Chemometrics

Introduction

Course 27411 Biological data analysis and chemometrics

Jens C. Frisvad

Fundamental disciplines in biological sciences

- Classification (theory: taxonomy)
  - Discrimination (diagnostics)
  - Identification
  - Nomenclature
- Cladification (theory: phylogeny)
- Modelling and predictions
- Tests and validations

Engineering

- The science by which the properties of matter and the sources of energy in nature are made useful to man in structures, machines and products
- Measurement techniques, computers, language, definitions, properties, chemistry, physics, mathematical hard modelling, statistics, chemometrics and many other disciplines are necessary to be a good engineer

X-metrics

- The use of multivariate statistics in the discipline X

- Psychrometrics (used in psychology)
- Taxometrics (used in taxonomy)
- Biometrics (used in biology)
- Technometrics (used in engineering)
- Chemometrics (used in chemistry)

Chemometrics

- Use of statistics and mathematics in chemical sciences, to measure and interpret chemical data (also used for biological data)

- Biometrics has often been restricted to univariate statistics and taxometrics to biosystematics

Other definitions of chemometrics:

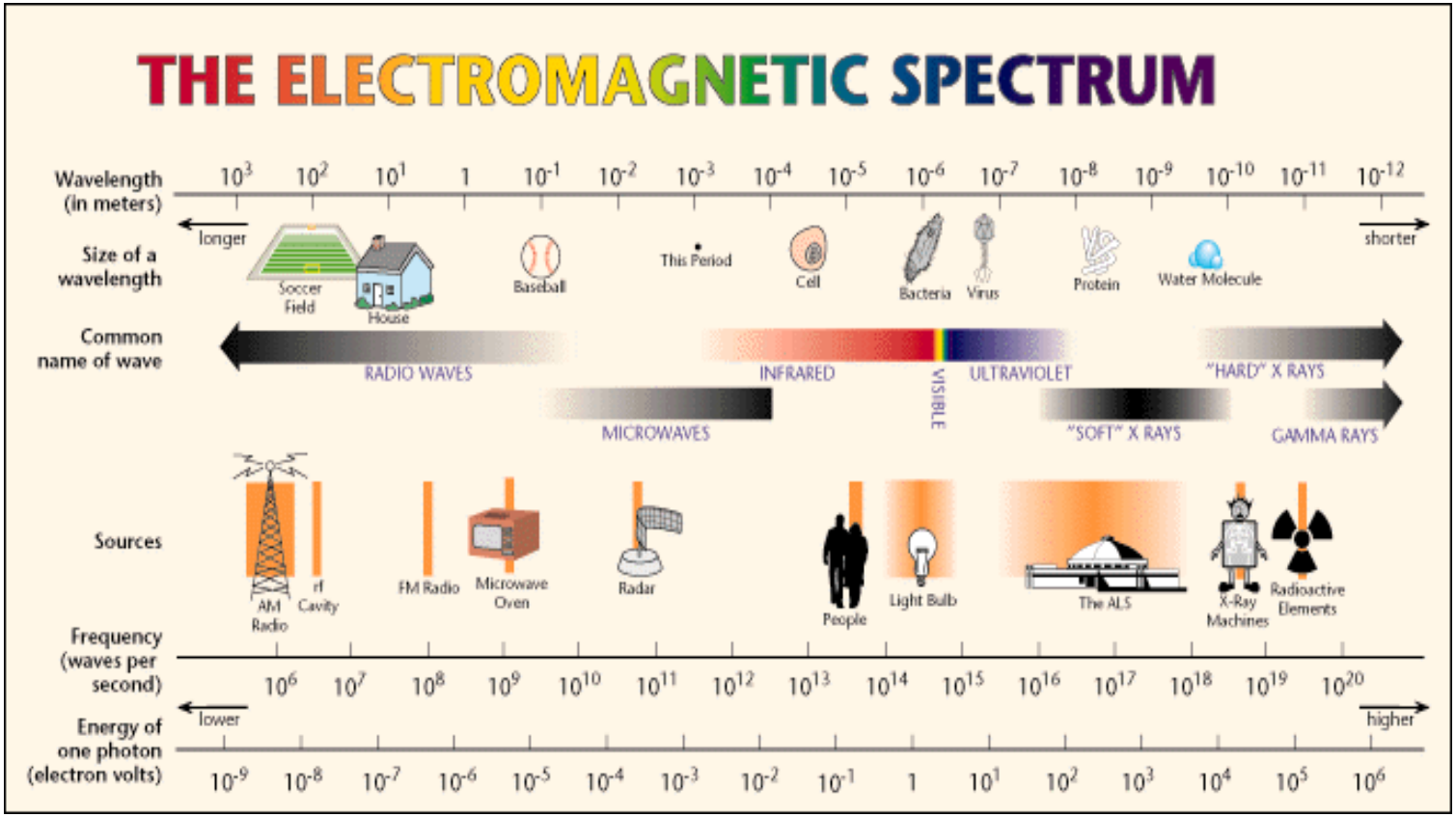
- Empirical interactive data-driven modelling in chemistry (induction and abduction)
- Exploratory and confirmatory data-analysis (hypothesis generating and hypothesis testing) in chemistry
- Predictive multivariate modelling in chemistry

Important disciplines in chemometrics

- Sampling, selection of objects and variables
- Clustering
- Ordination (projection from N dimensions to few dimensions, eigen vector based analyses)
- Multivariate regressions, calibrations and predictions
- Neural networks
- Validation (test set validation, boot-strapping, jackknifing, cross-validation)
- Graphical display and outlier analysis

Why use chemometrics?

