What Can Be Learnt from Experienced Data Scientists? A Case Study

Leah Riungu-Kalliosaari[1], Marjo Kauppinen[2], and Tomi Männistö[1(✉)]

[1] University of Helsinki, Helsinki, Finland
{leah.riungu-kalliosaari,tomi.mannisto}@helsinki.fi
[2] Aalto University, Espoo, Finland
marjo.kauppinen@aalto.fi

Abstract. Data science has the potential to create value and deep customer insight for service and software engineering. Companies are increasingly applying data science to support their service and software development practices. The goal of our research was to investigate how data science can be applied in software development organisations. We conducted a qualitative case study with an industrial partner. We collected data through a workshop, focus group interview and feedback session. This paper presents the data science process recommended by experienced data scientists and describes the key characteristics of the process, i.e., agility and continuous learning. We also report the challenges experienced while applying the data science process in customer projects. For example, the data scientists highlighted that it is challenging to identify an essential problem and ensure that the results will be utilised. Our findings indicate that it is important to put in place an agile, iterative data science process that supports continuous learning while focusing on a real business problem to be solved. In addition, the application of data science can be demanding and requires skills for addressing human and organisational issues.

Keywords: Data science · Software development · Service engineering

1 Introduction

Data science is defined as "a new interdisciplinary field that synthesises and builds on statistics, informatics, computing, communication, management and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology" [4]. The interdisciplinary nature implies that knowledge from different fields is needed in order to ensure successful outcomes, making data scientists valued members of teams in many different fields. In particular, there is a growth in the application of data science in software engineering [3]. For example, in 2015, Microsoft grew its 'data and applied science' discipline

M. Felderer et al. (Eds.): PROFES 2017, LNCS 10611, pp. 55–70, 2017.
https://doi.org/10.1007/978-3-319-69926-4_5

to over six hundred people and more than 1600 people were interested in data science work and signed up to data science related mailing lists [10].

Five years ago, Davenport and Patil [7] described the data scientist position as the sexiest job of the 21st century. In the recent past, the data scientist role has grown in both popularity and demand. However, there is a wide shortage of data scientist despite an increasing need for them across many fields [7]. In order to fill the growing gap, education institutions are also making efforts in educating future data scientists [14].

In order for data scientists to add the most value, they must be part of a team that encourages them to 'innovate with customer-facing products and services and not just to create reports and presentations' [7]. As part of a large Finnish research programme Need for Speed[1], we wanted to understand how data science can enable organizations to gain deep customer insight. We conducted a case study with one of the project partners whose data science team was involved in service and software development projects. We wanted to understand the activities involved in the data science projects along with the challenges associated with them. Hence, we focused on these research questions: (1) What are the key characteristics of the data science process applied in service and software development projects? and (2) What are the challenges of applying the data science process in the projects?

We present the results of the study in this paper. We found the data science process to be an agile, end-to-end and continuous learning process. We classified the challenges into three groups: (1) the demanding problems, e.g., difficulties in identifying relevant problems and measuring the impact of the results; (2) moderate problems e.g. unrealistic customer expectations; (3) mild problems such as poor data quality and differences in modelling and production technologies.

The rest of the paper is as follows: Sect. 2 takes a look at related research; Sect. 3 presents the research process; Sect. 4 presents the results as lessons learnt; Sect. 5 discusses the results and Sect. 6 concludes the paper.

2 Related Work

As data science continues to gain more prevalence in software engineering, so does the role of data scientists within organisations. The role and job titles of data scientists can vary greatly in practice. Kandel et al. [9] conducted interviews with 35 data analysts from 25 organisations, and they identified three analyst archetypes: hackers, scripters and application users. *Hackers* were proficient programmers and comfortable manipulating data. *Scipters* were experts in modeling and producing visualizations with software packages such as R and Matlab. *Application users* worked with smaller data sets using application such as SAS and SPSS.

[1] http://n4s.fi.

