Raman Data Analysis

GeoRaman 2012 International School

14th-16th of June 2012, Nancy

RAMAN SPECTROSCOPY APPLIED TO EARTH SCIENCES AND CULTURAL HERITAGE

N.Tarcea[1], Jürgen Popp [1, 2]

[1] Institute of Physical Chemistry, University of Jena, Germany

[2] Institute of Photonic Technology, Jena, Germany

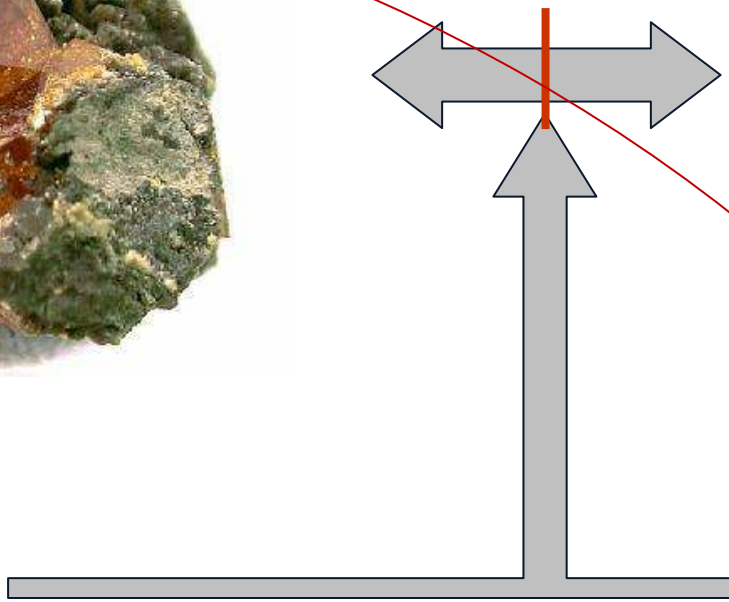
Overview

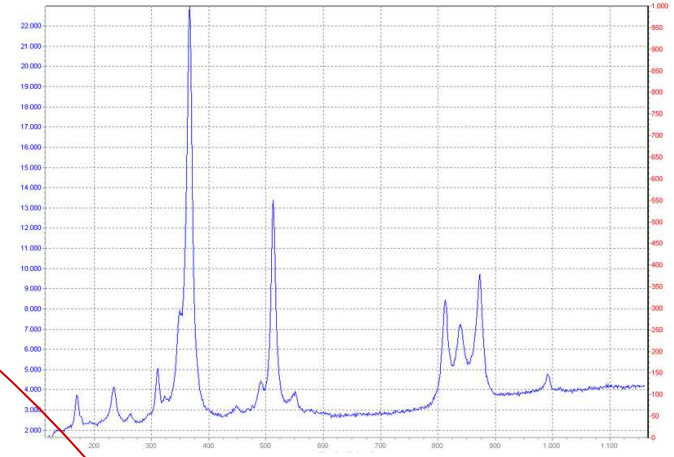
- Data Acquisition and storage
  - Acquisition
  - Storage & organization (spectral libraries)
- Pre-processing of Raman spectra
  - Spike removal
  - Smoothing/denoising
  - Background
  - Normalization
  - Spectral and intensity (re-)calibration
- Chemometry methods for Raman spectroscopy
  - Multivariate calibration algorithms (MLR, PCA, PCR)
  - Classification methods
  - Cluster analysis

Overview

- Data Acquisition and storage
  - Acquisition
  - Storage & organization (spectral libraries)
- Pre-processing of Raman spectra
  - Spike removal
  - Smoothing/denoising
  - Background
  - Normalization
  - Spectral and intensity (re-)calibration
- Chemometry methods for Raman spectroscopy
  - Multivariate calibration algorithms (MLR, PCA, PCR)
  - Classification methods
  - Cluster analysis

Data Acquisition



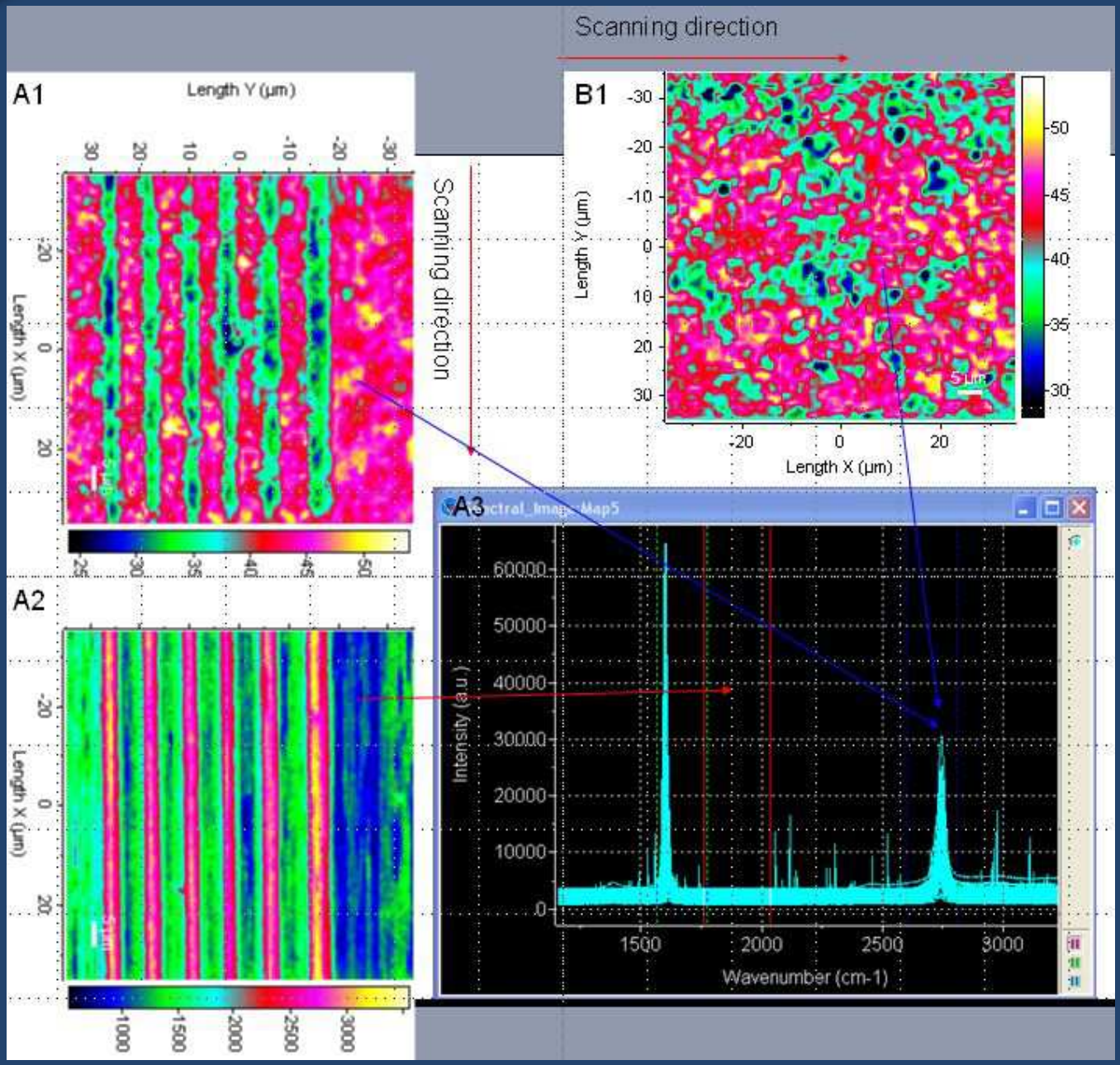


Data Acquisition

Effort needed to control the data acquisition process depends on intended data usage.

Will all Raman data be measured on a single device?
Do I need to compare data with data measured on other devices? (degree of comparison?)
Is data meant for a spectral database or building a robust calibration model (chemometrics) ?
....

Facts:
- vast majority of Raman devices make use of multichannel detection (e.g. using CCDs) - pixel uniformity?
- modern Raman instruments are delivered with complex software packages
  => degree of data preprocessing not always known (and/or understood) by the user
  e.g. automatic calibration
  spectral frames joining with offsets
  spike removal
  background subtraction, intensity corrections etc.
- spectral accuracy and stability changes (e.g. due to mechanical wear) - is your instrument still in specs?
- degraded spectral calibration by grating movement or environmental changes (lab temperature)
- instrument throughput characteristics might depend on light polarization (e.g. due to grating)
- crystal sample orientation
- laser wavelength might shift (e.g. diode laser)
- Raman band shape convoluted with the instrument function (Gauss, Triangular etc.)
- Variation in Raman spectra is induced by factors which are not always obvious
.....


Scanning direction
A1
Length Y (µm)
Length X (µm)
Scanning direction
B1
Length Y (µm)
Length X (µm)
5 µm
A2
Length X (µm)
A3
Intensity (a.u.)
Wavenumber (cm-1)

Data organisation/storage

Different types of Raman data:
- simple spectrum (intensity vs. relative wavenumber)
- hyperspectral data (multiple dimensions: intensity, wavenumber, spatial coordinates, time)

Different file formats:
- proprietary file formats (instrument manufacturer)
- industry standard formats (GRAMS format - .spc, ascii format)

Information saved within a datafile:
- spectral (hyperspectral) data
- metadata (instrument configuration parameters, user comments, user defined parameters)

Data organization:
- File system (project directories) data links in
  - experiment log book
  - descriptive filename/directories
  - extra metadata items (file headers)

Data organisation/storage

Relational databases (make use of SQL engines capabilities - MS Access, MySQL, MSSQL)

Standard approach

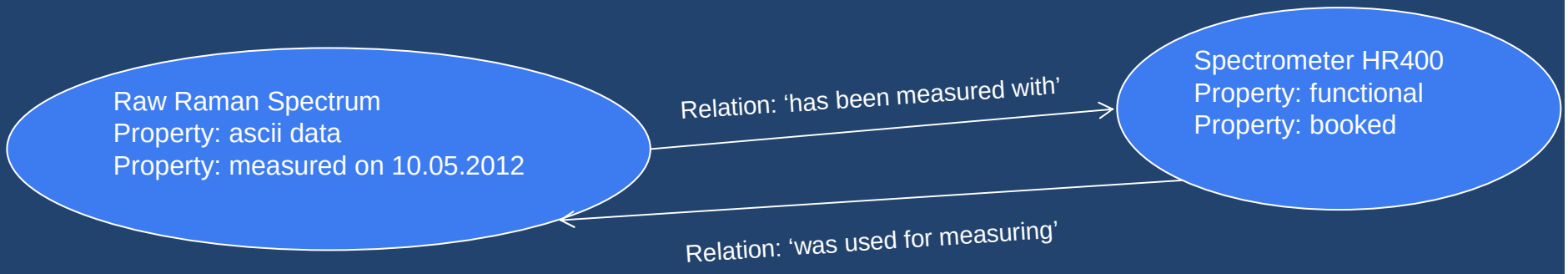
- each object type and each relation has a dedicated table
- rigid object structure
- effective data access and searching

Entity-Attribute-Value (EAV) implementation

- object structure and relations are defined as metadata inside the database
- tables hold data split on their type (strings, integers, pictures etc.)
- complex maintenance and user interface
- complex constructs for searching and data mining
- effective data access and searching
- flexible object structure

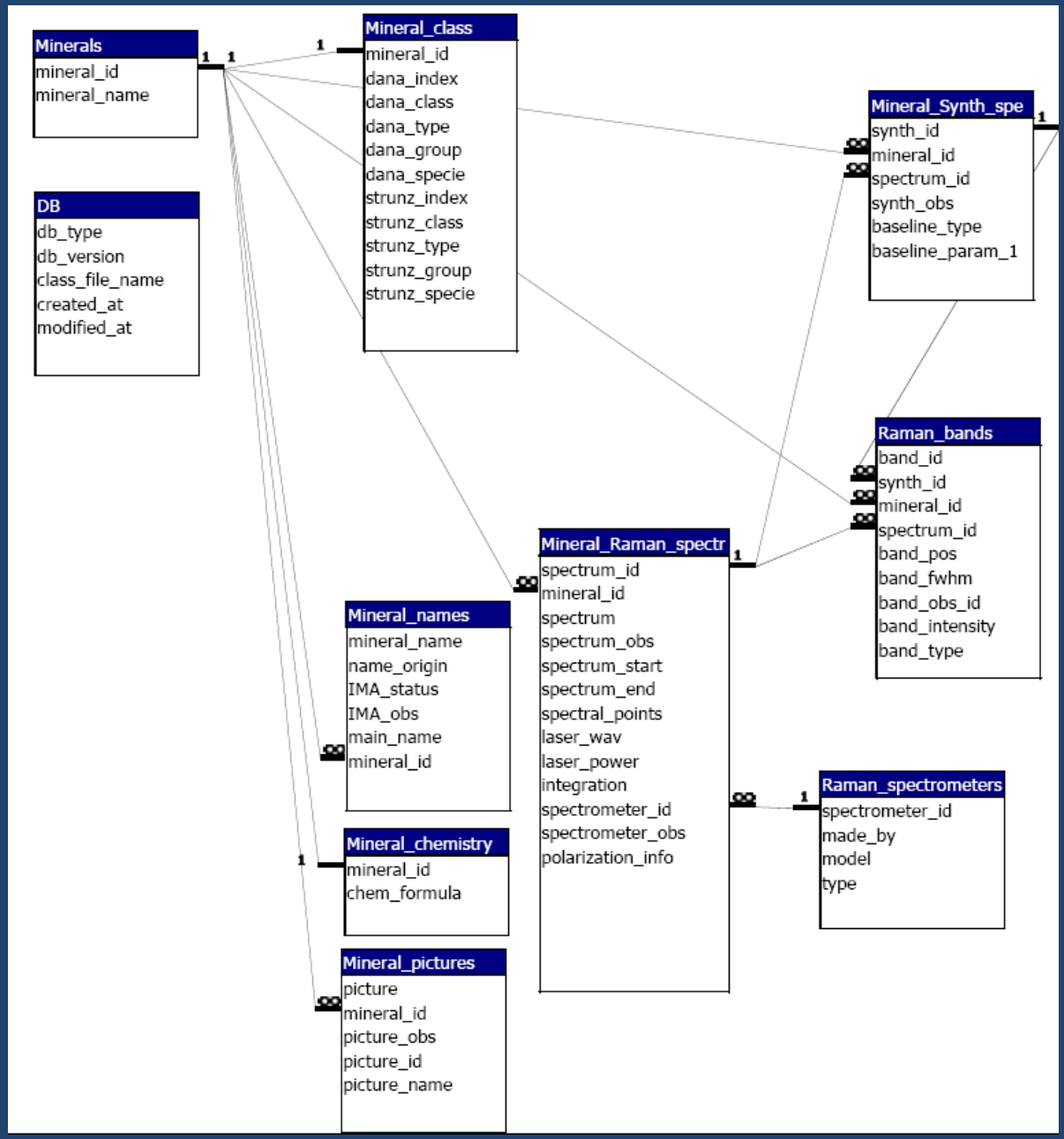
Graph databases (free schema database) (tools: Neo4J, InfoGrid, HyperGraph, AllegroGraph etc.)

- nodes, properties and edges in a graph like structure
- not in use for now by spectroscopic applications
- used for bio and medical data, social networks (facebook @ co.)



Databases

Mineral centric



Desktop Raman Database - [gypsum]

Database Search Edit View Window Help

mineral_test_db.mdb

main name: gypsum

gypsum

View Mineral Found

Mineral name
chalcedony
chert
chrysoprase
citrine
clevelandite
cyrtolite
davisonite
demantoid
dennisonite
flint
fluorapatite
gypsum
hornstone
hyacinth
jarcon
jasper
magnesite
malacon
melanite
moss agate
olivenite
plasma
prase

gypsum

Name\Class | Phys\Chem Properties | Raman spectra | Christalography | Observations

Mineral name	Main name	IMA status
gypsum	True	False
alabaster	False	False
satin spar	False	False
selenite	False	False

Name origin:

From the Greek, gyps meaning "burned" mineral.

IMA observations:

N/A

Strunz Clasification

VI/C.22-20

Class: **VI. - Sulfates, Chromates, Molybdates and Tungstates**

Type: **C. - Water-bearing sulfates without unfamiliar anions. cations of medium size**