- Complex systems with many interactions are common in science
- Indirect (and often non-destructive) observation of the world as it is
- An expansion of the human perception (the full electromagnetic spectrum)
- Chromatography and spectroscopy will always yield multivariate data


THE ELECTROMAGNETIC SPECTRUM
Wavelength (in meters)
10^{3} 10^{2} 10^{1} 1 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5} 10^{-6} 10^{-7} 10^{-8} 10^{-9} 10^{-10} 10^{-11} 10^{-12}
longer
shorter
Size of a wavelength
Soccer Field
House
Baseball
This Period
Cell
Bacteria
Virus
Protein
Water Molecule
Common name of wave
RADIO WAVES
MICROWAVES
INFRARED
VISIBLE
ULTRAVIOLET
"SOFT" X RAYS
"HARD" X RAYS
GAMMA RAYS
Sources
AM Radio
rf Cavity
FM Radio
Microwave Oven
Radar
People
Light Bulb
The ALS
X-Ray Machines
Radioactive Elements
Frequency (waves per second)
10^{6} 10^{7} 10^{8} 10^{9} 10^{10} 10^{11} 10^{12} 10^{13} 10^{14} 10^{15} 10^{16} 10^{17} 10^{18} 10^{19} 10^{20}
lower
higher
Energy of one photon (electron volts)
10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6}

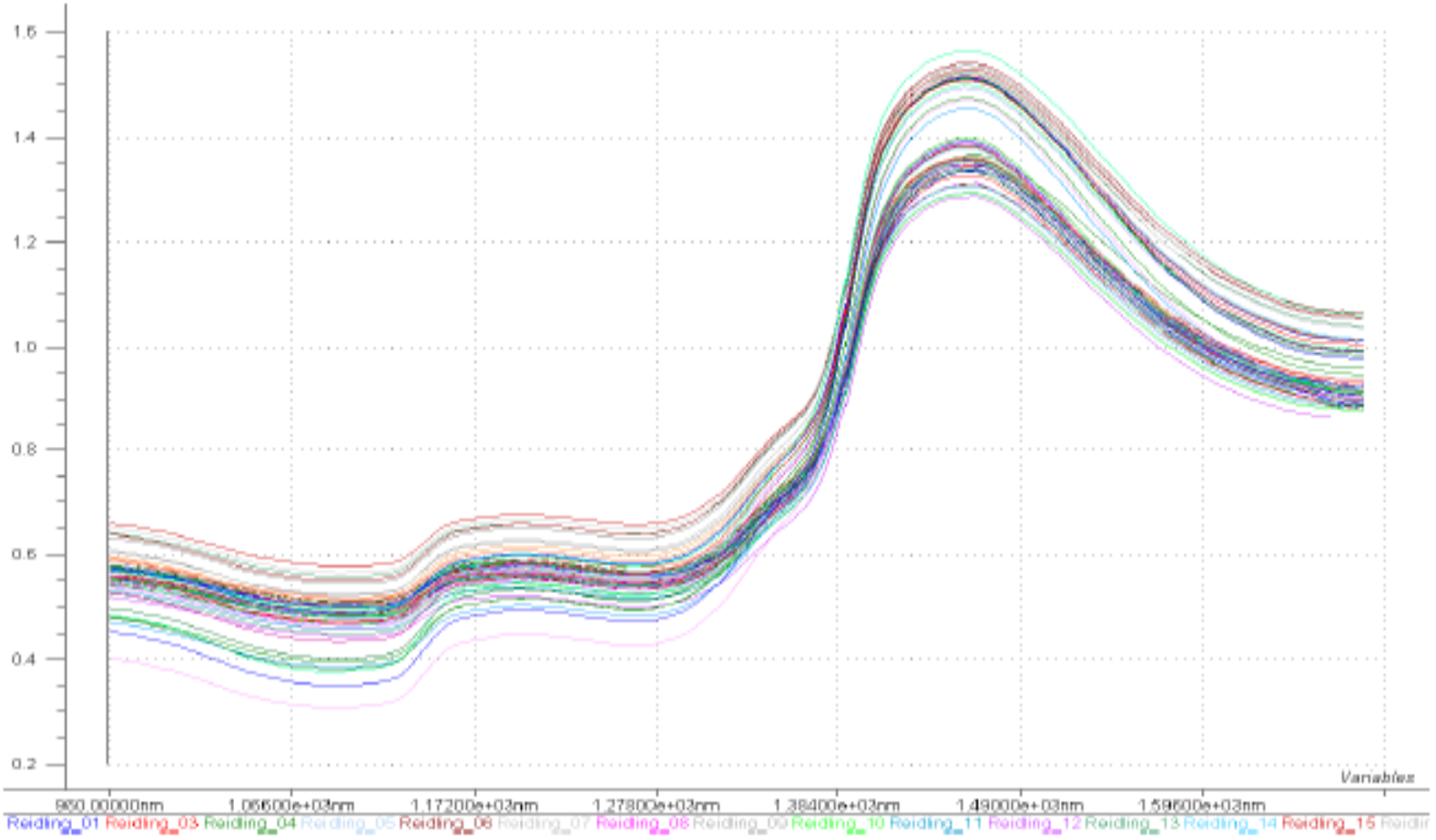
Most new analytical chemical methods give multivariate data

- IR and NIR
- UV-VIS & fluorescence
- NMR and ESR
- MS
- TLC
- HPLC
- GC
- CE
- CCC

