Exploring Project Management Methodologies Used Within Data Science Teams

Emergent Research Forum (ERF)

Jeffrey Saltz  
Syracuse University  
jsaltz@syr.edu

Nicholas Hotz  
Indiana University  
nhotz@umail.iu.edu

David Wild  
Indiana University  
djwild@indiana.edu

Kyle Stirling  
Indiana University  
kylestir@indiana.edu

Abstract

There are many reasons data science teams should use a well-defined process to manage and coordinate their efforts, such as improved collaboration, efficiency and stakeholder communication. This paper explores the current methodology data science teams use to manage and coordinate their efforts. Unfortunately, based on our survey results, most data science teams currently use an ad hoc project management approach. In fact, 82% of the data scientists surveyed did not follow an explicit process. However, it is encouraging to note that 85% of the respondents thought that adopting an improved process methodology would improve the teams’ outcomes. Based on these results, we described six possible process methodologies teams could use. To conclude, we outlined plans to describe best practices for data science team processes and to develop a process evaluation framework.

Keywords

Big Data, Data Science, Project Management, Process Methodology

Introduction

Data science is the analysis of data to solve problems and develop insights. It is viewed as a separate and distinct discipline emerging from computer science, statistics and information management. While there are many views on what to include in the field of data science, we adopt Saltz & Stanton’s (2017) definition, which includes the collection, preparation, analysis, visualization, management, and preservation of large collections of information. This definition embraces the notion that data science is more than just analytics. Moreover, we view big data, where the data sets are too large and/or complex for traditional data analysis techniques, as a subset of data science.

Unfortunately, companies struggle to realize value from investment in data science talent and projects. MIT reports that “While businesses are hiring more data scientists than ever, many companies are struggling to realize the full organizational and financial benefits from investing in data analytics” (Stein, 2015). Furthermore, Gartner notes, “that through 2017, 60% of big data projects will fail to go beyond piloting and experimentation, and will be abandoned” (Gartner, 2015). One key reason for this challenge in executing data science projects could be the lack of a defined process methodology to manage projects (Saltz, 2015). For example, Domino Data Lab blames “gaps in process and organizational structure, and inadequate technology” as the primary culprits (Domino Data Lab, 2017). In addition, John Akred, Co-founder and CTO of Silicon Valley Data Science, noted that “We’ve met a lot of data science teams that understand how to do the data science, but they don’t have any real method of managing the data science project” (Akred, 2016). Hence, our empirical research study explores the following research questions:

RQ1: What process do teams currently use?
**Data Science Team Process**

**RQ2:** Do teams think that using an improved / more consistent process would improve performance and if so, what methodologies might a team use to manage data science projects?

The rest of the paper first provides an initial overview of these findings.

**Background Context**

Data science is rapidly maturing beyond siloed data scientists into a team sport involving professionals with diverse skill sets spanning multiple domains (Spoelstra, Zhang, & Kumar, 2016), and as noted by Das et al (2015), “The exploratory and often ad-hoc nature of analytic demands and a distinct lack of established processes and methodologies make it difficult for Big Data teams to set expectations or even create valid project plans.” Despite the project management challenges of data science and the impact these have on project outcomes, there is little information available about how organizations are managing data science projects. For example, the first data science project management controlled experiment was not published until 2017 (Saltz, Shamshurin & Crowston, 2017), and in a broad literature review, minimal research was found on data science project management (Saltz & Shamshurin, 2016). The small amount of available research reveals that data science teams and organizations generally suffer from a low level of process maturity (Saltz, 2017) or rely on software engineering processes.

Teams with process immaturity are dependent on their senior data analytic leaders’ experience. This lack of robust team-based methodology is similar to how software was developed in the late 1960’s. The impact of process maturity and no systematic process is documented in other domains, such as software development. Specifically within data science, ad hoc processes can lead to numerous problems (Spoelstra, Zhang, & Kumar, 2016), such as thwarting team efficiency improvement (one can measure specific inefficiencies and improve that process in the future), slow information sharing (poor processes for storing, retrieving and sharing information wastes time in people look for information and are at risk for using the wrong version of data or metadata), delivering the “wrong thing” (lack of effective processes to engage with stakeholders increases the risk that teams will deliver something that does not satisfy stakeholder needs), lack of reproducibility (further building on past projects might be difficult given inconsistent preservation of relevant artifacts like data, packages, documentation, and intermediate results) and poor coordination (poor processes decrease coordination and can result in confusion, inefficiencies, and errors) and scope creep (without proper processes to determine what should be included and excluded in a project). These issues negatively impact project results. For example, organizations that relied on ad hoc processes (as opposed to planned processes) on big data initiatives were only half as likely to rate their projects as successful (Capgemini, 2014).

