Managing and Composing Teams in Data Science: An Empirical Study

Timo Aho
TietoEvry
Tampere, Finland
timo.aho@iki.fi

Terhi Kilamo
Computing Sciences
Tampere University
Tampere, Finland
terhi.kilamo@tuni.fi

Lucy Lwakatare
Computer Science
University of Helsinki
Helsinki, Finland
lucy.lwakatare@helsinki.fi

Tommi Mikkonen
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland
tommi.j.mikkonen@jyu.fi

Outi Sievi-Korte
Computing Sciences
Tampere University
Tampere, Finland
outi.sievi-korte@tuni.fi

Sezin Yaman
KPMG Finland
Helsinki, Finland
sezin.yaman@kpmg.fi

***Abstract*—Data science projects have become commonplace over the last decade. During this time, the practices of running such projects, together with the tools used to run them, have evolved considerably. Furthermore, there are various studies on data science workflows and data science project teams. However, studies looking into both workflows and teams are still scarce and comprehensive works to build a holistic view do not exist. This study bases on a prior case study on roles and processes in data science. The goal here is to create a deeper understanding of data science projects and development processes. We conducted a survey targeted at experts working in the field of data science (n=50) to understand data science projects' team structure, roles in the teams, utilized project management practices and the challenges in data science work. Results show little difference between big data projects and other data science. The found differences, however, give pointers for future research on how agile data science projects are, and how important is the role of supporting project management personnel. The current study is work in progress and attempts to spark discussion and new research directions.**

***Index Terms*—Data science, agile practices, teamwork, project management**

I. Introduction

The inherent issues in big data analytic systems can be tackled with more mature project management methodologies [1], [2]. Our seminal investigation into the practices of data science [3] found that experimentation is at the core of data science projects and that data science projects commonly utilize an iterative development approach. Furthermore, it was found out that multidisciplinary teamwork is present especially in larger projects where Proof-of-Concepts are utilized.

In this paper, our aim is to understand data science projects' processes and practices as well as team structure through a survey targeted at data science experts working in the field. The survey questions are based on the preliminary results gathered in our previous work [3] hence looking deeper into the topic with both a larger sample set and a more structural research approach. Specifically, we seek to understand what roles, tasks and processes data science projects consist of, and what are the main challenges of data science today.

Here, a note on terminology is in order as the field is somewhat mixed when it comes to using somewhat overlapping terms of big data, data science, data analytics, machine learning and data mining. Here, the term data science is used for extracting knowledge with multidisciplinary techniques such as statistics and machine learning from data sets big and small.

The survey indicates that data science projects suffer from a set of key challenges that expand our previous findings with how projects working with data work. Addressing these would improve the data science project processes and output. Further, we identify indicators of differences between data science projects with big data and those not utilizing big data. These indicators point direction for further research in identifying critical points of improvement in particularly big data projects. The research reported here is a work in progress. The results are mainly indicative due to the still small sample size and the local scope of the responses. We hope to inspire further research into the topic.

The rest of this paper is structured as follows. Section II presents the background of the paper and related work. Section III gives the research approach. The results of the study are given in Section IV. Section V discusses the results and Section VI presents threats to the validity of the study. Finally, we conclude the paper in Section VII.

II. Background and Related Work

It has been typical for data science [4] projects to follow their own processes and practices. For example, a recent 2018 survey by Saltz et al. [5] reported that 82% of data science teams did not follow any explicit project management process or practices, even though 85% thought such would be

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

beneficial. The processes have also been different from those that have been typical for example in the context of software development [6]. However data scientists seem to be moving towards consolidated practices and tools [3]. To succeed, this transition requires thorough reconsideration of data science organizations and their operations.

Big data and data science as concepts are intertwined [7]. In this paper, we refer to data science as a data-driven process of discovering knowledge from data by applying different techniques, such as machine learning. Data science work is primarily conducted by the data scientists, who conduct different activities, such as data cleaning, feature extraction, data analysis/modelling and result evaluation. Still, data scientists have been found to take on different roles and work with different types of profiles [8], [9], [10] in data science projects and teams. They also work with experts of other fields, e.g. business experts and software developers. Due to the varying nature of data science projects, different tools are also used [8].

Most commonly known methodologies targeted for data science work are KDD [11], CRISP-DM [12], and SEMMA[1]. Shafique and Qaier [13], [14] provide a comparison of these frameworks. Extensions (e.g. [15], [16]) on these methodologies are also available. Their goal is at tackling a number of different problems that the practitioners have identified. In the frameworks, the workflow of data science projects has multiple phases, for example preparation, modeling, and deployment.

Data science projects can be categorized based on their approach to data. One division is two-fold – routine data transformation and exploratory projects [17]. Data science projects can also be labeled based on the infrastructure and discovery dimension [18]. On these dimensions, the projects can be grouped into four categories – hard to justify, exploratory, well-defined and small data. Similarities in the work of data science teams to software development prior to the introduction of agile methodologies can also be seen [1]. Furthermore, in data science projects it can be difficult to estimate the budget and schedule and how successful the project will be [19]. Also quality assurance of the results is often considered insufficient, and the projects are reliant on individual effort instead of team work.

III. Research Approach

This study consists of a survey targeted at experts in the data science business domain. It continues to build knowledge on data science projects, teams, their practices and tools from our earlier multiple-case study [3], where the goal was to understand the typical process flow in data science projects and the role of the data scientist and teamwork. Here, the main research problem is: *How well structured in terms of roles, tasks and processes current data science projects are in their everyday work?*

[1]Available at https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=15.1

A. Research Questions

Specifically, the research questions we seek to answer are the following:

- RQ1: How structured are current data science projects?
  - RQ1.1: What kind of teams are involved in projects?
  - RQ1.2: What are the typical tasks?
  - RQ1.3: What kind of project management practices are utilized?
- RQ2: What are the typical challenges?