Interpreting chemical data

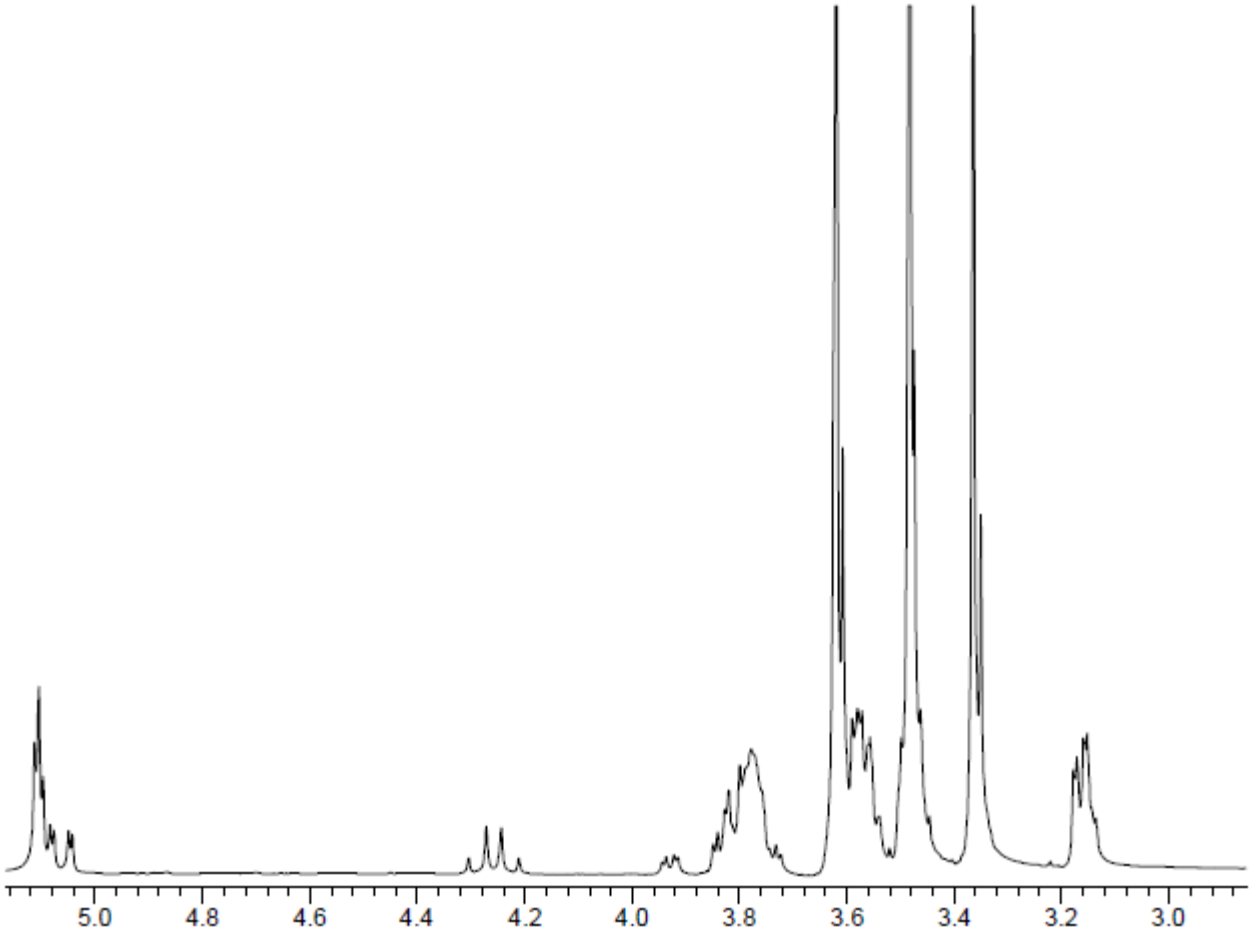
- Abduction
  - Sharp independent signals that can be interpreted:
    - NMR
    - IR
    - MS
- Induction
  - Soft overlayered signals (strong interaction, co-linearity)
    - UV
    - Fluorescence
    - NIR
    - FIA
  - Any chromatographic/spectrometric measurements on mixtures

NIR spectra, ”soft spectra”



From Jens Bo Holm Nielsen

H-NMR spectrum, sharp signals



No. variables > No. objects

- Classical statistical methods will not work in that case (for example multiple linear regression, linear learning machine etc.)

- Two solutions to this can be variable selection or classical statistics on scores from eigen vector analysis

Important methods in chemometrics (classification)

