Multivariate statistical analyses and machine learning for metabolomics

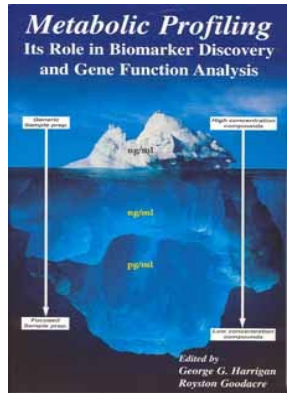
Roy Goodacre

School of Chemistry,
The University of Manchester,
Sackville Street, PO Box 88,
Manchester, M60 1QD

MANCHESTER 1824

Roy.Goodacre@manchester.ac.uk
www.biospec.net

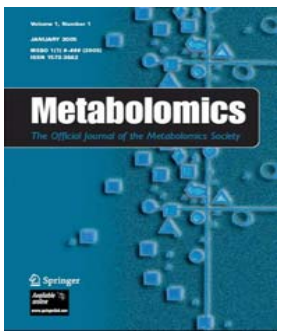
www.biospec.net



Postdocs: Dr Catherine Winder, Dr Robert Cormell, Dr Roger Jarvis, Dr Mohammad Afzaal, A.N. Other x 5,
With Collabs: Dr Seetharaman Vaidyanathan, Dr John Fletcher
Research Technicians: Jo Ellis, Steffi Schuler + A.N. Other

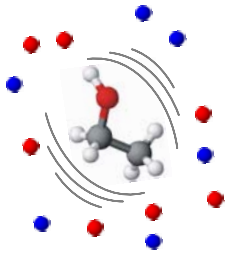
Research Students (PhD): Felicity Currie, Katherine Hollywood, Robben Jaber, Nicoletta Nicolaou, Soyab Patel, Ketan Patel, Emma Wharfe, Will Allwood, Will Cheung, Robert Coe.
From Oct 1: Nicola Wood, Dong Hyun Kim.

Research Students (MSc): Nicola Wood, Sadia Rabbini, Martin Coleston, Graham Mullard.



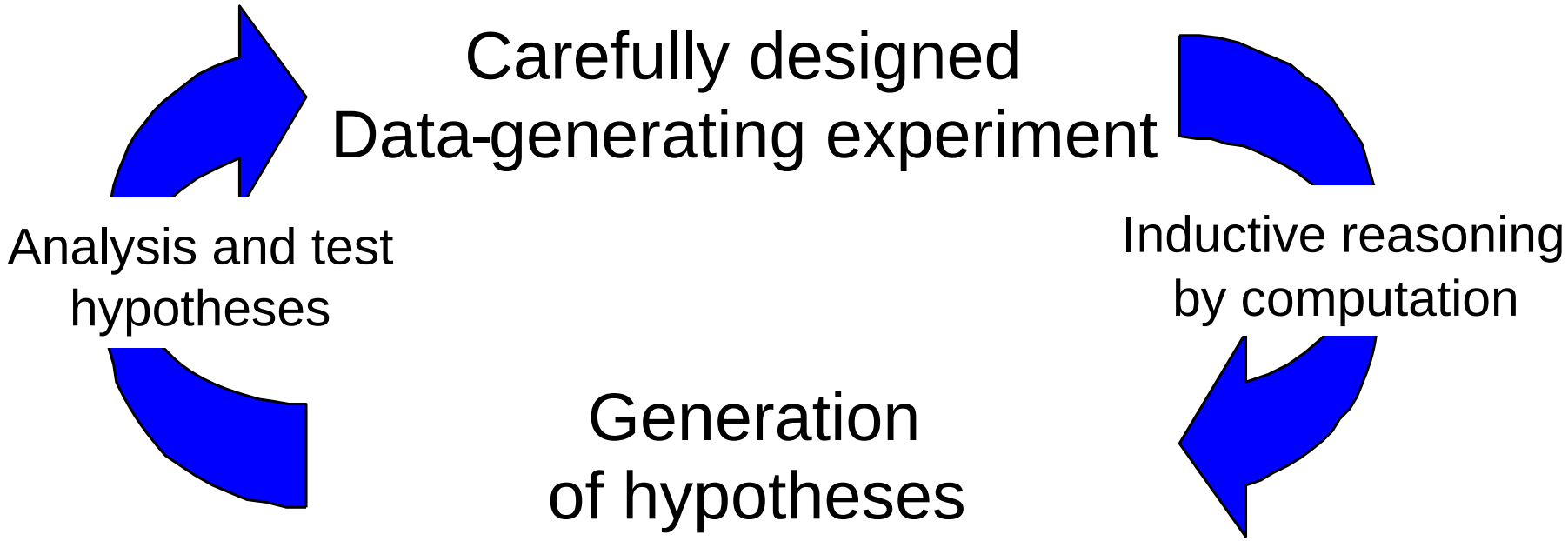
£s, €s, $s: BBSRC, EPSRC, RSC, EU F6, DEFRA, NERC, ORS

bbsrc
biotechnology and biological sciences research council



‘Top down’ metabolomics

Strategy: Inductive approach to knowledge discovery via holism



“Hiring a statistician after the data have been collected is like hiring a physician when the patient is in the morgue. He might be able to tell you what went wrong, but is unlikely to be able to fix it”

MANCHESTER 1824

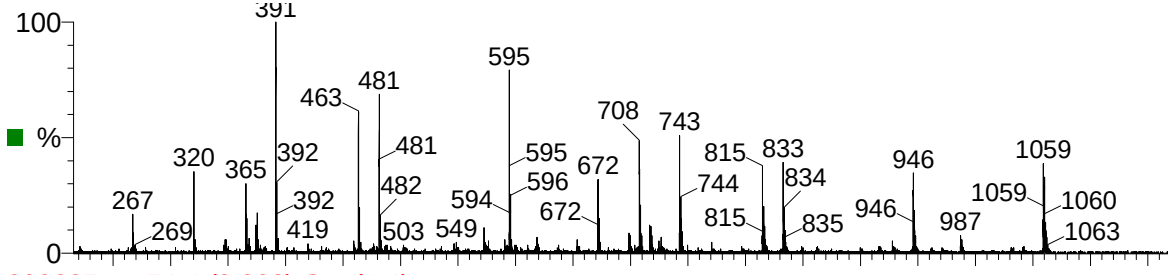
Goodacre, R. (2005) Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal of Experimental Botany* **56**, 245-254.

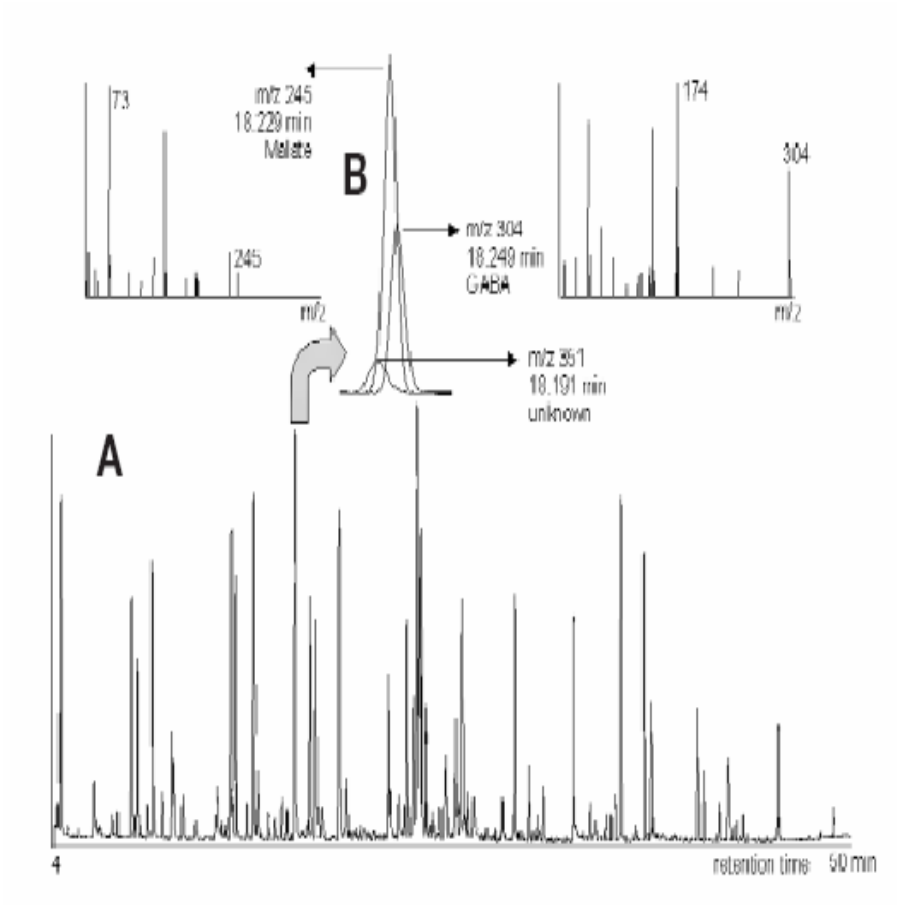
Data floods

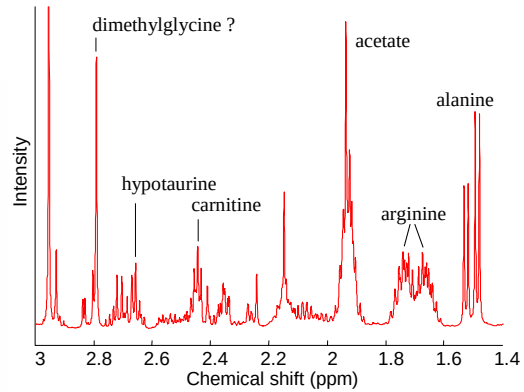
"Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom.
We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom"

