Some Perspectives on the History and Sociology of the Chemometrics Revolution - and some Suggestions for What the Future Holds

Prof. Steven Brown
Department of Chemistry and Biochemistry
University of Delaware, Newark, DE 19716 USA

Chimiométrie 2019
Montpellier, FRANCE
31 Jan 2019

“History does not repeat itself, but it does rhyme.”

–Mark Twain (from The Celebrated Jumping Frog of Calavaras County)

Goals of this Presentation

I'll present:

- A personal, (opinionated) survey of the history of chemometrics
- Some indication of sociological issues that directed changes in the field
- A short assessment of future directions in measurement chemistry and why these will cause changes in chemometrics/ data science
- A brief summary of the skill sets needed to succeed in chemometrics in the future

All with only a few equations, no real theory and almost no chemistry…

The Statistics Side of Chemometrics

The Statistical World

Dr. William S. Gossett ("Student")

BSc in analytical chemistry, Oxford
Worked at Guinness Brewing
Later obtained PhD in statistics from K. Pearson

Father of
- t-test
- studentized residuals (with K. Pearson)

Ignored by statisticians until Fisher "explained" the t-statistic in 1930

Source: Wikipedia

The Statistical World

Prof. Karl Pearson

Professor of Statistics at University College, London

Studied mathematics, law, and history (and other subjects) at Cambridge and Berlin
He initially took a faculty position in Germanics at Cambridge

father of
- Biometrika journal
- mathematical statistics
- first statistics department (UC London)
- biometrics field (mathematical genetics)
- metrology

- Egon Pearson

Arch-rival of Prof. Sir Ronald Fisher

Source: Wikipedia

The Statistical World

Prof. Egon Pearson

Prof. of Statistics at University College, London

Editor of Biometrika

Father of

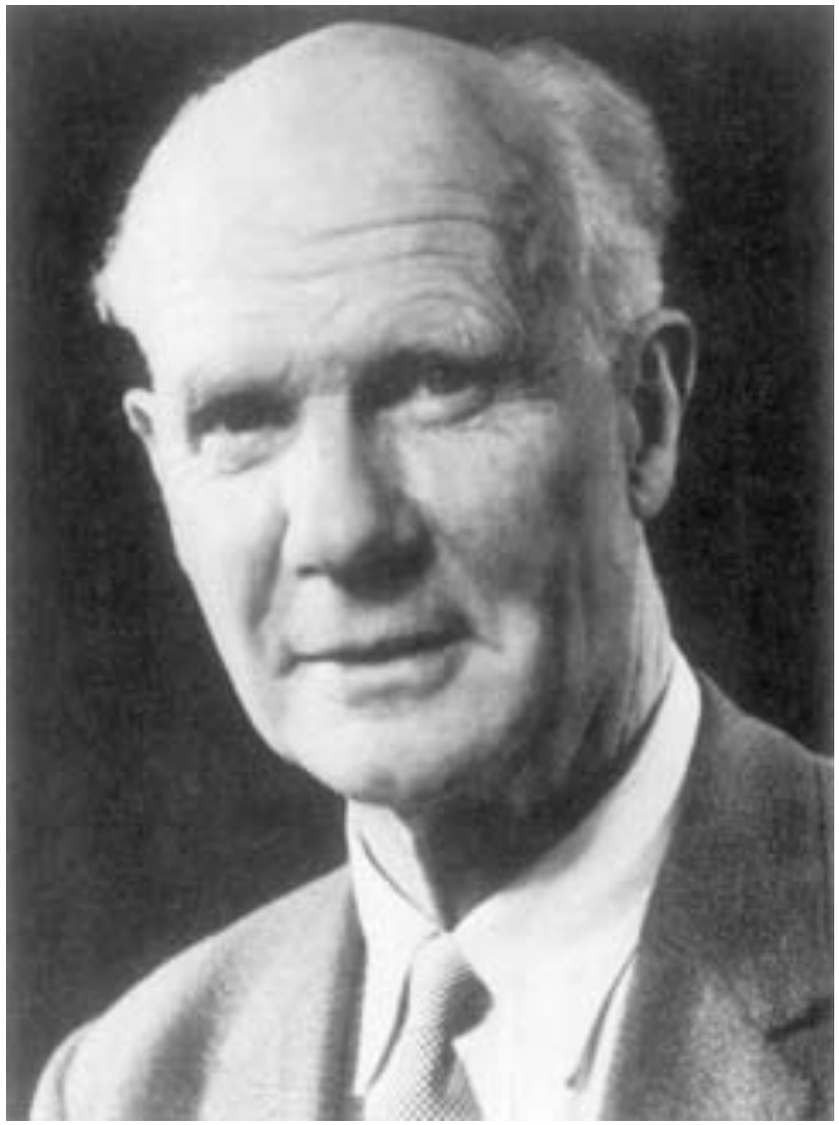
-Neyman-Pearson hypothesis testing

-Quality control and optimization methods

- Julia Pearson Box

Students:

-George Box



Source: Wikipedia

The Statistical World

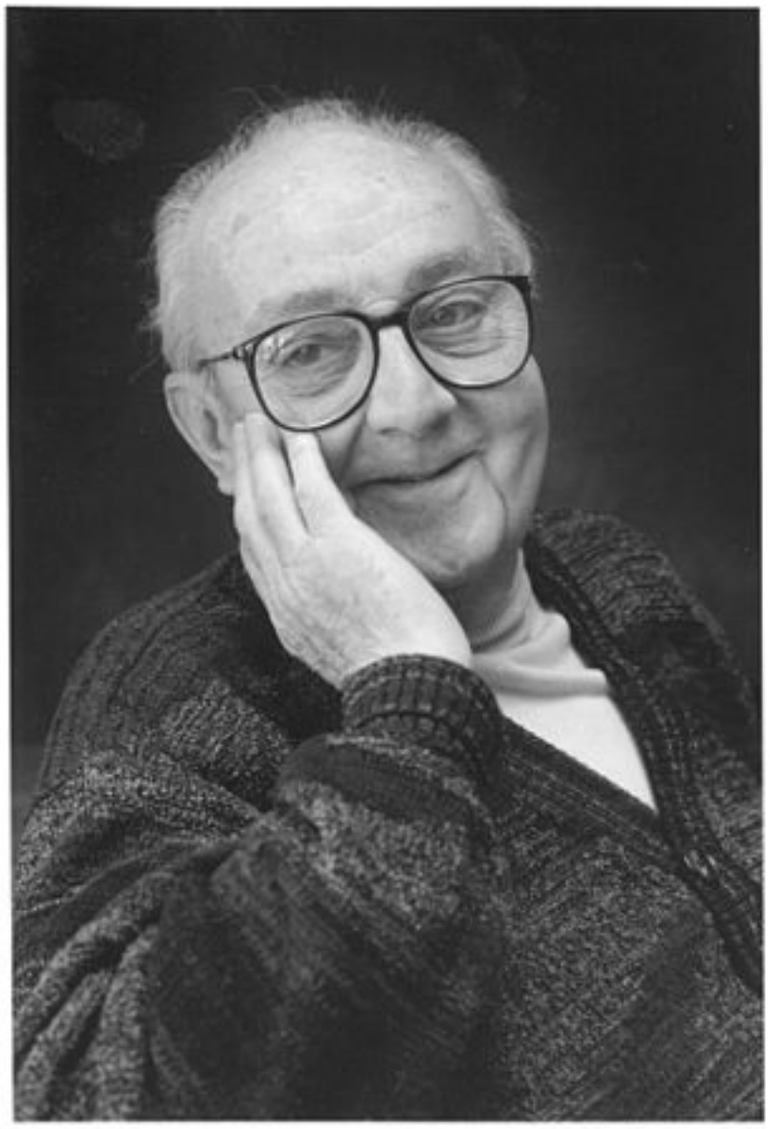
Prof. George H. Box

Chemist and (reluctant) statistician

BSc in Chemistry
Served in UK chemical corps in WWII, saw the value of designed experiments

Obtained PhD in statistics from E. Pearson

Became statistician at ICI, later Professor of Statistics at Univ. Wisconsin



Source: Wikipedia

The Statistical World

Prof. Herman Wold

statistician and econometrician, Univ. Uppsala

Member of Swedish Royal Academy

Member of selection committee for Nobel Prize in Economics

- the father of

(1) the Wold decomposition of stationary time series

(2) Projections to Latent Structures/Partial Least-Squares

(3) path modeling methodology

(4) Svante Wold

Source: Uppsala University

The Statistical World

Svante Wold

PhD in organic chemistry from Umea University. Post-doctoral position with George Box.

