THE REFLECTION PAPER

IST 644 | MANAGING DATA SCIENCE PROJECTS

Dan Tully

TABLE OF CONTENTS

1.0 ABOUT DATA SCIENCE PROJECTS

1.1 WHY MANAGING DATA SCIENCE PROJECTS IS IMPORTANT

“Data science project management is a methodology that involves managing projects using standardized methodologies. Without standardized methodologies for managing data science projects, teams often rely on ad hoc practices that are not repeatable, not sustainable, and unorganized. Such teams suffer from low project maturity without continuous improvements, poorly defined processes and checkpoints, or infrequent feedback.”[i]

Data science (DS) project management provides the methodology to recognize the warning signs of a bad process. A warning signs include too many concurrent projects with the same team. Stakeholders that do not trust the insights generated by the DS team are useful, productive, or the highest priority tasks for the company are other signs. Communcation with the stakeholders, management, and the DS team are predominant to the success of any DS project. The DS project management methodology provides the tools necessary to have a successful DS project.

There are many benefits to a well-defined process. Three I will discuss are: Team efficiency, result validity, and actionable insight. Team efficiency can be measured by stakeholder confidence that the team will be following a standard repeatable process. This process will drive process efficiency and high value analysis. Next, Results validity will help ensure accurate, fair, non-biased, and transparent results. Finally, actionable insight, through better communication with stakeholders, will provide analysis that can be used to make better business decisions.

With increased focus on efficiency, result validity, and actionable insight, the DS team will provide perceived benefits to the stakeholders and business. DS projects allows for organization confidence in developing an effective repeatable process to better understand the business needs, acquire, explore the available data, model (& evaluate) that data, and finally deploy (& communicate) the insight or actionable results to key stakeholders. Not every project is going to be successful, but learning from those mistakes, and improving the process iteratively will move the DS team, stakeholders, and the company from a defined state to a managed or even optimized process.

1.2 DATA SCIENCE PROJECTS ARE SIMILAR AND DIFFERENT WHEN COMPARED TO OTHER IT PROJECTS

Project management has many similar processes across all projects, including IT projects. Overall, these processes strive for team efficiency, result validity, and actionable insight. Although the phases may not all be called the same as these, they include the need to initiate, plan, execute, monitor/controlling, and deploy/close the project. These 5 phases are the ones outlined in the Project Management Body of Knowledge (PMBOK). Project Management sets a standard for directing and managing work in the project. "With project management, organizations have the ability to apply knowledge, processes, skills, and tools and techniques that enhance the likelihood of [project] success"[ii] These phases also set out to establish some quality assurance that the project, process, and products are repeatable, accurate, and actionable to fit the business needs. Team focus is particularly important, as each team member should understand the scope of work and have the appropriate skills to complete that work. All lifecycles discussed demonstrated the importance of communication and key stakeholder engagement. Discussing issues and changes with the key stakeholders will significantly increase your chances for a successful project. All project teams need to understand, control/mitigate, and communicate risk. Finally, there is a need to systematically close/deploy the projects. The deployment could be a one-time report/analysis or reoccurring. Documenting lessons learned should also be a similar step regardless of what type of project you are working on.

Many projects want to be agile! Agility could mean different things to different people when it comes to DS project management, but overall, it is the "The ability to think and understand readily and quickly[iii]" In contrast to rigid planning and processes, the Agile manifesto has 4 values which places a higher priority on collaboration, customer satisfaction, and adaptability particularly in software development. This manifesto can also be summarized to: Team efficiency, result validity, and actionable insight. The main principles are to provide iterative value early and continuous. Strong emphasis on a collaborative self-organized team environment. The overall benefits of agility include more relevant insights by defining tasks just before analysis, providing quicker delivery and feedback, while improving overall communication with the customer. It also

provides an opportunity to learn sooner in the process, which prevents wasting resources. The fundamental ways that DS projects differ from other IT projects are that DS is exploration focused, encounters unique data challenges, analysis is often non-linear, and outcomes are more ambiguous.

