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

(Version 5 as of 18.11.2014)

Dr. Diego Kuonen, CStat PStat CSci
Statoo Consulting
Statistical Consulting + Data Analysis + Data Mining Services
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 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.
About myself (about.me/DiegoKuonen)

- 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.
- CSci (‘Chartered Scientist’), Science Council, United Kingdom.
- Elected Member, International Statistical Institute, Netherlands.
- CEO, Statoo Consulting, Switzerland.
- Senior Lecturer in Statistics, Geneva School of Economics and Management, University of Geneva, Switzerland.
- President of the Swiss Statistical Society.
About Statoo Consulting (www.statoo.info)

- Statoo Consulting is a software-vendor independent Swiss consulting firm specialised in statistical consulting and training, data analysis and data mining services.
- Statoo Consulting offers consulting and training in statistical thinking, statistics and data mining in English, French and German.

⇝ Are you drowning in uncertainty and starving for knowledge?

⇝ Have you ever been Statooed?
Who has already been Statooed?

- Statoo Consulting’s clients include **Swiss and international companies** like

  - ABB;
  - Alcan Aluminium Valais;
  - Alstom;
  - Arkema, France;
  - AstraZeneca, United Kingdom;
  - Barry Callebaut, Belgium;
  - Bayer Consumer Care;
  - Berna Biotech;
  - BMW, Germany;
  - Bosch, Germany;
  - Cargill;
  - Crédit Lyonnais, United Kingdom;
  - CSL Behring;
  - F. Hoffmann-La Roche;
  - GfK Telecontrol;
  - GlaxoSmithKline, United Kingdom;
  - H. Lundbeck, Denmark;
  - Lombard Odier Darier Hentsch;
  - Lonza;
  - McNeil, Sweden;
  - Merck Serono;
  - Nagra Kudelski Group;
  - Nestlé;
  - Nestlé Frisco Findus;
  - Nestlé Research Center;
  - Novartis Pharma;
  - Novelis;
  - Pfizer, United Kingdom;
  - Philip Morris International;
  - Pioneer Hi-Bred, Germany;
  - PostFinance;
  - Procter & Gamble, Germany;
  - publisuisse;
  - Rodenstock, Germany;
  - Sanofi-Aventis, France;
  - SAP France and SAP Ireland;
  - Saudi Arabian Oil Company, Saudi Arabia;
  - Smith & Nephew Orthopaedics;
  - Swisscom Innovations;
  - Swissmedic;
  - The Baloise Insurance Company;
  - tl — Public Transport for the Lausanne Region;
  - Wacker Chemie, Germany;
  - upc cablecom;
as well as **Swiss and international government agencies, nonprofit organisations, offices, research institutes and universities** like

- King Saud University, Saudi Arabia;
- Lausanne Hotel School;
- Institute for International Research, Dubai, United Arab Emirates;
- Paul Scherrer Institute (PSI);
- Statistical Office of the City of Berne;
- Swiss Armed Forces;
- Swiss Federal Institute for Forest, Snow and Landscape Research;
- Swiss Federal Institutes of Technology Lausanne and Zurich;
- Swiss Federal Office for Migration;
- Swiss Federal Research Stations Agroscope Changins–Wädenswil and Reckenholz–Tänikon;
- Swiss Federal Statistical Office;
- Swiss Institute of Bioinformatics;
- Swiss State Secretariat for Economic Affairs (SECO);
- The Gold Standard Foundation;
- Unit of Education Research of the Department of Education, Canton of Geneva;
- Universities of Applied Sciences of Northwestern Switzerland, Technology Buchs NTB and Western Switzerland;
- Universities of Berne, Fribourg, Geneva, Lausanne, Neuchatel and Zurich.
‘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 10

2. Demystifying the ‘data science’ hype 19

3. What distinguishes data science from statistics? 36

4. Conclusion and opportunities (not only for statisticians!) 39

Copyright © 2001–2014, Statoo Consulting, Switzerland. All rights reserved.
‘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

• There is no question that ‘big data’ have hit the business, government and scientific sectors.

⇝ Indeed, the term ‘big data’ has acquired the trappings of religion!

⇝ However, there are a lot of examples of companies that were into ‘big data’ before they were called ‘big data’ — a term coined in 1997 by two researchers at the NASA.

• But, what exactly are ‘big data’?

⇝ In short, 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 architecture and most related challenges are IT focused!
The following characteristics — known as ‘the four Vs’ or ‘V^4’ — provide one standard definition of big data:

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

- **‘Variety’**: ‘data in many forms’ or ‘mixed 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) and data resolutions;

- **‘Velocity’**: ‘data in motion’, *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.
‘While its size receives all the attention, the most difficult aspect of big data really involves its lack of structure.’

Thomas H. Davenport, 2014

• ‘Volume’ is often the least important issue: it is definitely not a requirement to have a minimum of a petabyte of data, say.

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

िम Especially the combination of different data sources (such as combining company data with social networking data and public data) — resulting from ‘variety’ — provides a lot of insights and this can also happen with ‘smaller’ data sets.

• The above standard definition of big data is vulnerable to the criticism of sceptics that these four Vs have always been there.
‘The only thing really new about big data is the cool new name.’

Daniel J. Solove, September 16, 2014

Nevertheless, the above definition of big data provides a clear and concise ‘business’ framework to communicate about how to solve different data processing challenges.

