**Data Science Project Management Methodology: Making the Right Choice**

**IST 644**

**Anthony Champion**

Managing a data science project is important because the probability of success greatly increases when a systemic approach is taken. One of the most eye-opening statistics on the success of data science projects is illustrated by “85% of big data projects fails [1] and 87% of data science projects never make it to production [2]”. With such a high failure right, the question must be asked. Why are data science projects so unsuccessful in delivering a solution?

Based on my experience there are two main reasons data science projects fail. The first reason being a lack of understanding of how data science derives what they would deem the best model and associated output. The second being overall process integration and adoption by the end user. For a data science team to be highly functional there is an element of data exploration and working through all the associated use cases or hypothesis that needs to occur to achieve the best outcome.

There are many steps that go into making a data science model for example data gathering, data cleaning, model generation, model selection, model output analysis, etc . These high-level examples of process steps must be repeated as results and different use cases or theories are derived. A person could assume that data science projects are iterative in nature. This element forces data science to go back, repeat the process again and test the next potential answer to the use case. This repetitive nature is the reason why data science project management is so critical. There is a constant churn of activities that need to take place to get to the end point of delivering a true data science product. With such a high fail rate in data science projects today, not having a structured project management approach only makes the probability of success even lower.

If you were to go out and ask the business community what they knew about data science you would get a lot answers like lots of data, predictions, high-level math, coding, computers and technology. When you see the words associated with data science, its reasonable to say this sounds very similar to the Information Technology space (IT). When you go a level deeper, that is where most of the similarities stop. For simplicity, think of an IT solution as software being developed or implemented and data science as a solution to a question using data to answer.

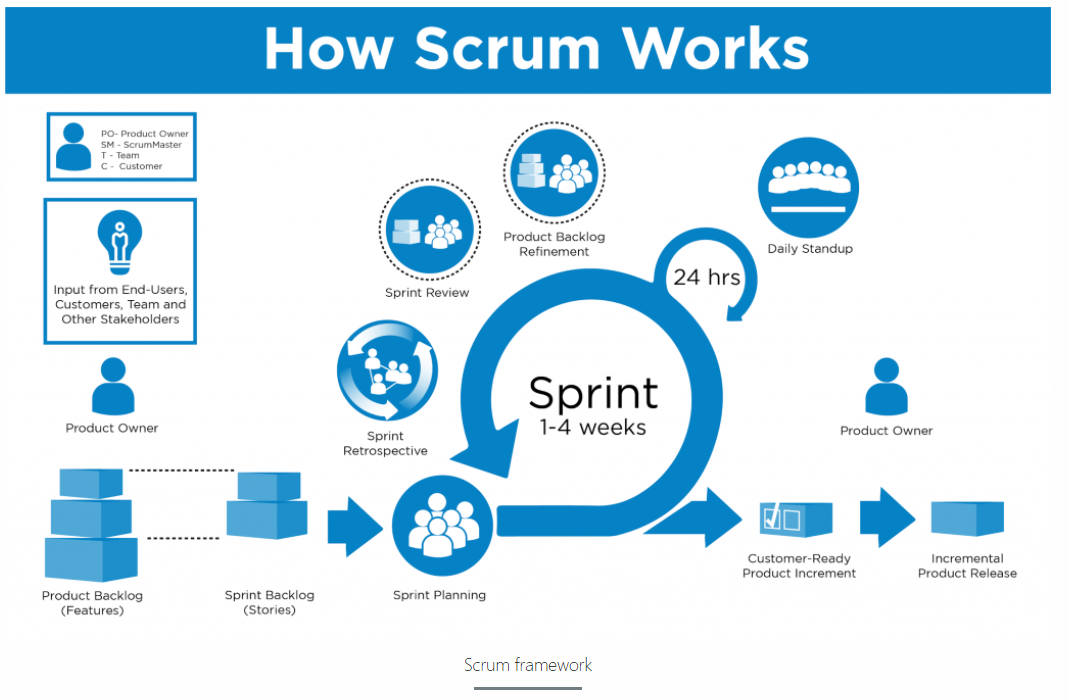
In a traditional IT project setting, a use case is created, and a technology or system is developed internally or purchased from a third part. Once that step has been completed the standard IT project management tools are deployed. Timelines and deliverables are finalized, and project team members understand the task at hand and when each individual piece is needed. IT project management takes a very rigorous approach to defining timelines and ensuring everyone is meeting those deadlines. During an IT implementation there are many meetings or touch points among the project team to communicate progress and risks to the over timelines. If you wanted to simplify an IT project it would be identify the technology business requirement, develop or acquire the technology, and implement technology. As an example this could a new ERP system (SAP), and new HR system (Workday), a new Payroll system (ADP), etc.

