BASICS OF CHEMOMETRICS

François Stevens
Juan Antonio Fernández Pierna

Walloon Agricultural Research Centre (CRA-W),
Valorization of Agricultural Products Department
Gembloux, Belgium

Basic definition of chemometrics

Application of mathematical and statistical methods to chemical measurements[1].

X-metrics

Bio-metrics → Biology

Psycho-metrics → Psychology

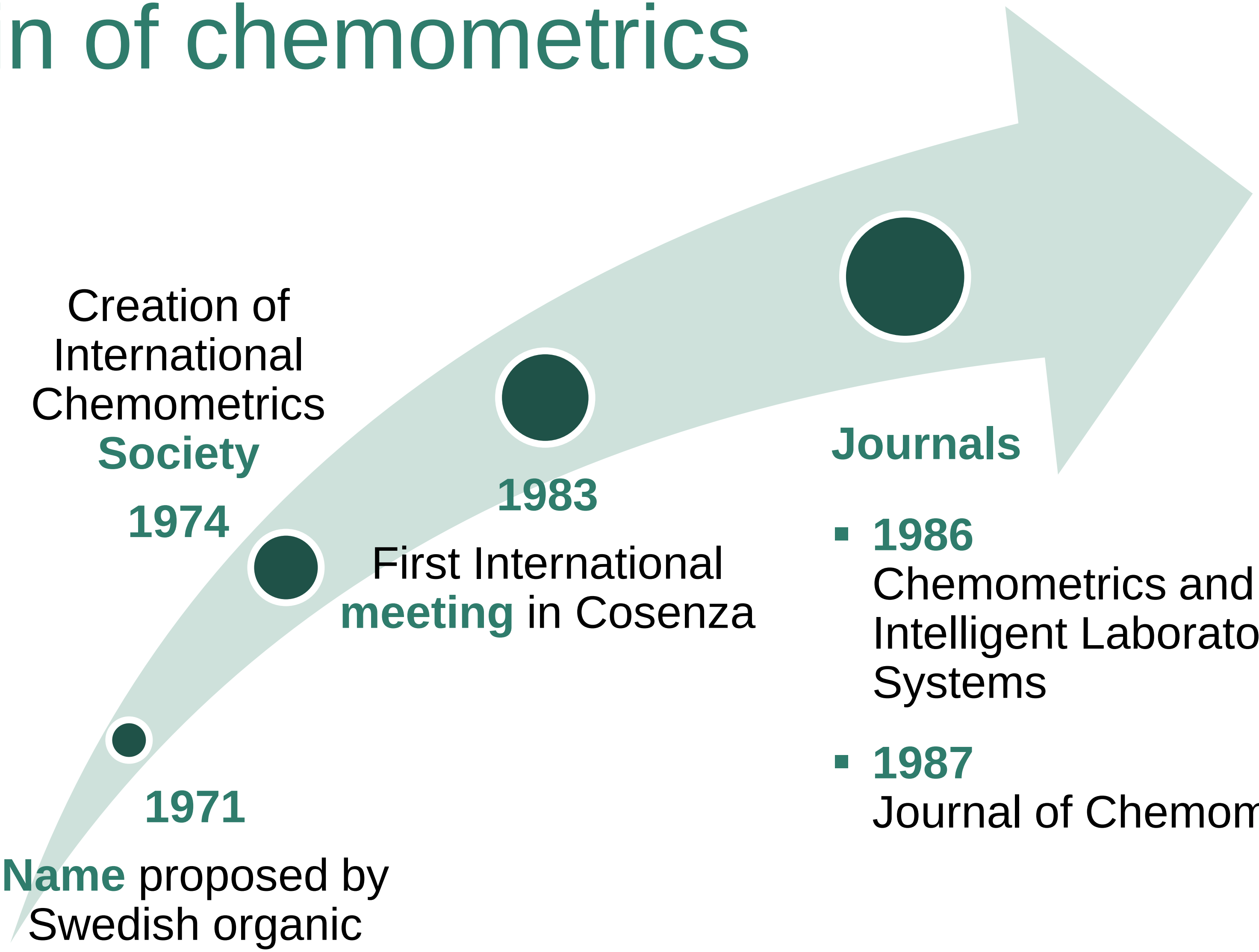
Chemo-metrics → Chemistry

[1] Kowalski, Anal. Chem. 1980, 52, 112R-122R

Historical origin of chemometrics

- Routine use of spectrometers for chemical analyses
- Rising application of multivariate statistics

≳ 1950



1971

Name proposed by Swedish organic chemist **Svante Wold**

Creation of International Chemometrics **Society**

1974

1983

First International **meeting** in Cosenza

Journals

- **1986** Chemometrics and Intelligent Laboratory Systems
- **1987** Journal of Chemometrics

Context of the last decades

Increase of the amount, quality and accessibility of **instruments**

Evolution of **computers** allowing faster acquisition and processing

Development of new tools and approaches

CHEMOMETRICS

Explosive growth of the amount and quality of **data**

Development of chemometric **softwares** and toolboxes

More complete definition

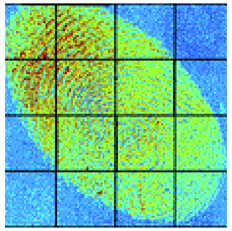
Chemical discipline that uses mathematics, statistics and formal logic to

1. design or select optimal experimental procedures
2. provide maximum relevant chemical information by analyzing chemical data
3. obtain knowledge about chemical systems.

Massart, D.L., et al. (1997) Data Handling in Science and Technology 20A, Handbook of Chemometrics and Qualimetrics Part A, p1.

Rapid development in multiple domains

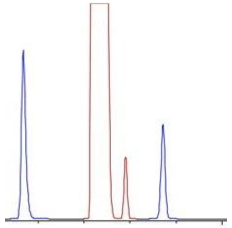
Process control and analysis



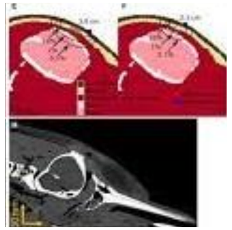
Forensic science

Reaction monitoring

Chromatographic optimisation

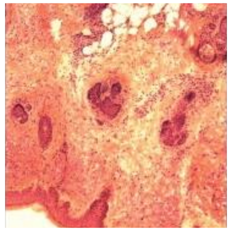


Biology Omics



Analytical Chemistry

Environmental monitoring

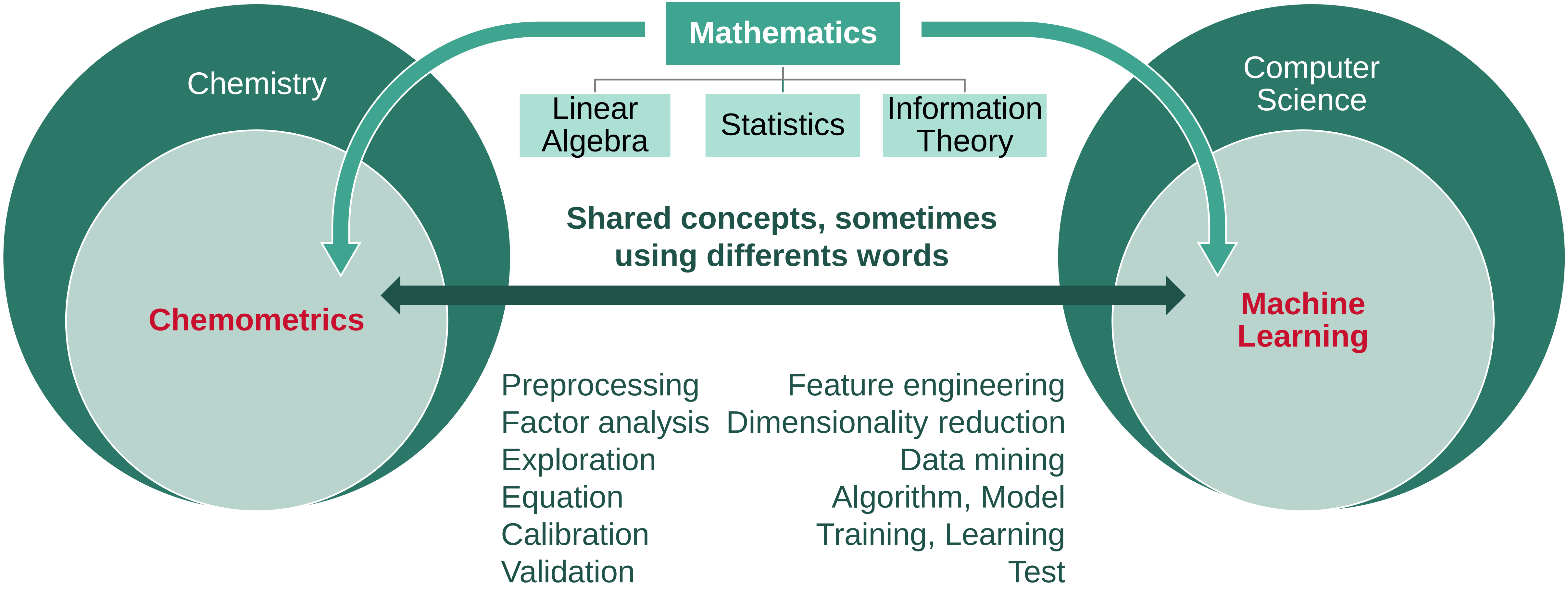


Clinical Science

Food analysis

and many others …

Chemometrics and machine learning



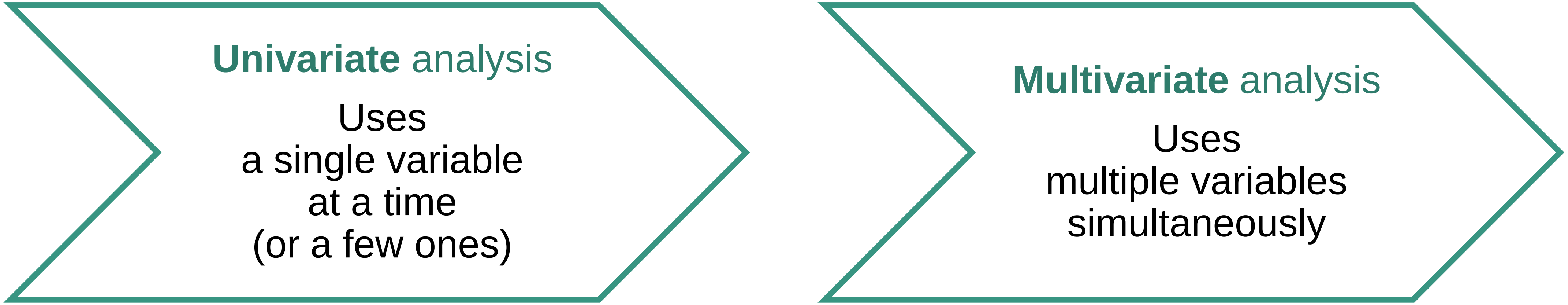
Shared concepts, sometimes using differents words

Preprocessing	Feature engineering
Factor analysis	Dimensionality reduction
Exploration	Data mining
Equation	Algorithm, Model
Calibration	Training, Learning
Validation	Test

Evolution of the methods in chemometrics

- The definition of chemometrics is traditionnaly associated with multivariate linear statistics
  - Multiple Linear Regression (MLR)
  - Principal Component Analysis (PCA)
  - Partial Least Squares (PLS)
  - …
- However, methods from the field of machine learning are now also considered as part of chemometrics:
  - Support Vector Machines (SVM)
  - Classification And Regression Trees (CART)
  - Artificial Neural Networks (ANN)
  - …

From univariate to multivariate analysis



Univariate analysis

Uses
a single variable
at a time
(or a few ones)

Multivariate analysis

Uses
multiple variables
simultaneously

- Reflectance at single spectral wavelength
- Height of one peak
- Spectral indices
- Integrated signal over spectral band

Full spectrum or spectral interval

Role of linear algebra (matrix and vector operations)

Linear algebra is the language of Chemometrics.

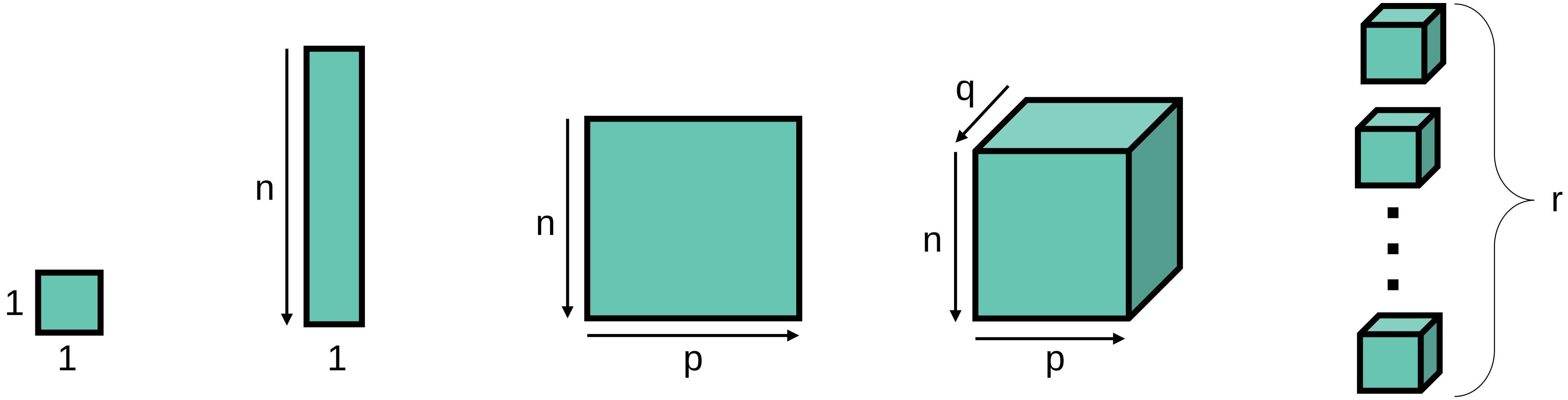
One cannot expect to truly understand most chemometric techniques without a basic understanding of linear algebra

Wise and Gallagher, 1998

Our objective of today

Grasp the fundamental principles of chemometrics *without equations* !

Typical structures of chemical data



Type	Scalar	Vector	2D Matrix	3D Matrix « hypercube »	4D Matrix
Size	1 × 1	n × 1	n × p	n × p × q	n × p × q × r
Example	Fixed room temperature	Reference values for one property	Matrix of spectra Reference values for multiple properties	Hyperspectral image Matrix of spectra at different timepoints 3D LC-MS plot	Hyperspectral video

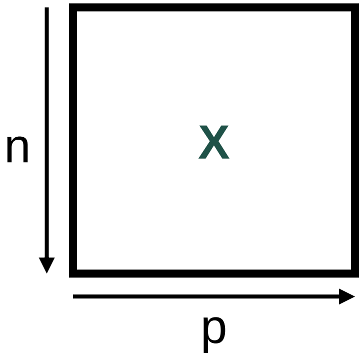
The data matrix (2D case)