Took faculty position at Umea University, started research in physical organic chemistry

Used the branding term “chemometrics” on a grant application in 1972

First paper mentioning the term “chemometrics” was published in 1973

Co-founded the International Chemometrics Society in 1975.

Developed SIMCA algorithm with M. Sjöström in 1977

Began using PLS in 1983-5

Co-founded Umetrics to sell SIMCA software and other services in 1977.

Students included:

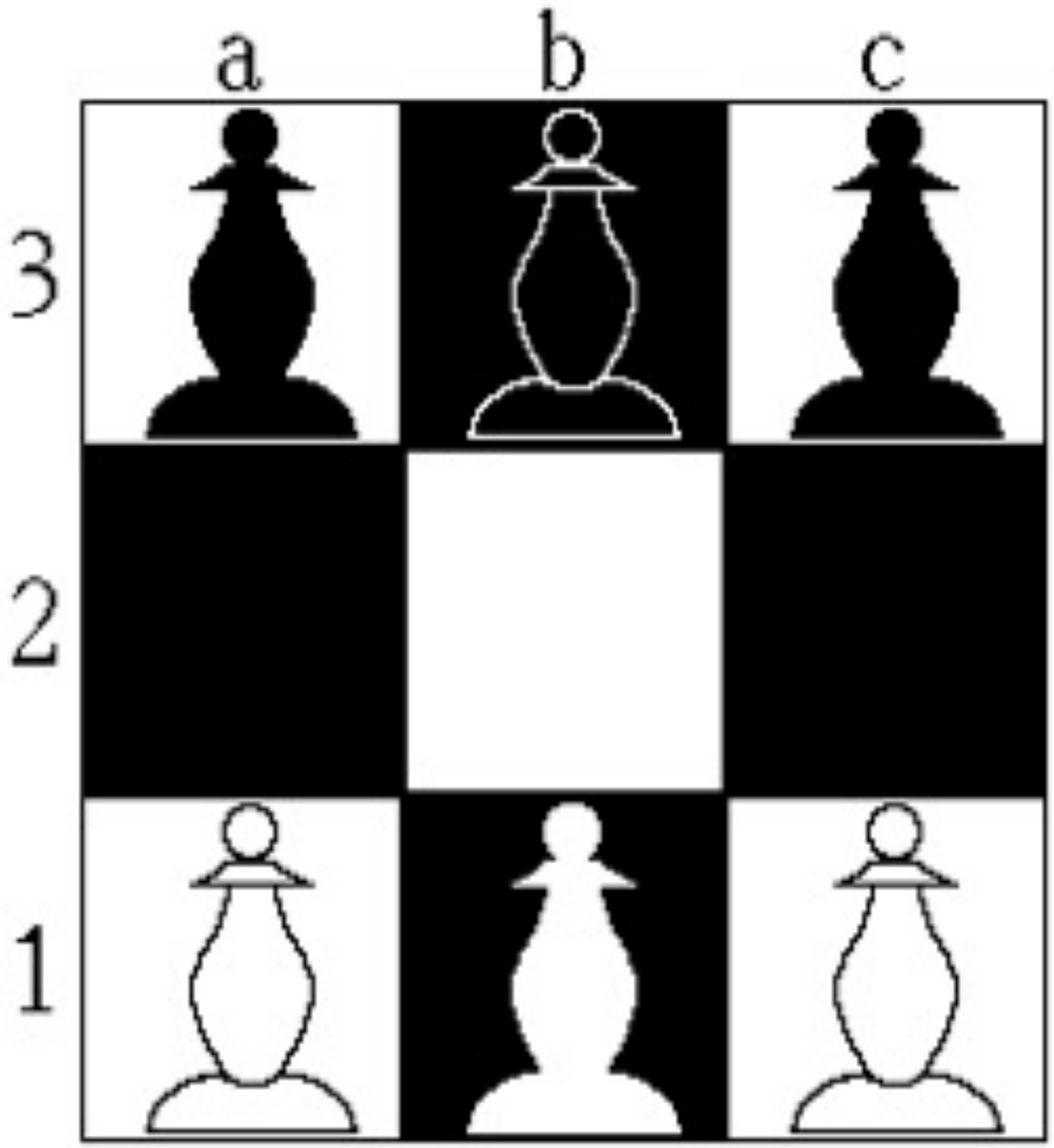
J. Trygg

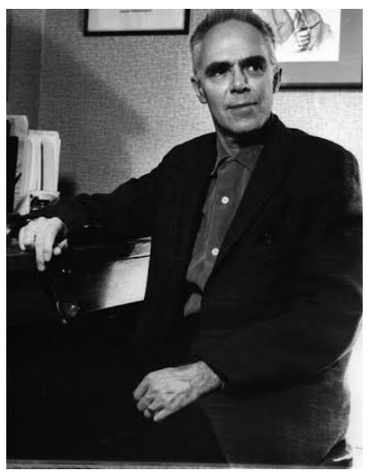
P. Geladi

Source: Umea University

The Heuristics Side of Chemometrics

The Heuristics World





Hexapawn Image modified from Gardiner, M. **Scientific American** 1962; March: 138-144.

Heuristic Classification

The linear discriminant function is

$$s = \mathbf{w} \cdot \mathbf{x} = ||w||\,||x|| \cos \Theta$$

This linear discriminant describes a direction in the (p+1) space, because Θ defines the angle between **x** and **w**.

For $Q < 90^o$, cos Θ is positive, while for $\Theta > 90^o$, cos Θ is negative. Thus, the problem of classification reduces to one of finding the equation of the separator for the two categories 1 and 2.

The classification process is based on iterative location of **w,** with negative feedback:

1) Calculate a trial discriminant s, which determines the category membership.

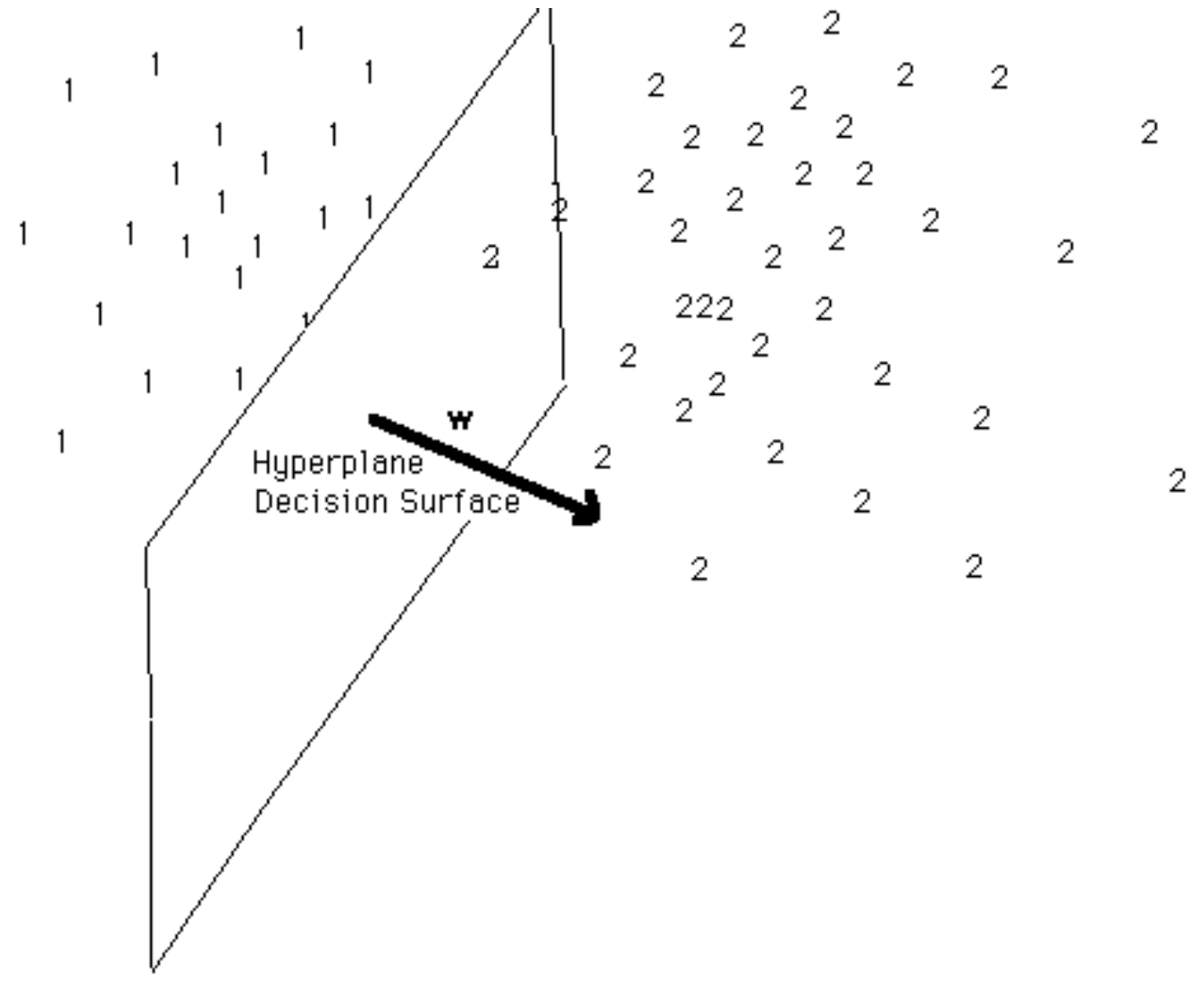
2) If s is correct, the trial discriminant vector **w** is left unchanged, but if the predicted category is wrong, **w** is altered

