PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

A COMPREHENSIVE SURVEY ON BIG-DATA RESEARCH AND ITS IMPLICATIONS – WHAT IS REALLY 'NEW' IN BIG DATA? - IT'S COGNITIVE BIG DATA!

Artur Lugmayr
Curtin University, artur.lugmayr@artur-lugmayr.com

Bjoern Stockleben
Univ. of Applied Sciences Magdeburg, Germany, bjoern.stockleben@hsmagdeburg.de

Christoph Scheib
Mathew Mailaparampil, Noora Mesia, Hannu Ranta, EMMi Lab., Tampere, Finland, mailaparampil@gmail.com

Recommended Citation

Lugmayr, Artur; Stockleben, Bjoern; and Scheib, Christoph, "A COMPREHENSIVE SURVEY ON BIG-DATA RESEARCH AND ITS IMPLICATIONS – WHAT IS REALLY 'NEW' IN BIG DATA? - IT'S COGNITIVE BIG DATA!" (2016). *PACIS 2016 Proceedings*. 248.
http://aisel.aisnet.org/pacis2016/248

A COMPREHENSIVE SURVEY ON BIG-DATA RESEARCH AND ITS IMPLICATIONS – WHAT IS REALLY 'NEW' IN BIG DATA? - IT'S COGNITIVE BIG DATA!

Artur Lugmayr, Visualisation and Interactive Media (VisMedia), School of Media and Creative Arts, Curtin University, Perth, Australia, artur.lugmayr@artur-lugmayr.com, www.artur-lugmayr.com

Bjoern Stockleben, Univ. of Applied Sciences Magdeburg, Germany, bjoern.stockleben@hs-magdeburg.de

Christoph Scheib, Mathew Mailaparampil, Noora Mesia, Hannu Ranta, EMMi Lab., Tampere, Finland, scheib_c@yahoo.com, mailaparampil@gmail.com, {noora.mesia|hannu.ranta@tut.fi}

Abstract

What is really 'new' in Big Data? – Big Data seems to be a hype that has been emerging during the past years. But it requires a more thorough discussion beyond the very common 3V (velocity, volume, and variety) approach. We established an expert group to re-discuss the notion of Big Data, identify new characteristics, and re-think what actually really is new in the idea of Big Data by analysing over 100 literature resources. We identified typical baseline scenarios (traffic, business processes, retail, health, and social media) as starting point, from which we explored the notion of Big Data from a different viewpoint. We concluded, that the idea of Big Data is simply not new, as well as we need to re-think our approach towards Big Data. We introduce a fully new way of thinking about Big Data, and coin it as the trend of 'Cognitive Big Data'. The publications introduces a basic framework for our research results. However, this work remains work-in-progress, and we will continue with a refinement of the Cognitive Big Data Framework in one future publication.

Keywords: Big Data, Scenarios, Data Research, Cognitive Big Data, Competitive Advantage, Theory.

1 INTRODUCTION

Recent advances in technology, the wave of the new digital economy, and the digitalization of industries let countless novel methods of technological possibilities, scientific research, social interaction, business intelligence, and data analytics emerge. All these trends together caused an exponential increase of generated data, including its new facets, forms, and processing speeds (Johnson 2012a). Besides creating new opportunities, the trend towards increasing amounts of data, new processes for handling vast amounts of data, and its related technology has been coined as "Big Data" during the recent years. Similar to the term Web2.0, the term Big Data tries to summarize a number of related (and non-related) trends and is a rather blurred term of discourse. Thus giving one accurate definition is a rather vain exercise, because of the issues related to society and technology (Ward and Barker 2013a).

In the public eye, Big Data seems to be a rather disruptive innovation, having given birth to wide range of new technologies, process, and possibilities. However, a wide range of technologies represent rather incremental innovations rather than true breakthroughs and novelties. Thus the notion of Big Data can be rather seen as a term for a powerful tool for knowledge creation, processing of huge amounts of data, and expressing hopes and fears of the new potentials. The goal of this research work, is to shade new light on the notion of 'Big Data', and it represents a discussion or discourse for its potential, and its impact. It shall shade a few more thoughts into its trends, and surveys the current state of the art, industrial applications, social motivations, and goes as far to describe the cognition of Big Data.

Thus what is really new in 'Big Data'? – Within the scope of this study we try to shade a different light on Big Data research through the thorough investigation of literature resources and on the example of three baseline scenarios for typical Big Data applications: real-time traffic data, business process management, health care, retail, and social media. Our main goal was to explore the cognitive elements of Big Data, and tie these to existing business applications.

Please also note, that due to the page limitations of this paper, the appendix (the evaluation of particular projects) can only be found online on: www.artur-lugmayr.org/Publications.

2 METHOD AND APPROACH

The approach of this paper is theoretical, thus we identified the most common fields of Big Data application in literature by conducting a thorough literature review based on over 60 resources. As follow up, we created several baseline scenarios where typical Big Data has been applied. In focus group discussions we refined these by cross-examining literature with theoretical approaches in data and computer science. These have been used to apply epistemological theories and led to the creation of an alternative view on Big Data, and a framework for its re-definition.