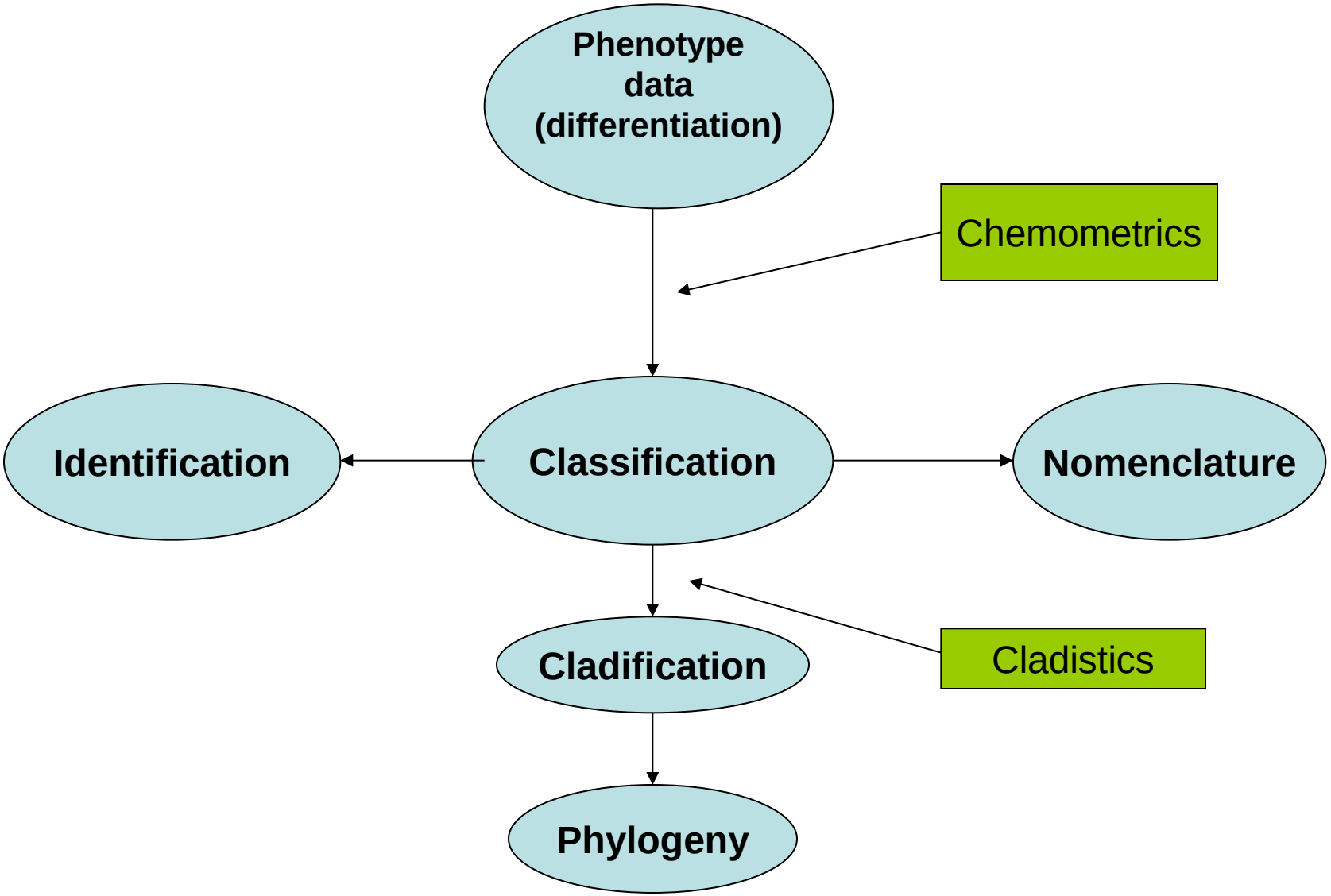
- Cluster analysis
  - Hierarchical clustering
  - Divisive clustering
  - Block-clustering and fuzzy clustering
- Ordination
  - Principal component analysis
  - Correspondence analysis
  - Multidimensional scaling

The Celestial Emporium of Benevolent Knowledge (encyclopedia from the 10. century)
An arbitrary and idiosyncratic classification

Classification of plants and animals:

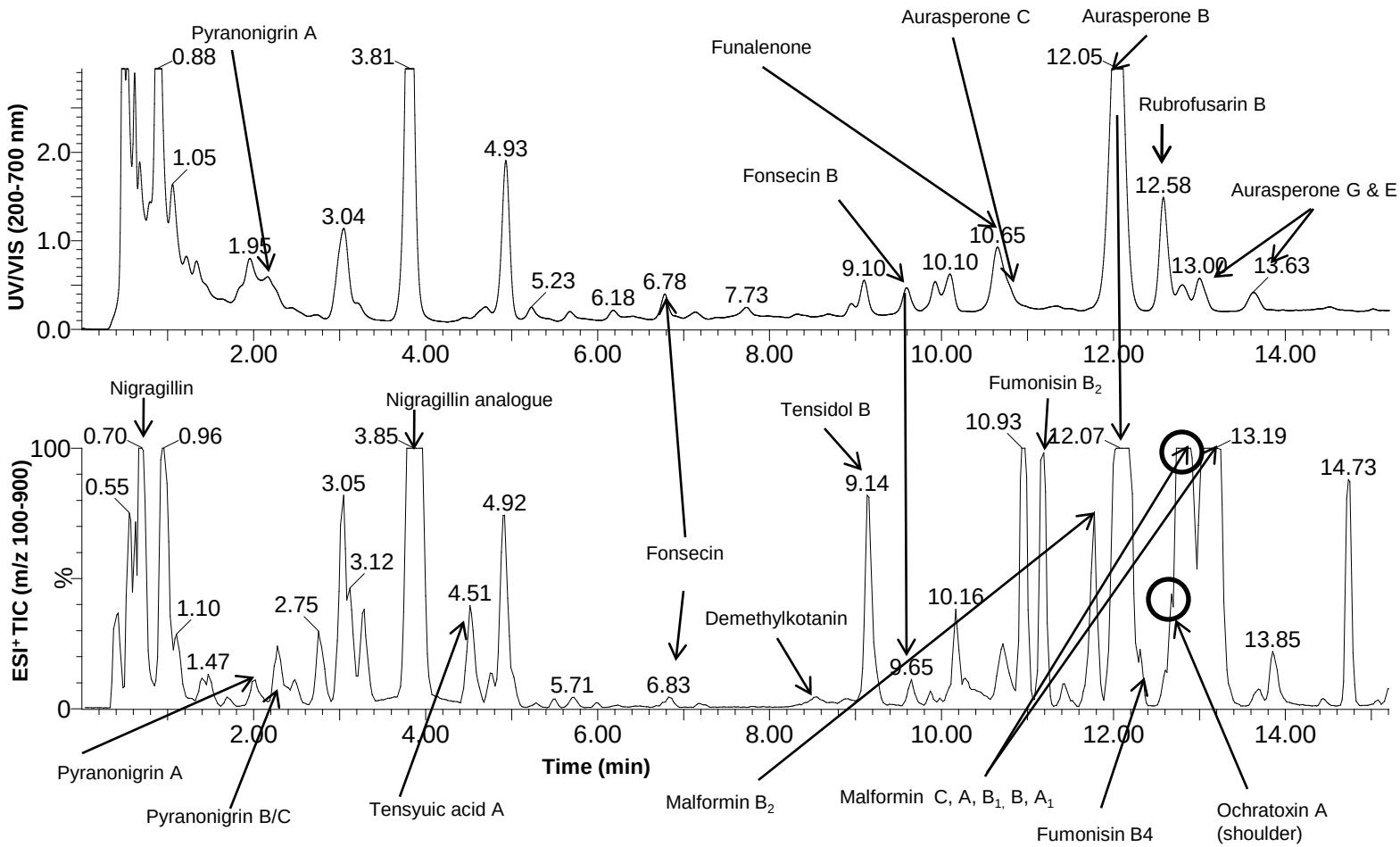
1. Those that belong to the Emperor
2. Embalmed ones
3. Those that are trained
4. Suckling pigs
5. Mermaids
6. Fabulous ones
7. Stray dogs
8. Those that are not included in this classification
9. Those that tremble as if they were mad
10. Innumerable ones
11. Those drawn with a very fine camel's hair brush
12. Others
13. Those that have just broken a flower vase
14. Those that resemble flies on a distance