More recently, Kim et al. [10] identified five emerging roles of data scientists in software development teams:

(1) "*Insight Providers*, who work with engineers to collect the data needed to inform decisions that managers make;"
(2) "*Modelling Specialists*, who use their machine learning expertise to build predictive models";
(3) "*Platform Builders*, who create data platforms, balancing both engineering and data analysis concerns;"
(4) "*Polymaths*, who do all data science activities themselves;"
(5) "*Team Leaders*, who run teams of data scientists and spread best practices."

Data science has the potential to improve software engineering in many ways. Begel and Zimmermann [1] surveyed the areas in which software engineers desired input from data scientists. They found 12 potential areas where data science could be applied namely, bug measurements, development practices, development best practices, testing practices, evaluating quality, services related to cloud computing and continuous delivery, customers and requirements, software development lifecycle, software development process, productivity, teams and collaboration, and reuse and shared components.

Handling of data and producing results involves different activities. These may include tasks such as discovering the data for analysis, wrangling or manipulating the data into an appropriate format, profiling data to ensure its quality and suitability for analysis, modelling the data, and reporting the results of the analysis [9]. Similarly, according to Fisher et al. [8], the analysis process may include five activities, i.e., acquiring data, choosing an architecture, shaping the data into the architecture, writing an editing code, and reflecting and iterating on the results. All these activities have challenges that can make data analysis an exhausting process.

Some of the existing challenges include data access restrictions, data quality issues, i.e., missing, incorrect or inconsistent data values, difficulties with identifying data sources and integrating data from multiple sources, problems with inferring the most important data while creating models and visualizations, and communication issues, e.g., while presenting the results [8,9].

The presence of data everywhere has led to a rapid growth of the data science field. Data-driven decision making is becoming increasingly critical while addressing different information needs in the software domain [3]. Critical and careful analysis of the problems should be practised in order to effectively apply data science interventions. As the goal in such interventions is not primarily to analyse data, but make the data useful for decision-making in relation to the business processes. It is of importance to consider the problems from a wider perspective than, e.g., data analytics only. Hence, our focus is on the data science process, i.e., the activities and tasks carried out while analysing data to produce actionable insights and outcomes.

3 Research Process

We conducted a qualitative study with experienced data scientists to understand their data science process along with its challenges (see Table 1 for an overview of our research process). We use the term 'experienced data scientist' because the participants had each been involved in data science or analytics type of work for 4–12 years (see Table 2). Despite the experience of the data scientists themselves, the team in question was new and worked on newly started data science projects. The data scientists were employees of an industrial partner Reaktor[2] in the Need for Speed programme. The industrial partner has 400 employees spread out in 4 offices across 3 continents. The company provides consultancy services in different areas with a connection to digital products and services. The data science team was composed of seven people.

At the beginning of the Need for Speed programme, the industrial partner hosted a workshop where its data science process was presented and discussed (Phase I, Table 1). After the workshop, collaboration between the researchers and the company was agreed upon. In addition, the presentation material was compared and linked with the findings from the focus group interview.

Our primary unit of analysis was the data science team. The work of the team was concretely characterised by examples from case projects. In addition, the informants also described the work of the team beyond the case projects.

Table 1. Research process

Phase	Theme	Method	Data	Informants
I	Overview of data science process	Workshop: presentations, discussions	5 slidesets	DS1, DS3, DS4
II	Characteristics and challenges related to data science process	Focus group interview	Audio recording, Post-it pictures	DS1, DS2, DS3, DS4
III	Validation of analytic interpretations (for Phase II), current situation	Feedback session, group interview	Slides, audio recording	DS1, DS4, Research manager

Next, we carried out a focus group interview (Phase II, Table 1). We chose the focus group method because it is suitable for gathering experiences and discovering new insights as well as allowing an in-depth discussion within a reasonable period of time [11,12]. The goal of the focus group was to know more about the data science process in the organization. The themes of the focus group included individual introductions, the company, the data science team,

[2] http://reaktor.com.

skills of a good data scientist, example projects, and lessons learnt (including challenges and success factors). Four researchers and four data scientists were present during the focus group interview. One researcher acted as the moderator and the others took notes and asked clarifying questions. The focus group was audio recorded and later transcribed for analysis. Details of the data scientists and the projects they had worked or were working on are shown in Table 2.

