Selected Aspects of 40 Years Applied Chemometrics

VARMUZA Kurt

Vienna University of Technology, Austria

Institute of Statistics and
Mathematical Methods in Economics

Laboratory for **C**hemo**M**etrics
www.lcm.tuwien.ac.at

Autumn School of Chemoinformatics
25 - 26 Nov 2015, Tokyo, Japan
26 Nov 2015, The University of Tokyo

Contents of Tutorial

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)

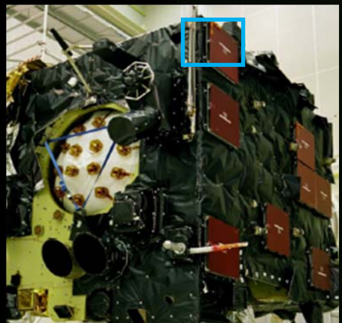
Supported by **Austrian Science Fund** (Project P26871-N20).

Collaboration with Peter **Filzmoser**, Irene **Hoffmann**, et al., and **COSIMA team** acknowledged..

This is an adjusted version of the lecture for presentation in web.

Rosetta mission (ESA) to **COMET**
67P/Churyumov-Gerasimenko - *Chury*

COSIMA



Launch 2 March 2004
Arrival 6 Aug 2014
Landing 12 Nov 2014

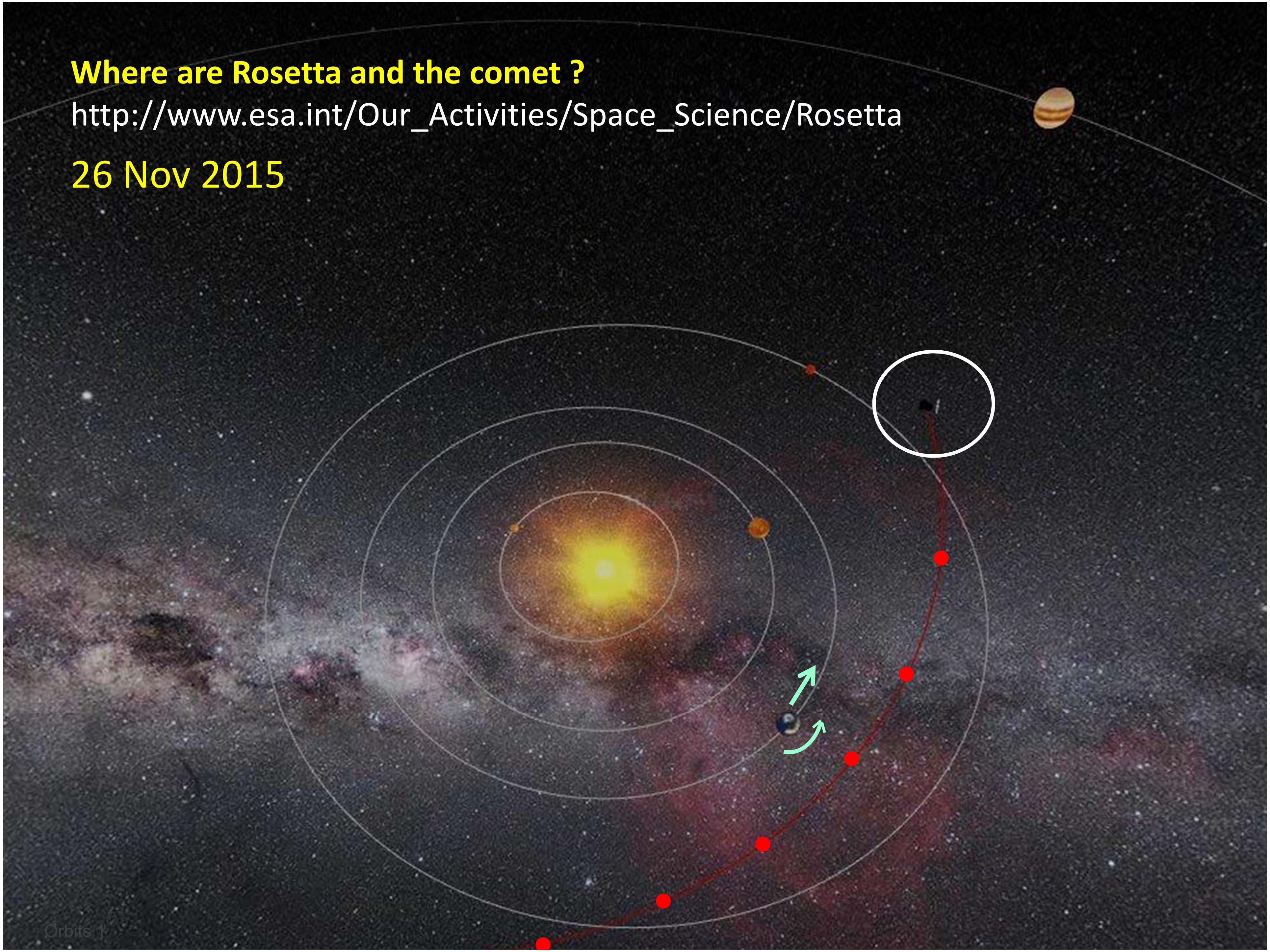
12 Aug 2015, ca perihelion (186 . 10^6 km from sun); 330 km from **comet**;
OSIRIS camera; animation, 17 pictures (ca 21 hours, incl. big outburst at 17:35 GMT).

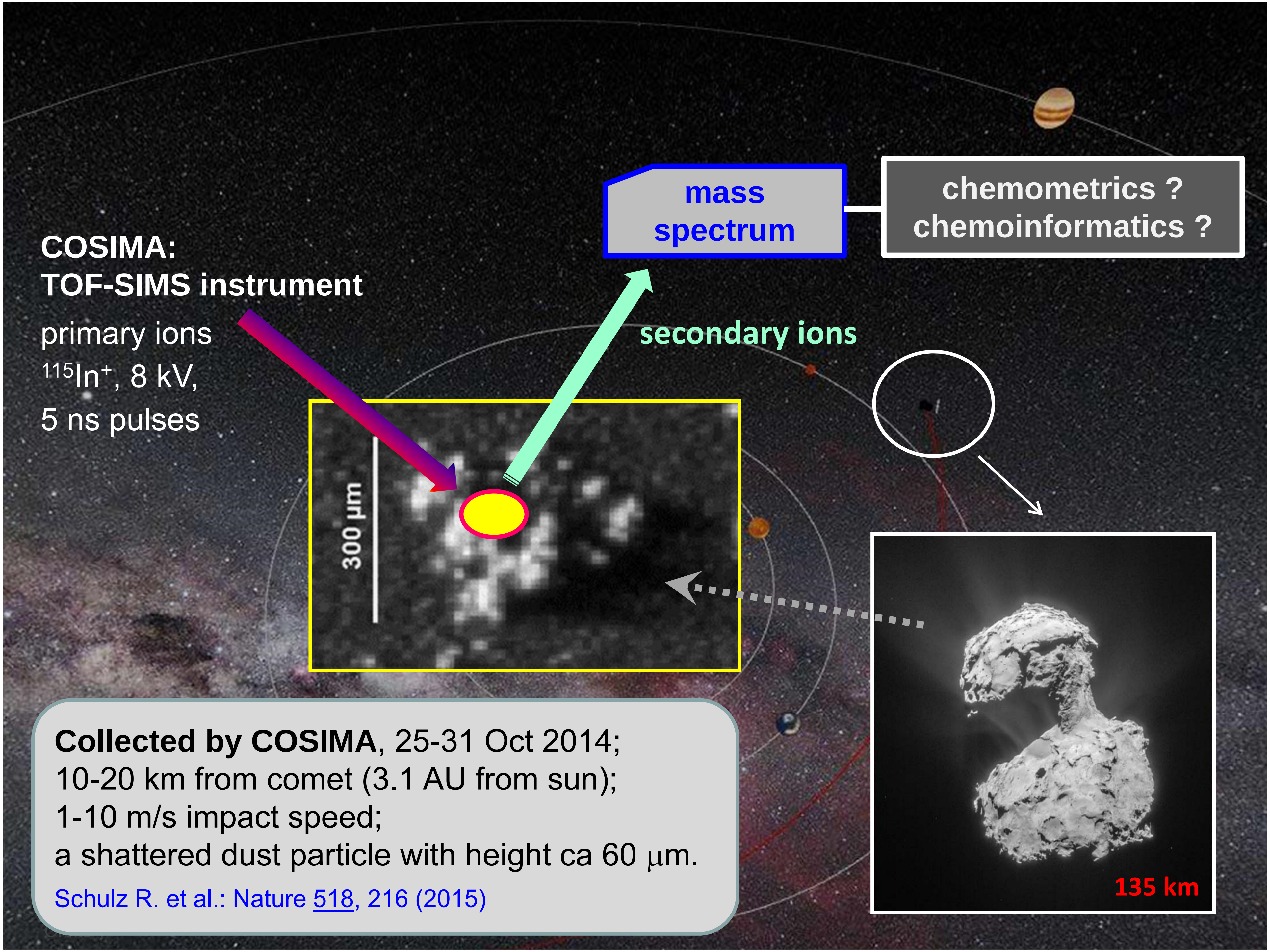
http://www.esa.int/spaceinimages/Images/2015/08/Approaching_perihelion_Animation

Where are Rosetta and the comet ?

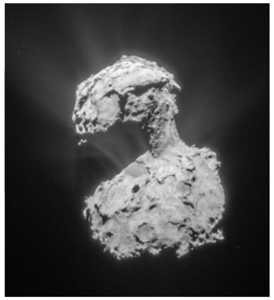
http://www.esa.int/Our_Activities/Space_Science/Rosetta

26 Nov 2015




mass spectrum
chemometrics ?
chemoinformatics ?
COSIMA:
TOF-SIMS instrument
primary ions
$^{115}In^{+}$, 8 kV,
5 ns pulses
secondary ions
300 µm
Collected by COSIMA, 25-31 Oct 2014;
10-20 km from comet (3.1 AU from sun);
1-10 m/s impact speed;
a shattered dust particle with height ca 60 µm.
Schulz R. et al.: Nature 518, 216 (2015)
135 km

Comet Material



… most **pristine** material in our solar system in the form of ice, mixed with dust, **silicates**, and **refractory organic** material (probably many different species) …

… aggregate of **pre-solar grains** (grains that existed prior to the formation of the Solar System), …

… comet material (water/organic/inorganic) may have been **the seed of life on earth …**

[1] Goesmann F. et al.: Science 349, issue 6247 (2015)
[2] Greenberg J. M. et al.: Space Sci. Rev., 90, 149 (1999)
[3] Kissel J. et al.: Space Sci. Rev., 128, 823 (2007)

Contents of Tutorial

1 Basics (history, strategies)

2 Empirical multivariate models (optimum complexity, evaluation)

3 One class classification

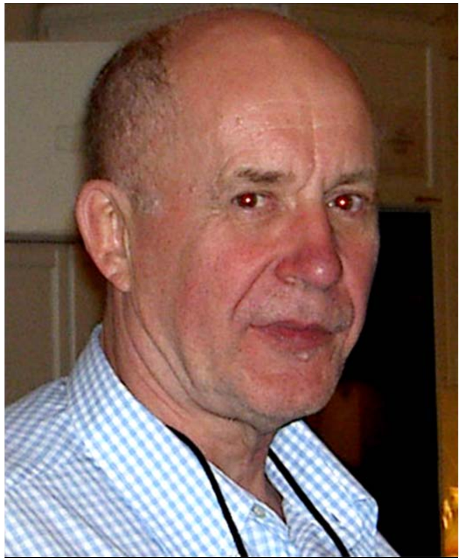
With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)

Chemometrics

- Uses methods from statistics, mathematics, and informatics,
- to extract relevant information from chemical / physical data,
- and to select or optimize chemical processes and experiments.

