A Swiss Statistician’s ‘Big Tent’ View on Big Data and Data Science

(Version 10 as of 19.10.2015)

Dr. Diego Kuonen, CStat PStat CSci
Statoo Consulting
Morgenstrasse 129, 3018 Berne, Switzerland

@DiegoKuonen  +  kuonen@statoo.com  +  www.statoo.info

Abstract

There is no question that big data have hit the business, government and scientific sectors. The demand for skills in data science is unprecedented in sectors where value, competitiveness and efficiency are driven by data. However, there is plenty of misleading hype around the terms ‘big data’ and ‘data science’. This presentation gives a professional Swiss statistician’s ‘big tent’ view on these terms, illustrates the connection between data science and statistics, and highlights some challenges and opportunities from a statistical perspective.
‘Statistics has contributed much to data analysis. In the future it can, and in my view should, contribute much more. For such contributions to exist, and be valuable, it is not necessary that they be direct. They need not provide new techniques, or better tables for old techniques, in order to influence the practice of data analysis.’

John W. Tukey, 1962
As of 19.10.2015:

Total views: 133,229

- Top 10 SlideShare presentations on data science and big data (#8 – September 2014)
- Top 24 SlideShare presentations on data mining (#16 – November 2014)
About myself

• PhD in Statistics, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
• MSc in Mathematics, EPFL, Lausanne, Switzerland.
• CStat (‘Chartered Statistician’), Royal Statistical Society, United Kingdom.
• PStat (‘Accredited Professional Statistician’), American Statistical Association, United States of America.
• CSci (‘Chartered Scientist’), Science Council, United Kingdom.
• Elected Member, International Statistical Institute, Netherlands.
• Senior Member, American Society for Quality, United States of America.
• CEO & CAO, Statoo Consulting, Switzerland.
• Senior Lecturer in Business Analytics and Statistics, Geneva School of Economics and Management (GSEM), University of Geneva, Switzerland.
• President of the Swiss Statistical Society (2009-2015).
@DiegoKuonen

- 30.11.2013: 3 followers
- 18.11.2014: 1’404
- 19.10.2015: 4’267

BIG DATA
Top 100 Influencers & Brands

Onalytica,
New York & London

#26
(January 2015)
http://goo.gl/4YSK9x
About Statoo Consulting (www.statoo.info)

• Founded Statoo Consulting in 2001.

• Statoo Consulting is a software-vendor independent Swiss consulting firm specialised in statistical consulting and training, data analysis, data mining and big data analytics services.

• Statoo Consulting offers consulting and training in statistical thinking, statistics, data mining and big data analytics in English, French and German.

⇝ Are you drowning in uncertainty and starving for knowledge?

⇝ Have you ever been Statooed?
‘Normality is a myth; there never was, and never will be, a normal distribution.’

Robert C. Geary, 1947
## Contents

1. Demystifying the ‘big data’ hype 9

2. Demystifying the ‘data science’ hype 18

3. What distinguishes data science from statistics? 36

4. Conclusion and opportunities (not only for statisticians!) 42
‘Data is arguably the most important natural resource of this century. ... Big data is big news just about everywhere you go these days. Here in Texas, everything is big, so we just call it data.’

Michael Dell, 2014
1. Demystifying the ‘big data’ hype

• ‘Big data’ have hit the business, government and scientific sectors.

⇝ The term ‘big data’ — coined in 1997 by two researchers at the NASA — has acquired the trappings of religion.

• But, what exactly are ‘big data’?

◊ The term ‘big data’ applies to an accumulation of data that can not be processed or handled using traditional data management processes or tools.

⇝ Big data are a data management infrastructure which should ensure that the underlying hardware, software and architecture have the ability to enable ‘learning from data’, i.e. ‘analytics’.
The following characteristics — ‘the four Vs’ — provide a definition:

- **Volume**: ‘data at rest’, *i.e.* the amount of data (⇝ ‘data explosion problem’), with respect to the number of observations (⇝ ‘size’ of the data), but also with respect to the number of variables (⇝ ‘dimensionality’ of the data);

- **Variety**: ‘data in many forms’, ‘mixed data’ or ‘broad data’, *i.e.* different types of data (*e.g.* structured, semi-structured and unstructured, *e.g.* log files, text, web or multimedia data such as images, videos, audio), data sources (*e.g.* internal, external, open, public), data resolutions (*e.g.* measurement scales and aggregation levels) and data granularities;

- **Velocity**: ‘data in motion’ or ‘fast data’, *i.e.* the speed by which data are generated and need to be handled (*e.g.* streaming data from machines, sensors and social data);

- **Veracity**: ‘data in doubt’, *i.e.* the varying levels of noise and processing errors, including the reliability (‘quality over time’), capability and validity of the data.
● ‘Volume’ is often the least important issue: it is definitely not a requirement to have a minimum of a petabyte of data, say.

