Cognitive big data: survey and review on big data research and its implications. What is really “new” in big data?

AQ: au

Artur Lugmayr, Björn Stockleben, Christoph Scheib and Mathew A. Mailaparampil

Artur Lugmayr is based at Visualisation and Interactive Media (VisMedia), Curtin University, Perth, Australia.
Björn Stockleben is based at Film University Babelsberg KONRAD WOLF, Potsdam, Germany.
Christoph Scheib and Mathew A. Mailaparampil are both based at the EMMi Lab., Tampere, Finland.

AQ: 1

Abstract

Purpose – *What is really “new” in Big Data? Big data seems to be a hyped-up concept that has emerged during recent years. But, it requires thorough discussion beyond the common 4V (velocity, volume, veracity and variety) approach.*

Design/methodology/approach – *The authors established an expert think tank to discuss the notion of Big Data, identify new characteristics and re-think what really is new in the idea of Big Data, by analyzing over 60 literature resources. They identified typical baseline scenarios (traffic, business processes, retail, health and social media) as a starting point from which they explored the notion of Big Data from different perspectives.*

Findings – *They concluded that the idea of Big Data is simply not new and recognized the need to re-think a new approach toward Big Data. The authors also introduced a five-Trait Framework for “Cognitive Big Data”, socio-technical system, data space, data richness, knowledge management (KM)/decision-making and visualization/sensory presentation.*

Research limitations/implications – *The viewpoint is centered on cognitive processes as KM process.*

Practical implications – *Data need to be made available in an understandable form for the right application context and in the right portion size that it can be turned into knowledge and eventually wisdom. The authors need to know about data that can be ignored, data that they are not aware of (dark data) and data that can be fully utilized for analysis (light data). In the foreground is human and machine understandability.’ – In form of Cognitive Big Data.*

AQ: 2

Social implications – *Cognitive Big Data implies a socio-technological knowledge system.*

Originality/value – *Introduction of cognitive Big Data as concept and framework.*

Keywords *Cognition, Data analysis, Information management, Knowledge-based systems*

Paper type *Research paper*

1. Introduction

In the public eye, “Big Data” seem to be a rather disruptive innovation, giving birth to a wide range of new technologies, production processes and perspectives for knowledge management (KM). At the same time, one might argue as well that the Big Data phenomenon is a mere incremental innovation which simply massively scales established methods of data processing. Both opinions can be backed with different examples of Big Data applications, and a key problem is that the term Big Data has grown beyond usefulness in professional context. We attempt to contribute to epistemology with innovative thoughts on typical scenarios for Big Data.

The typical discussion in Big Data research is centered on the 4V model: velocity, volume, variety and veracity. Velocity and volume are categories of scale and do not allow for a qualitative distinction of Big Data applications. The same holds more or less true for variety: The only difference in quality that could be argued is between applications that use a

Received 29 July 2016
Revised 6 September 2016
12 September 2016
28 September 2016
2 December 2016
Accepted 2 December 2016

DOI 10.1108/JKM-07-2016-0307

variety of data formats and sources and those that do not. Veracity again is a universal criterion that does not allow distinction, unless we consider the contrast between applications where veracity can be verified and where it cannot. This consideration already points out the necessity to shift the discussion from a technological focus to epistemology. The current discussion seems to be predominantly focused on Vs – problems which have been prevalent since the existence of computer science. We argue that we need to add this ultimate goal to the discussion: development of a theoretical framework to *gain more understanding and increasing perceptibility of underlying data*. In short, we refer to it as *Cognitive Big Data*.

We simply wanted to identify "*What is really new in Big Data?*", and it goes beyond current typical thinking and re-think current viewpoint:

- re-discussing and partially obsoleting the view of 3Vs (4Vs or 7Vs);
- investigating "data" in Big Data according the level of analysis;
- Big Data as socio-economic-technical system creating knowledge;
- visualization, sensory presentation and interaction as key for successful applications;
- cognitive Big Data as the natural consequence of simply data "conveyer belts"; and
- five traits for the characteristics of a cognitive Big Data framework.

With the introduction of the idea of *Cognitive Big Data*, being a socio-technical system capable of creating knowledge and support human understanding of data.

1.1 From a "Data conveyer belt" toward "Cognitive Big Data"

When Henry Ford invented the conveyer belt to mass produce cars and make them available to consumers at low cost, it led to a new societal revolution. To follow this metaphor, we believe that the conveyer belt of data processing already exists: we know how to process data, we know the issues around system scalability and we are able to apply these across many domains. Logistics, financial transactions, e-Commerce and digital manufacturing are just a few examples.

Data should be made perceptible, should support the knowledge acquisition process, create experiences and ease human cognitive load. "Cognitive Big Data" is about the interdependency of the two most capable data processing systems that exist – the human mind and computerized data processing systems. It is not solely about scalability: we introduce the "post-conveyer belt" model of information processing, rather than solely focusing on increasing throughput (i.e. focusing on scalability). Research on Big Data should focus more on knowledge and wisdom processing, and how data assists humans in their cognitive efforts.

1.2 Re-discussing today's view of Big Data: thinking beyond scalability and Vs

Recent advances in technology, the new digital economy and the digitalization of industries have inspired countless new methods of scientific research, social interaction, business intelligence and data analytics. All of these trends together have caused an exponential increase in generated data and introduced new ways of working with data, forms of computation and processing speeds (Johnson, 2012), addressing the traditional problem of scalability. The trend toward increasing amounts of data, new processes for handling vast amounts of data and related technology has been coined "Big Data" during recent years. Like the term "Web2.0", the term "Big Data" tries to summarize several related trends and increasingly proves unsuitable for the scientific discourse on the phenomena it comprises. Settling for one accurate definition is an impossible exercise and not only because of issues related to society and technology (Ward and Barker, 2013).

Current discussion around Big Data deals mostly with issues of scale: increasing data throughput, growing amounts of data and streamlining technical systems to facilitate data processing. The "data conveyer belt" remains a useful metaphor. The faster the belt runs, the more throughput and the cheaper the data product or service becomes for the consumer. Even though the amount of data appears to be exploding recently, the problem of scaling processing and storage capacity has dominated the hardware industry since its very beginning. Rather recently, cloud computing and virtualization introduced software-driven solutions to system scalability.