From a data science project standpoint there are many similarities to the IT project management approach but also many of differences. Data science projects are trying to answer a specific question or use case. An example of a simple data science question would be an airline company wanting to using net promoter score (NPS) to determine the impact on customer attrition. The IT side of this would be a technology that collects customer information and generates emails to airline customers once the flight is completed to collect and store NPS information. Both examples are working on a similar business case the desired outcome is very different.

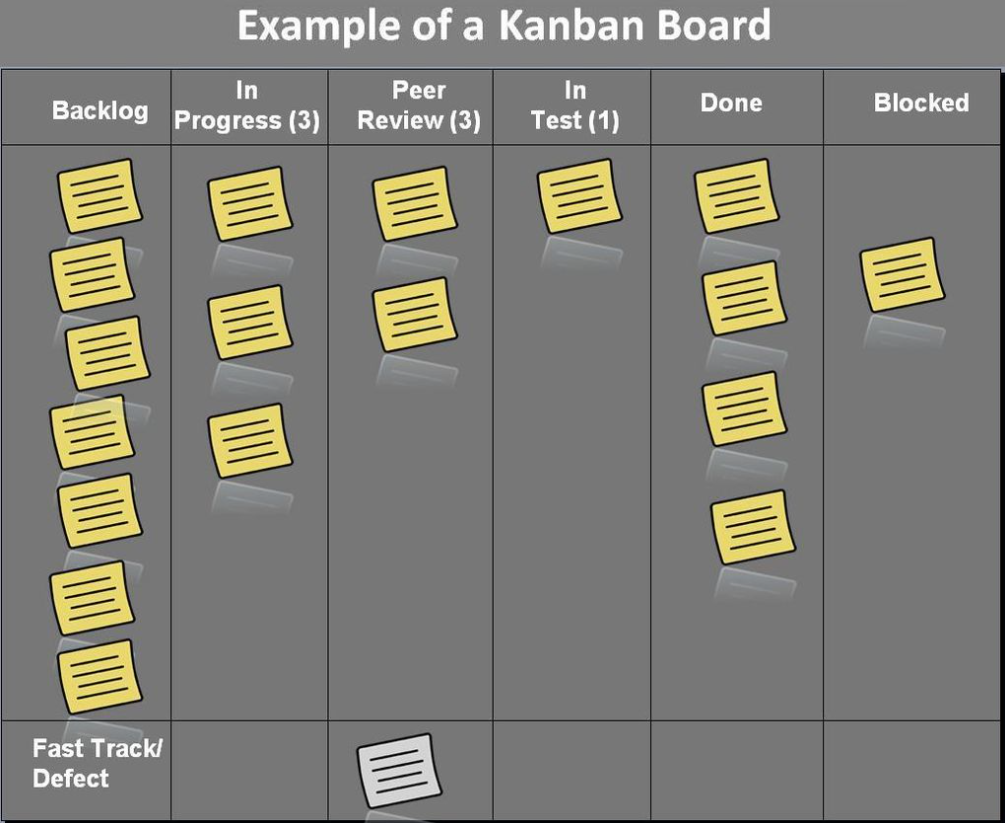
With IT projects occurring for the past seventy years, there has been a significant amount of project management expertise created. With that expertise comes many different approaches or frameworks that can be leveraged. As data science and big data analytics continues to mature. So has the research on how taking the key aspects of IT project management and tailoring them to the needs of data science project.

