*Emilio Ramos Monzalvo*

*Syracuse University*

*managing a data science project*

IST 600 - Managing Data Science Projects - Summer 2021

**Table of Contents**

[1 Introduction to Managing Data Science Projects 2](#_Toc78710610)

[2 Data Science vs. IT Projects 2](#_Toc78710611)

[3 Data Science Frameworks 3](#_Toc78710612)

[3.1 Types of Frameworks 3](#_Toc78710613)

[3.2 Selecting and Using Frameworks 4](#_Toc78710614)

[4 Data Driven Scrum (DDS) 6](#_Toc78710615)

[4.1 What is DDS? 6](#_Toc78710616)

[4.2 Rating DDS 6](#_Toc78710617)

[5 Ethics 7](#_Toc78710618)

[5.1 Ethical Dilemmas in Data Science 7](#_Toc78710619)

[5.2 Using Frameworks to Prevent Mishandling of Data 7](#_Toc78710620)

[6 FAQ 8](#_Toc78710621)

[6.1 How do you set a timeline for a Data Science Project? 8](#_Toc78710622)

[6.2 What is the difference between a Life Cycle and a Collaboration Framework? 8](#_Toc78710623)

[6.3 Is a Scrum Master / Framework Expert necessary? 9](#_Toc78710624)

[6.4 What is the best framework for my team? 9](#_Toc78710625)

# 1 Introduction to Managing Data Science Projects

As the world started generating zettabytes worth of data, companies started trying to figure out how to use it. This eventually led to the creation of a new field in the corporate world called Data Science. The term ‘corporate world’ is used because Data Science, Machine Learning and Artificial Intelligence have been in existence before personal computers were created. Within the last 20 years companies like Google and Facebook were some of the first to have implement it and to actual create financial gain from it. Once others started seeing the benefits of Data Science, the data industry boomed.

Due to the recent growth of Data Science, professionals from other career paths started to transition into Data Science roles. In my workplace, this includes careers like Physics, Chemistry, Statistics, Mathematics, and Computer Science. While all these careers are technical in nature, the diverse backgrounds also complicate how they work together as a team. This does not mean that Data Science teams can’t be efficient and can’t deliver exceptional value to a company. It means that new frameworks or guidelines need to be appropriated by a team to create Data Science Driven collaborations among diverse backgrounds.

The recent corporate integration and diverse backgrounds resulted in Data Science projects being compared to research projects. This is also because there is no guarantee that the next dataset a Data Science team must analyze will be similar to a previous dataset encountered. Another factor is the domain of the project’s subject. Even if there is a similar problem statement, the domain ranges from financial transactions to photos taken of planets in space. Additionally, data comes in different formats like photos, text, steaming, and table format where each format has its own appropriate way to be processed. All these factors need to be understood by the Data Science team for a complete solution. Naturally, this causes Data Scientist to approach a problem without knowing what they will encounter. Consequently, a Data Science project becomes a research project.

# 2 Data Science vs. IT Projects

IT and Data Science teams both require similar skill sets, namely, a programming background. In some companies, Data Science teams are even part of the IT department because of their similarities, but this often leads to management making the mistake of implementing IT project guidelines onto Data Science projects.

IT related projects tend to have a waterfall tendency. Looping back to a previous step in the process is avoided through the teams’ ability to build upon what has already been created. The teams use a Minimum Viable Product (MVP) strategy where the client gets an updated version with the most important features needed for the product to function. The strategy helps the IT teams track and plan what will be done next to build upon and create the final product.

In contrast, Data Science projects do not know what the solution will be from the start of a project. The teams’ approach to a project’s solution can be compared to a ‘I will know when I get there’ attitude making it unrealistic to create a project timeline with a specific time of completion. Furthermore, Data Science projects do not always end up with a tangible solution unlike IT projects. The data is the guide for the projects, so if the data is too messy, the project can come to an end before a solution is created or it will prolong the project by an indefinite amount.

# 3 Data Science Frameworks

## 3.1 Types of Frameworks

To better the effectiveness of a Data Science team, management is encouraged to find a pattern of communication and collaboration while also using standard steps necessary to complete a project. In Data Science these guidelines are captivated by two types of frameworks: Data Science Workflows and Collaboration Frameworks where one leans on the other for a successful use.

Data Science Workflows, also known as Data Science Life Cycles, are meant to guide the team on the usual flow of a Data Science project. This means the life cycle gives specific instructions on what the team should do before moving on to the next step. For example, if a Data Scientist skips the Pre-Processing step and goes directly to modeling, it is not likely that the model will have good results. The life cycle prevents these mistakes form happening by creating a flow of necessary steps to be completed before the data scientist moves on to modeling the data. Additionally, the customer should be informed of the life cycle the team is using to understand what the team is spending their time on. Due to the complexity of a Data Science project, a customer might be frustrated if he does not understand where his money is going towards. With the life cycle, the customer will be aware of the plan’s complexity and how it is necessary before a solution can be finalized.

