LUND UNIVERSITY
School of Economics and Management

Department of Informatics

Challenges and opportunities with using cognitive computing for mapping personal data in order to help comply with the GDPR

Master thesis 15 HEC, course INFM10 in Information Systems

Authors: Lukas Hedqvist
Henrik Månsson

Supervisor: Odd Steen

Examiners: Bo Andersson
Miranda Kajtazi

Challenges and opportunities with using cognitive computing for mapping personal data in order to help comply with the GDPR

AUTHORS: Lukas Hedqvist and Henrik Månsson

PUBLISHER: Department of Informatics, Lund School of Economics and Management, Lund University

PRESENTED: June, 2018

DOCUMENT TYPE: Master Thesis

NUMBER OF PAGES: 99

KEY WORDS: GDPR, Cognitive computing, mapping, data, unstructured

ABSTRACT (MAX. 200 WORDS):

As the world is becoming more digital, the use of data is rapidly increasing. To protect all EU citizens from privacy and data breaches, the regulations will become stricter. The General Data Protection Regulation (GDPR) will become effective on the 25th of May 2018. This will mean big structural changes for organizations, but it will also mean tougher requirements for gathering and maintaining personal data. The GDPR changes how personal data is defined, meaning that the same laws will apply to all personal data. Furthermore, the GDPR requires organizations to map and identify personal data in both structured and unstructured data sources before the it becomes effective. In this thesis, a case study of a project has been conducted, in which the challenges and opportunities with using cognitive computing for mapping personal data in order to help comply with the GDPR, have been identified. The findings present several challenges and opportunities with using cognitive computing in this specific context. Identified challenges were amongst others the need for manual work and human contributions. Identified opportunities were amongst others that it can understand structured and unstructured data and be learned to identify personal data.

Content

Tables

1 Introduction

1.1 The General Data Protection Regulation (GDPR)

As the world is becoming more digital and the use of data is rapidly increasing, so is the transition within organizations, using more digital systems to enhance the use of data. However, the use of personal data and the privacy and security it demands have long been discussed, and the 27 of April 2016, the European parliament announced that a new regulation regarding personal data will become effective on the 25th of May 2018 (REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016). The new regulation will be called the General Data Protection Regulation (GDPR) and will replace the old data directive 1995/46/EG (REGULATION (EU) 2016/679) and thus, also the Swedish Personal Data Act which is based on the old directive (Datainspektionen, 2018a). The aim of the GDPR is to protect all EU citizens from privacy and data breaches (REGULATION (EU) 2016/679) in an increasingly data-driven world.

The transition to the new regulation will mean big structural changes for organizations with all the obligations that the new regulation requires, but it will also mean tougher requirements for gathering and maintaining personal data. Currently, according to the old Swedish Personal Data Act (Personuppgiftslagen SFS 1998:204), personal data means all kind of information that is directly or indirectly referable to a natural person who is alive. With the old directive, there are simpler rules for unstructured data sources (Datainspektionen, 2018b). However, the GDPR changes how personal data is defined, meaning that the same laws will apply to all personal data, there will be no difference between structured and unstructured data (Datainspektionen, 2018b). The new definition is described in the new regulation (REGULATION (EU) 2016/679) as:

"'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". (REGULATION (EU) 2016/679, Article 4:1)

The aim of the new regulation is to protect all EU citizens from privacy and data breaches with more vigorous penalties for organizations violating the new regulation (REGULATION (EU) 2016/679). There are major differences between the old and the new regulation concerning personal data and how it needs to be handled. The new regulation distinctly states a number of requirements regarding personal data. However, the following requirements will be addressed in this thesis. The right to be forgotten, which entitles the data subject to have the data

controller erase his or her data if for example the personal data are no longer necessary or if the personal data have been unlawfully processed (REGULATION (EU) 2016/679, Article 17). The right to access, which entitles the data subject to obtain confirmation whether or not their personal data is being processed, where and for what purpose (REGULATION (EU) 2016/679, Article 15). The right to data portability, which entitles the data subject receive the personal data concerning him or her, in a structured, commonly used and machine-readable format and have the right to transmit the data to another controller (REGULATION (EU) 2016/679, Article 20). The lawfulness of processing, which states for instance, that processing shall be lawful only if the processing is necessary for the legitimate interests pursued by the controller or by a third party, except where such interest is overridden by the interest, rights and freedoms of the data subject (REGULATION (EU) 2016/679, Article 6). The records of processing activities, which requires that each controller shall maintain a record of processing activities under its responsibility, containing for instance the purposes of the processing, and a description of the categories of data subjects and of the categories of personal data (REGULATION (EU) 2016/679, Article 30).

1.2 Structured and unstructured data

Structured data refers to the tabular data found in spreadsheets or relational databases (Gandomi & Haider, 2015). Structured data is understood as data that is assigned to dedicated fields and can thereby be directly processed with computing equipment (Baars & Kemper, 2008). While structured data follow a specified format, unstructured or semi-structured data does not, it follows a variety of structures and could be large (Hurwitz, Kaufman, and Bowles, 2015). Text, images, audio, and video are examples of unstructured data (Gandomi & Haider, 2015). More specifically, unstructured data sources could be anything from data from documents, books, and journals articles but it could also be from clinical trials, customer support systems, satellite images, scientific data, radar or sonar data, mobile data, website content, and social media sites (Hurwitz, Kaufman, and Bowles, 2015). The semantics of unstructured or semi-structured data types are not explicitly defined, instead they must be identified and extracted with techniques such as text analytics, natural language processing, and machine learning (Hurwitz, Kaufman, and Bowles, 2015).

1.3 Case description of the Aigine project

Aigine AB are developing a cognitive computing system that can map personal data from unstructured data sources, in order to help comply with the GDPR. Aigine AB is the company who are providing the solution but Aigine is also the name of the cognitive computing system, that have been developed with the help of Elinar OY LTD. Elinar is a Finnish company and also a long time IBM business partner. They specialize in information management and have broad expertise and experience with artificial intelligence. Sambruk is a non-profit association that works with municipal business development and have around 100 members. Sambruk have financed the development of the specification of requirements for a suitable solution, together with 10 municipality members, one regional federation, and SKL (Sveriges kommuner och landsting) (Produktbulletin, 2018). The specification of requirements is designed to work from a municipality perspective. However, Aigine AB does not strictly provide their solution

only to municipalities in Sweden, it is aimed towards anyone who needs to improve the control of their unstructured data, and thus, be able to help comply with the GDPR.

The solution Aigine AB have developed works by first scanning unstructured data sources. The cognitive computing system then sorts out all data sources which does not contain any personal data and marks up all presumptive personal data that it finds. The cognitive computing system help the users find personal data and proposes the legal basis for keeping it, with the help of cognitive abilities in assistance with a knowledge database. The user or reviewer then assess and evaluate the findings of the cognitive computing system through the user-interface. Finally, the user's evaluation helps the cognitive computing system to learn and develop more accurate findings, and thus, become more effective.

1.4 Problem area

The GDPR changes how personal data is defined. The new definition of personal data has changed significantly from the previous data protection regulation. With the new definition there will be no difference regarding how structured and unstructured data is regulated. This means that the unstructured data in particular will become more strictly regulated, compared to the previous regulation. In the article: "The big unstructured data problem" published by Forbes, it is presented that analysts at Gartner estimated that upward of 80% of organizations data is unstructured (Rizkallah, 2017). Hurwitz, Kaufman, and Bowles (2015) also addresses the problem regarding the amount of unstructured data, which they argue is as much as 80 percent of all data and is increasing rapidly. Thereby, with the huge amount of unstructured data, it becomes a major difficulty for organizations to comply with the requirements of the GDPR. Furthermore, it requires organizations to map and identify personal data in both structured and unstructured data sources before the GDPR become effective. Rizkallah (2017) explains that the majority of organizations are struggling with managing and protecting the unstructured data, they do not understand how much they have or even what the unstructured data files contain. If organizations are not able to comply with the GDPR, there are penalties to be expected. Organizations in breach of the GDPR can be fined up to 4% of the annual global turnover or €20 Million, whichever is greater (REGULATION (EU) 2016/679). Similarly, organizations can be fined 2% of the annual global turnover for not having their records in order, not notifying the supervising authority and data subject about a breach, or not conducting impact assessment (REGULATION (EU) 2016/679). There are a number of ways of handling the unstructured data, much is dependent on the size of organization, amount of data, and the control of the data within the organization.

1.5 Research question

The research question is formulated with intent to identify the challenges and opportunities with using cognitive computing for mapping personal data to help comply with the GDPR. We aim to investigate the usage of cognitive computing as a technology in order to help organizations to comply with the GDPR, and thereby identify challenges and opportunities from an organizational perspective. Our research question is thus formulated as:

- ***What are the challenges and opportunities with using cognitive computing for mapping personal data in order to help comply with the GDPR?***

1.6 Purpose

The purpose of this study is to examine cognitive computing as a solution to handling the amount of unstructured data, to help become compliant with the GDPR. Furthermore, what the challenges and opportunities are with using it for mapping personal data. Both Hurwitz, Kaufmann, and Bowles (2015) and analysts at Gartner (Rizkallah, 2017), emphasize the huge amount of unstructured data within organizations. To comply with the GDPR, organizations need to map all their personal data. While there are many possible solutions to do this, we aim to investigate cognitive computing as a solution.

1.7 Delimitations

Our aim is not to provide guidelines or a solution on how to successfully use cognitive computing, but rather to identify the challenges and opportunities found in Aigine's project of mapping personal data to help comply with the GDPR. We do not aim to investigate whether using cognitive computing in this specific context is a suitable solution or not, but rather what challenges and opportunities it presents. We do not compare the Aigine project with other solutions to handle unstructured data, but rather compare challenges and opportunities found in their project with conducted literature. The study does not focus on how to fully comply with the GDPR or how its requirements should be interpreted, but rather how the Aigine project can help comply with the GDPR, mainly regarding unstructured data. With this in mind, we will not focus on all requirements of the GDPR, but rather those who are relevant to the Aigine project and unstructured data. We will not investigate or examine research on how to comply with the GDPR, or other studies on how unstructured data can be handled in correlation with the GDPR. We do not aim to study the structural changes required within organizations to implement a system, in this case a cognitive computing system.

2 Cognitive computing review

This chapter will provide with an overview of existing literature about cognitive computing. First, cognitive computing is defined. Followed by a description of cognitive computing, and applications and systems. Lastly, challenges and opportunities of existing literature will be presented. Finally, a theoretical model containing a summary of challenges and opportunities is presented. This chapter is constructed with the intent to provide an understand of cognitive computing.

2.1 Definitions of cognitive computing

Cognitive computing was originally referred to as artificial intelligence, however researchers began to use the term cognitive computing instead in the 1990s, to indicate that the science was designed to teach to think like a human mind, rather than developing an artificial system (Noor, 2015). The definition of cognitive computing is a topic of debate in the computing community (Chen, Argentinis, and Weber, 2016). Although there is not a well-accepted definition of cognitive computing, Watson (2017) argues that it generally refers to hardware or software that mimics the functioning of the human brain to improve decision making, and it is often associated with artificial intelligence. There have been many different definitions by researchers across time and contexts, according to Gutierrez-Garcia and López-Neri (2015) cognitive computing was first coined in 1995 by Valiant (1995), with the definition:

"*a discipline that links together neurobiology, cognitive psychology and artificial intelligence*" (Valiant, 1995, p.2)

A newer and more relevant definition of cognitive computing is presented by Demirkan, Earley and Harmon (2017):

"Cognitive computing refers to the development of computer systems modeled after the human brain, which has natural language processing capability, learn from experience, interact with humans in a natural way, and help in making decisions based on what it learns." (Demirkan, Earley, and Harmon, 2017, p.16).

2.2 Cognitive computing

Smart machines are becoming more like humans as a result of their ability to for example process natural language, and learn (Demirkan, Earley, & Harmon, 2017). Cognitive computing is about applying knowledge from cognitive science to build systems that simulate human thought process (Jones, 2017). It includes multiple disciplines, including machine learning, natural language processing, vision, and human-computer interaction (Jones, 2017). Cognitive computing has great potential to increase efficiency, reduce costs, and improve outcomes by

analyzing large amounts of data and performing routine tasks (Demirkan, Earley, & Harmon, 2017). Cognitive computing refers to systems that can learn at scale, reason with purpose, and interact with humans and other systems in a natural way (Demirkan, Earley, & Harmon, 2017). They can understand by processing unstructured information in similar ways as humans, which makes them understand patterns in language and sensory inputs, such as text and pictures (Mertens, 2017). They can reason and understand underlying concepts, infer and extract ideas, and form hypotheses, because they synthesize information rapidly to deliver appropriate and meaningful responses (Mertens, 2017). Moreover, cognitive computing systems learn and improve from every interaction and outcome (Demirkan, Earley, & Harmon, 2017; Mertens, 2017, Sheth, 2016), resulting in it building a deep knowledge base that is always updated (Mertens, 2017). Algorithms of cognitive computing systems interpret data by learning and matching patterns in a way that tries to mimic the human mind (Sheth, 2016).

The development of cognitive computing is a result of combining machine learning algorithms and traditional knowledge engineering, and applying breakthroughs from emerging technologies (Demirkan, Earley, & Harmon, 2017). Cognitive computing applications use tools such as natural language processing, intelligent search, image recognition, and decision analysis to adapt their underlying algorithms based on exposure to new data (Tarafdar, Beath, & Ross, 2017). In some cases, they do a better and much faster job than humans at for example recognizing patterns, performing rule-based analysis on very large amounts of data, and solving both structured and unstructured problems (Demirkan, Earley, & Harmon, 2017). There already exist services enabled by cognitive computing that we interact with, such as Apple Siri, IBM Watson, Google Go, Amazon Echo, and Microsoft Cortana (Demirkan, Earley, & Harmon, 2017). IBM Watson is a cognitive computing system that can process and reason about natural language and learn from documents, which defeated human champions in the game show Jeopardy! in 2011 (Sheth, 2016). The goal of cognitive computing is to increase productivity and creativity such as decision-making, innovation, connectivity, and augmentation, of both individuals and organizations (Demirkan, Spohrer, & Welser, 2016).

2.3 Applications and systems

Cognitive computing is a new and emerging field that enables entirely new ways of interaction between humans and computers, which works by simulating human reasoning processes and translates them into suited templates (Earley, 2015). Cognitive computing systems could be used for any activity that requires data in order to improve both processes and product, for example in training and customization (Earley, 2015). The cognitive computing system takes input from the user and provides an appropriate response, it can both answer questions and understand context (Earley, 2015). Decisions that have typically been made by humans can now be managed by cognitive computing systems, or with the assistance of cognitive computing systems (Earley, 2015).

Business functions throughout an enterprise can improve its efficiency dramatically with the use of cognitive computing (Finch, Goehring, and Marshall, 2017). Finch, Goehring, and Marshall (2017) mentions many departments that can be affected, ranging from core back office systems such as finance, IT, and HR to product development, manufacturing, and supply chain, to front office functions such as customer services, marketing and sales. Cognitive

computing enables faster and more effective operations, as well as new capabilities (Finch, Goehring, and Marshall, 2017). This could potentially result in gaining competitive advantages and thereby expand revenue opportunities (Finch, Goehring, and Marshall, 2017). Additionally, Coccoli and Maresca (2018) argues that the most promising feature of cognitive computing is its ability to manage huge quantities of information. They believe cognitive computing systems can easily overtake present solutions for both decision support and big data management. Another very important capability for companies to gain advantage is to be able to impart knowledge to employees with limited expertise (Hurwitz, Kaufman, and Bowles, 2015). Cognitive computing systems can learn from experienced practitioners and over time be fed newer data which expands the depth of knowledge, this can enhance the learning process for new practitioners which can gain a faster understanding of most up to date processes (Hurwitz, Kaufman, and Bowles, 2015). Moreover, as new best practices and new information emerge, the cognitive computing system can be trained, resulting in a dynamic knowledge source than can become a competitive advantage and differentiator for businesses (Hurwitz, Kaufman, and Bowles, 2015).

A well-known cognitive computing system is IBM Watson, which combines capabilities in natural language processing (NLP), machine learning techniques, analytics, hypothesis generation and evaluation (Hurwitz, Kaufman, and Bowles, 2015). IBM Watson gets smarter and gains insights with every user interaction and every time new information is acquired (Hurwitz, Kaufman, and Bowles, 2015). By combining all capabilities, IBM Watson is intended to help practitioners and professionals create hypotheses from data, accelerate findings, and determine the availability of supporting information for solving problems (Hurwitz, Kaufman, and Bowles, 2015). IBM Watson uses more than 100 advanced techniques for analyzing natural languages, this makes it potentially able to observe, interpret, and evaluate as humans do (Lee et al., 2016). Practitioners and professionals using IBM Watson in different industries find not only the question and answering capability to be a valuable aspect of the system, but rather the ability to discover, through human-machine interaction, which allows them to think and work differently (Kelly & Hamm, 2013). The ability for humans to interact with machines in a natural way, enables, according to IBM new ways of improving business outcomes (Hurwitz, Kaufman, and Bowles, 2015).

In a study conducted by Tarafdar, Beath, and Ross (2017) based on 51 initiatives and use cases for cognitive computing applications within industries across North America, Europe, and the Asia-Pacific region, of which about 70 percent of the initiatives were either in working proof of concept, in production, or in use. They found three different types of business objectives that these applications aimed towards, and in many cases each application aimed towards one of these three objectives: delighting customers, creating a superior experience for employees, or driving operational excellence. The cognitive computing applications with the business objective of operational excellence include sophisticated search and retrieval from large amount of information (Tarafdar, Beath, and Ross, 2017). The use of such applications was to enhance accuracy and reliability, and speed of the organization's core processes (Tarafdar, Beath, and Ross, 2017). Compared to those applications, the ones with the business objective of delighting customers were aimed towards fostering customer loyalty and engagement or by offering customers superior products or services (Tarafdar, Beath, and Ross, 2017). The last type of cognitive computing applications with the business objective of creating superior employee experience, include capabilities such as natural language processing,

pattern matching, text mining to help employees find solutions to problems (Tarafdar, Beath, and Ross, 2017).

2.4 Challenges with cognitive computing

Most cognitive computing systems have a focused approach, designed to learn and provide value to users in a specific field or industry, in every industry and every domain within that industry, there exists a specific vocabulary and its own historical knowledge (Hurwitz, Kaufman, and Bowles, 2015). Hurwitz, Kaufman, and Bowles (2015) says that it is a challenge for cognitive computing system builders to capture enough relevant knowledge to be useful and to represent it in a way that allows the system to add to the knowledge. For example, the domains include a lot of different types of objects, and each of the object types may have specific rules that guide their interaction and behavior (Hurwitz, Kaufman, and Bowles, 2015). To capture and represent the knowledge of an industry, experts who understand vocabularies and rules are required, furthermore they need to understand the industry well enough to be able to explain them in a way that they can be codified for machine processing (Hurwitz, Kaufman, and Bowles, 2015). Hurwitz, Kaufman, and Bowles (2015) suggests that taxonomies and ontologies with focus on a specific area of knowledge, should be defined. However, they argue that it is not possible to acquire enough knowledge to design a system that replicates an industry or market, not even with the assistance of industry experts. Because of this, most cognitive computing systems starts with a substantial subset of domain knowledge, and then dynamically enhance and refine it with experience or training (Hurwitz, Kaufman, and Bowles, 2015). Furthermore, cognitive computing applications fully depend on massive and continually growing amounts of data, to be able to return outputs that are matching the data (Tarafdar, Beath, & Ross, 2017).

Chen, Argentinis, and Weber (2016) explains that to enable language comprehension, cognitive computing systems must be supplied with relevant dictionaries and thesauri. Interpretation entails the ability to understand data, the language beyond the definitions of individual terms, to deduce the meaning of sentences and paragraphs (Chen, Argentinis, and Weber, 2016). Humans take in content and information by reading and recognizing words and then translate them into abstract meaning, likewise a cognitive computing system can leverage known vocabulary to deduce the meaning of new terms based on context, making it a key function to learn the language of a specific industry or domain (Chen, Argentinis, and Weber, 2016). Cognitive computing systems differs from keyword search and text analytics by having the ability to understand verbs, adjectives, and prepositions, which enables them to comprehend what the language actually means, instead of simply what it says (Chen, Argentinis, and Weber, 2016).

Kelly and Hamm (2013) also discuss the importance of language. They mean that language is what makes it possible for machines to help us understand the world better, as well as to engage with us. But language in this case is not as simple as for example English or Chinese. The special language in a specific domain of knowledge or expertise can consist of chemical symbols, medical images, or be embedded in legal terms (Kelly & Hamm, 2013). The challenge is to teach systems structure, vocabulary of spoken and written languages, as well as business-specific words and concepts (Kelly & Hamm, 2013).

Machine learning can be used to improve the accuracy of predictive models in a cognitive environment, as it enables the models to learn from data and enhance the knowledge base for a cognitive computing system (Hurwitz, Kaufman, and Bowles, 2015). Ensuring that the required amount of data is available is challenging, since cognitive computing applications need substantial amount of test data to train the model (Tarafdar, Beath, & Ross, 2017). Supervised learning and unsupervised learning are two traditional learning techniques (Hurwitz, Kaufman, and Bowles, 2015; Chen, Herrera & Hwang, 2018), which are based on closed training with data input (Chen, Herrera & Hwang, 2018). Supervised learning techniques usually use labeled data to train a model, while unsupervised learning uses unlabeled data (Hurwitz, Kaufman, and Bowles, 2015). Labeled data have some identification or tag that provides some information about the data, often provided by humans as part of the training (Hurwitz, Kaufman, and Bowles, 2015). However, the tag can be different in different conditions, meaning that the usability of information learned by machines is different for different users (Chen, Herrera & Hwang, 2018). Chen, Herrera and Hwang (2018) argues that these traditional learning techniques are unable to meet the requirements of sustainable improvement in the intelligence of machine. Supervised learning generally begins with an established set of data and a certain understanding of how that data is classified (Hurwitz, Kaufman, and Bowles, 2015). Algorithms are trained using preprocessed examples and then the performance of the algorithms is evaluated with test data (Hurwitz, Kaufman, and Bowles, 2015).

IBM Watson can show its full capability only through sufficient training (Murtaza et al., 2016). To do this, question-answer pairs needs to be produced in natural languages (Murtaza et al., 2016). Producing a sufficient number of question-answer pairs requires a lot of work. Machine learning techniques to generate question-answer pairs have been proposed by Serban et al. (2016), where resulting datasets are released for training large scale question-answer systems. To train a IBM Watson instance, the user submits a prepared question, and then IBM Watson suggests a relevant answer list (Lee et al., 2016). Additionally, the user can select a paragraph and specify some parts of an input document as the answer (Lee et al., 2016). Training IBM Watson can be tedious and may not be efficient if certain guidelines are not followed (Murtaza et al., 2016). According to Murtaza et al. (2016), the most important step in successful training and deployment of IBM Watson is to create well-formatted documents.

Many companies have invested in cognitive computing applications, however they have not experienced significant benefits (Tarafdar, Beath, & Ross, 2017). A reason for this could be a lack of understanding about how the applications can contribute to the company's business objectives (Tarafdar, Beath, & Ross, 2017). Business leaders must choose the right tools and make sure the required data is available to be able to take advantage of the applications ability to process massive amounts of data (Tarafdar, Beath, & Ross, 2017). It is challenging to identify the right data, and applications might require external data, real-time data, or the organization's internal data (Tarafdar, Beath, & Ross, 2017). A related challenge is to get the data into the right format, for the data to be accessible to its application it needs to be appropriately cleaned, formatted, and structured (Tarafdar, Beath, & Ross, 2017).

Hurwitz, Kaufman, and Bowles (2015) highlights the importance of security in cognitive computing applications, given the expansive nature of big data and the speed that data sources have to be deployed in many situations. Additional security is required when data comes from

real-time devices such as sensors and medical devices (Hurwitz, Kaufman, and Bowles, 2015). Security infrastructure must ensure security when data is in motion and distributed, since it is often culled from a variety of sources and then used for different purposes than it was originally intended for (Hurwitz, Kaufman, and Bowles, 2015). The security infrastructure needs to include the capability to anonymize data so that sensitive data such as social security numbers and other personal data can be hidden (Hurwitz, Kaufman, and Bowles, 2015). Otherwise, for example if patient information is moved to a big data application it might not have the sufficient level of protection to protect the private patient data (Hurwitz, Kaufman, and Bowles, 2015).

Business leaders must consistently supervise applications and allocate appropriate responsibilities between humans and machines, to take advantage of increased automation (Tarafdar, Beath, & Ross, 2017). To succeed with deploying cognitive computing applications, it is essential to have clarity regarding responsibilities between human users and the cognitive computing application (Tarafdar, Beath, & Ross, 2017). Generally, this means that the cognitive computing application will perform more routine tasks while humans take on more complex tasks (Tarafdar, Beath, & Ross, 2017). Although cognitive computing applications assumes some prior responsibilities, it also creates new tasks for humans, such as training and sustaining the cognitive computing application (Tarafdar, Beath, & Ross, 2017). Without supervision, a cognitive computing application will lose relevance over time and be able to handle less of its assigned tasks successfully, since organization's customers and products change over time and new policies and business rules might be required (Tarafdar, Beath, & Ross, 2017). Thus, organizations encounter the challenge of ongoing supervision of the cognitive computing application, to monitor the performance and regularly recalibrating it to result in correct outputs (Tarafdar, Beath, & Ross, 2017). Supervision is necessary to maintain and manage the quality, as well as to ensure that the cognitive computing application retains its accuracy and relevance (Tarafdar, Beath, & Ross, 2017).

2.5 Capabilities and opportunities with cognitive computing

No matter how smart humans might be, there are a lot of things that we simply cannot process in time to affect the outcome of a situation. Cognitive computing systems can help us overcome our limitations in many situations (Kelly and Hamm, 2013). Even if we are only in the early stages of cognitive computing, the coming decade will stretch the limits of what is possible with new software and innovations (Hurwitz, Kaufman, and Bowles, 2015). The possibilities of cognitive computing systems are almost endless (Coccoli, Maresca, and Stanganelli, 2016), its ability to handle complex tasks such as natural language processing, classification, data mining, and knowledge management enables the cognitive computing system to perform very sophisticated tasks and answer complex questions (Coccoli, Maresca, and Stanganelli, 2016; Sheth, 2016). Cognitive computing systems can easily perform tasks as extracting relationships from unstructured corpus, pattern recognition, speech to text and text to speech conversations, computer vision and machine learning (Coccoli, Maresca, and Stanganelli, 2016). Cognitive computing systems have the ability to automate tasks that previously required human interpretation (Tarafdar, Beath, and Ross, 2017). An example of this is image recognition to learn individuals' identities or queries in a call center, interpreted by natural language processing (Tarafdar, Beath, and Ross, 2017). Much of the cognitive computing

abilities is due to the large amount of data they are able to process and analyze (Coccoli, Maresca, and Stanganelli, 2016; Tarafdar, Beath, and Ross, 2017), and thus feed the machine learning algorithms they are based on (Coccoli, Maresca, and Stanganelli, 2016). Sheth (2016) describes that cognitive computing act as prosthetics for human cognition and is able to answer questions by analyzing massive amounts of data and can thereby help humans make certain decisions.

What especially makes a cognitive computing approach different according to Hurwitz, Kaufman, and Bowles (2015) is that these systems are not static, they are built to change and will continue to change dependent on the ingestion of more data and the capability to identify patterns and linkages between elements. Cognitive computing systems also enable a new level of engagement for the business leader, with intuitive interfaces between the system and the huge amount of data (Hurwitz, Kaufman, and Bowles, 2015). This could help organizations find, unknowingly in beforehand, associations and links between data elements that were not visible or present in the past (Hurwitz, Kaufman, and Bowles, 2015).

Hurwitz, Kaufman, and Bowles (2015), and Chozas, Memeti and Pllana (2017) argues that the whole interaction between the user and system will change dramatically with cognitive computing systems. The cognitive computing systems will not only help the user find new ways of answering and exploring questions about the business but will also learn gradually and provide expert assistance in many different fields of work (Hurwitz, Kaufman, and Bowles, 2015; Sheth, 2016). This learning ability according to Chen, Argentinis, and Weber (2016) is one of the greatest differentiators between cognitive and non-cognitive technologies.

Coccoli, Maresca, and Stanganelli (2016) explains, that in a scenario where users and systems interact naturally, striving spoken language, data collection and analysis becomes essential. It therefore becomes natural to collect, abstract and analyze considerable amounts of data, in order to realize intelligent systems, which are able to react to external factors in real time (Coccoli, Maresca, and Stanganelli, 2016). For example, as mentioned by Finch, Goehring, and Marshall (2017), a cognitive computing system is able to go through large amounts of structured and unstructured data to detect faster and more reliable answers within different activities. This could be, for example, cognitive computing applied to information security to detect fraud (Finch, Goehring, and Marshall, 2017).

By implementing cognitive computing, personnel can focus on more important business oriented or strategic initiatives and save hours of staff time and also reduce the actual detection and resolution time (Finch, Goehring, and Marshall, 2017). Kelly and Hamm (2013) shares the opinion on the effects of cognitive computing and also points out that cognitive computing applied to information technology promotes accelerated solution design. They argue that it supports for significantly faster, more efficient planning, development and testing of enterprise software.

