Factors that Influence the Selection of a Data Science Process Management Methodology: An Exploratory Study

Jeffrey Saltz
Syracuse University
jsaltz@syr.edu

Nicholas Hotz
Indiana University
nhotz@iu.edu

Abstract

This paper explores the factors that impact the adoption of a process methodology for managing and coordinating data science projects. Specifically, by conducting semi-structured interviews from data scientists and managers across 14 organizations, eight factors were identified that influence the adoption of a data science project management methodology. Two were technical factors (Exploratory Data Analysis, Data Collection and Cleaning). Three were organizational factors (Receptiveness to Methodology, Team Size, Knowledge and Experience), and three were environmental factors (Business Requirements Clarity, Documentation Requirements, Release Cadence Expectations). The research presented in this paper extends recognized factors for IT process adoption by bringing together influential factors that apply to data science. Teams can use the developed process adoption model to make a more informed decision when selecting their data science project management process methodology.

1. Introduction

Data science develops actionable insights from data by encompassing the entire life cycle of requirements, data collection, preparation, analysis, visualization, management and the preservation of large datasets [1]. This broad view embraces the notion that data science is more than just analytics in that it integrates a range of other disciplines including computer science, statistics, information management and most notably, big data engineering.

Much of the published data science research has focused on the technical capabilities required for data science; unfortunately, published research has not focused as much on the topic of managing data science projects [2]. For example, in a broad literature review, no research was found on improving data science team project management [3]. While there has been some recent research on this topic, data science project management is an area of research that is just starting to be explored.

This is true even though it has been observed that these projects are non-trivial and require well-defined processes [4]. In fact, it was recently noted that minimal research was available on the effectiveness and impact of the different possible methodologies that data science teams use. It was also noted that no research was identified that focused specifically on evaluating a methodology/framework that supports the design and implementation process of data science projects [5].

The research that does exist on data science project management reveals that data science teams generally suffer from immature processes, often relying on trial-and-error and Ad Hoc processes [6, 7, 8]. In fact, in a recent survey, 82% of the data scientists noted that they did not follow an explicit process; yet 85% of those respondents thought that their results would improve with a more systematic process methodology [9]. Furthermore, in Cao's discussion of data science challenges and future directions [10], it was noted that one of the key challenges in analyzing data includes developing methodologies for data science teams. Gupte [11] similarly noted that the best approach to execute data science projects must be studied.

Hence, not surprisingly, it has been reported that project management is a key challenge for successfully executing data science projects and that a key reason many data science projects fail is not technical in nature, but rather, the process aspect of the project [12]. For example, Espinosa and Armour [13] argue that task coordination is a major challenge for data projects. Likewise, Chen, Kazman and Kaziyev [14] conclude that coordination among business analysts, data scientists, system designers, development and operations is a major obstacle that compromises big data science initiatives. Angée et al. [4] summarized the challenge by noting that it is important to use an appropriate process methodology, but which, if any, process is the most appropriate is not easy to know.

Industry also acknowledges these challenges. For example, Domino Data Lab blames "gaps in process and organizational structure" as a primary culprit in project failure [15], and John Akred, Co-founder of Silicon Valley Data Science, explained that "We've met a lot of data science teams that understand how to do the data science,

URI: https://hdl.handle.net/10125/70728
978-0-9981331-4-0

but they don't have any real method of managing the data science project" [16].

Leveraging the Project Management Institute's [17] definition of project management ("a temporary endeavor undertaken to create a unique product, service or result") and the previously noted description of data science, a data science project management methodology (DS-PMM) is defined as:

> *A system of practices, techniques, procedures, and rules used to guide a temporary team-based endeavor that collects and analyzes data to solve problems by developing actionable insights.*

Thus, to help move the field forward, this research aims to help data science teams move beyond using an Ad Hoc process by providing a model that explains why teams select different DS-PMMs. By knowing the factors that influence the selection of a DS-PMM, a team could take a more structured approach to identify and select a process that works best given its specific situation. Therefore, this research explores the following key question:

What factors influence a team's selection of a DS-PMM?

The next section provides some background context on process methodologies used in data science projects as well as the factors teams use to select software development process methodologies. *Section 3* then summarizes the Technology-Organizational-Environmental framework employed in this study. Based on the data gathered in our interviews, *Section 4* notes the findings by describing the model's eight factors that drive a team to select a DS-PMM. Finally, *Sections 5 and 6* present the findings, conclusions, research limitations and possible next steps.

2. Background

This section first describes the six most common DS-PMMs that were identified in the literature. It then reviews research with respect to selecting a DS-PMM as well as the factors that teams use to select a software development process methodology since those factors might be similar to the factors used to select a DS-PMM.

