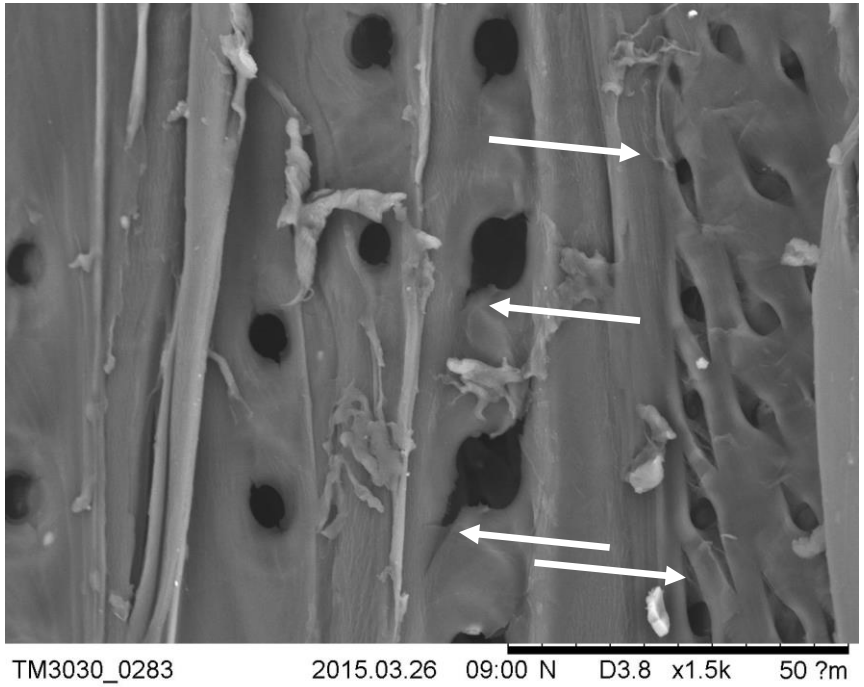
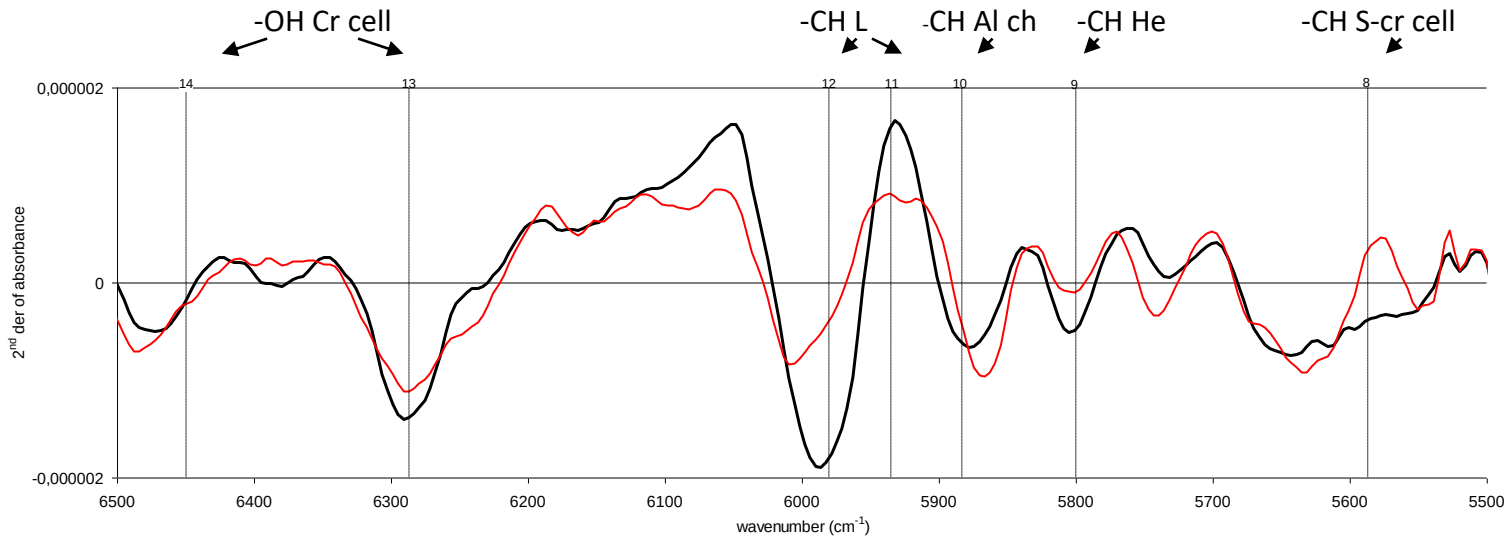
Modelling practice: data mining

Jakub & Anna Sandak
CNR-IVALSA, San Michele all Adige, Italy
University of Primorska, Koper, Slovenia

