

# ***Research in Statistics***

-

## ***What's the future?***

**David J. Hand  
Imperial College, London**

February 2009

***‘Prediction is very difficult, especially of the future’***

Neils Bohr

Compromise between predictions of such generality  
that they will almost certainly be right  
but utterly useless

*(‘the sun will shine tomorrow, or it will rain or snow’)*

and predictions of such precision  
that they would be very useful  
but may well be wrong

*(‘it will rain at 11.40 tomorrow’)*

All scientific forecasting is based on

- noting past experience
- noting similarities between past and present
- making assumptions of continuity
  - between similar incidents

## Structure of this talk:

- examine recent past of statistical research
- identify primary drivers
- and areas which are attracting increasing attention
- project from these to the future

# ***Applications, applications, applications***

## ***The past***

- very earliest roots (e.g. Gauss, least squares in *astronomy*)
  - experimental design, anova, in *agriculture*
  - correlation and regression in *human sciences*
  - latent factor models in *psychology*
  - change point detection in *quality control*
  - survival analysis in *medicine*
  - repeated measures in *medicine*
  - signal processing in *engineering*

***then widely applied in other areas***

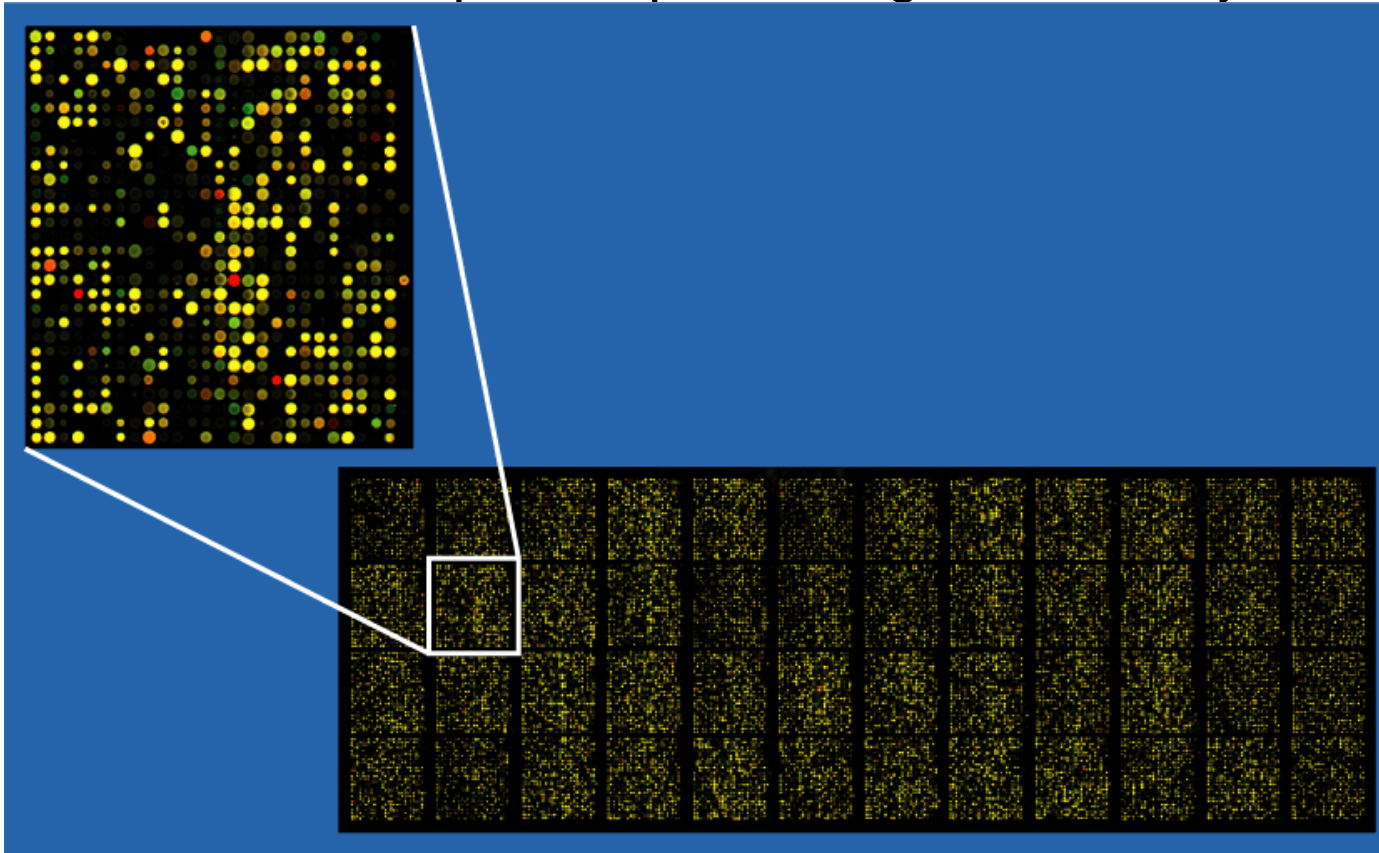
# ***The present***

Hot current application areas include

- molecular biology
- genomics
  
- physics

# Example: small-n-large-p problems

A c.40,000 probe spotted oligo microarray



(Source: Wikipedia: Microarray)

Classical statistical methods dramatically overfit

Shrinkage and penalisation needed

## Example: multiple testing

False discovery rate

Traditional methods control the probability of false positive (e.g. 5%)

FDR controls expected proportion of false positives *amongst those predicted as positive*

- Originally developed in medicine
- Heavily used in bioinformatics
- Spreading into other areas
  - evidence in support of big bang

A characteristic of bioinformatics and physics  
***large data sets***

Bioinformatics

- 40,000 spot microarray
- spectral data  
(e.g. mass spectrometry, 32,000 data bins)

Physics:

- astronomical databases,  $10^8$  objects
- particle experiments, e.g. single SLAC experiment:  
 $5 \times 10^{15}$  bytes

→ data mining

also driven by large data sets in

- retail business operations
- customer value management
- government

Leading to new problems  
Requiring new solutions

## **Housekeeping issues**

- too large to store in computer memory
