Multivariate analysis (chemometrics) - quality of multivariate calibration

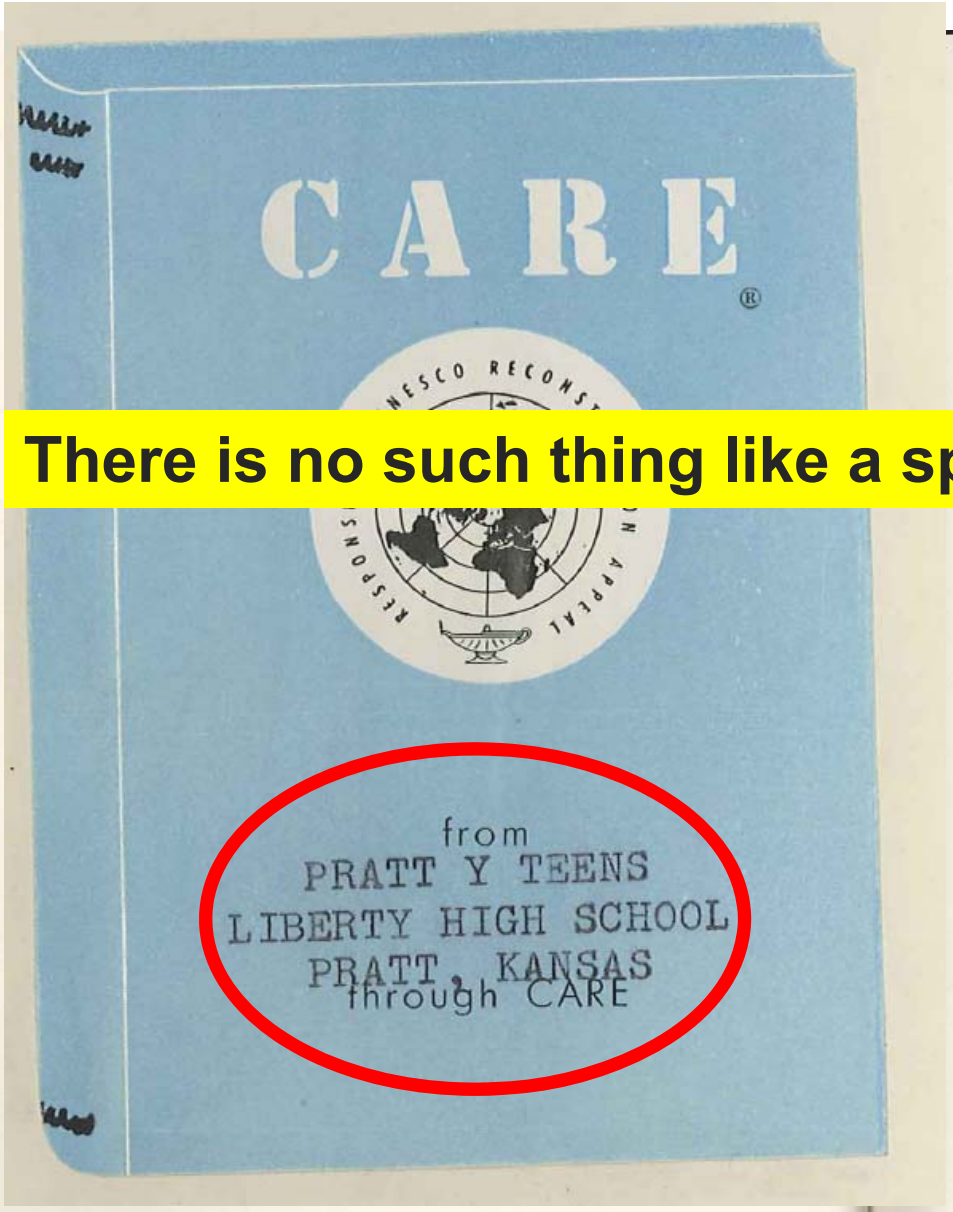
Wolfhard Wegscheider

with contributions by **Alessandra Rachetti**

15 May 2018

Outline

- Historical reminescence: how it all started
- Univariate – multivariate calibration
- Selectivity – specificity – solvability
- Current applications (analytes, signals, matrices, purpose)
- Bias vs. variance in calibration
- Predictive, parsimonious, explanatory models
- Uncertainty, control charts and traceability
- Conclusions



There is no such thing like a specific analytical method

CHEMISTRY OF SPECIFIC, SELECTIVE AND SENSITIVE REACTIONS

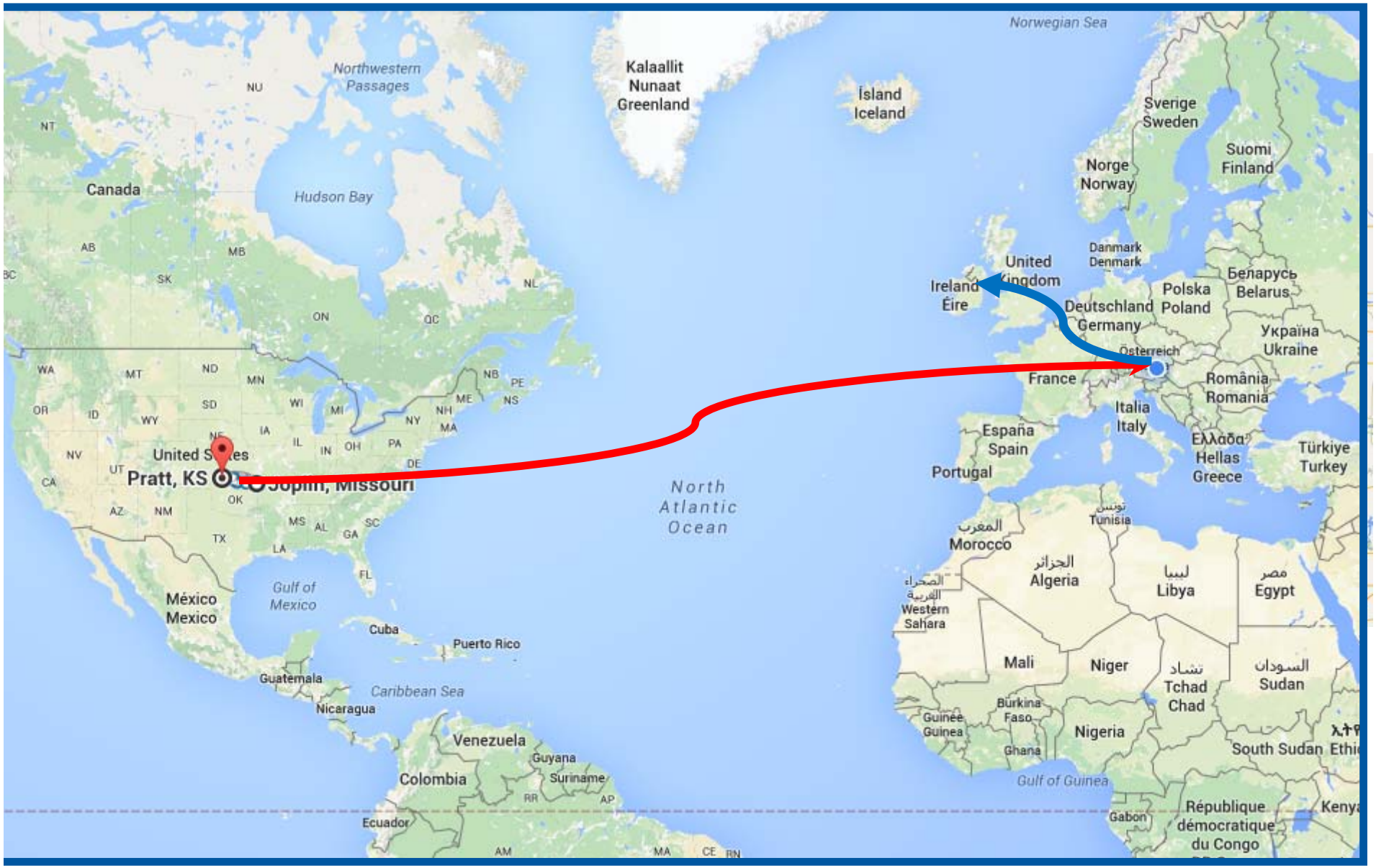
FRITZ FEIGL, Eng., Dr.Sc.
Laboratory of Mineral Products, Ministry of Agriculture, Rio de Janeiro, Brazil
Member of the Brazilian Academy of Sciences
Formerly Professor of Analytical and Inorganic Chemistry at the University of Vienna

TRANSLATED BY

RALPH E. OESPER
Professor of Chemistry, University of Cincinnati, Ohio

1949

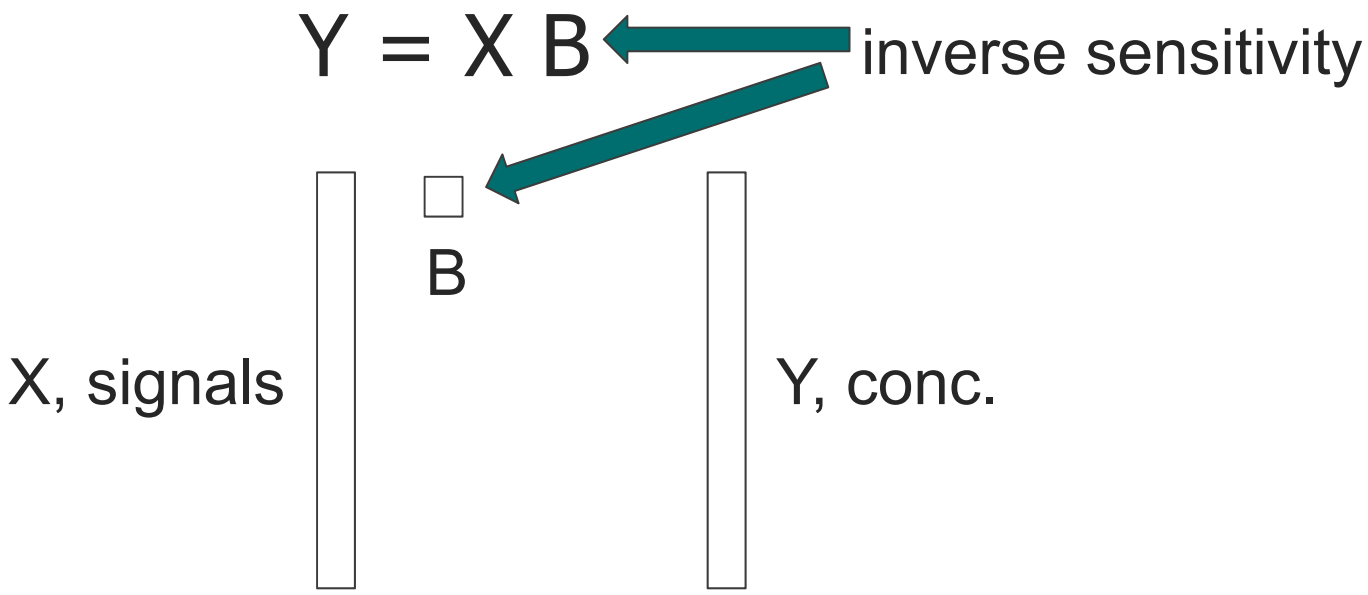
ACADEMIC PRESS INC., PUBLISHERS
NEW YORK, N. Y.


Norwegian Sea
Northwestern Passages
Kalaallit Nunaat Greenland
Ísland Iceland
Sverige Sweden
Suomi Finland
Norge Norway
Canada
Hudson Bay
Danmark Denmark
United Kingdom
Ireland Éire
Беларусь Belarus
Polska Poland
Deutschland Germany
Україна Ukraine
Österreich
France
România Romania
Italia Italy
España Spain
Portugal
Ελλάδα Hellas Greece
Türkiye Turkey
Pratt, KS
Joplin, Missouri
North Atlantic Ocean
تونس Tunisia
المغرب Morocco
الجزائر Algeria
ليبيا Libya
مصر Egypt
الصحراء الغربية Western Sahara
México Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Caribbean Sea
Nicaragua
Mali
Niger
تشاد Tchad Chad
السودان Sudan
Burkina Faso
Guinée Guinea
Nigeria
Ghana
South Sudan
Venezuela
Guyana
Suriname
Colombia
Gulf of Guinea
Ecuador
Gabon
République démocratique du Congo

Why multivariate?

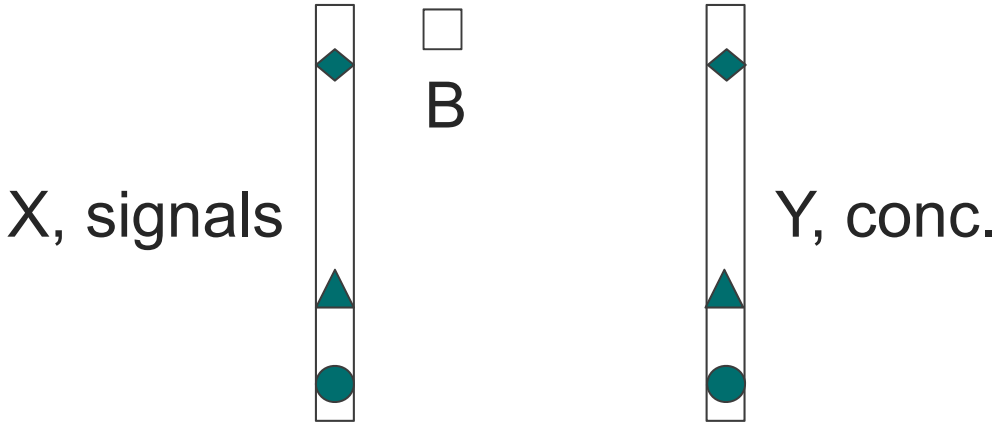
- Non-selective signals
- Many signals (quasi) simultaneously
- Image compression/comprehension/analysis
- More analytes/measurands at the same time
- Faster than classical procedures
- New analytical problems become accessible