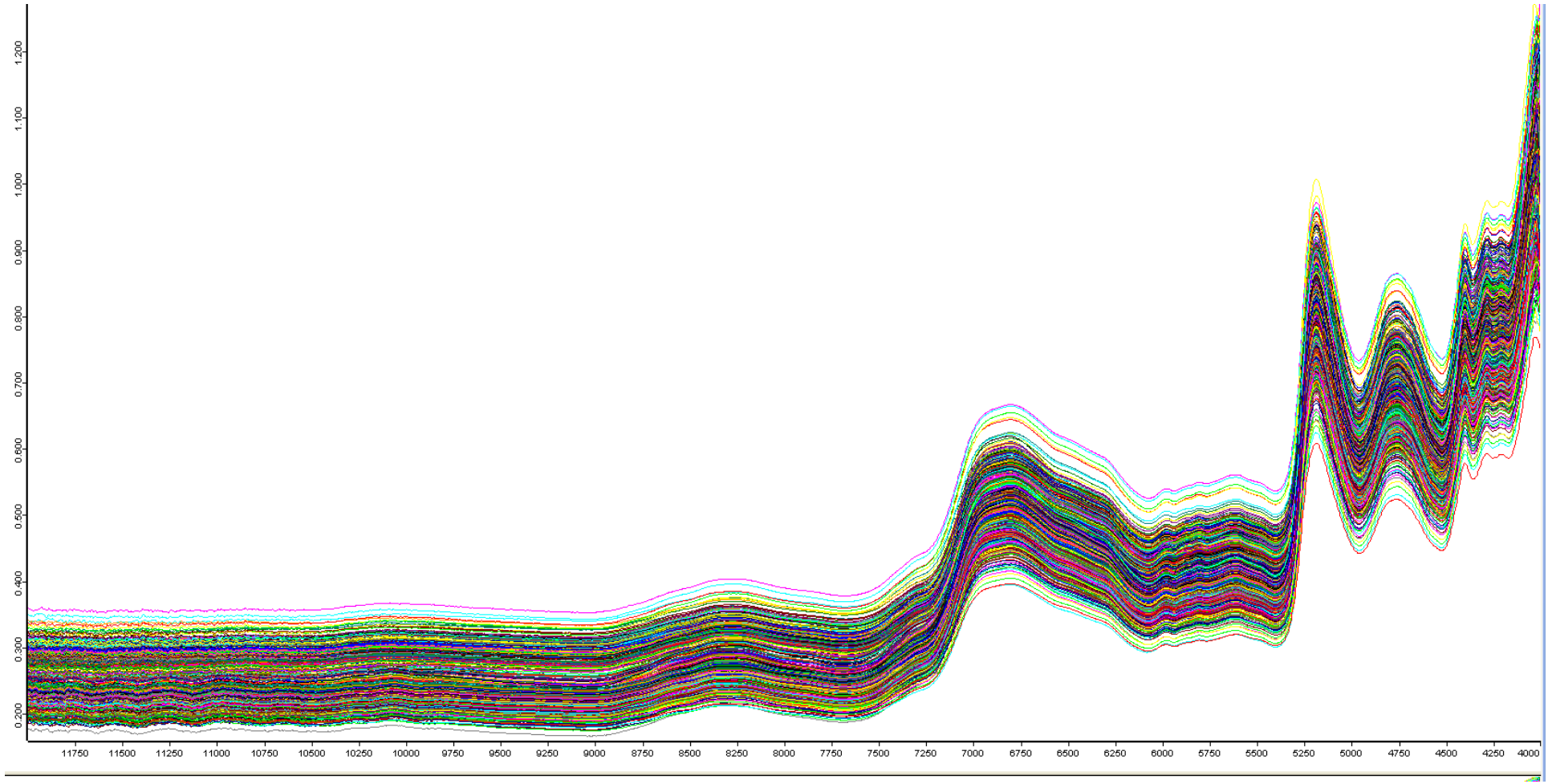
FP1407 Training School: Service life of modified wood - Understanding Test Methodologies