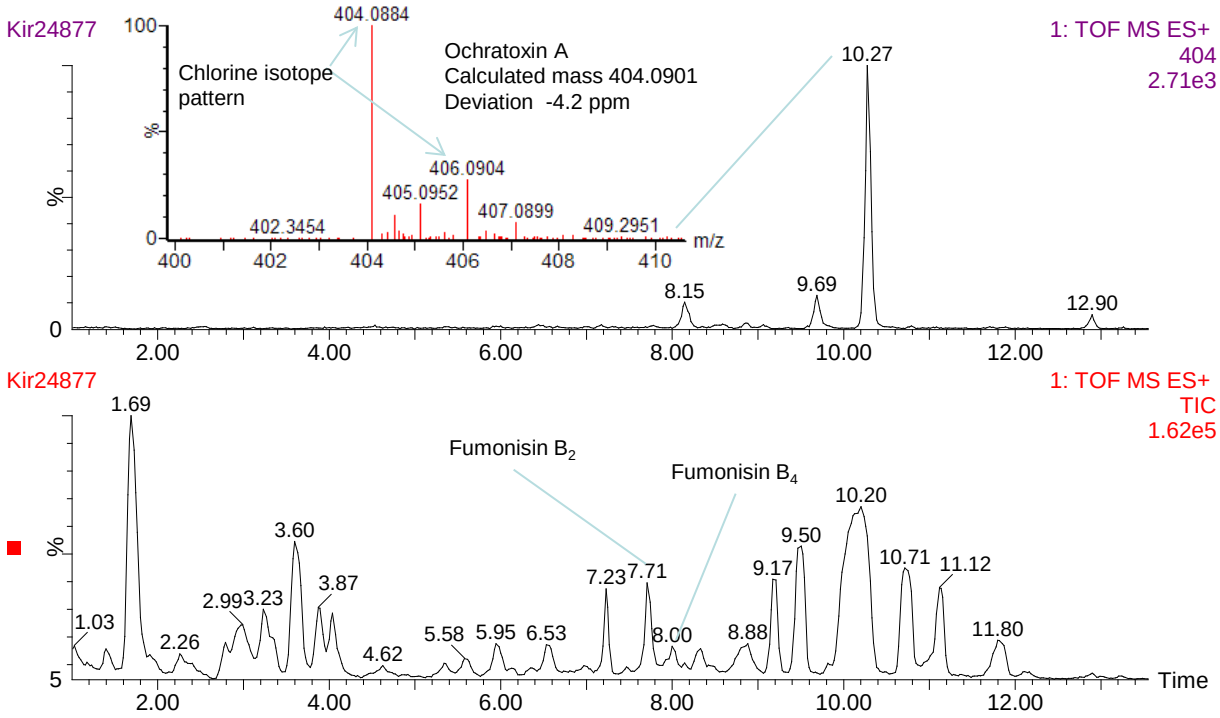
Classification is central!



