Ossi Ylijoki

BIG DATA – TOWARDS DATA-DRIVEN BUSINESS

ACTA UNIVERSITATIS LAPPEENRANTAENSIS 845

Ossi Ylijoki

BIG DATA – TOWARDS DATA-DRIVEN BUSINESS

Dissertation for the degree of Doctor of Philosophy to be presented with due permission for public examination and criticism in the Auditorium of Lahden Musiikkiopisto at Lahti, Finland on the 12th of April 2019, at noon.

Acta Universitatis
Lappeenrantaensis 845

Supervisors Professor Jari Porras
LUT School of Engineering Science
Lappeenranta-Lahti University of Technology LUT
Finland

Professor Vesa Harmaakorpi
LUT School of Engineering Science
Lappeenranta-Lahti University of Technology LUT
Finland

Reviewers Professor Nina Helander
Department of Industrial and Information Management
Tampere University of Technology
Finland

Docent Harri Jalonen
Department of Engineering and Business, Business Administration
Turku University of Applied Sciences
Finland

Opponent Professor Pauli Kuosmanen
Director, platforms and networks
Tampere University
Finland

ISBN 978-952-335-346-6
ISBN 978-952-335-347-3 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Lappeenranta-Lahti University of Technology LUT
LUT University Press 2019

Abstract

Ossi Ylijoki
Big Data – Towards Data-driven Business
Lappeenranta 2019
124 pages
Acta Universitatis Lappeenrantaensis 845
Diss. Lappeenranta-Lahti University of Technology LUT
ISBN 978-952-335-346-6, ISBN 978-952-335-347-3 (PDF), ISSN-L 1456-4491, ISSN 1456-4491

This research stems from the disruptive phenomenon known as digital transformation, i.e. the pervasive use of digital technologies in order to add value in business. As a side effect of digital transformation, vast amounts of various types of data are generated at a fast pace. This data is known as big data. The data is the root source of added value that businesses look for. Big data represents first and foremost a major paradigm shift, a new way to view businesses, enabled by related technology. However, the paradigm shift towards data-oriented business models and processes is challenging to incumbent enterprises. The aim of this research is to help incumbents to move towards data-driven business models and processes.

"The world is one big data problem."
– Andrew McAfee

This dissertation is based on articles published in scientific journals. The articles are presented in the Publications section of this dissertation. Each of the articles applied different research methods, which is justified by the fact that big data is an emerging concept. The approaches included a literature review, a survey and a case study, as well as algorithmic approaches. Together the articles explore the big data landscape from several angles, both from the theoretical and practical viewpoints.

The results can be viewed as a high-level framework that addresses the primary research question – understanding and utilising big data in the transformation process towards big data driven business – by explaining the phenomenon as well as the value creation processes and connecting theoretical aspects to practice. The theoretical foundations of this dissertation combine strategic management, data-driven innovations and big data in a way that helps to understand the digital transformation process.

This dissertation explains, how big data value creation mechanisms work. It helps to understand the nature of the big data phenomenon and provides building blocks and guidance for practitioners. The results suggest that big data must be seen as a business initiative instead of technological matter and strengthen the perception that big data in general and data-driven innovation in particular are potential sources of added value.

Keywords: thesis, dissertation, big data, digital transformation, datafication, digitalisation, business transformation, business value

Acknowledgements

Exploring the field of big data and writing this dissertation has been a wonderful, once-in-a-lifetime experience. It has been a journey I would never change. There have been many rewarding moments, such as the acceptance of my first academic article or simply having a small break-through idea after hours or days of "researcher's despair". Along the way many individuals have facilitated the process. In addition to individuals, several institutional organisations have helped me during my journey.

First, I would truly like to thank my principal supervisor, Professor Jari Porras whose sophisticated supervision I truly appreciate. The engagement and facilitation you provided were an incredibly important factor during my dissertation. All of the time I felt that I was in charge of the research, yet you were there when I needed support.

I am also extremely grateful to my co-supervisor, Professor Vesa Harmaakorpi, who believed in me in the first place. Professor Harmaakorpi helped my entrance into the academic world after many years in private business life.

This work was carried out in the School of Business and Innovation at Lappeenranta-Lahti University of Technology LUT, in Finland, between 2014 and 2019. I am grateful to the university for their organisational support.

I sincerely appreciate the financial support from Koulutusrahasto and the grants from Hämeen kauppakamari, Suomen Kulttuurirahasto, and Liikesivistysrahasto. The funding from these organisations greatly helped me to concentrate on the dissertation work.

I wish to acknowledge the role of the reviewers, Nina Helander and Harri Jalonen. Their valuable feedback helped to improve the quality of this manuscript. Moreover, the comments and insights gained from numerous anonymous reviewers greatly improved the quality of the articles included in the Publications section of this thesis.

I am also exceedingly grateful to my parents who undoubtedly have had an indirect influence on me taking this track by planting a positive attitude towards life-long learning. In addition, I would like to acknowledge the sage advice from my grandmother Hilja who taught me so much, among other things taking several viewpoints, developing an interest in understanding how things work and a "you can, just try harder" principle.

Last, but not least, I would like to thank my beloved wife Malla for her understanding and support during the process. This journey was a new stage for both of us that changed the dynamics of our lives. The positive attitude towards change is one of the values we share – and one of the reasons why I love you.

Ossi Ylijoki
April 2019
Lahti, Finland

To my grandma Hilja

Contents

Abstract

Acknowledgements

Contents

List of publications **13**

Terms and Definitions **15**

1 Introduction **17**
1.1 Background and Research Environment 17
1.2 Structure of the Dissertation 19
1.3 Objectives and Scope 20
1.3.1 Research Problem 21
1.3.2 Research Purpose 22
1.3.3 Research Questions 23
1.4 Research Approach 24
1.4.1 Research Philosophy 25
1.4.2 Research Design 26
1.4.3 Research Methods 27
1.5 Research Process 28
1.5.1 Data Collection 28
1.5.2 Data Analysis 29

2 Big Data **31**
2.1 Introduction to Big Data 31
2.1.1 Big Data Drivers 32
2.1.2 Data Sources and Types 35
2.1.3 Defining Big Data 39
2.1.4 Big Data Criticism 42
2.1.5 Challenges of Big Data 44
2.2 From Data to Actionable Insights 46
2.2.1 The Laws of Information 46
2.2.2 From Data to Knowledge 47
2.2.3 Big Data Value Generation 49
2.3 Big Data Impacts 53
2.3.1 Personal Level Impacts 53
2.3.2 Enterprise and Ecosystem Level Impacts 54
2.3.3 Society Level Impacts 55

3 Theoretical Foundation **59**
3.1 Strategic Management 59

3.2 Innovation Research 63
3.3 Information Systems Research 66

4 Research Contribution 71
4.1 Publication I – Current Interpretation of Big Data 71
4.2 Publication II – What Managers Think about Big Data 73
4.3 Publication III – Lessons Learned from Big Data Experiments 77
4.4 Publication IV – Driving Value with Innovations 81
4.5 Publication V – Driving Value with a Predictive Algorithm 84
4.6 Publication VI – A Recipe for Your Big Data Cook Book 85
4.7 Summary 89

5 Discussion 95
5.1 Theoretical and Practical Implications 95
5.2 Suggested Further Research 98
5.3 Reliability and Validity 100

6 Conclusions 103
6.1 Why Should Businesses Care about Big Data? 103
6.2 Where Should Businesses Look at to Avoid Pitfalls? 104
6.3 Who Should Be Involved in the Transformation? 105
6.4 How Should Big Data Be Positioned in Business and Research? 107

7 References 109

8 Publications

List of Figures:

Figure 1. An illustration of the paradigm shift of the business landscape. 18
Figure 2. Research overview. 20
Figure 3. Research approach components. 25
Figure 4. Example of text analysis tools of the KH Coder software. 30
Figure 5. The number of big data articles by subject area and year. 32
Figure 6. An exemplary illustration of Internet data volumes. 34
Figure 7. Global IP traffic growth forecast. 35
Figure 8. A visualisation of 3V definition of big data. 40
Figure 9. Virtual value creation process 49
Figure 10. Multidisciplinary approach of the research. 59
Figure 11. The VRIO tool. 60
Figure 12. Business canvas framework. 62
Figure 13. IT-enabled business transformation framework 63
Figure 14. Three forms of knowledge 64
Figure 15. Year-wise distribution of the found papers. 71
Figure 16. Evolution of the definition of big data. 72
Figure 17. The model explaining the big data intentions of executives. 75
Figure 18. Respondents' perceptions regarding big data by experience. 76
Figure 19. Co-occurrence map of the terms in the big data case study articles. 78
Figure 20. Research method of Publication IV. 81
Figure 21. Innovation as a mediator framework. 82
Figure 22. Algorithm follow-up principle. 84
Figure 23. AC/TC process – creating economic value with big data. 88
Figure 24. Research contribution from theory to practice viewpoint. 90

List of Tables:

Table 1. Information systems resources. 67
Table 2. A tool for evaluating IS resources with the RBV resource attributes 70
Table 3. T-test results – moderator effects on behavioural intentions. 75
Table 4. Big data case studies by application area. 77
Table 5. Guidelines for big data utilisation. 80
Table 6. Big data literature related to big data value creation. 86
Table 7. Research contribution as a whole in a nutshell. 91
Table 8. Tools for big data adoption. 93

List of publications

This thesis is based on the following articles. The rights have been granted by publishers to include the papers in dissertation.

I. Ylijoki, Ossi and Porras, Jari (2016). **Perspectives to Definition of Big Data: a Mapping Study and Discussion**. *Journal of Innovation Management*, Thematic Issue: "Boosting Innovation with Big Data", 4(1), pp. 69-91.

II. Ylijoki, Ossi and Porras, Jari (2018). **What Managers Think about Big Data**. *International Journal of Business Information Systems*. 29(4), pp. 485-501.

III. Ylijoki, Ossi and Porras, Jari (2016). **Conceptualising Big Data: Analysis of Case Studies**. *Intelligent Systems in Accounting, Finance and Management*, 23(4), pp. 295-310.

IV. Ylijoki, Ossi, Sirkiä, Jukka, Porras, Jari and Harmaakorpi, Vesa (2018). **Innovation Capabilities as a Mediator between Big Data and Business Model**. *Journal of Enterprise Transformation.* Accepted for publication. https://doi.org/10.1080/19488289.2018.1548396

V. Ylijoki, Ossi (2018). **Guidelines for Assessing the Value of a Predictive Algorithm - a Case Study**. *Journal of Marketing Analytics*, 6(1), pp. 19-26.

VI. Ylijoki, Ossi and Porras, Jari (2018). **A Recipe for Big Data Value Creation**. *Business Process Management Journal*. Accepted for publication. https://doi.org/10.1108/BPMJ-03-2018-0082

Author's contribution

Ossi Ylijoki is the principal author and investigator in papers I – VI.

Publication I. The author defined the research plan, gathered the articles for the review, performed the analysis of the articles and wrote the vast majority of the article.

Publication II. The author defined the research plan, designed and implemented the survey, analysed the results, wrote the vast majority of the article.

Publication III. The author defined the research plan, searched the articles, planned and conducted the content analysis process and wrote the vast majority of the article.

Publication IV. The author defined the research plan, selected the methods, planned and designed the framework, participated in the interviews and their analysis, and wrote the vast majority of the article.

Publication V. The author defined the research plan, designed and implemented the algorithm, wrote the article.

Publication VI. The author defined the research plan, searched the literature, planned and designed the framework and wrote the vast majority of the article.

Terms and Definitions

Term	Definition/explanation
Big data	Big data refers to high-volume, high-velocity and high-variety information assets.
Big data insights Big data analytics	Big data insights refer to analytics that are used to derive value from big data information assets. Big data analytics is used as a synonym.
Big data phenomenon	The big data phenomenon refers to the paradigm shift towards data-driven businesses and ecosystems.
Business model	A business model describes the rationale of how an organisation creates, delivers, and captures value (Osterwalder & Pigneur 2010).
Innovation	Innovation is a multi-stage process whereby organisations transform ideas into new/improved products, services or processes, in order to advance, compete and differentiate themselves successfully in their marketplace (Baregheh et al. 2009).
Data	Data refers to raw facts, or symbols, that "know nothing" (Ackoff 1989).
Digital trail	The digital trail is the data about a user that follows from the user's actions.
Digital transformation Digitalisation	The use of digital technologies to change a business model and provide new revenue and value-producing opportunities; it is the process of moving to a digital business (https://www.gartner.com/it-glossary/digitalization/). In this research, the term digitalization is used as a synonym.
Digitisation	The automation of existing manual and paper-based processes, moving from an analogue to a digital format.
Datafication	Information technology-driven sense-making process (Lycett 2013).
Information	Information is data that has been processed into a meaningful form for the problem at hand (Pigni et al. 2016).

Knowledge	Knowledge is the capacity to identify required problem-related information and to interpret it (Pigni et al. 2016).
Value	In this research, the impact to an (incumbent) enterprise, gained from investing in co-creation of (big) data assets and capabilities (Publication VI).

1 Introduction

"It is not necessary to change.
Survival is not mandatory."
– W. Edwards Deming

Digital transformation is a current megatrend that will have huge implications on society, industries, business ecosystems, companies and people. Some years ago sources such as (Gantz & Reinsel 2011; Manyika et al. 2011) predicted exponential growth in data volumes and huge potential benefits, including automated decision-making and the possibility to innovate new business models, products and services. Today, we can see that their predictions are becoming a reality, e.g. (Janssen et al. 2017; Roden et al. 2017). Digitalisation has already enabled enterprises to create new revenue streams. One example is the concept of predictive maintenance. Current technology enables a machine manufacturer, for example, to collect data from their installation base from all around the world, analyse the data in real-time, and identify potential forthcoming faults. This opens up several potential opportunities to create economic value. The manufacturer could use the collected data in their product development or they could offer customised services to the customer based on the analysis of how the customer currently uses the machine. There is also a significant value for customers. They avoid unexpected breaks in their production as the supplier is able to replace parts before they fail. This adds value to the customers, since unplanned breaks could affect their delivery time promises and costs.

1.1 Background and Research Environment

Digital transformation leads to data deluge. A machine manufacturer may collect billions of measurements from the sensors of their machines. In general, there are countless data sources that generate the data that we have termed big data. Big data refers to high-volume, high-velocity (real-time) and high-variety data. For a detailed discussion on big data, see Chapter 2. The number of mobile devices, especially smart phones is increasing, and this leads to the generation of location and other types of data. According to McAfee & Brynjolfsson (2012), "each of us is now a walking data generator".

Figure 1 shows a broad overview of the phenomenon behind this dissertation. Digital transformation is the megatrend behind the scenes, driving the paradigm towards a more data-driven economy. The transformation is ongoing (Weill & Woerner 2015), however, different industries and ecosystems are at different stages. In Figure 1, the dotted clouds represent ecosystems and industries at various stages in the transformation process. For example, the music industry has already gone through a "digitalisation storm" that has fundamentally changed the industry's value generation mechanisms and business models, whereas in the health care sector, the clouds are just arising on the horizon. Depending on strategic decisions, a single firm may position itself in different situations within its ecosystem as illustrated in the Figure 1.

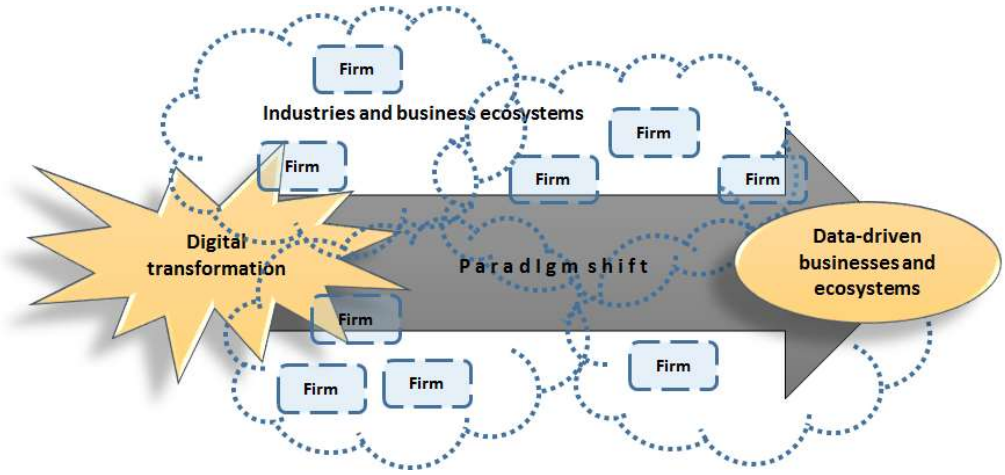


Figure 1. An illustration of the paradigm shift of the business landscape.

The root cause that drives digitalisation and thus the generation of big data is economic value (Mayer-Schönberger & Cukier 2013). Chapter 2 presents the concept of value and the value creation mechanisms of big data. For now, it is enough to say that the value potential of big data is huge. Various sources have estimated the significance and value of big data. For example, big data was number one on the list in a Finnish study (Linturi et al. 2013) that evaluated 100 of the most potential technological solutions for Finland. Manyika et al. (2011) claim, among other things, that big data represents an *annual* potential value of 250 billion euros in Europe's public sector administration and that retailers could increase their operating margins up to 60 % by utilising big data effectively.

The numbers speak for themselves, but they also indicate that big data is far more than a technological issue. It is unrealistic to assume that such huge numbers could be achieved just by applying a new technology to the current business models and processes. First and foremost big data is a major paradigm shift, a new way to view businesses, enabled by related technology, e.g. (Iivari et al. 2016). Enterprises must recognise relevant data-driven changes in their own business context. They must innovate and develop new capabilities in the changing environment. In other words, they must learn how to 1) adapt their business in the new situation and 2) how to develop their business models and operations in a way that leads to (sustainable) competitive advantage, e.g. (Iivari et al. 2016). In addition, as the predictive maintenance example shows, value comes in several form, which may be difficult to calculate by traditional means. These are not trivial tasks even in calm weather, not to speak of the turbulent, uncertain winds of digitalisation that we now sail in.

1.2 Structure of the Dissertation

This dissertation consists of two main parts: the introductory part and the publications containing Publications I-VI as appendices. The introduction sets the background. It positions the research and its objectives, provides a conceptual overview of big data as a phenomenon and lays out the theoretical foundations of the research. Moreover, it presents the key contributions of the research. The contributions come from the second part of the dissertation, as the Publications contain the key results of the research. Combined contributions from the articles provide answers to the research questions defined in the introduction part.

The first chapter of the introduction positions the research, defining the objectives and scope, the research approach and the research process. It introduces the purpose of the research – the main idea of this research – as well as specific questions that the dissertation addresses. It also presents the methods used, how the research process was organised, and discusses the background assumptions regarding the researcher's choices. In summary, this chapter helps the reader to understand *why this research was done, what the viewpoint of the research is, and how the research was conducted*.

The second chapter ("Big Data") discusses various aspects of big data, especially from the business value point of view. In order to understand how big data can be transformed into useful information, the underlying concepts and properties are presented. The chapter discusses questions such as what the driving forces are behind the data deluge, where all the data comes from, and what the consequences are. Moreover, the chapter briefly discusses big data criticism and some severe challenges that have been identified so far. As with any phenomenon, also the big data coin has two sides. After reading this chapter the reader should have a conceptual understanding of *what the big data phenomenon is, how big data creates value, and what the impact of big data is*. This helps in understanding the rest of the research.

Next, the theoretical foundations of the research are presented (chapter 3), followed by the key contributions of the research (chapter 4). This research draws its theoretical foundations from several disciplines. Each of the disciplines are introduced briefly, keeping in mind the applicability to this research and big data. The chapter "Research contribution" summarises the results from the Publications I-VI. The articles were published in peer-reviewed academic journals during the 2016 – 2018 period. By reading these chapters the reader should be able to understand the theoretical frame of the dissertation and the principles of how this research builds on previous research, as well as the contributions of this research.

The dissertation follows with a discussion chapter, which covers the theoretical and practical implications of the research. Moreover, the aspects of reliability and validity of the research are discussed as well as some thoughts about possible future research avenues. Finally, the dissertation ends with a conclusion. The conclusions are written as

a self-contained section that aims to give the reader a concentrated, "nutshell" view of this research.

1.3 Objectives and Scope

This section clarifies the need for the study, the intention of the study, and the specific issues that this dissertation addresses. The first section provided an overview of the challenges which the current paradigm shift towards data-driven business poses to enterprises. The research problem stems from this phenomenon. After discussing the research problem, including related factors and the research gap, the "Research purpose" -section identifies the major idea of the research as well as defines several key terms. Thus, it is perhaps the most important passage in positioning the research. Finally, narrowing the purpose statement, the specific research questions to be answered in this research are presented. Figure 2 presents an overview of the research: The research problem stems from the paradigm shift caused by digital transformation that leads to (big) data deluge. The purpose of the research is to explore the impact of big data on the context of a single firm. The research questions set the scope of the research by pinpointing the specific topics that this research addresses using a pragmatic approach. The following three sections ("Research Problem", "Research Purpose" and "Research Questions") discuss the matters shown in Figure 2 in detail.

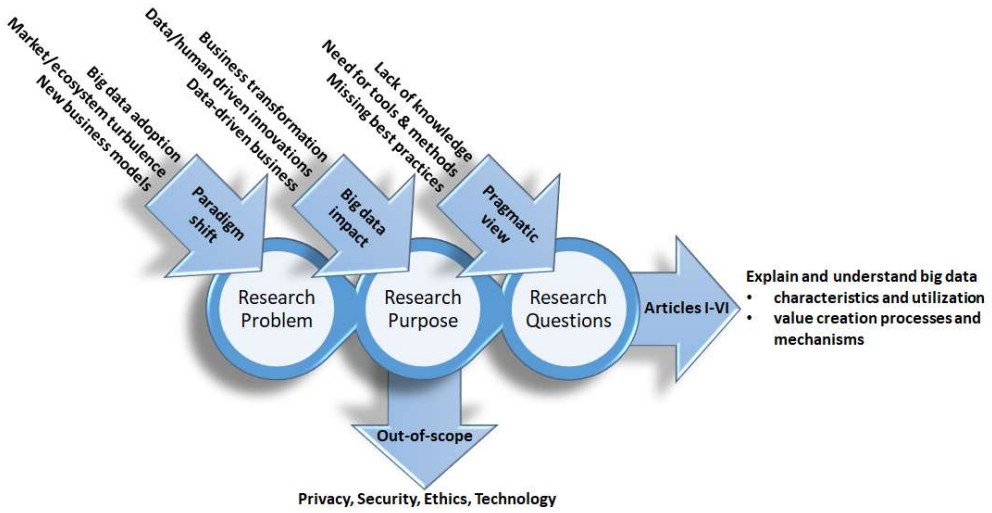


Figure 2. Research overview.

Most of the existing big data research is focused on specific, technology-related areas, such as hardware, data management software or analytics, which can be seen by looking recent surveys and reviews (Chen et al. 2014; Emani et al. 2015; Khan et al. 2014). Big

data literature also discusses topics such as security (Altshuler 2011; Berghel 2013), privacy (Eckhoff & Sommer 2014; Gehrke 2012; Stopczynski et al. 2014) or various organisational aspects (Dutta & Bose 2015; Martinez & Walton 2014; Mayer-Schönberger & Cukier 2013). All these areas are important aspects of big data. However, in order to stay focused, these topics are beyond the scope of this research. Each of the mentioned topics is wide and broad, and well worth separate studies.

1.3.1 Research Problem

The research problem stems from the *paradigm shift* caused by a disruptive phenomenon, digital transformation (see Figure 1 and Figure 2). The business landscape is becoming more and more uncertain. Most enterprises face the challenge of digitalisation in the years to come, if not already. The disruption of many current business models has already begun (Weill & Woerner 2015). Digitalisation, the related digital technologies and resulting vast volumes of data, big data, represent a major paradigm shift. The market turbulence is increasing and changes occur faster than ever (Mui & Carroll 2013).A changing environment poses not only risks, but opportunities as well. However, incumbent companies often fail to utilise new disruptive technologies as Christensen (2013) points out. Incumbents focus on serving their existing customers whereas new disruptive innovations tend to create new markets or disrupt existing business models. Global positioning system (GPS) technology created a huge market for location-based services and products; Über[1] introduced a new business model that has already disrupted incumbent taxi services in several countries.

Two things are common to these examples and apply in general as well. First, various new technologies enable innovation of new services. For this research however, the technologies as such are not interesting. This research focuses on the other common thing, data, and its utilisation in business context. The new services both create and heavily depend on data. As users consume the services, they generate vast volumes of various types of data. On the other hand, without real-time data many services are worthless. GPS based guidance requires a position in real-time; the Über service needs to operate in real-time in order to serve its customers.

As new business models and innovations emerge, incumbents must react. The minimum for a company is to avoid the major risks and to simply survive. Taking advantage of the opportunities in uncertain times is not an easy task. Incumbent companies have severe internal barriers against disruptive innovations, such as a restrictive mind-set, lack of discovery competencies and unsupportive organisational structures (Sandberg & Aarikka-Stenroos 2014). Moreover, incumbents often trust in their existing competencies in the new situation. Sainio (2005) pointed out that this is a false assumption that may lead to significant business risks. However, disruptive changes have always offered

[1] https://www.uber.com/en-FI/about/how-does-uber-work/

magnificent opportunities for new businesses as the rise of Internet and the electronic commerce explosion demonstrate.

One side of the problem is that under high levels of uncertainty old assumptions do not hold, and predicting the future becomes very difficult. Indeed, Christensen (2013) claims that with regard to disruptive innovations, any forecast should be deemed false. The Internet of Things (IoT) is a potentially disruptive concept that produces huge amounts of big data. Westerlund et al. (2014) discuss three major reasons that pose challenges regarding the development of an IoT business model: immature technologies, lack of standardisation, and evolving ecosystem participants. Looking at these reasons it is easy to understand the difficulties in forecasting. The same applies to big data. The Finnish Committee for the Future (Linturi et al. 2013) placed big data at top of the list of the 100 most promising disruptive technologies. But since the technologies, ecosystems and business models are still taking their first steps, only few enterprises can estimate the impact of big data in their own contexts. Others go blindfolded, they must react instead of engaging in proactive planning.

In a nutshell the research problem is as follows. The paradigm change towards data-oriented business models is challenging to incumbent enterprises. Old medicines do not work in the new situation; new ones must be developed. However, this is hard – if you do not understand the problem completely, it is hard to detect any solutions. Organisations will do their part in practice, but research needs to provide them with a deep understanding of the undercurrents beyond the hype, including the impacts of big data.

1.3.2 Research Purpose

The research approaches the research problem from the following angle. Convincing evidence exists claiming that data-driven businesses outperform their rivals (Dehning et al. 2003; McAfee & Brynjolfsson 2012; Porter & Millar 1985). It is clear that as big data adoption proceeds, the data-driven approaches become even more important. Various sources consider the potential of big data to be huge. Examples include (Davenport 2014; Linturi et al. 2013; Manyika et al. 2011; Mayer-Schönberger & Cukier 2013). The sources are unanimous that the *impact of big data* will affect almost every industry and every enterprise. The potential has been recognised also in the public sector, as governments have initiated big data strategies[2]. Most organisations, however, are not familiar with the data-driven approach, as e.g. Shen & Varvel (2013) point out. The fact that big data is a new, emerging concept underlines this. Moreover, current discussion is dominated by hardware and/or software vendors, who have their own interests. Scholars have

[2] E.g. European Big Data Value Strategic Research & Innovation Agenda:
http://www.nessi-europe.eu/Files/Private/EuropeanBigDataValuePartnership_SRIA__v099%20v4.pdf
U.S. Big data initiative:
https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

recognised the need for guidelines and frameworks for big data adoption (De Mauro et al. 2015; Fosso Wamba et al. 2015). Building on this background,

> the purpose of this research is to explore the big data landscape, cumulate the current big data theory, and provide new knowledge and tools that help enterprises innovate their transformation towards big data -driven businesses.

The purpose statement addresses the founding characteristics of big data and its implications on enterprises and ecosystems. Due to the novelty of the subject, many different interpretations exist regarding the most fundamental aspects of it, such as the definition of big data and its impact on various areas of life. In addition, this research aims to motivate enterprises to enter into transformation process. A clear understanding of the magnitude of change caused by digitalisation resulting in a big data deluge underlines the need of the transformation towards data-driven business.

The purpose statement also addresses the need for a systematic method for the business transformation process. As the digital transformation proceeds and produces more and more data, enterprises need to take advantage of the data. The transformation process is in its early stages, and many enterprises recognise the need for change, but lack methods and tools. This holds especially true for disruptive innovations such as big data. Incumbents are typically good at developing their current products and services, i.e. they excel at incremental innovations. Yet, they are incapable of making use of disruptive innovations (Christensen 2013).

1.3.3 **Research Questions**

This section presents the research questions that narrow the research purpose down to concrete, manageable pieces. The primary research question (RQ) of this dissertation is as follows:

> RQ: How can incumbent enterprises understand and utilise big data in their transformation process towards big data driven business?

The research question is broad. However, it can logically be divided into two sub-questions, one emphasising the importance of understanding the phenomenon, the other looking at the utilisation of big data in business contexts. These sub-questions are addressed in the Publications using pragmatic approaches discussed in the next section.

> SRQ1: What does an incumbent enterprise need to understand (regarding big data phenomenon) in its transformation process towards data driven business?

SRQ2: What are the key aspects that an incumbent enterprise must consider when utilizing big data in its value creation process?

The first sub-question (SRQ1) focuses on **explaining and understanding** the big data phenomenon. As always with new phenomena, numerous interpretations exist. This research sheds light on the subject by characterising big data, its possibilities and limitations. Understanding the basic concepts as well as examining the experiences of trailblazers provides useful knowledge and best practices for followers and may ease the transformation process.

The second sub-question (SRQ2) concentrates on the **value creation** processes and mechanisms. Committing to a data-driven approach is a start. Beyond that, several organisational and technical aspects, such as managerial and analytics capabilities, are required to convert the data into firm performance. The conversion process from data into economic value is by no means simple or straight-forward. Numerous factors, both internal and external to the firm, affect the process, causing either positive or negative impacts.

As a whole, this dissertation forms a coherent view that can be seen as a **high-level framework for big data adoption**. A pre-requisite of the adoption is to understand what big data is (and what it is not) as well as the potential benefits and risks. From the understanding of the phenomenon, enterprises might proceed to explore the experiences of trailblazers, which may help them to spark ideas and avoid pitfalls. As always, the transformation materialises through management decisions. Therefore, the behavioural intentions of executives with regard to big data are essential for the change to happen. Driving value creation with big data and analytics happens in a complex, non-deterministic process. Enterprises need to understand the process, as it helps them to avoid pitfalls, and focus on creating technical and organisational capabilities and skills that promote the value creation.

1.4 Research Approach

It is obvious that a researcher's background and previous experience affect how the approach to the subject area and the choices are made in the research. Creswell (2013) suggests that a researcher should explain the research approach of his/her research. The background of the author of this material is in computer science and software engineering disciplines. This, combined with experience in data warehousing, business intelligence and knowledge management, has certainly affected the choices, starting from the subject of the research, big data. The research is a natural continuum of the researcher's previous experience where data has played a central role.

Creswell (2013) also offers a framework that divides a research approach into three main elements, which are research philosophy, research design, and research methods. This research constructs the research approach from the three elements as shown in Figure 3.

The elements that best characterise this research are shown in the research approach circle in Figure 3. In short, multiple methods and designs are used in a pragmatic way in order to address the research questions.

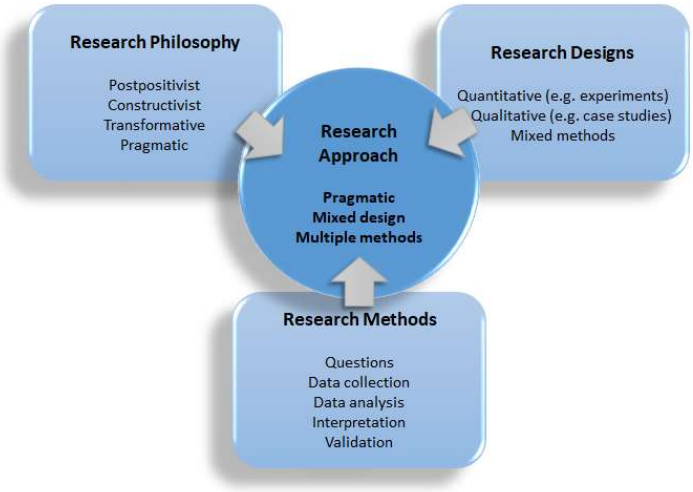


Figure 3. Research approach components, modified from Creswell (2013).

This chapter presents the research approach with background information regarding the researcher's choices. Understanding these assumptions has helped the researcher and hopefully will also assist the reader to set the research in the correct context.

1.4.1 Research Philosophy

The research philosophy refers to the philosophical assumptions that the researcher has or makes regarding the research. These assumptions influence the research and therefore they should be identified (Creswell 2013). Several scholars have contributed to the area of research philosophy, e.g. Chua (1986), Orlikowski & Baroudi (1991), and Easterbrook et al. (2008). Four commonly discussed paradigms are presented below (see also Figure 3):

- The postpositivist paradigm (a.k.a. the scientific method, empirical science, or positivist research) has long traditions. It represents a deterministic approach, in that causes at least probably lead to outcomes. The nature of the research is deductive. Starting from a theory, the researcher gathers, analyses and interprets data. The results either strengthen or weaken the theory.
- The constructivist (or interpretative) paradigm starts with an assumption that the relationships between technology, people and organisations are in constant change (Klein & Myers 1999). People create meanings and make sense of the

world in their historical and social contexts. Many qualitative researches use interpretative paradigm, building theories inductively as the research proceeds.

- The transformative paradigm is an umbrella term for several research directions that focuses especially on minorities, and includes political and social agendas in order to influence change; there is not a common body of existing literature (Creswell 2013). Change is essential in this paradigm. A researcher must include an action agenda for reform.

- The pragmatic paradigm focuses on practical matters, i.e. problems and solutions (Easterbrook et al. 2008). The focus is on solving the research problem and research questions. Researchers use different methods and approaches as they try to determine what works. Creswell (2013) notes several similarities between the pragmatic paradigm and the mixed methods research approach.

With regard to the research philosophy, the researcher could be identified as a pragmatist with a postpositivist flavour. The researcher considers this approach beneficial in a novel research area such as big data research, where the practices have not yet been formulated, i.e. there is no proof that certain methods produce more reliable results than others. When compared to single-method research, a novel area may be better explored using multiple methods. Different methods utilise different angles to the subject and may therefore reveal new insights. Moreover, the insights from Publications II and IV suggest that the nature of most big data implementations is experimental and iterative. This has similarities to the pragmatic research philosophy – focus on the problem and try to find the methods that work.

1.4.2 Research Design

The three main research approaches are quantitative, qualitative, and mixed methods approaches (Creswell 2013). Quantitative methodology is sometimes characterised as "numbers-related" or by as "closed-ended questions" -related, whereas qualitative research refers to a "word-related" or "open-ended questions" -related approach. However, the approaches are not isolated. Instead, they are more like a continuum where the quantitative approach lies at one end and the qualitative approach at another and the mixed methods approach falls in the middle. This research resides somewhere between the endpoints of the continuum; both quantitative and qualitative methods were used. A survey (Publication II), computer-aided content analysis (Publication III) and a case study (Publication V) characterise the research design and positioning in the continuum. On the whole, the research can be perceived as a qualitative study.

There are several possible research designs within each of the research approaches. For example, quantitative designs include experimental designs as well as survey-based designs. Qualitative research often employs case study designs. Other examples of qualitative research designs include grounded theory –based designs and narrative

research. Researchers decide which design or designs best meet the requirements of their research.

For this research, the design was selected in each case based on the best fit to the problem, considering the circumstances. For example, in Publication IV, the design selection was made from among several options. Action research (Baskerville 1999; Lewin 1947) or action design research (Sein et al. 2011) could have been used. As the goal of the study described in Publication IV was to develop an artefact, design science was selected instead of action research. The researcher considered that for developing a generic framework in a novel area, rapid experimental iterations with moderate user input was a proper approach. In addition, the researcher's previous experiences with numerous customer cases had shown that practitioners favour approaches presenting initial concepts or assumptions in order to provoke new ideas, instead of a clean sheet of paper as a starting point. The design science research method (Peffers et al. 2007) was a good match.

1.4.3 **Research Methods**

As one would expect from a researcher with a pragmatic research philosophy, different methods were used during the research. Each of the articles in this research applied different method, i.e. this is a multi-method research.

Publication I explored the concept of big data and the origin of the term. A natural method for this kind of research is a literature review. Big data is a wide topic that is discussed in various arenas and within multiple disciplines. On the other hand, the research is still at the early stages. Under these circumstances a mapping study is an appropriate approach to perform the review (Budgen et al. 2008; Kitchenham 2007). Therefore, a systematic mapping study was performed.

Publication II surveyed management's behavioural intentions regarding big data in Finnish enterprises using a well-established technology adoption model (Venkatesh et al. 2003). Surveys are commonly used to reflect the attitudes, opinions, and preferences of various audiences (Rea & Parker 2014). The usage of a web-based survey tool (Webropol) allowed a rapid turnaround time and cost-effective data collection.

Publication III looked at the big data case studies reported in peer-reviewed literature. The first stage of the research for this paper was to perform a mapping study in order to identify relevant case study articles. Again, principles defined by Budgen et al. (2008) and Kitchenham (2007) were followed. Then, computerised content analysis methods were used to identify common themes in the articles. Content analysis is an established methodology for investigating textual data (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 computerised content analysis have received the attention of scholars recently (Hao Hu et al. 2014; Lewis et al. 2013; Yu et al. 2014), as researchers wish to utilise new big data sources.

Publication IV explored the role of innovation as a mediator between big data and business models. The paper builds on the existing big data literature and the findings of Publications I and III. The research method of this paper applied a design science research method (DSRM) defined by Peffers et al. (2007).

Publication V used a case study approach. First, a decision-tree based machine learning algorithm was built using the R language package rpart (Therneau et al. 2017; Therneau & Atkinson 2017) for bid qualification. Second, cost-sensitive learning principles (Elkan 2001; Zadrozny & Elkan 2001) and classification matrices were used for business value assessment of the developed algorithm. Finally, the paper proposed guidelines and metrics for interpreting the impact in practical solutions.

Publication VI developed a process theory (Markus & Robey 1988; Soh & Markus 1995) based model of big data value creation in business contexts, reflecting the current big data literature in two widely used value creation frameworks (Ackoff 1989; Rayport & Sviokla 1995; Rowley 2007; Zeleny 1987) and arranged the results according to a process theory perspective.

1.5 Research Process

The research started at the beginning of August 2014. According to the initial research plan, the researcher began to explore the nature of big data. An interesting learning journey started. During the journey, the research plan was reviewed several times and changes were made according to the increased knowledge.

1.5.1 Data Collection

Several methods were used for data collection. Academic literature databases were naturally an important source in identifying relevant studies. Especially Publications I, III and VI grounded on comprehensive literature database searches. Principles defined by Kitchenham (2007; 2004) were used in the search processes. Publication I built an understanding of how big data is conceived. In addition to the databases, it was necessary to also look at other sources such as software and hardware vendors' materials, since practice was moving ahead compared to research.

A web-based survey was used to collect data for Publication II. The survey explored the attitudes and intentions of Finnish executives with regard to big data. Webropol was used as a tool. During the four-week time window in May-June 2016, 109 responses were received.

Data for Publication III was collected solely from literature databases as the purpose was to identify as many as possible big data case studies from academic sources. The search process revealed 33 peer-reviewed articles containing 49 big data cases.

For Publication IV, the purpose of the data collection was to gather information that could be used to evaluate and further develop the framework. Data was collected from three companies using personal, semi-structured interviews in March 2016.

Publication V was a case study, where a predictive algorithm was built and used for bid prediction. For learning the algorithm, data covering the last four years was collected from the related operative systems, such as a customer register, client meeting records and customer satisfaction surveys. Bids were collected monthly for six months. Several work meetings were arranged in order to gain comprehensive understanding of the data.

Three major literature databases (EBSCO, Scopus, Web-of-Science) were searched in order to identify relevant literature for Publication VI. Moreover, a comprehensive forward- and backward snowballing based on the title and abstract was performed for each of the papers that were considered relevant to the research.

1.5.2 **Data Analysis**

Data analysis methods were decided separately for each of the Publications. For Publication I, the first objective was determine the evolution of the definition of big data. Secondly, the aim was to identify gaps between the current definitions and common big data value propositions to identify missing perspectives of the definitions. The 62 papers included in the analysis were analysed manually as described in the Publication I.

Publication II surveyed the behavioral intentions of Finnish executives. The data was analysed using the R language and Microsoft Excel Data Analysis add-in. Regression analyses and t-tests were used to find statistically significant relationships between the variables and groups. In addition to inferential statistics, descriptive analyses and visualisations were used to present the results.

Content analyses of the 33 case study articles (Publication III) were analysed using text mining software. The goal was to identify common themes in the articles, i.e. to synthesise the findings of the big data cases selected for this study. This is a typical content analysis task that can be computerised. The KH Coder software[3], which has been used in over 900 research projects, was used in the analysis. First, the articles had to be converted to plain text as the software did not understand PDF-format. The principal analysis methods used were an analysis of key words in context (KWIC) and co-occurrence maps. Figure 4 presents a visual example of the tools. An iterative analysis approach revealed the themes. An additional analysis using KWIC and a manual inspection of the articles was performed to verify the correctness of the themes.

[3] KH Coder is free software for content analysis and text mining - http://khc.sourceforge.net/en/

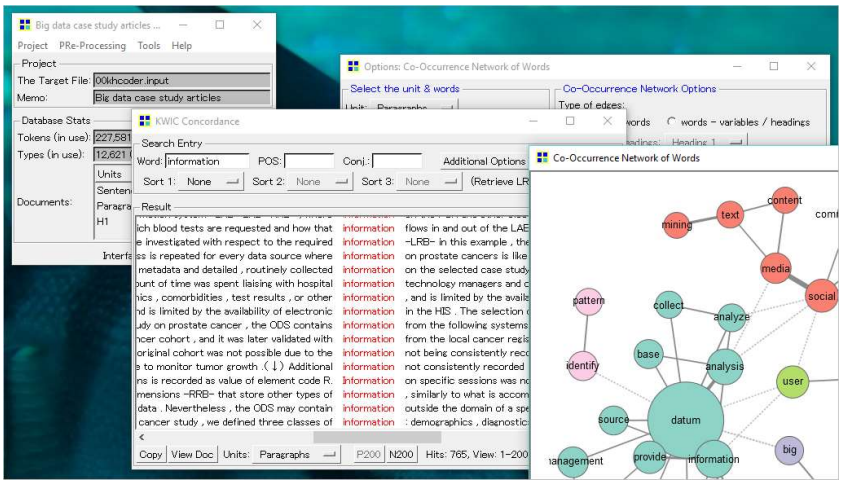


Figure 4. Example of text analysis tools of the KH Coder software.

Publication IV developed an artefact using the design science research methodology (DSRM) framework (Peffers et al. 2007). Current big data research, including Publications I and III, was analysed and integrated with findings from other disciplines, particularly from strategic management and innovation research in order to create an initial version of the artefact. Expert interviews were then used in the demonstration and evaluation phases to verify and further develop the artefact. The interview recordings for were analysed manually and translated into memos. The memos were then sent to the interviewees for review and possible corrections.

In the case study described in Publication V, the data analysis was done using the R language. Learning the algorithm, using it for bid prediction as well as the metrics proposed for follow-up are described in detail in the article.

The big data papers discussed in Publication VI were categorised manually. The topics discussed in the papers were determined based on the abstracts and, where required, full-text reviews. Then, the topics were arranged into categories, and both positive and negative factors towards the topics were identified. Based on the topical categorisation, the reflection against the two value creation frameworks was done, and a process-theory-based view explaining the conversion from data to firm performance was compiled.

2 Big Data

This chapter gives an overview of big data. Practitioners and public media have intensely discussed big data in the recent few years. As often with new, interesting concepts, the public discussion is somewhat distorted. Much hype is included and sometimes participants intentionally emphasise aspects that are beneficial to themselves (Publication I). One of the reasons for this is that big data is a new, emerging frontier as a research topic. Established theories and definitions are still emerging. The various aspects discussed in this chapter cover the phenomenon from three angles: what is the big data phenomenon, how does big data create value, and what is the impact of big data?

"We are all now connected by the Internet, like neurons in a giant brain."
– Stephen Hawking

2.1 Introduction to Big Data

An article published in Wired-magazine (Anderson 2008) suggested that big data will make scientific methods obsolete. This rather provocative article gained a lot of public and scholarly attention. The basic idea was that if you have more detailed data and good analytical tools, you do not need to perform causal or semantic analyses. "Correlation is enough." For many practical situations and decisions, indeed, correlation is enough. For example, many e-commerce sites run various product and user experience tests (A/B testing) every day on their sites, experimenting with which combinations are best with regard to increased sales. They do not ask why, they just rely on the data, i.e. how the combinations correlate with the sales figures. However, in many other contexts, for example in health care research, causality plays an important role.

In 2011 McKinsey Global Institute and IDC published reports (Gantz & Reinsel 2011; Manyika et al. 2011) that drew wide public attention to the potential value of big data. Since then there has been a lot of buzz in media. A countless number of blog texts have been published, newspaper articles regularly discuss the subject; numerous authors have published big data –related books. In addition, scholars have been activated. Big data is regularly discussed in various conferences, and new journals such as the "Journal of Big Data", "Big Data & Society" and "Big Data Research" have emerged.

Figure 5 shows the distribution of papers indexed in the SCOPUS literature database. Between 1998 and 2017, the number of indexed papers was 40,480. The search included all papers where the title, abstract or keywords contained the term "big data". Since 2011 there has been a surge in the number of the papers. The "Other" category covers 10 various subject areas, such as arts, neuroscience, and planetary sciences. The sum of percentages is more than 100 because one paper may have been categorised in several categories.

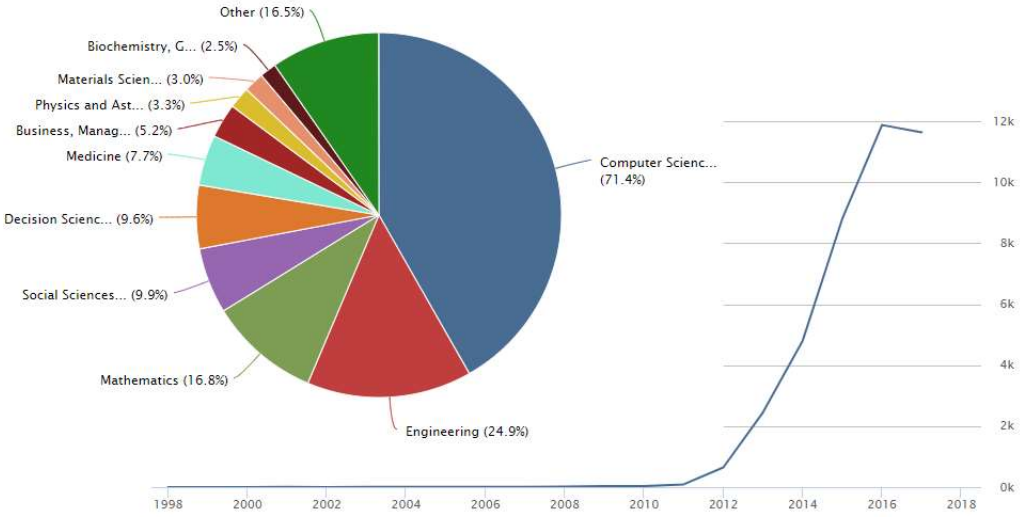


Figure 5. The number of big data articles by subject area and year.

Figure 5 clearly shows that big data research so far has focused on technical aspects. By subject type, most of the papers were categorised under computer science, engineering and mathematics. Business and management related big data papers represent only 5.2% – or approximately 2000 pieces – of the whole body of knowledge.

The following sections give an overview of the drivers and sources of big data. From there the discussion proceeds to the definition, criticism and challenges of big data. The basic concepts presented here help the reader to understand what big data is and what it is not.

2.1.1 Big Data Drivers

Technological developments such as personal computers, data communications, and collaboration tools allow both organisations and individuals to create local and global networks. These networks enable data sharing, social communications and emphasise the importance of individuals as change drivers. The uprisings in the Arab countries in 2011 were called “Facebook-revolutions” due to the centric role of social media platforms and unofficial, dynamic networks. People used their mobile phones and apps to communicate, collaborate and organise the events, as well as to pass around data such as text, photos and videos. They made decisions based on the information they shared and gathered in real-time using social media. Technology enables new, sometimes unpredicted ways of doing things.

Connectivity, IoT and mobility are commonly identified, e.g. (Davenport 2014; Mui & Carroll 2013; Van’t Spijker 2014), as key trends behind the data generation. Information

technology has enabled enterprises and people to **connect** for decades. The invention of the personal computer and the Internet have had a big influence on this. Enterprises integrate their processes with others. For example, many companies have created seamless logistics processes with their suppliers. A recent phenomenon is the exploitation of various social platforms. These enable individuals to easily connect with others, create unofficial networks that reach well over company boundaries. Van't Spijker (2014) underlines that this changes the ways employees communicate as well as the ways clients communicate. These new ways of communicating shape the businesses. At the same time these interactions create huge amounts of data.

The **Internet of Things** –concept (IoT) is another driver behind the data deluge. Among many others, Van't Spijker (2014) states that the IoT makes it possible to create products and services that have never before been possible. Energy companies now invoice their clients using real-time data from smart meters. However, the same data offers more possibilities. For example, companies could create energy consumption profiles of households and produce energy-savings information for their clients. Since each electrical device has a unique energy consumption profile, they could even analyse the data to see whether a client's washing machine wastes energy. Based on the data and analytics the company could then suggest a new machine that meets the customers' needs but saves energy and thus money. Smart meters as well as many other IoT applications are capable of generating detailed data in real-time. No human input is required; the IoT concept enables computers to sense the surrounding world by themselves. This autonomy, combined with the fact that IoT implementations are rapidly increasing is leading to the exponential growth of data from IoT sources.

The third change driver is **mobility**, e.g. (Davenport 2014; Mui & Carroll 2013; Van't Spijker 2014). Mobile devices, such as smart phones and tablets have had an enormous impact on peoples' lives. Mobility changes both individuals' lives and their way of working. Smart phones are easy to use, they have access to the Internet, and they have several sensors, e.g. cameras and GPS-sensors. Documenting a planning draft from a flip board is easy, just take a picture and send it to your colleagues. Moreover, humans are social beings; we want to share things with others. Foursquare[4] is a social sharing app that uses location data and helps you find restaurants etc. based on the suggestions of other people. Connectivity and mobility are closely related. With mobile devices, we are always connected and available – to home, to work, or to social platforms. This not only creates data, but also a new kind of dynamics regarding what we do and how we act.

Services enabled by change drivers create value for their members, e.g. pieces of information, cumulated knowledge or just pure fun, but they also emphasise the role of **platforms** in value generation. The platform cumulates the data the members create. Moreover, every time a member visits the platform, a pile of digital breadcrumbs is left behind. The owner of the platform has access to members' digital trails as well as to the

[4] https://foursquare.com/

data the members have created. LinkedIn[5] has more than 546 million members and probably knows more about career paths than most HR departments.

The resulting data volumes are huge. Figure 6 shows an example. This illustration presents, how Internet users (3.7 billion people in 2017) produced data using some popular services and platforms. The volumes are from 2017 and it should be noted that the numbers in the illustration are per *minute*.

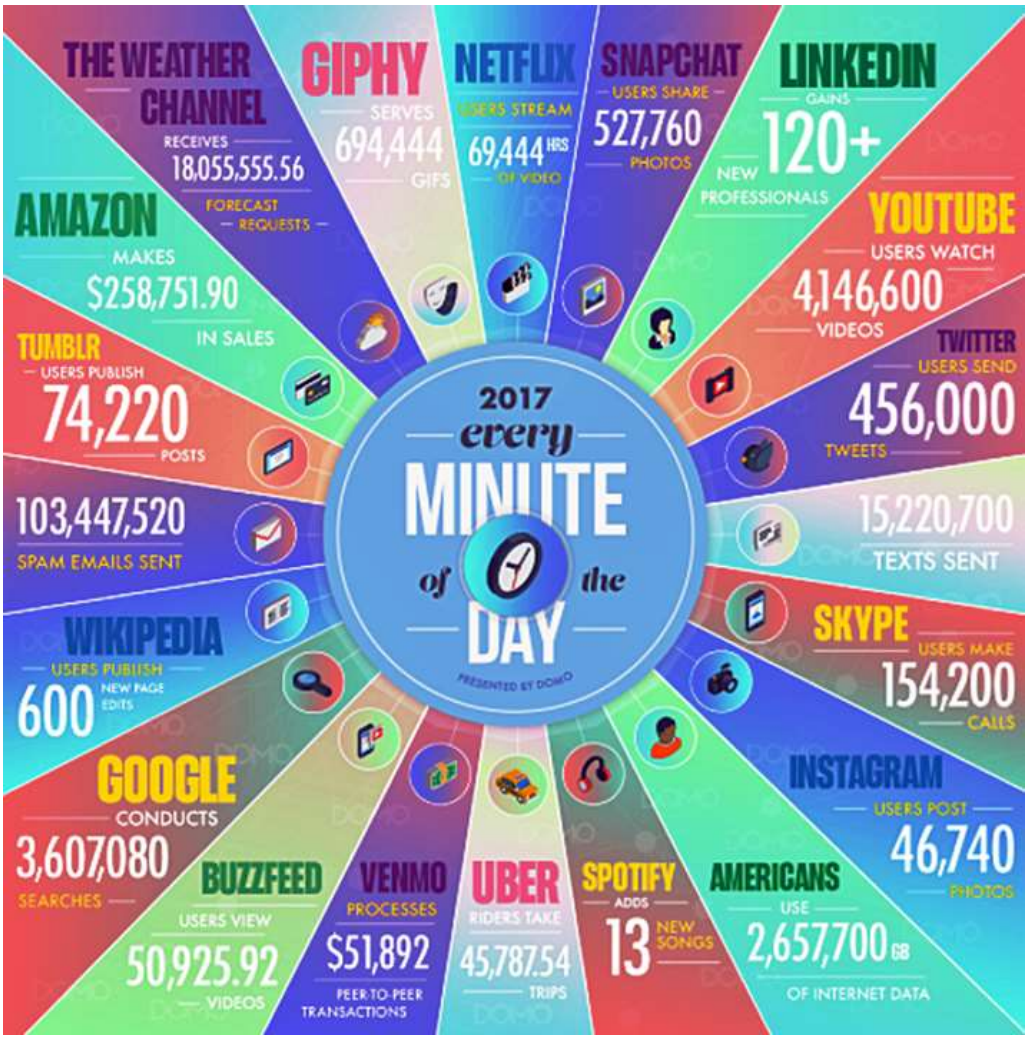


Figure 6. An exemplary illustration of Internet data volumes[6].

[5] LinkedIn is a business-oriented social networking service (https://about.linkedin.com/)
[6] The infographics was created by Domo, see: https://www.domo.com/learn/data-never-sleeps-5

Moreover, the growth shows no signs of slowing down. On the contrary, in many areas (such as IoT), the trend of the growth is exponential. The number of people who have Internet access is growing, the number of smartphones is increasing, and the IoT concept is emerging, just to name a few factors. Thus, the amount of data is growing at an ever-increasing pace. Figure 7 displays a forecast of annual IP traffic globally. Compound annual growth rate (CAGR) is 23 % per year.

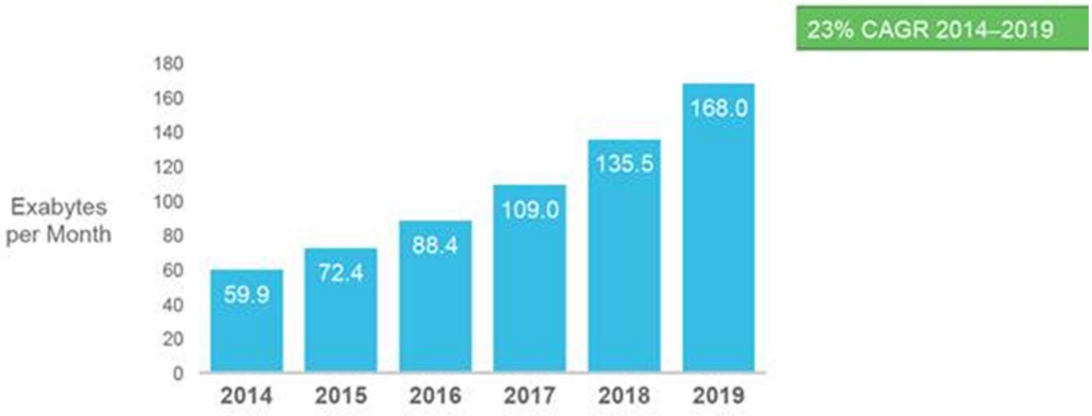


Figure 7. Global IP traffic growth forecast[7].

The volumes and growth numbers are impressive. However, an enterprise must look deeper in order to understand which data might be of value for its business. The next chapter takes a different perspective to clarify, where the data comes from.

2.1.2 Data Sources and Types

Looking at the data sources by type provides indications of where the data comes from. It also concretises how enterprises might utilise the big data drivers presented above. It should be noted that the following categorisations are indicative and partly overlapping, as many sources fit in several categories. For example, a picture taken with an enterprise owned phone, or web log files from enterprise's own web site.

Mobile data. Mobile devices have penetrated the market in recent years and this trend will continue. Gartner reports a 59 % year over year growth from 2014 to 2015 in global mobile data traffic and expects the traffic to increase more than three-fold by 2018, just

[7] The figure was taken from Cisco's white paper, available from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

in three years[8]. It is estimated that in 2018, of all the sold phones globally, smartphones represent almost 80 % (InternetSociety 2015). Moreover, tablet sales has overtaken PC sales. For enterprises, this increase in smart devices offers a number of possible data sources, such as geolocation, app usage patterns, and app generated transactions or interactions. Enterprises might gather data from external users as well as from internal users.

Social media data. Figure 6 shows that humans want to share. We share, ask, discuss, and read. Many enterprises already have Facebook, Twitter, or YouTube channels. Enterprises use them mostly for marketing. These channels offer a direct feedback mechanism for customers. Customers use these official channels, but they also discuss the enterprise, its products, and services elsewhere. Many enterprises also use social media applications such as chats or discussion boards internally. Social media data is typically high-variety and sometimes high-velocity data. One Facebook post may contain free text, links to external web pages and pictures or videos.