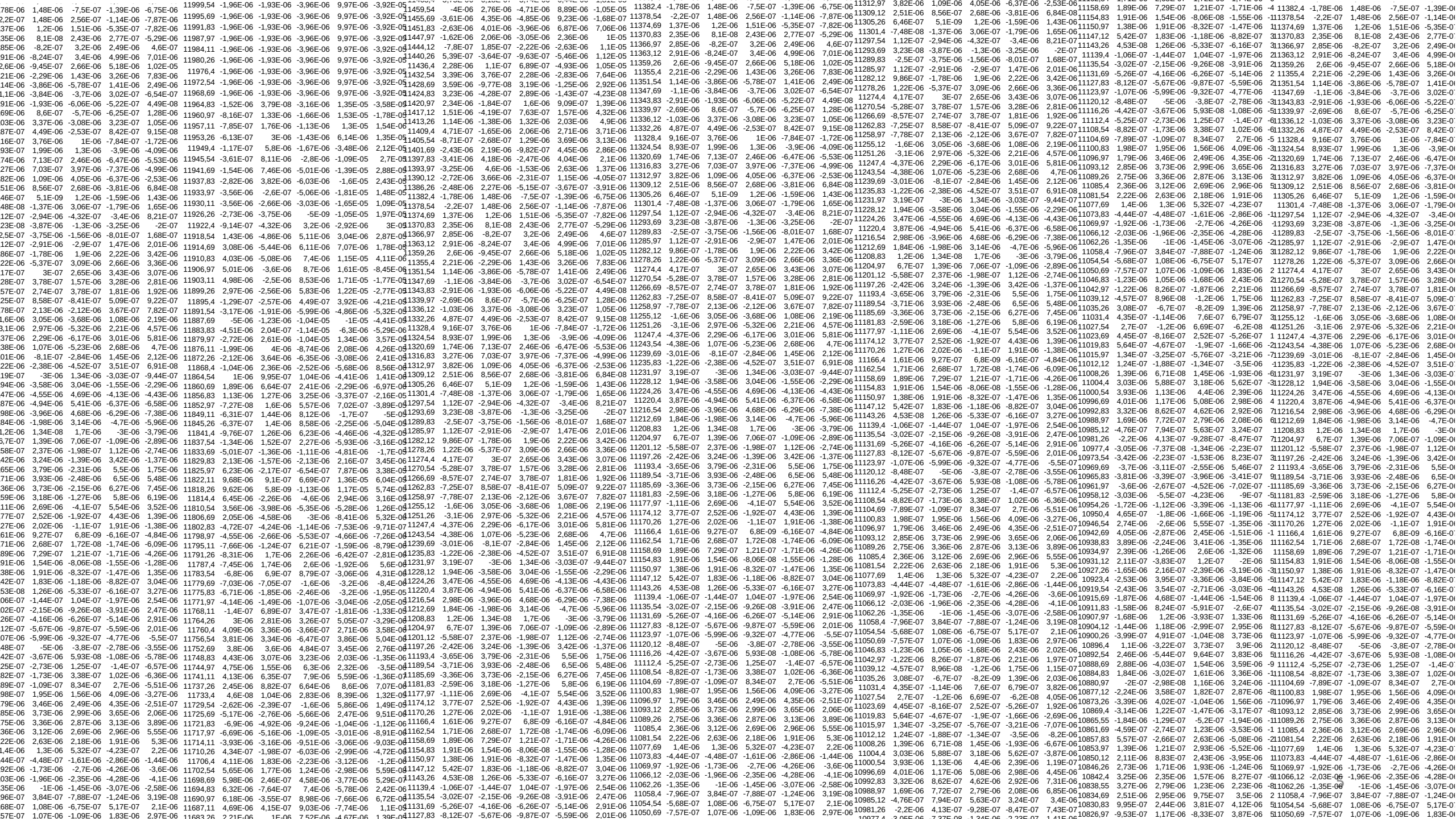

001A
28 days
001L


TM3030_0283
2015.03.26
09:00 N
D3.8
x1.5k
50 ?m


-OH Cr cell
-CH L
-CH Al ch
-CH He
-CH S-cr cell
14
13
12
11
10
9
8
0,000002
0
-0,000002
2nd der of absorbance
6500
6400
6300
6200
6100
6000
5900
5800
5700
5600
5500
wavenumber (cm-1)

When we have more spectra to be analyzed…





but…

Data is not the same as information

Too much data - too little information

(Harald Martens)

Chemometrics

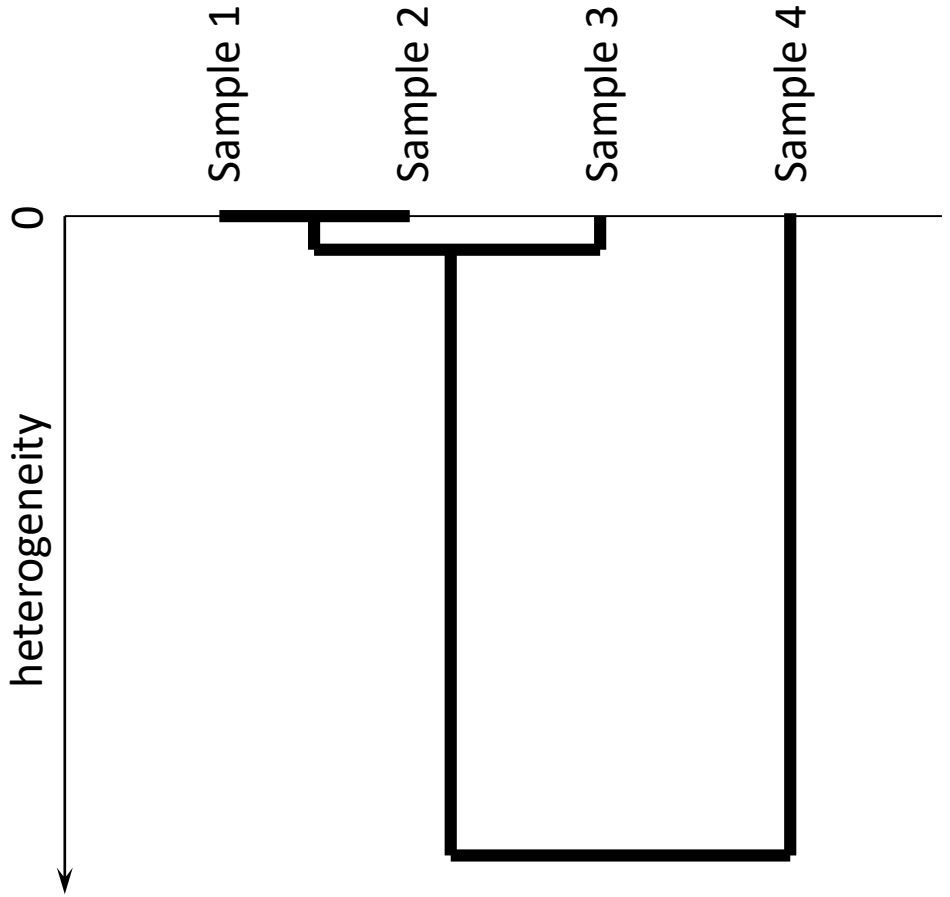
the **chemical** discipline that **uses** mathematical and statistical methods to design or select optimal procedures and experiments, and to provide maximum **chemical** information by analyzing **chemical** data

definition of the Chemometrics Society

Multivariate data analysis

- **Exploratory data analysis** (data mining) – attempts to find the hidden structure in large complex data sets
  - Cluster analysis
  - Principal Component Analysis
- **Regression analysis and Predictive Models** (developing the models from available data and predict desired response)
  - Partial Least Squares Regression
  - Multiplicative Linear Regression
- **Classification Models** (separation of group of object into one or more classes based on distinguished characteristic)
  - Cluster Analysis Test
  - Identity Test
  - SIMCA