Kelly and Hamm (2013) have proposed five different views of capabilities with cognitive computing, that will help us become smarter. They argue that cognitive computing offers complexity, expertise, objectivity, imagination, and sense making. As humans, we have difficulties processing large amount of information, in a short amount of time (Kelly and Hamm, 2013). With cognitive computing, we will be able to harvest insights from large amounts of

data and thereby also handle complex situations, which will not only make more accurate decisions but also better anticipate unintended consequences of actions, thereby, Kelly and Hamm (2013) express complexity. With expertise, Kelly and Hamm (2013) explains that cognitive computing systems also enables us to see the big picture and thus help us make better decisions. They point out that this is especially important when the experience is limited in a area or if we address problems that cut professional or practical domains. Cognitive computing systems also enables us to become more objective in our decision making (Kelly and Hamm, 2013). Kelly and Hamm (2013) explains that we all possess some kind of biases dependent on personal experience, background and intuition. Similarly, to objectivity, our previous experience, professional background, and training poses a difficulty to envision different settings and ranges of choice (Kelly and Hamm, 2013). They argue that cognitive computing systems will help us discover and explore new and contrarian ideas, and thereby express imagination (Kelly and Hamm, 2013). With sense making, Kelly and Hamm (2013) explains that we as humans can only make sense of so much information and that cognitive computing systems combined with sensors and analytic software will vastly extend and improve our ability to gather and process information, and thus, also enhance our sense making.

The development of cognitive computing has accelerated and just in the recent years, several tools and platforms have been created (Coccoli, Maresca, and Stanganelli, 2016). With the help of new tools of decision science, we will be able to apply new kinds of computing power to large amounts of data and thereby achieve insight into how things really work (Kelly and Hamm, 2013). The growing confidence in cognitive computing systems is another factor that Coccoli, Maresca, and Stanganelli (2016) points out. They argue that our degree of trust in cognitive computing systems in delicate areas such as economics and medicine have increased, resulting in an increased development within the field. They argue it is going in the direction of becoming more available and more affordable with great interest and substantial investments from companies, universities and research centers.

2.6 Literature review summary

In table 2.1 we have listed above mentioned challenges and opportunities presented from each author. This will provide an overview of the conducted literature.

Table 2.1: Literature review summary

Challenges	**Authors**
Capture enough relevant knowledge to teach cognitive computing systems a specific field or industry and define this in e.g. taxonomies or ontologies.	Hurwitz, Kaufman, and Bowles (2015)
Continuously enhance and refine cognitive computing systems with experience or training.	Hurwitz, Kaufman, and Bowles (2015), Tarafdar, Beath, and Ross (2017)

Supply cognitive computing systems with relevant dictionaries, thesauri, and vocabulary to enable language comprehension.	Chen, Argentinis, and Weber (2016), Kelly and Hamm (2013)
Perform learning techniques and training cognitive computing systems, with e.g. the required amount of data, and the correct format.	Tarafdar, Beath, and Ross (2017), Hurwitz, Kaufman, and Bowles (2015), Chen, Herrera, and Hwang (2018), Murtaza et al. (2016), Serban et al. (2016), Lee et al. (2016)
When using the cognitive computing system, choosing the right tools and make sure the required data is available. Identifying the right data, and in the right format.	Tarafdar, Beath, and Ross (2017)
Ensure security, e.g. to make sensitive data hidden, prevent unauthorized users to access sensitive data.	Hurwitz, Kaufman, and Bowles (2015)
When deploying cognitive computing applications, allocate appropriate responsibilities between human users and the application.	Tarafdar, Beath, and Ross (2017)
Supervise the cognitive computing system - monitor performance, recalibrate, maintain and manage the quality, and retain accuracy and relevance.	Tarafdar, Beath, and Ross (2017)
Opportunities	
Cognitive computing systems can handle very sophisticated tasks, such as natural language processing, classification, data mining, and knowledge management.	Coccoli, Maresca, and Stanganelli. (2016), Sheth (2016)
Cognitive computing's abilities is due to the large amount of data they are able to process and analyze.	Coccoli, Maresca, and Stanganelli (2016), Tarafdar, Beath, and Ross (2017)
Cognitive computing systems are not static, they are built to change.	Hurwitz, Kaufman, and Bowles (2015)
Ability for users and system to interact naturally	Coccoli, Maresca, and Stanganelli (2016)
Cognitive computing systems are able to go through large amounts of data and thereby detect faster and more reliable answers within different activities.	Finch, Goehring, and Marshall (2017)

Cognitive computing enables personnel to focus on more important business oriented or strategic initiatives.	Finch, Goehring, and Marshall (2017), Kelly and Hamm (2013)
Cognitive computing offers complexity, expertise, objectivity, imagination, and sense making	Kelly and Hamm (2013)
Cognitive computing is becoming more available and more affordable with great interest in the area	Coccoli, Maresca, and Stanganelli (2016)

3 Research Method

In the following chapter we will describe the methodology used when conducting our research. To enable as high replicability of this study as possible, we have described the following chapter as accurate and thoroughly as possible.

3.1 Background

With the definition of personal data that the GDPR introduces, there will be no difference between structured and unstructured data. This means that the same rules apply for both types of personal data. To comply with the regulation, regarding unstructured data, there are a lot of solutions. Much are dependent on the size of the organization, the number of systems, and the amount of personal data stored within the organization. The project Aigine is developing a cognitive computing system that is aimed to solve the problems with unstructured data. The goal of the cognitive computing system is to map personal data in unstructured data sources to help organization comply with the GDPR. While the usage of cognitive computing systems for various purposes in many industries have increased, we have not come across any similar solutions for mapping personal data in unstructured data sources. Because this solution is so unique, we want to investigate the challenges and opportunities of it.

3.2 Research strategy

The aim of this study is to answer the research question *What are the challenges and opportunities with using cognitive computing for mapping personal data to help comply with the GDPR?* To do this, a qualitative research strategy have been used. Recker (2013) argues that qualitative methods are strategies of empirical inquiry that investigate social and cultural phenomena within a real-life context. Since we want to examine the usage of cognitive computing for mapping personal data to help comply with the GDPR, to identify what challenges and opportunities arise, we see it fitting to investigate this in a natural setting using a qualitative method. More specifically, we decided to conduct a single-case study, which according to Recker (2013) involves intensive research on a phenomenon within its natural setting. Qualitative methods focus on text that captures records of what people have said, done, believed or experienced about a particular phenomenon, topic, or event (Recker, 2013). We decided to examine how different participants in a project, perceived using cognitive computing when mapping personal data to help comply with the GDPR, to identify what challenges and opportunities they encountered. Recker (2013) also argues that when in need to understand a phenomenon in context, qualitative methods are to prefer, instead of quantitative methods which focuses on measurement of specific aspects of a phenomenon.

3.3 Data collection

When we selected interviewees, our main objective was to conduct interviews with as many relevant people as possible in the selected project, as well as to have a spread of roles. We believe that this will help us identify a wider and more accurate result of the full project, rather than from a specific perspective. Interviewees with different competencies and backgrounds, and different roles in the project, have most likely perceived different challenges and opportunities from their own perspectives. The interviewees that were interviewed can be found below in table 3.1 It should be noted that our selection of interviewees is not simply employees at Sambruk, but rather a mix of consultants and experts from different companies who together formed the project team that executed the Aigine project. We did an expert sampling since Bhattacherjee (2012) argues that experts tend to be more familiar with the subject than non-experts, and opinions from a sample of experts are more credible than a sample that includes both experts and non-experts. The first project participant we contacted was Peter Mankenskiöld, who was the contact person of the project. He is an external consultant, working for Sambruk and was the project manager. Furthermore, he helped us get in contact with other relevant participants involved in the project. Claes-Olof Olsson is the executive official of Sambruk. Ari Juntunen was the project architect, externally hired from Elinar OY LTD, an IBM partner in Finland. Karl-Oskar Brännström is the president of Aigine AB and was the business process manager and legal advisor in the project. Hans Nyström was the solution architect and helped create the main architecture together, externally hired from IBM AB.

Table 3.1: Interviewees

Interviewee	**Role in project**	**Company**	**Duration**	**Date**
Claes-Olof Olsson	Executive official	Sambruk	28:07	2018-05-03
Peter Mankenskiöld	Project manager	Sambruk	1:09:10	2018-05-07
Ari Juntunen	Project architect	Elinar OY LTD	48:14	2018-05-14
Karl-Oskar Brännström	Business process manager and legal advisor	Aigine AB	54:20	2018-05-16
Hans Nyström	Solution architect	IBM AB	28:07	2018-05-17

3.3.1 Interview structure

The interview guide and subsequently the interviews were designed and performed as semi-structured. We did this, mainly to have a rich and open conversation, but also, as mentioned by Recker (2013), because it allows for more flexibility, both for the interviewer and the interviewee since new questions can be brought up based on how the conversation develops and thus, probe for more details or the possibility to discuss certain areas. The interviews were structured in the way that they followed the interview guide, which was designed with open questions followed, in some cases by supplementary questions to gain more insight into certain areas. Before conducting the actual interviews with the people involved in the project, we performed a pilot test. Pilot testing is according to Bhattacherjee (2012) is very important part of the research process but is often overlooked. Bhattacherjee (2012) argues that pilot testing helps to detect potential problems in the research and/or instrumentation, but also if the measurement instruments used are valid and reliable for measuring the study. After conducting the pilot test, we not only had more well formulated questions, but also an approximate time aspect of how long each interview was going to be, which helped for scheduling the interviews.

All interviews were conducted through phone calls. This was mainly because we wanted to adapt to what suited our interviewees best, but also because all the interviewees were in the final stages of their project and thus, had busy schedules. Another factor that argued for the phone interviews was the distance between us, in Lund, and the interviewees spread out across both Sweden and Finland.

Before each interview started we asked all interviewees for their consent regarding us recording the interviews to be able to transcribe and analyze afterwards, and whether they would like to be anonymized in our thesis or not. The initial questions of the interview were of general and introducing character, with questions such as background and education, followed by their role and involvement in the project. The questions after that were categorized by; the actual solution, the challenges, and the opportunities, to clearly get information based on our research question. After all questions had been answered and we were satisfied with the results, we asked if they had something to add.

3.3.2 Interview guide

Our interview guide was designed based on the conducted literature, and the research question. Question 1-2 focus on role and background. Question 3-4 focus on the solution and how it is adapted to help comply with the GDPR. Question 5 and 6 are the questions that are directly connected to our research question, where we aim to identify challenges and opportunities with using cognitive computing for mapping personal data to help comply with the GDPR. According to Kvale and Brinkmann (2009), interview guides for semi-structured interviews includes an outline of topics to be covered, with suggested questions. With this in regard, our key topics of the interview guide are the background and role, the solution, and most importantly the challenges and opportunities encountered in the project. The guide is designed with open questions, some suggested follow-up questions, with a possibility to add supplementary questions during the interview. The interview guide can be found below in table 3.2, containing every question as well as the purpose for each question.

Table 3.2: Interview guide

	Question	**Purpose**
1.	Could you please tell us more about your background and previous education?	To get to know the interviewee and find out their expertise and history.
2.	What is your role in the project?	To find out what part of the project the interviewee have been responsible or involved in.
2a.	Can you tell us about your involvement in the project?	To find out more specifically what the interviewee have been involved with and contributed with.
3.	How does your solution work?	To find out how the solution works and what functionality it has.
3a.	How have you adapted the system to the GDPR?	To find out specifically how they have designed the system to help comply with the GDPR.
4.	How will the system be implemented?	To find out how the finished system will be implemented in municipalities, to give us a better understanding of how it will be distributed.
5.	What has been challenging in the project?	To identify overall challenges in the project.
5a.	What technical challenges have you encountered?	To identify technical challenges specifically.
5b.	What challenges have you experienced with cognitive computing?	To identify challenges even more specifically, with cognitive computing. This is directly connected our research question.
6.	How will your solution change the municipalities' business?	To find out what the solution can do, leading the interviewees into talking about functionality and how it can facilitate businesses, and potentially opportunities.
6a.	What opportunities does your solution offer?	To identify what opportunities their solution offer.
6b.	What opportunities do you see with cognitive computing systems?	To identify more specifically what opportunities they see with cognitive computing systems. This is directly connected our research question.

6c.	What opportunities for future development do you see with your solution?	To identify what opportunities for future development they see with their solution. In a way, what further opportunities they see with cognitive computing systems, in addition to what they have developed in this project.
7.	Do you have anything that you would like to add?	To give the interviewees an opportunity to express anything they might think we have missed, or if they have some additional thoughts to add.

3.4 Analysis

3.4.1 Transcribing interviews

Once interview transcriptions have been made they tend to be regarded as a solid rock-bottom empirical data of an interview (Kvale & Brinkmann, 2009). However, the process of transcribing involves a number of judgements and decisions (Kvale & Brinkmann, 2009). For example, it can be hard to capture the tone of voice, body expressions, and irony, in textual format as compared to face-to-face interviews (Kvale & Brinkmann, 2009). To ensure the quality of our empirical data, we have followed some suggestions by Kvale and Brinkmann (2009). The first requirement for transcribing and interview is that it is recorded, furthermore the recording should be audible to the transcriber (Kvale & Brinkmann, 2009). When conducting the interviews, we recorded the audio with a phone and avoided background noise by performing the interview in a quiet place. We also made sure the recording was audible for transcription in our pilot test.

According to Kvale and Brinkmann (2009) there is no universal form or code for transcription of research interviews, the report should however state explicitly how the transcriptions were made. If there are several transcribers for the interviews of a single case study, it is important that they use the same procedures for typing (Kvale & Brinkmann, 2009). We split the workload of transcribing the interviews, and since there are two of us, we made sure beforehand that we were following the same procedures to be able to make cross-comparisons in the result and discussion sections. Kvale and Brinkmann (2009) argues that there are some standard choices to be made, for example, should statements be transcribed verbatim and word-by-word, including repetitions, noting expressions like “mh”s and similar? Should pauses, emphases in intonation, and emotional expressions like laughter or sighing be included? We decided to transcribe as detailed as possible word-by-word, but to exclude noises and sounds that did not contribute to the interview such as laughter, pauses, and expressions such as “hm”, “ah” and similar. Additionally, we made sure that both of us read all of the transcriptions afterwards to ensure the quality and to confirm that we had followed the same structure and procedures.

The transcriptions can be found as appendices 1-5, with belonging information such as interviewee name, role in the project, company, interview length, and the date of the interview.

The transcriptions contain the following columns; reference number, code, person, and questions and answers. To be able to refer to a specific item or answer in the transcriptions, we use a column that contains a unique reference number. The reference number is written in the format interviewNumber.answer, for example the reference number 2.16 means that it is the second interview and answer number 16. The person column contains the initials of the person talking, where "LH" and "HM" are the initials of the thesis authors.

3.4.2 Coding

Once the transcriptions were finished, we wanted to separate our interview answers into different categories of challenges and opportunities, to facilitate the presentation of the empirical findings, as well as to ease the structure of the discussion chapter. We used coding since Kvale and Brinkmann (2009) states that it provides structure and give overviews to often extensive interview texts. Kvale and Brinkmann (2009) suggests that coding can be either concept driven or data driven. Concept driven coding uses codes that have been developed in advanced by the researcher, by for example looking at material or existing literature (Kvale & Brinkmann, 2009). Data driven coding on the other hand means that the researcher starts without codes and develops them through reading of the material (Kvale & Brinkmann, 2009). Although we did not start entirely without codes or categories, we used a data driven coding approach. Since we are trying to identify challenges and opportunities with using a cognitive computing system for mapping personal data to help comply with the GDPR, we already knew beforehand that we had two major categories; challenges and opportunities. By using a data driven coding approach we were able to identify codes within these categories, where each one is a different challenge or opportunity. By using open questions in our interviews, we were able to identify the challenges and opportunities the interviewees had experienced, rather than to ask them leading questions about specific opportunities or challenges. When all of the interviews had been transcribed, we analyzed and compared the interview answers to identify challenges or opportunities with similar characteristics, making it possible to code these into according categories. To make sure we covered and coded every interesting empirical finding, both of us coded all of the transcriptions on our own, and then we compared the codings and together decided which to use. We believe that this helped improve the quality of our empirical findings. The codes are presented below in table 3.3 containing challenges, and table 3.4 containing opportunities. Additionally, each code can be found in the transcriptions (appendices 1-5) in its own column. For simplicity, we decided that every code connected to a challenge starts with a capital C, while every opportunity starts with a capital O. Additionally, we decided that the codes should consist of three letters each, simply for consistency.

Table 3.3: Codes for challenges

Challenge	Code
Cognitive computing systems in a new setting	CNS
Using the cognitive computing system	CUC
The learning process	CLP

Cognitive computing systems and the GDPR	CCG

Table 3.4: Codes for opportunities

Opportunity	Code
Increased efficiency	OIE
The learning ability	OLA
Compliance with the GDPR	OCG
Future opportunities with cognitive computing systems	OFO

3.5 Research quality

We have chosen to focus on reliability, validity, and ethics to establish the quality of our research. Bhattacherjee (2012), and Recker (2013) argues that these parameters are the top concerns in research quality, which is why they have been our top priority. We have seperated validity into internal and external aspects in order to discuss the generalizability of our research.

3.5.1 Validity

Recker (2013) explain that validity describes whether or not the collected data really measure what the researcher set out to measure. There are a number of attributes to define the quality in research. Bhattacherjee (2012) argues that the best research designs are those that can achieve high levels of internal and external validity. The internal validity, according to Bhattacherjee (2012), examines whether the observed change in a dependent variable is indeed caused by a corresponding change in hypothesized independent variable, and not by variables extraneous to the research context. The external validity, also called generalizability concerns the extent to which the observed associations can be generalized from the sample to the population or to other settings (Bhattacherjee, 2012). Yin (2009) explains, that case study research extends to the broader problem of making inferences and points out the importance in case studies is that the inference is correct, based on the interviews and the collected data for the study. To enhance the internal validity of our research we contacted our interviewees afterwards to confirm or further explain any questions or misunderstanding we encountered. Research with high levels of internal and external validity would be strong against spurious correlations, be generalizable to the population at large and inspire greater faith in the hypotheses testing (Bhattacherjee, 2012).

3.5.2 Reliability

Reliability means that, if another researcher follows the same procedures as described by a previous researcher and conduct the same case study again, the later researcher should arrive at the same findings and conclusions (Yin, 2009). Recker (2013) defines it in a similar way and says that reliability can be demonstrated with measures that provide consistently similar results. The goal of reliability is to minimize errors and biases in a study (Yin, 2009). There are various means to improve the reliability in a study, some of which we have used to improve the quality of our study. According to Yin (2009), one prerequisite for allowing another researcher to repeat an earlier case is to document the procedures necessary to conduct the study. With this in mind, we have tried to describe and document our procedures as detailed as possible in the research method chapter, while still making it easy to interpret. Bhattacherjee (2012) suggests that researchers should ask questions that respondents may know the answer to or issues that they care about to avoid ambiguous items in the result. When we selected interviewees, we tried to pick people who had relevant roles in the project and who were aware of what we were examining. Moreover, wording should be simplified to avoid being misinterpreted by respondents and generate misleading results (Bhattacherjee, 2012). We handled this when we designed the interview guide, where we aimed for easily interpreted and rather open questions. Additionally, we performed a pilot test with fellow students to make sure misinterpretations would not occur during our interviews.

3.5.3 Ethics

Throughout our study we have followed the ethical principles presented by Bhattacherjee (2012) to avoid violating the scientific principles of data collection, analysis, and interpretation. Subjects in a research project must be aware that their participation is voluntary, they should have the freedom to withdraw from the study, and they should not be harmed as a result of their participation or non-participation in the project (Bhattacherjee, 2012). Flick (2007) also highlights the importance to work on the basis of informed consent with participants, which means to inform them about the research and asking them to join the project, with a chance for the person to refuse participation. Interviews may treat sensitive topics in which it is important to protect the confidentiality of both the persons, subjects, and institutions mentioned in the interview (Kvale & Brinkmann, 2009). We informed the interviewees that their participation was voluntary and that they could exit the study without any repercussions. Moreover, we ensured that we treated the sensitive information in an appropriate manner by asking the interviewees if they wanted to be anonymized as suggested by Bhattacherjee (2012), or if we could use their names and roles in the project, as well if it was okay for us to record the interview. Kvale and Brinkmann (2009) also highlights the need for secure storage of recordings and transcripts, and of erasing the recordings when they are no longer in use. To protect the recordings and transcripts, we decided to store them locally on our personal devices and avoided cloud storage services as well as social media chat services and similar.

Researchers have ethical obligations to the scientific community on how data is analyzed and reported in their study (Bhattacherjee, 2012). Bhattacherjee (2012) provides a summarized code of conduct for the Association of Information Systems (AIS) where two items are especially important and must always be followed, namely, do not plagiarize, and do not fabricate

or falsify data, research procedures, or data analysis. This study is conducted with honesty and openness, and we have no intention of plagiarizing, fabricating, or falsifying any content.

4 Empirical findings

The empirical findings have been conducted through an analysis of interview transcriptions. The first part presents the challenges that were found, while the second part presents the opportunities. As described in the research method, the coding resulted in identifying different categories of challenges and opportunities, each of which will be described separately in their belonging part. The empirical findings contain references to the transcriptions (appendices 1-5) through the use of a reference number. To simplify the reading of this chapter, we have mentioned our interviewees by their first name only. The quotes presented in this chapter have been translated from Swedish to English except for quotes from interview number three, which was held in English. All original answers can be found in the transcriptions.

4.1 Challenges

4.1.1 Cognitive computing systems in a new setting

The challenge with applying cognitive computing systems in a new setting was mainly highlighted by Claes-Olof, but also a little by Ari. Claes-Olof pointed out the challenge with doing something that never have been done before, cognitive computing has never been tested in this specific setting (1.20). He added that this made it hard to convince stakeholders and customers. Municipalities especially, can be hard to convince to use new solutions, since they often like to see that somebody else has tested it successfully first, rather than being the first to try it (1.20). Additionally, in situations like this, municipalities often ask themselves why to use this specific solution, why IBM and so on (1.24). Another challenge related to the fact that this kind of solution has never been tested before was the cost and finding interested and dedicated stakeholders willing to finance the project (1.20). Moreover, setting up a reasonable pricing plan to make organizations and municipalities interested in acquiring it. For example, Ari said that if you are using standard IBM pricing for StoredIQ, which is one of the components in the solution, that would cost a bit over half a million euros just for the licensing, which would make it hard to convince municipalities to buy it (3.46).

Another challenge mentioned by Claes-Olof was about finding somebody who could actually build the solution and had the competence and resources to do it (1.22). Early in the project they had contact with IBM regarding their Watson concept and he said that while they were certain that the cognitive computing system could be built, however it was hard to find an IBM partner who actually could do it, since it had never been done before (1.22). He said that:

"*The challenge was to find an IBM partner who could actually create and set up this concept, and then Peter finally found a Finnish consulting company who constructed the system. For a while it was a bit like, yes it can be created but we do not know exactly how.*" (1.22)

4.1.2 Using the cognitive computing system

The challenge with using the cognitive computing systems was mentioned by four interviewees, namely Claes-Olof, Peter, Ari, and Karl-Oskar. They all had something to say about the usage of a cognitive computing system but from different perspectives, for example how it requires manual work, or how it requires human decision-making. Claes-Olof mentioned an example of when the solution had been tested in one of the municipalities and the managing director expressed disappointment because of the amount of manual work that was required, and that they had thought it would work automatically (1.14). Claes-Olof added that manual work is in fact required so that the cognitive computing system can learn and become more effective (1.14).

Karl-Oskar argued that the problem with unstructured data is that we do not know what it contains (4.41). Moreover, he said that this means that every data source could contain presumptive personal data, making every document a presumptive data source that requires legal basis to be kept, and it has to be listed in the records of processing activities (4.41). Peter also mentioned this problem, and how the combination of various information can form presumptive personal data, making it important to be able to identify all such information (2.14). Peter explained that such combinations could be for example unique winners of competitions, or something that results in getting an award, or a title perhaps given from education (2.14). Karl-Oskar continued with highlighting the amount of work required to go through each document containing presumptive personal data, since authorities can have thousands of databases containing millions of documents (4.41). Furthermore, they have come to the conclusion that human review is required, Karl-Oskar said:

"*Because human review is required. We cannot trust that the system is right, so a human review is required and especially legal decision-making is required that is connected to, do we have the right to keep this personal data? And if we have the right to keep it, with what legal basis do we have the right to keep it?*" (4.33)

Another challenge that is connected to using the cognitive computing system was brought up by Ari. He explained that every time something is added to the cognitive computing system, for example a new capability, or a new thing take out from documents, a sufficient volume of training data needs to be obtained for that, in this case from the municipalities (3.34). This means that the municipality need to be able to provide training data. If this is not done, the cognitive computing system will not be able to perform the new capability (3.34).

4.1.3 The learning process

The challenges regarding the learning process of the cognitive computing system was mentioned by all of the interviewees. It is no longer enough to only look for first name, last name, address or personal number, because with the GDPR, a combination of information can form personal data (1.16). Claes-Olof mentioned the challenge of identifying such combinations, and how important it is with the continuous and cooperative learning of the cognitive computing system, to acquire better knowledge and to be more efficient (1.16). Peter argued about

this in a similar way and said that it is rather easy to create searching algorithms for identifying personal numbers, e-mail addresses and so on (2.14). However, when it comes to for example awards in a competition, or titles from education, and the combination of such information that can create presumptive personal data, it is very hard to learn the cognitive computing system because you cannot add descriptions for every single competition (2.14). Instead he argued that the system need to learn to understand the combinations, to be able to find the presumptive personal data that needs to be identified (2.14).

When talking about identifying personal data correctly, both Peter and Ari said that the system will find an amount of false-positives (2.24; 3.38), meaning it will identify something it believe is personal data but it is not. The system can get confused by things such as the fact that a first name can be a surname, but also that two names in a row can be a first name and a surname, meaning it is one full name instead of two separate names (2.24). Ari explained that when the system is looking for people and names and so on, they use something he refers to as common words, which could be for example all lower-case words, words that start with a capital letter, or all caps (3.38). Furthermore, using only this type of analysis will generate a lot of false-positives, and it is even harder with multiple languages (3.38). He stated that the larger dictionary you have, the more false-positives you will get, because a name in one language can be a common word in another language (3.38). An example he gave was:

"*But the challenge is that, sales for example, it is an English word but also a last name in UK, Mr. Sales. There is a person, I am not sure about the name, but it is on the name list, which means that a lot of English documents get annotated wrongly. Because every time there is a word sales, it is last name.*" (3.38)

Regarding language, Hans argued that the system needs to be fully supported language-wise to be able to utilize the cognitive services, otherwise it will just be pattern recognition rather than cognitive (5.21).

Another challenge concerning the learning process is the amount of training. Ari described that for every new capability that is added in the system, sufficient volume of training data needs to be added to train the system, otherwise it will not have a clue about what to do (3.34). Claes-Olof also explained that in order for the cognitive computing system to be efficient it needs to learn (1.14). A similar aspect is pointed out by Karl-Oskar, who said that it requires training to understand context (4.37). He emphasized that it requires proper training and gives an example that a three week training sessions will not be enough, instead you rather want 5-10, maybe even 20 million evaluations (4.51). When discussing decision-making, Karl-Oskar argued that it is not enough that the cognitive computing system is as good as humans, it needs to be incredibly much better for us to be able to trust it (4.26).

4.1.4 Cognitive computing systems and the GDPR

While the whole project was about complying with the GDPR, this section presents the challenges more directly connected to GDPR compliance, such as architecture, and keeping records of processing activities. This was mentioned by Claes-Olof, Peter, and Karl-Oskar in the interviews. The main challenge addressed was concerning the architecture. Peter and Karl-Oskar both explained that most traditional AI-solutions is cloud based and that it does not work legally with the GDPR, evaluating personal data and cloud-based solutions does not go well together (2.14; 4.20). You cannot send personal data to the cloud to evaluate it (2.14). Karl-Oskar further explained that you cannot send data from municipalities to the cloud for evaluation since you cannot know what they contain, meaning they presumptively can contain personal data which according to the GDPR is not allowed to be handed to a third party (4.26). Data cannot leave the municipality (4.26), making it a big technical challenge of how to solve this locally (2.22). The challenge according to Peter, was finding the components to assemble this kind of system that is legally correct, and which makes it possible to perform the mapping of data while still keeping all the data on-premises (2.14).

Another challenge that was mentioned were connected to the records of processing activities. You need to list every data source that contains personal data in the records of processing activities, and while a database is considered a singular data source, Peter explained that when it comes to unstructured data every file or document can be a data source that has to be listed (2.18). Furthermore, he argued that when it comes to unstructured data, the searching, evaluating, decision-making, and listing is such a comprehensive task that it is almost impossible (2.18). Karl-Oskar talked about this challenge as well and said that since we cannot know what unstructured data contain, every document is presumptively a data source containing personal data, with different legal basis for each document (4.41). Moreover, he pointed to the fact that the GDPR requires keeping updated records of processing activities containing every data source, with information about what kind of personal data is kept, and with what legal basis (4.41). Furthermore, Karl-Oskar emphasized on the amount of work required to establish records of processing activities considering authorities can have a couple of thousands of databases, and millions of documents where each document presumptively is a data source (4.41). Claes-Olof also talked about the challenge with large amounts of data, that it can be hard to find and identify all data sources since there are so many data sources, and that every user has their own repositories and have emails saved in different ways and so on (1.16).

