Exploring the Process of Doing Data Science Via an Ethnographic Study of a Media Advertising Company

Jeffrey S. Saltz, Ivan Shamshurin
School of Information Studies
Syracuse University
Syracuse, NY, USA
{jsaltz, ishamshu}@syr.edu

***Abstract*—This paper presents the results of an ethnographic study focused on how data science projects were conducted within a global media advertising company. Observations, via embedding a researcher within the team, as well as more structured interviews and surveys, are documented. Recommendations to improve the current data science methodology within the company are also discussed. Overall, there had been little focus on the team's process methodology and the suggested process improvements would result in the company's data science projects having less risk and shorter timelines. Other big data teams might also benefit from reviewing and refining their work processes, but more work needs to be done to validate this assumption.**

Keywords-Data Science; Big Data; Process Methodology.

I. INTRODUCTION

During recent decades we have been witnessed a tremendous increase in amount of data that businesses produce. Along with the size of data, the complexity and variety are also increasing, including unstructured data. As evidence of the need for businesses to be able to analyze a large amount of data is that many tools are appearing for that purpose (e.g. IBM InfoSphere Warehouse with Unstructured Data Analysis capability [1], SAP Text Analytics [2]). Another example of the substantial growth of data science is that we can also observe a significant increase in the number of data science programs offered by universities [3] [4].

However, despite the fact that there is a strong need to do data science and big data projects, often times data science teams do not have an explicit data science team-based process methodology [5]. In other words, data science teams often do not have clear idea what steps should be done first, how long each phase of a project should take and which people with what skills should be involved in the project. There are also several questions data science projects need to address (such as what kind of analysis to perform, what technique to use and how to validate the results) that are currently answered in a case-by-case manner. Having an explicit data science methodology tailored to the particular type of company (size, domain of focus, type of tasks, etc.) may significantly improve the performance of data science projects [5].

In order to understand the methodology that one company is currently using, we conducted a study of a small global media advertising software company. The research questions that we pursued were:

RQ1: What is the current methodology that they follow?

RQ2: What are some possible ways to improve the current methodology (i.e., how to make the project more efficient in time and cost)?

II. RELATED WORK

Certainly, much has been written about the use of data science and algorithms that can generate useful results. In fact, within a corporate context, data is increasingly being viewed as a strategic resource for the organization [6]. Furthermore, Tiefenbacher explored big data success stories within industry, and noted that the combination of volume, variety and velocity (3Vs) can enable new and improved business models that have not been feasible in the past [7].

However, to gain a competitive advantage from this data, one must leverage the data via analysis and insight, and it has been noted that there are significant challenges in trying to leverage the data in a strategic manner [8, 9]. One of the challenges recently noted is the lack of focus on the process teams should use to actually do a data science project [5]. Hence, not surprisingly, Bhardwaj [10] noted that teams doing data analysis and data science work in an ad hoc fashion, using trial and error to identify the right tools, that is, at a low level of process maturity.

With respect to the process of doing a data science project, the current research has focused on describing data science as a step-by-step process. While this can provide an understanding of the tasks involved in analyzing data, these descriptions do not provide a specific methodology nor a description for how a data science team should operate. For example, Jagadish [11] described a process that includes acquisition, information extraction and cleaning, data integration, modeling, analysis, interpretation and deployment. Guo approached the problem from a slightly different perspective, and provided a data science workflow framework [12]. Guo's workflow defines several high-level phases: Preparation, Analysis, Reflection and Dissemination - with each phase having a specific series of steps that can be repeated within that phase.

One way to gain a better understanding of what might be an appropriate data science process methodology is to document case studies of how teams are actually doing data science, especially within a corporate context [5]. Hence, to help start the dialog of possible data science process methodologies, this rest of this paper documents the data science process used within one company and notes some of the key process improvement opportunities.

III. DATA COLLECTION

A. Background

An ethnographic study [13], in which one of the researchers was embedded within the data science team, was conducted within a global media advertising software company headquartered in New York City. The company had a total of 100 people distributed globally. This study was based out of their NYC office, which had 20 employees. In total, the extended data science team was divided across 3 offices and consisted of 9 people.