2.0 DATA SCIENCE FRAMEWORKS

2.1 OVERVIEW

Frameworks and/or lifecycles are great in that they provide a common vocabulary to describe what the team does in a project. They also provide a familiar process or steps for a team to use to progress through the project. They share a mental model of the steps to be taken. There are several lifecycles to consider (e.g., CRISP-DM, TDSP, OSEMN, SEMMA, etc.) on DS projects, below we will discuss a few. There is no absolute right or wrong answer when it comes to picking a lifecycle since each project has its own unique set of challenges.

Collaboration frameworks (e.g., Kanban, Scrum, and Data Driven Scrum) focus on improving team's collaboration and communication to understand goals and prioritize work. These frameworks not only define tasks and prioritize work but also solicit feedback from the stakeholder and seek process improvement. Coordination of communication with the DS team throughout the process will improve project success.

2.2 KNOWLEDGE DISCOVERY IN DATABASES (KDD)

KDD (KNOWLEDGE DISCOVERY IN DATABASES) emerged in 1989 to represent the overall process of collecting data and refining it to extract knowledge. "KDD is a multi-

step process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.[iv]" This process is summarized into five key steps: selection, pre-processing, transformation, data mining, and interpretation/evaluation.

In the first step, you will identify the goal from the customer's perspective while selecting the target data set or subset that could support the business objective. Next, you want to clean/normalize the data and/or handle any missing data. You will then utilize data mining methods (e.g., classifications, clustering, regression, etc.) to discover any hidden patterns in the data. Finally, you want to extract and document knowledge from the data while communicating the findings to interested stakeholders.

KDD is a logical, repeatable process which influenced many of the modern-day frameworks we will discuss later in this paper. This process also considers data storage, access, and scaling. "The ultimate goal is to extract high-level knowledge from low-level data. [lv]"

2.3 SEMMA

"SEMMA is an acronym that stands for Sample, Explore, Modify, Model, and Assess. It is a list of sequential steps developed by SAS Institute" SEMMA serves as another reference to the data science life cycle.

It begins with data sampling. When considering the right data set the team must evaluate that the data contains sufficient information to be relevant. In the explore phase, the team is looking for relationships between the variables, including abnormalities or outliers. The next phase, modify, is all about preparing the data for modeling. In the model phase, the DS team applies modeling techniques to achieve the desired outcome. Finally, in the Assess phase, the team is evaluating the reliability and usefulness of the model.

SEMMA is focused on the data and not as much focused on the business part of the process, at least up front. The last part discusses the usefulness of the created models, but a criticism would be the lack of business understanding up front prior to the analysis. These steps can seem directly focused on the data analysis or exploration performed by the DS team.

2.4 CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

CRISP-DM is touted as the most widely used analytic model. This model describes a common data mining approach used by many DS teams. This lifecycle describes 6 major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each of these phases includes defined tasks and deliverables. The process allows flexibility to reassess prior phase to increase understanding or even to refine data preparation for better modeling.

The initial focus is to understand the business needs and objectives to properly align the technical work. Data understanding makes it more likely that the availability of underlying data and potential connections will be discussed. The next step, data preparation, involves cleaning and munging the data - which is - frequently the most time-consuming part of a project. Many people associate data science with the modeling phase, which involves using methods like machine learning to extract knowledge from the data. The objective of the Evaluation phase is to determine the degree to which the model satisfies the business outcomes and may entail actions such as preliminary A/B test on a sample to gauge the model's effectiveness. Depending on the project, the deployment phase can range from one-time analysis to real-time prediction.

2.5 OSEMN

OSEMN is an acronym that stands for Obtain, Scrub, Explore, Model, and iNterpret. “[This OSEMN framework] are the steps that data scientists follow chronologically in a typical data science project.[v]” The first step is obtaining the data needed to complete the research. This step could be challenging depending on the format of the source data. The data could come from structured or unstructured data sources. Next is scrubbing the data, which has its own unique challenges. Most of the DS team time is spent in this area of the process. The DS team will spend time cleaning and munging the data to normalize the data set in preparation for analysis. Exploring the data will be an especially important phase to determine data type or using descriptive statistics and other tools to better understand the datasets identifying patterns and trends. Now that all the preparations are complete, the DS team can start modeling the data. “[U]se regression and predictions for forecasting future values, and classification to identify, and clustering to group values. [v]” The last step is iNterpret, this is more than just reading the results of the model. One essential skill the DS team needs is the ability to tell a clear and actionable story. The team must be able to brief the stakeholders on the results in a way that leaves the client(s) with actionable business intelligence.