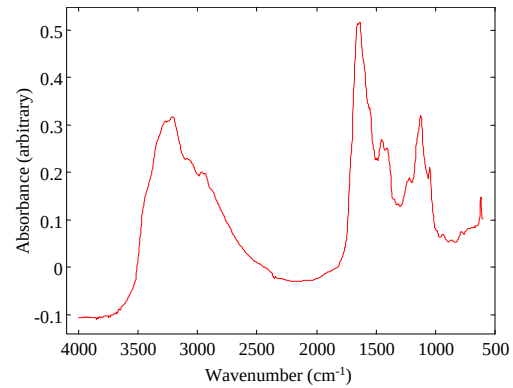
Henry Nix, 1990

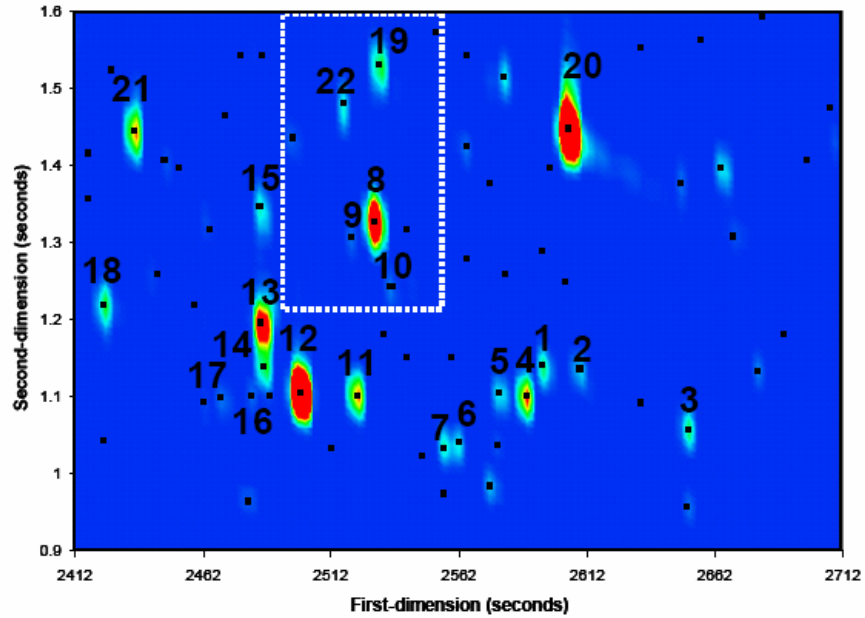
Data outputs





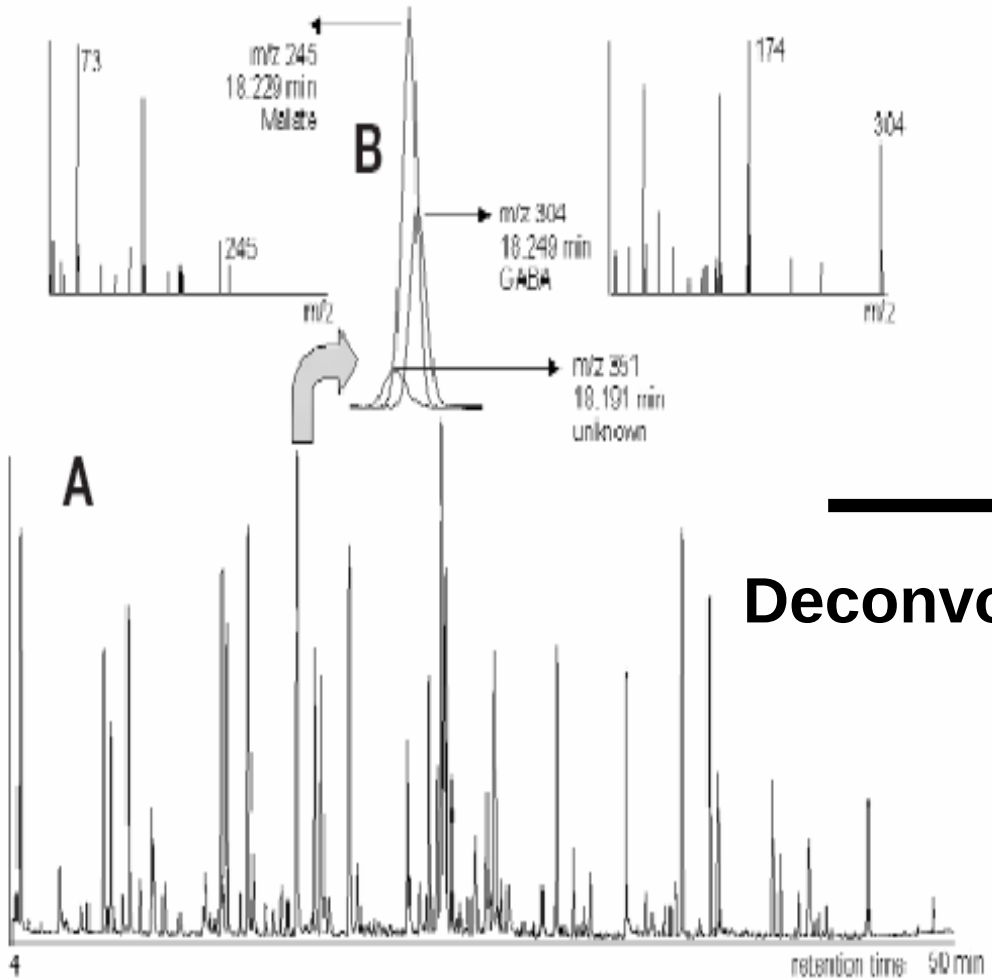






MANCHESTER 1824

Typical metabolic profile



Deconvolution

Metabolite	Conc
Glucose	0.1
Indole	0.001
Tryptophan	1.2
Ethanolamine	0.7
…	…
Metabolite #88	0.9
Metabolite #167	0.05

- Name of metabolite with its concentration

Data handling

Objects going down in different rows		**X-var 1** Metabolite or peak 1	**X-var 2** Metabolite or peak 2	**X-var 3** Metabolite or peak 3
Sample 1				
Sample 2…				

Input data

Metabolite	Conc
Glucose	0.1
Indole	0.001
Tryptophan	1.2
Ethanolamine	0.7
…	…
Metabolite #88	0.9
Metabolite #167	0.05

MANCHESTER 1824

Data handling

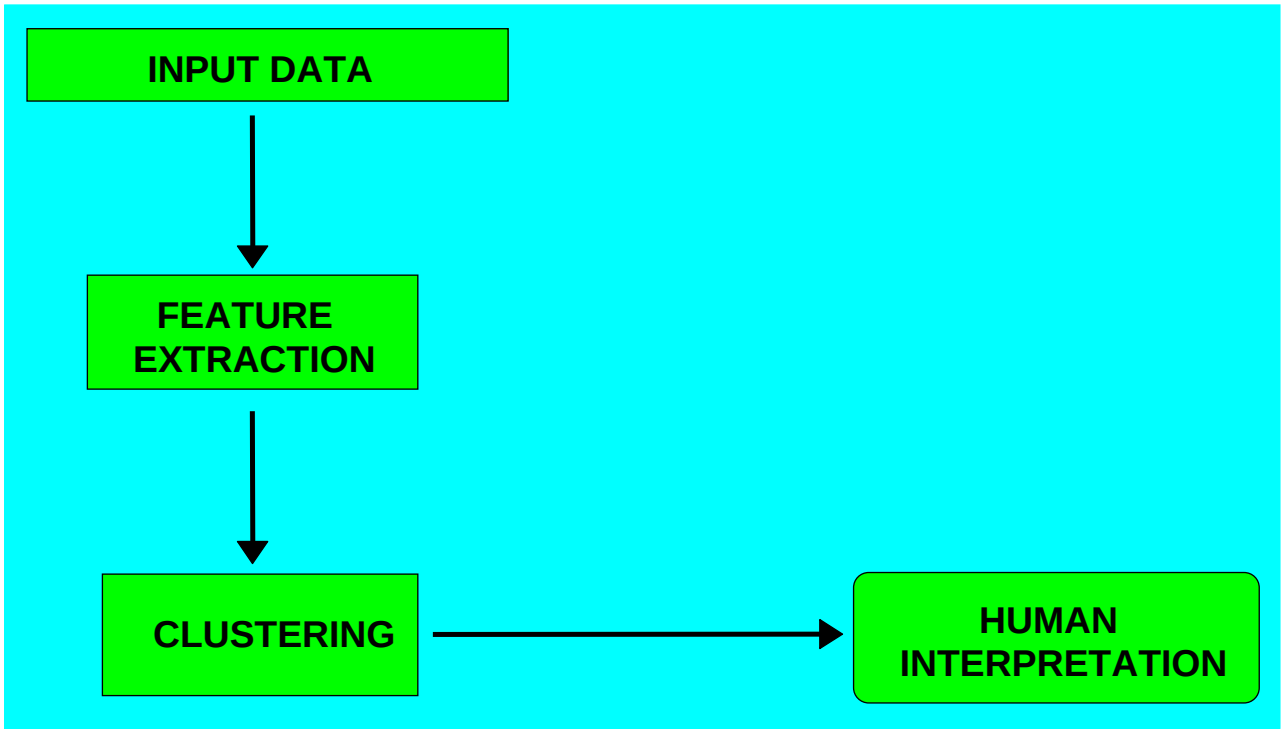
Objects going down in different rows		**X-var 1** Metabolite or peak 1	**X-var 2** Metabolite or peak 2	**X-var 3** Metabolite or peak 3	**Y-var 1** Lots of Metadata	**Y-var 2** Diseased or Healthy (Levels)
Sample 1					Species Age M/F BMI sampling processing etc, etc…	0 (control)
Sample 2…						1 (diseased)

Input data

Output data

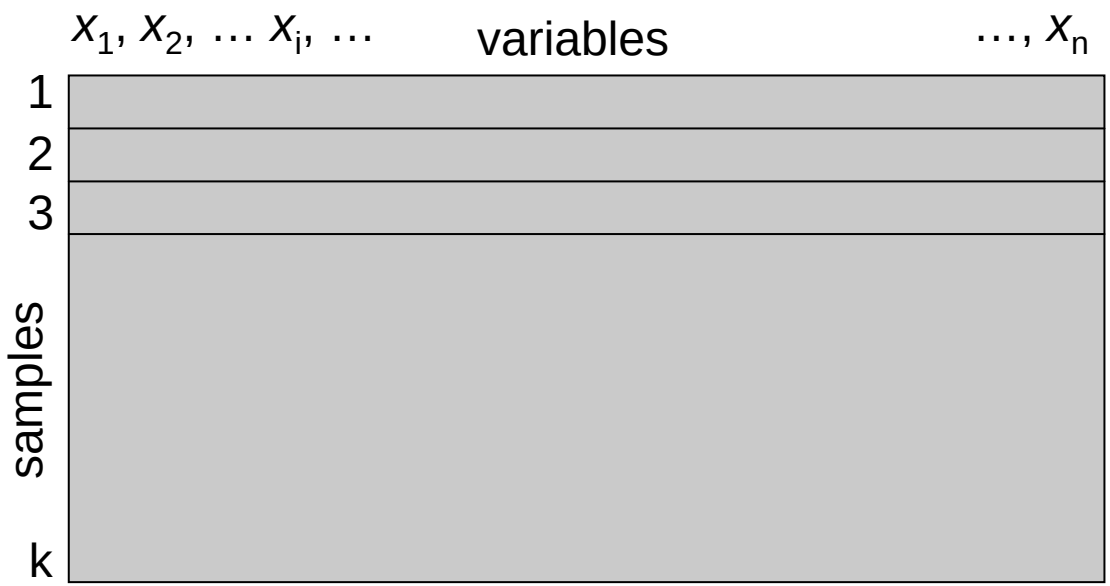
MANCHESTER 1824

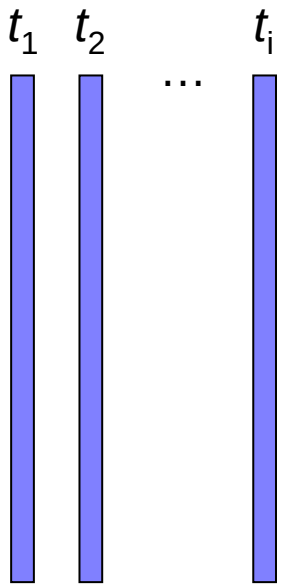
Unsupervised learning



- System is shown a set of inputs (spectra) and then left to cluster the spectra into groups
- This optimization procedure is usually simplification or dimensionality reduction.