B. Methodology

Information was collected via three different phases. At first, information was collected prior to one of the researchers being embedded within the data science team. Then, during a 9 week period, one of the researchers worked as part of the data science team, and in addition to collecting data and observing how the team functioned, actually helped the team with various tasks (such as data collection, data cleaning, implementing models and some data analysis). In this way, the culture of the organization and the challenges with how the team "did data projects" was experienced in a first-hand manner. There following 9 people were identified as the key members of the data science team:

- 2 Data Scientists
- 3 Data Operations
- 3 Software Developers
- 1 Data Engineer

As was previously noted, the team was divided across multiple locations. In particular, the software developers were in a different geographical location than the other team members, as was the data engineer. Information on how the team worked was collected via several mechanisms, including interviews, observations and a survey.

Specifically, the team was observed on an ongoing basis during the work-week, typically 8 – 10 hours per day. Information from group members was obtained via observing how the team interacted via email, phone calls, and other types of team communication. In addition, information from group members was also obtained via informal conversations. Finally, the VP of Data Science provided multiple interviews with respect to how their data science projects were performed. As part of these interviews, a semi-structured survey was also completed. This survey contained both specific Likert-type questions as well as more open-ended semi-structured questions.

C. Stakeholders

Figure 1 shows a summary of the stakeholders for the data science projects performed by the observed team. All stakeholders were measured on two criteria: interest in outcomes of the data science projects and their power to influence the project.

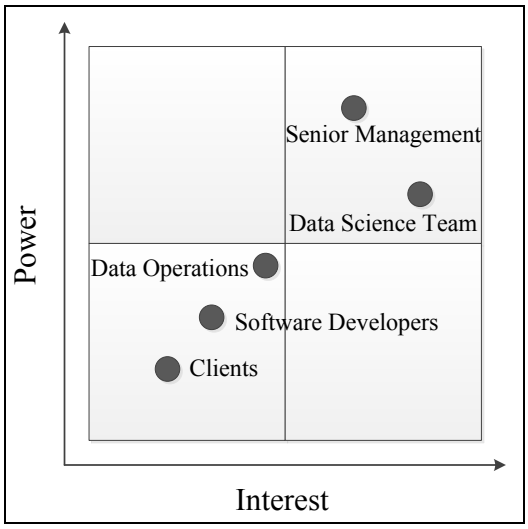

Figure 1. Stakeholder Analysis

- *Senior Management*: The management team had a high interest in the projects' outcome as well as the most influence / power. The interest of the management team was driven by their interest in trying to determine if the data science projects could solve business challenges.
- *Data Science Team*: The data science team had less power than the management team, but had the most interest in results of the projects – since it was their main task within the company. There interest was also driven by their desire to see how their methods work with real-world data, i.e. whether it works correctly and fast.
- *Data Operations*: The data operations team worked on pre-processing the data, so had some interest in the results but mainly just wanted to be kept informed of the project outcomes. They also had some influence, but not nearly as much as the previous two stakeholders.
- *Software Developers*: The developers had minimal power and interest in the data science projects. This was likely due to the fact that the software team supported multiple other teams.
- *Clients*: The clients had some interest in results of the project because they expected to use the project results, but clearly it was not "top of mind" for them. In addition, the clients had some influence on projects, typically via feedback which provided a certain amount of influence, especially for future projects.

IV. FINDINGS

A. Types of Projects

There were two types of data science projects observed within company, routine and exploratory projects. While both are briefly described below, in this paper focuses on

the projects of the second type (i.e. the more typical exploratory data science projects).

1) Routine Projects

Projects of the first type are routine efforts, which are performed on a regular basis. These projects typically include data preprocessing and the data operations group typically executes these projects. It was noted that these projects do have a standard process and well-defined methodology: each project receives data from a data provider and then transforms that data to another format. In the situation when errors are noted in the data, employees within the data operations group might need to contact the appropriate data providers. In effect, these projects include only preprocessing - one step of a typical data science project and it is probably not correct to consider this type of project as a data science effort. Analysis may be performed by the customer, i.e. after the company sends the transformed data to their customers.

