Big Data, Data Science and Analytics – The End of Statistics?

ICCSSA Breakfast Session

Paul Fatti

University of the Witwatersrand

24 July, 2015

Overview of my talk

- I will give informal definitions of these three concepts:
  - Big Data
  - Data Science
  - Analytics
- Their relationships with Statistics
  - Threat
  - Exciting opportunity

Big Data

- Typically large to very large datasets from large organisations
- Captured automatically
  - As a direct “product” of the organisation
    - Large Hadron Collider (15 petabytes per annum)
    - Square Kilometre Array (22 petabytes per day)
  - As a by-product of an organisation’s main product
    - Data on the credit card transactions at a bank
    - Data on Google searches

Big data as a by-product

- Any process producing goods has two products:
  - The goods themselves
  - Data on the goods produced, which can be used to monitor and improve the process
    - Quality control
    - Process optimisation
      - Chemical plants

Google “Flu Trends”

- Correlated Google’s top 50 million search terms with:
- The number of cases of ‘flu reported by the Centres for Disease Control (CDC) in the USA.
- Google’s predictions had only a day’s delay, compared to the week or more it took for the CDC to put together a story based on doctors’ reports.

Google "Flu Trends" (2)

- Quick, accurate and cheap
- No data sampling required
- Not based on any notion of causality
  - Theory free
  - "The end of theory"
  - "with enough data, the numbers speak for themselves"
- However, the performance of Flu Trends deteriorated over time and four years later its forecasts were way out
  - No theory behind the algorithm so Google could not analyse what went wrong

Big data is “all the data”

- Using bank customer transaction data to develop or update credit scoring models.
  - Have all the data on customers, but
  - no data on customers whom it had previously rejected.
- Developing a credit scoring model based only on the customers that the bank had accepted would not be appropriate for the population who apply for a loan/credit card.

Bigger does not necessarily mean better. 1936 USA Presidential Elections

- FD Rooseveld (Democrats) vs Al Landon (Republicans)
- Literary Digest Poll postal survey sent to several million households (addresses based on phone and car registries) and received 2.4 million returns
  - Predicted FDR would lose 44% to 56%
- George Gallup based his survey on only 3000 interviews selected on the basis of stratified random sampling
  - Predicted FDR would win by a clear margin
- Landslide victory for FDR

Features of Big Data

Claimed	Actually
• N = All	• Never is – for one, it excludes future data
• Found data	• Unknown biases • "Data exhaust"
• Predict future purchases of products (Target Stores)	• Unmeasured false positives

However

- Big data is here to stay
- Business wants to use it to improve its decisions and processes
  - Statisticians tend (and need) to be sceptical about found data
    - Bias – unlikely to be representative of the population
    - We need to develop theory and practices that allow us to get around these shortcomings.
    - Wonderful time to be a statistician, but
    - **We cannot afford to miss this opportunity!**

Data Science

- Data driven approach to Statistics, where you let the data lead, instead of first considering what statistical model is appropriate. (John Tukey, 1962)
  - Include the use of machine learning algorithms to draw information from large datasets
- Most people are coming to data science from non-statistical backgrounds.
- Not enough statisticians available (or willing) to move into the field.

Data Science v Statistics

- The challenge for both professions is to manage and draw valid inferences from the large data sets that are becoming increasingly available
- Data scientists often begin their journey from within computer science or the natural sciences, rather than the statistician's mathematical route.
- Both parties are motivated by what can be achieved with data, but crucially this curiosity is inspired from different angles.
- But within these divergent approaches lies the collaboration that will ultimately benefit both professions.

Data in the Natural Sciences

- Many fields of research in the natural and other sciences are awash with data, and scientists in these fields are often not up to the task of getting the most out of this data.
  - Bioinformatics (Genetics)
    - Very large datasets
  - In ecology, tracking the movements of animals through cell phone and other networks
    - Large amount of data and the challenge is to infer the animal behaviours from this data
    - Hidden Markov modelling to analyse this data and infer the behavioural states of the animals
    - Use this to assess the animals' response to, for example, drought or predators

Generating results for Business from large data sets

- Business doesn’t want data, it wants answers/information from it.
  - The challenge is to analyse these large datasets
    - Manage missing data, outliers
    - Use or derive statistical models appropriate for the data
    - Draw valid inferences from this data
    - Use these inferences to help management make better decisions
    - Improve or optimise their processes

Analytics

- “By analytics we mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models and fact-based management to drive decisions and actions”
- “Analytics relies on the simultaneous application of Statistics, Computer Programming and Operations Research to quantify (and improve) performance”

Analytics and Operations Research

- INFORMS, the Institute for Operations Research and the Management Sciences (USA), defines Analytics as:
  - "The scientific process of transforming data into insight for making better decisions"
- Some years ago INFORMS decided to "adopt" Analytics and now promotes it as one of its disciplines.
  - Is analytics a more sexy term than OR?
  - Certified Analytics Professional (CAP) Certification

Statistics and Analytics?

- Should Statisticians be concerned that Analytics is trying to usurp their position?
  - Or should we be pleased that it offers more opportunities and increases the market for statisticians?
- No chance of Statistics usurping Analytics!
- Perhaps Statistics should rather look at how to tackle the challenges posed by Big Data, Data Science and Analytics.

Challenges for Statisticians

- Need to recognise the paradigm shift
  - Traditionally data has been scarce and usually originated from designed experiments or surveys (and this is still mostly true)
  - Now we have large amounts of “found data” and we need to learn how to draw valid inferences from this type of data
    - Avoid invalid inferences
  - Once in a lifetime opportunity for Statistics to participate in this paradigm shift required for the analysis of big data.

Challenges for Statisticians (2)

- Need to make Statistics more accessible to the Data Scientists and Analytics professionals
- Put data analysis more to the fore in the teaching programmes and use it to motivate the theoretical aspects of Statistics
- Recognise that students find Statistics a subject that is “hard” to understand

Challenges for teaching statistics

- Many of today's students are technologically savvy and used to accessing information from the Internet.
- Teaching statistics at school:
  - Introduce scholars to descriptive and graphical analysis of data obtained from the Internet
    - Use the results of this analysis to suggest possible inferences
    - Thereafter more formal inference, based on elementary probability and distribution theory

University teaching

- Promote Statistics as being required for:
  - Analysing Big Data
  - Studying Data Science
  - Studying Analytics
- Require computer science and applied mathematics students to do at least a first year course in Statistics.
  - Encourage them to consider it as a second major

University Teaching (2)

- Include practical computing projects in all major statistics courses
- Ensure that statistics students learn one or more computer languages.
  - At Wits students are taught to use:
    - SAS in second year
    - R in third year
  - Excel in first year?

Final thoughts

- Big Data, Data Science and Analytics can either be seen as:
  - A threat to Statistics
  - Or a unique opportunity
- Let's make it the latter!

References

- Tim Harford (2014), “Big data: are we making a big mistake?”, *Significance*, 11, 5 (Dec 2014), 14-19
- Royal Statistical Society Video (May 2015), “Data Science and Statistics: different worlds?”, http://www.statslife.org.uk/audio-visual/video?jut1=2
- “Breaking the barriers between the domain and the data: an interview with Chris Wiggins” (June 2015) http://www.statslife.org.uk/features/2281-breaking-barriers-between-the-domain-and-the-data-an-interview-with-chris-wiggins