Philosophising data: A critical reflection on the 'hidden' issues

Jackie Campbell[1], Victor Chang[2], Amin Hosseinian-Far [3]
{[1]J.Campbell, [2]V.I.Chang, [3]A.Hosseinian-Far}@leedsbeckett.ac.uk

Abstract. This paper aims to critically reflect on the processes, agendas and use of Big Data by presenting existing issues and problems in place and consolidating our points of views presented from different angles. This paper also describes current practices of handling Big Data, including considerations of smaller scale data analysis and the use of data visualisation to improve business decisions and prediction of market trends. The paper concludes that alongside any data collection, analysis and visualisation, the 'researcher' should be fully aware of the limitations of the data, by considering the data from different perspectives, angles and lenses. Not only will this add the validation and validity of the data, but it will also provide a 'thinking tool' by which to explore the data. Arguably providing the 'human skill' required in a process apparently destined to be automated by machines and algorithms.

Keywords: Data Philosophy, Big Data, Epistemology, Ontology, Interpretivism, Positivism

1. Introduction and background

"Data is the new oil" (Thorpe, 2012) and Big Data can be used to know us better than we know ourselves (Mayer-Thurman and Cukier, 2013). Based on our Google searches, flu outbreaks can be predicted (known as nowcasting) (Google, 2014; Dugas; 2012); based on our social lifestyle data, the authorities will know when we will commit a crime before it has even been committed (Berk, 2009). These kinds of claims and threats are the current strap lines promoting the use and potential for Big Data and Business Intelligence (BI). Alongside this new research efforts in economics, Kahneman questions our basic skills as decision makers, claiming that human decisions are naturally biased and usually flawed (Kahneman, 2013). Kahneman argues (amongst other things) that if recent financial decisions had been based on data, the current financial crisis in the UK, US and now apparently China would never have happened, which have been supported by analyses undertaken by Chang (2014) in his studies.

The power of Big Data combined with the ability to ignore our own intuition could provide an entirely new paradigm to the processes by which we do business, research and make decisions. However, data still comes with its flaws: bias data, data with

quality issues, interpreted data, assumed data, mismatched data, subjective data, dated data, ethically questionable data, data from 'chosen' samples, data designed to prove a point, lies and statistics. Data need to be reprocessed, reorganised and restructured before any forms of analysis can be conducted. Contemplation of systems as black boxes and since the number of inflows and outflows through the black boxes is massive, a data flaw would be very challenging to spot. Hence, it would make it more difficult to bring about a new skill and area of concern; one which academics and professionals alike should develop and consider. That would be the ability to know the data and results for what they are, to understand the version of the 'truth' that this data represents, and to ensure the piece of information is not lost in the subsequent use and promotion of the findings.

This paper aims to critically reflect on the processes and use of Big Data by returning to the issues and considerations of smaller scale data analysis and research. The issues are unpacked with respect to the data findings and the search for the 'truth' or a perspective on the truth. The paper concludes that alongside any data collection and analysis the 'researcher' needs to be fully aware of the limitations of the data, by considering the data from different perspectives, angles, and lenses. Not only will this add the validation and validity of the data, but it will also provide a 'thinking tool' by which to explore the data. Arguably providing the 'human skill' required in a process apparently destined to be automated by machines and algorithms.

2. Big Data – The new oil or fools' gold?

There are some amazing data mining discoveries based on the leverage of Big Data; UPS used predictive analysis by monitoring and replacing specific parts they had saved on repair costs (FieldLogix, 2014). Deadly manhole explosions were predicted in New York (Ehrenberg, 2010). Walmart discovered that just before a hurricane, people in America bought an unusually large number of pop tarts (Hays, 2004). Data usually serves a purpose to prove or disprove hypotheses and theories, and excellent example of which is an investigation into the relationship between mobile phone usage and brain tumours (Frei et al, 2011). The Danish cell phone operators and health care service worked together to provide the data for a study into the relationship between brain tumours and cancer. The results showed no direct correlation. The benefits of adopting Big Data are as follows. First, the process of selecting data or population sampling is not required, as researchers can just take it 'all' and being persuaded to ignore data validity issues based on the 'general trend' provided by such a vast dataset. Second, Big Data processing can highlight the part of the datasets which reflects the core part of the problem. For example, medical data analysis can find the direct correlation between the lifestyle and genetic history with breast cancer.