2) Exploratory Projects

These projects are more typical data science projects, i.e. they include stages such as preprocessing, data analysis and implementation. They are both business and research oriented (e.g., papers are published). The data science team gets data from their data operations group and from other external sources, e.g. from publicly available sources. However, no standard methodology is used. Total duration of these projects can vary from a week to a year. For example, one project that included data analysis of US Census data took roughly one year, and most of that time was focused on pre-processing the data. Other projects spent almost no time on pre-processing the data, and focused almost exclusively on data analytics. Hence, we cannot explicitly say the typical percentage for each stage of a data science process (such as preprocessing, data analysis and implementation).

B. Current Process Methodology

During the ethnographic study the following roles and processes for doing data science projects were observed.

1) Roles

There were four teams actively involved in the data science project. However, team responsibilities were only implicitly defined, based on each person / group's organizational position within the company. Below is a brief description of each of these teams, which had one or more people, and their respective roles:

- *Data Science*: Prepares and explores the data and generates insight from the data, including tasks such as data mining and data visualization. This team included the data scientist who was the embedded observer.
- *Data Operations*: Data transformation and preparation for the data analysis (i.e., for use by the data science team).
- *Software Development*: Develop software tools to help the data science team perform data analysis; they also supported and improved previously developed tools.
- *Data Engineering*: Supports and improves the existing system and participates in some data science projects.

2) High Level Process Description

There was no explicitly defined and documented process. Rather, the people on the team had a basic understanding of each of their roles, and collectively, had an implicit understanding of the process to be followed. As shown in Figure 2, at a high level, the process consisted of three main steps preparation, analysis and dissemination.

The *preparation* phase consists of two parts: business context and data preparation. The goal of the business context part of the process is to understand the needs of the company. The data science team meets with senior management to define the goals and tasks of the project. The data preparation aspect focuses on getting data from data providers. The current process does not include any planning of human resources (such as resource allocation) or the definition of any project milestones (such as project deadlines).

Next, the *Analysis* phase is focused on defining what to do and then doing it (i.e., the data analysis). This, for example, included items such as determining if a new tool or method needs to be developed. After this was done, the analysis and modeling is typically performed. The last step within the analysis phase was generating insights from the data.

The last phase of the project was *Dissemination,* which includes communicating the results to senior management (done by data science team), and, as appropriate, sharing the results with customers.

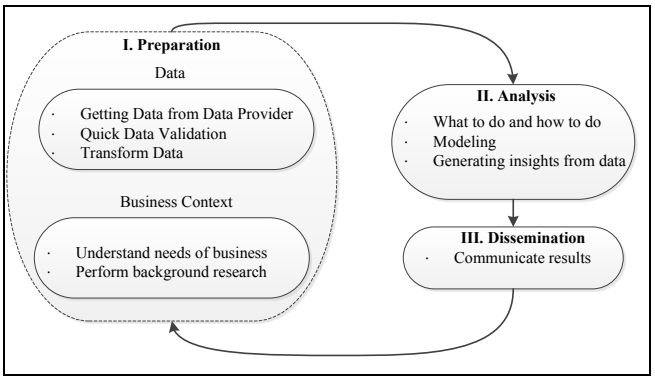

Figure 2. High Level Data Science Process

3) Process Flow Description

A more detailed flow of the process is presented in Figure 3, which includes the roles (i.e. functional groups) involved within each step.

