Conceptualizing Big Data: Analysis of Case Studies

Article *in* Intelligent Systems in Accounting Finance & Management · May 2016

DOI: 10.1002/isaf.1393

Ossi Ylijoki
Finanssi-Kontio Oy
8 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Jari Porras
Lappeenranta – Lahti University of Technology LUT
186 PUBLICATIONS **615** CITATIONS
SEE PROFILE

Conceptualizing Big Data: Analysis of Case Studies

OSSI YLIJOKI[a*] AND JARI PORRAS[a]
[a] *School of Business and Management, Lappeenranta University of Technology, Finland*

*Corresponding author, address: Puistokatu 1C38, FI-15100 Lahti, Finland;
e-mail: ossi.ylijoki@phnet.fi; tel. +358442387442

ABSTRACT. Digitization and the related datafication produce huge amounts of data. Organizations have started to exploit these new data in order to gain benefits. Exploring this ”big data jungle” is a new area for both scholars and practitioners, and the experiences of early adopters are valuable. This paper analyses big data use cases described in the academic literature by using computerized content analysis methods. Based on the analysis results, we have conceptualized themes and guidelines of big data in the context of an organization, thus contributing to the emerging research of big data. In addition to the realized benefits, the case studies reveal issues regarding technology, skills, organizational culture, and decision-making processes. The paper also points out several new research avenues, acts as a reference collection to big data case studies found in academic sources, and bridges theory and practice by pointing out several topics that practitioners should consider.

Keywords: big data, case study, content analysis, digital transformation, digitization

1. Introduction

Today, new digital technologies produce vast amounts of various types of data (Gantz and Reinsel, 2011), often referred to as big data. From the point of view of technology, big data are different from traditional transaction data, requiring new data management and analysis technologies (Laney, 2001). More importantly, several sources, including (Davenport, 2014; Manyika et al., 2011; Mayer-Schönberger and Cukier, 2013) claim that big data have potentially huge effects on many industries. Technology and data drives change, and as e.g. (Dehning et al., 2003; Sainio, 2005) suggest, companies must link their strategy with technology. The business environment is changing. However, it is difficult to forecast the impacts at the micro level, as digitization and data deluge are a new, emerging phenomenon.

The effects of this phenomenon are different for each company. As an example, self-driving cars[1], which will invade the markets in the future, will have significant effects on various firms, like car dealers and insurance companies. However, the potential and the challenges that a car dealer faces will differ significantly from those of an insurance company. Realizing the potential implies that this new, data-driven paradigm will affect companies’ strategies and business models heavily. Several excellent pieces of work exist on business transformation. Venkatraman (1994) builds a framework that helps understand the effects of the transformation. Christensen (2013) explains clearly how incumbent companies fail

[1] E.g. Google: http://googleblog.blogspot.fi/2015/05/self-driving-vehicle-prototypes-on-road.html, Nissan: http://abcnews.go.com/Technology/nissan-driving-car-ready-2020-ceo/story?id=31120512, or Volvo: http://www.wired.com/2015/02/volvo-will-test-self-driving-cars-real-customers-2017/

constantly in utilizing new, disruptive technologies. Sainio (2005) shows that companies are often well aware of new, emerging technologies, but neglect linking the technologies with their strategies.

There are some trailblazers, Google and Amazon being the most obvious examples, which have built their business models around data. These kind of examples, as well as some previous studies, e.g. (Dehning et al., 2003; McAfee et al., 2012; Porter and Millar, 1985) indicate that companies utilizing data heavily gain competitive advantage over their less data-driven rivals. However, the data-driven approach is still a new paradigm for most organizations (Shen and Varvel, 2013). In addition, established companies have their own history, processes, and capabilities. They just cannot turn their existing structures and business models upside down at once. The transformation takes time. When established firms start to explore the possibilities of big data, they can learn from the experiences and methods of the early adopters. Several studies, e.g. (De Mauro et al., 2015; Wamba et al., 2015) recognize the need for guidelines and a conceptual framework for big data. One way towards this goal is to examine the experiences of real big data projects. In this article, we use computerized text analysis methods to analyze a number of big data case studies documented in academic publications.

The key contribution of this article is that we synthesize the findings (benefits and challenges) of our case study analysis to a set of generic themes and guidelines. This contributes to the research on big data by conceptualizing existing practices and pointing out several new research avenues. In addition, this work bridges practice and theory, acts as a reference collection to currently known, peer-reviewed big data case studies, and benefits practitioners by providing guidelines and experiences from the early adopters of big data.

2. Big Data Case Studies

This section describes the research process we used to identify big data case studies. We used the systematic mapping study approach presented by Kitchenham (2007). Our goal was to identify well-documented big data case studies in the academic literature. Well-documented in this context means a peer-reviewed, high quality source. In order to cover the area broadly, we performed a systematic mapping study. According to Kitchenham (2007), mapping studies are designed to give a broad overview of a research area. Mapping studies have typically have broad research questions. Our target (research question) was simple: to locate as many big data case studies documented in peer-reviewed sources as possible, and to capture the common concepts and lessons learned in these use cases. Figure 1 gives an overview of the search process.

2.1 Search Strategy

Big data is a multi-disciplinary phenomenon. Unlike some other subject areas, big data related articles cannot be found only in certain highly focused forums. Although there are some new journals that focus on big data, various publications in many research domains discuss the topic. Big data is an emerging, multi-disciplinary research area.

Initial search. First, in order to identify a representative set of well-documented studies, we searched for cases in literature databases at the end of August 2015. We studied four literature databases: Scopus, ProQuest, Web of Science, and EBSCO, using a rather broad search terms ("'big data' and 'case study'"), and limiting our search to peer-reviewed papers. The reason for this was to avoid "commercially-oriented" cases, i.e. we wanted to stick to papers that had gone through scientific evaluation. In addition, we filtered the results to contain only papers written in the English language. These searches revealed 281 papers.