Camera data. Every smartphone has a camera. Cameras are common everywhere: in public places, retail stores, private properties, and factories. Analysing data from pictures and videos is increasing. One great example comes from professional sports. The NBA Stats[9] service provides basketball fans with detailed player and team statistics from every NBA basketball game. Each venue has six high-speed video cameras, which record every move of the ball and each of the players as well as the movement of officials. The service collects and analyses data in real-time, at 25 frames per second. This data, combined with manually gathered live-statistics from each game enables the teams and coaches to analyse every second of the game afterwards in a three-dimensional model and provides extremely detailed statistics to the fans. NBA Stats has more than 25 petabytes of video data.

Sensor data. Cars, phones, machinery, environment, buildings, military technology – sensors are already everywhere. In most cases, sensors are relatively inexpensive and easy to install. The Internet of Things (IoT) concept is rapidly increasing the number of sensors that are connected to Internet. Sensor data is often high-volume and high-velocity data. For example, a self-driving car or a drone produces – and consumes – vast volumes of data every second.

Web data. Web logs and click-through analyses are common techniques that enterprises use to find information regarding the visitors to their web sites. These techniques provide measurable data, such as where the users came from, how long they stayed, or what pages they visited. In addition to this, the web offers other possibilities. Data from external sources is often either available or accessible. Some sites offer mechanisms such as application programming interfaces (API) or RSS-feeds that could be used to gather data.

[8] Gartner's press release available from: http://www.gartner.com/newsroom/id/3098617
[9] NBA Stats service, see: http://stats.nba.com/

Web-crawlers, i.e. programs that extract data from web pages, are useful if no programmable interface is available.

Enterprise data. Organisations perform transactions that generate "traditional" data such as orders and invoices that they use, e.g. in their operations and reporting. This data is familiar to them, and enterprises have systems such as enterprise resource planning (ERP) and reporting solutions that help them to utilise this data. However, as already discussed, enterprises could utilise many new data sources, both internal and external. One internal area is the collection of so-called dark data. Often enterprises gather and store data that they never actually use. Some of the dark data lays in transactional databases. However, most of the dark data are unstructured, such as word processing documents, Excel files, or emails. It is estimated[10] that 80 – 90 % of all data in an enterprise are unstructured. Analysing this data typically requires adding some structure to the data as well as text mining capabilities. Unstructured data amounts are growing[11] and enterprises allocate storage at it – but neglect to utilise it.

Geolocation data. Maps and location data is useful for many purposes. For example, enterprises visualise reporting data on maps and follow their fleet using location data. Social media apps utilise geolocation data, e.g. the social media application Foursquare shows restaurants and other points of interest on a map. It can even show where your friends are based on the GPS data of their smart phones. The growing number of smart phones, combined with new mobile device apps is one driver that increases the amount of available geolocation data. Moreover, almost any data can be augmented with location data.

Open data. Governments and other organisations increasingly open their data to others. As an example, the U.S. Government's open data service[12] contains more than 190,000 data sets. Accordingly, several Finnish organisations[13] have open interfaces to their data. Utilising these data sources provides opportunities to provide new services. Enterprises could augment existing data with this open data gathered from open data sources or even innovate new services that make use of the open data.

Another categorisation is to divide the *data* by type into transactions, interactions, and observations (Moniruzzaman & Hossain 2013). This categorisation approaches the data from enterprise point of view.

Transactions result from the internal activities of an enterprise. They indicate actions that enterprises must record in order to run their businesses, or for legal reasons. For

[10] http://www.webopedia.com/TERM/U/unstructured_data.html
[11] http://wikibon.org/wiki/v/The_Growth_and_Management_of_Unstructured_Data
[12] https://www.data.gov/
[13] Finnish open data examples:
- Valtion Rautatiet (http://rata.digitraffic.fi/api/v1/doc/index.html),
- Ilmatieteen laitos (https://ilmatieteenlaitos.fi/avoin-data),
- Maanmittauslaitos (http://www.maanmittauslaitos.fi/avoindata/aineistoluettelo).

example, a purchase order or a sales invoice represent transactional data. Enterprises store transactional data typically in relational databases. Transactions are structured data. However, the number of transactions depends on the case, as well as the frequency. A consulting company may send only a few hundred invoices per year, whereas a large web shop may produce thousands of payments per day. Transactional data is a by-product of the business, controlled and managed by the company, and is relatively easy to utilise, e.g. in reporting due to the tabular structure.

Interactions are non-transactional activities between the enterprise and other parties. Typically, enterprises are most interested in customer interactions. For example, interactions take place when a customer calls a helpdesk, sends an email inquiry, or browses the company's web site. Enterprises are interested in interactions, since they enable better understanding of the customers' behavior. However, for several reasons, enterprises utilise interactional data sources much less than transactions. The variety of data is rich, which in many cases requires new capabilities such as new software or text-mining skills. Deriving insights from a click-through analysis of a web site and delivering the results to end users is not a trivial task. Besides, calculating the business case may be challenging. The costs are direct and thus relatively easy to calculate. The benefits, however, are often indirect or qualitative, which makes the calculations difficult.

The third data type in this classification are observations. **Observations** refer to gathering and analysing data about the actions people take, rather than collecting the transactions or interactions they create. Tao et al. (2014) explore bus rapid transit passenger behavior using smart card data. Instead of being interested in transactions (smart card payments), they observe the behavioral patterns of the passengers. The time of the ride, how often the passengers ride, how many passengers there are at certain times and so on. By augmenting this data with other data, such as location data, holiday calendars, air temperatures and rainfall data, they build visualisations and analyses that help to understand the usage patterns and dynamics of urban traveling. These insights help to plan smarter and more rapid bus transit systems. Digital platforms are another example of the power of observations. Amazon's recommender system gathers observations from several sources and produces suggestions to the potential buyer. The system produces almost a third of the company's revenues (Van't Spijker 2014).

Observations take advantage of the **digital trail**. Each of the user's actions contains metadata that can be used to profile the user. Comments and likes describe the interests of the user; while date, time and location indicate when and where she is (digitally) active; on-line networks such as LinkedIn contacts or Facebook friends offer data from the user's social connections. This research defines digital trail as follows:

> The digital trail is the data *about* the user that follows from the user's actions.

Gantz & Reinsel (2011) claim that the digital trail[14] is not only much greater than the information a person creates herself, but that the trail is growing rapidly as well. Gathering and analysing this data can create insights and value, as the above discussed observations examples show. Therefore, enterprises are increasingly interested in the digital trails of users. It seems that environments or platforms with lots of users and action are becoming important sources of digital breadcrumbs. Moat et al. (2014) gives several examples of "predicting the present" such as incoming tourist numbers, motor vehicle sales and stock market trading volumes where the correlation between real-world events and search engine data have been identified.

Utilising the digital trail often requires composition of "breadcrumbs", i.e. small data bits coming from several sources. This has consequences both for enterprises and for users. For enterprises, the trail is technically challenging to create due to various data types and possibly advanced analytics requirements. Users have very few means to review their digital trail, as there is no way to find out what data is used and by whom.

2.1.3 Defining Big Data

The term "big data" is not new. It has been used both in research and non-research papers for quite a long time. Back in 1997 it was used in the context of visualising large data sets (Cox & Ellsworth 1997). In 1998 it was used in a hardware-related presentation (Mashey 1998) and also in the data mining context (Weiss & Indurkhya 1998), and 2003 in combination with statistics (Diebold 2003). In the early days, big meant the size, and all these sources recognised and referenced big data with the increasing volumes of data.

The year 2001 can be considered a major milestone in the definition of big data. Laney (2001) described three essential dimensions of big data: *volume*, *velocity* and *variety* (so-called 3V definition). Volume refers to the ever-increasing amounts of data. Velocity indicates the need of capturing and analysing high-speed or bursty data in (near) real-time, or else the value may be lost. Variety relates to different types of data, be it structured or non-structured, such as social media posts or videos. Figure 8 visualises the 3V definition. In recent years, scholars and especially practitioners have developed numerous big data definitions. No consensus regarding the definition exists. For a detailed discussion about big data definitions and the history of the term, see Publication I in the "Publications" section.

[14] Several different terms, such as digital shadow, cyber shadow, electronic footprint and digital footprint are used in discussions. This research prefers the term digital trail.

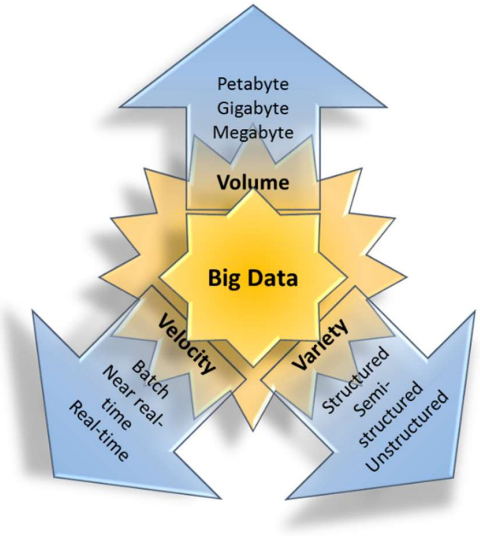


Figure 8. A visualisation of Laney's (2001) 3V definition of big data.

According to the suggestions made in the Publication I, this research distinguishes between data related actions, data usage related actions, and the big data phenomenon. The following definitions are used.

Big data refers to high-volume, high-velocity and high-variety information assets.

This definition for data related actions follows from the original 3V definition (Laney 2001). Defining big data as high-volume, high-velocity and high-variety information assets gives the essential dimensions for the data. Based on the analysis of existing definitions (Publication I), there seems to be a common understanding that from the viewpoint of data, the 3Vs are the very essence of big data. It should be noted that, in line with the findings of Publication I, additional dimensions like value or veracity are omitted, as they would lead to logically flawed definition.

The definition suggests that data is an asset. This is to emphasise the importance of data in modern businesses. Data meets the criteria of an intangible asset (IFRS 2015) with the exception of value measurement – data is an identifiable, non-monetary, non-physical, potentially valuable resource that a firm produces or harvests from different sources. Although reliable measurement of the value may be difficult or even impossible, making big data mentally equal to other assets is justifiable due to the its increasing importance.

Big data insights refer to analytics that are used to derive value from big data information assets.

Analytics needs to be separated from data, because mixing data and its intended usage leads to flawed definitions. This is the case with several current definitions. Value, for example, must be derived from the data by using analytics. There is no value in plain data as such (Ackoff 1989), although big data assets are *potentially* valuable. Value is also case-dependent. A certain piece of information may be worthless to one company but highly valued by some other enterprise, or in another context.

The big data phenomenon refers to the paradigm shift towards data-driven businesses and ecosystems.

The definition of the big data phenomenon emphasises the disruptive nature of big data. When describing the phenomenon, big might refer to the big and pervasive paradigm change that is driven by data. As big data will affect most industries, ecosystems and enterprises, the change definitely is big. Accordingly, from this point of view, data would refer to cultural or organisational changes that are required, i.e. a new data-oriented mindset. In order to utilise big data possibilities to full extent, enterprises must drive towards more data-driven culture and apply decision-making processes that are based on data. Approaching big data as a phenomenon increases understanding and shifts the focus of the discussion from technical to business related topics. Enterprises must review the impact of big data on their business model to reveal the risks and opportunities. Dealing with disruptive technologies such as big data is more about business transformation than about managing data, hardware or software. Although proper technology is required to deal with big data, the potential business transformation is the most important aspect. The aspect of strategic importance of big data is currently either neglected, as technical topics are preferred, or confused with the data aspects, which has led to vague definitions, for example.

As the number of stakeholders and parties increases, common definitions become more and more important. Good definitions enable common understanding between parties coming from different backgrounds. The definitions above are an improvement on the current situation where inconsistent definitions blur the view and cause misunderstandings. However, the researcher recognises that even these definitions are not exact, but subject to different interpretations.

Netflix[15] is an enterprise that serves as a clarifying example of the definitions. The company was founded in 1997 to offer movie rentals. At the beginning, they used a web shop combined with traditional pay-per-rental business model, physically sending DVDs to customers. A couple of years later they dropped the per-title renting and introduced

[15] Netflix (https://www.netflix.com) aims to be a global Internet TV network offering movies and TV series commercial-free.

monthly subscriptions. By paying a fixed amount per month, a customer could keep the DVDs as long as they liked, i.e., with no late fees or due days. At the same time, they gathered data and developed analytics that personalised the offering on their web site for each customer. In 2007, Netflix introduced a streaming service that allowed customers to download movies on-demand over the Internet, instead of the previous postal delivery. In the following years Netflix grew, while overall DVD sales fell. At the end of 2017, Netflix had more than 117 million global subscribers.

So how does this history meet big data? To put it simply, the company currently builds its business model on big data. Amatriain (2013) explains Netflix's recommender system and analytics behind it. The recommender system personalises the offerings based on high-volume, high-velocity, and high-variety data that is gathered from several sources, including ratings from users, search terms, and social media data. Analytics and rapid testing cycles allow data-driven decisions. Amatriain (2013) states that the recommender system is at the core of Netflix's offering and is thus a key component of their business success. In other words, the recommender system is an essential component of their business model. With regard to the discussion about definitions above, all three definitions and the asset aspect can be identified as well as the logic between them. The transformation from traditional business to a data-driven company represents the big data phenomenon. Analytical algorithms derive the value from the data, representing big data insights. However, without the data the algorithms are useless. Thus, big data is an important information asset for Netflix.

2.1.4 Big Data Criticism

With any new phenomenon critical voices arise. Big data is not an exception. One of the main sources of the criticism is the term itself. Fox and Do (2013) claim that the term is vague, with indeterminate boundaries. Their observation is obvious; a few terabytes of data may be big for one enterprise, but a no-brainer for another. In addition, the boundary changes over time. Volumes that ten years ago were considered big are nowadays normal. However, a proportional definition of the size, i.e. what is considered big, is practical and easy to understand. From a practitioner's view, things are clear: Whenever you have so much data that it exceeds the capabilities of your current hardware or software, you have a problem to solve. Without an exact definition of the size, the term does not become outdated over time and it is applicable to many situations. Defining "big" in the current context is practical, although it comes with the risk of misunderstanding.

Another aspect related to the term is that the word "big" is frequently associated with volume, which effectively leads to the perception that big data is only about high volumes of data. Ignoring the two other dimensions, velocity and variety, underestimates the challenges they pose. Many, if not most, big data sources produce unstructured data, such as textual data from social media, or video data from surveillance cameras. Managing big data requires new technical capabilities that stretch beyond managing the volumes alone. Unstructured data does not fit neatly in the relational data models that are common today. Sensors, for example in a manufacturing process, can produce high-velocity data. Real-

time processing and analytics are virtually impossible to do with traditional data warehouse architectures that many enterprises currently have.

Yet another consideration regarding the term is that the word data can be misleading. Data itself usually is not interesting. Firms and other organisations are far more interested in value, i.e. the insights derived from the data. Big data sounds like a technical term, which may be confusing. Although there may be significant technical challenges, the most important factors of big data relate to business transformation and cultural changes.

In their insightful article "Critical questions for big data", Boyd & Crawford (2012) raise six critical viewpoints regarding big data. Their first claim is that the numbers do not speak for themselves, as many big data enthusiasts argue. Big data specialised tools have their limitations, which need to be understood as well as the possibilities. Secondly, they point out that the objectivity and accuracy of the data is a myth. Data management involves activities such as cleaning, filtering, and interpretation that require subjective decisions. Data may contain errors, originate from an unreliable source, or parts of the data may be missing. They also argue, "bigger data is not always better data". This is to say that in practice big data often means a sample, not the entire data set. In these cases, there are the same risks regarding biased samples as in any sample-based experiment. The fourth claim addresses contextual factors. For example, results derived from Facebook networks are not applicable to personal networks. The context must be taken into account. The next viewpoint emphasises ethical questions. Under what circumstances and for what purposes is the usage of the data ethical? Their sixth observation is that access to the big data sources causes inequality. For example, social media enterprises have access to all of the data they have gathered, but others may have to pay for the access. According to the authors, this leads to a digital divide between "big data rich" and "big data poor".

Although Boyd and Crawford (2012) raise their questions in the research data context, they are applicable to the enterprise data context as well. For example, is it ethical for a betting company to exploit their data about gaming habits of a gambling addicted person? Is it ethical for a bank to profile its customers by analysing their bank transactions without customers knowing it? Or, when an autonomous vehicle confronts a situation where an accident is inevitable, the vehicle must make a decision based on the available data. What are the grounds and morality behind the algorithm, or artificial intelligence that attempts to minimise the damage? Will different manufacturers try to optimise their passenger safety by making ethically questionable decisions? As a further example, Martin (2015) discusses the ethical questions related to the secondary usage of data. She identifies big data related issues at the company, ecosystem and industry level and suggests guidelines to preserve sustainable practices. Clarke (2016) presents different scenarios where big data quality (or veracity) might effectively ruin the expected value, the point being that the moral responsibility of both scholars and practitioners is to apply reality checks before promoting the positive effects of big data.

2.1.5 Challenges of Big Data

This section gives an overview of the generic challenges and risks with big data from the viewpoint of an enterprise. Although most of these topics are beyond the scope of this research, they are briefly listed here as the issues are real. It is also important to notice, that the opportunities for making use of big data require awareness and activities not only from enterprises but from various parties of society. Citizens, philosophers and policy makers face interesting, and difficult, questions already, and increasingly in the near future.

Technical possibilities are almost limitless, and it is difficult to find well-justified economic reasoning against digital transformation. It seems obvious that "*everything that can be digital will be"*[16]. One key issue is that technology advances much faster than our understanding of the impacts. Many issues identify themselves in various ways in real-world cases, such as regulatory stipulations (is it legal to implement an Über-style passenger service?), organisational challenges (does the transformation create inequality problems?), or ethical problems (how do customers react, if we base our insurance prices on the digital trail that the user has left behind?). Thus, ethical and moral questions become increasingly important.

Practitioners must consider **privacy** in every big data project. Privacy is a broad topic and privacy concerns get a lot of attention both from scholars, e.g. (Eckhoff & Sommer 2014; Gehrke 2012; Stopczynski et al. 2014) and the media. The digital trail and especially the difficulty in managing it is obviously one of the concerns for individuals. Losing anonymity is one aspect; exposing information unintentionally is another. For example, by analysing our tweet history, we might find data that exposes our social network, political and other interests, behavioral patterns or other things that we had not thought of. For what purposes, if any, is it right to profile users based on the data left behind their activities? This viewpoint has gained huge publicity due to Facebook improperly leaking data from tens of millions of user accounts to political consultancy company Cambridge Analytica[17]. Moreover, it is worth noting that even simple "breadcrumbs" such as knowing that the visitor uses an iPhone or has visited the site several times may be used to e.g. price optimisation. Sometimes the public opinion makes the data unavailable, or at least unsanctionable. Combining the data from speed limits, mobile phone locations and time could produce a lot of speeding tickets. However, most people would consider this a serious affront to their privacy.

Big data **security** is another broad challenge and is widely discussed among scholars, e.g. (Altshuler 2011; Subashini & Kavitha 2011; Sullivan 2014). Newspapers and the media have reported several data thefts in many areas of life[18]. Stealing data such as credit card

[16] http://digital-archaeology.org/anything-that-can-be-digital-will-be/

[17] E.g. http://www.bbc.com/news/technology-43649018

[18] For more about data breaches, see e.g.
- https://www.rt.com/news/363642-websites-outage-ddos-attack/
- http://www.forbes.com/sites/moneybuilder/2015/01/13/the-big-data-breaches-of-

numbers and personal information potentially leads to economic loss or identity theft. Enterprises must plan for the security from day one in their big data projects.

ICT related challenges. All big vendors invest huge amounts in their big data technology development. Although the technologies will mature over time, so far many big data implementations face challenges either directly related to technology or related to data. Managing the volumes is one issue, e.g. (Dutta & Bose 2015; Krumeich et al. 2014), data inconsistencies and poor data quality are another, e.g. (Mathew et al. 2015; O'Leary 2013). In addition, many technologies that enterprises currently have in place, do not support big data. For example, many enterprises have implemented data warehousing/business intelligence (DW/BI) solutions to support their organisational information needs. While DW/BI systems work well in environments with (reasonable amounts of) relational data, it is easy see a number of reasons why these systems have limited capabilities with big data by just looking at the definition of big data: Storing huge volumes of data in relational databases is not cost efficient (80 – 90 % overhead due to indices etc.), architecture is based on batch processing instead of real-time analysis (velocity is an issue) and unstructured data does not fit well in relational databases or current business intelligence tools (variety is an issue).

Organisation related challenges include decision-making and skills. Improved decision-making is a key intention in many current big data projects, e.g. (Bärenfänger et al. 2014; Dutta & Bose 2015). Analytics and insights should be embedded into processes and decision-making routines (Bekmamedova & Shanks 2014). These intentions implicate the adoption of a data-driven organisation culture. This may be a difficult and long process due to organisational resistance – old habits die hard. Skill-shortages are another organisational challenge, which it is also related to new technologies. New data types, such as social media posts or text documents, require new kinds of analytics. This is a multifaceted issue: in addition to new technology, organisations need new talent, both business oriented and technology skilled, e.g. (Phillips-Wren & Hoskisson 2015; Shen & Varvel 2013). The term data scientist has emerged to describe the role. Davenport (2014) has analysed the qualities required from a data scientist. The role combines a scientist, a consultant, an analyst and a business domain expert. It is clear that very few people exist, who qualify to this role.

There are **business model related challenges** also. Disruptive innovations are challenging for incumbent enterprises due to several reasons (Christensen 2013). Incumbents steer their resources according to their customers. Big, profitable customers are preferred. Under normal conditions, i.e. under rather stable business environment, this is reasonable. However, disruptive innovations do not fit under normal conditions. Often they do not initially meet the needs of current customers, include technical risks and are less profitable than current products and services. Secondly, markets for disruptive innovations are initially small and volatile, i.e. insignificant for larger companies. And

2014/#554b00fc3a48
- http://www.networkworld.com/article/3011103/security/biggest-data-breaches-of-2015.html

since the markets are "non-existent", it is very difficult to analyse and forecast them. All this applies to big data as well as long as we speak about new, disruptive products or services. Any typical incumbent enterprise requires a business plan for new product or service development that contains evidence regarding market potential, customer segments and development costs. Creating such a plan is virtually impossible (except if the plan is imaginary). Indeed, according to Christensen (2013) any forecast with regard to disruptive innovation is deemed to be false.

2.2 From Data to Actionable Insights

This chapter explains the mechanisms that lay the grounds for understanding how data creates value. The chapter covers three topics. The characteristics of data as well as the differences between data and tangible assets are discussed. Then the concepts that explain how data is expected to generate value according to current perceptions are explained, followed by a presentation of the areas where big data is supposed or has already proven to be able to add value.

"It is a capital mistake to theorize before one has data."
– Sir Arthur Conan Doyle

2.2.1 The Laws of Information

Moody & Walsh (1999) presented seven principles or "laws" of information. These principles define characteristics that separate data from other assets. Understanding of the principles is a pre-requisite for evaluating or measuring the value of data. The laws are as follows.

1. Data is infinitely shareable. The same data can be shared between any number of parties with equal value. This is not possible with most of the resources. E.g. sharing staff between business areas means that each area receives part of the staff. With data, each area can have all of the data, i.e. gain all the value available. Therefore, from the enterprise point of view value is cumulative: the more people use the data, the more value can be extracted.

2. The value of data increases with use. The reason stems partly from the first law, partly from the fact that most of the costs of data consists of data management (gathering, harmonising and storing). The cost of using the data (marginal cost) is usually very low. Therefore, the more people use the data, the better value-cost ratio.

3. Data is perishable. The value of data typically deprecates over time. The trend is towards real-time usage, but the value depends on the use case. A real-time bus schedule provides value for potential passengers, historical sales figures from the last three years provide value for sales managers.

4. The value of data increases with accuracy. Accurate information is more valuable. However, this is case-dependent. Bank transactions require 100 % accuracy, whereas some erroneous home phone numbers in an HR system's employee data might not be significant.

5. The value of data increases when combined with other data. Combining customer demographics with their buying patterns can lead to more targeted marketing and increased sales.

6. More is not necessarily better. Excess information leads to information overload and reduced understanding.

7. Data is not depletable. For most resources the following holds: the more you use, the less you have. However, using the data often generates new, derived data as a result of summaries or analyses.

Moody & Walsh published their article before the big data era, but the laws can be applied to big data as well. Especially combining (law 5) and deriving new data, i.e. analytics (law 7) have gained a lot of attention in the big data era. Sharing (law 1), presenting the results in appropriate format (law 6) and actually using the data (law 2) may as well play an important role in explaining big data value. Big data is perishable like any other data (law 3). Some, like Mayer-Schonberger & Cukier (2013), have argued against the importance of accuracy (law 4) on the grounds that if all the data is available at a detailed level, erroneous data have a minimal effect on the analysis. On the other hand, having all the data (n = all) is often impossible, and data quality is a frequent concern in big data projects, e.g. (Halamka 2014; Mathew et al. 2015; O'Leary 2013).

The importance of detailed data is, however, not reflected in the laws. Gathering details allows a more precise view of the phenomenon. More detailed data enables companies to reveal common patterns behind the observations. For example, gathering large volumes of geolocation data about passengers or vehicles allows visual representation of traveling patterns (Cai et al. 2014; Tao et al. 2014). Moreover, detailed data about customer valuations and behaviour allows personalisation of services (Amatriain 2013) through the micro-segmentation of customers. The number of observations relates to the completeness of the view: the more observations, preferably from different sources, the more comprehensive view of the phenomenon. Thus, the value of data increases with the detail. Paradoxically, however, the value of a single data point decreases as the amount of data increases. Yet, the amount of available detail increases the value. Law 6 must of course be taken into account when presenting the results.

2.2.2 From Data to Knowledge

The term **datafication** is used to describe the data deluge and rising importance of data. Datafication can be seen as a "information technology driven sense-making process" (Lycett 2013). A pre-requisite for this is that the data must be separated ("harvested") from its original context in the real world and augmented with appropriate additional data

(Piccoli & Pigni 2013). For example, GPS data points harvested from a bus could be augmented with other data such as date, time, bus identifier, driver and route information. After these steps, the data is in a "liquid" form, in the sense that it has a context, which means that it can be moved and used for another purposes (so-called secondary usage of data). Analytics and visualisation are typical steps required in the sense-making process. The augmented, contextualised – or liquid – GPS data can now be used for several secondary purposes, such as real-time schedule visualisation or driving style analyses when combined with fuel consumption data.

Previous research presents models that explain how data generates value. Ackoff (1989) presented a **data-information-knowledge-wisdom** –hierarchy (DIKW), which has been widely used and studied (Rowley 2007). The DIKW hierarchy is useful as it conceptualises the value creation process to an abstract model that helps understanding of the process. Pure GPS coordinates are simply symbols, i.e. raw data that "knows nothing". When the symbols are augmented with time and other attributes, they become information that provides answers to questions such as who, what, where and when. When was bus number 123 at the stop number 234? The next step, knowledge creation, happens when the information is given meaning. Knowledge is applied to contextual information that can answer questions such as how and why. Why is the fuel consumption variation so high, how can we reduce overall consumption? Knowledge becomes wisdom when it provides insights that help to understand and act on matters. If the analysis of combined data (knowledge) shows that fuel consumption varies because of the driving style, decisions according to that insight can be taken.

Another useful framework is the **virtual value chain** (VVC). This framework generalises the stages that are required when creating value from data. The stages (or steps) can be identified in real-world applications and therefore this framework provides a mental model that connects implementations and value creation processes. Building on Porter's value chain, Rayport and Sviokla (1995) defined VVC that consists of a five-step value creation process (Figure 9). The steps are gathering, organising, selecting, synthesising, and distributing. Publication I suggested separating the data and its usage. This distinction can be applied to the VVC. The steps of gathering and organising are data-related, and they cover aspects such as data acquisition from sensors, integration with other data, and data storage. The steps selecting, synthesising and distribution depend on data usage. These are activities such as filtering data for analysis or represented as artefacts such as analytical models, data visualisation, and information delivery tools. Value is expected to increase as data items from various sources are combined to form meaningful information chunks in the VVC process. A recent article (Miller & Mork 2013) applies the value chain to big data, suggesting current technologies for each of the steps.

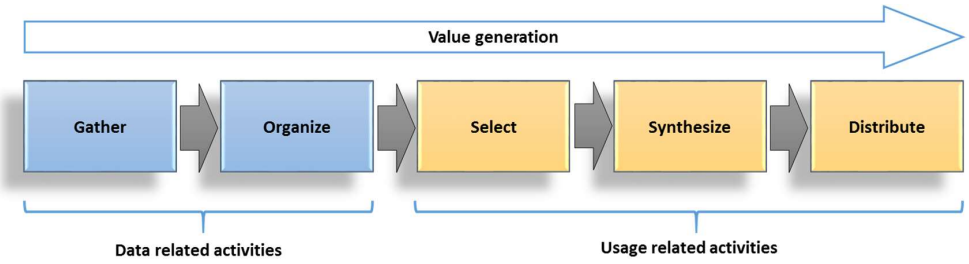


Figure 9. Virtual value creation process, modified from Rayport & Sviokla (1995).

Data does not become value automatically. Adding value requires an **innovation** to take place. Humans play an important role in driving the innovation process. One view of value addition is to look at the relations between humans, innovations and big data. This approach was discussed in Publication IV. Many innovations are human-driven, but increasingly data is a source of innovation. As a simple example, a human-driven innovation might start with a thought "If we had detailed data about how our customers use our software and what they think about it, we might be able to improve the user experience". This might then lead to logging users' actions, collecting social media posts, helpdesk data, and interviewing users. After analysis of the data improvements could be planned and implemented. Data-driven innovation takes place when insights drawn from the data, either manually or automatically, lead to improvements. A/B-testing is one example of data-driven innovation, predictive maintenance is another. Human-driven and data-driven innovations are not mutually exclusive. Instead, they can and should support each other.

Another human aspect to the value generation is related to the laws of information. It does not matter how much data you have or how sophisticated your analytics are if nobody uses the results. As Moody & Walsh (1999) state, people have to know that the data exists; where to find it; how to access it; and how to use it. This is one of the crucial challenges in the transformation to more data-driven forms of organisation. One solution for this is to embed analytics and insights into processes and decision-making routines (Bekmamedova & Shanks 2014).

2.2.3 Big Data Value Generation

The basic rationale behind the added value of big data is simple. *Current technology trends lead to generation of huge amounts of detailed, real- or near real-time data of various types. Enterprises then gather, combine and analyse that data in order to* ***create value***. As the amount of data grows and availability becomes easier and less expensive, enterprises begin to increasingly take advantage of it.

The laws of information show that **value can be difficult to measure**. Direct economic benefits, like additional revenues due to sales of data, are measurable. Indirect benefits

can be impossible to measure, such as the economic value of more informed decision-making due to the availability of real-time data. Moreover, data can drive qualitative benefits, such as customer satisfaction. Although the ultimate goal for an enterprise is to generate economic profit, defining value simply using economic terms provides a narrow view in the context of data. In order to understand the benefits of data, intangible factors need to be taken into account.

In spite of the measurement challenges, there is convincing evidence that data adds value. During the last two decades, trailblazers such as Google and Amazon have developed practical big data solutions[19]. These solutions have proved that they add value to their businesses. In fact, the trailblazers have built their business models on big data solutions. Previous studies, e.g. (Dehning et al. 2003; McAfee & Brynjolfsson 2012; Porter & Millar 1985) show that companies utilising data heavily gain a competitive advantage over their less data-driven rivals. However, enterprises are not the only ones that benefit. Über serves an example of the benefits that digitalisation and big data can offer to consumers. The estimated consumer surplus of Über in U.S was 6.8 billion dollars in 2016[20].

An obvious but relevant point to notice is that not all of the available data is of value for an enterprise. Careful analysis of the data sources, their availability and value for the company help to create a useful big data subset. Having a clear idea of how the data would benefit the enterprise is a good start. Gathering data without knowing what to do with it is in most cases a liability rather than an asset. The first thing for an enterprise is to recognise new data sources and their potential to the enterprise. The question to ask is: what new data sources exist that might add value to our business? Or, approaching the problem from another direction: what (possibly valuable) data do we have that we currently do not utilise?

Clearly, **value is a case-dependent matter.** An insurance company needs different data about its customers than a car dealer does. Both are interested in customer preferences and driving habits but for different reasons. The insurance company wants to assess the risks of their customers and follow the company's online reputation. Valuable data sources for these purposes include sensor data (e.g. on-board diagnostics data like speed, heavy acceleration or braking), external transactions data (e.g. alcohol purchases), geolocation data (e.g. driving in urban or rural areas, and kilometres driven) and social media data (e.g. the tone of discussions in Facebook). The car dealer concentrates on marketing and lead generation. They might use social media channels to strengthen their brand through visual stories or engage customers with interactive content and build two-way communications with customers (e.g. answers to questions or chats). Both of the

[19] See e.g. https://mapr.com/blog/5-google-projects-changed-big-data-forever/ and https://aws.amazon.com/about-aws/

[20] A joint study by the University of Chicago, University of Oxford and Uber https://news.uchicago.edu/article/2016/09/16/big-data-gives-insight-appeal-services-uber

companies can potentially benefit from big data, but each of them must do their own thinking on how to achieve it.

Apart from the different needs there is another question related to value: the availability of the data. The data may even be unavailable for several reasons, e.g. a competitor may have control over the data. For the data that is available, the cost of the data may be high due to the purchase price[21]. The data may require complex management and analytic skills, such as parsing the digital trail from different types of data gathered from various data sources. The users may be reluctant to give access to the information, at least unless they get something in exchange. For example, an insurance company may be able to obtain onboard diagnostics data from drivers, if they give payment discounts to careful drivers.

An interesting question is **how does big data add value?** Vargo and Lusch (2004) proposed a new perspective for value creation. They claimed that the old dominant value creation logic that focused on tangible resources has been gradually replaced by a new service-dominant market logic. Focal areas of this perspective are 1) developing competencies, knowledge and skills that may create competitive advantage, 2) identifying stakeholders or customers that may benefit from the developed assets and 3) develop customized value propositions together with the customers (Vargo & Lusch 2004). Since then the service-dominant value creation perspective has been widely studied, e.g. (Akaka & Vargo 2014; Grönroos & Voima 2013; Lusch et al. 2007; Payne et al. 2008). The service-dominant view emphasizes competency development, customization and co-creation in the value creation processes.

The service-dominant view of value creation seems to align well with big data assets. The results of Publications VI and III underline the need for competency and skills development in order to derive value from data. Correspondingly, the conclusions of Publication I show that the value is case-dependent, i.e. customized. Publication IV developed a framework that requires co-operation between IT and business persons. In this sense, the business side can be viewed as an internal customer, who has an active role in the value creation. Therefore, this research relies on service-dominant value creation perspective and refers to value as the impact to an (incumbent) enterprise, gained from investing in co-creation of (big) data assets and capabilities (see Publication VI).

However, extant big data literature introduces several, partly overlapping viewpoints that explain how big data drives value creation. The viewpoints are mainly based on trailblazer experiences and **there is a need for further research in this area**. The following paragraphs summarize typical value creation approaches presented in recent big data literature.

[21] See (Dalessandro et al. 2014) for a discussion about the cost of data and a method to estimate the added value.

Manyika et al. (2011) identify five broad categories to leverage big data in value creation. Big data can be used to create transparency (timely, detailed information to users). Accurate experiments become possible with more detailed data. Thirdly, big data helps in segmenting customers, for example. Automated or supported decision making could substantially improve decisions. Finally, big data helps to innovate business models, products and services.

Schmarzo (2013) uses the 3V definition as a starting point. According to his view, the capability to combine and analyse all the details of many kinds of data in real-time brings the benefits. In other words, volume, variety, velocity and advanced analytics explain the value potential of big data. These factors enable the business benefits: new products, services, business models and improved decision making.

Mayer-Schonberger & Cukier (2013) also share the view that the details add value. In addition, they emphasise correlation analysis; in many practical decision-making situations causality is not required. Moreover, they separate primary and secondary usage of the data. Primary usage means using the data for its original purpose, e.g. monitoring a motor's temperature for overheating. Using the temperatures for something else, such as for predictive maintenance purposes, is an example of secondary usage. According to their view, the secondary usage of the data is the key to adding value to the business. Based on the secondary usage, they perceive three scenarios for adding value: Data owners can monetise the data they own, i.e. sell the data. Secondly, those who have capabilities to analyse the data, can monetise their competence. Finally, those with a "big data mind set" can innovate new products and services that utilise big data.

Van't Spijker (2014) explains that big data enables enterprises to innovate new business models that create value in new ways. He lists five value generation models for data-driven business: "1) selling data directly, 2) innovating products through data, 3) swapping commodity offerings into value-added services, 4) creating interaction in the value chain, and 5) creating a network of value based on data exchange". He also notes that analysing the digital trail that users leave behind them at platforms can create significant value.

Publication III analysed big data case studies found in academic literature. The 49 cases covered most of the areas identified above. Another explanation is presented in Publication VI: a process theory based "recipe", i.e. a model, for value creation. The model explains, how big data investments convert to economic performance, and why the conversion sometimes fails.

Yet another way to look at the value of data is crime. Media articles about data breaches are common, as discussed above. Although there may be other motives than money, such as resentment, presumably in most cases the basic driver is the fact that data is valuable. Credit card data or health records are worth money in criminal markets. Industrial espionage can cut product development costs effectively. With regard to value, Volkswagen's "emission-gate" provides an enlightening example. The sensors and

related software of certain Volkswagen models purposely produced false CO2 emission data when tested. Lower emissions, of course, are a major benefit both economically and for marketing. When the cheating was revealed, the CEO had to leave the company and eventually the total costs for the company are expected to be more than 16 billion euros.

2.3 Big Data Impacts

"The future is not what it used to be."
– Yogi Berra

Digitalisation affects all levels of society. The impact of big data can be seen through society as well, since big data follows from digitalisation. This chapter discusses the impact on individuals, enterprises, ecosystems and society. It should be noted that the impact assessment here covers only a few aspects of these broad topics; the intention is to provide a brief introduction to the topics. However, the impact on enterprises (and ecosystems) is of course implicitly included throughout this research.

2.3.1 Personal Level Impacts

The trend towards a **personal data-driven culture** is ongoing. People have always gathered data, such as facts and references (opinions), before they make decisions. The Internet and search engines like Google make the gathering process much easier and faster than before. In addition, with personal computers and mobile devices more and more people are able to access data and collaboration networks. This development has "democratised" the access to knowledge, at least when compared to the age before the Internet. In developed countries such as Finland almost everyone has the possibility to search and access data relevant to their decisions. This does not equal better decisions, of course, but at least it makes them possible. Moreover, the technologies boost **personal productivity**. A few on-line searches may reveal much more data on the subject at hand than a whole day's visit to the library before the Internet era.

Another consequence of digitalisation and big data is the effect on job markets. Digitalisation creates new, technology and information systems related jobs[22] while destroying many other types of jobs, such as those in service, sales and office work (Frey & Osborne 2013). It seems that at least in the transition phase the number of jobs lost will be larger than the jobs created; e.g. the World Economic Forum (2016) predicts the seven million jobs will be lost while two million jobs will be created by 2020. Those with "traditional" job skills will find the effects negative, whereas for those, who are able to apply technology and big data, the coming years will be good. For example, just 10-15 years ago there were no such occupations as mobile app developer, social media manager

[22] For example, Gartner predicted 4.4 million new, big data related jobs to be created worldwide in only three years (http://www.gartner.com/newsroom/id/2207915)

or cloud computing specialist. The transformation creates a **structural unemployment problem**.

As an example, self-driving cars will make most taxi and bus drivers redundant. At the same time, new job opportunities will open to create the infrastructure and services that the automation requires. Building the services requires different skills from driving ability. The drivers are not likely to be able (or willing) to transform themselves to become digitally skilled ICT professionals. At the same time, three out of four CEOs consider (digital) skills shortage a serious concern to their business[23].

2.3.2 Enterprise and Ecosystem Level Impacts

An increasing number of enterprises see big data as a major opportunity to create value. However, the data-driven approach is still a new paradigm for most organisations (Shen & Varvel 2013). Data-driven business models and decision-making requires a new approach to innovation, skills and business management.

Innovation management capabilities are relevant to all enterprises. Dyer et al. (2011) present four principles that the most innovative organisations apply: 1) innovation is everybody's job, 2) both incremental and disruptive innovations must be considered, 3) innovations are best developed in small teams, and 4) managed risks must be taken. These are by no means trivial requirements: e.g. (Sandberg & Aarikka-Stenroos 2014) identify restrictive mind-sets, lack of discovery competences and unsupportive organisational structures as three main internal barriers to radical innovations in incumbent firms.

As discussed earlier, digitisation and big data require new, digitally skilled talent. Enterprises must consider, how to **find and recruit potential skills**. Kane et al. (2015) surveyed a large number of business executives, managers and analysts around the world. One of their findings was that the vast majority of these people want to work in enterprises that are digital leaders. This is an important finding for enterprises as there is a shortage of digitally skilled employees. A digital leader is more likely to attract job candidates with digital skills and thus it has better chances of hiring people who are capable of creating or capturing value from data. Another aspect that enterprises must consider is transparency. Current and potential employees discuss companies with their peers. A potential job applicant now has access to peer reviews and even insider opinions. People discuss company related matters on social media platforms. In addition, services such Glassdoor[24] enable employees to anonymously review their current companies.

[23] PwC 18th annual global CEO survey 2015 (http://www.pwc.com/gx/en/hr-management-services/publications/assets/people-strategy.pdf)

[24] Glassdoor (https://www.glassdoor.com/index.htm) is a jobs and recruiting site that allows you to see which employers are hiring, what it's really like to work there according to employees, and how much you could earn.

Start-ups typically apply lean methods, prefer low fixed costs and rely on ecosystem partners. This emphasises the role of data in many ways. Lean methods mean lower hierarchies and more networks. Low fixed costs means external resources instead of a company's own resources. Ecosystems mean more partners. The data must flow smoothly between parties, creating common understanding, transparency and trust. Due to volatile business environment undergoing rapid changes, also incumbents should employ agile business practices, and rely on partners and ecosystems (Weill & Woerner 2015).

In the era of big data, the role of **ecosystems** becomes more important for the reasons stated above. The business ecosystem concept was coined by Moore (1993), who described the life-cycle ("evolutionary stages") of a business ecosystem. The participants of a business ecosystem depend on each other and the success of the ecosystem is crucial to each party (Afuah 2000). A recent study suggested that firms should develop digital ecosystems in order to face the challenges of digitisation (Weill & Woerner 2015). Iansiti and Levien (2004) propose a framework that helps an enterprise to assess the health of the ecosystem as well as their own position and role in the ecosystem. Efficient use of data is an asset for an ecosystem. An interesting question (beyond the scope of this research) is how the ecosystem structures and its participants support the sharing of data. After all, the participants are competitors in many situations. For a discussion about business models in situations where competing companies collaborate, see e.g. Ritala et al. (2014).

One future ecosystem is built around self-driving cars. Currently this ecosystem is at birth stage, a term coined by Moore (1993), when all the major car manufacturers are innovating their value propositions and protecting the details of their key ideas from competitors. The ecosystem is still rather volatile; new entrants are coming into the market, technologies and software solutions are developing rapidly. As the market is large and the expansion of the ecosystem is just around the corner, many enterprises consider the steps required as worth taking. There are numerous opportunities for sensor manufacturers and software companies and so on. However, each enterprise must evaluate their position and role in the ecosystem in order to know what strategies to select. For example, concentrating on expensive light detection and ranging (LIDAR) sensors may be attractive now. However, within a few years this will be a commodity market. A commodity strategy does not work well in complex, innovative ecosystems (Iansiti & Levien 2004) like the self-driving car ecosystem. Developing value-added services for a certain narrow niches based on the secondary usage of the data that self-driving cars produce offers an example of a new business model innovation that a software company might choose.

2.3.3 Society Level Impacts

Digitalisation and big data affect society in several ways, including social implications, regulation and productivity. According to a recent study (Frey & Osborne 2013), almost half (47 %) of U.S. jobs are at high risk of being computerised in the next few decades. Moreover, their study categorises an additional 19 % of the jobs as medium risk

occupations, i.e. the expectation is that only one third of the occupations are relatively safe with regard to digital transformation. The greatest digitalisation impact will hit areas such as service, sales and office work. Although these are rough estimates, Frey and Osborne (2013) show, where we are heading. This trend has **social implications**. Some will benefit from the transformation, but many will be made redundant. Division lines across society may increase; digitally-capable and digitally-incapable people may perceive their future fundamentally differently. Where those who are capable see opportunities, others may perceive that the future has nothing to offer. Political decisions must be taken to minimise the negative effects.

Regulation of big data culminates in privacy. On the other hand, current regulations should be reviewed in order to make it easier for enterprises to enter the market with data-driven innovations. Big data and open data resources potentially create economic activity, which benefits society. At the same time, policy-makers must consider potential privacy issues. An iconic example of this are health records. It is obvious that if various parties were allowed to share patient data freely, it would probably benefit patients and prevent fraud and other crimes, but the question is where to draw the line. Who should be allowed to see the data and on what grounds?

There are also other, complicated regulation related big data matters that need attention from policy-makers. Mayer-Schönberger & Cukier (2013) discuss the threats of actions based on probability and the decisions taken by algorithms. **Probability-based actions** refer to situations where actions against (usually unwanted) behavior are initiated before the actual behavior takes place. For example, one possible scenario is that the law enforcement personnel could arrest an individual based on data suggesting that the person will probably commit a crime. At least in Finland and most other democratic countries one key principle of legislation is that consequences are related to actions, not to intentions. Their second concern, **automated decision-making**, is actually happening already. For example, banks routinely use algorithms that invalidate a customer's credit card, if suspicious activity is found. However, there are relevant questions that must be discussed: On what data do the algorithms base their decisions? What is the decision-making logic? How far can this development progress?

Productivity, the root cause behind digitalisation and big data, benefits society. For example, it is estimated that automated traffic in Finland could save approximately 10 – 20 billion euros per year and free up to 100 billion euros of capital[25]. The same report adds, however, that the savings will not be realised without changing current legislation and procedures. Another Finnish study (Linturi et al. 2013) evaluated the 100 most potential technological solutions for Finland. Big data was considered the number one, most promising new technology. Manyika et al. (2011) claim that big data represents a 250 billion euro potential in *annual* value in Europe's public sector administration. For

[25] Linturi 2013: Loppuraportti - automaattisen liikenteen metropolivisio (https://www.sovelto.fi/application/files/9214/2331/4161/Loppuraportti_automaattisen_liikenteen_metropolivisio.pdf)

productivity reasons everything that can be digitised, will be. And because digitalisation leads to a data deluge, (big) data will be the new oil for modern economies.

3 Theoretical Foundation

This research uses a multi-disciplinary view. A multidisciplinary approach is required to explain how enterprises can adopt a more data-driven approach in their businesses, as big data has profound effects on many aspects of an enterprise. Figure 10 illustrates the approach. Strategic management must be considered, as big data enables new ways of doing business. New ways indicate changes, which in turn require innovation. Thus, innovation research is one discipline related to this research. Big data naturally relates to information technology, and driving value from data requires human actions. This leads to information systems research. The combination of these three disciplines, information systems research, innovation research and strategic management research, is at the core of the theoretical playground relevant to this study.

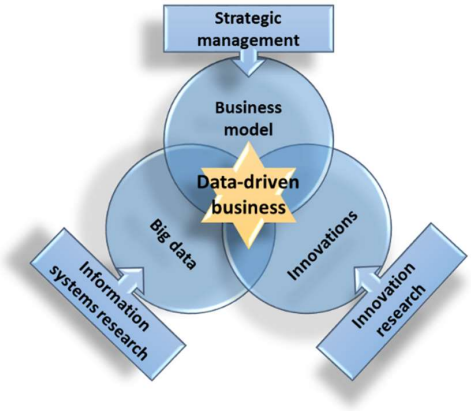


Figure 10. Multidisciplinary approach of the research.

3.1 Strategic Management

Michael Porter's value chain and five forces model, e.g. (Porter 1991; Porter & Millar 1985), are well-known strategic management frameworks. These models apply to the industry or business level and they have been applied to big data value creation as well, e.g. (Porter & Heppelmann 2014; Schmarzo 2013). The value chain is an internal view of the business, whereas the five forces model provides an outside-in industry view. Porter's models explain the source of competitive advantage by two factors: low cost or differentiation. Low costs allow competitive pricing options, which ultimately might put rivals with similar offerings out of business. Personal computer manufacturer Dell quickly gained market share by developing an effective low-cost business model that utilised disruptive technologies, namely the Internet and e-commerce (Dyer et al. 2011). Differentiation is a required alternative to the low-cost approach, as not everyone can

compete with prices. Differentiation simply means offering some unique features that are of value to customers who are willing to pay a premium. As an example, consider Apple. Porter's framework emphasises competition and rivalry. The goal for a business is to gain a leading position or even a monopoly in a selected industry. By beating the competitors in activities that are crucial to competition a business can achieve a sustainable competitive advantage.

A resource-based view of the firm or RBV (Wernerfelt 1984) is another established strategic management framework. This combines the internal and external views of the previous models. In the resource-based view of a firm, resources and capabilities are key concepts that explain the firm's competitive advantage. Each enterprise has a different set of resources and capabilities, since each has a different history, skills and organisation culture. Resources can be tangible, intangible or organisational capabilities (Collis & Montgomery 1995). An enterprise may own a hardware environment suitable for big data processing (tangible resources), it may have a good brand name or home-build analytical algorithms (intangible resources), and an organisational culture that supports experimenting and data-driven decision making (organisational capabilities). Teece (2007) introduced the concept of dynamic capabilities, defining them as

> "the firm's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments",

which complements the RBV model by explaining how firms renew their competencies in order to adapt to the changing business environment.

In order to gain sustainable competitive advantage, the competencies must meet the following criteria (Barney 1991): They must be valuable (i.e. able to create rents), rare, costly to imitate, and non-substitutable. Later, Barney (1995) attuned the last attribute into the form of the question "Is a firm organised to exploit the full competitive potential of its resources and capabilities?" Barney's model is often referred as the VRIO-tool. Figure 11 illustrates how the key attributes (criteria) of the model can be used as a tool. This tool is useful for assessing both current capabilities and those resulting from the planned changes.

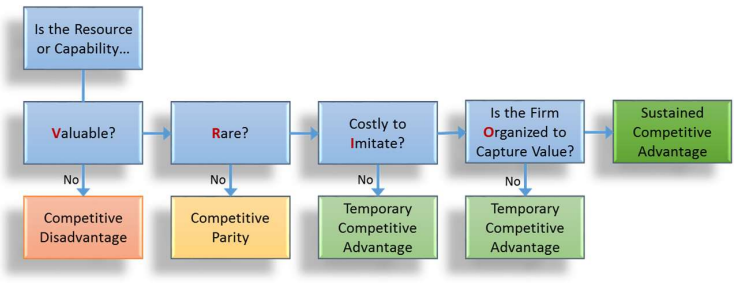


Figure 11. The VRIO tool, modified from Barney (1991; 1995).

This research relies on the RBV for the following reasons. In general, the RBV is a widely accepted theory that is capable of explaining why some enterprises are more profitable than others. For incumbents the RBV is understandable and tools such as the VRIO are practically applicable. Next, rather than beating the competition, value co-creation and the role of ecosystems is becoming more important, e.g. (Weill & Woerner 2015). The participants in a business ecosystem depend on each other and the success of the ecosystem is crucial to each party (Afuah 2000). Iansiti & Levien (2004) note that trying to extract the maximum value from a network that an enterprise does not control is a fundamentally flawed strategy. The RBV meets these conditions better than e.g. Porter's framework that underlines fierce competition. Finally, an implicit concept of the multidisciplinary approach of this research are **capabilities**. As digitalisation and big data re-shape the business landscape, firms must develop new organisational capabilities in order to take advantage of the opportunities. An enterprise cannot simply gain significant benefits from big data by just perceiving it as a technical exercise.

Capabilities are related to the business model. The term **business model** is often used interchangeably with strategy (Burkhart et al. 2011). Many definitions exist for both of these terms, e.g. (Amit & Zott 2001; Casadesus-Masanell & Ricart 2010; Teece 2010; Timmers 1998). However, there seems to be some consensus that these concepts differ from each other (Zott et al. 2011). Strategy focuses on competition and product-market matters. A business model describes how the strategy is implemented, i.e. how the company creates value. As an example, described by Dyer et al. (2011), Dell conquered the PC markets by developing a disruptive business model. Incumbent PC manufacturers used retail chains, whereas Dell utilised new technologies and sold the devices directly to the end-customers through web shops. This new business model gave Dell significant benefits. They achieved a much lower cost structure by eliminating retailers from the chain. The incumbents could not match the lower prices. For the incumbents, adopting the direct-selling mechanism was almost impossible, since they had no capabilities to deal with the end customers and selling directly would damage their current retailer sales chains. Dell's product-market strategy was essentially the same as their incumbent rivals. They offered standard PCs to various customer segments; however, their business model was much more capable of creating value for both to the end-customers and to Dell.

With regard to understanding the big data impact, a business model is an essential concept. This research uses the following definition for a business model, adopted from Osterwalder & Pigneur (2010):

> "A business model describes the rationale of how an organisation creates, delivers, and captures value".

Big data can drive innovations that may change the existing business models or create new ones. The exploitation of big data will transform businesses, i.e. the ways resources

and capabilities are arranged; it will change the ways businesses operate and make decisions. Eventually this will lead to new capabilities such as a more data-driven organisation culture. A business model can be concretised using frameworks such as the business model canvas (Osterwalder & Pigneur 2010). Approaching big data with the business model canvas enables companies to link the impacts to business functions.

Furr & Dyer (2014) present a slightly modified version of the business canvas framework. Their business model components, shown in Figure 12, include the solution (which is further divided to value proposition and pricing strategy), cost structure (activities and resources), and customer acquisition (relationships and channels). This version focuses especially on the components that are most relevant to innovations, offering an appropriate tool for evaluating the impact of big data. Using a tool like the business model canvas helps to remain focused on the relevant issues and questions. For example, using the tool to compare a proposed business model with existing ones, or a company's own or a competitor's, reveals differences and possible weak spots.

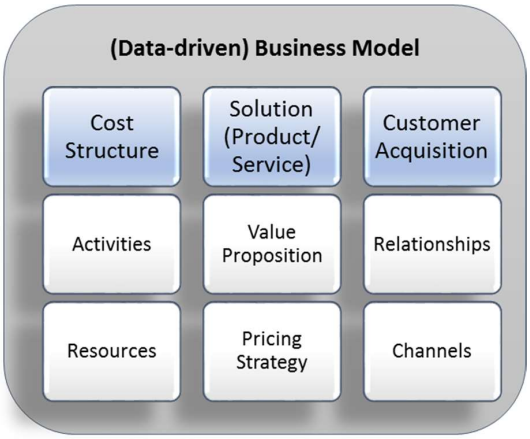


Figure 12. Business canvas framework (Furr & Dyer 2014).

It should be noted that most of the current business model frameworks take the viewpoint of a single enterprise. Although it is beyond the scope of this research, an interesting point of view would be to explore the business models of (possible) big data ecosystems. New, data-related ecosystems are emerging, providing data services (data-as-a-service) or analytics (analysis-as-a-service) to the parties of the system (Chen et al. 2011). One related example is given by Westerlund et al. (2014), who discuss the challenges of defining ecosystem business models for IoT and make some propositions for the future.

Changing the business model is equal to **business transformation**. Venkatraman (1994) created a framework for IT-enabled business transformation. His five-level model, shown in Figure 13, basically states that the greater the potential business value, the greater business transformation is required. Thus, the model implies that the value of IT is based

on firm's resources and capabilities. Implementing evolutionary changes, such as an isolated, specific-purpose piece of software or integrating a firm's systems may be valuable, e.g. in terms of reduced costs. However, reaching the maximal benefits requires more revolutionary approaches, namely redesigning the processes, redefining the business network, or coming up with a new business scope. In the turbulent era of digitalisation, firms should innovatively seek business model innovations and then leverage IT to create the required capabilities.

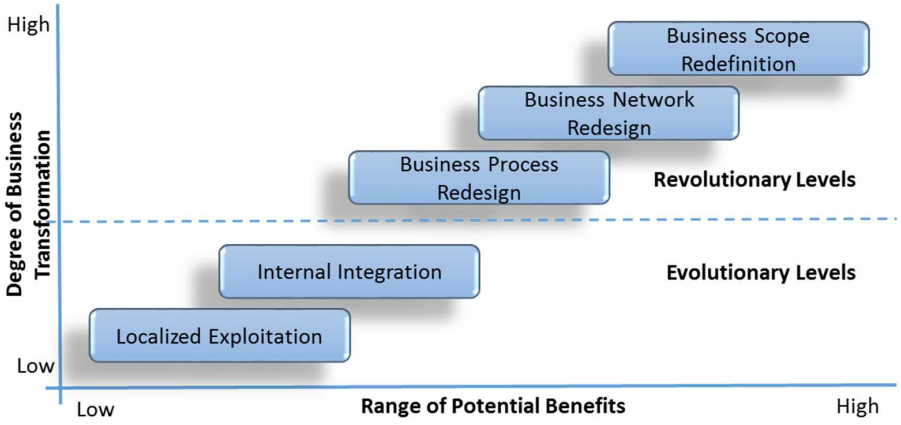


Figure 13. IT-enabled business transformation framework (Venkatraman 1994).

Venkatraman's business transformation framework aligns with the RBV and the business model canvas. It complements the strategic management part of the theoretical foundation of this research. Together these components form a coherent model that builds both on established strategic management principles and provides tools that help understand the impact of big data in the business context.

3.2 Innovation Research

Business transformation requires innovations and management decisions. Digitalisation implies changes and thus forces firms into a renewal process, which requires innovating. Innovation and change are tightly related; to innovate effectively aims to change something. Many different perspectives towards the concept of innovation have been taken in several disciplines. Thus, there are different definitions of the term, e.g. (Ettlie & Reza 1992; Schumpeter 1942; West & Anderson 1996). This research uses the definition by Baregheh et al. (2009) that views innovation as a process and includes strategic aspects:

> "Innovation is the multi-stage process whereby organisations transform ideas into new/improved products, service or processes, in order to

> advance, compete and differentiate themselves successfully in their marketplace".

The **human-driven approach to innovation** states that innovations stem from people. A good example of a person utilising this approach was Steven Jobs. Dyer et al. (2008) claim that by training certain skills systematically, anyone can accelerate associative thinking, which is a key concept of human-driven innovation. The four discovery skills are: observing (especially looking for surprises), networking with people from different backgrounds, experimenting a lot, and questioning the status quo. Innovative people spend more time exercising these skills than others. New ideas emerge from self-transcending knowledge (Scharmer 2001) as a result of associations. Scharmer extends the traditional explicit vs. tacit knowledge categorisation by dividing tacit knowledge in two, see Figure 14. Tacit-embodied knowledge is hidden in the working processes, whereas self-transcending knowledge ("tacit knowledge prior to its embodiment") is the ability to recognise the forthcoming.

