Interesting MV case stories in the European Pharmaceutical, Paper and Pulp, and other industries.

Svante Wold, Umetrics Inc., talk at MACC symp., Hamilton, Ont., Oct 3, 2000

1. Introduction – the current situation, trends, objectives
2. Potential value in multivariate data
3. Tools (briefly and simply)
4. R & D – examples, values, … **Autos, PaP, Pharma, …**
   - Masses of data
   - Chemical Structure ↔ Properties, Biological Activity, Tox, Environm.
5. Processes – examples, values, investment, conclusions
   - **Paper and Pulp, Minerals, Oil, Chemicals, Polymers, Autos, Semiconductors, Pharmaceuticals ….**
6. Conclusions – new technologies = threats & opportunities

1. Two strong trends – how to deal with them ??

1. More demands – create more value !!!
   - higher throughput, higher yield, less personnel, ….
   - better quality
   - lower costs
2. More data – “information explosion”
   - controllers – in more places, more frequent measurements
   - sensors – in more places
   - Spectroscopy, chromatography, images – thousands of variables
   - HTS, Combinatorial Chemistry (Pharma), ….
3. Using the potential in multiple data will create value.
   - needs some investments
   - return is very good
   - threat (ignorance & chaos) and opportunities (insight & value)

The ladder of value creation

Measurements (WDWW)

Data

Information

Decision

Action

Value

UMETRICS

*1987

Software

Training

Consulting

Implementation

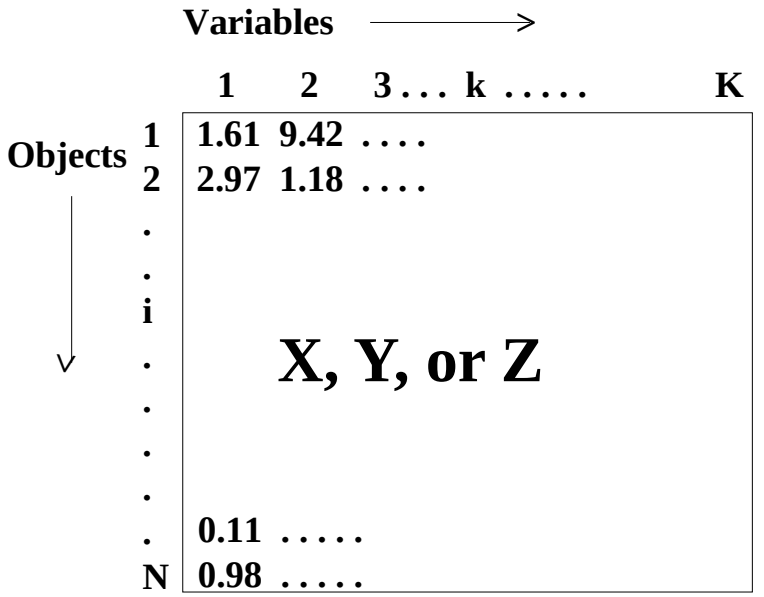
2. Investigation of complicated relationships/processes involves the generation and analysis of ***tables of data***

Objects

- Process time points
- Analytical samples
- Experimental trials (runs)
- Reactions
- Compounds
- Individuals, …..

Variables

- From spectra (NMR, NIR, MS, …)
- From separation (GC, HPLC, EF, …)
- T, P, flows
- Other; kinetic, quantum mechanical, structure descr. (size, lipophil, …), any curve form



Data table -- matrix -- denoted by **X**, **Y**, **Z**, ….

2b. Potential Values in Multivariate Data

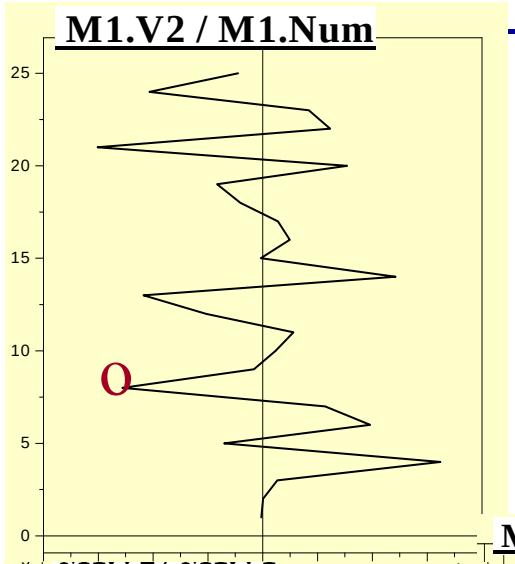
- Diagnostics, patterns -- good / bad -- where ? – why ?
- Multivariate relationships, forecasts, optimization.

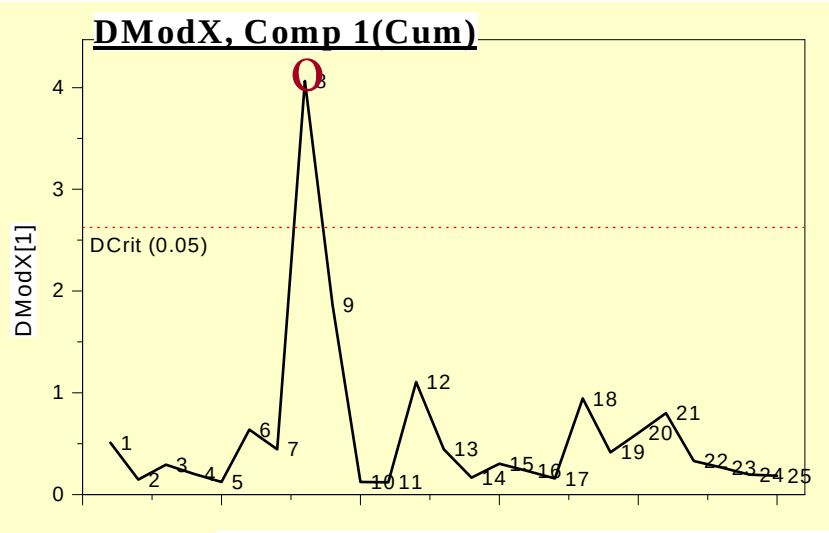
- Better use of research & development data (Automobiles, Chemicals, Pharma, Pulp and Paper, Semiconductors, ….)
  - more reliable conclusions, faster results
- Feedback to process – any continuous or batch process
  - fast on-line analysis, speed up, removal of bottle necks
  - process monitoring, decreased variability
  - early fault detection & classification; knowledge, less down-time
  - early quality assurance, parametric release (Pharma)

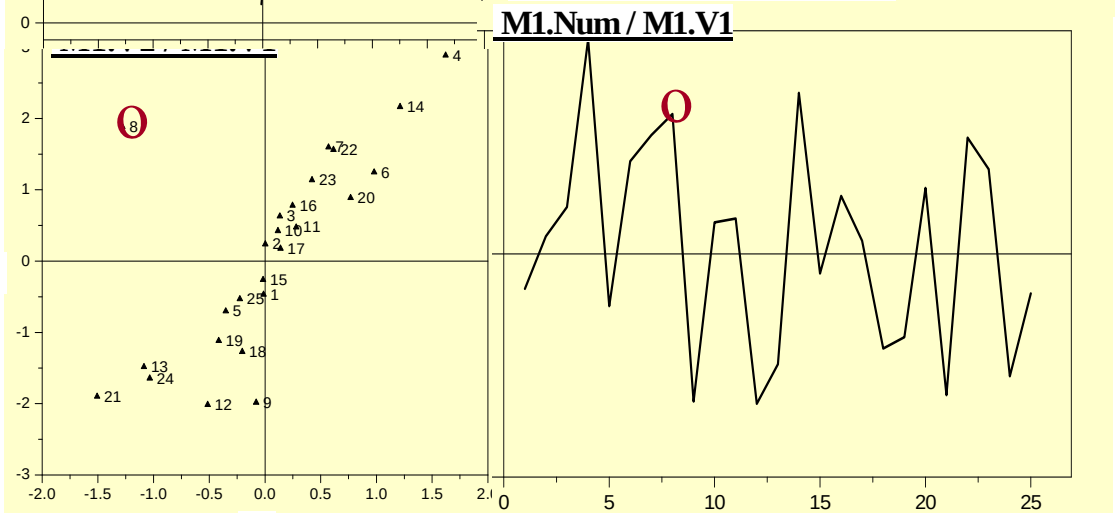
3. Do not look at data one variable at a time; MVA
Do not manipulate one factor at a time; DoE

- Inefficient
- Often misleading – confusing

Two correlated variables measured on a process

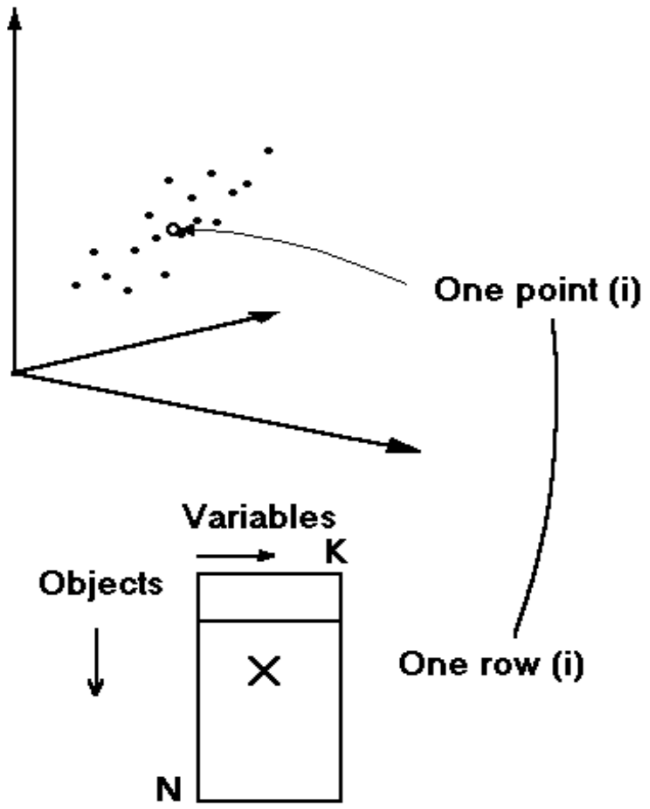






A multivariate chart, as above, catches the upset

Multivariate analysis by means of projections
PCA, PLS, …..