1.3 A new definition of Big Data: cognitive Big Data

Big Data applications are undoubtedly one of the main drivers for these developments; yet, these solutions tend to narrow our view on the scalability of data processing. In this paper, we explore the realm of Big Data beyond scalability and suggest a draft model that characterizes Big Data applications by how they generate insights and influence decision making. We aim to shift the discussion from the perception of a "conveyer belt of data processing" – which has existed for decades – toward models that are post-data conveyer belt and introduce new thoughts on thinking about the current 3/4V focused view. Our view of data processing is as follows.

1.3.1 Def. cognitive Big Data. Human understandable data are supporting mental capabilities as socio-technical system, in the right application context, right granularity with knowledge and wisdom processing in the foreground. Data need to be re-categorized into data that can be ignored, unaware data (dark data) and data that can be fully utilized in the analysis process (light data).

2. Method and approach

This paper follows a theoretical approach and seeks to shed light on the qualitative not quantitative differences between the current Big Data applications. It is based on previous publications (Lugmayr *et al.*, 2016), as well as on an online resource, which is accompanying this publication (Lugmayr, 2017), where the complete method leading to these conclusions is described in detail. We reviewed over 60 seminal publications (which can be found on the accompanying website[1]), and, based on these, we have selected five baseline scenarios representing the variety of applications currently subsumed under the term Big Data. Through a cross-disciplinary examination through focus groups, we reached the conclusions presented in this article:

- Base Scenario 1 (BS1) – Real-Time Traffic Data and Decision-Making (Intel, 2013).
- Base Scenario 2 (Bs2) – Business Process Management in Logistics and IoT Scenarios (Davenport and Dyché, 2013; Lugmayr, 2010; Lugmayr *et al.*, 2012).
- Base Scenario 3 (BS3) – Consumer Satisfaction Data for Healthcare Services (Davenport and Dyché, 2013).
- Base Scenario 4 (BS4) – Big Data as Cross-Domain Analysis in Retail or for Epidemic Prediction (Mayer-Schönberger and Cukier, 2013; Dugas *et al.*, 2012; Ginsberg *et al.*, 2009).
- Base Scenario 5 (BS5) – Social Media Data and Online Social Network Analysis.

3. Re-thinking "Data" in "Big data": dark data, gray data, light data and data spots

First, we need to investigate the idea of "data" as such and where data is coming from. In statistical analysis, it is typical to create samples and various methods have been developed to mitigate the issue of bias in the selected samples. In Big Data research, data collection is not only restricted to samples but also the idea is to collect as much potentially related data as possible, in many forms and types. This raises questions about the representability of the data collected in relation to its application context, as well as awareness of data that is available for investigation, and which data might be missing without us even knowing.

Complete observation of a system is usually only possible with a narrow research question that can be answered relying on internal data such as complete sales records covering all customers of a company. In the case of complete observation, we can assume bias-free data. For the scope of this article, we would like to define data completeness from a systems theoretical viewpoint:

> We consider a dataset complete if it represents the whole state of a defined system in such a way that the system's behavior could be reproduced if the data was applied to an identical system in a different state.

Using this definition, we conclude that data completeness is prone to error because of human judgment during the process of defining a system model, its parameters and its limitations for a particular application context. Thus, data completeness is a matter of system definition and defining its borders, which is not possible in many practical application cases. A prime example is social media, where only the respective companies have access to the full data sets and decide how much data is exposed to the public for free or with paid access. Twitter offers different plans to access 1, 10 or 100 per cent of user data, respectively (González-Bailón *et al.*, 2012), with no certainty about how the data are selected from the main unit.

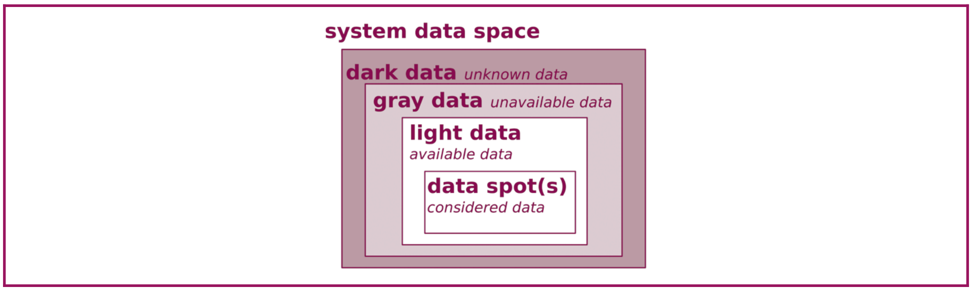
F1 The model depicted in Figure 1 assumes that the relation between main unit and samples may be obscure in Big Data scenarios. It distinguishes between four types of data:

1. *Light data* are the data that are available; it is ready to be accessed and used at any time.
2. *Data spots* are subsets of the light data that are considered in the analysis. As described within the scope of this article, an important challenge in Big Data is deciding which part of the data to look at.
3. *Gray data* are data that are not available to us but that we can make qualified assumptions about and that we know is part of the system we are analyzing.
4. *Dark data* are unknown data that we cannot qualify or quantify in any way. They are the data which we do not know that we do not know about.

We further postulate the following characteristics:

- presence of dark data increases uncertainty of a system model;
- presence of gray data decreases the accuracy of a system model; and
- selection of data spots out of the light data increases the discovery of small patterns and weak signals

Figure 1 Data spaces in Big Data



It should also be noted that the discrimination between a complete and an incomplete observation is far from trivial and depends on the question that must be answered and the goal of the analysis. Google Flu Trends, for example, relies on flu-related search queries without cross-referencing other data to identify flu trends. Yet, the definition of what is or is not a flu-related query is difficult. This also should not be confused with full coverage of all people searching for flu-related medical advice, as the web alone has countless further sources, and there are numerous additional sources of information beyond the web.

F2 Figure 2 shows an abstract description of the data spaces of the five baseline scenarios mentioned earlier. Below follows as detailed discussion of the implication of these very different data space configurations.