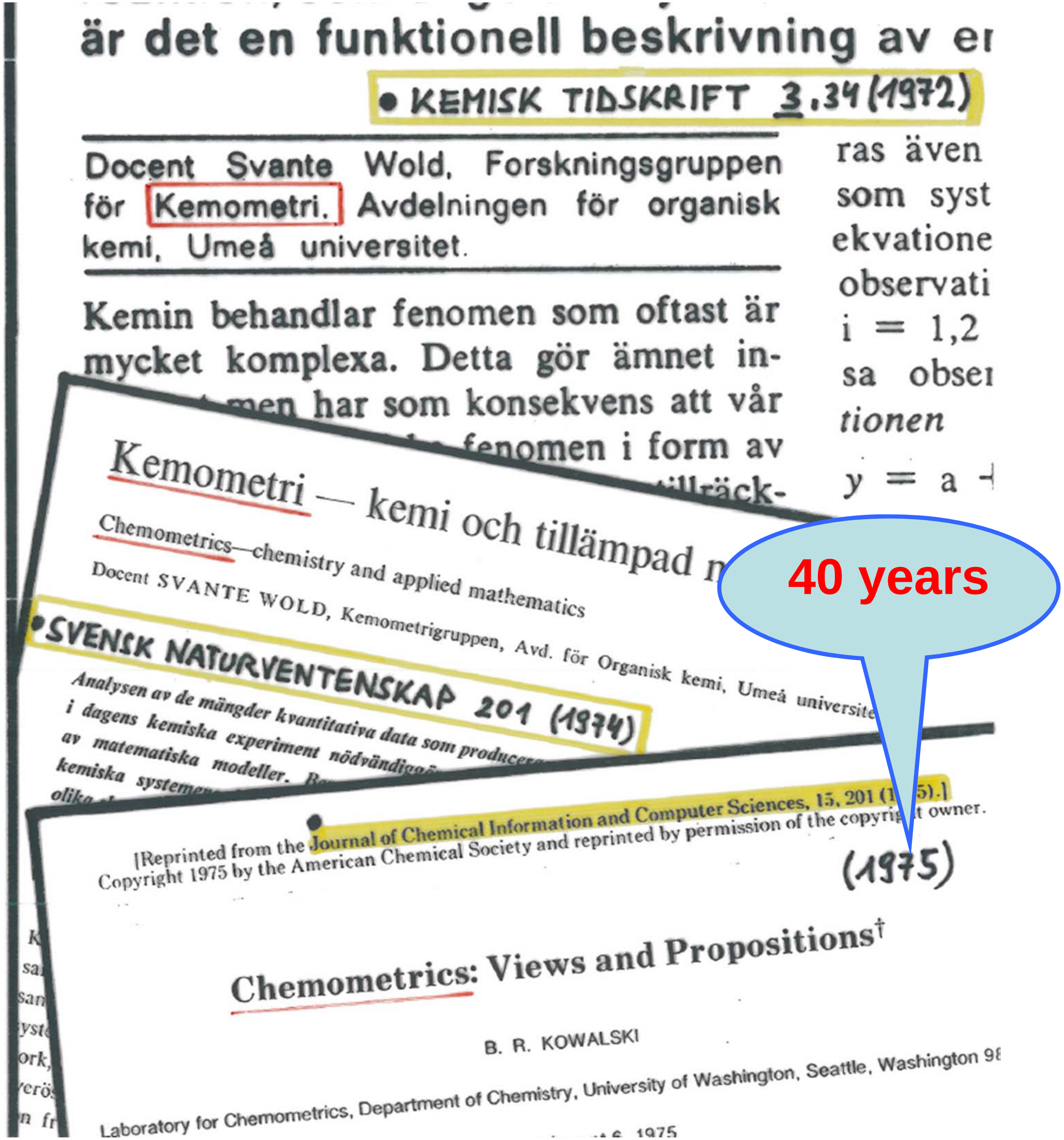
Perhaps a part of ChemoInformatics

Mostly using multivariate data

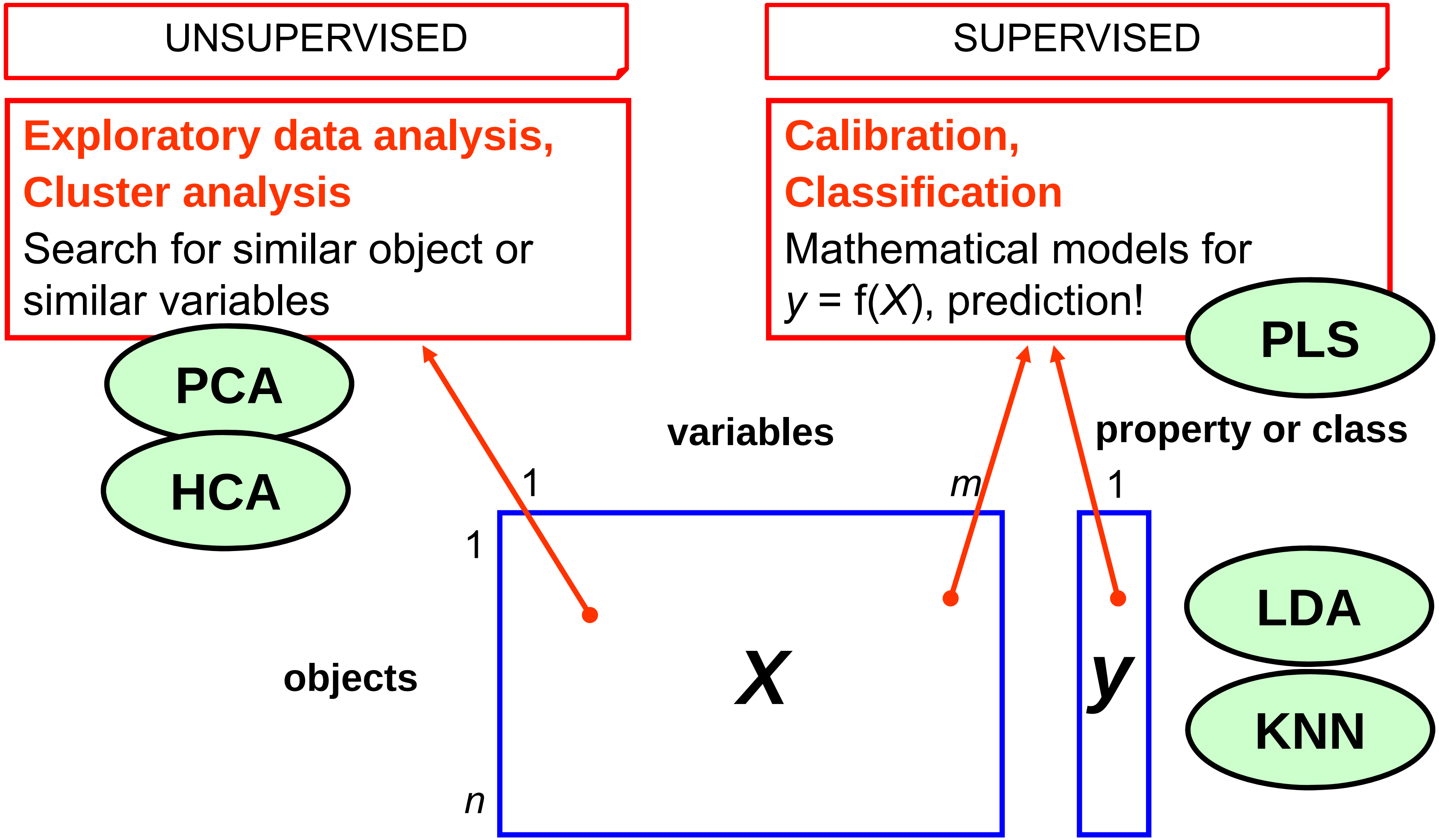


Svante Wold
Lappeenranta, 2007

Bruce Kowalski
(1942 - 2012)
Loen, 2011



Multivariate data analysis



Typical: collinear variables, and often $m > n$

Multivariate data analysis

- **Exploratory data analysis**
- **Multivariate calibration**
- **Multivariate classification**

... data-driven ...

... empirical models !

Some aspects for empirical models (in chemometrics)

- More variables than objects ($m > n$)
- Multicollinearity
- Parsimonious (interpretation, understanding)
- Tested (for new cases, domain, performance)
- Robust (data distribution, outliers)

Trial and error …

Contents of Tutorial

1	Basics (history, strategies)
2	**Empirical multivariate models (optimum complexity, evaluation)**
3	One class classification

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)

Optimum model complexity

Optimum Model

Underfitting

model too simple

Overfitting

model too much adapted to calibration data

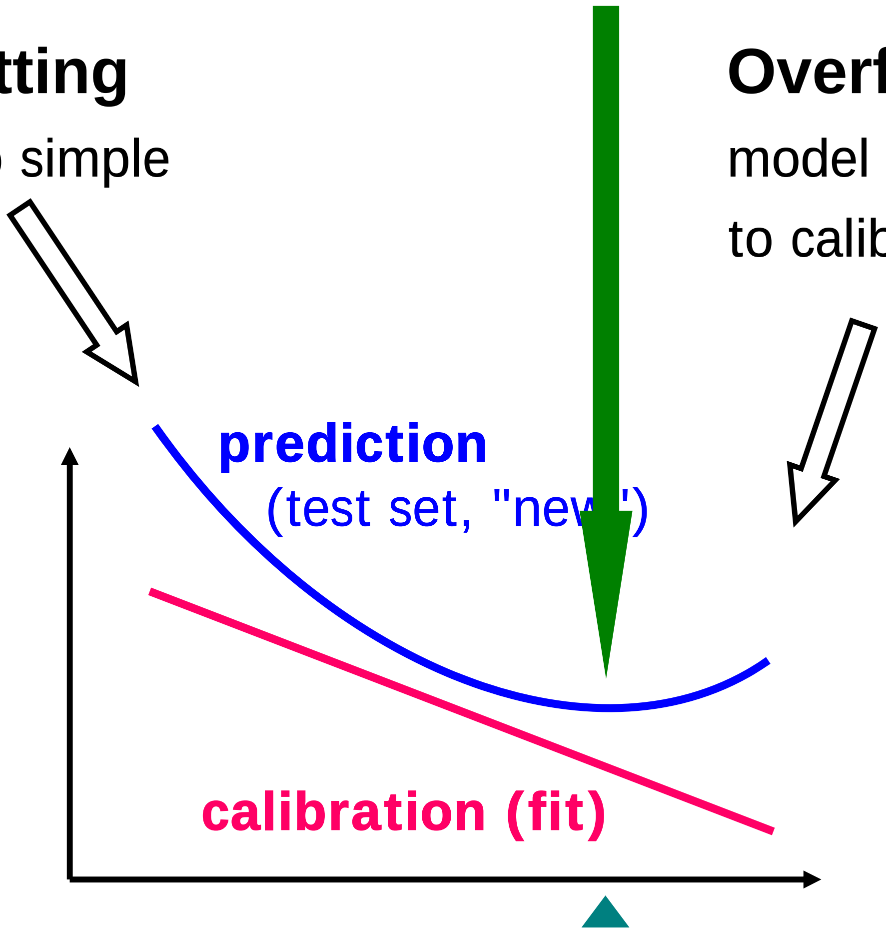
Error

mean squared error

$MSE = (1/n)\ \Sigma e_i^2$

***MSE*_C_, _c_alibration**

***MSE*_P_, _p_rediction**

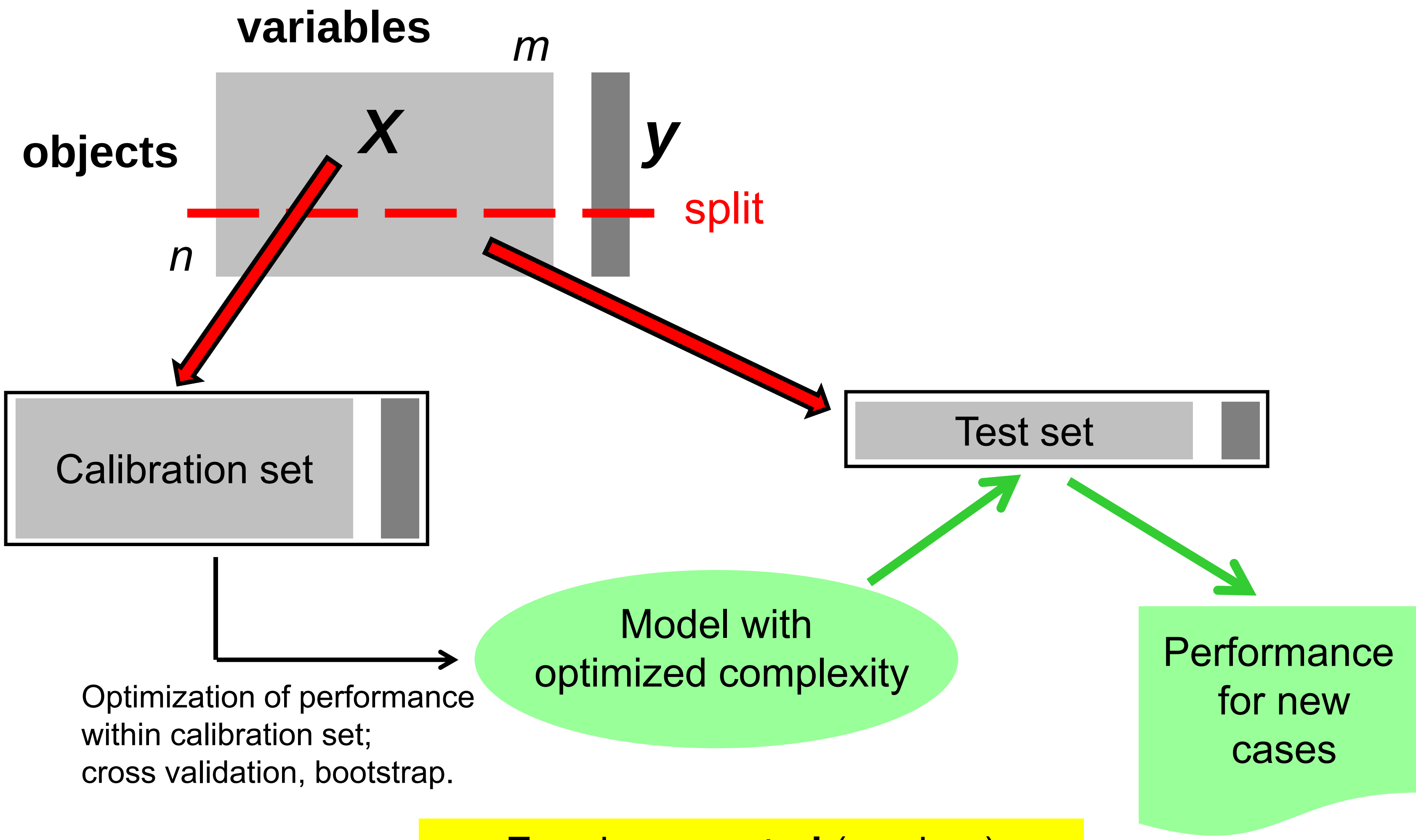


Complexity of model

No. of PCA/PLS-components,
no. of variables,
non-linearities

Multivariate data analysis

Optimization and Evaluation (separated)



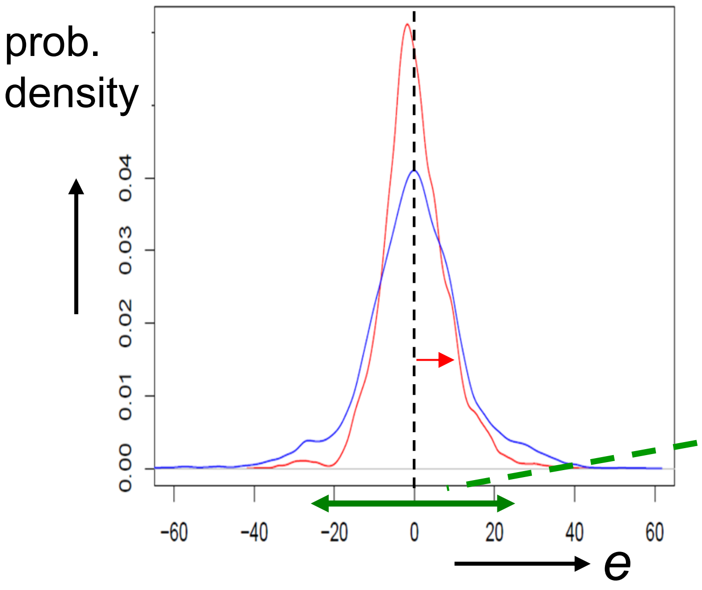
E.g., by **repeated** (random)
double cross validation (rdCV)

Estimation of model performance: CALIBRATION

y_i	reference ("true") value for object *i*
$\hat{y}_i$	calculated (predicted) value (test set !)
$e_i = y_i - \hat{y}_i$	prediction error for object (residual)
$i = 1 \dots z$	*z* is the number of predictions

Specify: ☞ which data set (calibration set, **test set**)
☞ which strategy (cross validation, ...)