$$\mathbf{w} = \mathbf{w} + c\,\mathbf{x}$$

where the correction factor c reflects the discriminant vector w by the same amount that is was in error.

$$c = \frac{-2s}{\mathbf{x} \cdot \mathbf{x}^T}$$

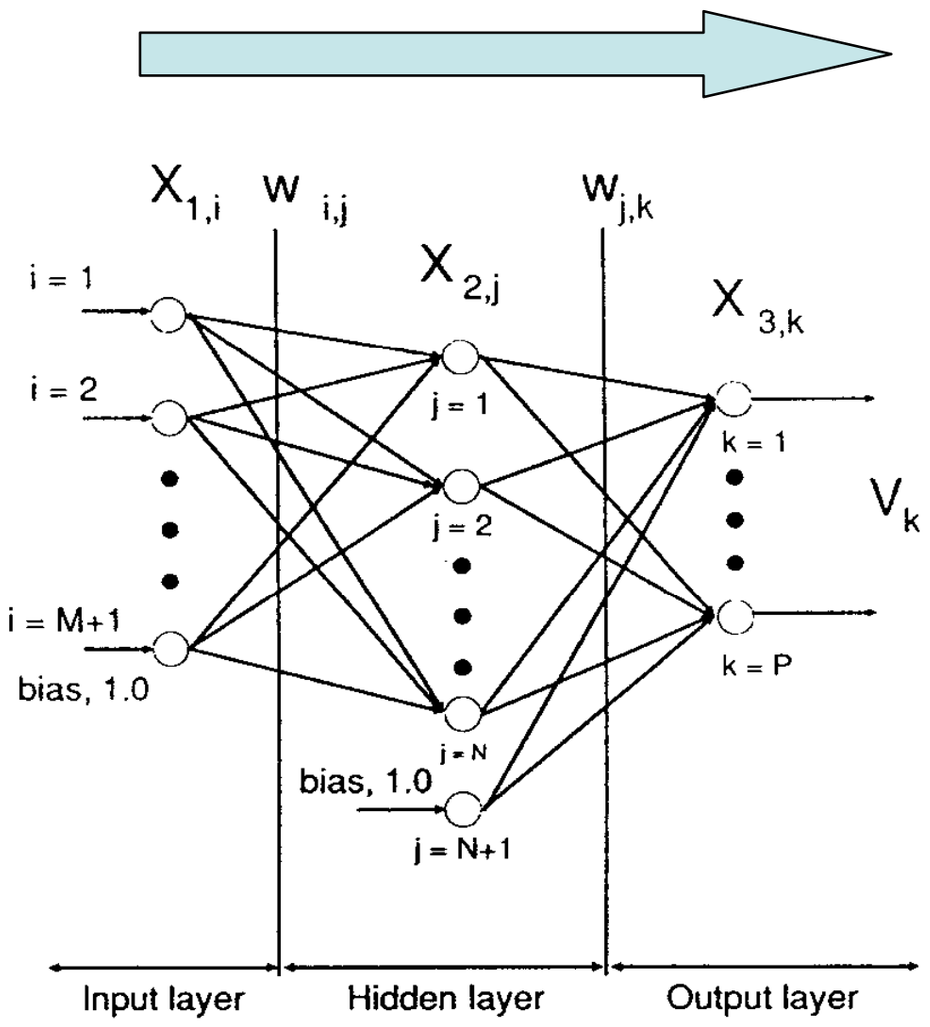
3) This process is repeated until all patterns are classified correctly.

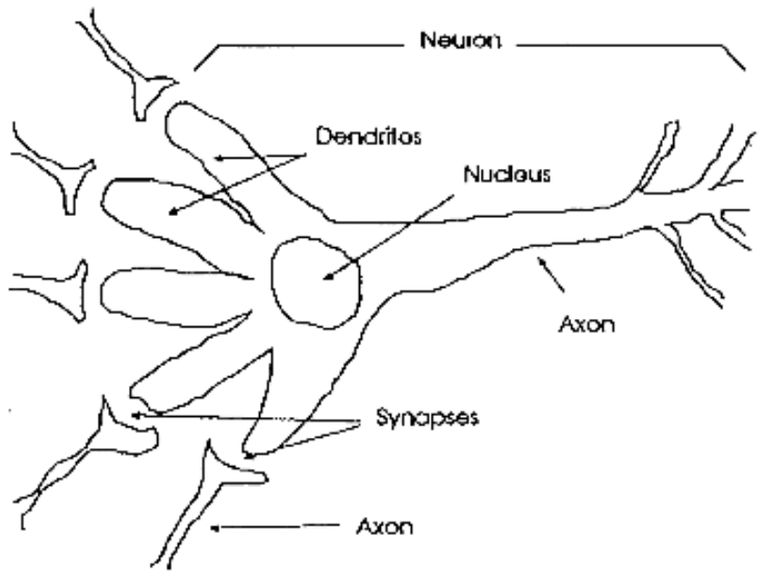


Classification of Categories 1 and 2 by the Linear Learning Machine

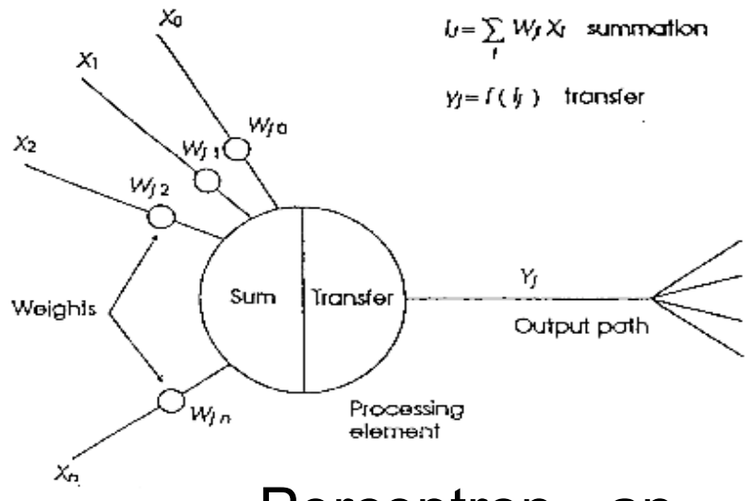
Neural Network Classifier

data flow





Multi-layer, feed-forward perceptron classifier – an Artificial Neural Network (ANN)



Perceptron - an artificial neuron

The Heuristics World

Source: Brian Rohrback

The Heuristics World

Prof. Thomas Isenhour

PhD in analytical chemistry from Cornell

Took faculty position at Univ. Washington
Mass spectroscopist - he *opposed* Kowalski's chemometrics research at first - later saw the value and changed his focus to expert systems and linear learning machine heuristics

Co-directed change in J. Chem. Inf. Comp. Sci. and ACS COMP division to reflect mathematical methods in database searches, etc.- *but not measurement-oriented chemometrics*

Did not participate in any chemometrics conferences or publish in any chemometrics journals, and soon became an administrator.