The Collaboration Framework serves a different purpose outside of Data Science. The collaboration framework is used for the team to communicate effectively, manage customer expectations, and to keep track of the project’s goals. Frameworks like Agile lead to increase communication among the team members while also creating a feedback loop from the customer to make sure the solution we are working towards is aligned with the customer’s goal. The goal of a project should be to create an actionable insight, but as Data Scientist are not domain experts, their understanding might be misguided if there is no customer communication. All in all, the frameworks allow the user to create confidence among all the project’s members that the team is moving in the right direction.

## 3.2 Selecting and Using Frameworks

The traditional frameworks like Gantt Charts and Waterfall have showed a lack of effectiveness in Data Science projects as highlighted above. Therefore, when the time comes to choose a framework, there are specific Data Science frameworks to help the teams. Firstly, the team must choose a Data Science Workflow that best fits their project. Life cycles tried to encompass all the usual steps a Data Scientist takes to create a final solution and how the steps flow with each other. Nonetheless, the complexity of a project might require additional steps that might not be addressed by all life cycles like Model Monitoring and Ethical Liability Monitoring.

The most popular life cycle includes CRISP-DM or a variation of it. CRISP-DM is broken down into six phases that capture most of the key elements of Data Science pipeline. For each step, there are key MVPs that are meant to create milestones for the team to present to the customer and assure the process is not flawed. This first step in CRISP-DM is the Business Understanding step. Here, the team should focus on understanding the problem statement and working with the customer to make sure the data is available and to plan how the insight can be used in the business. The next step is Data Understanding where the team pulls the data and works with Subject Matter Experts to understand column names and the flow of the data. For example, if the data is lagging by 1 day, but they want a real time model, the team would be unable to implement it. Up next is the Data Preparation step where the team usually spends most of the time cleaning data to make it ready for the modeling stage. Once the model is created, the team evaluates the model and if the model performs up to the customer’s standard, then the team can move on to the implementation phase. Now, if the model was not performing well, CRISP-DM also creates guidelines a backward propagation step where the team redesigns previous stages to improve the final solution.

For the Collaboration Framework, there are two popular methodologies. The first one, Scrum, focuses on significant team communication by demanding regular stand-up meetings among the team. The intension is to prevent the team from wasting time on hopeless efforts by creating a constant channel of communication. The team would also be aware of other’s efforts, so the work does not get duplicated. Finally, Scrum creates a flexible environment for the team to know when to quit or go back on the workflow. Scrum uses an MVP-like methodology where the team focuses on the next increment to the project. This is done in pre-defined time intervals where the team presents the new updated version of a solution and gets feedback from the customer. If the customer likes the solution, the team can move forward. Otherwise, since there is constant input from the customer, the team can easily change trajectories. Additionally, Scrum introduces the role of a Scrum Master. This role focuses on making sure the Scrum methodology is not being forgotten or ignored by the team by ensuring everyone is communicating and following the pre-defined path.

The second most popular collaboration framework is called Kanban. Kanban focuses on the creating a backlog of tasks for the team to complete. The backlog is displayed neatly in a Kanban Board where the team can see what needs to be done in order of importance without having any time bounds on when to deliver the final solution. In contrast to Scrum, Kanban allows more work flexibility and creates communication among the team using the Kanban Board rather than constant meetings. While Kanban does not provide specific guidelines on customer communication, it does create a stress-free environment for the developers. The more experienced members can grab the harder tasks from the backlog which allows the most junior members to take the less complex tasks which leads to the project moving forward faster.

# 4 Data Driven Scrum (DDS)

## 4.1 What is DDS?

Data Driven Scrum (DDS) is the solution created to take the best qualities of Scrum and Kanban and apply them specifically to Data Science projects. From Kanban, DDS inherits its backlog and Kanban Board while from Scrum it inherits its MVP way of thinking while also creating a Scrum Master like role and having Interval related meetings. Most frameworks were created prior to the Data Science corporate boom which makes them inefficient or outdated in some cases, but some of their ideologies can still be used which are inherited by DDS.

On the other hand, DDS’ main difference from other frameworks is the dynamic intervals that are meant to match the intensity and complexity of the tasks planned for that interval. The steps and tasks defined in CRISP-DM have different complexity and importance levels which frameworks like Scrum and Kanban do not take into consideration. In other words, the steps like the Business Understanding stage are crucial to move on to the next step, but it is also not complex enough to take two weeks before the next interval planning meeting where the team decides how to move forward by having a backlog of tasks ranked by ‘T-Shirt Size’ rankings.