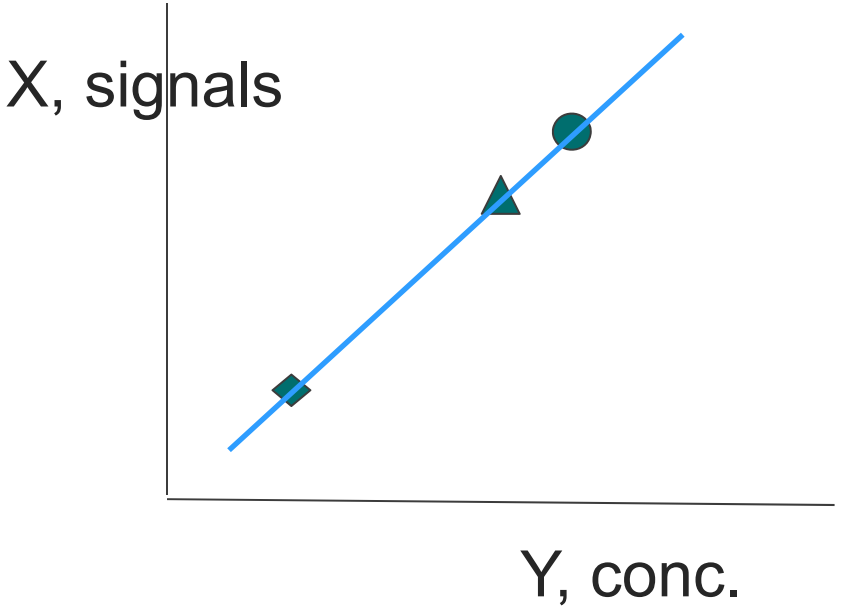
Univariate Calibration



Univariate Calibration

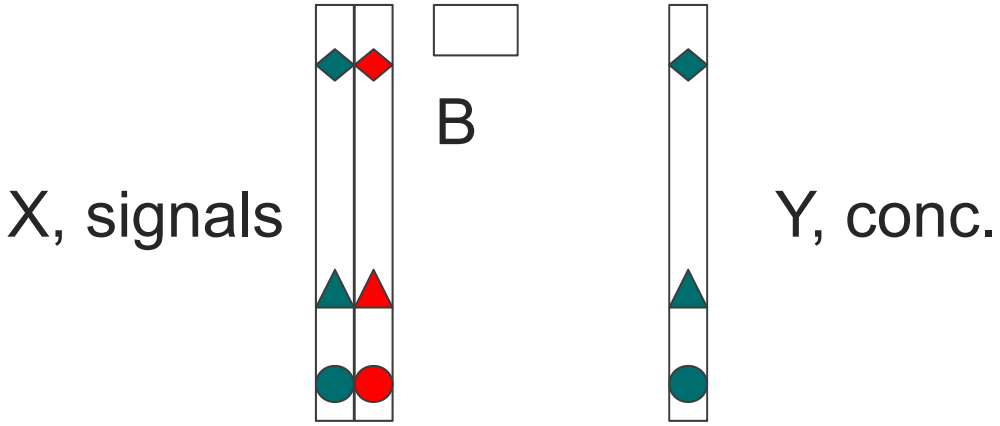
$$Y = X\,B$$

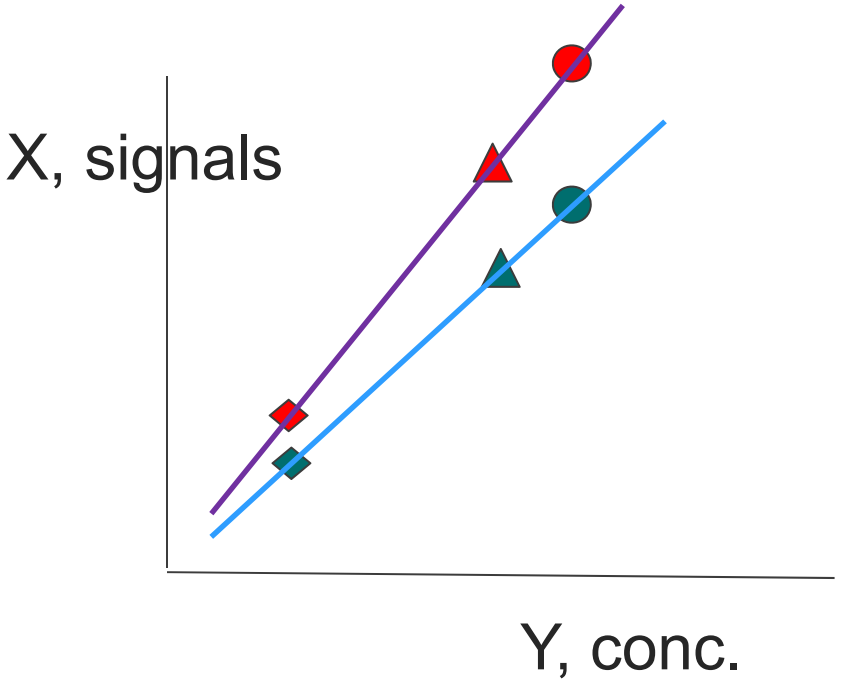




Univariate Calibration
two signals, one measurand

$$Y = X\,B$$

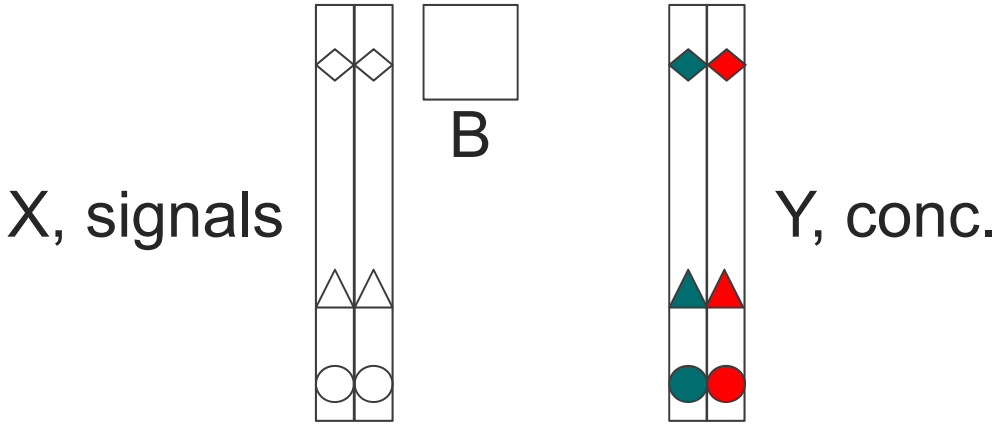


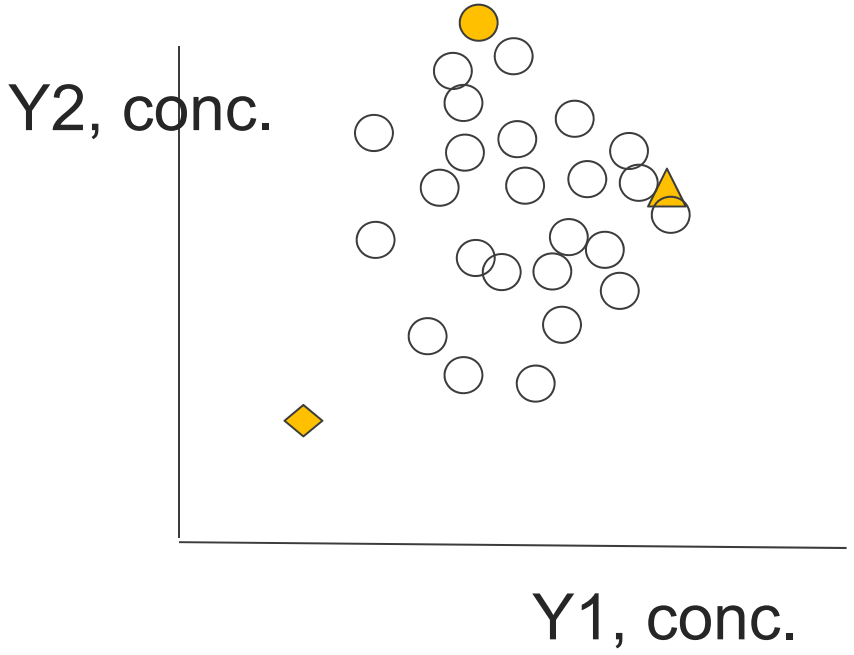


Multivariate Calibration

two signals, two measurands

$$Y = X\,B$$

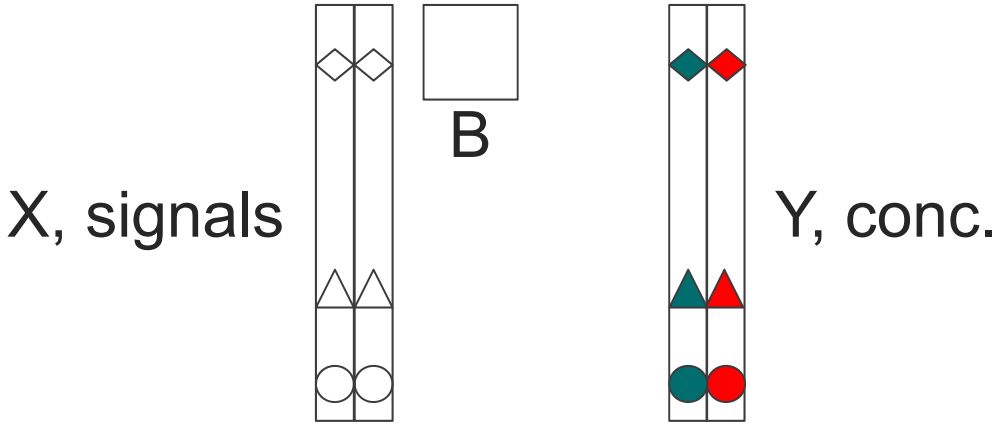


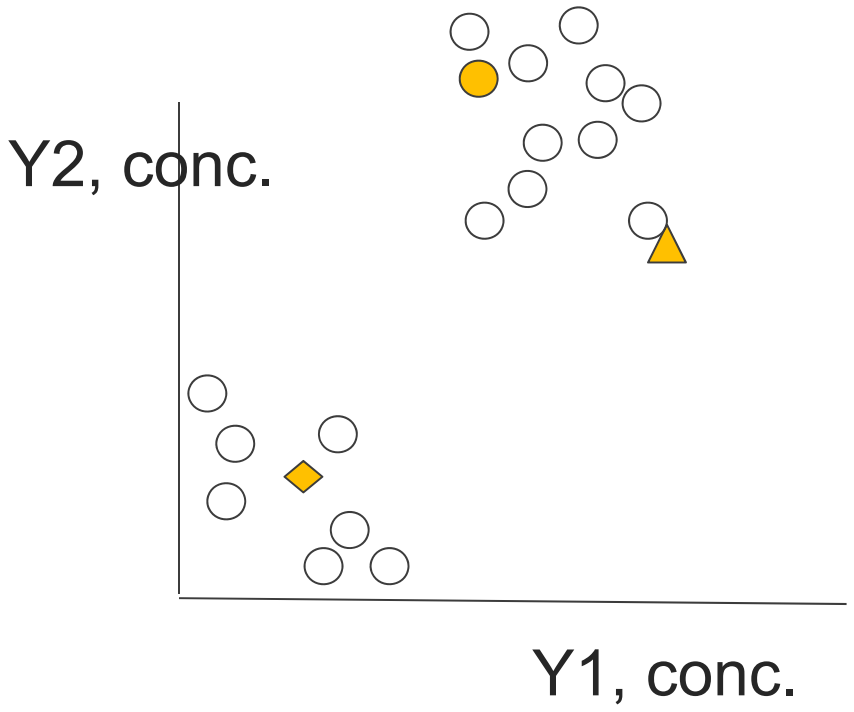


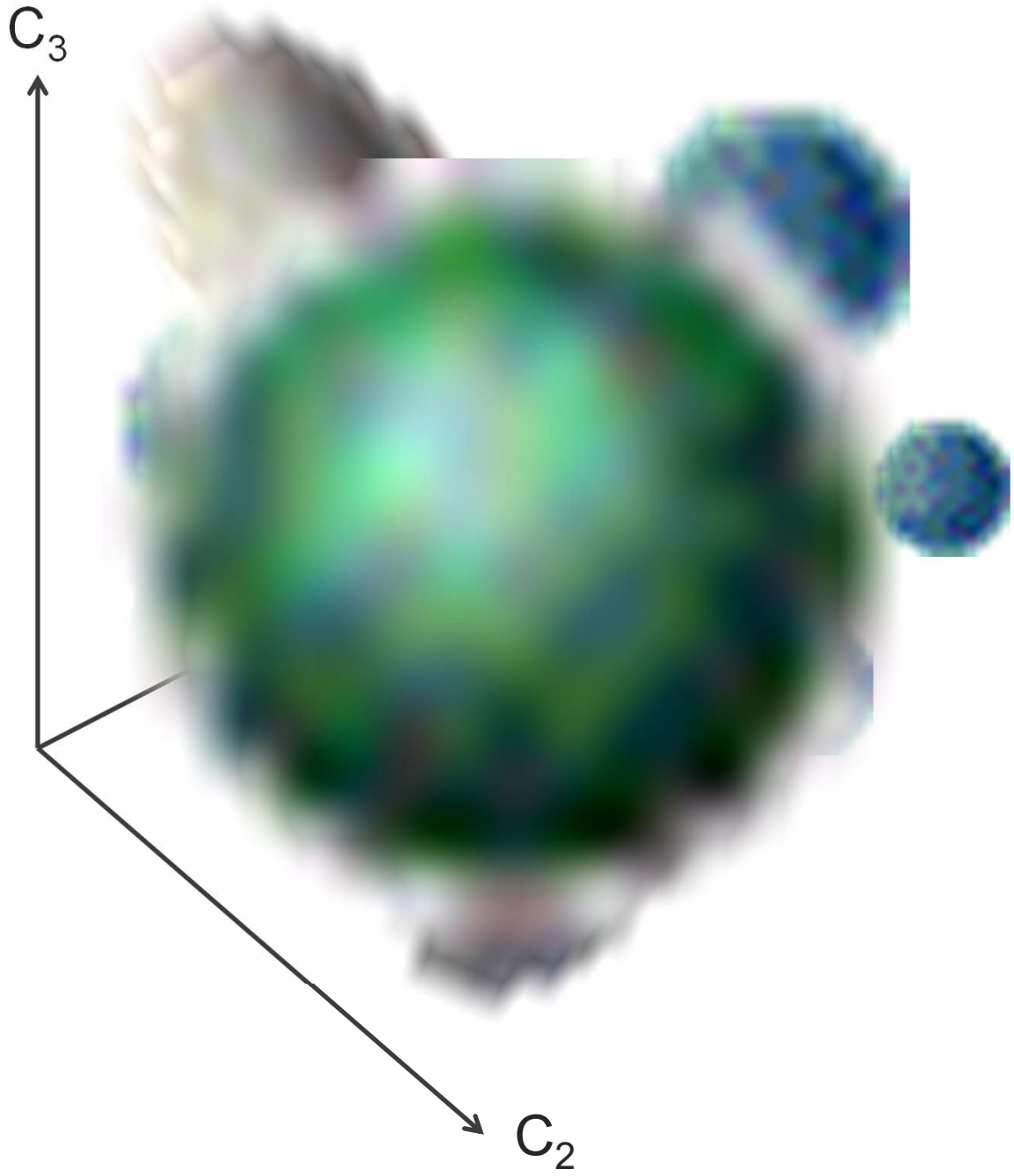
Multivariate Calibration

two signals, two measurands

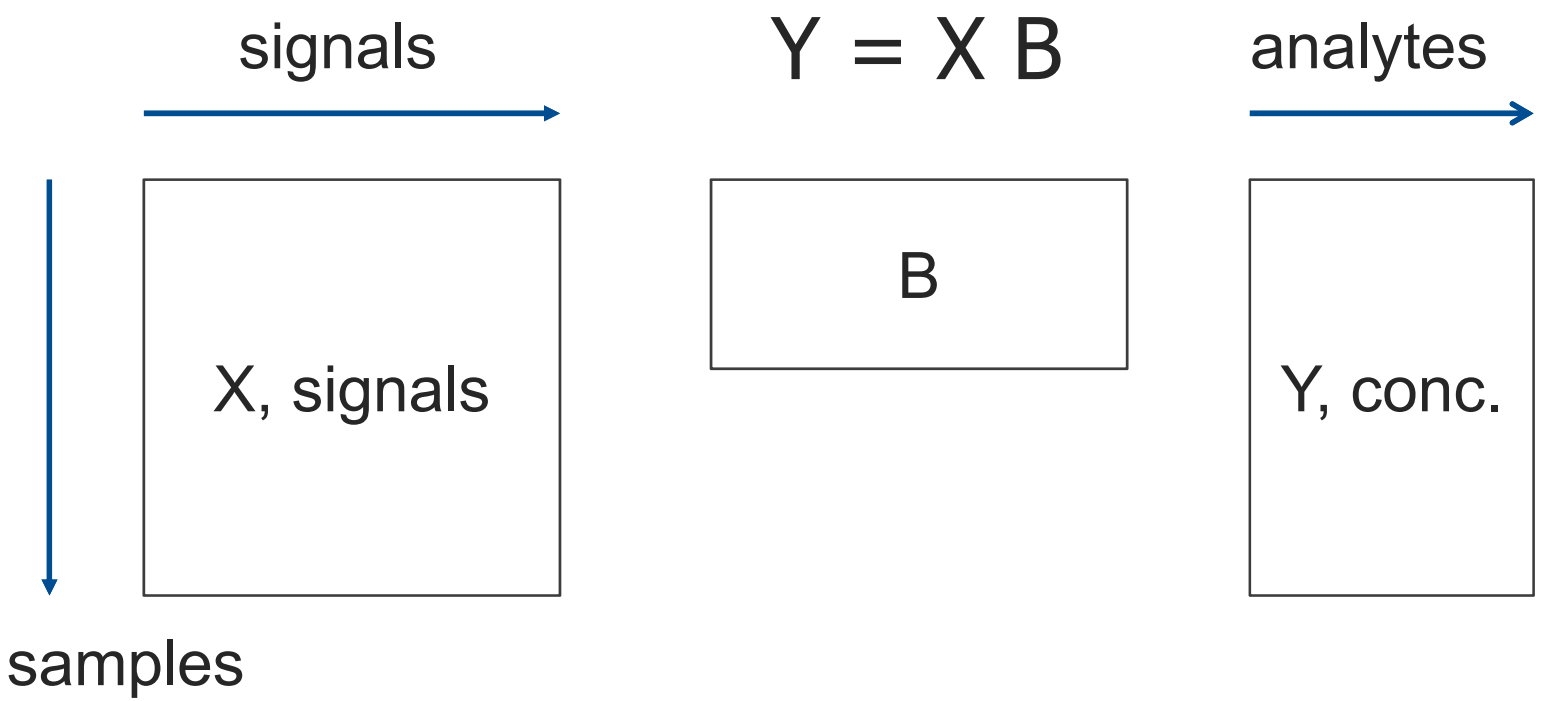
$$Y = X\,B$$



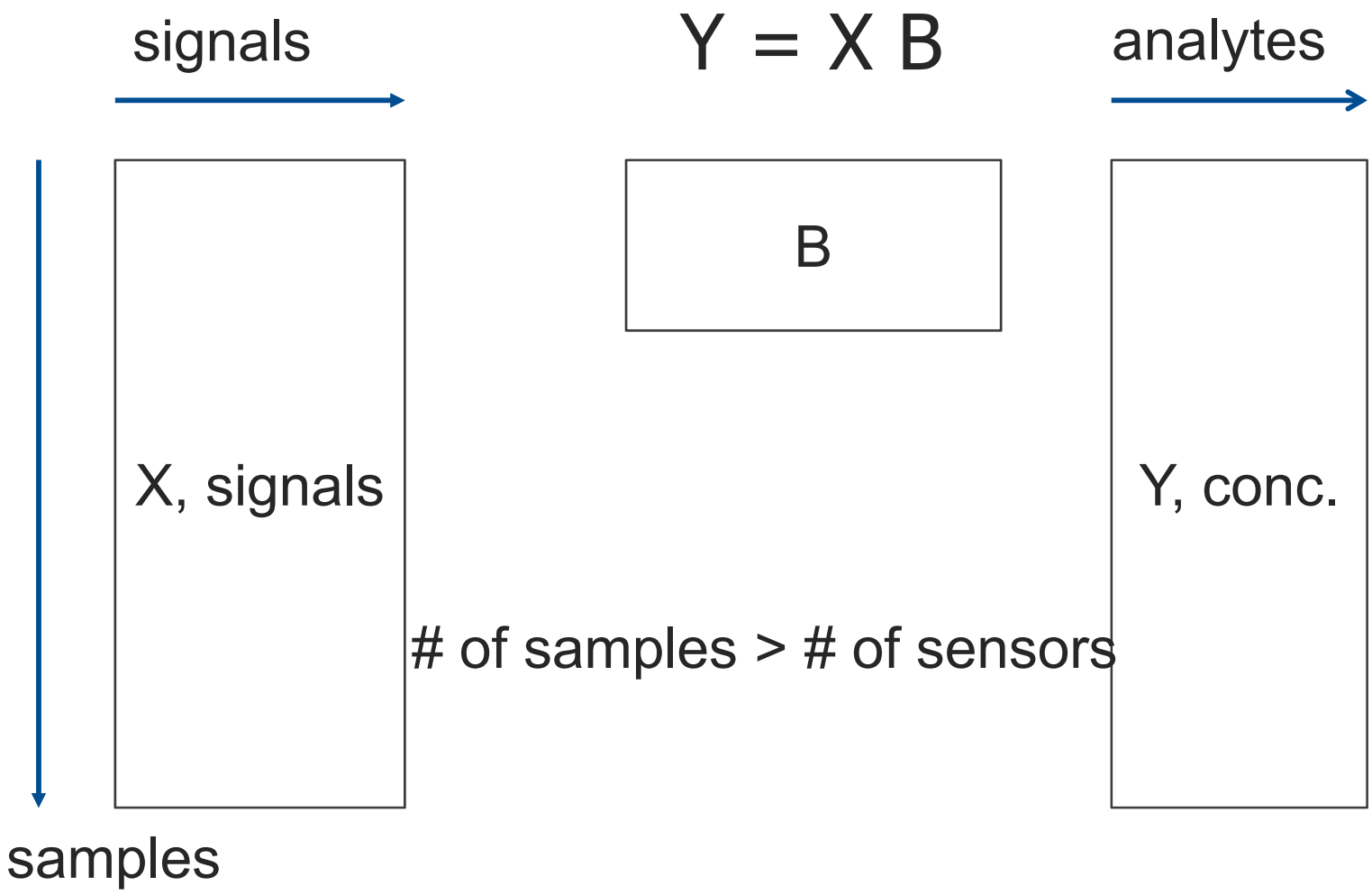




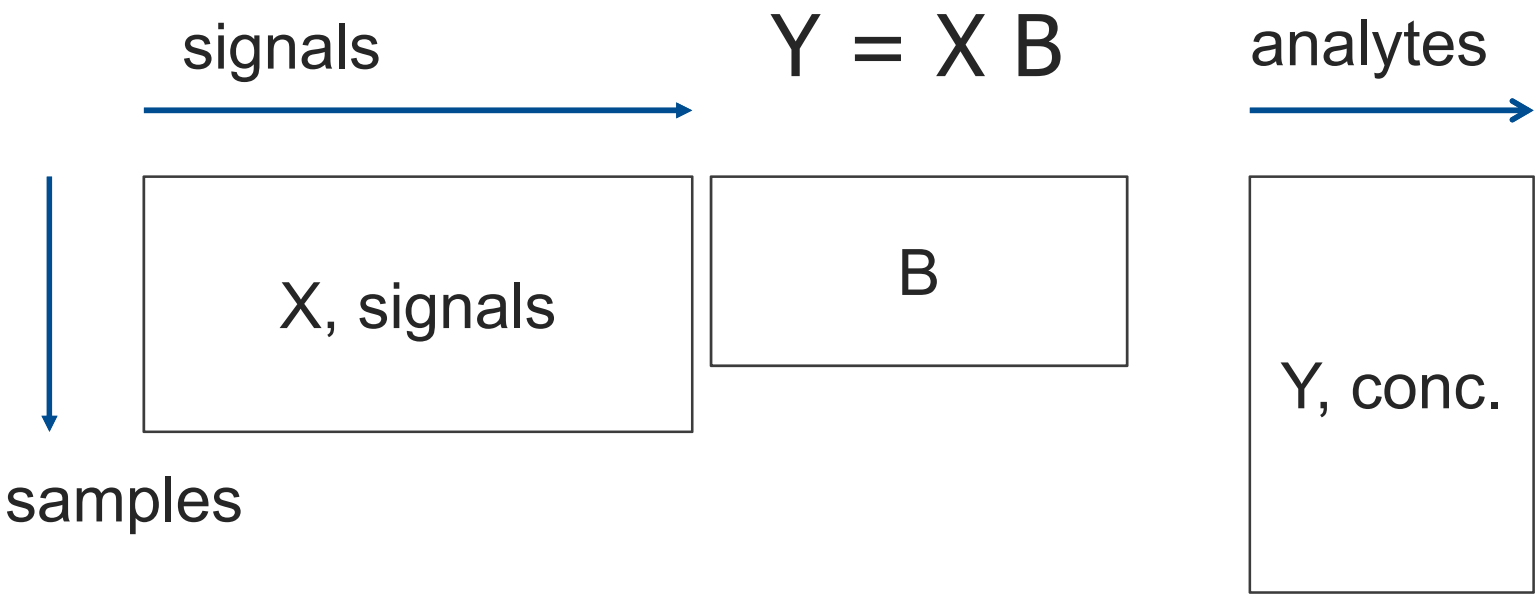
Classical Regression



Classical Regression



~~Classification~~



# of samples < # of sensors

Anal. Chem. 1985, 57, 63–69

Spectrophotometric Multicomponent Analysis Applied to Trace Metal Determinations

namely, in the ultraviolet, visible, and infrared spectral range. Limitations imposed by data reduction schemes based on ordinary multiple regression are shown to be overcome by means of partial least-squares analysis in latent variables.

Quantitative spectrophotometric analysis of mixture components is featured for systems with low spectral selectivity, namely, in the ultraviolet, visible, and infrared spectral range. Limitations imposed by data reduction schemes based on ordinary multiple regression are shown to be overcome by means of partial least-squares analysis in latent variables. The influence of variables such as noise, band separation, band intensity ratios, number of wavelengths, number of components, number of calibration mixtures, time drift, or deviations from Beer's law on the analytical result has been evaluated under a wide range of conditions providing a basis to search for new systems applicable to spectrophotometric multicomponent analysis. The practical utility of the method is demonstrated for simultaneous analysis of copper, nickel, cobalt, iron, and palladium down to 2×10^{-6} M concentrations by use of their diethyldithiocarbamate chelate complexes with relative errors less than 6%.

ordinary MR even in the sense that deviations from linearity or background changes must not be considered explicitly.

Principal component analysis (PCA) is used to model the absorbance matrix obtained from multivariate calibration measurements into principal components that are then fitted to the concentrations of the components by ordinary regression methods.

A newer approach describes additionally the concentration matrix by principal components and relates these components to the principal components of the absorbance matrix. This method known as partial least squares analyses (PLS) (*14*) is preferable to conventional PCA as the information from the calibration solution is better used as it reflects a criterion of the similarity of the sample to the calibration set and as the method is easily implemented on minicomputers (*12*, *14*).