2.1. DS Project Management Methodologies

Below, are six common DS-PMMs encountered in published research:

- **Kanban**: Visually represents tasks on a board and achieves agility by minimizing work-in-progress.
- **Scrum**: Divides work into sprints (mini-projects up to one month long), defines four meetings (daily standup, sprint planning, sprint review and sprint retrospective) and three roles (product owner, development team and scrum master).
- **Research-Agile Hybrid**: Starts with an open-ended research phase (to do exploratory analysis) followed by a more formalized agile phase (typically similar to Scrum). The entire process can be iterated as needed.
- **Waterfall-Agile Hybrid**: Blends elements of Waterfall/phased processes (to do tasks such as data repository buildout) and elements of agile phases (typically similar to Scrum) to incrementally deliver insights. These phases can be concurrent or phased with Waterfall typically proceeding Scrum in the project life cycle.
- **CRISP-DM**: The *CRoss Industry Standard Process for Data Mining* has six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. It is a phased approach for data mining but with some flexibility that encourages a team to loop back to a previous phase when needed.
- **Ad Hoc**: The term used for groups that do not follow a process or use a process that is undocumented.

In short, Kanban and Scrum are agile approaches that focus on rapid delivery and progressive elaboration of requirements while Waterfall and CRISP-DM emphasize significant up-front planning. Research is a more open-ended approach. The hybrid approaches blends the other approaches mentioned [18].

2.2. Research on Selecting a DS-PMM

A recent research effort explored the factors that can influence a team to use, or not use, a data science process methodology [19]. It found eight positive factors with respect to relative advantage and compatibility and two negative factors with respect to complexity. However, this research did not explore the factors driving the use of one methodology versus other methodologies, but rather, the use of any DS-PMM (as compared to using an Ad Hoc process methodology). Verma and Bhattacharyya [20] reported on the factors driving the adoption of big data analytics initiatives, but not the process used to do the big data analytics projects.

In addition, one other study explored the strengths and weaknesses of different DS-PMMs [21]. In that study, Kanban was shown to be more effective than Scrum within a data science context mainly due to Kanban's ability to better handle exploratory analyses, as compared to Scrum that requires time-boxing sprints.

Beyond this, there has been no identified research relating to how teams select their DS-PMM. The lack of research in this area is consistent with Ahmed et al.'s [5] observation, in that Ahmed's research did not identify any other research that was specifically focused on evaluating a DS-PMM.

2.3. Software Team Selection Factors

The factors teams use to select a process methodology has been explored within the field of software engineering. Since data science and software development have some commonalities (ex. creating code), and some differences (ex. data science's focus on more open-ended exploratory analysis), there might (or might not) be different factors when selecting a process methodology across these two domains. Regardless, to provide a broader context of process methodology selection factors, the key factors that influence a software team's selection of a project management methodology are summarized below.

Vijayasarathy and Butler [22] identified organizational (annual revenue and number of employees), project (project budget and criticality of the effort), and team factors (number of teams and team size) that drove the use of different software development methodologies. An earlier effort also explored the factors for choosing a software development methodology [23]. This work noted that there were many possible software process methodologies and choosing which one to use was not an easy task, but was very important since the success rate of software projects increases with a methodology that caters to the specific characteristics of a project. The analysis identified key factors that influence the selection of a methodology for a specific project, including: clarity of the initial requirements, accurate initial estimation of costs and development time, incorporation of requirements changes during the development process, obtaining functional versions of the system during the development process, software criticality, development costs, length of the delivery time of the final system, system complexity, communication between customers and developers, and size of the development team.

Finally, during the deployment of a process methodology, organizations typically focus more on the technical rather than the equally important human aspects of process model selection [24]. Via a literature review and interviews with industry professionals, these researchers identified 27 different success factors across four categories: Organization, People, Process, and Product.

3. Theoretical Framework

From an IT perspective, innovation refers to a new practice or operational idea [25]. Hence, from a theoretical perspective, the selection and use of a DS-PMM is a *process innovation*. Oliveira and Martins [26] noted that most studies on IT adoption leverage one of two frameworks, either the Technology-Organization-Environment (TOE) framework [27] or the Diffusion of Innovation (DOI) framework [28].

Many, such as Verma and Bhattacharyya [20], have suggested that TOE is more appropriate because it includes the organizational context that can influence the adoption and implementation of that process [29]. The TOE framework thus provides a useful way to distinguish between the innovation's inherent qualities and the adopting organization's motivations, capabilities, and broader environmental context [30]. In addition, the TOE framework brings together the technology and the organization focus, something unique among the models.

According to DePietro, Wiarda, & Fleischer [27], the three contextual factors (technology, organization and environment) present both opportunities and constraints that can influence the firm's level of technological innovation. Technical factors describe both the technologies and practices (i.e. processes) that influence individual, organization and industry adoption [30]. Organizational context represents the internal factors of an organization influencing innovation adoption and implementation [27]. The environmental context represents the environmental conditions in which the organization conducts its business, service or process and can include the demands of trading partners and customers, professional associations, as well as legal frameworks [31].