Cluster analysis: CA

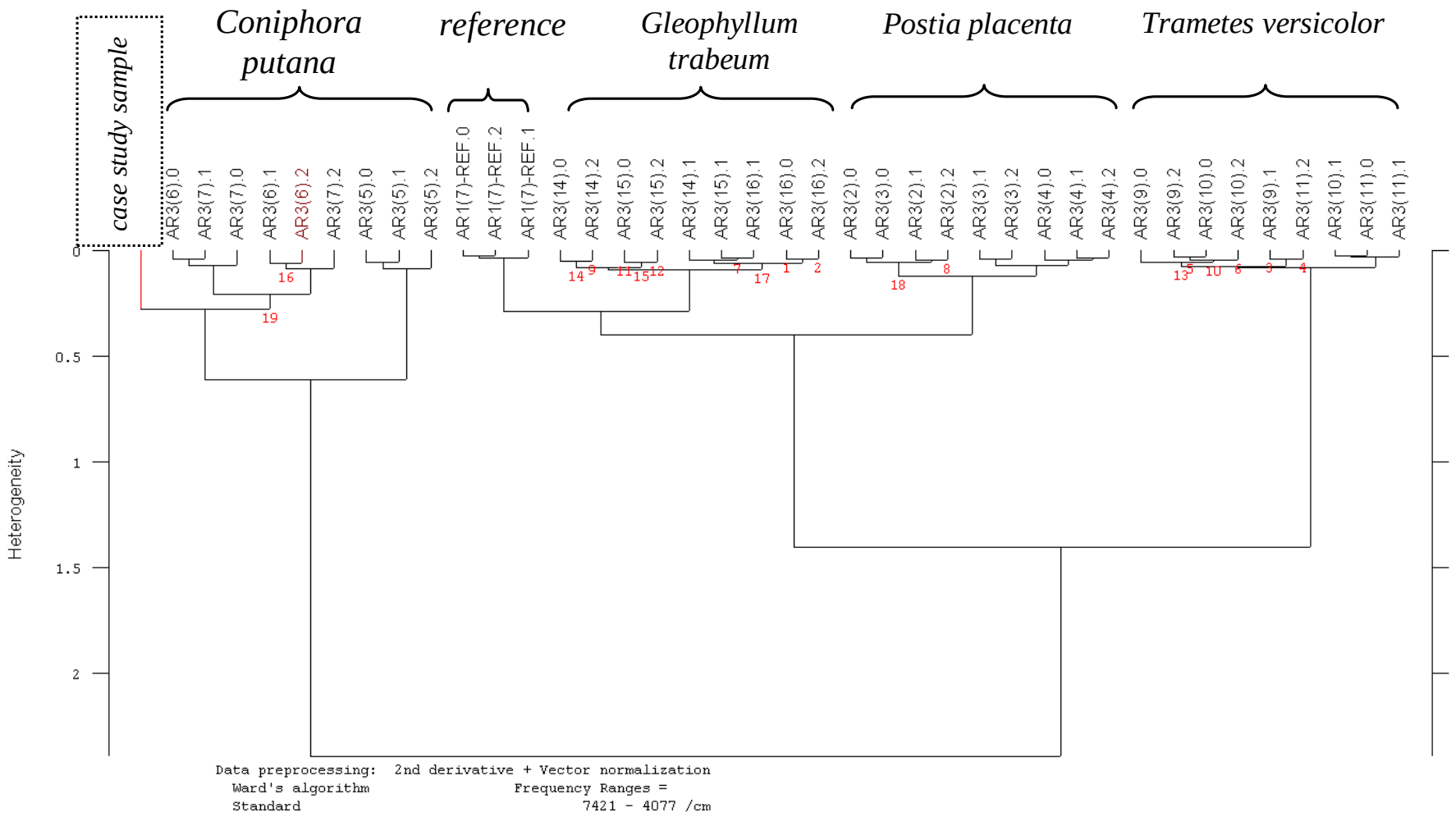


analyze spectra for their similarity, divide most similar spectra into groups, which are called clusters or classes

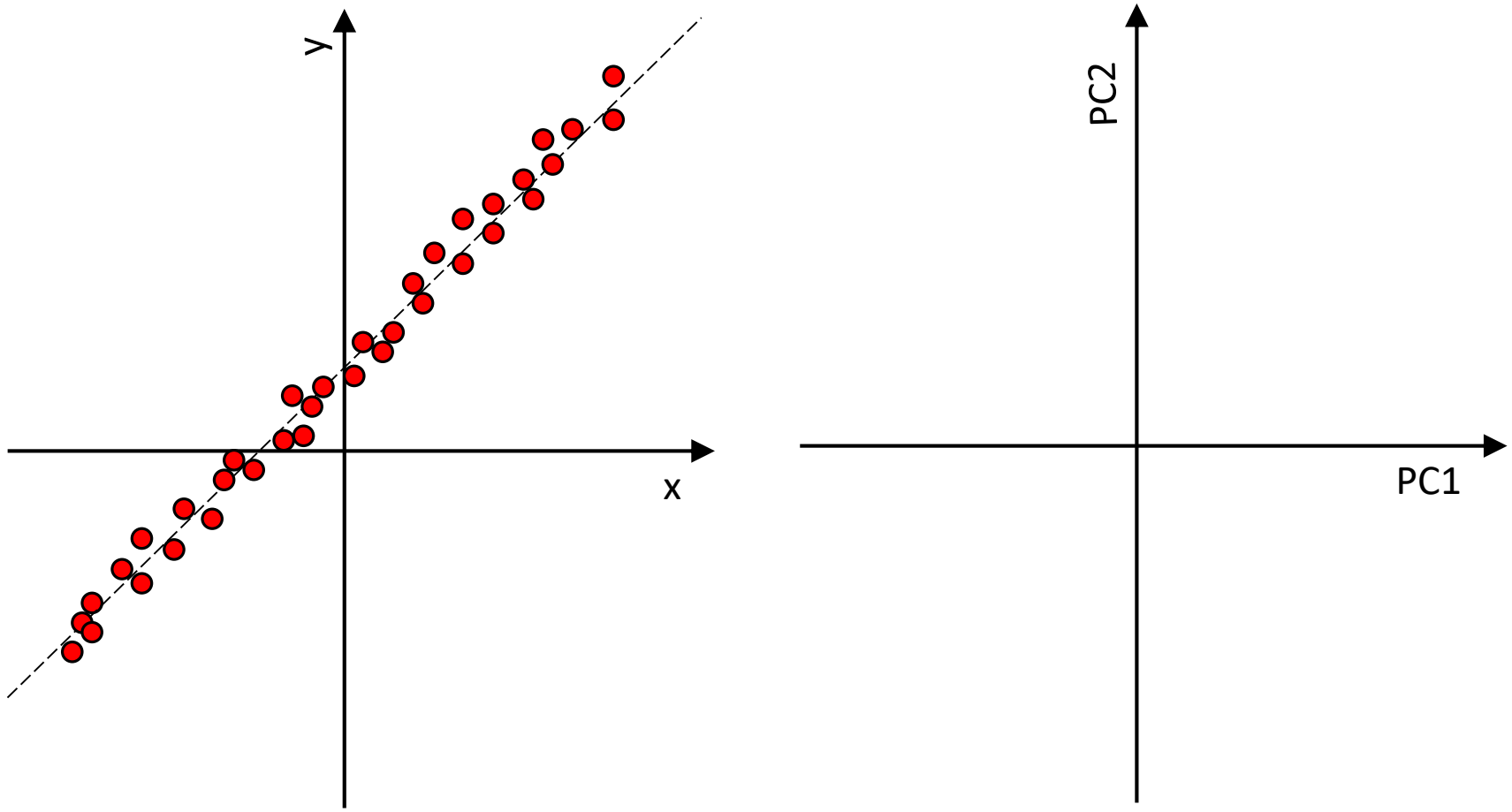
The spectral distance – heterogeneity explains the similarity between the spectra