Applications of the PLS method for SMA were described for spectrofluorimetry (*14*) and near-infrared (near-IR) spectroscopy (*15*). The present paper demonstrates the ad-

SINGLE- AND MULTI-CHANNEL DETECTION FOR GENERALIZED QUANTITATIVE ANALYSIS IN CASES OF UNRESOLVED CHROMATOGRAPHIC PEAKS

MATTHIAS OTTO

Department of Chemistry, Bergakademie Freiberg, 9200 Freiberg (German Democratic Republic)

WOLFHARD WEGSCHEIDER* and ERNST P. LANKMAYR

Institute for Analytical Chemistry, Micro- and Radio-chemistry, Technical University of Graz, Technikerstrasse 4, A-8010 Graz (Austria)

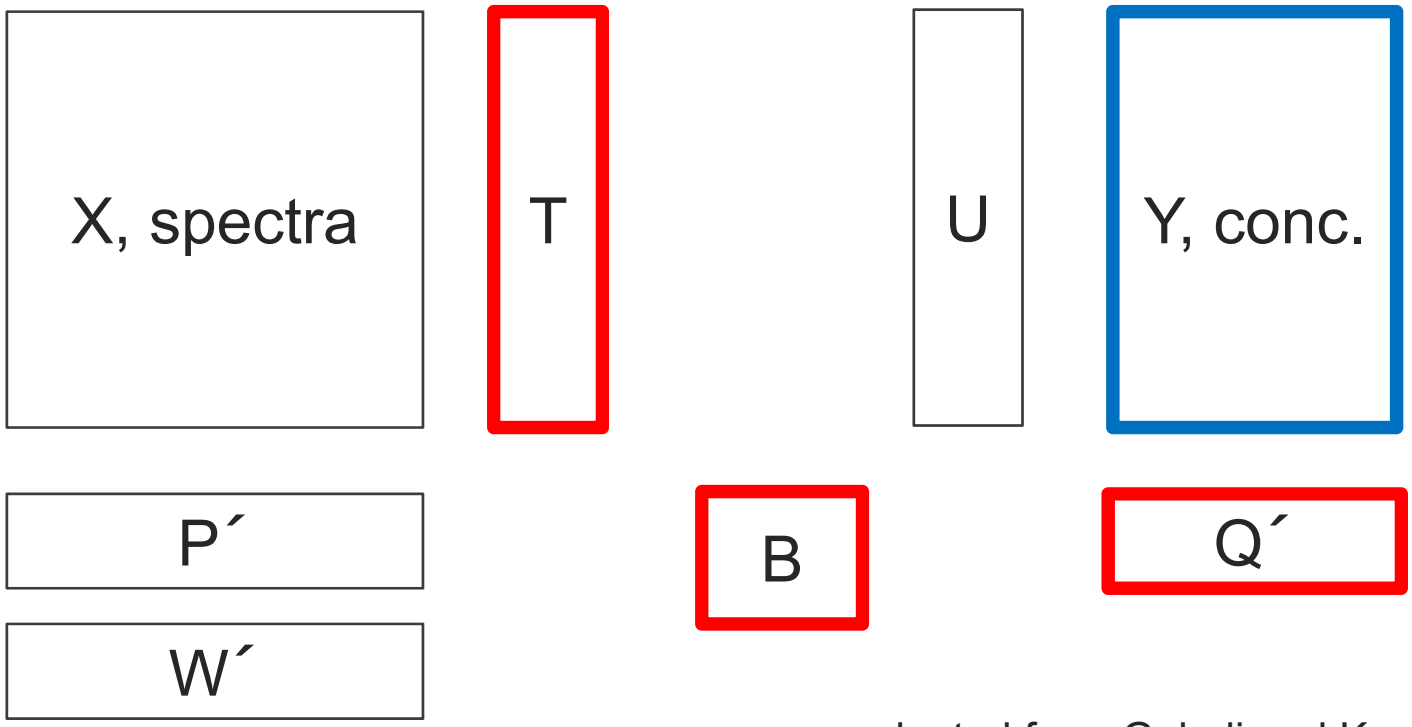
(Received 7th September 1984)

SUMMARY

Computerized quantification of components under overlapping chromatographic peaks is done by calibration of chromatograms against component mixtures. For conventional (single-channel) detectors, the limitations of earlier methods based on ordinary multiple regression, can be circumvented by data reduction with the aid of principal component analysis with the partial least-squares approach. Simulation studies show that the method can be applied even when there is severe peak overlap, unstable baseline, noisy chromatograms or non-linear detector response. Advantages in the quantification of fused peaks by means of multichannel detectors are outlined. Present limitations on the quantitative

Classical Regression vs. PLS (partial least squares)

$$Y = T\,B\,Q'$$



adapted from Geladi and Kowalski, 1986

Purpose	Principle	Measurand (SI ?????)
Adulteration/authenification of coffee	FT-IR & FT-NIR	Glucose, starch, chichory, barley, …
Composition of coffee	FT-NIR	caffeine, theobromine, theophylline, moisture, ash, lipid
Degree of roasting of coffee	FT-NIR	Effect of roasting, prediction of roasting degree
Waste water characterization	UV-vis	Total suspended solids, chemical oxygen demand (COD)
Honey	^{1}H-NMR, HPLC-UV	Origin and phenolic compounds
Olive oil/ adulteration	PTR-MS	Botanical origin
Olive oil	SESI-MS	Geographical origin
Cheese	HS-PTR-MS	Geographical origin
Apples, plums, tomato, mushrooms	Backscattering images	Firmness and elastic modulus
Banana	Backscattering images	Ripeness and chilling injury
Tablets, capsules, liquids, suspensions	Raman	Active pharmaceutical ingredient (API) content

Outline

- Historical reminescence: how it all started
- Univariate – multivariate calibration
- Selectivity – specificity – solvability
- Current applications (analytes, signals, matrices, purpose)
- **Bias vs. variance in calibration**
- Predictive, parsimonious, explanatory models
- Uncertainty, control charts and traceability
- Conclusions

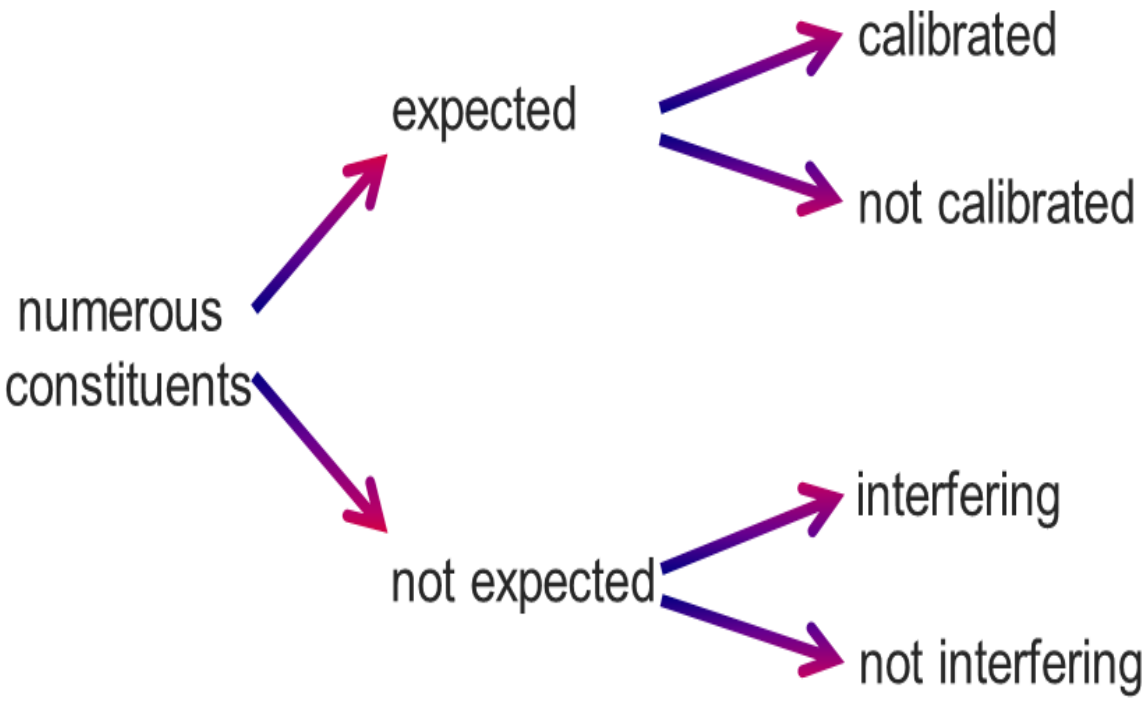
Example: P in various steels by OES

- Trace element
- 699 standard samples
- Multielement procedure
- Signal generation process very complex
- Expansion to non-linear effects
- Linear model from 140 variables
- Variance/bias trade-off gives (only) 13 of those 140 variables

