Modernising Data Science Project Management:

A Comparative Analysis of CRISP-DM and Data Driven Scrum

Colin Bowers

Department of Continuing Education, McMaster University

DAT 202 Data Management

Instructor: Zaki Eltwaishi

July 14, 2024

Introduction

Data science projects encompass a wide range of domains including data analysis, machine learning (ML) models, deep learning, and artificial intelligence (AI) applications. These projects are inherently more ambiguous and uncertain than traditional software projects due to their exploratory nature and the iterative process of model refinement. There are a plethora of reports indicating that most of these projects fail. For instance, according to Venture Beat, "87% of data science projects will never make it into production." [1] Indeed, the need to challenge traditional project management methodologies for data science is clear.

This paper reviews two prominent frameworks for managing these projects: CRISP-DM (Cross Industry Standard Process for Data Mining) and Data Driven Scrum (DDS). It will highlight the shortcomings of CRISP-DM, provide an overview of traditional Scrum and its limitations, introduce DDS as a modern alternative, and explore how DDS can be integrated with CRISP-DM.

CRISP-DM

CRISP-DM, developed in the 1990s, is a widely used methodology for data mining projects, consisting of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. According to a 2020 survey by Data Science Process Alliance, CRISP-DM remains the most commonly used framework for managing data science work (49% of all responses). [2]

Despite its widespread use, CRISP-DM has several limitations. First, it lacks mechanisms for team coordination and communication, which are crucial for the collaborative nature of data science projects. Second, the framework does not emphasize continuous stakeholder engagement, leading to potential misalignment with business goals. Third, CRISP-DM's waterfall-like structure can result in delays to value delivery. Lastly, the framework has not been updated since its inception, making it less suitable for the high volume, real-time, heterogeneous nature of modern Big Data.

Scrum

Scrum is an agile framework that organizes work into fixed-time iterations known as sprints. Each sprint typically lasts between one to four weeks, during which a team attempts to deploy new or enhanced functionality known as a Product Increment. Scrum emphasizes regular inspection and adaptation cycles to validate progress and plan subsequent iterations. This iterative process ensures continuous improvement and alignment with the needs of the customer which address a few of the shortcomings of CRISP-DM.

While Scrum excels in software development, it faces challenges in data science projects. For instance, fixed-length sprints can lead to challenges by forcing teams to fit unrelated tasks into a sprint, potentially delaying feedback and compromising quality. Furthermore, some data science tasks, like data collection and model evaluation, may require more time than a sprint allows. Additionally, Scrum's emphasis on delivering new features for end users does not align well with the exploratory nature of data science, where business value is often with gaining new knowledge or enriching existing data.

Data Driven Scrum

Data Driven Scrum (DDS) is a modern framework designed to address the unique challenges of data science projects by integrating agile principles with data science workflows. It retains much of the core concepts of Scrum, such as the roles, artifacts and events, but with adaptations to fit the data science context. For instance, Scrum's Product Owner role, the person empowered to drive the direction of the project, is present in DDS. In the context of DDS, they decide which components of the project should be built, the order in which to build them, and which aspects of them to observe and analyze.

There are some key differences between DDS and Scrum described in the remainder of this section.

Hypothesis-driven development. Unlike traditional Scrum, which focuses on delivering features to end users, DDS adds a concept around the creation and testing of hypotheses. Backlog items in DDS can be framed as a hypothesis or experiment, guiding the development and analysis process. This approach ensures that every cycle, or iteration, is aimed at generating new insights or validating assumptions, which is more applicable in the context of data science.

Task breakdown. Backlog items are decomposed into at least three essential tasks: create, observe, and analyze. The creation task involves developing a new data asset such as an enriched dataset, a predictive model or an analytical report. The observation task is about inspecting and gathering data on the performance, correctness or business value that the new asset offers. Finally, the analysis task focuses on interpreting the results to inform subsequent iterations. This structured approach ensures that each iteration not only produces tangible outcomes but also generates valuable knowledge that drives the project forward.

Capability-based Iterations. Traditional Scrum relies on fixed-length sprints, typically lasting one to four weeks. In contrast, DDS employs capability-based iterations, where the length of an iteration is determined by the completion of the empirical process (i.e the create-observe-analyze cycle described above) rather than a set time period.