Graduated:

B.R. Kowalski
P.C. Jurs
P. deB. Harrington

Source: Wikipedia

The Heuristics World

Prof. Bruce Kowalski

BS in mathematics and chemistry, Millikan University
PhD in Analytical chemistry, University of Washington

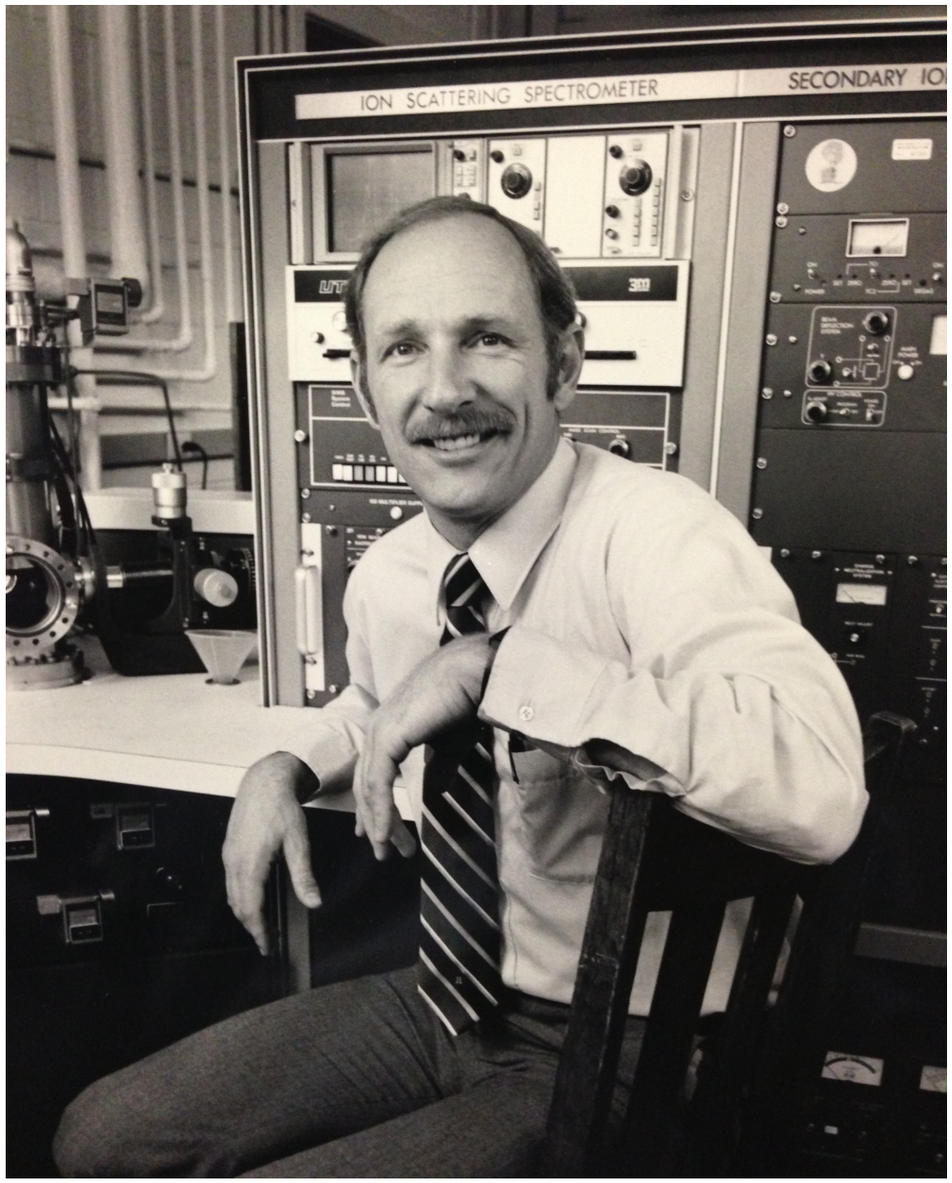
Co-founded the International Chemometrics Society
Organized the NATO ASI in Chemometrics
Co-founded the Journal of Chemometrics
Co-authored several texts on chemometrics
Co-founded Infometrix

Founded the Center for Process Analytical Chemistry (CPAC)
First named professor at UW

Father of
-NAS methods (with K. Booksh)
-multiway methods (with E. Sanchez)
-heuristics in chemistry (with C. Bender)
-multi-algorithms in chemistry for “big data” (> 16 kb)

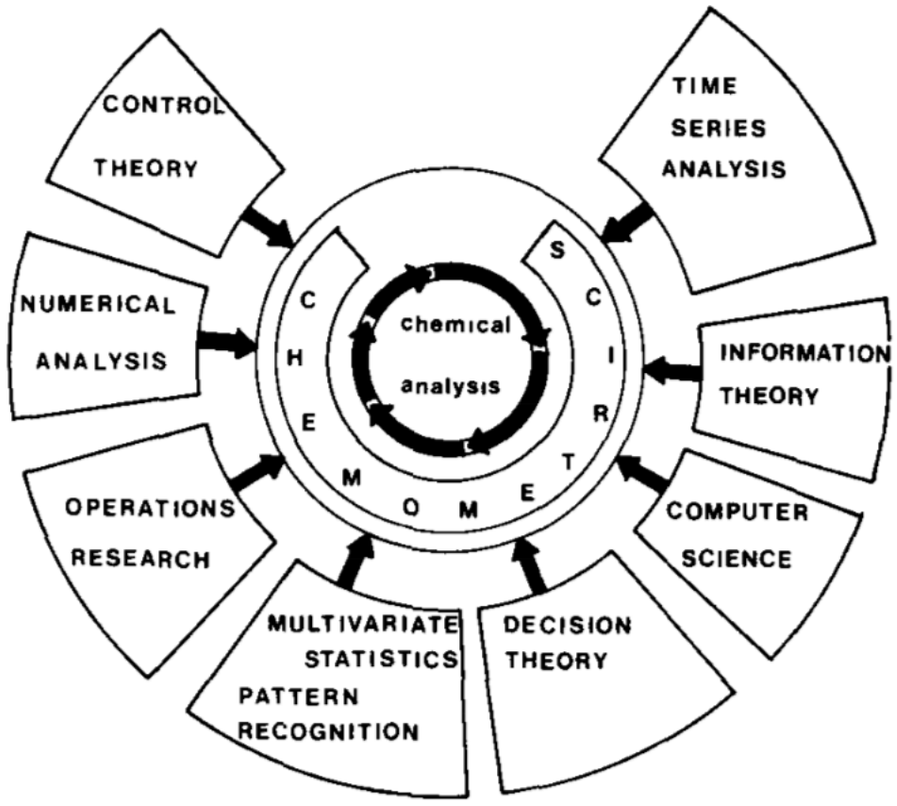
Students included:

-D. Duewer
-S. Brown
-K. Booksh
-J. Kalivas
-M.B. Seasholz



Source: University of Washington

Chemometrics as an Interface

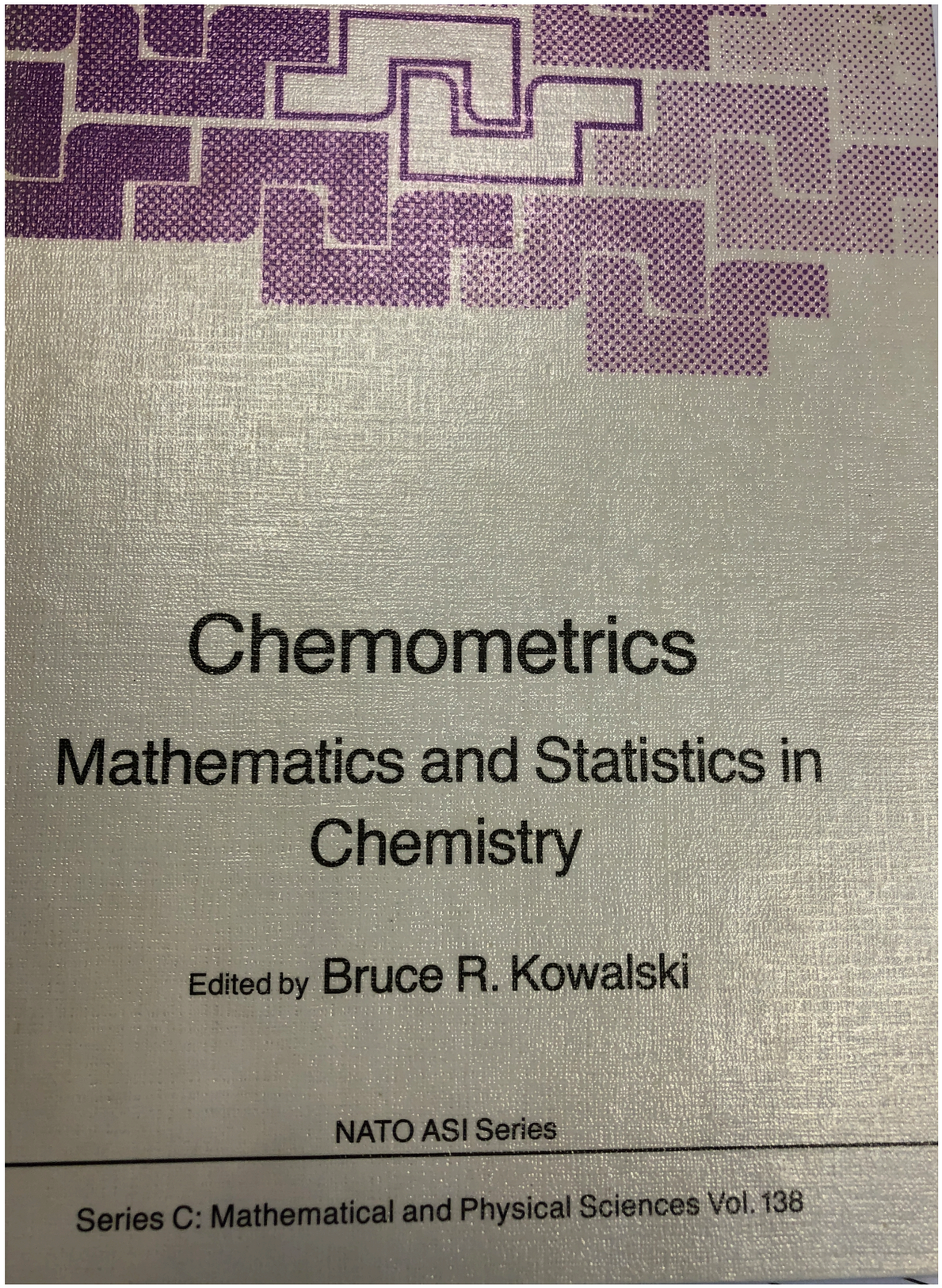


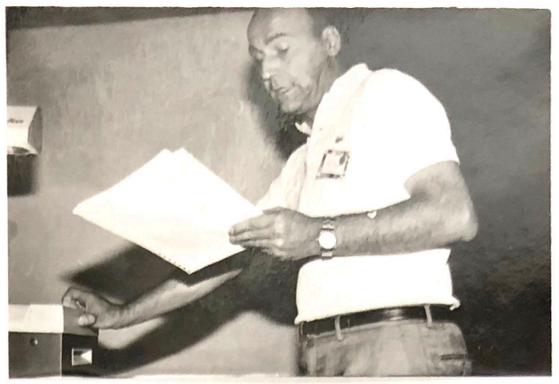
Source: Vandeginste, B.G. Teaching chemometrics, Analyt. Chim. Acta. 150 (1983) 199-206.

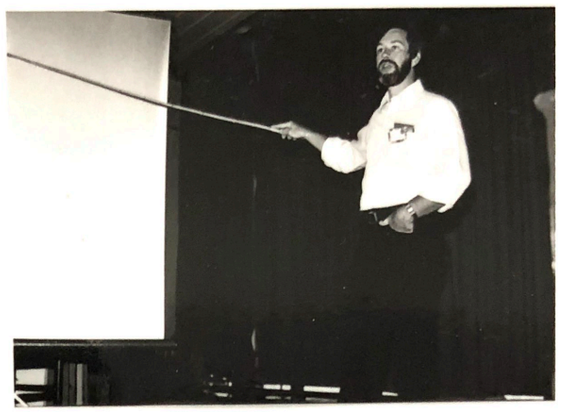
Statistics and Heuristics Meet

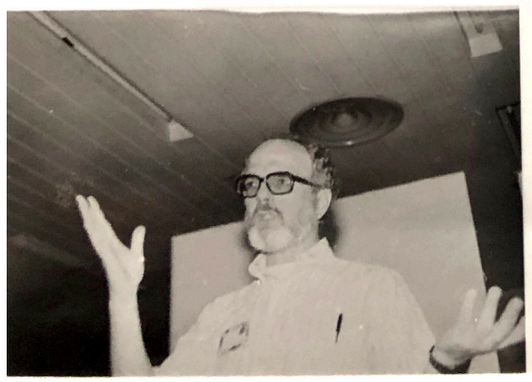
2-week NATO Advanced Study Institute
Università della Calabria
Cosenza, Italy
September, 1983
Organizer: B.R. Kowalski

Experimental Design - S. Hunter and B. Hunter
ANOVA - G. Latorre
Spatial Analysis - J.C. Davis
Detection theory - L. Currie
Filtering of Noise - H.C. Smit
Sampling - G. Kateman
3-D Graphics - S. Grotch
Control of Chemical Processes - L. Ricker
Optimization - S. Deming
Linear Models and Matrix Least Squares - S. Deming
Data Analysis in Food Chemistry - M. Forina and S. Lanteri
Cluster Analysis - L. Kaufman and D.L. Massart
Multivariate Analysis with SIMCA - S. Wold, et al.
Multivariate Calibration - H. Martens and T. Naes
Teaching Chemometrics - B. Vandeginste