3.1 The available data set is complete and fully considered in analysis (BS1 and BS2)

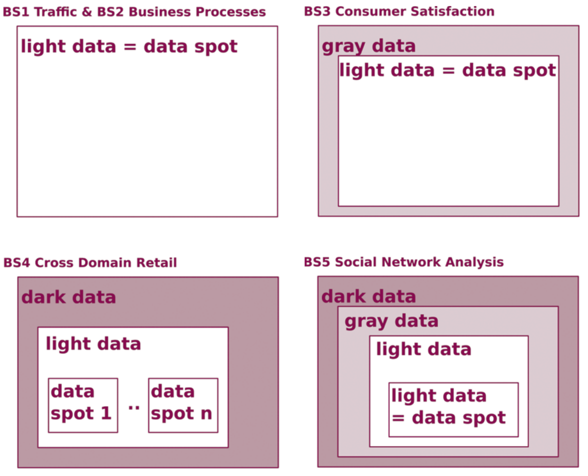
In BS1 (Traffic), the data space is filled with light data, which is fully considered. All data that are necessary to measure and influence the traffic control system of an entire city are available and considered. Once the system is up and running, it makes autonomous decisions on traffic control and optimization. Completeness in this sense does not mean that all possible traffic-related data have been taken into consideration. For example, BS1 does not mention access to navigation information from individual cars, which would greatly enhance the predictive quality of the traffic model. What makes this system complete is the fact that from the different kind of entities that were included in the system, for example, all the city's traffic lights, all instances are covered, not just a subset of them. Additionally, it is important that the system can fulfil its purpose by considering the chosen entities. To illustrate this, imagine that only traffic lights would be controlled but not automated signaling of traffic deviations. The latter would add gray data to the system and would vastly decrease the system's capability to route traffic dynamically.

BS2 (business processes) considers business processes which are human-designed and as such usually constitute a closed system. Thus, the related business entity can have full access to all business process related data, meaning the system data space is filled with light data only.

3.2 The available data set is incomplete (BS3)

BS3 (consumer satisfaction) is an example of a Big Data application working on an incomplete body of data. The extraction of emotional cues forms only one of many sources

Figure 2 Data space descriptions of baseline scenarios



for customer satisfaction, in one unique situation of customer contact (i.e. via the call center). Even if the reliability in BS3 of detection is high, it does not capture the customers' long-term satisfaction, and it certainly does not capture the customers who do not call at all. Still, in this case, the degree of incompleteness of the data can be assessed, i.e. the characteristics of the unknown data can be estimated. The data space is thus filled with light data, which is fully considered, and gray data that represents the data missing from a holistic customer satisfaction model. It can be argued that such a model is too complex to be qualified in its entirety and thus likely involves a lot of dark data, but this is a matter of goal definition and the question of whether the system definition is sufficient to reach these goals.

3.3 Unawareness of possible data (BS4)

BS4 (cross domain retail) also shows a Big Data scenario working on incomplete data, but it differs from BS3 in that the incompleteness cannot be described, resulting in dark data. While BS3 describes a well-defined system with estimable unknowns, BS4 cross-relates different specific data sources without a complete causal model in the background. Just as Google searches for flu symptoms are not the only indicator for detecting a flu epidemic, Wal-Mart sales are not exclusively driven by weather phenomena. Different data spots are correlated to gain viable insights. However, the data sources in this case have not been selected according to a pre-defined system and causal model but because of the mere fact that they correlated sufficiently in the past. A causal model may be inducted in hindsight but does not affect the usefulness of the application.

3.4 Awareness of data which is/is possibly missing (BS5)

BS5 (social network analysis) is a very generic scenario, which may comprise multiple data space configurations. However, this is the only one of the baseline scenarios that includes aspects where automated decisions are made on incomplete data in the presence of dark data. In the area of content recommendation, this is the rule not the exception. No social media platform can cover an individual as a whole partly because of the reason that "personality" is an incredibly complex concept, partially because people act differently in different social contexts. Curating myriads of individual social media feeds and targeted advertisements is a task that cannot be done manually and requires an automated approach using an incomplete data set. The risk in this case is that users become locked into a model that the platform generated for them, and they get stuck in a recursive loop of suggestions – in a "filter bubble" as Pariser (2011) puts it.

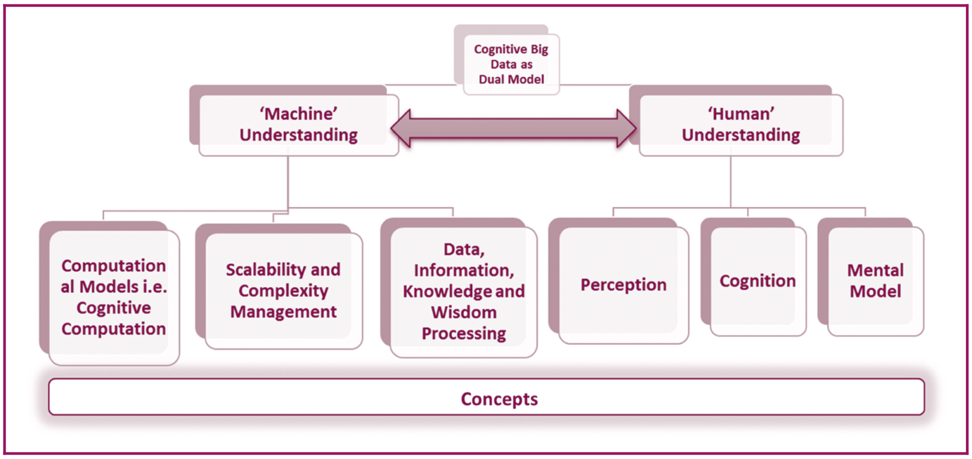
4. Two distinct "Mental models" in cognitive Big Data

Many expectations and promises about the changes Big Data will bring to businesses and private lives have been discussed in scientific research (Brynjolfsson *et al.*, 2011). But is Big Data indeed a paradigm shift that we are observing or simply a notion enriching well-known existing methods such as machine learning or statistics? In this section, we provide a theoretical and epistemological perspective to examine the disruptiveness of Big Data and develop several criteria for the categorization of its applications (Figure 3). F3

4.1 Gaining a deeper understanding of underlying data beyond the Vs

To start the discussion, Floridi (2012) argued that the real epistemological challenge of Big Data is finding small patterns in data sources. In his eyes, scaling data processing systems just leads to even more data and does not solve the challenge of Big Data processing but rather, amplifies it, which is also the case in BS 4 (epidemic prediction). The likelihood of finding patterns and connections between data increases with the volume of available data, but the amount of data that is actually considered in the analysis does not necessarily increase the possibility of identifying patterns in data sets. In fact, patterns can be identified both in data subsets or large datasets (Floridi, 2012). To identify more advanced patterns

Figure 3 Cognitive Big Data as dual model


Cognitive Big Data as Dual Model
'Machine' Understanding
'Human' Understanding
Computational Models i.e. Cognitive Computation
Scalability and Complexity Management
Data, Information, Knowledge and Wisdom Processing
Perception
Cognition
Mental Model
Concepts

that can be turned into knowledge through analysis and interpretation, it is important to identify and decide which parts of the data are relevant and which can be neglected (Bollier and Firestone, 2010).