B. Data Collection and Analysis

The survey was based on the preliminary results in our previous work [3]. The survey questions were additionally iterated to reflect surveys used in state of the art research and to allow answering in a reasonable time (maximum 15 minutes). A pilot of the survey was run and the questions were improved based on the pilot feedback. The survey was distributed through social media channels as well as the researchers' direct contacts over a timeline of five weeks with a theoretical total reach of 5000 (excluding LinkedIn group reach). The timeline, method and reach are shown in Table I.

Between March and April 2021, 50 respondents anonymously answered the survey. The survey contained different types of questions, e.g., Likert-scale as well as open questions. However, only a total of five (5) answers were entered into the open-ended questions and they were excluded from the study as too small a data set. Thus, the data we studied further was quantitative. For quantitative analysis, we transformed the dataset into long form and cleansed it for example of any typos. This made the application of descriptive statistical methods possible.

As respondents were given the freedom of choosing multiple roles reflecting their job functions, individual respondents and the roles have M:1 relationship. Similarly, the other survey questions that allowed respondents to select multiple options (positions and data types, data science and project management tasks and challenges) created a rich dataset where respondents are linked to multiple data points. In order to investigate the relationships between these data points, we trained a sparse recommender matrix using a R package [20]. In effect, we transformed the dataset into long format, retrieved sparse matrices and trained a sparse matrix recommender and investigated the respondents that are clustered together based on provided profiles.

The recommender matrix is a person-tag matrix, where each row corresponds to a person participated in the survey and each column corresponds to tags retrieved from survey answers (such as "role:data scientist" is a tag for the survey question where the participant chose data scientist as his/her job). Furthermore, we restricted the recommender matrix with certain tags (e.g., "role:data scientist", "role:data analyst" and "challenges") in order to capture clusters associated with these profiles. In order to visualize these clusters, we used community graph plots [21] in the tool Wolfram Mathematica.

Time	Media	Method	Reach
Week 1	Twitter	Tweet and 9 retweets	$\sim$ 3000
Week 1	LinkedIn	Posts by two researchers	$\sim$ 1100
Week 1	LinkedIn	Reshared post by one researcher	500
Week 1	Direct contact	Emails to focus group	10
Week 1	Expert network	Sharing in Slack channel	$\sim$ 100
Week 2	LinkedIn	Post by two researchers	$\sim$ 800
Week 2	LinkedIn	Direct tag in post comments	63
Week 2	LinkedIn	Sharing link to six groups	$\sim$ 900 000
Week 2	Expert network	Sharing in industry network	$\sim$ 100
Week 3	Twitter	Tweet and one retweet	$\sim$ 200 reach
Week 3	LinkedIn	Post by a researcher	$\sim$ 1000
Week 5	LinkedIn	Direct tag in post comments	37

TABLE I: Distribution of the survey in terms of timeline, channels and theoretical reach

IV. Results

While presenting the results of the survey, we will first focus on the team structure and roles, second on the project management practices, third on the perceived challenges and finally on the relationships between data points with multiple options. We will highlight the effect of the size of the data.

A. Team structure and Roles

While the respondents represented a variety of industrial verticals as shown in Table II, the majority, 34% of respondents, were in consultancy and technology/ICT. The various roles of the 50 respondents are listed in Table III. The majority of the respondents were data scientists (50%).

Many data scientists also assigned themselves with other roles. Such roles included data analyst (8%), data engineer (6%), and management (8%). Furthermore, 10% of the respondents reported having more than two distinctive roles. Software related roles – software architect or software developer – were involved clearly more rarely. However, they also appeared combined with the data scientist role. When combining all roles with several elements, it becomes the most common role (38%). The respondents were further asked about how often certain roles are involved in their data science projects. A breakdown of these answers overall is given in Fig. 1.

Juxtaposing the roles in big data projects with other data science projects shows some differences. This is shown in Figure 2. Data scientist and engineer roles are more commonly present in projects where big data is handled. In projects handling big data, there were no answers where the roles of data scientist and data engineer were marked as never to appear in the teams. Data scientist is present at least sometimes when handling big data. The roles of the project manager and product owner are also more relevant with big data. Especially product owner is again a role that appears in big data projects at least rarely.

For the size of the organizations the respondents were working in, Table IV gives the results. Most commonly the respondents worked in small and medium-sized companies (42%) or in large enterprises (42%). However, 16% of the respondents represented a micro enterprise.

Table V depicts the size of the development team size. The most typical development team size was 2–5 persons (32%). Teams of 6–9 persons was the next most common with 14% and a small team of less than two persons was the case for 12% of the respondents.

Table VI shows the reported lengths of the data science projects. The typical length reported was 3–12 months (38%) on average. However, 18 % reported the project length to vary between projects. Short projects 1–2 weeks and project length of more than a year or with continuous product development both had an 8% share in the answers.

The types of data being worked on in the data science projects was reported to constitute tabular data (N=44), time-series data (N=37), image data (N=13), video data (N=6), audio (N=1), Telecom binary data (N=1) and geographic data (N=1). This represents a diverse set of use cases in data science projects.

Figure 3 gives the tools used by the respondents. The most common tools were the ones used for analytics coding, such as Python and R. These were used by 90% of the respondents (45 responses). Query languages such as SQL and computational notebooks such as Jupyter/Databricks were both used in 70% (35) of the responses. Version control systems, e.g. Git, were also relatively common with 62% (31 responses). Big data processing was used by 26% of the respondents.

The tasks that took the most effort in data science projects were related to preprocessing the data, i.e., data manipulation, aggregation and cleaning unstructured data. Formulating and communicating a clearly defined problem/training objective was the next tasks requiring effort. Getting data from external data sources and access to data storages was considered the third most laborous task. In summary, Figure 4 shows all tasks with the most effort according to the respondents.

B. Project management practices

A particular interest for us was to find out how well-managed data science projects currently are. Project management practices are of particular importance in the effort to improve big data processes. We have here compared responses from those who are engaged in big data projects with the rest of the sample (Figure 5).

