Tiina Strengell

COMPETITIVENESS FROM DATA AND ANALYTICS: REQUIRED COMPETENCY IN ORGANIZATION

UNIVERSITY OF JYVÄSKYLÄ
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2017

ABSTRACT

Strengell, Tiina
Competitiveness from Data and Analytics: Required Competency in Organization
Jyväskylä: University of Jyväskylä, 2017, 91p.
Information Systems Science, Master's Thesis
Supervisor: Holtkamp, Phillipp

The purpose of this master's thesis is to determine the competency requirements of data analytics and identify the most beneficial competencies that create customer value. This study is implemented using a mixed-method approach, which combines qualitative and quantitative methods. A literature review is at the base of the empirical research. The empirical study is done using questionnaires, in which both structured and unstructured questions are used. A total of 18 participants from three different companies are involved in the study. Descriptive statistics and graphs are used to summarize the answers to the structured questions. Descriptions are compared with the literature review findings. Analysis of open questions is carried out using the Competence Performance Theory of Korossy.

Competency requirements of data analysts found in empirical research had much in common with the requirements described in the literature. The major exceptions to this are programming skills and machine learning. Their importance is emphasized in the literature. In the empirical data, the knowledge varies quite a lot. Nevertheless, respondents do not see it as a problem. Aside from a few technical skills, which are repeated in all datasets, other skills also vary.

The results justify the main competencies needed for the creation of customer value, which are data architecture design, business planning, and knowledge of the company's strategic planning. These competencies were raised as competencies both for individuals and the organization. Individuals need the skills mentioned above, but the company's contribution is also essential. Generally speaking, personal traits and business skills were appreciated more than technical skills.

Keywords: competency requirements, organizational performance, analytics

TIIVISTELMÄ

Strengell, Tiina
Competitiveness from Data and Analytics: Required Competency in Organization
Jyväskylä: Jyväskylän Yliopisto, 2017, 91s.
Tietojärjestelmätiede, pro gradu tutkielma
Ohjaaja: Holtkamp, Phillipp

Tässä pro gradu – tutkimuksessa tarkoituksena on selvittää mitä kompenssivaatimuksia data analytiikassa on ja mitkä kompetenssit ovat tärkeimpiä asiakkaalle luodun arvon kannalta. Tutkimuksessa on hyödynnetty mixed method – lähestymistapaa, jossa yhdistetään määrillisen ja laadullisen tutkimuksen elementtejä. Empiirinen tutkimus pohjautuu kirjallisuuskatsauksen tuloksiin. Empiirinen tutkimus toteutettiin sähköisin kyselylomakkein, jotka sisälsivät sekä strukturoituja että avoimia kysymyksiä. Tutkimukseen osallistui 18 henkilöä kolmesta eri yrityksestä. Strukturoitujen kysymysten vastaukset valmisteltiin kuvailevien tilastomenetelmien avulla ja analysoitiin verraten löydöksiä kirjallisuuskatsauksen tuloksiin. Avointen vastausten analysointi tehtiin Korossyn Competence-Performance -teorian pohjalta.

Data-analyytikoilta vaadittavat kompetenssit olivat kirjallisuudessa ja tutkimuksen empiirisessä aineistossa pääosin samankaltaisia. Suurin poikkeus tähän oli ohjelmointitaidot ja koneoppiminen. Niiden tärkeyttä painotettiin kirjallisuudessa. Empiirisessä aineistossa keskimääräinen osaaminen ei ollut erityisen korkea ja vaihteli melko paljon. Siitä huolimatta osallistujat eivät kokeneet sitä ongelmaksi. Muu tekninen osaaminen vaihteli melko paljon, lukuun ottamatta muutamaa teknistä taitoa, jotka toistuivat kaikissa aineistoissa.

Tulosten perustella asiakkaan arvon luonnin kannalta tärkeimmät kompetenssit ovat data-arkkitehtuurin suunnittelu, tietämys datan ja liiketoiminnan yhdistämisestä sekä yrityksen strateginen suunnittelu. Nämä kompetenssit tulivat esiin sekä yksilöiden että yrityksen kompetensseina. Yksilöt tarvitsevat edellä mainittua osaamista, mutta myös yrityksen panos on välttämätön. Yleisesti ottaen, henkilökohtaisia piirteitä ja liiketoiminnallista osaamista arvostettiin teknistä osaamista enemmän.

Hakusanat: kompetenssivaatimukset, organisaation suorituskyky, analytiikka

FIGURES

TABLES

CONTENTS

1 Introduction

It has been understood for decades that organizations are consumers, managers, and distributors of data. Essential elements of organization operations are the rules for gathering, storing, communicating, and using data (Feldman & March 1981). Information is gathered and used because it helps to support decision making. By analyzing data, decisions can be based on tight data-based insights rather than intuition or experience. (Davenport 2015.)

Nowadays, almost all devices can be connected to the internet, and the amount of collected data has increased exponentially. New technologies are collecting more data than ever before in every industry and in every part of the world (LaValle et al. 2010.) However, rich data in itself is not valuable. In terms of organizational competitiveness the value lies in how organizations can turn customers' needs into algorithms and algorithms into action (Sondergaard 2015). The drive to get the full value from the massive amounts of information organizations already have has become one of the most important strategic issues (Davenport & Prusak 1998; LaValle et al. 2010).

Analytics techniques have been studied for a long time among scientists. Recently, these techniques have gotten a foothold in business life, and through that new challenges have been encountered leading to the emergence of data science. (LaValle et al. 2010; Davenport 2015; Gartner 2016.) With regard to the exploitation of data, the terms such as *data*, *data analytics* and its types, *data science*, *big data*, and *advanced analytics* have been the subject of considerable discussion. These terms are essential to this thesis. Definitions are introduced in chapter 4. The descriptions are based on many sources. In this study, the persons who actively participate in designing and implementing the use of data and analytics are called "data analysts."

The ability to transform data into insights about customers' motivations and to turn those insights into strategy increasingly separates the winners from the losers (van den Driest, Sthanunathan & Weed 2016). Analyzing data can also be used to do things more efficiently and faster. It can even enable new business models (Van der Aalst 2012). It is known that there are enormous possibilities. The possibilities of analytics have received considerable attention, but practical ways to achieve competitiveness have been discussed less. The problem at the moment in many organizations is how to utilize these opportunities efficiently in order to gain competitive advantage. To achieve an advantage, competency and actions on the part of organizations and individuals are required.

The abovementioned issues have been investigated by analyzing how organizations can utilize competencies. As an example, Sandberg (2000, 9) says, organizational actions are always based on competence. It makes competency a good starting point for development. When all the organizational actions are somehow based on competency, the key for success in organizations is the efficient usage of competencies. In this thesis, the definition of competence and competency relies on many sources. To summarize, competence or competency differs from intelligence or achievements. It includes not only skills but also knowledge, abilities, behaviors, and personal characteristics, which are connected to performance and can be improved. (Taylor 1911; McClelland 1973; Dubois 1993; Sandberg 2000.) In addition to individual competencies, organizational competencies have also been included in the review (Boyatzis 2008).

In this study, underlying competencies in data analytics are evaluated in order to analyze how organizations can utilize those possibilities to gain competitive advantage. Also, the most beneficial competencies to create customer value are resolved. Nonacademic discourses, such as many company publications and consultant reports, provide a great deal of tacit knowledge on the topic. For this reason, in this thesis some insights from nonacademic sources are also taken under review. One example of the generalized nonacademic definition is the Venn diagram taken from Conwey (2010). It is perhaps the most used description of needed competency in data science. Data analytics is developing with such rapidity that academic research has difficulty keeping up with the progress.

This study is mixed method research, which combines qualitative and quantitative methods (Johnson & Onwuegbuzie 2013). The empirical study is based on literature review. In this study, the purpose of the literature review is to build an overall picture of competency and the terms related to data analytics. The literature review also creates an overview of competencies that are needed in the work of a data analyst. In addition to creating an overview of the subjects, the literature review is also the basis of a questionnaire that is used for collecting empirical data.

1.1 Goal of the Research and Research Questions

This thesis gives an overview of past research concerning competencies, commonly used terms in data analytics, and needed competencies in data analytics. This study also introduces a competency identification process for assessing competency requirements in an organization. Based on these element, required competencies regarding data analytics empirically in organizations are investigated. The main goal is to understand which competencies related to data analytics are most beneficial to create customer value.

Research questions are:

1. What are the competency requirements in data analytics?
   - The answer will be addressed by comparing the competencies found in the literature and empirically identified competencies.

2. Which are the most beneficial competencies to create customer value with data analytics?
   - The answer will be addressed by defining competence-performance space based on the empirical material.

1.2 Structure of the thesis

This study is organized as follows: introduction, methodology, literature, empirical part, and discussion. The first chapter describes the frame of this investigation. In the introduction section, the importance of this study, the goal of the research and research questions are explained. Second and third sections cover additional literature and the basis for the empirical study.

The second section introduces literature related to competence. Based on literature, the definition and usage of competence or competency will be introduced. Next, the competencies identified in the literature will be reviewed. After competence and competency have been defined, modeling competency in organizations will be discussed. The section on competence also includes an introduction to the theories that will be used in the analysis of the empirical material.

In chapter three, literature related to data analytics will be introduced. Most of the relevant articles and studies in the data analytics area have been implemented in the United States. This is indicated by the research executed by Capgemini with MIT Management Sloan Research (2016). It reveals that at the moment US companies are the most successful in their operational analytics initiatives as well as the most advanced.

The data analytics part starts by find defining the terms *data* and *big data*. After that, there will be a short history of data analytics, in which the term will be described. In order to understand what kind of competencies are needed, it is important to define the goal to be achieved. After that, analytics is explained in more detail by describing the types of analytics. In that section, also the terms *advanced analytics* and *data science* will be explained.

Finally, we get to the key issue: competency needed in data analytics. In section 3.2, the skills and other competencies, which could be required for exploitation of data analytics, will be introduced. In section 3.2, empirical material collected by Pinola (2015) is also presented.

Research methodology and research process are introduced in chapter 4, and quality aspects of the project are also discussed. Chapter 5 summarizes findings derived from the literature review and empirical materials. Chapters 6 and 7 provide answers to the research questions and a discussion of the results and future areas of study.

2 Competence-Oriented Research

Nowadays, especially in dynamic work domains, a crucial issue in the organization is how to develop human competence at work in the way that enables an organization to remain viable. There are many conflicting definitions concerning competence and competency. These are presented in next chapter.

2.1 Definition of Competence and Competency

When developing work competence we need to understand what it constitutes. The issue is not new, but only recently has the concept of competence been used more systematically in management (Sandberg 2000).

Taylor (1911) was one of the first to address the large difference between the least and most competent workers accomplishing their work. Contrary to previous claims, he believed that the most efficient and most competent workers are not readymade. Under systematic management, the best person rises to the top more certainly and more rapidly. *Scientific management*, the term Taylor uses, includes the idea that the only way to expand productivity is to increase the efficiency of workers. (Taylor 1911.)

Taylor was one of the pioneers who found that things other than cognitive ability make a difference in individual's performance. Still, until the 1970s most organizations believed that only deep technical skills raise performance. In the 1970s the same topic has been investigated by researchers such as McClelland (1973), who suggested that it would be better to examine the competence than intelligence or achievements because it would predict performance better. In random testing, he proposed to identify the competency domains that are most likely to be able to do a particular job well. (McClelland 1973.)

After that, competency research has continued, and competence or competency have often consisted of job analysis. Job competency has been defined as an underlying characteristic of an employee that results in effective or superior performance in a job. It can be defined as a capability to use knowledge, skills, abilities, behaviors, and personal characteristics to successfully perform critical work tasks or operate in a given role or position. Personal qualities may be the mental, intellectual, cognitive, social, emotional, attitudinal, physical or psychomotor attributes necessary to perform the job. (Dubois 1993.)

Boyatzis' (1982) definition includes both internal and external constraints, personal capability or talent, and organizational environment. According to him, the behaviors and performance are the outcomes of the competencies, such as motives, knowledge, thinking styles, and social roles. The findings of Sandberg (2000) suggest that workers' ways of conceiving of their work constitutes competence, in addition to attributes such as skills and knowledge. In his study, it was empirically demonstrated that variation in performance is related not only to the attributes of those who are regarded as the most competent but to the way they conceive of their work. For example, the conceptions about engine optimization were more hypothetical than factual. (Sandberg 2000.)

Spencer (1993) models competency as an iceberg, which is shown in figure 1. Spencer defines five hidden and visible competencies and asserts that five essential features are associated with superior performance in a certain situation. The features of competency are motives, personal characteristics, self-concept, knowledge, and skills. Motives lead behaviors. Traits are physical features that give consistent responses to a position or information. Self-perception is a personal attitude or value. Knowledge includes information that a person has, and skill is the ability to perform a particular task. (Spencer 1993.)

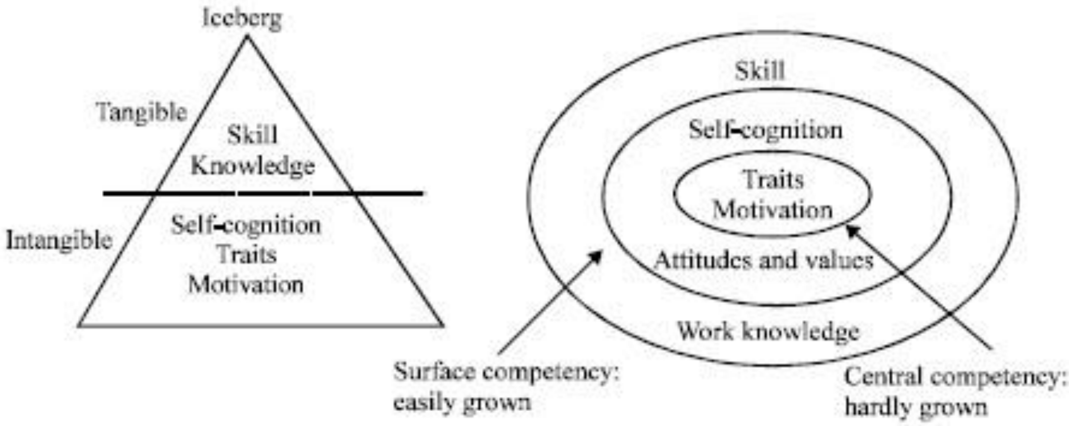

FIGURE 1 The Iceberg Model of Competence (Spencer & Spencer 1993, 11)

In conclusion, competency differs from intelligence or achievements. It includes not only skills but knowledge, abilities, behaviors, and personal characteristics, which are connected to performance and can be improved. (Taylor 1911; McClelland 1973; Dubois 1993; Sandberg 2000.) Also, performance is also dependent on the context (Sandberg 2000). In this study differentiation of competencies are in line with KSAOs (knowledge, skills, abilities and other characteristics) research (Schmitt & Chan 1998; Lucia & Lepsinger 1999). Also, Spencer & Spencer (1993) and Lay et al. (2006) have classified competencies into more stable characteristics, such as personal abilities and motivation, and more variable characteristics, such as skills and knowledge.

In this thesis, *competency* is used to describe a set of knowledge, skills, and attitudes to solve a problem in a given context. By contrast, the term *competence* is used to describe a particular knowledge item, skill, or attitude.

In human resource management, the competency approach is based on identifying, defining, and measuring individual differences regarding work-related constructs, particularly the capacities that are critical to successful job performance. Competency modeling can be used as a planning tool by organizations to deal with changes in technology, changes in the workforce, and changes in the very nature of the work the organization does. (Lucia & Lepsinger

1999.) More about modeling competencies will be discussed after defining organizational competency.

2.2 Organizational Competency

As submitted above, competitive advantage depends on the ability to activate organizational resources. In strategic management, organizational behavior and human resource management have been concentrated on the internal capabilities of organizations, including a particular focus on employees' competency (Vakola, Soderquist, Prastacos 2007).

However, competence can be divided into two components relating to its scope: individual and organizational. Organizational competency is more than just the collective competency of all the individuals. It is the combination of individual competencies, which are connected to strategic, structural, and cultural factors. It is also the availability of suitable resources that help an organization to compete. (Wagner 2012.)

As also mentioned above, Boyatzis (1982; 2008) is of the opinion that maximum performance occurs when a person's capability or talent is consistent with the job demands and the needs of the organizational environment. The theory used in this approach is a simple contingency theory, as presented in figure 2. Boyatzis (2008) describes a person's talent in terms of his or her values, vision, and personal philosophy, knowledge, competencies, life and career stage, interests, and style. He describes job demands as the role responsibilities and tasks needed to be performed. In addition, aspects of the organizational environment are predicted to have a substantial impact on the performance. The organizational environment consists of culture and climate, structure and systems, the maturity of the industry, strategic positioning within it, and core competencies. It also includes aspects of the economic, political, social, environmental, and religious milieu surrounding the organization. (Boyatzis 2008.)

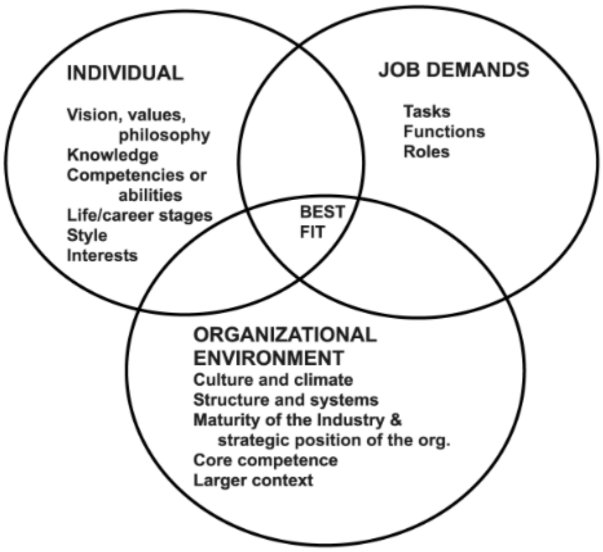

FIGURE 2 Theory of Action and Job Performance (Boyatzis 2008, 7)

In this study, the focus is on researching competency mostly at the individual level. But the importance of organizational competency cannot be ignored and so it is also considered in this thesis. The next section introduces how competency can be modeled.

2.3 Competency Models

Mansfield (1996) describes a competency model as a detailed, behaviorally accurate description of the skills and traits that employees need to have to be productive in a job. When beginning to form a competency model from single competence, a decision on what model would fit best our needs must be made. Attention can be focused on the roles, on a real position, or on a single person. Also, a combination of a role and competency that has been useful in a real position can be focused on. (Lucia & Lepsinger 1999.)

Often when identifying a competency model, there has been an orientation toward the skills needed to continue doing what the organization already does. In times of frequent change, it is important also to take a forward-looking and proactive approach. Competency management has also to be integrated into organization strategy. Not only high skilled employees but, even more importantly, employees who can adapt to changes, learn quickly, commit themselves to continuous development, and communicate effectively are the most valuable. (Athey & Orth 1999; Rodrigues at al. 2002.)

2.3.1 Competence Performance Approach

As described above, competence is connected to performance. Competence is thought of as a psychological or mental property. Performance, by contrast, refers to an actual event. This distinction between competencies and performance outcomes has been known for a very long time and has influenced traditional approaches to competency management. (Boyatzis 1982; Lucia & Lepsinger 1999.)

In addition to the previously mentioned studies, the topic has been investigated extensively also in linguistics. Hymes (1971), the father of communicative competence, found that linguistic competence and performance differ from one another. For example, if you make grammatical mistakes and know that they are mistakes, your performance does not match your competence. On the other hand, if you don't know they are mistakes, your competence matches your performance.

In this study, Pinola's (2015) master's thesis has been taken as a basis. In his thesis, he presented data scientist competency using Havelka and Merhout's (2009) competence framework. Havelka and Merhout's framework is suitable for describing and categorizing the competencies. This thesis also presents the results of a literature review and parts of the empirical research in the framework.

However, when the aim is also to analyze or develop an organization's performance based on competencies, another approach is needed.

2.3.2 Knowledge Spaces Theory

Doignon and Falmange (1985) developed the knowledge spaces theory, which is a good base for further development. Knowledge spaces theory is a mathematical theory for the efficient assessment of knowledge. The theory conceptualizes knowledge as a large, specified set of questions or problems. (Doignon & Falmange 1985; Doignon & Falmange 2012.)

Knowledge states describe every topic that a person is capable of solving. Knowledge states for a given subject are organized into a knowledge space, which is a structure specifying the surmise relations between knowledge states. From the answer to some questions, we can surmise a question about some other element. (Doignon & Falmange 2012.) An example of using knowledge spaces theory could be a teacher helping a student determine which courses would be appropriate to the student's career. By considering a basic set of questions, called "domains," which are broad enough to give representative coverage of the field, asking questions (this can also be called "problem" or "item") and hearing student's answer, the teacher can formalize a picture of student's knowledge state. Here, *knowledge state* means all the problems that the student can solve in ideal conditions. By collecting all the knowledge states, a knowledge structure can be created, as presented in figure 3. (Doignon & Falmange 2012.)

A simplified example of a knowledge structure for a domain Q = {a, b, c, d, e}, is given in figure 3**.** The knowledge structure in the figure contains 11 states. Among them are also the domain Q and the empty set ∅. When the student has answered correctly questions a, b, and c, we can surmise that he or she can also answer question d correctly. (Doignon & Falmange 2012.)

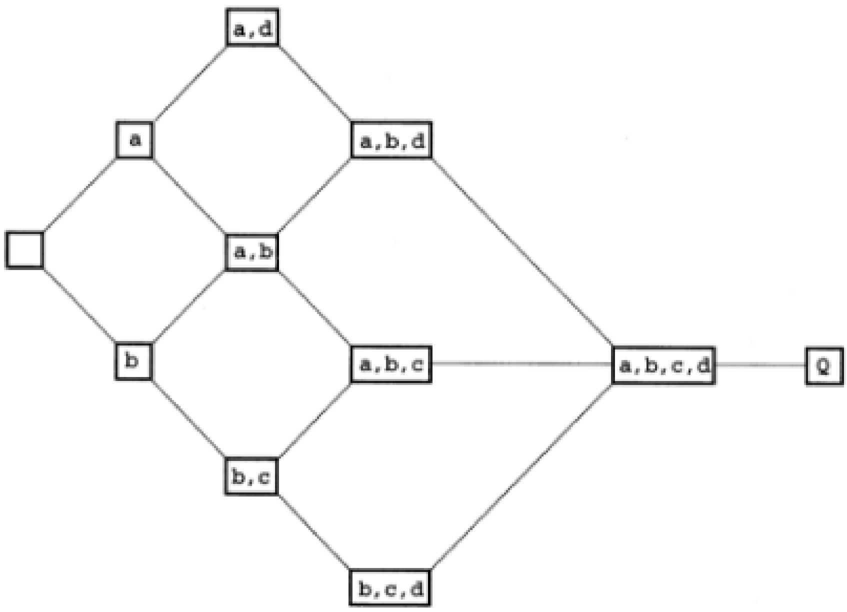

FIGURE 3 Simple Example of Knowledge Structure (Doignon & Falmange 2012, 4)

Because we want to both describe knowledge structures and reflect performance related to certain competencies meaningfully, it is good to review the theory from Korossy (1999).

2.3.3 Competence Performance Theory

Competence Performance Theory (CPT) originates from knowledge spaces theory but is enriched with integration between competence and performance. (Korossy 1999.) There are two levels in CPT. The aim is to make a clear distinction between competence (skills, ability) and performance. In this model, the selected knowledge domain is modeled in the form of a competence space. A model of performance space is based on an empirical representation. (Korossy 1999.)

Competence (ability, skills) level describes the presence that may explain the observable behavior. *Performance* level describes the observable behavior of the person on the set of domain-specific problems. The basic concept is a mathematical structure termed a *diagnostic*. This creates the correspondence between the competence and the performance level. When the family of competence states and the family of performance states are union-stable and the set of competence states are mapped onto the set of performance states with a union-preserving function, term *union-stable diagnostics* is used. (Korossy 1999.)