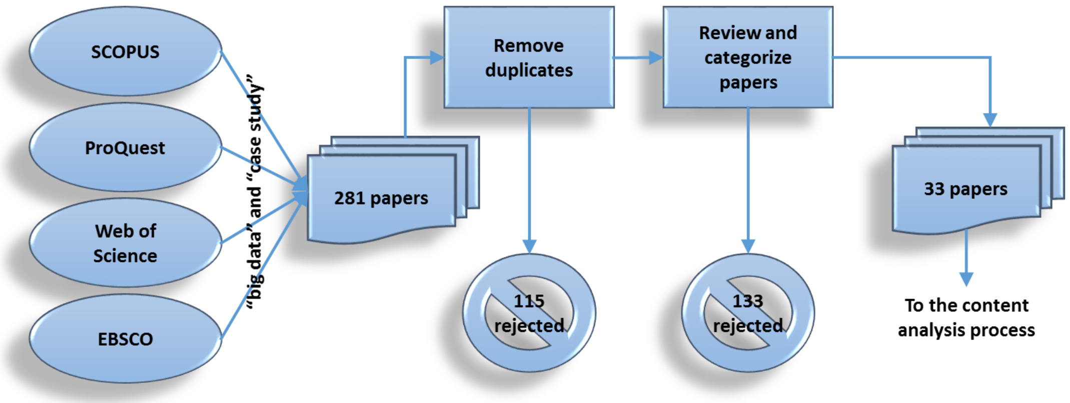

Figure 1. Search process for big data case studies.

Exclusion criteria. First, we removed 115 duplicate papers from the initial result set. Then we reviewed and categorized the remaining 166 articles. By reading the abstracts, and whenever necessary the introduction and conclusions, we categorized the studies to be either *case-focused* or *non-case-focused*. This led to the exclusion of 120 studies, which focused on developing e.g. new algorithms, methods or frameworks. These papers verified or clarified their contributions typically with an experimental prototype, proof-of-concept, or something similar. Altogether, their focus was on developing something, not describing a case study. We rejected additional 13 papers for various reasons: the paper contained a hypothetical case (3 studies), we could not access or find the paper (9), or the paper was in Spanish (1).

As a result, the search process revealed 33 peer-reviewed big data case study papers containing in total 49 case studies due to three multi-case studies (Bärenfänger et al., 2014; Kowalczyk and Buxmann, 2014; Wehn and Evers, 2015). Appendix 1 lists the papers and provides a short contextual description of each paper. Next, we analyzed the articles describing big data cases. For the analysis, we used a quantitative natural language processing software to identify common concepts and themes. Finally, we analyzed the results of the text-mining phase and formulated a set of guidelines.

2.2 Characteristics of the Case Studies

The found cases represented different application domains, from education to business, and from healthcare to entertainment. This indicates that big data affects every aspect of life. Table 1 lists the number of cases categorized by the ISIC classification of the UN (UnitedNations, 2008). ISIC has 21 categories; we identified at least one big data case in 15 (71%) of these categories. Transportation, especially intelligent transport systems -related studies, and various healthcare studies represented the highest numbers of cases (8 and 6, respectively). Several industries were also well represented with four or five cases each: manufacturing, retail, finance, and information -related cases.

Table 1. Big data case studies by application area.

Application area (categories adopted from (UnitedNations, 2008))	**Number of cases**
A-Agriculture, forestry and fishing	1
B-Mining and quarrying	-
C-Manufacturing	5
D-Electricity, gas, steam and air conditioning supply	2
E-Water supply; sewerage, waste management and remediation activities	-
F-Construction	3
G-Wholesale and retail trade; repair of motor vehicles and motorcycles	5
H-Transportation and storage	8

I-Accommodation and food service activities	2
J-Information and communication	4
K-Financial and insurance activities	4
L-Real estate activities	-
M-Professional, scientific and technical activities	2
N-Administrative and support service activities	1
O-Public administration and defense; compulsory social security	1
P-Education	3
Q-Human health and social work activities	6
R-Arts, entertainment and recreation	2
S-Other service activities	-
T-Activities of households as employers; undifferentiated goods- and...	-
U-Activities of extraterritorial organizations and bodies	-

All the papers were recent, which is not surprising, since most organizations are still taking their first steps with big data. Figure 2 presents the number of the case study articles per publishing year of the *paper* (not the cases). Note that we did our searches at the end of August 2015, which explains the relatively low number of studies published in 2015.

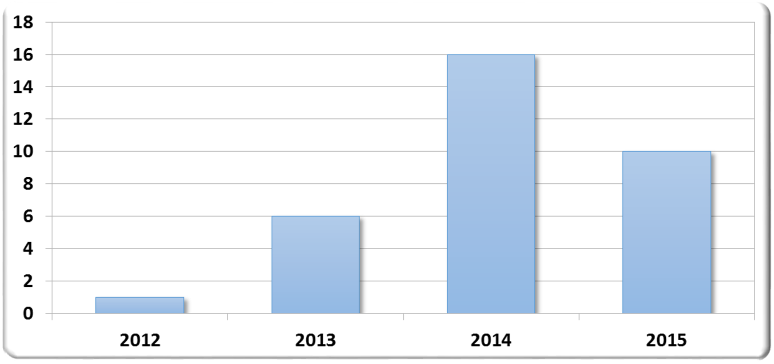

Figure 2. Number of big data case study articles by the publishing year.

As with the application area, also the geographical distribution of the cases was wide, representing five continents (figure 3). Companies based in North America and Europe represented a majority of the cases with 29 instances. Beyond that, there were cases from Asia, Australia and Africa. One of the studies, a multi-case study of 12 cases shown as “n/a” in figure 3, did not report the origin of the cases.

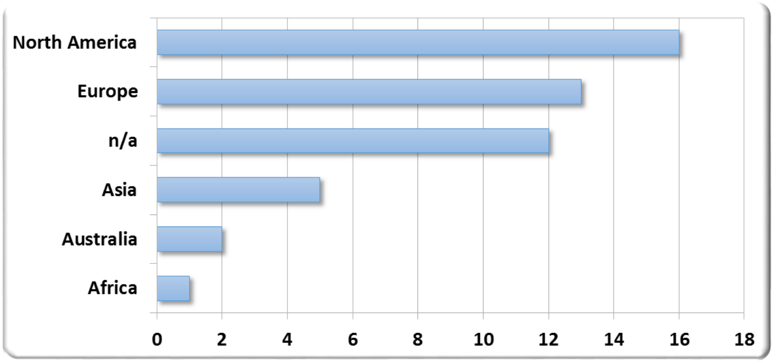

Figure 3. Geographical distribution of the big data cases.