Projection of the data

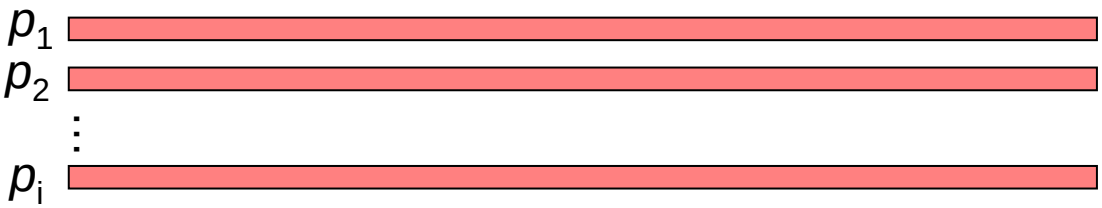




scores (*t*)
summarise variation in samples

uncorrelated orthogonal

variance: $t_1 \geq t_2 \geq \ldots \geq t_i$



loadings (*p*)
summarise variation in variables

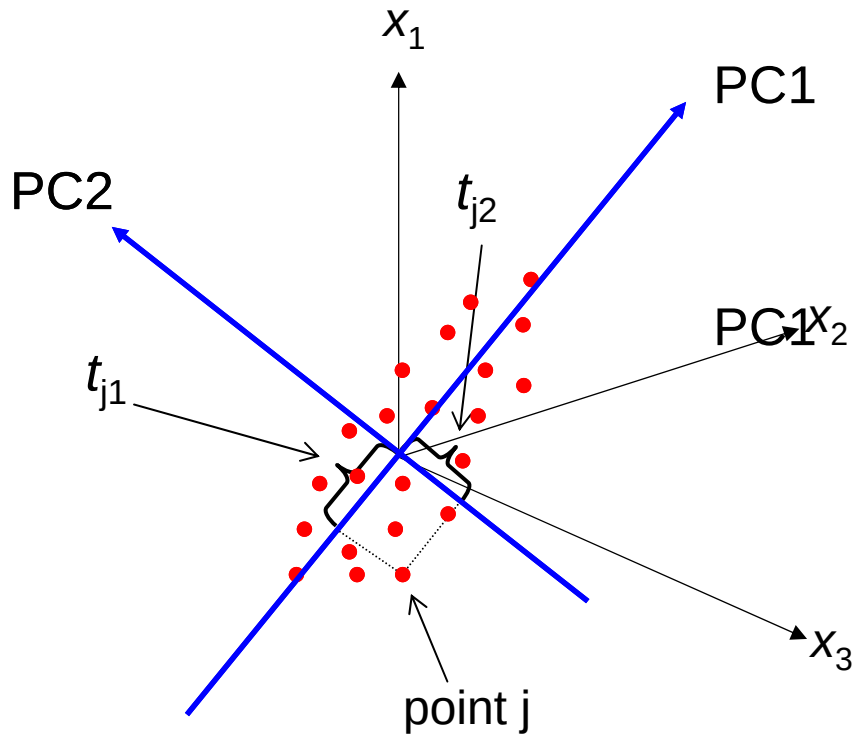
scores = loadings × data

$t_1 = p_1x_1 + p_2x_2 + \ldots + p_ix_i + \ldots + p_nx_n$

MANCHESTER 1824

PCA

- Finds structure in data
- Rotate to uncover *maximum* correlations with respect to *natural* variation

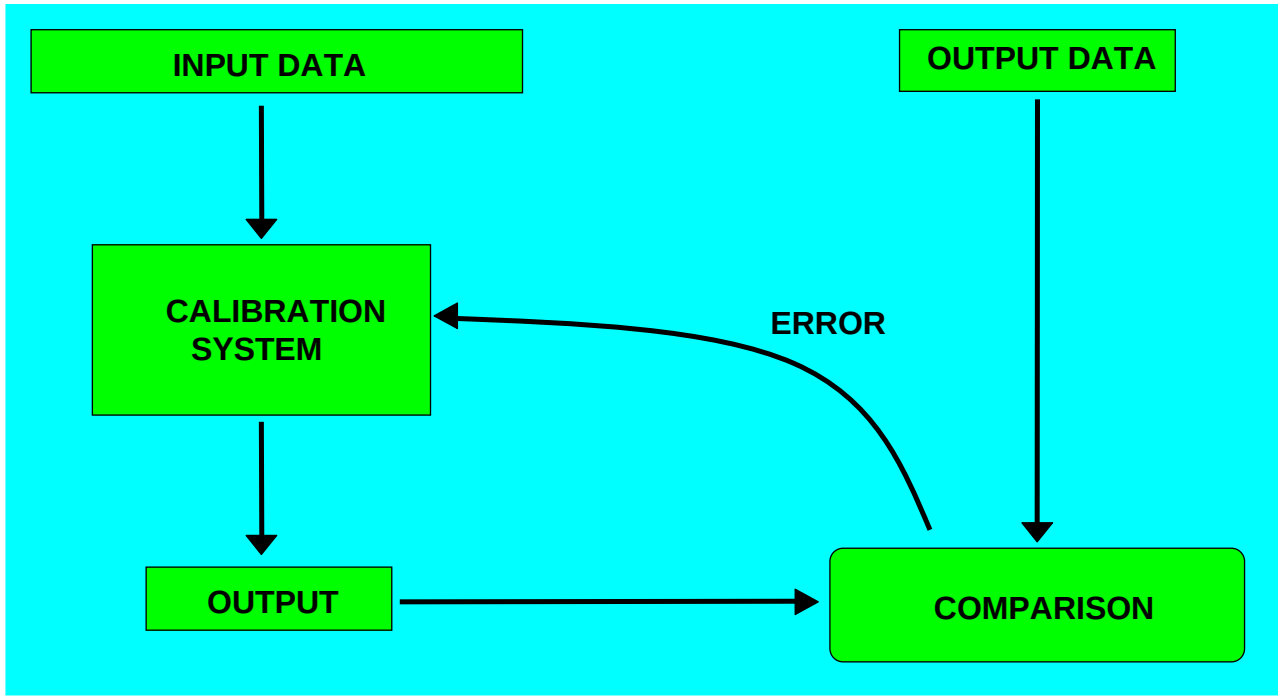


PC1 = 1st principal component
describes largest variance
goes through variable origin space
t_{j1} = score for point j = distance
from projection of that point
onto PC1 from origin

PC2 = 2nd principal component
describes 2nd largest variance
goes through variable origin space
t_j2 = score for point j

MANCHESTER 1824

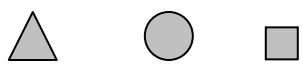
Supervised learning

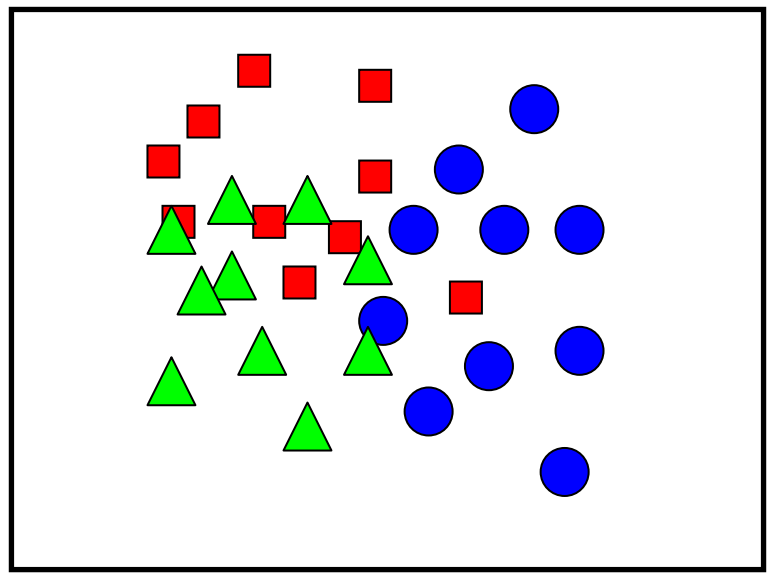


- The goal is to find a mathematical model that will correctly associate the inputs with the targets
- Usually achieved by minimising the error between the target and the model's response (output).

Discriminant function analysis (aka, canonical variates analysis)

- Uses uncorrelated inputs *a priori* information
- Projection based on:
  - Minimises within group variance
  - Maximises between group variance
- Test by projection of ‘unknown’ samples



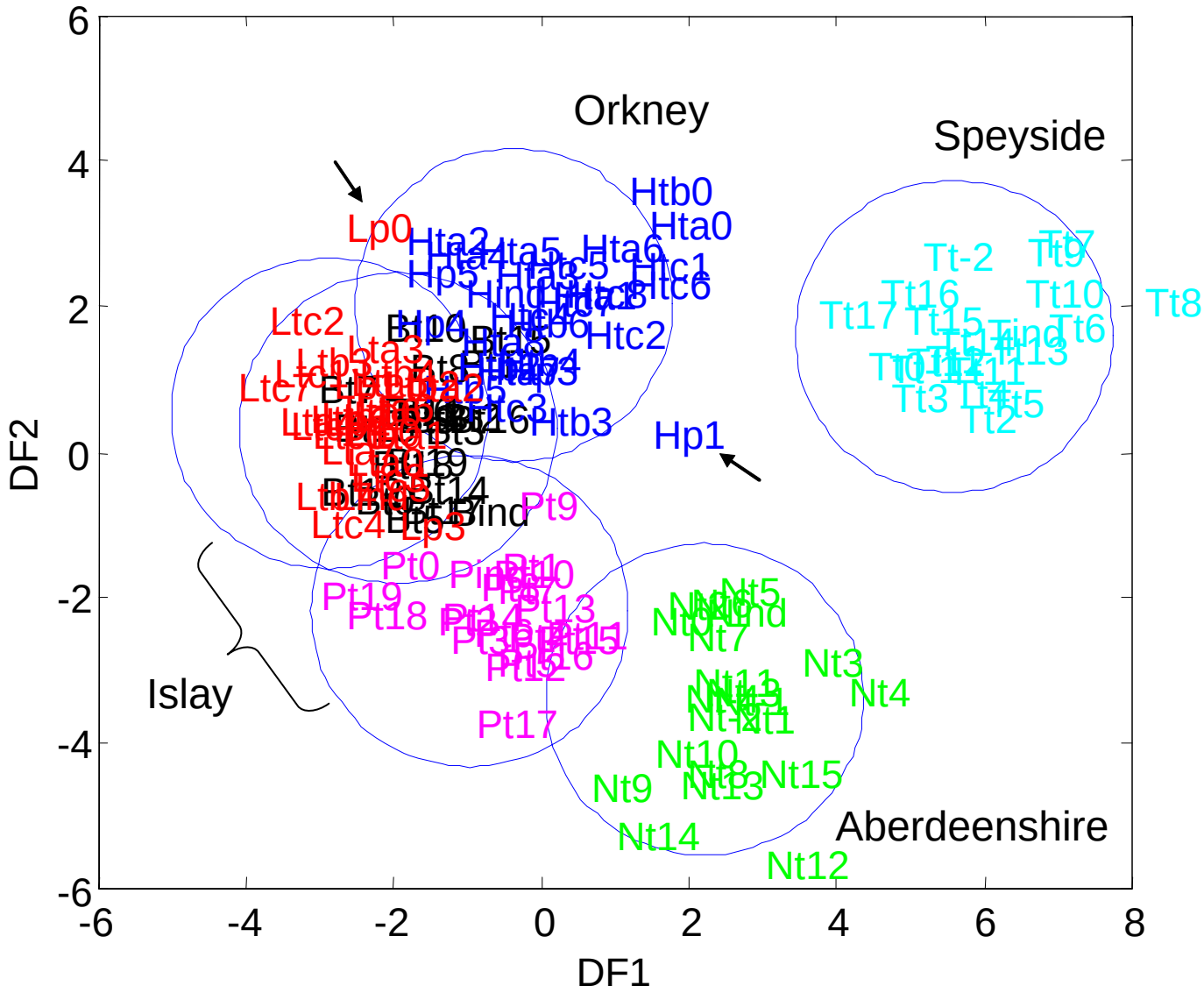


- Statistical significance: χ^2 confidence limits

MANCHESTER 1824

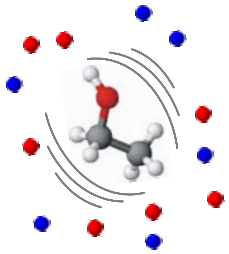
Peat

- 6 groups
- Circles = 95% $\chi 2$ confidence limits
- Arrows represent outlier samples that were from upper horizon of the peat depth profile