A short example clarifies the terms. Explicit knowledge about a smart phone refers to an understanding of the specifications and physical attributes of the device. Manufacturing processes contain a great deal of know-how, i.e. tacit-embodied knowledge. Self-transcending knowledge is the process of thoughts that led to the idea of the smart phone in the first place. In the business environment, recognising emerging opportunities are of course important. New ideas, seeds of innovations, often emerge sub-consciously, when our brain associates things with one another.

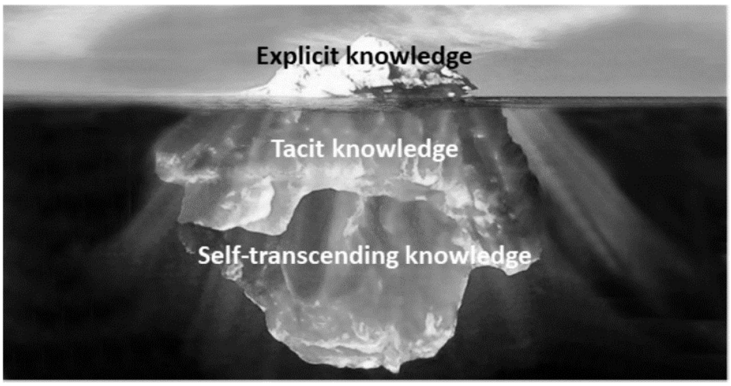


Figure 14. Three forms of knowledge (Scharmer 2001).

Enabling human-driven innovation requires that individuals use part of their time in seemingly unproductive discovery-skills development. This is important for developing self-transcending knowledge. Another requirement is that the organisation's culture and management practices must support risk-taking and allow mistakes. Dyer et al. (2011) present four principles that the most innovative organisations apply: 1) innovation is

everybody's job, 2) both incremental and disruptive innovations must be considered, 3) innovations are developed in small teams, and 4) managed risks must be taken. These are by no means trivial requirements: e.g. (Sandberg & Aarikka-Stenroos 2014) identify a restrictive mind-set, lack of discovery competences and unsupportive organisational structures as the three main internal barriers to radical innovations in incumbent firms. Human-driven innovations may require new data sources or processing capabilities, which creates a link between innovation and big data.

Another approach relies on data and automation. A novel approach termed **innovation automation** is data-driven. Data-driven innovation suggests that the innovation processes could and should be automated (Shaughnessy 2015). This approach puts technology and (big) data at the core of the innovation processes. More and more data is becoming available, technological and analytical capabilities are increasing, and data processing costs are decreasing. Utilising automation and vast volumes of different data will produce a more holistic view, leading to more data-driven decisions and more agile innovation processes. An example of this approach is Netflix, where data-driven innovation, combined with rapid experiments, significantly speeds up the innovation process (Amatriain 2013). Correspondingly, Manyika et al. (2011) claim e.g. that manufacturing firms can reduce their product development costs by 20-30% as well as achieve up to a 50% faster time-to-market cycle by utilising big data. One case example on utilising data to drive innovations is documented by Jetzek et al. (2014). However, innovating with big data is still in its infancy, although data-driven innovation may enable business model innovations, i.e., there is a connection between the business model and big data.

A modern enterprise should utilise both human and data-driven innovations. The paradigm shift forces enterprises to innovate, re-think their products, services and eventually business models. Innovation capabilities and management actions lay the foundation to the business transformation. However, the kind of activities that productive innovation processes require are in contrast with the current procedures of most incumbent firms. Since Taylor (1911) published his principles of scientific management, incumbents have organised themselves to be efficient. They concentrate on minimising all kinds of waste (such as allocating time to "unproductive" activities) and avoiding errors, which is reasonable as such. Paradoxically these procedures effectively kill the seeds of innovation and leads to a culture that neglects the capabilities required to deal with innovations. This indicates that focusing on the organisational aspects of innovation is an important factor for a successful business change for any enterprise.

The nature of innovation is related to business transformation. Christensen (2013) distinguishes between **incremental and disruptive innovations**. Incremental innovations refer to enhancements of existing products or services whereas disruptive innovations refer to new technologies that potentially outperform existing solutions. Digitalisation and big data are potentially disruptive, as the examples like Dell or Google show. However, the effects of emerging disruptive innovations are difficult to predict. In fact, according to Christensen (2013), any forecast with regard to the disruptive innovations is deemed to be false. In spite of this, it is important to understand the

difference. Disruptive innovations require different business models and capabilities from incremental innovations. Disruptive technologies are especially difficult for incumbent firms as discussed above (in the "Challenges of Big Data" section).

This research underlines the importance of recognising the disruptive potential of big data. Innovation management and processes are key elements in the process of making use of big data and enterprises must consider the takeaways of innovation research discussed here. Although it is not an easy task, continuous improvement of both human and data-driven innovation processes are one of the required elements of the transformation; innovating with big data is much more than the old adage "Think out-of-the-box!".

3.3 Information Systems Research

Information systems research draws from behavioural and design sciences, exploring the combination of technology, organisation, and people. A wide variety of methods are applied, and no single common theoretical perspective can be identified. However, there seems to exist a common set of philosophical assumptions, mostly based on natural sciences (Orlikowski & Baroudi 1991). In addition, it is widely accepted that information systems research should both make theoretical contributions and provide assistance to practitioners in solving current problems (Benbasat & Zmud 1999; Iivari 2003).

Combining theory and practice is not always straightforward and some researchers like Orlikowski & Iacono (2001) have proposed that information systems research should concentrate more on the core matter – the information technology artefact. One of the two established research approaches in information systems research, design science, focuses on artefacts. Hevner et al. (2004) state that design science "creates and evaluates IT artefacts intended to solve identified organisational problems". Building on previous design science literature, Peffers et al. (2007) defined the design science research method (DSRM) framework, which provides a process model for creating artefacts.

Another direction relies more on the behavioural sciences and thus underlines the role of humans. For example, Sein et al. (2011) proposed an approach that combines action research (Lewin 1947) and design science. Action design research, as they named the approach, emphasises the view that the user organisation shapes the IT artefacts during their development and usage. Thus, combining user-intervention centric action research and IT artefact-oriented decision science provides a more complete approach than either of them alone. Action design research aims to improve organisational capabilities over time and requires long-term commitment from the participating firms.

Indeed, one of the strengths of information systems research stems from the combination of behavioural and design sciences; technology and behaviour are inseparable in an information system. For a researcher with a pragmatic worldview – like the author of this dissertation – the principles of information systems research resonate well. Moreover, in novel areas such as big data research and implementations both new artefacts and human

contributions are required to explain and capture the value of data. Information systems research is a versatile discipline for this purpose.

Although previous research has convincingly demonstrated the value of information systems, the connection between competitive advantage and information systems is not obvious. Many resources such as financial assets, hardware capable of processing big data or the value of a patent can be related to competitive advantage quite straightforwardly. However, information system resources like a big data solution typically contribute indirectly. A fine-tuned customer analytics system may have a direct impact on sales, but it also may have indirect effects, such as better customer understanding which in turn leads to better decisions, for example. These kinds of indirect relationships are difficult to assess and valuate.

The RBV explains how tangible or intangible resources and core competencies create a firm's competitive advantage. Wade & Hulland (2004) have studied the combination of the RBV and information systems. They conclude that the RBV is a useful paradigm for information systems research, because it provides a valid model for evaluating the value of information systems. Based on a review of information systems research papers they proposed a set of key information systems resources, followed by a description of those resources using traditional resource attributes used in the RBV literature. This research establishes the connection between information systems resources and the RBV by building on their ideas as discussed below. Information system capabilities are summarised in the Table 1.

Table 1. Information systems resources, modified from Wade & Hulland (2004).

Inside out	Spanning	Outside in
• IS infrastructure	• IS business partnerships	• External relationship management
• IS technical skills	• IS planning and change management	• Market responsiveness
• IS development	• Utilisation of big data assets	
• IS operations		

Inside out capabilities are internally focused, consisting of e.g. IT technology and cost related resources. Outside in capabilities emphasise customer relationship creation and competitor understanding. The third category, spanning capabilities, integrates the two other capabilities, e.g. by linking IT with business processes. The capabilities and categorisation shown in Table 1 were adopted from Wade & Hulland (2004) with the exception of big data assets in the spanning capabilities category. The value potential of big data justifies the addition. Big data is technical by nature, but the value is realised only when business functions make use of it. Therefore, it belongs to the spanning category.

Each of the capabilities listed in Table 1 are briefly presented below in the context of big data and the seven laws of information (Moody & Walsh 1999) that were discussed earlier. Moreover, the capabilities are reflected in the VRIO model (Barney 1995), also discussed earlier.

IS infrastructure. Some enterprises, such as Amazon, have developed their own infrastructure[26] that has significantly added value to the business, but this is rare. In most cases infrastructure is a mandatory pre-requisite for information systems and thus subject to cost optimisation. Cloud based infrastructure platforms have become popular in recent years due to their flexible cost structures. With regard to the laws of information, enough resources should be allocated to manage the details. Besides that, IS infrastructure is hardly a significant source of competitive advantage for an average Finnish enterprise, as it is not rare nor is it a non-imitable resource, i.e. the choices are available to competitors as well.

IS technical skills. Technical skills and knowledge are human-driven capabilities that can add significant value to information systems development and management. Big data skills, especially analytical talent, are scarce today, and will also be in years to come, which underlines these capabilities. Technical skills relate to several laws of information. Ensuring accuracy (related to law 4), combining data from various sources (law 5), delivering appropriate visualisations and representations (law 6), implementing analytics (law 7), and delivering detailed data when required are examples of tasks that require technical skills. Technical capabilities related to big data can be valuable, rare and difficult to imitate, i.e. a potential source of competitive advantage.

IS development should be future-oriented in the sense that new technologies may potentially add value to a business. This involves being aware of technology trends and experimenting as well as planning the life-cycle of currently used technologies. Development capability relates also to IS planning and change management, avoiding declining technologies and architectural dead-ends is preferable. Being able to keep up with the increasing pace of change and turbulence of the current business environment is potentially a significant source of value. Agile methods such as Scrum[27] have gained popularity in recent years. With regard to big data, technologies are evolving rapidly and combining new architectures with legacy systems can be challenging. Delivering accurate data (law 4) that is integrated in legacy systems (law 5) at the right time (law 3) in appropriate format (law 6) is a potential source of competitive advantage – a rare and hard to imitate capability that creates value.

IS operations are mostly about managing the operations as cost-efficiently as possible while providing appropriate quality to the business. This includes, for example, minimising downtime, budget control and maintenance costs like annual software and license fees. Cost efficient operations are especially important if the firm seeks cost

[26] See e.g. https://aws.amazon.com/what-is-aws/
[27] For an overview of Scrum, see e.g. https://www.scrumalliance.org/

leadership position among its competitors. Producing accurate (law 4) and when appropriate detailed data about operations and combining (law 5) it with e.g. benchmark data is always valuable. IS operations are no exception. However, the options in this area are available to everyone. IS operations should be cost-effective enough, compared to competitors, but it seems that no significant advantage can be gained here.

IS business partnerships represent internal relations between IS function and the business areas of the enterprise. The more important the role of IS becomes to a business, the more important this is. The value of data increases with use (law 2) and sharing tends to multiply the value (law 1). Moreover, one of the findings of Publication II was that the interplay between IT management and the management group of the company seems to play an important role in big data adoption. With regard to the VRIO model, enterprises that focus on making use of big data as a joint effort between IS and business, and organise their processes accordingly, are probably able to generate value in terms of operational efficiency, for example.

IS planning and change management. Planning and managing technology architectures that are future-proof, i.e. flexible enough to adapt to future changes cost-effectively is a valuable capability. Currently one of the main challenges is how to mix new big data related software and technology with existing architectures. Rapid big data technology development by all major hardware and software vendors does not make things easier, although maturing technologies are beneficial as such. The dominant big data architecture now is Hadoop-based[28], but the future is hard to predict. This capability is related to IS development. In many practical cases planning, development and change management are tightly interwoven in a single, agile process.

Big data utilisation. Throughout this dissertation the researcher considers (big) data as an asset, i.e. a potentially valuable resource that should be taken seriously. The capability to utilise available data resources requires business domain understanding as well as technical skills. All of the seven laws of information apply to big data utilisation. Added value and potential competitive advantage should be evaluated case-by-case. The impact can be significant in terms of added value and a competitive edge, as numerous examples discussed e.g. in the Publications show.

External relationship management represents the firm's capability to manage relationships with external stakeholders such as hardware and software vendors. Many enterprises have outsourced most of their ICT activities. Developing and managing these relationships effectively can contribute to the firm's performance. Moreover, in the era of digitisation and big data the role of ecosystems increases. As with internal partnerships, increased use (law 2) and sharing (law 1) tends to add to the value of data. Partners may have pieces of data that combined (law 5) produce useful insights.

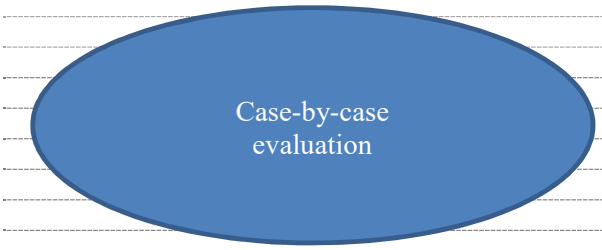
[28] For an overview of Hadoop, see e.g. http://hadoop.apache.org/

Market responsiveness. Information systems should be able to quickly respond to market changes. Rapid system implementations, efficient information delivery and the ability to align with strategic changes are typical factors that can add value. This competency relates to development and planning capabilities. Strategic changes are easier if information systems can adapt rapidly and thus support the changes. For example, systems should be able to quickly integrate data from a new acquisition or an external data source. Information sharing (law 1) and efficient usage across organisation (law 2), and detailed data when appropriate apply here.

The added value of big data is case-dependent (Publication I). Therefore, trying to create a generally applicable method that valuates information systems resources is difficult, even meaningless. However, evaluating the competencies with regard to a certain case is both useful and rather easy to perform. Table 2 presents a model inspired by Wade & Hulland (2004) that can be used as a tool. With IS competencies in the rows and RBV resource attributes in the columns as defined by Barney (1991, 1995) the added value of information systems can be evaluated when a concrete case is identified. With a concrete case in mind resources or competencies such as data assets or required algorithms can be identified in more detailed level, which then enables their assessment with the resource attributes.

Table 2. A tool for evaluating IS resources with the RBV resource attributes, modified from Wade & Hulland (2004) and Barney (1995).

IS resource/competency	**V**alue High/med/low	**R**are? Yes/no	**I**mitability Easy/difficult	**O**rganised? Yes/no
IS infrastructure				
IS technical skills				
IS development				
IS operations				
IS business partnerships				
IS planning and change management				
Utilisation of big data assets				
External relationship management				
Market responsiveness				



Information systems research considers both humans and technology as important factors for technology adoption. This research focuses on big data value creation at the enterprise level. In an enterprise humans make the decisions that drive the adoption of big data and related technology. Defining the IS resources and connecting them with the RBV provides a vehicle to evaluate the potential competitive advantage of IS in general and big data in particular. Therefore, information systems research meets the needs of this research as one of the theoretical building blocks.

4 Research Contribution

This chapter concentrates on the results of the research. The following sections discuss key findings of each publication as well as their contribution to the whole.

4.1 Publication I – Current Interpretation of Big Data

By looking at the current definitions of big data Publication I explored how big data is currently understood by scholars and practitioners. Big data is an emerging research area where common terminology is still evolving. Different perspectives on the research area and varying terminology exist, but a common definition for big data does not exist. A (theoretical) definition is a proposal for understanding the meaning of a term. Good definitions improve the quality of communication significantly and enable common understanding between participants from different backgrounds. Moreover, analysing the definition(s) provides insights into the phenomenon.

Numerous definitions of big data can be found in the literature, as well as among practitioners. Technology vendors, the public sector, private companies, consumers, and policy makers, among others, have interests in the field. Therefore, big data definitions were searched broadly in major reference databases (Scopus, ProQuest, and Web of Science) as well as Internet sources. At the end of the search process phase, 62 papers were identified to contain a definition of big data. The year-wise distribution of these papers is presented in Figure 15. Although the first definition was presented more than 10 years ago, the discussion on the definition of big data started only a few years back.

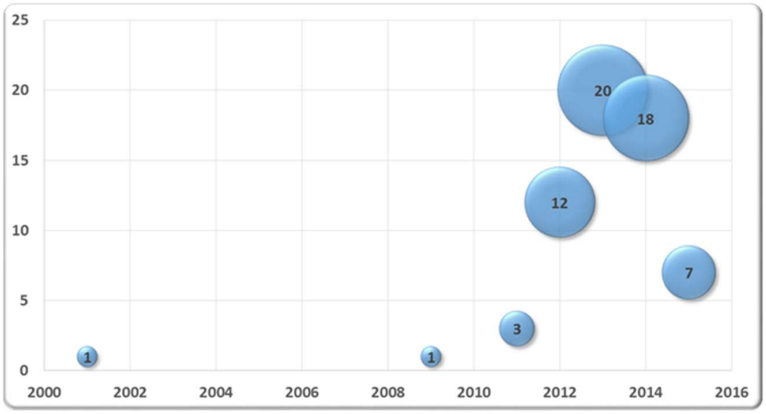


Figure 15. Year-wise distribution of the found papers.

The year 2001 can be considered a major milestone in the definition of big data. In that year Laney (2001) described three essential dimensions of big data: volume, velocity and

variety. These three dimensions are widely used in big data definitions. During the following decade, trailblazers such as Google and Amazon developed practical big data solutions. These solutions have proved to add value to their businesses. In fact, the trailblazers built their business models on big data solutions. An article published in 2008 in the Wired magazine (Anderson 2008) aroused public interest in the use of big data and its effects in science. The next significant milestone was in 2011, when McKinsey Global Institute and IDC published reports (Gantz & Reinsel 2011; Manyika et al. 2011) that drew wide public attention to the potential value of big data. Since then a number of newspaper articles, scientific papers and books on big data have been published.

The included 62 papers were arranged by their publishing date, and each paper was inspected against previously published definitions. If the paper contained a new definition or added some new elements to the existing definitions, it was considered to be a new definition. This analysis resulted in 17 different definitions. These 17 definitions have similarities in the sense that many of them aim to widen the 3V definition to cover technical and especially business aspects. The rest of the papers (45) contained definitions which were essentially covered in earlier papers. The fishbone diagram in Figure 16 gives an overview of the evolution. The bones show the essential additions of all 17 different definitions, i.e. new aspects or components that each definition adds.

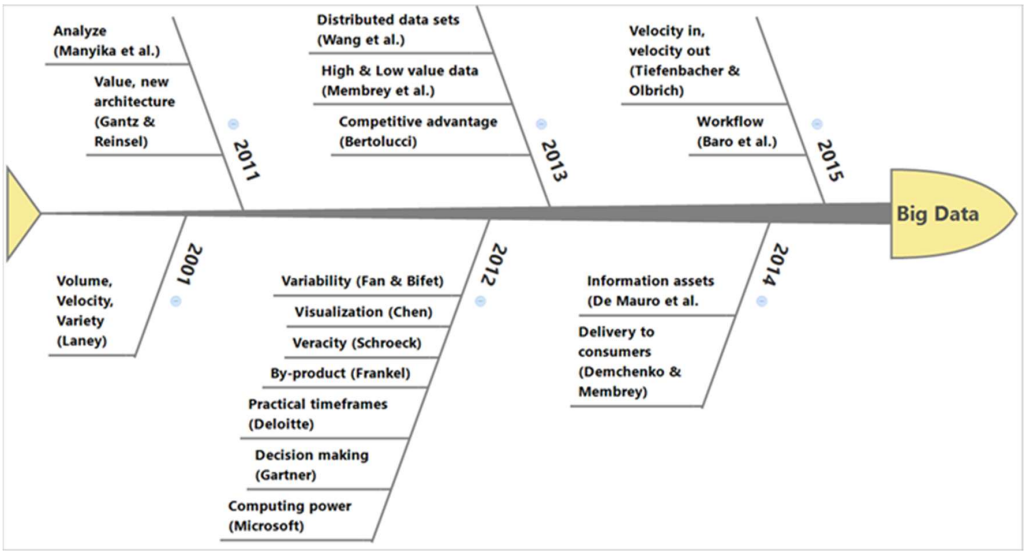


Figure 16. Evolution of the definition of big data.

An analysis of the definitions revealed that several of them have logical incoherencies. Value, for example, must be derived from the data by using analytics, there is no value in plain data as such (Ackoff 1989). Value is also case-dependent. A certain piece of information may be worthless to one company but highly valued by some other firm or in another situation. Value reflects the usage of data. Mixing data related and data usage

related factors leads to vague definitions. Although usage-related factors, such as value, are important, data and its usage should be separated. Another important point to note is that current definitions neglect the disruptive nature of big data. Moreover, scholars discuss various big data related aspects such as privacy, security or policy-making. The discussions implicate that the terminology should be developed further. The key contributions of the study are:

- Although there are various opinions on what big data is, the 3V definition by Laney (2001) contains three dimensions (volume, velocity, variety), which are common to most definitions. In addition to these dimensions, many definitions include technical parts and components related to the intended usage of the data, such as analysis or decision-making.

- Many of the definitions are logically inconsistent, which is one reason for the vagueness of the term big data. A typical flaw is to include both the data and its intended usage in the definition. We suggest that they should be separated. The term big data should cover data-related aspects, whereas a new term *big data insights* should be used when discussing data usage -related activities.

- The current definitions do not consider several important aspects of the big data phenomenon, such as security and privacy, or its disruptive nature. These are not characteristics of big data, but they are important factors of the big data phenomenon that both scholars and practitioners must consider. We suggest that a new definition for big data as a phenomenon should be developed.

Publication I contributes to the whole by setting the stage in terms of basic concepts. It helps to understand the characteristics, or dimensions, of big data, how big data is currently understood by scholars and practitioners, and the limitations of big data definitions.

4.2 Publication II – What Managers Think about Big Data

Management leads the change and sets the pace of digital transformation; therefore, the attitudes and intentions of executives towards big data are important. Publication II investigated the behavioral intentions of Finnish executives with regard to big data. The purpose of the research was to determine the behavioral intentions of business management with regard to big data and explore the factors that explain these intentions.

The instrument used for the data gathering was a survey. The population of the research included executives of large Finnish companies, including companies listed on the Helsinki Stock Exchange, as well as the largest private enterprises. The answers were collected by using a general-purpose online survey tool, Webropol.

Venkatesh et al. (2003) presented a technology acceptance model, which was used as the theory in our research. Their research summarises the findings of the technology

acceptance model (TAM) (Davis et al. 1989) and its several extensions in a "unified theory of acceptance and use of technology (UTAUT)" (Venkatesh et al. 2003). Several information systems studies have applied the UTAUT model, e.g. (Eckhardt et al. 2009; Koivumäki et al. 2008; Tsourela & Roumeliotis 2015; Verhoeven et al. 2010). The model has four principal constructs: performance expectancy, effort expectancy, social influence and facilitating conditions. Based on extant big data literature the following hypotheses were developed.

- H1a: The generic potential of big data has a positive effect on the respondent's behavioural intentions.
- H1b: Company-specific expected benefits of big data have a positive effect on the respondent's behavioural intentions.
- H2: Low perceived complexity in big data utilisation has a positive effect on the respondent's behavioural intentions.
- H3: Social pressure has a positive effect on the respondent's behavioural intentions.
- H4: Perceived technological and organisational capabilities have a positive effect on the respondent's behavioural intentions.

Likert scales (Likert 1932) were used to test the hypotheses. Each of the constructs was tested by using four or more statements. Each statement had five response alternatives: strongly disagree, disagree, neutral, agree, and strongly agree. In the analysis phase the verbal alternatives were replaced by numbers from 1 to 5, accordingly. The statements were developed by using the findings in the current big data literature mentioned above, and propositions stated by Venkatesh et al. (2003).

We received 109 completed questionnaires from 82 companies (45 % of the companies in the survey population). These companies represented 90 billion euros in current turnover (median 301 million euros) and 213,000 employees in 15 different industries. Most of the companies (63) employed more than 250 people. 34 % of the responses came from manufacturing companies, followed by wholesale and retail (12 %), information and communication (12 %), and finance and insurance (10 %). The respondents were members of the management group of their companies (86.8 %), IT managers (7.5 %), or line-of-business executives (5.7 %). Three of the respondents did not expose their role.

A regression analysis of the mean values revealed that three of UTAUT's constructs (performance expectancy, effort expectancy and social influence) had a significant effect on the behavioural intentions of the executives. We did not find the facilitating conditions effect to be statistically significant. Moreover, the analysis did not support the assumption that the generic potential of big data would influence the respondents. Therefore, the data did not support hypotheses H4 and H1a, whereas hypotheses H1b, H2 and H3 were supported.

We performed three separate two-sample t-tests assuming unequal variances to test the following moderators: gender, age and experience (Table 3). The mean age (49.6) was used to divide the respondents into two age groups. Experience in the context of this study was measured by asking whether the respondent had participated in a big data project or not. Moreover, we analysed the means of the constructs by experience in order to find out the moderator effect on individual constructs

Table 3. T-test results – moderator effects on behavioural intentions (INT).

	n	Mean	Variance	p-value
Big data experience	52	4.43	0.257	<0.001
No big data experience	53	3.54	0.587	
Age ≤ 50	53	3.91	0.604	0.335
Age > 50	53	4.05	0.619	
Female	21	4.15	0.444	0.232
Male	87	3.95	0.619	

(H0: no difference in means, 2-sided test)

On the basis of the regression analysis, t-tests and mean analysis described above, we drew the model shown in Figure 17. This statistically valid model explains 48.4 % of the variance in the behavioral intentions of the executives.

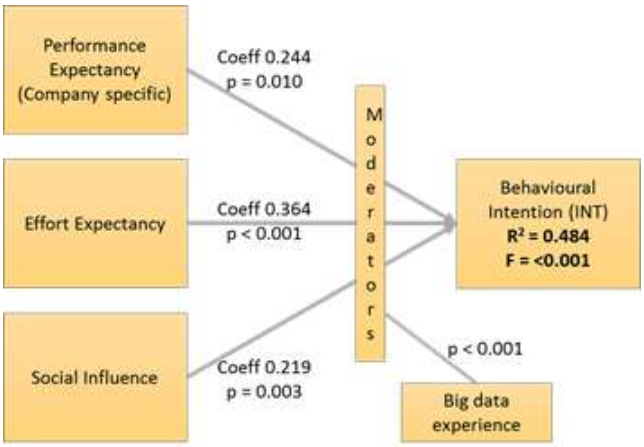


Figure 17. The model explaining the big data intentions of executives.

Figure 18 shows the perception differences of the respondents by experience. Neutral responses have been excluded, i.e. the graph includes those respondents who had clear opinions. While both experienced and non-experienced respondents had positive performance expectations, the non-experienced ones considered the effort required to be high compared to the experienced ones. The overall perception (INT) of big data was very positive among those who had some experience with big data– nearly every respondent promoted big data in their organisation.

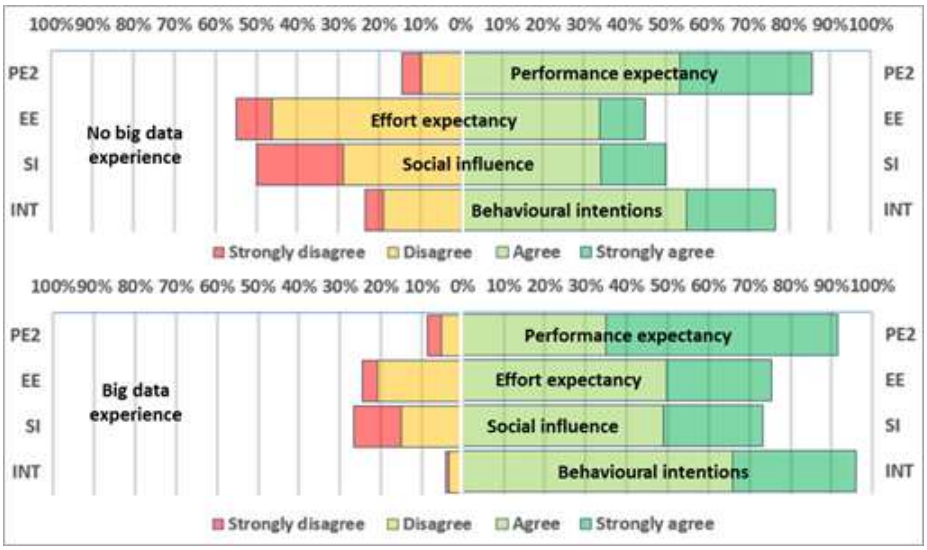


Figure 18. Respondents' perceptions regarding big data by experience.

According to the results, executives have high expectations of big data. Both experienced and non-experienced respondents perceived big data as a vehicle to add value, e.g. to develop more efficient processes, add value to current products or services, and increase customer understanding.

Social influence and IT management seem to play an important role in big data adoption. The respondents who had experience of big data considered management group members and IT management as important social influencers. On the other hand, non-experienced respondents expected big data to be a complex matter, which is of less importance to the management group and IT management. This is an interesting observation, which supports the perception that big data is much more than a technical exercise. However, the IT management seems to play an important role in the differences.

While the attitude towards big data was highly positive in general, the least potential was seen in innovating new products or services, i.e. the respondents did not fully conceive of the disruptive potential of big data in their own business context. This may indicate a small steps approach; executives take cautious, experimental steps towards big data, trying to avoid unnecessary risks. Another possibility is that the companies lack capabilities that are required to identify disruptive innovations. The respondents represented incumbents, who had developed their processes, capabilities and culture over time to perform well in a less data-oriented environment.

Publication II contributes to the whole by adding an organisational aspect to the value creation. Investigating how executives perceive the possibilities of big data helps to understand the scale and speed of the transformation. Publication II also sheds light on

the social aspects that influence big data adoption. These aspects are important to practitioners, as they give indications of where the surrounding world is heading, and help executives identify factors that affect their attitudes.

4.3 Publication III – Lessons Learned from Big Data Experiments

Publication III analysed big data use cases described in the academic literature using computerised content analysis methods. Based on the analysis results, conceptualised themes and guidelines of big data in the context of an organisation were defined. 33 peer-reviewed big data case study articles were identified in major literature databases. Three of these articles were multi-case studies and thus the articles covered 49 big data implementations in total. All the papers were recent, which is not surprising, since most organisations are still taking their first steps with big data.

Table 4. 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 defence; 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 organisations and bodies	-

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 4 lists the number of cases categorised by the ISIC classification of the UN (UnitedNations 2008). ISIC has 21 categories; at least one big data case in 15 (71%) of these categories was identified. As with the application area, also the geographic distribution of the cases was wide, representing five continents. Companies based in North America and Europe represented a majority of the cases with 29 instances.

Content analysis is an established methodology for investigating textual data (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 computerised content analysis have received the attention of scholars recently, e.g. (Han Hu et al. 2014; Lewis et al. 2013; Yu et al. 2014), as researchers wish to utilise 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 computerised content analysis to be a proper method for revealing common big data concepts and lessons learnt in the articles. We visualised the results with KH Coder software using co-occurrence maps. Co-occurrence maps build on the idea that words are related to the concepts they are connected to (Ryan & Bernard 2003). Osgood (1959) was among the first scholars to use co-occurrence matrices to reveal connected concepts in textual data.

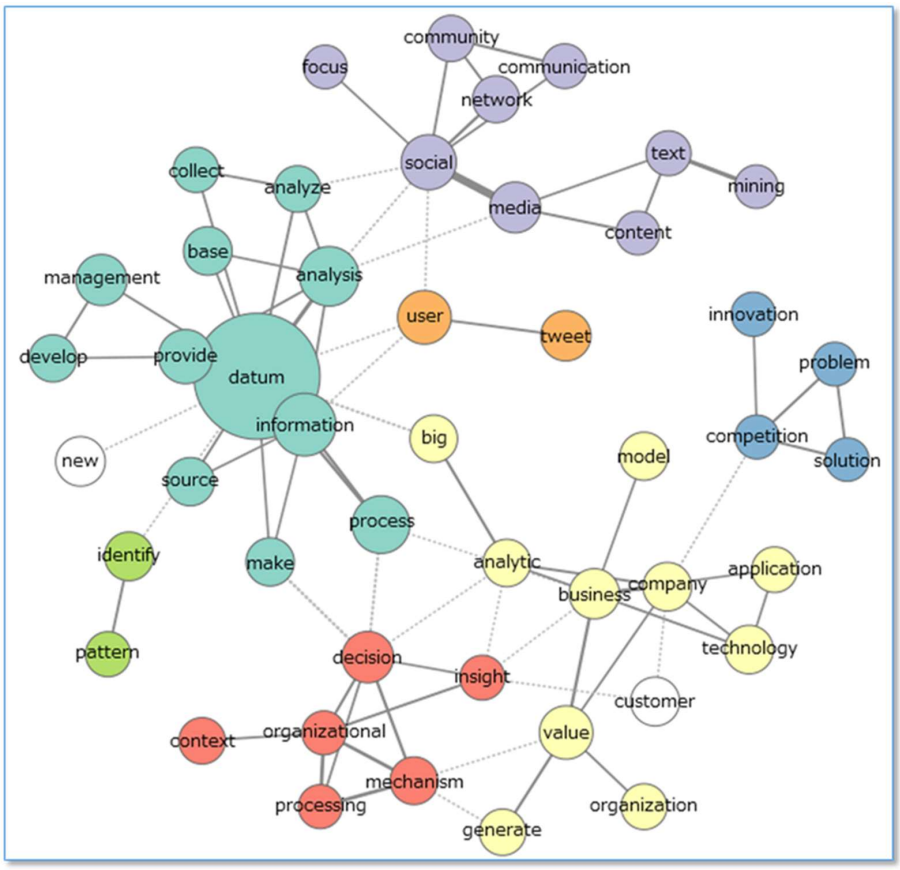


Figure 19. Co-occurrence map of the terms in the big data case study articles.

Figure 19 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 colours distinguish the themes. The three business (or data usage) -related themes are:

- Decision-making (in red on the map). Several studies discussed enhancing the decision-making processes, enabling data-driven decision-making, or providing actionable insights to managers.
- Innovation (blue). Big data was seen as an enabler for data-driven innovation and faster innovation cycles.
- Business value (light yellow). According to the studies, big data is a vehicle to creating new value. The studies recognised positive results and opportunities, such as a business model that was based on big data, energy and cost savings, business transformation, increased revenue and customer satisfaction, better transparency over operations, generating value from the secondary use of data, and deeper understanding of real events. 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) throughout the whole lifecycle of data, from the sources to the analytics, is a central aspect of any big data project. Some studies mentioned that managing the volumes of data is a key challenge. Moreover, the case studies pointed out additional aspects that need to be addressed, such as data inconsistencies and poor data quality. Several studies also reported concerns regarding potential security and/or privacy issues.
- New data sources (purple). In several cases organisations utilised data from outside their own organisation, such as Facebook and Twitter data, blog texts and user reviews, or data collected from mobile apps.

Table 5 synthesises the findings. 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, realising the value is much more a business transformation initiative than a technical issue.

Table 5. Guidelines for big data utilisation.

Theme	Guidelines	Examples
Decision-making	• Embed analytics into decision-making processes. • Be prepared for organisational side-effects.	(Bekmamedova & Shanks 2014), (Cai et al. 2014), (Dutta & Bose 2015), (Phillips-Wren & Hoskisson 2015)
Innovation	• Trust the data. • Search for new methods.	(Amatriain 2013), (Jetzek et al. 2014), (Martinez & Walton 2014), (Ciulla et al. 2012)
Business value	• Look for value in various directions; experiment with the data. • Enable business transformation with the data. • Consider secondary usage of the data.	(Amatriain 2013), (Bettencourt-Silva et al. 2015), (Dutta & Bose 2015), (O'Leary 2013), (Prescott 2014)
Data management	• Expect to face technical and data-related challenges. • Plan for security.	(Dutta & Bose 2015), (Halamka 2014), (Krumeich et al. 2014), (Prinsloo et al. 2015), (Shen & Varvel 2013)
New data sources	• Experiment with new data types. • Consider potential privacy issues.	(He et al. 2013), (Marine-Roig & Clavé 2015), (Yu et al. 2014)

Several studies, e.g. (Davenport 2014; Manyika et al. 2011; Mayer-Schönberger & Cukier 2013) have made claims that big data causes pervasive changes, which will affect almost every sector of life. The cases confirmed the claims, at least partly. Clearly, big data applications are emerging in various walks of life. The studies recognised positive results and opportunities. However, several studies also reported challenges such as data inconsistencies and poor data quality, security and/or privacy issues, missing analytics strategies, lack of leadership, lack of data-driven organisational culture, and the need for new analytics and technology skills. These challenges reflect the disruptive nature of big data. They are indications of the major shifts required; changes that affect not only technical platforms and skills, but they also – and more importantly – influence the organisational culture, decision-making processes and management functions.

Publication III contributes to the whole in three ways. First, it cumulates current knowledge by identifying central the themes discussed in big data case studies. Second,

it provides guidance to practitioners by pointing out factors that early adopters of big data have found significant. Third, it acts as an example of big data technique – text mining – that can add value. These outcomes support the purpose of the whole by increasing understanding and providing tools and methods for big data adoption.

4.4 Publication IV – Driving Value with Innovations

Publication IV presents a framework that explains the role of innovation capabilities as a mediator between big data and the business model of a firm. The framework helps in understanding how big data and innovations are shaping business models in the digital transformation. Digitalisation implies business changes that are technology driven. Thus, the article takes a multi-disciplinary approach to the topic, combining strategic management and information systems research. Both managerial actions and technology are required to drive innovations that may add value to the business.

The study followed the design science research method (DSRM) framework (Peffers et al. 2007), which provides a nominal process for conducting and evaluating design science research in information systems. Figure 20 shows how the DSRM framework was applied in the study.

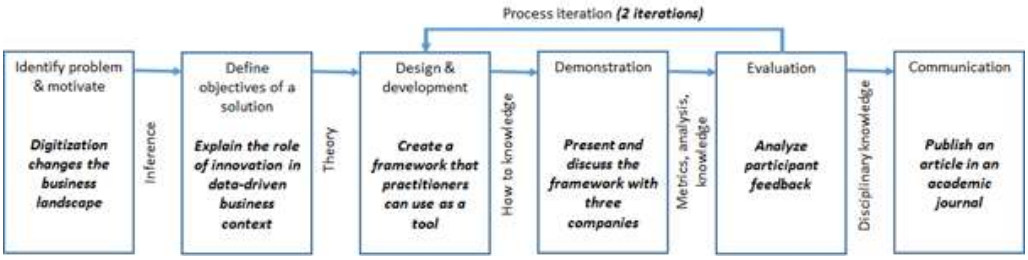


Figure 20. Research method of Publication IV, applied from Peffers et al. (2007).

The framework, shown in Figure 21, was developed based on the existing literature. The key components of the framework are the business canvas model by Furr & Dyer (2014) and the 3V definition of big data (Laney 2001), supplemented with a categorisation element for each of the dimensions. These components add the viewpoint from strategic management and information systems research. This provides a sufficiently detailed starting point for discussions from different angles, while still preserving the big picture.

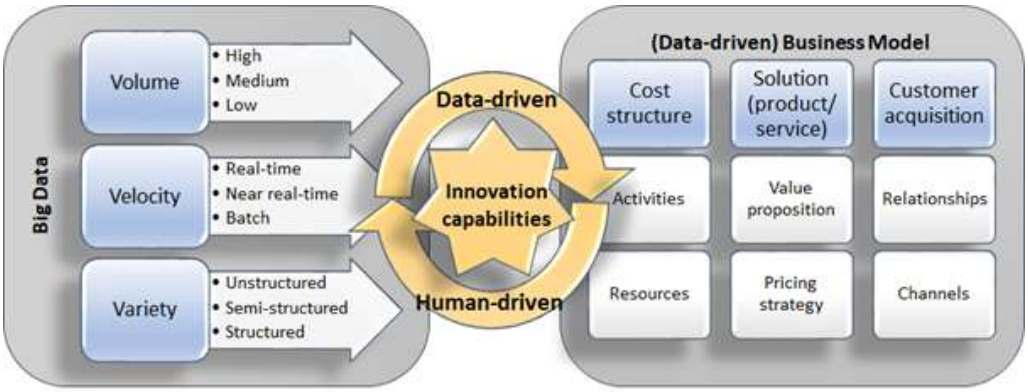


Figure 21. Innovation as a mediator framework.

The third, and crucial, component of the framework adds innovations. The data deluge and changing business models challenge incumbents to develop new capabilities. Approaches using resource-based theory have been popular among big data researchers recently, (Braganza et al. 2017; Gupta & George 2016; Mikalef et al. 2018). The innovation as a mediating framework adds the role of innovation capabilities in shaping the business model to the conversation. Moreover, the framework distinguishes between data-driven and human-driven innovations. The *human-driven approach to innovation* states that innovations stem from people (Harmaakorpi & Melkas 2012). New ideas – the seeds of innovations – often emerge sub-consciously when our brain associates things with one another. Another way to view innovations relies on data and automation. A novel approach, called *innovation automation*, is data-driven. Data-driven innovation suggests that the innovation processes could, and should, be automated (Shaughnessy 2015). This approach puts technology and (big) data at the core of the innovation processes.

The purpose of the demonstration and evaluation phase was to assess the applicability of the framework in practical situations. Another objective was to gather feedback in order to develop the framework further. First, the framework was presented to two big data intensive firms. The aim was to establish the usability of the framework in firms that are in the quite early stages of exploiting big data. In the second iteration, a "post-mortem" examination of a real-world big data case was performed. Reflecting the case against the framework provided insights into how the framework might have helped the company to understand the impact of big data.

Companies can choose different practical approaches to the framework according their situation and objectives. Increasing the role of data in current business model, e.g. by automating delivery processes to reduce transaction costs requires a different approach than developing a new, data-driven model. Based on the evaluation phase of our framework, we synthesised an exemplary usage scenario as follows.

1. Assess the current situation and big data impact. This can be done by looking at the business model components (see Figure 21) one by one and creating scenarios of potential big data effects. The effects can be either risks or opportunities; some of the effects are such that they cannot be influenced, some of them are within the reach of the company.

2. Define business objectives. Clear goals and ownership are best practices for any business development initiative. This phase might be iterative, innovating back and forth between the business model and big data. What internal or external data could affect the business? What data do we need in order to best run our business? Novel ideas or long-term, strategic goals often require experimentation for verification.

3. Evaluate/ideate big data value potential. Many innovations are human-driven, but increasingly, data is a source of innovation. Humans ideate things that require gathering and combining new and existing data in novel ways. Accordingly, experimenting with data may reveal insights that spark ideas. Human-driven and data-driven innovations are not mutually exclusive. Instead, they can, and should, support each other, effectively creating an innovation loop. Our framework provides support for these activities by explaining the theoretical background and enablers for an effective innovation process.

4. Identify required datasets and capabilities. Identifying data that is beneficial to the business is a two-way operation. As one may intuitively think when looking at our framework (Figure 21), human-driven innovation comes down to a question: "What data do we need to fulfil this business need?" On the other hand, data-driven innovations ask: "Do we have data, or can we access data that is beneficial to our business?" Once the datasets are identified, the modified 3V model shown in Figure 21 can be used to classify and categorise the data at a more detailed level. This, in turn, reveals possible gaps in ICT capabilities.

Publication IV contributes to the whole both from theoretical and practitioner viewpoints. The theoretical viewpoint connects strategic management, innovations and information systems research into a framework that offers a logical way to organise perspectives on technological change. Most of the current big data studies so far approach the subject from the viewpoint of single discipline. As the phenomenon is ubiquitous, multi-disciplinary approaches may fertilise new thinking. With regard to the research questions, Publication IV explains the role of innovation in the big data value creation process. It helps to understand the role of human and data-driven innovations, and innovation capabilities in an organisational context. For practitioners, Publication IV offers a framework that can be used to develop a systematic approach towards the development of big data capabilities. Moreover, understanding the theoretical background of the innovation process in the big data context will help practitioners to focus on developing the capabilities and methods that best support the transformation towards data driven business models.

4.5 Publication V – Driving Value with a Predictive Algorithm

Publication V first presented a case study, where a predictive algorithm is used to qualify bids. Following this, classification matrices were used to estimate the business value of the algorithm, followed by practical guidelines.

The subject of the case study was a large information technology service provider, offering systems integration, consulting, and outsourcing services. The number of bids, i.e. projects and services in the pipeline, can consist of several thousand at any given point in time. Due to the heterogeneous offerings, and the size of the pipeline, identifying winning bids as early as possible is a key challenge for the sales force management and for sales efficiency.

Data for the predictions was gathered from several sources. The history from the past four years, including: customer register data, customer satisfaction surveys, and client meeting records were combined into a harmonised data set with almost 80,000 records. From this dataset, 22 variables were constructed and used for the predictions. The predictive model was based on decision trees and built using the R language.

The accuracy of the algorithm was verified every month by comparing predicted outcomes with actual outcomes from the previous four months. Based on the results, the model may be re-trained. Figure 22 illustrates the principle. Field evidence shows that the model's accuracy is 65–67%, when we look at the pipeline as a whole. However, the accuracy in the early phases of sales, i.e. in the bid qualification and planning stages, is significantly higher, and was 78–86%.

Figure 22. Algorithm follow-up principle.

The primary goal of setting the problem was to identify winning opportunities at early stages. This helps to avoid inefficient sales force allocation, additional costs, and inaccurate sales forecasts. Based on the field evidence, the model gives a correct prediction roughly four times out of five, i.e. it produces the desired results. A secondary goal was to evaluate the bids according to the same criteria. It was self-evident that if the primary goal was met, also the secondary goal would be met. The algorithm treats every

case according to same rules and variables. Thus, the model effectively creates a decision-making baseline for sales management and teams.

Classification matrices and management accounting principles were then used to demonstrate the value in a way that would be easier to digest for executives than technical details. Due to business restrictions, only the building blocks of the business case instead of the actual business case could be presented. Classification matrices combined with usual management accounting practices provide the tools for business case calculation. As an example, a five-step process for calculating savings of predicted no-win bids was presented.

Guidelines based on the approach taken were then proposed. The suggestions can be summarised as follows.

- Be bold in bid elimination. When the accuracy of the algorithm has been verified, trust the results.
- Defer the implementation of a production-ready solution until the algorithm has been proven to produce results.
- Play with the data. Understanding the data inside and out is essential for defining the prediction variables, interpreting the results, and identifying potential biases.
- Create an initial benchmark for accuracy, by comparing predictions to actual results with classification matrices.

Moreover, the article suggested developing key metrics – *the prediction accuracy percentage, the potential for salesforce re-allocation, the predicted hit rate, and the percentage of lost opportunities* – for evaluating the performance of the algorithm, and its business effects. The metrics are indicators of trends and act as tools for continuous improvement instead of being exact accounting figures.

Publication V contributes to the whole by examining value creation in practice, especially from the viewpoint of justifying the algorithm in business terms. Business executives typically prioritise investments based on added value. Being able to concretise the value of an algorithm may be a decisive factor between a go or no-go decision. Moreover, the design and development of a machine learning algorithm, gathering data from several source systems, and building a concept how to "productise" the solution offered the researcher an opportunity to reflect theory and practice.

4.6 Publication VI – A Recipe for Your Big Data Cook Book

Publication VI develops a process theory-based model of big data value creation in a business context from the viewpoint of a single firm. The model provides a "recipe" for converting big data investments into firm performance. The recipe helps practitioners to

understand the ingredients and complexities of the value creation processes, and it explains how big data investments translate into economic performance, and why the conversion sometimes fails. This helps to focus on success factors which promote positive performance.

Process theory based approaches are rare in current big data research. However, for most organisations big data adoption equals business transformation, which effectively means uncertainty. Process theories tend to perform well in situations where the outcome is uncertain (Markus & Robey 1988). As Publication VI attempts to build a holistic end-to-end view of big data value generation, a setting which is obviously influenced by a huge number of internal and external factors, process theory is a good candidate for research methodology.

As a theoretical context Publication VI builds applies the IT value creation model by Soh & Markus (1995) and two value creation theories: the data, information, knowledge, wisdom (DIKW) hierarchy (Ackoff 1989; Zeleny 1987) and the virtual value creation (VVC) framework (Rayport & Sviokla 1995). The IT value creation model provides a starting point for the processes and the value creation models explain how the value cumulates. Current big data literature relevant to the study context is categorised topically (see Table 6). Moreover, the contingency factors that may either promote or demote the value creation are identified from the literature (+/- column).

Table 6. Topical categorisation of big data literature related to big data value creation.

Topic	Sub-topics	Focal areas	+/- factors
Data and technology	• **Big data**, e.g. (Gandomi & Haider 2015; Kitchin & McArdle 2016; Laney 2001; Pigni et al. 2016; Ylijoki & Porras 2016). • **Technology**, e.g. (Boncea et al. 2017; Dutta & Bose 2015; Krumeich et al. 2014)	• Characteristics of big data, where it comes from, or how it can be used to add value. • Technical processing of big data, e.g. data management frameworks and analytical tools.	• Data & technology management, e.g. (Alguliyev et al. 2017; Alharthi et al. 2017; Boncea et al. 2017; Piccoli & Pigni 2013; Tiwana 2014)
Capabilities	• **Analytics**, e.g. (Akter & Fosso Wamba 2016; Arora & Malik 2015; Najjar & Kettinger 2013; Fosso Wamba et al. 2017) • **Innovation**, e.g. (Brynjolfsson & McAfee 2012; Gobble	• Technical and human related skills. • Organisational capabilities.	• Data quality, e.g. (Ardagna et al. 2016; Chae et al. 2014; Hazen et al. 2017; Janssen et al. 2017; Merino et al. 2016; Vidgen et al. 2017) • Resource availability, e.g. (Davenport 2014;

Topic	Sub-topics	Focal areas	+/- factors
	2013; Hartono & Sheng 2016; Iddris 2016; Jetzek et al. 2014; Zhan et al. 2017) • **Information management**, e.g. (Dutta & Bose 2015; Mithas et al. 2011; Tallon et al. 2013)		Janssen et al. 2017; Shah et al. 2012).
Big data impacts	• **Decision-making**, e.g. (Janssen et al. 2017; Manyika et al. 2011; Sharma et al. 2014) • **Operational efficiency**, e.g. (Bärenfänger et al. 2014; Dutta & Bose 2015; Roden et al. 2017) • **Product/service innovation**, e.g. (Davenport 2014; Gobble 2013; Manyika et al. 2011; Mayer-Schönberger & Cukier 2013) • **Business model innovation**, e.g. (Chen et al. 2011; Iivari et al. 2016; Van't Spijker 2014)	• Organisational change management. • Management perceptions. • Industry/ecosystem aspects.	• Data maturity, e.g. (Anand et al. 2016; Comuzzi et al. 2016; Dutta & Bose 2015) • Organisation culture, e.g. (Anand et al. 2016; Sandberg & Aarikka-Stenroos 2014; Shah et al. 2012) • Competition, e.g. (Huberty 2015; Pousttchi & Hufenbach 2014; Weill & Woerner 2015) • Privacy/security factors, e.g. (Clarke 2016; Newell & Marabelli 2015; Sullivan 2014) • Regulation, e.g. (Keen et al. 2013; Truyens & Van Eecke 2014)

Next, the literature is directed towards the IT value creation process. Organisations must make investments in order to achieve big data assets (**asset creation process**). We adopt an idea from Soh & Markus (1995), who presented the IT conversion process for turning IT expenditures into IT assets. Expenditures in this context are economic investments and spending that aims to, or supports, big data asset creation.

In order to make use of the information, firms must develop resources and capabilities related to the assets (**capability creation process**). In the resource-based view of the firm (Wernerfelt 1984) resources and capabilities explain the firm's competitive advantage. As an example, a firm must have a hardware environment suitable for big data processing (tangible resources), develop analytical algorithms (intangible resources), and develop a data-oriented organisation culture (organisational capability).

A company uses its capabilities and knowledge in a **transformation process** to produce valued outcomes. Teece (2007) introduced the concept of dynamic capabilities, which explains how companies renew their competencies when adapting to business turbulence. Big data capabilities typically produce intermediate effects. For example, Tallon et al. (2013) state that the effects of data governance are linked to industry-specific, intermediate results instead of firm-level effects such as profitability. Moreover, aspects like the organisational context and managerial actions play a crucial role in the value creation process (Müller & Jensen 2017).

Finally, we must address the fact that big data impacts are intermediate results, i.e. we need to link the impacts to actual performance metrics. The performance of a firm depends on the **competition process**, which connects the firm to its industry, ecosystem, competitors and customers.

Figure 23 presents the model, consisting of four sequential processes. Big data assets connect the asset creation process to the capability creation process. Big data capabilities connect the capability creation process to the transformation process, which in turn links to the competition process via the impacts of big data. Using process theory terminology, the inputs and outputs are necessary but not sufficient conditions with regard to big data value creation.

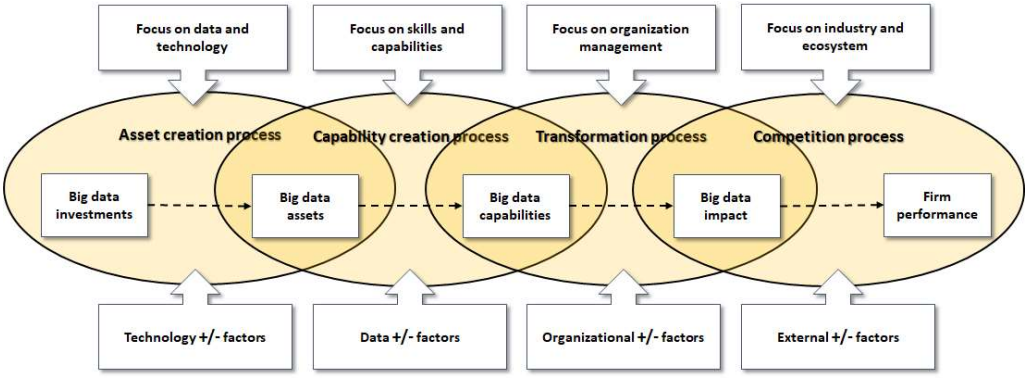


Figure 23. AC/TC process – creating economic value with big data.

Each of the processes may be affected by conditions that either favour or impede the process, such as data quality (Janssen et al. 2017; Vidgen et al. 2017), immature

technologies leading to problems (Boncea et al. 2017), organisational maturity regarding big data (Comuzzi et al. 2016), and organisational silos (Bärenfänger et al. 2014; Sharma et al. 2014). In order to be successful, these factors must be favourable. To mitigate the risks related to technology, data, organisational and external conditions, organisations must apply procedures that support the processes, as highlighted by the focus areas shown in Figure 23.

Publication VI contributes to the whole by offering a holistic, end-to-end process model from investment to performance. This addresses the research sub-question concerning the big data value creation processes and mechanisms because the model explains *how* big data investments are transformed into improved economic performance, and it helps to understand *why* investments sometimes fail. Moreover, the model bridges current big data theory and practice. It is comprehensive and based on solid theoretical foundations, yet it is an appropriate framework also for practical situations. For example, it can be used to identify success factors that organisations should focus on in order to mitigate negative impacts.

4.7 Summary

The disruption of many current business models is already in progress (Weill & Woerner 2015). New digital technologies produce vast amounts of various types of data (Gantz & Reinsel 2011), referred to as big data. However, the data-driven approach is still a new paradigm for most organisations (Shen & Varvel 2013). While the approach represents a significant opportunity for enterprises, e.g. (Davenport 2014; Manyika et al. 2011; Mayer-Schönberger & Cukier 2013), according to SCOPUS, only 5.2 % of papers that refer the term “big data” represent business and management categories. This research approaches the domain not only from the strategic management perspective, but also from innovation research and information systems perspectives.

Making use of big data requires companies to develop new organisational and managerial capabilities, e.g. (Akter et al. 2016; Anand et al. 2016; Arora & Malik 2015; Comuzzi et al. 2016; Mithas et al. 2011) and integrate analytics into core business and decision-making processes (Bekmamedova & Shanks 2014; Davenport et al. 2012; Isik et al. 2011). Thus, big data represents a paradigm shift from the IT side towards the business, which justifies a multi-disciplinary, holistic approach instead of e.g. a technical view.

Together the articles, presented in the Publications section, explore the big data landscape from several angles. The two main perspectives are theoretical and practical. Bridging theory and practice is necessary relevant in understanding the transformation process. Publications I, III and VI cumulate current big data theory by:

- explaining basic concepts and their issues,
- concentrating current big data experiences into guidelines, and
- explaining the value conversion process from data to firm performance.

Correspondingly, Publications II, IV and V are more practice-oriented. They:

- explore current management attitudes toward big data,
- present a framework for big data driven innovation management, and
- offer tools and best practices that help practitioners to avoid pitfalls.

The results of this research answer the primary research question (RQ) – understanding and utilising big data in the transformation process towards big data driven business – by explaining the phenomenon as well as the value creation processes as described in Table 7, and by connecting theoretical aspects to practice as shown in Figure 24. Section 5.1 which describes the theoretical and practical implications discuss the results in more detail.

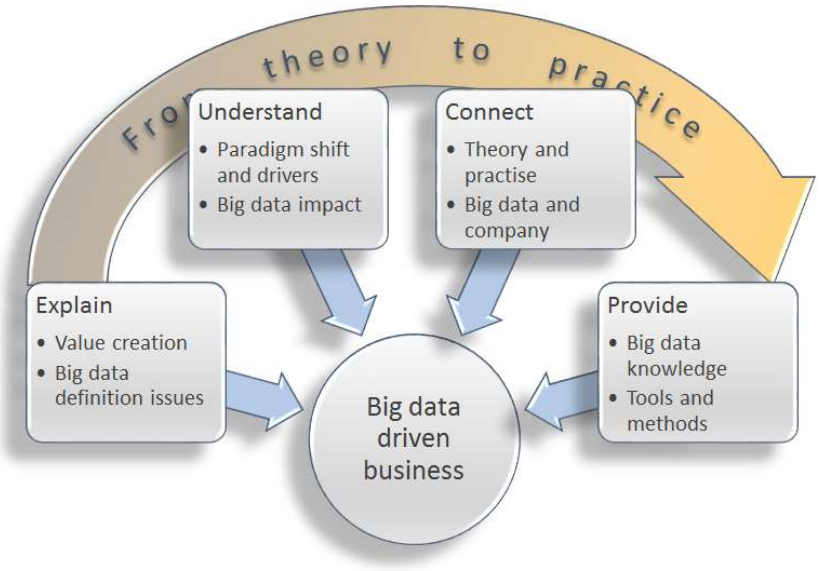


Figure 24. Research contribution from theory to practice viewpoint.