- Data shaped as a table, **X**
- Space with K axes (K-space)
  K = number of variables (col.s)
  Each obs. (process time point)
  is a point in this space
  Data Table = swarm of points
- Multivariate analysis
  - finding structures in M-space
  - describing them (math & stat)
  - using them for problem solving

4. R & D – examples, values, …

- Pharmaceutical examples – lots of data
  - Product optimization
  - Analytical data
  - QSAR (quantitative structure-activity relationships), Bioinformatics
  - Value: efficiency; fewer experiments, less time, reliability
    1 year = $ 1 billion, 1 day = $ 4 millions

- Tech chem example – structure – activity
  - Pharmaceuticals, Chemistry, Polymers, ..
  - Value: efficiency; fewer experiments, less time, reliability
    1 year = $ 10 millions to 1 billion

Optimizing the properties of tablets
AstraZeneca, Mölndal, Sweden

Tablet properties:

- Hardness ⇐
- release rate, disintegration time ⇐
- availability of the active substance
- content uniformity
- friability, etc.

Tablet properties depend on:

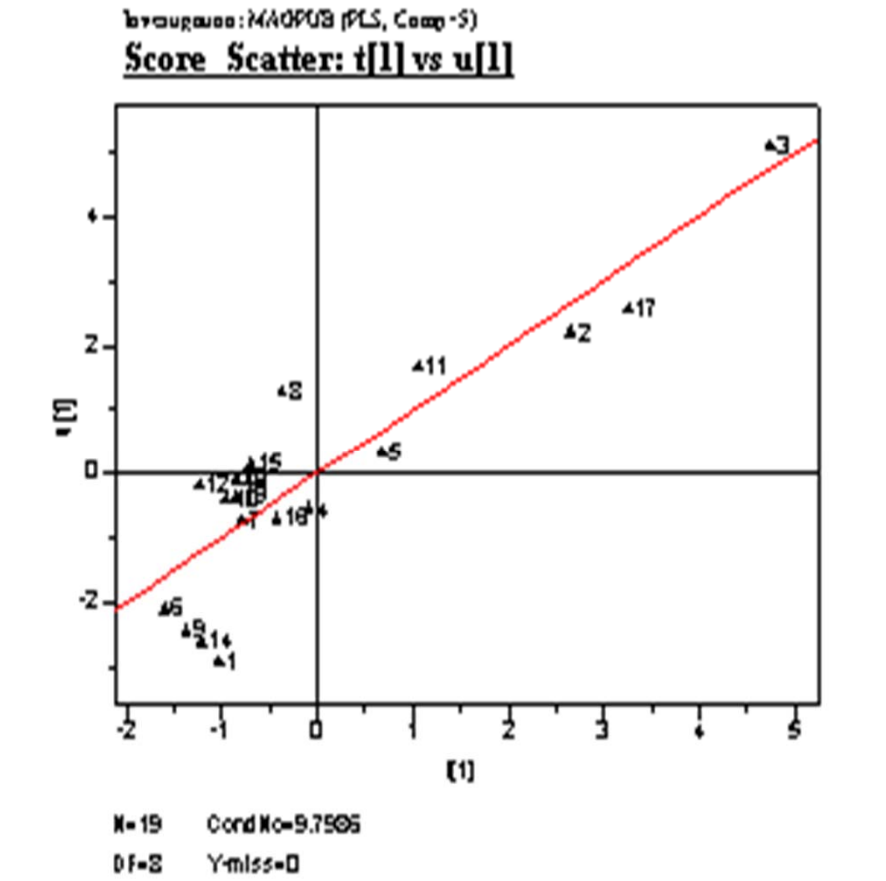
- The composition of the tablet [4 constituents]

  proportions of excipients (binder, filler,diluents, disintegration, lubricants, etc) used in the formulation of the tablet
- Processing conditions [3 process variables]

Do not change one factor at a time – use DoE – reliable, efficient (N=17)

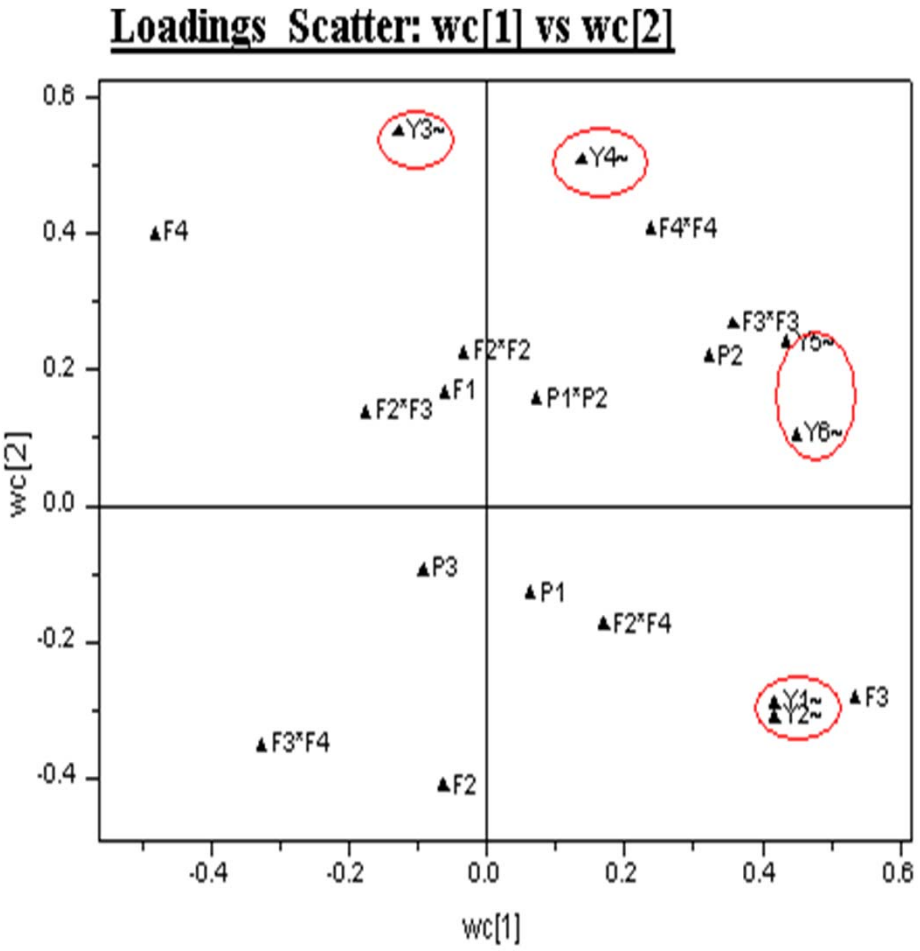
P1	P2	F1	P3	F2	F3	F4	Y1 …..
200	10	0.02	0.71	0.2	0.3	0.5	
200	10	0.02	0.71	0.07	0.93	0	
400	50	0.005	0.71	0	1	0	
400	50	0.005	0.71	0.13	0.3	0.57	
400	10	0.02	0.71	0.2	0.8	0	
400	10	0.02	0.71	0	0.53	0.47	
200	50	0.02	0.71	0	0.3	0.7	
200	50	0.02	0.71	0.2	0.63	0.17	
400	10	0.005	1	0	0.3	0.7	
400	10	0.005	1	0.2	0.63	0.17	
200	50	0.02	1	0.2	0.8	0	
200	50	0.02	1	0	0.53	0.47	
200	10	0.02	1	0	1	0	
200	10	0.02	1	0.13	0.3	0.57	
400	50	0.02	1	0	0.3	0.7	
400	50	0.02	1	0.2	0.3	0.5	
400	50	0.02	1	0.07	0.93	0	

