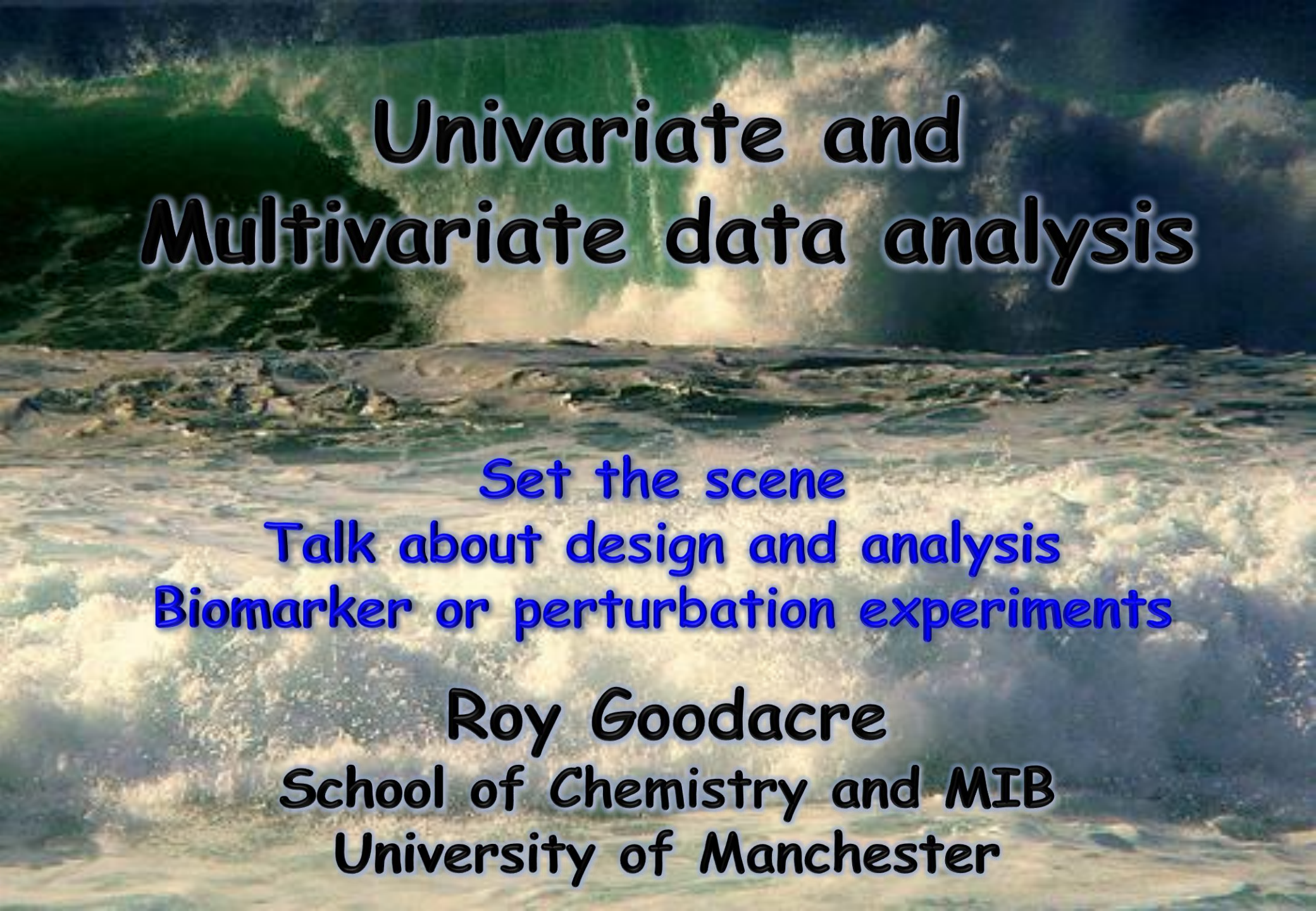
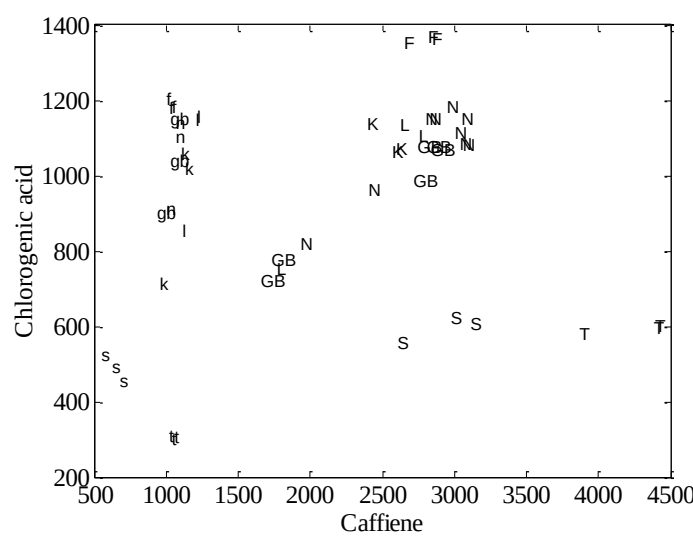


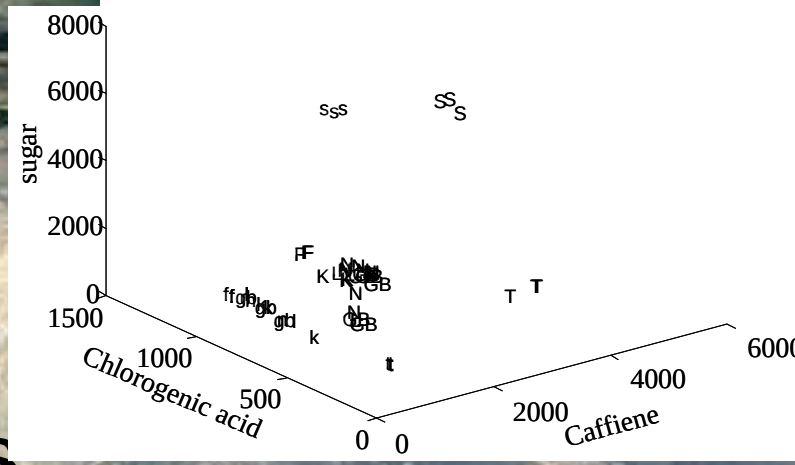
Univariate and Multivariate data analysis

Set the scene
Talk about design and analysis
Biomarker or perturbation experiments

Roy Goodacre
School of Chemistry and MIB
University of Manchester

Mega- & Multi-variate data





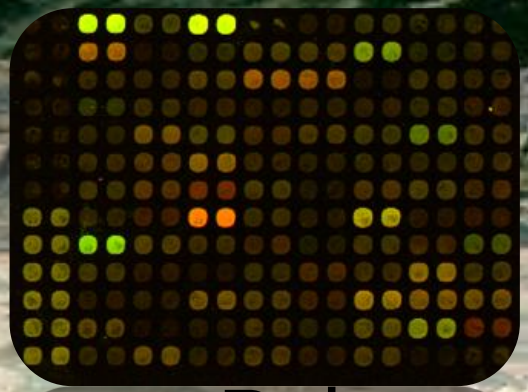
A sample resides somewhere in 2D or 3D space

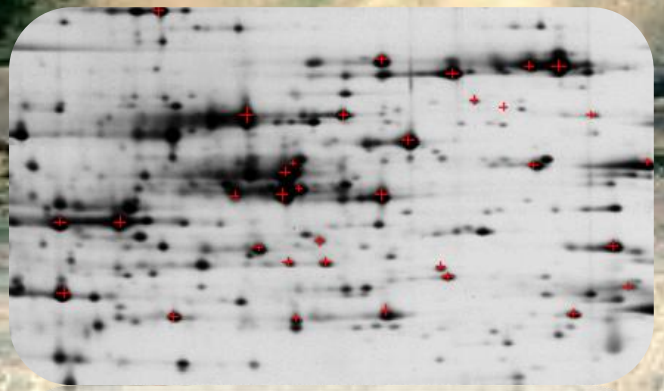
But if one collects 100 variables…

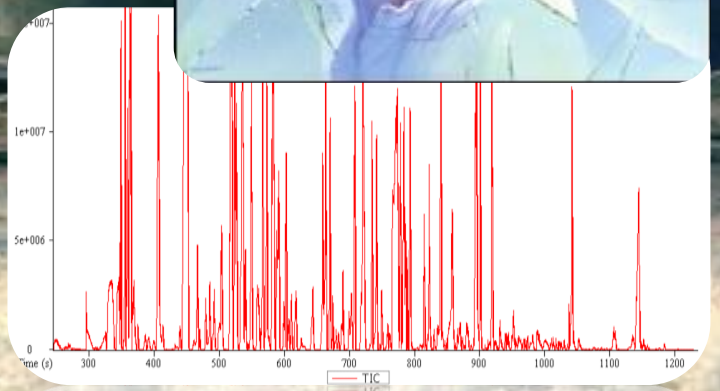
Need to visualise 100 D space!

underlying theme of multivariate analysis (MVA) is thus *simplification* or *dimensionality* reduction