4.2 Opportunities

4.2.1 Increased efficiency

All interviewees except Hans emphasizes that cognitive computing systems offers efficiency, however, they all present different forms of efficiency. Peter explained that there are similar solutions for personal data in unstructured data sources, however, those solutions only enables a visualization of your situation, and, what they actually needed was some form of decision support to be able to go through all documents that could contain personal data (2.14). The cognitive computing system will help find data but also help create more effective analysis of

all structured and unstructured data which will help if they need to protect but also if they need to find data (1.26). This information classification and control of the data will help create better process descriptions and thus enable much greater ability to streamline the whole organization (2.26). This structure that the solution enables, opens up according to Ari also for extremely powerful possibilities to make traditional business analytics much more accurate and much more powerful (3.49). Ari compared the solution with manual work and pointed out the large amount of time difference between both solutions, he argued that this is the ultimate goal of cognitive computing, that it can understand things in a similar fashion to humans and thus replace highly educated and skilled individuals to do more important tasks (3.49). He was also clear to point out that the largest cost for every municipality is the time spent on people trying to find data (3.53). Karl-Oskar agreed on the fact that the new solution will require less work by professionals and highlights the fact that the solution do not need legally qualified personnel (4.49), he says:

"It is still a gigantic challenge but then with it connected, we no longer need legally qualified personnel, we do not need lawyers to do it because we both have the system that actually gives concrete suggestions on the legal support but also knowledge databases which contextually train and leads the personnel through this work." (4.49)

The solution will, according to Peter provide much help, it will enhance and accelerate the process of finding personal data, it will read through documents much faster than a human and to some extent actually understand what it reads (2.14). The fact that the solution will find duplicate or even multiple stored data is Claes-Olof certain of (1.26). He explained that multiple stored data in some cases is unnecessary and that this solution enables structuring and thereby also the possibility of making the whole infrastructure much more effective (1.26).

4.2.2 The learning ability

The learning ability of cognitive computing systems is mainly mentioned and discussed by Claes-Olof, Peter and Karl-Oskar. Peter explained that after every time the cognitive computing system runs, it returns presumptive personal data (2.14). However, the next time it runs, the results from the cognitive computing system could be much more extensive, based on inputs from reviewers who are using the system (2.14). This means as mentioned by Karl-Oskar, Peter and Claes-Olof that the cognitive abilities continuously improve throughout the usage of the cognitive computing system as users gives feedback and evaluate the findings (4.28; 1.14; 2.14). Karl-Oskar gave an example of this, that in the beginning the cognitive computing system will not be especially good and have an accuracy of maybe 80% but the continuous human reviewing and feedback will function as a kind of training session (4.28).

Karl-Oskar explained that the cognitive computing system is used for filtering out personal data and reducing the workload initially and thereafter gradually speeding up the process of the human review (4.26). He further explained that the cognitive computing system cannot be trusted to be right initially, but their hopes are that the system will over time be much better than the human ability (4.26). He compared the cognitive computing system as a strainer

which collects documents containing personal data, and as the cognitive computing system is being used and learned the strainer will collect more and more (4.28).

Peter pointed out that even if all users have similar operations, everyone who is using the cognitive computing system will contribute to develop its learning and thus improve its efficiency, which means, that all users will be able to access the contributions to the cognitive computing system (2.2). This is nothing specific for any line of business when it comes to the learning of the cognitive computing system (2.2). The learning process works, according to Peter, by uploading to the central unit what have been learned locally, and then processed by the cognitive computing system to improve the next outcome of the findings (2.14). For this to work properly, the user interface must be easy to use and intuitive (2.14).

This architecture with the combined learning of the cognitive computing system is something that Karl-Oskar believe could be used in many different areas (4.51). He argued that it could be used in many different industries where you have large quantities of data but also large growing quantities of data, he mentioned for example the bank and finance industry as suitable for this kind of cognitive computing system (4.51).

4.2.3 Compliance with the GDPR

This section will present the opportunities of using a cognitive computing system to help comply with the GDPR. According to Peter, most organizations have not focused on mapping their data, even if that has been one of the top priorities before the GDPR becomes effective (2.16). Instead they have focused on how they could handle the structured data and change the processes of working with the structured data after the 25th of May, when the GDPR become effective (2.16). He believes that many organizations have turned their backs on the unstructured data, mainly because it is to complicated to manage (2.16).

Karl-Oskar explained that early in the project they established the need for a metadata warehouse to use as a record of processing activities (4.41). All documents which have been screened gets classified by which legal basis they have and for how long they can save that specific data (4.41; 1.16; 2.20). The metadata warehouse created is connected to each document and thus the record of processing activities, which is a requirement of the GDPR (4.41). This will help not only with controlling which documents have already been screened (2.20), but also with requirements that the GDPR advocates (4.41; 2.20), for example the right to be forgotten (4.41). However, it can also help with what kind of legal basis they have to keep the personal data (2.20), Peter pointed out the example of a concern notification from the social services in which the person in question is not able to access that information (2.20). The record of processing activities can easily capture where and on which places a specific person's information is stored, and thus is able to present it if someone asks (4.41). With the record of processing activities, the solution creates a greater value for the customer by more easily comply with the GDPR (4.41).

The annotation code that is needed for the cognitive learning is usually a component located in the cloud (4.26). However, for complying with the GDPR, the annotation code has been located locally together with the Aigine installation (4.26). Karl-Oskar explained that for the system to comply with the GDPR they are not able to send any personal data, neither to the cloud service or any other third party (4.26). He further explained that for the system to be compliant with the GDPR the annotation code is transformed into machine code, which is then sent to the cognitive abilities located in the cloud (4.46). The machine code is unreadable, it is not possible to extract any information or personal data from this code (4.46). Karl-Oskar explained that they have benchmarked the machine code with people within the industry and gotten an acknowledgement that the machine code is non-recreatable (4.26). Karl-Oskar explained the machine code as:

"You change, for example the words to digits which tell that here we have a noun in genitive and we have a capital letter, and a word with four letters and with a capital letter, that is the type of code we send." (4.26)

This system helps organizations meet the requirements of the GDPR, which Karl-Oskar explained, that they believe would be impossible without this kind of tool or system, in which he argued Aigine is unique (4.49).

4.2.4 Future opportunities with cognitive computing systems

The future opportunities with cognitive computing systems are several and discussed by all interviewees. Peter and Karl-Oskar pointed out that the system could be applied to new forms of data (2.20; 4.53). They both mentioned the ability of applying the system to structured data in databases, with screening of tables and contents so that they are free from any personal data (2.20; 4.53). Peter said:

"And I am totally convinced that I could do the same type of screening of a table in a database as I do with a filesystem." (2.20)

Karl-Oskar explained, that there is a problem with screening personal data in structured data sources, but actually unaddressed today (4.53). He further explained that they have chosen not to address structured data, simply because it is already a competitive market (4.53).

Both Ari and Claes-Olof addressed the ability of the definitions (1.28; 3.51), Ari explained that they are creating more and more capable common definitions which now is focused on privacy data (3.51). However, he believes that they will be able create a system that will be able to abbreviate from any type of document (3.51). The infrastructure the system enables is another opportunity that Hans mentioned (5.32). He argued that it will for sure enable us to build a lot of timesaving and legally secure solutions (5.32).

4.3 Summary of empirical findings

Four categories of cognitive computing system challenges were identified from the conducted empirical data. First, cognitive computing systems in a new setting concern challenges such as applying cognitive computing systems in a setting that never have been done before, convincing stakeholders, cost and financing, and finding competence and resources to build the solution. Second, using the cognitive computing system concern challenges such as that it requires much manual work, human review, decision-making, and municipalities need to be able to provide training data. Third, the learning process concern challenges such as the cognitive computing system need to be able to identify personal data, and combinations of information that can form personal data. Additionally, the problem with identifying many false-positives because of dictionaries and language, and that the cognitive computing system require proper training and a sufficient volume of training data. Fourth, cognitive computing systems and the GDPR concern challenges directly connected to GDPR compliance, such as architecture, and identifying data sources to list in the records of processing activities with legal basis.

Additionally, four categories of cognitive computing system opportunities were identified from the conducted empirical data. First, increased efficiency concern opportunities such as identifying, visualizing, and make decisions on presumptive data, decrease the amount of human work required, and enhance and accelerate the processes of finding personal data. Second, the learning ability concern opportunities such as continuous improvement, accuracy, and efficiency, and collaborative learning. Third, compliance with the GDPR concerns opportunities of using a cognitive computing system to help comply with the GDPR and how the solution is GDPR compatible, with for example keeping records of processing activities, and using annotation code. Fourth, future opportunities with cognitive computing systems concern future opportunities such as applying the cognitive computing system to other forms of data such as structured and being able to abbreviate from any type of document.

5 Discussion

This chapter will present our discussion where we will discuss the empirical findings against conducted literature as well as our own reflections and thoughts. The discussion is based on the same structure as in the empirical findings. However, "the learning process" and "the learning ability" will be discussed i the same section. Similarly, "the cognitive computing systems and the GDPR" will be discussed together with "compliance with the GDPR".

5.1 Cognitive computing in a new setting

The challenge with applying cognitive computing systems in a new setting in the Aigine project is highlighted by both Claes-Olof and Ari. To clarify, the new setting in this case is mapping personal data from unstructured data sources to help comply with the GDPR. Our empirical findings presented that because it has never been done before, firstly, it was hard to convince stakeholders (1.20; 3.46), and municipalities seldom want to be the first to try something (1.20; 1.24). Additionally, there was a challenge of cost (1.20; 3.46), and financing (1.20). It was however pointed out that although it had never been done before, they were certain it could be done even if it was a bit troubling to find partners that had the required resources and competence (1.22). In a way, we believe that this demonstrates the potential and capability of cognitive computing. Clearly, it contains many capabilities that can be applied to various contexts. It is not necessarily the technology that lacks in capability, but it can be argued that there is a lack of competence and resources amongst developers or architects in the industry of cognitive computing. This is probably due to the fact that the field is rather young, where it has been argued that the term was first coined in 1995 (Valiant, 1995), and one of the most well-known breakthroughs in the field happened only a few years ago, when IBM Watson competed in Jeopardy! in 2011(Sheth, 2016).

Organizational challenges such as convincing stakeholders, costs, partners and similar, is not something we have focused on in our thesis, since we believe that the empirical findings about these aspects is probably applicable for many other kinds of systems as well. However, even though our empirical findings concerning these aspects might be rather generic, they do still shed light to the challenges that can be expected when developing and implementing a cognitive computing system like Aigine. These aspects are still interesting to highlight due to the fact that a project like this has never been done before. Cognitive computing is an emerging area that shows great potential (Demirkan, Earley, & Harmon, 2017), and it is interesting to see it be applied to a highly relevant problem that requires much attention. As discussed by Finch, Goehring, and Marshall (2017), a cognitive computing system is able to go through large amounts of both structured and unstructured data to detect faster and more reliable answers within different activities. We believe that it is a very interesting solution to use cognitive computing to map personal data to help comply with the GDPR and get better control of unstructured data, which we have pointed out is a significant and increasingly growing problem.

5.2 Using cognitive computing

Claes-Olof, Peter, Ari, and Karl-Oskar all had something to say about the challenge of using the cognitive computing system Aigine, when they talked about e.g. what was required, from their own perspectives. Claes-Olof told us that the cognitive computing system had been tested and apparently it requires a lot of manual work which led to disappointment from users, who had thought that it would work automatically (1.14). Claes-Olof himself was well-aware that manual work is required so that the system can learn and become more effective (1.14). This is in accordance with what was found in the conducted literature, that cognitive computing systems learn and improve from every interaction and outcome (Demirkan, Earley, & Harmon, 2017; Mertens, 2017, Sheth, 2016). Additionally, Hurwitz, Kaufman, and Bowles (2015) mean that cognitive computing systems can learn from experienced practitioners and over time be fed newer data which expands the depth of knowledge. We believe it is important that the municipalities understand that the manual work required is not only for the cognitive computing system to work for the moment and take correct decisions, but rather to improve the system, make it more effective over time. This will in the end result in less manual work required from human users. This is similar to what was found in literature by Demirkan, Earley, and Harmon (2017) who means that cognitive computing has great potential to increase efficiency, reduce costs, and improve outcomes by analyzing large amounts of data and performing routine tasks.

Karl-Oskar argued that the challenge with unstructured data is that they do not know what it contains (4.41), and moreover him and Peter discussed how this gets problematic since it means that every document can contain presumptive personal data (4.41; 2.14). It is evident that the Aigine project had a clear GDPR focus, and even though cognitive computing offer great capabilities, there was a concern from project participants such as Karl-Oskar and Peter to identify all the personal data that existed in the unstructured data sources, and especially the combinations of various information that can be defined as personal data (2.14). As we see it, the concern mainly existed because of the new tricky definition of personal data that the GDPR states, and the challenge of going through every single unstructured data source to identify personal data, rather than not trusting the cognitive computing system. We think that, what they really found the most challenging about this was to get their cognitive computing system good enough to be able to identify all the presumptive personal data, and combinations of information that forms presumptive personal data. Additionally, following this Karl-Oskar highlighted the amount of work required to go through each document containing presumptive personal data (4.41). As presented in the empirical findings, the cognitive computing system will not be especially good in the beginning and have an accuracy of maybe 80% (4.28). Thus, we can imagine that it will in fact require a fair amount of work when using the cognitive computing system, at least in the beginning. It might also be challenging to stay sharp and make good decisions for every presumptive personal data the user is faced, especially if they are similar to one another and there are huge amounts of them. The goal is however that the cognitive computing system will be able to decrease the amount of time required from human users significantly as discussed by Peter, who said that it can read through documents much faster than humans and that it will save time, especially as it gets more efficient (2.14). Interestingly enough, previous literature has mentioned the goal of cognitive computing in a similar way, that it is to increase productivity and creativity such as decision-making, innovation, connectivity, and augmentation, of both individuals and organizations (Demirkan, Spohrer, & Welser, 2016). Furthermore, Demirkan, Earley, and Harmon (2017) even argues that in some

cases cognitive computing systems do a better and much faster job than humans at for example recognizing patterns, performing rule-based analysis on very large amounts of data, and solving both structured and unstructured problems.

An empirical finding that we found interesting was as Karl-Oskar expressed, that they cannot trust the cognitive computing system to be absolutely correct, and that human review is required, especially for the legal decision-making when they have to decide if they have the right to keep the personal data and if so, with what legal basis (4.41). When it comes to identifying personal data, we believe that it is a good choice to include human review as a last step in the process, especially considering the new definition of personal data that the GDPR introduce, meaning that combinations of information can form personal data. However, as presented in the literature the cognitive computing system will learn over time which will make it more accurate (Demirkan, Earley, & Harmon, 2017; Mertens, 2017, Sheth, 2016). We believe that while the cognitive computing system might finally be good enough to identify all personal data, the human review should still exist. It should also be taken into account that cognitive computing systems does not necessarily have to take decisions themselves, Earley (2015) means that decisions that have typically been made by humans can now be managed by cognitive computing systems, or with the assistance of cognitive computing systems. With this in mind, it can be argued that the role of an assistant might fit them better than to finalize the decision, in this specific context of taking legal decisions to comply with a new regulation. The human review will ensure the safety of the personal data as an extra precaution, while ensuring that organizations comply with the GDPR. Compared to a solution without a cognitive computing system, human users will still be relieved of a huge amount of workload. It is also our opinion that when it comes to legal decision-making you can never be too cautious. Imagine if something went wrong and there is only a cognitive computing system to blame, when instead you instead could have appointed human users to ensure that the decision was taken correctly.

Another challenge that is connected to using the cognitive computing system is brought up by Ari. Every time for example a new capability is added to the cognitive computing system, the municipality need to provide a sufficient amount of training data (3.34). This affects the usage of the cognitive computing system in the way that if organizations want to add functionality to the system they need to provide training data, otherwise the cognitive computing system will not be able to perform (3.34). Connections to this can be found in the literature. Murtaza et al. (2016) strengthens what Ari said, and says that a cognitive computing system such as IBM Watson can show its full capability only through sufficient training. Additionally, Tarafdar, Beath, and Ross (2017) argues that business leaders must choose the right tools and make sure that the required data is available, and that it is challenging to identify the right data and get it into the right format. Providing sufficient and the right data is something that is known from previous research and while it might be challenging, we argue that it is a one-time thing worth working hard for to get a capability working, and then from there the ability can improve as it learns from each interaction.

5.3 Increased efficiency

The empirical findings of the increased efficiency of the Aigine project, were mostly focused on three different areas; accuracy, saving time, and minimizing cost. Peter, Claes-Olof, Ari, and Karl-Oskar did all present different views on how the Aigine solution would provide organizations with increased efficiency, however they all present rather different views on efficiency. Ari compared the Aigine solution with doing it manually and pointed out the large amount of time difference between both solution, he argued that this is the ultimate goal of cognitive computing, that it can understand things in a similar fashion to humans and thus replace highly skilled individuals to do more important tasks (3.49). Similarly, Karl-Oskar explained that the solution will require less work by professionals or even legally qualified personnel (4.49). This is in accordance to what Finch, Goehring, and Marshall (2017) explained, that by implementing cognitive computing, personnel can focus on more important business oriented or strategic initiatives and thereby save hours of staff time and also reduce the actual detection and resolution time. Cognitive computing systems have the ability to automate tasks that previously required human interpretation (Tarafdar, Beath, and Ross, 2017). By reducing personnel, they will also save costs, as Ari explained that the biggest cost will be spent on people trying to find data (3.53), if they were to do it manually. However, we believe that cognitive computing systems are not the cheapest systems around, neither to buy, develop or to use as a service. We are not sure about the cost of the Aigine solution, but we know that for municipalities in Sweden the cost will be based on the number of people working for the municipality (4.45). However, in contrast to similar solutions such as Elinar AI-miner which costs 5 Swedish crowns per scanned document (4.43), or that a license to IBM StoredIQ costs more than half a million euros (3.46). We believe that the pricing of Aigine is more suitable for a system like this, and this will allow for unlimited usage for the municipalities, rather than paying for each document scanned, which will make it substantially cheaper. Regardless, our growing confidence in machines in delicate areas such as economics and medicine have only increased (Coccoli, Maresca, and Stanganelli, 2016), and thereby, Coccoli, Maresca, and Stanganelli (2016) argue that it is going in the direction of becoming more available and more affordable with great interest and substantial investments from companies, universities and research centers. We would argue in the same manner as Coccoli, Maresca, and Stanganelli (2016), we believe that organizations will see the potential of cognitive computing systems and that the development will increasingly advance, with more available and affordable solutions.

When it comes to the accuracy, Peter explained that the information classification and the control of data will help create better process descriptions and thereby enable much greater ability to streamline the whole organization (2.26). He also explains that the solution will enhance and accelerate the processes of finding data by reading through documents much faster than a human and to some extent actually understand what it reads (2.14). This also means that the accuracy he discusses also supports the efficiency of saving time. Claes-Olof agreed with Peter, that the solution will help find data but argues more towards the fact that the cognitive computing system will help create more effective analysis of the data and thereby enhance the processes of both protecting data and finding data (1.26). Claes-Olof was also keen on pointing of the fact the he is sure the solution will find duplicates of or even multiple stored data, which in some cases are unnecessary. However, it allows for structuring the data and thereby also the possibility of making the whole infrastructure much more effective (1.26). These factors they present is some of the cognitive abilities which cognitive computing

systems enables, for example, the ability to process and analyze huge amounts of data (Coccoli, Maresca, Stanganelli, 2016; Tarafdar, Beath, and Ross, 2017; Finch, Goehring, and Marshall, 2017). The solution accuracy Peter and Claes-Olof presented is something that Kelly and Hamm (2013) also discuss, they argue that cognitive computing applied to information technology promotes significantly faster, more efficient planning. Another interesting factor that enables opportunities for cognitive computing that only Ari mentioned, is the structure that the solution enables (3.49). He argued that it opens up the possibilities of making traditional business analytics much more accurate and powerful (3.49). Having all data, both structured and unstructured clearly accessible, enables of course business analytics to be much more accurate and powerful which will also help make more and better decisions. Even if no previous research we have investigated for this study mentions this, we believe this opportunity is pretty obvious, the more data accessible to an analytics software the more accurate and precise the outcome is going to be.

5.4 The learning process and learning ability with cognitive computing

As presented in the empirical findings, the challenges regarding the learning process of the cognitive computing system were mentioned by all of the interviewees, and the opportunities with the learning ability were discussed by Claes-Olof, Peter and Karl-Oskar. The learning aspect of cognitive computing was one of the most interesting areas in our thesis, since it was found to be both challenging to train the cognitive computing system and get it to learn what was wished for, but also it opens up for opportunities since the cognitive computing system is able to learn over time, get more efficient, which results in more accurate usage and thereby more efficient work. Our empirical findings showed that the challenges were mostly focused on training the cognitive computing system before starting to use it and getting it to identify personal data accurately. The opportunities however, had more focus on the continuous learning, improving over time, and finally resulting in improved efficiency. These were all things that were covered and discussed in the conducted literature.

One challenge that was identified in the empirical findings was related to the new definition of personal data that the GDPR introduces. Claes-Olof explained that it will not be enough to only look for first name, last name, address or personal number, because with the GDPR a combination of information can form personal data (1.16). Both Claes-Olof and Peter stressed the importance and challenge of teaching cognitive computing system to be able to understand such combinations, to be able to find presumptive personal data (1.16; 2.14). Peter also added that it is not possible to for example add descriptions for each item necessary, because there are so many (2.14). The specific issue about finding combinations of information that can form personal data is not something found in literature since a study like this, to our knowledge, have never been conducted before.

Tarafdar, Beath, and Ross (2017) argues that ensuring that the required amount of data is available is challenging, since cognitive computing applications need substantial amount of test data for training. They specifically highlight the challenges of getting the right data, as well as getting the data into the right format. In the Aigine project, both Ari and Karl-Oskar

emphasized that the cognitive computing system require proper and sufficient amount of training (3.34; 4.51). Ari focused on sufficient training data for every new capability that is added (3.34), much likely because he was the project architect and participated in the development of Aigine. Karl-Oskar on the other hand talked more generally about training but gave an example that a three-week training session will not be enough, instead you rather want 5-10, maybe even 20 million evaluations (4.51). From this we can conclude that this challenge is presented both in the literature and in the investigated project. We have not gone into more depth about how to solve it or similar, but rather to identify the challenge.

The project participants were however positive to the learning ability of the cognitive computing system. Although Karl-Oskar said that the cognitive computing system might only have around 80% accuracy in the beginning he said that the continuous human reviewing and feedback will function as a kind of training session (4.28). Hurwitz, Kaufman, and Bowles (2015) also says that cognitive computing systems can learn from experienced practitioners and over time be fed newer data which expands the depth of knowledge. This is interesting in the aspect that it is not only about traditional training to get a capability to function well, but rather, organizations will be able to capture expertise from an experienced practitioner to a cognitive computing system, which can possibly take the role as an assistant and provide more support than a human could do. Peter and Claes-Olof also said that the cognitive abilities continuously improve throughout the usage of the cognitive computing system as users gives feedback and evaluate the findings (2.14; 1.14). The continuous learning and improvement in accuracy is according to us one of the top capabilities of cognitive computing, and is mentioned by for example Demirkan, Earley, and Harmon (2017), Mertens (2017), and Sheth (2016) in the literature. Moreover, this capability makes the system more accurate and could help organizations to cope better with the requirements in the GDPR. What we found especially interesting with the learning ability of Aigine was that every user of the system will contribute to the learning process to improve its efficiency together, and then all the users will be able to access the new improvements, as pointed out by Peter (2.2). He described that the learning process works by uploading to the central unit what have been learned locally, and then processed by the cognitive abilities to improve the next outcome of the findings (2.14). We have not found any literature that mentions the architecture of collaborative learning in the way that Aigine does. We do however find it interesting to discuss how this works if Aigine were to be applied in other industries, something different from municipalities, in the aspect of for example business-specific vocabularies and such. Especially since Kelly and Hamm (2013) argues that it is challenging to teach systems structure, vocabulary of spoken and written languages, as well as business-specific words and concepts (Kelly & Hamm, 2013). Since the cognitive computing system learns through annotation code instead of actual words it might work well anyway, but it would be interesting to investigate and measure how collaborative learning works across various industries and organizations, maybe different countries as well. In the specific context of personal data to comply with the GDPR, and with using annotation code for learning, it would probably work rather well. If the cognitive computing system were to be applied to identify other information than personal data or in other contexts as proposed by some project participants, for example Karl-Oskar who says Aigine could be used in many different areas and industries where you have large quantities of data (4.51), we cannot be sure how it would work without investigating. This is something to look into for future researchers, and something that would have to be measured practically when the cognitive computing system is implemented across a variety of industries and contexts. If collaborative learning similar to the

architecture used in the Aigine project would work well across various industries and contexts, it would open up for many business opportunities.

The learning ability of the cognitive computing system opens up opportunities to reduce workload by filtering out personal data, and thereafter gradually speeding up the process of the human review (4.26). Furthermore, while the cognitive computing system might not be accurate in the beginning, Karl-Oskar compared it as a strainer that will collect, or identify, more and more personal data (4.28). This ability has been presented in the literature as well, that cognitive computing systems can learn and over time be fed newer data which expands the depth of knowledge (Hurwitz, Kaufman, and Bowles, 2015). Claes-Olof is one of the interviewees who really emphasizes how important it is with the continuous and cooperative learning of the cognitive computing system, to acquire better knowledge and to be more efficient (1.16). Something that is interesting to highlight is that Karl-Oskar argues that it is not enough that the cognitive computing system is as good as humans, it needs to be incredibly much better for us to be able to trust it (4.26). Here we would like to highlight what was mentioned by Tarafdar, Beath, and Ross (2017), that to succeed with deploying cognitive computing applications, it is essential to have clarity regarding responsibilities between human users and the cognitive application. Furthermore, this usually means that the cognitive application will perform more routine tasks while humans take on more complex tasks (Tarafdar, Beath, & Ross, 2017). With this in mind, we believe that it is safer to let the cognitive computing system act as an assistant than to take decisions in the future, especially since we are talking about personal data, protecting the integrity of citizens, and complying with the GDPR.

Another challenge that were mentioned by the interviewees concerned languages and dictionaries. When talking about identifying personal data correctly, both Peter and Ari said that the system will find an amount of false-positives (2.24; 3.38), meaning it will identify something it believe is personal data but it is not. Peter for example explained that the system can get confused by things such as the fact that a first name can be a last name, but also that two names in a row can be a first name and a last name that forms a full name (2.24). Ari went into specifics a bit more and explained that when the system is looking for people and names and so on, and stated that the larger dictionary you have, the more false-positives you will get, because a name in one language can be a common word in another language (3.38). Hans as well, argued briefly that the system needs to be fully supported language-wise to be able to utilize the cognitive services, otherwise it will just be pattern recognition rather than cognitive (5.21). The conducted literature also mentioned the challenges with language and dictionaries. Chen, Argentinis, and Weber (2016) argued that to enable language comprehension, cognitive computing systems must be supplied with relevant dictionaries and thesauri. According to Kelly and Hamm (2013), the challenge is to teach systems structure, vocabulary of spoken and written languages, as well as business-specific words and concepts, especially since the special language in a specific domain of knowledge or expertise can consist of chemical symbols, medical images, or embedded in legal terms. While we did found literature highlighting the challenge of language and supporting relevant dictionaries, we did not find something strengthening Ari's statement that the larger dictionary you have, the more false-positives you will get. It can also be argued that false-positives are not necessarily only bad. Obviously, it requires more amount of work from human users, but it also means that there probably is a better chance of finding all personal data. Moreover, as the cognitive computing system learns over time thanks to the human reviews, the amount of work required from human users will decrease, even if the dictionary is large. With this argument, it can be argued that it is worth

having a large dictionary in the beginning because it could potentially identify more, or at least as much personal data, and then over time the cognitive computing system will require less amount of work from humans either way.

5.5 Cognitive computing to help comply with the GDPR

The main challenge for the Aigine solution was mainly concerning that architecture. Since most similar solutions are cloud based, complying with the requirements of the GDPR and evaluating personal data with a cloud-based system is difficult (2.14; 4.20). The problem is that the data, in this case from municipalities cannot leave the municipality (2.14). You cannot to send personal data to the cloud, to evaluate it (2.14). Or as for the Aigine system, any data of unstructured form, because you do not know in beforehand if it contains any personal data (4.26). This is in correlation with the requirement presented in article 6f in the GDPR which states that processing shall be lawful only if the processing is necessary for the legitimate interests pursued by the controller or by a third party, except where such interest is overridden by the interest, rights and freedoms of the data subject (REGULATION (EU) 2016/679, Article 6). That means, that a big challenge was to create a solution that is legally correct and makes it possible to perform the mapping of data while still keeping all data on-premises (2.14).

For the cognitive computing system to gradually learn and perform better over time, it continuously needs the annotation code form each scanned document. Karl-Oskar explained that this usually is a component located in the cloud (4.26). However, as previously mentioned, no personal data can leave the municipality in accordance with article 6f in the GDPR, as mentioned before. This meant, for complying with the GDPR, the annotation code has been located locally together with the Aigine installation (4.26). The solution they came up with, to have a GDPR-compliant cognitive computing system, was to transform the annotation code into machine code, which is then sent to the cognitive abilities located in the cloud (4.46). Karl-Oskar explained that the machine code is unreadable, that it is not possible to extract any information or personal data from this code (4.46). To ensure the effectiveness of the machine code, they have benchmarked the machine code with people within the industry and gotten the acknowledgement that the machine code is non-recreatable (4.26). Neither of these challenges or solutions have been in presented in the conducted literature. We believe that this mainly have to do with the fact that the Aigine solution is a rather unique project and, to our knowledge, not many similar solutions have been studied. But also, because the GDPR becomes effective the 25th of May 2018, and before that there were no regulations regarding disclosure of personal data to a cloud service or any other third party of interest, which mean that similar solutions are either under development or have just been finished. We believe the solution with converting the annotation code to machine code is not something especially new or something unique. However, for this kind of cognitive computing system with its abilities located in the cloud, it is a simple, yet powerful solution.