While it was overall common to have teams of 2–5 members suggesting a good agile practice, the average length of a data science project in respective teams suggest otherwise as common responses were for 3–12 months and 3–8 weeks. On

Industry	N	%
Consultancy	17	34%
Technology/ICT	17	34%
Government	4	8%
Transport and mobility	2	4%
Finance	1	2%
Forestry	1	2%
Health	1	2%
Hospitality	1	2%
Manufacturing	1	2%
Marketing	1	2%
Media	1	2%
Research on multiple sectors	1	2%
Sports	1	2%
Telecommunication	1	2%
Total	50	100%

TABLE II: Organization Industry

Role	N	%
Data scientist	14	28%
Data analyst	5	10%
More than 2 distinctive roles	5	10%
Data scientist/analyst	4	8%
Data scientist/management	4	8%
Data scientist/engineer	3	6%
Data/software engineer	3	6%
Lead/Head/Director	3	6%
Software developer/engineer	1	2%
System administrator/solution architect	1	2%
Project manager	1	2%
Data analyst & lead/head/director	1	2%
Data engineer	1	2%
Software developer/engineer & lead/head/director	1	2%
System Administrator	1	2%
Machine learning intern	1	2%
Student	1	2%
Total	50	100.0%

TABLE III: Roles of Survey Respondents

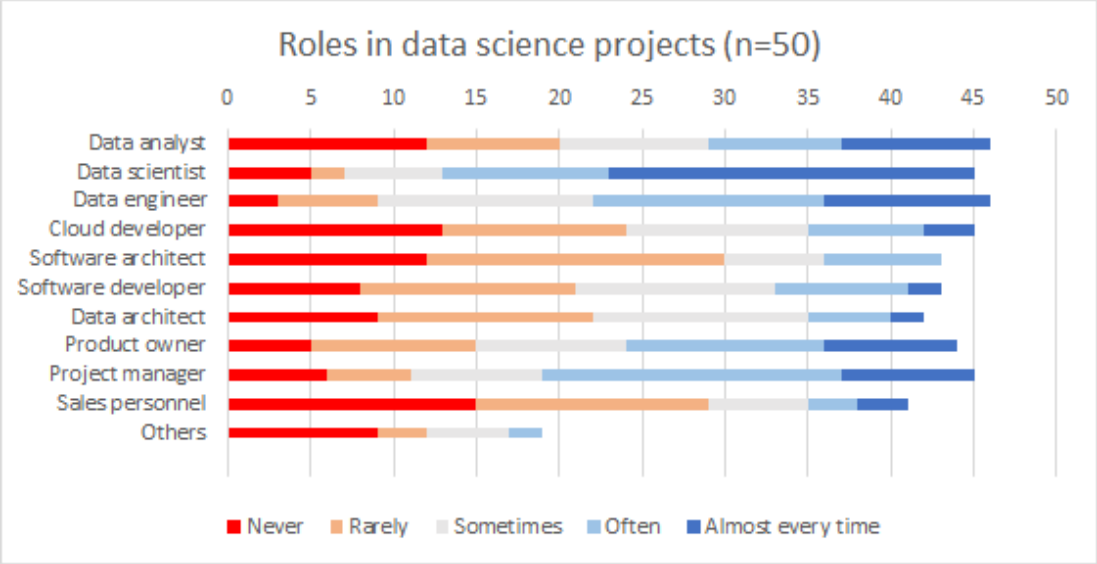

Fig. 1: Roles involved in data science projects

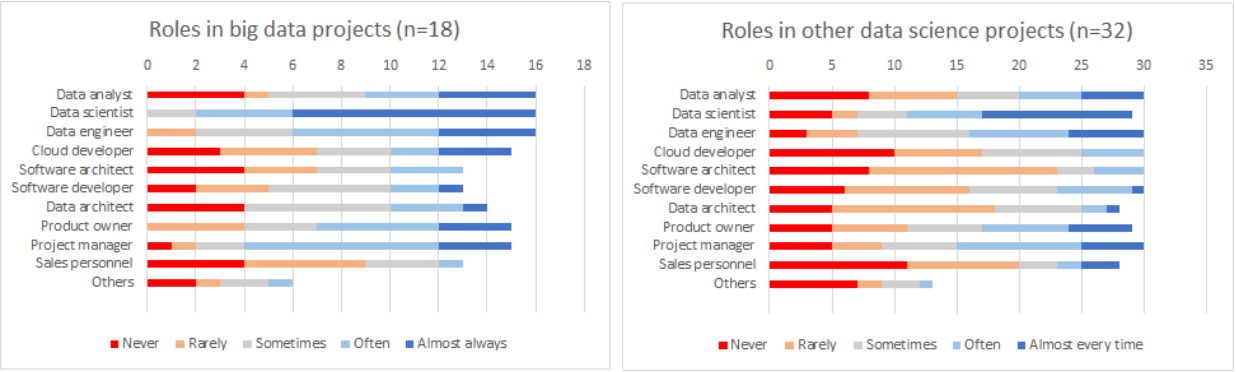

Fig. 2: Roles appearing in big data projects compared to other data science projects

project management practices, among the responses, daily or weekly meetings were utilized with nearly all respondents regardless of whether big data was utilized or not. Prioritization of work was also commonly utilized. As seen in Figure 5, in big data projects prioritization was used with 55% of the sample, while for data science not specifically using big data,

Size (employees)	N	%
1–10	8	16%
11–50	15	30%
51–200	6	12%
201–1000	10	20%
More than 1000	11	22%
Total	50	100%

TABLE IV: Organization size

Size	N	%
Less than 2 persons	6	12%
2–5 persons	32	64%
6–9 persons	7	14%
10–15 persons	0	0%
16–20 persons	2	4%
¿20 persons	3	% 6
Total	50	100.0%

TABLE V: Development team size

Duration	N	%
1–2 weeks	4	8%
3–8 weeks	13	26%
3–12 months	19	38%
More than a year	3	6%
It varies from project to project	9	18%
Continuous product development	1	2%
Total	50	100%