It is interesting that 70% of companies say they are keeping data, but they do not know what for (Avanade, 2012). In terms of the amount of data being produced, current estimates are that 90% of the world's data as of 2013 was created in 2012 and 13 alone, and there will be 44 times as much by the year 2020 (ScienceDaily, 2013). It is being called 'the new oil', yet clearly

companies are aware they have the raw materials, but do not know how to 'spin' the oil. Big Data, in terms of project lifecycle and requirements misses out on the 'data collection' and 'data design' stages that would be perceptible in a research project (Bazeley, 2013). This means that data analysts are more commonly brought in to 'resuscitate' (Bazeley, 2013) the data, not an impossible task, but more challenging and arguably more costly than a data collection and design process targeted at specific aims, objectives and requirements (Grbich, 2013).

3. Bigger data = bigger problems?

Volume, velocity and variety are the three elements of Big Data. Hence, the size, amount and growth of the data available can influence how data can be used. This does not excuse the data quality problems and challenges. Data quality issues can cause complications: a girl died in South America after receiving a donor transplant of the wrong blood group in data error (NYTimes, 2004), the outdated land data showed the area hit by Hurricane Katrina in 2005 to be of low risk of flooding causing many of the residents insurance to be invalid (Campbell et al, 2006). There are data related issues reported frequently, but even on a personal level almost anyone can tell you of a data quality problem they have experienced.

The consequence of these errors may or may not be significant, and data storage facilities aim to reduce data quality problems by enforcing constraints (which is not null, check, data types, etc.) and the relational database model is well designed to protect the data [Date, 2004; Attaran & Hosseinian-Far, 2011]. Big Data however does not lend itself to the relational database model. The preferred data stores are hugely denormalised structured star schema models (Inmon, 2005; Kimball, 2003) or for unstructured data completely denormalised databases such as NoSQL or Hadoop (Smith, 2007). Big Data claims that data quality issues are minimalized due to the vast amount of data (Mayer-Thurman and Cukier, 2013). However they are still there even if the nature of the algorithms ignores them their very presence could be of consequence and should still be considered.

Knowing our requirement and expectation from the data is important for Big Data science. There is a difference in asking for a bank balance and a likelihood of rain, one expects an entirely accurate answer and the other is surrounded by a number of assumptions and potentials. Which takes us back to the original objectives of the data analysis – what is the information we are looking for? - General trend or absolute, qualitative or quantitative? These are considerations that researchers and data analysts spend some time deciding, exploring and identifying as part of the process. Does Big Data benefit from such consideration? Judging by the number of companies who have collected the data and don't know what to do with it, it would be challenging to provide a positive answer to this question.

The next few sections will consider the data lifecycle in more details in an attempt to unpack some of the considerations and issues with each phase in a data analysis project.

3.1. The Data 'project'

"Whenever we use data we are forced to think" (Bazeley, 2013) and this is a good start to any project – thinking and considering what is sought. Ideally the aims and objectives of data

analysis should be identified at the start of the project and then lead on to appropriate data collection phase. The characteristics of a 'Big Data' project can make this more complicated. The project is likely to have many stakeholders, each with their own requirements and hypothesis. Not all these stakeholders may fully appreciate the nature of the project, increasing the likelihood of the project requirements changing as the 'users' understand the capability and potential of the system. This is a recognised approach to data warehouse projects – to start small and 'grow' as the potentials are realised (Smith, 2014). This lack of project requirements is a basis of the argument between Kimball and Inmon as to whether to bring over 'all' the data (at the lowest granularity) into the data warehouse (Kimball) or to just bring over the required data (Inmon) (Kimball, 2003; Inmon, 2005). Kimball arguing that by bringing all the data warehouse, it can be more flexible and supportive to the ever changing requirements and questions of a company. This is because that they learn to appreciate the potential of the data and their business changes (Kimball, 2003). Inmon prefers to acknowledge that the data marts of a data warehouse will be rebuilt many times over its course, leading to ever evolving designs and data extractions to support a company in a more 'as and when' style (Inmon, 2005).