2.6 Harvard

Harvard's data science workflow consists of 5 steps: ask an interesting question, get the data, explore the data, model the data, and communicate / visualize the results. The part of this workflow that is interesting is the first step of asking an interesting question. In the scientific method all research starts with asking a question. DS teams should find it refreshing that this workflow incorporates the core scientific process to start the workflow. The next step, getting the data, is common for most of these workflows. The unique challenges the DS teams have with it, although not overtly mentioned, is cleaning/normalizing the data is quite an elaborate process. Next, we get into exploring the data which allows the DS team to better understand the dataset. Lastly, communicating the results with visualization is also very important. The greatest discovery could have been made but without the proper presentation of those results, the significance could still go unnoticed.

This workflow starts with asking an interesting question which makes it unique, it is not that popular in the DS community. The remaining steps are familiar to many of the other workflows which is why this life cycle does not stand out much amongst the others. I am also unclear why they do not specifically call out cleaning or munging the data as that is one of the most time-consuming steps in the DS project. I do appreciate that communicating results by effectively telling the story is emphasized.

2.7 Domino

The Domino's life cycle consists of six steps: Ideation, Data Acquisition & Exploration, Research & Development, Validation, Delivery, and monitoring. There are three guiding principles of this framework. First, expect iteration and embrace it; projects should anticipate the need to go back to an earlier phase. Next, create components that can be reused in other projects to enable compounding collaboration. Finally, consider the model's assessment requirements and that it frequently needs to be recreated to preserve artifacts.

This framework has educated steps identified through the process. The flow chart is easy to follow, and it encourages the DS teams to move back through previous steps when necessary. It is machine learning operations focused which aims to deploy models that are reliable and efficient. This lifecycle is deemed more for marketing publicity and not widely used by other DS teams outside of Domino's. I do appreciate that Domino's advertises its use of DS and Artificial Intelligence to make better business decisions.

2.8 Uber

Uber created a simple yet compounded framework. On the simple side, there are four main steps: Define, Protype, Production, and Measure. Once you start looking at each step you begin to realize how more comprehensive it becomes. This first step defines the importance of understanding the business need and introduces the concept of a minimum viable product (MVP). An MVP is a prototype of a product that has just enough features for early adopters to use it and to provide feedback on how to improve it in the future. The next step is prototype, in this step the team is collecting and preparing the data. They are also training and evaluating models. This is an iterative process to refine the data to fine-tune the model results. In the next step, production, the DS team is deploying the models and making predictions from them. In the last step, measure, the DS team monitors predictions and analyzes insights from the model to determine next steps.

The Uber model is not widely used except by Uber. Uber's operation is highly data centric as it receives all kinds of data from its users daily. This data must be processed into actionable business intelligence. A particular focus is on the scalability of "analytics cluster compute resources which process hundreds of petabytes of data every day.[vi]"

2.9 TDSP

This Microsoft framework consists of 5 stages: business understanding, data acquisition & understanding, modeling, deployment, and customer acceptance. This framework also identifies 4 team roles and documented 10 artifacts.

The first of the four team roles outlined is the Group manager. The Group manager is responsible for the entire DS unit with the enterprise. The second is the Team lead, who manages the DS team. Third is the Project lead who directs the daily activities of the DS team on a specific project. Finally, the individual contributor who executes the project.

Some of the artifacts in the first step can include the charter document or data dictionary. In the second step, the DS team will produce the data quality report and solutions architecture. Next, they set features and document the initial model report. In the second to last step, the DS team creates a status dashboard and final modeling report. In the last step, the team provides an exit report.