Figure 4 describes the position of HR Manager based on the study of Ley and Albert (2003). It is a simplified view of Korossy's approach and shows the competence-performance structure for the HR Manager. There are defined tasks for the position (1.1, 2.1, 4.2, 5.4) and the competencies required to accomplish these tasks (A, B, I, J, L, M, O, P). In the right side, learning paths that proceed from the bottom to the top structure are shown. According to theory, in the figure is a set P of tasks (A, B, I, J, L, M, O, P), which have to be carried out in a position. Subsets of P are called performance states if a person can accomplish the tasks (performances) in the state. Performance space P is a collection of the performance states associated with a particular union. (Ley and Albert 2003.)

Competencies	Tasks 1.1	2.1	4.2	5.4
A			X	
B		X	X	
I		X	X	X
J		X	X	X
L		X		
M			X	
O	X	X	X	
P	X	X	X	

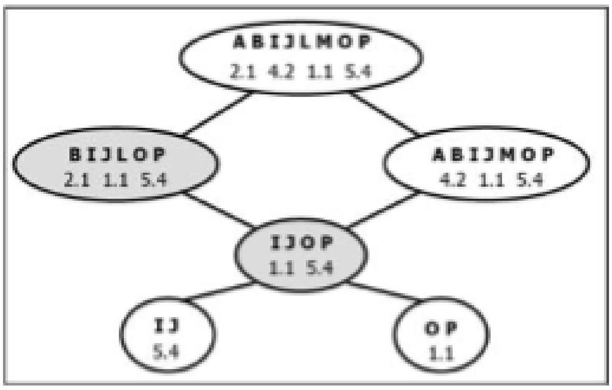

FIGURE 4 Part of Competence-Performance Structure for the HR Manager. (Lay and Albert 2003, 1505.)

As we stated above, CPT can be used as a diagnostic framework for a competence diagnosis or goal-oriented learning processes in organizations (Korossy 1999). The next chapter presents a description of an example of the competence-performance approach in practice.

2.3.4 Competence Performance Theory in Practice

One example of validating the competence-performance approach of Korossy (1999) has been presented by Ley and Albert (2003). In their study, it is shown how the methodology and resulting structures can be validated in an organizational setting. The issue in the study of the Ley and Alberts is quite similar to the topic of this thesis and is thereby well-suited to the frame of the empirical study. Therefore, it is described here in detail.

Ley and Albert (2003) consider skills management in knowledge-based organizations where job requirements change frequently and employees are responsible for their work processes. Their study is based on documents employees have created as part of their usual work assignments. They have evaluated these documents based on the knowledge spaces and its extension of Competence Performance Theory. As we noted above, the advantage of the competence-performance approach is that competency can help to predict the performance outcomes and provide an explanation for differences in performance. For example, missing competencies can explain why an employee is not able to complete a task. (Ley & Albert 2003.) Table 1 describes summarized how the methodology for validating competence-performance structures was employed in Ley and Albert's (2003) study.

TABLE 1 Steps for Deriving and Validating Competence-Performance Structure (Ley & Albert 2003, 1507)

Steps of a competency identification process	
1	Examining the purpose and setting of competency modelling
2	Selecting a position and a group of employees
3	Selecting the set of performance outcomes
4	Performing competency elicitation
5	Obtaining individual competence-performance structures
6	Obtaining organisational competence-performance structures
7	Validating the results

Ley and Albert start the case study by examining the purpose and setting of competency modeling. One purpose of their study was to provide input for the strategy formulation process. Their aim was also to explicate what kind of knowledge and skills project managers had acquired in the projects conducted thus far. They wanted to provide more transparency of the available expertise in the organization. An aim of the study was also to provide insight into the job requirements of project managers in order to develop expertise in future. (Ley & Albert 2003.)

The second step was to select a position and a group of employees whose expertise would be evaluated. In this case study, Ley and Albert wanted to use project managers' profiles. Seven employees were involved. The work outcomes were chosen on which performance can be assessed. Ley and Albert decided to use documents that had been written by the project managers as the performance outcomes. (Ley and Albert 2003.)

The next task was to determine which competencies were used to create the the documents. Ley and Albert (2003) evaluated skills and knowledge that were used to create some but not all of the documents. After eliciting competencies, it was time to obtain competence-performance structures for individuals. Each respondent created matrixes, which contained competencies used in creating documents. Based on matrixes, the competence-performance structures for individuals could be derived. (Ley and Albert 2003.)

The next step is to determine the organizational competence-performance structures from the matrixes described by the individuals. Ley and Albert chose to use a content analysis to obtain an organizational competence-performance structure. At first, they categorized competencies found by respondents to similar sets. Based on the categorization, they could draw a conclusion on an organizational level. In the second round, participants were asked to make document-competency assignments for the same documents used in the first part. The study was implemented in a questionnaire-type format and included competencies taken from first part of the study. The results of the survey were then combined into an organizational matrix. These organizational matrixes help to develop competency and support development planning of employees by describing the competencies that need to be developed. Table 2 shows a subset of competencies that are needed when communicating with customers and team members. Table 2 also shows a surmise function σ for the competence space. It seems reasonable to assume that competency E ("Effective interview techniques") is needed as a prerequisite for all the other competencies shown in Table 2. Competence C ("Understanding goals of others"), instead, is independent of the others. (Ley & Albert 2003.)

TABLE 2 Surmise Function σ for Competence Space (Ley & Albert 2003, 1511)

$e \in E$	Short Description	$\sigma(e)$
A	Communication about customer requirements	{A,D};{A,B,F}
B	Discussing ideas and concepts on an informal level	{A,B,D};{A,B,F}
C	Understanding goals of others	{C}
D	Discussing a common practice in a team	{A,D};{D,G}
E	Employing effective interview techniques	{A,B,C,D,E,F,G}
F	Presenting and selling own ideas	{F,G};{A,B,F}
G	Defining goals and persuading others	{F,G};{D,G}

Competencies listed in the table could be used in plans for the organization. Competence space can be obtained based on the table, in the same way as presented in figure 4. Using competence-performance structure, for example, performance weakness can be assessed, and required competencies can be derived. (Ley & Albert 2003.)

This topic will be returned to when describing the research process. The next chapter explains relevant issues for this study regarding data analytics.

3 Data Analytics Oriented Research

We are living in a world where the amount of data is increasing rapidly. However, raw data has no value by itself. Instead, the information contained in it must be extracted. In many organizations, *analytics* is a trending buzzword. There is a growing interest in understanding data and taking advantage of data, which is available in enormous quantities.

Next, the most basic terms concerning the subject, data, and big data will be introduced. After that, to understand what data analytics includes and how the data can create advantages for organizations, data analytics history will be reviewed. After that, other related terms will be defined.

3.1 Definition of Terms

3.1.1 Data

There are an enormous number of definitions of *data*. Zins (2007) documents 130 of definitions of *data* formulated by 45 scholars in his article. In Zins' article, data is represented by three other essential building blocks: information, knowledge, and wisdom. (Zins 2007.) These basic structures are described as a hierarchical model often depicted as a pyramid with data at its base and wisdom at its apex. It is called the data–information–knowledge–wisdom (DIKW) hierarchy. The pyramid is shown in figure 5. (Rowley 2007.)

Data (the plural form of the Latin word *datum*) can be described as the facts that are the result of observation or measurement (Landry et al., 1970). Stonier (1997) defines data as a series of disconnected facts and observations. Davenport & Prusak (2000) define data as a set of discrete, objective facts about events.

Data can be converted to information by analyzing, cross-referring, selecting, sorting, summarizing, or organizing it. Patterns of information can be worked into a coherent body of knowledge. According to Stonier, knowledge consists of an organized collection of information. Such information patterns form the basis of the kind of insights that we call wisdom. (Stonier 1997.)

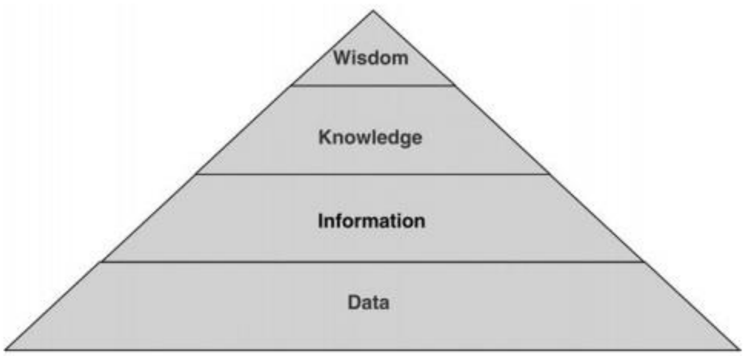

FIGURE 5 DIKW Hierarchy (Rowley 2007, 163)

According to Davenport and Prusak (2000), unlike data, information has meaning. It is meant to change the way the receiver perceives something and to have an impact on his or her judgment and behavior. They note that information must inform; data makes a difference.

In contrast to Rowley (2007), Stonier (1997) and some other researchers (Davenport & Prusak 2000) identify only three entities: data, information, and knowledge. They believe that firms have enough difficulty distinguishing among three related concepts. However, they are not inclined to address more. Instead of more entities, they incorporate higher-order concepts such as wisdom and insight into knowledge. According to Davenport and Prusak (2000), knowledge is a mix of framed experience, values, contextual information, and expert insight. It provides a framework for evaluating and incorporating new experiences and information. They also reveal that knowledge often becomes embedded in organizational routines, processes, practices, and norms, but also in documents or repositories. (Davenport and Prusak 2000.)

Based on the previous views, in this thesis data is conceived as objective facts about events or observations. The aim is to find competencies that are needed to convert data into knowledge or wisdom so that it is most valuable to organizations.

3.1.2 Big Data

One buzzword of the day is "Big Data." Big data has been described by Sagisogu and Sinanc (2013) as massive datasets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. Big data has also been considered as data coming from various channels including sensors, satellites, social media feeds, photos, video and cell phone, and GPS signals (Rich 2012). Big data analytics has been described as a process of research into massive amounts of data to reveal hidden patterns and secret correlations (Sagiroglu & Sinanc 2013).

Some scholars and practitioners have defined big data in terms of three Vs: volume, velocity, and variety, which distinguishes it from regular-sized data. *Volume* refers to the quantities of data, which grows by about 2.5 exabytes each day, and that number is doubling every 40 months or so. *Velocity* indicates the speed of data creation, and it is even more important than the volume. Real-time information can provide a competitive advantage for a company and help it to be much more agile than its competitors. *Variety* means the different forms of data available. Data can be structured, unstructured, or semi-structured. It can be from many sources. Data can be from messages, updates, and images posted to social networks, readings from sensors, GPS signals from cell phones, and so on. (Russom 2011; McAfee et al. 2012; Kwon & Sim 2013; Sagiroglu & Sinanc 2013; Gartner 2016.) In addition to these Vs, a fourth "V," for *value*, has been added to emphasize the importance of the benefits for business (Vesset et al. 2012). White (2012) has suggested that a fifth dimension, *veracity,* should be added to prior definitions of big data. Veracity highlights the importance of quality data and the level of trust in various data sources. White points that the quality of the data is important because when data integrated with other data and information, a false correlation could result in the organization making an incorrect analysis of a business opportunity. (White 2012.)

Quitzau (2013) from IBM has added to the definition of big data considerations about open data. Open data is provided with no cost and no license constraint. He adds another "V," *visibility*, which raises issues of privacy and security. Value and visibility are also separate from other big data Vs. Quitzau offers a few new considerations regarding open data, which can create opportunities for companies but also more challenges. (See figure 6)

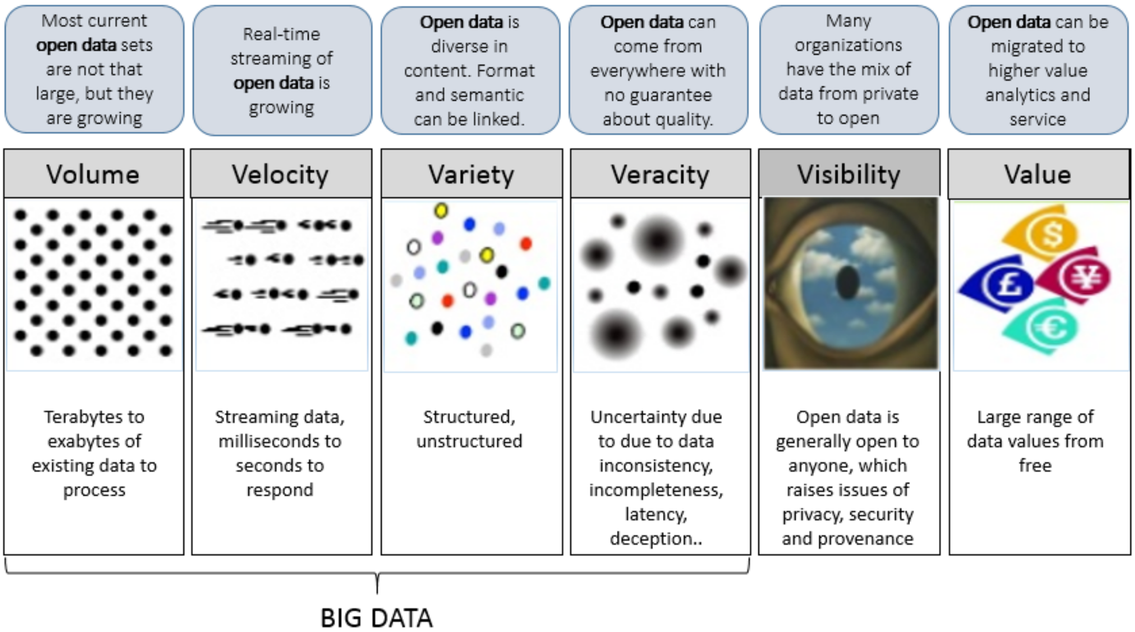

FIGURE 6 The six Vs of Big Data (Adapted from Quitzau 2013, 24)

While more data has been recorded in the past few years than in all previous human history, it has presented an enormous opportunity and challenge for organizations (IBM 2013; Waller & Fawcett 2013). However, according to the research of Davenport and Dyche (2013), many large firms view the vast amount of available data as something they have been wrestling with for a while and find it to be business as usual, and not something revolutionary. More than volume, the aspect of data that interests large firms is variety. Companies can have a much more complete picture of their customers and operations by combining unstructured and structured data. (Davenport & Dyche 2013.)

3.1.3 Data Analytics

A good starting point for figuring out what data analytics encompasses is a review of history. Use of data analytics can be at very different levels in different companies. A description of the history helps us to understand how analytics can help businesses.

Analytics before Big Data

In studying the literature, the first mentions of data analytics were found in 1954. At that time, George Smith, CEO at UPS, raised a question concerning the importance of making decisions by analyzing problems based on data rather than intuition (UPS Company Profile in Analytics 2010). In 1962 John W. Tukey wrote an article titled "Future of Data Analysis." He considered himself to be a statistician but said that his central interest was data analysis. He wanted to use statistics to analyze real-world problems. His article presents more questions and general considerations related to analytics than direct and clear answers, but it was a good starting point for the development of data analytics. (Tukey 1962.) Shortly after that Peter Naur (1966) suggested that the term *datalogy* be used to express the idea that science revolved around data and treatment. The alternative term he defined clearly a little later was the *science of data,* or *data science*. (Naur 1974).

In the late 1990s Thomas Davenport noticed the enormous interest in knowledge. Even if the subject was age-old, firms suddenly began to understand that a more than casual (or even subconscious) approach to corporate knowledge is required if they want to succeed. It was also noticed that firms were looking for best practices, new ideas, creative synergies, and breakthrough processes. These kinds of results could come only from making efficient use of knowledge. (Davenport 1998.)

In the 2000s, the practice of analytics started to mature. In 2001 William S. Cleveland published an article titled "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." Cleveland proposed data science as an independent field of study. He also named six areas in which he believed data scientists should be educated: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory. (Cleveland 2001)

The period from the beginning of the use of analytics tools until about 2009, is characterized as the era of "business intelligence." It can also be called Analytics 1.0. It was the time for real progress in gaining an objective, deep understanding of important business phenomena and giving managers the fact-based comprehension needed to make decisions based on data rather than intuition. In that period, data about production processes, sales, customer interactions, and more was recorded, aggregated, and analyzed for the first time. (Davenport 2013; Davenport & Dyche 2013.)

The key in that time was new computing technologies. New competencies were required to manage data. Datasets were relatively small in volume and static enough in velocity to be segregated in warehouses for analysis. However, preparing a dataset for inclusion in a warehouse was not easy. Preparing data for data analysis took a long time, and analysis itself required relatively little time. Because the whole analysis process was slow, often taking weeks or months to perform, it was vital to figure out the right questions on which to focus. Organizations saw analytics as a competitive advantage and used it to make better decisions based on what happened in the past, but they offered no explanations or predictions. (Davenport 2013.)

The Second Wave of Analytics

The next period started when internet-based and social-network firms (e.g., Google, eBay) began to amass and analyze new kinds of information. The period from 2005 to 2012 is called Analytics 2.0. Big Data was noticed, and it changed the nature of analytics and computing architectures. 2.5 quintillion bytes of data was created each day and storage of data, and high-speed processing became cost-effective. Powerful tools for mass data processing become accessible, and the need for professionals to utilize them grew exponentially. (Davenport 2013; Davenport & Dyche 2013; Vesset et al. 2012.)

At the same time, it became common for the most advanced businesses to use analytics. For example, Google applied algorithms to web searches. In 2005 Google Analytics was introduced, the online experience started to be more personalized, and real-time analytics became more common. Also, unstructured data began to have analytical value. (Davenport 2013; Davenport & Dyche 2013.)

It was noticed that big data can bring about dramatic cost and time reductions. It also helped to offer brand-new products and services based on big data or, like traditional analytics, support for internal business decisions. Even small benefits could provide large payoffs when adopted on a large scale. (Davenport & Dyche 2013.)

The first movers recognized advantages of using data. It led not only to an impressive level of hype but also prompted an unprecedented acceleration of new offerings. For example, LinkedIn built a strong infrastructure and hired smart, productive data scientists. They helped create numerous data products, including People You May Know, Jobs You May Be Interested In, Groups You May Like, Companies You May Want to Follow, Network Updates, and Skills and Expertise. (Davenport 2013.) Netflix created the Network Prize, which could optimize movie recommendations for customers. The company is now using big data to help in the creation of proprietary content, including the *House of Cards* series. (Leonard 2013; Davenport & Dyche 2013.)

Big data could not fit or could not be analyzed fast enough on a single server, so Hadoop, an open source software framework for fast batch data processing across parallel servers, was developed. Companies turned to a new class of databases known as NoSQL to deal with unstructured data. Public and private cloud computing environments were used to store and analyze information. Other technologies introduced during this period include "in memory" and "in database" analytics for fast processing. Machine-learning methods developed and were used to generate models rapidly from the fast-moving data. Also, visual presentations of the information took a step forward. (Davenport 2013.)

Competencies needed for utilization of Analytics 2.0 were different than in the Analytics 1.0 period. The next-generation quantitative analysts started to be called data scientists. They wanted to work on new product offerings and help to shape the business. During the 2.0 period, pioneering data firms began to invest in analytics to support customer-facing products, services, and features. They attracted customers through better search algorithms, recommendations from friends and colleagues, suggestions for products to buy and highly targeted ads. All of these were based on enormous amounts of data and driven by analytics. (Davenport 2013.)

The Third Wave of Analytics

Next new era of analytics, Analytics 3.0 begin when other large organizations started to follow suit. Today, data analytics is widespread not only in information firms and online companies that can create products and services from analyses of data but every company in every industry. Increasing amounts of data are now available about making things, moving things, consuming things, or working with customers because every device, shipment, and consumer leaves some data. There is the ability to analyze those sets of data for the benefit of customers and markets. There is also the ability to embed analytics and optimization into every business decision made on the front lines of an organization's operations. (Davenport 2013.)

Big data is still a popular concept, but Analytics 3.0 combines the best features of 1.0 and 2.0. It is a blend of big data and traditional analytics. Although it is early days for this new model, the traits of Analytics 3.0 are becoming apparent in large organizations. (Davenport 2013.) As Davenport and Dyche discovered in their study, in which they interviewed 20 large organizations, in not even one of these large organizations, was big data being managed separately from other types of data and analytics. That integration was leading to a new management perspective on analytics, which is called Analytics 3.0. (Davenport & Dyche 2013). Next, some attributes that describe Analytics 3.0 organizations will be introduced.

Multiple Data Types, Often Combined

Organizations need to integrate large and small volumes of data from both internal and external sources. Structured and unstructured formats are also combined to yield new insights in predictive and prescriptive models. For example, the goal can be to improve the efficiency of the company's route network, to lower the cost of fuel, and to decrease the risk of accidents. Data can be added from sensors to logistical optimization algorithms, allowing companies to monitor key indicators such as fuel levels, container location and capacity, and driver behavior. (Davenport 2013.)

A New Set of Data Management Options

In the 1.0 era, IT organizations had well-organized operational data in the data warehouses. In the 2.0 era, a new set of options became available, and they were focused mostly on Hadoop clusters and NoSQL databases. Now there are even more of choices: data warehouses, databases, big data appliances, environments that combine traditional data query approaches with Hadoop (these are sometimes called Hadoop 2.0), vertical and graph databases, and so on. The old formats are still available, but new processes are needed to move data and analysis across staging, evaluation, exploration, and production applications. (Davenport 2013; Davenport & Dyche 2013.)

Faster Technologies and Methods of Analysis

Big data technologies include a variety of technologies, such as hardware/software architectures, including clustered parallel servers using Hadoop/MapReduce, in-memory analytics, in-database processing, and so on. These all are considerably faster than previous generations of technology for data management and analysis was. In addition, "agile" analytical methods and machine-learning techniques are being used to produce insights at a much faster rate. The challenge in the 3.0 era is to adapt operational, product development, and decision processes to take advantage of what the new technologies and methods can bring forth. (Davenport 2013; Davenport & Dyche 2013.)

Integrated and Embedded Analytics

Models in Analytics 3.0 are often embedded into operational and decision processes. That dramatically increases their speed and impact. Analytics can be embedded into fully automated systems through scoring algorithms and analytics-based rules. Integrating analytics into systems and processes means greater speed. For decision makers, it makes it harder to avoid using analytics—which is usually a good thing. (Davenport 2013.)

Analytics can also be built into consumer-oriented products and features. Previously, analytics were used mostly to support internal decisions. It is still useful, but recently companies have started using data and analytics to create new products and services. Digital players such as Google and LinkedIn but also mainstream firms such as GE and several large banks are pursuing data products. Managers need to understand Analytics 3.0 to be aware of this a new option. (Davenport 2015B.)

Data Discovery

Enterprise data warehouses were initially intended to facilitate exploration and analysis. However, getting data from them is time-consuming. Companies need a capable discovery platform for data exploration, to develop products and services based on data. Data discovery environments make it possible to determine the essential features of a dataset without a lot of preparation. (Davenport 2013.)

Cross-Disciplinary Data Teams

In online firms and start-ups, a data scientist or data analyst is often able to operate with relative independence. In larger and more conventional organizations, they must collaborate with a variety of other roles to ensure that analytics matches data. In this picture, interdisciplinary expertise is needed, and often traditional analytics experts will be placed outside of their comfort zone. Still, a rare individual can address all the questions. The combined team takes on whatever is needed to get an analytical job done, with frequent overlap among roles. (Davenport 2013.)

Chief Analytics Officers

Organizations need senior management oversight to be competitive in analytics. For that reason, companies have created “chief analytics officer” roles to superintend the building and use of analytical capabilities. (Davenport 2013.)

Prescriptive Analytics

There are different types of analytics: descriptive, diagnostic type, predictive, and prescriptive. The new wave of analytics includes all previous types, but it emphasizes the last one. Prescriptive models provide a high level of operational benefits but require high-quality planning and execution in return. For example, if the system gives incorrect routing information to drivers, it won’t be around for long. Executives have said that they have spent much more time on change management issues than on algorithm and systems development. (Davenport 2013; Davenport & Dyche 2013.)

