Synthesize what you have learned via the creation of a paper, which could be thought of as a book / guide for others (or yourself later, when you might forget a bit of this content) who want to manage / lead data science teams.

There is a data problem example that is my all time favorite. The WWII Statistical Research Group was charged with identifying why specific bombers had poor survivability. The sponsor had provided the group extensive data on damage records of all the B-17’s that were hit and returned to base in the past year. This statistical group poured over the data in order to determine how to improve the survivability of this aircraft but the business objectives had tradeoffs. Adding additional armor made the planes heavier and therefore slower with reduced fuel efficiency (and hence a shorter operational range). The group need to identify precisely which part of the plane to add armor to in order to improve its survivability. Data provided by the stakeholder showed holes in the planes concentrating on the wings and the fuselage. Mathematician Abraham Wald informed the group that whatever solution they identified with this data it would be incorrect. The group was looking at damage to planes that returned to the base. These planes survived and therefore the types of damage these planes were seeing wasn’t catastrophic. The group needed to look where the damage wasn’t. That was likely where the planes that didn’t return were being shot and that is why they didn’t return. What this group experienced and what this vignette represents is a disconnect between business problem and data understanding and how a data science team can bias results with their perspectives. So apparently in data science problems, seek to be the individual who looks where there aren’t any holes.

The outline of the guide should include, at a minimum:

1. Why managing a data science project is important (~1 page)

Because not everyone believes, I think it is first necessary to articulate why it is necessary to have a codified set of processes for any project first. The pace of business and operations within the industry that I am a part of is greatly expanding on a regular basis. Complexity is growing on a continuous and incremental basis. Without structure, the complexity can transition to chaos. But with a structure, a simple set of rules and guidelines, the growing complexity can be manageable. Using the notion of taking structured and disciplined actions and applying them to complex problems makes them less complex and instead, a series of simple and digestible tasks. Without structure, chaos and inefficiencies are more likely to have negative impacts on productivity, output, organizational climate, morale and impede potential. In essence, a defined process leads to greater team efficiency. With steps and processes that are repeatable and structured, they support long term execution with the perception of higher value to the organization.

With the acknowledgement that a structure is necessary for project management, it is possible to analyze the nuances of why this argument is especially pertinent for data science projects. There is evidence that people believe that data science projects don’t produce value or deliver outcomes - “Through 2022, only 20% of analytic

insights will deliver business outcomes” (Gartner, 2019). When analyzing this through the lens that more than half of data science organizations do not have or follow established and standardized processes (Only 48% of data science organizations have established standardized processes – Corinium Execute survey (2020)), it gives credence to the notion presented above that agnostic to the industry, structure can make the complex manageable. Defined processes help to mitigate biases and variabilities that impact results and valid results endorse the work of the data science team.

Not only does the complexity of data science belie the need for a structured process but its popularity does as well. With the increasing realization of the potential of data science comes the increased request for support without a corresponding bump in available resources. Competing requirements with limited requirements, complex data sets with high levels of unknowns can lead stakeholders and organizational leadership to believe that data science is not effectively producing. Instead, it is a matter of structure, processes and prioritization to streamline estimation of value of a project, estimation of the effort to complete that project and then assessing the risk of completing those in lieu of other projects. By estimating the effort and value of projects as well as assessing risk in a structured manner, the data science team can ensure work is scoped in line with the business objectives and that the data collection and analysis is focused towards actionable insights to benefit the stakeholders and the organization.

1. How is a data science project different from other IT projects? How might it be similar? (~1 page).

First, similarities. From the most basic point, both data science and IT are similar fields, built on a foundation of science, computers and mathematics. This basic similarity is magnified as data science can no longer be executed outside of a digital IT environment. Both data science and IT also rely on a nested framework that knits together the stakeholder needs with the capability of the teams products or deliverables. This is executed best with a unifying framework that makes those results reliable and reproduceable. It is this notion of reliable and reproducible results that truly nest these two fields together and put them into the realm of science. With a framework of understanding, testing, validation and production, both data science and IT are able to produce reliable and reproducible deliverables to address a business objective.