ROWS

Samples

Observations

OBJECTS



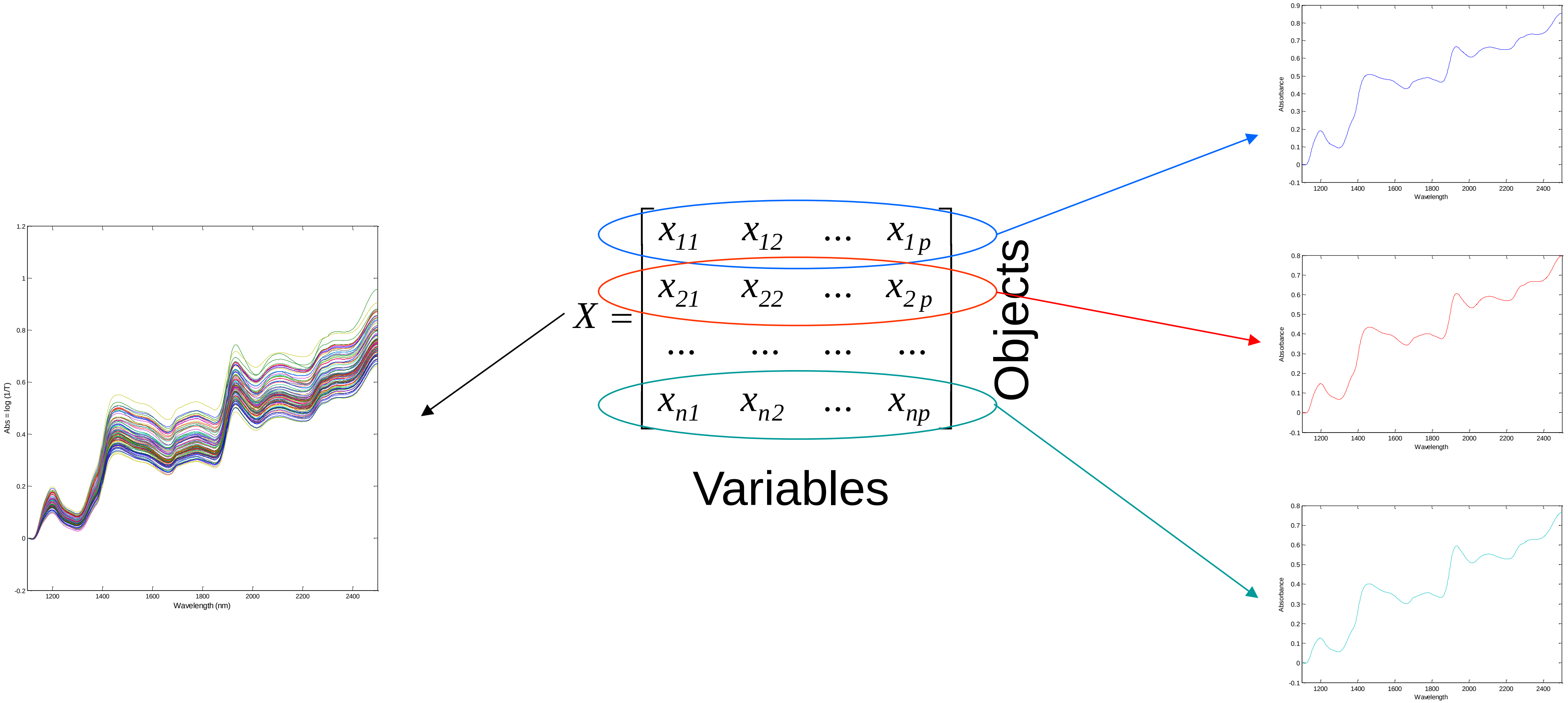
COLUMNS

Spectral variables

Wavelengths

VARIABLES

The spectral matrix in spectroscopy

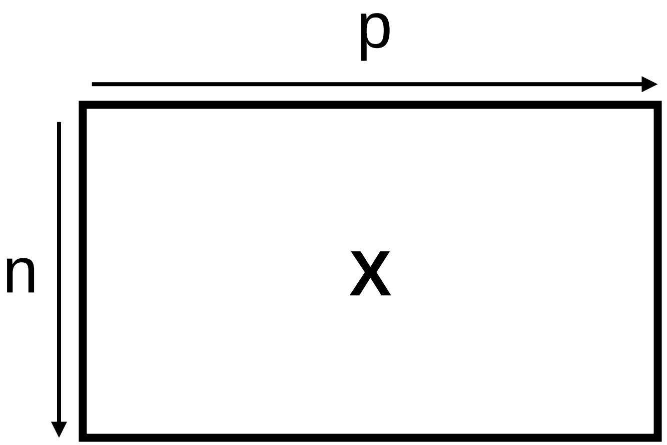


$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

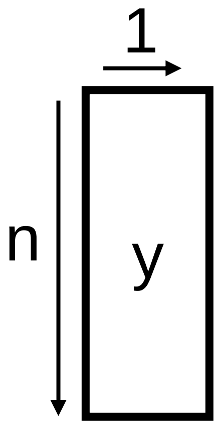
Objects

Variables

Data matrix and reference values



Data matrix

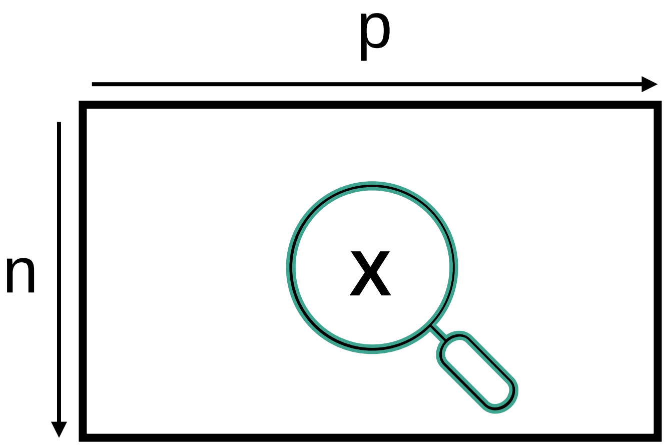


Reference values

- *Experimental* data
- Typically a matrix of spectra from vibrational spectroscopy or hyperspectral imaging

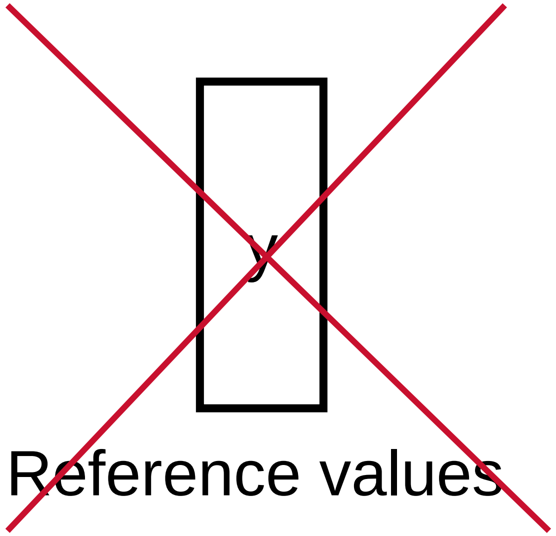
- Values of a given property for each object, considered as *ground truth*
- Generally obtained from reference methods such as wet chemistry or mass spectrometry
- May also come from visual observation or known metadata (country of origin, species, variety, …)

Unsupervised approaches



Data matrix

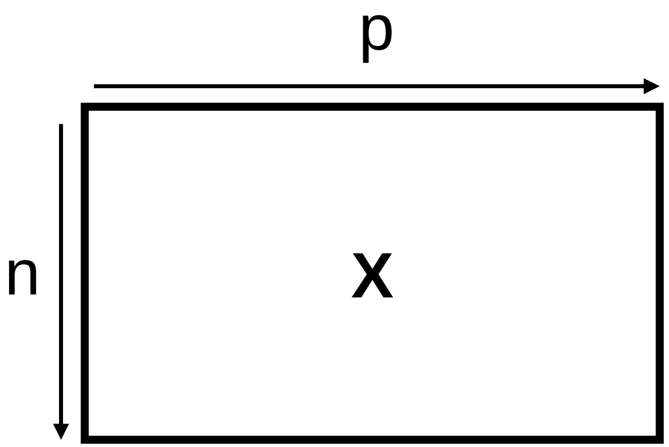
Analysed using data exploration methods



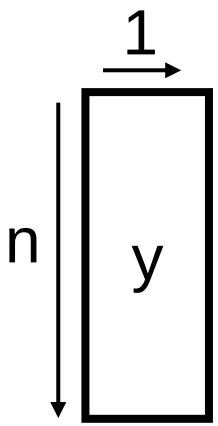
Reference values

Not available or not exploited

Supervised approaches



Data matrix



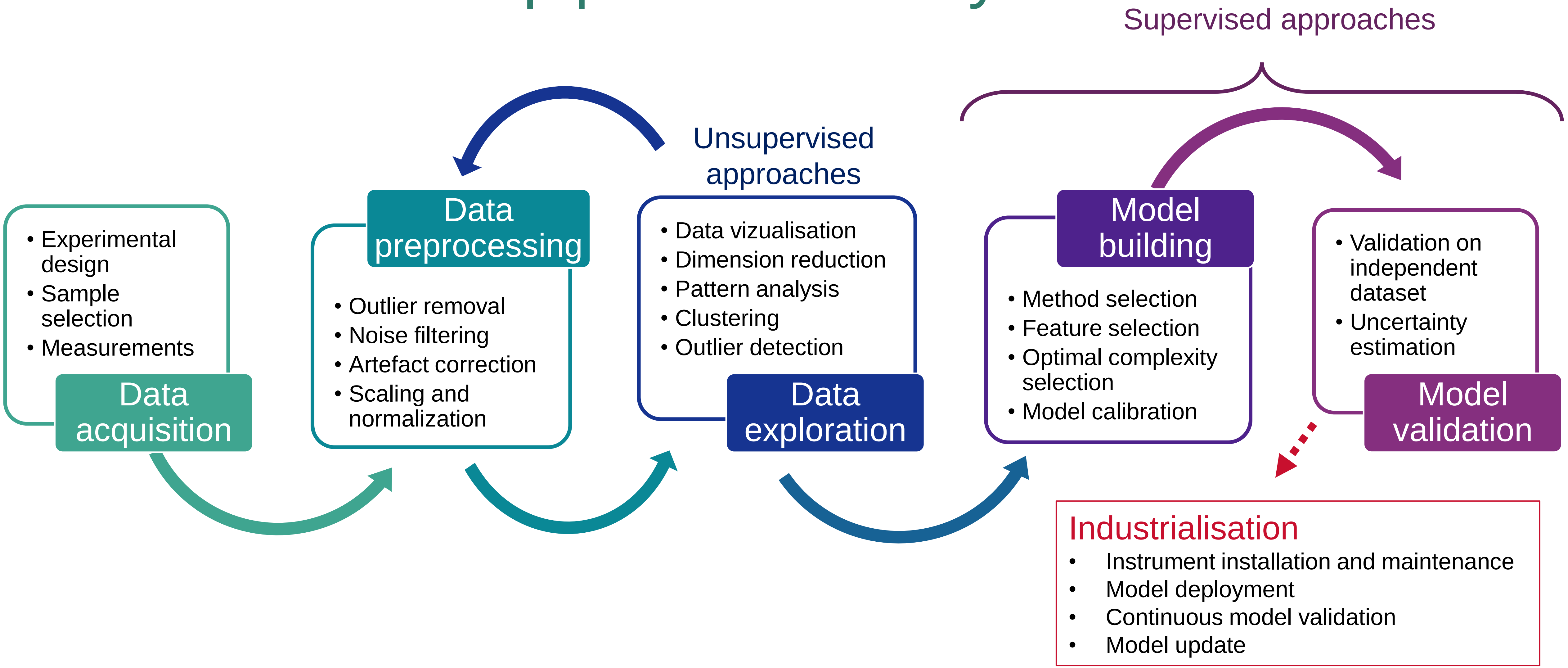
Reference values

Used as *explanatory variables…*

Used as *response variable…*

… in a predictive model (regression or classification)

Chemometric pipeline of analysis



Supervised approaches

Unsupervised approaches

Data acquisition

- Experimental design
- Sample selection
- Measurements

Data preprocessing

- Outlier removal
- Noise filtering
- Artefact correction
- Scaling and normalization

Data exploration

- Data vizualisation
- Dimension reduction
- Pattern analysis
- Clustering
- Outlier detection

Model building

- Method selection
- Feature selection
- Optimal complexity selection
- Model calibration

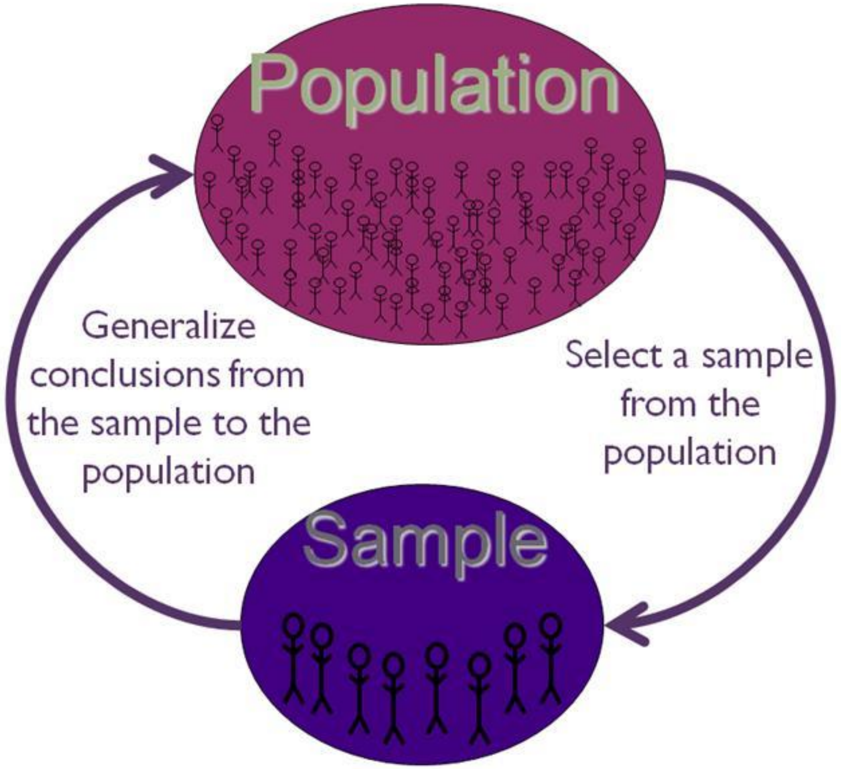
Model validation

- Validation on independent dataset
- Uncertainty estimation

Industrialisation

- Instrument installation and maintenance
- Model deployment
- Continuous model validation
- Model update

Data acquisition: sampling

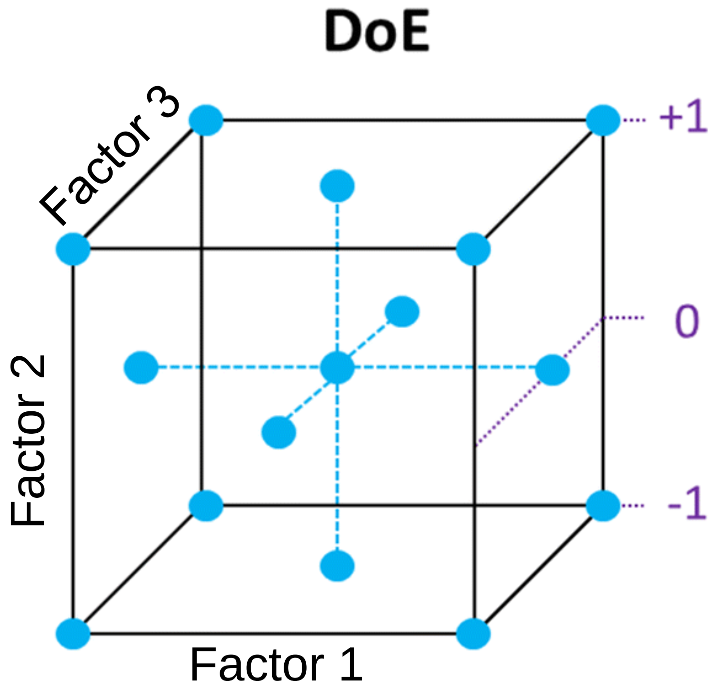

Population
Generalize conclusions from the sample to the population
Select a sample from the the population
Sample

Each case = a new challenge
Beware sample heterogeneity !!

Data acquisition: experimental design

When data are collected during a controlled experiment

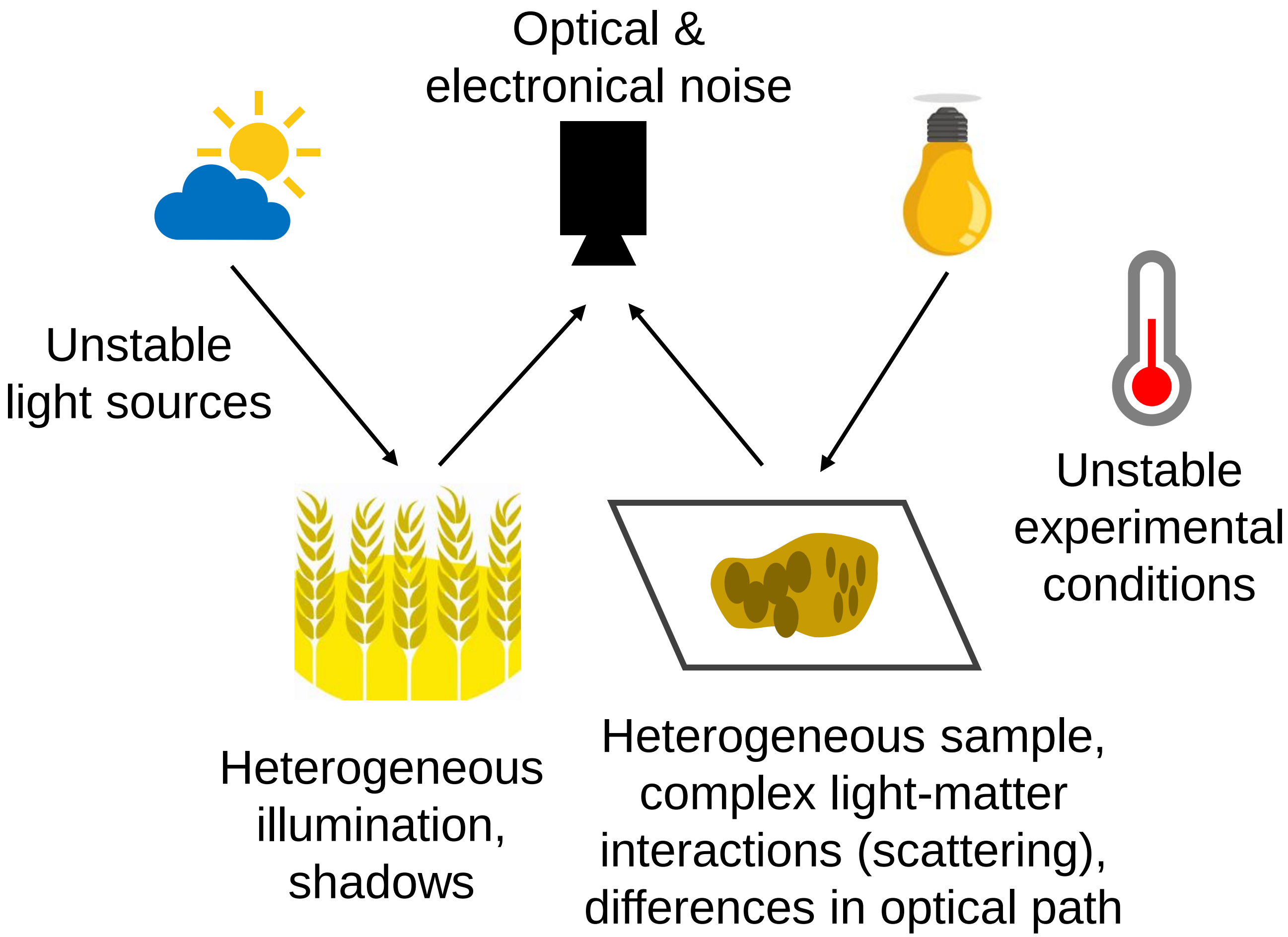
Objective Optimizing the coverage of all the factors of variation and their interactions, within constraints of experiment duration.