Distribution of prediction errors



bias = mean of prediction errors e_i

SEP = standard deviation of prediction errors e_i
= **S**tandard **E**rror of **P**rediction

CI = confidence interval, $CI_{95\%} \approx \pm 2*SEP$

User friendly ! All in units of *y* !

Estimation of model performance

SEP (or any other performance criterion)
must NOT be considered as a single number.

Depends on

- the used objects, and variables;
- the random split in a CV (or a bootstrap);
  repetitions highly recommended, e. g., rdCV.

It is an estimation.

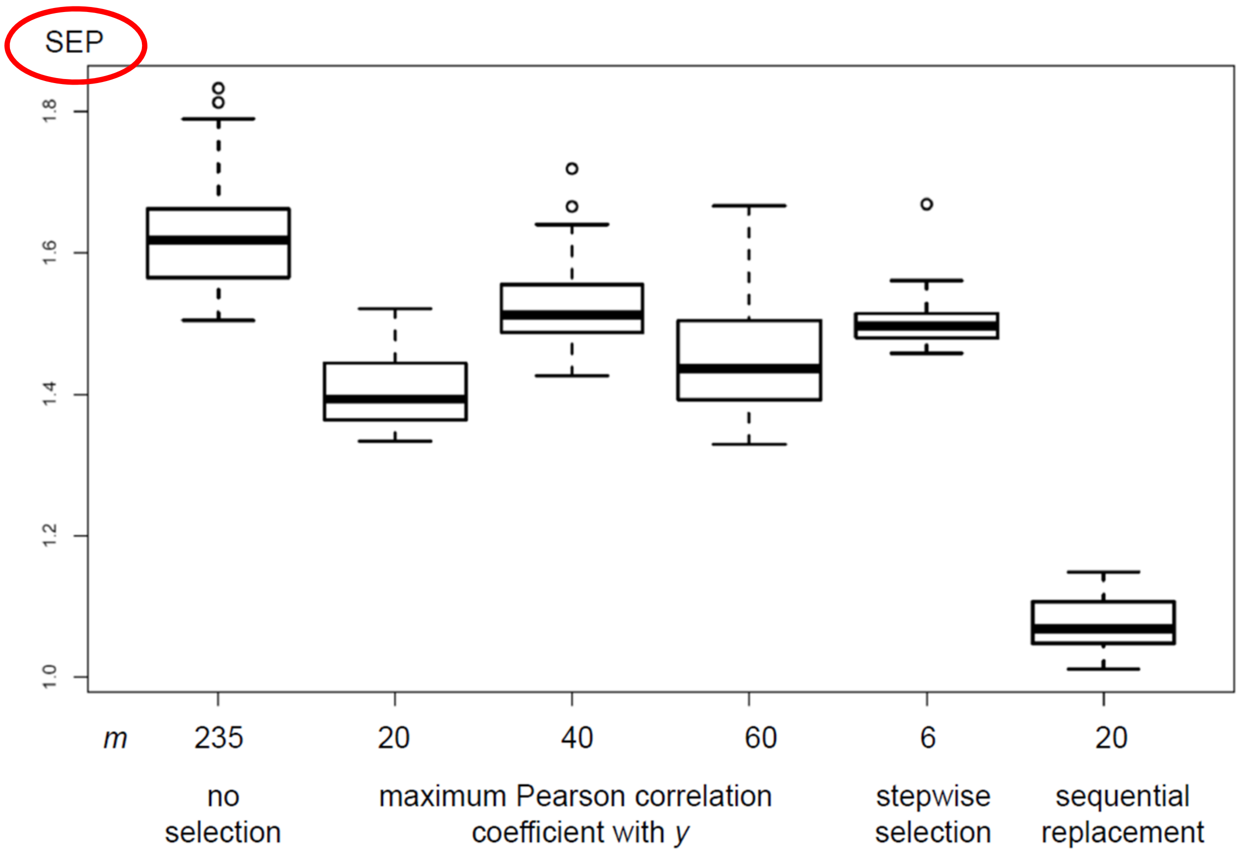
It has a distribution (variation) - boxplots recommended.

Multivariate data analysis - EXAMPLE

Estimation of model performance: CALIBRATION

X *n* = 166 **fermentation samples** (cereals), centrifuged
m = 235 **NIR** absorbances, 1115 - 2285 nm (step 5 nm), 1st deriv., (7 points, 2nd order)

Y **ethanol** content, reference method HPLC; 21.7 - 88.1 g/L



PLS regression,
[ethanol] = f(NIR abs.)

rdCV
30 repetitions,
4 and 7 segments

Comparison of variable selection methods.

Liebmann B., Friedl A., Varmuza K.: *Anal. Chim. Acta* 642 (2009) 171.
Varmuza K., Filzmoser P.: In Khanmohammadi M. (ed.), *Current Applications of Chemometrics*, Nova Science Publishers, New York, USA (2015), p. 15.

Estimation of model performance: CLASSIFICATION

Class assignment table
(binary classification)

	assigned class 1	assigned class 2	sum (no. of objects)
true class 1	$\boxed{n_{11}}$	n_{12}	n_1
true class 2	n_{21}	$\boxed{n_{22}}$	n_2
sum	$n_{\rightarrow 1}$	$n_{\rightarrow 2}$	n

Predictive ability class 1 $P_1 = n_{11}/n_1$

class 2 $P_2 = n_{22}/n_2$

Average predictive ability $P = (P_1 + P_2)/2$

Avoid: Overall predictive ability $= (n_{11} + n_{22})/n$

Extraterrestrial Material

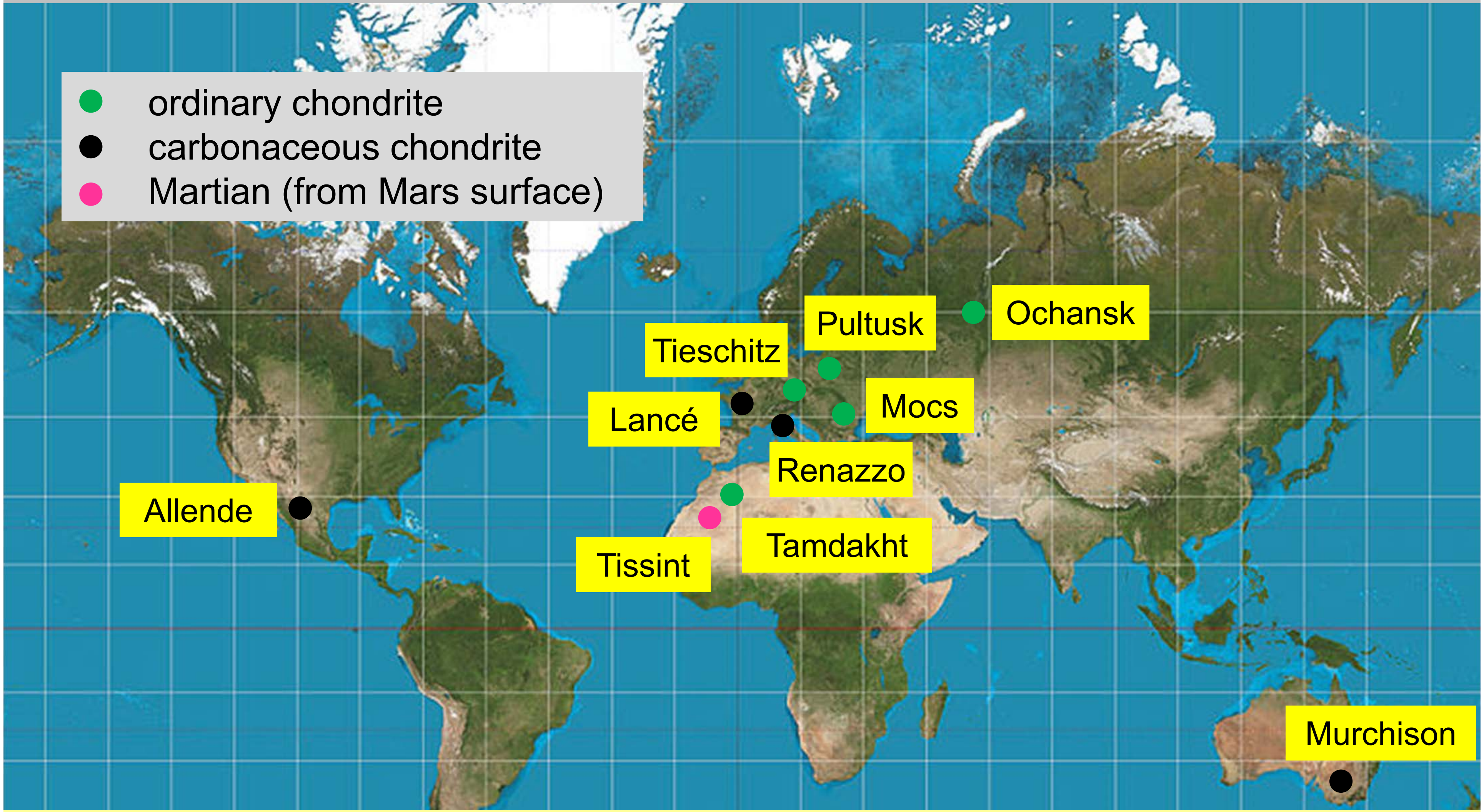
remote optical spectroscopy

- **Collected in space and brought safely to Earth (Stardust [near comet], Hayabusa [asteroid surface])**
- **Measurements in space (Rosetta, Mars, Moon, …)**
- **Coming autonomously (meteorites, ca 40,000 t/year)**
  ***Finds* and *Falls* (witnessed, observed, samples)**

Multivariate data analysis - EXAMPLE

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

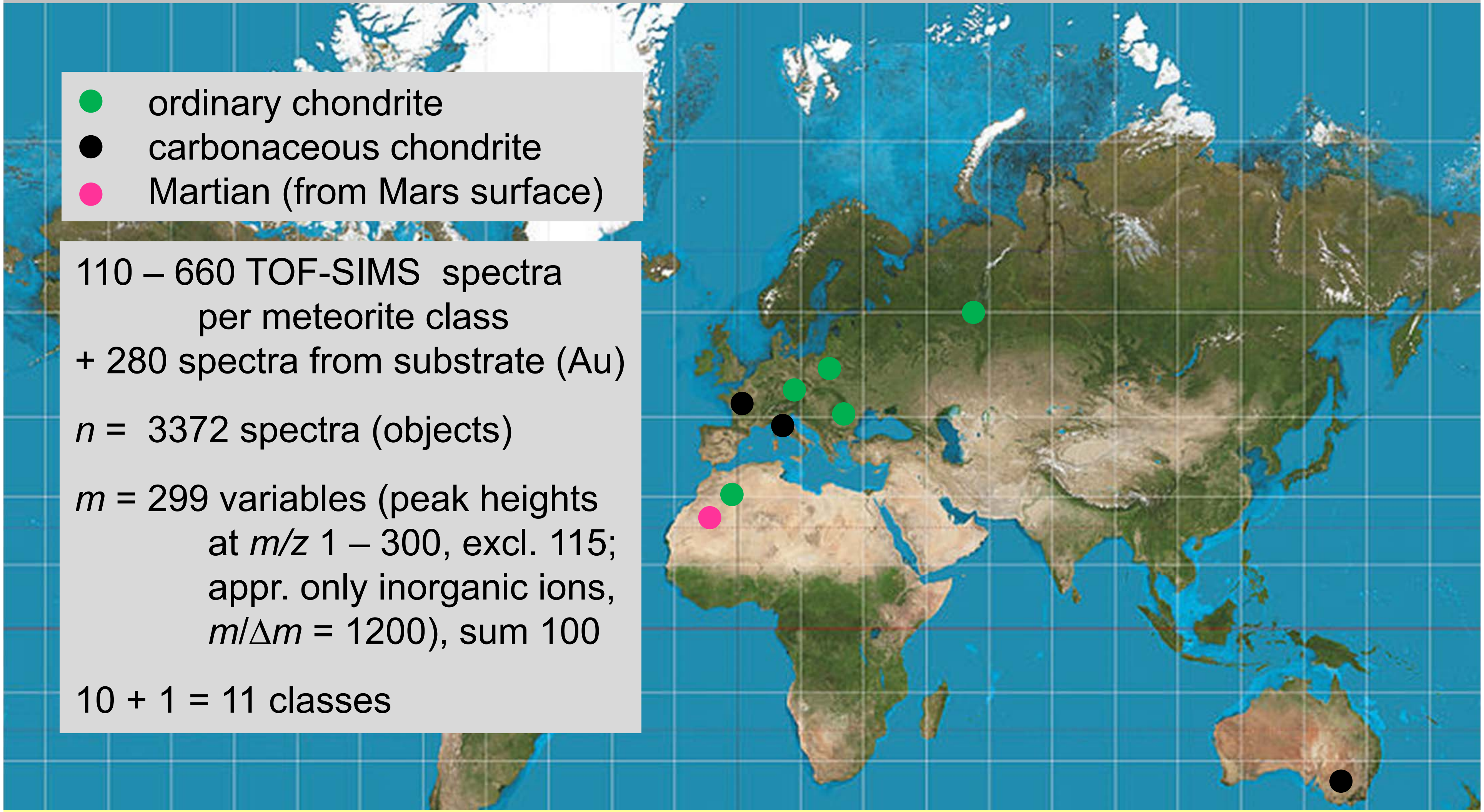


Samples from Natural History Museum (NHM) Vienna: **10 meteorites**

Multivariate data analysis - EXAMPLE

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS



- ordinary chondrite
- carbonaceous chondrite
- Martian (from Mars surface)

