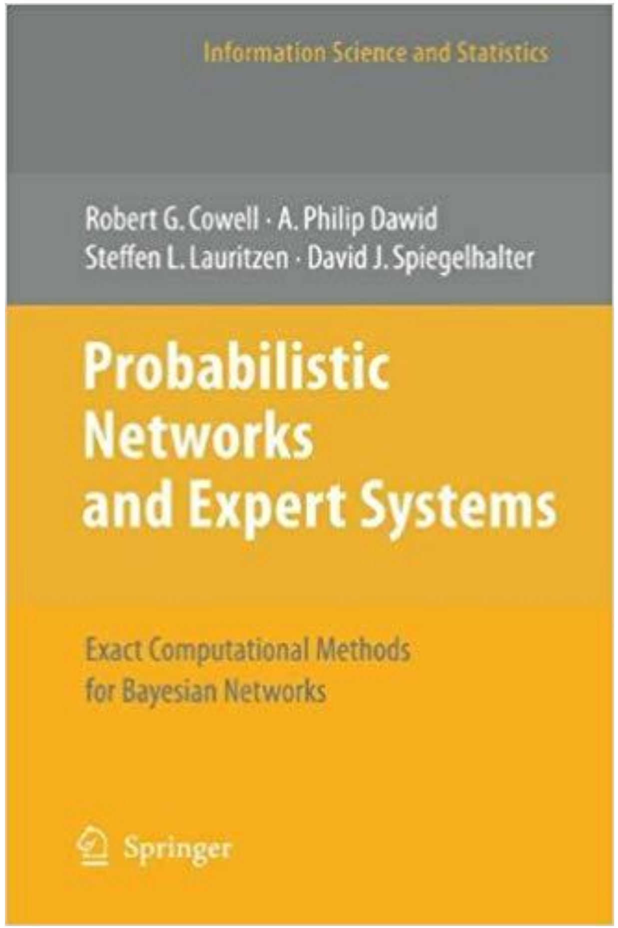
Statistics for data science: what are the essentials?

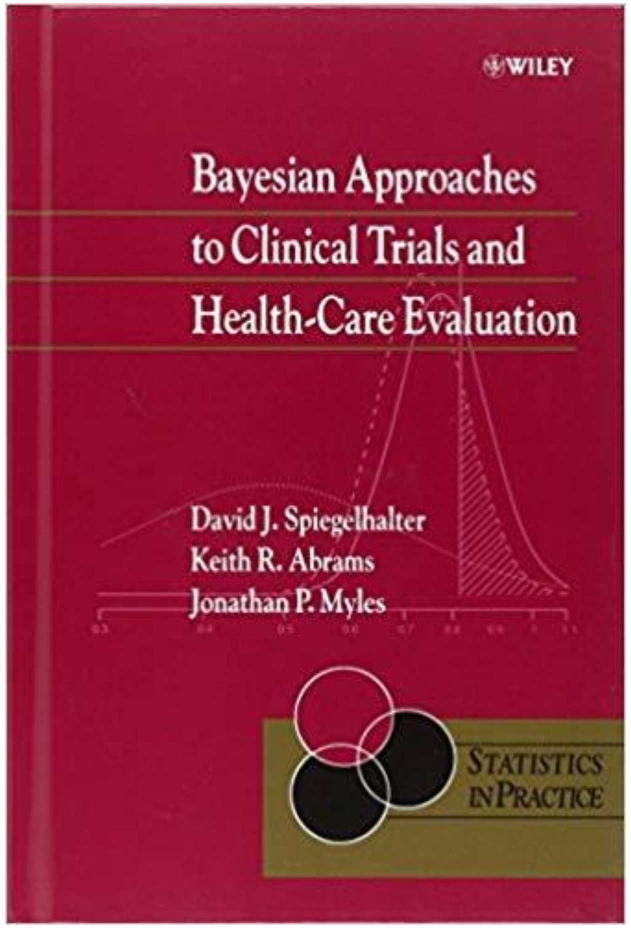
David Spiegelhalter

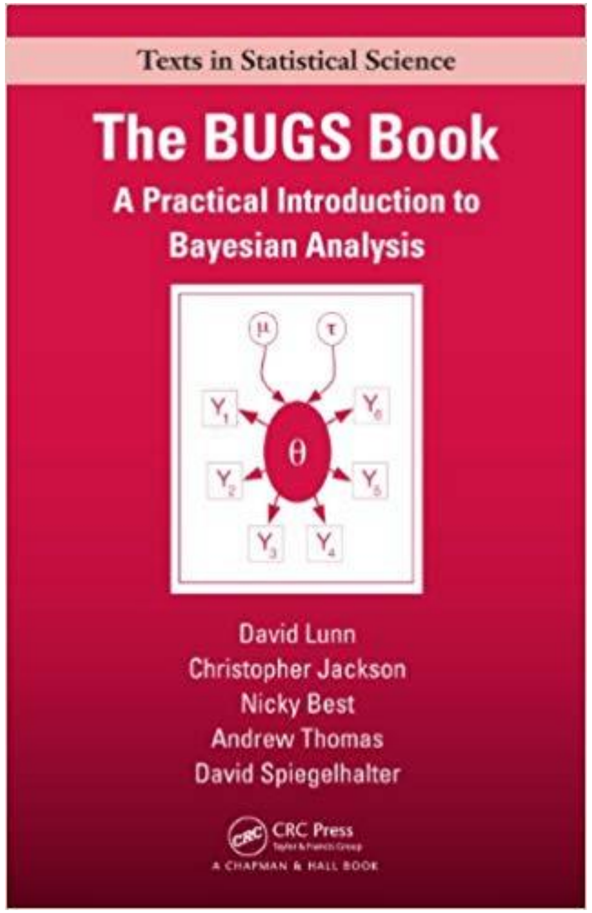
Chair of the Winton Centre for Risk & Evidence Communication, University of Cambridge

President, Royal Statistical Society (2017-2018)

ISM Consortium for Human Resource Development of Statistical Experts in Japan. February 2022