Data floods





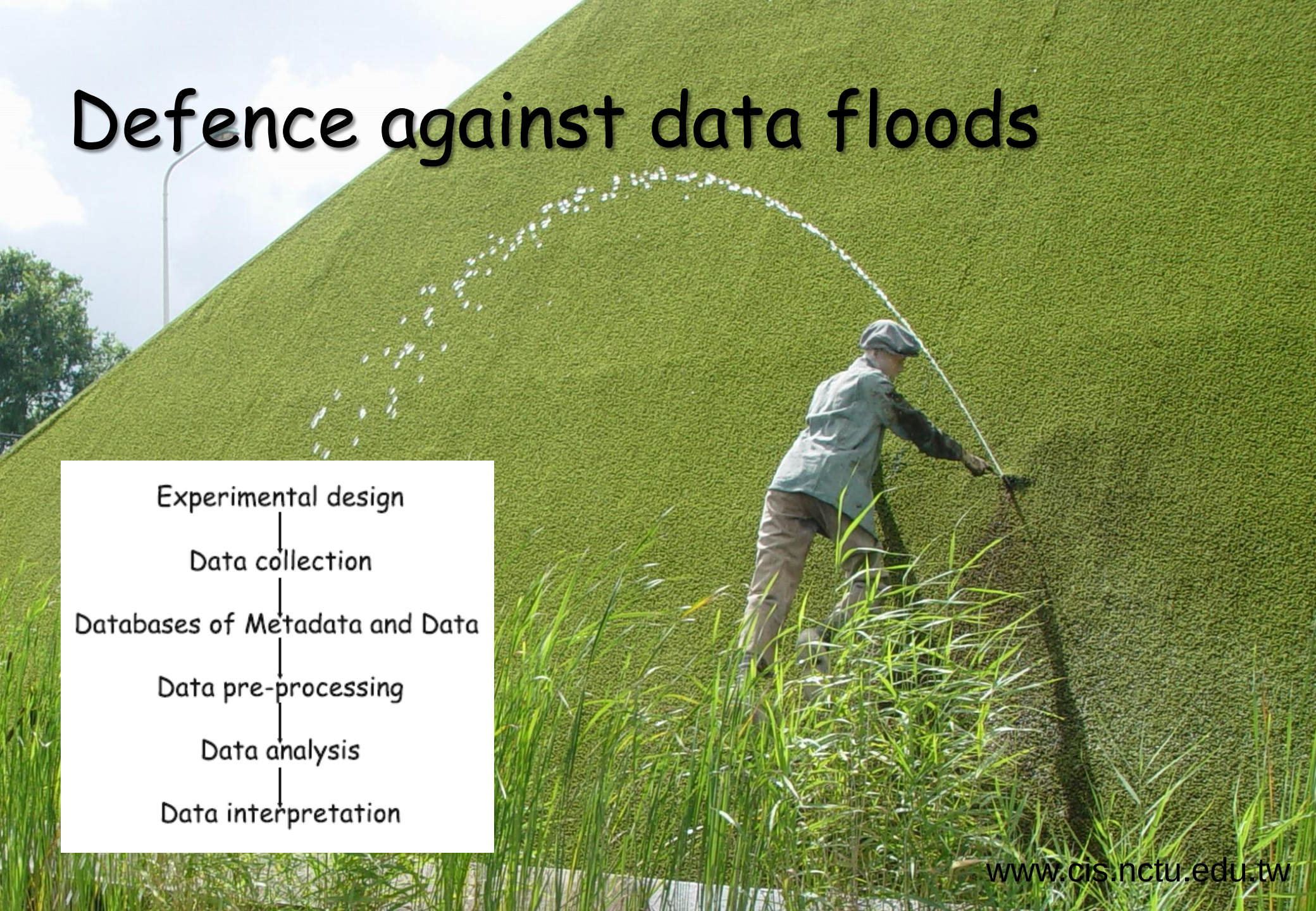


Pairs:

Identifier: transcript / protein / metabolite

Quant Info: concentration or ratio

Defence against data floods



Is this a good experimental design?

- Liver failure from plasma
- Metabolome measured with GC-MS & LC-MS

Table 1. Demographic Information of the Healthy Group and Liver Failure Patient Group Investigated[a]

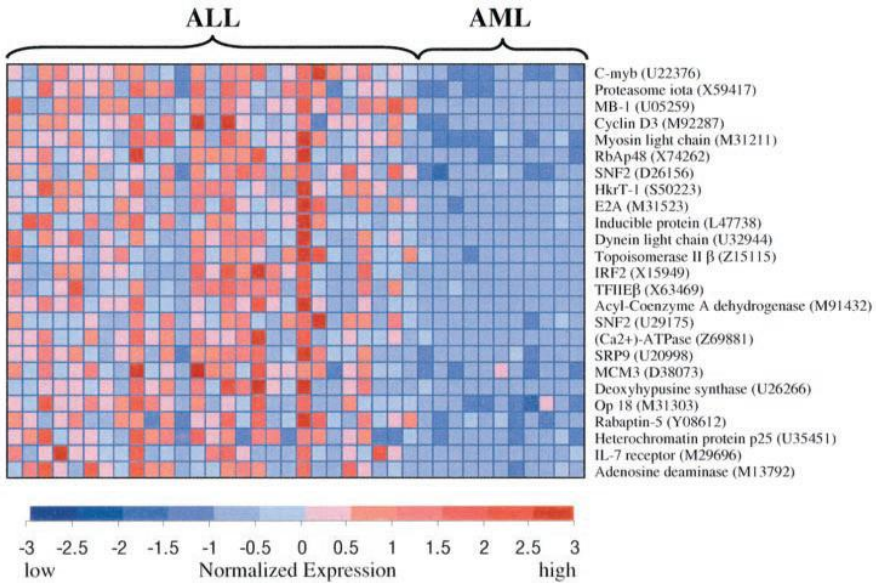
	healthy group ($n = 23$)	patient group ($n = 24$)
Gender (male/female)	15/8	21/3
HBsAg	Negative	Positive
Age (year)	27.39 ± 9.24	46.77 ± 13.35
ALT (U/L)	<40	172.63 ± 147.49
TB (μmol/L)	<12	457.33 ± 135.48
PT (s)	<14	26.06 ± 15.14
MELD score	/	24.68 ± 8.38

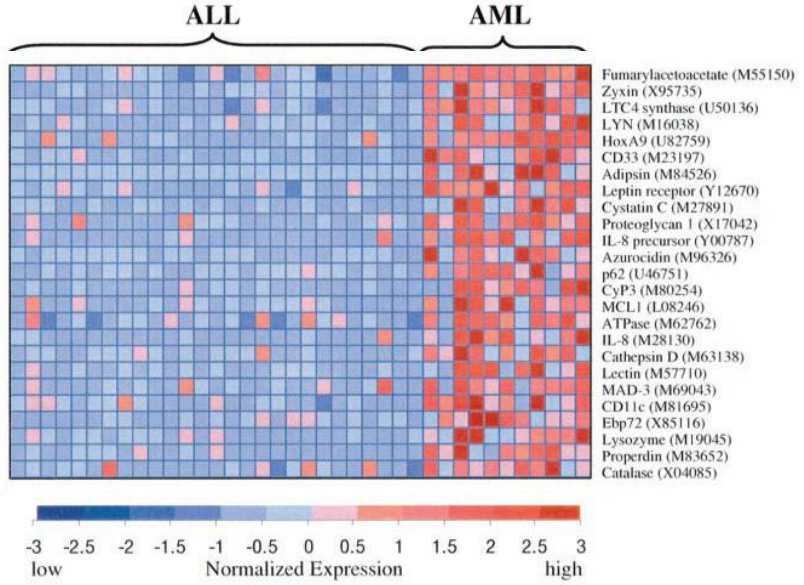
[a] Abbreviations: ALT, alanine aminotransferase; TB, total bilirubin; PT, prothrombin time; MELD, model for end-stage liver disease. The value is represented as the form of mean ± SD.

Lawton, K.A. (2008) *Pharmacogenomics* **9**, 383-397

Is this a good experimental design?

- 27 acute lymphoblastic leukemia — Childhood
- 11 acute myeloid leukemia — Adult
- Affymetrix arrays with 6,817 probes





Golub *et al*. (1999) *Science* **286**, 531-537

Sample size bias

- Should try to have even cohort sizes
- Using a large sample size does not directly address bias, although it can reduce statistical uncertainty by providing a smaller confidence interval around a result.

Is this a good experimental design?