110 – 660 TOF-SIMS spectra per meteorite class
+ 280 spectra from substrate (Au)

n = 3372 spectra (objects)

m = 299 variables (peak heights at m/z 1 – 300, excl. 115; appr. only inorganic ions, $m/\Delta m$ = 1200), sum 100

10 + 1 = 11 classes

Samples from Natural History Museum (NHM) Vienna: **10 meteorites**

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

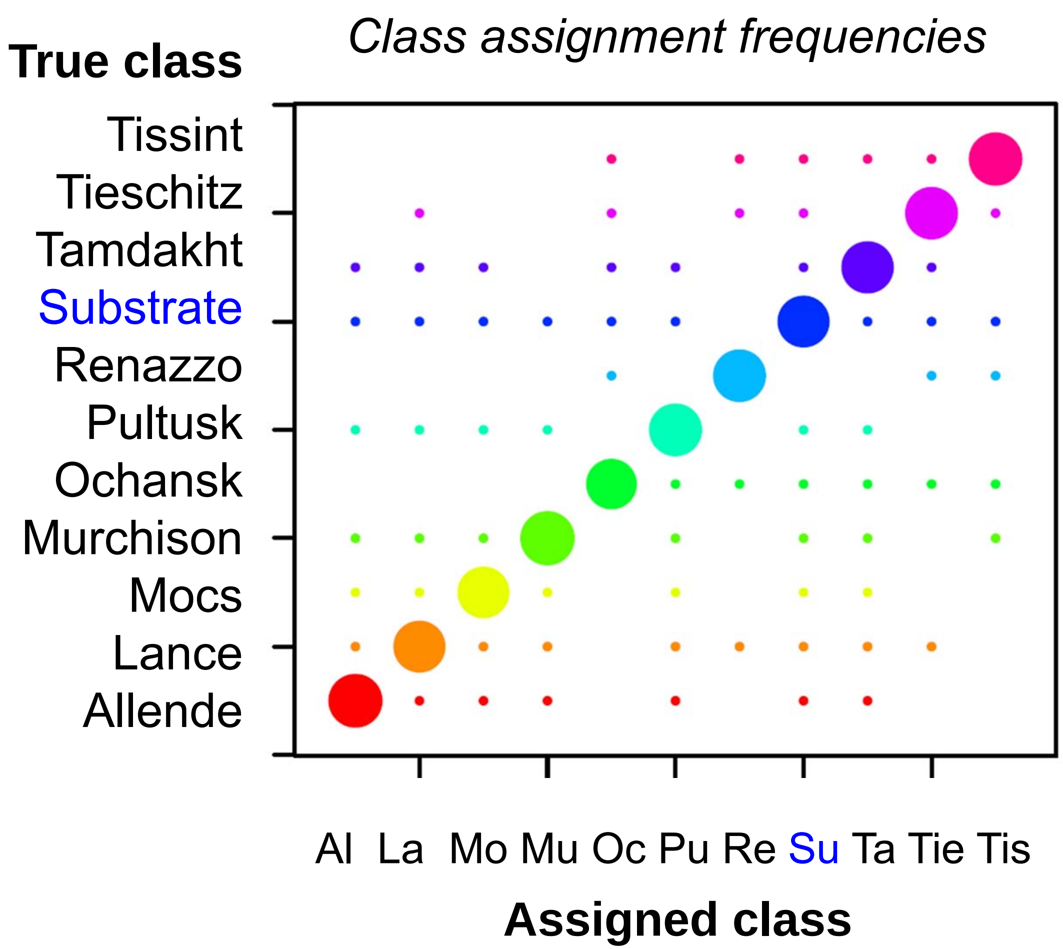
KNN classification,

Euclidean distance,
rdCV strategy
(20 repetitions,
2 and 5 segments),

optimum no. of
neighbors = 1

Predictive abilities
(mean of 20 repetitions)
per meteorite class:
90 – 97 %

Total mean: 94 %



Multivariate data analysis - EXAMPLE

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

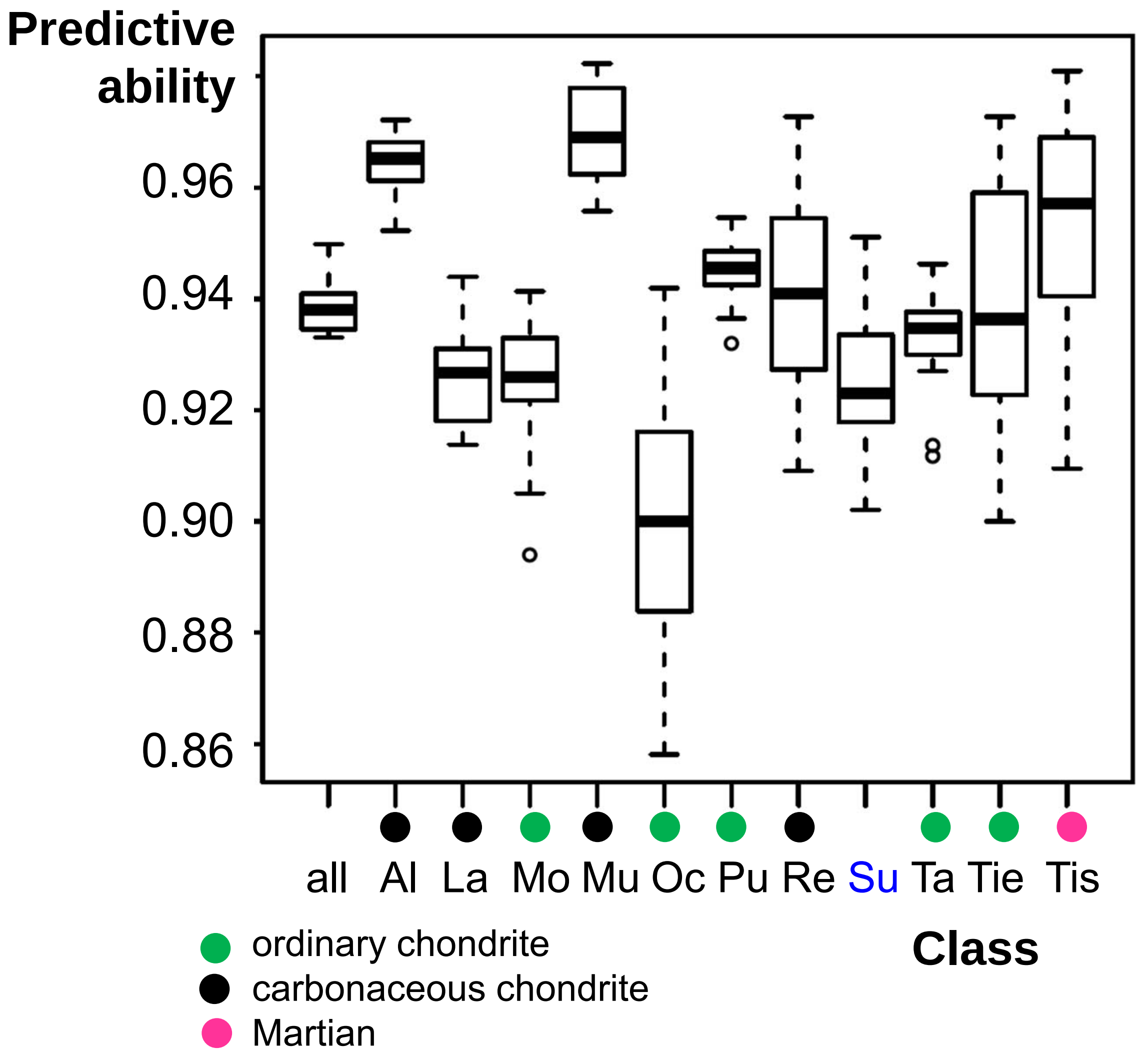
KNN classification,

Euclidean distance,
rdCV strategy
(20 repetitions,
2 and 5 segments),

optimum no. of
neighbors = 1

Predictive abilities
(mean of 20 repetitions)
per meteorite class:
90 – 97 %

Total mean: 94 %



Multivariate data analysis - EXAMPLE

Estimation of model performance: CLASSIFICATION

Classification of meteorites by TOF-SIMS

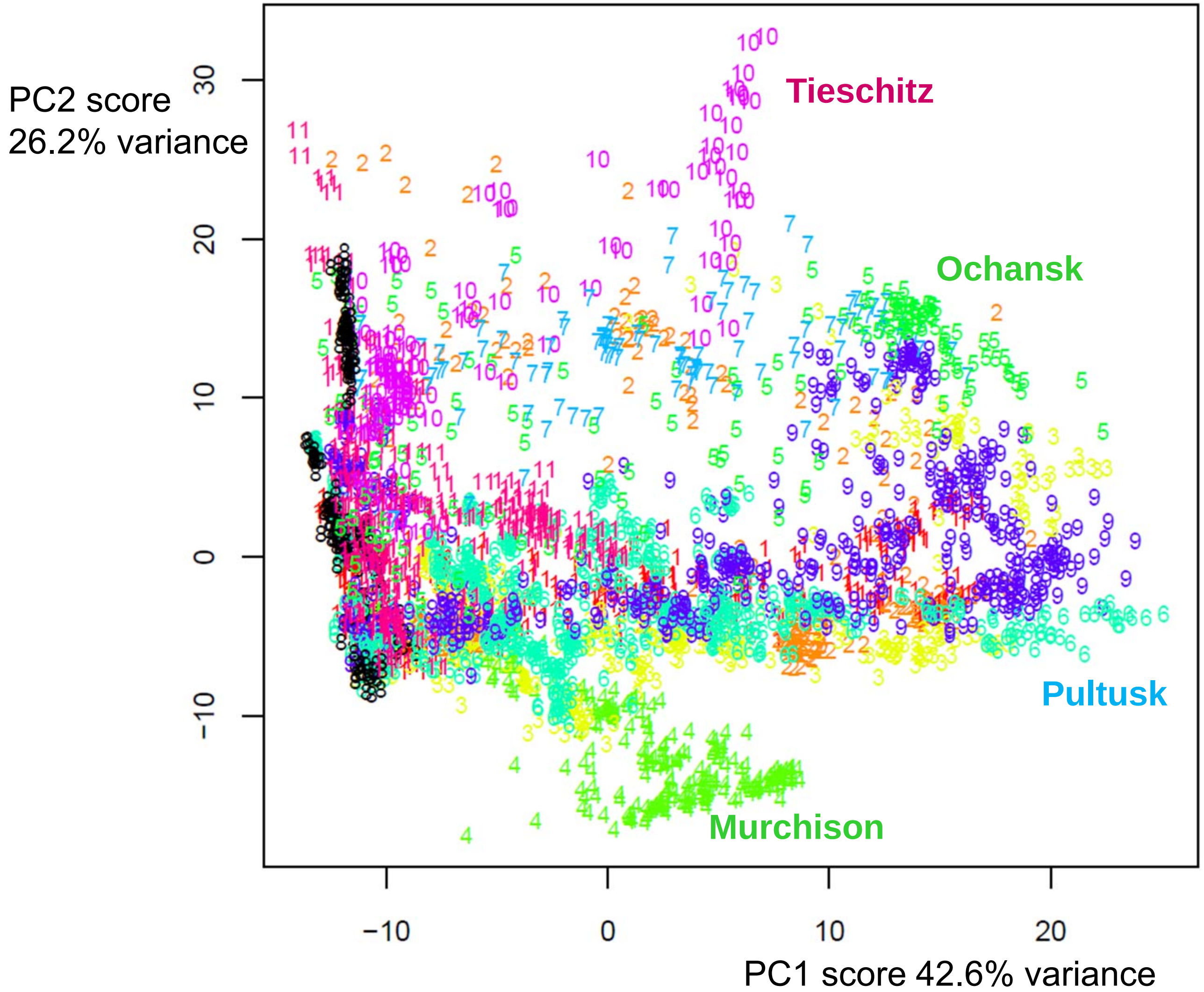
PCA

X (3372 x 299),

row sum = 100

Classes

11 Tissint
10 Tieschitz
9 Tamdakht
8 Substrate
7 Renazzo
6 Pultusk
5 Ochansk
4 Murchison
3 Mocs
2 Lance