Appendix 1 lists the case study papers. A brief description of the case context with industry categorization provides basic information of the cases.

3. Content Analysis of the Case Study Papers

Content analysis is an established methodology for investigating textual data (see e.g. (Berelson, 1952; Holsti, 1969; Krippendorf, 1989). Weber (1990) defines content analysis as a repeatable, systematic procedure that reduces the many words of a text to much fewer content categories. Novel applications of computerized content analysis have received the attention of scholars recently, e.g. (Hu et al., 2014; Lewis et al., 2013; Yu et al., 2014), as researchers wish to utilize new big data sources. In our case, manual coding of the texts of the 33 articles would have been a time-consuming job, and therefore we considered computerized content analysis to be a proper method for revealing common big data concepts and lessons learnt in the articles.

We had no pre-defined categories or themes. By using the data-driven approach, we just drew the patterns from the articles with the analysis software. "Let the data speak", as Mayer-Schönberger and Cukier (2013) put it. As a tool we chose an open source software, KH Coder[2]. It supports several text analysis methods described in content analysis studies, and more than 900 research projects have used the software.

Stemler (2001) suggests word counting and key words in context (KWIC) -analysis as a starting point of a content analysis. KH Coder can count words, and more: it uses Stanford POS Tagger (Toutanova et al., 2003) for tagging and lemmatization of words, i.e. it recognizes parts of speech (such as nouns, verbs, and adjectives) and converts words to their base format. This, combined to word frequency counting functionality, provided us a good basis for the analysis. We used word frequencies and the KWIC analysis to create a so-called "stop-word" list. Stop words are common words that exist in almost every sentence. Stop words are not included in further analyses, as they do not add information; on the contrary, they make the results more difficult to perceive. E.g. Wilbur and Sirotkin (1992), Yang and Pedersen (1997) and Yang and Wilbur (1996) discuss automatic identification of stop words. We had to use the manual method, as the KH Coder does not support automation. However, the KWIC analysis tools of the KH Coder proved to be an efficient means to ensure whether the word was relevant or not in the context that we were interested in.

We visualized the results with KH Coder software using co-occurrence maps.[3] Co-occurrence maps build on the idea that words are related to the concepts they are connected to (Ryan and Bernard, 2003). Osgood (1959) was among the first scholars to use co-occurrence matrices to reveal connected concepts in textual data.

Figure 4 shows the co-occurrence map that resulted from the analysis of the 33 big data case study articles (representing 49 cases) after several analysis iterations. The map revealed five main themes and two sub-themes. The different colors distinguish the themes. We labelled the themes based on the following: First, based on the virtual value creation process (Rayport and Sviokla, 1995), we distinguished between data and data usage, as suggested by Ylijoki and Porras (2016). Three of the main themes are business or organization -related, representing the usage or utilization of data. Two of them are ICT- and data-related, technical themes. Then we

[2] KH Coder is a free software for quantitative content analysis or text mining - http://khc.sourceforge.net/en/

[3] A few notes that clarify the interpretation of the map: When plotting a map, the KH Coder uses the method explained in (Fruchterman and Reingold, 1991). This algorithm may plot nodes side by side, but unlike e.g. multi-dimensional maps, this does not necessarily indicate co-occurrence. Instead, edges (lines) indicate co-occurrence: if a line connects the nodes (words), co-occurrence exists. For example, in figure 4, the terms 'customer' and 'organization' are close to each other, but there is no co-occurrence between them, since there is no line between the words. Accordingly, a strong co-occurrence between the terms 'value' and 'generate' exists, as there is a thick line between them. The thicker the line, the stronger the co-occurrence is. The dotted lines show co-occurrence between terms that belong to different communities (i.e. themes). The size of the plot indicates the frequency of the term, 'data' being obviously the term used most frequently in the articles. The color-coding indicates the communities (sub-graphs) that are relatively close to each other. The KH Coder offers several methods for indicating patterns. We used the modularity method defined in (Clauset et al., 2004). This method builds on the principle that there are many edges within the communities and only a few between them.

decided the label for each theme based on KWIC-analysis and manual inspection of the articles. The co-occurrence of words within a theme is presented with a solid line between the words.