Table 2. Details of focus group participants

Participant	Background	Experience (years)	Examples of customer projects
DS1	Theoretical physics, data mining	12	Personalisation, optimisation; make predictions
DS2	Machine learning, CS, statistics	4	Change detection, make recommendations, produce more tailored advertisements
DS3	Machine learning, statistics	8	Marketing campaigns, make recommendations, location analysis
DS4	Psychology, IS, machine learning	11	Segmentation; make recommendations, improve revenue and user experience

The data scientists were given post-it notes where they wrote notes related to the discussed themes. The post-it notes were collected, placed on a white board and a picture was taken that would be used to support the analysis.

After the analysis, we held a two-hour feedback workshop session (Phase III, Table 1). Regarding the research process and its validation strategy, the feedback session also acted as member checking [5]. The goal was to present the results of the analysis from the focus group session and get feedback from the data scientists. The company's research manager, two most experienced data scientists (DS1 and DS4), and three researchers were present during the feedback session. The feedback session was also audio recorded and transcribed for analysis.

We analysed the data iteratively using the thematic analysis approach [6]. To guide our analysis, we used the pre-existing themes of interest discussed in the focus group interview, i.e., key characteristics of the data science process, challenges, success factors, example projects, and skills of a good data scientist. We iterated and refined the codes as we discussed with each other during the analysis as well as after the feedback session. We also used material obtained from the company to supplement our analysis, e.g., presentation slides. In this paper, we present the analysed themes related to the data science process, its characteristics and challenges.

4 Lessons Learnt

4.1 Data Science Process

The organisation had defined a data science process. During a Need for Speed programme workshop, the organisation presented the data science process on a high abstraction level. During the focus group and feedback sessions, the study participants provided more details about the process composed of six steps (Fig. 1): conceptualization, problem definition, data collection and preparation, modelling, evaluation and validation, and deployment and utilization of results.

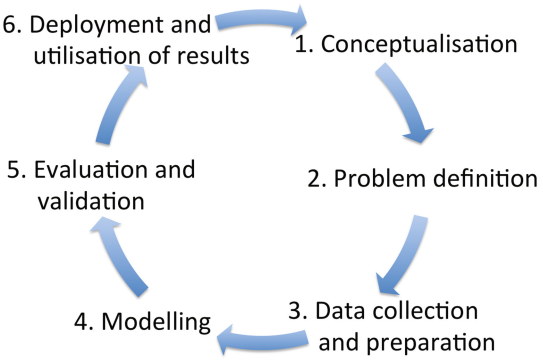

Fig. 1. Overall data science process of the case company.

Conceptualization: The main focus of this activity is the business problem. This involves interacting with the customer in order to assess the customer's understanding of (1) the business problem and (2) data science as a solution to the business problem. The business problem should be described clearly, putting the business targets and constraints into consideration, so as to develop the appropriate solution. The data scientists stressed the importance of knowing the customer's understanding of data science because it helped in preparing to address different customer expectations. One participant emphasised this:

It's important [for the customer] to understand the possibilities and limitations, really understanding what you are able to do and not do with data science. [DS3]

Problem definition: This activity focuses on the data science solution to the identified business problem. The business problem is formalised into an analytically solvable problem. One data scientists explained that many customers needed help to '*translate the [business] problem into a computational or mathematical problem'* [DS1]. Successful problem definition therefore calls for a lot of interaction between the customer and the data scientist.

A good data science solution starts by understanding who the customer or end-user is. This helps to know how the data science solution will be applied.

With this knowledge, the data scientists said that it was the best way to provide an optimum solution.

Data collection and preparation: The end result is determined by the data at hand. Hence, this makes collecting the data and preparing it for efficient use a vital aspect. In order to make this a fruitful endeavour, the data scientists wished that not only would the data be handed over to them, but that they would also be granted access to the actual data collection process. This would grant them the opportunity to improve the data collection process which they believed would have significant impact on the results.