Although TOE has not been used when exploring process adoption within a data science context, it has been employed in related areas such as enterprise resource planning, knowledge management system, customer relationship management, data warehousing, business intelligence and cloud computing [20]. Thus, the TOE framework has a solid theoretical basis, consistent empirical support and the potential of application to IT adoption [26]. Given these advantages, the TOE framework was selected to be used for this research effort.

3.1. Technical Factors

Data science projects have often been described via the "4 Vs". Specifically, the data's volume (size of data to be analyzed), variety (number of sources and type of data), velocity (speed of data collected/generated that needs to be analyzed), and veracity (the trustworthiness of the data). However, the 4 Vs are sometimes not sufficient to describe a project. Hence, we focus on Saltz et al's [37]

characterization of data science projects, in which there are two key dimensions. One key attribute is the level of discovery and the other is the level of technical infrastructure required for the project (this second attribute includes, from a project management perspective, the four Vs).

The level of discovery, typically due to the need for exploratory data analysis, is the general process of discovering insights from, identifying the value of, or assessing the validity of a data set. In addition, the level of technical infrastructure can cause significant project management challenges. Furthermore, it has been observed by others, such as Verma and Bhattacharyya [20], that "getting the right data in time and trusting the information for decision making was found as a critical issue". Hence, data collection and cleaning, which is the amount of time, effort and coordination needed to collect and clean data is likely a significant factor that impacts which DS-PMM a team selects. Hence, it is hypothesized:

H1. The level of exploratory analysis required within a project is a key factor in the selection of a DS-PMM.

H2. The level of technical infrastructure required within a project is a key factor in the selection of a DS-PMM.

H3. The level of data collection and cleaning required for a project is a key factor in the selection of a DS-PMM.

3.2. Organizational Factors

Team size is a factor that has often been identified as a factor driving the selection of the methodology used by a team [22, 23]. In addition, the knowledge and experience of a team, while also not specific to data science, has been noted as a key factor to successfully deploy a process methodology for software development [24]. Receptiveness to the methodology is also likely a key factor in that teams that are receptive to using a methodology will be more likely to adopt that methodology, compared to teams that resist adopting a methodology. This has been noted in other contexts, such as Bayona-Oré et al. [24], who describe three components of receptiveness (positive attitude toward change, motivation for the use of processes, and willingness to learn new skills). Thus, we hypothesize the following:

H4. The size of the project team is a key factor in the selection of a DS-PMM.

H5. The knowledge and experience of the team is a key factor in the selection of a DS-PMM.

H6. The team's receptiveness to a methodology is a key factor in the selection of a DS-PMM.

3.3. Environmental Factors

Business requirement clarity is the extent to which the final project deliverables are agreed upon and understood early in the project. It has been noted that the team will need to be able to iterate if the requirements are not clearly understood/defined, and thus it was identified as a factor driving the selection of the methodology used by software teams [23].

In addition, the documentation requirements for a project have been observed as a key project attribute by Bayona-Oré et al. [24], in the context of software projects. However, there are sometimes documentation requirements that are unique to data science. For example, there might be a documentation need that ensures that the data used in a machine learning model was acquired in a proper way and that there were the necessary processes in place to ensure that there was no bias in the output of the predictive algorithms [32]. Finally, release cadence expectations, which is the rhythm at which the team needs to deliver output, varies drastically based on the project, customer needs and organizational expectations, and has also been mentioned as a key project characteristic by Geambaşu and Bayona-Oré [23]. Hence, we hypothesize the following:

H7. The business requirement clarity is a key factor in the selection of a DS-PMM

H8. The document requirements for a project is a key factor in the selection of a DS-PMM

H9. The release cadence expectation of a project is a key factor in the selection of a DS-PMM

3.4. Conceptual Model

Based on our previously noted hypotheses, as shown below in *Figure 1*, our conceptual model includes nine factors across the three TOE themes. Each factor can positively or negatively impact the willingness of teams to adopt a given process methodology.

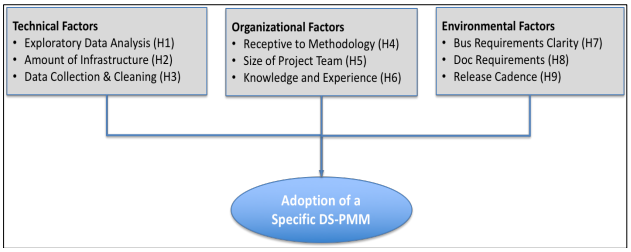

Figure 1: Conceptual DS-PMM Selection Model

4. Methodology

An exploratory interpretive approach was conducted to investigate the adoption decision of a DS-PMM for a data science project. An interpretive paradigm allows the researchers to develop insight and understanding into the issues related to the adoption decision of an innovation at an organizational level [33]. Furthermore, a qualitative approach is more appropriate in the context of this study as it is naturally associated with the epistemological assumptions of the interpretive paradigm and can be used to thoroughly examine a complex phenomenon in its natural setting [34].