As a project, data analyses and research and statistics have been around for a while. They have revolved around reasonably small scale data, often collected by a few people with specific intentions as to the purpose of the data, according to the requirements of the project, sample data sets are selected through appropriate data analysis techniques (Baseley, 2013). It seems that Big Data projects are currently treated and understood by most companies to be the same as data analysis projects, though with much more data. This is suggested that this view is not correct. It should go even further to separate Big Data projects into two further generic types:

1. Big 'data analysis' projects i.e. the same kind of data analysis that has been around for years; looking at sales patterns, looking for clusters, time based analysis, just with more data. And
2. 'Big Data' analysis projects i.e. projects which are searching through the large sets of randomly related data for correlations and/or something of interest e.g. data mining.

"Research is messy" (Kincheloe, n.d). Sometimes it is compared to a patchwork quilt; each 'patch' representing a small part of the 'whole' of the research, each its own separate 'research project' designed with considered data collection and analysis methods to investigate the question (Kincheloe, n.d.). It is suggested that within either type of 'Big Data' project, there still be the smaller investigations. It could well be that a project would like to do both as "whenever we use data, we are forced to think" (Bazeley, 2013), and in thinking, we wonder how the sales relate to the weather and how the weather improves our mood and how our mood affects our spending …..Which is all of interest and worthy of investigation, but each of those variables (sales, weather, mood, spending) lend themselves to different identifications, classifications and objectivity.

Academic research projects are often categorised as being Qualitative or Quantitative. Qualitative research is an approach where tend to focus on thinking about the quality of things

rather than the quantity (Bazeley, 2013). It is often used to look for 'causality' – if one event causes another event – and to 'understand' or 'explore'. The findings are not absolute, but in understanding we can aim to improve. Quantitative research is generally looking for more absolute truths much of the analysis is based around correlations (Marsh and Elliot, 2008). Note the difference between the statements 'with their lack of education Bill and Stephen struggled to find meaningful work' and 'low education predicts poor employment rates'. The first uses people and emotive wording to encourage us to relate and question, the second reads as if to present the facts. Potentially both 'analysis' come from the same data set. The first seems more like qualitative data – it is known that Stephen 'struggles' and he wanted 'meaningful' employment. The second quantitative data – based on employment and education statistics (Bazeley, 2013). The aim of the research here was important, if it was to see if low education affects employability then to discover that 'Low education predicts poor employment rates' is useful. If we are looking to 'understand the difficulties of low education on men' (within a specific age bracket/area …) then the first statement goes some way towards that. This example provides a good illustration of where initial research (into employability statistics and education standards) can provide a rationale for further research (understanding the difficulties).

3.2. Data collection

Any kind of data analysis will involve data collection. An appropriate methodology may be selected based on the requirements of the data investigation. Academics spend a lot of time effort and research in identifying a methodology in keeping with the objectives of the research (Bazeley, 2013). A Case study approach could be used to investigate a specific 'case', or a 'grounded theory' approach used to collect data and see where it takes us (Cresswell, 2007). More radically, a 'phenomenological' approach would aim to explore a phenomenon by 'bracketing off' all prior opinions, perceptions, tand heories on the research (Cresswell, 2007). Industry projects are generally not considered for such theoretical methodologies; however it can be argued that often the approach is implicit in the project. Phenomenology can be defined as 'looking for anything of interest' (all data). By analysing data relating to employability and education – a case study is forming. Researchers 'collect' data via narrative extraction, interviews, focus groups, observations, surveys, etc. Industry, of course, uses these techniques as well, perhaps not with such a thorough evaluation of the methods and rigorous design of tools. Although more agile and with speedier life cycle!

Academics are forced to consider the limitations, validity and bias of their data. They consider their rationale for and position in the research; they critique the effect this will have on the data. In business, a data analysis/mining specification would likely address the same points. Research is often driven by personal interest (Bazeley, 2013), often it puts the researcher as an 'insider' to the research – this can create ethical and validation issues (Cresswell, 2007). The personal interest could well be driving a personal opinion or hypothesis that is subconsciously 'set up' to discover (Crawford and Boyd, 2011). Discussing and looking for concerns does not eliminate them and can sometimes lead to well critiqued arguments. It does however, encourage us to think about the data in other ways

which may lead to greater understanding.

The argument with 'Big Data' projects is that the biases, filters, codifications in the data are less transparent. The data may be 'bought' in, the design of the original data collection, may be lost or not available, and the codifications may not be as it was sought. So the data is compromised and in turn our requirements become compromised. To use twitter data to evaluate modern language, has its limitations as offensive language has been removed (Mayer-Thurman and Cukier, 2013). Google data, insurance data, National Health Services data belong to a certain demographic (is the demographic effectively a 'case study'?). It is worth noting that which organisations (National Health Services, government) have non-selective data sets, offering some of the nearest N=all datasets available in a country.