⇝ Bigger challenges are ‘variety’ and ‘velocity’, and most important is ‘veracity’ and the related quality of the data.

⇝ Indeed, big data come with the data quality challenges of ‘small’ data along with new challenges of its own!

● The above definition of big data is vulnerable to the criticism of sceptics that these four Vs have always been there.

⇝ Nevertheless, the definition provides a clear and concise framework to communicate about how to solve different data processing challenges.

⇝ But, what is new?
‘Scientists have long known that data could create new knowledge but now the rest of the world, including government and management in particular, has realised that data can create value.’

Sean Patrick Murphy, 2013

Source: interview with Sean Patrick Murphy, a former senior scientist at Johns Hopkins University Applied Physics Laboratory, in the *Big Data Innovation Magazine*, September 2013.
‘The data revolution is giving the world powerful tools that can help usher in a more sustainable future.’

Ban Ki-moon, United Nations Secretary-General, August 29, 2014
‘To get the full business value from big data, companies need to focus less on the three Vs of big data (volume, variety, velocity) and more on the four Ms of big data: ‘Make Me More Money’!’

Bill Schmarzo, March 2, 2015
‘Do not focus on the ‘bigness’ of the data, but on the value creation from the data.’

Stephen Brobst, August 7, 2015

~⇒ The 5th V of big data: ‘Value’.
‘Data are an infrastructural resource — a form of capital that cannot be depleted. ... Data can be used and re-used to open up significant growth opportunities, or to generate benefits across society in ways that could not be foreseen when the data were created.’

OECD’s Directorate for Science, Technology and Innovation (STI),
OECD STI Policy Note, October 2015
‘Data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or a recommendation for action.’

W. Edwards Deming, 1942
2. Demystifying the ‘data science’ hype

- The demand for ‘data scientists’ — the ‘magicians of the big data era’ — is unprecedented in sectors where value, competitiveness and efficiency are data-driven.

- The Data Science Association defined in October 2013 the term ‘data science’ within their Data Science Code of Professional Conduct as follows:

  ◊ **Data science** is the scientific study of the creation, validation and transformation of data to create meaning.

  Source: www.datascienceassn.org/code-of-conduct.html
‘Data-Driven Decision making’ (DDD) refers to the practice of basing decisions on the analysis of data, rather than purely on gut feeling and intuition:

Let data drive decisions, not the Highest Paid Person's Opinion.

#HowGoogleWorks

HowGoogleWorks.net
‘One by one, the various crises which the world faces become more obvious and the need for hard facts [facts by analyzing data] on which to take sensible action becomes inescapable.’

George E. P. Box, 1976
• Data science has been dubbed by the *Harvard Business Review* (Thomas H. Davenport and D. J. Patil, October 2012) as

‘the sexiest job in the 21st century’

and by the *New York Times* (April 11, 2013) as a

‘hot new field [that] promises to revolutionise industries from business to government, health care to academia’.

⇝ But, is data science really new and ‘sexy’?
• The term ‘data science’ was originally coined in 1998 by the statistician Chien-Fu Jeff Wu when he gave his inaugural lecture at the University of Michigan.

〜 Wu argued that statisticians should be renamed data scientists since they spent most of their time manipulating and experimenting with data.

• In 2001, the statistician William S. Cleveland introduced the notion of ‘data science’ as an independent discipline.

〜 Cleveland extended the field of statistics to incorporate ‘advances in computing with data’ in his article ‘Data science: an action plan for expanding the technical areas of the field of statistics’ (International Statistical Review, 69, 21–26).
• Although the term ‘data scientist’ may be relatively new, this profession has existed for a long time!

◊ For example, Johannes Kepler published his first two laws of planetary motion, which describe the motion of planets around the sun, in 1609.

⇝ Kepler found them by analysing the astronomical observations of Tycho Brahe.

⇝ Kepler was clearly a data scientist!
‘Data science in a way is much older than Kepler — I sometimes say that data science is the ‘second oldest’ occupation.’