- Pre-eclampsia: Pregnancy-induced hypertension
- Metabolomics: GC-MS of serum

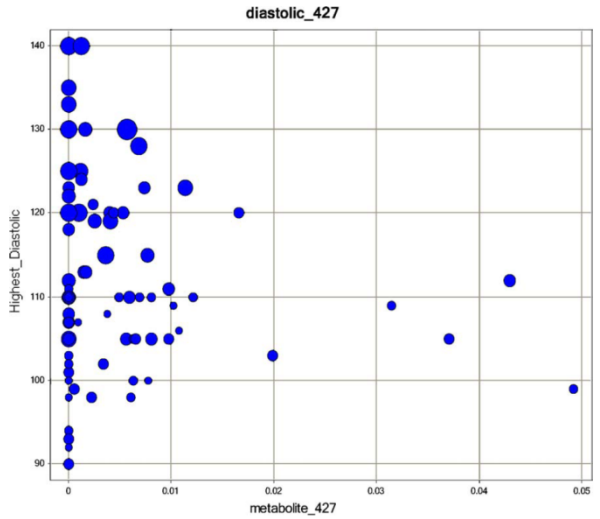
Demographic data for patients from whom plasma samples were taken

	Normal outcome $n=87$	Preeclampsia $n=87$
Age	30 (19–43)	31 (19–41)
Parity	0 (0–2)	0 (0–2)
BMI (weight/height^2)	25 (19–46)	26 (18–46)
Max (S) BP (mm Hg)	122 (96–147)	162 (138–220)*
Max (D) BP (mm Hg)	80 (60–93)	110 (90–140)*
Delivery gestation (weeks + days)	40 + 4 (34 + 3 to 42 + 0)	37 + 0* (26 + 3 to 41 + 1)
Birth weight (g)	3420 (2380–4420)	2410 (590–4300)*
IBR (centile)	34 (10–99)	8 (0–99)*

Median (range).
Pre-eclampsia vs normal outcome.
*$p < 0.0001$.

Measuring BP?

Markers found did not correlate with BP



Kenny L.C. *et al*. (2005) *Metabolomics* **1**, 227-234

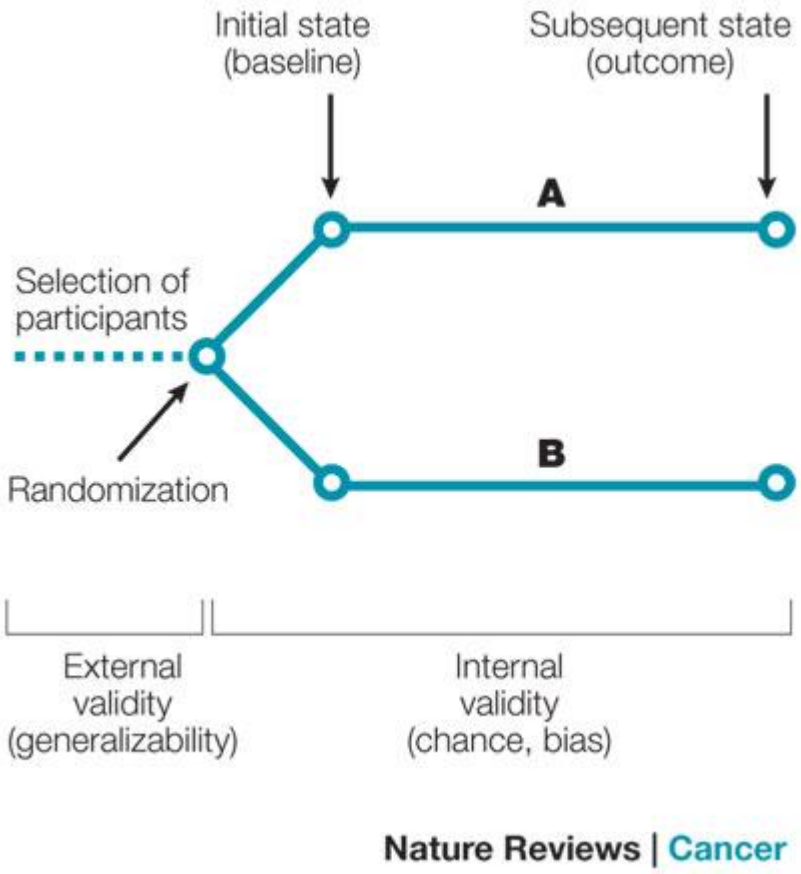
Ronald A. Fisher (1938)

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of"

* and this is why most claimed research findings are false

*Broadhurst, D. & Kell, D.B. (2006) *Metabolomics* **2**, 171-196

Statistician needed at onset



- Was randomisation successful: Check!
- In case-control studies only non-random thing should be what you are testing for

Ransohoff, D.R. (2005) *Nature Reviews Cancer* **5**, 142-149

All about Null hypothesis (H_0)

- Given the test scores of two random samples of guilty people and innocent people, does one group differ from the other?
- A possible null hypothesis is that the mean score for guilty is the same as the mean innocent score. In other words H_0: $\mu_1 = \mu_2$

	Null hypothesis (H_0) is true	**Null hypothesis (H_0) is false**
Reject null hypothesis	False positive [Type I error]	True positive [Correct outcome]
Fail to reject null hypothesis	True negative [Correct outcome]	False negative [Type II error]

- Type I error can be thought of as “convicting an innocent person”
- Type II error “letting a guilty person go free”

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…				

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

Input data → **Chemometrics**

Unsupervised methods

- Use X data only
  - X data = transcript/protein/metabolite levels
  - Inputs to some analysis method
- Most common methods
  - Principal components analysis (PCA)
  - Clustering methods
  - Kohonen neural networks

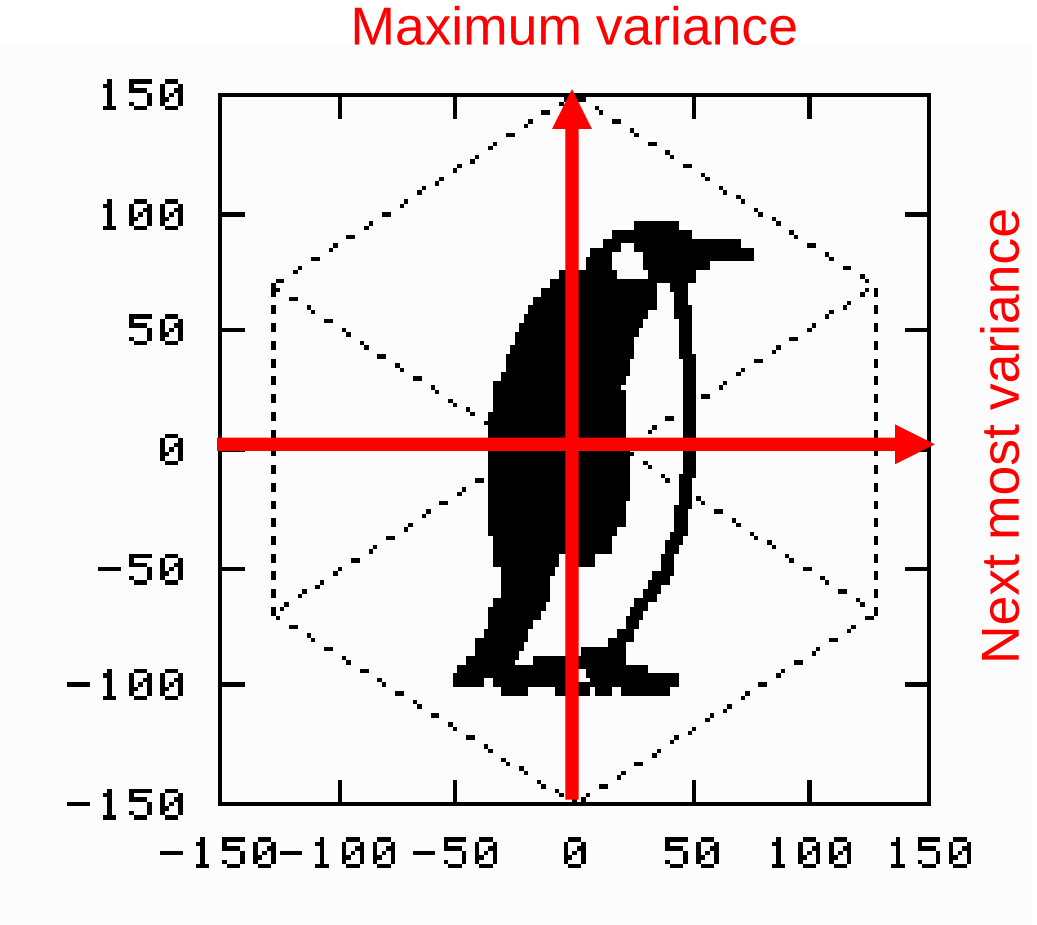
Uncovering correlations in data

- Correlations between *x* variables are confusing.
- Need to examine the *structure* within data sets, rather than using them blindly.
- Finding such structure by hand can be extremely difficult, even in relatively simple cases.