PLS plot of t1 vs u1



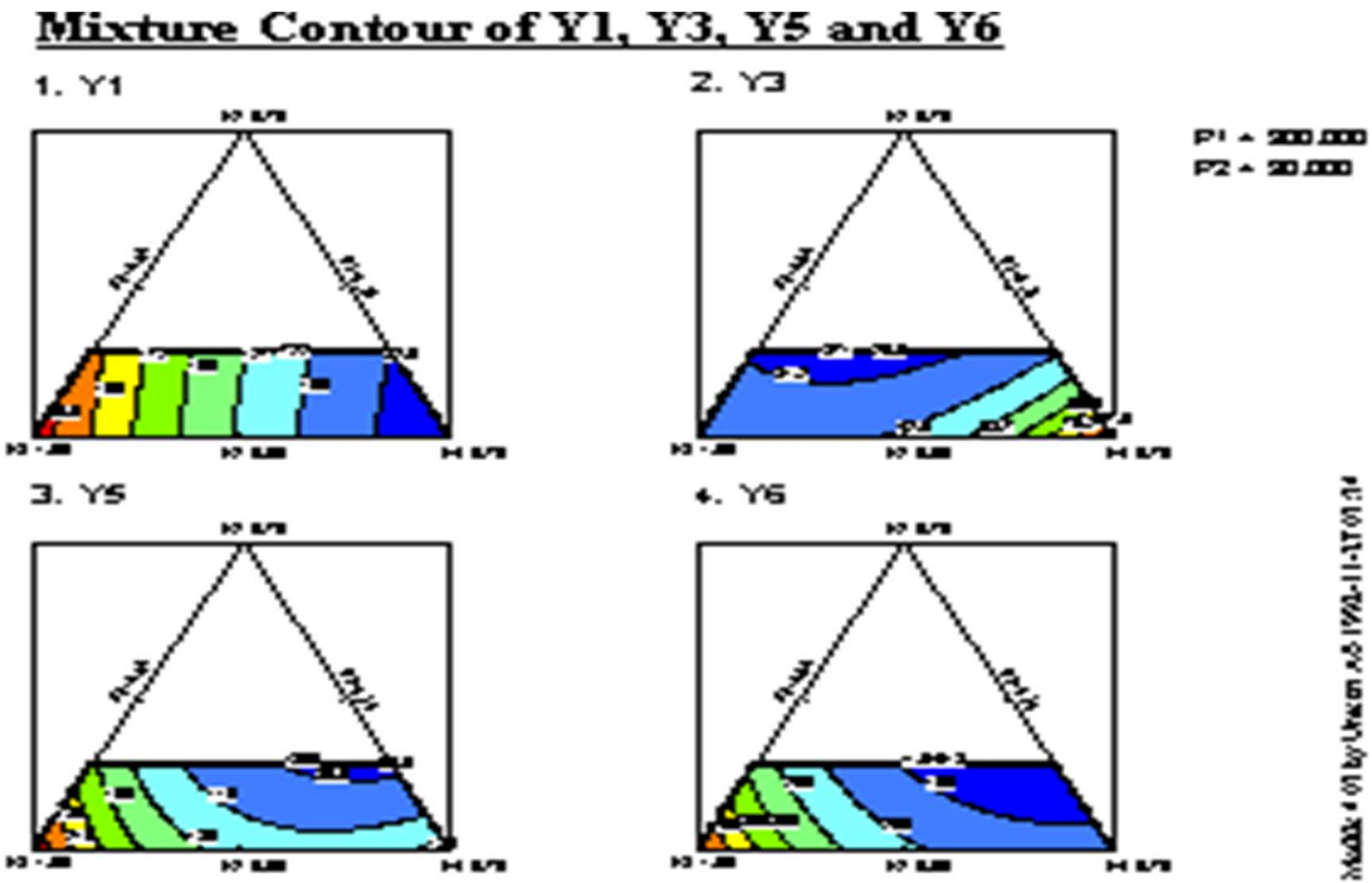
This shows a good relationship with some scatter around the line

PLS plot of wc1 vs wc2, explaining 70% of SS



- Hardness measures (Y1 and Y2) are positively correlated and negatively correlated to Y3
- Y5 and Y6 correlate positively
- F1 and P3 little importance.
- F4 and F3 very important
- $F3^2$, $F4^2$, and the interaction
- F3*F4 are the only important second order terms (not indep.)

Contour plots of 4 responses
Y1, Y3 Y5 and Y6



Optimizer: Simplex search of the fitted model

Simplex Runs

	1	2	3	4	5	6	7	8	9	10	11	12	13
	P1	P2	F1	P3	F2	F3	F4	Y1	Y2	Y3	Y4	Y5	Y6
1	399.9524	10.0086			0.1999	0.5795	0.2206	143.7739	164.8717	26.8303	32.3925	60.3087	94.38
2	399.5985	10.0105			0.2000	0.5774	0.2226	143.1349	164.2258	26.8425	32.3634	60.1464	94.02
3	399.9977	10.0002			0.2000	0.5781	0.2219	143.4375	164.5365	26.8417	32.3418	60.1501	94.07
4	399.8923	10.0028			0.2000	0.5768	0.2232	143.0724	164.1696	26.8552	32.3141	60.0382	93.85
5	399.9997	10.0021			0.2000	0.5773	0.2227	143.2212	164.3208	26.8527	32.3168	60.0682	93.92
6	399.8949	10.0016			0.2000	0.5783	0.2217	143.4407	164.5374	26.8381	32.3568	60.1801	94.12
7	399.5276	10.0097			0.1999	0.5767	0.2234	142.9150	164.0035	26.8564	32.3521	60.0898	93.91
8	399.2264	10.0047			0.2000	0.5761	0.2239	142.7037	163.7847	26.8489	32.3647	60.0728	93.84

Conclusion

- Best Conditions:

  P1= 398, P2=10, F2=0.2, F3=0.52 and F4 = 0.28

  Resulting in the following predicted responses

  Y1= 130, Y2=150, Y3=28, Y4=31, Y5= 55, Y6=85

- **Confirmed by experimentation**

- **This result could only be achieved by experimental design & MVA**
- **Value -- time saved; weeks to months**

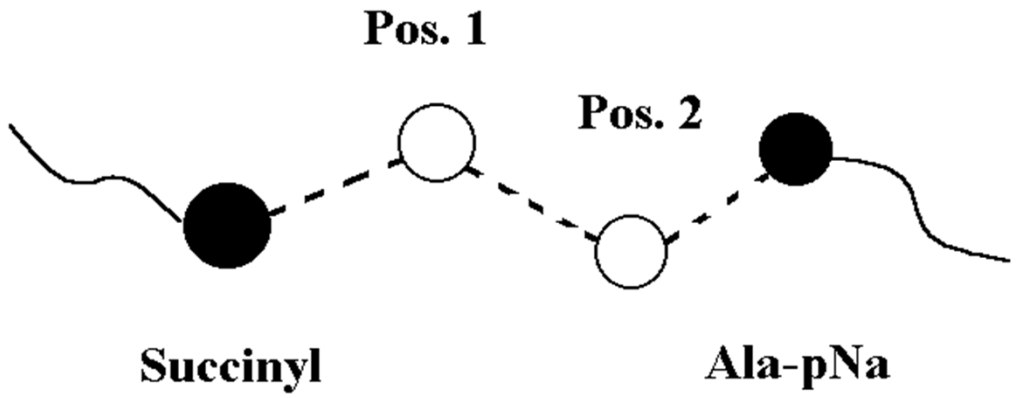
Optimizing chemical structure (academic example)
QSAR ex: Porcin pancreatic elastase

Y1 = log($\mathbf{k_{cat}}$)

acylation of enzyme & p-NitroAniline (pNa) release

Y2 = log ($\mathbf{k_{cat}}$ / $\mathbf{K_m}$)

K_m = substrate – enzyme binding



Nomizu et al., Int.J.Pept.Protein Res. **42** (1993) 216-226. **N = 89** substrates (peptides) varied in two positions.