New Ways of Deciding and Managing

To stay competitive and to power the data economy with analytics, companies need new approaches to decision making and management. Directors must have experience in the exploitation of the data and support data-driven experimentation. Small-scale experimentation needs to be used systemically with rigorous controls to permit the determination of cause and effect. It is always better to arrange limited experiments than wholesale changes, which can turn out badly. Managers should establish guidelines for the cases, in which the early warning leads to decisions and activities. (Davenport 2013; Davenport & Dyche 2013.)

Using prescriptive analytics also requires changes in the way frontline workers are managed. Companies will gain unprecedented visibility into the activities of employees who wear or carry sensors. Workers will undoubtedly be sensitive to this monitoring. Just as analytics that is intensely revealing of customer behavior can cause discomfort, overly detailed reports of employee activity have a certain “creepiness” factor. (Davenport 2013; Davenport & Dyche 2013.)

Creating Value in the Data Economy

Even Analytics 3.0 would not represent the ultimate form of competing on analytics, but it can be viewed as the point when participation in data analytics went mainstream. Online companies did not need to reconcile or integrate big data with traditional sources of information and the analytics performed on it. For the most part, they did not have those traditional sources. They also did not need to merge big data technologies with traditional infrastructures because those infrastructures did not exist. But, as Davenport (2013) notes, each of these companies now has its own version of Analytics 3.0. The big data model was a huge step forward, but businesses that want to prosper in the new data economy must once again fundamentally rethink how the analysis of data can create value for themselves and their customers because big data will not provide an advantage for much longer. (Davenport 2013; Davenport & Dyche 2013.)

Smart, Connected Products
To the list of attributes Davenport and Dyche (2013) present could be added smart, connected products, which offer new capabilities. Porter et al. (2015) are not talking about the Internet of Things (IoT), in which separate physical components are connected to the web. They are talking about smart, connected products in which IT is embedded. Smart products have many of the same physical components that products have always had as well as new features that make them more intelligent. They also generate new data flows, new ways to store and manage product data, and new ways to create a competitive advantage using analytics. (Porter et al. 2015.)

3.1.4 Types of Analytics

Companies can benefit from many types of analytics. Gartner classifies analytic types into four categories: descriptive, diagnostic, predictive, and prescriptive analytics (Laney 2012). Types or maturity of analytics has also been classified, for example, by SAP (SAP 2013), Strategy At Risk (Strategy At Risk 2012) and Herman et al. (2013). These models have some differences, such as in the model of SAP (SAP 2013) the descriptive analytics has been separated into smaller parts, such as raw data, cleaned data and standard, and ad hoc reports. However, the idea behind the classifications is the same in all these models. The types of analytics are described in the frames of Gartner model (Laney 2012). Different types are illustrated in figure 7.

Descriptive Analytics: What Happened?
Most businesses start using analytics with descriptive analytics. It means that they use data to understand past and current business performance and make informed decisions. Descriptive analysis techniques categorize, characterize, consolidate, and classify data to convert it into useful information for understanding and analyzing business performance. Data can be summarized into meaningful charts and reports, for example, about budgets, sales, revenues, or cost. Reports can be standardized or customized, and reports can also be viewed in more detail and queries can be made (e.g., to understand the impact of an advertising campaign, to review business performance to find problems or areas of opportunity, or to identify patterns and trends in data). Descriptive analysis can also be the basis of predictive analytics. (Evans & Lindner 2012.)

Diagnostic Analytics / Generic Predictive Analytics: Why Did It Happen?
Gartner includes diagnostic analytics for analytics types. With it, the reason for what happened can be diagnosed. In the maturity model from SAP (2013), it is called generic predictive analytics. Diagnostic or generic predictive analysis is exploratory data analysis of existing or additional data to discover the root causes of a problem. (Banerjee, Bandyopadhyay & Acharya 2013; Laney 2012; SAP 2013.)

Predictive Analytics: What Might Happen?

Predictive analytics or predictive modeling examines historical data and combines it with rules and algorithms. It can detect patterns or relationships in data. After that, it extrapolates relationships forward in time. The purpose of predictive analytics is to analyze past performance in an effort to predict the future. (Evans & Lindner 2012.)

Predictive analytics is a broad term that describes a variety of statistical and analytical techniques. Advanced prediction techniques are a combination of statistics, data mining, and machine learning to find meaning from large amounts of data. There are two major types of predictive analytics: supervised and unsupervised. (Eckerson 2007.) Chapelle, Schölkopf, and Zien (2006) add semi-supervised learning to the types of predictive analytics.

Predictions can be based on the training sample of previously solved cases, or historical data that contains the results you are trying to predict. This kind of process of creating predictive models is called *supervised learning* or learning with a teacher. Approaches to supervised learning include classification, regression, and time-series analysis. For example, classification can be used, if you want to know which customers are likely to respond to a new direct mail campaign. The results of past campaigns can be used to train a model to identify the characteristics of individuals who responded to that campaign. Regression is used in forecasting. Variance analysis and time-series analysis in addition to regression analysis understands the unique properties of time and calendars. (Eckerson 2007; Chapelle et al. 2006.)

Unsupervised learning does not use previously known results to train its models. Rather, it directly infers the properties of the probability using descriptive statistics to examine the natural patterns and relationships that occur within the data. One well-known example of unsupervised learning and its associated technique is market basket analysis. It can be used for identifying products and content that go well together. (Eckerson 2007; Chapelle et al. 2006.)

Semi-supervised learning is halfway between supervised and unsupervised learning. It means that the algorithm is provided with some supervision information but not necessarily for all examples. (Chapelle et al. 2006.)

Analytic experts build analytic models using a variety of techniques. Examples of these are neural networks, decision trees, linear and logistic regression, naive Bayes, clustering, and association. Models can be implemented using a variety of algorithms with unique characteristics that are suited to different types of data and problems. When creating efficient analytic models, analysts need to know how to determine which models and algorithms to use. Fortunately, many analytic workbenches automatically apply multiple models and algorithms to a problem to find the combination that works best. This advance has been made it possible to create relatively effective analytical models without requiring deep expertise. (Eckerson 2007.)

Prescriptive Analytics, Optimization, Simulation: How Can We Make It Happen?

The promise of prescriptive analytics is that it enables decision-makers not to look only into the future of their mission critical processes and see the opportunities (and issues) that are potentially out there. It also presents the best course of action to take advantage of that foresight promptly. (Basu 2013.)

Prescriptive analytics goes beyond describing, explaining, and predicting. It associates alternatives with the prediction of outcomes to help analysts think about what might happen in the future and then to optimize their approach to achieve the best response or action necessary to achieve business objectives, given the limited resources of the enterprise business. Prescriptive analytics can use hybrid data, a combination of structured and unstructured data, and business rules to predict what lies ahead and to prescribe how to take advantage of this predicted future without compromising other priorities. (Riabacke et al. 2012; Danielson & Ekenber 2012; Banerjee et al. 2013.*)* Examples of prescriptive analytics are recommendation systems such as those used by Netflix, Google, and Amazon.

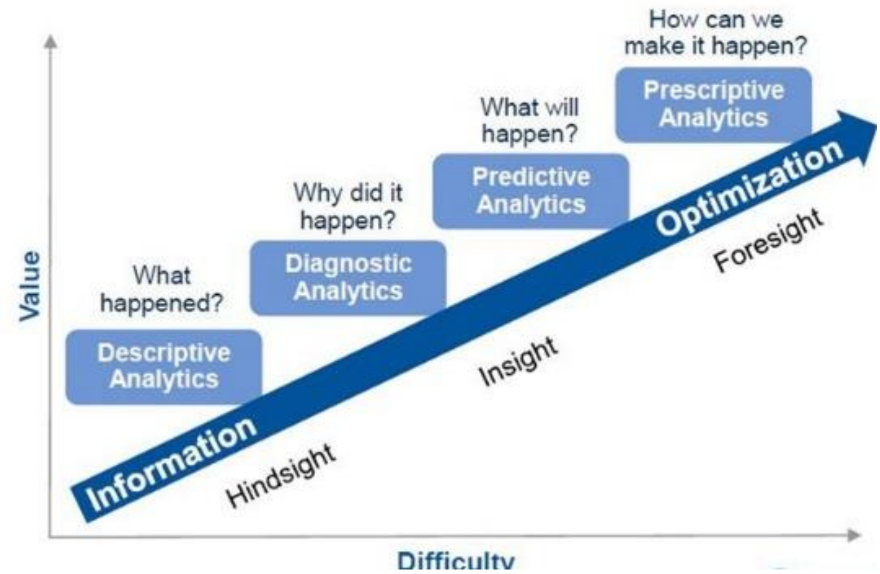

FIGURE 7 Analytics Types according to Gartner (Laney 2012, 33)

Advanced Analytics

In addition to the most commonly classified types, advanced analytics is often spoken about. Its definition depends on the date and context. In 2003 it was defined as being the same as predictive analytics (Bose 2009). Recently it has been described as solving problems combining predictive analytics and prescriptive analytics. Often it is also used as a general term that describes advanced analytics as a broad category of inquiry that can be used to help drive changes and improvements in business practices. (Rouse 2013.)

According to Gartner, advanced analytics means the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools. It is typically beyond those of traditional business intelligence (BI). Advanced analytics discover deeper insights, make predictions, or generate recommendations. Techniques used in advanced analytics include data/text mining, machine learning, pattern matching, forecasting, visualization, semantic analysis, sentiment analysis, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, and neural networks. (Gartner 2016, IT Glossary.)

3.1.5 Data Science

Data science is increasingly commonly used term nowadays. Advanced analyzing techniques had been used already long time amongst scientist, but once they were noticed in business life, the use of term data science began. Also the availability of big data and major advancements in machine intelligence created significant new questions and opportunities. (Davenport 1998; Dhar 2013; Agarwal & Dhar 2014.)

Herman et al. (2013) describe data science briefly as the art of turning data into actions. Provost and Fawcett (2013) are of the opinion that data science consists of principles, processes, and techniques for understanding phenomena via the analysis of data that can be automated. According to them, the ultimate goal of data analysis is to improve decision making across the firm, thus directly benefiting business. (Provost & Fawcett 2013A.)

When considering the differences compared with statistics and other existing sciences, the raw material (i.e., the data) is initially heterogeneous and unstructured. It is often coming from networks that have complex relationships among their entities. Analysis of that kind of data, including the combination of different types of data, requires integration, interpretation, and sense-making. It is increasingly derived through tools from computer science, linguistics, econometrics, sociology, and other disciplines. (Dhar 2013.)

From an engineering perspective, scale matters in that it changes the traditionally used methods and models, which are somewhat inadequate for knowledge discovery. They are optimized for fast access and summarization of data, given what the user wants to ask, or a query. In data science, instead, patterns can be discovered in a massive set of data even when users lack a well-formulated query. Discovery asks "What pattern satisfies this data?" whereas database query would ask "What data satisfies this pattern (query)?" Often in data science, the main aim is to find interesting and robust patterns that satisfy the data, where "interesting" is something unexpected and "robust" is a pattern that is expected to occur in the future. (Dhar 2013.)

The terms *data science* and *data mining* are often used interchangeably. Data science is a more broadly used term than data mining. Techniques used in data mining (e.g., machine learning, pattern recognition, artificial intelligence, neural networks) are the same as in data science, but data science also involves much more than just data-mining algorithms. (Provost and Fawcett 2013.)

Clear differences between data analytics and data science could not be found from scientific articles. The difference does not appear to be clear, and the definition of Herman et al. (2013) for data science, "the art of turning data into actions," could describe either terms. However, as said before, use of the term *data science* came into use at the time when advanced analyzing techniques, big data, and major advancements in machine intelligence began to create significant new questions and opportunities. (Dhar 2013; Agarwal & Dhar 2014.) Data analytics was already in use at an earlier date. Its use has not been limited to big data or the most advanced technologies. It also includes traditional analytics. When organizations are analyzing small data, and utilizing other than the most advanced technologies can be beneficial, in this study we prefer to use the term *data analytics* instead of *data science*, unless directly referring to the literature. The same applies to use of *data analyst* and *data scientist*. (Davenport 2011; Davenport & Dyche 2013.)

3.1.6 Data-Driven Organization

Nowadays there is a lot of talk about data-driven organizations. Data-driven decisions, made by evidence are better decisions than those based on intuition. It's as simple as that (Patil 2011; McAfee et al. 2012). In data-driven organizations, behaviors, practices, and beliefs are consistent with the principle that business decisions at every level are based on analysis of data (Kiron et al 2012). Patil & Mason (2015) describe that "A data-driven organization acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on, and develop new products, and navigate the competitive landscape."

In data-driven organizations, for example, before accepting a new service offering or operational approach, the leader asks for the analytics to support it. Such organizations engage in working with high value by using analytics for day-to-day operations. (Kiron et al 2012.)

3.2 Data Analytics Competency

As mentioned earlier, making sense of data and turning it into knowledge has been discussed by scientists, statisticians, computer scientists and others for years. Especially in recent years, the need for people who are fluent in working with data has grown exponentially. Those professionals who make insights from data by analyzing it are often called data scientists in the literature. As mentioned earlier, the title became used more frequently when big data came along. At first, this usage was common in firms like Google, eBay, LinkedIn, and Facebook, in which big data analytics was the only focus of analytics. (Davenport 2013; Granville 2014.) Granville (2014) even talks about "the real data scientists," who have many required skills and are working with big data. Instead, according to Granville (2014), "fake" data scientists can be, for example, statisticians who have not been involved in the big data revolution.

In many other firms, big data must integrate with more traditional sources of data and perform the analytics upon them. In these companies, naturally, the need is much broader than "real" data scientists. In these companies, other titles in the data analytics area are used more often, such as statistician, software engineer, developer, business analyst, big data architect, data engineer and so on. (Hammerbacher 2009; Davenport & Dyche 2013; Granville 2014; Miller 2014.) In this study, of interest are all the people who work daily in analytics area. In this thesis, they are called data analysts, if not directly referring to the literature. The competencies that are important for data analysts will be described in the next section.

3.2.1 Models and General Descriptions

Drew Conway created one of the well-known definitions of data science. He described data science using a Venn diagram based on his experience (figure 8). It represents the main skill set of a data scientist. According to Conwey (2010), three areas of expertise are essential for data scientists to master. Hacking skills mean that for acquiring and cleaning the data, a data scientist should be able to manipulate text files at the command-line, to understand vectored operations, and to think algorithmically. At the time when data is a commodity that is traded electronically, it is appropriate to use the term *hacking*. For extracting insight from the data, data scientists need to be able to apply appropriate math and statistics methods, which requires at least a baseline familiarity with those tools. Because data and math and statistics only provide machine learning, data science, motivating questions and hypotheses that can be applied to the data and tested with statistical methods are also necessary, and for this, substantive expertise is needed. The danger zone in the diagram means that people who have deep hacking skills and substantive expertise "know enough to be dangerous." (Conwey 2010.)

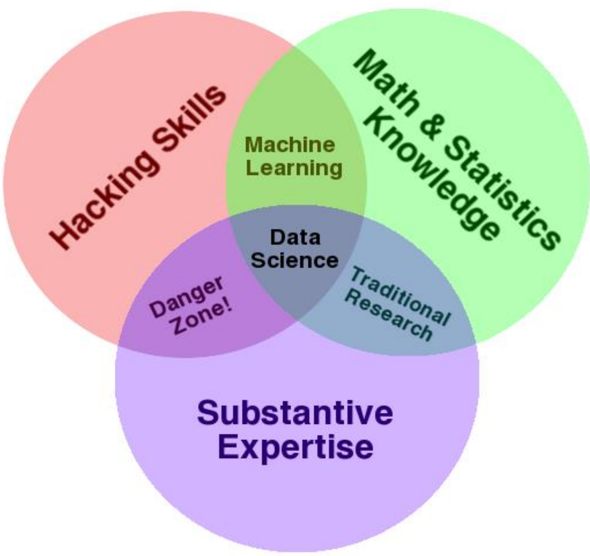

FIGURE 8 Data Science Venn Diagram (Conway 2010)

As said before, Conway's diagram is not a scientific description of the issue. Some have criticized it for overlooking some important things. However, it is a good starting point for a discussion of the skills required by data scientists and data analysts.

Many data scientists are formally trained in computer science, math, or economics, but they can emerge from any field that has a strong data and computational focus (Patil 2011; Davenport & Patil 2012). Harris, Murphy and Vaisman (2013) add to the list of more specific fields such as statistics, machine learning, databases, operations research, business intelligence, and social and physical sciences. Universities rarely provide any meaningful learning experiences designed to develop an understanding of the data science concept. (Harris et al. 2013.*)* Recently some universities have improved their degree programs and created new offerings. The brightest data scientists have come from very different backgrounds. Facebook’s and LinkedIn’s teams have included, for example, experimental physicists, oceanographers, computational chemists, and neurosurgeons, which is maybe the most surprising. (Davenport & Patil 2012; Patil 2011.)

3.2.2 Personal Traits

In many cases when evaluating essential competencies, the focus is on technical competencies. However, some personal capabilities are also extremely. These cannot be easily developed, but when considering an applicant's suitability for the job, it is advisable to take personal capabilities into account (Spencer & Spencer 1993).

Personal capabilities that characterize the most creative scientist in any field are curiosity and associative thinking. These traits are required to clarify the questions and hypotheses data scientists trying to solve. Data scientists have to be creative so that they can look at problems in innovative ways. (Patil 2011; Davenport & Patil 2012.)

Davenport and Patil (2012) published a well-known article concerning the competency of data analysts titled “Data Scientist: The Sexiest Job of the 21st Century.” In the article, they tell about the long career already of data scientist Jonathan Goldman, who in 2006 while working at LinkedIn, the business networking site, devised a way to link users’ connections with the names of people they hadn’t yet connected with but seemed likely to know. After exploring messy data and unwieldy analysis, he began to form theories, testing hunches, and finding patterns that allowed him to predict which network a given profile would land in. When LinkedIn’s top management recognized this as a good idea, it soon became a standard feature. Thanks to this one feature LinkedIn started to grow. (Davenport & Patil 2012.) As a result, the people-you-may-know feature has become a critical part of every social network’s offering. Facebook, Netflix, Google, and many other successful companies have their own versions of predicting of the unknown link.

The story of Jonathan Goldman is an example of the job of data scientist. He had an idea and formed a hypothesis about the idea, and he solved it proactively. He may have reviewed some earlier studies on the subject (e.g., the article of Watts and Strogatz from 1998 "Collective dynamics of 'small-world' networks." and Liben-Nowell and Kleinberg's study from 2004 "The Link Prediction Problem for Social Networks"), or maybe he just iterated the idea using his extensive know-how until it was ready to be presented to management. (Davenport & Patil 2012.)

Davenport and Patil (2012) describe a data scientist as a high-ranking professional with the training and curiosity to make discoveries in the word of big data. They also assert that data analysts need analytical skills. They need autonomy and but also need to be "on the bridge," responding to management issues with their managerial colleagues in real time. Data scientists inevitably face technical limitations. In these situations, they need to be creative, determined, and pressure tolerant when a great deal is expected of them. (Davenport & Patil 2012.)

Pinola's master's thesis (2015) focused on data scientists' competency through job advertisements. Pinola's (2015) study found that the most frequently mentioned competencies in the job advertisements for data scientists were creativity, pressure tolerance, and flexibility. A few job advertisements also mentioned that data scientists need to be able to operate in a changing environment and to be passionate. (Pinola 2015.)

3.2.3 Professional Skills

More than anything, data scientists make inventions while examining the data. Problem-solving skills are essential. As Davenport & Patil (2012, 73) describes, data analysts need to "understand how to fish out answers to important business questions from today's tsunami of unstructured information." (Davenport & Patil 2012.) Also Dhar (2013) states that the key differentiator to being an effective data scientist is the ability to formulate problems in a way that results in effective solutions. He also uses the term *computational thinking* to describe the core skills needed by data scientists in the coming decade. Wing (2006) also predicts that it will be a fundamental skill used by everyone in the world. The term *computational thinking* is originated by Papert (1996) and elaborated by Wing (2006). Barr and Stephenson (2011) have investigated the concept and capability of computational thinking. They identified next capabilities in their study: data collection, data analysis, data representation and analysis, abstraction, analysis and model validation, automation, testing and verification, algorithms and procedures, problem decomposition, control structures, parallelization, and simulation. These recommendations for computational thinking are made in the same spirit as the needed skills identified in other sources.

One point that was mentioned in many sources was that only deep technical competence is not enough and may be even less important than substantive competence. Waller and Fawcett (2013) and Provost and Fawcett (2013), for example, write, that professionals must have both analytical skills as well as business and management understanding. Data scientists need expertise concerning data-mining algorithms, but they also must be able to view business problems from a data perspective (Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013). The work should be focused on issues where it is easy to determine return on investment (ROI). Using analytics should be ROI-driven, and organizations should invest in technology only when it solves real business problems (Davenport & Dyche 2013). Data scientists need also be able to help decision makers' shift from ad hoc analysis to an ongoing conversation with data. Data scientists need to know business issues and have empathy for customers. (Patil 2011; Davenport & Patil 2012.)

Data scientists also need to know how to communicate in a way that all the stakeholders understand, both verbally and visually (Davenport & Patil 2012; Patil 2011). The requirements concerning reporting and communicating with stakeholders have changed from only showing charts and tables to storytelling (Davenport & Patil 2012; Davenport & Dyche 2013; Driest et al 2016).

Data scientists need to be able to identify rich data sources and create structures to contain large quantities of formless data. They need to be able to make analysis possible by joining data with other elements and incomplete data sources. They also clean the resulting dataset. (Patil 2011; Davenport & Patil 2012.)

In the study of Davenport & Dyche (2013) revealed that there is a real need for combining data scientist skills with traditional data management competencies. Many of the challenges concern the data itself. Solid knowledge of data architectures, metadata, data quality and correction processes, data stewardship and administration dashboards, master data management hubs, matching algorithms, and a host of other data-specific topics are important. According to the research, however, respondents on the "business side" (as opposed to IT) did not mention these data management capabilities, and in fact sometimes complained when IT groups tried to impose them. (Davenport & Dyche 2013.)

One of the needed competencies related to data analytics is change management knowledge. As prescriptive analytical models are embedded into key operational processes, someone needs to bring about needed changes in roles, process designs, and skills. (Davenport & Dyche 2013.)

3.2.4 Technical Knowledge

When speaking about technical knowledge needed for data analytics, the most basic, universal skill of data scientist is the ability to write code so that they can, for example, develop prototypes (Patil 2011). Also, machine learning, at least basic skills, seems to be becoming a necessity as companies try to make sense of massive datasets and seek to build automated and predictive decision systems (Patil 2011; Provost & Fawcett 2013). Dhar (2013) believes that knowledge about machine learning must build on more basic skills that fall into the following classes:

- Statistics
    - Especially Bayesian statistics, which requires a working knowledge of probability, distributions, hypothesis testing, and multivariate analysis
- Computer science skills
    - Awareness of how data is represented and manipulated by computers. Knowledge of data structures, algorithms, and systems, including distributed computing, databases, parallel computing and fault-tolerant computing. Together with scripting languages (e.g., Python and Perl), the above skills are the basic building blocks required for dealing with datasets.
- Knowledge about correlation and causation
    - A clear idea of the distinction between correlation and causality and the ability to assess which models are desirable and practical in different settings.

As said earlier, when the data can be heterogeneous and unstructured, techniques and tools from computer science, linguistics, econometrics, sociology and other disciplines can be needed (Dhar 2013). In addition, data analysts need technical skills to design and implement all the earlier discussed types of analytics. Data analysts need to know which models and algorithms to use and find the combination that is the most effective for a particular situation. In order to find the best analytical model, deep technical expertise is needed. Fortunately, workbenches that can help data analysts find effective combinations have been developed. (Eckerson 2007.)

Miller (2014) points out that data analysts need to master technical skills such as math and statistics, machine learning, predictive analytics, computer science and programming, in addition to core information systems, database, data warehousing, and data mining skills. He points out that businesses and other organizations need data professionals who understand basic machine learning concepts and can apply existing algorithms to a solution but that those organizations are also striving to differentiate themselves in an increasingly competitive market and will need hard-to-find, highly skilled, machine-learning data scientists who can invent innovative machine-learning algorithms. (Miller 2014.)