2.1 Baseline Scenarios for the Scope of this Study

To allow a different kind of discussion around the theme Big Data, we approached our study by identifying four scenarios in different application areas. These acted as baseline for this study, and we utilized these to identify features, characteristics, and approaches for a new definition of Big Data:

- **Real-Time Traffic Data and Decision Making** (based on (Intel 2013)): real-time features, and the opportunities that emerge from traffic monitoring for Big Data research have let us to adopt this as one of the baseline scenarios. One study that we have been considering was the traffic management system of the city Zhejiang, which attempts to avoid traffic jams and accidents. Through a centralized data management system, the city was able to monitor, provide a highly efficient traffic routing system, and manage local traffic by applying Hadoop technology in combination with Intel's Xeon architecture (Hadoop n.d.). The system can recognize register plats form a collection of over 2.6 billion records in less than a second, and is based on more than a terabyte of collected data per month;

- **Business Process Management** (based on (Davenport and Dyché 2013)): the second identified base scenario related to enable industry to increase their performance and the efficiency of processes through Big Data. We selected UPS as scenario, as UPS tracks package movements and transactions in resulting to over 16 petabytes of storage data by serving over 16 million packages to customers, and tracking over 8 million of these per day. Real-time vehicle tracking to monitor daily performance, online maps, and route optimizations allow significant savings for UPS;
- **Health Care Big Data** (based on (Davenport and Dyché 2013)): a second scenario has been described in (Davenport and Dyché 2013): utilization of Big Data as one of the key performance indicators of United Healthcare, to provide additional consumer satisfaction. For this purpose call center data is collected, and analysed utilizing Hadoop and NoSQL. Customer satisfaction evaluation is based on the recognition of emotion in voice;
- **Big Data as Cross-Domain Analysis in Retail or for Epidemics Prediction**: this baseline scenario relates to the correlation of different data-source as e.g. weather data and sales. One study has been conducted by Walmart and Kellogg's Pop-Tarts (Mayer-Schönberger and Cukier 2013), which attempted to correlate weather history with sales transactions prior to hurricanes. A similar cross-domain analysis has been conducted by Google Flu Trends, analysing influenza alike illnesses to predict the spreading of influence ahead of governmental instructions as e.g. the US Center of Disease Control (see e.g. (Dugas et al. 2012) and (Ginsberg et al. 2009));
- **Social Media Analysis:** social media analysis has been well researched, and can be considered as a prime example in Big Data research. Twitter network analysis, integration of social media data into business processes, and the analysis of social media as part of customer relationship management systems are one of the prime concerns in this use-case.

These non-exhaustive scenarios where the starting point for our exploration, and the discussions as part of the expert groups. We analysed these scenarios, and attempt to shade a fully new way on how to think about Big Data research. However, this work is still research-in-progress, and we aim at a more thorough analysis after the research project is completed.

2.2 Literature and Related Works

To accomplish this study, we have examined over 100 literature sources, where the most significant resources are enlisted in the appendix of this publication. This table classifies literature resources, identifies the key-aspects, and research contribution of each single publication. Please note, due to the restrictions of the length of the paper, the list of investigated publications can be downloaded from: www.artur-lugmayr.org/Publications.

3 REVISING THE VISION OF 'BIG DATA' – AN EPISTOMOLOGICAL APPROACH

Many expectations and promises, what changes Big Data will bring to businesses and private lives has been discussed in scientific research, as e.g. in (Brynjolfsson et al. 2011). Thus, is 'Big Data' indeed a paradigm shift that we are observing, or simply a notion enriching well-known methods? Within the scope of this section, we provide a systems theoretical and epistemological perspective to examine the disruptiveness of Big Data, and develop several criteria for the categorization of its applications.

3.1 Knowledge and Wisdom Processing rather than Velocity, Volume, Variety (3V)

The 3V (eventually a 4th V – Veracity) discussion around the theme of Big Data is rather short sighted. Computational power increases, methods for processing data will become more sophisticating, as well as data science will let data structures emerge to process a wide variety of data sources. Thus, the key issue in Big Data Research needs to be centred also outside this very limited viewpoint. Big Data processing is about finding patterns, knowledge patterns, and eventually wisdom. In accordance with the discussion addressed in (Floridi 2012), the epistemological challenge of Big Data is about finding patterns in data sources. The likelihood of finding patterns and connections between data increases with increasing amounts of data – but the size of data volume is not the relevant indicator to increase the possibility to identify patterns in data sets. In opposite, patterns can be identified in data sub-sets or large data-sets (Floridi 2012). To identify more advanced patterns as e.g. knowledge or wisdom, it's also an important issue to provide possibilities to identify and decide which parts of the data is of relevancy, and which can be neglected (Bollier and Firestone 2010) . We conclude, that the relevant indicator in Big Data research requires to be tied to the capability to identify knowledge and wisdom patterns, independent of data volume, variety, or velocity. The indicator for the analysis is based on complexity management, adapting available computational power, velocity, and data variety. Thus, the identification of a meaningful subset and data granularity upon which computational methods are applied to gain insights into knowledge and eventually wisdom patterns shall act as indicator in Big Data research.