Teams that use a well-defined systematic process tend to leverage methodologies predominately designed for software development such as the software development lifecycle (SDLC), Scrum, or Kanban. While these can overcome shortcomings of ad hoc project management, more research is needed to understand whether software engineering project management processes are appropriate for data science. A nascent argument is that they might not effectively handle nuances of data science including its exploratory nature, the increased ambiguity of requirements, statistical challenges in validating results, and the increased focus on data gathering, cleansing, and modeling (Saltz, 2015) (Godsey, 2017).

**Methods**

To better ensure the reliability of the results and to examine perceptions from multiple perspectives, 78 professionals were surveyed with different roles across multiple organizations – both from industry and not-for-profit organizations. These surveys, which followed the structure described by Myers and Newman (2007), focused on the current methodology data scientists used and whether an improved project management process would benefit their results. The participants were identified via direct outreach at the IEEE Big Data 2017 conference and by reaching out to other known data science professionals. Since the survey participants are experts in their field, and because domain experts typically offer significant insights of the desired domain, the total number of interviewees can be low (Bogner, Littig & Menz, 2009).
Findings

In total, 54% of the 78 participants identified as data scientists, 22% as data science team leads and 24% as data engineers. In terms of team size, 13% worked mainly by themselves, 40% worked in a team of 2–3 people, 37% worked in a team of 4–6 people and 10% worked in a team with more than 6 people.

As shown in Table 1, 82% of the respondents did not follow an explicit process, in that they either were not sure what process they used or just ‘figured it out’ as they went along. Perhaps more interesting, 85% of the respondents answered “yes” to “Do you think you would have more effective projects (ex. less work, improved insights) if you used an improved / more consistent process to do a big data project”. These results were consistent across the type of methodology that the team was currently using (or not using).

<table>
<thead>
<tr>
<th>Usage of each methodology</th>
<th>Would an Improved Process Help?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Based on an Agile methodology</td>
<td>12 (15%)</td>
</tr>
<tr>
<td>Based on CRISP-DM</td>
<td>2 (3%)</td>
</tr>
<tr>
<td>Figure it out as we go</td>
<td>46 (59%)</td>
</tr>
<tr>
<td>Not sure</td>
<td>18 (23%)</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>78 (100%)</td>
</tr>
</tbody>
</table>

Table 1: Would an improved process help your results?

Established Approaches to Managing Data Science Projects

Based on our survey responses and research, we identified several possible approaches teams could use to manage projects. In this section, we provide some context for each of the approaches.

Scrum has become the most commonly used agile approach with over 12 million practitioners (scrum.org, 2017). Scrum has three roles (product owner, scrum master and the development team) and divides a larger project into a series of mini-projects, each of a consistent and fixed length ranging from one week to one month. Each mini-project cycle, called a sprint, starts with a sprint-planning meeting where the product owner defines and explains the top feature priorities. The team forecasts what increments they can deliver by the end of the sprint and then create an execution plan. During a sprint, the team coordinates through daily ‘standup’ meetings. At the end of each sprint, the team demonstrates the results to stakeholders and solicits feedback during sprint review. These increments should be potentially releasable and meet the pre-defined definition of done. To close a sprint, the team inspects itself and plans for how it can improve in the next sprint during the sprint retrospective (Sutherland & Schwaber, 2017). Scrum has started to be used within a data science context, often with promising results (Chen et al, 2016; Grady, Payne & Parker, 2017).

Kanban started as an automotive supply chain and inventory control system in the 1950s in Japan and has since been adopted by other industries including software engineering (Brechner, 2015). Kanban starts with a list of potential tasks that are placed in the initial ‘To Do’ column on a Kanban board. In a simple three column (or three bin) Kanban board, when the team decides to start working on the task, the Kanban card (ex. a sticky note) is moved from the ‘To Do’ to the ‘Doing’ column. When the task is complete, it is moved to the ‘Done’ column. Teams also might split Doing into ‘In Development’ and ‘Testing’ breakouts (Brechner, 2015). Uncompleted work, known as work-in-progress (WIP), is an investment (in team effort) whose value has yet to be realized. To reduce the investment in WIP, Kanban teams set WIP Limits that define the maximum number of tasks that can simultaneously exist in given columns (Brechner, 2015). Kanban is less prescriptive than Scrum as roles, meetings, and time boxes are not defined. Some initial results on using Kanban within a data science context have been promising and better than a scrum-based approach (Saltz, Shamshurin & Crowston, 2017).