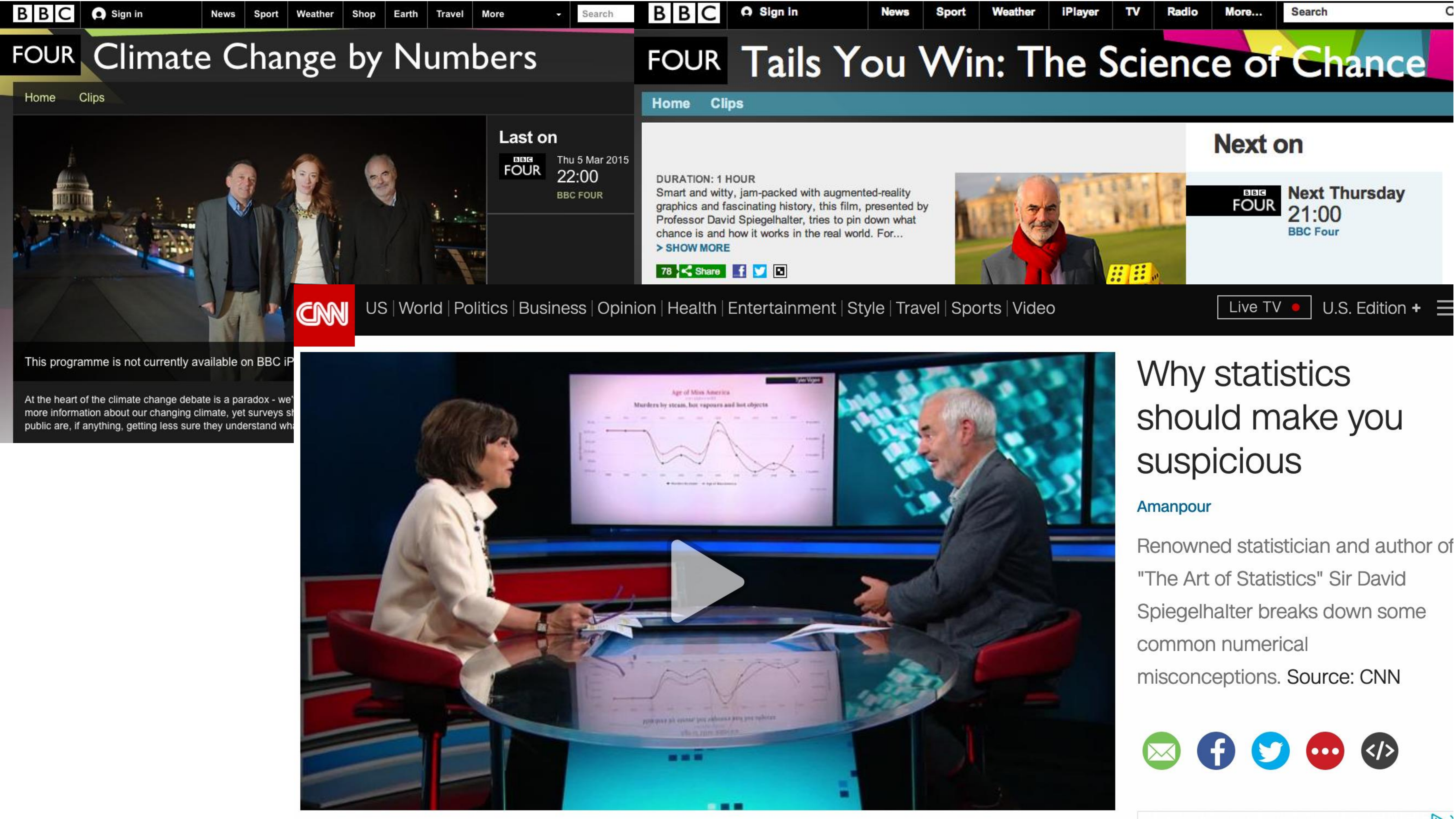




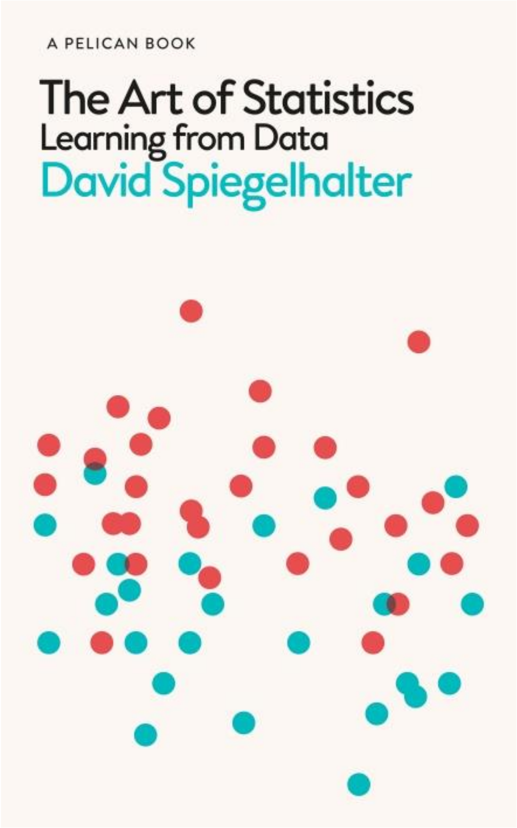


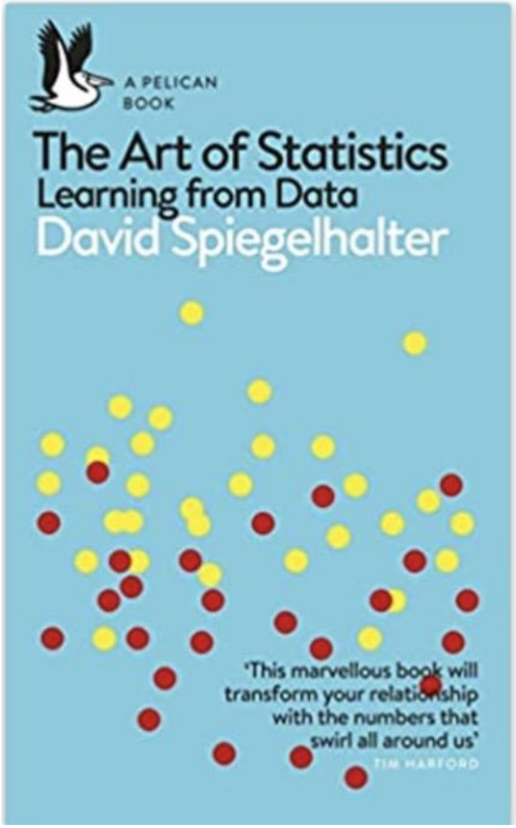
I used to do statistical methodology....

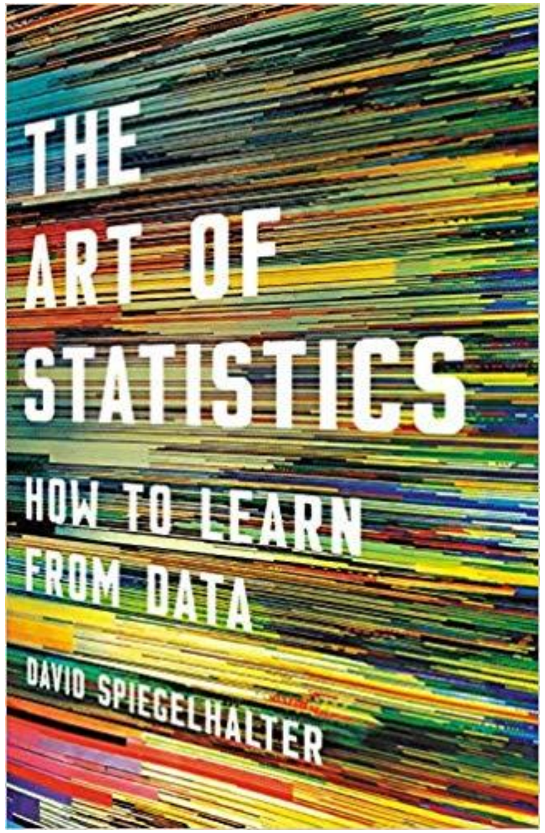
until I was philanthropically funded in 2007.....


BBC Sign in News Sport Weather Shop Earth Travel More Search
FOUR Climate Change by Numbers
Home Clips
Last on
BBC FOUR Thu 5 Mar 2015 22:00 BBC FOUR
This programme is not currently available on BBC iP
At the heart of the climate change debate is a paradox - we'
more information about our changing climate, yet surveys sh
public are, if anything, getting less sure they understand wha
BBC Sign in News Sport Weather iPlayer TV Radio More... Search
FOUR Tails You Win: The Science of Chance
Home Clips
DURATION: 1 HOUR
Smart and witty, jam-packed with augmented-reality graphics and fascinating history, this film, presented by Professor David Spiegelhalter, tries to pin down what chance is and how it works in the real world. For...
> SHOW MORE
78 Share
Next on
BBC FOUR Next Thursday 21:00 BBC Four
CNN US World Politics Business Opinion Health Entertainment Style Travel Sports Video Live TV U.S. Edition +
Why statistics should make you suspicious
Amanpour
Renowned statistician and author of "The Art of Statistics" Sir David Spiegelhalter breaks down some common numerical misconceptions. Source: CNN


XINZHI
一念之差
关于风险的故事与数字
[英] 迈克尔・布拉斯兰德
戴维・施皮格哈尔特 著 威治 译
The Norm
Chronicles
Stories and Numbers
about Danger
74
新知
文库
統計学はときにセクシーな学問である
デビッド・シュピーゲルハルター 著
David Spiegelhalter
石塚直樹 訳
SEX
BY
NUMBERS
世界的権威は
数値をどう読むか
What Statistics Can Tell Us About Sexual Behaviour
ライフサイエンス出版


A PELICAN BOOK
The Art of Statistics
Learning from Data
David Spiegelhalter


A PELICAN BOOK
The Art of Statistics
Learning from Data
David Spiegelhalter
'This marvellous book will transform your relationship with the numbers that swirl all around us'
TIM HARFORD


THE ART OF STATISTICS
HOW TO LEARN FROM DATA
DAVID SPIEGELHALTER

Interpreting data is not easy

INTRODUCTION

The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.

— Nate Silver, *The Signal and the Noise*[1]

The traditional statistics course

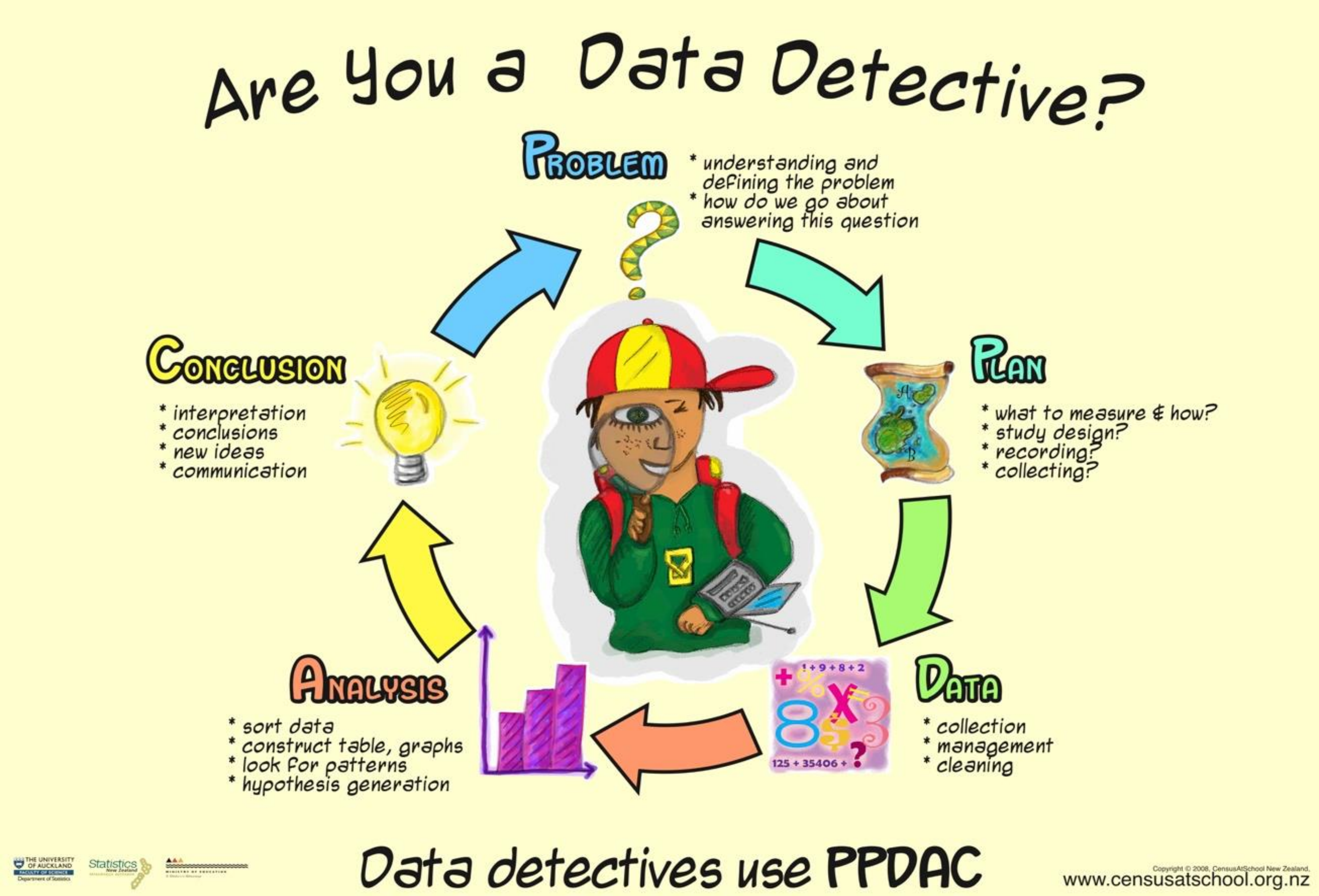
- Describing data with summary statistics
  - *dull*
- Probability theory for drawing random observation from a population distribution
  - *difficult and mathematical*
- Probability theory for distributions of summary statistics
  - *mathematical and incomprehensible*
- Formulae for statistical tests
  - *mathematical, unmotivated, just a bag of tools*
- (If lucky) Examples of using statistical models in real life.

A ‘modern’ statistical course

- Motivate by problem solving
- Start with visualisation and exploring data
- Focus on what can be reasonably learned from data, biases in data, concluding causation, etc
- Models and algorithms
- Assessing uncertainty through re-sampling data (‘bootstrap’)
- Probability theory as neat way of turning random variation into uncertainty about what is true
- Hypothesis testing and its potential problems
- Bayesian methods

All these rather abstract, challenging, ideas are there to help answer real questions

- The ’data cycle’
- eg PPDAC (promoted in New Zealand)


Are You a Data Detective?
Problem
* understanding and defining the problem
* how do we go about answering this question
Plan
* what to measure & how?
* study design?
* recording?
* collecting?
Data
* collection
* management
* cleaning
Analysis
* sort data
* construct table, graphs
* look for patterns
* hypothesis generation
Conclusion
* interpretation
* conclusions
* new ideas
* communication
1 + 9 + 8 + 2
125 + 35406 +
Data detectives use PPDAC
THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE
Department of Statistics
Statistics New Zealand
Copyright © 2008, CensusAtSchool New Zealand.
www.censusatschool.org.nz

Looking at data

What was the pattern of Harold Shipman's murders?

