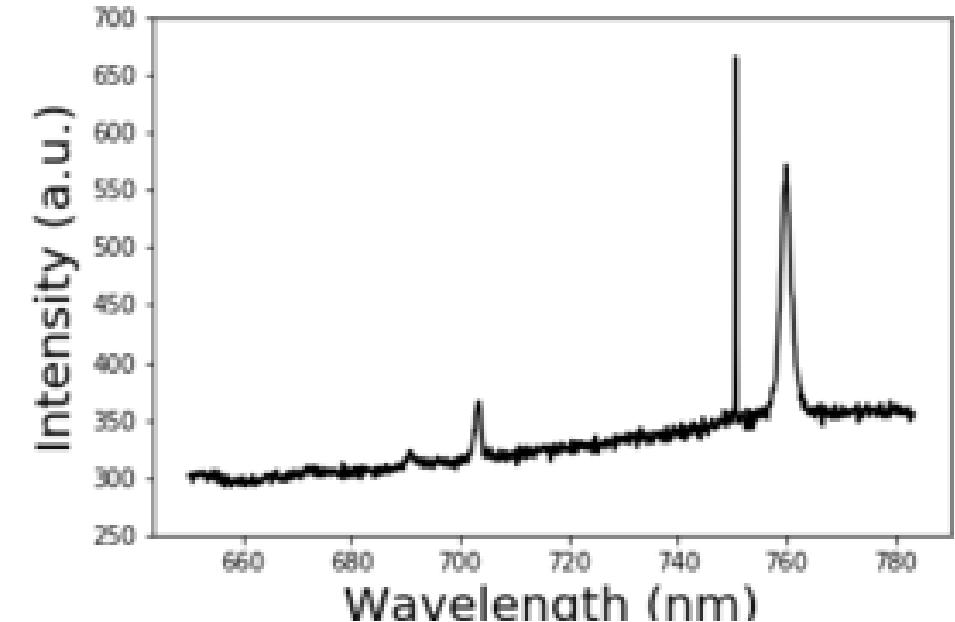


<https://towardsdatascience.com/removing-spikes-from-raman-spectra-8a9fdda0ac22>

Spike removal

Strategies:

- visually/manually (remove the spike through comparison);
- mathematically:
 - missing point polynomial filter
 - robust smoothing filter
 - moving window filter
 - wavelet transform methods
- reduction of spike events by special design of the instrument



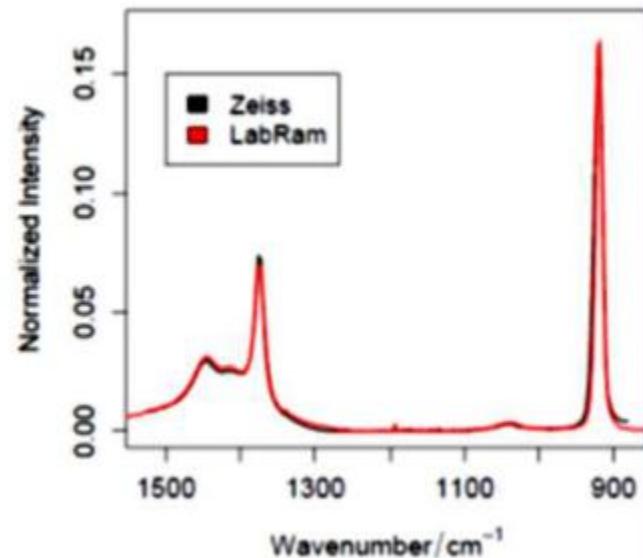
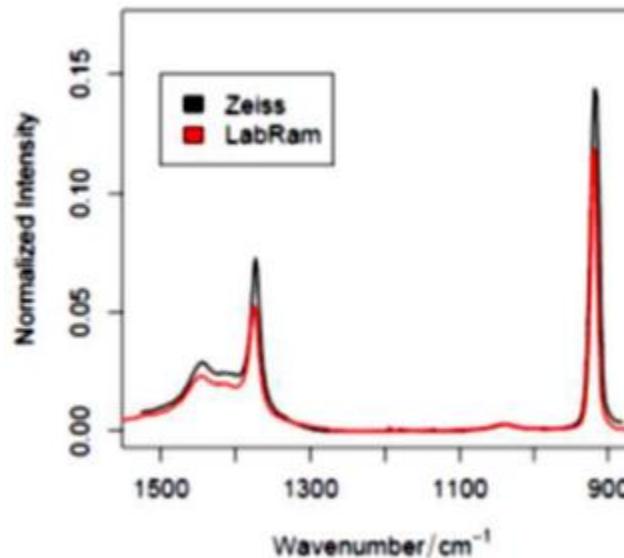
<https://towardsdatascience.com/removing-spikes-from-raman-spectra-8a9fdda0ac22>

Intensity calibration

Corrects the recorded Raman spectrum with the intensity response of a SRM

- Intensity-calibrated lamps;
- NIST fluorescence standards for Raman.

Instrument variation can be reduced,
but very hard to completely eliminate

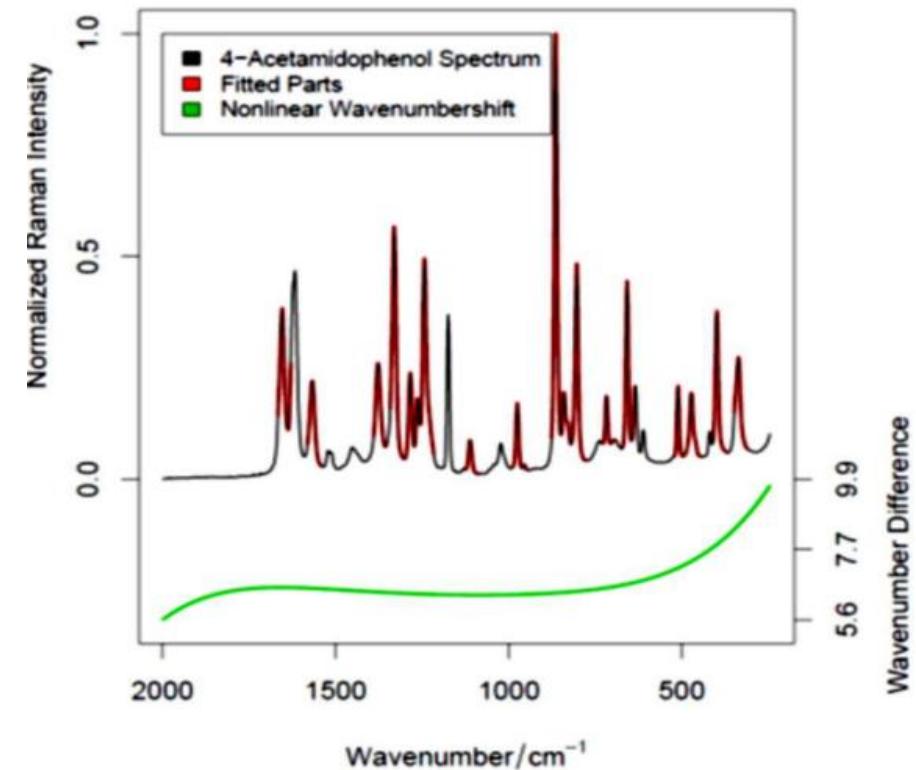


Wavenumber calibration

Find the correct wavenumber axis and remove the wavenumber shifts

- Atomic emission lines (Ne, Hg, Ar, Kr lamps, plasma lines of the gas lasers);
- Shift calibration standards (indene, polystiren, paracetamol, naphthalene, etc.).

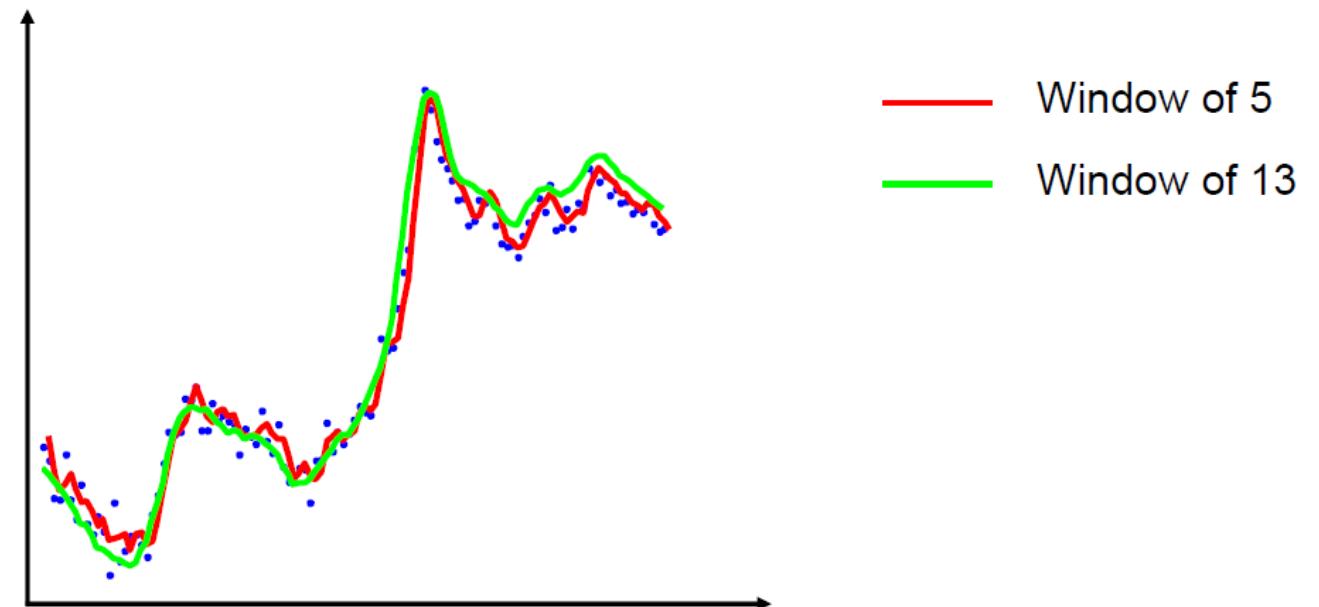
Instrument variation can be reduced,
but very hard to eliminate



Smoothing

It is (typically) based on the interpolation in small windows of a ***polynomial of n degree*** (Savitzky-Golay methodology) but there are several methods:

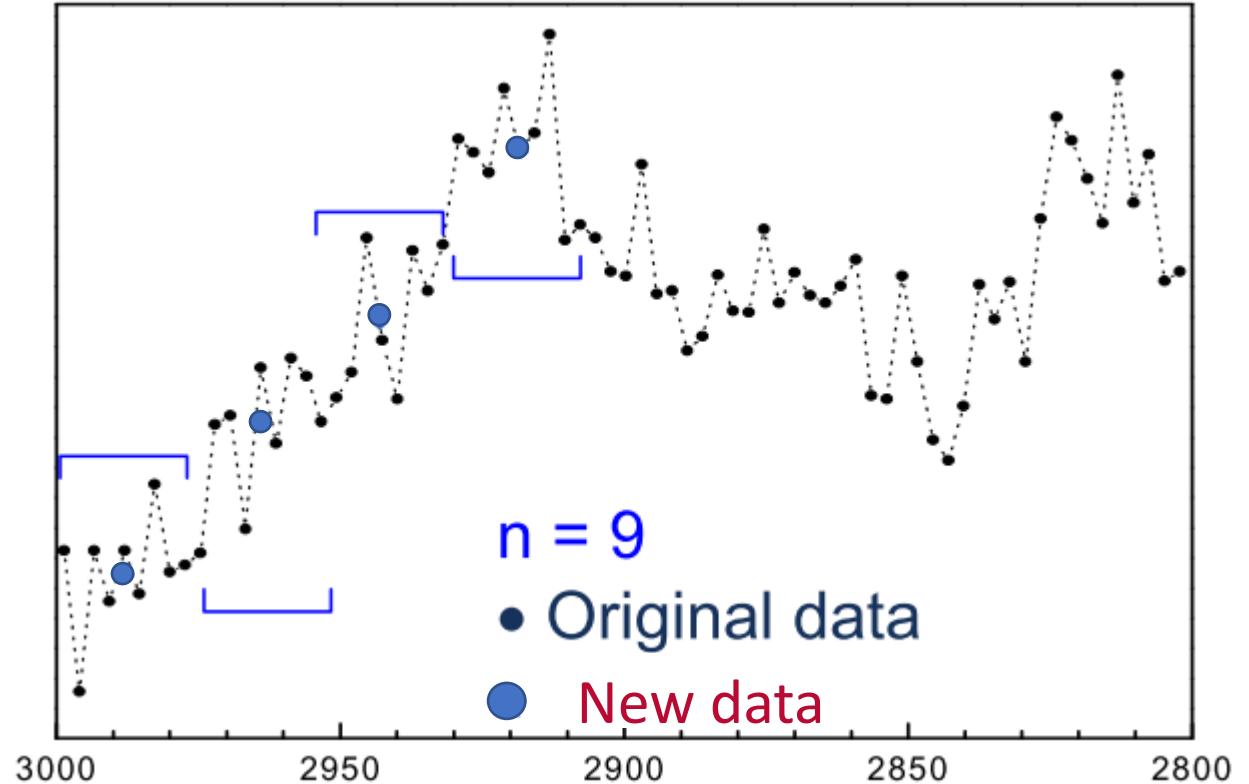
- Average;
- Moving Average;
- Moving Median;
- Moving Polynomial (Savitzky-Golay);
- Fourier Filter.



Smoothing

Average smoothing:

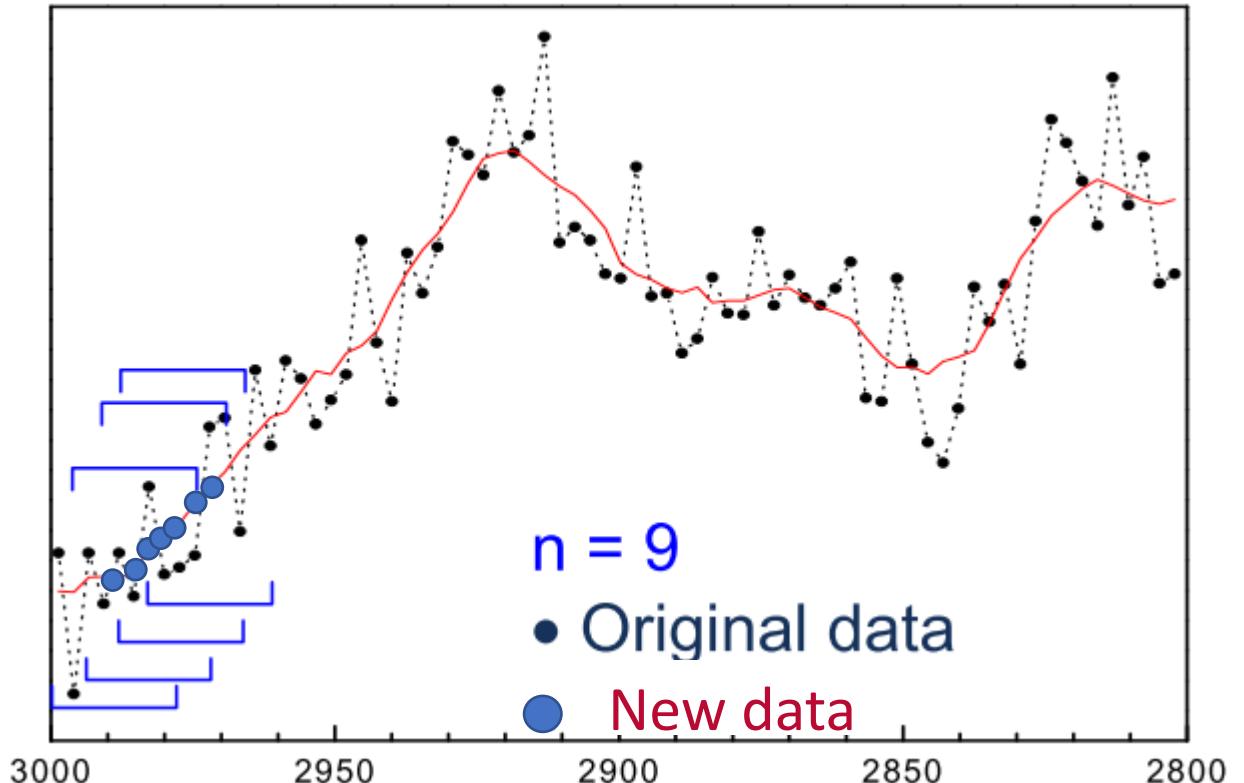
- reduces the data points;
- information can be lost/altered



Smoothing

Moving Average:

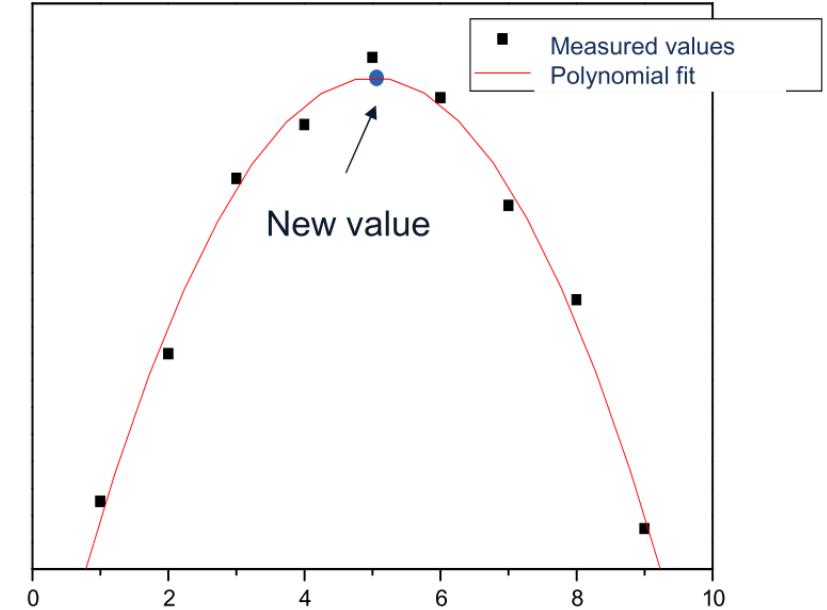
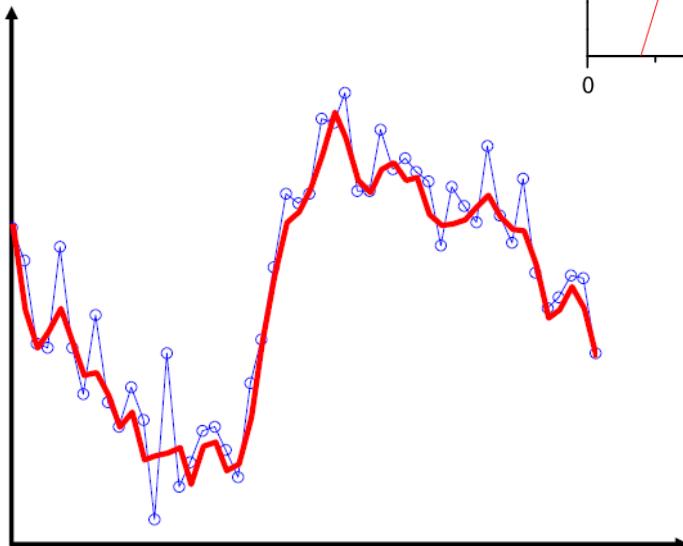
- data points number remains;
- information loss increases with increasing moving window size;
- 'tails' effect.



Smoothing

Savitzky-Golay:

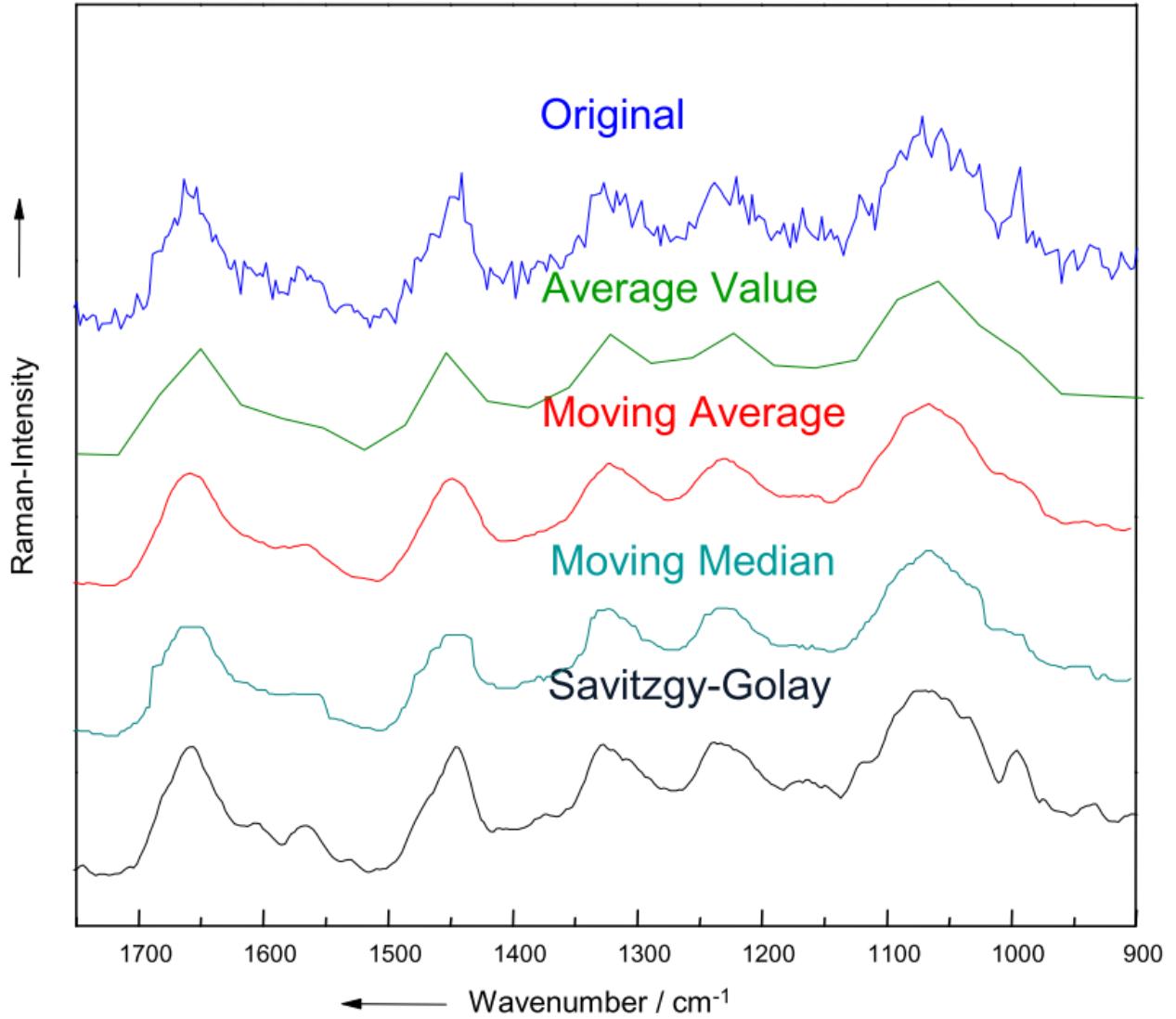
- a polynomial of small orders ($n=1, 2, 3$) fits the data;
- information loss increases with increasing moving window size.



Smoothing

Window of 13 points used with different smoothing algorithms

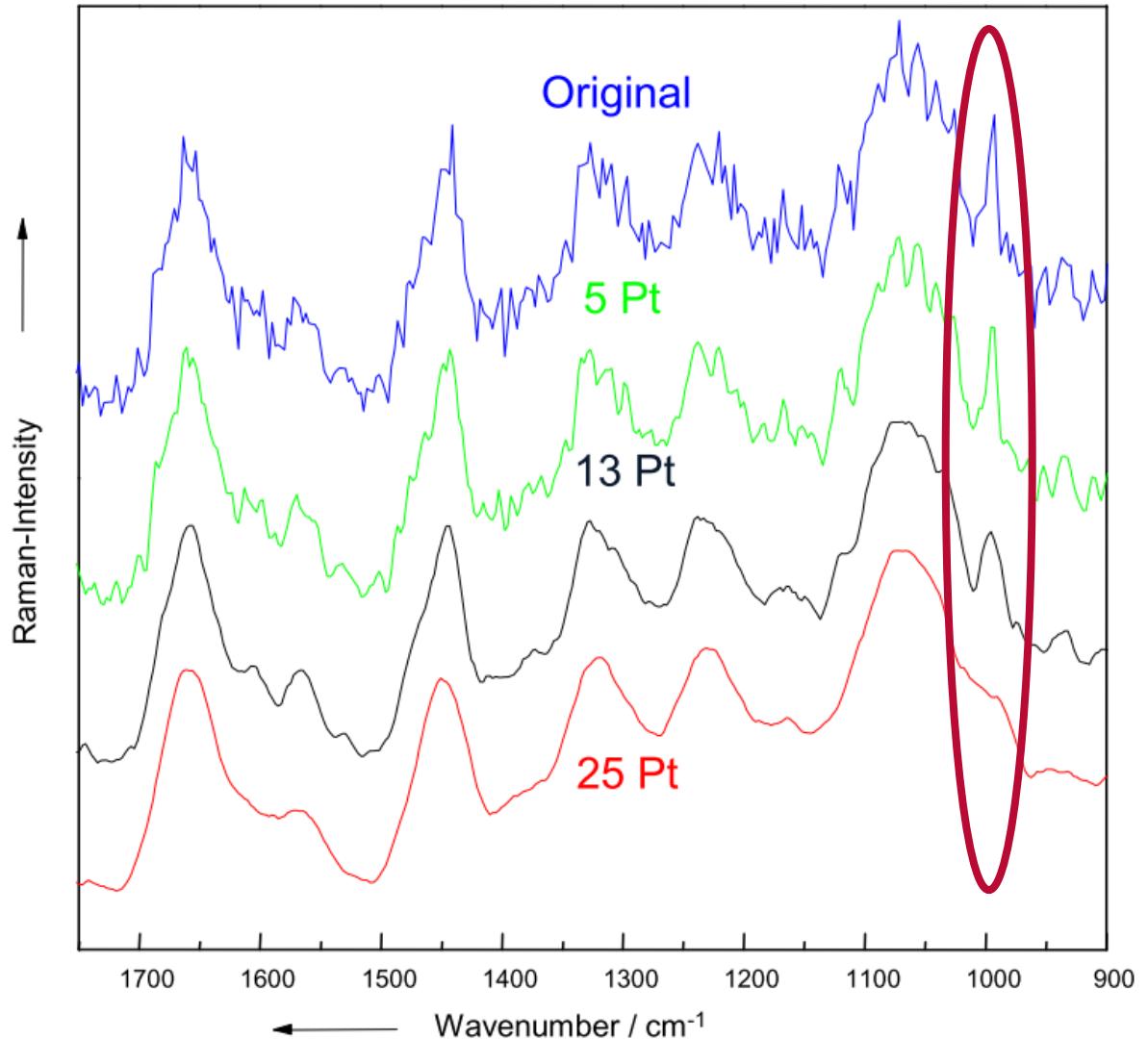
If the band shape is important then use Savitzky-Golay smoothing:
 it preserves the band shape better than the other methods



Smoothing

Different smoothing window sizes
for Savitzky-Golay method

Width of Raman bands increases with the
smoothing window (and the loss of
information increases, too)



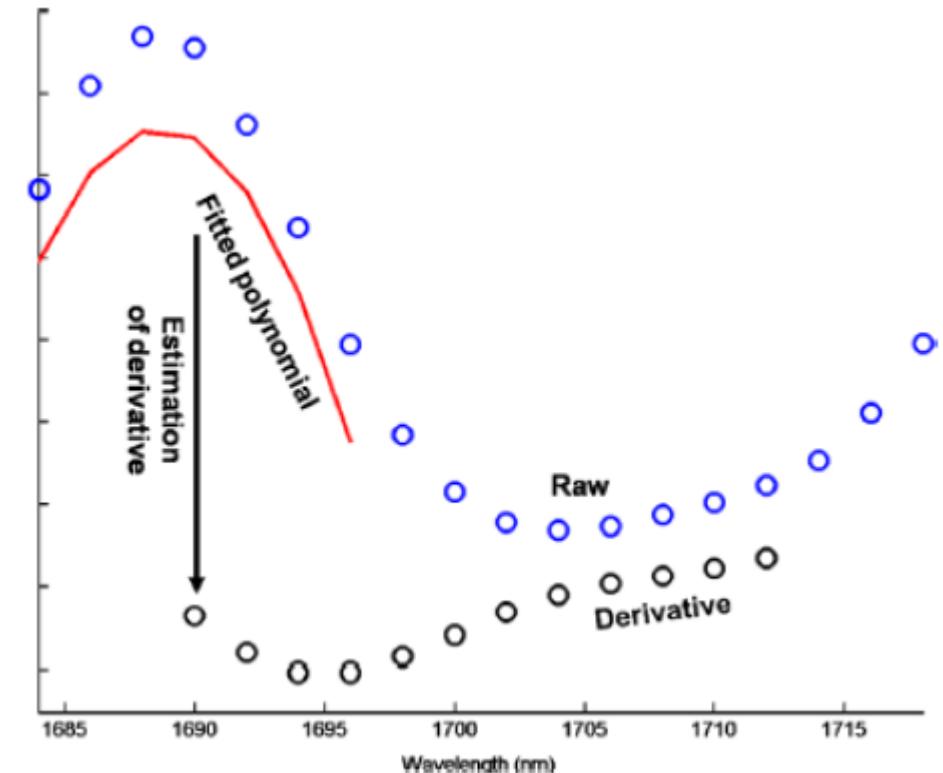
Background removal

Derivatives:

They **minimize noise and transform the spectra**.

There are three parameters to optimize:

- **Window size:** smoothing step;
- **Polynomial degree:** second degree preferred;
- **Derivative degree:** risk of 'creating' new peaks.

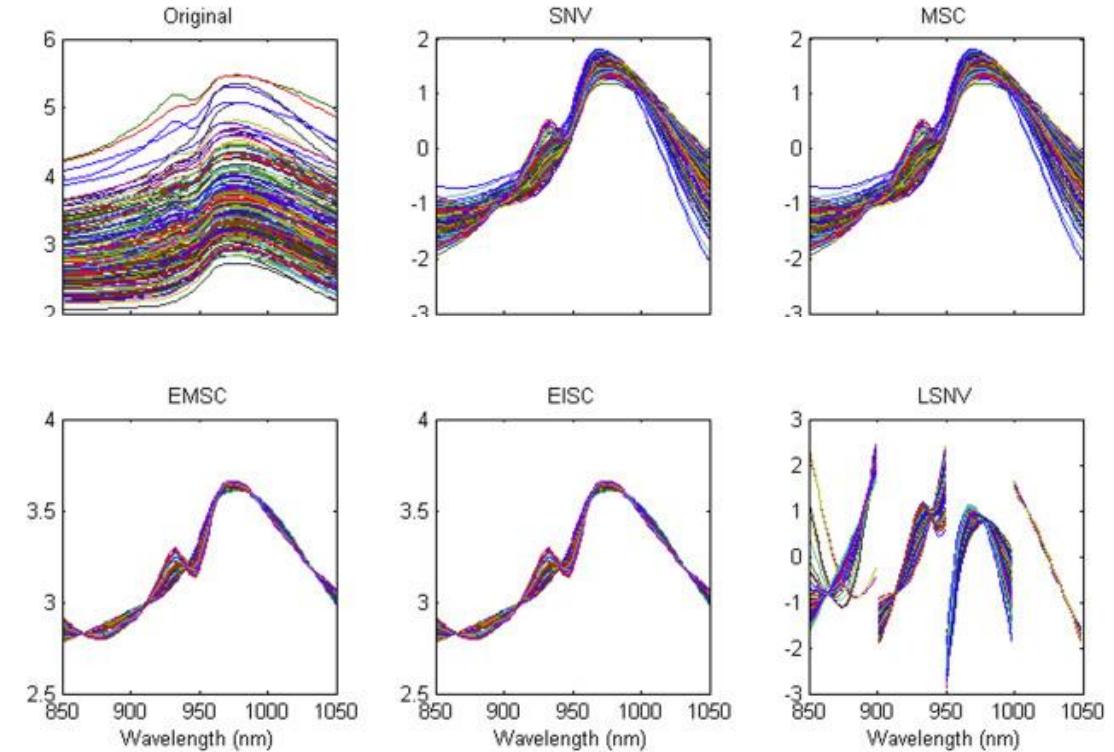


Normalization

Eliminate systematic differences among measurements

Raman spectra show intensity differences because of:

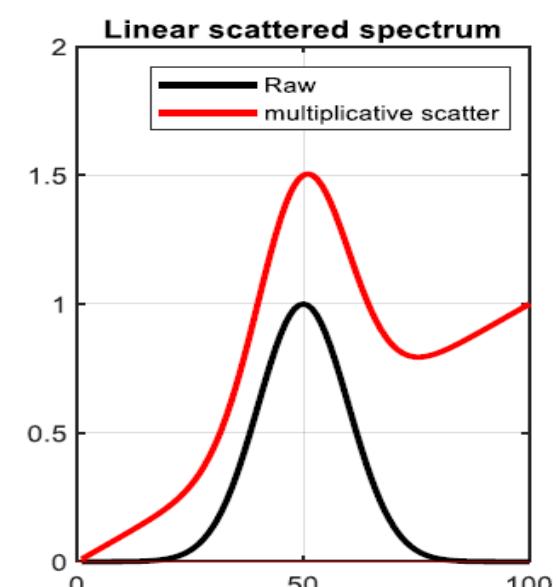
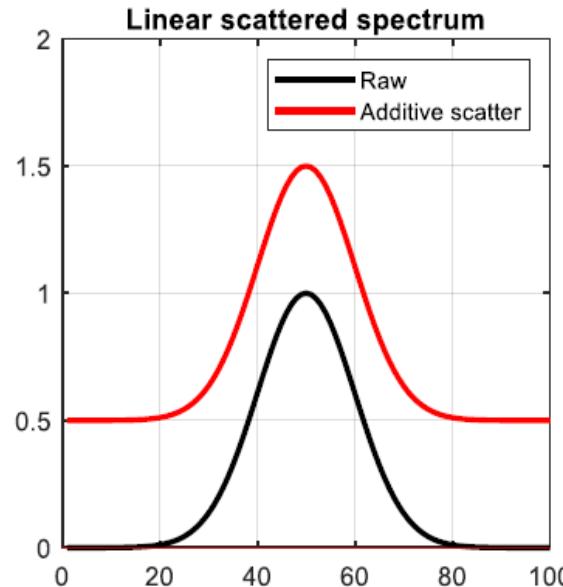
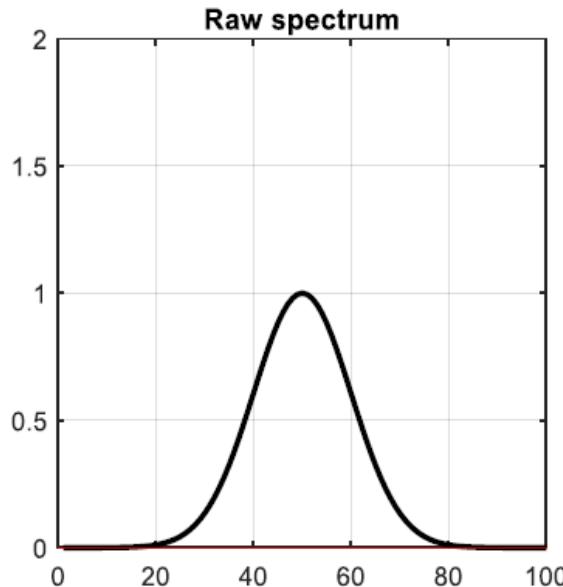
- Changing laser power;
- Differences in focusing depth;
- Sample volume differences.



Normalization

It is fundamental when:

- measuring concentrations with Raman based on calibration curves;
- building chemometric models.

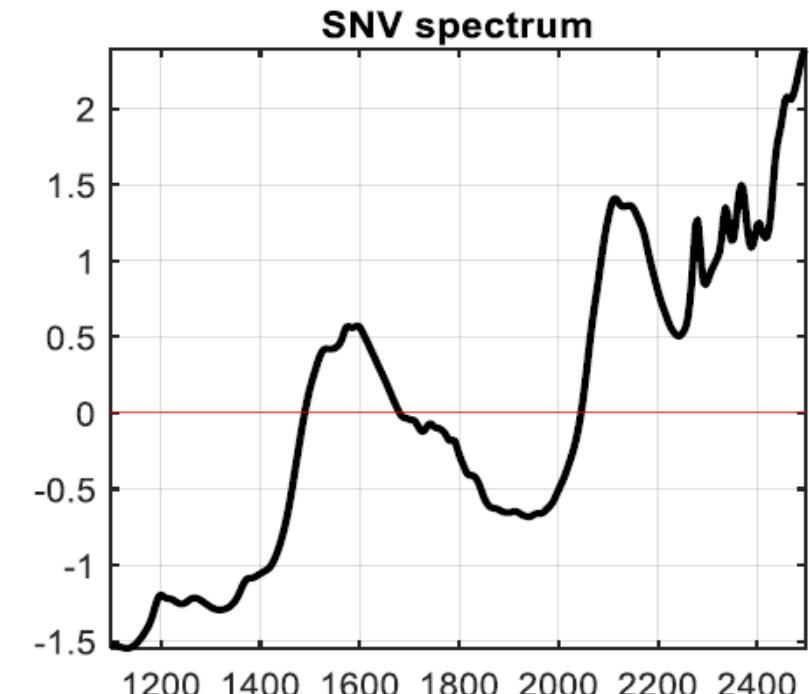
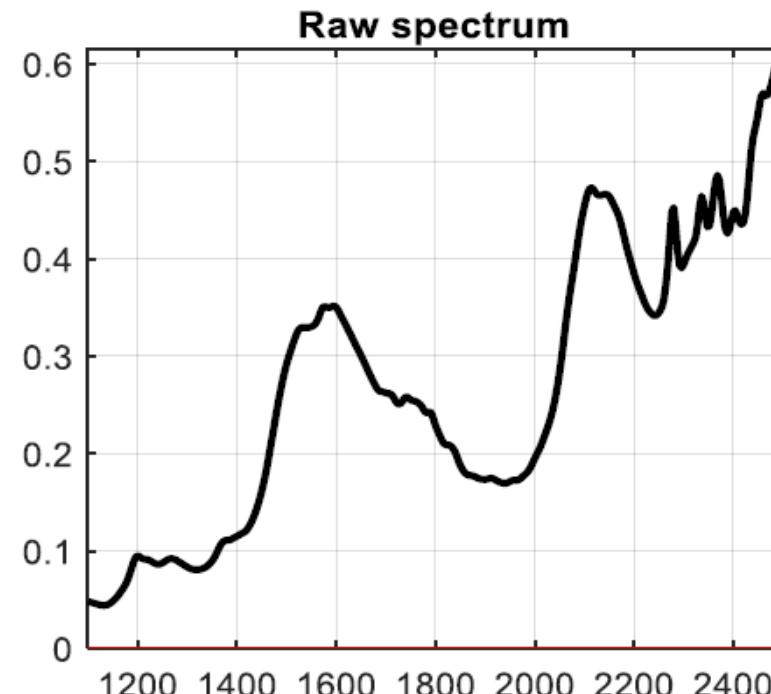


Standard Normal Variate (SNV)

It does **NOT** change the shape of the spectra

It is applied to every single row

$$\hat{x}_m = \frac{(x_m - \text{mean}(x_m))}{\text{std}(x_m)}$$

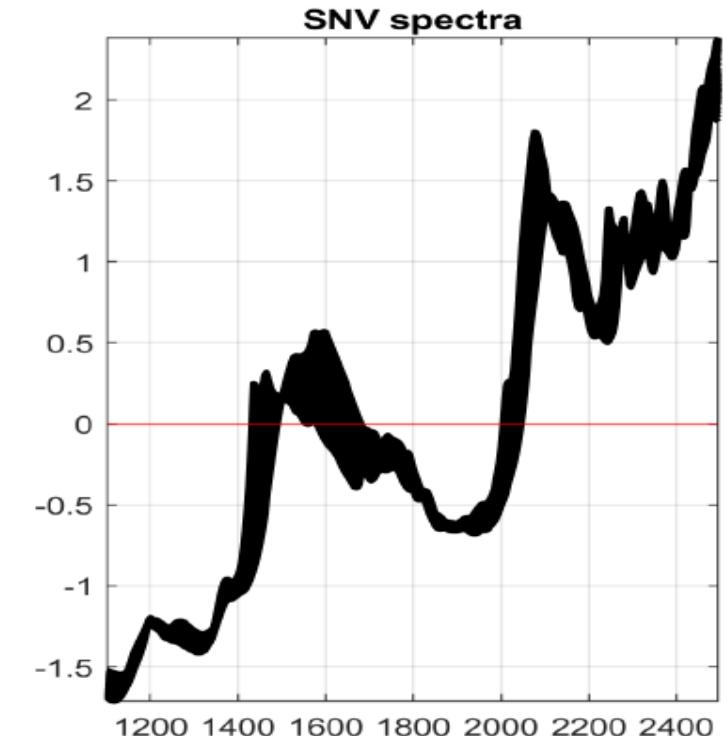
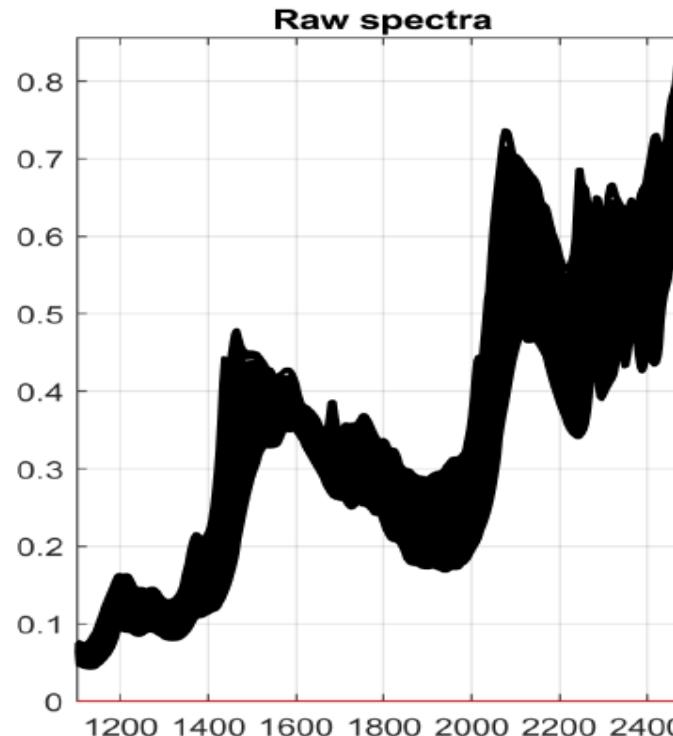


Standard Normal Variate (SNV)

It does **NOT** change the shape of the spectra

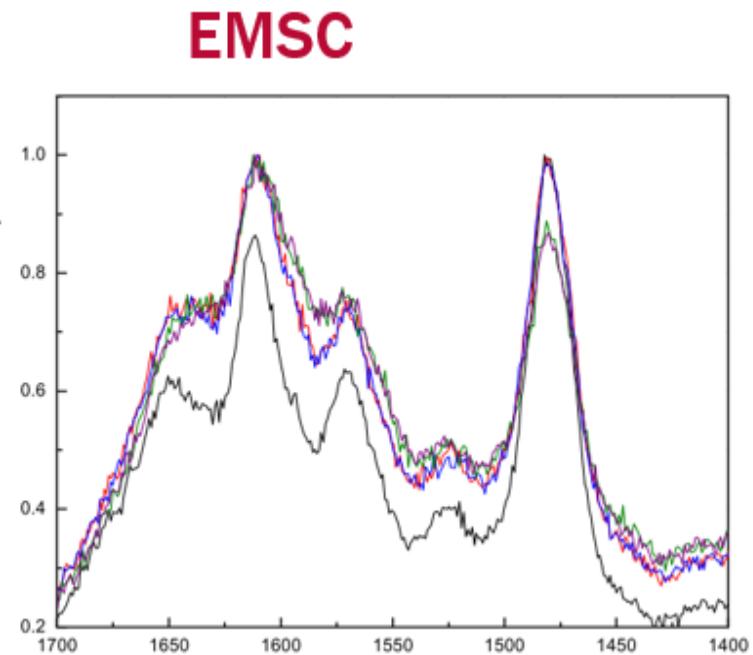
It is applied to every single row

$$\hat{x}_m = \frac{(x_m - \text{mean}(x_m))}{\text{std}(x_m)}$$



Multiplicative Scatter correction (MSC) and the extended version (EMSC)

- It removes constant effects in one spectra (scattering)
- It does MIGHT change the shape of the spectra
- A reference spectrum is needed: the mean, the median,
- or an external spectrum



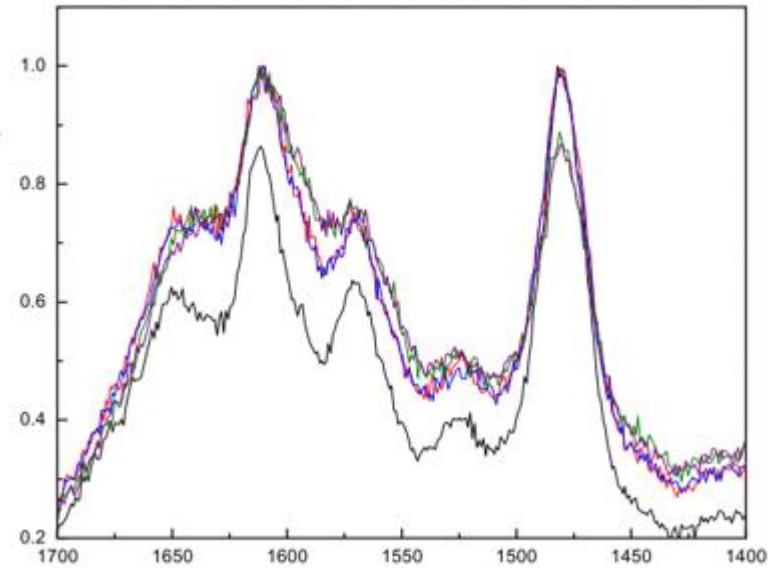
Multiplicative Scatter correction (MSC) and the extended version (EMSC)

Given a matrix $\mathbf{X}(M \times N)$ with M spectra measured at N wavelengths, for every spectrum:

- 1) Regress \mathbf{x}_m over the reference \mathbf{x}_{ref} . Calculate the slope and offset
- 2) Correct \mathbf{x}_m as follow:

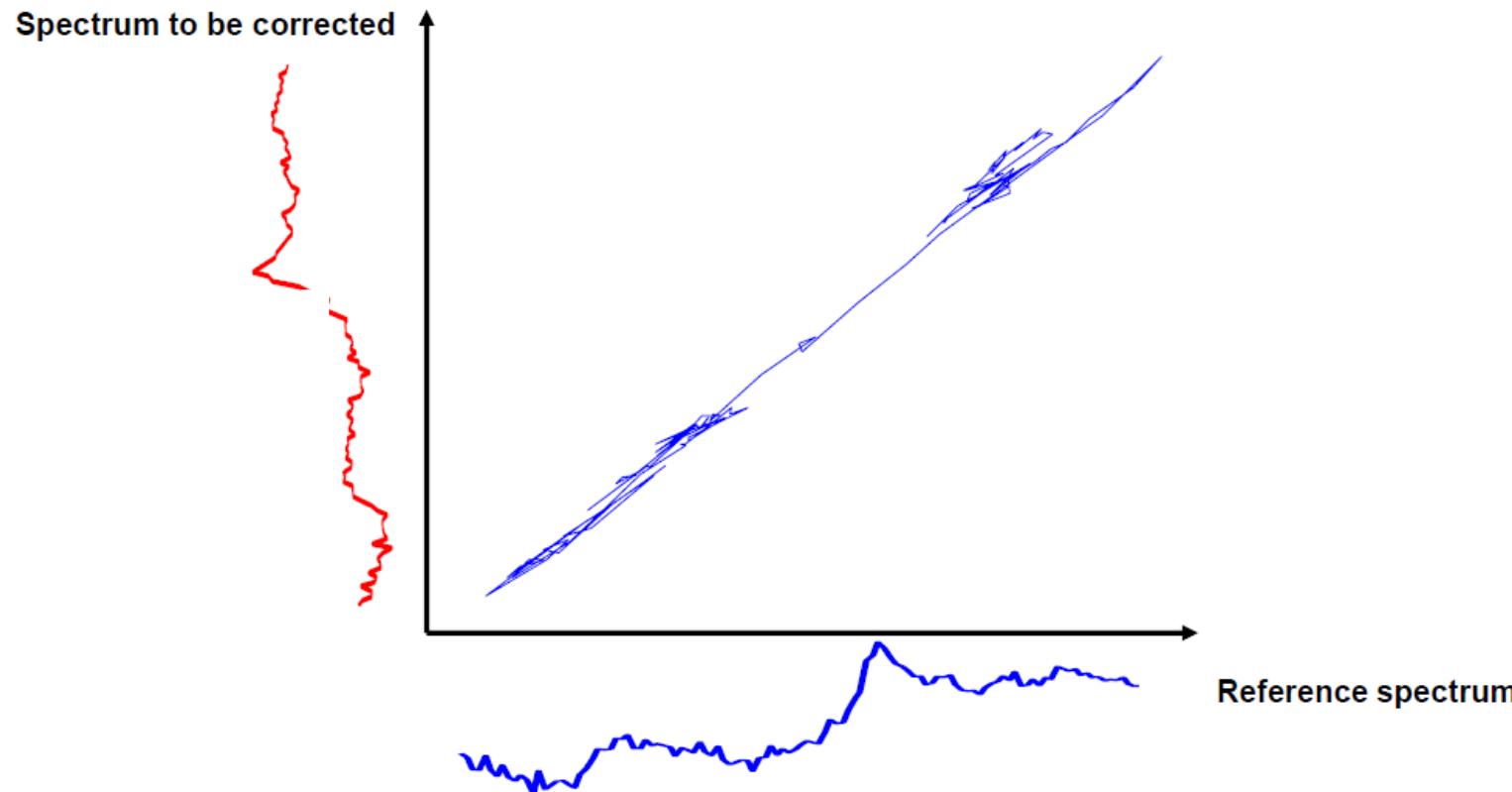
$$\hat{x}_n = \frac{(x_m - \text{offset})}{\text{slope}}$$

EMSC

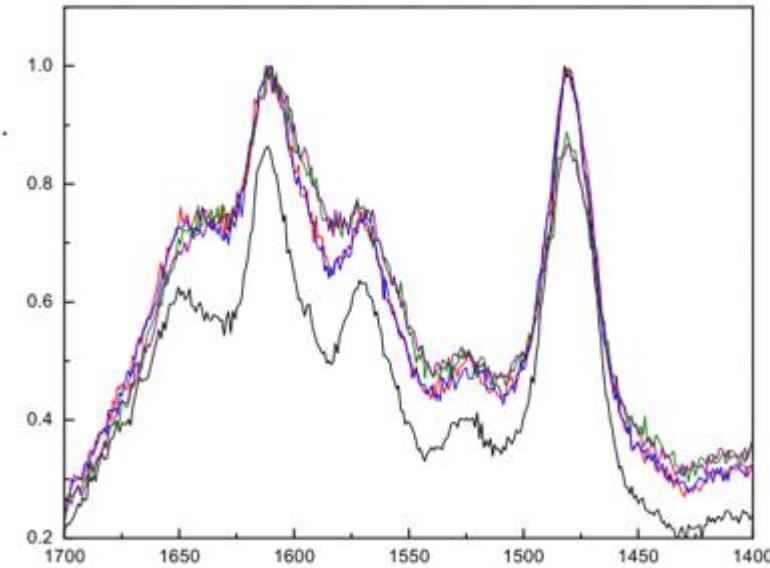


Multiplicative Scatter correction (MSC) and the extended version (EMSC)

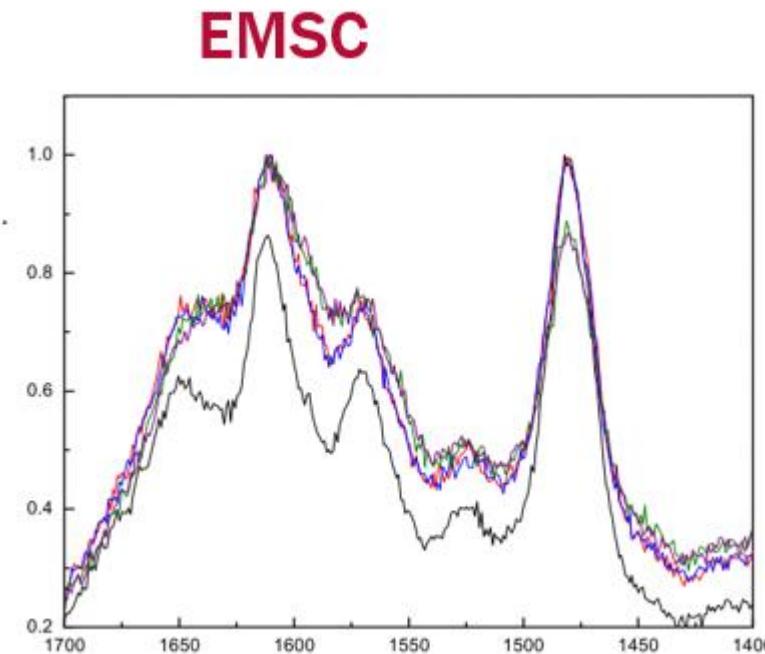
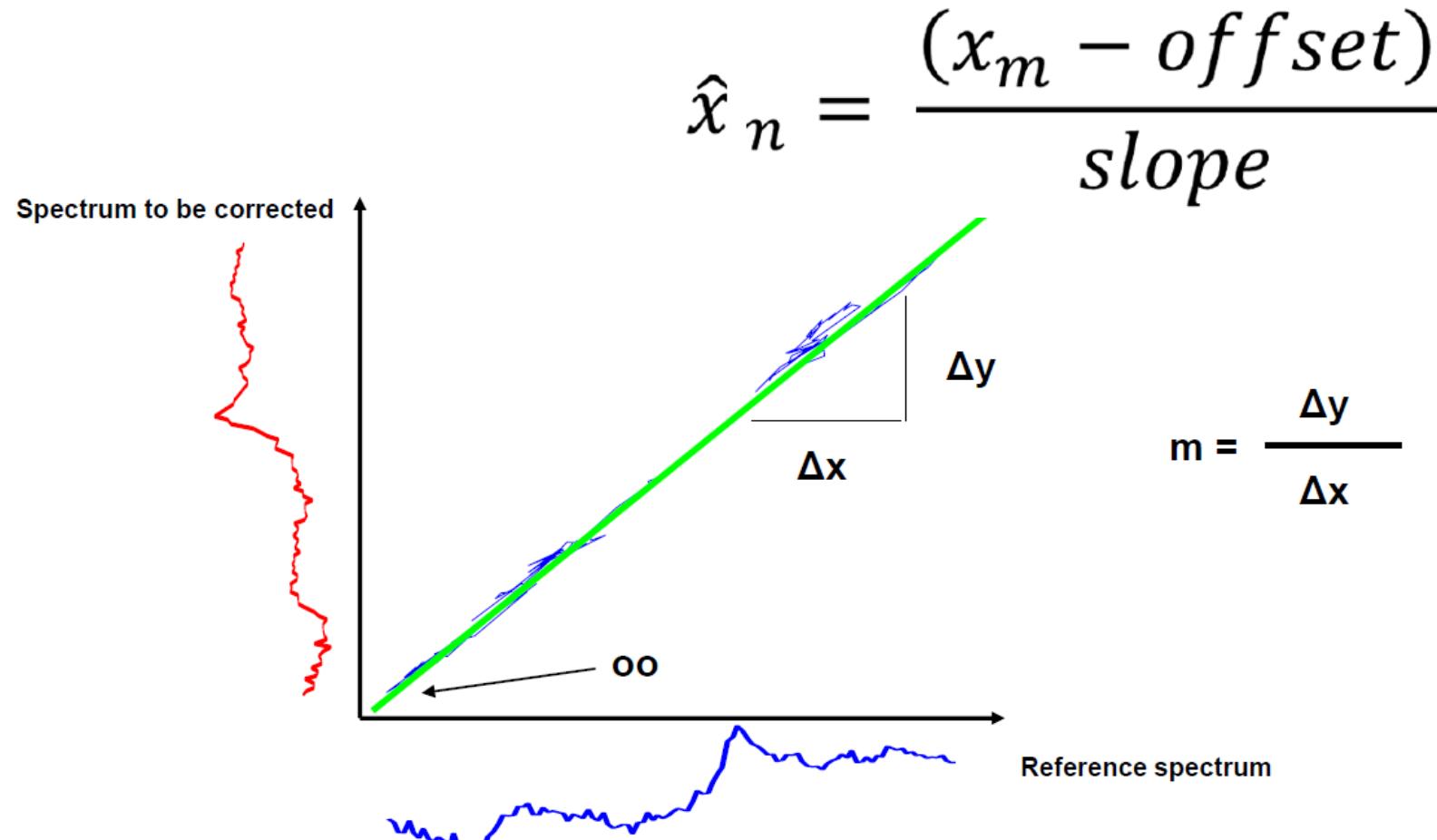
$$\hat{x}_n = \frac{(x_m - \text{offset})}{\text{slope}}$$



EMSC

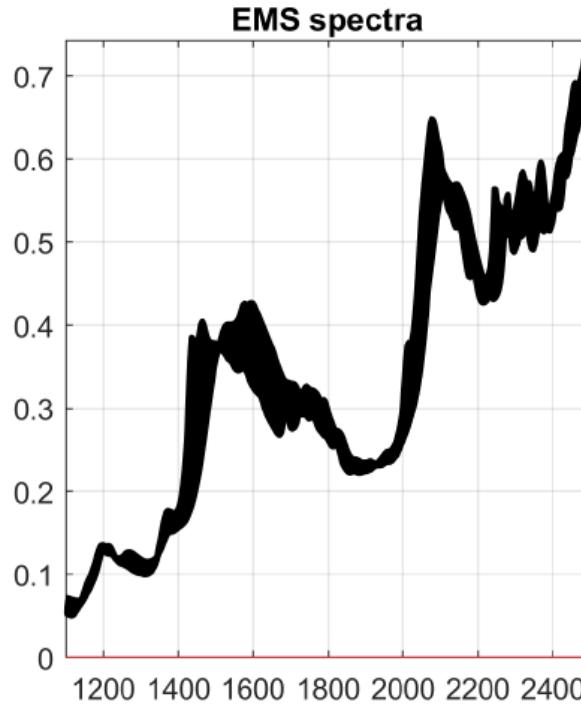
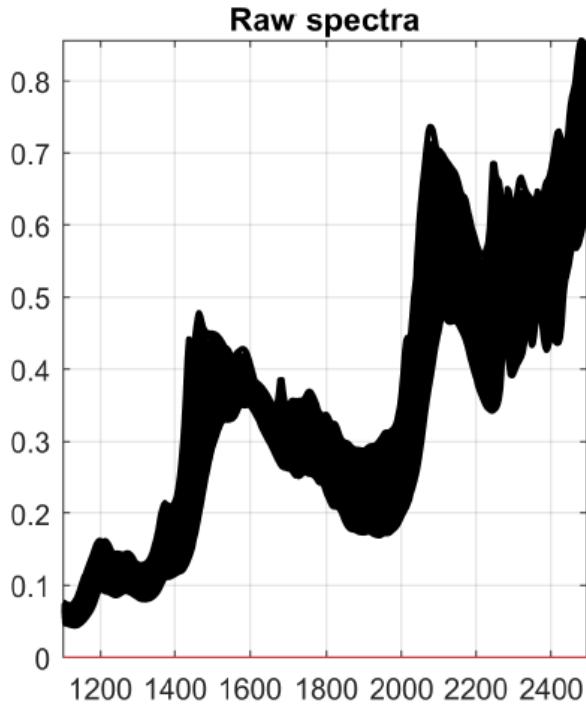


Multiplicative Scatter correction (MSC) and the extended version (EMSC)

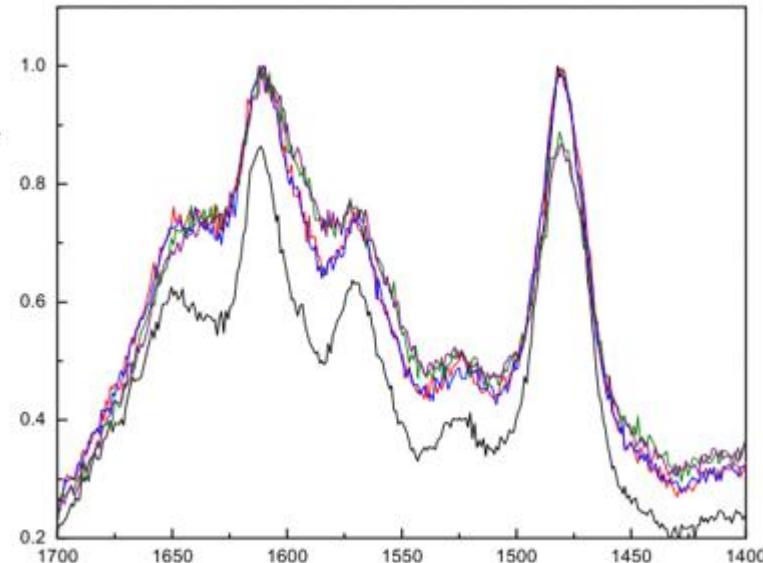


Multiplicative Scatter correction (MSC) and the extended version (EMSC)

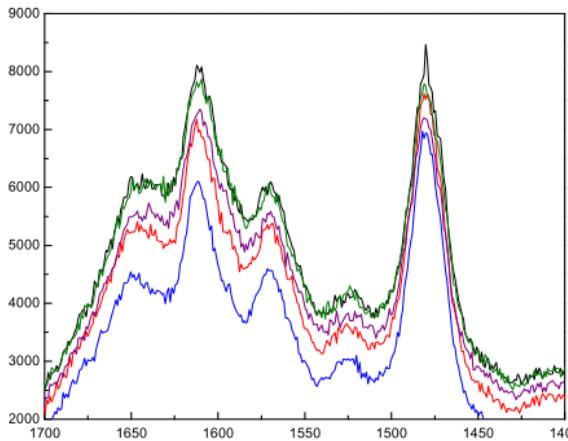
$$\hat{x}_n = \frac{(x_m - \text{offset})}{\text{slope}}$$



EMSC

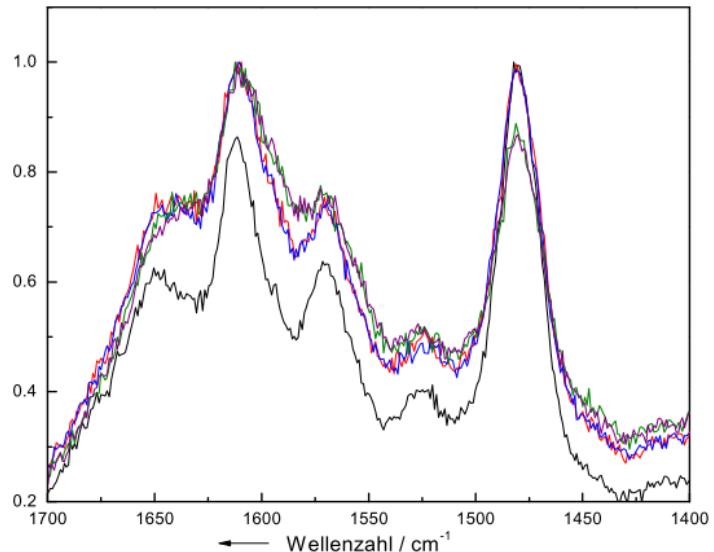


Multiplicative Scatter correction (MSC) and the extended version (EMSC)

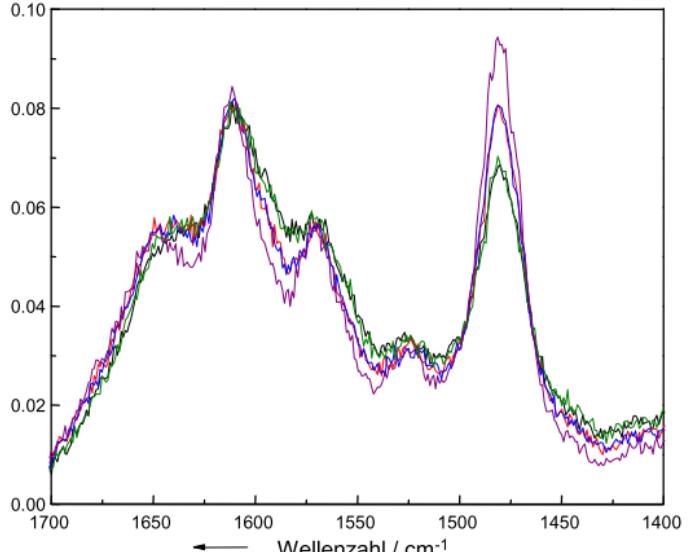


Original data

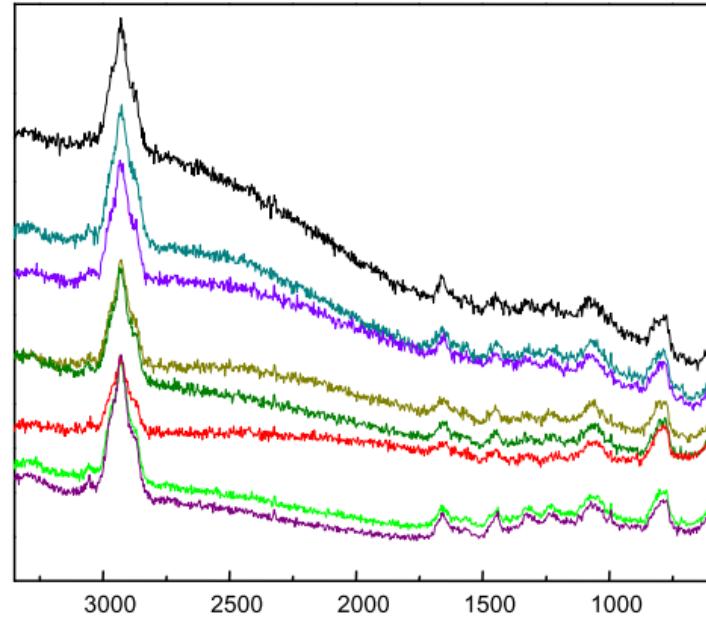
EMSC



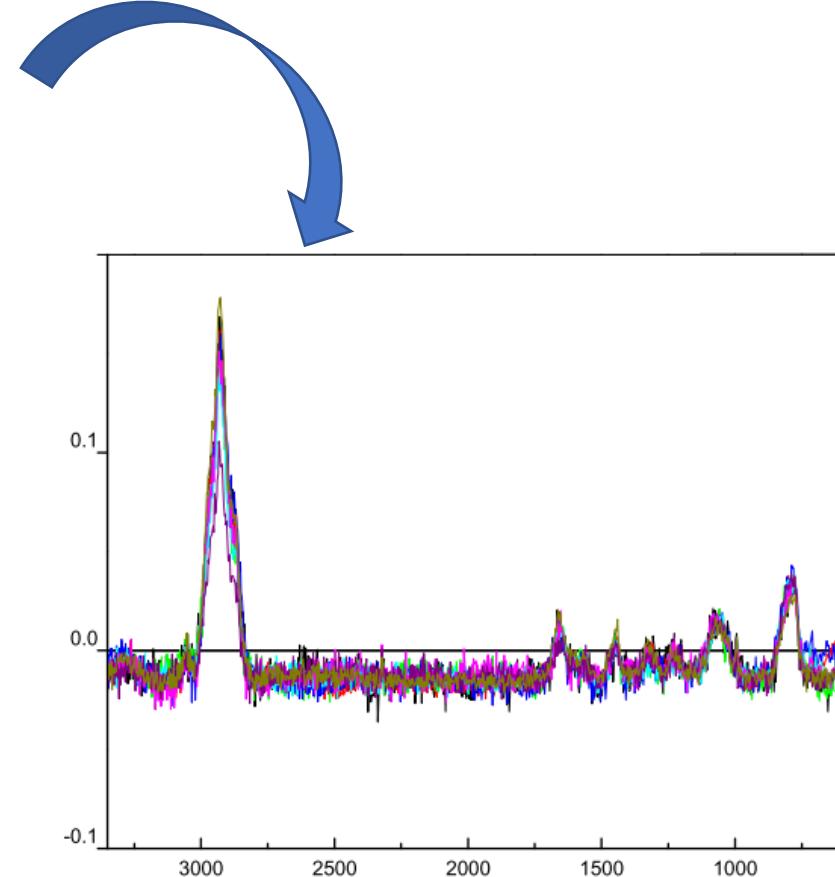
SNV



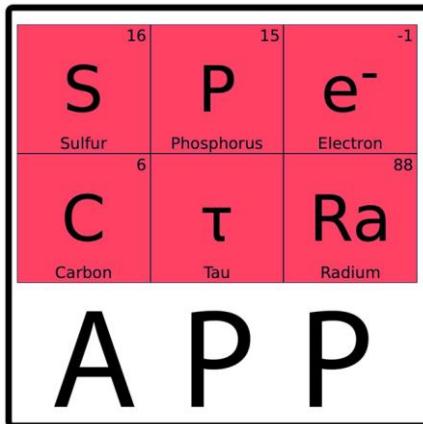
Pre-processing combinations



SNV + baseline

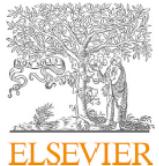


Pre-processing combinations



Pre-processing combinations

Chemometrics and Intelligent Laboratory Systems 202 (2020) 104029



Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components



Agnieszka Martyna ^{a,*}, Alicja Menzyk ^a, Alessandro Damin ^b, Aleksandra Michalska ^c,
 Gianmario Martra ^b, Eugenio Alladio ^{b,d,e}, Grzegorz Zadora ^{a,c}

^a University of Silesia in Katowice, Faculty of Science and Technology, Institute of Chemistry, Forensic Chemistry Unit, 9 Szkołna, Katowice, 40-006, Poland

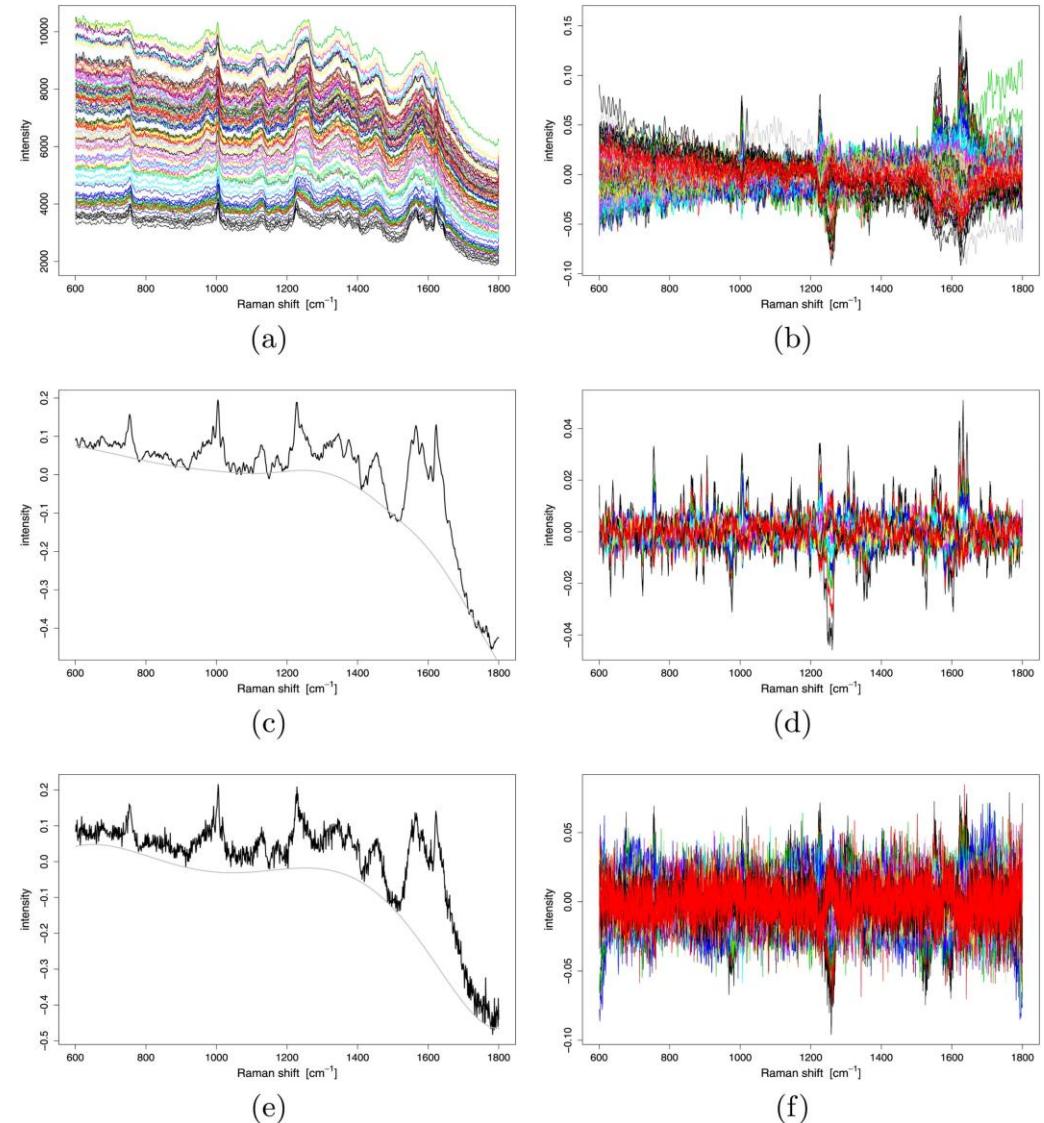
^b Dipartimento di Chimica, Università degli Studi di Torino, 7 Via Pietro Giuria, Torino, 10125, Italy

^c Institute of Forensic Research in Krakow, 9 Westerplatte, Krakow, 31-033, Poland

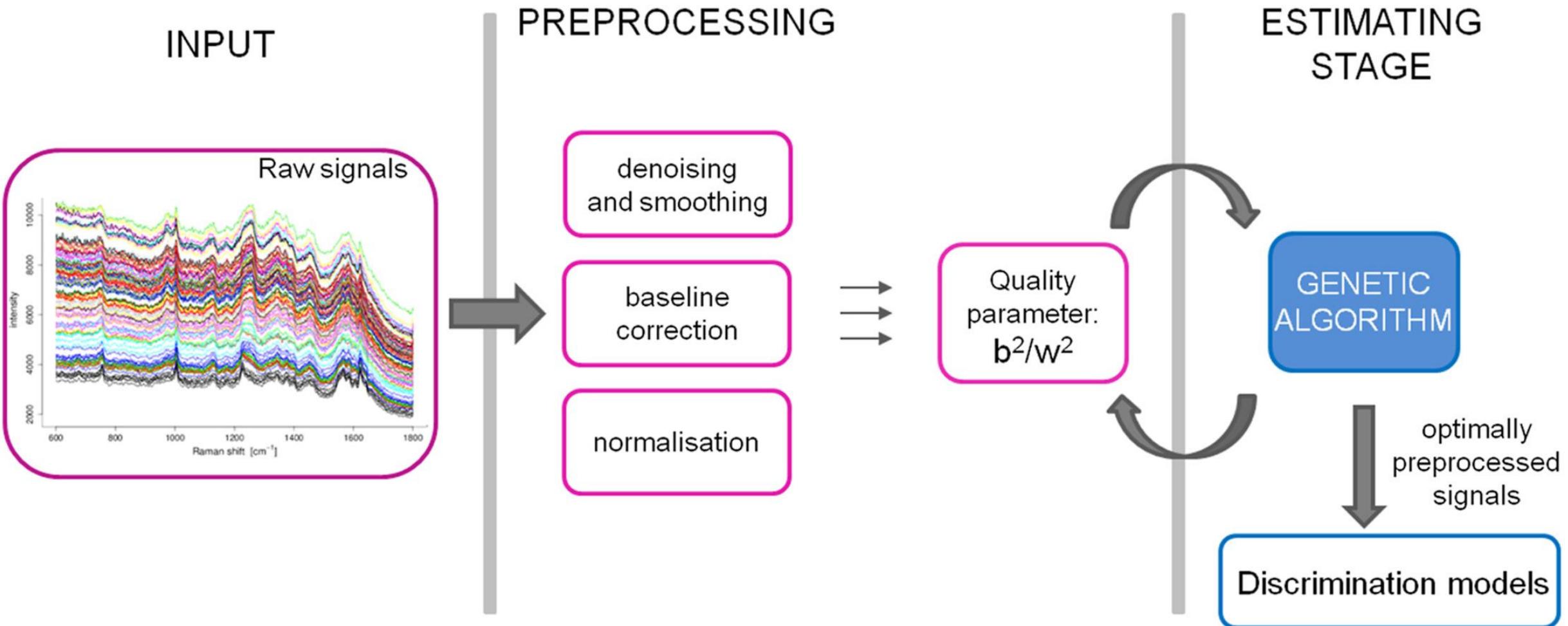
^d Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", 10/1 Regione Gonzole, Orbassano, 10043, Torino, Italy

^e Reparto CC Investigazioni Scientifiche di Roma, Sezione di Biologia, 119 Viale Tor di Quinto, Rome, 00191, Italy

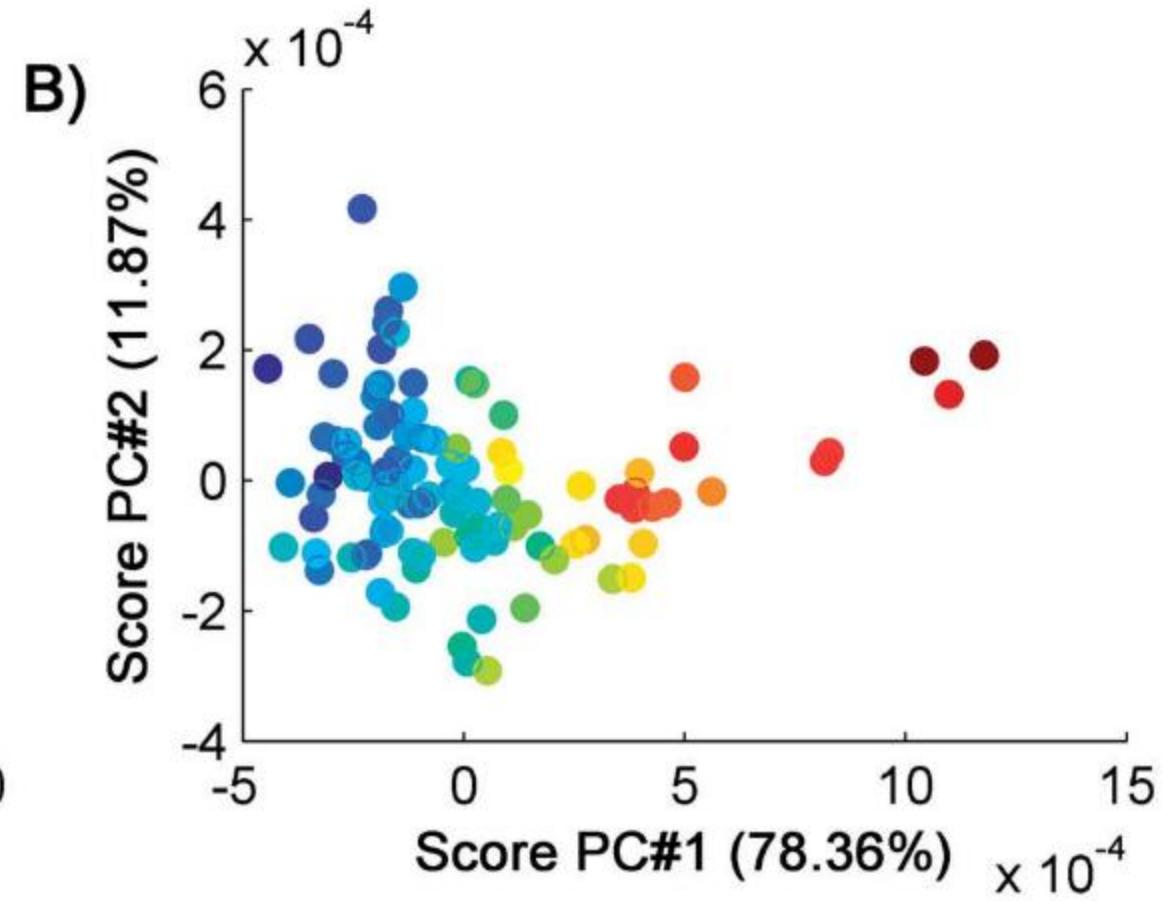
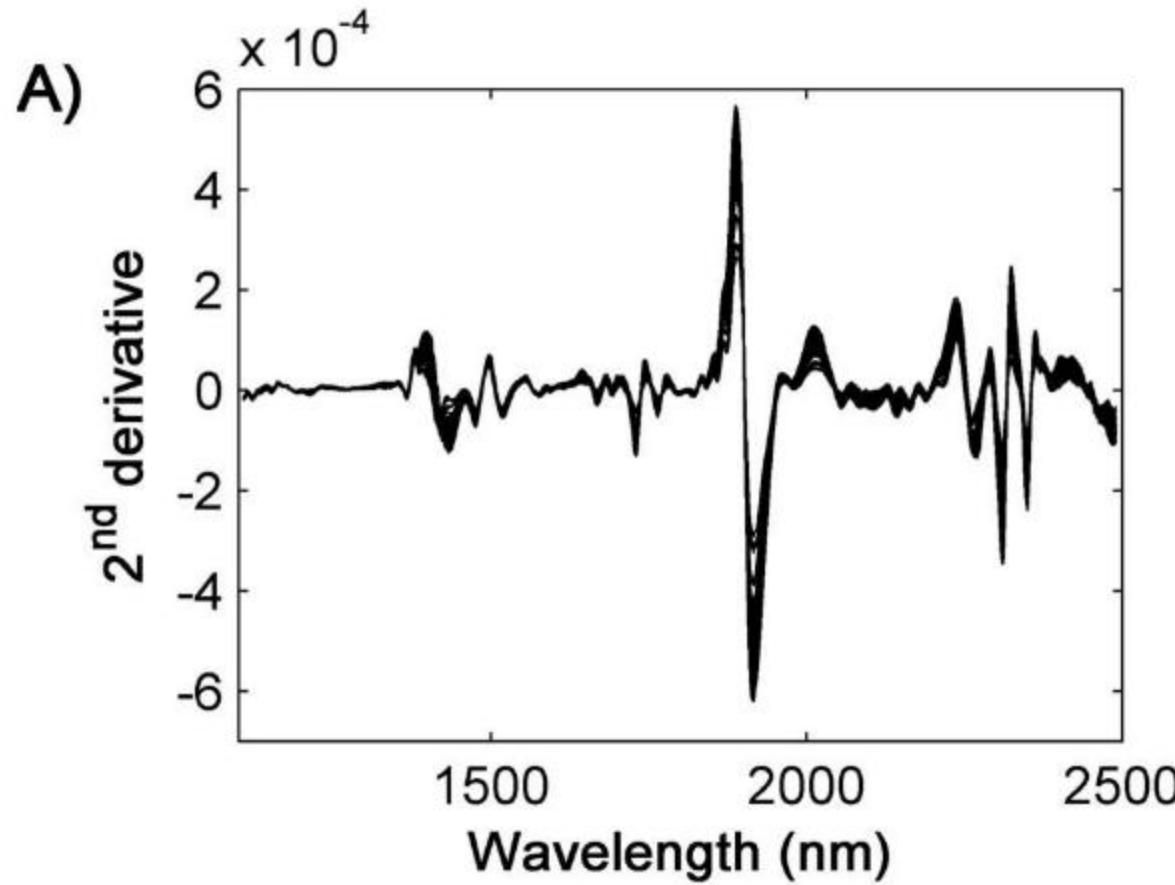
<https://doi.org/10.1016/j.chemolab.2020.104029>



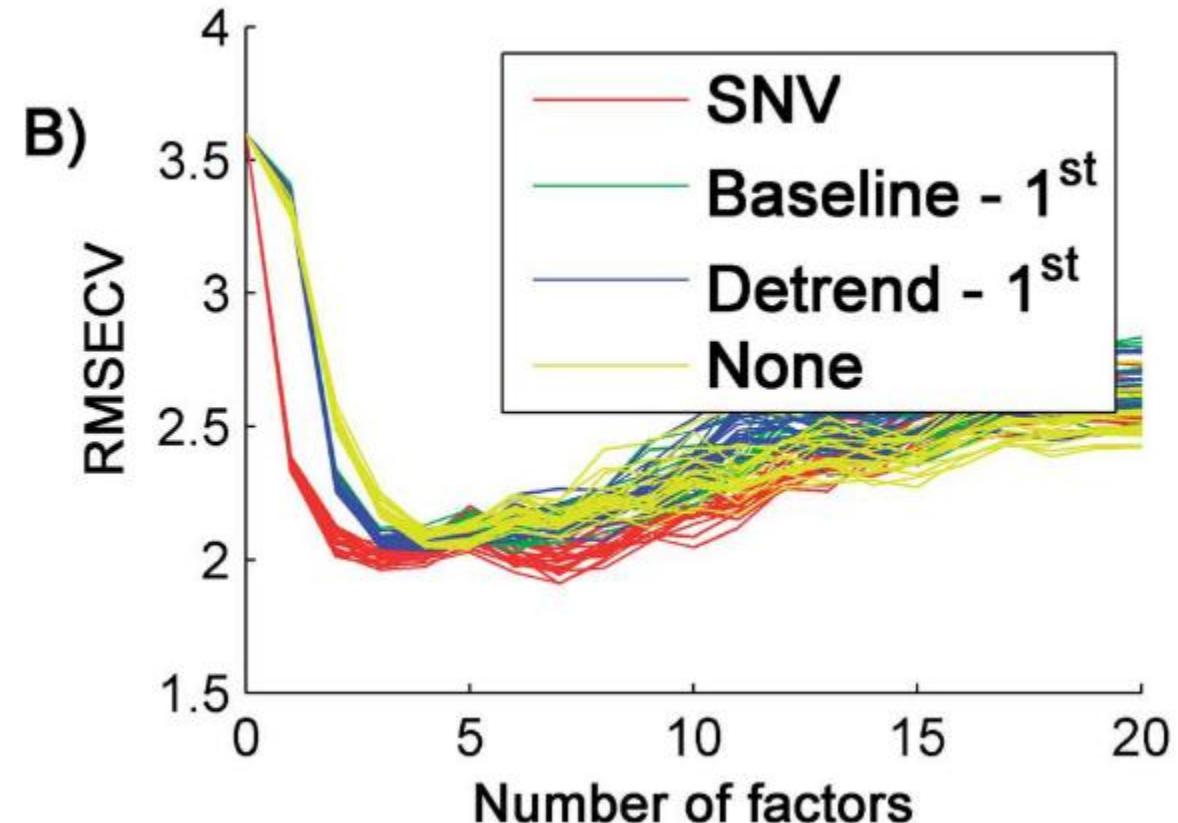
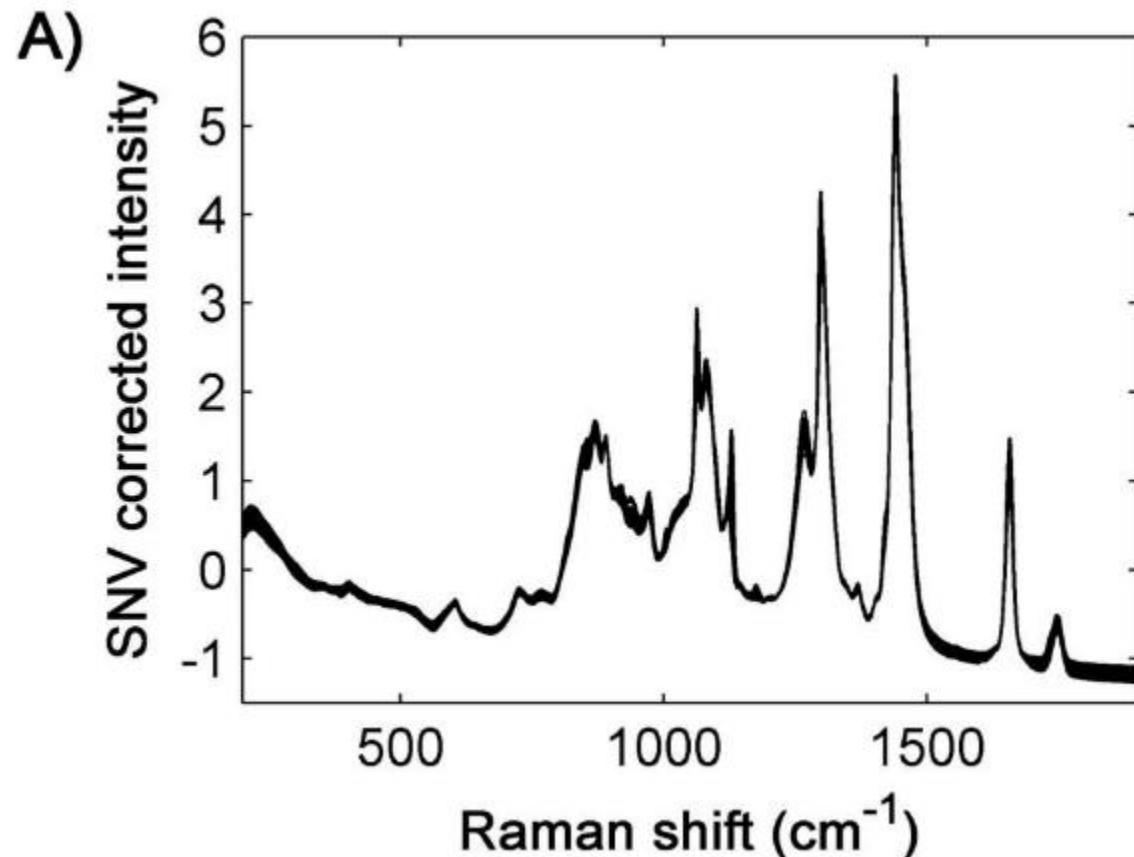
Pre-processing combinations



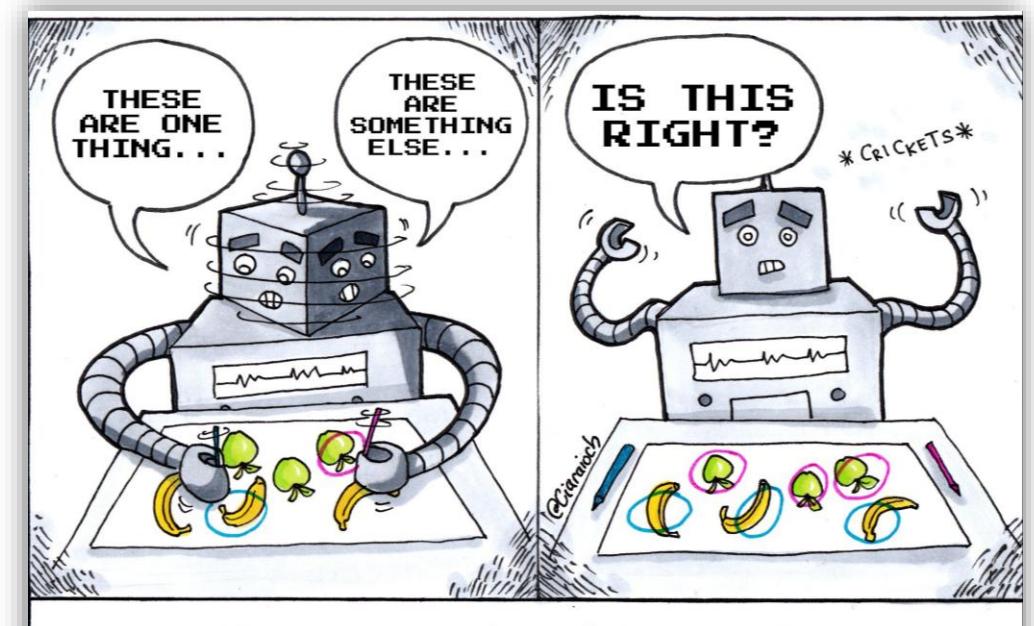
Some examples



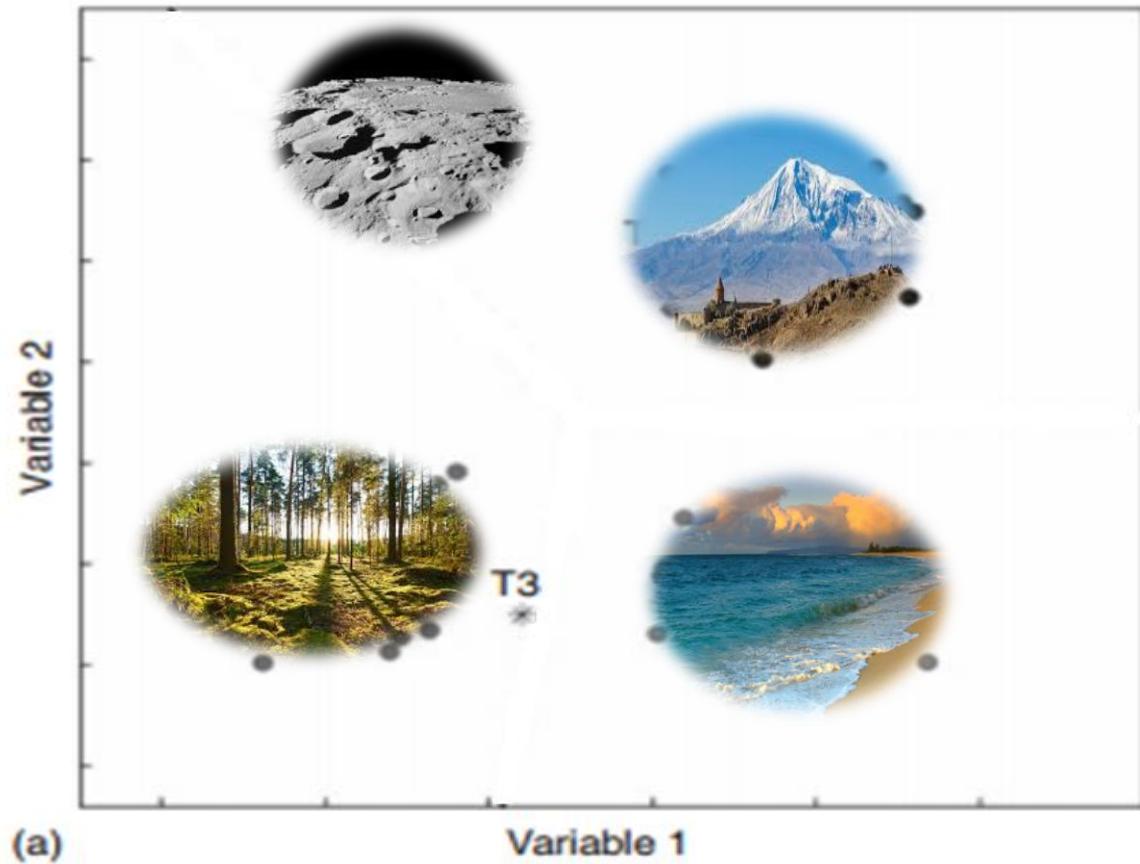
Some examples



Unsupervised learning

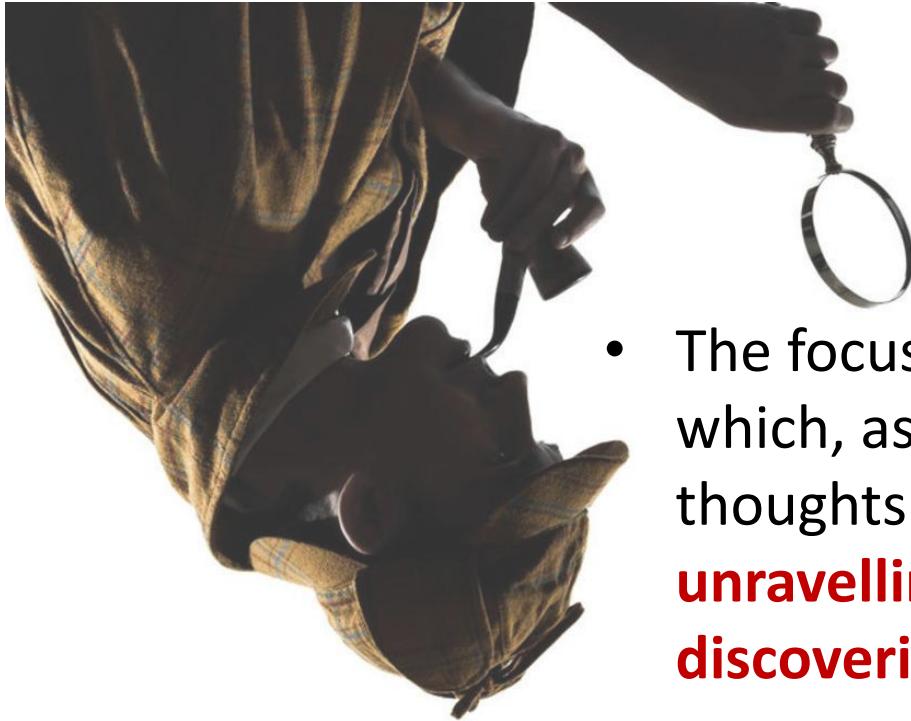


Unsupervised Learning



- **PCA (Principal Component Analysis)**
 - **CA (Cluster Analysis)**
- **t-SNE (t-distributed stochastic neighbor embedding)**

PCA - exploration

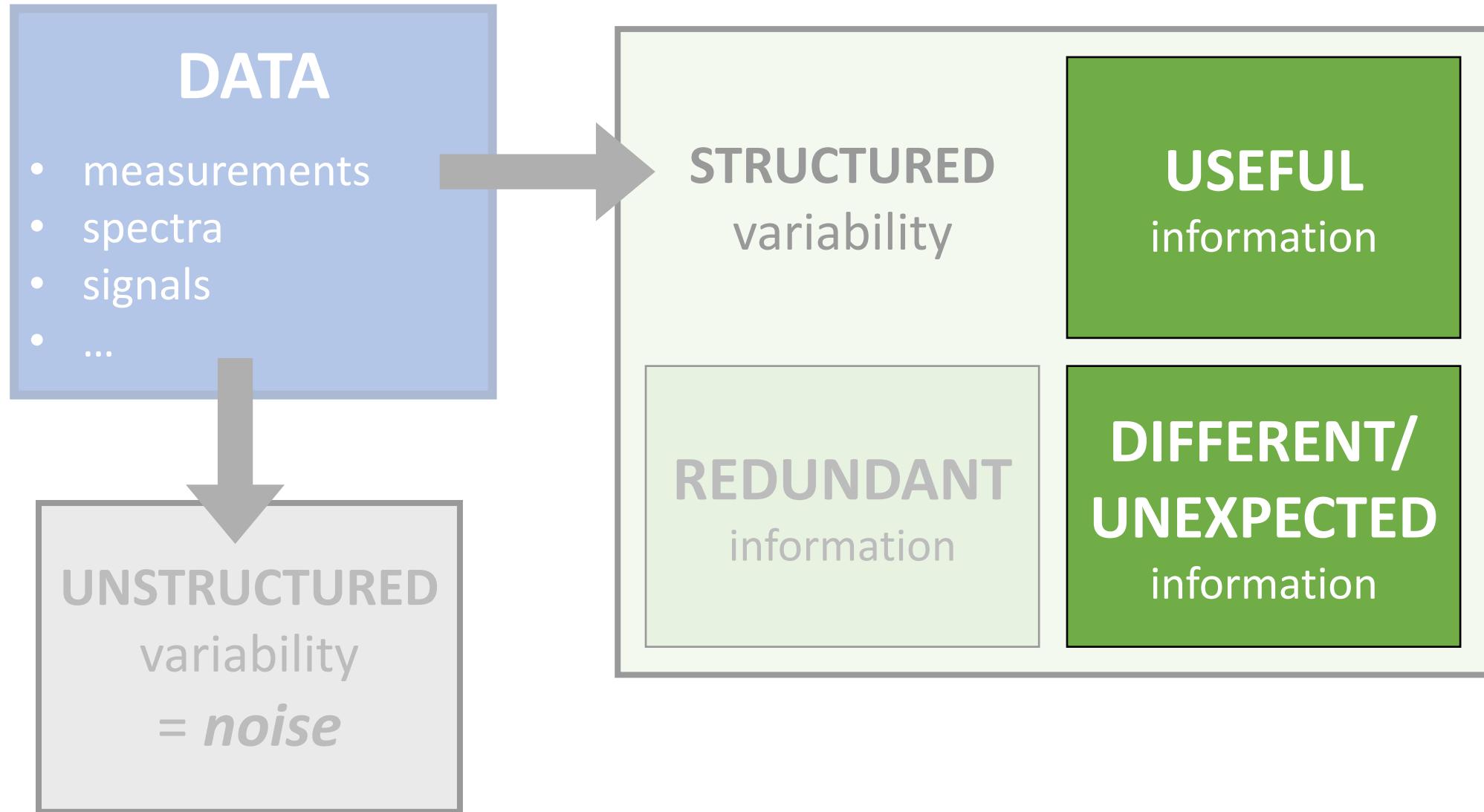


- Inductive **data-driven** attitude with a minimum of *a priori* hypothesis.
- The focus is on “**exploration**”, which, as a word, recalls exotic thoughts and feelings, such as **unravelling** mysterious threads or **discovering** unknown worlds.