Explain variability of concentrations in terms of signal and matrix

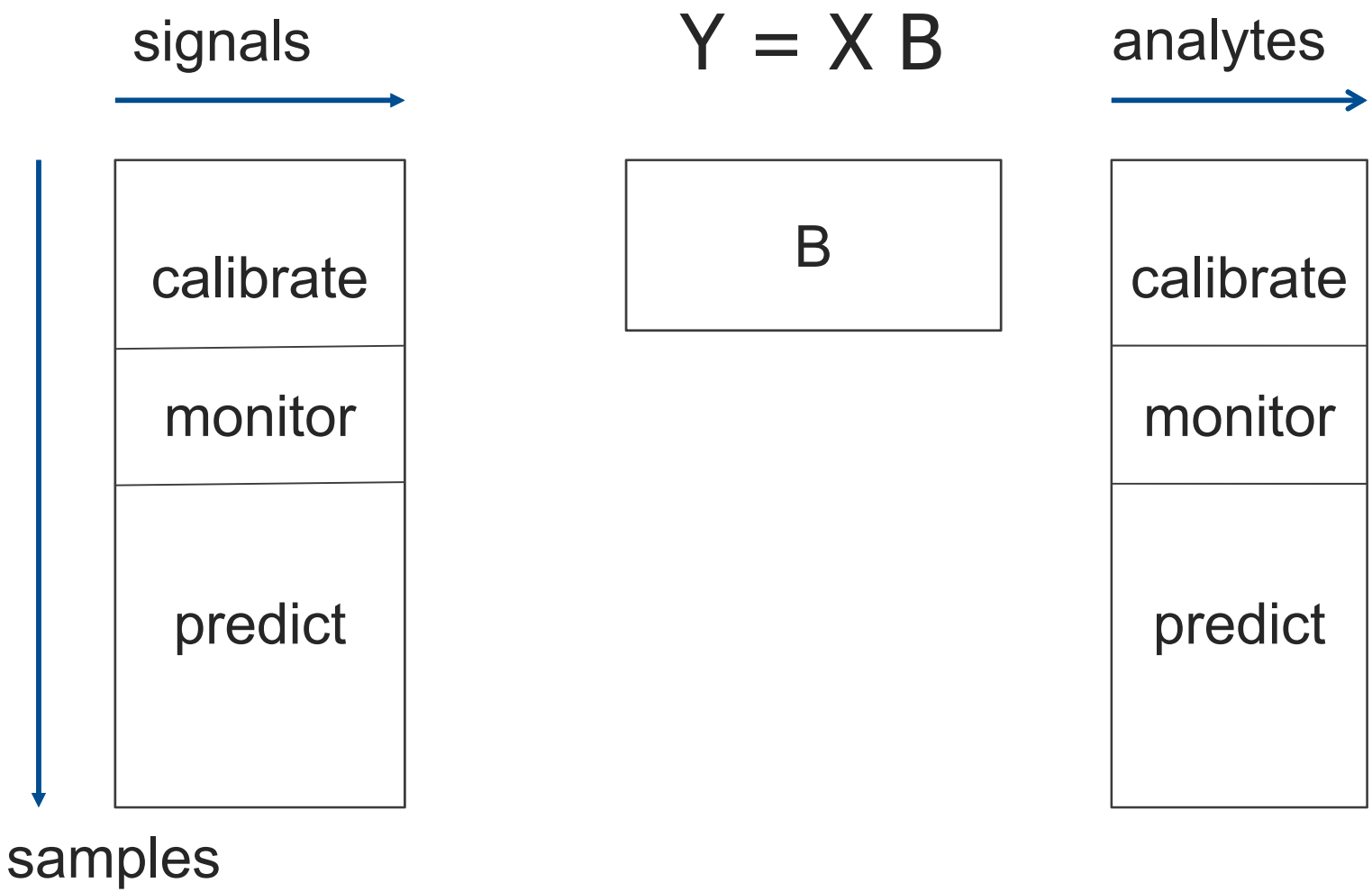
- $c_A = f(I_j, c_i)$
  - Concentration of an analyte is function of (potentially) all sensor signals and (potentially) all constituents in the sample
  - Left side: predicted concentration
  - Right side: „variables“

Consequences
in complex samples



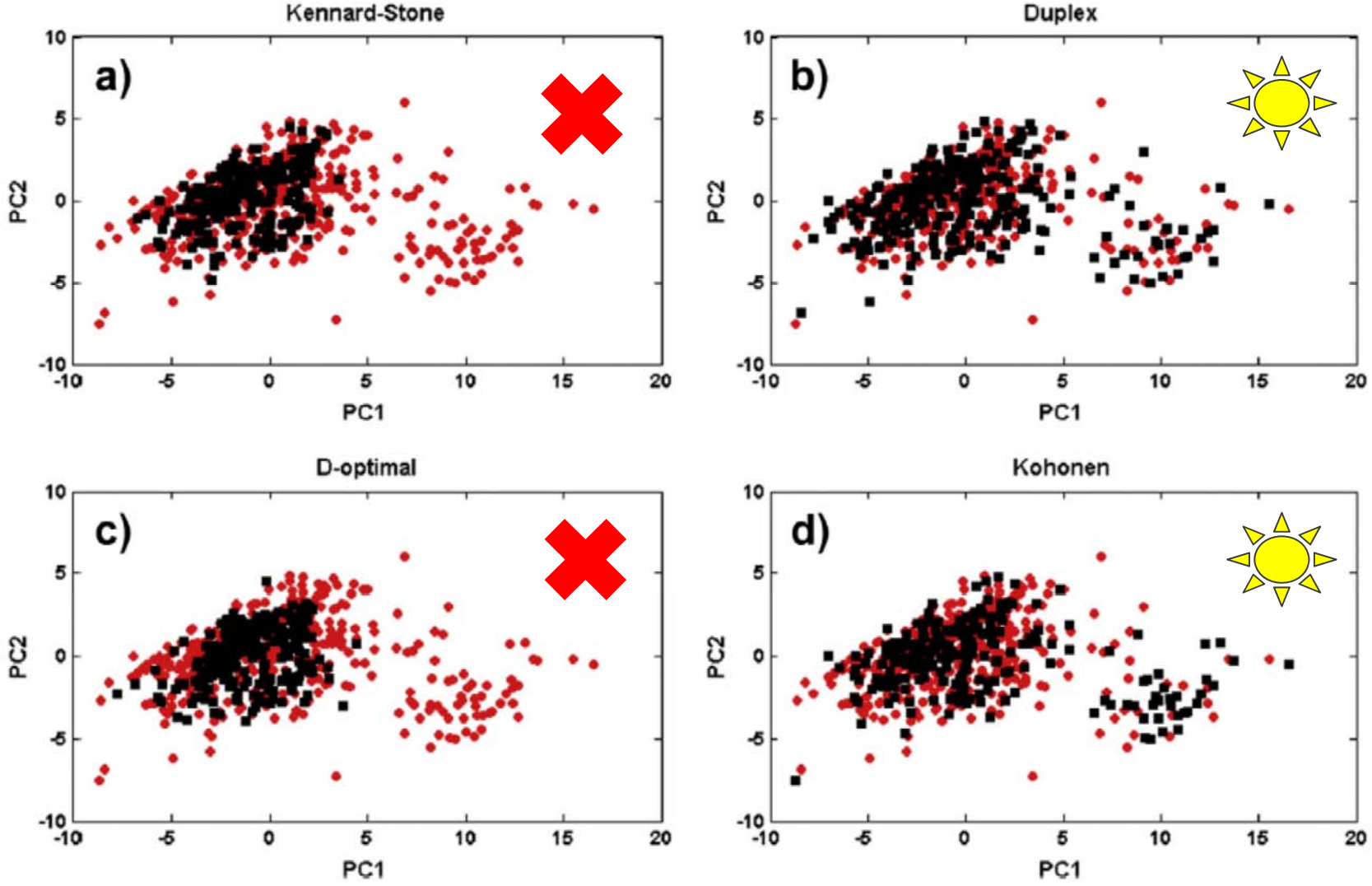
univariate calibration	multivariate calibration
C „sen" p sa e	Multiple „sensors" per sample
	o.k.
Su nt sele ty ?	o.k.
N e e	Effect acceptable ?
o.k.	o.k.

Validation
as a sampling problem



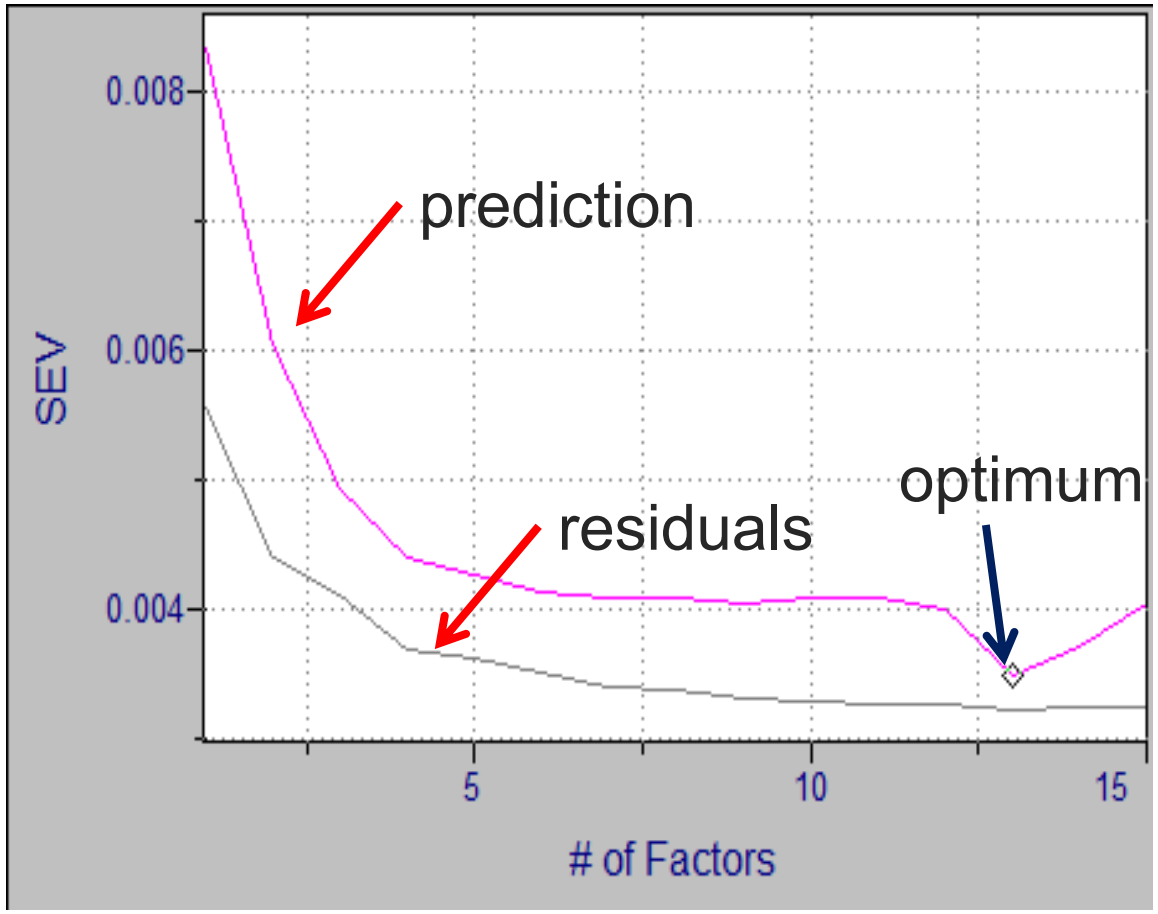
Effect of four selection procedures

Westad & Marini, 2015



Variance/bias trade-off: optimum from cross-validation

- Select some of the standards and predict others
- 13 variables from a total of 140 are left

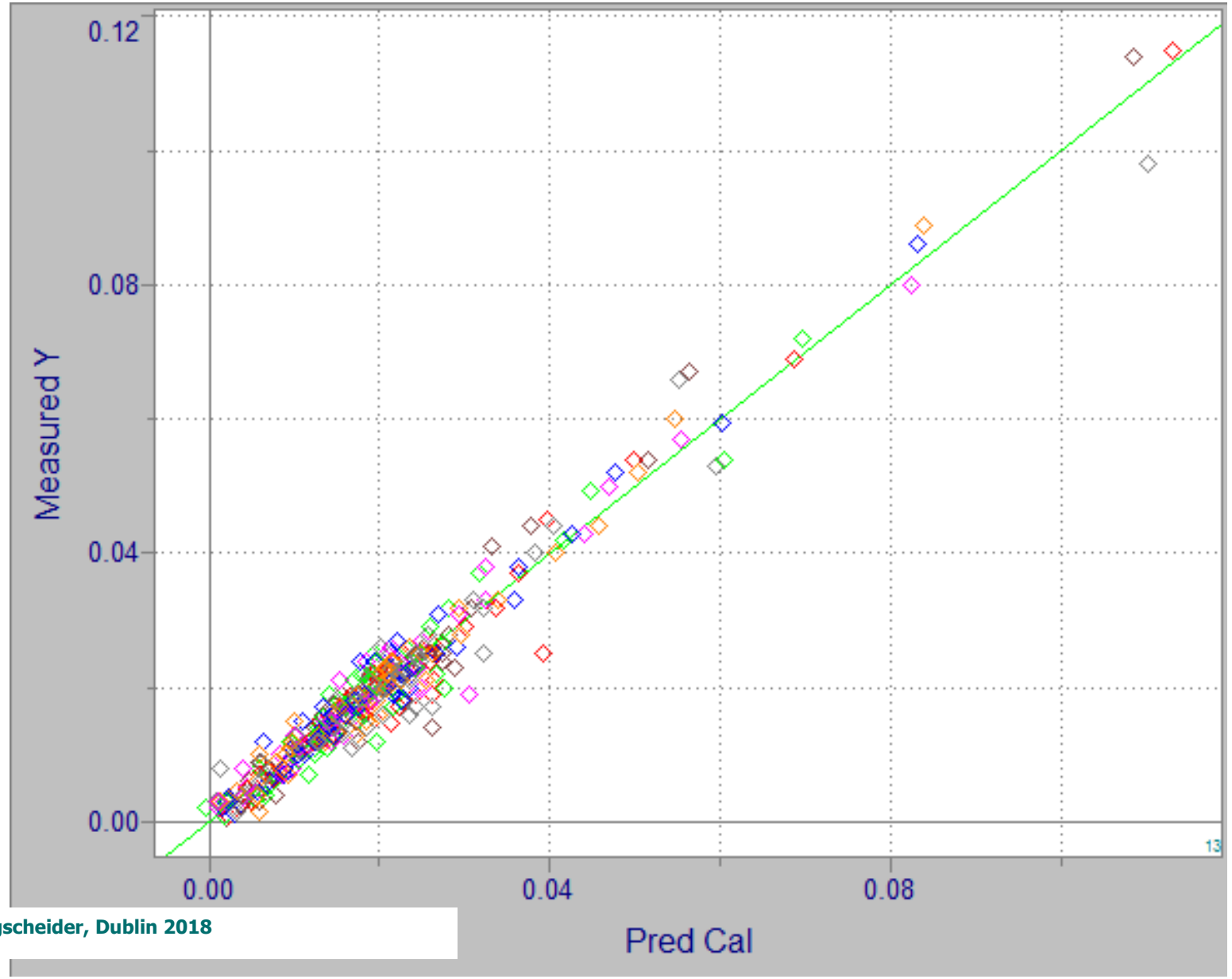


Consequence of bias/variance trade-off:

- In calibration, we cannot distinguish between systematic and random errors

- Tragedy for quality assessment?
- No, we work with a „combined effect“ for measurement uncertainty

