Evaluating Data Science Project Agility by Exploring Process Frameworks Used by Data Science Teams

Sucheta Lahiri
Syracuse University
sulahiri@syr.edu

Jeffrey Saltz
Syracuse University
jsaltz@syr.edu

Abstract

The lack of effective team process is often noted as one of the key drivers of data science project inefficiencies and failures. To help address this challenge, this research reports on semi-structured interviews, across 16 organizations, which explored data science agile framework usage. While 62% of the organizations reported using an agile framework, none actually followed the Scrum Guide (or any other published framework), but rather, each organization had defined their own process that incorporated one or more aspects of Scrum. The other organizations used a proprietary / ad-hoc approach, often based on a proprietary data science life cycle. In short, while many data science teams are trying to be agile, they are adapting existing frameworks to work within a data science context. Future research could explore how data science teams can best achieve agility, perhaps via new agile frameworks that address the unique data science project management challenges.

Keywords: Agile, Data Science, Team Process

1. Introduction

Data science is often identified with the 5 "Vs", which describes the data in terms of volume, velocity, variety, veracity, and variability [1]. This view of data science might weave a narrative that data science project success is purely technical in nature. Indeed, most data science research is centered on the technical features and capabilities of predictive modeling.

However, a key challenge for successful data science project completion is not technical in nature, but rather, driven by the process used to execute data science projects [2]. For example, it has been noted that most data science projects do not leverage well-defined process methodologies [3][4]. Thus, it is not surprising that poor coordination and collaboration are notable challenges that have led to data science project failure [3][5]. It has also been noted that data scientists need more awareness with respect to the process frameworks that could be used to do data science work [6].

In short, to help improve the success of data science projects, more research is required to better understand how data science project management could be improved. With this in mind, it is helpful to explore how current data scientists execute and manage data science projects. Towards this end, via in-depth semi-structured interviews, this research explores the agility of data science teams, and more specifically, the process frameworks used by teams working on data science projects.

To help achieve these goals, this research explores the following three research questions:

- **RQ1**: How are data science projects managed?
- **RQ2**: What are the key process challenges in executing data science projects?
- **RQ3**: What are the best practices suggested by data scientists?

The aim of this research is to understand how data scientists work while they execute data science projects. As a result of this study, we highlight current team processes, key challenges that occur while working on data science projects, and areas of suggested process improvement.

The rest of this paper has the following structure: Section 2 provides some additional background context. Section 3 highlights the methodology of the research approach, data collection, and analysis. Section 4 presents the findings of the study. Section 5 provides a summary as well as limitations and potential future research.

2. Background

This section briefly provides additional context on data science process challenges and team process frameworks currently used by data science teams.

URI: https://hdl.handle.net/10125/103424
978-0-9981331-6-4

2.1. Data Science Project Execution Challenges

Studies confirm that most data science projects fail to match with the desired success criteria. For example, it was reported that 87% of data science projects fail to reach the production environment [7]. A different report, which analyzed 21 different case studies, noted that data science projects fail due to naïve processes and ad-hoc project methodologies that often do not go well with data science projects, and ultimately become the reason for the project to fail [5]. In another example of the need for improve data science team process, a survey of data scientists reported that 82% of the data scientists noted that they did not adhere to any specific process methodology, but 85% of the survey respondents thought that using a better process would lead to data science project success [6].

One reported challenge relating to data science project management is the length of the time required to completing the project is difficult to predict [8][9]. Other challenges that have been noted include communication challenges with stakeholders, which often leads to project goals that to not align with the business. Agility should help with these challenges.

2.2. Project Management Methodologies

The three most commonly used approaches for managing and coordinating data science projects are CRISP-DM, Scrum and Kanban [10]. Each of these are briefly described below.

2.2.1 CRISP-DM: The Cross-Industry Standard Process for Data Mining [11] was defined in the 1990s and has often been found to be the most commonly used framework and *de facto* standard for data science projects [12].

Specifically, CRISP-DM defines six phases of a project (business understanding, data understanding, data preparation, modelling, evaluation and deployment). Typically, when a project uses CRISP-DM, the project moves from one phase (such as data understanding) to the next phase (*e.g.*, data preparation). However, as the team deems appropriate, the team can go back to a previous phase. In a sense, one can think of CRISP-DM as a waterfall model for data mining [13]. However, the framework provides little guidance on how to know when to loop back to a previous phase, iterate on the current phase, or move to the next phase.