4.2 Two distinct "Mind models" – machine correlation vs human causation

Anderson (2008) addresses the primacy of correlation over causation straightforwardly as "the end of theory", as in Big Data it is possible to test hypotheses "in the wild" and derive future events from past patterns, without the need of a causal theory explaining why they should happen. Yet, Patrick W. Gross, Chairman of the Lovell Group, counters this view, saying that "in practice, the theory and the data reinforce each other" (Bollier and Firestone, 2010). The high probability of spurious correlations is frequently mentioned as an epistemological problem of Big Data.

Yet, how do we decide whether a correlation is spurious? Simon (1954), in "Spurious Correlations", explains that we need to rely on two types of deductive assumptions: logical assumptions, such as that preceding events cannot be caused by later events, and the assumption that other environmental variables do not interfere with the correlation to be tested. Simon argues that these assumptions are a priori, as they are not founded in statistics but are otherwise empirical and certainly not arbitrary. Nevertheless, 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 proof of causality but rather based on their viability for a certain purpose, using the term "viability" as defined by (Glasersfeld, 1998). Thus, we would discriminate solely between viable and non-viable correlations. The downside of these terms is that there is no absolute truth-value to them. Whether a correlation is viable can only be decided in the context of a specific purpose. Such a context may be the desire to fit the observation made into a larger model or a goal to be reached.

4.3 Dual cognitive Big Data model

The notion of Cognitive Big Data implies a dual model: on one side, it shall support humans in gaining understanding of data and the links between the different data spots, and, on the other side, it also involves a shift towards learning machines and IT systems that are able to learn, understand and sense human intentions. These machines will eventually be capable of understanding the cognitive processes of humans and using this to turn data into perceptible assets.

Some of these ideas go back to the idea of cognitive computer science, where human cognition is simulated through intelligent algorithms to assist humans. Thus, all issues that we discuss currently regarding Big Data will shift toward issues of "understand[ing] nature and improv[ing] the human condition, [and] oceans of structured and unstructured data [lays only] the groundwork" (Kelly, 2015). Information system research needs to consider these trends and support the knowledge creation process inside companies through emerging possibilities, in particular adding machine intelligence and machine-supportive systems. Additional focus needs to be on visual data analytics and visualization to ease the human cognitive load of understanding data.

We see the human and machine as two corresponding entities. Both are building models of observed phenomena:

1. *Correlation model (machines)*: Machines primarily seek to make sense of the world through applying algorithms to the data space and enriching the data space through digitalization of the environment through computational models, without the sole goal of creating causal explanations.
2. *Causal model (humans)*: Humans primarily seek to make sense of the world through turning observations into assets of the experience space in their minds. In principle, observations are validated, filtered, and matched with past experiences through a causal model.

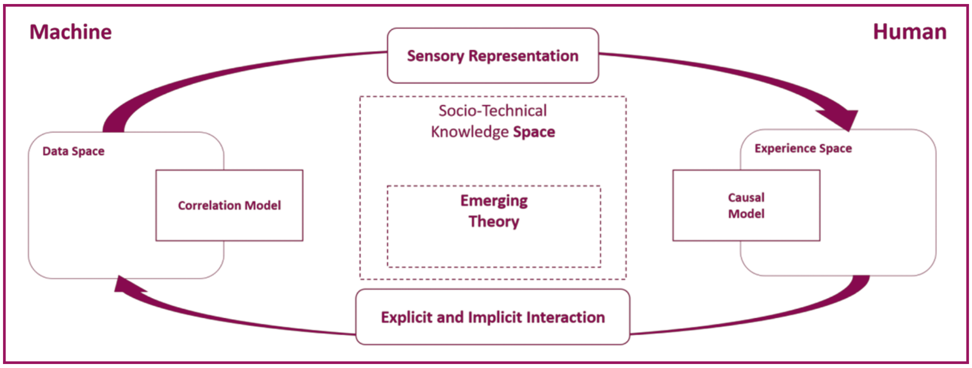
We would strongly recommend other works as (Pinker, 1999; Piaget, 1978; Lugmayr, 2012; Luhmann and Baecker, 2002) to understand the human mind. We would also like to state that machines are capable of computing causal models, for example, Bayesian networks and would like to refer to literature such as (Pearl, 2000) for the interested reader. In this article, we assume that humans primarily seek to build causal models, whereas machines apply computational methods without an inherent "desire" for causal explanations with currently available technologies.

5. A socio-technical knowledge system

We assume that humans always seek to build causal models that are validated against the experience of the respective human being, as shown by Heider and Simmel (1944). In practice, we need to consider social processes, where humans interact and communicate their experiences through communication, as a socio-knowledge system to create a common causal model. For the sake of simplicity, we consider this socio-knowledge system as one entity within our framework. By involving machines, these systems extend towards socio-technical knowledge systems (Figure 4). F4

The machine enriches its data space by building correlation models rooted in the data space. Unlike humans, machines do not have an inherent desire for causal explanations.

Figure 4 Cognitive Big Data between human and machine understanding



They can work with very complex computational models, utilizing mathematic models of causality to refine the data space iteratively. Data itself are "dead matter" which does not provide any inherent grounded truth to verify or gain insights through causal relations. The creation of causal relations, sense-making and validation with current technology is an exclusively human principle. Another important difference is that a correlation model is the result of induction from the data, whereas the human causal model is a result of deduction from existing explicit and implicit models already prevailing in the experience space. Eventually the future shows different pathways, and visions become reality as claimed by Kurzweil (2006).

A draft framework for Cognitive Big Data is depicted in Figure 4. As human and machine interact, there is continuous translation between the correlation model and the causal model working as a cybernetic system. From the perspective of human beings, data visualization provides the essential stimuli to which the human may react. These reactions reconfigure the machine's causal "mental model" through codification, in the form of rules or programs and a section of data. The higher the degree of machine intelligence, the closer the way both models operate shifts together. The results of this interaction decrease the required cognitive load and make data more accessible for users other than trained data analysts.