Measurement of P in steels



General problems with complex samples

- Require complex calibration (or superb selectivity/separation)
- Little insight
- Lack of intuitivity
- Highly variable results of calibration (but not necessarily on predictions):
  - Size of „coefficients“
  - Variables selected

Outline

- Historical reminescence: how it all started
- Univariate – multivariate calibration
- Selectivity – specificity – solvability
- Current applications (analytes, signals, matrices, purpose)
- Bias vs. variance in calibration
- **Predictive, parsimonious, explanatory models**
- Uncertainty, control charts and traceability
- Conclusions

Calibration is only one step…
…. we must seek understanding

- Predictive
- Parsimonious
- Explanatory
- Beware of lurking variables and spurious correlations
- Consider influential observations

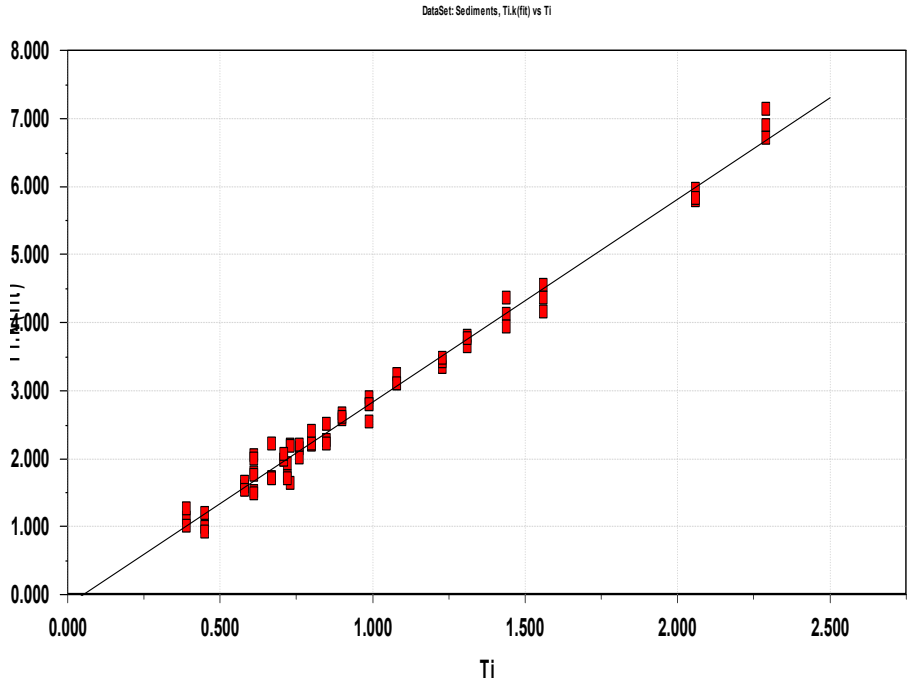
Quality of prediction and indicators

- Quality of prediction: RMSECV …. Root mean squared error of cross validation
- Indicators: which variables are important for prediction
  - VIP: modelling power on x and y
  - X-weight… describes covariance of x and y
  - SR… selectivity ratio: influence on y
  - OR… orthogonal ratio: non-influence on y
  - Reproduced correlation

Model data: XRF on sediments

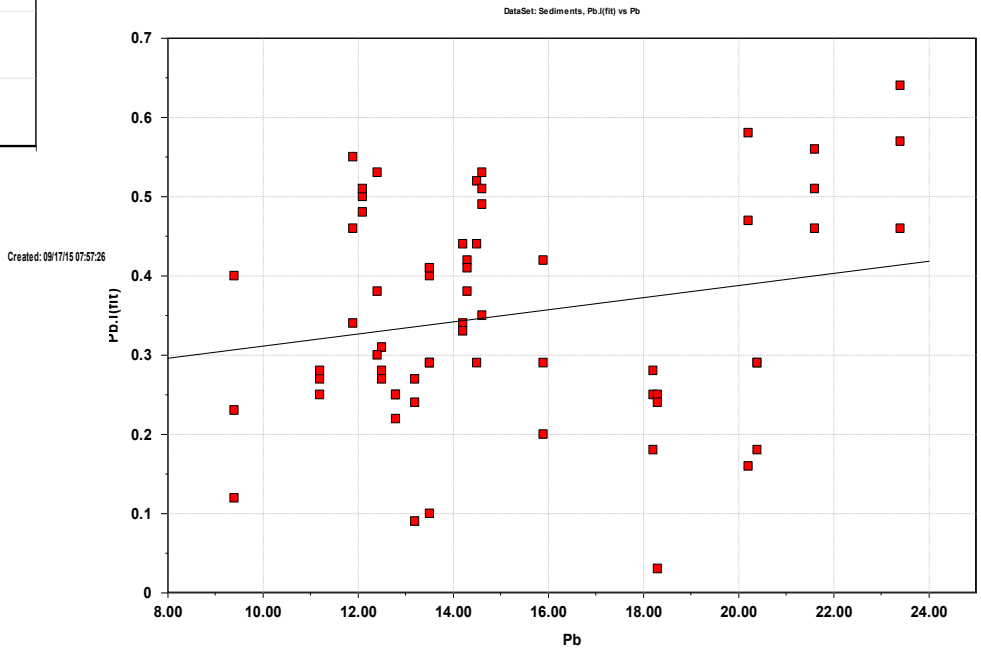
- Advantage: „completely“ understood on basis of first (physical) principles
- Remaining problems: inhomogeneity, mineralogical effects, grain size distribution
- Scientific background is geochemical
- Are there any accounts of the Nb-anomaly?
- COLTAN research