Examples of chemical data

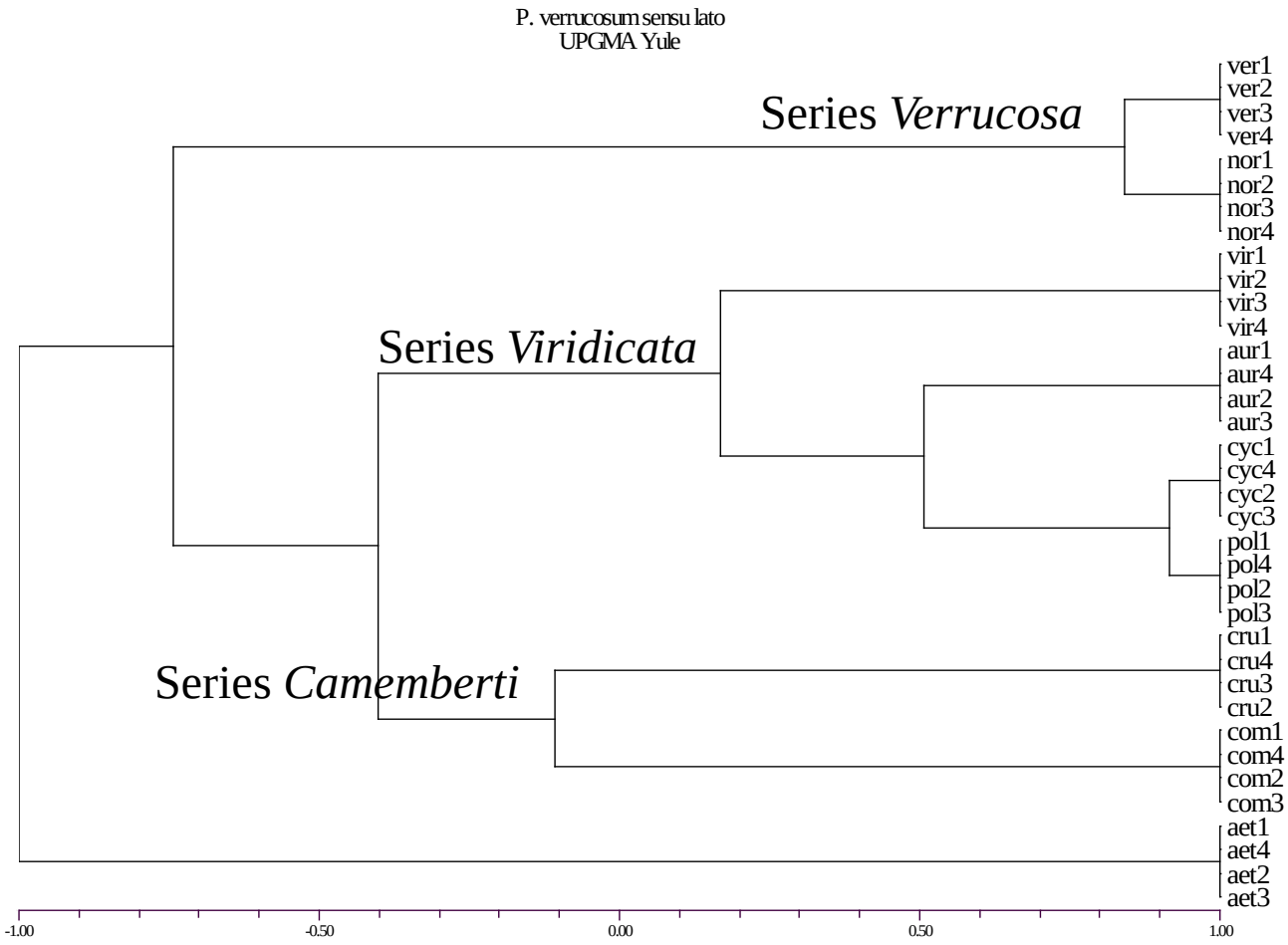
Aspergillus niger secondary metabolite HPLC profile (sharp signals)



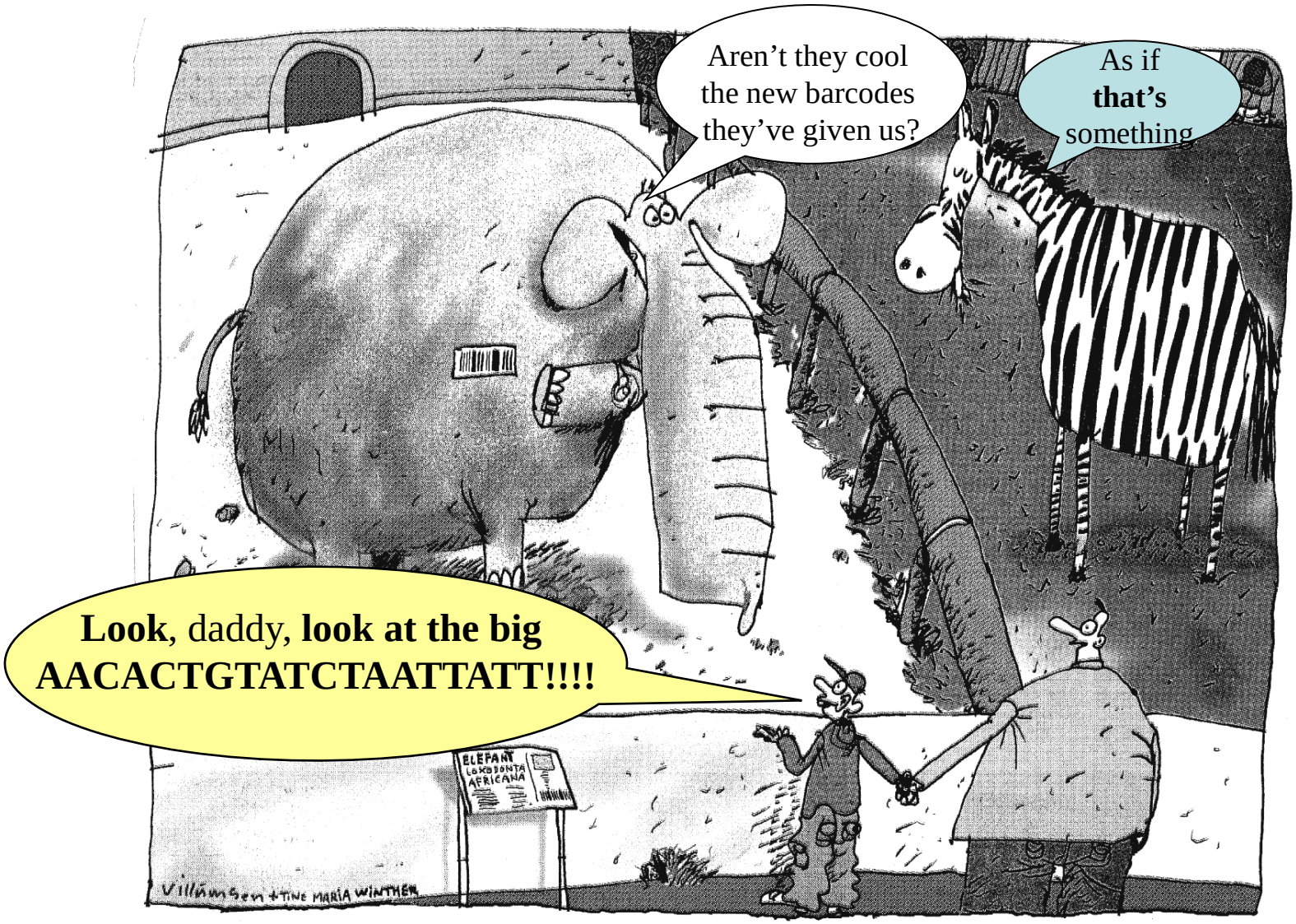


HPLC-ESI$^+$ chromatograms (Luna C_{18} (2) column) *A. niger* NRRL 3122 from YES agar.
Upper extracted ion chromatogram m/z 404 and lower total ion chromatogram (m/z 100-900)

Clustering of some common Penicillia based on 31 extrolite biosynthetic families



Biosystematics: Genome or phenome?