- keeps on coming - streaming data

## **Inferential issues**

- tiny effects are significant

## **Data quality issues**

- especially anomaly detection
- contamination
- missing values
- automated analysis must cope with poor quality

## Some new areas:

- environmental science and climate change
  - detecting
  - monitoring
  - forecasting
- medicine
- imaging: new modalities

# ***Computer technology***

## ***Enabling new tools***

- new data capture technologies
  - sidestep slow, error-prone manual entry
- massive data storage capacities
  - UK NHS health records
  - telemetry from satellites
- extensive application of older methods
  - multivariate statistics
- development of entirely new kinds of methods
- impact on the level of statistical debate

## Examples of new methodology

- simulation
- Bayesian methods
- multiple models: bootstrapping, random forests, etc
- multilevel models
- model search: tree methods
- anomaly detection

# Official statistics

## Drivers

- automated data capture
- large data sets

## File merging, data fusion, and data integration

### Merging *public*

- electoral rolls
- censuses
- surveys by national statistical offices

### And *private*

- purchasing patterns
- banking transaction patterns
- medical records
- cellphone records
- websurfing traces

# Opportunities

**Example:** Australian study linking records of long haul flights to records of deep vein thromboembolism:

*one long haul flight a year increases your risk of DVT by 12%*

**Example:** *'HBOS, Britain's biggest mortgage lender, is pressing the Government to force local authorities to provide banks with details of council tax arrears' in a drive to improve credit scoring.*

*London Times, 5th August 2005*

# And challenges

- data security
- confidentiality
- privacy

***'We stand at the brink of an information crisis. Never before has so much information about so many people been collected in so many different places. Never before has so much information been made so easily available to so many institutions in so many different ways and for so many different purposes.'***

Simson Garfunkel *Database Nation*

Solutions need a mix of ethical and technical considerations

The linking of data files is an unstoppable process:

Imagine the situation a few years from now:

- *Use supermarket food purchase details to deduce eating habits*
- *Link to home address and name via credit card*
- *Link to medical records via name and address*
- *Link to insurance company records*

- ⇒ ***epidemiologist's dream***: huge natural experiment for determining what eating habits are risky
  