3.2 Completeness of Observed Data

In statistical analysis the problem of bias in the sample is addressed by randomization and other methods. In Big Data, data collection is not restricted to samples, but basically to collecting as much potentially related data as possible. This raises questions about the representability of the data collected. It appears safe to assume two basic cases: The observed data can be either complete or incomplete with regard to the research question to be answered. Complete observation is usually only possible with a narrow research question that can be answered relying on internal data like the complete sales records covering all customers of a company. In the case of complete observation we can assume bias-free data.

In practically all cases that rely on external data, we have to assume incomplete observation and thus a bias. A prime example is social media, where only the respective companies have access to the full data sets and decide on how much data is exposed to the public for free or paid access. Twitter used offers to offer different plans to access 1, 10 or 100% access respectively (González-Bailón et al. 2012), with no certainty about how the data is selected from the main unit. Yet even with the open access APIs, there may be a bias depending upon which particular API is used (González-Bailón et al. 2012). The challenge is to take this bias into account when interpreting the results of data analysis and, if possible, assess its impact on the results.

It should be noted that the discrimination between a complete and an incomplete observation is far from being definite and depends a lot on the question that shall be answered. In the Google Flu Trends example, we can be sure that Google has access to the entirety of flu-related queries, yet the definition of what is a flu-related query and what not is not trivial. As well this should not be mistaken with a full coverage of all persons searching for flu-related medical advice, as the web alone has countless further sources and there are countless more sources of information beyond the web.

3.3 Meaning and Quality of Data

Next to the obvious implications of the idea of data completeness in Big Data, it deems useful to revisit further established categories of data quality. For this purpose we use the seminal data quality framework introduced by Wang and Strong in 1996 (Wang and Strong 1996). This model was developed in 1996, so well before the term Big Data was coined. The framework has been developed to assess data quality from the perspective of data customers, so it rather applies to processed data, not data acquired through data-mining. Yet this might indicate a major shift already, that companies start actively mining their own data instead of solely relying on external providers.

Intrinsic DQ	**Relevance in Big Data**
Believability	In Big Data, this is a difficult criteria, as it implies a “plausibility” check by a human actor. As many Big Data applications just work with correlations and people expect to find unexpected correlations in the first place, believability cannot be a criterion to assess DQ in Big Data.
Accuracy	This is an ambivalent criterion in Big Data. While of course more accuracy is always better, there are supposedly two ways to increase accuracy in Big Data: First, the accuracy of single data items can be increased and second, the number of total data items.
Objectivity	Data objectivity can be assessed first by considering the bias of the sample. A huge sample does not automatically mean more objectivity, rather it is about the decision about which part of the data to look at and which to ignore. Second, objectivity is about what is actually measured and what not, i.e. whether all information that could possibly confirm or contradict a research question is actually included in data item.
Reputation	Reputation refers to the data source. As Big Data relies a lot on data scraping and we are dealing with large amounts of unstructured data, this criterion seems to fit more when talking of data already processed by intermediaries, i.e. if you actually buy data.
Contextual DQ	
Value-added	Traditionally, data would be acquired with respect to a certain problem. In contrast, the hope in Big Data lies partly in the discovery of new problems. Thus it is hard to tell, whether certain data will add value or not.
Relevancy	Also relevancy can be defined only with respect to a pre-defined problem.
Timeliness	Expectation towards Big Data is real-time in many cases.
Completeness	Completeness of a data set can only be determined in relation to a research question, yet Big Data approaches promise that the questions would emerge from examining all available data. As discussed earlier in this paper, completeness in Big Data means all the data there is in the domain researched and completeness is the exception, not the rule.
Appropriate Amount of Data	The demand for appropriateness of data volume tries to balance the effort of data analysis with the expected gain. In Big Data, more data always holds the promise of more and unexpected gain. Also technology and processes for analysis of complex data sets have vastly increased, so it could be argued that the idea of Big Data renders this criterion void.
Representational DQ	
Interpretability	If we follow the turn from causality towards correlation, interpretability is not an issue of the data set as such. It rather applies to the results of the analytic algorithms, i.e. to refined and abstracted data intended for human decision making.
Ease of understanding	As interpretability, also “understanding” is a human category, not applicable to data that is processed automatically. This also has to be applied to the interpretational framework provided by Big Data analysis tools.
Representational consistency	The use of unstructured data is often cited as a key element of Big Data. Reaching representational consistency is thus a requirement towards the data processing algorithms, not to the source data itself.
Concise representation	Data sets in Big Data do not have to be human readable, so again, this is a requirement towards the output of a Big Data processing system, not the source data.
Accessibility DQ	
Accessibility	The data mining technologies developed in recent years have made data accessible that would not even have been considered a data source a decade ago. Thus accessibility is another criterion that has to be considered in relation to the state-of-the-art in data processing technology. In Big Data, we have open and closed data sources, as well as dynamic and static data sources. While e.g. certain census data may be open, it is not updated frequently and quickly growing old and possibly unusable. Social media data is available in real-time, but filtered through the different access policies of social media services. The ideal accessible Big Data data source is: -Updated in real-time, -Accessible in real-time, -Exposes the complete available data for a certain request, -Open and free
Access Security	Access security refers to aspects like exclusivity of access to certain data as a competitive advantage, as well as the technical security of data storage and access. Both aspects play an important role in Big Data. On one hand, many sources of unstructured data are openly available on the internet and given enough processing power, any traceable transaction on the internet can be turned into data. On the other hand, this is what everybody is doing, so a real competitive advantage can only arise from exclusive access to certain data sources. The primary gate keepers at the moment are large social media companies, which creates a new kind of digital divide into data-rich and data-poor [59]. The second aspect of technical security of Big Data storage is important not only to keep competition from illegal data retrieval, but because trust of

	customers is a particularly valuable resource in this domain. Security breaches involving customer data seriously affect the readiness to share private data with companies.