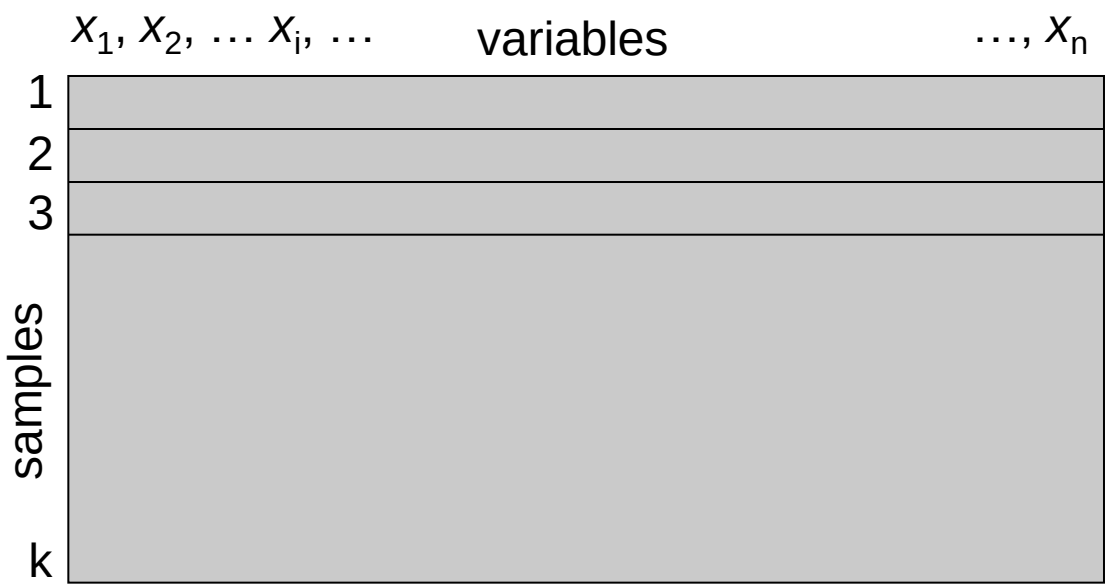
→ use projections

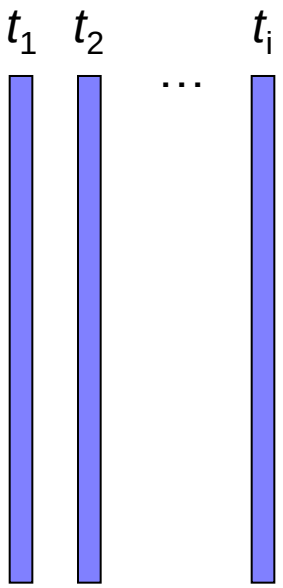
Who's there?

- Data → random mess?
- On rotation of the data …



Projection of the data

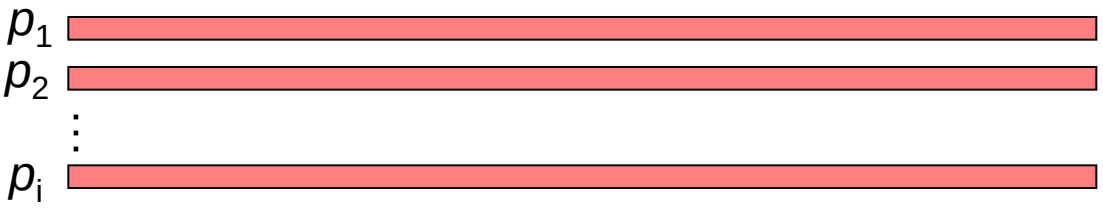




scores (*t*)
summarise variation in samples

uncorrelated orthogonal

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



loadings (*p*)
summarise variation in variables

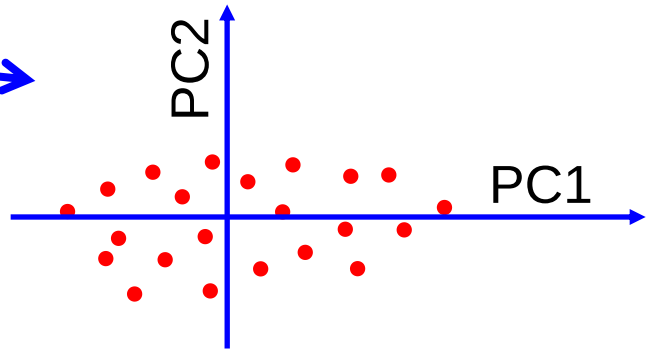
scores = loadings × data

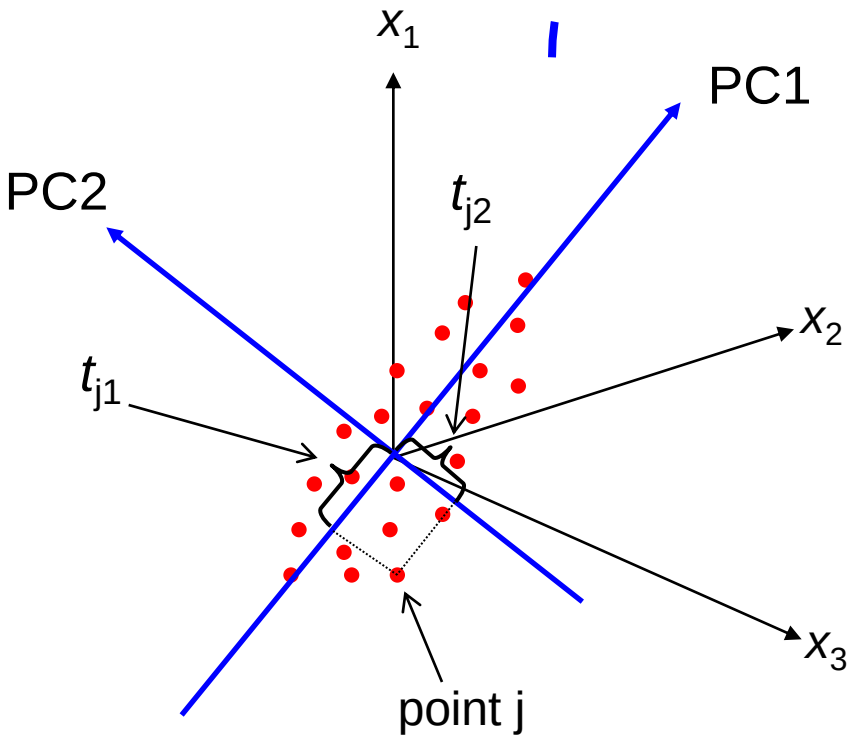
$t_1 = p_1x_1 + p_2x_2 + \ldots + p_ix_i + \ldots + p_nx_n$

Principal components analysis

- Finds structure in such data sets.
- An old method
  - Karl Pearson (1901) → Hotelling (1933)
- Rotate to uncover *maximum* correlations
- First axis placed along the most *natural variation*
- Second axis orthogonal to this to find 2nd highest correlation, and so on
- Plot axes → spot major underlying structures automatically

PCA



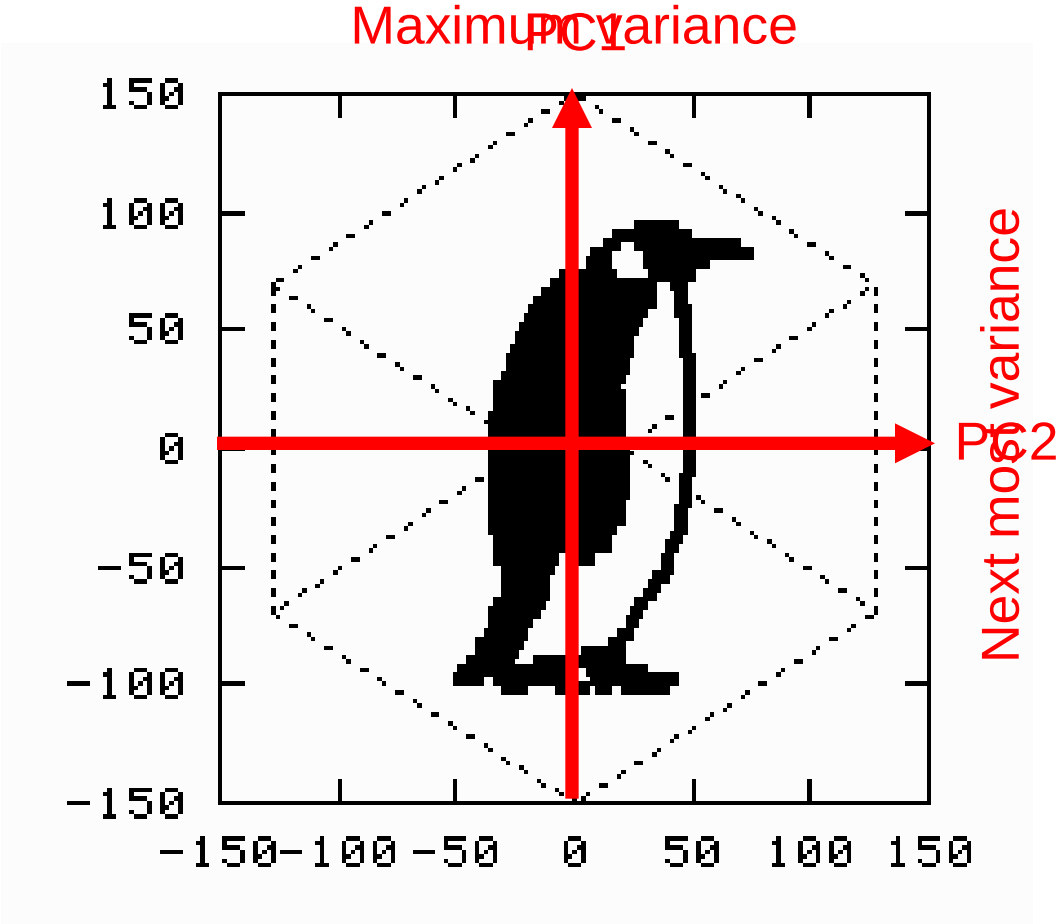


PC1 = 1st principal component
describes largest variance
goes through variable origin space
t_{j1} = score for point j = distance
from projection of that point
onto PC1 from origin