Modelling: When the data is in good shape for analysis, the data scientists then manipulate the data using different data analysis and modelling techniques. Depending on the problem, modelling aims at describing what has happened, diagnosing why something has happened, predicting what will happen or providing guidance on how to make something happen. Often, the models are demonstrated using visualisations.

Evaluation and validation: The data scientists need to provide results that are reliable and relevant to the business problem. The participants were very interested in knowing the effectiveness of their results and therefore desired to obtain feedback from the real end users, not just from the business stakeholders or domain experts.

Deployment and utilization of results: It is essential that the results are put into use so as to assess their impact. Continuous and consistent monitoring is imperative along with a feedback loop that enables the end users to communicate their thoughts about the results. One participant [DS1] emphasised that *tight collaboration with the end result user* was very important.

4.2 Characteristics of the Data Science Process

Agility: Data science projects are exploratory and iterative in nature. Following an agile approach helps to manage customers? expectations and produce useful results. The data scientists said that their way of working resonates well with the agile approach.

There's a lot in common that you can really apply...Like always [in software development], do the MVP ["Minimum Viable Product"]...start iterating quick and try to have lots of communication and have the end user involved. [DS4]

...agility fits very well [with our] approach because we have to start with something and then actually try to produce as quickly as possible some kind of insight or results and then learn from those results and build on top of that. [We also] learn the environment that the customer has. Then actually I think it?s more visible also to the customer [that] we are producing something useful. [DS3]

Data science problems have to deal with a degree of uncertainty. The agile approach provides the opportunity to address the unexpected changes along the way.

Continuous learning process: The agile approach supports continuous learning throughout a project. It is important that both the data science team and the customer have the opportunity to learn during the process. The data scientists want to work with domain experts in order to gain good understanding of the application domain and the problem to be solved. The customer can learn what kind of results can be gained from the application of data science and how to utilize the results. It should be everyone's aim to 'learn by doing' [DS1] and use the new knowledge to improve the end results and possibly 'inspire some other ideas' [DS1].

End to end process: This means that the data scientists start the project by first understanding the customer and the customer's problem. This entails evaluating the relevance of the business problem. It also important that end-users are willing to utilize the solution. This calls for understanding the problem from the end-user?s point of view in order to provide the appropriate solution.

...we have sort of tried to formulate our way of getting into projects that go on and we really want to put an emphasis on the starting point or the end usage point, of who is going to use this result and how. And we start from there and then go backwards and do what we can and then try to improve it always...really start from the end user. [DS1]

4.3 Challenges

We present the challenges as they were experienced by the data scientists in different phases of the data science process. Table 3 shows an overview of the challenges.

Table 3. Overview of the challenges

Data Science Process Phase	Challenges
Conceptualization	Unrealistic customer expectations, communicating uncertainty
Problem definition	Identifying the right problem, limited interaction with domain experts, preference for tools as a solution
Data collection and preparation	Limited access to the data collection process, poor data quality, lack of cooperation from all required parties
Modelling	Lack of the required computational resources, differences in modelling and production technologies
Evaluation and validation	Lack of feedback from the end user
Deployment and utilization of results	The results are not utilised, what is the impact of the results?

Conceptualization. The challenges of this activity had to do with unrealistic customer expectations and communicating uncertainty.

Unrealistic customer expectations: The participants found that most customers did not have a realistic view of data science and its capabilities. In order to sell their solutions, tool vendors had propagated a tools-driven approach in the market. Hence, the customers expected quick solutions, mostly in the form of tools or systems but not recommendations or guidelines to aid in decision making. This led to a tendency to acquire tools without clearly knowing the initial problem for which to use the tools.

...people have need for data science but they don't understand it...then the other thing is that the market is kind of saturated by vendors who don't really sell data science in the sense that we understand it. [DS4]

If customers did not understand data science well, it made it difficult for them to view the problem correctly, hence hindering how well they could conceptualise the problem. The participants strongly advocated for a data-driven approach and had to employ some effort in getting the customer to gain the appropriate focus on the problem.