- An **immediate, direct** and **easy-to-understand** visual representation of the information contained in the data is sought.
- The **focus on the data**, rather than the hypothesis → **let your data talk!**
- **VERY IMPORTANT!** → exploratory data analysis can never be the whole story, it only is the first step, the foundation stone of the whole data analysis.

PCA - exploration



Why using PCA?

- We study **phenomena that can not be directly observed**
 - Underlying factors that govern the observed data (from a physico-chemical point-of-view)
- We want to **identify and operate with underlying latent factors rather than the observed data**
 - E.g. topics in news articles
 - Transcription factors in genomics
- We want to **discover and exploit hidden relationships**
 - “beautiful car” and “gorgeous automobile” are closely related
 - So are “driver” and “automobile”
 - But does your search engine know this?
 - Reduces noise and error in results



Why using PCA?

- **We have too many observations and dimensions (Big Data)**
 - To **reason** about or obtain insights from
 - To **visualize**
 - Too much **noise** in the data
 - Need to “reduce” them to a **smaller set of factors**
 - **Better representation** of data without losing much information
 - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition.
- **Combinations of observed variables** may be more effective bases for insights, even if physico-chemical meaning is obscure

Why using PCA?

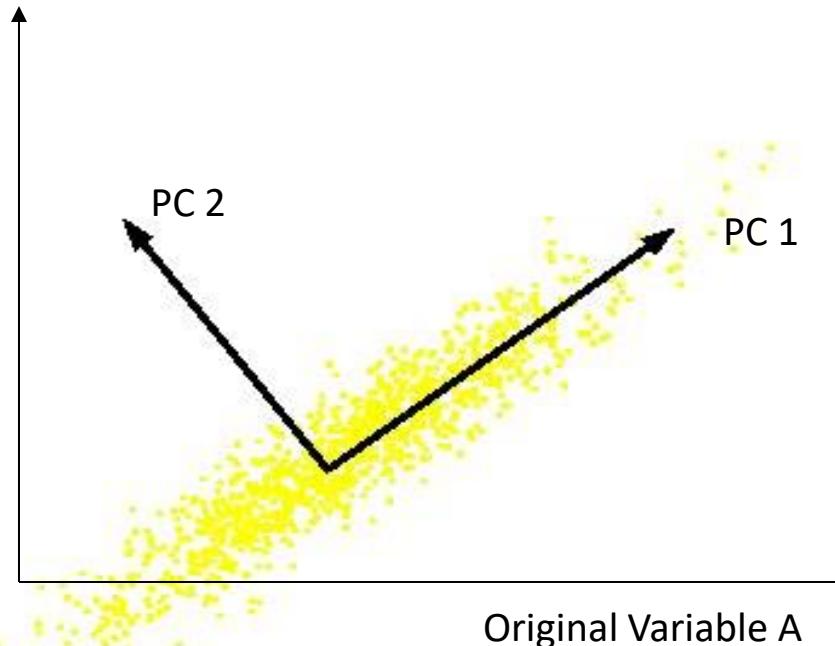
- Discover a **new set of components/dimensions/axes against which to represent, describe or evaluate the data**
 - For more effective reasoning, insights, or better visualization
 - Reduce noise in the data
 - Typically, a **smaller set of factors: dimension reduction**
 - Better representation of data without losing much information
 - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition
- **Components are combinations of observed variables**
 - May be more effective bases for insights, even if physico-chemical meaning is obscure
 - **Observed data are described in terms of these components** rather than in terms of original variables/dimensions

How to interpret PCA

- **Areas of variance** in data are where items can be best discriminated and key underlying phenomena observed
 - Areas of highest “signal” in the data
- **If two items or dimensions are highly correlated or dependent**
 - They are likely to represent highly related phenomena
 - If they tell us about the same underlying variance in the data, combining them to form a single measure is reasonable
 - Parsimony
 - Reduction in Error
- So we want to **combine related variables**, and **focus on uncorrelated or independent ones**, especially those along which the observations have high variance
- We want a smaller set of variables that explain most of the variance in the original data, in more compact and insightful form

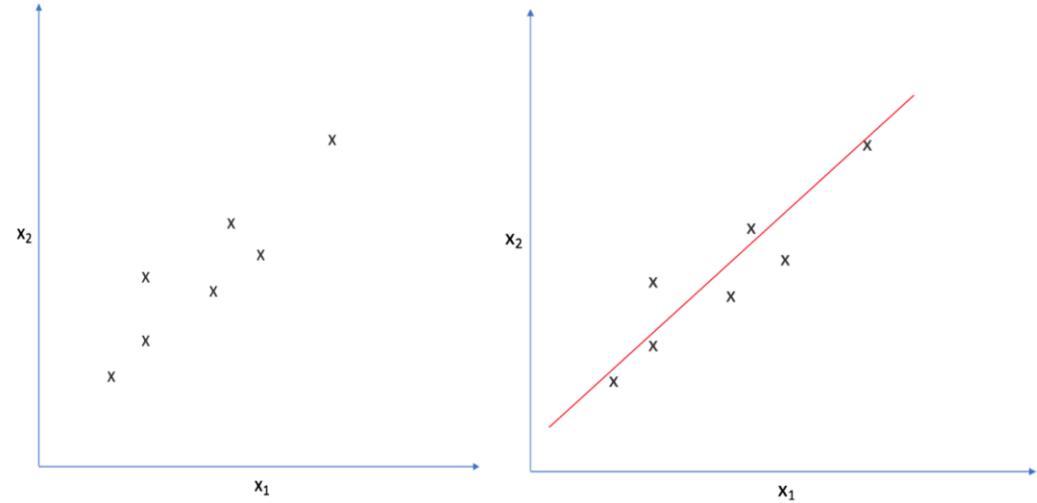
Computing the PCA model

What are the new axes?



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

Computing the PCA model



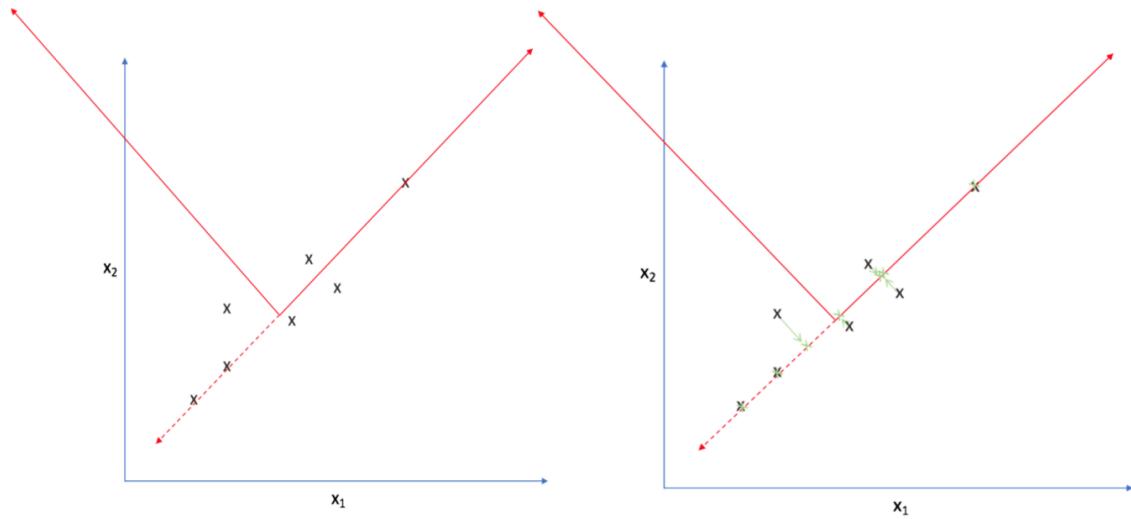
First principal component is the direction of **greatest variability (variance)** in the data

Second is the next orthogonal (uncorrelated) direction of greatest variability

So first remove all the variability along the first component, and then find the next direction of greatest variability

And so on ...

Computing the PCA model



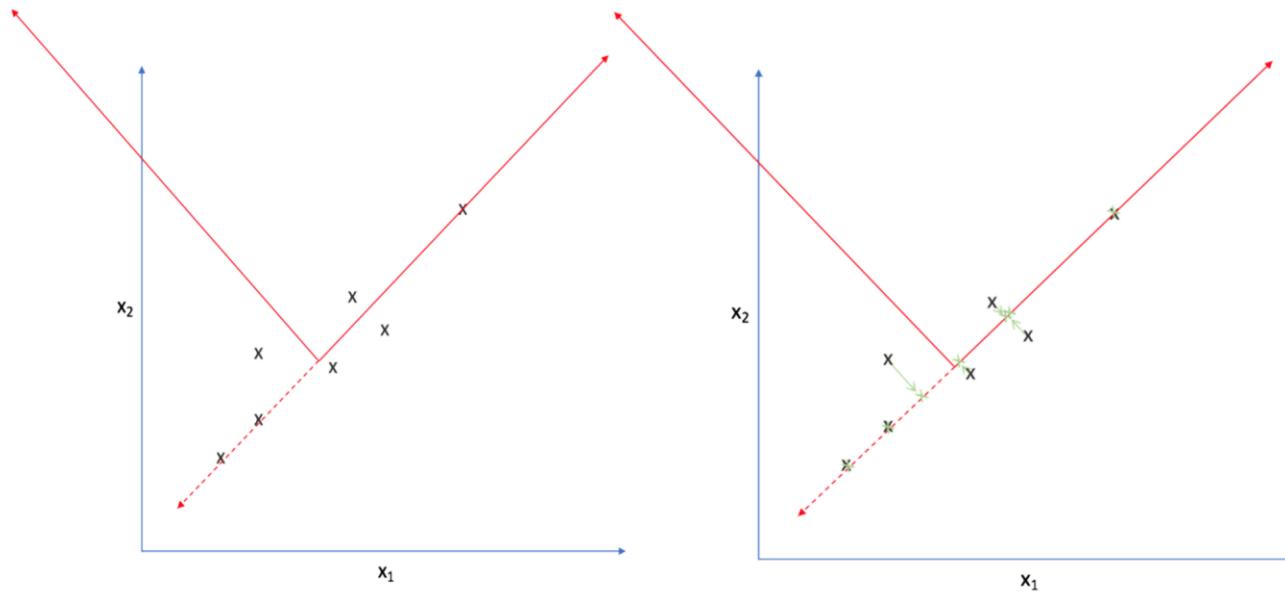
First principal component is the direction of **greatest variability (variance)** in the data

Second is the next orthogonal (uncorrelated) direction of greatest variability

So first remove all the variability along the first component, and then find the next direction of greatest variability

And so on ...

Computing the PCA model



Principle

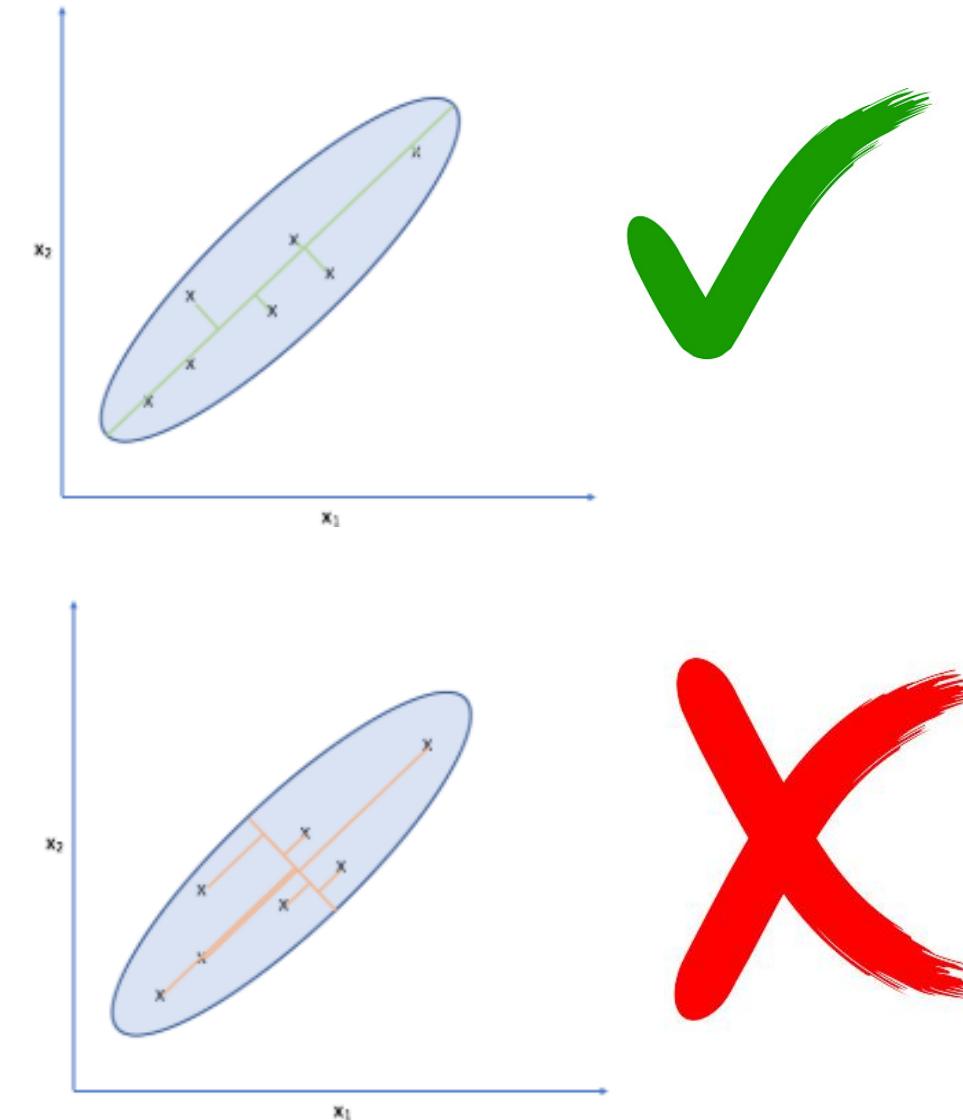
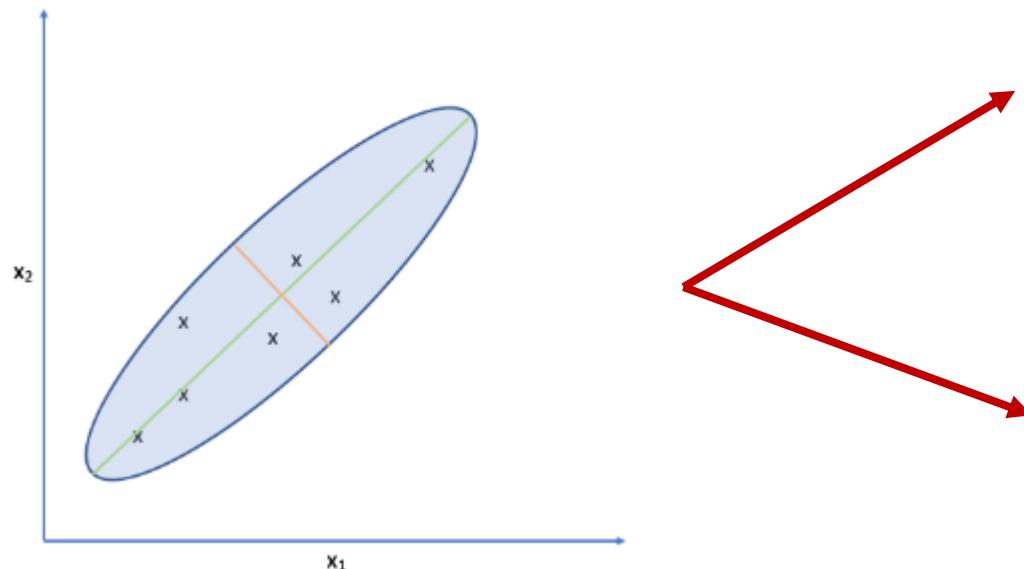
Linear combination and projection method to reduce the number of parameters;
Transfer a set of correlated variables into a new set of uncorrelated variables;
Map the data into a **space of lower dimensionality**.

Properties

It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables;
New axes are orthogonal and represent the directions with maximum variability.

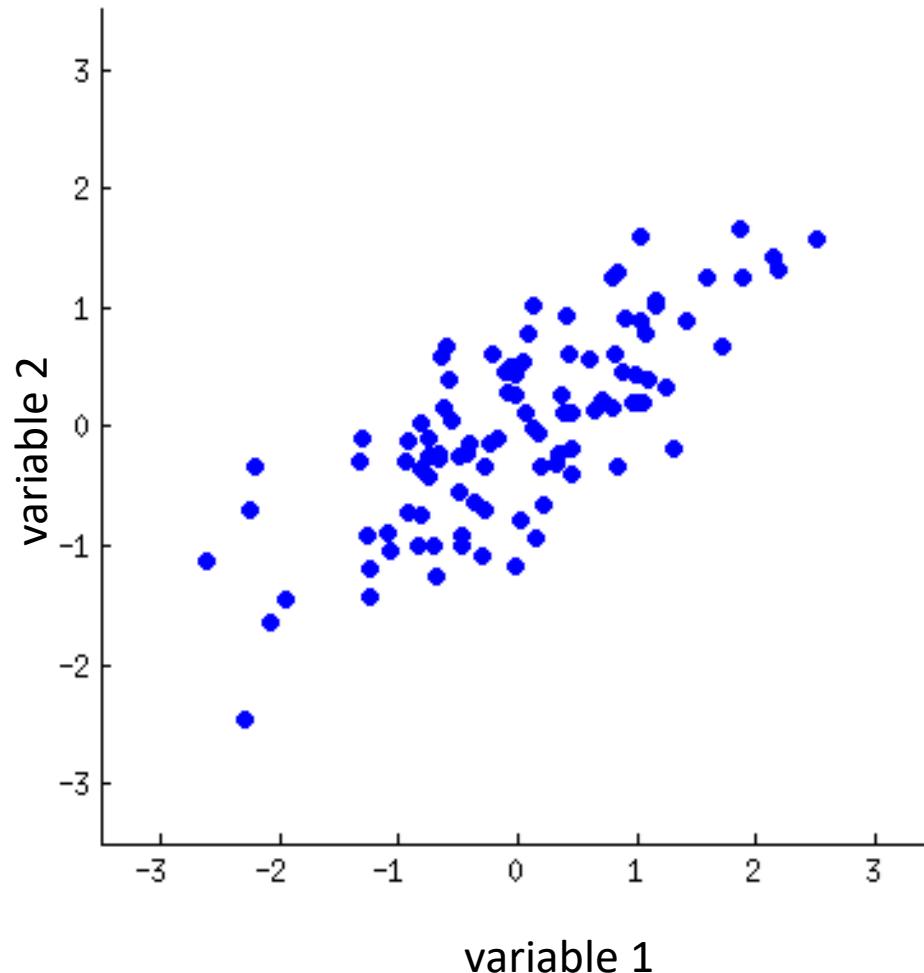
PCA – in theory

$$\text{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

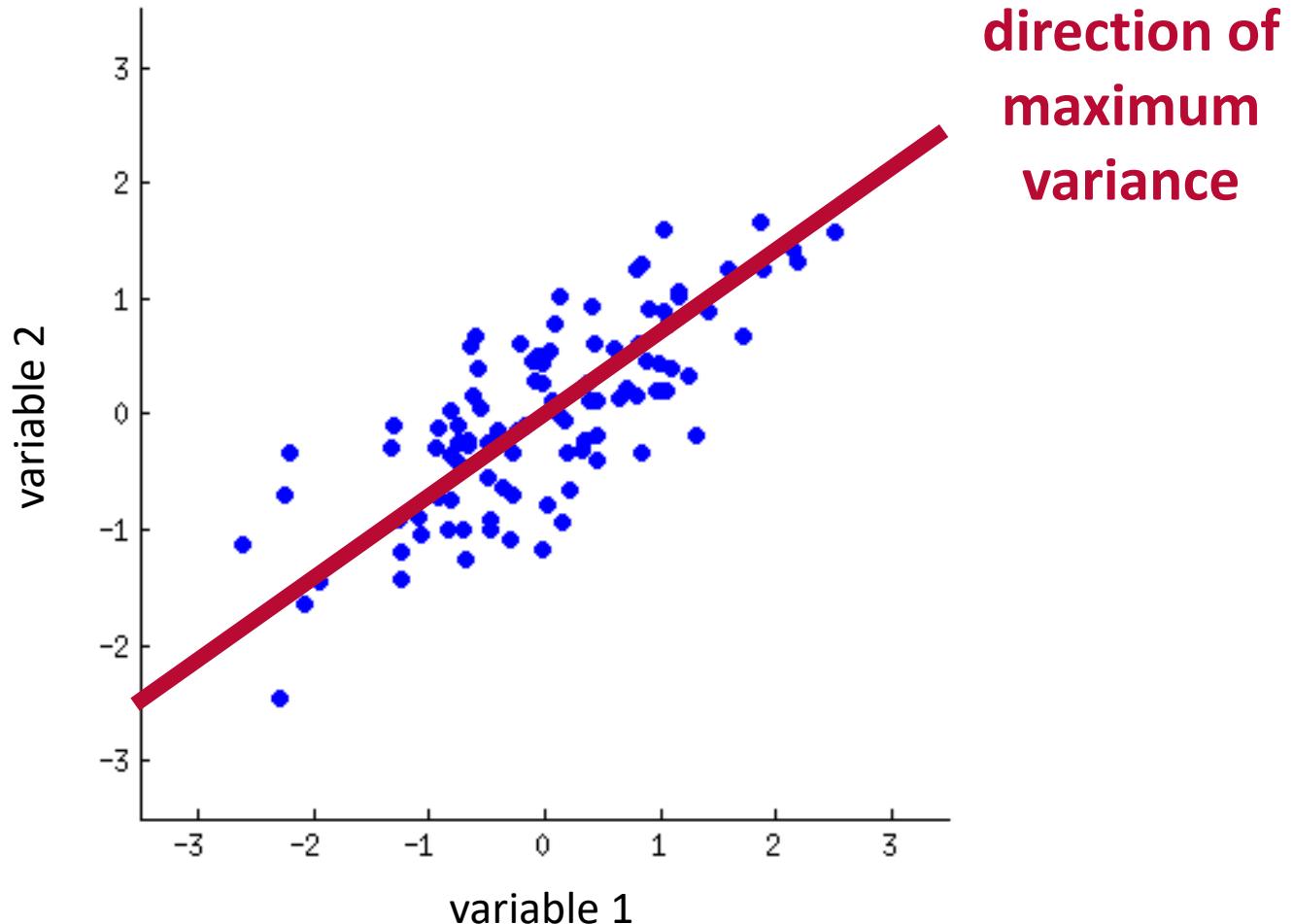


PCA – in theory

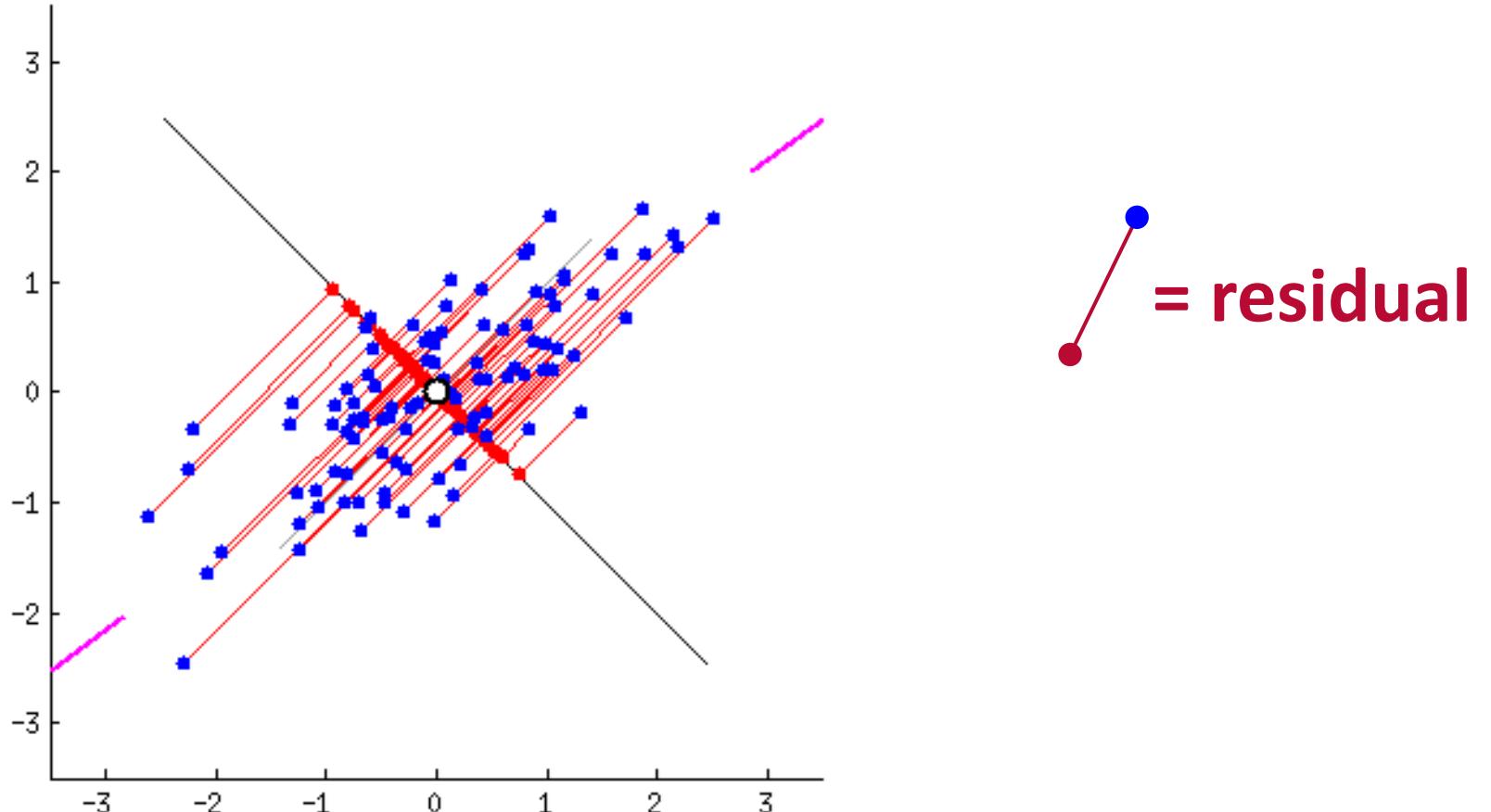
Most common algorithms: **NIPALS, SVD, ...** → they all provide the same results!



PCA – in theory

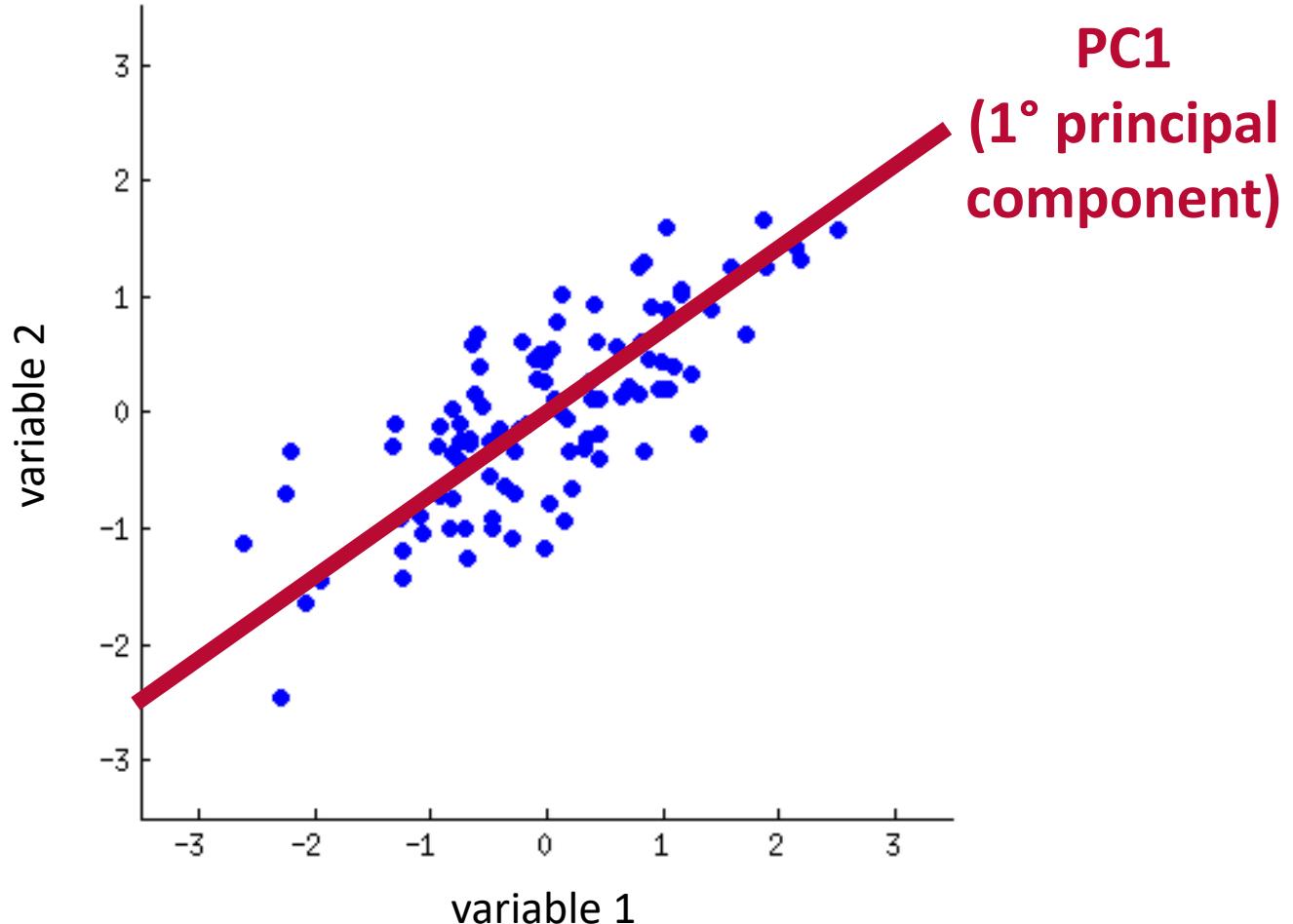


PCA – in theory



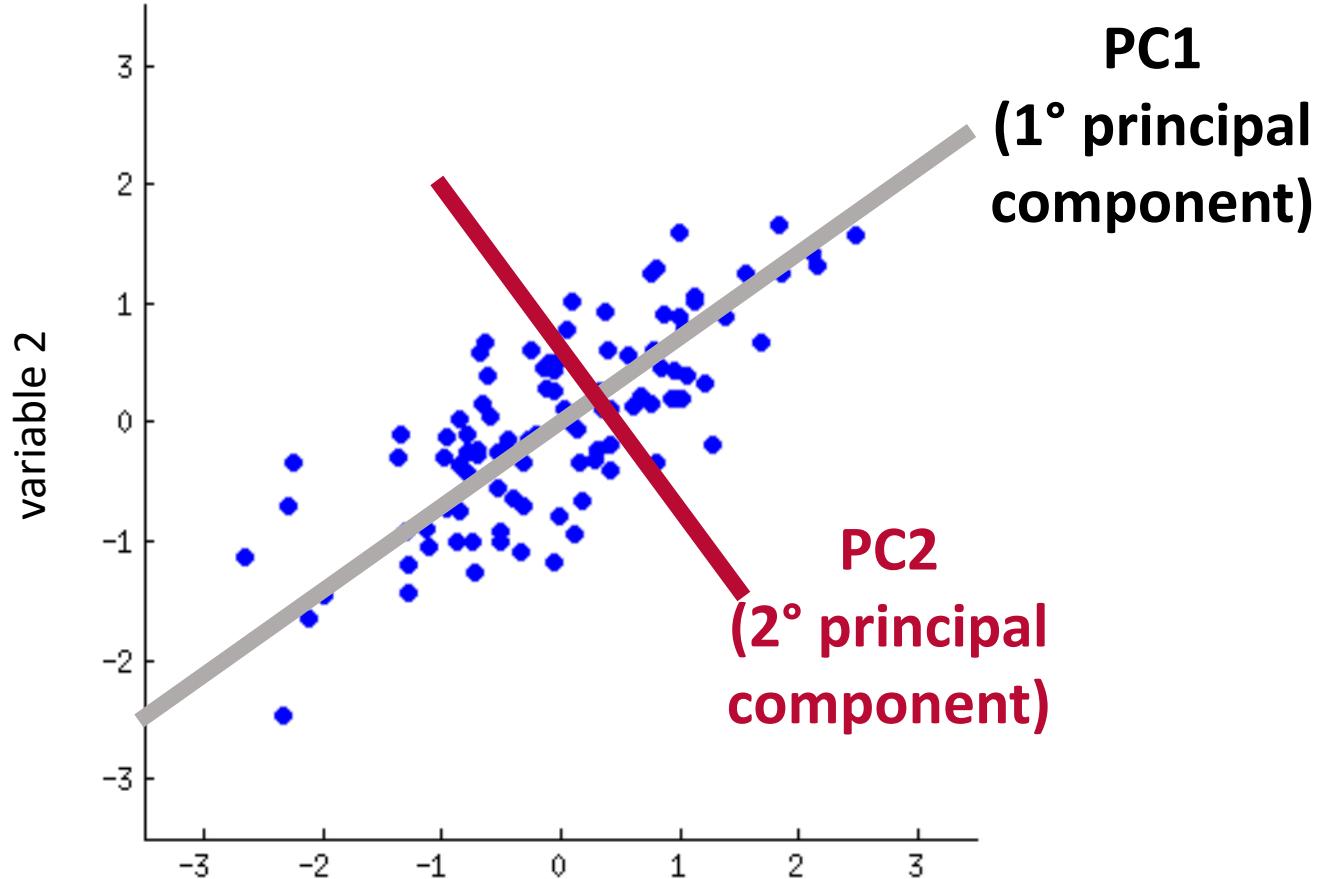
Animations from:
<https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues>

PCA – in theory



PCA – in theory

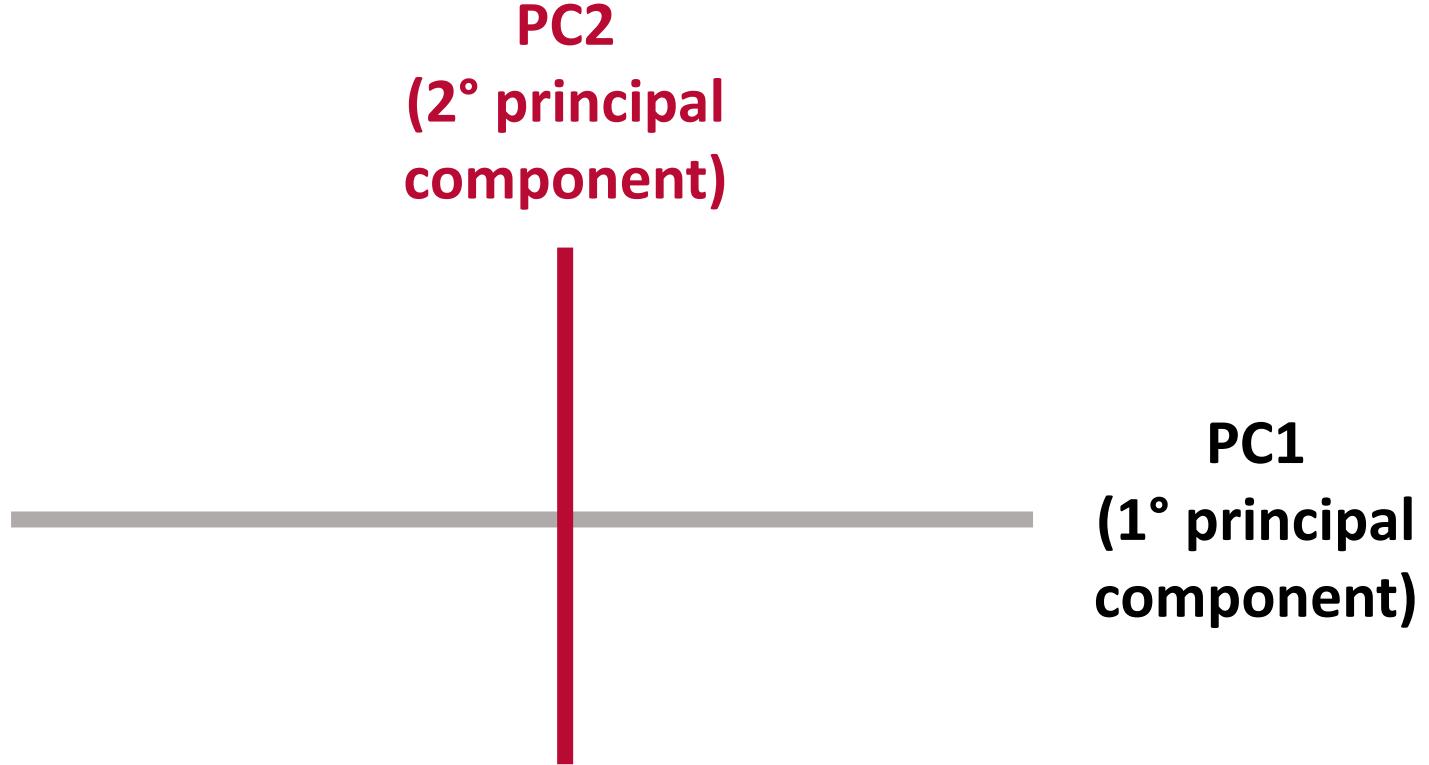
Orthogonal!



PCA – in theory

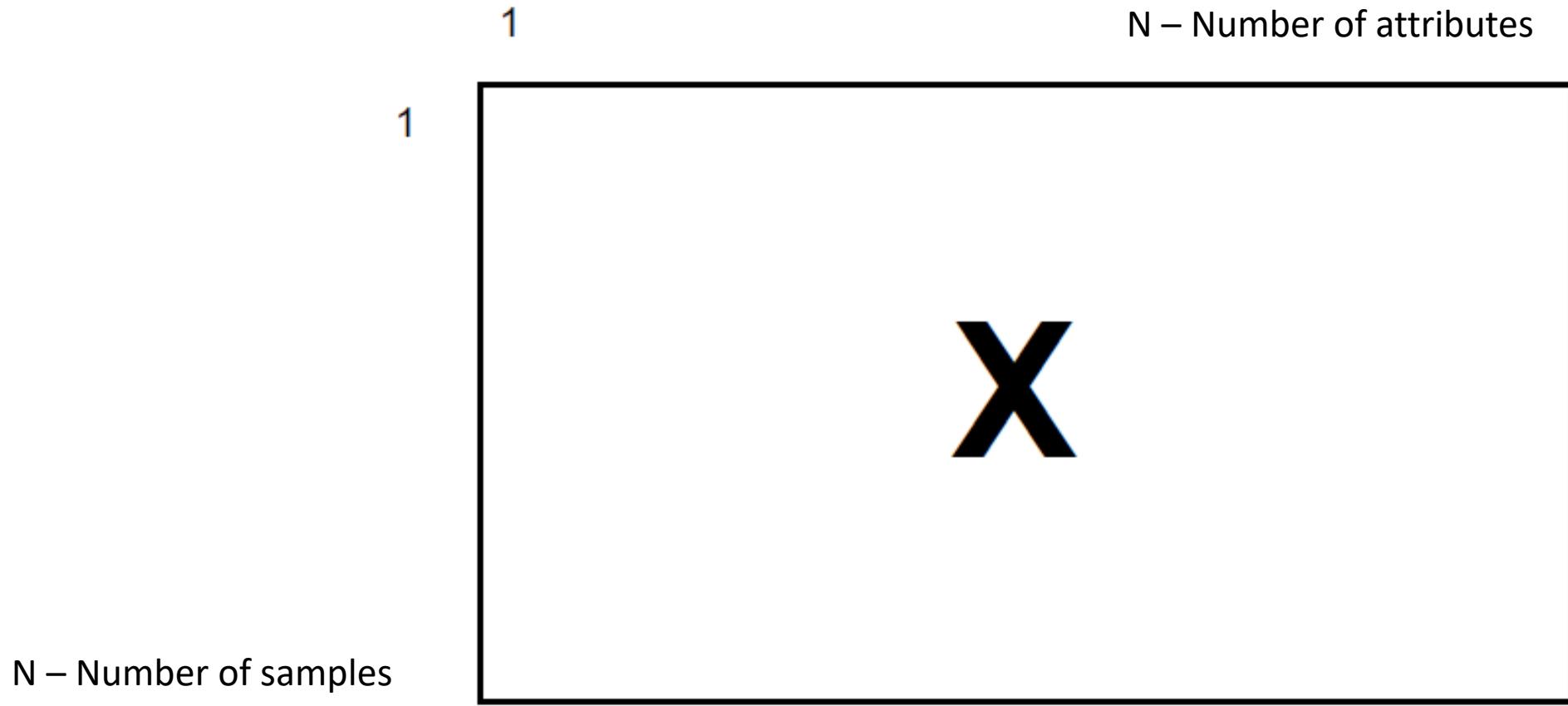
Orthogonal!

The result:



The samples are now described in a new reference system, or better,
in a **space of lower dimensions**.

PCA – in theory



PCA – in theory

$$X = TP' + E$$

P: Loadings
T: Scores E: Error

From a mathematical point of view, the **Principal Components** are built as **linear combinations of the original variables** (attributes, features)

$$t_{i1} = p_{11}x_{i1} + p_{21}x_{i2} + p_{31}x_{i3} + \dots + p_{m1}x_{im} = \mathbf{x}_i \mathbf{p}_1$$

$$t_{i2} = p_{12}x_{i1} + p_{22}x_{i2} + p_{32}x_{i3} + \dots + p_{m2}x_{im} = \mathbf{x}_i \mathbf{p}_2$$



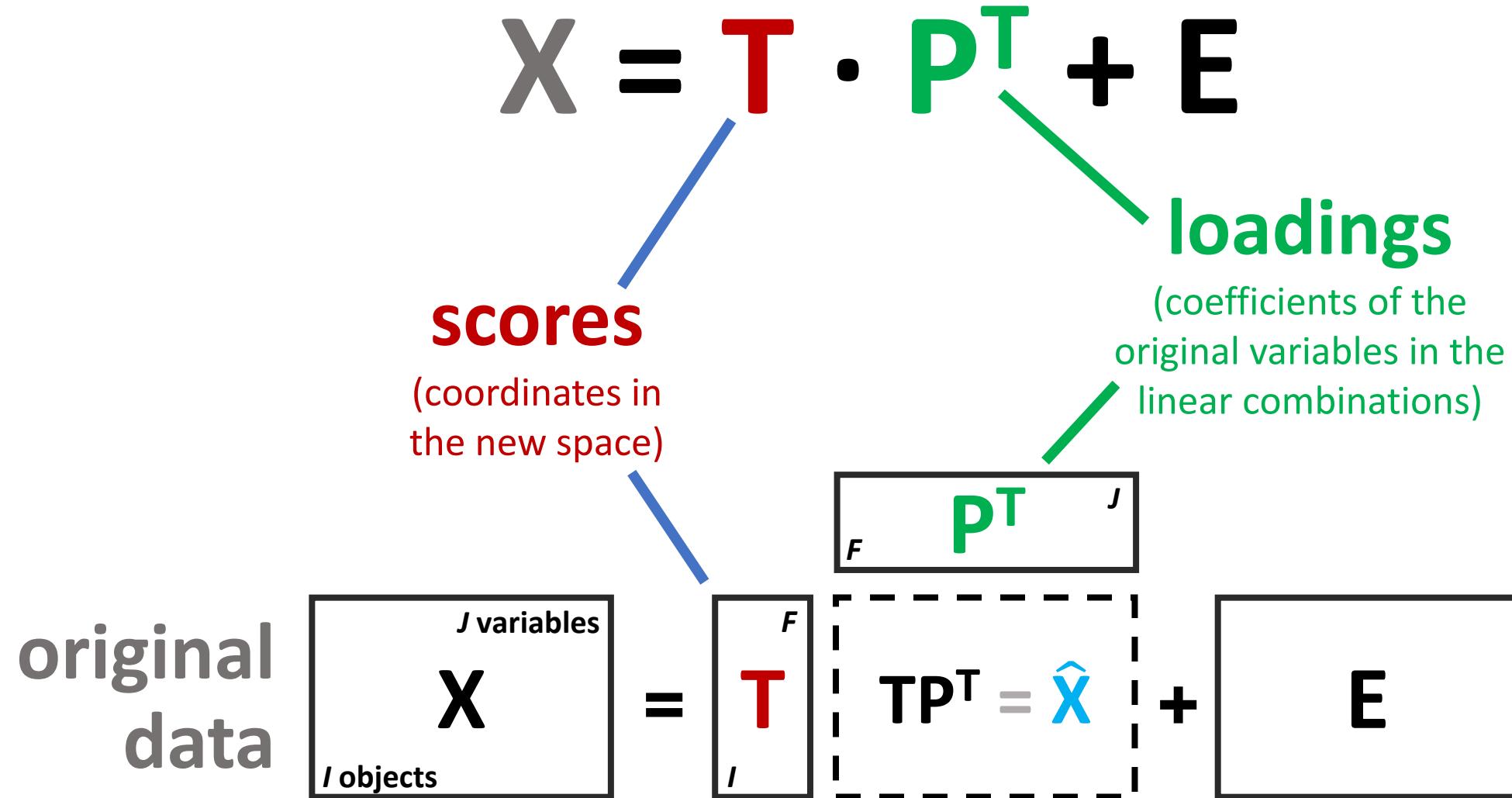
PCA = EXPLORATION

SCORES = SAMPLES

LOADINGS = VARIABLES

**SCREE PLOT = VARIANCE =
INFORMATION**

PCA – in theory



PCA – in theory

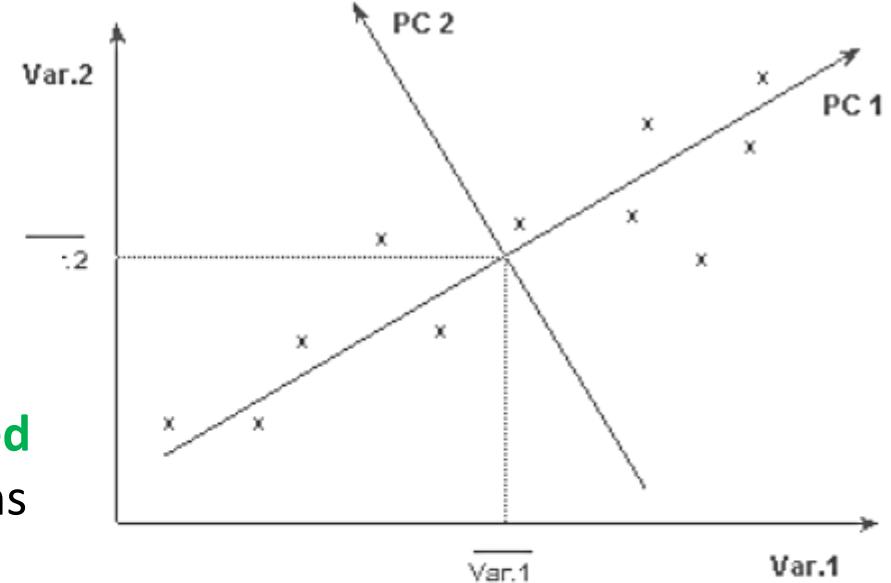
Mathematically: **Diagonalization** of the covariance matrix of the data (X)

$$\text{diag}(S) = \text{diag} \left[\frac{\mathbf{X}_c^T \mathbf{X}_c}{n-1} \right]$$

The **diagonalization** of the **covariance matrix** provides the definition of a **diagonal matrix Λ (p,p)** named as **eigenvalues matrix**, whose diagonal elements are called **eigenvalues λ_m** , ranked in decreasing order, and a **loadings matrix L (p,M)**, whose columns are the **eigenvectors l_m** of the covariance matrix

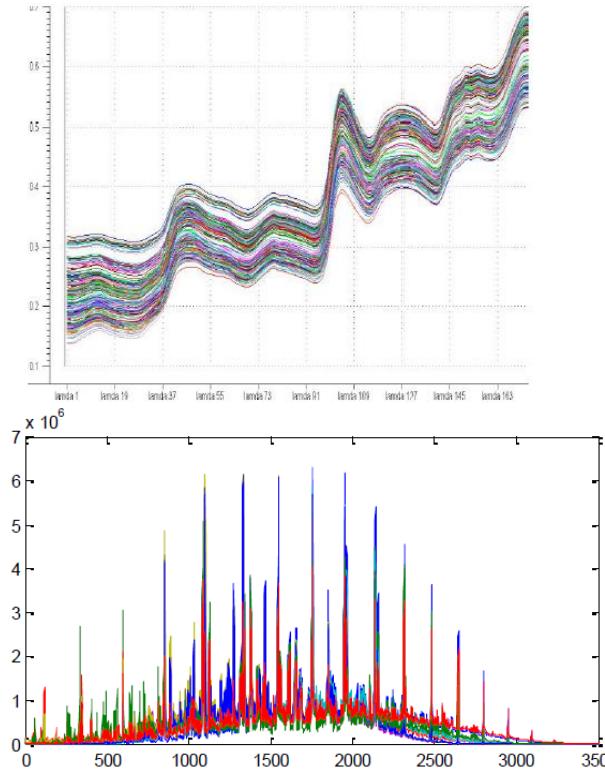
$$S = L \cdot \Lambda \cdot L^T = \sum_p \lambda_p l_p l_p^T$$

$$(n, M) = (n, p) (p, M)$$



PCA: $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}$

PCA – in theory



original
data

$$X = T \cdot P^T + E$$

scores
(coordinates in the new space)

loadings
(coefficients of the original features in the linear combinations)

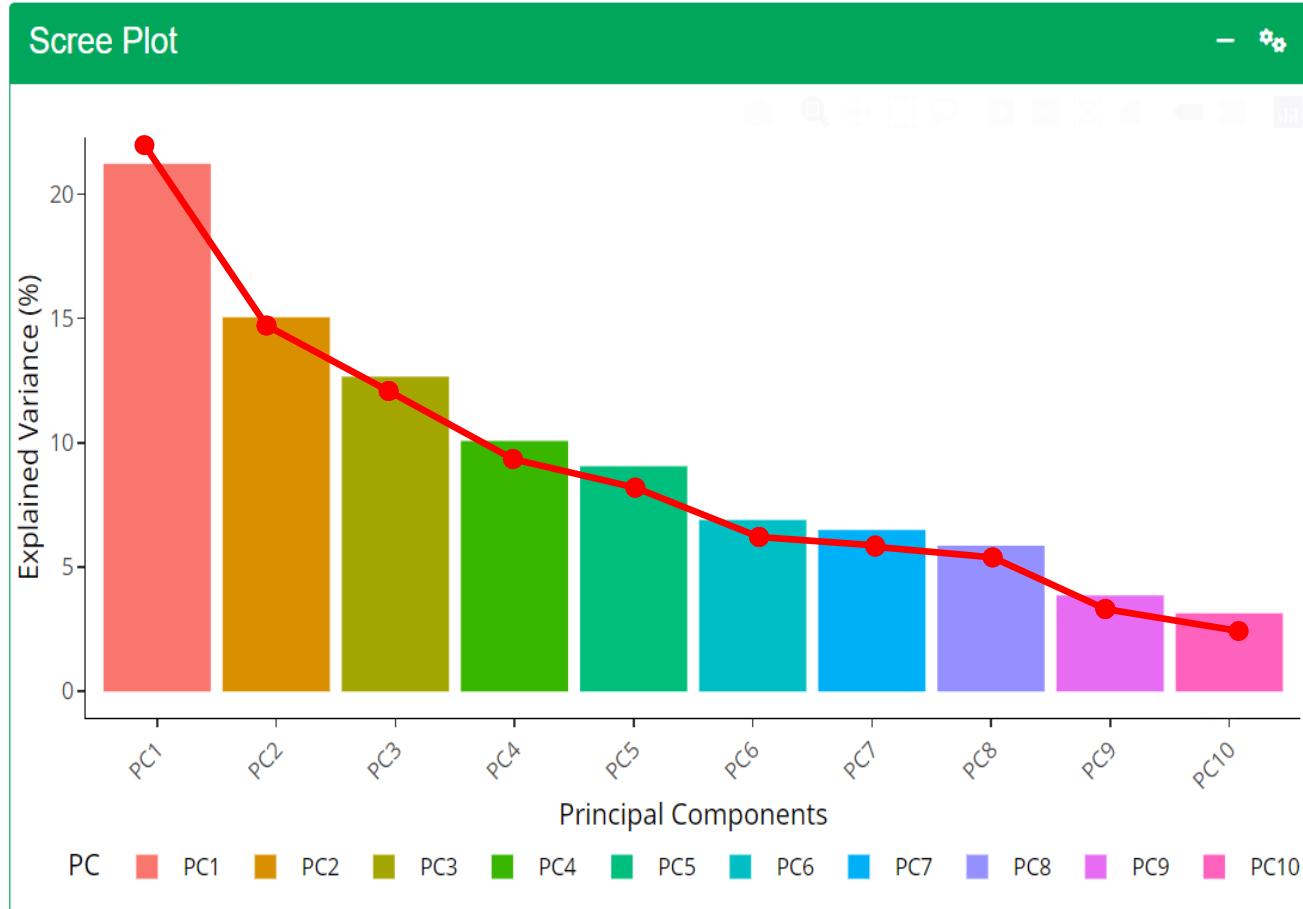
X = T P^T E
 $I \times J$ $F \times F$ $J \times I$

T = T P^T E
 $I \times F$ $F \times J$ $I \times I$

P^T = T P^T E
 $J \times F$ $F \times J$ $J \times I$

\hat{X} = $T P^T$ E
 $I \times J$ $F \times J$ $I \times I$

PCA – select the number of components

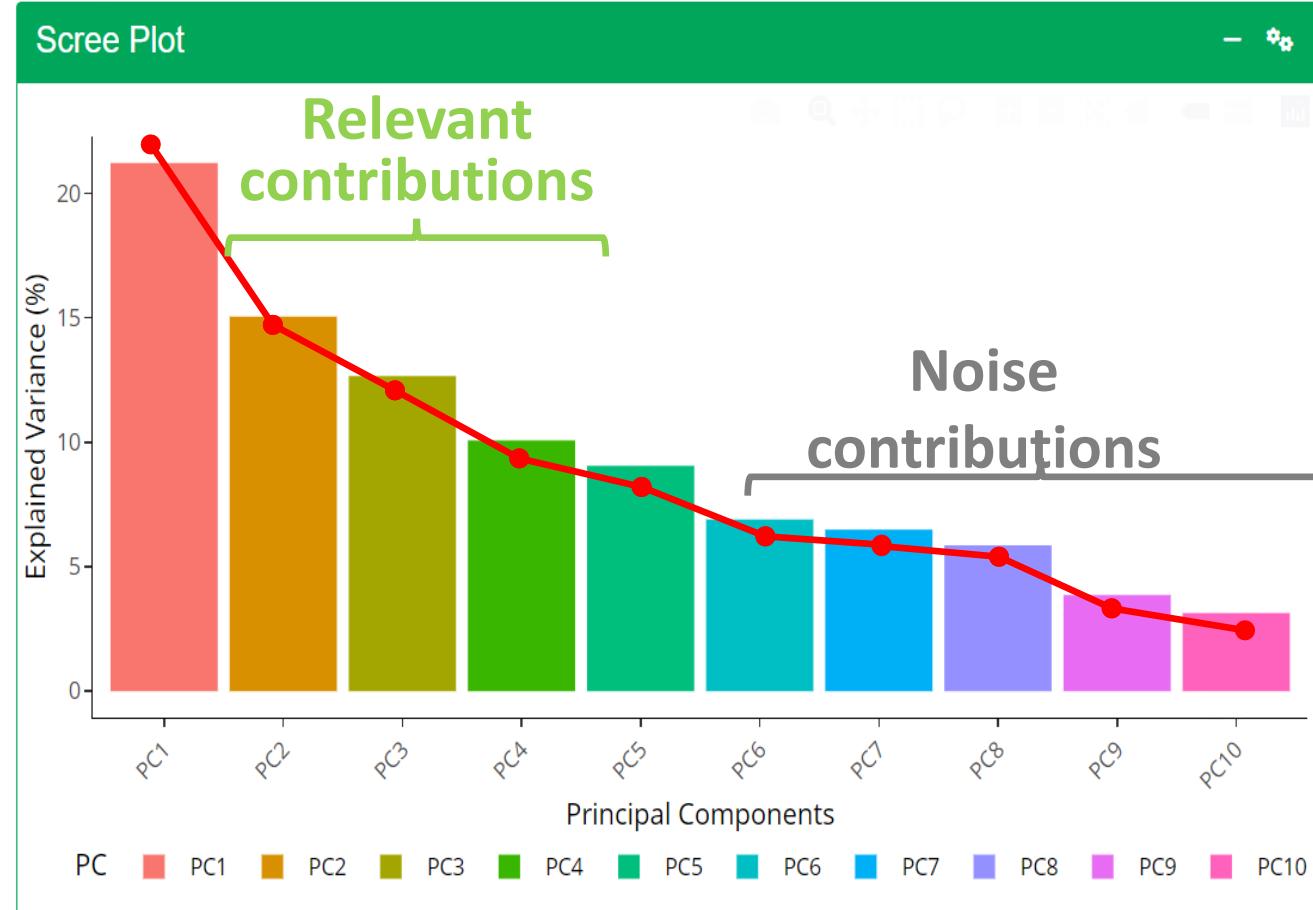


This is the eigenvalues “**scree plot**”.

Each PC has a corresponding eigenvalue ●.

The magnitude of the eigenvalue is related to the variance explained by the PC.

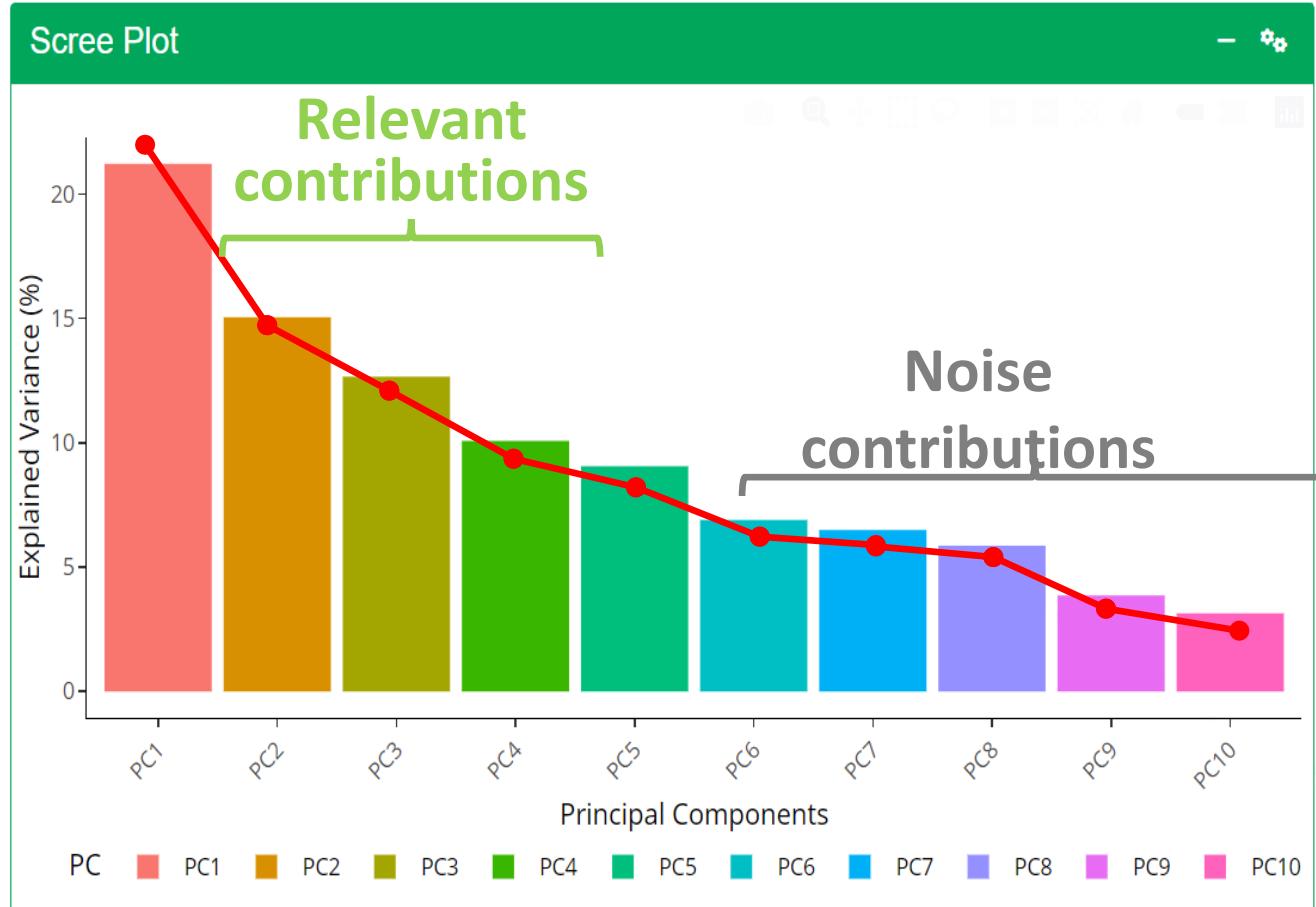
PCA – select the number of components



The magnitude of the eigenvalue is related to the variance explained by the PC.

It is possible to estimate the correct number of components, i.e. how many of them carry information, and how many are related to random variability (= noise!).

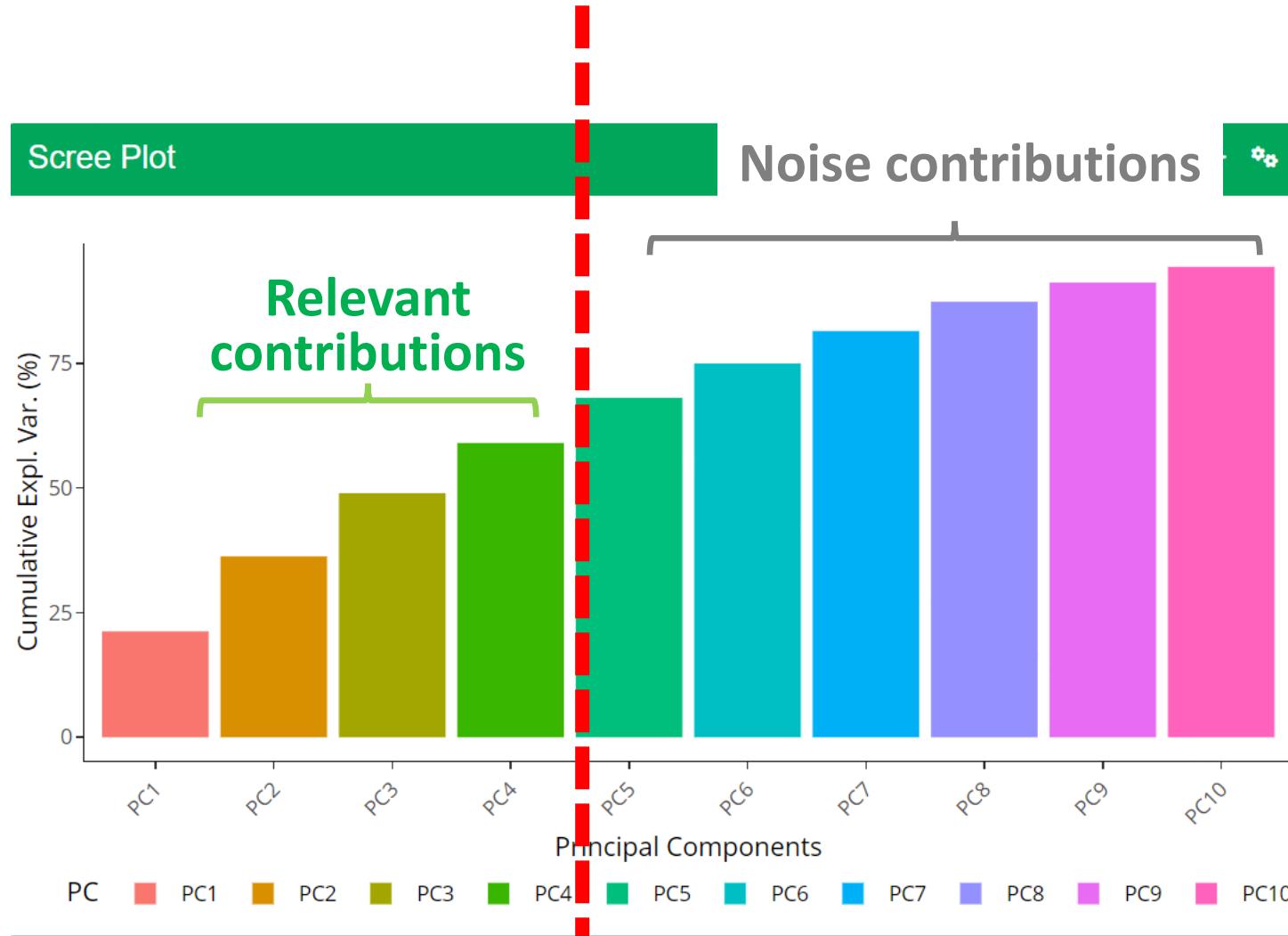
PCA – select the number of components



It is important to consider that often there is not a correct number of components, but also choosing:

- **Too few → underfitted model**
(i.e. not all relevant information is included)
- **Too many → overfitted model**
(i.e. the model perfectly describes the data, and would not model well new samples → becomes useless!)

PCA – select the number of components

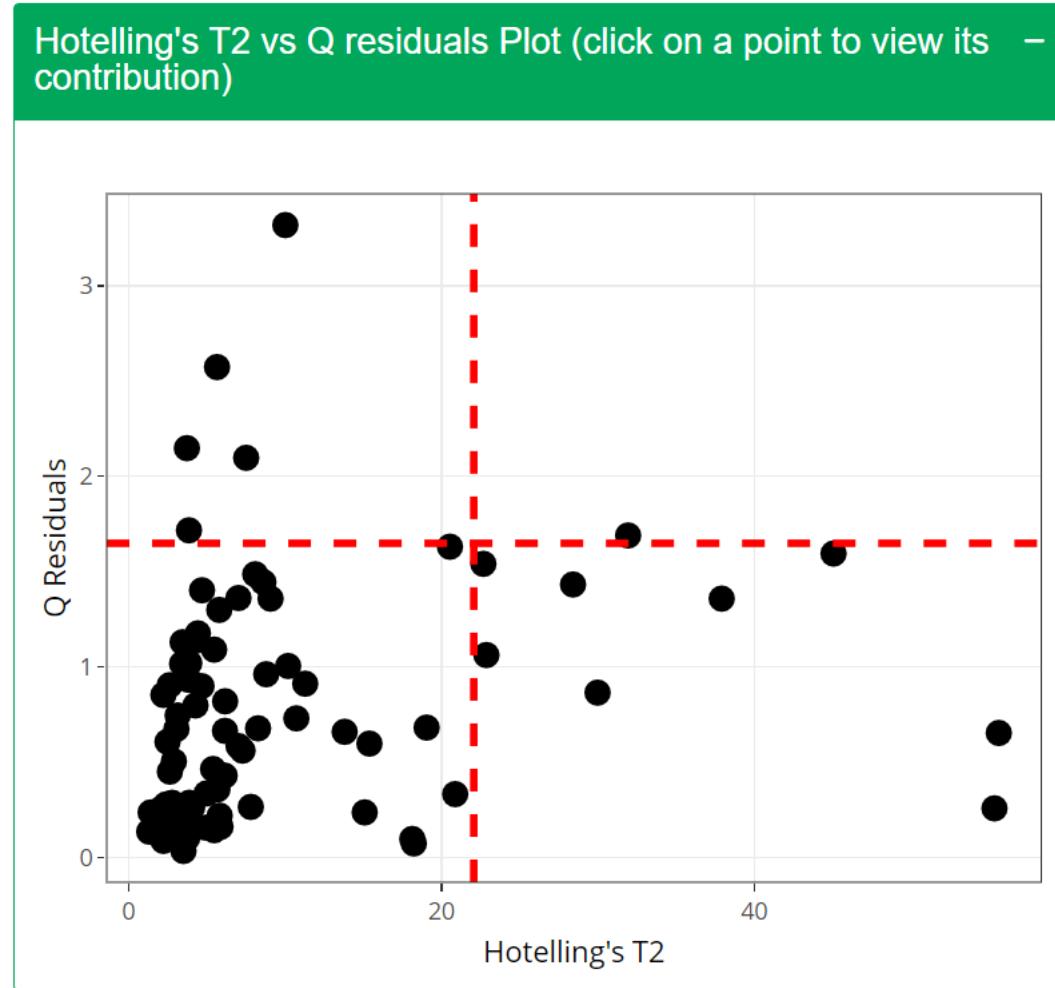


Another way to put it, is the **cumulative variance bar plot**, which can be used to estimate the number of components by looking where a **plateau** is reached.

A "conservative" choice:
fit a new model using 4 components.

Anomalous data

Q residuals



T² residuals ("Hotelling")

This is the **residuals plot**.

$$X = T \cdot P^T + E$$

The **Q residuals** are related to the "distance from the model's plane".