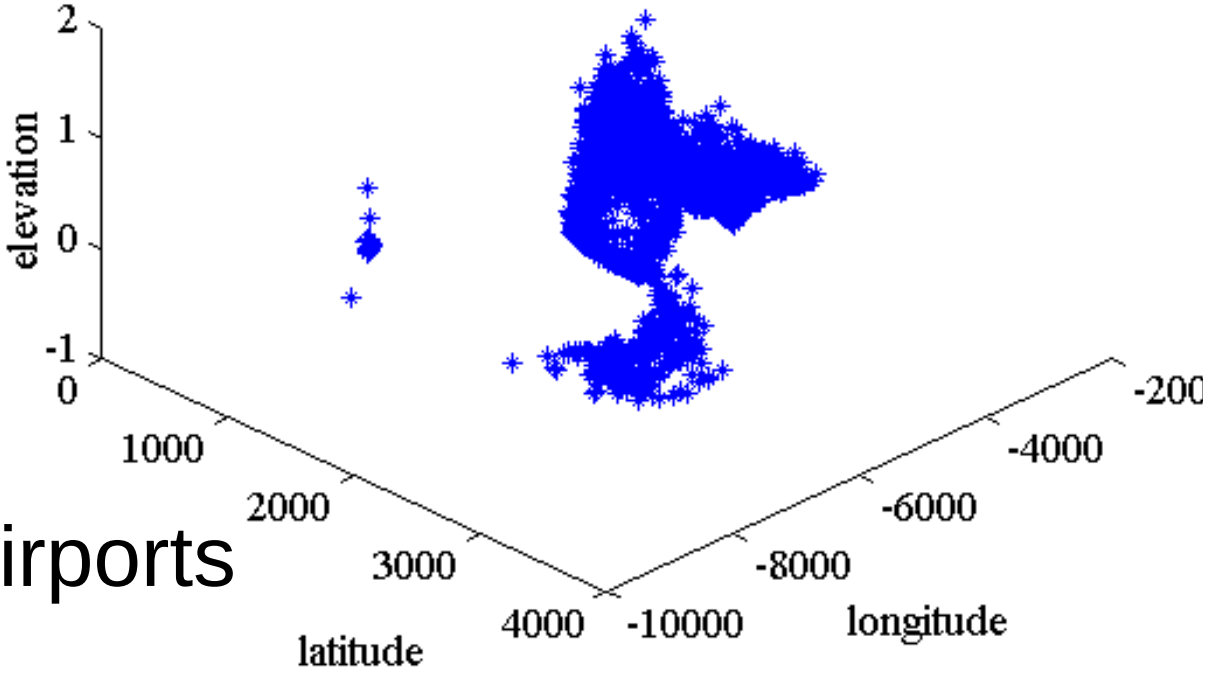
PC2 = 2nd principal component
describes 2nd largest variance
goes through variable origin space
t_{j2} = score for point j

Who's there?

- Data → random mess?
- On rotation of the data …
- Uncovered where variance and correlations are



USA airport data



- Data from 5036 airports
- For example
  - DURANGO, CO
    - Airport information on 10 August, 2000
    - **Latitude**: 37-12-11.442N (37.2031783)
    - **Longitude**: 107-52-09.103W (-107.8691953)
    - **Elevation**: 6684 ft. / 2037.3 m

http://www.airnav.com/airport/00C

PCA loadings

- *PC scores*
  - $t_1 = -0.2311x_1 + 0.9729x_2 - 0.0001x_3$
  - $t_2 = 0.9729x_1 + 0.2311x_2 - 0.0001x_3$
  - $t_3 = 0.0001x_1 + 0.0001x_2 + 1.0000x_3$
- *% explained variance*
  - $t_1 = 91.65\%$
  - $t_2 = 8.35\%$
  - $t_3 = 0\%$

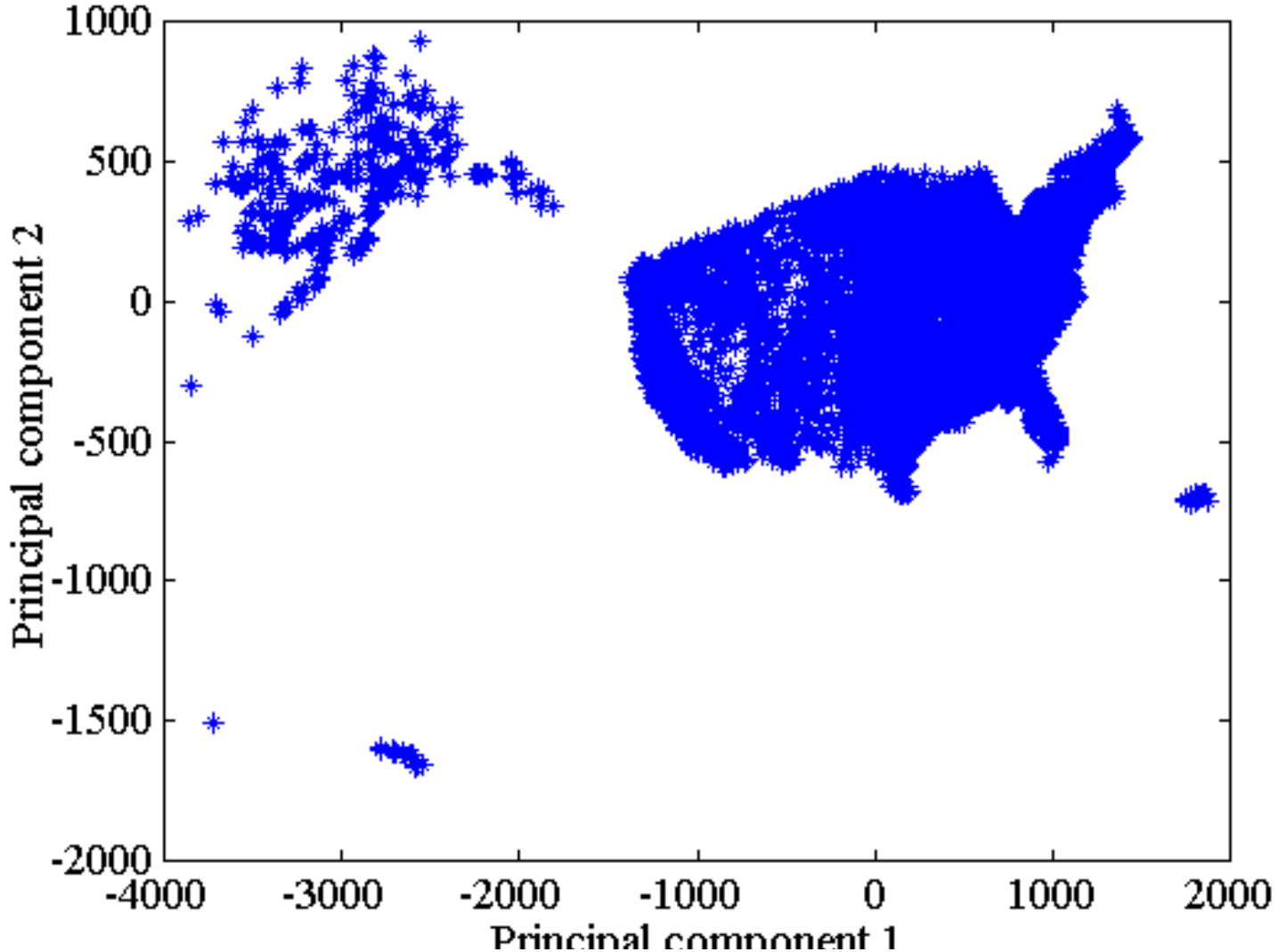
Longitude

Latitude

Elevation

∴ PCA removes 'noise'