In each communication cycle, the models lose their context and are translated in the respective other form. By creating sensory representations (e.g. visualizations), the correlation model is removed from the data it is based on. Just a fraction of the original data is presented. Often, these data are presented in the form of summaries and inferred higher-level concepts. When humans perceive visualizations, they are turned into memories of the human's experience spaces, and the abstracted patterns are validated against a network of memories and experiences, which we call a causal model.

If both models do not match, either one must be adapted or the data space has to be reconfigured. In the latter case, the causal model is stripped from its human experience context and translated into a correlation model through codification. An equilibrium is reached when the causal (or correlational) model equals its own translated and retranslated iterations; and when the transformation of one model to the other through codification or sensory representation no longer alters any part of either model. More simply, the socio-technological knowledge space is balanced, and new application-specific knowledge is created. Even more simply, the human experience space and the machine data space are synchronized, and contain knowledge in a human- and machine-useable form.

6. Visualization, sensory presentation and meaning as key challenge

Referring to the system data space model (Figure 1), rendering sensory representations of both light data and data spots appears as a straightforward exercise in data visualization. Data dashboards show all the data that is available to the analyst. However, they do not show what is not available or not known, creating a false impression of completeness. To preserve information about the limits of a system model, we should include information on what we do not know, the dark and gray areas of the system data space. This calls for new forms of visualization or other sensory representations, possibly introducing a new kind of "modest" dashboards that convey information about its own limits and applicability.

Considering the arguments above, the research focus needs to shift towards a more cognitive and perceptive approach, where the emphasis is on cognition, perception, and data understanding to create knowledge and eventually wisdom; in the "machine's mind" as well as in the "human mind". This implies reconsidering the current simplified model of Big Data, towards a focus on the question of how human cognition, and the process of acquiring new knowledge and understanding principles, can be supported; it also requires re-thinking how computers can apply artificial intelligence and other methods to increase the understanding of their environment.

Thus, the job of a business analyst or a data analyst inside a corporation will need to be re-centered to become the role of a “perception data designer” who understands ways to make humans understand principles, knowledge, and wisdom that emerges from manifold corporate data sources. Data visualization is a current buzzword that many people relate with charts, histograms, arc diagrams, or colorful social network graphs. But we are talking about quite more far reaching ideas – more conceptual representation through internet of Things or smart media, increasing perception and going far beyond simple visualizations through computer graphics.

It is not only about the passive act of displaying information. We must consider the complete process of making Big Data perceptible, including human cognition, communication and the actions humans undertake to gain more understanding of knowledge and wisdom from Big Data. This process is visualized in Figure 5 based on the original diagram from (Spence, 2001), where information visualization is discussed. It is about transforming a computer-generated correlation model into a causation model that can be understood by humans. How humans are able to interact with the various models, to gain more insights, is also an issue. The process of making Big Data perceptible and understandable requires computers that can codify and interpret data, which is currently a major issue in the domain of Cognitive Computation. Linking data, and the visualization of correlations between data spots, allows more insights and makes data perceptible and cognitive, as currently conducted in social network analysis.

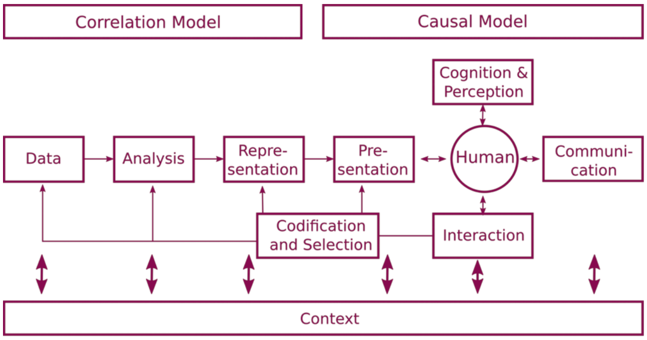
7. Consequences of cognitive Big Data in knowledge management

In terms of KM, the primary question is whether the insights we gain from data analysis converge into a consistent mental model or not. Further, what implications does it have on decision-making if they do not? A mental model requires a human understanding of causal relationships. These relationships have to be plausible to the person creating the mental model, i.e. they have to be consistent with the person’s prior knowledge. This is the reason why mental models tend to be stable: they act as a filter for the interpretation of data.

7.1 5 trait model for characterization of cognitive Big Data rather than Vs

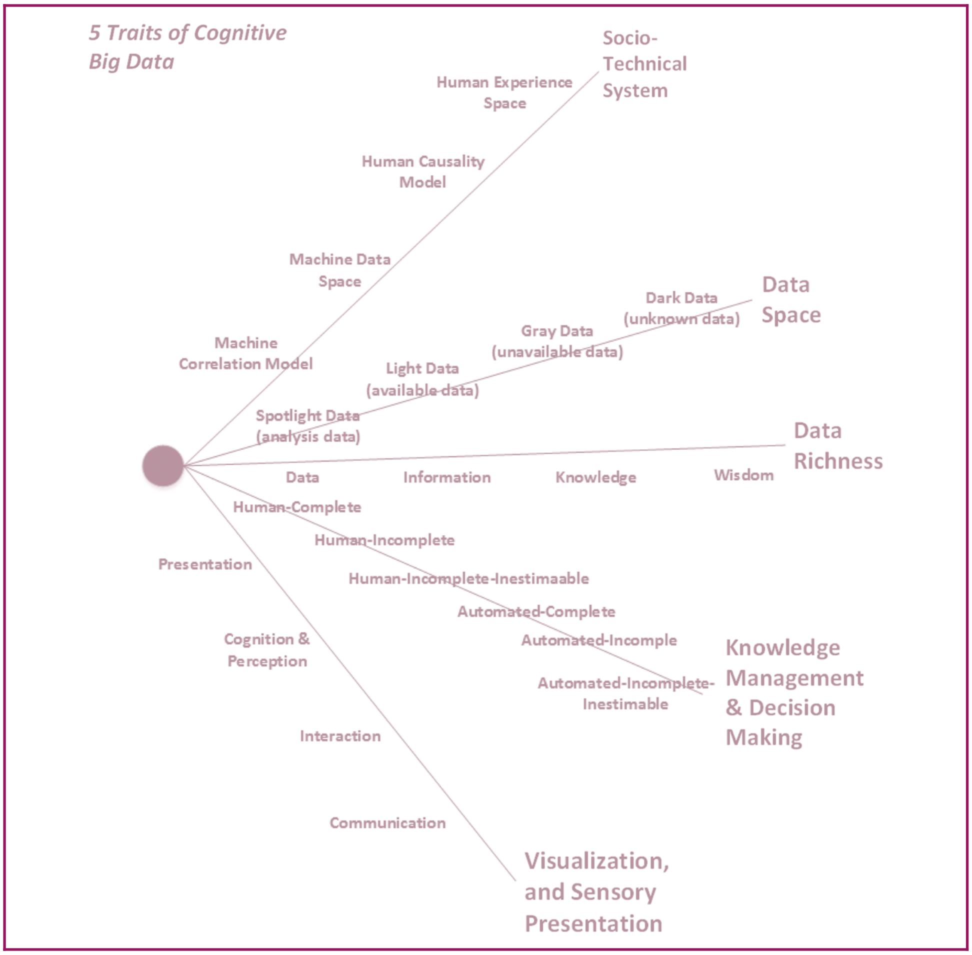
Cognitive Big Data postulates Big Data analysis as a cybernetic system involving both human and machine as illustrated in Figure 6. Both rely on a distinct body of knowledge: machines rely on a vast data space, while humans rely on a highly abstract “experience space”. At the point of equilibrium, the system creates congruent knowledge shared