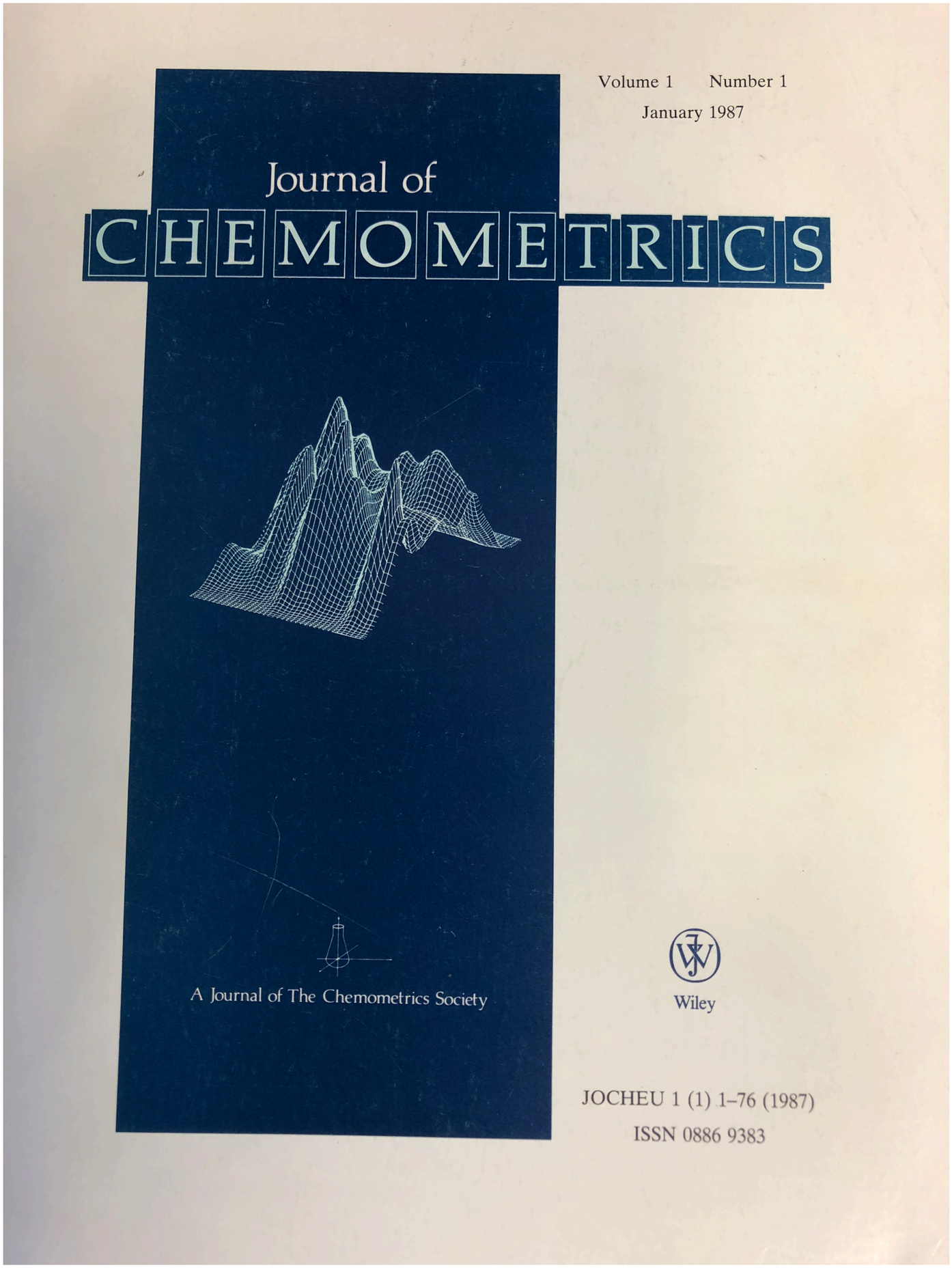


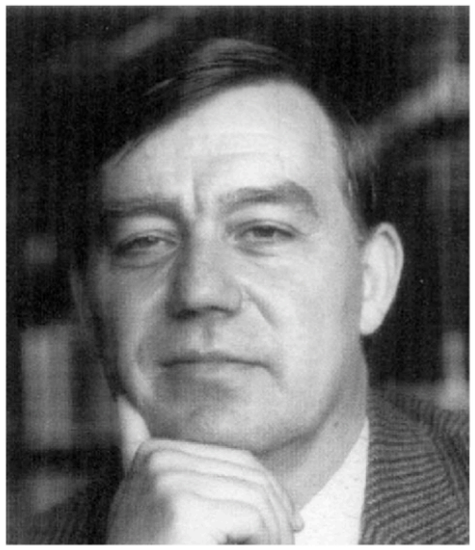




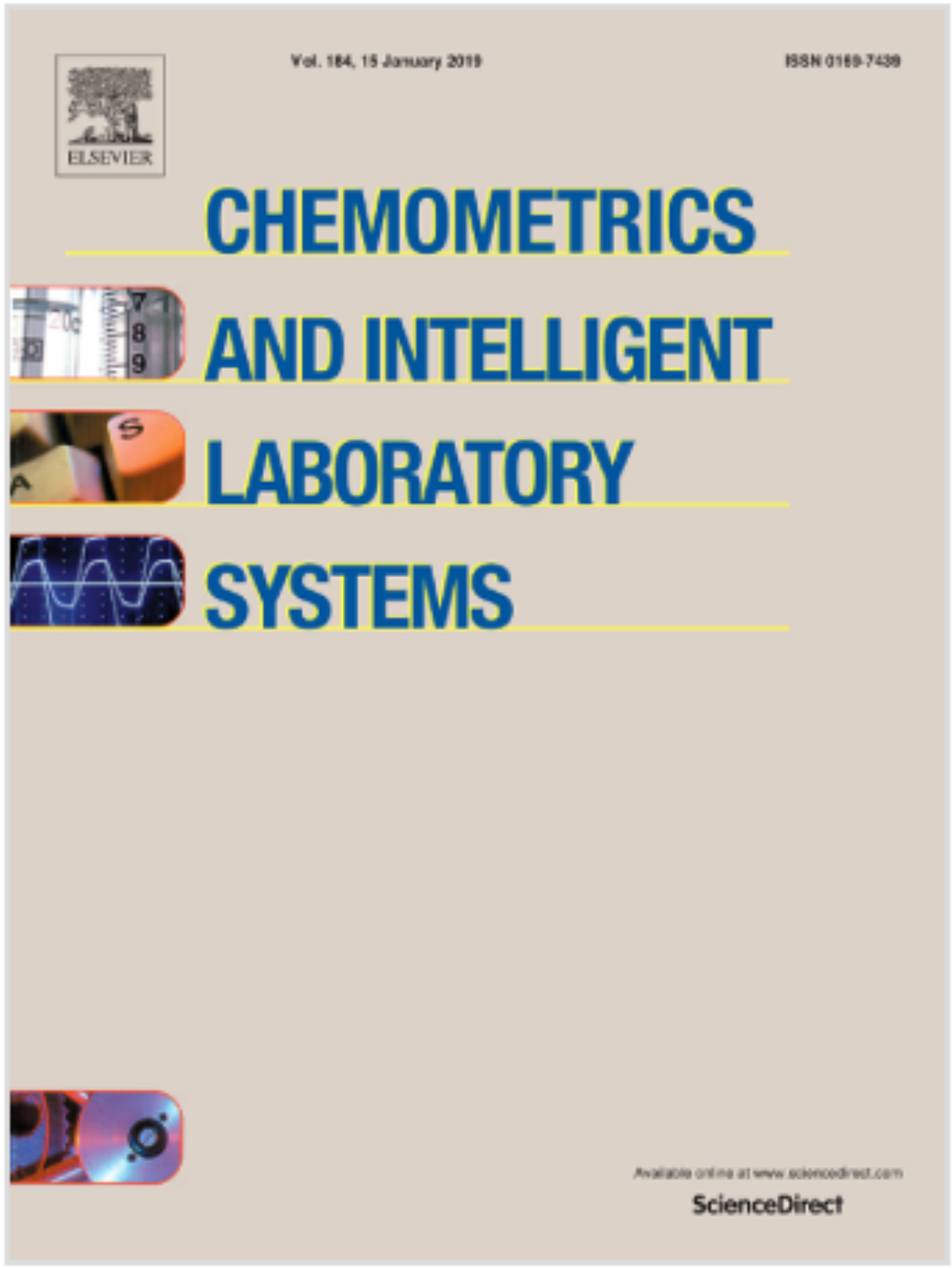
Photos courtesy of Paul Geladi

Two Chemometrics Journals


Volume 1 Number 1
January 1987
Journal of
CHEMOMETRICS
A Journal of The Chemometrics Society
Wiley
JOCHEU 1 (1) 1–76 (1987)
ISSN 0886 9383