2.2.2 Scrum: Scrum is the most commonly used agile framework with over 12 million practitioners [14].

Scrum divides a larger project into a series of mini projects, called "sprints", each of a consistent and fixed length, typically one to two weeks long. Scrum teams have three roles: the product owner, the development team, and the scrum master.

Each sprint starts with a sprint planning meeting where the product owner explains the top items from the product backlog, which is an ordered list of product development ideas. The development team forecasts what items from the product backlog they can deliver by the end of the sprint and then makes a sprint plan to develop a product increment that includes the selected product backlog items. During a sprint, the team coordinates closely and holds a daily meeting to identify potential roadblocks.

At the end of each sprint, the team demonstrates the newly developed product increment to stakeholders and solicits feedback during their sprint review. To close a sprint, the team inspects itself and plans for how it can improve in the next sprint during the sprint retrospective. Throughout the process, the scrum master acts as a servant leader and coach to help everyone effectively implement Scrum [15].

2.2.3 Kanban: Kanban is a lean approach that focuses on maximizing value and minimizing waste in production processes. While agile practices such as Scrum have a well-defined process framework to structure work, Kanban has no such specified process framework nor any specifically defined roles. Rather, Kanban defines a set of principles which include: visualize the workflow, limit work-in-progress, measure and manage flow, make process policies explicit, and improve collaboratively/implement feedback loops [16]. Each team is free to use any process framework that supports/encourages these Kanban principles.

Two key strengths of Kanban are that (1) it visually represents work on a Kanban board with work items flowing across the columns of increasing work status completion, typically starting with a 'to do' column and ending with a 'done' column, and (2) it aims to minimize work-in-progress, often with WIP limits. Minimizing WIP enables a lean approach (by focusing on reducing the time it takes to complete a task) and also enables agility (since possible tasks are re-prioritized each time a new task starts).

3. Methodology

3.1. Data Collection

Data scientists were recruited from diverse backgrounds. Many factors of diversity were considered, including the nature of business, the

maturity of the person and the organization, and if the organization was private, government or public.

We applied snowball sampling to contact data scientists from public and private organizations. Snowball sampling is also referred to as 'chain referral sampling' [17] which is extensively used for qualitative research studies. In snowball sampling, data scientists were identified who later referred other potential participants they considered to be pertinent for this study. Snowball sampling method worked well for this study because the goal was to identify data science professionals who had a thorough knowledge of how data science projects were managed, across a variety of contexts. The sampling was not restricted to only those organizations that had a robust data science methodology for project execution. For example, recently launched the data science teams were also included in the study.

Professional networking sites, social media, and digital channels were used to identity pertinent participants for the study. The 16 data scientists who were identified, from public and private firms, all worked in the United States. The study was approved by the IRB, and all the interviews were conducted via Zoom calls.

The online interviews lasted from 30 to 120 minutes. Otter and Zoom transcription services were used to transcribe the interviews. The procedure of reviewing the transcripts were double layered and Zoom transcription and Otter transcription were both used to make sure that all the words were captured and transcribed correctly.

3.2. Study Participants

The data scientists had on five to 32 years of work experience. The data scientists worked for North American and European organizations. Two data scientists worked for government enterprises, 14 data scientists worked for private firms. Table 1 provides a summary of data scientists participants.

Table 1: Participant Summary

ID#	Experience (in years)	Industry
P1	25 yrs	Public–Defense
P2	32 yrs	Private–Conglomerate
P3	22 yrs	Private–Supply Chain (ex-marine)
P4	30 yrs	Private–Healthcare
P5	29 yrs	Private–IT and Services
P6	13 yrs	Private–Management Consulting
P7	7 yrs	Private–Finance
P8	25 yrs	Public–Veteran's affairs
P9	12 yrs	Private–Consulting
P10	16 yrs	Private–Finance
P11	30 yrs	Private–Asset Management
P12	20 yrs	Private–Media and Finance
P13	5 yrs	Private–Finance
P14	8 yrs	Private-Finance
P15	30 yrs	Private-Manufacture (automotive)
P16	19 yrs	Private-Entertainment

3.3. Data Analysis

The method of "general inductive analysis" [18,19], also evident in grounded theory [20], with coding was used as the methodology to determine themes for this study. Being inductive in nature, the process started from several rounds of thorough reading of the interview transcripts drafted for 16 data scientists without having any preemptive sets of themes in mind. The raw data was then sanitized to remove the superfluous words, which enabled the researchers to focus on the terms that pertain to participant's opinions and thoughts. The authors performed these exercises independently to address biases and transparency, and later conducted brainstorming meetings to discuss similar and dissimilar themes. Due diligence was done to make sure that the themes created with the data align with the research questions of this study.