Group: **22. - Cations of very big size**

Specie: **20. - Gypsum**

Dana Clasification

29.6.3.1

Class: **29. - Hydrated Acid and Sulfates**

Type: **6. - Hydrated Acid and Sulfates where A XO4 - x(H2O)**

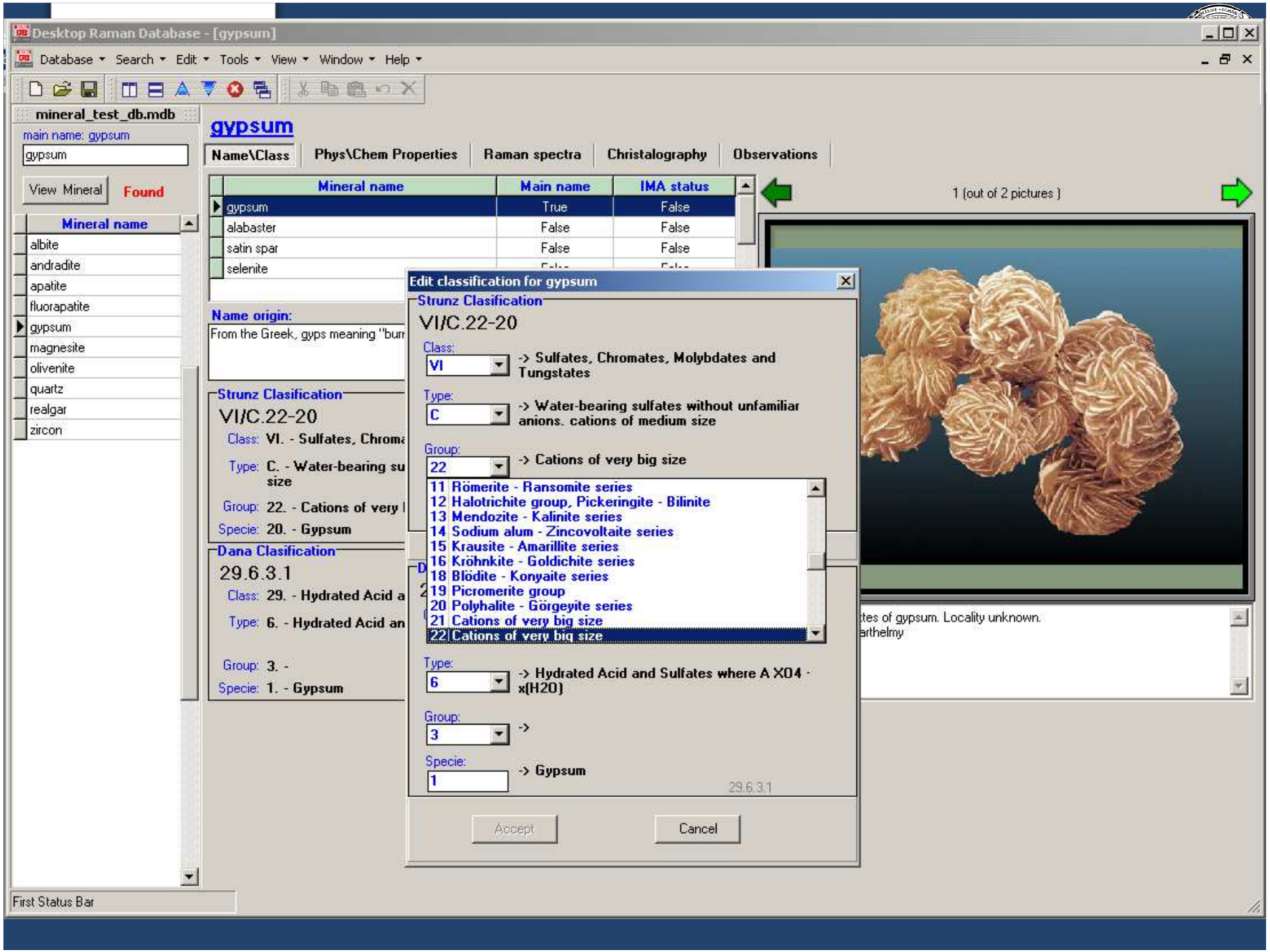
Group: **3. -**

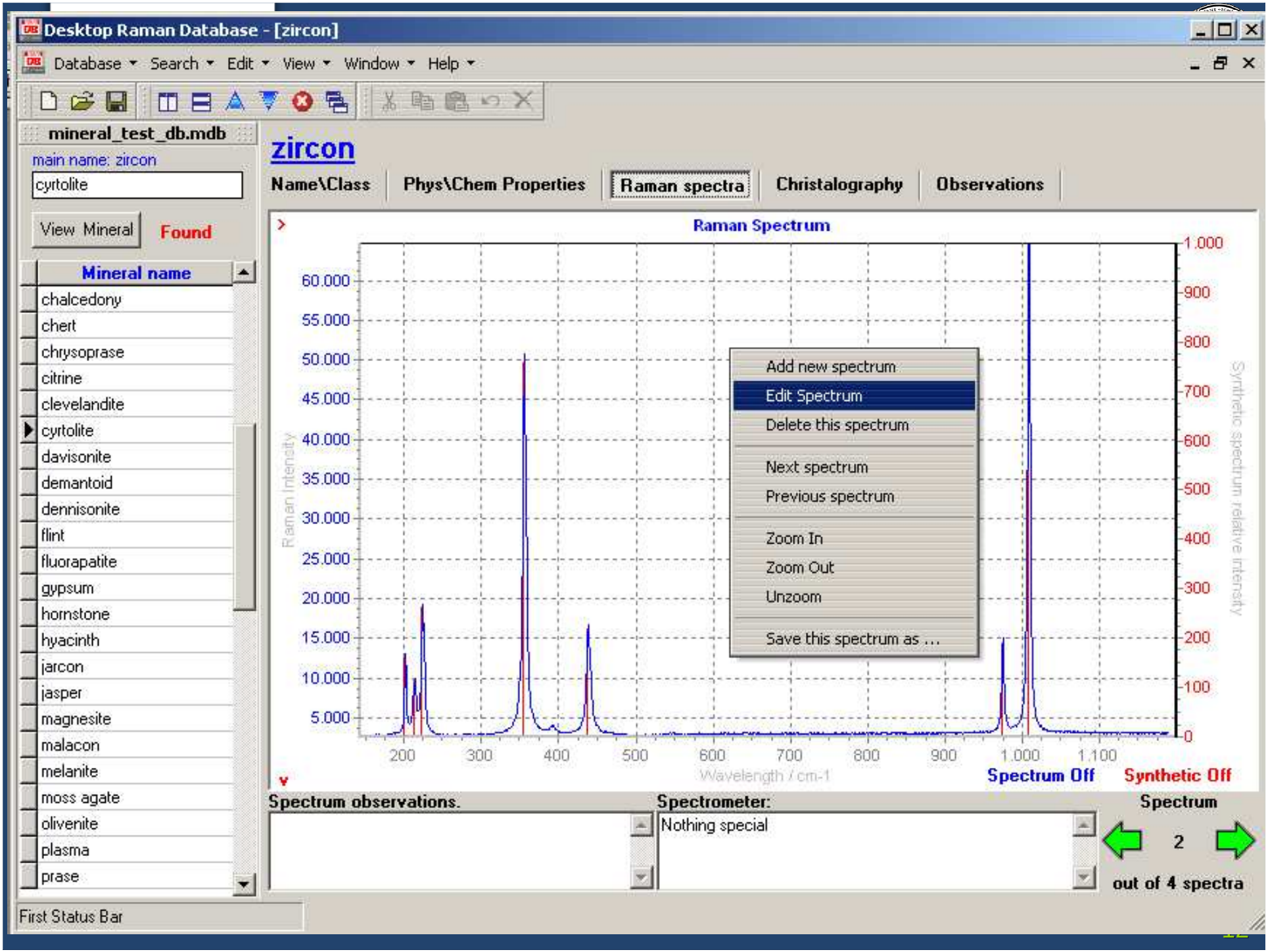
Specie: **1. - Gypsum**

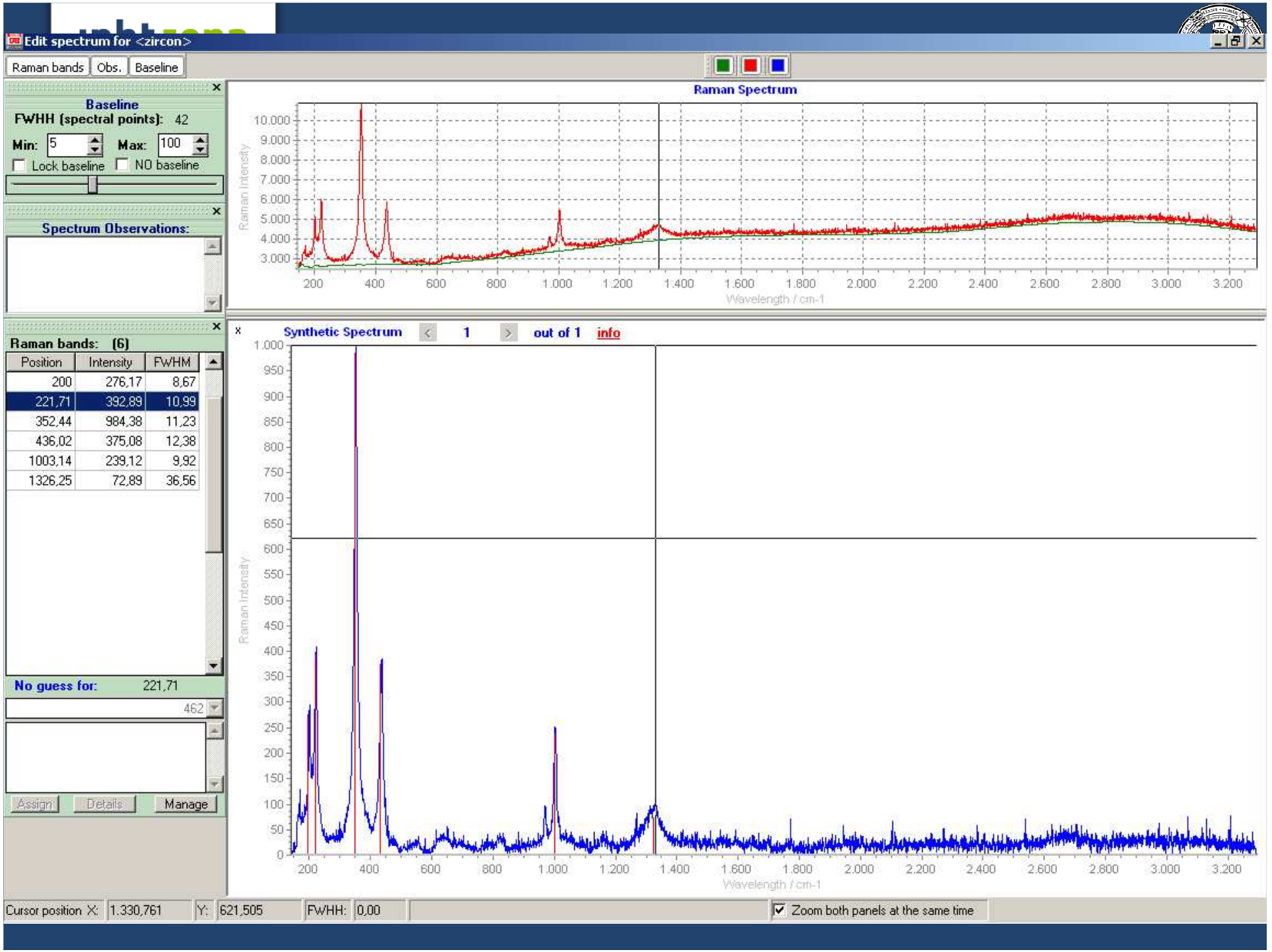
1 (out of 2 pictures)

6 cm bladed rosettes of gypsum. Locality unknown.
Photo by Dave Barthelmy

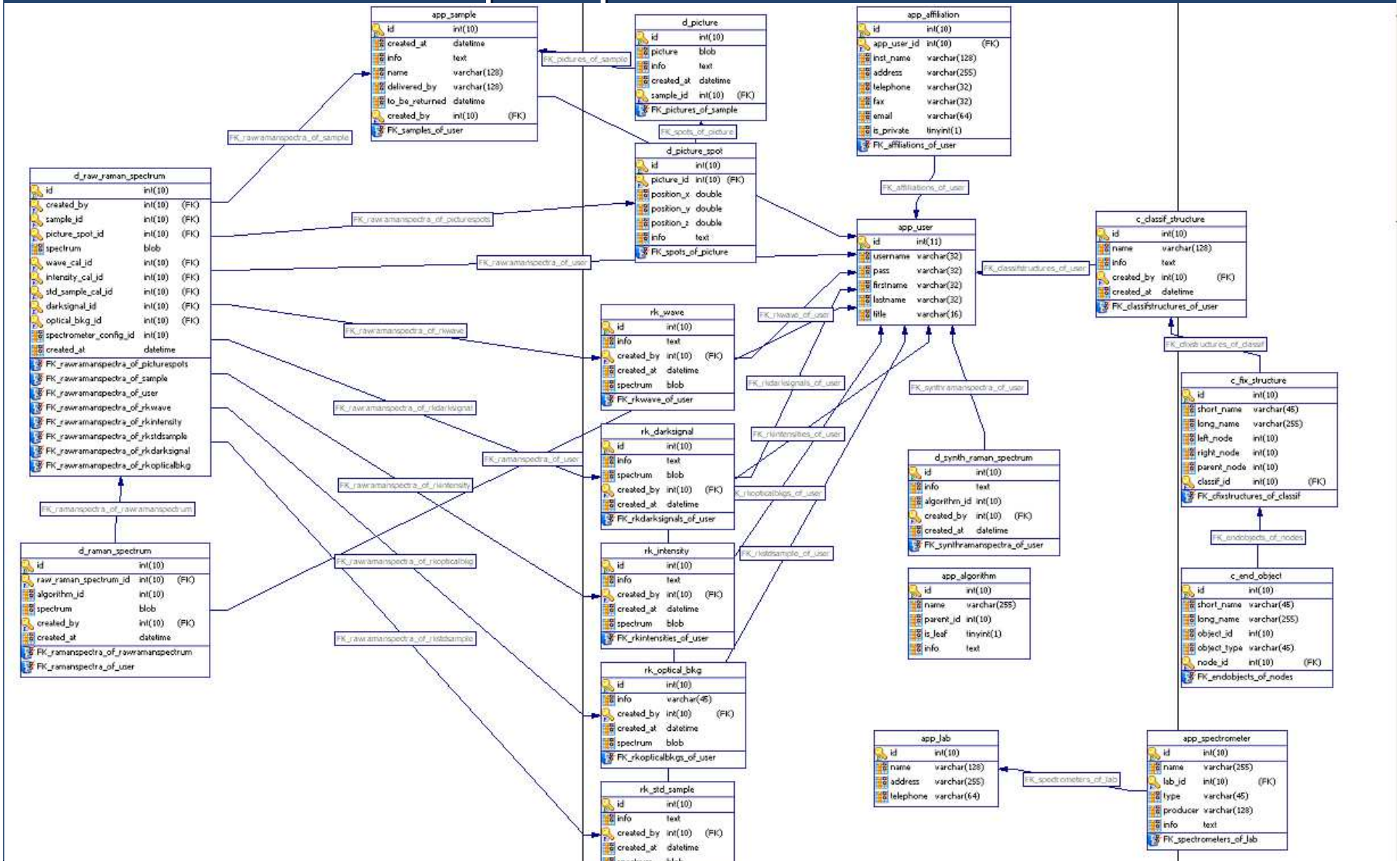
First Status Bar


Desktop Raman Database - [gypsum]
Database Search Edit Tools View Window Help
mineral_test_db.mdb
main name: gypsum
gypsum
View Mineral
Found
Mineral name
albite
andradite
apatite
fluorapatite
gypsum
magnesite
olivenite
quartz
realgar
zircon
gypsum
Name\Class
Phys\Chem Properties
Raman spectra
Christalography
Observations
Mineral name
Main name
IMA status
gypsum
True
False
alabaster
False
False
satin spar
False
False
selenite
1 (out of 2 pictures)
Name origin:
From the Greek, gyps meaning "burn
Strunz Clasification
VI/C.22-20
Class: VI. - Sulfates, Chroma
Type: C. - Water-bearing su
size
Group: 22. - Cations of very l
Specie: 20. - Gypsum
Dana Clasification
29.6.3.1
Class: 29. - Hydrated Acid a
Type: 6. - Hydrated Acid an
Group: 3. -
Specie: 1. - Gypsum
tes of gypsum. Locality unknown.
arthelmy
Edit classification for gypsum
Strunz Clasification
VI/C.22-20
Class:
VI
-> Sulfates, Chromates, Molybdates and Tungstates
Type:
C
-> Water-bearing sulfates without unfamiliar anions. cations of medium size
Group:
22
-> Cations of very big size
11 Römerite - Ransomite series
12 Halotrichite group, Pickeringite - Bilinite
13 Mendozite - Kalinite series
14 Sodium alum - Zincovoltaite series
15 Krausite - Amarillite series
16 Köhnkite - Goldichite series
18 Blödite - Konyaite series
19 Picromerite group
20 Polyhalite - Görgeyite series
21 Cations of very big size
22 Cations of very big size
Type:
6
-> Hydrated Acid and Sulfates where A XO4 - x(H2O)
Group:
3
->
Specie:
1
-> Gypsum
29.6.3.1
Accept
Cancel
First Status Bar

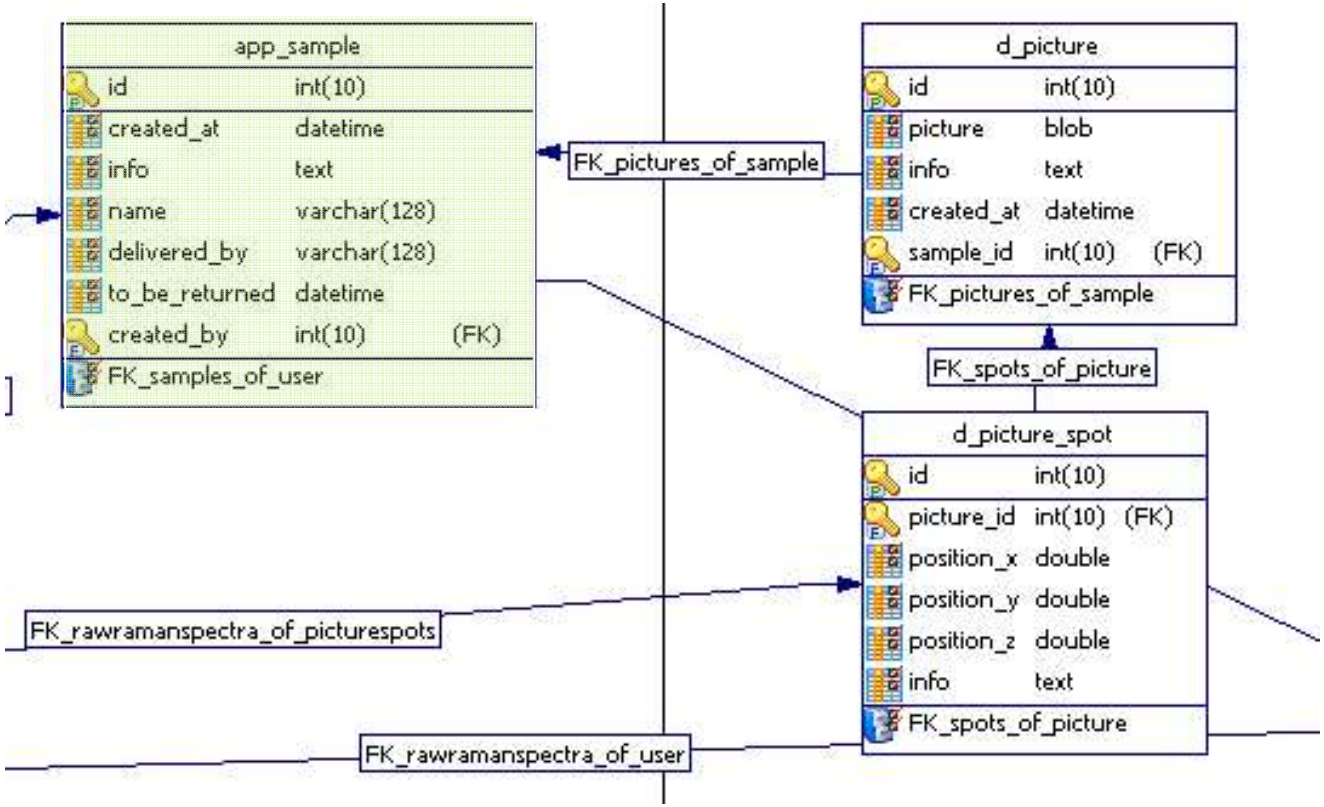

Desktop Raman Database - [zircon]
Database
Search
Edit
View
Window
Help
mineral_test_db.mdb
main name: zircon
cyrtolite
View Mineral
Found
Mineral name
chalcedony
chert
chrysoprase
citrine
clevelandite
cyrtolite
davisonite
demantoid
dennisonite
flint
fluorapatite
gypsum
hornstone
hyacinth
jarcon
jasper
magnesite
malacon
melanite
moss agate
olivenite
plasma
prase
zircon
Name\Class
Phys\Chem Properties
Raman spectra
Christalography
Observations
Raman Spectrum
Add new spectrum
Edit Spectrum
Delete this spectrum
Next spectrum
Previous spectrum
Zoom In
Zoom Out
Unzoom
Save this spectrum as ...
Raman Intensity
Synthetic spectrum relative intensity
Wavelength / cm-1
Spectrum Off
Synthetic Off
Spectrum observations.
Spectrometer:
Nothing special
Spectrum
2
out of 4 spectra
First Status Bar