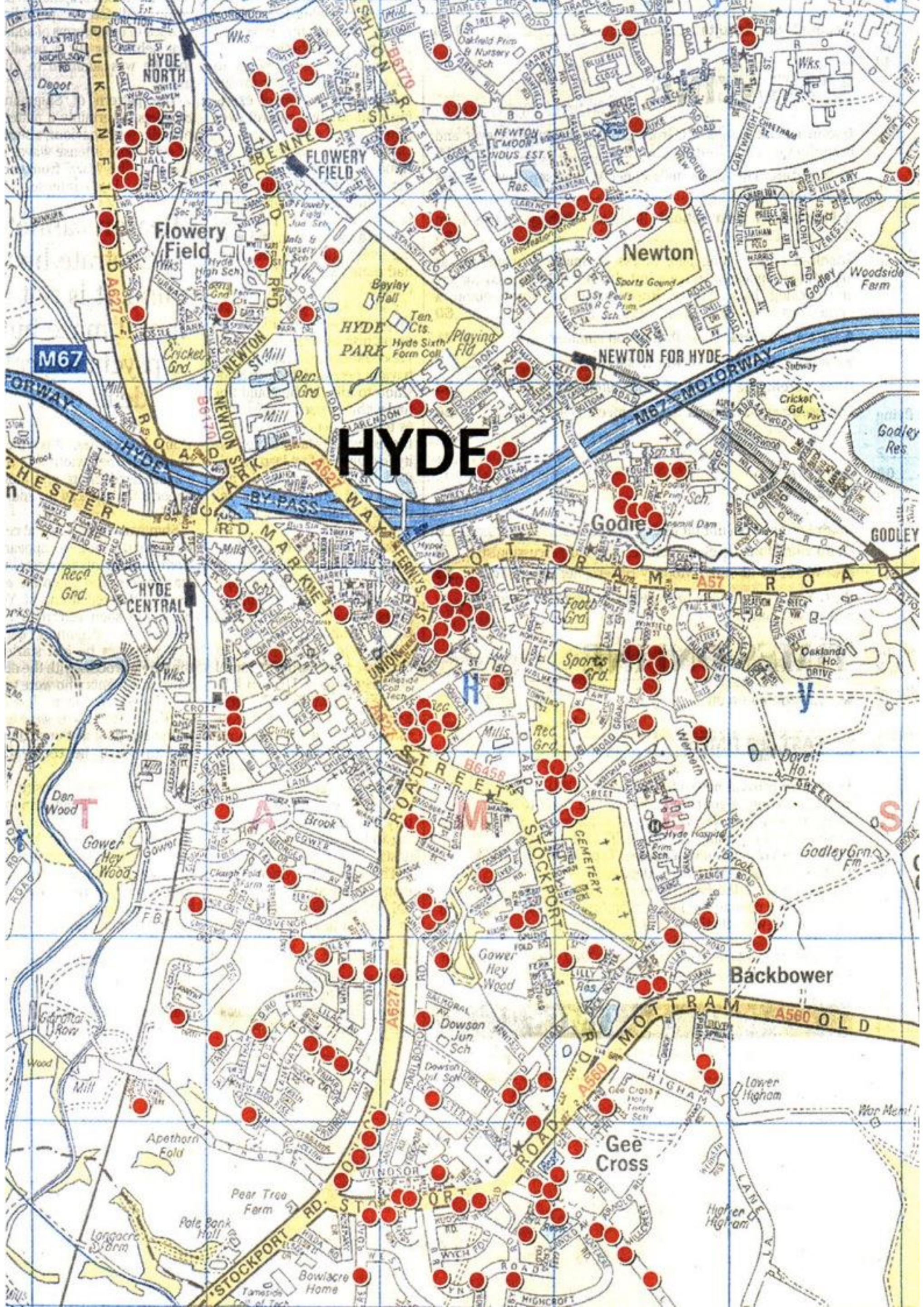
‘I have nothing to hide’

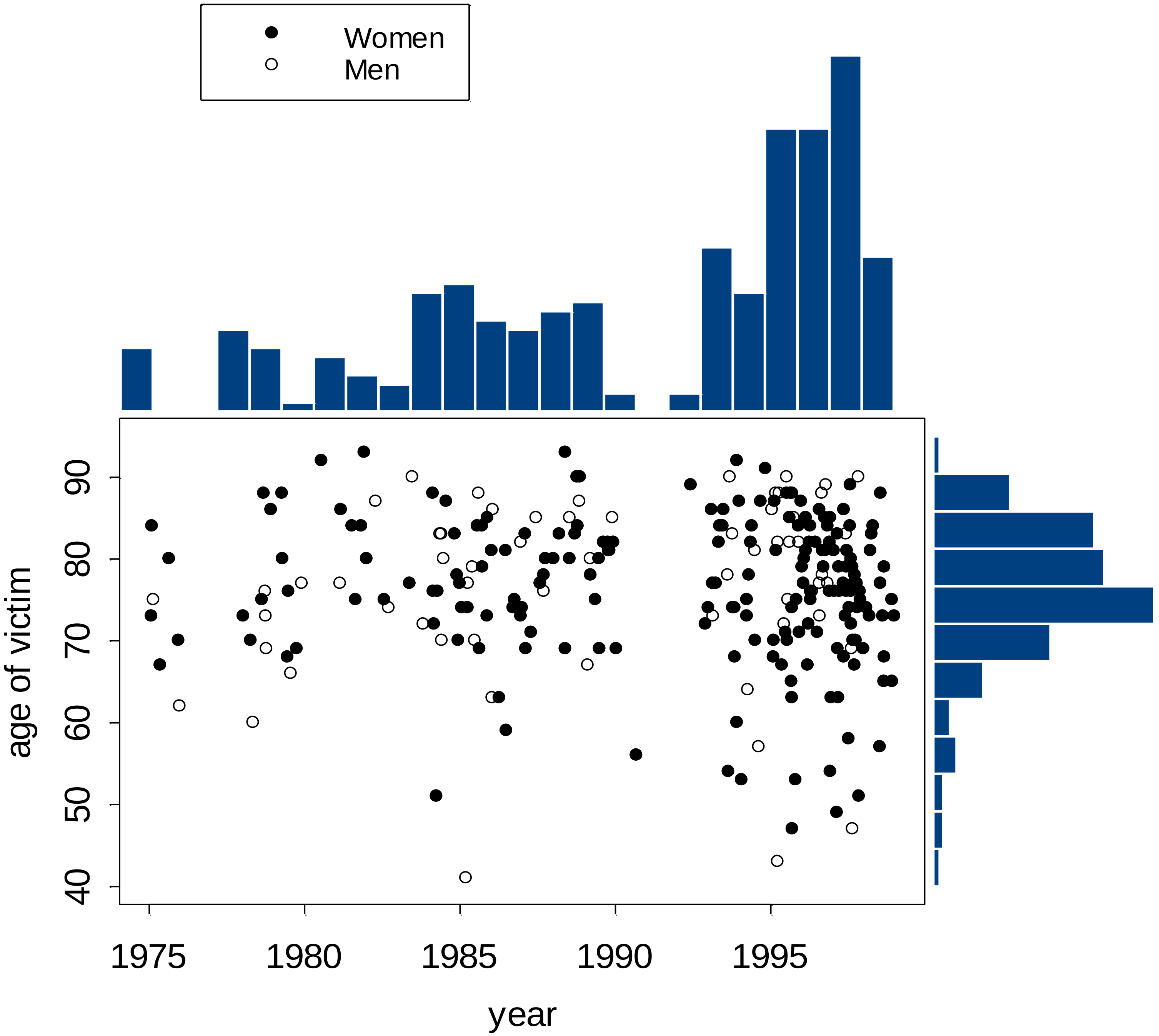
Dr Harold Shipman, general practitioner, on his arrest in September 1998

Shipman Inquiry July 2002:

215 definite victims,

45 probable




Women
Men
age of victim
40
50
60
70
80
90
1975
1980
1985
1990
1995
year

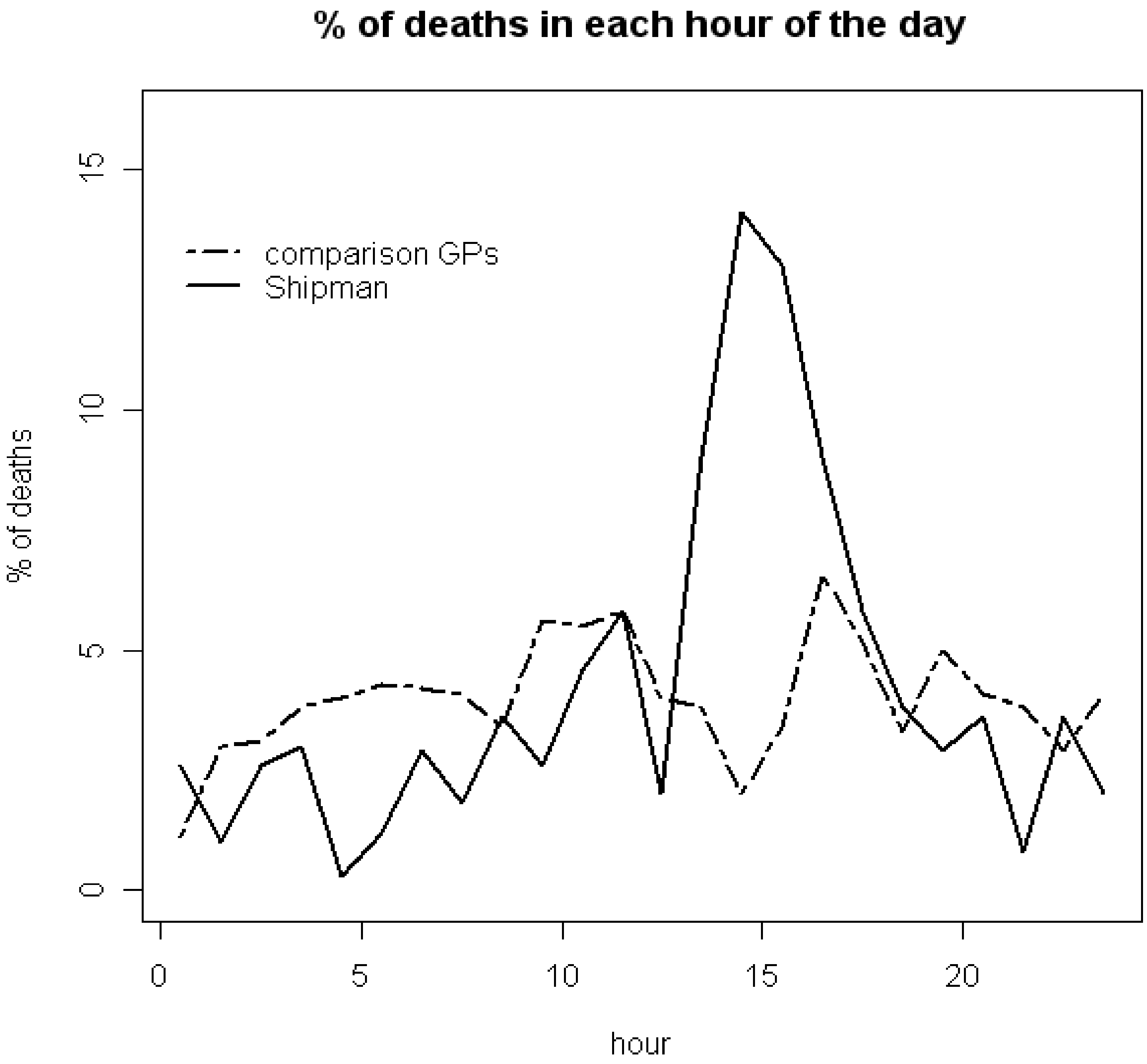
Looking at data
What was the pattern of Harold Shipman's murders?