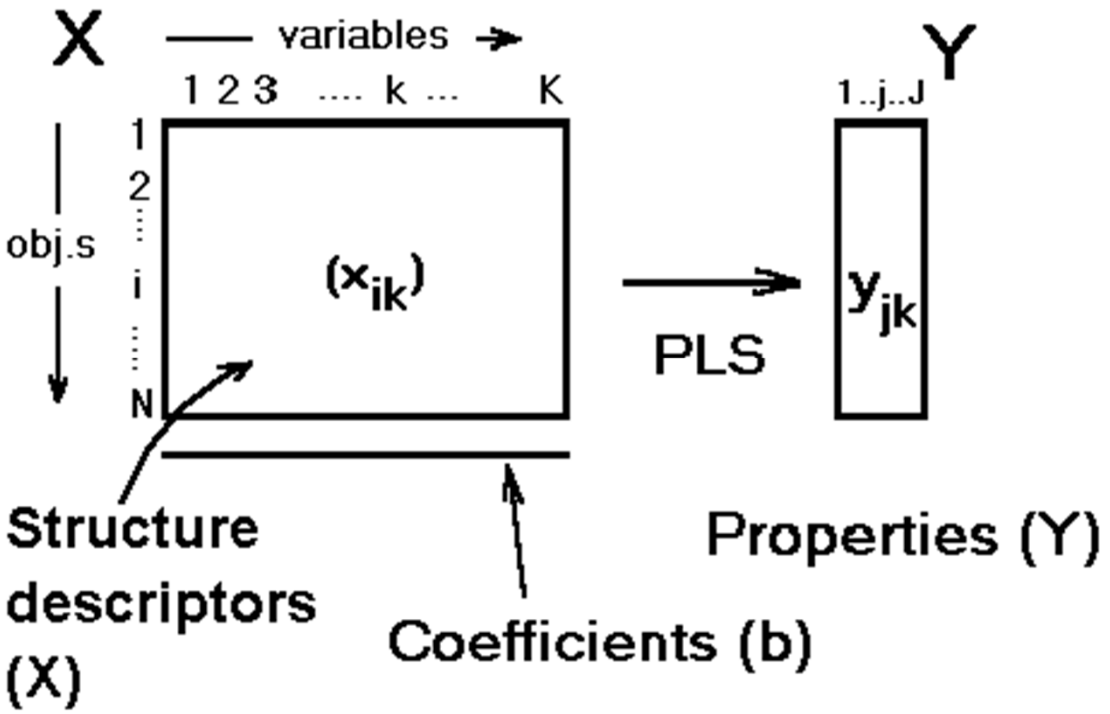
Here training set = **32** (D-opt design, quadratic model), pred. set = **57**

Essential step:
Translate structural variation → Numbers

Here, each varying peptide position

is translated to 4 numbers

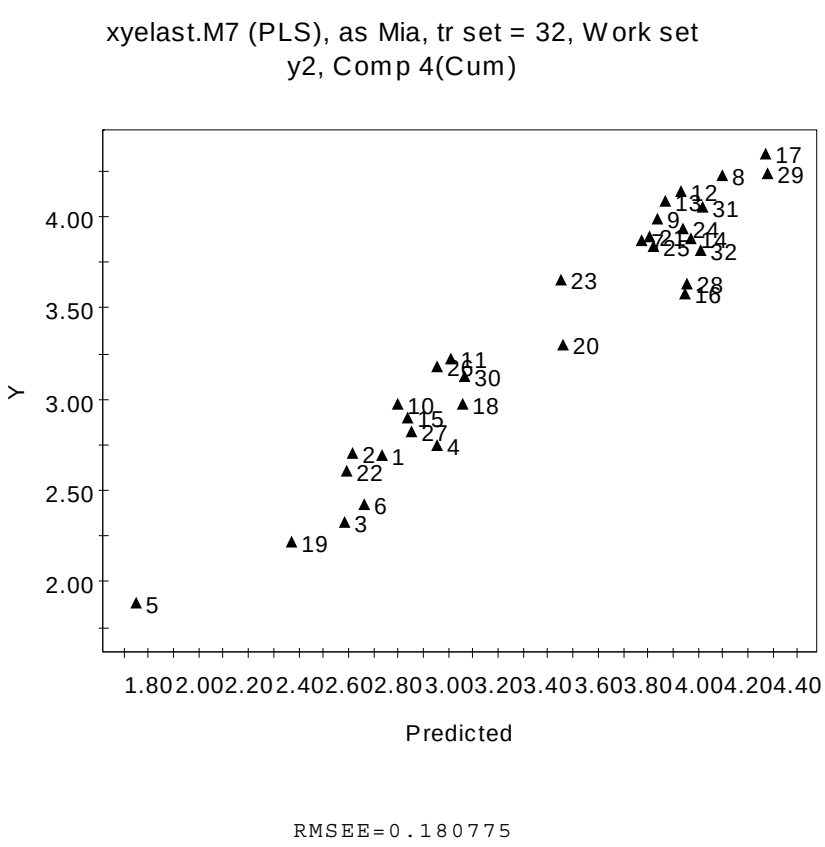
[the z-scales of the Umeå Chemometrics Group]

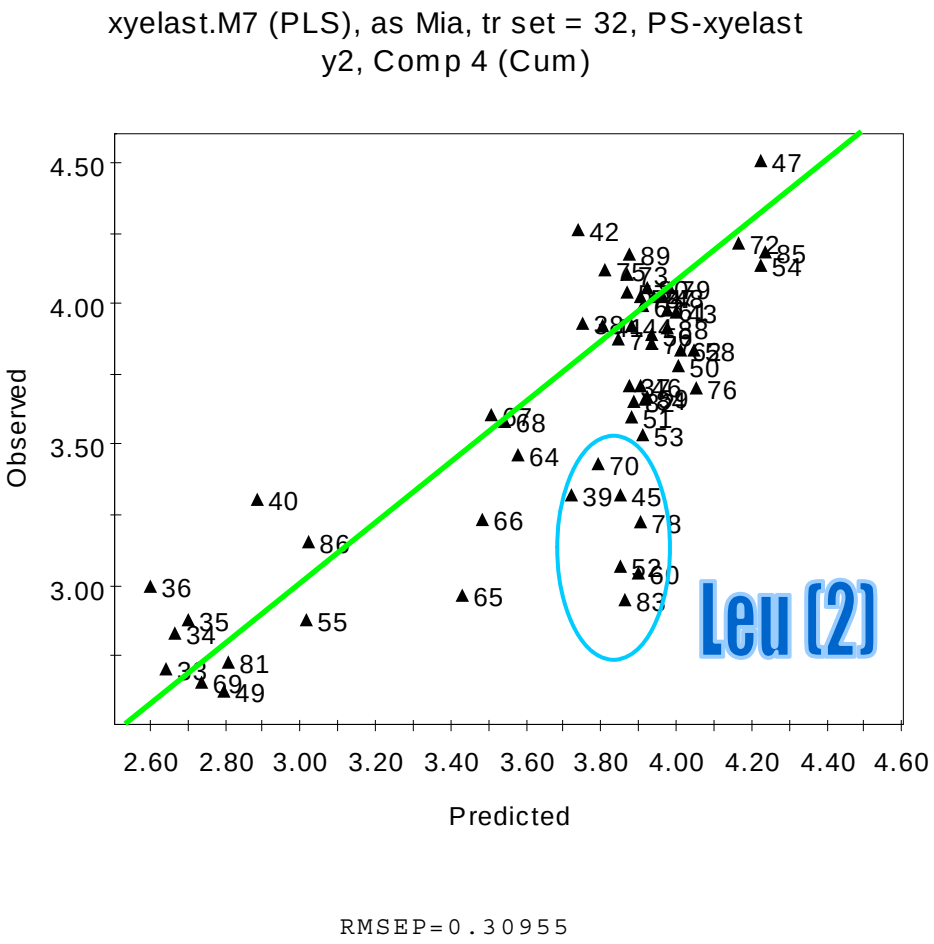


Before the PLS modelling, the X-matrix is expanded with some quadratic and interaction terms [design !]

Observed vs Predicted, y2, TS = left, PS = right

4 PLS components (CV); $R^2 = 0.93$, $Q^2 = 0.78$





PLS coefficients, elastase substrates

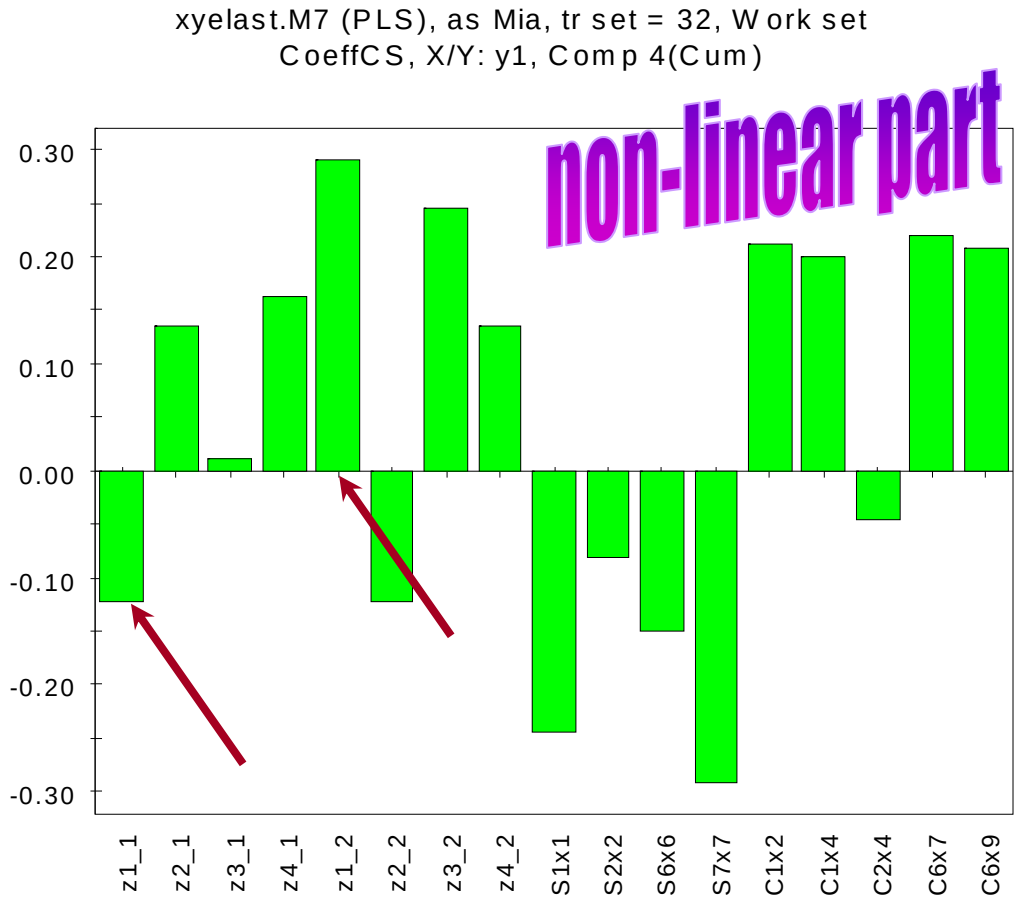
z1 = hydrophilicity

z2 = size, polarizab.

z3 = polarity

z4 = ??

z5 = ????