TABLE VI: Length of data science project

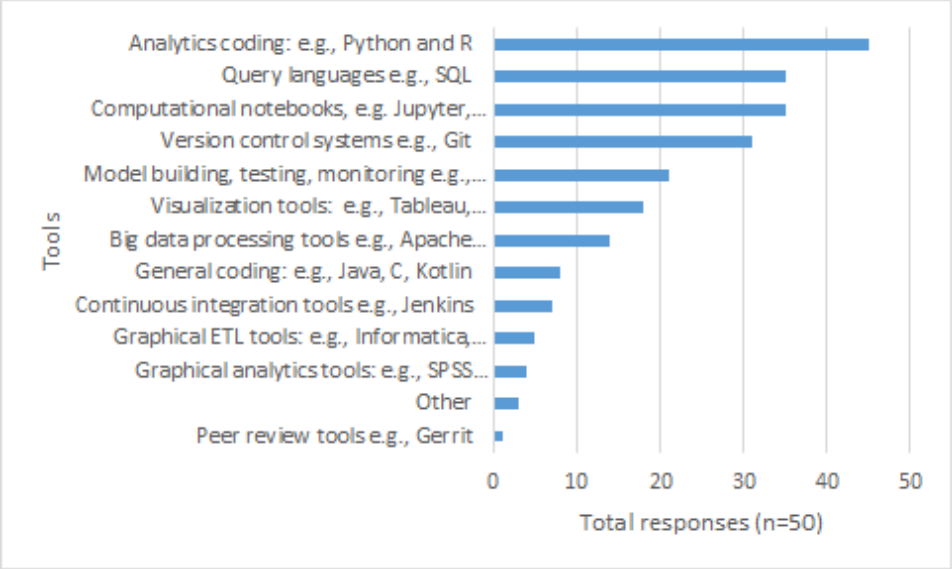

Fig. 3: Used tools

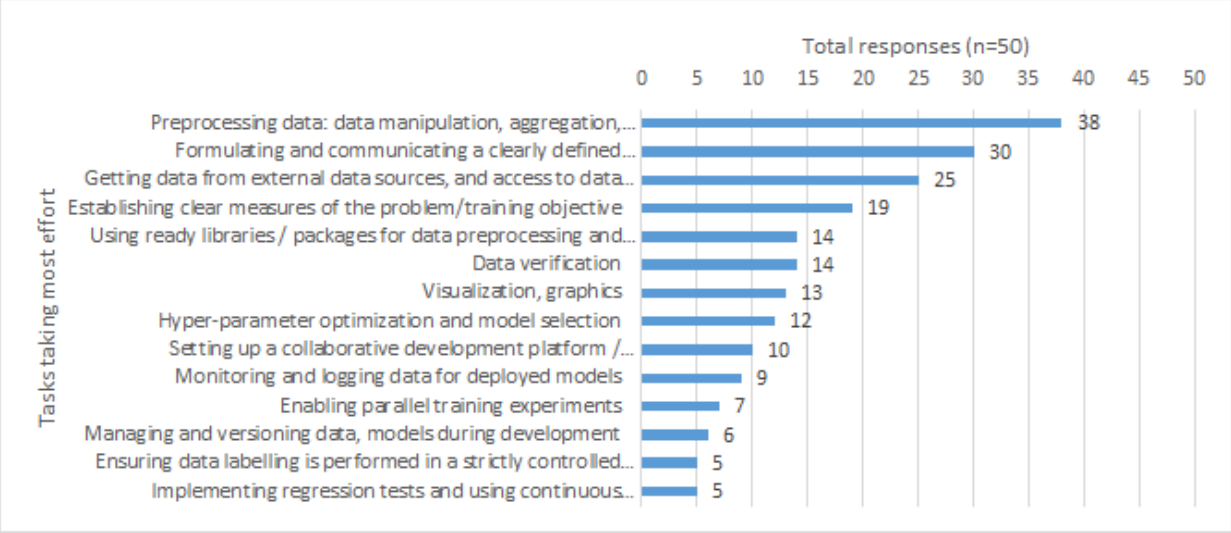

Fig. 4: Tasks with the most effort

over 62% utilized prioritization. Kanban boards were used by around 50% of all respondents, and code/result reviewing was also fairly common. However, other practices were much less frequently applied. Sprints were used by around 39% of big data projects and 31% of other projects, while retrospective meetings were used in approximately 28% of big data projects and 25% of others. Finally, while evidence so far has indicated that data scientist are increasingly merged into teams with various kinds of specialists, cross-functional working was quite rarely utilized (22% equally in both groups).

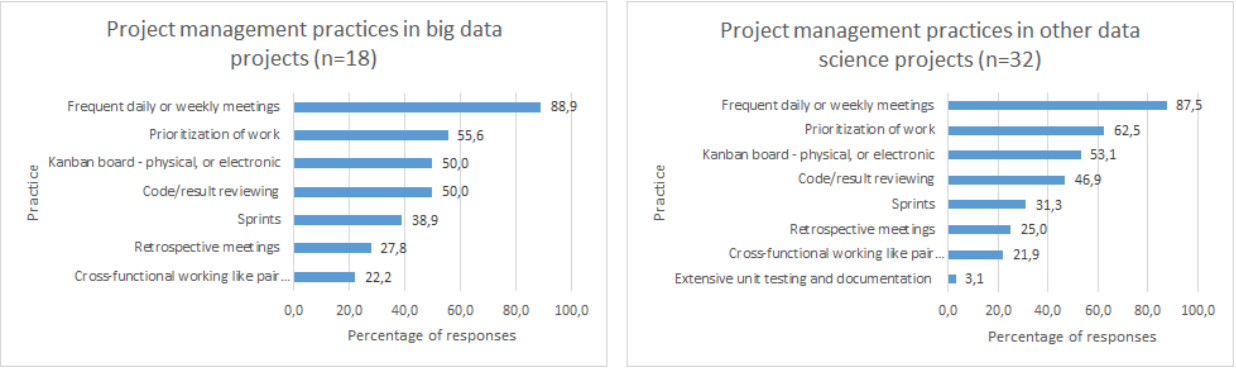

Fig. 5: Project management practices

C. Challenges

As seen in Figure 6, common challenges are observed in relation to both tasks in data science projects and project management practices. Unclear project goals and changing targets was a common challenge for tasks in data science projects. Technical development practices such as CI/CD pipelines were typical challenges from the project management practices point of view. Estimating the project timeline as well as poor data quality were among the challenges for more than half of the respondents. Half of the respondents reported overselling or overly optimal initial belief in the results as a challenge.