Harrison, B., Ellis, J., Broadhurst, D., Reid, K., Goodacre, R. & Priest, F.G. (2006) Differentiation of peats used in the preparation of malt for Scotch whisky production using Fourier transform infrared spectroscopy, *Journal of the Institute of Brewing*, submitted.

MANCHESTER 1824



Target encoding for PLS, ANNs, etc…

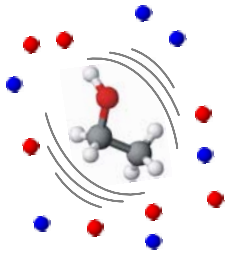
- Usually binary encoded:

Known bacteria

New isolates

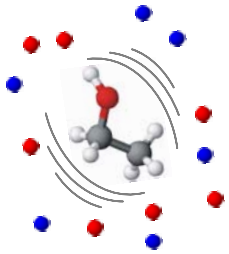
	A	B	C	identity

Easy look up table



Supervised methods are powerful…

- Learn from experience
- Generalise from previous examples to new ones
- Perform pattern recognition on complex multivariate data.
- Make errors
  - usually because of badly chosen data
  - tanks from trees…



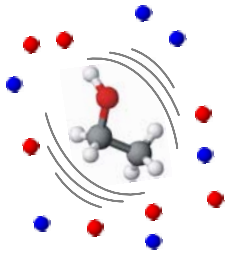
What have I measured that is important

Consider healthy *vs*. diseased

- 2 class problem
- Collect metabolite data from both classes:
  - Encompass biological variation
    - Even distribution
  - Try to keep number of samples the same
    - Statistics are easier / more rigorous

→ next do inductive bit…

MANCHESTER 1824



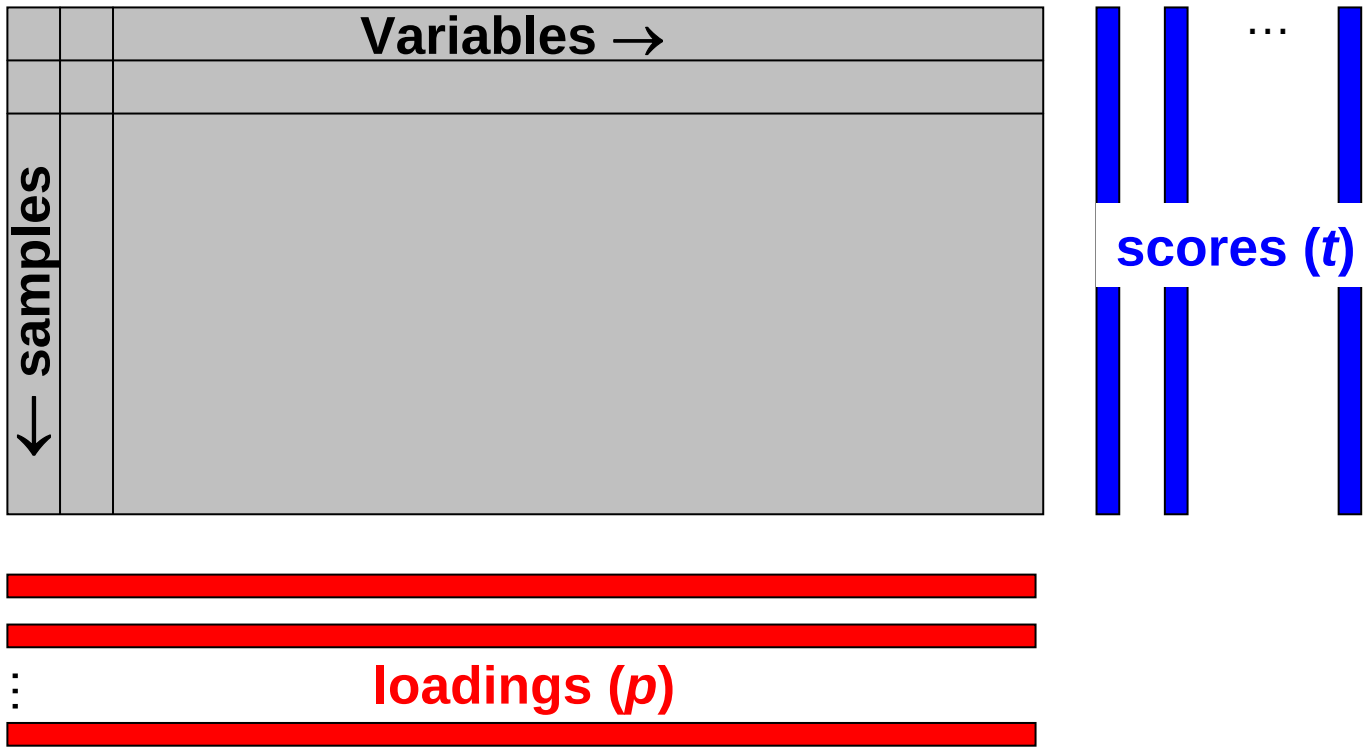
Easy strategies first

- Stare and compare!
  - Always wise to actually look at the data…
- Difference spectra
  - Avg(diseased) - Avg(healthy)
- Univariate analyses
  - Analysis of variance (ANOVA)
  - T-test, etc

MANCHESTER 1824

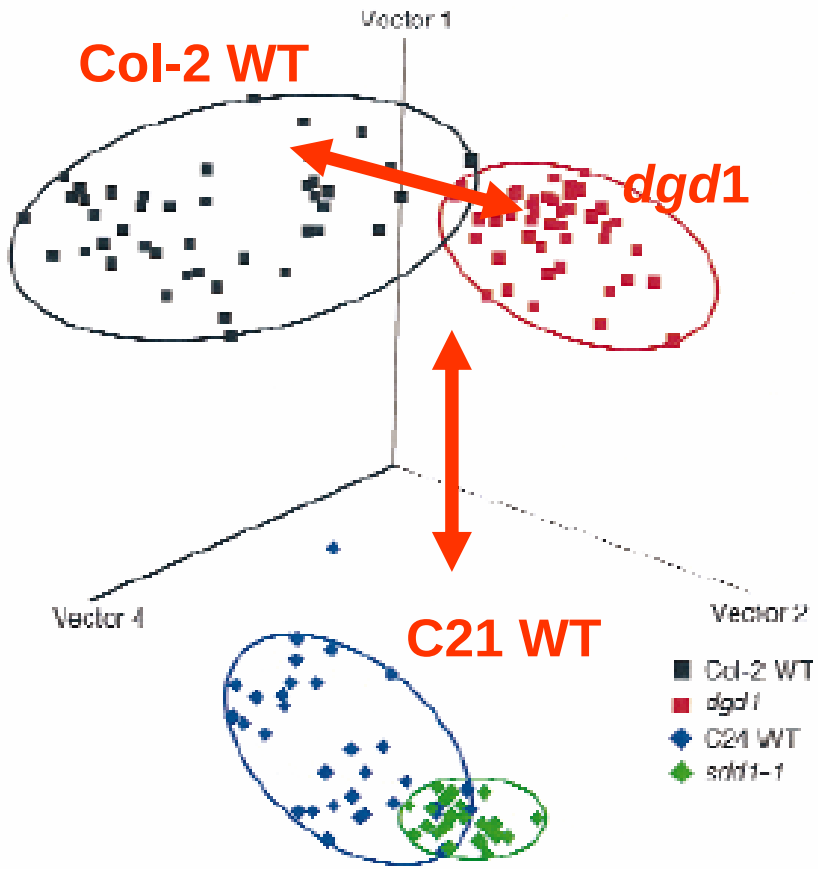
Loadings plots

- Use multivariate analyses and inspect loadings



Types of loadings plots

- Unsupervised
  - PCA
  - SOFM
- Supervised: discriminatory
  - DFA (CVA)
  - PLS-DA
- Supervised: regression
  - PLS
- Supervised: tree-based
  - CART



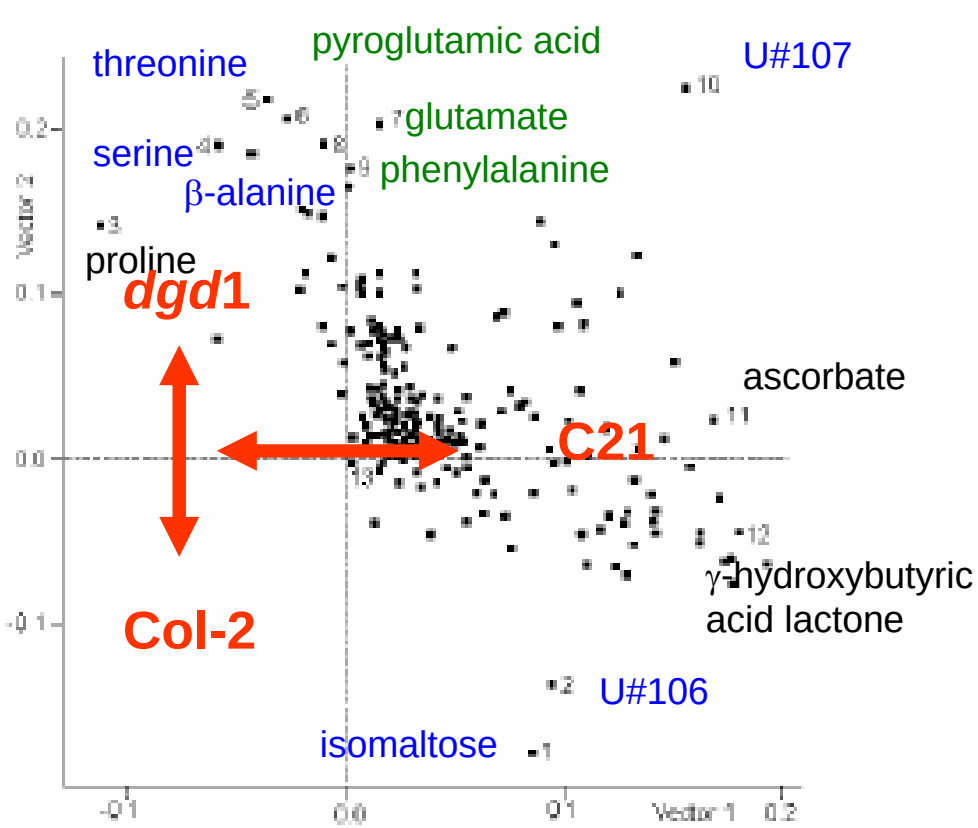
Fiehn *et* al. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* **18**, 1157-1161.