Figure 5 Data visualization and perception process



Source: Extended from Spence (2001)

Figure 6 Cognitive Big Data: five traits model



between machine and human, i.e. knowledge that is biunique between both entities. At its best, the system is at this point of equilibrium, and the human understanding is congruent with the machine's understanding; this kind of biunique knowledge leads to a system in which both machine and human can understand and judge decisions made by the other, avoiding disconnects where humans or machines lose control. One example is the microeconomic procrastination systems used by Amazon, Google, or Airbnb to analyze consumers' buying behavior – these systems judge, and then predict, consumer behavior.

7.2 Focus on human understandable correlation models

The classic model of knowledge levels may soon become obsolete. Humans have the desire to understand how things work, and only accept a correlation if they can find a plausible explanation (e.g. visualizing epidemics in BS4). This means they are more likely to ignore a correlation they cannot explain. Algorithms do not feel this desire, as they do not care if a correlation is spurious in human eyes or not. They simply decide based on 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 and ignoring any attempt to explain them. This allows very fast evolution of models about the world (or a part of it that a particular Big Data application considers). These models will always remain less complex than the real world, and lag behind reality. Big Data systems can only measure what has already happened, and the future role of humans in the techno-social ecosystem of Big Data remains open to further research.

In addition, let us consider the creation of more resilient automation through more resilient algorithms. Algorithms may show unpredictable behavior when based on dynamic, incomplete data. They may have *tipping points* (Gladwell, 2006), at which their behavior changes disruptively. While the focus so far has been on creating algorithms for data analysis, we may need to create algorithms that analyze such algorithms, e.g. self-learning algorithms evolving based on induced rules, or on a societal level, an open algorithm movement, similar to the open data movement that is currently taking place.

7.3 The impact of cognitive Big Data on knowledge management and computerized vs human decision-making

While Big Data technologies have vastly increased our means of processing vast amounts of data, the capacities of human cognition do not increase (e.g. in BS 1 drivers make the final decisions). In Big Data, automation through algorithms helps us to decide how to focus our attention, i.e. to decide what to look at, implicitly ignoring anything else. What we as users see is always a representation of aggregated data on a distinct abstraction level. This level can be anything, from statistical figures describing the dataset as a whole, to a single data unit. Yet we can only focus our attention on the whole dataset on a high abstraction level, or on very few single data units on a concrete level. Therefore, the algorithms used, and their configuration, restrict what we may or may not see in a dataset. For example, preset thresholds determine what will surface as a pattern, and what will 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 cannot recognize their potential meaning for applications. The higher the velocity of a Big Data application, the more it has to rely on algorithms, first for deciding which to data to observe and second for deciding how to react to its observations. BS 1 (real-time traffic data and decision-making) and the emergence of smart self-driving cars and the ethics dilemma how to react in traffic accidents.

7.4 Big data applications as knowledge creation ecosystem enriching humanities knowledge

Scenarios as BS 3 (consumer satisfaction for healthcare) or BS4 (epidemic prediction) require data about humanities. Big Data are a cultural, scholarly and technological phenomenon creating knowledge ecosystems, and influencing how we think about knowledge itself (Boyd and Crawford, 2012). The research field of Digital Humanities, for example, investigates human knowledge and cultural aspects in humanities (Lugmayr and Teras, 2015). Privacy, and the digital footprint we leave behind every single day (Michael and Miller, 2013). All the positive aspects of personalization, for example, through smart chat agents as Microsoft's Cortana, and workforce productivity monitoring, might end up becoming "big brother". Issues such as data ownership and selling data (Boyd and Crawford, 2012) are impacting our social lives.

7.5 New cognitive technologies as enabler for cognitive Big Data

Each of the BSs have an underlying new technological enabler (i.e. BS 5 social media, a new media form). Big Data research is currently peaking on its plateau of expectations, when illustrated on a hype cycle. Many new technologies will need to emerge before promises will be able to be kept. A new way of thinking is required, so that Big Data does not become a disappointment, and will shift towards a stable position on its plateau of productivity. Models and frameworks like ours will ensure that Big Data will shift towards maturity. Big Data will support decision making in business intelligence systems across business areas and provide a competitive advantage.

7.6 "Algorithmic regulation" for enabling humans to keep with the pace of computational power