- **Problem**: can more detail tell us more about what Shipman did?
- **Plan**: compare actual times at which his patients died with the times of deaths recorded by other local GPs
- **Data**: a huge exercise requiring examination of death certificates
- **Analysis**: simple plotting.....

People die at all hours

People die at all hours

- but not Shipman's victims

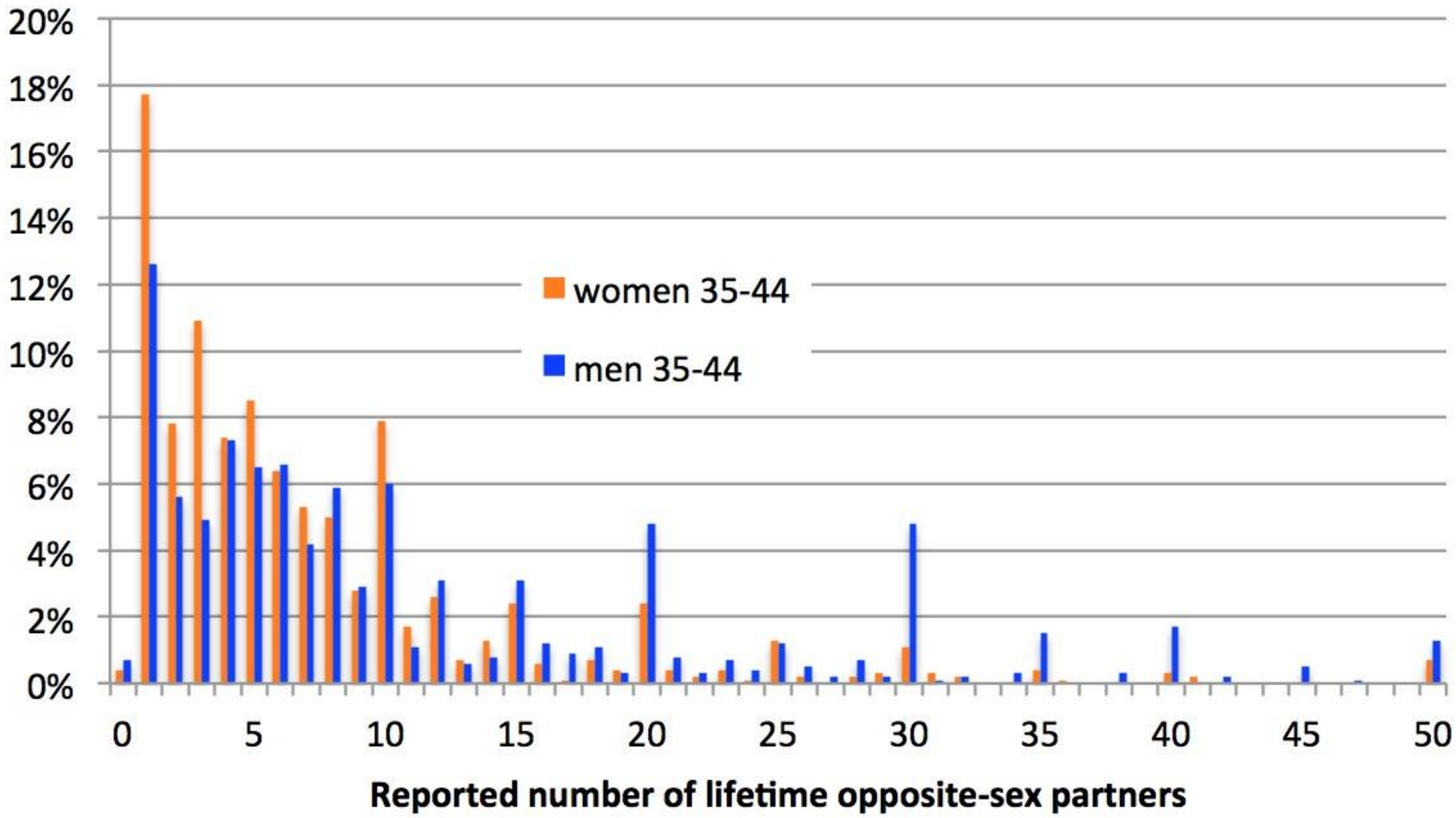


Inference and bias

How many sexual partners have people in Britain had in their lifetime?

- **Problem**: cannot know this as a fact
- **Plan**: survey in which people are carefully asked about the sexual activity (Natsal)
- **Data**: reports of numbers of partners
- **Analysis**: plotting and summary statistics

How many sexual partners do people report?



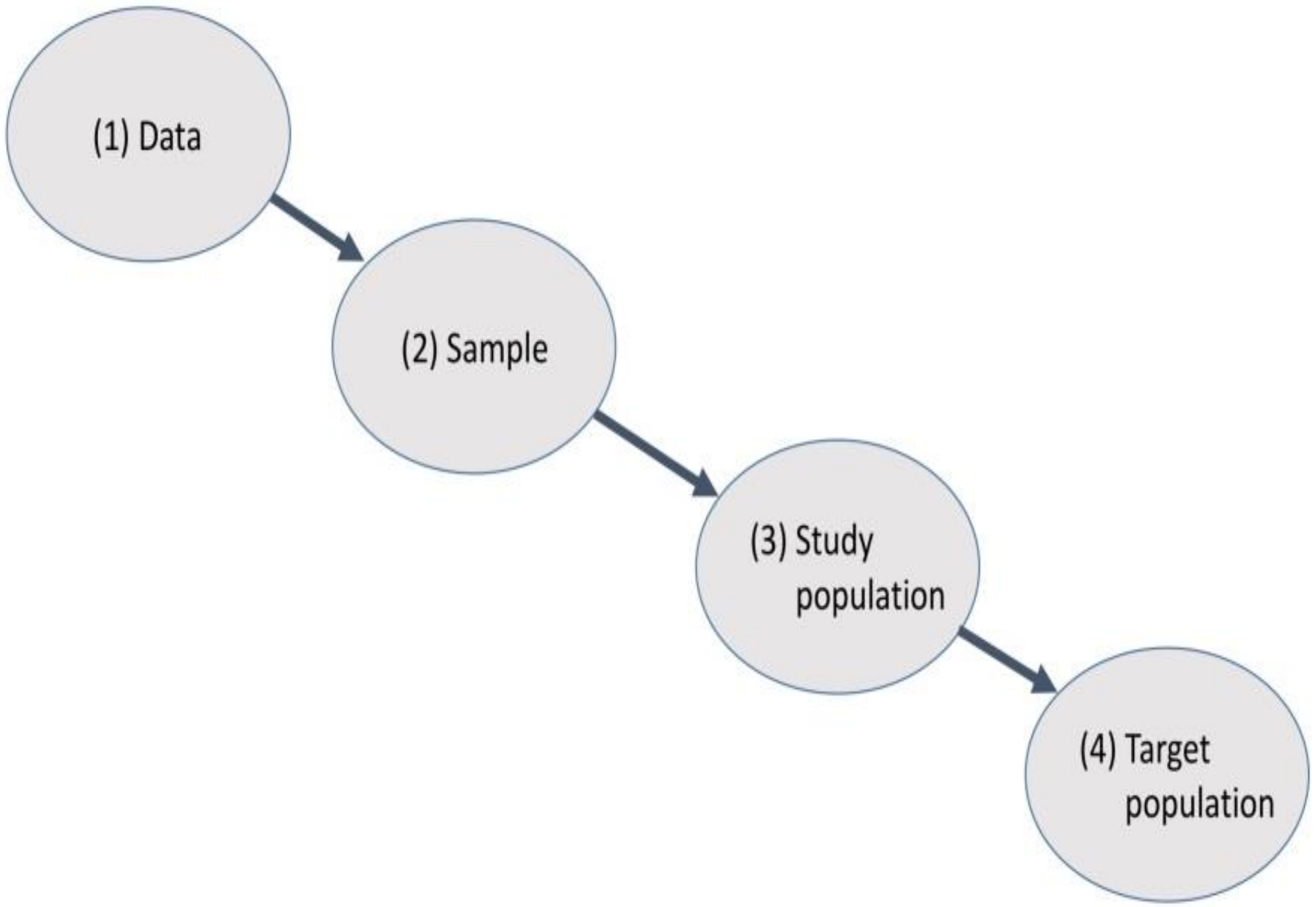
Inference and bias

How many sexual partners have people in Britain ***really*** *had in their lifetime?*

Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

- **Conclusions**: can we generalise this to the whole population?????

Induction: the stages in generalising from data



- ***1 to 2.*** How reliable are the reports?
- *Poor memory, social acceptability bias etc*

- ***2 to 3.*** How representative is the sample of those eligible for the study?
- *Random sampling of families (soup), 66% response*

- ***3 to 4.*** How close does the study population match the target population?
- *No people in institutions, etc*

Causation (or correlation)
The power of the press release....