2.10 SCRUM

Scrum is one of the most popular collaboration frameworks. Scrum is concentrated on creating and carrying out an incremental delivery, referred to as a sprint. The main idea

is that work products are broken down and at the end of each increment (usually 1-4 weeks) a product is delivered to the stakeholders for critique. Plans are revised for the next increment based on the feedback. During the sprint, the whole team has a daily standing scrum meeting for 15 mins to answer three questions: what you accomplished yesterday, what will you do today, and what obstacles are in your way?

The self-managed Scrum team also has three main defined roles. The *Product owner* establishes and assigns priority to the Product Backlog's list of potential product features. As a servant leader, the *Scrum Master* facilitates the entire process. Finally, the *Development team* is a cross-functional team of 5 – 9 individuals that complete sprints to deliver priority items. There are no changes to the priority items during a sprint, the team must complete the sprint before making changes. However, in unique situations the team can cancel the sprint.

There are five events that must take place. Prior to the sprint execution, there is sprint planning. This is where the team sets the plan for the sprint. As previously mentioned, the Sprint iteration is a smaller subset of the priority focused project tasks. During the sprint, the team iterates through a four-phase cycle: requirements, design, code, and test. The daily scrum meeting is used to keep the team focused on the sprint goal and to inspect progress. After the sprint, the team has a sprint review. Here they discuss the deliverable work product along with the desired goals. They get feedback on the work from the stakeholders. Lastly, the team holds a retrospective meeting which is completely focused on the process. Some questions may include: Did the team follow the scrum process? What went right or wrong? Is the team getting the necessary resources to be effective?

There are three main process artifacts: the product backlog, sprint backlog, and increment. The Product backlog is the product goal or a list of desired items to complete to reach that goal. Next is the sprint backlog, this is the team's plan or goal for the sprint iteration. Finally, the increment, this is the output deliverable achieved during the sprint which brings the project closer to its overall product goal.

2.11 LEAN

This concept or philosophy was developed to eliminate waste in the manufacturing process. It has been adopted by other business competencies, including Information Technology. Lean has three foundational guides: Understand the customer needs, Limit work in process (WIP) to avoid unnecessary holdups, and to continue streamlining activities to better serve the customer. The three key Lean data science principles are: iterate often, validate, and continuous improvement.

The *iterate often* principle establishes an MVP or minimum viable increment (MVI). The MVP/MVI reduces uncertainty, waste, and increases innovation. The team focus on *validate* hinges on understanding the value of what will provide the greatest impact to the product with the least effort. The final principle, *continuous improvement*, is learning from failure/success and empowering the team to adapt as needed.

2.12 KANBAN

Kanban (means “billboard” in Japanese) is a commonly found implemented tool in many business competencies, including Information Technology. Kanban has two primary principles: Visualize the flow and limit WIP. The first principle starts with a potential list of tasks not started. Those tasks, when assigned, progress to in process/being worked. Finally, when the work is done (and validated) the task is identified as complete. The second principle, limit WIP, sets a limit to the number of tasks assigned. This WIP limit prevents teams/individuals from being overwhelmed.

Some additional principles include providing opportunities to identify resources needs, encouraging the team members to see the “big picture.” Increase individual ownership in the process or at the very least they see how their contributions add value to the overall project. Individuals may seek professional development to expand their skills if they see a need to help in other areas of the project. The principle could also measure lead and cycle times to help determine how quickly similar types of tasks can be accomplished.

Kanban does not define policies (ex. WIP limits, definition of complete – quality of deliverables, how to prioritize tasks, etc.). It also does not specifically define the team's role, timelines, or meeting cadence. It does however improve communication, enable agility, and adaptation. Kanban boards are visually pleasing and easy to understand/follow. They allow anyone to quickly understand the progress of the tasks in the defined project.

2.13 DATA DRIVEN SCRUM (DDS)

DDS focused on addressing two key challenges when trying to use scrum for data science. The task estimation is unreliable, leading to a sprint length being more dynamic instead of static throughout the project. The key pillars of DDS include only high-level estimation, decoupling meetings from an iteration, and allowing capability-based iterations.