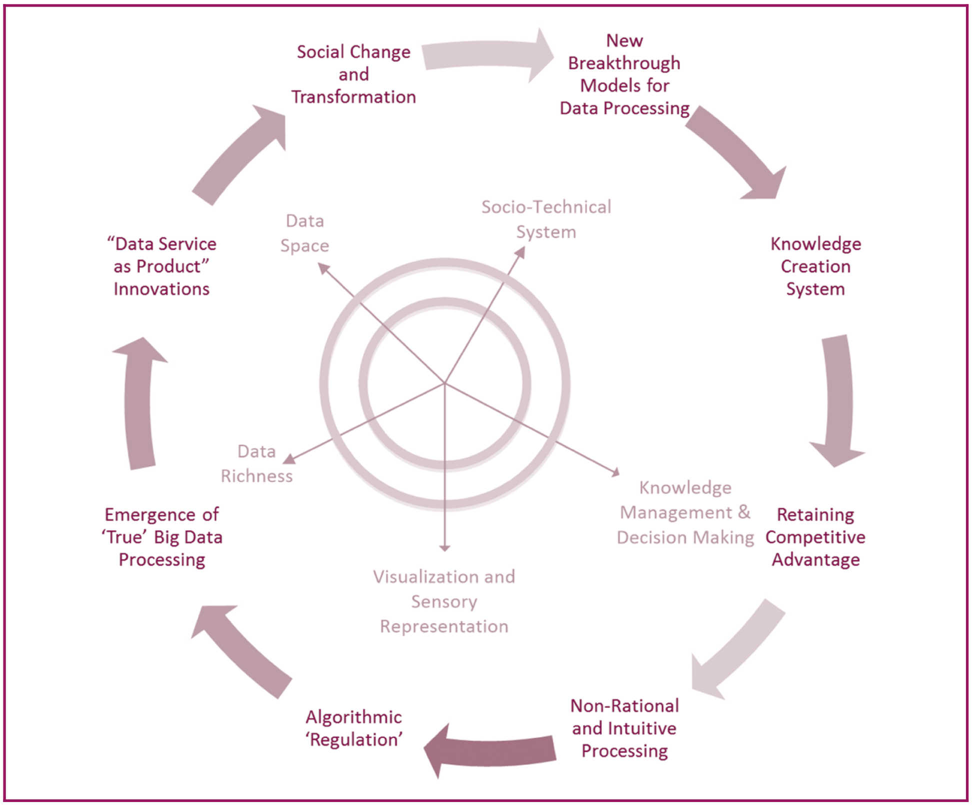
In the finance industry, automated trading algorithms are today's stockbrokers. This illustrates the complexity of automatic and algorithmic data processing. As suggested by Tim O'Reilly, who predicts the dawn of "algorithmic regulation", it may be necessary to introduce regulations that limit calculations in real-time. In financial industries, this trend is already taking place – we argue that it is only a matter of time until real-time processing in business applications is also regulated. (Steinbrune, 2002, *The Cybernetic Theory of Decision: New Dimensions of Political Analysis*) argues that rational assumptions in politics lead to random, context-less and systematic decisions, where a non-rational way of problem solving would be for the good of the population.

8. Conclusion

Cognitive Big Data enables competitive advantages and leads to innovative new business areas. It is obvious that any new innovation will lead to new business activities focusing on collecting and interpreting complex information from internal and external sources and will enable many new business niches for newcomers (Johnson, 2012). Figure 7 depicts the key-ideas of this article. F7

Therefore we contributed with a new way of thinking about Big Data within the scope of this article. We proposed a new classification scheme in form of a five traits model, and discussed a new way to classify Big Data scenarios dependent on the composition of their data spaces. We suggested Cognitive Big Data as an iterative communication process between man and machine, and the resulting mental models. This model also means to inspire more research in the domain of data visualization and data capture from both,

Figure 7 Cognitive Big Data framework


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 Space
Socio-Technical System
Knowledge Management & Decision Making
Visualization and Sensory Representation
Data Richness

cognitive sciences and user experience angles. Thus, shared knowledge between humans and machines can only emerge at the inter-section of their very distinct approaches to capture and understand the world.

Note

1. Additional material: www.artur-lugmayr.com

References

Anderson, C. (2008), "The end of theory", *Wired Magazine*, available at: www.simulex.it/Materiali/Articoli/Chris%20Anderson-The%20end%20of%20theory.doc

Bollier, D. and Firestone, C.M. (2010), "The promise and peril of big data", Aspen Institute, Communications and Society Program, Washington, DC, available at: www.ilmresource.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

Boyd, D. and Crawford, K. (2012), "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, Vol. 15 No. 5, pp. 662-679, available at: www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878

Brynjolfsson, E., Hitt, L. and Kim, H. (2011), "Strength in numbers: how does data-driven decision making affect firm performance?", available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486

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

Dugas, A.F., Hsieh, Y.-H., Levin, S.R., Pines, J.M., Mareiniss, D.P., Mohareb, A., Gaydos, C.A., Perl, T.M. and Rothman, R.E. (2012), "Google flu trends: correlation with emergency department influenza rates and crowding metrics", *Clinical Infectious Diseases*, Vol. 54 No. 4, pp. 463-469, doi: 10.1093/cid/cir883.

Floridi, L. (2012), "Big data and their epistemological challenge", *Philosophy & Technology*, Vol. 25 No. 4, pp. 1-3, available at: www.springerlink.com/index/5550128750462P0W.pdf

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457 No. 7232, pp. 1012-1014, doi: 10.1038/nature07634.

Gladwell, M. (2006), *The Tipping Point: How Little Things can Make a Big Difference*, Little, Brown and Company.

Glasersfeld, E. (1998), "Konstruktion der Wirklichkeit und des Begriffs der Objektivität", *Einführung in den Konstruktivismus*, Piper, München, pp. 9-39.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J. and Moreno, Y. (2012), "Assessing the bias in communication networks sampled from twitter", available at: http://arxiv.org/abs/1212.1684

Heider, F. and Simmel, M. (1944), "An experimental study of apparent behavior", *The American Journal of Psychology*, Vol. 57 No. 2, pp. 243-259, available at: www.jstor.org/stable/10.2307/1416950

Intel (2013), *Improving Traffic Management with Big Data Analytics*, Intel Corporation.

Johnson, J.E. (2012), "Big Data + Big Analytics = Big Opportunity", *Financial Executive*, Vol. 28 No. 6, pp. 50-53, available at: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=77604799&site=ehost-live

Kelly, J. (2015), *Computing, Cognition and the Future of Knowing*, IBM Research and Solutions Portfolio, available at: www.research.ibm.com/software/IBMResearch/multimedia/Computing_Cognition_WhitePaper.pdf

Kurzweil, R. (2006), "Reinventing humanity: the future of machine-human intelligence", *The Futurist*, pp. 39-48.

Lugmayr, A. (2010), "Introduction to the business processes with ambient media - challenges for ubiquitous and pervasive systems", in Yu, Z., Liscano, R., Chen, G., Zhang, D. and Zhou, X. (Eds), *Ubiquitous Intelligence and Computing*, Vol. 6406, pp. 125-137.

Lugmayr, A. (2012), “Connecting the real world with the digital overlay with smart ambient media – applying Peirce’s categories in the context of ambient media”, *Multimedia Tools and Applications*, Vol. 58 No. 2, pp. 385-398, doi: 10.1007/s11042-010-0671-3.