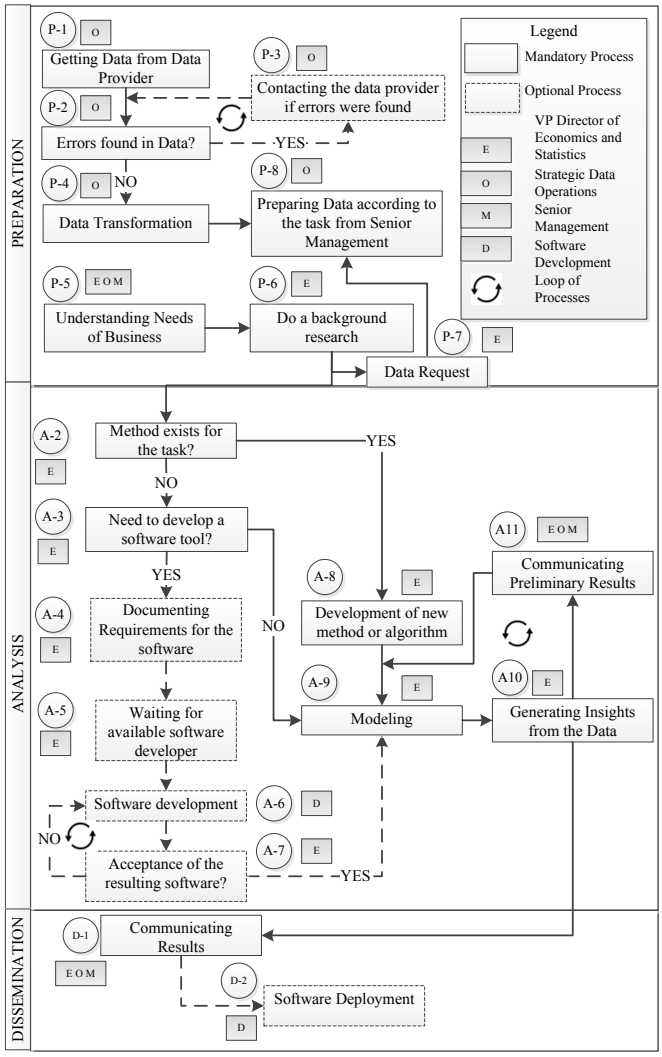

Figure 3. Detailed View of Processes

The preparation phase starts with understanding the needs of business performed by data science and operation teams. Then the data science team conducts an evaluation to determine whether the task can be solved by currently existing techniques or tools. Meanwhile, the other aspect of the preparation phase, which focuses on the data, typically starts with a 3rd party data provider (step P-1 in the Figure 2) that provides the data to the data operations team. These initial steps are performed on regular basis, in that P-1, P-2, P-3 and P-4 do not depend on other projects. The data provider sends raw data in a format that is difficult to work with (often in a standard format for advertising data providers). Then the Data Operations team transforms the data into a format that can be easily analyzed (step P-4). If during this stage errors in data are found then the Data Operations team will contact the data provider to solve the issue (step P-3). This data validation is done, for example, by visually observing a random sample of data and checking whether the source attribute names are correctly mapped to the target attribute names. Step P-7 represents the joining of the two preparation work steams – where the data science team requests specific data, in a specific format, that was obtained by the data operations team. Some other projects involve using external publicly available data. For example, there were two projects which comprised analysis of datasets from the US Census Bureau and the US Bureau of Labor Statistics. We did not present this in our diagram because the diagram highlights what was observed during our study.

The next phase is Analysis. The first item to note is that there is no discussion about the possible questions to explore, as VP of Data Science independently decides this question. However, depending on the analysis to be done, this phase starts with one of the two following options. If there is no current algorithm or method available that the company needs (as determined by the VP of Data Science) or, if the existing algorithm is not efficient enough, then the data science team develops the new method. In this case, the project includes the software development phases (steps A-4, A-5, A-6, A-7). In other words, the data science team documents the software requirements, sends those requirements to the software developers (who are remote from the data scientists), and then the data science team waits for software to be developed. Note that the software team supports many groups, so there is a prioritization process that is not fully transparent to the data science team. The software team can also be asked to develop, for example, helpful tools for generating test datasets or tools for running new methods. Also, if required, the software team can work on embedding new analytic results into the production system which the company's customers use for accessing data and doing basic data marketing analysis, e.g. cross-tabbing, clustering and optimal target audience selection. The software developers work on the task and then present the results back to the data science team. That team reviews the work and may accept the result or ask the software team to improve the requested software; hence there is a small loop (step A-7). Typically the reviewing and QA (Quality Assurance) testing process goes as follows. First, in terms of calculated output, it is either correct or not, therefore QA is comprised of checking the output of the tool to "hand calculated" results. However, in terms of the design and allowing the software to run in different modes, the review process is more agile and incremental. The next step in the analysis phase is modeling (A-9) and generating insights from the data (A-10), which are both performed by the data science team. Currently the VP of Data Science performs the quality assurance task. This includes QA of preprocessing, making sure that all the assumptions about models and data made were correct, as well as ensuring that no errors were made in interpretation of the results. This QA

process is not documented, but rather, relies on the significant expertise of the VP of Data Science. During the analysis phase, regular updates are provided to senior management by the data science team (step A-11). Hence, we note the process loop in this part of the diagram.