An empirical study of Pinola's (2015) concerning needed competencies of data scientists reported that the skills most frequently mentioned in job advertisements for data scientists were statistics, machine learning, and programming skills. Data mining, algorithms, and predictive analytics were also mentioned in many job advertisements. (Pinola 2015.)

Surprisingly, security and privacy are discussed very little in the academic literature. As Davenport and Dyche (2013) say when describing the current state of the use of big data in big companies, data security approaches are not yet fully developed in the Hadoop cluster environment. However, when developing connected products, security concerns become amplified. Security must be embedded as a first principle across the value chain, and data professionals need broader and deeper skills in information security (Miller 2014; Porter et al. 2015), individual privacy, and ethical use of data (Miller 2014).

3.2.5 Tools and Techniques for Data Analytics

In academic discourses, the necessary tools and techniques are not often precisely defined. It has also been mentioned that it is important for an organization to have an integrated information platform as well as a robust set of tools and skills concerning the tools. At least they get support for acquiring tools and expertise that bring state-of-art analytical capabilities into the organization. (Kiron et al. 2012.) When listing skills required by data scientists, it was also mentioned that they often fashion their tools and toolkits to accomodate their technical limitations (Davenport & Patil 2012; Patil 2011). However, in the analytics area, ready-made alternative tools are used frequently. The need for tools depends on the situation and on the skills of the user.

The study of Davenport and Dyche (2013) revealed that existing analytics or technology groups often add big data functions to their missions. Technology architectures and skills for big data in big companies are evolving and integrating with existing structures. Big companies have often made significant investments in their data warehouses and have neither the resources nor the will to simply replace an environment that works well and is doing what it was designed to do. For many of these companies, a coexistence strategy that combines the best of legacy data warehouse and analytics environments with the new power of big-data solutions are the best options. (Davenport & Dyche 2013.)

The most used technology for storing large and diverse amounts of data is Hadoop. The study of Davenport and Dyche (2013) revealed that many big company executives see Hadoop as a low-cost alternative for the archival and quick retrieval of large amounts of data. Hadoop uses a processing engine called MapReduce to distribute data across disks and to apply sophisticated computational instructions to that data (McAfee et al. 2012; Davenport & Dyche 2013). Conwey (2011) and Davenport (2014) indicates the need for competency in Hadoop as well.

Just as data varies with the business application, the code used to manipulate and process the data can vary. Required skills might include natural language processing or text mining techniques, video or image analytics, and visual analytics. Many of data analysts can code in scripting languages like Python, Pig, and Hive or many are simply strong programmers with some analytical skills. (Davenport & Dyche 2013.) In the literature came out also tools and techniques for analytics like R, SQL, NoSQL, Excel, Impala, SAS, IBM DB2, Vertica, Tableau, and Watson (Kiron et al 2012; Davenport & Dyche 2013).

In the master's thesis of Pinola (2015) the expertise in tools and technologies most frequently requested in job advertisements were R, Python, SQL, Java, and Hadoop. SAS, Hive, C++, Pig, and Matlab were also mentioned quite often. (Pinola 2015.)

Patil (2011) points out that tools are important because they allow automation, which saves time. Something as simple as reducing the turnaround time on a complex query from “get the result in the morning” to “get the result after a cup of coffee” represents a massive increase in productivity. If queries run overnight, analysts can only afford to ask questions when they already think they know the answer. If queries run in minutes, analysts can more easily experiment and be creative. (Patil 2011.)

It is rare that a single individual has all the holistic skills mentioned above. In practice, there is a team of experts who have different areas of expertise, and analytics tasks can be shared among them. (Davenport & Dyche 2013; Harris et al 2013; Waller & Fawcett 2013; Maruyama et al 2015.) Even when there are people with different kinds of competencies on a team, Patil believes the silos that have traditionally separated data people from engineering, from design, and from marketing, don’t work when an organization is building data products. He argues that every data project is a new experiment. It never would work to have a waterfall process in which one group defines the product, another builds visual mock-ups, a data scientist preps the data, and a group of engineers builds it to some specification document. (Patil 2011.)

3.2.6 Organizational Competencies

Even though skilled employees are crucial to competing in analytics, they are not enough to succeed. A strategy that considers data as a core business asset is also required. (Miller 2014.) When Davenport & Dyche (2013) investigated the role of big data in big companies and needed skills in organizations, the competencies whose approaches seemed most effective and likely to succeed had close relationships between the businesses groups addressing big data and the IT organizations supporting them. Patil (2011) points out that the interaction between the analytic teams and the rest of the corporate culture is a key factor when organization tries to get benefits from data. Data products should be built in all parts of the company, and there should not be any of the silos that have traditionally separated data people from engineering, from design, and from marketing. (Patil 2011.) Data-oriented culture has also been highlighted when have been discussed achieving the competitive advantage comparing to competitors (Kiron et al 2012).

The study of Ransbotham et al. (2016), which included more than 2,000 interviews with managers at global companies, revealed that many managers who have difficulty obtaining a competitive advantage are still experimenting with what they can do with data and have yet to capitalize on data use across their organizations. The changeover from pilot project to organization-wide deployment can be difficult. It requires costs and effort, and raises broader questions about whether the analytics will be used consistently by decision makers. Companies should prepare for the robust investment and cultural changes that are required to achieve sustained success with analytics. This includes, for example, expanding the skill set of managers who use data and the commitment and hard work necessary to execute and maintain a successful analytics strategy. (Ransbotham et al. 2016.)

Van den Driest et al. (2016) assert that the new source of competitive advantage is customer centricity. The organization must have a deep understanding of customers' needs. To succeed, an organization must fulfill these needs better than anyone else. Data is required to accomplish this. As already mentioned, what increasingly separates the winners from the losers is the ability to transform data into insights about consumers' motivations and to turn those insights into strategy. Van den Driest et al. (2016) speak about an "insights engine," which requires innovative organizational capabilities.

4 Methodology and Research Process

The following section describes the methodological aspects of this study. The first chapter introduces research methodology before introducing research process and addressing the appropriateness of the methodology for this study.

4.1 Research Methodology

The aim of this study is to identify the competency requirements in data analytics and to clarify the influences of competency on performance in data analytics. This study is implemented using a mixed-method approach. The research questions of this thesis can best be answered by combining qualitative and quantitative methods (Teddlie and Tashakkori 2003).

The mixed-method approach has been regarded as the third methodological movement. Mixed-methods research is defined as a type of research that combines the elements of qualitative and quantitative approaches in any stages of the investigation (e.g., research questions development, data collection, analysis state) for the broad purposes of breadth and depth of understanding and corroboration. (Teddlie & Tashakkori 2003; Johnson & Onwuegbuzie 2004; Johnson et al. 2007; Venkatesh et al. 2013.)

The most common background of the mixed-method study is shown to be pragmatism. It focuses specifically on practice and research questions and not so much on methods. Methods are selected based on practical considerations about what works in order to obtain answers to questions. (Tashakkori & Teddlie 2009.) Also in this investigation, the selection of methods was made on the previous basis.

Venkatesh et al. (2013) offers a set of guidelines for conducting and evaluating mixed-methods research. They are introduced and discussed in Chapter 4.3. (Venkatesh et al. 2013.)

Methodological integration in the multi-method study can be implemented in many ways. When designing the study, one has to find its specific methodological integration. Methods can be integrated using different methods sequential (e.g., findings from one approach inform the other) or concurrently (i.e. independent of each other). When using concurrent design, quantitative and qualitative data are collected and analyzed in parallel. Results from analysis can be merged for a complete understanding of a phenomenon or to compare individual results. When using sequential design, quantitative and qualitative data collection and analyses are implemented in different phases, and each is integrated into a separate phase. (Teddlie and Tashakkori 1998; Venkatesh et al. 2013.) Venkatesh et al. (2013) suggest that scholars develop a design strategy in keeping with their research questions and objectives.

In this study, methods have been integrated sequentially and concurrently. At first, terms and required competency in data analytics are defined based on a literature review. Findings from the literature review are the basis for the empirical study. The literature review is also directly at the base of empirical research, in such a way that the list of competencies that are collected from the literature is used as a base for structured questions. The empirical study is done using questionnaires, in which both structured and unstructured questions are used. More detailed description of the utilization of the methods is provided in the next section when the competency identification process is explained.

In the chapters in which the competence- and analytics-oriented literature is described, background and analytical frames for the research is also provided. However, the theory and practice are iteratively combined in the literature section and is not limited to the empirical part. The practice in literature section is described in examples concerning the subject. Next, the research process, including a strategy for designing and analyzing will be presented.

4.2 Competency Identification Process

The research process is based on the research of Ley and Albert (2003), in which they identified competency of project manager using Competence Performance theory from Korossy (1999). It is introduced in chapter 2.3.4. The use of the theory is, however, modified so that it suitable for this study. Used procedures in the research process will be described next.

4.2.1 Examining the Purpose and Settings of Competency Modeling

Thesis planning began in the spring of 2016. At work and in the predictions made by Research Company Gartner could be seen much discussion about big data, intelligent algorithms, and data analytics in general (Sondergaard 2016). It was known that there are lots of possibilities, but it was not clear how to benefit. Determining the required competency related to data analytics seemed to be good starting point for the development.

4.2.2 Selection a Position and Group Employees

The next phase was to select a position and a group of employees. In the beginning, the target position was defined as data scientists in the selected company. However, the literature review revealed that it is rare that a single individual has all the needed competencies of a data scientist. In practice, the expertise and analytics tasks are often shared among a team of experts who have different kinds of skills. (Harris et al. 2013; Waller & Fawcett 2013; Maruyama et al. 2015.) Harris at al. (2013) found that data scientists can be clustered into subgroups, each with a different mix of skills. Also, other reasonable job titles, such as business analyst and data analyst, are used to describe the person who participates in designing and implementing analytics (Patil 2011). In addition, the job title data scientist became more common when the use of big data increased during the second wave of analytics and thereby is often combined with the analytics of big data (Davenport 2014). This study do not focus exclusively on big data analytics.

The desired participants for the study were those who actively participate in designing and implementing the use of data and analytics. As already mentioned, in this thesis participants are called data analysts. The participants were selected from three different companies.

4.2.3 Selecting the Performance Outcomes

Ley and Albert (2003 selected specific work outcomes on which performance could be assessed. In this study, we arranged the identification of competencies a little differently because the participating companies did not want to provide confidential information. Performance outcome was defined as work that creates customer value. It can be, for example, a change in financial conditions, behaviors, processes, and programs as well as any other benefits that result from the analysis (Davenport 2001).

4.2.4 Competency Elicitation using Literature Review

Instead of arranging two rounds of interviews, as Ley and Albert (2003) did in their study, in this study the competency elicitation performed by implementing a literature review and questionnaire study. A detailed description of the literature review and an empirical study using a questionnaire will be presented next.

Literature Review

The main aim of the literature review is to determine which competencies are currently found to be needed in the data analytics field. The purpose was to construct a list of competencies that could be required in data analytics so that it could be used as a basis for the next steps of research. The main aim divided into parts and began by finding out what kind of terminology is used.

The literature review followed instructions from Webster & Watson (2002), who provide suggestions for executing a literature review. When determining the source material for the review, they recommend carrying out a systematic search to ensure that a relatively complete set of relevant literature will be accumulated. They suggest that the major contributions are likely to be in the leading journals. Also, reviewing the citations of key articles can help to determine which ones should be included in the review. (Webster & Watson 2002.)

In this study, sources were primarily selected by following Webster and Watson's suggestions (Webster & Watson 2002). Source material was sought mainly by using the Google Scholar search engine, which also shows citations. Magazines were reviewed using the "Publication forum" (www.julkaisufoorumi.fi). Some papers were also found by exploring the references of the articles, and some were found incidentally when browsing the news of the day. When searching for relevant articles, keywords such as *competency*, *competence*, *capability*, *skill*, *big data*, *analytics*, *organizational*, *benefits* and *data* were combined.

When getting acquainted with the literature, it was revealed that even if there is quite a lot of common talk about the data analytics and the lack of data scientists, there has been little academic research implemented regarding the competencies needed in data analytics. Conclusive or clear definitions for the essential concepts (such as competence/competency or data analytics) were not found in the literature. However, it is essential to know the meaning of these terms and the theoretical background for their use when developing organizations or analyzing research results. Therefore, the review of the literature split into parts.

The structure of the review, according to Webster & Watson (2002), should be concept-centric prefer than author-centric. In this study, literature is also synthesized by discussing one concept at a time. This created some challenges, especially when compiling the needed competencies in data analytics. In that section, some general models of data analytics competency are presented separately before going into more detail. As Webster & Watson (2002, 18) point out "a review succeeds when it helps other scholars to make sense of the accumulated knowledge on a topic. We believe that sense-making is enhanced when a review is logically structured around the topic's central ideas and makes good use of tables and figures to convey economically the key findings and relationship."

As a result of the literature review, definitions of the terms used in this study were described. At first competencies and data analytics literature were handled separately. The aim of sections 2 and 3 is to clarify the terms and theoretical foundation used in this study. Chapter 5.1 describes the main results of the literature review.

4.2.5 Clarifying the Competency of Data Analyst based on Empirical Study

The empirical study was based on the results of the literature review. Data analytics competencies, listed in the previous phase were compiled from the questionnaire.

Designing the Survey

Designing the questionnaire was the most complicated process in this study. Without sufficient design and planning, the weaknesses are frequently not recognized until the results are interpreted, if at all. As Oppenheim (2000) presents in his article "Questionnaire design, interviewing and attitude measurement." most surveys go through a cycle of stages. In this study, the phases described next could be distinguished. (Oppenheim 2000.)

The most time-consuming step was when the general aims of the research lead to operationalized goals. Even after the subject and general aims and theories to be investigated were decided, modifying the questions and variables required many rounds of designing. Fortunately, my supervisor and the representative of the company who participated in this study gave good advice about designing the questionnaire. Finally, questions were finalized, and the rating scale was determined. Care was taken to ensure that questions could be answered quickly and without compromising confidentiality. (Oppenheim 2000.)

Interviews would have been an alternative to the questionnaire. Interviews could have acquired more information. Often, participants don't take time to give substantial answers to open questions. However, using a questionnaire seemed the better option for this study because interviews would have been difficult to arrange. Questionnaires also appeared to be the best way to assess the competencies that participants already have. The results could be used quite easily in the next phase of the research when designing competence-performance structures.

After a few planning cycles, the questionnaire was tested by one potential participant. Some revisions were made based on pilot testing (e.g., to the rating scale) and tried them out again. The Dreyfus et al. (1986) rating scale from novice to expert was selected. This scale is based on the evidence that skills are gained through experience, instruction, and imitation, and through experiential learning to gain the capacity to respond intuitively in complex situations. To the original five stages, we added a sixth option of ''don't know/not needed." Some of the participants were selected at the beginning of the planning and some were chosen just before sending questionnaires. The questionnaires and a cover letter were sent using Webropol or via an email link. There were different kinds of cover letters for different companies. The questionnaire is attached as Appendix 1.

Preparing the Data for Analysis

Survey was answered by 18 participant. The survey was prepared for the analysis using Excel software and a reporting tool called Webropol. Demographics were analyzed using Webropol reports. Structured questions, which helped to clarify the current state of competency, were prepared for analysis using Excel. Data was summarized using descriptive statistics and graphs. Based on descriptions, results were compared with the literature review findings.

Obtaining Competence-Performance Structures

Open questions were prepared for analysis by creating a separate competence-performance structure. Competence levels were derived from the answers to the questions. Performance levels were constructed based on the individuals and on principle (i.e., particular skills are required to perform particular tasks). Because particular duties were not specified, the competence-performance structure applied only to the job of data analyst in a general way. The competence-performance space was derived from the competence-performance structure. The competence-performance structure and space will be presented and explained in more detail in section 5.2.3.

4.3 Quality Aspects of the Research

In order to guarantee the quality of the mixed-method study, evaluation of quantitative and qualitative research separately, using their own criteria, is often required. In addition, it has been said that quality of the study needs to be evaluated based on criteria from mixed methods. (Teddlie and Tashakkori 2003; Creswell & Clark 2011.) Instead, Venkatesh et al. (2013) provide an integrated framework for validating inferences in mixed methods research. They are of the opinion that conducting high-quality quantitative and qualitative studies in mixed methods research does not necessarily guarantee high inference quality of mixed methods research. Instead, for providing the holistic insights on the phenomena, the focus should be on the quality of integrative inferences or meta-inferences. Next, the quality of this study has been evaluated based on the integrated framework. (Venkatesh et al. 2013.)

Venkatesh et al. (2013) emphasize that designing the research should be started by evaluating the appropriateness of a mixed-methods approach to the research. The purposes of the study will be described next. After assessing the appropriateness of mixed-methods design and quality aspects will be discussed.

4.3.1 Appropriateness of Mixed Methods Approach

As Venkatesh et al. (2013) note, before undertaking mixed-methods research, researchers need to carefully consider the appropriateness of employing such an approach in their study. Venkatesh et al. (2013) describe a comprehensive list of reasons for using the mixed-method approach in a research inquiry instead of a single-method approach. They have adapted these guidelines from other studies, such as Teddlie and Tashakkori (2003). The reasons are complementarity, completeness, developmental, expansion, corroboration or confirmation, compensation, and diversity. The purposes concerning this study are described in Table 3. In this study, mixed methods are used to gain complementary views of data analyst competence. Using structured questions in the questionnaire also provide richer explanations of the findings from the qualitative data. The method is used sequentially and also offers the developmental purpose for the use of mixed methods. (Venkatesh et al. 2013)

TABLE 3 Purposes for using Mixed-Method Approach (Adapted from Venkatesh et al. 2013)

	Description	Purpose of this study
Complementarity	Using mixed methods can achieve complementary views about the same phenomena.	Mixed methods are used in order to gain complementary views of data analyst's competence. The current state of competency provides additional information when evaluating the required competencies. In addition, the competency requirements presented in the literature complete the picture.
Completeness	Using mixed-method designs complete the picture of a phenomenon.	The quantitative data from the structured questions of the questionnaire complete the picture of required competencies and provide richer explanations of needed competency, which is resolved also based on qualitative data from the literature review and from open questions of the questionnaire. Methods are also combined to give a description that is more generalized.
Developmental	Questions for one section emerge from the deduction of a previous one, or one strand provides hypotheses to be tested in next section.	Methods have been used sequentially and parallel. Questionnaire (especially the structured part) is based on the findings of the literature review.
Expansion	Purpose can be to find out or extend the understanding obtained in a previous strand of study.	Expansion using different methods is not the main purpose of this study. The purpose of mixing methods is more to complete the picture than expanding it.
Corroboration/ confirmation	Purpose can be to assess the credibility of inferences obtained from one approach.	Confirmation or collaboration is not the main purpose of the use of the mixed method in this study.
Compensation	Purpose can be to enable compensating for the weaknesses of one approach by using the other.	In this study, the purpose is not that methods compensate one another. On the other hand, there would not be enough material for only quantitative analysis.
Diversity	The purpose is to obtain divergent views of the same phenomenon.	Diversity is not the purpose of using the mixed method in this study.

4.3.2 Design Quality

Design quality can be improved by selecting methods and designing the research so that appropriate answers can be obtained to the research questions (Venkatesh et al. 2013). The **adequacy of the design** degree can be guaranteed by implementing the design components for the qualitative and quantitative parts with acceptable quality and rigor. Concerning qualitative research, *credibility* and *dependability* are discussed. Quantitative research focuses on *reliability* and *internal validity.* (Teddlie and Tashakkori 2009.)

The results of qualitative research are credible if alternative explanations are convincingly ruled out. In this study, *credibility* in the literature review is guaranteed by selecting not only parts of the studies or articles, but by considering them as a whole and by searching also other articles by the same authors. Credibility of survey questions could have been assessed in advance, but when the questions were left without clarification, more comprehensive answers were obtained. As a whole, it could be even better credibility when the problems were not limited to individual competencies. *Dependability* has been ensured by following the selected and previously validated research process. (Venkatesh et al. 2013) The process for identifying competency was described in chapter 4.2.

Design adequacy of quantitative part of the research can be assessed by considering reliability and internal validity issues, such as Venkatesh et al. (2013) advises. *Reliability* means repeatability or consistency of the research. In this study, the reliability of structured questions was ensured by setting clear questions and response options. More specifically, it was not evaluated or tested, but presumably, the same results would come from structured questions if the survey was repeated. However, the sample size was not big enough to create a generally applicable picture. In considering the *internal validity* of the quantitative part of the study, there were considerations of the equivalence of competencies compared to the literature review. This was not, however, related not only to quantitative research. (Venkatesh et al. 2013.) The internal validity may be affected by many factors, such as time and the rating scales. In this study, the duration of the empirical study was relatively short, and the use of rating scales was discussed carefully. (Shadish et al. 2002.)

Analytic adequacy of the research can be defined as the degree to which qualitative data analysis procedures are appropriate and adequate to provide plausible answers to the research questions. Qualitative research can be considered by evaluating theoretical validity and plausibility. (Venkatesh et al. 2013.) *Theoretical validity* describes the extent to which the theoretical explanation developed fits the data and, therefore, is credible and defensible (Maxwell 1992). Venkatesh (2013) has been described *plausibility* according to Sandelowski (1986) so, that it is achieved when the findings of the study are suitable for the material from which they have been derived. This study is implemented based on the process introduced by Ley and Albert (2003), and practices are thus credible and defensible. However, the theoretical validity and plausibility could not be fully ensured because the open replies had to be coded into categories. However, predetermined, and at least once tested (in the previous studies), categories helped to ensure theoretical validity and plausibility. The quantitative part was analyzed using only averages and medians to describe the current knowledge. Analysis methods regarding quantitative data were very simple. The material was also too small to draw any accurate conclusions. Hence it can be said that *statistical conclusion validity* has not been considered in this study. (Venkatesh et al. 2013.)

4.3.3 Explanation Quality

According to Venkatesh et al. (2013), another aspect of the integrative framework of inference quality is explanation validity. **The quality of qualitative inferences** can be analyzed by evaluating credibility, confirmability, and transferability. When analyzing the open questions of the survey, as said earlier, *credibility* was not easy to ensure. Some of the answers to the questions were not clearly classifiable, and when the competencies derived from the answers, misconceptions cannot be completely ruled out. *Confirmability* describes the level of the results that could be confirmed or corroborated by others. In this study, the confirmatory level has been strengthened by using references to the literature in addition to the empirical material when drawing conclusions. The fact that the material has been collected from more than one company and analysis has also been carried out based on literature, increases the level of *transferability.* (Venkatesh et al. 2013.) However, concerning the transferability, the researcher cannot specify the transferability of findings. He or she can only provide sufficient information that can then be used by the reader to determine whether the findings apply to the situation. (Lincoln and Guba 1985.)

The quality of quantitative interferences is determined by how closely interpretations from the quantitative analysis follow the relevant findings and the state of knowledge in the field are consistent with theory and are generalizable. According to Venkatesh et al. (2013) indicators of quality include *internal validity*, *statistical conclusion validity,* and *external validity*. Internal and statistical conclusion validity in this research were described previously in section 5.2. External quality has been described as the extent to which the results can be generalized to other settings and groups. (Venkatesh et al. 2013.) As mentioned earlier, in this study the findings of quantitative part of the study cannot be generalized because of the number of respondents. In this study, the results derived from the structured questions have been used to give a broader view of required competency in data analytics. Even if the findings from the quantitative part are not generalizable, they are mostly consistent with the theory and the state of knowledge in the field.

The quality of integrative inference is determined by assessing *inference transferability, integrative efficacy,* and *integrative correspondence*. Inference transferability describes the degree to which meta-inferences from mixed methods research can be generalized or transferred to other contexts or settings. (Venkatesh et al. 2013.) In this study, using mixed methods considerably increases the generalizability or transferability of the findings. As said earlier, when discussing of the appropriateness of mixed methods in this study, different methods were used in order to gain a complementary view of competency in data analytics. Useful information for evaluating the required competencies can be obtained by collecting information about the current state of competency.