Overlapping iterations. In Scrum, teams work on one sprint at a time. However, DDS allows for multiple iterations to occur simultaneously. For example, while a team is waiting for the results of observing a recently deployed predictive model, they might begin preparing data or setting up a new experiment. This overlapping approach is better suited for the exploratory and hypothesis-driven nature of data projects.

Integrating DDS with CRISP-DM

This section describes the concepts of "horizontal" and "vertical" slicing project phases as an approach to integrating DDS with a lifecycle framework like CRISP-DM.

Horizontal Slicing. This approach involves performing all tasks within a particular phase of the CRISP-DM framework before moving on to the next phase. For example, imagine a project that aims to reduce churn with three deliverables: an analysis on customer segmentation, a model to predict if a given customer will churn, and a product recommendation engine. A team working horizontally would complete the entire Data Preparation phase for each of the three deliverables before moving on to the Modelling phase and begin developing their first working prototype.

Vertical Slicing. On the other hand, vertically slicing a project would focus on a small piece of business value, progressing through all phases, for a single deliverable as quickly as possible. The resulting deployment should be the simplest version of a testable product as possible, also referred to as a Minimal Viable Product (MVP). For example, the team may choose to create a predictive model using basic cleaning techniques of available data and a simple algorithm such as Random Forest. The next iteration may attempt to improve this model with feature engineering or experimenting with a different algorithm.

Comparison. Horizontal slicing will maintain a more complete project plan throughout the life of the project by offering a more comprehensive view overall. However, it presents several challenges. First, the only meaningful value that is delivered to a customer or to the business is at the very end of the project. Second, this approach often results in a "big bang" deployment of all deliverables simultaneously which can be overwhelming to manage unforeseeable technical issues in production environments. Finally, model evaluation and feedback from business users is delayed which severely limits the team's ability to validate the progress being made.

Hybrid Approach. A hybrid approach could involve an initial phase of preparing and understanding the complete picture of the business needs for the overall project (horizontal), while then proceeding to an iterative development of a machine learning model (vertical). This hybrid approach will offer the benefits of providing business value quickly while maintaining alignment between the project's purpose and the goals of the organization.

Conclusion

Modern data science projects are becoming increasingly complex, with a majority, according to Gartner, "remaining alchemy, run by wizards." [3] The responsibility of managing these projects must extend beyond development teams and become a priority for managers and leaders. Innovation is essential to develop new methodologies that address the unique challenges inherent in data science. Data Driven Scrum (DDS) offers a viable alternative to traditional practices, integrating agile principles with a focus on hypothesis-driven development and continuous learning. As data science continues to evolve, combining frameworks like CRISP-DM and DDS will be critical to ensure projects are not only completed on time but also deliver meaningful insights. It is vital for organizational leaders to adopt these modern methodologies to stay competitive and effectively harness the power of their data.

References

1. VB Staff, (2019, July 19), “*Why do 87% of data science projects never make it into production?*”, Venture Beat, https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/

2. J. Saltz, (2024, April 10), “*CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*”, Data Science Process Alliance, https://www.datascience-pm.com/crisp-dm-still-most-popular/

3. B. T. O’Neill, (2019, July 23), “*Failure rates for analytics, AI, and big data projects = 85% – yikes!*”, Designing for Analytics, https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/

Previous Work

J. Saltz, N. Hotz, & A. Sutherland, (2022). “*Achieving Lean Data Science Agility Via Data Driven Scrum*”. In Proceedings of the Hawaii International Conference on System Sciences (HICSS). Available from: https://scholarspace.manoa.hawaii.edu/items/ec333162-1c18-455a-b5b4-c9e81bf7ddb7

J. S. Saltz, "*CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps*," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2337-2344, doi: 10.1109/BigData52589.2021.9671634. Available from: https://ieeexplore.ieee.org/document/9671634

K. Schwaber, J. Sutherland, (2020, November), “*The Scrum Guide*”, Scrum Guides, https://scrumguides.org/scrum-guide.html

Alternative Sources of Previous Work

J. Saltz, (2024, April 10), “*Data Driven Scrum*”, Data Science Process Alliance, https://www.datascience-pm.com/data-driven-scrum/

N. Holtz, (2024, April 28), “*What is CRISP DM?*”, Data Science Process Alliance, https://www.datascience-pm.com/crisp-dm-2/