Finally, the dissemination phase occurs, where the insights and results that were generated (step A-10) are communicated to the senior management (D-1) at the final meeting of the project. Some projects may also include modification of the previously existing analytics module or deployment of a new module with new functionality to the system that the customers are already using (step D-2).

4) Current Process Maturity

The Capability Maturity Model (CMM) comprises five levels of maturity of processes that might exist within organization [14]. The goal of using CMM is to understand the process maturity for software development teams. However, one can also use CMM for other domains. For example, CMM has been adapted for data science management [15]. In that work, the CMM levels were mapped to those for scientific data management.

Based on observations of the company's data science processes, the current process maturity observed within the company falls somewhere between the first and the second level category (i.e. some items are documented and repeatable). The reason for this is the fact that many projects are unique in terms of ideas and methods, and that they are believed to be creative. Hence, it is believed that it would be hard to fit those into a standard model and hard to document the processes in detail. Another reason is that often times the analysis is performed by only one individual – the VP of Data Science. In this case, there is no perceived need to document how the analysis was performed and what assumptions were made about data. The results may be documented via published papers, where the results can be presented to the statistical or media/advertising community. One final item to note with respect to process maturity, it was reported that other offices within the company likely have standard documented processes for data projects.

V. DISCUSSION

A. Observed Issues / Challenges

As was stated previously, routine projects do have well-defined methodology and they are working without any known issues. However, the study did reveal several issues and challenges in the current data science project methodology.

First, there were no specific milestones or deadlines for the whole data science project or for any individual phase of the project, including remote work of the software developers. Creating project management deadlines for the remote software team would likely speed up their performance because it will guarantee that the team that started to work on the task would most likely not be switched to another project, which is definitely currently causing delays. It is not clear if milestones for the whole project would be useful (ex. increased communication across the team, escalation of potential issues).

A different challenge is related to better project organization and planning. For example, during one of the projects, the team realized that in order to implement a good prediction model, they needed to have extra data with more detailed information (i.e., the data science team needed more data). The team did not have time to get the data, and therefore they decided to use a less effective model. If the team was aware about this issue at the start of the project, then they might have been able to obtain the desired data.

Another issue is related to the fact that the developers are geographically distributed and there is a 10 hour time difference between data science team in NYC and the office where developers are located. Currently, whenever the data science team needs to have a task completed, they send a request to the developers, but the developers typically respond the following day. A related issue with respect to the remote developers is the fact that the developers are involved in several projects at the same time. Therefore the scheme above includes waiting for the availability of the software developers (step A-5). Once the data science team needs a bug to be fixed or to have a new tool developed, that team must wait for a developer who is familiar with that particular task. If the developers are busy because of other projects, then the data science team needs to either wait for a resource to become available or ask, for example, to have a bug fixed by a person who is not familiar with the task. This clearly causes delays.

Lastly, the study revealed that some steps of the data transformation are manually performed (ex. step P-4). The problem with that is not only is there a delay in the project but also the manual process might cause errors. It is not possible to make the process fully automated but there are ways to decrease the degree of manual work.

B. Possible Process Improvements

In addition to documenting the organization's current state process methodology, based on our observations, the following recommendations were suggested as process improvements. The suggestions were shared with the VP of Data Science. The goal of developing the suggestions, and then sharing them with the VP of Data Science, was to gain insight into the viability of doing these process improvements.

1) Documenting the current process

By documenting and then sharing the current process, team members might be able to offer suggested process improvements, or at the least, understand the end-to-end process. This could easily be done by leveraging the work done via this study. This would be especially useful if the data science team grows.

2) Better structuring developer interactions

As of now, the software team is spread across many projects and if the data science team needs a bug to be fixed, or have the software team implement a new capability (e.g., change an interface or add functionality), then the data science team often has to wait (to let the developers finish other projects). Therefore, the company should plan all the software development in advance.

3) Imposing deadlines

Establishing project milestones would likely be especially valuable for tasks done by the software developers. But milestones and dates would likely prove helpful across the team in that deadlines would help set expectations about what level of effort is required, and during the project, what level of effort is required to finish the project.