The data always represents something and this is where a smaller case study can identify more clearly what it is representing, and where the data collection methods of an academic researcher should be understood and defended rigorously. On a larger scale the decisions as to which data was to be collected still occurred, the biases in that data will still exist based on who designed the original systems, the intention of the original systems, the original stakeholders and business drivers. To put it bluntly, politics, personality and culture will create biases in this data.

3.3. Data cleansing and transformation

Data analysis consists of data collection, data reduction and data visualisation (Bazeley, 2013). A key stage in data warehousing projects is the 'cleansing and transformation' stage of the ETL (extract transform and load) process (Kimball, 1996; Inmon 2005). ETL involves stripping data of 'bad data' by dealing with missing values, outliers or other data quality issues see Won et al's paper on 'dirty data' for list of numerous possible data quality issues (Won et al, 2003). Data may be 'coded', or aggregated (so losing the detail) or 'transformed' to matching units (American/UK date or degree Celsius/Fahrenheit). Data errors can even be introduced at this stage for example a crime is bracketed into a 'crime_type' such as 'robbery' or violent crime', is the same crime counted twice if it was both? Is it logged as two crimes - a violent crime and a robbery? These decisions are decided by experts and stakeholders and are as a result of processes procedures and legalities and also sometimes some strange reasoning - but they can become invisible in the data. The intention is to provide the data in the same format so it can be compared. Ideally the 'metadata' about the data will be recorded and stored – i.e. where the data has come from, the date the data was obtained, the filters the data has been through etc. In many cases this metadata will form part of the data analysis. It is vital to 'reduce' the data to be useful and provide input to queries and reports. Kimball and Inman's arguments become very real in this phase of the project the decisions made providing potential for less detailed and more bias results.

In becoming 'Big Data' the data and system experts can be lost and new assumptions may made about the data which compromise the validity.

3.4. Data analysis

Data analysis techniques should suit the research questions (Bazeley, 2013). Commonly considered data mining (for Big Data) are association, classification, sequences and correlation (Marakas, 2003), still any

type of data analysis can be performed on a large amount of data as long as the technique is appropriate for the data set. This argument works both ways – in that any technique can be used on a smaller dataset. However there are data issues that can affect all analysis independent of technique. Sometimes data doesn't work together, it may not have a 'primary key' field such as "employee_id" to link the data or it may be stored by different granularities e.g. weather data being held by year is not useful to link with sales data by month. This is a strength of the data warehouse design – it is recommended to always include time (year, financial year, season, month, week, day, hour, time – as desired) to provide a link to the data. Much Big Data works by linking GIS data gained from cell phone data (Gobble, 2013). However, real problems can occur when data is collected from open source repositories in a genuine hope to link it with other data to investigate.
Another problem with Big Data is testing and validating the results. Still the bigger the data set, the less 'combined' the data and the more complex the code and the harder it is to spot whether the data (and results) retrieved are correct especially to a novice. The algorithms used are less transparent and another source of error. In returning to the Google 'nowcasting' prediction of flu it has been since undermined in its validity based (in the main) on the algorithms used (Hal, 2014).

3.5 Data visualisation

Data visualisation is a technique to process datasets and present them in a way that people can understand implication and interpretation of data analysis easily (Alsufyani et al., 2015). It provides benefits to the businesses since the stakeholders and decision-makers can understand the complexity more easily and thus can make better judgement on their decisions. Data visualisation has been used in industry for decades. For example, financial services use visualisation to present the daily data indexes of stock market exchanges, credit rating and foreign exchange rates, which include the quantity of transactions, volume of trading and daily revenues and losses.

Adding to the issue of data validation – assuming the results are correct. Providing a powerful visualisation can often skew the data further 'artistic license' helpful to produce an aesthetically pleasing visualisation, can often work better on a subset of the data, or with outliers removed or further categorising of the data. Tables, bar charts, line graphs and pie charts may be 'boring' but they do provide a reasonably straight forward, recognised and meaningful way to represent data (Yau, 2011).