Table 7 presents a condensed view of the Publications and their contribution as a whole. Column "Motivation and results" describes why the Publication is important for the whole of this research and presents the key result of the Publication. Moreover, column "RQ's" links the Publications to the sub research questions (see section 1.3.3 "Research Questions") of this research.

Table 7. Research contribution as a whole in a nutshell.

Publication	Publication type	Motivation and results	RQ's
Publication I	A systematic mapping study.	Big data is an emerging research area where common terminology is still evolving. Different perspectives to the research area and terminology exist, but a common definition for big data does not exist. The paper reviews the current status in big data definitions, discusses the shortcomings of the current definitions, and proposes possible solutions for the shortcomings.	SRQ1
Publication II	A survey concerning the behavioural intentions of Finnish executives with regard to big data.	Management leads the change and sets the pace; therefore, the attitudes and intentions of executives towards big data are important in the transformation process. In addition to the generally positive attitude towards big data, the results reveal statistically significant differences between respondents with big data experience compared to the inexperienced ones. The role of IT management seems to play an important role in the differences.	SRQ2
Publication III	Computerized content analysis of big data use cases described in the academic literature.	Exploring the "big data jungle" is a new area for both scholars and practitioners, and the experiences of early adopters are valuable in order to develop best practices and avoid pitfalls. Publication develops conceptualized themes and guidelines of big data in the context of an organization based on the analysis. Moreover, realized benefits as well as issues regarding technology, skills, organizational culture and decision-making processes are revealed.	SRQ1, SRQ2
Publication IV	A multi-case study.	Linking the opportunities of big data and the business transformation imperative resulting from digitalization leads to a situation where incumbent firms must re-think and innovate in their business models, and create new capabilities in order to stay competitive in their business ecosystem.	SRQ1, SRQ2

Publication	Publication type	Motivation and results	RQ's
		A conceptual, practitioner-oriented framework that explains the role of innovation capabilities as a mediator between big data and business model, and points out the role of human and data-driven innovation capabilities in the big data value creation process. The framework was field tested with three companies and refined based on their feedback.	
Publication V	Application of classification matrices for algorithm impact assessment.	Predictive algorithms are increasingly used to support decision-making. Understanding the costs and benefits of a predictive model is an important aspect for businesses. However, algorithms are abstract, and their impact oftentimes remains vague. The results show, how to apply classification matrices for business value assessment of an algorithm. Moreover, the paper proposes guidelines and metrics for interpreting the impact in practical solutions. The results are based on a machine learning algorithm written and field tested in a real-world case by the author.	SRQ2, SRQ1
Publication VI	A process theory-based model of big data value creation.	Organizations are complex structures and value creation involves various stakeholders, functions and processes. This brings in uncertainty. It is not guaranteed that a project will deliver its desired outcomes. Some of big data initiatives will inevitably fail, whereas others may lead to sustained competitive advantage. The result is a holistic, process theory-based model, which consist of four probabilistic processes and provides a "recipe" for converting big data investments into firm performance. Moreover, it helps to understand, why the conversion sometimes fails.	SRQ2, SRQ1

Moreover, the purpose statement of this research promised to provide tools that help enterprises in their transformation towards data-driven business. These tools were

presented in various contexts in the articles. Therefore, to give an overview to the reader, Table 8 presents a condensed summary of the tools. For details, refer to articles. The level of abstraction of the tools is quite high, thus their usage in practical situations needs facilitating.

Table 8. Tools for big data adoption.

Tool type	Description	Reference
Process model	An end-to-end process model for big data value creation.	Publication VI, Fig. 4
Framework	A model to organize perspectives around data- and human-driven innovation.	Publication IV, Fig. 2 and Section 4.2.2
Themes & guidelines	Key topics to be focused and addressed in big data implementations.	Publication III, Table II
Method & guidelines	Generic method and guidelines for algorithm impact valuation.	Publication V
Concept	Open-source based content analysis of unstructured data	Publication III

In addition to the tools listed in Table 8, the whole dissertation can be seen as a high-level framework for big data adoption.

5 Discussion

This chapter starts with a reflection of scientific and practical implications. It shows how this study links to existing research and discusses who could benefit from the results of this research and how. Following the theoretical and practical implications, suggestions for further research are proposed. Indeed, there are many potential new research avenues, as the subject is a relatively new research topic. Finally, the chapter ends with a discussion about reliability and validity aspects of the research.

5.1 Theoretical and Practical Implications

Big data in business contexts is an emerging area of research. Most of the research so far focuses on technical aspects (see Figure 5), such as advanced algorithms or managing unstructured data. There are many big data related management aspects and organisational areas that would benefit from sound theoretical foundations. However, in these areas theory and practise go hand in hand. Several existing theories could be applied in the big data context and, as the practices develop, new theoretical advances could be made that better take into account the impact of big data on the managerial and organisational context. In the early stages of big data research and its application, numerous topics require attention. This research sheds light on some theoretical aspects, points out a number of new research directions, and acts as a high-level framework for big data adoption. The following paragraphs also explain Figure 24 (presented in the previous Section) in detail.

This research **explains**, how big data value creation mechanisms work. Previous research has created several models that can be applied to big data as well. The DIKW model (Ackoff 1989) serves as a useful abstraction of the value creation process. The VVC framework (Rayport & Sviokla 1995) conceptualises the steps that are required in the value creation process. Building on these foundations, this research applies the laws of information (Moody & Walsh 1999) to big data in order to explain the value creation primitives. Moreover, several partly overlapping explanations of the value adding mechanisms are presented as they may potentially add significant value in certain situations. These concepts are presented in the section "From Data to Actionable Insights". Publication VI explains the value creation mechanisms and related factors using a process theory view.

This research also explains, why one of the key challenges in current big data research is the definition of big data. This issue is addressed in Publication I. Various definitions, e.g. (Frankel 2012; Gantz & Reinsel 2011; Laney 2001), are not problematic as such. Different viewpoints may require different definitions. The problem with the definitions of big data is that many of them are inconsistent. A typical flaw is to combine case-dependent elements such as value or veracity in the definition. This immediately raises questions like "value for whom?" or "veracity for what purpose?" which cannot be answered by just looking at the data. This research contributes by recognising the

problems in current terminology and providing definitions that address the most common pitfalls (see the section "Defining Big Data"). The definitions proposed separate data and its usage. Using two different definitions for different purposes brings two benefits: logical problems can be avoided, and the definitions are clear, easier to understand. Moreover, a definition of the broader phenomenon is proposed. Many important aspects such as security and privacy, e.g. (Altshuler 2011; Berghel 2013; Eckhoff & Sommer 2014; Gehrke 2012; Stopczynski et al. 2014), must be taken into account when dealing with big data. In addition, the disruptive effects of big data shape the structures and processes in various industries. The big data phenomenon shifts the paradigm towards data-driven businesses. This change is reflected in the definition of big data phenomenon (section "Defining Big Data"). This research proposes definitions to some of the key terms but there is a lot more to do.

This research attempts to **understand** the disruptive nature of big data. Explaining the value creation mechanisms and proposing clarifications to the terminology increases our understanding of the phenomenon. In order to properly understand the impact of big data, this research also adopts a wider perspective. The forces behind the paradigm shift – connectivity, the Internet of Things and mobility (Van't Spijker 2014) – drive the change towards more data-oriented business environment. New emerging ecosystems, innovative products, services and business models require enterprises to develop new capabilities. This applies to start-ups as well as incumbents. Start-ups must take advantage of potential opportunities and be agile, pivoting in new directions if their original offering or business model fails.

In the spirit of Christensen (2013), this research pinpoints the fact that incumbents also need to focus on disruptive technologies, not just attempting to be more and more efficient. They must re-invent their innovation capabilities. Big data in general and data-driven innovation in particular are potential sources of new value (Publication IV). The impact of big data is pervasive and has implications for most companies, business ecosystems and industries, e.g. (Manyika et al. 2011; Weill & Woerner 2015). Moreover, the impact on the individual and society level will be significant. Each of the articles as well as chapter "Big Data" underline and justify the disruptiveness. Understanding the phenomenon is required for successful application of big data in the business context and several new research avenues are available for scholars with regard to understanding the impact of big data from different perspectives.

Connecting theory and practice plays an important role in this research. Bridging the gap helps both scholars and practitioners. In a new research area, such as big data, practitioners often lead the pack. Research benefits from attempts that concentrate or generalise current practices. Publication III provides concentrated insights from actual case studies. These generalisations offer building blocks for. more theoretical frameworks that reach the core concepts and develop a deep understanding of the phenomenon. Furthermore, practitioners can use the results as guidelines that help to avoid typical pitfalls in big data adoption. Another example of how this research bridges theory and practice is the discussion of the definition of big data. This points out the discrepancies

in the currently used definitions, offering possibilities for new studies. In addition, the discussion aids practitioners in understanding what big data is and what it is not. Software and hardware vendors have different viewpoints and vague terminology does not make things easier for practitioners.

Another connection that this research investigates is the linkage between big data and the company context. The resource-based view of the firm (Wernerfelt 1984) states that the resources and competencies conform the competitive edge of the firm, while Wade and Hulland (2004) define a frame that can be used to measure the strategic value of information systems resources. Combining these findings with resource attributes defined by Barney (1995), this research proposes a model (see Table 2) that can be used to evaluate the value of information system resources and big data in a company and case context. The model acts as a concrete and practical connection between theory and practice. Moreover, it supports one of the key claims of this research – that the value of big data is case-dependent.

This research **provides** new knowledge about big data phenomenon. This knowledge is acquired by applying existing theories and disciplines to the emerging phenomenon. It shows that the basic or primitive mechanisms behind the big data value creation processes are the same as with any data. The DIKW hierarchy (Ackoff 1989) and the virtual value chain (Rayport & Sviokla 1995) apply. However, there are significant differences when we look at the broader scope. Big data is disruptive by nature. Many traditional industries are already facing turbulent times as new entrants invade the market using new, data-driven business models. New ecosystems are emerging. The magnitude of the phenomenon is pervasive. The pace of the change is rapid, as Publication II – and the daily news – indicate. This research investigates the phenomenon from several angles, providing knowledge that helps to put it into a context.

In addition to new knowledge, this research provides tools and methods for practitioners in order to ease their efforts in big data adoption. As a whole, it offers a coherent model that explains the big data phenomenon and helps to understand the implications in the business context. This research also shows, how relevant existing theoretical foundations such as Venkatraman's (1994) IT-enabled business transformation framework or Barney's (Barney 1995) VRIO-model can be used in conjunction with big data. Other tools and methods, such as the business model canvas (Osterwalder & Pigneur 2010), the laws of information (Moody & Walsh 1999), themes and guidelines (Publication III) or the tool for evaluating IS resources (Table 2) have been modified or created to meet the specifics of big data. In practical applications many of the tools and methods need to be adapted to current situation in more detailed level. Yet, the researcher believes that the presented toolbox can add value to the process.

As a summary, the results contribute to current big data research by conceptualising existing practices and knowledge in a way that can be used as a high-level framework for big data exploitation in a business context. The theoretical foundations of this dissertation combine business, innovations and big data in a way that helps in understanding the

transformation process. From the practice point of view, experiences from the field show that early adopters of big data report both positive experiences and technical challenges (Publication III). Moreover, the pace of the change is fast (Publication II) and the disruption of many industries is already ongoing (Weill & Woerner 2015). The paradigm is changing; we are heading towards data-driven businesses. So far only a limited set of theoretical models and best practices exist. There are many areas for further research.

5.2 Suggested Further Research

The terminology and taxonomy around key concepts of and related to big data should be investigated deeply. The number of related parties increases as the adoption of big data proceeds. It is common that people from different backgrounds have difficulties in understanding each other. Without definitions, abstract terms such as a *digital trail* or *machine learning* may have as many interpretations as there are participants in a conversation. Well-defined terminology enables parties to find common understanding and improves the quality of communication. Further research might concentrate on developing a coherent model of the terms in the big data domain. In addition to providing a detailed definition and the meaning of the terms, the connections between them as well as their relationships to related concepts should be covered.

There is an urgent need for research that involves the effects of big data on decision-making and organisational side-effects. Publication II showed that the adoption of big data is ongoing at a fast pace. Several papers, e.g. (Dutta & Bose 2015; Phillips-Wren & Hoskisson 2015; Shen & Varvel 2013) indicate challenges in managing the transformation. For many individuals change is a threat, and this leads to change resistance. Change resistance is familiar to most executives and they recognise it well. However, this may not be the case with the big data implications in decision-making processes and their effects on the organisational culture. Decision-making processes will change as the utilisation of big data increases. Some (Bettencourt-Silva et al. 2015; Isik et al. 2011) suggest that decision-making support systems should be tightly integrated with operative enterprise systems. The reasoning behind this makes sense. Specialists and customer-facing personnel make a large number of smaller, or operative, decisions. They would benefit from more accurate, detailed or timely data, which would lead to more informed decisions. This kind of development would reduce the role of middle management with regard to decision-making. Moreover, algorithms might take care of routine decisions and suggest alternatives to humans. These changes may sneak in and the resulting effects may take managers by surprise. Mayer-Schönberger and Cukier (2013) discuss automated decision-making, wondering whether this development could lead to “data dictatorship”. This raises questions as to whether there are limits to automated decision-making or what data and rules the decisions are based on. A better understanding of changing decision-making processes and the effects of automated decision-making would help to manage the transformation process.

A significant potential value of big data is transparency (Manyika et al. 2011), which stems from the laws of information (Moody & Walsh 1999). Increased transparency, however, leads to organisational side-effects that must be considered. For example, subsidiaries typically resist exposing their internal affairs to the headquarters. This is a situation that the researcher has experienced several times in practice. The claim is that they know local circumstances better. This may be the case sometimes, but often the claim is just an excuse. It is obvious that local personnel see increased transparency as a threat. If the headquarters has the same information as the subsidiary, they might question local decisions. Indeed, one scenario is that decision-making becomes more centralised, possibly leading to a harder, number-based culture that ignores human aspects. The pros and cons of transparency and its consequences to the organisational culture would be worth studying.

Big data adoption is related to business models. A few papers focus on data-related business models, e.g. (Chen et al. 2011; Van't Spijker 2014). Iansiti and Levien (2004) state that stand-alone business models do not work in turbulent environments. Networked or ecosystem-based models can be more useful. In practice, new big data related ecosystems are emerging. However, papers discussing ecosystem business models in general and big data related ecosystem business models in particular are still rare. This brings up new research avenues. Some examples of subjects for further studies are presented below.

Changing business ecosystems link directly to the capabilities of the company. Enterprises must create new capabilities in order to adapt to changing environments. Moreover, they might have more ambitious goals, such as gaining competitive advantages by developing big data ecosystem related capabilities. The section "Enterprise and Ecosystem Level Impacts" in chapter "Big Data" presents some aspects of relevance to business model and capability related topics. Publication II showed that big data adoption is progressing rapidly in Finland. It is highly probable that big data ecosystems are emerging as well. Exploring and explaining how data in general and particularly big data can add value to an ecosystem and what the key capabilities that an incumbent should develop would be good candidates for future studies.

Another direction for new studies would be the impact of social relations. Humans drive change forward; they make decisions that promote (or demote) change and big data adoption. Furthermore, they develop new business models. Publication II indicated that the attitudes of IT management may play an important role in big data adoption. Investigating the social relations between the executives, both inside and outside the company, and especially the role of IT management could shed light on the key enablers in big data adoption and the obstacles to it.

There are also more technical aspects that relate to business models, ecosystems and the value generation of big data. Information sharing is essential for any business ecosystem. In order to effectively share data both systems integration and data compatibility must be developed. A service oriented architecture (SOA) is one architectural style that enables

software to be constructed in a business-oriented way. Alkkiomäki (2016) proposed an ontology that enables enterprises to connect SOA as a part of their business model. Extending the viewpoint to the ecosystem level could help to understand the value generation process. Moreover, it would help to justify the costs related to building the services within the ecosystem. It would also pave the way towards data-driven innovation (Shaughnessy 2015), which was briefly discussed above in the Section "Innovation Research".

Trailblazers such as Netflix already exploit data-driven innovation methods (Amatriain 2013). Combining big data and innovation into an automated process is a new concept. Researching how the process should be constructed, what components are essential, and the theoretical foundations of the value creation mechanisms in the process would greatly help practitioners. Another direction for future research is more technical. According to the definition (see the section "Innovation Research") an idea must be put into practice to become an innovation. Creating an innovation automation platform is a technical challenge. SOA can be one approach to solve data integration issues, while proprietary analytics or integration with engines like IBM's Watson[29] might be useful for producing insights from the data. Studying the practices and solutions that the trailblazers have implemented would benefit incumbents.

5.3 Reliability and Validity

This research uses a mixed methods approach, yet the nature of the research is qualitative as discussed earlier. It intends to reach out to understand, describe and explain big data in the context of enterprises. Qualitative research has no fixed set of designs. Therefore, several different procedures can be used to check the accuracy of the results. Indeed, over the years several different approaches have been proposed. For example, Yin (2011) and Creswell (2013) present numerous best practices that can be used to increase the reliability and validity of research.

Reliability refers to the consistency of the researcher's approach. According to Yin (2011) researchers should document their procedures in detail. This enables others to repeat or at least review the research. Maxwell (1992) uses the term *descriptive validity* which is rather closely related to reliability. Descriptive validity can be evaluated by assessing the factual accuracy of raw data gathering. This research aims to be as transparent as possible in order to help the assessment of its reliability. Examples of the selected approach include the following:

- Documentation of the literature review processes (Publications I and III) at a detailed level so that the process can be repeated.

[29] "IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data." http://www.ibm.com/watson/

- Use of commonly used statistical algorithms (Publications II and V) which makes it easier to repeat the analysis process. Moreover, this helps to understand and interpret the results, which in turn makes it easier to evaluate the reliability of the study.
- Member checking of the interviews (Publication IV) ensures the correct interpretation of the transcripts.
- Computerised text mining (Publication III) increases the reliability of the study in two ways. Algorithmic coding removes the potential bias of manual coding errors. Secondly, repeating and verifying the analysis procedure can be done effortlessly.

Validity refers to the procedures that the researcher uses to ensure the accuracy of the findings (Creswell 2013). The research started with a "questions first" –approach. This approach is a commonly used best practice (Yin 2011). The questions stem from the paradigm shift caused by digitalisation, which has already started. It can be seen that this shift has been – and increasingly is – disruptive by nature. Disruptive changes are challenging to incumbent firms. The questions rose from the desire to understand the phenomenon. Additional scoping came from the background of the researcher. The pragmatic worldview and history of adding value with data led to the purpose and design of the research. That said, the process was not pre-determined, but iteratively focused instead. This is common in qualitative studies (Maxwell 1996). Several design decisions, such as deciding on the methodologies and re-designing the research questions used in the articles were made during the research process. Some of the changes were based on feedback from the reviewers, sometimes they were seen as necessary as the understanding and knowledge of the phenomenon increased. Overall the approach was pragmatic and involved setting a goal and selecting the tools along the way as required by the current task and viewpoint.

This iterative approach, of course, includes the possibility of bias. The researcher may – inadvertently or not – influence the findings. This is related to research integrity. According to Yin (2011), research integrity corresponds to the researcher's "truthful positions and statements". Disclosing personal roles and traits is one possible way to demonstrate the integrity of the research. The implications of big data are various, there are threats as well as possibilities. The background of the researcher has certainly had an effect on the selected viewpoint or research lens. While recognising the threats, such as privacy concerns or data dictatorship, the research focuses on creating value with big data in a business context. This is based on the researcher's previous experiences in practice. As many of these experiences have shown that firms are able to derive value from data, it was natural to start exploring the big data phenomenon from this viewpoint. A researcher with a different background might have made different choices regarding the viewpoint of the research.

Another concern related to research integrity that should be disclosed is the role of the researcher's employer in the context of the research. It is commonly recognised, e.g. (Creswell 2013; Yin 2011), that researchers should be cautious with so-called "backyard studies", i.e. when studying their own organisations. From the outset of this study the researcher decided that although it would be possible to carry out the research either in-house or in co-operation with the employer, this would not be the case. The research is interesting for both parties, but the employer has played no role in the research, nor in the selection of the study subject.

Besides the above mentioned measures, **triangulation** (Mingers 2001), i.e. mixing methods and looking at the phenomenon from different perspectives has been an important vehicle to ensure the validity of the research. For example, the value of big data has been approached in Publications II, III and VI as well as in the introductory part of the dissertation. Publication III revealed that firms had been able to gain new value from big data, Publication II showed that Finnish big data practitioners were satisfied with their big data projects. Publication VI and the introductory part looked for the mechanisms and value creation processes. All these different angles point in the same direction, i.e. that the generic value potential proposed in various sources, e.g. (Linturi et al. 2013; Manyika et al. 2011; Mayer-Schönberger & Cukier 2013), really can be achieved at an enterprise level.

Generalisability is the extent the results can be applied in other contexts. Maxwell (1992) views generalisability as one aspect of research validity. Creswell (2013) agrees on the meaning of the term, but in contrast to Maxwell's view, he states that generalisability and validity are different things. Nevertheless, the two mentioned scholars among many others see generalisability as an important aspect of research. The most obvious generalisable part of this research is the value conversion process presented in Publication VI. Moreover, Publication II, which surveyed management's attitudes towards big data in Finnish private sector enterprises can be generalised. This survey could be repeated in other countries or in the public sector. Moreover, it could be repeated over time to determine how the attitudes in question develop. Another part subject to generalisation is the method used in Publication III. The computerised text mining approach described in the article is generic; it could be applied to any set of texts to reveal common themes or clusters. In addition, the findings of Publication I point out the terminology related challenges of big data in a generic way. Understanding the importance of distinguishing between big data, its value and the phenomenon apply quite universally.

6 Conclusions

Don't be surprised, if big data takes you by surprise.

Digital transformation is justified simply by economic value. Enterprises are eagerly utilising digital transformation in numerous ways, such as increasing their operational efficiency, creating new services and products, or innovating new business models. Organisations are replacing manual routines with digital processes and algorithms. Garbage bins have sensors that indicate when they need to be emptied, chat bots backed up with artificial intelligence take care of customer service activities, and companies can access world-wide markets through e-commerce platforms. The scale and speed of the transformation is breath-taking. The first video was uploaded onto YouTube in 2005. Now, 300 hours of video are uploaded every *minute*[30]. Apple's iPhone, the first smart phone with significant market penetration, was released in 2007. Ten years later, 90 % of Alibaba's Single's Day[31] sales, which generated 812 million e-commerce orders in just *one day*, were made using mobile devices[32]. The number of connected objects, ranging from light bulbs to cars, is expected to be almost 21 billion in 2020, roughly four times more than five years earlier[33]. Increasing digital activity means more data. Currently, enormous amounts of various types of data are generated by countless sources, often in real-time. Moreover, we are still at the early stages of the transformation.

6.1 Why Should Businesses Care about Big Data?

Several studies, many of them referenced in the articles, show that in general, data-intensive firms perform better than their rivals. It is most likely that the trend towards more data-intensive business will strengthen this phenomenon as the data deluge proceeds. YouTube on mobile devices alone reaches more people in the 18-49 years age group than any cable network in the United States. Money will follow the audience. Many enterprises are capable of building successful marketing campaigns around traditional media, such as television. Now they are increasingly learning how to the same on YouTube and other social media channels. Those who best understand and utilise the data available on the current platforms, will be most successful. For example, profiling and identifying potential customers using social media data, combined with an e-commerce site might create a world-wide market for a niche product.

According to Publication I, big data has numerous definitions, some of them logically flawed. Nevertheless, most definitions contain three dimensions – volume, velocity, and variety – which characterise big data. This dissertation defines big data as "high-volume,

[30] https://fortunelords.com/youtube-statistics/
[31] http://fortune.com/2018/11/09/alibaba-singles-day-china/
[32] https://techcrunch.com/2017/11/11/alibaba-smashes-its-singles-day-record/
[33] https://www.gartner.com/newsroom/id/3165317

high-velocity and high-variety information assets" (p. 40). High volumes typically consist of large amounts of detailed data. Velocity refers to real-time data. Variety has two interpretations: some sources consider this to refer to different types of data, such as transactions, text or video, whereas others see variety as data combined from various sources. Different, detailed types of data from various sources gathered in real-time are elements that are connected to data value creation.

However, big data as such is hardly valuable. Even selling the data, not to speak of more advanced uses, requires that it must be extracted from the original sources and augmented with other fragments of data, such as the date, location, or other identifiers that contextualise, or explain the origins, of data. In other words, the data must be converted into information before it is useful for value creation. The extraction and augmentation process is not trivial in most cases, as discussed e.g. in Publication III. Due to the effort required and uncertainties of the process (see Publication VI), we must treat big data as a potential asset.

C1. Big data is a *potentially* valuable asset.

With regard to the research questions of this dissertation, conclusion C1 underlines the characteristics of big data and the requirements behind the value creation. The potential may be huge, but to realise it as added business value, organisations need to put in serious effort both in technical and non-technical areas. The conclusion may be simplified and may even sound lame. However, without proper understanding of the factors behind C1, such as the characteristics and nature of the phenomenon (Publication I), challenges (Publication III), or organisational and process aspects (Publications II and VI), a big data initiative does not convert into a transformation process and thus fails to fulfil the value proposition. This research dives deep into the big data phenomenon, providing insights that help to understand the nature, potential and requirements of big data. Thus, it helps incumbents in their transformation processes.

6.2 Where Should Businesses Look at to Avoid Pitfalls?

Most of the big data discussion deals with technical aspects such as analytics. However, the data is the root source of any possible value that big data initiatives may create. Moreover, often the data must be gathered in a timely fashion, or it is lost forever. No analytics software, regardless of how sophisticated it is, can derive value if the required data is not available. It all comes down to the data. Thus, the first thing is to concentrate on the data. What data do we have, or can we access? What will it require to extract value from the data?

Although big data should be considered an asset, it does not convert automatically into economic value. A large share of the current big data literature assumes that the data is available. In practice, this is rarely true. Data typically resides locked in silos such as operative system databases, document repositories, or external systems. Publication III

reported technical and data related challenges, such as tedious integration with legacy systems or difficulties in harmonising data, and that these aspects increase implementation costs. Although getting the data from the silos and converting it into harmonised information can be hard, it must be done. Otherwise it will not be possible to apply analytics to combined datasets, which is a key concept in big data value creation.

C2. Data residing in silos does not equal data assets.

Besides getting the data out of the silos, veracity, i.e. the trustworthiness or quality of the data, is an important factor in value creation. As Publication I showed, value is context dependent. A single error in one patient's medical record may lead to fatal consequences, whereas several errors are insignificant in records collected from millions of patients for statistical analysis. However, from the data usage perspective it is important to know the quality of data.

"Garbage in, garbage out" is a well-known adage among computer scientists that, when applied to big data context, states that the output of an algorithm can only be as accurate as the data entered in the processing. Indeed, veracity is actively studied by big data scholars (Publication VI) and the literature suggests various methods and measures for ensuring veracity. This research touched on the subject as well. Publication V developed metrics for measuring the accuracy of a developed algorithm, which is an indirect measurement of veracity. If the algorithm remains unchanged, but the accuracy changes over time, it indicates changes in the veracity.

C3. Knowing the veracity of the data is a cornerstone for a reliable big data solution.

With regard to the research questions, conclusions C2 and C3 highlight two crucial viewpoints that must be addressed in the transformation process towards big data driven business. Converting the data into a liquid form, i.e. extracting it from silos is a pre-requisite for secondary usage such as analytics or data monetisation. Correspondingly, understanding the veracity of the data highly affects its value. Implementing automated decision-making based on data makes no sense until the organisation knows and implements follow-up procedures for the quality of data. Conclusions C2 and C3 emphasise the two viewpoints discussed in the articles (especially Publications III. V and VI) as well as in numerous studies cited in this research. Organisations may find this research useful in creating a proper understanding of these aspects.

6.3 Who Should Be Involved in the Transformation?

The role of humans at all levels of an organisation plays a crucial role in adopting new information technology products and services. Especially the management's role in leading the change is important. Managers cannot force changes, but without their support the transformation becomes impossible. Publication II examined how Finnish executives

perceive big data. The findings show that managers have a positive attitude towards big data in general, and especially if they already have experience with big data. Other scholars have found that the management's support and commitment are essential success factors driving changes. However, this requires learning new skills. For example, some may find it hard to trust the data instead of their intuition.

Identifying and leading the development of new capabilities related to big data value creation is another matter that managers must focus on. Publication VI discusses many of these capabilities. Being able to answer business questions by applying analytics, such as text mining (Publication III) or predictive algorithms (Publication V), is a capability that includes both technical and social skills. Technical skills are required, but in order to make a business impact, the results must be communicated and explained in "business language". Making data-driven decisions is impossible without adequate organisational maturity, as well as analytic and data processing capabilities. One essential aspect are innovation capabilities, which were the subject of Publication IV. Transformation means change and change means innovation. Data-driven innovations require new thinking and experimenting with data.

C4. In the context of a single, incumbent enterprise, deriving economic value from big data requires committing to developing new capabilities.

For an incumbent enterprise identifying and developing the required capabilities is a long-term commitment. Not everything can be changed at once. Economic resources are one limiting factor, the learning curve regarding the required skills are another. However, often organisational factors and culture are those that most hamper the change. Incumbents organise themselves in order to be effective in a relatively stable environment. As a consequence, their organisations become change resistant. They may be well aware of disruptive technologies but either ignore or do not perceive the connection to their business until it is too late. From the perspective of a single firm, this can lead to catastrophic results. Kodak went into bankruptcy only a few years after they had categorised digital photography as a marginal phenomenon. Large hotel chains must have seen AirBnB long before it went big, but despite of that the potential of digital technology and data took them by surprise. Moreover, fearing and resisting change is a built-in feature of human nature. Managers and employees sticking to their comfort zones, in-house politics, sub-optimisation and personal interests among other organisational factors are all involved in managing the change.

Conclusion C4 approaches the research questions of this study from the perspective of understanding the big data phenomenon. Especially Publications III, IV and VI discuss this viewpoint. The organisation must actively seek new possibilities and support the necessary and related capability creation. This research helps to identify the gaps in current and future capabilities and skills.

6.4 How Should Big Data Be Positioned in Business and Research?

The paradigm shift towards data-driven business effectively means that big data – and data in general – are sneaking into the core areas of business. Therefore, big data implementations actually become business initiatives. They must be viewed either as programmes of continuous improvement, or new business development activities. If we look at big data implementations from the IT side, they are always custom-made, tailored solutions. They must be tightly interwoven with business. Defining goals, identifying and interpreting the required data, and verifying the results produced by algorithms require tight cooperation between technical and business people.

Another, and more important difference compared to traditional IT projects, is uncertainty. Business outcomes are by definition uncertain. Due to the business uncertainty, considering big data implementations as business initiatives effectively means that the outcome of any big data activity becomes uncertain as well. Publication VI discussed the value creation process in business contexts. Numerous factors, some of them internal, others external to the firm, can either promote or demote the process.

C5. Big data implementations are business initiatives, not technology projects.

With regard to the research questions of this study, the key point of conclusion C5 is to understand that the value creation process of big data is not a straightforward technology implementation. Instead, it is a business development process related to numerous factors that are beyond the control of the process. Publications II and III discuss some aspects of this issue; Publication VI takes a broader view, making this observation clearly visible.

The business environment is becoming more turbulent and fast-paced due to the digital transformation. Examples from customers requiring faster responses to start-ups attempting to disrupt existing business models underline the importance of information. Creating big data assets along with related capabilities and positioning those assets at the core of the business characterise the paradigm shift towards data-driven business.

7 References

Ackoff, R.L., 1989. From Data to Wisdom. *Journal of Applied Systems Analysis*, 16(1), pp.3–9.

Afuah, A., 2000. How much do your co-opetitors' capabilities matter in the face of technological change? *Strategic Management Journal*, 21(3), pp.377–404.

Akaka, M.A. & Vargo, S.L., 2014. Technology as an operant resource in service (eco) systems. *Information Systems and e-Business Management*, 12(3), pp.367–384.

Akter, S. et al., 2016. How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Production Economics*, 182, pp.113–131.

Akter, S. & Fosso Wamba, S., 2016. Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2), pp.173–194.

Alguliyev, R.M., Gasimova, R.T. & Abbaslı, R.N., 2017. The Obstacles in Big Data Process. *International Journal of Modern Education & Computer Science*, 9(3), pp.28–35.

Alharthi, A., Krotov, V. & Bowman, M., 2017. Addressing barriers to big data. *Business Horizons*, (in press).

Alkkiomäki, V., 2016. The role of service-oriented architecture as a part of the business model. *International Journal of Business Information Systems*, 21(3), pp.368–387.

Altshuler, M. Y.; Aharony N.; Pentland A.; Elovici Y.; Cebrian, 2011. Stealing Reality: When Criminals Become Data Scientists (or Vice Versa). *IEEE Intelligent Systems*, 26(6), pp.22–30.

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*, pp.2201–2208.

Amit, R. & Zott, C., 2001. Value creation in e-business. *Strategic Management Journal*, 22(6/7), pp.493–520.

Anand, A., Sharma, R. & Coltman, T., 2016. Four steps to realizing business value from digital data streams. *MIS Quarterly Executive*, 15(4), pp.259–277.

Anderson, C., 2008. The end of theory: The data deluge makes the scientific method obsolete. Wired Magazine, 23-Jun-2008. Available at: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/.

Ardagna, C.A., Ceravolo, P. & Damiani, E., 2016. Big data analytics as-a-service: Issues and challenges. In *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, pp. 3638–3644.

Arora, D. & Malik, P., 2015. Analytics: Key to go from generating big data to deriving business value. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*. IEEE, pp. 446–452.

Baregheh, A., Rowley, J. & Sambrook, S., 2009. Towards a multidisciplinary definition of innovation. *Management Decision*, 47(8), pp.1323–1339.

Bärenfänger, R., Otto, B. & Österle, H., 2014. Business value of in-memory technology–multiple-case study insights. *Industrial Management & Data Systems*, 114(9), pp.1396–1414.

Barney, J., 1991. Firm resources and sustained competitive advantage. *Journal of management*, 17(1), pp.99–120.

Barney, J.B., 1995. Looking inside for competitive advantage. *The Academy of Management Executive*, 9(4), pp.49–61.

Baskerville, R.L., 1999. Investigating information systems with action research. *Communications of the AIS*, Vol. 2, Article 19.

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*, pp.3728–3737.

Benbasat, I. & Zmud, R.W., 1999. Empirical research in information systems: the practice of relevance. *MIS Quarterly*, 23(1), pp.3–16.

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

Berghel, H., 2013. Through the PRISM darkly. *Computer*, 46(7), pp.86–90.

Bettencourt-Silva, J.H. et al., 2015. Building Data-Driven Pathways From Routinely Collected Hospital Data: A Case Study on Prostate Cancer. *JMIR medical informatics*, 3(3), pp.1–21.

Boncea, R. et al., 2017. A Maturity Analysis of Big Data Technologies. *Informatica Economica*, 21(1), pp.60–71.

boyd, danah & Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), pp.662–679.

Braganza, A. et al., 2017. Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research*, 70(1), pp.328–337.

Brynjolfsson, E. & McAfee, A., 2012. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*, The MIT Center for Digital Business.

Budgen, D. et al., 2008. Using mapping studies in software engineering. In *Proceedings of PPIG*. pp. 195–204.

Burkhart, T. et al., 2011. Analyzing the business model concept—a comprehensive classification of literature. *Thirty Second International Conference on Information Systems, Shanghai 2011*.

Cai, H. et al., 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(December), pp.39–46.

Casadesus-Masanell, R. & Ricart, J.E., 2010. From strategy to business models and onto tactics. *Long Range Planning*, 43(2), pp.195–215.

Chae, B.K. et al., 2014. The impact of advanced analytics and data accuracy on operational performance: A contingent resource based theory (RBT) perspective. *Decision Support Systems*, 59, pp.119–126.

Chen, M., Mao, S. & Liu, Y., 2014. Big data: A survey. *Mobile Networks and Applications*, 19(2), pp.171–209.

Chen, Y. et al., 2011. Analytics ecosystem transformation: A force for business model innovation. *SRII Global Conference (SRII), 2011 Annual*, pp.11–20.

Christensen, C., 2013. *The Innovator's Dilemma: When new technologies cause great firms to fail*, Boston, Massachusetts: Harvard Business Review Press, 5th edition.

Chua, W.F., 1986. Radical developments in accounting thought. *Accounting review*, 61(4), pp.601–632.

Ciulla, F. et al., 2012. Beating the news using social media: the case study of American Idol. *EPJ Data Science*, 1(1), pp.1–11.

Clarke, R., 2016. Big data, big risks. *Information Systems Journal*, 26(1), pp.77–90.

Collis, D.J. & Montgomery, C.A., 1995. Competing on Resources: Strategy in the 1990s.

Comuzzi, M. et al., 2016. How organisations leverage Big Data: a maturity model. *Industrial Management & Data Systems*, 116(8), pp.1468–1492.

Cox, M. & Ellsworth, D., 1997. Application-controlled demand paging for out-of-core visualization. *Proceedings of the 8th conference on Visualization'97*, pp.235–244.

Creswell, J.W., 2013. *Research design: Qualitative, quantitative, and mixed methods approaches*, Sage publications.

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

Davenport, T.H., Barth, P. & Bean, R., 2012. How big data is different. *MIT Sloan Management Review*, 54(1), p.43.

Davis, F.D., Bagozzi, R.P. & Warshaw, P.R., 1989. User acceptance of computer technology: a comparison of two theoretical models. *Management science*, 35(8), pp.982–1003.

Dehning, B., Richardson, V.J. & Zmud, R.W., 2003. The value relevance of announcements of transformational information technology investments. *MIS Quarterly*, 27(4), pp.637–656.

Diebold, F.X., 2003. Big Data'Dynamic factor models for macroeconomic measurement and forecasting. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society (edited by M. Dewatripont, LP Hansen and S. Turnovsky)*. pp. 115–122.

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

Dyer, J., Gregersen, H. & Christensen, C., 2008. Entrepreneur behaviors, opportunity recognition, and the origins of innovative ventures. *Strategic Entrepreneurship Journal*, 2(4), pp.317–338.

Dyer, J., Gregersen, H. & Christensen, C., 2011. *The Innovator's DNA*, Boston, Massachusetts: Harvard Business Review Press.

Easterbrook, S. et al., 2008. *Selecting empirical methods for software engineering research* F. Shull, J. Singer, & D. I. K. Sjøberg, eds., London: Springer.

Eckhardt, A., Laumer, S. & Weitzel, T., 2009. Who influences whom? Analyzing workplace referents' social influence on IT adoption and non-adoption. *Journal of Information Technology*, 24(1), pp.11–24.

Eckhoff, D. & Sommer, C., 2014. Driving for big data? Privacy concerns in vehicular networking. *IEEE Security & Privacy*, 12(1), pp.77–79.

Elkan, C., 2001. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*. Lawrence Erlbaum Associates Ltd, pp. 973–978.

Emani, C.K., Cullot, N. & Nicolle, C., 2015. Understandable Big Data: A survey. *Computer Science Review*, 17, pp.70–81.

Ettlie, J.E. & Reza, E.M., 1992. Organizational integration and process innovation. *Academy of Management Journal*, 35(4), pp.795–827.

Fosso Wamba, S. et al., 2017. Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, pp.356–365.

Fosso Wamba, S. et al., 2015. How "big data"can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, pp.234–246.

Fox, S. & Do, T., 2013. Getting real about Big Data: applying critical realism to analyse Big Data hype. *International Journal of Managing Projects in Business*, 6(4), pp.739–760.

Frankel, D.A., 2012. Big Data and Risk Management. *Risk Management*, 59(8), p.13. Available at: http://www.rmmagazine.com/2012/10/01/big-data-and-risk-management/.

Frey, C.B. & Osborne, M.A., 2013. *The future of employment: how susceptible are jobs to computerisation*, Oxford, United Kingdom: University of Oxford.

Furr, N. & Dyer, J., 2014. *The Innovator's Method* N. Furr & J. Dyer, eds., Boston, Massachusetts: Harvard Business Review Press.

Gandomi, A. & Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp.137–144.

Gantz, J. & Reinsel, D., 2011. *Extracting value from chaos*, IDC. Available at: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

Gehrke, J., 2012. Quo vadis, data privacy? *Annals of the New York Academy of Sciences*, 1260(1), pp.45–54.

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

Grönroos, C. & Voima, P., 2013. Critical service logic: making sense of value creation and co-creation. *Journal of the academy of marketing science*, 41(2), pp.133–150.

Gupta, M. & George, J.F., 2016. Toward the development of a big data analytics capability. *Information & Management*, 53(8), pp.1049–1064.

Halamka, J.D., 2014. Early Experiences with big data at an academic medical center. *Health affairs*, 33(7), pp.1132–1138.

Harmaakorpi, V. & Melkas, H., 2012. *Practice-based Innovation. Insights, Applications and Policy Implications.* V. Harmaakorpi & H. Melkas, eds., Berlin, Germany: Springer.

Hartono, R. & Sheng, M.L., 2016. Knowledge sharing and firm performance: the role of social networking site and innovation capability. *Technology Analysis & Strategic Management*, 28(3), pp.335–347.

Hazen, B.T. et al., 2017. Toward understanding outcomes associated with data quality improvement. *International Journal of Production Economics*, 193, pp.737–747.

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(3), pp.464–472.

Hevner, A.R. et al., 2004. Design science in information systems research. *MIS Quarterly*, 28(1), pp.75–105.

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

Hu, H. et al., 2014. Toward scalable systems for big data analytics: A technology tutorial. *Access, IEEE*, 2, pp.652–687.

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(4), pp.1896–1912.

Huberty, M., 2015. Awaiting the second big data revolution: from digital noise to value creation. *Journal of Industry, Competition and Trade*, 15(1), pp.35–47.

Iansiti, M. & Levien, R., 2004. Strategy as ecology. *Harvard business review*, 82(3), pp.68–81.

Iddris, F., 2016. Innovation capability: A systematic review and research agenda. *Interdisciplinary Journal of Information, Knowledge, and Management*, 11, pp.235–260.

IFRS, 2015. *International Financial Reporting Standard for Small and Medium-sized Entities*, IFRS Foundation Publications Department, London, United Kingdom.

Iivari, J., 2003. The IS core-VII: Towards information systems as a science of meta-artifacts. *Communications of the Association for Information Systems*, 12(1), pp.568–581.

Iivari, M.M. et al., 2016. Toward ecosystemic business models in the context of industrial internet. *Journal of Business Models*, 4(2), pp.42–59.

InternetSociety, 2015. *Internet Society Global Internet Report 2015* M. Kende, ed., Internet Society. Available at: http://www.internetsociety.org/globalinternetreport/assets/download/IS_web.pdf.

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(4), pp.161–176.

Janssen, M., Voort, H. van der & Wahyudi, A., 2017. Factors influencing big data decision-making quality. *Journal of Business Research*, 70(1), pp.338–345.

Jetzek, T., Avital, M. & Bjorn-Andersen, N., 2014. Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research*, 9(2), pp.100–120.

Kane, G. et al., 2015. *Strategy, Not Technology, Drives Digital Transformation*, MIT Sloan Management Review and Deloitte University Press, July 2015.

Keen, J. et al., 2013. Big data+ politics= open data: The case of health care data in England. *Policy & Internet*, 5(2), pp.228–243.

Khan, N. et al., 2014. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*, 2014, pp.1–18.

Kitchenham, B., 2007. *Guidelines for performing systematic literature reviews in software engineering*, Keele University.

Kitchenham, B., 2004. *Procedures for performing systematic reviews*, Keele University.

Kitchin, R. & McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), pp.1–10.

Klein, H.K. & Myers, M.D., 1999. A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS quarterly*, 23(1), pp.67–93.

Koivumäki, T., Ristola, A. & Kesti, M., 2008. The perceptions towards mobile services: an empirical analysis of the role of use facilitators. *Personal and Ubiquitous Computing*, 12(1), pp.67–75.

Krippendorf, K., 1989. *Content analysis. In International ensyclopedia of communication* E. Barnouw et al., eds., New York, NY: Oxford University press.

Krumeich, J. et al., 2014. *Towards planning and control of business processes based on event-based predictions. In Business Information Systems. BIS 2014.* W. Abramowicz & A. Kokkinaki, eds., Springer, Chan: Springer.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6, pp.70–73.

Lewin, K., 1947. Frontiers in group dynamics II. Channels of group life; social planning and action research. *Human relations*, 1(2), pp.143–153.

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(1), pp.34–52.

Likert, R., 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Linturi, R., Kuusi, O. & Ahlqvist, T., 2013. Suomen sata uutta mahdollisuutta: Radikaalit teknologiset ratkaisut. *Eduskunnan tulevaisuusvaliokunnan julkaisu*, 6, p.2013.

Lusch, R.F., Vargo, S.L. & O'brien, M., 2007. Competing through service: Insights from service-dominant logic. *Journal of retailing*, 83(1), pp.5–18.

Lycett, M., 2013. Datafication: Making sense of (big) data in a complex world. *European Journal of Information Systems*, 22, pp.381–386.

Manyika, J. et al., 2011. *Big data: The next frontier for innovation, competition, and productivity* J. Manyika & M. Chui, eds., McKinsey Global Institute. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Marine-Roig, E. & Clavé, S.A., 2015. Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management*, 4(3), pp.162–172.

Markus, M.L. & Robey, D., 1988. Information technology and organizational change: causal structure in theory and research. *Management science*, 34(5), pp.583–598.

Martin, K.E., 2015. Ethical Issues in the Big Data Industry. *MIS Quarterly Executive, Forthcoming*.

Martinez, M.G. & Walton, B., 2014. The wisdom of crowds: The potential of online communities as a tool for data analysis. *Technovation*, 34(4), pp.203–214.

Mashey, J.R., 1998. Big Data... and the Next Wave of InfraStress. http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf. Available at: http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

Mathew, P.A. et al., 2015. Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, 140, pp.85–93.

De Mauro, A., Greco, M. & Grimaldi, M., 2015. What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*, 1644, pp.97–104.

Maxwell, J., 1996. *Qualitiative Research Design: An Interpretative Approach*, Sage.

Maxwell, J., 1992. Understanding and validity in qualitative research. *Harvard educational review*, 62(3), pp.279–301.

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(10), pp.61–67.

Merino, J. et al., 2016. A data quality in use model for big data. *Future Generation Computer Systems*, 63, pp.123–130.

Mikalef, P. et al., 2018. Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and e-Business Management*, 16(3), pp.1–32.

Miller, H.G. & Mork, P., 2013. From data to decisions: a value chain for big data. *IT Professional*, 15(1), pp.57–59.

Mingers, J., 2001. Combining IS research methods: towards a pluralist methodology. *Information systems research*, 12(3), pp.240–259.

Mithas, S., Ramasubbu, N. & Sambamurthy, V., 2011. How information management capability influences firm performance. *MIS quarterly*, 35(1), pp.237–256.

Moat, H.S. et al., 2014. Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences*, 37(01), pp.92–93.

Moniruzzaman, A. & Hossain, S.A., 2013. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*.

Moody, D.L. & Walsh, P., 1999. Measuring the Value Of Information-An Asset Valuation Approach. In *ECIS*. Frederiksberg, Denmark, pp. 496–512.

Moore, J.F., 1993. Predators and prey: a new ecology of competition. *Harvard business review*, 71(3), pp.75–83.

Mui, C. & Carroll, P., 2013. *The new killer apps - how large companies can out-innovate start-ups*, United States: Cornerloft Press.

Müller, S. & Jensen, P., 2017. Big data in the Danish industry: application and value creation. *Business Process Management Journal*, 23(3), pp.645–670.

Najjar, M.S. & Kettinger, W.J., 2013. Data Monetization: Lessons from a Retailer's Journey. *MIS Quarterly Executive*, 12(4), pp.213–225.

Newell, S. & Marabelli, M., 2015. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification. *The Journal of Strategic Information Systems*, 24(1), pp.3–14.

O'Leary, D.E., 2013. Exploiting big data from mobile device sensor-based apps: Challenges and benefits. *MIS Quarterly Executive*, 12(4), pp.179–187.

Orlikowski, W.J. & Baroudi, J.J., 1991. Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, 2(1), pp.1–28.

Orlikowski, W.J. & Iacono, C.S., 2001. Research commentary: Desperately seeking the IT in IT research—A call to theorizing the IT artifact. *Information systems research*, 12(2), pp.121–134.

Osgood, C.E., 1959. *The representational model and relevant research methods. In Trends in Content Analysis* I. Pool, ed., Urbana, United States: Illinois Press.

Osterwalder, A. & Pigneur, Y., 2010. *Business model generation: a handbook for visionaries, game changers, and challengers*, Hoboken, New Jersey: John Wiley & Sons.

Payne, A.F., Storbacka, K. & Frow, P., 2008. Managing the co-creation of value. *Journal of the academy of marketing science*, 36(1), pp.83–96.

Peffers, K. et al., 2007. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), pp.45–77.

Phillips-Wren, G. & Hoskisson, A., 2015. An analytical journey towards big data. *Journal of Decision Systems*, 24(1), pp.87–102.

Piccoli, G. & Pigni, F., 2013. Harvesting External Data: The Potential of Digital Data Streams. *MIS Quarterly Executive*, 12(1), pp.53–64.

Pigni, F., Piccoli, G. & Watson, R., 2016. Digital data streams. *California Management Review*, 58(3), pp.5–25.

Porter, M.E., 1991. Towards a dynamic theory of strategy. *Strategic management journal*, 12(S2), pp.95–117.

Porter, M.E. & Heppelmann, J.E., 2014. How smart, connected products are transforming competition. *Harvard Business Review*, 92(11), pp.11–64.

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

Pousttchi, K. & Hufenbach, Y., 2014. Engineering the value network of the customer interface and marketing in the data-rich retail environment. *International Journal of Electronic Commerce*, 18(4), pp.17–42.

Prescott, M., 2014. Big data and competitive advantage at Nielsen. *Management Decision*, 52(3), pp.573–601.

Prinsloo, P. et al., 2015. Big (ger) data as better data in open distance learning. *The International Review of Research in Open and Distributed Learning*, 16(1). Available at: http://www.irrodl.org/index.php/irrodl/rt/printerFriendly/1948/3203.

Rayport, J.F. & Sviokla, J.J., 1995. Exploiting the virtual value chain. *Harvard business review*, 73(6), pp.75–85.

Rea, L.M. & Parker, R.A., 2014. *Designing and conducting survey research: A comprehensive guide*, San Francisco, CA: John Wiley & Sons.

Ritala, P., Golnam, A. & Wegmann, A., 2014. Coopetition-based business models: The case of Amazon. com. *Industrial Marketing Management*, 43(2), pp.236–249.

Roden, S. et al., 2017. Big data and the transformation of operations models: a framework and a new research agenda. *Production Planning & Control*, 28(11-12), pp.929–944.

Rowley, J.E., 2007. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science*, 33(2), pp.163–180.

Ryan, G.W. & Bernard, H.R., 2003. Techniques to identify themes. *Field methods*, 15(1), pp.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.

Sandberg, B. & Aarikka-Stenroos, L., 2014. What makes it so difficult? A systematic review on barriers to radical innovation. *Industrial Marketing Management*, 43(8), pp.1293–1305.

Scharmer, C.O., 2001. Self-transcending knowledge: sensing and organizing around emerging opportunities. *Journal of Knowledge Management*, 5(2), pp.137–151.

Schmarzo, B., 2013. *Big Data: Understanding how data powers big business*, Indianapolis, Indiana: John Wiley & Sons.

Schumpeter, J.A., 1942. *Socialism, Capitalism and Democracy*, Harper and Brothers.

Sein, M. et al., 2011. Action design research. *MIS Quarterly*, 35(1), pp.37–56.

Shah, S., Horne, A. & Capellá, J., 2012. Good data won't guarantee good decisions. *Harvard Business Review*, 90(4), pp.23–25.

Sharma, R., Mithas, S. & Kankanhalli, A., 2014. Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, 23(4), pp.433–441.

Shaughnessy, H., 2015. *Shift: A User's Guide to the New Economy*, Boise, Idaho: Tru Publishing.

Shen, Y. & Varvel, V.E., 2013. Developing data management services at the Johns Hopkins University. *The Journal of Academic Librarianship*, 39(6), pp.552–557.

Soh, C. & Markus, M.L., 1995. How IT creates business value: a process theory synthesis. *ICIS 1995 Proceedings*, pp.29–41.

Stopczynski, A. et al., 2014. Privacy in sensor-driven human data collection: A guide for practitioners. *arXiv preprint arXiv:1403.5299*.

Subashini, S. & Kavitha, V., 2011. A survey on security issues in service delivery models of cloud computing. *Journal of network and computer applications*, 34(1), pp.1–11.

Sullivan, C., 2014. Protecting digital identity in the cloud: Regulating cross border data disclosure. *Computer Law & Security Review*, 30(2), pp.137–152.

Tallon, P.P., Ramirez, R.V. & Short, J.E., 2013. The information artifact in IT governance: toward a theory of information governance. *Journal of Management Information Systems*, 30(3), pp.141–178.

Tao, S. et al., 2014. Exploring Bus Rapid Transit passenger travel behaviour using big data. *Applied Geography*, 53, pp.90–104.

Taylor, F.W., 1911. *The Principles of Scientific Management*, JSTOR.

Teece, D.J., 2010. Business models, business strategy and innovation. *Long Range Planning*, 43(2), pp.172–194.

Teece, D.J., 2007. Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. *Strategic management journal*, 28(13), pp.1319–1350.

Therneau, T., Atkinson, B. & Ripley, B., 2017. Rpart: Recursive Partitioning,(2013). R package version 4.1-3. Available at: https://cran.r-project.org/web/packages/rpart/rpart.pdf.

Therneau, T.M. & Atkinson, E.J., 2017. An introduction to recursive partitioning using the RPART routines. *Rochester: Mayo Foundation*.

Timmers, P., 1998. Business models for electronic markets. *Electronic Markets*, 8(2), pp.3–8.

Tiwana, A., 2014. Separating Signal from Noise: Evaluating Emerging Technologies. *MIS Quarterly Executive*, 13(1), pp.45–61.

Truyens, M. & Van Eecke, P., 2014. Legal aspects of text mining. *Computer Law & Security Review*, 30(2), pp.153–170.

Tsourela, M. & Roumeliotis, M., 2015. The moderating role of technology readiness, gender, and sex in consumer acceptance and actual use of Technology-based services. *The Journal of High Technology Management Research*, 26(2), pp.124–136.

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

Van't Spijker, A., 2014. *The New Oil: Using Innovative Business Models to Turn Data Into Profit*, Basking Ridge, New Jersey: Technics Publications.

Vargo, S.L. & Lusch, R.F., 2004. Evolving to a new dominant logic for marketing. *Journal of marketing*, 68(1), pp.1–17.

Venkatesh, V. et al., 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly*, 23(4), pp.425–478.

Venkatraman, N., 1994. IT-enabled business transformation: from automation to business scope redefinition. *Sloan management review*, 35(2), pp.73–87.

Verhoeven, J.C., Heerwegh, D. & De Wit, K., 2010. Information and communication technologies in the life of university freshmen: An analysis of change. *Computers & Education*, 55(1), pp.53–66.

Vidgen, R., Shaw, S. & Grant, D.B., 2017. Management challenges in creating value from business analytics. *European Journal of Operational Research*, 261(2), pp.626–639.

Wade, M. & Hulland, J., 2004. Review: The resource-based view and information systems research: Review, extension, and suggestions for future research. *MIS quarterly*, 28(1), pp.107–142.

Weber, R.P., 1990. *Basic content analysis* M. Lewis-Beck, ed., Sage.

WEF, 2016. *The Future of Jobs*, Geneva, Switzerland: World Economic Forum.

Weill, P. & Woerner, S.L., 2015. Thriving in an increasingly digital ecosystem. *MIT Sloan Management Review*, 56(4), pp.27–34.

Weiss, S.M. & Indurkhya, N., 1998. *Predictive data mining: a practical guide*, San Francisco, California: Morgan Kaufmann Publishers.

Wernerfelt, B., 1984. A resource-based view of the firm. *Strategic Management Journal*, 5(2), pp.171–180.

West, M.A. & Anderson, N.R., 1996. Innovation in top management teams. *Journal of Applied Psychology*, 81(6), pp.680–693.

Westerlund, M., Leminen, S. & Rajahonka, M., 2014. Designing Business Models for the Internet of Things. *Technology Innovation Management Review*, 4(7), pp.5–13.

Yin, R., 2011. *Qualitative Reseach from Start fo Finish*, New York, NY: The Guilford Press.

Ylijoki, O. & Porras, J., 2016. Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management*, 4(1), pp.69–91.

Yu, K. et al., 2014. Mining hidden knowledge for drug safety assessment: topic modeling of LiverTox as a case study. *BMC bioinformatics*, 15, pp.1–8.

Zadrozny, B. & Elkan, C., 2001. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Seventh ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 204–213.

Zeleny, M., 1987. Management support systems: Towards integrated knowledge management. *Human Systems Management*, 7(1), pp.59–70.

Zhan, Y. et al., 2017. A big data framework for facilitating product innovation processes. *Business Process Management Journal*, 23(3), pp.518–536.

Zott, C., Amit, R. & Massa, L., 2011. The business model: recent developments and future research. *Journal of Management*, 37(4), pp.1019–1042.

Publication I

Ylijoki, Ossi and Porras, Jari
Perspectives to Definition of Big Data: a Mapping Study and Discussion

Reprinted with permission from
Journal of Innovation Management
Vol. 4(1), pp. 69-91, 2016

Journal of Innovation Management
JIM 4, 1 (2016) 69-91
HANDLE: http://hdl.handle.net/10216/83250

Perspectives to Definition of Big Data: A Mapping Study and Discussion

Ossi Ylijoki[1*], Jari Porras[1]

[1]School of Business and Management, Lappeenranta University of Technology, Finland
ossi.ylijoki@phnet.fi, jari.porras@lut.fi

Abstract. Big data is an emerging research area where common terminology is still evolving. Different perspectives to the research area and terminology exist, but a common definition for big data does not exist. We have performed a systematic mapping study in order to identify different big data definitions and their perspectives. As a result, we present a state-of-the-art review of the current status in big data definitions, discuss the shortcomings of the current definitions, and propose possible solutions for the shortcomings. The paper contributes to the emerging big data research by analyzing current definitions of big data from different perspectives, suggesting enhancement to the terminology as well as pointing out new research avenues. In addition, the article helps new researchers and practitioners to understand what big data is, and bridges the knowledge between theory and practice.

Keywords. Analytics, Big Data, Big Data Definition, Business Model, Datafication, Digitization, Knowledge Management.