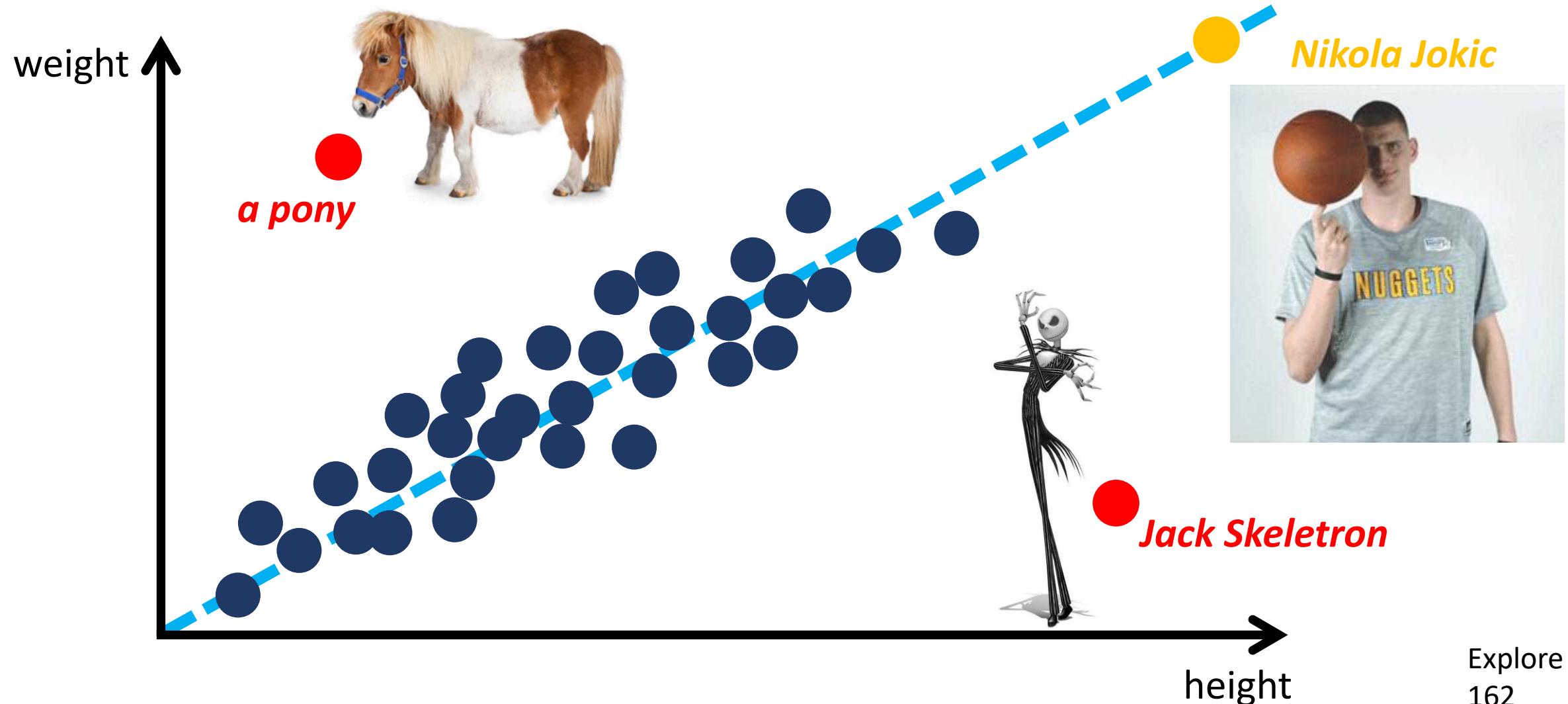
- the sample's information is not well modelled
- this is where we look for outliers

The **T² residuals** are related to the "distance within the model"

- the sample is well modelled but has "extreme" score values

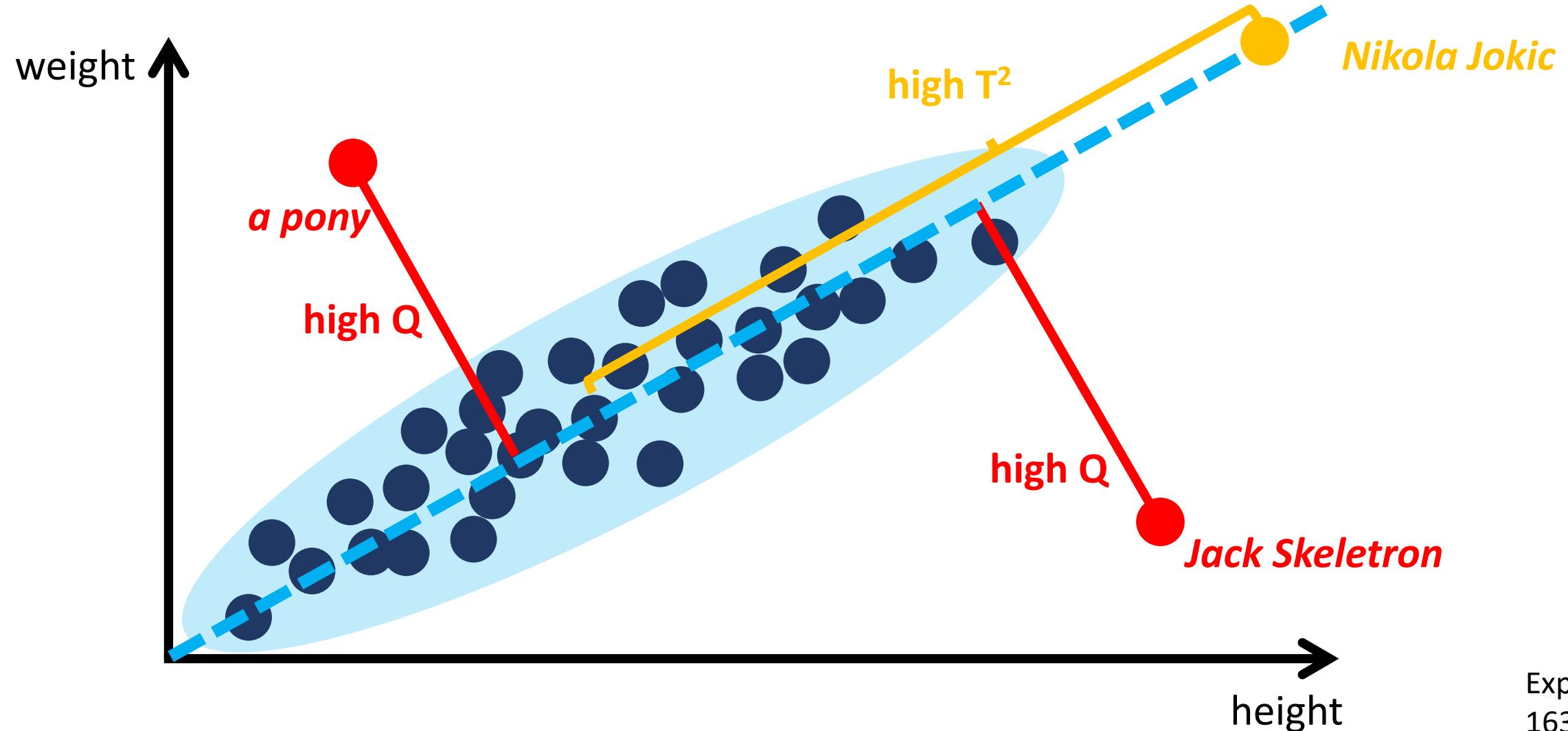
PCA – understanding the residuals

Another perspective: **a model of the human population**



PCA – understanding the residuals

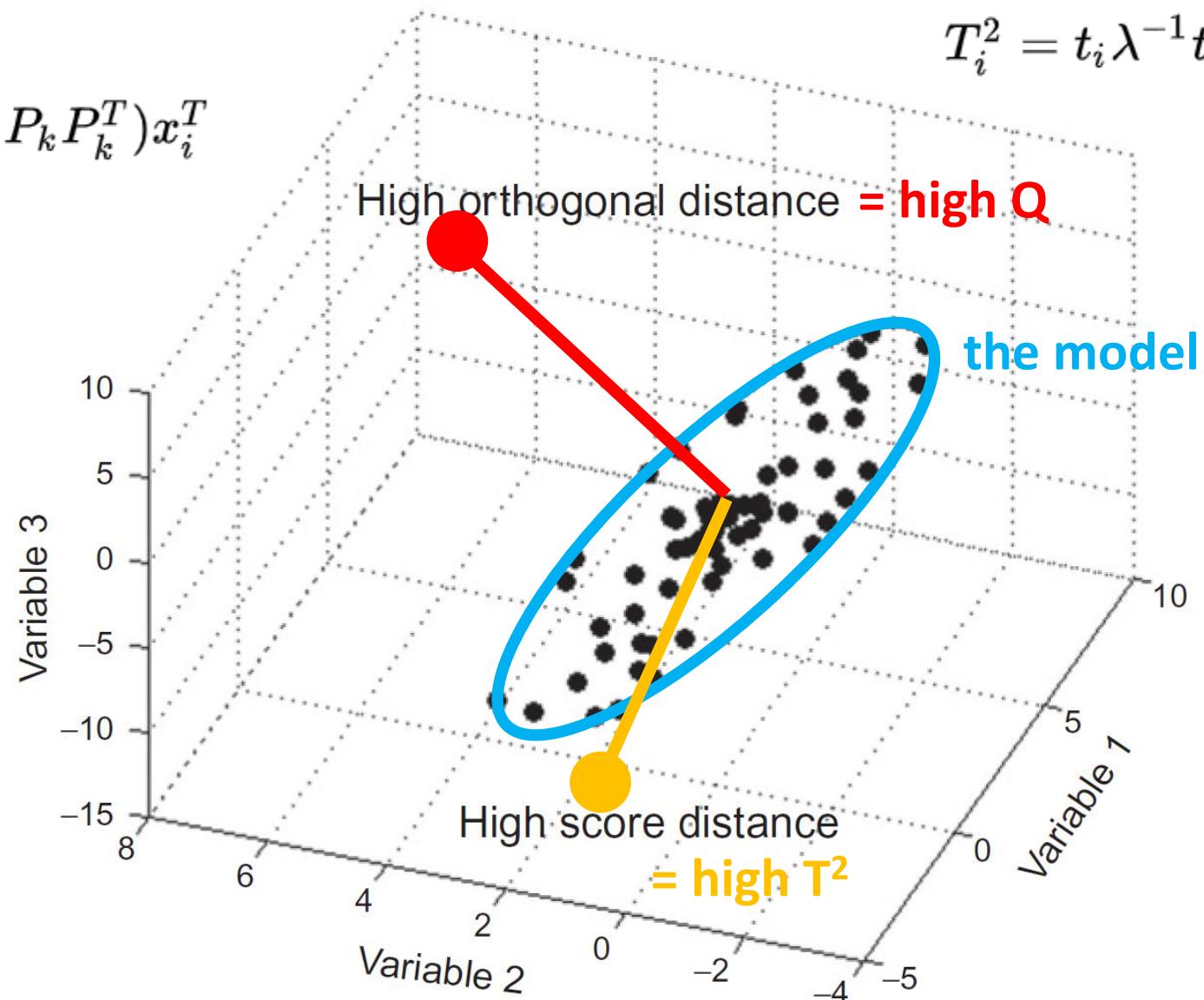
Another perspective: a model of the human population



PCA – understanding the residuals

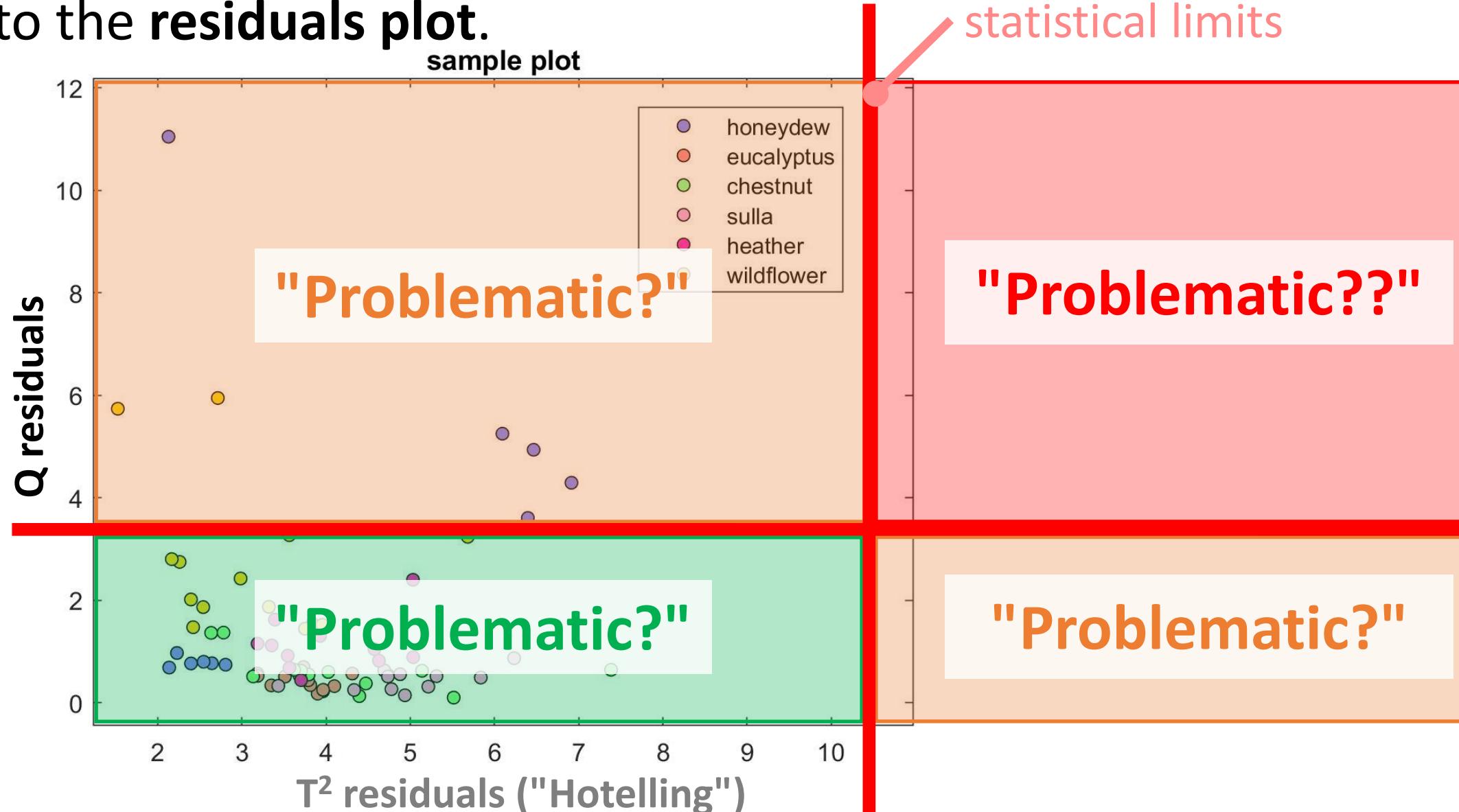
$$Q_i = e_i e_i^T = x_i (I - P_k P_k^T) x_i^T$$

$$T_i^2 = t_i \lambda^{-1} t_i^T = x_i P_k \lambda^{-1} P_k^T x_i^T$$

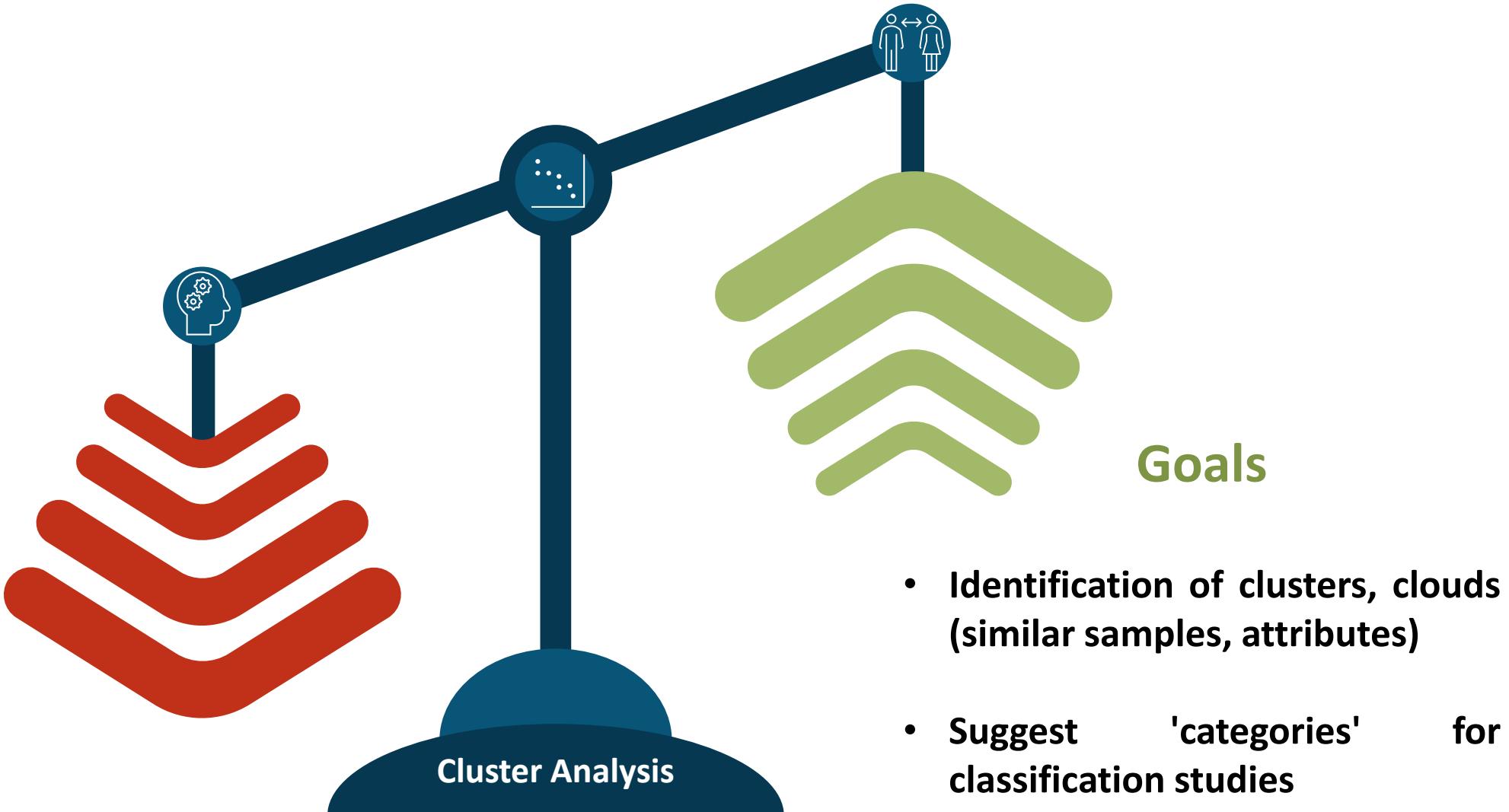


PCA – understanding the residuals

Back to the residuals plot.



Cluster analysis - introduction



Cluster analysis - introduction

Watch your
step, the SCALES
change



your
OUTCOMES!

Cluster analysis - introduction

Study of the **similarity** among objects

Attention to the **scale**!!!

(the variables may have a different nature and/or scale):
use a suitable scaling

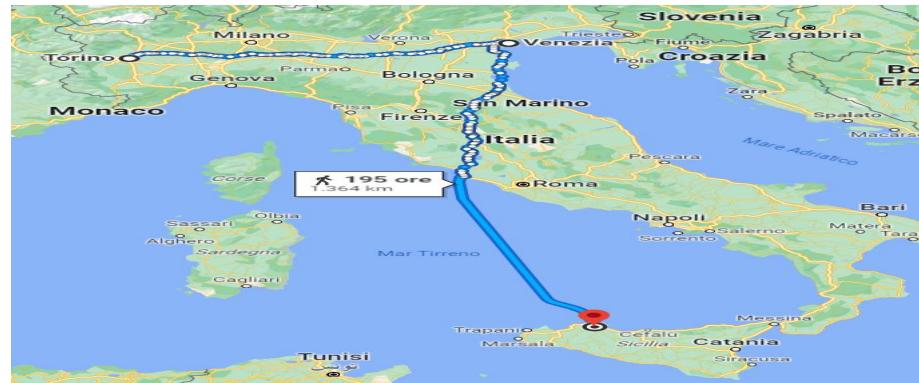
→ avoid that the proximity depends on the scale of the variables

Cluster analysis - introduction

Attention to the **scale!!!**

(the variables may have a different nature and/or scale):
use a suitable scaling

→ **avoid that the proximity depends on the scale of the variables**



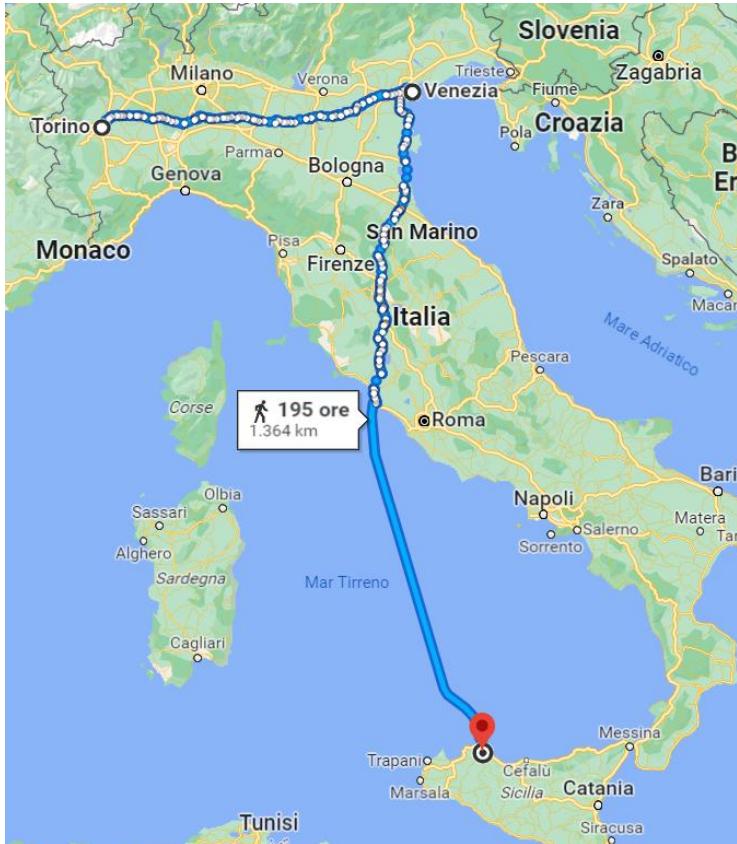
Cluster analysis - introduction

Attention to the **scale!!!**

(the variables may have a different nature and/or scale):

use a suitable scaling

→ **avoid that the proximity depends on the scale of the variables**



Distances

$$1. \quad d_{st} = \sqrt{\sum_{j=1}^p (x_{sj} - x_{tj})^2}$$

$$2. \quad d_{st} = \sum_j |x_{sj} - x_{tj}|$$

$$3. \quad d_{st} = \max_j |x_{sj} - x_{tj}|$$

$$4. \quad d_{st} = \sum_{j=1}^p \frac{|x_{sj} - x_{tj}|}{(x_{sj} + x_{tj})}$$

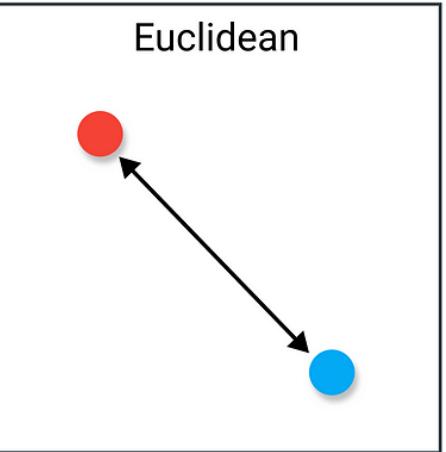
$$5. \quad d_{st} = \frac{\sum_j |x_{sj} - x_{tj}|}{\sum_j (x_{sj} + x_{tj})}$$

$$6. \quad d_{st} = \sqrt{r \sum_j |x_{sj} - x_{tj}|^r}$$

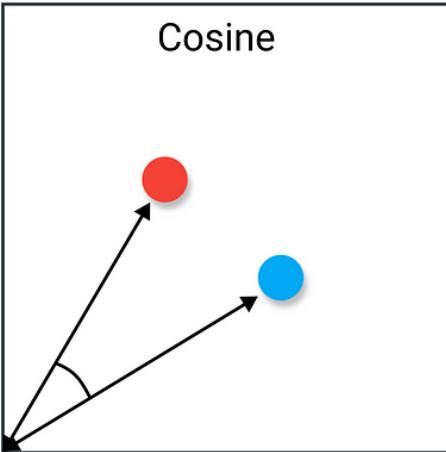
$$7. \quad d_{st} = \sqrt{(\mathbf{x}_s - \mathbf{x}_t)^T \mathbf{S}^{-1} (\mathbf{x}_s - \mathbf{x}_t)}$$

$$8. \quad d_{st} = \sqrt{\sum_j \frac{(x_{sj} - x_{tj})^2}{s_j^2}}$$

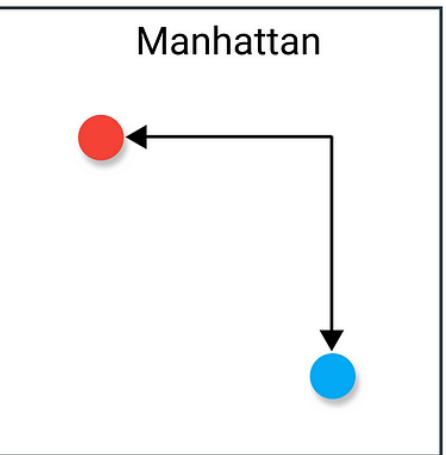
Euclidean distance



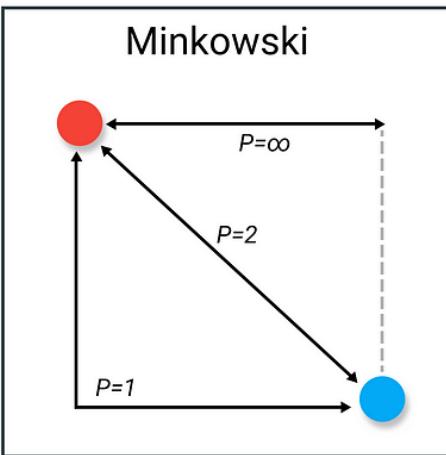
Manhattan distance



Chebyshev/Lagrange distance



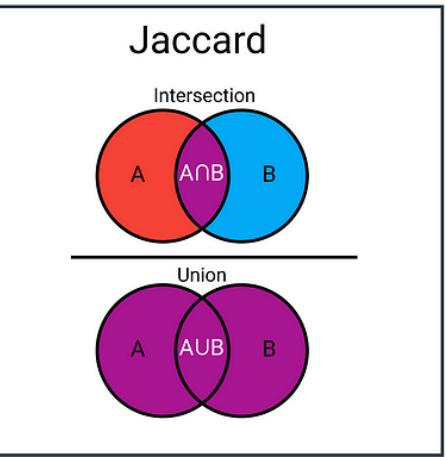
Camberra distance



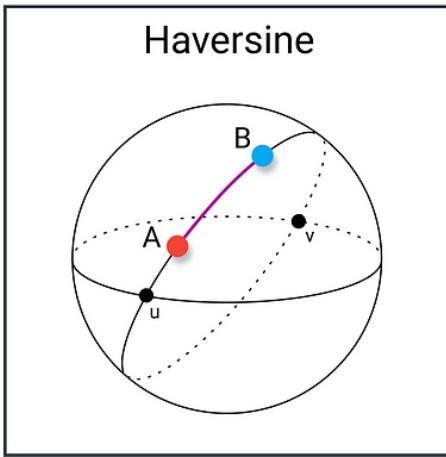
Lance-Williams distance

	Hamming																		
A	<table border="1"> <tr><td>1</td><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>↑</td><td></td><td></td><td></td><td>↑</td><td></td></tr> <tr><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td></tr> </table>	1	0	1	1	0	0	↑				↑		1	1	1	0	0	0
1	0	1	1	0	0														
↑				↑															
1	1	1	0	0	0														

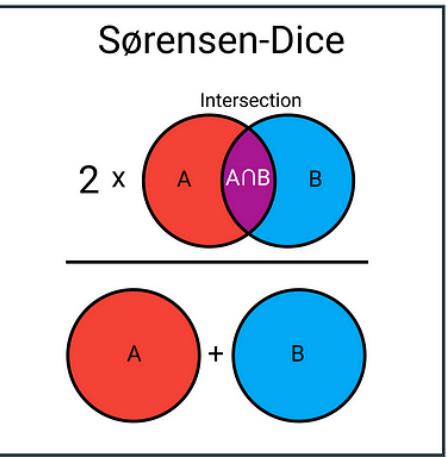
Minkowski distance



Mahalanobis distance



Pearson distance



Distances

$$1. \quad d_{st} = \sqrt{\sum_{j=1}^p (x_{sj} - x_{tj})^2}$$

$$2. \quad d_{st} = \sum_j |x_{sj} - x_{tj}|$$

$$3. \quad d_{st} = \max_j |x_{sj} - x_{tj}|$$

$$4. \quad d_{st} = \sum_{j=1}^p \frac{|x_{sj} - x_{tj}|}{(x_{sj} + x_{tj})}$$

$$5. \quad d_{st} = \frac{\sum_j |x_{sj} - x_{tj}|}{\sum_j (x_{sj} + x_{tj})}$$

$$6. \quad d_{st} = \sqrt[r]{\sum_j |x_{sj} - x_{tj}|^r}$$

$$7. \quad d_{st} = \sqrt{(\mathbf{x}_s - \mathbf{x}_t)^T \mathbf{S}^{-1} (\mathbf{x}_s - \mathbf{x}_t)} \quad \text{Mahalanobis distance}$$

$$8. \quad d_{st} = \sqrt{\sum_j \frac{(x_{sj} - x_{tj})^2}{s_j^2}}$$

Euclidean distance

Manhattan distance

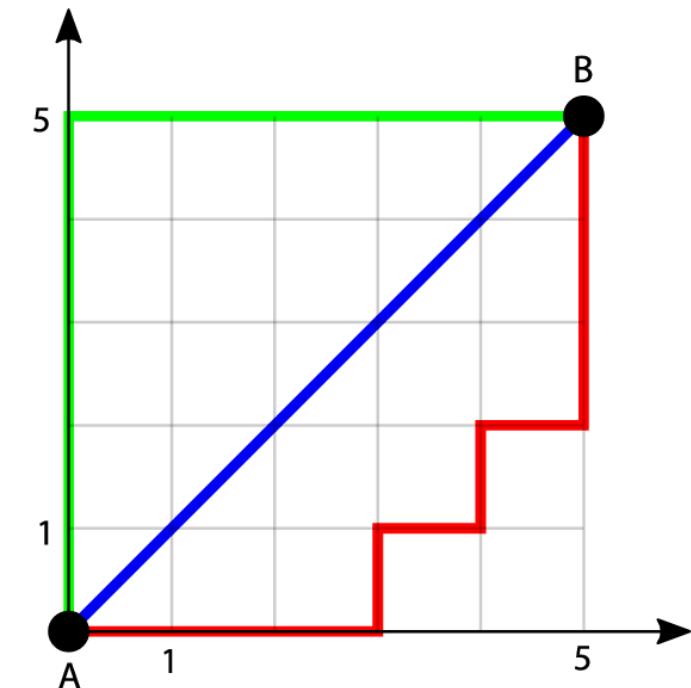
Chebyshev/Lagrange distance

Camberra distance

Lance-Williams distance

Minkowski distance

Pearson distance



— Euclidean distance

— Manhattan distance

Kendall and Spearman distances are non-parametric (rank)

Similarity

The **similarity** between two samples (i and j) or two variables is defined as:

$$s_{ij} = 1 - \frac{d_{ij}}{d_{MAX}}$$



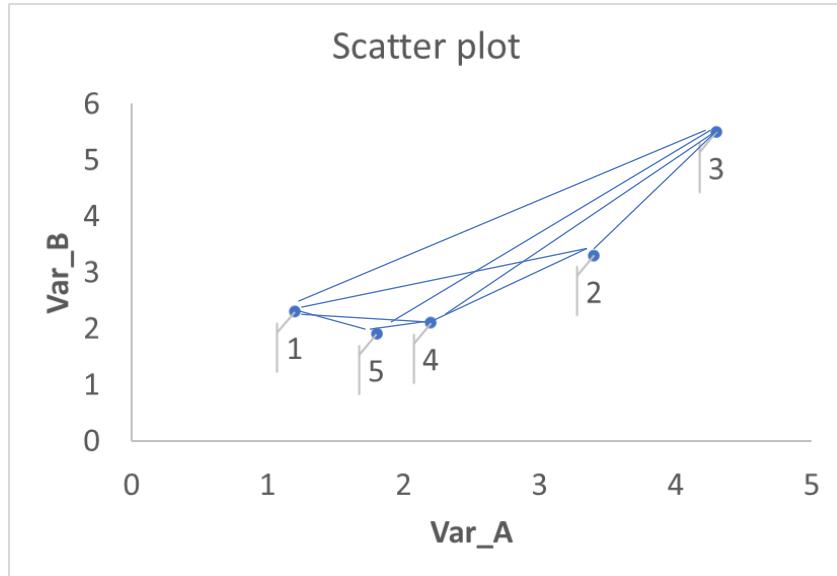
d_{ij} is the **distance**(Euclidean, Mahalanobis, Minkowski...) between two samples (variables);

d_{MAX} is the **max distance** between two samples (variables) of the dataset;

Two samples with the max distance have a similarity = 0

Similarity

CALCULATING THE DISTANCES AMONG OBJECTS



BUILDING OF THE DISTANCE MATRIX

DISTANCE MATRIX

0.00	2.42	4.46	1.02	0.72
2.42	0.00	2.38	1.70	2.13
4.46	2.38	0.00	4.00	4.38
1.02	1.70	4.00	0.00	0.45
0.72	2.13	4.38	0.45	0.00

$$s_{ij} = 1 - \frac{d_{ij}}{d_{MAX}}$$

SIMILARITY MATRIX

1.00	0.46	0.00	0.77	0.84
0.46	1.00	0.47	0.62	0.52
0.00	0.47	1.00	0.10	0.02
0.77	0.62	0.10	1.00	0.90
0.84	0.52	0.02	0.90	1.00

Clustering techniques

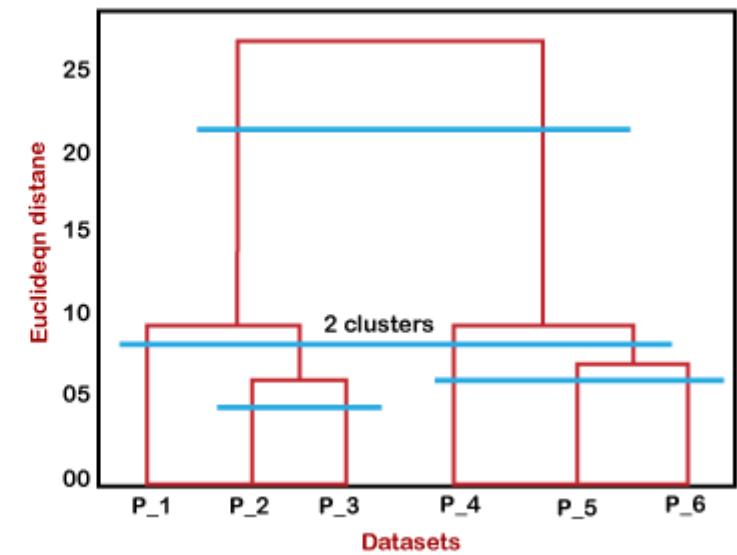
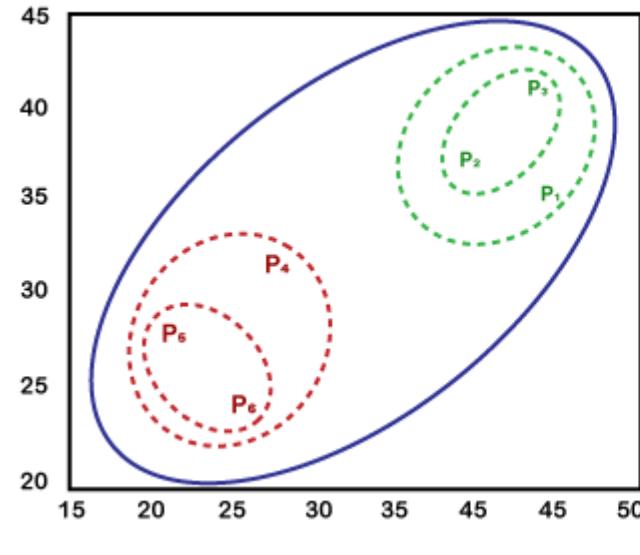
Divided into:

a) Hierarchical

a₁) Agglomerative

a₂) Divisive

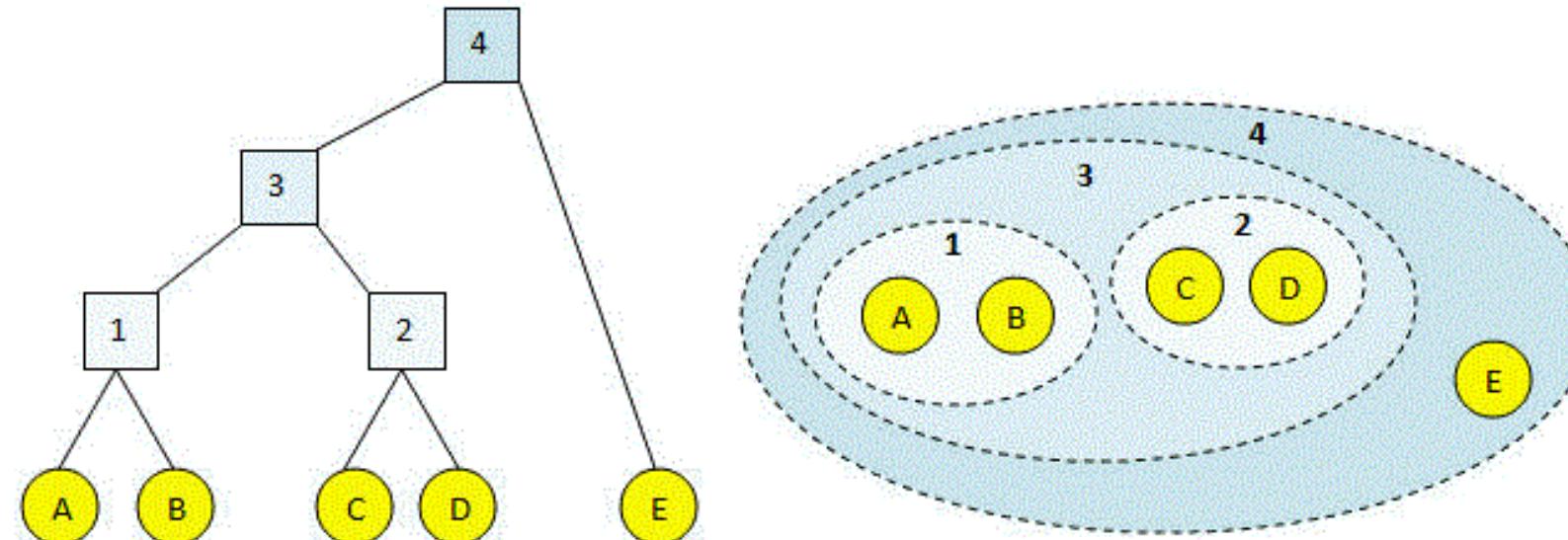
b) Non-hierarchical



Hierarchical (agglomerative)

- Each object is initially considered a cluster
- Gradually the objects are brought together in ever larger clusters

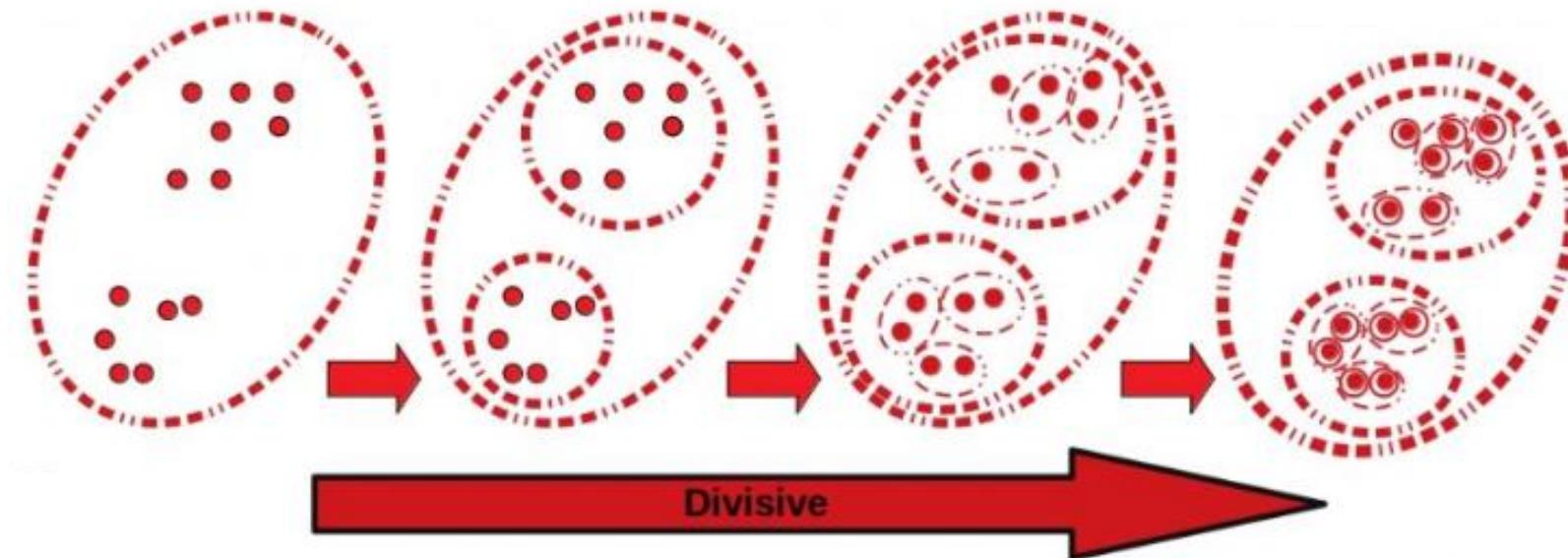
The hierarchy is a consequence of the fact that the larger clusters are always obtained by merging the smaller ones (with all their objects).



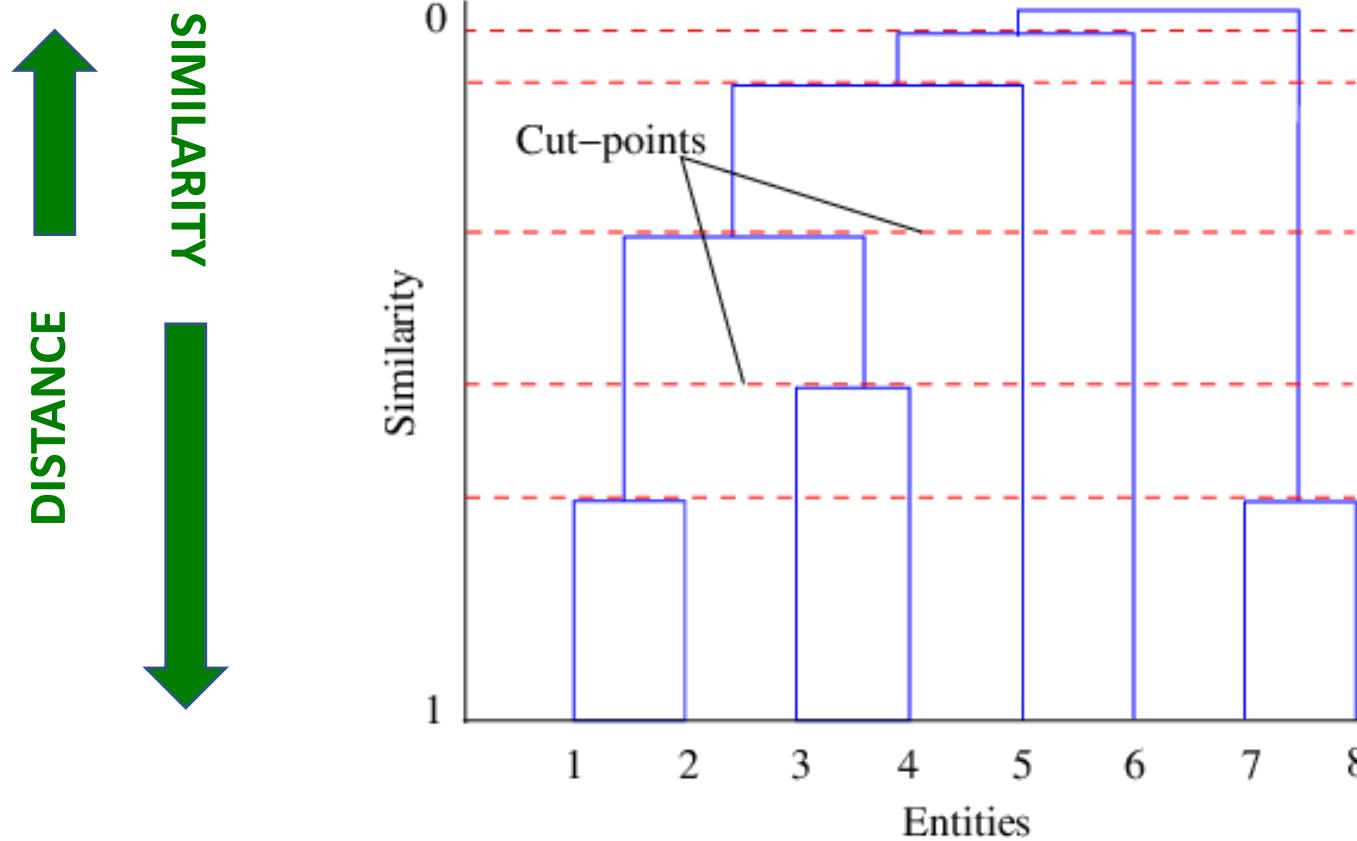
Hierarchical (Divisive)

- It starts with a large cluster containing all the objects
- Gradually the objects are inserted in small clusters obtained as subsets of the initial one

The hierarchy is a consequence of the fact that the larger clusters are always obtained by merging the smaller ones (with all their objects).



Hierarchical (Agglomerative) - dendrogram



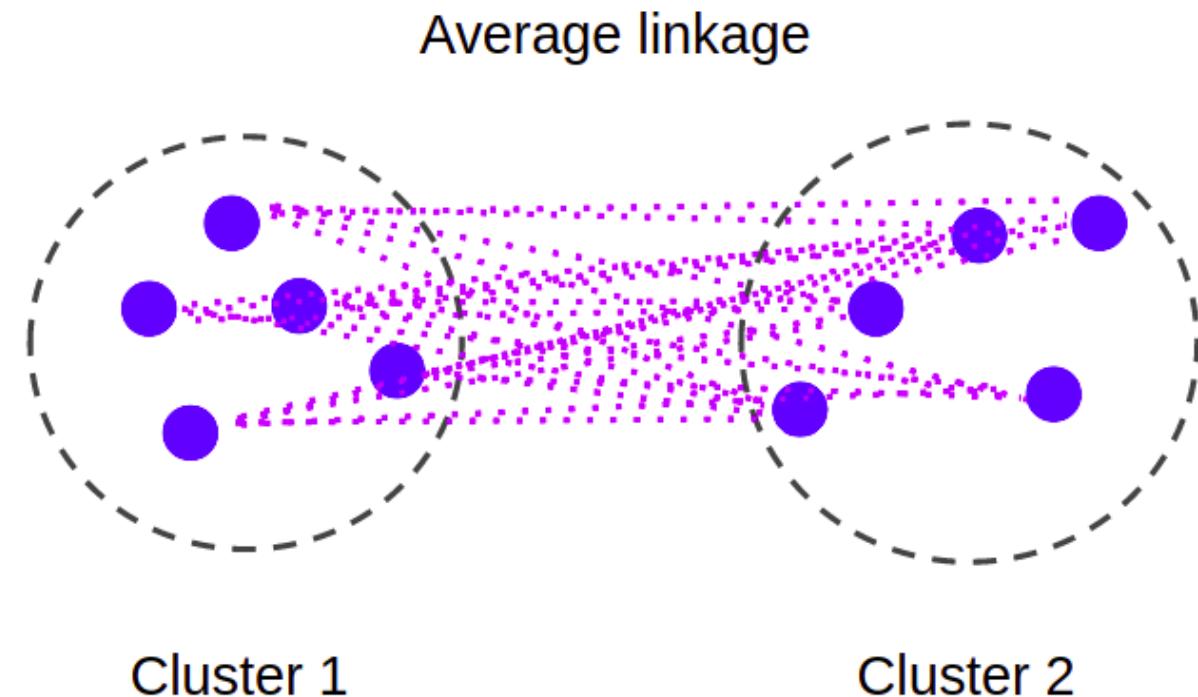
Normally the longer branches indicate the best separated clusters

The **INTERPRETATIVE PHASE** is more important than the statistical test

Hierarchical (agglomerative)

Unweighted average linkage method

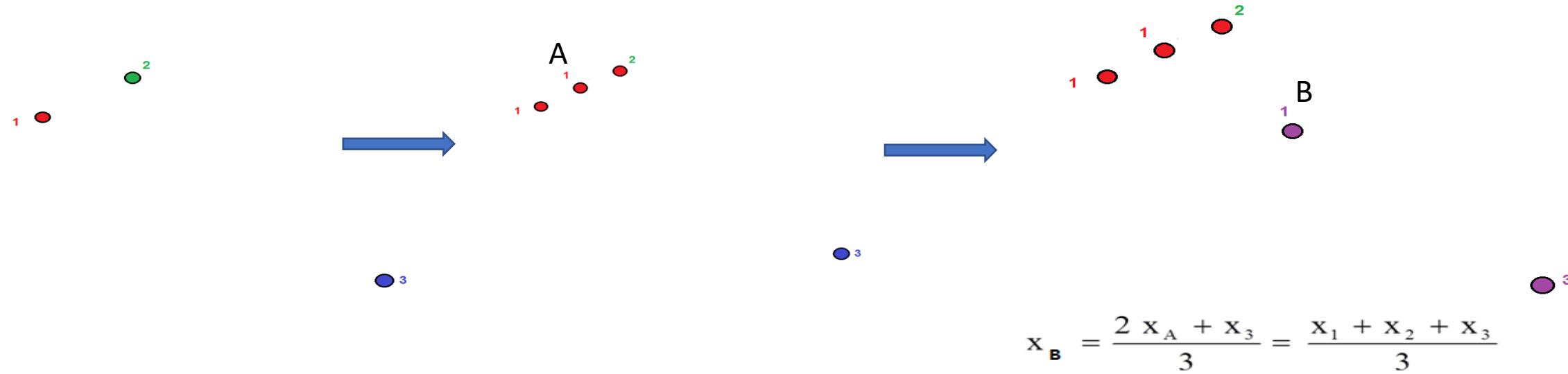
Also known as UPGMA (Unweighted Pair Group Method with Arithmetic mean). The percentage of the number of points of each cluster is calculated with respect to the number of points of the two clusters if they were merged.



Hierarchical (agglomerative)

Weighted average linkage method

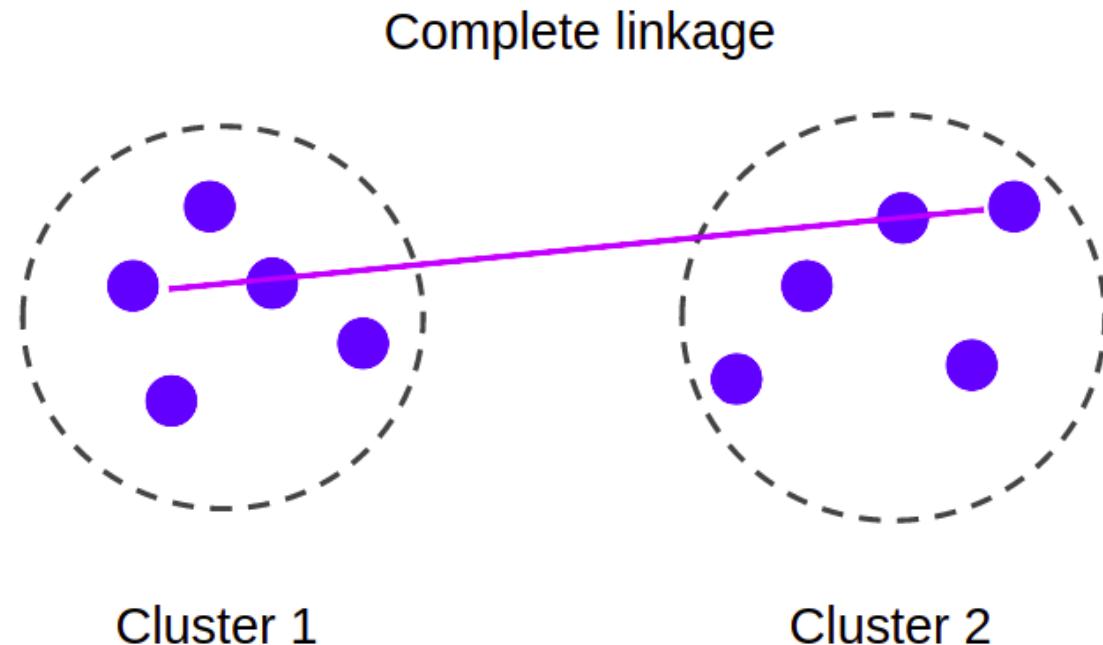
Also known as WPGMA (Weighted Pair Group Method with Arithmetic mean). The individual points of the two clusters contribute to the aggregated distance between a smaller and a bigger cluster.



Hierarchical (agglomerative)

Complete linkage method

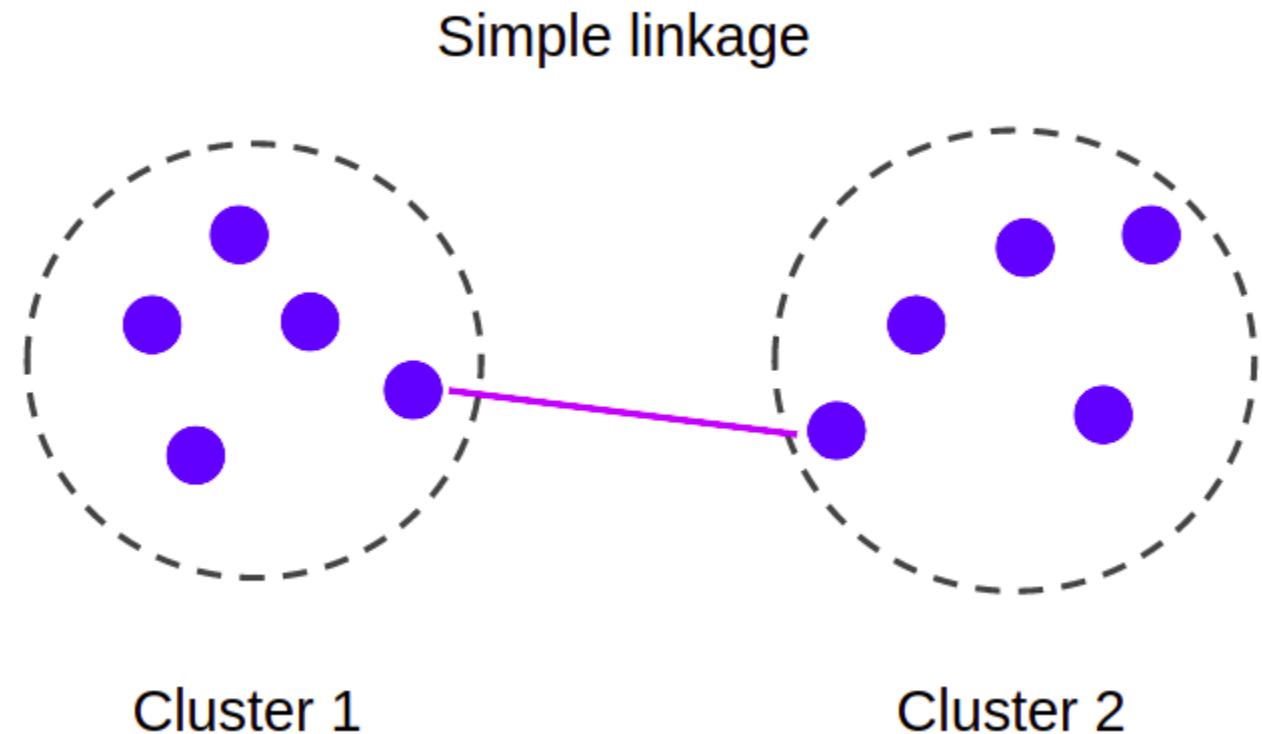
Also referred to as Furthest Neighbor (FN), Farthest Point Algorithm, or VoorHees Algorithm. The distance between clusters is defined by the distance between their furthest members. **This method is computationally expensive.**



Hierarchical (agglomerative)

Single linkage method

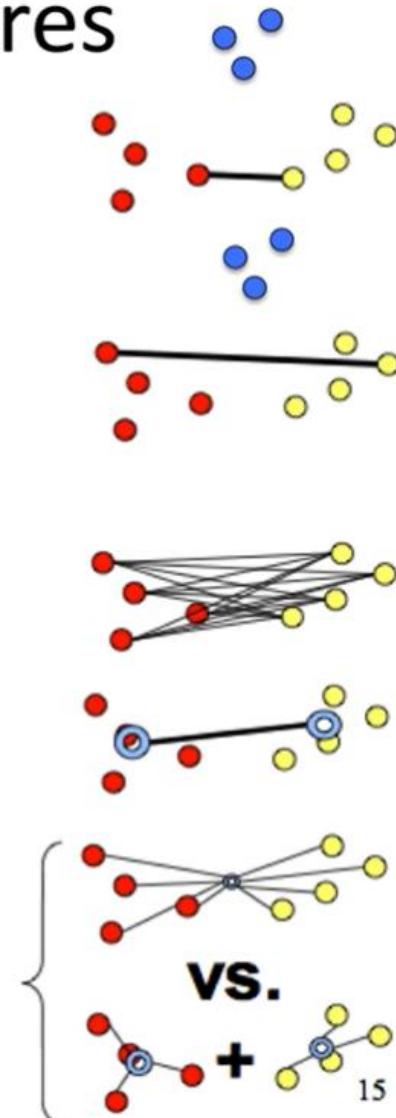
Also referred to as Nearest Neighbor (NN). The distance between clusters is defined by the distance between their closest members.



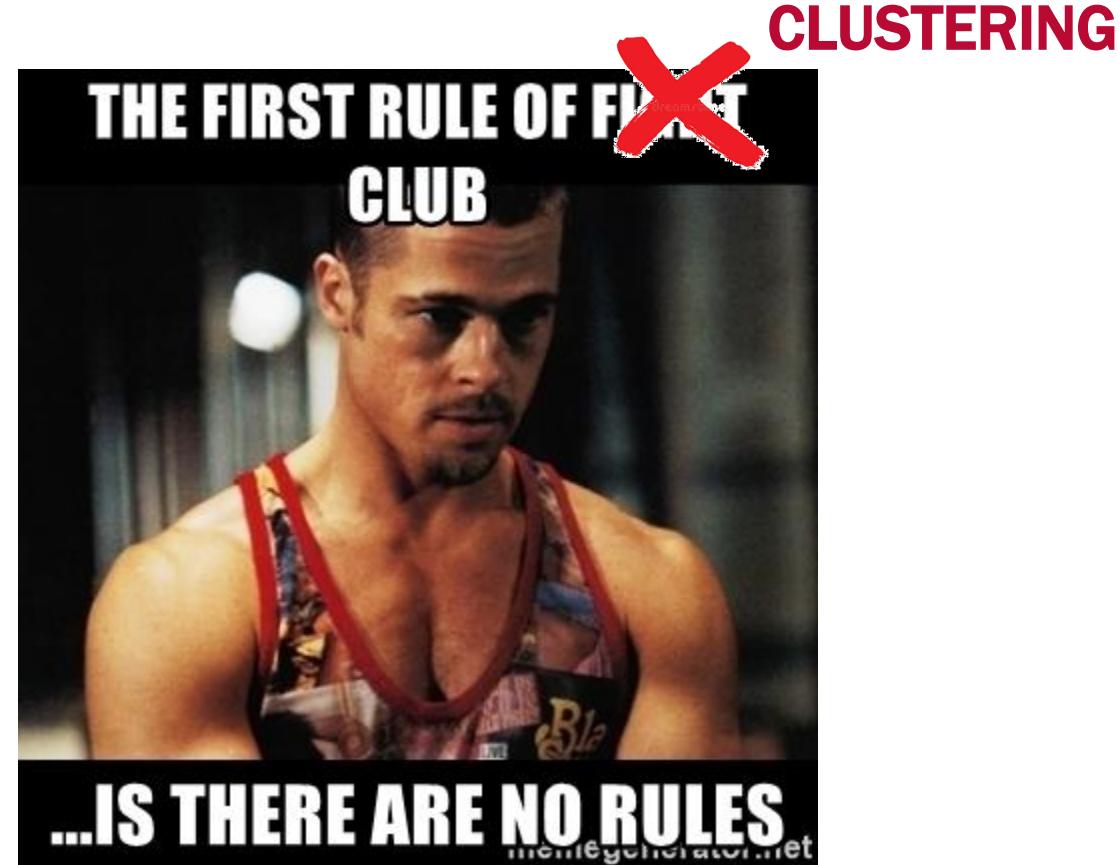
Hierarchical (agglomerative)

Cluster distance measures

- **Single link:** $D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
 - distance between closest elements in clusters
 - produces long chains a→b→c→...→z
- **Complete link:** $D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$
 - distance between farthest elements in clusters
 - forces "spherical" clusters with consistent "diameter"
- **Average link:** $D(c_1, c_2) = \frac{1}{|c_1|} \frac{1}{|c_2|} \sum_{x_1 \in c_1} \sum_{x_2 \in c_2} D(x_1, x_2)$
 - average of all pairwise distances
 - less affected by outliers
- **Centroids:** $D(c_1, c_2) = D\left(\left(\frac{1}{|c_1|} \sum_{x \in c_1} \vec{x}\right), \left(\frac{1}{|c_2|} \sum_{x \in c_2} \vec{x}\right)\right)$
 - distance between centroids (means) of two clusters
- **Ward's method:** $TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$
 - consider joining two clusters, how does it change the total distance (TD) from centroids?

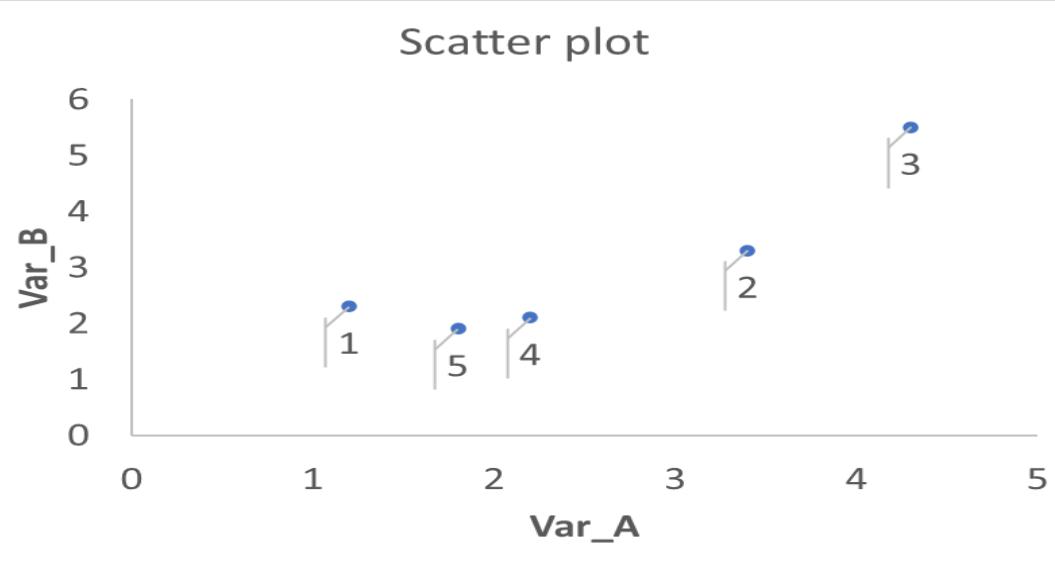


Clustering rules...



Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DATASET

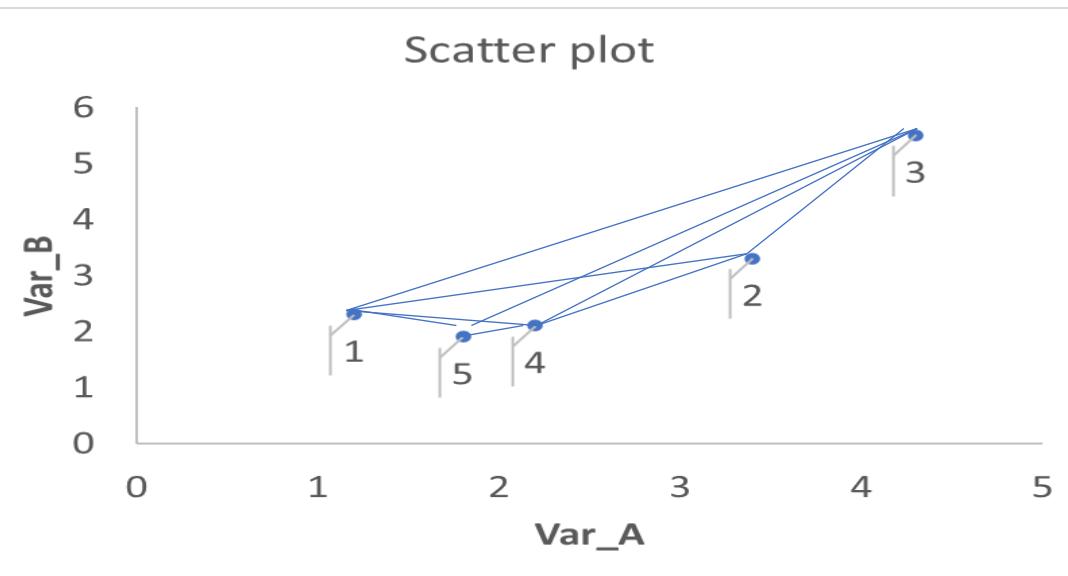
Sample	Var_A	Var_B
1	1.2	2.3
2	3.4	3.3
3	4.3	5.5
4	2.2	2.1
5	1.8	1.9

SIMILARITY MATRIX

(EUCLIDEAN DISTANCE HERE)

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

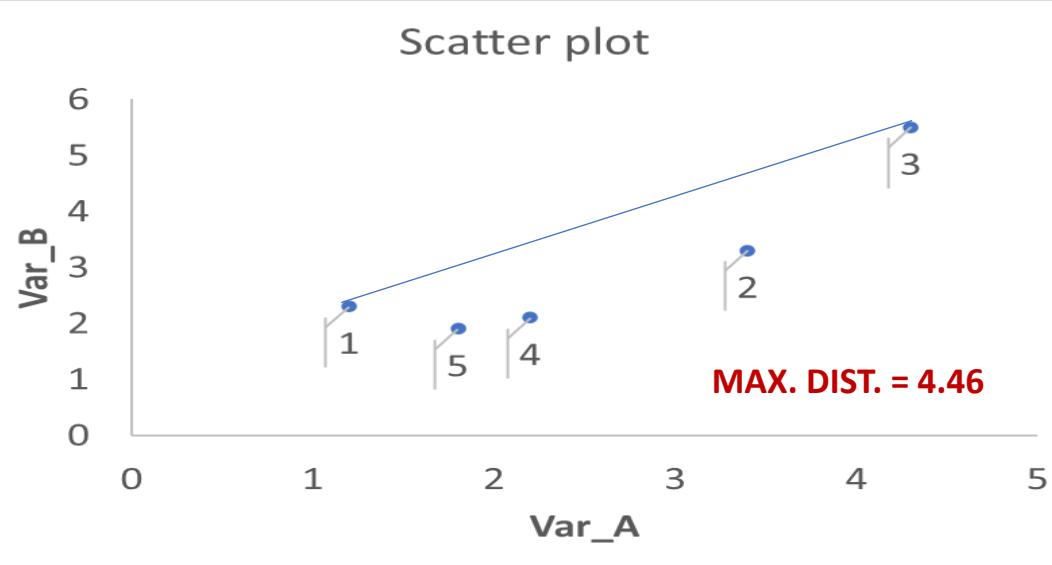
0.00	2.42	4.46	1.02	0.72
2.42	0.00	2.38	1.70	2.13
4.46	2.38	0.00	4.00	4.38
1.02	1.70	4.00	0.00	0.45
0.72	2.13	4.38	0.45	0.00

SIMILARITY MATRIX

(EUCLIDEAN DISTANCE HERE)

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

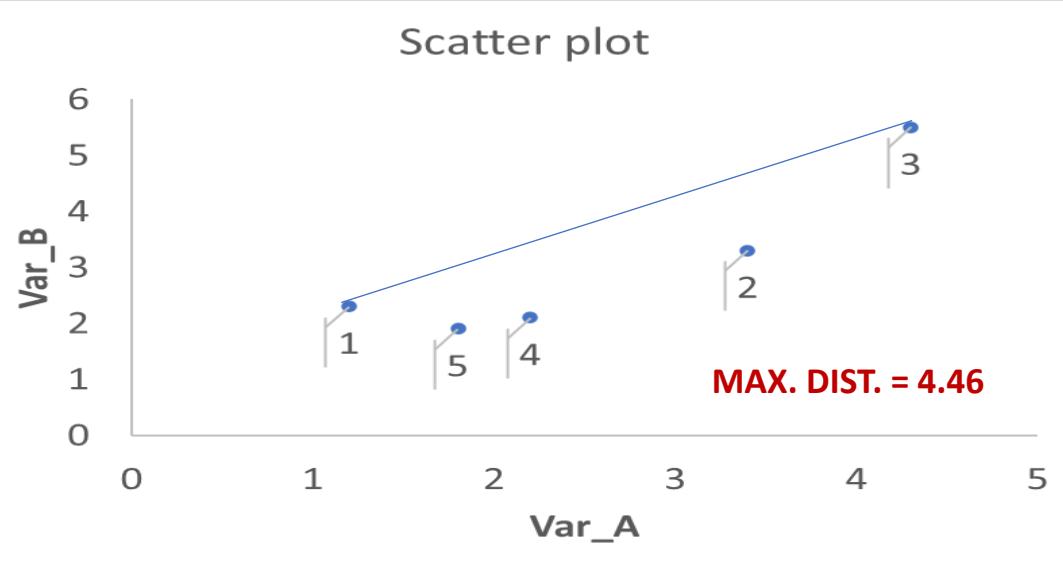
0.00	2.42	4.46	1.02	0.72
2.42	0.00	2.38	1.70	2.13
4.46	2.38	0.00	4.00	4.38
1.02	1.70	4.00	0.00	0.45
0.72	2.13	4.38	0.45	0.00

SIMILARITY MATRIX

(EUCLIDEAN DISTANCE HERE)

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



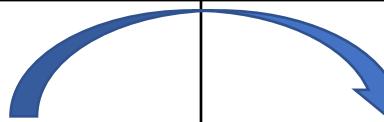
DENDROGRAM



DISTANCE MATRIX

0.00	2.42	4.46	1.02	0.72
2.42	0.00	2.38	1.70	2.13
4.46	2.38	0.00	4.00	4.38
1.02	1.70	4.00	0.00	0.45
0.72	2.13	4.38	0.45	0.00

SIMILARITY MATRIX



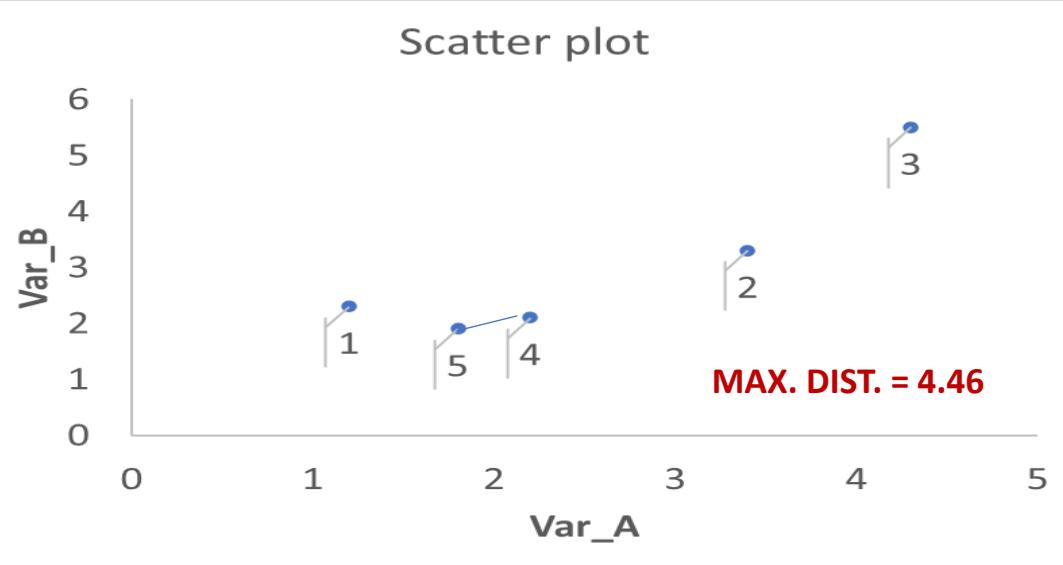
MAX. DIST. = 4.46

$$s_{ij} = 1 - \frac{d_{ij}}{d_{\text{MAX}}}$$

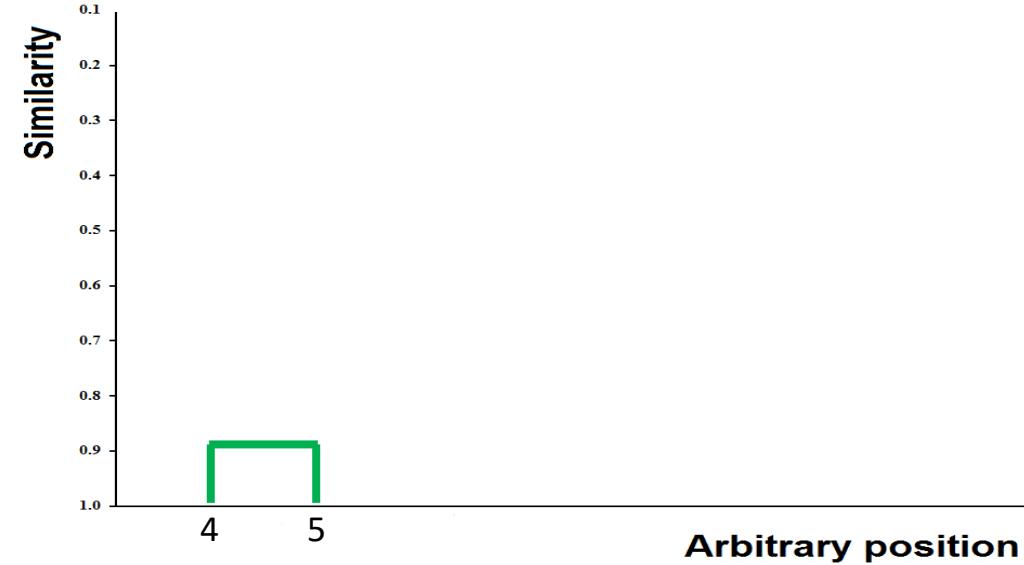
1.00	0.46	0.00	0.77	0.84
0.46	1.00	0.47	0.62	0.52
0.00	0.47	1.00	0.10	0.02
0.77	0.62	0.10	1.00	0.90
0.84	0.52	0.02	0.90	1.00

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

0.00	2.42	4.46	1.02	0.72
2.42	0.00	2.38	1.70	2.13
4.46	2.38	0.00	4.00	4.38
1.02	1.70	4.00	0.00	0.45
0.72	2.13	4.38	0.45	0.00

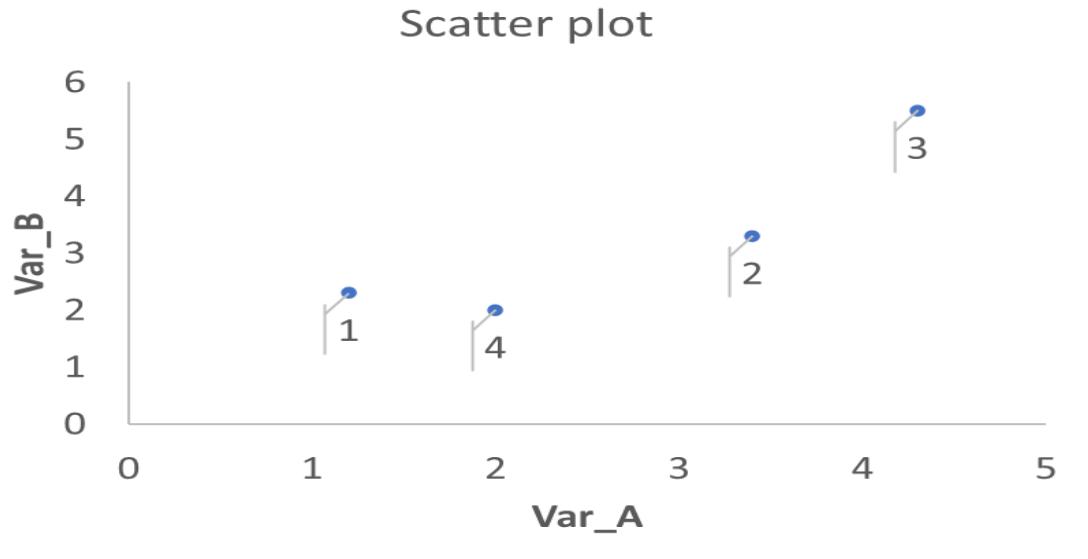
MAX. DIST. = 4.46

SIMILARITY MATRIX

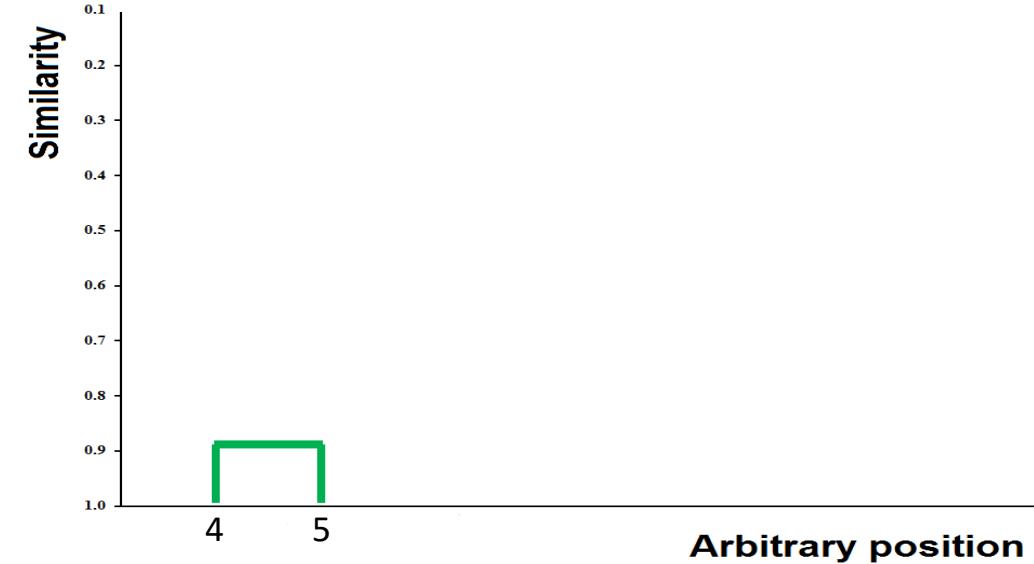
1.00	0.46	0.00	0.77	0.84
0.46	1.00	0.47	0.62	0.52
0.00	0.47	1.00	0.10	0.02
0.77	0.62	0.10	1.00	0.90
0.84	0.52	0.02	0.90	1.00

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DATASET

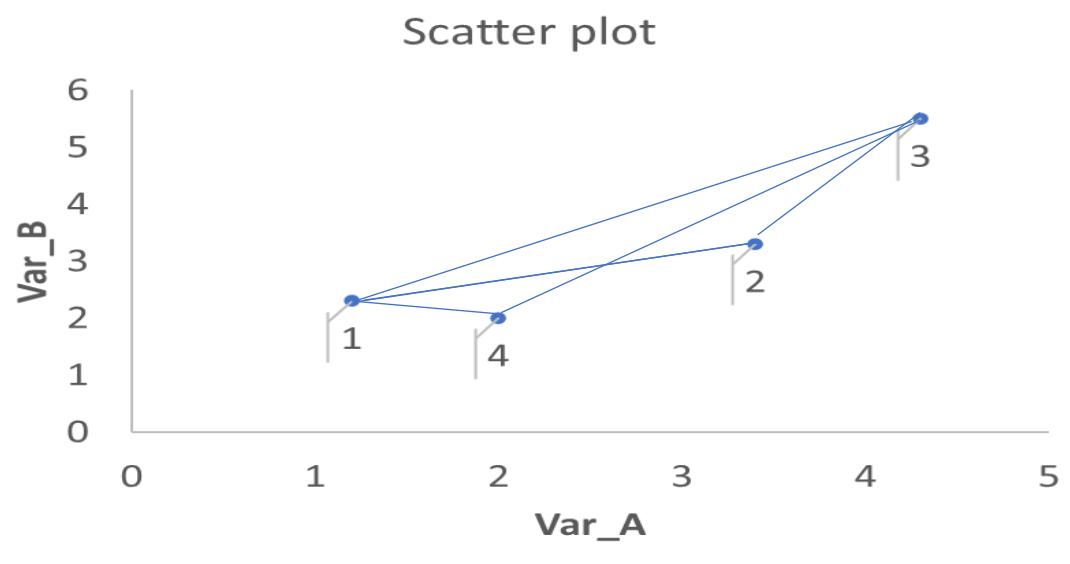
Sample	Var_A	Var_B
1	1.2	2.1
2	3.4	3.3
3	4.3	5.5
4	2.0	2.0

MAX. DIST. = 4.46

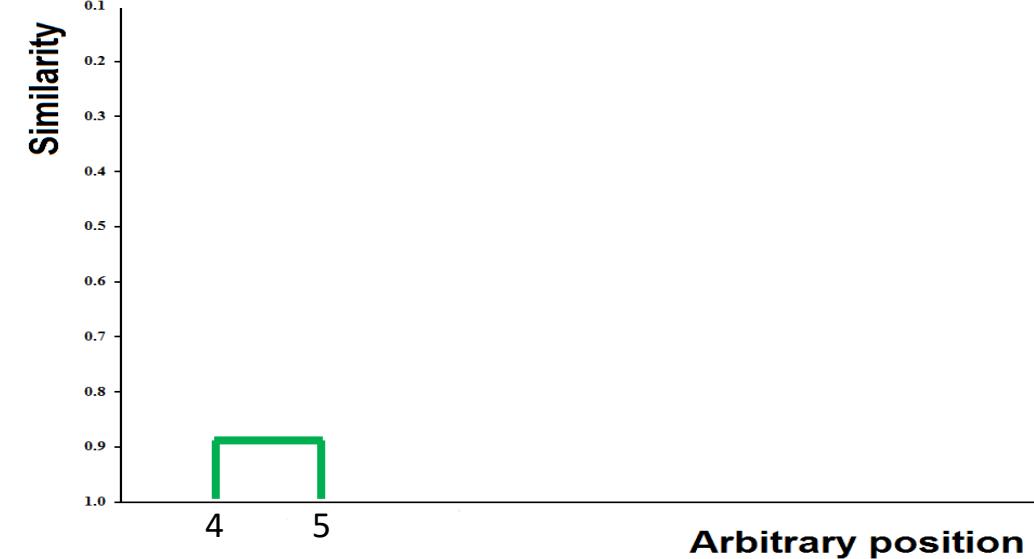
SIMILARITY MATRIX

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

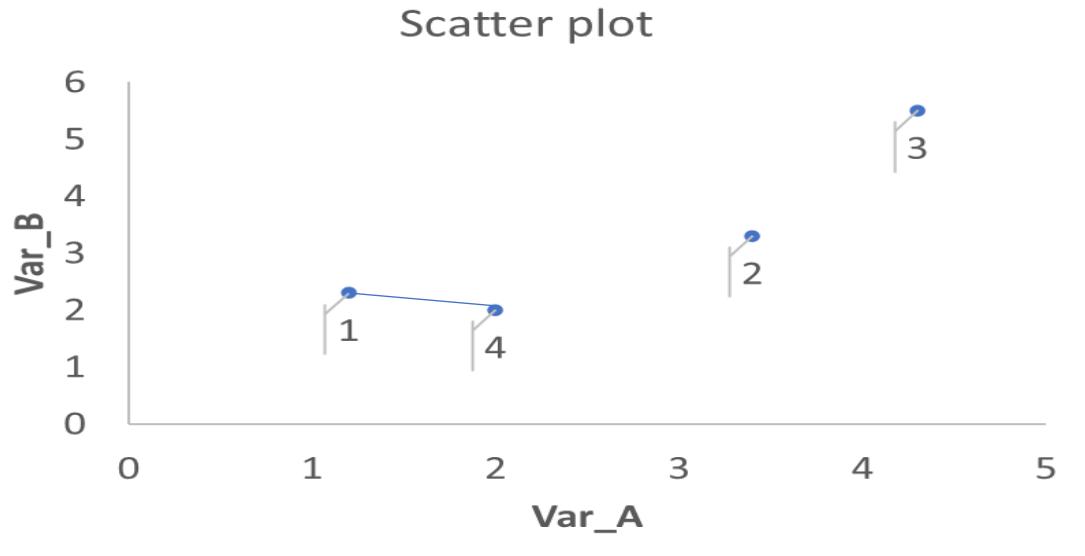
0.00	2.42	4.46	0.85
2.42	0.00	2.38	1.91
4.46	2.38	0.00	4.19
0.85	1.91	4.19	0.00

MAX. DIST. = 4.46

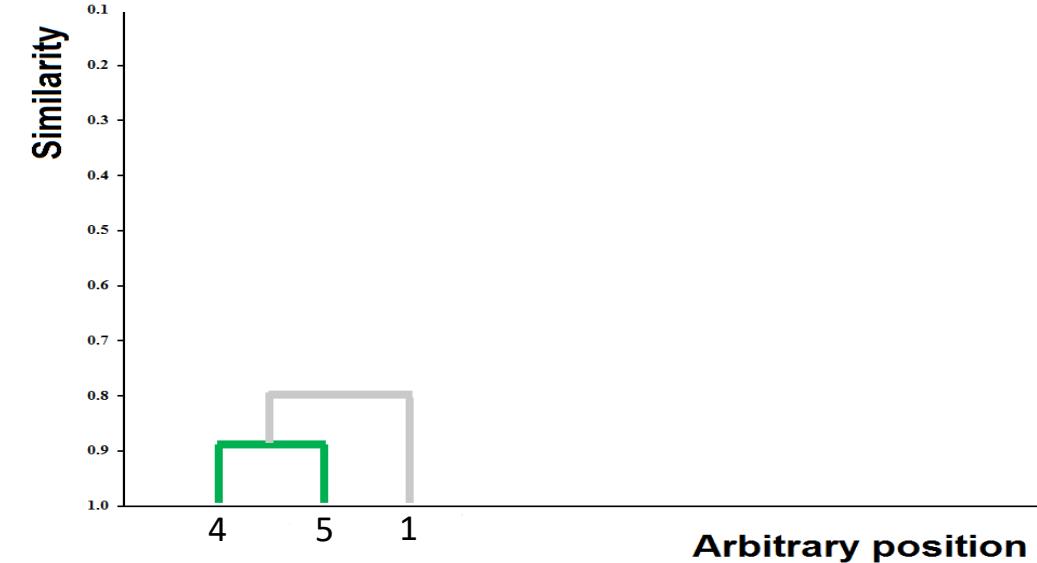
SIMILARITY MATRIX

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

0.00	2.42	4.46	0.85
2.42	0.00	2.38	1.91
4.46	2.38	0.00	4.19
0.85	1.91	4.19	0.00

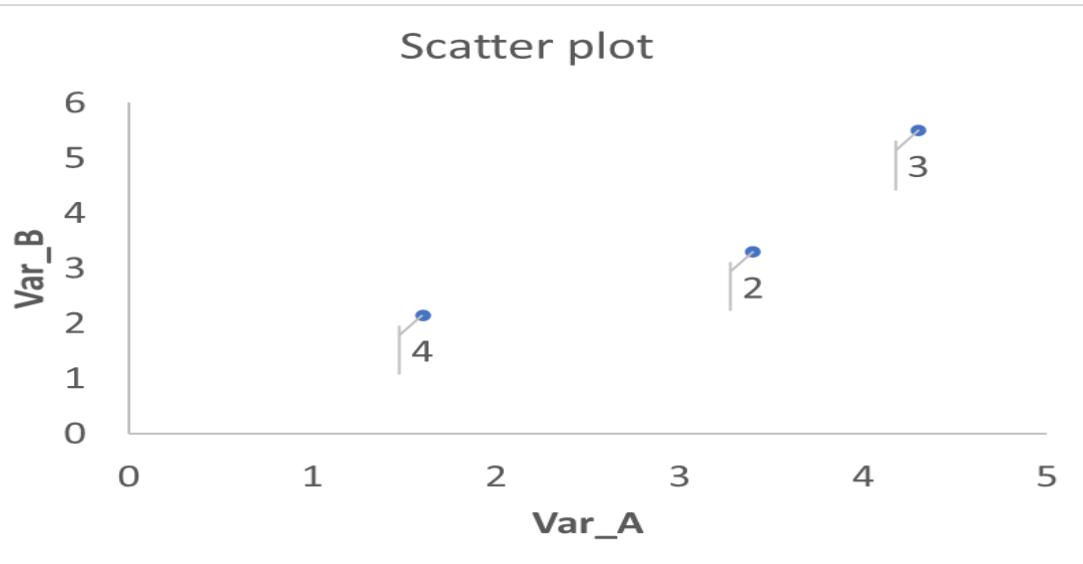
MAX. DIST. = 4.46

SIMILARITY MATRIX

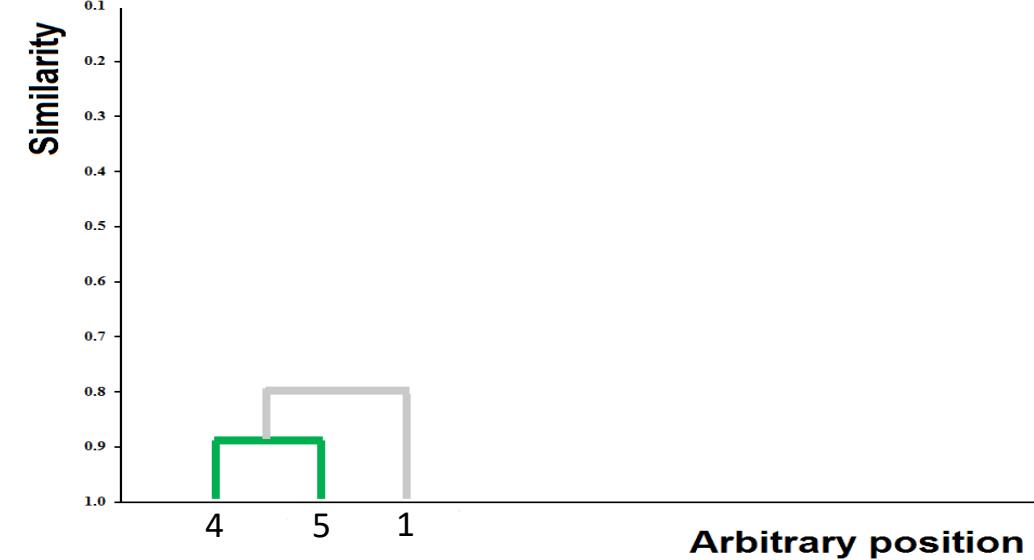
1.00	0.46	0.00	0.81
0.46	1.00	0.47	0.57
0.00	0.47	1.00	0.06
0.81	0.57	0.06	1.00

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DATASET

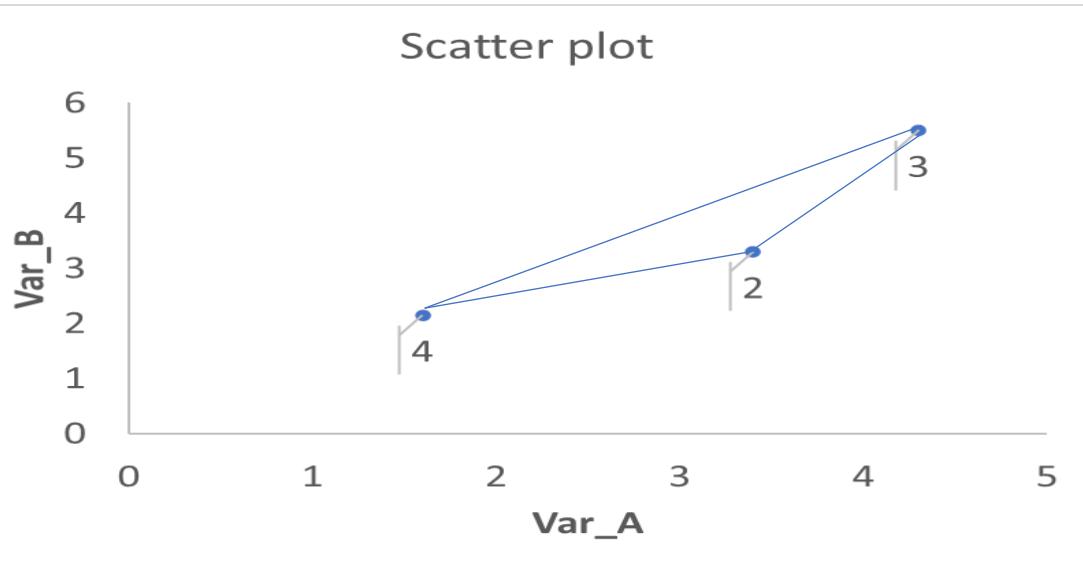
Sample	Var_A	Var_B
4	1.6	2.15
2	3.4	3.3
3	4.3	5.5

MAX. DIST. = 4.46

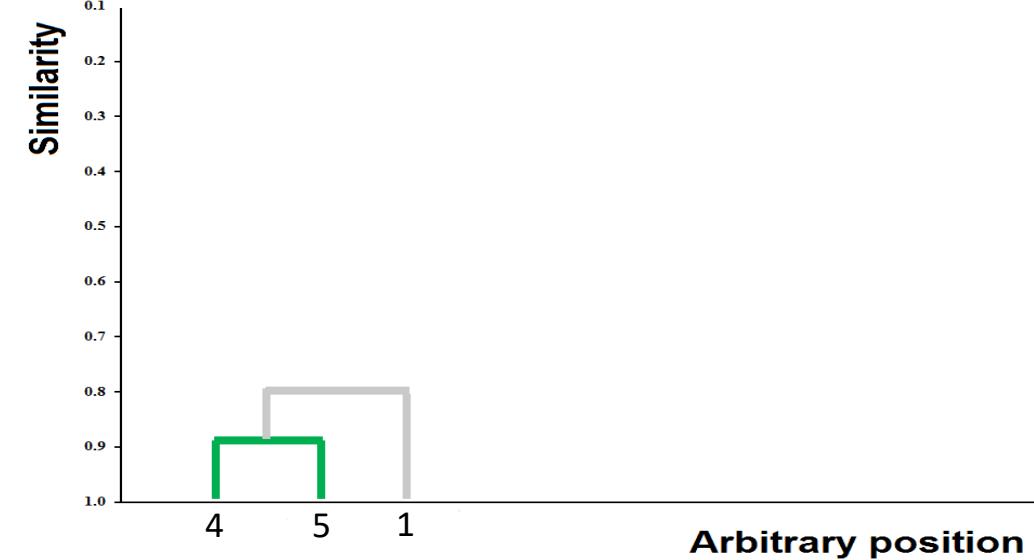
SIMILARITY MATRIX

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

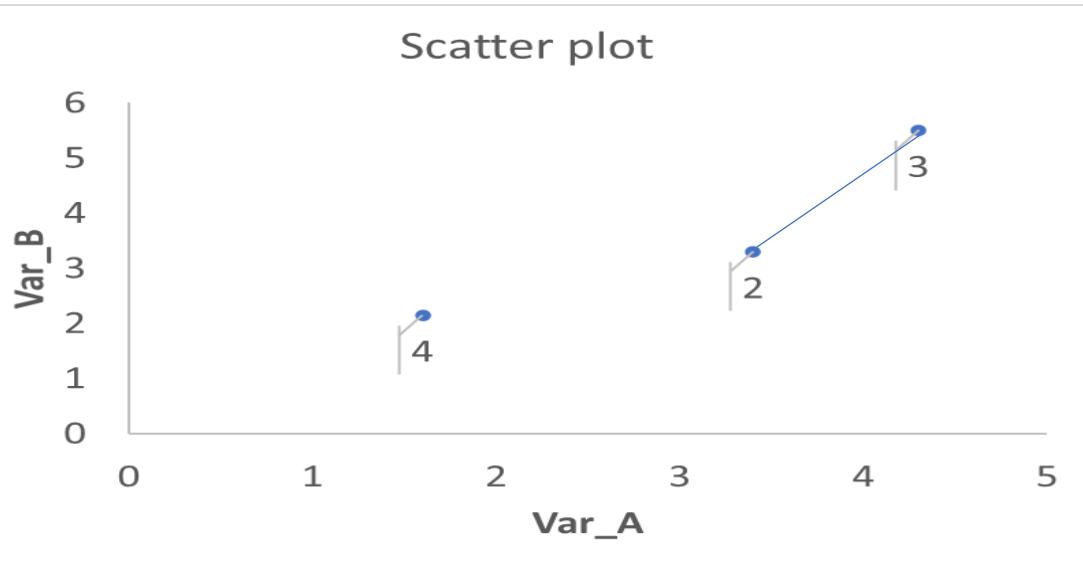
0.00	2.51	4.30
2.51	0.00	2.38
4.30	2.38	0.00