4) Process Automation

One of the main work tasks for the data group is to transform datasets from one data format to another one. According to the interviews and observations, some stages of this work are currently performed manually. One of the reasons for that is there are problems with different coding schemes, e.g. names of variables are not consistent throughout the datasets. Datasets come from different sources and they might have the same variables and columns but these may be written in several different ways. Also abbreviations may occur. One suggestion would be to help enable the process automation would be to implement automatic or semi-automatic data transformation, e.g. with using NLP for fuzzy matching of variable names.

5) Better Preparation to Understand Requirements

One of the biggest challenges in a data science project is that often the team does not know what data might be needed or have the required data to build a good model. Based on the projects that the company recently faced, they did have this challenge. There are two strategies that might be helpful. First, prior to starting the project, establish a connection with the end client, to better understand what would be of value to the actual business user. If this is not possible, one could try to establish those connections during the project (but often there is not enough time). In addition to better understand the client needs, the second strategy is to do a "data needs" analysis. So, for example, prior to starting a project, list all of the parameters of the model and map them to the sources (e.g. A goes from dataset1, B goes from dataset2, C comes from model T). Of course, one challenge is that models and parameters often evolve during the project.

6) Possible Longer Term Recommendations

One item to note is that several processes are not defined. This is not currently an issue, since the team is small, and the processes not yet defined are done by VP of Data Science. However, as the team grows, this might become an issue. Hence, in addition to the suggestions above, other process improvements would make sense if / when the data science team grows (ex. prioritization of possible analysis's to be performed).

C. Feedback on Suggestions

Feedback from the company, with respect to discussing the suggestions to improve the process methodology, is summarized in Table 1. During this discussion, the top five suggestions were rated on a 1-4 scale, with the following meaning:

1. The suggestion makes no sense / impossible to implement
2. The suggestion does not make much sense / hard to implement
3. The suggestion makes sense / easy to implement
4. The suggestion makes a lot sense / very easy to implement

TABLE I. FEEDBACK ON SUGGESTIONS
(RATINGS ARE ON A 1-4 SCALE, WITH 4 GETTING THE HIGHEST SUPPORT)

Suggestion	Make sense?	Can Implement?	Short or Long Term?
Documenting the current process	4	4	short
Better structuring developers interactions	3	2	long
Imposing deadlines	4	3	long
Process Automation	4	3	long
Better Preparation	3	3	short

As noted in Table I, most of the suggestions made sense to the organization. For example, documenting the current process was thought to be easy to implement and was perceived to add value - especially for new data analysts who will join the team when the team will grow. Imposing deadlines also made sense, especially for the software development team, but it was thought it would be hard to apply to the data analysis phase due to creative nature of this work. Automation was thought to probably be one of the most necessary and helpful suggested improvements for the company's data processes. This will require hiring new people to solve, develop and test the tools for automation. Finally, better preparation is mostly about improving data collection. While it is hard to know in advance what data the project will need, thinking about this early in the project would be helpful, and the company hopes to apply this suggestion in the near term.

D. Effective Practices Observed

During study several "micro practices" were identified that were effective and could be generalized and potentially applied to other companies. These include:

- *Pre-processing:* There was often a step, that was standardized when possible, to do as much processing prior to data analysis. Within the company, this was a very mature, structured and documented process except in situations when complicated preprocessing was required in this case, it was performed in an ad-hoc method by the VP Director of Data Science.
- *Frequent dialog with senior management:* The data science team had frequent discussions with Senior Management as well as with the Data Operations team. These discussions helped the Data Science team prioritize tasks and refine the analysis to be done during a project.
- *Engaging Senior Management*: It was very helpful that the senior management team was interested and engaged about the data science projects and also understood the main concepts and general ideas used in the projects.
- *Using a defined SDLC with the software team:* The software team uses a traditional software development life cycle (ex. documenting requirements, quality assurance testing). This has helped ensure that the data science teams communication with the software developers is effective and fast.

VI. Conclusion

This paper reports on the results of an ethnographic study within a global media advertising company. During this study, we observed that the data science team did not follow any specific methodology to work on their data science projects. For example, the projects typically did not have project schedules, which among other issues, made working with remote team members difficult.