1 Introduction

Digitization is a current megatrend, meaning that digital technologies are integrated into our everyday life. The use of digital technologies enables the connection of different services and automation of many processes. Although digitization itself is an important technological (r)evolution, it enables even more fundamental change: datafication. An increasing number of devices and sensors are constantly connected to the Internet. Cameras, mobile phones, tablets, various applications and services running on them produce wide varieties of digital data. This data generation phenomenon is called datafication (Mayer-Schönberger and Cukier, 2013). Lycett (2013) defines datafication as a "sense-making process", which emphasizes the value generation aspect. Digitization and datafication make it possible to capture different situations, actions, or even series of events in the form of data. A vague term "big data" describes the data resulting from datafication. This phenomenon has widespread effects.

As an example, let us consider quadcopters. Amazon and DHL[1], among others, are prototyping these small flying devices for delivering goods to customers. Quadcopters

1 Amazon Prime Air. http://www.amazon.com/b?node=8037720011. Accessed 28th April 2016.
DHL: http://www.theguardian.com/technology/2014/sep/25/german-dhl-launches-first-commercial-drone-delivery-service. Accessed 28th April 2016.

ISSN 2183-0606
http://www.open-jim.org

gather vast volumes of different types of data in real-time (e.g. sensor readings, video and geolocation data) in order to be able to perform their tasks autonomously. They analyze and use the data in many tasks, such as avoiding collisions and orienting their way to the destination. In addition, they synthetize and distribute data. Sending data, such as location and altitude to the command center is essential for the fleet management.

As the fleet of a firm might contain thousands of quadcopters, this represents a real-world big data problem. In general, there are numerous *technical challenges* to conquer for organizations that wish to benefit of big data, see e.g. (Ma et al., 2013; Chen et al., 2014; Kambatla et al., 2014). So far, humans supervise most quadcopter experiments, but due to rapid technical advances, it is obvious that in the near future these little flying machines will become autonomous. Hardware and software vendors are investing heavily in their offerings, so this area is progressing rapidly.

Big data resulting from digitization is seen as a *significant opportunity*, see e.g. (Manyika et al., 2011; Mayer-Schönberger and Cukier, 2013; Schmarzo, 2013; Davenport, 2014). Big data is considered as a key enabler that can be used to generate value in private companies and public organizations. Governments have initiated big data strategies[2]. Examples of the benefits include creating new business opportunities, boosting R&D activities, and supporting decision making (Amatriain, 2013; Mehta et al., 2013; Lee et al., 2014). Quadcopters, among other technological solutions, can be used to save costs and even enable new business models for many organizations, both in the private and the public sector. There are naturally also questioning voices that criticize the big data paradigm and value proposals (see e.g. (boyd and Crawford, 2012; Fox and Do, 2013).

Datafication and big data are disruptive technologies that have widespread implications on the society. Technology vendors, the public sector, private companies, consumers, and policy makers, among others, have interests in the field. Moreover, as the number of stakeholders and parties increases, common understanding of the terminology and concepts becomes more and more important. Unfortunately, big data is a volatile term now. Different definitions of big data can be found in the literature, as well as among practitioners. A (theoretical) definition is a proposal for understanding the meaning of a term. It should be observable, clear (i.e., unambiguous) and complete. Good definitions improve the quality of communication significantly and enable common understanding among participants from different backgrounds. To put it simply, a good definition equals clarity.

The purpose of this article is – considering the broad implications of big data on the

2 E.g. European Big Data Value Strategic Research & Innovation Agenda. http://www.nessi-europe.eu/Files/Private/EuropeanBigDataValuePartnership_SRIA__v099%20v4.pdf. Accessed 28th April 2016.
U.S. Big data initiative: https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf. Accessed 28th April 2016.

society, organizations, and individuals – to shed light on the definition of big data. As the method, we use 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 broad research questions. Our research questions are:

- What kind of definitions of big data exist in research papers and among practitioners?
- How has the definition of big data evolved?
- How do the definitions reflect the different characteristics and perspectives of big data?

2 Literature Search

Our initial search covered three major reference databases: Scopus, ProQuest, and Web of Science. We considered this as a good starting point, as these databases index a broad range of papers, covering both technical and business fields. Figure 1 gives an overview of the search process. In addition to wide research questions, Kitchenham (2007) suggests that mapping studies should use rather loose search criteria. We searched the databases (title, keywords, abstract) by using ("big data" and "definition") as a search string. All papers written in English and indexed up to 02-Sep-2015 were included in the initial result set. No additional limitations were set. A total of 479 papers were identified. Next, we removed duplicate articles (117).

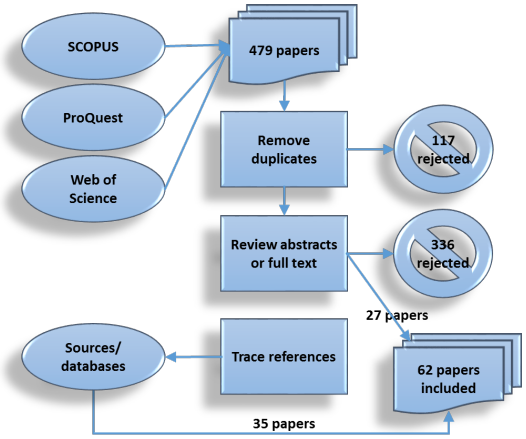


Fig. 1. Search process.

After removing the duplicates, we read the abstracts, and where necessary, the whole text of each of the resulting papers. We categorized the papers by using the following inclusion/exclusion criteria: If the paper contains a definition of big data, include it, otherwise reject it. Due to the loose search criteria, a number of papers defining other

things than big data were included in the initial search. Papers that obviously did not meet the eligibility criteria were rejected. If the decision was not clear, we performed a full text review, and the paper was either included or excluded on the basis of the review. Additional 17 papers were excluded because they were either commercial, high-price reports or they could not be found. As a result of this phase, 27 papers were included in the result set.

In the reference-tracking phase, we searched for additional papers on the basis of citations in the included papers (backward snowballing). Possibly interesting references were checked in the article context, and if still promising, they were tracked from databases or web sources, including Google Scholar and various web pages. If the article met the eligibility criteria, it was included. We identified additional 35 papers in this phase.

At the end of the search process phase, we had identified 62 papers that contained a definition of big data. The year-wise distribution of these papers is presented in Figure 2. It seems that although the first definition was presented more than 10 years ago, the discussion of the definition of big data started only a few years back. These papers and their definitions were examined further.

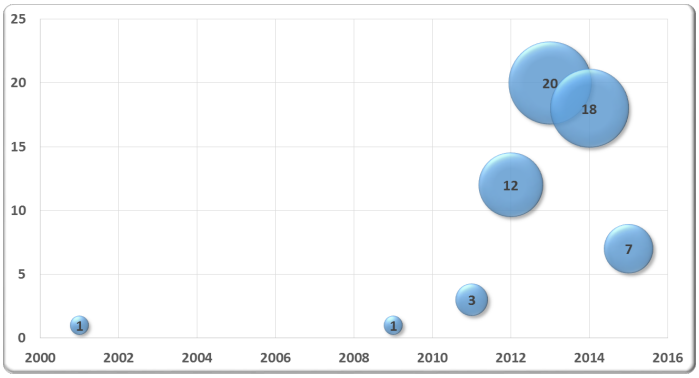


Fig. 2. Year-wise distribution of found papers.

3 Analysis of the Definitions

The first part of the analysis covers the evolution of the definition of big data. Definitions, their existence in time, as well as similarities and differences are presented. This analysis reveals what perspectives (or components) various participants have added to the definition over time. The second part of our analysis identifies gaps between the current definitions and big data value propositions, in order to find out what perspectives are still missing.

3.1 Evolution of the Term Big Data

The term "big data" is not new. It has been used both in research and non-research papers for quite a long time. Back in 1997 it was used in the context of visualizing large data sets (Cox and Ellsworth, 1997). In 1998 it was used in a hardware-related presentation (Mashey, 1998) and also in the data mining context (Weiss and Indurkhya, 1998), and 2003 in combination with statistics (Diebold, 2003). In the beginning, big meant the size and all these sources recognized and referenced big data with the increasing volumes of data. However, year 2001 can be considered as a major milestone in the definition of big data. Laney (2001) described three essential dimensions of big data: *volume*, *velocity* and *variety*. Operating with a swarm of autonomous quadcopters requires the management of high-volume, high-velocity (real-time) data that have many types (variety).

During the following decade, trailblazers like Google and Amazon developed practical big data solutions. These solutions have proved to add value to their businesses. In fact, the trailblazers build their business models on big data solutions. An article published in 2008 in the Wired magazine (Anderson, 2008) aroused public interest in the use of big data and its effects in science. The next significant milestone was 2011, when McKinsey Global Institute and IDC published reports (Gantz and Reinsel, 2011; Manyika et al., 2011) that drew wide public attention to the potential value of big data. Since then a number of newspaper articles, scientific big data papers and books have been published.

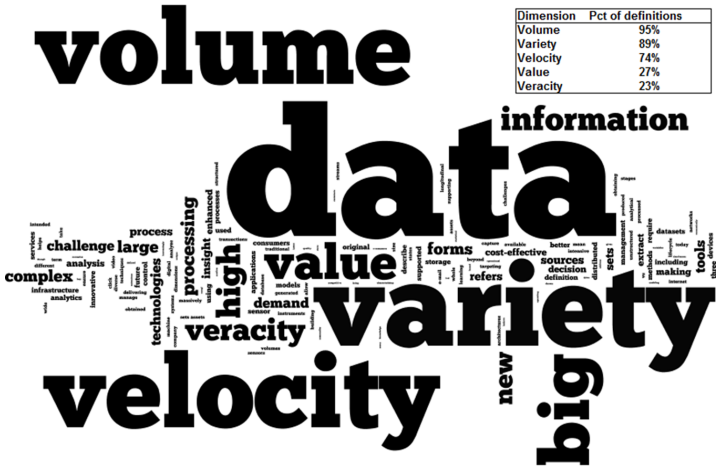


Fig. 3. Most common big data characteristics.

We considered Laney (2001) to be the one to offer the first real definition, although the term big data had been used earlier. In our analysis of the studies, we could not identify references to earlier papers. Naturally, Laney must have been influenced by

earlier work, but his paper was the first to introduce the three big data dimensions: volume, variety and velocity. Most of the definitions rely at least partly on the 3V definition by Laney (2001). Figure 3 shows the most common characteristics used in the definitions of big data (see Appendix 1 for details of the definitions). 95% (59 occurrences out of 62) of the papers identified volume as a key characteristic of big data. In addition, the papers considered variety (55 occurrences) and velocity (46) to be typical big data factors. Value (17) and veracity (14) had also caught attention. These five dimensions dominate the current definitions of big data.

The included 62 papers (see Appendix 1 for details) were arranged by their publishing date, and each paper was inspected against previously published definitions. If the paper contained a new definition or added some new elements to the existing definitions, it was considered to be a new definition. This analysis resulted in 17 different definitions. These 17 definitions have similarities in the sense that many of them aim to widen the 3V definition to cover technical and especially business aspects. This is quite a natural consequence with regard to the big data value proposal. However, wide definitions can be problematic, and some essential aspects of big data are still lacking. We will discuss these aspects below. The rest of the papers (45) contained definitions essentially covered in earlier papers. Appendix 1 presents details of the definitions.

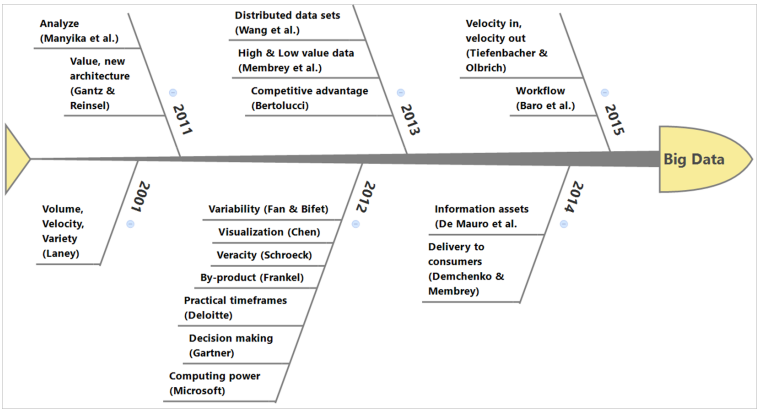


Fig. 4. Evolution of the definition of big data.

The fishbone diagram in figure 4 gives an overview of the evolution. The bones show essential additions of all 17 different definitions, i.e. new aspects or components that each definition adds. Laney (2001) presented the original, so-called 3V definition of big data. The Vs come from volume, velocity and variety. Volume refers to ever-increasing amounts of data. Velocity indicates the need to capture and analyze high-speed or bursts of data in (near) real-time, or else the value may be lost. Variety is related to different types of data, be it structured or non-structured, such as social media posts or a video.

The 3V definition was the de-facto big data standard until 2011, when both Manyika et al. (2011) and Gantz and Reinsel (2011) published their reports. Manyika et al. (2011) emphasize the potential value of big data, but curiously enough, their definition focuses on data volume including only a hint ("analyze") of the value. Also, when compared to Laney (2001), Manyika et al. (2011) have left out velocity and variety. Gantz and Reinsel (2011) include the three Vs, and add value extraction and new architectures. They have also decided to define big data technologies instead of big data. This approach allows them to balance the definition between data, technology and business components with a reasonable logic.

The big data hype was at its peak in the years 2012 and 2013. Several aspects of big data were discussed, such as privacy, security, (business) value, and veracity. We identified seven definitions from 2012 that were either completely new, like the one by Microsoft (2012), or added new components to existing definitions (Gartner, 2012; Schroeck et al., 2012; Fan and Bifet, 2013), and three from the year 2013. After that date we identified four more additions. Two of these later definitions (Demchenko, DeLaat, et al., 2014; Baro et al., 2015) note the importance of delivering the results to consumers. This analysis showed that the evolution of the definition started with data and especially data volumes, and then the discussion shifted to infrastructure topics, followed by the (business) value of data. Finally, more fine-grained aspects, like data delivery and collaboration, appeared.

3.2 Definitions vs. Big Data Value Chain

An interesting question is how the 17 different big data definitions reflect the significant value proposal of big data? Several frameworks explain how data adds value. One of the first of such models is the Virtual Value Creation (VVC) framework presented by Rayport and Sviokla (1995). This framework describes five steps that are required to create value from data: gather, organize, select, synthesize, and distribute (see Figure 5). The steps gather and organize are data-related, and they cover aspects like data acquisition from sensors, integration with other data, and data storing. The steps select, synthesize and distribution depend on data usage. They are activities like filtering data for analysis, or represented as artifacts like analytical models, data visualization, and information delivery tools. Value is expected to increase as data items from various sources are combined to form meaningful information chunks in the VVC process.

A quadcopter reads its current location from the GPS sensor and combines it with the destination information (gather, organize). Based on the analysis, it may take a decision to change its direction (select, synthetize). At frequent intervals, the copter sends data (e.g. location) to the command center (distribute). This simple VVC process adds value, as it enables the copter to work autonomously. However, taking a helicopter view by looking at the whole fleet instead of one quadcopter, it becomes clear that much more value is available. The command center systems gathers data from each of the copters and other sources, e.g. from delivery orders (gather, organize). An analytical model calculates the routes (select, synthetize) and sends instructions (like pick-up and delivery addresses) to each of the copters (delivery). This automated VVC process creates value from the data by producing optimal routes, maximizing the number of deliveries and increasing efficiency.

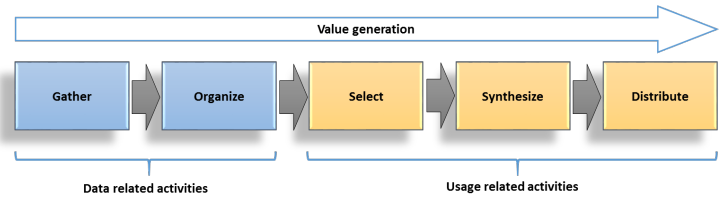


Fig. 5. Virtual value creation process.

Table 1 maps the 17 different definitions to the big data value chain. Together the current big data definitions cover all phases of the value chain. However, most of the definitions cover only parts of the chain. There are two definitions that consider all five phases, those of Demchenko, DeLaat, et al. (2014) and Baro et al. (2015).

Note that the table shows which phases of the value chain the new perspective of each definition emphasizes. This is for clarity: many of the definitions cover also other phases, e.g. Demchenko, DeLaat, et al. (2014) have also covered other steps. However, the new perspective of their definition is the delivery aspect, and therefore only the distribute phase is included in Table 1.

Table 1. Mapping the new perspectives of the big data definitions to the value chain.

Authors	New perspective	Gather (data)	Organize (data)	Select (usage)	Synthesize (usage)	Distribute (usage)
(Laney, 2001)	Volume, velocity, variety	x	x			
(Manyika et al., 2011)	Analyze			x	x	
(Gantz and Reinsel, 2011)	Value, new architecture		x	x	x	
(Microsoft, 2012)	Computing power		x	x	x	
(Gartner, 2012)	Decision making			x	x	
(DeloitteConsulting, 2012)	Practical timeframes		x	x	x	
(Frankel, 2012)	By-product	x				
(Schroeck et al., 2012)	Veracity	x	x	x	x	
(Chen et al., 2012)	Visualization				x	
(Fan and Bifet, 2013)	Variability	x				
(Wang et al., 2013)	Distributed data sets		x			
(Membrey et al., 2013)	High & low value data		x	x		
(Bertolucci, 2013)	Competitive advantage				x	
(Demchenko, DeLaat, et al., 2014)	Delivery to consumers					x
(De Mauro et al., 2015)	Information assets		x			
(Tiefenbacher and Olbrich, 2015)	Velocity in, velocity out	x	x	x	x	
(Baro et al., 2015)	Workflow					x

4 Discussion

As can be seen in the definitions and analysis, big data can mean different things, depending on the selected viewpoint. Some perceive big data as a technical challenge, others view it as a vehicle to increase efficiency or profits. In this section we will show that combining data and its intended usage leads to vague definitions, and consider how the disruptive nature of big data should be taken into account.

4.1 Separate Data and Its Usage

Our analysis revealed that several definitions have logical incoherencies. Value, for example, must be derived from the data by using analytics, there is no value in plain data as such (Ackoff, 1989). Value is also case-dependent. A certain piece of information may be worthless to one company but highly valued by some other firm or in another situation. For example, quadcopter flight details are much more valuable in case of an accident than in a normal situation. This is emphasized by Mayer-Schönberger and Cukier (2013) who state that the value of big data is in the secondary uses of the data. For veracity, analytics is required to determine whether the data is relevant for the planned usage. As important factors as value and veracity are in practice, they do not define the characteristics of big data, but instead they reflect the *usage* of the data. Vague definitions are typically hard to understand as they raise questions that cannot be answered coherently. This will lead to different interpretations and misunderstandings.

The original 3V definition (Laney, 2001) leaved the business effects out. This is one of the main reasons why many new definitions have emerged. Both technology vendors and enterprises have an interest to add a value proposition. Companies see big data as a vehicle to gain value, vendors naturally like to justify the costs of their offerings with potential benefits. A natural tendency would be to add a value component to the definition. However, as discussed above, value is not a characteristic of data. Definitions should be clear and unambiguous. Therefore, adding data usage to the definition is not a good idea, as the definition would become unambiguous, and coherency would be lost.

Our suggestion to the problem is that the *data and its usage should be separated*. Data is similar to oil: when combined with data management and analytics processes it provides organizations with value. Analytics and data usage are of course essential elements in successful big data exploitation. However, from the definition point of view, combining data and its usage is like combining oil and engine into one single definition. Separating big data from its intended usage clarifies the inconsistencies of the definitions and helps us to understand the plain characteristics of big data. As the purpose of data usage is to realize the potential value of the data, we propose the term *big data insights* to be used in any context in data usage -related activities (see also figure 5).

4.2 Other Perspectives – Big Data as a Phenomenon

In addition to technical and value aspects, scholars have focused on several other perspectives to big data, such as privacy, security (Altshuler, 2011; Berghel, 2013; Lu et al., 2014), and policy-making (Keen et al., 2013; Blume et al., 2014; Truyens and Van Eecke, 2014). None of the current definitions of big data consider these. These aspects are not characteristics of big data; we do not suggest that these aspects should be included in the definition. Instead, they are aspects that help to understand big data as a phenomenon. Moreover, these perspectives are important, as failing to consider them can drive an organization to difficulties.

Another, even more important aspect is that the current definitions neglect the disruptive nature of big data. On the basis of the literature it seems obvious that in the future, big data will have significant impacts on businesses (Manyika et al., 2011; Schmarzo, 2013; Davenport, 2014). Big data is seen as a technology that can have huge impacts on most industries and enterprises. Data-driven companies can achieve significant benefits (McAfee et al., 2012), but transformational business changes (Dehning et al., 2003) are required to achieve full competitive advantage from big data. The impact of big data will be significant, but the nature of the change is even more important. The effects of big data on firms, ecosystems and industries will be disruptive (Earley, 2014; Fan and Gordon, 2014; Kim et al., 2014). Industry structures are changing, and new business opportunities are emerging. On the other hand, this means that also competitors may be able to invent new business models, not to speak of new entrants, which will increase the turbulence effectively. The impacts of big data may – and will – be positive for some organizations, negative for others. Due to the disruptive nature of big data, companies must review their business models in order to reveal possible threats and opportunities. Moreover, as the disruptive drivers are technological by nature, these technologies and their potential effects must be linked with strategy.

We suggest that a *new definition for big data as a phenomenon should be considered*. For clarity and coherency, the definition of big data should cover only data and data management aspects (like the 3V definition). The phenomenon of big data is a broad concept that deserves a definition of its own. Instead of defining big data, the definition should consider several important perspectives of it. In our opinion, this definition should include the disruptive nature and strategic importance of the phenomenon. Adding these elements would help managers to understand the importance of the matter. This opens a new research avenue. Discussing and defining the nature and relations between various perspectives would build understanding of the broader context of big data, big data as a phenomenon.

5 Conclusions

Our aim was to shed light to the concept of big data, especially from the following viewpoints:

- What kind of definitions of big data exist in research papers and among practitioners?

- How has the definition of big data evolved?
- How do the definitions reflect the different characteristics and perspectives of big data?

A systematic mapping study was conducted in order to find answers to these questions. We made a search in major reference databases, search engines, and web sources containing both technical and business topics. A total of 62 sources were included in the result set. With regard to our research questions, we chose a broad search strategy in order to cover a wide range of possible sources. We identified 17 different definitions of big data that together presented a clear picture of the current situation and evolution of the definition, thus providing answers to our first and second research questions. We also compared the current definitions with various characteristics of big data. We found that the current definitions do not cover several perspectives that are discussed among big data scholars and practitioners, which answers our third research question. In addition, we identified several logically incoherent definitions. This clouds the matter further, as these definitions raise new questions, which will typically lead to ambiguous answers.

5.1 Results

This study revealed 17 different big data definitions from 62 relevant source papers. Each of the papers was analyzed against previously published definitions. If the paper contained a new definition or added some new elements to the existing definitions, it was considered as a new definition. The key contributions of this study are:

- Although there are various opinions on what big data is, the 3V definition by Laney (2001) contains three dimensions (volume, velocity, variety), which are common to most definitions. In addition to these dimensions, many definitions include technical parts and components related to the intended usage of the data, such as analysis or decision-making.
- Many of the definitions are logically inconsistent, which is one reason for the vagueness of the term big data. A typical flaw is to include both the data and its intended usage in the definition. We suggest that they should be separated. The term big data should cover data-related aspects, whereas a new term *big data insights* should be used when discussing data usage-related activities.
- The current definitions do not consider several important aspects of the big data phenomenon, such as security and privacy, or its disruptive nature. These are not characteristics of big data, but they are important factors of the big data phenomenon that both scholars and practitioners must consider. We suggest that a new definition for big data as a phenomenon should be developed.

In addition, this study bridges the knowledge between theory and practice. We have presented the history and the state-of-the-art of the definition of big data. This will help new researchers and practitioners to understand the different perspectives of big data, as well as the limitations of the current definitions. Therefore, we hope that this paper will also stimulate discussion about the terminology and help parties coming from different backgrounds to understand each other and communicate their

reasoning clearly.

5.2 Limitations

We recognize that an uncountable number of various definitions of big data exist in the "Internet jungle", e.g. in blog postings and discussion forums. However, due to limited resources, identifying and analyzing all or even most of them would be impossible, and therefore we have filtered blogs and forums out. Another limitation is that we have excluded all non-English language sources.

5.3 Suggestions for Further Studies

There are several possible topics for further studies, including the following. It is clear that there is a need to develop the terminology and taxonomy further (including related terms, such as big data analytics, big data phenomenon, and veracity) in order to create common understanding of the key concepts and their relationships in the area of big data. Another interesting research avenue would be to investigate the effects of big data on organizations' business models or decision-making processes, organizational structures, and culture.

6 References

Ackoff, R.L., 1989. From Data to Wisdom. Journal of Applied Systems Analysis 16, 3–9.

Akerkar, R., 2014. Analytics on Big Aviation Data: Turning Data Into Insights. International Journal of Computer Science and Applications 11, 116–127.

Altshuler, M. Y.; Aharony N.; Pentland A.; Elovici Y.; Cebrian, 2011. Stealing Reality: When Criminals Become Data Scientists (or Vice Versa). IEEE Intelligent Systems 26, 22–30.

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.

Anderson, C., 2008. The end of theory: The data deluge makes the scientific method obsolete. Wired Magazine, 23-Jun-2008.

Ashraf, J., Hussain, O.K., Hussain, F.K., 2015. Making sense from big RDF data: OUSAF for measuring ontology usage. Software: Practice and Experience 45, 1051–1071.

Balar, A., Malviya, N., Prasad, S., Gangurde, A., 2013. Forecasting consumer behavior with innovative value proposition for organizations using big data analytics. Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on 1–4.

Baro, E., Degoul, S., Beuscart, R., Chazard, E., 2015. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Research International 2015, 1–9.

Benjamins, V.R., 2014. Big Data: from Hype to Reality? Proceedings of the 4th

International Conference on Web Intelligence, Mining and Semantics (WIMS14) 2.

Berghel, H., 2013. Through the PRISM darkly. Computer 46, 86–90.

Bertolucci, J., 2013. Big Data: A Practical Definition. Informationweek - Online.

Blume, G., Scott, T., Pirog, M., 2014. Empirical Innovations in Policy Analysis. Policy Studies Journal 42, S33–S50.

boyd, danah, Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society 15, 662–679.

Cackett, D., 2013. Information Management and Big Data: A Reference Architecture. Oracle Corporation. Accessed 28th April 2016. http://www.oracle.com/technetwork/topics/entarch/articles/info-mgmt-big-data-ref-arch-1902853.pdf.

Chen, H., Chiang, R.H., Storey, V.C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly 36, 1165–1188.

Chen, M., Mao, S., Liu, Y., 2014. Big data: A survey. Mobile Networks and Applications 19, 171–209.

Collins, E., 2014. Intersection of the cloud and big data. IEEE Cloud Computing 1, 84–85.

Cox, M., Ellsworth, D., 1997. Application-controlled demand paging for out-of-core visualization. Proceedings of the 8th conference on Visualization'97 235–ff.

Cuzzocrea, A., Song, I.-Y., Davis, K.C., 2011. Analytics over large-scale multidimensional data: the big data revolution! Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP 101–104.

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. AIP Conference Proceedings 1644, 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.

DeloitteConsulting, 2012. The insight economy: Big data matters—except when it doesn't. Deloitte. Accessed 28th April 2016. http://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-us-ba-insight-economy-10012012.pdf.

Demchenko, Y., DeLaat, C., Membrey, P., 2014. Defining architecture components of the Big Data Ecosystem. Collaboration Technologies and Systems (CTS), 2014 International Conference on 104–112.

Demchenko, Y., Grosso, P., De Laat, C., Membrey, P., 2013. Addressing big data issues in scientific data infrastructure. Collaboration Technologies and Systems (CTS), 2013 International Conference on 48–55.

Demchenko, Y., Gruengard, E., Klous, S., 2014. Instructional Model for Building Effective Big Data Curricula for Online and Campus Education. Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on 935–941.

Demchenko, Y., Ngo, C., Laat, C. de, Membrey, P., Gordijenko, D., 2014. Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure, in: SecureDataManagement. Springer, pp. 76–94.

Diebold, F.X., 2003. Big Data'Dynamic factor models for macroeconomic measurement and forecasting, in: Advances EconomicsEconometricsTheory Applications, Eighth World CongressEconometricSociety(edited M. Dewatripont, LP HansenS.Turnovsky). pp. 115–122.

Earley, S., 2014. The Digital Transformation: Staying Competitive. IT Professional 16, 58–60.

Emani, C.K., Cullot, N., Nicolle, C., 2015. Understandable Big Data: A survey. Computer Science Review 17, 70–81.

EMC, 2012. Big Data-as-a-Service. EMC Solutions Group. Accessed 28th April 2016. http://www.emc.com/collateral/software/white-papers/h10839-big-data-as-a-service-perspt.pdf.

Fan, W., Bifet, A., 2013. Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter 14, 1–5.

Fan, W., Gordon, M.D., 2014. The power of social media analytics. Communications of the ACM 57, 74–81.

Ferrando-Llopis, R., Lopez-Berzosa, D., Mulligan, C., 2013. Advancing value creation and value capture in data-intensive contexts. Big Data, 2013 IEEE International Conference on 5–9.

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

Frankel, D.A., 2012. Big Data and Risk Management. Risk Management.

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35, 137–144.

Gantz, J., Reinsel, D., 2011. Extracting value from chaos (No. 1142), IDC iview. IDC. Accessed 28th April 2016. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

Gardner, E., 2013. The HIT approach to big data. Health data management 21, 34–36.

Gartner, 2012. Gartner IT Glossary - Big Data. Accessed 28th April 2016. http://www.gartner.com/it-glossary/big-data.

Gerhardt, B., Griffin, K., Klemann, R., 2012. Unlocking value in the fragmented world of big data analytics, Cisco Internet Business Solutions Group, June. Cisco Internet Business Solutions Group (IBSG).

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U., 2015. The rise of "big data" on cloud computing: review and open research issues.

Information Systems 47, 98–115.

Hu, H., Wen, Y., Chua, T.-S., Li, X., 2014. Toward scalable systems for big data analytics: A technology tutorial. Access, IEEE 2, 652–687.

Jacobs, A., 2009. The pathologies of big data. Communications of the ACM 52, 36–44.

Jin, X., Wah, B.W., Cheng, X., Wang, Y., 2015. Significance and challenges of big data research. Big Data Research 2, 59–64.

Kambatla, K., Kollias, G., Kumar, V., Grama, A., 2014. Trends in big data analytics. Journal of Parallel and Distributed Computing 74, 2561–2573.

Keen, J., Calinescu, R., Paige, R., Rooksby, J., 2013. Big data+ politics= open data: The case of health care data in England. Policy & Internet 5, 228–243.

Kim, G.-H., Trimi, S., Chung, J.-H., 2014. Big-data applications in the government sector. Communications of the ACM 57, 78–85.

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.

Lamont, J., 2012. Big data has big implications for knowledge management. KM World 21, 8–11.

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

Lee, J., Kao, H.-A., Yang, S., 2014. Service innovation and smart analytics for industry 4.0 and big data environment. Procedia CIRP 16, 3–8.

Li, T.Z., Wang, S.H., Ma, J., 2014. Study on Fair Definitions and Application Modes of Big Data, in: AppliedMechanics Materials. Trans Tech Publ, pp. 606–613.

Lin, P.P., 2014. What CPAs Need to Know about Big Data. The CPA Journal 84, 50.

Lu, R., Zhu, H., Liu, X., Liu, J.K., Shao, J., 2014. Toward efficient and privacy-preserving computing in big data era. Network, IEEE 28, 46–50.

Lycett, M., 2013. Datafication': Making sense of (big) data in a complex world. European Journal of Information Systems 22, 381–386.

Ma, M., Wang, P., Chu, C.-H., 2013. Data management for internet of things: challenges, approaches and opportunities. Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing 1144–1151.

Madden, S., 2012. From databases to big data. IEEE Internet Computing 16, 4–6.

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.

Mashey, J.R., 1998. Big Data... and the Next Wave of InfraStress. Accessed 28th April 2016. http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

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., Davenport, T.H., Patil, D., Barton, D., 2012. Big data: The Management Revolution. Harvard Business Review 90, 61–67.

Mehta, S., Pimplikar, R., Singh, A., Varshney, L.R., Visweswariah, K., 2013. Efficient multifaceted screening of job applicants. Proceedings of the 16th International Conference on Extending Database Technology 661–671.

Membrey, P., Chan, K.C., Demchenko, Y., 2013. A disk based stream oriented approach for storing big data. Collaboration Technologies and Systems (CTS), 2013 International Conference on 56–64.

Meng, L., Meng, L., 2014. Application of Big Data in Higher Education, in: 2ndInternationalConference TeachingComputationalScience. Atlantis Press.

Microsoft, 2012. The Big Bang: How the Big Data Explosion Is Changing the World. Accessed 28th April 2016. http://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/.

O'Leary, D.E., 2013. Artificial intelligence and big data. IEEE Intelligent Systems 28, 96–99.

Pandey, S., Tokekar, V., 2014. Prominence of mapreduce in big data processing. Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on 555–560.

Pospiech, M., Felden, C., 2013. A Descriptive Big Data Model Using Grounded Theory. Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on 878–885.

Provost, F., Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. Big Data 1, 51–59.

Rayport, J.F., Sviokla, J.J., 1995. Exploiting the virtual value chain. Harvard business review 73, 75.

Richards, G., 2014. Let us clarify what we mean by Big Data; We first need a clear definition. The Ottawa Citizen, 02-Aug-2014.

Sagiroglu, S., Sinanc, D., 2013. Big data: A review. Collaboration Technologies and Systems (CTS), 2013 International Conference on 42–47.

Schmarzo, B., 2013. Big Data: Understanding how data powers big business. John Wiley & Sons.

Schneider, R.D., 2012. Hadoop for dummies, Special Edition, John Willey & sons. John Wiley & Sons Canada, Ltd.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., Tufano, P., 2012. Analytics: The real-world use of big data, IBM Global Business Services, Somers. IBM. Accessed 28th April 2016. https://www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf.

Spiess, J., T'Joens, Y., Dragnea, R., Spencer, P., Philippart, L., 2014. Using big data to improve customer experience and business performance. Bell Labs Technical Journal 18, 3–17.

Stonebraker, M., Robertson, J., 2013. {Big Data is'Buzzword Du Jour;'CS Academics'Have the Best Job'}. Commun. ACM 56, 10–11.

TataConsultancyServices, 2013. The Emerging Big Returns on Big Data. Tata Consultancy Services. Accessed 28th April 2016. http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf.

TheIrishTimes, 2013. Get to grips with the streams of data you generate. Irish Times.

Tiefenbacher, K., Olbrich, S., 2015. Increasing the Value of Big Data Projects–Investigation of Industrial Success Stories. System Sciences (HICSS), 2015 48th Hawaii International Conference on 294–303.

Truyens, M., Van Eecke, P., 2014. Legal aspects of text mining. Computer Law & Security Review 30, 153–170.

Vossen, G., 2014. Big data as the new enabler in business and other intelligence. Vietnam Journal of Computer Science 1, 3–14.

Wang, W., Zhou, X., Zhang, B., Mu, J., 2013. Anomaly detection in big data from UWB radars. Security and Communication Networks 8, 2469–2475.

Ward, J.S., Barker, A., 2013. Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821.

Weiss, S.M., Indurkhya, N., 1998. Predictive data mining: a practical guide. Morgan Kaufmann.

Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., Celi, L.A., 2015. Big data in global health: improving health in low-and middle-income countries. Bulletin of the World Health Organization 93, 203–208.

Xin, N.Y., Ling, L.Y., 2013. How we could realize big data value. Instrumentation and Measurement, Sensor Network and Automation (IMSNA), 2013 2nd International Symposium on 425–427.

Zhang, J., Chen, Y., Li, T., 2013. Opportunities of innovation under challenges of big data. Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on 669–673.

Appendix 1 – Included Papers

Definitions of big data sorted by date. The Definition column contains either a direct quotation from the paper or essential parts of the given/referenced definition. The New perspective column indicates what new component or aspect the definition has added to the previous ones.

Authors	Date	Definition	New perspective
(Laney, 2001)	2001-Feb	"E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity, and variety."	volume, velocity, variety
(Jacobs, 2009)	2009-Aug	"data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time."	-
(Manyika et al., 2011)	2011-May	"Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."	analyze
(Gantz and Reinsel, 2011)	2011-Jun	"Big Data technologies describe a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis."	value, new architecture
(Cuzzocrea et al., 2011)	2011-Oct	"enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios"	-
(Microsoft, 2012)	2012-Feb	"Big data is the term increasingly used to describe the process of applying serious computing power – the latest in machine learning and artificial intelligence – to seriously massive and often highly complex sets of information."	computing power
(Lamont, 2012)	2012-Apr	Volume, variety, velocity	-
(Madden, 2012)	2012-May	"...it means data that's too big, too fast, or too hard for existing tools to process."	-
(Gartner, 2012)	2012-Jun	"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."	decision making
(Gerhardt et al., 2012)	2012-Jun	Volume, variety, velocity	-
(EMC, 2012)	2012-Jul	"Big data refers to new scale-out architecture that address these needs. [quickly processing more	-

Authors	Date	Definition	New perspective
		varied, more complex, and less structured data] Big data is fundamentally about massively distributed architectures and massively parallel processing, using commodity building blocks to manage and analyze the data."	
(Schneider, 2012)	2012-Sep	Volume, variety, velocity	-
(DeloitteConsulting, 2012)	2012-Oct	"Big data generally refers to datasets so large and complex they create significant challenges for traditional data management and analysis tools in practical timeframes."	practical timeframes
(Frankel, 2012)	2012-Oct	”the volumes of structured and unstructured data produced as a by-product of operating a company”	by-product
(McAfee et al., 2012)	2012-Oct	Volume, variety, velocity	-
(Schroeck et al., 2012)	2012-Oct	"...characterizing three dimensions of big data – “the three Vs:” volume, variety and velocity. And while they cover the key attributes of big data itself, we believe organizations need to consider an important fourth dimension: veracity."	veracity
(Chen et al., 2012)	2012-Dec	big data and big data analytics have been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies.	visualization
(Fan and Bifet, 2013)	2012-Dec	Volume, Velocity, Variety, Value, Variability	variability
(Cackett, 2013)	2013-Feb	Volume, velocity, variety, value	-
(Gardner, 2013)	2013-Mar	”Volume, velocity, variety”	-
(O’Leary, 2013)	2013-Mar	”Big data isn’t just volume, variety, and velocity, though; it’s volume, variety, and velocity at scale.”	-
(Provost and Fawcett, 2013)	2013-Mar	For this article, we will simply take big data to mean datasets that are too large for traditional data-processing systems and that therefore require new technologies.	-
(TataConsultancyServices,	2013-Mar	Volume, variety, velocity	-

Authors	Date	Definition	New perspective
2013)			
(Wang et al., 2013)	2013-Mar	" Big Data refers to large, diverse, complex, longitudinal, and distributed data sets generated from instruments, sensors, Internet transactions, e-mail, video, click streams, and other digital sources available today and in the future"	distributed data sets
(Demchenko et al., 2013)	2013-May	"...we intend to propose wider definition of Big Data as 5 Vs: Volume, Velocity, Variety and additionally Value and Veracity."	-
(Membrey et al., 2013)	2013-May	"Extensions to the (3V) model that take Value into account are then proposed and discussed. … However recording the data does not bring any value to the company. It only becomes valuable once that data is used or processed. ... High Value Data (HVD) is data that has a known benefit from its storage. ... Low Value Data (LVD) is data that is stored in the anticipation that value will be drawn from it in the future."	High & low value data
(Sagiroglu and Sinanc, 2013)	2013-May	Volume, variety, velocity	-
(Zhang et al., 2013)	2013-Jul	Volume, velocity, variety, value	-
(Bertolucci, 2013)	2013-Aug	Big data is about "building new analytic applications based on new types of data, in order to better serve your customers and drive a better competitive advantage."	competitive advantage
(Ward and Barker, 2013)	2013-Sep	"Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning."	-
(Stonebraker and Robertson, 2013)	2013-Sep	In summary, big data can mean big volume, big velocity, or big variety.	-
(Ferrando-Llopis et al., 2013)	2013-Oct	Volume, velocity, variety, veracity	-
(TheIrishTimes, 2013)	2013-Nov	"A simple definition is that it gives organisations insights into data which they don't already have and does that in a way that helps them improve their operational efficiency and helps them make better decisions."	-
(Vossen, 2014)	2013-Nov	Volume, velocity, variety, veracity	-

Authors	Date	Definition	New perspective
(Balar et al., 2013)	2013-Dec	Volume, variety, velocity	-
(Xin and Ling, 2013)	2013-Dec	Volume	-
(Pospiech and Felden, 2013)	2013-Dec	Volume, variety, velocity	-
(Chen et al., 2014)	2014-Jan	Volume, variety, velocity	-
(Spiess et al., 2014)	2014-Feb	Volume, variety, velocity	-
(Ashraf et al., 2015)	2014-Apr	Volume, velocity, variety, value, veracity	-
(Pandey and Tokekar, 2014)	2014-Apr	Volume, variety, velocity	-
(Blume et al., 2014)	2014-May	Volume	-
(Collins, 2014)	2014-May	Volume, variety, velocity	-
(Demchenko, DeLaat, et al., 2014)	2014-May	"Big Data (Data Intensive) Technologies are targeting to process high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) new data models (supporting all data states and stages during the whole data lifecycle) and new infrastructure services and tools that allow obtaining (and processing) data from a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices."	delivery to consumers
(Demchenko, Ngo, et al., 2014)	2014-May	Volume, velocity, variety, value, veracity	-
(Benjamins, 2014)	2014-Jun	Volume, variety, velocity	-
(Hu et al., 2014)	2014-Jun	Volume, velocity, variety, value	-

Authors	Date	Definition	New perspective
(Li et al., 2014)	2014-Jun	" Big Data refers to large, diverse, complex, longitudinal, and distributed data sets generated from instruments, sensors, Internet transactions, e-mail, video, click streams, and other digital sources available today and in the future"	-
(Lu et al., 2014)	2014-Jul	Volume, variety, velocity	-
(Meng and Meng, 2014)	2014-Jul	Volume, velocity, variety, value	-
(Richards, 2014)	2014-Aug	Big Data is commonly defined as data that cannot be processed by standard database systems.	-
(De Mauro et al., 2015)	2014-Sep	"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."	Information assets
(Lin, 2014)	2014-Nov	Volume, variety, velocity	-
(Akerkar, 2014)	2014-Dec	"Using big volume, big velocity, big variety data asset to extract value (insight and knowledge), further confirm veracity (quality and trustworthiness) of the original data and the acquired information, that demand cost-effective, novel forms of data and information processing for enhanced insight, decision making, and processes control. Additinally, those demands are supported by new data models and new infrastructure services and tools which are able to procure and process data from a variety of sources and deliver data in a variety of forms to several data and information consumers and devices."	-

Authors	Date	Definition	New perspective
(Demchenko, Gruengard, et al., 2014)	2014-Dec	"Big Data (Data Intensive) Technologies are targeting to process high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure highveracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) new data models (supporting all data states and stages during the whole data lifecycle) and new infrastructure services and tools that allow obtaining (and processing) data from a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices."	-
(Hashem et al., 2015)	2015-Jan	Volume, velocity, variety, value	-
(Tiefenbacher and Olbrich, 2015)	2015-Jan	Volume, variety, velocity (in), visibility, veracity, virtue (= value), velocity (out)	velocity in, velocity out
(Baro et al., 2015)	2015-Feb	"Volume: Log(n * p) ≥ 7" (n=statistical individuals, p=nbr of variables) Properties: "Great variety, High velocity, Challenge on veracity, Challenge on all aspects of the workflow, Challenge on computational methods, Challenge on extracting meaningful information, Challenge on sharing data, Challenge on finding human experts"	workflow
(Jin et al., 2015)	2015-Feb	Volume, velocity, variety, value, veracity	-
(Wyber et al., 2015)	2015-Mar	Volume, velocity, variety, veracity	-
(Gandomi and Haider, 2015)	2015-Apr	Volume, variety, velocity	-
(Emani et al., 2015)	2015-May	Volume, velocity, variety, veracity	-

Publication II

Ylijoki, Ossi and Porras, Jari
What Managers Think about Big Data

Reprinted with permission from
International Journal of Business Information Systems
Vol. 29(4), pp. 485-501, 2018.

What managers think about big data

Ossi Ylijoki* and Jari Porras

School of Business and Management,
Lappeenranta University of Technology,
Lahti, Finland
Email: ossi.ylijoki@phnet.fi
Email: jari.porras@lut.fi
*Corresponding author

Abstract: Digitisation progresses rapidly, producing vast amounts of big data. Companies can innovate their business models by using big data and related technologies. Some industries and companies are already on their way towards more data-driven businesses, but for most organisations this is an uncharted territory. The process is first and foremost a business transformation issue. Management leads the change and sets the pace; therefore the attitudes and intentions of executives towards big data are important in the transformation process. This survey concerns the behavioural intentions of Finnish executives with regard to big data. Building on a well-established technology acceptance model we explored the factors that explain the intentions. According to the results, executives intend to take actions that will promote the utilisation of big data. They have either experienced or expect big data to be beneficial to their business, especially with regard to current products or services, streamlining of processes and increasing customer understanding. In addition to the generally positive attitude towards big data, the results reveal significant differences between respondents with big data experience compared to the inexperienced ones. The role of IT management seems to play an important role in the differences.

Keywords: big data; behavioural intentions; business information systems; business transformation; business model; digitisation; Finland; innovation; management survey; management attitudes; technology acceptance model.

Reference to this paper should be made as follows: Ylijoki, O. and Porras, J. (2018) 'What managers think about big data', *Int. J. Business Information Systems*, Vol. 29, No. 4, pp.485–501.

Biographical notes: Ossi Ylijoki is a doctoral student at Lappeenranta University of Technology (LUT). He has 15+ years of experience in knowledge management, business intelligence and data warehousing in various expert and management positions in business life. His research interests are big data, information systems, and software engineering as a value driver.

Jari Porras is a Professor of Software Engineering (especially Distributed Systems) at the Lappeenranta University of Technology (LUT). He has conducted research on parallel and distributed computing, wireless and mobile systems and services as well as sustainable ICT. In recent years he has focused his research on human and sustainability aspects of software engineering. He is actively working in international projects and organisations.

1 Introduction

Technology enables innovations. Christensen (2013) distinguishes between incremental and disruptive innovations. Incremental innovations create value within current business models, typically by adding new features to existing products and services. This is what happens every day at incumbent companies. Disruptive innovation is a fundamentally different approach that often makes current solutions redundant. It potentially changes current business models and thus disrupts incumbent businesses. Examples like AirBnB or Über show that digital technologies, big data and advanced analytics are capable of enabling disruptive innovations. Disruptive innovations can damage current business models severely, like Über has done to traditional taxi services in several countries, or even make them redundant, as point-of-sale scanners and bar codes in retail stores did to the data analysis business model that was based on manual store audits in the late 1980's.

The disruption of many current business models is already in progress (Weill and Woerner 2015). Digitisation is changing the current business landscape rapidly in most industries. While the overall trend is clear – due to reasons of productivity reasons everything that can be digitised, will be – it is hard to predict the consequences at the organisation level. Organisations' have different capabilities and they live in different business ecosystems. However, it is certain that as digitisation proceeds, the amount of (big) data will explode. Exploiting data becomes a necessity due to the evidence supporting the claims that data adds value to business (see, e.g., Dehning et al., 2003; McAfee and Brynjolfsson, 2012). Moreover, trailblazers like Google, Amazon or Über have shown their value. New business models are emerging that are based on the value of data (Bucherer and Uckelmann, 2011; Chen et al., 2011; Leminen et al., 2012; Van't Spijker, 2014). Data has to be considered as a valuable asset.

As a new, more data-driven business environment emerges, the incumbents must adapt themselves to it. Many excellent papers about IT-enabled business transformation have been written (e.g., Loebbecke and Picot, 2015; Porter and Heppelmann, 2014; Venkatraman, 1994). These papers provide methods, frameworks and insights that help executives and scholars to understand various aspects of the transformation. In addition, technologies like social networking sites or service-oriented architectures can be used to support the change (Alkkiomäki, 2016; Hartono and Sheng, 2016). Managing the transformation, as well as disruptive innovations require solid leadership and vision (Dyer et al., 2008; Earley 2014). In this study we are interested in management's perceptions as the transformation and big data adoption starts from the leaders. Therefore, the purpose of this research is to:

- find out the behavioural intentions of business management with regard to big data
- explore the factors that explain these intentions.

Next we explain how the research was done. We present the theoretical model, connect it to the big data context and explain the operationalisation of the model. From the methods we move on to the results and analysis. After introducing and evaluating the relevance of the sample, we present our findings and the statistics we used for data analyses. The paper ends with a discussion of the findings.

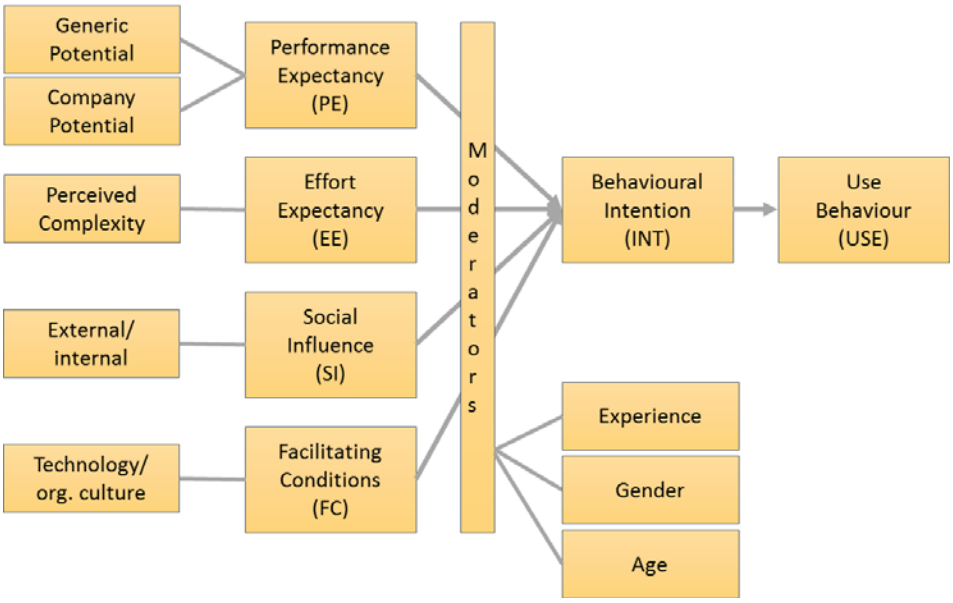
2 Research method

We chose survey as the instrument of data gathering. Surveys are commonly used to reflect the attitudes, opinions, and preferences of various audiences (Rea and Parker 2014). In order to find out management intentions regarding big data, we approached executives' in large Finnish private companies.

2.1 Theory

Technology adoption models are an established research area (Hu et al. 1999). Venkatesh et al. (2003) have presented a technology acceptance model, which we used as the theory in our research. Their research summarises the findings of the Technology Acceptance Model (Davis et al., 1989) and its several extensions to a "unified theory of acceptance and use of technology (UTAUT)" (Venkatesh et al., 2003). Several information systems studies have applied the UTAUT model (e.g., Eckhardt et al., 2009; Koivumäki et al., 2008; Tsourela and Roumeliotis, 2015; Verhoeven et al., 2010). The model has four principal constructs: performance expectancy, effort expectancy, social influence and facilitating conditions (Figure 1). The UTAUT model uses the constructs to explain the behavioural intentions of individuals. Drawing the logic from the behavioural sciences, the model then expects the intentions to turn into actual behaviour. In addition to the constructs, moderators like gender can affect the intentions.

Figure 1 Research model (see online version for colours)



Building on existing big data research, we applied the UTAUT constructs (Figure 1) to the context of our study as follows:

- *Performance expectancy* is "the degree to which an individual believes that using the system will help him or her to attain gains in job performance" (Venkatesh et al. 2003). With regard to big data this equals to how the respondents expect to benefit from big data. In general, the value proposition of big data is huge (Manyika et al., 2011). Ylijoki and Porras (2016) investigated big data case studies concluding that business benefits can be achieved. Therefore, we expected that both generic potential (including 'hype') and presumed company-specific benefits of big data will have an effect on the respondents' behavioural intentions.

  H1a The generic potential of big data has a positive effect on the respondent's behavioural intentions.

  H1b Company-specific expected benefits of big data have a positive effect on the respondent's behavioural intentions.

- *Effort expectancy* is "the degree of ease associated with the use of the system" (Venkatesh et al., 2003). Several studies have identified organisational side-effects, like changes in decision-making processes or cultural changes (Cai et al., 2014; Dutta and Bose, 2015; Phillips-Wren and Hoskisson, 2015). These kinds of side-effects require additional effort to manage changes and thus they affect the manager's workload. We asked the respondents about their personal feelings regarding the complexity in utilising big data.

  H2 Low perceived complexity in big data utilisation has a positive effect on the respondent's behavioural intentions.

- *Social influence* is "the degree to which an individual perceives that important others believe he or she should use the new system" (Venkatesh et al., 2003). In our research context, the important others included peers (i.e., other members of the management group), the board of directors and competitors. We were interested in whether the manager's social networks inside and outside the company influence her/his perception of big data.

  H3 Social pressure has a positive effect on the respondent's behavioural intentions.

- *Facilitating conditions* are "the degree to which an individual believes that an organisational and technical infrastructure exists to support use of the system" (Venkatesh et al., 2003). Numerous studies report technical challenges with big data (Dutta and Bose, 2015; Halamka, 2014; Krumeich et al., 2014; Mathew et al., 2015). Organisational factors like innovation capabilities (Dyer et al., 2008; Furr and Dyer, 2014) and data-oriented organisation culture (Amatriain, 2013; Mayer-Schönberger and Cukier, 2013) help companies to adopt big data.

  H4 Perceived technological and organisational capabilities have a positive effect on the respondent's behavioural intentions.

We followed Venkatesh et al., (2003) model and used the following moderators: age, gender and experience with big data (i.e. whether the respondent had participated in a big data project) to evaluate their possible effects on an individual's opinions. The moderators allowed us to filter the results in order to find differences among different groups of respondents.

2.2 Operationalisation

We created Likert scales (Likert, 1932) to test our hypotheses. Each of the constructs was tested by using four or more statements (see the Appendix for details). Each statement had five response alternatives: strongly disagree, disagree, neutral, agree, and strongly agree. In the analysis phase the verbal alternatives were replaced with numbers from 1 to 5, accordingly. The statements were developed by using the findings in current big data literature mentioned above, and propositions stated in Venkatesh.

The population of the research included executives of large Finnish companies. Small, 'traditional' companies are important for the economy in terms of turnover and employment as well, but at least for now they have very limited capabilities of taking advantage of big data. Therefore, we concluded companies listed in Helsinki Stock Exchange, as well as the largest private companies to be a representative set. This selection of incumbent companies covered all the large companies in terms of turnover (as well as a few smaller firms) and virtually all industries in the Finnish private sector. As the management group is the key decision-making organ in any company, we identified the members of the management groups of the companies. Moreover, because big data is an information systems -related matter, we decided to include IT managers even if they were not members of the company's management group.

As a sampling frame we used emails to all members of the population described above. Each member of the population had the same chance of being selected, which effectively resulted to a simple random sample. Thus, we believe that there is no bias regarding to under coverage or sampling.

The answers were collected by using a general-purpose on-line survey tool (Webropol) during a four-week time window in May–June 2016. The language of the survey was Finnish and the statements were translated into English for this article. The survey was pre-tested with a small number of people who did not belong to our population. The usage of the web-based tool allowed a rapid turnaround time and cost-effective data collection for this cross-sectional study.

3 Results and analyses

Over the years scholars have discussed how to analyse Likert responses. We used the following principles when analysing our survey data. Clason and Dormody (1994) define Likert-type items as single statements that use, e.g., five-point response alternatives. Likert-type items fall in the ordinal measurement level. Correspondingly, they defined Likert scale as a composite score of a set of several statements. Boone and Boone (2012) suggested that Likert scales should be analysed at the interval level of measurement. We used regression analysis and two-sample t-tests to analyse our Likert scales and ordinal level measures to describe results that were based on Likert-type items (statements). In addition, we report means of the statements, as this "has become current practice" (Rea and Parker 2014). The scales and statements that we used are listed in the Appendix.

The consistency of the constructs was measured with Cronbach's alpha (Cronbach 1951). On the basis of the test we removed one statement from the effort expectancy construct and one statement from the social influence construct in order to increase consistency. The resulting alpha values and the number of statements per construct are shown in Table 1.

Table 1 Cronbach's alpha values

Construct		*Items*	*Alpha*
INT	Behavioural intention of the respondent	5	0.883
PE1	Big data performance expectancy in general	4	0.796
PE2	Big data performance expectancy at the respondent's company	6	0.753
EE	Big data effort expectancy	4	0.623
SI	Social influence	5	0.773
FC	Big data facilitating conditions at the respondent's company	7	0.684

With regard to the sample, we are confident that we have a simple random sample, which represents the population. Finland Chamber of Commerce has studied the demographics of the management groups of Finnish companies (Linnainmaa and Turunen, 2014). Our respondents' gender and age distributions shown in Figure 3 aligned well with their statistics. Moreover, the respondent distribution by industry aligns with the population distribution (Figure 2).

3.1 Respondent profiles

We received 109 completed questionnaires from 82 companies (45% of the companies in the population). These companies represented 90 billion turnover (median 301 million euros) and 213,000 employees in 15 different industries (Figure 2). Most of the companies (63) employed more than 250 people. 34 % of the responses came from manufacturing companies, followed by wholesale and retail (12%), information and communication (12%), and finance and insurance (10%). The column named 'of population' displays the percentages calculated from the population.

Figure 2 Companies and industries of the respondents

	Companies	***Executives***
Population	184	1 104
Respondents	82	109
Response rate	45%	10%

Respondent's company size by personnel	***Companies***	***Turnover (m€)***	***Headcount***
Large (250+ persons)	62	87 900	210 350
Midsize (50-249 persons)	14	1 609	2 159
Small (1-49 persons)	6	19	90

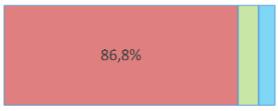
Respondent's industry	***% of respondents***	***of population***
C-Manufacturing	34%	37%
G-Wholesale and retail trade	12%	16%
J-Information and communication	12%	12%
K-Financial and insurance activities	10%	10%
O-Public administration, defense; compulsory social sec	7%	3%
H-Transportation and storage	6%	3%
F-Construction	5%	4%
Others	14%	16%

The respondents were members of the management group of their companies (86.8%), IT managers (7.5%), or line-of-business executives (5.7%). Three of the respondents did not

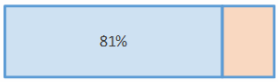
expose their role. Four out of five respondents were male. A typical respondent was a middle-aged man as the age distribution in Figure 3 shows.