1 Allende



Contents of Tutorial

1 Basics (history, strategies)

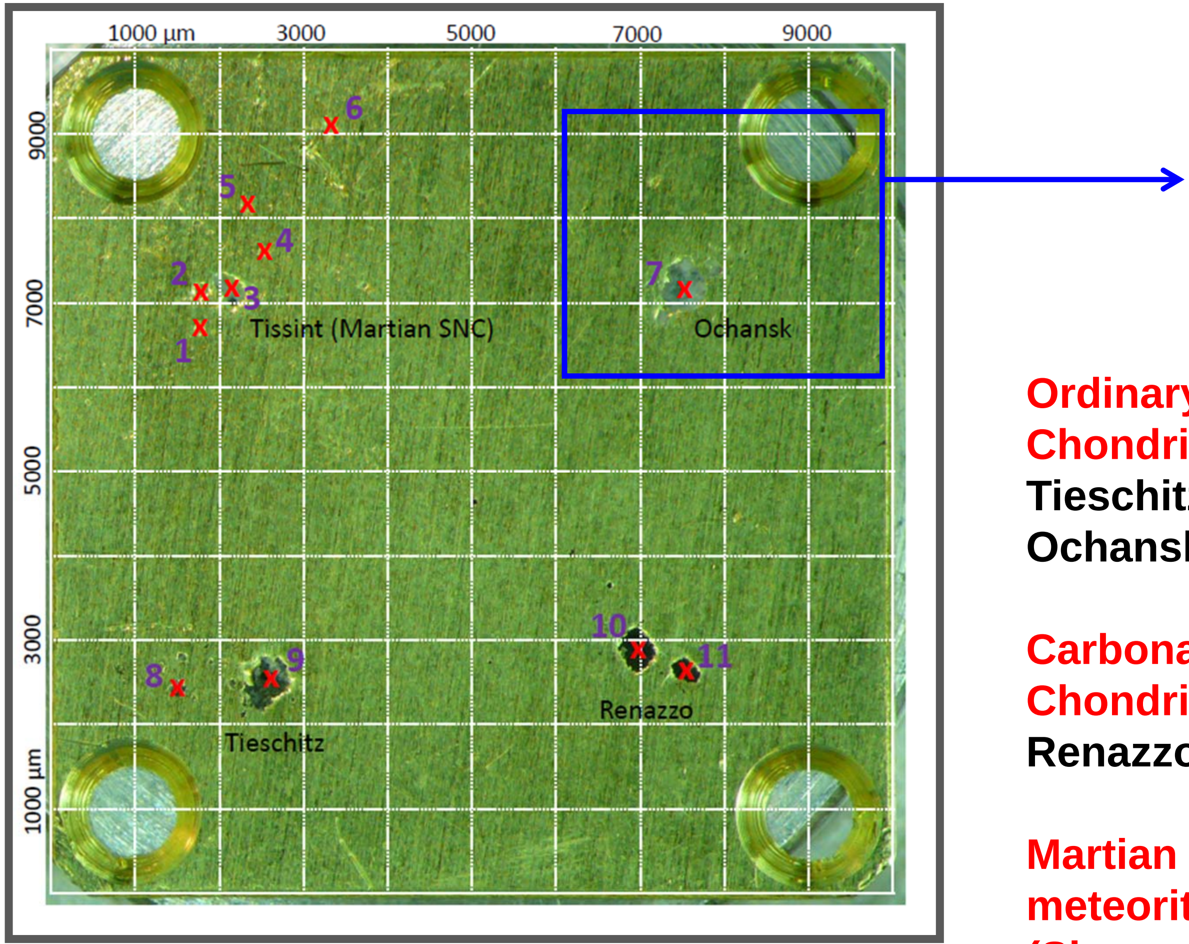
2 Empirical multivariate models
(optimum complexity, evaluation)

3 One class classification

With examples from TOF-SIMS measurements on meteorite samples and cometary dust particles (Rosetta)

TOF-SIMS on Meteorite Grains

Meteorite grains prepared on a gold foil (10 mm x 10 mm)



Samples
Christian Köberl,
Franz Brandstätter,
Ludovic Ferrière,
Natural History Museum Vienna

Preparation
Cécile Engrand,
Univ. Paris Sud (Orsay)

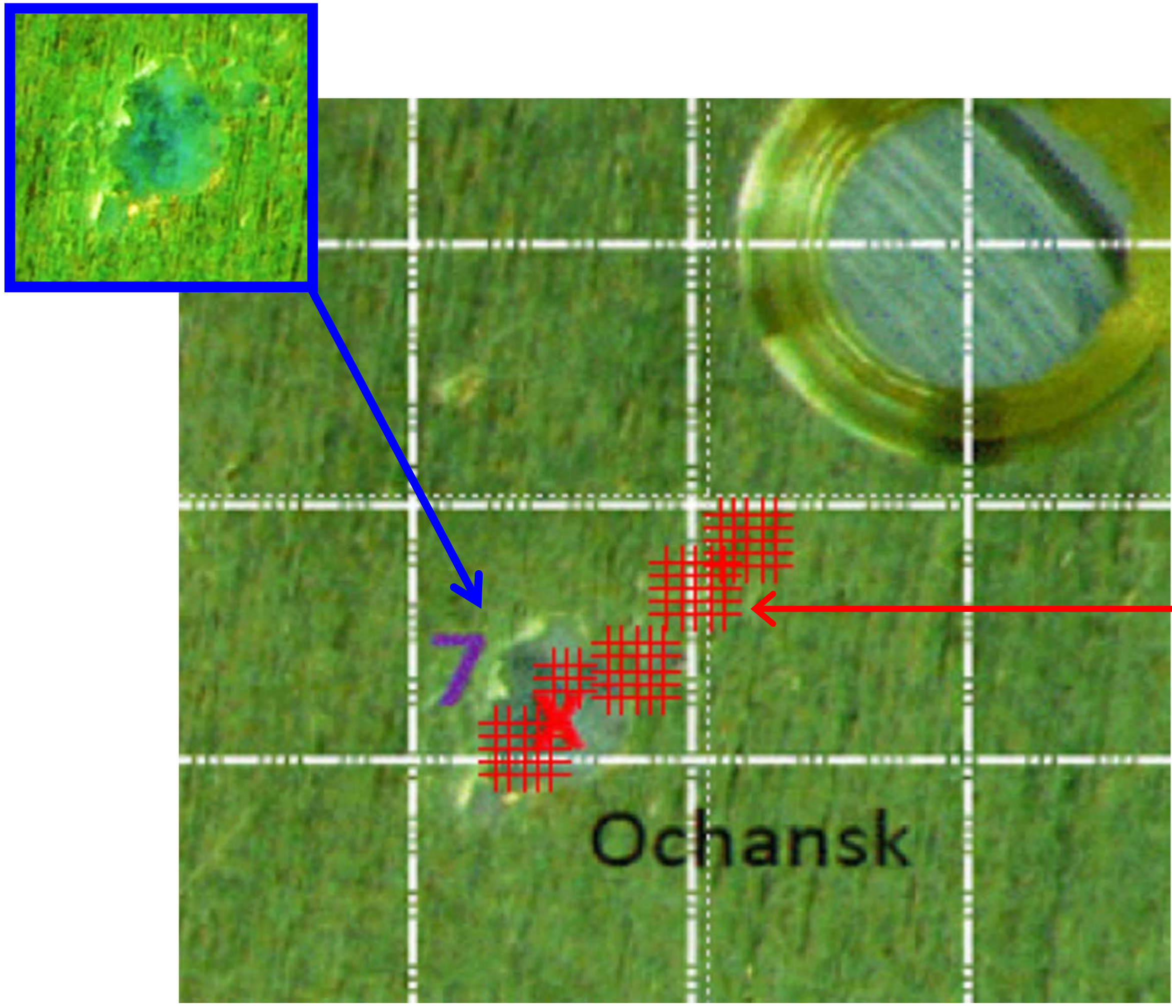
TOF-SIMS (COSIMA twin)
Martin Hilchenbach,
Max Planck Institute for Solar System Research, Göttingen

Ordinary Chondrites
Tieschitz
Ochansk

Carbonaceous Chondrite
Renazzo

Martian meteorite (Shergottite)
Tissint

TOF-SIMS on Meteorite Grains

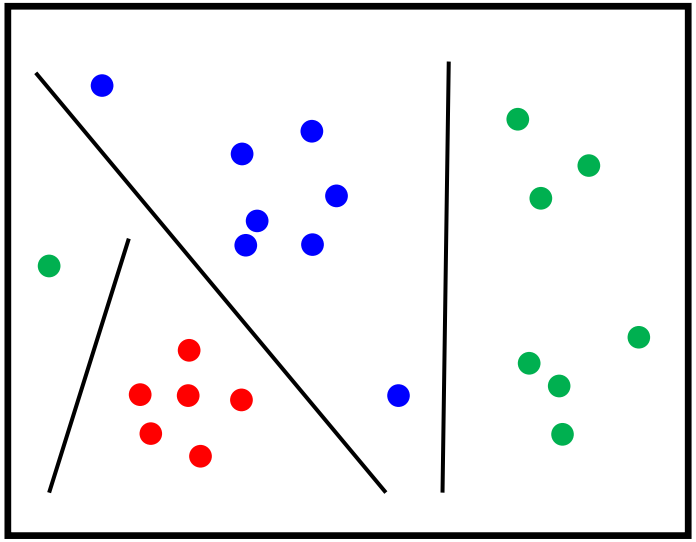


1000 μm

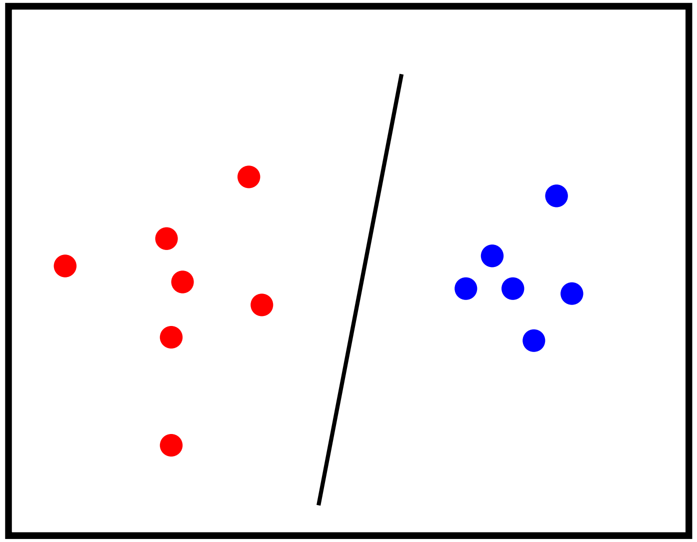
Photographic picture of a target with a meteorite grain.

TOF-SIMS measuring positions (155 query spectra)

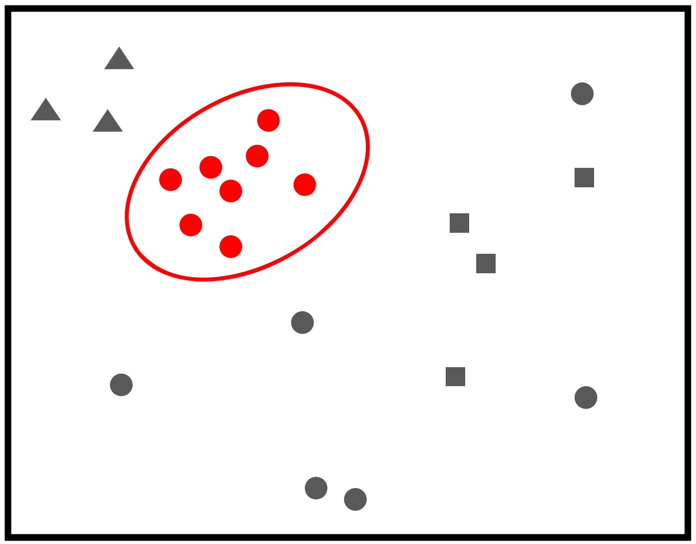
63 background (Off grain) spectra



Multi-class classification

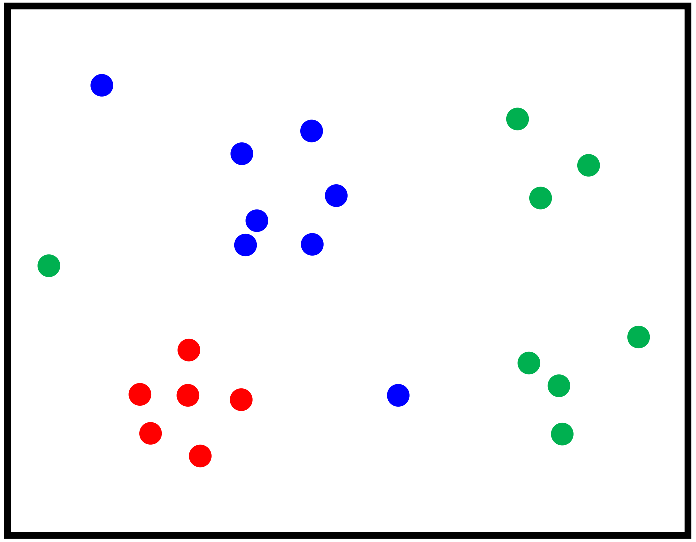


Binary classification

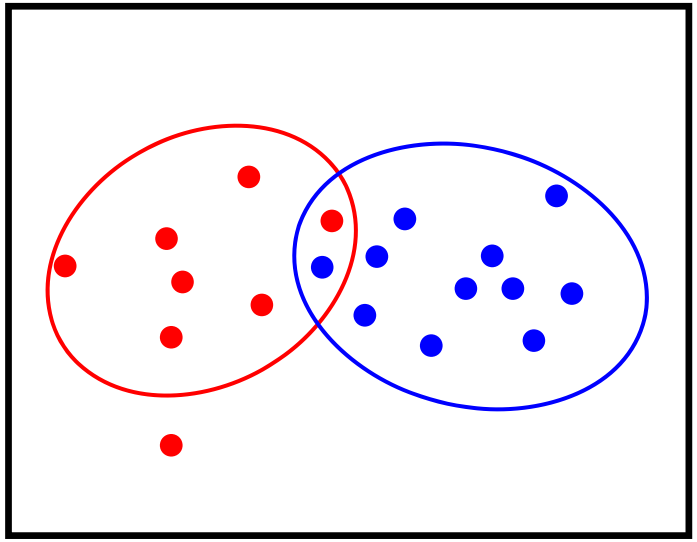


One-class classification

Only one ***target*** class,
all others (outlier)