During the inductive analysis, the authors created two separate spreadsheets, each of which had themes gleaned through full text transcripts. The analysis was done on a weekly basis by the authors, and the themes were discussed together in brainstorming meetings to determine similar and dissimilar themes. The similar themes were matched together to create bigger themes, and dissimilar themes were further discussed to come up with broader themes or context driven outlier themes. Prominent themes, based on the number of times shared as data by data scientists, were used to derive conclusions, and context driven idiosyncratic themes were used to highlight specific scenarios.

4. Findings

4.1. Current Process Practices

In the interview protocol, the data scientists were asked how their data science projects were managed and how the team coordinated their activities. *Table 2* provides a summary of the project management process for each of the organizations. Note that each organization had a unique process.

Table 2: Team Project Management Process

ID#	How Data Science Project is Managed
P1	A proprietary process with a designated 'mission owner' to determine the business value, be a sounding

	board to data scientists, allocate budget and resources. A phased approach was used, which included prototyping and transitioning to production environment
P2	An agile Kanban process that consists of swim lanes and columns, project prioritization, deciding project problem, development of project plan collaboratively made with budgetary and personnel chart, decentralized accountability, vetting the plan with development team to finalize hardware, software, team size and roles
P3	A proprietary life cycle focused framework focused on understanding the business problem/opportunity, evaluating the data, building a model, model, model testing and evaluation, as well as model lifecycle management
P4	Use Scrum concepts, including attempting sprints as well as a three-prong data life cycle, which includes (1) create a uniform data model, then (2) publish the data and (3) data governance
P5	Agile approach, including concepts such as a daily 15-minute call to discuss goals and roadblocks collaboratively and dedicated project process focus
P6	Aspects of Agile, such as product-oriented delivery and a process expert, combined with a 3-phase project life cycle
P7	CRISP-DM is extensively used, conceptual understanding of design and business architecture, requirement gathering, breaking down into different tasks, with parts of Scrum (mainly the role of a product owner).
P8	Proprietary process focuses on the: what and why questions, along with a streamlined PMO framework, bigger projects split into smaller chunks and then reported directly to management, consolidated report for smaller projects
P9	Proprietary Life cycle focused, which starts with requirement understanding, data collection, data engineering, model design, model training, model validation, model deployment, model performance. The process was thought to be agile, but no specific agile concepts were part of their defined process.
P10	No specific project management framework but stated that they used Agile loosely (no agile concepts were part of their defined process). Prioritization based on federal reserve requirements for MRA, MRIA and frequency of reviews needed, asking critical questions
P11	Agile, projects prioritized on dollar value, bigger projects divided into smaller projects to publish quick turnaround deliverables
P12	Agile, communication with stakeholders and identifying the risk ahead of time, stand-ups to identify roadblocks, documentation and recording of items, rotation of scrum masters having a dedicated scrum master for every sprint
P13	No project management framework, internal project management tool as a central system created by technology team for models to engage developers and stakeholders, project map to review each stage
P14	No specific project management framework, four criteria framework: (1) quality improvement (2) throughout improvement (3) cost reduction strategies (4) predictive maintenance
P15	No project management framework, projects selected though 'push' and 'pull' strategy, added to pipeline shortlisted basis two criteria: (1) checklist based on 100 previously done projects (2) four criteria informal framework, identifying purpose, participants, business leads, data scientists treated as project managers
P16	Agile, firm culture considered important, defining goals and the success criteria of the project, ingress and egress decided for determining tolerance, maturity of data, creating MVP, exploratory data analysis (EDA)

As can be noted from Table 1 and summarized in Table 2, 62% of the participants (10 of 16 organizations) discussed the use of an agile framework to manage data science projects. However, as noted by [19], a rigorous study of agile development requires the measurement of the use of individual agile practices, even in qualitative settings.