In DDS the iterations use a visual board (e.g., Kanban) to view the project progress. The team focuses on work on prioritized item(s) (e.g., MVP/MVI). The iteration is task-

based not time-boxed. Lastly, each iteration validates a specific lean question. During the iteration, the DDS focuses on executing Create, Observe, and Analyze. During the first step, create, the team focuses on what is the business objective(s) and what is the goal product to meet the objective(s). The next step is to observe, which are a set of outcomes that will be measured. The last step is to analyze those results and plan for the next iteration.

DDS has a product increment (PI) meeting which is time-boxed but decoupled from the iteration. The PI meetings are used to prioritize iterations to best achieve the business goal. PI meetings are collaborative which helps stakeholder expectations and provides a framework for teams to engage the stakeholders. PI increases transparency, communication, and sets realistic expectations for team and stakeholders.

The self-managed DDS team also has three main defined roles. The *Product owner* establishes and assigns priority to the Product Backlog's list of potential product features. As a servant leader, the *Process Expert* facilitates the entire process. Finally, the *Development team* is the cross-functional team who, through iterations, deliver Product Backlog Items (PBIs).

There are five events that must take place. Prior to the iteration execution, there is an Iteration review. This occurs regularly and is where the team sets the plan for the iteration. The iteration consists of priority focused project tasks or PBIs. The daily meeting is used to keep the team focused on the iteration goal and to inspect progress. An Incremental meeting is established to discuss the deliverable work products along with the desired business goals. They get feedback on the work from the stakeholders. Lastly, the team holds a retrospective meeting which is completely focused on the process.

There are three main process artifacts: the PBIs, Item backlog – prioritization of PBIs, and a task board (e.g., Kanban). The PBIs are the product goals to create/observe/analyze. Next is the item backlog, this is the team's plan or goal for the iteration (the prioritized list of PBIs). Finally, the task board is the visual representation of progress on work items.

2.14 INTEGRATE DIFFERENT TYPES OF FRAMEWORKS

Integrating frameworks can be immensely helpful to a successful data science project. The goal of the lifecycle and/or frameworks is to perpetuate team buy in on a project. These frameworks reduce process complexity by providing standard vocabulary, setting

expectations, and establishing collaborative environments. They encourage the teams to incrementally improve a product or even the project process.

The concept of integrating frameworks can be described as applying or overlaying a framework on top of the original framework to fill the shortcomings of that original framework. What I mean is that some frameworks or lifecycles are more focused on DS team data science analysis (e.g., OSEMN) rather than the timelines of the project while other frameworks (e.g., SCRUM) are more focused on the timeliness of the project rather than the product analysis. In theory, combining the frameworks (e.g., OSEMN & SCRUM) would produce a framework with structured time and thorough analysis. That sounds great, but as we discussed time-boxing data science analysis is not the ideal plan for successful DS projects due to its exploratory nature.

2.15 SELECTING ONE (OR MORE) FRAMEWORKS FOR DS PROJECT

CRISP-DM and DDS seem like the best combination of a detailed DS lifecycle along with a focused collaboration framework. Together they complement each other very nicely and provide an optimal environment for success on your next DS project. For starters, the CRISP-DM lifecycle is among the most popular in the DS community. As we outline in section 2.4 above, it has a simple to explain approach but very thorough analytical process. The DDS framework, as discussed in section 2.13, provides a "scrum-like" structure with strategic deviations which make it better suited for DS projects.

When choosing the frameworks, you want frameworks that complement one another. So, one DS lifecycle and one collaborative framework. You do not want two DS lifecycle frameworks because all that would do is baffle the DS team. There would be two different vocabularies, documented artifacts, and overall processes that would cause redundancy and unnecessary hesitation. Complementing the DS lifecycle would be a good collaboration framework which would help establish timelines for project deliverable and increase transparency and communication.

Mixing and matching the DS lifecycle and collaboration frameworks is more of an art than science. You want to select the frameworks that work best for your team and project. There really is no right or wrong answer, well a wrong answer could be - your team is not familiar with the framework selected/implemented. The team must buy-in to whatever framework you select. To obtain that buy-in the framework must be easy for the collective team to use and they must see the advantage of using it.