Summary, elastase substrates;
Value ⇐ model predicting how to modify molecule

- Substrates (substituted tripeptides) described by 8 variables
- N=32 training compounds selected by D-optimal design
- PLS model explains and predicts (57 additional compounds) about 90 % of the variation of log k_{cat} and log (k_{cat} / K_M)
- Some systematic misfit is detected
- Model coefficients are interpretable and indicate how to modify structure to improve activity

- Maria Sandberg et al., "New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 AA.s." J.Med.Chem. **41** (1998) 2481- 2491

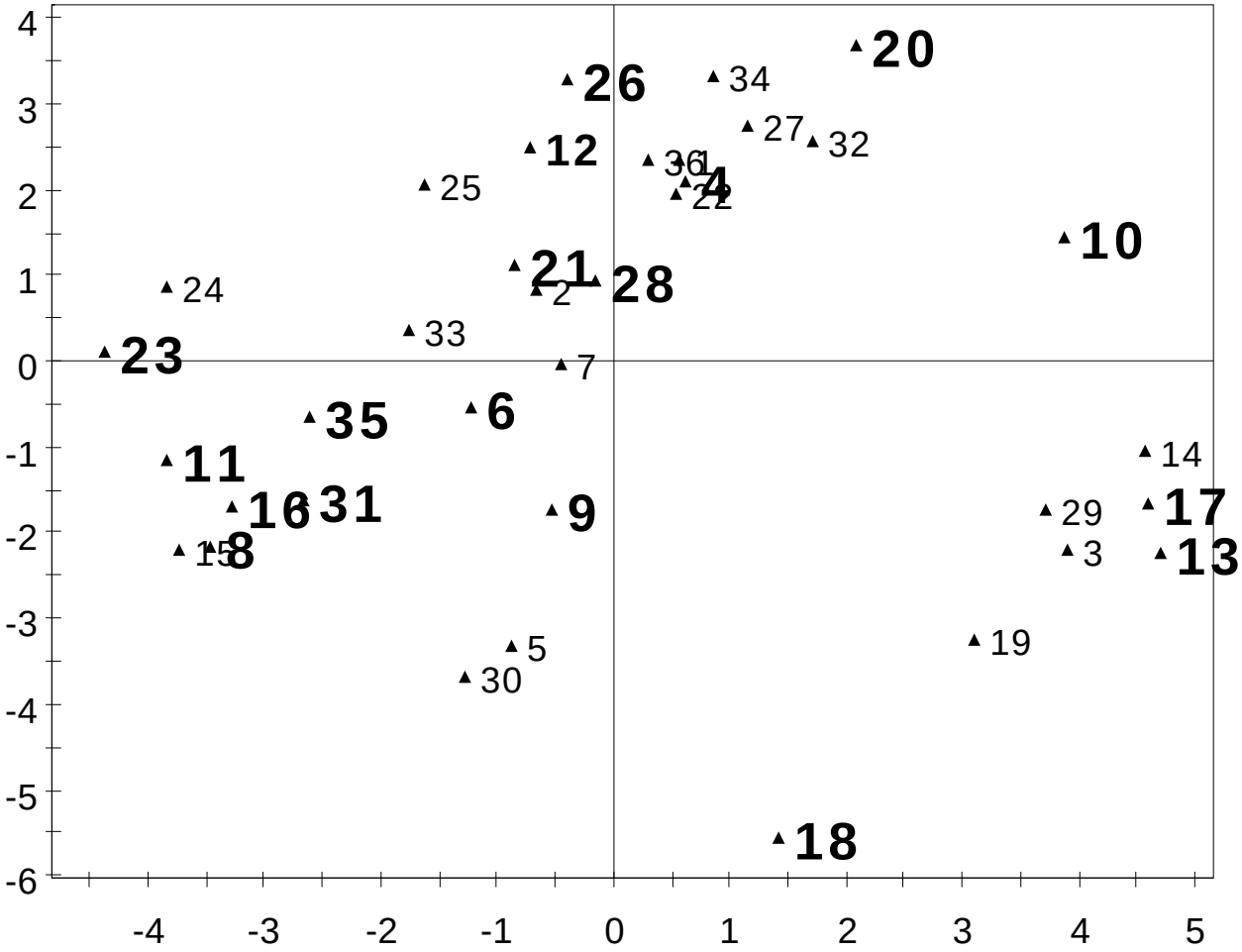
Non-ionic detergents -- Akzo-Nobel, Sweden

- N=36 technical detergents
- Environmental toxicity is becoming a concern
- A map is wanted of the toxicity of potential detergents.
- All potential detergents cannot be tested

- A representative set – N=18 -- is selected (multivariate characterization + design), tested, and modelled.
- Model used to predict further detergents (validation)
- Value = preparedness, knowledge about alternatives

- Å Lindgren, et al., QSAR 15 (1996) 208-218 & P13

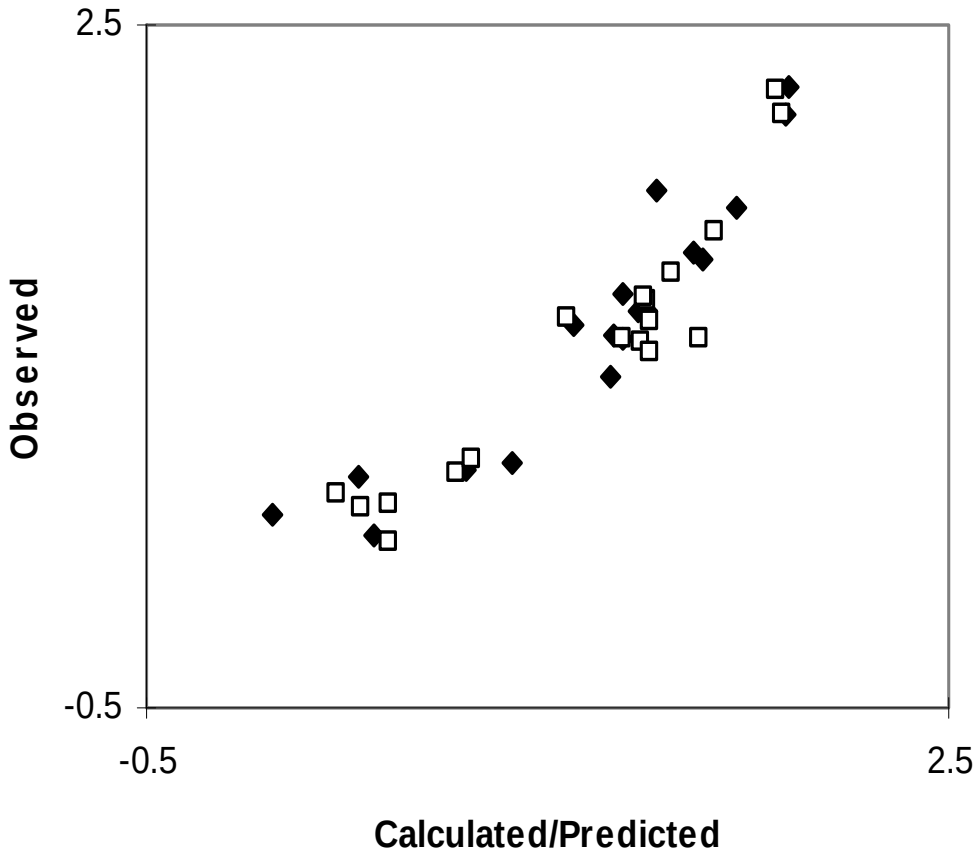
x.1B Design in t1 and t2 (PC scores 1 & 2)



$19 \Rightarrow 4$

x.1B. Toxicity of non-ionic detergents

♦ = training set = predictions



Example of
non-specific activity.
Single model for all data

Reactivity (detergent
effect) is “specific”
needs 5 models,
one per group.