The spectral distance 0 means that the spectra are identical

CA example


case study sample
Coniphora putana
reference
Gleophyllum trabeum
Postia placenta
Trametes versicolor
AR3(6).0
AR3(7).1
AR3(7).0
AR3(6).1
AR3(6).2
AR3(7).2
AR3(5).0
AR3(5).1
AR3(5).2
AR1(7)-REF.0
AR1(7)-REF.2
AR1(7)-REF.1
AR3(14).0
AR3(14).2
AR3(15).0
AR3(15).2
AR3(14).1
AR3(15).1
AR3(16).1
AR3(16).0
AR3(16).2
AR3(2).0
AR3(3).0
AR3(2).1
AR3(2).2
AR3(3).1
AR3(3).2
AR3(4).0
AR3(4).1
AR3(4).2
AR3(9).0
AR3(9).2
AR3(10).0
AR3(10).2
AR3(9).1
AR3(11).2
AR3(10).1
AR3(11).0
AR3(11).1
Heterogeneity
0
0.5
1
1.5
2
Data preprocessing: 2nd derivative + Vector normalization
Ward's algorithm
Standard
Frequency Ranges =
7421 - 4077 /cm

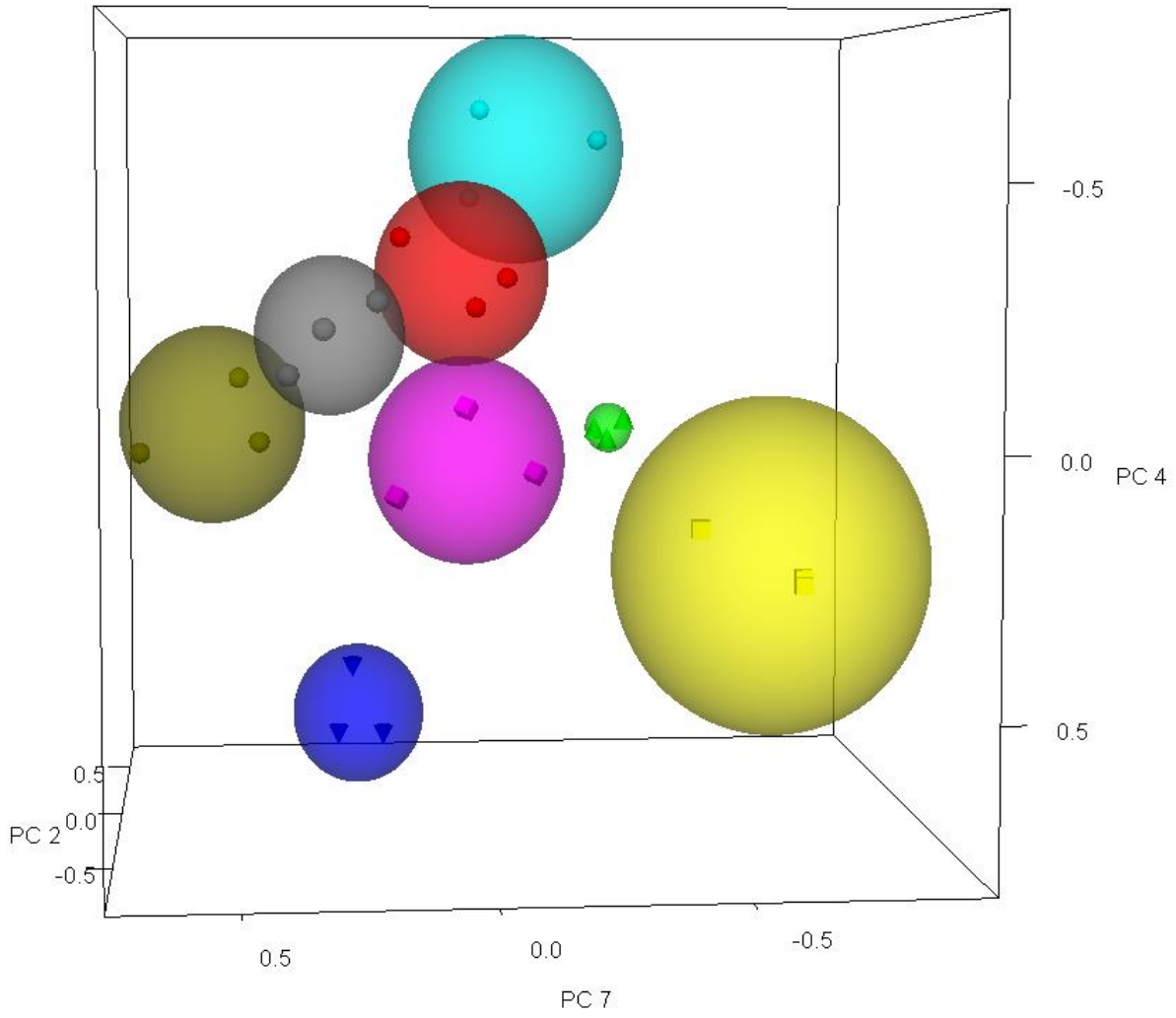
Principal Component Analysis



It is used for de-correlation of highly correlated data to reduce multidimensional data set to lower dimensions