Hence, to help provide a better understanding of the agile practices in use by each of the organizations, Table 3 shows which of the organizations used which of the key agile practices (such as a daily meeting, the use of iterations, getting feedback on the iterations, and a defined process improvement approach).

As can be seen in Table 3, none of the teams used the full Scrum framework, but rather picked one or two agile concepts to integrate into their process.

Table 3: Use of Agile by data scientists

Checkpoints of agile	P2	P4	P5	P6	P7	P9	P10	P11	P12	P16
15-minute daily meeting		✔	✔							
Team delivers incremental value (iterations / sprints)		✔						✔		
Iterations are sprints with the same fixed time box interval		✔						✔		✔
A Product Owner provides feedback				✔	✔				✔	
There is on-going, team-wide process improvement effort	✔	✔								
A Kanban Board is used to track the work in the sprint	✔									

More generally, through an analysis of the process used across the sixteen organizations, four data science process approaches were identified with respect how teams managed their data science

projects. These approaches, shown in Table 4, ranged from an Ad-hoc/proprietary process to a process with well-defined iterations.

Table 4: Approach to Managing DS Projects

Approach	ID#
Ad-hoc / Proprietary Process	P8, P10, P13, P14, P15
Proprietary Life Cycle Focused Process	P1, P3, P7, P9
Proprietary Process with Some Agile Concepts	P2, P5, P6 P7
Proprietary Process with Well-Defined Iterations	P4, P11, P12, P16

In the rest of this section, these four project management approaches are described in more detail.

4.1.1 Ad Hoc/Proprietary Process: Teams that used an ad-hoc/proprietary process did not try to leverage a life cycle approach, nor use agile concepts. Some of the teams focused on trusting the data science team lead, while others had a process to help ensure consistent information sharing across the team.

For example, P13's team used a proprietary tool to implement their framework, which focused on keeping team members up to date. P13 notes the use of a traditional project plan, with defined dates/milestones that are tracked. In fact, their proprietary tool acted as a central database for their project and each phase had defined checkpoints that the team must follow:

…this system was built to keep all stakeholders [up to date], including the model developers, so that they know that … there is the current status of the review and so on. So, after every milestone in the Project Map, the review lead or any other team member can go and update the status saying this part is done and then we are in the next part…

P10's organization, on the other hand, did not follow a formal project management framework (but did think of themselves as agile). Rather, P10's organization started with broad questions, such as:

- *Who are we answering to?*
- *Who is going to finally sign off the outcome of the project?*
- *What are our objectives?*
- *What are our risks?*

4.1.2 Proprietary Life Cycle Focused Process: These organizations used a proprietary process, which leveraged a phased approach. In a phased approach, each phase is part of a data science life cycle. While some organizations leveraged a CRISP-DM like life cycle, other organizations started from scratch to define the phases that made the most sense for their organization.

With this in mind, these organizations might be considered waterfall-like in their process (since they were focused on each project going through each phase of their life cycle). However, the teams often tried to be somewhat agile, such as going back to a previous phase.

For example, P7's organization leveraged CRISP-DM for their life cycle. As previously noted, CRISP-DM is the most commonly used data science life cycle framework, and consists of the following phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

P9, on the other hand, used a proprietary life-cycle focused team process. Their life cycle started with (1) requirement understanding, then (2) data collection and (3) data engineering. Next was building the model (4), and then training the model (5). This is followed by model validation (6), and finally deployment of model in the production environment (7). After deployment, the performance is monitored and documented (8). They called their project management framework ML flow or machine learning flow. P9 stated that their ML flow was used in conjunction with Agile, but in reality, their life cycle steps were given precedence over agility, and the team typically followed a phased-based process, based in their ML flow life cycle.

4.1.3. Proprietary Process with Some Agile Concepts: Within the organizations that used a proprietary process with some agile concepts, most of those organizations used aspects of Scrum that they tailored to their needs. In short, none of the organizations in this category adhered to the Scrum Guide, nor did try to use the concept of an iteration.

For example, P2's team viewed themselves as agile, since the team worked collectively to define a project plan, as well as the fact that interactions across the team were viewed as more important than the process artifacts. Furthermore, P2 used a Kanban-like process that consisted of swim lanes and columns to track project progress, where work was pulled by each team member. However, they also defined an overall project plan, with dates and deliverables. Hence, one could view this team as a Kanban-like team with some additional structure.