Å Lindgren and M.Sjöström
Chemom. Intell. Lab. Syst.
23 (1994) 179-189

Value

- Resulting model allows the selection of technical detergent that has specific properties and low toxicity
- Environmental preparedness
- Insight

5. Processes -- lots of data, use for knowledge

- Ex. 1; Slow unreliable traditional measurement $\Rightarrow$ fast, reliable (multivariate) measurement

  Value: Process understood, scrap rate 30 % $\Rightarrow$ 0 %

- Ex. 2; Integration of multivariate measurements, optimization, MSPC

  Value: Improvement of profit by more than 20 %

Correlation of Sensory & Analytical Data in Flavour Studies into Scotch Malt Whisky

James S. Swan and David Howie

Pentlands Scotch Whisky Research Ltd,
84 Slateford Road, Edinburgh EH11 1QU, Scotland

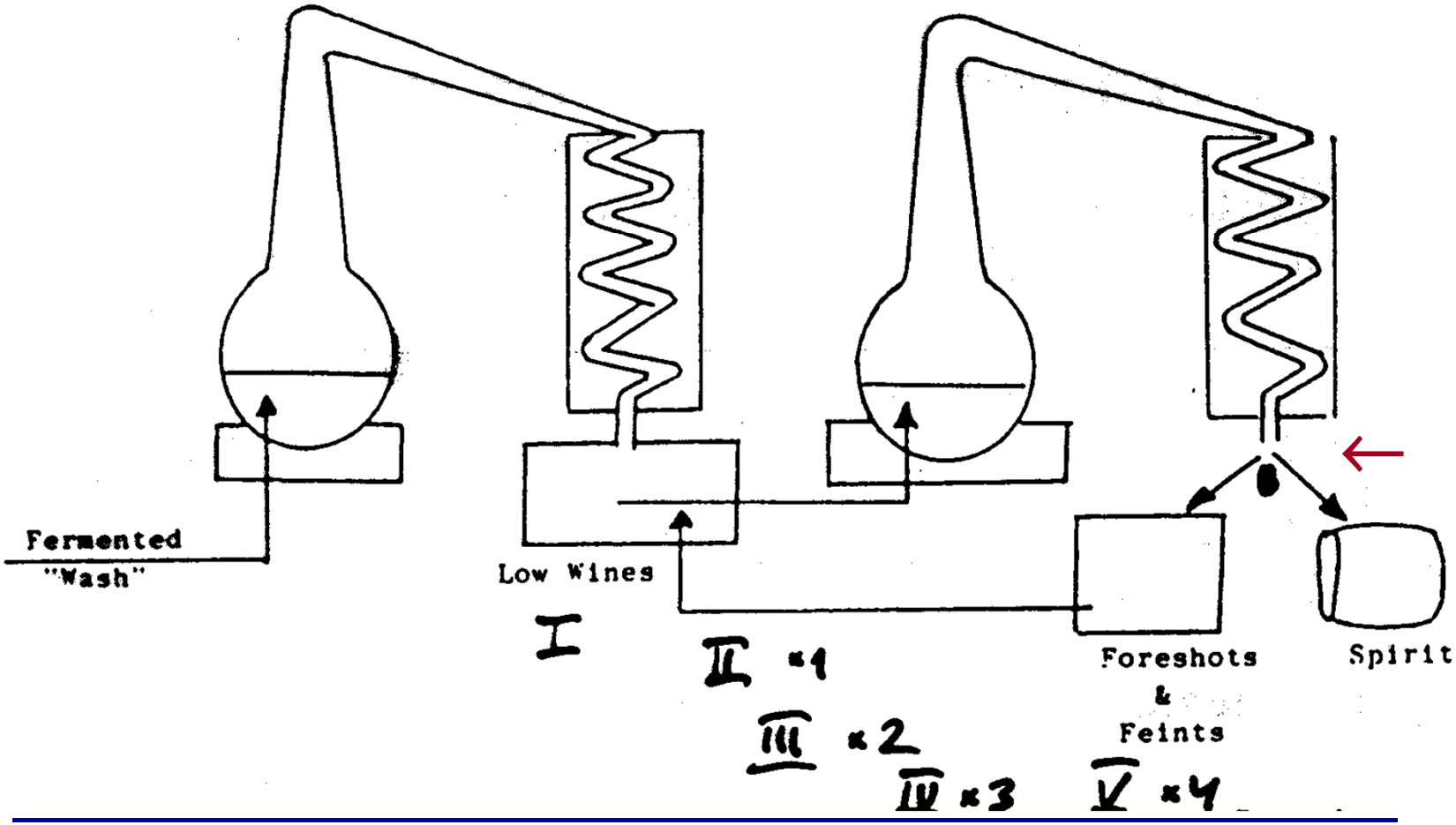
Proceedings of the Alko Symposium on FLAVOUR RESEARCH OF ALCOHOLIC BEVERAGES, Helsinki 1984, ed. by L. Nykänen & P. Lehtonen. Foundation for Biotechnical and Industrial Fermentation Research 3 (1984): 291–300

Scotch malt whisky is a noteworthy example of a product in which the flavour impact is the result of a complexity of interactions between a large number of compounds. To connoisseurs the tradition behind the product is an integral part of their enjoyment but in the modern world the maintenance of traditional quality relies increasingly upon an understanding of the flavour construction and the contribution made by various production processes.

Analytical studies such as those using capillary column gc.ms have enabled more than 280 compounds in matured whisky to be identified (1) and recent decades have seen the development of sensory

Whisky Process: Double distillation (copper vats)