Likewise, needed competence presented in the literature complete the picture of required competency in data analytics.

Integrative efficacy refers to the degree to which interferences of each strand of research inquiry are integrated into a theoretically consistent meta-inference (Venkatesh et al. 2013). In this study, the inquiries cannot be separated. The process of the identification is defined based on the study of Ley and Albert (2003) so all the blocks of this study are necessary for the outcome.

The last criterion of the quality of integrative inference is i*ntegrative correspondence.* According to Venkatesh et al. (2013), integrative correspondence is important to ensure that researchers employ a mixed methods research approach in keeping with an overarching research objective. Quantitative and qualitative studies should be conducted to achieve common research objectives so that these justify a high degree of integrative correspondence. Venkatesh et al. (2013) are of the opinion that if these inference quality criteria are met, researchers can feel confident about the overall inference quality of mixed methods research and be able to report findings from mixed methods research. Results of the study are described next.

5 Results

This chapter provides an overview of results from the literature review of this study and results from the empirical part of Pinola's (2015) master's thesis. In addition, results from the empirical study of this thesis are described. In chapter 6, the results from the empirical study of this thesis will be analyzed based on the literature.

5.1 Results from Literature Review

The aim of the literature review is to create a basis for empirical research by introducing an overview of past research concerning competencies, commonly used terms in data analytics, and competencies needed in data analytics. In this study, the competency identification process for assessing competency requirements in an organization is also introduced.

The definition of the concepts related to data analytics is introduced in chapter 3.1. *Data* and *data analytics* are quite old terms, but their use in organizations has changed over time as understanding and technologies have developed. *Big data*, *data science*, and *data-driven organizations* are more recently coined terms. In all its simplicity, when data (small data or big data) is to be converted into knowledge or wisdom so that it is most valuable to organizations, different types of data analytics techniques (which can also be called *data science* when big data is used) are needed. When organizations engage in working with high quality by using data and analytics in day-to-day operations, it can be called a data-driven organization. (Kiron et al. 2012.)

The primary purpose of the literature review is to construct a list of competencies that could be needed in data analytics so that it could be used as a basis for the next steps in research. In practice, the results of the literature review were used to design the questionnaire. The list of necessary competencies was constructed based on the results of the empirical part of Pinola's (2015) master's thesis and other literature regarding the competency of data analysts.

In practice, the master's thesis of Pinola (2015), in which he focused on job advertisements that mentioned data-scientist competencies and that was taken as the basis of this study. He investigated data scientists' competency through 94 job advertisements. However, the results are presented here because taking them into account, the picture of required competencies is complemented. In addition, some of the literature sources found by Pinola in his review of the literature were reused. Some articles, which Pinola used in his study, were excluded from this study because the articles were not available anymore or because they were not scientific articles according to Julkaisufoorumi. Reused parts are set in *italics* in the tables.

The categorization of competencies is based on Havelka and Merhout's (2009) competence framework. The framework classifies the skills required by IT professionals into four categories: personal traits, professional skills, business knowledge, and technical knowledge. It is quite an abstract framework and describes categories, but not any specific competencies within them. In this thesis, the categories were modified so that business knowledge was merged into the professional skills category. In addition, a category for organizational competencies was added. The same categorization is used in all chapters.

Next, the results of the literature review and results from Pinola's (2015) empirical study are introduced. As mentioned, the design of the questionnaire was based on the literature review. For the questionnaire, competencies that emerged from the literature review and from Pinola's (2015) master's thesis were selected. The most frequently mentioned competencies chosen for the questionnaire, but some less frequently mentioned were also included. Because of the length of the form, some competencies had to be left out and others were combined. For example, communication and reporting were combined into reporting. The final questionnaire can be seen in Appendix 1. All the parts, including Pinola's empirical part of the master's thesis, the literature review for this study, and the empirical material of this study, are presented in chapter 6.

5.1.1 Personal Traits

The personal traits that are most commonly mentioned in the literature are shown in table 4. Personal traits include all the characteristics of the person that persist in different situations. These traits are more difficult to develop than other kinds of competencies.

Based on the literature, data analysts need, above all, to be passionate problem-solvers. This trait was mentioned in three sources. (Patil 2011; Davenport & Patil 2012; Dhar 2013.) Data analysts need to be creative so they can look at problems in innovative ways; this trait was also mentioned in three sources (Davenport & Patil 2012; Van den Driest 2016; Patil 2011). Social skills were also mentioned in two articles (Davenport & Patil 2012; Davenport 2012). Davenport & Patil (2012) also mention that data scientists need to be pressure tolerant, aware of real time, curious, proactive, determined, productive, eager to learn new things, and determined.

TABLE 4 Most Frequently Mentioned Personal Traits Based on the Literature [1]

Personal Traits		
[2]	**Competence**	**Sources**
3	*Passionate problem-solver,* Problem-solving skills	(*Davenport & Patil 2012*; Dhar 2013; Patil 2011)
3	Creative, Innovative	(Davenport & Patil 2012; Van den Driest 2016; Patil 2011)
2	Good team player, *Strong social skills*	*(Davenport & Patil 2012;* Davenport 2012)
1	Pressure tolerant	(Davenport & Patil 2012)
1	Real-time awareness	(Davenport & Patil 2012)
1	Curious	(Davenport & Patil 2012)
1	Proactive	(Davenport & Patil 2012)
1	Productive	(Davenport & Patil 2012)
1	Eager to learn new things	(Davenport & Patil 2012)
1	*Determined*	*(Davenport & Patil 2012)*

The results of Pinola's (2015) empirical study are described in table 5. The number in parentheses is the number of times a competence was mentioned in job advertisements. The most frequently mentioned personal traits in job advertisements were creativity, pressure tolerance, and flexibility. In some cases, the terms *passionate, curious, proactive*, and *productive data scientist* were mentioned. Ability to learn and ability to work independently seem also to be essential characteristics.

TABLE 5 Personal Traits Occurred in a Job Advertisements of Data Scientist (Pinola 2015, 40)[3]

	Job Advertisement (Pinola 2015)
Personal Traits	Creative (16)
	Pressure tolerant (10)
	Flexibility (10)
	Able to operate in a changing environment (5)
	Passionate (5)
	Curious (4)
	Proactive (4)
	Productive (4)
	Ability to learn (4)
	Able to work independently (4)
	Willingness to travel (2)

Based on the literature research, personal characteristics received significantly less attention compared to professional or technical skills. This was also true in job advertisements (Pinola 2015). When comparing competencies found in the

[1] Results taken from Pinola (2015) are set in *italics.*
[2] The number in the column indicates how many times competence was mentioned in the literature.
[3] Number in the parentheses indicates the number of times the competence was mentioned in job advertisements.

literature to the competencies mentioned in job advertisements, it can be seen that some of the characteristics are the same. Creativity and pressure tolerance were quite high in both. In the job advertisements were also sought data scientists who had the willingness to travel, the ability to learn and were flexible (Pinola 2015).

To the questionnaire were selected following competencies: able to work independently, determined, passionate problem solver, eager to learn new things, curious, creative, pressure tolerant, able to work in a changing environment, proactive, productive, good team player, willing to take responsibility.

5.1.2 Professional Skills

Professional skills category includes all the competencies, which can be learned. These are the skills and knowledge related to working. The competencies in this category do not require technical skills but are more general work-related skills.

The most important professional skills based on previous studies are subject knowledge or business knowledge from the data perspective (Conway 2011; Chen ym. 2012; Davenport & Patil 2012; Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013; van den Driest 2016). These skills were mentioned in seven articles. As Dhar (2013) notes, the key differentiator to being an effective data scientist is the ability to formulate problems in a way that results in effective solutions. That cannot be accomplished without strong subject knowledge.

Also, analytical skills were highlighted as necessary competence for data analyst (Davenport & Patil 2012; Davenport & Dyche 2013; Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013). In the context of analytical skills, computational thinking was also discussed (Dhar 2013). Analytical skills were mentioned in five sources and computational thinking in two sources. These two were combined into a single competence because computational thinking was discussed in the context of analytical skills.

Management or decision-making understanding was mentioned in three sources. According to the literature, data analysts have to help decision makers, which requires management understanding. Reporting skills were also mentioned to be a necessary competence for data analysts in three sources. In this context, storytelling seems to be a useful way of communicating and reporting the findings. (Chen ym. 2012; Davenport & Patil 2012; Davenport & Dyche 2013; Driest et al 2016.) Communication skills were also mentioned in two sources. It could be combined with reporting skills, in which some sources it seemed to be included but was mentioned separately in others.

TABLE 6 Most Frequently Mentioned Professional Skills Based on the Literature [4]

Professional skills		
[5]	**Competence**	**Sources**
7	*Subject knowledge/business understanding*	(*Davenport & Patil 2012; Chen ym.2012; Conway 2011*; Dhar 2013; van den Driest 2016; Provost & Fawcett 2013; Waller & Fawcett 2013)
7	Analytical skills, Computational thinking	(Wing 2006; Barr and Stephenson 2011; Davenport & Patil 2012; Davenport & Dyche 2013; Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013)
3	Management understanding, decision management	(Waller & Fawcett 2013; Provost & Fawcett 2013; Miller 2014)
3	Reporting skills, storytelling	(Davenport & Patil 2012; Davenport & Dyche 2013; Driest et al 2016)
2	Communication skills	(Davenport & Patil 2012; Chen ym.2012)
1	Strategic data management	(Miller 2014)
1	Agile methods	(Patil 2011)
1	Change management	(Davenport & Dyche 2013)

Table 7 shows the results of an empirical study concerning professional skills of data scientist (Pinola's 2015). The number in parentheses is the number of times the competence was mentioned job advertisements. Pinola's (2015) empirical study reports that the most frequently mentioned competencies in the job advertisements of data scientists were analytical skills, communication skills, subject or business knowledge, and team working skills. Documentation skills and management skills were also mentioned quite often in the job advertisements as were entrepreneurial spirit and process know-how. (Pinola 2015.)

TABLE 7 Professional Expertise Occurred in Job Advertisements for Data Scientists (Pinola 2015, 40)[6]

	Job Advertisement (Pinola 2015)
Professional skills	Analytical skills (53)
	Communication skills (49)
	Subject knowledge, business knowledge (39)
	Team working skills (25)
	Reporting skills (2), documentation skills (22)
	Management skills (9)
	Entrepreneurial spirit (4)
	Process know-how (4)
	Strategical planning (2)
	Agile methods (2)

[4] Results taken from Pinola (2015) are set in *italics.*

[5] The number in the column indicates how many times competence was mentioned in the literature.

[6] Number in the parentheses indicates the number of times the competence was mentioned in job advertisements.

The term *change management* did not appear in job advertisements even though their importance was mentioned in the literature (Davenport & Dyche 2013). Instead, *entrepreneurial spirit* and *process know-how* were mentioned in the job advertisements collected by Pinola (2015) but were not found from the literature.

The questionnaire focused on management skills, reporting skills, math and statistics knowledge, company's strategic planning, problem-solving skills, business planning related to data analytics, project management skills, agile methods, and data architecture design. When selecting competencies, it was thought that analytical skills are associated with problem-solving skills, which are why they were combined. Also, communication skills were considered to be associated with reporting skills. Team working skills were excluded from this section because the competence of being "good team player" was included in personal traits. Different kinds of business knowledge was hard to distinguish. In the questionnaire, they were classified into business planning related to data analytics, management skills, and the company's strategic planning. In retrospect, subject knowledge could have been added separately, as it was thought to be contained in business planning related to data analytics.

5.1.3 Technical Knowledge

The technical knowledge category includes all technical-related competencies that are not directly related to a specific tool or technology. When searching for information about the needed technical skills of data analyst, the most frequently mentioned technical skills were programming and machine learning skills. As Patil (2011) notes, programming seems to be a universal skill for data scientists. It was mentioned in four articles as well as machine learning (Patil 2011; Davenport & Patil 2012; Davenport & Dyche 2013; Miller 2014). Statistics and using unstructured data were mentioned almost as often (Conway 2011; Davenport & Patil 2012; Davenport 2013; Dhar 2013; Miller 2014). Some skills were merged into broader categories, and some, such as statistical techniques, were not described separately. The category of algorithms was included with math skills. However, in the questionnaire, some tools and technologies were described in more detail.

In the literature, frequent mention was made of integration of traditional and big data sources of information in companies that have traditional information infrastructures (Davenport & Dyché 2013; Davenport 2013). Security, privacy, and ethical use of data was mentioned, particularly in the sense that lack of skills and attention to these issues concerns at the moment (Miller 2014; Porter et al. 2015). Also in the literature, predictive analytics, simulation, network analysis and real-time analysis was mentioned twice.

TABLE 8 Most Frequently Mentioned Technical Skills Based on the Literature

Technical skills		
Times mentioned	**Competence**	**Sources**
4	*Programming skills*	(*Patil 2011; Davenport & Patil 2012;* Davenport&Dyche 2013; Miller 2014)
4	Machine Learning	(*Conway 2010*; Dhar 2013; Harris, Murphy & Vaisman 2013; Miller 2014)
3	*Statistics*	(*Conway 2011*, Dhar 2013; Miller 2014)
3	Using unstructured data	(Banerjee et al 2013; Davenport 2013; Davenport & Dyche 2013)
2	Math, Algorithms	(Dhar 2013; Miller 2014)
2	Integration of traditional and big data	(Davenport & Dyché 2013; Davenport 2013)
2	Security, Privacy and ethical use of data	(Miller 2014; Porter et al. 2015)
2	Predictive analytics	(Dhar 2013; Miller 2014)
2	Simulation	(Barr and Stephenson 2011; Gartner 2016, IT Glossary)
2	Network analysis	(Dhar 2013; Davenport 2013)
2	Real-time analytics	(Davenport 2013; Davenport & Dyche 2013)

Table 9 shows the results of Pinola's (2015) empirical study concerning the technical skills of data scientists. As mentioned, he investigated data scientists' competency by evaluating 94 job advertisements. The number in parentheses is the number of times the competence was mentioned job advertisements. The most often in the job advertisements mentioned skills were statistics, machine learning, and programming skills. Also, data mining, algorithms, and predictive analytics were mentioned in many job advertisements.

TABLE 9 Technical Skills Occurred in a Job Advertisements of Data Scientist (Pinola 2015, 41)

	Job Advertisement (Pinola 2015)
Technical knowledge	Statistics (59)
	Machine learning (50)
	Programming skills (38)
	Data mining (33)
	Algorithms (28)
	Predictive analytics (27)
	Visualization (13)
	ETL (11)
	Mathematics (8)
	Econometrics (7)
	Network analysis (4)
	Big data -systems (4)
	Simulation (4)
	Building recommender systems (3)
	Cloud computing (2)
	Data warehouses (2)

When selecting competencies for the questionnaire, data mining and ETL were excluded even though they were included in Pinola's (2015) study. When testing the questionnaire respondents said that these concepts were too broad and unclear. Algorithms were included in the mathematics category. In retrospect, it could have been added separately. On the other hand, for the end result, it is probably not a relevant issue. Security, privacy, and ethical use of data did not appear in job advertisements even though they were mentioned in the literature (Davenport & Dyche 2013; Miller 2014; Porter et al. 2015). Because of the importance of security issues, it was added to the questionnaire.

5.1.4 Tools and Technologies

Tools and technologies were separated from other technical skills because of clarity. This category includes specific tools and technologies used by the data analyst.

In the literature, Hadoop, which is a technology for storing large and diverse amounts of data, was most widely spoken of technology (Davenport & Dyche 2013; McAfee et al. 2012; Conwey 2011; Davenport 2014). It was mentioned in four articles. Other frequently mentioned tools and technologies were scripting languages like Python, Pig, and Hive. Tool such as R, SQL, NoSQL, Excel, Impala, SAS, IBM DB2, Vertica, Tableau, and Watson were also cited (Kiron et al. 2012; Davenport & Dyche 2013). These are listed in table 10.

TABLE 10 Tools and Technologies Mentioned in the Literature[7]

Tools and Technologies	
Competence	**Sources**
Hadoop	(*Conwey 2011; McAfee et al. 2012;* Davenport 2013; Davenport & Dyche 2013)
R	(Davenport & Dyche 2013)
Python	(Davenport & Dyche 2013)
SQL	(Davenport & Dyche 2013)
NoSQL	(Davenport & Dyche 2013)
Watson	(Kiron et al. 2012)
Pig	(Davenport & Dyche 2013)
Hive	(Davenport & Dyche 2013)
Excel	(Kiron et al. 2012)
Impala	(Davenport & Dyche 2013)
SAS	(Davenport & Dyche 2013)
IBM DB2	(Davenport & Dyche 2013)
Vertica	(Davenport & Dyche 2013)
Tableau	(Davenport & Dyche 2013)

Regarding expertise in tools and technologies, Pinola (2015) found that R, Python, SQL, Java, Hadoop were mentioned most frequently in job advertisements. SAS,

[7] Results taken from Pinola (2015) are set in *italics*.

Hive, C++, Pig, and Matlab were mentioned quite often. Table 11 shows the most frequently mentioned competencies in job advertisements. The number of times the competence was mentioned in job advertisements is given in parentheses.

TABLE 11 Tools and Technologies Mentioned in Job Advertisements for Data Scientists[8] (Pinola 2015, 41)

	Job Advertisement (Pinola 2015)
Tools and technologies	R (46)
	Python (46)
	SQL (37)
	Java (37)
	Hadoop (35)
	SAS (24)
	Hive (22)
	C++ (21)
	Pig (19)
	Matlab (18)

A few tools and technologies were added to the questionnaire: R, Python, SQL, Hadoop, Java, Hive, Power BI, SAS, SPSS, Tableau, Matlab, Pig, Spark and SQL Server. SPSS, Spark and SQL Server did not emerge from the literature or job advertisements, but they were added because in test investigations these were reported to be generally used.

5.1.5 Organizational Competency

In addition to individual competencies, organizational competencies were mentioned in the literature. Although initially the study focused on individual competencies, as the investigation progressed the importance of organizational competencies became apparent. This category includes the needs of the organizational environment.

The most commonly in the literature were identified the benefits of data-oriented culture (Patil 2011; McAfee et al. 2012; Kiron et al. 2012; Patil & Mason 2015; Ransbotham et al. 2016). According to the literature, it can also be concluded that it is extremely important that analytics be connected to a company's business (Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013). As stated already in the definition of big data, value emphasizes the importance of the benefits for business (Vesset et al. 2012). In addition, analytics strategy was mentioned in two sources (Miller 2014; Ransbotham et al. 2016). Other organizational competencies mentioned in the literature are described in table 12.

[8] The number in parentheses is the number of times a competence was mentioned in job advertisements.

TABLE 12 Organizational Competencies mentioned in the Literature

Organizational Competency		
9	**Competence**	**Sources**
5	Data-oriented culture	(Patil 2011; McAfee et al. 2012; Kiron et al. 2012; Patil & Mason 2015; Ransbotham et al. 2016)
4	Analytics connected to company's business	(Vesset et al. 2012; Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013)
2	Strategy that considers a data as a core business asset (Miller 2014), Commitment and hard work required to execute and sustain a successful analytics strategy (Ransbotham et al. 2016)	(Miller 2014; Ransbotham et al. 2016)
1	Close relationships between IT and business	(Davenport & Dyche 2013)
1	Capitalizing on data use across the organization	(Ransbotham et al. 2016)
1	Skillset for managers who use data	(Ransbotham et al. 2016)
1	Customer centricity	(Van den Driest et al. 2016)

5.2 Results from Empirical Study

This chapter describes the results of the empirical study. Demographics is discussed first, followed by the current competency of participants and the most beneficial competencies.

5.2.1 Demographics

Invitations for the survey were sent to three different companies. The participants were selected based on their job tasks. The criteria were that respondents must participate in designing and/or implementing the use of data and analytics in their work. One of the companies has the separate analytics team, but in two other companies, the data analysts were not assembled into one team which made the search for target persons more difficult. Finally, 18 respondents from three different companies participated in the survey. Table 13 describes demographics of the participants.

Most of the participants were male. Only two were females. Age was mostly (89%) between 26 and 45, but two of the participants were between 46 and 55. The level of education was remarkably high. Most of the participants (61%) had master’s degrees, and 28% of the respondents had doctorate degrees.

[9] The number in the column indicates how many times competence has been mentioned in the literature.

TABLE 13 Demographics of the Respondents

Gender	Female	2	11%
	Male	16	89%
Age	26 - 35	8	44.5%
	36 - 45	8	44.5%
	46 - 55	2	11%
Highest degree	Bachelor's degree	2	11%
	Master's degree	11	61%
	Doctorate degree	5	28%

The field of education	Information and service management / economics and business administration
	Technology management
	Information technology
	Media engineering
	Information and communication technology
	Industrial engineering and management
	Control and automation engineering
	Telecommunications technology
	Mathematics and engineering
	Applied mathematics
	Technical physics and mathematics / computational engineering
	Physics
	Physics
	Chemistry, economics, materials science
	Astrophysics
	AI
	Machine learning and data mining
	Tax law
Title	Analytics designer
	Senior analytics designer
	Senior business analyst
	Product owner
	Product owner / senior developer
	Manager
	BI lead architect
	BI consultant
	Integration specialist / BI specialist
	QA specialist
	Big data developer
	Senior developer, advanced analytics
	Developer
	Developer
	Programmer
	Consultant
	Consultant
	n/a

Even if none of the participant's title was a data scientist, the field of their education corresponds well with the descriptions found in the literature of data scientist. Data scientists were often formally trained in computer science, mathematics, or economics, statistics, machine learning, databases, operations research, business intelligence, social or physical science, but they can emerge from any field that has a strong data and computational focus (Patil 2011; Davenport & Patil 2012; Harris et al. 2013).

5.2.2 Current Competency

Current competency of participants was surveyed by collecting the data from structured questions of the questionnaire. The questions were separated into four themes, which were adapted from the framework of Havelka & Merhout (2009). This chapter explains the findings in each category.

Personal Traits

Based on the empirical research, data analysts are usually, above all, passionate problem-solvers and eager to learn new things. They are also able to work independently and are curious and determined. The median of these competences was 5 on a rating scale of 1 to 5, meaning that the competence describes them very well. The standard deviation was only between 0.38 and 0.62, except determined, which deviation was higher (0.91). All 18 participants answered all questions concerning personal traits. The results are shown in figure 9.

Based on the study, it can be argued that data analysts are also creative and able to work in a changing environment. They are good team players who are usually willing to take responsibility. Productive, pressure tolerant, and proactive were not mentioned, although these traits also described the data analysts who participated in the study.

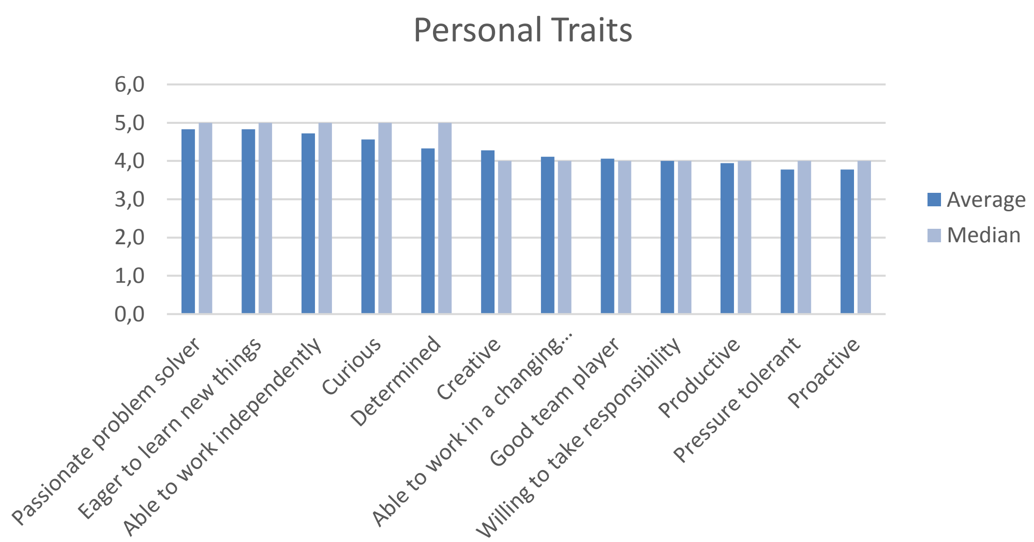

FIGURE 9 Averages and Medians of the Answers when asked to evaluate Personal Traits

Professional Skills

The level of professional expertise was determined by asking participants to evaluate their expertise in certain competencies on a scale of 1 to 5, where 1 = Novice, 2 = Beginner, 3 = Intermediate, 4 = Advanced and 5 = Expert. An addition option of "Don't know/Not needed" was also included. In this section, the option "Don't know/Not needed" was chosen by two of the participants when they evaluated their expertise in business planning related to data analytics and company's strategic planning. For other competencies, evaluation was obtained from all the 18 respondents.