In many ways, the differences are more glaring than the similarities. Data science projects just have different profile attributes. The requirements are generally ambiguous and difficult to estimate and for that reason, very difficult to manage timelines and expectations for an organization or stakeholders. Very often, data or data problems are more exploratory, with an effort to simply find what the data and the models indicate. This exploratory nature makes it much harder to scope the big picture as well as all the sub tasks associated with it. It is not uncommon to approach a data science project thinking strongly that the direction or trajectory of the work will go one way and the data will end up taking you in a completely different direction. This may be due to data accuracy (or inaccuracy), data availability (or lack thereof) or data simply presenting different outcomes or analytic insights than expected. In a traditional IT project, there may be some risk associated with unexpected events or challenges due possibly to time or resources but for the most part, the components of the project can be laid out and estimated ahead of time. There is a relatively set path on what will happen and how they will cycle. That foreknowledge is just not feasible in a data science project.

The data challenges mentioned above can lead to some ambiguous and imprecise outcomes. As we are humans, there is more room for bias which can complicate the process and invalidate results. If we have an imprecise understanding of our data needs, we will be unsure of when and if we collect all our needed data. These two present significant challenges to validating result and without valid results, there are no actionable insights.

Data science projects just has a different profile attributes. Data science projects are vague at best and are far less likely to follow the path that they are expected.

1. Provide a description of the types of frameworks that could be used on data science projects (~1 page). Provide an explanation of the different key frameworks that might be used - using the type of framework structure defined above (~1/2 page per framework).

The first item may need to be addressed is the question of a data science life cycle framework. This answers the most basic question “what you do to accomplish the data science project” and may really be more important to the external organization or stakeholders. Generally, these types of lifecycle frameworks are broken into the steps of a project and use a workflow to establish a common vocabulary and describe what the team does during a project.

Common examples of this type of framework include Cross Industry Standard Process for Data Science (CRISP\_DM), Obtain, Scrub, Explore, Model and iNterpret (OSEMN), Harvard’s Phased-Based Workflow, Domino’s Phased-based Workflow, Uber’s Phased-Based Workflow and Team Data Science Process (TDSP). The most widely used process is CRISP-DM which relies on six primary iterative phases that enable projects to move back and forth between phases as needed or cycle through all phases repeatedly. Each phase has pre-defined tasks and a set of deliverables though roles or a meeting structure are not identified. At the center of the lifecycle is the data, highlighting the role of data throughout the process and its importance throughout all phases of the cycle. Below is a snap shot of the phases with the fluid iteration between said phases.



***Figure 1: CRISP-DM Lifecycle Framework***

Phases generally begin with Business Understanding (as in CRISP-DM) or something similar for the other cycles, to provide the initial focus to help ensure alignment of business needs and data analysis and prevent the situation of a data science team moving into a project without a clear understanding of its objectives. Admittedly, CRISP-DM has a more deliberate process for this synergy than the others while the uber and TDSP models does seek to understand customer needs and problems. The next piece of this shared understanding is the data side of the problem, enabling the team to identify the state of available data, assess it and whether or not it will support the business objectives. CRISP-DM, Domino, TDSP all explore the data but CRISP-DM is the one that focuses on a concerted linkage directly and deliberately between the business problem and the data understanding. The next phase of a lifecycle takes on data preparation which corresponds to the laborious task of data munging/data cleaning, which can take a disproportionately large percentage of the time. Following this, modeling is the next phase which incorporates the data science techniques intended to bridge the data from information to insight. This insight finally occurs in the evaluation phase where the goal is to assess the model as it related to the business objectives. The final phase is deployment or delivery and monitoring which incorporates the longer-term application of getting products to the customer and maintenance and monitory of the project and codifying the results into a result. Key to the CRISP-DM and all the other framework is the ability to transition between phases to some degree, sometimes as needed and sometimes as a full iteration of the process.

While CRISP-DM is the most prevalent, the other listed and discussed lifecycle frameworks rely on a similar phased framework that follow a similar pattern of: you figure out what you’re trying to answer, gather the data, explore the data, model the data, produce some results though this were differences between them. The table below summarizes some of the key points of the subsequent frameworks but there is a reason that CRISP-DM is the most widely known: it is visually easy to understand and therefore more prevalent and succinctly encapsulates the phases. This gets back to an issue specified earlier, if something appears complex then individuals and teams are less inspired to adopt it.