Two model elements: Ti and Pb

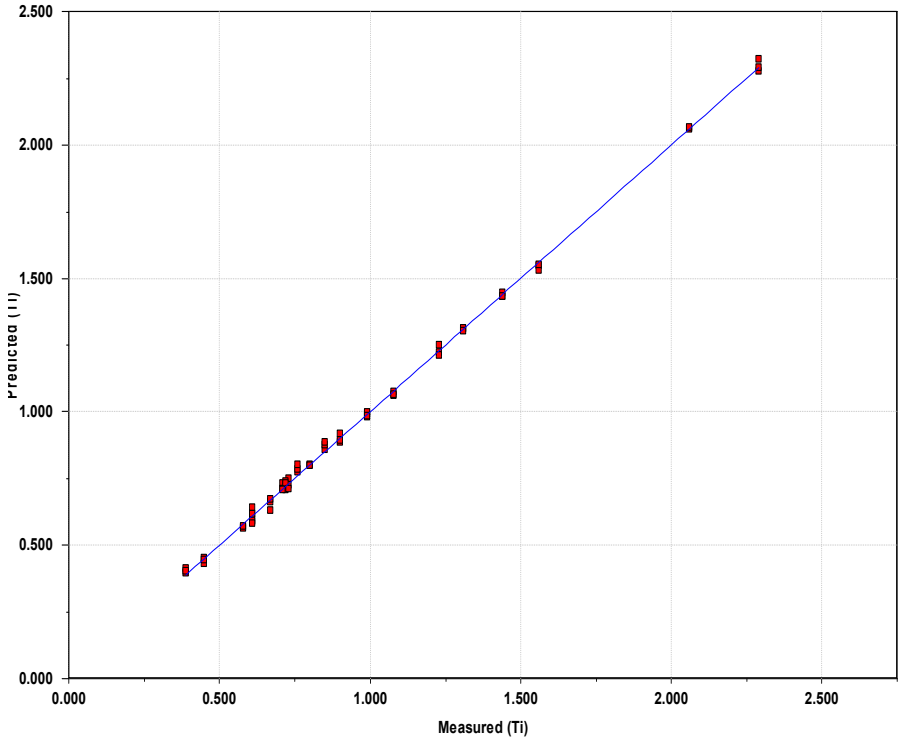


C_{Ti} vs $I_{Ti\ K}$

C_{Pb} vs $I_{Pb\ L}$

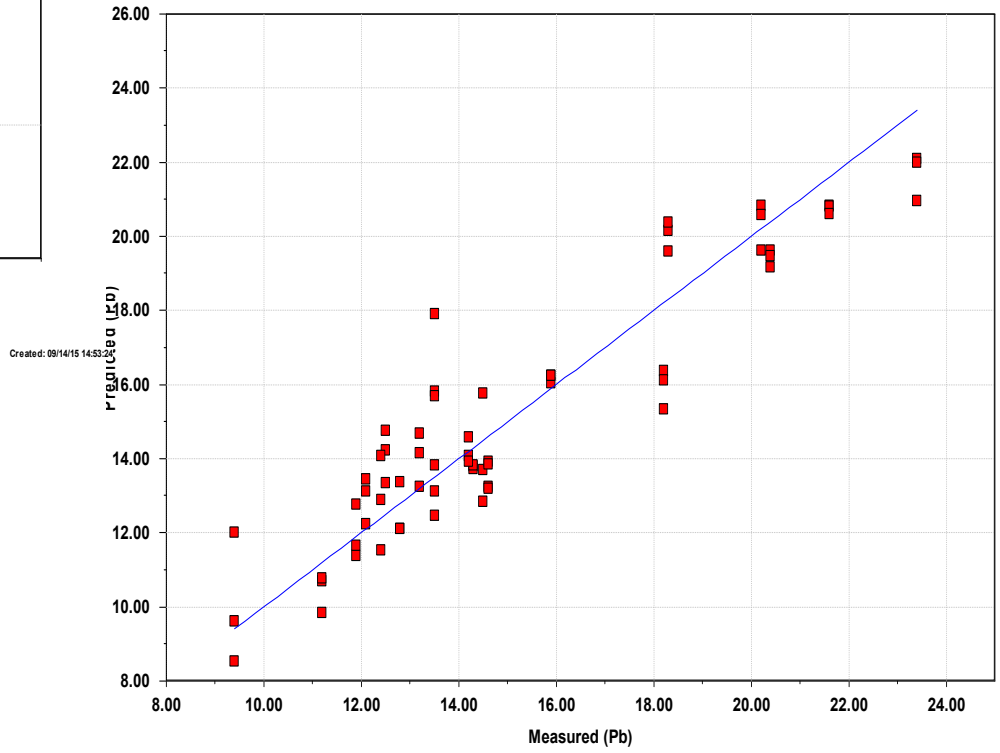


PLS on Ti and Pb

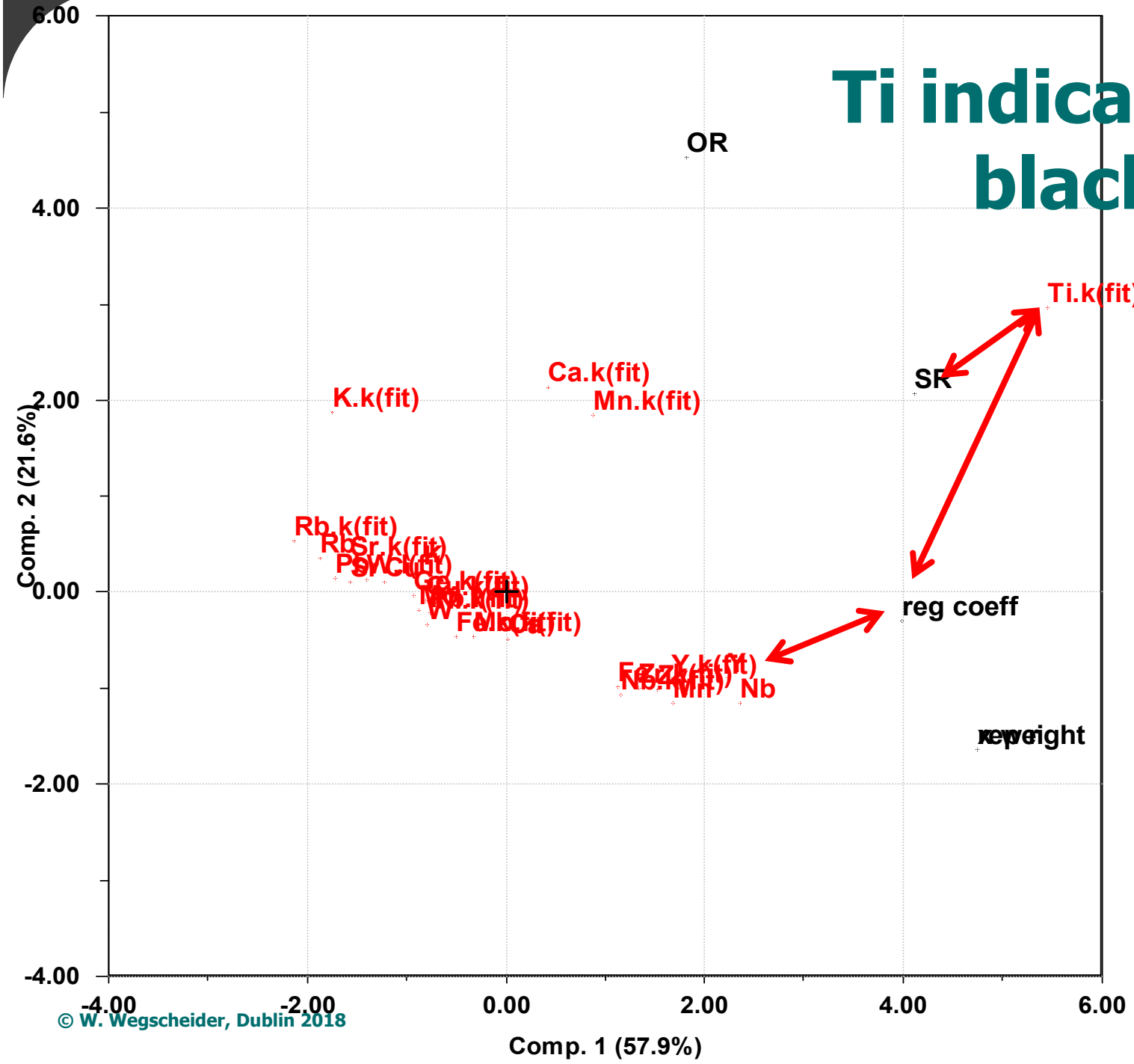


$C_{Ti\ meas}$ vs $C_{Ti\ pred}$

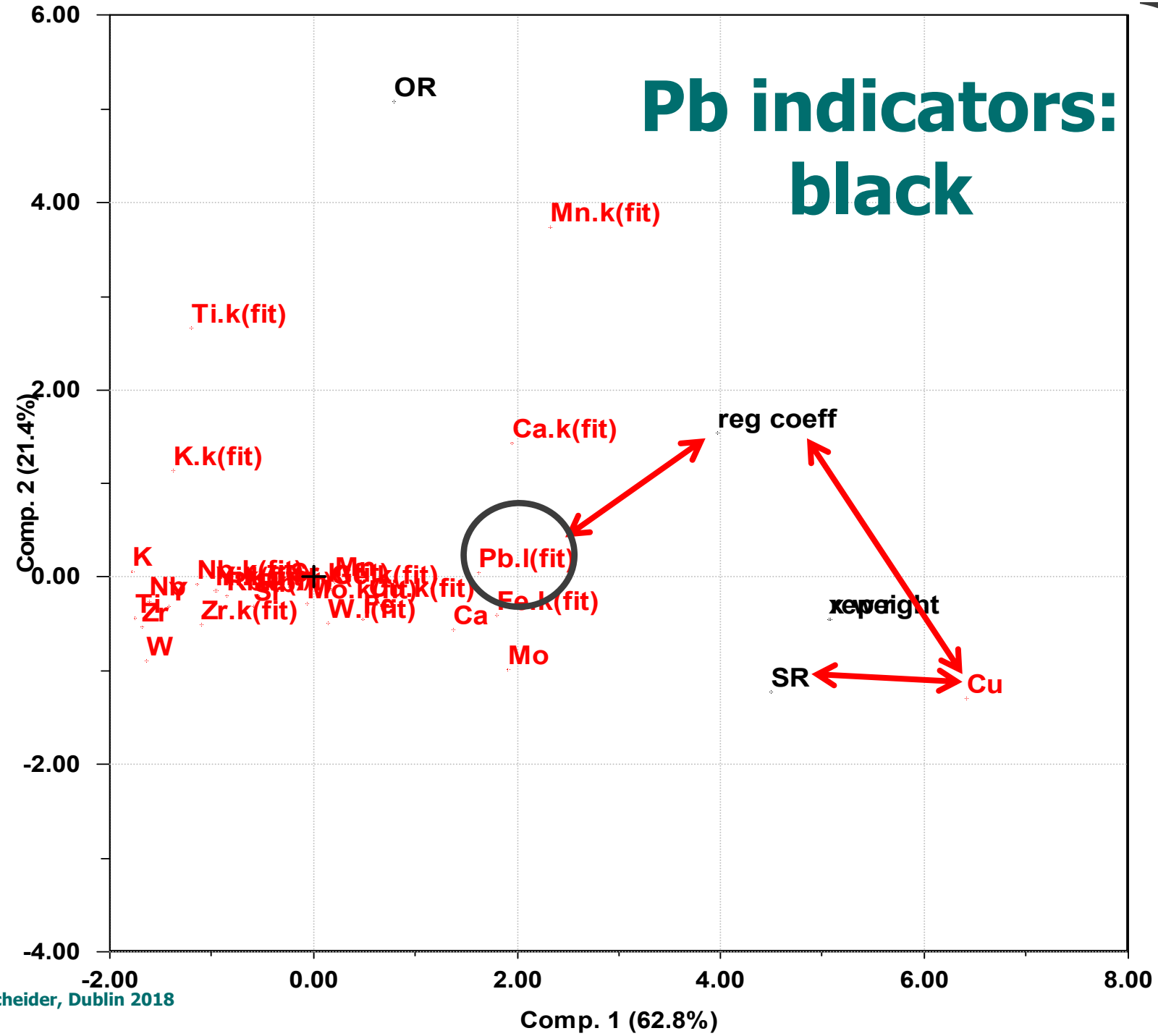
$C_{Pb\ meas}$ vs $C_{Pb\ pred}$



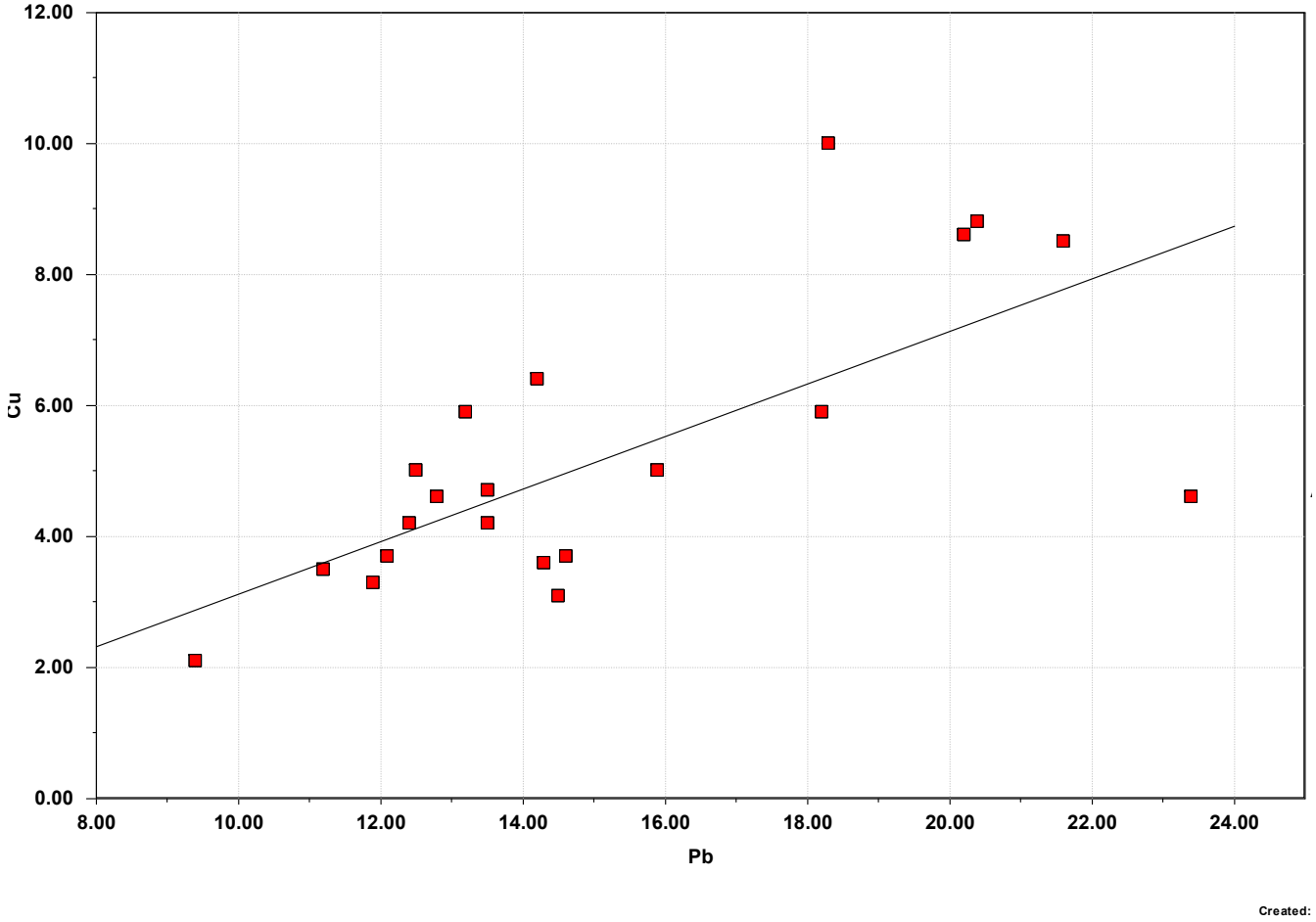
Ti indicators: black



Pb indicators: black



Pb vs Cu: spurious correlation



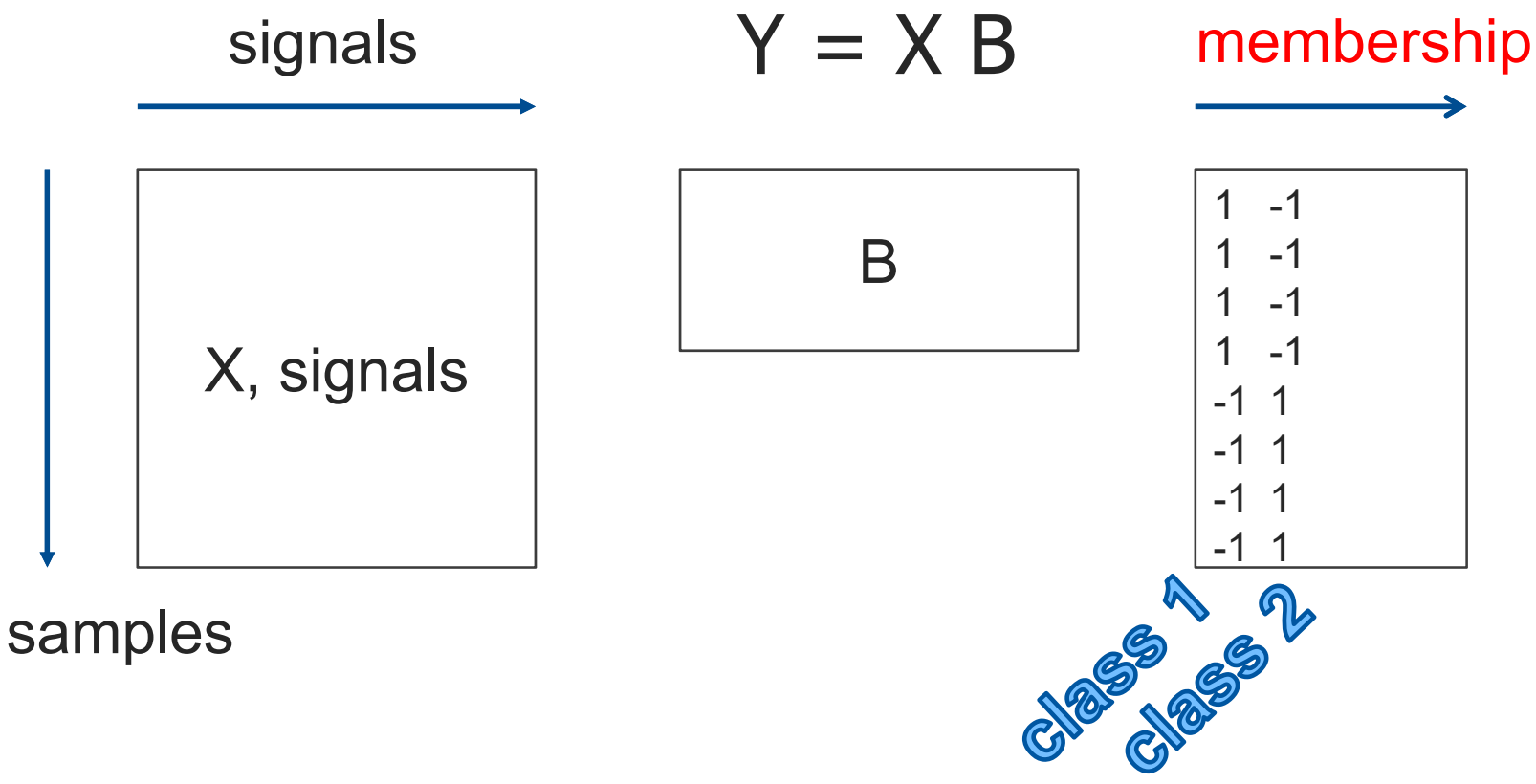
Where do spurious correlations come from?

Calibration models depend

- on the physics/chemistry of the measurement process, but also
- on the internal correlations in the calibration set (here: from geochemistry)

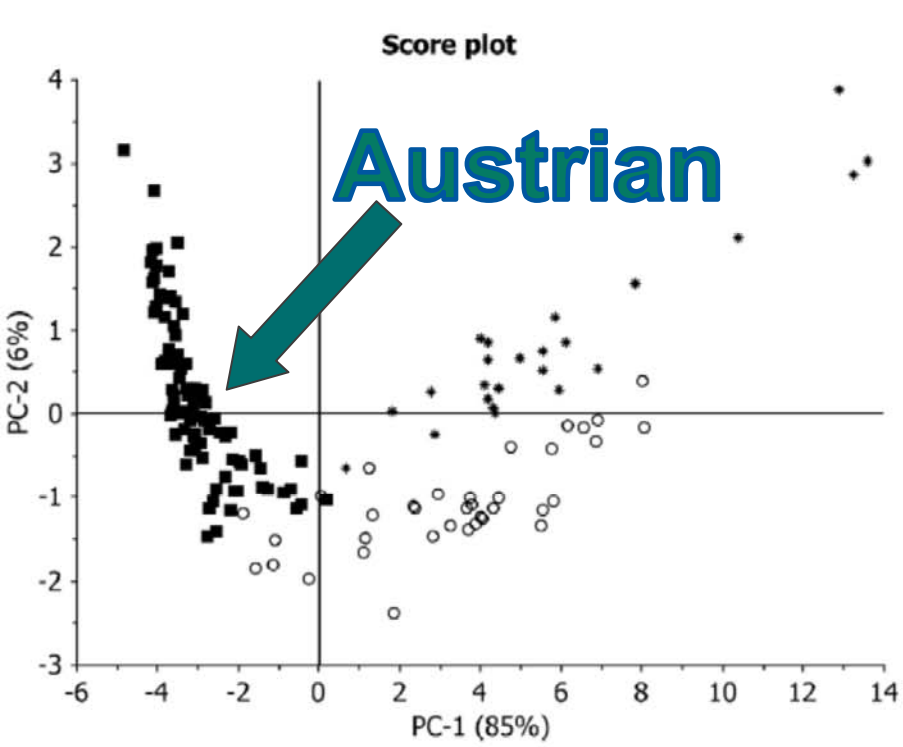
We actually have a sampling problem

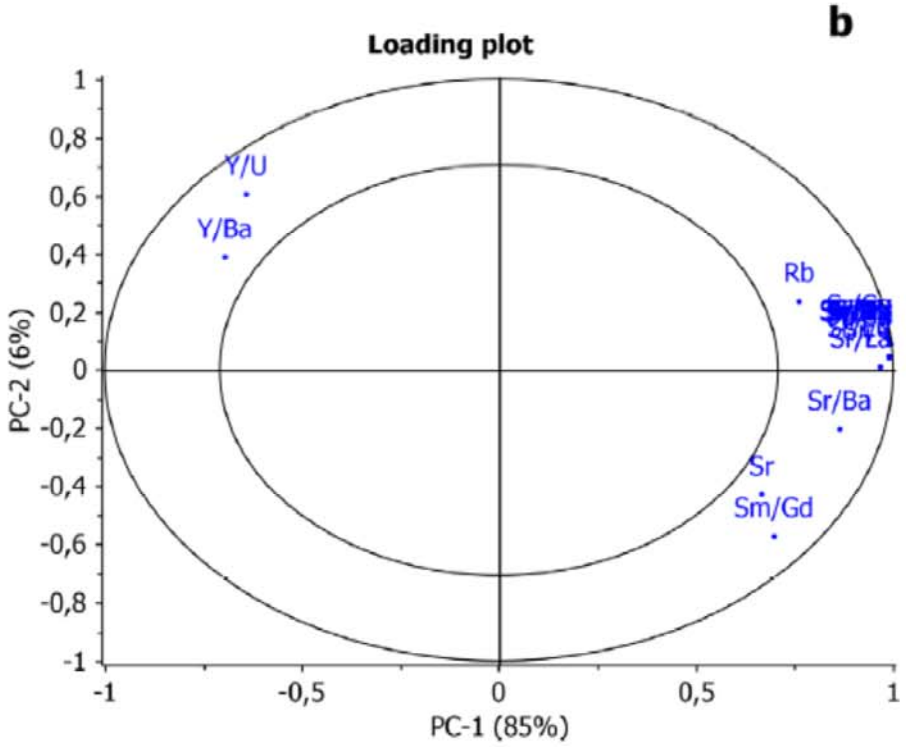
Regression for classification



Origin of Styrian pumpkin seed oil

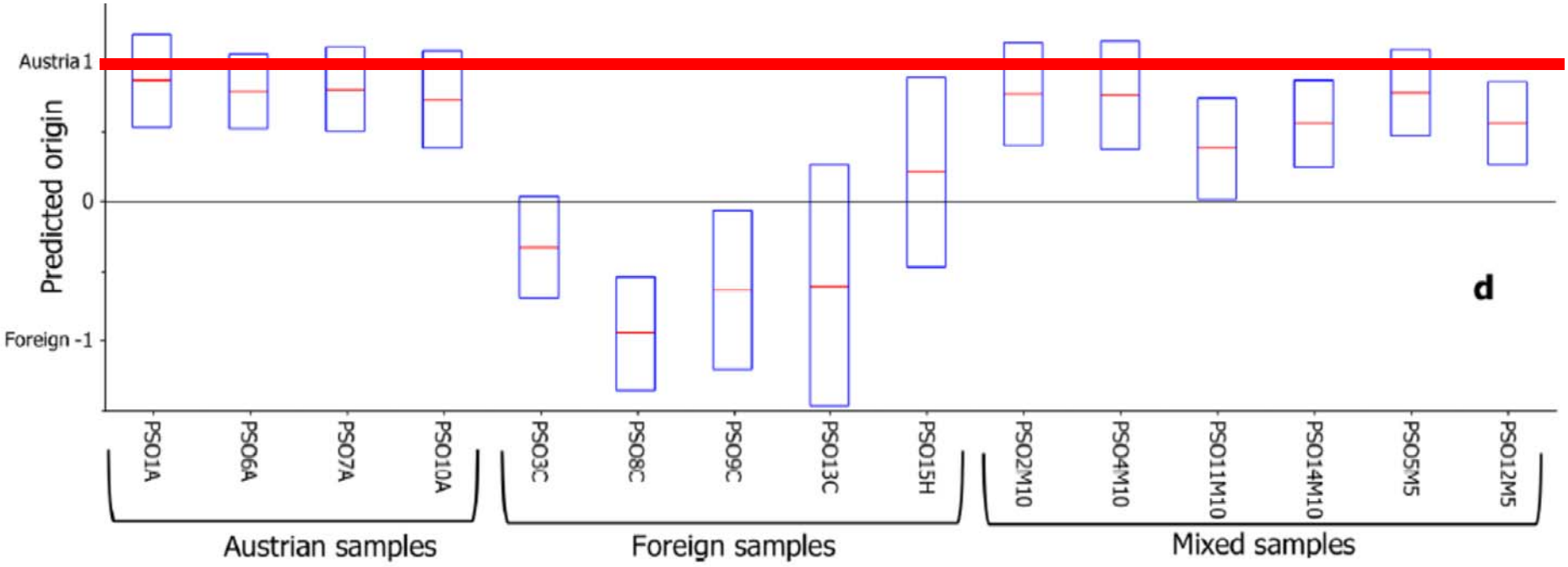
PGI (protected geographical indication)





Zettl et al., 2017

148 oils, 37 elemental variables, external validation set from Austrian Food Authority