3.5.1 Lies and statistics

Lastly the data concerns that is described above represent genuine issues that are difficult to control, spot and prevent. They are suspected human errors either unintentionally or intentionally.
Unintentionally people are generally bad at statistics and decision making (Kahneman, 2011). When given the introduction to a young American man as "Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has little need for order and structure and a passion for detail." Is Steve more likely to be a librarian or a farmer' (Kahneman, 2013, p.7). Most people will say librarian, however in American there are 20 more male farmers to each male librarian so

statistically he is more likely to be a farmer (Kahneman, 2013). Another of Kahneman's thought provoking examples is "A bat and ball cost £1.10; the bat costs a pound more than the ball. How much does the ball cost? " (Kahneman, 2013, p.44) – the answer is 5 pence, but the answer we are drawn to giving is 10 pence. We have subconscious biases – we believe a suspect with 'previous' is more likely to commit a crime than a suspect with no criminal record (Viktor Mayer-Schönberger and Cukier, 2013). Of course computers will provide statistically correct answers to these questions – but will we trust them? Software will not do analysis for you, nor can it think for you (Bazeley, 2013). Theories may change but the reality is the same - consider the theory that once the world was flat.

Data findings are used to add value and validity to arguments. By intentionally employing some of the above 'techniques' or issues the data and findings can be manipulated, massaged, presented in such a way as to give favourable results. Discussion of the dark side of data mining and ethics is not included in this paper they are slowly being realised somewhere behind this tidal wave of discovery and can be readily found elsewhere.

3.5.2 Is it an accessible commodity or a commodity which creates a 'digital divide'?

To be using Big Data well requires resources – physical and human. The fact that so many companies are storing the data but not actively using it suggests they do not have the resource or understanding to analyse the data. Current literature implies that data warehousing/data mining projects are iterative and experimental - perhaps not a 'one off' project that can be costed and resourced (McAfee and Brynjolfsson, 2012). The functionality contained in the data mining whilst having huge potentials for many departments do not belong solidly in one department (Smith, 2007) – making the 'ownership' of the project difficult. The skills that have been traditionally linked with data analysis such as mathematics, statistics and programming - whilst still required – are only associated with one stage of the BI project. A successful BI project will benefit as much from systems analysis and creative thinking skills (Smith, 2007). The cost of resourcing the projects, buying the data and training or recruiting staff in relevant skills is significant. Perhaps only realistic for the larger, profitable companies so creating a greater divide between Companies who can and cannot utilise BI (Crawford, 2011).

If the power of Big Data is to reach the 'man on the street' or even the small businesses it seems likely it would be at an unquestionable, 'black box' – data in, recommendation out level. Whether we would want or trust it is another thing. Some people want to know the future, like predictions. Would it be useful to be given a 'recommended job role' based on any data? a holiday destination? a partner ? We already have all of these systems and in general we treat them with the respect they deserve. We ask – where did they get this information? We happily input the wrong data to get results we prefer. There are people quite happy to put faith is far worse and even disproven prediction techniques. Actual predictions are rarely given –we get 10% chance of rain a 50% chance of living until we are 90. These systems support our intuitions and still we remain in control, British society is one of freedom, learning by mistakes, innovation and creativity. Some people may trust the data as some astrology, but the policy of data the management

will have to prove itself for people to actively allow it to affect their lives.

3.5.3 Predicting the likely trends and the future

Data visualisation can blend with business intelligence to understand the market trend and predict the likely movement based on analysis of historical data and user behaviours. It involves with mathematical modelling that use sophisticated technology based on Cloud Computing and Big Data methods to calculate complex derivatives and statistical analysis, whereby results have been presented in a form of visualisation within seconds. The benefits allow the businesses to make more accurate and rapid decisions to stay competitive, or to maximise short windows of opportunities. As demonstrated by Chang (2014), Business Intelligence as a Service (BIaaS) can predict the likely trend of a certain investment under fulfilling certain conditions. Predicting the future trends can offer benefits for businesses to understand more about their customer behaviours and preferred choices, which can be blended with decision-support systems to provide a more effective combination for businesses. However, the limitations are as follows. Firstly, it requires analysts with strong programming and system administration skills to achieve data visualisation. Secondly, most of the tools available in the market are either expensive or not easy to use. It will take years of substantial development for businesses to get affordable and easy to use tools for visualisation.