MAX. DIST. = 4.46

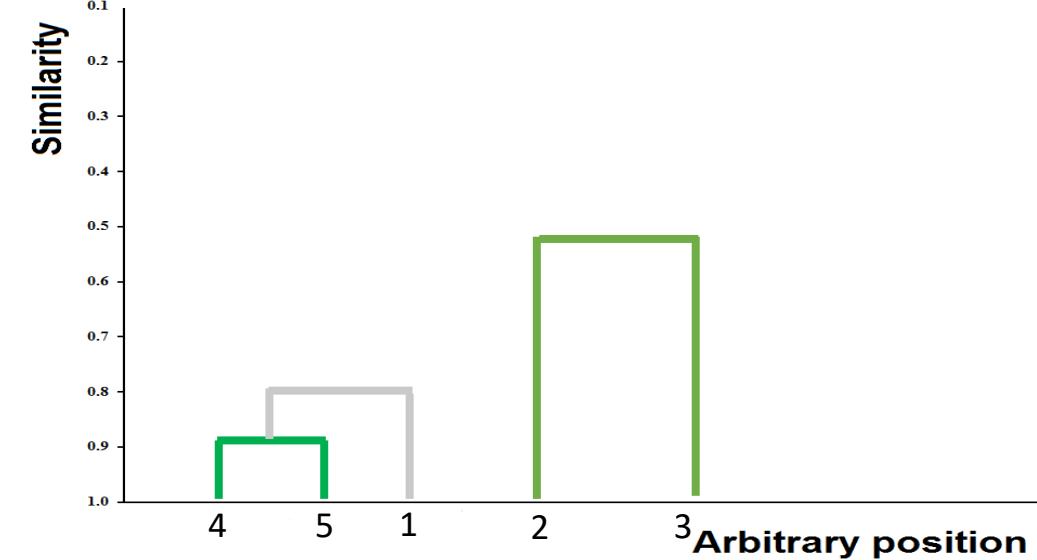
SIMILARITY MATRIX

Example: Agglomerative clustering – unweighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DISTANCE MATRIX

0.00	2.51	4.30
2.51	0.00	2.38
4.30	2.38	0.00

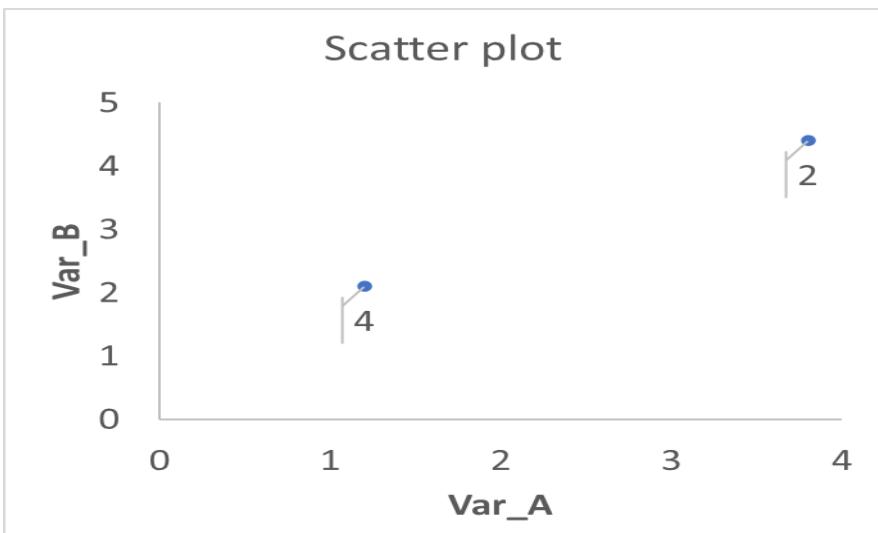
MAX. DIST. = 4.46

SIMILARITY MATRIX

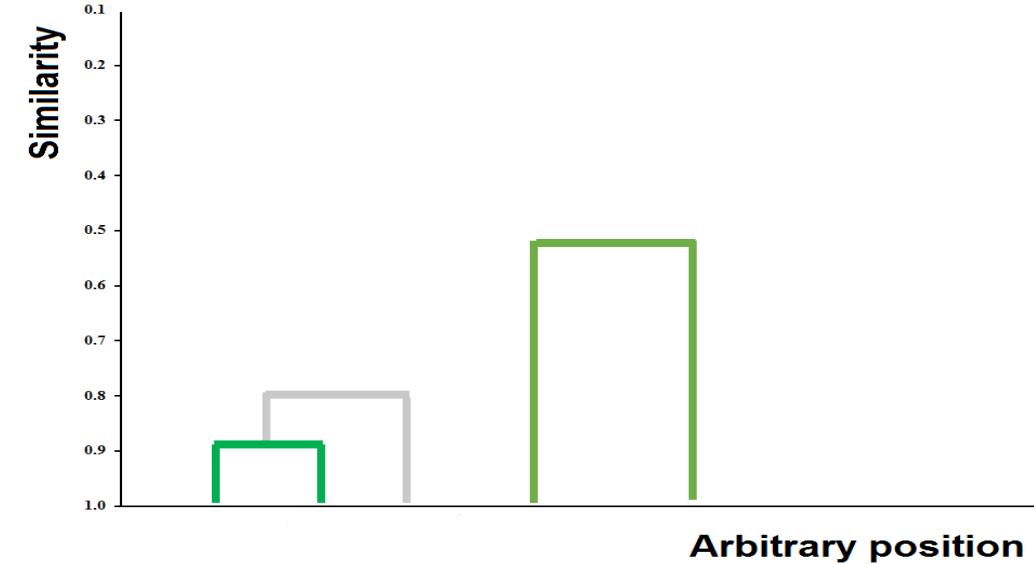
1.00	0.44	0.00
0.44	1.00	0.47
0.00	0.47	1.00

Example: Agglomerative clustering – weighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES



DENDROGRAM



DATASET



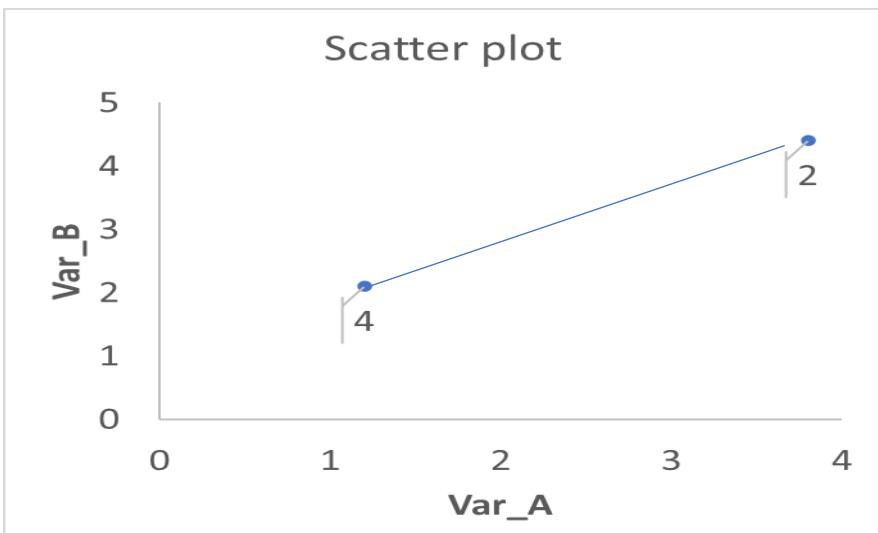
SIMILARITY MATRIX

MAX. DIST. = 4.46

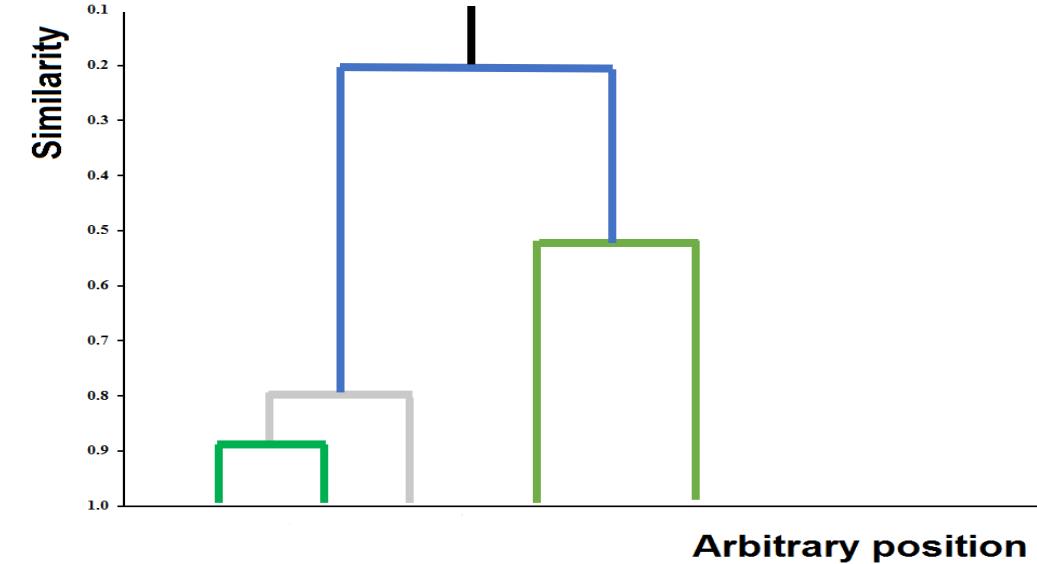
Sample	Var_A	Var_B
4	1.2	2.1
2	3.8	4.4

Example: Agglomerative clustering – weighted average linkage

DATA IN THE SPACE OF THE MEASURED ATTRIBUTES

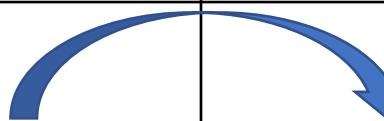


DENDROGRAM



DISTANCE MATRIX

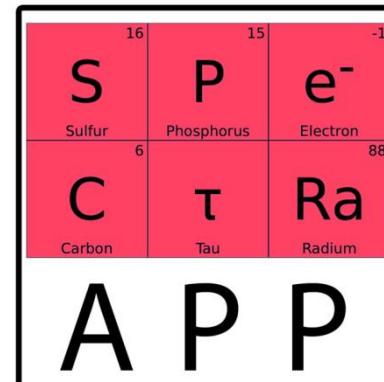
0.00	3.47
3.47	0.00



MAX. DIST. = 4.46

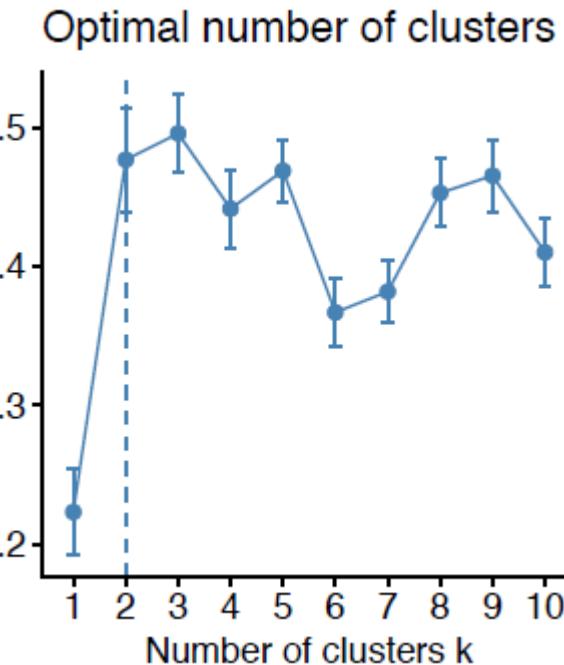
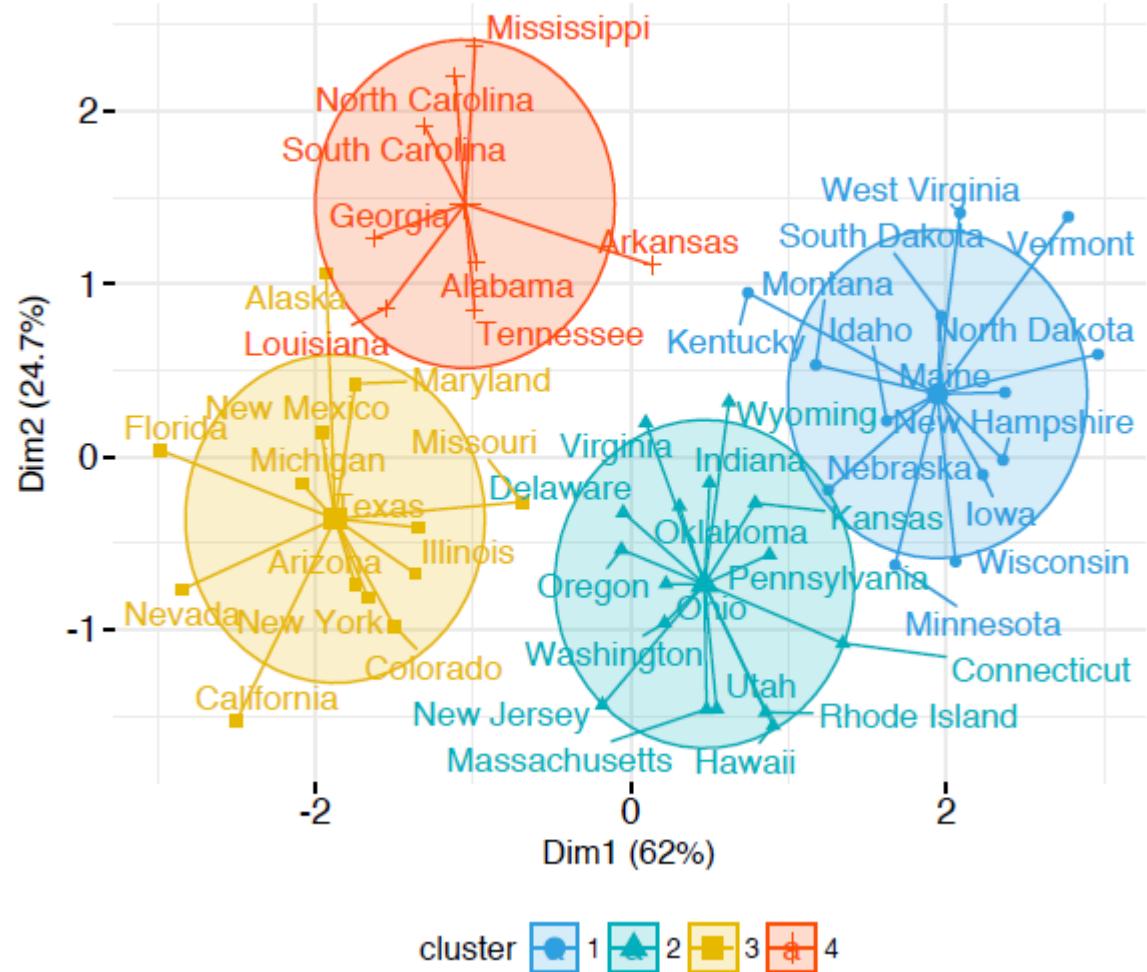
SIMILARITY MATRIX

1.00	0.22
0.22	1.00



Other techniques

Partitioning Clustering

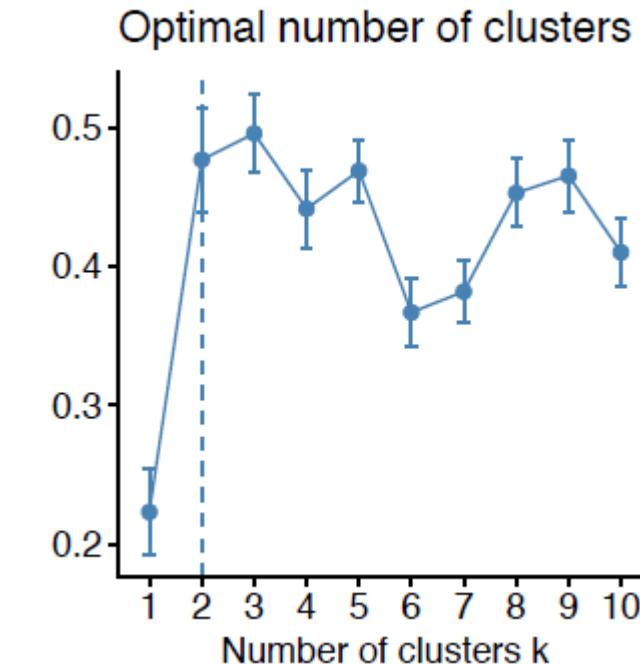
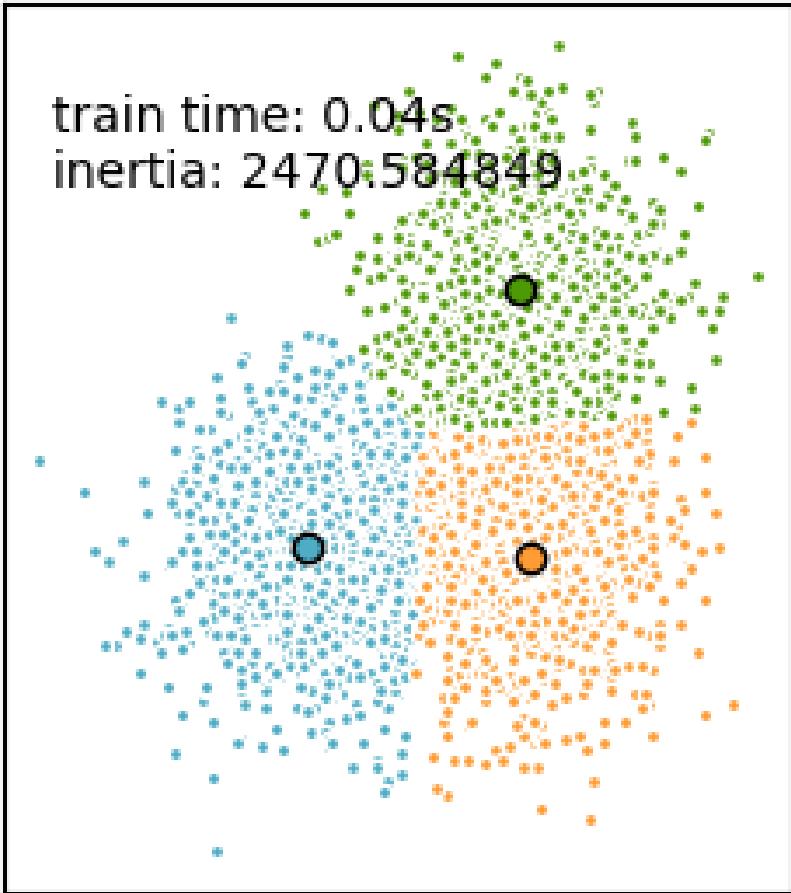


The number of N groups must be established in advance

k-means, CLARA
(Clustering Large Applications)

Other techniques

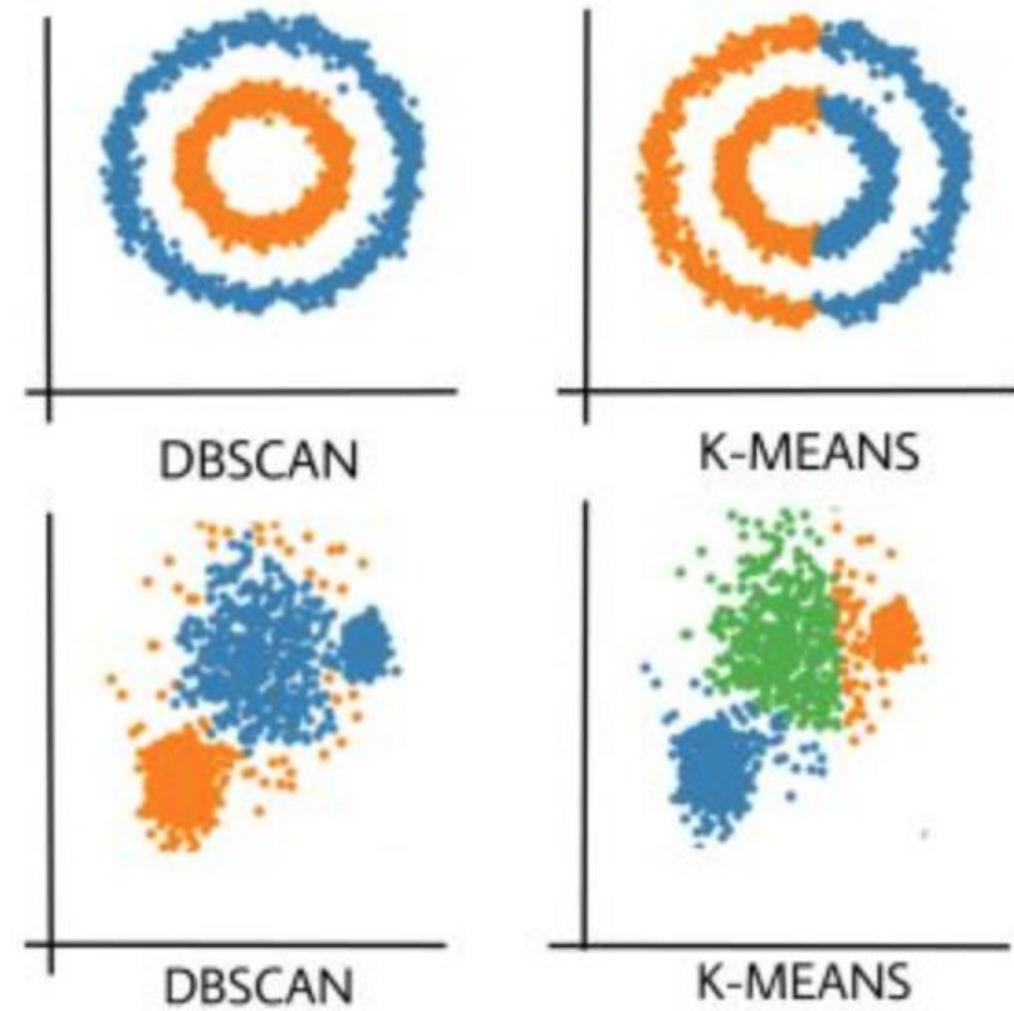
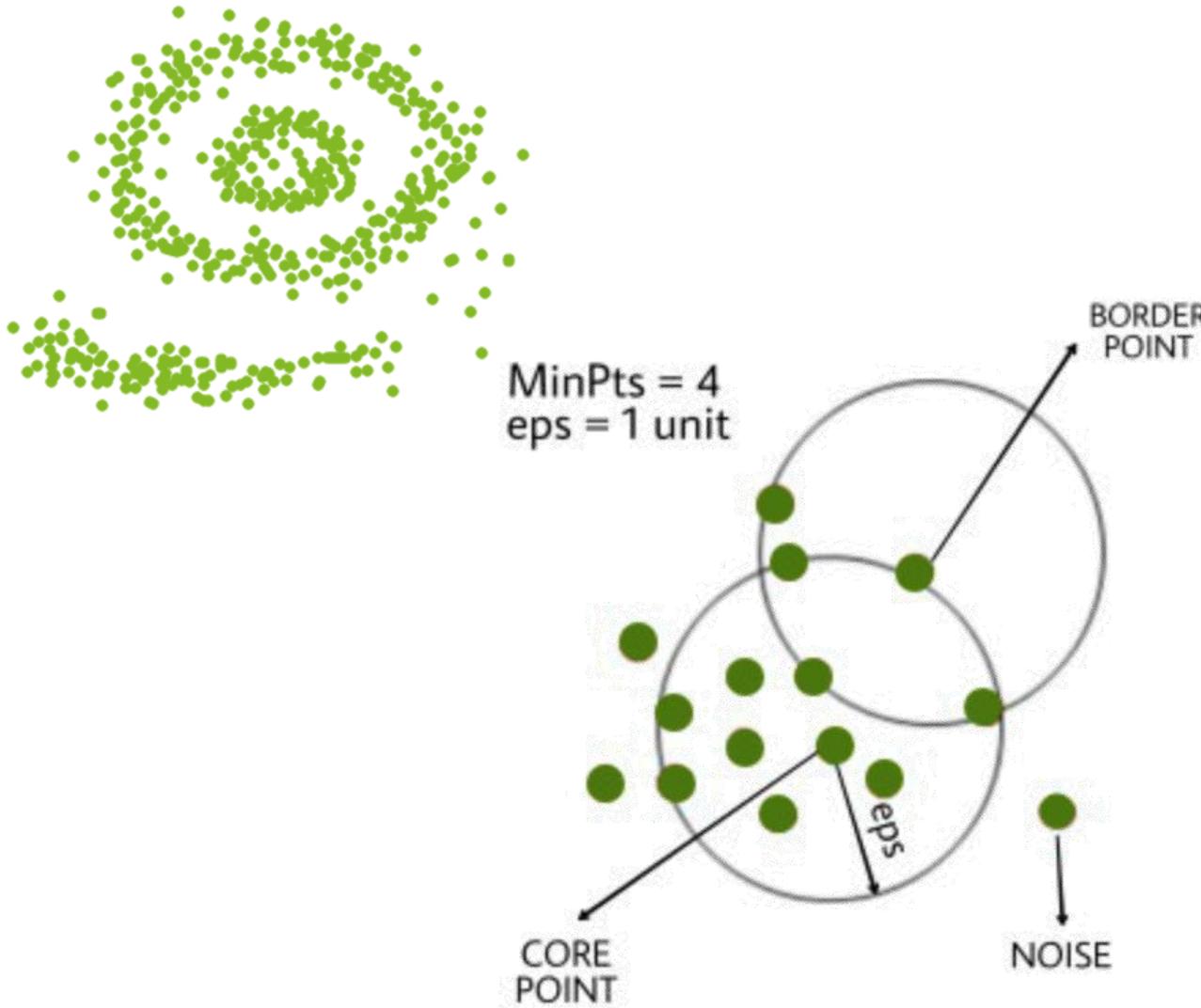
KMeans



The number of N groups must be established in advance

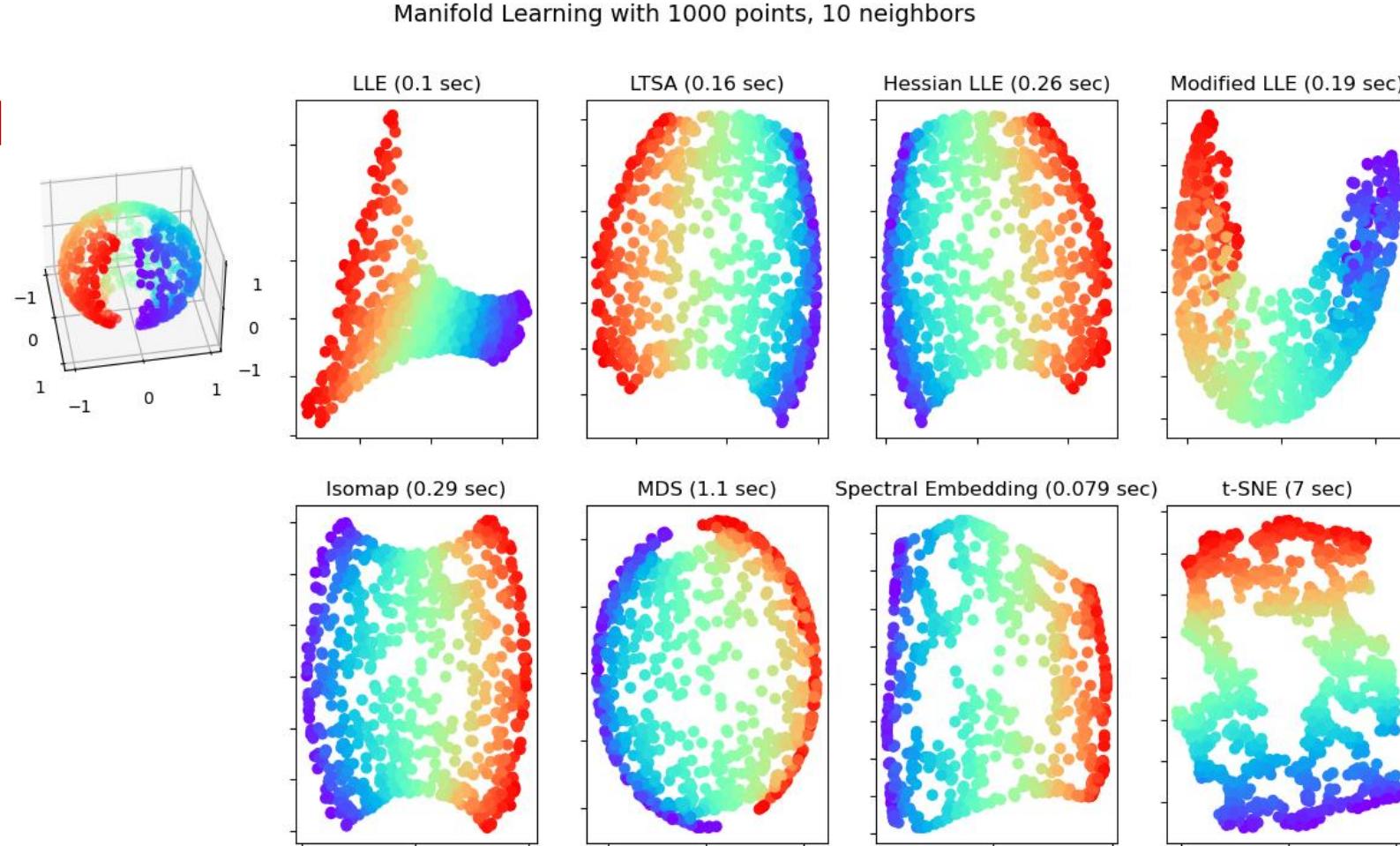
k-means, CLARA
(Clustering Large Applications)

Other techniques

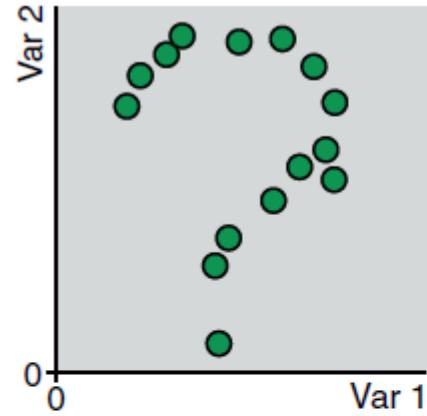


Uniform manifold approximation and projection (UMAP)

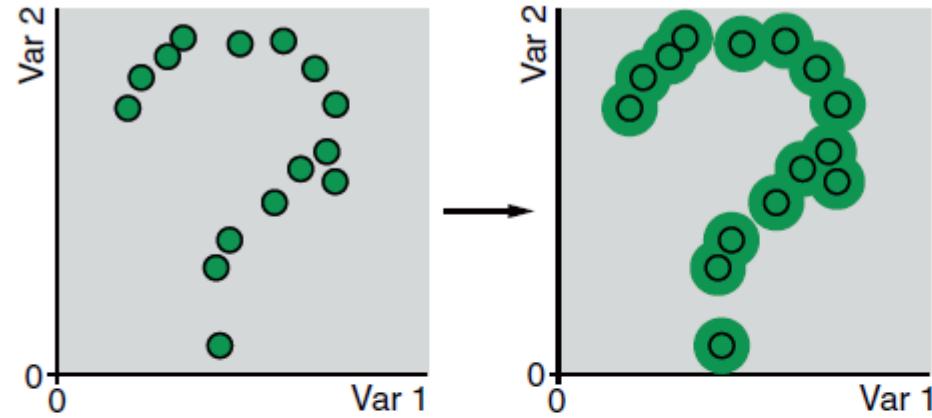
nonlinear
dimensional
reduction
algorithms



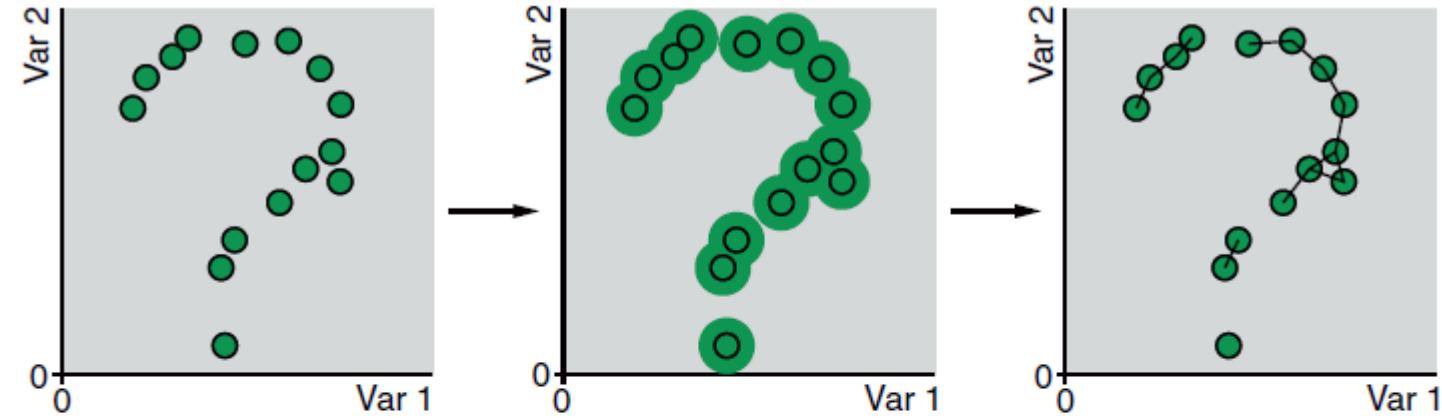
Uniform manifold approximation and projection (UMAP)



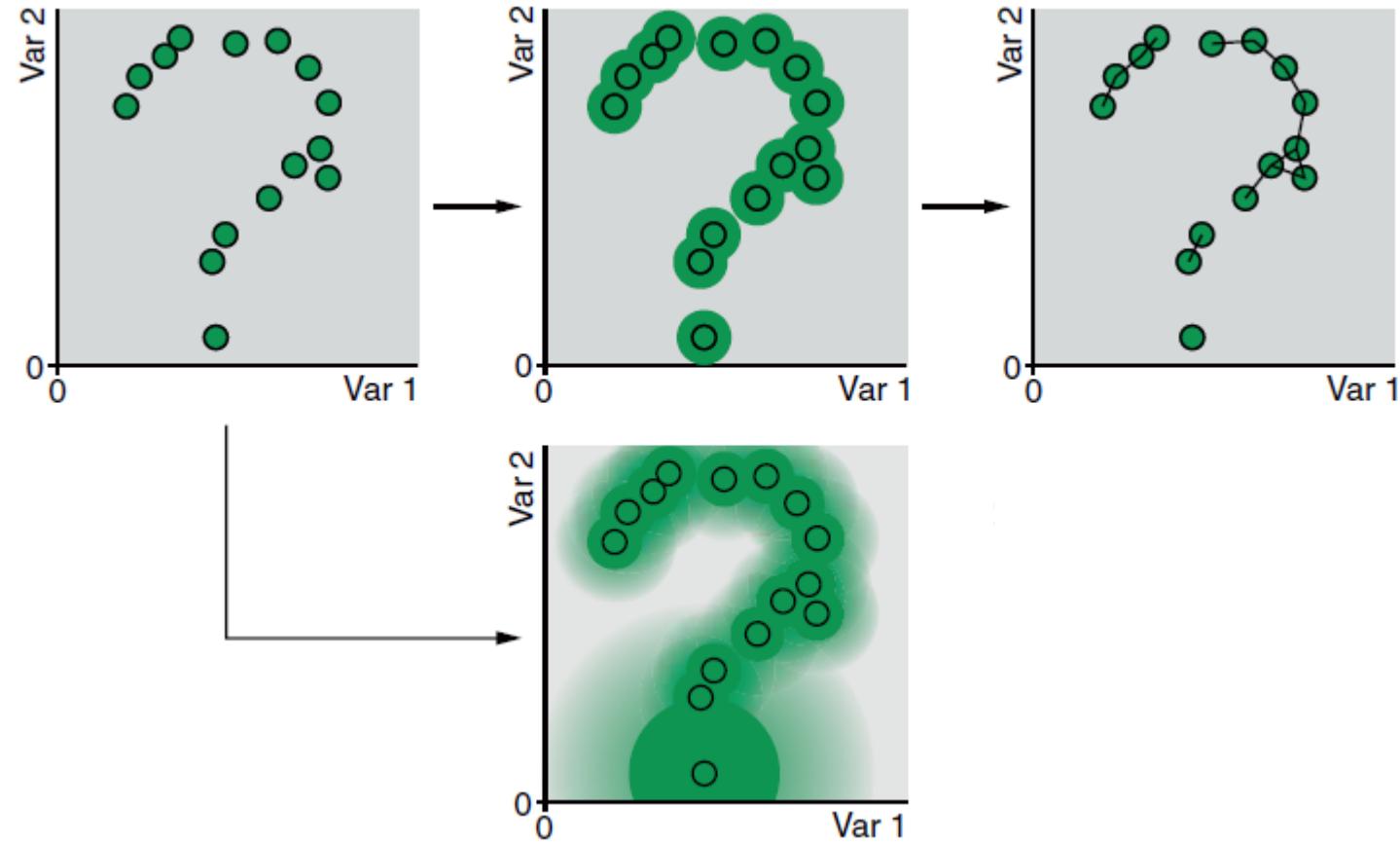
Uniform manifold approximation and projection (UMAP)



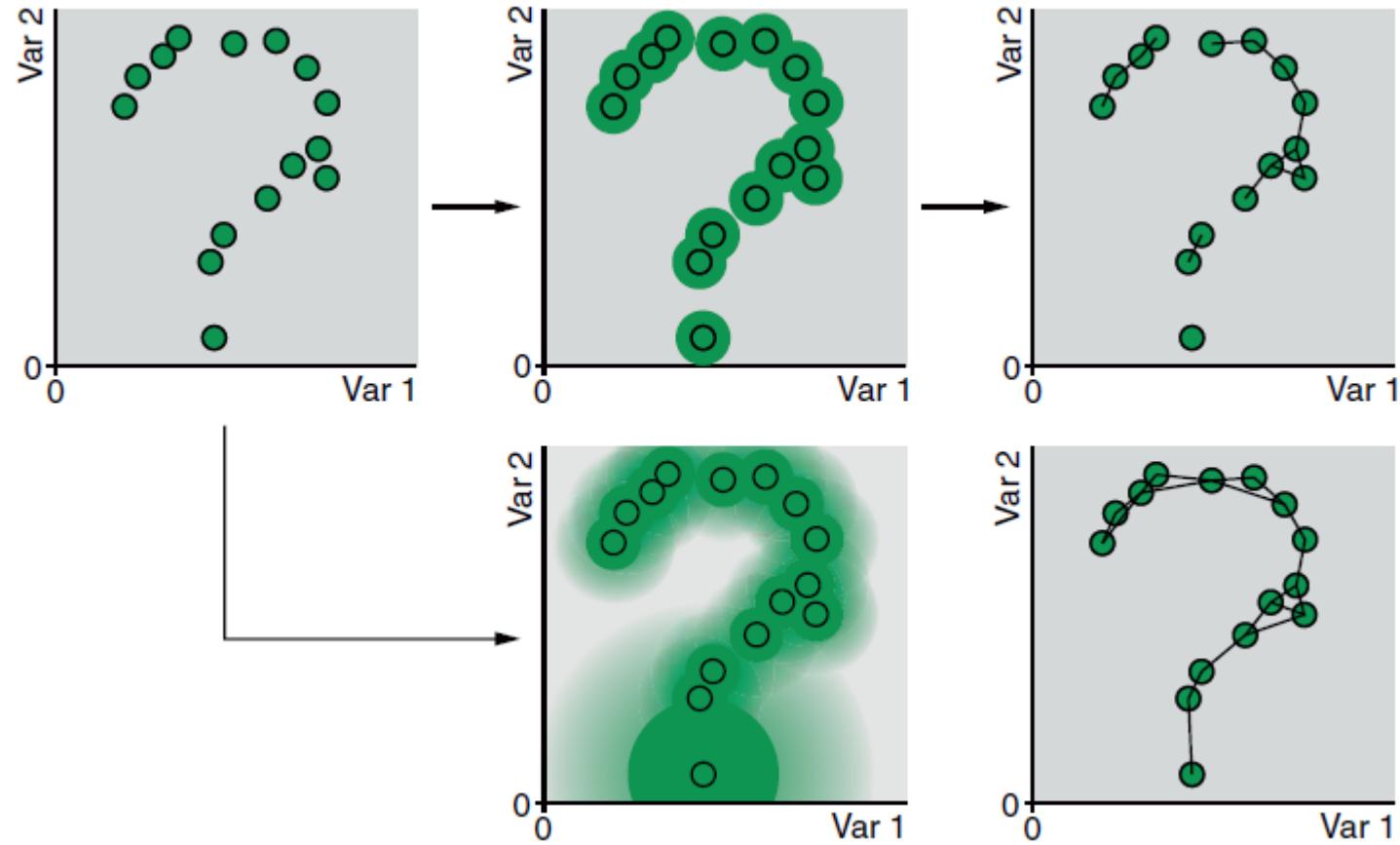
Uniform manifold approximation and projection (UMAP)



Uniform manifold approximation and projection (UMAP)



Uniform manifold approximation and projection (UMAP)

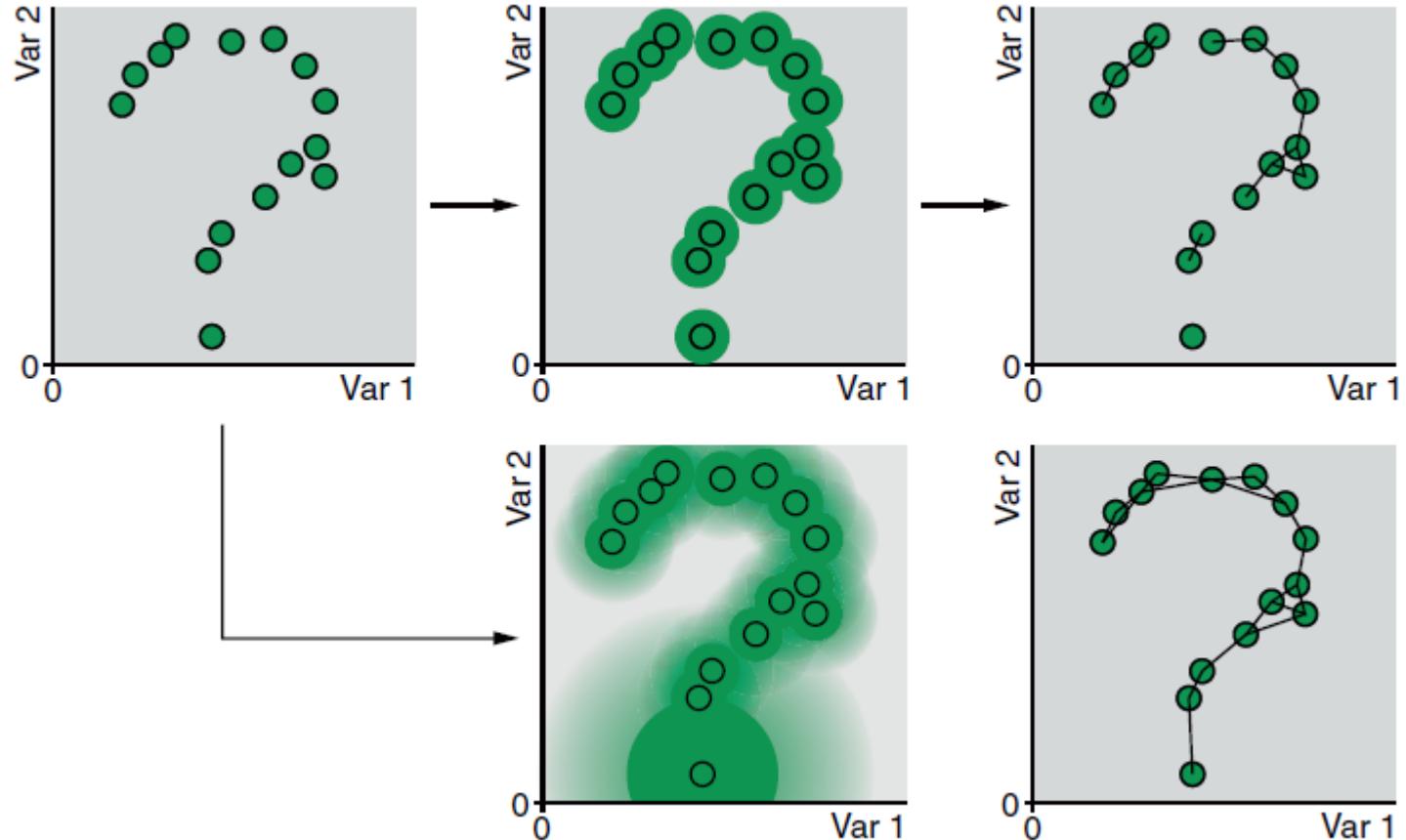


Uniform manifold approximation and projection (UMAP)

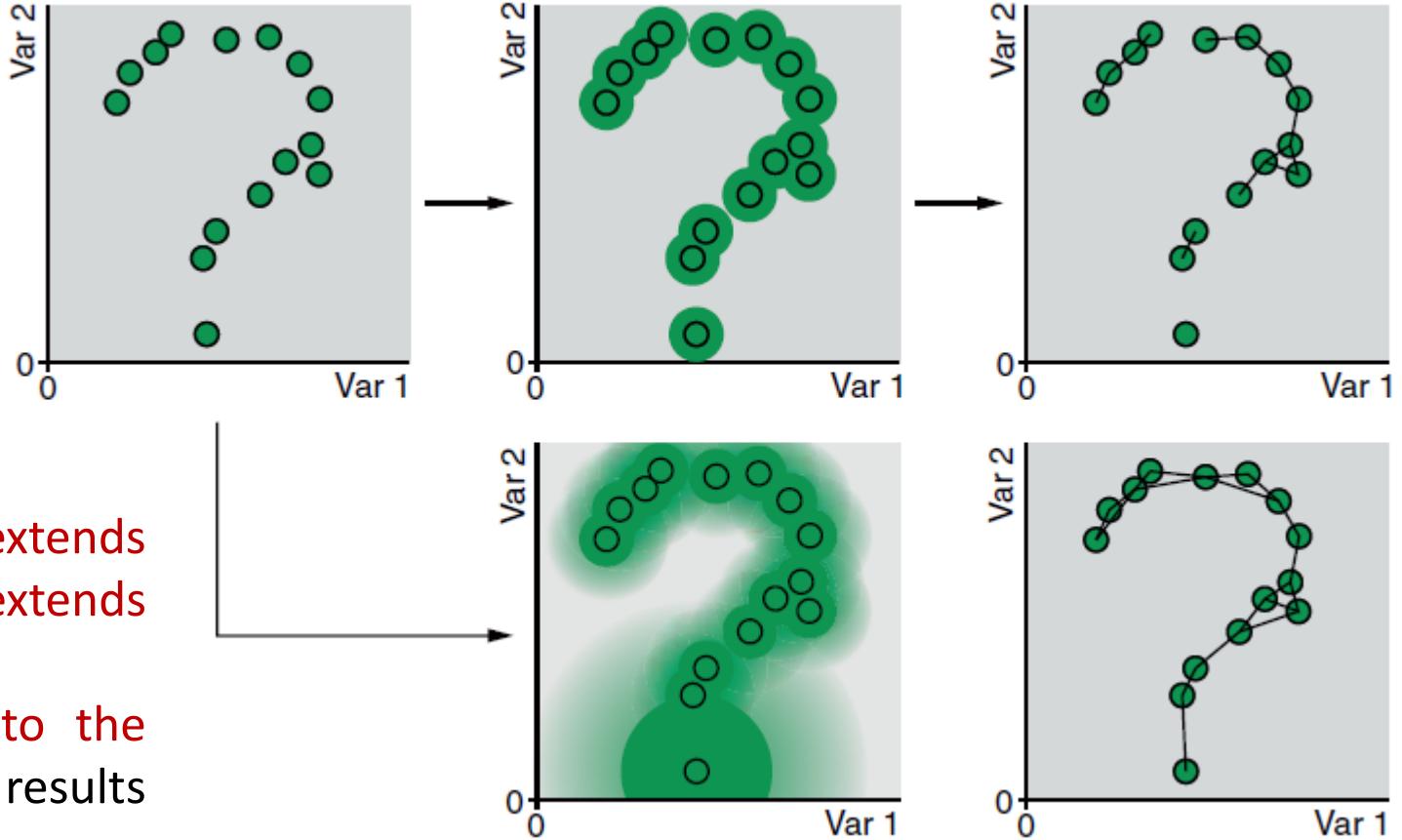
UMAP expands a search region around each case.

A naive form of this is shown in the top, where the **radius** of each search region is the same.

When cases with overlapping search regions are connected by edges, there are no gaps in the manifold.

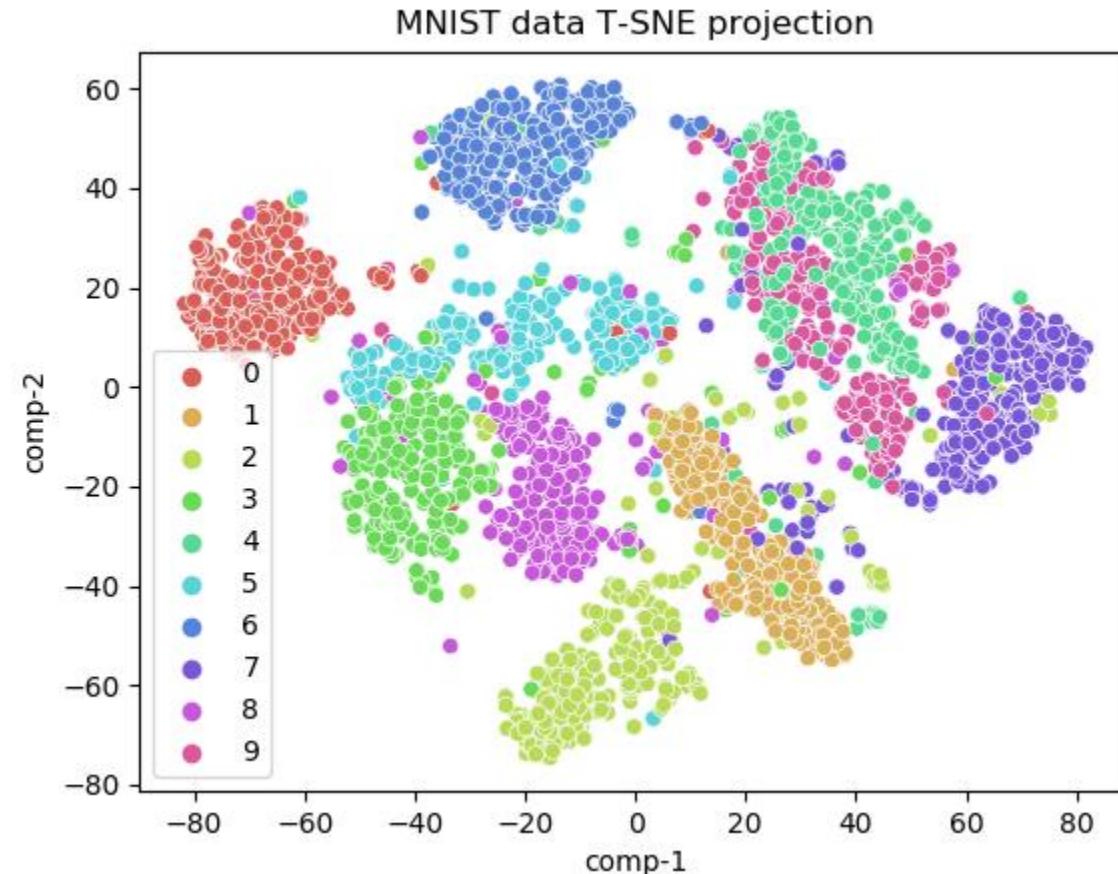


Uniform manifold approximation and projection (UMAP)



Distributed stochastic neighbor embedding (t-SNE)

nonlinear
dimensional
reduction
algorithms



t-SNE

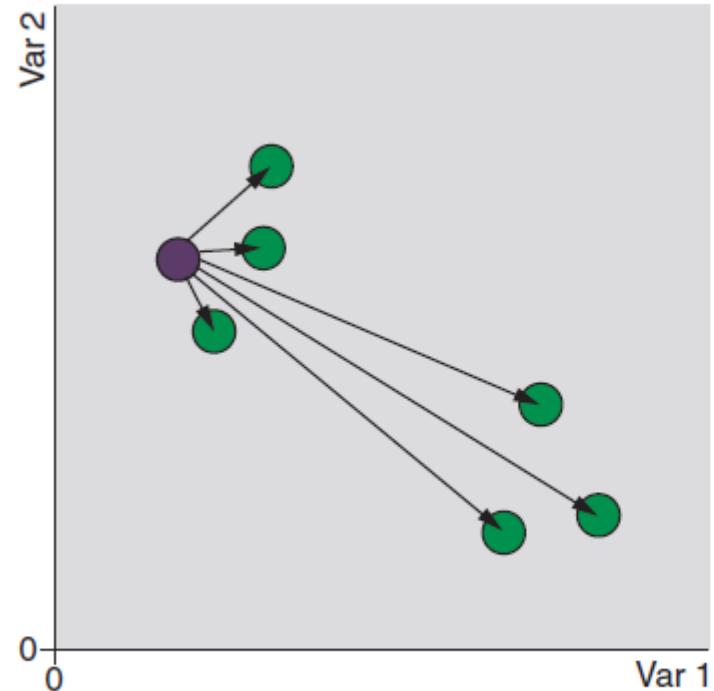
t-distributed stochastic neighbor embedding (t-SNE)

I'm glad people shorten it to t-SNE (usually pronounced "[tee-snee](#)," or occasionally "[tiz-nee](#)"), not least because when you hear someone say it, you can say "[bless you!](#)," and everyone laughs (**at least the first few times**)



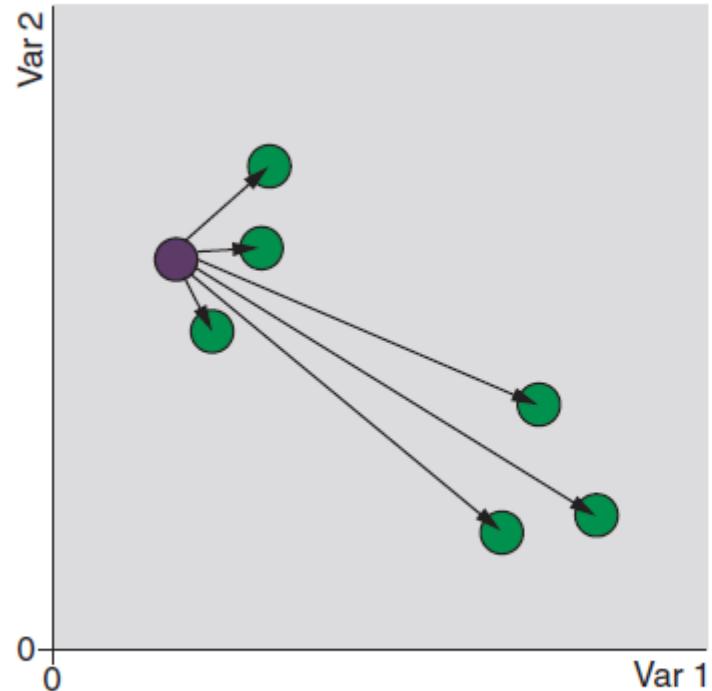
t-SNE

How does this work (simply!) - distances



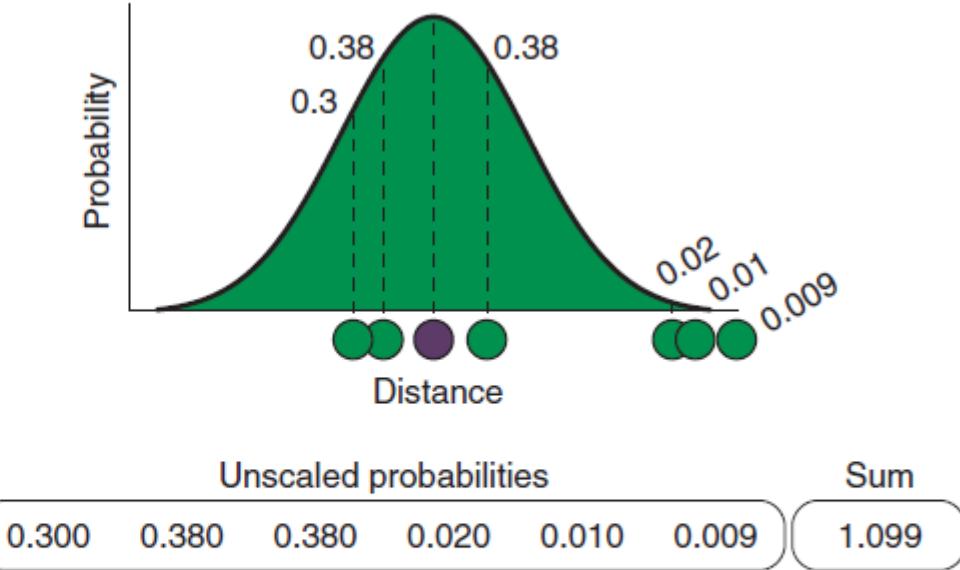
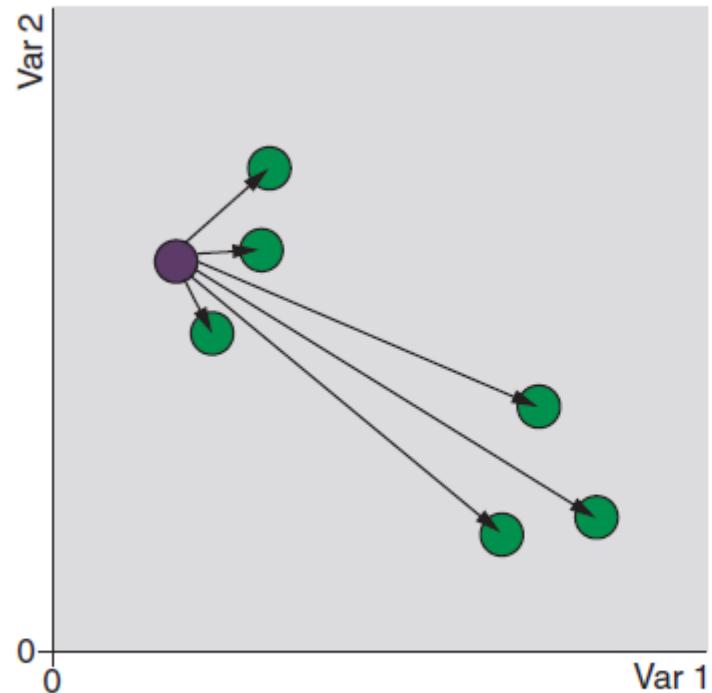
t-SNE

How does this work (simply!) – t-distribution



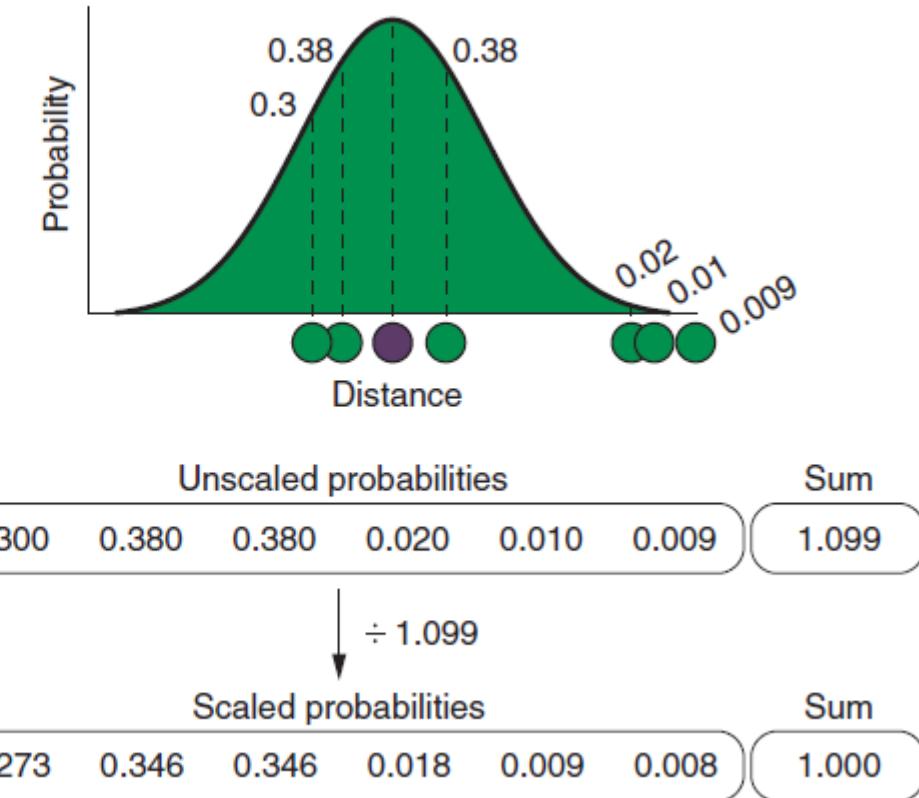
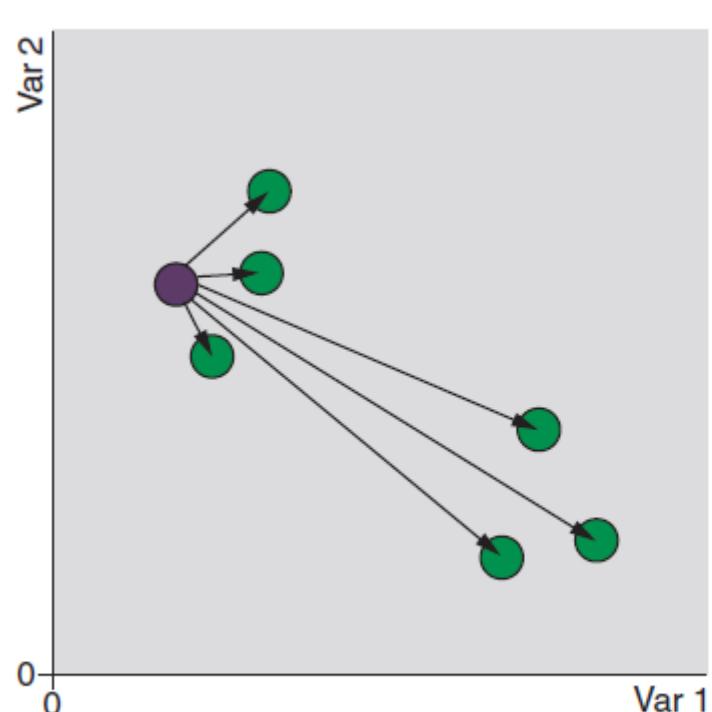
t-SNE

How does this work (simply!) – probabilities



t-SNE

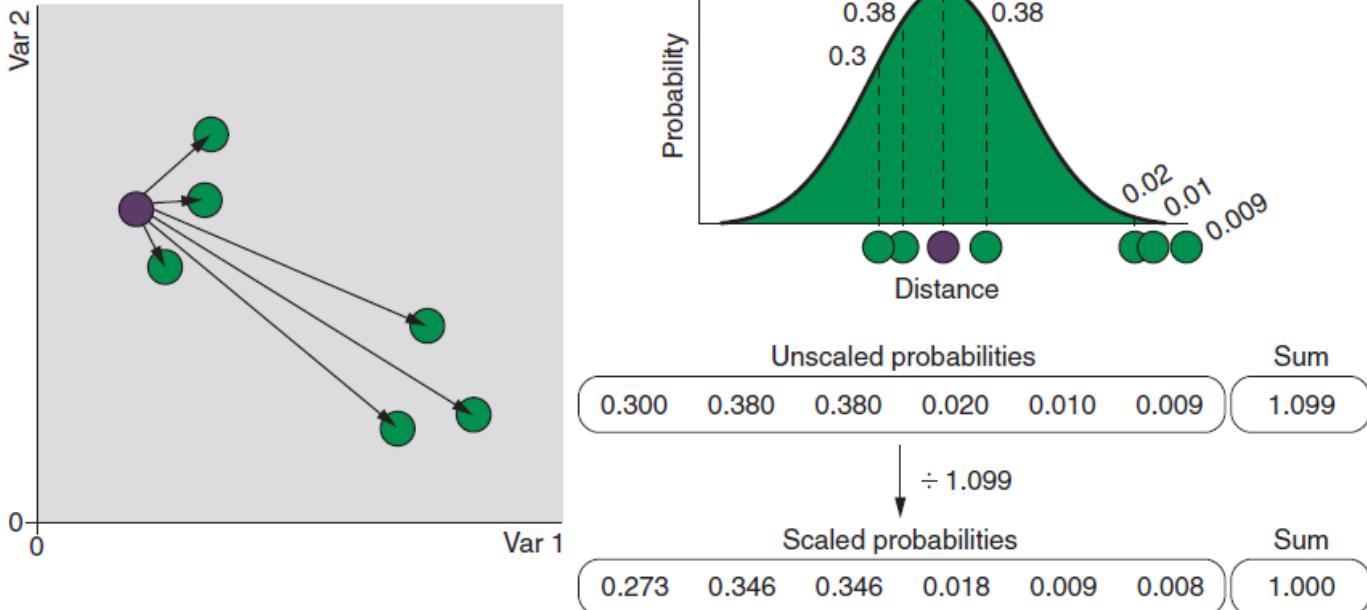
How does this work (simply!) – probabilities



t-SNE measures the **distance** from each case to every other case, **converted into a probability by fitting a normal distribution** over the current case. These probabilities are scaled by dividing them by their sum, so that they add to 1.

t-SNE

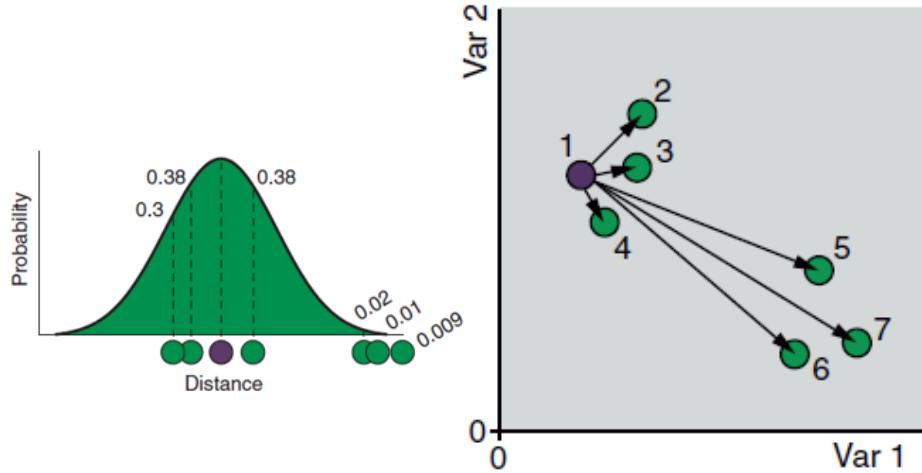
How does this work (simply!) – probabilities



It doesn't need to be two axes, but it commonly is. This is because humans struggle to visualize data in more than **two dimensions** at once, and because, **beyond three dimensions, the computational cost of t-SNE becomes more and more prohibitive**

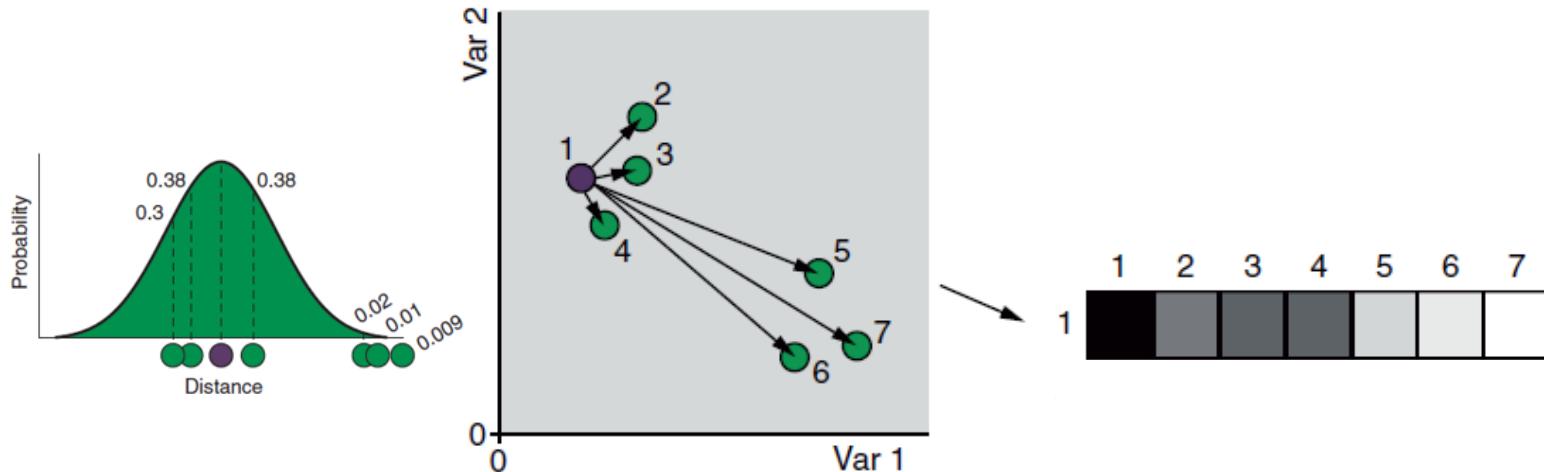
t-SNE

The correlation matrix (on scaled probabilities – distances)



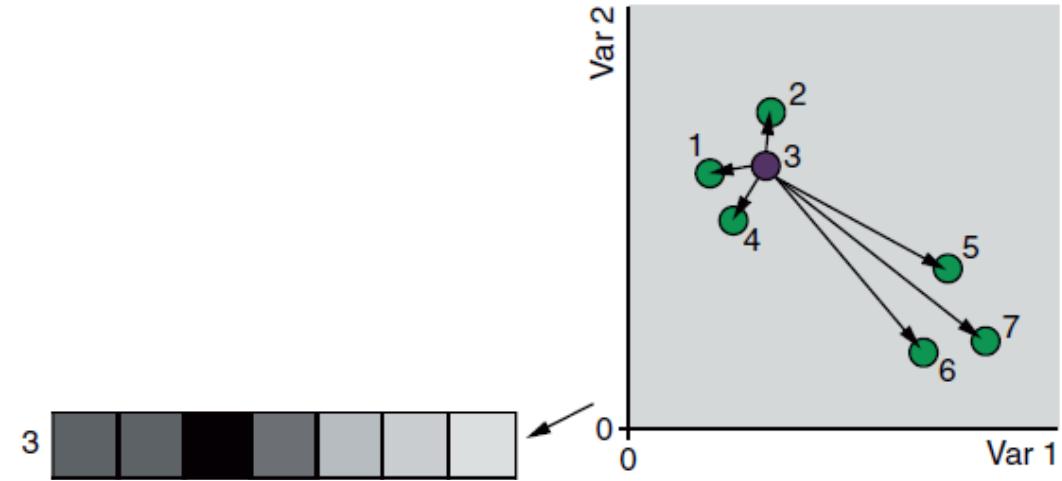
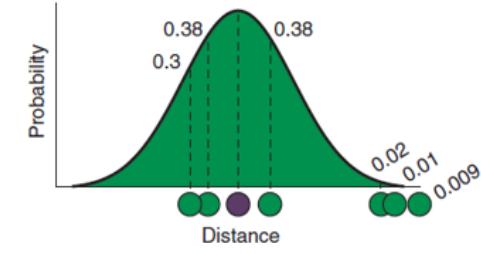
t-SNE

The correlation matrix (on scaled probabilities – distances)



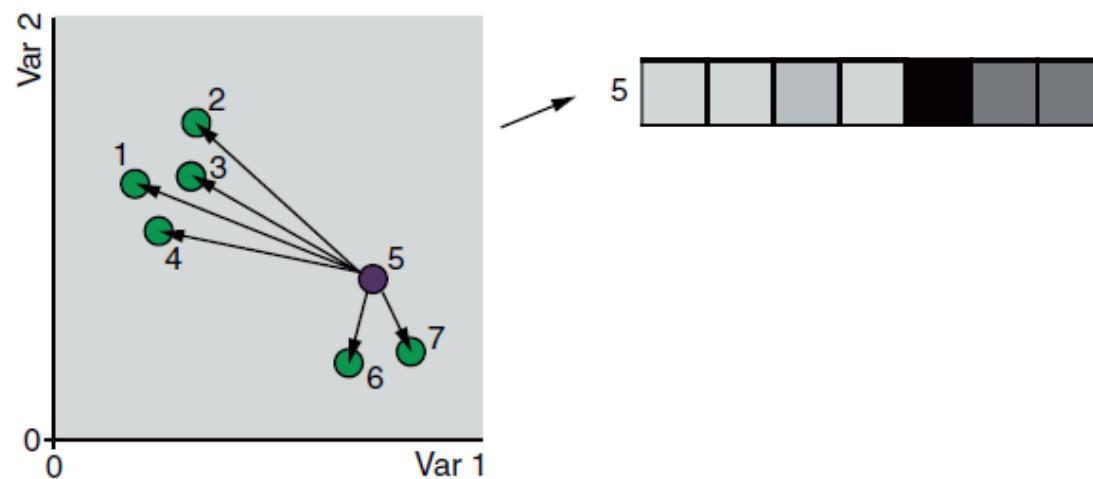
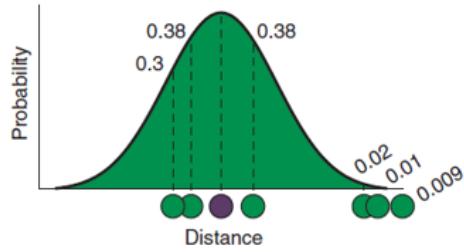
t-SNE

The correlation matrix (on scaled probabilities – distances)



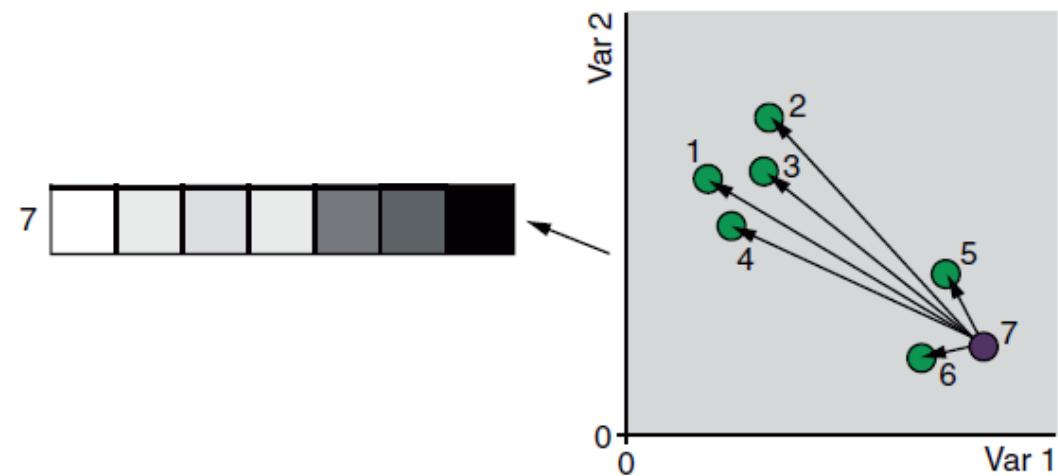
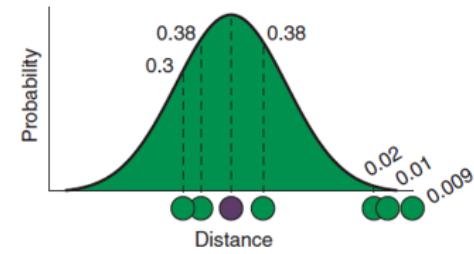
t-SNE

The correlation matrix (on scaled probabilities – distances)



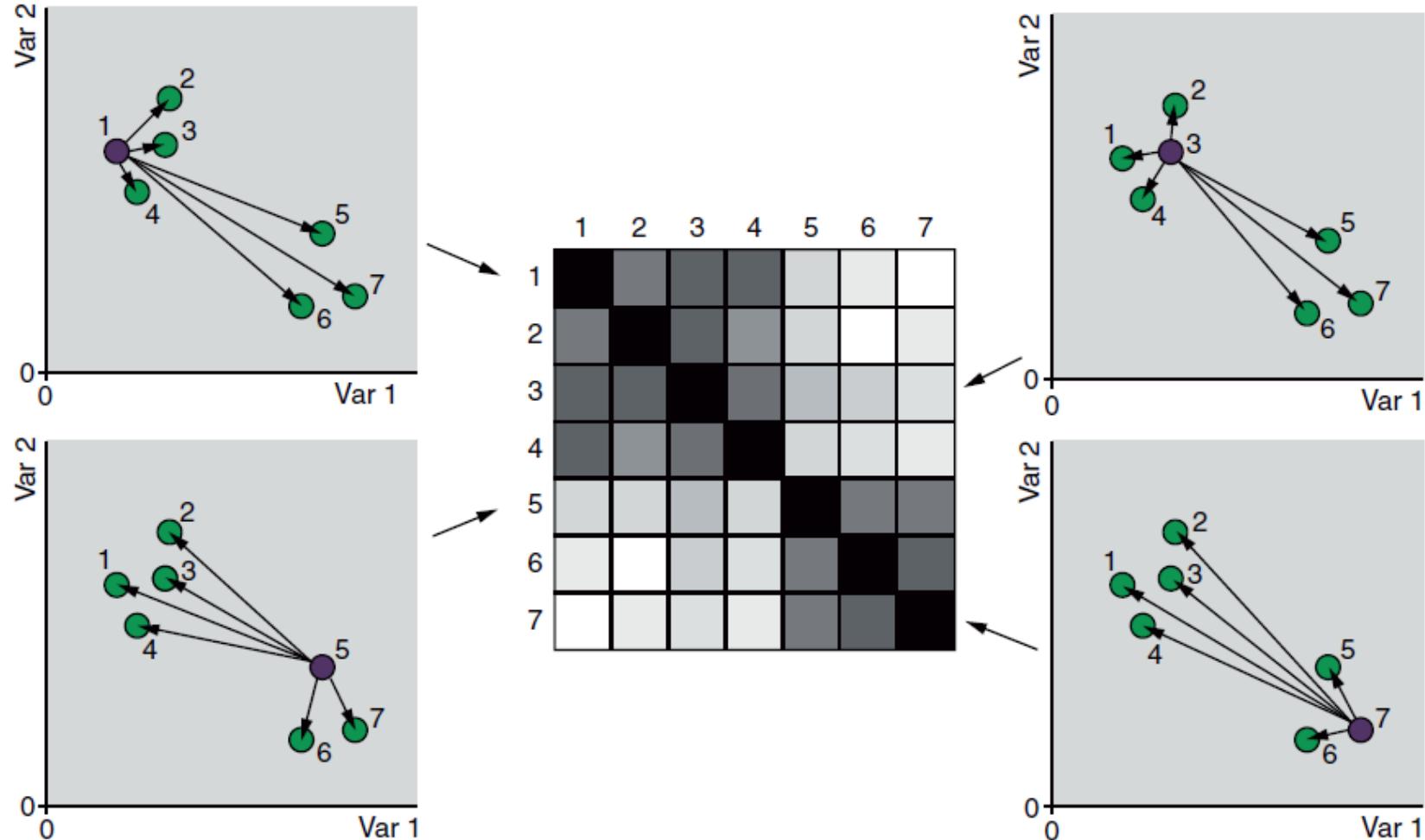
t-SNE

The correlation matrix (on scaled probabilities – distances)



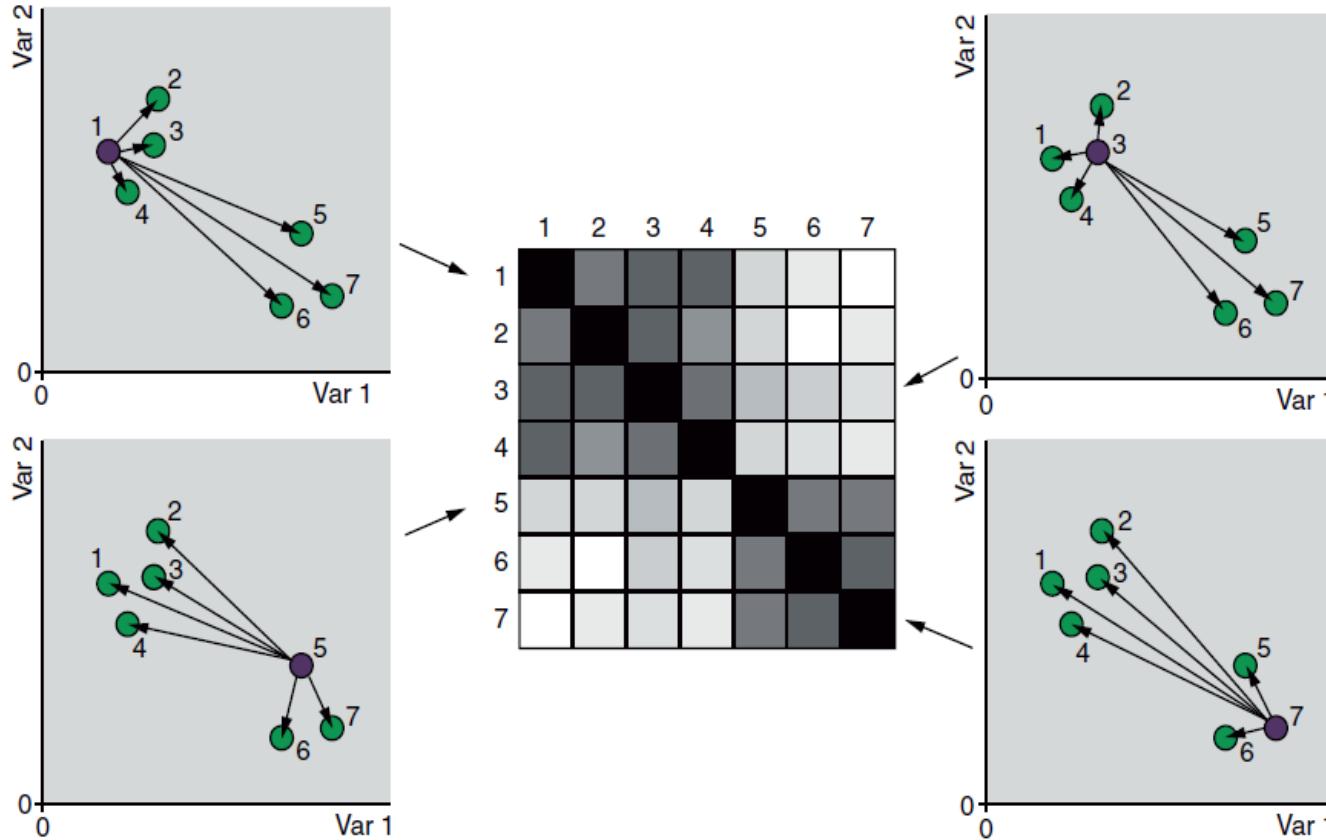
t-SNE

The correlation matrix (on scaled probabilities – distances)



t-SNE

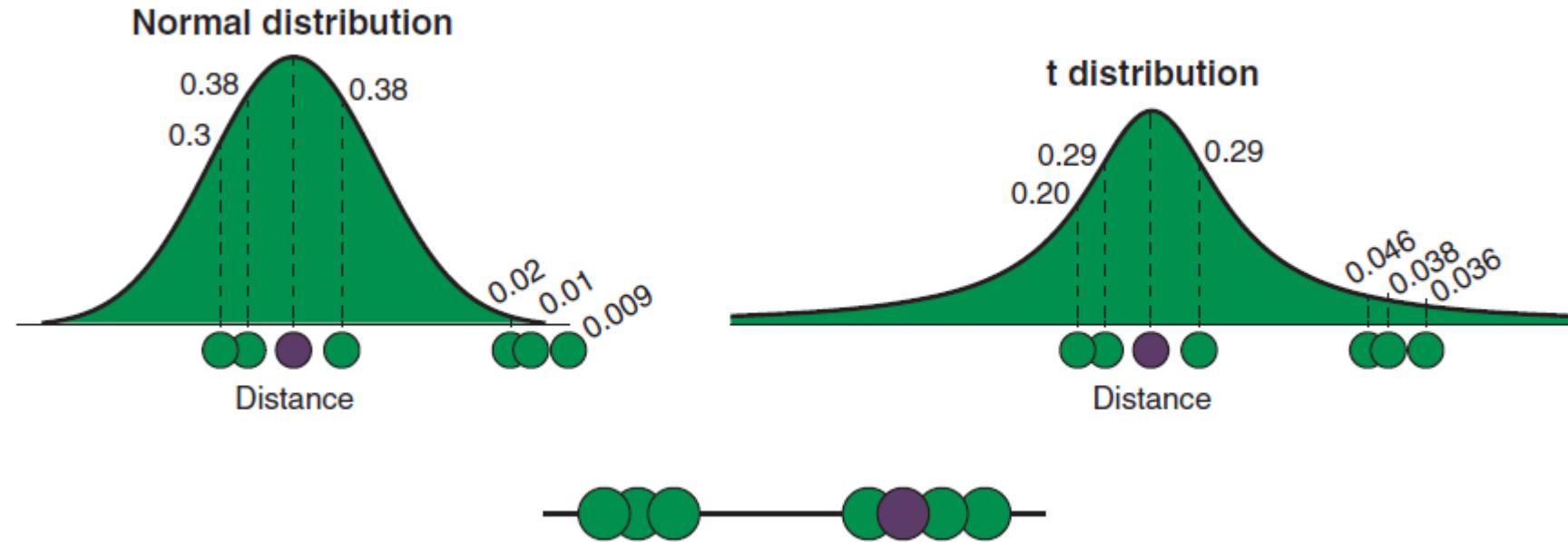
The correlation matrix (on scaled probabilities – distances)



The scaled probabilities for each case are stored as a matrix of values. This is visualized here as a **heatmap**: the closer two cases are, the darker the box is that represents their distance in the heatmap.

t-SNE

The t-distribution



When converting distances in the lower-dimensional representation into probabilities, **t-SNE fits a Student's t distribution** over the current case instead of a normal distribution. The Student's t distribution has longer tails, meaning **dissimilar cases are pushed further away to achieve the same probability** as in the high-dimensional representation.