***Table 1: Lifecycle Framework Quick Comparison***

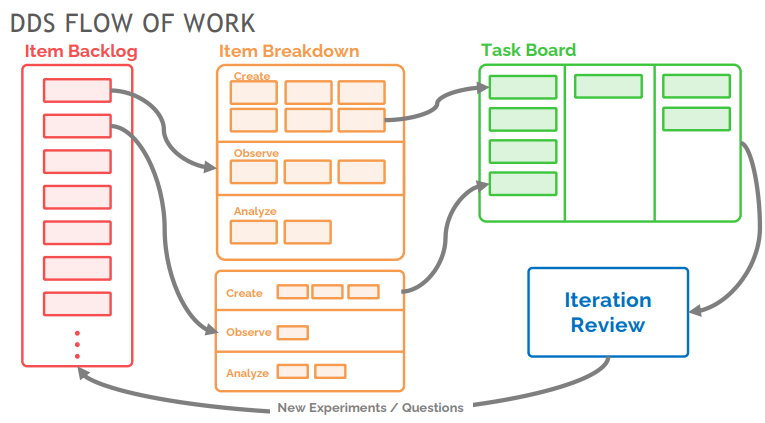
After deciding upon a lifecycle framework, the decision transitions to deciding upon what agile coordination or collaboration framework to use. This answers the most basic question of “how do you execute the data science project” and is really important to the internal team. The collaboration framework emphasizes guidelines such as prioritization, distribution of labor, solicitation of feedback, and a focused and standardized methodology to reflect and improve the process. This type of framework allows the team to manage the project and all its components.

The basis of these frameworks are agility. For data science teams, this could mean breaking the herculean data tasks into bite sized pieces and then evaluating progress and getting stakeholder input to make sure that the progress is in line with expectations, or it could mean producing a minimally viable product (MVP) and then soliciting feedback. The three most common collaboration frameworks for data science are Scrum, Kanban and Data Driven Scrum (DDS).

Scrum is extraordinarily popular and commonly used but presents some significant challenges for a data science project. Scrums rely on time-based sprints which define and execute equal bite-sized pieces of a project call sprints. These sprints last between one and four weeks and break work into a burst to execute within a pre-determined amount of time. There is no agility within or between sprints of a project though there can be flexibility between projects as lessons learned can be applied to make changes as needed. In scrum, roles as well as meetings (ceremonies) and artifacts (deliverables) are pre-defined. The roles, meetings and deliverables are all intended to streamline and optimize all possible efficiencies in the process. The primary challenge with data science is the difficulty in time estimation and time completion of data science tasks.

Kanban is an implementation of lean and can best be described as a manner of visualizing and organizing tasks of a project without the time-based iteration of Scrum. It relies on a board with a series of columns and tasks make their way through the columns to completion. An oversimplified Kanban board may have columns for “to do”, “working” and “done” and as the tasks are initiated, they move from “to do” into “working” and then move again into “done” upon completion. Another key to Kanban is the WIP or work in progress limits which limit the number of items that can reside within each column of a Kanban board. This limit prevents a team from having too many items in the works and becoming spread too thin. While Kanban is not a project lifecycle, it can work hand in hand with frameworks like CRISP-DM to manage task progress. Kanban does not enforce timelines for the completions of tasks, have team roles or identify meetings to execute. The visual representation has been known to improve communications, enable agility and make teams more adaptable.

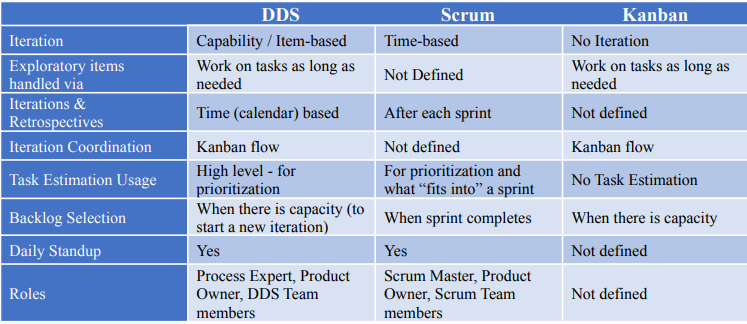
The third collaboration framework is the DDS or data driven scrum, which takes the concepts of scrum and makes them more applicable for a data science project. The first key difference between DDS and scrum is that the iterations are capability based as opposed to time based so sprints, or with DDS iterations, are based on a chunk of work to be completed as opposed to specific period of time. DDS relies on four main principles: agile is really a sequence of iterations, each iteration should focus on an idea or task and how to evaluate the results of that task, the basis for an iteration should include the full cycle of that task from initial idea to results, and finally the end of the iteration is marked by some empirical output and not just a specific amount of time elapsed. Further, DDS relies on a create, observe, analyze cycle that applies to the analytic properties to each phase of the lifecycle framework to ensure that tasks are codified to each element of the task. The flow of work in DDS follows a prioritized item backlog that is used to create an item breakdown of all the sub or implied tasks of that one task. The flow of work is shown below in figure 2.