MANCHESTER 1824

Types of loadings plots

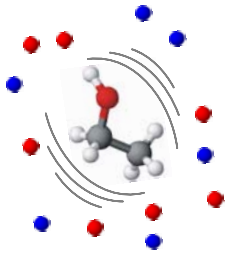
Univariate
$p < 0.01$ *t*-test
p not signif.

- Unsupervised
  - PCA
  - SOFM
- Supervised: discriminatory
  - DFA (CVA)
  - PLS-DA
- Supervised: regression
  - PLS
- Supervised: tree-based
  - CART



But not always that easy…

MANCHESTER 1824

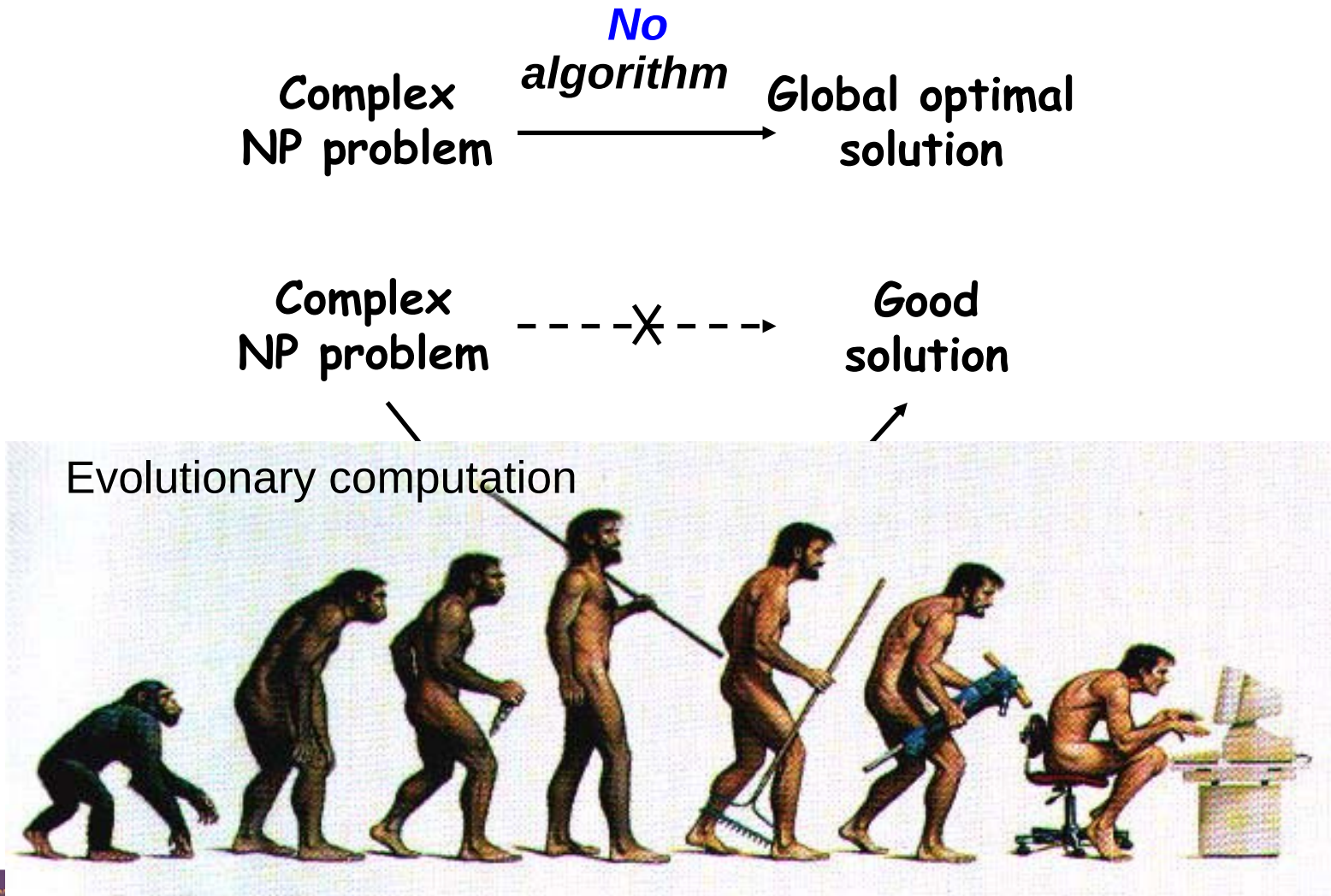


Modest 250 metabolites

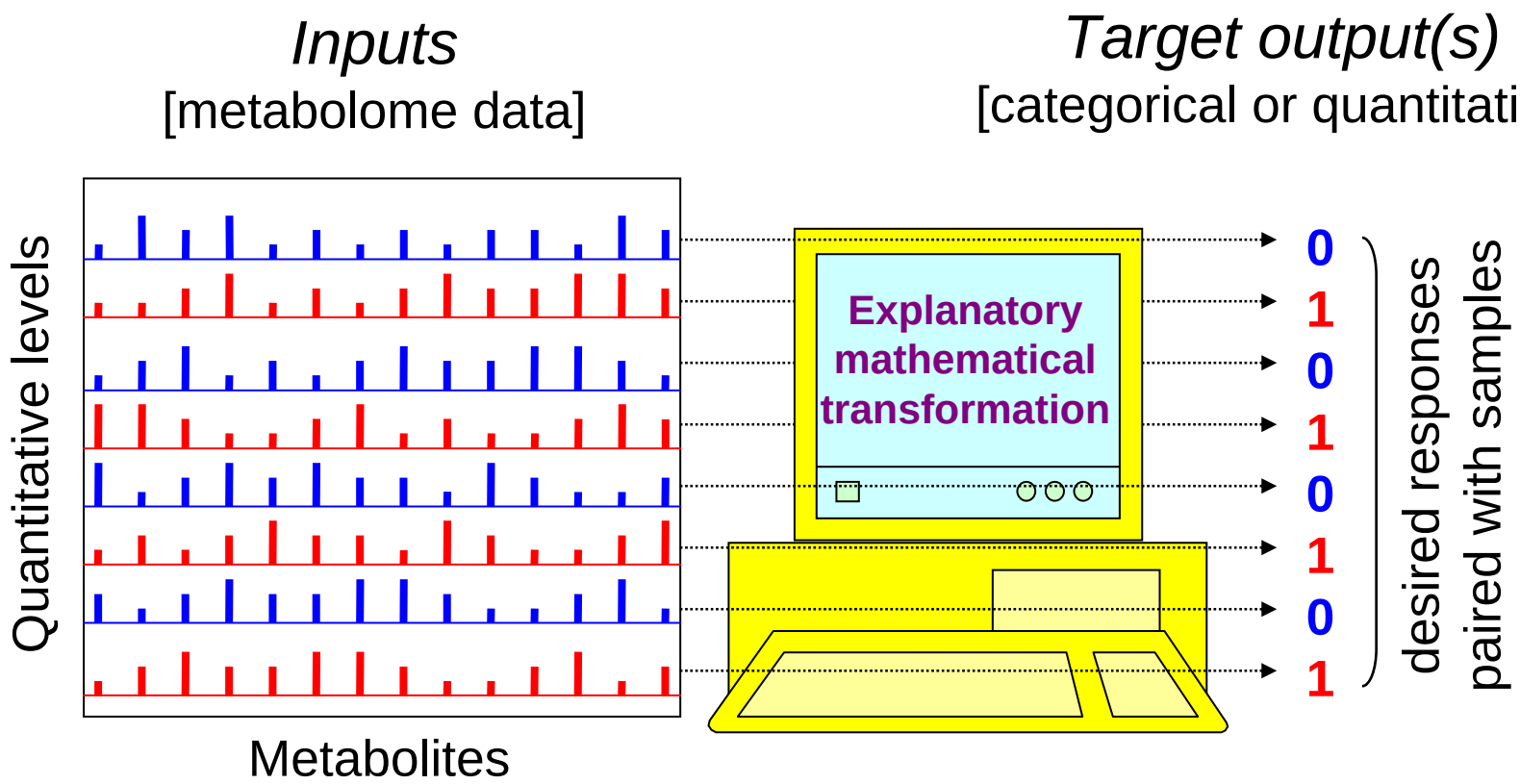
- 2 class problem
  - Healthy *vs*. Diseased
- To use or not to use?
- Yes / No, 250 times = 2^{250} or 1.8×10^{75}
- PC does say 10×10^{6} orderings every second it would still take > 3×10^{62} years

MANCHESTER 1824

Houston we have a problem…



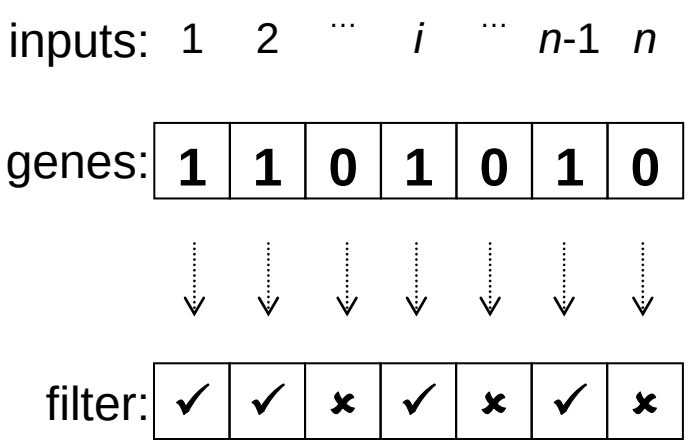
Explanatory machine learning

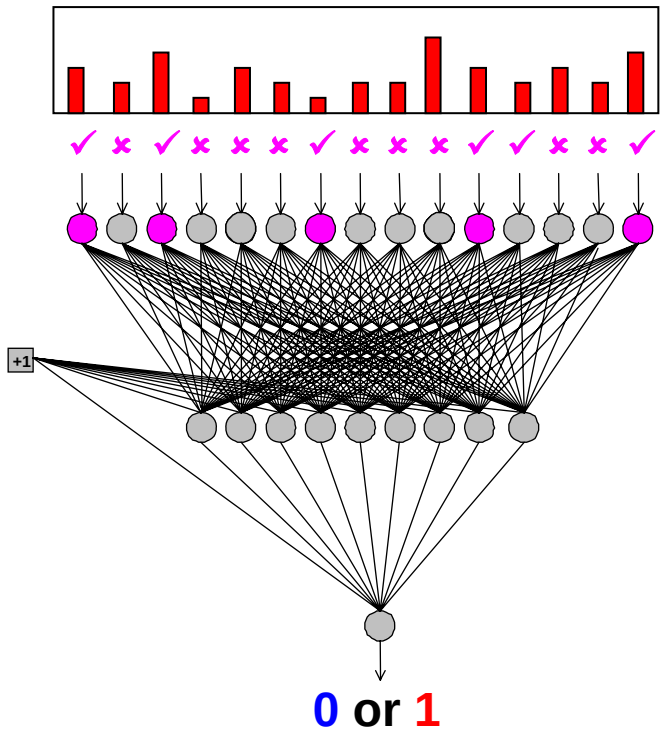


MANCHESTER 1824

Genetic algorithm

- Chromosome with *n* genes selects which of *n* inputs to use in DA, MLR, PLS or ANN