An effective team needs both a proper workflow and collaboration framework. Your framework may need adjusting if you see some of the common signs of a poor process. One sign would be the DS team does not perceive it is working on the highest priority items. The DS team lacks focus or understanding of what the business needs are, which is another sign. Next, Insights from the DS team analysis are not perceived as useful. Lastly, there is a noticeable disconnect or lack of communication between the DS team and the stakeholders.

2.16 REVIEW 3 LIFECYCLES AND 3 COORDINATION FRAMEWORKS

...

framework	scale	useful	explain
CRISP-DM	5	This lifecycle is most useful for data science projects	I selected the highest rating because CRISP-DM is touted as the most widely used analytic model. This model describes a common data mining approach used by many DS teams. This lifecycle describes 6 major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each of these phases includes defined tasks and deliverables. The process allows flexibility to reassess prior phase to increase understanding or even to refine data preparation for better modeling. The steps are easy to understand and provide the team with a common vocabulary. This process follows steps that align to a logical DS process. This is probably because this framework was developed in collaboration with several companies and not tailored to any one specifically.
OSEMN	3	This lifecycle is useful for data science projects	I selected this average rating for OSEMN since it is primarily focused on data science specific activities. Obtain, Scrub, Explore, Model, and INterpret are all commonly known vocabulary and processes for DS teams. It is easy for the team to understand and implement. The reason it did not receive a higher score is that this workflow is so focused on the DS tasks and analysis, it could easily lose focus on the business understanding before the DS activities begin as well

			as other significant activities during and after the project is complete. There is no emphasis on communication with stakeholders, looping back for iterative improvement, or deployment monitoring.
TDSP	4	This lifecycle is more useful for data science projects	This TDSP received a higher-than-average rating because of all that it has to offer. TDSP has 5 stages: business understanding, data acquisition & understanding, modeling, deployment, and customer acceptance. This framework also identifies 4 team roles and documented 10 artifacts. I really like the start and end of the project to be identified. This fully embraces the definition of a project from the PMBOK Guide, "A project is a temporary endeavor undertaken to create a unique product, service, or result. [ii]" It is not perfect as it lacks specific focus on evaluation, validation of the data, or communication with stakeholders except at the end when customer acceptance is sought. If you wait until the end to get stakeholder feedback, you may have missed the essence of the business need.
Scrum	4	This coordination framework is more useful for data science projects	I rated SCRUM above average because Scrum is one of the most popular collaboration frameworks. Scrum is concentrated on creating and carrying out an incremental delivery, referred to as a sprint. The main idea is that work products are broken down and at the end of each increment (usually 1-4 weeks) a product is delivered to the stakeholders for critique. Plans are revised for the next increment based on the feedback. Scrum has 5 events, 3 identified roles, and 3 artifacts. I did not rate it as perfect because it poses challenges for data science: the time-based iteration is difficult because DS tasks are hard to estimate due to their exploratory nature. There are tasks that require more time than others. Product backlog re-prioritization only after the sprint is complete.
Data Driven Scrum (DDS)	5	This coordination framework is most useful	This DDS framework scored the highest because it was built specifically for DS projects. This framework took all the best of SCRUM, but it strategically modified it to better support DS projects. Like SCRUM, DDS has 5 events, 3 identified roles, and 3 artifacts. The vocabulary and how the framework is applied differs from SCRUM. DDS also uses a visual

		for data science projects	board, focusing work on the specific item(s) during an iteration. Iterations are task-based. Each iteration validates a specific question. The fact that it closely matches the SCRUM process is also a distinct advantage because it brings a sense of familiarity to the process. DS teams can explain the similarities and differences easily since some stakeholders may already be familiar with the SCRUM framework.
Kanban	4	This coordination framework is more useful for data science projects	I love Kanban boards because of their simplicity. With only two primary principles, Visualize the flow and limit WIP, it is extremely easy to understand and implement. Even the more complex Kanban boards, done correctly, are easy to follow. It provides a quick visualization of work making it quite easy to identify bottlenecks and hold-ups in the process. WIP limits prevent the work from piling up on one team or individual keeping a steady level of work and not overwhelming anyone. There are clear downsides to this framework so you will have to integrate it with another framework to have a successful DS project. It is missing timelines for meeting and/or collaboration. It is more of a useful tool used with another framework than a standalone one.