- Tip: if you want a robust predictive model, allow some variability in the acquisition of your training sample: different varieties, different storage conditions, different operators, …

Why is preprocessing required?

- The spectra contain relevant information and noise
- Noise is unwanted variation, artifacts resulting from different processes
- We need to remove as much noise as possible without altering relevant information



Optical & electronical noise

Unstable light sources

Unstable experimental conditions

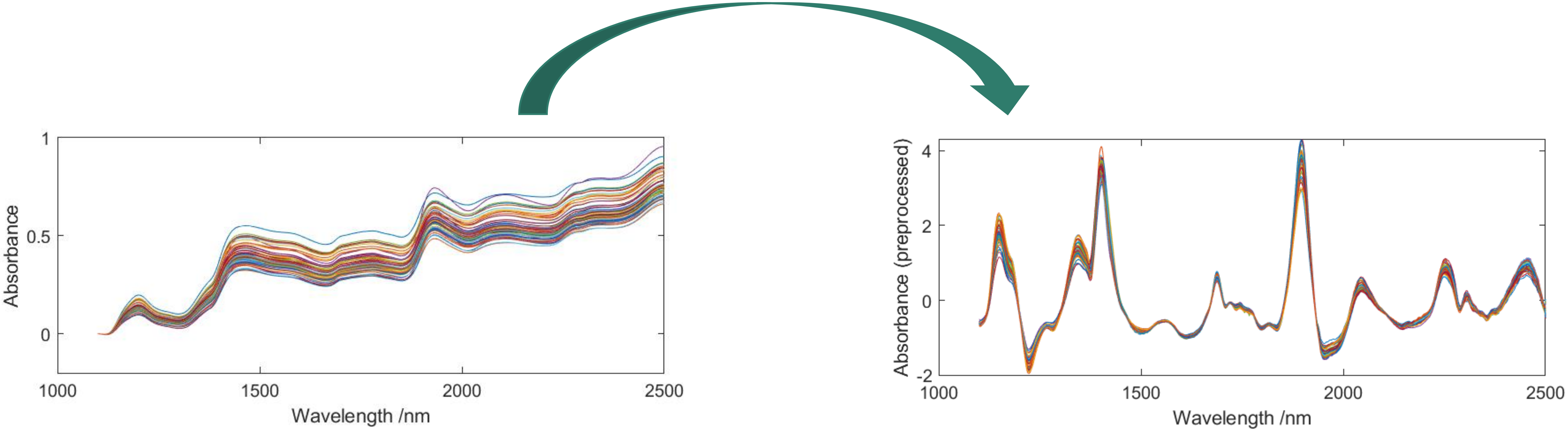
Heterogeneous illumination, shadows

Heterogeneous sample, complex light-matter interactions (scattering), differences in optical path

Example of preprocessing

Dataset of NIR spectra of wheat kernels

- Spectral derivative → highligths spectral bands
- SNV normalization → corrects for differences in optical path lengths



Data exploration: Principal Component Analysis (PCA)

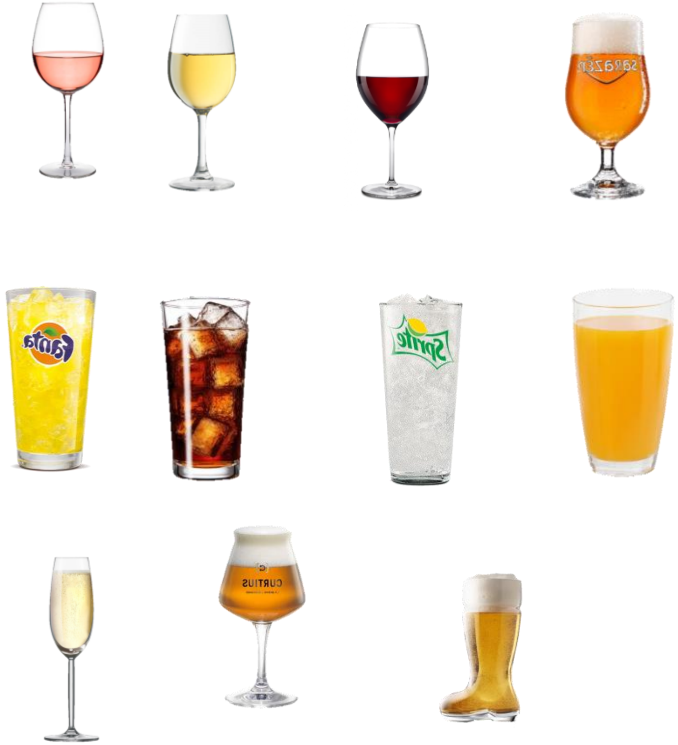
- With PCA, we create new variables (PC’s) as linear combinations of the original variables

- The PC’s are uncorrelated and ordered so that the first few retain most of the variation present in all the original variables

PCA: symbolic example

Objective

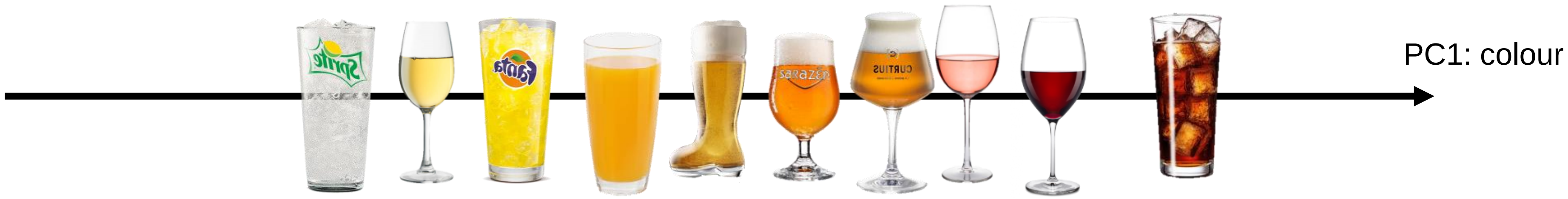
Find the factors capture the maximum of variability among these objects



PCA: symbolic example

Objective

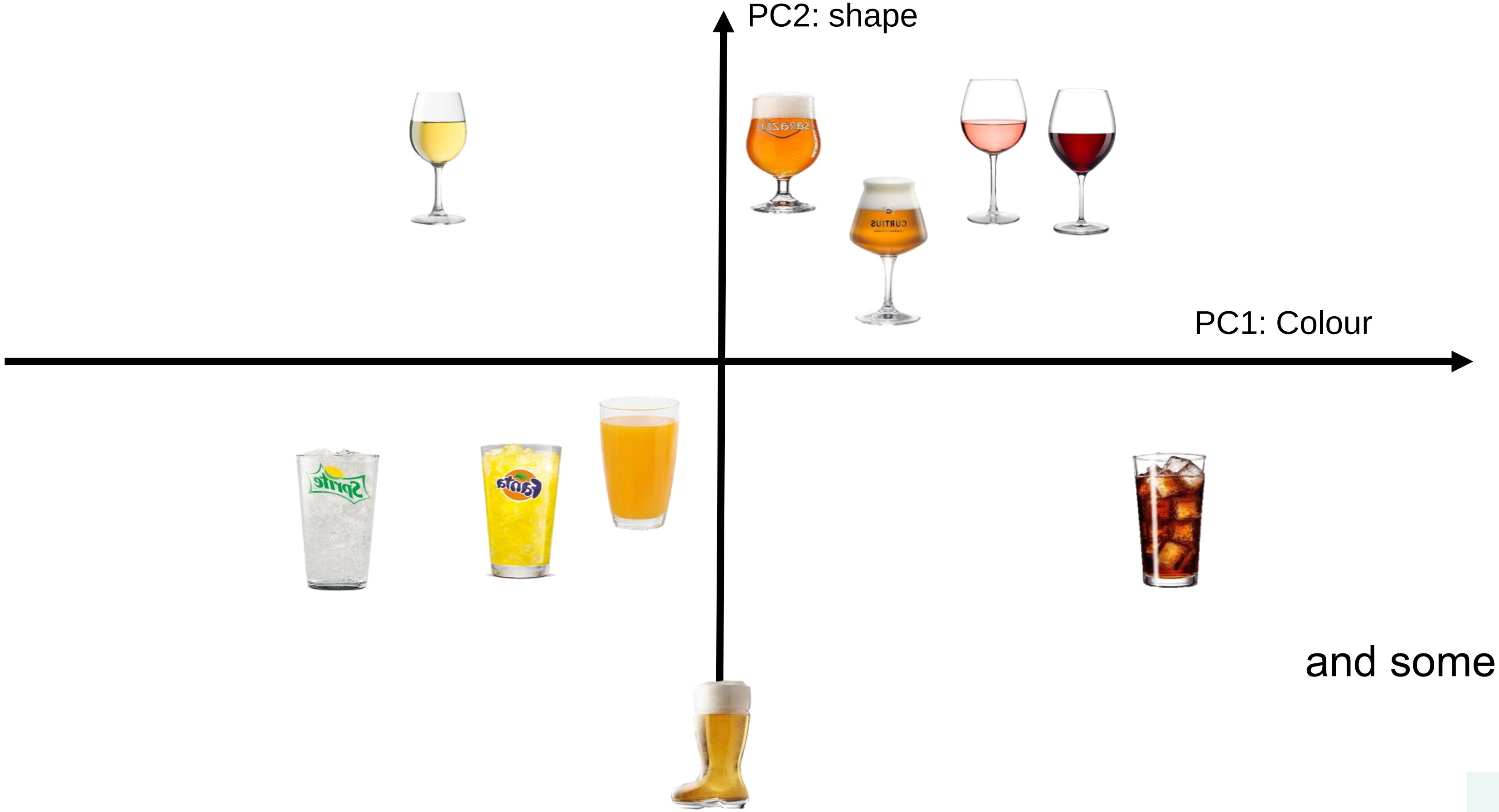
Find the factors capture the maximum of variability among these objects



PCA: symbolic example

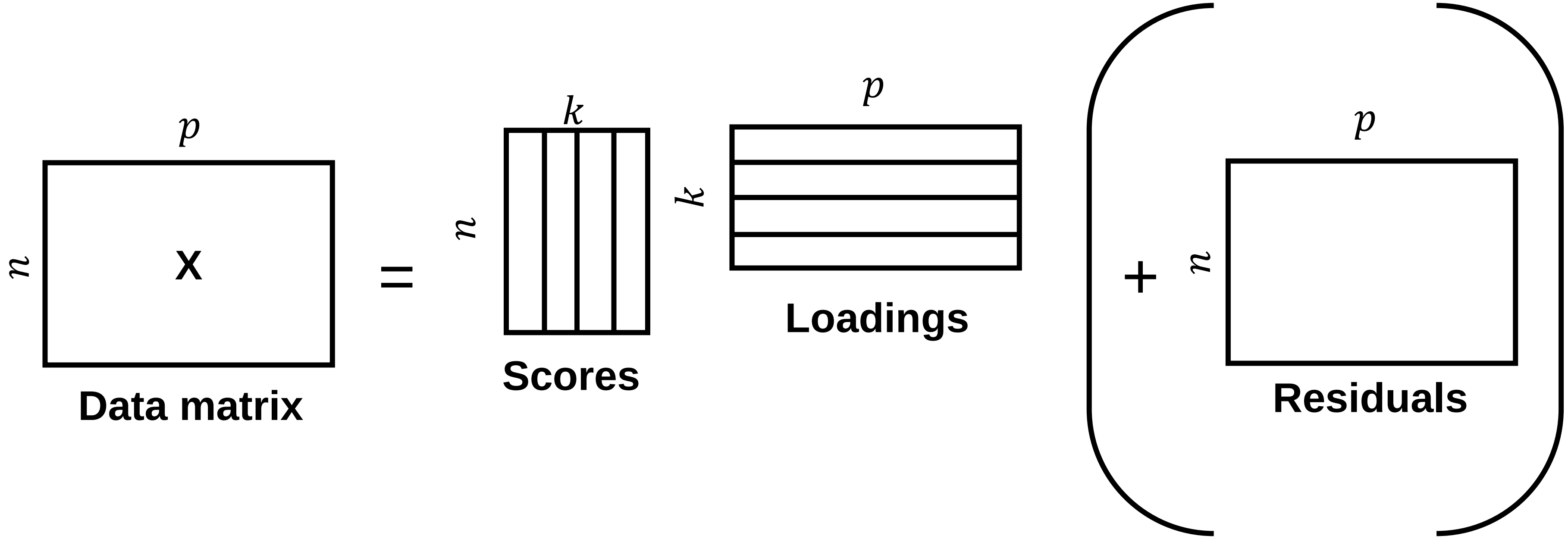
Objective

Find the factors capture the maximum of variability among these objects



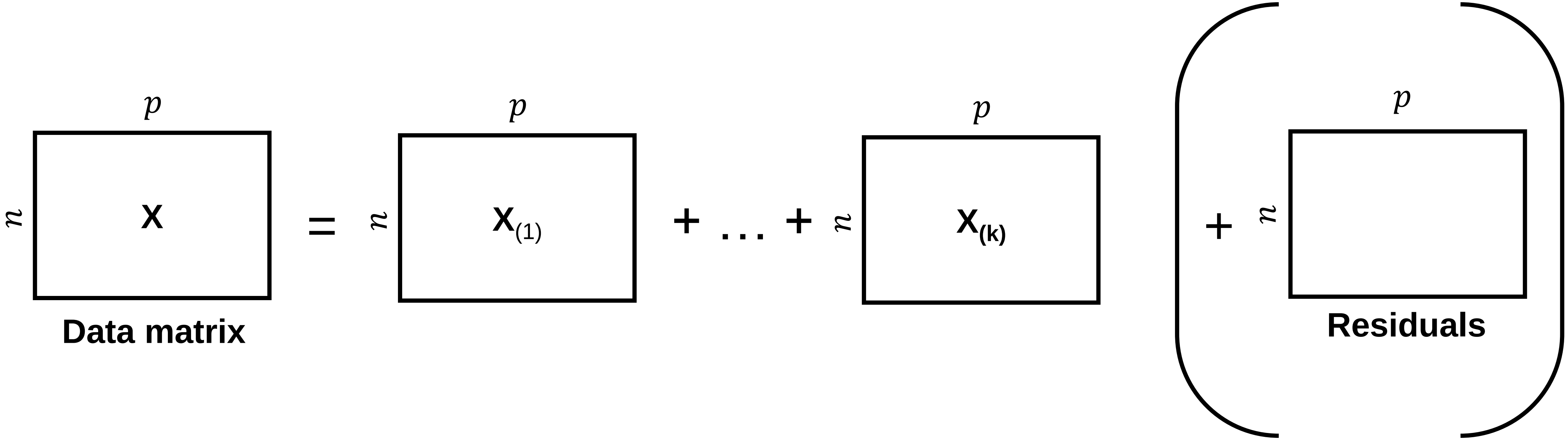
and some more PCs…

PCA: decomposition into scores and loadings



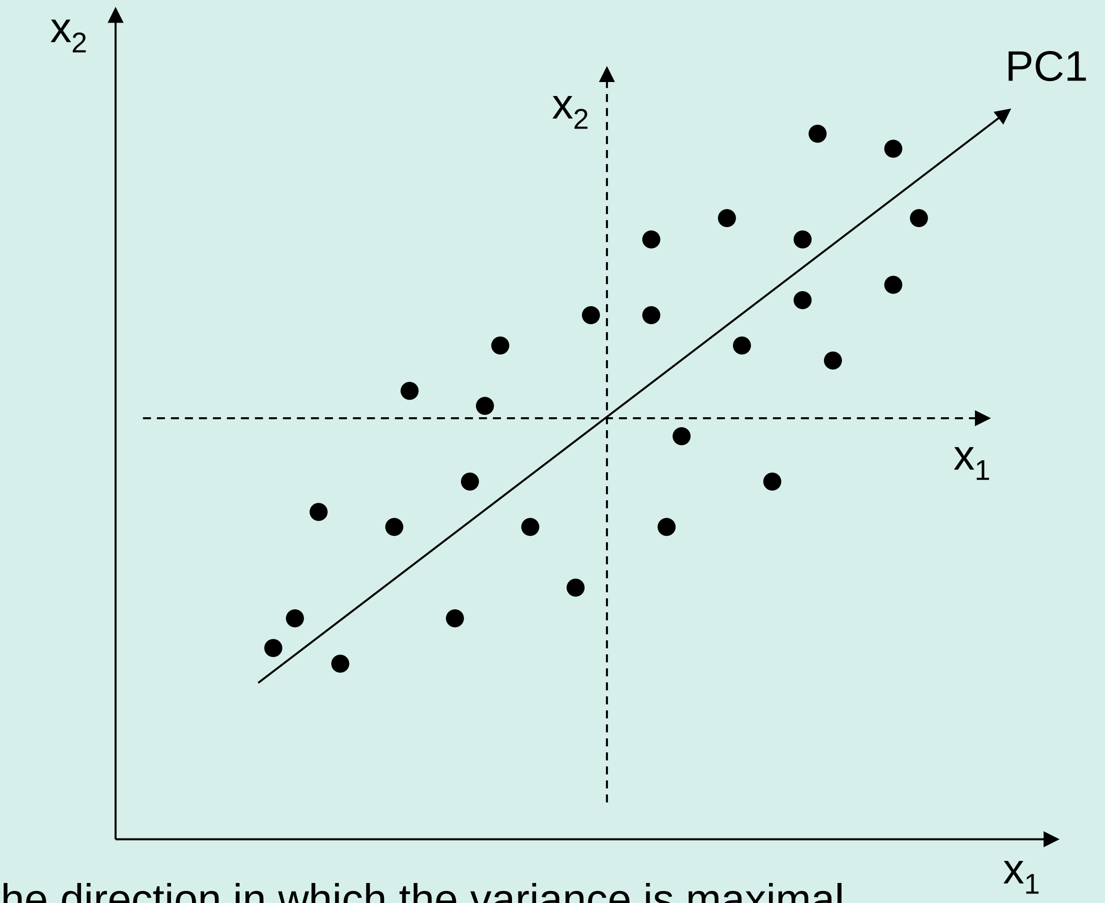
- The scores represent the values of the new factors for the observations
- The PCA model is described univocally by the loadings