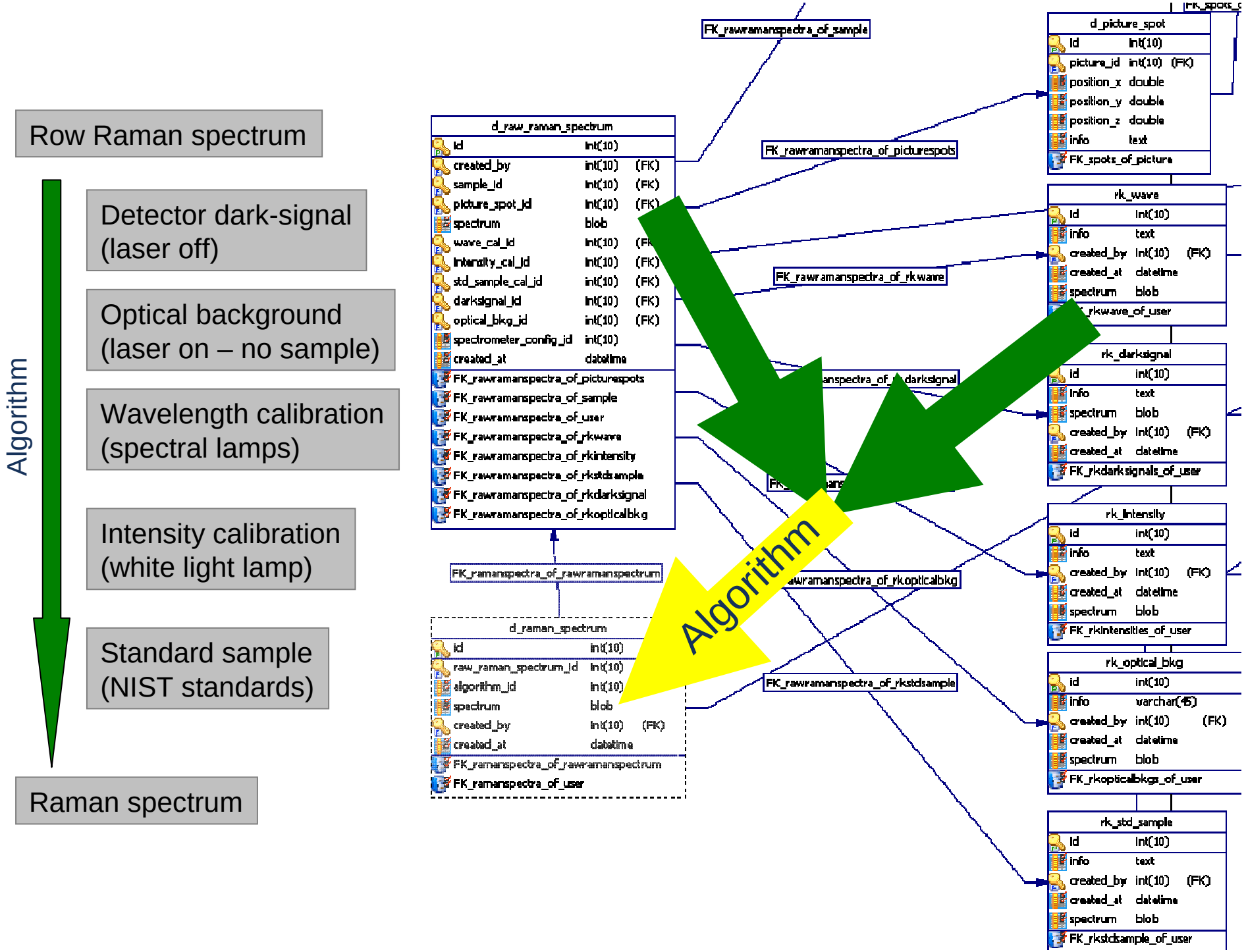

Edit spectrum for <zircon>
Raman bands
Obs.
Baseline
Baseline
FWHH (spectral points): 42
Min: 5
Max: 100
Lock baseline
NO baseline
Spectrum Observations:
Raman bands: (6)
Position Intensity FWHM
200 276,17 8,67
221,71 392,89 10,99
352,44 984,38 11,23
436,02 375,08 12,38
1003,14 239,12 9,92
1326,25 72,89 36,56
No guess for: 221,71
462
Assign
Details
Manage
Raman Spectrum
Raman Intensity
Wavelength / cm-1
Synthetic Spectrum
1
out of 1
info
Raman Intensity
Wavelength / cm-1
Cursor position X: 1.330,761
Y: 621,505
FWHH: 0,00
Zoom both panels at the same time

Databases – sample/spectrum centric



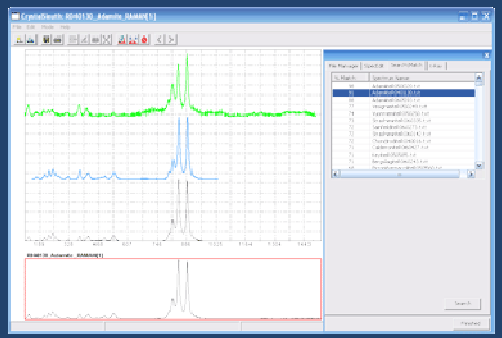
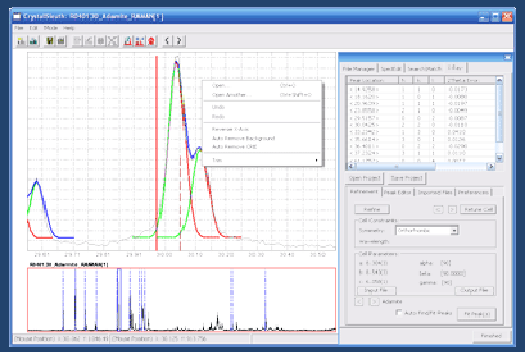
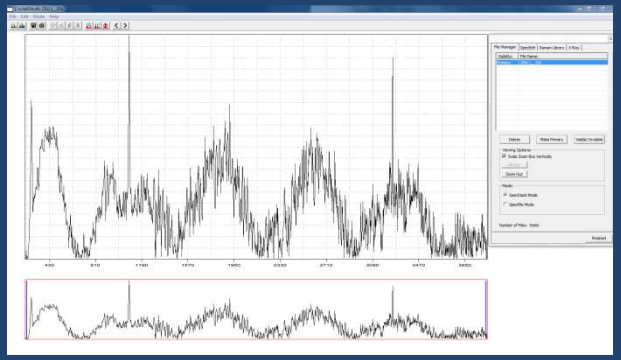
Databases – sample/spectrum centric




Row Raman spectrum
Algorithm
Detector dark-signal (laser off)
Optical background (laser on – no sample)
Wavelength calibration (spectral lamps)
Intensity calibration (white light lamp)
Standard sample (NIST standards)
Raman spectrum
Algorithm
d_raw_raman_spectrum
id int(10)
created_by int(10) (FK)
sample_id int(10) (FK)
picture_spot_id int(10) (FK)
spectrum blob
wave_cal_id int(10) (FK)
intensity_cal_id int(10) (FK)
std_sample_cal_id int(10) (FK)
darksignal_id int(10) (FK)
optical_bkg_id int(10) (FK)
spectrometer_config_id int(10)
created_at datetime
FK_rawramanspectra_of_picturespots
FK_rawramanspectra_of_sample
FK_rawramanspectra_of_user
FK_rawramanspectra_of_rkwave
FK_rawramanspectra_of_rkintensity
FK_rawramanspectra_of_rkstdsample
FK_rawramanspectra_of_rkdarksignal
FK_rawramanspectra_of_rkopticalbkg
FK_ramanspectra_of_rawramanspectrum
d_raman_spectrum
id int(10)
raw_raman_spectrum_id int(10)
algorithm_id int(10)
spectrum blob
created_by int(10) (FK)
created_at datetime
FK_ramanspectra_of_rawramanspectrum
FK_ramanspectra_of_user
FK_rawramanspectra_of_sample
FK_rawramanspectra_of_picturespots
FK_rawramanspectra_of_rkwave
FK_rawramanspectra_of_rkopticalbkg
FK_rawramanspectra_of_rkstdsample
d_picture_spot
id int(10)
picture_id int(10) (FK)
position_x double
position_y double
position_z double
info text
FK_spots_of_picture
rk_wave
id int(10)
info text
created_by int(10) (FK)
created_at datetime
spectrum blob
rk_darksignal
id int(10)
info text
spectrum blob
created_by int(10) (FK)
created_at datetime
FK_rkdarksignals_of_user
rk_intensity
id int(10)
info text
created_by int(10) (FK)
created_at datetime
spectrum blob
FK_rkintensities_of_user
rk_optical_bkg
id int(10)
info varchar(45)
created_by int(10) (FK)
created_at datetime
spectrum blob
FK_rkopticalbkgs_of_user
rk_std_sample
id int(10)
info text
created_by int(10) (FK)
created_at datetime
spectrum blob
FK_rkstdsample_of_user

Free available Raman Spectra Databases

RRUFF Project (http://rruff.info) - more than 3000 different samples (more than 2000 minerals)
+ IR and X-Ray diffraction spectra. CrystalSleuth software for offline search

RAMIN – Raman Spectra Database of Minerals and Inorganic Materials is available at http://riodb.ibase.aist.go.jp/rasmin/E_index.htm and has 576 Raman spectra of minerals and 1022 of other inorganics as end of 2011.

Mineral Raman Database at University of Parma, Italy
http://www.fis.unipr.it/phevix/ramandb.php ~200 Raman spectra of minerals.

Some others smaller free accessible databases:
ColoRaman (http://oldweb.ct.infn.it/~archeo) pigments,
Handbook of Minerals Raman Spectra (http://www.ens-lyon.fr/LST/Raman),
Minerals Raman spectra by Department of Earth Science in Siena (http://www.dst.unisi.it/geofluids-lab),
Database of SFMC (Société Française de Minéralogie et Cristallographie) specially dedicated to minerals (http://wwwobs.univ-bpclermont.fr/sfmc/ramandb2/index.html).

Commercially available Raman Spectra Databases

Raman data libraries can be also purchased at instrument manufacturer (Horiba, Witec, Renishaw, Thermo Scientific, etc.) and are delivered also with most of the portable/hand-held Raman devices.

Other commercially available databases:

Fiveash Data Management Inc. (http://www.fdmspectra.com)
600 Raman spectra of organics, 250 of pharmaceuticals
6051 minerals spectra

Thermo Fisher Scientific Inc. (http://ramansearch.com)
over 16000 Raman spectra

Bio Rad (http://www.bio-rad.com)
4465 Raman spectra of polymers, inorganics and processing chemicals

Sigma-Aldrich (http://www.sigmaaldrich.com)
14033 FT-Raman spectra

S. T. Japan Europe GmbH (http://www.stjapan-europe.de)
8.694 searchable Raman spectra of polymers, food additives, food packaging, solvents, biochemicals, hydrocarbons, pesticides, dyes, pigments, pharmaceuticals, minerals and inorganics, etc.

why so little compared with other techniques (IR, XRD) ?

Overview

- Data Acquisition and storage
  - Acquisition
  - Storage & organization (spectral libraries)
- Pre-processing of Raman spectra
  - Spike removal
  - Smoothing/denoising
  - Background
  - Normalization
  - Spectral and intensity (re-)calibration
- Chemometry methods for Raman spectroscopy
  - Multivariate calibration algorithms (MLR, PCA, PCR)
  - Classification methods
  - Cluster analysis

Pre-processing of Raman spectra

Why?

•Scope: reduce or eliminate irrelevant, random and systematic variations in the data

e.g. - signal intensity variations linked to laser intensity
- low SNR
- high background (fluorescence)
- spikes

Spike removal

Spikes:

- single events (mostly caused by cosmic rays) hitting the detector
- are always positive peaks of narrow bandwidth
- random position on CCD and random in time

Spike removal:

- multiple accumulations (eliminates the spike through comparison, robust summation etc.)
- mathematical
  - missing point polynomial filter
  - robust smoothing filter
  - moving window filter
  - wavelet transform methods
- reduction of spike events by special design of the instrument (Zhao, 2003)

Assumptions:

- spike band width much smaller than Raman band width
- Raman maps de-spiking: spatially adjacent spectra are similar

Smoothing

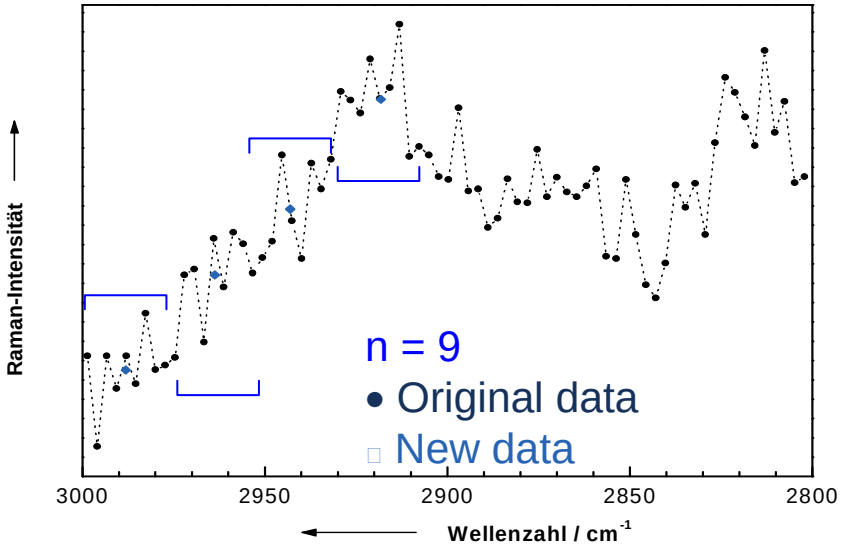
- Scope: Reduce the noise→ Improvement of SNR
- Noise is random and it changes with a higher frequency than the Raman signals
- Methods:
  - Average
  - Moving Average
  - Moving Median
  - Moving Polynomial (Savitzgy-Golay)
  - Fourier Filter

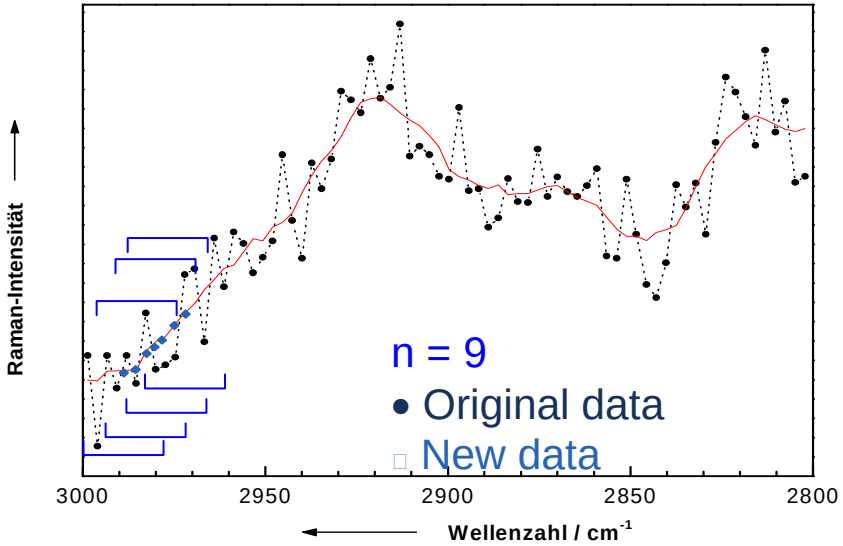
} Works on intervals of **n** points

Smoothing

Average

- reduces the data points
- information is lost

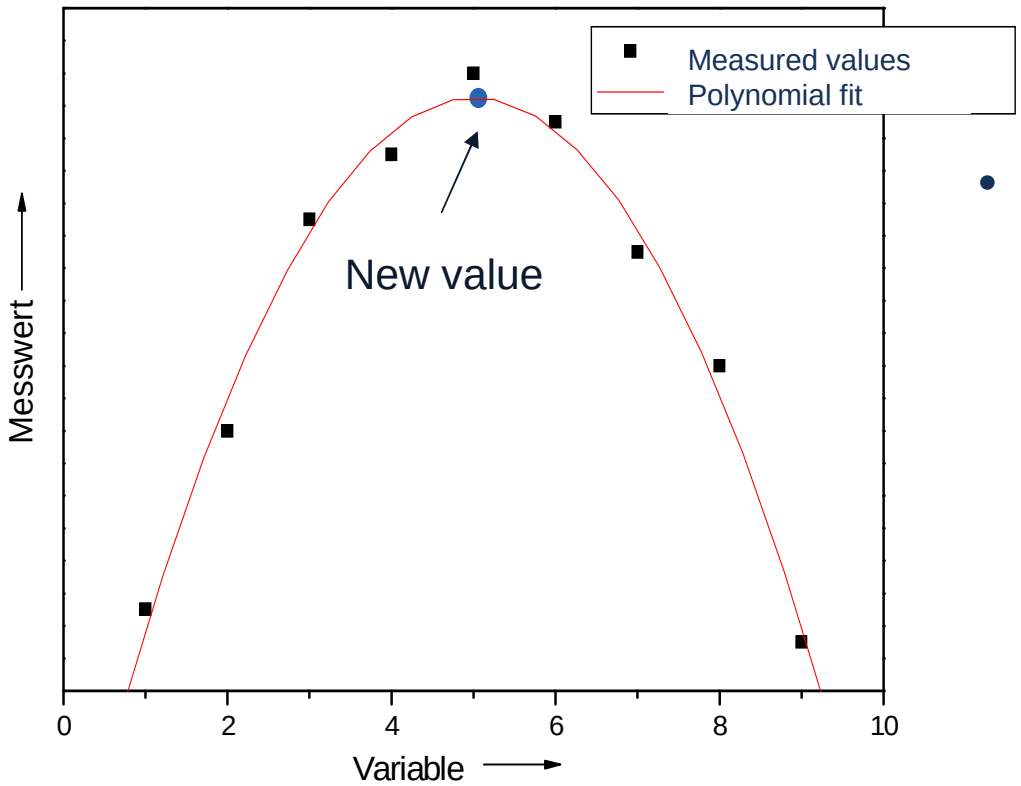




Moving Average

- data points number remains
- information loss increases with increasing moving window size
- ‘end’ effects

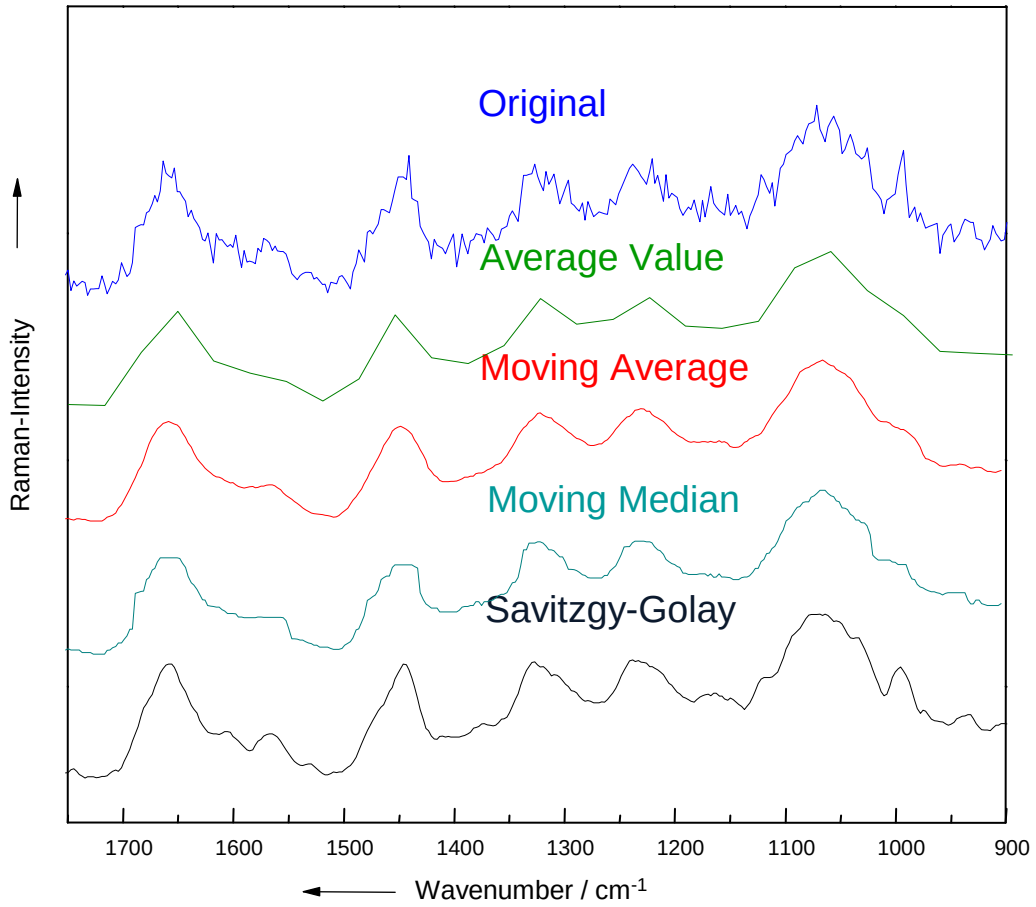
Smoothing



- Moving Polynomial (Savitzgy-Golay):
  - Polynomial of small order is fitted through data points