In a different example, P4 talked about their customized Scrum in terms of the stakeholder communication and meetings. P4 highlighted having global meetings which was outside of their daily Scrum meeting.

…because of the geographic teams, global team, rather…so we have a global team that spread across the whole geography of the world…so that to ensure that those communications take place, what I wound up doing was I wound up having another meeting outside of the Scrum meeting…which the Scrum Master and the project manager chair, to make sure that all the blockers were resolved … So, we did that other layer of meeting…

In yet a different example, P5's organization also leveraged the concept of a daily meeting. However, the methodology was customized with a project management committee and other meetings:
… And there's a project management committee that looks at various projects and gets a report on, you know, what parts of it [project] are on time, what parts of it are ahead of schedule, and what parts are behind schedule…customization would be that sometimes they have two meetings a day…

Furthermore, P5's daily meeting was more of a status meeting, but still did focus on the typical Daily meeting, with questions such as:

- *What did you do yesterday?*
- *What would you do today?*
- *What are the roadblocks for the projects?*

P6 also viewed their approach as Agile, where they focused on what they referred to as POD (product-oriented delivery). Under POD, teammates worked together to bring agility and efficiency. The organization leveraged some agile concepts, including the notion of a Scrum Master (even though the team did not fully follow Scrum). However, much of their process was a more traditional phased project approach with separate teams doing business analysis, development, and testing:
…And these teams are usually in PODs. So, you have a project manager, you have an overall lead, for instance, or lead is kind of like a manager or senior manager level, who's leading a strategy. The project manager is obviously manager managing the project. And then you have a scrum master. So that's more really around output efficiency, and maybe making sure that obviously, milestones are happening according to plan that all the divided labor or work have been…umm…[pause]…moving in the right pace, and that there's no issues. And then you get your business analyst's level, or we call it generalists, the people who do the discovery and analysis piece of work that understand the gap assessment. And then you have obviously the testing folks or technology who really help you in that case…

4.1.4 Proprietary Process with Well-Defined Iterations: Teams in this category identified the benefit of small iterations. Note that these iterations were often not Scrum sprints. For example, for some teams, their iterations were planned in advance, and for other teams, iterations did not have a fixed length time-boxed duration.

For example, for P11's team, a big project was divided into smaller increments that could provide quicker deliverables. This was driven by the fact, that the team was concerned with the fact that big projects require significant time and money to execute, which was difficult to deliver. Hence, the bigger projects were divided into smaller, more attainable, efforts.
… So, the goal is to see less of these big elaborate projects starting and running over budget and leading nowhere to, to agile way of working and project management, where you're setting smaller goals, [and] quicker deliverables …

For P4, the framework was a proprietary modification of Scrum, with pre-defined sprints. For example, P4 described a typical project where they created the project plan, with ten defined sprints. P4's process also could be considered a phased approach, especially with respect to the data management, where P4's organization adopted three phased process. First, the team creates a uniform data model, then they publish the data and finally they ensure governance of the data.

4.2. Project Management Challenges

During the interviews with the 16 data scientists, there were questions focused on the challenges faced while executing data science projects.

Three key challenges were identified with respect to managing the process. Table 5 highlights these three major themes that emerged during the interviews.

Table 5: Data Science Project Challenges

Theme	Project Challenges	ID#
Culture	Impact of organizational culture on the importance of the project and how stakeholders view data science projects	P1, P6, P15, P16
Project Uncertainty	It can be difficult to know how long a task will take, or even if the project will deliver results.	P1, P3, P4, P5, P9, P12, P14
Ethics Oversite	The process doesn't explore potential ethical situations (ex. Data privacy, model fairness)	P2, P3, P11

Beyond these themes, there were other challenges noted. For example, P13 noted that overfitting of the model was one of their main challenges.
…overfitting is like a very, very common risk of machine learning problems where the model does not really generalize well, it gives you like a very good performance on training data, but on any other out of sample data, it does not really do well…

There was no solution suggested from the framework point of view that could mitigate the issue of overfitting. Though technical in nature, a process life cycle could be suggested to mitigate this type of technical challenge for data science projects.

The rest of this section describes these three themes in more detail.

4.2.1 Challenge with Culture in the organization: A team's culture can impact the team's ability to improve their process and how the data science team works with others (such as stakeholders and software development teams).