it can separate set of input data into groups of peculiar similarities

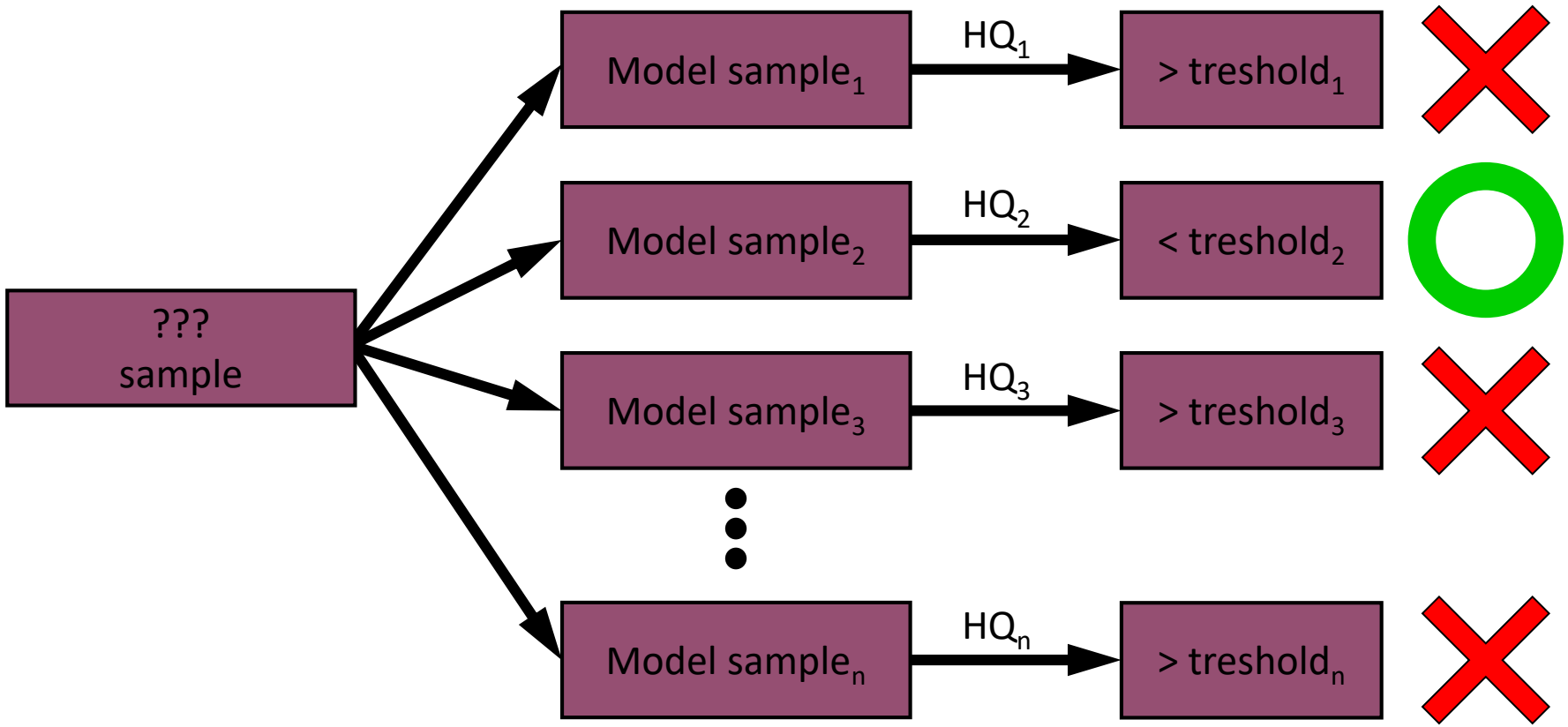
PCA example



PCA of NIR spectra of non coated samples exposed to South (4 years of natural weathering);

- Reggio Emilia
- Roma
- Loninghens
- Macerata
- Udine
- Trento
- Lecce
- standard