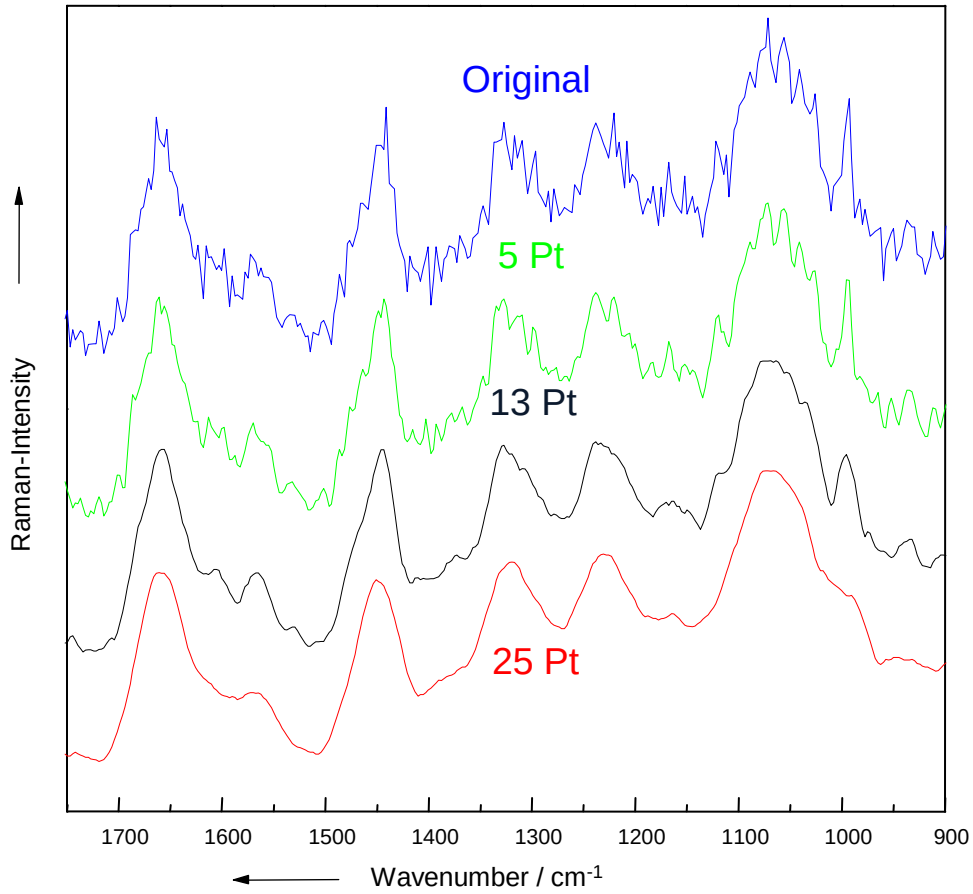
Smoothing

Window of 13 points used with different smoothing algorithms



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

Smoothing



Different smoothing window sizes for Savitzgy-Golay (Polynomial)

Width of Raman bands increases with smoothing window size

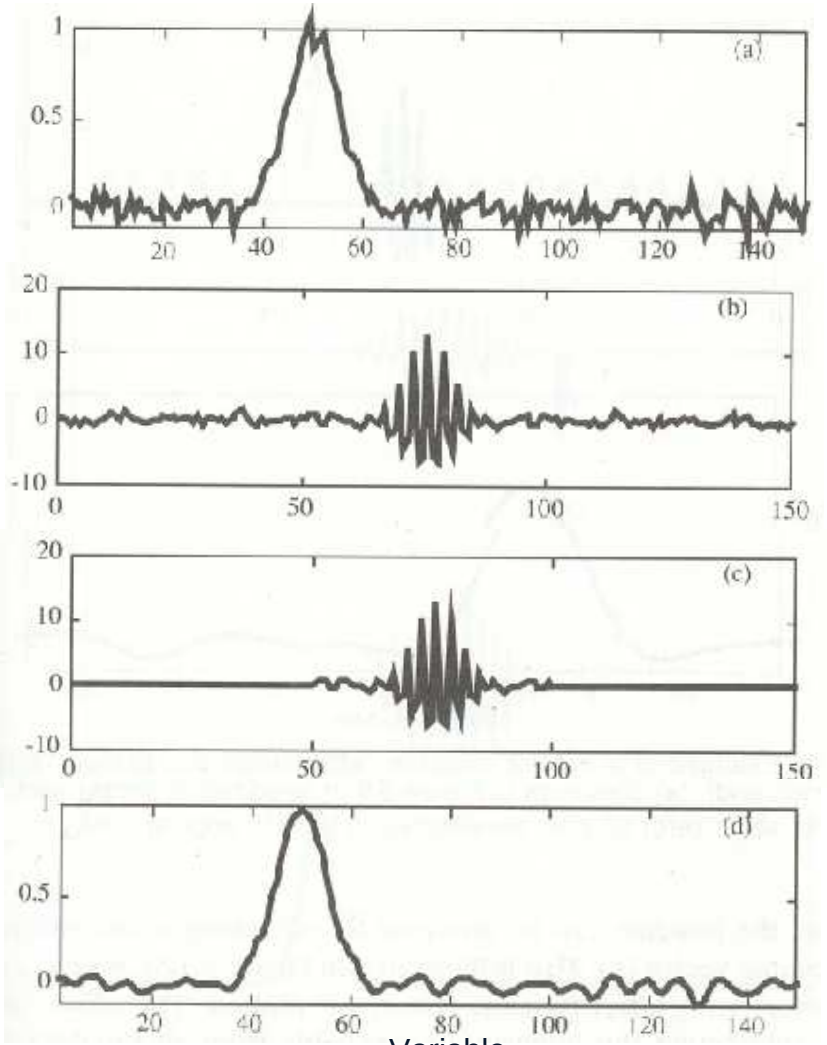
Smoothing

Spectrum is Fourier-transformed

Transformed spectrum is multiplied with a filtering function (reduce or eliminate the influence of high frequency terms in FT, e.g. Boxcar

Another FT (back transformation)

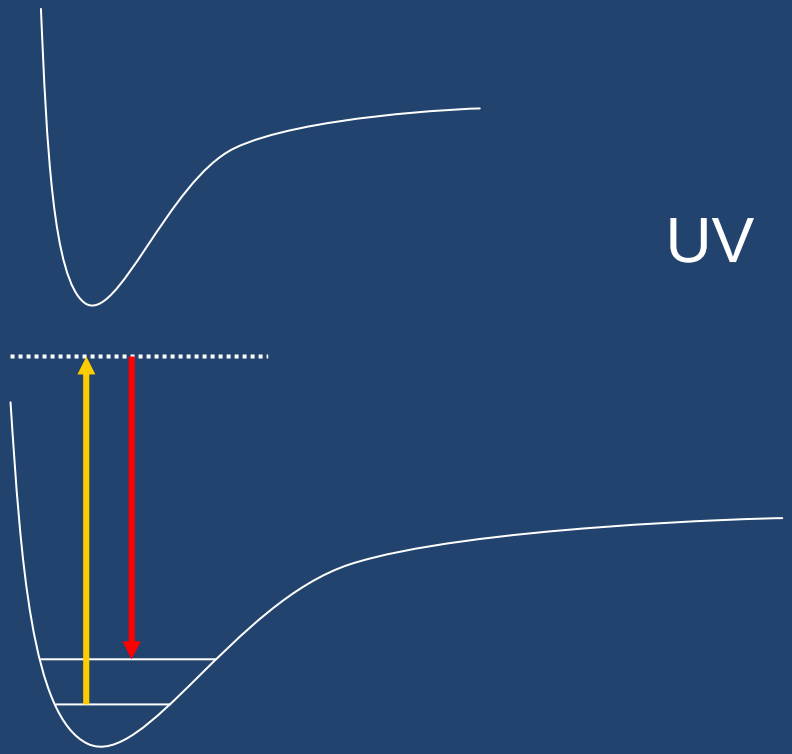
Eliminating too many frequencies will lead to
- looses of spectral information
- enlargement of band widths



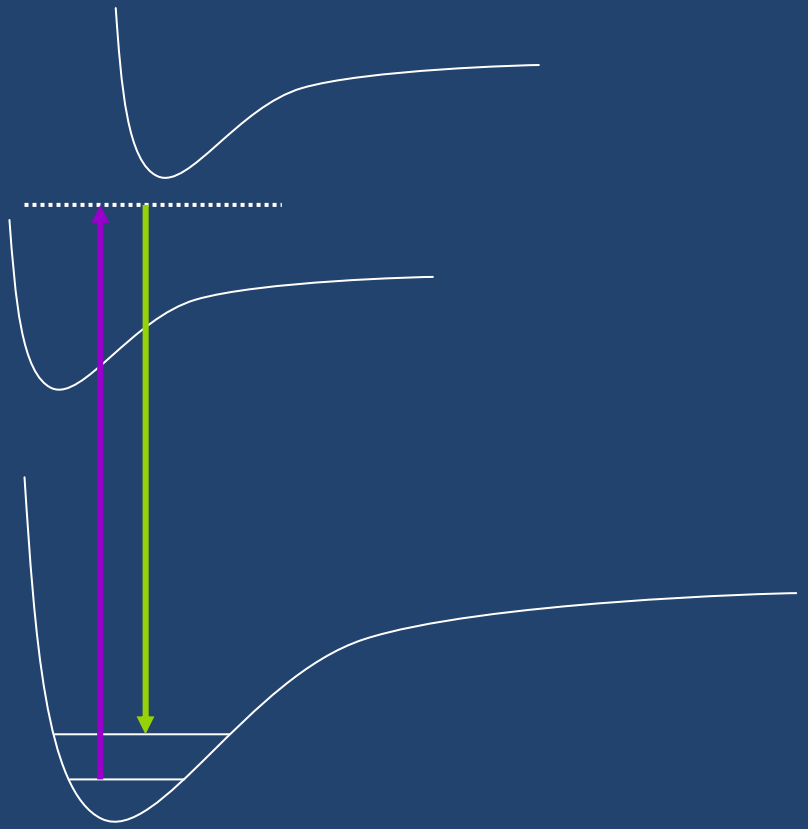
Images: K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Practical Guide* (John Wiley & Sons, New York, 1998).

a) Chose appropriate excitation wavelength: near infrared or UV laser lines

Near IR



UV



b) For low-wavenumber modes look at anti-Stokes side since fluorescence starts in most cases only in the Stokes region

c) Use nonlinear technique: coherent anti-Stokes Raman spectroscopy (CARS)

d) Fluorescence quenching by means of surface enhanced Raman scattering (SERS)

e) Bleaching

f) Rejection of fluorescence by means of Shifted Excitation Raman Difference Spectroscopy (**SERDS**) (equivalent with recording a spectrum and shifting the same spectrum digitally with computer)

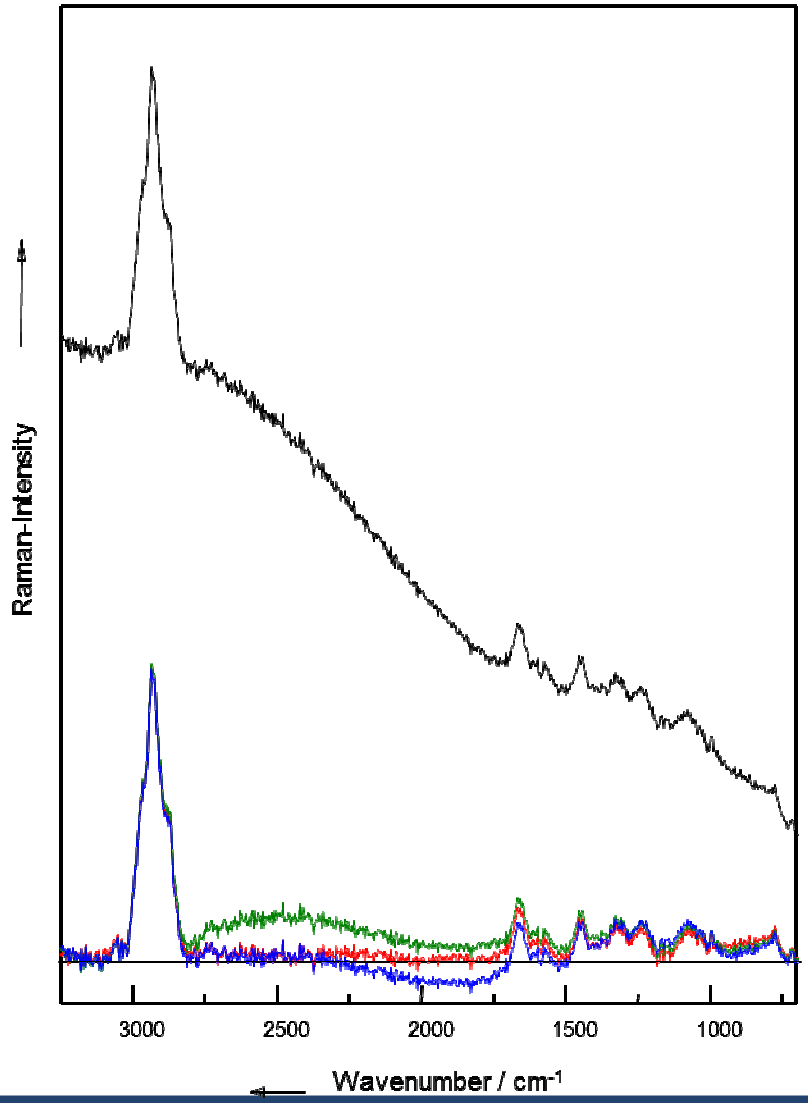
(see e.g. P.A. Mosier-Boss et al., Appl. Spectrosc. **49**, 630 (1995)

Background

- Background sources:
  - Fluorescence and other laser induced emissions (sample+instrument optics)
  - Environmental light sources
  - Instrument specific background (e.g. dark current)

- Methods of background correction:
  - Polynomial $r = \tilde{r} + \alpha + \beta x + \gamma x^2 + \ldots$
    - Offset correction (Subtract a constant, α)
    - Subtract a line $\alpha + \beta x$
    - Approximate a polynomial through basis points $\alpha + \beta x + \gamma x^2 + \ldots$
    - Derivate $r' = \tilde{r}' + 0 + \beta + \gamma x + \ldots$
    - FFT with high pass filter
    - Wavelet transformation techniques
    - Free form baseline

Background – baseline support points



- Correct Baseline
- Wrong baselines:
  - Left-over background
  - Overfitting (negative values)

A baseline should not cut into Raman band signal strength

Background – Derivate

Advantages: Baseline support points are not needed

Spectrum: $r = (r_1, r_2, r_3, r_4, \ldots)$

Methods:

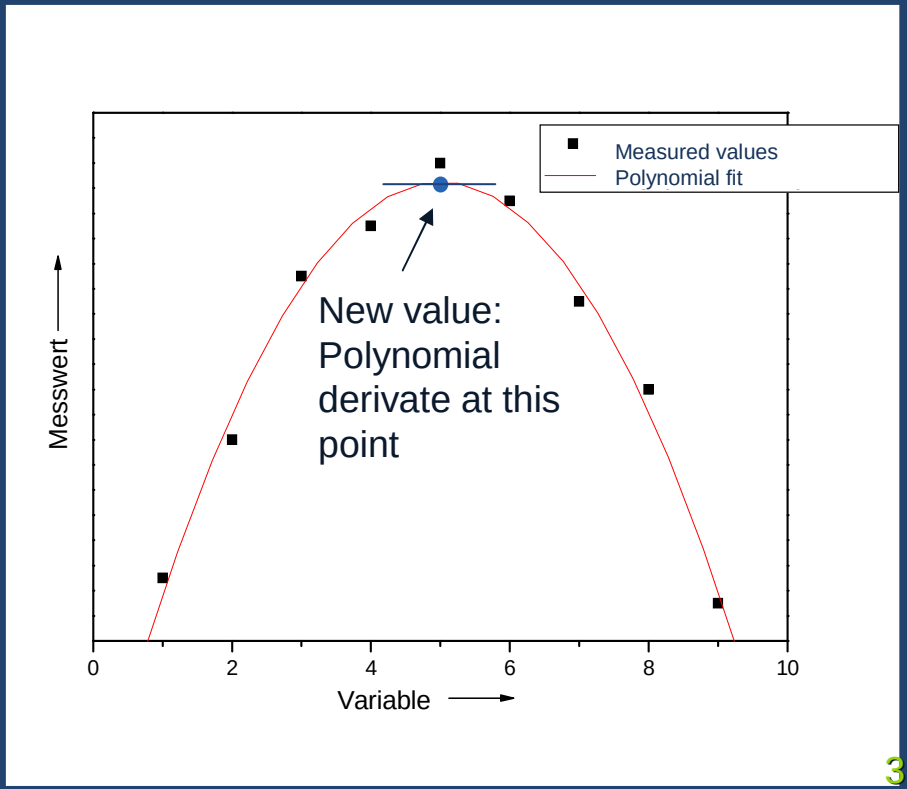
- Moving simple difference:

$$r' = (r_2 - r_1, r_3 - r_2, r_4 - r_3, \ldots)$$

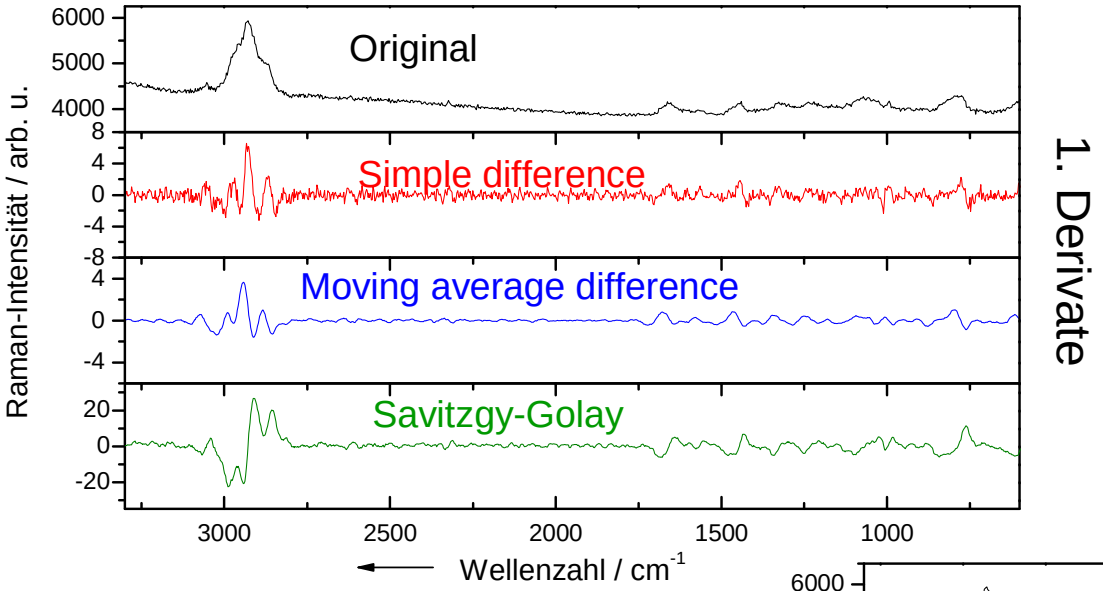
- Moving average difference:
  - window size, average values
  - e.g. window size 3:

$$r' = \left(\frac{r_2 + r_3 + r_4}{3} - \frac{r_1 + r_2 + r_3}{3}, \ldots\right)$$