But, what is new?
“Big Data’ ... is the simple yet seemingly revolutionary belief that data are valuable. ... I believe that ‘big’ actually means important (think big deal). 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.’

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.
‘Suddenly it makes economic sense to try to extract value from all this data out there.’

Sean Owen, 2014

Source: interview with Sean Owen, Cloudera’s director of data science, in Research Live, May 19, 2014.
‘The term ‘big data’ is going to disappear in the next two years, to become just ‘data’ or ‘any data’. ’

Donald Feinberg, 2014
‘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 driven by data.

• The Data Science Association defined in October 2013 the terms ‘data science’ and ‘data scientist’ within their *Data Science Code of Professional Conduct* as follows (see www.datascienceassn.org/code-of-conduct.html):

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

  - A **data scientist** is a professional who uses scientific methods to liberate and create meaning from raw data.
• ‘Data-Driven Decision making’ (DDD) refers to the practice of basing decisions on data, rather than purely on intuition:

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

#HowGoogleWorks

HowGoogleWorks.net
• 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
• 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 can be defined as the science of ‘learning from data’ (or of making sense out of data), and of measuring, controlling and communicating uncertainty.

- It includes everything from planning for the collection of data and subsequent data management to end-of-the-line activities such as drawing conclusions of numerical facts called 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 uncertainty and with the study of 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’!

‘The terms ‘data science’ and ‘data mining’ often are used interchangeably, and the former has taken a life of its own as various individuals and organisations try to capitalise on the current hype surrounding it.’

Foster Provost and Tom Fawcett, 2013
‘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.
But, what is data mining?

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or structures or models or trends or relationships in data to enable data-driven decision making.

~ Non-trivial: it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers.

~ Valid: the patterns hold in general, i.e. being valid on new data in the face of uncertainty.

~ Novel: the patterns were not known beforehand.

~ Potentially useful: lead to some benefit to the user.

~ Understandable: the patterns are interpretable and comprehensible — if not immediately then after some postprocessing.
The data science (or data mining) Venn diagram

‘Statistics has been the most successful information science. Those who ignore statistics are condemned to re-invent it.’

Brad Efron, 1997
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) 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’.
  ⇝ Knowledge discovery.
• 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 output?’.

⇝ 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
4. Conclusion and opportunities (not only for statisticians!)

- Data, and the capability to extract useful knowledge from data, should be regarded as key strategic assets!

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

- Extracting useful knowledge from data to solve ‘business’ problems must be treated systematically by following a process with reasonably well-defined stages.

  → Like statistics, data science (or data mining) is not only modelling and prediction, nor a product that can be bought, but a whole iterative problem solving cycle/process that must be mastered through interdisciplinary and transdisciplinary team effort.
‘If I had only one hour to save the world, I would spend fifty-five minutes defining the problem, and only five minutes finding the solution.’

Albert Einstein
• One might rightly be **sceptical** about big data because the idea is still fuzzy and there is an awful lot of **marketing hype** around.

~~> However, **big data are here to stay** and will, over time, **impact every single ‘business’**!

~~> The ‘age of big data’ will (hopefully) be a **golden era for statistics**.

• **There is convincing ‘evidence’** that data-driven decision making and big data technologies will **substantially improve ‘business’ performance**.

• **Statistical rigour is necessary** to justify the inferential leap from data to knowledge.

~~> That is, **at the heart of extracting value from big data lies statistics**.

• **Lack of expertise in statistics** can lead (and has already led) to **fundamental errors**!
‘We are in the era of big data, and big data needs statisticians to make sense of it.’

Eric Schmidt and Jonathan Rosenberg, 2014

‘Statisticians have spent the past 200 years figuring out what traps lie in wait when we try to understand the world through data. The data are bigger, faster and cheaper these days — but we must not pretend that the traps have all been made safe. They have not.’

Tim Harford, 2014

Source: Tim Harford’s article ‘Big data: are we making a big mistake?’ in the Financial Times Magazine, March 28, 2014.
‘There are a lot of small data problems that occur in big data. They do not disappear because you have got lots of the stuff. They get worse.’

David Spiegelhalter, 2014
The challenges from a statistical perspective include:

- the need to think really hard about a problem and to understand the underlying mechanism that drive the processes that generate the data (⇝ ask the ‘right’ question(s)!);

- 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?);
– 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 **visualisation** of the data;

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

– the identification of (and the controlling for) **confounding factors**;
- 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 (\(\leadsto\) 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’), and the replicability and reproducibility of findings;

- the nature of uncertainty (both random and systematic);

- the balance of humans and computers.

‘Data and algorithms alone will not fulfil the promises of big data. Instead, it is creative humans who need to think very hard about a problem and the underlying mechanisms that drive those processes.’

David Park, 2014
‘We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big data equation as the numbers themselves.’

Kate Crawford, 2013
‘The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. ... Data-driven predictions can succeed — and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves.’

Nate Silver, 2012

‘As you plan your big data analytics journey, make sure you know your business goals, understand the data life-cycle and have a plan for delivering insights to everyone who needs them. Taking a holistic approach to big data analytics will be an important step in addressing technical challenges and reaching your business goals.’

John Whittaker, 2014
‘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

e-mail kuonen@statoo.com

@DiegoKuonen

web www.statoo.info

/Statoo.Consulting