For respondents who chose to elaborate using free-text the important problems emphasized further the following problems in data science project: lack or insufficient data, including poor data quality, long query and model training times and the lack of skills and knowledge, including the inability to interpret results or how different features affect the results.

D. Experiments and Respondent Communities

Experiment 1: Data analyst, scientists and engineers. In attempt to understand how these three specific roles, i.e., data analysts, engineers and scientists, responded to the questions about their positions, team-sizes, data science tools they use and the challenges they face, we identified 2 communities as depicted in Figure 7. The smaller community consists of 21 respondents. Although there are respondents who provided multiple roles defining their job function, a clear majority (67%) chose the role ”data scientist”. The outliers that can be detected from the Figure 7 in this community are the ones who reported other roles, e.g, ”person:7” being system administrator. While one third of this community works with a team of 2–5 members, the second most common team size is 6–9 members (28%). As for their most common work tasks, 9 respondents (42% of the community) report ”Getting data from external data sources, and access to data storages”, 12 respondents (57%) selected ”preprocessing the data” and 13 respondents (62%) ”formulating and communicating a clearly defined problem”. One third of this community utilizes data visualization tools (e.g., Power BI, Tableau), 12 respondents utilize computational notebooks (57%), 17 (80%) utilize analytics coding tools (e.g., R or Python) and 9 (31%) utilizes general coding. Only 4 respondents (19%) utilize model building tools.

The second community, depicted with red nodes in Figure 7 consists of 29 respondents. 48% of this community are people reporting more than one role (e.g., both data analyst and data engineer), and roles different than ”data analysts, engineers and scientists”, such as directors or project managers, were chosen by 8 respondents (28%). The majority in this community is in teams of 2–5 members (86%). The outlier ”person:45” that can be seen in Figure 7 is a machine learning intern. Most of the respondents in this community report ”preprocessing the data” (86%) as their main task and both ”Getting data from external data sources, and access to data storages” and ”formulating and communication a clearly defined problem” are reported by around half of the community (55%). Almost everyone in this community utilizes analytics coding tools (96%), and a clear majority uses computational notebooks (79%). While 17 (59%) utilize model building tools, 16 people utilize version control systems (55%) and 15 (51%) utilize general coding tools (e.g., Kotlin). Only 10 (34%) utilizes visualization tools.

When we look into the challenges we see interesting differences between these two communities. While 38% of the first community reports ”Overly optimal initial belief in the results / overselling” as a major challenge, the percentage goes up the 59% in the second community. Unclear project goals pose a challenge for the first community with just 47%, while for the second community it is a challenge for 76%. While the first community reports challenges with ”Technical development practices: version control, CI/CD pipelines, DevOps” for 38%, the percentage goes down to 17% in the second community. Although we have limited data size, we may interpret that the first community of data scientists commonly work on preprocessing the data, data analytics tasks and come across challenges in data pipelines. The second community is a mixture of respondents with multiple reported roles that can also work with general coding tools besides the data analytics and data modelling, and they are having challenges regarding unclear project goals and overselling.

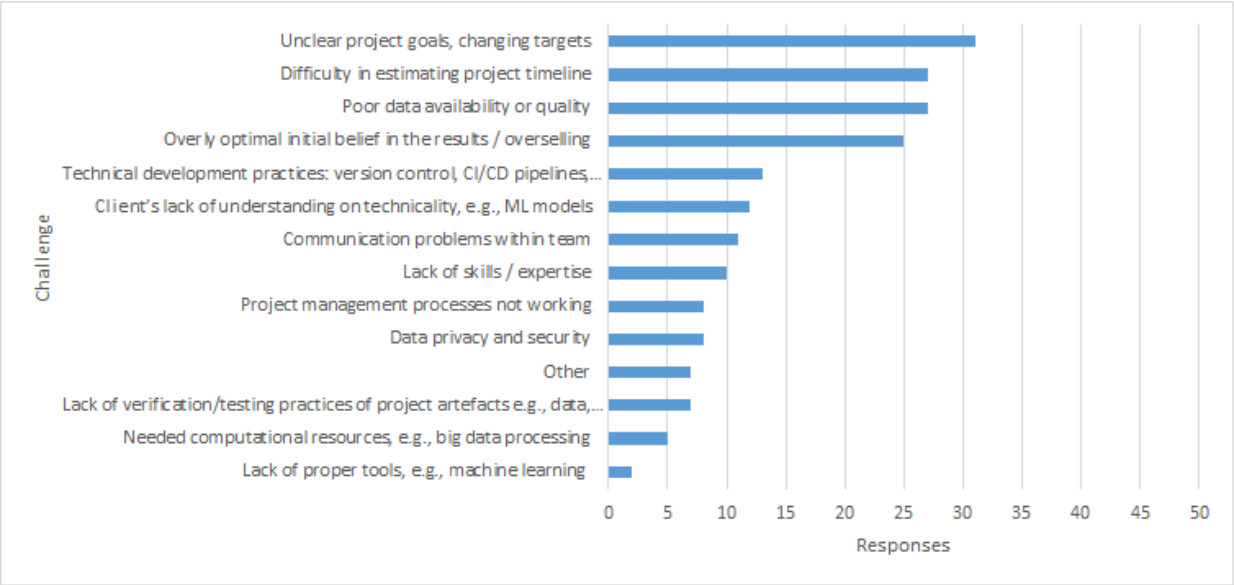

Fig. 6: Challenges

Overall, although we could observe two communities, we don't see significant differences between these two communities with respect to the roles of data scientist, analyst and engineer, but rather overlapping tasks, use of tools and challenges faced. This might indicate that the role title does not consistently define the essence of the work conducted at data science projects.