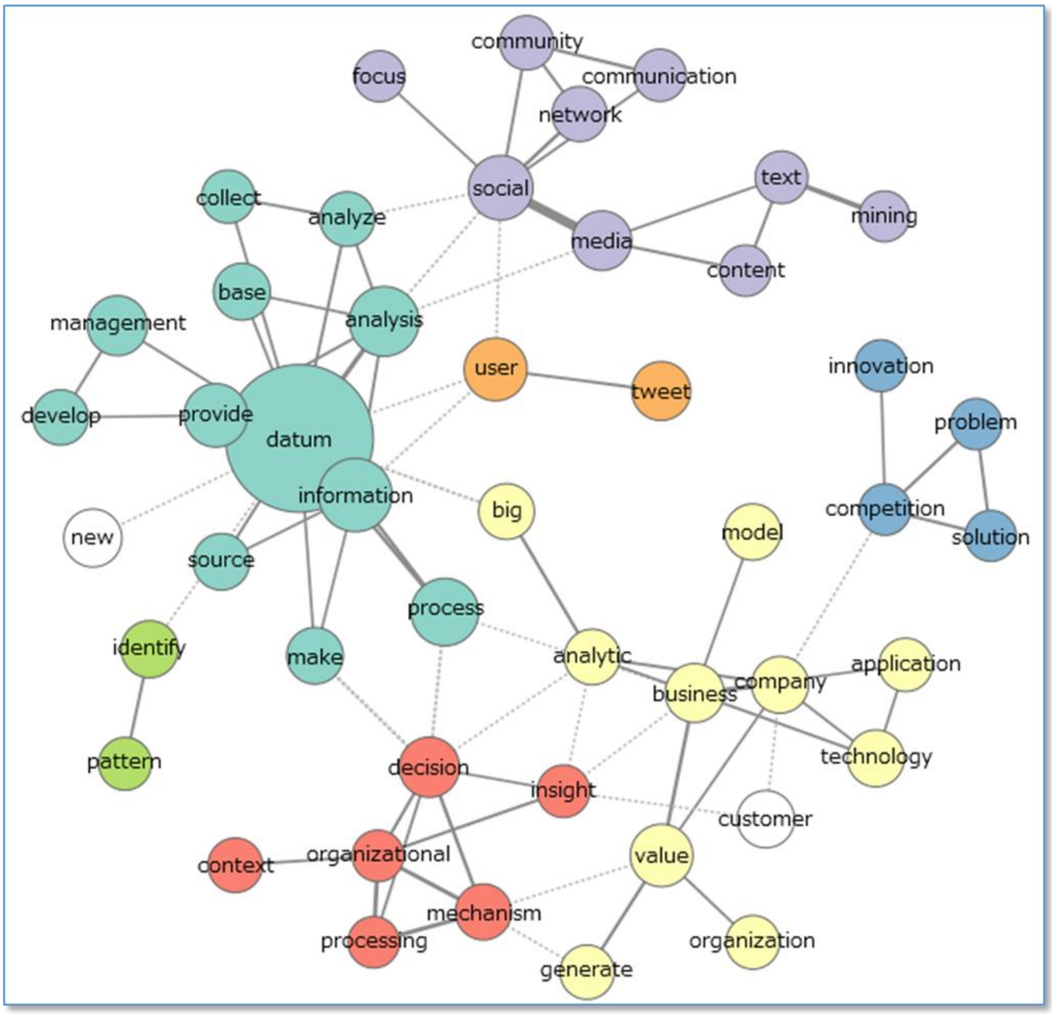

Figure 4. Co-occurrence map of the terms in big data case study articles.

The three business (or data usage) -related themes are:

- **Decision-making** (red color in the map). Several studies discussed enhancing the decision-making processes, enabling data-driven decision-making, or providing actionable insights to managers (Bärenfänger et al., 2014; Dutta and Bose, 2015; Krumeich et al., 2014). Several studies (Cai et al., 2014; Tao et al., 2014; Kalakou et al., 2015) also investigated transportation or passenger patterns, providing insights into planning and decision-making. Embedding analytics and insights into processes and decision-making routines is important (Bekmamedova and Shanks, 2014). However, according to the case studies, there are challenges to overcome in this area, such as lack of data-driven organizational culture (Shen and Varvel, 2013; Dutta and Bose, 2015), missing analytics strategy, and lack of leadership (Phillips-Wren and Hoskisson, 2015).
- **Innovation** (blue). Big data was seen as an enabler for data-driven innovation and faster innovation cycles (Amatriain, 2013; Jetzek et al., 2014). In addition, (Martinez and Walton, 2014) reported successful and cost-efficient usage of crowd-sourced big data analytics, and (Ciulla et al., 2012) used social media data to predict the winner of a song contest.
- **Business value** (light yellow). According to the studies, big data is a vehicle to create new value. The studies recognized positive results and opportunities, such as a business

model that was based on big data (Amatriain, 2013), energy and cost savings (Dobson et al., 2014; Jetzek et al., 2014; Mathew et al., 2015), business transformation (Prescott, 2014), increased revenue and customer satisfaction (Dutta and Bose, 2015), better transparency over operations (Bärenfänger et al., 2014), generating value by secondary use of data (Bettencourt-Silva et al., 2015), and deeper understanding of real events (Crampton et al., 2013; Hu et al., 2014). The other side of this coin is that there are challenges related to the technical themes.

The two ICT-related themes cover data and analytics, new data sources, and data management aspects.

- **Data management** (cyan) through the whole lifecycle of data, from the sources to the analytics, is a central aspect. In general, the volume, variety, and velocity of big data can be challenging for data management and technology (Laney, 2001). Companies are experimenting with new technologies (Bärenfänger et al., 2014). Some studies mentioned that managing the volumes of data is a key challenge (Krumeich et al., 2014; Dutta and Bose, 2015). Moreover, the case studies pointed out additional aspects that need to be addressed, such as data inconsistencies and poor data quality (O’Leary, 2013a; Halamka, 2014; Mathew et al., 2015). Several studies also reported concerns for potential security and/or privacy issues (Halamka, 2014; Martinez and Walton, 2014; Stephansen and Couldry, 2014; Bettencourt-Silva et al., 2015). Applying proper analytics to the vast amounts of data is the key in gaining value and insights. New data types, such as social media posts or text documents, require new kinds of analytics. This is a multi-faceted issue: in addition to new technology, organizations need new talent, both business-oriented and technology-skilled (Shen and Varvel, 2013; Phillips-Wren and Hoskisson, 2015; Prinsloo et al., 2015).
- **New data sources** (purple). In several cases organizations utilized data from outside their own organization, such as Facebook and Twitter data (He et al., 2013), blog texts and user reviews (Marine-Roig and Clavé, 2015), or data collected from mobile apps (O’Leary, 2013a; Papenfuss et al., 2015). They had been able to extract value from these external sources. The data are freely available, but requires quite a lot of processing, as described e.g. in (Marine-Roig and Clavé, 2015).

In addition to the five main themes, the map in figure 4 shows a few sub-themes. The Pattern-identify theme is related to data. The KWIC-analysis showed that these keywords were mostly used when the articles discussed revealing patterns from data. The User-tweet theme rose from articles in which Twitter-analyses were discussed. The keyword 'new' was used in various contexts, but mostly in conjunction with data. Accordingly, the keyword 'customer' was mostly used in contexts that discussed a company’s customer insights.

The map also shows several dotted lines between the words that belong to different themes. This indicates that the themes are inter-related. For example, there are several relations between the nodes that belong to decision-making, business value and data management themes. Concrete indications of these relations are the challenges regarding decision-making and data management. The linkages reflect the disruptive impact of big data and the inevitable business transformation process. The case study articles provide minimal information on *how* to solve the challenges, which opens new research avenues like those listed in the conclusions section. Data management theme inter-relates also with new data sources theme. This is intuitively obvious. However, as many organizations lack the required analytical and technical capabilities, they will turn to external vendors, and new data or analytics related services will emerge.