Even though there will be no difference between structured and unstructured data in the GDPR, Peter argued that most organizations have not focused on mapping their data, even if that has been one of the top priorities for many organizations (2.16). Instead, he argued that

they have focused on how they handle the structured data and changed the processes of working with it after the 25th of May, when the GDPR become effective (2.16). He believes that many organizations have turned their backs on the unstructured data, mainly because it is to complicated to manage (2.16). However, the same rules will apply for both types of data, and the GDPR requires among other things, to have a record of processing activities. To cope with this requirement, the Aigine solution established early in the project the use of a metadata warehouse to use as a record for processing activities (4.41). All documents which have been screened gets classified by which legal basis they have and for how long they can keep that specific data (4.41; 1.16; 2.20). The metadata warehouse helps, not only with controlling which documents have been screened, but also with the requirements that GDPR advocates (4.41; 2.20). That means, as described by Karl-Oskar, that the record of processing activities can easily capture where and on which places someone's personal data is stored, and easily present it. We believe the metadata warehouse they have created as a record of processing activities will thereby easily handle some of the most significant requirements of the GDPR, such as the right to access, the right to be forgotten, and the right to data portability. However, while the Aigine solution only focus on the unstructured data, we believe that the solution will easily be adaptable to also handle structured data, so that all forms of data within the organization can be handled in the same system. This means less systems to cope with, and all personal data, structured and unstructured, within the same record of processing activities.

5.6 Future opportunities with cognitive computing

The future opportunities presented by the interviewees in the empirical findings were more aimed towards future opportunities with the actual solution and of its infrastructure. Both Peter and Karl-Oskar expressed that fact that the Aigine solution could easily be applied to new forms of data, not necessarily just unstructured data (2.20; 4.53). They both mentioned the ability of applying the system to structured data in databases, with screening of tables and contents so that they are free from any personal data (2.20; 4.53). Karl-Oskar were also keen to explain that they have only focused on unstructured data because it previously was unregulated, and is not an area prioritized by other system providers (4.53). However, we believe that adapting the solution to work with structured data could have incremental benefits. As we previously explained, the metadata warehouse used as the record for processing activities, can in our mind be easily adapted to be used with the structured data as well and thereby control all personal data within the organization in the same system, and thereby more easily cope with the requirements that the GDPR advocates, both for structured and unstructured data. And the fact that Peter and Karl-Oskar do not see any problems with applying the solution to structured data only strengthen our previous suggestion.

Ari and Claes-Olof also addressed the ability of the definitions (1.28; 3.51), Ari explained that they are creating more and more capable common definitions which now is focused on privacy data (3.51). However, he believes that we will be able create a system that will be able to abbreviate from any type of document (3.51). The infrastructure the system enables is another opportunity that Hans mentioned (5.32). He argued that it will for sure enable us to build a lot of timesaving and legally secure solutions (5.32). The abilities of cognitive computing sys-

tems they all present, with many capabilities such as pattern recognition, data mining, and natural language processing will according to us only further enable new and more capable solutions to enhance efficiency in organization within many different industries.

6 Conclusion

We have identified several challenges and opportunities in our case study. Because we examined a specific case in a specific context, our findings are applicable for using cognitive computing for mapping personal data in order to help comply with the GDPR. Some of the findings are new while some are in correlation with conducted research, highlighted in the discussion chapter.

The research question we aim to answer is:

- ***What are the challenges and opportunities with using cognitive computing for mapping personal data in order to help comply with the GDPR?***

To answer our research question, the following challenges have been identified:

- Cognitive computing requires manual work, so it can learn and become effective.
- It is challenging to increase the accuracy of cognitive computing abilities enough to identify all personal data.
- Cognitive computing requires sufficient training data.
- Cognitive computing requires huge amounts of human contributions for it to become as accurate as humans.

To answer our research question, the following opportunities have been identified:

- Cognitive computing can understand structured and unstructured data.
- Cognitive computing can be learned to identify personal data.
- Cognitive computing can learn continuously and thus improve its accuracy.
- Cognitive computing can increase efficiency and thus reduce workload, save time, and reduce costs.

Appendix 1

Name: Claes-Olof Olsson

Role in project: Executive official

Company: Sambruk

Time: 28:07

Date: 2018-05-03

Refe-rence number	Code	Per-son	Questions and answers
1.1		LH	Vi tänkte börja med att fråga om det är okej att vi spelar in det här samtalet?
1.2		CO	Yes, absolut.
1.3		LH	Vill du vara anonym? Eller är det okej att vi anger ditt namn och din roll och så vidare?
1.4		CO	Ja, det är okej att använda namn.
1.5		LH	Kanon!
1.6		LH	Kan du börja med att beskriva din bakgrund och utbildning lite grann?
1.7		CO	Okej, om vi ska ta det från början så är jag civilingenjör från LTH, elektronik. Jobbade ursprungligen ett antal år inom processtyrning på Alfa Laval och har gått över sen till IT-sidan som konsult på det som idag heter Capgemini. Och därefter ett par år på Stenalines engelska bolag som IT-chef där och strax innan millennieskiftet så flyttade jag tillbaka till Sverige och var IT-direktör i Malmö stad fram till för ungefär 10 år sedan, när jag valde att sluta och bli egenföretagare, så jag är konsult och jobbar som sådan sen drygt 10 år. Jag har då haft detta uppdraget som övergripande ansvarig för föreningen Sambruk verksamhet, som sagt i 10 år. Jag var med och startade det, föreningen Sambruk och under min tid i Malmö stad eftersom jag redan då tyckte att Sveriges kommuner behöver sam-arbeta mer och bättre och det fanns ingen sådan organisation då som erbjöd den, så att säga forumet för samarbete för verksamhets-utveckling.

1.8		LH + HM	Intressant!
1.9		LH	Vad har du då haft för roll i det här projektet?
1.10		CO	Jag har ju varit, så att säga samordnande, det här började, ni har intervjuat Peter sen tidigare, vår projektledare?
1.11		LH	Ja, vi har snackat med Peter, precis.
1.12		CO	Så han har gett bakgrunden, så att det här egentligen började som en projektidé som vi försökte intressera Vinova för. Vi har haft flera FU-projekt som har finansierats eller delfinansieras av Vinova under åren och för första försöken med detta måste ha varit 2 kanske till och med 3 år sedan, ungefär så. Vår idé, det var Peter som var initiativtagare, eller idéknäckare till det här, att vi skulle sätta upp ett, vi kallade ett datorlabb och det fanns den sortens utlysningar inom Vinova då. Det var två sådana projektansökningar som vi fick avslag på och när vi sen hade den diskussionen uppe med ett par andra intressenter kring just det här området, så kom de fram till att det är skulle man kunna använda, just för att utforska och att hitta ostrukturerad data, så personuppgifter. Så då satte vi upp det projektet och jag har normalt den rollen när det kommer projektidéer, de flesta projektidéer kommer ju som, kan vi säga förfrågningar eller problemställningar från våra kommunmedlemmar men i det här fallet så var det ju mer en, ja det var en organisation som är med i föreningen Sambruk eller blev medlem på grund av detta projektet och tyckte det var så intressant. Så att jag, man kan säga var övergripande styrgruppsansvarig när vi formerade projektet och, så att säga erbjöd föreningen Sambruks samarbetskoncept och plattform att samla in fler intressenter och, så att säga fördela och organisera arbetet som sen då Peter Mankenskiöld är projektledare för och driver det operativt, så att säga.
1.13		LH	Skulle du kunna berätta hur lösningen fungerar?
1.14	CUC CLP OLA OCG	CO	Det får du faktiskt ta med Peter för att jag har alltså ännu inte sett den i verkligheten och blivit lite, projektet skulle varit färdigt, det skulle ha varit levererat, ett färdigt koncept och vi skulle se hur det fungerade och affärsmodellen kring det här, hur våra medlemmar och andra intressenter ska (avropa, oklart om rätt ord) det hela. Men en av dem företagen som projektpartner som inte lyckats leverera det här för två veckor sen och så vitt jag vet så är Peter igång just nu med att sätta upp det här hos en kommun för att söka ut det hela. Det lilla jag, alltså, som jag förstår det så är det ju just att man sätter upp AI lösningen, dels lokalt, alltså man kan säga en frontend i organisationen i det här fallet, i vårt projekt, alltså i en kommuns IT-miljö. Man definierar ett antal parametrar som ska göra att

			AI lösningen ska kunna identifiera bland en oerhört stor mängd data, är detta en personuppgift eller ej. Sen kör man igenom det här och så levererar då verktyget, AI-verktyget en analyserad lista som kommer fram till att, ja det här en personuppgift och det här behöver ni då fokusera på att hitta, helt enligt dataskyddsförordningen, ni måste motivera var det finns för laglig grund för den här behandlingen, ni måste definiera varför ni använder, och ni hur länge ni använder och så vidare. Och de här bitarna samlar man sen in och så att grundbulten i AI-systemet är ju sen att, eller verktyget är att det lär sig då att den här sortens definitioner hittar den här sortens personuppgifter, kopplat till de här lagliga grunderna och så lär sig systemet efter hand då och blir bättre ju fler gånger det körs. Den här diskussionen hade vi inledningsvis med en av kommunerna där vi körde och där den administrativa direktören blev lite besviken för han sa det, att det verkar ju som att det krävdes en jävla massa manuellt arbete för att få det här att fungera och trodde liksom att det här skulle gå automatiskt. Och det har både jag och Peter replikerat som jag sa att, båda artificiell och mänsklig måste läras upp först innan den blir effektivt och det är ju verkligen så med datorbaserade intelligens eller de lösningar, det lär sig efterhand. Det som är nytt i den här lösningen som Peter beskriver för mig är att det här centrala kunskaps repository, det finns då i molnet men det handlar ju inte om då personuppgifter utan det, så att säga den här samlade lärdomen om, vad är en personuppgift, hur ska den tolkas, hur ska det behandlas. Själva utsökningen av data görs helt lokalt, man kan kalla det för client server lösning, om ni varit med på den tiden då det var hett, för 10-15 år sedan?
1.15		LH	Hur kommer lösningen implementeras sen då, för den ska väl till flera kommuner?
1.16	CCG CLP OCG	CO	Ja, det här sätts ju upp då som ett avrop som koncept där man i princip abonnerar den här tjänsten, man får ett, man kan säga ett installationspaket för att installera lokalt som sen körs mot den här molnbaserade repository:t dvs kunskapsdatabasen och sen kan kommunen eller en annan organisation men kommuner är oftast väldigt väldigt stora datahanterare. En av dem vi provat på en medelstor kommun, de hade ungefär 4TB data och då pratar vi bara sånt som är ostrukturerad data, alltså i mailservrar, i C och M och X diskar och allt vad det är. Det är det som är svårt att hitta och identifiera, dels så är det spritt på oerhört många människor, det kan ju vara 1000 användare även i en rätt så liten kommun. Och var och en har lagt upp sina egna lagringsplatser och har mycket email sparat på olika sätt och så vidare. Det behöver man ju policy mässigt först besluta om hur ska det är hanteras och då behöver man det här stödet och det är ju inte bara så att man behöver köra igenom och få koll på det till den 25 maj utan det här blir ju viktigt varje dag efter den 25 maj också när man kan få en fråga från en individ,

			har ni som organisation någon personuppgifter om mig? Då ska man söka igenom 4TB data till exempel och leta rätt, finns den här personen på något sätt? Det speciella med AI-lösningen är ju också att det räcker liksom inte att leta efter förnamn, efternamn, adress och inte heller via personnummer utan enligt dataskyddsförordningen så kan ju en kombination av två inte helt säkerställda identiteter så att säga kombineras till att bli en personuppgift i dataskyddsförordningens definition så att säga. Det är ju därför det är viktigt att den här AI-lösningen nu ständigt lär sig och det är ju det också kopplat till föreningen Sambruks grundbult och vision, att vi jobbar tillsammans, får bättre kunskap och lär både oss och detta fallet systemen mer och mer och blir effektivare.
1.17		HM	Implementerar ni de nya system stegvis då, dvs kommun för kommun eller hur gör ni?
1.18		CO	Ja, det kan man säga. Det finns ingen samlad inköpskanal, man gör upphandlingar ibland gemensamt men varje kommun har ju en stor och ofta värnad frihet att liksom satsa på, i sin takt på sina egna lösningar och det är det som vi försöker hjälpa till med från föreningen att man implementerar i sin egen takt men man ska ju också samtidigt kunna göra saker ting tillsammans här i utrednings-kravspecifikation och framställningsfasen. Sen har vi ingen rådighet eller styrning över när kommunerna, hur kommunerna eller hur snabbt de implementerar den här lösningen utan de är fria att avropa den från nästa vecka när den konceptet levereras från projektet.
1.19		LH	Om vi ser till utmaningar, vad i projektet har varit utmanande?
1.20	CNS	CO	Ja den stora utmaningen, att det här aldrig har prövats förr, på det här sättet och det är därmed en abstrakt liten okänd materia och man kan säga organisationer i allmänhet och kommuner i synnerhet är ibland lite avvaktande inställda till det. Jag brukar beskriva som att kommunfolk vill gärna stå upp och skryta om att de varit först med någonting i Sverige men helst ska någon annan ha känt efter provat på en sådan här lösning först. Så att det är kombinationen, vi skapar någonting som inte finns förut och vi ska övertyga våra kommunmedlemmar att det här är fantastiskt bra så att man ska avropa och beställa det snabbt nu, helst innan den 25 maj naturligtvis men även efteråt sen. Jag kan lägga till en sak på utmaningar så är det ju rent ekonomiskt. Vi har inga färdiga eller extra pengar så oss att finansiera ett sådant här projekt utan alla våra projekt finansieras av att ett antal kommuner anmäler sig som intresserade av att vara med i ett sådant här projekt och så bidrar dem också till projektbudgeten. Just eftersom det här är nytt och oprövat så har det

			varit lite, har utmaningen varit att hitta tillräckligt många intresserade och engagerade kommuner som ska finansiera den projektbudgeten vi satte upp i början.
1.21		LH	Om vi tänker tekniska utmaningar, har ni upplevt någonting där?
1.22	CNS	CO	Ja, som jag sa inledningsvis att det har väl Peter berättat att vi inledde ju ett tidigt samarbete med IBM och det var ju deras AI-koncept Watson kring med produkter så att säga och de var övertygade att det här gick att göra men när det verkligen skulle realiseras faktiskt så hade inte IBM själv dem resurserna och knappt den kompetensen. Då var utmaningen att hitta en IBM partner som kunde faktiskt tillverka och sätta upp det här konceptet och då hittade Peter till slut en finsk konsultfirma som har konstruerat själva lösningen. Ett tag var det lite att, jo det går att realisera men vi vet inte riktigt hur.
1.23		LH	Om vi går då ganska specifikt, ser du några speciella utmaningar med den här faktiska AI-lösningen om vi tänker systemet i sig, som till exempel IBM Watson, har ni upplevt några utmaningar just med det systemet?
1.24	CNS	CO	Nä inte rent tekniskt, det är kanske en del mer mentala utmaningar, att övertyga intressenter och de som är intresserade av det här att verkligen satsa på det här, det finns ju alltid människor som, när vi kommer och presenterar en lösning säger, jaha varför valde ni just den här tekniska lösningen och varför valde ni just IBM och så vidare. Men det kallar jag för som sagt mer en mental utmaning än en rent tekniskt för att så vitt vi vet, Peter och hans projektteam har både själva tittat på många olika lösningar och fått …18.00... från omvärlden men har konstaterat att det finns ingen motsvarighet till det här någonstans idag.
1.25		LH	Hur kommer då lösningen förändra själva verksamheten, vilka möjligheter erbjuder systemet?
1.26	OIE	CO	Först och främst är det just att kunna underlätta dramatiskt, att hitta igen och strukturera den här enormt stora mängden ostrukturerad data och som alla konsulter, jurister och andra människor som diskuterar och som kan någonting om dataskyddsförordningen så hävdar vi ju alla att det är ju inte bara att undgå det här att det kommer, krav och ordning och reda och att man ska ha laglig grund för hantering av personuppgifter utan till sist så kommer det här också resultera i en effektivare administrativ verksamhet och har man ett effektivare administration så har man också ett mycket bättre stöd för, om vi kallar det för kärnverksamheten. Det kommer ju det här verktyget då att hjälpa till med, dels med att man ska hitta data, man får en betydligt effektivare analys av hur ens stora mängd strukturerade och ostrukturerade data ser ut och som hjälper när

			man behöver lägga extra kraft på både skydda känslig data men också kanske hitta. Jag är övertygad om att man kommer hitta en stor mängd dubbel lagrad data eller multipelt lagrad som ligger på olika ställen som inte behöver finnas på massa olika ställen och därmed får man ju också möjlighet att strukturera och effektivisera hela sin infrastruktur, alltså IT-infrastruktur om hur man lagrar och behandlar grunddata. Just i kommunal verksamhet så är ju, skulle jag säga en förkrossande majoritet av all data som kommunen hanterar är ju personuppgifter. Kommunen är ju till för sina kommuninvånare och mycket kretsar kring kommuninvånarnas behov och av service, tjänster och stöd på olika sätt.
1.27		LU	Ser ni några framtida utvecklingsmöjligheter i den lösningen ni har tagit fram?
1.28	OFO OIE	CO	Den började som sagt som en spin off på det här som Vinova kallar för datadrivna labb och vi hade då ambitionen att använda det här för att lite skynda på en bredare användning av öppen data. Det tror vi fortfarande att det här kan göra, alltså inte bara handla om personuppgifter utan man kan ju sätta in en definition av vad som helst och få ut det. Jag tror personligen, och det har jag och Peter pratat om sen de här första projekten, sökningarna att strävan hittills har ju varit väldigt mycket att man ska skapa gemensamma standarder och begreppskataloger så att man ska kunna få ut olika former av information som öppna data. Så ska företag kunna bygga nya affärsidéer på att få ut öppna data, framför allt för den offentliga verksamheten. Jag tror att den här lösningen kommer visa vägen där man istället för att först definiera hur, ska vi säga målsystemet eller källan ska strukturera sin data för att man ska läsa ut den och sen behandla den. Så gör man så att säga tvärt om, man definierar med hjälp av den här AI-konceptet eller verktyget, det är den här sortens data jag vill få ut och sen kan den ligga lagrad strukturerad och definierad i princip hur som helst och var som helst i en offentlig IT-miljö till exempel, så får man ut det. Det kräver mycket mindre jobb och framför allt behöver man inte som idag lägga en massa tid och resurser på olika projekt från olika offentliga initiativ där man först ska komma fram till att ja..., när det gäller tidtabeller till exempel från olika lokala och regionala transportföretag, offentliga sådana, så måste tidtabell informationen vara strukturerad exakt så här. Det är en av anledningarna till att det har gått så långsamt. Vi har drivit sådana projekt också, Peter har varit projektledare en gång tidigare och han och hans kollegor höll på i ett och ett halvt år och kom fram till att två definitioner och det var…, redovisningsdata som egentligen är extremt väldefinierade från början och så var det uppgifter om kommunernas miljöförvaltningar gör allt som oftast tillsyn på restauranger och så vidare. Det kan ju vara intressant för oss som privatpersoner att det kan finnas en lista av vilka restauranger som har, inte bara godkänt käk utan godkänd

			miljö överhuvudtaget. Det tog som sagt Peter och hans kollegor drygt ett år att samla ett antal kommuner för att komma överens om, okej, såhär strukturerar vi den här datan. Det har liksom vi pratat om, eller haft det här projektet och gått i ordentligt med 10 kommuner och publicerat att det här är liksom en standard för öppna data kring det här. Så är det 280 kommuner som inte har varit med som inte har liksom gjort hemjobbet än. Den här sorten initiativ håller man på med olika kommuner runt om i Sverige och så lanserar man, nu har vi publicerat den här sortens information som öppna data. Det hjälper ju oftast inte företagen för de vill ha informationen på samma sätt från alla kommuner om de ska kunna göra några intressanta applikationer kring det hela. Så om man vänder på hela det här resonemanget och säger vi definierar hur AI systemet ska hitta data kring restaurang inspektioner till exempel, matinspektioner, så kan det systemet leta i 290 kommuners miljöförvaltningars system som säkert har det på 200 olika sätt i alla fall och hitta den här data och strukturera den och så får man ut det som om det hade varit gemensam standard för de här dataset:en från början.
1.29		LH	Nu svarar du redan lite på det men jag tänkte ändå fråga. Ser du några ytterligare möjligheter med just ett sådant här AI system då?
1.30		HM	Vi tänker väl egentligen, Watson erbjuder ju otroliga lösningar, finns det ytterligare funktioner ni kan baka in i lösningen?
1.31		CO	Det är klart att det finns men vi har inte kommit så långt så att vi har hunnit liksom brainstorma kring det. Jag är så gammal så att jag liksom sett ett antal historiska exempel på att något snille tar fram en viss lösning och det gäller ju särskilt IT-sidan, i tron med att det här löser det här specifika problemet och sen någon annan som tar hand om den här lösningen och gör något helt annat av det. Jag brukar ta Mark Zuckerberg som exempel, som verkar ha gått full cirkel nu eftersom det sas iallafall att han började konstruera Facebook för att kunna dejta tjejer på universitetet där han jobbade och liksom två miljarder användare senare så är han nu tillbaka till att nu ska han skapa en dejtingapp av Facebook. Men under tiden har det ju blivit något helt annat av den här appen som han tog fram.
1.32		HM	Det var faktiskt allt vi hade.
1.33		LH	Ja, vi känner oss klara. Vi tänkte fråga, har du något som du vill tillägga kring projektet eller ämnet eller?
1.34		CO	Nä, jag är nöjd av att fått ut det så här.

Appendix 2

Name: Peter Mankenskiöld

Role in project: Project manager

Company: Sambruk

Time: 1:09:10

Date: 2018-05-07

Reference number	Code	Person	Questions and answers
2.1		LH	Är det okej att vi spelar in?
2.2		PM	Ja det är klart.
2.3		LH	Vill du vara anonym eller är det okej att vi anger ditt namn, din roll och så vidare?
2.4		PM	Nä men det är väll klart jag måste vara offentlig, officiell.
2.5		LH	Perfekt. Nu vet du ju redan vår bakgrund så vi kan köra direkt på frågorna. Kan du ta och berätta lite om din bakgrund och tidigare utbildning?
2.6		PM	Jo, jag har en.. Jag har pluggat ekonomi och systemvetenskap. Och sen så har jag ett antal olika typer utav utbildningar i samband med anställningar. Jag började mitten på 80-talet, 82 faktiskt, på SE-banken och har jobbat som banktjänsteman och har alltid haft ett stort teknikintresse. Så där jobbade jag vidare och fick då systemvetenskapsutbildningen via banken. Det var på den tiden som organisationer fortfarande kunde betala för utbildning, rätt fascinerande. Så att där har jag suttit och jobbat med utveckling också faktiskt utav en del system och utifrån det så har jag vidareutvecklat ett antal saker. Efter banktiden så har jag haft ett antal egna företag med olika inriktningar på leverans utav effektiviserande verksamhetsstödssytem, skulle man kunna säga. Det här har jag hållt på med sedan 20 år tillbaka typ, i olika former. Jag har alltid varit väldigt verksamhetsorienterad och för min del så har också då det här med informationsförvaltning och informationsstrukturer varit det som har varit den genomgående faktorn, det som har ständigt återkommit, därför att det är det jag tror är en utav de viktigaste pusselbitarna för att få en effektiv organisation.
2.7		LH	Okej. Vad har du för roll i det här projektet?

2.8		PM	Projektledare
2.9		LH	Kan du berätta lite mer om din delaktighet?
2.10		PM	Ja, det är ju så att jag är fristående konsult men har då jobbat som underkonsult för bland annat då, eller inhyrd konsult ska jag säga, för bland annat då SKL har jag jobbat hos också kopplat till informationsförvaltning med samordningen öppna data för kommuner och landsting. I det arbetet har jag lyft fram ett antal olika aktiviteter som jag tror är viktiga, bland annat då att få jobba med standardisering utav framförallt outputen, alltså hur man då ska presentera öppna data på ett gemensamt sätt. Det är en utav de olika bitarna, sen så är det också naturligtvis kopplat till när det då är strukturen, alltså hur ska vi kunna hitta den här informationen som ska vara tillgänglig för publicering. I det så har det krävts att man faktiskt har en koll på vilka informationsmängder man har och att man har möjlighet att klassificera dem. Men som sagt, SKL-arbetet var en sak, och när jag då hade ett antal olika aktiviteter som skulle jobbas vidare med, bland annat då det här med standardisering, så ville man från SKLs sida att det skulle flyttas över till föreningen Sambruk. Så det var föreningen Sambruk som sökte medel hos Vinnova för att jobba med den typen av frågor. Så där har jag kommit in på föreningen Sambruk och har då haft ett antal sådana projekt som är kopplade till just det här med informationsförvaltning. I det så har vi sökt medel för någonting kallades för datadriven innovation. Det var egentligen då som embryot till det här projektet dök upp, det var som vi kom fram till att det här är något bra exempel på vad man skulle kunna göra med datadriven innovationslabb. Vinnova var inte speciellt sugna på att finansiera det här projektet, så att det har finansierats via medlemmarna istället, vilket innebär att vi har haft en runda där vi har haft ett antal, 10 kommuner i dagsläget, och sen ett regionsförbund, som är med och finansierar den tiden som vi har lagt ner från i höstas fram till strax efter årsskiftet i projektet. Min roll har ju då varit som sagt initiativtagare till hela projektet men också då projektledare, och jag har då varit den som jobbat med att försöka få in medel från de här olika medlemskommunerna. Så vad har jag inte gjort i projektet skulle man kunna fråga? Det har varit väldigt mycket av ett enmans-projekt det här faktiskt, eftersom väldigt mycket bygger på de idéerna. Jag tror jag berättat för er om det här tidigare men jag kan berätta det igen. En sak som jag såg som väldigt viktigt är just det här med inventering för att kunna få koll på sina datamängder. För det är ju någonting som jag jobbat med som sagt de senaste 20 åren och tyckt att det här är någonting som är väldigt viktigt att man verkligen vet. Min uppfattning är att vad vi har gjort i Sverige är ju väldigt mycket utav det som man skulle kunna kalla för engelskans digitizing. Vi har alltså inte jobbat med digitalisering i ordets bemärkelse utan bara digitisering, om man ska försvenska det. Vi har bara lyft in de olika formulären som vi

			har haft analoga, har vi lyft in i datamaskinen. Det är ju inte riktigt digitala processer, utan det är ju bara att man har flyttat in processen som den är in i ett elektroniskt flöde. Det kunde naturligtvis underlätta rätt så mycket inledningsvis i den här datoriseringseran, men det krävs ju väldigt mycket mer för att jobba med digitala processer. Framförallt så har vi gjort fel när vi har lyft in blanketterna och vi har byggt olika typ utav system, så har vi låtit de olika leverantörerna bygga informationsstrukturen precis enligt eget huvud, och specifikt för varje system som utvecklats. Det är rätt så svårt, för har man jobbat med någon form utav systematiserad datautveckling så vet man ju att det viktigaste är ju att ha struktur på det man jobbar med. Det är ju inte riktigt det vi har gjort med informationsstrukturen i Sverige. Det är ju inte ett svenskt fenomen utan såhär är det ju i princip överallt. Så att nu då, när vi har en utveckling som är fascinerande med digitalisering och så vidare, så krävs det ju mer och mer utav, att ha koll på den informationen vi har, som jag ser det. Men, jag lyfte fram för ett par år sedan en otroligt enkel modell för att kunna identifiera, den är inspirerad utav norska Drifi som då hade en.. När de jobbade med öppna data-frågan så jobbade de framförallt då med något som de kallade för rödljus-modell, och det är den som jag har tagit till mig. Det är rätt så enkelt. Kan du klassificera dina datamängder, och då pratar vi alltså om datanivå, inte på datakälla, alltså systemnivå. Om du kan strukturera och klassificera informationen utifrån ett perspektiv där rött innebär att det är data som inte är lämplig att publicera därför att den innehåller olika former utav personuppgifter eller sekretessbelagd information och så vidare. Den gula datan, det som är gulmarkerat är det som då är ämne för beslut, alltså du måste fatta beslut om det innan du kan göra någonting utav det. Det kan vara arbetsmaterial, eller andra saker då som ännu inte är klassificerade skulle man kunna säga. Sen så har du det som är grönt uppmärkt, det som är lämpligt att publicera. Och det här låter ju jätteenkelt, och det är klart man skulle kunna göra det här, men då är frågan vilken organisation, om vi pratar offentlig sektor, som har mycket annat att göra, tycker att man har råd eller tid att göra en sådan klassificering i enkom för att publicera öppna data, ställer jag frågan. Också känner jag direkt att då har ni någon idé eller inte, jag tror inte att det finns någon nämligen som tycker att det här är självklart att det är värt det för vi ser ju att öppna data ger ju en sådan otrolig genomslag i vår ekonomi och så vidare, jag tror inte på det. Utan det kommer vara som så att, aa men det där låter jättebra, den ska vi titta på någon gång när vi har tid, och det är ingen som har tid. Så mitt förslag då var egentligen till de som har lyssnat på mig, det var då att, man kanske ska titta på GDPR, därför att de som jobbar med GDPR-projekten måste göra den här genomlysningen. De måste göra den här klassificeringen, fast de måste göra den på en djupare nivå, men fortfarande på datanivå, där de alltså måste göra en klassificering utifrån informationssäkerhet. Att då lyfta in den här rödljus-modellen i samband med