Socioeconomic position and the risk of brain tumour: a Swedish national population-based cohort study

Amal R Khanolkar,[1,2] Rickard Ljung,[2] Mats Talbäck,[2] Hannah L Brooke,[2] Sofia Carlsson,[2] Tiit Mathiesen,[3] Maria Feychting[2]

- abstract:
  - *We observed consistent associations between higher socio-economic position and higher risk of glioma*
- press release
  - *High levels of education linked to heightened brain tumour risk*
- Daily Mirror…

Mirror

Why going to university increases risk of getting a brain tumour

23:30, 20 JUN 2016 | BY ANDREW GREGORY

Highly educated people are more likely to suffer from brain tumours than those who do not progress as far in their education

13 SHARES 1 COMMENT

Regression, prediction and algorithms

Who was the luckiest person on the Titanic?

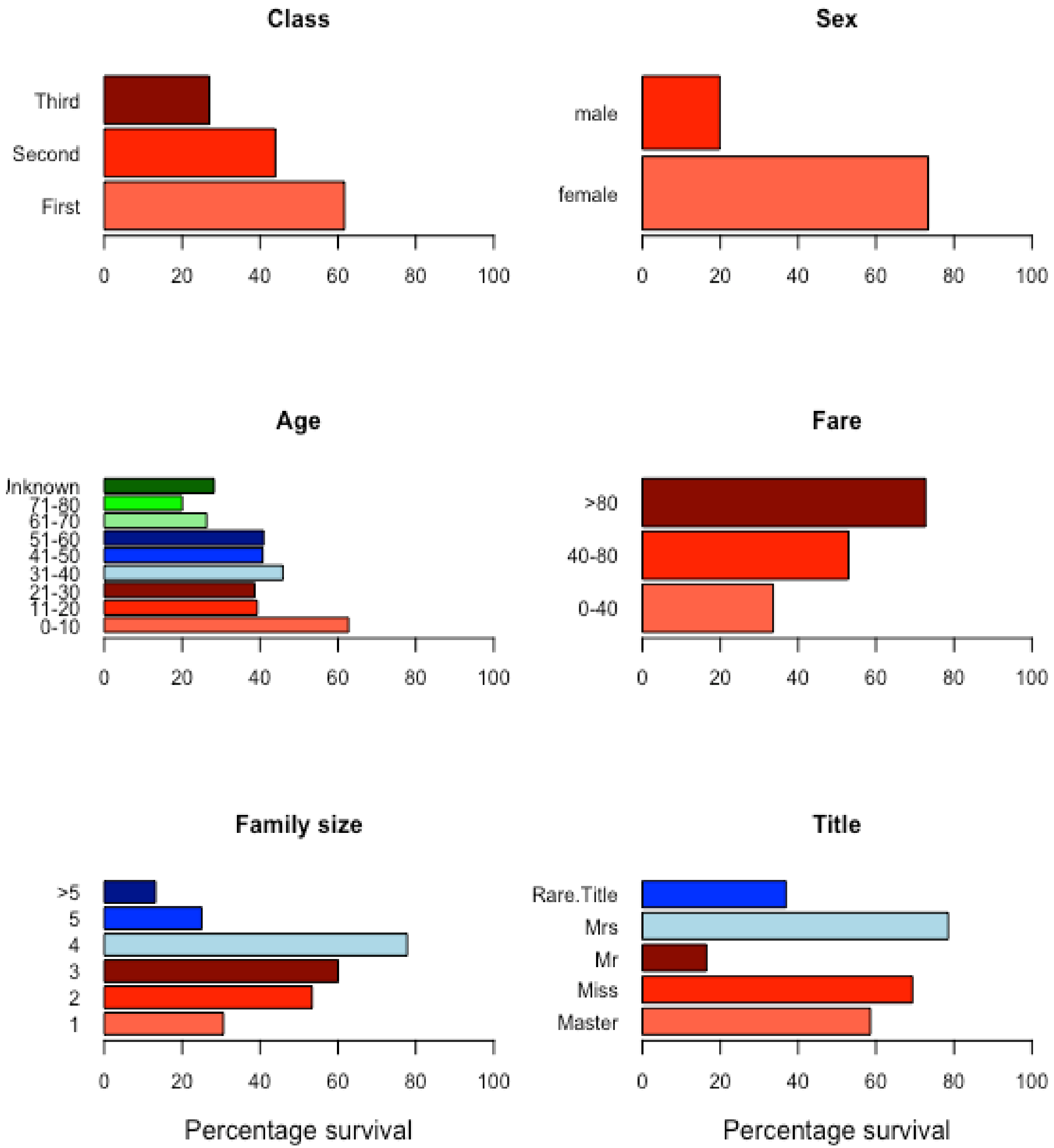
Ilfracombe, North Devon

William Somerton’s entry in a public database of 1309 passengers (39% survive)

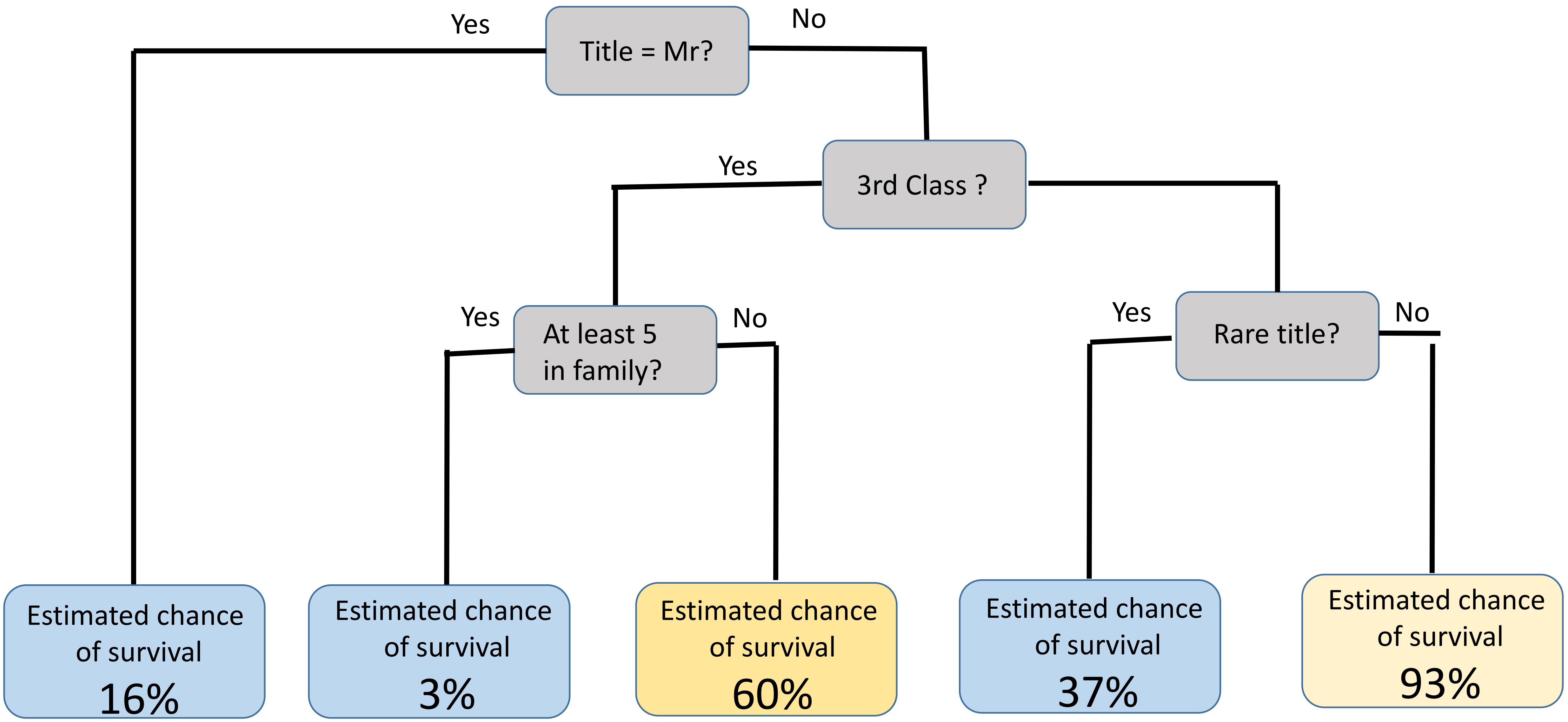
pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body
3	0	Somerton, Mr. Francis William	male	30	0	0	A.5. 18509	8.0500		S		
3	0	Spector, Mr. Woolf	male		0	0	A.5. 3236	8.0500		S		
3	0	Spinner, Mr. Henry John	male	32	0	0	STON/OQ. 369943	8.0500		S		
3	0	Staneff, Mr. Ivan	male		0	0	349208	7.8958		S		
3	0	Stankovic, Mr. Ivan	male	33	0	0	349239	8.6625		C		
3	1	Stanley, Miss. Amy Zillah Elsie	female	23	0	0	CA. 2314	7.5500		S	C	
3	0	Stanley, Mr. Edward Roland	male	21	0	0	A/4 45380	8.0500		S		

- Challenge: can we build an algorithm that will accurately predict who survives the Titanic?
- Based on factors in data-base, produce either a yes/no judgement, or a probability of survival
- Split the data-base of 1309 passengers at random into a **training set** (70%) on which to build algorithms, and a **test set** (30%) to assess how good it is.
- Currently over 59,000 entries in a similar online Kaggle competition

Unsurprising factors predict survival



A simple classification tree



How good is my algorithm?