The study was focused on determining if there was a need for an improved methodology for doing data science projects within this small marketing company as well as if any suggestions made would make sense to the company's management team. Overall, it was clear that there had been very little focus on the team's process methodology prior to our study. As is typical of many data science and big data teams, their focus was on "the analytics". Our work demonstrated that a small incremental effort on improving the team's process methodology proved to be beneficial to this organization. This is probably not unique to this organization, in that other teams would also likely benefit from reviewing and refining their work processes, but more work needs to be done to validate this assumption.

While we have documented the organization's current data science process, which answers our first research question ("What is the current methodology that they follow?"), this process does not have a formal name, nor was the process well documented. In summary, we were able to identify several effective "micro practices" that might be useful within other organizations. We also noted several possible process improvements (which answers our second research question). These observations might also be applicable to other data science work teams.

Interestingly, while several suggestions were well received by the organization, some of the other suggestions were not adopted, largely due to the small size of the current organization. Specifically, the current data science team involved in the analysis phase of the project is often small and sometimes just the VP of Data Science (this might not be the case for projects in offices at other locations). Therefore, one person often performed the key steps within the analysis phase of the project. In this case, there was no perceived need to document processes such as validating assumptions about the data or choosing an appropriate analytical method. However, since the company is successful and growing, there is a significant chance that the data science team will need to grow, and in this case, then it was agreed that a more precisely defined methodology would be appropriate.

Finally, the most pressing next step to better understand potentially useful methodologies for data science projects would be to study additional organizations. Specifically, it would be interesting to examine if the suggestions and feedback from this study are related to the current size, organizational structure or domain of the company, and if there are any patterns observed across the organizations doing data science projects.

References

[1] "InfoSphere Warehouse", DOI=http://www-01.ibm.com/software/data/infosphere/warehouse/unstructured-data-analysis.html

[2] "Exploit full-text search and analyze unstructured data – with text analytics for Big Data", DOI=http://www.sap.com/solution/big-data/software/text-analytics/index.html

[3] O'Neil, M., "As Data Proliferate, So Do Data-Related Graduate Programs," The Chronicle of Higher Education, February 2014, DOI= http://m.chronicle.com/article/As-Data-Proliferate-So-Do/144363

[4] Violino, B., "The Hottest Jobs In IT: Training Tomorrow's Data Scientists," Forbes, June 2014, DOI=http://www.forbes.com/sites/emc/2014/06/26/the-hottest-jobs-in-it-training-tomorrows-data-scientists/

[5] Saltz, J.,"The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness," 2015 IEEE International Conference on Big Data, in press.

[6] Wade, M., and Hulland, J., "Review: The Resource Based View and Information Systems Research: Review, Extension, and Suggestions for Future Research," MIS Quarterly, 28(1), 2004, pp. 107–142.

[7] Tiefenbacher, K., Olbrich, S. , "Increasing the Value of Big Data Projects--Investigation of Industrial Success Stories," 48th Hawaii International Conference on System Sciences (HICSS), 2015.

[8] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A., "Big Data: The Next Frontier for Innovation, Competition, and Productivity," 2015, McKinsey & Company.

[9] Chen, H., Chiang, R., and Storey, V., "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly Executive, 36(4), pp. 1165-1188.

[10] Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A., Madden, S., and Parameswaran, A., "DataHub: Collaborative Data Science & Dataset Version Management at Scale," Biennial Conference on Innovative Data Systems Research (CIDR), 2015.

[11] Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J., Ramakrishnan, R., Shahabi, C.,"Big data and its technical challenges," Commun. ACM 57, 2014, pp. 86-94.

[12] Guo, P., "Data Science Workflow: Overview and Challenges," Commun. ACM Blog, Oct. 2013, DOI= http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

[13] Hammersley, M.,"Ethnography: problems and prospects', Ethnography and Education, vol. 1, no. 1, 2006, pp. 3–14

[14] Paulk, M. C., Curtis, B., Chrissis, M. B., Weber, C., "Capability maturity model Version 1.1," IEEE Software, 10(4), 1993, pp 18-27.

[15] Crowston, K., Qin. J., "A capability maturity model for scientific data management: Evidence from the listeratute," Proc. American Society for Information Science and Technology, 2011.