3.0 ETHICS IN DATA SCIENCE

"New technologies often raise new moral questions...There are no currently agreed on responses to these questions. Nonetheless, it is extremely important to confront them and to attempt to work out shared ethical guidelines. Where agreement is not possible, it is important to attend to the competing values in place and to specifically articulate the underlying assumptions at work in different models.[vii]"

We discussed in class that a bias is a deviation from expectation in the data or outcome. We also discussed four types of fairness: max profit, group unaware, demographic parity, and equal opportunity. In the max profit type, the groups are not treated the same as the groups have different thresholds or standards. The next, group unaware, all groups are held to the same threshold. Demographic parity, the number of

a particular demographic will be represented in every group the same. Lastly, the equal opportunity, the proportion selected is the same for each group.

3.1 POTENTIAL ETHICAL SITUATIONS IN DATA SCIENCE

Ethical situations in the form of bias and fairness can easily enter data science projects. One ethical situation could be individual privacy. Personally identifiable information (PII) is defined as any information that can be used to identify a specific individual. There are laws (e.g., The Privacy Act of 1974, Privacy Act 1988, the Fair Credit Reporting Act, etc.) protecting PII disclosure which must be understood and followed while understanding the business needs. If analysis of PII data is necessary, teams should provide reasonable security of the data and limit access. Background checks on the DS team could be required and/or a non-disclosure agreement could be initiated.

Ethical use of data is another challenge we see every day. Just because the data is available does the use of that data for any project aligned with the how or why the data was collected in the first place? Consent due to power imbalances can exploit an individual's personal data for profit. This happens every day online. Companies sell individual information and trends for profit. You can reduce the ethical dilemma by providing a disclosure to the individual or by providing the individual with options (I.e., as we see on about every website with cookies acceptance).

Data and algorithms can reinforce stereotypes, prejudices, and other human biases. If data is selected from a particular region, it could reflect primarily one demographic only. It is important our DS teams document their work to be transparent when using training models, to prevent mistakes or unethical decisions without justification or accountability.

The issues surrounding artificial intelligence (AI), which includes machine learning, is another current concern. Preventing bias or unfairness from entering the results of AI is difficult since the collection of data can be unsupervised, lacking transparency, and human judgment. One article[viii] suggests 7 steps to reduce these risks:

1. Identify existing infrastructure that a data and AI ethics program can leverage.
2. Create a data and AI ethical risk framework that is tailored to your industry.
3. Change how you think about ethics by taking cues from the successes in health care.
4. Optimize guidance and tools for product managers.
5. Build organizational awareness.
6. Formally and informally incentivize employees to play a role in identifying AI ethical risks.
7. Monitor impacts and engage stakeholders.

4.0 FAQ

4.1 Question: How do you prioritize work?

Answer: You can prioritize work in 9 steps:

1. Define Objectives: Understand business objectives & ask clarifying questions.
2. Define Metrics: set milestones or goals to effectively measure progress.
3. Ideate: Know the product features & collect feedback (e.g., from stakeholders)
4. Estimate Value: identify the potential impact compared to business goals.
5. Estimate Effort: identify what level of effort will be required for each item.
6. Assess Uncertainty: risk-based approach; measure the risk & uncertainty
7. Identify Dependencies & Synergies: determine prerequisite & Lessons learned
8. Prioritize Value / Effort: MVP/MVI focused
9. Repeat!: Prioritize frequently. Cycles will vary

4.2 Question: What are three types of scaling agile teams or products?

Answer: There are three types of scaling agile teams or products:

1. Multiple teams working on a single product (Scaling teams) - Coordinating DS teams with other teams on the same product/project.
2. Multiple teams working on multiple products (Scaling teams) - This is useful for coordinating multiple data science teams. Each team manages one product.
3. Scaling the solution (Scaling products) - this is going from proof of concept to large production

4.3 Question: What are some types of bias that are found in data science?

Answer: Here are some of the types of bias you may encounter[ix] :

1. Confirmation bias: People are less critical of Data Science that supports their prior beliefs rather than challenges their convictions.
2. Rescue bias: This bias involves selectively finding faults in an experiment that contradicts expectations.