Identity test

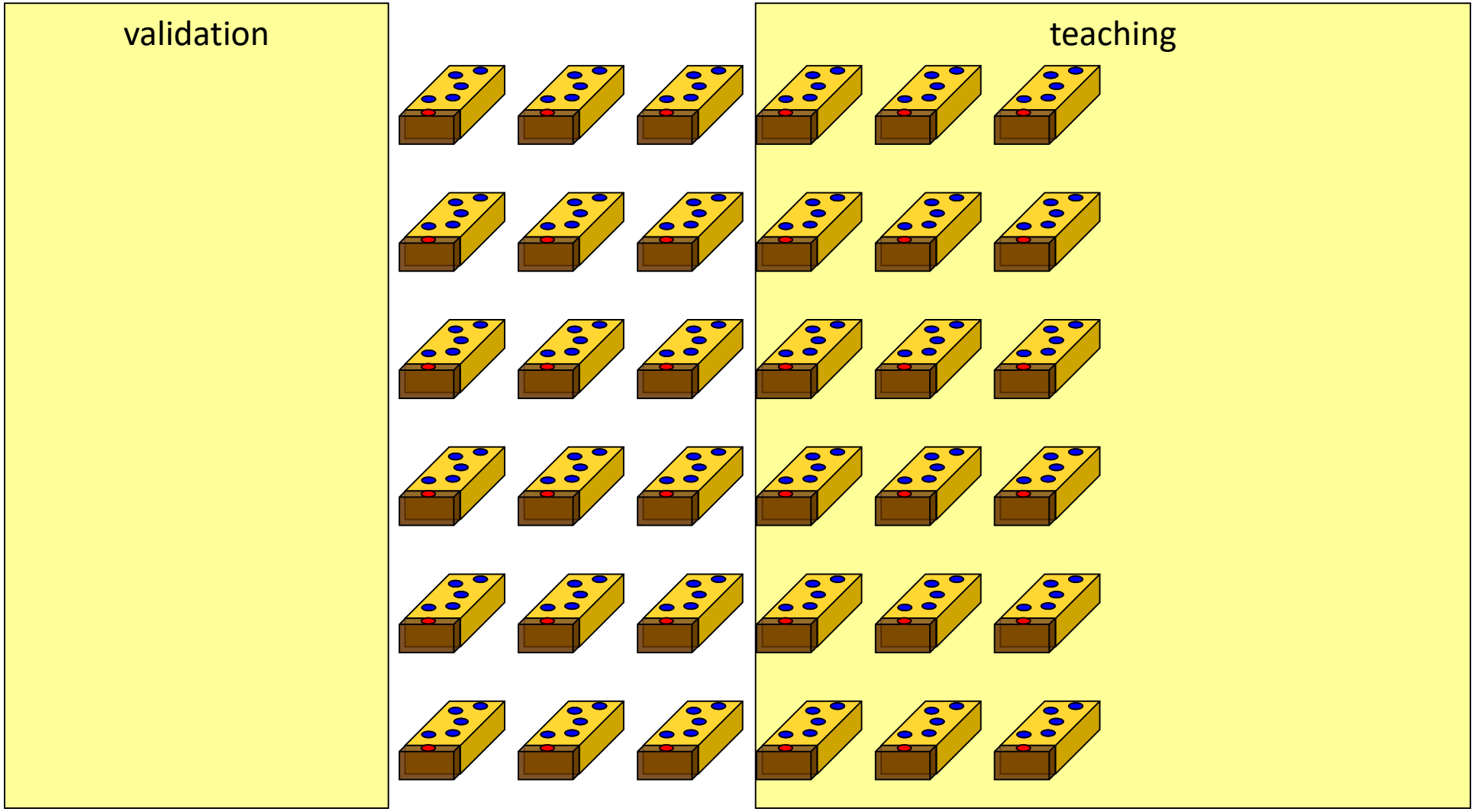


Compare the unknown spectrum with all reference spectra

The result of comparison between two spectra is the spectral distance called hit quality

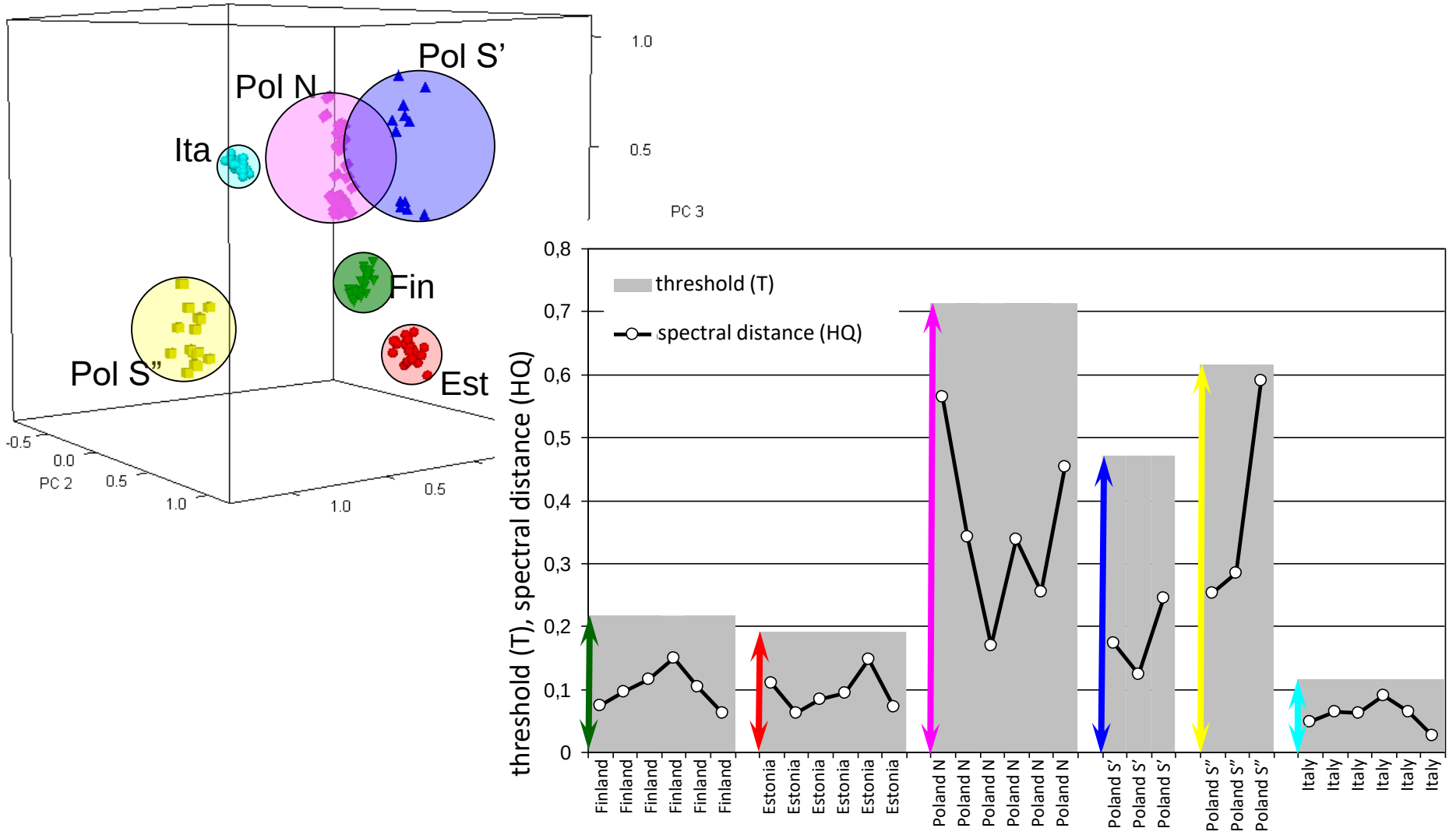
The better spectra match the smaller is spectral distance; HQ for identical spectra is 0

Identity test (IT)