Table 1. Quality of Big Data (based and extended from (Wang and Strong 1996)).

Thus the broader picture of this way of discourse is simply that many criteria address the idea that data should be human-readable and human-interpretable. In a way, back in the 1996s, data was already considered as an answer, but not as a resource to solve the problem. This is expressed by the demand that data should carry distinguishable meaning. Therefore we conclude, that in Big Data research the notion of meaning is a category that has been only introduced after data analysis, or manifested by far too less in data research till date. Thus, meaning is not inherited inside data, but is emerging through the application of processing algorithms, identification of knowledge/wisdom, or patterns as based on data.

3.4 Correlation, Causation, and Predictability

Chris Anderson addresses the primacy of correlation over causation as a bit more prosaic as "the end of theory", as we could test our hypotheses directly “in the wild”. We possibly could derive future events from past patterns without having a theory of why they should happen. The high probability of spurious correlations (Bollier and Firestone 2010) are frequently mentioned as an epistemological problem of Big Data.

Yet how do we decide whether a correlation is spurious? Herbert Simon (Simon 1954): Spurious Correlations explains that we need to rely on two types of a priori assumptions, logical assumptions such as preceding events cannot be caused by later events and the assumption that other environmental variables do not interfere with the correlation to be tested. Although Simon argues that these assumptions are a priori as they are not founded in statistics, but are otherwise empirical and by far not arbitrary. Still it remains that we cannot judge the causality of a correlation without relying on prior empirical experiences, which limits us in both analysis and decision-making.

In Big Data it might make sense to differentiate correlations not into spurious and genuine types by a proof of causality, but rather on their viability for a certain purpose, using the term of viability as defined by Glasersfeld (Glasersfeld 1998). Thus we would discriminate solely viable and non-viable correlations. The downside of these terms is that there is no absolute value (truth) in them, but it can only be decided in the context of a purpose, whether a correlation is viable. Such a context may be the desire to fit the observation made into a larger model or a goal to be reached.

We shall give an example to illustrate this: A classic example is the legend that somebody pointed out a correlation between the wages of Presbyterian ministers in Massachusetts and the rum prices in Havana – simply people ‘lie based on statistics. While there cannot be any causal explanation by logic or empirical experience, the question is not whether this correlation is spurious. The question is whether we could use this correlation to forecast raises in rum prices or increasing demand for rum in Massachusetts and build a successful trading business upon this information. If so, we do have a viable correlation. Both rum prices and minister wages are triggered by general developments in world economy, so a correlation viable for this purpose can be doubted in this case.

However viability does tell anything whether the successful solution is optimal and whether there are alternatives. Taking a look again at the Walmart example, obviously Pop-Tarts sales rose when they were placed at the front of the store. So obviously Big Data analysis has helped to increase revenues here. Yet it can be assumed that anything positioned well visibly and in large supply will sell better. Had they looked for causal explanations - maybe people like to buy fancy, but practical food to cheer up their minds during hurricanes - they would have had much broader options to increase sales of possibly more than one product.

A more general view on this example suggests the main difference between acting upon correlation versus causality: In the first case we just amplify a phenomenon inherent to the system, in the second case causal information empowers us to change the system. In other words, correlations can be used to improve the performance of existing systems, while true innovation can only happen if we make causal assumptions. Revisiting Anderson's statement, it seems that the strength of Big Data is not to render the idea of causality void, but to point us to search for causality in places we never considered in the first place. There exist many other examples, as e.g. the promise of Big Data applications delivering more accurate and earlier predictions than usual statistical means. An often cited example is the case of Google Flu Trends already described earlier in this paper.

3.5 Automated Decision Making vs. Data Observability

While Big Data technologies have vastly increased our means of processing vast amounts of data, human attention remains constant. In Big Data, automation through algorithms help us to decide to focus our attention, i.e. to decide what to look at, implicitly ignoring anything else. What we see as users is always a representation of aggregated data on a distinct abstraction level. This level can be anything from statistical figures describing the data-set as a whole to a single data unit. Yet we can focus our attention only to the whole data-set on a high abstraction level or to very few single data units on a concrete level. Therefore, the algorithms used and their configuration restrict a priori what we may see or may not see in a data-set. For example, pre-set thresholds determine what will be surfaced as a pattern and what not. Correlations too weak for the threshold will not come to our attention, even though they might carry valuable hints for certain applications. The algorithm can only identify patterns, but does not reason about their potential meaning for applications. The higher the velocity of a Big Data application is, the more it has to rely on algorithms, first for the decision on what to data to observe, second for decisions on how to react upon the observation.