Experiment 2: Project practices. We performed another experiment in attempt to understand how the top 3 project management practices cluster around industries, challenges, tools and team-sizes as depicted in Figure 8. The smallest community (depicted as purple) consists of 13 participants, mostly with data scientist and data engineering roles. Only 4 participants here reported the industry they are in as consultancy and the rest reported as various areas, including government, sport, telecommunication, media and manufacturing. Majority of this community (85%) have frequent daily of weekly meetings and a majority (70%) reports that unclear project goals is the common challenge in data science projects. 3-8 weeks of project duration is the most common (85%) for this community.

The second community depicted with yellow nodes in Figure 8 consists of 14 participants, mostly with multiple roles including data analysts, data engineers, software developers, and also 4 Lead/Head/Director and project managers. Half of the community reports that the approximate team size is 2-5 people. A clear majority of this community (86%) is in the industry of Technology/ICT and they work as a part of in-house teams (92%). While other project management practices are also reported, the top practise is "Kanban board - physical or electronic" (79%). As the typical length of a data science project, this community reports on "It varies significantly from project to project" (43%) as the most common answer.

The third and biggest community consists of 22 participants. While the reported roles vary, the most common role is "data scientist", reported by half of the community (50%). Person:7 has been an outlier, with the role of "system administrator" and the industry of "Hospitality". Majority of this community states that the industry they work in is "Consultancy" (60%) and half of the community (50%) states that they work as "an external consultant". All participants of this community reported that they are involved in "frequent daily or weekly meetings" (100%). Majority of the community (60%) reported that 3-12 months is the most typical project length. 14 participants in this community (64%) reports "poor data availability or quality" as a major challenge, while 13 participants (60%) report on "unclear project goals, changing targets", and exactly half of the community select "overly optimal initial belief in the results / overselling" as the biggest challenges in the data science projects.

Overall, we observe that the industries the respondents work in might be correlated with the project management practices that are followed. We observed that the first community, depicted as purple nodes in Figure 8 mainly consisted of respondents from industries other than consultancy and ICT, such as government, sports, transportation an the like. This community favors frequent daily or weekly meetings and often works with 3-8 weeks long data science projects. On the other hand, we saw that the majority of the second community (yellow nodes in Fig.8) had respondents from Technology/ ICT who mainly work in-house, and project duration might significantly vary for their work in comparison to the other communities. The use of Kanban baoards is reported to be the most common project management practice for this community. The majority of the last community (red nodes in Fig.8) consisted of consultants and they all reported that "frequent daily or weekly meetings" is the most common project management practise.

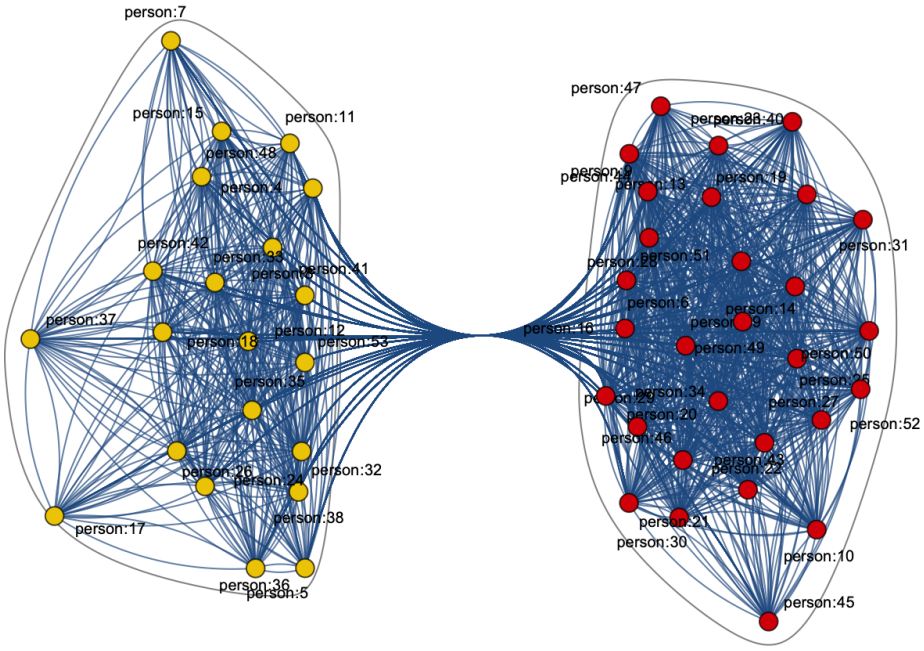

Fig. 7: Communities based on the tags ”role: data analyst, data engineer or data scientist”