			det och få med det som en del i en matris, det är egentligen inga större problem. Jag försökte få in det här i ett projekt som heter Klassa som SKL har gjort för ett antal år sedan, 3 närmare bestämt. Klassa bygger alltså på en informationssäkerhetsklassificering som bygger på ISO-27000, det var alltså innan GDPR. Problemet som jag ser med Klassa och problemet med ISO-27000 det är att man klassificerar på systemnivå och inte på datanivå. Och, vi måste då klassificera på datanivå för att tar vi då exempelvis ett betygssystem i en skola så är det per definition i Klassa och ISO-27000 anses då vara ett högrisksystem, eftersom det finns väldigt mycket personuppgifter i det. Men om vi tittar på det utifrån ett rödljusmodell-system och vi tittar på det utifrån datamängder så ser vi det att på en viss aggregerad nivå så skulle en vy kunna vara helt grön, alltså den skulle kunna vara lämplig att publicera för att den är enbart ren statistik. De skulle dessutom kunna underlätta statistikutlämnandet till de myndigheter som samlar in statistiken. Och det här är något som jag har verkligen haft som min käpphäst, att man hittar de olika bitarna som faktiskt kan underlätta. Jag har massor med exempel på just det. Jag vet inte om vi har pratat om informationslämning till SCB. En utav de sakerna som jag hittade på SKL, det första jag tänkte på.. Jag kommer tillbaka till det här med öppna data och den strukturen. För det är ju såhär, det absolut bästa skulle ju vara att 290 kommuner.. När jag började med det här för 5 år sedan så var det ungefär 5 stycken kommuner som publicerade öppna data överhuvudtaget. En del ansåg att allting som fanns på webben var öppna data, och det är ju en definitionsfråga naturligtvis. Men, skillnaden var ju enorm, dels då de här 5 utav 290 som publicerade någonting överhuvudtaget. Stockholms Stad publicerade mer än hälften av all öppna data som publicerades överhuvudtaget i Sverige. Vilket innebär då att från noll till en sådan stor mängd publicerad öppna data så är det väldigt svårt att hitta någon form utav genomsnitt.
2.11		LH	Kan du beskriva hur er lösning fungerar?
2.12		PM	Ja det kan jag göra. Vi behöver inte gå in på bakgrunden mer?
2.13		LH	Jag tror vi har en del bakgrund nu.
2.14	CCG CUC CLP OIE OLA	PM	Ja, för att jag tror en sak som är viktig som jag inte tror att jag berörde, det var ju just det här med att när vi tittade på vad som fanns på marknaden. För det var ju rätt så viktigt. Det som finns idag på marknaden, det finns ett antal verktyg som gör att du kan titta på hur din situation ser ut, alltså vad gäller då personuppgifter i ostrukturerad data till exempel. Det som är problemet med den typen utav uppgifter eller den typen utav system som finns är de bara visualiserar hur situationen ser ut, alltså tydliggör eller synliggör för dig vilken utmaning du har. Men det som behövs är ju också naturligtvis någon form av beslutsstöd, någon form av arbetsstöd, så att du

			faktiskt kan beta igenom den här högen med dokument. Det var ju det som egentligen saknades. Det var en utav de viktigaste bitarna, att det inte fanns något som hjälpte till att minska den här mängden dokument med misstänkta förekomster utav personuppgifter. En annan sak som var väldigt viktig det var ju utifrån både juridiskt och tekniskt perspektiv.. Jag har ju sedan tidigare, kanske nämnde litegrann då i det här datadrivna labbidén, så har jag haft en tanke på att just artificiell intelligens ska kunna hjälpa till som ett stöd i att underlätta beslutsstödssytem och så vidare. Att underlätta besluten genom att kunna föreslå dels då de olika, i det här fallet förekomster utav personuppgifter där man föreslår vad man tror själv är personuppgifter, alltså vad maskinen tror själv. Där människan då kan komplettera och rida över men också då när det gäller beslut utav den rättsliga grund, eller den laga grunden som finns för att man faktiskt har personuppgifter i olika system. För jag tror ju det att i det här så finns det en väldigt stor hjälp vi kan få utav någon som kan läsa igenom dokument mycket snabbare än en människa kan göra, och faktiskt förstå vad man läser, i viss mån. Det är precis det som jag har varit ute efter, det är därför som jag då har sett, också då AI-möjligheten att kunna komplettera med. Projektet hette ju då, väldigt lång titel faktiskt på projektet, men det är ju då hjälpen att söka igenom komplexa datamängder med stöd av AI. Så AI har funnits med hela tiden från början. Där stötte vi på lite patrull ska jag inte säga men utmaning eftersom det faktiskt är så att de flesta AI-lösningar bygger faktiskt på någonting som är molnbaserat. Och just det här med att granska personuppgifter och molnbaserade lösningar det går liksom inte ihop. Det fungerar inte rent juridiskt sett. Eftersom du kan inte hålla på och skicka upp personuppgifter i en molntjänst för att göra olika typer av bedömningar. Så det har varit en av de stora utmaningarna, och det här har varit en av de sakerna som vi har brottats med hårdast egentligen med de arkitekterna som varit involverade på IBM. Dels för att hitta de olika komponenterna som ska kunna sättas samman för att göra den här typen utav systemlösning som vi har varit ute efter, men också då för att kunna göra en rättssäker lösning som gör att vi faktiskt har en möjlighet att göra genomlysningar och bearbeta de genomlysningarna och fortfarande bibehålla alla data på, vad ska vi säga, på engelska heter det on-premises, på svenska heter det lokalt. Det är ju det som är den stora biten, att försöka få den här paketeringen, och det är det vi har lyckats med, genom då våra finska vänner Elinar som vi har fått kontakt med från IBM, de har pekat på att det är en IBM partner som enligt IBM är spetskompetensen på AI i Skandinavien. Jag är benägen att hålla med, de är otroligt duktiga och de har verkligen gjort väldigt mycket och de har dessutom lyckats.. De hade en lösning som hette Elinar AI Miner som då gjorde delar utav det som vi var ute efter. Vi har också tittat på den traditionella renodlade IBM produkten sedan ett antal år tillbaka, som heter Stored IQ, som är en väldigt duktig motor på att göra det som man kallar för Crawling.

			Som går igenom och läser igenom olika filer framförallt då i ostrukturerad data för att kunna göra olika hanteringar och bedömningar utav det. Det som är intressant då det är ju att man kan föra samman de här olika bitarna, för att Stored IQ i sig är ungefär, den är mycket mer omfattande än de här enkla crawling-verktygen som finns att ladda ner gratis och så vidare naturligtvis, men det är fortfarande så att den visualiserar situationen, den visualiserar hur stort ditt problem är, men den gör inte lösningen om man säger så. Men med hjälp utav Elinars AI Miner så hade vi möjlighet att göra den kopplingen så man skulle kunna få hjälp med lösningen. Men då var det fortfarande problemet att AI Minern i sig är en centralt baserad molntjänst. En sådan AI-funktion bygger ju väldigt mycket på.. Den har ju GPU istället för CPU, det är ju grafiska process som jobbar fram det här underlaget. Det finns någon server som heter Minsky. Det här är ju maskiner som är jättekraftfulla och kostar därefter. Så det är ingenting man normalt slänger in på vilken organisation som helst, utan det är det som grejen med AI att det är en stor central funktion med väldigt mycket datakraft. Men det som man har gjort nu, det är att man lyckats bryta ut en del utav AI-funktionaliteten, framförallt det som är annotationstekniken för just igenkännandet utav mönster. Mönsterigenkänningen och kognitivt lärande som då handlar om dels naturligtvis att ha de olika igenkänningsmönsterna för vad som kan vara definitionen av en personuppgift, men också då vad som är kopplingen till hur man hanterar, alltså hur olika former utav personuppgifter ofta kopplas till olika former utav laga stöd, alltså olika typer utav beslut kring laga stöd. Och det är ju rätt intressant därför att då lagrar vi ju, då samlar vi den annotationsdelen lokalt. Vi bygger sedan, och gör en anonymiserad version som också då krypterat kommunicerar med den centrala motorn. Så vi har en lokal AI-komponent och sen så har vi en central AI-motor som då kommunicerar med varandra, men inte med personuppgifter utan kommunicerar bara annotationskod. Det gör ju också det att det vi lär oss lokalt, kommer vi då att kunna föra upp. Den delen som är.. Det lärandet som vi har gjort, skickas upp till den centrala enheten och bearbetas och gör då att nästa upplaga utav algoritmen, alltså där man läser igenom de olika dokumenten, så har man lättare att hitta definitioner på personuppgifter, misstänkta personuppgifter, eller presumtiva personuppgifter, därför att man får en bättre.. Vi kan ta ett exempel. Att en personuppgift är ett personnummer enligt PUL, det kan vara en adress, det kan i princip vara en yrkesroll och så vidare, det finns ett antal sådana färdiga saker som det relativt enkelt går att bygga olika typer av sökalgoritmer för. Hur personnummer ser ut, hur en e-postadress ser ut och så vidare, det är ju relativt enkelt. Men, när vi sen kommer till nästa steg då, så är det så att om vi tar ett exempel på att någon som kom tvåa i vasaloppet 2017, det är ju faktiskt en presumtiv personuppgift. Det pekar ut en specifik människa. Just det här med hur man sätter samman, en placering i en specifik tävling och så vidare. Det går inte att

			lägga in alla typer av tävlingar som vi har som har unika vinnare eller liknande, utan det är just själva sammansättningen utav orden som gör att AI kan känna igen något som kan motsvara någon form utav utmärkelse. Vilket ju också då skulle kunna vara delar utav en utbildning där man kanske också får en titel som då kanske är rätt unik. Jag vet personligen att jag fick titeln som certified agile master och jag var ensam om det i Sverige åtminstone en halvtimme, beroende på att mitt test rättades först. Det är en sådan grej, det finns massor av saker som kan bli presumtiva personuppgifter som faktiskt är viktiga att kunna lyfta fram. Så i det här kan tänka sig att, om vi då tittar på den här första genomlysningen så får vi fram rätt många dokument där vi då har presumtiva personuppgifter som finns, och så betar vi av dem. Nästa inkrementella genomlysning som görs så kommer den här listan att bli längre därför att vi har ju samtidigt ett antal olika användare runt om i landet som använder systemet och berättar för motorn alltså AI:t, vad som är en presumtiv personuppgift, så då kanske vi hittar fler presumtiva personuppgifter. Men samtidigt så har vi möjlighet att beta av högen. Det är egentligen det som är målet med projektet, att lyfta fram kravspecifikationen på hur det här ska kunna se ut. Och det är så, användarstöd, enkelt, intuitivt, stöd för beslut, och självklart då också att vi ska kunna dra nytta utav det här lärandet, att alla lär upp och alla får tillgång till det som maskinen lärt sig.
2.15		LH	Hur har ni då anpassat lösningen efter GPDR, om man ser till de olika kraven som det ställer?
2.16	OCG	PM	Det har varit en väldigt tydlig sak i hela projektet eftersom GDPR är ju det som varit.. Alltså, det är inför GDPR som vi har sagt att vi måste ha kolla på våra ostrukturerade datamängder. Och det kan jag säga då, om vi tänker tillbaka till den här rödljus-modellen när jag då sa till olika åhörare i samband med öppna data happenings, så sa jag just det att häng på era GDPR-projekt för att de måste göra den här genomlysningen. Problemet var bara det att det fanns i princip inga sådana projekt. Därför att man gjorde på ett helt annat sätt, och de som satte igång med GDPR jobbade inte med att göra genomlysning och inventera vilka datamängder man hade trots att det står först på de flesta listorna. Utan de man gjorde det var att man började titta på hur ska vi hantera personuppgifter i våra strukturerade datakällor framgent. Alltså, hur ska vi förändra våra processer för bearbetning av personuppgifter från och med den 25e maj. Och inte ta reda på hur såg det ut innan, vad har vi för problemhög som ligger här, utan man har framförallt fokuserat på vad som är framåt och många.. Det känns som att väldigt många har blundat för den här framförallt ostrukturerad data, därför man anser att den är, enligt min uppfattning, alldeles för komplicerad. Det är för mycket. Det är som en liknelse jag hörde från en kollega, det är ju knappast så att du vaknar på morgonen och går till din att göra-lista och först

			på listan så står fixa ozon-hålet, därför det gör man inte. Det står inte där. Det är en omöjlig uppgift. Det här har också varit en omöjlig uppgift, eller iallafall nästintill omöjlig uppgift så man har struntat i det. Det går inte ens att uppskatta hur stort det är.
2.17		HM	Vet du om några andra lösningar som andra företag har tagit till?
2.18	CCG	PM	Ja. Till exempel så är det så att det finns fenomen som heter DLP, det är Data loss eller leak prevention system. Dom jobbar ju inte så mycket med att lysa igenom datamängder som ligger lagrade utan de jobbar framförallt med trafiken, men i viss mån också då datamängder. De har ju olika typer utav.. Det finns ett system som heter Force Point till exempel, som jobbar med de här frågorna. Men det handlar väldigt mycket om genomlysning och rapportering. Så att de system som jag har hittat, de som jag har tittat på, det är alltifrån, det handlar dels om den här typen av system som ofta är då kopplat till olika former utav informationssäkerhet. Du har även då i det sammanhanget de som jobbar med virusskydd, som också då har lyft in.. För det här är ju litegrann samma epitet skulle man kunna säga. Man jobbar litegrann på samma sätt. Alltså att man har en definitionslista som man använder när man söker igenom olika filer för att se om man kan få träffar på de här definitionerna. Så att det finns ett antal sådana, men återigen är det framförallt kopplat till att göra inventering och rapportera om vart någonstans du har problem. Vilket också gör att när du har gjort en sådan genomlysning och sen så får du sitta och jobba med den här omöjliga högen med informationsmängd som du ska traggla dig igenom. Du har inget stöd för vilket laga.. alltså för granskning och beslut, och du har framförallt inte det som är också en utmaning och det är ju då själva dokumentationen. Eftersom, till skillnad från, om vi nu tittar på GDPR så är det så att där har ju då varje systemlösning och databas är ju en datakälla, vilket innebär att du kan ha miljontals entries i den datakällan men du behöver egentligen bara peka ut datakällan som en informations.. Alltså där det finns personuppgifter, enligt då din registerförteckning. Men när det gäller ostrukturerad data så kan ju varje fil, varje dokument, kan ju vara en datakälla enligt GDPR vilket innebär att den ska då dokumenteras i registerförteckningen. Det är ju det som gör att den här, alltså sökning, granskning, beslutsfattande och dokumentation är en så omfattande fråga så det är nästintill omöjligt. Det är därför som jag, vi, tyckte att det var så otroligt viktigt att få in de här stegen i ett sådant system som går att genomföra. Det var det som var själva.. Utgångsläget var att det går att genomföra, och det är det som vi har haft med oss hela tiden. Det här går att genomföra och vi ska försöka få fram en sådan lösning. Så ja, GDPR har ju funnits med hela tiden men också det som har varit det viktigaste är ju att ha.. Underlätta arbetsbördan inför GDPR.

2.19		HM	Om vi då tänker er lösning nu igen. Hur kommer den implementeras till kommunerna eller era kunder?
2.20	OCG OLA OFO OIE	PM	Sambruk har ju inte kunder i det fallet utan det är ju.. Målet var att ta fram en kravspecifikation på hur det här skulle kunna se ut för att kunna ta fram en tjänst. Och tjänsten är nu framtagen av två IBM-partners kopplat till IBM, så IBM har ju varit med hela tiden i den här processen. Implementationen bygger på.. Det här var också en sak som var väldigt viktig och det var den tekniska kravställningen. Det är ju det att det är inga små saker som är installerade utan det är ju rätt omfattande grejer. Ett utav problemen är att den första uppfattningen är att, ja men det här är ett antal arbetsdagar för att kunna få in den här Stored IQ och kunna lyfta in alla de här komponenterna. Men där har vi lyckats få ett fantastiskt paket med fyra stycken virtuella servrar bestående utav de här två komponenterna som är Stored IQ som är IBM produkten, och det som är Aigine och det är själva den lokala AI-komponenten. Den här nedladdningsbiten, det är ett paket på ungefär 20GB som då ska installeras på fyra stycken virtuella servrar i den lokala miljön hos respektive kommun. Det har vi nu testat, själva installationspaketet med uppkoppling mot en datakälla bara för att se att de bitarna fungerar och vi klarade oss under två timmar. Vilket är rätt.. Jag är jätteimponerad, och de tekniker som varit med är också imponerade. Så det är lite häftigt att kunna få paketeringen såpass bra. Sen är det naturligtvis så att nedladdning utav 20GB data tar en stund så det har vi inte räknat med i de här två timmarna. Det beror ju helt på vilken typ utav lina du har. Men målet som var, det var egentligen, det absolut bästa skulle ju vara om man skulle kunna ha allting på en USB-sticka, komma med det till kunden och stoppa in det i maskinen och sen köra och klicka på go och sen ha hela installationen färdig. Vi är nästan där, vi behöver lite handpåläggning för att kunna namnge de olika virtuella servrarna och så vidare. Men vi har kommit långt. Sen så, när man väl har gjort det så börjar man då koppla upp sig mot de olika datakällorna som man ska lysa igenom. Det är ju då det som man betecknar som data shares då med olika.. Dels då gemensamma fillagringsytor. Men också då det som kan vara e-postkataloger, som då ska lysas igenom. Då är det så att du kopplar upp dig med någon form av superuser identitet så att du har läsrättigheter, du gör en.. Du anger vem det är som är ansvarig för respektive datakälla, alltså för respektive enhet som du kopplar upp dig mot. Det gör man då med hjälp av en e-postadress. Det är den delen som är den tekniska biten, sen sätter maskinen igång och snurrar och gör sitt arbete så att säga. Därefter så genereras det då, per datakälla som genomlyses, så genereras ett e-postmeddelande som går till respektive ansvarig med en länk till där dokumenten finns, vilket är i Aigine, i det gränssnittet. Där får du din lista som då kommer vara rätt så omfattande från början. Och du klickar på ett doku-

			ment, du får.. Det första som du möts utav är de presumtiva personuppgifter som är uppmärkta, från början i dokumentet. Det gör också då att det underlättar genomläsningen därför att du ser vart man har hittat förekomster eller presumtiva förekomster. Nästa steg är att du ser något som du själv anser att, men vänta nu, det här är ju faktiskt en definition av en personuppgift som du då kan markera och trycka på det du anser att det ska vara, om det är ett namn eller ett personnummer eller vad det nu är för något. Sen så kommer då den biten som handlar om, och det gör du samtidigt då, att du markerar det som är anledning till att de här dokumenten finns. Men det kan också vara så att du har ett dokument där det finns ett antal saker som är uppmärkta, för att ibland kan det ju vara så att ett ortnamn kan vara ett efternamn eller tvärtom. Det kan också vara så att att du har andra saker som man uppfattar som något som skulle kunna vara en personuppgift men som kanske inte är det för att det är en del i löpande text och så vidare. Då ska ju du som tjänsteman ha möjlighet att säga, nä men vänta nu, det här dokumentet innehåller inga personuppgifter. Och då ska du ju kunna rensa bort de markeringarna som finns, alltså de förslagen som finns. Och det innebär ju då att, om vi går händelserna litegrann i förväg, så innebär det då också att när du har markerat ett dokument att här finns det inga personuppgifter. Då kommer det nästa gång du gör en genomlysning så kommer du inte få träff på det dokumentet. Där skiljer sig också det här verktyget gentemot de befintliga verktygen, därför att, det som vi gör är att vi bygger upp ett metadata-lager. Och metadata-lagret har flera funktioner, en av funktionerna är naturligtvis att se till så att du inte får träff på dokument som redan är granskade. Och de kan alltså vara granskade och bedömda att det inte finns personuppgifter. De kan vara granskade och bedömda att det finns personuppgifter men det finns laga stöd. Och det är ju det som metadata-registret ser till så att du inte får träff på de dokumenten igen så att din hög med dokument minskar. Samtidigt som träffsäkerheten gör att högen kanske ökar också, alltså listan blir större tack vare att man hittar fler presumtiva förekomster. Men målet är ju att den ska minska så att du kan bearbeta, beta av den här högen mer och mer vartefter tiden går. Så ungefär så är det tänkt att systemet fungerar. Det finns en tydlig avgränsning i projektet. I dagsläget eftersom vi faktiskt skriver 7e maj, så kan vi då säga att det finns en jättetydlig anledning till de avgränsningarna. En var ju då att vi bara skulle fokusera på ostrukturerad data. Jag vill ju naturligtvis, jag menar, för att gå tillbaka till historiken. Det som jag berättade om öppna data så är det så att normalt sätt så publicerade vi inte öppna data från ostrukturerad data utan då publicerade vi från strukturerad data, naturligtvis, från deras system. Därför tycker jag att det finns en stor vikt för att kunna få in den strukturerade datan i den här bedömningen också. Eftersom då kan du ha kontroll på alla dina personuppgifter, alla dina datakällor, och använda det här till mer än bara GDPR. Och jag är helt övertygad om att jag ska kunna

			göra samma typ av genomlysning utav en databastabell som jag gör utav ett filsystem. Och alltså göra samma identifiering och samma dokumentation runt omkring det. Det är teorin som finns med. Men på grund utav två saker, dels att de flesta kommuner som jag pratade med inledningsvis ansåg att de hade rätt bra koll på de strukturerade datakällorna. Hur det är med den saken det kan jag inte uttala mig om men det var så de flesta sa, att de hade koll. Då får vi leka med tanken att de har det. Och dessutom så var det tiden. Eftersom vi troligtvis skulle vara tvungen att använda andra typer utav genomlysningsverktyg, alltså crawler-verktyg, för databastabellerna. Därför att jag är inte säker på att den här Stored IQ verkligen klarar av det, jag har inte fått det bekräftat. Men det var en utav de viktigaste avgränsningarna som vi hade för att kunna klara av tidsperspektivet. Det andra var att vi bara skulle hantera det som är innan 25e maj. Om ni har sett då den här processmodellen som jag beskrev för er tidigare, som ni kan klicka in och leka med, som finns på hemsidan. Då är det så att där finns en klocka, och klockan är 25e maj. Så att själva Sambruks-projektet fokuserar på det som händer innan 25e maj, för att få igång det här arbetet, för att kunna få den biten att fungera. Sen har det dykt upp, under resans gång naturligtvis, en utav de viktigaste frågorna är ju, men vänta nu om vi bygger upp ett sådant här metadata-lager och vi har hela den här genomlysningen fulltext, index och allt sådant. Borde vi inte kunna använda det här systemet då för när en medborgare kommer och vill veta vad har ni om mig. För det är ju det som de flesta kommuner är rädda för, den utmaningen. För att gå igenom alla typer utav system och även fillagringsytor för att ta reda på vilka förekomster, alltså var någonstans finns det information om just den här medborgaren. Och sen då naturligtvis förklara vilket laga stöd har vi att ha det. Och dessutom naturligtvis om det nu är så att vi har hittat information eller det finns, det kanske finns då en rättslig förpliktelse för oss att ha information. Det kan ju vara vad som helst, är det socialtjänsten så kan det ju vara en orosanmälan som kommer in och då har ju faktiskt den medborgaren inte rätt att få reda på den informationen, och så vidare. Så att det här är ju, det finns ju massor med sådana saker som kommer komma fram i nästa skede. Alltså när vi har passerat den 25e maj. Och det är ju rätt intressant för då har vi ju redan byggt upp, det är också därför det är så viktigt att iallafall utifrån mitt perspektiv att få in strukturerad data. Ett exempel som man skulle kunna titta på också, det som är så intressant här är att nu har vi byggt upp något som är helt systemoberoende. Och om vi nu säger såhär då att du i en kommun eller i vilken organisation som helst egentligen har ett jättestort gammalt pappersarkiv med massor av dokument, och om vi då säger att du skulle kunna jobba med digitalisering utav de pappersarkivet genom att skanna in, OCR läsa, alltså inte skanna in för bilder utan OCR läsa in det och göra om det till en pdfa-format. Så kan du ju sedan köra pdfa-formatet, alltså

			hela det arkivet, hela den fillagringsytan, med då ett antal pdfa-dokument. Så kan du köra det genom Aigine och gör samma typ utav klassificering som du gör på word-dokument eller vad som helst. Och ju mer sådana här grejer du gör desto mer får du kontroll över dina informationsmängder och det är ju också naturligtvis i nästa skede kan du ju använda det här för dels rensning naturligtvis utav data, därför du kanske också här kommer fram till att allting som vi bedömt som inte innehåller personuppgifter och som är då över en viss ålder, det kan vi ta bort. Det går inte trycka kontroll A och delete, men däremot kan du göra en sådan rensningsfunktion. Det kan ju också vara så att du gör en bedömning när du gör en granskning att här har vi för visso personuppgifter men vi kanske då ska göra någon form av maskningsfunktion, dvs vi ska maska viss data för att göra den tillgänglig. Det innebär ju inte att man gör en överstrykning med svart färg, eftersom det här någonting som ska göra tillgängligt elektroniskt. Då måste man hitta andra modeller och det är också en sådan sak som man kan lägga till som funktion. För du har ju väldigt mycket grundinformation kring det här.
2.21		LH	Om vi går till utmaningar, vad i projektet har varit utmanade?
2.22	CCG OLA		Oj, tid har varit en av de stora faktorerna. Att få saker och ting att hända, sen är det ju att gå från ord till handling, det är väll också en sådan sak. Men en utav det viktigaste bitarna, jag har ju jobbat väldigt mycket med, också att försöka få in medel till själva projektet, asså Sambruksprojektet, för att snurra det. Vilket innebär att jag har jagat medfinansiärer med ljus och lykta. Det har tagit en del tid, hade vi haft den financieringen klar till exempel från någon central aktör typ Vinova, Internetstiftelsen osv som jag även har naturligtvis även varit i kontakt med, hade det varit klart från början då hade det ju naturligtvis varit lite lättare. Då hade man kunnat ägna betydligt mer tid till övrig dokumentation och övrig kravspecificering vilket kanske innebär att vi hade klarat oss något tidigare, det vet jag inte, men det har varit en av utmaningarna. Sen är det ju självklart så att, det har ju stått och balanserat naturligtvis eftersom vi också har haft den här utmaningen kopplat till juridiken som jag var inne på. Asså, hur och vida vi kan ha/måste ha det här installerat lokalt. Den här tekniska utmaningen har varit stor. Jag kan säga då att det är ju tack vare Elinar då som jag ser att vi faktiskt har kommit dit. Men det har också gjort, det har skapat en möjlighet, och det är det som är så fascinerande, ett traditionellt AI-projekt går ju ofta ut på att du jobbar med, du börjar med att lära AI:et, du börjar med ett projekt där du ska gå igenom ett antal dokument och lära AI:et vad den ska bli duktig på egentligen. Vilket innebär att du bygger också då en dokumentation kring det här och det som du skapar där blir ju en intellektuell ägodel för just den organisationen som driver projektet, men det blir ju en proprietär information. Ta som exempel, en bank, vi säger Handelsbanken gör ett AI-projekt och så går de

			igenom då och skapar det här som jag med underlaget och så vidare och så vidare, som då blir det proprietära för Handelsbanken och sen jobbar man vidare med det och så AI:et bättre och bättre. Men du håller det fortfarande inom den organisationens väggar. Skulle sen någon annan stor bank, SEB banken till exempel göra ett liknande projekt, även med samma leverantör så blir det ett nytt lära AI projekt, därför man kommer ju inte kunna ta det av det som Handelsbanken lärde sig. Även om man har väldigt likartade verksamheter och det är ju det som vi brutit nu, vi gör ju det att alla som jobbar med systemet lär upp systemet löpande under arbetets gång. Men alla som använder systemet, oavsett om det är en kommun, landsting, statlig myndighet eller privat företag kommer att kunna ta del utav dem lärdomarna. Så vi gör alltså inget som är bransch unikt eller branschspecifikt på det sättet i lärande delen. Det blir ju också, vi har två saker som blir lite unika då för AI-projekt, dels att vi har en lokal AI-komponent som gör all genomlysning, all genomlysning, all bearbetning, alla listor, alla söklistor, alla beslutsunderlag, all dokumentation är lokalt hos respektive kommun. Men det som är själva anotationstekniken är den som vi delar med oss utav. Så det är ju andra delen som blir unik, eller andra och tredje som blir unika. Så det är dels ett nytt sätt hantera AI-lösningar men också så är det ett stöd som faktiskt ger, ger dig stöd utanför de här normala ramarna som du har och jobbar i systemets system eller stuprörets stuprör.
2.23		LH	Har ni upplevt några utmaningar specifikt med kognitiva system?
2.24	CLP	PM	Det är självklart vi har gjort det. Just det här som jag var inne på tidigare, jag menar maskinen kommer att tro att vissa saker som du, som finns skrivna i dokument är personuppgifter fast det inte är det. Just det här med att lära sig då att ibland kan ju ett förnamn vara ett efternamn och även tvärtom. Det kan ju också vara ett sådant lite förvirrande för en maskin att förstå att det inte två förnamn som kommer efter varandra utan det är faktiskt ett förnamn och ett efternamn som utgör ett namn till exempel. En sak som jag höll på att glömma är naturligtvis, det som det finns stöd för också som vi har lyft fram som har varit rätt så viktigt, det är ju att, om vi nu säger att vi har ett dokument så kan det ju vara som så att, vi pratar ju om presumtiva personuppgifter, ett dokument behöver ju inte bara handla om en person utan vi kommer alltså kunna lägga till flera personer. Vi kan markera upp den här informationen innehåller till person nummer 1, den här informationen till person nummer 2 och så vidare. Vilket gör att du kan få mycket bättre kontroll över vad du faktiskt har. Men det finns ju utmaningar, absolut. Jag tror säkert vi kommer komma fram till ännu fler och just det här med att hitta de olika förekomsterna tror jag är en av de sakerna som vi kommer se en utveckling på. Jag tror att det kommer bli bättre och bättre