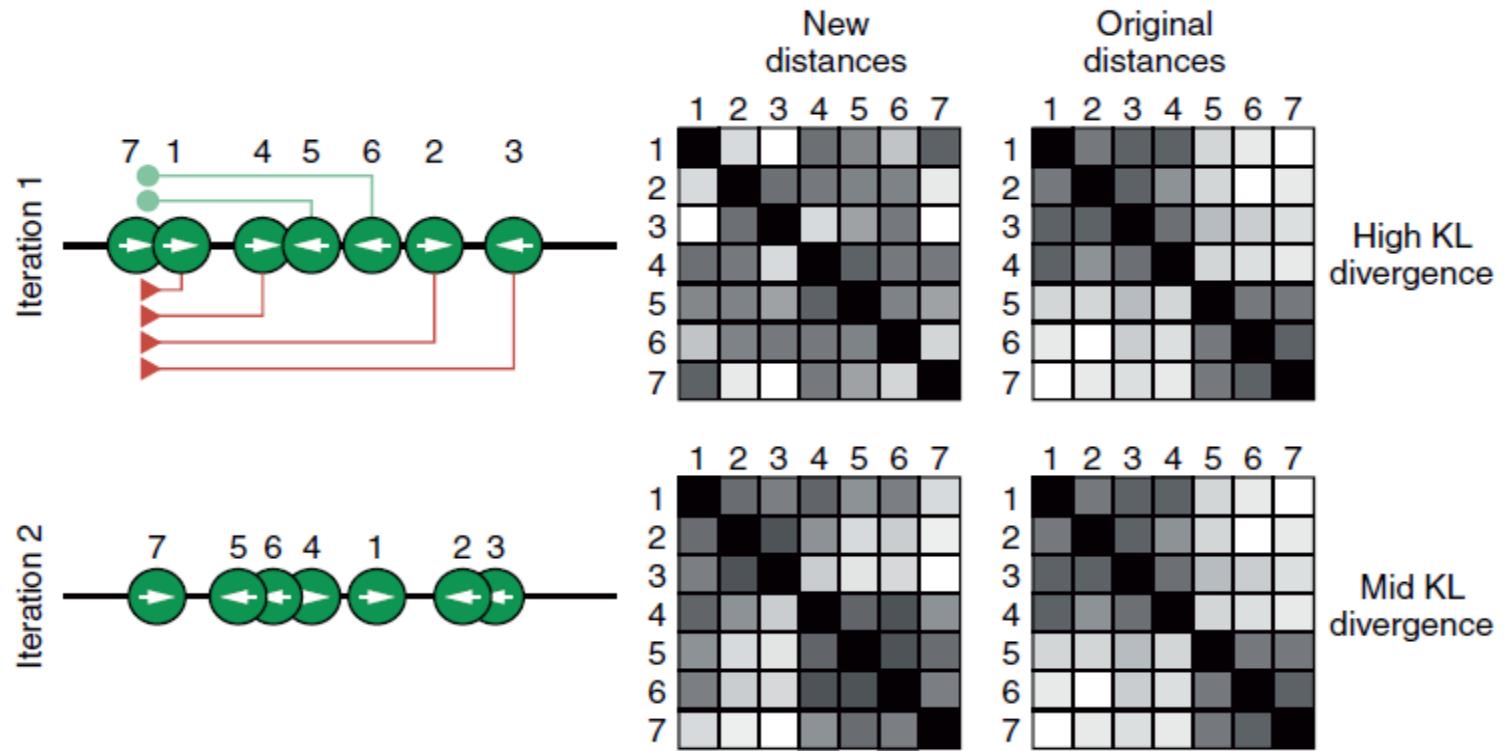
t-SNE

Several iterations



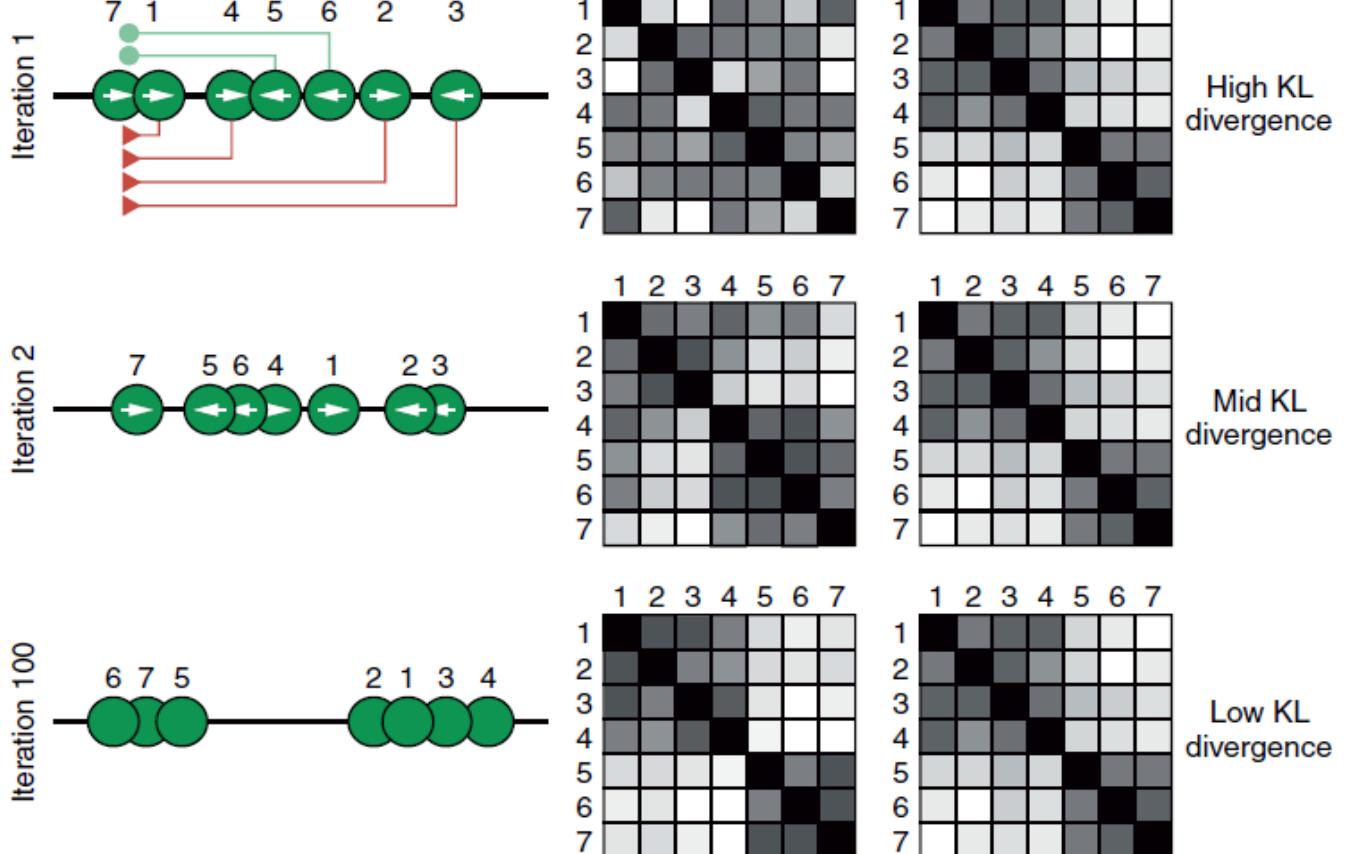
t-SNE

Several iterations



t-SNE

Several iterations



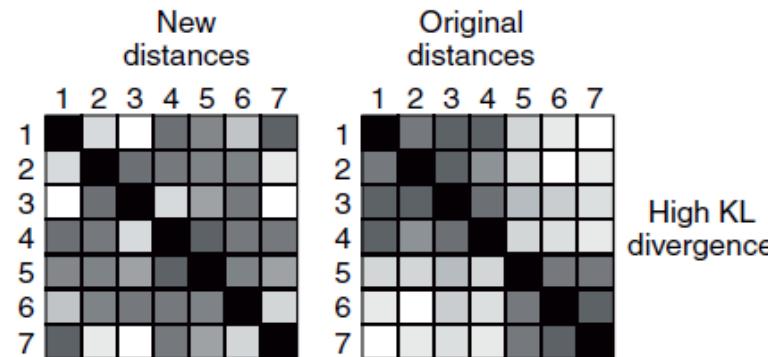
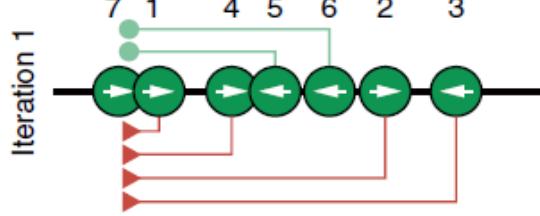
Cases are randomly initialized over the new axes (one axis is shown here).

The probability matrix is computed for this axis, and the cases are shuffled around to make this matrix resemble the original, high-dimensional matrix by minimizing the Kullback-Leibler (KL) divergence.

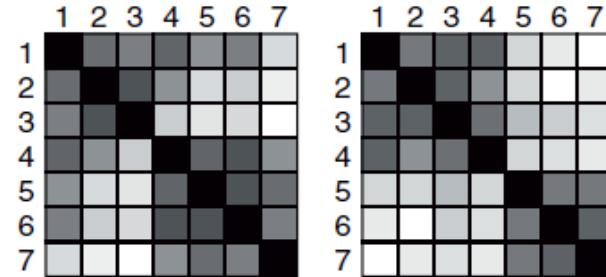
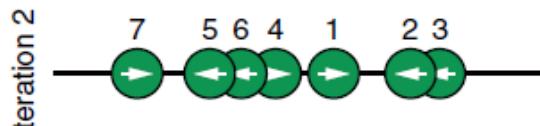
During shuffling, cases are attracted toward cases that are similar to them (lines with circles) and repulsed away from cases that are dissimilar (lines with triangles).

t-SNE

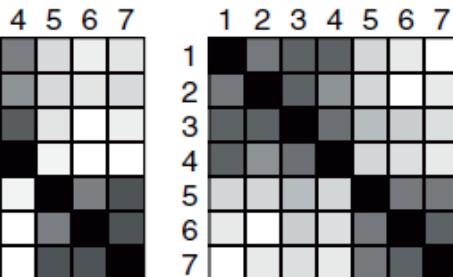
Several iterations



High KL divergence

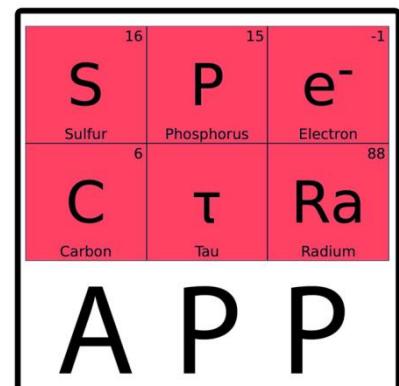


Mid KL divergence



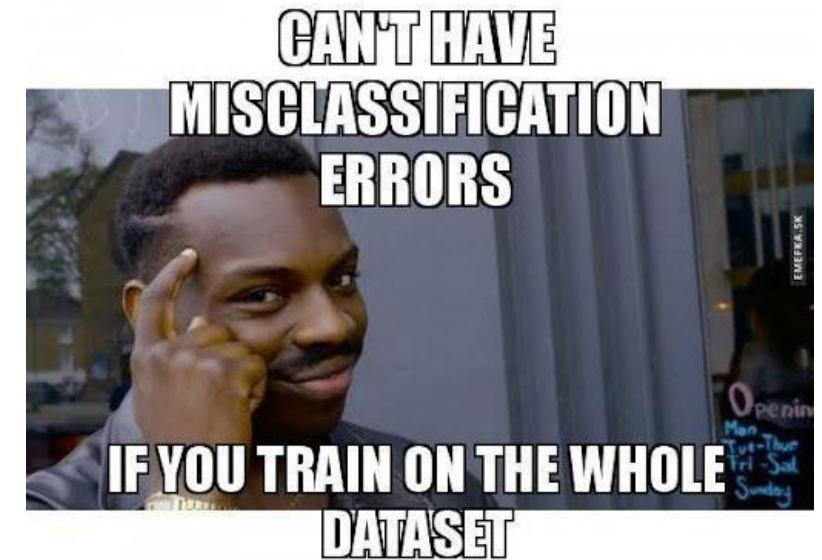
Low KL divergence

The difference between the two matrices is measured using a statistic called the **Kullback-Leibler divergence**, which is large when the matrices are very different and zero when the matrices are perfectly identical.

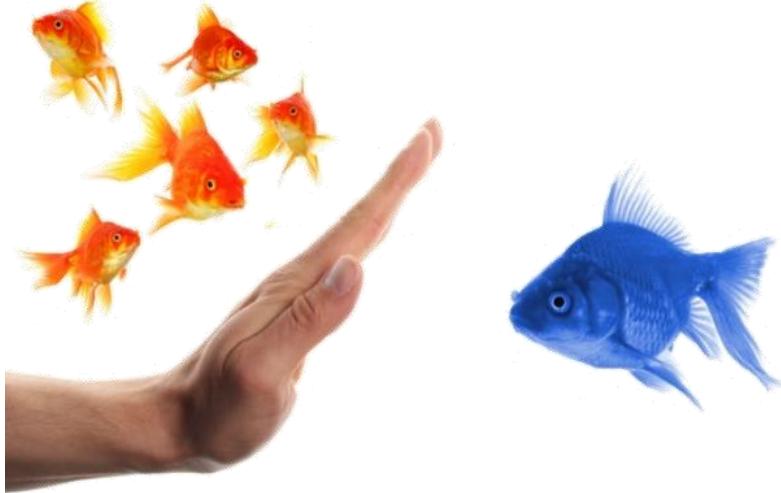




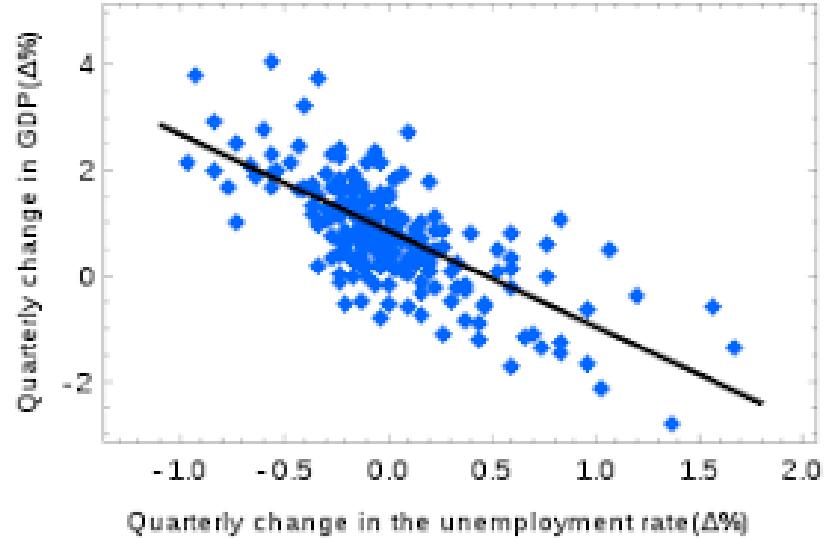
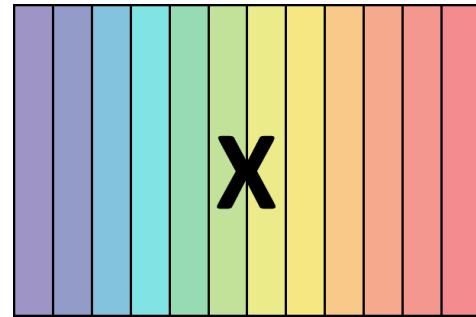
Validation



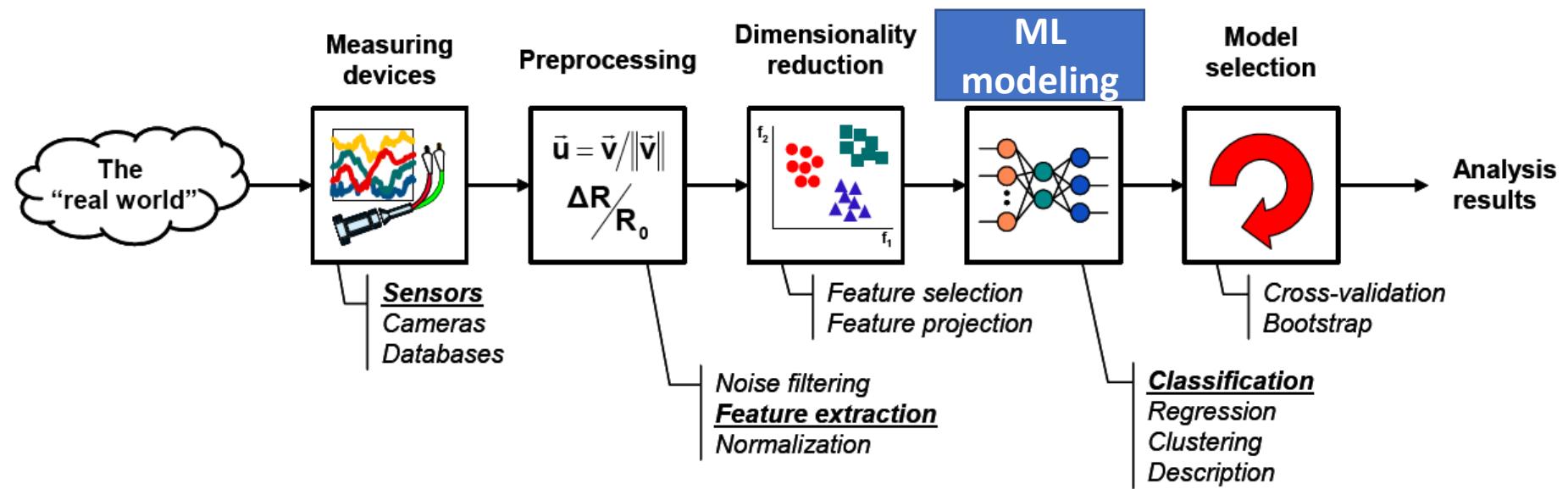
Supervised models



Regression/classification
(supervised)



Supervised models

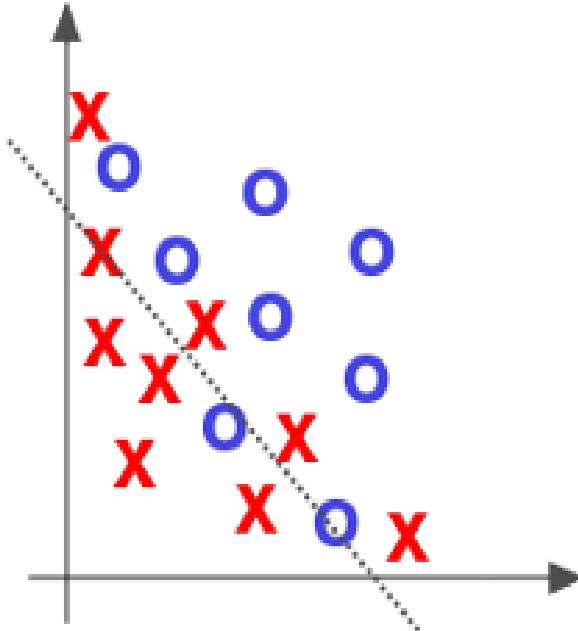


Models validation

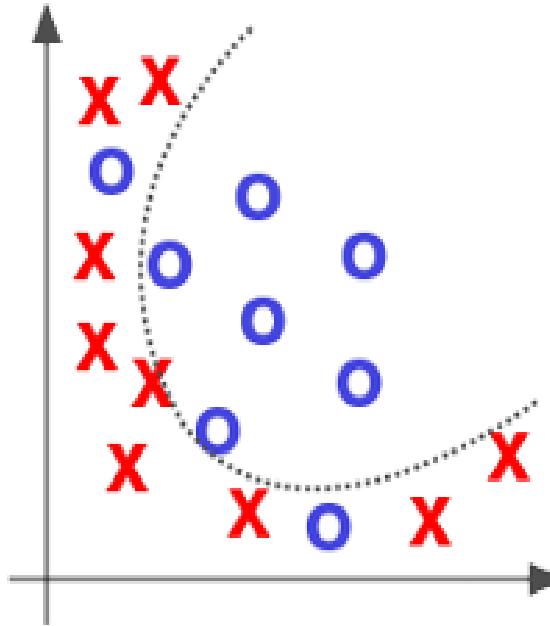
Model
Validation



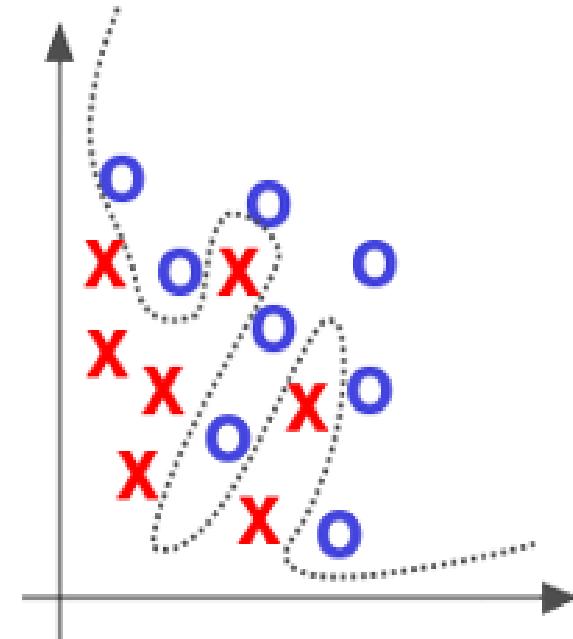
Models validation



Underfitting

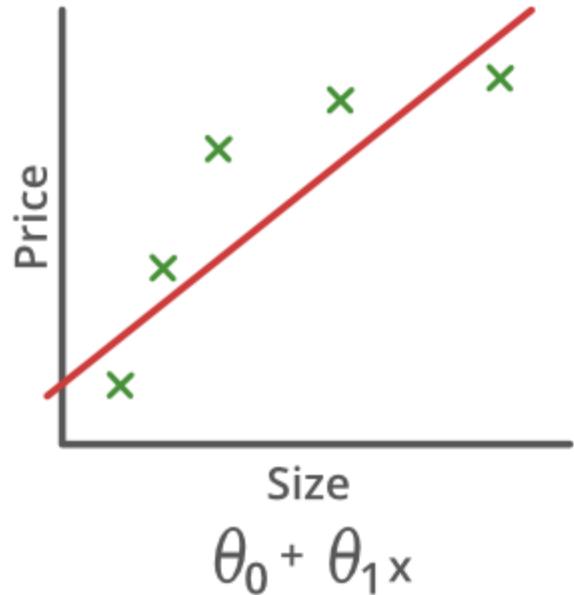


Fitting

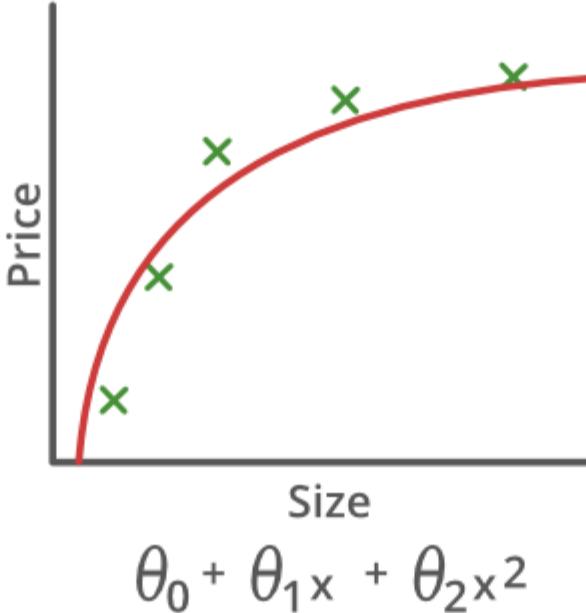


Overfitting

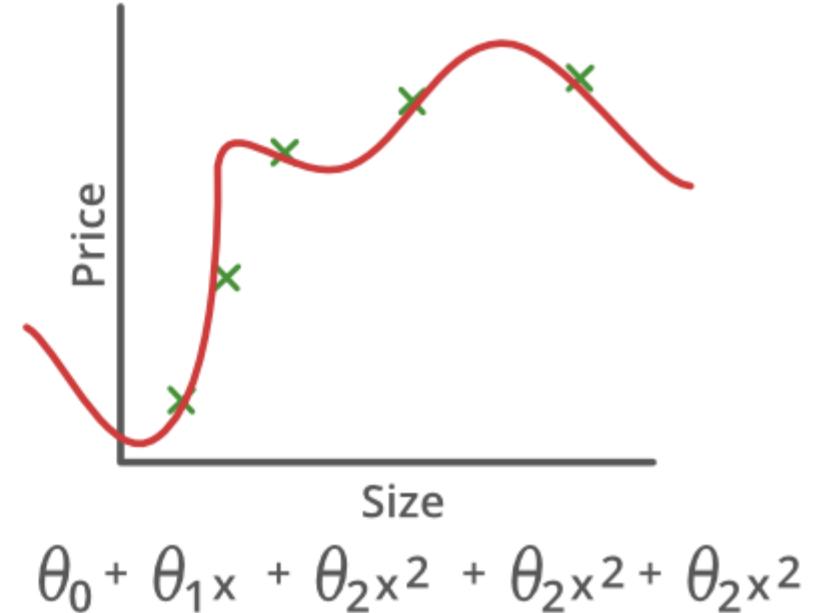
Models validation



Underfitting



Fitting



Overfitting

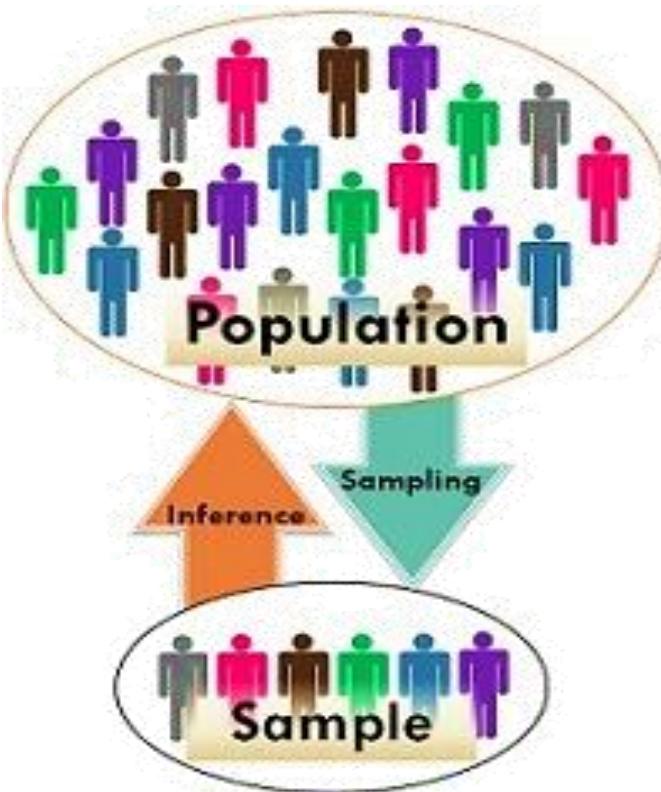
Models validation – sampling (data splitting)



If not built correctly, it is the riskiest approach!
The data should be selected as representative as possible.

Models validation – sampling (data splitting)

An experiment can theoretically be repeated infinite times: these infinite repetitions constitute a family, an infinite population of possible repetitions. In reality, an experiment is repeated a limited number of times; these repetitions constitute a **STATISTICAL SAMPLE** or simply a sample extracted from the infinite population.



Infinite-Abstraction

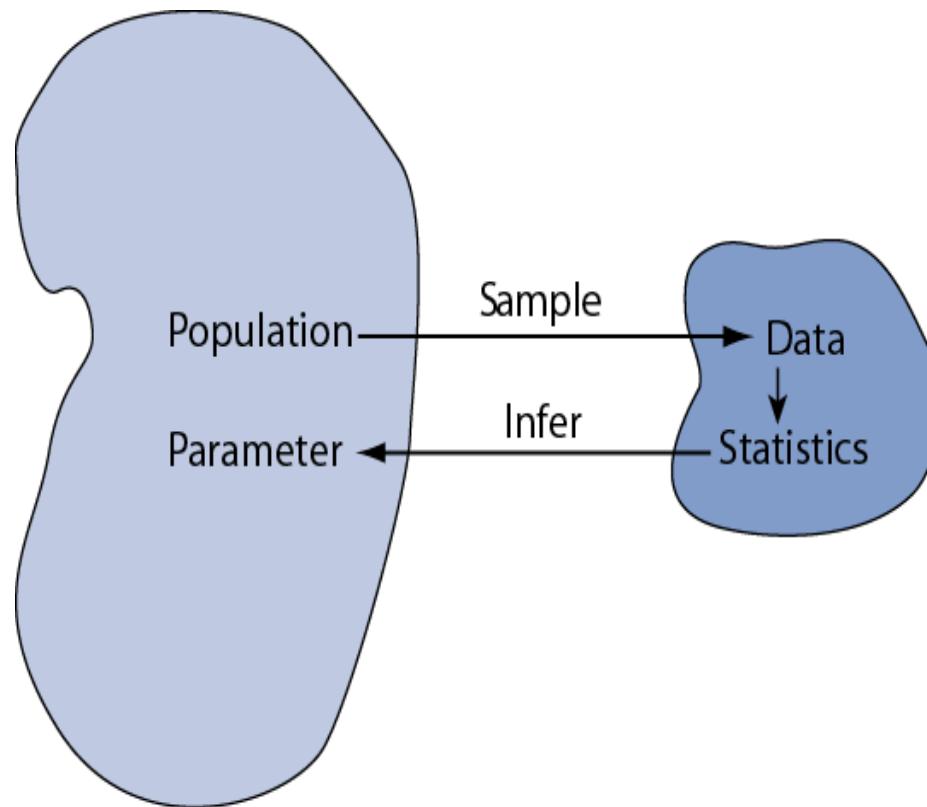
↑
Limited-Reality

A sample represents a population

Models validation – sampling (data splitting)

Inferential statistics represents the act of generalizing data and knowledge from a sample to arrive at a population/scenario with a calculated degree of confidence and significance.

For the population,
the parameters are
sought...



... using statistics
based on samples

Models validation – sampling (data splitting)

Population

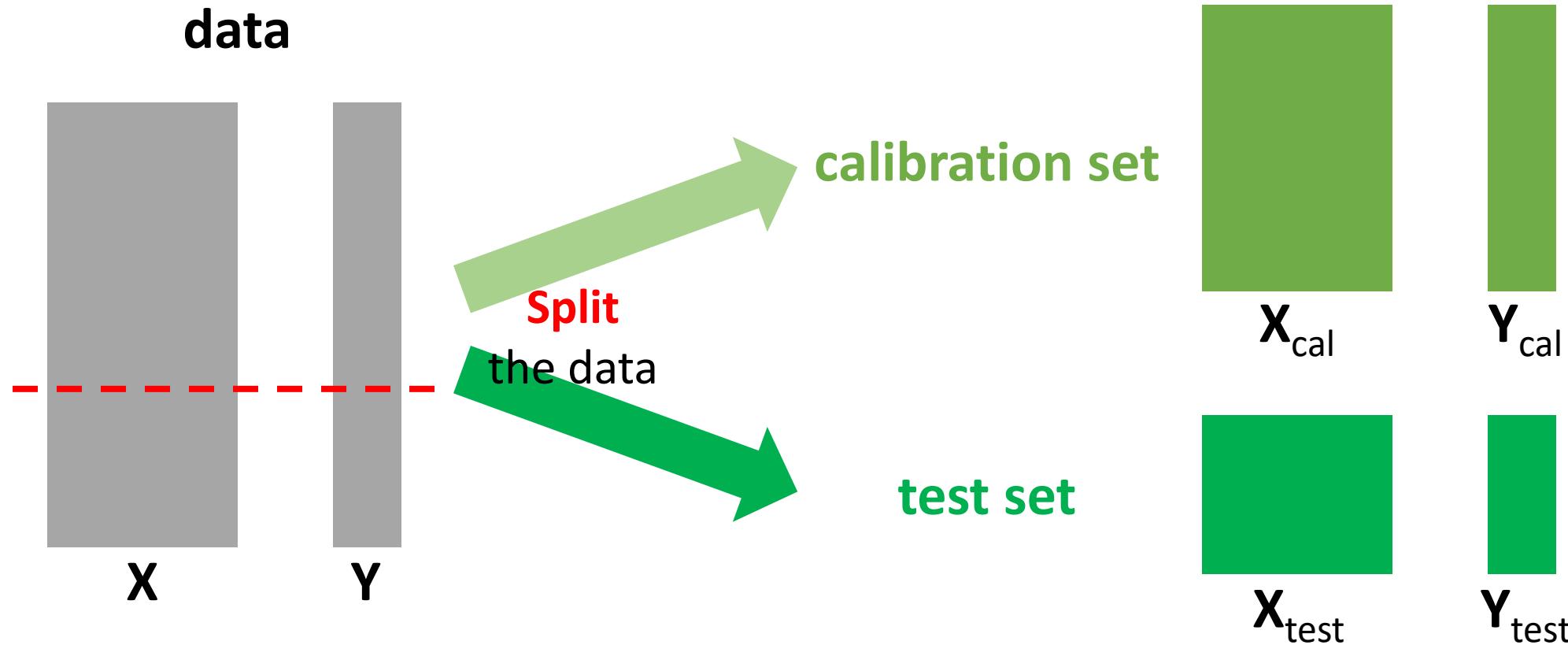


Sample



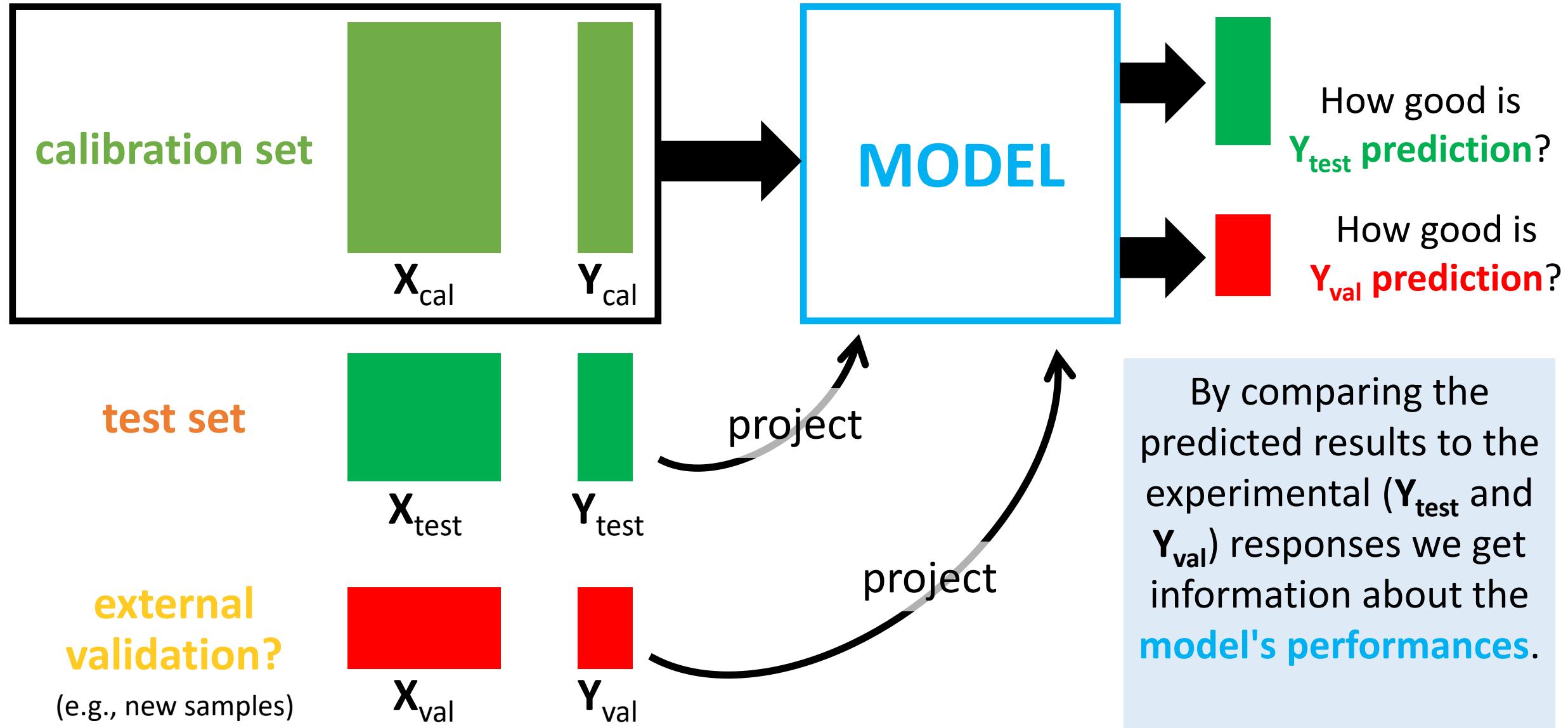
Statisticians use different types of sampling:
Random; Systematic; Accidental; Intentional; Stratified.

Models validation – sampling (data splitting)



VERY IMPORTANT: this can be done when enough samples are available, since the **calibration set** should contain enough information to build a **reliable model**.

Models validation – sampling (data splitting)

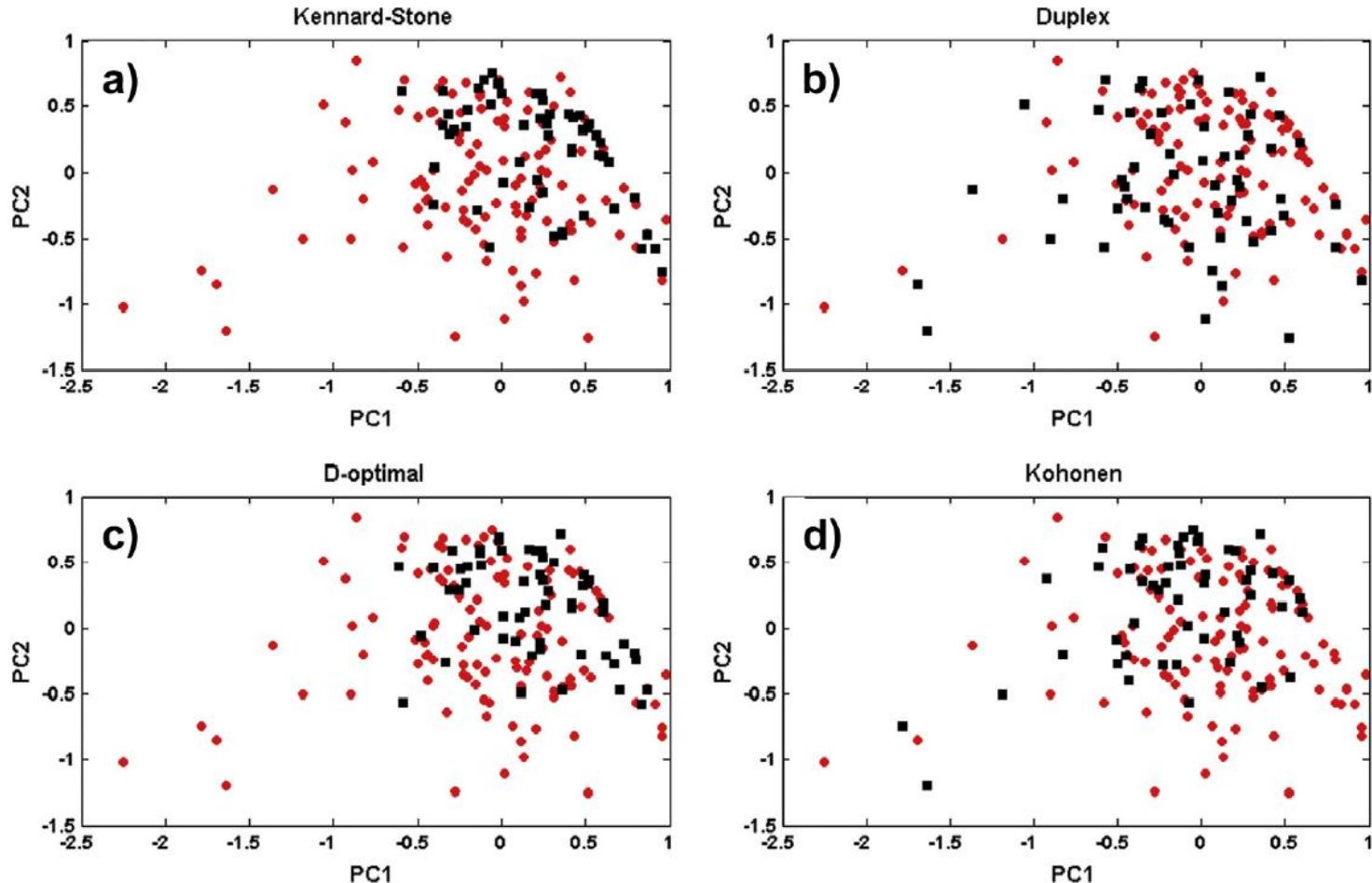
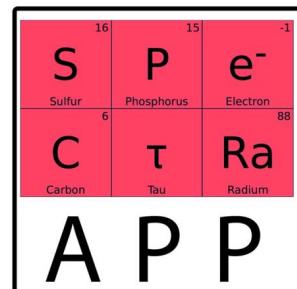


Models validation – sampling (data splitting)

Many algorithms for selecting a test set from the complete dataset are available.

Kennard-Stone, Duplex, D-optimal and Kohonen, for example, all select different "versions" of the test set, since their functioning is based on different ways of inspecting the samples distributions.

In the figure, from the complete dataset (●), the black squares (■) are selected.



Cross-validation

- Cross-validation can be considered an "**internal validation**", as the model performances are evaluated directly from the data, without a test set.
- It is also used to assess **how robust** is the information contained in the data, in relation to the modelling aims.
- An important piece of information that can be obtained using cross-validation regards **the system's dimension** or, in other words, the number of principal components (in PCA) or latent variables (in PLS and PLS-DA) that should be chosen in order to correctly model the data.
- Moreover, it can **serve as an actual validation method** when not enough samples are available (in general, <40).

Cross-validation

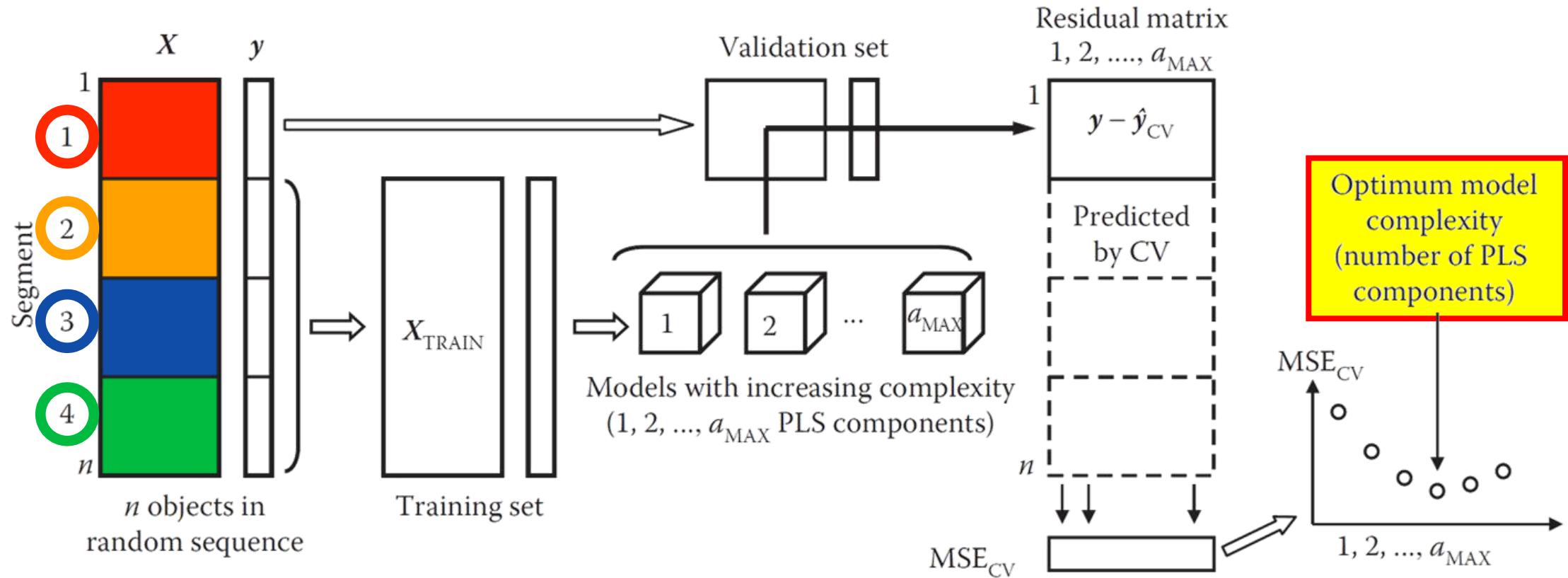


FIGURE 4.5 CV with four segments (leave-a-quarter-out) applied to estimation of the optimum complexity of the model.

Cross-validation

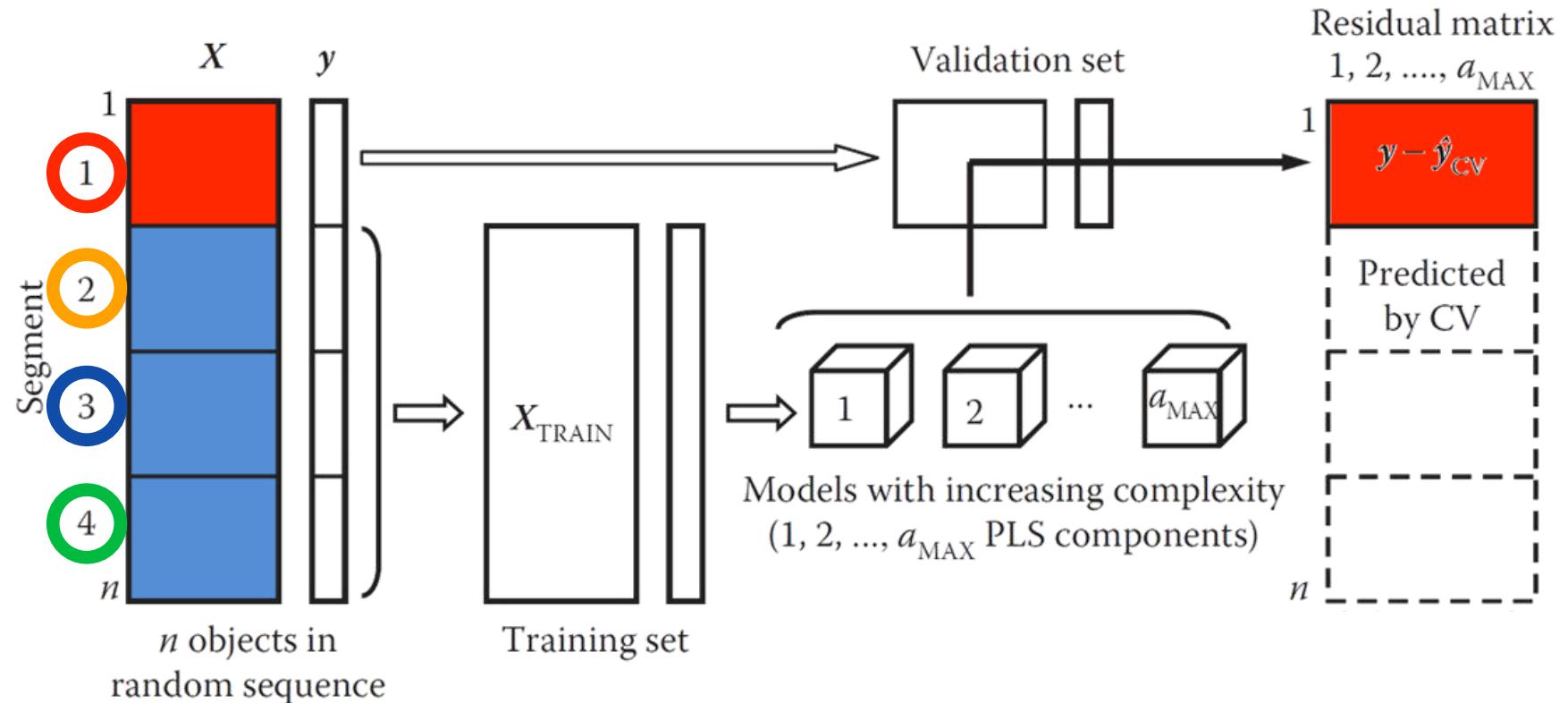


FIGURE 4.5 CV with four segments (leave-a-quarter-out) applied to estimation of the optimum complexity of the model.

Cross-validation

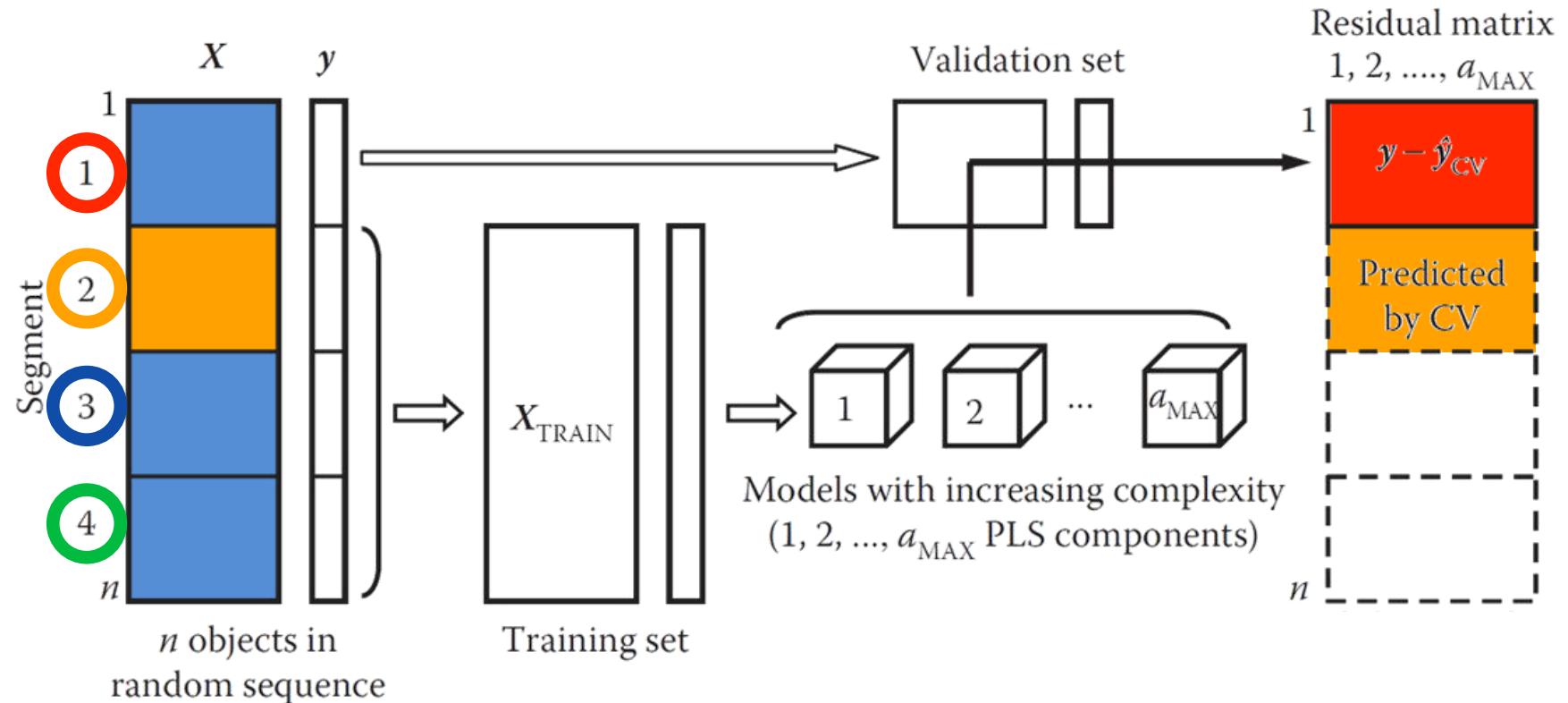


FIGURE 4.5 CV with four segments (leave-a-quarter-out) applied to estimation of the optimum complexity of the model.

Cross-validation

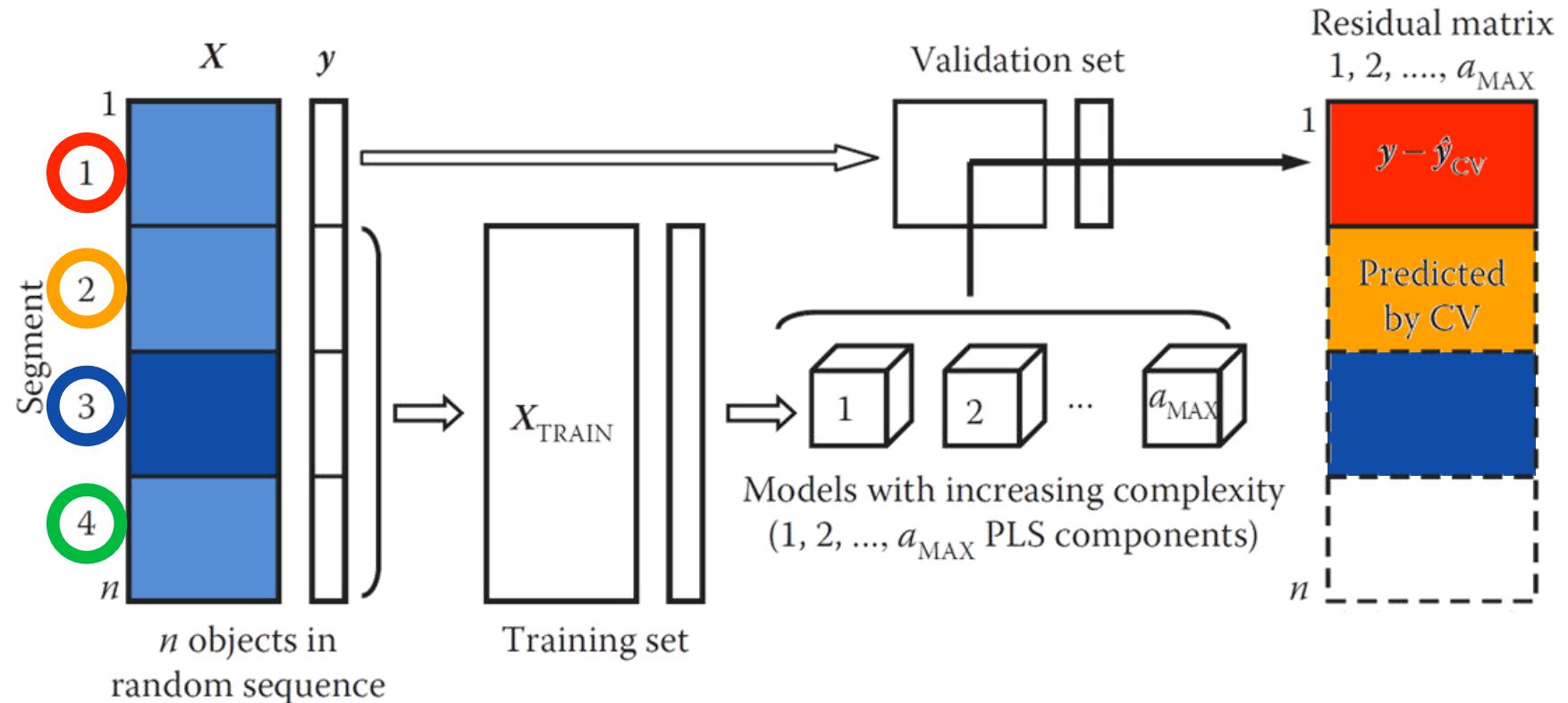


FIGURE 4.5 CV with four segments (leave-a-quarter-out) applied to estimation of the optimum complexity of the model.

Cross-validation

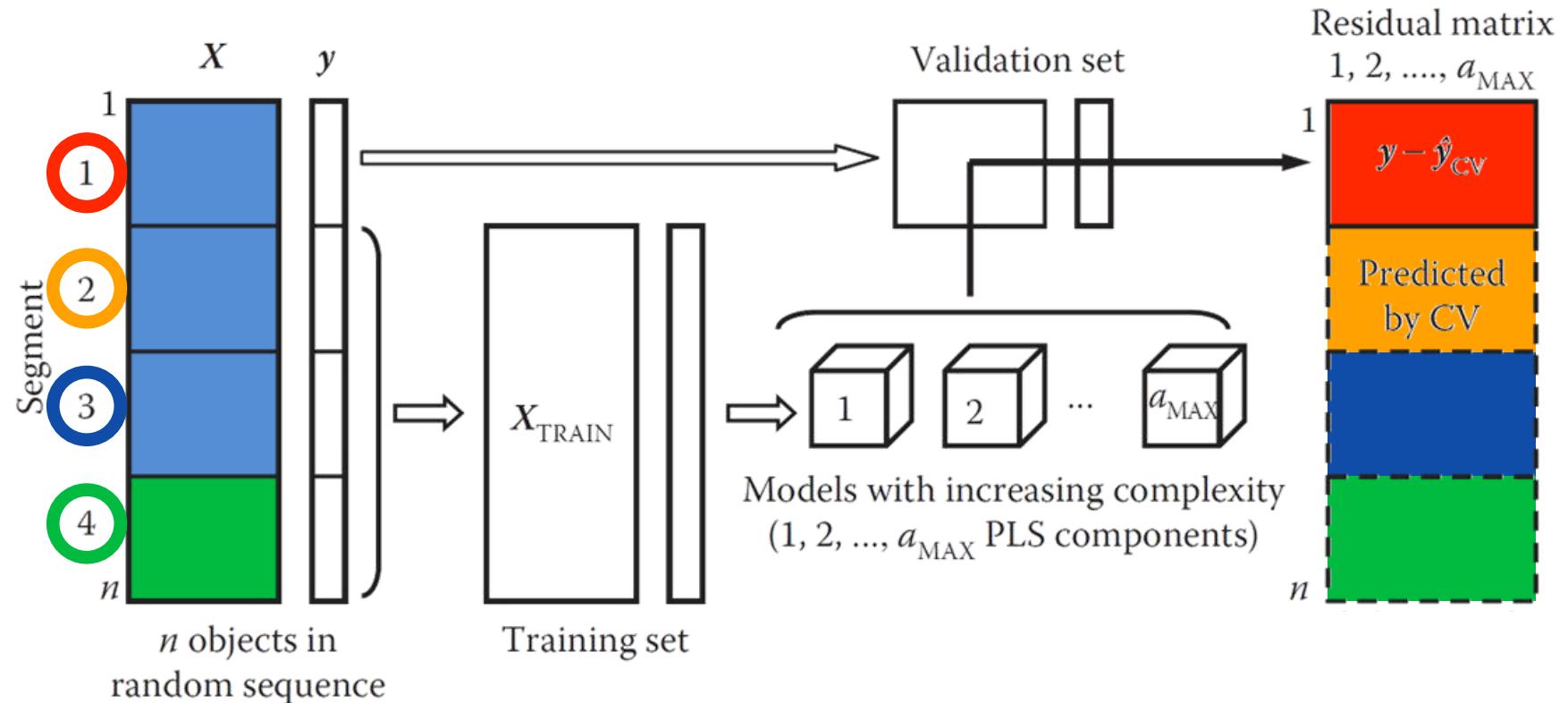


FIGURE 4.5 CV with four segments (leave-a-quarter-out) applied to estimation of the optimum complexity of the model.

Cross-validation

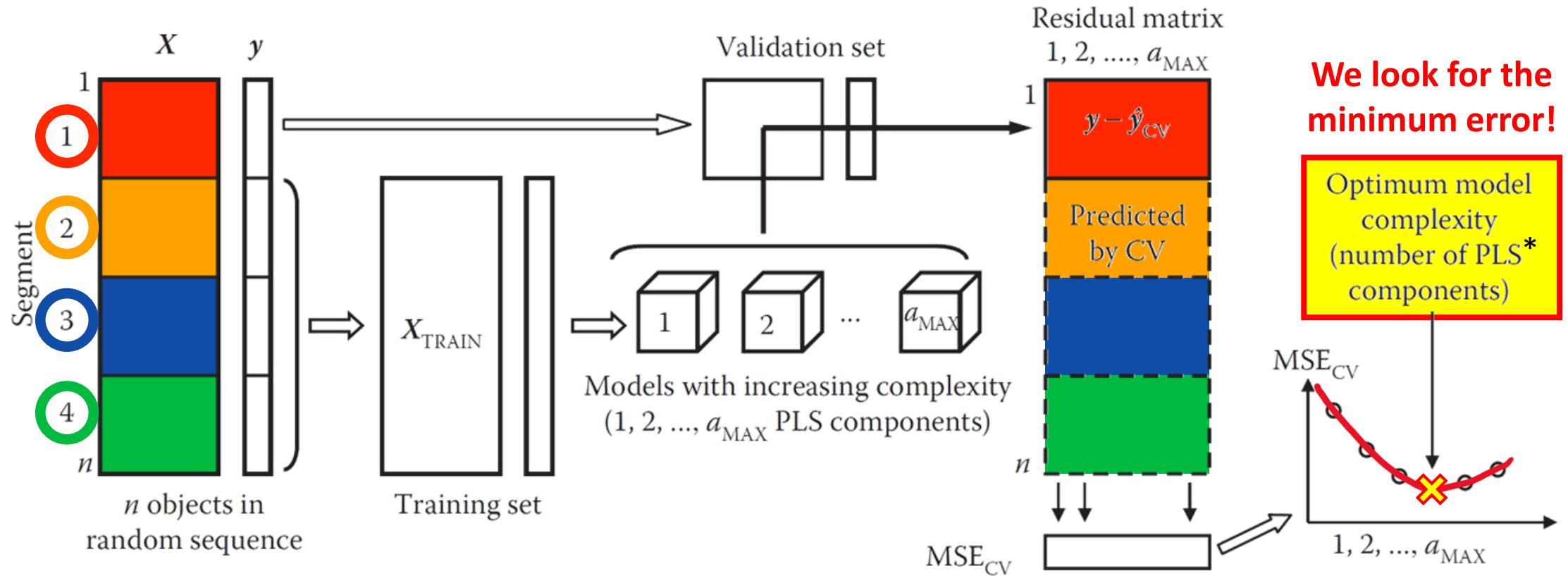
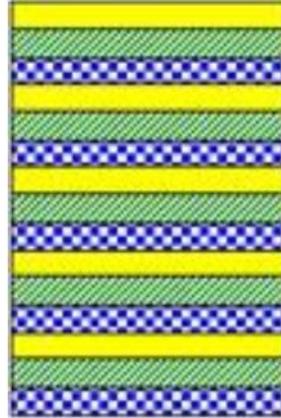
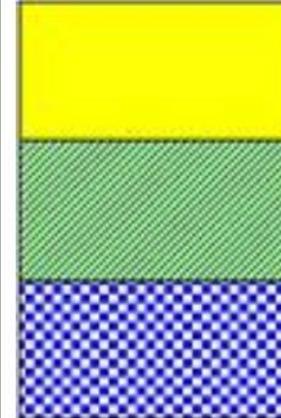


FIGURE 4.5 CV with four segments (leave-a-quarter-out) applied to estimation of the optimum complexity of the model.

* it can be any method (PCA, PLS, PLS-DA, ...)

Cross-validation

Venetian Blinds	Contiguous Blocks	Random Subsets	Leave-One Out
			

VERY IMPORTANT: if you have **replicates** of the same sample,
these must be kept together! (i.e. in the same CV block)

Otherwise, your model is going to be overoptimistic, since information about the same sample is going to be present in both the calibration and test sets (i.e. the CV blocks)

...the two sets would not be independent!

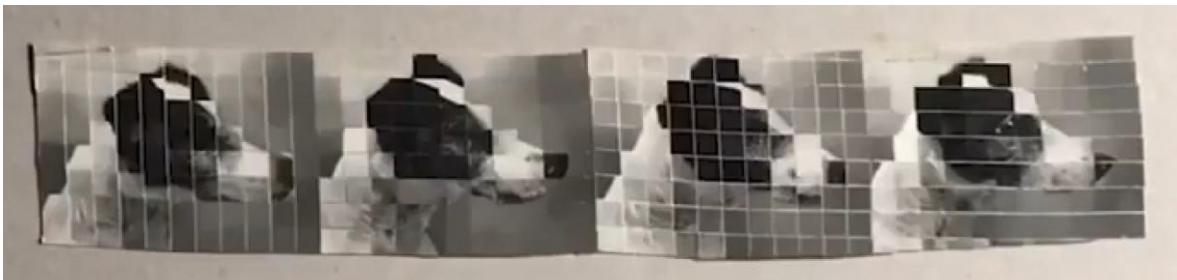
Cross-validation

A visual representation of **why it is important to select representative subsets** during any validation step.



↑ Your dataset

Four representative subsets! →



Validation mantra

- When do we use cross-validation?

ALWAYS!



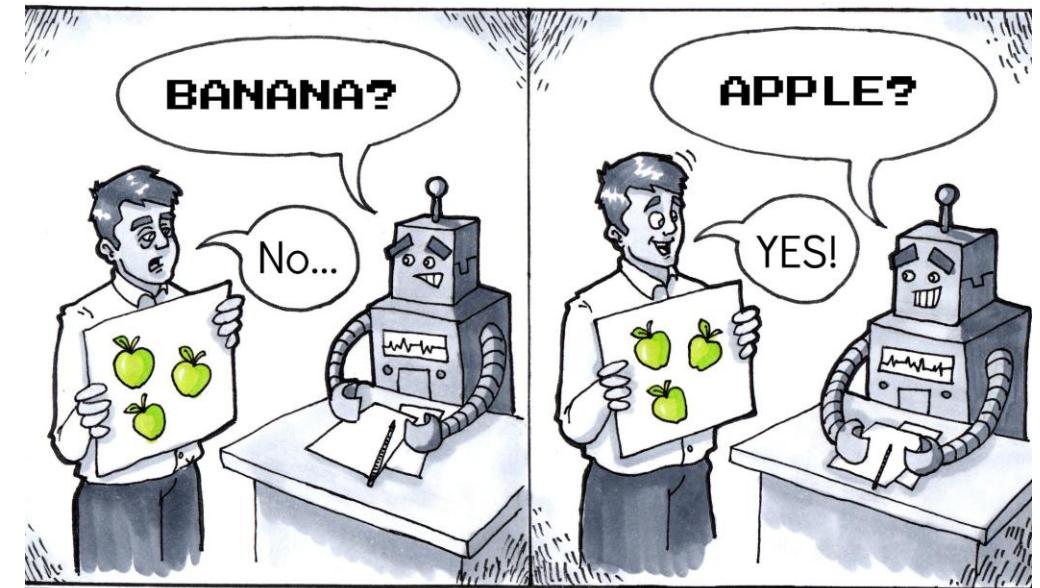
- When do we split the dataset into two?

If we have enough samples!

- When do we know that we have enough samples?

Experience!

Supervised learning

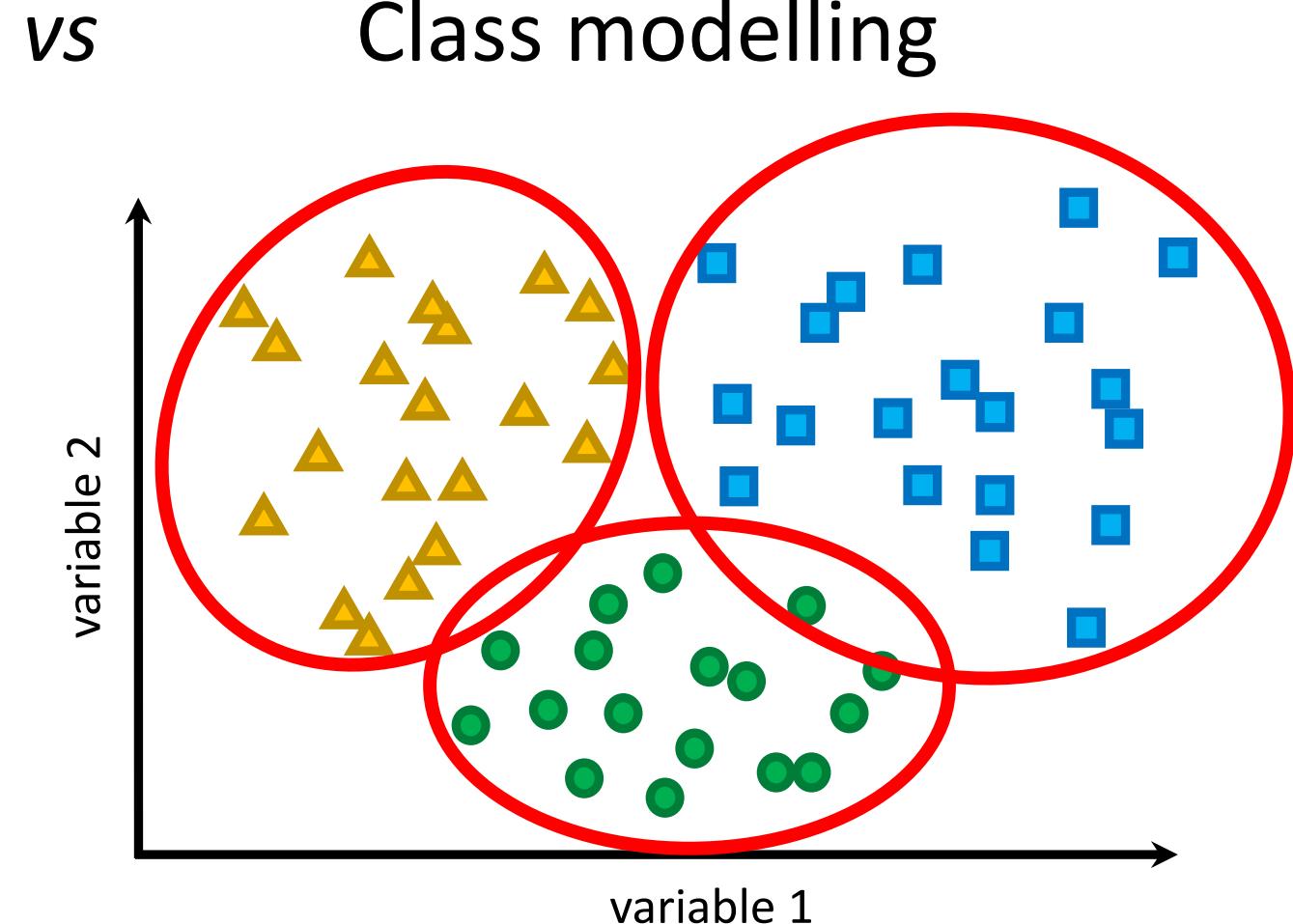
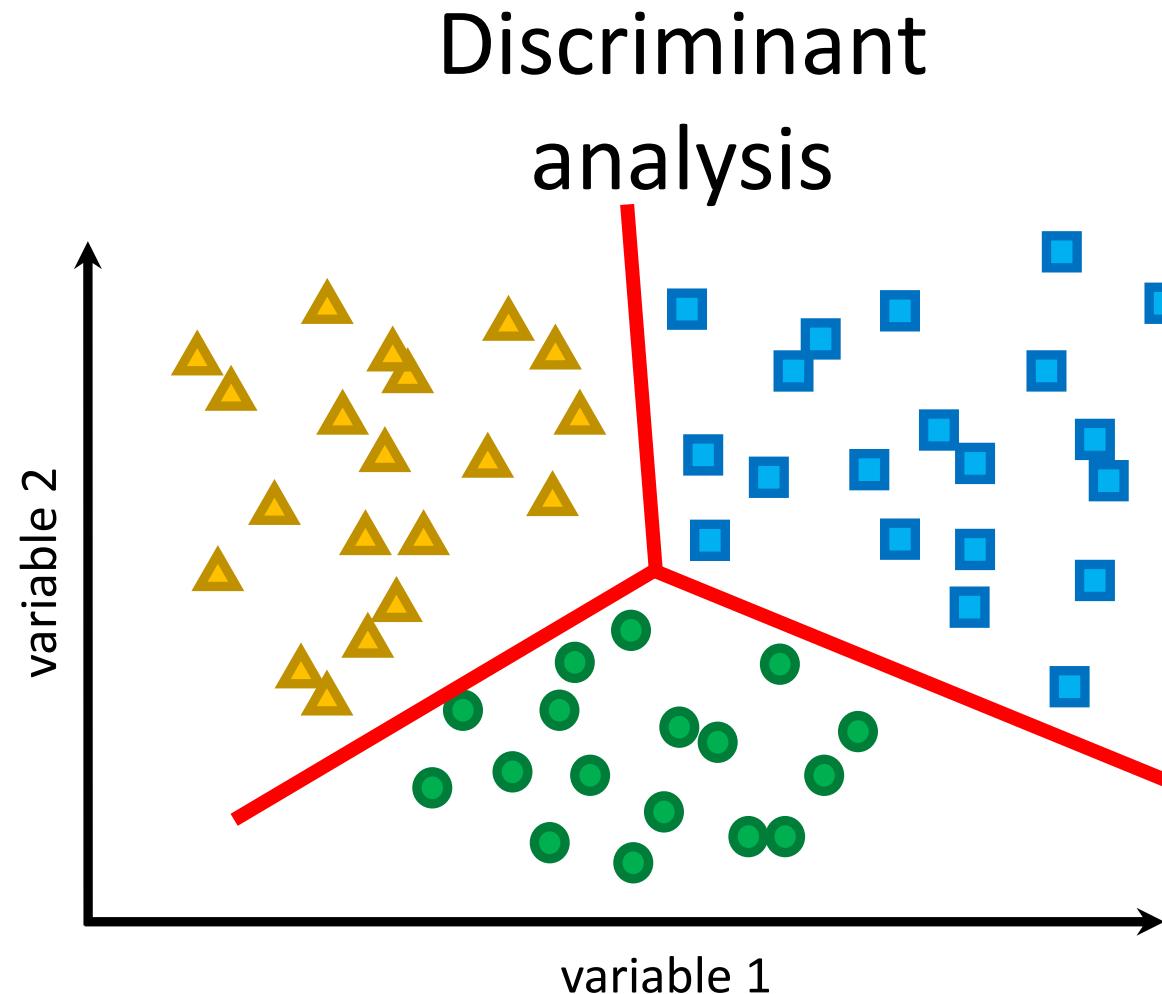


Supervised Learning

Supervised models - classification

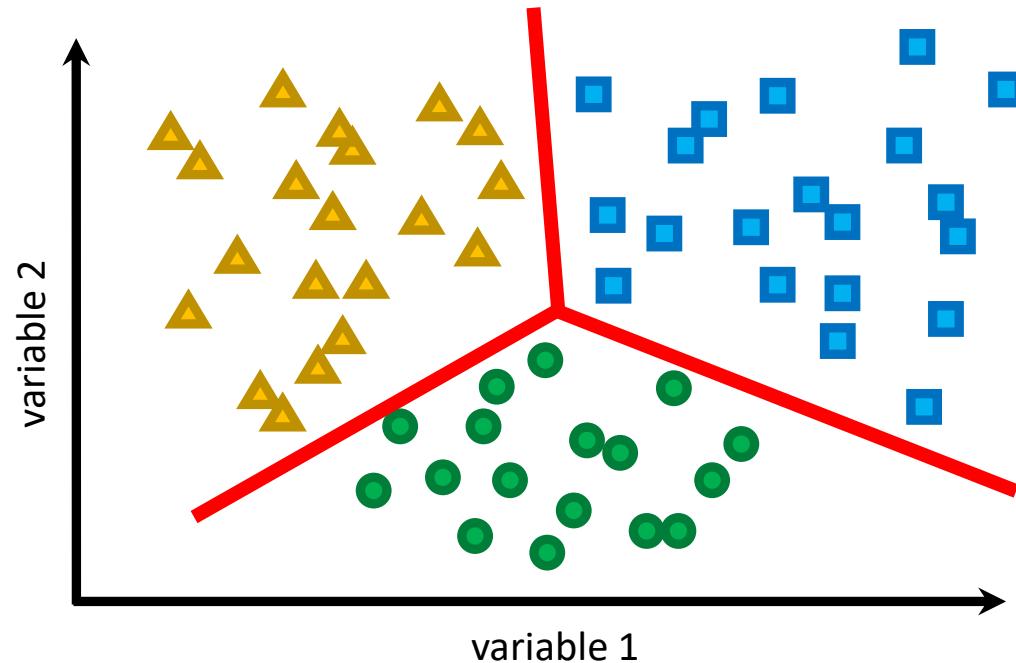


Multivariate Classification



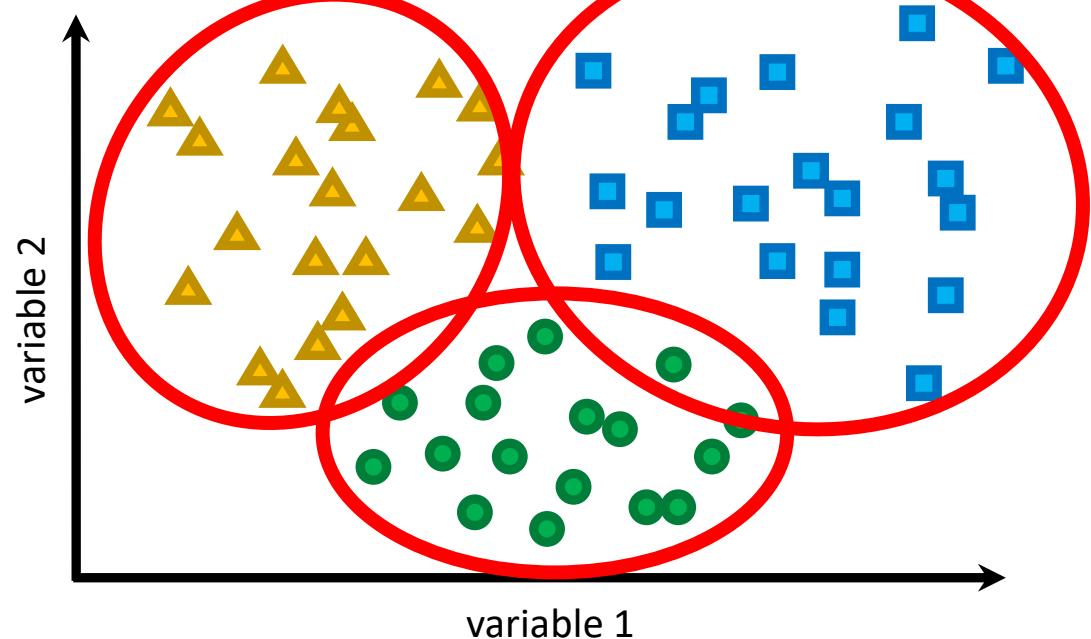
Multivariate Classification

Discriminant analysis



VS

Class modelling



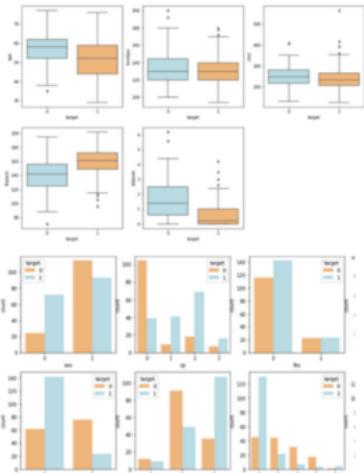
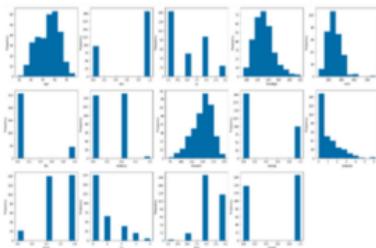
- It is a **single model**.
- Each sample is always assigned to one class.
- There are no chances to obtain multiple assignments or to not assign a sample at all!

- It is an **ensemble of models**.
- Each sample can either be assigned:
 - to no class (rejected by all classes)
 - to one class
 - to more than one class

Machine Learning Algorithms - Classification

Exploratory Data Analysis (EDA)

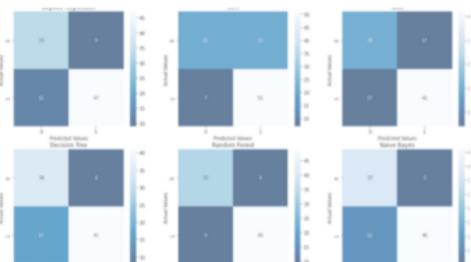
- 1) Histogram: `df.plot(kind = 'hist')`
- 2) Box Plot: `sns.boxplot()`
- 3) Grouped Bar Chart: `sns.countplot()`



Model Evaluation

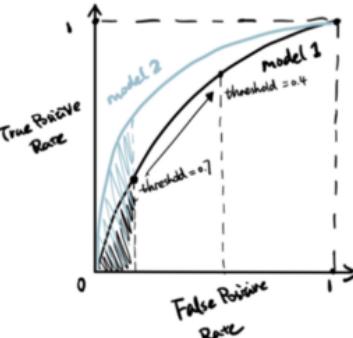
Confusion Matrix

```
confusion_matrix(y_test, y_pred)
```

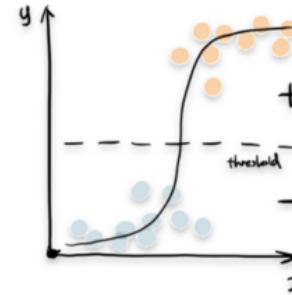


ROC & AUC

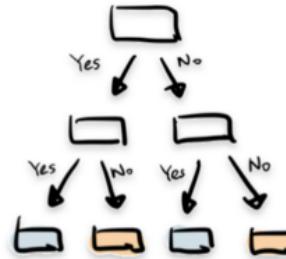
```
metrics.auc(fpr, tpr)
```



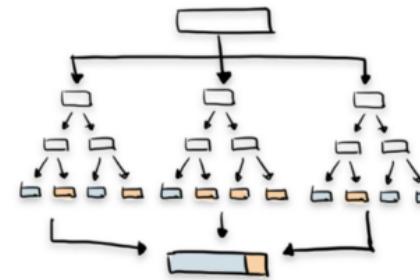
Logistic Regression



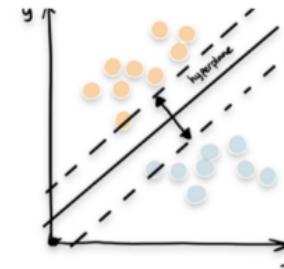
Decision Tree



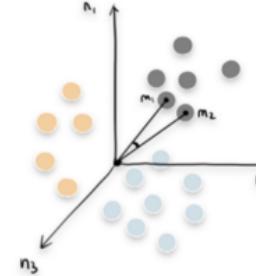
Random Forest



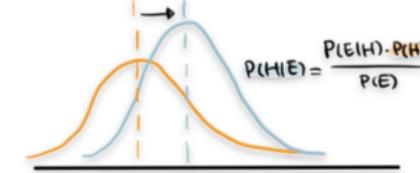
Support Vector Machine



K Nearest Neighbour



Naive Bayes



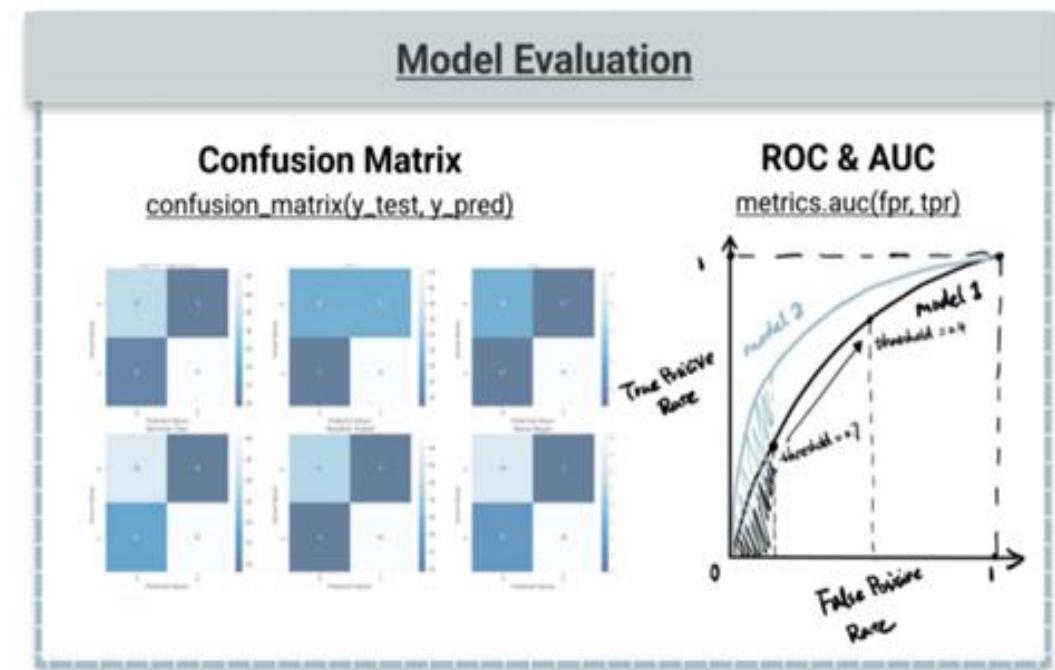
Classification & Class Modeling – Model Evaluation

Whatever technique is used to build a classification model, the evaluation of the model performance is not postponed to a stage subsequent to model building, but rather it is continuously involved within the model generation and optimization process. Some techniques simply requires the optimization of some model parameters, other techniques implies iterative trial-and-error procedures, but in both contexts recurrent evaluation of the developing model is necessary.

Commonly used evaluation tools presented in this course are the following:

Class models evaluation methods:

- 1) Confusion Matrix – Loss Matrix
- 2) Non-Error-Rate – Misclassification risk
- 3) Sensitivity & Specificity
- 4) Receiver Operating Characteristic (ROC) Curve
- 5) Weighted Gini Impurity
- 6) Entropy & Cross-Entropy Cost Function



Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

TP = true positive

Number of **P** samples correctly classified as **P**.

		Predicted	
		P	N
Real	P	TP	FN
	N	FP	TN

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

TP = true positive

Number of **P** samples correctly classified as **P**.

TN = true negative

Number of **N** samples correctly classified as **N**.

		Predicted	
		P	N
Real	P	TP	FN
	N	FP	TN

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

		Predicted
		P N
Real	P	TP FN
	N	FP TN

TP = true positive

Number of **P** samples correctly classified as **P**.

TN = true negative

Number of **N** samples correctly classified as **N**.

FP = false positive

Number of **N** samples wrongly classified as **P**.

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

		Predicted
		P N
Real	P	TP FN
	N	FP TN

TP = true positive

Number of **P** samples correctly classified as **P**.

TN = true negative

Number of **N** samples correctly classified as **N**.

FP = false positive

Number of **N** samples wrongly classified as **P**.

FN = false negative

Number of **N** samples wrongly classified as **N**.

Evaluation of Classification Models – Confusion Matrix

The **Confusion Matrix** is a useful tool to evaluate the performance of a classification model. It can be used with both the training set and the evaluation set, as in both cases the true class of each object is known in advance. Whatever the objects' set, the matrix lines represent the true class of the objects, while the columns represent the class to which the objects have been assigned.

In the example reported below, a set of $N=30$ objects belonging to the classes A (10 objects), B (12), and C (8) are subjected to a classification model that predicts the class of assignment. 12 objects are assigned to the class A, 11 to B, and 7 to C, with the distribution evidenced by the matrix.

For example, only 8 objects (out of 12) belonging to the class B are assigned correctly, while 2 of them are assigned to the class A and other 2 to the class C. On the other hand, 1 object of the class A and 2 objects of the class C are incorrectly assigned to the class B. Apparently, 11 objects are assigned to the class B, but only 8 of them are really belonging to the class B. The correctly assigned objects are on the matrix diagonal.

The simplest parameter used to summarize the results reported in a confusion matrix is the percentage of objects that have been assigned correctly: the **non-error rate (NER%)**.

Confusion Matrix		Assigned Class (a)			N
		A'	B'	C'	
True Class (t)	A	9	1	0	10
	B	2	8	2	12
	C	1	2	5	8
	N'	12	11	7	30

$$\begin{aligned} NER\% &= \frac{\sum_g c_{gg}}{N} \times 100 \\ &= \frac{9 + 8 + 5}{30} \times 100 = 73.3\% \end{aligned}$$

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

		Predicted	
		P	N
Real	P	TP	FN
	N	FP	TN

Sensitivity (Sn) of P class - RECALL

- Ability to avoid FN (P wrongly classified as N)
- The higher, the better!

$$Sn = \frac{TP}{TP + FN}$$

With two classes, the sensitivity of one class is equal to the specificity of the other class!

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

		Predicted	
		P	N
Real	P	TP	FN
	N	FP	TN

Specificity (Sp) of P class

- Ability to avoid FP (N wrongly classified as P)
- The higher, the better!

$$Sp = \frac{TN}{FP + TN}$$

With two classes, the specificity of one class is equal to the sensitivity of the other class!