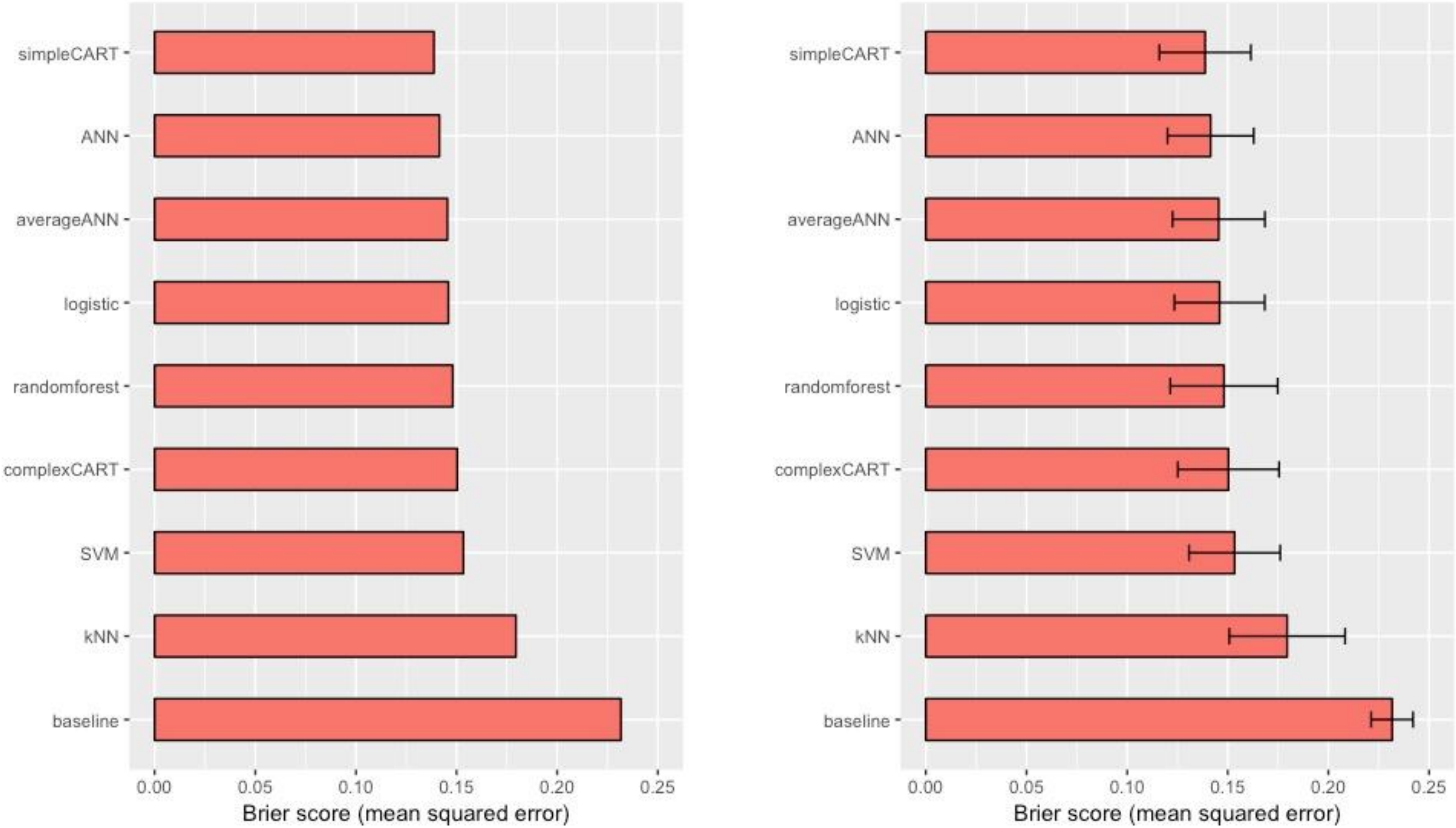
- 'Accuracy' is a very crude way of judging an algorithmic prediction

- Better to use the probabilities provided

- If probability *p* is given to an event X (0,1), then the Brier score is $(X - p)^2$

Performance of a range of methods on the test set

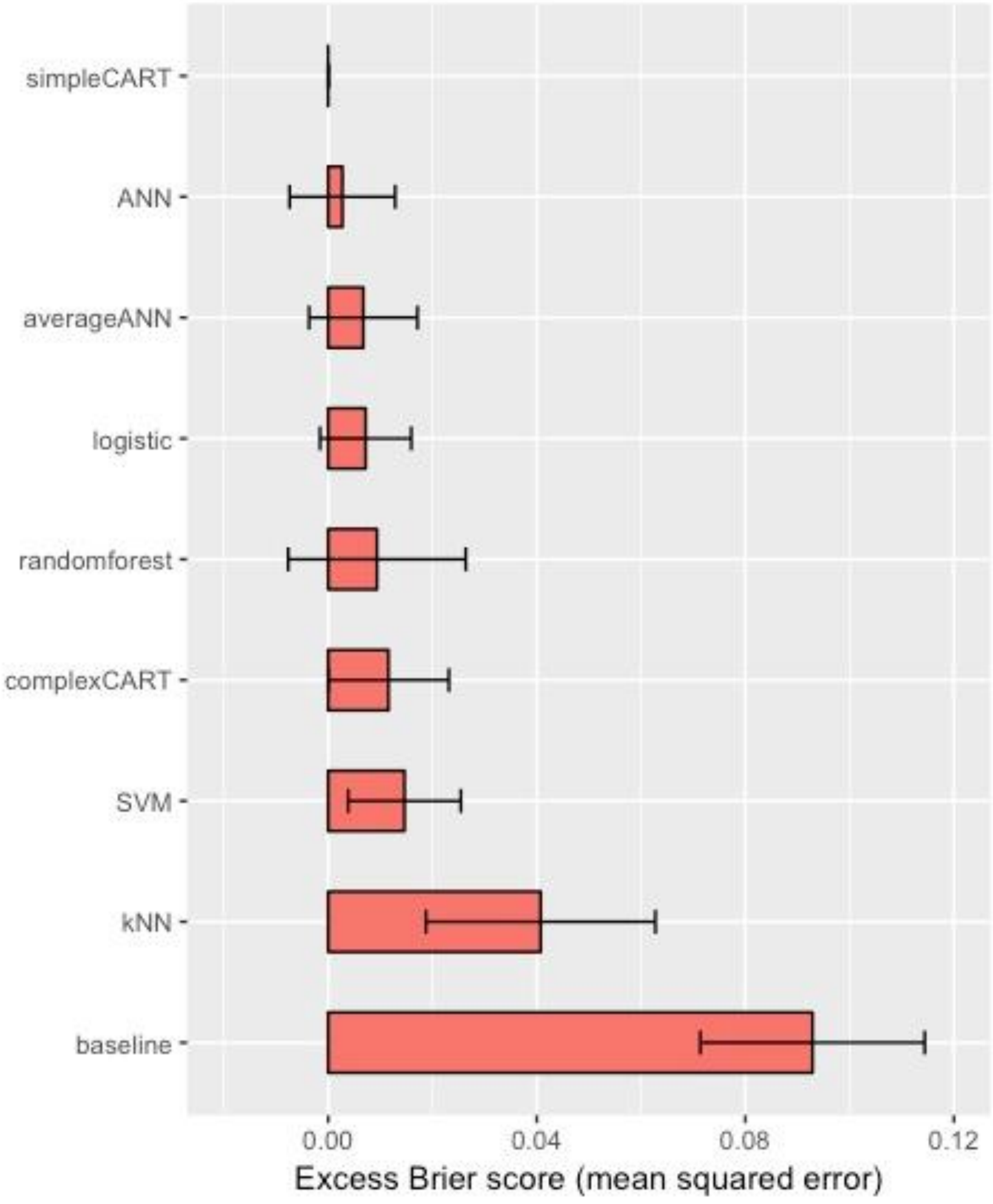
Method	Accuracy (high is good)	Brier score (low is good)
Everyone has a 39% chance of surviving	0.639	0.232
All females survive, all males do not	0.786	0.214
Simple classification tree	**0.806**	**0.139**
Classification tree (over-fitted)	0.806	0.150
Logistic regression	0.789	0.146
Random forest	0.799	0.148
Support Vector Machine (SVM)	0.782	0.153
Neural network	0.794	0.146
Averaged neural network	0.794	0.142
K-nearest-neighbour	0.774	0.180



simpleCART
ANN
averageANN
logistic
randomforest
complexCART
SVM
kNN
baseline
0.00
0.05
0.10
0.15
0.20
0.25
Brier score (mean squared error)
simpleCART
ANN
averageANN
logistic
randomforest
complexCART
SVM
kNN
baseline
0.00
0.05
0.10
0.15
0.20
0.25
Brier score (mean squared error)

- Potentially a very misleading graphic!
- When comparing, need to acknowledge that tested on same cases
- Calculate differences and their standard error

- How confident can we be that simple CART is best algorithm?

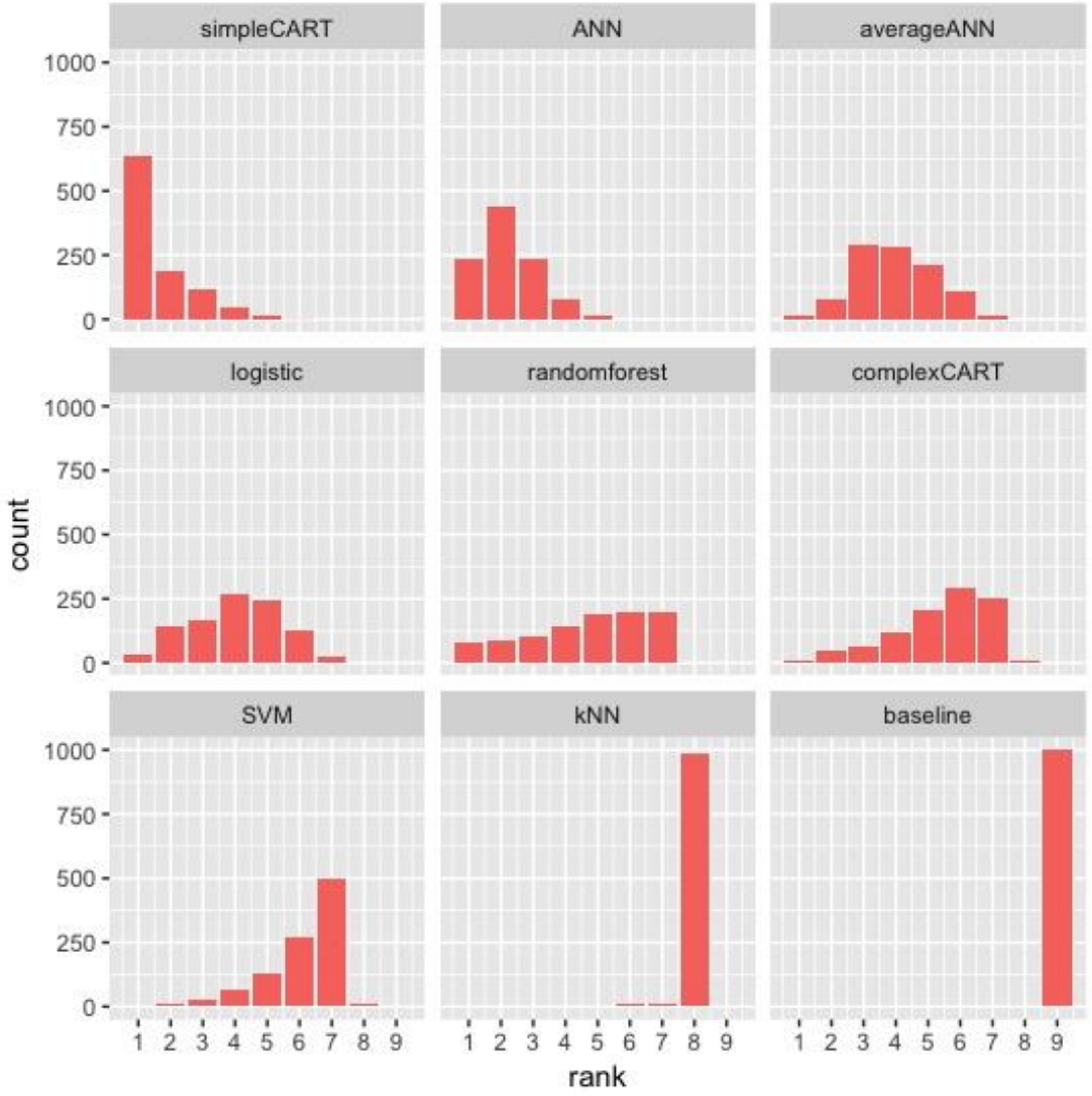


Ranking of algorithms

- Bootstrap sample from test set (ie sample of same size, drawn with replacement)
- Rank algorithms by performance on the bootstrap sample
- Repeat ‘000s of times