			men vi kommer ha utmaningar inledningsvis, det är jag helt övertygad om. Sen är det väl så naturligtvis, en annan utmaningen kan vara en teknisk utmaning att man ser att oj, nu ska vi ha in någonting som gå igenom våra system och allting sånt där och det är väl en sådan sak som varit lite dialog runt omkring. Det som de flesta har varit oroliga för är naturligtvis att det här är en molntjänst, eftersom dessa lösningar brukar vara det.
2.25		LH	Om vi istället går in på möjligheter, vad ser du för möjligheter med kognitiva system?
2.26	OIE	PM	Ja det finns en vansinnig mängd möjligheter med. Om vi då säger med det jag berättade för er om hur min bakgrund har sett ut och vad jag har jobbat med och vad som har varit min drivkraft runt omkring just det här med informationsstruktur så ser jag möjligheter med att faktiskt göra med att få kontroll på informationen på ett sätt som varit näst intill omöjligt tidigare. Så gör det så att vi faktiskt kan börja titta på möjligheten till att faktiskt bygga gemensamma arkitekturkomponenter. Jag ser möjligheten att kunna få, det är en annan sak som jag ser som väldigt viktig i business process management, vanlig processorientering i en organisation så ser jag en av de viktigaste bitarna är ju inte vad man gör innan varje process steg utan det är ju vad man tar in och vad man lämnar efter sig, alltså input och output. Det är ju där det är det viktigaste att man kommer överens för det är det som gör att de olika stegen kan agera separat och man faktiskt kan få effekt utav de olika bitarna. Med den här typen av informationsklassificering, med den här typen av kontroll på den informationen och det behov av information som finns så kan vi göra våra processbeskrivningar betydligt bättre. Och med våra processbeskrivningar som är bättre beskrivna så har vi väldigt större möjlighet att effektivisera all vår verksamhet och då gör det att man kan då se till att de här, de som tidigare har varit ett slags mantra, att man då ska försöka digitalisera det som då lämpar sig att digitalisera medans det som krävs mankraft för ska man då fokusera på den mankraften. Här skapar vi förutsättningar för det, som vi då kan få tydligt var någonstans vi kan effektivisera med hjälp av digitalisering i tal om processer. Då gör vi någonting annat än att bara flytta in blanketter i datorn. Det gör ju också det att då kan man börja titta på mycket mycket mer på det som jag anser är en väldigt viktig del i processorientering, det är ju avvikelsehanteringen som i dagsläget väldigt ofta hanteras som en helt egen process, blir ju egentligen då innovationen i de befintliga processerna. Då har vi tid att jobba med effektivisering, då har vi tid att titta på vad är det vi är ute efter, hur ska vi kunna effektivisera verksamheten, hur ska vi kunna göra någonting. Om vi nu säger att vi tar fram förslag på olika gemensamma processbeskrivningar då kan man ju börja jobba kanske likartat och använda processbeskrivningar som kravspecifikationer för nya systemupphandlingar. Som

			gör då att, det handlar ju mer om att utifrån ett kommunalt perspektiv till exempel då med 290 organisationer som jobbar med lite grann samma typ ut av struktur, det är ju samma saker som görs inom respektive kommun. Kan man då enas om att man faktiskt gör saker enligt en specifik processbeskrivning då finns det en möjlighet att få bättre systemstöd och bättre gemensam struktur vilket innebär att man kan bygga också business intelligence system mellan, asså för kommuner då, där du kan jämföra resultaten med olika kommuner även om man använder olika system. Så att möjligheterna med just den här typen av hantering och den fortsatta hanteringen tror jag är hur stor som helst. Som sagt, en ut av tankarna som jag hade också tidigare är just det här med den här genomlysningen av strukturerade system och få fram då den klassificeringsmodellen där man ska kunna identifiera det här jag var inne på med det här betygssystemet, där man ser liksom att här uppe när vi agregerade på den här nivån så skulle man kunna publicera på det här sättet. Då kan man också föreslå ett gemensamt sätt att publicera, vilket innebär då att alla kommuner till exempel eller alla skolor skulle kunna publicera data på samma sätt enligt samma struktur. Det kan vi alltså få hjälp utav, det skulle ta ungefär två manveckor för en hyfsat duktig person att göra en sådan definition, alltså på en given datamängd. Det vet jag i alla fall, jag har jobbat med den typen av projekt kopplat till då, till exempel leverantörskontroller men då ska man få maskinerna att föreslå ett gemensamt sätt att publicera gemensam information och då kan vi också bygga upp det som skulle kunna kallas en gemensam informationsutbyteskatalog.
2.27		HM	Er lösning som ni nu har kommit fram till, hur kommer den förändra framtiden, alltså hur kommer ni utveckla den vidare?
2.28		PM	Jag ser ju det att det blir ju naturligtvis behovsstyrt. Men det som jag ser framför allt det är ju att, det handlar dels om att få in fler funktioner som även finns då visuellt i den här processbilden, vad som kan bli då i förvaltningsstegen, dels strukturerad data, dels att få in fler funktioner då som stöttar och som sagt återigen i första hand så är det ju kommunerna som vi har jobbat mot eftersom det här ett Sambruksprojekt från början så det är kommunerna som vi har kravställt för men det är också under resans gång som vi kommit fram till det inte är någon större skillnad för att alla har samma utmaningar inför GDPR. Men att man då underlättar det som kommer in, dvs förfrågningar som kommer in från allmänheten och hantering av registervård osv osv. Det finns, där finns det flera saker som kommer in då, som kommer kunna förädla egentligen systemet mer och mer.
2.29		HM	Jag tror vi känner oss klara där från vårt håll, om inte du känner att du har något speciellt att tillägga?

2.30		PM	Nä jag kan babbla på ett par timmar till om vi hade haft den tiden.

Appendix 3

Name: Ari Juntunen

Role in project: Project architect

Company: Elinar OY LTD

Time: 48:14

Date: 2018-05-14

Reference number	Code	Person	Questions and answers
3.1		LH	So, first of all is it okay if we record this interview so we can transcribe it and use it?
3.2		AJ	Sure, yes!
3.3		LH	Great! Do you want to be anonymous or is it okay that we use your name and your role?
3.4		AJ	You can use my name, no problem.
3.5		LH	Perfect, thanks! Okay, so can we start with, can you tell us a bit about your background and education?
3.6		AJ	Well, I started computer computer science in Helsinki University 1993 and I have still two courses to do, so that I would graduate myself. We founded Elinar in 1994 so I have been kind of busy ever since. My current title is CTO, I am essentially in charge of technologies and directions we take and also I deplus as a AI architect. I have very long background with content strategic processes, everything regarding unstructured data. Which means managing, archiving, capturing, content analytics, all of that. So that is in short.
3.7		HM	I was just wondering, were you part of the founders of Elinar?
3.8		AJ	Yes.
3.9		LH	Okay, so this project with Aigine, what has your role been?
3.10		AJ	Well, I have been in a kind of, some of our guys they are what one would call highly technical and the gentlemen of Aigine are pretty much on the other end of the spectrum. So, I have been translating Aigine's to tech talk and trying to translate tech talk to Aigine, so that they will have very high quality solution that fulfills the business they have. Obviously I was heavenly involved before we even

			heard about Aigine on Elinar AI development. And that comes back done to the unstructured because we are very much focused on the unstructured and a bit over three years ago we saw a lot of use cases for running AI that can make sense of unstructured data and we started developing Elinar AI for that purpose. Now we have developed technology that is also being the core of Aigine solution that allows us to understand unstructured data in similar fashion to what humans do.
3.11		HM	Could you tell us a bit about your involvement in the project?
3.12		AJ	Do you mean Aigine of Elinar AI?
3.13		HM	Aigine
3.14		AJ	Well, I have been, essentially what I said translating business to IT and IT to business but I have not been doing much besides being kind of a bridge in the project, because the only thing that Aigine actually needed was the couple of new things on UI, because we had the AI core in place already. And I have been developing the AI core, so there are other people that have been developing the functional stuff that Aigine needs for their solution. AI core has been completely untouched on this project, because there have been nothing kind of new that AI needs to do because it is flexible and basically when you train you feed in (oklart String-data) and properties we are going to extract and that have been extracted. But, most of the work in Aigine project have been on adapting the user experience in to the visional feature of Karl-Oskar.
3.15		HM	Am I right if I understand it that your AI miner is like the basis of there solution as well?
3.16		AJ	Yes. It is a base, of course there are other things like functioning but it needs like mouth to eat something and so on, but it is the brains.
3.17		HM	How many people have been involved in developing the solution, for Aigine then?
3.18			Now I need fingers, 7-8 personers besides me.
3.19			Are they mainly developers or?
3.20			Ofcourse there is a project manager involved but yes many developers. The roles we have are kind of focused. We have one guy who deals with UI, another who deals with data preparation and so on, and other ones who deal with IBM stored IQ which is part of the solution as well. And other people who are dealing with packaging of everything. One of the key points on this solution is that you need to to be able to deploy it into ours. And, creating a packaging

			and developing, to do that is of course also major task on a project like this.
3.21		HM	Could we say that been like the project leader from Elinar's side of view?
3.22		Aj	We have a project manager who does the daily stuff. I am missing that part of brain that allows me to manage projects. But, I have been project architect is probably a better term. I have been over-seeing the things that the team does and making sure that the pieces fit together. My brain does not do scheduling and tracking so we need other people to do that.
3.23		LH	Okay, could you tell us a bit how the Aigine solution works?
3.24		AJ	Yes, so basically we have three major components. One is IBM stored IQ, that is IBM primare data management product, that is data. It has capability to connect to many different data sources, crawl through them and index them. So our tool does not, and we decided very early that we are not building the capability to do the crawling and that stuff because there are already very high quality programs. So in this case IBM stored IQ and a IBM person connects stored IQ to the data sources, like share disks, sharepoints wherever you have possible GDPR data. Then Stored IQ crawls through them and creates an index. When the index is created there is an entity on Stored IQ infoset and it is basically, infoset is kind of the working when you do something on storedIQ you do it at infoset level, so nothing before that is comfortable so you have all this created infosets. It is a collection of files, so you create for example, I want everything from this data source first to be part of this infoset and then they done enhance of classify action using a special custom code that we have added to StoredIQ and then StoredIQ extract text layers from every single file, from intosets and sends them to our AI. Then our AI responds to if this is GDPR or not. About the same time we take the text layer and store it on our side. And this is now, the next part is where we have to do the major kind of UI development for Aigine. Because now we create essentially records of every single file that was broad and the contents GDPR according to our AI. But we can not know for sure if the client has legal basis for storing this file. For example, if somebody has a applied for a job at us, if he got the job according to law in Finland we have to store those, everything about that person for a 100 years. But if the were 20 people applying for a job, it only applies to the one that were selected, the other 19 people we have to remove this data in three months even without GDPR, according to corporate legislation. Out AI miner has no way of telling if the person who applied for a job and whos files we found. Actually, was or is employed by this company. So, we have records on AI miner side

			and then superuser can assign based on structure everything below this point is assigned to Ari but this project, then I go in and I see 40 000 files and I see that this is a big job, I need a couple of my project managers do deal with this projects. Then I assign all the projects for example to all project managers who have been running the project because they know the documentation, I do not know them. So essentially all the files get assigned to a person, that person then has to see it, they open the file, check it, if we have legal base for example municipality, there are a lots of documents, they have required to store them for a very long time even if they content is privacy data. Those are very quick and at some point it is possible to automate that but it does not damage them in its initial release. Because we, essentially we need some statical facts before we can automate this kind of decisions. For example, if we have certain areas and AI says that this is privacy data and the legal basis for example the functions of municipalities like, is planning or whatever it is. Nobody ever changed anything since we can start also make a decision to start our admitting that they had everything, as far as they looked decisions list and then system executes a script that removes all those files that we no longer have legal basis for. That is in short. And for each document our AI will generate a personal record for every single person that was found. And it also is into context so for example in a 500 page document I can exist 60 times with slightly different adoption. Because it is not just, you know, sometimes the decision cannot be made at the definite level, there might be persons of documents that we are no longer allow to keep and there could be portions of documents of just that document that we have to keep for hundred years or whatever. Which means that then these will have go back and edit those documents and remove those portions of we are not allowed by law to keep any more. And then the document is recrawled and he does the approval for the edited document, to keep it. This is not a one time exercise its something we start doing and keep doing essentially forever because we will per date need this documents created are unidentified and you have to sent compact. So the recrawling of all the data sources much faster computationally less intense because we keep, has coded for every single document so if document has not changed , but there has pre dececual then we are not doing the document preparation and analysis for the document again. But Aigine still might to need to approve it multiple times because in some cases, for example we could have emergencies, we could have sent document essentially to two cases and one case for example has resulted in official decision and the other one was canceled. For canceled cases we might not have the premises for this document but for the one if it becomes part of the official decision document for decision making kind of a document package then will have set for 7 or 8 years.

3.25		HM	Can I just confirm one thing, the actual like computer/processing power, that is cloud based, right?
3.26		AJ	Yes and no. The architecture is split, there are four virtual machines stored locally, three of them are StoredIQ and one of them is our AI miner. We do a lot of preprocessing computing at our AI miner and then for this case, the solution itself could say its cloud but what i would say is that it is private cloud. So essentially, itself its a dedicated AI for the municipalities, the instances is completely for them and the way we have implemented in our AI miner is that we are not actually sending anybody's private information to Elinar, at all. But yes, I would say that it is a private cloud because we are making dedicated instances and we will probably actually end up having several instances in Sweden.
3.27		HM	Do you know how the solution is going to be implemented in like the municipalities?
3.28		AJ	Yes, I believe I have a very good understanding with it. What do you want to know about it?
3.29		HM	I was just thinking like, I mean you almost answered it already, they have 4 virtual machines and then the cloud based, what do you call it, is it a computer or is it?
3.30		KB	Yes, its AI inferencing service. So essentially there will be multiple instances but in Sweden, but now its in Finland. Data volumes are going to be significant so once we understand. Many of these solutions will be hosted by the local companies like Tieto and Atea and CGI so most likely we will have to place the inferencing servers close by these data centers or within those data centers.
3.31		HM	Okay, so let's move on to challenges. What in this Aigine project would you say have been challenging?
3.32		AJ	For Elinar the most challenging thing has been the IT vs business, terminology and talk. Aigine is completely business talk, they do not have, they do not talk IT. And then we have this developer at the far end who do not talk any business.
3.33		LH	Could you repeat that, sorry?
3.34	CLP CUC	AJ	So, we have Peter on other side who talks pretty much business language. Then we have these developers who for example developed the UI and they do not talk any business they talk javascript, HTML5. So there are communication issues, which I have been trying to bridge with various levels of success. Sometimes I have been more successful than sometimes. That is probably the most challenging part. I we looked a little bit outside the actual project time

			at Aigine project, creating an AI kind of framework that can work reliably with privacy data, it obviously had its own conjunctions as well. Also one challenge on Aigine project that we been facing for a while is to obtain the training data from the municipalities for the AI. Because every time Aigine want to add something new, a new capability, a new thing that they want to take from the documents, we need to obtain the training data first. So AI needs to be trained, so that kind of for every time Aigine decides to add something to the common definitions we have to be able to obtain sufficient volume training data, otherwise AI won't have a clue about it.
3.35		HM	If we look at the technical challenges, it there anything you like to add?
3.36		AJ	Well, we had some coding issues which delayed one part of the project quite a bit in combining IBM StoredIQ and our AI miner. Because in this Aigine case, what we had already in place was that we were sending a text layer, happily to AI miner but in Aigine case we also needed the full half and filename for the file and actually being able to extract that from IBM storedIQ internals proved to be trivial.
3.37		LH	Okay, so you already mentioned the training data but have you experienced any other challenges with cognitive systems specifically?
3.38	CLP	AJ	You know, maybe I explain a bit how we apply either IBM Watson natural language understanding or in this case we have used our own on-prem miner. So, essentially what we do when we receive a file, or text layer file, first we need to split this into overlapping windows. Because business files can be very short one-pagers, not much text, or they can be 500 pages. So, we have to break the documents into appropriate size windows. Because when you develop AI, special requirements, it is not practical to send 500 pages on training, because challenge is we are pretty much limited on the GPU accelerators, that is 16GB. And in 16GB you have to fit in the model and personal training data that you are about to burn into a network. And the larger the network is, the less space you have for training data which makes the batch size smaller. And if we have very large documents it will take forever because we have to actually log every single document one by one to the card. So, we can process batches and the bigger batch the better. So, when we use windowing we can fill the GPU much more efficiently. So, first we split the documents into overlapping windows, because when it comes to privacy data, my details can exist in a document multiple times, but they don't span, a single record of me, of my private identification data, does not span for 10 pages. It could be all 10 pages about me, but the actual identification data that has some value for sensitive data for example, that is typically within a page

			or two. And there can be multiple records, multiple people within those pages. Now, when we split the document into windows, we also pre-process, we look at each word, and this is now the traditional text analytics way, which is in a sense bad if you only do it this way. We look at word, we see okay this is Peter, that's a first name. This is Mankenskiöld, this is last name. This is Ari, again it is first name. And here we seem to have a date, here we have an integer, here we have a phone number, here we have an e-mail address and so on. And then we have lots of what I call common words, so they are all lower case words, or start with capital letter, or they are all caps. And we have quite a few of these kind of typings for words. But if you use this kind of analysis only, and you try for example to pull out people, names, you get large quantity of false-positives. Because the problem is that GDPR is not local, which means that for example if a person living in UK has supplied something from the municipality, also that persons' data need to be extracted. And then we get into the situation where we have a lot of names that are names in one language and common words in other languages. So the larger dictionary you use for names, the more you will have false-positives. And that is actually, this situation was actually what caused Elinar to start developing AI, because we have been doing text analytics using IBM content analysis and Watson explorer, and Watson NLU, for a long time. But the challenge is that, sales for example, it is an English word but also a last name in UK, Mr. Sales. There is a person, I am not sure about the name, but it is on the name list, which means that a lot of English documents get annotated wrongly. Because every time there is a word sales, it is last name. And you generate.. This also happens with Swedish documents, happens with Finnish documents, happens with every single language. The larger dictionaries you are using, the larger you get false-positives. So this is now where the AI comes in. We feed this, the original text, and another layer that is just describing every single word, to recurrent neural network, that has multiple layers. And then, it will be able to tell, based on the samples that it has seen in the past, that this, this is actually privacy data. The sales here is not. But part of these.. this is a record, we have these 16 dates in the document, but only this one is a birth date.
3.39		LH	Okay, but is this accomplished through a specific dictionary you have created or is it through the learning capability of the cognitive system?
3.40		AJ	No no, this is the AI part. So, the preparation is done through the dictionaries. Data preparations, descriptions, the kind of the traditional language models do that. And we use Watson NLU or our custom tokenizer to create a description of the document or actually description of the window. Because we are working with splitting the document in windows. And then our AI takes the description

			text and tells, okay, on this, you have these three records, three persons mentioned, and here we have birth date, this is actually his home address.
3.41		LH	Alright, interesting.
3.42		AJ	And that is the key what our AI does. Obviously we also can, occasionally, determine even if it is not in a dictionary, or if it is written wrong. Like, I am working with another client and we have this.. Came up with a document with a Finnish lady, her first name is Paula and last name was, now I'm changing it because I don't want to disclose, but say Mäkinnen for example. But, the person who was writing the name did not have Ä. The person was sitting in Germany, so they do not have Ä. So, it turned out to be Makinnen, which is not on the list, because it was spelled wrong. This is especially when you are dealing with the.. It is a fairly common thing, that when you have these, you have one part of the name correct but the other part is incorrect because it is not on the list because it is missing characters or whatever letters you use. And the writer who create the document did not have access to a keyboard that could print those in the documents. So, what our AI does, it takes text as data, and humans annotates, Paula Mäkinnen, Makinnen in this case. Makinnen is not on the list, but based on the text around, it learns that very commonly, since this is described as a word that starts with a capital letter, this actually needs to be mapped in a privacy record. So it learns fairly quickly this kind of out-of-dictionary combinations, provided it can see samples of them.
3.43		HM	If we look at the opportunities instead. So, what opportunities do you see with the Aigine solution?
3.44		AJ	Well, I think most of the companies in Sweden, in Nordics and Europe will need a solution like this. Most companies need it. So the business opportunity is huge. And the way.. They did not open up how they are pricing their solution. Has Peter told you about the pricing?
3.45		LH	Well, we talked a lot with Peter..
3.46	CNS	AJ	So, you have an idea of the pricing structure. So it is fixed per month price essentially. Now, let's take a municipality that has terabytes of files. If you are using standard IBM pricing for StoredIQ for example, that is going to cost half a million euros just for StoredIQ licensing, actually a little bit over. But the model Aigine has now made, which IBM allows them to sell it substantially cheap. And this solution is kind of the door opening solution, essentially. Because license.. it is not for example.. They do not get the full access to all the StoredIQ capabilities. So, once they have bought this, we can start selling other value that comes from

			StoredIQ or related products to these clients, which means that the business opportunity is very, very big.
3.47		LH	Okay, if we look at cognitive systems specifically. Instead of business opportunities, what opportunities would you say comes with cognitive systems?
3.48		HM	Like, compared to regular, standard systems.
3.49	OIE	AJ	Well, one of the key things that we see is that.. Let's take standard text analytics for example. Like I said, it will generate a lot of false-positives, that is the way it is. Now when we are adding artificial intelligence, we are actually.. Because essentially, what we are doing is, we are replacing the decision-making that is made by humans. If either of you pull out a document in Swedish or even Polish, you can say from that document, looking through that document and say that this is privacy data most likely. You can do the call. Now, we are giving the computer similar capability to analyze things, to understand in this context that this is most likely privacy data or some other valuable business information. And that opens up.. Well, the problem is kind of that.. We can start doing things with the clients that have been undoable in the past. Because they would have required so much human effort, that this case would have been very difficult. And now we can create the structure, we can mimic decision-making or decision-making capability of humans in a large scale. Which means that we can for example.. We are working with unstructured data at the moment. The volume of unstructured data is exploding, and about 70% of the data companies have is unstructured. But, they can use currently only the 30% which is structured data in their solutions. For example what we did for a pension insurance company, was that, because in pension insurance business it is very bad if an individual dies early. Because it will cost a lot of money for the company where this person is working and also for the pension insurance company. And they had used IBM SSPS which is kind of an advanced analytics tool, to create a statistical model based on the structured data they had. But the problem was that the model was not accurate enough, it was not able to predict, because so much things were in unstructured data. Medical.. Doctors knows, medical certificates contain a lot of data about this individuals capability to work, and also background information if something has changed. They asked us; can you create an AI that will go through all the doctor certificates and tell if the person has met with a doctor, that something on their work has changed. For example like the location has changed or their time changed, for example person used to be working a lot of day shifts but now he is forced to do night shifts, and he is tired because of that. Or the boss changed, or it kind of changes. The people tell the doctor; okay now I am tired because of this. And we created an AI that were able to

			go through all the doctor certificates and tell there is no indication of any kind of change, or that there is this kind of change found. And being able to create this kind of structure on unstructured data, opens up extremely powerful possibilities to make the traditional business analytics much more accurate and much more powerful. Doing those manually for example for a pension insurance company, the hire a doctor to read through all six million certificates in their system, to go through every single certificate, and imagine how much time that will take. And now, this is the ultimate goal of cognitive. We can understand things in a similar fashion to humans. And for example, if there would be.. They would hire ten doctors to go through six million certificates, and it would take them two years. Complete waste of time, highly educated and skilled individuals. Now we can take one very good doctor, make him sit down for a week, and say okay, tell us, is there a change on this one? Is there a change on this one? And if there is a change, what kind of change? And then we take this expertise we have captured from specialists a make an AI that can repeat, repeat, repeat.. And the same thing applies to the Aigine project. We take expertise from the municipalities then we are removing the human having to understand and look at every single document, if there is any privacy data on this; no, okay, next one, no, next one, yes, now you have to look at it. This has privacy data.
3.50		HM	Alright, so if we go back to the Aigine project. What opportunities do you see for future development with that solution?
3.51	OFO	AJ	Well, there are a lot of them. Obviously.. but.. some of them they are core kind of business for Aigine. They are creating more and more capable common definitions. Right now it is focused on privacy data. But shortly we can for example create a system that will be able to abbreviate from any type of document...
3.52		LH	Could you repeat that last part?
3.53	OIE	AJ	In municipalities business you have to have.. You have to point that this document belongs to this business process and this decision-making step and so on.. So when you create a document and write it into a archive. The largest cost by far for every single municipality is not that you buy some licenses for the system, or pay out some money for someone to run the system, it is the time people spend trying to get data into the system, or trying to find data. With this kind of approach, we can for example eliminate the time to get documents into the systems almost completely. Because they are clever enough to say it contains these problems, this step, and these things were found from the documents. So instead of people having to type them in, we can direct this kind of information to somebody. So instead of environmental planning needing to have 20 people,

			they can streamline their document records keeping where they could keep 15, so they could put five to do something more valuable. You know, it is surprisingly long time that people spend in putting documents into some records keeping system. And obviously when we have an AI that can drain this kind of information, it is also so convenient to back this records keeping system. That is kind of the plan. Also, case management. Because pretty much everything in municipalities requires a case, you need to have a reason why you do something.
3.54		HM	Yes, I think that was our last question actually. Is there anything you would like to add to this?
3.55		AJ	No. But if you come up with more questions later just send me an e-mail and we will do another call.
3.56		HM	Yes. There were some part where we maybe did not understand everything, so maybe we will follow up with some questions about some areas in the interview, if that is okay?
3.57		AJ	Yes
3.58		LH	Perfect, thank you very much!
3.59		HM	Yes, thank you! Bye
3.60		AJ	Thank you! Bye

Appendix 4

Name: Karl-Oskar Brännström

Role in project: Business process manager and legal advisor

Company: Aigine AB

Time: 54:20

Date: 2018-05-16

Reference number	Code	Person	Questions and answers
4.1		LH	Är det okej att vi spelar in samtalet, så att vi kan transkribera och använda oss det lättare?
4.2		KB	Absolut.
4.3		LH	Perfekt. Vill du vara anonym eller kan vi använda ditt namn och din roll och så vidare?
4.4		KB	Jag behöver inte vara anonym men om ni använder mitt namn vill jag gärna kunna läsa igenom det innan ni publicerar.
4.5		LH	Absolut. Okej så, kan du börja med att berätta om din bakgrund och tidigare utbildning?
4.6		KB	Jag heter som sagt Karl-Oskar Brännström. Är någonting så ovanligt som en naturvetare som blev jurist. Jag läste juristlinjen på handelshögskolan i Göteborg, där jag också läste mitt sista år på Chalmers entreprenörsskola. Så jag har hoppat litegrann mellan institutionerna, inriktningarna, och efter genomgången utbildningen så började jag driva bolag. Först som ensamstående konsult i någonting som jag kallar strukturkapital, byggande av strukturkapital, alltså start-up jurist, eller start-up juridik, med stort fokus just på strukturkapital. Alltså sådant som inte bara är pengar i bolag utan mer rutiner och processer och avtal, metoder och vad det nu kan tänkas vara för någonting. Och hamnade i den rollen inne på IT branschen och har under närmare 15 års tid då drivit IT bolag i Sverige. Även ut på en global marknad, och har därifrån jobbat då med det som kallas business process management, alltså digitalisering utav verksamhetsprocesser, optimering och digitalisering av verksamhetsprocesser. Väldigt ofta med en touch utav juridik givetvis eftersom jag jobbat mycket med processer som kräver någon form utav compliance och har någon form utav grundlagsstöd. Så det är vad jag har gjort.