Figure 3 Characteristics of the respondents (see online version for colours)

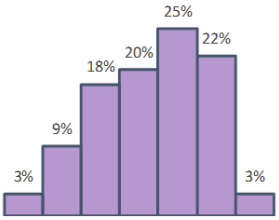
Respondent's role	*Count*	*%*
Member of the mgmt board	92	86,8%
IT management	8	7,5%
Other management	6	5,7%
Unknown	3	-



Gender	*Count*	*%*
Male	87	81%
Female	21	19%
Unknown	1	-



Age category	*Count*	*%*
-35	3	3%
36-40	10	9%
41-45	19	18%
46-50	21	20%
51-55	27	25%
56-60	23	22%
61-	3	3%
Unknown	3	-



3.2 Assessment of the model

Regression analysis of the mean values revealed that three of UTAUT's constructs (performance expectancy, effort expectancy and social influence) had a significant effect on the behavioural intentions of the executives. We did not find the effect of facilitating conditions statistically significant. Moreover, the analysis did not support the assumption that the generic potential of big data would influence the respondents. Therefore, the data did not support hypotheses H4 and H1a, whereas hypotheses H1b, H2 and H3 were supported. Table 2 presents the results of the regression analysis after the removal of the facilitating conditions and generic potential of big data. The overall F-significance was < 0.001, and all p-values were less than 0.05, i.e., the result is statistically significant.

Table 2 Results of regression analysis – significant constructs

Construct	*Coefficients*	*Std error*	*t stat*	*p-value*	*Lower 95%*	*Upper 95%*
Performance expectancy (H1b)	0.244	0.092	2.637	0.010	0.061	0.427
Effort expectancy (H2)	0.364	0.071	5.157	<0.001	0.224	0.505
Social influence (H3)	0.219	0.072	3.059	0.003	0.077	0.361

Regression analysis has some assumptions that must be met for the results to be reliable: linear relation between the response variable and predictors, normality, homoscedasticity, independence of errors, sufficient number of observations and absence of outliers. As the scatterplots of the residuals were linear and evenly varied, the linearity and

homoscedasticity assumptions were met. Due to simple random sample and sufficient number of observations the assumption of the sample size and independence of errors were met. Histograms of the residuals showed that the distribution was normal. Finally, we checked the number of outliers using standardised residuals and found that there were only very few of them. As a conclusion, we found that the assumptions hold.

We performed three separate two-sample t-tests assuming unequal variances to test the following moderators: gender, age and experience (Table 3). Mean age (49.6) was used to divide the respondents into two age groups. Experience in the context of this study was measured by asking whether the respondent had participated in a big data project or not. Contrary to the UTAUT suggestion, we did not find significant difference with regard to gender and age. With experience, however, the results were clear. Those having big data experience had significantly more intentions of acting.

Table 3 T-test results – moderator effects on behavioural intentions (INT)

	n	*Mean*	*Variance*	*p-value*
Big data experience	52	4.43	0.257	< 0.001
No big data experience	53	3.54	0.587	
Age ≤ 50	53	3.91	0.604	0.335
Age > 50	53	4.05	0.619	
Female	21	4.15	0.444	0.232
Male	87	3.95	0.619	

Note: H0: no difference in means, two-sided test.

Next, we analysed the means of the constructs by experience in order to find out the moderator effect on individual constructs. Table 4 summarises the responses by construct and experience. Means, standard deviations, standard errors and confidence intervals for each of the constructs are shown. None of the confidence intervals of the constructs overlap between the groups, i.e., experience moderates each of the constructs.

Table 4 Responses by construct and experience

Construct		*No experience (n = 53)*						*Experience (n = 52)*					
		Mean	*SD*	*SE*	*Cl 95%*			*Mean*	*SD*	*SE*	*Cl 95%*		
INT	Behavioural intention of the respondent	3.54	0.77	0.11	3.33	--	3.76	4.43	0.51	0.07	4.29	--	4.57
PE2	Big data performance expectancy	3.84	0.76	0.11	3.63	--	4.05	4.27	0.64	0.09	4.09	--	4.45
EE	Big data effort expectancy	2.94	0.79	0.11	2.72	--	3.16	3.64	0.75	0.10	3.43	--	3.85
SI	Social influence	2.96	0.98	0.14	2.69	--	3.23	3.52	0.78	0.11	3.30	--	3.73

On the basis of the regression analysis, t-tests and mean analysis described above, we drew the model shown in Figure 4. This statistically valid model explains 48.4% of the variance in the behavioural intentions of the executives. In the next section we look more

closely at the statements behind the constructs as well as the differences between the groups of experienced and non-experienced respondents.

Figure 4 The model explaining the big data intentions of executives (see online version for colours)

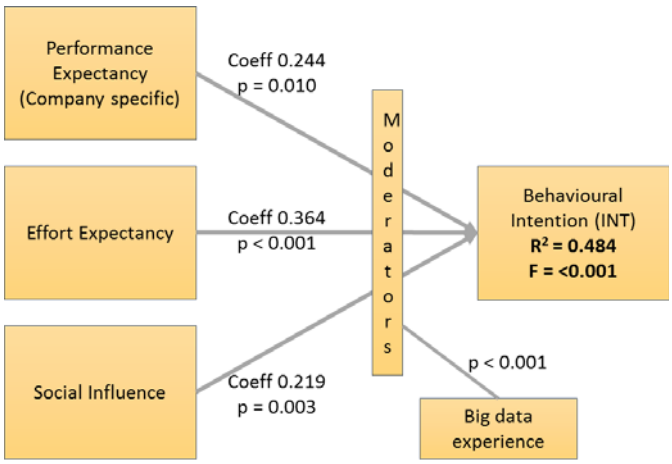
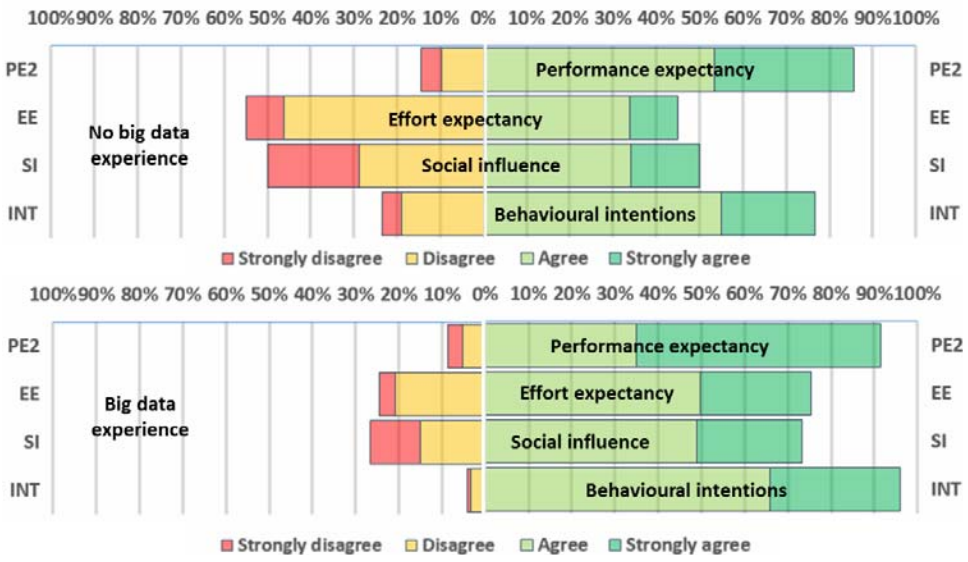


Figure 5 Respondents' perceptions regarding big data by experience (see online version for colours)



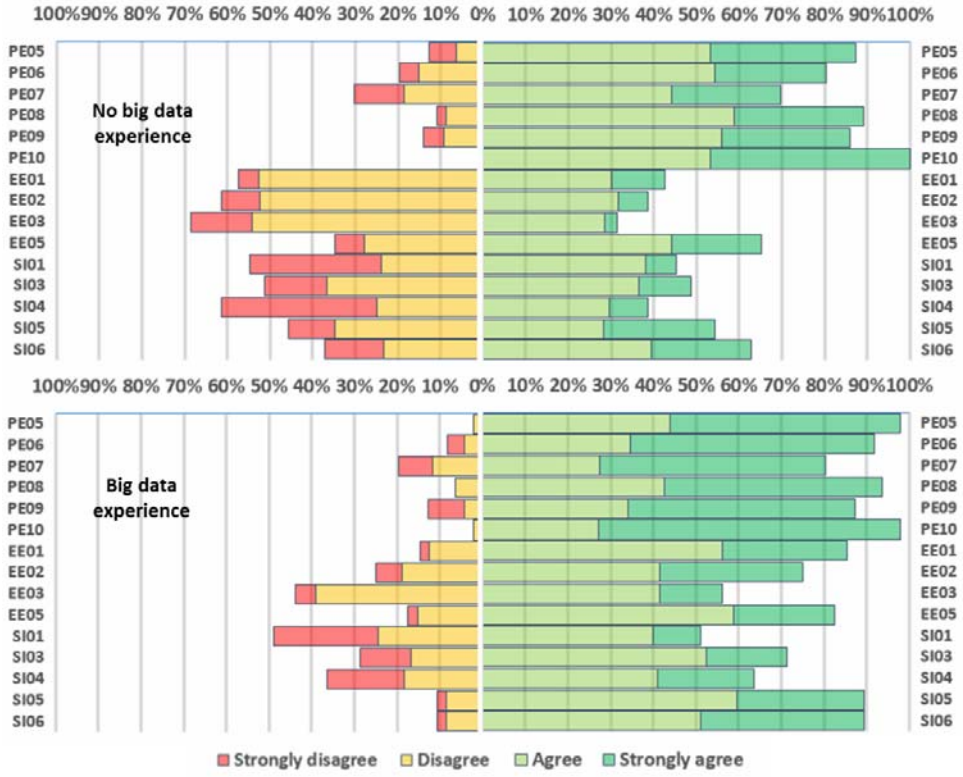
3.3 Factors explaining intentions

Figure 5 shows the perception differences of the respondents by experience. Neutral responses have been excluded, i.e., the graph includes those respondents who had clear opinions.

While both experienced and non-experienced respondents had positive performance expectations, the non-experienced ones considered the effort required to be high compared to the experienced ones. The overall perception (INT) of big data was very positive among those who had big data experiences – nearly every respondent would promote big data in their organisation.

Figure 6 presents the responses from both groups by statement. The performance expectancy scale (PEnn), which we used to measure company specific benefits contained six statements. All of the statements were related to the business model. A business model describes how an organisation creates, delivers, and captures value (Osterwalder and Pigneur, 2010). Both experienced and non-experienced respondents perceived the value of big data to be high with respect to the business model components.

Figure 6 Respondents' perceptions by statement (see online version for colours)



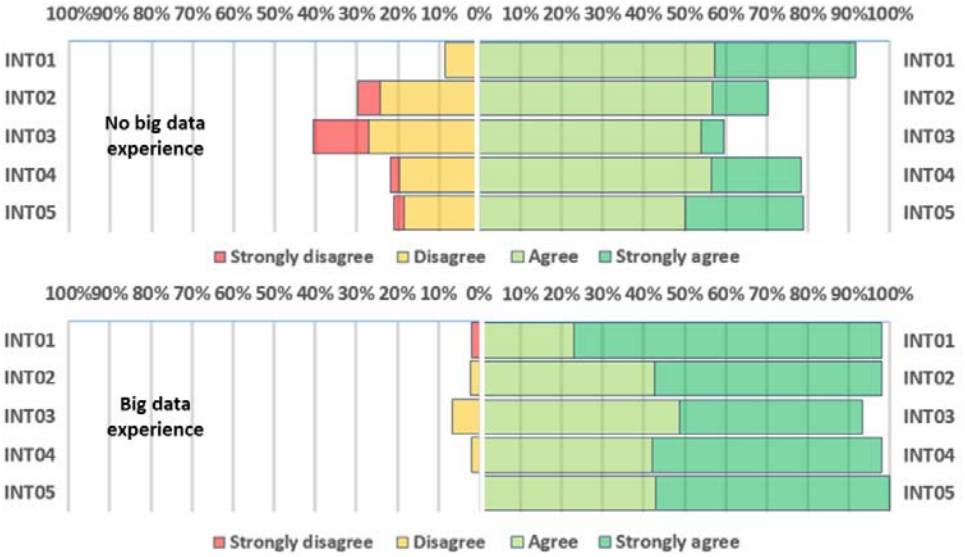
Our effort expectancy scale (EEnn) consisted of four statements that measured the personal effort required to utilise big data whereas in the social influence scale (SInn) five statements probed the impact of peers. See the Appendix for the verbal representation of each statement. Almost six out of ten respondents with no big data experience considered that utilising big data in their area of responsibility is difficult (EE01), whereas more than eight out of ten of the experienced ones disagreed with this. The same tendency applied to the perceived difficulty of measuring the impact of big data (EE02). Another noteworthy difference could be seen in the role of peers. The social influence of other management group members and IT management was high among the

experienced respondents. Nine out of ten of the experienced respondents agreed that both the management group and IT management consider utilisation of big data as an important matter (SI05, SI06).

3.4 Behavioural intentions

The intentions of the respondents were surveyed by using a Likert scale containing five statements (Appendix). The statements (INTnn) measured how active and willing the respondent was to promote big data in his/her organisation. Figure 7 presents detailed results.

Figure 7 Respondents' intentions to promote big data by experience (see online version for colours)



Almost all respondents in both groups agree that big data was an important factor in their area of responsibility (INT01). However, there was a clear difference between the groups. Four out of ten non-experienced respondents were not actively looking for information about big data whereas almost all of the experienced ones were. The experienced respondents indicate strong agreement towards actions that would promote the utilisation of big data in their organisation: all respondents either agreed or strongly agreed with the statement "I make decisions that drive the utilisation of big data in my company" (INT05).

4 Discussion

According to the results, executives have high expectations on big data. Both experienced and non-experienced respondents perceive big data as a vehicle to add value, e.g., to develop more efficient processes, add value to current products or services, and increase customer understanding. These expectations were reflected on their behavioural

intentions. Looking at the responses to individual statements (Figure 6), it is obvious that especially experienced respondents expected to gain business value from big data. This, along with the intention to promote big data, indicates that managers considered their big data projects successful. Moreover, it shows that the managers had recognised the connection between big data and the business models of their companies.

Our study offers further evidence to the claim that the business transformation process towards more data-driven business is already ongoing. Almost half of the respondents reported that they had participated in a big data project. Although those who were familiar with the survey domain may have been more probable to answer, the number of experienced respondents was high. Moreover, as also the non-experienced respondents had positive perceptions about big data, it seems that the pace of change is rapid.

Rapidly changing business landscape puts pressure on organisations. New competencies, both technical and organisational, must be acquired. Data and analytics –related skills can be at least partly outsourced, but organisational changes like adopting a more data-driven decision-making culture must be managed internally. Several studies have identified challenges in managing the transformation (Dutta and Bose, 2015; Phillips-Wren and Hoskisson, 2015; Shen and Varvel, 2013). Our survey did not focus on challenges, but as the experienced respondents seemed to be satisfied with their big data projects, either the challenges had been tackled or they had not surfaced yet. Or, companies have taken small steps towards big data and have thus avoided the issues.

Social influence and IT management seem to play an important role in big data adoption. The respondents who had big data experience considered management group members and IT management as important social influencers. On the other hand, non-experienced respondents expected big data to be a complex matter, which is of less importance to the management group and IT management. This is an interesting observation, which supports the perception that big data is much more than a technical exercise.

However, big data is related to many technical areas where IT management has expertise. The expert opinions of IT personnel affect the attitudes of executives. This claim was supported by high correlation (0.60) between the statements INT05 (“I make decisions that drive the utilisation of big data in my company”) and SI06 (“Our IT management considers the utilisation of big data an important issue”). This raises several questions, for example: Is the business model of these companies such that it does not benefit of utilisation of big data? Is the focus of IT management solely in technical aspects? Do these companies lack of data management capabilities? Do business executives recognise the impact of the IT management?

Contrary to the non-experienced respondents the experienced ones perceived big data as more valuable and relatively easy to adopt. However, the least potential was seen in innovating new products or services, i.e., the respondents did not fully conceive of the disruptive potential of big data in their own business context. This may indicate a small steps approach; executives take cautious, experimental steps towards big data, trying to avoid unnecessary risks.

Another possibility is that the companies lack capabilities that are required to identify disruptive innovations. The respondents represented incumbents, who had developed their processes, capabilities and culture over time to perform well in a less data-oriented environment. Incumbents arrange their operations as efficiently as possible according to

the principles defined by Taylor (1911). This kind of environment, however, is not an ideal seedbed for disruptive innovations. Dyer et al. (2008) and Furr and Dyer (2014) suggest that innovations require different capabilities. Sandberg and Aarikka-Stenros (2014) state that restrictive mind-set, lack of discovery skills and unsupportive organisational structures are main reasons that prohibit disruptive innovations. If this is the case, i.e., that incumbents have a shortage of capabilities related to disruptive innovations, it underlines the need for understanding and explaining the theoretical background of big data value creation processes.

4.1 Suggestion for further studies

The big picture seems to be that both experienced and non-experienced respondents perceive big data as an important factor for the business, but the non-experienced ones consider the effort required to be high. One reason to this may be the social influence of peers, especially the impact of IT management. This factor could be addressed in a further study. In addition, the study or studies might explore big data adopters regarding, e.g., value capturing and business model innovation; identification and management of organisational changes and potential problems; and the role of IT as social influencer, including the nature of communication between IT and business management.

4.2 Limitations of the study

Our sample (as well as the population) consisted of Finnish executives in the private sector. This should be taken into account when considering the generalisation of the results. Companies in other geographic areas or the public sector may face different conditions. Still, we believe that our results are significant in the private sector context. Many Western economies are in a similar situation as Finland with regard to digitisation and big data.

5 Conclusions

In this survey we explored management's behavioural intentions towards big data and the factors that explain their intentions. Based on an established technology acceptance model, we built a model that explained 48.4 % of the variance of the executives' behavioural intentions. The results showed that company-specific performance expectancy, effort expectancy and social influence had a positive effect on management attitudes regarding big data, which supported our hypotheses H1b, H2 and H3. However, we did not find support to generic big data performance expectancy (hypotheses H1a) or the effect of facilitating conditions (H4).

Moreover, we found that the adoption of big data is ongoing, as half of the respondents had big data experiences. The overall attitude towards big data was highly positive, especially among the experienced respondents. Ten out of ten of the experienced respondents either agreed or strongly agreed with the statement "I make decisions that drive the utilisation of big data in my company". The respondents saw process streamlining, product/service enhancements and increased customer understanding as the most potential big data application areas. In addition to new, detailed insights into how

executives perceive big data and what their intentions are, our research proposes a new research avenue.

References

Alkkiomäki, V. (2016) 'The role of service-oriented architecture as a part of the business model', *International Journal of Business Information Systems*, Vol. 21, No. 3, pp.368–387.

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 and Knowledge Management*, pp.2201–2208.

Boone, H.N. and Boone, D.A. (2012) 'Analyzing likert data', *Journal of Extension*, Vol. 50, No. 2, pp.1–5.

Bucherer, E. and Uckelmann, D. (2011) 'Business models for the internet of things', *Architecting the Internet of Things*, p.253.

Cai, H. et al. (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*, December, Vol. 33, pp.39–46.

Chen, Y. et al. (2011) 'Analytics ecosystem transformation: A force for business model innovation', *SRII Global Conference (SRII), 2011 Annual*, pp.11–20.

Christensen, C. (2013) '*The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, , 5th edition, Harvard Business Review Press.

Clason, D.L. and Dormody, T.J. (1994) 'Analyzing data measured by individual Likert-type items', *Journal of Agricultural Education*, Vol. 35, No. 4, p.4.

Cronbach, L.J. (1951) 'Coefficient alpha and the internal structure of tests', *Psychometrika*, Vol. 16, No. 3, pp.297–334.

Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. (1989) 'User acceptance of computer technology: a comparison of two theoretical models', *Management Science*, Vol. 35, No. 8, pp.982–1003.

Dehning, B., Richardson, V.J. and Zmud, R.W. (2003) 'The value relevance of announcements of transformational information technology investments', *Mis. Quarterly*, Vol. 27, No. 4, pp.637–656.

Dutta, D. and Bose, I. (2015) 'Managing a big data project: the case of Ramco Cements Limited', *International Journal of Production Economics*, Vol. 165, No. 4, pp.293–306.

Dyer, J., Gregersen, H. and Christensen, C. (2008) 'Entrepreneur behaviors, opportunity recognition, and the origins of innovative ventures', *Strategic Entrepreneurship Journal*, Vol. 2, No. 4, pp.317–338.

Earley, S. (2014) 'The digital transformation: staying competitive', *IT Professional*, Vol. 16, No. 2, pp.58–60.

Eckhardt, A., Laumer, S. and Weitzel, T. (2009) 'Who influences whom? Analyzing workplace referents' social influence on IT adoption and non-adoption', *Journal of Information Technology*, Vol. 24, No. 1, pp.11–24.

Furr, N. and Dyer, J. (2014) *The Innovator's Method*, J. Nathan & Dyer Furr, ed., Harvard Business Review Press.

Halamka, J.D. (2014) 'Early experiences with big data at an academic medical center', *Health Affairs*, Vol. 33, No. 7, pp.1132–1138.

Hartono, R. and Sheng, M.L. (2016) 'Knowledge sharing and firm performance: the role of social networking site and innovation capability', *Technology Analysis & Strategic Management*, Vol. 28, No. 3, pp.335–347.

Hu, P.J. et al. (1999) 'Examining the technology acceptance model using physician acceptance of telemedicine technology', *Journal of Management Information Systems*, Vol. 16, No. 2, pp.91–112.

Koivumäki, T., Ristola, A. and Kesti, M. (2008) 'The perceptions towards mobile services: an empirical analysis of the role of use facilitators', *Personal and Ubiquitous Computing*, Vol. 12, No. 1, pp.67–75.

Krumeich, J. et al. (2014) 'Towards planning and control of business processes based on event-based predictions', in *Business Information Systems*, pp.38–49, Springer.

Leminen, S. et al. (2012) 'Towards iot ecosystems and business models', in *Internet of Things, Smart Spaces, and Next Generation Networking*, pp.15–26, Springer.

Likert, R. (1932) 'A technique for the measurement of attitudes', *Archives of Psychology*.

Linnainmaa, L. and Turunen, A. (2014) *Keskuskauppakamarin naisjohtajaselvitys* [online] http://kauppakamari.fi/wp-content/uploads/2014/11/keskuskauppakamarin-naisjohtajaselvitys-2014.pdf (accessed 28 March 2017).

Loebbecke, C. and Picot, A. (2015) 'Reflections on societal and business model transformation arising from digitization and big data analytics: a research agenda', *The Journal of Strategic Information Systems*, Vol. 24, No. 3, pp.149–157.

Manyika, J. et al. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, in J. Manyika and M. Chui (Eds.), McKinsey Global Institute [online] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed 28 March 2017).

Mathew, P.A. et al. (2015) 'Big-data for building energy performance: lessons from assembling a very large national database of building energy use', *Applied Energy*, Vol. 140, pp.85–93.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, V. Mayer-Schönberger and K. Cukier (Eds.), Houghton Mifflin Harcourt.

McAfee, A. and Brynjolfsson, E. (2012) 'Big data: the management revolution', *Harvard Business Review*, Vol. 90, No. 10, pp.61–67.

Osterwalder, A. and Pigneur, Y. (2010) *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*, John Wiley & Sons, Hoboken, New Jersey.

Phillips-Wren, G. and Hoskisson, A. (2015) 'An analytical journey towards big data', *Journal of Decision Systems*, Vol. 24, No. 1, pp.87–102.

Porter, M.E. and Heppelmann, J.E. (2014) 'How smart, connected products are transforming competition', *Harvard Business Review*, Vol. 92, No. 11, pp.11–64.

Rea, L.M. and Parker, R.A. (2014) *Designing and conducting survey research: A Comprehensive Guide*, John Wiley & Sons.

Sandberg, B. and Aarikka-Stenroos, L. (2014) 'What makes it so difficult? A systematic review on barriers to radical innovation', *Industrial Marketing Management*, Vol. 43, No. 8, pp.1293–1305.

Shen, Y. and Varvel, V.E. (2013) 'Developing data management services at the Johns Hopkins University', *The Journal of Academic Librarianship*, Vol. 39, No. 6, pp.552–557.

Taylor, F.W. (1911) *The Principles of Scientific Management*, JSTOR.

Tsourela, M. and Roumeliotis, M. (2015) 'The moderating role of technology readiness, gender, and sex in consumer acceptance and actual use of Technology-based services', *The Journal of High Technology Management Research*, Vol. 26, No. 2, pp.124–136.

Van't Spijker, A. (2014) *The New Oil: Using Innovative Business Models to Turn Data Into Profit*, Technics Publications, Basking Ridge, New Jersey.

Weill, P. and Woerner, S.L. (2015) 'Thriving in an increasingly digital ecosystem', *MIT Sloan Management Review*, Vol. 56, No. 4, pp.27–34.

Venkatesh, V. et al. (2003) 'User acceptance of information technology: toward a unified view', *MIS Quarterly*, Vol. 27, No. 3, pp.425–478.

Venkatraman, N. (1994) 'IT-enabled business transformation: from automation to business scope redefinition', *Sloan Management Review*, Vol. 35, No. 2, pp.73–87.

Verhoeven, J.C., Heerwegh, D. and De Wit, K. (2010) 'Information and communication technologies in the life of university freshmen: an analysis of change', *Computers & Education*, Vol. 55, No. 1, pp.53–66.

Ylijoki, O. and Porras, J. (2016) 'Conceptualizing big data: analysis of case studies', *Intelligent Systems in Accounting, Finance and Management*, Vol. 23, No. 4, pp.295–310, DOI: 10.1002/isaf.1393.

Appendix

Table A1 presents the means, standard deviations, standard errors and confidence intervals for each construct used in the survey. Constructs PE1 and FC were excluded from the final model due to statistical insignificance in regression analysis.

Table A1 All respondents (N = 109)

		Mean	*SD*	*SE*	*Cl 95%*		
INT	Behavioural intention of the respondent	3.99	0.78	0.07	3.84	--	4.14
(PE1)	(Big data performance expectancy in general)	4.13	0.77	0.07	3.98		4.27
PE2	Big data performance expectancy at the respondent's company	4.07	0.73	0.07	3.93		4.21
EE	Big data effort expectancy	3.31	0.84	0.08	3.15	--	3.47
SI	Social influence	3.26	0.92	0.09	3.08	--	3.43
(FC)	(Big data facilitating conditions)	3.37	0.65	0.06	3.25	--	3.49

Table A2 presents the differences between respondents who had big data experience and those who did not. Mean values and their difference for each of the statements as well as for each of the constructs are given. Statements SI02 and EE04 were discarded based on Cronbach's alpha tests.

Table A2 Respondents grouped by big data experience

		No (N = 53)	*Yes (N = 52)*	*Diff*
INT	*Behavioural intention of the respondent*	*3.54*	*4.43*	0.89
INT01	Big data is insignificant in my area of responsibility. (*)	4.04	4.69	0.65
INT02	I influence actively matters that enable utilisation of big data in my area of responsibility.	3.34	4.37	1.03
INT03	I actively look for information regarding big data.	3.08	4.13	1.05
INT04	I have ideated, how to utilise big data in my area of responsibility.	3.66	4.46	0.80
INT05	I make decisions that drive the utilisation of big data in my company.	3.60	4.48	0.88
PE2	*Big data performance expectancy at the respondent's company*	*3.84*	*4.27*	0.43
PE05	With big data we can add value to our current products/services.	3.91	4.44	0.53
PE06	Big data does not help us to innovate new pricing models. (*)	3.72	4.29	0.57
PE07	We cannot innovate new products/services with big data. (*)	3.43	4.04	0.61
PE08	Big data helps us to develop more efficient processes.	3.92	4.25	0.33
PE09	Big data does not help us to utilise our resources better. (*)	3.79	4.08	0.29
PE10	Big data helps us to better understand our customers.	4.25	4.54	0.29
SI	*Social influence*	*2.96*	*3.52*	0.56
SI01	Our rivals have created services/products based on big data that threaten our business.	2.74	2.90	0.16
(SI02)	(Our rivals are ahead us with regard to the utilisation of big data.)			
SI03	Rivals from other industries that utilise big data penetrate to our industry.	2.96	3.40	0.44
SI04	Our board of directors requires us to utilise big data.	2.58	3.27	0.69
SI05	Our management group considers the utilisation of big data as an important issue.	3.21	3.96	0.75
SI06	Our IT management considers the utilisation of big data as an important issue.	3.29	4.04	0.75
EE	*Big data effort expectancy*	*2.94*	*3.64*	0.70
EE01	Utilising big data in my area of responsibility is difficult. (*)	2.94	3.90	0.96
EE02	The impact of big data is difficult to measure in my area of responsibility. (*)	2.78	3.71	0.93
EE03	A big data project produces quick results in my area of responsibility.	2.68	3.17	0.49
(EE04)	(Utilising big data requires me to learn new management practices.)			
EE05	People in my area of responsibility are capable to utilise big data.	3.36	3.77	0.41

Note: (*) = reverse scale

Publication III

Ylijoki, Ossi and Porras, Jari
Conceptualizing Big Data: Analysis of Case Studies

Reprinted with permission from
Intelligent Systems in Accounting, Finance and Management
Vol. 23(4), pp. 295-310, 2016

INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1393

CONCEPTUALIZING BIG DATA: ANALYSIS OF CASE STUDIES

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

SUMMARY

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. Copyright © 2016 John Wiley & Sons, Ltd.

1. INTRODUCTION

Today, new digital technologies produce vast amounts of various types of data (Gantz & 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) and Mayer-Schönberger and Cukier (2013), claim that big data have potentially huge effects on many industries. Technology and data drives change, and as Dehning, Richardson, and Zmud (2003) and Sainio (2005), for example, 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 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 kinds of examples, as well as some previous studies (e.g. Porter

* Correspondence to: School of Business and Management, Lappeenranta University of Technology, Lappeenranta, Finland. E-mail: ossi.ylijoki@phnet.fi

[1]For example, 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; Volvo: http://www.wired.com/2015/02/volvo-will-test-self-driving-cars-real-customers-2017/.

& Millar, 1985; Dehning *et al.,* 2003; McAfee & Brynjolfsson, 2012) indicate that companies utilizing data heavily gain a competitive advantage over their less data-driven rivals. However, the data-driven approach is still a new paradigm for most organizations (Shen & 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, Greco, & Grimaldi, 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 analyse 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 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 multidisciplinary 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

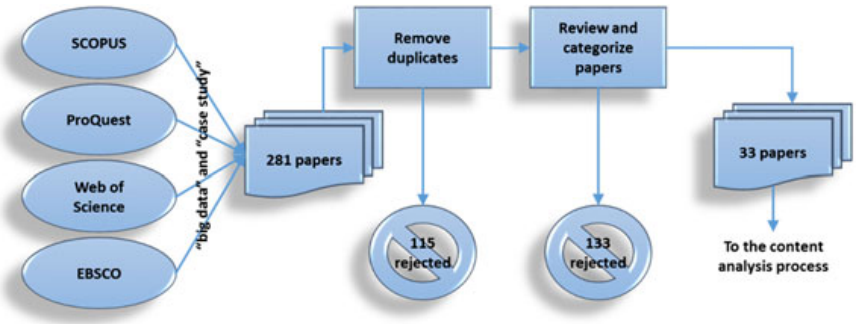


Figure 1. Search process for big data case studies

DOI: 10.1002/isaf

focus on big data, various publications in many research domains discuss the topic. Big data is an emerging, multidisciplinary 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; that is, 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.

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, for example, developing 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 an additional 13 papers for various reasons: the paper contained a hypothetical case (three studies), we could not access or find the paper (nine), or the paper was in Spanish (one).

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, Otto, & Österle, 2014; Kowalczyk & Buxmann, 2014; Wehn & Evers, 2015). Appendix A lists the papers and provides a short contextual description of each paper. Next, we analysed 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 analysed the results of the text-mining phase and formulated a set of guidelines.

2.2. Characteristics of the Case Studies

The cases found represented different application domains, from education to business, and from health care to entertainment. This indicates that big data affects every aspect of life. Table I lists the number of cases categorized by the ISIC classification of the UN (United Nations, 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 health-care studies represented the highest numbers of cases (eight and six respectively). Several industries were also well represented with four or five cases each: manufacturing, retail, finance and information-related cases.

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.

As with the application area, the geographical distribution of the cases was also 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.

DOI: 10.1002/isaf

Table I. Big data case studies by application area

	Application area (categories adopted from United Nations (2008))	No. 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 defence; 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 services-producing activities of households for own use	—
U	Activities of extraterritorial organizations and bodies	—

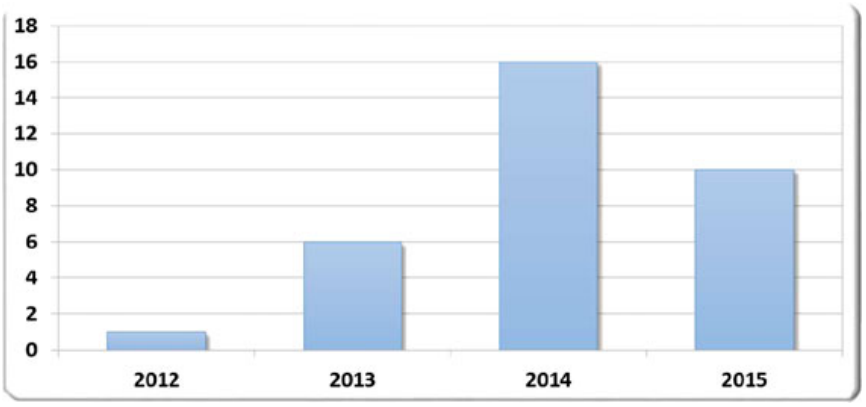


Figure 2. Number of big data case study articles by the publishing year

Appendix A 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 (e.g. Berelson, 1952; Holsti, 1969; Krippendorf, 1989). Weber (1990) defines content analysis as a repeatable,

DOI: 10.1002/isaf

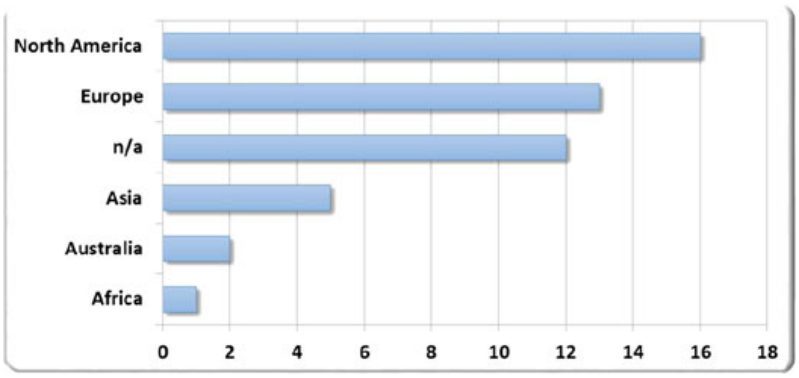


Figure 3. Geographical distribution of the big data cases

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. Lewis, Zamith, & Hermida, 2013; Hu, Ge, & Hou, 2014; 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 predefined 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 the 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 the Stanford part-of-speech tagger (Toutanova, Klein, Manning, & Singer, 2003) for tagging and lemmatization of words; that is, it recognizes parts of speech (such as nouns, verbs and adjectives) and converts words to their base format. This, combined with 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. For example, 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.

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

DOI: 10.1002/isaf

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 & 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. Different colours are used to distinguish the themes. We labelled the themes based on the following. First, based on the virtual value creation process (Rayport & 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 information and communications technology (ICT)- and data-related, technical themes. Then we 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.

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

- **Decision-making** (red colour in the map—see online image). Several studies discussed enhancing the decision-making processes, enabling data-driven decision-making, or providing actionable insights to managers (Bärenfänger *et al.,* 2014; Krumeich, Jacobi, Werth, & Loos, 2014; Dutta & Bose, 2015). Several studies (Cai *et al.,* 2014; Tao, Corcoran, Mateo-Babiano, & Rohde, 2014; Kalakou, Psaraki-Kalouptsidi, & Moura, 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 & 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 & Varvel, 2013; Dutta & Bose, 2015), missing analytics strategy, and lack of leadership (Phillips-Wren & Hoskisson, 2015).
- **Innovation** (blue—see online image). Big data was seen as an enabler for data-driven innovation and faster innovation cycles (Amatriain, 2013; Jetzek, Avital, & Bjorn-Andersen, 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—see online image). 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, Tilson, Tilson, & Haas, 2014; Jetzek *et al.,* 2014; Mathew *et al.,* 2015), business transformation (Prescott, 2014), increased revenue and customer satisfaction (Dutta & Bose, 2015), better transparency over operations (Bärenfänger *et al.,* 2014), generating value by secondary use of data (Bettencourt-Silva *et al.,* 2015)

[3]A few notes that clarify the interpretation of the map. When plotting a map, KH Coder uses the method explained in Fruchterman and Reingold (1991). This algorithm may plot nodes side by side, but unlike multidimensional maps, for example, 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 Fig. 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 colour coding indicates the communities (sub-graphs) that are relatively close to each other. KH Coder offers several methods for indicating patterns. We used the modularity method defined in Clauset, Newman, and Moore (2004). This method builds on the principle that there are many edges within the communities and only a few between them.

DOI: 10.1002/isaf

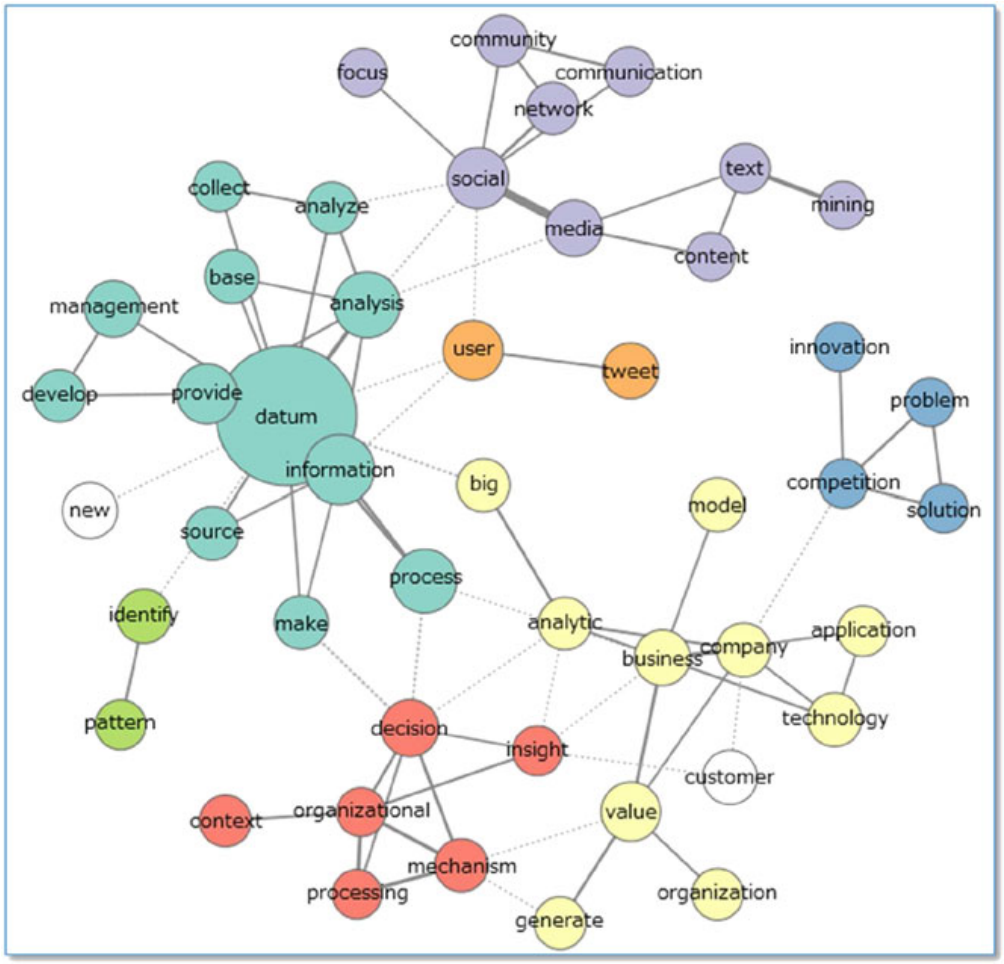


Figure 4. Co-occurrence map of the terms in big data case study articles

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—see online image) 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 & Bose, 2015). Moreover, the case studies pointed out additional aspects that need to be addressed, such as data inconsistencies and poor data

DOI: 10.1002/isaf

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 & Walton, 2014; Stephansen & 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 multifaceted issue: in addition to new technology, organizations need new talent, both business oriented and technology skilled (Shen & Varvel, 2013; Phillips-Wren & Hoskisson, 2015; Prinsloo *et al.,* 2015).

- **New data sources** (purple—see online image). In several cases organizations utilized data from outside their own organization, such as Facebook and Twitter data (He, Zha, & Li, 2013), blog texts and user reviews (Marine-Roig & Clavé, 2015), or data collected from mobile apps (O'Leary, 2013a; Papenfuss, Phelps, Fulton, & Venturelli, 2015). They had been able to extract value from these external sources. The data are freely available, but require quite a lot of processing, as described in Marine-Roig and Clavé (2015) for example.

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 arose 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 interrelated. 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 Section 5. The data management theme interrelates also with the 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 Section 1, 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 its business based on the data it collects, and boosts its sales by making customer-specific, data-driven recommendations. One case study (Prescott, 2014) reported a *business transformation process* leading to regaining a competitive advantage. There was also one case where a company-wide data-driven approach was taken (Dutta & Bose, 2015). In this case, a large steel manufacturer reshaped its processes and functions to take advantage of data, which resulted in significant business benefits. On the other hand, it 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 & Chen, 2014; Hu *et al.,* 2014) had

DOI: 10.1002/isaf

analysed social media data in order to identify signs or clues of, for example, rising 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 & 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 II synthesizes the findings of our analysis. The final 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.

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, Jones, & Sidorova, 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 not only to understand but also to support these changes. Managing the change and the organizational side effects requires training and new managerial skills.

Table II. 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.	Ciulla *et al.* (2012); Amatriain (2013); Jetzek *et al.* (2014); Martinez and Walton (2014)
Business value	• Look for value in various directions; experiment with the data. • Enable business transformation with the data. • Consider secondary usage of the data.	Amatriain (2013); O'Leary (2013a); Prescott (2014); Bettencourt-Silva *et al.* (2015); Dutta and Bose (2015)
Data management	• Expect to face technical and data-related challenges. • Plan for security.	Shen and Varvel (2013); Halamka (2014); Krumeich *et al.* (2014); Dutta and Bose (2015); Prinsloo *et al.* (2015)
New data sources	• Experiment with new data types. • Consider potential privacy issues.	He *et al.* (2013); Yu *et al.* (2014); Marine-Roig and Clavé (2015)

DOI: 10.1002/isaf

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 its 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 & 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 emissions 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 into 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 computerized content analysis suggested that appropriate software tools are efficient and cost-effective (compared with 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 is 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 & 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 analysed 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 Brynjolfsson (2012), has shown that data-driven

DOI: 10.1002/isaf

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 on 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 also – and more importantly – influence the organizational culture, decision-making processes and management functions. Previous studies discuss many of these challenges at a general level. Based on current big data implementations as described in peer-reviewed literature, our study adds insights at a 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 II) 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 study 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 at a detailed enough level to answer the *how* question. This opens new research avenues. We point out some of these avenues shortly.

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?

APPENDIX A: BIG DATA CASE STUDY ARTICLES

Table A. I is a summary of the case study articles we analysed. The table contains 33 different studies and 49 big data cases (due to three multi-case studies) representing five continents. We identified

DOI: 10.1002/isaf

A.1 Big data case study articles

Reference	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 five large European companies from different industries		Manufacturing (2) Wholesale and retail Electricity 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 analysing 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 checkpoint times, using Lisbon airport as the case	H	Transportation and storage	Portugal
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 K G I H	Information and communication (2) Financial and insurance activities (3) Wholesale and retail (2) Accommodation and food service activities Transportation and storage (2)	n/a

(*Continues*)

DOI: 10.1002/isaf

A.1 (Continued)

Reference	Context	Application area (ISIC)		Country
		Q	Human health and social work activities	
		C	Manufacturing	
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 USA; 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)	Analysing behavioural 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
Prescott (2014)	Nielsen regaining their competitive advantage by using data and analytics	H	Transportation and storage	USA
Prinsloo *et al.* (2015)	Unifying and analysing 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, two cases	F	Construction	UK, Netherlands
Yu *et al.* (2014)	Text mining (topic modelling) 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

DOI: 10.1002/isaf

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 (United Nations, 2008). Country is the origin of the organization subject to the study.

REFERENCES

Amatriain X. 2013. Beyond data: from user information to business value through personalized recommendations and consumer science. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM: New York, NY; 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. In Proceedings of the 47th Hawaii International Conference on System Sciences. IEEE Computer Society: Los Alamitos, CA; 3728–3737.

Berelson B. 1952. Content Analysis in Communication Research. US Free Press: New York.

Bettencourt-Silva JH, Clark J, Cooper CS, Mills R, Rayward-Smith VJ, 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 AS, 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: New York.

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 ME, Moore C. 2004. Finding community structure in very large networks. *Physical Review E* **70**: 1–6.

Crampton JW, Graham M, Poorthuis A, Shelton T, Stephens M, Wilson MW, 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: Boston, Massachussets.

De Mauro A, Greco M, Grimaldi M. 2015. What is big data? A consensual definition and a review of key research topics. In AIP Conference Proceedings 1644. AIP Publishing: Madrid, Spain; 97–104.

Dehning B, Richardson VJ, Zmud RW. 2003. The value relevance of announcements of transformational information technology investments. *MIS Quarterly* **27**: 637–656.

Dobson G, Tilson D, Tilson V, Haas CE. 2014. Quantitative case study: use of pharmacy patient information systems to improve operational efficiency. In Proceedings of the 47th Hawaii International Conference on System Sciences. IEEE Computer Society: Los Alamitos, CA; 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.

Fang S, Xu LD, 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. *IEEE Transactions on Industrial Informatics* **10**: 1596–1605.

Fruchterman TM, Reingold EM. 1991. Graph drawing by force-directed placement. *Software—Practice & Experience* **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. IDC Iview, IDC 1142. https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (accessed 22 April 2016).

DOI: 10.1002/isaf

Halamka JD. 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 OR. 1969. Content Analysis for the Social Sciences and Humanities. Addison-Wesley: Reading, MA.
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 MC, 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, Ver. 2.3. EBSE Technical Report EBSE-2007-01. Keele University/Durham Unversity.
Kolowitz BJ, Lauro GR, 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 Encyclopedia 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 Business Information Systems, Abramowicz W, Kokkinaki A (eds). Springer International Publishing: Switzerland; 38–49.
Laney D. 2001. 3D data management: controlling data volume, velocity and variety. META Group Research Note 6. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 22 April 2016).
Lewis SC, 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 AH. 2011. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute.
Marine-Roig E, Clavé SA. 2015. Tourism analytics with massive user-generated content: a case study of Barcelona. *Journal of Destination Marketing & Management* **4**(3): 162–172.
Martinez MG, Walton B. 2014. The wisdom of crowds: the potential of online communities as a tool for data analysis. *Technovation* **34**: 203–214.
Mathew PA, Dunn LN, Sohn MD, 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 DE. 2013a. Exploiting big data from mobile device sensor-based apps: challenges and benefits. *MIS Quarterly Executive* **12**: 179–187.
O'Leary DE. 2013b. Big Data, the Internet of Things and Internet of Signs. *Intelligent Systems in Accounting, Finance and Management* **20**: 53–65.
Osgood CE. 1959. The representational model and relevant research methods. In Trends in Content Analysis, de Sola PI (ed). University of Illinois Press: Urbana, IL; 33–88.
Papenfuss JT, Phelps N, Fulton D, Venturelli PA. 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 ME, Millar VE. 1985. How information gives you competitive advantage. *Harvard Business Review* **63**(4): 149–160.
Prescott ME. 2014. Big data and competitive advantage at Nielsen. *Management Decision* **52**: 573–601.

DOI: 10.1002/isaf

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**(1). http://www.irrodl.org/index.php/irrodl/article/view/1948/3203 (accessed 22 April 2016).

Rayport JF, Sviokla JJ. 1995. Exploiting the virtual value chain. *Harvard Business Review* **73**(6): 75–85.

Ryan GW, Bernard HR. 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: Lappeenranta, Finland.

Shen Y, Varvel VE. 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 HC, 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 CD, Singer Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In NAACL'03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologyvol. 1.. Association for Computational Linguistics: Stroudsburg, PA; 173–180.

United Nations. 2008. International Standard Industrial Classification of All Economic Activities (ISIC), Rev.4. United Nations: New York.

Venkatraman N. 1994. IT-enabled business transformation: from automation to business scope redefinition. *Sloan Management Review* **35**: 73–87.

Wamba SF, 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 RP. 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 WJ, Sirotkin K. 1992. The automatic identification of stop words. *Journal of Information Science* **18**: 45–55.

Yang Y, Pedersen JO. 1997. A comparative study on feature selection in text categorization. In ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann: San Francisco, CA; 412–420.

Yang Y, Wilbur J. 1996. Using corpus statistics to remove redundant words in text categorization. *Journal of the Association for Information Science and Technology* **47**: 357–369.

Ylijoki O, Porras J. 2016. Perspectives to definition of big data: a mapping study and discussion. *Journal of Innovation Management* **4**(1): 69–91.

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.

DOI: 10.1002/isaf

Publication IV

Ylijoki, Ossi, Sirkiä, Jukka, Porras, Jari and Harmaakorpi, Vesa
Innovation Capabilities as a Mediator between Big Data and Business Model

Reprinted with permission from
Journal of Enterprise Transformation
Accepted for publication

JOURNAL OF ENTERPRISE TRANSFORMATION
https://doi.org/10.1080/19488289.2018.1548396

Innovation capabilities as a mediator between big data and business model

Ossi Ylijoki [a], Jukka Sirkiä [a,b], Jari Porras [a], and Vesa Harmaakorpi [a]

[a]School of Business and Management, Lappeenranta University of Technology, Lappeenranta, Finland; [b]Business Administration, Saimaa University of Applied Sciences, Lappeenranta, Finland

ABSTRACT
The digital transformation is forcing organizations to change towards more data-driven business models. In this paper, we propose a conceptual framework that explains the role of innovation capabilities as a mediator between big data and business model. Using the design science research method approach, we built the framework based on the existing literature. We then applied the framework to the real-world context with three firms and refined it based on the feedback. This study contributes to big data research by pointing out the role of human and data-driven innovation capabilities in the big data value creation process. The developed framework is practitioner oriented, offering a systematic approach towards the development of big data capabilities.

KEYWORDS
big data; business model; business transformation; business value; datafication; digital transformation; framework; innovation

1. Introduction

Linking the opportunities of big data and the business transformation imperative resulting from digitization leads to a situation where incumbent firms must re-think and innovate in their business models and create new capabilities in order to stay competitive in their business ecosystem. Digital technologies, such as social media, cameras or open web sources, produce vast amounts of data. The number of "things" (mobile devices, sensors, etc.) that are connected to the Internet is rising rapidly. Humans generate more and more digital breadcrumbs; we are actually becoming "walking data generators" (McAfee & Brynjolfsson, 2012). The business landscape is becoming turbulent, and changes can happen rapidly. The disruption of many current business models has already begun (Weill & Woerner, 2015).

The data generation phenomenon is called datafication. Datafication can be seen as an "information technology-driven sense-making process" (Lycett, 2013). As datafication proceeds, more and more data are produced,

CONTACT Ossi Ylijoki ossi.ylijoki@phnet.fi School of Business and Management, Lappeenranta University of Technology, PL 20, Lappeenranta 53851, Finland.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujet

processed and transferred. These data are often called big data. Although big data is understood in various ways, the importance of utilizing data in modern businesses is undeniable. Several studies show that data-intensive firms are more effective and profitable than their less data-driven rivals (see e.g. Brynjolfsson, Hitt, & Kim, 2011; Dehning, Richardson, & Zmud, 2003; McAfee & Brynjolfsson, 2012).

Business ecosystems (Iansiti & Levien, 2004; Moore, 1993) and data-related business models are emerging, providing services like data (data-as-a-service), analytics (analysis-as-a-service) or Internet of Things-related offerings (Chen et al., 2011; Leminen et al., 2012). Many new startup firms follow the trailblazers like Google or Amazon and formulate their entire business models on top of or around the data. Data are a valuable asset for these kind of firms. Start-ups heavily rely on data in their decision-making processes and product and service innovations.

For most incumbents, however, a data-driven approach is still a new and unexplored area. Technology vendors have introduced frameworks that aim to help organizations in their big data efforts. A typical approach is to offer "how to" examples or success stories where business benefits have been achieved by utilizing a specific technology stack. In addition to practitioners, scholars are increasingly investigating different aspects of big data, such as the value creation mechanisms (Hartmann et al., 2016; Wixom, Yen, B., & Relich, 2013) and organizational performance (Akter, Wamba, Gunasekaran, Dubey, & Childe, 2016; Ren et al. 2017; Wamba, Gunasekaran, Akter, Ren, Dubey, & Childe, et al. 2017).

The purpose of this study is to present a practitioner oriented framework that explains the role of human and data-driven innovation capabilities as a mediator between big data and the business model of a firm. Our framework helps in understanding how big data and innovations are shaping business models in the digital transformation. The term mediator is used to reconcile problems and promote development between innovation capabilities, big data and business model. The framework offers ways to organize perspectives on the transformation. This helps practitioners to focus on developing the capabilities and methods that best support the transformation towards data-driven business models.

2. Research method

As big data is an emerging area for both scholars and practitioners, the results of the research should benefit both theoretical and practical viewpoints. Therefore, a natural research discipline for the study is information systems (IS) research. IS research draws from behavioral and design sciences, exploring the combination of technology, organization and people. It

should both make theoretical contributions and provide assistance to practitioners in solving current problems (Benbasat & Zmud, 1999; Iivari, 2003).

Design science is one of the two established paradigms in IS research. Hevner et al. (2004) give a clear goal for design science research in their definition: "Design science creates and evaluates IT artifacts intended to solve identified organizational problems." Their description of artifacts covers tangible products, constructs, models and methods. Thus, our framework fits into their definition. We followed the design science research method (DSRM) framework (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), which provides a nominal process for conducting and evaluating design science research in information systems. Figure 1 shows the DSRM framework as we applied it in our study. Starting by identifying the problem, we followed the framework in our study. This paper represents the last stage, communication. The steps or stages, of the DSRM framework can be identified also by looking at the organization of this paper.

After building the framework, we used three real-world case scenarios for evaluation and to develop the artifact further. First, we tested the framework with two big data intensive firms. Secondly, we analyzed one mature stage big data implementation in order to find out how the framework might have benefited the implementation. We then refined the framework based on the feedback from both iterations. The companies utilize big data and look for opportunities to develop data driven business models. The firms are in different stages of utilizing big data and represent different industries. They focus on potential benefits in cost savings, product development and revenue generation. Thus, the three firms offer different viewpoints to the evaluation of the framework.

Other research methods, like action research (Baskerville, 1999; Lewin, 1947) or action design research (Sein et al., 2011) provide applicable methods for dealing with organizational change as well. However, we found that the design science DSRM framework would meet the needs of our research. We were developing an artifact (the framework) and therefore preferred design science to action research. As our goal was to develop a generic framework in a novel area, we preferred rapid iterations with moderate

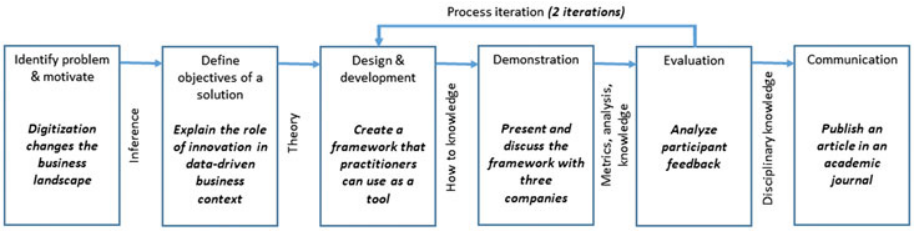


Figure 1. Research method, applied from Peffers et al. (2007).

user input instead of their long-term commitment. In addition, practitioners favor approaches that present initial concepts or assumptions in order to provoke new ideas, instead of a clean sheet of paper as a starting point. Experimenting is a common technique for iterative development. The development, demonstration and evaluation stages of DSRM form a cycle that fits well to rapid experimenting without overloading the participating firms.

3. Building the framework

Linking the opportunities of big data (as a consequence of datafication) and the business transformation imperative (resulting from the digital transformation) leads to a situation where incumbent firms must re-think (innovate) their business models and create new capabilities in order to stay competitive in their business ecosystem. The objective of the developed framework is twofold:

- Advance big data research by addressing the role of innovation capabilities in data-driven business context.
- Assist practitioners in their big data initiatives by offering a framework that helps them in developing required capabilities.

3.1. Theoretical background

The framework treats innovation capabilities as a mediating factor between the business model and big data. Therefore, the nature of the framework is multi-disciplinary. The elements are related to strategic management and information systems research.

The value chain and five forces model, e.g. (Porter, 1991; Porter & Millar, 1985), are commonly used strategic management frameworks. The former is an internal view of the business, whereas the latter provides an outside-in industry view. Porter's models explains the source of competitive advantage by two factors: low cost or differentiation. Another established strategic management framework is the resource-based view (Wernerfelt, 1984). It combines the internal and external views of previous models. The resource-based view aligns well with data-oriented business environments, where digital ecosystems, e.g. (Weill & Woerner, 2015), are increasingly used to add value.

IS research draws from behavioral and design sciences, exploring the combination of technology (including data), organization and people. This research should both make theoretical contributions and provide assistance to practitioners in solving current problems (Benbasat & Zmud, 1999;

Iivari, 2003). One of the strengths of information systems research stems from the combination of behavioral and design sciences; technology, data and behavior are inseparable in an information system.

Digitalization implies business changes that are technology driven. Thus, strategic management and IS research provide a broader view to the topic. Moreover, change requires innovation. Start-ups and incumbents search for new, technology and data driven innovations. We discuss how data- and human-driven innovations act as a mediator between (big) data and business model later in this paper. In this section, we briefly present and define the building blocks from the above described disciplines that our framework relies on, namely the concepts of business model, capabilities, innovation and big data.