- (ranks actual *algorithm* – if want to rank *methods*, need to bootstrap training data too, and reconstruct algorithm each time)

Distribution of true rank of each algorithm



Probability of 'best':

63% simpleCART
23% ANN
8% randomforest

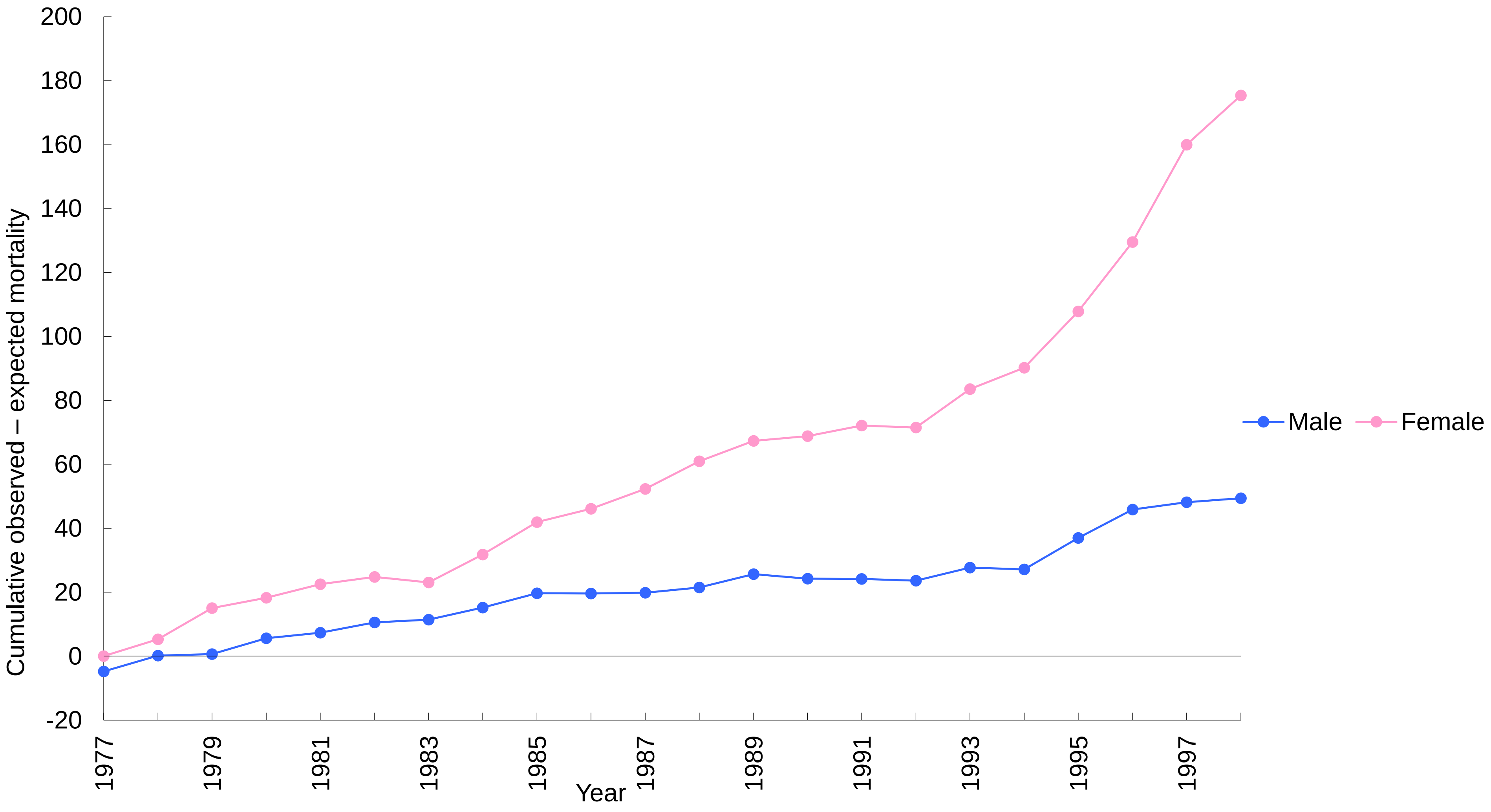
Who was the luckiest person on the Titanic?

- Karl Dahl, a 45-year-old Norwegian/Australian joiner travelling on his own in third class, paid the same fare as Francis Somerton
- Had the lowest average Brier score among survivors – a very surprising survivor
- He apparently dived into the freezing water and clambered into Lifeboat 15, in spite of some on the lifeboat trying to push him back.

- Hannah Somerton was left just £5, less than Francis spent on his ticket.

Hypothesis testing
Could Harold Shipman have been caught earlier?

- Using mortality rates from local GPs, calculate how many deaths he would have been **expected** to observe each year, under the **null hypothesis** that his mortality rates were normal.
- Subtract **expected** from **observed** number to get **excess mortality**

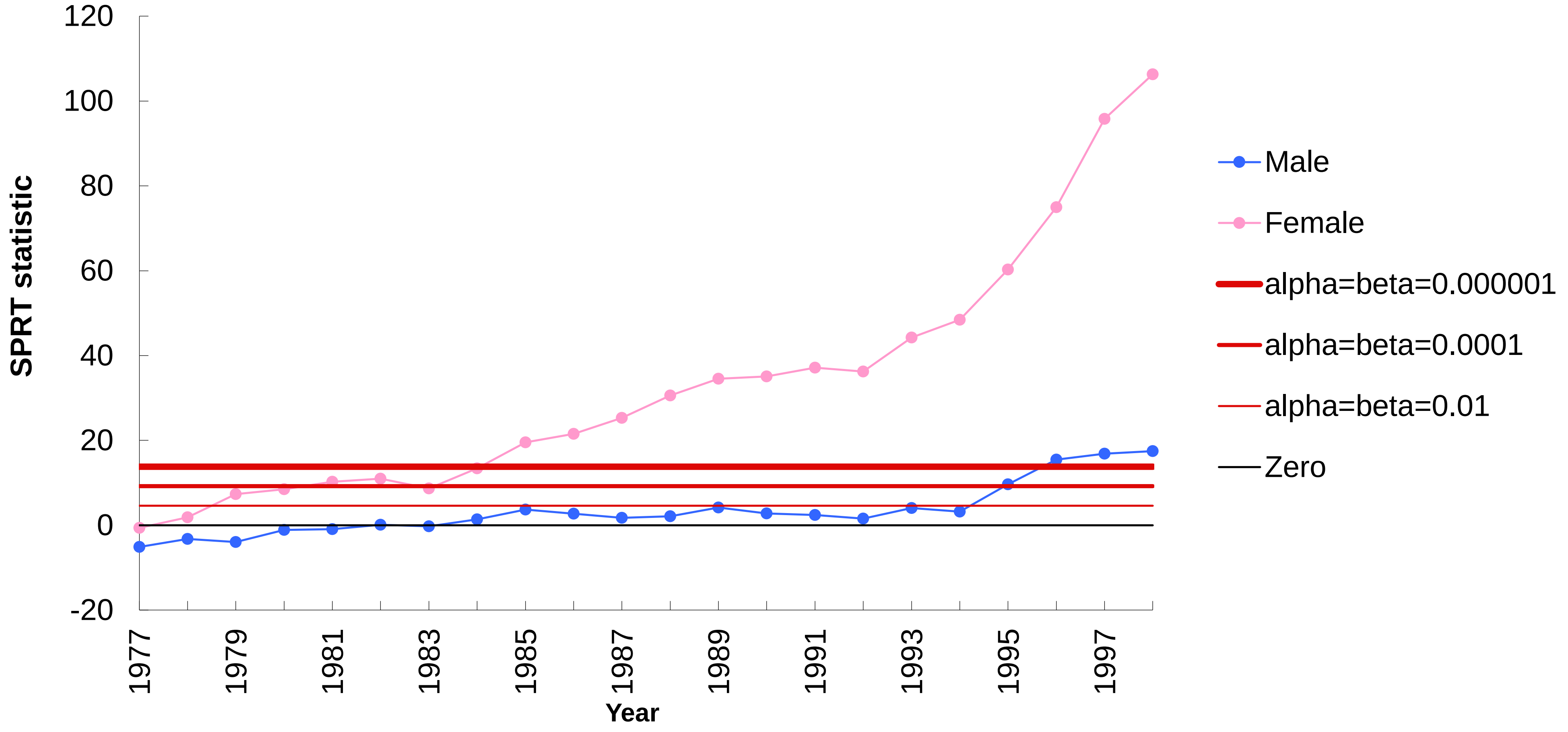


(NB: Shipman Inquiry total of definite or probable victims:
189 female > 65, **55** male over 65)

Hypothesis testing
Could Harold Shipman have been caught earlier?