PCA: contribution of principal components



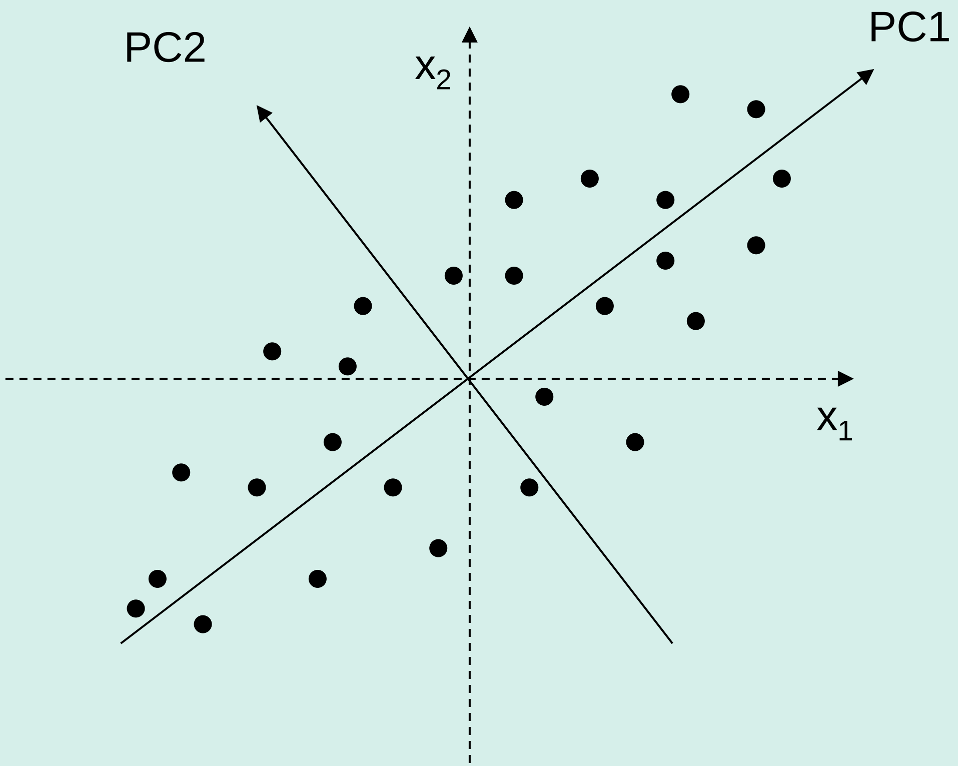
The spectra are the sum of the contribution of k principal components, plus remaining variation considered as not relevant

Geometric perspective of PCA



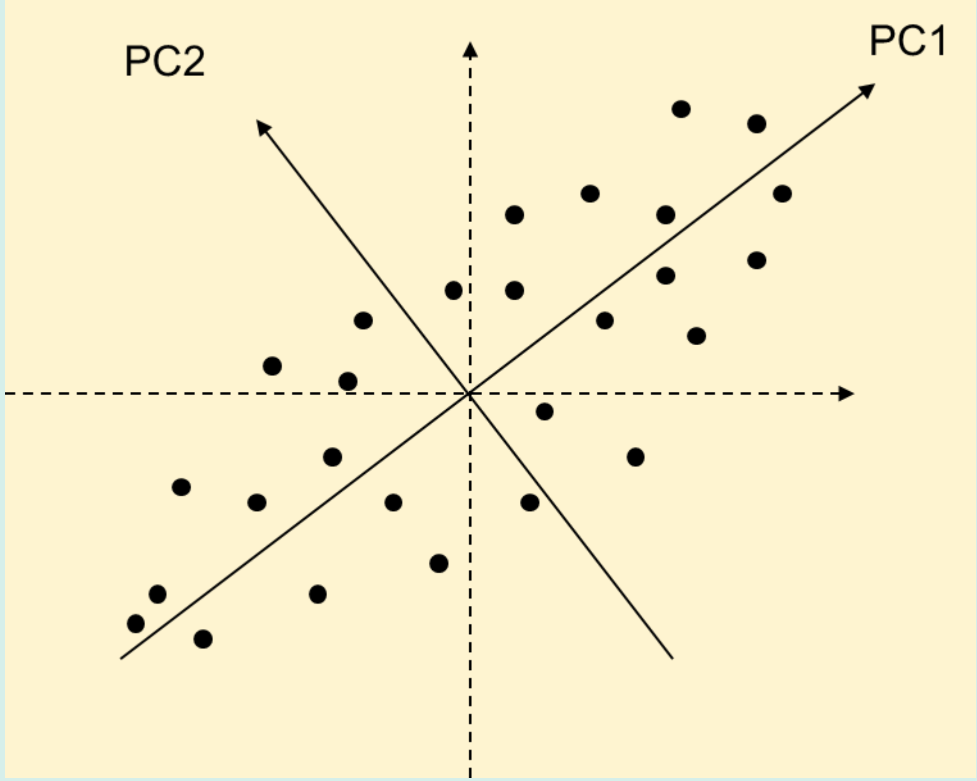
Find the direction in which the variance is maximal

Geometric perspective of PCA



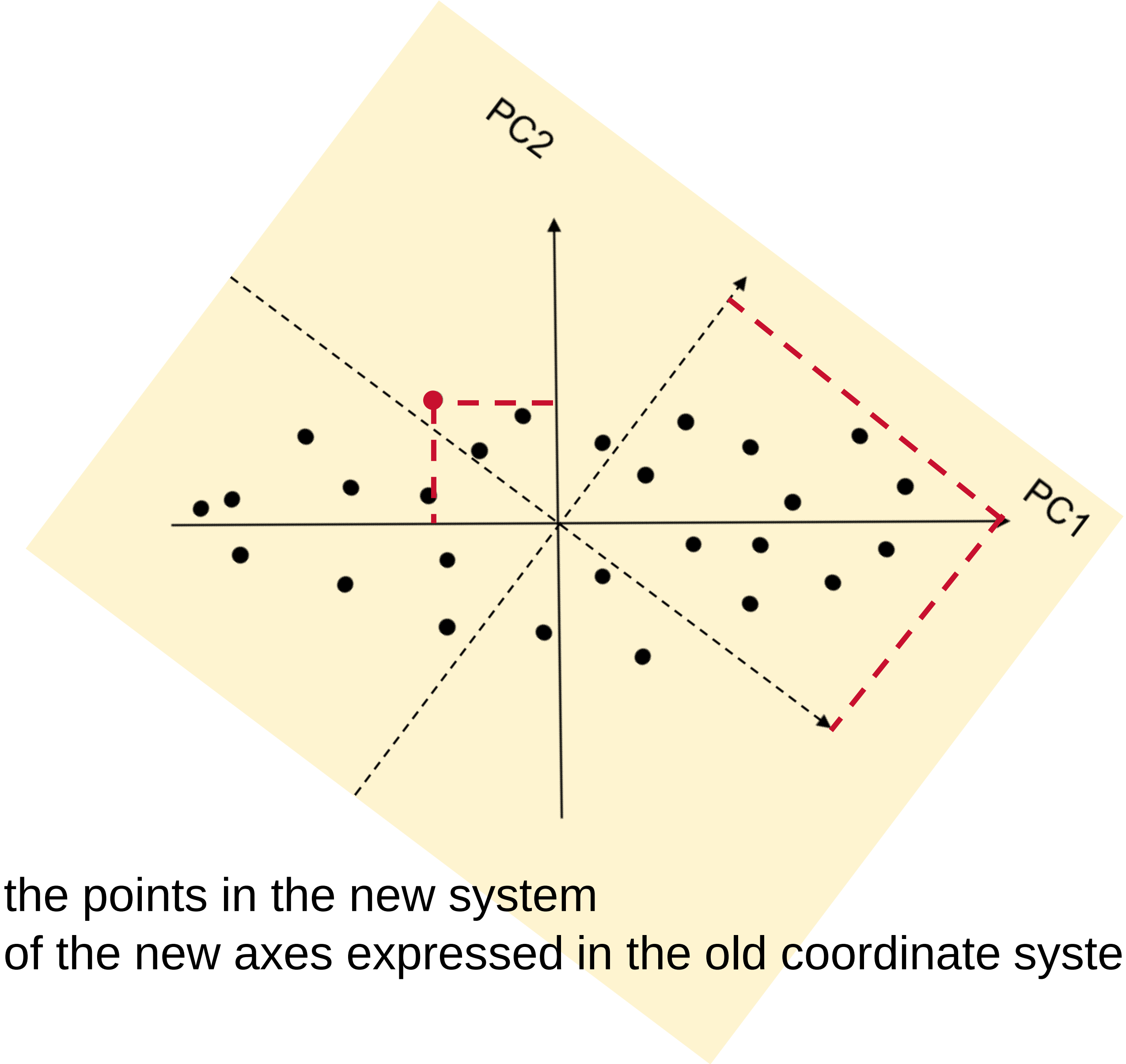
And repeat this for the remaining variance

PCA: scores and loadings



Rotate the space

PCA: scores and loadings



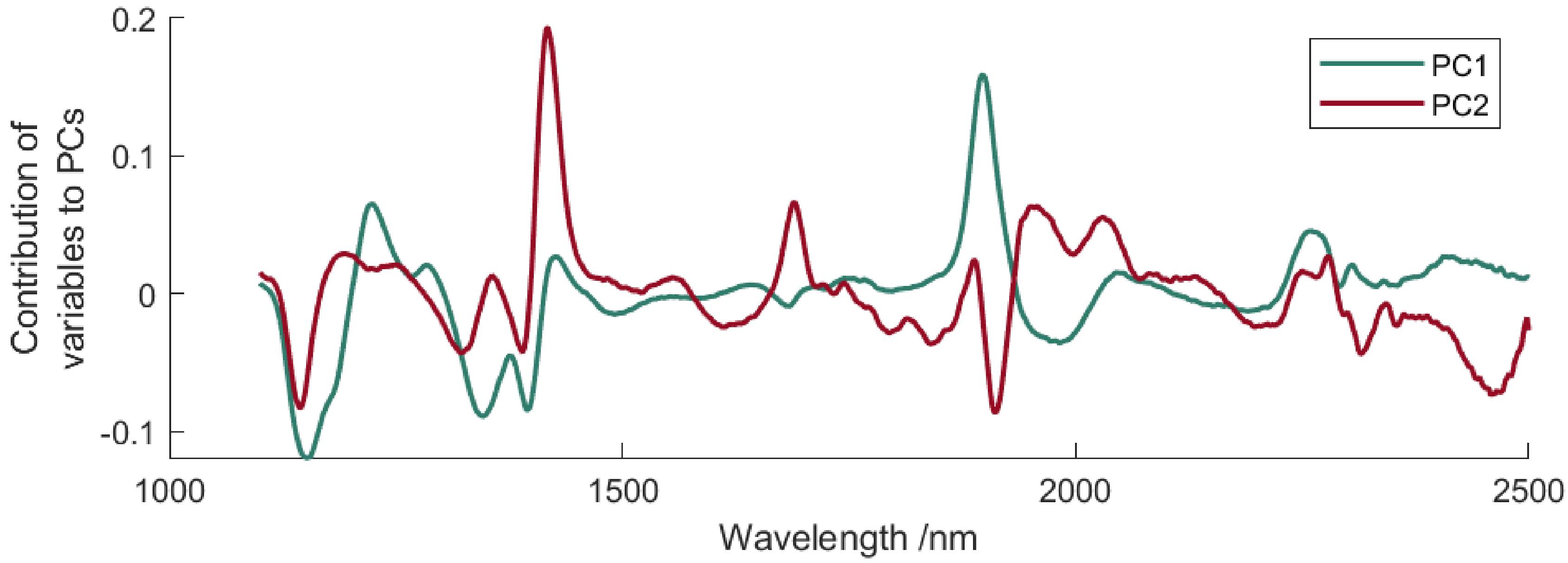
Scores - coordinates of the points in the new system
Loadings - coordinates of the new axes expressed in the old coordinate system

The loadings with a spectroscopic example

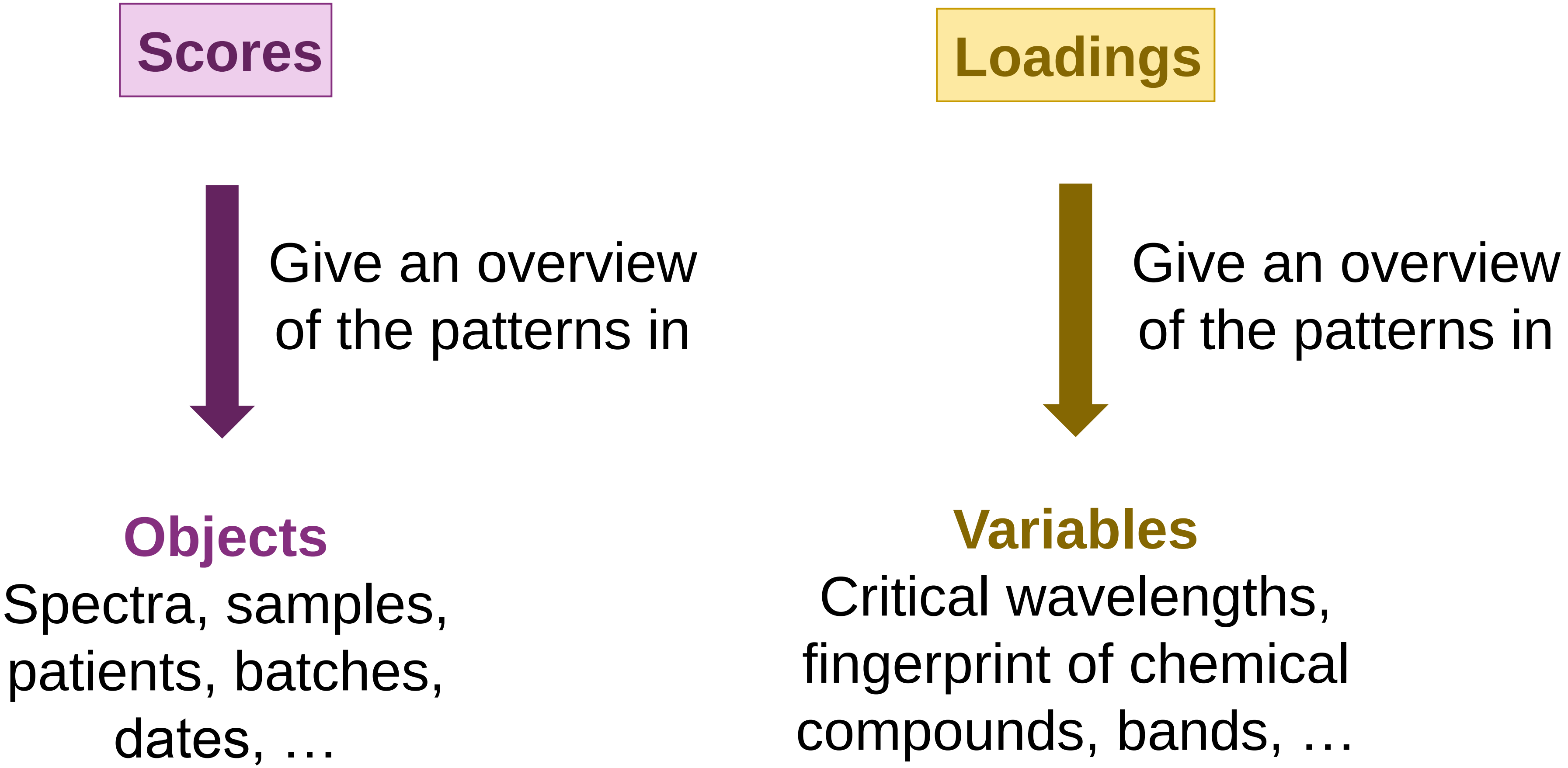
The loadings can be interpreted as

- the coordinates of the PCs in the original space
- the contribution of each original variable to each PC