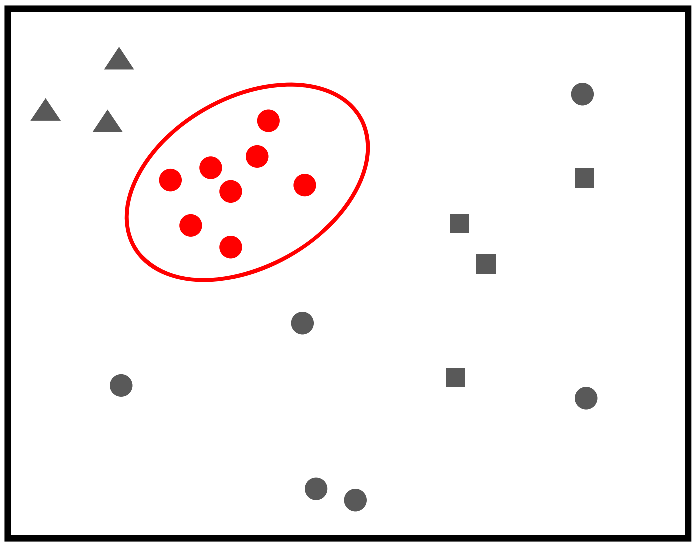


Multi-class classification



Binary classification

SIMCA (S. Wold):
PCA model of each class

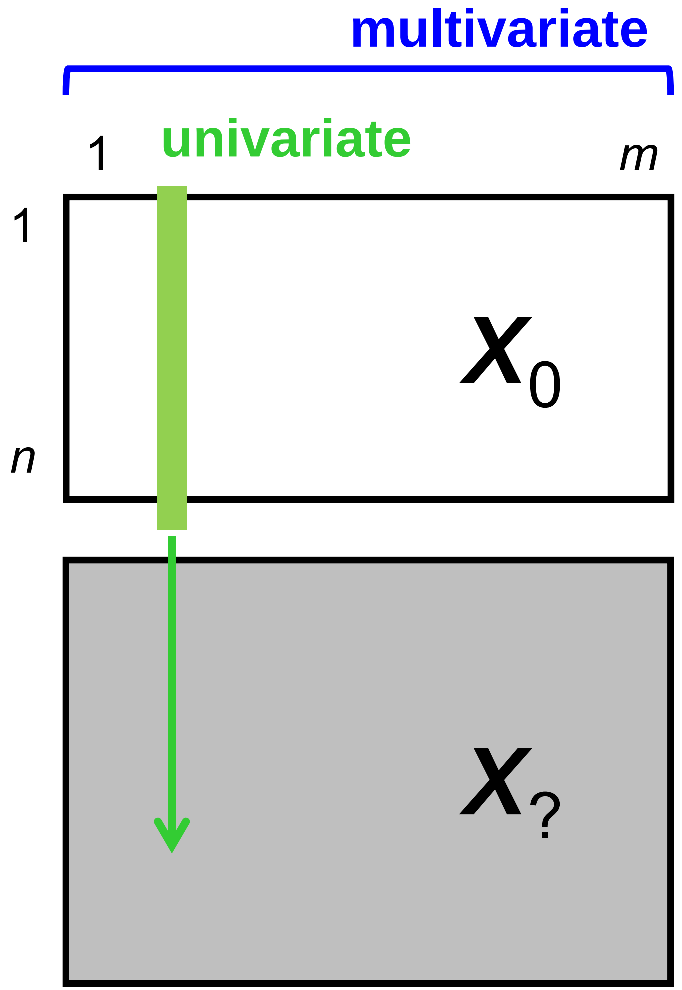


One-class classification

Only one ***target*** class,
all others (outlier)

On-grain – Off-grain

Recognition of potentially relevant spectra (TOF-SIMS)



Target class spectra
(background, off-grain)

Query spectra
off-grain (background)
or not (potentially relevant, on-grain)

Recognition of potentially relevant spectra (TOF-SIMS)

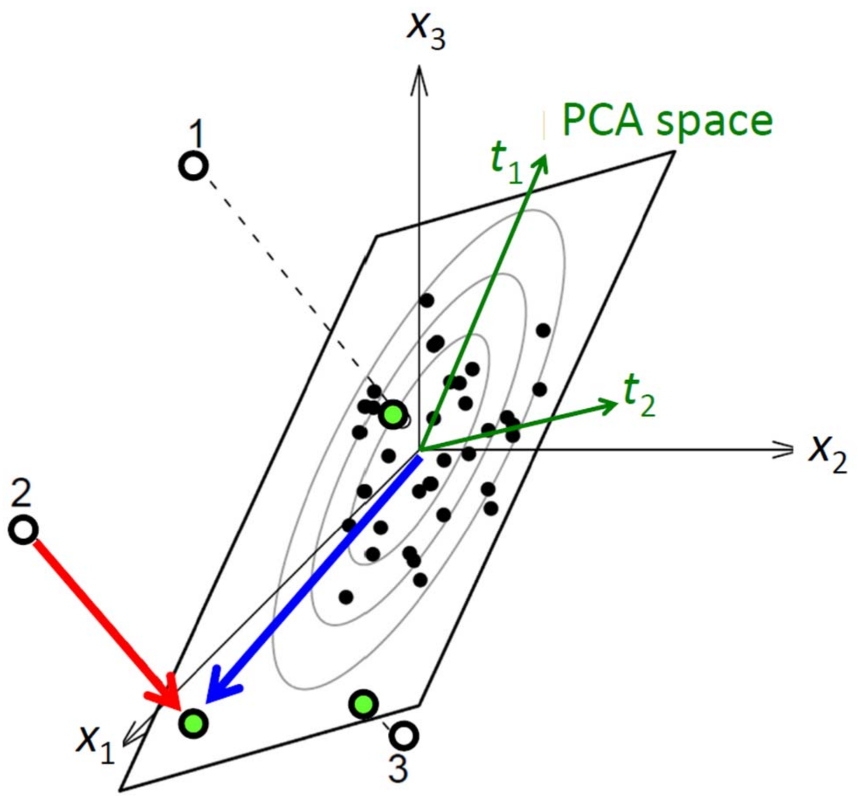
- Univariate; intensity of a selected ion (element, e. g., Fe, …)
- Ratios of variables (or other ‚simple' heuristic combinations)
- **One-class classification** (*target* class = off-grain) *[supervised]*
  1 **PCA: orthogonal and score distance**
  2 **KNN distance distribution**
  - Weights from sparse and robust PLS-DA *[supervised]*
  - Cluster analysis *[unsupervised]*
- Deconvolution
  - NMF (nonnegative matrix factorization) *[unsupervised]*

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

1

Distances to PCA model made from *Off-grain spectra*



Demo scheme

***Target* class: X_0** , m = 3 variables;

PCA model with A = 2 components (scores t_1 and t_2);

- ● Projection of X_0-points into the PCA model (plane, defined by t_1 and t_2)
- O Query points 1, 2, 3 in x-space
- ● Projections of query points into the PCA plane

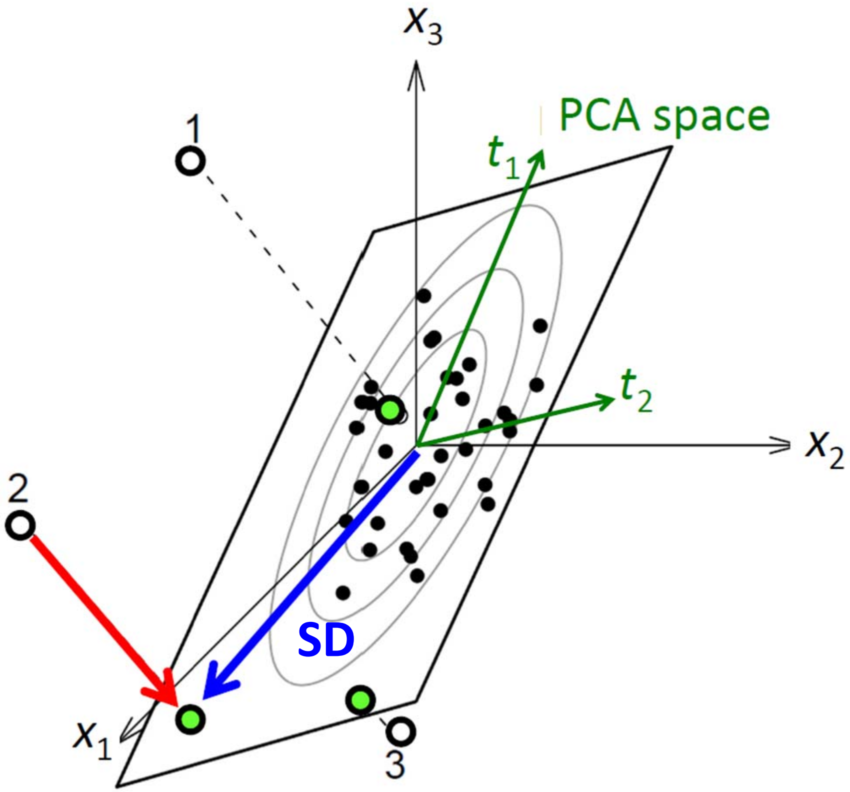
- Xu Y., Brereton R.: J. Chem. Inf. Model., **45**, 1392 (2005)
- Pomerantsev A.L.: J. Chemom., **22**, 601 (2008)
- Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA (2009)

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

1

Distances to PCA model made from *Off-grain spectra*



Demo scheme

***Target* class: X_0** , m = 3 variables;

PCA model with A = 2 components (scores t_1 and t_2);

- Projection of X_0-points into the PCA model (plane, defined by t_1 and t_2)
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

Score distance (SD)
= *Mahalanobis* distance from center, measured in the PCA space (plane).
Describes the distance to the center (of the background spectra) in PCA score space, considering the covariance structure of the x-variables.

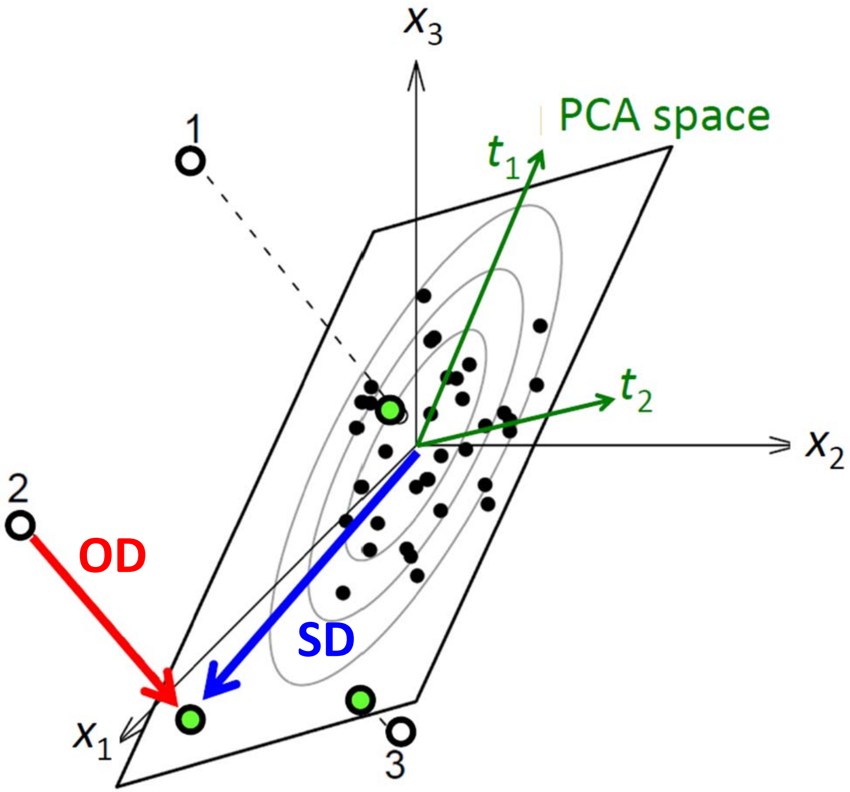
- Xu Y., Brereton R.: J. Chem. Inf. Model., **45**, 1392 (2005)
- Pomerantsev A.L.: J. Chemom., **22**, 601 (2008)
- Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA (2009)

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

1

Distances to PCA model made from *Off-grain spectra*



Demo scheme

***Target* class: $\boldsymbol{X}_0$** , m = 3 variables;

PCA model with A = 2 components (scores t_1 and t_2);

- Projection of X_0-points into the PCA model (plane, defined by t_1 and t_2)
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

Orthogonal distance (OD)

= Distance in x-space between point and its projection onto the PCA space.

Describes information loss by projecting into the A-dimensional PCA score space.

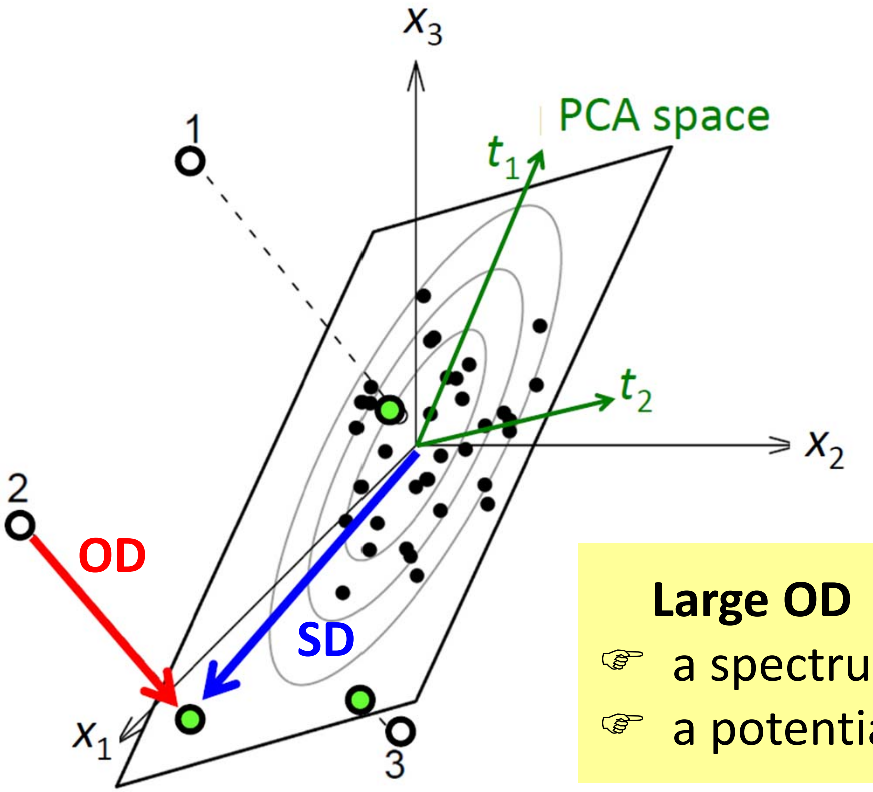
- Xu Y., Brereton R.: J. Chem. Inf. Model., **45**, 1392 (2005)
- Pomerantsev A.L.: J. Chemom., **22**, 601 (2008)
- Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA (2009)

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

1

Distances to PCA model made from *Off-grain spectra*



Demo scheme

***Target* class: X_0** , m = 3 variables;

PCA model with A = 2 components (scores t_1 and t_2);

- Projection of X_0-points into the PCA model (plane, defined by t_1 and t_2)
- Query points 1, 2, 3 in x-space
- Projections of query points into the PCA plane

Large OD - AND/OR large SD - indicate an *outlier*;

☞ a spectrum not belonging to the background group,

☞ a potentially relevant spectrum.