Politiken, 16/3 2003

Methods used in cladistics

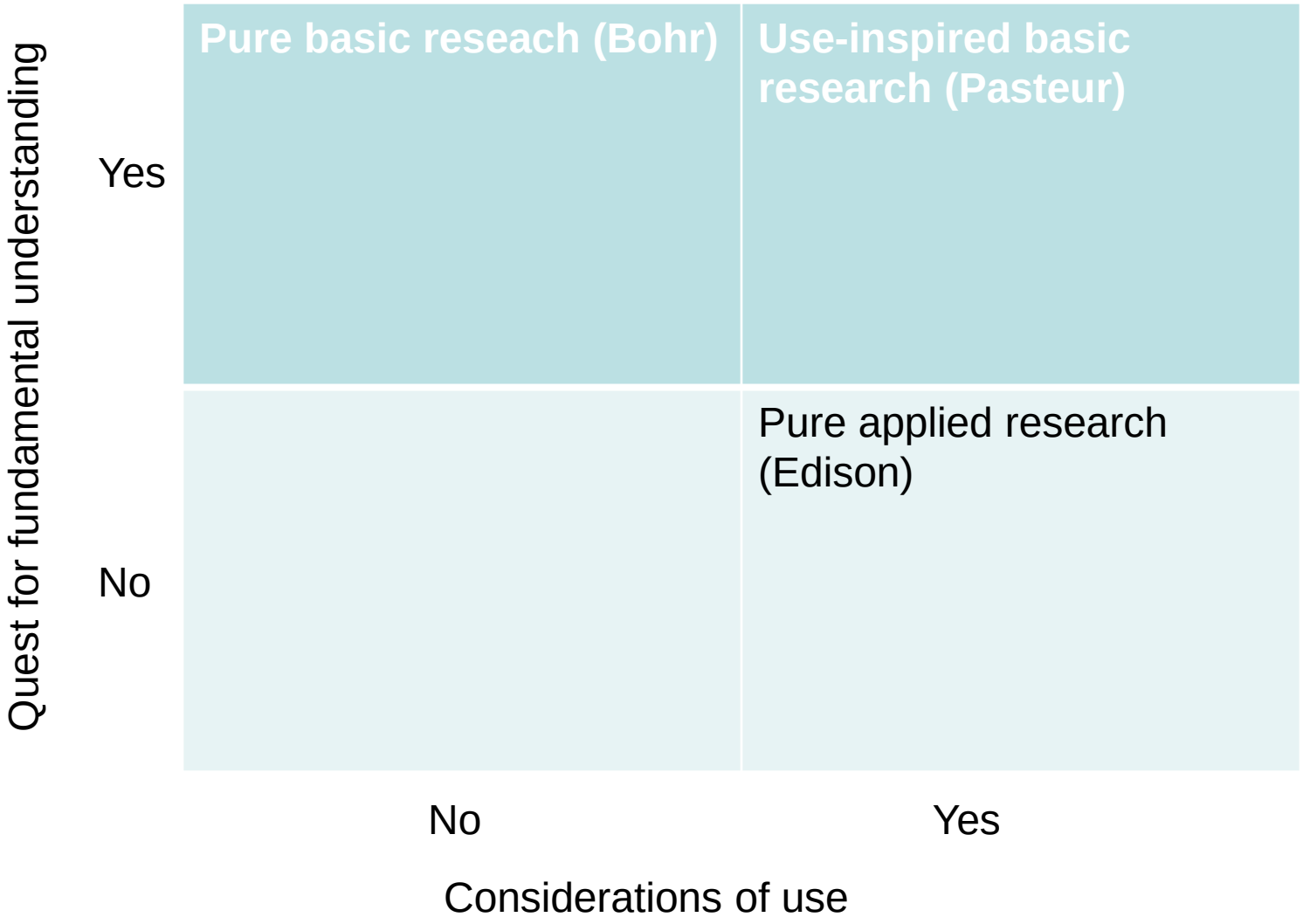
- Parsimony
- Maximum likelihood
- Nearest neighbour
- Bayes analysis

- Validation: Often boot-strapping

Important methods in chemometrics (regresssion)

- Regression
  - MLR (multiple linear regression)
  - PCR (principal component regression)
  - PLSR (partial least squares regression)
  - RR (ridge regression)
- Neural Networks

The different kinds of reseach



The scientific method

- Hypothesis
- Prediction
- Test and validation
- Repeat

The scientific method is a recursive system of matching theory with observation

A hypothesis is a tentatively held conjecture for the purposes of developing predictions of empirical observations

The scientific method (deduction)

- Discovery, observations, ideas, intuition, former results
- Propose a hypothesis and connect it with logical derivatisations from known theory and propose a mathematical model, prove by several experiments/observations (tests) and also try to disprove hypothesis (**deduction**: from the general to the specific)

The scientific method (induction)

- Gather many objects and measure by a series of features.
- Classify and find latent features.
- Predict by regression.
- Validate.
- Connect with known theory and set up hypotheses or set up experimental designs to find important features and their dimensionality
- (**induction**: from the specific to the general)

Abduction, deduction and induction

- Science often exhibits a subtle interplay between abduction, induction and deduction. Abduction is a common process of creating new generalizations, theories and hypothesis. Deduction takes a hypothesis to make a specific prediction. ”Then” induction is used to fit the evidence to the hypothesis.