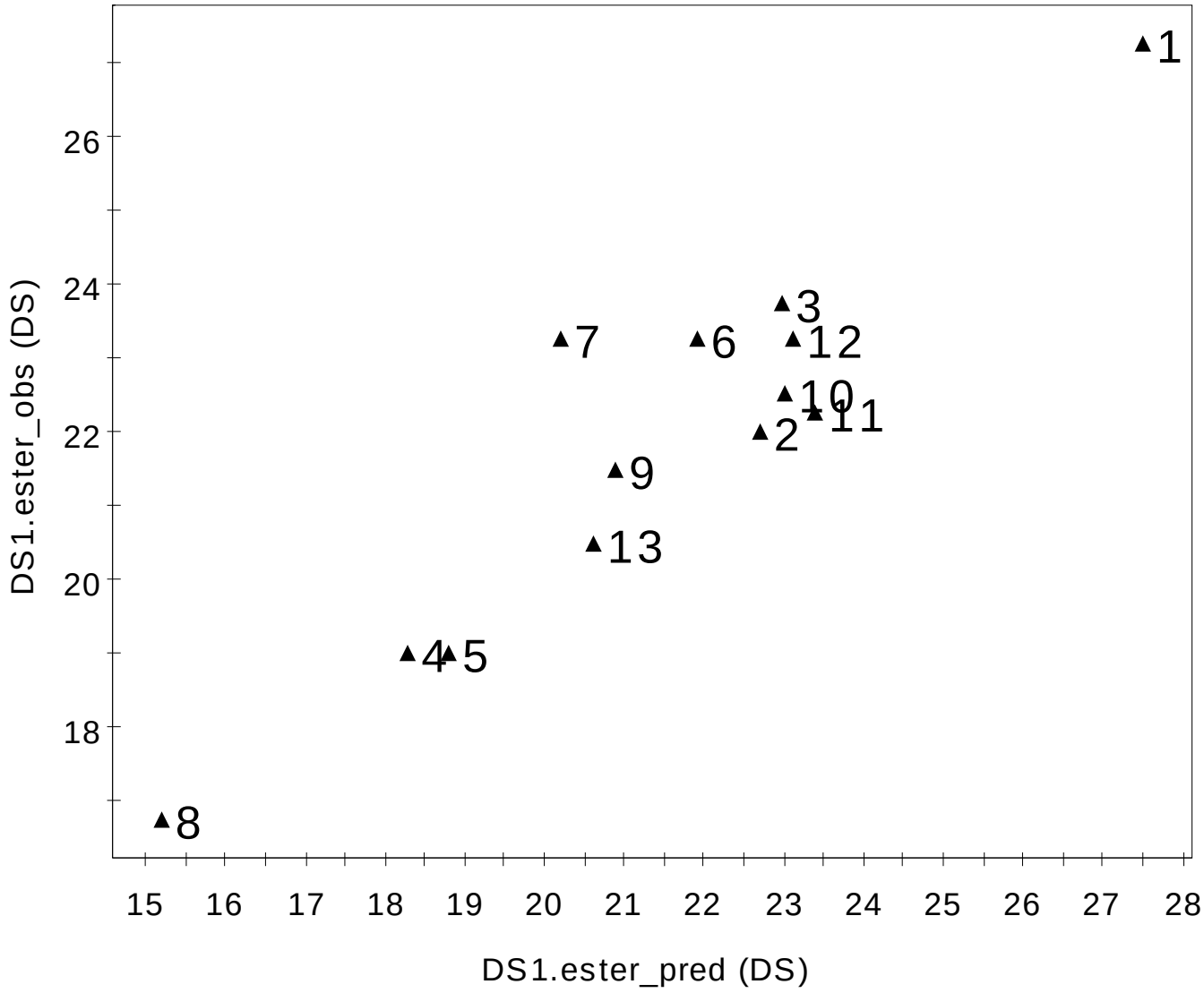
Critical point: amount of reflux ←



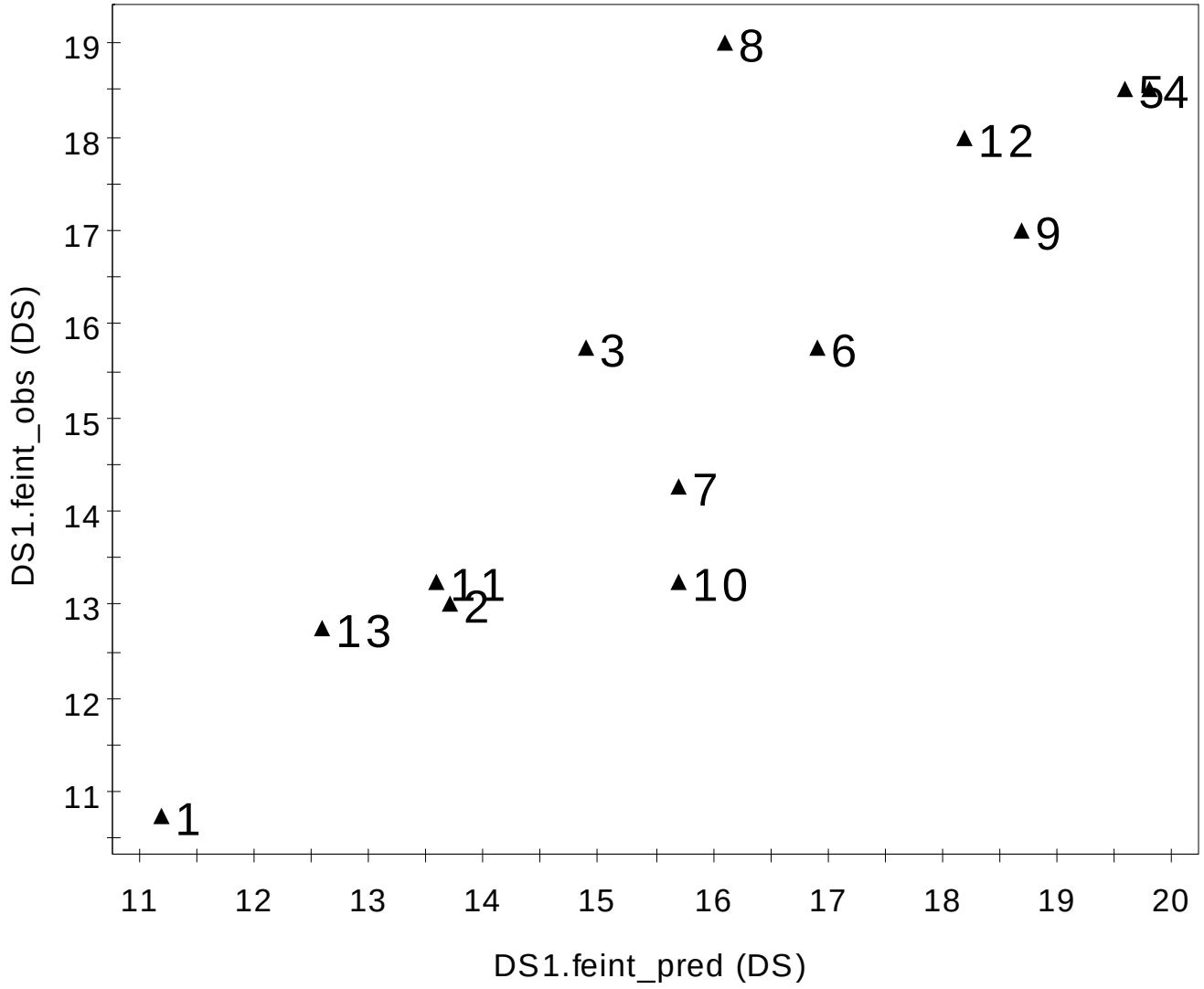
Initial analysis

- **Historical data** N ≈ 500, K = 49 (log GC peaks)
- $R^2 = Q^2 \approx 0.97$
- **Overfit ?**
- Cheating ?
- Additional experiments (designed)
- 2 more GC.s – raw material (wash) & intermediates (lw)

Ester pred/obs (first 12)



Feinty, pred/obs (first 12)



Summary and Epilogue

Summary

- GC's used to characterize raw material and intermediates
- Extensive experimentation (DOE) on process
- Multivariate calibrations; GC vs 12 taste scales
- Still-men learnt GC & micro-computers
- Process well understood and controllable

Epilogue

- 3 years later -- final quality control of product in barrels
- Stable quality at slightly improved level, scrap rate = 0 %

Prediction in particle-board industry

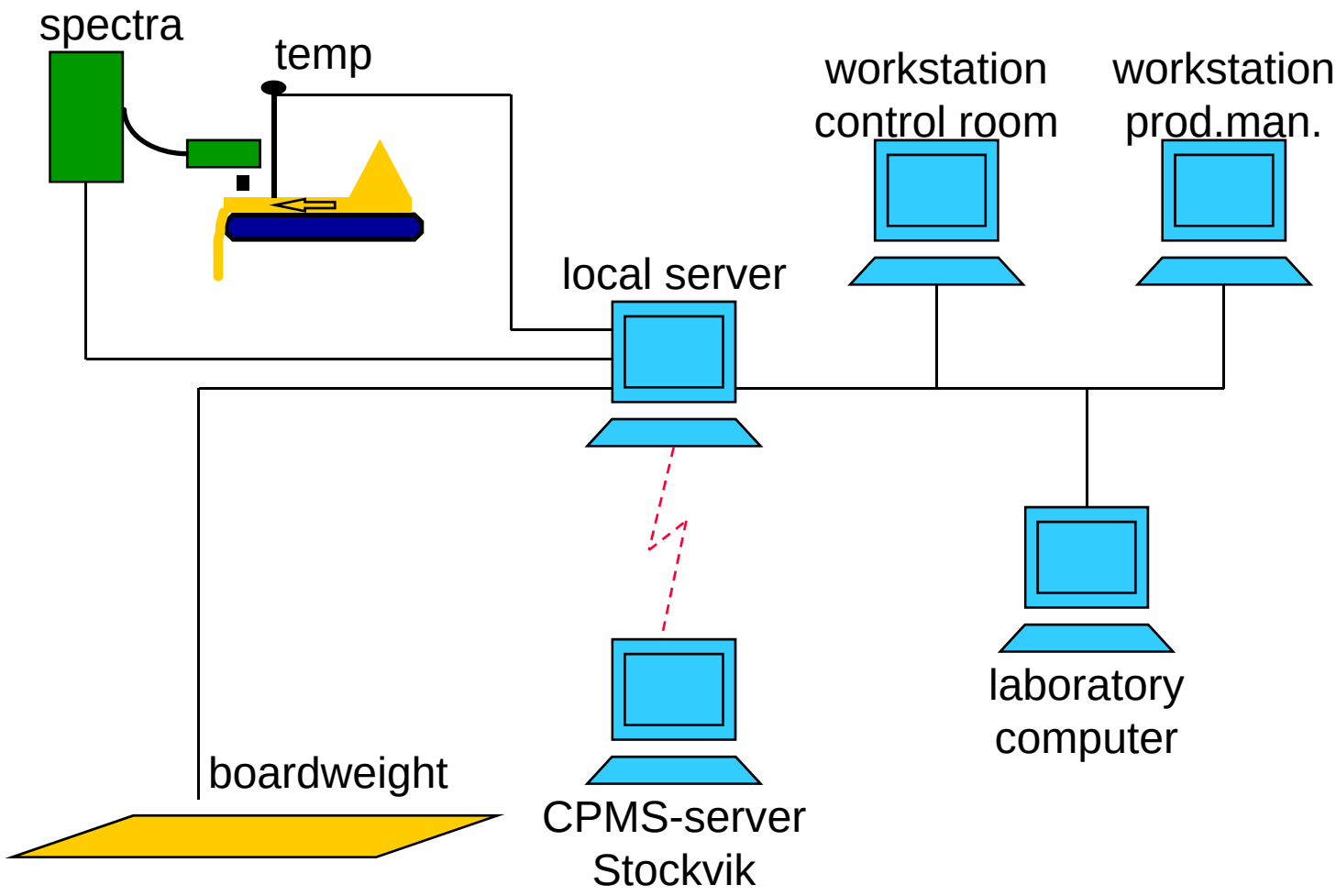
- This is a specially designed product developed jointly by Casco Products AB and Umetrics AB.