4.7		HM	Okej, så vi fortsätter med din roll i projektet då?
4.8		KB	Ja om vi tittar på rollen i Sambruksprojektet, så dels har jag haft en kontinuerlig kontakt med Peter Mankenskiöld som är projektledaren, och jag känner ju även människor inom Sambruk och kommun-Sverige eftersom jag har varit både leverantör till dom och är presumtiv leverantör. Jag har under ganska många år rest runt till Sveriges kommuner och pratat mycket om digitalisering och vad digitala processer egentligen innebär. Och har utifrån det också då konstaterat att du kan med digitalisering så kan du bara nå en viss nivå utav effektivisering, eftersom du fortfarande har mänskliga resurser som förväntas producera inom ramen för varje aktivitet. Och har därför då under de senaste åren tittat både på AI och på RPA. Och i den rollen, med den i bakgrunden så ville Peter att jag skulle kliva in och dels då svara för den juridiska delen, eller den juridiska hållbarheten utav den kravställningen som man tog fram i projektet. Men även att ta med det som är mina erfarenheter från att ha byggt digitala processer för en halv miljon människor, alltså hur får man saker och ting att sen fungera i en praktisk kontext. För att verktyg är alltid en sak, att få dem att funka är en helt annan sak. Vad är ni läser för något, vad läser ni för program?
4.9		LH	Vi läser en magister i informationssystem, och vi har en kandidat i systemvetenskap.
4.10		KB	Ja okej, men då vet jag att ni är systemvetare, att ni kommer från det hållet, ja, bra. Men av den anledningen så tog Peter med mig då i projektet för att titta på hur ska det här se ut, och det är väll det som jag har jobbat väldigt mycket med, att dels förstå hur de här lösningarna och hur informationsflöden fungerar, hur de identifierade produkterna med relevanta funktioner både konfigureras och presenteras för det som blir slutanvändare.
4.11		LH	Okej, kan du beskriva lite hur Sambruks, eller Aigines lösning fungerar?
4.12		KB	Det kan jag absolut göra. Nu är det ju viktigt att komma ihåg att det Sambruk gjorde, det var att de tog fram en kravspecifikation, och det är inom ramen för det som jag varit involverad i projektet. Alltså att ta reda på hur ska den här arkitekturen, eller hur måste den här arkitekturen se ut för att dels den ska vara juridiskt korrekt både utifrån GDPR men även sekretess- och offentlighetslagen. Och också då vad är det för kravställning som krävs på de ingående komponenterna och vad är det man måste bygga för värden för att det här ska vara rimligt att göra överhuvudtaget. Den kravspecifikationen har ju sen utvecklats inom ramen för Aigine och för att ta fram den faktiska lösningen. Så att Aigine och Sambruk är egentligen två olika saker där Aigine är den direkta fortästtningen på att

			realisera dom slutsatser som man kom fram till i Sambruksprojektet. Så frågan är om ni skriver om Sambruksprojektet eller om ni skriver om Aigine?
4.13		HM	Ja, alltså Aigine, både Aigine och Sambruk.. Aigine är ett företag, eller hur?
4.14		KB	Aigine är ett företag.
4.15		HM	Och det här projektet då?
4.16		KB	Projektet är ett projekt som syftade till att ta fram en kravspecifikation eller insikter om hur den här arkitekturen och vilken funktionalitet som krävs för att det här ska fungera överhuvudtaget. Och sen är Aigine bolaget som har byggt den, byggt produkten, eller byggt tjänsten utifrån då slutsatserna.
4.17		HM	Okej, och det är Aigine då som har beställt lösningen från Elinar?
4.18		KB	Det är Aigine som har beställt och bekostat och kravställt lösningen från Elinar, och även de delar som finns med från IBM i det här. Och det byggs också av mjukvara som inte har existerat tidigare överhuvudtaget som blir då bryggor mellan de olika produkterna för att få till fungerande digitala flöden i en organisation.
4.19		HM	Okej, jag vill bara reda ut en sak till. För vi pratade med Peter också, och det är kommunerna som finansierat det här, eller hur? Eller?
4.20	CCG	KB	Nja, kommunerna har till viss del finansierat kravställningen. Alltså det som är frågan; hur ska det här se ut, hur måste det fungera. Som ett konkret exempel när man konstaterar att AI-lösningen både från IBM men också från Microsoft och Google hela tiden bygger på att man tankar upp klartextdata till AI-lösningen som ligger i molnet. Och där har man väll konstaterat att inom ramen för projektet då att så kan man inte jobba eftersom det kan utifrån distriktsombudsmannen anses vara utlämnande av allmän handling. Kopplat då till problematiken att man givetvis kan hantera personuppgifter, men det är ju sammanhanget ett mindre problem än att vi bryter mot en svensk grundlag. Så att, första konstaterandet var att en traditionell AI-lösning går överhuvudtaget inte att använda inom offentlig sektor. Och det är ju en slutsats som man fick i projektet eftersom då den rättsutredningen är bekostad utav kommunerna, sen ligger det en investering på mellan 5-10 miljoner för framtagandet av tjänsten och den är helt privatfinansierad, den har inte kommunerna varit med och finansierat överhuvudtaget. Alltså för att få det här att fungera, att vi bygger om AI-funktionalitet på ett sådant sätt att det fungerar juridiskt vilket innebär att vi lägger annoteringskodningen lokalt, som gör att vi aldrig skickar någon data i klartext längre. Och det är

			ju en hel ombyggnation av hur AI fungerar, så att, och den har inte bekostats utav kommuner. Utan realiseringen utav dom slutsatserna är privatfinansierad.
4.21		LH	Okej. Men någonstans i helhet så är det ju hela projektet vi är intresserade av och framförallt hur AI liksom.. Alltså utmaningar och möjligheter med AI, hur det kan underlätta och så vidare.
4.22		KB	Ja men det förstår jag.
4.23		LH	Så det blir väll kanske lite mer åt Aigine då.
4.24		KB	Ja om ni vill komma i mål, eller alltså realiseringen är ju Aigine. Och det är ju där vi också har stött på.. Alltså att en kravspecifikation har konstaterat att, ja men såhär måste det fungera, det är ju en sak. Men att sedan få det att fungera på det viset när man då dessutom arbetar med befintliga produkter, för det var ju nästa grej. Vi konstaterade ganska tidigt i projektet att det finns ingen möjlighet att göra det här från scratch. För det är så avancerade funktioner, bara en sådan fråga om kompatibilitet åt olika datakällor, där vi konstaterade att StoredIQ som vi har använt oss av från IBM har stöd för över 400 typer av datakällor. Samma sak med filformat, att bygga någonting som kan läsa filformat när StoredIQ stödjer närmare tusen filformat. Att det måste byggas på befintliga programvaror för att det överhuvudtaget ska fungera både när det kommer till resurstid, men också med ledtid givetvis. Så att det konstaterades ganska tidigt att det här går att göra utifrån en kravspecifikation, för vi kan inte sitta och vänta 10 år på det utan vi måste hitta något smartare sätt att realisera funktionaliteten snabbare.
4.25		HM	Yes. Hur har ni då anpassat den här lösningen efter GDPR?
4.26	CCG CLP OCG OLA OIE OLA	KB	Ja, den absolut viktigaste insikten som jag nämnde, den har ju då bäring på hur data eller alltså data i klartext hanteras. Där vi väldigt tidigt konstaterade att data kan och får inte lämna myndigheten, i det här fallet kommunen. Det går alltså inte att.. Juridiskt så kan vi inte skicka ut data från en kommun upp till en molnleverantör. Och det här beror ju dels på att vi vet inte vad det är för innehåll i de dokument som vi skannar, och som vi behöver skanna. Vilket innebär att de presumtivt kan innehålla personuppgifter, och där är ju GDPR väldigt tydlig med hur man får lämna ut personuppgifter till tredje man. Utifrån det, den grundinställningnen att vi vet inte om det innehåller personuppgifter vilket innebär att skicka ut det utanför miljön det är ett brott mot GDPR, presumtivt brott mot GDPR redan där. Det är nummer ett. Men nummer två är eftersom man inte vet vad dokumenten innehåller så vet vi inte heller då huruvida det finns anledning för sekretessbeläggning över vissa delar av informationen. Och där, som jag nämnde så har ju JO varit väldigt tydligt med att man kan inte fatta bulkbeslut om utlämnande av offentlig

			handling utan varje akt som lämnas ut måste genomgå en så kallad och men- och sekretessprövning. Och där hamnar vi också i baktakt, för om man tankar upp allting till en AI-motor i molnet så har vi ju redan gjort utlämnandet, informationen har redan lämnat innan vi har gjort men- och sekretessprövningen, och särskilt i ljuset av transportstyrelse-skandalen så var det ett ganska snabbt konstaterande att så kan vi inte heller göra. Så den traditionella AI-lösningen med skanning i molnet fick vi helt enkelt revidera och konstatera, okej, vad kan vi göra för att säkerställa att det aldrig lämnar någon information utan bara data egentligen, eller egentligen kod. Vad är det vi behöver för att kunna få igång ett kognitivt lärande. Och det är ju därför vi använt oss utav Elinar som har erfarenhet av att bryta isär IBMs Watson-lösning i tidigare projekt. Så att där någonstans så stannade ju projektet i förhoppningen om att IBM skulle kunna realisera den kravställda lösningen, där IBM i princip sa nej tyvärr det kan vi inte, för vi har inte den kompetensen utan ni måste gå utanför våra väggar. Och då kom vi då i kontakt med Elinar. Så det vi konkret har gjort i det fallet, det är att annoteringskodningen, alltså kodningen utav annotering, annoteringskod som krävs för kognitivt lärande, den sker lokalt i kundens miljö. Och det är ju en komponent som normalt ligger i molnet, men den har vi då lagt lokalt och ligger tillsammans med Aigine-installationen. Vilket innebär att det vi skickar upp för det kognitiva lärandet, för den kognitiva bedömninen, det är ren kod, maskinkod kan vi säga, eller ja, annoteringskod. Som inte går att utläsa och där du inte kan få fram varken information eller personuppgifter. Och där vi då har tittat på hur den koden ser ut och vi har benchmarkat den mot lite folk i branschen och sagt okej, håller det här? Eller är er bedömning att det går att återskapa? Där vi har fått då kvittot på att, nej det här.. När man gör på det här viset så går det inte att återskapa. För du byter till exempel ut ord mot siffror som talar om att här har man ett substantiv i genitiv och vi har en stor bokstav, och ett ord på fyra bokstäver med stor bokstav, det är den typen utav kod som skickas. Så att det var ju en stor utmaning och har varit en stor utmaning. Det andra som vi har gjort inom ramen för Aigine handlar ju om det kontinuerliga lärandet, där vi konstaterade att traditionella AI-projekt är just projekt. Vilket innebär att man definierar en uppgift för AI:t och sen så sätter man upp en training session där man lägger stor mantid till att träna upp AI:s kognitiva förmåga och när de väl är upptränade så applicerar man dom på en specifik uppgift. Där vi konstaterade att det är ett väldigt stort stuprörstänk. Det vi var intresserade av, det var ju att.. De kognitiva förmågorna, de förbättrade kognitiva förmågorna skulle komma alla till del. Och det är ju också så som det fungerar då med annoteringskoden, det är att alla kunder som gör genomlysning, när den här annoteringskoden skapas så använder vi ju det faktiskt mänskliga bedömandet, de mänskliga bedömningarna lokalt hos kunden för att träna de kognitiva förmågorna. Som ett exempel, AI har sagt såhär att; jag tror att

			Upplands Väsby är ett personnamn för det har den karaktäristik som krävs för ett personnamn. Det är två ord efter varandra med stor begynnelsebokstav. Där en människa vid en granskning säger då; nej här har du fel. Det är ju precis den annoteringskoden vi lyfter upp då till molnet för att.. Och använder då för träning. För i nästa genomlysning så vet AI:t att nej, Upplands Väsby, det är inte en personuppgift. Och dom kognitiva förmågorna kommer alla till det, så att vi har tagit bort det som kallas training session, och vi använder istället det faktiska handläggararbetet i den här digitala processen. Den faktiska granskningen av AI:s resultat för den kognitiva träningen. Och det är ju också tämligen udda att man gör på det viset, utan som sagt normalt sätt är det seriellt, man har en training session och när man anser att det är tillräckligt bra då ersätter man människor med AI:t. Så gör inte vi här utan vi använder AI:t för att filtrera ut och minska arbetsbördan initialt, att snabba på hela den mänskliga granskningen som sen måste ske. Och det är vår absoluta bedömning, att den måste ske. Vi kan inte lita på att maskinen i det här läget har rätt. Men när vi hela tiden över tid då ser en förbättring över träffsäkerheten från AI:t, så vår förhoppning är att över tid så ska AI:s förmåga vara betydligt mycket bättre än den mänskliga förmågan. Men där är det ju som så att det räcker inte med att det är lika bra utan det måste vara otroligt mycket bättre för att vi ska våga lita på en maskin. Det är ju litegrann som diskussionen med självkörande bilar att dom är redan idag säkrare än människor men det räcker liksom inte, utan de måste vara så otroligt mycket bättre för att vi ska våga lita på maskiner.
4.27		LH	Ja, precis.
4.28	OLA	KB	Så det är ju också en väldigt annorlunda del som vi har gjort då inom ramen för Aigine vilket innebär att vi tränar de kognitiva förmågorna men vi har en iterativ process där de nya förmågorna hela tiden används i kundens miljö. Och appliceras på deras datamängd. Så vi hamnar ju i ett läge att om man ser AI:s genomlysning som en sil som ska fånga dokument som innehåller personuppgifter, så kommer vi över tid se att den fångar mer och mer. Att initialt så är den ganska.. Så är den inte så bra. Säg att den har en träffsäkerhet på 80% men hela tiden eftersom vi ger feedback på dom bedömningar som sker genom den mänskliga handläggarens vidimering och granskning av dokumentet så blir själva arbetet en training session. Så vi brukar säga, vi brukar använda analogin att istället för att pamflett med instruktion för någonting så försöker vi bygga kungliga svenska biblioteket, att samla kunskap istället och se till att den kunskapen blir tillgänglig för alla.
4.29		LH	Ja okej.

4.30		HM	Det är väll också så att alla som använder systemet kommer tillsammans bidra till att det blir bättre också?
4.31		KB	Det är precis det som är tanken. Och det är därför vi inom Aigine så, vi har ju nu presumtiva kunder som säger att; nej men vi vill inte dela med oss utav det där, utan vi vill få del utav det som tränas men vi vill inte dela med oss. Då har vi sagt tyvärr, det går liksom emot hela grundtanken. Tanken är att alla bidrar och alla får del.
4.32		LH	Ja, just det. Men okej, säljer ni då så att säga till samma.. Säljer ni bara till kommuner eller hur kommer den..?
4.33	CUC OCG OIE	KB	Nej nej, erbjudandet går ut till i princip alla organisationer som har behov av GDPR-granskning av sin ostrukturerade data, så det är kommuner, myndigheter, landsting, och det är framförallt hela den privata sektorn. Anledningen till att vi har ett kommunfokus det är ju att, dels att Sambruk tog det här initativet, men sen även det faktum att i jämförelse med privat sektor så har kommuner bättre tillgång till resurser som kan göra den här handläggningen. För det krävs en mänsklig granskning. Vi kan inte lita på att systemet har rätt så det krävs en mänsklig granskning och det krävs framförallt juridiskt beslutsfattande som ju då är kopplat till, har vi rätt att spara på den här personuppgifter? Och om vi har rätt att spara den, med vilken laglig grund har vi rätt att spara den? Och det är ju också någonting som vi då, ja, som en trainee som står bakom alla dessa som håller koll på då, okej, hur görs bedömning av laglig grund. Så AI:t föreslår, ger ju då både förslag på, okej, det här är vad vi tror är personuppgifter vilket då snabbar på genomläsningen av dokumenten. Men AI:t ger ju också förslag på laglig grund, eftersom vi även håller koll på att okej, den här typen av dokument det har vi tidigare sagt att det är en myndighetsutövning, ja då kan vi ge ett exempel på en myndighetsutövningen, att det här är socialtjänstlagen som är den myndighetsutövning och den laggrund som gör att vi har rätt att spara på personuppgifterna. Och det i sin tur får ju då effekten att de människor som gör granskningen kan vara andra typer utav människor. De behöver inte ha en kvalificerad juridisk utbildning, utan de får ett stöd i verktyget. Både då med AI:s förslag på laglig grund men också att vi bygger in en kunskapsdatabas och lyfter upp relevanta riktlinjer kontextuellt. Men det är ett erbjudande som går till alla organisationer globalt. Ungefär så kan man säga.
4.34		LH	Ja jag tänkte bara, för om ni sparar, eller vad man ska säga, annotationskoden. Är det.. Funkar det då i alla industrier och alla språk och så vidare?
4.35		KB	Ja det gör det. Annotationskoden fungerar som så att det ingår en information om vilket språk det är.

4.36		LH	Ja okej.
4.37	CLP	KB	Sen är det givetvis så att om vi har 100 kunder i Sverige som kör på stenhårt och sen har vi en kund i Frankrike så kommer ju AI:s förmåga att förstå den franska texten vara betydligt mycket sämre än den svenska texten. Och den kräver ju träning då för att förstå det kontextuella. Det där var ju också en sådan diskussion inom Aigine-projektet, inte inom Sambruks-projektet eftersom Sambruk har haft ett nationellt fokus. Där vi konstaterade ganska snabbt att det finns ingen anledning att bryta upp det i språk för det kan mycket väl vara så att du har en organisation som har dokument på både svenska och danska och engelska. Så att språkidentifieringen är väldigt viktig men det finns ingen anledning att applicera sig, att okej den här installationen gäller för Sverige. Eftersom svenska företag kan ha en majoritet av dokumenten på engelska.
4.38		HM	Okej. Skulle du kunna berätta lite om hur lösningen kommer att implementeras?
4.39		KB	Jo det kan jag göra, men jag är inte riktigt färdig än. Det finns en till väldigt viktig del i det här.
4.40		LH	Ja okej, kör.
4.41	CUC CCG OCG	KB	Det är ju då kopplat till Sambruks-projektet där vi konstaterade att det som kallas ostrukturerad data har ju tidigare varit oreglerat. Det är en sanning med modifikation. Det har varit reglerat men det har funnits en så kallad missbruksregel som gör att det har inte funnits någon anledning att adressera den ostrukturerade datan överhuvudtaget. Det som händer med GDPR är att den här missbruksregeln tas bort, vilket innebär att all typ utav data ska hanteras på samma sätt. Man gör ingen skillnad mellan databaser och fillagring, eller arkivskåp. Utan man säger att allting som går att teoretiskt söka igenom hamnar under GDPR. Även fysiska arkivskåp och det är väll också viktigt att komma ihåg. Och med den slutsatsen kan man då konstatera att så som vi hanterat våra databaser så har vi ju varit väldigt tydliga med att datakällor utgörs som en databas där vi kan ta reda på vilka tabeller som innehåller presumtiva personuppgifter. Problemet med ostrukturerad data är att vi har ingen aning om vad de innehåller för något. Det i sin tur gör att varje dokument är presumtivt en egen datakälla vilket innebär att den kan ha eller innehålla personuppgifter som har olika laglig grund för varje dokument. Och där innehåller GDPR en väldigt tydlig reglering kring registerförteckning, att man måste ha förteckning kring samtliga sina datakällor med information om vilka typer av personuppgifter de innehåller och framför allt vilken laglig grund man har för att spara på dem. Och det här är ju då direkt viktigt grundande på om man inte har en uppdaterad registerförteckning. Och det arbetet med att upprätta en registerförteckning med svensk databaser är ganska mycket jobb

			om man tittar på en myndighet som har ett par tusen databaser. Men det blir betydligt mycket större om man betänker att de har miljoner dokument där varje dokument presumtivt är en egen datakälla. Så att i projektet så konstaterade vi ganska tidigt att vi måste hitta ett sätt, om vi nu gör en genomlysning så måste vi bygga ett metadata lager som vi kan använda som registerförteckning. Det är ju också vad vi gör i Aigine, att vi markerar upp och klassificerar varje dokument, det räcker inte med att vi identifierar att här finns det personuppgifter utan vi tittar också då på, okej vad är den lagliga grunden för att spara, hur länge ska vi spara, vad är det för gallringsrutin som gäller och så vidare. Så det metadata lager vi bygger upp kopplat till varje dokument blir då den viss typ av registerförteckning som kravställs av GDPR. Och fördelen med det är ju då att vi i ett förvaltningsskede, när vi är klara med inventering ut av all den ostrukturerade datan är att vi kan använda det här metadata lagret för dem förvaltningsskyldigheter som kommer. Till exempel, registerutdrag eller rätten att bli bortglömd eller vad det nu kan tänkas vara för någonting. Så att istället för att göra en slagning på ett par miljoner dokument så kan vi en slagning i vår registerförteckning och se att ja du existerar på 73 olika platser med de här lagliga grunderna. Med den hanteringen kan vi också då se till att, det finns vissa lagliga grunder som gör att du inte bör tala om att du finns där, till exempel pågående brottsutredningar eller orosanmälningar inom socialtjänsten eller liknande. Du kan ju inte ringa till kommunen och säga att jag vill veta om det finns någon orosanmälan för det får dem inte lämna ut. Så med den registerförteckning vi bygger så kommer vi också kunna effektivisera ordentligt hanteringen ut av individuella rättigheter. Och inte minst det som är det absolut viktigaste, det är ju GDPR börjar gälla 25 maj sen så gäller den för tid och evighet, det är ungefär så man får räkna med. Där det inte längre räcker med att ha koll den 25 maj utan vi måste ha koll över tid och genom de kognitiva förmågorna och det metadata lager vi bygger så har vi möjlighet att monitorera våra datakällor och vår fillagringshistorik och så vidare. Har vi sagt att i den här katalogen så ligger det inga, finns det inga personuppgifter så därför har vi ingen behörighetshantering på den. Med hjälp utav metadata lagret och de kognitiva förmågorna så kommer vi kunna bevaka den typen utav eller alla typer av fillagringsutrustning så att hoppsan här kom det nu ett dokument som i strid med våra interna policys faktiskt innehåller personuppgifter och kunna lyfta upp det till en avvikelsehantering, att okej, det här dokumentet ska inte finnas här, vad ska vi göra, ska vi maska det, ska vi ta bort det eller ska vi flytta på det? Så att det metadata lager som vi bygger upp det behövs då för registerförteckningen med GDPR men i och med att vi skapar det så skapar det ett långsiktigt värde för kunden, att man kan använda det här i sin förvaltningen av GDPR. Så det var en annan sak som var väldigt viktig men en tredje sak, eller femte kanske det blir. Har ju med att göra med leveransen, vi använder oss utav Elinar AI-miner,

			vi använder oss utav StoredIQ, vi använder oss av en enklare workflow funktion, vi använder oss av IBM Watson som ligger i molnet. Vi har en rad komponenter som är extremt kompetenta. Vi fick då en tidsuppskattning att en installation av StoredIQ hos kund skulle ta någonstans mellan 80-120 timmar och då kunde vi kanske snabbt konstatera att det är helt omöjligt. Det kunderna är intresserade av idag är att få ordning på sin ostrukturerade data och man är inte alls intresserad ut av IT-projekt som drar 1000 wattimmar. Så en annan del i kravställningen har varit att minimera tröskeln för installation och komma igång och där vi har sagt att tjänsten ska vara så low-cut som det bara är möjligt man ska kunna leverera nytta på så få timmar som möjligt. Allting som riskerar att bli ett projekt läggs utanför, till exempel AD koppling vilket vi har stöd för rent teknisk men konstaterat att en AD koppling är ja, det kan gå fort det kan ta 10 timmar men det kan också ta 100 timmar. Där har vi sagt direkt att den typen ut av add on funktionalitet naturligtvis är väldigt bra men som skapar en osäkerhet och som tarvar resurser hos kunden måste hamna utanför och måste vara en option, vi kan inte bygga lösningen på att det krävs att det finns ett krav på att ha osäkerhet på timkonsumerande tjänster. Så där någonstans har ni vad vill uppnå för någonting, sen finns det en till del i det här och det handlar ju om min erfarenhet med digitala processer och användbarhet över huvud taget. Så att ett stort fokus har vi lagt på att vi måste ha ett gränssnitt som andas 2018 och som skapar ett förtroende hos slutanvändaren eftersom de kommer ju luta sig på den här, de kommer ju luta sig på AI:s bedömning. Och vi kommer inte ha tillgång till de här handläggarna och det är precis som att vi inte vill ha AI-projekt som drar iväg i tid, så vill vi inte ha ett krav på att vi måste genomföra på plats en utbildning med klicka här och klicka här och klicka här utan gränssnittet måste vara intuitivt. Och vi måste ha workflow, eftersom sitter du på ett par miljoner dokument så är det inte rimligt att en människa ska sitta och granska det utan det måste finnas en delegeringsfunktionalitet för att kunna sprida ut det här på flera ögon.
4.42		HM	Då går vi vidare. Du har ju tagit upp ganska mycket utmaningar som vi är intresserade utav också, har ni haft några tekniska utmaningar annars som du skulle vilja tillägga?
4.43		KB	Ja det har vi haft, stora tekniska utmaningar. Och det är ju kopplat till det jag har pratar om här, AI-projekt fungerar ju inte på det här viset. Vi har ju i stora delar fått bryta isär både Elinar AI-miner och IBM:s Watson för att förändra arkitekturen och flödet ut av information i de programvarorna. Likaså med StoredIQ har varit stora problem med att eftersom vi gör iterativa sökningar när de kognitiva förmågorna förbättras så vill vi att det ska göras en ny sökning och StoredIQ är en programvara som normalt vis sätts på per terabyte ut av skannad data.Vilket innebär att det har helt och hållet att det är

			användarstyrt, att en användare som ska mappa, en administratör som ska mappa upp datakällor och trycka på en knapp för att den där sökningen ska kunna ske. Och det har ju vi givetvis inte varit intresserade av, utan vi har ju varit intresserade av att när de kognitiva förmågorna har förbättras så pass mycket att det påkallat en ny sökning så ska det ske automatiskt och workflow motorn ska kliva in och delegera de träfflistor som finns där vi har fått nya träffar. De har också varit en stor teknisk utmaning, att få StoredIQ som då är en administrativ, som kräver en administratör som trycker på knappen och accepterar en kostnad på 150 000 per terabyte till att istället bli en tjänst som ligger och snurrar hos kunden. Det metadata lagret är också en sådan sak där vi konstaterar att metadatalagret måste ligga lokalt hos kunden men också kunna födas centralt ifrån, i och med att vi har kommunal rättsutveckling inom GDPR som det vi anser är relevant metadata kanske inte är relevant imorgon, vi måste kunna förändra metadata modeller centralt ifrån. Det har varit en utmaning. Användargränssnittet, med att få till en fungerande workflow med intuitiva användargränssnitt har varit en jätteutmaning. Det här är ju väldigt avancerade verktyg för välutbildade certifierade specialister där vi har velat bygga om det till att någon helt utan utbildning ska klara av att använda det. Sen har det ju varit en kommersiell utmaning och det är ju att förändra hela prismodellen för hur man tar betalt för de här tjänsterna. Som exempel som jag nämnde nu med StoredIQ de tar 150 000 per terabyte som skannas. Vi räknar med en kommun har ungefär 35 terabyte data som kommer behövas skannas 2-4 gånger per månad, vilket innebär att den prissättningsmodellen fungerar inte utan vi måste ha en helt annan prismodell. Elinars AI-miner prissatt per genomskannat dokument som kostar då fem kronor per dokument, som då inte är baserat på datamängd utan på antal dokument. Så en kommersiell utmaning har då varit att få alla dessa parter att acceptera en prismodell som är rimlig för kommuner och rimliga för det syfte som tjänsten har.
4.44		HM	Vad har ni kommit fram till då?
4.45		KB	Vi har en prissättning baserat på anställd. Och framför allt då typ av verksamhet, vi har konstaterat att kommuner och landsting oavsett om det är Sverige, Frankrike eller Tyskland har väldigt många hands-on arbetare som gör annat än att producera dokument. Så vi gör en differentierad prissättning så att kommuner och landsting betalar en lägre kostnad per anställd, myndigheter och företag som återanvänder och producerar många dokument per anställda har en högre prissättning per anställd. Men det utgår inte från mängden data eller antal dokument utan det utgår helt enkelt på storleken på din organisation eftersom vi har gjort presumtioner att ju fler anställda tjänstemän du har desto fler dokument producerar du.