Levels of knowledge

- One (few) example(s): ”Laymans science”
- Neural networks and validation
- X-metrics and validation
- Statistics and distributions
- Mathematical exact modelling
- Essentialism

Technology and science

”It’s alright in practice, will it ever work in theory?”

- Theory (Plato): Know what (clever people)
- Practice: Know how (skilled people)

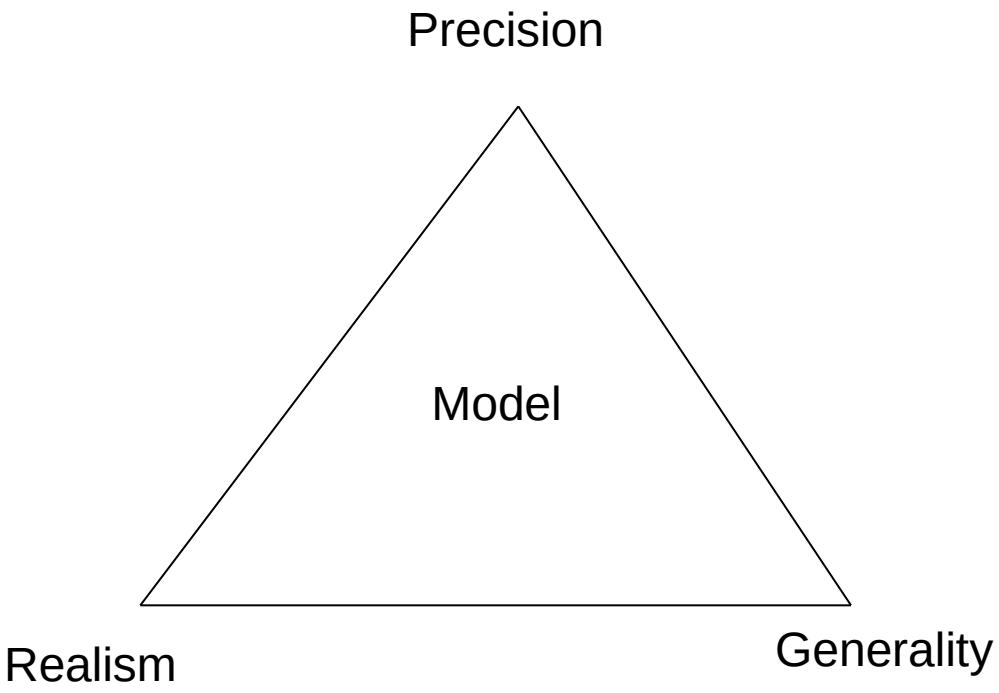
Advantages of technology (applied research)

- Holistic, not reductionistic
- Context driven, not subject driven
- Mission-oriented research, not ”blue skies”
- Team work, not individual scholar
- Divergent, not convergent thinking
- Decisive criterion: does it work?

Beware of pure reductionism

We must reject this primitive and almost cannibalistic delution about knowledge, that an understanding of something requires first that we dismantle it, like a child who pulls a watch to pieces and spread out the wheels in order to understand the mechanism” (Thom, 1975)

Models are not reality




Ceci n'est pas une pipe.

Systematic generalization

(hierarchical reductionism)

”We very soon got six yards to the mile. Then we tried hundreds yeards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to a mile!”

”Have you used it much?” I enquired.

”It has never been spread out, yet, ” said Mein Herr: ”the farmers objected: they said it would cover the whole country, and shut out the sunlight! So now we use the country itself, as its own map, and I assure you it does nearly as well”

(Lewis Carroll, Sylvie and Bruno, 1893)

Chemometrics and science?

- Find an important problem in your field of interest (FOI) to which there is yet no solution (think!). Propose a preliminary hypothesis.
- Observe and measure within the FOI
- Use statistical and multivariate design
- Propose a hypothesis (think!)
- Experiments and/or observations: tests and predictions based on proposed model
- Reject hypothesis or accept it for the time being

This course

- In this course you will have hands-on experience in how to treat data with a lot of features (variables) measured on several or a lot of objects

Learning objectives of the course

- Give an overview of important chemometric methods
- Identify situations where exploratory data-analysis is required
- Describe and use different forms of scaling, transformation and normalization
- Understand and describe the difference between classification and regression
- Understand and describe the difference between clustering and ordination
- Apply and interpret principal component analysis (PCA) on multivariate data
- Apply and interpret the principles of validation and outlier detection
- Use and interpret cluster analysis
- Describe, apply and interpret multiple linear regression (MLR) and ridge regression (RR) and where to apply them in two data matrix problems
- Describe, apply and interpret principal component regression (PCR) and partial least squares regression (PLSR) and where to apply them in two data matrix problems
- Apply and interpret correspondence analysis
- Describe the method metric multidimensional scaling

Programs used

- R, R-Studio (free software)

- NTSYS (Exeter publishing) version 2.2
  - A whole package on the methods used most frequently: used earlier in this course
  - (http://www.exetersoftware.com)(USA)

- UNSCRAMBLER version 10.3 (CAMO, Norway)
  - (you can buy your own version (but it is expensive)) (http://www.camo.com)

Book used

- Lattin, L, Carroll, J.D., Green PE:
- Analyzing multivariate data, Thomson, Pacific Grove, CA, USA, 2003, 556 pp.
- Recommended: E-boks on R

- + a little extra reading material, especially Romesburg (1984) on cluster analysis