Respondents rated the highest expertise in problem-solving skills which was already mentioned in the previous section. Respondents evaluated their skills as intermediate or higher also regarding reporting skills, mathematics and statistics knowledge, business planning related to data analytics, data architecture design, project management skills, agile methods and management skills. The knowledge of their companies' strategic planning was a bit below intermediate, and the answers varied quite a bit (standard deviation 1,18). Also, the expertise in data architecture design, business planning related to data analytics and management skills varied quite much (standard deviations between 1,09 and 1,2). In the other skills, variability among the responses was less. The results are shown in figure 10.

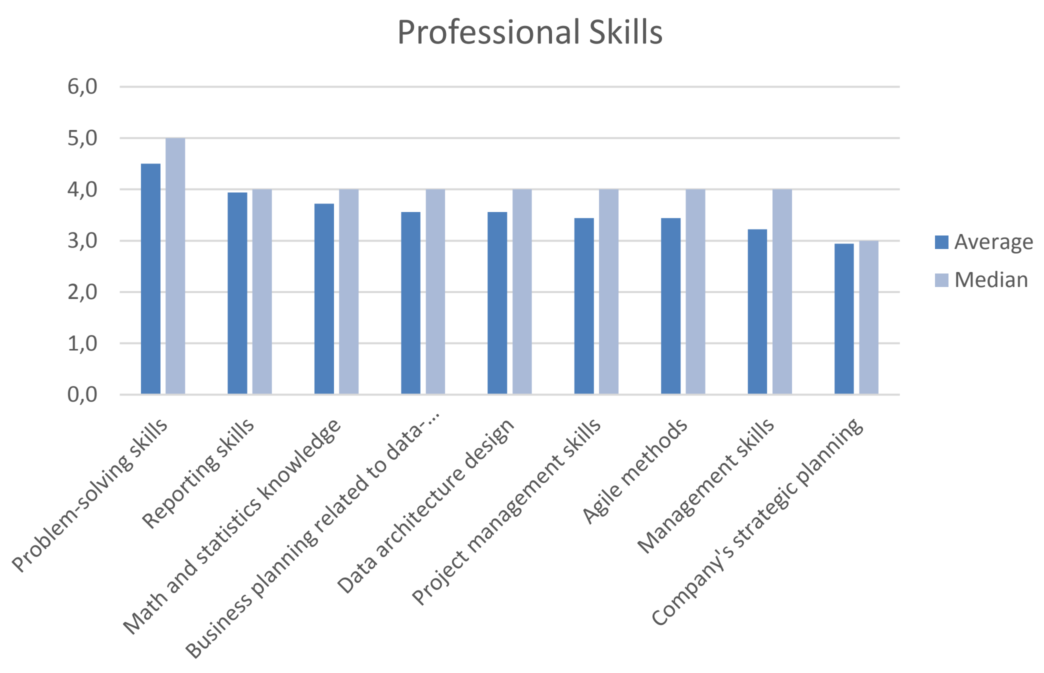

FIGURE 10 Averages and Medians of the Answers when asked to evaluate Professional Skills

Technical Skills

Technical skills varied significantly more than personal traits or professional skills. Depending on the skill, the number of responses varied from 13 to 18. The evaluation was obtained from all the 18 respondents for preparing data for analytics, databases, data warehouses, and processing of IoT (Internet of Things) data. Figure 11 can be shows missing responses, which reflect the "Don't know/Not needed" answers.

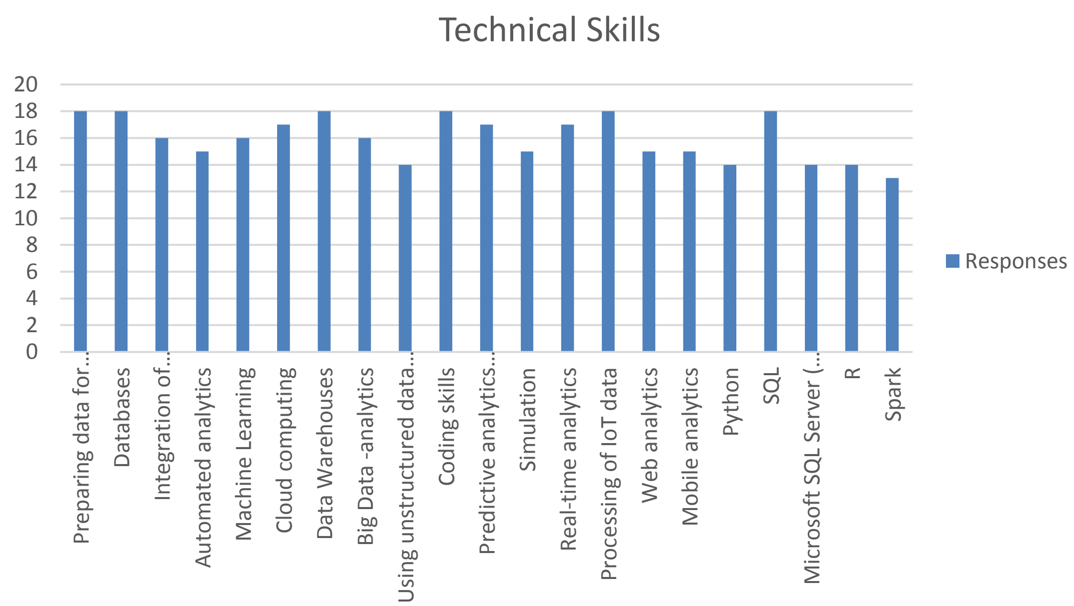

FIGURE 11 Number of Responses for Technical Skills

Technical competencies in which participants had the greatest expertise were preparing data for analytics (see figure 12). Almost all respondents' expertise in databases and the integration of traditional (structured, centralized data) and big data (unstructured, distributed data) analytics were at least at the intermediate level. Competency in automated analytics, machine learning, cloud computing, data warehouses, big data analytics, using unstructured data in analytics, coding skills, predictive analytics, and simulation was intermediate on average. Remarkably, coding skills, which have been cited as one of the data analysts' basic competencies, were not significantly high. Out of 18, 4 respondents (22%) reported novice-level coding skills. A second skill, to which attention was drawn, is machine learning, and 6 of 18 respondents (33%) rated themselves as novices or advanced beginners.

In automated analytics, machine learning, cloud computing, and Big Data–analytics, standard deviation of the responses was 1.20 – 1.28, when, for example, standard deviation of responses in the competency of preparing data for analytics was 1.0. This indicates that there were quite a lot of differences in the expertise of these competencies.

Expertise in real-time analytics, processing IoT data, web analytics, and building recommender systems was mostly at the intermediate level. In mobile analytics, security and privacy competency level of the average respondent was advanced beginner.

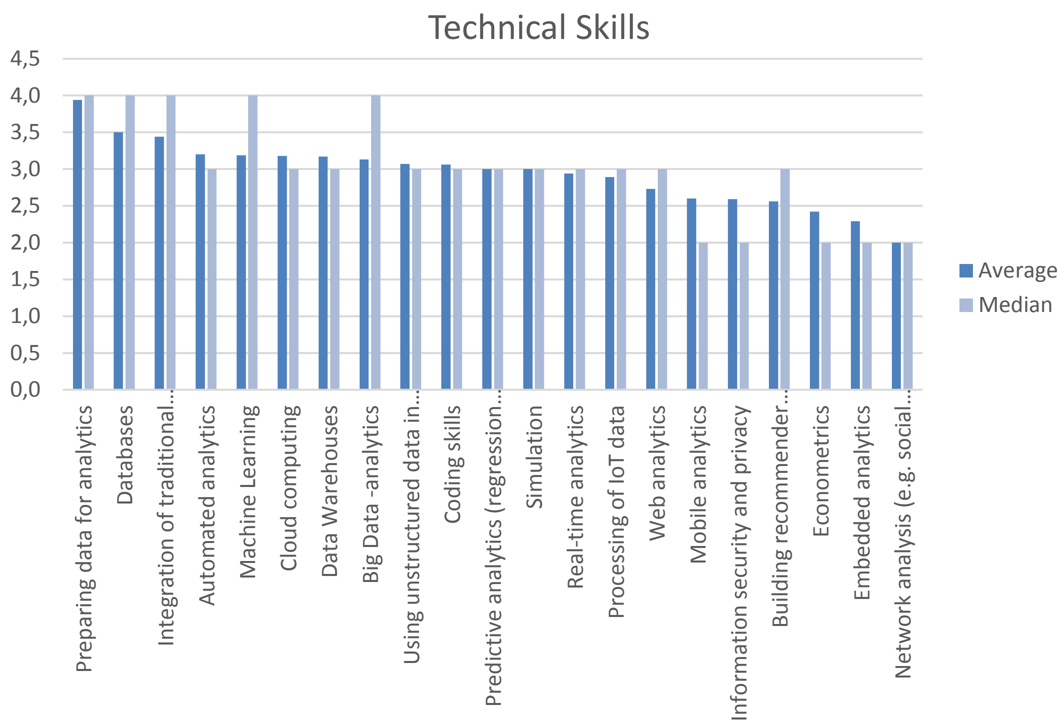

FIGURE 12 Participant's level of Expertise in Technical Skills

Tools and Technologies

The expertise of tools and technologies varied significantly more than competency in other categories. The amount of "Don't know / Not needed" –answers were also larger than in others (figure 13). This in itself was presumably, because a number of alternative tools and technologies is huge nowadays.

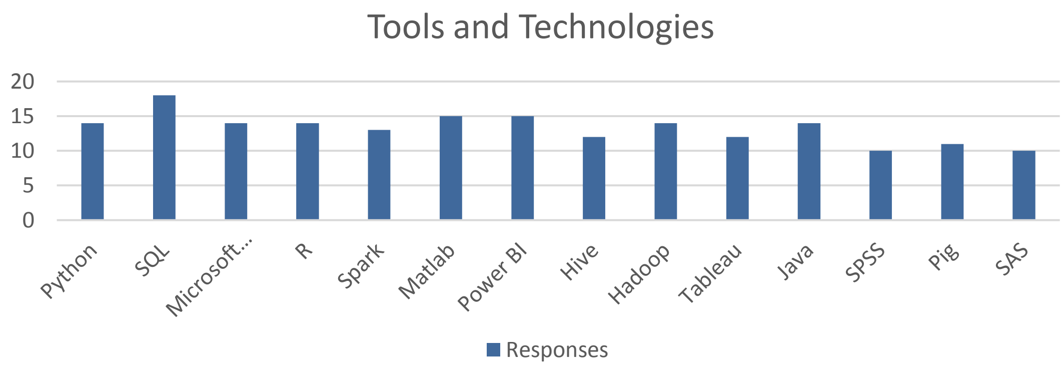

FIGURE 13 Number of responses for Technical Skills

As shown in figure 14, Python and SQL seem to be most commonly used tools or technologies in the data analytics area. Half of all participants described their expertise in Python and SQL as being at an advanced level. According to results from the empirical study, Microsoft SQL Server seems to be the basic tool used for data analysts, and most of the data analysts can use it to some extent. Use of other tools, such as R, Spark, Matlab, Power BI, Hive, Tableau and Hadoop, varied quite a bit. Some of the participants can use tools fluently, but others did not know them at all or don't need them in their work. Expertise in Hadoop was reported by 8 respondents (44%) as being at an advanced beginner level (2 in the scale of 0 to 5) or lower.

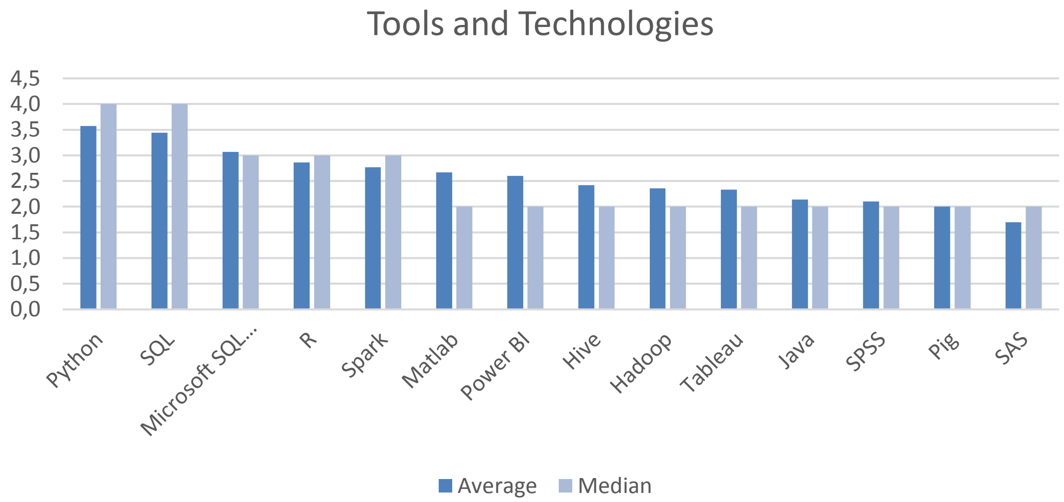

FIGURE 14 Participant's Level of Expertise in Tools and Technologies

In addition, respondents mentioned other tools, including Alteryx, IBM Infosphere streams, QlikView, QlikView Sense, Mixpanel, IBM Watson explorer, IBM Watson content analytics, Cognos, Scala, Disco/PRoM, Excel PowerPivot, IDL, Cognos, and TARGIT.

5.2.3 Most Beneficial Competencies

Most beneficial competencies in data analytics were the basis of responses to open-ended questions of the questionnaire. The responses were analyzed and categorized into competencies, which were identified based on the literature review. When responses were analyzed, they were collected into a matrix. Responses to the open questions were formulated into the competence-performance structure. Table 14 shows the competence-performance structure obtained from three respondent's answers to the open questions of the questionnaire.

Competencies level was derived from the answers to the questions. "X" in the structure means a competence derived from the answer to the question: "Based on your experience which has been the most important competence to create customer value with data analytics?" "S" means a competence derived from the question "Which competencies would help to eliminate or reduce the above-mentioned problems?". Some required competencies also came from answers to questions "Which problems have you encountered in your work related to data analytics and the creation of customer value with it?", which are marked as "P" in the structure, and "Do you have other ideas how the company could better take advantage of analytics-related competencies?" *Performance level* is constructed based on individuals ("Person 1", "Person 2,"etc.) and on the principle that performing well in tasks, a certain person requires particular competencies. Because specific duties were not identified, the competence-performance structure obtained only generalized results.

According to Ley and Albert (2003) based on the structure, it could be surmised which tasks respondents would be able to accomplish. Even if the case is not exactly the same as in the study of Ley and Albert (2003), the theory is well suited to our purpose. In this study, when the performance level includes all the required tasks of certain data analysts (which are described as persons 1, 2, 3, and 4 in the matrix and as numbers in figure 15), the structure can be used to analyze the most beneficial competencies.

TABLE 14 Competence-Performance Structure Regarding Most Beneficial Competencies of Data Analysts

	Passionate problem solver	Reporting skills	Company's strategic planning	Business planning related to data-analytics	Subject knowledge	Data architecture design	Integration of trad and Big Data	
Person 1				X		X, P		Business planning related to analytics Data architecture design
Person 2		X	S	X	X			Reporting skills Companies strategic planning Business planning related to data analytics Subject knowledge
Person 3	X		P	X		P	P	Passionate problem solver Company's strategic planning Business planning related to data-analytics Data architecture design Integration of trad. and big data
Person 4			X	X	X	P, S		Company's strategic planning Business planning related to data-analytics Subject knowledge Data architecture design

In this case, based on the competence-performance structure we can surmise that, for example, for the tasks of Person 1, the most beneficial competencies are “Business planning related to analytics“ and “Data architecture design.” These are the prerequisites, which, can be interpreted as the most beneficial competencies. Competence-performance space that combines all competencies mentioned in the open questions was obtained based on the matrix. It is shown in figure 15.

In the competence-performance space, the boxes describe the most beneficial competencies based on the answers to open questions of persons. The most beneficial competencies can be seen in the bottom of the hierarchy. These competencies are also included in upper person’s competencies. When all the mentioned competencies are included in the competence-performance space, there can be seen many similarities between competencies of the persons. However, quite a number of boxes are not connected to others. They are not connected because a particular person has mentioned competencies that do not entirely fit into competencies mentioned by some other person. It can be confusing when some of the frequently mentioned competencies, such as communication skills, are not at the bottom of the hierarchy.

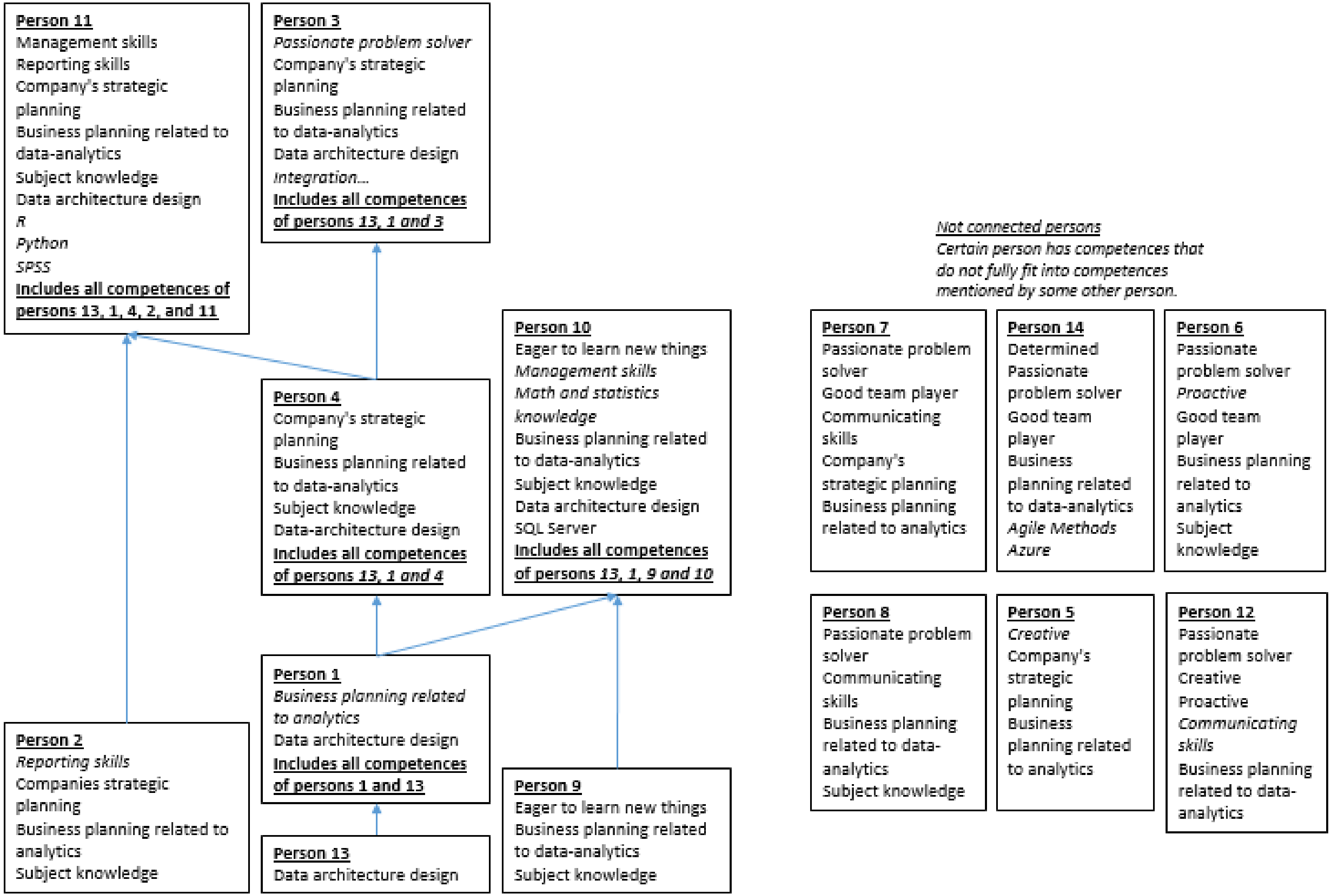

FIGURE 15 Competence-Performance Space of the Most Important Individual Competencies to Create Customer Value

Based on the competence-performance space it can be seen that there are many similarities between the respondents. Furthermore, if we exclude personal traits, which are more difficult to improve, and concentrate on examining professional and technical skills, which are easier to develop through education or training, we find more connections. Competence-performance space of most beneficial professional and technical skills is shown in figure 16.

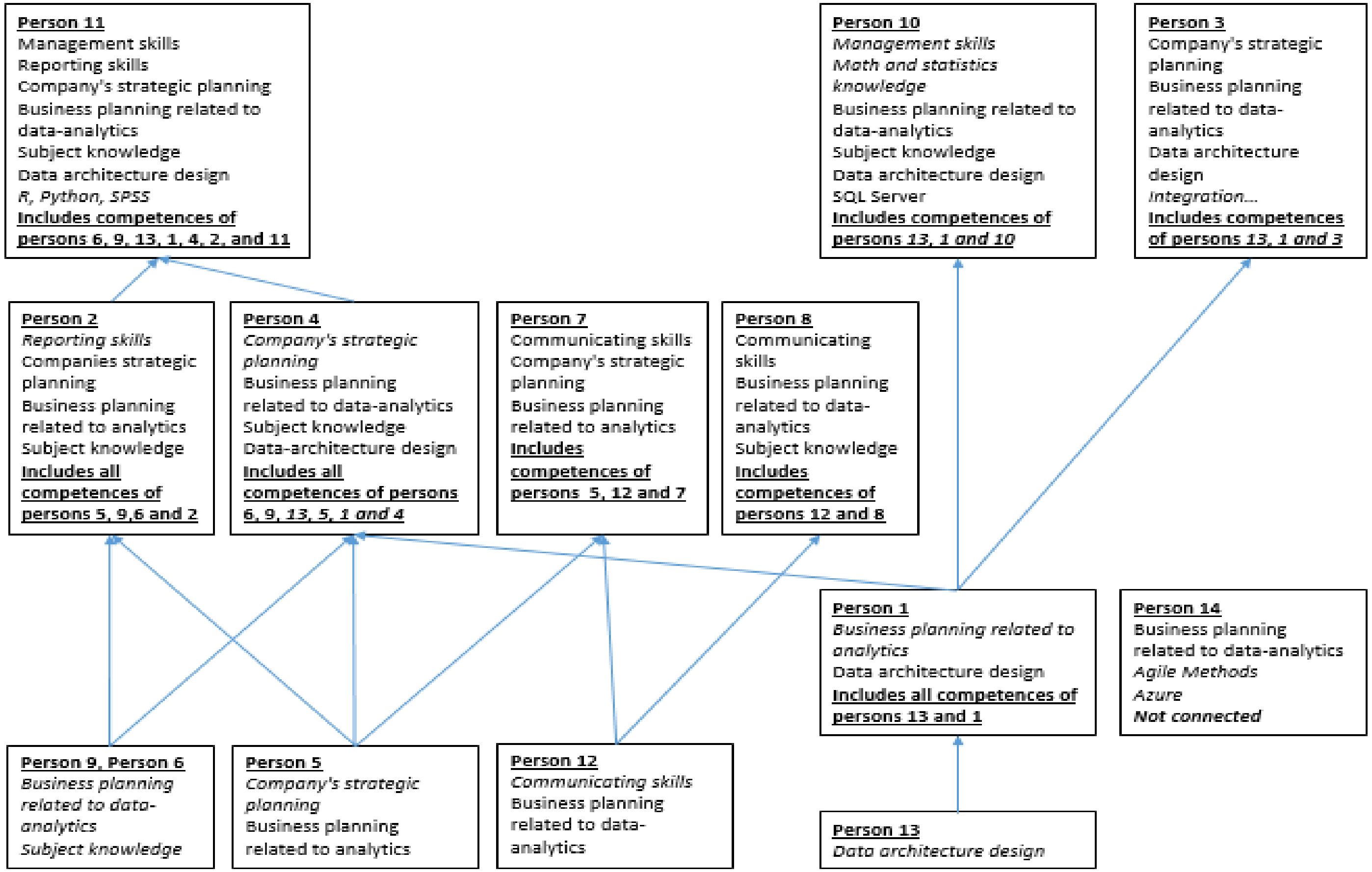

FIGURE 16 Competence-Performance Space of the Most Important Professional and Technical Competencies

In the base of the structure can be found “data architecture design.” Half of answers given could be interpreted as showing that data architecture one of the most beneficial competencies of the individual. Also, respondents highlighted other data-related issues. For example, data availability is often problematic, data quality and integrity is poor, and users have limited knowledge about the data they use every day. Regarding previous issues, the understanding of data architecture and other data related knowledge is beneficial.