Specifically, as shown in *Figure 2*, a two-phased approach was used. The first phase developed a theoretical model based on a literature review, which was discussed in the previous section. The second phase leveraged qualitative interviews to validate and refine the theoretical model.

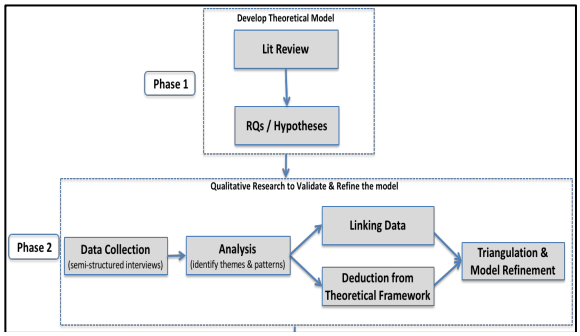

Figure 2: Research Design and Evaluation

4.1. Data Collection

Fourteen organizations were selected to be part of the study. They were identified via a selective sampling method to ensure that there was diversity across several theoretically salient factors [35], including organizational size, data science team size, team project role and business domain.

Contextual details about each interviewee are summarized in *Table 1*. The *Primary DS-PMM* is based on the primary process used, even if the team did not explicitly label their process. *Team Size* includes the people working on their data science projects across a variety of roles including data scientists, data engineers, software engineers, business analysts, product managers and consultants. All the interviewees with managerial or executive titles had 5+ years of experience leading technical teams. Data Science experience ranged from G who just completed his first big data project to H who had 20+ years' experience conducting data science research and managing a data science company.

Table 1: Interviewee Summary

ID	Interviewee Role	Primary DS-PMM	Industry	Team Size	# of Employees
A	Machine Learning Lead	Research-Agile	IT Services	8	100,000
B	Project Manager	Waterfall - Agile	Consulting	6	100
C	Algorithmic Trader	Ad Hoc	Capital Markets	13	500
D	Data scientist	Ad Hoc	Venture Studio	1	250
E	Product Manager	Scrum	IT Services	9	15,000
F	Data Manager	Waterfall -Agile	Pharmaceutical	8	50,000
G	Program Manager	Scrum	Biotechnology	8-10	10,000
H	Chief Scientist	Research-Agile	IT Services	4	30
I	President	Scrum	Consulting	10	10
J	Lead Data Scientist	Ad Hoc	Media	6	1,000
K	Senior Manager	CRISP-DM	Consulting	15	15,000
L	Data Science Manager	CRISP-DM	Financial Services	40 - 45	250,000
M	Data Science Manager	CRISP-DM	Financial Services	10	2,500
N	Chief Data Officer	Kanban	City Government	2 - 6	1,000

4.2. Data Analysis Process

One-on-one, semi-structured interviews were conducted either in person or via phone/video calls. The open-ended, semi-structured interview enabled the authors to ask probing and follow-up questions, allowing for a more in-depth understanding of the phenomenon under investigation. Each interview lasted 30 to 60 minutes. The objective for these interviews was to collect information about the different factors of technological, organizational and environmental context influencing the adoption of a DS-PMM. During each interview, the initial questions covered the participants' background, roles and responsibilities. Then, the focus shifted to understanding the interviewees' thoughts and practices with respect to how their teams executed data science projects, their process methodology, and why they used a specific methodology. Their project challenges were also explored, including the challenge in using a process methodology and the key characteristics of their projects. However, the interviews did not cover specific algorithms, nor the technologies used.

The analysis of the interviews leveraged the guidelines suggested by Braun and Clarke [36] for thematic analysis of qualitative data, which involves six steps: familiarizing oneself with the data, generating initial code, searching for themes, reviewing themes, defining and naming themes and producing a report.

5. Findings

As shown below in *Table 2*, the analysis identified eight factors across the three TOE themes. Each factor can positively or negatively impact the willingness of teams to adopt a given process methodology.

Table 2: Summary of Findings

Theme	Hypothesis	Factor	Support	Key Factor
Technology	1	Exploratory Data Analysis	Yes	A, D, E, H, J
	2	Technical Infrastructure Amount	No	--
	3	Data Collection & Cleaning	Yes	L, M
Organization	4	Receptiveness to Methodology	Yes	A, G, H, J, L, M, N
	5	Size of Project Team	Yes	A, C, D, F
	6	Knowledge and Experience	Yes	A, C, D, G, H, J, N
Environment	7	Business Requirements Clarity	Yes	B, D, F, I, N
	8	Documentation Requirements	Yes	B, F, K
	9	Release Cadence Expectations	Yes	C, E, I, L

Figure 3 shows our refined model based on these supported factors. This derived model helps to explain how teams selected their DS-PMM. The rest of this section describes each factor and how that factor influenced each organization's selection of a DS-PMM.