Communicating uncertainty: Due to the exploratory nature of data science, it is not always easy to predict the results. The conceptualization process also involved getting the customer to have an open mind towards what the results might imply. It was difficult for the participants to get the customer to understand and accept the inherent uncertainty of the outcome. This resulted in prolonged initial negotiations that were not always fruitful in closing the deals.

...often times, it is that you [i.e., the data scientist] really cannot say beforehand that—okay this is the result and that is what you will get. Basically because the outcome is very vague. You [i.e., the customer] use the money and you don?t know what you are investing [in]. [DS1]

Problem Definition. The main challenge here was identification of an essential problem to be addressed. The other issues were the limited interaction with the domain experts and the customers' overemphasis on tools.

Identifying an essential problem: A correct problem should be one that is solved by the obtained results. The participants had a great desire to produce useful results. However, it was often that the customers could not clearly explicate the problem in the first place.

...in many cases, you notice that your customer has collected data, but what to do with the data is unclear. And then there are lots of things we can actually calculate from the data but, all of them are not useful ones. So you really should find the useful thing and then concentrate on that. Then we would try to make the point that okay—in a way such data collection is not enough but you really need to find the correct problem that you actually need to solve. [DS3]

Limited interaction with domain experts: In most cases, the domain experts would be the ones to evaluate and sometimes use the data science results.

When defining the problem, the participants expressed that it was important to have input from the customers' domain experts. The domain experts know the problem best and are able to describe it very well—but their input was not readily available.

We might have a communication problem with the customer since we're not experts on the domain. We don't know what their problems are. And on the other hand they might not be aware of what we could do. [DS1]

The other one [i.e., problem] is how much we can actually communicate with the domain expert. [DS2]

Preference for tools as a solution: The participants found that there was a general bias towards tools and products in the market. Tools were seen as easy solutions to the problems as they were easy to acquire, were well-defined, easy to start using, and were perceived with less uncertainty. This hindered the customers' attitudes towards more thorough problem solving that data science requires.

I think many times the products are preferred to in a way because if you don't know the field then you actually think [of a product]. Because it's a product you can teach anybody to use it. But that's not really the case because if you don't know what you are doing or you don't know what the problem you are solving is, you put rubbish [in] and get rubbish out. I think it goes for why [the] typical thinking [is] okay, we buy a tool and then everybody can use it. [DS3]

Data Collection and Preparation. The challenges encountered during this activity are as follows.

Limited access to the data collection process: The participants were uncomfortable with being seen as magicians that could unravel wonderful discoveries from any sort of data without knowing its context. Not only did the participants want to have access to the data, but they also felt that understanding the process through which the data was collected would be useful in evaluating the problem and achieving the desired results.

...data is produced by some process. And, what we really need to do is understand the process or, preferably intervene with the process so that we get measurements that we really are after. Not so that there's some shadow on the wall [and] we try to deduce from that—we want to set up the whole thing. [DS4]

Poor data quality: There were several factors that compromised the data quality, such as the data being random and subpar, incorrect formatting and missing attributes, values and information. One participant gave an example:

But just as a practical example, it was not a data science project per se but in one project they had this legacy database of users where they only had one field for name. And then you had one to three first names and then several different variations of surnames and then we spent two weeks to build the engine that parsed the names to extract a surname. And even after two weeks, we got like two per cent of errors. [Research manager]

The way the data was gathered might also have had a negative effect, especially if it was collected without knowledge or intention of its use in the future.

...the data is originally not for the use that we [intend] but it has been collected for other purposes, maybe as log [data] and it's a side product of a process, and it's supposed to be somehow, [a] gold mine of insights. Or useful for some specific purpose. [DS2]

The data is often scattered around the organizations, the quality is poor. [DS1]

During the feedback session, the participants said that the data quality problem was improving. This was mainly because the market was becoming more informed about data science, hence investing effort and resources to collect meaningful data that could be utilised in the future and for different purposes.

Lack of cooperation from all required parties: We observed that some customer organizations had internal issues that hindered the participants' involvement in the projects. The issues mainly stemmed from the lack of a shared vision for the data science project amongst different departments in the customer organizations. This made it especially difficult to gather or have access to the required data.