For example, P16 highlighted the issue of culture across the institution, and how it impacts the team's data science process and results:
…every enterprise has very different culture. And some culture change, like I have been in companies, which have been in legacy where data or data science or data driven, have been an afterthought. And now in this world as they compete with companies or enterprises that have been born in the data driven world. So, it's, it's…it's like rescaling of the whole enterprise…

P1 also noted this cultural challenge, and how it can impact the stakeholder's view of data science project requirements:
…Then the other thing is just understanding what data science and automated solutions can and can't do for you. It's been a cultural challenge for our folks…

4.2.2 Project Uncertainty: Many teams noted the challenge of working on projects with high uncertainty.

For example, P4 noted the uncertainty in the data science project:
I have very rarely seen people start big data and data analytics project, knowing ahead of time, what kind of analytics they are going to perform of this data? Most of them start with let's get the data together. And we'll figure out what we can do …

Going from prototype to production was an uncertain process for P1, who noted that it required a whole list of transition related questions:
…how am I transitioning that [prototype] into production? Who gets to watch and manage and…and see that thing live beyond that prototyping phase? So, we will think about that fairly early on in a project's development. And we will then help the client identify what that pathway looks like …

P3, who has experience with both private and public organizations, highlighted several challenges in executing data science projects, but the highest ranked challenged focused on the ability to access data, which can impact timelines and model accuracy:
…I think if I were to list out the three biggest challenges, they would be number one, access to the data. So oftentimes, either it's difficult, or data can't be shared or won't be shared, or when they are shared. They're so messy. So, number one is the data…

4.2.3 Ethics Oversight:
Another consistent challenge noted during the interviews was the lack of oversight on data privacy, and more generally, ensuring an ethical data science model was developed.

For example, P6 highlighted the pressing concern of data privacy and GDPR (as well as PII information):
…data, privacy has become a huge deal a little bit, as you know, in the last three, four years now, with GDPR, and some other nuances of it, all versions will be in other regions. So, the risk of how we keep data, how are we using other people's data, how long we're going to be using it for and who else we're going to be giving it to, are becoming very important these days…

A similar data concern was focused on trust in external data sources, which was brought up by P1. In other words, the organization was not very cognizant about the data sources that were procured:
…the other big challenge we have is data, most of what, you know, mature AI organization focuses in on this data, we have, we do a good job in this regard for classified data sources, because we care a lot about what we collect, how we collect it, and how we protect it…So we are, surprisingly not as diligent about data sources that we buy…

In a different example, P2 noted that their process methodology did not include how to ensure ethics oversight as well as other external the associated potential risks:
… now mind you, again…Kanban, agile and so on, … they're not at all useful for managing the external risk of your communication, change, management ethics, and so on. And so, you kind of have to pick the risk you're looking at…

4.3. Process Best Practices

Table 6 shows the four key best practices that were identified via an analysis of the interviews:

Table 6: Process Best Practices by Data Scientists

Best Practice	Explanation	ID#
Ensure Effective client communication	Work to make sure the problem is understood and that clients have their needs met	P1, P2, P3, P7, P11
Identify and use stakeholders effectively	Know who are the stakeholders and get feedback from the stakeholders	P1, P9, P11
Refine IT Project management approaches to Data Science	Do not assume that a process framework that works for software development can work for data science project	P2, P3, P12
Define project success criteria	Proactively define how the team will know if the project was a success	P4, P5, P13

The rest of this section describes each of these best practices in more detail.

4.3.1. Ensure Effective Client Communication: Effective communication with stakeholders was deemed as a best practice by many. For example, P2 talked about the importance of working and communicating and 'connecting" with their stakeholders:

I think it's a risk that, after everything you and I have talked about…the data, the project management, the analytic methods, the technologies, and everything like that…all of that doesn't matter. It doesn't matter at all. If it ultimately does not connect with the end user, or the person that's going to be using it … then they're not ever going to really crack it open.

Furthermore, P1 highlighted explainability as something that was very important in their current project management process, and a significant issue if this was note done correctly. P1 suggested comprehendible language to explain the models to the social scientists that can mitigate the risk of loss of information currently seen in the existing project management process:

…And so, as I was reading some of the language written into what we're trying to communicate to folks [management] about what they're supposed to do with their products…we are using a lot of techno-babble to explain to again…social scientists, what it is that we're asking them to do. And it's a losing proposition there. Yes…You sound super smart. Congratulations! But nobody can use the information because they don't understand what it means.