The most common IT project management framework is called scrum. This has become the most widely used agile approach by software development teams. According to [3] “70% of software teams use scrum or a scrum hybrid”. Scrum is a framework that takes a larger project and breaks them down into a smaller subset of activities. This subset of activities is commonly referred to as a sprint. A sprint is defined by [4] as, “a short, time-boxed period when a scrum team works to complete a set amount of work. Sprints are at the very heart of scrum and agile methodologies and getting sprints right will help your agile team”. These sprints have a hard deadline associated with the completion of the task and each project team member understands their respective deliverables. Each member of the project under the scrum methodology is put into the following roles: Product Owner, Scrum Master, and development team.

The other key attribute of a scrum approach is the level of communication. Daily there are meetings to review the product backlog of activities, the progress made on the completion of the current sprint, and what is needed to complete the current sprint. This type of approach is holding those accountable for their respected area of the scrum tasks within the sprint. There is an activity that occurs at the end of each sprint called retrospective, according to the scrum guide the sprint retrospective is [5] is “a recurring meeting held at the end of a sprint used to discuss what went well during the previous sprint cycle where the team evaluates itself and makes needed adjustment to improve prior to the start of the next sprint”. This type of team reflection is important in a project management setting. This will help with team members development and increase overall efficiency as the project continues to progress. Below is a visual representation of the scrum framework flow.



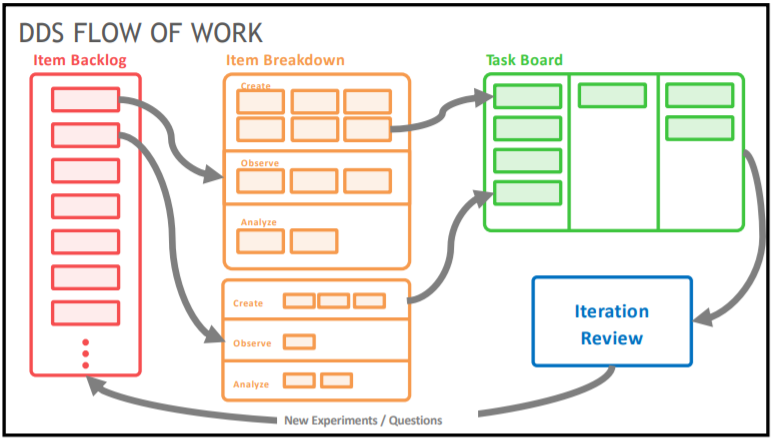
Another popular project management framework is called Kanban. This is another agile approach that leverages work items visually on what is referred to as a Kandan board. The Kanban board shows the work passing through each stage of the workflow. The goal of expressing the project visually is to try and eliminate bottlenecks in the process. This can also be referred to as trying to identify and eliminate all work in process (WIP). When activities stall or are in a WIP state (not moving towards completion) it is adding waste. Just like in the scrum framework Kanban also leverages heavy communication and transparency. Everyone can see the task at hand, what task is coming next, and what is causing potential delays. Below is a visual representation of a Kanban Board.



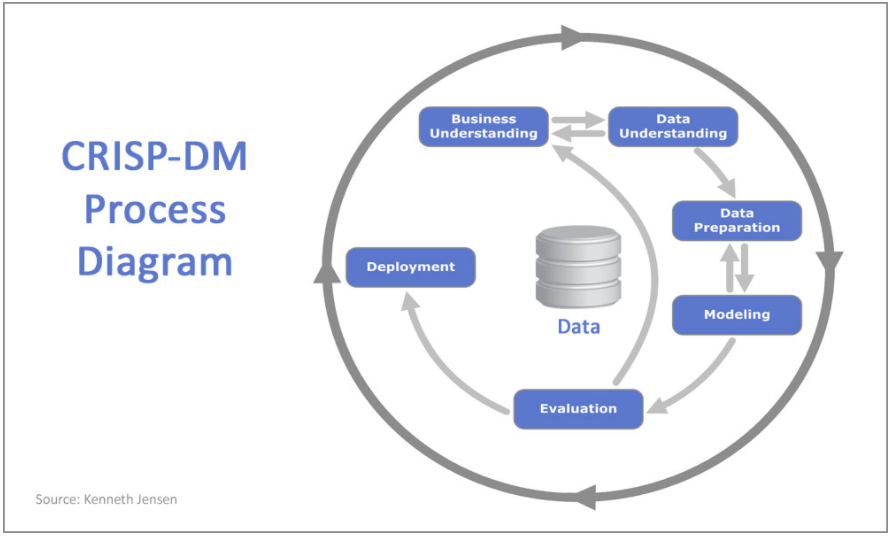
During the class IST 600 Managing Data Science Projects, a lifecycle framework was introduced called Data Driven Scrum (DDS). Which is a hybrid of scrum and Kanban methodologies, while also taking into consideration fundamental data science needs to increase the probability of success. There are three key principles to DDS, define capability-based iterations, decouple meetings from iterations, and create high-level item estimations.