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

		Predicted
		P N
Real	P	TP FN
	N	FP TN

Accuracy (Acc)

- Estimation of the model error
- The higher, the better!

$$\text{Acc} = \frac{TP + TN}{\text{samples}}$$

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

		Predicted
		P N
Real	P	TP FN
	N	FP TN

Non Error Rate (NER)

- Mean of sensitivities
- Model capability to correctly classify
- The higher, the better!

$$NER = \text{mean}(Sps)$$

Error Rate (ER)

- Estimation of model error
- The lower, the better!

$$ER = 1 - NER$$

Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**

$$SENSITIVITY = \frac{TP}{TP + FN} \cdot 100\%$$

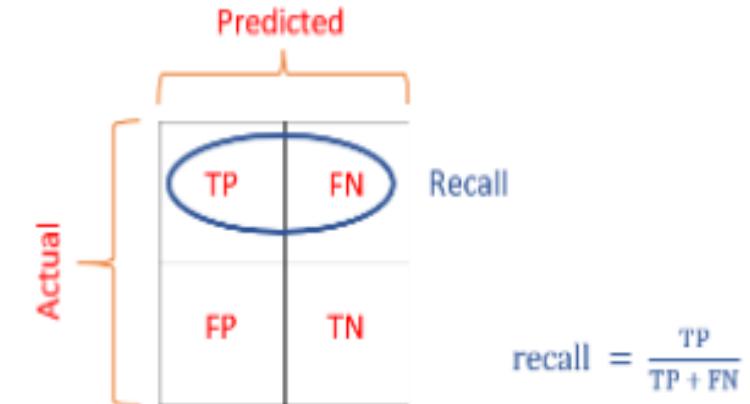
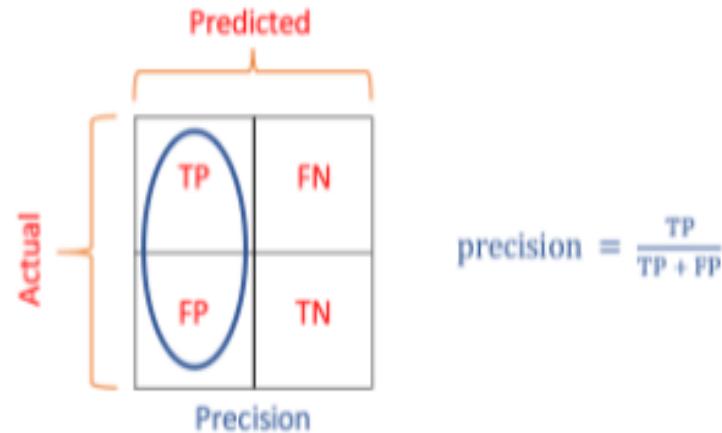
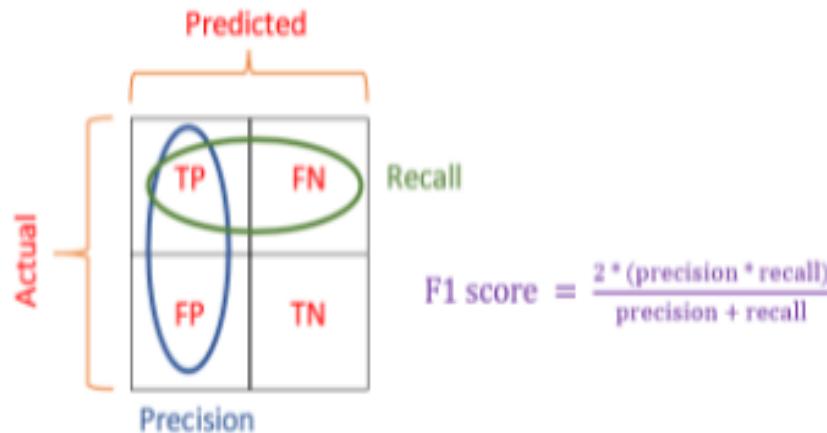
$$SPECIFICITY = \frac{TN}{TN + FP} \cdot 100\%$$

$$EFFICIENCY = \sqrt[2]{SENSITIVITY \cdot SPECIFICITY}$$

Sensitivity is an **experimental measure of the confidence level of a specific constructed model**.

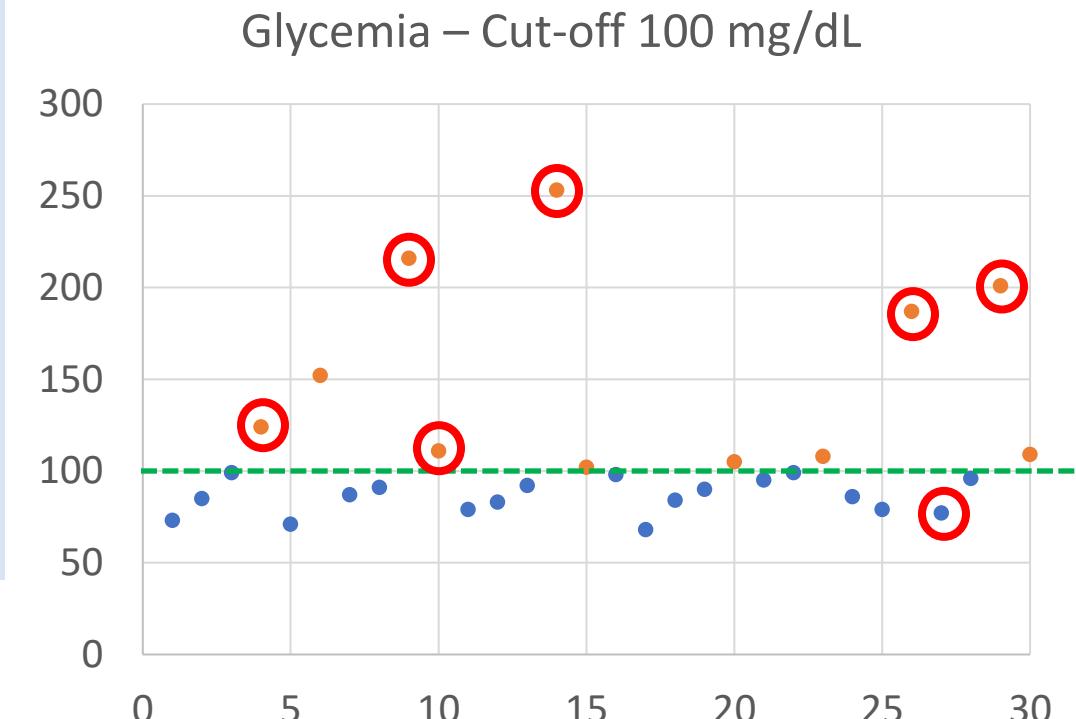
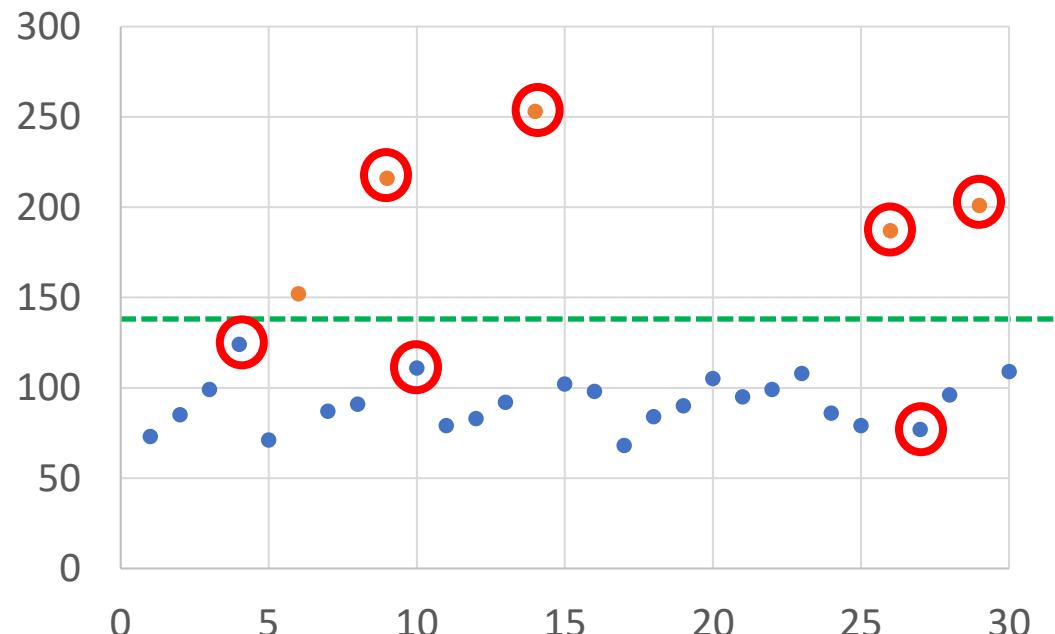
Classification measures (performance metrics)

All “classification measures” are derived from the **confusion matrix**



Evaluation of Classification Models – Receiver Operating Characteristic Curves

When the response of a diagnostic test is a certain value of a continuous variable (for example, colesterolemia, glycemia, etc.), then a “cut-off” decision threshold is chosen to classify the samples as either positive or negative. In the alongside figure, a cut-off of 100 mg/dL has been chosen and the samples above this threshold are classified as positive (red dots), while the samples below 100 mg/dL are classified as negative (blue dots). Thus, 11 samples out of 30 are “positive” and 19 are “negative”. If we choose a cut-off of 140 mg/dL, only 5 samples are “positive” and 25 “negative” (figure below).



Suppose that only the patients whose samples are red-circled have diabetes, while the others are healthy. In the first case, **with 100 mg/dL** cut-off, we obtain: **6 TP, 5 FP, 18 TN, 1 FN – Sn 86% - Sp 78%**. **With 140 mg/dL** threshold: **4 TP, 1 FP, 22 TN, 3 FN – Sn 57% - Sp 96%**. By increasing the cut-off value, the sensitivity decreases while the specificity increases.

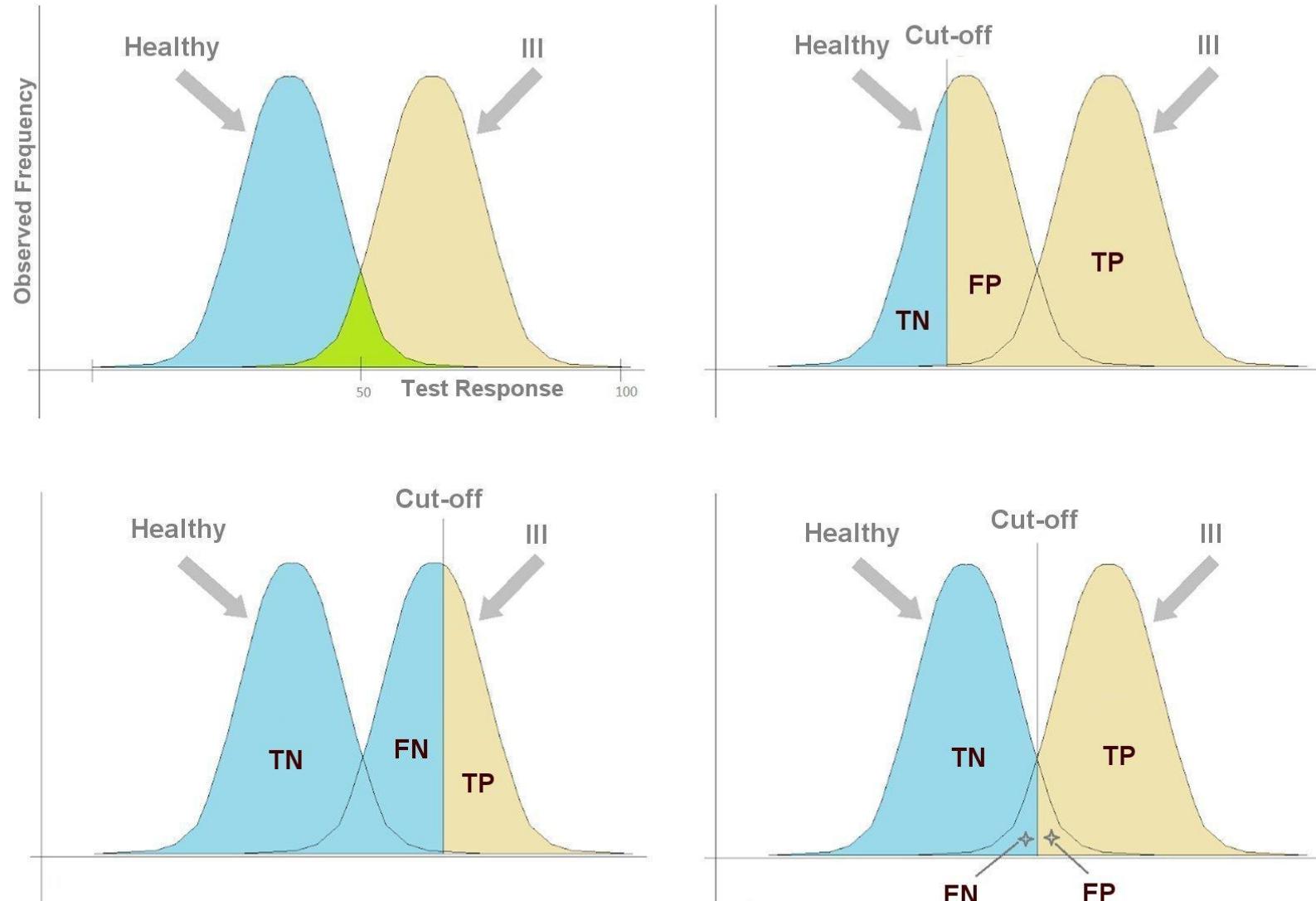
Evaluation of Classification Models – Receiver Operating Characteristic Curves

This is a common feature: if high values of the chosen biomarker indicate a pathological condition, by increasing the cut-off value, the sensitivity decreases while the specificity increases.

The **Receiver Operating Characteristic Curve (ROC curve)** is a diagram reporting sensitivity vs. specificity as a function of the variable cut-off value.

$$Sn_g = \frac{TP}{TP + FN}$$

$$Sp_g = \frac{TN}{TN + FP}$$

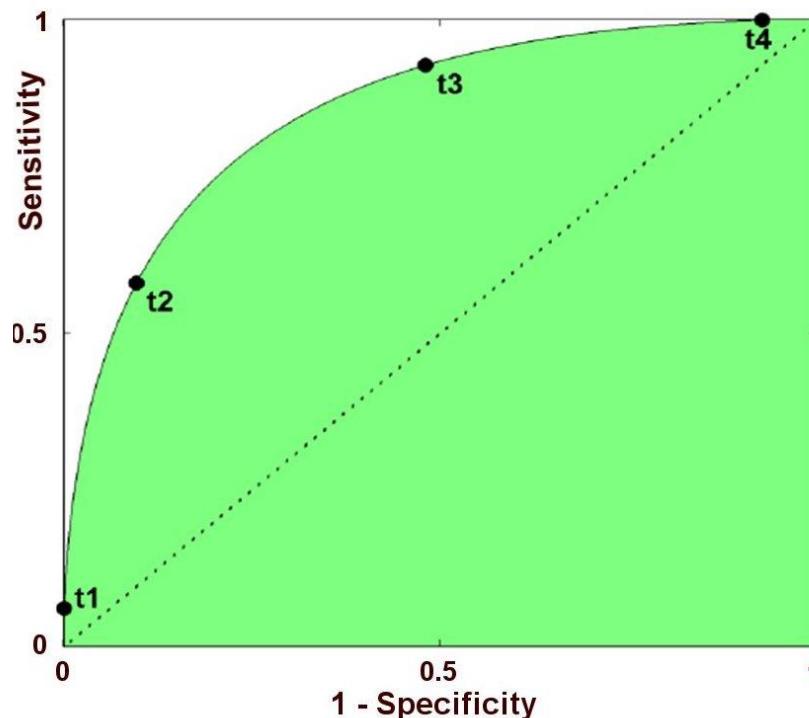
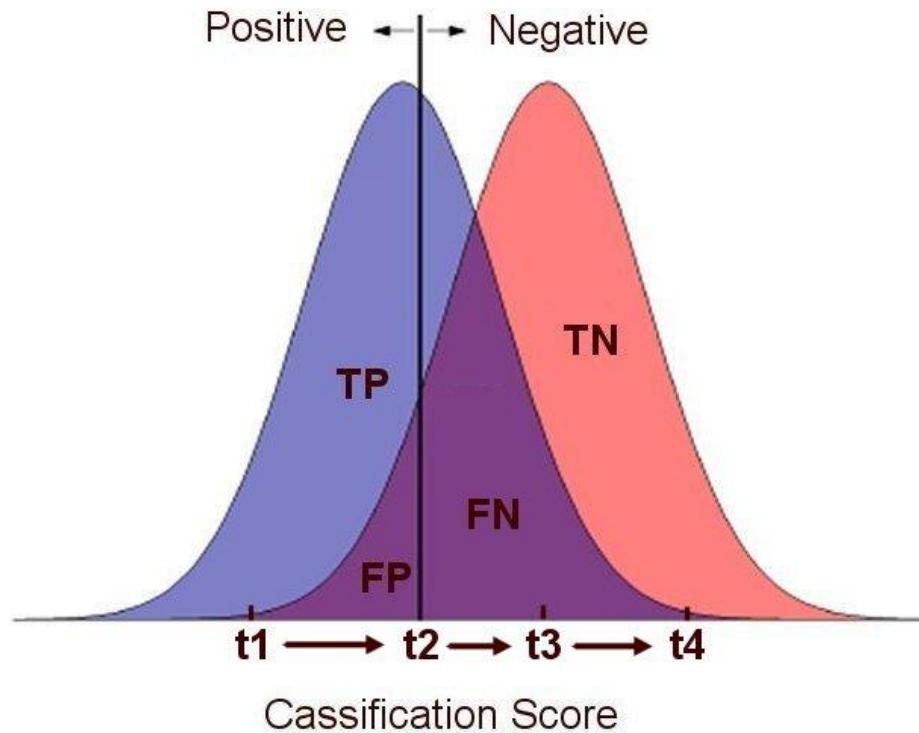


Evaluation of Classification Models – ROC curves

The **ROC curve** is a graphical plot reporting on the x-axis the **complement of specificity** ($1 - Sp_g$) and on the y-axis the **sensitivity** (Sn_g) resulting from a progressive increment of the cut-off (threshold) value.

Note that conventionally the ROC curves are represented in the context where low values of the biomarker (test) correspond to the pathological state (positive test), while high values indicate an healthy condition (negative test).

Under these circumstances, the ROC curve starts in the lower left corner of the graph with low sensitivity and high specificity, and finishes in the upper right corner with high sensitivity and low specificity. It is used to evaluate the best classification score, i.e., the cut-off values yielding the best compromise between sensitivity and specificity.



$$Sn_g = \frac{TP}{TP + FN}$$

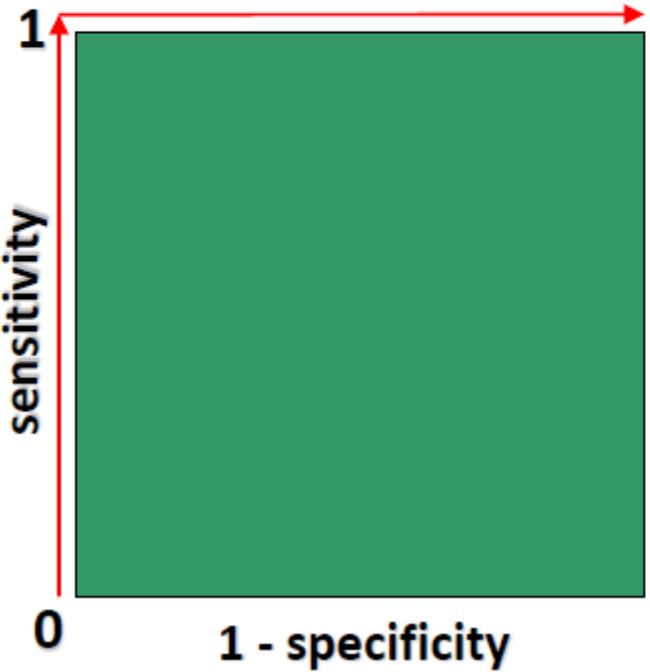
$$Sp_g = \frac{TN}{TN + FP}$$

$$1 - Sp_g = \frac{FP}{TN + FP}$$

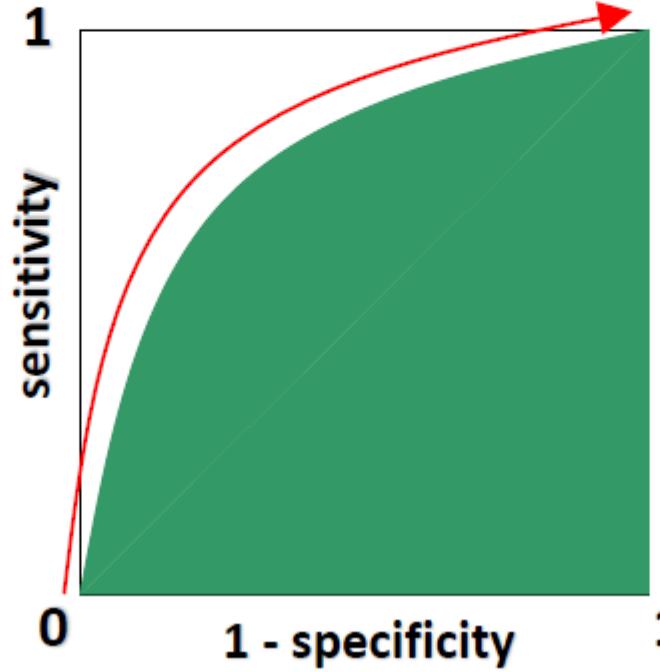
Evaluation of Classification Models – ROC curves

For each analyzed class we get a ROC curves, describing the class discrimination on the basis of specificity and sensitivity values

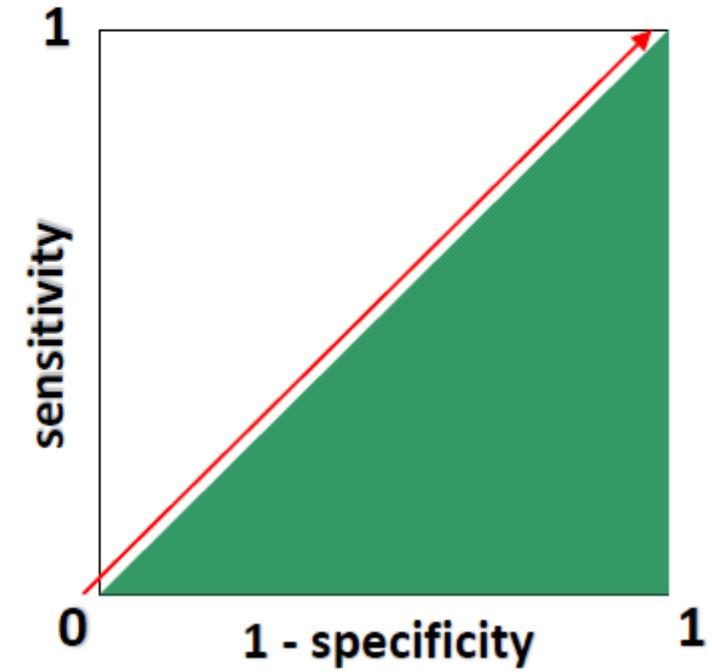
Perfect discrimination



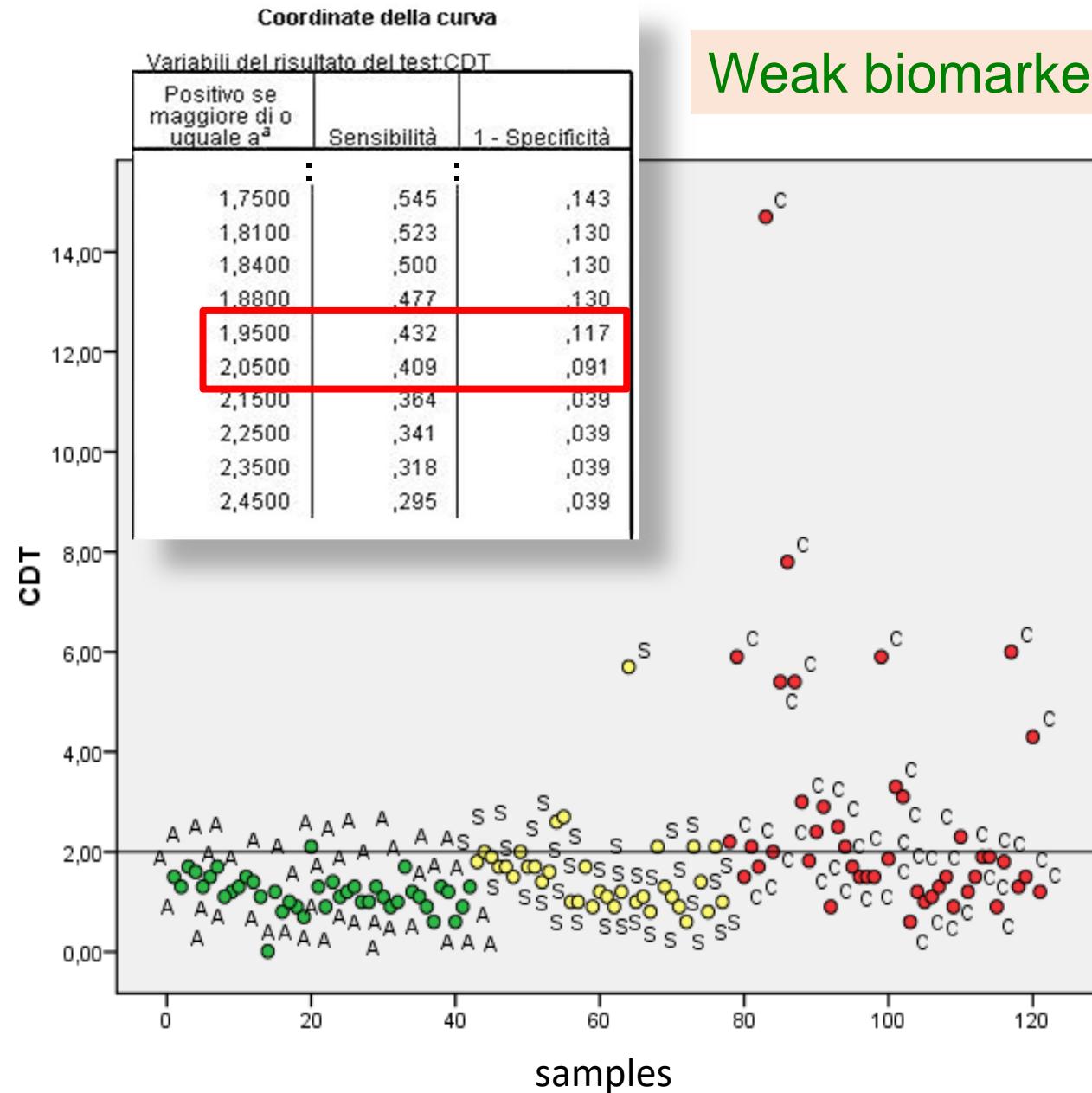
Medium discrimination



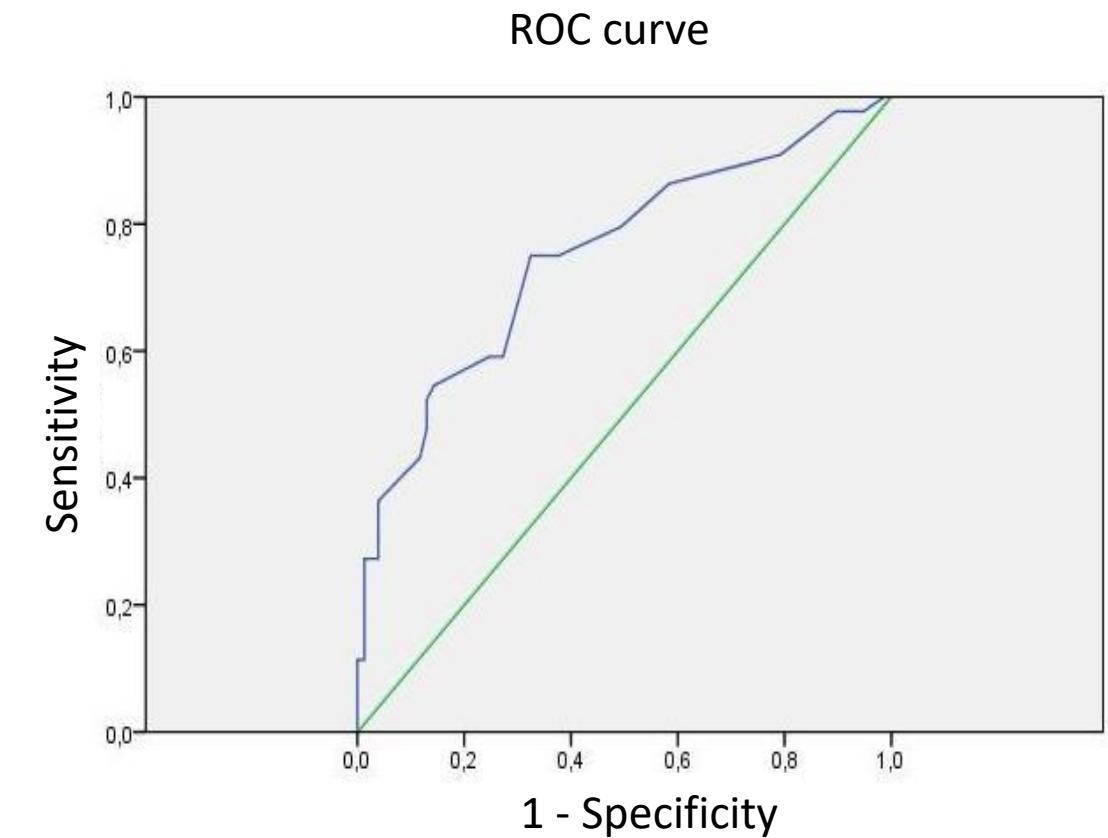
Bad discrimination



Evaluation of Classification Models – ROC curves



Weak biomarker: CDT to assess chronic alcohol abuse



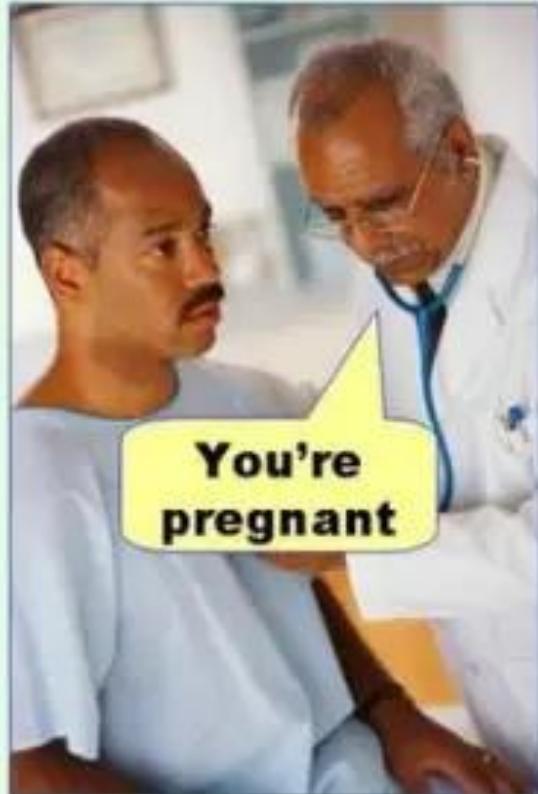
A = abstinentes

S = social drinkers

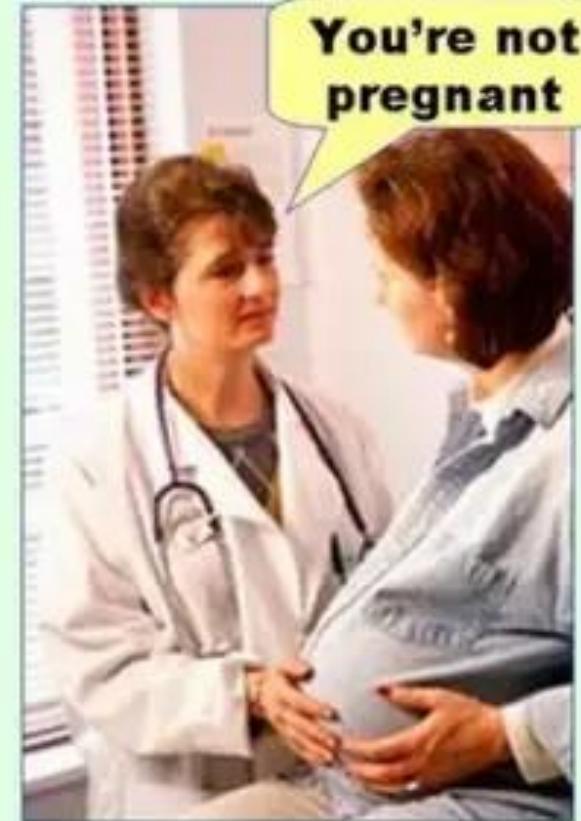
C= chronic alcohol abusers

It's a trade-off world...

Type I error
(false positive)

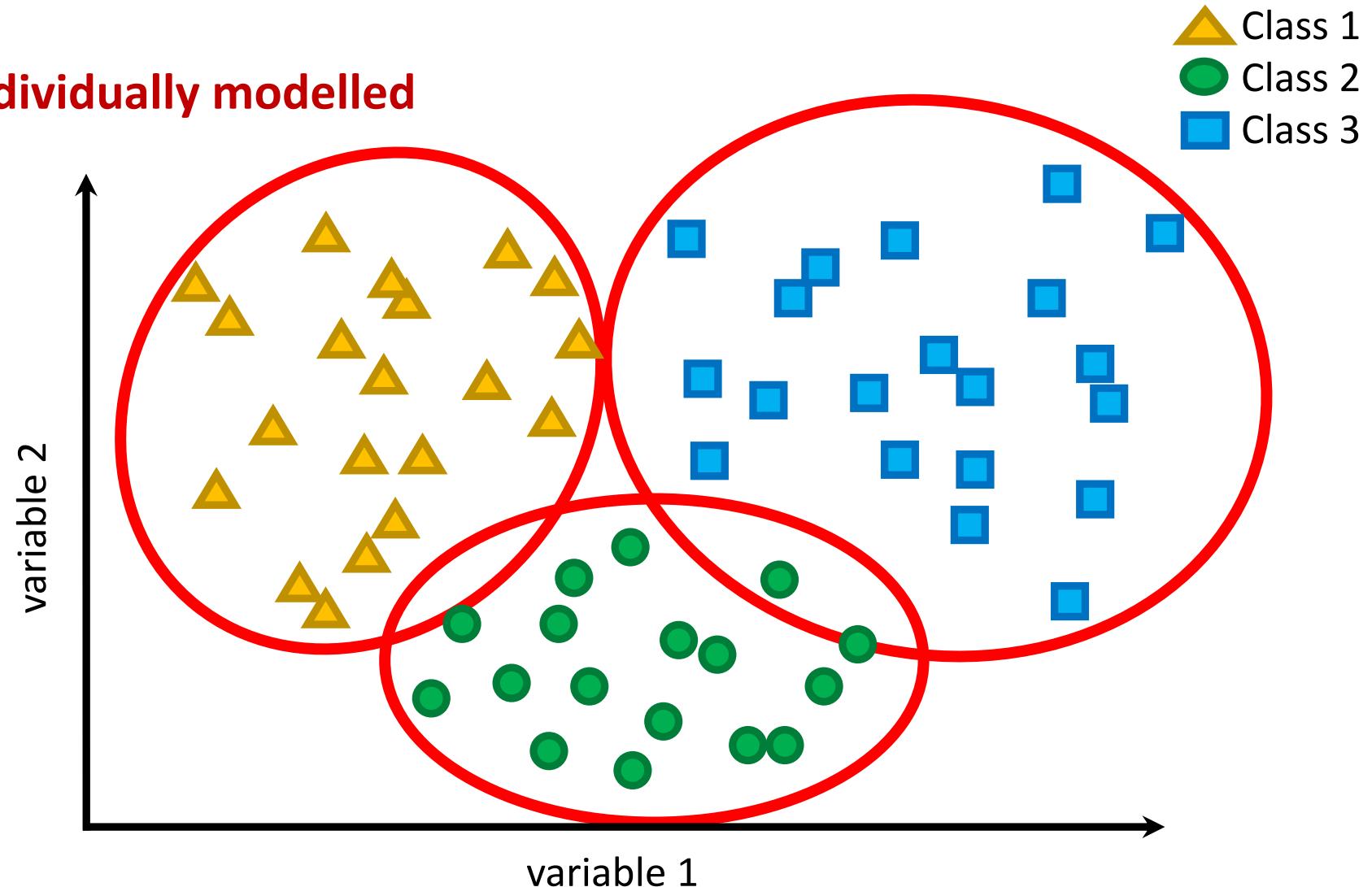


Type II error
(false negative)



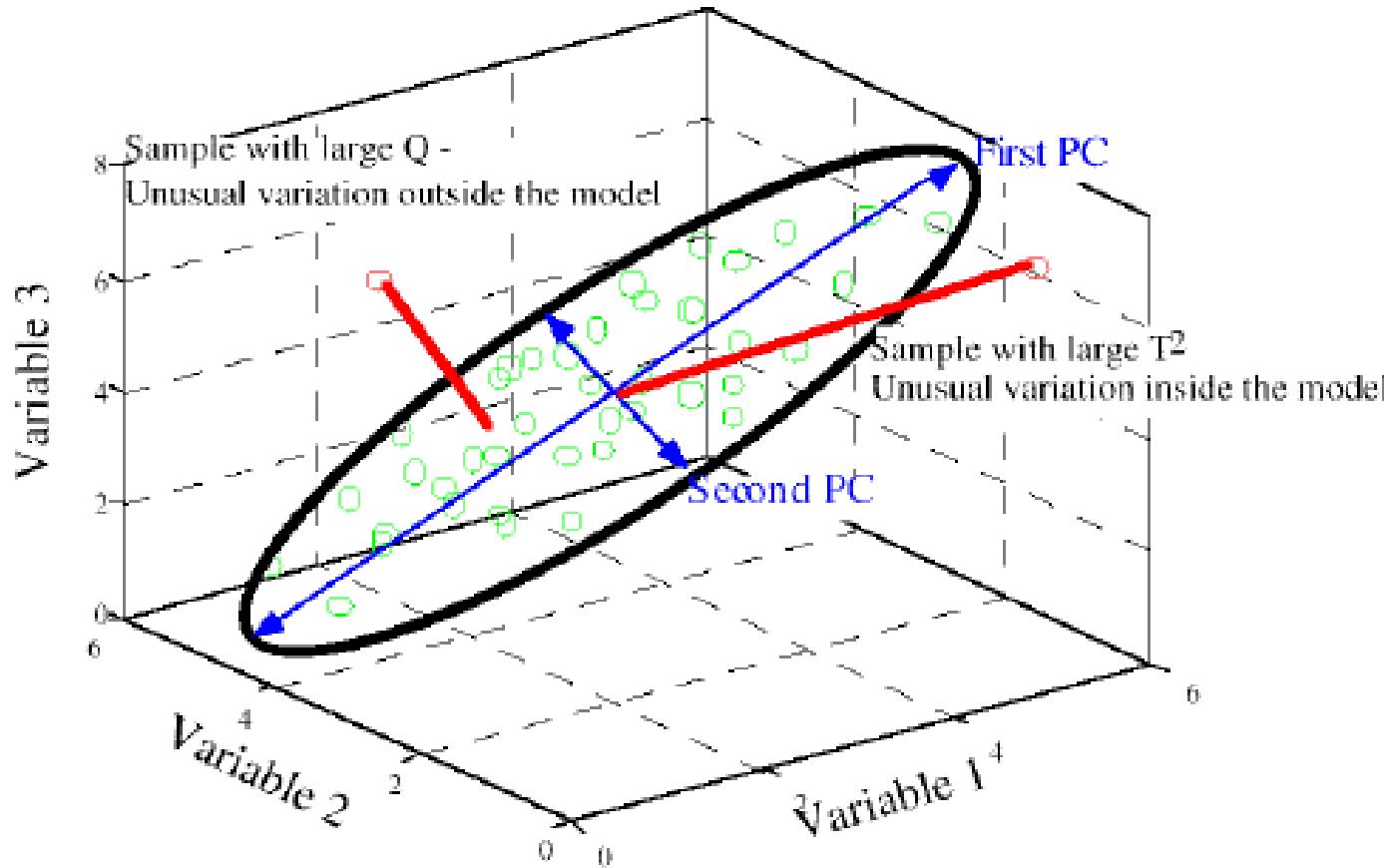
SIMCA – soft-independent model of class analogy (class modeling)

→ each class is **individually modelled**



SIMCA – soft-independent model of class analogy (class modeling)

Residuals: Q residuals and Hotelling's T²



Distance from plane/hyperspace of the model – variables/loadings

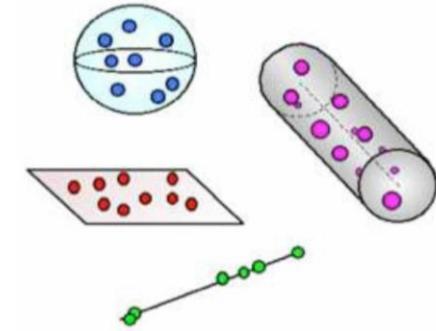
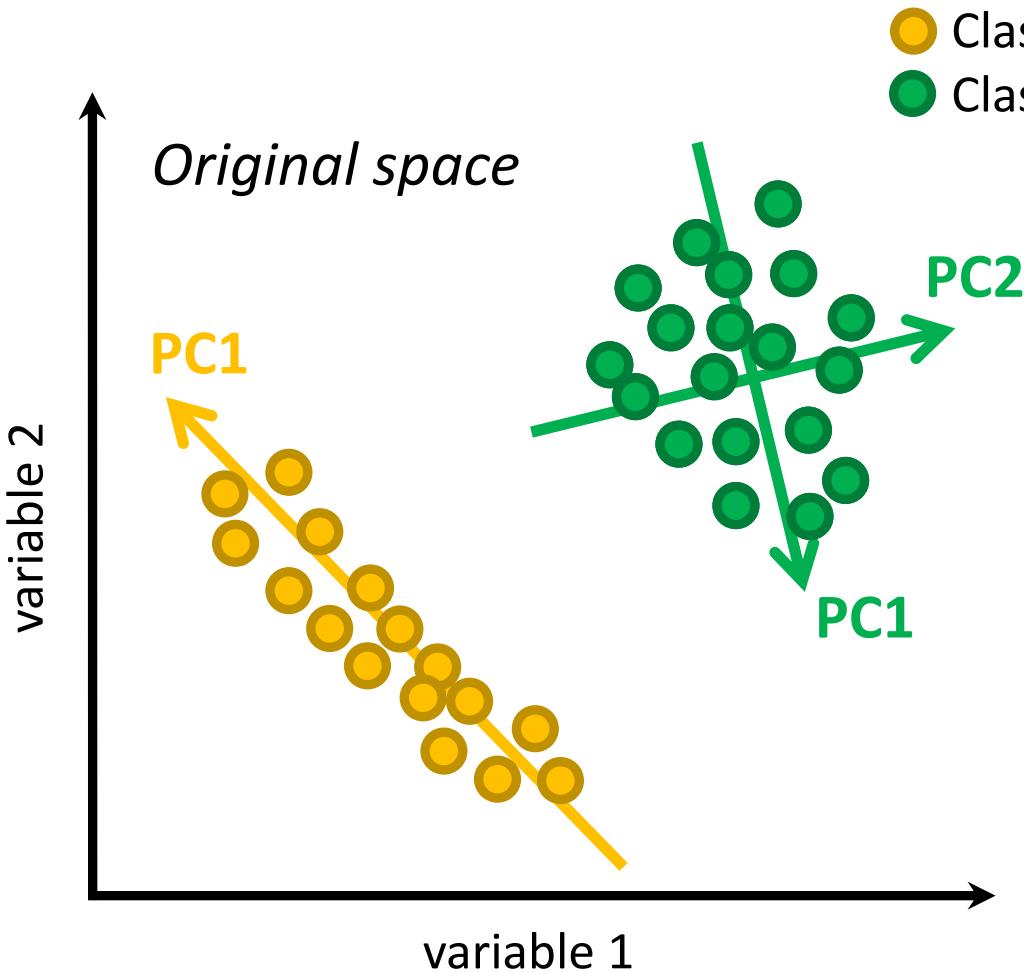
$$Q_i = e_i e_i^T = x_i (I - P_k P_k^T) x_i^T$$

Distance from the center of the model (centroid) – campioni/scores

$$T_i^2 = t_i \lambda^{-1} t_i^T = x_i P_k \lambda^{-1} P_k^T x_i^T$$

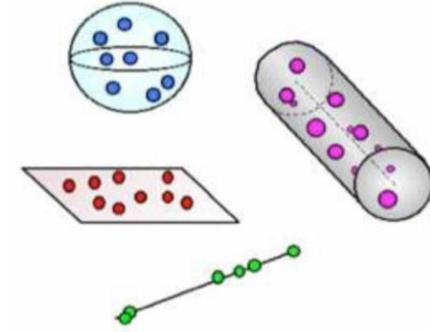
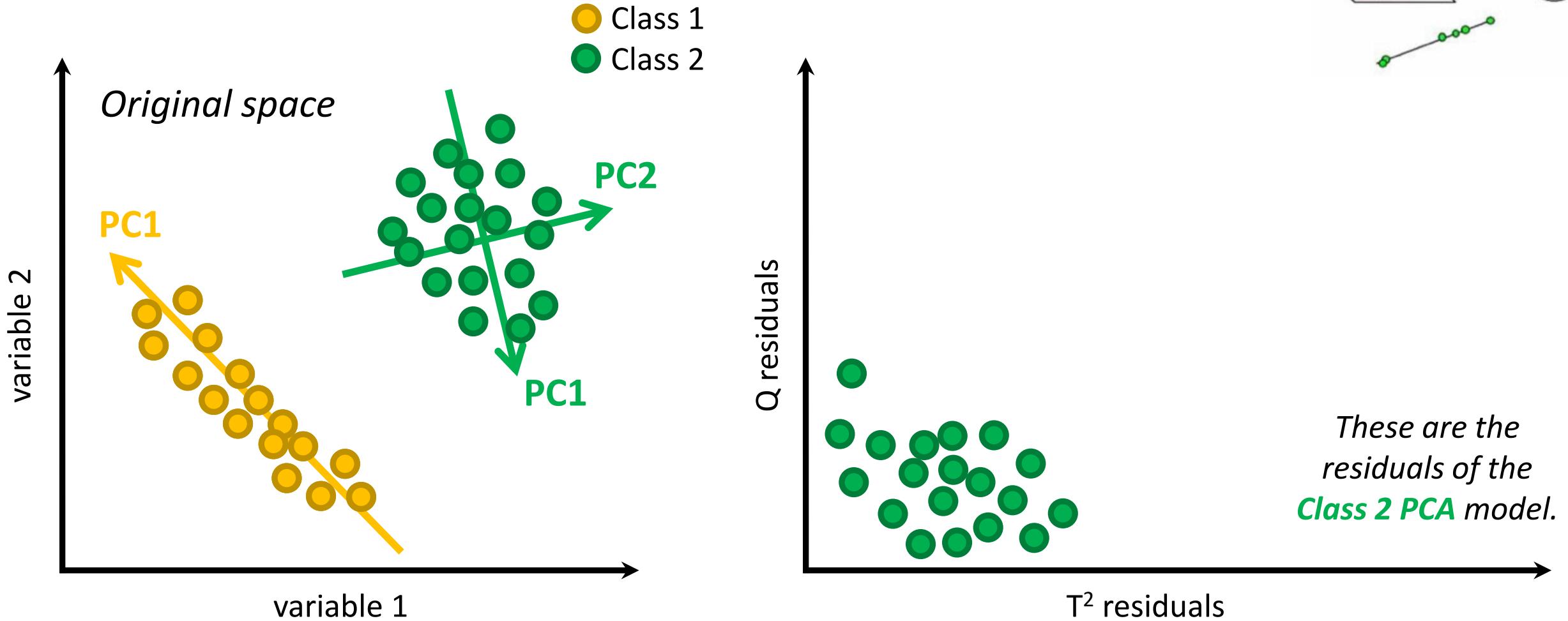
SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)



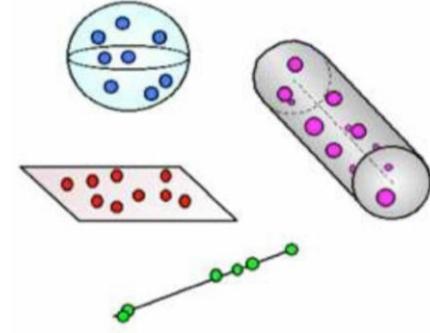
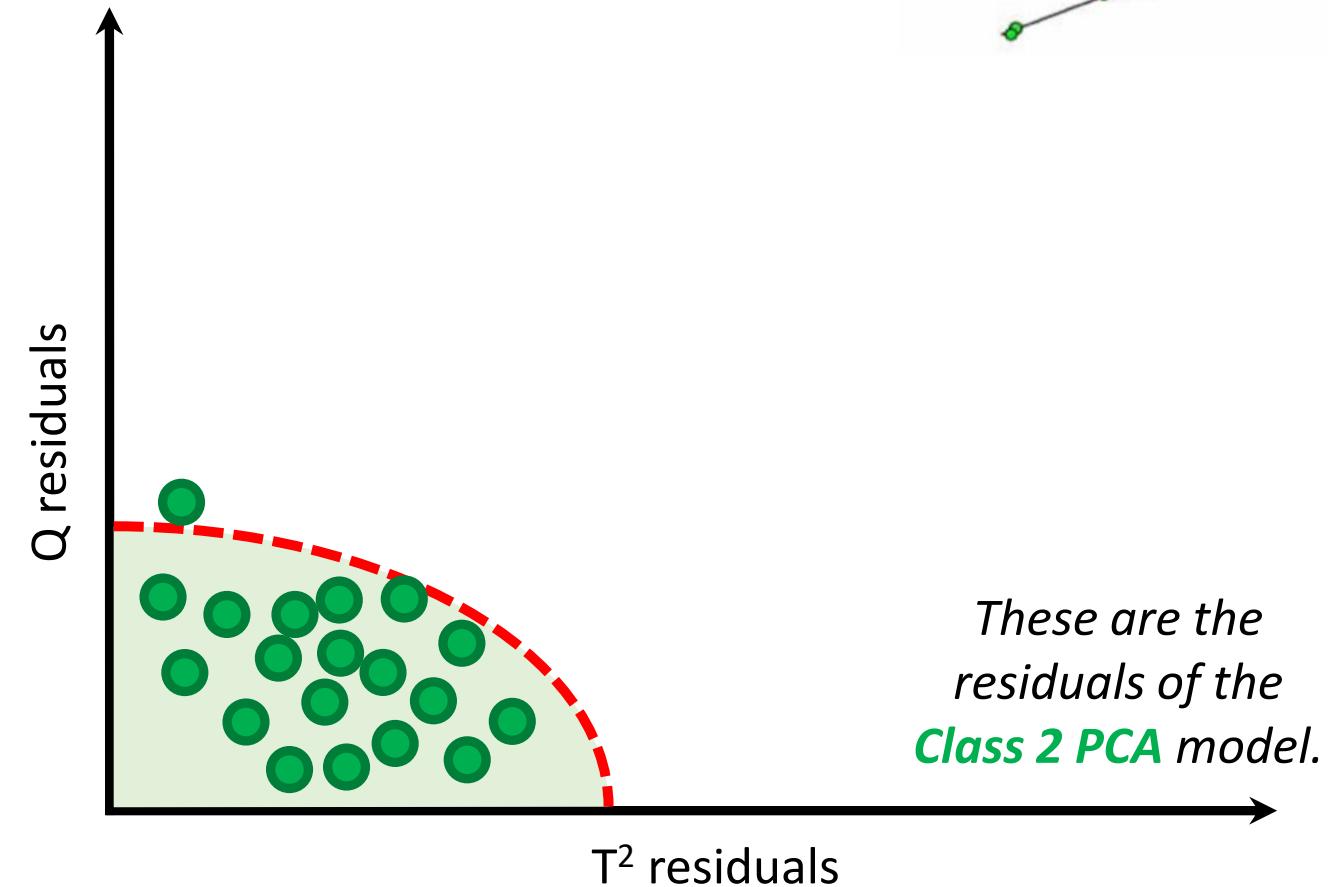
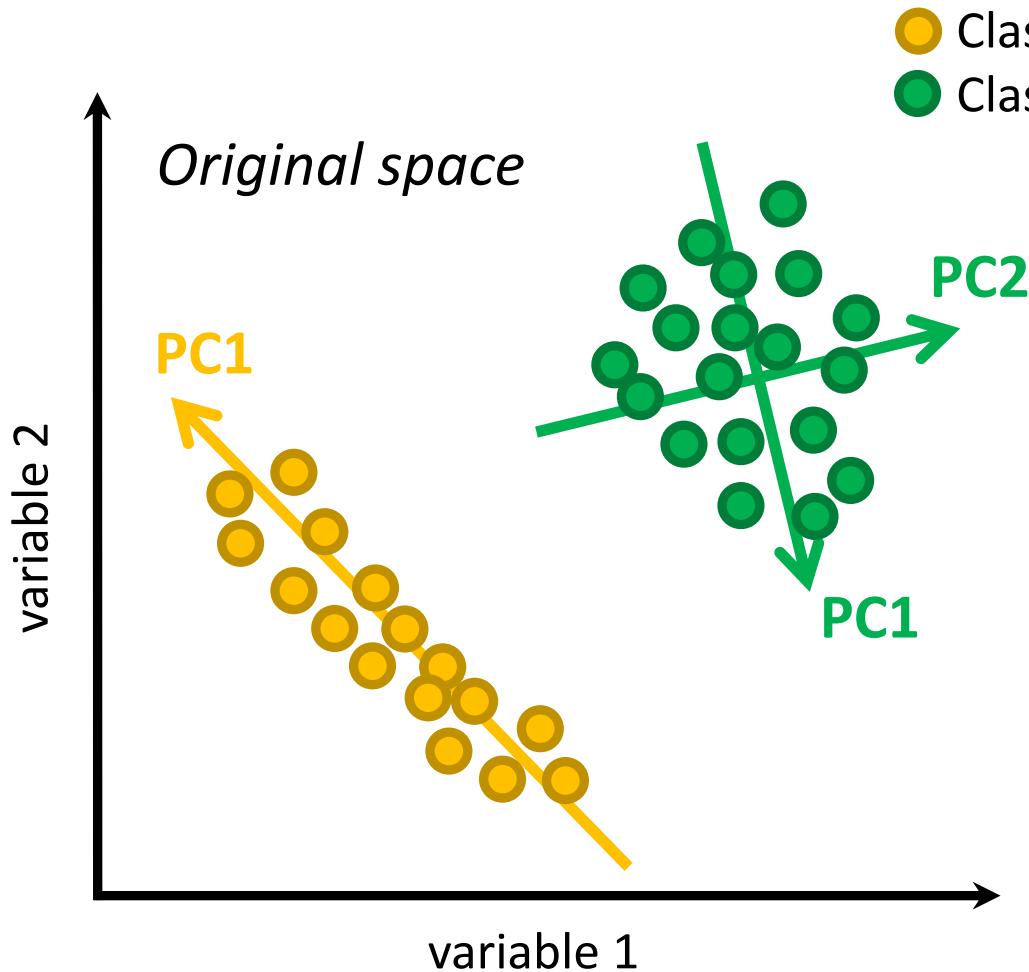
SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)



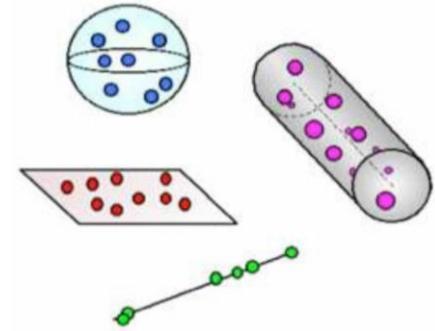
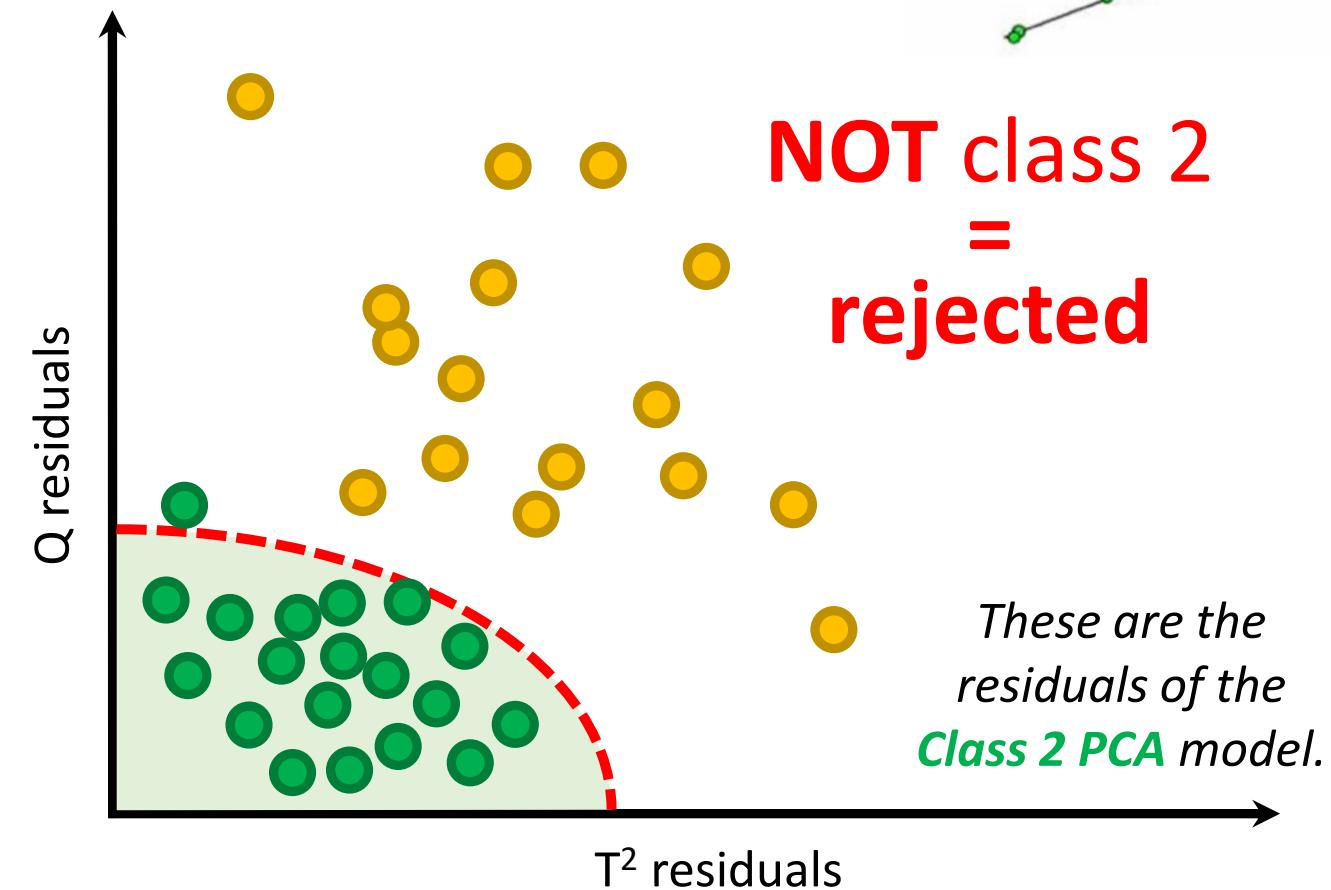
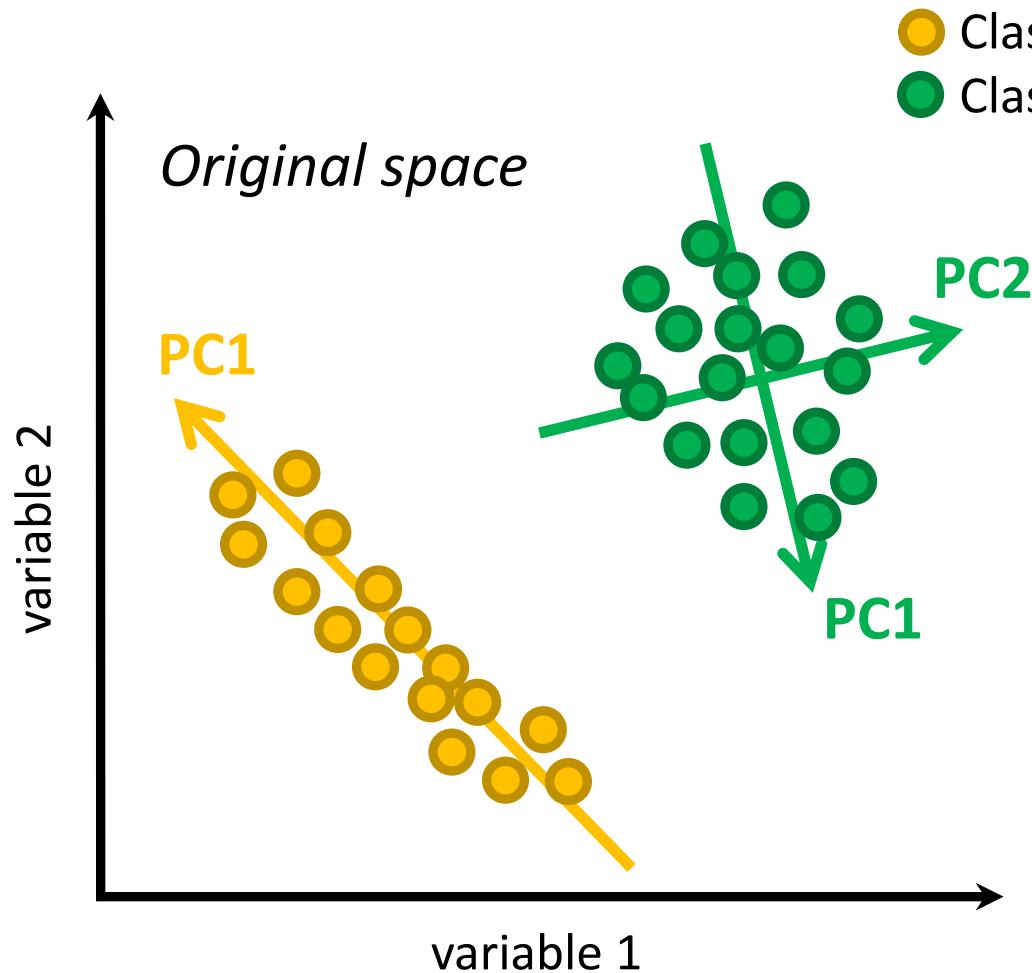
SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)



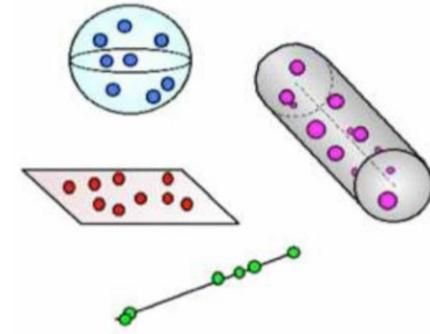
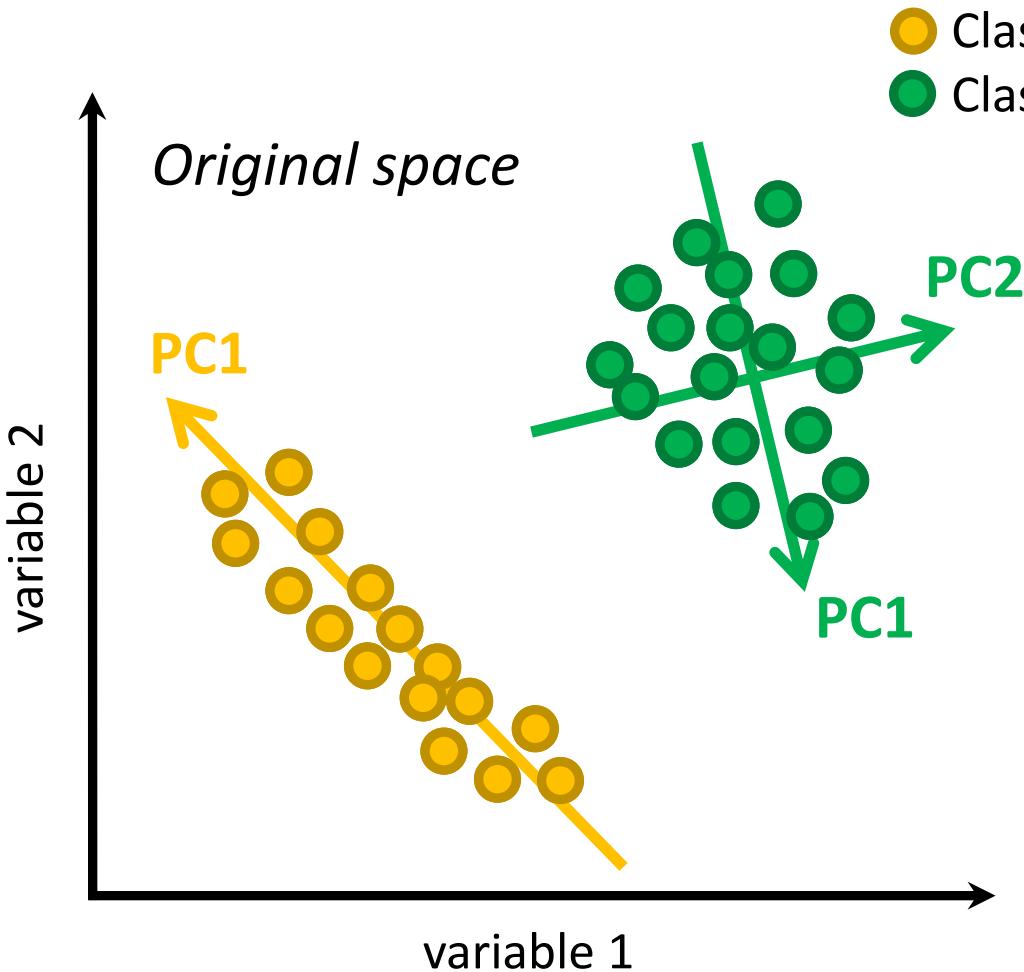
SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)



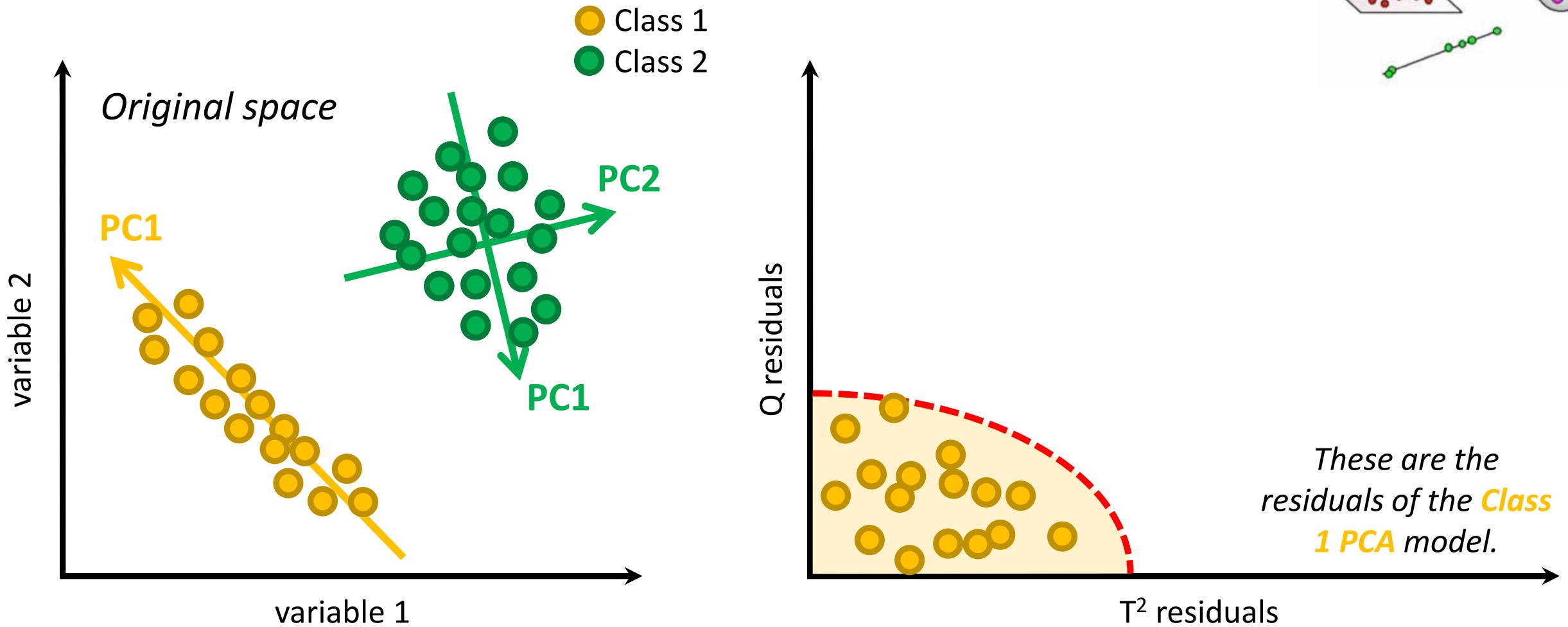
SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)



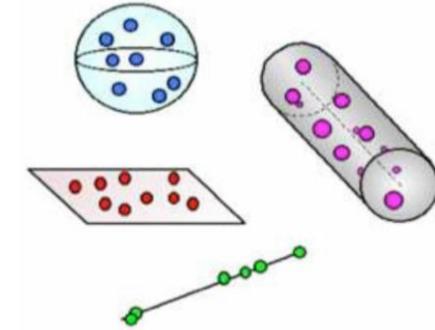
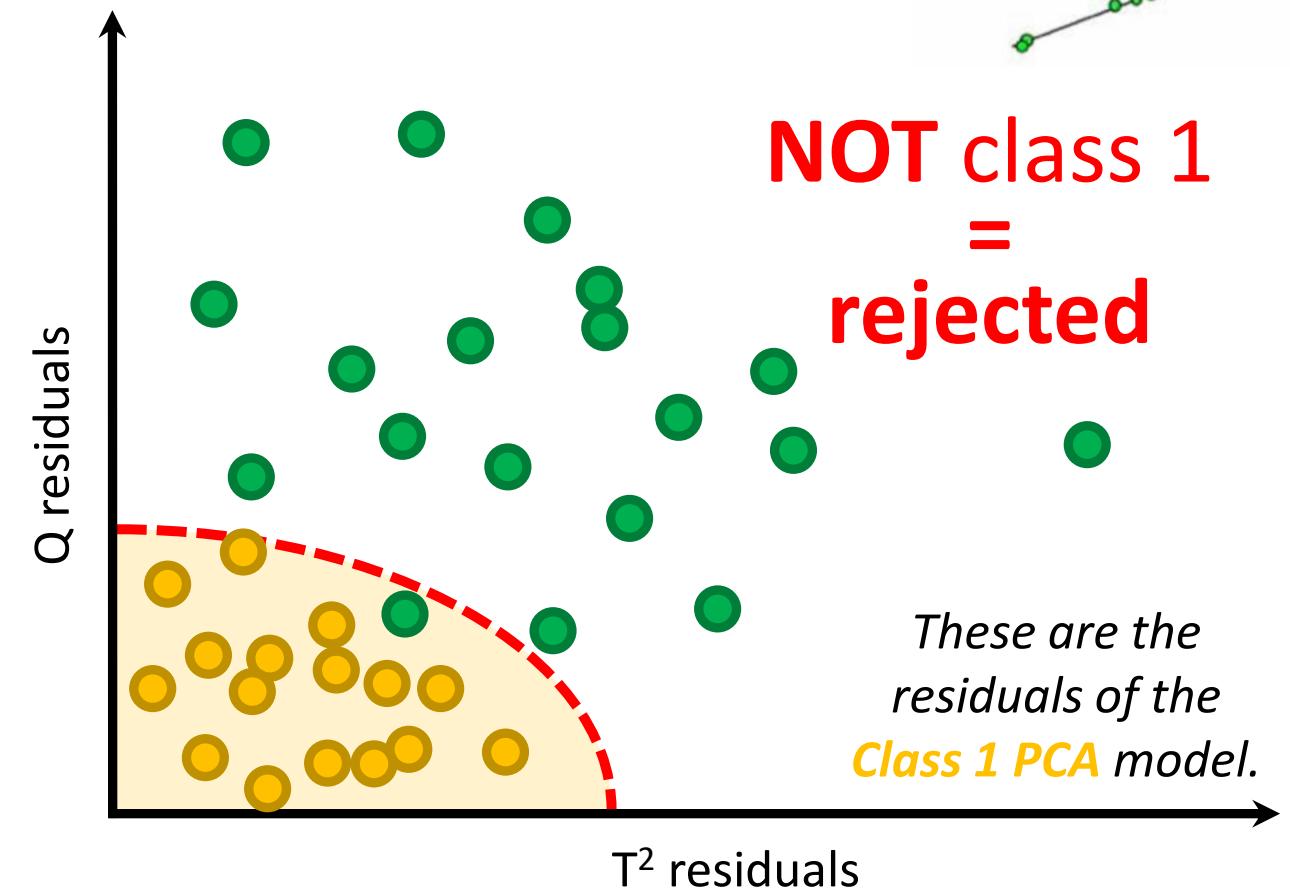
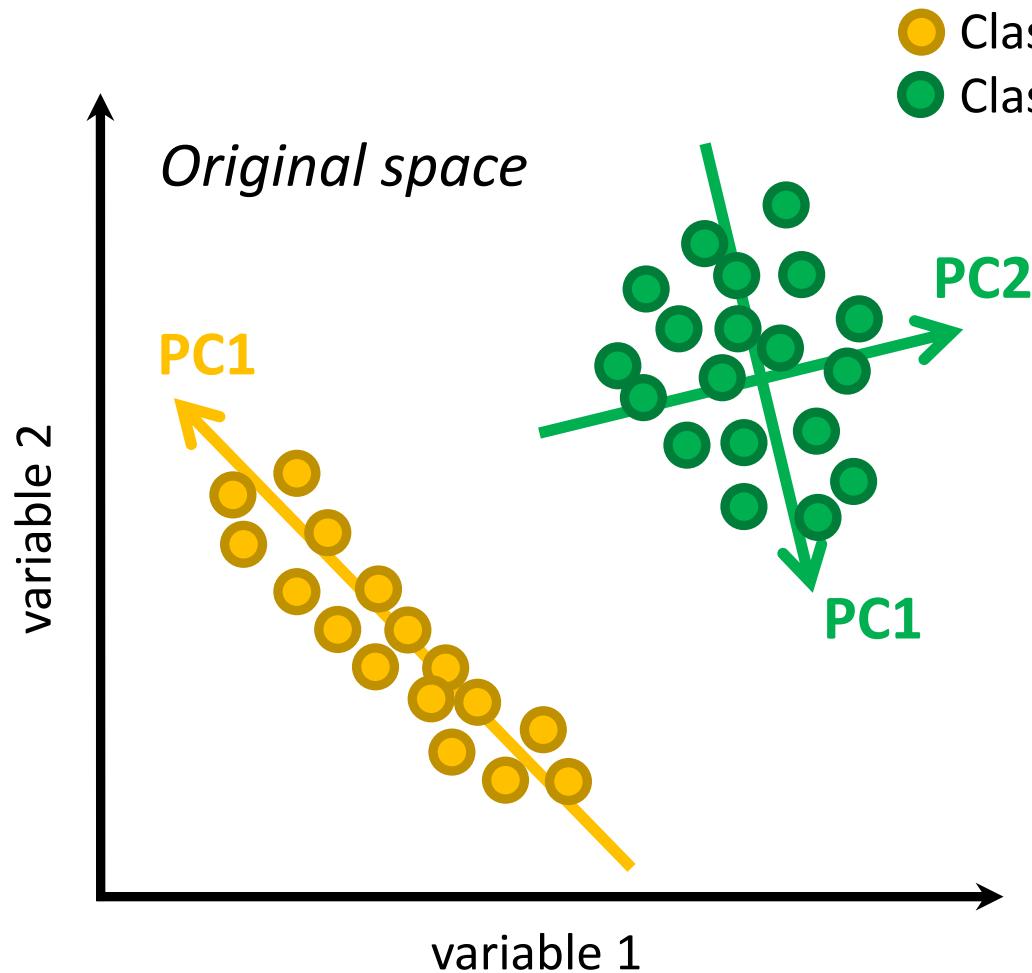
SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)



SIMCA – soft-independent model of class analogy (class modeling)

- each class is **modelled independently** by PCA
- residuals are used for classification (--- inclusion/exclusion)

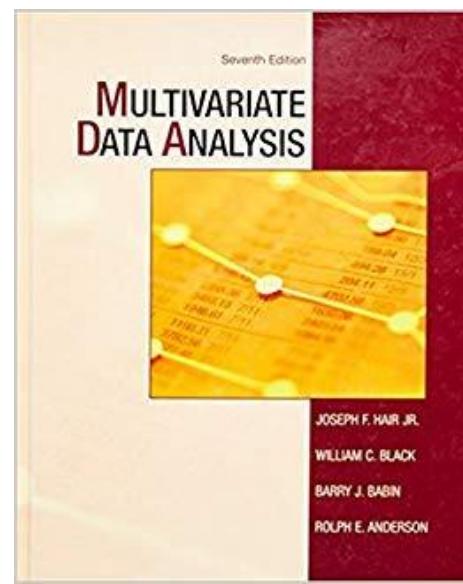


Classification models

The classification and modeling methods can be divided into 3 classes.

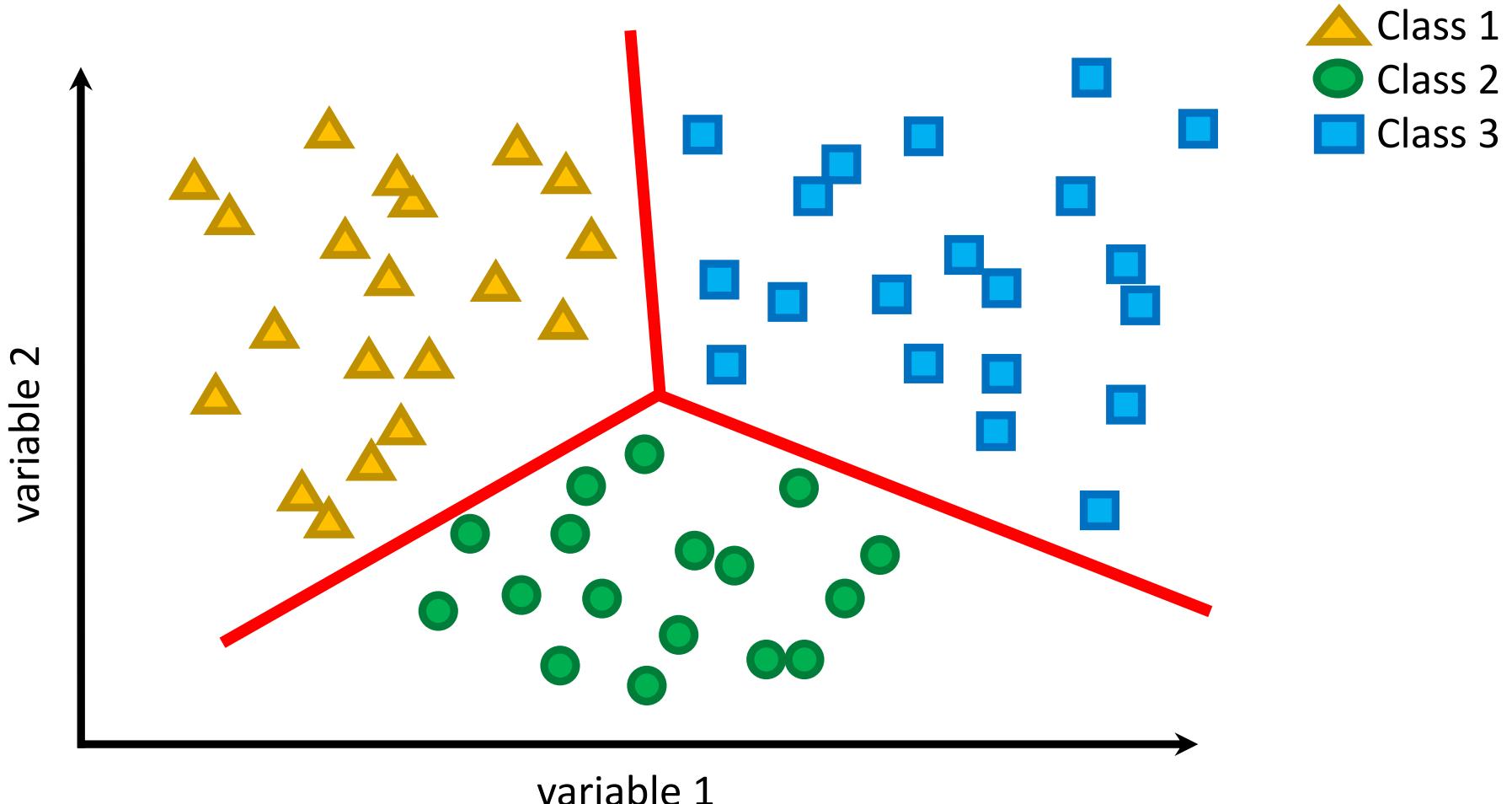
There are such methods:

- **DISTANCE-BASED** (possibility to use different metrics / distances);
- **PROBABILITY-BASED** (based on the estimation of probability distributions);
- **EXPERIENCE-BASED** (iterative, with trial-and-error procedures).



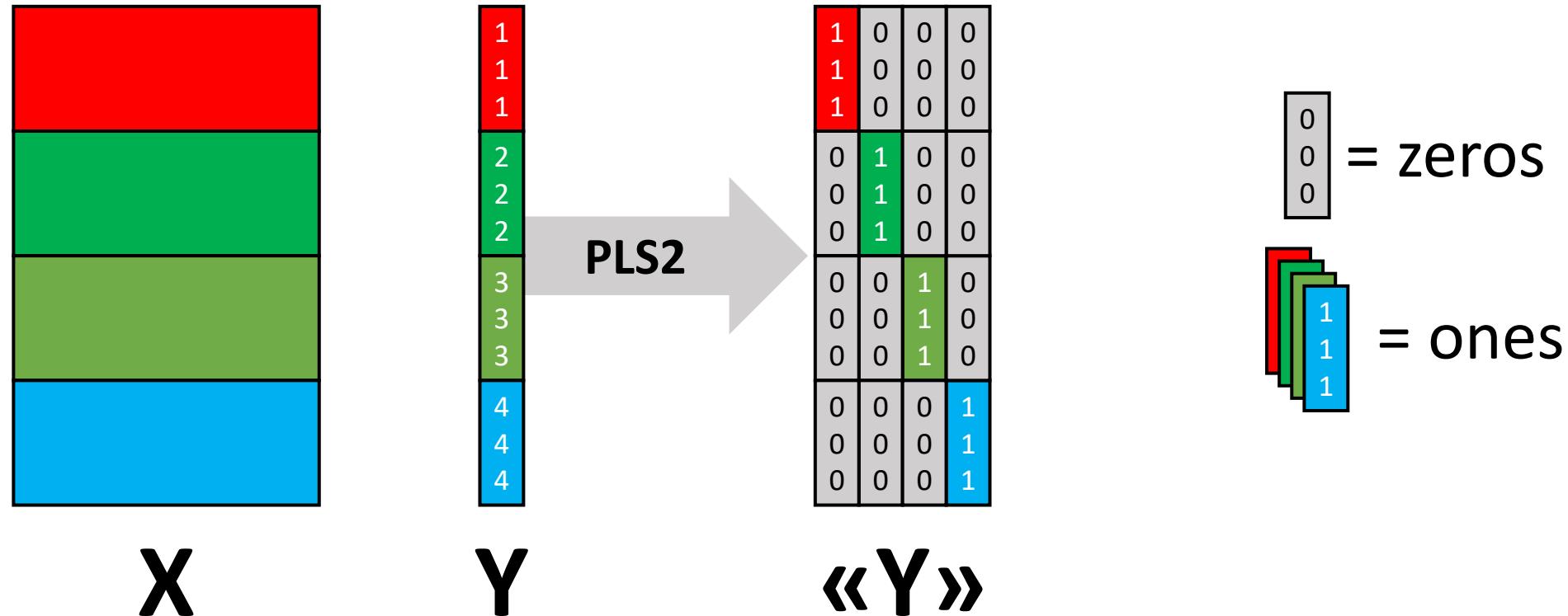
Discriminant analysis (classification)

→ **boundaries** between classes are searched



Discriminant analysis (classification)

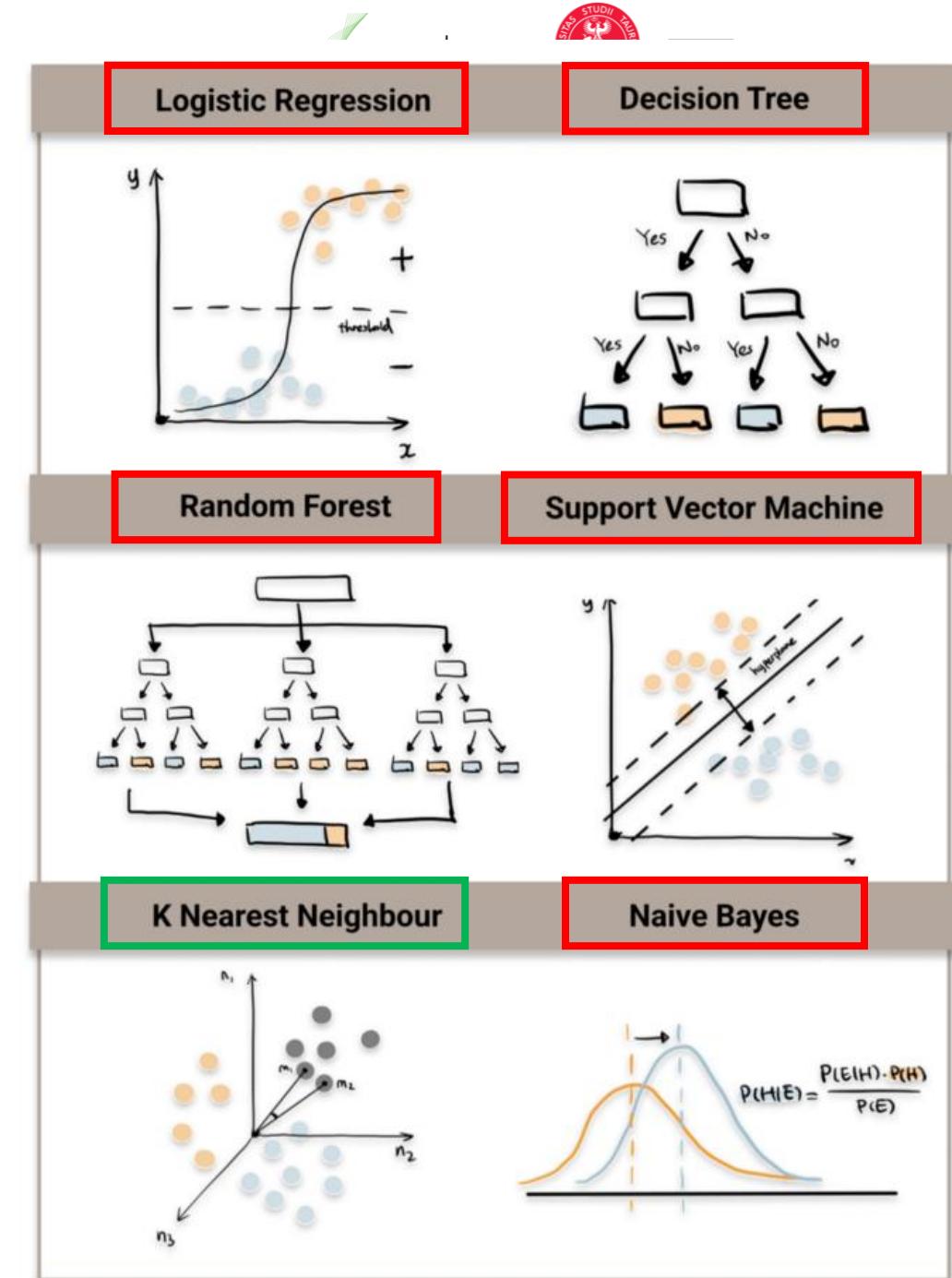
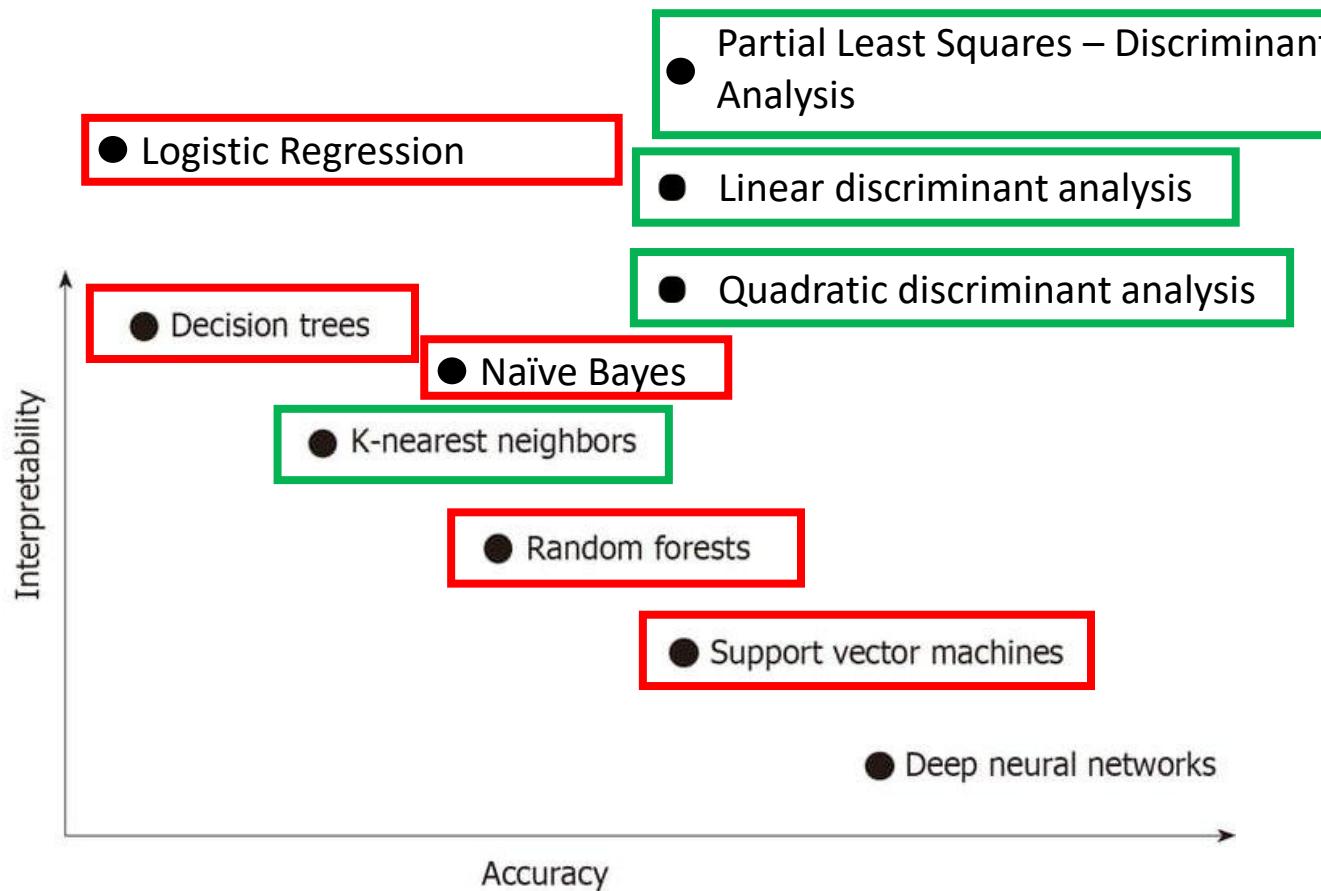
Y is a “dummy matrix” which encodes the class information (i.e. a qualitative property).