One thing is that often the processes are lateral in the organization so that they [spread across] different branches of the organization. So there's IT and marketing and someone else involved and it's often hard to get [them] working [together]. [DS4]

Modelling. There were a couple of challenges related to this activity.

Lack of the required computational resources: During the focus group interview, the participants mentioned having difficulties with getting access to the IT resources and computational environments that they needed for modelling the results, particularly if the data could not be moved from the company premises.

More than so, it?s difficult to get the IT resources, both the data and the computational environment that we need. Often it?s difficult to get either of them or at least one of them. [DS1]

During the feedback session, the participants pointed out that the situation had improved due to cloud solutions becoming readily acceptable and accessible.

Differences in modelling and production technologies: Sometimes, there was a difference between the modelling technology and the one in which the results are applied. This led to difficulties with integrating the results in the customer's environment and required more time, effort and money. In the end, this would limit the impact of the results.

Evaluation and Validation. The main challenge here was an apparent gap between the data scientists and the end users of the results. The people who ordered the project and thus got the results, e.g., the business experts, were not necessarily the actual end users acting on or using the results.

Lack of feedback from the end user: There is a difference between the feedback received from the business or domain experts working in the customer company, and the real end users of the results. If the real end users are not connected to the data scientists, it makes hard for the data scientists to actually assess the progress of their results.

This is actually the number one [problem], [lack of] tight collaboration with the end result user. [DS1]

Deployment and Utilization of Results. The data scientists were sometimes frustrated by how the customers handled the project outcomes. Sometimes, the results were not put into use which meant that the participants would never know the real impact of the results.

The results are not utilised: Sometimes, the results were not applied. This was due to factors, such as (1) lack of cooperation between different departments, e.g., marketing and IT, (2) the business stakeholders failed to facilitate the utilization of the results if they did not understand, were not fully convinced or they did not feel confident about the results.

...I think most of the failures that we [have] had are because the results are just never [used]. They are ready and nobody ever uses them for anything...like I said, most of the time the problem is really to get the results into use. [DS1]

What is the impact of the results? As a result of the outcomes not being utilised, the participants found it difficult to know, measure or observe the effectiveness of the results.

For the results to be useful, they [i.e., customers] have to accept that—well—things are how they are, not how people thought they would like them to be. [DS4]

On the other hand, the participant quoted above [DS4] pointed out the fact that in order to effectively measure the impact, one would require an experimental setup which is usually *'expensive and technically heavy'* to put in place. This means some considerations have to made with respect to investments towards experimentation.

Summary of the Challenges. The challenges we have presented above reflect the complications of applying data science in software and service engineering as experienced by the study participants. We classified the challenges into three groups, i.e., difficult, moderate, and mild problems. The groups were according to the perceived ability to solve them, as observed during the analysis. Table 4 summarises the challenges.

The difficult problems were those considered hard to solve. They comprised of human and organisational aspects which are always not easy to resolve. These problems also seemed to be more out of the participants' control, even though the participants considered them to be very important. The moderate problems were seen as somewhat solvable with some persistent intervention from the participants. The mild problems, such as those related to data quality, computational resources, and modelling issues, were seen as clear and easily solvable.

Table 4. Summary of the challenges

Problem Group	Challenges
Difficult	Communicating uncertainty, identifying essential problems, lack of cooperation from all required parties, lack of feedback from the end user, the results are not utilised, what is the impact of the results?
Moderate	Unrealistic customer expectations, limited interaction with domain experts, preference for tools as a solution, limited access to the data collection process
Mild	Poor data quality, lack of required computational resources, differences in modelling and production technologies

The human and organisational nature of the difficult problems is an indication of immature markets, which have spread extremely fast to many new application domains. Some of these problems can be expected to fade with time as the misconceptions about data science get clearer and data scientists become integrated as members of software and service development teams.

5 Discussion

The goal of this study was to gain understanding on how data science can be applied in software development organisations. The results are based on a qualitative case study approach. This paper presents the process that the experienced data science team of the case study company recommends to be used with customers. The paper also describes the key characteristics of the process and challenges encountered in practice when data science projects were conducted with customers.