- Savitzgy-Golay:
  - Polynomial fitted in moving window
  - New value is the polynomial derivate on that point



Background – Derivate



Simple Difference:

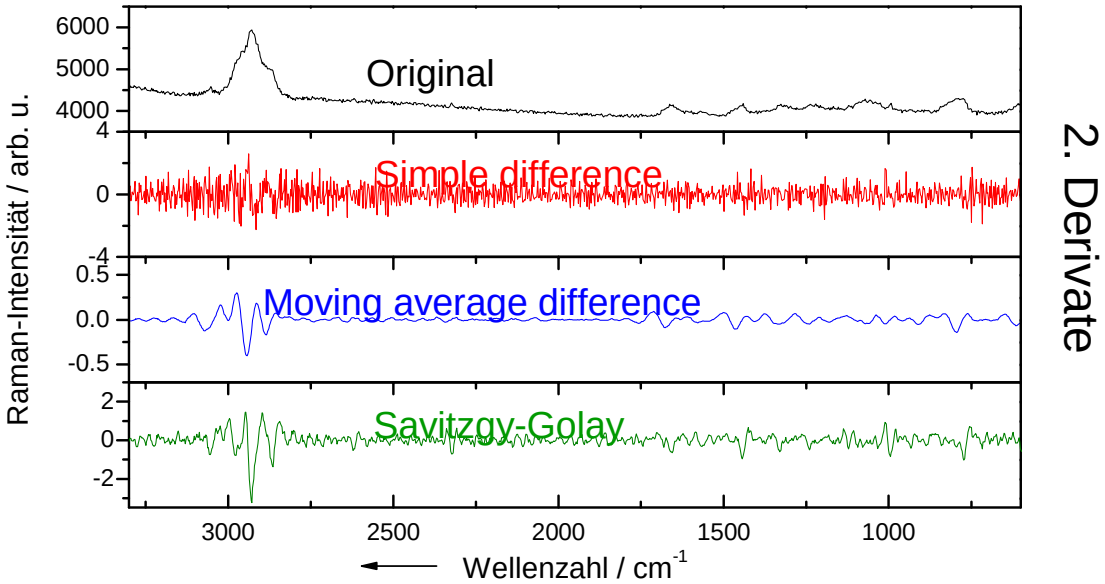
- SNR decrease

Moving average difference:

- smooth
- Bands are shifted

Savitzgy-Golay:

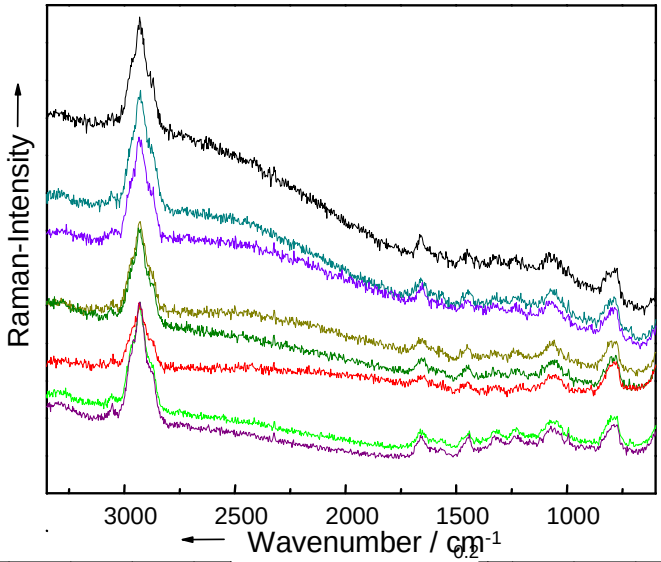
- smooth
- Band shape better preserved



Normalization

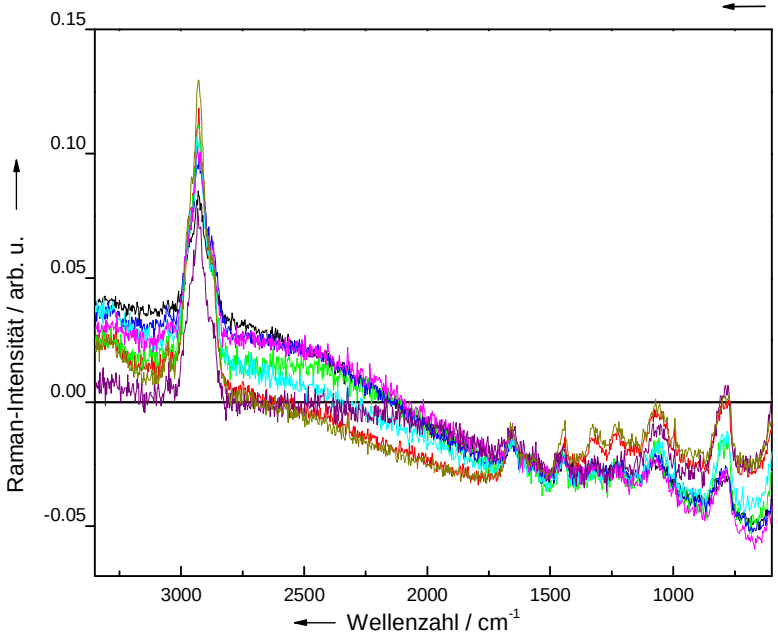
- Eliminate systematic differences among measurements
  - Raman spectra show intensity differences because of e.g.:
    - Changing laser power
    - Differences in focusing depth
    - Sample volume differences
- For measuring concentrations with Raman based on calibration curves do not normalize your spectra on bands dependent on concentration!
- 2 Possible normalization methods: Vector normalization and Min-/Max-Normalization
- Perform baseline correction before normalization

Normalization

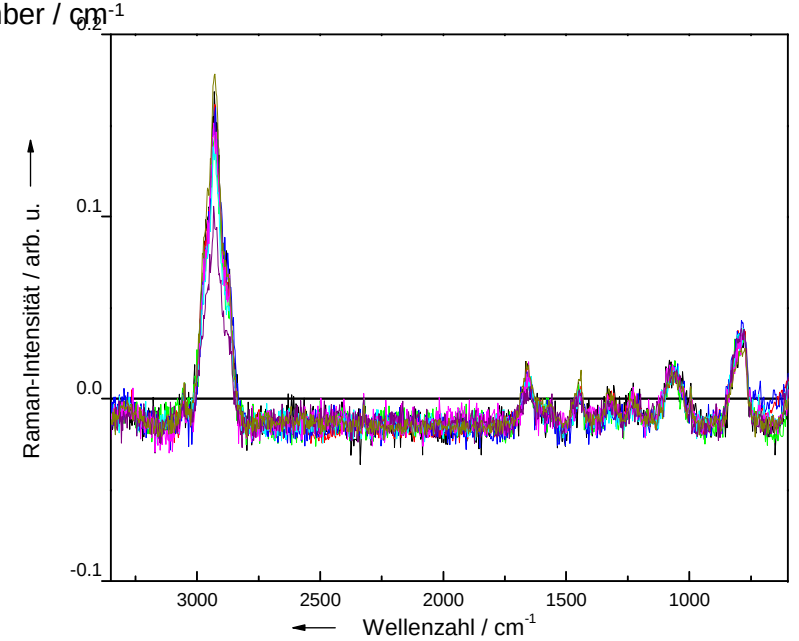


Raw data

Nomalization without baseline correction

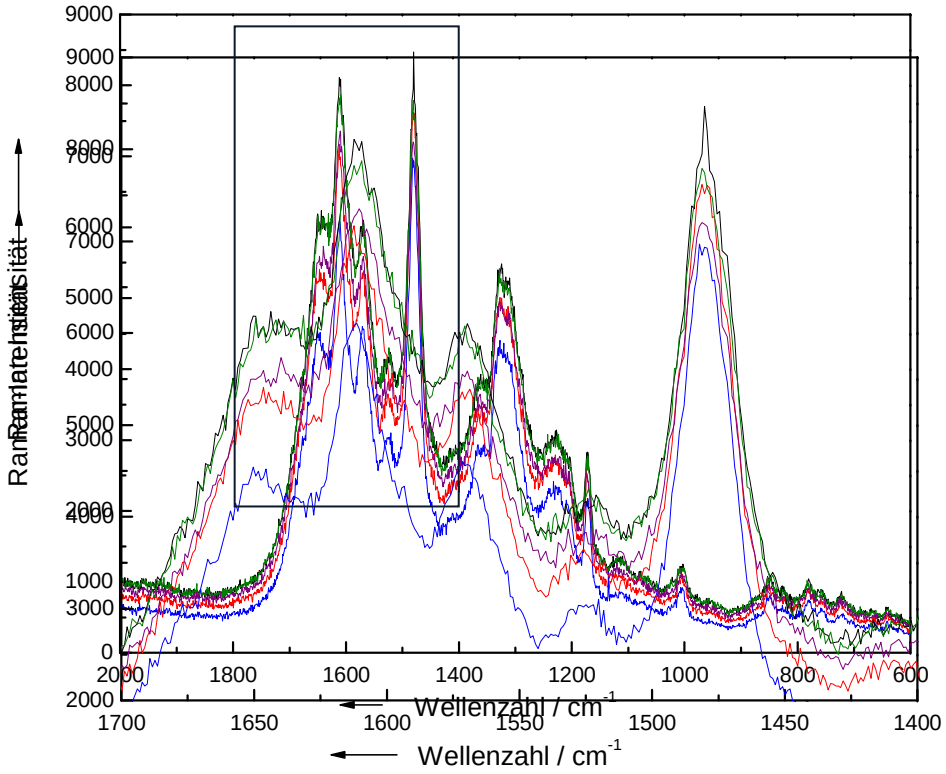


Nomalization with baseline correction

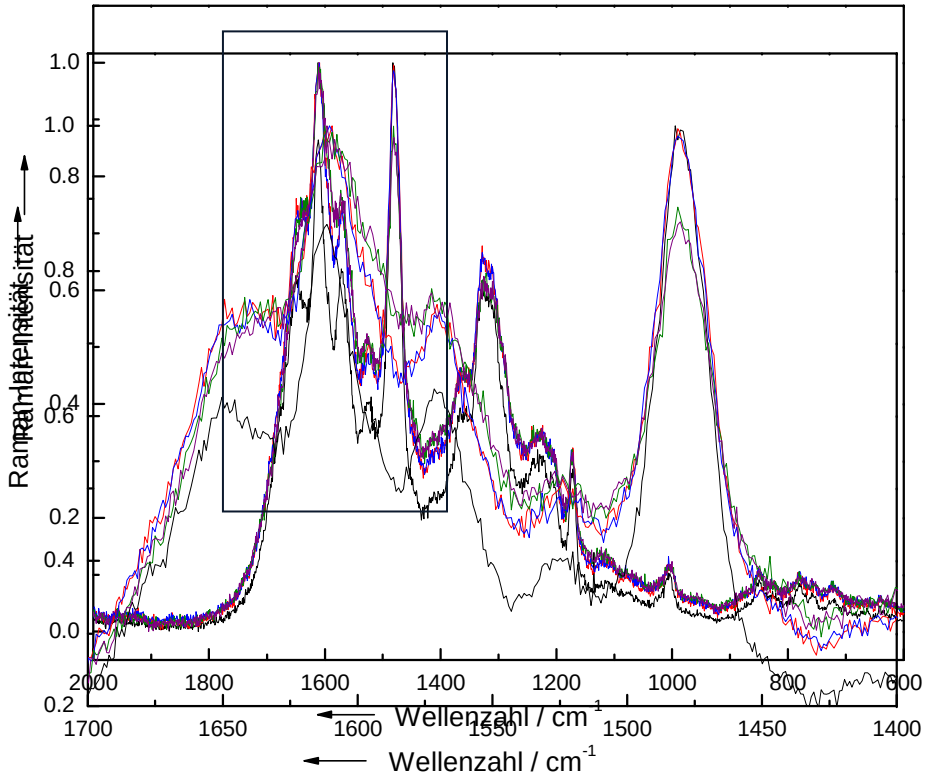


Normalization – Min/Max method

Original data



Normalized data



$$I_{norm} = \frac{I - I_{\min}}{I_{\max}}$$

Vector Normalization

- Calculates the average intensity value for all (chosen) wavenumber:

$$a_m = \frac{\sum_k a(k)}{k}$$

- Intensities will be centered:

$$a'(k) = a(k) - a_m$$

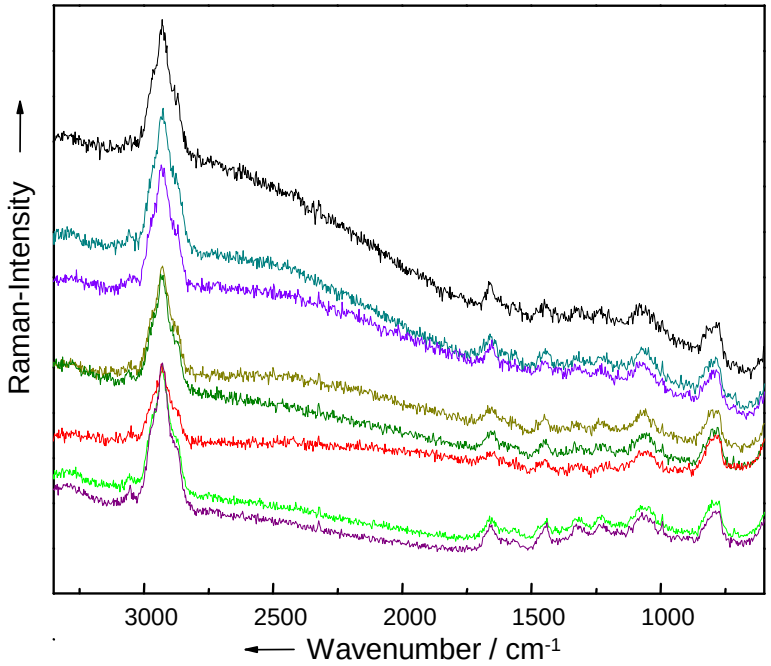
- Centered intensities will be divided by the length of the spectrum (as vector)

$$a''(k) = \frac{a'(k)}{\sqrt{\sum_k (a'(k))^2}}$$

- The new spectrum vector will have a length of 1.