3.1.1. Business model

The concept of business model has been vividly discussed in the strategic management discipline. It relates to how an organization arranges its functions in order to create value. The term business model is often used interchangeably with strategy (Burkhart et al., 2011). Many definitions exist for both terms (e.g. Amit & Zott, 2001; Casadesus-Masanell & Ricart, 2010; Teece, 2010; Timmers, 1998). However, there seems to be consensus that these concepts differ from each other: strategy is more focused on competition and product-market matters, whereas the business model describes how the strategy is implemented (Zott, Amit, & Massa, 2011).

Osterwalder and Pigneur (2010) define the business model as follows: "A business model describes the rationale of how an organization creates, delivers, and captures value." They also present a business canvas framework that serves as a tool when analyzing how an enterprise creates value. Furr and Dyer (2014) present a simplified business canvas framework. It includes three main elements: solution (which is further divided for value proposition and pricing strategy), cost structure (activities and resources) and customer acquisition (relationships and channels).

3.1.2. Capabilities

Capabilities are related to resources and the resource-based view (RBV) of a firm (Wernerfelt, 1984). In the resource-based view of a firm, resources and capabilities are the key concepts that explain a firm's competitive advantage. Teece (2007) introduced the concept of dynamic capabilities, defining them as "the firm's ability to integrate, build and reconfigure internal and external competences to address rapidly changing environments," which complements the RBV by explaining how firms renew their competencies in order to adapt to the changing business environment.

Big data is a disruptive new technology. Incumbent firms typically rely on the idea that their current capabilities are relevant with regard to new technologies, but often this is just based on assumptions (Sainio, 2005). Accordingly, incumbents often fail to take advantage of disruptive technologies, even when they are well aware of them (Christensen, 2013). These are the main caveats for incumbents and therefore a deeper understanding of the datafication phenomenon and correct recognition of the required capabilities are crucial steps when a firm assesses the effects of big data.

3.1.3. Innovation

In order to innovate effectively and develop an innovative organization culture, the firm needs to understand the nature of innovation. Innovation and change are tightly related: to innovate effectively aims at changing something. Many different perspectives towards the concept of innovation have been taken in several disciplines. Thus, there are different definitions of the term (e.g. those suggested by Ettlie & Reza, 1992; Schumpeter, 1942; West & Anderson, 1996). After a review of existing definitions, Baregheh, Rowley, & Sambrook, (2009) provided a synthesized definition that views innovation as a dynamic capability: "Innovation is the multi-stage process whereby organizations transform ideas into new/improved products, service or processes, in order to advance, compete and differentiate themselves successfully in their marketplace."

3.1.4. Big data

Laney (2001) described three essential dimensions of big data: "volume, velocity and variety" (the 3 V definition). Volume refers to ever-increasing amounts of data. Velocity indicates the need to capture and analyze high-speed data in (near) real-time or else the value may be lost. Variety relates to different types of data, be it structured or non-structured, such as social media posts or a video. In recent years, scholars and especially practitioners have developed numerous big data definitions. Ylijoki and Porras (2016b) present a detailed discussion about big data definitions and the history of the term, concluding that the abovementioned dimensions are not only used in most of the definitions, but also form a logically coherent set. For the purposes of this research, the 3 V definition is adequate.

3.2. Innovation capabilities as mediator framework

The value proposition of big data is that companies can gain benefits by making use of it. Indeed, new, data-driven business models are emerging, like the data-as-a-service or analytics-as-a-service approaches discussed by

Chen et al., 2011). To give another example, van't Spijker (2014) lists five value generation models for data-driven business: 1) selling data directly, 2) innovating products through data, 3) swapping commodity offerings into value-added services, 4) creating interaction in the value chain and 5) creating a network of value based on data exchange. One example of data monetization is presented by Najjar and Kettinger (2013), who have investigated the process, benefits and drawbacks of the data monetization process in a large drug retailing firm. Internet of Things-related business models are evolving (Bucherer & Uckelmann, 2011; Chan, 2015; Leminen et al., 2012; Westerlund, Leminen, & Rajahonka, 2014). These examples clearly indicate that big data affects business models.

We adopt the business canvas model by Furr and Dyer (2014) into our framework, shown in Figure 2, as it covers essential elements from the innovation point of view. The business canvas model provides a (sufficiently) detailed starting point to understand the potential impact of big data in business context. Organizations can develop different scenarios using the canvas as a tool. It helps in concentrating on different viewpoints, such as streamlining core activities or developing new pricing strategies with data and analytics. On the other hand, it helps in seeing the big picture, as it covers all core functions of an organization. Moreover, we use the 3 V definition of big data (Laney, 2001), supplemented with a categorization element for each of the dimensions, in order to ease the use in practice.

The data deluge and changing business models are challenging incumbents to develop new capabilities. Approaches using resource-based theory have been popular among big data researchers recently, (Braganza, Brooks, Nepelski, Ali, & Moro, 2017; Gupta & George, 2016; Mikalef et al., 2017). We make no exception to the approach. However, we wish to add to the conversation the role of innovation capabilities in shaping the business model, as shown in Figure 2. In the big data context, we distinguish between "data- and human-driven" innovations.

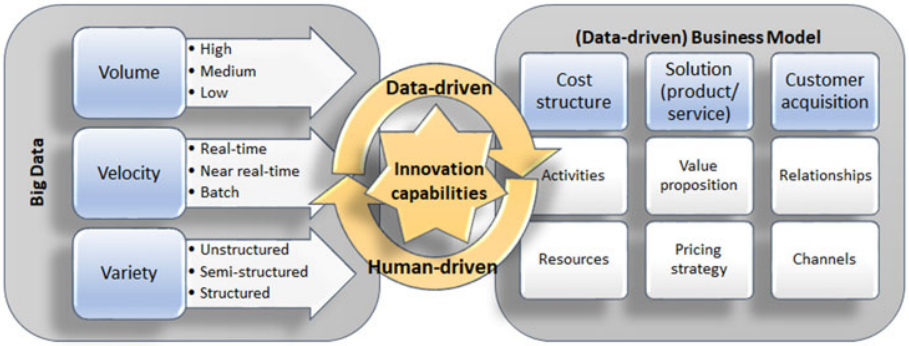


Figure 2. Innovation capabilities as a mediator framework.

The "human-driven approach" to innovation states that innovations stem from people (Harmaakorpi & Melkas, 2012). New ideas – the seeds of innovations – often emerge sub-consciously, when our brain associates things. Dyer, Gregersen, and Christensen (2008) claim that by training certain skills systematically, anyone can accelerate associative thinking. The four discovery skills are: observing (especially looking for surprises), networking with people from different backgrounds, experimenting a lot and questioning the status quo. Innovative people spend more time exercising these skills than others. New ideas emerge from self-transcending knowledge (Scharmer, 2001) as a result of associations. Self-transcending knowledge is future-oriented knowledge. It could be considered to be tacit knowledge prior to its embodiment. A good example of a person utilizing this approach was Steve Jobs.

Enabling human-driven innovation requires that individuals may use part of their time in seemingly unproductive discovery-skills development. This is important for developing self-transcending knowledge. Another requirement is that the organization's culture and management practices must support risk-taking and allow mistakes. Dyer, Gregersen, and Christensen (2011) present four principles that the most innovative organizations apply: 1) innovation is everybody's job, 2) both incremental and disruptive innovations must be considered, 3) innovations are developed in small teams and 4) managed risks must be taken. These are by no means trivial requirements: e.g. Sandberg and Aarikka-Stenroos (2014) identify a restrictive mindset, lack of discovery competences and unsupportive organizational structures as the three main internal barriers to radical innovations in incumbent firms.

Another way to view innovations relies on data and automation. A novel approach, "innovation automation," is data-driven. Data-driven innovation suggests that the innovation processes could and should, be automated (Shaughnessy, 2015). This approach puts technology and (big) data at the core of the innovation processes. More and more data becomes available, technological and analytical capabilities are increasing and data processing costs are decreasing. Utilizing automation and vast volumes of different data will produce a more holistic view, leading to more data-driven decisions and more agile innovation processes. For example, big data might be used for early recognition of trends, combined with automatized simulations of possible scenarios in order to fertilize people's thinking and, in turn, new human innovations. An example of innovation automation approach is Netflix, a company offering on-demand movies and television series.

Data-driven innovation, especially when combined with rapid experiments, can significantly speed up the innovation process, as the Netflix (Amatriain, 2013) example shows. Manyika et al. (2011) claim e.g. that

manufacturing firms can reduce their product development costs by 20–30% as well as achieve up to 50% faster time-to-market cycle by utilizing big data. One case study of utilizing data to drive innovations is documented by Jetzek, Avital, and Bjorn-Andersen(2014). However, for most incumbents, innovating with big data is in its infancy.

Innovation capabilities are at the core of the transformation of a firm. Therefore, understanding the nature of innovation is a prerequisite of a fertile innovation process. However, the kind of activities that efficient innovation processes require are in contrast with the current procedures of most incumbent firms. Incumbents organize themselves to be effective. They concentrate on minimizing all kinds of waste (like allocating time to seemingly unproductive activities such as developing discovery skills) and avoiding errors. This indicates that focusing on the organizational aspects of innovation capabilities is an important factor for a successful business transformation.

4. Evaluating the framework

The purpose of demonstration and evaluation phases was to assess the applicability of the framework in practical situations in business. Another objective was to gather feedback in order to develop the framework further.

4.1. Demonstration

The following sections describe the demonstration phase. In addition to presenting our approach and the companies, we provide a synthesis of the findings.

4.1.1. Demonstration process

First, we presented the framework to two big data intensive firms. One operates in several countries, offering cloud-based human resource (HR) solutions, and the other is a startup company offering SaaS-based integrations. Both firms act as "data hubs" and deal with large amounts of data. This gives them opportunities to develop even smarter data-related services or products, like predictive HR signals or benchmarking analyses. Using the proof-of-concept approach, we demonstrated the usage of the framework through semi-structured interviews. The people we interviewed were members of the executive boards of their firms. Our aim was to find out the usability of the framework in firms that are in the quite early stages of exploiting big data. In the second iteration, we performed a "postmortem" examination of a real-world big data case. Reflecting the case against the

framework, we explored how the framework might have helped the company to understand the impact of big data.

4.1.2. Demonstration contexts

The arrangements for the demonstrations of the framework were as follows: we prepared materials for the interviews beforehand. The material consisted of an overview presentation of the framework and a brief questionnaire to assess the current situation. As orientation material, we sent a set of preliminary questions to the interviewees a couple of days before the interviews. The duration of each interview was around two hours. We recorded the interviews and took notes. At the end, we asked for feedback, both verbal and written. Afterwards, we analyzed our notes and outlined them in memos. Any unclear points were checked in the recordings. The interviewees reviewed the memos.

The first iteration was to introduce the framework to the HR solutions provider and the integration services provider. Two interviewers participated in both interviews. Sympa, the HR solutions provider, is an established firm in the HR business. Their key offering is Sympa HR, a cloud-based tool for managing employment relationships, from recruitment to termination of employment. The firm is an established company that already has a large number of customers in several countries. However, they are seeking rapid growth. Their management has discussed big data, although they have not yet formulated a clear big data strategy.

The other firm, Flashnode, is a Finnish startup offering services for interconnecting and synchronizing data. The firm's business model relies on highly standardized integrations, which challenges the traditional tailored interface development approach. A key objective of automated integrations is to reduce the manual work done by client companies. They have an aggressive growth strategy. Their position as a data hub offers various opportunities to utilize big data. The company accumulates huge amounts of data in terms of transaction volumes.

The second iteration was the postmortem examination of an existing big data initiative. A bus company operating 430 buses in urban and suburban areas wanted to improve their cost efficiency. A potential target for this purpose was to reduce fuel costs, as these costs represent the company's second largest expense. Meeting the business goal would require a transformation process consisting of training and motivating of the drivers, and a technical solution to collect detailed fuel consumption data. We applied the same procedures as in the first iteration, with the exception that only one interviewer was present. During the analysis of the case, we interviewed the owner of the big data initiative (the chief technical officer,

CTO) and inspected the software vendor's project documentation. The bus firm treated the initiative as a business development program

4.1.3. Demonstration findings

Here are the key findings of research and interviews of three different firms. The HR solution provider firm sees several areas where big data can add value to their business, such as integrating external data, utilizing metadata that their platform generates, and creating a "Sympa tribe" ecosystem for their customers. Moreover, they increasingly utilize sensor data, such as access control logs and working time recordings.

In the second case the provider of system integration the data is internally mostly applied to product development prioritization. Based on the big data, internal analyzes have also been carried out to compare customer-specific data volumes and other business customers trends. In this case, customer data is enriched with other external sources. Based on the above, the purpose is to develop a reporting service that provides benchmarking data to customers. For example, how does our business compare to industry in general in terms of transactions of subscription volumes.

In the case of the bus company they set the business goal as follows: to educate and motivate the drivers to change their driving habits, which will in turn lead to reduction in fuel consumption. They appointed the CTO to lead the program. They consulted potential software and hardware vendors and selected the technologies. The implementation started with a pilot project, which integrated the technologies using data from ten buses. After an assessment of the pilot, a full-scale implementation project followed. During the following months, the project team implemented a production-ready system and solved several data quality-related issues. At the same time, the firm created a change management plan for the release and roll-out. After two years of use, the system has proved to be a success. It has met the original business goal. The system and its usage have costs, but the net result is a significant cost reduction due to permanently lower fuel consumption. Moreover, the firm has recorded remarkable improvement with regard to the traveling experience in customer satisfaction surveys. The passengers have noticed that the drivers drive smoothly. A third achieved benefit is, of course, the reduction of carbon-di-oxide emissions.

4.2. Evaluation

In this section we first analyze our findings from the demonstration. Next, we offer building blocks for practitioners by describing an exemplary process and discussion on how to apply the framework in practical situations.

4.2.1. Evaluation of the demonstration

The first iteration of the demonstration phase confirmed that the framework provides a mental model for evaluating the effects of big data in the business context. The interviewees found the framework useful for this purpose. The HR solution provider emphasized that the framework helps to create a systematic approach that will save time. Managing a growing business is hectic; they could not tie key resources to long consulting projects. The integration services provider stated that the framework offers guidance and provides understanding of big data adoption. They saw big data as a strategic issue. Applying the framework would help identify the required changes, as the business model aspect of the framework emphasizes the strategic importance of big data. Another key aspect of the framework is focusing on the importance of data, as "you can always build or buy algorithms, but you cannot generate data from the void."

In both firms, we noted that the framework helped keep the discussion focused. It helped in keeping concentrated on one topic at a time. Moreover, the discussion regarding new ideas was lively. Although this may be related to personal characteristics, the framework facilitated the ideation. Both firms considered the framework useful in their feedback. After the first iteration, we reflected the discussions in our framework. From the theory point of view, no new aspects were recognized. By looking at the practical side, the interviewees appreciated our orientation material, as it helped them to perceive the framework.

In the second iteration, we analyzed the bus firm's big data initiative. The initiative started when the firm noticed that the current technology enabled them to collect detailed data from the buses. This idea led ultimately to the big data initiative. The management of the company considered that the big data initiative would strengthen their competitive situation due to increased operational efficiency.

The firm set a clear objective for the program, although they did not define an explicit value for it. They also set a member of the executive board (the CTO) to lead the program. They focused on meeting the business goal, i.e. reducing fuel consumption. The firm also recognized real-time monitoring needs on the horizon. However, as the fuel consumption analysis did not require real-time processing, they decided to postpone the implementation to later phases. The project team, consisting of members of the firm and a software vendor, identified the required datasets and drew up a cost-effective architecture without paying attention to additional features. In the pilot phase, neither the firm nor the software vendor considered real-time needs. The project members may have discussed possible future scenarios in their coffee breaks, but the project did not consider those scenarios officially. This ultimately led to a situation where further development to meet the real-time needs is difficult.

Our framework would have helped the bus firm to understand the situation better at the beginning of the program. The owner of the big data initiative shared this view and saw that the framework would have been helpful. A better understanding of the data might have helped them to earlier identify certain data-related issues that reflected on the attitudes of middle management. Moreover, the framework would have helped to develop capabilities to tackle the data issues. Another capability-related matter was the shortage of analytical capabilities, which has been a hindrance to making use of the gathered data. In addition, using the framework in the early phases would probably have stressed the future needs. This, in turn, would probably have led to a different architecture or technology selection in the project phase.

The two companies that had less experience with big data paid only a little attention to developing capabilities, whereas the postmortem indicated that more attention should have been paid to capability development during the project. Another observation, although an expected one, was that the usage of the framework benefits from facilitation. Preparing supporting material for the interviews and group work adds a middle layer between the framework and the daily operations. This helps the participants connect the framework concretely with their own context. Using a few questions or brief examples that concretely link to the firm's current situation would concretize the subject and help link big data to the firm's business context even better than a generic approach.

The demonstration phase confirmed our initial assumption that simplicity is a virtue when providing building blocks to practitioners. This assumption in mind we carefully considered in the first place the balance between theoretical completeness and practical viewpoints in the presentation of the framework. E.g., the simplified business canvas model (Furr & Dyer, 2014) we chose, covers essential components from innovation point of view. However, it should be noted that extension to full business model canvas (Osterwalder & Pigneur, 2010) can easily be done (by scholars or practitioners) whenever required.

4.2.2. Suggestions for making use of the framework

Companies can choose different practical approaches to the framework according their situation and objectives. Increasing the role of data in current business model, e.g. by automating delivery processes to decrease transaction costs requires a different approach than developing a new, data-driven model. Based on the evaluation phase of our framework, we synthesized an exemplary usage scenario as follows.

1. Assess current situation and big data impact: This can be done by looking at the business model components (see Figure 2) one by one and

creating scenarios of potential big data effects. Effects can be either risks or opportunities; some of the effects are such that they cannot be influenced, some of them are within the reach of the company.

2. *Define business objectives.* Clear goals and ownership are best practices for any business development initiative. Especially in the case of new things solid leadership is important to achieve the set goals. This phase might be iterative, innovating back and forth between the business model and big data. What internal or external data could affect our business? What data do we need in order to best run our business? Novel ideas or long-term, strategic goals often require experimenting for verification.
3. *Evaluate/ideate big data value potential.* Many innovations are human-driven, but increasingly, data are a source of innovations. Humans ideate things that require gathering and combining new and existing data in novel ways. Accordingly, experimenting with data may reveal insights that spark ideas. Human-driven and data-driven innovations are not mutually exclusive. Instead, they can, and should, support each other, effectively creating an innovation loop. Our framework provides support for these activities by explaining the theoretical background and enablers for an effective innovation process.
4. *Identify required datasets and capabilities.* Identifying data that is beneficial to the business is a two-way operation. As one may intuitively think when looking at our framework (Figure 2), human-driven innovation comes down to a question: "What data do we need to fulfill this business need?" On the other hand, data-driven innovations ask: "Do we have data or can we access data that is beneficial to our business?" Once the datasets are identified, the modified 3V model shown in Figure 2 can be used to classify and categorize the data at a more detailed level. This, in turn, reveals possible gaps in ICT capabilities.

As another building block, and from the IT point of view, the 3 V model of big data (Laney, 2001) is a good starting point. Adding categorization elements to the dimensions (see Figure 2) and classifying any available or required data source (internal or external) accordingly helps to identify potential development needs of the current IT platform. For example, a business need may need to harvest and analyze social media data. The framework helps to translate the business need into an IT-related requirement: Do we have the hardware and software that is required to gather and process (relatively) low volumes of unstructured data in near real-time? It should, however, be noted that there are technical and data-related challenges (Benabdellah et al., 2016; Ylijoki & Porras, 2016a). Technical challenges are hardware- and software -related, such as managing vast

volumes of data with a Hadoop cluster, or using a not only SQL (NoSQL) database to store unstructured data. Analytical capabilities are human-centric, like building a predictive analytics model or interpreting a business need into an algorithm. A firm can develop technical and analytical capabilities in-house, or it can leverage service providers.

5. Conclusion

The paradigm shift towards more data-intensive business landscape is inevitable. Companies must consider the combination of big data, innovations and potential value against the required transformation when they plan their big data initiatives. For incumbents, the transformation may be more revolutionary than evolutionary, which implies a complicated process. Our practitioner oriented framework helps in understanding the role of innovations and (big) data in the digital transformation. Understanding these factors forms the basis for informed management decisions regarding business transformation and the capabilities the transformation requires.

In this paper, we have presented a multi-disciplinary framework that contributes to research by pinpointing the role of human and data-driven innovation capabilities as a mediator between big data and the business model. The framework combines elements from strategic management, innovations research and resource-based theory. It helps to understand the role of human and data-driven innovations, and innovation capabilities in an organizational context. Thus, the framework acts as a mental model that offers a way to organize perspectives on the digital transformation, especially from the practitioner's point of view. Understanding the theoretical background of the innovation process in the big data context will help practitioners to focus on developing the capabilities and methods that best support the transformation towards data driven business models.

ORCID

Ossi Ylijoki http://orcid.org/0000-0001-8107-0796

References

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., and Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Production Economics*, *182*, 113–131.

Amatriain, X. (2013). Beyond data: from user information to business value through personalized recommendations and consumer science. Paper presented at Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, California, USA, October 27, November 01, 2013 (pp. 2201–2208).

Amit, R., and Zott, C. (2001). Value creation in e-business. *Strategic Management Journal*, *22*(6-7), 493–520.

Baregheh, A., Rowley, J., and Sambrook, S. (2009). Towards a multidisciplinary definition of innovation. *Management Decision*, *47*(8), 1323–1339.

Baskerville, R. L. (1999). Investigating information systems with action research. *Communications of the Association for Information Systems*: *2*(19). DOI: 10.17705/1CAIS.00219

Benabdellah, A. B., Asmaa, B., Imane Z, and El Moukhtar. (2016). Big data for supply chain management: opportunities and challenges. Paper presented at IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) 29 Nov.–2 Dec. 2016, Agadir, Morocco.

Benbasat, I., and Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly*, *23*(1), 3–16.

Braganza, A., Brooks, L., Nepelski, D., Ali, M., and Moro, R. (2017). Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research*, *70*, 328–337.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in Numbers: How does data-driven decision-making affect firm performance?" (2011). ICIS 2011 Proceedings. 13. https://aisel.aisnet.org/icis2011/proceedings/economicvalueis/13

Bucherer, E., and Uckelmann, D. (2011). Business Models for the Internet of Things. In: Uckelmann D., Harrison M., Michahelles F. (eds) *Architecting the Internet of Things*. Berlin, Heidelberg: Springer.

Burkhart, T. K., Julian, W., and Dirk, L. P. (2011). Analyzing the business model concept—A comprehensive classification of literature. Paper presented at Thirty Second International Conference on Information Systems, December 2–5, Shanghai China 2011.

Casadesus-Masanell, R., and Ricart, J. E. (2010). From strategy to business models and onto tactics. *Long Range Planning*, *43*(2-3), 195–215.

Chan, H. C. (2015). Internet of Things business models. *Journal of Service Science and Management*, *8*(4), 552–568.

Chen, Y. K., Jeffrey, C. M., and Abrams, C. (2011). Analytics ecosystem transformation: a force for business model innovation. Paper presented at SRII Global Conference (SRII), 2011 Annual 29 March; 2 April 2011, San Jose, CA, USA (pp.11–20).

Christensen, C. (2013). *The innovator's dilemma: When new technologies cause great firms to fai*. Boston, Massachusetts: Harvard Business Review Press, 5th edition.

Dehning, B., Richardson, V. J., and Zmud, R. W. (2003). The value relevance of announcements of transformational information technology investments. *MIS Quarterly*, *27*(4), 637–656.

Dyer, J., Gregersen, H., and Christensen, C. (2008). Entrepreneur behaviors, opportunity recognition, and the origins of innovative ventures. *Strategic Entrepreneurship Journal*, *2*(4), 317–338.

Dyer, J., Gregersen, H., and Christensen, C. (2011). *The Innovator's DNA*. Boston, Massachusetts: Harvard Business Review Press.

Ettlie, J. E., and Reza, E. M. (1992). Organizational integration and process innovation. *Academy of Management Journal*, *35*(4), 795–827.

Furr, N., and Dyer, J. (2014). *The innovator's method*. Boston, Massachusetts: Harvard Business Review Press.

Gupta, M., and George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, *53*(8), 1049–1064.

Harmaakorpi, V., and Melkas, H. (2012). *Practice-based innovation. insights, applications and policy implications*. Berlin, Germany: Springer.

Hartmann, P., Zaki, M. E., Feldmann, N., and Neely, A. D. (2016). Capturing value from big data–A taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, *36*(10), 1382–1406.

Hevner, A. R., et al. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Iansiti, M., and Levien, R. (2004). Strategy as ecology. *Harvard Business Review*, *82*(3), 68–81.

Iivari, J. (2003). The IS core-VII: towards information systems as a science of meta-artifacts. *Communications of the Association for Information Systems*, *12*(1), 568–581.

Jetzek, T., Avital, M., and Bjorn-Andersen, N. (2014). Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*(2), 100–120.

Laney, D. (2001). 3D data management: controlling data volume, velocity and variety. *Meta Group Research Note*, 6, 70–73.

Leminen, S., et al. (2012). Towards IoT ecosystems and business models. In, *Internet of Things, smart spaces, and next generation networking*. A. Sergey, B. Sergey, K. Yevgeni, (Eds.), St. Petersburg, Russia: Springer, pp. 15–26.

Lewin, K. (1947). Frontiers in group dynamics II. Channels of group life; social planning and action research. *Human Relations*, *1*(2), 143–153.

Lycett, M. (2013). Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems*, *22*(4), 381–386.

Manyika, J., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity. In* J. Manyika and M. Chui, (Eds.), McKinsey Global Institute. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

McAfee, A., and Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 61–67.

Mikalef, P., et al. (2017). Big data analytics capabilities: A systematic literature review and research agenda. *Information Systems and e-Business Management*, *16*(3), 1–32.

Moore, J. F. (1993). Predators and prey: A new ecology of competition. *Harvard Business Review*, *71*(3), 75–83.

Najjar, M. S., and Kettinger, W. J. (2013). Data monetization: Lessons from a retailer's journey. *MIS Quarterly Executive*, *12*(4), 213–225.

Osterwalder, A., and Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers*, Hoboken, NJ: John Wiley & Sons.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Porter, M. E. (1991). Towards a dynamic theory of strategy. *Strategic Management Journal*, *12*(S2), 95–117.

Porter, M. E., and Millar, V. E. (1985). How information gives you competitive advantage. *Harvard Business Review*, *64*(4): 149–160.

Ren, S. J.-F., et al. (2017). Modelling quality dynamics, business value and firm performance in a big data analytics environment. *International Journal of Production Research*, *55*(17), 5011–5026.

Sainio, L.-M. (2005). *The effects of potentially disruptive technology on business model - a case study of new technologies in ICT industry*. Lappeenranta, Finland: Lappeenranta University of Technology.

Sandberg, B., and Aarikka-Stenroos, L. (2014). What makes it so difficult? A systematic review on barriers to radical innovation. *Industrial Marketing Management*, *43*(8), 1293–1305.

Scharmer, C. O. (2001). Self-transcending knowledge: Sensing and organizing around emerging opportunities. *Journal of Knowledge Management*, *5*(2), 137–151.

Schumpeter, J. A. (1942). *Socialism, capitalism and democracy*. United States: Harper and Brothers.

Sein, M., et al. (2011). Action design research. *MIS Quarterly*, *35*(1), 37–56.

Shaughnessy, H. (2015). *Shift: A user's guide to the new economy*, Boise, Idaho: Tru Publishing.

Teece, D. J. (2010). Business models, business strategy and innovation. *Long Range Planning*, *43*(2-3), 172–194.

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*(13), 1319–1350.

Timmers, P. (1998). Business models for electronic markets. *Electronic Markets*, *8*(2), 3–8.

Van't Spijker, A. (2014). *The new oil: Using innovative business models to turn data into profit. Basking Ridge*, NJ: Technics Publications.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J-F., Dubey, R., and Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356–365.

Weill, P., ans Woerner, S. L. (2015). Thriving in an increasingly digital ecosystem. *MIT Sloan Management Review*, *56*(4), 27–34.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal*, *5*(2), 171–180.

West, M. A., and Anderson, N. R. (1996). Innovation in top management teams. *Journal of Applied Psychology*, *81*(6), 680–693.

Westerlund, M., Leminen, S., and Rajahonka, M. (2014). Designing business models for the Internet of Things. *Technology Innovation Management Review*, *4*(7), 5–13.

Wixom, B. H., Yen, B., and Relich, M. (2013). Maximizing value from business analytics. *MIS Quarterly Executive*, *12*(2), 111–123.

Ylijoki, O., and Porras, J. (2016a). Conceptualizing big data: Analysis of case studies. *Intelligent Systems in Accounting, Finance and Management*, *23*(4), 295–310.

Ylijoki, O., and Porras, J. (2016b). Perspectives to definition of big data: A mapping study and discussion. *Journal of Innovation Management*, *4*(1), 69–91.

Zott, C., Amit, R., and Massa, L. (2011). The business model: Recent developments and future research. *Journal of Management*, *37*(4), 1019–1042.

Publication V

Ylijoki, Ossi

Guidelines for Assessing the Value of a Predictive Algorithm - a Case Study

Reprinted with permission from
Journal of Marketing Analytics
Vol. 6(1), pp. 19-26, 2018

J Market Anal (2018) 6:19–26
https://doi.org/10.1057/s41270-017-0027-1

ORIGINAL ARTICLE

Guidelines for assessing the value of a predictive algorithm: a case study

Ossi Ylijoki[1]

Revised: 29 December 2017 / Published online: 15 January 2018

Abstract Predictive algorithms are increasingly used to support decision making. Understanding the costs and benefits of a predictive model is an important aspect for businesses. However, algorithms are abstract, and their impact oftentimes remains vague. We present a case study, where a machine-learning algorithm is used for bid qualification. We show how to apply classification matrices for business value assessment and propose guidelines and metrics for interpreting the impact in practical solutions.

Keywords Business case · Business value · Big data · Case study · Machine learning · Predictive analytics

Introduction

The focus of active sales management research has been shifting over time from understanding the selling process to sales force automation, and from sales reporting to predictive sales analytics. The seven steps of selling paradigm (Dubinsky 1981) is a well-known and widely used framework for explaining and understanding the selling process. The process and its steps—(1) prospecting, (2) pre-approach, (3) approach, (4) presentation, (5) overcoming objections, (6) closure, and (7) follow-up—are discussed in numerous scholarly sources (e.g. Dwyer et al. 2000; Long et al. 2007; Moncrief and Marshall 2005). In the literature, the number of stages varies, as well as the meaning of each step, but the basic approach holds. Understanding the selling process enables sales force automation.

The sales process of our case study organization is organized according to seven steps of selling paradigm and sales force automation principles. Sales force automation is an active research area, where substantial business benefits have been achieved. Flanagan (1995) presents case study examples, pinpointing the efficiency and information value benefits gained from automation. Scientific approaches, such as return on sales investment, or science-driven selling, have been proposed (Ledingham et al. 2006; Lukes and Stanley 2004). These approaches have shown impressive sales and productivity growth percentages. Lawrence et al. (2010) presents two information systems developed for increasing the efficiency and productivity of a global corporation. One of the systems aims to detect new sales opportunities, whereas the other focuses on allocating sales resources. They estimate that the business impact of the two solutions is hundreds of millions of dollars annually.

Automation brings benefits, but synchronizing sales with operations is difficult (Cooper and Budd 2007). Dealing successfully with production, resource utilization and delivery times require a steady, or at least predictable flow from sales. Therefore, companies typically coordinate their sales activities closely, for the most part organizing the sales process so that it roughly follows the seven steps selling paradigm. Independent of the exact number of steps, the process forms a sales funnel (aka. *sales pipeline*). In the funnel, part of the prospects turn into offers, and eventually some of the offers realize as new revenue. Information systems and sales force automation enable companies to gather detailed data from their sales process. Coordination between sales and operations requires sophisticated information as well (e.g. Cooper and Budd 2007; Bhattacharyya 2014). Moreover, factual data

✉ Ossi Ylijoki
ossi.ylijoki@phnet.fi

[1] School of Business and Management, Lappeenranta University of Technology (LUT), Puistokatu 1C38, 15100 Lahti, Finland

enable the creation of objective metrics (e.g. Beam 2006; Morgan and Rego 2006). Our case study organization seeks quantitative analysis of the sales pipeline and aims to identify winning opportunities as early as possible in order to direct resources more efficiently.

In recent years, new quantitative sales pipeline data analytics have been proposed in areas like bid qualification and pipeline prioritization (e.g. D'Haen and Van den Poel 2013; Eichhoff and Maass 2014; Greenia et al. 2014; Yan et al. 2015). While there is no doubt that these analytics potentially add value to business, the viewpoint of the papers is more or less technical, focusing on matters like prediction accuracy or statistical significance. Management is willing to promote big data initiatives (Ylijoki and Porras 2017), and there is an increasing number of papers dealing with data-driven approaches (e.g. Amatriain 2013; Dutta and Bose 2015; Wixom and Ross 2017). Still, if we look at the business side, technical capability is not enough. Business executives think of added value. Value is a concrete measure which helps in prioritizing investments. Being able to concretize the value of an algorithm is often a decisive factor between a go or no-go decision.

In this paper, we first present a case study, where a predictive algorithm is used to qualify bids. We then use classification matrices to estimate the business value of the algorithm, followed by practical guidelines, and discussion about the applications and limitations of our method.

Case study presentation

The subject of our case study was a large information technology service provider, which we shall refer to as "SWCo". SWCo provides systems integration, consulting, and outsourcing services. They have a large customer base, ranging from small, local clients to international, multi-industry companies. The nature of their offerings is heterogeneous as well, varying from small up-selling offers to product offerings, project implementations, and large outsourcing bids. Most of the opportunities (the terms bid and opportunity are used as synonyms in this paper) are unique and complex. The number of bids, i.e. projects and services in the pipeline, is several thousand at any given point in time. Due to heterogeneous offerings, and the size of the pipeline, identifying winning bids as early as possible is a key challenge for sales force management and sales efficiency.

The sales process of SWCo

SWCo uses a sales pipeline model that has seven stages (Table 1). For comparison, the last column of Table 1 shows corresponding steps of the seven steps paradigm (Dubinsky 1981). The pipeline starts with lead identification. Every employee of the company can record a lead, i.e. a potential sales opportunity, in the CRM. In stage 01, business development personnel perform an initial screening of the opportunities, and decide to either approve or discard the lead. Starting from stage 02, the opportunities and the results of each stage are reviewed both from the viewpoint of win potential and internal controls, such as profitability and legal aspects.

Each opportunity is categorized using a status code. Possible status code values are "Open", "Cancel", "Lost", and "Won". The status of a new lead is "Open". The status of opportunities eliminated from the pipeline during the sales process is set to "Cancel". When an opportunity is won, the status is set to "Won". Accordingly, if a competitor wins, the status is set to "Lost". Thus, the final status is either "Won", "Lost", or "Cancel".

The sales process is well organized and controlled. However, applying new technologies to bid qualification might bring benefits, compared to the current situation. The problem setting is as follows.

- Identifying winning opportunities at early stages is difficult. Stages 03–06 account by far for most of the sales costs of SWCo. Currently, a relatively high percentage of the opportunities survive to the later stages. This leads to inefficient sales force allocation, additional costs, and inaccurate sales forecasts.
- The basic metrics (number of opportunities/stage, win rate etc.) of the sales pipeline are known. However, no quantitative metrics are available, which would support qualification. Leads are qualified one by one in qualifying meetings with varying participants. This means that in practice decision making is emphasized by personal views and valuations.

An optimal solution would provide a baseline, where (1) winning bids are identified as early as possible and (2) all bids are evaluated using the same, quantitative criteria. This baseline would act as an insightful basis for decision making to sales management and sales teams.

Machine-learning solution for bid qualification

The objective of bid qualifying is to identify winning opportunities from the current sales pipeline. From the revenue point of view, a cancelled bid equals a loss. Thus, we can reduce the problem to a binary classification problem—the end result is either a win, or a no-win. We believe that this will not introduce bias to predictions. Bids are cancelled either by customer or SWCo. Customer cancellations are out of reach for SWCo. When SWCo cancels, the root reason is that either the offering does not meet customer needs, or commercial terms would be

Table 1 Sales pipeline framework at SWCo

Stage	Explanation	Status	7 steps paradigm
00 Lead/suspect	A potential sales opportunity has been identified	Open	Prospecting
01 Identification	Lead screening done and qualifications planned (for approved opportunities)	Open, Cancel	Pre-approach
02 Qualification	Opportunity qualifications done, qualified opportunities move to bid planning	Open, Cancel	Approach
03 Bid planning	Authoring of the solution proposal starts, sales force allocated	Open, Cancel	Approach
04 Proposal	Offer authored and reviewed internally, open questions resolved (with customer)	Open, Cancel	Sales presentation
05 Client decision	Offer delivered and presented to customer	Open, Cancel	Sales presentation/handling objections
06 Negotiation and signature	Contract negotiations	Won, Lost, Cancel	Handling objections/close

unacceptable to SWCo. Thus, the probability of loss is significant, and cancelled bids can be treated as losses.

Data and prediction variables

We combined our data from several sources. Opportunity history from the past four years, customer register data, customer satisfaction surveys, and client meeting records were combined into a harmonized dataset with almost 80000 records. From this dataset, we built 22 variables that are used for the predictions. Some of the variables are static, such as customer's industry, some of them are dynamic in the sense that they are time-related, like the age of an opportunity.

Predictive model

After experimenting with Naïve Bayes and two decision tree-based methods, and various subsets of the dataset, we built a model based on decision trees, using the R language package rpart (Therneau et al. 2013; Therneau and Atkinson 2017). Rpart uses unsupervised, tree growing–pruning approach, i.e. it first grows the tree to its maximum size and then removes the least important branches, based on given complexity parameters. We tested the model as follows, using both the whole history and several subsets, such as a certain customer segment or offering type.

1. Select randomly, and without replacement, 20% of the historical bids as a test set.
2. Build the model using the rest of the history.
3. Run the test set against the model, verify percentage of correct predictions.
4. Repeat steps 1–3 four more times, having a new test set each time.

Operationalization

Figure 1 presents a conceptual view of the proposed, production-stable pipeline prediction solution at SWCo. Once per month new data are extracted from source systems, harmonized, and stored in a relational database. Historical data are cumulated, and sales pipelines are stored as monthly snapshots. Next, predictions are run for each bid in current pipeline and the results are stored in the relational database.

The relational database provides data for reporting and visualization purposes. Sales teams, as well as business units, have a new vehicle for sales pipeline analysis. Moreover, as the pipeline snapshots are stored, it is easy to find out the actual accuracy of the past predictions at any given time. Comparing the outcome of closed opportunities with predicted outcomes gives factual data, e.g. for further development of the predictive model or cost/benefit analysis.

Prediction accuracy

The accuracy of the algorithm is verified every month by comparing predicted outcomes with actual outcomes from

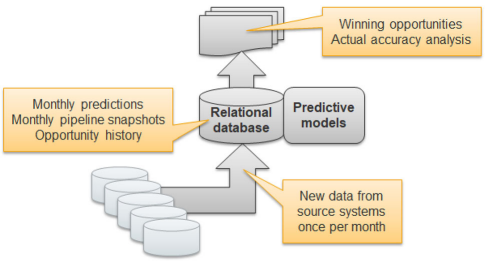


Fig. 1 SWCo's bid prediction concept

the previous four months. Based on the results, the model may be re-trained. Figure 2 illustrates the principle. Field evidence shows that the model's accuracy is 65–67%, when we look at the pipeline as a whole. By looking at the results by stage, we noticed that the accuracy varies, e.g. combined stages 02 (Qualification) and 03 (Bid planning) gives 78–86%. It is worth noting that stages 05 and 06 have a lower actual accuracy percentage (60% and 62%, respectively). At these stages, numerous interactions, such as phone calls and emails, take place between SWCo and the customer. Moreover, several other aspects, such as competitors' actions or personal selling abilities, play an important role. None of these are reflected in our data, which assumedly affects the accuracy.

The primary goal of our problem setting was to identify winning opportunities at early stages. Based on the field evidence, the model gives correct prediction roughly four times out of five, i.e. it produces desired results. The main caveat here is the risk of lost opportunities by elimination of bids, which the algorithm deems as no-wins, but which actually could be wins. In case of large bids this is a significant business risk. It turned out that excluding exceptionally large bids (Z-score $> 2\sigma$) actually leads to higher prediction accuracies in all stages. This finding provides an insight to decision making: rely on the algorithm on smaller bids, put more human effort on larger ones.

Considering the second part of our problem setting, the model evaluates the pipeline according to the same criteria. Thus, the model effectively creates a decision-making baseline for sales management and teams. However, the baseline could be enriched by collecting data in a more systematic way at the early stages, especially at stage 01 (Identification). In addition to qualitative data, collecting quantitative data with a standardized set of questions would help harmonizing the outcomes of the lead screening, which would support the review meetings at later stages. Moreover, it would assumedly lead to higher and less volatile accuracy in predictions at stage 01. Monat (2011) presents a set of commonly available lead characteristics, which could be used as a basis of the quantitative question set.

As mentioned, the costs at SWCo cumulate strongly in the later stages of the sales process. Eliminating non-promising opportunities at the early stages allows concentrating in potentially winning bids, and leaves room for seeking new opportunities. However, while this general statement makes sense, it raises a question, how do you assess, or demonstrate the value?

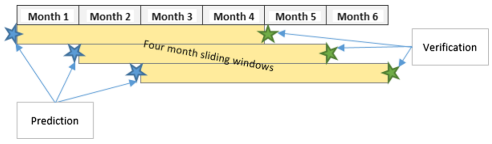


Fig. 2 Algorithm follow-up principle

Value assessment of the algorithm

As predictive algorithms have increasingly been applied to sales management, we first present a current literature overview of algorithmic applications in this field. We continue by demonstrating how classification matrices, combined with management accounting rules, can be applied to assess the business value of an algorithm.

Background: data and analytics in sales management

Sales force automation has been under investigation since the 1990 s, as we saw in the Introduction section. We searched the academic literature databases for studies dealing with predictive analytics and sales management. We identified five broad domains in recent research, where data and analytics have been applied. Table 2 presents these domains, and gives examples of studies that have explored the area.

Various different algorithmic approaches have been used. Naïve Bayes has been applied to four of the five domains: bid qualification, bid pricing, pipeline prioritization, and new sales opportunity identification (Greenia et al. 2014; Yan et al. 2015; Lawrence 2003; Megahed et al. 2015; Saxena et al. 2016). Variations of regression analyses—logistic, linear, kernel—have been applied particularly to sales team allocation, but also to new sales opportunity identification (Lawrence et al. 2010; Kawas et al. 2013; Varshney and Singh 2013). Other approaches include predicting bid pricing using Monte Carlo simulations or support vector machines (Akkiraju et al. 2014; Petruseva et al. 2016), pipeline prioritization using a combination of nearest neighbour method, logistic regression, neural networks, and decision trees (D'Haen and Van den Poel 2013).

Some of the studies (e.g. Lawrence et al. 2010; Greenia et al. 2014; Yan et al. 2015), reported business benefits, and many others implicitly contained proposals of potential business benefits. Nevertheless, the studies did not provide any guidance on how to measure the cost and benefits of the proposed algorithms.

Applying classification matrices

Comparing the predicted results with actual outcomes can be done using a classification matrix. We lay our approach to cost-sensitive learning foundations (Elkan 2001;

Table 2 Analytical approaches in sales management

Domain	Purpose/context	Examples
Bid qualification	Identify winning opportunities of the current sales pipeline using quantitative methods	Eichhoff and Maass (2014), Greenia et al. (2014), and Yan et al. (2015)
Bid pricing	Determine price that maximizes expected profit and winning probability of an opportunity	Lawrence (2003), Akkiraju et al. (2014), and Petruseva et al. (2016)
Pipeline prioritization	Identify opportunities of the current sales pipeline that are most likely to win	D'Haen and Van den Poel (2013), Greenia (2014), and Megahed et al. (2015)
New sales opportunities	Discover new prospects or neglected current opportunities based on customer's potential spending	Lawrence et al. (2010), D'Haen and Van den Poel (2013), and Saxena et al. (2016)
Sales team allocation	Allocate available sales resources so that revenue is maximized	Lawrence et al. (2010), Megahed et al. (2015), Kawas et al. (2013), and Varshney and Singh (2013)

Zadrozny and Elkan 2001). Table 3 presents a binary classification matrix for possible outcomes of our algorithm. If the algorithm predicted a no-win, and the actual outcome was also a no-win, the prediction was correct (true negative). Correspondingly, a true positive represents a correct prediction. False negatives and false positives are misclassified bids. *Ntn, Nfn, Nfp*, and *Ntp* represent the number of bids in each class. The total number of bids $Ntot = Ntn + Nfn + Nfp + Ntp$.

Predictions are not 100% accurate. Counting the share of true negatives and true positives of all bids gives us useful metrics for calculating the prediction accuracy percentage = $(Ntn + Ntp)/Ntot * 100$. The accuracy percentage constitutes a solid basis for business decision making. Knowing that the accuracy is, for example, 80% means that relying on the prediction gives expected the result four times out of five. While this accuracy may be unacceptable in domains like predicting rare diseases, it is highly useful in many practical business scenarios, including bid qualification.

Another useful metric is hit rate, i.e. the share of wins. Actual hit rate is easily interpreted from the classification matrix ($Ntp/Ntot * 100$). For predictions, we need to take into account the number of incorrectly predicted wins, i.e. false positives. The predicted hit rate can be calculated using the following procedure.

1. Calculate the percentage of predicted wins *Pwin* in the prediction (= predicted wins/bids in pipeline * 100)
2. Calculate the percentage of actual false positives ($Nfp/Ntot * 100$) in the past.
3. Expected hit rate = *Pwin*—actual false positives.

Table 3 A classification matrix for bid qualification

	Actual no-win	Actual win
Predicted no-win	True negative (*Ntn*)	False negative (*Nfn*)
Predicted win	False positive (*Nfp*)	True positive (*Ntp*)

One business consideration is the risk of lost opportunities due to misclassification as no-win (false negatives). A classification matrix with actual, historical data provide data for estimation. The percentage of actual false negatives ($Nfn\% = Nfn/Ntot * 100$) against predictions tells us the risk level. If the *Nfn%* has been e.g. 5–6% in the past, we can assume that it will hold also in the near future. In other words, for each 100 opportunities, we will most probably miss five or six due to incorrect prediction. This is data-based, hard evidence, at least when compared to personal opinions and valuations that may change over time, and case by case. Thus, *Nfn%* offers a coherent basis for decision making. Moreover, personal judgement, or expertise may always override predictions, as long as the elimination of bids is not fully automated.

The classification matrix also helps concretizing the value of the algorithm in monetary terms, hours, or percentage of total costs. A straightforward method is to allocate all sales costs according to the bid distribution in the matrix. Once the allocations are done, we can estimate the potential effects of the algorithm in monetary terms. The true positive bids generate new revenue, and therefore, the costs are justified from the business perspective. In this sense, we can ignore the sales costs of these bids (*Ntp*). The true negatives are no-win bids, as was predicted. The sales costs of these bids (*Ntn*) represent the potential from the efficiency perspective. Costs can be translated to hours using a standard hourly rate.

This provides insights for decision making. Depending on the case, the true negatives potentially represent a significant efficiency impact. For example, if we strictly follow the prediction, and eliminate all of the predicted no-win bids, we "spare" these sales costs. In practice, this means that we can re-allocate part of our sales force for securing the promising opportunities, or searching for new opportunities. The calculation process goes as follows:

1. Allocate total sales costs to a cost matrix according to actual outcomes from the past.

2. Perform predictions, i.e. run the algorithm against current sales pipeline.
3. Calculate the percentage of predicted no-wins in the prediction.
4. Calculate no-win sales costs = predicted no-wins percentage * total sales costs.
5. Potential hours for re-allocation = no-win sales costs/standard hourly rate.

Obviously, applying the procedure to the early stages of the pipeline maximizes the benefits, as most of the sales costs cumulate in the later stages. Moreover, it is reasonable to include only those bids, where the expected closing date is in the near future. For instance, consider bids at stages 01, 02, and 03, with the closing date within the next 3 months.

Guidelines and metrics

Some takeaways and guidelines can be drawn from the case study. We suggest developing key metrics for evaluating not only the performance of the algorithm, but also its business effects. We start with the guidelines and continue with metrics.

Be bold in bid elimination

When the accuracy of the algorithm has been verified, trust the results. Following the predictions allows focusing on potential bids. However, it should be noted that in practice there are circumstances, e.g. based on competition, where it is reasonable to go against a prediction. Moreover, it is advisable to apply expert consideration in order to mitigate the risk of losing potential opportunities. However, overriding the prediction should be a well-justified, case-by-case decision.

Defer the implementation

Defer the implementation of a production-ready solution until the algorithm has been proven to produce results. The data are different in each case, which means that results are not guaranteed. The less effort there is, the less business risk. Starting with a mock-up or prototype minimizes the risk. We took snapshots of the data from the source systems, focused on identifying potential variables, and developing the predictive model. That said, be prepared to manage expectations, when a prototype produces promising results. Algorithms are a hot, interesting topic, but for a production-stable solution you need the whole iceberg, not just the tip. Automating data gathering, ensuring data quality, driving organizational change, processes for prediction, and metrics for follow-up require quite a lot of work and coordination.

Play with the data

Understanding the data inside and out is essential for defining the prediction variables, interpreting the results, and identifying potential biases. Look for subsets of data, where the accuracy of the algorithm is at its best, discuss with domain experts about data recording practices etc. This is a learning experience, characterized by a trial and error process.

Create an initial benchmark

Create an initial benchmark for accuracy, by comparing predictions to actual results with classification matrices. This should be done without sales personnel involvement. The predictions affect the behaviour of your sales force, which potentially leads to biased results. Once you have a "neutral" baseline, you have a reliable reference point for follow-up.

Measure the effects regularly

This is essential for ensuring the reliability of the predictions. Using appropriate metrics ensures that we identify factors such as changes in data recording procedures, sales force behaviours, or changes at markets that affect our algorithm.

Table 4 presents a summary of metrics based on the classification matrix approach discussed above. Note that the metrics are indicators for trends and act as tools for continuous improvement instead of being exact accounting figures. The measurement interval is 1 month. All metrics aim to support sales management's decision making. Regular measurement, follow-up of the trend, and comparison to the initial benchmark allows us to be confident that the predictions constitute a sound basis for decisions.

Discussion and conclusions

We propose applying classification matrices as a separate step in similar fashion meta-learning algorithms apply cost matrices (e.g. Domingos 1999; Sheng and Ling 2007), instead of including the monetary elements, like the re-allocation potential, or the costs of lost opportunities in the predictive algorithm. This has implications that are important in practical implementations. The separation allows the usage of numerous "out-of-the-box" algorithms that are cost-insensitive. In most cases, fewer modifications equal less implementation costs. It also enables the

Table 4 Trend metrics for algorithm follow-up

Metric	Purpose/application
Prediction accuracy percentage (PA%) = *(Ntn + Ntp)/Ntot* * 100	Used to measure the overall performance of the algorithm. Calculate PA% monthly using actuals, e.g. from the last four months
	In the near future you can assume the PA% to be roughly the same as the one calculated from recent actuals
Potential for salesforce re-allocation (P%) = *Ntn/Ntot* * 100	Used to measure the efficiency of sales. Calculate P% monthly using actuals, e.g. from the last four months. You can also translate this metric into hours using the 5-step calculation process above
Predicted hit rate (PH%) = *Pwin − Nfp/Ntot* * 100	Used to approximate the expected wins. Calculate PH% monthly using actuals, e.g. from the last four months. By observing the previous predictions and actuals, you can approximate the hit rate in the near future. If your pipeline contains outliers, you might remove exceptional bid values, i.e. Z-score outside +/− 2σ
Percentage of lost opportunities (LO%) = *Nfn/Ntot* * 100	Used to approximate the risk of potentially lost opportunities when eliminating bids based on the algorithm. Calculate LO% monthly using actuals, e.g. from the last four months. Expect the LO% to be roughly the same in the near future
Trends	For each metric, record the results—predicted and actual—monthly and follow up the trend. Monitoring the trends gives indications regarding how your organizational performance develops over time

separation of algorithm implementation and management accounting, which eases the division of work. Management accounting expertise is often available in-house, whereas specialists capable of building predictive analytics applications are a scarce resource. Loose coupling between the algorithm and costing components eases change management, e.g. sales cost allocation rules may be changed without touching the algorithm. Moreover, the separation allows applying the method to existing solutions.

Algorithms or bid qualification decisions are, of course, just a fraction of sales prediction. There are several related organizational and personal factors, like management support, or cross-functional communications. Davis and Mentzer (2007) provide a framework that views sales forecasting management as an organizational capability. In this view, our method can be seen as a piece of technology and information processes that can help companies create a shared interpretation of the current sales pipeline. This should lead to more informed decisions, and more accurate sales forecasts.

Moreover, individual factors are always involved in sales predictions. Bonney et al. (2016) explore how sales persons valuate options. Their results show that sales persons are prone to use available resources for overvaluing (compared to non-salespeople), in order to keep their options open for the future. This can be costly, but oftentimes there are not enough fact-based data for informed reallocation decisions. An algorithmic bid qualification method offers a baseline for discussion and decisions. The value potential here lies in efficiency improvements.

It is easy to forecast that the usage of algorithms is becoming more and more common in sales management. This underlines the importance of data: predictive algorithms live on data. Data often reside in silos, meaning that gathering, harmonizing, and processing the data represents a significant effort. Moreover, a great deal of potentially useful data are unstructured, like sales presentations and documents. One potential improvement for producing high quality data for sales predictions is to first use a systematic approach for metrics definition (e.g. Monat 2011; Neely et al. 1997), and then record the data during the sales process. Another approach would be to utilize the metadata generated by the ever-increasing B2B sales platforms.

As a conclusion, we presented a case study, where classification matrices and management accounting practices are used to assess the effects of a machine-learning algorithm. Related guidelines and metrics aim to help practitioners. We developed the method in a case study dealing with the bid qualification domain. Besides that, the method can be applied to other problem domains. In the sales management domains, prioritizing the pipeline and identifying new opportunities are obvious examples. We can also imagine applications in many other classification problems, like fraud detection or predictive maintenance. Being able to assess and concretize the value with proper metrics is an important factor in business decision making. The method presented in this article can be used both a priori, e.g. for business case evaluation, and a posteriori, for assessing the realized value.

References

Akkiraju, R., M. Smith, D. Greenia, S. Jiang, T. Nakamura, D. Mukherjee, et al. 2014. On pricing complex IT service solutions. In *IEEE Global Conference* (*SRII*), 2014 *Annual SRII*, 55–64.

Amatriain, X. 2013. Beyond Data: From User Information to Business Value Through Personalized Recommendations and Consumer Science. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 2201–2208. ACM

Beam, C. 2006. How to Assess Your Sales Pipeline. *Consulting to Management* 17 (2): 18–21.

Bhattacharyya, S. 2014. Improving Inventory Demand Forecasting by Using the Sales Pipeline: A Case Study. *The Journal of Business Forecasting* 33 (1): 7–11.

Bonney, L., C.R. Plouffe, and M. Brady. 2016. Investigations of Sales Representatives Valuation of Options. *Journal of the Academy of Marketing Science* 44 (2): 135–150.

Cooper, M.J., and C.S. Budd. 2007. Tying the Pieces Together: A Normative Framework for Integrating Sales and Project Operations. *Industrial Marketing Management* 36 (2): 173–182.

D'Haen, J., and D. Van den Poel. 2013. Model-Supported Business-to-Business Prospect Prediction Based on an Iterative Customer Acquisition Framework. *Industrial Marketing Management* 42 (4): 544–551.

Davis, D.F., and J.T. Mentzer. 2007. Organizational Factors in Sales Forecasting Management. *International Journal of Forecasting* 23 (3): 475–495.

Domingos, P. 1999. Metacost: A General Method for Making Classifiers Cost-Sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155–164. ACM.

Dubinsky, A.J. 1981. A Factor Analytic Study of the Personal Selling Process. *Journal of Personal Selling & Sales Management* 1 (1): 26–33.

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

Dwyer, S., J. Hill, and W. Martin. 2000. An Empirical Investigation of Critical Success Factors in the Personal Selling Process for Homogenous Goods. *Journal of Personal Selling & Sales Management* 20 (3): 151–159.

Eichhoff, J.R., and W. Maass. 2014. Functional Design Space Representations for Lead Qualification Situations. In *Design Computing and Cognition'12*, 529–547.

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning. *International Joint Conference on Artificial Intelligence* 17: 973–978.

Flanagan, P. 1995. Getting the Paper Out of the Marketing & Sales Pipeline. *Management Review* 84 (7): 53–55.

Greenia, D.B., M. Qiao, and R. Akkiraju. 2014. A Win Prediction Model for IT Outsourcing Bids. In *IEEE Global Conference* (*SRII*), *2014 Annual SRII*, 39–42.

Kawas, B., M.S. Squillante, D. Subramanian, and K.R. Varshney. 2013. Prescriptive Analytics for Allocating Sales Teams to Opportunities. In *IEEE 13th International Conference on Data Mining Workshops* (*ICDMW*), 211–218.

Lawrence, R.D. 2003. A Machine-Learning Approach to Optimal Bid Pricing. In *Computational Modeling and Problem Solving in the Networked World*, 97–118.

Lawrence, R., C. Perlich, S. Rosset, I. Khabibrakhmanov, S. Mahatma, S. Weiss, et al. 2010. Operations Research Improves Sales Force Productivity at IBM. *Interfaces* 40 (1): 33–46.

Ledingham, D., M. Kovac, and H.L. Simon. 2006. The New Science of Sales Force Productivity. *Harvard Business Review* 84 (9): 124–132.

Long, M.M., T. Tellefsen, and J.D. Lichtenthal. 2007. Internet Integration into the Industrial Selling Process: A Step-by-Step Approach. *Industrial Marketing Management* 36 (5): 676–689.

Lukes, T., and J. Stanley. 2004. Bringing Science to Sales. *Marketing Management* 13 (5): 36–41.

Megahed, A., G.-J. Ren, and M. Firth. 2015. Modeling Business Insights into Predictive Analytics for the Outcome of IT Service Contracts. In *IEEE International Conference on Services Computing* (*SCC*), 515–521.

Monat, J.P. 2011. Industrial Sales Lead Conversion Modeling. *Marketing Intelligence & Planning* 29 (2): 178–194.

Moncrief, W.C., and G.W. Marshall. 2005. The Evolution of the Seven Steps of Selling. *Industrial Marketing Management* 34 (1): 13–22.

Morgan, N.A., and L.L. Rego. 2006. The Value of Different Customer Satisfaction and Loyalty Metrics in Predicting Business Performance. *Marketing Science* 25 (5): 426–439.