4. Conclusions

To conclude the following points are put forward:

- There are different types of projects; ‘Big Data’ projects, projects with a lot of data and smaller data analysis project. These project requirements will be fundamentally different and should be resourced and run accordingly. Consider each question and choose a method as appropriate.
- Where there is data, there will be higher chances of error. Actively distrust the data – question it, look for bias and issues at each stage of the process, look at the data through a different ‘lens’ – Your argument might be different from the original one. An expert team will recognise, document and make transparent the potential issues in validity.
- Data provides an appropriate thinking tool – which can help the user discover more, innovate and move a business forward. Exploit this by exploring theories and theorists.
- Lastly, there is so much data not being used. It is suggested that this has to indicate a general requirement to develop methodologies, standards, architecture, vertical alignment and software solutions to the field. A challenge which would benefit from bringing together skill sets from both industry and academia.

References

Alsufyani, R., Safdari, F., & Chang, V. (2015). Migration of Cloud Services and Deliveries to higher Education. In Emerging Software as a Service and Analytics 2015 Workshop (ESaaSA 2015), in conjunction with CLOSER 2015, Lisbon, PT, 20 - 22 May 2015.

Attaran, H., Hosseinian-Far, A. (2011) A novel technique for object oriented relational database design. IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS). London, UK.

Avanade (2012). [Online]. Available from<http://www.avanade.com/Documents/Research%20and%20Insights/avanade-big-data-executive-summary-2012.pdf />[Accessed 25th March 2014].

British Computer Society (2013) *About BCS*. BCS: The Chartered Institute for IT. [Online]Available from: http://www.bcs.org/category/5651> [Accessed 7th February 2013].

McAffee, A. (2013). Datas biggest challenge convincing people not to trust their judgement. *Harvard Business Review*. [Online] Available from<http://blogs.hbr.org/2013/12/big-datas-biggest-challenge-convincing-people-not-to-trust-their-judgment/>[Accessed 25th March 2014].

Bazeley, P. (2013) *Qualitative data analysis : practical strategies / Pat Bazeley*. London : SAGE, 2013.

Bennett, B. (2013) Big Data rewards. *New zealand management*, 60(3), pp.28-34.

Biesdorf, S., Court, D. and Willmott, P. (2013) Big Data: What's your plan? *McKinsey Quarterly*,(2), pp.40-51.

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and Social problems* (1) pp. 231-242.

Campbell, R et al. (2006) GIRO data quality working paper. [Online] Available from: <www.actuaries.org.uk/system/files/documents/pdf/Francis.pdf> [Accessed 25th March 2014].

Chang, V. (2014). The business intelligence as a service in the cloud. Future Generation Computer Systems, 37, 512-534.

Chui, M., Manyika, J. and Kuiken, S. V. (2014). What executives should know about 'open data'. *McKinsey Quarterly*,(1), pp.102-105.

Cumbley, R. and Church, P. (2013) Is "Big Data" creepy? *Computer Law & Security Review*, 29(5), pp.601-609.

Crawford, K. Boyd, D. (2011) A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. *Oxford Internet Institute*. September 21, 2011. [Online]. Available from:< http://softwarestudies.com/cultural_analytics/Six_Provocations_for_Big_Data.pdf > [Accessed 25th March 2014].

Creswell, J. W. (2013). *Qualitative inquiry & research design : choosing among five approaches* / John W. Creswell. Thousand Oaks, Calif. ; London : SAGE, c2013.

Date, C. J. (2004). *An introduction to database systems / C.J. Date*. London : Addison-Wesley, c2004.

Dugas, A. F. et all (2012). Google Flu trends: correlation with Emergency Department influenza rates and crowding metrics CAD Advanced Access January 8th 2012.

Ehrenberg, R. (2010). Predicting the next deadly manhole explosion. Wired July 7, 2010. [Online]. Available from:< http://www.wired.com/wiredscience/2010/07/manhole-explosions/>/[Accessed 25th March 2014].

FieldLogix (2014). [Online]. Available from:< http://www.fieldtechnologies.com/gps-tracking-systems-installed-in-ups-trucks-driver-efficiency/[Accessed 25th March 2014].

Frei, P et al. Use of mobile phones and Brain tumours: Update of the Danish cohort study. BMJ , 2011, 343. [Online]. Available from:<http://www.bmj.com/content/343/bmj.d6387/>[Accessed 25th March 2014].

Hal, H. (n.d). News: [dn25217] Google Flu Trends gets it wrong three years running. *New Scientist*. p. 24. doi:10.1016/S0262-4079(14)60577-7

Hays, C. L. (2004). What Wal-Mart Knows about customers habits. New York times November 14th 2004. [Online]. Available from:< http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0/>/[Accessed 25th March 2014].