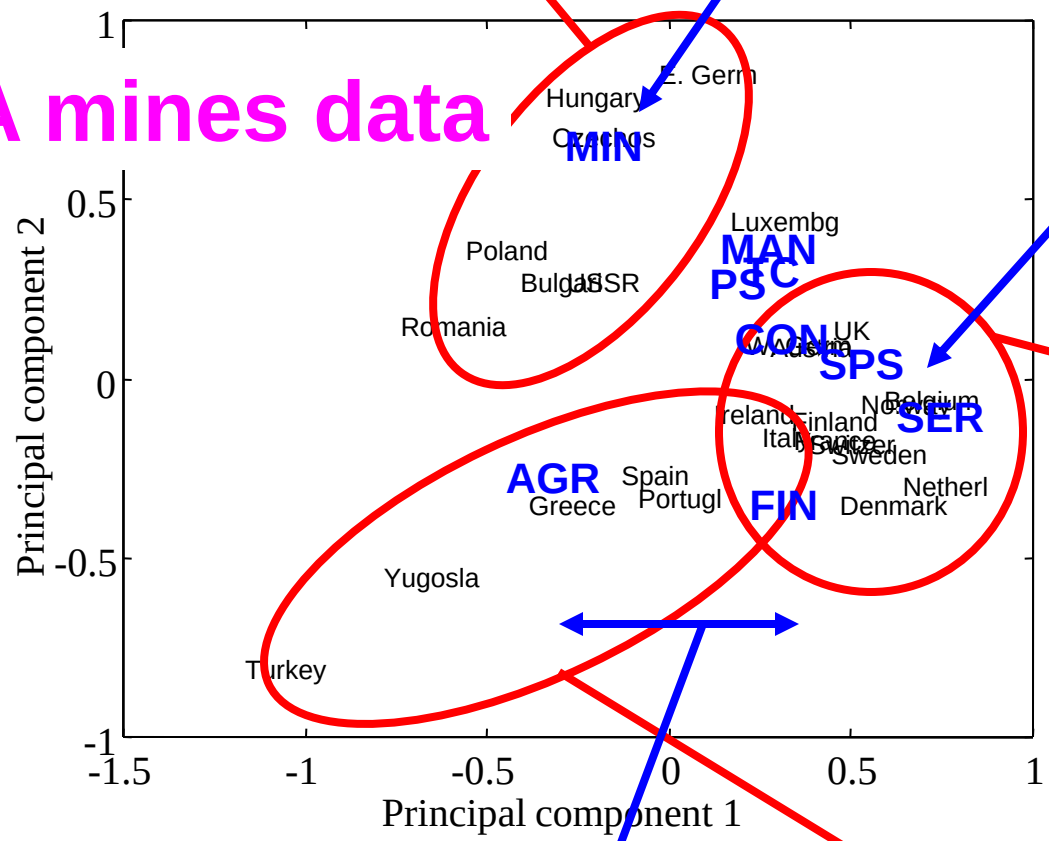
PCA scores plot



European employment data, 1979

- 26 European countries
- 9 variables for each
  - agriculture, mining, manufacturing, power supplies, construction, service industries, finance, social and personal services, transport and communications
- PC1 and PC2
  - account for 67.3% of total variance

PCA



Communist

Mining

Finance/
Service Ind.

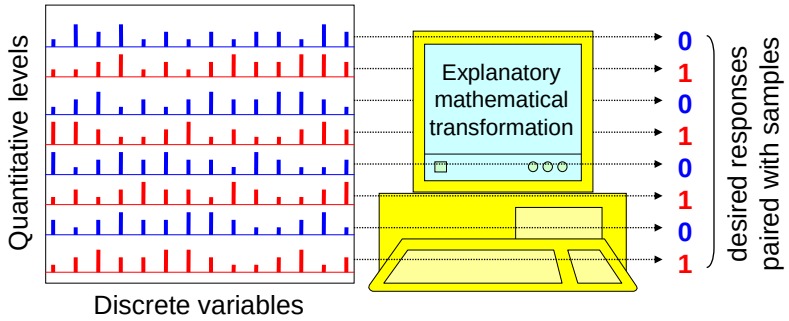
∴ PCA mines data

Strong
economy

Mediterranean

1^y vs 2^y & 3^y
industries

Predictive analyses



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

Supervised learning methods

- Use X and Y data
  - X data are mRNA/protein/metabolite levels as Inputs
  - Y data are targets as Outputs
- Analysis can be:
  - Univariate
  - Multivariate
  - Must be validated

Univariate testing methods

	Compare means	Compare median	Multivariate extension
One factor, 1 or 2 groups	Student's t-test and its variants	Wilcoxon rank sum test and its variants	Hotelling's t^2 test and its variants
One factor, multiple groups	One-Way ANOVA	Kruskal-Wallis ANOVA	mANOVA
Two factors, multiple groups	Two-Way ANOVA	Friedman test	N/A
Multiple factors, multiple groups	N-Way ANOVA	N/A	N/A

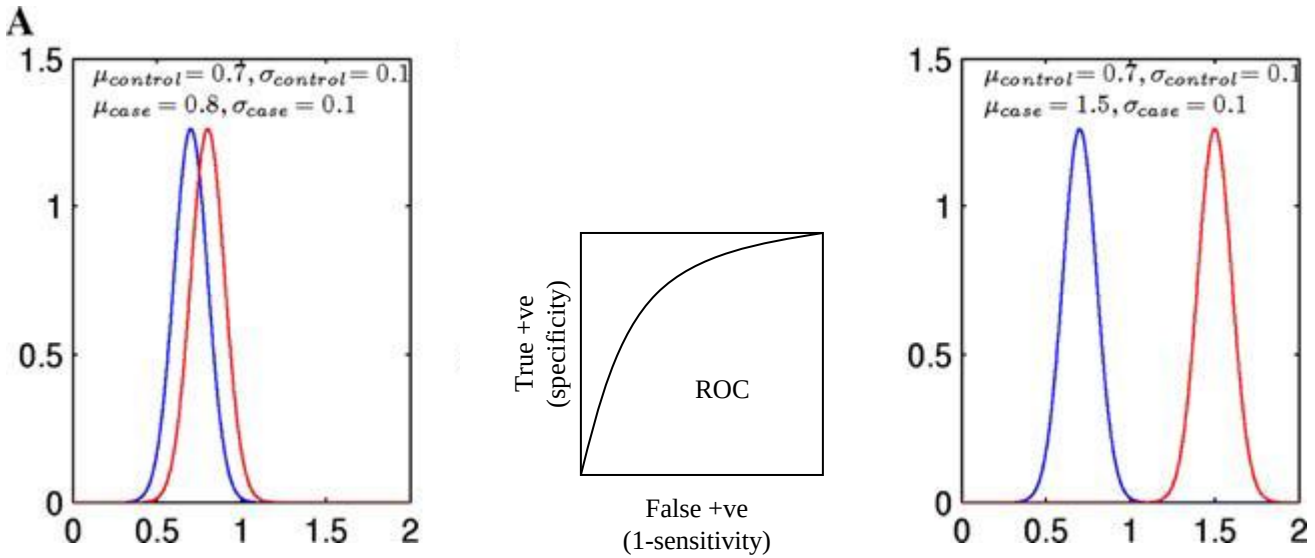
Tests comparing means known as parametric test, more powerful but lest adaptive.
Tests comparing medians known as non-parametric, less powerful but more adaptive.