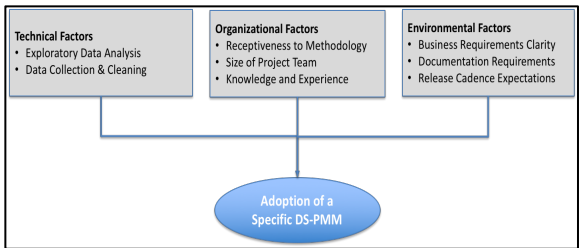

Figure 3: Factors influencing the Adoption

5.1. Technical Factors

5.1.1 Exploratory Data Analysis

This type of analysis is central to many data science projects (e.g., it helps the team understand the problem-solution space). However, scope and schedule management are challenging because data exploration often lacks a clear set of required tasks. This creates problems with time-boxing, which led two teams to explicitly state their aversion to using Scrum (which uses time-boxes known as "sprints"). For example, a product manager (E) explained that her data science team did not "*know the level of work that would be required until after they've gotten to get into data*" which led the team to miss their sprint commitments. The product manager believed that a process without strict time-boxing would work better, but the team still used Scrum to match the cadence of other teams (See *5.3.3 Release Cadence Expectations*). Meanwhile, without such external constraints, the chief scientist at Company H refused to use time-boxing approaches because "*You cannot put a schedule on insights*."

While strict time-boxing can be challenging for highly explorative work, approaches that provide too much time freedom without deadlines might foster environments where tasks linger longer than required. Consequently, a data scientist at Company D explained that a balanced approach was needed:

> "*I do believe that having a little bit of free space for a deliverable today or tomorrow really opens up your ability to think creatively since if you put people under constraints […], that can really choke creativity. But at the same time, you can't let people have an unlimited leash*."

As this data scientist typically worked independently on highly explorative work and did not want to apply undue processes upon himself (See *5.2.2 Team Size*), he balanced the creative freedom and time constraints on a case-by-case basis in an unstructured Ad Hoc format. In contrast, Team A consisted of eight members, many with PhDs, who conducted deep learning research. The researchers wanted flexibility to conduct their research but management wanted structure to ensure the work is productionized. As such, they settled on a Research-Agile methodology that allowed for unstructured research time followed by a Scrum-like process to productionize their work.

5.1.2 Data Collection and Cleaning

Surprisingly, data collection and cleaning challenges were only mentioned in two interviews (L and M). A large financial services company (L) selected a highly-structured CRISP-DM approach to manage its data collection process, which was thought to be appropriate due to the long duration of their data buildout, but as noted by their senior data scientist, this was not ideal:

> "*We needed all the data, with history, in order to understand which attributes would be helpful in improving our predictive analytics model. However, our IT team kept saying that was too expensive and asking for a prioritized list of data requirements, which made no sense*"

As can be seen, Organization L struggled with the fact that sometimes data requirements are unknown, since part of the project was to determine what data might be relevant and which data attributes were of value. This is challenging because significant time and money is

required to store potentially useful data, but the value of that data in helping to create an accurate model might not be clear. Because of this, while the team desired a more agile approach, due to long lead times to collect, clean and store the targeted data, this was not seen as a viable alternative.

In a related example, a large investment management organization (M) also had lengthy projects to build large data repositories. They chose a CRISP-DM process because the phased approach allowed the team to establish clearly defined requirements. These well-documented requirements helped the team work through the challenging systems creation and data collection project phases.

5.2. Organizational Factors

5.2.1 Receptiveness to Methodology

Five organizations (A, C, G, H, I) expressed resistance to follow certain (or even *any*) methodologies. A program manager at Organization G explained data scientists' resistance:

> "*in the early stages of discovery there is a big resistance to process […] having a set way to walk and do your job constrains thinking outside the box and stifles creativity*".

Moreover, the chief scientist from Company H did not feel that any of the existing common methodologies were appropriate for data science, in that he thought that even a light-weight process such as Kanban "*adds an unnecessary managerial burden that doesn't really result in any improved outcomes.*" Teams such as these tended to prefer Ad Hoc processes or lightweight processes like Research-Agile. However, G, I, and L chose more structured approaches due to other factors such as large *Team Size*, ambiguous *Business Requirements Clarity* and rigid *Release Cadence Expectations*.

Organization J did not consciously resist specific methodologies, but rather, didn't understand the importance of using an appropriate DS-PMM. Hence, they did not explore which methodology might make the most sense and just used an Ad Hoc process. In contrast, three teams (B, F, E) were very receptive to various methodologies, which enabled them to choose from a broader set of DS-PMMs. They each reported that they regularly shift DS-PMMs to fit the project's needs.

5.2.2 Team Size

Larger teams tended to desire more structured DS-PMMs to handle their greater coordination challenges. This was noted by a manager from Organization L who stated that:

> "*due to our size and being geographically and organizationally divided, a process, any process, is better than no process*".