Furthermore, P7 highlighted the need to ensure that the project goals were well understood:

I think what really matters is requirement gathering and the understanding of the book of work, right? I think 80% of the project success is if you can…if you can understand the art correctly.

4.3.2. Identify and use stakeholders effectively: While effective communication with stakeholders is related to the first theme previously discussed, this theme had more of a focus on ensuring the project correctly identified stakeholders, and then got feedback from the stakeholders.

For example, P1's process defined 'mission owners' who brought the idea of projects to the table. The first and the foremost objective of the mission owner within P1 organization was to focus on understanding the problem to solve and identifying stakeholders to resource optimization:

…what kind of staff members do you have available to work on the project? Do you have 'mission owners' that actually are responsible for the outcomes of that particular mission?...

For P9, stakeholders included a range of roles, including: (1) client partners (2) domain experts (3) and technology experts.

For P3, the key stakeholder was the data scientist who could understand the technical aspects of the project. A blend of EQ, AQ and IQ was mentioned by P3 as a skill to be pursued by the data scientist to maintain a 'smooth' process rather than 'fast' and contradictory in the realm of competitive data science domain. The smooth approach consisted of following up with the teammates about the project expectations and goals to be met.

4.3.3. Refine IT Project management approaches to Data Science:

This best practice acknowledges that there is not a well-known appropriate process for data science projects, and as such, teams need to adapt existing project management approaches (from other domains) for data science projects. However, the data science team should plan on adapting that existing approach.

For example, while P3 was not a proponent of any one specific framework, P3 noted that that having a fundamental idea of traditional project management concepts as well as agile concepts would provide data scientists a better understanding with respect how to best manage data science projects:

… I'm not a fan of anybody who was too or overly orthodox and following either one [Scrum or Project

Management Framework]. I think traditional project management frameworks are very waterfall, very linear…. And anyone who tries to do a project like this, and a typical waterfall PMP method is going to fail. I think the folks who can adapt and can learn from the PMP principles and, you know, lightly seasoned them, or federate them with what they understand around agile and Scrum, that they'll do far better…

Similarly, P2 also noted that there is a difference between IT project management and data science project management:
…there are, you know, project management, best practices. … And, of course, the big, big difference between project management for IT and project management for data science, is that it [IT] is predictable…

Furthermore, P8 observed that there was a lack of knowledge and skills to determine how to best combine the frameworks in an optimal way:
…I think a lot of our program managers on the federal side, don't have the knowledge to really manage the program, and they rely on the contractor to tell them what they need to do. And they kind of rubber stamp it…

4.3.4. Define project success criteria: Finally, several organizations noted the need to ensure everyone understood what it meant for the project to be a success.

For example, P4 emphasized the need for success criteria that can drive a better understanding of the project before that project starts:
…First thing you have to do for a given project is… you have to establish, what is my success criteria? What am I trying to get out of this? When do I say this project has succeeded?

P13 also emphasized the importance of success criteria, and that the criteria is well documented:
…there is some sort of documentation associated with it, which kind of details the model scope, the criteria for success, the model performance, any benchmark model has been varied, and how the team is planning to properly monitor the models…

However, P4 noted that most teams do not define success criteria:
…And I have to tell you, and matter of fact, I encourage you to ask this question to many other participants. How do you establish your success criteria for a given project? And you'll see there lies the problem, majority of them don't even do that…

5. Conclusion

5.1. Summary

In this qualitative study, we interviewed 16 data science practitioners to explore data science project management processes across 16 different organizations.

This study found that while many teams want to be agile, or think of themselves as agile, most data science teams use a proprietary framework that was not agile in nature. In fact, in this study, none of these custom frameworks incorporated the key agile concept of doing short iterations, getting feedback on those iterations, and then prioritizing future work based on that feedback.

More specifically, to address the first research question (*How are data science projects managed?*), an analysis of these semi-structured interviews identified four data science project management approaches, ranging from an ad-hoc process to an adapted agile process with defined iterations. However, while many of the teams leveraged key concepts from common frameworks such as Scrum and CRISP-DM, none of the teams fully followed any of these frameworks.

To address the second research question (*What are the key process challenges in executing data science projects?*), three common challenges were identified across the organizations. The first challenge was the need for an appropriate organization culture, such as when and how to engage stakeholders. It also includes the fact that there are many uncertainties when executing a data science project, which the teams project management framework needs to be able to address. Finally, ensuring ethics is something that was noted to be missing across almost all of the organizations. Note that these challenges could be eliminated, or at least mitigated with an appropriate data science team process.