The recommended data science process consists of six activities. The first activity focuses on understanding customers? business problem and their expectations for the project. The second step is to translate the business problem into a computational or mathematical problem. The following two activities cover data collection and modelling tasks. During the fifth activity of the data science process, the results are evaluated and validated with the customers and end users. Finally, it is essential to ensure that the results are put into use and their impacts are assessed. Some of the activities of this process are similar to activities mentioned in other data science analysis processes, i.e., discovering the data [8,9], modelling the data [8,9], and reflecting and iterating on the results [8].

Based on the interview study of 16 data scientists, Kim et al. [10] found that data scientists at Microsoft worked on three activities: (1) data collection, (2) data analysis, and (3) data use and dissemination. The authors also point out that this list is not complete, but an overview of the activities they identified from their study. When comparing the list of the three activities with the data science process described in this paper, the main difference is that the data

scientists of our case study highlighted especially the importance of identifying a real business problem that can be translated into a computational problem.

According to the experienced data scientists of our case study, identifying essential problems to be solved by data science is one of the most difficult challenges in their work. Similarly, Zhang et al. [15] report that it is often easy to start from some datasets, apply certain data analysis techniques and make some observations that actually do not help practitioners. One of the main lessons Zhang et al. learned was that it is important to first identify essential problems and then obtain the right dataset to help solve the problems.

Another difficult challenge that data scientists can face in practice is that it is not easy to communicate and get the customer to understand the uncertainty of outcomes from data science projects. According to the experienced data scientists, it is often so that they cannot state precisely at the beginning of the project what results the customer will get. In order to solve this challenge and also other challenges, such as identifying essential problems and managing unrealistic customer expectations, the experienced data scientists recommended the agile and iterative data science process. This lesson from our case study supports the lesson learned by Zhang et al. [15]. Based on a case study conducted at Microsoft, they report that creating software analytics solutions for real-world problems is an iterative process. They also point out that it is important to work in an agile way to build a quick feedback loop with practitioners and to identify essential problems early.

From the perspective of research, the main contribution of this paper is that it describes a rather large set of challenges that are based on the experiences of the data scientists who have worked in customer projects. An increasing number of companies are interested in applying data science. Therefore, it is important that software engineering and data science researchers can develop solutions to these challenges in close collaboration with practitioners. It is also important that challenges related to the application of data science in software development projects will be investigated in different kinds of companies and contexts. For example, Kim et al. [10] plan to conduct a large-scale survey to quantify data science tasks identified in their interview study and describe the challenges associated with data science work. It will be interesting to compare the results of the survey with the results of our case study.

From the perspective of practice, the paper offers an overview of the six data science activities. The results also suggest that the data science process should be an agile, continuous learning and end-to-end process. Continuous learning means that data scientists need to gain iteratively a good understanding about the business problem and application domain. In addition, customers need to learn what kind of insights can be gained from the application of data science and what these insights mean in practice. The end-to-end process means that it starts from the discovery of relevant problem and covers the activity where the results from the application of data sciences are actually used and their impacts are evaluated.

Threats to Validity. As this study is a case study and descriptive in nature, there is little evidence to support any causal relationships, thus the internal validity is not the main concern of this study. However, the results do include knowledge constructs that could be interpreted having some causal characteristics, such as the claims from the informants that iterative approach to design science process would help to overcome certain challenges. These are clearly the views of the informants and thus to be taken with appropriate caution if interpreted as guidelines to follow. On the other hand, however, the informants were data science experts, who have encountered the challenges in their work and thought for the possible solutions beyond the interview sessions of this study, so their claims may be more valid and justified than random opinions.

In terms of construct validity, the richness of the data from multiple interviewees and member checking the results with the informants significantly reduce the risk that major issues would have been misunderstood by the researchers. However, one issue on construct validity may rise from the varied definitions or understandings of the term data science, particularly as its interpretation beyond this study may differ from the semantics captured between the informants and the researchers, which is broader than, e.g., data collection and analytics only (see Fig. 1). To build a basis for the credibility [13], the interviews were audio recorded, transcribed and analysed using Atlas.ti as the tool.