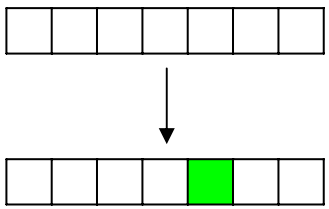


GA: *in silico* evolution

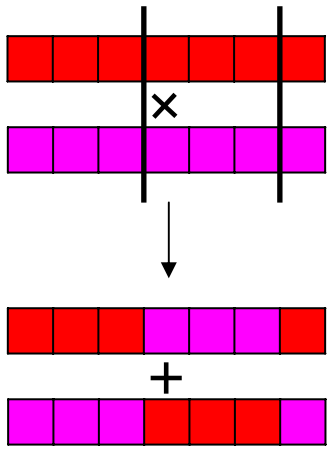
Population of solutions

Assess fitness function of population

Mutation

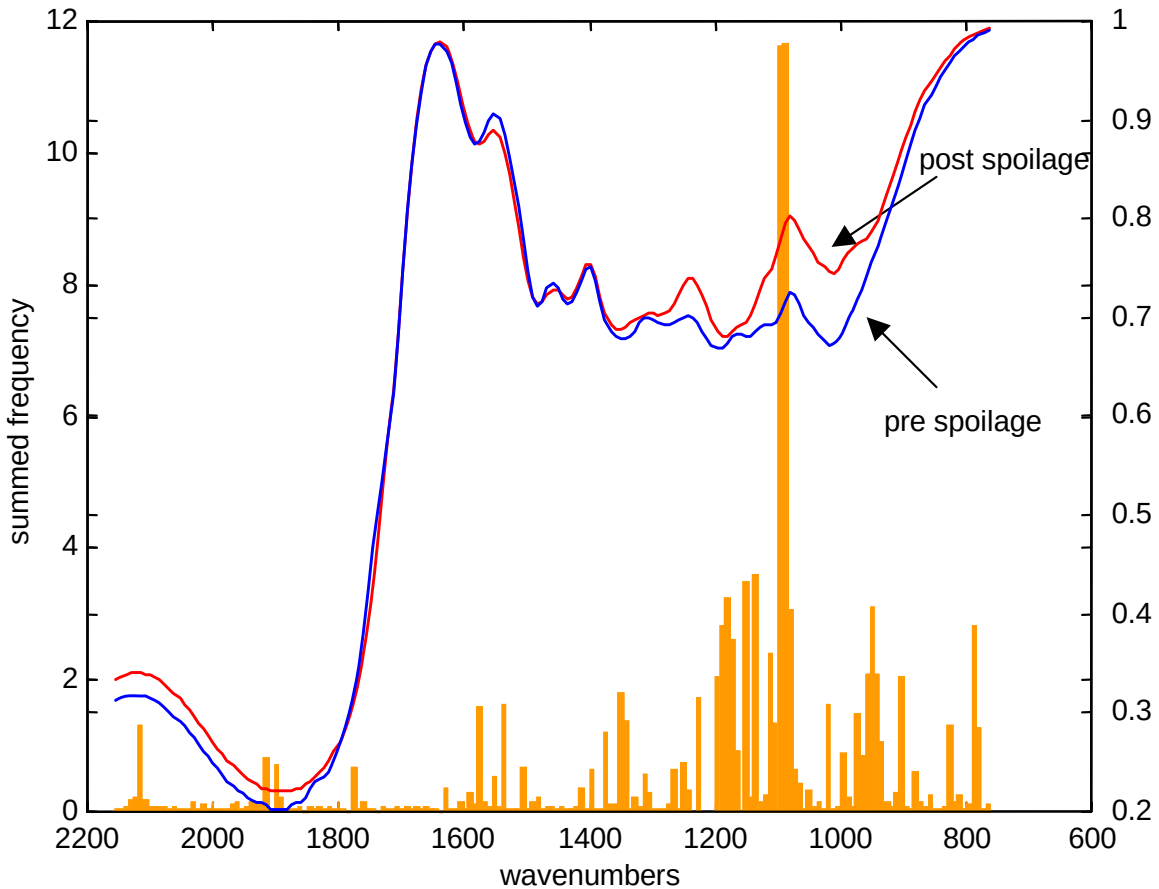


Crossover



Natural selection

Frequency of input selections



Spoilage at 10^7 =

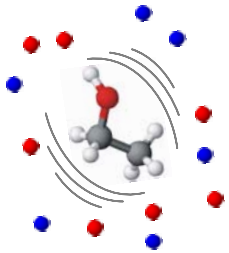
$$\frac{1096\ cm^{-1}}{1683\ cm^{-1}} =$$

$$\frac{\text{Free amines (C-N stretch)}}{\text{Amide I (C=O vibration)}} =$$

Proteolysis of meat by microbes detected by FT-IR

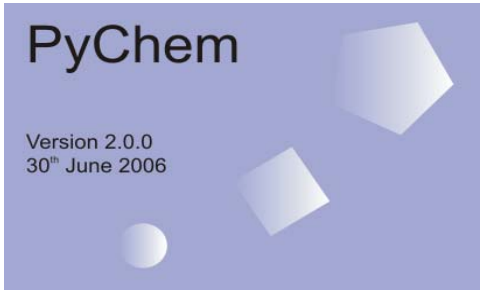
Ellis, D.I., Broadhurst, D., Kell, D.B., Rowland, J.J. & Goodacre, R. (2002) Rapid and quantitative detection of the microbial spoilage of meat using FT-IR spectroscopy and machine learning. *Applied and Environmental Microbiology* **68**, 2822-2828.

MANCHESTER 1824



Overall conclusions on hypothesis generation

- Many approaches can be used ranging from
  - Simple → stare and compare, ANOVA etc
  - Multivariate → loadings: PCA, DFA, PLS
  - More complex → evolutionary computation
- Need to design experiment carefully
- Hypotheses might not always be correct
  - But a good place to start when initial knowledge is non-existent.



PyChem v2.0.0 Beta (http://pychem.org.uk)

- Standalone WinXP **graphical** application for MVA
- Preprocessing algorithms incorporated
- Many standard chemometrics algorithms
  - PCA, HCA, DFA, PLSR, PLS-DA.
  - Feature selection using a variety of genetic algorithm tools
- Coded in Python (http://python.org/) the software is both **FREE** and **OPEN SOURCE**. Source code can be downloaded at http://sourceforge.net/projects/pychem/.
- Can be used to analyse continuous or discrete data such as transcriptomic, metabolomic (vibrational spectra, GC-MS, LC-MS etc…)

Written by Dr Roger Jarvis

PyChem 2.0.0

File Tools Help

Experiment Setup | Spectral Pre-processing | Principal Components Analysis | Cluster Analysis | Partial Least Squares Regression | Discriminant Function Analysis | GA - Discriminant PLS | GA-DFA Clustering