Inmon, W.H. (2005) Building the data warehouse/W.H. Inmon. Indianapolis, IN; Chichester:Wiley 2005.

Gang-Hoon, K. I. M., Trimi, S. and Ji-Hyong, C. (2014) Big-Data Applications in the Government Sector. *Communications of the ACM*, 57(3), pp.78-85.

Gobble, M. M. (2013) Big Data: The Next Big Thing in Innovation. *Research Technology Management*, 56(1), pp.64-66.

Google (2014). [Online]. Available from:<http://www.google.org/flutrends//> [Accessed 25th March 2014].

Grbich, C. (2013) *Qualitative data analysis : an introduction / Carol Grbich.* London : SAGE, 2013. 2nd ed.

Harding, J. (2013) *Qualitative data analysis from start to finish / Jamie Harding.* London : SAGE, 2013.

Hsinchun, C., Chiang, R. H. L. and Storey, V. C. (2012) Business Intelligence And Analytics: From Big Data To Big Impact. *Mis Quarterly*, 36(4), pp.1165-1188.

Kahneman, D. (2013) *Thinking, fast and slow / Daniel Kahneman*. New York : Farrar, Straus and Giroux, 2013. First paperback edition.

Karlgaard, R. (2013) Big Data's Promise Messy, Like Us. *Forbe*s, 191(11), pp.36-36.

Kincheloe, J. (n.d). On to the next level: Continuing the conceptualization of the bricolage, *Qualitative Inquiry*, 11(3), 323-350.

Kim, W., Choi, B., Hong, E., Kim, S., & Lee, D. (n.d). A taxonomy of dirty data. *Data Mining And Knowledge Discovery*, *7*(1), 81-99.

Kimball, R. Ross M (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, Solutions from the Expert. New York. Wiley & Sons, 1996.

Laughlin, L. S. (2014) Buying Into Big Data. *Fortune*, 169(2), p.48.

Marakas, G. M. (2003). *Modern data warehousing, mining, and visualization : core concepts / George M. Marakas*. Upper Saddle River, N.J. : Prentice Hall, c2003.

Marsh, C. and Elliott, J. (2008) *Exploring data : an introduction to data analysis for social scientists.* Cambridge : Polity, 2008.

2nd ed. / Catherine Marsh and Jane Elliott.

Mayer-Thurman, C. C. and Cukier, K. (2013) *Big Data : a revolution that will transform how we live, work and think* / Viktor Mayer-Schönberger and Kenneth Cukier. London : John Murray, 2013.

McAfee, A. and Brynjolfsson, E. (2012) Big Data: The Management Revolution. (cover story). *Harvard Business Review,* 90(10), pp.60-68.

NYTimes (2004). Jessica Santillan. [Online] Available from:<http://topics.nytimes.com/top/reference/timestopics/people/s/jesica_santillan//>[Accessed 25th March 2014].

Ross, J. W., Beath, C. M. and Quaadgras, A. (2013) You May Not Need Big Data After All. *Harvard Business Review*, 91(12), pp.90-98.

ScienceDaily (2013) Big Data, for better or worse: 90% of world's data generated over last two years. May 22nd 2013. [Online] Available from<http://www.sciencedaily.com/releases/2013/05/130522085217.htm/>[Accessed 25th March 2014].

Schultz, B. (2013) Big Data In Big Companies. *Baylor Business Review*, *3*2(1) Fall2013, pp.20-21.

Smith, D. (2007). Data Model Overview. Teradata. [Online] Available from: <www.teradata.com/.../Data-Model-Overview-Modeling-for-the-Enterprise-...>[Accessed 25th March 2014].

Thorpe. J (2012). Data Humans and the New Oil. *Harvard Business Review*. [Online]. Available from:<http://blogs.hbr.org/2012/11/data-humans-and-the-new-oil/?> [Accessed 25th March 2014].

Treiman, D. J. (2009) *Quantitative data analysis : doing social research to test ideas /* Donald J. Treiman. San Francisco, Calif. : Jossey-Bass, c2009.

1st ed.

Vizard, M. (2013) Big Data Experience Is Much Needed--and Wanted. *CIO Insight*, pp.1-1.

Yau, N. (2011). *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics.* Indianapolis, Wiley, 2011.