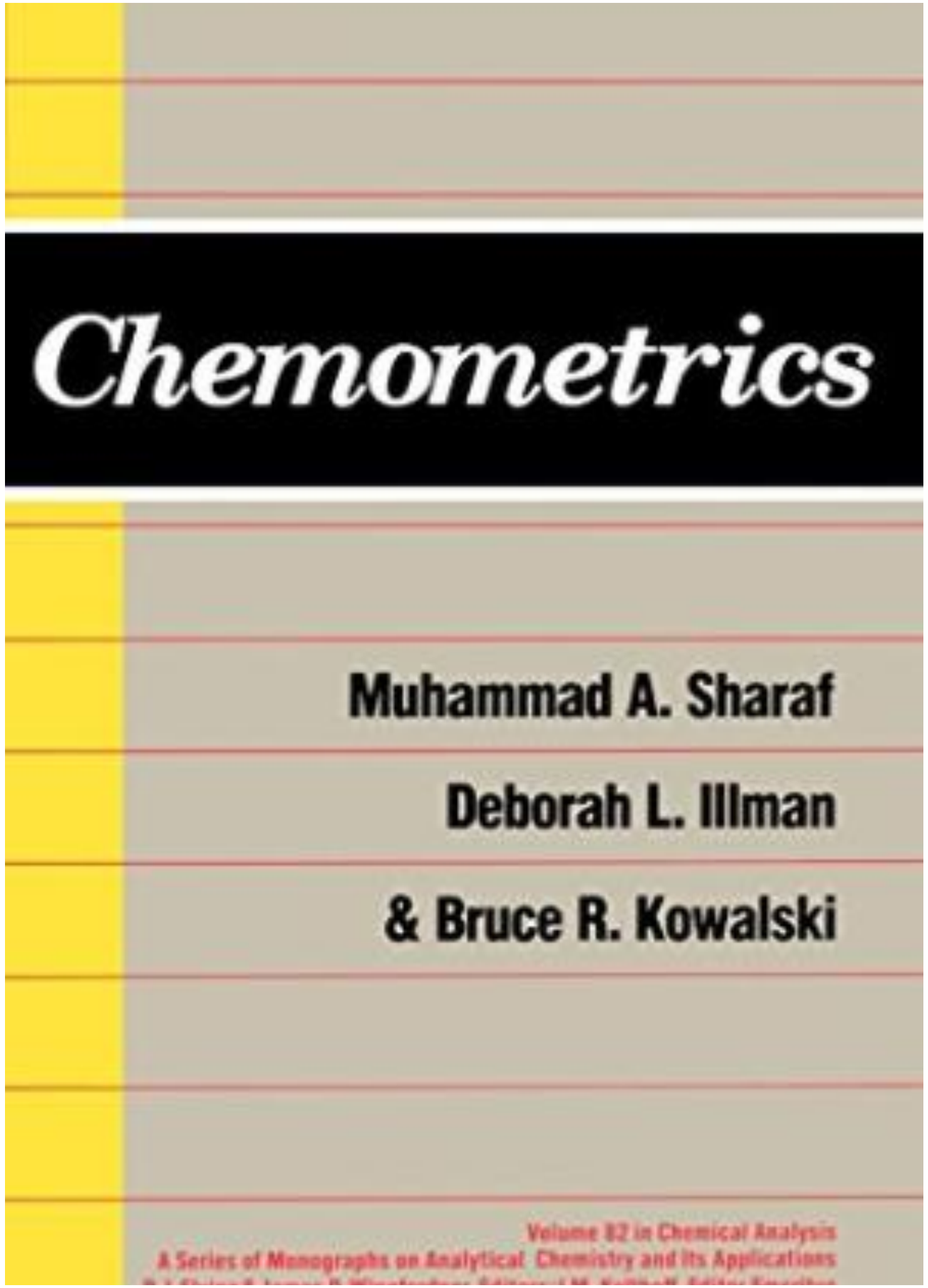


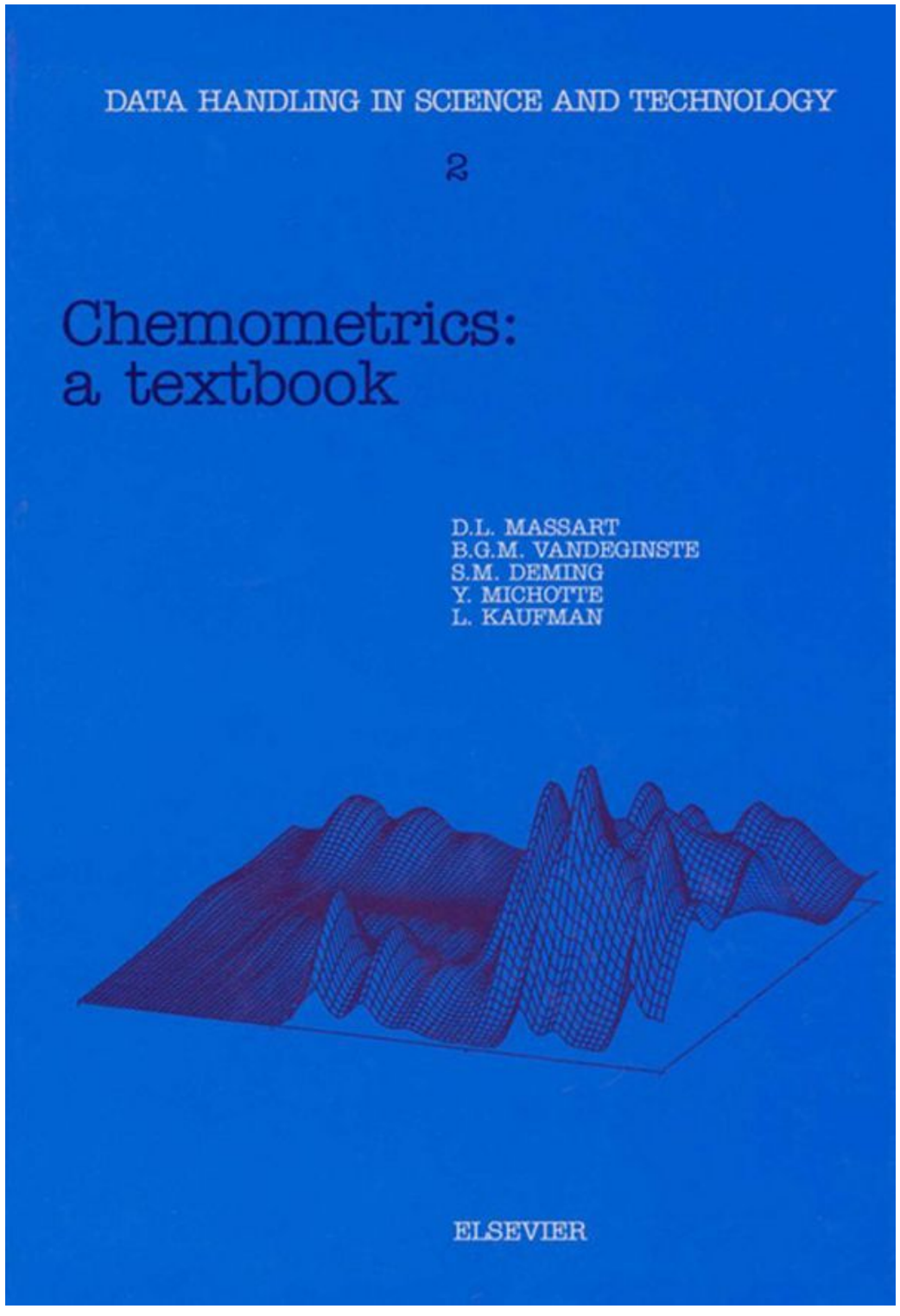
Source: Vandeginste, B.
Obituary: D. Luc Massart,
Chemom. Intell. Lab. Syst. 81 (2006) 1-2.

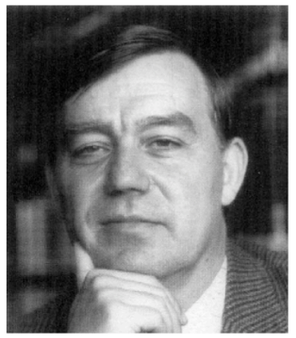


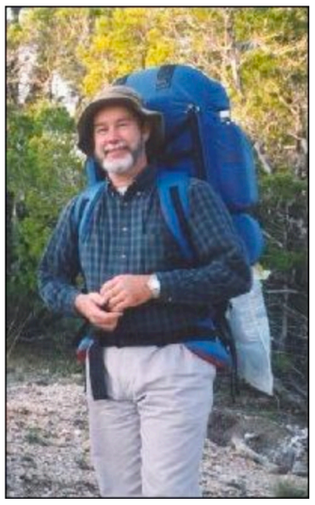
Two Chemometrics Books

Source: Wikipedia









Source: Wikipedia, Deming Personal Page, Chemolab

Difficulties with Statistics

Latent variable methods caused friction with statisticians, especially with George Box, Jerry Sachs, and others

These and other classical statisticians left a Gordon Conference (~1993) when a speaker (Aloke Phatak) began a presentation about latent variable methods.