Finally, four best practices were identified, with addresses the third research question (*What are the best practices suggested by data scientists?).* These best practices included (1) ensuring effective client communication, (2) identifying and using stakeholders effectively, (3) adapting an IT PM approach for data science and (4) defining project success criteria.

5.2. Limitations & Potential Next Steps

One of the biggest limitations of this study was that the all the interviewees were located in North America. A global perspective of data scientists and project management in a multi-cultural and distributed organization would be worthwhile.

Another limitation was the sample size. While 16 interviews provided a broad cross section of organizations, interviews or surveys with additional organizations would be helpful to ensure these findings are generalizable.

5.3. Potential Next Steps

Beyond research to create a larger sample size of organizations, future research should explore the reason teams only used a subset of the key agile concepts. This could have been driven by the lack of knowledge on how to effectively use these frameworks, or by the frameworks not meeting the needs of the organizations.

Future research could also explore if centralized or decentralized structures are preferred while managing data science projects. Additional research might also explore how to integrate ethical guidelines into the data science process. This is related to exploring how data science project risks could be addressed via improved project management methodologies [20].

Finally, based on the fact that none of the teams used a well-known project management framework, it is likely that new frameworks, which address the specific needs of data science projects, need to be defined and evaluated. These frameworks could support the unique needs of data science projects.

6. References

[1] Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231-1247.

[2] Saltz, J. S., & Hotz, N. (2020, December). Identifying the most Common Frameworks Data Science Teams Use to Structure and Coordinate their Projects. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 2038-2042). IEEE.

[3] Martinez, I., Viles, E., & Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research*, *24*, 100183.

[4] Saltz, J., Hotz, N., & Sutherland, A. (2022). Achieving Lean Data Science Agility Via Data Driven Scrum. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*.

[5] Saltz, J. S., & Krasteva, I. (2022). Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science*, *8*, e862.

[6] Saltz, J., Hotz, N., Wild, D., & Stirling, K. (2018). Exploring project management methodologies used within data science teams. In *24th Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*. Association for Information Systems.

[7] VentureBeats. (2019). Why do 87% of data science projects never make it into production?, https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

[8] Qadadeh, W., & Abdallah, S. (2020, March). An Improved Agile Framework For Implementing Data Science Initiatives in the Government. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)* (pp. 24-30). IEEE.

[9] Shahapurkar, S. (2016). *Crossing the chasm: Deploying machine learning analytics in dynamic real-world scenarios*. Arizona State University.

[10] Saltz J. (2022). CRISP-DM is still the most popular framework for executing data science projects, www.datascience-pm.com/crisp-dm-still-most-popular/

[11] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz TP, Shearer C, Wirth R. 2000. CRISP-DM 1.0: Step-by-step data mining guide. Chicago: SPSS, Inc, https://www.the-modeling-agency.com/crisp-dm.pdf

[12] Martínez-Plumed, F. et al., (2021). "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048-3061, 1 Aug. 2021, doi: 10.1109/TKDE.2019.2962680.

[13] Gao, J., Koronios, A., & Selle, S. (2015). Towards a process view on critical success factors in big data analytics projects. *Twenty-first Americas Conference on Information Systems (AMCIS)*, Puerto Rico.

[14] West, D., 2017. Scrum Guide Update, Scrum.org, www.scrum.org/resources/blog/scrum-guide-update-november-2017

[15] Sutherland, J., & Schwaber, K. (2017, November). The Scrum Guide. Retrieved from scrumguides.org: www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-GuideUS.pdf

[16] Anderson, D., "Kanban: Successful Evolutionary Change for Your Technology Business". Sequim, WA: Blue Hole Press, 2010.

[17] Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research*, *10*(2), 141-163.

[18] Thomas, D. R. (2003). A general inductive approach for qualitative data analysis.

[19] Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, *27*(2), 237-246.

[20] Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Sage publications.

[21] Tripp, J., Saltz, J., & Turk, D. (2018, January). Thoughts on current and future research on agile and lean: Ensuring relevance and rigor. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[22] Lahiri, S., & Saltz, J. (2022, January). The Risk Management Process for Data Science: Gaps in Current Practices. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.