Gregory Piatetsky-Shapiro, November 26, 2013
‘I keep saying the sexy job in the next ten years will be statisticians.’

Hal Varian, 2009

‘And with ongoing advances in high-performance computing and the explosion of data, ... I would venture to say that statistician could be the sexy job of the century.’

James (Jim) Goodnight, 2010
‘I think he [Hal Varian] is behind — using statistics has been the sexy job of the last 30 years. It has just taken awhile for organisations to catch on.’

James (Jim) Goodnight, 2011
‘The demand for competent statisticians who can tease out the facts by analyzing data, planning investigations, and developing the necessary new theory and techniques, will continue to increase.’

George E. P. Box, 1976
• Looking at all the ‘crazy’ hype around the terms ‘big data’ and ‘data science’, it seems that ‘data scientist’ is just a ‘sexed up’ term for ‘statistician’.

⇝ It looks like statisticians just needed a good marketing campaign!
But, what is ‘statistics’?

Statistics is the science of ‘learning from data’ (or of making sense out of data), and of measuring, controlling and communicating uncertainty.

It is a process that includes everything from planning for the collection of data and subsequent data management to end-of-the-line activities such as drawing conclusions of data and presentation of results.

Uncertainty is measured in units of probability, which is the currency of statistics.

Statistics is concerned with the study of data-driven decision making in the face of uncertainty.
• However, data science is not just a rebranding of statistics, large-scale statistics or statistical science!

• Data science is rather a rebranding of ‘data mining’!

⇝ But, what is ‘data mining’?

◊ Data mining is the non-trivial process of identifying valid (i.e. being valid on new data in the face of uncertainty), novel, potentially useful, and ultimately understandable patterns or structures or models or trends or relationships in data to enable data-driven decision making.
The data science (or data mining) Venn diagram

‘The awesome nerds’

‘It is already time to kill the ‘data scientist’ title. ... The data scientist term has come to mean almost anything.’

Thomas H. Davenport, 2014

Source: Thomas H. Davenport’s article ‘It is already time to kill the ‘data scientist’ title’ in the Wall Street Journal, April 30, 2014.
3. What distinguishes data science from statistics?

• Statistics traditionally is concerned with analysing primary (e.g. experimental) data that have been collected to explain and check the validity of specific existing ideas (hypotheses).

⇝ Primary data analysis or top-down (explanatory and confirmatory) analysis.

⇝ ‘Idea (hypothesis) evaluation or testing’.

• Data science (or data mining), on the other hand, typically is concerned with analysing secondary (e.g. observational or ‘found’) data that have been collected for other reasons (and not ‘under control’ of the investigator) to create new ideas (hypotheses).

⇝ Secondary data analysis or bottom-up (exploratory and predictive) analysis.

⇝ ‘Idea (hypothesis) generation’.
The two approaches of ‘learning from data’ are complementary and should proceed side by side — in order to enable proper data-driven decision making.

**Example.** When historical data are available the idea to be generated from a bottom-up analysis (e.g. using a mixture of so-called ‘ensemble techniques’) could be

‘which are the most important (from a predictive point of view) factors (among a ‘large’ list of candidate factors) that impact a given process output (or a given indicator)?’.

⇝ Mixed with subject-matter knowledge this idea could result in a list of a ‘small’ number of factors (*i.e.* ‘the critical ones’).

⇝ The confirmatory tools of top-down analysis (statistical ‘Design Of Experiments’, DOE, in most of the cases) could then be used to confirm and evaluate this idea.

⇝ By doing this, new data will be collected (about ‘all’ factors) and a bottom-up analysis could be applied again — letting the data suggest new ideas to test.
‘Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.’

John W. Tukey, 1980
Data-driven decision making and scientific investigation (Box, 1976)

‘Experiments may be conducted sequentially so that each set may be designed using the knowledge gained from the previous sets.’

George E. P. Box and K. B. Wilson, 1951

⇒ Scientific investigation is a sequential learning process!

⇒ Statistical methods allow investigators to accumulate knowledge!
‘Statistics has been the most successful information science. Those who ignore statistics are condemned to re-invent it.’

Brad Efron, 1997
4. Conclusion and opportunities (not only for statisticians!)

- Decision making that was once based on hunches and intuition should be driven by data (⇒ data-driven decision making).

- Despite an awful lot of marketing hype, big data are here to stay and will impact every single domain!