They highlight features that explain the more the variability in the dataset



Data exploration in spectroscopy



Scores

Give an overview of the patterns in

Objects

Spectra, samples, patients, batches, dates, …

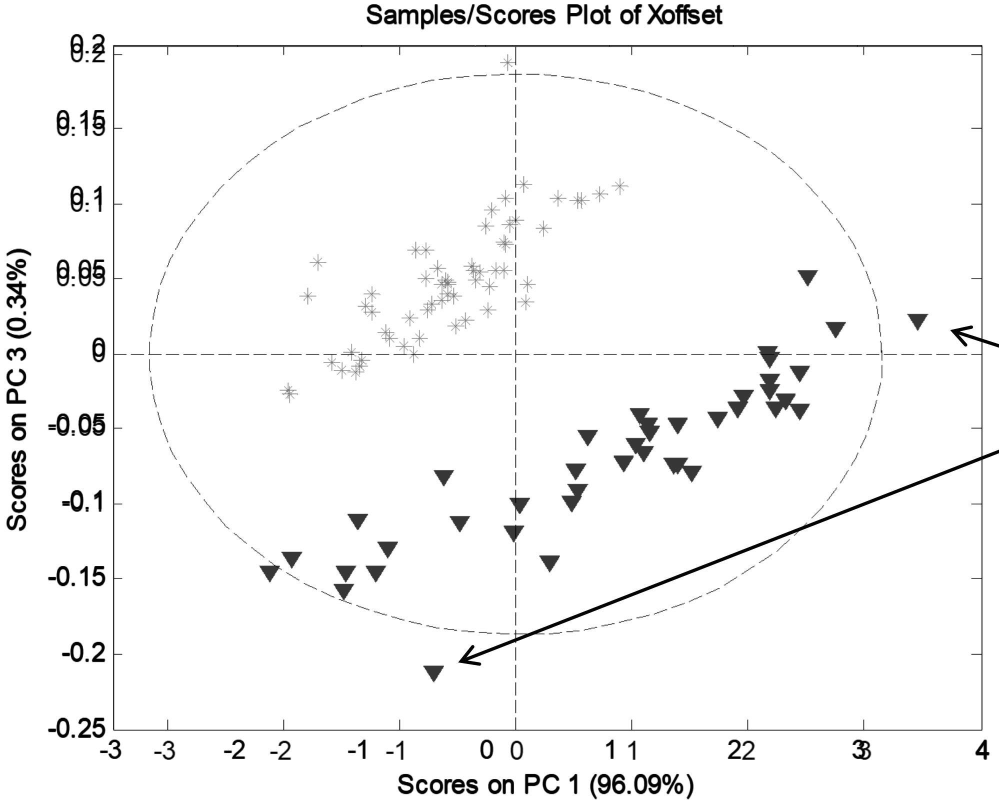
Loadings

Give an overview of the patterns in

Variables

Critical wavelengths, fingerprint of chemical compounds, bands, …

PCA scores: clustering and outliers

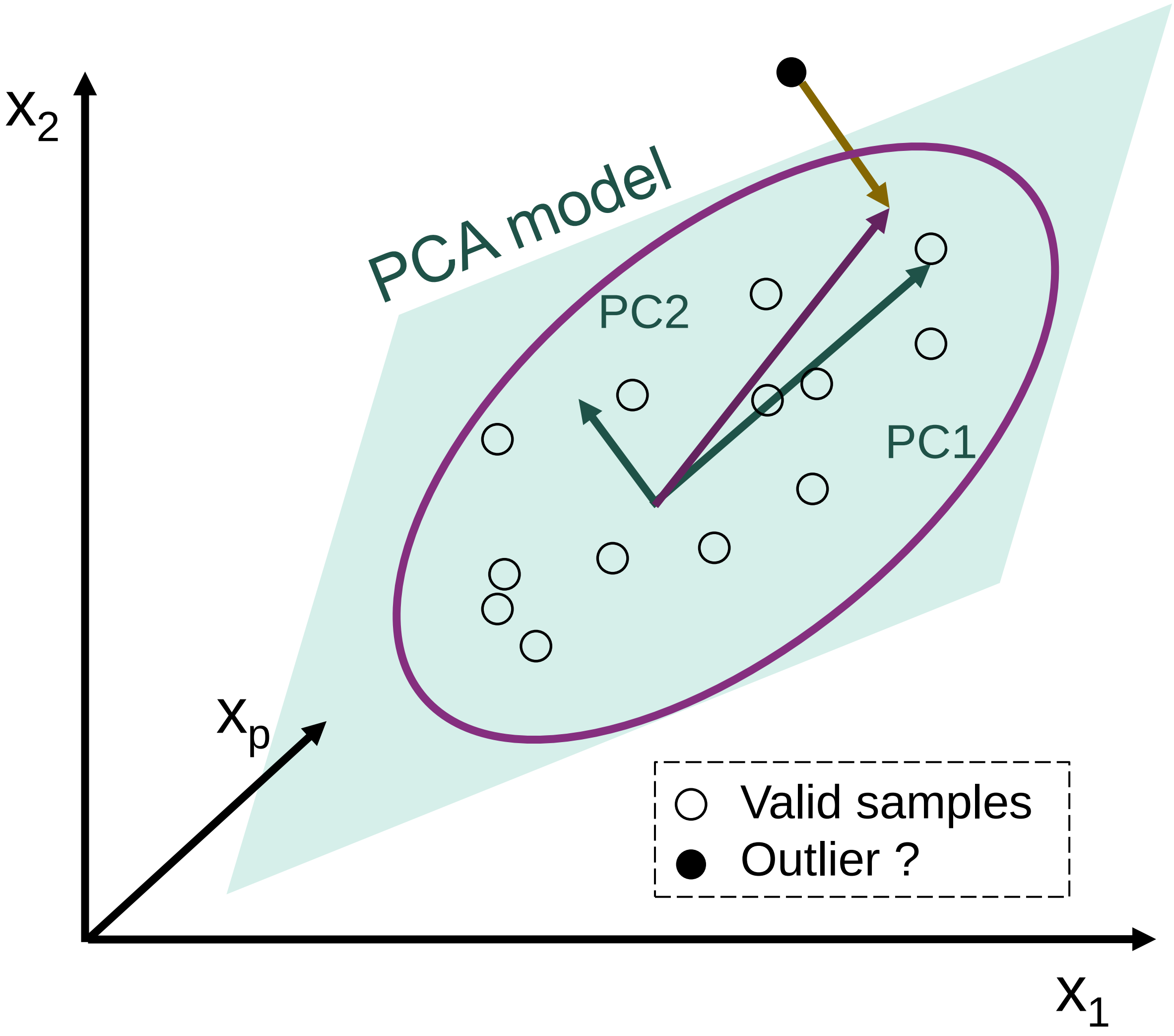


Clusters

Outliers (?)

Object space

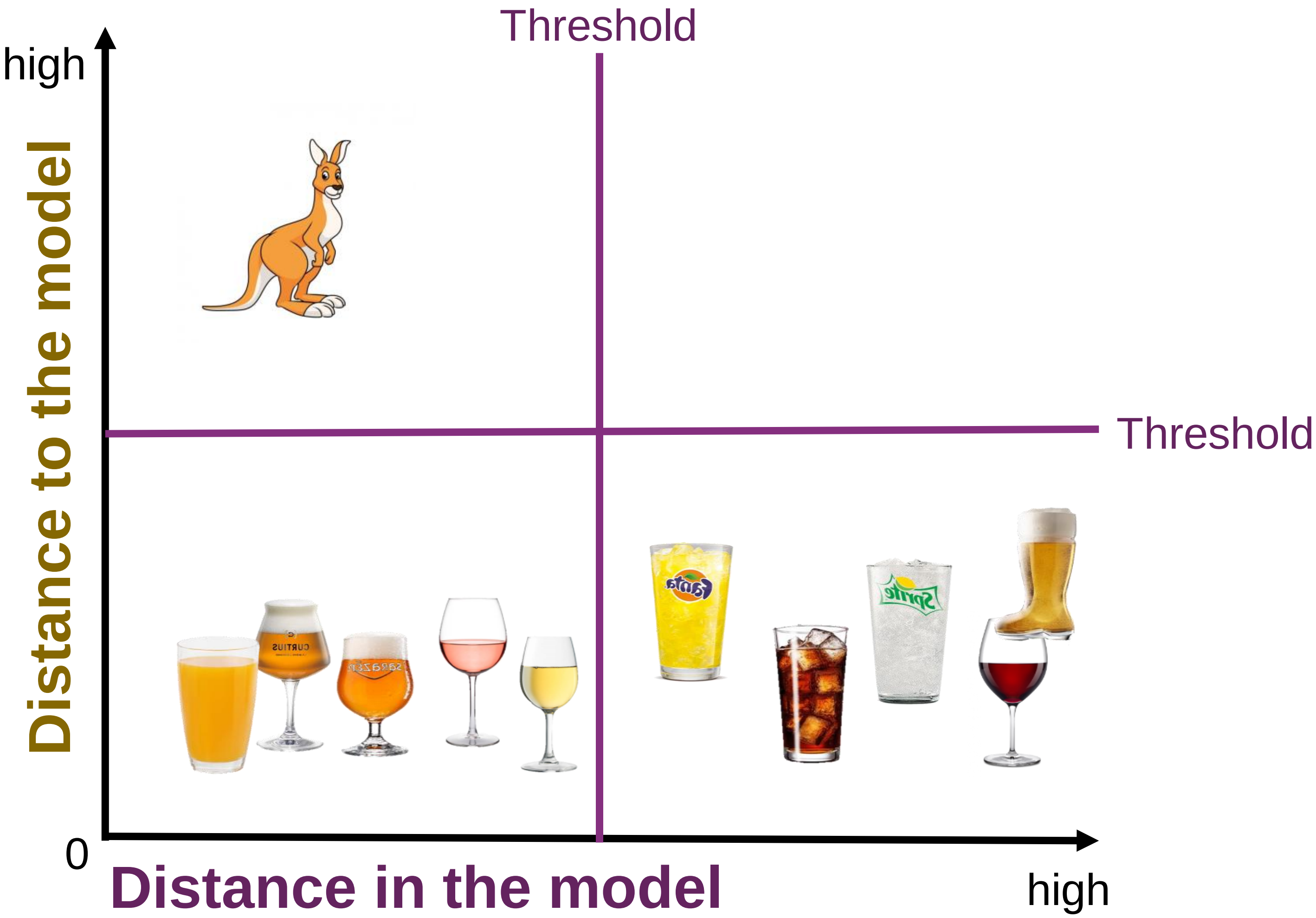
PCA and X-outlier detection



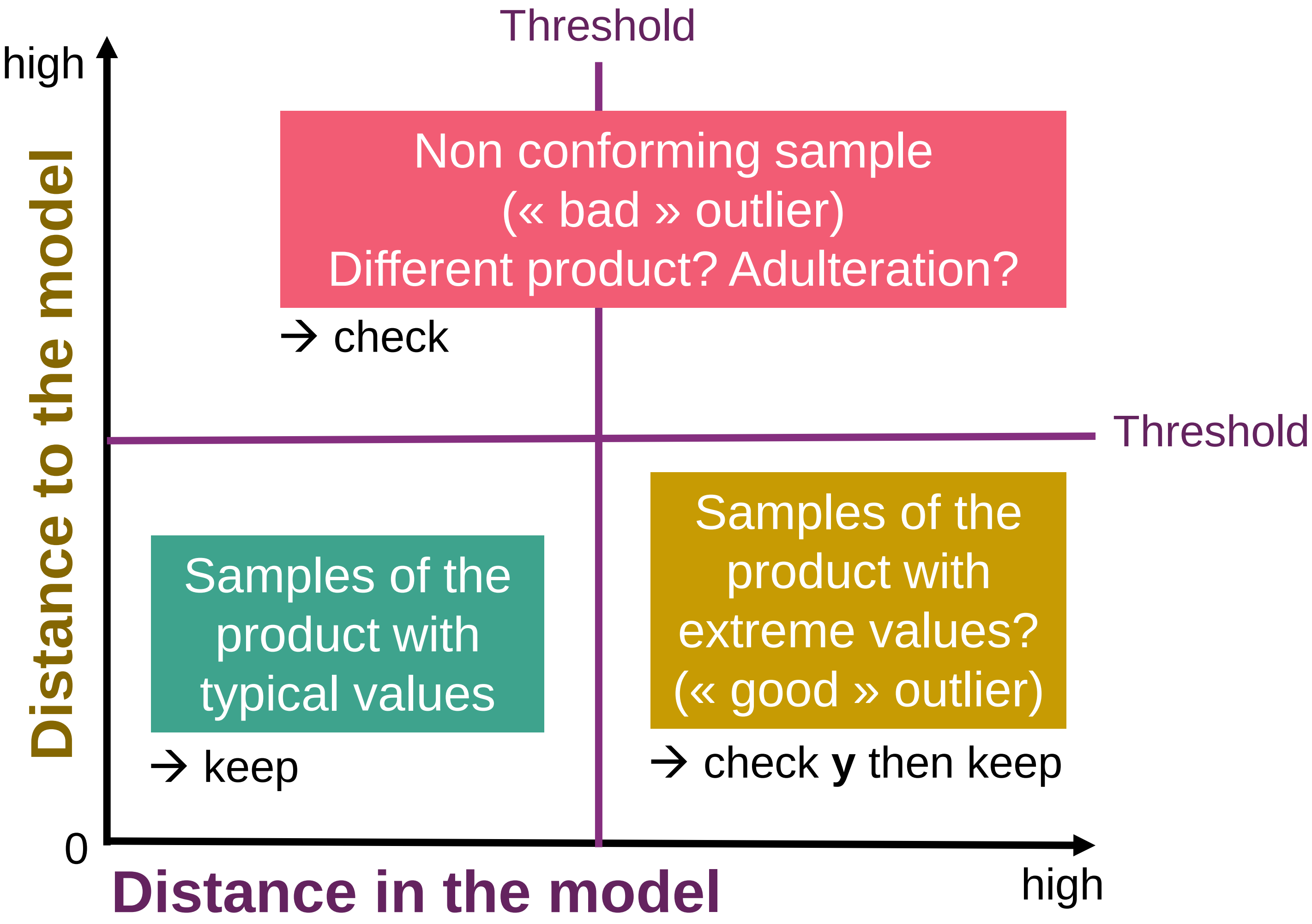
- **Distance to the model**
  Q residuals, maximum residuals, …

- **Distance in the model**
  Mahalanobis, GH, Hotelling’s T², …

X-outliers with PCA model



Interpretation of X-outliers



Non conforming sample
(« bad » outlier)
Different product? Adulteration?

→ check

Samples of the product with typical values

→ keep

Samples of the product with extreme values?
(« good » outlier)

→ check **y** then keep

PCA summary

Swiss army knife of chemometrics !!

Multiple advantages

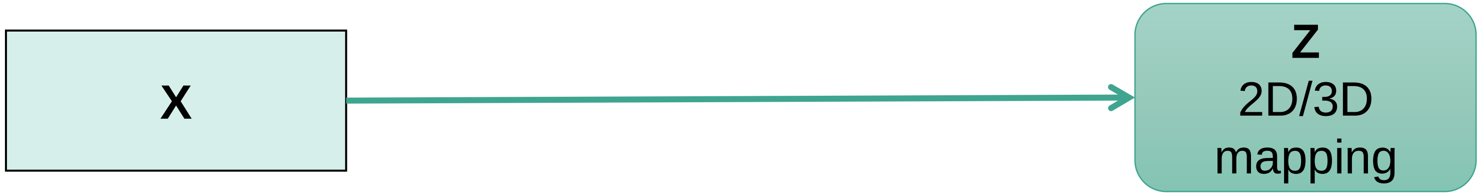
- Within objects
  - Identify clusters, highligh the effect of external factors
  - Detect outliers
- Identify important or useless variables and their relations
- Remove noise (preprocessing)
- Reduce dimensionality
  - Decrease storage requirements
  - Accelerate further processing

But

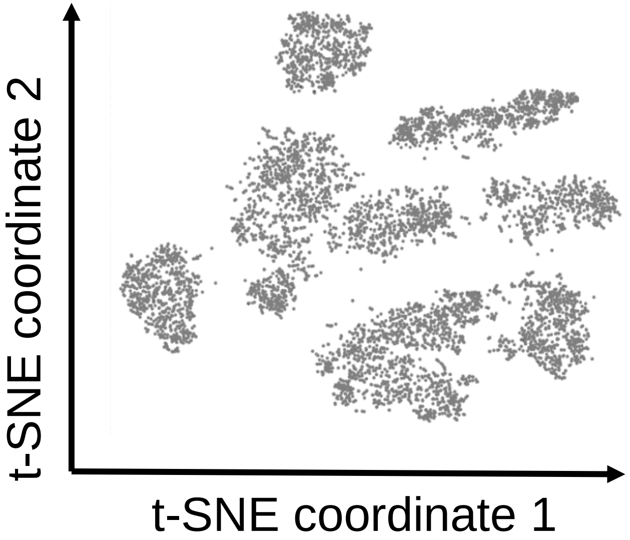
- It requires some expertise to make correct exploration and interpretations
- It is a linear method, only fitting linear variation (often sufficient with vibr. spectroscopy)

Specific vizualization methods: t-SNE, UMAP, …

Aim: visualize a high dimensional dataset into a single 2D map while preserving at best the relationships of similarity between objects


X

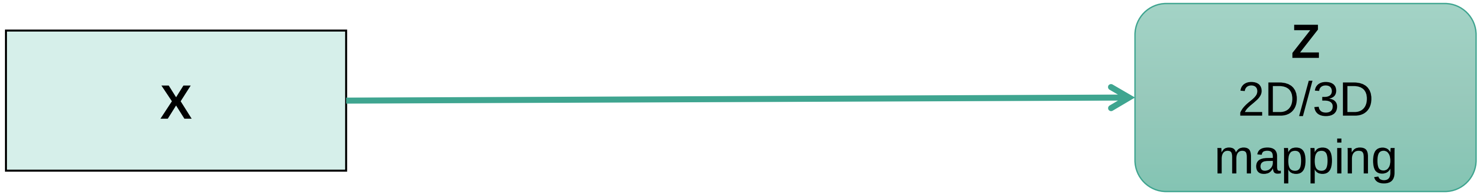
Z
2D/3D
mapping



van der Maaten, Laurens & Hinton, Geoffrey. (2008). **Visualizing data using t-SNE**. *Journal of Machine Learning Research*. 9. 2579-2605.