4. Discussion and Lessons Learned

The themes we discovered pointed out three essential business aspects – decision-making, innovation, and business value – related to big data. Regarding these themes, many of the cases reported positive results. However, to meet the big data value proposal discussed in the Introduction section, business transformation and new business models are required. We could identify in the articles one case where the *business model was based on big data* (Amatriain, 2013): Netflix runs their business based on the data they collect, and boosts their sales by making customer-specific, data-driven recommendations. One case study (Prescott, 2014) reported a *business transformation process* leading to re-gaining competitive advantage. There was also one case where a company-wide data-driven approach was taken (Dutta and Bose, 2015). In this case, a large steel manufacturer reshaped their processes and functions to take advantage of data, which resulted in significant business benefits. On the other hand, they also faced challenges, such as organizational resistance towards the change. The rest of the cases were more function-specific, limited-scope initiatives that brought benefits to certain operations, e.g. marketing, or social media -related experiments. Several cases, e.g. (Crampton et al., 2013; Cheng and Chen, 2014; Hu et al., 2014) had analyzed social media data in order to identify signs or clues of e.g. raising trends or other emerging actions. For a discussion of Internet of Signs, see (O'Leary, 2013b). One aspect in data-driven innovation is secondary use of data, which means that the data are used to another purpose than it was originally collected for. Some of the sources, e.g. (Mayer-Schönberger and Cukier, 2013) claimed that the *secondary usage of data* has huge potential. We identified one case (Bettencourt-Silva et al., 2015), where this kind of data usage was clearly recognized and utilized. Of course, these findings must be compared against the fact that most of the organizations were taking their first steps on the big data path.

Technology and software vendors typically emphasize the business aspects, and especially their positive effects. As our results point to the same direction, our study confirms the hype partly. What the hype typically leaves out is that changing the organization to a more data-driven one will have effects on the organizational culture and decision-making processes, as the challenges related to the decision-making theme indicate. Moreover, several studies reported technical challenges, especially with the data volumes.

Table 2 synthesizes the findings of our analysis. The examples column includes examples of the articles related to the theme. The case studies showed that the value proposal of big data is significant. However, realizing the value is much more a business transformation initiative than a technical issue. Organizations need to consider these aspects carefully in their big data experiments.

Table 2. Guidelines for big data utilization.

Theme	Guidelines	Examples
Decision-making	• Embed analytics into decision-making processes. • Be prepared for organizational side effects.	(Bekmamedova and Shanks, 2014), (Cai et al., 2014), (Dutta and Bose, 2015), (Phillips-Wren and Hoskisson, 2015)
Innovation	• Trust the data. • Search for new methods.	(Amatriain, 2013), (Jetzek et al., 2014), (Martinez and Walton, 2014), (Ciulla et al., 2012)
Business value	• Look for value in various directions; experiment with the data.	(Amatriain, 2013), (Bettencourt-Silva et al., 2015), (Dutta and

	• Enable business transformation with the data. • Consider secondary usage of the data.	Bose, 2015), (O’Leary, 2013a), (Prescott, 2014)
Data management	• Expect to face technical and data-related challenges. • Plan for security.	(Dutta and Bose, 2015), (Halamka, 2014), (Krumeich et al., 2014), (Prinsloo et al., 2015), (Shen and Varvel, 2013)
New data sources	• Experiment with new data types. • Consider potential privacy issues.	(He et al., 2013), (Marine-Roig and Clavé, 2015), (Yu et al., 2014)

Data and analytics should be embedded into the decision-making processes. Taking advantage of analytic software suggestions and decision support information should be a habit in a data-driven organization. This is possible only if the information is easily available in the normal decision-making context. A recent study suggests that tight integration to enterprise systems is a success factor to business intelligence solutions (Isik et al., 2011). However, this can be a difficult task. For example, middle management and specialists make important operative decisions. Although the cases did not discuss this matter, it is obvious that embedding analytics into their working context and to legacy systems can be a complex and expensive task. It would require significant changes to legacy systems, often combining new and old technologies. Another aspect to consider is that the organizational side effects of the data-driven approach can be significant. Several of the case studies reported challenges in this area. In order to gain benefits, a data-driven organizational attitude is required, but the organizational culture often hinders the change. In addition, utilizing data may lead to changes in the decision-making processes. Managers need to not only understand but also support these changes. Managing the change and the organizational side effects requires training and new managerial skills.

The data-driven innovation method requires rapid testing of many new hypotheses and ideas, gathering data from the tests and – most importantly – relying on the data that results from the tests. This kind of process is described in (Amatriain, 2013). Netflix runs several tests simultaneously in order to improve their services. Although in this case the services are digital, the principle is general. Instead of concentrating on finalizing one solution at a time, a better approach might be to test several primitive prototypes with the customers at the same time. The feedback would help to improve the solution, to ensure that the solution really is something that the customers need, and to speed up the innovation pace (Furr and Dyer, 2014). However, relying on data and an experimental, more customer-centric innovation method requires the organizational culture to allow mistakes and uncertainty. Many ideas simply do not make it, and the more disruptive the idea, the more difficult it is to calculate the business case.

The business value of big data value potential is case-dependent. According to our analysis of the case studies, big data can drive business value and innovation. The cases reported various opportunities in different areas. However, the opportunities are case-dependent, so each organization must do their thinking in order to find out how to add value with data. What is the business problem that we are trying to solve with big data? One important aspect to consider is the secondary usage of data. As organizations generate and harvest more and more data, opportunities will open to utilize the data in new, unexpected ways that can generate value. For example, a factory that must collect real-time emission data for regulatory purposes might be able to use the same data for another purpose, such as process monitoring.

The data management challenge of big data is real. Several of the studies reported significant issues with data volumes. New technologies are rapidly emerging, and organizations should be able to integrate these into their current infrastructure. This requires architectural and technical talent, money, and company policies that allow new vendors and technologies to enter to the playground. Security issues are obvious: where there is value, there is a potential fraud. Data protection must be secured from the source to the presentation. Security must be planned and built into the systems. Many of the case studies recognized potential problems in these areas. The case studies also recognized challenges in data quality and the shortage of analytic capabilities. These are partly technical issues, but they also require business talent.