4.46		LH	Tänker, om vi var på utmaningar och tänker specifikt på kognitiva system så nämnde du träningen och lärandet, har ni upplevt något mer än det?
4.47		KB	Nej det har vi inte men då ska du tänka att det är inte vi som sitter på de tekniska utmaningarna. Det här ju det som har lyft sig och som har problematiskt i projektet som har lett till förseningar och krismöten och sådär. Det är fullt möjligt att det har varit en rad andra tekniska problem kopplat till det här men det har nog snarare Elinar koll på. Hårdvaran har ju varit en utmaning, alltså i och med att vi inte, vi vet inte hur mycket data som kommer processas och vi vet inte hur ofta. Den här AI eller kognitiva lärandet snurrar ju på specifik hårdvara det går liksom inte bara att ha en vanlig serverfarm för det. Så dimensioneringen har varit en utmaning, vad är det vi ska räkna med, vad ska vi ha för serverfarm och för hårdvara för att kunna hantera tio kunder. Så där har vi helt enkelt gjort estimat och sen så vet vi att vi kanske får anledning att revidera det.
4.48		HM	Du har ju nästan svarat på det här, men vår fråga är i alla fall, hur kommer lösningen förändra kommunernas verksamhet?
4.49	OCG OIE	KB	Det är en svårt fråga för vi befinner oss i ett läge där kommuner inte adresserar det här problemet över huvud taget, det finns ingen kommun som har sjösatt ett projekt för att granska sin ostrukturerade data. Det är i sig uppseendeväckande eftersom det är ett frivillig avsteg ifrån lagen och det såg vi hos transportstyrelsen att det kanske inte är det bästa man gör. Men det förändrar det på så sätt att man faktiskt har en möjlighet att uppfylla kraven i GDPR, vilket vår bedömning från första början har varit att utan den här typen av hjälpmedel, där Aigine är en unik lösning, så finns det ingen tjänst att göra det här, det går helt enkelt inte. Det är för stora datamängder, det är för tidskonsumerande att det finns inget som kan få fram 500 manår i sin befintliga verksamhet. Då ska vi även betänka att de 500 manåren måste då om du inte har ett stöd måste då tas från juridiskt skolad personal, det måste i princip vara jurister som gör de här bedömningarna. Och 500 manår juristtid per kommun det går liksom inte, det hamnar i lådan omöjligt. Det är lite därför, jag brukar säga det att frågan om ostrukturerad data är som ozonhålet, om ni kommer ihåg det? Ozonhålet är liksom fortfarande där men det är ju inte det vi tänker på när vi vaknar på morgonen, så vi kan ju inte göra något åt det i vilket fall som helst, vi funderar istället på vilka skor vi ska ha. Eller om vi ska ta bilen eller åka elcykel för att göra vår insats för miljön men ozonhålet är fortfarande där. Det som Aigine presenterar och det Aigine förändrar det är ju möjligheten att göra någonting åt ozonhålet eller i det här fallet den ostrukturerade datan på ett sätt som faktiskt blir hanterbart dels då genom att, våra simuleringar visar på en effektiviseringgrad mot 97%. Det är ju fantastiskt men då går ju fortfarande från 500 manår till 1 manår och 1

			manår är ju fortfarande sjukt mycket. Det är fortfarande en gigantisk utmaning men då med den kopplat till att vi inte längre behöver juridiskt kvalificerad personal, vi behöver inte jurister som gör det eftersom vi har både systemet som faktiskt ger konkreta förslag på det lagliga stödet men också kunskapsdatabaser som kontextuellt tränar personalen och leder personalen genom det här arbetet. Där någonstans så blir det hanterbart.
4.50		LH	Vi har varit inne på det en hel del nu men om vi går vidare, bygger vidare på möjligheter och tänker vad ser du eller ni för möjligheter med just kognitiva system?
4.51	CLP OLA OFO	KB	Vi ser väldigt stora möjligheter och Aigine lösningen bygger vi ju nu för GDPR men vår målsättning är inte att stanna där. Den här arkitekturen med gemensamt lärande där alla får del av nya kognitiva förmågor är ju applicerbara på en rad andra områden där du har, för det första där du har stora datamängder men du har stora växande datamängder. Jag nämnde monitoreringen i nästa skede där vi inte befinner oss idag men vi bygger ju Aigine också för att kunna ha hand om monitoreringen ut av dina datakällor inom organisationen. Så där har vi ju inom bank och finans till exempel insiderlagstiftning där Aigine kan appliceras på precis samma sätt, alltså att identifiera insider information ordentligt utav kommunikation i din datamängd. Företagshemligheter precis samma sak, nationell säkerhet, alla typer av mönsterigenkänning men där vi inte då applicerar training session och sen att ersätta mänskliga resurser med ett halvt bra tränat AI utan bygga det på att vi hela tiden, vårt faktiska arbete används för gemensamt träna AI:et. Så arkitekturen är tänkt att kunna användas för i princip vad som helst. Inom skolväsendet, ni skriver magisteruppsats att kunna identifiera plagiat, det kan man applicera Aigine på direkt. Du har lärare som sitter och rättar och går igenom uppsatser hela tiden och då räcker det med att så här känner vi igen ett plagiat, det är den typen av mönsterigenkänning som Aigine kan göra, out of the box egentligen. Det är bara det att vi applicerar den på en annan typ av, fast egentligen ingen anna typ ut av datakälla utan vi har annan metadata för GDPR är ju intresserad av personuppgifter, för insiderinformation är vi intresserade av insiderinformation och för företagshemlighet är vi intresserade av företagshemlighet och så vidare, när det är då metadata som ligger till grund för mönsterigenkänningen. Så att det är ju breddningen horisontellt så att säga sen har vi ju den vertikala breddningen och det är ju att som jag nämnde med Aigines arkitektur så inom en tämligen snar framtid så kommer vi ha 10 miljoner granskade dokument, vilket innebär att de kognitiva förmågorna kommer vara i Aigine vara otroligt mycket skarpare än någon människa kan prestera. Så att det är givetvis så att, är det bara kommuner? Nej, vi börjar med kommuner så att vårt erbjudande till företag är att givetvis när vi har ett AI som bräcker den bästa tjänsteman i bedömning så kan vi plocka bort

			människan. Men det kräver att det tränas och tränas ordentligt, det räcker inte med tre veckors training session utan du vill redan ha 5-10 kanske 20 miljoner bedömningar.
4.52		HM	Om vi ser era framtida utvecklingsmöjligheter med er lösning, vad ser ni för möjligheter då?
4.53	OFO OCG OIE	KB	Om vi tittar specifikt för GDPR så är det dels hela förvaltningsfasen som har att göra med monitorering och individuella rättigheter, gallring, E-arkivs-frågor och så vidare. Där vi redan idag har på ritbordet men även handfast har börjat ta fram den typen ut av lösningar kopplat till metadatan vi bygger. Men sen är det ju bredden mot strukturerade datakällor. Aigine kan ju appliceras precis lika gärna på en databas för att hålla koll på att de tabeller som du har sagt ska vara fria från personuppgifter verkligen är det. Det är ett problem som egentligen är oadresserat idag, vi har valt att inte gå mot de strukturerade datakällorna av den anledningen att det redan är konkurrenssatt av alla systemleverantörer. Men det räcker ju med att du har en bifogad fil i en databas så har du ju skapat en ostrukturerad datakälla i din strukturerade datakälla och det är inte så många som har tänkt på det. Samma sak med, har du fritextfält så har du fritextfält så är det ostrukturerat, du har ingen aning om vad som kommer in där. Många sitter trygga med att här har vi ett fält för personnummer och det är maskat på det här sättet så vi vet exakt var det ligger någonstans och sen så har dem en fritextbeskrivning som gör att det som är strukturerat också i stora drag också håller ostrukturerad information. Det är ju inom ramen för GDPR men som sagt vi ser ju både vertikalt och horisontell vidgning av erbjudandet. Sen är det ju lite så att “sky is the limit”, vi har ju en försäljnings och distributionsstrategi som bygger på att vi inte själva ska leverera den här tjänsten utan att det ska göras utifrån de som sitter ute hos kund, alltså våra stora återförsäljare och leverantörer i Sverige och även globalt och det är väll vår förhoppning att det i kunddialogen ska kunna identifiera andra saker än det vi suttit och brainstormat och identifierat behov. Vi vet till exempel, med Aigine kan vi hantera våra fysiska arkiv genom inskanning till pdf så kan vi göra de här genomlysningarna även på fysiska arkiv. Det är ju en sådan kunddriven grej som vi inte hade tänkt på förrän kunden sa det, att vänta nu kan vi inte göra så här och så svar på frågan att jo det är klart vi kan. Ge oss en digital informationskälla så kan vi applicera hela Aigine tänket på det.
4.54		HM	Är projektet helt klart nu eller hur ligger ni till?
4.55		KB	Vi ligger väll till så att vi ska få leverans idag av sista versionen så förhoppningen är att vi börja kunna installera skarpt hos kund under nästa vecka.

4.56		HM	Ah okej, coolt!
4.57		KB	Ja! Det håller jag med om. Ni var med på att jag sa planen är att...
4.58		HM	Ja vi hörde från Peter att det har blivit lite förseningar och sådär?
4.59		KB	Ja det har blivit förseningar och förseningar speciellt av teknisk karaktär men framför allt handlar det mycket om användargränssnitt. Det är som jag sa, det här är mjukvara och leverantörer som är vana att jobba med högutbildade certifierade användare. Jag kommer från en verklighet där vi håller på med avvikelsehantering inom landstingen där du har undersköterskor som ska kunna använda gränssnitt som verktyg och det är den nivån jag anser man måste lägga sig på. Det måste bli ett stöd, det kan inte vara så att vi måste skicka in folk i en skolsal i 16 timmar för att de ska kunna klara av att använda verktyget. Det är det vi jobbar med, användargränssnitt.
4.60		HM	Det var faktiskt alla frågor vi hade. Är det någonting du vill tilllägga?
4.61		KB	Nej jag har nog sagt det mesta. Vi fick ju någon form ut av resa av vad Sambruk gjorde och vad vi har försökt åstadkomma med Aigine och vad är det för insikter vi har haft efter vägen och vilka utmaningar. Sen talar man om insikter så när vi började titta från ett Sambruksperspektiv så konstaterade vi att oj vi har en stor effektiviseringsmöjlighet här genom att filtrera bort saker och ting vi inte behöver granska men där tillägg ut av markerade ord i dokumenten, kunskapsdatabas, AI som ligger till grund för lagens stöd och liknande så har vi i Aigine projektet ökat både effektiviseringgraden i så många fler steg från ett digitalt processperspektiv. Vi har tittat på vilka moment som vi måste gå igenom och hur kan vi tillhandahålla funktionalitet för att effektivisera varje moment och varje steg.
4.62		HM	Ah precis, det är ju kanon!
4.63		KB	Då har ni lite grann.
4.64		HM	Ja, vi tackar så jättemycket för all information vi fick!

Appendix 5

Name: Hans Nyström

Role in project: Solution architect

Company: IBM AB

Time: 28:07

Date: 2018-05-17

Reference number	Code	Person	Questions and answers
5.1		LH	Kan du börja med att berätta lite om din bakgrund och tidigare utbildning?
5.2		HN	Jag har jobbat på IBM i 20 år nu i år och blev headhuntad från ett annat teknikföretag som hette Frontech år 98, där jobbade jag 3-4 år tror jag. Har jobbat i olika roller som konsult både på Frontech och på IBM, jag har jobbat i Europa-roller, jag har jobbat i en global roll som hade med några partners att göra och nu jobbar jag som IT-arkitekt då på mjukvarudelen av IBM. Mitt kundsegment är centrala myndigheter typ försäkringskassan, skatteverket och polisen och så. Där täcker jag alla våra mjukvaruprodukter, så jag är generalist, vi har ju hundratals produkter så det är ju svårt att vara specialist på alla. Och min utbildning är egentligen, Stockholms universitet, jag är humanist inte IT-tekniker, utan jag har läst psykologi och socialantropologi till exempel. Sen halkade jag in på IT-spåret via en polare då för 25-30 år sedan när internet började komma och bli stort. Så på den vägen är det, och det måste jag säga har varit en väldig nytta av, att kunna hantera människor i möten och i projekt och så. Så hantera människor och grupper av människor istället för att vara en djävel på C++ till exempel. Så det är min bakgrund.
5.3		HM	Vi gör då en case study på Sambruks projekt eller Aigine då, så vi tänkte vilken roll har du haft i det projektet?
5.4		HN	Just nu har jag ingen roll. Jag var med tidigt i diskussionen mellan Peter Mankenskiöld och Karl-Oskar och dem var tidiga och sa att vi har en idé om hur vi skulle kunna skannat Sambruks medlemmar, deras IT-infrastruktur efter ostrukturerad data och med hänsyftning på GDPR, hjälpa dem att städa upp. Det lät ju intressant, så då hamnade jag i, eller blev inkallad i det initiala mötet och fick lite idéer och tankar om det här. Vi kom väldigt bra överens och vi ville liksom åstadkomma samma sak och trodde på samma idé då. Så det

			blev några möten till och sen kom vi fram till att, okej hur ska ett sådant här system se ut, vilka grundkomponenter skulle behöva ingå för att möjliggöra den här processen som vi vill implementera. Då tog jag fram den första arkitekturen för systemet som nu Aigine säljer.
5.5		LH	Så du har mest varit del i framtagandet av kravspecifikationen mer än du har varit med i, vad ska man säga, realiseringen av Aigine?
5.6		HN	Ja precis, så kan man faktiskt säga. När det kom till programmering och faktisk implementering så vände vi oss till Elinar, vår finske partner och sa har ni tid och resurser att hjälpa Aigine med det här, och det hade dem. Jag var med i början där och sen har jag och Peter då och då, men jag har inte varit en aktiv del i projektet just nu, jag var med och satte grunden och sen...
5.7		HM	Okej, då ska vi se här om vi kanske ska anpassa intervjun lite...
5.8		HN	Haha, det blir en kort intervju.
5.10		HM	Nja, vår nästa fråga är, hur fungerar den faktiska lösningen? Jag vet inte riktigt hur delaktig du har varit i den processen.
5.11		LH	Du kanske kan beskriva vad som krävs för att lösa det här, hur har du tänkt att lösningen ska fungera blir det ju mer.
5.12	OLA	HN	Ja precis. Det bygger ju på att vi har en existerande programvara som är jätteduktig på att skanna filytor och system efter ostrukturerad information, dokument och så. Sen kan den highlighta massa grundläggande saker som filstorlekar, massa metadata information och så och sammanställer det på snyggt sätt. Den kan även leta i dokumenten efter typer av objekt som använder personnummer eller så. Problemet är om vi ska bygga då en lösning för 290 kommuner som ska använda det här så vill vi inte få, vi vill inte ha 290 systemadministratörer som sitter och gör exakt samma jobb utan vi vill ju att i takt med att lösningen används så ska den bli smartare och bättre på att identifiera känslig information och föreslå för administratören eller den som kör systemet att det här måste ni åtgärda. Det var det ena kravet då. Det andra är att när man har kört en skanning som man kör då dagligen eller en gång i veckan eller något, så det är ju ingen one-off grej som man kör inför den 25 nu utan det här ska ligga och snurra kontinuerligt för att säkerställa att man är GDPR-compliant även i framtiden. Då kan man ju inte ha en situation där systemet varje vecka varnar för samma dokument, om en människa har blivit varnad och tittat på doumenten och bedömt att nä men de här ser ut att vara felaktiga enligt GDPR, men de är okej därför att vi har en motivering till varför de finns där. Då vill vi undanta dem ur söklistan på ett bra sätt eller den typen av dokument eller kategorin eller den typen av innehåll när man har sagt att det

			här är godkänt då ska de inte dyka upp i nästa skanning. Och då är tricket då att hur externaliserar vi information om informationen på ett sätt så att vi inte själva bryter mot dataskyddslagen och så. Och det är där Elinar kommer in i bilden, där tappar jag tråden, då blev det implementation och där har inte jag varit med. Och ostrukturerad information den sekreteten berättade vi i första steget då. I nästa steg är det då strukturerad information.
5.13		HM	Om vi då går vidare. Vad i projektet har varit utmanande? Det som då du har varit med och hjälp till med.
5.14		HN	Vad som har varit utmanande?
5.15		HM	Yes.
5.16		HM	Vi tänker ju lite så här att, eller vår uppfattning är att det här är ett ganska unikt projekt och från er sida är det många som gör, många företag som gör liknande lösningar med IBM Watson?
5.17		HN	Utmaningen tror jag, det är inte specifikt på det här projektet att det är liksom någon. Det är ju tekniska svårigheter och det är saker som måste lösas, så är det ju alltid. Men möjligen kan jag säga att svårigheten det är att, det är någon form av inbyggt i innovationsprocessen på något vis, att dela med sig, att överföra en vision och få alla ingående i olika möten att få samförstånd och vart man vill och att kommunicera sin vision, det kan vara svårt. För att folk är väldigt snabba på att hoppa till slutsatser liksom att jaja den här typen av lösning pang, och så har man låst sig mentalt sig för det. Då kan det vara svårt att se den större bilden, det är jävligt flummigt men det är något vi, det är väll den största svårigheten när vi pratat om det här internt både på IBM och med andra. Att få dem att förstå vad det är vi vill göra.
5.18		HM	Skulle du säga att den här lösningen är unik eller är detta något som flera andra företag gör liknande lösningar, med IBM Watson och liknande?
5.19		HN	Det finns andra liknande lösningar globalt som bygger på Watson, det gör det. Och det finns garanterat, nu är jag inte påläst på liksom competition data men det finns garanterat andra leverantörer som gör liknande skanningar med hjälp av AI-funktioner, varierande grad beroende på språk och sådär. Men det som jag fastnade för med Peters och Karl-Oskars vision, det vi kom fram till det var att alla fokuserar på den 25 maj och GDPR deadline men de tänker mycket mycket längre och det här blir liksom en systemkomponent i alla företag och myndigheters systemflora. Det perspektivet känns unikt och det kan ju vara bara att marketing förenklar alla leverantörers budskap men det har varit liksom grundinställningen från början med det här projektet att det är ska säkerställa compliance

			långsiktigt, det är inte bara en skott vi kör nu för att pusta ut den 25 maj liksom och säga ja vi klara det. Det längre perspektivet tycker jag är viktigt.
5.20		LH	Om vi tänker lite mer tekniska utmaningar, vad är det egentligen som är utmanande med att använda kognitiva system? Det kan ju väldigt mycket men vad är utmaningen med det?
5.21	CLP	HN	Ja det finns ju flera problem. Det första som man tänker på är språket, det är ju en sak att ha ett system som förstår, till exempel Watson förstår de stora språken som engelska, spanska, franska och även japanska tror jag, det är väldigt bra. Som en liten marknad då som Sverige är, eller Norden för den delen. Så är det svårt, det är svårt att få ett fullt utvecklat språkstöd på systemnivå på sådan nivå därför att för Watson-ledningen så är Sverige bara ett litet land och ingen jättemarknad så det är inte värt att prioritera liksom. Men tekniskt så är det språkstödet och att kunna utnyttja de kognitiva tjänsterna fullt ut då måste man ha ett fullt språkstöd annars blir det inte kognitivt utan det blir mönsterigenkänning.
5.22		LH	Är det då alltså Watsons förmåga att förstå så att säga det svenska språket, eller är det mer användarnas förmåga att använda sig av Watson?
5.23		HN	Nej det är Watsons förmåga att förstå svenska språket. Nu visade det sig, och det har vi gjort i många andra fall och tester att vi har sagt att vi bryr oss inte om vi har något svenskt språkstöd eller inte utan vi kör ändå. Och då visade det sig att algoritmerna som är Watson klarar det ganska lysande. Så att vi har gjort chatbots därför och demos och grejer för både myndigheter och företag där Watson kan svara även utanför det som är faktiskt scriptat. Den förstår liksom kontexten och kan generera svar ändå.
5.24		HM	Ja okej, coolt.
5.25		HN	Ja, däremot så kan vi inte lämna support på.. Eftersom det inte finns något officiellt supportat språkstöd på svenska så blir det en svårare införsäljning därför att vi kan säga att det funkar, men vi kan inte garantera att det funkar.
5.25		LH	Ja okej.
5.26		HN	Den andra tekniska.. På tekniska utmaningar. Den andra tekniska utmaningen det är ju att förklara och förstå, få kunder att förstå, hur data skickas till IBMs molntjänster och hur det hanteras där. Det är många som förstår att när vi säger att vi lagrar inte data, utan vi får den data som kommer in, vi processar den, och vi skickar ett svar.

			Då är det alltid någon påläst som säger att, ja fast cachar ju informationen, eller åtminstone så cachas den i liksom något CPU minne någonstans under en kort tid. Och då kan man ju säga att den är lagrad. Och där.. Någonstans går det ju en gräns för vad man menar med lagring. Det är en svår diskussion. IBM garanterar ju att, om vi säger att vi inte lagrar informationen, då gör vi inte det, inte ens i cachar, och cachar vi den så är det något tidsintervall där vi rensar cachen. Jag har det inte i huvudet just nu men det är någon sekund eller något.
5.27		LH	Okej. Jag tänkte sen, du nämnde det här med inte bara inför GDPR utan även det framtida kontinuerliga användandet. Jag tänker liksom, hur kommer lösningen förändra i det här fallet kommunernas verksamhet? Om vi tänker lite möjligheter och förändring?
5.28		HN	Inför dataskyddsförordningen tror jag att det kommer att göra så att kommunerna tvingas tänka efter och se över sin datahantering. De vill inte vara i vetskapen om att de har långa listor med dokument som ligger på fel plats eller känsliga GDPR-synpunkter utan det kommer på sikt att styra upp datalagrings- processer och strukturer och att öka liksom medvetandegraden att det är viktigt hur vi hanterar data. Det är inte bara en wordfil vi lägger på en fileshare någonstans utan, det innehåller personinformation och det ska hanteras med respekt. Det är väll en effekt tror jag.
5.29		HM	Om vi går in lite mer på möjligheter då istället. Vad ser ni för möjligheter med att använda kognitiva system, som då IBM Watson?
5.30	OIE	HN	Oj, ja det finns ju allt från högt och lågt. Vi har.. När man jobbar med myndigheter, stor myndigheter som Arbetsförmedlingen och Försäkringskassan och även Polisen och Skatteverket så finns det tiotusentals handläggare som sitter och svarar på telefonsamtal och mail och skriver brev och även face-to-face möten med medborgare. Och allt egentligen som man kan göra för att effektivisera deras jobb och avlasta dom gör att handläggningen går fortare. Och då kan man säga att, det kan vara enkla saker som att en medborgare fyller i ett myndighetsformulär på webben och beskriver sin problemställning i en fritextruta. Där har vi implementerat Watson som sitter och lyssnar på det som skrivs in och kan då förstå vad ärendet handlar om och för-populera fält i webbformuläret, eller i ärendeformuläret. Det sparar kanske några sekunder bara men några sekunder gånger kanske 30-40 tusen handläggare totalt på de här myndigheterna, det blir tid till sist. Det är ett väldigt väldigt enkelt exempel. Men det behöver ju inte vara stora självtänkande system som löser alla världens problem åt alla människor utan det kan vara pyttesmå saker som gör att någonting går lite lite fortare, eller att handläggare får förslag på nästa åtgärd eller vilken regel som ska

			gälla, på ett enhetligt sätt över landet. För det är ett annat myndighetsproblem att man bedömer.. Saker i Skåne på Arbetsförmedlingen bedöms annorlunda än saker i.. Samma sak bedöms i Norrland därför att handläggarna har olika utbildning och erfarenhet och ålder och olika arbetsbelastning och sådant. Och det är inte rättssäkert. Så det vi kan göra för att säkerställa en likhet inför lagen liksom, det är värt det. Då blir det så att även små förbättringar kan bli väldigt väldigt viktiga. Det är inte jättesexigt men det är viktigt.
5.31		LH	Om vi tänker.. Vi har en fråga som är; vilka framtida utvecklingsmöjligheter ser ni med lösningen? Alltså finns det någonting, vi har pratat lite om det, men finns det någonting mer den kan göra? Något mer den kan lösa? Något som kan läggas till?
5.32	OFO	HN	Ja vi pratade ju om strukturerad data till exempel. Det är väll det exemplet jag har just nu. Sen.. På den här infrastrukturen så kan man säkert bygga massor med tidssparande och rättshöjande, rättssäkerhets-höjande, åtgärder. Utanför GDPR-scopet liksom. Man kan tänka sig att man ser en påminnelse av, om det står att en handläggare ska återkomma i något ärende står det i en fritext någonstans, i en anteckning. Då kan man tänka sig att systemet snappar upp det och påminner relevant handläggare om att kommer du ihåg att du skrev här den 17e maj att du skulle återkomma i augusti? Så det går att hitta på hur mycket som helst. Men jag har ingenting just nu. Det är nog mer, där ska ni nog prata med Peter och Karl-Oskar, de är säkert fulla med idéer.
5.33		HM	Ja vi har pratat med dom också.
5.34		LH	Ja precis, vi har pratat men vi tänker höra om det finns olika tankar från olika roller liksom.
5.35		HM	Jag tänkte på en annan fråga också. Hur ser du på liksom, framtiden inom kognitiva system? Tror du fler företag och organisationer kommer att använda sig av det eller, hur ser du på utvecklingen där?
5.36	OIE OCG	HN	Ja, det är ju en viss hype kring kognitiva system just nu. Delvis påeldad av IBM naturligtvis. Jag tror att om något år bara så är det ingenting vi pratar om egentligen utan det bara finns, det är självklart att vi använder kognitiva system. Det är nästa.. Att bara använda datorer för att göra saker snabbare, den övre gränsen på något vis har vi nått. Utan, med ökade datamängder och IT-baserad kommunikation mellan både människor och system och företag och alla aktörer så måste vi.. Så kommer det att bli självklart att någon extern kraft, en AI måste hjälpa oss sortera, eller analysera, eller föreslå åtgärder, eller någon förbättring. Vi kan inte bara göra saker snabbare liksom, utan vi måste göra det smartare också. Så därför kommer AI att bli en naturlig del i, i princip i alla system. Jag vet

			inte om ni tänker på det när ni använder Android telefoner eller Google-tjänster överhuvudtaget så dom vet ju vart vi är någonstans och så, vilken restaurang som ligger närmast eller. Det är en sådan, vi tänker inte ens på det längre och det är bara något år gammal, gamla funktioner. Och det gör livet enklare liksom. Så att det kommer att gå blixtsnabbt. Hypen kommer att gå över alldeles snart och det kommer att bli självklart att vi har de här systemen.
5.37		HM	Ja, det var faktiskt allting vi hade, alla frågor. Är det någonting du vill tillägga eller så?
5.38		LH	Om ämnet, eller projektet, eller lösningen?
5.39		HN	Nja bara en framtidsspaning som jag själv tycker är otroligt spännande. Det är att, jag vet inte om ni känner till IBM Q?
5.40		HM	IBM Q?
5.41		HN	Ja Q som bokstaven Q.
5.42		HM	Nää?
5.43		HN	Det finns, och det är också i sin linda just nu men det är kvantdatorer. Det är våran kvantdator. Det finns, man kan gå till IBM.com/q tror jag. Och där får man tillgång till, där kan man utföra beräkningar på vår kvantdator. Det finns ett API där. Och tänk er en värld av kvantdatorer och kompetenta kognitiva system. Det blir spännande.
5.44		HM	Ja verkligen.
5.45		LH	Det blir det.
5.46		HN	Och det kommer vi, eller jag fyller 50 i år så att.. Men vi kommer att hinna, jag kommer att hinna uppleva det under min livstid. Det blir väldigt spännande.
5.47		LH	Ja verkligen, kraftfullt.
5.48		HN	Mm, det var min avslutande tanke.
5.49		LH	Alright.
5.50		HM	Vi tackar så jättemycket för all information vi fick.
5.51		LH	Ja verkligen.
5.52		HM	Superschysst!
5.53		HN	Ja tack ska ni ha, tack hej!

5.54		LH	Hej hej!

References

Baars, Henning; Kemper, Hans-Georg (2008) "Management support with structured and unstructured data: An integrated business intelligence framework", *Journal of Information systems management*, No.25, Pages 132-148

Bhattacherjee, A. (2012). *Social Science Research: Principles, Methods, and Practices* (2 ed.): USF Tampa Bay Open Access Textbooks Collection.

Chen, Y., Argentinis, J. E., & Weber, G. (2016). IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics*, *38*(4), 688-701.

Chen, M., Herrera, F., & Hwang, K. (2018). Cognitive Computing: Architecture,Technologies and Intelligent Applications. *IEEE Access*, *6*, 19774-19783.

Chozas, A. C., Memeti, S., & Pllana, S. (2017). Using Cognitive Computing for Learning Parallel Programming: An IBM Watson Solution. *Procedia Computer Science*, *108*, 2121-2130.

Coccoli, M., & Maresca, P. (2018). Adopting Cognitive Computing Solutions in Healthcare. *Journal of e-Learning and Knowledge Society*, *14*(1).

Coccoli, M., Maresca, P., & Stanganelli, L. (2016). Cognitive computing in Education. *Journal of e-Learning and Knowledge Society*, *12*(2).

Datainspektionen (2018a). *The Personal Data Act.* [ONLINE] Available at: *https://www.datainspektionen.se/in-english/legislation/the-personal-data-act/*. [Accessed 6 April 2018]

Datainspektionen (2018b). *Samma regler för alla*. [ONLINE] Available at: https://www.datainspektionen.se/lagar--regler/dataskyddsforordningen/samma-regler-for-alla/. [Accessed 30 May 2018].

Demirkan, H., Earley, S., & Harmon, R. (2017). Cognitive Computing. *IT Professional*, *19*(4), 16-20. doi:10.1109/MITP.2017.3051332

Demirkan, H., Spohrer, J., & Welser, J. (2016). Digital Innovation and Strategic Transformation. *IT Professional*, *18*(6), 14-18. doi:10.1109/MITP.2016.115

Earley, S. (2015). Executive roundtable series: machine learning and cognitive computing. *IT Professional*, *17*(4), 56-60.

Finch, G., Goehring, B., & Marshall, A. (2017). The enticing promise of cognitive computing: high-value functional efficiencies and innovative enterprise capabilities. *Strategy & Leadership*, *45*(6), 26-33.

Flick, U. (2007). *Managing quality in qualitative research*. Thousand Oaks, Calif. : Sage Publications, 2007.

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

Gutierrez-Garcia, J. O., & López-Neri, E. (2015). Cognitive computing: A brief survey and open research challenges. In *Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), 2015 3rd International Conference* on (pp. 328-333). IEEE.

Hurwitz, J., Kaufman, M., & Bowles, A. (2015). *Cognitive computing and big data analytics*. John Wiley & Sons.

Jones, M. T. (2017). *A beginner's guide to artificial intelligence, machine learning, and cognitive computing*. [ONLINE] Available at: https://www.ibm.com/developerworks/library/cc-beginner-guide-machine-learning-ai-cognitive/index.html [Accessed 6 April 2018].

Kelly III, J., & Hamm, S. (2013). *Smart Machines: IBMs Watson and the Era of Cognitive Computing.* Columbia University Press.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing.* California, US: SAGE, 230-43.

Lee, J., Kim, G., Yoo, J., Jung, C., Kim, M., & Yoon, S. (2016). Training IBM Watson using Automatically Generated Question-Answer Pairs. *arXiv preprint arXiv*:1611.03932.

Mertens, J. (2017). IBM Study: More than Half of CHROs See Cognitive Computing as a Disruptive Force in the Next Three Years. *Workforce Solutions Review*, *8*(3), 27-31.

Murtaza, S. S., Lak, P., Bener, A., & Pischdotchian, A. (2016, January). How to effectively train IBM Watson: Classroom experience. *In System Sciences (HICSS), 2016 49th Hawaii International Conference* on (pp. 1663-1670). IEEE.

Noor, A. K. (2015). Potential of cognitive computing and cognitive systems. *Open Engineering*, 5(1), 75-88.

Produktbulletin (2018). *Inventering, filtering, delegering, granskning, klassificering och dokumentation av personuppgifter i ostrukturerad data med hjälp av gemensamt kognitivt lärande*. [Online]. Available at: http://aigine.se/files/Aigine_Productbulletin_EC_swedish.pdf [Accessed 10 May 2018].

Recker, J. (2013). *Scientific research in information systems : a beginner's guide.* Berlin ; New York : Springer, cop. 2013.

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC

(General Data Protection Regulation) Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679 [Accessed 6 Apr. 2018]

Rizkallah, Juliette. (2017). *The Big (Unstructured) Data Problem.* [ONLINE] Available at: https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#16ddb612493a
[Accessed 23 May 2018]

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., & Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv*:1603.06807.

SFS 1998:204. Personuppgiftslag. Stockholm: Justitiedepartementet

Sheth, A. (2016). Internet of Things to Smart IoT Through Semantic, Cognitive, and Perceptual Computing. *IEEE Intelligent Systems, 31*(2), 108-112. doi:10.1109/MIS.2016.34

Tarafdar, M., Beath, C., & Ross, J. (2017). ENTERPRISE COGNITIVE COMPUTING APPLICATIONS: OPPORTUNITIES AND CHALLENGES. *IT Professional*, doi:10.1109/MITP.2017.265111150

Valiant, L. G. (1995, October). Cognitive computation. In *ANNUAL SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE* (Vol. 36, pp. 2-5). IEEE Computer Society Press.

Watson, H. (2017). The Cognitive Decision-Support Generation. *Business Intelligence Journal*, 22(2).

Yin, R. K. (2009). *Case study research and applications: Design and methods.* 4th ed. Sage publications.