Use: Raw data / Processed data / Principal Component scores

Apply 10 principal components

Extract 3 discriminant functions

Apply full cross validation

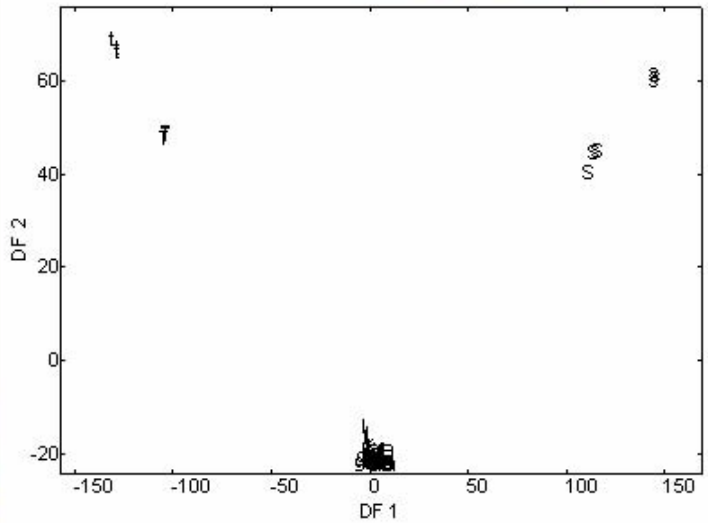
Run DFA

Export

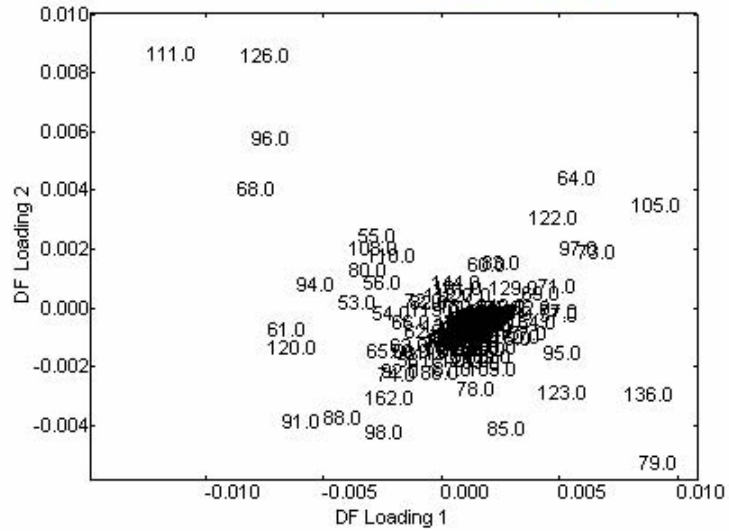
Goto PCA

Goto GA-DFA

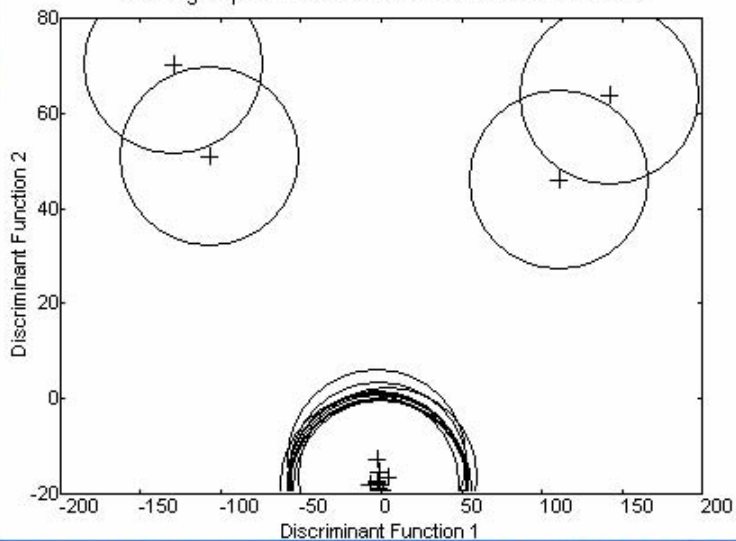
Discriminant Function 1 vs. 2



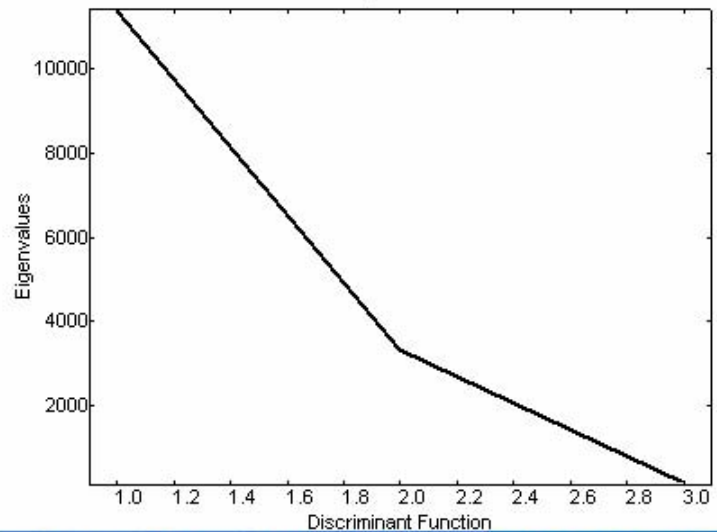
Discriminant Function Loading 1 vs. 2



Mean group centres with 95% confidence intervals



Eigenvalues



start 11:52

PyChem 2.0.0

File Tools Help

Experiment Setup | Spectral Pre-processing | Principal Components Analysis | Cluster Analysis | Partial Least Squares Regression | Discriminant Function Analysis | GA - Discriminant PLS | GA-DFA Clustering

Use: Raw data
Processed data
Principal Component scores
Apply 10 principal components

Extract 3 discriminant functions

Apply full cross validation

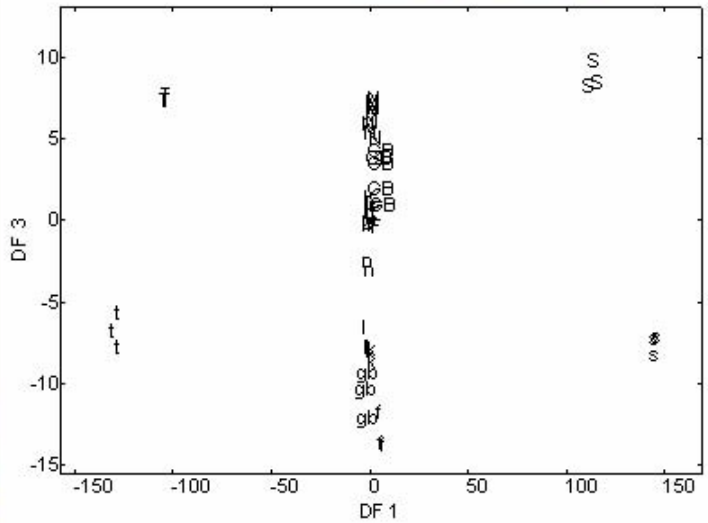
Run DFA

Export

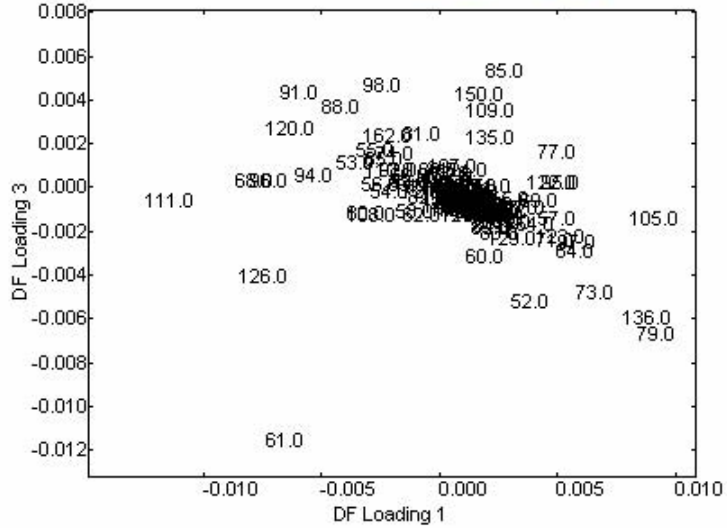
Goto PCA

Goto GA-DFA

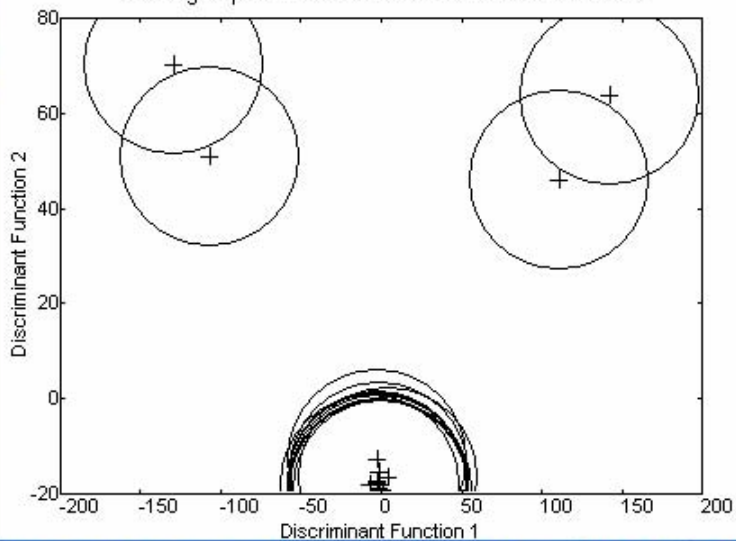
Discriminant Function 1 vs. 3



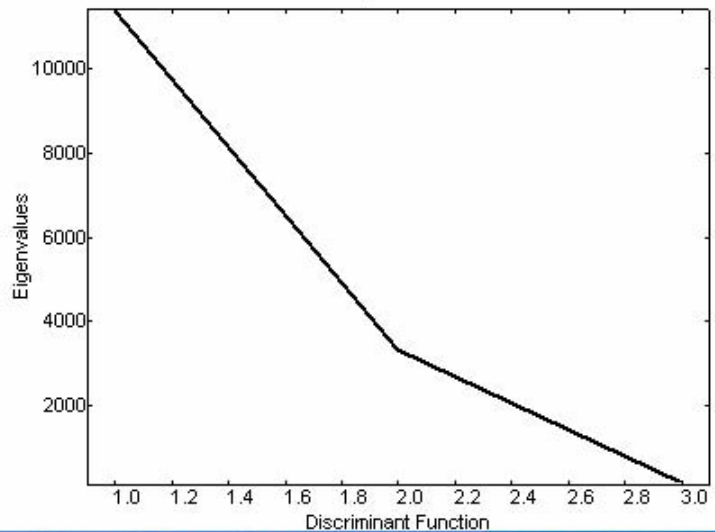
Discriminant Function Loading 1 vs. 2



Mean group centres with 95% confidence intervals



Eigenvalues



start 11:53

PyChem 2.0.0

File Tools Help

Spectral Pre-processing | Principal Components Analysis | Cluster Analysis | Partial Least Squares Regression | Discriminant Function Analysis | GA - Discriminant PLS | GA-DFA Clustering | GA-PLS Calibratic

GA Parameters

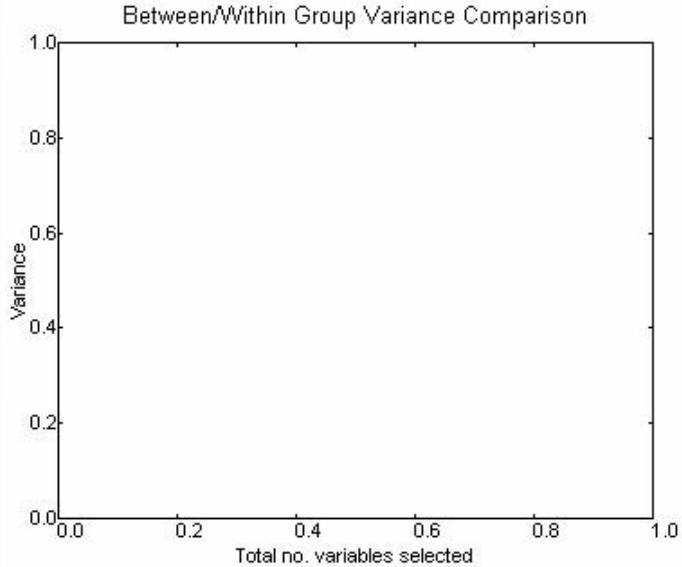
No. vars. from	3	No. vars to	3
No. inds.	5	No. runs	100
Xover rate	0.8	Mut. rate	0.4
Insert. rate	0.8	Max. DFs	2
Max. gen.	500	Rep. until	20

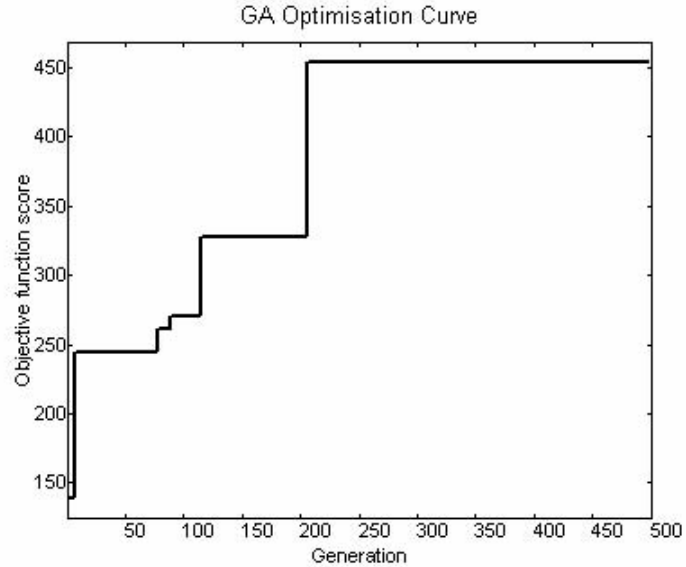
Raw data Processed data

Run GA-DFA Export Results

GA Results

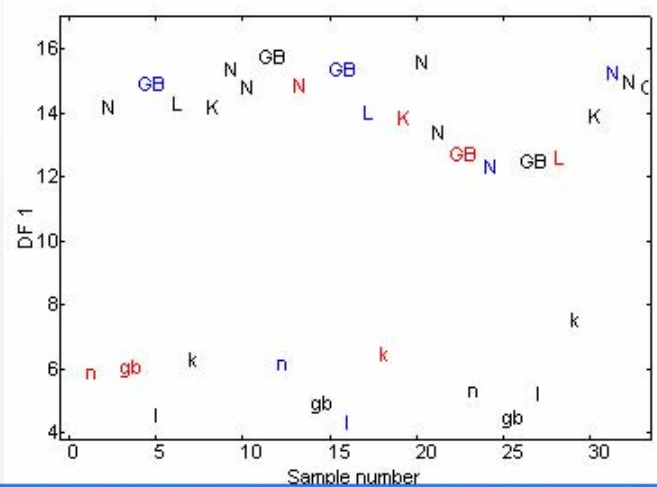
- 3 variables
  - #1 [58. 94. 109.] 352.29
  - #2 [58. 94. 109.] 352.29
  - #3 [58. 94. 109.] 352.29
  - #4 [109. 117. 136.] 454.11
  - #5 [58. 94. 109.] 352.29
  - #6 [57. 109. 120.] 324.76
  - #7 [58. 94. 109.] 352.29
  - #8 [69. 109. 136.] 327.45
  - #9 [109. 117. 136.] 454.11
  - #10 [58. 67. 109.] 343.35
  - #11 [58. 73. 109.] 336.46
  - #12 [109. 117. 136.] 454.11
  - #13 [89. 109. 117.] 392.02
  - #14 [58. 94. 109.] 352.29
  - #15 [58. 94. 109.] 352.29
  - #16 [58. 73. 109.] 336.46
  - #17 [67. 117. 136.] 331.19
  - #18 [109. 117. 136.] 454.11
  - #19 [82. 117. 136.] 338.26
  - #20 [58. 94. 109.] 352.29
  - #21 [58. 67. 109.] 343.35
  - #22 [69. 109. 136.] 327.45
  - #23 [109. 117. 136.] 454.11