☐ Xu Y., Brereton R.: J. Chem. Inf. Model., **45**, 1392 (2005)

☐ Pomerantsev A.L.: J. Chemom., **22**, 601 (2008)

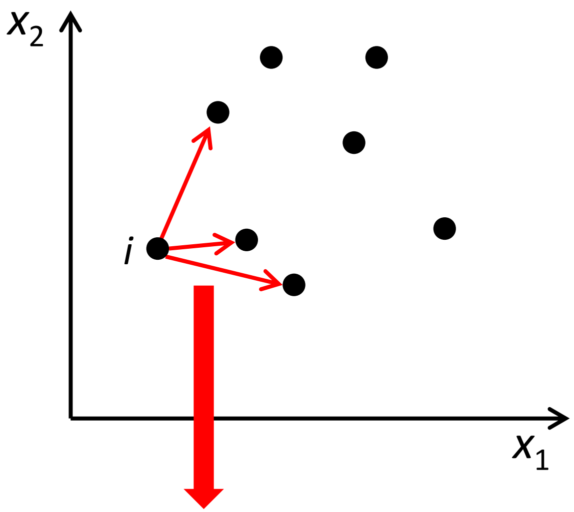
☐ Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA (2009)

Recognition of potentially relevant spectra (TOF-SIMS)

One-class classification

2

Mean KNN distances within the *Off-grain spectra*



Target class (off-grain);
n objects ●

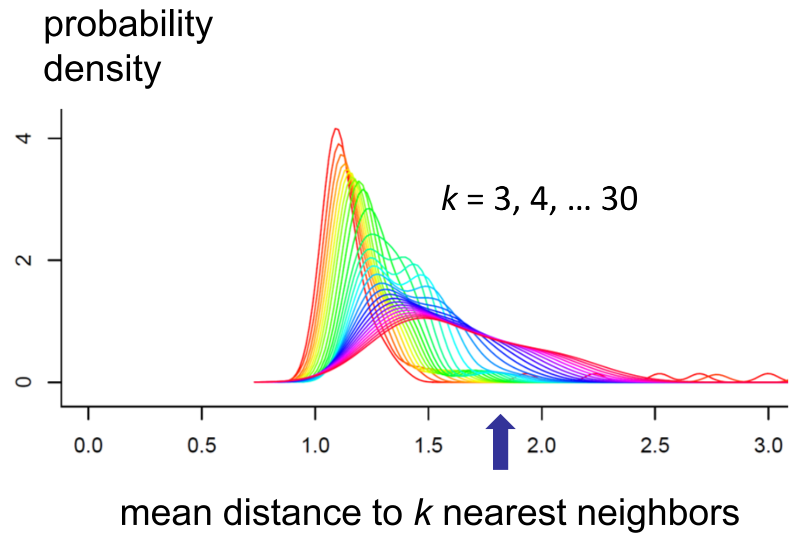
E. g., consider k = 3 neighbors
(k to be optimized)

(1) mean distance to k nearest neighbors
of object i

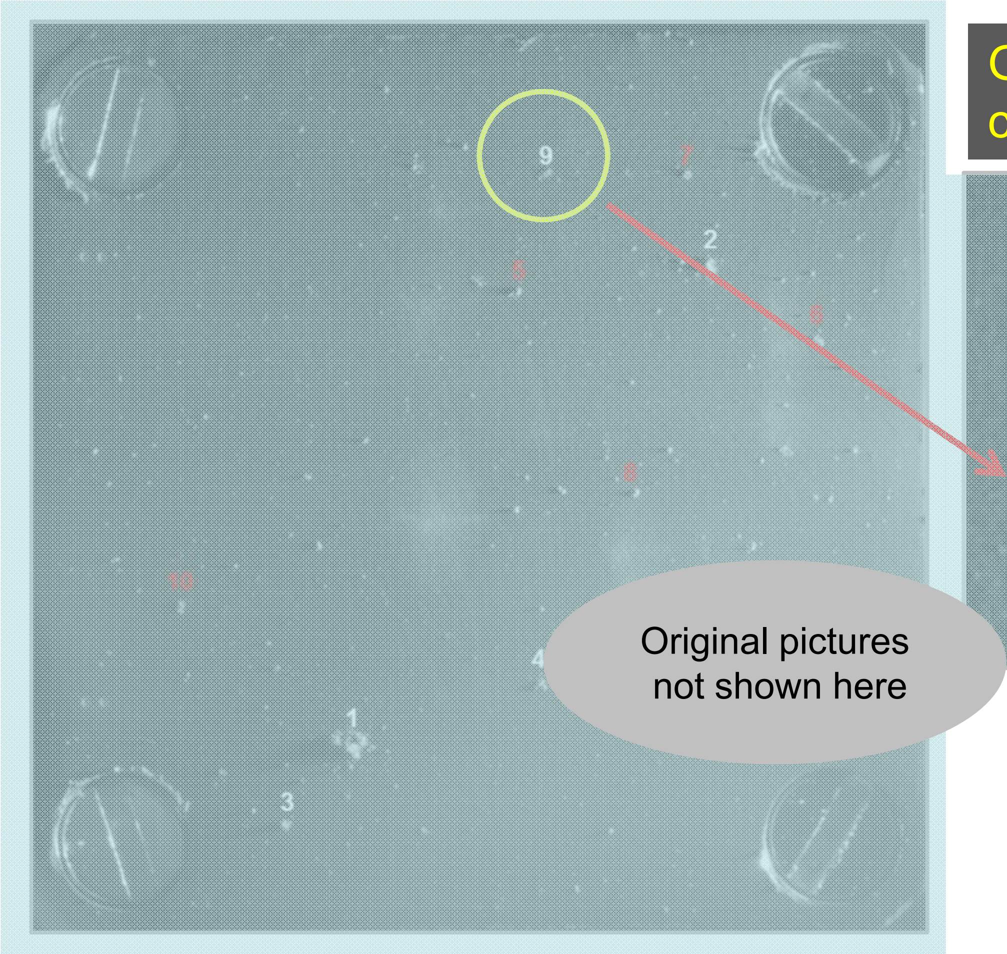
(2) For all objects of *target* class, $i = 1 \dots n$

(3) Distribution of the mean distances

(4) **Cutoff value** (quantile 0.99)

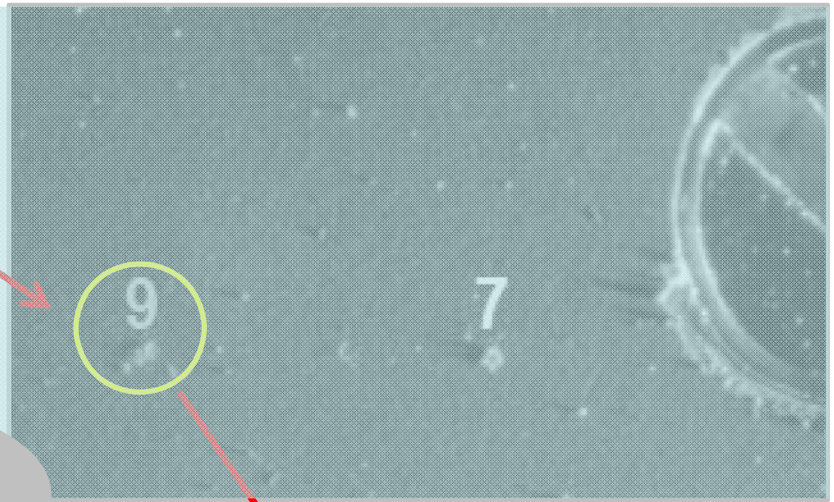


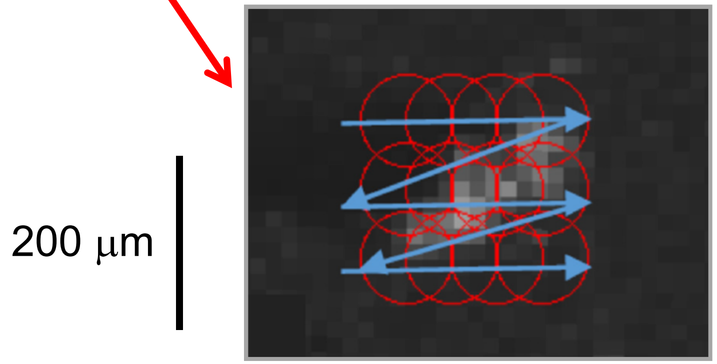
COSIMA: Au target with collected comet particles (grains)



Target *2D0* (Au black), [1 cm x 1 cm],
collection Aug – Dec 2014 (94 days),
ca 3 AU from sun; 50-100 km from comet.
[Yves Langevin (Paris, Orsay)]

Grain *D0NIA* (ca 200 μm diameter),
collected 18-24 Aug 2014.



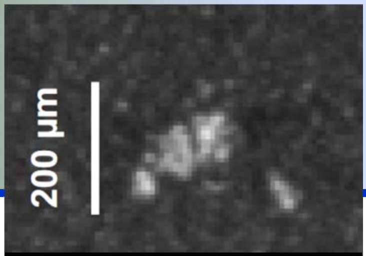


4 x 3 TOF-SIMS spectra scanned
(7 Sep 2014, 12:14 - 13:22)

Recognition of potentially relevant spectra (TOF-SIMS)