	Complete Observation	**Incomplete Observation**
Automated Decision Making	Traffic Observations	Algorithmic Financial Trading
Human/Fuzzy Decision Making	Visualisation of corporate data	Real-Time Feedback to Newspaper Editors

Table 2. Automated Approaches towards Defining Big Data: Illustration of the different automated (algorithmic) approaches towards Big Data, including a few practical examples.

From this point of view, we can distinguish between the following 4 possible features of Big Data applications:

- **Automated-complete:** Many real-time applications rely exclusively on automated (algorithmic) decision-making, with human users action as supervisors only. When acting on complete data, algorithms can be considered reliable and predictable, so that once the system is in place and configured, it could basically run unsupervised. A simple example for this kind of applications are automated recommendation systems used by online retailers, but also advanced traffic control applications can be counted into this category, depending on the way traffic flow is measured.

- **Automated-incomplete:** Automated decision-making that is based on incomplete data is prone to unpredictable behaviour. This is for example the case in real-time finance transactions, where all actors have different incomplete data about the market. These systems need constant human supervision and adjustments.

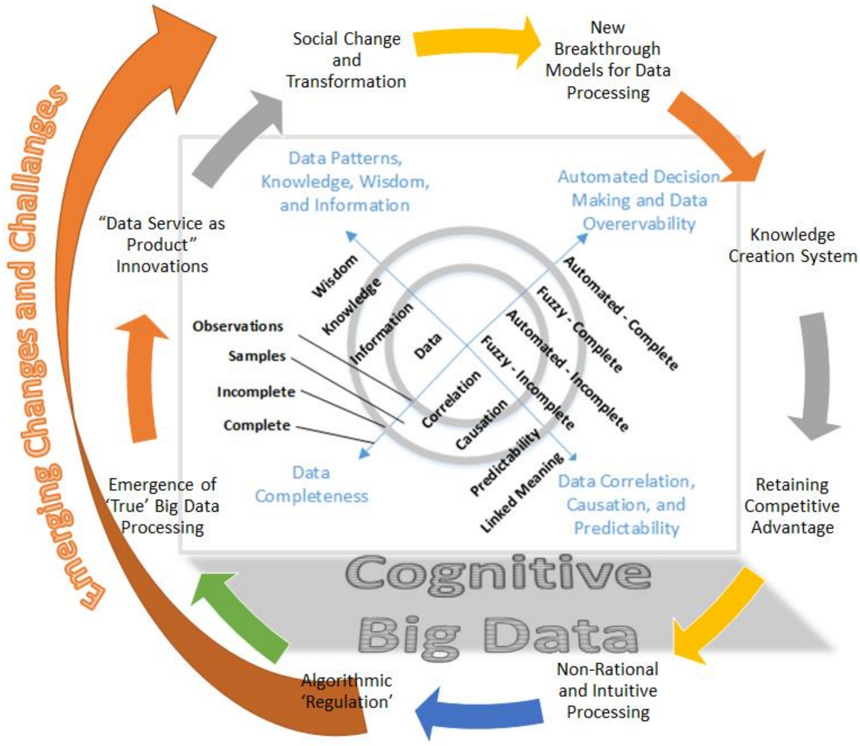
- **Human-complete:** These kind of applications mostly applies to the visualisation of organisation-internal data for purposes of strategic planning or operative controlling. They are informative and demand for suitable representation of the analysed data.

- **Human-incomplete:** These kind of Big Data applications do inform decisions of human users. They are applied in areas where no meaningful automated decisions can be taken upon incomplete data.

The prime challenge is to find data representations that optimally support decision-making (Lurie and Mason 2007). An example are social media monitoring tools used to optimise the reach of media content.

4 DISCUSSION

We summarized our view on Big Data in diagram with four paired dimensions. On the one axis we are look at the relation between *Data Completeness* and *Decision Making*, on the other we see the relation between *Rule Induction* and *Knowledge Levels*. In comparison with the prevailing 3V model of big data we think that our model is a much more suitable foundation to discuss both epistemological and ethics implications of big data. It might as well help to shed a light on the real technological challenges of big data to be solved. This is necessary because the 3V model addresses scalability challenges only. If the amount of data grows, new and more powerful technology is needed to process it. This leads to a perpetual race, but can never be solved for once and for all. This makes the 3V model a perfect marketing tool, but does not fit as a theoretical model for future scientific research on big data. The model is illustrated in Figure 1, and includes the Cognitive Big Data Model in it's core, and the associated emergent changes and challenges.