$$\sum_k \left(a''(k)\right)^2 = 1$$

Vector Normalization



Baseline correction

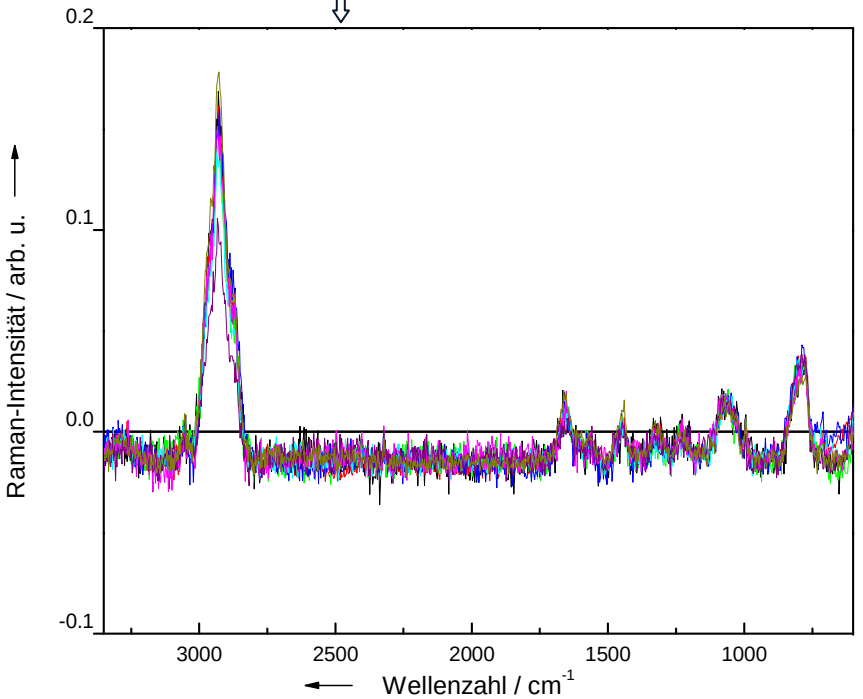
⇩

Centering

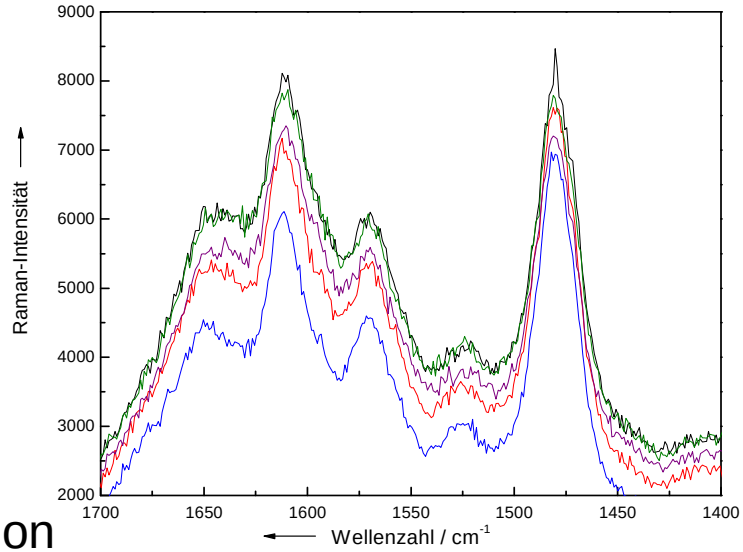
⇩

Vector normalization

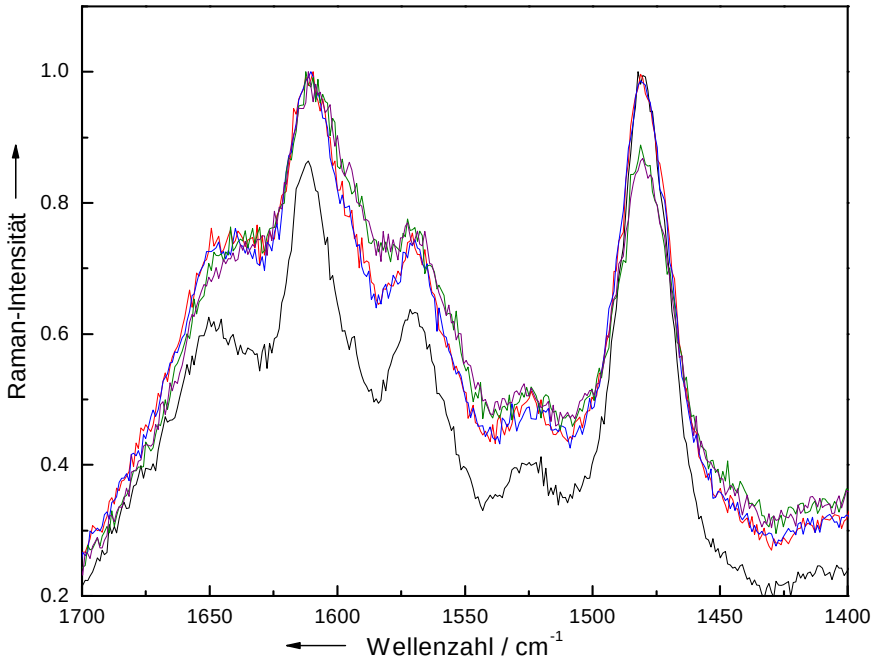
⇩



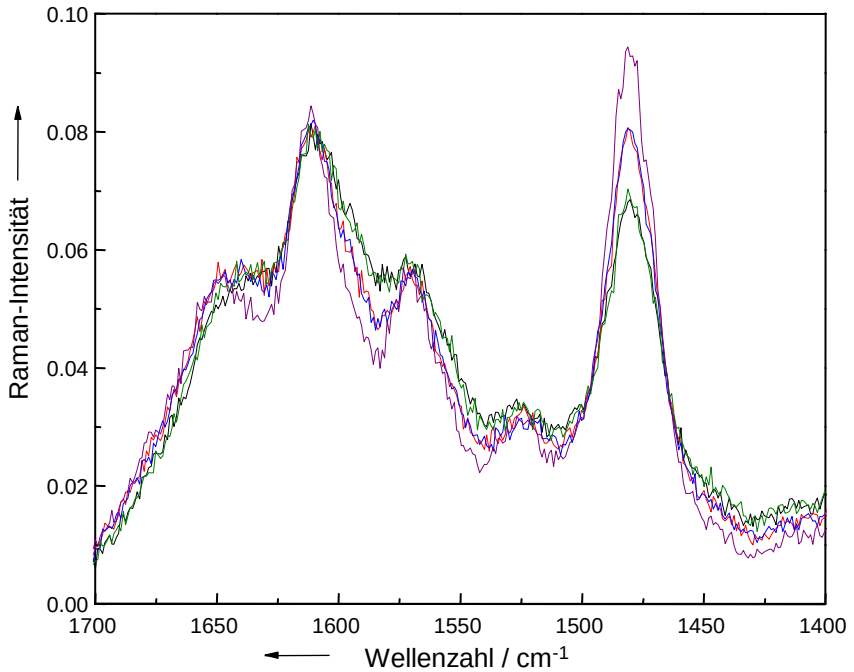
Original data



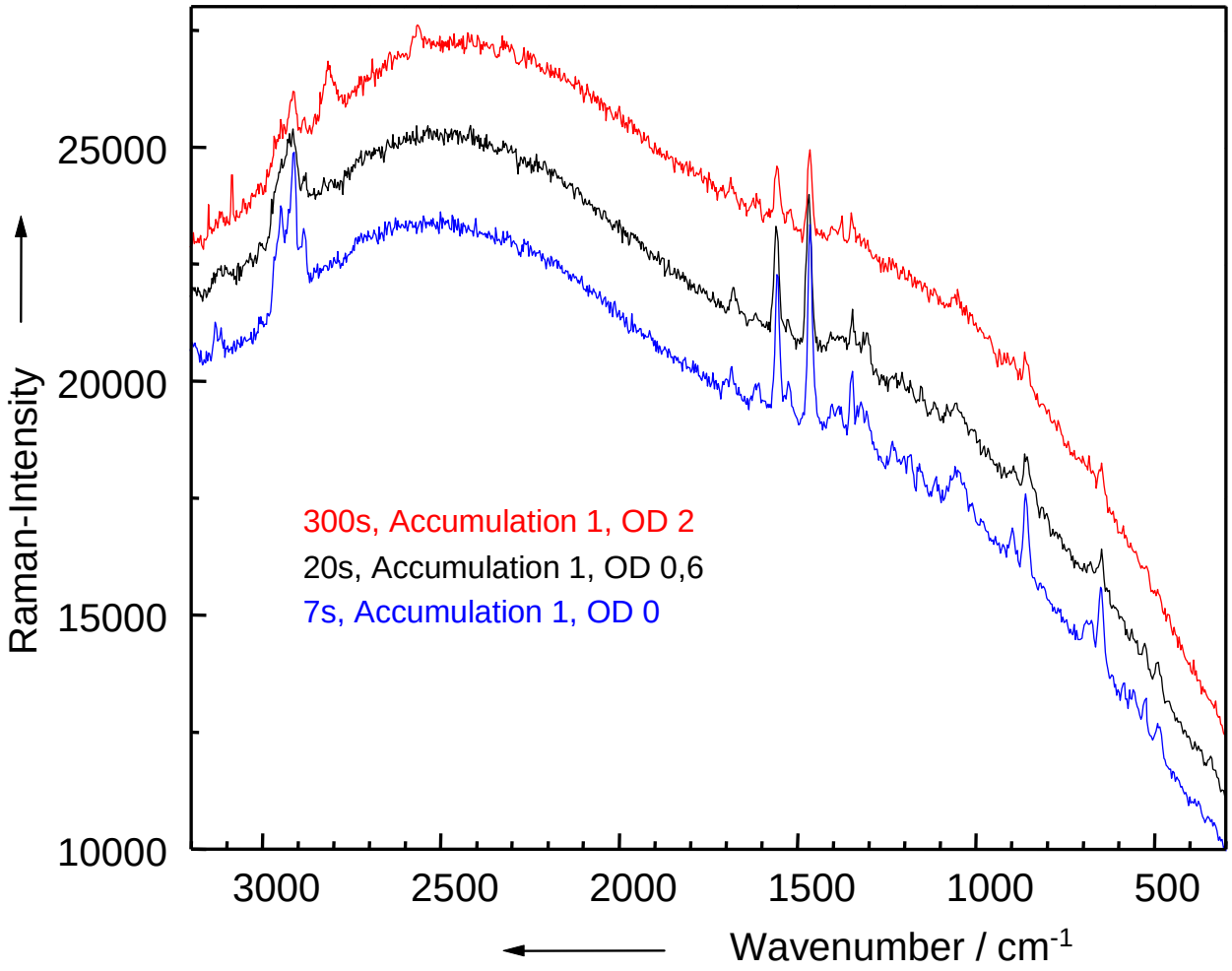
Min-/Max-Normalization



Vector normalization

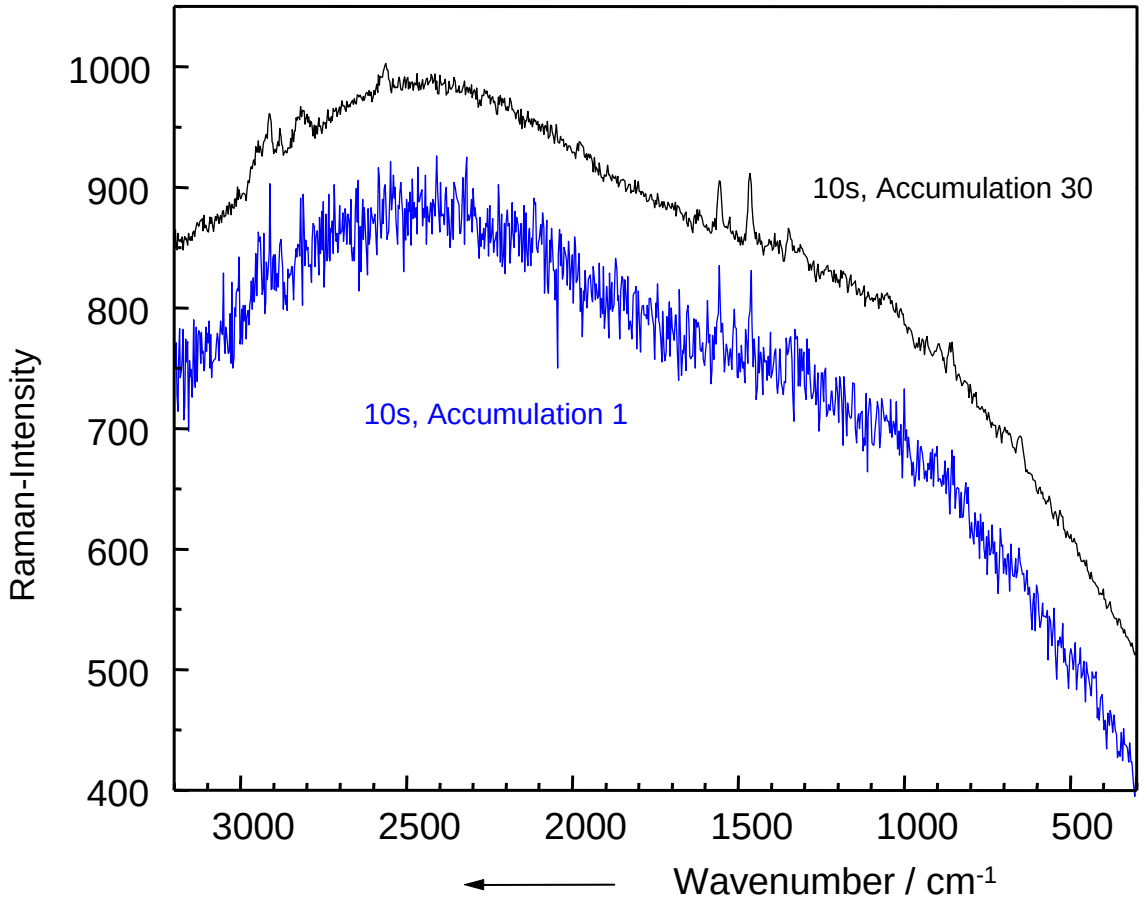


Higher laser power leads to better SNR as longer integration time.
But: photo degradation of the sample.



$$I_R \propto I_{Laser}$$

- Improvement of SNR through multiple accumulations



Spectral and Intensity (re-)calibration

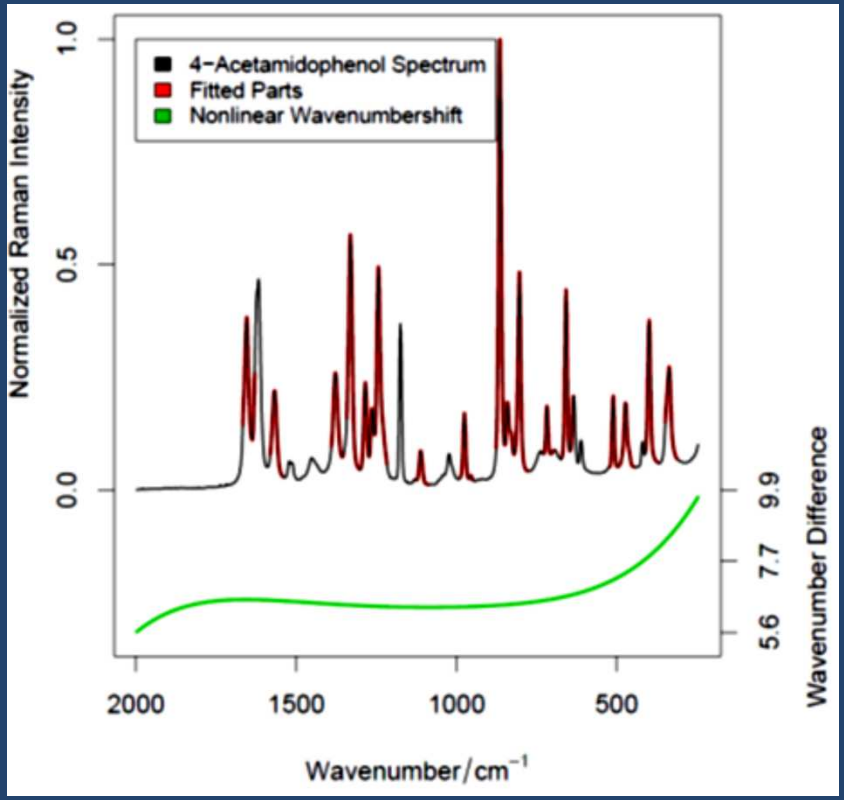
Spectral Calibration

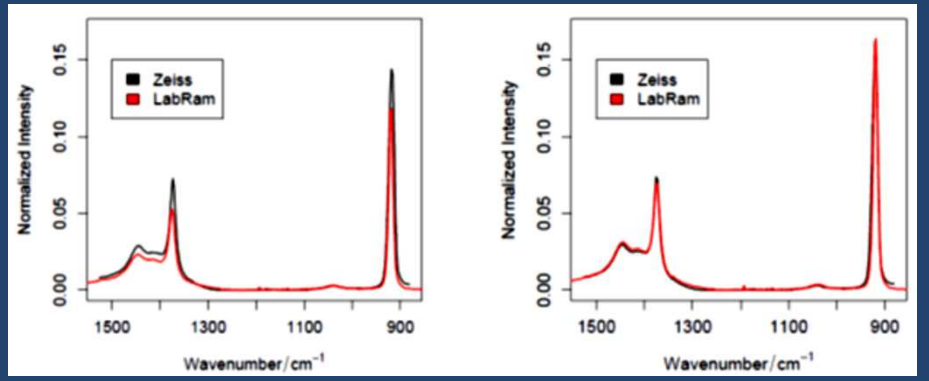
- Atomic emission lines (Ne, Hg, Ar, Kr lamps, plasma lines of the gas lasers)
- Shift calibration standards (indene, polystiren, paracetamol, naphtalene etc.)

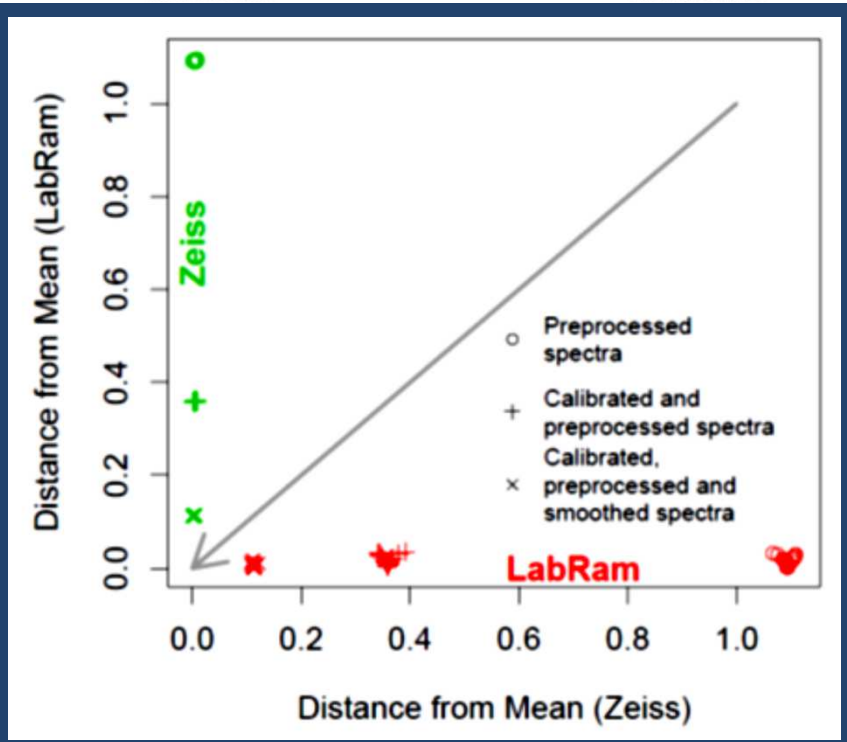
Intensity Calibration

- Intensity calibrated lamps
- NIST fluorescence standards for Raman

Spectral and Intensity (re-)calibration







Instrument variation induced in Raman data can be reduced but very hard to completely eliminate

=> Spectrometer identification by analyzing the Raman spectra ☺

Overview

- Data Acquisition and storage
  - Acquisition
  - Storage & organization (spectral libraries)
- Pre-processing of Raman spectra
  - Spike removal
  - Smoothing/denoising
  - Background
  - Normalization
  - Spectral and intensity (re-)calibration
- Chemometry methods for Raman spectroscopy
  - Multivariate calibration algorithms (MLR, PCA, PCR)
  - Classification methods
  - Cluster analysis

Chemometrics

K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Practical Guide* (John Wiley & Sons, New York, 1998).

H. Martens and T. Naes, *Multivariate Calibration* (John Wiley & Sons, Chichester, 1993).

R. Henrion and G. Henrion, *Multivariate Datenanalyse* (Springer Verlag, Berlin, 1994).

J. Gasteiger and T. Engel, (Eds.) *Chemoinformatics*, (Wiley-VCH, Weinheim, 2003).

1972 Svante Wold und Bruce R. Kowalski establish the new term „Chemometry“

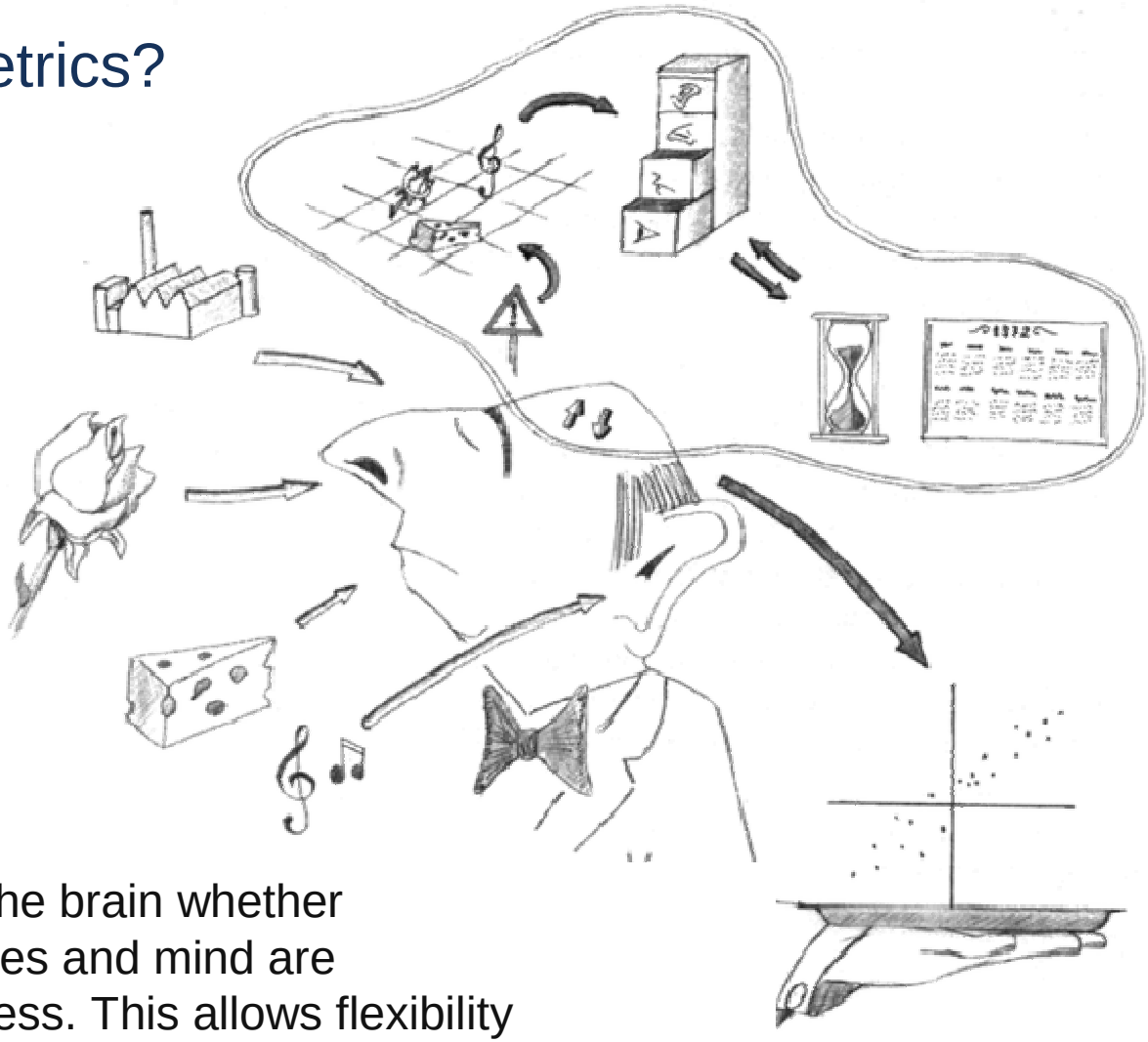
1974 International Chemometrics Society

Definition:

“the chemical discipline that uses mathematical and statistical methods
(a)to design or select optimal measurement procedures and experiments, and
(b)to provide maximum chemical information by analyzing chemical data.”