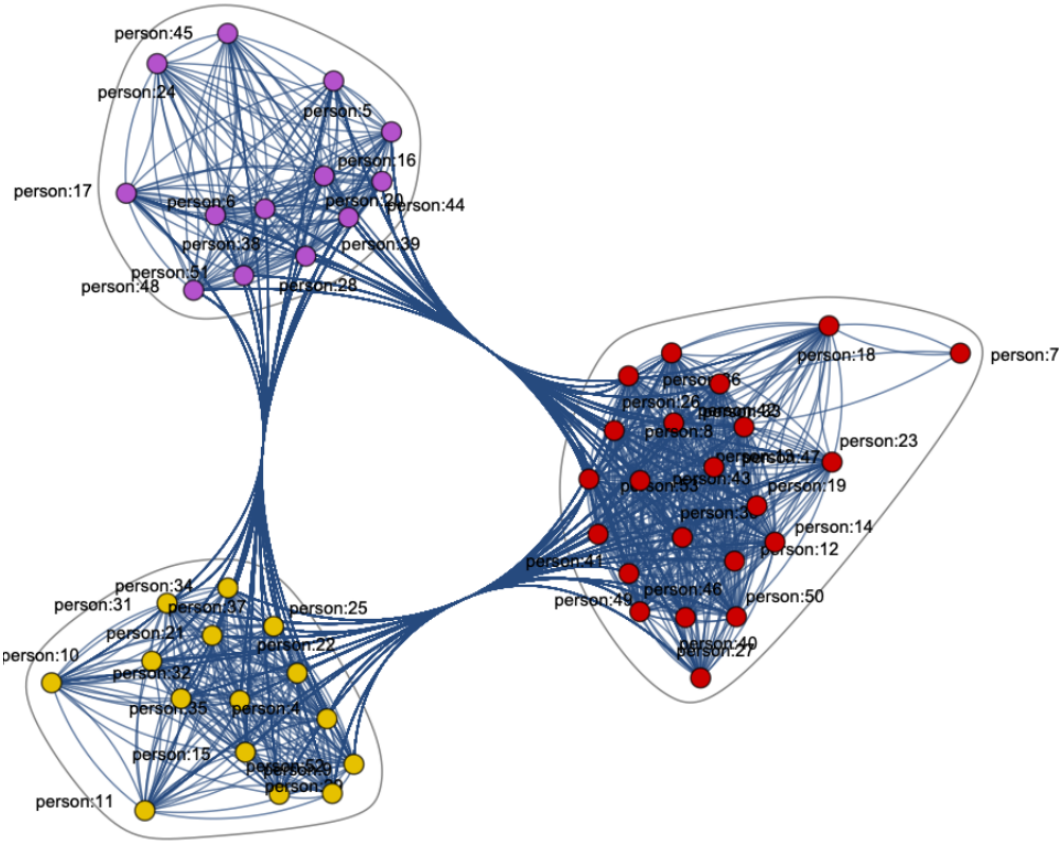

Fig. 8: Communities based on the top 3 project management practices.

V. Discussion

Our earlier work found that data science projects come in many flavors. Some have a very strong machine learning element/aspect, whereas some others are mostly about finding, fetching, and organizing data. Some projects are executed by a team of two to get to a prototype level and test feasibility, while some others are part of a large software development process from the start. Furthermore, data scientists may be consultants working very focused on a specific corner for a short period of time, or if the project is in-house, the target is different. [3]

Results presented here provide further support that there is no definition that fits all data science projects. Experimentation revealed distinct communities even from this rather small sample size. Thus, when considering how to improve the processes for (big) data science projects, we should not attempt for a one process fits all kind of answer, either, but carefully consider the distinct nature of the projects. Our survey further indicates that working with big data does not significantly impact how data science projects are run. Attempting a similar experiment with big data as a defining label, we were not able to create any similar communities as were revealed with roles and project management practices. Yet, there are some interesting indicators.

In big data projects, there were relatively more sprints used than with other data science projects, but fewer retrospective meetings compared to the popularity of sprints (Figure 5). This indicates that while big data projects utilize an iterative approach, they are not following, e.g., Scrum principles. Does the amount of data naturally drive for iterative development more strongly than projects using small data? How do data scientist interpret the term ”sprint” here? Iterative development and comparability to software engineering were identified as key elements of data science projects already in our earlier work [3], and these results strongly imply that further research is required on this.

Another key element in data science projects is the utilization of multi-disciplinary teams, and having a variety of roles in the team. A particular element we discovered in our earlier work is the role of data engineer. Our survey results confirm the many-faceted role of data engineer. On the one hand, tasks considered to be those belonging to a data engineer, are ones that require a lot of effort from our respondents (see Figure 4). Data engineers are also an inherent part of the project teams in all data science projects, but even more so in big data projects (see Figure 2). Yet, data engineer was a role solely chosen by just one respondent, while 10 (20%) more chose data engineer in addition to some other role. This type of ”polymath”, where a data scientist is involved in most or all activities during the data science project has also been identified by Kim et al. [10]. Another key element is the large variance in how teams are composed. While there are small proof-of-concept projects where the team may only be composed of a couple of data scientists [3], typically there a large variety of people involved [8]. Our survey results confirm the wide variety of team members - and also how difficult it is to distinguish the roles of ”data scientist” or ”data analyst”, and how big an impact the nature of data science project really has on the composition of the team. In Figure 1 we can see that there are several answers saying that either the role ”data analyst” or ”data scientist” is *never* involved in data science projects. Taking a closer look at big data projects, however, the presence of data scientist is more apparent, and also the role of data engineer is more clearly recognized. Interestingly, also cloud developer is more often present in big data projects - perhaps an indicator of the technical needs of storing and handling the large volume of data. Further, the role of product owner is more prominent in big data projects. Is the role of product owner connected to the more popular usage of sprints? More research is required to understand this.

In conclusion, let us return to the original research questions and their answers in the light of study results.

RQ1: How structured are current data science projects? We found that there is large variance on what kind of teams are involved in data science projects (RQ1.1). Even having the role of data scientist or analyst is not given, and the role of data engineer is often mixed with other roles. Managerial roles appear to be more prominent in projects utilizing big data than in other data science projects. The typical tasks, in turn contain very much ”data engineering” (RQ1.2), which on the other hand highlights the importance of a data engineer, but raises some questions on why this role is not more clearly present in the form of a distinct team member. Typical project management practices (RQ1.3) involve frequent meetings and prioritization, but even Kanban boards and code/result reviewing were only used by half of the respondents.

RQ2: What are the typical challenges? The most typical challenges relate to unclear project goals and changing targets and difficulties in estimating the project timeline. We have previously identified that setting goals and definition of done are essential elements for a selected development approach in data science projects, while managing uncertainty and changes in goals are essential elements of the inherent experimentation involved with data science [3]. Clearly, better-defined and managed processes are required to alleviate these challenges. The third most common challenge relates to used data, but the fourth most common challenge (selected by 50% of the respondents) regards overly optimal initial belief in results/overselling. This challenge can be seen to be brought by not having strong enough communication and collaboration between sales and the data scientists/analyst (or the sales deliberately overselling or dismissing warnings from data scientists). Sales personnel was most often not involved in data science projects (see Figure 1), and our results indicate that this should be changed.