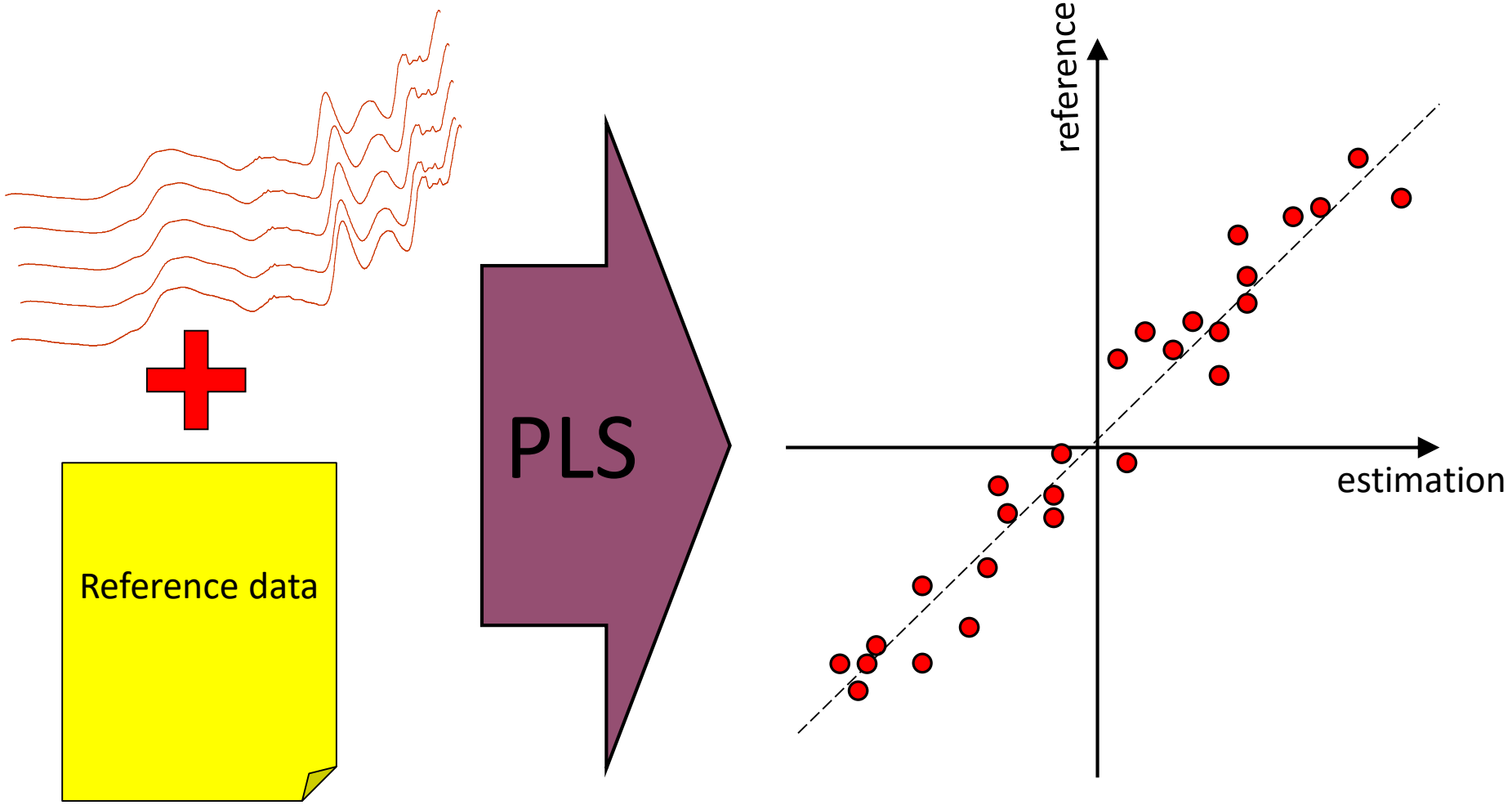


20% for validation (80% for building model)

Identity test (IT)



Partial Last Squares

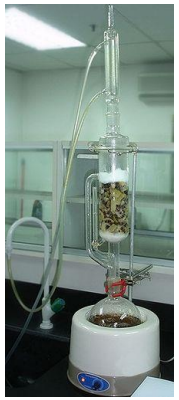


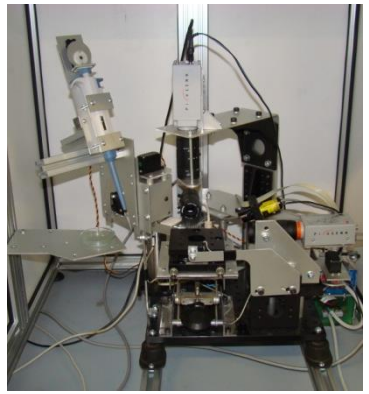
How it works?

spectra

+

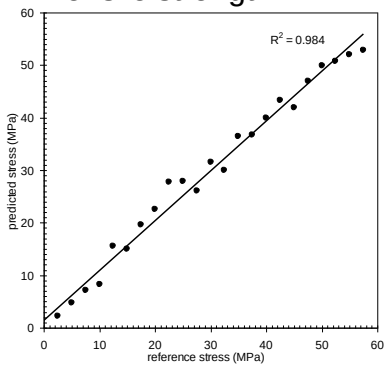
reference data



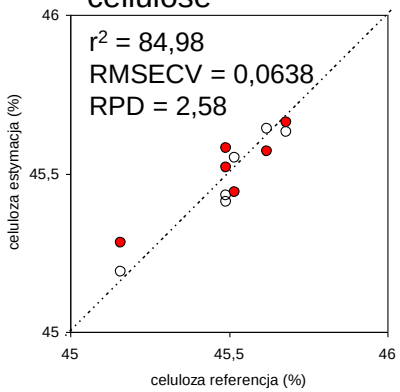


calibration (PLS)

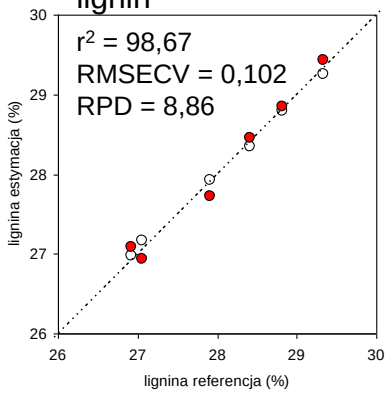
Tensile strength



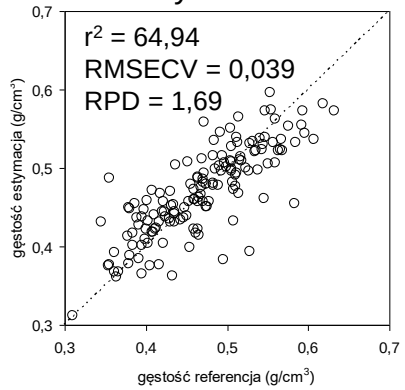
cellulose



lignin



density



PLS in practice