(Mathias Otto 1999 -'Chemometrics Statistics and Computer Application in Analytical Chemistry.‘)

Why use chemometrics?

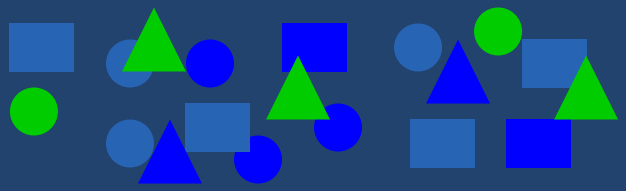


"It all comes together in the brain whether we like it or not, our senses and mind are part of the research process. This allows flexibility and creativity, but calls for checks against wishful thinking."

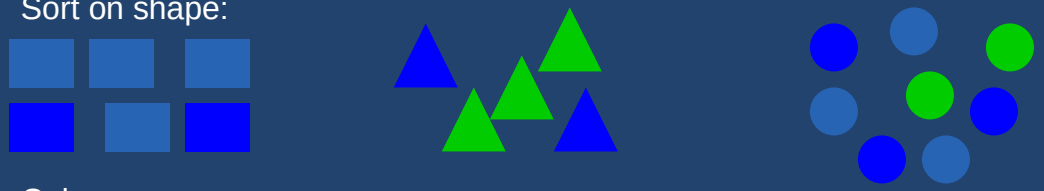
H. Martens, M. Martens, 2001

Chemometrics

- The task of sorting objects migt seem sometimes trivial



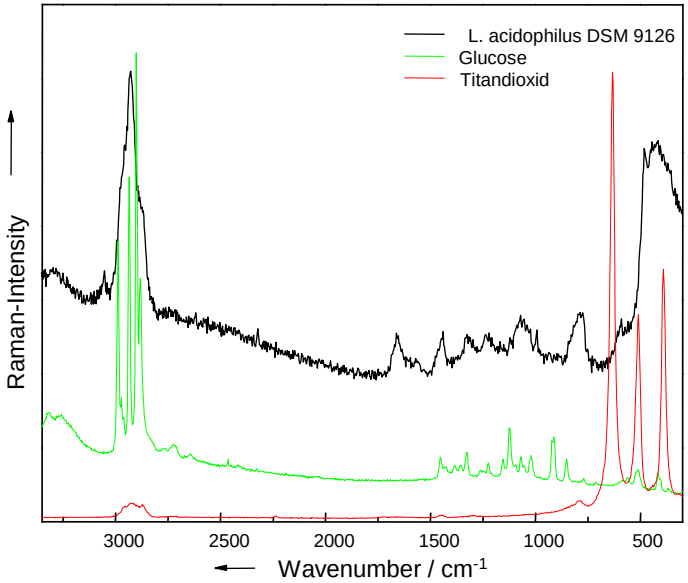
- Sort on shape:



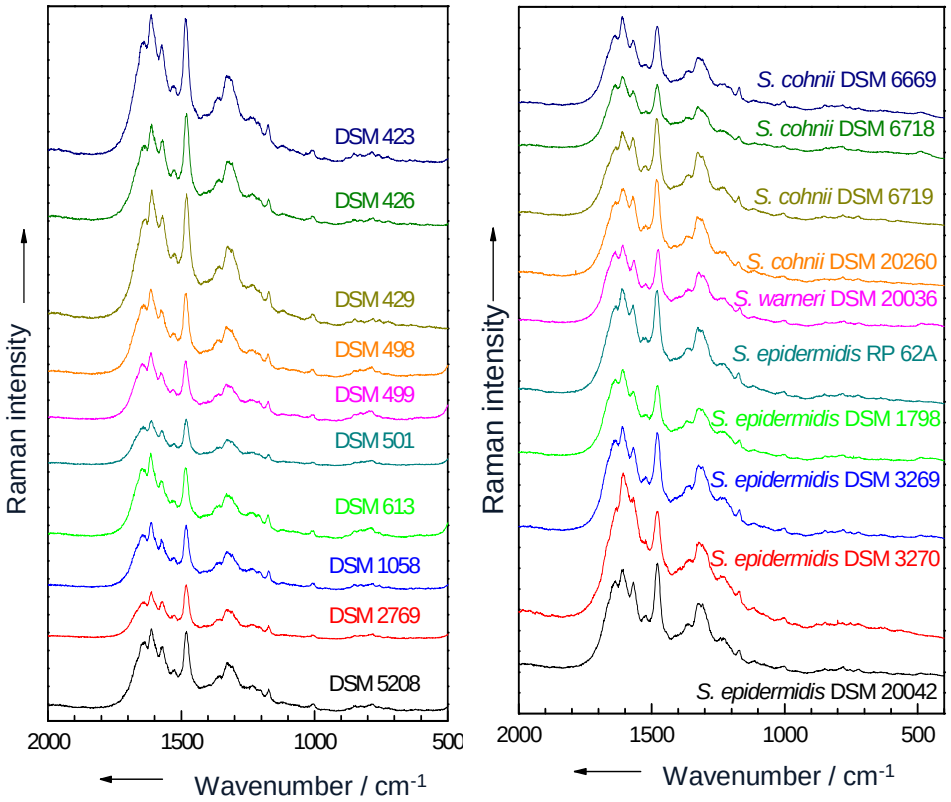
- Color:

- Shape and color:

Chemometrics



For very different spectra the task is easy



For larger data set and more resembling data ?

Chemometric methods

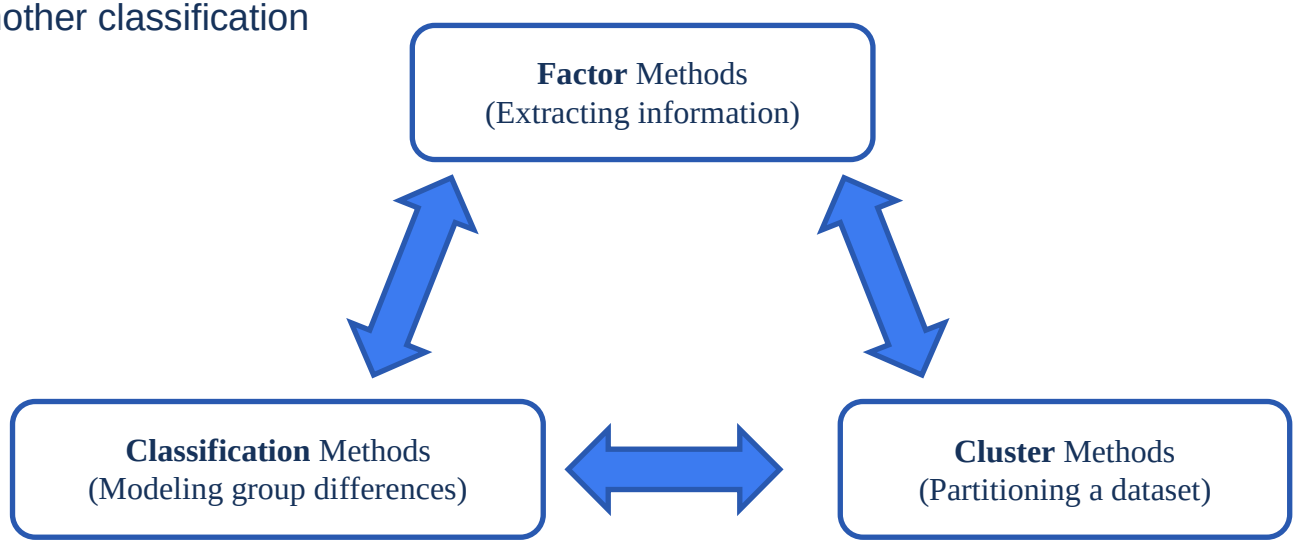
Supervised methods (require a set of well characterized samples – calibration model)

- Multiple Linear Regression (MLR)
- Partial Least Square Regression (PLS)
- Linear Discriminant Analysis (LDA)

Unsupervised methods (do not require *a priori* knowledge on the samples)

- Principal Component Analysis(PCA)
- Cluster Analysis (CA)
- Factor Analysis

Another classification



Chemometric methods

Multivariate calibration Algorithms - needs calibration model (validation)

- Multiple Linear Regression (MLR) – simple regression extended to multiple var. (deconvolutions)
- Principal Component Analysis(PCA) – unsupervized (reduce dimensionality to PC)
- Principal Component Regression (PCR) – linear regression using PCA
- Partial Least Squares Regression (PLS) – like PCR with other projection rules
- …

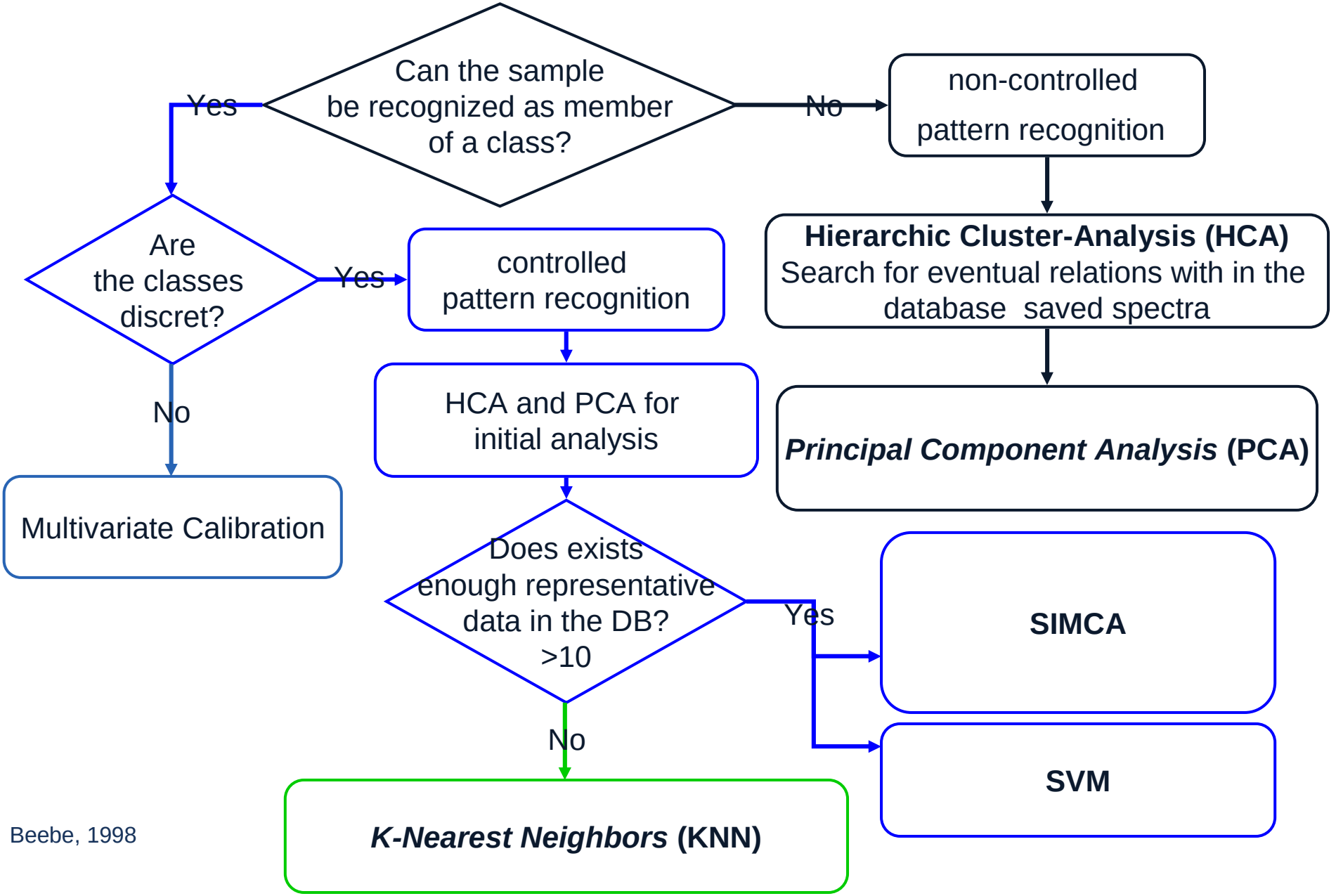
Classification Methods

- Linear Discriminant Analysis (LDA) – reduce dimensionality by preserving class discrimination
- Gaussian Mixture Modeling (GMM) – weighted sums of gaussian distributions
- Artificial Neuronal Network (ANN) – layered processing elements (neurons)
- Support Vector Machines (SVM) – try to find the borders between two classes
- …

Cluster Analysis –widely used for Raman image analyses (hyperspectral data in general)

- Hierarchical Cluster Analysis (HCA) – cluster based on distances among spectra
- Fuzzy –c – means Cluster Analysis (c-means CA) – like HCA & k but with membership function
- Fuzzy –k – means Cluster Analysis (k-means CA) –min D inside clusters and max between them
- Gaussian Mixture Modeling (GMM)
- …

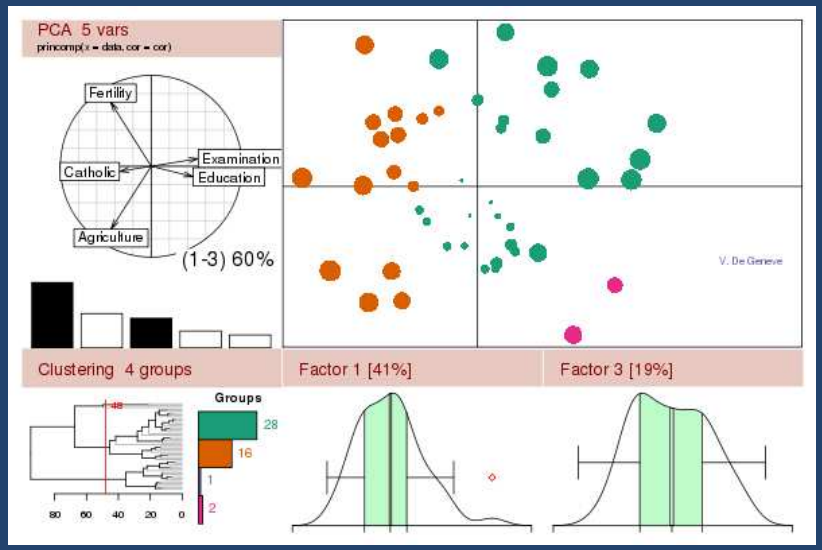
+ nonlinear versions of each (SVM with kernel-trick …)


Can the sample be recognized as member of a class?
Yes
No
non-controlled pattern recognition
Hierarchic Cluster-Analysis (HCA)
Search for eventual relations with in the database saved spectra
Principal Component Analysis (PCA)
Are the classes discret?
Yes
No
controlled pattern recognition
HCA and PCA for initial analysis
Multivariate Calibration
Does exists enough representative data in the DB? >10
Yes
No
SIMCA
SVM
K-Nearest Neighbors (KNN)

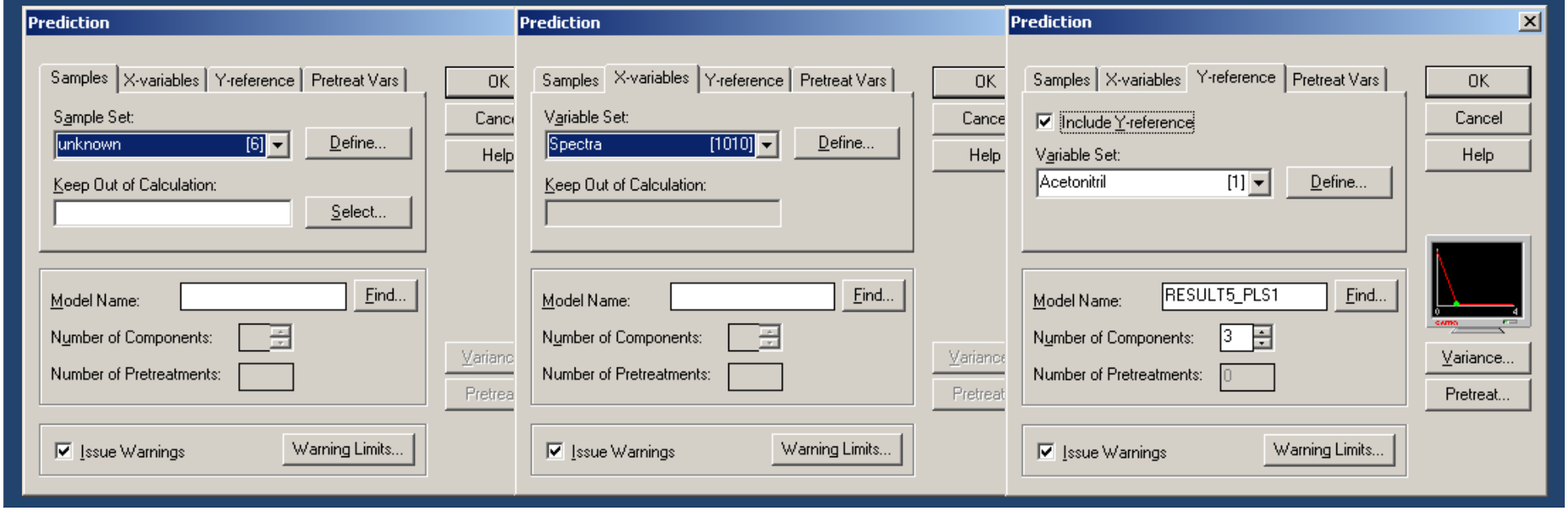
Beebe, 1998

Chemometric tools

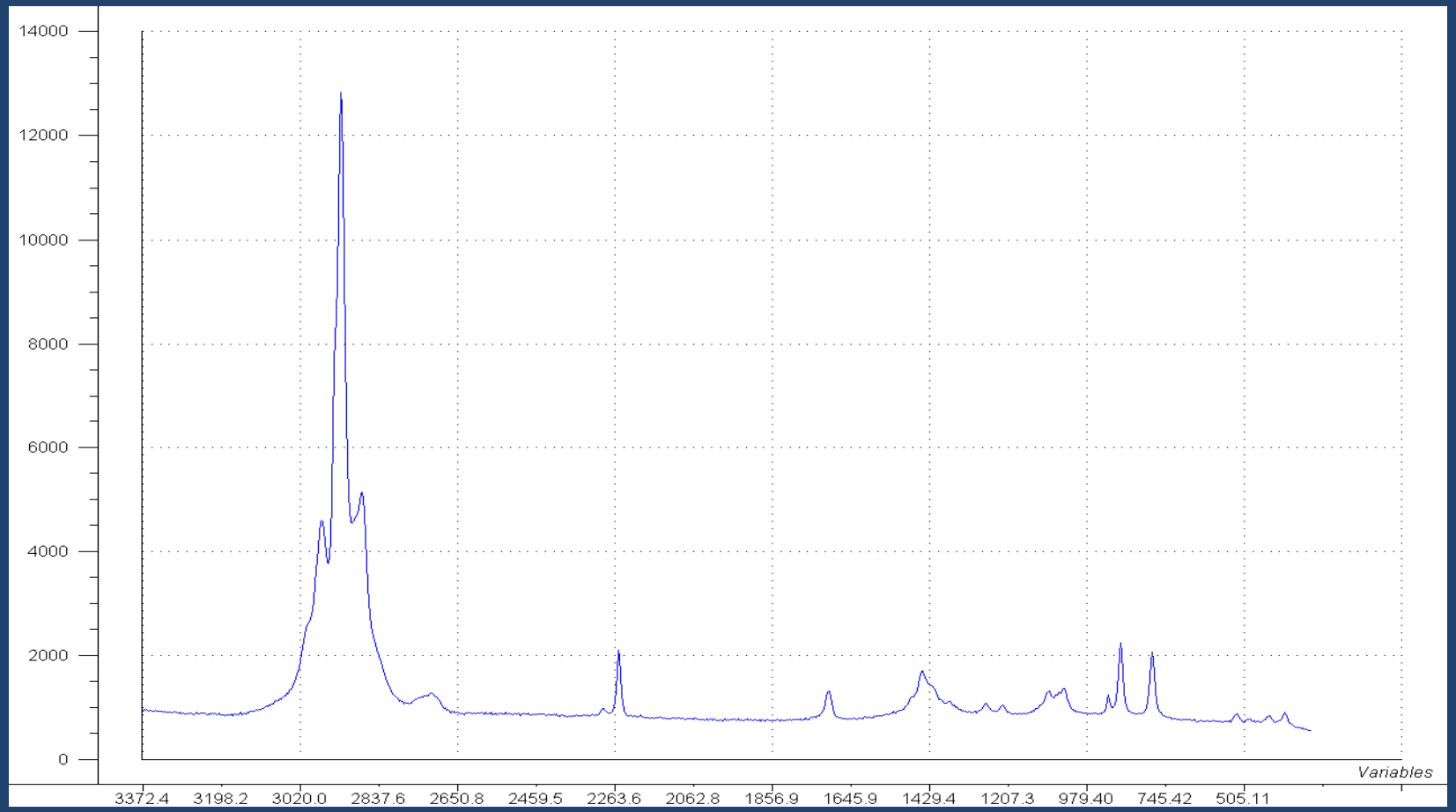
R (free open source statistics software)
http://www.r-project.org/