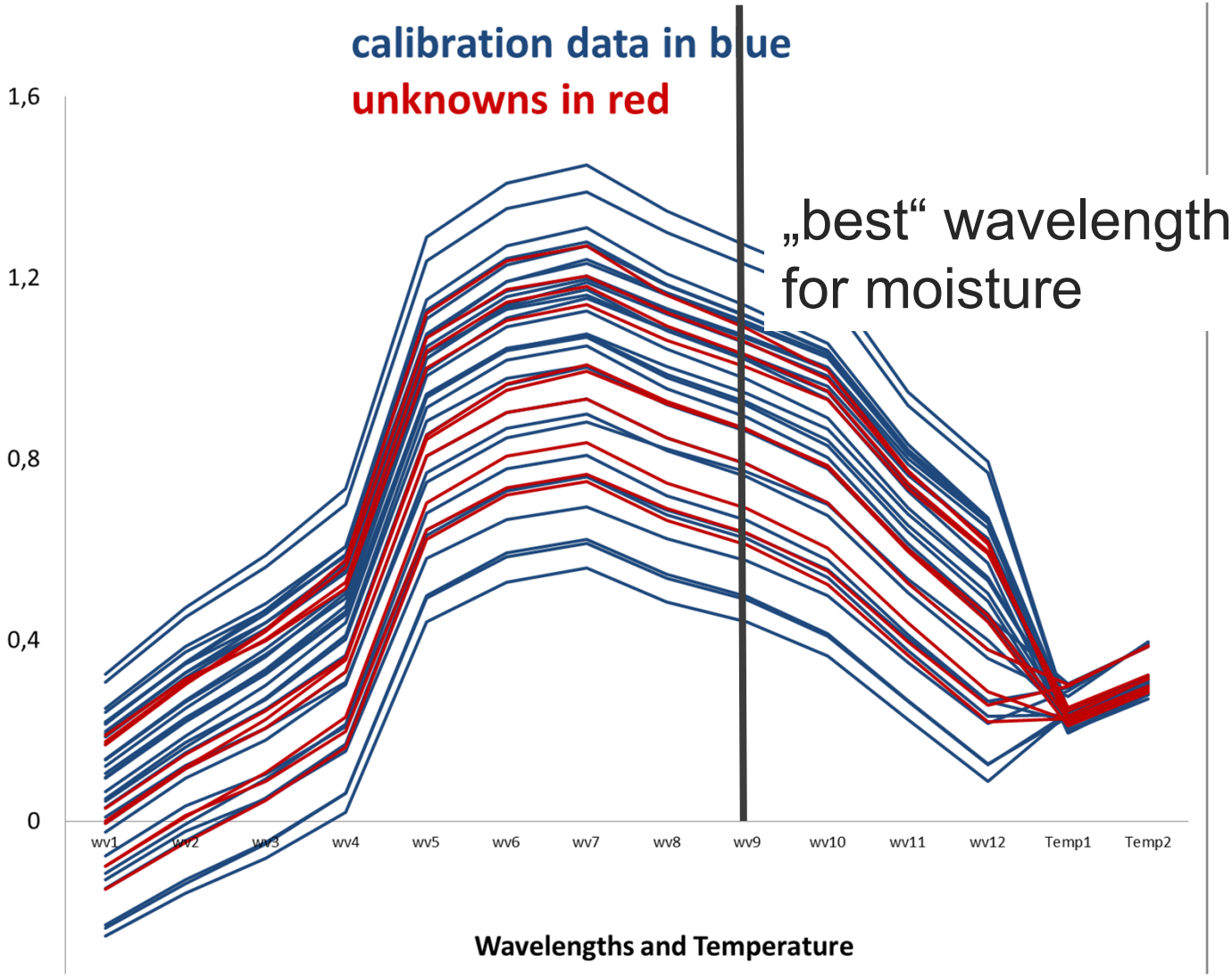
Zettl et al., 2017

Outline

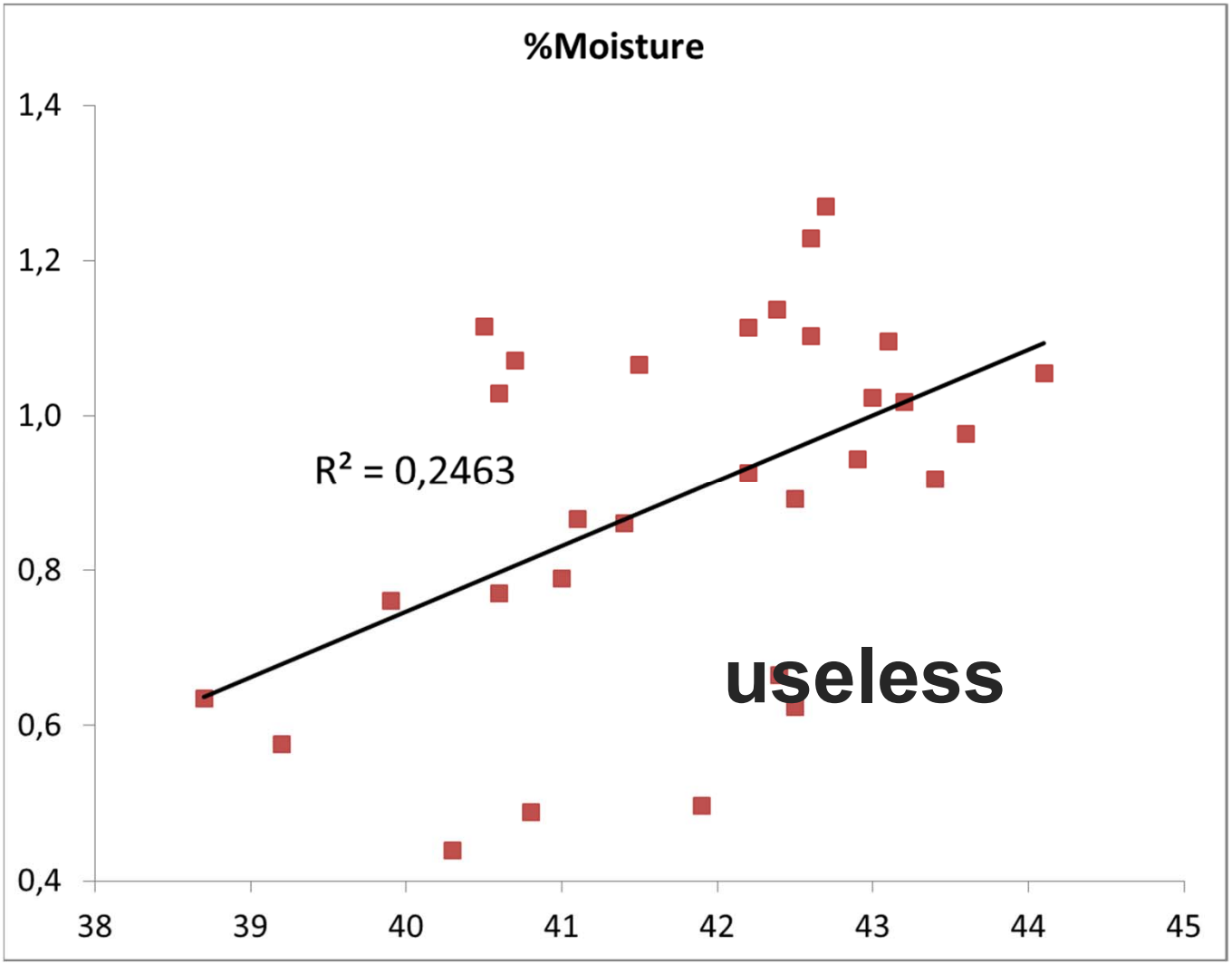
- Historical reminescence: how it all started
- Univariate – multivariate calibration
- Selectivity – specificity – solvability
- Current applications (analytes, signals, matrices, purpose)
- Bias vs. variance in calibration
- Predictive, parsimonious, explanatory models
- **Uncertainty, control charts and traceability**
- Conclusions