New data sources can provide value. Several of the case studies mined out value from tweets or other textual data, as did we. Our own experiments with the computerized content analysis suggested that appropriate software tools are efficient and cost-effective (compared to manual coding). This makes content analysis a viable option also for practitioners. The main caveat here is that text analysis requires knowledge in the theory and methods of content analysis. Another consideration are the tools. According to (Isik et al., 2011), users are dissatisfied with external data capabilities of current tools. However, integrating new, external data sources would also improve user satisfaction. From the privacy point of view, combining analytics and data from several sources can lead to unpredicted privacy issues. Companies must consider the public opinion as well as the governing policies and legislation.

5. Conclusions

Several studies, e.g. (Manyika et al., 2011; Mayer-Schönberger and Cukier, 2013; Davenport, 2014) have made claims that big data causes pervasive changes, which will affect almost every sector of life. In this study, we analyzed 33 peer-reviewed papers describing 49 big data use cases. The cases confirmed the claims, at least partly. Clearly, big data applications are emerging in various areas of life. The studies recognized positive results and opportunities, such as new business models, energy and cost savings, cost-efficient open innovation, business transformation, or deeper understanding of real events for decision support. Previous research, like McAfee and Brynjofsson (2012), have shown that data driven decisions add value to the business. Our research used a different methodology and a different research set, but the results point to the same direction, supporting the results of the previous research.

However, several studies also reported of challenges like data inconsistencies and poor data quality, security and/or privacy issues, missing analytics strategy, lack of leadership, lack of data-driven organizational culture, and the need of new analytics and technology skills. These challenges reflect the disruptive nature of big data. They are indications of major shifts required; changes that affect not only technical platforms and skills, but they also – and more importantly – influence the organizational culture, decision-making processes and management functions. Previous studies discuss many of these challenges in general level. Based on current big data implementations as described on peer-reviewed literature, our study adds insights at more concrete level, providing practitioners best practices and guidance to avoid common pitfalls.

We used computerized content analysis to extract knowledge from the raw text of the case study papers. Using the computerized approach with open-source tools enables organizations to experiment with text analysis. The results of the computer analysis must be processed further and proved to be useful. We interpreted the results of a co-occurrence map to five named themes and verified the results against case study papers. These insights enabled us to conceptualize the findings to a set of guidelines (see Table 2) that point out several essential aspects that organizations must consider in their big data experiments. These guidelines emphasize that dealing with big data is a complicated task, which requires addressing technical, business-related and organizational issues.

In this research we created a set of guidelines stating *what* organizations should consider when dealing with big data. Another viewpoint is *how* to tackle the topics. This is an important question especially for practitioners. However, only a few articles discussed the case studies on a detailed enough level to answer the *how* question. This opens new research avenues. We point out some of these avenues below.

For researchers focusing on big data topics in the business context, this study offers a collection of big data case studies to start with, and several possibilities for further research. In addition to several technical questions, there are many open questions related to business transformational effects of big data, including the following.

- Understanding business transformation processes behind digitalization and big data. How does datafication drive the change in different industries? How can an organization adapt to the changes in industry structures and ecosystems? What is required to manage the change effectively?
- What are the effects of big data on the decision-making processes of the organization? What organizational effects does this have? How should an organization integrate big data analytics effectively to the existing business processes and workflows?
- How does big data enable innovation? What are the driving forces behind the new, data-based innovation processes? How should an organization arrange its innovation method to be effective in the big data era?
- What methods and processes are efficient when organizations start to explore big data? How do the existing infrastructure and company policies match with big data experimenting? How could companies evaluate various options quickly in order to decide which of them are promising, and what kind of risks they contain?

6. References

Amatriain, X., 2013. Beyond data: from user information to business value through personalized recommendations and consumer science. *Proceedings of the 22nd ACM international conference on information & knowledge management* 2201–2208.

Bärenfänger, R., Otto, B., Österle, H., 2014. Business value of in-memory technology–multiple-case study insights. *Industrial Management & Data Systems* 114, 1396–1414.

Bekmamedova, N., Shanks, G., 2014. Social Media Analytics and Business Value: A Theoretical Framework and Case Study. *System Sciences (HICSS), 2014 47th Hawaii International Conference on* 3728–3737.

Berelson, B., 1952. Content analysis in communication research. US Free Press, New York.

Bettencourt-Silva, J.H., Clark, J., Cooper, C.S., Mills, R., Rayward-Smith, V.J., De La Iglesia, B., 2015. Building Data-Driven Pathways From Routinely Collected Hospital Data: A Case Study on Prostate Cancer. *JMIR medical informatics* 3, 1–21.

Cai, H., Jia, X., Chiu, A.S., Hu, X., Xu, M., 2014. Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet. *Transportation Research Part D: Transport and Environment* 33, 39–46.

Cheng, Y.-C., Chen, P.-L., 2014. Global social media, local context: A case study of Chinese-language tweets about the 2012 presidential election in Taiwan. *Aslib Journal of Information Management* 66, 342–356.

Christensen, C., 2013. The innovator’s dilemma: when new technologies cause great firms to fail. Harvard Business Review Press.

Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., Vespignani, A., 2012. Beating the news using social media: the case study of American Idol. *EPJ Data Science* 1, 1–11.

Clauset, A., Newman, M.E., Moore, C., 2004. Finding community structure in very large networks. *Physical review E* 70, 1–6.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W., Zook, M., 2013. Beyond the geotag: situating "big data"and leveraging the potential of the geoweb. *Cartography and geographic information science* 40, 130–139.

Davenport, T., 2014. Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press.

De Mauro, A., Greco, M., Grimaldi, M., 2015. What is big data? A consensual definition and a review of key research topics, in: AIPConferenceProceedings. pp. 97–104.

Dehning, B., Richardson, V.J., Zmud, R.W., 2003. The value relevance of announcements of transformational information technology investments. *Mis Quarterly* 27, 637–656.

Dobson, G., Tilson, D., Tilson, V., Haas, C.E., 2014. Quantitative case study: Use of pharmacy patient information systems to improve operational efficiency. *System Sciences (HICSS), 2014 47th Hawaii International Conference on* 4220–4228.

Dutta, D., Bose, I., 2015. Managing a Big Data project: The case of Ramco Cements Limited. *International Journal of Production Economics* 165, 293–306.

E. Prescott, M., 2014. Big data and competitive advantage at Nielsen. *Management Decision* 52, 573–601.

Fang, S., Da Xu, L., Zhu, Y., Ahati, J., Pei, H., Yan, J., Liu, Z., 2014. An integrated system for regional environmental monitoring and management based on internet of things. *Industrial Informatics, IEEE Transactions on* 10, 1596–1605.