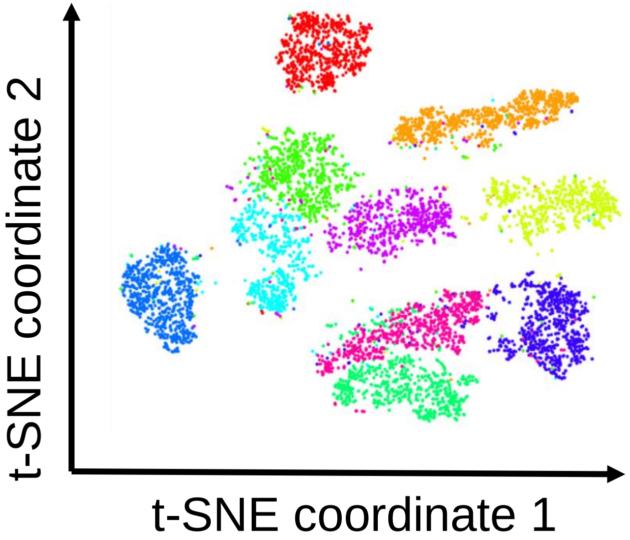
Specific vizualization methods: t-SNE, UMAP, …

Tip: if we color the objects in the t-SNE map according to a categorical reference variable, then it allows checking the influence of this variable



X

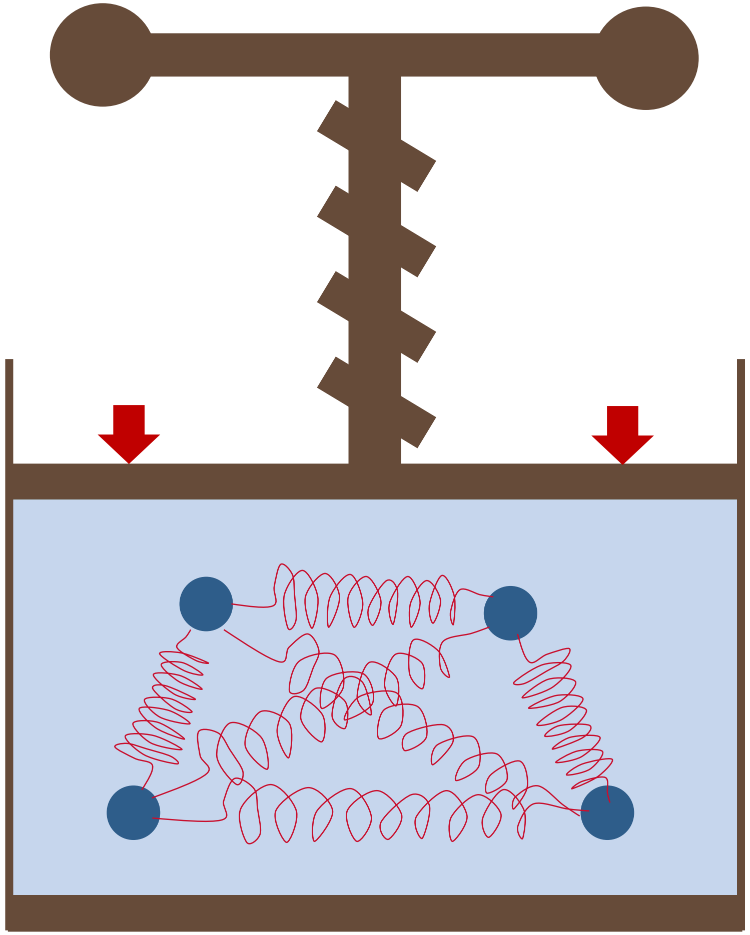
Z
2D/3D
mapping



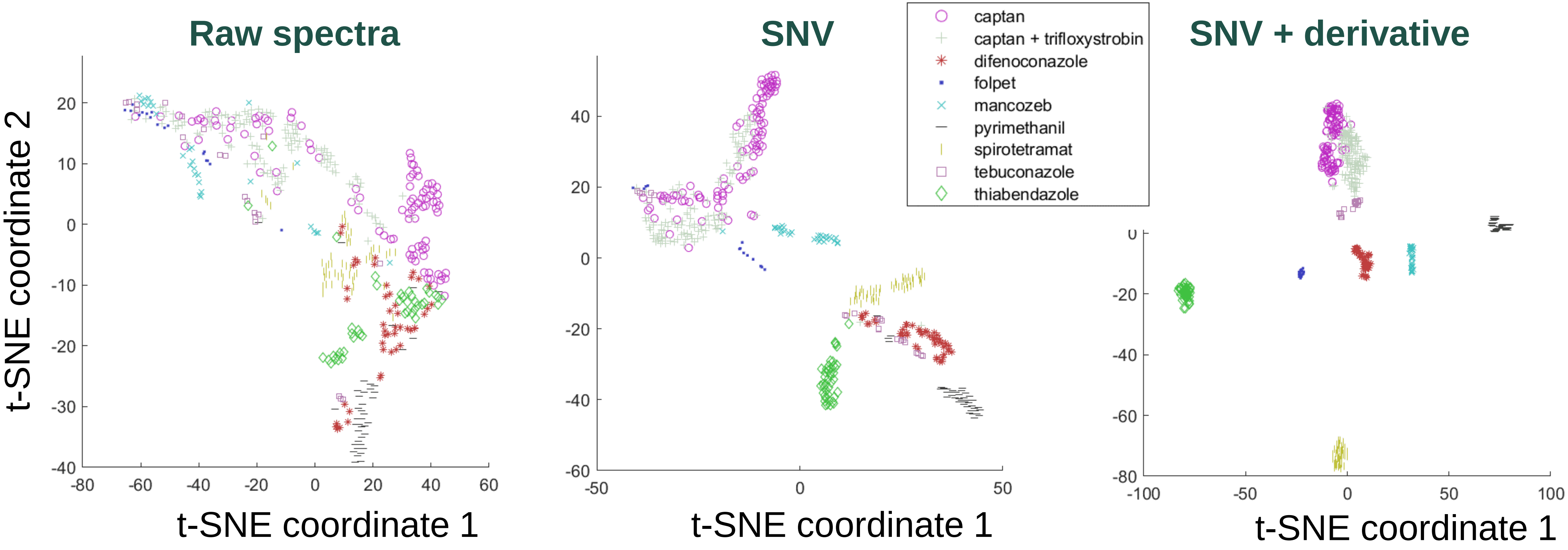
van der Maaten, Laurens & Hinton, Geoffrey. (2008). **Visualizing data using t-SNE**. *Journal of Machine Learning Research*. 9. 2579-2605.

t-SNE analogy

- n objects (balls) float in dimension p, undergoing forces from other objects
- The more different are the objects, the more repulsive are the forces
- The dimensionality is reduced progressively, leading to an optimal reorganisation, until reaching dimension 2, the « mapping »



Pesticides example: choice of preprocessing



Raw spectra

SNV

SNV + derivative

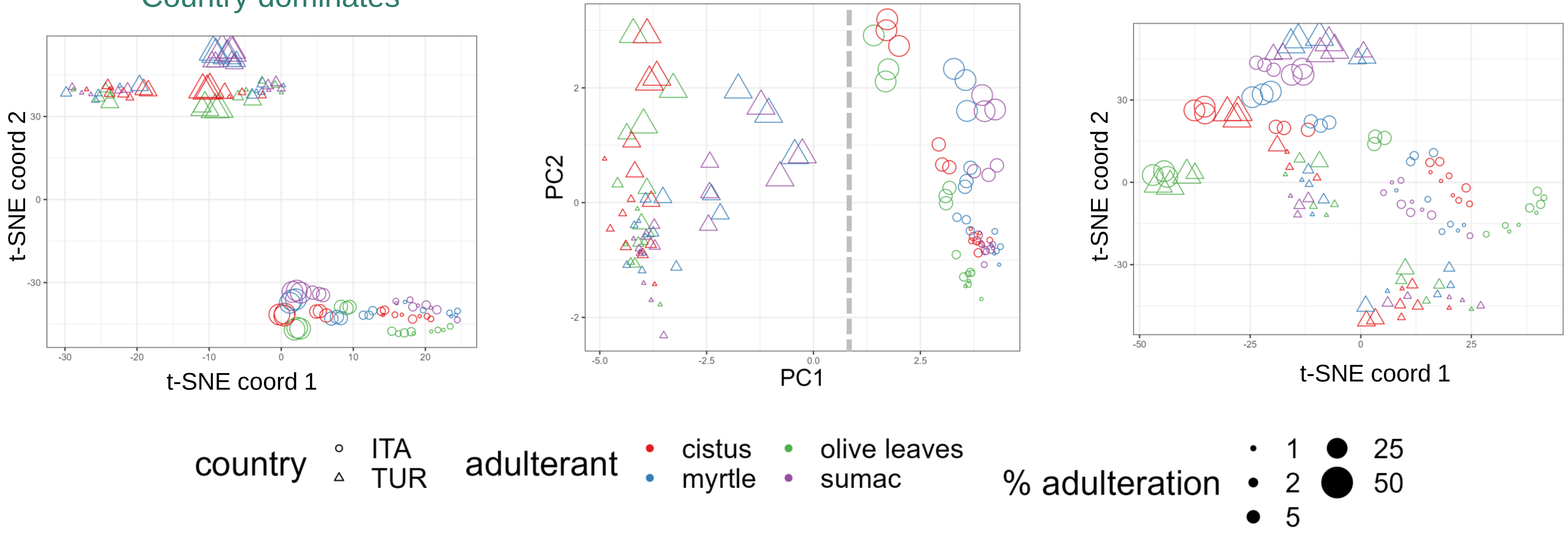
→ t-SNE can help choosing the preprocessing pipeline by indicating which one provides the best separation between the classes in **Y**

Combine t-SNE and PCA – oregano dataset

t-SNE on the 100 first PCs
Country dominates

Scores of PC1 and PC2
PC1 largely explains country

t-SNE on PCs **2**-100
Effect of contaminant clearer



Stevens F, Carrasco B, Baeten V, Fernández Pierna JA. **Use of t-distributed stochastic neighbour embedding in vibrational spectroscopy**. Journal of Chemometrics. 2024; 38(4):e3544. doi:10.1002/cem.3544

Regression: many methods

Method	Regression	Discrimination
Multiple linear regression (MLR) With regularization: ridge, lasso, elasticnet	+	+
Principal component analysis (PCA)	PCR	SIMCA
Partial least squares (PLS)	PLSR	PLSDA
Support vector machine (SVM)	SVMR	SVMDA
Local methods	Local PLS, …	K-nearest neighbors (kNN)
Classification and regresion tree based methods (random forest, XGBoost, …)	+	+
Artificial neural networks (ANN)	+	+

The general regression framework

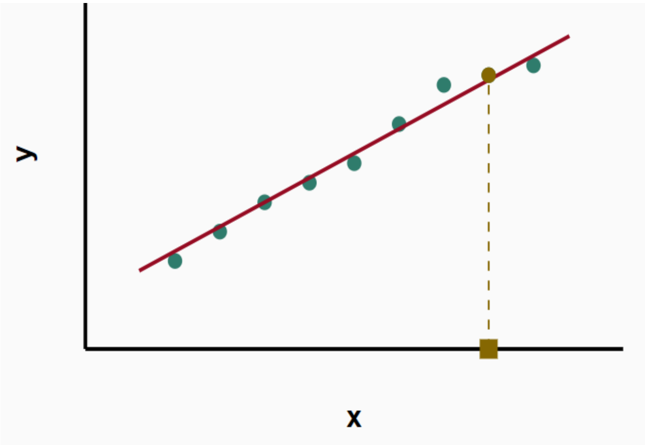
The process

The light ($\mathbf{X}$) is a function of the matter ($\mathbf{y}$)

$\mathbf{X}$ = F($\mathbf{y}$, .)

The general framework

$$\mathbf{y} = \hat{f}(\mathbf{X}) + \boldsymbol{\varepsilon}$$



The linear framework

$$\boldsymbol{y} = \boldsymbol{X}\hat{\mathbf{b}} + \boldsymbol{\varepsilon}$$

$$y_i = x_{i1}\hat{\mathrm{b}}_1 + x_{i2}\hat{\mathrm{b}}_2 + \cdots + x_{ip}\hat{\mathrm{b}}_p + \varepsilon_i$$

$\hat{\mathbf{b}}$ = the model

sorry…

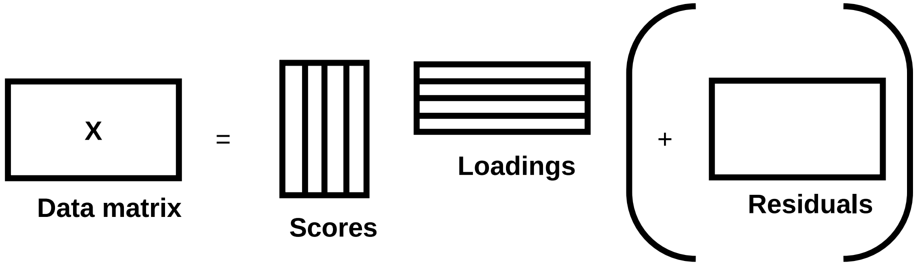
The problem of multicollinearity

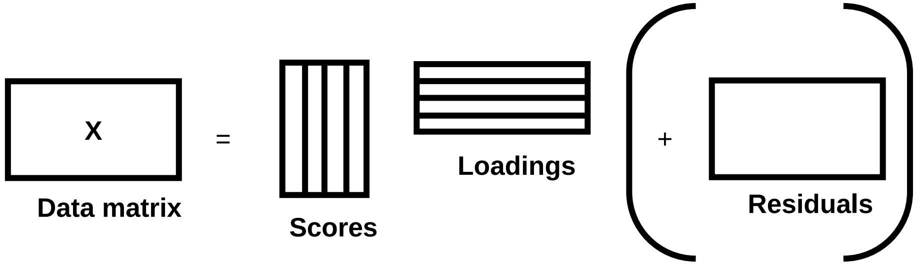
Linear regression does not work when multicollinearity is present

→ this leads to unstable models that fails in future predictions

One solution is to compress the data into independent factors using a method like PCA and apply the linear regression on the scores

Regression: PCR vs. PLSR





	Factors name	Criterion for selecting factors
PCR	Principal components	Maximize variance
PLSR	Latent variables	*Maximize covariance with y*

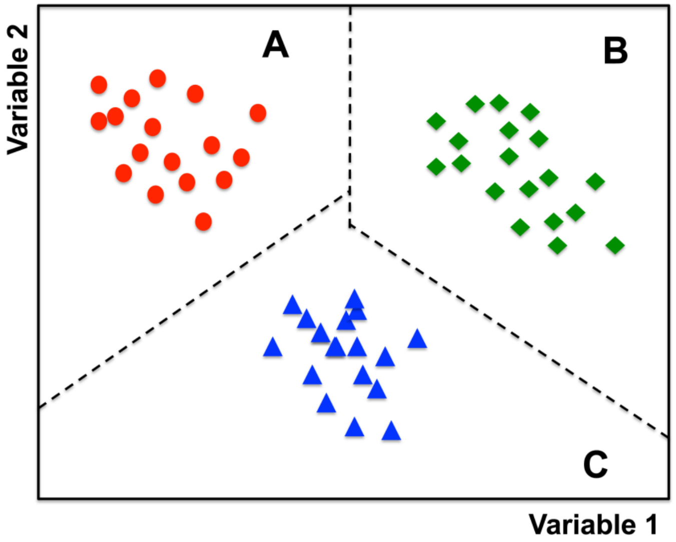
Advantages of PLSR over PCR

- With PCR the first factors are not necessarily the ones that best explain $\boldsymbol{y}$
- Actually, the factors that best explain $\boldsymbol{y}$ could have a very low variance in $\boldsymbol{X}$ and appear late in the list of factors
- With PLSR, the ability to explain $\boldsymbol{y}$ is taken into account in the selection and in the ranking of the factors
- PLSR is thus able to better fit the calibration dataset and to better predict future samples while using less factors than PCR

Classification: discriminant modelling

- This group of methods implicitly or explicitly tries to find the boundaries which separate the different classes in the multidimensional space.

- The corresponding outcome is always the classification to one of the available categories.