Lugmayr, A. (2017), available at: www.artur-lugmayr.com/newtiki/Publications

Lugmayr, A. and Teras, M. (2015), “Immersive interactive technologies in digital humanities: a review and basic concepts”, *Proceedings of the 3rd International Workshop on Immersive Media Experiences, ImmersiveME ’15*, *ACM*, *Brisbane*, pp. 31-36, doi: 10.1145/2814347.2814354.

Lugmayr, A., Zou, Y., Stockleben, B., Lindfors, K. and Melakoski, C. (2012), “Categorization of ambient media projects on their business models, innovativeness, and characteristics - evaluation of Nokia Ubimedia MindTrek Award Projects of 2010”, *Multimedia Tools and Applications*, Vol. 66 No. 1, pp. 1-25.

Lugmayr, A., Stockleben, B., Scheib, C., Mailaparampil, M., Mesia, N. and Ranta, H. (2016), “A comprehensive survey on big data research and it’s implications - what is really ’new’ in Big Data? It’s cognitive big data”, *Proceedings of the 20th Pacific-Asian Conference on Information Systems (PACIS 2016)*, available at: www.pacis2016.org/abstract/Index

Luhmann, N. and Baecker, D. (2002), *Einführung in die Systemtheorie*, Vol. 2, Carl-Auer-Systeme-Verlag.

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

Michael, K. and Miller, K.W. (2013), “Big Data: new opportunities and new challenges”, *Browse Journals & Magazines*, Vol. 46 No. 6.

Pariser, E. (2011), *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*, Penguin Publishing Group, available at: https://books.google.com.au/books?id=wcalrOI1YbQC

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.

Piaget, J. (1978), *Success and Understanding*, Harvard University Press, Cambridge, MA.

Pinker, S. (1999), “How the mind works”, *Annals of the New York Academy of Sciences*, Vol. 882 No. 1, pp. 119-127.

Simon, H.A. (1954), “Spurious correlation: a causal interpretation”, *Journal of the American Statistical Association*, Vol. 49 No. 267, pp. 467-479, available at: www.tandfonline.com/doi/abs/10.1080/01621459.1954.10483515

Spence, R. (2001), *Information Visualization*, Vol. 1, Springer-Verlag.

Steinbrune, J.D. (2002), *The Cybernetic Theory of Decision: New Dimensions of Political Analysis*, Princeton University Press, Princeton, NJ.

Ward, J.S. and Barker, A. (2013), “Undefined by data: a survey of big data definitions”, available at: http://arxiv.org/abs/1309.5821

About the authors

Artur Lugmayr is Professor at Curtin University, Australia, where he teaches and supervises students in visualization technologies, interactive media, media business and ubiquitous media. Artur was Professor for digital media management in Tampere, Finland 2009-2014 establishing the Entertainment and Media Management Lab. (EMMI Lab). and the New Ambient Multimedia Lab. 2004-2009 (NAMU Lab.). Artur holds a Dr-Techn degree (Information Technology) and is pursuing his Dr-Arts studies at Aalto Univ., Helsinki, Finland in motion pictures. He was visiting scientist in Singapore, Brisbane, Austria, Ghana; since 2000 raised/involved in 1.7+ MEUR funding (excl. 2014/2015 applications); 170+ publications; 24+ invited keynotes; and 27+ invited guest lectures. He founded and chairs the Association for Information Systems (AIS) Special Interest Group “AIS SIG eMedia” and the International Ambient Media Association (iAMEA) and is active member of the ACM TVX steering board, IFIP TC 14 for Entertainment Computing and IEEE IG MENIG. More about me on www.artur-lugmayr.com. Artur Lugmayr is the corresponding author and can be contacted at: artur.lugmayr@artur-lugmayr.com

Björn Stockleben is Professor for new media production at Film University Babelsberg *KONRAD WOLF*. His current research areas comprise management of interdisciplinary teams, data-driven media production, online collaboration and non-linear AV media. He holds a master's degree in Media Sciences, Media Technology and Computer Sciences from Technische Universität Braunschweig and Braunschweig University of Art in 2003. He used to work as coordinator of the Master programme Cross Media at University of Applied Sciences Magdeburg-Stendal. Before, he worked as technical manager and UX consultant in EC and ESA funded research projects such as HBB-NEXT, ARTES-COTV and TV-Ring at Rundfunk Berlin-Brandeburg innovation project group. He coordinates the Erasmus+ partnership "OnCreate" on creative processes in online collaboration. He has lectured in media theory and human-machine interaction at various universities and is a PhD student in Media Management at TU Tampere, Finland.

Christoph Scheib holds a master degree in Business Administration from the Europa University Viadrina, Germany. Besides that, he has been studying international Business at the University of Tampere, Finland. After his studies, he has been heading the Business Intelligence Unit of an international telecommunications company in Finland for several years. Christoph has been focusing very intensively on the topics of BI, while working for an international consultancy in Finland. Currently, he is based in Munich, Germany and is working as Delivery Manager for an international Consultancy.

Mathew A. Mailaparampil holds a master's degree in software and communications engineering from the Tampere University of Technology, Finland and a bachelor degree in electrical engineering from the University of Kerala, India. He co-authored four publications while working as a research-assistant at the Tampere University of Technology. Along with several years of experience from the telecommunications industry as a project manager, senior localization engineer and quality manager he has a green belt in lean Six Sigma and PRINCE2 certification in project management. He is based in Tampere, Finland and is setting up a start-up in the technology sector.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 1

AQau—Please confirm the given-names and surnames are identified properly by the colours.
🟥= Given-Name, 🟦= Surname
The colours are for proofing purposes only. The colours will not appear online or in print.

AQ1— Note that to conform to the journal guidelines, please provide a sentence that uses the phrase “the purpose of this study/paper…” or “this study/paper aims to…” in the first line of the Purpose of the Abstract.

AQ2— Please note that the following sentence is unclear as given. Please consider revising the sentence for clarity: In the foreground is human and machine understandability.’ – In form of Cognitive Big Data.

AQau> Yes, the authors are correctly spelled.

AQ1> The purpose of this paper is to introduce and define Cognitive Big Data as concept. Furthermore, it investigates what is really “new” in Big Data, as it seems to be a hyped-up concept that has emerged during recent years. The purpose is also to broaden the discussion around Big Data far beyond the common 4V (velocity, volume, veracity, and variety) model.

AQ2> Please change “In the foreground is human and machine understandability” into “In the foreground is the extension of human mental capabilities and data understandability.”