Fruchterman, T.M., Reingold, E.M., 1991. Graph drawing by force-directed placement. *Softw., Pract. Exper.* 21, 1129–1164.

Furr, N., Dyer, J., 2014. The Innovator's Method. Harvard Business Review Press.

Gantz, J., Reinsel, D., 2011. Extracting value from chaos (No. 1142), IDC iview. IDC.

Halamka, J.D., 2014. Early Experiences with big data at an academic medical center. *Health affairs* 33, 1132–1138.

He, W., Zha, S., Li, L., 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management* 33, 464–472.

Holsti, O.R., 1969. Content analysis for the social sciences and humanities. Addison-Wesley.

Hu, H., Ge, Y., Hou, D., 2014. Using web crawler technology for geo-events analysis: A case study of the Huangyan Island incident. *Sustainability* 6, 1896–1912.

Isik, O., Jones, M.C., Sidorova, A., 2011. Business intelligence (BI) success and the role of BI capabilities. *Intelligent systems in accounting, finance and management* 18, 161–176.

Jetzek, T., Avital, M., Bjorn-Andersen, N., 2014. Data-driven innovation through open government data. *Journal of theoretical and applied electronic commerce research* 9, 100–120.

Kalakou, S., Psaraki-Kalouptsidi, V., Moura, F., 2015. Future airport terminals: New technologies promise capacity gains. *Journal of Air Transport Management* 42, 203–212.

Kitchenham, B., 2007. Guidelines for performing systematic literature reviews in software engineering (No. EBSE-2007-01), Technical report, Ver. 2.3 EBSE Technical Report. EBSE. Keele University.

Kolowitz, B.J., Lauro, G.R., Venturella, J., Georgiev, V., Barone, M., Deible, C., Shrestha, R., 2014. Clinical Social Networking—A New Revolution in Provider Communication and Delivery of Clinical Information across Providers of Care? *Journal of digital imaging* 27, 192–199.

Kowalczyk, M., Buxmann, P., 2014. Big Data and Information Processing in Organizational Decision Processes. *Business & Information Systems Engineering* 6, 267–278.

Krippendorf, K., 1989. Content analysis. *International ensyclopedia of communication* 1, 403–407.

Krumeich, J., Jacobi, S., Werth, D., Loos, P., 2014. Towards planning and control of business processes based on event-based predictions, in: BusinessInformationSystems. Springer, pp. 38–49.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. *META Group Research Note* 6, 70.

Lewis, S.C., Zamith, R., Hermida, A., 2013. Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media* 57, 34–52.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Marine-Roig, E., Clavé, S.A., 2015. Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management*. doi:http://dx.doi.org/10.1016/j.jdmm.2015.06.004i

Martinez, M.G., Walton, B., 2014. The wisdom of crowds: The potential of online communities as a tool for data analysis. *Technovation* 34, 203–214.

Mathew, P.A., Dunn, L.N., Sohn, M.D., Mercado, A., Custudio, C., Walter, T., 2015. Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy* 140, 85–93.

Mayer-Schönberger, V., Cukier, K., 2013. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., 2012. Big data: The Management Revolution. *Harvard Business Review* 90, 61–67.

O'Leary, D.E., 2013a. Exploiting big data from mobile device sensor-based apps: Challenges and benefits. *MIS Quarterly Executive* 12, 179–187.

O'Leary, D.E., 2013b. Big Data, the Internet of Things and Internet of Signs. *Intelligent Systems in Accounting, Finance and Management* 20, 53–65.

Osgood, C.E., 1959. The representational model and relevant research methods. *Trends in content analysis* 33–88.

Papenfuss, J.T., Phelps, N., Fulton, D., Venturelli, P.A., 2015. Smartphones Reveal Angler Behavior: A Case Study of a Popular Mobile Fishing Application in Alberta, Canada. *Fisheries* 40, 318–327.

Phillips-Wren, G., Hoskisson, A., 2015. An analytical journey towards big data. *Journal of Decision Systems* 24, 87–102.

Porter, M.E., Millar, V.E., 1985. How information gives you competitive advantage.

Prescott, M., 2014. Big data and competitive advantage at Nielsen. *Management Decision* 52, 573–601.

Prinsloo, P., Archer, E., Barnes, G., Chetty, Y., Van Zyl, D., 2015. Big (ger) data as better data in open distance learning. *The International Review of Research in Open and Distributed Learning* 16.

Ryan, G.W., Bernard, H.R., 2003. Techniques to identify themes. *Field methods* 15, 85–109.

Sainio, L.-M., 2005. The Effects of Potentially Disruptive Technology on Business Model - A Case Study of New Technologies in ICT Industry. Lappeenranta University of Technology.

Shen, Y., Varvel, V.E., 2013. Developing data management services at the Johns Hopkins University. *The Journal of Academic Librarianship* 39, 552–557.

Stemler, S., 2001. An overview of content analysis. *Practical assessment, research & evaluation* 7, 137–146.

Stephansen, H.C., Couldry, N., 2014. Understanding micro-processes of community building and mutual learning on Twitter: a "small data"approach. *Information, Communication & Society* 17, 1212–1227.

Tao, S., Corcoran, J., Mateo-Babiano, I., Rohde, D., 2014. Exploring Bus Rapid Transit passenger travel behaviour using big data. *Applied Geography* 53, 90–104.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* 173–180.

UnitedNations, 2008. International Standard Industrial Classification of All Economic Activities. United Nations.

Venkatraman, N., 1994. IT-enabled business transformation: from automation to business scope redefinition. *Sloan management review* 35, 73–73.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G., Gnanzou, D., 2015. How "big data"can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics* 165, 234–246.

Weber, R.P., 1990. Basic content analysis. Sage.

Wehn, U., Evers, J., 2015. The social innovation potential of ICT-enabled citizen observatories to increase eParticipation in local flood risk management. *Technology in Society* 42, 187–198.

Wilbur, W.J., Sirotkin, K., 1992. The automatic identification of stop words. *Journal of information science* 18, 45–55.

Yang, Y., Pedersen, J.O., 1997. A comparative study on feature selection in text categorization, in: ICML. pp. 412–420.