Emerging Changes and Challanges
Social Change and Transformation
New Breakthrough Models for Data Processing
Knowledge Creation System
Retaining Competitive Advantage
Non-Rational and Intuitive Processing
Algorithmic 'Regulation'
Emergence of 'True' Big Data Processing
"Data Service as Product" Innovations
Data Patterns, Knowledge, Wisdom, and Information
Automated Decision Making and Data Overervability
Data Correlation, Causation, and Predictability
Data Completeness
Wisdom
Knowledge
Information
Data
Observations
Samples
Incomplete
Complete
Correlation
Causation
Predictability
Linked Meaning
Automated - Complete
Fuzzy - Complete
Automated - Incomplete
Fuzzy - Incomplete
Cognitive Big Data

Figure 3. Cognitive Big Data: Illustration of our Framework for thinking about Big Data – Cognitive Big Data is the new direction in research and many emergent changes and challenges on various levels

4.1 Cognitive Big Data Framework

We have been identifying a set of challenges, which are tied to the Cognitive Big Data model:

- On the *Data Completeness* dimension, we see that it is highly important to better understand what is not the obvious from the data. With incomplete data from possibly a multitude of sources we cannot apply classic statistics as we know it, as even one half of the main unit bears a high degree of uncertainty if we cannot describe what the other half might be;

- Related to that, on the *Decision Making* dimension we apply algorithms displaying unpredictable behaviour when the data base just changes slightly. How can we know where our algorithms have their *tipping points* (Gladwell 2006) and how can we create more resilient algorithms? While the focus so far has been on creating algorithms for data analysis, we might have to create algorithms that analyse such algorithms, especially when we are thinking about self-learning algorithms which evolve based on inducted rules. This may lead from the open data movement to an open algorithms movement.

- The classic model of *Knowledge Levels* might become obsolete. Humans have the desire to understand how things work and only if they can find a plausible explanation, they accept a correlation. This in turn means they are more likely to ignore a correlation they cannot explain. Algorithms do not feel this desire; they do not care if a correlation is spurious in human eyes or not. They simply decide based on the viability: If a rule inducted from an identified correlation leads to a successful result, the rule will be considered viable. In a way, this is a very pure ontological viewpoint, considering just the facts, ignoring any attempt to explain. This allows a very fast evolvement of models about the world (or a part of it that a particular big data application considers). However, these models will always remain less complex than the real world and they will always lag behind reality, as big data systems can only measure what already happened. It remains to be researched what the future role of humans is in the techno-social eco-system of big data will be.

- Under the perspective of *Rule Induction* this means that explaining causality is not anymore a requirement to achieve predictability.

4.2 Emergent Challenges of Cognitive Big Data

Within the scope of this section, we collected our implications of emergent changes and required considerations in discussing Cognitive Big Data. These are briefly described in the following:

- **Big Data requires New Models to be still 'New' in Decades to come:** large volumes of data, high velocity of data processing, as well as a wide variety of data is existing since the beginning of computation. New emerging technologies, as e.g. new computational models, new computer architectures, algorithm research will solve many issues related to these bottlenecks. Thus, if Big Data should still remain a trend that is 'new', it requires a re-thinking about the approach. Figure 1 illustrates the key components of our framework, which will remain valid if these new technologies are emerging;

- **Big Data Enables Competitive Advantage and Leads to Innovative New Business Areas:** From a business perspective, Big Data – independently if we define it as discussed in this paper, or based on common publications - is a main tool in catalysing and supporting business processes. It's doubtful, that the increasing amount of information will change how businesses will make use of the new possibilities for business intelligence analysis and to support business processes (Rogers 2011). It is obvious, that any new innovation will lead to new businesses activities focusing collecting and interpreting complex information from corporate internal and corporate external sources. It's a growing area, which will enable many new business niches for newcomers (Johnson 2012b);

- **Social Impact of Big Data as Knolwedge Creation Eco-System:** Big Data is not solely a technological phenomonon, Big Data is a cultural, scholarly, and technological phenomenon creating knowledge eco-systems, and how we think about knowledge itself (Boyd and Crawford 2012). The research field of Digital Humanities i.e. investigates human knowledge, cultural aspects, as well as it utlizes latest computational method to make knowledge in humanities available

(Lugmayr and Teras 2015). Privacy, and the digital footprint we are leaving behind every single day by using computers and smartphones (Michael and Miller 2013) might render all the positive aspects of recomender systems, personalization techniques, and workforce productivity monitoring towards a 'big brother' system where we are constantly monitored. Other issues, such as ownership of data, and companies offering data sets to researchers (Boyd and Crawford 2012) are another issue impacting our social life and increases. Nevertheless, examples such as smart cities (Calabrese 2011) or open data archieves by governmental organisations illustrate the benefit, as well as the backdraft of these developments;

- **Big Data is a Hype, and Expectations High – 'Real' Big Data Processing will Require New Technologies:** Big Data research is currently peaking it's plateou of expections, when illustrated on a hype cycle. Viewing Big Data from this perspective, many new technologies will need to emerge before promises will be able to be kept. Thus a new way of thinking, as we discussed within the scope of this paper, is required, that Big Data does not become a disillusion, and will shift towards a stable position on it's plateu of productivity. Models and frameworks like ours, will ensure, that Big Data will be shifting towards maturity. Viewing Big Data on the basis of Porter's framework of Five Forces (QuickMBA n.d.), Big Data will be supporting decission making in business intelligent systems accross business areas and provide competitive advantage. Currently investments, new processes required to be established 'Big Data' inside businesses, and it's limitations are still high. Nevertheless, Big Data will increase competitive advantage;

- **Towards 'Algorithmic' Regulation and Non-Rationality:** in financial industries, automated trading algorithms are today's stock brockers. This example illustrates the complexity of automatic and algorithmic data processing. As sugested by e.g. Tim O'Reilly – who predicts the dawn of 'algorithmic regulation' – it might be required to introduce regulations that limit calculations in real-time. In financial industries this trend is already taking place today – it might be a matter of time, until real-time processing in the context of business applications is regulated as well. One other example is illustrated in (*The Cybernetic Theory of Decision: New Dimensions of Political Analysis* 2002), who introduced systems theory into political thinking. He argues, that rational assumptions in politics lead to random, contextless, and systematic decisions, where a non-rational way of problem solving would be for the good of the populations and other participants.