Our study is conducted with the case company only, although through their customer projects, the results cover data science challenges beyond the case company only. The external validity or transferability of the results beyond the case would be based on the assumption that the informants would have encountered challenges that are not particular or stemming from the context of the case company only. That is, it is very much possible that the challenges identified have relevance beyond the case as well as the ideas proposed by the informants for alleviating the challenges. However, it is clear that the potential application of the results in other cases essentially expects a knowledgeable person or persons with good expertise in their own domain in order to interpret and apply the results in their context.

6 Conclusions

This study contributes to the growing interest in data science across different disciplines, specifically service and software engineering. It helps both researchers and practitioners to understand the applicability of data science in service and software development and be informed about some of the impending challenges.

The difficult problems identified comprised of human and organisational aspects, whereas the problems such as poor data quality and modelling issues were not seen as primary concerns for the data science process. Our results also indicate that it is important to establish an agile and lightweight data science process that supports continuous learning while focusing on a real business problem. The experienced data scientists highlighted that it is not enough to focus on data collection and modelling. Instead, you really need to find the relevant problem that you actually need to solve and can be solved by applying data science.

Our future work will focus on the factors influencing the successful application of data science in service and software development projects. In addition, we are interested in investigating how customers experience the application of data science in service and software development projects.

Acknowledgments. This work was supported by TEKES as part of the N4S Program of DIMECC (Digital, Internet, Materials & Engineering Co-Creation). We would also like to thank the case company Reaktor for the possibility to conduct this research.

References

1. Begel, A., Zimmermann, T.: Analyze this! 145 questions for data scientists in software engineering. In: ICSE, pp. 12–22 (2014)
2. Bener, A., Misirli, A.T., Caglayan, B., Kocaguneli, E., Calikli, G.: Lessons learned from software analytics in practice. In: The Art and Science of Analyzing Software Data, pp. 453–489 (2015)
3. Bird, C., Menzies, T., Zimmermann, T.: Past, present, and future of analyzing software data. In: The Art and Science of Analyzing Software Data, 1st edn., pp. 1–13 (2015)
4. Cao, L., Science, D.: A comprehensive overview. ACM Comput. Surv. 59(3) (2017). Article No 43
5. Creswell, J.W.: Research Design-Qualitative, Quantitative, and Mixed-Methods Approaches, 4th edn. SAGE, California (2014)
6. Cruzes, D., Dyba, T.: Recommended steps for thematic synthesis in software engineering. In: International Symposium on Empirical Software Engineering and Measurement (ESEM), pp. 275–284 (2011)
7. Davenport, T.H., Patil, D.J., Scientist, D.: The Sexiest Job of the 21st Century, Harvard Business Review, pp. 70–76 (2012)
8. Fisher, D., DeLine, R., Czerwinski, M., Drucker, S.: Interactions with big data analytics. Int. Mag. **19**(3), 50–59 (2012)
9. Kandel, S., Paepcke, A., Hellerstein, J.M., Heer, J.: Enterprise data analysis and visualization: an interview study. IEEE Trans. Vis. Comput. Graph. **18**(12), 2917–2926 (2012)
10. Kim, M., Zimmmermann, T., DeLine, R., Begel, A.: The emerging role of data scientists on software development teams. In: ICSE, pp. 96–107 (2016)
11. Kontio, J., Lehtola, L., Bragge, J.: Using the focus group method in software engineering: obtaining practitioner and user experiences. In: ISESE, pp. 271–280 (2004)
12. Liamputtong, P.: Focus Group Methodology-Principles and Practices. SAGE, California (2011)
13. Patton, M.Q.: Qualitative Research & Evaluation Methods, 3rd edn. SAGE, California (2002)
14. Strawn, G.: Data Scientist, IT Pro, pp. 55–57. Computer.org
15. Zhang, D., Han, S., Dang, Y., Lou, J.-G., Zhang, H., Xie, T.: Software analytics in practice. IEEE Softw. **30**(5), 30–37 (2013)