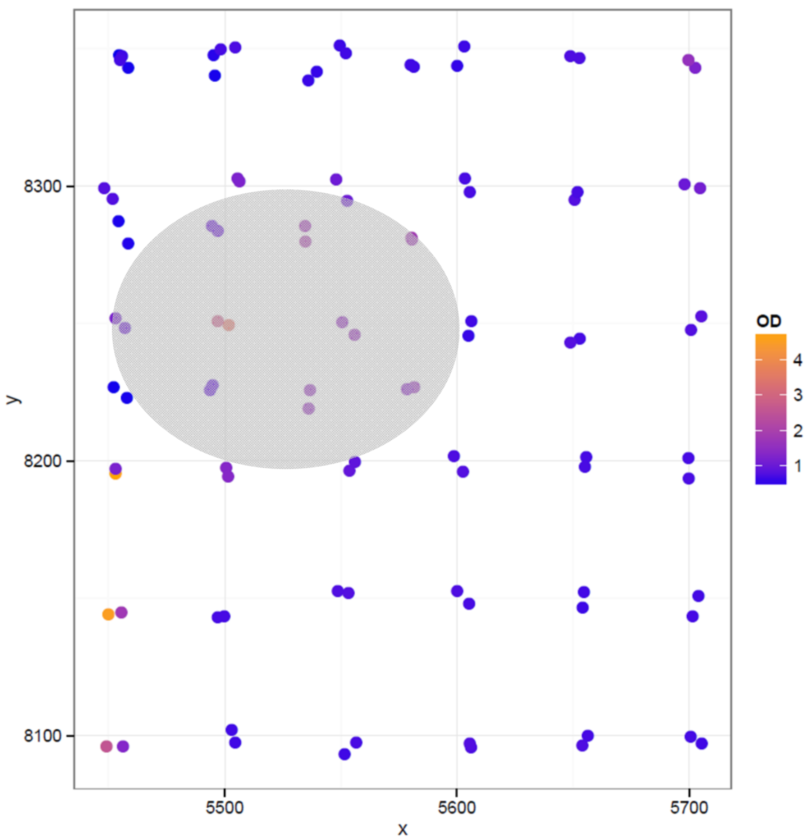
One-class classification

OD (Orthogonal Distance), KNN mean distance

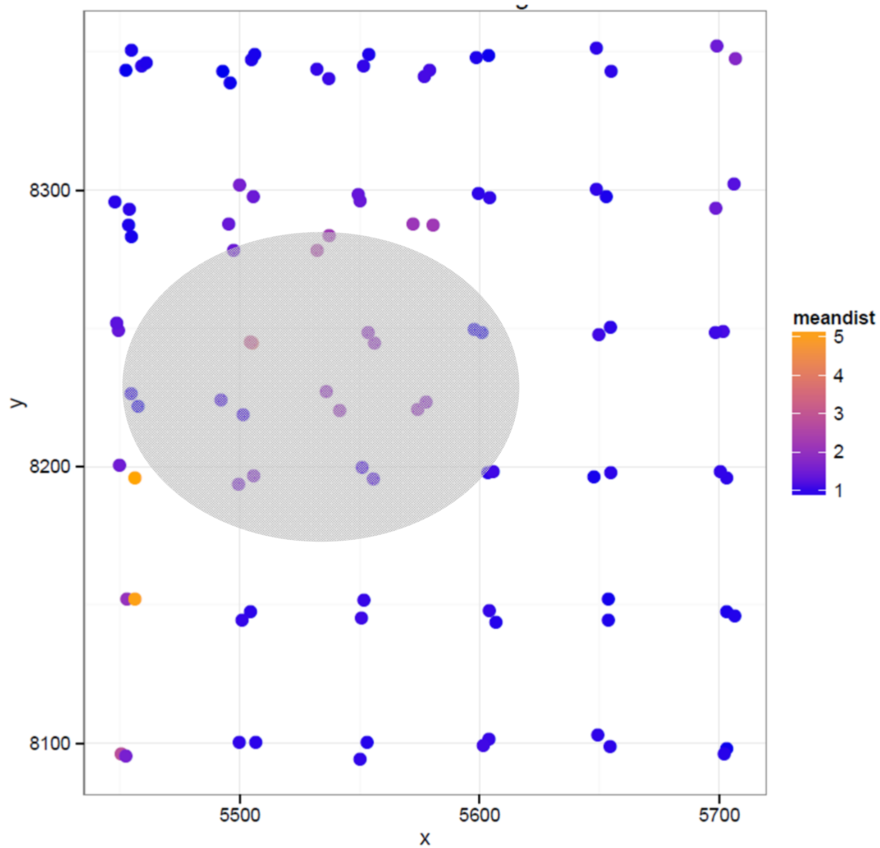


Particle DONIA, Aug 2014

OD



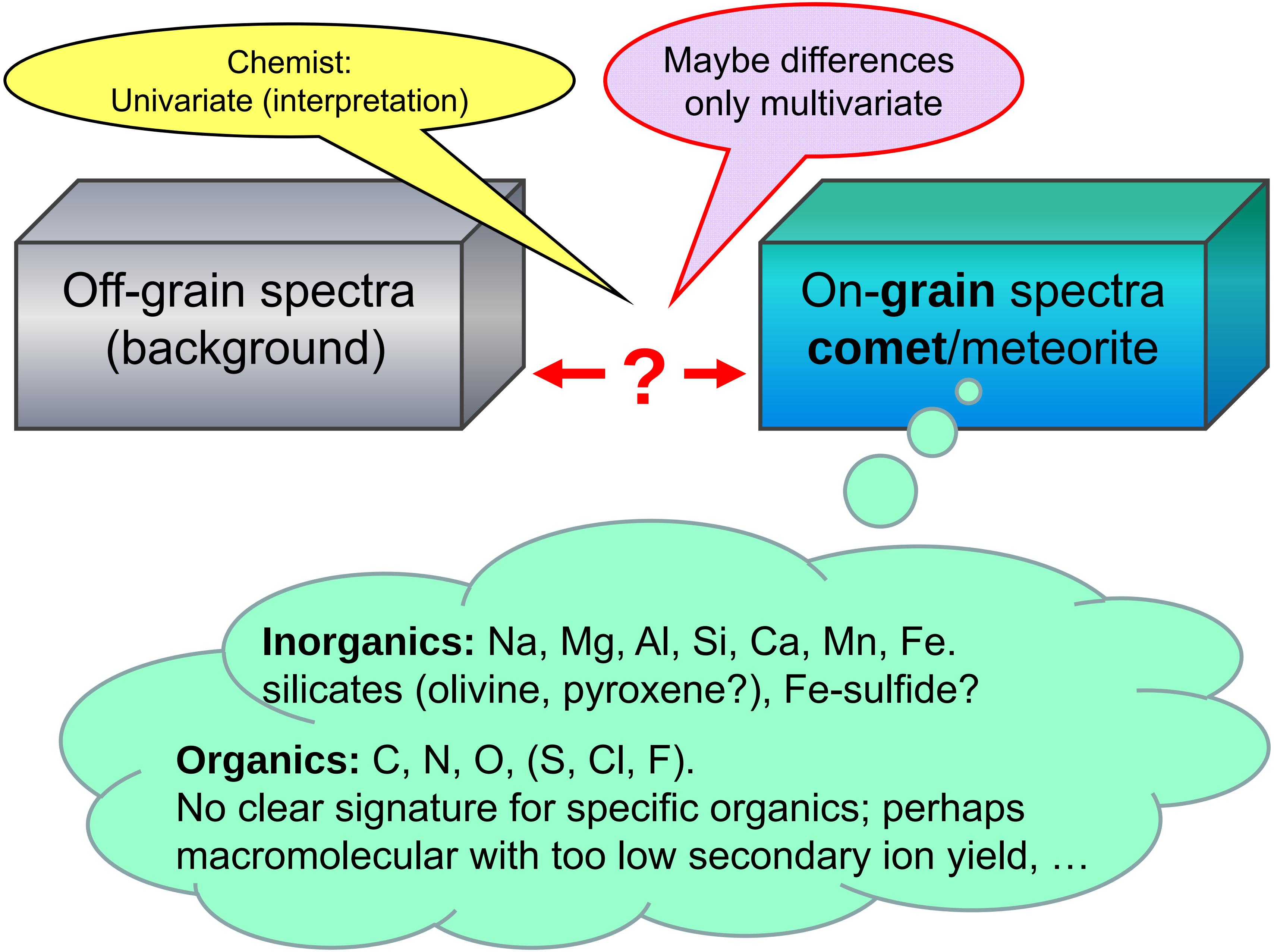
KNN (k = 10)



14 PC components

Off-grain (*target* class), $n1$ = 59; **Query spectra, $n2$ = 96** (plot); m = 3437 variables (inorganic ions, sum 100)

COSIMA: TOF-SIMS Mass Spectra of Comet Particles



Rosetta: Mass Spectra of Gas Phase

GC-MS instrument *ROSINA* (Orbiter)

N_2, O_2

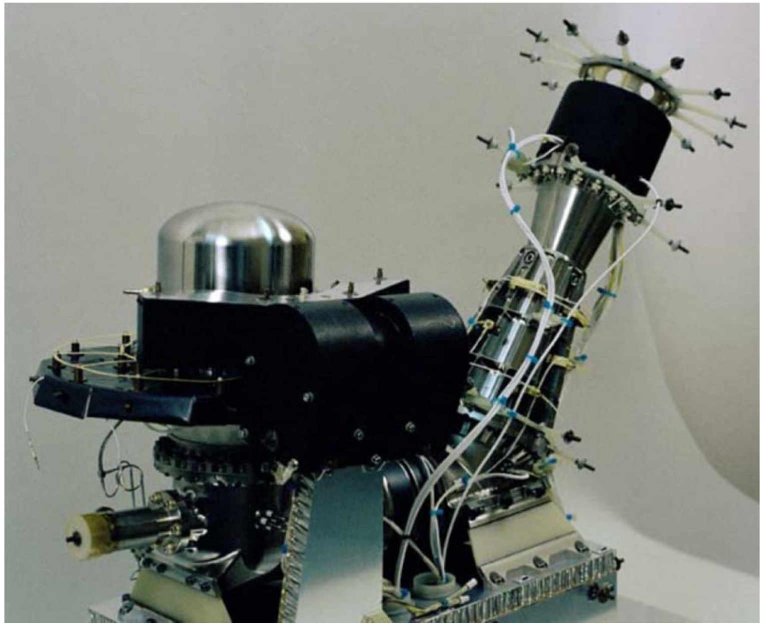
H_2O

CO, CO_2, CH_4

CH_3OH, CH_2O

NH_3, HCN

H_2S, CS_2, SO_2



Balsiger H. et al.:
Space Sci. Rev. 128, 745-801 (2007).
ROSINA - Rosetta Orbiter Spectrometer for Ion and Neutral Analysis.
$m/\Delta m$ 9000 (50% peak height);
m/z 12-150; double-focusing magnet ms.

Altwegg K. et al.: Nature 2014, 2015; University Bern

Rosetta: Mass Spectra of Gas Phase

GC-MS instrument *COSAC* (*Philae* Lander)

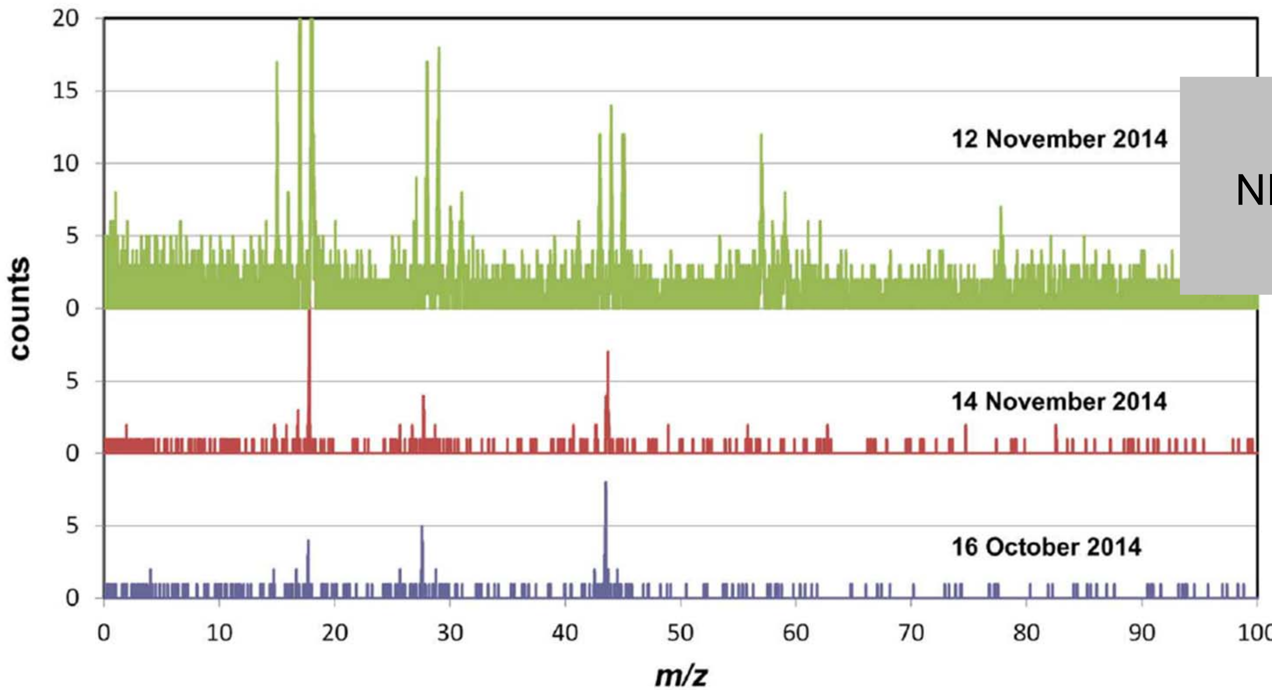


Fig. 1. Mass spectra taken by COSAC in "sniffing mode." Top (green): spectrum taken 25 min after first touchdown; the m/z 18 peak reached a height of 330 counts, but the spectrum is truncated to show smaller peaks more clearly; middle (red): final spectrum, taken 2 days later at the current Philae position; bottom (blue): first spectrum, obtained in orbit 27 days before landing, from a distance of 10 km.

„fit", based on NIST mass spectra database

Goessmann F. et al.: Science, 349, issue 6247 (2015); MPS Göttingen

Rosetta: Mass Spectra of Gas Phase

GC-MS instrument *COSAC* (*Philae* Lander)

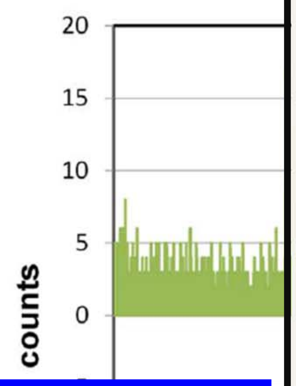


Table 1. The 16 molecules used to fit the COSAC mass spectrum.

Name	Formula	Molar mass (u)	MS fraction	Relative to water
Water	H_2O	18	80.92	100
Methane	CH_4	16	0.70	0.5
Methanenitrile (hydrogen cyanide)	HCN	27	1.06	0.9
Carbon monoxide	CO	28	1.09	1.2
Methylamine	CH_3NH_2	31	1.19	0.6
Ethanenitrile (acetonitrile)	CH_3CN	41	0.55	0.3
Isocyanic acid	HNCO	43	0.47	0.3
Ethanal (acetaldehyde)	CH_3CHO	44	1.01	0.5
Methanamide (formamide)	$HCONH_2$	45	3.73	1.8
Ethylamine	$C_2H_5NH_2$	45	0.72	0.3
Isocyanomethane (methyl isocyanate)	CH_3NCO	57	3.13	1.3
Propanone (acetone)	CH_3COCH_3	58	1.02	0.3
Propanal (propionaldehyde)	C_2H_5CHO	58	0.44	0.1
Ethanamide (acetamide)	CH_3CONH_2	59	2.20	0.7
2-Hydroxyethanal (glycolaldehyde)	CH_2OHCHO	60	0.98	0.4
1,2-Ethanediol (ethylene glycol)	$CH_2(OH)CH_2(OH)$	62	0.79	0.2

91 isomers
(MOLGEN)

201 ion formulae
$C_cH_hN_nO_o$
(m/z 1 – 62),
potential fragment ions

?

Goessmann F. et al.: Science, 349, issue 6247 (2015); MPS Göttingen

Organics in Extraterrestrial Material

Fall 28 Sep 1969
Carbon-rich chondrite, total ca 100 kg,
considered to be similar to comet material

High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed $\boxed{\text{40 years}}$ after its fall

30 mg freshly broken meteorite sample

extraction by methanol, acetonitril, toluene, ...

Fourier transform ion cyclotron resonance mass spectrometer (FTICR-MS);
electrospray ionization (ESI),
only quasi molecular ions $(M+H)^+$, $(M-H)^-$
$m/\Delta m$ ca 10^6

→

Peak list mass spectra

Suitable for highly complex organic mixtures (one peak per brutto formula)

Schmitt-Kopplin P. et al. (Helmholtz-Zentrum, Munich): PNAS **107**, 2763 (2010)

Organics in Extraterrestrial Material

Identified molecular formulas
mass range 150 – 1000

ESI(-), methanol extract

CHO	2,022
CHOS	3,340
CHNO	4,021
CHNOS	4,814
Sum	14,197
Mass peaks	>150,000

Typical carbon-rich meteorites contain 3 – 5 % C.

70 % macromolecular
30 % soluble

→ Total estimated >50,000

Millions of different chemical compounds (isomers) with elements C, H, O, N, S, P.

Extraterrestrial chemodiversity very high

Schmitt-Kopplin P. et al. (Helmholtz-Zentrum, Munich): PNAS **107**, 2763 (2010)

Selected Aspects of 40 Years Applied Chemometrics

Everything should be made ***as simple as possible****,*
but not simpler.

Data in chemistry, …

Sometimes CHEMOMETRICS helps,
but not always.

Thank you for your interest

Autumn School of Chemoinformatics
25 - 26 Nov 2015, Tokyo, Japan
26 Nov 2015, The University of Tokyo