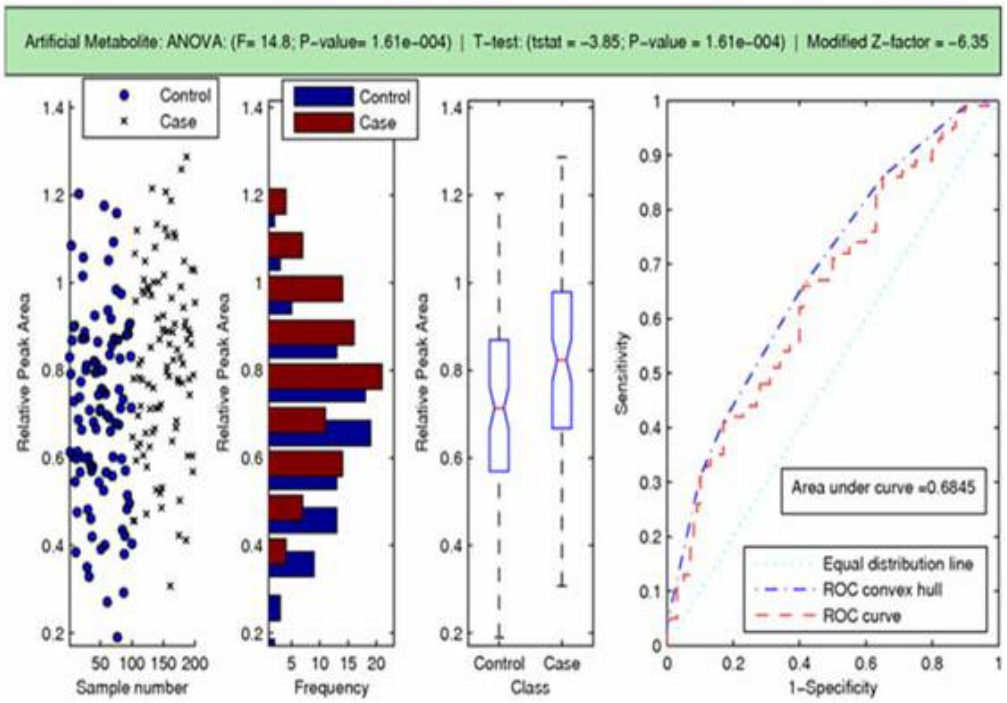
Thanks to Dr Yun Xu

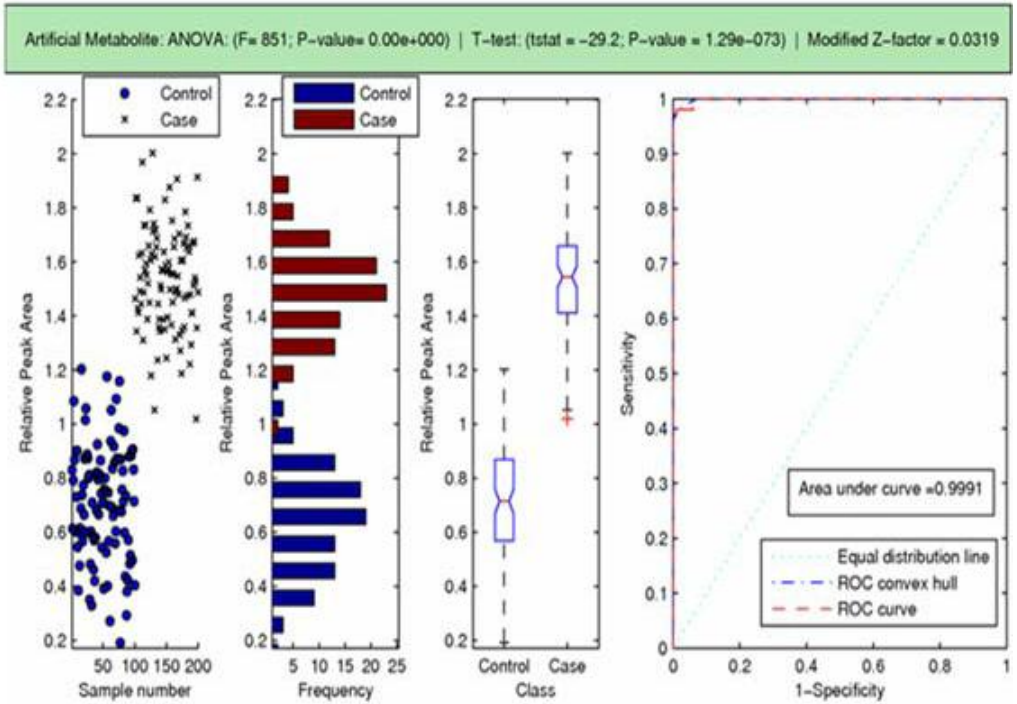
Extended to multivariate data

- For each measurement perform a test.
- Null hypothesis is that the mean metabolite level for the diseased cohort is the same as the mean for the healthy control group

	Null hypothesis (H_0) is true	**Null hypothesis (H_0) is false**
Reject null hypothesis	False positive [Type I error]	True positive [Correct outcome]
Fail to reject null hypothesis	True negative [Correct outcome]	False negative [Type II error]







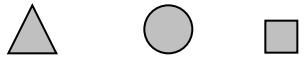
Broadhurst, D. & Kell, D.B. (2006) *Metabolomics* **2**, 171-196

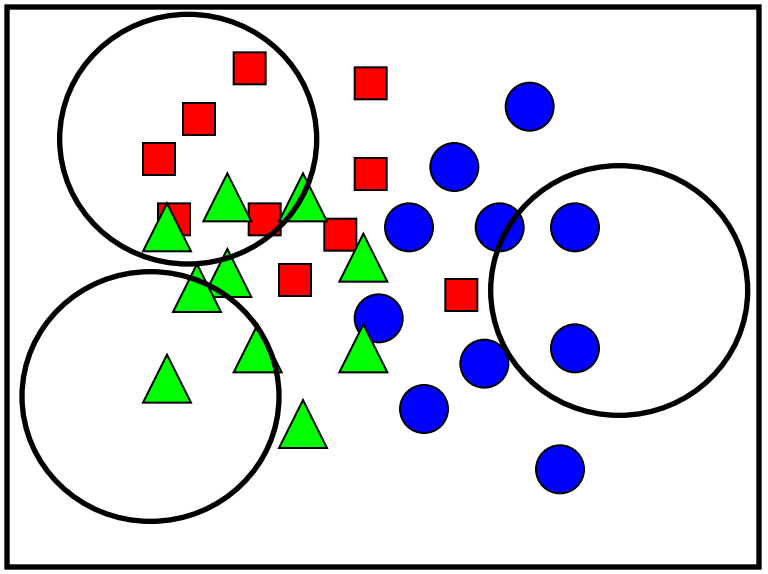
Multivariate analysis methods

- Most common methods
  - (Fisher or PLS) discriminant analysis
  - Partial least squares (PLS) regression
- Outputs
  - Scores plots
  - Target outputs or ‘labels’ → Identification
- Projection based
  - Just like PCA but this time with respect to some label (from the Y data)

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

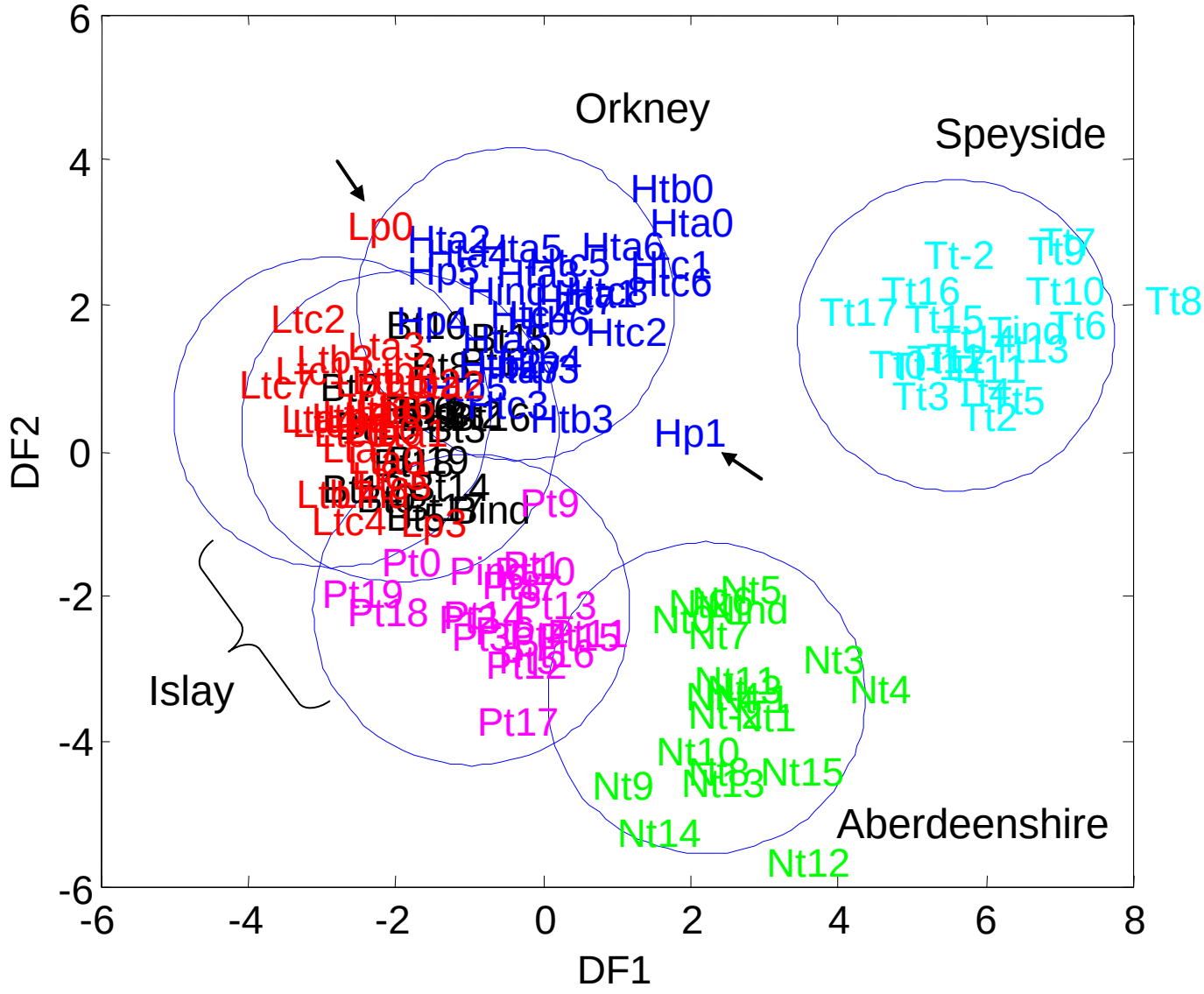




- Statistical significance: χ^2 confidence limits

Peat

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



Harrison, B. *et al*.. (2006) *Journal of the Institute of Brewing* **112**, 333-339.