Neely, A., H. Richards, J. Mills, K. Platts, and M. Bourne. 1997. Designing Performance Measures: A Structured Approach. *International Journal of Operations & Production Management* 17 (11): 1131–1152.

Petruseva, S., P. Sherrod, V.Z. Pancovska, and A. Petrovski. 2016. Predicting Bidding Price in Construction Using Support Vector Machine. *TEM J* 5 (2): 143–151.

Saxena, N., S. Arumugam, and C. Roy. 2016. Deal RADAR (Real-Time Abandonment Detection and Recourse). *Journal of Information and Optimization Sciences* 37 (5): 819–838.

Sheng, V., and C. Ling. 2007. Roulette Sampling for Cost-Sensitive Learning. *Machine Learning* 724–731

Therneau, T., B. Atkinson, and B. Ripley. 2013. Rpart: Recursive Partitioning. R Package Version 4.1-3. https://cran.r-project.org/web/packages/rpart/rpart.pdf.

Therneau, T.M., and E.J. Atkinson. 2017. *An Introduction to Recursive Partitioning Using the RPART Routines*. Rochester: Mayo Foundation.

Varshney, K.R., and M. Singh. 2013. Dose-Response Signal Estimation and Optimization for Salesforce Management. In *IEEE International Conference on Service Operations and Logistics, and Informatics* (*SOLI*), 328–333.

Wixom, B.H., and J.W. Ross. 2017. How to Monetize Your Data. *MIT Sloan Management Review* 58 (3): 9–13.

Yan, J., M. Gong, C. Sun, J. Huang, and S.M. Chu. 2015. Sales Pipeline Win Propensity Prediction: A Regression Approach. In *IEEE IFIP/IEEE International Symposium on Integrated Network Management* (IM), 854–857

Ylijoki, O., and J. Porras. 2017. What Managers Think about Big Data. *International Journal of Business Information Systems* (forthcoming)

Zadrozny, B., and C. Elkan. 2001. Learning and Making Decisions When Costs and Probabilities Are Both Unknown. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 204–213. ACM.

Ossi Ylijoki M.Sc. (Tech.) is a doctoral student at Lappeenranta University of Technology. He has 15+ years of experience in knowledge management, business intelligence and data warehousing in various expert and management positions in business life. His research interests are big data, information systems as a value driver, and software engineering.

Publication VI

Ylijoki, Ossi and Porras, Jari
A Recipe for Big Data Value Creation

Reprinted with permission from
Business Process Management Journal
Accepted for publication

A recipe for big data value creation

Ossi Ylijoki and Jari Porras
LUT School of Engineering Science, Lappeenranta University of Technology, Lappeenranta, Finland

Abstract

Purpose – The purpose of this paper is to present a process-theory-based model of big data value creation in a business context. The authors approach the topic from the viewpoint of a single firm.
Design/methodology/approach – The authors reflect current big data literature in two widely used value creation frameworks and arrange the results according to a process theory perspective.
Findings – The model, consisting of four probabilistic processes, provides a "recipe" for converting big data investments into firm performance. The provided recipe helps practitioners to understand the ingredients and complexities that may promote or demote the performance impact of big data in a business context.
Practical implications – The model acts as a framework which helps to understand the necessary conditions and their relationships in the conversion process. This helps to focus on success factors which promote positive performance.
Originality/value – Using well-established frameworks and process components, the authors synthetize big data value creation-related papers into a holistic model which explains how big data investments translate into economic performance, and why the conversion sometimes fails. While the authors rely on existing theories and frameworks, the authors claim that the arrangement and application of the elements to the big data context is novel.

Keywords Big data, Business value, Capabilities, Data assets, Digital transformation, Process theory
Paper type Research paper

Received 12 March 2018
Revised 23 August 2018
Accepted 24 September 2018

1. Introduction

Digital transformation is a current megatrend. Consequently, vast amounts of data are generated in what is known as datafication (Lycett, 2013; Mayer-Schönberger and Cukier, 2013). Datafication leads to several implications. Among them are inevitably increasing demand of information technology (IT), i.e. hardware and software investments required to process the data, and the fact that companies increasingly utilize information, derived from the data that IT enables them to gather from various internal and external sources. Data have become an asset.

The effects of applying IT in a business context have been studied extensively. It is commonly accepted (Barua *et al.*, 1995; Melville *et al.*, 2004; Soh and Markus, 1995) that the effects of IT cannot be measured directly in terms that describe organizational performance, such as profitability or productivity. Instead, according to the studies, IT influences processes and functions which then have performance impacts that maybe positive or negative. The value of information in a business context has also been widely studied (Dehning *et al.*, 2003; Hitt and Brynjolfsson, 1996; McAfee and Brynjolfsson, 2012; Porter and Millar, 1985) and the positive effects have been shown repeatedly. The effective utilization of information is undeniably a source of competitive advantage. Nevertheless, while the impact in general is clear, single initiatives may still fail. Value creation processes are at the core of our study. In Section 2, we discuss the theoretical foundations that we build on.

Scholars have studied big data from various angles. The majority of the research works have focused on technology, e.g. processing technologies or algorithms[1]. However, there is also a significant body of knowledge viewing the topic from a business perspective. As with IT in general, big data adds value indirectly. For example, Sharma *et al.* (2014) investigated how analytics influences decision making, which in turn affects the firm's performance. Accordingly, Wamba *et al.* (2017) looked at the impact of analytic capabilities, Mithas *et al.* (2011) underlined information management capabilities, and Merino *et al.* (2016) discussed data quality aspects. These, among other factors, influence the data-driven approaches applied in organizations.

Business Process Management Journal

1463-7154
DOI 10.1108/BPMJ-03-2018-0082

Organizations are complex structures and value creation involves various stakeholders, functions and processes. This brings in uncertainty. It is not guaranteed that a project will deliver its desired outcomes. Some of big data initiatives will inevitably fail, whereas others may lead to sustained competitive advantage. Under conditions of uncertainty, where the outcome sometimes occurs and sometimes does not, process theories can provide powerful explanations (Markus and Robey, 1988). In Section 3, we present a comprehensive overview of current big data research and reflect on how this links to process theory principles.

The purpose of this research is to compose a process-theory-based model of big data value creation. We present the model in Section 4. Using theoretically solid building blocks from other scholars, we add current big data research and arrange the elements in a novel way. The model contributes to big data research by offering a holistic, end-to-end view of big data value creation processes.

2. Theoretical context

This section briefly discusses the theoretical foundations relevant to the context of our study. We present two established and widely studied frameworks that conceptualize the value creation process. The first is a hierarchical structure (Ackoff, 1989; Zeleny, 1987) describing how data cumulates to form knowledge and the second one (Rayport and Sviokla, 1995) is a sequential process addressing steps required in data processing, starting from data gathering continuing on to distribution to end users. Both frameworks are frequently applied in big data research (e.g. Braganza *et al.*, 2017; Bumblauskas *et al.*, 2017; Miller and Mork, 2013).

A common approach in big data research is to use variance theories. Section 3.1 (see Table II) contains several examples of studies applying variance theory (e.g. Akter *et al.*, 2016; Müller and Jensen, 2017). Variance theories can produce excellent results, as shown by many big data studies focusing on certain aspects in the value creation process, such as the impact of analytics or data quality. However, the broader the view, the more difficult it becomes to manage the increasing number of conditions that influence the processes. We propose a process-theory-based perspective as an alternative or complimentary perspective for a better understanding of the big data value creation processes as whole. Thus, we begin with an introduction to process theories before continuing with the frameworks. Our aim is to extend a process-theory-based value model with the data value creation frameworks.

2.1 IT and business value – variance and process theory views

Process theories differ from variance theories in several ways. The differences are summarized in Table I, adapted from Mohr (1982) and Soh and Markus (1995). Markus and Robey (1988)

	Process theory	Variance theory
Definition	Causation consists of necessary conditions occurring in a particular sequence in which change and random events play a role	The cause is necessary and sufficient to produce the effect
Outcomes	Discrete occurrences	Variables
Role of time	Crucial, conditions must occur in a particular sequence	Meaningless, conditions can occur in any order
Assumptions	In addition to necessary conditions, external factors must be favorable for the outcomes to occur	Outcomes occur whenever necessary and sufficient conditions are present
Logical form	If not-X (necessary conditions), then not-Y (outcome) Cannot be extended to "more X" or "more Y"	If X (independent variable, necessary and sufficient conditions), then Y (dependent variable) If more X, then more Y

Table I.
Differences between process and variable theories

discussed the differences: the outcome of a variance theory is a variable, whereas a process theory produces a discrete occurrence. Variance theories assume that the outcomes occur invariably as long as the necessary conditions are present. Therefore, more input leads to more output. Process theories assume that in addition to the necessary conditions, external factors affect the process and, thus, the necessary conditions are not enough to ensure the outcomes. Moreover, process theories consider time as crucial, whereas variance theories are independent of time.

For most organizations big data adoption equals business transformation, which effectively means uncertainty. Process theories tend to perform well in situations where the outcome is uncertain (Markus and Robey, 1988). Therefore, we believe that in addition to variance theory applications, process-theory-based approaches might be useful for big data scholars as well.

Soh and Markus (1995) presented a process-theory-based model that explains how IT creates value. The model consists of three processes: the first process converts IT expenditure into IT assets, the second process converts assets into IT impacts, and the third process connects the impacts to organizational performance. The processes are sequential in time, and the outcome of each process is nondeterministic. For example, increasing IT expenditure does not necessarily convert to more or better IT assets (e.g. due to inefficient IT management), not to mention improved organizational performance. That is, expenditure is a necessary, but not sufficient input to ensure the desired outcome of the conversion process.

2.2 Data and business value

Zeleny (1987) and Ackoff (1989) presented conceptualized hierarchies for transforming data into wisdom. Since then the hierarchy concept has been widely used and studied (Rowley, 2007). A common interpretation of the hierarchy contains four levels: data, information, knowledge and wisdom, hence, the acronym DIKW. Data consist of raw facts, or symbols, that "know nothing" at the bottom of the hierarchy. Information is data given a context, typically answering to questions like who, what, where and when. The next level, knowledge, is defined as information combined with understanding and capability, or simply know-how. Wisdom is placed at the top of the hierarchy, accumulating knowledge to create intuition and interpretations, i.e. it has more to do with human capabilities than with systems. Higher levels of the hierarchy depend on the lower levels, i.e. wisdom cannot cumulate without knowledge and so on. According to Rowley (2007), the four abovementioned levels are included in virtually all studies dealing with the hierarchy. Moreover, they are always arranged in the same order. It should also be noted that although the terms data, information and knowledge are sometimes used interchangeably, they are distinct concepts (Pigni *et al.*, 2016).

Another useful framework is the virtual value chain (VVC). Building on Porter's value chain, Rayport and Sviokla (1995) defined the VVC framework which consists of a five-step value creation process. The steps are gather, organize, select, synthetize and distribute. This framework generalizes the stages that are required when creating value from data. The stages (or steps) can be identified in real-world applications. Thus, this framework provides a mental model that connects implementations to the value creation processes.

Big data challenges IT to develop new technology and human-related skills and capabilities. For example, Laney (2001) declared that the 3Vs (volume, variety, velocity) that characterize big data requires IT to innovate new data management technologies. Janssen *et al.* (2017) stated that due to the technical properties – the 3Vs + veracity – big data require special attention and governance mechanisms to ensure data quality. Correspondingly, Merino *et al.* (2016) underlined that conceptual and technological aspects not only differentiate big data from traditional business intelligence but also makes data

quality management difficult. Big data has implications to information systems research as well: e.g. Abbasi *et al.* (2016) pointed out numerous research opportunities, such as paradigmatic considerations and adapting models proposed in the era of scarce data to the current big data era.

However, what really differentiates big data from IT are the nature and impact of big data. Big data is a disruptive innovation (e.g. Davenport, 2014; Manyika *et al.*, 2011; Mayer-Schönberger and Cukier, 2013) that potentially has a huge consequences on business (e.g. Manyika *et al.*, 2011; McAfee and Brynjolfsson, 2012; Van't Spijker, 2014; Weill and Woerner, 2015). Making use of big data requires companies to monitor data flows (instead of static data sets) and integrating analytics into core business processes (Davenport *et al.*, 2012). Thus, big data represents a paradigm shift from the IT side toward the business. Big data initiatives require new organizational and managerial capabilities that drive the transformation toward data-driven business (e.g. Anand *et al.*, 2016; Comuzzi and Patel, 2016; Sandberg and Aarikka-Stenroos, 2014).

2.3 Reflecting the theory in a big data context

The VVC framework is perfectly applicable in a big data context. Indeed, several authors have applied it in their research. Braganza *et al.* (2017) presented an archetype big data business process that builds on the VVC framework. Miller and Mork (2013) presented a variation of the VVC, linking the value chain with current standards and technologies. Janssen *et al.* (2017) presented yet another application of the VVC in their case study regarding factors that influence decision making.

The DIKW hierarchy is useful in business contexts in general, as it conceptualizes the value creation process in an abstract model that helps build an understanding of the process. Several scholars have employed the hierarchy in a big data context. For example (Abbasi *et al.*, 2016; Bumblauskas *et al.*, 2017) presented frameworks which explain the steps of transforming information into knowledge using the hierarchy. Although VVC and DIKW frameworks are separate perspectives they can be seen related: data and information are related to gathering, organizing and partly selection (context) stages, while knowledge and wisdom are processes represented by synthetizing and distribution stages of the VVC framework. These can be further linked to the process-theory-based IT value creation model of Soh and Markus (1995).

Figure 1 summarizes the theoretical background of our research. All three models that we build on are sequential in nature. The models overlap or limit each other. However, the overlapping shown in Figure 1 is indicative. The models shed light on the value creation processes from three different perspectives. The information systems view, i.e. VVC framework, provides building blocks for software implementations, the DIKW perspective explains the knowledge creation process, and the IT value creation theory enlightens the related processes.

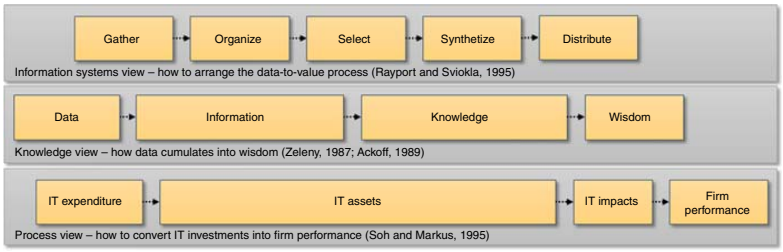


Figure 1.
Theoretical context of the study

When we reflect on the big data value creation with the process theory view, as shown in Figure 1, it becomes intuitively clear that converting big data investments into improved performance has similarities to the IT value creation process presented by Soh and Markus (1995). However, for understanding the process in a big data context, we need to identify the big data specific processes, their outcomes and relationships as well as external factors which may influence the outcomes. We are especially interested in converting the assets into impacts with regard to big data value creation.

3. From big data assets to big data impacts

In this section, we present the topics and conceptualize the core processes related to big data value creation. We build on the DIKW and VVC frameworks, linking our conceptual views to these models. Moreover, we discuss the necessary and sufficient conditions, reflecting the literature and our conceptualization in the process theories. We start with an overview of current big data literature relevant to our research.

3.1 Current big data literature

Scholars are actively working on various aspects of big data. Although the majority of the studies focus on technical aspects, an increasing number of papers deal with big data implications in a business context. We searched literature databases (EBSCO, Scopus, Web of Science) for papers relevant to our study, i.e. papers where title, abstract or keywords contains "big data" and ("business value" or "value creation" or "business innovation" or "firm performance" or "digital data stream"), limiting the search to articles written in English language and business categories. Moreover, we performed two-tier forward and backward snowballing for the papers which we considered relevant based on the title and abstract. Although no literature search can be exhaustive, we believe that Table II contains a representative intersection of current big data literature for business contexts. We identified the most discussed areas (data and technology, capabilities and big data impacts) from the papers. Based on this analysis, we arranged the literature according to topics, shown in the first and second columns in Table II. Some of the papers are listed more than once, as they discussed more than one topic. In addition, we identified factors that could either promote or demote the value creation process. These contingency factors are shown in the last column of Table II.

The literature, topics and concepts presented in Table I, as well as their relations to big data value creation processes, are discussed in the following sections. We arrange the literature around process models that together explain big data value creation from the process theory perspective. Because the literature points out the importance of big-data-related capabilities and big data impacts, we focus on the processes related to these aspects. The first category, data and technology, represents the investments and assets embodied in the process theory discussed above. Capabilities are also related to the assets of the process theory and are seen as important phase for achieving the impacts. Starting from big data assets, we move on to capabilities required to convert the assets into impacts.

3.2 Big data assets

Companies invest in data management activities, software and hardware, expecting to convert the investments made into valuable assets. The data meet the criteria of an intangible asset (IFRS, 2015) with the exception of value measurement – data are identifiable, non-monetary, non-physical, potentially valuable resources that a firm produces or harvests from different sources. In order to derive information assets from (big) data, they must be harvested and put into a relevant context (Piccoli and Pigni, 2013). Harvesting relates to actions by which the data are extracted from its sources.

Topic	Sub-topics	Focal areas	+/− factors
Data and technology	Big data (e.g. Gandomi and Haider, 2015; Kitchin and McArdle, 2016; Laney, 2001; Pigni *et al.*, 2016; Ylijoki and Porras, 2016) Technology (e.g. Boncea *et al.*, 2017; Dutta and Bose, 2015; Krumeich *et al.*, 2014)	Characteristics of big data, where it comes from, or how it can be used to add value Technical processing of big data, e.g. data management frameworks and analytical tools	Data and technology management (e.g. Alguliyev *et al.*, 2017; Alharthi *et al.*, 2017; Boncea *et al.*, 2017; Piccoli and Pigni, 2013; Tiwana, 2014)
Capabilities	Analytics (e.g. Akter and Fosso Wamba, 2016; Arora and Malik, 2015; Najjar and Kettinger, 2013; Fosso Wamba *et al.*, 2017) Innovation (e.g. Brynjolfsson and McAfee, 2012; Gobble, 2013; Hartono and Sheng, 2016; Iddris, 2016; Jetzek *et al.*, 2014; Zhan *et al.*, 2017) Information management (e.g. Dutta and Bose, 2015; Mithas *et al.*, 2011; Tallon *et al.*, 2013)	Technical and human-related skills Organizational capabilities	Data quality (e.g. Ardagna *et al.*, 2016; Chae *et al.*, 2014; Hazen *et al.*, 2017; Janssen *et al.*, 2017; Merino *et al.*, 2016; Vidgen *et al.*, 2017) Resource availability (e.g. Davenport, 2014; Janssen *et al.*, 2017; Shah *et al.*, 2012)
Big data impacts	Decision making (e.g. Janssen *et al.*, 2017; Manyika *et al.*, 2011; Sharma *et al.*, 2014) Operational efficiency (e.g. Bärenfänger *et al.*, 2014; Dutta and Bose, 2015; Roden *et al.*, 2017) Product/service innovation (e.g. Davenport, 2014; Gobble, 2013; Manyika *et al.*, 2011; Mayer-Schönberger and Cukier, 2013) Business model innovation (e.g. Chen *et al.*, 2011; Iivari *et al.*, 2016; Van't Spijker, 2014)	Organizational change management Management perceptions Industry/ecosystem aspects	Data maturity (e.g. Anand *et al.*, 2016; Comuzzi and Patel, 2016; Dutta and Bose, 2015) Organization culture (e.g. Anand *et al.*, 2016; Sandberg and Aarikka-Stenroos, 2014; Shah *et al.*, 2012) Competition (e.g. Huberty, 2015; Pousttchi and Hufenbach, 2014; Weill and Woerner, 2015) Privacy/security factors (e.g. Clarke, 2016; Newell and Marabelli, 2015; Sullivan, 2014) Regulation (e.g. Keen *et al.*, 2013; Truyens and Van Eecke, 2014)

Table II.
Topical categorization of big data literature

Contextualizing the data typically involves augmenting the data points with additional attributes, e.g. a timestamp, location or client number. In other words, data are converted into a liquid form (Lycett, 2013) that can be moved and used outside its original context.

There are numerous different big data definitions (Gandomi and Haider, 2015; Mayer-Schönberger and Cukier, 2013). The most common elements, or dimensions, of the definitions include volume, velocity and variety, followed by veracity and value (Ylijoki and Porras, 2016). We introduce the dimensions below. It should be noted that the following merely presents commonly used characterizations of the dimensions – we are not trying to define what big data is. All these Vs are regularly used in the current big data literature. We could say that they constitute the building blocks of big data assets:

- Volume refers to the increasing amount of data (Laney, 2001). A typical approach is to express the volume by using the total size using gigabytes, terabytes or petabytes (Chen *et al.*, 2012; Frankel, 2012; Lamba and Singh, 2017; Manyika *et al.*, 2011). Kitchin and McArdle (2016) noted that huge amounts of data can result either from large files, such as video clips, or from a huge number of observations, e.g. from sensor readings.

- Velocity refers to the increasing pace at which data are generated (Laney, 2001). The pace may be a steady data flow (Gandomi and Haider, 2015), such as real-time GPS coordinate readings once per second. On the other hand, velocity can refer to volatile, periodic or even unpredictable data flows (Kitchin and McArdle, 2016), such as monthly updates, or a data burst on social media or phone networks caused by a major accident.
- Variety refers to data formats and structures (Laney, 2001). Some of the data are structured, such as in records in a relational database, but the vast majority of the data are in non-structural formats (Cukier, 2010; Gantz and Reinsel, 2011). Managing semi-structured data, for instance social media posts, or unstructured data, such as text or video, requires different approaches than dealing with structured data.
- Veracity relates to "the importance of addressing and managing for the uncertainty inherent within some type of data" (Schroeck *et al.*, 2012). Forecasting the weather or economic factors for the next year is example of this uncertainty. Moreover, veracity covers data quality aspects, which are, of course, an important factor for instance in decision making (Hazen *et al.*, 2014).
- The fifth V, value, was proposed by Gantz and Reinsel (2011). Value is clearly linked to the context where the data are used. In the context of our study, value equals economic value, which is ultimately realized as improved firm performance.

The focus of our research is on data, not technology. However, big data cannot be converted into information without technology assets. Big data may require significant investments in technology (Anand *et al.*, 2016). New technologies advance at a fast pace, and a wide variety of big-data-related technology stacks have emerged in recent years. Technical frameworks like Hadoop (http://hadoop.apache.org) or Amazon EMR (https://aws.amazon.com/emr/) emphasize data management and technical implementation. They do not consider other aspects such as the business or organization, but they provide companies with building blocks for technology platforms.

Technical frameworks essentially implement the data management steps (gather, organize) and may provide tools for the next two steps (select, synthetize) of the VVC framework (Rayport and Sviokla, 1995). In practice, the terminology and the number of steps varies, but the generic approach can be identified just by looking at the data flows and processing.

From the data point of view, creating assets realizes the first step of DIKW hierarchy (Ackoff, 1989), i.e. it converts data into information that is in a liquid form. This makes it possible to apply analytics, and use the information for other secondary purposes (Mayer-Schönberger and Cukier, 2013) that can potentially add value.

Guo *et al.* (2017) presented an automated competitor analysis tool that provides a concrete example, how to apply big data assets and analytical capabilities to support operational decision making. Their approach is to automatically harvest data from external sources and then apply natural language processing and machine learning algorithms to monitor firm's market position and competitors. With regard to Vs, their data are obviously in various formats, requiring quite a lot processing and "clean-up" before use. Although volume and velocity are moderate in terms of big data, automation makes it possible to harvest and process more data in almost real-time. The role of veracity in this setting relates mainly to the accuracy of the analytics (a special case of data quality). The fifth V, value, comes from cost efficiency and up-to-date, high-resolution competitor analyses that enable better decisions. This example demonstrates also that big data and analytics require competencies that are of short supply in many enterprises. For example, implementing an automated data harvesting from various external sources requires information management skills. Applying natural language processing and machine learning algorithms requires mastering concepts and tools, i.e. analytic capabilities.

3.3 Capability creation process

In order to make use of the information, firms must develop resources and capabilities related to the assets. In the resource-based view of the firm (Wernerfelt, 1984), resources and capabilities explain the firm's competitive advantage. As an example, a firm must have a hardware environment suitable for big data processing (tangible resources), develop analytical algorithms (intangible resources), and develop a data-oriented organization culture (organizational capability). Developing these competencies can be seen as a process. Teece (2007) introduced the concept of dynamic capabilities, which explains how companies renew their competencies when adapting to business turbulence. In addition to technical skills, the extant big data literature considers other types of capabilities as important outcomes, including analytical skills, information management capabilities and innovation capabilities (see also Table I). Figure 2 illustrates the capability creation process.

Analytical skills and capabilities turn information into knowledge, i.e. here we climb to the next level of the DIKW hierarchy. Thus, big data analytic capabilities have gained much attention recently (Akter and Fosso Wamba, 2016; Arora and Malik, 2015; Fosso Wamba *et al.*, 2017). Analytical capabilities are human centric, such as building a predictive analytics model, or interpreting a business need into an algorithm. A firm can develop analytical capabilities in-house or it can leverage service providers. In their data monetization case analysis, Najjar and Kettinger (2013) proposed three alternative paths toward these capabilities. From the practice point of view, analytical capabilities are a bottleneck, as data scientists are a scarce resource (Davenport, 2014; Janssen *et al.*, 2017).

Using the VVC terminology, applying analytics involves selecting and synthetizing the data into meaningful results. Information management and governance capabilities play a central role in the process. Mithas *et al.* (2011) stated that information management capabilities have an influential role in developing other capabilities that drive firm performance improvements. Tallon *et al.* (2013) found that while factors like IT standardization, data ownership rights and responsibilities enable realization of value, there are also factors that delay or impede value creation, such as legacy IT systems with data silos, and lack of integration. Moreover, Anand *et al.* (2016) pointed out the role of resource allocation and the role of top management in capability creation. Executives should promote a data-driven culture, and the organization should be able to allocate resources dynamically for data-based innovations.

Innovation capabilities are required to apply knowledge derived from data. Given certain conditions, people and organizations can be innovative (Dyer *et al.*, 2011). Digital transformation forces companies focus on data, and there are plenty of examples on how to utilize data. Zhan *et al.* (2017) built a framework that shows how big data can help companies accelerate their product innovation processes. Jetzek *et al.* (2014) presented a case study, where a company drives innovation by making use of open data.

3.4 Transformation process

A company uses its capabilities and knowledge to produce valued outcomes. Big data capabilities typically produce intermediate effects. For example, Tallon *et al.* (2013) stated that the effects of data governance are linked to industry specific, intermediate results instead of firm-level effects such as profitability. Moreover, aspects like the organizational context and managerial actions play a crucial role in the value creation process (Müller and Jensen, 2017).

Figure 2.
Converting big data assets into big data capabilities

Current big data literature recognizes four main areas (see also Table I) where big data can add value: decision making, operational efficiency, product/service innovations and business model innovations (Figure 2).

Big data analytics can influence decision-making processes, which improves the decision-making quality (Janssen *et al.*, 2017; Sharma *et al.*, 2014), i.e. analytics is related to better decisions. Moreover, the speed of decisions has increased due to real-time analytics (Bärenfänger *et al.*, 2014), even up to automated decision making (Mayer-Schönberger and Cukier, 2013). The agility of an organization increases with the speed of decisions, which assumedly causes a positive impact, at least when combined with quality decisions.

With regard to operational efficiency, big data has been connected to various factors that contribute toward positive impacts. Lower personnel costs, higher inventory accuracy (Bärenfänger *et al.*, 2014), better delivery accuracy (Dutta and Bose, 2015) and incremental improvements in operations (Roden *et al.*, 2017; Ylijoki and Porras, 2017) are examples of achieved impacts. Studies also pinpoint the importance of distribution (the last step in the VVC framework) in terms of insightful and visual user interfaces (Bärenfänger *et al.*, 2014; Dutta and Bose, 2015). These are related especially to decision making and operational efficiency. The interfaces should help to make correct interpretations of the analytics outcomes (Janssen *et al.*, 2017).

Manyika *et al.* (2011) expect up to 20–30 percent savings in product development and even 50–60 percent faster time-to-market cycles for companies that make use of big data in their product/service innovation processes. Data are a vehicle to developing new products, or to adding value to old products (Gobble, 2013), e.g. by adding sensors to existing machines for data gathering in order to create a predictive maintenance solution. Companies also innovate and offer data-based services, such as data-as-a-service or analytics-as-a-service (Chen *et al.*, 2011).

In addition, big data serves as a seedbed for new business models. In fact, many start-up firms base their business models on data. Digitalization and big data have changed the business landscape, forcing companies to innovate and review their existing business models (Weill and Woerner, 2015; Woerner and Wixom, 2015). Scholars are active in this area. Value co-creation and co-capture with business model considerations and related challenges have been studied, particularly in the context of the Internet of Things (Iivari *et al.*, 2016; Saarikko *et al.*, 2017). Business models by the data-driven trailblazers (Amazon, Google, Facebook, Apple) have been investigated (Walton, 2012) as well as models which are based on collaboration between competing firms (Ritala *et al.*, 2014).

4. Big data value creation recipe á la process theory

In order to realize big data impacts, companies need to have big data assets and capabilities. In the capability creation process, information turns into knowledge. The DIKW hierarchy states that information must come first, i.e. knowledge cannot be created without information. Correspondingly, in the transformation process, firms must have big data capabilities to produce impactful results. Moreover, transforming assets into impacts without the appropriate capabilities is impossible. Thus, the capability creation process and transformation process are sequential in nature.

To achieve a holistic view of big data value creation, we need to consider two additional factors. The first is that firms must make investments in order to achieve big data assets (asset creation process). We adopt an idea from Soh and Markus (1995), who presented the IT conversion process for turning IT expenditures into IT assets. Expenditures in this context are economic investments and spending that aims to, or supports, big data asset creation. Similarly to other processes, the big data asset creation process is subject to external factors. These factors, often technical in nature, can positively or negatively affect the output.

Finally, we must address the fact that big data impacts are intermediate results, i.e. we need to link the impacts to actual performance metrics. The performance of a firm depends on the competition process, which connects the firm to its industry, ecosystem, competitors and customers. As discussed in the previous section, there is a strong body of knowledge indicating that big data impacts, such as improved decision making or more agile product innovation cycles, indeed relate to performance. The impacts are a necessary condition for improved performance due to big data, but not sufficient, as any competition process is subject to numerous external factors.

These two additional processes complement the value creation model. Figure 3 presents the model, consisting of four sequential processes. Big data assets connect the asset creation process to the capability creation process. Big data capabilities connect the capability creation process to the transformation process, which in turn connects to the competition process via the impacts of big data. Using process theory terminology, the inputs and outputs are necessary but not sufficient conditions with regard to big data value creation.

Each of the processes may be affected by conditions that either favor or impede the process. In order to be successful, these factors must be favorable. To mitigate risks related to technology, data, organizational and external conditions, organizations must apply procedures that support the processes, as highlighted by the focus areas shown in Figure 3. The extant big data literature listed in Table II discusses many of these success factors (Figure 4).

Taken together, the four processes together with affecting +/− factors (discussed in Table III, see also Table II) and success factors combine a recipe for big data value creation. The model explains how big data and related technologies can contribute to firm performance. It takes into account the necessary conditions, such as big data assets and capabilities, while it also recognizes they are not sufficient on their own: each of the processes is probabilistic, i.e. it may fail to produce desired outcomes due to additional conditions that are unpredictable, and often difficult to manage in variance-theory-based approaches. A process-theory-based model allows us to build a holistic end-to-end view. Numerous variance-theory-based studies that cover certain parts of the whole support the view. We named the model AC/TC, taking first letters from each process and adding a slash to indicate the technology/performance conversion.

Figure 3.
Converting big data capabilities into big data impacts

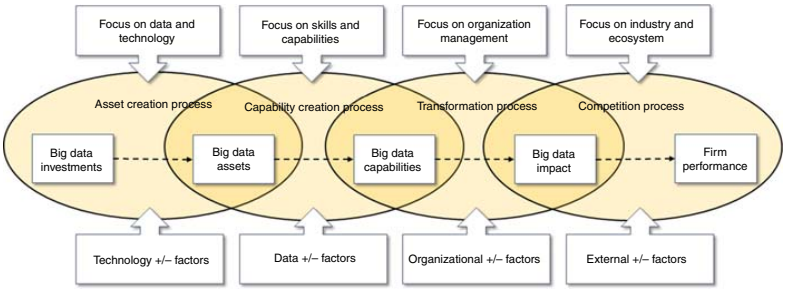


Figure 4.
AC/TC process – creating economic value with big data

Process	Contingency factors (+/−)
Asset creation	Integrating huge amounts of high-velocity data from various sources is challenging in terms of technology (Dutta and Bose, 2015; Krumeich *et al.*, 2014). In the context of a single firm, the output depends on factors and random events that the process cannot manage. For example, big data technologies are expensive and of varying maturity (Alharthi *et al.*, 2017; Boncea *et al.*, 2017). Immature technology can seriously damage the asset creation process. Pouring more money into the big data asset creation process does not automatically lead to more or better, big data assets
Capability creation	The outputs of the capability creation process depend on the inputs, i.e. big data assets are necessary conditions for the process to produce outputs. However, they are not sufficient on their own – there is no guarantee that more input produces more output. Scholars have observed several factors, such as organizational factors inhibiting radical innovations in incumbent firms (Sandberg and Aarikka-Stenroos, 2014), as well as data quality concerns weakening the veracity of data (Janssen *et al.*, 2017; Vidgen *et al.*, 2017), immature technologies leading to problems (Boncea *et al.*, 2017), or simply a lack of analytic resources that may hamper the process. From the viewpoint of a single firm, these external factors add uncertainty to the process. The outcomes should be viewed as discrete events. In some cases, the process will produce unsatisfactory results due to external factors. Thus, we can state that big data assets are necessary, but not sufficient for producing big data capabilities
Transformation	While it is clear that big data capabilities are a necessary condition for big data impacts, the current body of knowledge makes it clear that they are not sufficient on their own. A number of organizational factors affect the process, which can either promote or reduce the impacts of data. Several aspects have been discussed in the big data literature. For example, the lack of a data-driven culture (Dutta and Bose, 2015), scarce analytic resources (Janssen *et al.*, 2017; Shah *et al.*, 2012), insufficient organizational maturity regarding big data (Comuzzi and Patel, 2016), and organizational silos (Bärenfänger *et al.*, 2014; Sharma *et al.*, 2014) are considered challenges. Moreover, executives are better at managing capital, brands and people than data – and one out of five managers relies on intuition rather than data (Shah *et al.*, 2012)
Competition	The competition process is out-of-scope for this research. However, it is clear that a number of external factors which are out of the reach of the firm, such as competitors with new business models, regulations or customer behavior may support or hamper the performance

Table III.
Contingency factors affecting the processes

Figure 5 reflects our process model in the theoretical context of our study. We have replaced the generic IT value generation model (Soh and Markus, 1995) with a more specific model that applies to big data value creation processes. As in Figure 1, the overlapping in Figure 5 is indicative. In the preceding sections, we have linked our model to current big data literature. The model aligns well with existing theories created before the era of big data, while providing new insights into the big data value creation processes.

5. Concluding remarks

Our research acknowledges well-established and widely used theories created by other scholars. The elements our model builds on – DIKW hierarchy, VVC framework and process

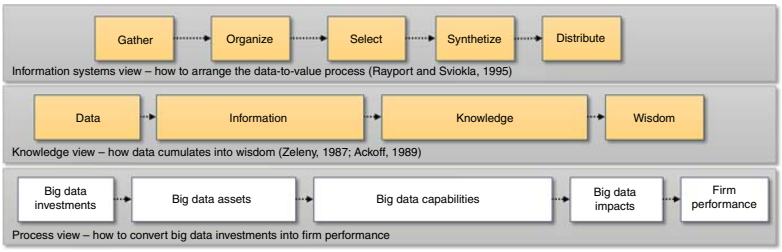


Figure 5.
AC/TC process model in the theoretical context of this study

theory – are not new. However, we claim that the way we rearrange and apply these elements to the big data context is novel. We support our model with a comprehensive overview of current big data literature.

The model contributes to big data research in several ways by providing a holistic, end-to-end process model from investment to performance. First, the model explains how big data investments are transformed into improved economic performance. Additionally, the model helps to understand why investments sometimes fail. Second, the model can be used to point out gray areas in current big data research. Reflecting the current body of knowledge in the model reveals areas which are less studied. Third, the model bridges current big data theory and practice. It is comprehensive and based on solid theoretical foundations, yet it is an appropriate framework in practical situations. For example, it can be used to identify success factors that organizations should focus on in order to mitigate negative impacts. Besides technical and data-related factors, organizations must focus on managerial and organizational aspects to gain performance improvements.

Our model is limited to a business context, specifically the perspective of a single firm utilizing big data in order to create economic value. As our focus was on the organization level, one potential area for future research is to explore big data value creation at the ecosystem or network level. For example, platform ecosystems or Internet of Things related value networks introduce new interactions and conditions, such as data ownership or cross-company collaboration, to the conversion process. Another avenue for future research would be at the boundaries of the model. Organizational factors play a crucial role in the transformation, thus, management activities of become crucial. Investigating, for example, success factors and pitfalls in capability creation and transformation processes would require a multi-disciplinary approach. As the data-driven paradigm becomes more and more pervasive, organizational factors become increasingly important, which will highlight the role of social sciences in explaining the impacts of big data.

In this study, we have developed an end-to-end process model which attempts to explain how big data creates economic value in a business context. Our synthetization of big data value creation-related theoretical papers, surveys and case studies, in combination with process components, provides a recipe for explaining how and why big data investments convert into performance.

The recipe defines four sequential and probabilistic processes with inputs and outputs, through which value creation takes place. In addition to the necessary conditions defined by the processes, factors which affect the processes must be favorable for performance improvement. Understanding the whole process helps scholars to identify potential research gaps and complex relationships in the conversion process. Practitioners may use the model to avoid pitfalls and identify best practices, especially related to data, as well as organizational and managerial factors that stimulate desired outcomes.

Note

1. For example, SCOPUS literature database index covers almost 50,000 papers containing the term "big data" in title, abstract or keywords. The vast majority of the papers are categorized under technical subject areas, such as computer science or mathematics.

References

Abbasi, A., Sarker, S. and Chiang, R.H. (2016), "Big data research in information systems: toward an inclusive research agenda", *Journal of the Association for Information Systems*, Vol. 17 No. 2, pp. 1-25.

Ackoff, R.L. (1989), "From data to wisdom", *Journal of Applied Systems Analysis*, Vol. 16 No. 1, pp. 3-9.

Akter, S. and Fosso Wamba, S. (2016), "Big data analytics in E-commerce: a systematic review and agenda for future research", *Electronic Markets*, Vol. 26 No. 2, pp. 173-194.

Akter, S., Wamba, S.F., Gunasekaran, A., Dubey, R. and Childe, S.J. (2016), "How to improve firm performance using big data analytics capability and business strategy alignment?", *International Journal of Production Economics*, Vol. 182 No. 8, pp. 113-131.

Alguliyev, R.M., Gasimova, R.T. and Abbaslı, R.N. (2017), "The obstacles in big data process", *International Journal of Modern Education & Computer Science*, Vol. 9 No. 3, pp. 28-35.

Alharthi, A., Krotov, V. and Bowman, M. (2017), "Addressing barriers to big data", *Business Horizons* (in press).

Anand, A., Sharma, R. and Coltman, T. (2016), "Four steps to realizing business value from digital data streams", *MIS Quarterly Executive*, Vol. 15 No. 4, pp. 259-277.

Ardagna, C.A., Ceravolo, P. and Damiani, E. (2016), "Big data analytics as-a-service: issues and challenges", *IEEE International Conference on Big Data (Big Data), IEEE*, pp. 3638-3644.

Arora, D. and Malik, P. (2015), "Analytics: key to go from generating big data to deriving business value", *IEEE First International Conference on Big Data Computing Service and Applications (BigDataService), IEEE*, pp. 446-452.

Bärenfänger, R., Otto, B. and Österle, H. (2014), "Business value of in-memory technology–multiple–case study insights", *Industrial Management & Data Systems*, Vol. 114 No. 9, pp. 1396-1414.

Barua, A., Kriebel, C.H. and Mukhopadhyay, T. (1995), "Information technologies and business value: an analytic and empirical investigation", *Information Systems Research*, Vol. 6 No. 1, pp. 3-23.

Boncea, R., Petre, I., Smada, D.-M. and Anagrama, A.Z. (2017), "A maturity analysis of big data technologies", *Informatica Economica*, Vol. 21 No. 1, pp. 60-71.

Braganza, A., Brooks, L., Nepelski, D., Ali, M. and Moro, R. (2017), "Resource management in big data initiatives: processes and dynamic capabilities", *Journal of Business Research*, Vol. 70 No. 8, pp. 328-337.

Brynjolfsson, E. and McAfee, A. (2012), *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*, The MIT Center for Digital Business.

Bumblauskas, D., Nold, H., Bumblauskas, P. and Igou, A. (2017), "Big data analytics: transforming data to action", *Business Process Management Journal*, Vol. 23 No. 3, pp. 703-720.

Chae, B.K., Yang, C., Olson, D. and Sheu, C. (2014), "The impact of advanced analytics and data accuracy on operational performance: a contingent resource based theory (RBT) perspective", *Decision Support Systems*, Vol. 59, pp. 119-126.

Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from big data to big impact", *MIS Quarterly*, Vol. 36 No. 4, pp. 1165-1188.

Chen, Y., Kreulen, J., Campbell, M. and Abrams, C. (2011), "Analytics ecosystem transformation: a force for business model innovation", *Annual SRII Global Conference (SRII)*, pp. 11-20.

Clarke, R. (2016), "Big data, big risks", *Information Systems Journal*, Vol. 26 No. 1, pp. 77-90.

Comuzzi, M. and Patel, A. (2016), "How organisations leverage big data: a maturity model", *Industrial Management & Data Systems*, Vol. 116 No. 8, pp. 1468-1492.

Cukier, K. (2010), "Data, data everywhere", *The Economist*, February 25, available at: www.economist.com/node/15557443

Davenport, T. (2014), *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*, Harvard Business Review Press.

Davenport, T.H., Barth, P. and Bean, R. (2012), "How big data is different", *MIT Sloan Management Review*, Vol. 54 No. 1, pp. 43-46.

Dehning, B., Richardson, V.J. and Zmud, R.W. (2003), "The value relevance of announcements of transformational information technology investments", *MIS Quarterly*, Vol. 27 No. 4, pp. 637-656.

Dutta, D. and Bose, I. (2015), "Managing a big data project: the case of Ramco cements limited", *International Journal of Production Economics*, Vol. 165, July 1, pp. 293-306.

Dyer, J., Gregersen, H. and Christensen, C. (2011), *The Innovator's DNA*, Harvard Business Review Press.

Fosso Wamba, S., Gunasekaran, A., Akter, S., Ren, S.J.-f., Dubey, R. and Childe, S.J. (2017), "Big data analytics and firm performance: effects of dynamic capabilities", *Journal of Business Research*, Vol. 70, pp. 356-365.

Frankel, D.A. (2012), "Big data and risk management", *Risk Management*, Vol. 59 No. 8, p. 13, available at: www.rmmagazine.com/2012/10/01/big-data-and-risk-management/

Gandomi, A. and Haider, M. (2015), "Beyond the hype: big data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

Gantz, J. and Reinsel, D. (2011), "Extracting value from chaos", IDC, available at: www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Gobble, M. (2013), "Big data: the next big thing in innovation", *Research and Technology Management*, Vol. 56 No. 1, pp. 64-66.

Guo, L., Sharma, R., Yin, L., Lu, R. and Rong, K. (2017), "Automated competitor analysis using big data analytics: evidence from the fitness mobile app business", *Business Process Management Journal*, Vol. 23 No. 3, pp. 735-762.

Hartono, R. and Sheng, M.L. (2016), "Knowledge sharing and firm performance: the role of social networking site and innovation capability", *Technology Analysis & Strategic Management*, Vol. 28 No. 3, pp. 335-347.

Hazen, B.T., Boone, C.A., Ezell, J.D. and Jones-Farmer, L.A. (2014), "Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications", *International Journal of Production Economics*, Vol. 154, pp. 72-80.

Hazen, B.T., Weigel, F.K., Ezell, J.D., Boehmke, B.C. and Bradley, R.V. (2017), "Toward understanding outcomes associated with data quality improvement", *International Journal of Production Economics*, Vol. 193, pp. 737-747.

Hitt, L.M. and Brynjolfsson, E. (1996), "Productivity, business profitability, and consumer surplus: three different measures of information technology value", *MIS Quarterly*, Vol. 20 No. 2, pp. 121-142.

Huberty, M. (2015), "Awaiting the second big data revolution: from digital noise to value creation", *Journal of Industry, Competition and Trade*, Vol. 15 No. 1, pp. 35-47.

Iddris, F. (2016), "Innovation capability: a systematic review and research agenda", *Interdisciplinary Journal of Information, Knowledge, and Management*, Vol. 11, pp. 235-260.

IFRS (2015), *International Financial Reporting Standard for Small and Medium-sized Entities*, IFRS Foundation Publications Department, London.

Iivari, M.M., Ahokangas, P., Komi, M., Tihinen, M. and Valtanen, K. (2016), "Toward ecosystemic business models in the context of industrial internet", *Journal of Business Models*, Vol. 4 No. 2, pp. 42-59.

Janssen, M., van der, V.H. and Wahyudi, A. (2017), "Factors influencing big data decision-making quality", *Journal of Business Research*, Vol. 70 No. 1, pp. 338-345.

Jetzek, T., Avital, M. and Bjorn-Andersen, N. (2014), "Data-driven innovation through open government data", *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 9 No. 2, pp. 100-120.

Keen, J., Calinescu, R., Paige, R. and Rooksby, J. (2013), "Big data+ politics = open data: the case of health care data in England", *Policy & Internet*, Vol. 5 No. 2, pp. 228-243.

Kitchin, R. and McArdle, G. (2016), "What makes big data, big data? Exploring the ontological characteristics of 26 datasets", *Big Data & Society*, Vol. 3 No. 1, pp. 1-10.

Krumeich, J., Jacobi, S., Werth, D. and Loos, P. (2014), "Towards planning and control of business processes based on event-based predictions", *Business Information Systems*, pp. 38-49.

Lamba, K. and Singh, S.P. (2017), "Big data in operations and supply chain management: current trends and future perspectives", *Production Planning & Control*, Vol. 28 Nos 11/12, pp. 877-890.

Laney, D. (2001), "3D data management: controlling data volume, velocity and variety", *META Group Research Note*, Vol. 6, pp. 70-73.

Lycett, M. (2013), "Datafication': making sense of (big) data in a complex world", *European Journal of Information Systems*, Vol. 22, pp. 381-386.

McAfee, A. and Brynjolfsson, E. (2012), "Big data: the management revolution", *Harvard Business Review*, Vol. 90 No. 10, pp. 61-67.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), "Big data: the next frontier for innovation, competition, and productivity", in Manyika, J. and Chui, M. (Eds), McKinsey Global Institute, available at: www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Markus, M.L. and Robey, D. (1988), "Information technology and organizational change: causal structure in theory and research", *Management Science*, Vol. 34 No. 5, pp. 583-598.

Mayer-Schönberger, V. and Cukier, K. (2013), *Big Data: A Revolution that will Transform How we Live, Work and Think*, Houghton Mifflin Harcourt, Boston, MA.

Melville, N., Kraemer, K. and Gurbaxani, V. (2004), "Information technology and organizational performance: an integrative model of IT business value", *MIS Quarterly*, Vol. 28 No. 2, pp. 283-322.

Merino, J., Caballero, I., Rivas, B., Serrano, M. and Piattini, M.A. (2016), "A data quality in use model for big data", *Future Generation Computer Systems*, Elsevier, Vol. 63 No. 12, pp. 123-130.

Miller, H.G. and Mork, P. (2013), "From data to decisions: a value chain for big data", *IT Professional*, Vol. 15 No. 1, pp. 57-59.

Mithas, S., Ramasubbu, N. and Sambamurthy, V. (2011), "How information management capability influences firm performance", *MIS Quarterly*, Vol. 35 No. 1, pp. 237-256.

Mohr, L.B. (1982), *Explaining Organizational Behavior*, Jossey-Bass, San Francisco, CA.

Müller, S. and Jensen, P. (2017), "Big data in the Danish industry: application and value creation", *Business Process Management Journal*, Vol. 23 No. 3, pp. 645-670.

Najjar, M.S. and Kettinger, W.J. (2013), "Data monetization: lessons from a Retailer's journey", *MIS Quarterly Executive*, Vol. 12 No. 4, pp. 213-225.

Newell, S. and Marabelli, M. (2015), "Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of 'datification'", *The Journal of Strategic Information Systems*, Vol. 24 No. 1, pp. 3-14.

Piccoli, G. and Pigni, F. (2013), "Harvesting external data: the potential of digital data streams", *MIS Quarterly Executive*, Vol. 12 No. 1, pp. 53-64.

Pigni, F., Piccoli, G. and Watson, R. (2016), "Digital data streams", *California Management Review*, Vol. 58 No. 3, pp. 5-25.

Porter, M.E. and Millar, V.A. (1985), "How information gives you competitive advantage", *Harvard Business Review*, Vol. 63 No. 4, pp. 149-160.

Pousttchi, K. and Hufenbach, Y. (2014), "Engineering the value network of the customer interface and marketing in the data-rich retail environment", *International Journal of Electronic Commerce*, Vol. 18 No. 4, pp. 17-42.

Rayport, J.F. and Sviokla, J.J. (1995), "Exploiting the virtual value chain", *Harvard Business Review*, Vol. 73 No. 6, pp. 75-85.

Ritala, P., Golnam, A. and Wegmann, A. (2014), "Coopetition-based business models: the case of Amazon. com", *Industrial Marketing Management*, Vol. 43 No. 2, pp. 236-249.

Roden, S., Nucciarelli, A., Li, F. and Graham, G. (2017), "Big data and the transformation of operations models: a framework and a new research agenda", *Production Planning & Control*, Vol. 28 Nos 11/12, pp. 929-944.

Rowley, J.E. (2007), "The wisdom hierarchy: representations of the DIKW hierarchy", *Journal of Information Science*, Vol. 33 No. 2, pp. 163-180.

Saarikko, T., Westergren, U.H. and Blomquist, T. (2017), "The internet of things: are you ready for what's coming?", *Business Horizons*, Vol. 60 No. 5, pp. 667-676.

Sandberg, B. and Aarikka-Stenroos, L. (2014), "What makes it so difficult? A systematic review on barriers to radical innovation", *Industrial Marketing Management*, Vol. 43 No. 8, pp. 1293-1305.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. and Tufano, P. (2012), *Analytics: The Real-world Use of Big Data*, IBM Global Services, available at: www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf

Shah, S., Horne, A. and Capellá, J. (2012), "Good data won't guarantee good decisions", *Harvard Business Review*, Vol. 90 No. 4, pp. 23-25.

Sharma, R., Mithas, S. and Kankanhalli, A. (2014), "Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations", *European Journal of Information Systems*, Vol. 23 No. 4, pp. 433-441.

Soh, C. and Markus, M.L. (1995), "How IT creates business value: a process theory synthesis", *ICIS 1995 Proceedings*, pp. 29-41.

Sullivan, C. (2014), "Protecting digital identity in the cloud: regulating cross border data disclosure", *Computer Law & Security Review*, Vol. 30 No. 2, pp. 137-152.

Tallon, P.P., Ramirez, R.V. and Short, J.E. (2013), "The information artifact in IT governance: toward a theory of information governance", *Journal of Management Information Systems*, Vol. 30 No. 3, pp. 141-178.

Teece, D.J. (2007), "Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance", *Strategic Management Journal*, Vol. 28 No. 13, pp. 1319-1350.

Tiwana, A. (2014), "Separating signal from noise: evaluating emerging technologies", *MIS Quarterly Executive*, Vol. 13 No. 1, pp. 45-61.

Truyens, M. and Van Eecke, P. (2014), "Legal aspects of text mining", *Computer Law & Security Review*, Vol. 30 No. 2, pp. 153-170.

Van't Spijker, A. (2014), *The New Oil: Using Innovative Business Models to Turn Data Into Profit*, Technics Publications, Basking Ridge, NJ.

Vidgen, R., Shaw, S. and Grant, D.B. (2017), "Management challenges in creating value from business analytics", *European Journal of Operational Research*, Vol. 261 No. 2, pp. 626-639.

Walton, N. (2012), "Four-closure: how amazon, apple, Facebook & Google are driving business model innovation", *Chinese Business Review*, Vol. 11 No. 11, pp. 981-988.

Weill, P. and Woerner, S.L. (2015), "Thriving in an increasingly digital ecosystem", *MIT Sloan Management Review*, Vol. 56 No. 4, pp. 27-34.

Wernerfelt, B. (1984), "A resource-based view of the firm", *Strategic Management Journal*, Vol. 5 No. 2, pp. 171-180.

Woerner, S.L. and Wixom, B.H. (2015), "Big data: extending the business strategy toolbox", *Journal of Information Technology*, Vol. 30 No. 1, pp. 60-62.

Ylijoki, O. and Porras, J. (2016), "Perspectives to definition of big data: a mapping study and discussion", *Journal of Innovation Management*, Vol. 4 No. 1, pp. 69-91.

Ylijoki, O. and Porras, J. (2017), "What managers think about big data", *International Journal of Business Information Systems* (in press).

Zeleny, M. (1987), "Management support systems: towards integrated knowledge management", *Human Systems Management*, Vol. 7 No. 1, pp. 59-70.

Zhan, Y., Tan, K.H., Ji, G., Chung, L. and Tseng, M. (2017), "A big data framework for facilitating product innovation processes", *Business Process Management Journal*, Vol. 23 No. 3, pp. 518-536.

Corresponding author
Ossi Ylijoki can be contacted at: ossi.ylijoki@phnet.fi

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

807. DABIRI, MOHAMMAD. The low-cycle fatigue of S960 MC direct-quenched high-strength steel. 2018. Diss.

808. KOSKELA, VIRPI. Tapping experiences of presence to connect people and organizational creativity. 2018. Diss.

809. HERALA, ANTTI. Benefits from Open Data: barriers to supply and demand of Open Data in private organizations. 2018. Diss.

810. KÄYHKÖ, JORMA. Erityisen tuen toimintaprosessien nykytila ja kehittäminen suomalaisessa oppisopimuskoulutuksessa. 2018. Diss.

811. HAJIKHANI, ARASH. Understanding and leveraging the social network services in innovation ecosystems. 2018. Diss.

812. SKRIKO, TUOMAS. Dependence of manufacturing parameters on the performance quality of welded joints made of direct quenched ultra-high-strength steel. 2018. Diss.

813. KARTTUNEN, ELINA. Management of technological resource dependencies in interorganizational networks. 2018. Diss.

814. CHILD, MICHAEL. Transition towards long-term sustainability of the Finnish energy system. 2018. Diss.

815. NUTAKOR, CHARLES. An experimental and theoretical investigation of power losses in planetary gearboxes. 2018. Diss.

816. KONSTI-LAAKSO, SUVI. Co-creation, brokering and innovation networks: A model for innovating with users. 2018. Diss.

817. HURSKAINEN, VESA-VILLE. Dynamic analysis of flexible multibody systems using finite elements based on the absolute nodal coordinate formulation. 2018. Diss.

818. VASILYEV, FEDOR. Model-based design and optimisation of hydrometallurgical liquid-liquid extraction processes. 2018. Diss.

819. DEMESA, ABAYNEH. Towards sustainable production of value-added chemicals and materials from lignocellulosic biomass: carboxylic acids and cellulose nanocrystals. 2018. Diss.

820. SIKANEN, EERIK. Dynamic analysis of rotating systems including contact and thermal-induced effects. 2018. Diss.

821. LIND, LOTTA. Identifying working capital models in value chains: Towards a generic framework. 2018. Diss.

822. IMMONEN, KIRSI. Ligno-cellulose fibre poly(lactic acid) interfaces in biocomposites. 2018. Diss.

823. YLÄ-KUJALA, ANTTI. Inter-organizational mediums: current state and underlying potential. 2018. Diss.

824. ZAFARI, SAHAR. Segmentation of partially overlapping convex objects in silhouette images. 2018. Diss.

825. MÄLKKI, HELENA. Identifying needs and ways to integrate sustainability into energy degree programmes. 2018. Diss.

826. JUNTUNEN, RAIMO. LCL filter designs for paraller-connected grid inverters. 2018. Diss.

827. RANAEI, SAMIRA. Quantitative approaches for detecting emerging technologies. 2018. Diss.

828. METSO, LASSE. Information-based industrial maintenance - an ecosystem perspective. 2018. Diss.

829. SAREN, ANDREY. Twin boundary dynamics in magnetic shape memory alloy Ni-Mn-Ga five-layered modulated martensite. 2018. Diss.

830. BELONOGOVA, NADEZDA. Active residential customer in a flexible energy system - a methodology to determine the customer behaviour in a multi-objective environment. 2018. Diss.

831. KALLIOLA, SIMO. Modified chitosan nanoparticles at liquid-liquid interface for applications in oil-spill treatment. 2018. Diss.

832. GEYDT, PAVEL. Atomic Force Microscopy of electrical, mechanical and piezo properties of nanowires. 2018. Diss.

833. KARELL, VILLE. Essays on stock market anomalies. 2018. Diss.

834. KURONEN, TONI. Moving object analysis and trajectory processing with applications in human-computer interaction and chemical processes. 2018. Diss.

835. UNT, ANNA. Fiber laser and hybrid welding of T-joint in structural steels. 2018. Diss.

836. KHAKUREL, JAYDEN. Enhancing the adoption of quantified self-tracking wearable devices. 2018. Diss.

837. SOININEN, HANNE. Improving the environmental safety of ash from bioenergy production plants. 2018. Diss.

838. GOLMAEI, SEYEDMOHAMMAD. Novel treatment methods for green liquor dregs and enhancing circular economy in kraft pulp mills. 2018. Diss.

839. GERAMI TEHRANI, MOHAMMAD. Mechanical design guidelines of an electric vehicle powertrain. 2019. Diss.

840. MUSIIENKO, DENYS. Ni-Mn-Ga magnetic shape memory alloy for precise high-speed actuation in micro-magneto-mechanical systems. 2019. Diss.

841. BELIAEVA, TATIANA. Complementarity and contextualization of firm-level strategic orientations. 2019. Diss.

842. EFIMOV-SOINI, NIKOLAI. Ideation stage in computer-aided design. 2019. Diss.

843. BUZUKU, SHQIPE. Enhancement of decision-making in complex organizations: A systems engineering approach. 2019. Diss.

844. SHCHERBACHEVA, ANNA. Agent-based modelling for epidemiological applications. 2019. Diss.

ISBN 978-952-335-346-6
ISBN 978-952-335-347-3 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491
Lappeenranta 2019