Company F also understood this factor and considered the project team size in the selection of its DS-PMM. It used a Waterfall-Agile approach for medium and large-projects but found Ad Hoc to be appropriate for projects with fewer team members. Likewise, the data scientist from D felt that his current Ad Hoc processes were sufficient for himself (since he was a team of one) but wanted to select a more structured approach as he grew his team. Similarly, the interviewee from Company H believed that a Research-Agile approach worked for small teams but that "*there really isn't a standard for how to manage and scale up data science teams*."

5.2.3 Knowledge and Experience

A knowledgeable team primarily consists of members with many years of data science experience who know what they need to do to be productive, even without significant management oversight or processes. Three interviewees (A, C, H) suggested that less structured processes tend to be selected by these mature, knowledgeable and experienced teams, while structured processes are often deemed necessary for teams with junior staff.

For example, a machine learning technical lead (A) felt that his team's loose Research-Agile methodology isn't for everyone but was effective for his team of highly motivated senior researchers. Likewise, an algorithmic trader (C) thought his team was productive despite using an Ad Hoc process because it was comprised of highly-motivated senior staff who did not need much guidance; yet, he noted that:

> "*a structured approach—especially for the juniors—really helps*" because "*you can't just throw someone who is new to the environment into a loose academic environment and expect them to produce*"

Meanwhile, the chief scientist at Company H chose a Research-Agile approach because he felt it was a natural fit for data science but recognized that his choice was challenging for his junior staff whom he had to "spoon feed" because they were unable to "work independently." He was actively searching for a more effective approach but could not find one that he felt did not conflict with the natural process of data science (See *5.2.1 Receptiveness to Methodology*).

5.3. Environmental Factors

5.3.1 Business Requirements Clarity.

Requirements can sometimes be stable and clear, such as when the program manager at G described the requirements for his recent big data project as "obvious";

however, such clarity is often the exception as most interviewees (B, C, D, E, I, J, K, L, M, N) revealed requirements ambiguity to be a challenge. When this factor was perceived as very important, those teams tended to select either an agile approach (which were designed to progressively elaborate requirements throughout the project), or an Ad Hoc approach (due to not being aware of a better option, or due to being opposed to following a specific process - see *5.2.1 Receptiveness to Methodology)*.

For example, as explained by the data scientist at Company D, because data science is "so much more ambiguous" than other domains such as software engineering, he selected an Ad Hoc process so that he could be flexible in responding to changing business needs. Furthermore, as the president of a bioinformatics consulting company (I) explained:

> "*the customer doesn't really know what they want or what is possible, and the data scientists don't know what is going be helpful and how to communicate that*."

The team countered this problem by selecting Scrum and using two-week sprints to deliver small units of value and solicit customer feedback as to whether they are on track to meet their intended needs. Moreover, the interviewee from Organization N noted that they often did not have a clear view of what to do, and how to do it, and as a result, thought that:

> "*we needed a process that could easily handle our ambiguous requirements*"

Hence, the team selected Kanban to focus on one task at a time, without needing to accurately scope that effort, and then, based on those results, select the next task to be done. However, despite requirements ambiguity, some organizations still used less agile approaches. For example, Company B still chose Waterfall-Agile (due to *Documentation Requirements* as discussed next) and Company L selected CRISP-DM (due to *Release Cadence Expectations*).

5.3.2 Documentation Requirements

Documentation requirement challenges were not mentioned by most interviewees but were critical in three companies' DS-PMM selection process (B, F, K). The pharmaceutical company (F) had to be able to prove their results with a very high degree of certainty to comply with their company's quality control and with Food and Drug Administration requirements. Meanwhile, the consulting firms (B and K) had to comply with local and regional government documentation requirements.

Although most other factors provide reasons for a team to select an agile methodology, extensive documentation requirements led these three teams to choose more traditional or hybrid approaches. Company F alternated between a series of two-week development sprints "to get to 'this is good enough'" and "a three-month Waterfall-type production cycle tacked onto that" for validation. Company B also chose a Waterfall-Agile approach, using Waterfall to manage customer-facing activities and documentation while simultaneously coordinating development with Scrum. Seeking a process with a well-structured documentation process, Company K selected CRISP-DM partly because it includes a Waterfall-style of cascading documentation reports throughout the project lifecycle.

5.3.3 Release Cadence Expectations

Different methodologies are designed to support different release cadences. Agile approaches like Kanban and Scrum can deliver rapidly while methodologies like CRISP-DM (that require detailed upfront planning) or Research-Agile and Waterfall-Agile (which require extensive research or planning before transitioning to release cycles) are unable to support rapid delivery, especially early in the project lifecycle. Depending on how they are implemented, Ad Hoc approaches can support rapid releases.

Consequently, when given the choice, teams that needed rapid releases chose agile and Ad Hoc approaches. For example, the capital markets trading team (C) used Ad Hoc to quickly respond to market conditions with minimal process constraints, and the biotechnology consulting firm (I) selected Scrum to provide frequent value delivery to its customers' set cadence. On the other hand, neither Team A nor Team H faced significant release cadence constraints. As such, both selected Research-Agile without concern for its slower release cadence.