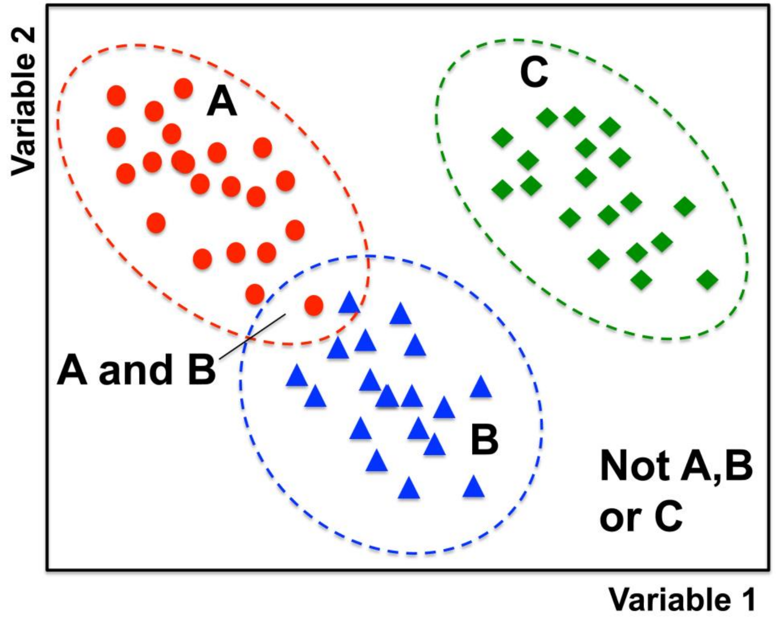


A classical discriminant model is constructed based on differences among classes studied, and a new sample is always assigned to one of these classes.

Ex. PLS-DA

Classification: class modelling

- This group of methods focuses on looking for similarities among samples belonging to the same class.
- Each category is modeled individually.
- A sample can be assigned to one class, to more than one class or to no class at all.

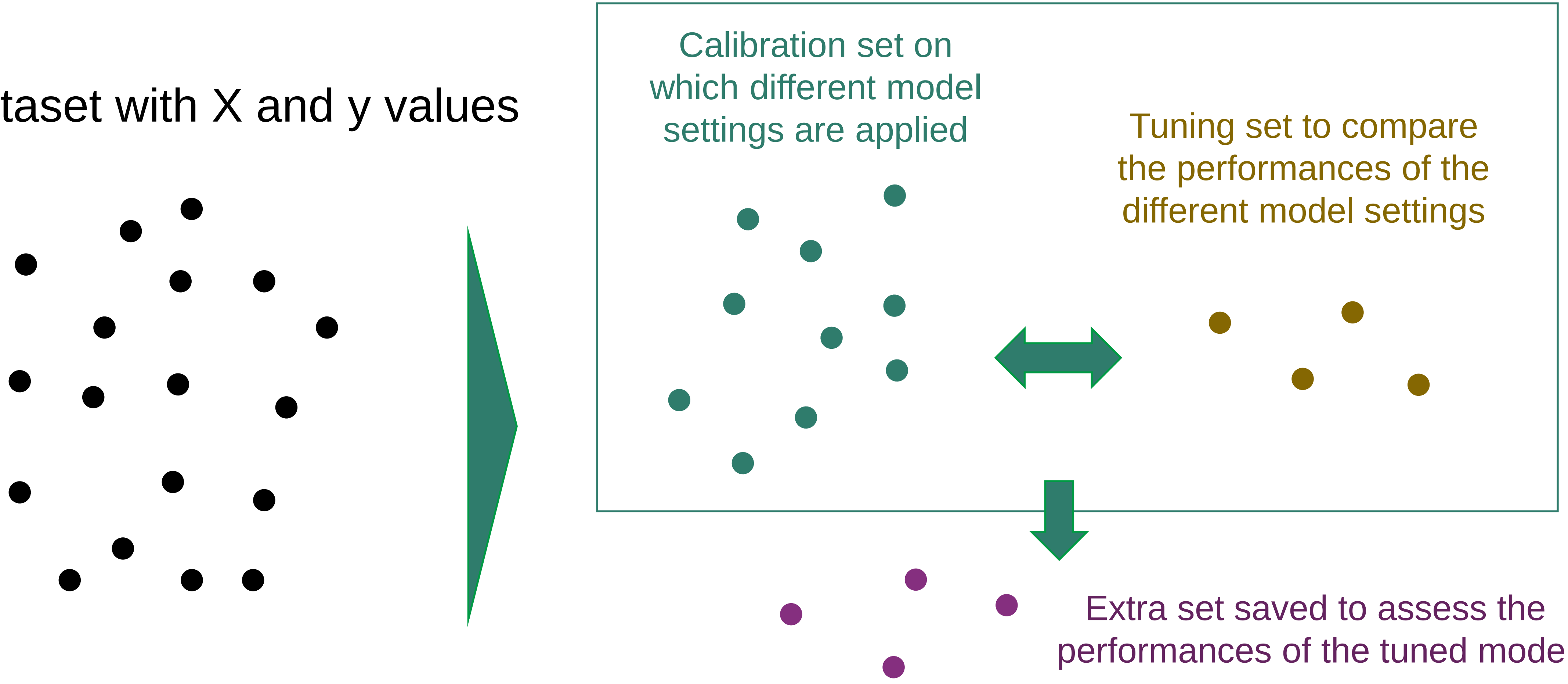


A class-model is constructed individually for each of the classes studied, based on the similarities among samples from the same class.

Ex. SIMCA

Predictive modelling and validation

A dataset with X and y values



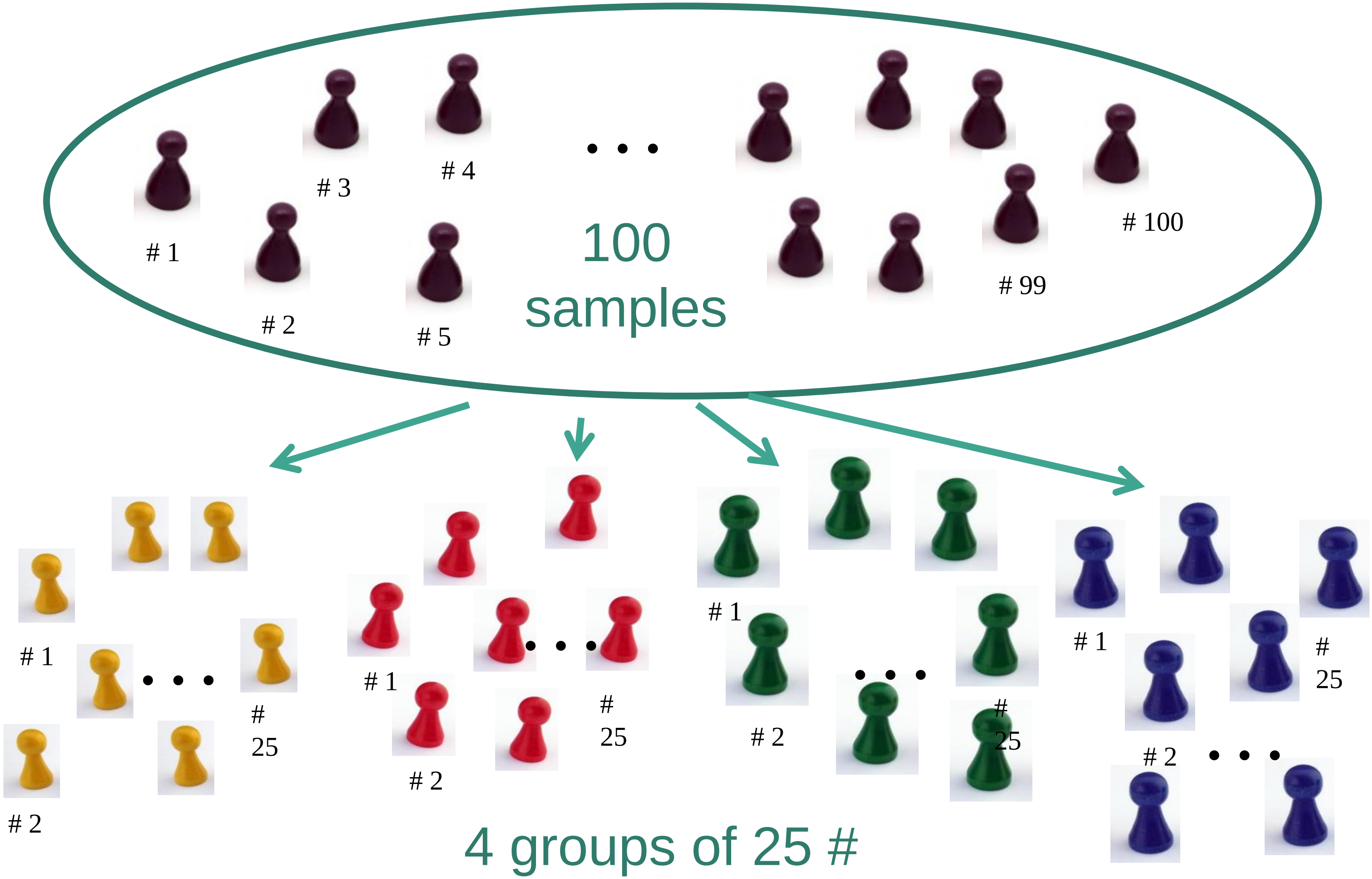
Calibration set on which different model settings are applied

Tuning set to compare the performances of the different model settings

Extra set saved to assess the performances of the tuned model

The assessment set should be independent from the rest of the data !

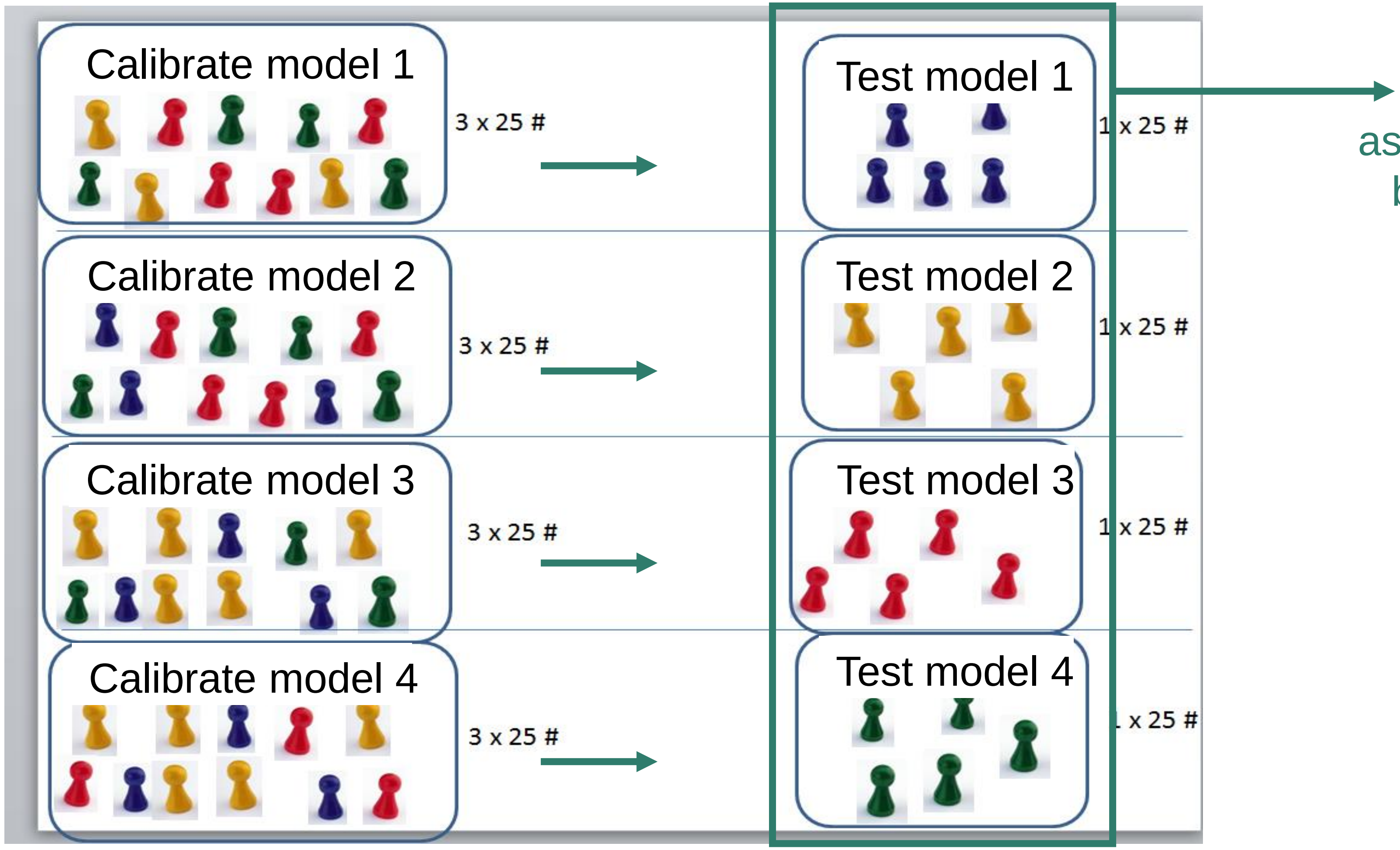
Tuning model with k-fold cross-validation



4 groups of 25 #

Tuning model with k-fold cross-validation

For a given value of the complexity hyperparameter(s)

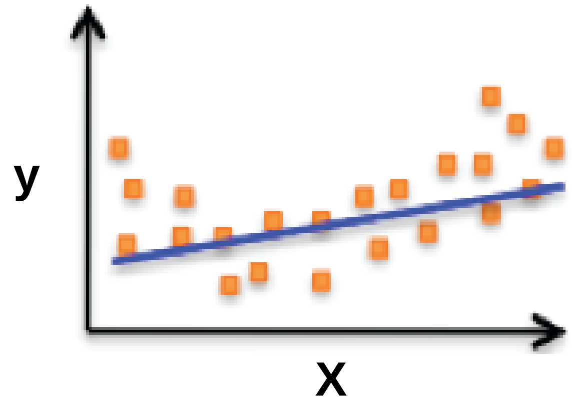


Join predictions and assess model performance by comparing predicted and (reference) values

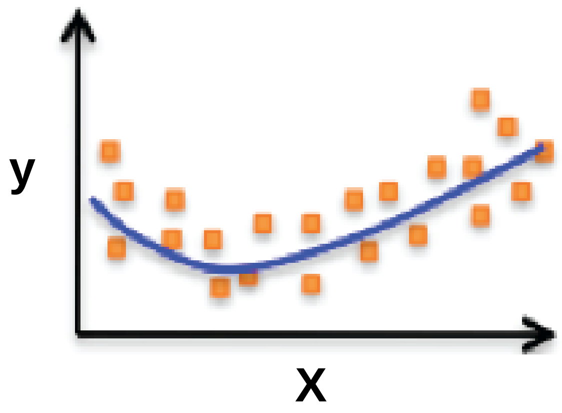
Under- and overfitting

Too simple model | Good model | Too complex model

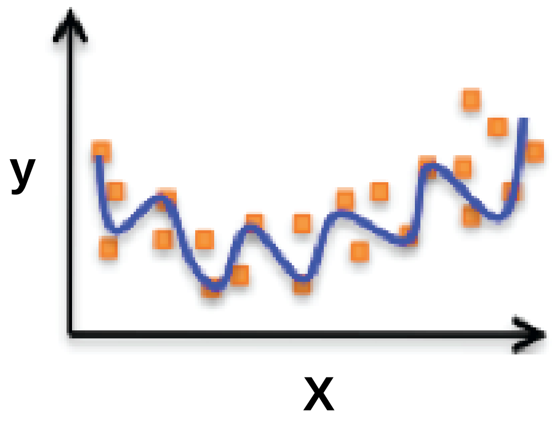
Underfitting



”Relying on incorrect assumptions
and missing relevant relations
leads to poor prediction with systematic error”