NIR spectra of brick cheese

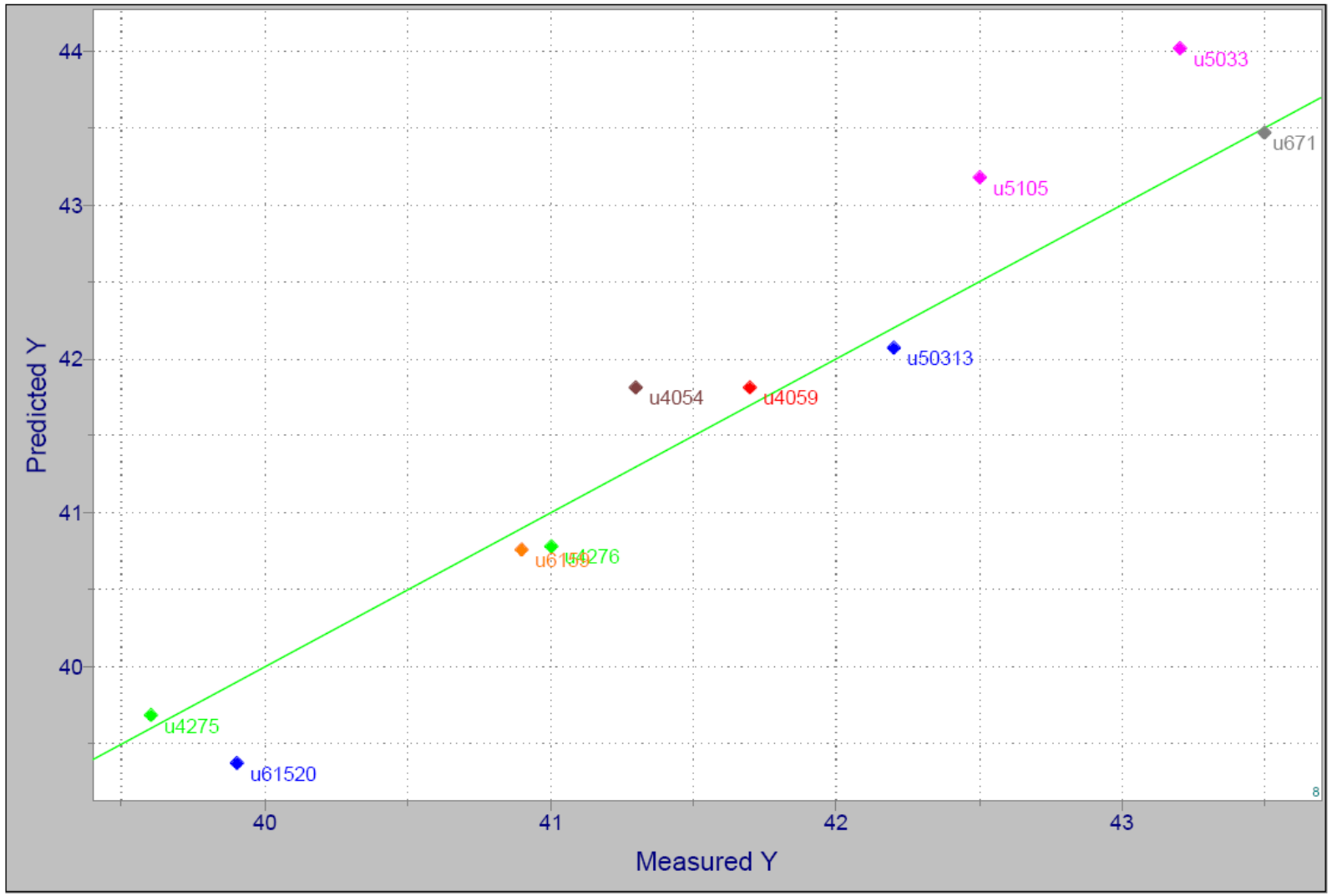
Infometrix Applications Overview 1996



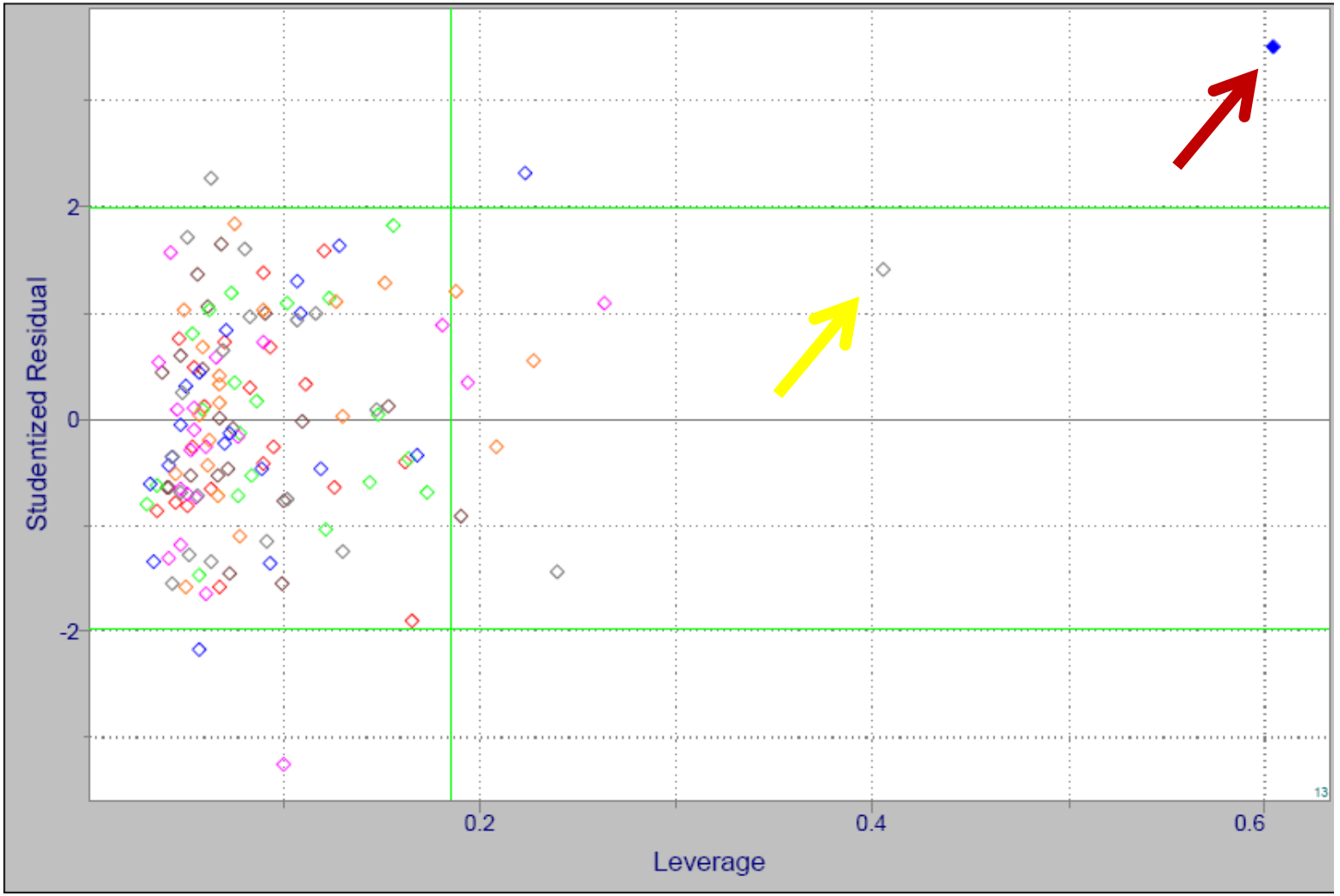
Calibration on wavelength 9



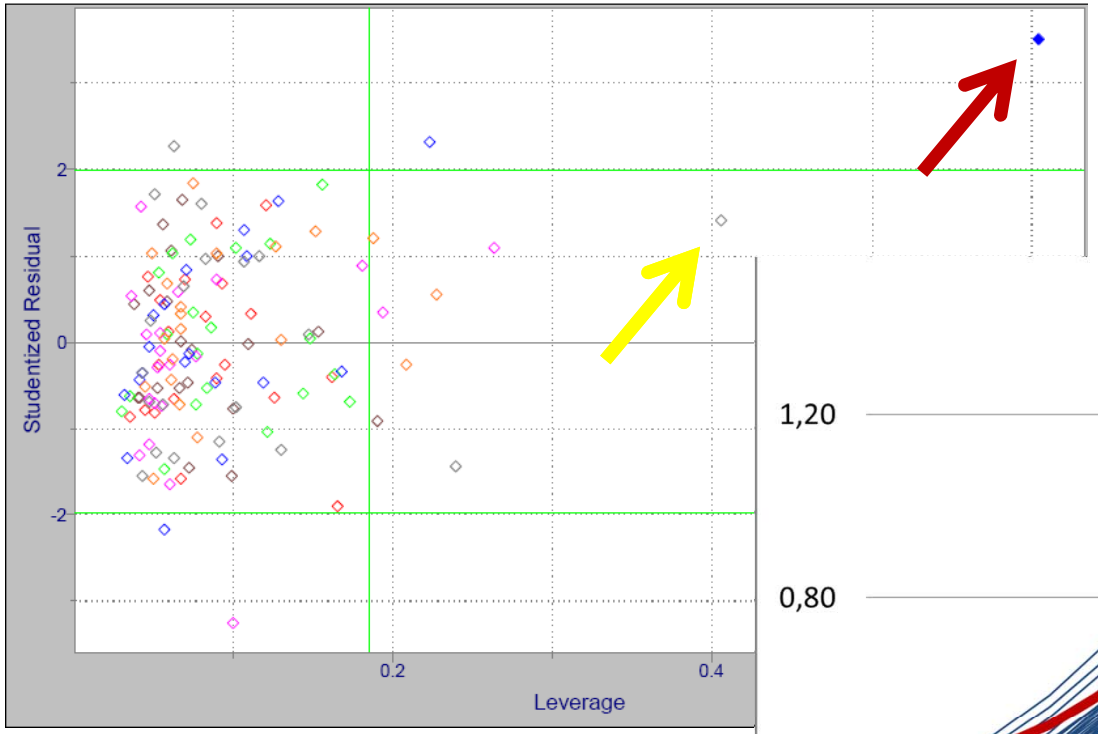
Multivariate prediction of moisture

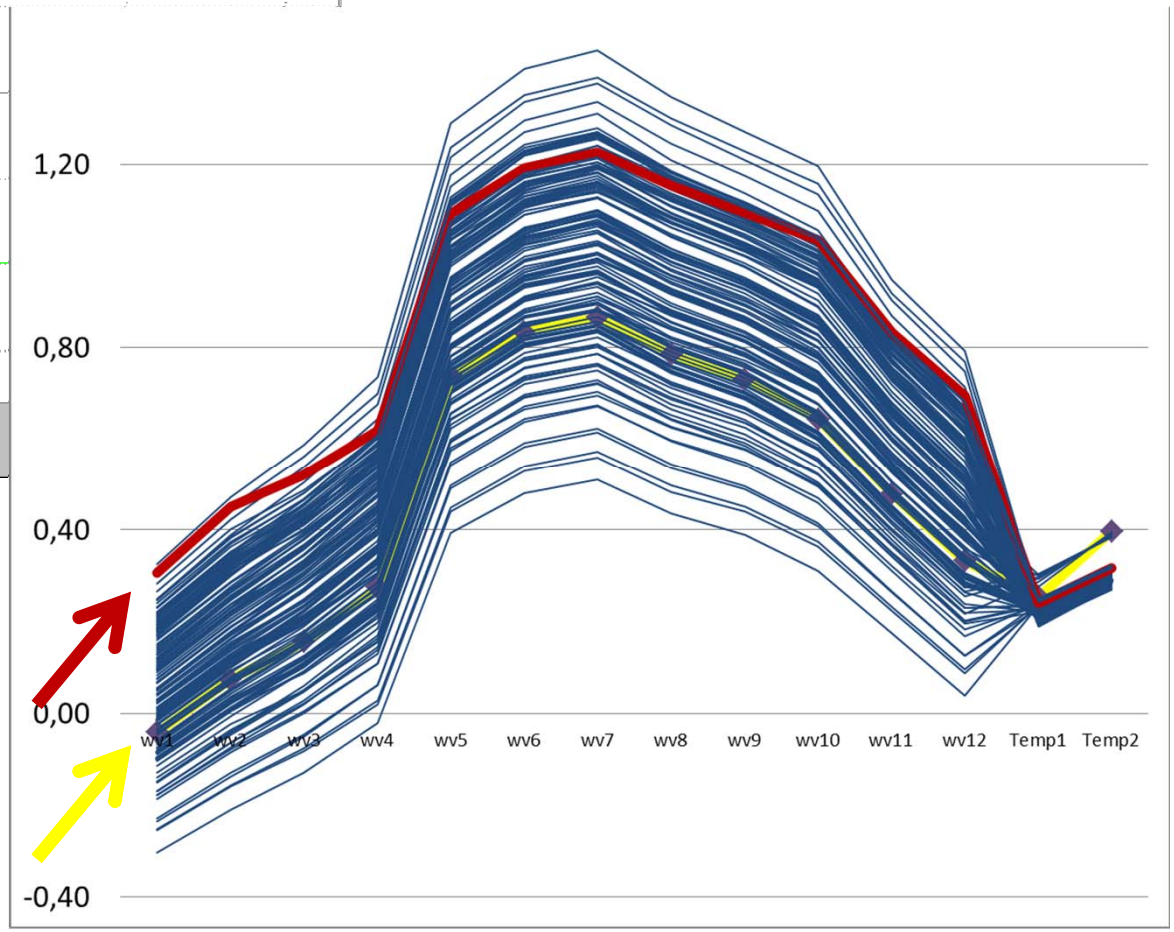


h ..... leverage



h ..... leverage vs. spectra





IUPAC Approach to „uncertainty“

Pure Appl. Chem. 78, No. 3, pp. 633–661, 2006

$$\left[s(\mathrm{c})\right]^2 = hs_{\mathrm{c}}^2 + h(s_{\mathrm{r}}/S_n)^2 + (s_{\mathrm{r}}/S_n)^2$$

$\left[s(\mathrm{c})\right]^2$ variance of prediction

hs_{c}^2 leverage * var. of conc. of standards

$h(s_{\mathrm{r}}/S_n)^2$ leverage * var. of (signals/sensitivity)

$(s_{\mathrm{r}}/S_n)^2$ variance of (signals/sensitivity)

Effect as extra-RMSEP
in % moisture

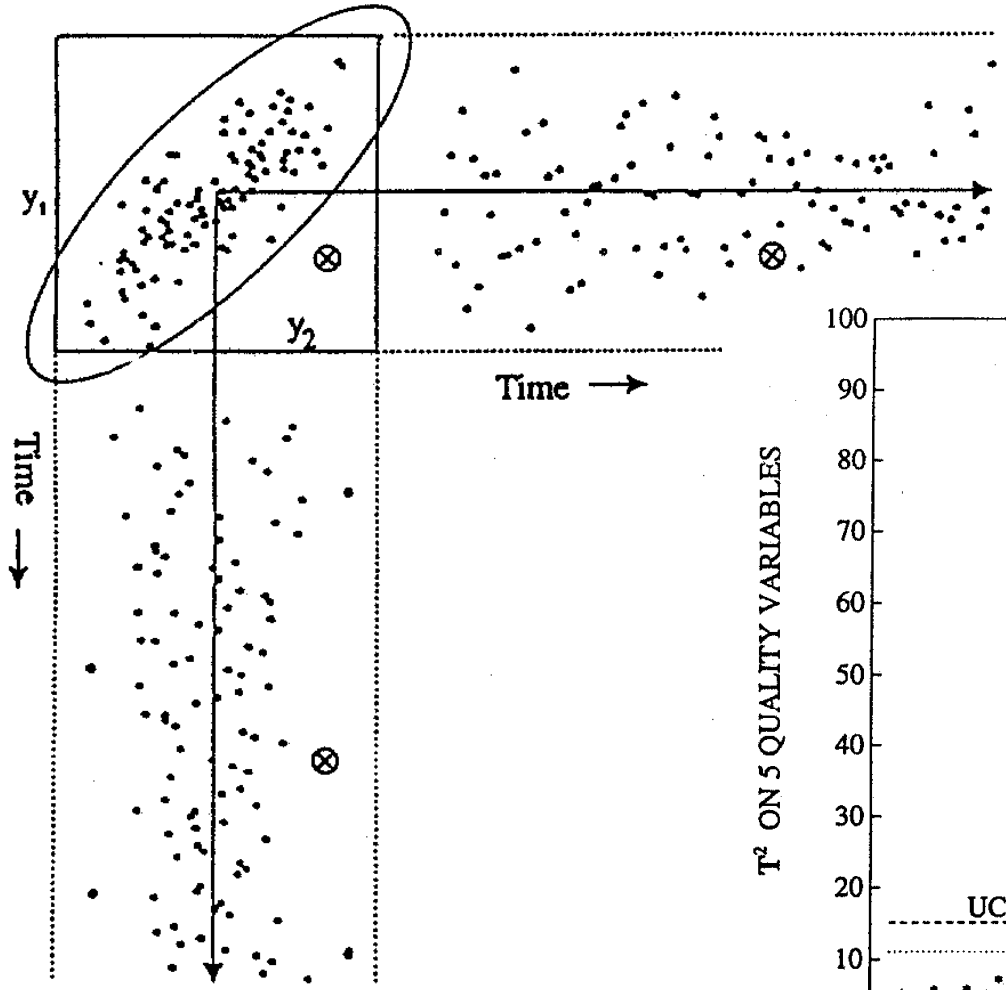
Noise of 2% added to	Spectra of standards	Spectra of unknowns	Concen-tration of standards	Sum (IUPAC)	All three
Effect (in %moisture)	0.194				

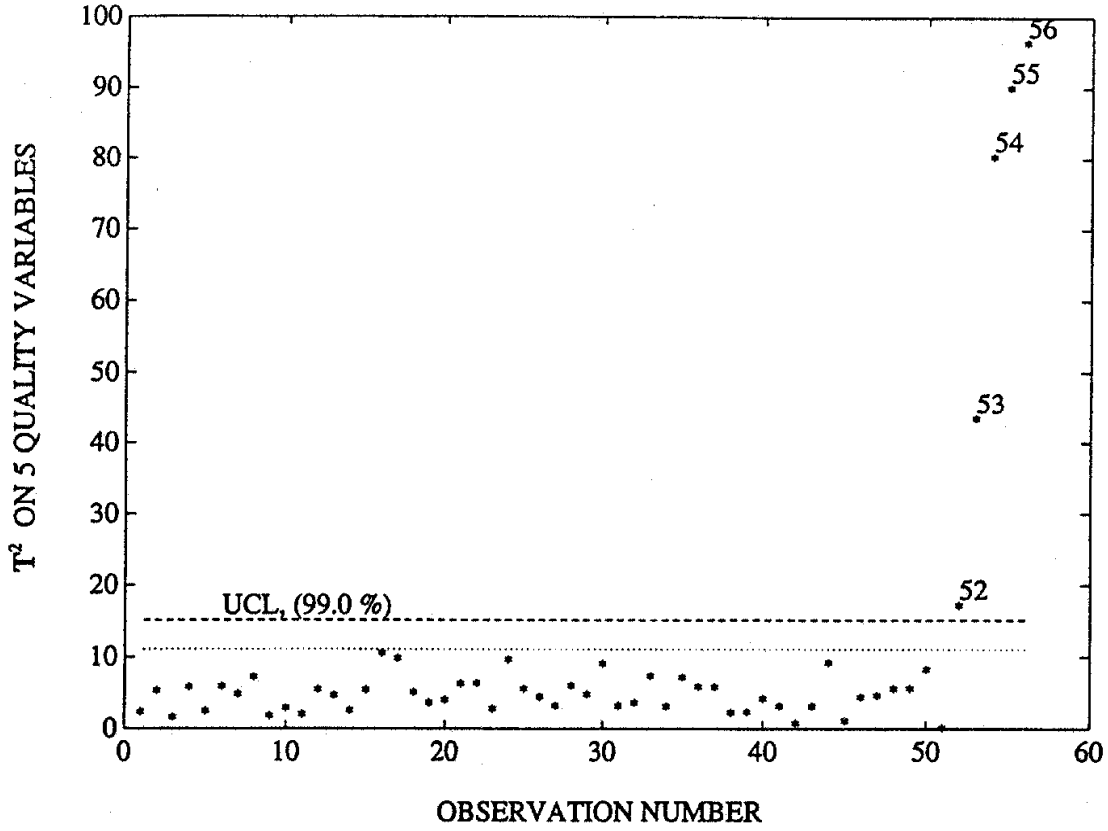
RMSEP – **r**oot **m**ean **s**quared **e**rror of **p**rediction

- The IUPAC-proposed summation does not work
- Expected reason: no allowance for significant correlations
- Full Monte-Carlo procedure is required instead

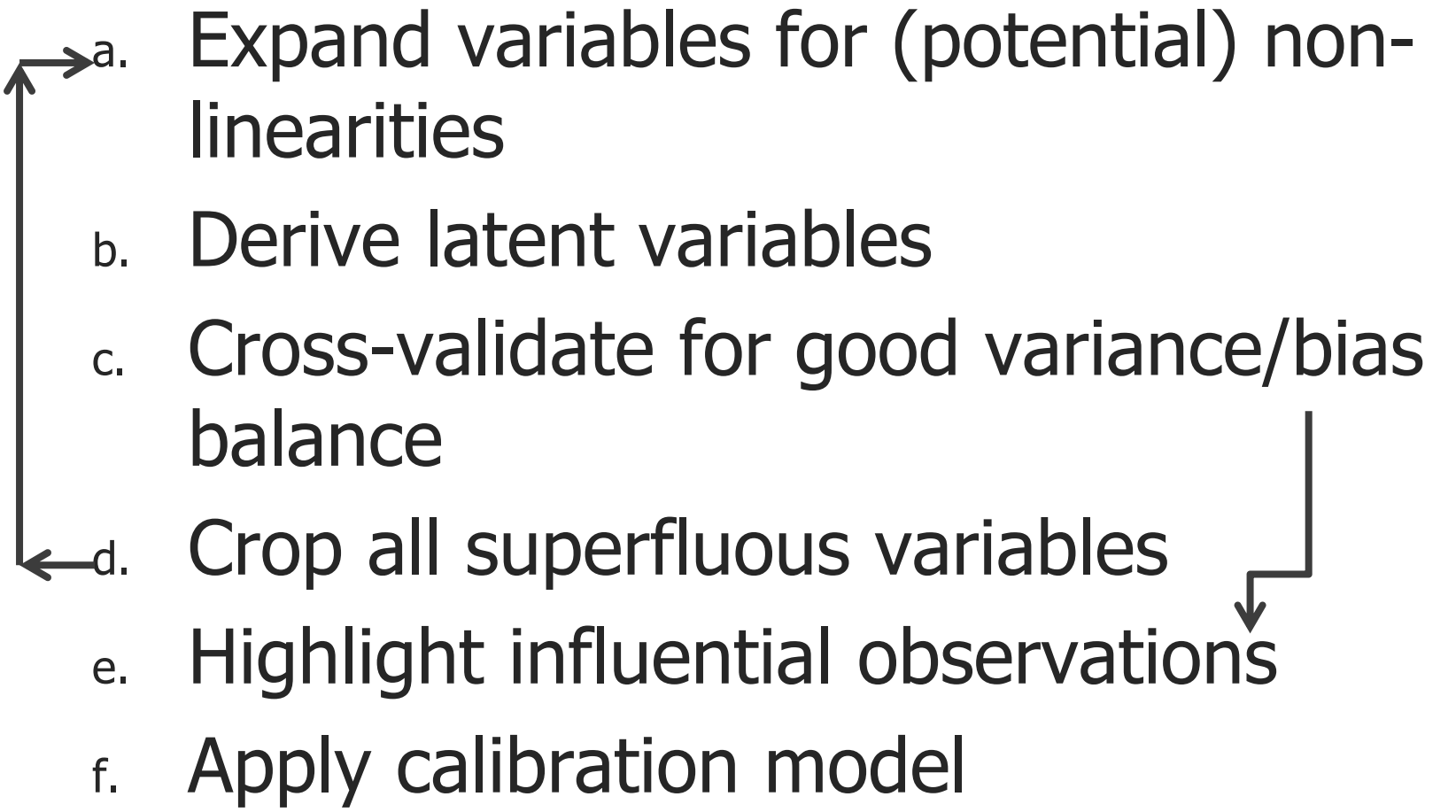
Multivariate control charts

Kourti & MacGregor, 1995





Automation to achieve „traceable“ calibration models



Conclusions

- Multivariate calibration will be even more prominent in the future
- So far, too little has been invested in the understanding of the calibration models
- Traceability (and reproducibility) of calibrations require a fixed and hopefully automated protocol
- Many univariate QC measures are not useful

Acknowledgements

- Böhler Co, Kapfenberg, for cooperation in metals analysis
- Department of Applied Geosciences, Leoben, for cooperation on sediment samples
- O. Kvalheim (Bergen, NOR) for making available the newest version of Sirius

Thank you for your interest