- ⇒ ***corporate ideal***: understanding of customer behaviour, and ability to maximise profit
  
- ⇒ ***individual's nightmare?***, as insurance companies withdraw cover because of what the customer bought in the supermarket

*‘There was of course no way of knowing whether you were being watched at any given moment.... It was even conceivable that they watched everybody all the time.’*

George Orwell, 1984

## ***Example: Dr Harold Shipman***

- World's most prolific serial killer?
  - 236 murders between 1978 and 1998
  - overdoses of painkillers to elderly female patients
- detected when a healthy 81 year old died suddenly
- her lawyer daughter noticed her mother had recently changed her will in favour of Shipman

But it could have been detected earlier using on-line statistical surveillance tools?



## ***Example: Detecting terrorists***

One current approach: watch for money laundering

e.g. Since 9/11 the UK has frozen £80m of terrorist assets

### ***Waste of effort?***

- Madrid bombings cost less than \$1000
- London bombings cost less than £100
- in 1997, the IRA threatened to stop the Grand National horse race: estimated loss in millions: cost about 2p
- the first bomb attack on the WTC cost about \$25,000
- 9/11 attack cost nothing beyond pilot training and air tickets

## ***But, while planning their attacks***

- terrorists must live somewhere
- they may need training (e.g. to pilot an aircraft)
- they need cell phones, cars, hotels, food, transport

***Often obtain the money through petty fraud***

## **Example:**

Nydia Velázquez won the US New York Democratic Party nomination to the US House of Representatives in 1992

She was the first Puerto Rican woman elected to the House of Representatives

But, in 1991, during a bout of depression, she attempted suicide

Someone trawled through her medical records and sent details of this to the newspapers in an attempt to derail her election campaign

Is the law keeping up?

Disclosure control

Credit scoring in retail banking

# Excluded characteristics in credit scoring

Aim:

avoid discrimination on grounds unrelated to risk

Approach:

exclude gender, religion, age, marital status etc from retail risk models

- 1) If aim is to obtain ***most accurate estimates of probability of default***
  - then include all characteristics in models
  
- 2) If aim is to obtain ***most accurate estimates of probability of default which can be obtained from information unrelated to excluded characteristics***
  - include all in model, and make decisions in subspace orthogonal to excluded characteristics

## Example:

Outcome  $y$ ;

allowed characteristic  $x_1$

excluded characteristic  $x_2$

Ground truth:  $y = \alpha + \beta_1 x_1 + \beta_2 x_2$

### ***Current legislation:***

$$y = \alpha + \gamma_1 x_1$$

Treat  $x_2$  classes ***equally***:

$$y = \alpha + \beta_1 x_1$$

Treat  $x_2$  classes ***fairly*** (best estimate of  $y$ ):

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

International standards and classifications

The march of globalisation

Promoting mutual understanding between countries  
Decreasing suspicion arising from ignorance

Experiments for determining social policy

# Data quality

***No data are perfect***

Particular danger with merged files and privacy issues

***Who bears the responsibility when inaccurate data leads the police to suspect an innocent man of being a terrorist, and he is shot dead?***

## Example:

*The Lancet, 2002: discovery of patterns in proteomic spectra that could distinguish between women with ovarian cancer and those without*

Ovarian cancer is often fatal

*But* closer examination: observed difference probably due to differences in ways the samples were treated

Think of the women who were confidently told they did not have cancer, on the basis of this analysis of faulty data

# Statistics and Computing, 1993

*Special issue on the future of statistical research 'over the next ten years'*

Ole Barndorff-Neilsen, John Chambers, John Copas, Bradley Efron, John Gower, Jack Hibbert, Pierre Legendre, David Moore, John Nelder, Donald Rubin, Richard Smith

## Predictions included:

- issues of massive data sets
- need to train statisticians more broadly
- 'greater statistics'
- new kinds of data
- the impact of the computer
- international systems and standards
- data quality
- enabling researchers to use and understand methods
- simulation
- graphics
- Bayesian methods
- spatial models

# Conclusion

Statistics mainly driven by exogenous factors

- application domains
- the computer

My predictions:

- massive data sets
- anomaly detection
- data quality
- new kinds of data
- experiments and social policy

***END***