Unscrambler (PCA, PLS, MLR)
http://www.camo.com/

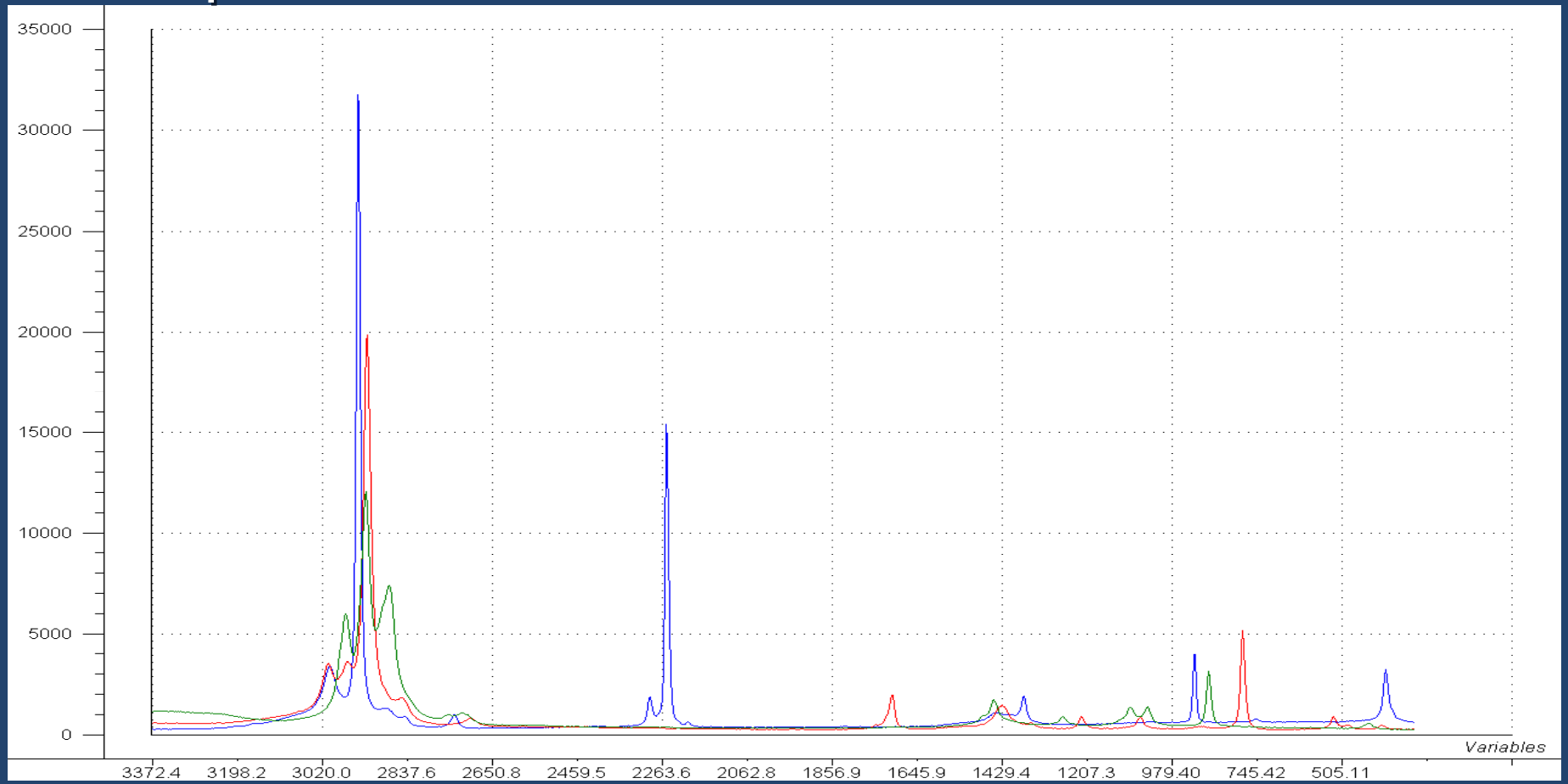


Example PCA



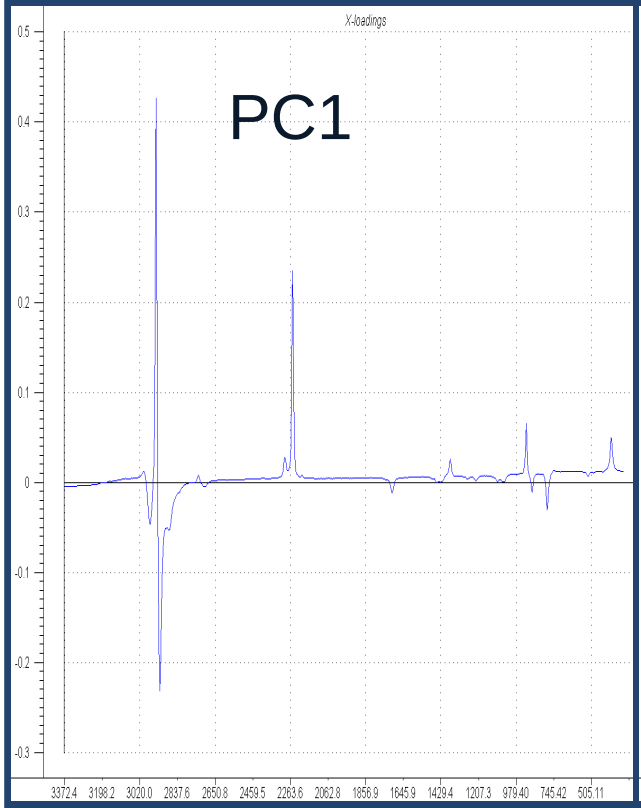
Raman spectra (3372 bis 298 cm^{-1})
How many components and which concentrations

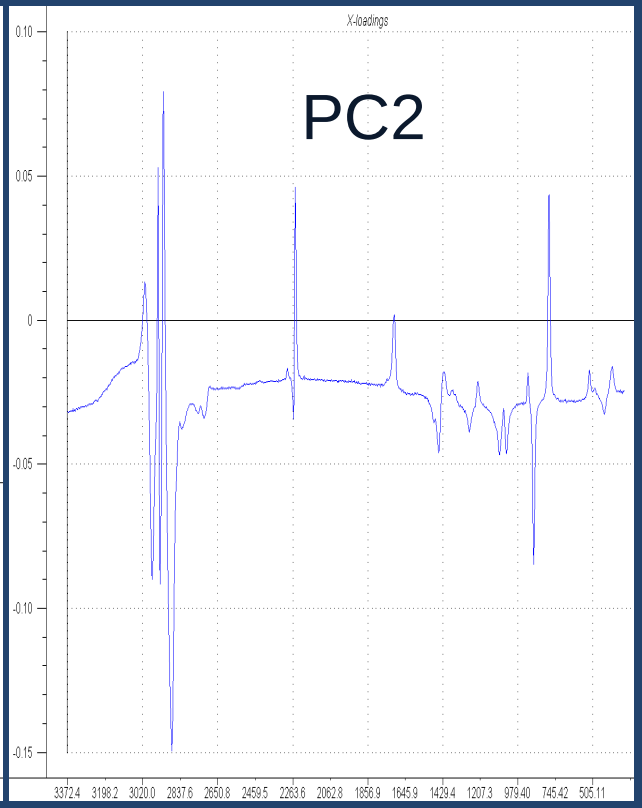
Example PCA

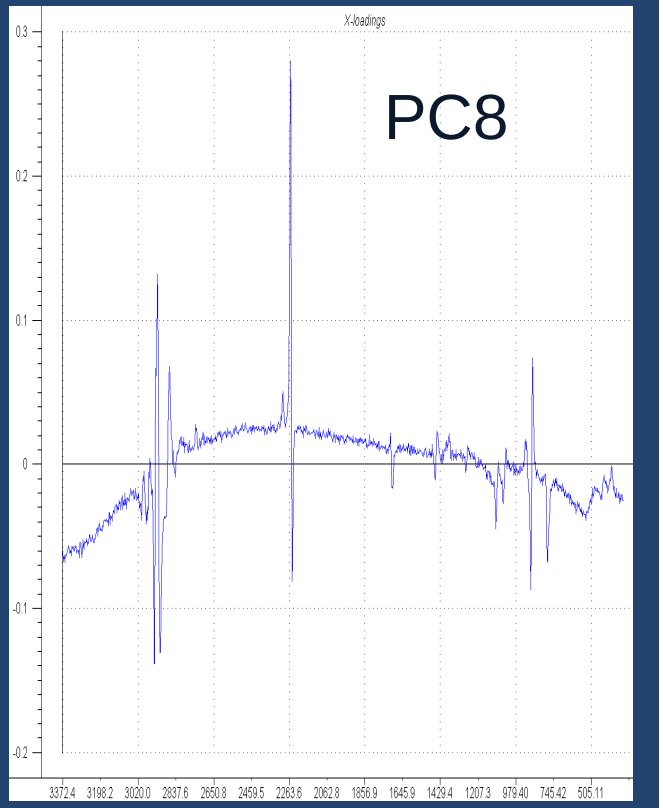


- A qualitative inspection, shows 3 components in the spectrum
  - Blue: Acetonitrile
  - Red: Acetone
  - Green: Ethanol

Initial tries - PCA

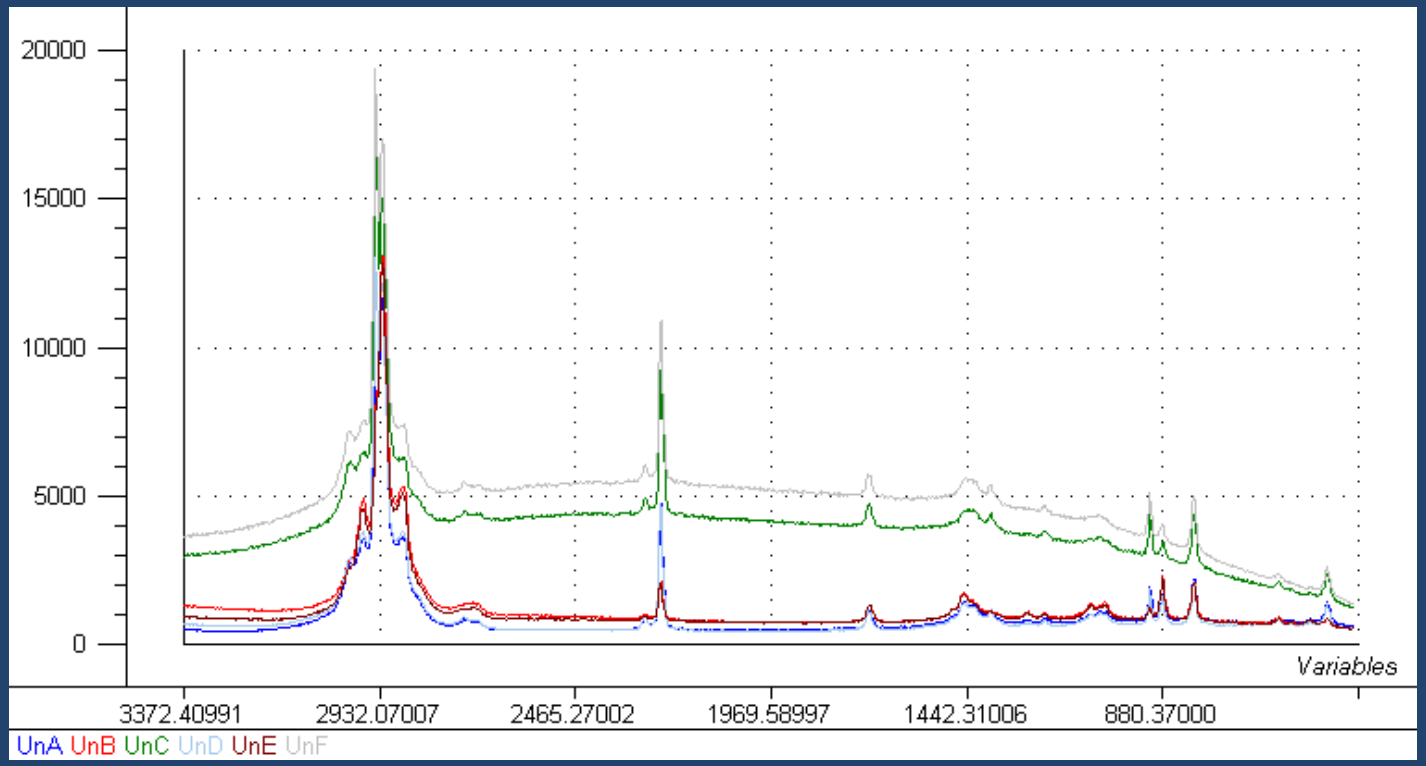




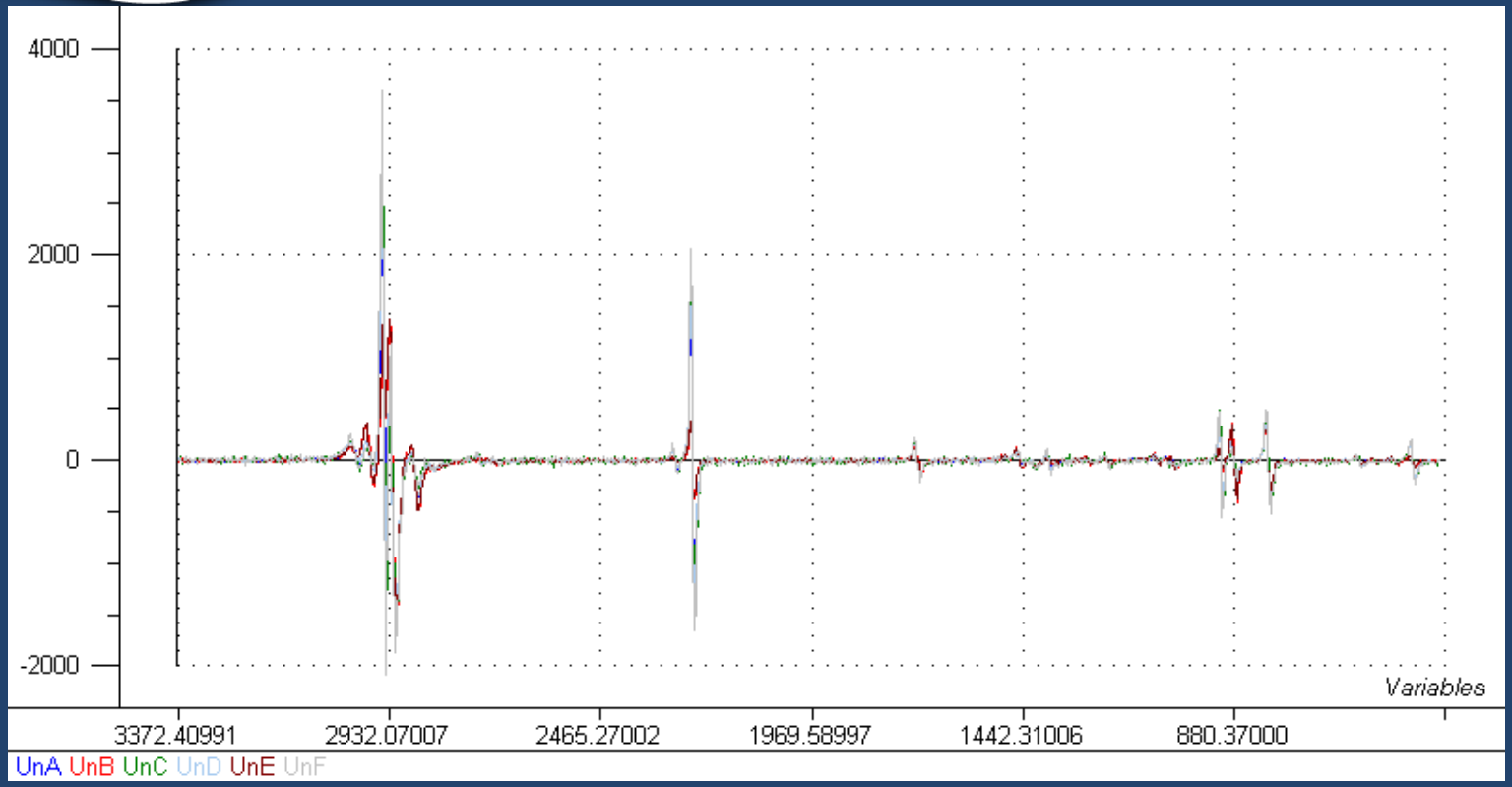


Die Factor loadings shows which spectral regions are important for a given PC:

- PC1: positiv: Acetonitrile
  negativ: Acetone
- PC2: positiv: Acetone
  negativ: Ethanol
- PC8: large noise component → for model not important



real concentration	PCR	PLS1	PLS2
Un A = 33.0 E	31.9 (4.3)	33.9 (3.6)	31.9 (4.3)
Un B = 11.0 E	12.1 (5.3)	12.2 (3.6)	12.1 (5.3)
Un C = 40.0 E	33.5 (34.9)	30.5 (36.1)	32.9 (34.8)
Un D = 33.0 E	32.1 (2.6)	32.2 (2.2)	32.1 (2.6)
Un E = 11.0 E	14.2 (5.1)	15.1 (3.9)	14.2 (5.1)
Un F = 40.0 E	34.2 (45.4)	28.5 (47.8)	33.3 (45.1)



real concentration	PCR	PLS1	PLS2
Un A = 33.0 E	30.1 (3.5)	29.9 (2.8)	29.2 (3.3)
Un B = 11.0 E	11.8 (3.7)	11.8 (2.6)	11.8 (2.8)
Un C = 40.0 E	36.5 (3.7)	36.5 (2.7)	35.4 (3.8)
Un D = 33.0 E	30.4 (3.8)	30.2 (2.4)	30.5 (2.7)
Un E = 11.0 E	11.5 (3.1)	11.5 (2.5)	11.8 (2.7)
Un F = 40.0 E	37.7 (4.7)	37.7 (3.1)	37.9 (3.2)