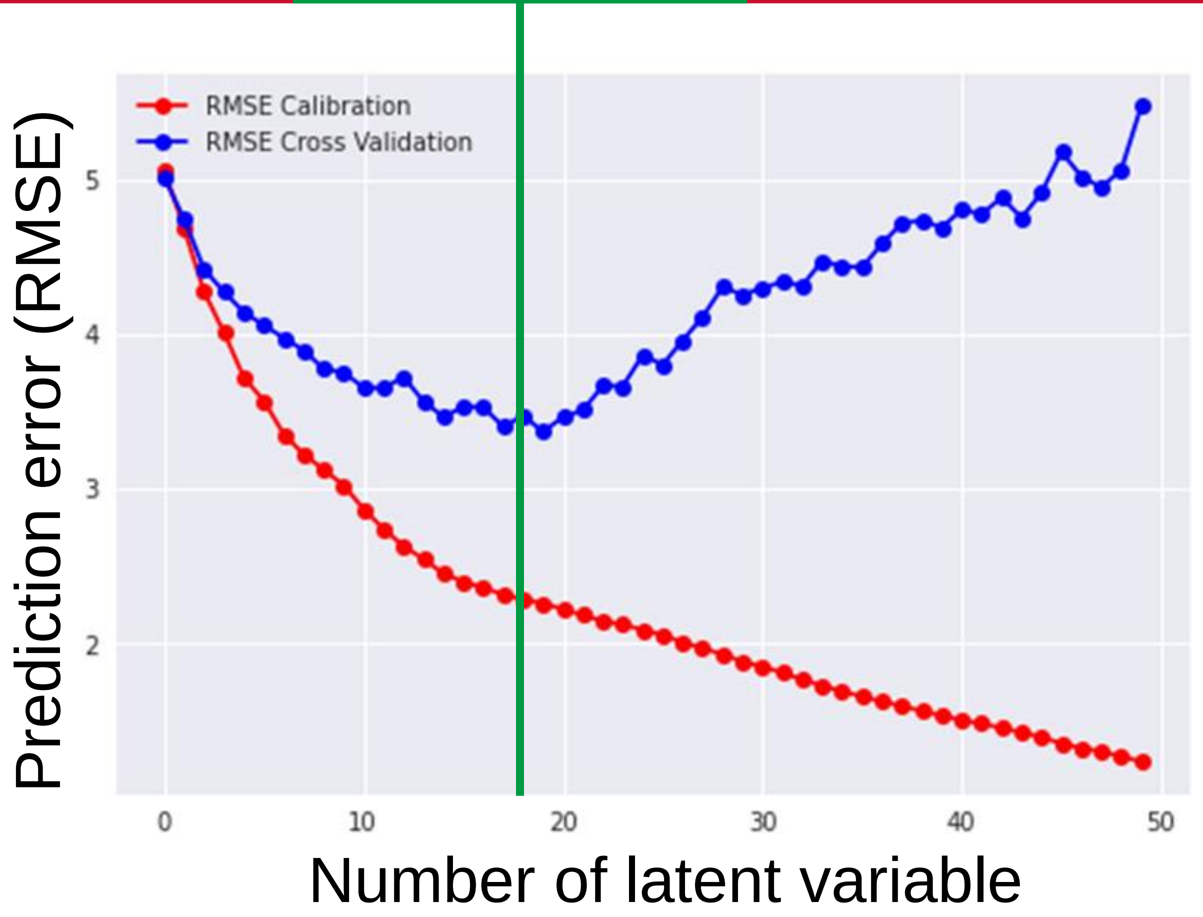
Overfitting



”Small variations in calibration data
might result in a completely different
model being generated”

PLS model tuning

Too simple model | Good model | Too complex model



Cross-validation

Objective estimate of the performance of prediction on new samples

Calibration

Local methods

Considering an large database of spectra and associated reference values

For *each spectrum* whose prediction is aimed

1. Select spectra located in a neighbourhood (typically using Euclidian or Mahalanobis distance)
2. Fit a predictive model on these neighbourhood spectra
3. Predict the y of the new sample with this model

Strategy of local methods

Local methods exploits the fact that

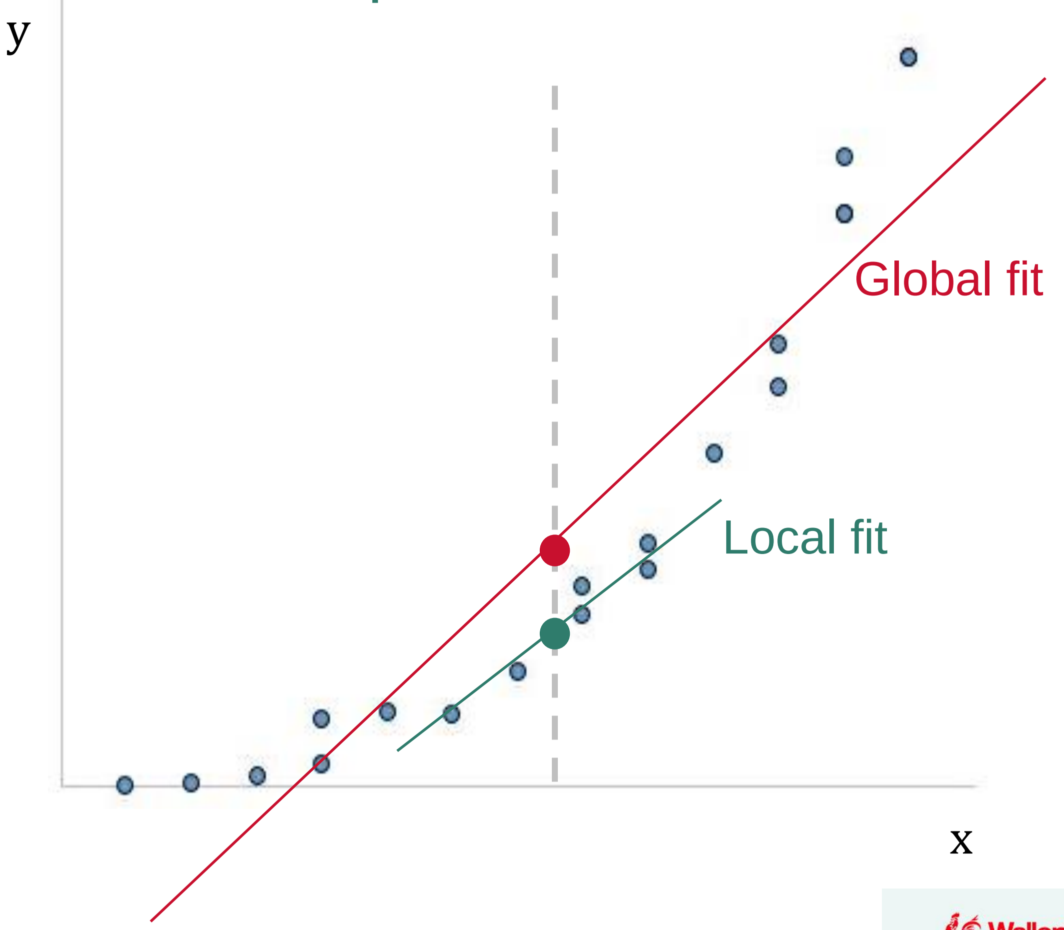
non-linear trends

may be well approximated

locally

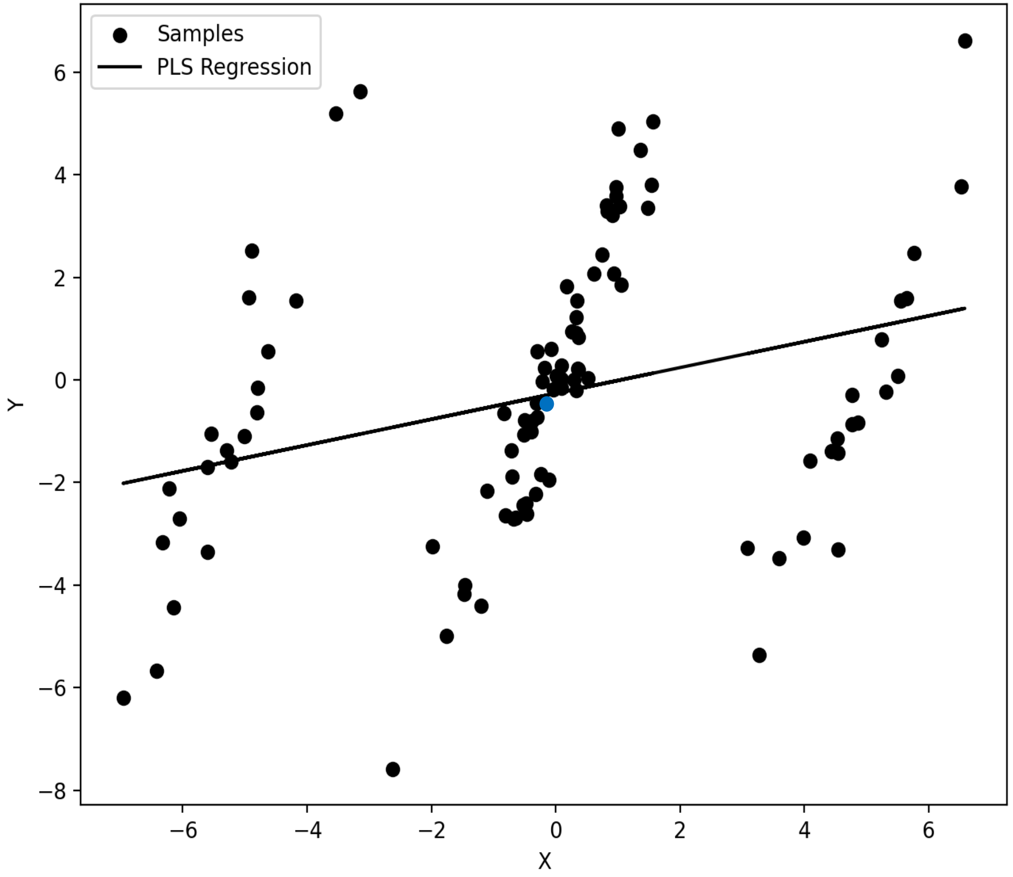
by a linear model

Simplified 2D illustration

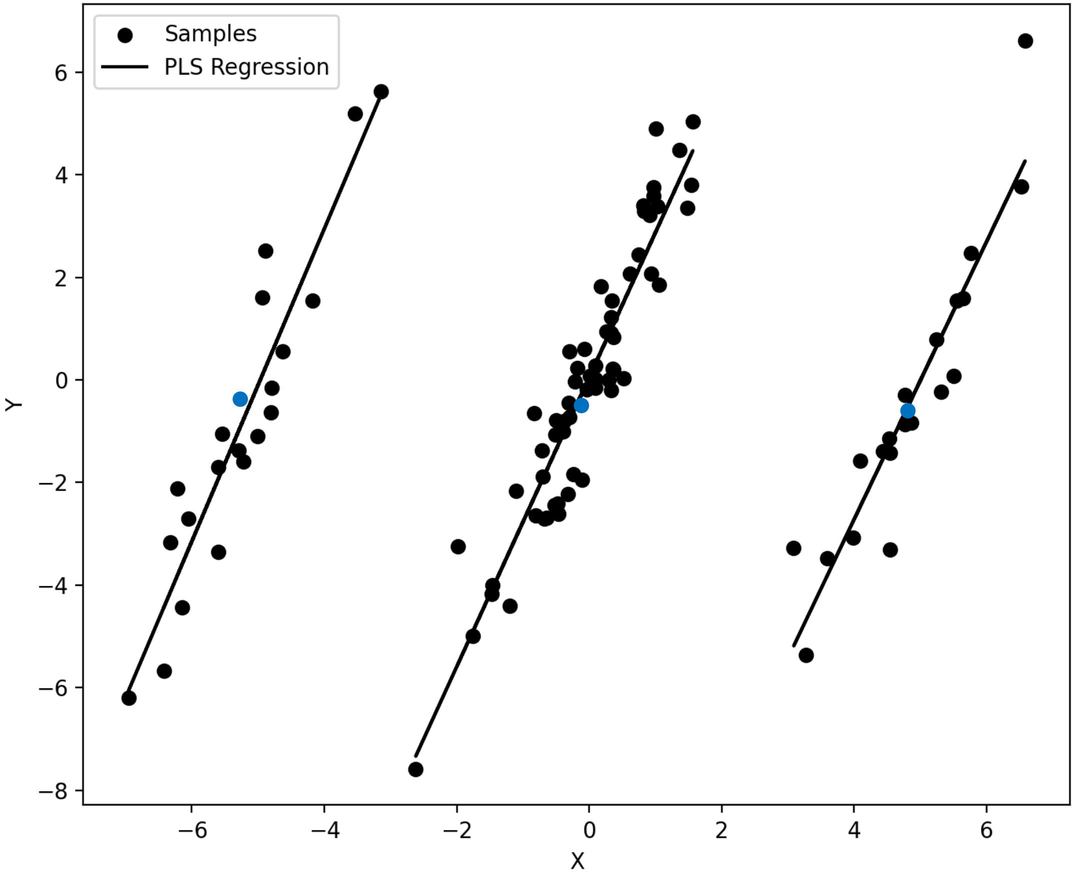


Strategy of local methods

Global PLS



Local PLS



Advantages of local methods

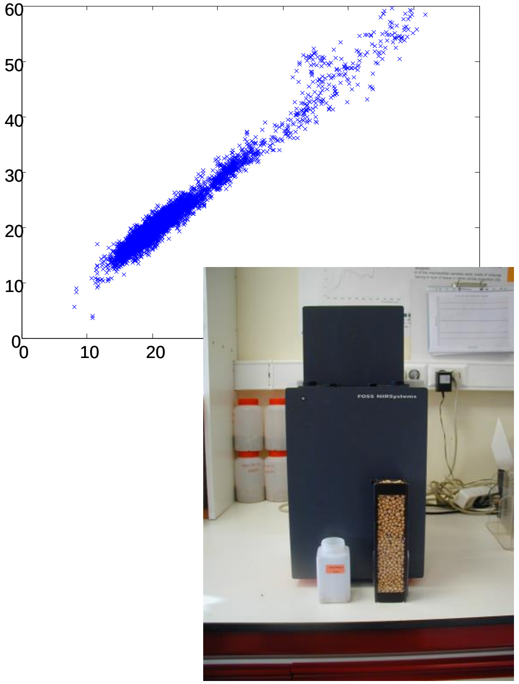
- Deal with non linearities
- Work with a multi-product library
- No need to develop and maintain individual calibration models
- Ideal for cloud predictions
- The library can be protected and compressed (example: PCA)

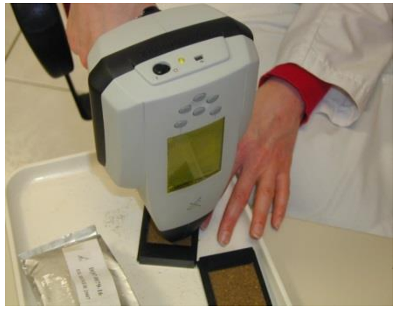
But keep in mind

- Requires a library at disposal
- Prediction may be slower than with the global method

Transfer between instruments

CALIBRATION TRANSFER FROM DISPERSIVE INSTRUMENTS TO HANDHELD SPECTROMETERS

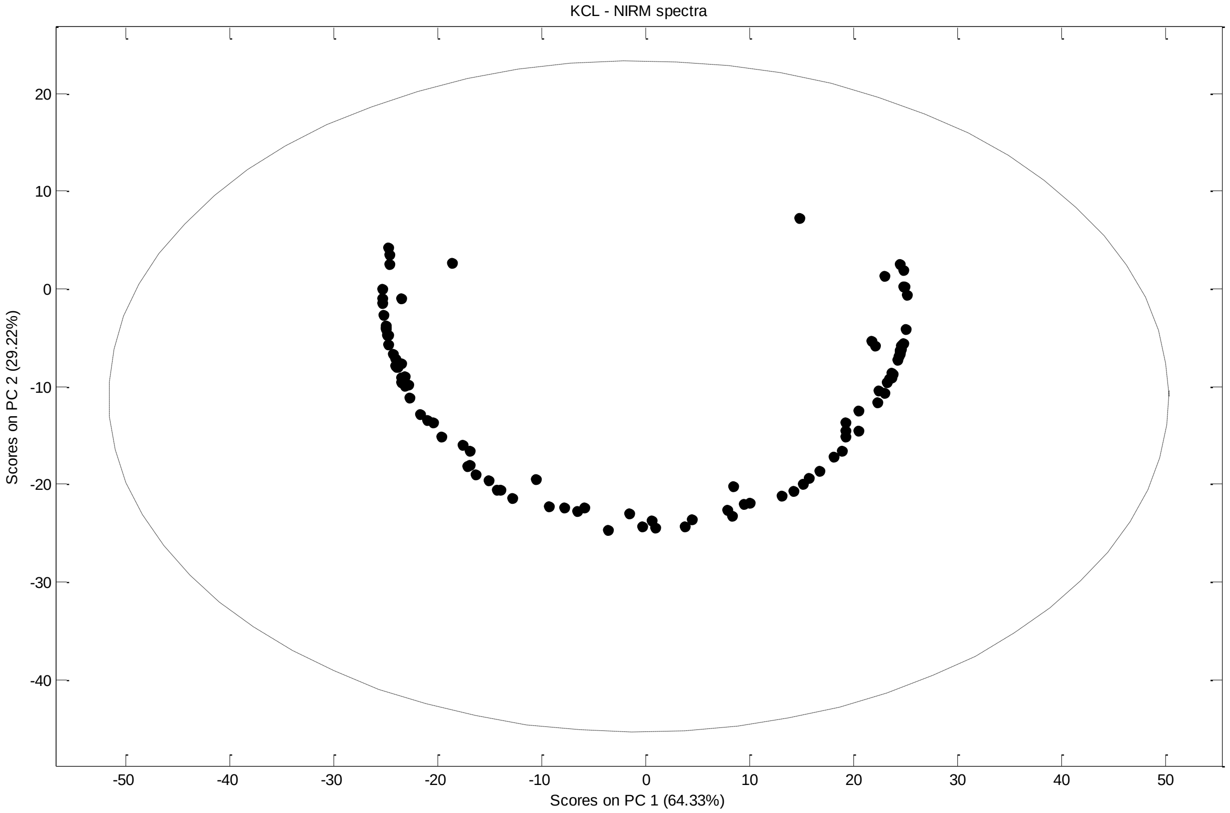




Protein, fat, fiber & starch in feed

'Calibration Transfer from Dispersive Instruments to Handheld Spectrometers', J.A. Fernández Pierna, P. Vermeulen, B. Lecler, V. Baeten, P. Dardenne. Applied Spectroscopy 64 (6) (2010)

Thank you for your attention
Do you have questions?



François Stevens
Juan Antonio Fernández Pierna

Walloon Agricultural Research Centre (CRA-W),
Valorization of Agricultural Products Department
Gembloux, Belgium