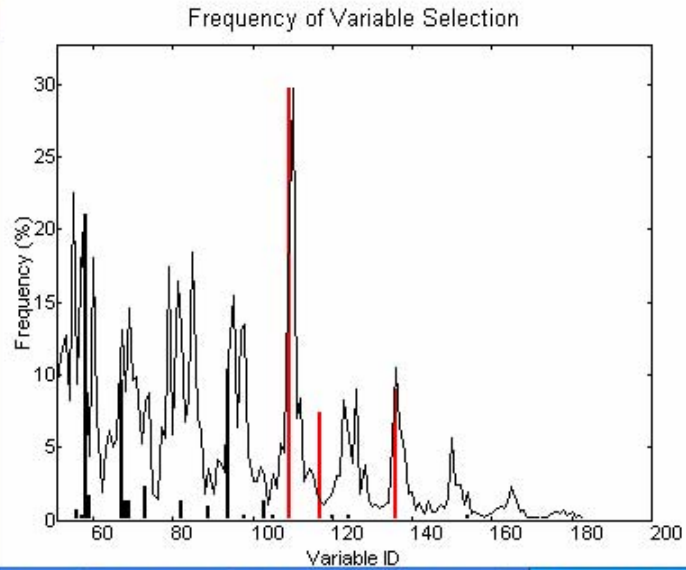




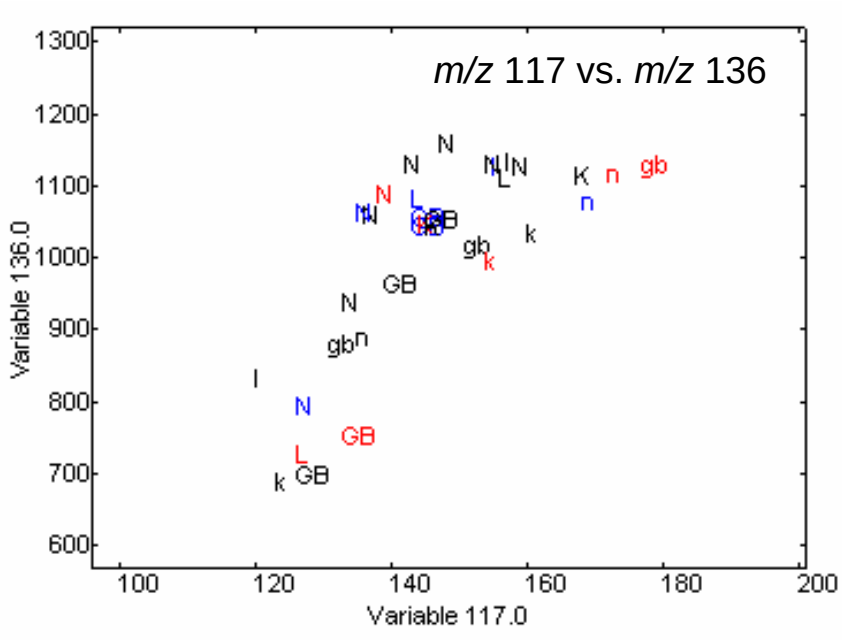
2D Ordination Plot | Eigenvalues | Spectral Loadings | Feature Plot

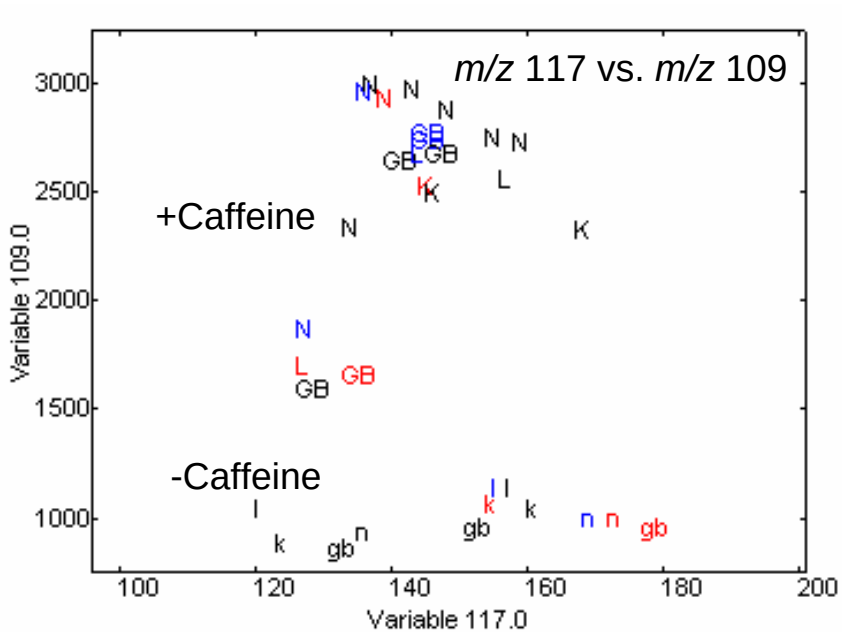
Discriminant Function 1 v. 1

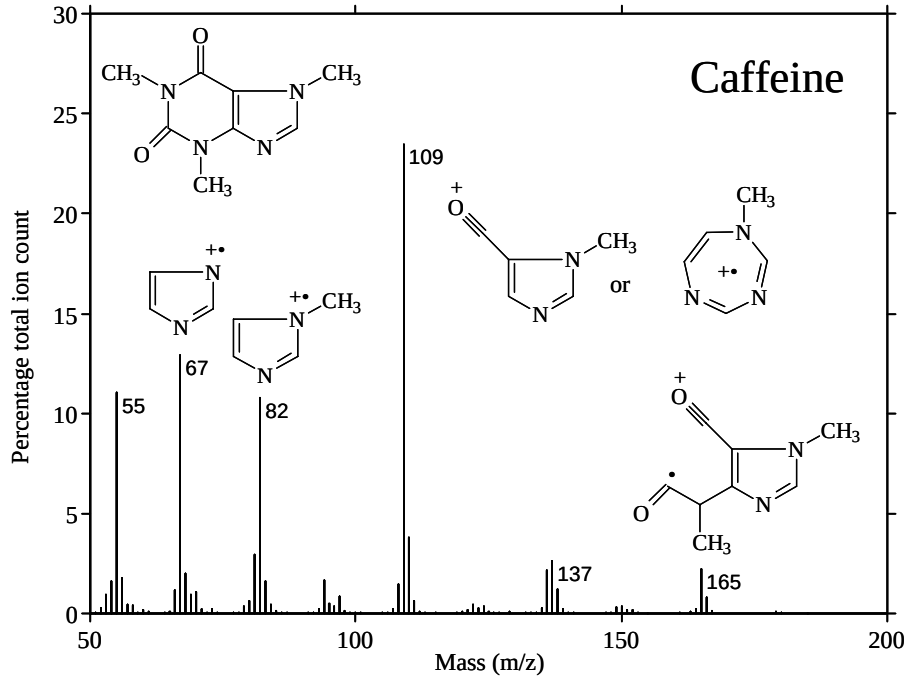


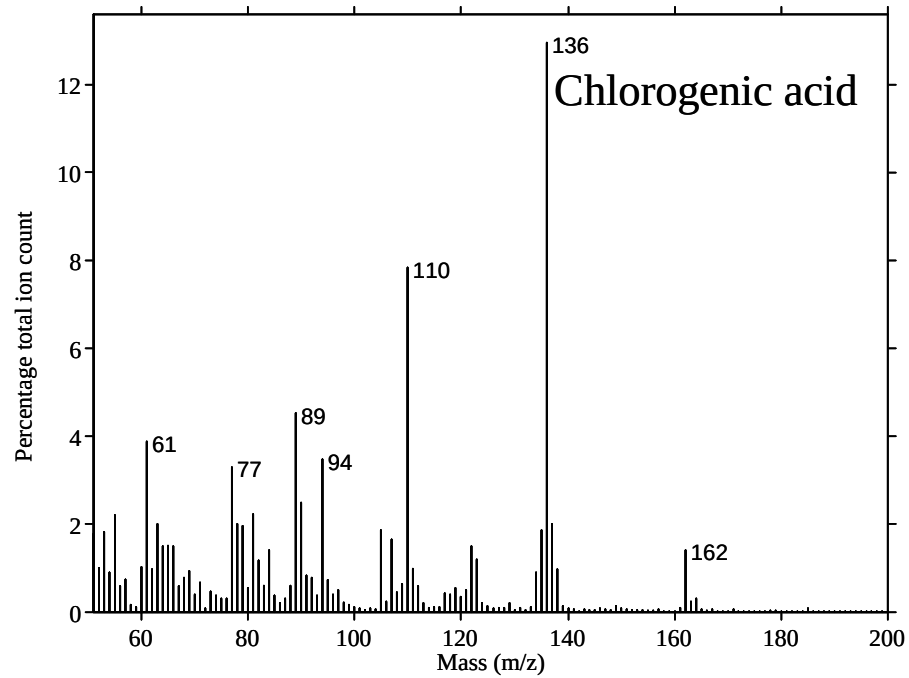


start 2 W... Micr... Me... PyC... Peg... Micr... 16:46


m/z 117 vs. m/z 136
Variable 136.0
Variable 117.0


m/z 117 vs. m/z 109
+Caffeine
-Caffeine
Variable 109.0
Variable 117.0


Caffeine
Percentage total ion count
Mass (m/z)
109
67
55
82
137
165
or


Chlorogenic acid
Percentage total ion count
Mass (m/z)
136
110
89
61
77
94
162

On (stamp) collecting data

"Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house"

Jules Henri Poincaré (1854-1912) *La Science et l'hypothèse*

Having the parts list doesn't mean that you know how an organism works