3. ‘Time will tell’ bias: Taking time to gather more evidence should increase our confidence in a result.
4. Orientation bias: This reflects a phenomenon of experimental and recording error being in the direction that supports the hypothesis.
5. Cognitive bias: This is the tendency to make skewed decisions based on pre-existing factors rather than on the data and other hard evidence.
6. Selection bias: This is the tendency to skew your choice of data sources to those that may be most available, convenient and cost-effective for your purposes.
7. Sampling bias: This is the tendency to skew the sampling of data sets toward subgroups of the population.
8. Modelling bias: This is the tendency to skew Data Science models by starting with a biased set of assumptions about the problem.

4.4 Question: What are Popular Data Visualization Tools that Businesses Should Use?

Answer: The 10 most popular data visualization tools that businesses should use includes[x]:

1. Tableau
2. Dundas BI
3. JupyteR
4. Zoho Reports
5. Google Charts
6. Data Wrapper
7. Power BI
8. QlikView
9. Highcharts
10. Plotly

4.5 Question: What is the Best Software for Project Management of 2023?

Answer: The best software for project management of 2023[xi]:

- ClickUp: Best for Agile Development Teams
- Monday.com: Best for Startups on a Tight Budget
- Asana: Best for Collaboration Tools
- Zoho Projects: Best for Integrations
- Smartsheet: Best for Workflow Automation

- Notion: Best for Content Creators
- Airtable: Best for Data-Driven Companies
- Teamwork: Best for Client-Facing Service Providers
- Wrike: Best for Artificial Intelligence Features
- Jira: Best for Product Development Teams

---

[i] Canuma, P. (2023). Data Science Project Management [The New Guide For ML Teams]. *neptune.ai*. https://neptune.ai/blog/data-science-project-management

[ii] Project Management Institute. A Guide to the Project Management Body of Knowledge (PMBOK® Guide)—Fifth Edition (ENGLISH) . Project Management Institute. Kindle Edition.

[iii] *agility - Quick search results | Oxford English Dictionary*. (n.d.-b). https://www.oed.com/search/dictionary/?scope=Entries&q=agility

[iv] Techopedia. (2017, August 18). *What is Knowledge Discovery in Databases (KDD)? - Definition from Techopedia*. https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd

[v] Lau, C. H. (2021, December 7). 5 steps of a Data Science Project Lifecycle - towards Data Science. *Medium*. https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

[vi] *Uber's highly scalable and distributed Shuffle as a service | Uber blog*. (2022, July 7). Uber Blog. https://www.uber.com/blog/ubers-highly-scalable-and-distributed-shuffle-as-a-service/

[vii] *Ethics and Data science*. (n.d.). Data Science. https://datascience.stanford.edu/research/research-areas/ethics-and-data-science

[viii] Blackman, R. (2020, October 15). *A practical guide to building ethical AI*. Harvard Business Review. https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

[ix] Ridge, A. E. (2015, July 26). *How Do I Avoid Bias In My Data Science Work?* Guerrilla Analytics: Book, Speaking and Training. https://guerrilla-analytics.net/2015/07/26/bias-in-data-science/#:~:text=Reducing%20Bias%201%20Track%20your%20data%20sources%20and,understanding%20as%20they%20evolve%20with%20the%20project.%20

[x] Zaveria. (2022). 10 Most Popular Data Visualization Tools that Businesses Should Use. *Analytics Insight*. https://www.analyticsinsight.net/10-most-popular-data-visualization-tools-that-businesses-should-use/

[xi] Rudder, A. (2023, July 20). *10 Best Project Management Software Of 2023*. Forbes Advisor. https://www.forbes.com/advisor/business/software/best-project-management-software/