Big data:
the unexpected,
the unpredictable,
and the unwanted

David J, Hand
Imperial College, London
Disclaimer: The views expressed in this presentation are those of invited contributors and not necessarily those of the Institute and Faculty of Actuaries. The Institute and Faculty of Actuaries do not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation. The information and expressions of opinion contained in this presentation are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the Institute and Faculty of Actuaries.
Roots of actuarial and statistical work:

- Probability theory, from games of chance
- Data summarisation, from official statistics
- Observed regularities in data
  - Quetelet
  - Sir Edmund Halley and the first life table
  - John Graunt: *Bills of Mortality*
- etc
Increasing number of data sources and types with different properties

- Surveys
- Panel data
- Administrative data
- Transaction data
- Web scraped data
- Social media data
- ....

Experimental/observational
Big data
Open data
etc etc
Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Chris Anderson

Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough.

Chris Anderson

That's right – as far as it goes
But it does not go far enough
Statistical models
Making decisions in the modern world:

- Subjective
- Theory-driven models
- Data-driven models

Models vs algorithms

- Algorithms for prediction
- Models for understanding (and prediction)
Theory-driven example

Model the relationship between the height from which a stone is dropped and the time it takes to hit the ground

\[
\begin{align*}
\rightarrow \text{Data} & \quad \left( H_i, t_i \right) \quad i = 1, \ldots, n \\
\rightarrow \text{Model} & \quad t = \sqrt{\frac{2H}{a}} + \varepsilon \quad H = a \left( t + \varepsilon \right)^2 / 2
\end{align*}
\]
Weaknesses of theory-driven models

- Need a theory
- OK in theory rich domains
- Less so in other domains
- Bias, if you get theory wrong
**Data-driven example**

Logistic regression model for the probability that someone will purchase a product based on the values of their characteristics $x_1, x_2, \ldots, x_n$

$$
\log \left( \frac{p}{1-p} \right) = \sum_{j=1}^{d} \beta_j x_j
$$
Weaknesses of data-driven models

Assume:

- the future is like the past
- choose criterion to fit model to data
- good quality data
- no selection bias
- no gaming, feedback, etc
- ...

Leading to

- Brittleness: due to changes of circumstances
- Horses: algorithm choosing non-human features
The future is like the past

Sources: BBA, CCRG (Access cards only) added from 1974, Building Societies from 1996
Disconnect between productivity and typical worker compensation,* 1948–2013

Cumulative percent change since 1948

- Productivity
- Hourly compensation

Failure to model range of variation:

→ Predictive models may break down
→ Regression to the mean because underestimating variation
Example: Spurious correlations

Correlation = 0.992
Need to choose criterion to fit model to data

Optimum error rate: top-left to bottom-right
Optimum Gini: bottom-left to top right
Risk of mismatch between aims and algorithms
Different cluster methods revealing different shapes
Long sausage shapes vs compact clusters
Campbell’s law: *The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor*  
- e.g. schools enter for public exams only those expected to excel  
- e.g. ambulance response times

Bevan and Hamblin, 2009
Adding substance
Data driven model

Age, income, other loans, etc.
Theory-driven model

AGE

AMT

CGC

QUALITY

Age, income, other loans,....

CCJs, defaults, months in arrears, parking tickets, speeding fines,....
Data quality
“The majority of insurers around the world are failing to ensure the data that feeds their artificial intelligence (AI) systems is accurate, potentially undermining their business decisions.”

The Actuary, 25th April 2018
“Two students suffered ‘life threatening reactions’ when they were given enough caffeine for 300 cups of coffee. … spent several days in ICU…dialysis… Should have been given 0.3g of caffeine. Instead they were given 30g.”

The Times, 26 January 2017

The Mars Climate Orbiter
Launched 1998, but communication lost on September 1999 when the spacecraft trajectory brought it too close to Mars … because one of the software teams forgot to convert Imperial units to SI units
“Poor data quality costs the US economy around $3.1 trillion per year”

Source: IBM
Causes of poor quality data

Human error:

Shares in J-Com losing $200m after a broker tried to sell 610,000 shares for 1 yen each, instead of 1 share for 610,000 yen.

Poor data collection:

Peak at 11 November 19

Fabrication of data?

Scientific fraud

Source: Steen RG, Casadevall A, Fang FC (2013)
Berry and Linoff (2000) example:

“The data is clean because it is automatically generated – no human ever touches it”

But it turned out that 20% of transactions had

“arrived before they were sent .... not only did people never touch the data, but they didn’t set the clocks on the computers either”
Not merely human error
Bad data can occur in an unlimited number of ways

Cannot check a billion values by hand

The computer is a necessary intermediary
Maintain a healthy scepticism

Twyman’s Law:

*Any figure that looks interesting or different is usually wrong*
Non-response and refusals

The magazine survey which asks readers one question:

“Do you reply to magazine surveys?”

And discovers that apparently all the readers reply to surveys.

The Actuary, July 2006, editorial:

“A couple of months ago I invited all 16,245 of you to participate in our online survey concerning the sex of actuarial offspring.”

“... Well, I’m pleased to say that a number of you (13, in fact) replied to our poll.”
Other aspects of bad data:

*relevance,*

*timeliness,*

*consistency,*

*coherence,*

*availability,*

*and accessibility*
It is also a good rule not to put overmuch confidence in the observational results that are put forward until they have been confirmed by theory

If the data can speak for themselves

They can also lie for themselves

David Hand
Longevity
1871:
Male: 41.4
Female: 44.5
Gap: 3.3 years

Increasing life expectancy likely due to health improvements in young population e.g. childhood immunisation

Increasing life expectancy likely due to health improvements in older population e.g. heart disease treatment

Source: Decennial Life Tables, ONS
“A stunning increase in the life expectancy of New Yorkers over the past 20 years, compared with the rest of the country, has been driven by sharp declines in deaths from AIDS, homicide, smoking-related illnesses and, in a surprising twist, an increase in the numbers of immigrants, a new study has found.

…

The magnitude of the gains recalls those that followed major public health improvements, like the advent of sewage systems at the end of the 19th century.”
“Immigrants have much lower rates of smoking, AIDS and alcohol-related illnesses than native-born Americans, he said. The significant fall in homicides, down by 77 percent in the city since 1990, and death rates from AIDS, down by 85 percent over the same period, have helped drive improving life expectancies, according to the study.

…

foreign-born people live longer and life expectancy is pulled up by that…”
The economic parallel

- The productivity paradox
- Changing work patterns
- The gig economy

Competition from new types of source
Changing shopping habits
- Declining footfall
- Consumers spending less in stores but more on the internet
- £1 in 5 is now spent with online retailers
- Kodak and Polaroid’s film-based model destroyed by digital photography
- Blockbuster Video rental vs Netflix streaming
- Amazon vs Toys R Us, Maplin, etc

Established companies miss disruptive innovations because they give their customers what they want - instead of coming up with new things they don’t yet know they want

The next big thing is not the current big thing
“With enough data, the numbers speak for themselves”

Chris Anderson in *Wired* magazine, 2008

“the most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves, …”

Alfred Marshall, Inaugural Lecture to Chair in Political Economy, Cambridge, 1885
thank you !