## 4.2 Rating DDS

While DDS brings a Data Science driven guidelines for project management, it is increasing the complexity of an already complex field. For a new framework to be adopted by a team, the main driver can be the complexity of the transition. With DDS, the team will have to go through training sessions and will create outside stress for the developers.

If the team was already familiarized with Scrum and Agile in Data Science, the transition to DDS will be easier, but will create skepticism as to why the team is transitioning to another version of Scrum if it is already being used. Conversely, if Scrum has not been satisfactory for the team, there will be hesitation to implement a new framework with Scrum being a major part of it. Furthermore, teams with no experience with Scrum will have a much harder time transitioning to a complex framework like DDS. It would be simpler to start the team on Scrum, then transition to DDS once the team becomes familiar with Scrum guidelines. Nonetheless, Scrum as the first step creates conundrum by going back to the earlier question of whether transitioning to DDS from Scrum is necessary.

Theoretically, DDS would offer more flexibility and compatibility to the complexity of Data Science projects, but it fails to provide practicality for implementation in a work environment. Scrum has an already proven record within the industry even if it fails to directly address some Data Science related problems. The addition of a Kanban Board to a Scrum Framework provides a simpler format to be adopted while keeping the additional complexity to a minimum. When creating a new framework, the overlaying result should not be ranked or selected just by how many compatibility points were addressed regarding the project type. It should be judged by minimal amount of effort needed for the framework to be adopted while still increasing the productivity of a project. In that regard, the **Data Driven Scrum framework would score three out of five points** while Scrum would earn half a point more for its reduced complexity in the implementation.

# 5 Ethics

## 5.1 Ethical Dilemmas in Data Science

As data can be collected from every subject imaginable in human nature, Data Scientist inherit all the political and ethical responsibilities when mining, analyzing, and modeling data. Ironically, Data Scientist are not trained to think about these implications making it easier for end solutions to overlook them. Examples of data mishandling can have unprecedented effects from altering political opinions by breaking the privacy of individuals to racial discrimination based on image classification. Therefore, it is important for Data Science teams to plan for these problems prior to the project start by adding specific guidelines to a project framework.

## 5.2 Using Frameworks to Prevent Mishandling of Data

Due to the novelty of Data Science, current Life Cycles or Collaboration Frameworks fail to address ethical responsibilities, but it does not mean that they cannot be adapted to include it. Project life cycles offer a generalized step by step guide on the usual flow of a Data Science project. Most cycles like CRISP-DM start with a stage like the Problem Understanding stage where the team should start thinking about possible ethical dilemmas that should be addressed throughout the project. Nevertheless, the team should not stop tracking possible ethical dilemmas after the first stage. There is a possibility of ethical problems arising at every stage of a life cycle like creating a bias at the Modeling stage or merging personal information databases illegally at the Data Cleaning stage. This means that a life cycle like CRISP-DM should include key deliverables at the end of every stage to make sure the integrity of the data is kept throughout the whole project. In other words, at the end of a phase, the team needs to make sure there is no liability if the team moves forward with the solution with what was used or created during the phase.

# 6 FAQ

## 6.1 How do you set a timeline for a Data Science Project?

Due to the similarities of a Data Science Project to a Research Project, coming up with an exact timeline is unrealistic. The similarities of problem statement do not directly translate to the data itself. The Data Science will pick up a dataset and try to tell a story based on it, but the dataset is the tool that does the talking. Unforeseen data integrity problems will cause projects to be prolonged indefinitely until the concerns are understood and solved. Nonetheless, the client can become frustrated with the intricacy of a Data Science project’s timeline, so it is important to provide the client with a taste of what can be achieve if the project ends with a successful solution. This can be done by including early key deliverables by working on a small subset of data and providing an analysis to showcase progress.

## 6.2 What is the difference between a Life Cycle and a Collaboration Framework?

A project management guideline can be generalized into two types of frameworks. The first is the Project Life Cycle which provides the typical and essential steps that a project must go through before finishing a project while also creating a visual flow of how these steps interact with each other. Additionally, it provides specific key deliverables that are meant to protect the team from picking up additional work later in the project while also giving the client an understanding of what has been accomplished. A Collaboration Framework builds on top of the Life Cycle to create a project environment with increased communication, flexibility and understanding. It accomplishes this by providing key guidelines on when the team should meet, how the team should keep track of progress and who should be part of the project planning. The two types of frameworks are not required for a team’s success in a project, but they are recommended as both are intended to increase the prosperity of a Data Science team.

## 6.3 Is a Scrum Master / Framework Expert necessary?

Collaboration Frameworks sometimes include an additional role which can be described as a framework expert.

## 6.4 What is the best framework for my team?

There is no one framework which will address every aspect of every project a team will encounter in their life span. Nonetheless, a team should pick a project management framework that fills the gaps of their current process and that does not add significant stress and complexity.