Defining a capability-based iteration, there are times it makes sense that certain activities last a shorter amount of time than others. An example of this would be gathering and scrubbing data which could take weeks versus analyzing data which could take days. Why this is important is due to the many activities that take place prior to any model generation. Separating task by capability versus deliverable enables data science projects to still have creative flexibility. The next area of focus is decoupling meeting from iterations. Due to the length of time each iteration long or short [7] meetings should be based on a logical time-based window, not linked to each iteration”. The final principal which separates this framework from the others is to create a high-level item estimation. Instead of providing a detailed timeline of when specific activities will start or end. DDS takes the approach of generating detailed task estimations. This approach uses high level estimations to help drive prioritization of future iterations. From a process standpoint this methodology starts with a product backlog and then identifies the needed tasks to complete and then puts them into an item breakdown. The item breakdown puts task into categories of create, observe, and analyze. Once the task is identified in the backlog and further processed in the item breakdown. It is then shown visually in the Task Board and communicated daily among the project team. Again, this process is iterative in nature and once a task is completed that determines if more activities or different questions needs to be answered.

The team structure is broken down into product owner (similar to scrum), process expert (similar to Scrum Master), and then team member. Below is a visual representation of the DDS framework process flow.



The most common data science lifecycle framework is CRISP-DM. The stands for Cross Industry Standard Process for Data Mining. This framework breaks the process into [6] “six distinct phases, business understanding, data understanding, data preparation, modeling, evaluation, and deployment”. Below is a visual representation of this iterative framework., notice that prior to the deployment phase, the project can loop back and forth between the different phases if the desired outcome was not achieved. Said a different way, no deployment for inadequate results.



On a scale from one to five how likely would I suggest using DDS. Considering I have already told my data science team about it, the answer is five. All of the data scientists said the same thing, there is a clear need for something like this in data science. The other piece of feedback I thought was interesting was that as a team we don’t really follow a process. For the most part we have a general idea on how to start and what tasks need to be completed. In all honesty we’re making it up as we think of different items to address. In my time in this current role (11 months) we’ve never taken a step back and thought through all the needed activities and worked towards a systematic approach to completion. In small pockets we do this, but only a one-to-two-week view to get us to the next step.

Regarding is DDS better or worse than Scrum, the answer is DDS is better than. Scrum has an extremely rigid timelines and because of this it creates an environment where the path to least resistance is what data science will deliver. Whereas DDS levels sets up front what the approach needed to give the best chances of finding a quality solution has to be. The idea of creating capability-based iteration and being transparent on the different approaches needed to address the use cases does two things. One, it allows the data science teams to have time to perform exploratory ideas and two educates those involved that the best answer will be derived by testing different hypothesis. The other key aspect as noted above is the rigid timelines associated with scrum do not apply to DDS. A high-level estimation of time needed to complete the task are identified and communicated to all parties involved

In my opinion I think time restrictions are the biggest factor in picking a data science project management framework. When I look at the different methodologies outlined earlier (Scrum, Kanban, DDS, and CRISP-DM), I would argue they’re all very similar. The biggest similarity being that each one is trying to improve communication and coordination of the project. The fundamental difference between them is how the Data Driven Scrum methodology takes a high-level approach to defining time need to complete a task. As well as identifying what task are driving capability changes versus a task needed to move the project along i.e. data sourcing.