Classification Methods

Classification & modeling methods can be divided into 3 classes

- **DISTANCE-BASED**
- **PROBABILITY-BASED**
- **EXPERIENCE-BASED (iterative, trial & error)**

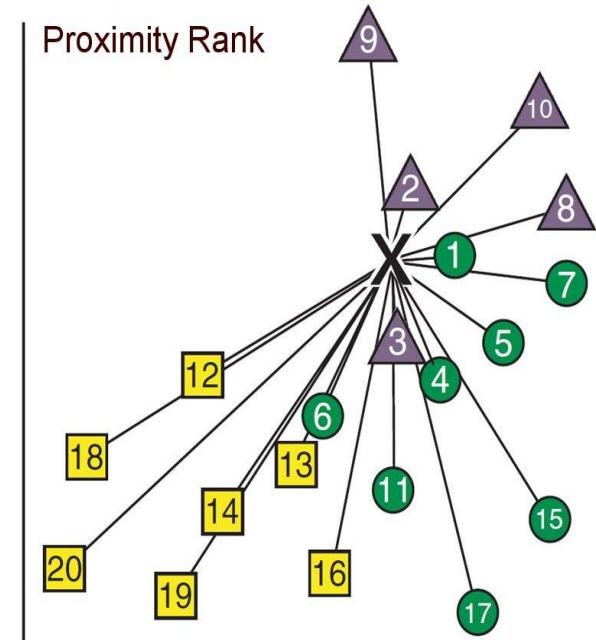
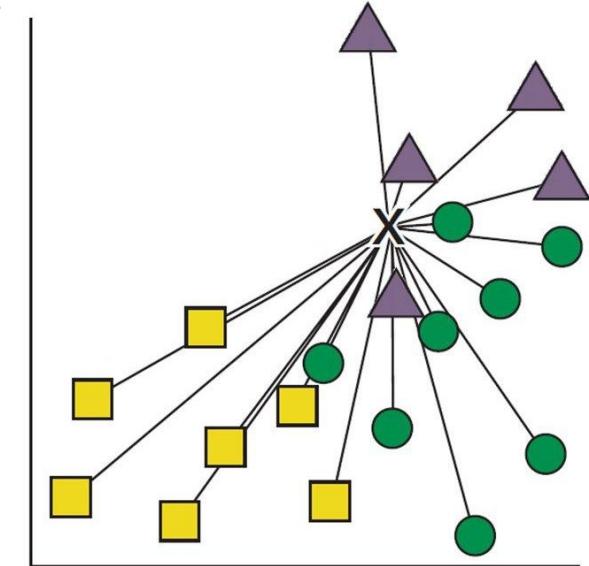


Classification Methods – K-NN (K-Nearest Neighbors)

X = object to be classified

The **K-NN method** is a simple non-parametric classification method (i.e., it disregards the specific continuous variables' value & their distribution when a class is assigned) that uses the concept of analogy (closeness). K-NN is based on (I) the choice of a method to calculate the distance between couples of objects (generally, Euclidean) and (II) the selection of a K number of neighbors, that determine the class assignment applying the majority principle. The following steps should be respected:

- a) Data scaling
- b) Choice of the algorithm to calculate distances
- c) Selection of the K-number of neighbors that governs the class assignment
- d) Calculation of the Distance Matrix
- e) Identification of the K closest neighbors for each object
- f) Assignment of each object to the most represented class among the K closest neighbors
- g) In practice, the K parameter is varied so as to find the K-value yielding the best classification outcome
- h) Note: the class assignment is made by considering all the objects of the training set.



Classification Methods – K-NN (K-Nearest Neighbors)

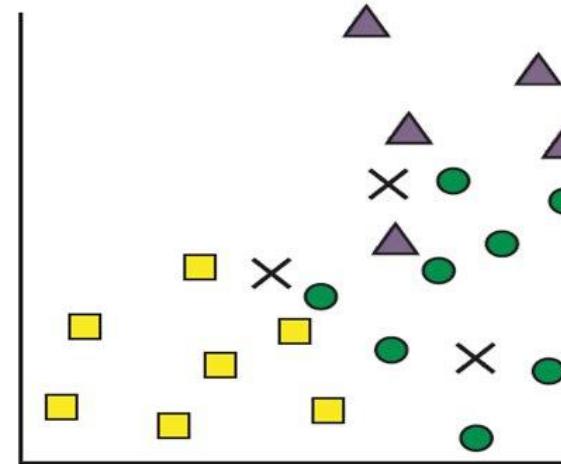
In the example reported alongside, 3 objects (x) have to be assigned to a class, using K-NN with $K = 1, 3$, and 5 , respectively.

For 2 objects out of 3, the class assignment changes depending on the chosen K-value.

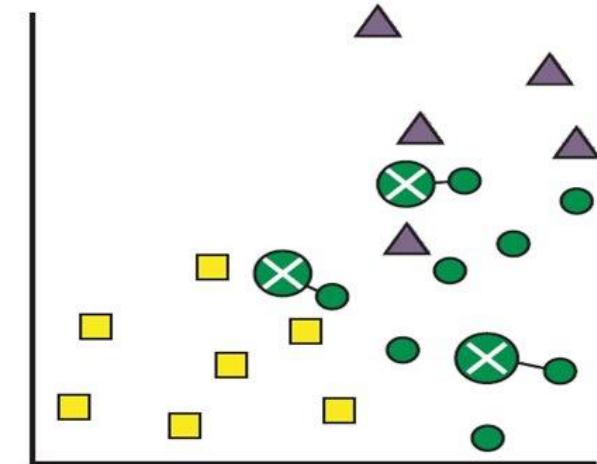
In a multidimensional space, the algorithm selected to calculate the distances also plays an important role.

Typically, the training set provides the basic data from which to calculate the distances of the new objects (either from the evaluation set or unknown). The evaluation set is used to test the different K-values using NER% or other performance test. Also the objects of the training set can be tested by an assignment-verification process, but each tested object should be included among the K-neighbors. Overfitting results are obtained, especially for low k-values (for $K=1$, NER% = 100).

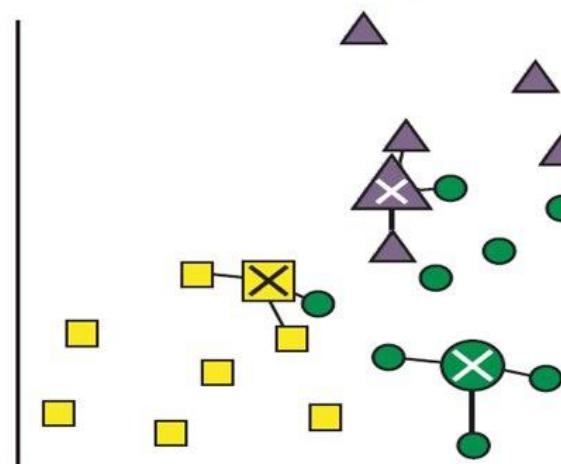
Before classification



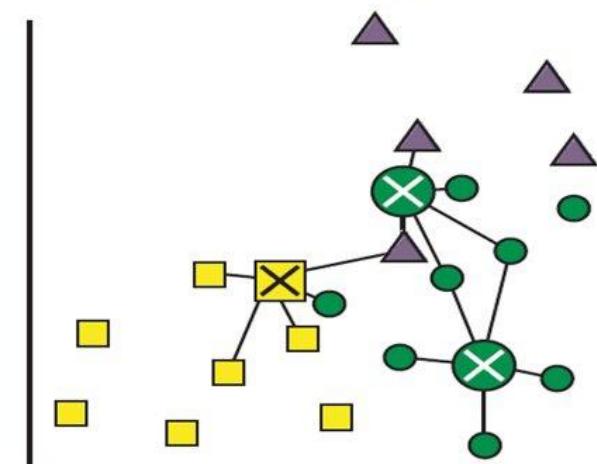
1-nearest neighbor



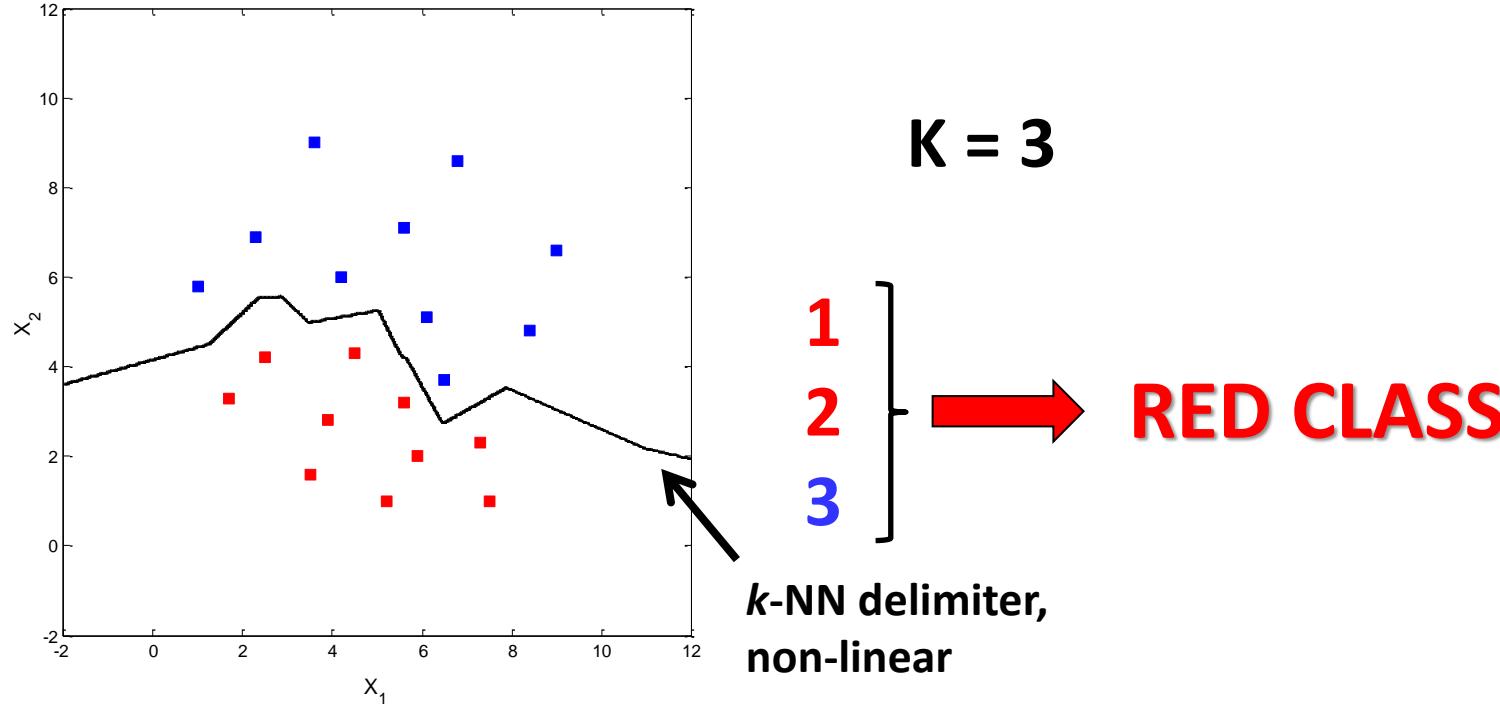
3-nearest neighbors



5-nearest neighbors

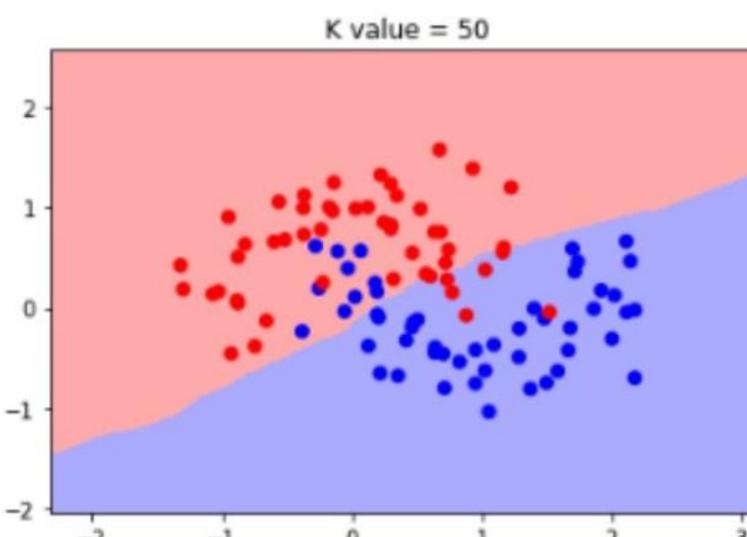
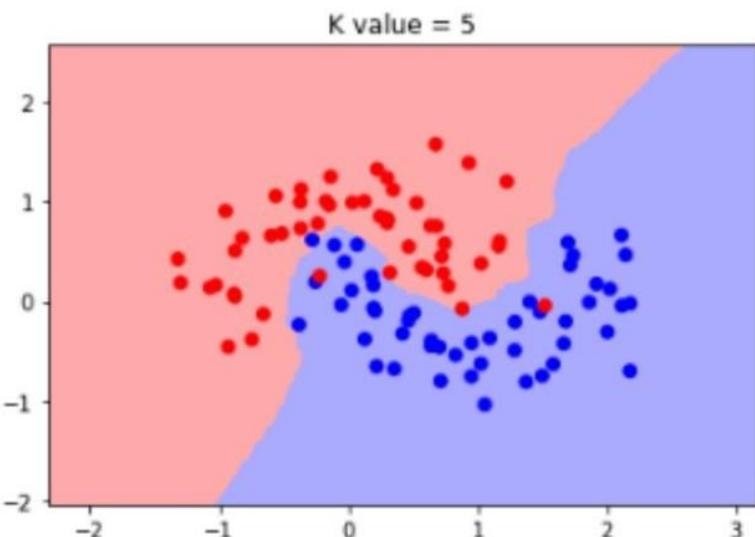
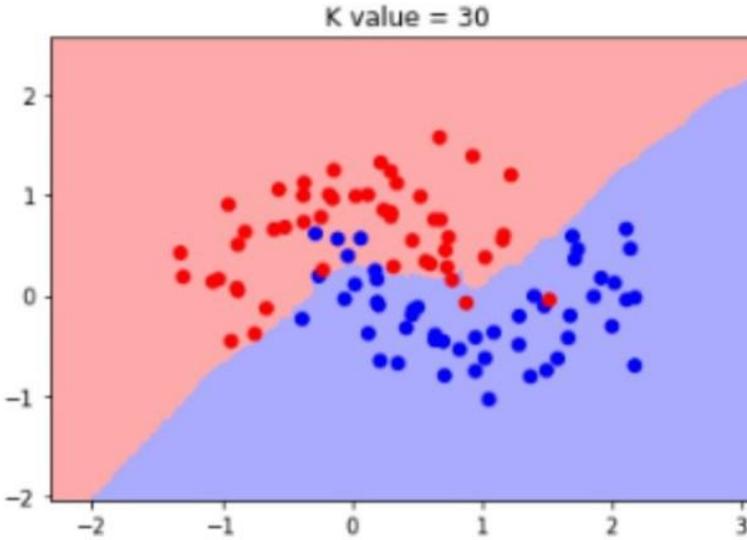
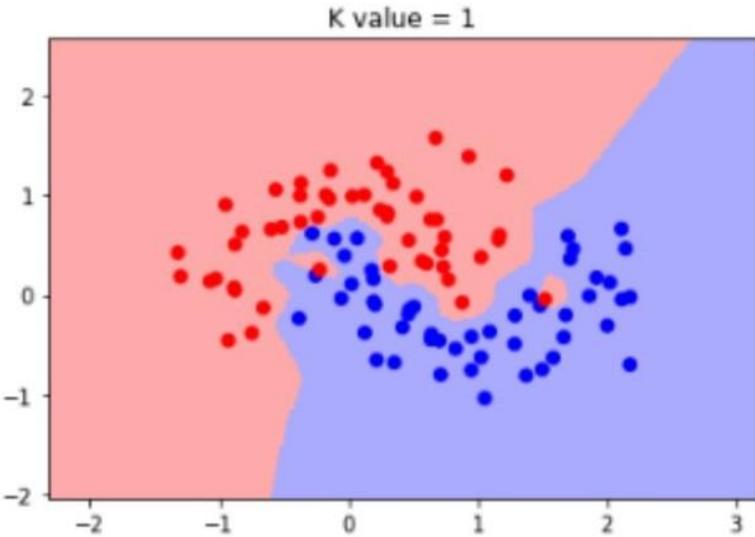


Classification Methods – K-NN (K-Nearest Neighbors)



In the bidimensional space represented in figure, a broken-line delimiter can be drawn. Each segment corresponds to a tie-line. For example, for $K = 3$, as in figure, the 1st and 2nd closest objects belong to different classes, while two other objects of opposite classes equally occupy the 3rd place (they are equally distant from the segment). Each segment ends corresponds to a change in at least one of the four objects under consideration. In a p-multidimensional space the delimiter is defined in $(p-1)$ dimensions and is impossible to visualize.

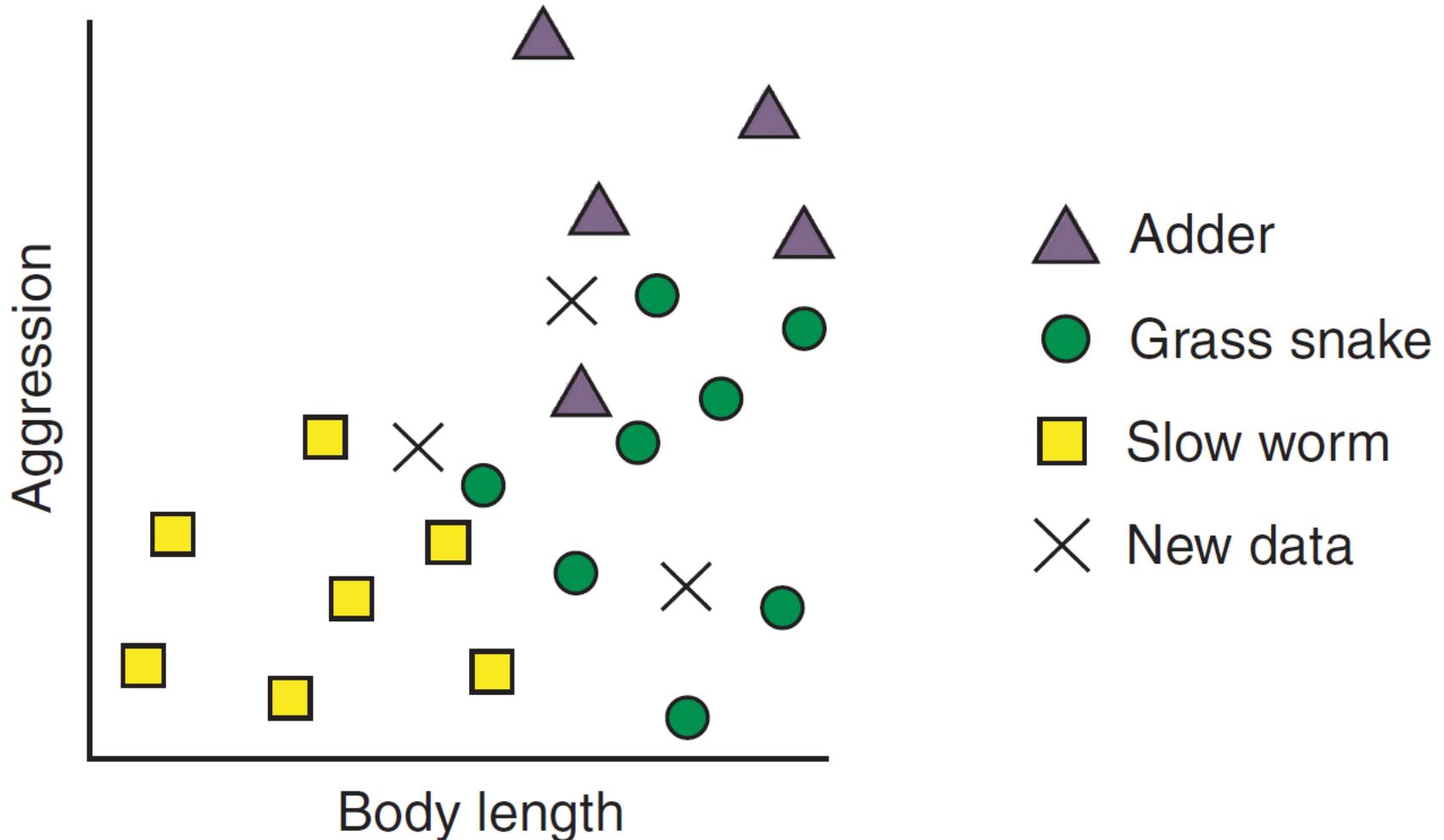
Classification Methods – K-NN



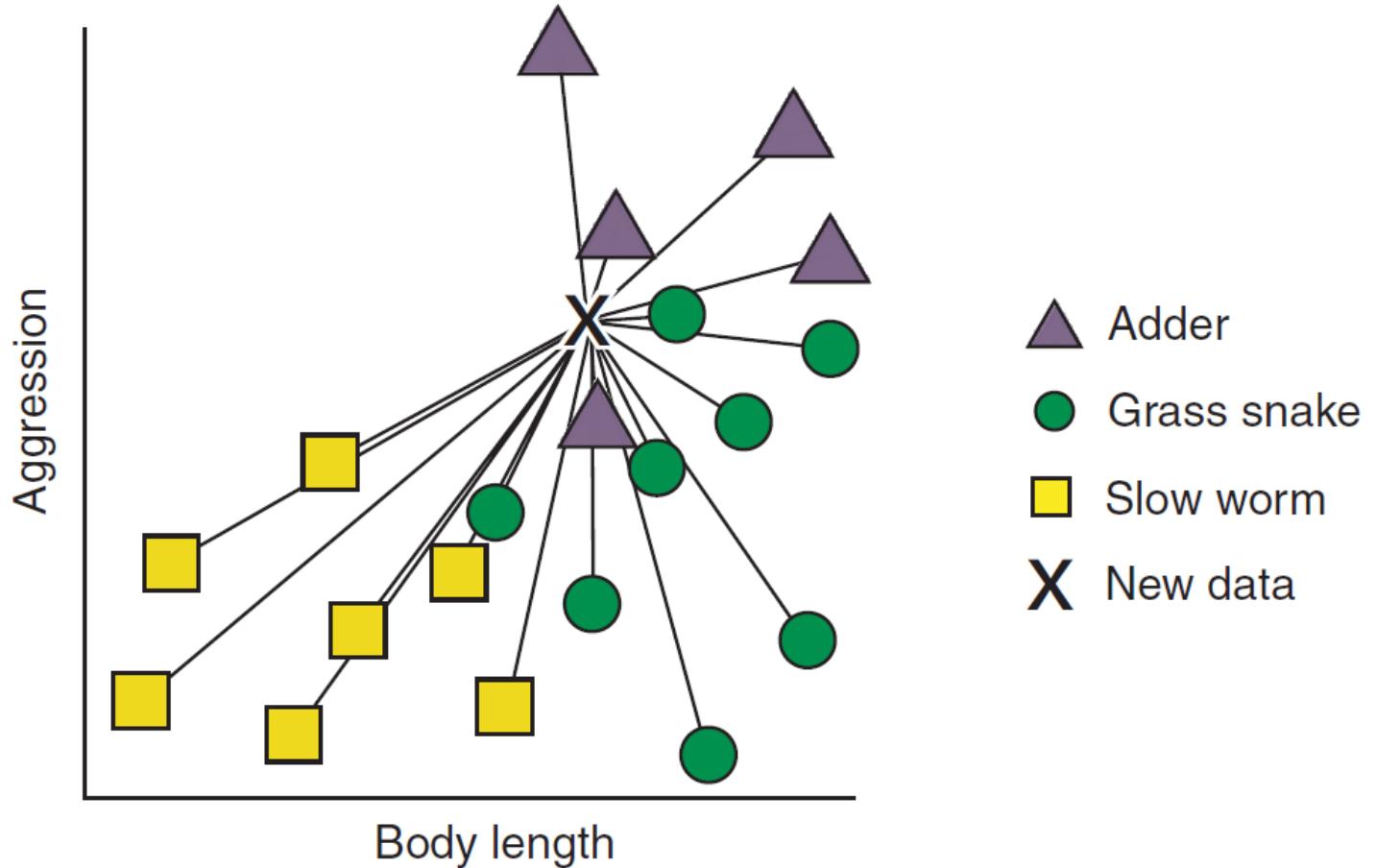
For low K-values, the delimiter is extremely jagged, which generally indicates overfitting. At high K-values, the delimiter tends to flatten down, and sometimes does not provide adequate separation of the objects distribution in the variables' space (underfitting, as in the figure for K=30 and K=50). In the example shown in figure, K=5 represents a realistic compromise between overfitting and underfitting.

A class modelling version of K-NN exists (ACMT) and uses as class delimiters either the mean or the median of the K-nearest objects of the same class.

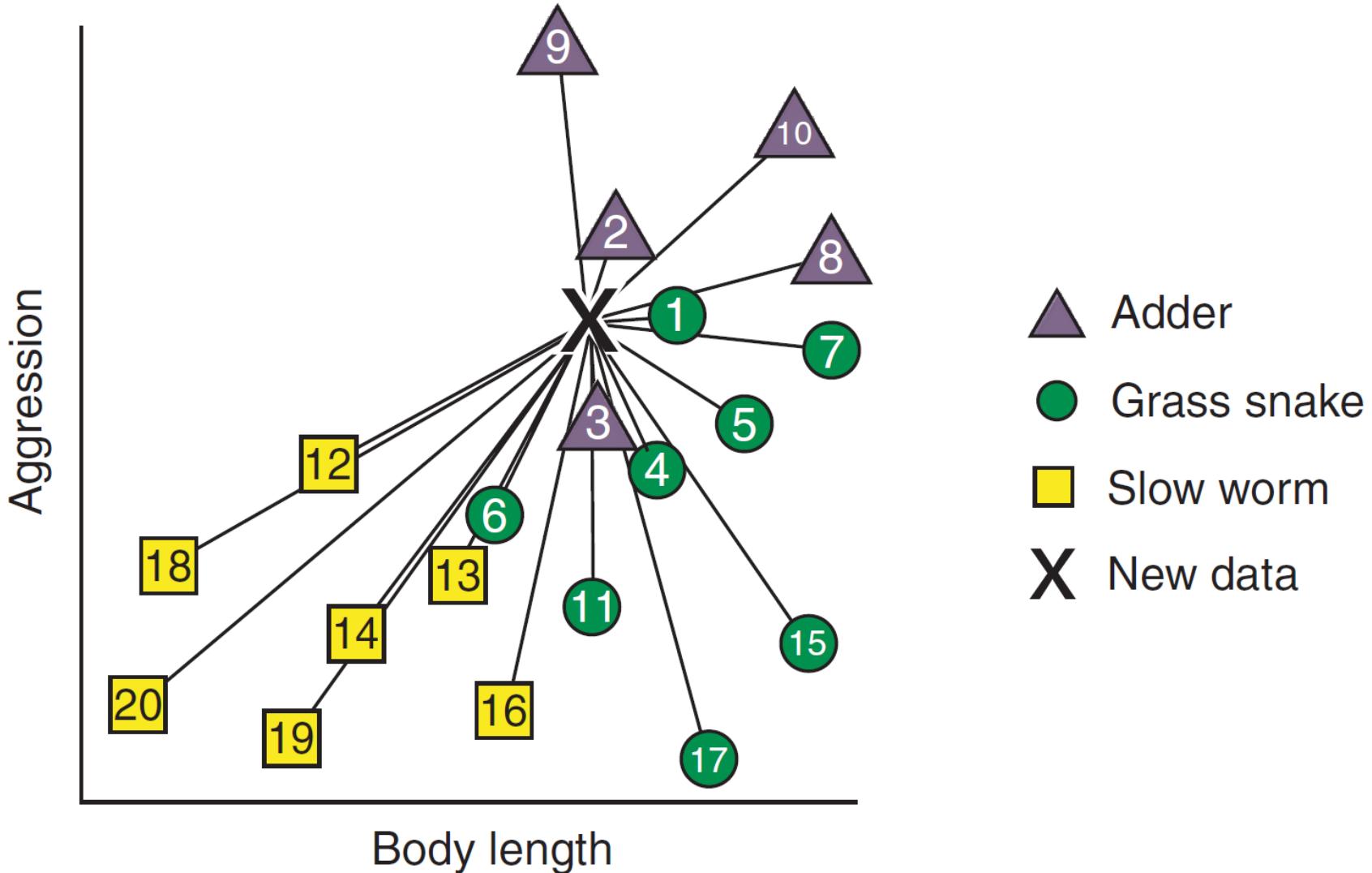
Classification Methods – K-NN



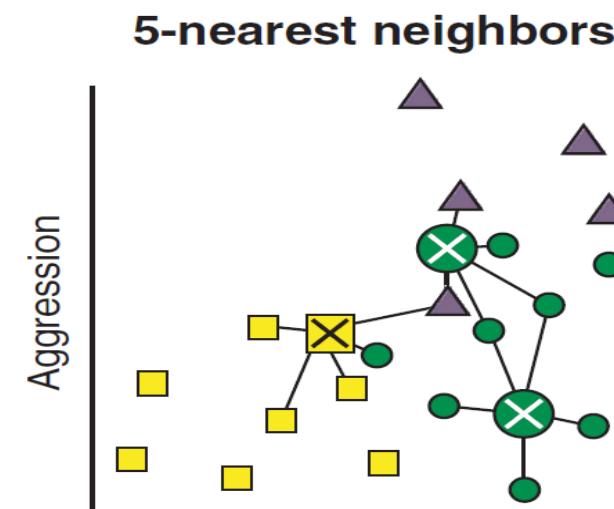
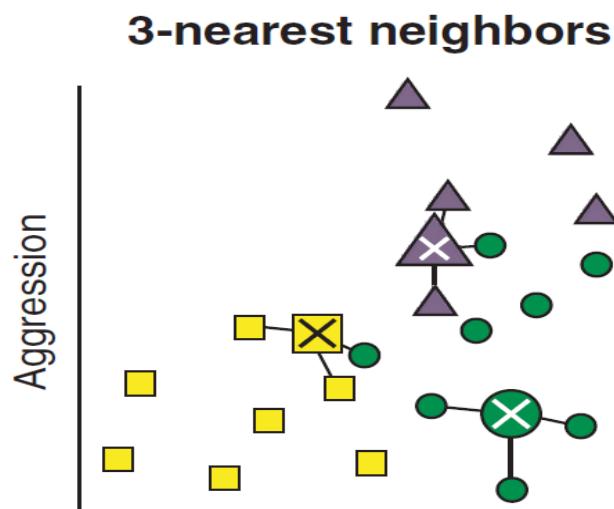
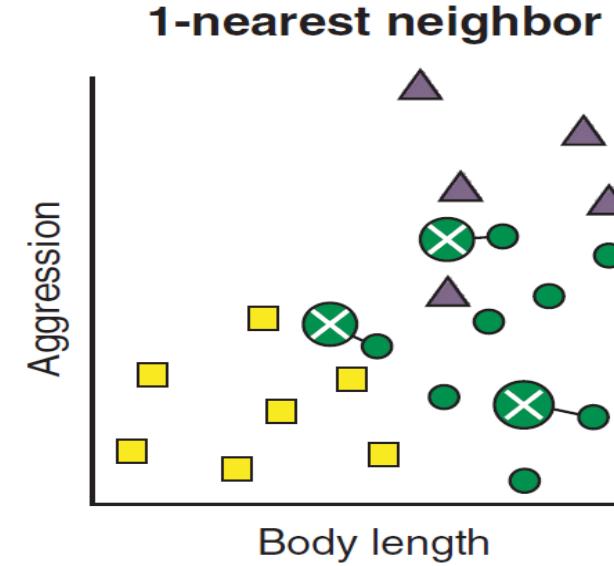
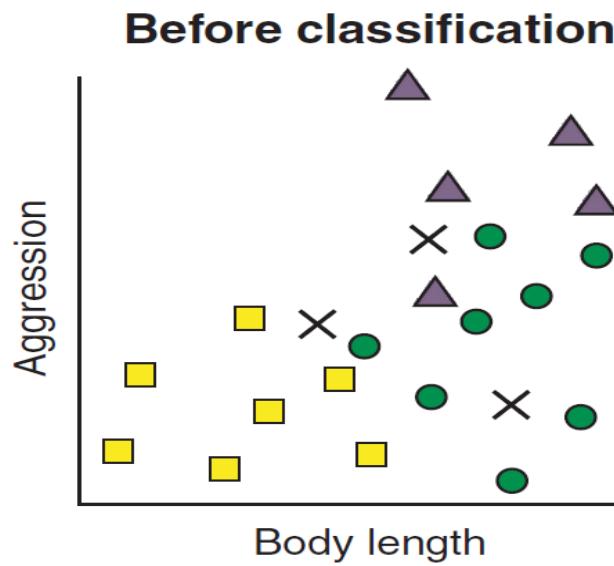
Classification Methods – K-NN



Classification Methods – K-NN



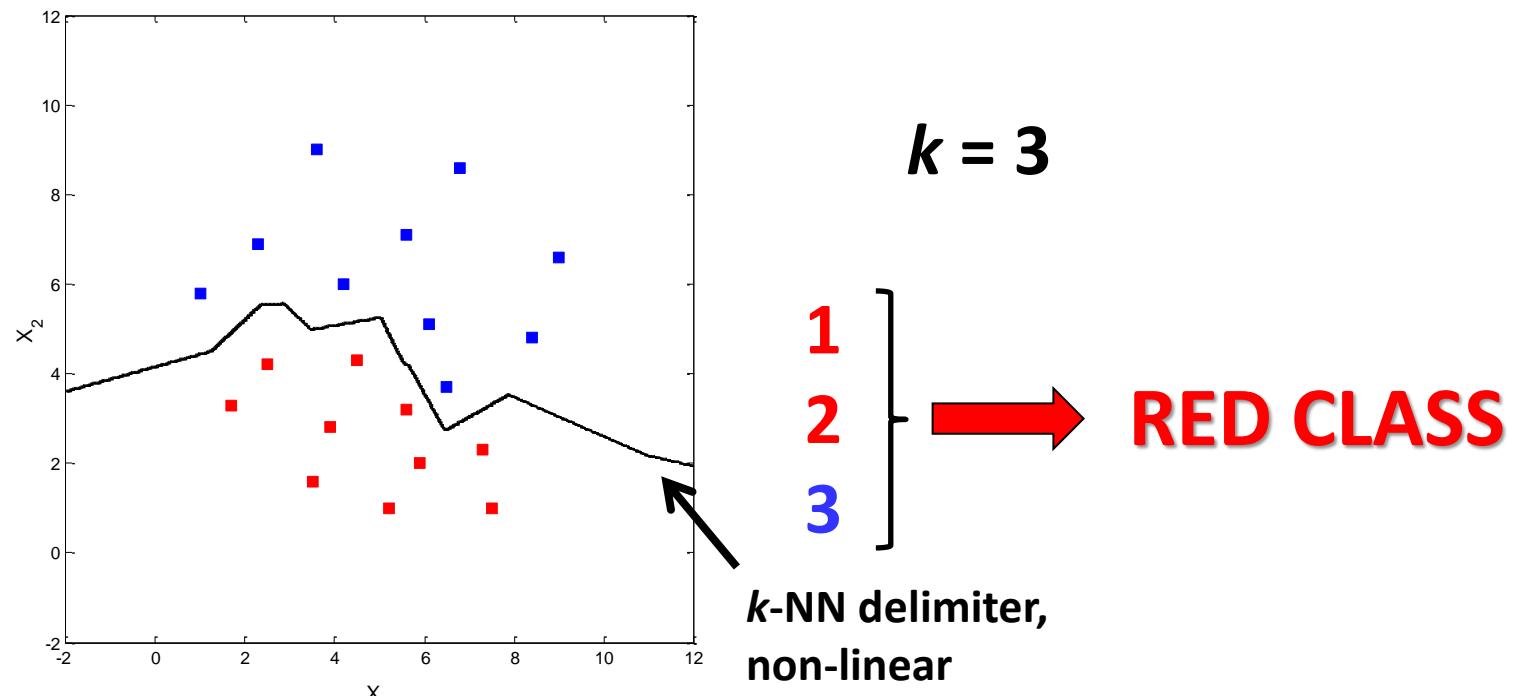
Classification Methods – K-NN



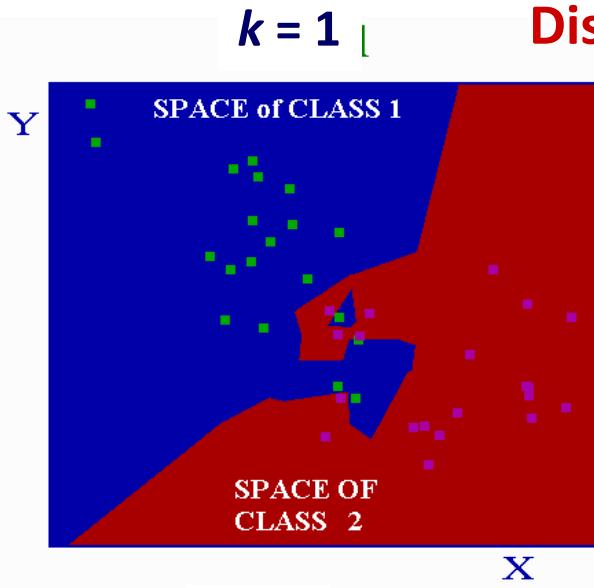
K-NN IS A DISTANCE-BASED APPROACH

NON-LINEAR CLASSIFICATION technique, useful for exploring samples in the vicinity of a specific sample of interest.

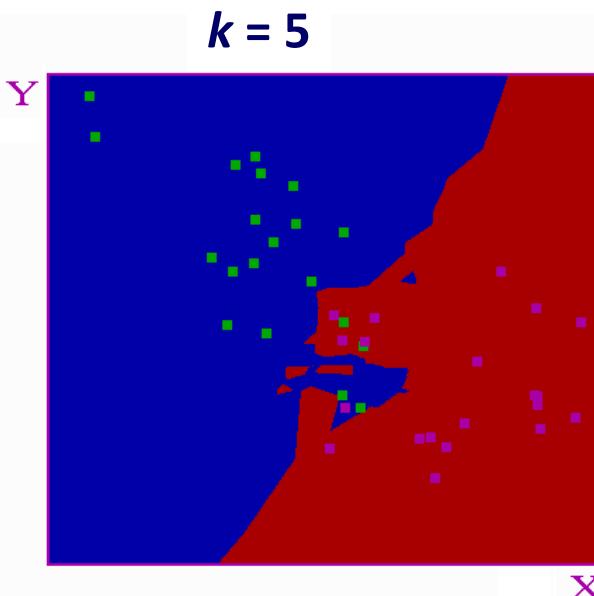
The **distances** among **all the samples** in the training set are calculated, and the samples are sorted accordingly. Samples are ranked based on the **nearest k samples**.



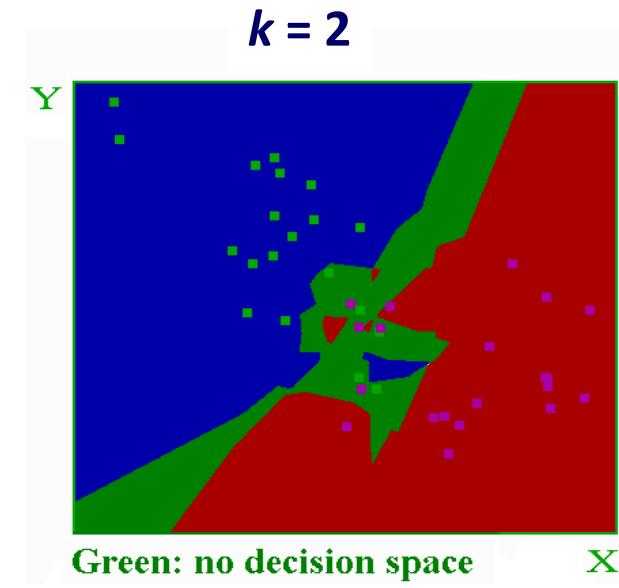
Classification Methods – K-NN



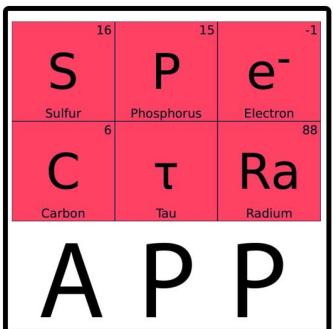
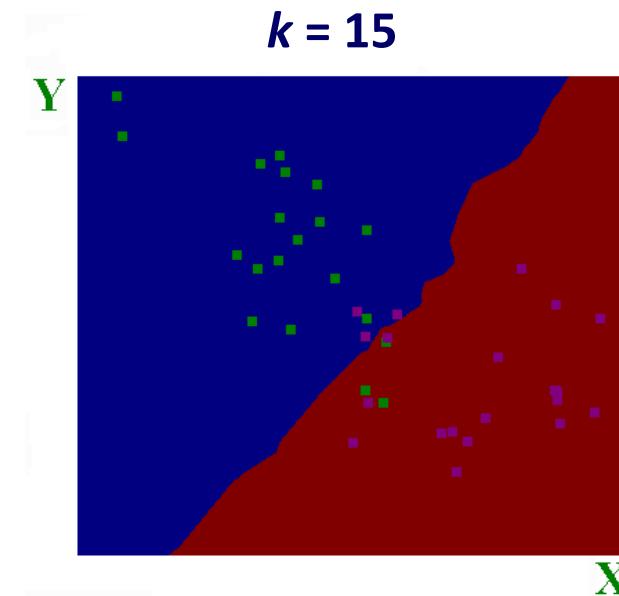
The choice of
 k value is crucial!



k varies the degree
of smoothing of the
delimiter.



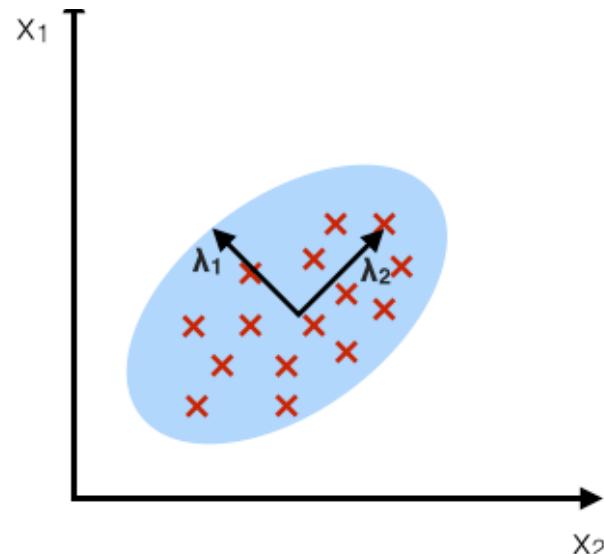
Green: no decision space



Linear Discriminant Analysis (LDA)

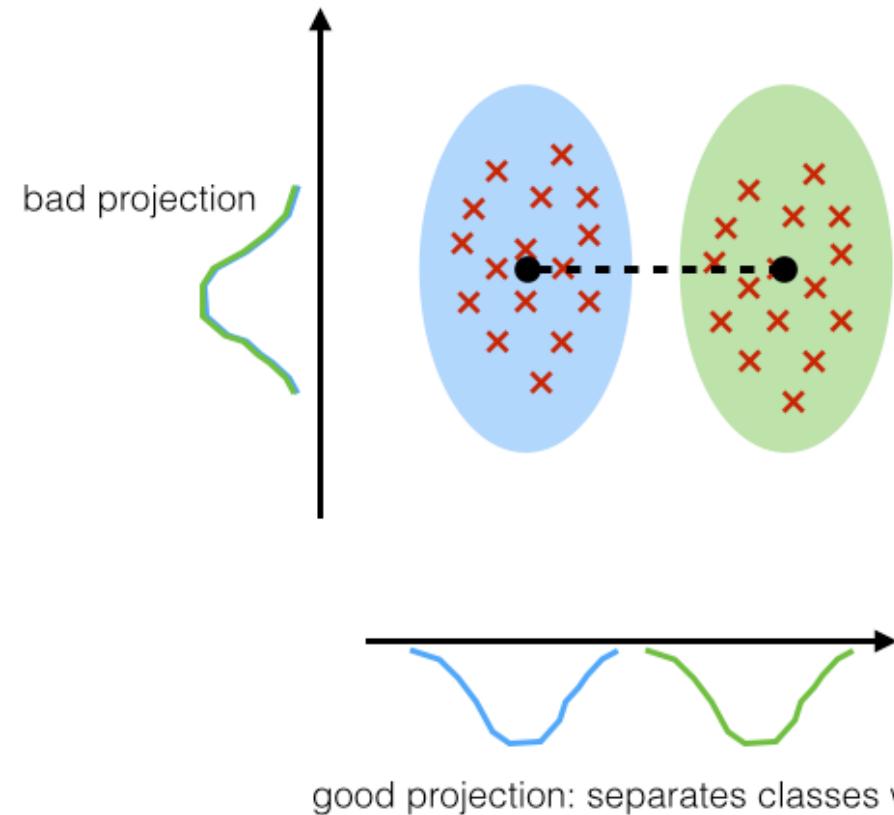
PCA:

component axes that maximize the variance



LDA:

maximizing the component axes for class-separation



Probability-based approaches

Also known as Bayesian approaches

Their goal is to evaluate, for a given sample x , the a-posteriori probability that the sample belongs to a given class.

This probability is calculated using **Bayes' theorem**:

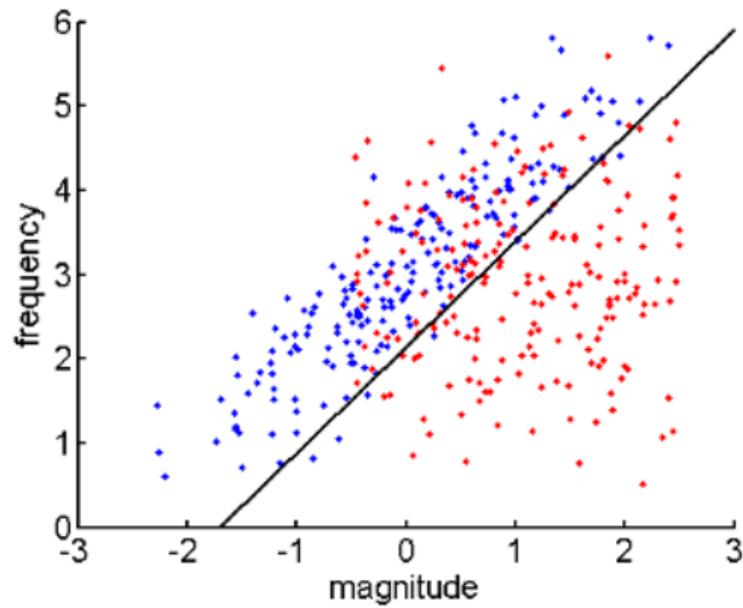
$$p(c | \mathbf{x}) = \frac{p(c) \cdot f(\mathbf{x} | c)}{\sum_c p(c) \cdot f(\mathbf{x} | c)}$$

where:

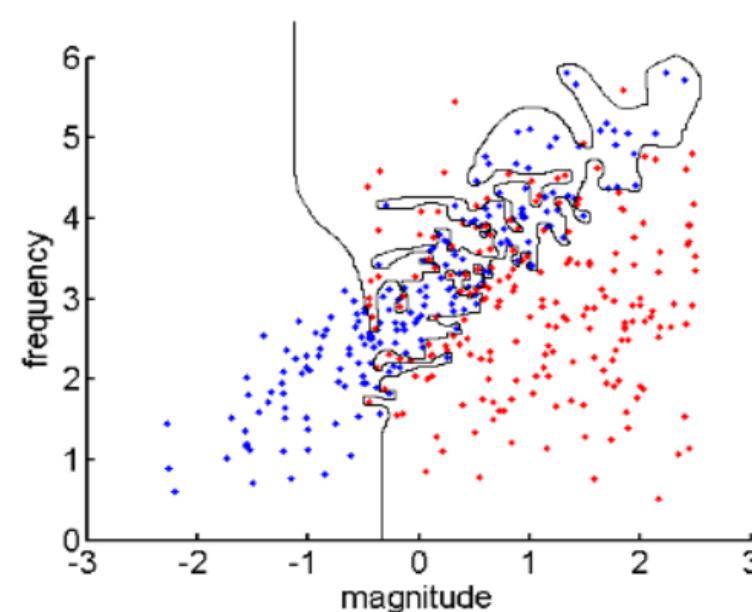
- **$p(c)$** is the **a priori probability** of class c ;
- **$f(\mathbf{x} | c)$** is the conditional probability (density) of \mathbf{x} → the probability that X has a value equal to x , if it is true that it belongs to class c ;
- **$p(c|x)$** is the posterior probability of class c → the probability that the object belongs to class c after observing a value equal to x .

Probability-based approaches

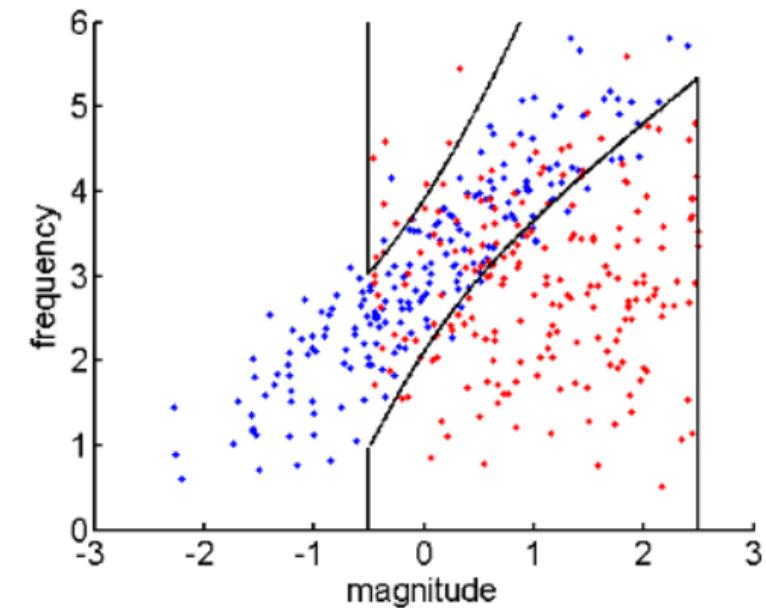
Linear (LDA)



Polynomial (QDA,...)



Maximum Likelihood Estimation



Probability-based approaches

The discrimination methods classify the samples into one of the N available classes.

To build the model we start from the **Bayes Rule** for which:

“A sample must be assigned to the class for which it is greater
its probability of belonging ”

$$p(c | \mathbf{x}) = \frac{f(\mathbf{x} | c)}{\sum_c f(\mathbf{x} | c)}$$

Probability-based approaches

Important points of model construction:

1. A sample is classified in class C whereby the $p(c|x)$ value is higher.

$$p(c|x) = \frac{f(\mathbf{x}|c)}{\sum_c f(\mathbf{x}|c)}$$

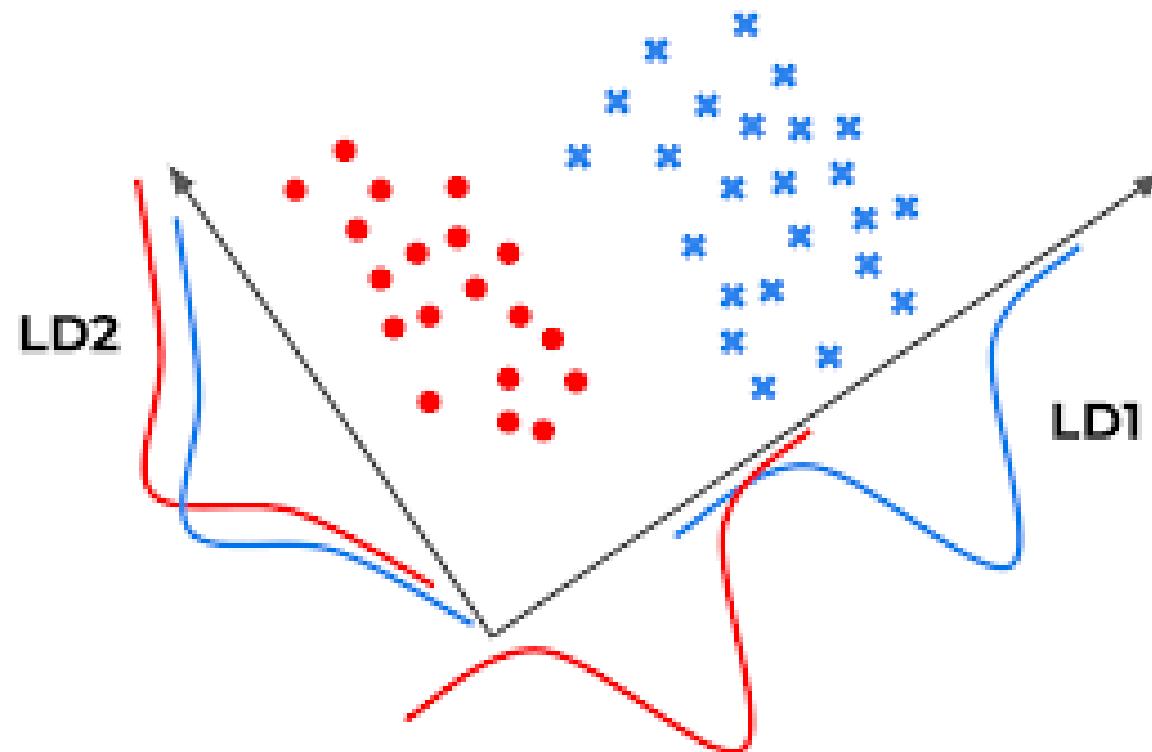
2. In the case of 2 classes, the **DELIMITER** is defined with the x values for which:

$$p(c_1|x) = p(c_2|x) \quad \text{i.e.} \quad p(c_1) f(x|c_1) = p(c_2) f(x|c_2).$$

3. In the case of more than 2 categories, the delimiter is defined by comparing the various classes 2 by 2.

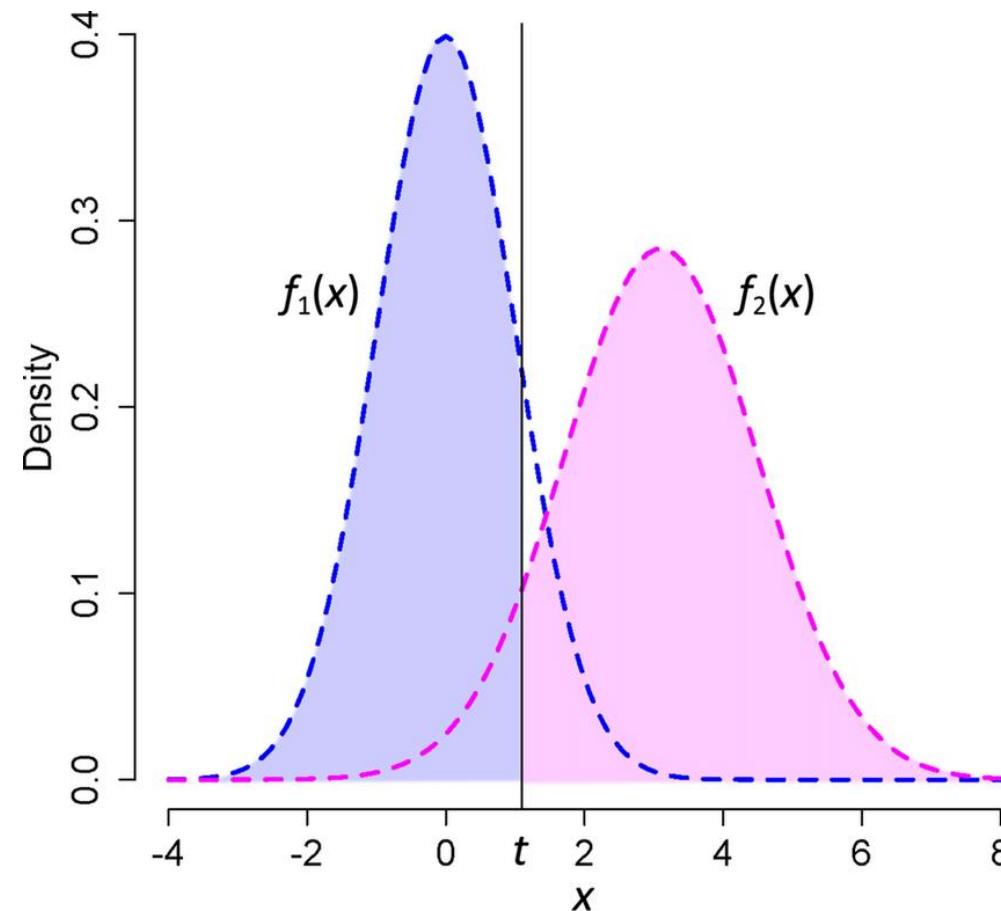
Probability-based approaches

The more the classes are separated for a given variable (based on the parameter estimation of their **probability distributions** → **frequency distributions**), the better the model will work!



Probability-based approaches

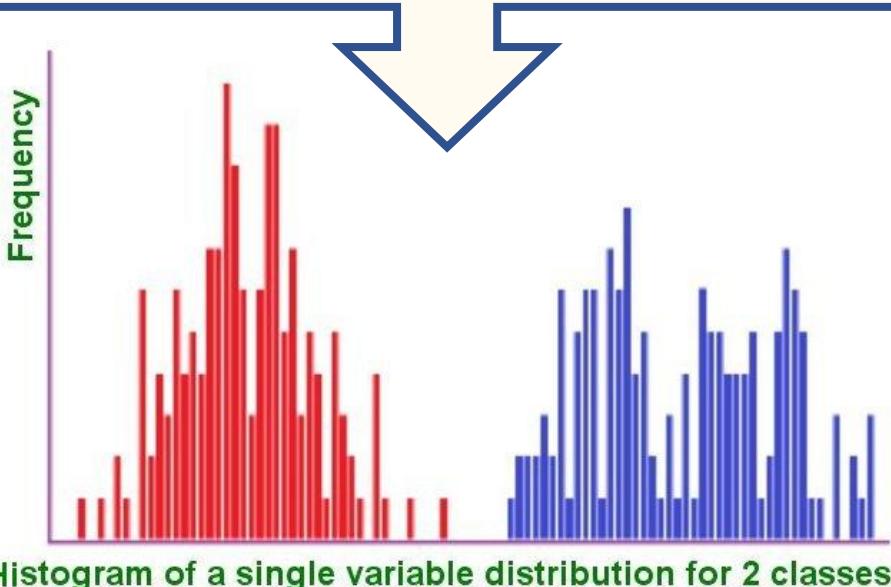
Since they are **PARAMETRIC** type models, the distributions are evaluated using **location** and **dispersion** parameters.



Classification Methods – Discriminant Analysis

Discriminant Analysis includes several probability-based methods founded on the estimation of probability distributions. They are also known as Bayesian's methods, with reference to Bayes' rule affirming that “an object must be assigned to the class for which it is greater its probability of belonging”.

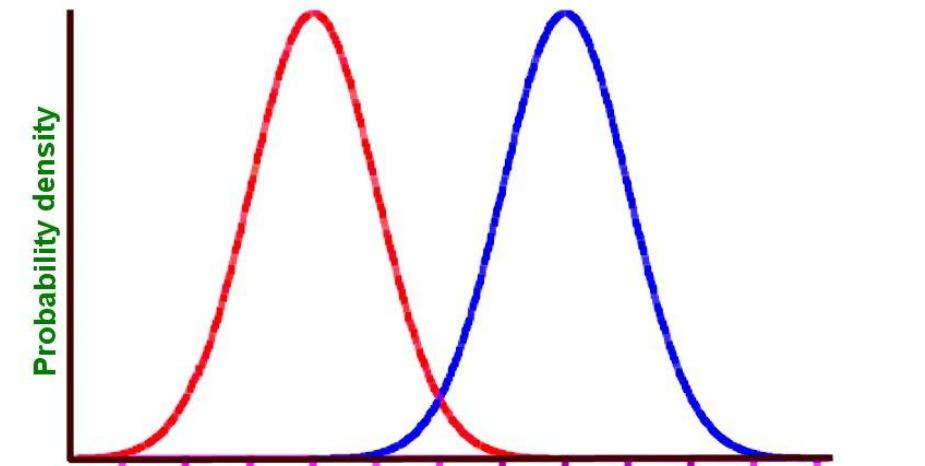
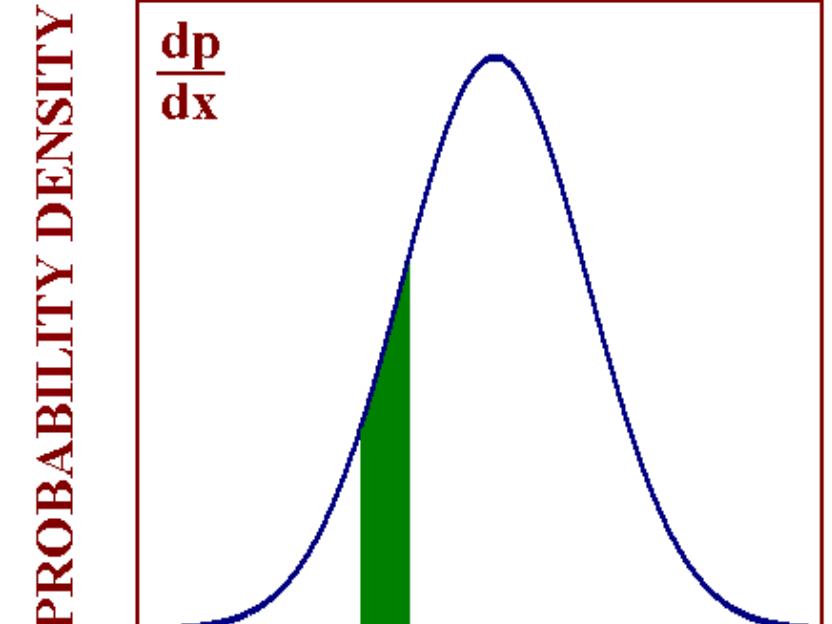
Actually, in real experimental cases each variable's distribution is likely to appear as the histogram below, and the generating continuous curve has to be postulated.



$$p_{a,b} = \int_a^b \frac{dp}{dx} dx$$

Postulated:
Gaussian distribution

Mean, Variance



Linear Discriminant Analysis (LDA)

The **first multivariate classification** approach, defined by R.A. Fisher in 1936.
 Starting from the available data, the centroids of the various classes are calculated and a combined variance (**pooled** variance - variance-covariance matrix) is defined.

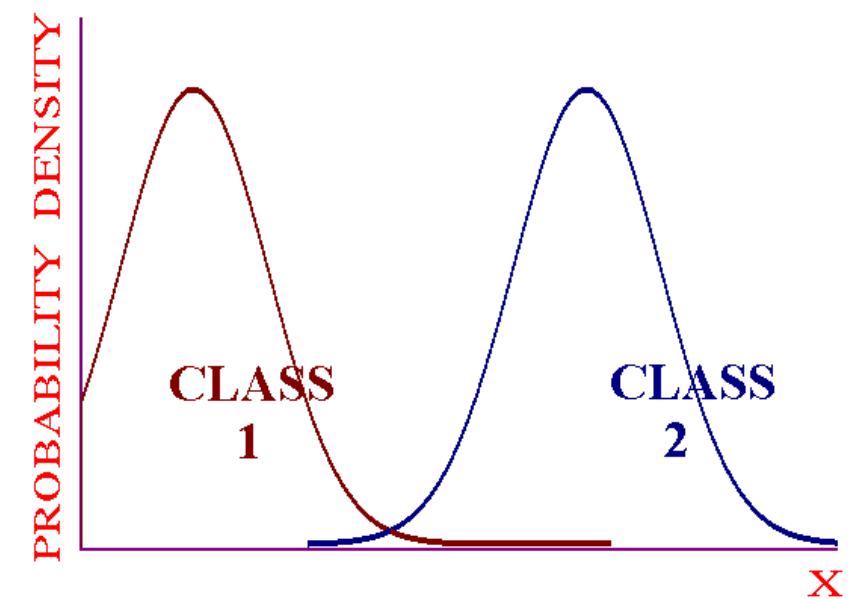
The probability distributions, for each class, are evaluated
 on the basis of 2 hypotheses:

1) Normal (multivariate) distribution for each class;

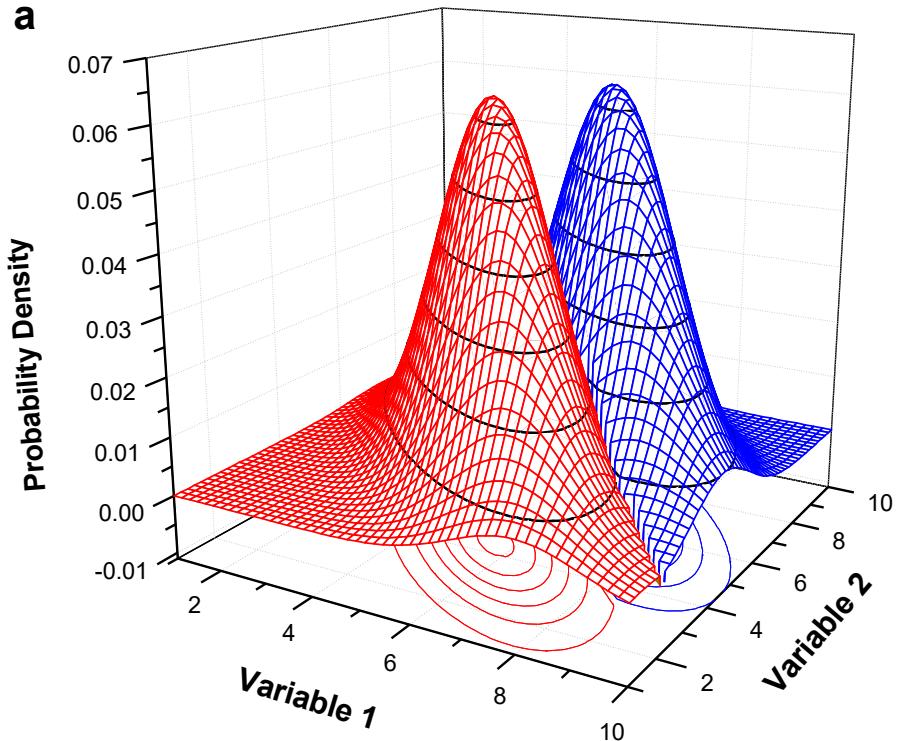
$$p(g|\mathbf{x}_i) \propto \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{S}_g|} e^{-\frac{1}{2}(\mathbf{x}_i - \bar{\mathbf{x}}_g)^T \mathbf{S}_g (\mathbf{x}_i - \bar{\mathbf{x}}_g)}$$

2) Same variance-covariance for all classes (G).

$$\mathbf{S}_i = \mathbf{S}_j = \mathbf{S} = \frac{\sum_{g=1}^G (n_g - 1) \mathbf{S}_g}{N - G}$$

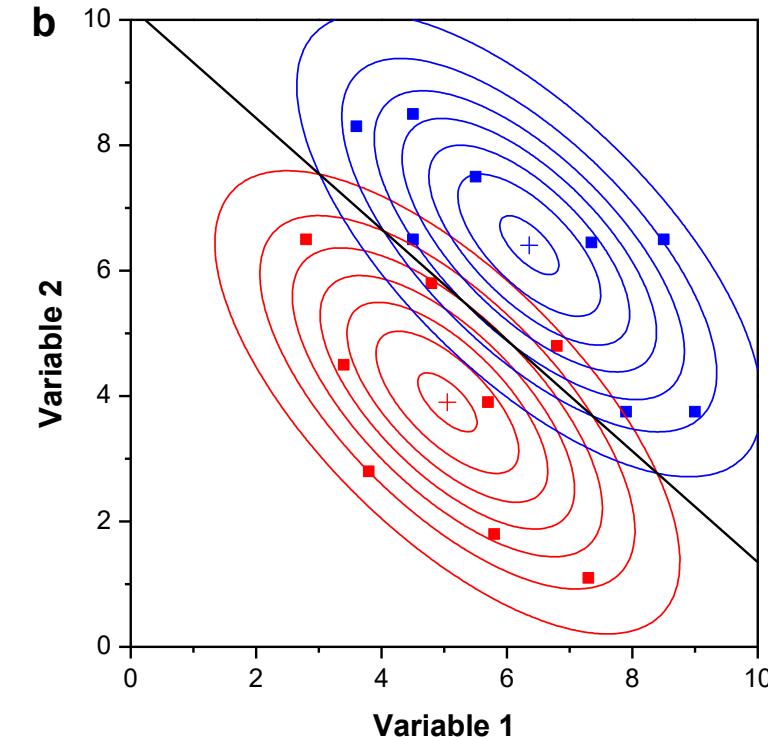


Classification Methods – Linear Discriminant Analysis (LDA)



A bivariate example

**Two-dimensional probability
distributions for 2 classes**

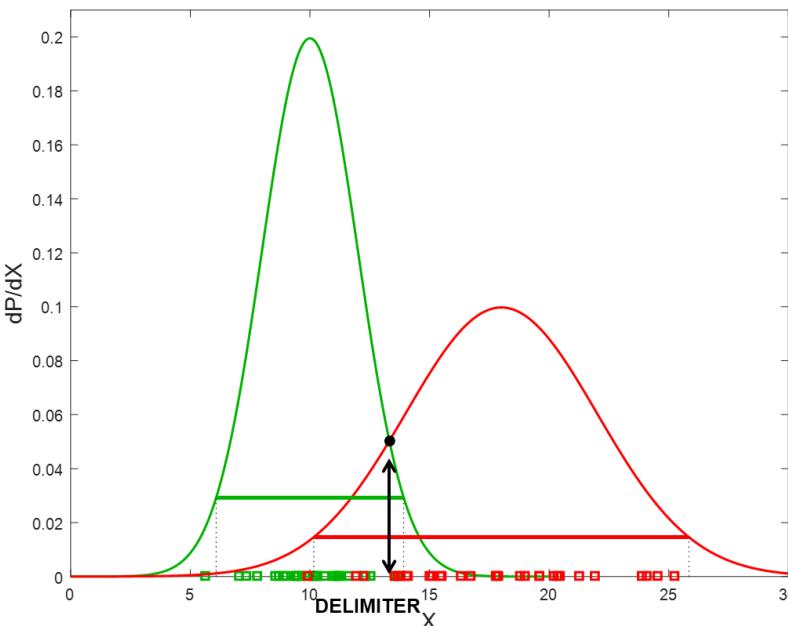


LDA delimiter for 2 classes

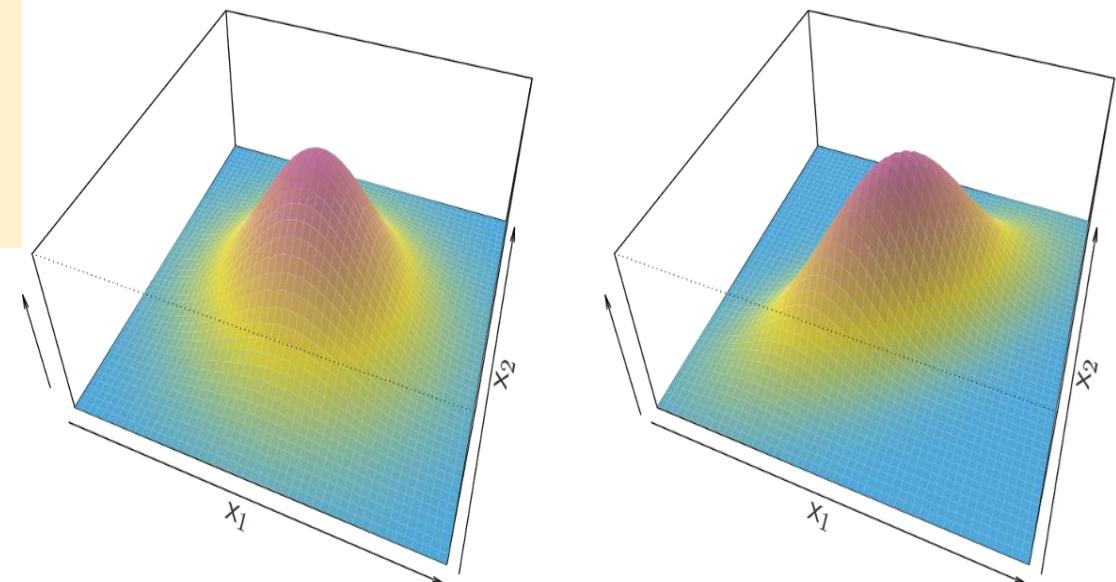
Classification Methods – Quadratic Discriminant Analysis (QDA)

As an alternative to LDA, another discriminant technique is **Quadratic Discriminant Analysis (QDA)**, which maintains only one constraint: (1) all variables should have Gaussian distribution for all classes. Thus, means and variances are still calculated for all classes, but variances-covariances are not pooled anymore and separated covariance matrices are calculated for each class. Delimiter equation:

$$P_g f(g|x_i) = P_g \frac{e^{-\frac{1}{2}(x_i - \bar{x}_g)^T S_g^{-1} (x_i - \bar{x}_g)}}{\sqrt{(2\pi)^p |S_g|}} = P_h \frac{e^{-\frac{1}{2}(x_i - \bar{x}_h)^T S_h^{-1} (x_i - \bar{x}_h)}}{\sqrt{(2\pi)^p |S_h|}} = P_h f(h|x_i)$$



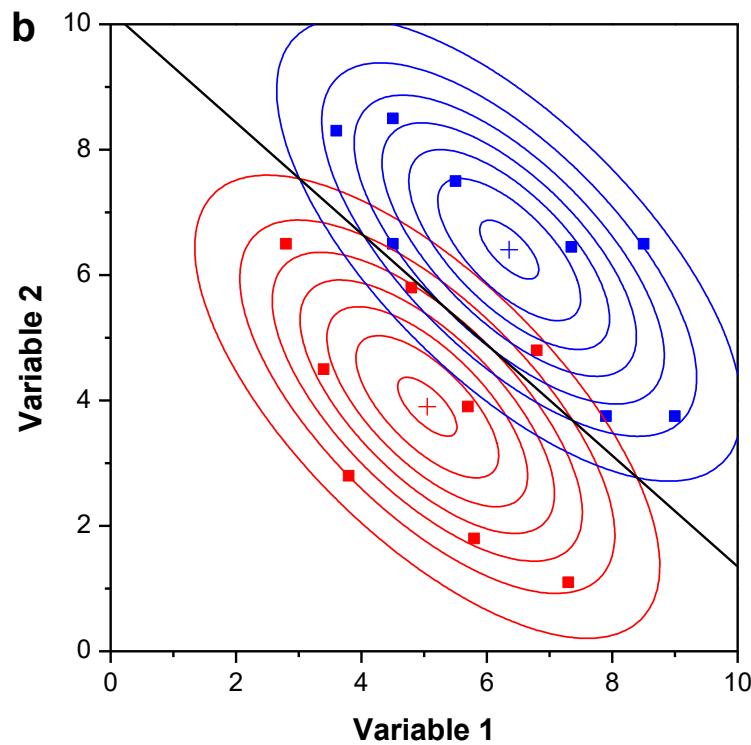
Examples of QDA models for one- and two-dimensional data



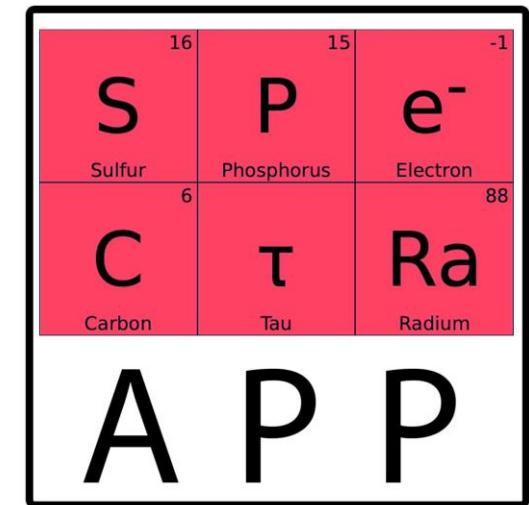
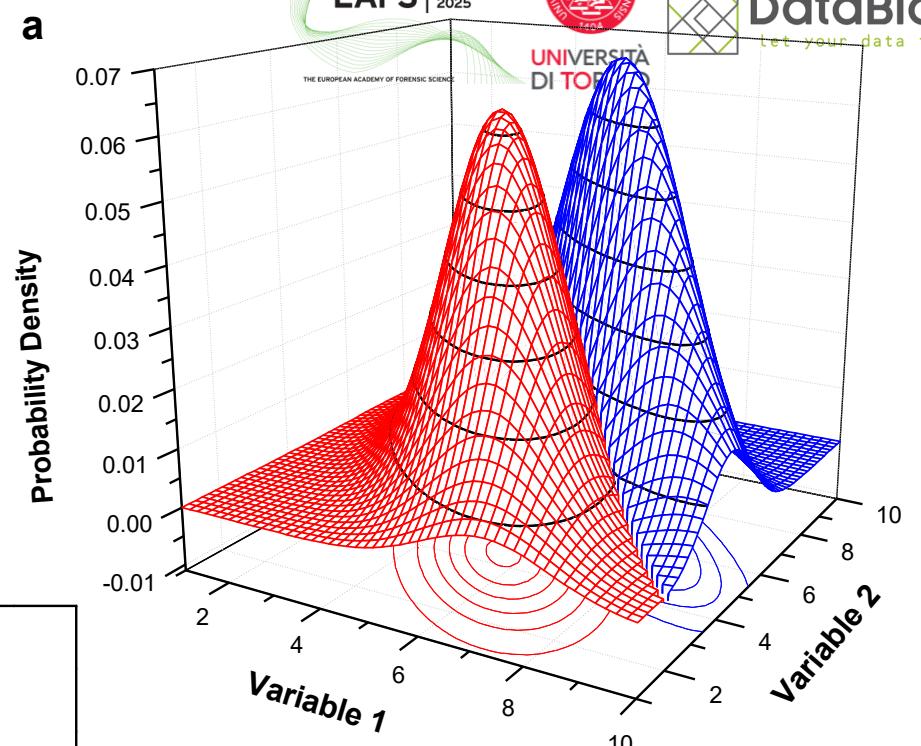
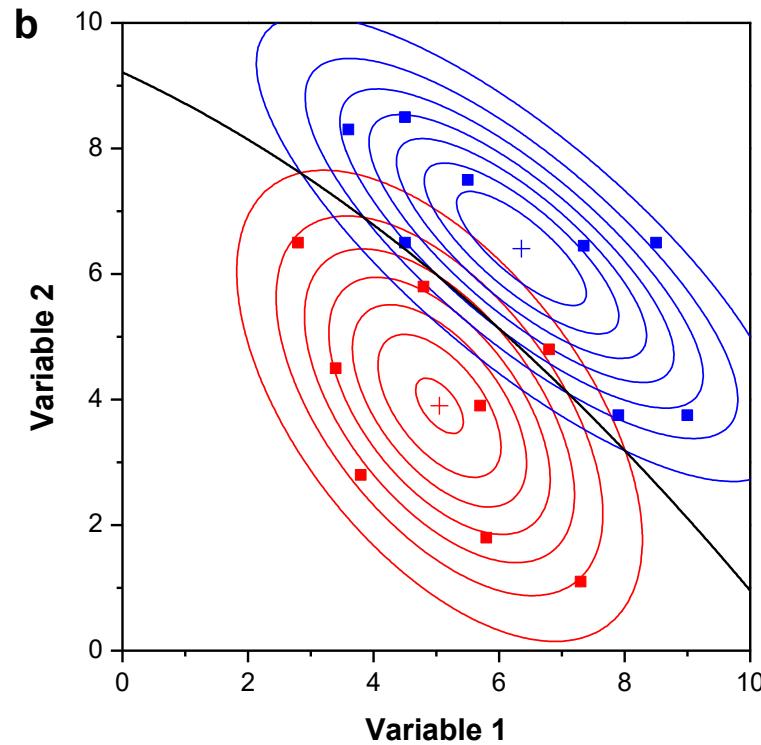
Classification Methods – LDA vs. QDA

SAME DATA – DIFFERENT MODELS

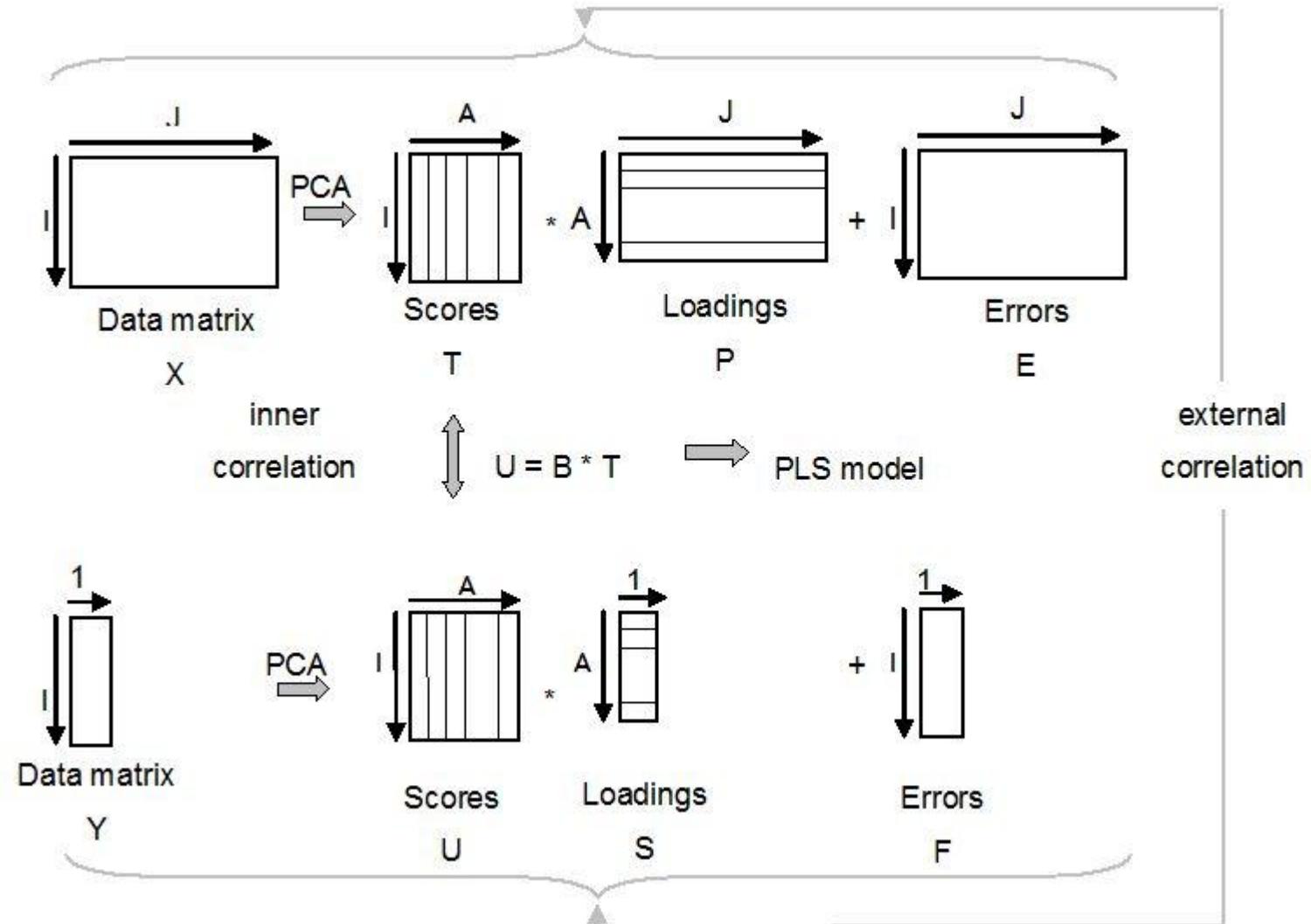
LDA



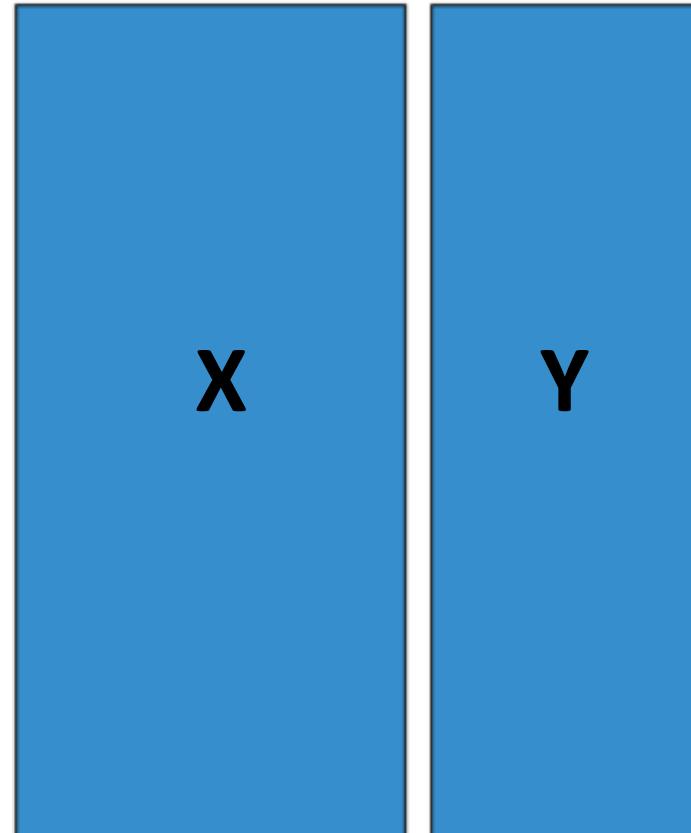
QDA



Partial Least Squares – Discriminant Analysis PLS - DA (if Regression – PLS-R)

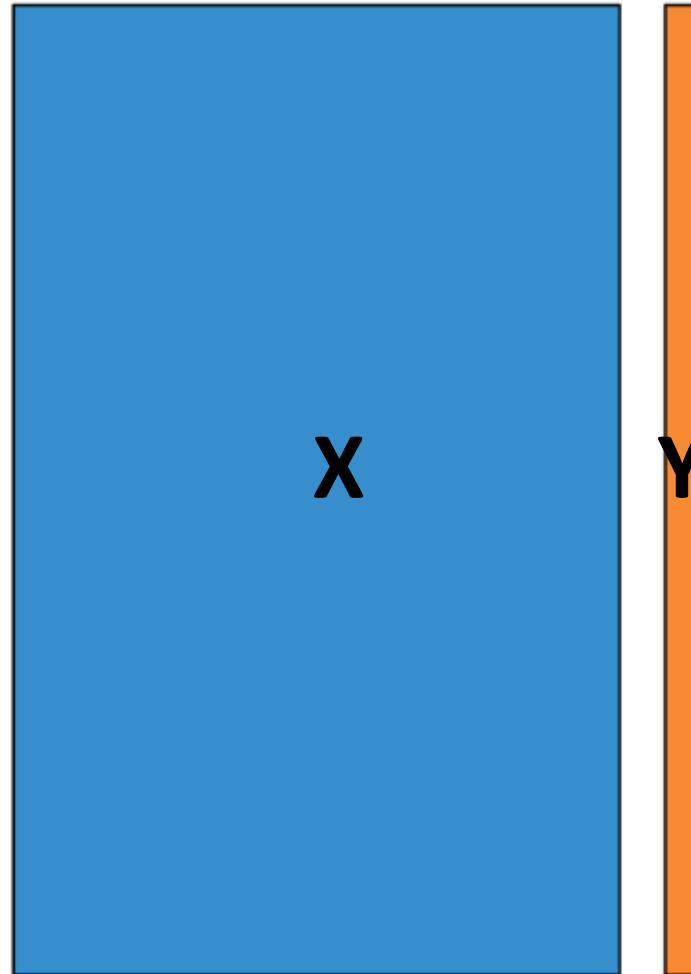


Partial Least Squares – R (Regression – PLS-R)