Production of the right quality at every moment

- Casco Products is the responsible originator of this concept, CPMS

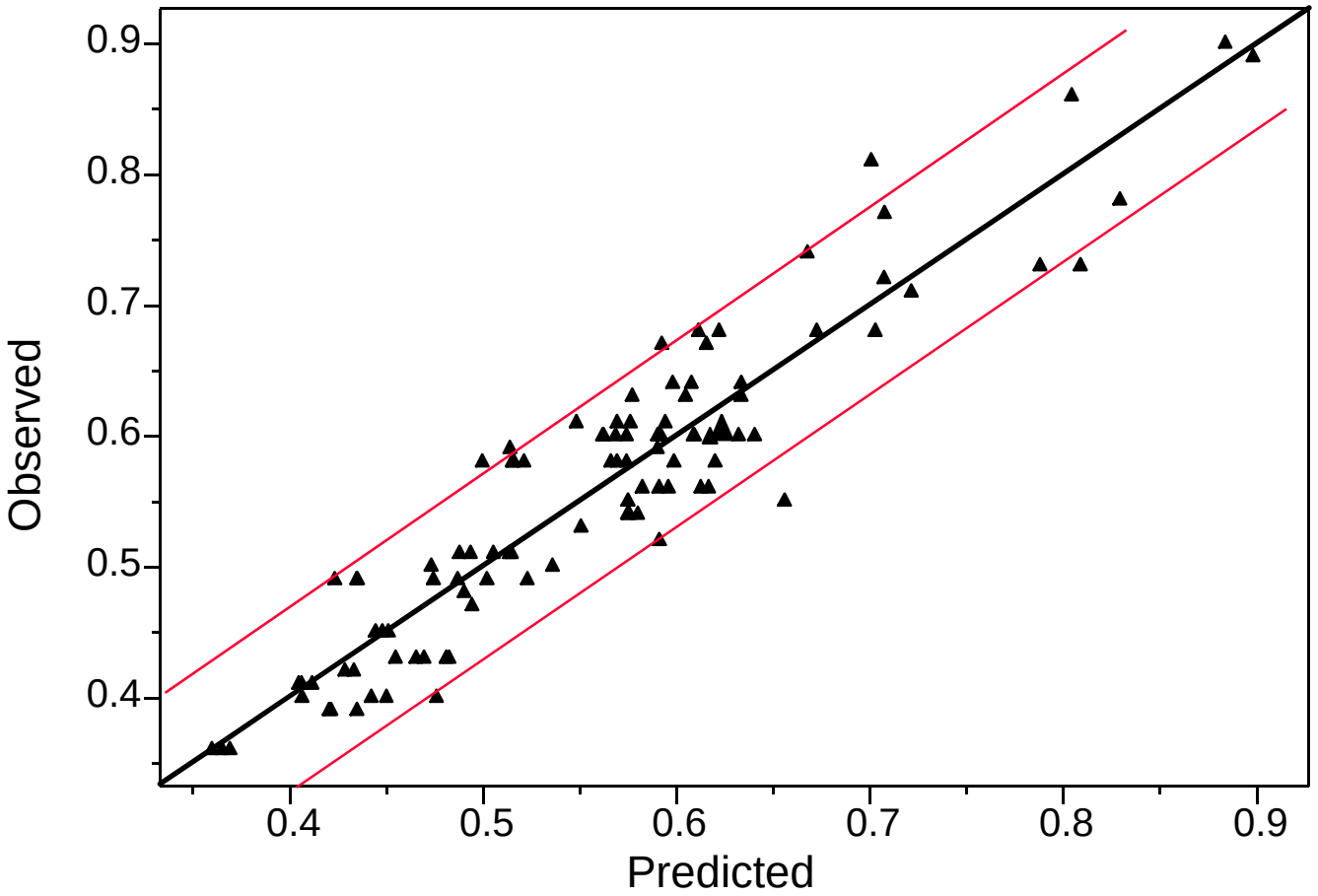
- **Problems:** The raw material varies,
  The quality measurement of the final product is a lab test obtaines hours after the board is done.
- **The keys:** NIR spectroscopy
  experimental design (selected chips + process)
  multivariate models.

System layout, CPMS - Byggelit Storuman Line 2



Prediction in particleboard industry

Safety margin, IB

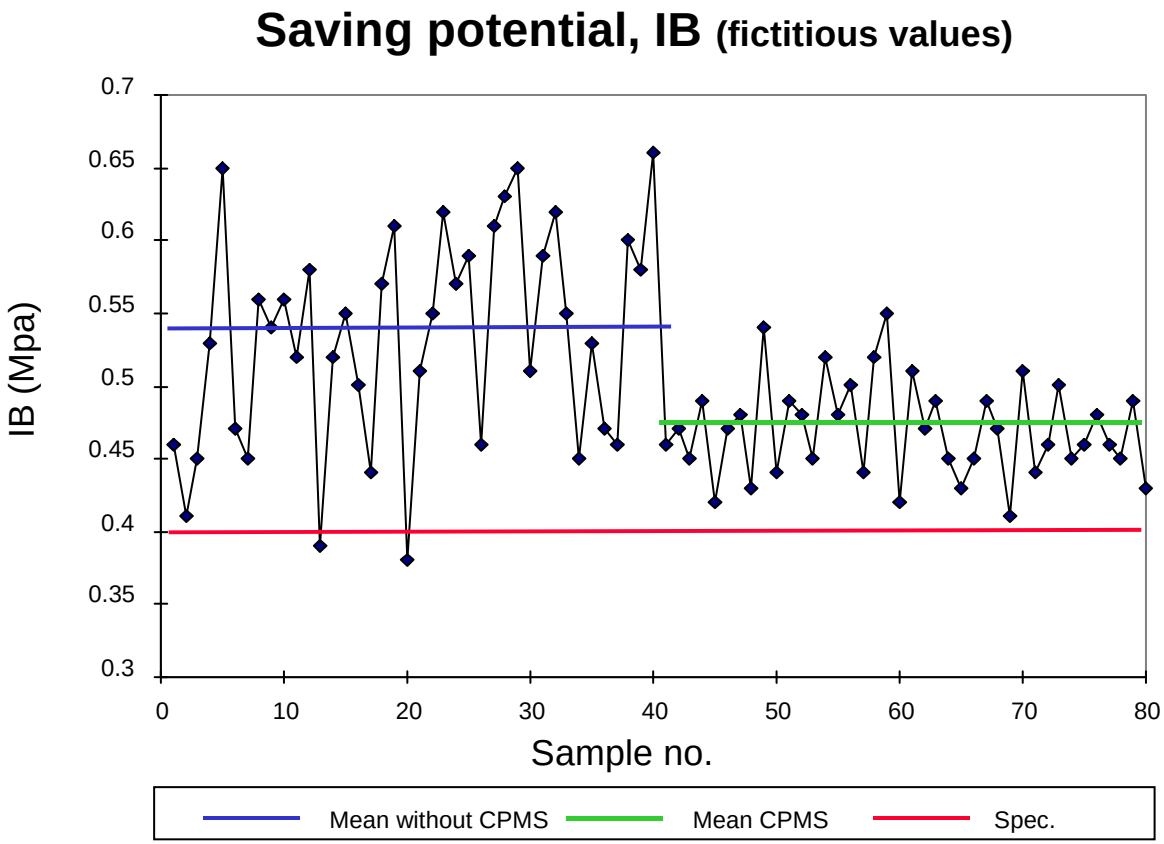


Prediction in particleboard industry -- CPMS

The production target for IB could be reduced and the fluctuations in quality minimised.

New target depends on the model prediction error.

The saving potential has been estimated to approximately 1 billion SEK in Europe



Casco Products

Guidelines to successful implementation

- Good process knowledge
- Allocation of enough resources
- Efficient data handling environment
- Awareness of the precision in the quality measurements
- Commitment from the producing part
- Well defined steps of the implementation
- Use design of experiments (operational region) & MVA

  Look at and manipulate many variables simultaneously,

  NOT one at a time

The ladder of value creation

UMETRICS

Measurements (WDWW)

Data

Information

Decision

Action

Value

*1987

Software

Training

Consulting

Implementation

MV characterization	Structure, Tablet, Process, Materials
Indirect measurements	Whisky (GC), Chips (NIR), Compds
MV modelling [PCA, PLS]	T, QSAR, GC-taste, Process-prop.
Chemical structure optimization	Elastase substrates, Detergents

6. Conclusions

new technologies = threats & opportunities

- Masses of good data have large potential value
  - Use appropriate data generating approach (DoE)
  - Measure multiple and relevant data – more is better
  - Use methods of analysis developed for multivariate data
  - Use common sense & clearly stated objectives
  - This creates value
- Some investment is necessary
  - personnel; adequate level and training
  - time; for DoE, interpretation, feedback, action
  - hardware and software; data bases, DoE, MVA
- ROI is 50 to more than 200 %
- Train is leaving the station – get on !

Some pioneers in Europe and other continents

- Akzo-Nobel, Perstorp, Foss, Karlshamn, Carlsberg, BASF, Hoechst, Wacker, Honeywell (Allied), …
- AstraZeneca, Pharmacia-Upjohn, Novo, Rhone-Poulenc, SKB, Novartis, Merck, Melacure, Camitro, Arqule, …
- LKAB, Boliden, SSAB, Noranda, Avesta-Sheffield, Dofasco, …
- Exxon, Statoil, Norsk Hydro, BP, Lubrizol, Petro-Canada, Shell, ..
- ASSI, MoDo, SCA, Stora-Enso,

  Tembec, Weyerhaeuser,…
- ABB, Harris, Ericsson, IBM, TEA, …
- Volvo, Ford,

THE END

Thanks for Your attention