⇒ The ‘age of big data’ could (and will hopefully) be a new golden era for statistics.

- Statistical principles and rigour are necessary to justify the inferential leap from data to knowledge.

⇒ At the heart of extracting value from big data lies statistics!
‘We are in the era of big data, and big data needs statisticians to make sense of it.’

Eric Schmidt and Jonathan Rosenberg, 2014

• Lack of expertise in statistics can lead (and has already led) to fundamental errors!

⇝ Large amounts of data plus sophisticated analytics do not guarantee success!

⇝ Historical results do not guarantee future performance!

• The key elements for a successful (big) data analytics and data science future are statistical principles and rigour of humans!

• Statistics, (big) data analytics and data science are aids to thinking and not replacements for it!
Do not neglect the following four principles that ensure successful outcomes:

– use of **sequential approaches** to problem solving and improvement, as studies are rarely completed with a single data set but typically require the sequential analysis of several data sets over time;

– having a strategy for the project and for the conduct of the analysis of data (**‘strategic thinking’**);

– carefully considering data quality and how data will be analysed (**‘data pedigree’**); and

– applying sound **subject matter knowledge** (**‘domain knowledge’**), which should be used to help define the problem, to assess the data pedigree, to guide analysis and to interpret the results.
‘The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.’

John W. Tukey, 1986
Some challenges from a statistical perspective include:

- the ethics of using and linking (big) data, particularly in relation to personal data, *i.e.* ethical issues related to privacy (*⇝* ‘information rules’ need to be defined), confidentiality (of shared private information), transparency (e.g. of data uses and data users) and identity (*i.e.* data should not compromise identity);

- the provenance of the data, *e.g.* the quality of the data — including issues like omissions, data linkage errors, measurement errors, censoring, missing observations, atypical observations, missing variables (*⇝* ‘omitted variable bias’), the characteristics and heterogeneity of the sample — big data being ‘only’ a sample (at a particular time) of a population of interest (*⇝* ‘sampling/selection bias’, *i.e.* is the sample representative to the population it was designed for?);
‘To properly analyze data, we must first understand the process that produced the data. While many ... take the view that data are innocent until proven guilty, ... it is more prudent to take the opposite approach, that data are guilty until proven innocent.’

Roger W. Hoerl, Ronald D. Snee and Richard D. De Veaux, 2014

– the visualisation of the data, e.g. using and developing effective graphical procedures;

– spurious (false) associations (⇝ ‘coincidence’ increases, i.e. it becomes more likely, as sample size increases, and as such ‘there are always patterns’) versus valid causal relationships (⇝ ‘confirmation bias’);

– the identification of (and the controlling for) confounding factors (⇝ ‘confounding-edness’, i.e. the attribution of the wrong factors to success and the superficial learning of observational data);
- multiple statistical hypothesis testing for explanation (not prediction!) with tens of thousands or even millions of tests performed simultaneously, often with complex dependencies between tests (e.g. spatial or temporal dependence), and the development of further statistically valid methods to solve large-scale simultaneous hypothesis testing problems;

- the dimensionality of the data (curse of dimensionality’, i.e. data become more ‘sparse’ or spread out as the dimensionality increases), and the related usage and development of statistically valid strategies for dimensionality reduction, e.g. using ‘embedded’ variable subset selection methods like ‘ensemble techniques’;
- the validity of generalisation (avoid ‘overfitting’, i.e. interpreting an exploratory analysis as predictive);

- the replicability of findings, i.e. that an independent experiment targeting the same question(s) will produce consistent results, and the reproducibility of findings, i.e. the ability to recompute results given observed data and knowledge of the data analysis pipeline;

- the nature of uncertainty (both random and systematic);
– the balance of humans and computers.

‘There’s no better way to lose money really quickly than through automated analytics. So I think we have to be very careful to not totally step out of the picture and let things go awry.’

Thomas H. Davenport, April 30, 2014
‘Business is not chess; smart machines alone can not win the game for you. The best that they can do for you is to augment the strengths of your people.’

Thomas H. Davenport, August 12, 2015
‘Most of my life I went to parties and heard a little groan when people heard what I did. Now they are all excited to meet me.’

Robert Tibshirani, 2012

Have you been Statooed?

Dr. Diego Kuonen, CStat PStat CSci
Statoo Consulting
Morgenstrasse 129
3018 Berne
Switzerland

email    kuonen@statoo.com
web      www.statoo.info