Another area of importance that goes under looked is the ethical issues that will arise through out many data science projects. There are data science related actives that if many people knew were happening would take issue. Let’s look at the company Facebook and Instagram. Their entire business model is predicated around users logging into the platform and sharing content, watching videos, seeing different advertisements, etc. These companies make no money if you don’t use their platform. With millions of customers and the ability to access and track unlimited data. They use data science to develop models to get you to log on to their platform as often as possible to watch content. Have you ever wondered how strategic the alerts are and how often you log in once you receive them? That’s because they have made a data science model that’s only purpose is to determine how many alerts can they send per day to get maximum customer usage. Most people would agree this is customer manipulation, but what is the line that data science should not cross?

The possibility for ethical issues to arise in data science are limitless, that is why it needs to be on the forefront of each data science team members mind. One of the common themes that can been seen in most frameworks is communication. In most frameworks, communication is around the project and deliverables. This also needs to the be the place where potential ethical conflicts are discussed. According to [8] there are three questions for each data science project that need to be asked. The first is to collect minimal data, and aggregate what’s there. An example of this would be “One-way companies can harness this power while heeding privacy worries is to aggregate their data...if the data shows 50 people following a particular shopping pattern, stop there and act on that data rather than mining further and potentially exposing individual behavior.” Identify and scrub sensitive data, restricting access and implementing data privacy issues allows data science teams to stay ahead of any potential breaches to sensitive information. Every data science team needs to have a plan in place in case the insight generated backfires. It may take time, but all results need to be constantly reviewed to ensure no bias or misrepresentation of the data is happening.

**Frequently Asked Questions:**

**How can we measure performance?**

There are many ways to measure the performance of a data science lifecycle project. The following performance measures are how we define success:

* Was the data science product deployed?
* Is the data science activity used by the customer?
* Was the implementation timeline achieved?

**How often do you change lifecycles?**

Data science projects are iterative in nature, so is data science as a whole. If the data science team is leveraging a specific framework and its yielding results in a systemic way, then the only change is to make slight enhancements internally to hopefully improve outcomes. If at any time our data science team feels a framework is restricting creativity or success, then we will always revaluate the need for change.

**Why is a framework important in Data Science?**

Data Science solutions can go in many different directions. Rarely is a model produced that doesn’t undergo many iterations. The goal is that with every iteration something is learned, and the data science team is one step closer to the best possible answer. A data science framework allows team members not to waste any energy at the task at hand. Without a logical flow of how to produce output, the chance for success is greatly lowered.

## References

## **[1] White, A. (2019, January 03). Our Top Data and Analytics Predicts for 2019.** [**https://blogs.gartner.com/andrew\_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/**](https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/)

## **[2] VB Staff. (2019, July 19). 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/**](https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/)

## **[3] digital.ai., (2021). Scrum Methodology.** [**https://digital.ai/glossary/scrum-methodology**](https://digital.ai/glossary/scrum-methodology)

## **[4] Rehkorf, M. (2021). *What are sprints?*** [**https://www.atlassian.com/agile/scrum/sprints**](https://www.atlassian.com/agile/scrum/sprints)

## **[5]** [**https://www.scrum.org/resources/what-is-a-sprint-retrospective**](https://www.scrum.org/resources/what-is-a-sprint-retrospective)

**[6]** [**https://www.datascience-pm.com/crisp-dm-2/**](https://www.datascience-pm.com/crisp-dm-2/)

**[7] Saltz, J., Hotz, N. and Sutherland, A. (2021). *Achieving Lean Data Science Agility Via Data Driven Scrum***

**[8] Taylor, J. (2020). 3 Key Ethics Principles for Big Data and Data Science** [**https://www.kdnuggets.com/2016/07/ethics-principles-big-data-science.html**](https://www.kdnuggets.com/2016/07/ethics-principles-big-data-science.html)