More than a half of all respondents expressed the need of seeing the connection between business and data. Knowledge regarding business planning related to analytics and company’s strategic planning seem to be crucial competencies for data analysts. As respondents mentioned, data analysts need to understand the data and its business value for the customer. They need to know the business user and transfer the needs of data, and also show the value of the solution for the customer. It was also mentioned that “even the best expertise in usage of data analytics tools and technologies is of very limited value without considerable knowledge in the field of the application.” One of the respondents summarized the need for business understanding and subject knowledge as follows:

> *“Data analytics is polarized: customers do not understand how analytics can be utilized, and do not know how to buy it. Tenderers are difficult to understand customers' needs in depth unless a lot of discussions and providers cannot, therefore, sell analytics. Analytics applications are always far, "processed", and require a lot of understanding of the topic.”*

Reporting or communicating skills were mentioned in six responses. Data analysts need to “collect data and build a model to get their feedback until satisfied results agreed by scientists and customers.” Often problems in communicating seem to be between the client and data analyst. Obviously, in situations in which the data analyst is a consultant. Lack of real customer input poses problems, and the solutions would be to face the customers directly. When sales pick a potential customer, it would be beneficial to take a data analyst in the discussion from the beginning. Also in other cases, according to some respondents, real customer input and flatter organizations could address unrealistic expectations and problems concerning the use of time in relevant matters. Even if it is certain that data analysts need to have good social skills and be excellent communicators, these issues do not depend only on the competency of the individual, but on organizational problems as well.

Other competencies, which were mentioned as the most beneficial for data analyst when creating customer value were certain personal traits, such as being passionate problem solvers, creative, proactive, determined and eager to learn new things. Regarding the professional expertise of management skills, mathematics and statistics knowledge, agile methods were mentioned. Technical competency in general and some specific competencies, such as integration of traditional (structured, centralized data) and Big Data (unstructured, distributed data) analytics, R, Python, SPSS, Microsoft SQL Server, and Azure, were mentioned.

In addition to competence-performance space, required individual competencies were identified from open-ended questions together with the most beneficial organizational competencies (see table 15). Together with the results from other previous steps they are also discussed in section 6.

Organizational Competency

In addition to individual competencies, the survey produced valuable ideas regarding the activities of organizations. These were collected separately. As mentioned, the most frequently mentioned problem in the data analytics area according to the empirical study seems to be data. Data can, for example, be in poor form, stored poorly for reporting or the database, or data warehouse can be filled with unnecessary data. One respondent described the problems associated with data as follows:

> *"Companies are not willing to invest money on analytics or advancing their data collection/storage/usage. This results that no good analysis can be made and data that they have is not in a shape that could be used to do any analytics. Also, all data is scattered in different places and connecting them for gaining insight might be impossible (e.g. different customer numbers in different systems - data from these systems cannot be combined)."*

One respondent suggested that the data model should be restructured so that needs concerning reporting and analysis could be accomplished in the future. After that, or in those companies that do not have problems mentioned above, the respondent's suggestion could be taken advantage of:

> *"Different data streams could be combined more efficiently. When data collection is reliable and collected data is structured or can be made structured, it can be connected with other sources."*

To structure data models, a strategy is needed. Participants emphasized this topic, especially when thinking about problems in the data analytics area and their solutions. They mentioned a lack of strategic focus and unclear goals. The respondents also reported that analytics is poorly connected to company business and work is focused on irrelevant matters. When the focus is clear, time could be devoted to more essential subjects. A clear strategy could also facilitate the lack of time, of which three respondents reported. In addition to strategy, the importance of the data-driven culture rose from the answers.

Change management is also important to carry out carefully. As one of the participants mentioned, "the normal inertia towards organizations willing to change their processes and thinking" is a problem. He also wrote that the "education and training budget should always be on the same level as an investment to technology and tools." Another respondent raised similar issues.

The results of most beneficial competencies are collected in the table 20. In next chapter, the results of the literature review and empirical study will be combined. They are used to answer the research questions.

6 Discussion

This chapter looks at how this study accomplishes its objectives and provides answers to the research questions. The aim was to give an overview of past research concerning competencies, mostly used terms in data analytics and competency in data analytics. Competence and competency are defined and competency models are discussed in chapter 2. Mostly used terms in data analytics and competency of data analyst based on literature are described in chapter 3. The competency identification process, used in this study to assess the most beneficial competencies, is introduced in section 2.3.2. Its application in this study is described in section 4.2.

In section 6.1 competencies required in data analytics are introduced. After introducing required competencies which, based on this study, create the most customer value in the data analytics area are discussed.

6.1 What are the Competency Requirements in Data Analytics?

This research question was investigated by comparing the literature and empirically identified individual competencies. Results are described in this section and summarized in the tables. Tables compare the most frequently mentioned competencies in job advertisements, based on the thesis of Pinola (2015), the most often frequently competencies based on the literature review of this thesis, and the results from the empirical study. In the tables, competencies are listed in one section, which Pinola (2015) introduced in his thesis. Pinola's literature section results were also taken as a basis. In practice, some of the sources were reused, but other articles were excluded from the study because they were no longer available or were not scientific articles according to *Julkaisufoorumi*. Reused parts are written in *italics* in the tables. In the tables, an empirical study has been categorized into two sections. Current competency describes the results from the structured questions of the study and most beneficial competencies are described based on answers to the open questions of the questionnaire.

Individual competency requirements are introduced first. The empirical study also highlights organizational competencies that are essential for the exploitation of analytics. These are summarized at the end of this section.

6.1.1 Individual Competency Requirements

Personal Traits

Data analyst need to be passionate problem solvers, who are also creative and curious. To succeed in challenging tasks, they need to be eager to learn new things and be determined. Data analyst need to be able to work independently, but the work also requires good social and team working skills. In many cases, they are also required to be pressure tolerant and able to work in the changing environment. (Patil 2011; Davenport & Patil 2012; Dhar 2013; Van den Driest 2016.)

Results are summarized in table 15. Table compare most often mentioned competencies in job advertisements, based on the thesis of Pinola (2015), most often mentioned competencies based on the literature review of this thesis and results from the empirical study.

TABLE 15 Personal Traits of Data Analyst[10]

	Competencies in Job Advertisements collected by Pinola (2015)	**Competencies in Literature**	**Current competency based on empirical study** (level of expertise)	**Most Beneficial based on Empiria** (times mentioned)
Personal Traits	*Creative (16)*	*Passionate problem-solver (Davenport & Patil 2012), Problem-solving skills (Davenport & Patil 2012*, Dhar 2013, Patil 2011)	Passionate problem-solver (4,8)	Passionate problem solver (6)
	Flexibility (10)	Creative, Innovative (Davenport & Patil 2012, Van den Driest 2016; Patil 2011)	Eager to learn new things (4,8)	Good team player (3)
	Able to operate in a changing environment (5)	Good team player (Davenport & Patil 2012), Strong social skills (Davenport 2012)	Able to work independently (4,7)	Eager to learn new things (2)
	Passionate (5)	Pressure tolerant (Davenport & Patil 2012)	Curious (4,6)	Creative (2)
	Curious (4)	Real-time awareness (Davenport & Patil 2012)	Determined (4,3)	Proactive (2)
	Proactive (4)	Curious (Davenport & Patil 2012)	Creative (4,3)	Determined (1)
	Productive (4)	Proactive (Davenport & Patil 2012)	Able to work in a changing env. (4,1)	
	Ability to learn (4)	Productive (Davenport & Patil 2012)	Good team player (4,1)	
	Able to work independently (4)	Eager to learn new things (Davenport & Patil 2012)	Willing to take responsibility (4,0)	
	Willingness to travel (2)	*Determined (Davenport & Patil 2012)*	Productive (3,9)	

[10] Results taken from Pinola (2015) are set in *italics*.

In the literature and job advertisements, personal characteristics received significantly less attention compared to professional or technical skills. It may be that these features receive less attention due to the fact that, according to Iceberg Model of Competence (see section 2.1), these are more difficult to develop than working skills (Spencer & Spencer 1993). However, based on empirical research, certain personal traits are one of the most beneficial competencies. When applying for employees, it is particularly important to pay attention to personal traits of job applicant precisely because they are stable and cannot be easily developed.

Professional skills

Results revealed that subject knowledge and business knowledge related to data analytics are really important competencies for data analysts. Even if, based on the empirical research, the knowledge of business planning related to analytics and company's strategic planning seems not to be at a particularly high level, it would be important to see business issues from a data perspective. (Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013.) Analytics should always be ROI-driven and organizations should invest in technology only if it solves real business problems (Davenport & Dyche 2013). Although close collaboration with the business side is necessary, in order to ask the right questions, the data analyst needs to have an understanding of the subject and the business.

In order to ask the right questions, analytical skills are also required. In that context computational thinking is also important. Excellent communication skills are crucial to the data analyst. They need to know how to communicate verbally and visually in a way that stakeholders understand. Storytelling is recommended. (Patil 2011; Davenport & Patil 2012; Davenport & Dyche 2013; Davenport 2014; Driest et al 2016.) All the necessary professional skills of data analyst are summarized in table 16.

TABLE 16 Professional Skills of Data Analyst[11]

	Competencies in Job Advertisements collected by Pinola (2015) (times mentioned)	**Competencies in Literature** (from most often mentioned)	**Current competency based on empirical study** (level of expertise)	**Most beneficial based on empiria** (times mentioned)
Professional skills	*Analytical skills (53)*	*Subject competence / Business understanding* *(Davenport&Patil 2012; Chen ym.2012; Conway 2011;* Dhar 2013; van den Driest 2016; Provost&Fawcett 2013; Waller&Fawcett 2013)	Problem-solving skills (4,5) (Analytical skills could also include Creativity 4,3)	Business planning related to analytics (13)
	Communication skills (49)	*Analytical skills* (Davenport&Patil 2012; Davenport&Dyche 2013; Dhar 2013; Provost&Fawcett 2013; Waller & Fawcett 2013), Computational thinking (Barr and Stephenson 2011; Dhar 2013, Wing 2006)	Reporting skills (3,9)	Subject knowledge (7)
	Subject knowledge, Business knowledge (39)	Management understanding (Waller & Fawcett 2013, Provost&Fawcett 2013), Decision management (Miller 2014)	Business planning related to analytics (3,6)	Company's strategic planning (6)
	Team working skills (25)	Reporting skills, storytelling (Davenport&Patil 2012; Davenport&Dyche 2013; Driest et al 2016)	Project management, Agile methods (3,4)	Communicating skills (3)
	Reporting skills (2), Documentation skills (22)	*Communication skills* *(Davenport & Patil 2012, Chen ym.2012)*	Management skills (3,2)	Management skills (2)
	Management skills (9)	Strategic data management (Miller 2014)	Company's strategic planning (3,2)	Reporting skills (2)
	Entrepreneurial spirit (4)	Agile methods (Patil 2011)		Agile Methods (1)
	Process know-how (4)	Change Management (Davenport&Dyche 2013)		
	Strategical planning (2)			
	Agile Methods (2)			

The same professional competencies were mentioned in all parts of the study. The desired level of expertise of the participants in relation to these competencies, however, differed. These differences can be used to define areas of development (e.g., knowledge of company's strategic planning).

[11] Results taken from Pinola (2015) are set in *italics.*

Technical knowledge

Data analysts need to master technical skills, such as mathematics and statistics, data architecture design, and databases and data warehouse knowledge (Conway 2011; Dhar 2013; Miller 2014). Based on the empirical study, programming and machine learning skills were not at a particularly high level among data analysts who participated in this study. Instead, in literature, the need for strong programming skills was highlighted relatively often (e.g., Davenport & Dyche 2013; Miller 2014). Nevertheless, these were not mentioned when describing problems regarding analytics, solutions for the problems or most valuable competencies for creating customer value. On that basis, it appears that they are not among the most important skills. Although, it may be that some programming skills are needed.

Competency in data management seems to be required by data analysts. Among study participants, competency in subjects related to data management, such as data architecture design, databases, integration of different kind of data and preparing data for analytics was generally in a quite good level. However, issues related to data management were mentioned when describing the problems and solutions to these problems, and it was also listed as one of the most valuable competencies.

On the basis of empirical results and the literature, competence in predictive analytics, automated and real-time analytics, cloud computing, and network analysis could be added to the list of requirements. The results concerning the technical knowledge of data analysts from all the parts of the study are summarized in table 17.

TABLE 17 Technical knowledge of Data Analyst[12]

	Competencies in Job Advertisements collected by Pinola (2015) (times mentioned)	**Competencies in Literature** (from most often mentioned)	**Current competency based on empirical study** (level of expertise)	**Most beneficial based on empiria** (times mentioned)
Technical knowledge	*Statistics (59)*	*Programming skills (Davenport & Patil 2012;* Davenport & Dyche 2013; Patil 2011; Miller 2014)	Preparing data for analytics (3,9)	Data architecture design (6)
	Machine Learning (50)	*Machine Learning* (Conway 2010; Dhar 2013; Harris, Murphy & Vaisman 2013; Miller 2014)	Math and statistical knowledge (3,7)	Math and statistics knowledge (1)
	Programming skills (38)	*Statistics (Conway 2011;* Dhar 2013; Miller 2014)	Data architecture design (3,6)	Technical skills generally (1)

[12] Results taken from Pinola (2015) are set in *italics.*

	Competencies in Job Advertisements collected by Pinola (2015)	Competencies in Literature	Current competency based on empirical study (level of expertise)	Most beneficial based on empiria (times mentioned)
Technical knowledge	*Algorithms (28)*	Math (Dhar 2013; Miller 2014) Algorithms (Miller 2014)	Integration of traditional and Big Data analytics (3,44)	
	Predictive analytics (27)	Integration of traditional and Big Data (Davenport & Dyché 2013; Davenport 2013)	Automated analytics (3,2)	
	Visualization (13)	Security (Miller 2014; Porter et al 2015), privacy and ethical use of data (Miller 2014)	Machine Learning (3,2)	
	ETL (11)	Predictive analytics (Dhar 2013; Miller 2014)	Cloud computing (3,2)	
	Mathematics (8)	Simulation used in advanced analytics (Gartner 2016, IT Glossary), Simulation skills needed in computational thinking (Barr and Stephenson 2011)	Data warehouse (3,2)	
	Econometrics (7)	Network analysis (Dhar 2013; Davenport 2013)	Big Data -analytics (3,1)	
	Network analysis (4)	Real-time analytics (Davenport 2013; Davenport&Dyche 2013)	Using unstructured data in analytics (3,1)	
	Big Data -systems (4)		Coding skills (3,1)	
	Simulation (4)	Database management (Miller 2014)	Predictive analytics (regression analysis, neural networks, naive bayes, decision trees..) (3,0)	
	Building recommender systems (3)	Data warehouse (Miller 2014)	Simulation (3)	
	Cloud computing (2)	*Hacking skills (Conway 2011)*	Real-time analytics (2,9)	
	Data Warehouses (2)	*Ability to implement a technical system (Chen et al. 2012)*	Processing of IoT data (2,9)	
		Preparing data for analyses (Davenport 2013)	Security and privacy (2,6)	

	Competencies in Job Advertisements collected by Pinola (2015)	**Competencies in Literature** (from most often mentioned)	**Current competency based on empirical study** (level of expertise)	**Most beneficial based on empiria** (times mentioned)
Technical knowledge		Automated analytics (Patil 2011)	Building recommender systems (optimation, simulation, suggestions) (2,5)	
		Econometrics (Dhar 2013)	Econometrics (2,4)	
		Cloud computing (Davenport 2013)	Embedded analytics (2,3)	
		Network analysis used in advanced analytics (Gartner 2016, IT Glossary)	Network analysis (e.g. social networks) (2)	
		Embedded analytics (Davenport & Dyche 2013)		

The list of required technical skills could be used when developing the competency of individuals or the organization itself. The most needed competencies, such as knowledge of data architecture or mathematical and statistical knowledge are quite clear, and other requirements could be discussed and determined using the list.

The differences between current competency and required competencies could be used to define areas for development within organization or for individuals. Here, targets of development for data analysts could be competence in data architectures, programming and machine learning, or security and privacy knowledge.

Tools and Technologies

Concerning tools and technologies, the basic requirements for data analysts seem to be competence in Python, SQL, and R. Competence in SQL Server also seems to be needed in many cases. In contrast to the literature, expertise in Hadoop was not really high. It was not mentioned as important or problematic. By contrast, in the literature, the importance of competence in Hadoop was mentioned frequently (Davenport & Dyche 2013; McAfee et al. 2012; Conwey 2011). Other required expertise in tools and technologies is determined by what tools are used in the company. The results concerning tools and technologies are summarized in table 18.

TABLE 18 Tools and technologies of Data Analyst[13]

Competencies in Job Advertisements collected by Pinola (2015) (times mentioned)	**Competencies in Literature** (from most often mentioned)	**Current competency based on empirical study** (level of expertise)	**Most beneficial based on empiria** (times mentioned)
R (46) *Python (46)* *SQL (37)* *Java (37)* *Hadoop (35)* *SAS (24)* *Hive (22)* *C++ (21)* *Pig (19)* *Matlab (18)*	Hadoop (Conwey 2011; McAfee et al. 2012; Davenport & Dyche 2013; Davenport 2013) R Python SQL NoSQL Pig Hive Excel Impala SAS IBM DB2 Vertica Tableau (Kiron et al. 2012; Davenport & Dyche 2013)	<u>Current competency from highest to lowest:</u> Python SQL Microsoft SQL Server (SSIS, SSAS, SSRS..) R Spark Matlab Power BI Hive Hadoop Tableau Java SPSS Pig SAS	SQL Server (1) Azure (1) Python (1) R (1) SPSS (1)

6.1.2 Organizational Requirements

In addition to competent data analysts, organizations also need to have a strategy regarding data collection, storage, and usage. Data needs to be structured so that reporting and analysis can be accomplished in the future. Organizational requirements are summarized in table 19.

As Miller (2014) states, differentiating in business requires a strategy that considers data as a core business asset. The strategic focus should be clear, and analytics should be connected to company's business (i.e., Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013). It also helps to focus the work of data analyst on relevant matters. Also, the importance of having a data-driven culture has been raised by many researchers (Kiron et al. 2012, Ransbotham et al. 2016). Behaviors, practices, and beliefs should be consistent with the principle that business decisions at every level are based on data analysis (Kiron et al 2012).

When creating new and changing the processes, it is important to carry out change management carefully (Davenport & Dyche 2013; Ransbotham et al. 2016). Ransbotham et al. (2016) are of the opinion that companies should prepare for the robust investment and cultural change that are required to achieve sustained success with analytics. It includes, for example, expanding the skill set of managers who use data. Commitment and hard work are required to execute and sustain a successful analytics strategy. The study also revealed that organizations need to invest in education and training, not only technology and tools.

[13] Results taken from Pinola (2015) are set in *italics.*

Also, as mentioned, communication between the customer/management / business side and the data analyst must be done transparently and so that goals are clear for both. In addition to descriptive reporting and communication of data analysts, the customer side should have realistic expectations and enough time. (Davenport & Dyche 2013.)

TABLE 19 Organizational Requirements for Utilization of Data

	Competencies in Literature	**Most Beneficial Based on Empirical Study**
Organizational competencies	Data-oriented culture (Patil 2011; McAfee et al. 2012; Kiron et al. 2012; Patil & Mason 2015; Ransbotham et al. 2016)	Ensuring that data model fulfill the needs concerning reporting and analysis
	Analytics connected to company's business (Vesset et al. 2012; Dhar 2013; Provost & Fawcett 2013; Waller & Fawcett 2013)	Strategic focus; Strategy, in which data is considered as a core resource
	Strategy, that considers a data as a core business asset (Miller 2014), Commitment and hard work required to execute and sustain a successful analytics strategy (Ransbotham et al. 2016)	Clear and direct communication between customer/user/management and data analyst
	Close relationships between IT and business (Davenport & Dyche 2013)	Data-driven culture
	Capitalizing on data use across the organization (Ransbotham et al. 2016)	Change management Teaching the client organization to use the data in the daily work
	Skill set for managers who use data (Ransbotham et al. 2016)	
	Customer centricity (Van den Driest et al 2016)	

Boyatzis (1982, 2008) is of the opinion that maximum performance is believed to occur when the person's capability or talent is consistent with the needs of the job demands and organizational environment (figure 2). In addition to individual and organizational competency, he describes job demands including role responsibilities and tasks needed to be performed (Boyatzis 2008). Job demands were not mentioned in the empirical study, but these are also good to remember.

6.2 Which Are the Most Beneficial Competencies to Create Customer Value with Data Analytics?

The most beneficial competencies were identified based on competence performance space. It was formed by applying the Competence Performance Theory of Korossy (1999) for answers to open questions in the questionnaire. The results are discussed in section 5.2.3. and can be seen in table 15. Here, the topic is discussed briefly and summarized in a table.

As Davenport & Dyche (2013) mention, many of the challenges faced by big data centers concern the data itself. According to this study, the most beneficial competencies for data analysts in current situation seems to be data architecture modeling and business planning related to data analytics. The key for benefitting from data analytics is well-structured data model, from which the reporting and analysis needs of the future is possible. When the data model has been carefully designed, almost any tool can be used, and data can be integrated more efficiency. Also, knowledge regarding business planning related to analytics, subject knowledge, and company's strategic planning are really beneficial competencies for data analysts. Data analysts need to understand the data and its business value for the customer.

Many of the most important requirements require knowledge and action from both individuals and the organization. For example, to get data models into shape, a strategy and resources from the organization are necessary in addition to the competence of individuals. The most beneficial competencies are given in table 20. The number in parentheses is the number of times a competence was mentioned by respondents.

TABLE 20 Most Beneficial Competencies to Create Customer Value [15]

Personal Traits
Passionate problem solver (6)
Good team player (3)
Eager to learn new things (2)
Creative (2)
Proactive (2)
Determined (1)
Professional Skills
Business planning related to analytics(13)
Subject knowledge (7)
Company's strategic planning (6)
Communicating skills (3)
Management skills (2)
Reporting skills (2)
Agile Methods (1)
Technical skills
Data architecture design (6)
Math and statistics knowledge (1)
Technical skills generally (1)
Integration of traditional (structured, centralized data) and Big Data (unstructured, distributed data) (1)
Tools and Technologies
SQL Server (1)
Azure (1)
Python (1)
R (1)
SPSS (1)

Organizational Competencies
Ensuring that data model fulfill the needs concerning reporting and analysis
Strategic focus, Strategy in which data is considered as a core resource
Clear and direct communication between customer/user/management and data analyst
Data-driven culture
Change management

15 The number in parentheses is the number of times a competence was mentioned by respondents.

7 Conclusion and Suggestions for Future Research

The Competence Performance Theory of Korossy (1999) seems to be a really useful technique for competency identification. The theory could be used in the investigation of the necessary competency in any area. However, in this study, without the knowledge of current competency, results would have been quite limited. Combining the most beneficial competencies, which resulted from open questions using Competence Performance Theory, with the knowledge of current competency, the development plans for individuals or organizations could be derived. By comparing the competencies, the development targets can be found and the knowledge of the development targets can be used (e.g., for individual's development plans or strategic planning of the organization). The results could also be used to determine what kind of employees are needed. The most beneficial competencies can be used to create a short list, and if additional information is needed, a more comprehensive list of requirements can be developed.