X – Quantitative
(continuous)

Partial Least Squares – DA (Discriminant Analysis – PLS-DA)



X – Quantitative (continuous)

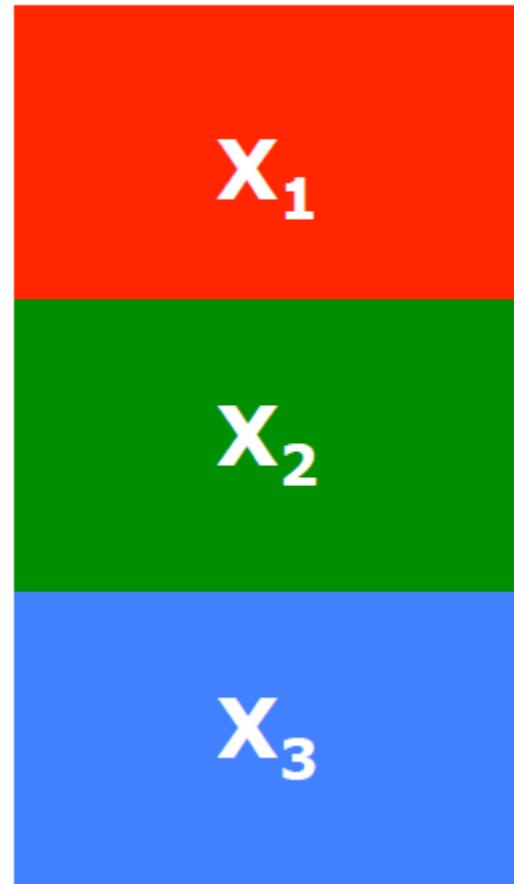


Y – Qualitative (DA) / Qualitative (Regression)



Partial Least Squares – DA (Discriminant Analysis – PLS-DA)

Dataset ‘Training’



Classes

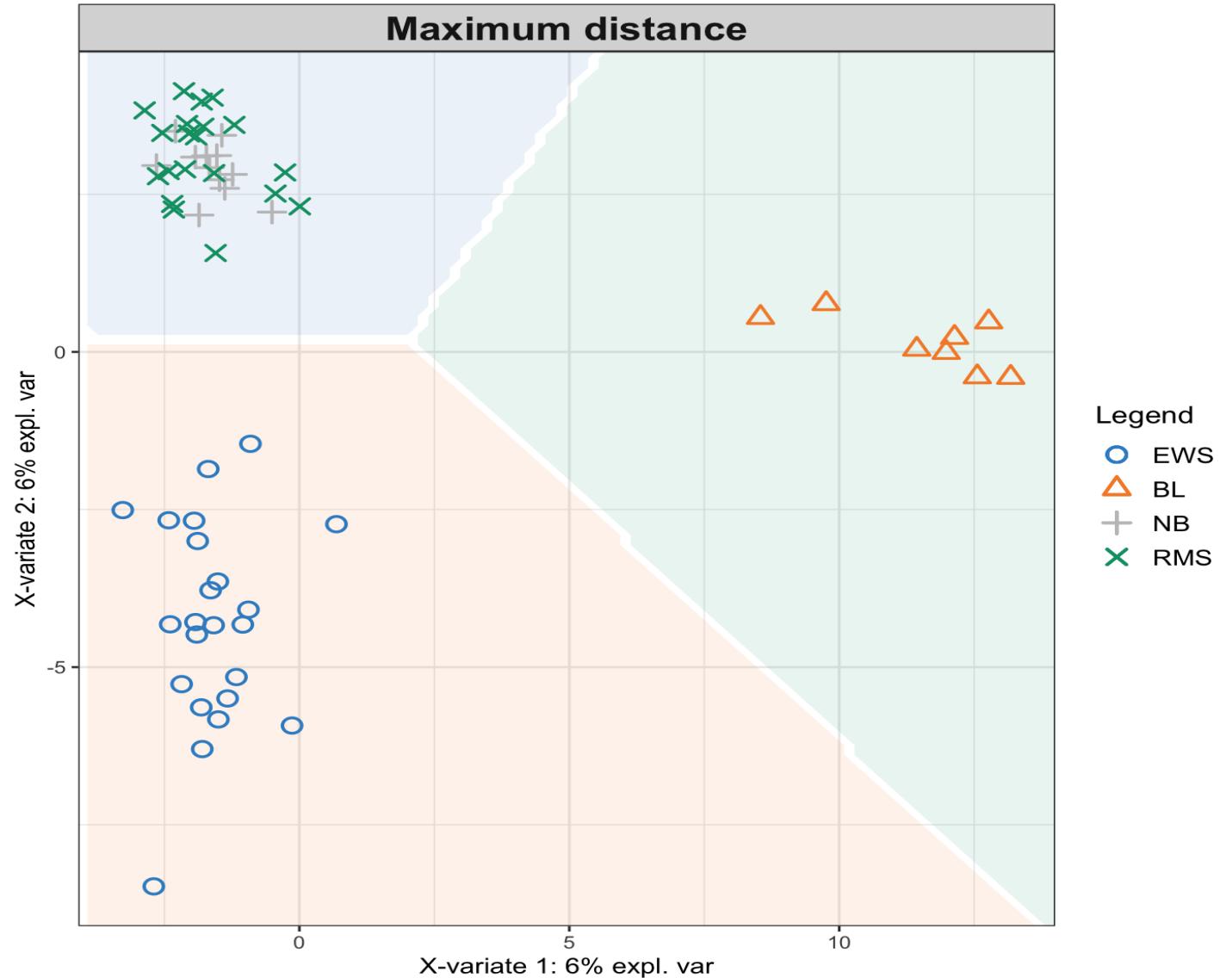
1
1
1
1
1
2
2
2
2
2
3
3
3
3
3

Class matrix

1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

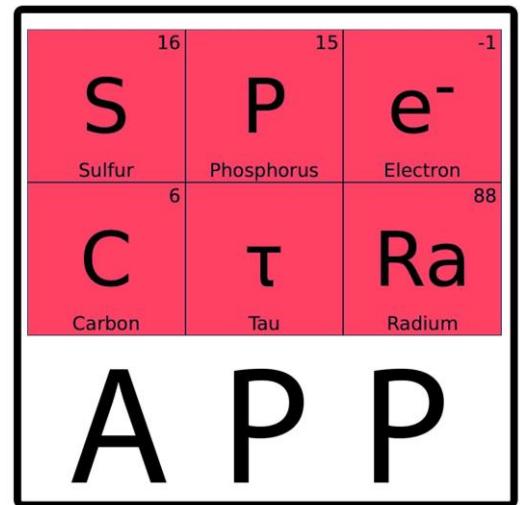
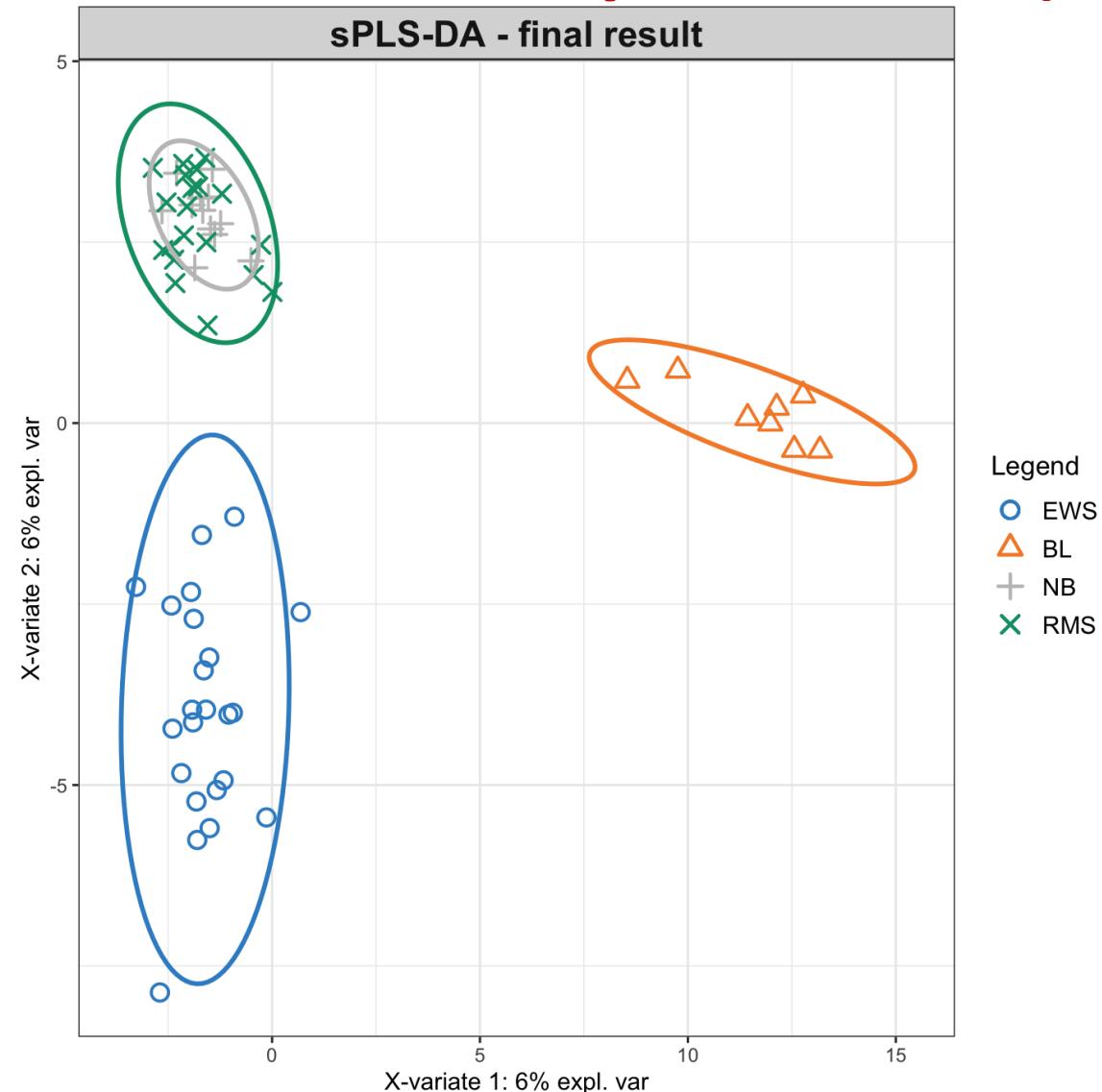
Partial Least Squares – DA

(Discriminant Analysis – PLS-DA)

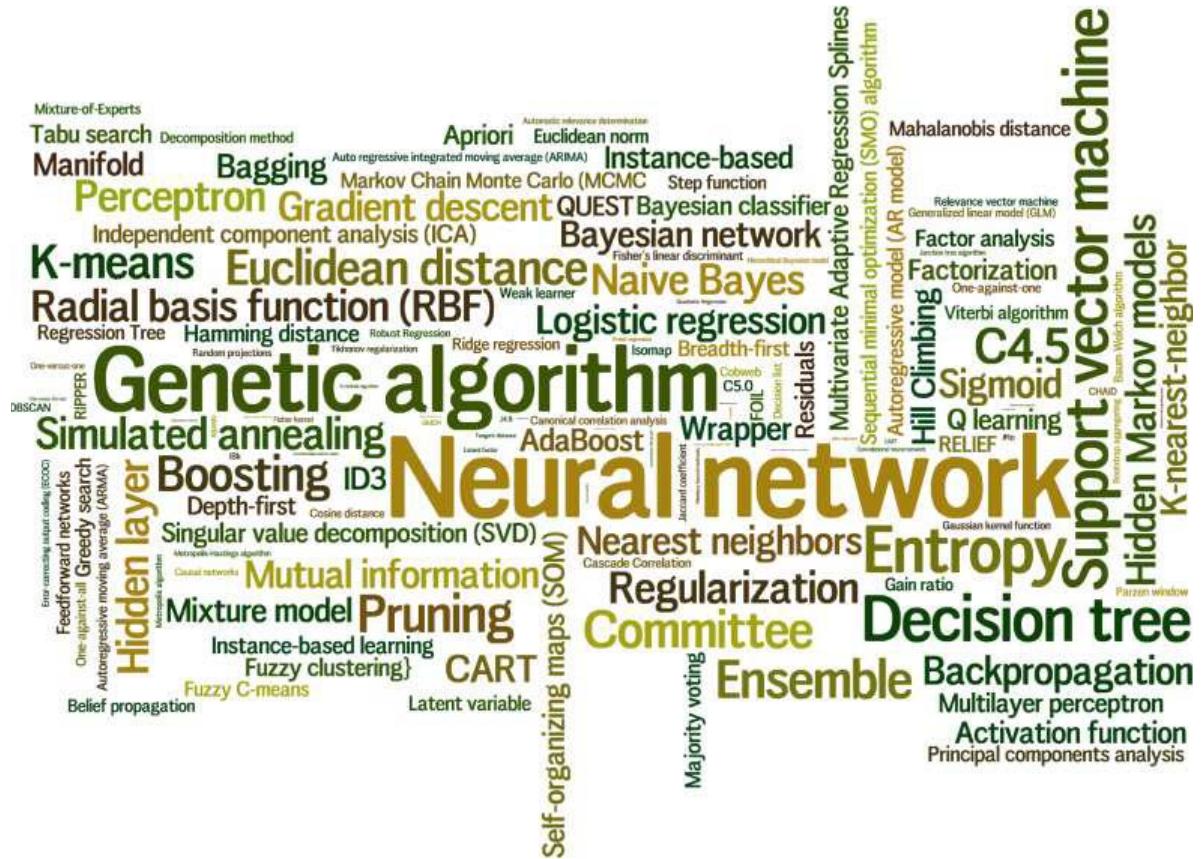


Partial Least Squares – DA

(Discriminant Analysis – PLS-DA)



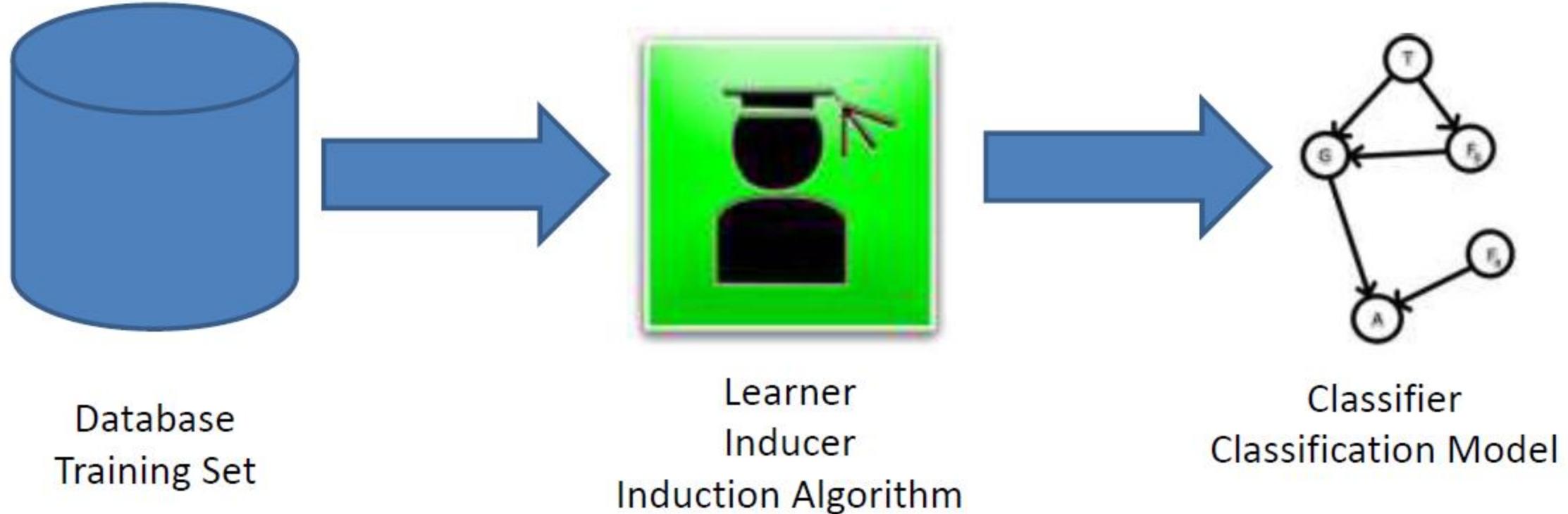
Experience-based – ML approaches



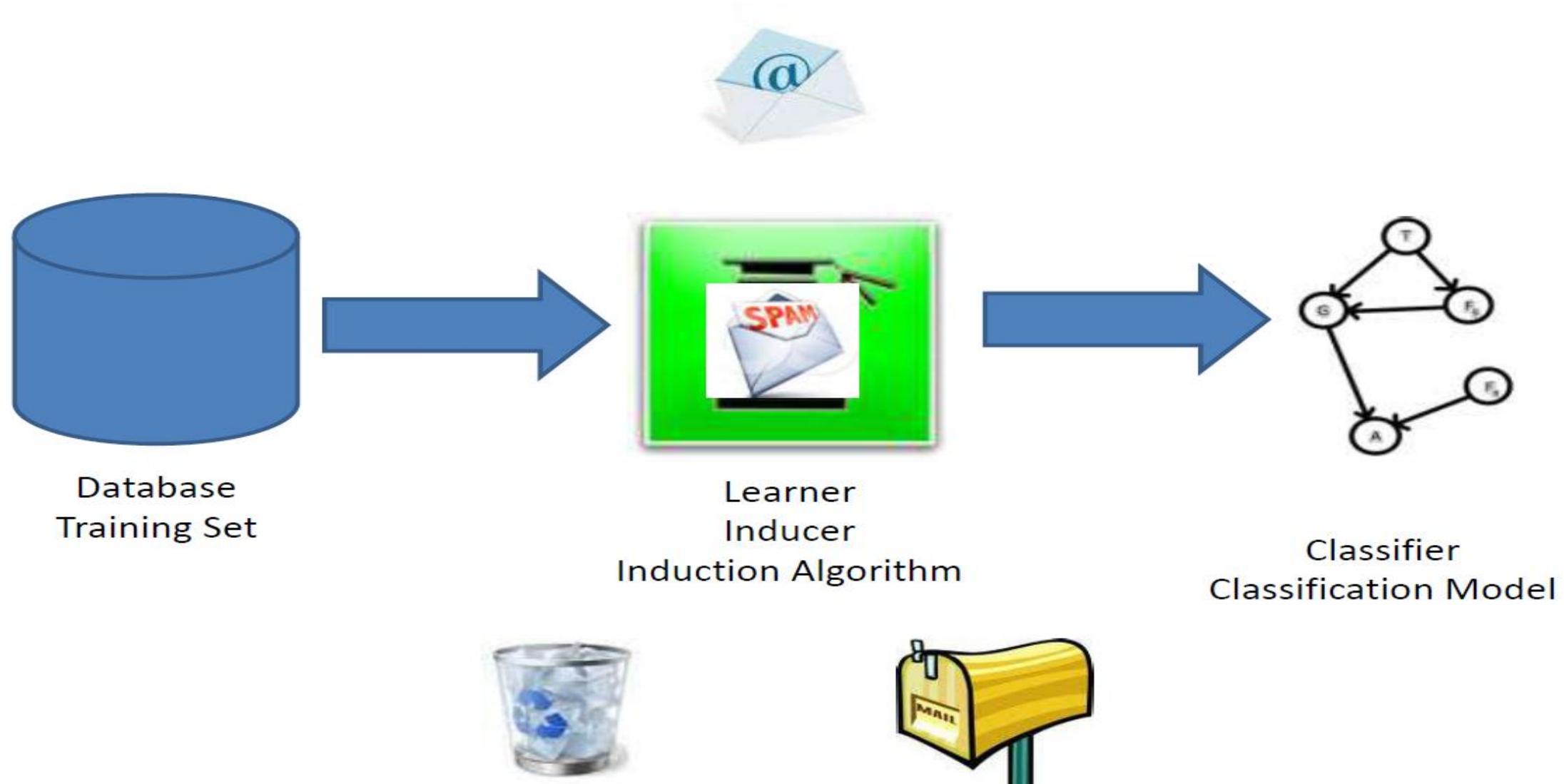
Experience-based – ML approaches



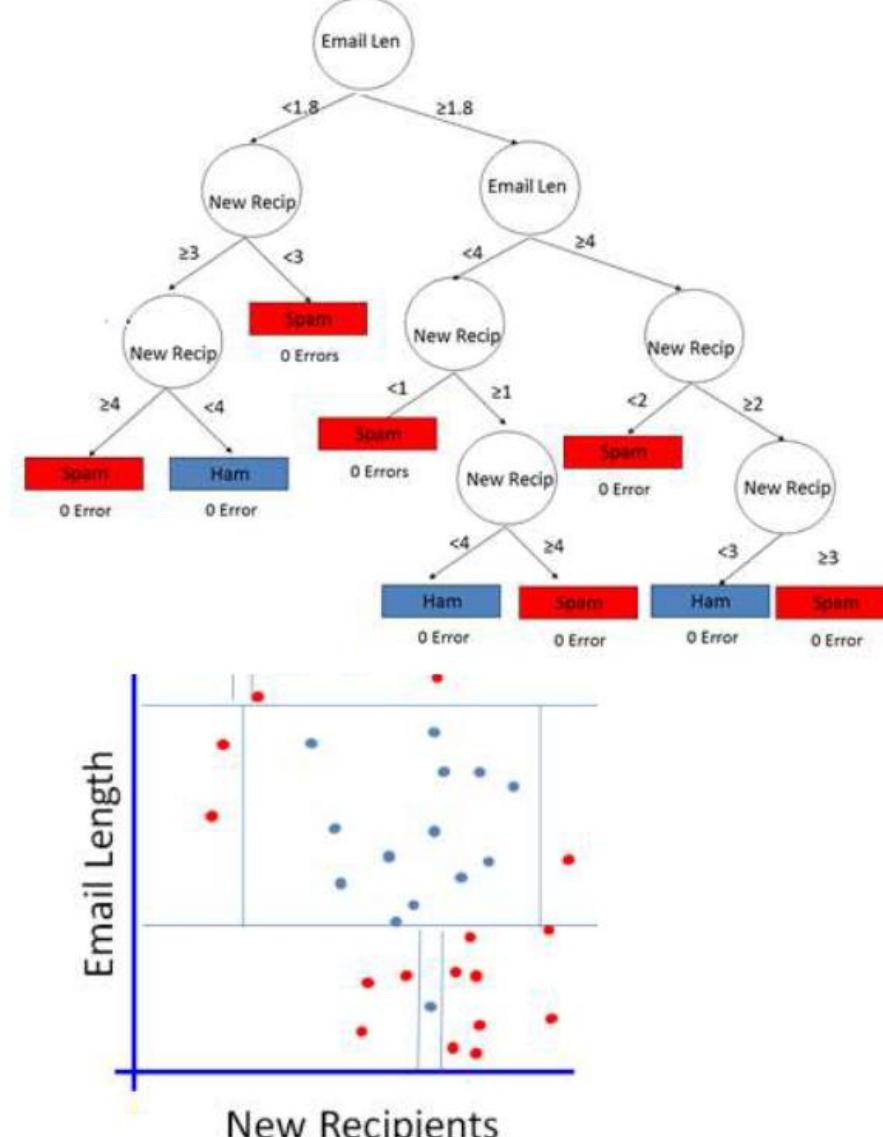
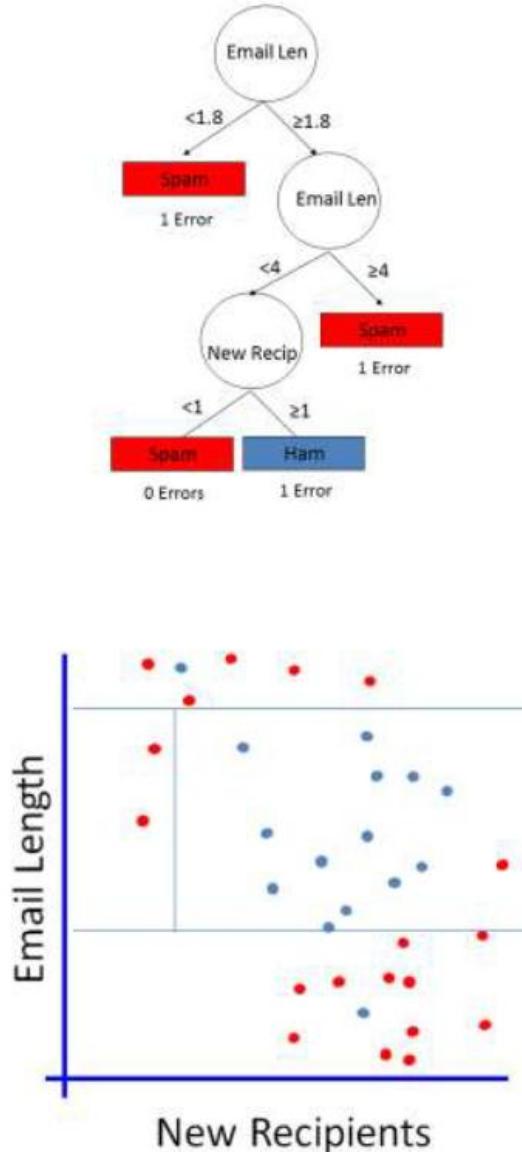
Experience-based – ML approaches



Experience-based – ML approaches



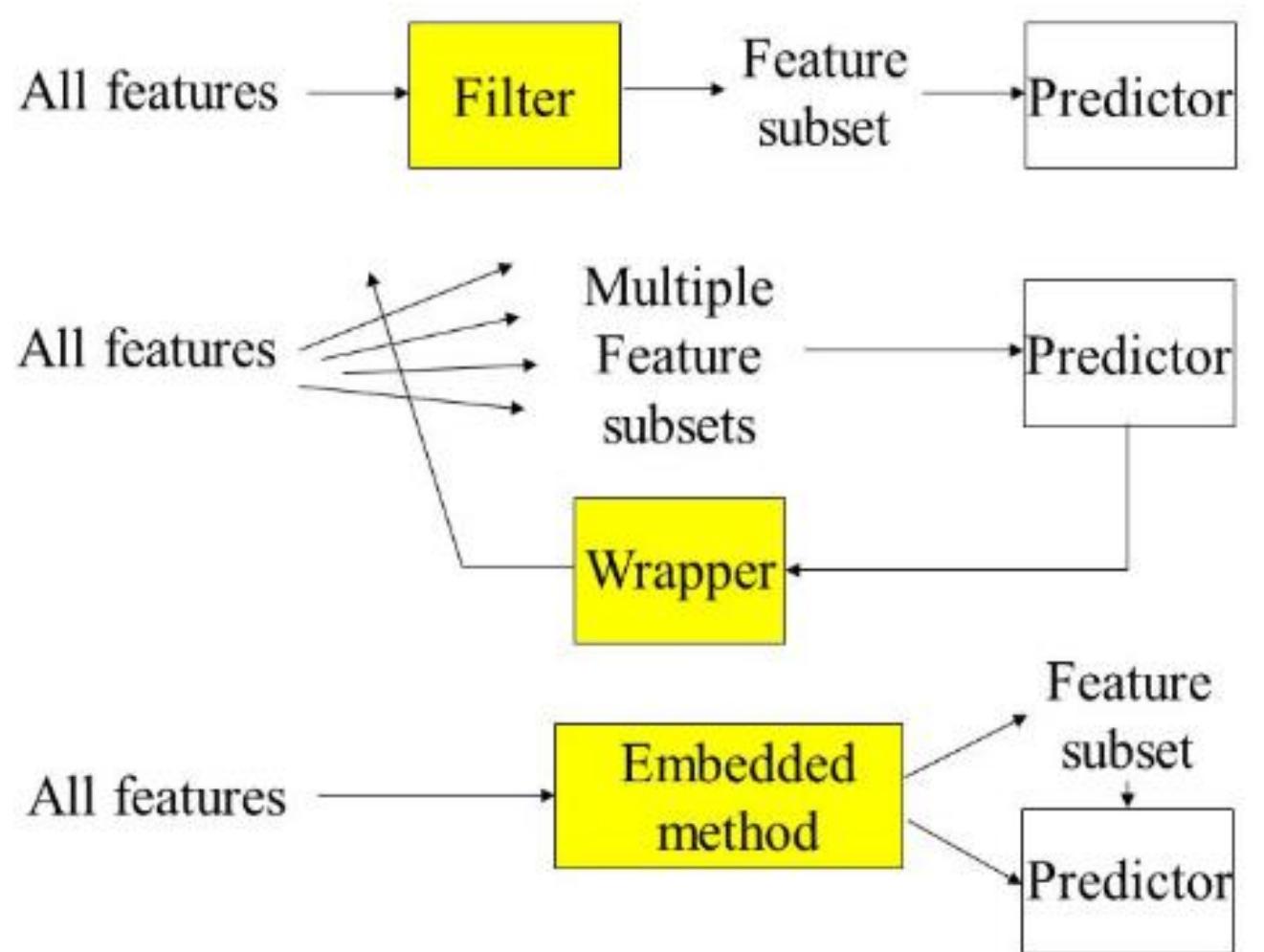
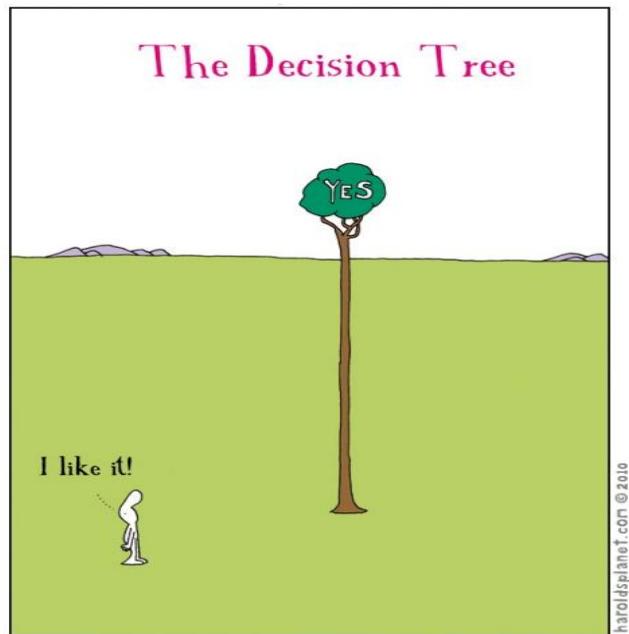
Decision tree – several possible trees...



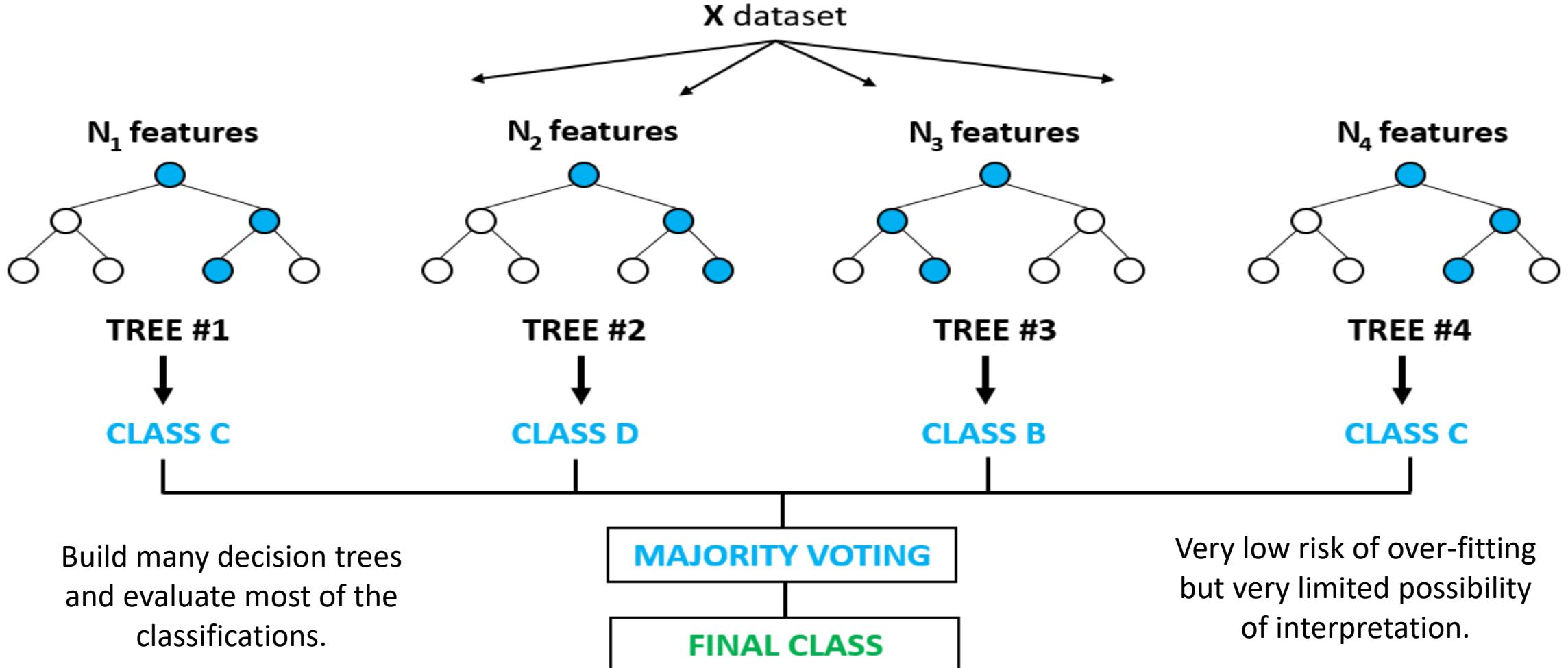
Embedded models (also for feature selection)

Several methods!

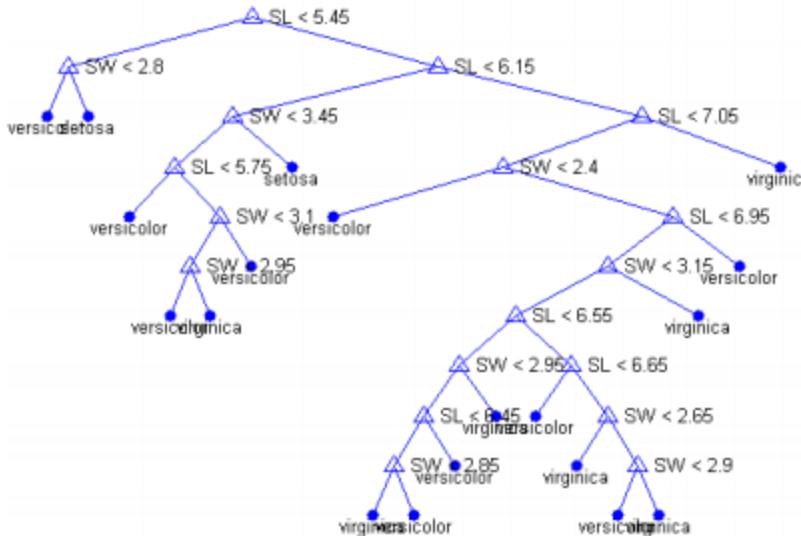
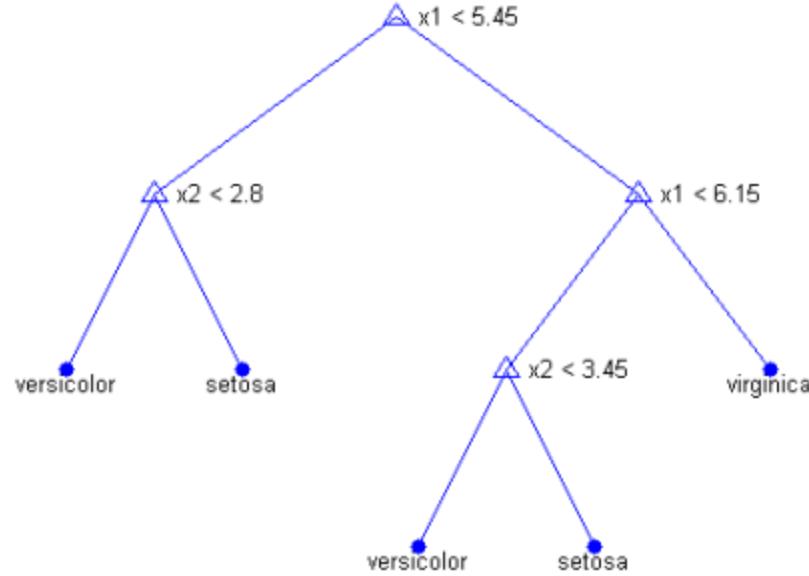
Examples: **random forests**



Random forests



Random forests



They can provide several different results.

Random forests - issues

1. Very time consuming (Big Data).
2. Risk of over-fitting (always test your models).
3. The interpretation is not easy.



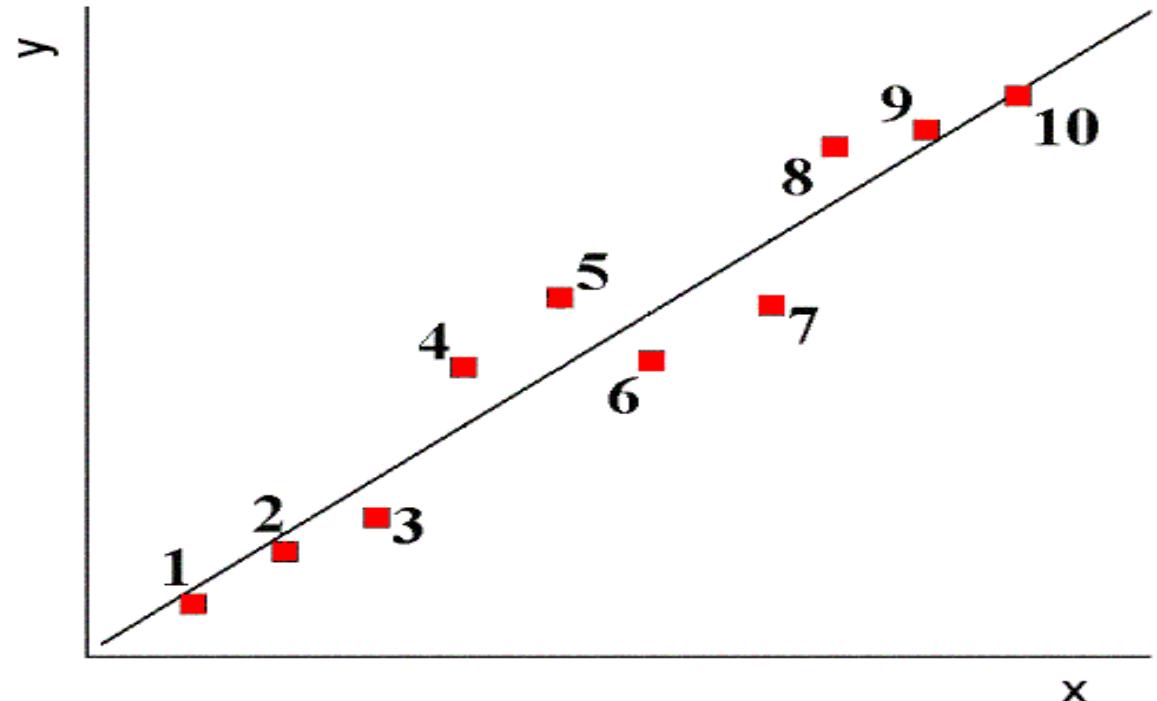
Supervised models - regression



Supervised Machine Learning

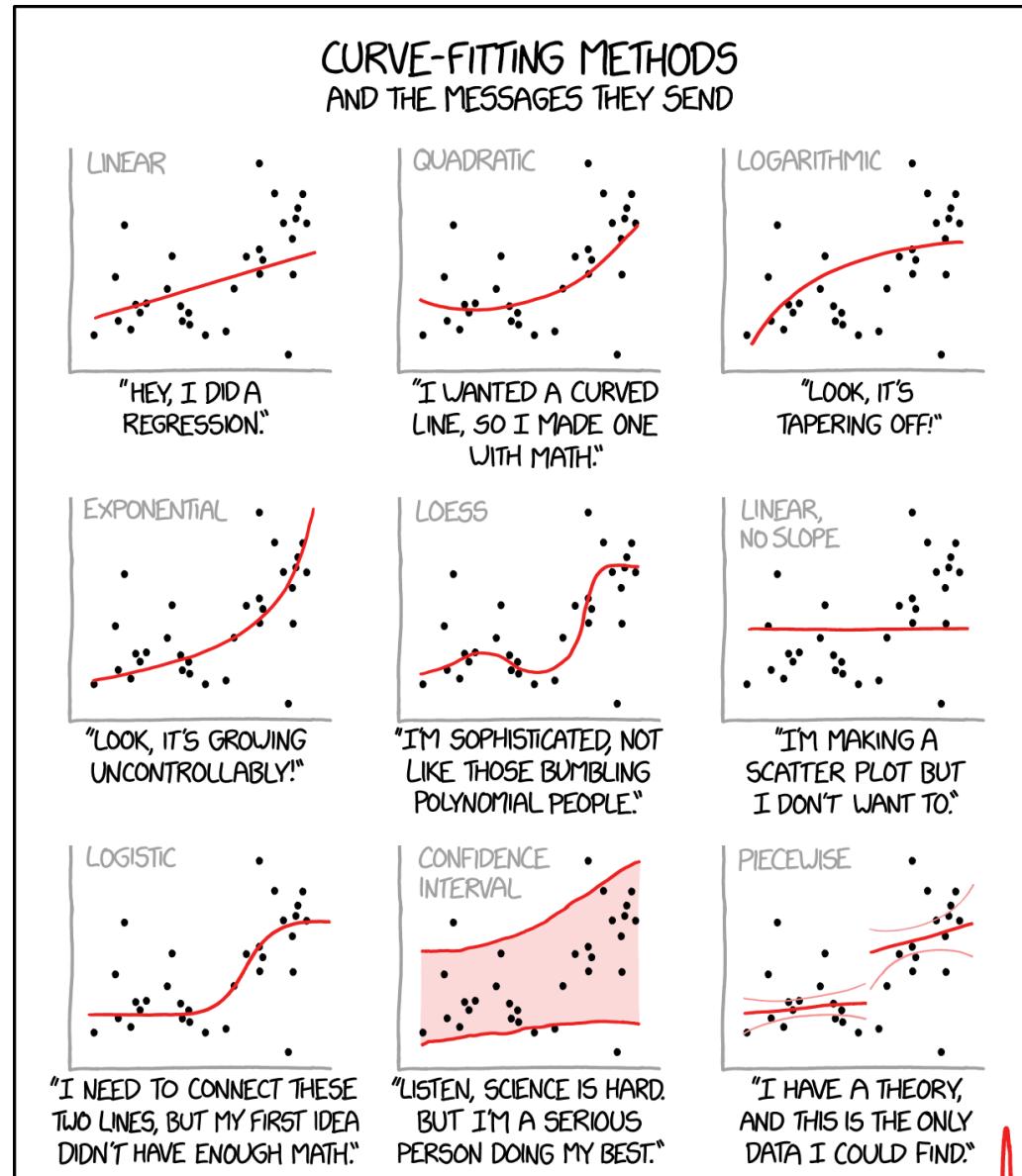
Regression models

- **Linear** ($y = b + mx$);
- **Polynomial** ($y = b + mx + nx^2$);
- **Multiple Linear** ($y = b + m_1x_1 + m_2x_2 + m_3x_3 + \dots + m_nx_n$);
- **PCR** (Principal Component Regression);
- **PLS-R** (Partial Least Squares – Regression);
- etc...



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_k X_{ki} \quad i = 1 \dots n$$

OLS: Ordinary Least Squares Regression



OLS: Ordinary Least Squares Regression

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_k X_{ki} \quad i = 1 \dots n$$

Response –
predicted/estimated
value
(dependent variable)

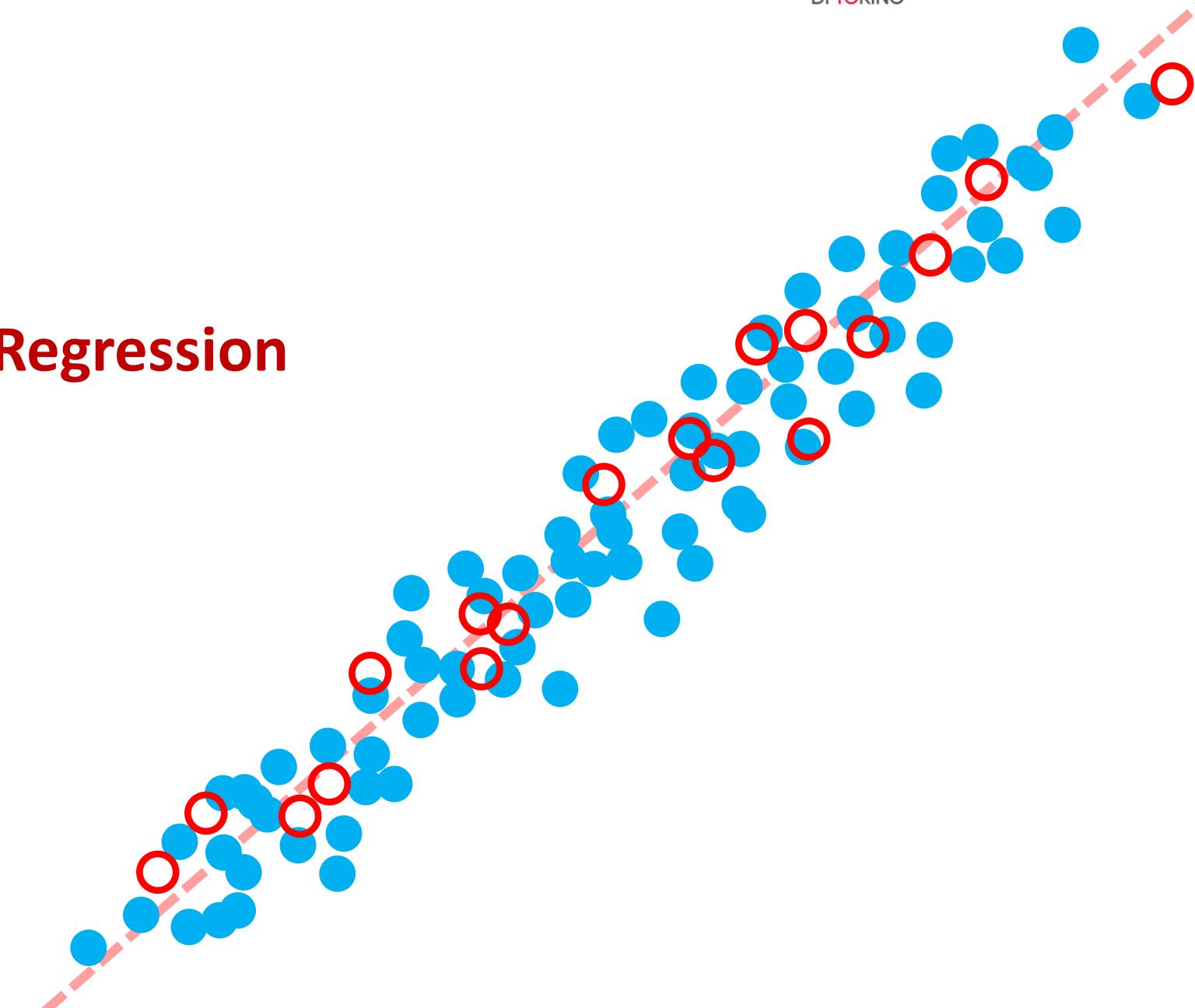
Intercept
($X=0$)

Predictors/features/
attributes/variables
(measured,
independent)

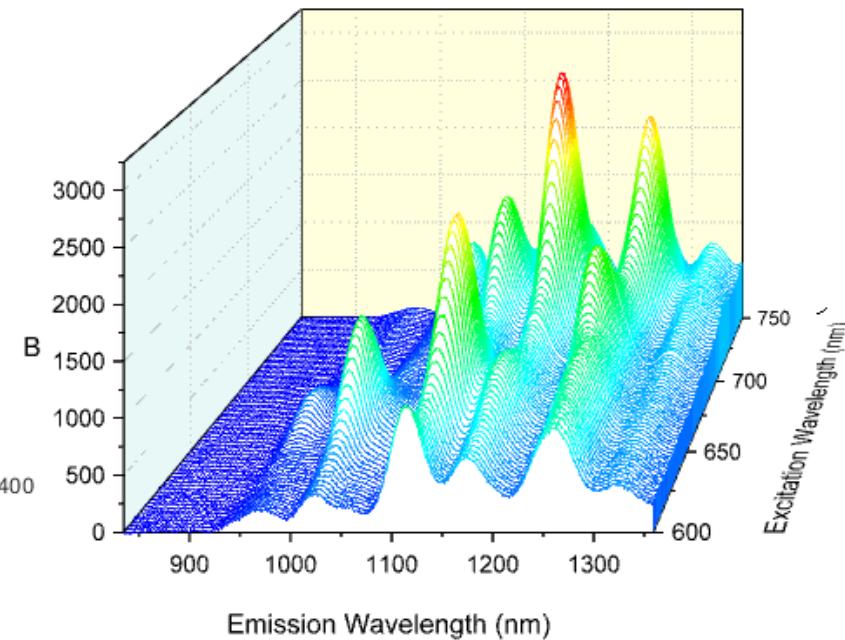
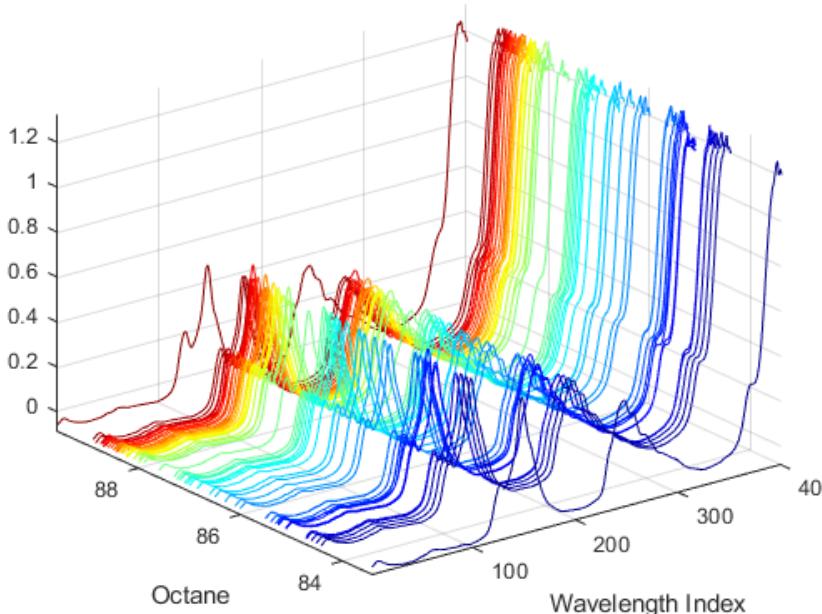
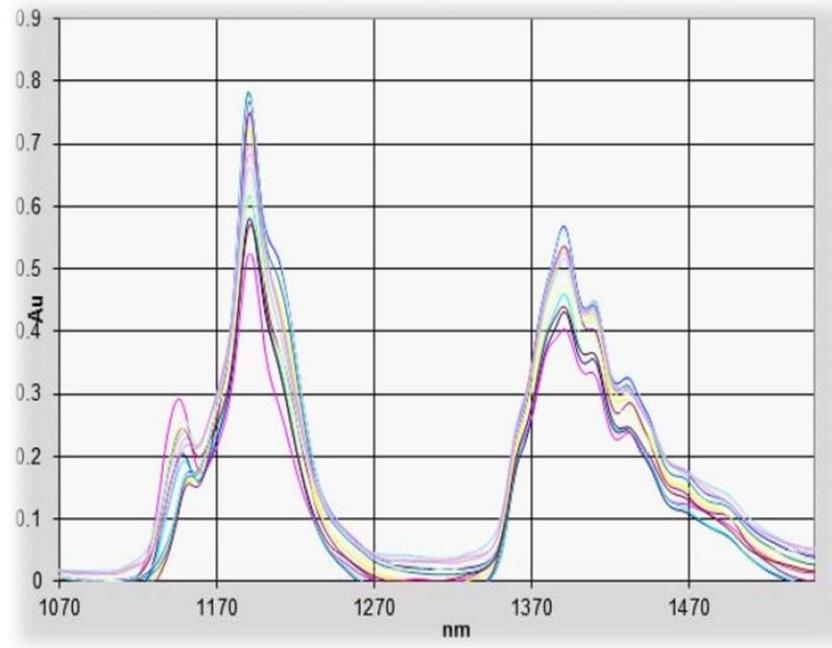
Regression coefficient
(slopes) for each
predictor

ML regression models

Multivariate Regression



Supervised ML - Regression



*The selectivity problem – too much information!
It is necessary to use a multivariate approach*

Supervised ML - Regression

The traditional OLS approach

LS - MLR

...as well as Multiple Linear Regression...

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

...it cannot be used when you have many variables!

Supervised ML - Regression

MLR provides:

- predicted values
- regression coefficients
- "diagnostic" graphs

But if there are **many correlated/collinear variables** (linear combination of the original variables):

- **MLR provides unstable regression equations;**
- they are very complex to process → **their use is no longer needed.**

Supervised ML - Regression

Problems with MLR

- MLR is conceptually simple and generalizes univariate LS regression.
BUT
- The core of MLR is the calculation of the model parameters by the LS approach, according to:

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Due to the term $(\mathbf{X}^T \mathbf{X})^{-1}$, it can't be applied to ill-conditioned matrices:
 - Correlated variables
 - More variables than samples



- Use of Latent variables (**BILINEAR MODELING**):
 - Few
 - Orthogonal

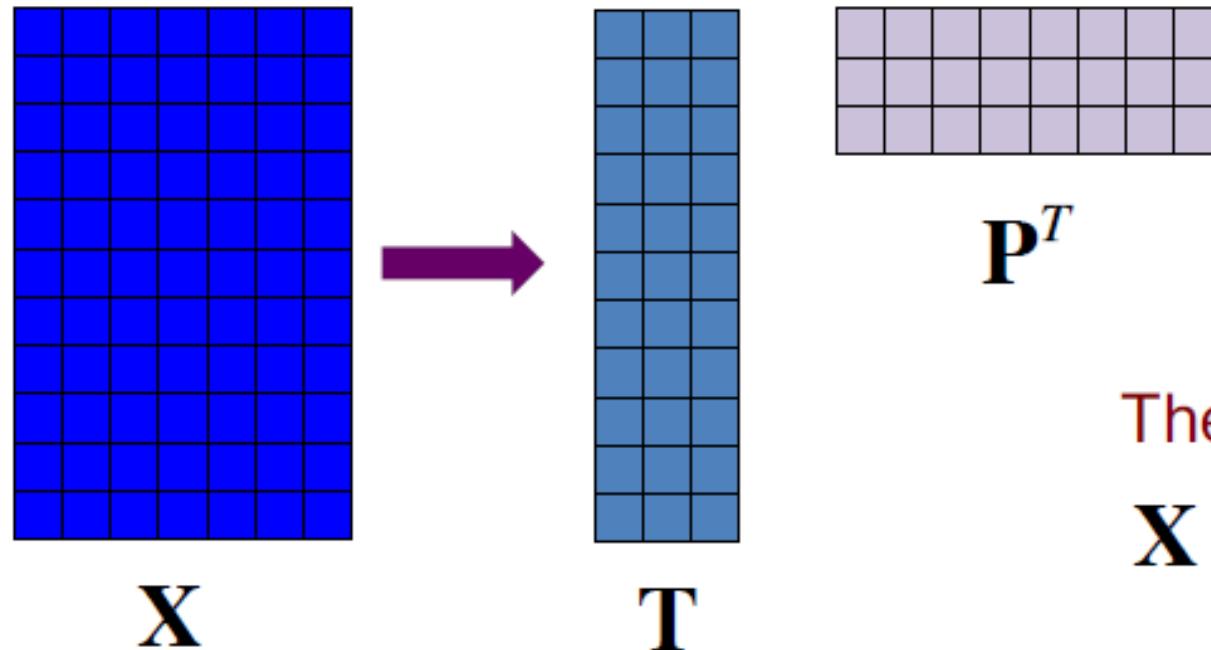
Supervised ML - Regression

- Select the variables that are most significant (**stepwise methods**)
- Compress the variables and direct them towards the "zones" of the most relevant data:
 - PCR – Principal Component Regression;
 - PLS – Partial Least-Squares Regression.

Principal Component Regression (PCR)

2-steps procedure:

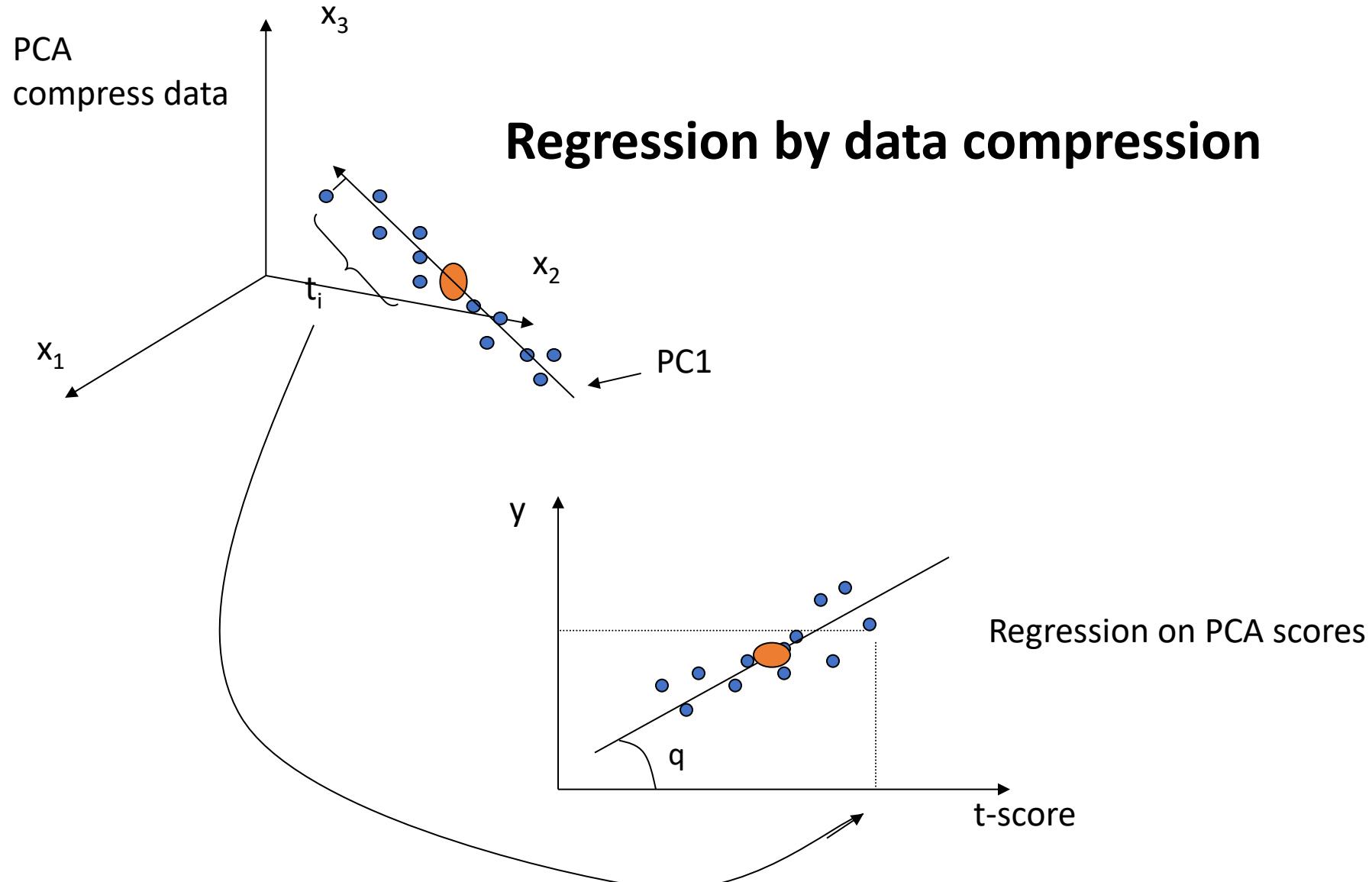
1. Build a **PCA** model;
2. Doing **MLR** on PCA scores.



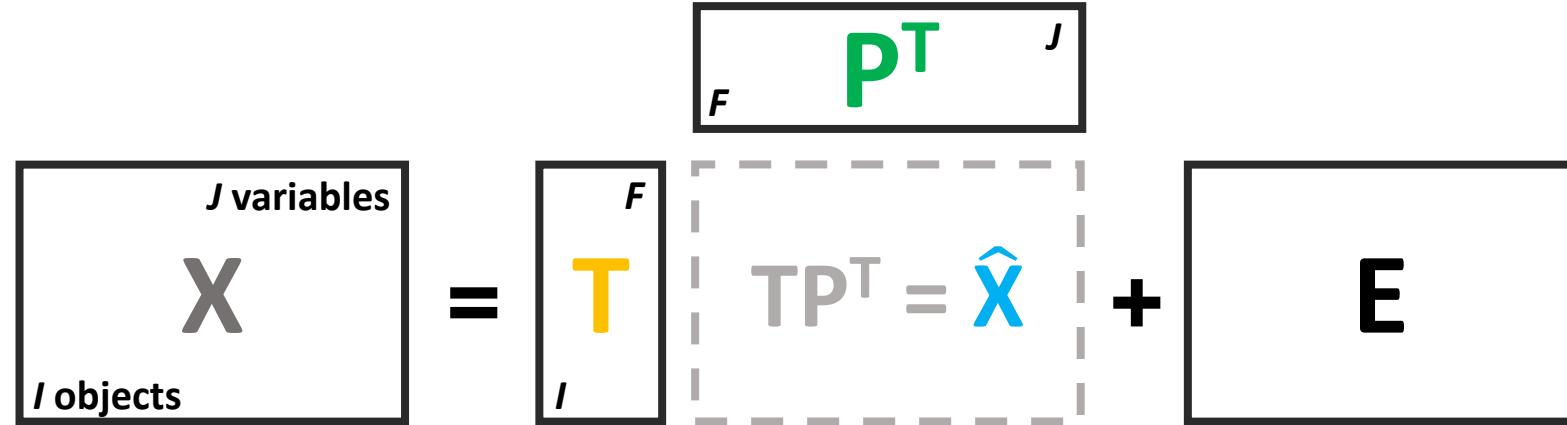
The PCA model:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_X$$

Principal Component Regression (PCR)



Principal Component Regression (PCR)



→ using PCA **T scores**, calculate the **b regression coefficients** (a vector!)

$$X = T \cdot P^T + E$$

$$b = (T^T T)^{-1} T^T y \quad (\text{predict } \underline{\text{one}} \text{ property})$$

$$B = (T^T T)^{-1} T^T Y \quad (\text{predict } \underline{\text{more}} \text{ properties})$$

Principal Component Regression (PCR)

A PCA model:

- is very simple
- reduces the dimensionality of X (data compression!)
- eliminates the collinearity on X (as it is included into the PCs)

→ using PCA **T scores**, calculate the **b regression coefficients** (a vector!)

$$X = T \cdot P^T + E$$

$$b = (T^T T)^{-1} T^T y \quad (\text{predict } \underline{\text{one}} \text{ property})$$

$$B = (T^T T)^{-1} T^T Y \quad (\text{predict } \underline{\text{more}} \text{ properties})$$

Principal Component Regression (PCR)

$$Y = TB + E$$

Just like in the MLR case, we want to obtain a **model** in the form of an equation, with which y predicted values can be obtained from new samples x (computing T, first!).

→ using PCA **T scores**, calculate the **b regression coefficients** (a vector!)

$$X = T \cdot P^T + E$$

$$b = (T^T T)^{-1} T^T y \quad (\text{predict } \underline{\text{one}} \text{ property})$$

$$B = (T^T T)^{-1} T^T Y \quad (\text{predict } \underline{\text{more}} \text{ properties})$$

Principal Component Regression (PCR)

$$Y = TB + E$$

Just like in the MLR case, we want to obtain a **model** in the form of an equation, with which y predicted values can be obtained from new samples x (computing T, first!).

However, the **main limitation of PCR** is that the coefficients are computed starting from a PCA model, which *by definition* only describes...

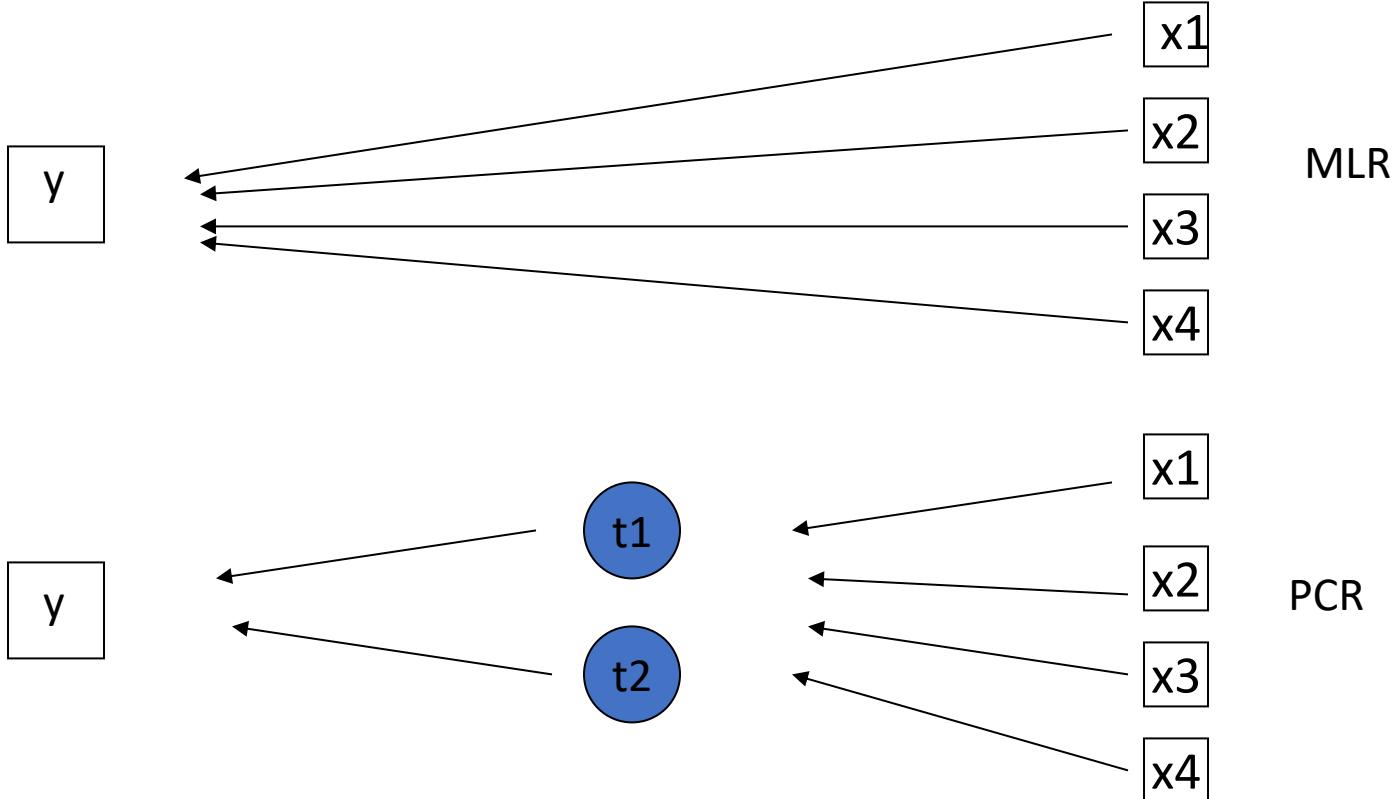
the variability of the original data X, independently from the variability of Y.

Therefore, it is not assumed that what it is modelled by the PCA model would be correlated with the response Y!

Principal Component Regression (PCR)

- Use the main components;
- Solves the problem of collinearity, providing stable results;
- Provides graphs for interpretation (scores and loadings);
- Easily explainable;
- Diagnostics on outliers;
- Easily adaptable and editable;
- **But uses only X to determine components**

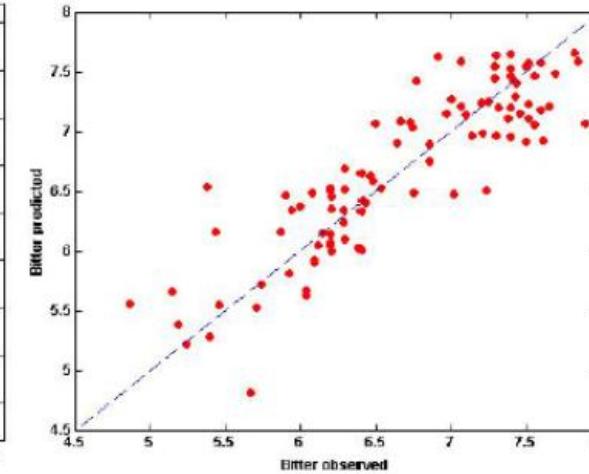
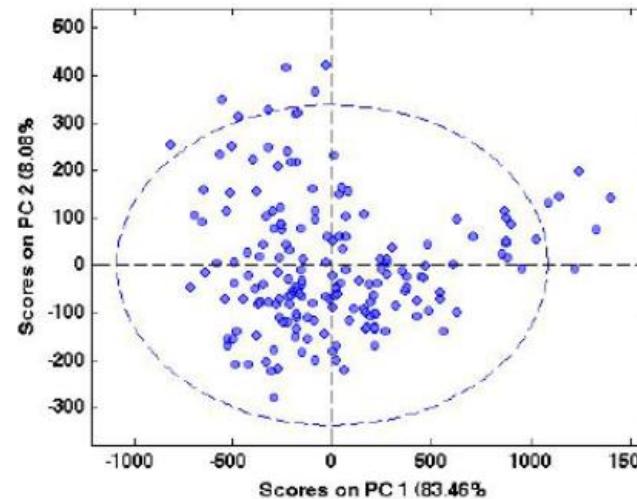
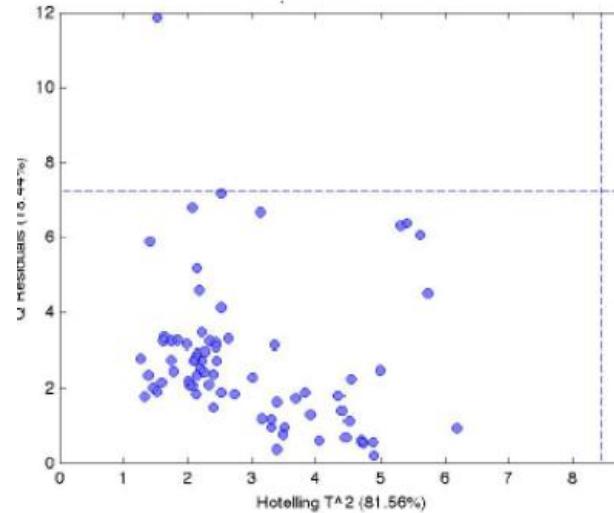
Principal Component Regression (PCR)



Principal Component Regression (PCR)

Considerations

- What do we get:



- **DRAWBACK:** Information in \mathbf{Y} is not actively used for the definition of the scores:
 - Maximum explained variance doesn't necessarily mean maximum correlation with \mathbf{Y} .

Partial Least-Squares (PLS) Regression

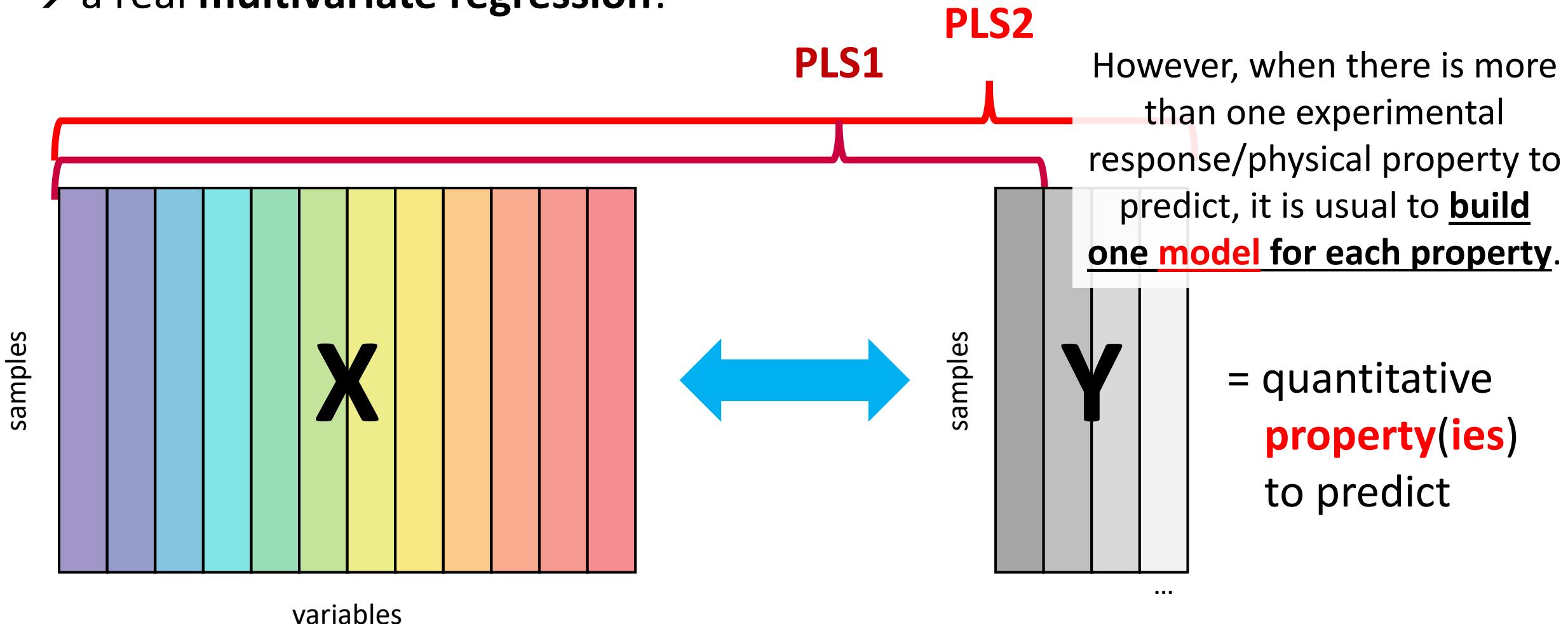
For each factor/component/latent variable

- PCR
 - Maximizes the **variance** of linear combinations of \mathbf{X}
- PLS
 - Maximizes the **covariance** of linear combinations of \mathbf{X} and \mathbf{Y}

Each factor is subtracted before moving on to
the calculation of the next factor

Partial Least-Squares (PLS) Regression

→ a real multivariate regression!

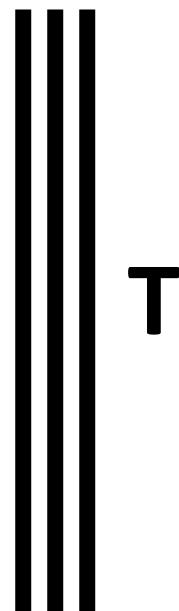
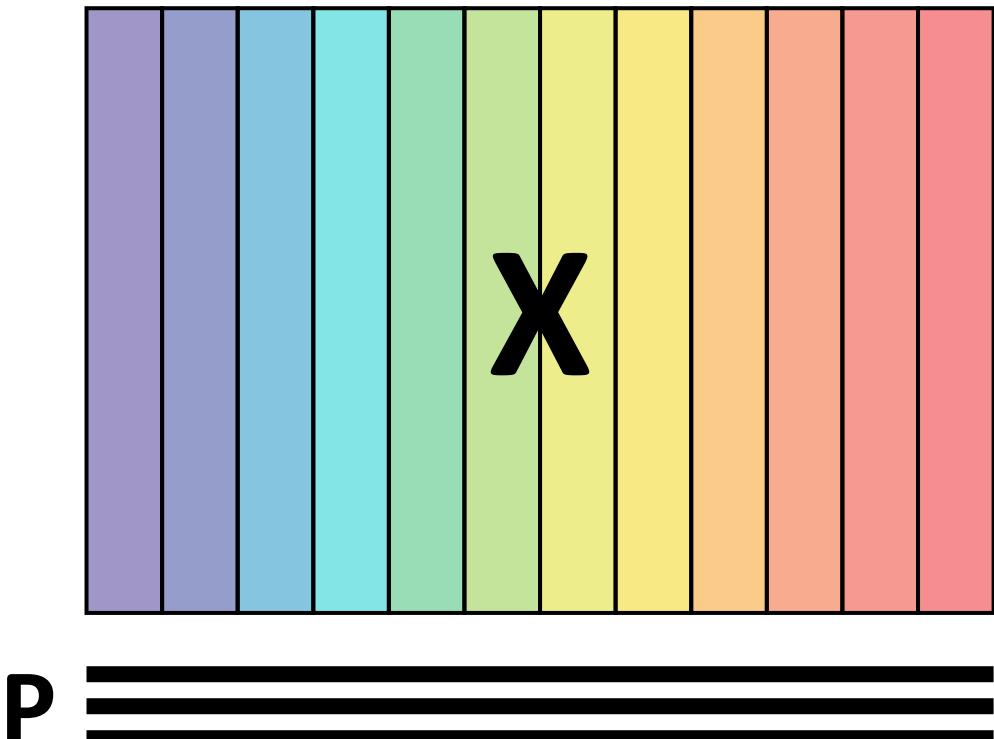


Partial Least-Squares (PLS) Regression

→ correlation between X and Y

by maximizing the COVARIANCE between t and u

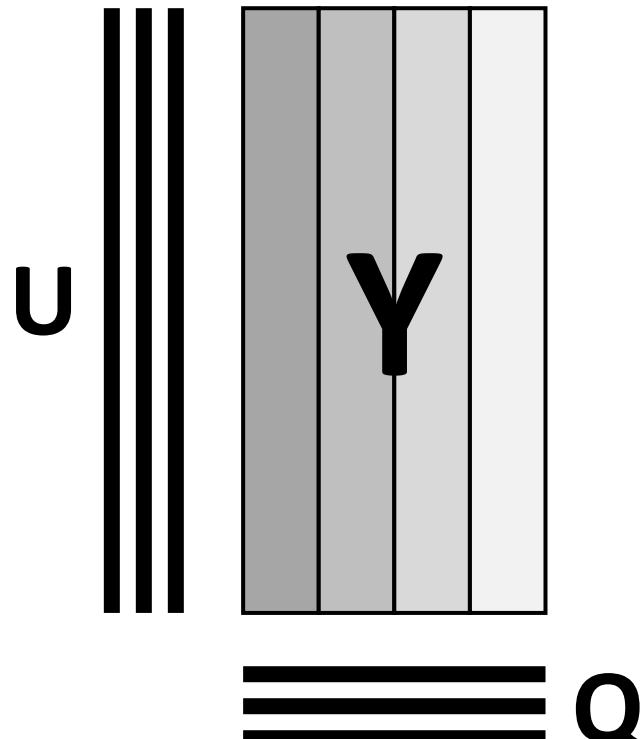
t and u carry the information about the samples!



$$u_1 = r_1 t_1$$
$$u_2 = r_2 t_2$$

...

r = scores regressors



$$Y = UQ^T + F$$

Partial Least-Squares (PLS) Regression

→ !!! PLS is NOT two individual PCA models for X and Y

→ PCA and PLS may look the same, but:

PCA seeks for **VARIANCE** to model X

PLS seeks for **COVARIANCE** between X and Y

regression
coefficients

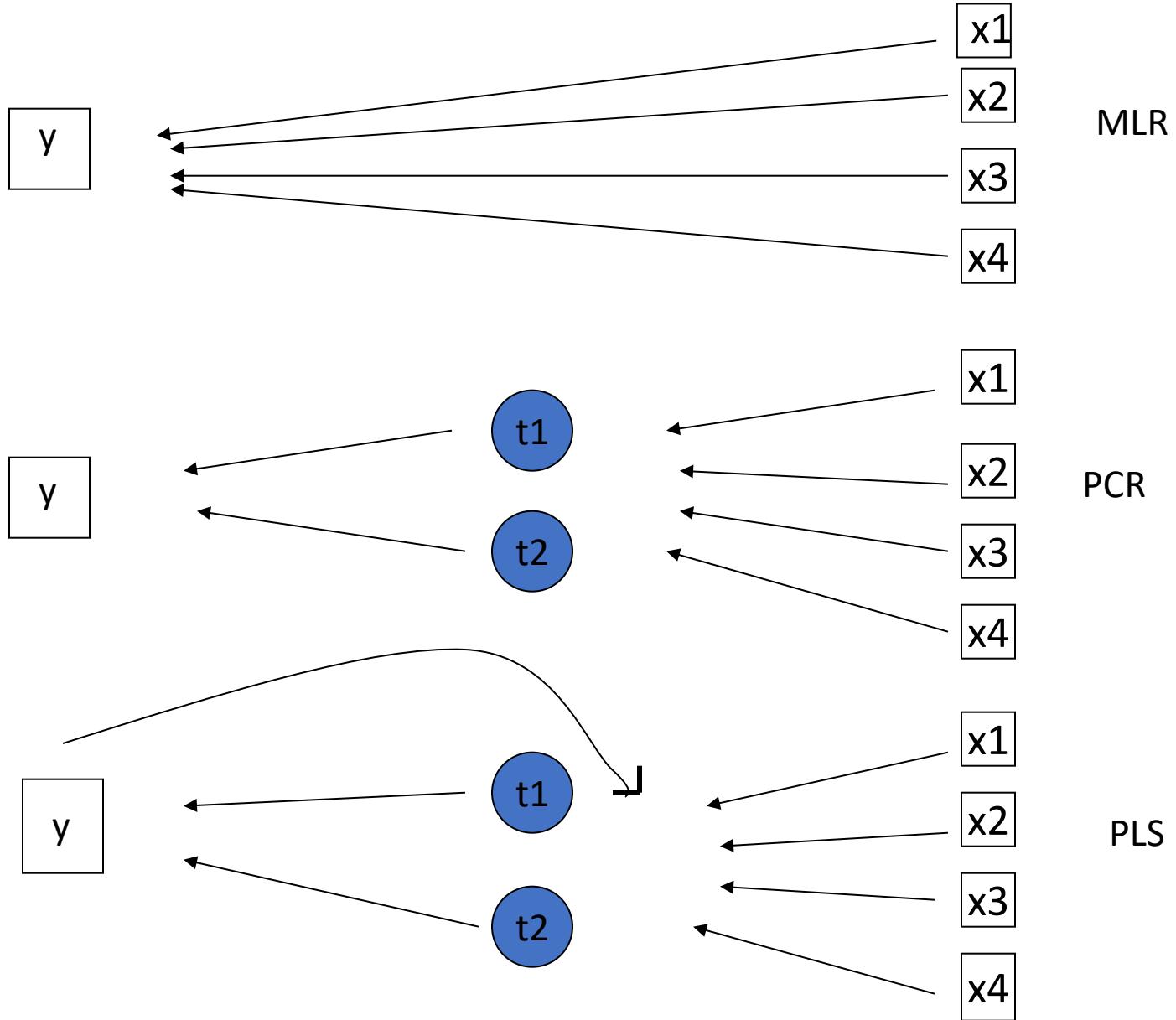
$$\left. \begin{array}{l} \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \end{array} \right\} \quad \mathbf{U} = \mathbf{R}\mathbf{T}$$



$$\boxed{\mathbf{Y} = \mathbf{X}\mathbf{B}}$$

another model!

Partial Least-Squares (PLS) Regression



Partial Least-Squares (PLS) Regression

PLS

$$\left. \begin{array}{l} X = T P^T + E \\ y = T_q + f \end{array} \right\} \text{ CENTERED } X \text{ AND } y$$

SET $a=1$

(i) MAXIMIZE COVARIANCE BETWEEN
 $t_a = X w_a$ AND y

(ii) FIND p_a AND q_a AND SUBTRACT THE
FACTOR

$$\begin{aligned} X - t_a p_a^T &\rightarrow X \\ y - t_a q_a &\rightarrow y \end{aligned}$$

Single y PLS-R

$$\max_w (\text{cov}(t, y)) \quad \text{with: } t = Xw$$

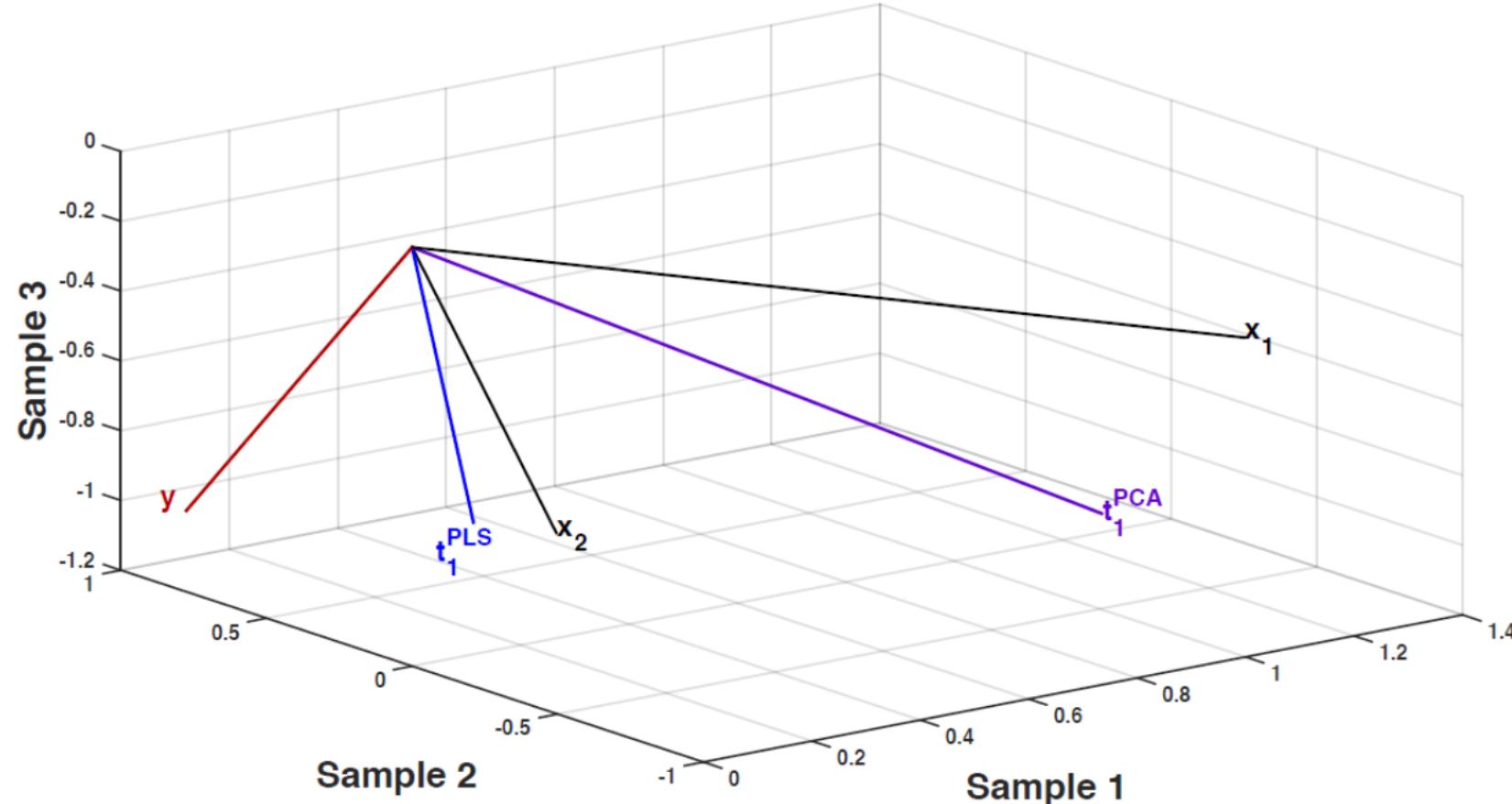
INCREASE a BY 1 AND CONTINUE
UNTIL $a = A$

Partial Least-Squares (PLS) Regression

- Easy to calculate;
- Provides stable solutions;
- Provides scores and loadings;
- Often the number of components (**latent variables**) is lower than those calculated with PCR;
- Better predictions can be made;
- Can be used for **more than one Y response simultaneously (PLS2)**.

Partial Least-Squares (PLS) Regression

The PLS criterion graphically explained



- With respect to PCA, scores are «rotated» towards y