Yang, Y., Wilbur, J., 1996. Using corpus statistics to remove redundant words in text categorization. *JASIS* 47, 357–369.

Ylijoki, O., Porras, J., 2016. Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management (in press).*

Yu, K., Zhang, J., Chen, M., Xu, X., Suzuki, A., Ilic, K., Tong, W., 2014. Mining hidden knowledge for drug safety assessment: topic modeling of LiverTox as a case study. *BMC bioinformatics* 15, S6.

Appendix 1 – Big Data Case Study Articles

Table 1 is a summary of the case study articles we analyzed. The table contains 33 different studies and 49 big data cases (due to 3 multi-case studies) representing five continents. We identified academic, peer-reviewed articles in major literature databases (ProQuest, SCOPUS, Web-of-Science, and EBSCO) covering business and technical topics at the end of August 2015.

The context column describes the focus area of the study. Application area is the categorization of the case(s) that the article reports, according to ISIC classification (UnitedNations, 2008). Country is the origin of the organization subject to the study.

Appendix 1 – Table 1. Big data case study articles.

Paper	**Context**	**Application area (ISIC)**	**Country**
(Amatriain, 2013)	Netflix recommender system.	J-Information and communication	USA
(Bekmamedova and Shanks, 2014)	Marketing campaign using social media.	K-Financial and insurance activities	Australia
(Bettencourt-Silva et al., 2015)	Secondary usage of routinely collected patient data.	Q-Human health and social work activities	UK
(Bärenfänger et al., 2014)	In-memory computing business value assessed in 5 large European companies from different industries.	C-Manufacturing (2) G-Wholesale and retail D-Electricity H-Transportation	n/a (Europe)
(Cai et al., 2014)	Taxi trajectory data used to reveal travel patterns in order to help the planning of public charging infrastructure.	H-Transportation	China
(Cheng and Chen, 2014)	Analysis of Twitter communities during the presidential election in Taiwan in 2012.	O-Public administration	Taiwan
(Ciulla et al., 2012)	Predicting the American Idol competition results by using Twitter analysis.	R-Arts, entertainment and recreation	USA
(Crampton et al., 2013)	Social and spatial analysis of geotagged tweets following the 2012 NCAA championships.	R-Arts, entertainment and recreation	USA
(Dobson et al., 2014)	Cost reductions in a hospital by process analytics.	Q-Human health and social work activities	USA
(Dutta and Bose, 2015)	Big data initiative in a manufacturing company.	C-Manufacturing	India
(Fang et al., 2014)	An integrated system for monitoring regional environmental data (collecting, storing and analyzing temperature-related data).	M-Professional, scientific and technical activities	China
(Martinez and Walton, 2014)	By adopting a crowdsourcing approach to data analysis (using Kaggle), Dunnhumby were able to extract information from their own data that was previously unavailable to them.	G-Wholesale and retail	UK
(Halamka, 2014)	Analysis and experiences of new big data possibilities and challenges in a hospital.	Q-Human health and social work activities	USA
(He et al., 2013)	Social media marketing in the pizza industry.	G-Wholesale and retail	USA
(Hu et al., 2014)	The Huangyan Island incident was studied by using a web crawler technology and text analysis.	M-Professional, scientific and technical activities	(Huangyan Island, South China Sea)
(Jetzek et al., 2014)	Case Opower: generating value from open data. Saving energy by offering benchmark information to consumers.	D-Electricity, gas, steam and air conditioning supply	USA
(Kalakou et al., 2015)	Simulation for planning airport terminals and reducing passenger check-in and security	H-Transportation and storage	Portugal

Paper	Context	Application area (ISIC)	Country
	checkpoint times, using Lisbon airport as the case.		
(Kolowitz et al., 2014)	Using social technologies to construct dynamic provider networks, simplify communication, and facilitate clinical workflow operations.	Q-Human health and social work activities	USA
(Kowalczyk and Buxmann, 2014)	Multi-case study, 12 big companies from various industries.	J-Information and communication (2) K-Financial and insurance activities (3) G-Wholesale and retail (2) I-Accommodation and food service activities H-Transportation and storage (2) Q-Human health and social work activities C-Manufacturing	n/a
(Krumeich et al., 2014)	Big data experiments and challenges of a steel factory.	C-Manufacturing	Germany
(Lewis et al., 2013)	A case of news sourcing on Twitter combining text mining and manual methods.	J-Information and communication	USA
(Marine-Roig and Clavé, 2015)	Tourism and city strategy planning and marketing in the Barcelona region by using big data analytics.	N-Administrative and support service activities	Spain
(Mathew et al., 2015)	Case study of the largest database of building energy data in US; aiming at enabling energy savings.	F-Construction	USA
(O'Leary, 2013a)	A mobile device application collecting data that the city of Boston uses to facilitate road infrastructure management.	H-Transportation and storage	USA
(Papenfuss et al., 2015)	Analyzing behavioral patterns in fishing by using mobile app -generated data.	A-Agriculture, forestry and fishing	Canada
(Phillips-Wren and Hoskisson, 2015)	Case Choice-hotels customer analytics (CRM, Twitter).	I-Accommodation and food service activities	USA
(E. Prescott, 2014)	Nielsen re-gaining their competitive advantage by using data and analytics.	H-Transportation and storage	USA
(Prinsloo et al., 2015)	Unifying and analyzing data (360 000 students, courses, programs etc.) at the University of South Africa (Unisa).	P-Education	South Africa
(Shen and Varvel, 2013)	New data management services platform implementation at Johns Hopkins University. Aims to increase data and knowledge sharing.	P-Education	USA
(Stephansen and Couldry, 2014)	A case study where a departmental Twitter account was used to create a community of practice (students and teachers) and to enable mutual learning beyond the classroom.	P-Education	UK
(Tao et al., 2014)	Big data visualization case in bus-rapid-transit in order to understand passenger travel dynamics and plan capacity.	H-Transportation and storage	Australia
(Wehn and Evers, 2015)	Planning and managing flooding situations, 2 cases	F-Construction	UK Netherlands
(Yu et al., 2014)	Text mining (topic modeling) applied to text documents in order to improve drug safety by finding drugs susceptible to acute liver failures.	Q-Human health and social work activities	USA