In contrast, interviewees from two companies (E and L) felt like they were forced to use a sub-optimal process to comply with externally directed release cadences. Team E, despite continually missing sprint commitments, used Scrum with two-week sprints to match the release cadence of the software development teams. Meanwhile, the manager at L explained that:

> "*I would have liked to use a more agile approach, but I felt forced to use a methodology that worked with our IT's delivery schedule*"

Hence, this team used a CRISP-DM-like approach to synchronize their releases with the Waterfall process mandated by their IT team.

6. Discussion

Based on previous studies, this paper presented nine hypotheses about what drives a data science team's decision to adopt a specific DS-PMM. Eight of these nine hypotheses were corroborated via the TOE framework and interviews with 14 organizations.

The eight factors from these eight hypotheses are grouped into organizational, environmental and technological themes. Note that while the technical factors are somewhat unique to data science projects, the organizational and environmental factors apply to many fields. However, these factors might have more or less importance for data science as compared to other domains.

6.1. Limitations & Potential Next Steps

This empirical study has limitations that could be addressed through additional research. For example, this is an exploratory study conducted through in-depth interviews. Hence, the results can further be verified through a quantitative survey-based research.

In addition, while the organizations in this study varied across a range of dimensions, there were still limitations in the sample, such as all the organizations were based in the United States. Therefore, future research could explore additional organizations to help to refine and validate the model. Specifically, it is not clear whether the lack of support for H2 (The level of technical infrastructure required within a project is a key factor in the selection of a DS-PMM) is due to the factor itself not being important or from the limited sample size.

Furthermore, while this research focused on understanding why teams selected a specific DS-PMM, future research could explore these identified factors in greater detail such as the relative importance of each of these factors Similarly, a cross-factor synthesis could help determine the relationship among the various factors.

Finally, future research could also explore creating new or hybrid methodologies that address the weaknesses of some of the existing methodologies identified during this investigation.

6.2. Implications

As organizations try to leverage data science for insight and competitive advantage, the size of data science projects and project teams continue to grow. In addition, the results of those analysis are also of increasing importance. Hence, selecting an appropriate DS-PMM is of growing importance.

In addition to identifying eight factors that influence the decision to adopt a specific DS-PMM, another key outcome of this study was the exploration of how an organization could select an appropriate DS-PMM. Thus, the results of this study provide a guideline to managers who are either in the process of selecting a DS-PMM or have already selected a DS-PMM but might consider selecting a more appropriate DS-PMM. Just as there is no one algorithm that should be used for all data science problems, this research suggests that there is no single DS-PMM that should be selected for all data science projects.

This research also provides a vehicle to understand the unique nuances of DS-PMM selection as compared to software engineering process selection. Some factors, such as *Exploratory Data Analysis*, are more critical to data science. Meanwhile, others, such as the *Documentation Requirements*, are also important in other domains, such as software engineering, but might have unique nuances in data science projects. Yet other factors, such as project *Team Size*, are not specific to data science and have been identified via previous research efforts for software engineering.

Armed with a broad understanding of possible project management approaches, lessons from other companies, and the eight factors that can impact process adoption, this research enables teams to more effectively convert data science investments into actionable insights by using an appropriate DS-PMM for that particular project team. By using the model, teams can explicitly identify the key factors impacting process selection for their project. The result should be a more informed decision that leverages these eight factors for selecting a DS-PMM.

7. References

[1] Saltz, J. and Stanton, J. (2017). *An Introduction to Data Science*. SAGE Publications.

[2] Ransbotham, S., Kiron, D. and Prentice, P. K. (2015). Minding the analytics gap. *MIT Sloan Management Review*, *56*(3), 63.

[3] Saltz, J. and Shamshurin, I. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. *2016 IEEE International Conference on Big Data* (pp. 2872-2879).

[4] Angée, S., Lozano-Argel, S. I., Montoya-Munera, E. N., Ospina-Arango, J. D., & Tabares-Betancur, M. S. (2018, August). Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects. In *International Conference on Knowledge Management in Organizations* (pp. 613-624).

[5] Ahmed, B., Dannhauser, T., & Philip, N. (2018). A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects. In *2018 10th Computer Science and Electronic Engineering* (pp. 11-14).

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

[7] Gao J., Koronios A. and Selle S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects, *Twenty-first Americas Conference on Information Systems (AMCIS)*.

[8] Saltz, J. and Shamshurin, I. (2015). Exploring the process of doing data science via an ethnographic study of a media advertising company, in *Big Data (Big Data), 2015 IEEE International Conference on*, pp. 2098-2105: IEEE.

[9] Saltz, J., Hotz, N., Wild, N. and Stirling, K. (2018). Exploring Project Management Methodologies Used Within Data Science Teams. *Americas Conference on Information Systems (AMCIS)*.

[10] Cao, L. (2017). Data science: challenges and directions. *Communications of the ACM*, *60*(8), 59-68.