With this paper we wanted to contribute with a different viewpoint towards the way of thinking about Big Data, and introduced the idea of Cognitive Big Data. This research work is still in process, and we will contribute with a deeper discussion in a follow-up publication.

Please also note, that due to the page limitations of this paper, the appendix (the evaluation of particular projects) can only be found online on: www.artur-lugmayr.org/Publications.

References

Aissi, S., P. Malu, and K. Srinivasan. 2012. E-Business Process Modeling: The Next Big Step. Computer.

Anderson, C. 2008. The end of theory. Wired magazine 16.

Bollier, D., and C. M. Firestone. 2010. The promise and peril of big data. . Aspen Institute, Communications and Society Program Washington, DC, USA.

Boyd, D., and K. Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society 15:662–679.

Brankovic, S. (n.d.). Big Data Analytics And Its Epistemological Relevance.

Brynjolfsson, E., L. Hitt, and H. Kim. 2011. Strength in numbers: how does data-driven decisionmaking affect firm performance? Available at SSRN 1819486.

Bughin, J., M. Chui, and J. Manyika. 2010. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. McKinsey Quarterly.

Buytendijk, F. 2013. The Philosophy of Postmodern Business Intelligence. Business Intelligence Journal 18:51–55.

Calabrese, F. 2011. Smart Cities – How Can Data Mining and Optimization Shape Future Cities? IBM Research and Development.

Chen, H., R. H. L. Chiang, and V. C. Storey. 2012. Business Intelligence and Analytics: From Big Data To Big Impact. MIS Quarterly.

Courtney, M. 2012. THE LARGING-UP OF BIG DATA. Engineering & Technology (17509637) 7:72–75.

Davenport, T. H., and J. Dyché. 2013. Big Data in Big Companies. International Institute for Analytics. . International Institute For Analytics.

Dugas, A. F., Y.-H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman. 2012. Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. Clinical Infectious Diseases 54:463–469.

Floridi, L. 2012. Big data and their epistemological challenge. Philosophy & Technology:1–3.

Fox, B. 2012. Thought Leaders: Business Intelligence. Using big data for big impact. Health Management Technology 33:32.

Fox, S., and T. Do. 2013. Getting real about Big Data: applying critical realism to analyse Big Data hype. International Journal of Managing Projects in Business 6:739–760.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457:1012–1014.

Gladwell, M. 2006. The tipping point: How little things can make a big difference. . Little, Brown.

Glasersfeld, E. 1998. Konstruktion der Wirklichkeit und des Begriffs der Objektivität. Pages 9–39 Einführung in den Konstruktivismus. . Piper, München.

Gobble, M. M. 2013. Big Data: The Next Big Thing in Innovation. Research Technology Management 56:64–66.

González-Bailón, S., N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. 2012. Assessing the bias in communication networks sampled from twitter. arXiv preprint arXiv:1212.1684.

Gopalkrishnan, V., D. Steier, H. Lewis, and J. Guszcza. 2012. Big Data, Big Business: Bridging the Gap. Pages 7–11 Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. . ACM, New York, NY, USA.

Gorman, S. P. 2013. The danger of a big data episteme and the need to evolve geographic information systems. Dialogues in Human Geography 3:285–291.

Graham, M. 2010. Neogeography and the palimpsests of place: Web 2.0 and the construction of a virtual earth. Tijdschrift voor economische en sociale geografie 101:422–436.

Hadoop, A. (n.d.). .

Hsinchun, C., R. H. L. Chiang, and V. C. Storey. 2012. BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. MIS Quarterly 36:1165–1188.

HUFF, A. 2013. BIG DATA II: BUSINESS INTELLIGENCE. Commercial Carrier Journal 170:53–58.

IBM. 2013. Big Data Helps City of Dublin Improve its Public Bus Transportation Network and Reduce Congestion. . IBM.

Intel. 2013. Improving traffic management with big data analytics. . Intel Corporation.

Johnson, J. E. 2012a. BIG DATA + BIG ANALYTICS = BIG OPPORTUNITY. Financial Executive 28:50–53.

Johnson, J. E. 2012b. BIG DATA + BIG ANALYTICS = BIG OPPORTUNITY. Financial Executive 28:50–53.

LaValle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Krushwitz. 2011. Big Data, Analytics and the Path From Insights to Value. MITSloan Management Review.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science 343:1203–1205.