CRISP-DM (Cross Industry Standard Process for Data Mining), defined in the 1990s, describes six major iterative phases (business understanding, data understanding, data preparation, modeling, evaluation and deployment), with some high-level iteration allowed between the steps (Chapman, et al., 2000). Typically, when using this framework, the team progresses through the different phases. As needed, the team can...
“loop back” to a previous phase (ex. more data preparation) and can define useful milestones. In a sense, one can think of CRISP-DM as a waterfall model for data analysis. CRISP-DM has been consistently the most commonly used methodology for Knowledge Discovery in Database (KDD) projects (Haffar, 2015), - the general process of discovering knowledge in data through data mining. However, there has been a reported decrease within the community of people using CRISP-DM, and an increase in people using their own methodology (Piatetsky, 2014).

**Possible Emerging Approaches to Managing Data Science Projects**

Additionally, some organizations are formalizing and publishing specific project management approaches that typically combine elements of agile and CRISP-DM. While these are not widespread in use, they hint towards a promising future of commonly accepted methodologies, similar to what is available for software development efforts but that are specific to data science. In this section, we describe three of these emerging methodologies.

**Hybrid Methodologies** combine multiple approaches. For example, one hybrid approach uses aspects of Scrum with aspects of waterfall/CRISP-DM. In fact, there have been many that have advocated for this type of approach in other contexts, such as Nelson & Stolterman (2014), who believe that crystalline processes such as waterfall and liquid processes such as Scrum should simultaneously co-exist. Using this approach, one might, for example, use a CRISP-DM or waterfall-like approach for large infrastructure aspects of a project (ex. developing a large data warehouse). This could also enable an initial focus on understanding the business and data context at the start of the project. Then, the team could use an agile Scrum or Kanban approach to iteratively leverage that large infrastructure. Schmidt and Sun (2018) discuss one successful such example of this type of hybrid approach.

**Team Data Science Process** (TDSP), launched in 2016 by Microsoft, is “an agile, iterative, data science process for executing and delivering advanced analytics solutions.” Its lifecycle is similar to CRISP-DM and its process coordination takes several elements from Scrum including the backlog artifact, sprints, and clearly-defined team roles (Microsoft, 2017). TDSP’s project lifecycle is like CRISP-DM and includes five iterative stages: Business Understanding, Data Acquisition and Understanding, Modeling, Deployment and Customer acceptance (Microsoft, 2017). TDSP addresses the weakness of CRISP-DM’s lack of team definition by defining four distinct roles and their responsibilities during each phase of the project lifecycle: Group Manager, Team lead, project lead and individual team contributor.

**Knowledge Discovery in Data Science** (KDDS) expands upon CRISP-DM to address big data problems by defining additional integration with management processes. Specifically, KDDS defines four distinct phases: assess, architect, build, and improve and five process stages: plan, collect, curate, analyze, and act (Grady, 2016) and Grady has reported positive results in its use (Grady, Payne & Parker, 2017).

**Conclusion**

We note that 80% of our survey respondents work in a team and 82% of our respondents did not use any well-defined process methodology. In addition, 85% of the respondents thought that their data science efforts would improve if they used a systematic process methodology. Moreover, minimal research is published on the benefits and strengths of applying traditional or emerging project management methodologies to data science. Consequently, further research is needed to support the emerging subfield of data science project management. Perhaps part of the reason that many teams use an ad hoc approach is that there is no clear agile process that is designed for data science projects. Hence, many teams used an ad hoc approach either because they did not consider using a process methodology or were unaware of suitable approaches. What is clear is that data science teams are starting to consider improving the process they use to execute data science projects. Hence, our research identified and described six different process methodologies that could be used within a data science context.

We plan to continue to explore data science project management, via additional semi-structured interviews, which will help us better understand how teams select and execute their current methodology. Equally important, we also plan to develop an evaluation model that would score each methodology based on how effectively it handles the key challenges of data science such as business uncertainty, data uncertainty, result uncertainty, stakeholder management, team management, data collection, and
infrastructure challenges. Finally, we are also working to identify best practices that would be of value to data science teams. Hopefully, this will enable teams to more effectively convert data science investments into actionable insight via an approach that fits their specific challenges.

REFERENCES


Saltz, J. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness, IEEE International Conference on Big Data (Big Data).