VI. Limitations and Threats to Validity

There are several threats to the validity of the study [22], [23]. They are addressed here together with possible mitigation factors. Specifically for this study, construct validity, external validity and the reliability of the work are addressed.

Construct validity: Construct validity considers how well research investigates what it means to investigate. In this study, construct validity is threatened by how representative the respondents are of data science projects in general. The threat is mitigated by a relatively large theoretical reach of the study. However, the sample size is small in respect to the size of the field and the theoretical reach. The respondents have self-selected to respond which is a mitigating factor for validity of the results.

External validity: External validity refers to how well the study results can be generalized beyond the scope of the study. While this is a study continuing prior work, the sample size of the study is still small. This threatens the overall generalizability of the results and makes the clustering especially only indicative. Further research is needed. However, we believe the study results make way to gaining better understanding of data science projects as well as building the methodology for them.

Reliability: The main threats to the reliability of the results are related to the limited sample size. As the recommender was trained with a the sample we gathered from the 50 persons, we consider the clustering to be indicative at this stage rather than conclusive. Still, we believe that the clusters give valuable insight into data science projects and can act as a starting point for further research. To enable to replication of the study, the full dataset is given online. [2]

VII. Conclusions

In our previous work, we created a conceptual model of data science projects, identifying key elements and factors defining data science projects and needing addressing when engaged in them. In this study, our goal was to deepen our understanding with a larger sample by conducting a survey. Due to a small number of responses, the results from the survey should be considered as indicative, but they do provide us with indicators for future research on the topic, as well as confirming the previous findings. We hope to initiate discussion and further research particularly on the following.

First, to improve processes for (big) data science projects, the framing of the project should be clearly identified. There is no one process that fits all data science. What the critical defining factors and possible parameters for processes are, require more research. Secondly, the roles involved in projects, while already widely studied, deserve further in-depth look. Particularly, what is the role of data engineer and how it manifests in practice? Finally, project management practices and their application needs more insight. Project management practices and the presence of supporting individuals – in particular project manager and product owner – are fundamental for improving processes.

References

[1] N. W. Grady, J. A. Payne, and H. Parker, "Agile big data analytics: AnalyticsOps for data science," in *IEEE International Conference on Big Data*, 2017.

[2] C. Hill, R. Bellamy, T. Erickson, and M. Burnett, "Trials and tribulations of developers of intelligent systems: A field study," in *IEEE Symposium on Visual Languages and Human-Centric Computing*, 2016.

[3] T. Aho, O. Sievi-Korte, T. Kilamo, S. Yaman, and T. Mikkonen, "Demystifying data science projects: A look on the people and process of data science today," in *International Conference on Product-Focused Software Process Improvement*. Springer, 2020, pp. 153–167.

[4] W. Van Der Aalst, "Data science in action," in *Process mining*. Springer, 2016, pp. 3–23.

[5] J. Saltz, N. Hotz, D. Wild, and K. Stirling, "Exploring project management methodologies used within data science teams," in *Americas Conference on Information Systems*, 2018.

[6] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects," KDnuggets, 2014, retrieved June 2020 from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[7] H. E. Brady, "The challenge of big data and data science," *Annual Review of Political Science*, vol. 22, pp. 297–323, 2019.

[8] A. X. Zhang, M. Muller, and D. Wang, "How do data science workers collaborate? roles, workflows, and tools," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW1, May 2020. [Online]. Available: https://doi.org/10.1145/3392826

[9] P. Pereira, J. Cunha, and J. P. Fernandes, "On understanding data scientists," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 2020, pp. 1–5.

[10] M. Kim, T. Zimmermann, R. DeLine, and A. Begel, "Data scientists in software teams: State of the art and challenges," *Transactions on Software Engineering*, vol. 44, 2018.

[11] R. J. Brachman and T. Anand, "The process of knowledge discovery in databases: A first sketch," in *AAAI Workshop on Knowledge Discovery in Databases*, 1994.

[12] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000.

[13] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *International Journal of Innovation and Scientific Research*, vol. 12, 2014.

[14] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: A parallel overview," in *IADIS European Conference on Data Mining*, 2008.

[15] S. Angée, S. I. Lozano-Argel, E. N. Montoya-Munera, J.-D. Ospina-Arango, and M. S. Tabares-Betancur, "Towards an improved ASUM-DM process methodology for cross-disciplinary multi-organization big data & analytics projects," in *International Conference on Knowledge Management in Organizations*, 2018.

[16] N. W. Grady, "KDD meets big data," in *IEEE International Conference on Big Data*, 2016.

[17] J. S. Saltz and I. Shamshurin, "Exploring the process of doing data science via an ethnographic study of a media advertising company," in *IEEE International Conference on Big Data*, 2015.

[18] J. Saltz, I. Shamshurin, and C. Connors, "Predicting data science sociotechnical execution challenges by categorizing data science projects," *Journal of the Association for Information Science and Technology*, vol. 68, 2017.

[19] J. S. Saltz, "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness," in *IEEE International Conference on Big Data*, 2015.

[20] A. Antonov, "Mapping sparse matrix recommender to streams blending recommender," 2019, retrieved October 2021 from https://github.com/antononcube/MathematicaForPrediction/blob/master/Documentation/MappingSMRtoSBR/Mapping-Sparse-Matrix-Recommender-to-Streams-Blending-Recommender.pdf, https://github.com/antononcube/R-packages/tree/master/SMRMon-R.

[21] W. Mathematica, "Community graph plots," retrieved October 2021 from https://reference.wolfram.com/language/ref/CommunityGraphPlot.html.

[22] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, 2008.

[23] R. K. Yin, *Case Study Research: Design and Methods*, 5th ed. SAGE Publications, 2013.

[2] https:/github.com/sgizm/DS_survey_opendata