Leng, Y., and L. Zhao. 2011. Novel Design of Intelligent Internet-of-Vehicles Management System Based on Cloud-Computing and Internet-of-things. International Conference on Electronic & Mechanical Engineering and Information Technology.

Lin, T. C. 2013. The New Investor. UCLA Law Review 60.

Lugmayr, A., and M. Teras. 2015. Immersive Interactive Technologies in Digital Humanities: A Review and Basic Concepts. Pages 31–36 Proceedings of the 3rd International Workshop on Immersive Media Experiences. . ACM, Brisbane, Australia.

Lurie, N. H., and C. H. Mason. 2007. Visual representation: Implications for decision making. Journal of Marketing 71:160–177.

Mayer-Schönberger, V., and K. Cukier. 2013. Big Data. . Houghton Mifflin Harcourt.

McAfee, A., and E. Brynjolfsson. 2012. Big Data: The Management Revolution. Harvard Business Review.

Michael, K., and K. W. Miller. 2013. Big Data: New Opportunities and New Challenges. IEEE Computer Society.

Mukherjee, S., I. Pan, and K. N. Dey. 2009. Traffic Organization by Utilization of Resources Through Grid Computing Concept. World Congress on Nature & Biologically Inspired Computing.

Naphade, M., G. Banavar, C. Harrison, J. Paraszczak, and R. Morris. 2011. Smarter Cities and Their Innovation Challenges. Computer June:32–39.

Ortiz, J. R., H. Zhou, D. K. Shay, K. M. Neuzil, A. L. Fowlkes, and C. H. Goss. 2011. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. PloS one 6:e18687.

Power, D. J. 2008. Understanding data-driven decision support systems. Information Systems Management 25:149–154.

Power, D. J., and R. Sharda. 2007. Model-driven decision support systems: Concepts and research directions. Decision Support Systems 43:1044–1061.

Provost, F., and T. Fawcett. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data 1:51–59.

QuickMBA. (n.d.). Porter's Five Forces - A Model for Industry Analysis.

Rajpurohit, A. 2013. Big Data for Business Managers - Bridging the gap between Potential and Value. IEE International Conference on Big Data.

Rogers, S. 2011. BIG DATA is Scaling BI and Analytics. Information Management (1521-2912) 21:14–18.

Roopa, T., A. N. Iyer, and S. Rangaswamy. 2013. CroTIS - Crowdsourcing based Traffic Information System. IEEE International Congress on Big Data.

Schaefer, S., C. Harrison, N. Lamba, and V. Srikanth. 2011. Smarter Cities Series: Understanding the IBM Approach to Traffic Management. . IBM.

Shi, W., J. Wu, S. Zhou, L. Zhang, Z. Tang, Y. Yin, L. Kuang, and Z. Wu. 2009. Variable Message Sign and Dynamic Regional Traffic Guidance. IEEE INTELLIGENT TRANSPORTATION SYSTEMS MAGAZINE Fall:15–21.

Simon, H. A. 1954. Spurious Correlation: A Causal Interpretation*. Journal of the American Statistical Association 49:467–479.

STEVENS-HUFFMAN, L. 2013. Profit from Big Data. Smart Business Philadelphia 8:12–15.

Stockleben, B., and A. Lugmayr. 2013. Issues and Topics to Consider for Information Management Research in eMedia Industries. International Series on Information Systems and Creative eMedia No 2 (2013): Proceedings of the 6th International Workshop on Semantic Ambient Media Experience (SAME).

The Cybernetic Theory of Decision: New Dimensions of Political Analysis. 2002. . Princeton University Press.

The Meme Hustler | Evgeny Morozov | The Baffler. (n.d.). .

Tien, J. M. 2013. Big data: Unleashing information. Journal of Systems Science and Systems Engineering 22:127–151.

Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. arXiv:1403.7400 [physics].

Wang, R. Y., and D. M. Strong. 1996. Beyond Accuracy: What data quality means to data customers. Journal of Management Information Systems 12:5–33.

Wang, W. Q., X. Zhang, J. Zhang, and H. B. Lim. 2012. Smart Traffic Cloud: An Infrastructure for Traffic Applications. IEEE 18th International Conference on Parallel and Distributed Systems.

Ward, J. S., and A. Barker. 2013a. Undefined By Data: A Survey of Big Data Definitions. arXiv preprint arXiv:1309.5821.

Ward, J. S., and A. Barker. 2013b. Undefined By Data: A Survey of Big Data Definitions. arXiv preprint arXiv:1309.5821.

Watson, H. J., and O. Marjanovic. 2013. Big Data: The Fourth Data Management Generation. Business Intelligence Journal 18:4–8.

Wong, A. K., and Y. Wang. 2003. Pattern discovery: a data driven approach to decision support. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 33:114–124.

Xiong, G., K. Wang, F. Zhu, C. Cheng, X. An, and Z. Xie. 2010. Parallel Traffic Management for the 2010 Asian Games. IEEE INTELLIGENT SYSTEMS May/June:81–85.

Zhang, H.-S., Y. Zhang, Z.-H. Li, and D.-C. Hu. 2004. Spatial–Temporal Traffic Data Analysis Based on Global Data Management Using MAS. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 5:267–275.