[11] Gupte, A. (2018). *Determining Critical Success Factors for Big Data Projects* (Doctoral dissertation, Purdue University).

[12] Ponsard, C., Majchrowski, A., Mouton, S., & Touzani, M. (2017). Process Guidance for the Successful Deployment of a Big Data Project: Lessons Learned from Industrial Cases. In *IoTBDS* (pp. 350-355).

[13] Espinosa, J. A. and Armour, F. (2016). The Big Data Analytics Gold Rush: A Research Framework for Coordination and Governance. *Hawaii International Conference on System Sciences (HICSS)*.

[14] Chen, H., Kazman, R. and Haziyev, S. (2016). Agile Big Data Analytics for Web-based Systems: An Architecture-centric Approach, *IEEE Transactions on Big Data*.

[15] Domino Data Lab. (2017). *The Practical Guide to Managing Data Science at Scale.* Retrieved from: https://www.dominodatalab.com/wp-content/uploads/domino-managing-ds.pdf

[16] Akred, J. (2016). *Using Agile development techniques for data science projects*. (B. Lorica, Interviewer), https://www.oreilly.com/ideas/using-agile-development-techniques-for-data-science-projects

[17] Project Management Insitute. (2017). *A Guide to the Project Management Body of Knowledge (PMBOK Guide) - 6th Edition.* Newtown Square, Pennsylvania: Project Management Intitute, Inc.

[18] Pressman, R. S., and Maxim, B. R. (2015). *Software Engineering: A Practitioner's Approach.* Singapore: McGraw - Hill.

[19] Saltz, J. (2018). Identifying the Key Drivers for Teams to Use a Data Science Process Methodology, Proceedings of the 26th European Conference on Information Systems (ECIS).

[20] Verma, S., and Bhattacharyya, S. (2017). Perceived strategic value-based adoption of Big Data Analytics in emerging economy: A qualitative approach for Indian firms, *Journal of Enterprise Information Management*, Vol. 30 Issue: 3, pp.354-382

[21] Saltz, J., Shamshurin, I. and Crowston, K. (2017). Comparing data science project management methodologies via a controlled experiment. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

[22] Vijayasarathy, L. R. and Butler, C. W. (2016). Choice of software development methodologies: Do organizational, project, and team characteristics matter? *IEEE software*, *33*(5), 86-94.

[23] Geambaşu, C. V., Jianu, I., Jianu, I. and Gavrilă, A. (2011). Influence factors for the choice of a software development methodology. *Accounting and Management Information Systems*, *10*(4), 479-494.

[24] Bayona-Oré, S., Calvo-Manzano, J. A., Cuevas, G., and San-Feliu, T. (2014). Critical success factors taxonomy for software process deployment. *Software Quality Journal*, *22*(1), 21-48.

[25] Lind, M. and Zmud, R. (1991). The influence of a convergence in understanding between technology providers and users of information technology innovativeness, *Organization Science*, Vol. 2 No. 2, pp. 195-217.

[26] Oliveira, T. and Martins, M. F. (2011). Literature Review of Information Technology Adoption Models at Firm Level. *The Electronic Journal Information Systems Evaluation*, 14(1), 110– 121.

[27] DePietro, Rocco, Wiarda, Edith & Fleischer, Mitchell (1990). "The context for change: Organization, technology, and environment," in Tornatzky, L. G. and Fleischer, M. (Eds.) The processes of technological innovation, Lexington Books: Lexington, MA., pp. 151-175. https://is.theorizeit.org/wiki/Technology-organization-environment_framework.

[28] Rogers, E.M. (1995). Diffusion of Innovation, *Free Press*, New York, NY.

[29] Baker, J. (2011). The technology-organization-environment framework, in Dwivedi, Y., Wade, M. and Schneberger, S. (Eds), *Information Systems Theory: Explaining and Predicting Our Digital Society*, Springer, New York, NY, pp. 231-246.

[30] Rui, G. (2007). Information Systems Innovation Adoption among Organizations a Match-Based Framework and Empirical Studies, *National University of Singapore*, Singapore.

[31] Deephouse, D. (1996). Does isomorphism legitimate? *Academy of Management Journal*, Vol. 39 No. 4, pp. 1024-1039.

[32] Saltz, J. S., and Dewar, N. (2019). Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics and Information Technology*, 1-12.

[33]Irani, Z., Ezingeard, J., Grieve, R. and Race, P. (1999). Case study approach to carrying out information systems research: a critique, *International Journal of Computing Applications and Technology*, Vol. 12 No. 2, pp. 190-198.

[34] Cornford, T. and Smithson, S. (2006). *Project Research in Information Systems: A Student's Guide*, 2nd ed., Palgrave Macmillan, New York, NY.

[35] Eisenhardt, K. (1989). Building theories from case study research, *Academy of management review*, vol. 14, no. 4, pp. 532-550.

[36] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

[37] Saltz, J., Shamshurin, I., & Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology*, *68*(12).