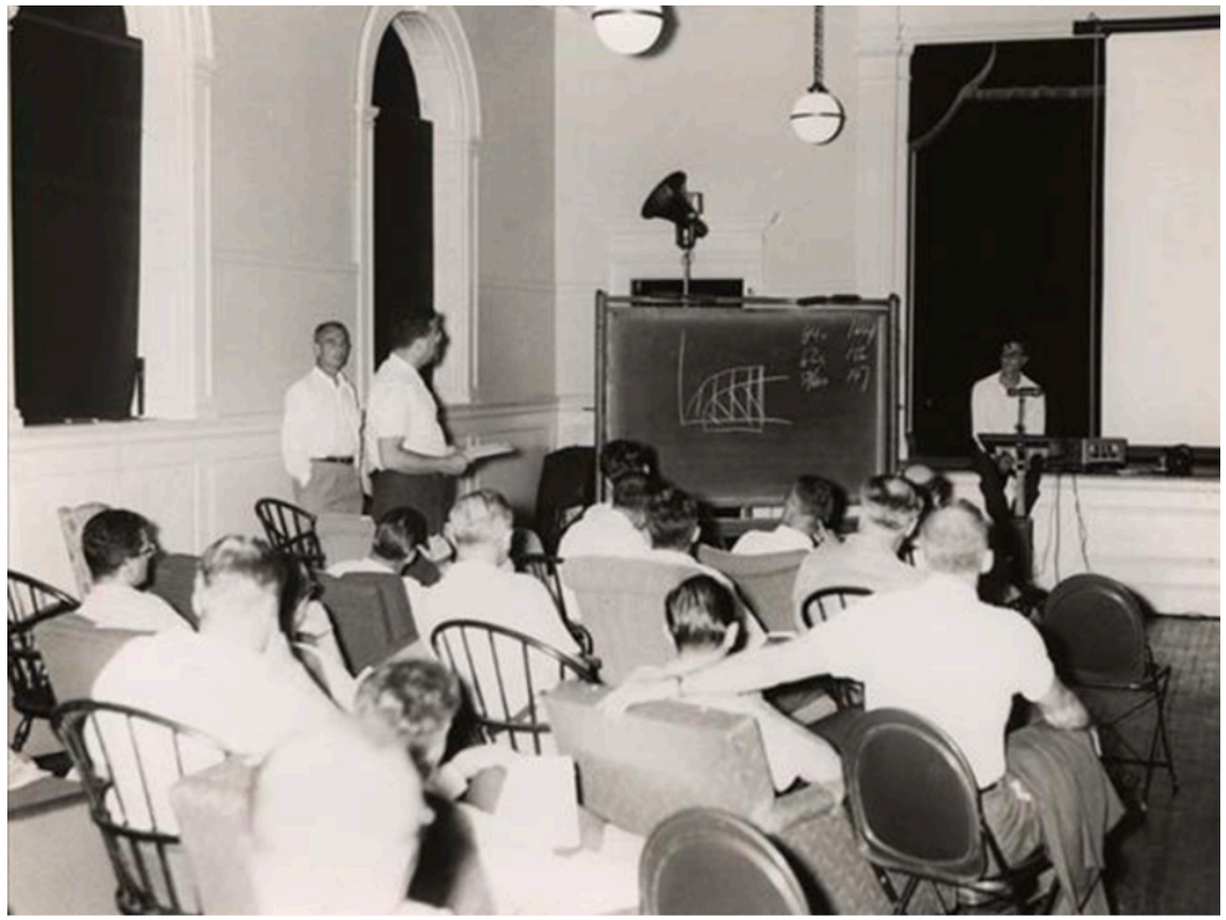
Statisticians were troubled about:

-lack of rigor (theory) for latent variable methods

-fast review in chemometric journals

The Statistics in Chemistry and Chemical Engineering GRC ended in 2006.

Statisticians soon “explained” PLS and other latent variable methods, but some remain troubled by “ad hoc” machine learning methods of Breiman and Tukey.



Source: Gordon Research Conferences

“Machine learning is statistics minus any checking of models and assumptions.”

– Prof. Brian D. Ripley (talking about the difference between machine learning and statistics) useR! Wien (May 2004)

Difficulties with Funding

- In 1980-2000, analytical chemistry and statistics were experiencing trouble - they were challenged as not very “pure”

- Academic departments began to drop these areas of research

- Funding for doing analyses or for data analysis was not available - many practitioners of “applied data analysis” relocated to food science, agriculture, or forensics

- Chemometrics largely missed out on participating in the genomics revolution in bioinformatics where the focus was on heuristics and classification, not on regression

Data Science, Big Data and Predictive Analytics

Recently, the situation for chemometrics and all areas of data science changed - for two big reasons:

- “Big data” became available through improved internet capabilities and it became desirable to explore these sets using machine learning modeling of “large” (10-100 Gb) to very large (10 Tb ++) data sets using parallel computation on GPU sets to discover trends or other hidden features

- “Predictive analytics” involving use of machine learning heuristics to develop predictive models for continuous quantities and for categorical quantities where first-principles models did not work

Now, the field of data science and research in the domain areas involved in data science are both in high demand.

But, this area is moving incredibly rapidly, and developments appear almost daily.

What's Next?

Machine learning heuristics, of course!

Why will we need Machine Learning?

New instrumentation - or combinations of established types of instruments - will become common.

Data and model fusion will become the norm in measurement-based chemistry as it already has in many other fields.

Generating huge amounts of data will also become more common as the new capabilities and instruments lead to new questions.

Classical chemometric methods are not sufficient to deal with these new kinds and amounts of data.

Classical chemometric practice - making a “toolbox” or a new method and showing it on a few data sets - will be less important. Software will become ubiquitous, especially for machine learning.

What’s Next?

“(Statistics) ought to be concerned with data analysis. The field should be *defined as a set of problems* (as are most fields) *rather than as a set of tools*, namely those problems that pertain to data.” (emphasis added by SDB)

- Tukey, J. The future of data analysis, Ann. Statist. 33 (1962) 1-67.

Where will Machine Learning be Used?

Like it or not, machine learning methodology will be used to address new chemical problems arising from the use of new measurements and sensors, for example:

- Nonlinear modeling
- Complex classifications
- Spatial and temporal analyses
- Modeling of fused, large data sets
- Non-quantitative, predictive applications

What Skills are Needed?

1**. Machine learning/computational statistics/research skills:**

- Math skills, including multivariate calculus and (mathematical) analysis, matrix algebra, statistics, differential equations, numerical analysis
- Programming skills, with expertise in **2-5** computer languages

2. Domain Knowledge:

- Strong skills in chemical metrology
- Deep knowledge of the physics/biology behind analytical measurements

3. “Soft” skills:

- Documenting and verifying research work
- Working in teams
- Working with “customers”
- Writing reports and presenting work at different technical levels to different audiences


CAC-2006
Bruce Kowalski
UNITED STATES

Acknowledgements

I thank Marcia Ferreira and Brian Rohrback for contributing their and others' images. I appreciate the images sent from the ASI by Paul Geladi.

Comments and remembrances from Maynarhs daKoven, Esq. were helpful in preparing this talk.

Prof. Brown's research is supported by the United States National Science Foundation, under Grant 1506853.

Source: ARTHUR Manual, 1977

“In matters controversial,
My perception’s rather fine.
I always see two points of view:
The one that’s wrong -
And mine.

See you later,
A.R.T.H.U.R.”

– ARTHUR, 1977

Some Additional Reading

Breiman, L. Statistical modeling: The two cultures. Statistical Science. 16 (2001) 199-215.

Geladi P.; Esbensen, K. The start and early history of chemometrics: Selected interviews. J. Chemometrics 4 (1990) 337-354.

Geladi P.; Esbensen K. The start and early history of chemometrics: Selected interviews, part 2. J. Chemometrics 4 (1990) 389-412.

Brown,S.D. Has the chemometrics revolution ended? Some views on the past, present and future of chemometrics, Chemometrics and Intelligent Laboratory Systems. 30 (1995) 49–58.

Brown, S.D. Information and data handling in chemistry and chemical engineering: the state of the field from the perspective of chemometrics, Computers & Chemical Engineering. 23 (1998) 203–216.

Brown, S.D. The chemometrics revolution re-examined, J. Chemometrics. 31 (2016) e2856–23. doi:10.1002/cem.2856.

Friedman, J.H. Data mining and statistics: What's the connection? Proceedings, 29th Symposium on the Interface Between Computer Science and Statistics (1998). [Available from docs.salford-systems.com]

Brereton, R.G. A short history of chemometrics: A personal view, J. Chemometrics. 28 (2014) 749–760.

Ziliak, S.T. Guinnessometrics: The economic foundation of “Student’s t, J. Economic Perspectives. 22 (2008) 199–216.

Lindsay, R.K.; Buchanan, B.G.; Feigenbaum, E.A.; Lederberg, J. Applications of Artificial Intelligence for Organic Chemistry. The DENDRAL Project. McGraw-Hill: San Francisco, CA (1980). [Available from https://profiles.nlm.nih.gov/ps/access/BBALAF.pdf]