Target Output for PLS

- Usually binary encoded:

Known diseases

New sample

	A	B	C	identity
X	1	0	0	$\rightarrow$ A
Y	0	1	0	$\rightarrow$ B
Z	0	0.2	0.8	$\rightarrow$ C

Easy look up table

Target Output for PLS

- Can be quantitative:

Level of disease; e.g., Gleason Grades

Patient	Grade	Diagnosis
A	0	→ No prostate cancer
B	1	→ Well differentiated cells ∴ less aggressive
C	3	→ Moderately differentiated cells
D	5	→ Undifferentiated cells ∴ fast growing + aggressive

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…

→ Use validation

Validation uses data resampling

- Resampling methods
  - Training set and a Monitoring set
  - Select subsets from the training data, while keeping the training pairs together
- External validation
  - Do the experiment again with a different cohort

Goodacre R. *et al*. (2007) *Metabolomics* **3**, 231-241

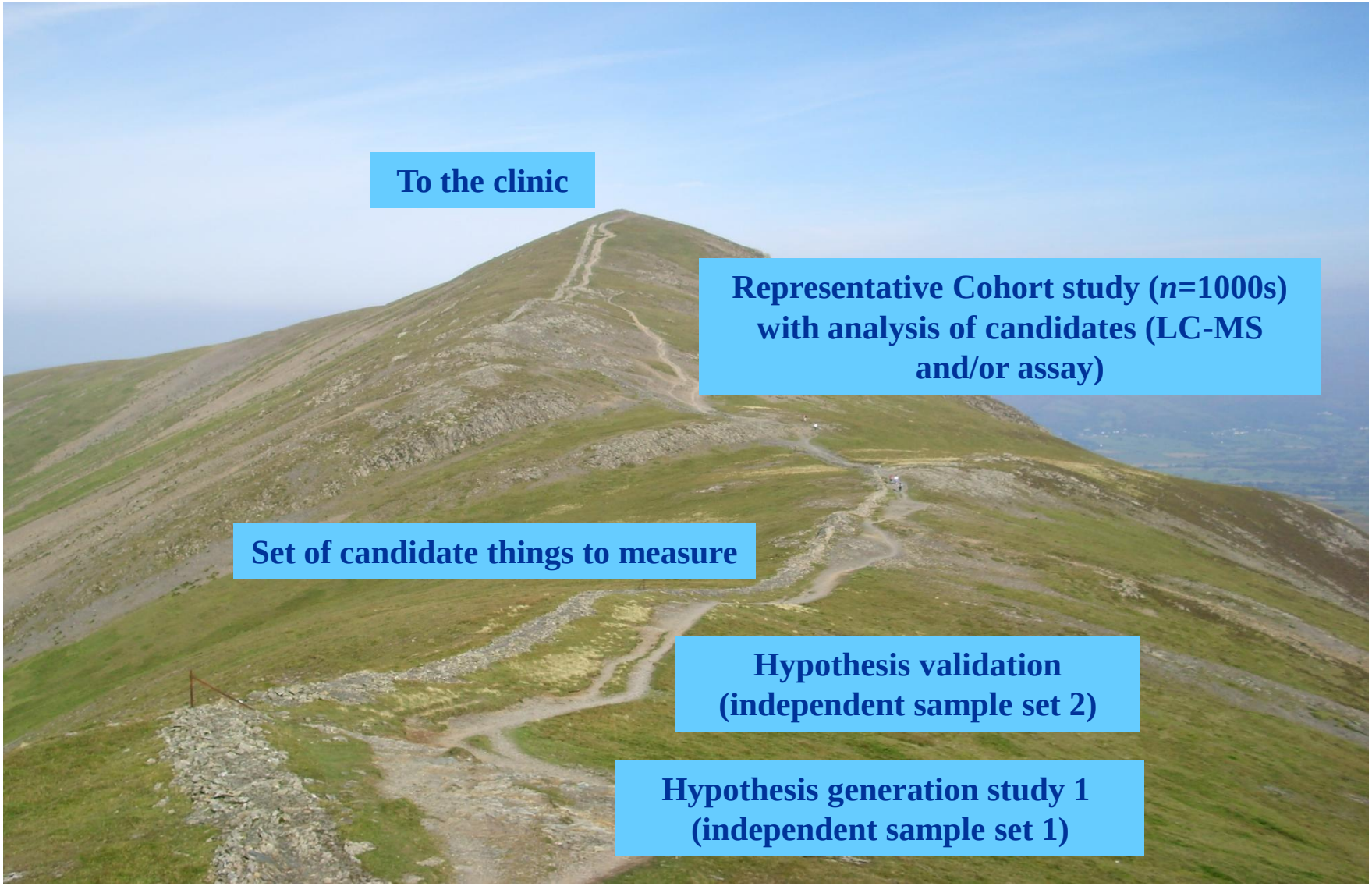
Resampling approaches

- Leave-one-out validation (LOO)
  - Single training pair (X-data and Y-data) left out and rest used for training
  - Repeat until all samples have been left out once
- K-fold validation
  - Split data into slices:
    - one slice monitors the model
    - remainder used as the training set

Better still

- Bootstrapping
  - ‘On average’ 36.8% samples were used for testing and 63.2% samples for training
  - Do many times (say 1000)
  - Do statistics on test data only
- Permutation testing
  - Null distribution using lots of permutation tests
  - Use same data but answer (Y-data) is permuted

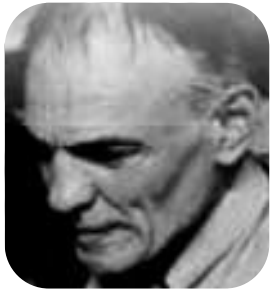
Biomarker Discovery: From lab to bedside



Increasing sample numbers and throughput

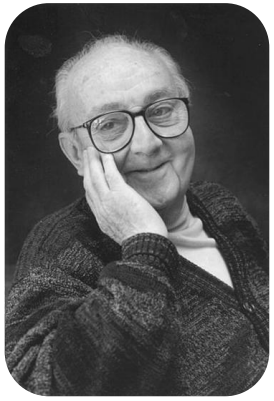
Conception (objectives, collaborations, design of experiment)

Data rich environment needs statistical analyses



"Statistics are like bikinis.
What they reveal is suggestive,
but what they conceal is vital"

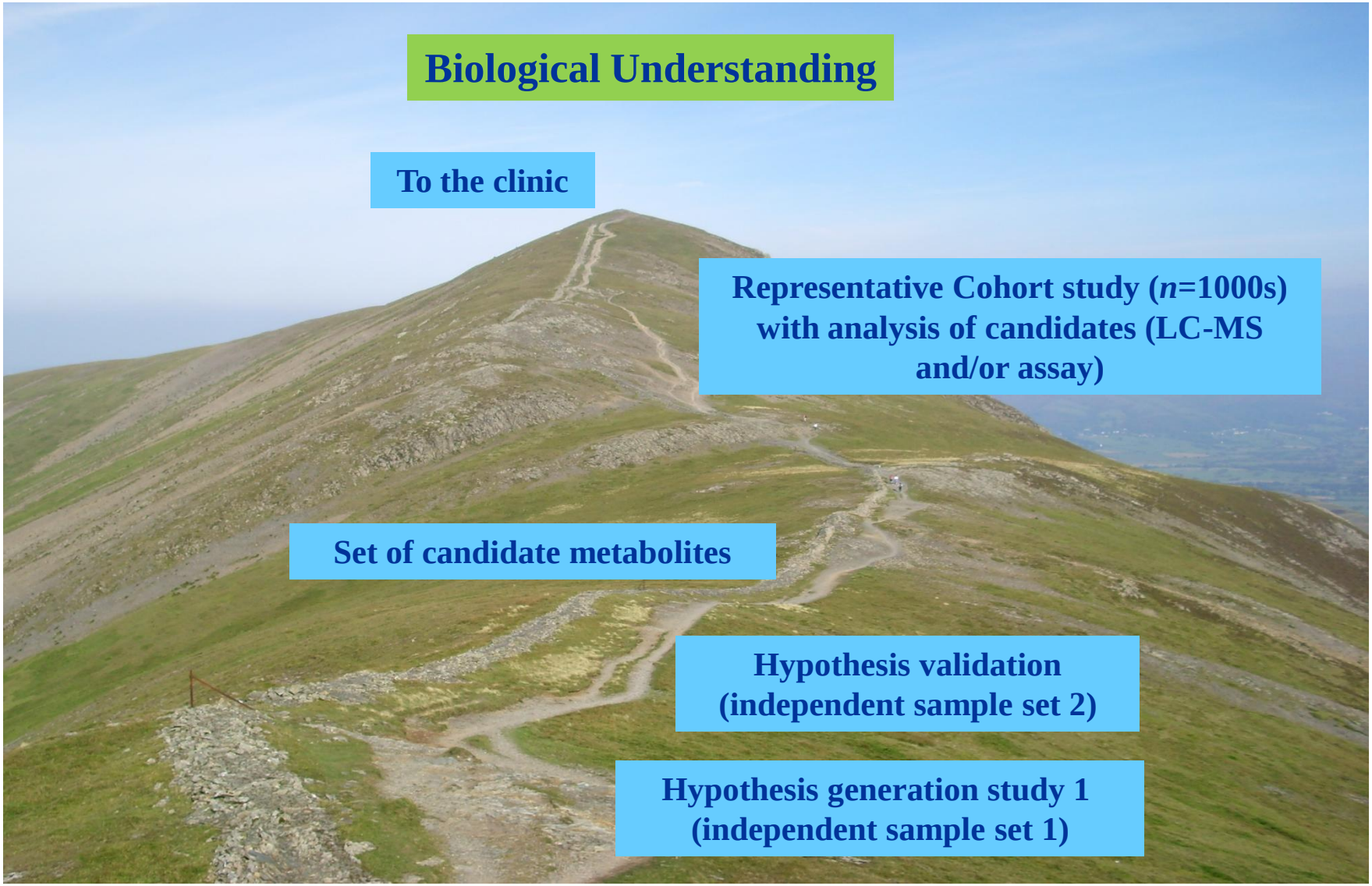
Aaron Levenstein



"All models are wrong,
but some are useful"

George E. P. Box

Biomarker Discovery: From lab to bedside



Increasing sample numbers and throughput

Conception (objectives, collaborations, design of experiment)