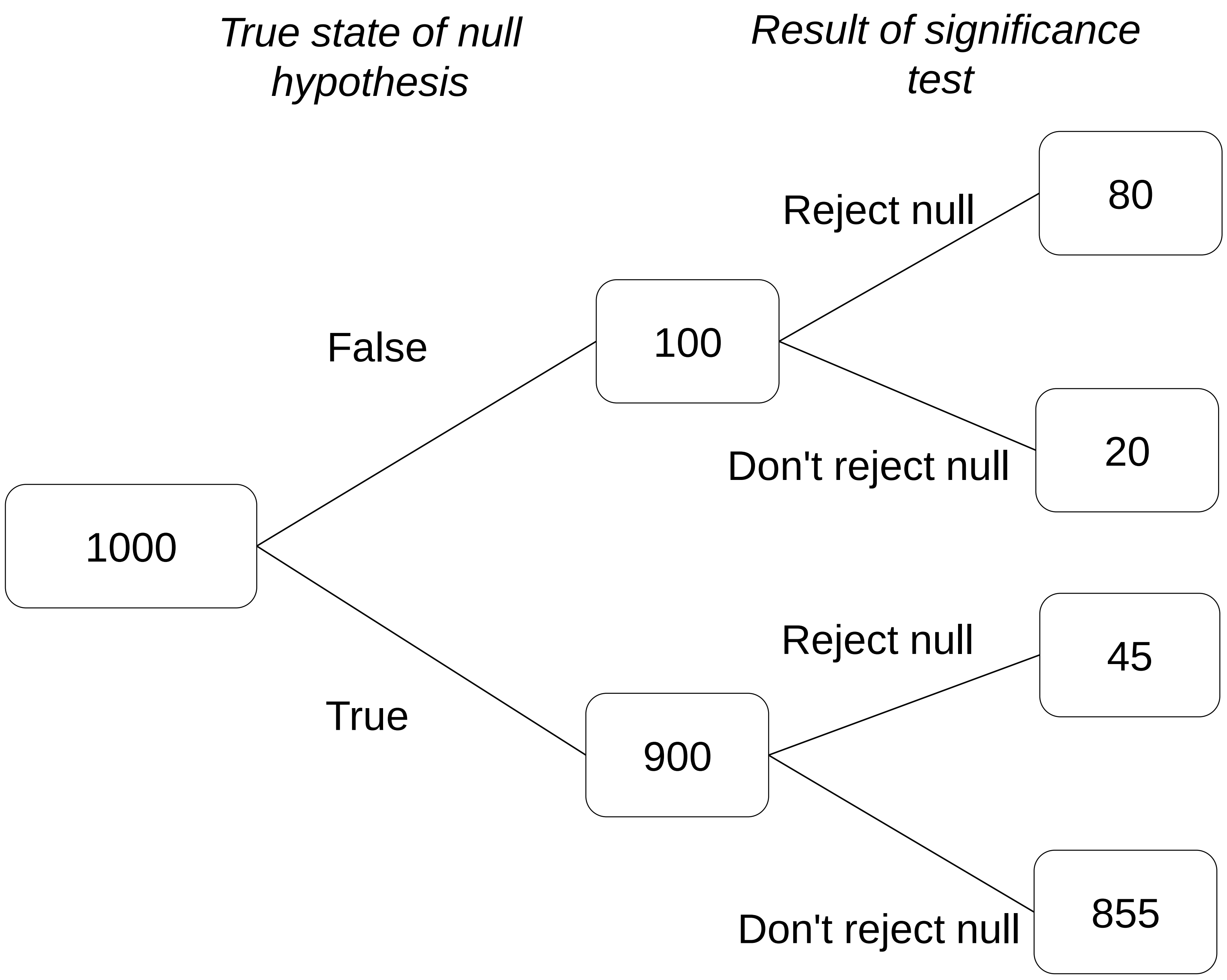
- But when to ‘blow the whistle’?
- This are two possible types of error -
  - **Type I error:** falsely accuse an innocent person
  - **Type II error:** miss someone with true increased risk
- This is an example of a *hypothesis test* used throughout science. Again, two possible types of error
  - **Type I error:** falsely claim an ’effect’ when nothing is there (ie the *null hypothesis* is true)
  - **Type II error:** miss a true effect
- Generally, we want to
  - control the probability of a Type I error at a low value (α)
  - make experiments large enough to make Type II errors rare (β)

Shipman: “Sequential probability ratio test” (SPRT) older females would have set off ‘alarm’ in 1985, after only 40 deaths

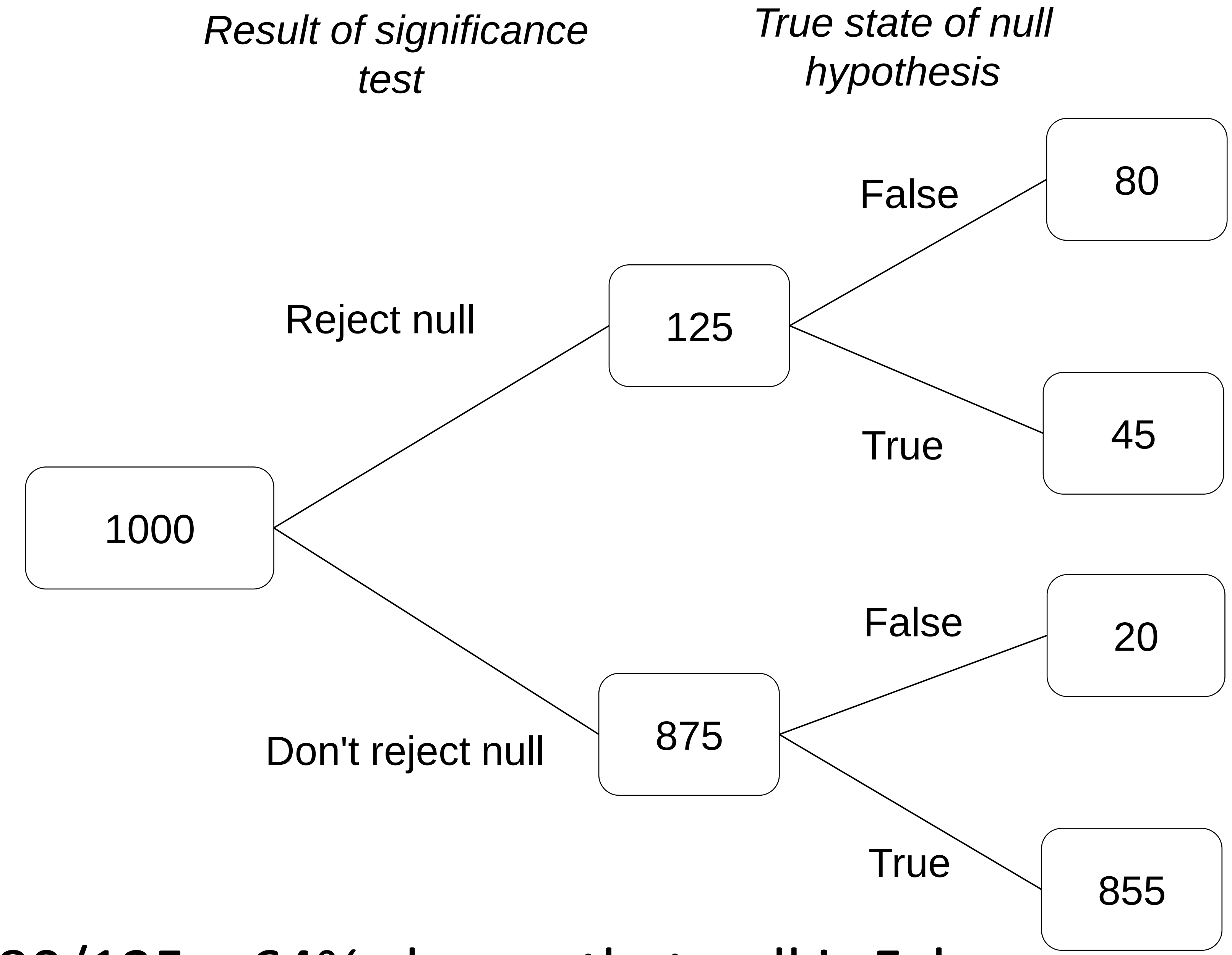


Probability and Bayes

- 1000 experiments
- Assume that in 10%, there is really an effect
- Size (α) = 5%
- Power ($1-\beta$) = 80%



- ‘reverse the tree’



- If reject the null, only 80/125 = 64% chance that null is False

Bayes theorem

- Initial odds that null hypothesis False = 10 / 90
- After 'significant' results, final odds that null hypothesis False = 80/ 45
- *Likelihood ratio* = $\dfrac{\text{Pr(significant result | null hypothesis False)}}{\text{Pr(significant result | null hypothesis True)}}$
- $= \dfrac{\text{Power}}{\text{Size}} = \dfrac{1-\beta}{\alpha} = \dfrac{0.80}{0.05} = 16$
- Bayes theorem:

the initial odds for the hypothesis x the *likelihood ratio*

= the final odds for a hypothesis.

$$\frac{10}{90} \times \frac{80}{5} = \frac{80}{45}$$

Probability and Bayes

What is the probability that the skeleton in a Leicester car park was really Richard III?

A recent case

- On Saturday 25 August 2012, archeologists started digging in a car park in Leicester – the site of Grey Friars friary

- In a few hours they found their first skeleton

- This was later claimed to be Richard III

nature
COMMUNICATIONS

ARTICLE

Received 5 Aug 2014 | Accepted 21 Oct 2014 | Published 2 Dec 2014

DOI: 10.1038/ncomms6631 OPEN

Identification of the remains of King Richard III

Turi E. King[1,2], Gloria Gonzalez Fortes[3,4,*], Patricia Balaresque[5,*], Mark G. Thomas[6], David Balding[6], Pierpaolo Maisano Delser[1], Rita Neumann[1], Walther Parson[7,8], Michael Knapp[9], Susan Walsh[10,11], Laure Tonasso[5], John Holt[12], Manfred Kayser[11], Jo Appleby[2], Peter Forster[13,14], David Ekserdjian[15], Michael Hofreiter[3,4] & Kevin Schürer[16]

$$\text{Likelihood ratio} = \frac{\text{probability of evidence, if skeleton were Richard III}}{\text{probability of evidence, if someone else}}$$

Suggested 'verbal equivalents' for bands of likelihood ratios

Value of likelihood ratio	Verbal equivalent
>1–10	Weak support for proposition
10–100	Moderate support
100–1000	Moderately strong support
1000–10,000	Strong support
10,000–1,000,000	Very strong
>1,000,000	Extremely strong

Standards for the formulation of evaluative forensic science expert opinion

Association of Forensic Science Providers

Evidence	Likelihood ratio (conservative estimate)	Verbal equivalent
Radiocarbon dating AD 1456–1530	2	Weak support
Age and sex of skeleton	5	Weak support
Scoliosis	212	Moderately strong support
Post-mortem wounds	42	Moderate support
mtDNA match	478	Moderately strong support
Y chromosome not matching	0.2	Weak evidence against
Combined evidence	6.5 million	More than extremely strong support

Conclusions – statistics for data science

- Motivate by problem solving
- Start with visualisation and exploring data
- Focus on concepts: what can be reasonably learned from data, biases, causation, etc
- Models and algorithms
- Probability can come much later
- Conditional probability / Bayes theorem taught through ‘expected frequency trees/