Based on this study, it seems that organizations are not using small data or big data exclusively but are integrating the two. As Davenport & Dyche (2013) noted, they are creating a version of 3.0 analytics. As the old base must be taken into account when creating analytics, challenges in adapting operational, development, and decision processes can be seen (Davenport 2013; Davenport & Dyche 2013). As stated at the beginning, there are enormous possibilities to do things more efficiently and faster (Van der Aalst 2012). However, based on this study it can be noticed that there are some basic things that should be considered first, such as strategy and data models.

Related to the competency of individuals, one difference between the descriptions from the literature and data analysts who participated in this study was an expertise in programming skills and expertise in Hadoop. These kinds of results can, however, be attributed to the fact that the sample in this study was relatively small. In addition, even though the sample was small, there were some strong programmers. It may be that tasks are divided among different roles on the team. However, in this study respondents could not be divided into certain competency-based profiles. In this study, for example, the same persons who had advanced or expert competence in coding and machine learning could also be advanced or expert in company's strategic planning. Also, the title of the person who had the previous competency could, for example, as well be a manager or a developer. This is a topic that could be explored in more detail in the future.

Almost all participants indicated that there is a need for deeper knowledge regarding business planning related to analytics. In this context, the so-called citizen data scientist movement is creating a lot of excitement at the moment in non-academic discussions. *Citizen data science* is a term that Gartner introduced in its hype cycle in 2015 (Gartner 2015). The point of the hype is not that highly skilled data analysts are not needed anymore. They still will have a crucial role, but competency in data analytics seems to be also needed outside of the IT lab and across the organization so that more meaningful analytics is possible. Also based on this study, it appears that there is a need for people who put insights into action in organizations and raise awareness and understanding of how to use analytics to improve the business. (Marr 2016.)

Research on the subject could continue in many areas. This study did not explain the kinds of tasks that data analysts undertake. It would be interesting to see what kind of data analytics they actually are implementing. It could give a different perspective on which competencies are required. Tools also received limited attention. In the future, that subject could be studied more broadly. For example, it would be interesting to determine which tools can be used without programming skills. It would also be interesting to study how relevant programming skills are in these days when there are also versatile software platforms, such as Watson and Azure.

8 References

Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3), 443-448.

Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential. *Vikalpa*, 38(4), 1-11.

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: what is involved and what is the role of the computer science education community? *ACM Inroads*, 2(1), 48-54.

Basu, A. (2013). Five pillars of prescriptive analytics success. *Analytics Magazine*, 8-12.

Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, 109(2), 155-172.

Boyatzis, R. E. (1982). *The competent manager: A model for effective* performance. John Wiley & Sons.

Boyatzis, R. E. (2008). Competencies in the 21st century. *Journal of management development*, 27(1), 5-12.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. MIT Press Cambridge, MA

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, 36(4), 1165-1188.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1), 21-26.

Conway, D. (2010). The data science Venn diagram. Drewconway.com Retrieved 2.10.2016 from http://drewconway.com/the-lab/

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research (2nd ed.)*. Thousand Oaks, CA: Sage

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Harvard Business Press.

Davenport, T. H., & Prusak, L. (2000). Working knowledge: how organizations manage what they know. *Ubiquity*, *2000*(August), 6. Retrieved 17.1.2017 from http://ubiquity.acm.org/article.cfm?id=348775

Davenport, T. H., Harris, J. G., David, W., & Jacobson, A. L. (2001). Data to knowledge to results: building an analytic capability. *California Management Review, 43(2)*, 117-138.

Davenport, T. H., & Dyché, J. (2013). Big data in big companies. *International Institute for Analytics*.

Davenport, T. H. (2013). Analytics 3.0. *Harvard Business Review*, *91*(12).

Davenport, T. (2015). *The insight-driven organization: Management of insights is the key.* Deloitte University Press.

Davenport, T (2015B) 5 Essential Principles for Understanding Analytics, *Harvard Business Review*.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56(12),*

64-73.
Van den Driest, F., Sthanunathan, S., & Weed, K. (2016). Building an insights engine. *Harvard business review*, 94(9), 15.
Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*.
Doignon, J. P., & Falmagne, J. C. (2012). *Knowledge spaces.* Springer Science & Business Media.
Dubois, D. D. (1993). Competency-Based Performance Improvement: A Strategy for Organizational Change. *HRD Press*.
Dreyfus, H., Dreyfus, S. E., & Athanasiou, T. (2000). *Mind over machine*. Simon and Schuster.
Eckerson, W. W. (2007). Predictive Analytics–Extending the Value of Your Data Warehouse Investment. TDWI Research: Renton.
Evans, J. R., & Lindner, C. H. (2012). Business analytics: the next frontier for decision sciences. *Decision Line*, *43*(2), 4-6.
Feldman, M. S., & March, J. G. (1981). Information in organizations as signal and symbol. *Administrative science quarterly*, 171-186.
Gartner (2015) Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor. Retrieved 12.2.2017 from http://www.gartner.com/newsroom/id/3114217
Gartner (2016) IT Glossary. Retrieved 21.09.2016 from http://www.gartner.com /it-glossary/
Granville, V. (2014). *Developing analytic talent: Becoming a data scientist.* John Wiley & Sons.
Hammerbacher, J. (2009). Information platforms and the rise of the data scientist. *Beautiful Data*, 73-84.
Harris, H., Murphy, S., & Vaisman, M. (2013). *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work.* O'Reilly Media.
Havelka, D. & Merhout, J.W. (2009). Toward a Theory of Information Technology Professional Competence. *The Journal of Computer Information Systems*, 50(2), 106-116.
Herman, M., Rivera, S., Mills, S., Sullivan, J., Guerra, P., Cosmas, A., & Kim, M. (2013). *The field guide to data science*. Booz Allen Hamilton.
Hirsjärvi, S., Remes, P. & Sajavaara, P. (2009). *Tutki ja kirjoita (15. uud. painos).* Helsinki: Tammi.
Hymes, D. (1971). Competence and performance in linguistic theory. *Language acquisition: Models and methods,* 3-28.
Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 33(7), 14-26.
Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, 1(2), 112-133.
Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2012). Analytics: The widening divide. *MIT Sloan Management Review*, *53*(2), 1.
Korossy, Klaus, et al. (1999) Modeling knowledge as competence and

performance. *Knowledge spaces: Theories, empirical research, and applications*, 1999, 103-132.

Kwon, O., & Sim, J. M. (2013). Effects of dataset features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847-1857.

Landry, B.C., Mathis, B.A., Meara, N.M., Rush, J.E., & Young, C.E. (1970). Definition of some basic terms in computer and information science, *Journal of the American Society for Information Science*, 24(5), 328–342

Laney, D. (2012). Information Economics, Big Data and the Art of the Possible with Analytics. Retrieved 6.3.2017 from: https://www-01.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/Gartner_Doug-%20Analytics/$file/Gartner_Doug-%20Analytics.pdf

LaValle, S., Hopkins, M. S., Lesser, E., Shockley, R., & Kruschwitz, N. (2010). Big Data, Analytics: The new path to value. *MIT Sloan Management Review*, *52*(1), 1-25.

Leonard, A. (2013). How Netflix is turning viewers into puppets. *Salon.*

Ley, T., Albert, D. (2003) Identifying employee competencies in dynamic work domains: methodological considerations and a case study. *J. UCS*, 2003, 9.12: 1500-1518.

Liben-Nowell, D., & Kleinberg, J. (2004). The Link Prediction Problem for Social Networks.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75). Sage.

Lucia, A. D., & Lepsinger, R. (1999). *Art & Science of Competency Models*. San Francisco, CA: Jossey-Bass.

Mansfield, R. S. (1996). Building competency models: Approaches for HR professionals. *Human Resource Management* (1986-1998), 35(1), 7.

Marr, R. (2016) How The Citizen Data Scientist Will Democratize Big Data – *Forbes.* Retrieved 12.01.2017 from http://www.forbes.com/sites/bernardmarr/2016/04/01/how-the-citizen-data-scientist-will-democratize-big-data/#32b0bf124557

Maruyama, H., Kamiya, N., Hicughi, T., & Takemura, A. (2015). Developing Data Analytics Skills in Japan: Status and Challenge. 日本経営工学会論文誌, 65(4E), 334-339.

Maxwell, J. A. 1992. "Understanding and Validity in Qualitative Research", *Harvard Educational Review* (62:3), pp. 279-300

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. *Harvard Bus Rev*, *90*(10), 61-67.

McClelland, D.C. (1973) Testing for competence rather than for "intelligence". *American psychologist 1: 1-14.*

Miller, S. (2014). Collaborative Approaches Needed to Close the Big Data Skills Gap. *Journal of Organization Design,* 3(1), 26-30.

Naur, Peter. (1966). "The science of datalogy". *Communications of the ACM. 9 (7): 485.*

Naur, Peter. (1974). *Concise survey of computer methods.* Studentlitteratur, Lund, Sweden. Retrieved 12.1. 2016 from http://www.naur.com/Conc.Surv.html

Oppenheim, A. N. (2000). *Questionnaire design, interviewing and attitude measurement.* Bloomsbury Publishing.

Papert, S. (1996). An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning, 1(1), 95-123.*

Patil, D. J. (2011). *Building data science teams.* O'Reilly Media, Inc.

Patil, T. H., & Davenport, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review.*

Patil, D.J. & Mason, H. (2015). *Data Driven: Creating a Data Culture*. Sebastopol, CA: O'Reilly Media.

Pinola, T. (2015). Datatieteilijän kompetenssien määritteleminen.

Porter, M. E., Heppelmann, J. E., Ignatius, A. (2015). How smart, connected products are transforming companies. *Harvard Business Review*, 93(10).

Provost, F. & Fawcett, T (2013A). Data Science and its Relationship to Big Data and Data-driven Decision Making. *Big Data Journal*, Vol. 1, No. 1, March 2013

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* O'Reilly Media, Inc.

Quitzau (2013). Retrieved 17.12.2016 from http://www.slideshare.net/AndersQuitzauIbm/ibm-open-data

Ransbotham, S., Kiron, D., & Prentice, P. K. (2016). Beyond the Hype: The Hard Work Behind Analytics Success. *MIT Sloan Management Review,* 57(3).

Riabacke, M. Danielson, M. and Ekenber, L. (2012). State-of-the-Art Prescriptive Criteria Weight Elicitation. *Advances in Decision Sciences.*

Rich, S. (2012). Big Data Is a 'New Natural Resource'. Retrieved 17.12.2016 from http://www.govtech.com/policy-management/Big-Data-Is-a-New- Natural-Resource-IBM-Says.html.

Rouse, M. (2013, October 1). What is advanced analytics? – Definition from WhatIs.com. Retrieved 9.10.2016, from http://searchbusinessanalytics.techtarget.com/definition/advanced-analytics

SAP (2013). Retrieved 17.12.2016 from https://blogs.sap.com /2013/03/05 /predictive-analysis-enables-operational-insights/

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

Russom, P. (2011). Big data analytics*. TDWI Best Practices Report, Fourth Quarter, 1-35.*

Sagiroglu, S., and Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 42-47). IEEE.*

Sandberg, J. (2000). Understanding human competence at work: an interpretative approach. *Academy of management journal*, 2000, 43.1: 9-25.

Sandelowski, M. (1986). "The Problem of Rigor in Qualitative Research," *Advances in Nursing Science* (8:3), pp. 27-37.

Shadish, W. R., Cook, T. D., and Campbell, D. T. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Boston: Houghton-Mifflin.

Sondergaard, Peter (2015). Gartner Analyst Keynote, Symposium/IT Expo, Orlando 2015. Retrieved 5.5.2016 from http://www.gartner.com/smarterwithgartner/the-arrival-of-algorithmic-business/

Spencer, L. M., & Spencer, P. S. M. (1993). *Competence at work: Models for superior performance*, John Wiley & Sons, Inc.

Stonier, T. (1997). *Information and meaning – An evolutionary perspective.* Berlin: Springer

Strategy At Risk (2012). Retrieved 17.12.2016 from http://www.strategy-at-risk.com/2012/06/05/budgeting-revisited/

Taylor, F.W. (1911). The Principles of Scientific Management. Retrieved 17.12.2016 from http://elibrary.kiu.ac.ug:8080/jspui/bitstream/1/516/1/The%20Principles%20Of%20Scientific%20Management.pdf

Teddlie, C., and Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. *Handbook of mixed methods in social & behavioral research, 3-50.*

Teddlie, C., and Tashakkori, A. (2009). *Foundations of Mixed Methods Research*, Thousand Oaks, CA: Sage Publications.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics,* 33(1), 1-67

UPS Company Profile, Corporate Profile: Analytics At UPS March /April 2010, *Informs Analytics*. Retrieved 16.9.2016 from http://analytics-magazine.org /corporate-profile-analytics-at-ups/

Vakola, M., Soderquist, K. E., & Prastacos, G. P. (2007). Competency management in support of organisational change. *International Journal of Manpower*, 28(3/4), 260-275.

Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS quarterly,* 37(1), 21-54.

Vesset, D., Woo, B., Morris, H. D., Villars, R. L., Little, G., Bozman, J. S., ... & Eastwood, M. (2012). *Worldwide big data technology and services 2012–2015 forecast. IDC Report*, 233485.

Wagner, R. (2012). Organizational competence in project management—new perspectives on assessing and developing organizations. *Journal of Project, Program & Portfolio Management*, 3(1), 45-57.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), 77-84.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' net works. *Nature, 393(6684), 440-442.*

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly,* xiii-xxiii.

White, M. (2012). Digital workplaces Vision and reality. *Business information review*, *29*(4), 205-214.

Wing, J. M. (2006). Computational thinking. *Communications. ACM* 49, 33–35.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology*, *58*(4), 479-493.

APPENDIX 1 – QUESTIONNAIRE

Competencies related to Data analytics

Dear respondent,

This survey addresses the importance of data science and analytics competencies and their relation to creating customer value. Answers provide an input to the pro gradu thesis, which is made in the Information Technology Faculty of the University of Jyväskylä.

I hope that you spend a little of your time and answer all the questions. Answering the survey takes approximately 10-15 minutes. It is possible to post a link to your e-mail address if you wish to continue filling the questionnaire at a later time. This can be done by "Keskeytä" button.

The survey data is confidential and the data will be reported only in the aggregate and no individual will be identified.

For further information regarding the survey, contact via e-mail tjstreng@student.jyu.fi.

Thanks in advance,
Tiina Strengell

Hyvä vastaaja,

Tämä tutkimus käsittelee datatieteeseen ja analytiikkaan liittyvien kompetenssien tärkeyttä ja niiden suhdetta asiakkaan saamaan hyötyyn. Vastauksia käytetään aineistona pro gradu -työssä, joka tehdään Jyväskylän yliopiston Informaatioteknologian tiedekunnassa.

Toivon, että käytät hieman aikaasi ja vastaat kaikkiin kysymyksiin. Kyselyyn vastaaminen vie keskimäärin 10-15 minuuttia. Mikäli tahdot jatkaa kyselyn täyttämistä myöhemmin uudestaan, voit lähettää linkin sähköpostiin. Tämä tapahtuu "Keskeytä" -painikkeella.

Kyselyn tiedot ovat luottamuksellisia eikä yksittäisen vastaajan tietoja voi eritellä tuloksista.

Saadaksesi lisätietoja kyselystä, voit olla yhteydessä sähköpostitse tjstreng@student.jyu.fi.

Yhteistyöstä etukäteen kiittäen,
Tiina Strengell

Demographics / Taustatiedot *

Gender / Sukupuoli

◯ Female / Nainen

◯ Male / Mies

Age / Ikä *

◯ < 25

◯ 26-35

◯ 36-45

◯ 46-55

◯ 56-65

◯ >65

Title / Nimike *

____

What is the highest degree or level of school you have completed? / Ylin suorittamani tutkinto *

◯ Comprehensive school / Peruskoulu

◯ Upper secondary level / Toinen aste

◯ Bachelor's degree / Alempi korkeakoulututkinto

◯ Master's degree / Ylempi korkeakoulututkinto

◯ Licentiate or doctorate degree / Lisensiaatin tai tohtorin tutkinto

The field of education / Koulutusala *

____

____

____

Personal capabilities / Henkilökohtaiset ominaisuudet *

Evaluate how well the following competencies describe you in the scale of 1-5, where 1 = Does not describe me at all, 2 = Does not describe me very well, 3 = Describes me somewhat, 4 = Describes well and 5 = Describes very well.

Arvioi kuinka hyvin seuraavat kompetenssit kuvaavat sinua asteikolla 1-5, jossa

1 = Ei kuvaa lainkaan, 2 = Ei kuvaa kovin hyvin, 3 = Kuvaa jonkin verran, 4 = Kuvaa melko hyvin ja 5 = Kuvaa erittäin hyvin.

	1	2	3	4	5	Don't know
Able to work independently / Kykenevä työskentelemään itsenäisesti	○	○	○	○	○	○
Determined / Päättäväinen	○	○	○	○	○	○
Passionate problem solver / Intohimoinen ongelmien ratkaisija	○	○	○	○	○	○
Eager to learn new things / Innokas oppimaan uutta	○	○	○	○	○	○
Curious / Utelias	○	○	○	○	○	○
Creative / Luova	○	○	○	○	○	○
Pressure tolerant / Paineensietokykyinen	○	○	○	○	○	○
Able to work in a changing environment / Kykenevä toimimaan muuttuvassa ympäristössä	○	○	○	○	○	○
Proactive / Oma-aloitteinen	○	○	○	○	○	○
Productive / Aikaansaava	○	○	○	○	○	○
Good team player / Hyvä tiimityöskentelijä	○	○	○	○	○	○
Willing to take responsibility / Halukas ottamaan vastuuta	○	○	○	○	○	○

Professional expertise *

Evaluate your expertise of the following competencies in the scale of 1-5, where 1 = Novice, 2 = Beginner, 3 = Intermediate , 4 = Advanced and 5 = Expert.

Arvioi osaamistasi seuraavien kompetenssien osalta asteikolla 1-5, jossa 1 = Noviisi, 2 = Aloittelija, 3 = Keskitasoinen, 4 = Edistynyt ja 5 = Expertti.

	1	2	3	4	5	Don't know/ Not needed
Management skills / Johtamistaidot	○	○	○	○	○	○
Reporting skills / Raportointitaidot	○	○	○	○	○	○
Math and statistics knowledge / Matemaattinen ja tilastollinen osaaminen	○	○	○	○	○	○
Company's strategic planning / Yrityksen strateginen suunnittelu	○	○	○	○	○	○
Problem-solving skills / Ongelmanratkaisutaidot	○	○	○	○	○	○
Business planning related to data analytics / Data-analytiikkaan liittyvä liiketoiminnan suunnittelu	○	○	○	○	○	○

	1	2	3	4	5	Don't know/ Not needed
Project management skills / Projektinhallintataidot	○	○	○	○	○	○
Agile methods / Ketterät menetelmät	○	○	○	○	○	○
Data architecture design / Data-arkkitehtuurin suunnittelu	○	○	○	○	○	○

Technical skills / Tekniset taidot *

Evaluate your expertise of the following competencies in the scale of 1-5, where 1 = Novice, 2 = Beginner, 3 = Intermediate, 4 = Advanced and 5 = Expert.

Arvioi osaamistasi seuraavien kompetenssien osalta asteikolla 1-5, jossa 1 = Noviisi, 2 = Aloittelija, 3 = Keskitasoinen, 4 = Edistynyt ja 5 = Expertti.

	1	2	3	4	5	Don't know/ Not needed
Coding skills / Koodaustaito	○	○	○	○	○	○
Databases / Tietokannat	○	○	○	○	○	○
Data Warehouses / Tietovarastot	○	○	○	○	○	○
Integration of traditional (structured, centralized data) and Big Data (unstructured, distributed data) analytics / Perinteisen ja Big Data -analytiikan integrointi	○	○	○	○	○	○
Predictive analytics (regression analysis, neural networks, naive bayes, decision trees..) / Ennustava analytiikka	○	○	○	○	○	○
Building recommender systems (optimation, simulation, suggestions) / Suosittelujärjestelmien rakentaminen	○	○	○	○	○	○
Web analytics	○	○	○	○	○	○
Mobile analytics / Mobiilianalytiikka	○	○	○	○	○	○
Using unstructured data in analytics (videos, images, sounds, text..) / Strukturoimattoman datan käyttö	○	○	○	○	○	○
Big Data -analytics / Big Data -analytiikka	○	○	○	○	○	○
Econometrics / Ekonometria	○	○	○	○	○	○
Preparing data for analytics / Datan valmistelu analytiikkaa varten	○	○	○	○	○	○
Simulation / Simulointi	○	○	○	○	○	○
Network analysis (e.g. social networks..) / Verkostoanalytiikka	○	○	○	○	○	○
Cloud computing / Pilvilaskenta	○	○	○	○	○	○

Information security and privacy / Tietoturva- ja tietosuoja	○	○	○	○	○	○
Processing of IoT data / IoT -datan käsittely	○	○	○	○	○	○
Embedded analytics / Sulautettu analytiikka	○	○	○	○	○	○
Automated analytics / Automatisoitu analytiikka	○	○	○	○	○	○
Real-time analytics / Reaaliaikainen analytiikka	○	○	○	○	○	○
Machine Learning / Koneoppiminen	○	○	○	○	○	○

Technologies and tools / Teknologiat ja työkalut *

Evaluate your expertise of the following competencies in the scale of 1-5, where 1 = Novice, 2 = Beginner, 3 = Intermediate, 4 = Advanced and 5 = Expert.

Arvioi osaamistasi seuraavien kompetenssien osalta asteikolla 1-5, jossa 1 = Noviisi, 2 = Aloittelija, 3 = Keskitasoinen, 4 = Edistynyt ja 5 = Expertti.

	1	2	3	4	5	Don't know / Not used
R	○	○	○	○	○	○
Python	○	○	○	○	○	○
SQL	○	○	○	○	○	○
Hadoop	○	○	○	○	○	○
Java	○	○	○	○	○	○
Hive	○	○	○	○	○	○
Power BI	○	○	○	○	○	○
SAS	○	○	○	○	○	○
SPSS	○	○	○	○	○	○
Tableau	○	○	○	○	○	○
Matlab	○	○	○	○	○	○
Pig	○	○	○	○	○	○
Spark	○	○	○	○	○	○
Microsoft SQL Server (SSIS, SSAS, SSRS..)	○	○	○	○	○	○

Please indicate other data analytics tools, techniques, and technologies you have experience with.
Luettele muut data-analytiikan työkalut, tekniikat ja teknologiat, joista sinulla on kokemusta.

____

____

Based on your experience which has been the most important competencies to create customer value with data analytics?
Mitkä data-analytiikkaan liittyvät kompetenssit ovat kokemuksesi mukaan olleet tärkeimpiä asiakkaalle luodun arvon kannalta?

Which problems have you encountered in your work related to data analytics and the creation of customer value with it?
Millaisia data-analytiikkaan ja sen avulla luotavaan asiakashyötyyn liittyviä ongelmia olet kohdannut työssäsi?

Which competencies would help to eliminate or reduce the above mentioned problems?
Mitkä kompetenssit auttaisivat poistamaan tai vähentämään edellä mainittuja ongelmia?

Do you have other ideas how the company could better take advantage of analytics-related competencies?
Kuinka yrityksessä voitaisiin mielestäsi hyödyntää analytiikkaan liittyvää osaamista nykyistä paremmin?

Other comments?
Muita kommentteja?