***Figure 2: Data Driven Scrum Flow of Work***

They are visualized on a task board such as a Kanban board and then reviewed upon completion before cycling back to the next item in the item backlog. This frame has shown proof to refine delivery expectation as well improves communications and provides a more focused review at the end of each iteration.

Below is a summary table of the three frameworks together including the manner of iteration, speed or cycle of iterations, estimation procedure and meeting and role specifics.



***Table 2: Collaboration Framework Comparison***

1. Explain how to select a framework - hints on how / why to choose the different alternatives (~2 pages)

There is a very insightful TeD talk about leadership on the golden circle that talks about how to effectively get your point across in business and life. The talk can best be summarized that most people explain to people first by saying what they want, and then maybe how they can help and then possibly why it’s beneficial to help them. But the speaker of the TeD talk postulates that the great leaders and influencers in the world throughout history led with the why. Why should you make change? Why should you support a particular project? Once they’ve conveyed the why, they then discuss the how. How will employing this new change the way you’re doing business? Or how is this new way different and innovative from the rest? And finally, what that change or project is. I think the direction of that information flow is critical to not only understanding how to choose a framework but also how to gain buy-in on that framework.

We’ve discussed why a well-defined process, such as a standardized framework, can benefit a data science project: it increases team efficiency, produces valid results as well as actionable insights. As defined by the CRISP-DM approach, business understanding includes determining the business approach, assessing the situation, determining the data mining goals and producing the project plan. These tasks would greatly reduce the belief that a data science team is not productive and it would additionally alleviate the data science teams misunderstanding of expectations. Another problem that a lifecycle could mitigate would be data understanding including data collection, data description, data exploration and verification of data quality. This step would both help the data science team and the rest of the organization. It would help the data science team understand the scope of the data available to them, which is particularly important in a resource constrained environment.

So that addresses more of the why (and admittedly some of the what). But how should the framework be determined. I believe that many of the frameworks seek not to necessarily provide an end-to-end solution but instead a set of tools (or an integrated set of tools) to handle all the steps of a traditional workflow. At an individual and organizational level, I believe most people pick and choose ideas that support what they’re already doing, what they want to achieve, or what works or already exists within their organization.

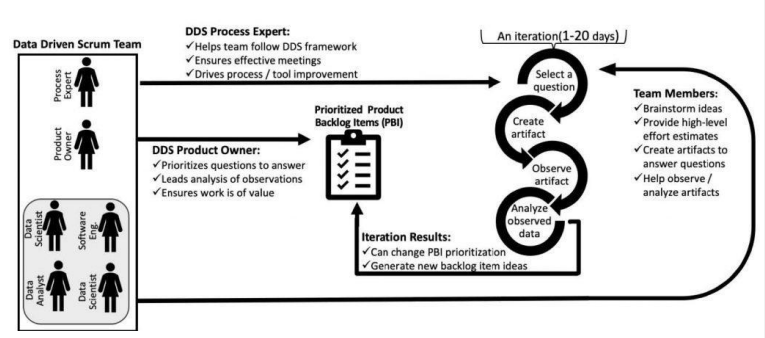
In addition to what an organization is already comfortable with, a data science team needs to take into consideration what their weakness are and that has the potential to shape their framework choices. This necessitates an unvarnished introspection and realistic assessment of individual and team strengths and weaknesses. It needs to address key questions/introspections that will help assess where your organization and team mindset is. What is the current level and interest in participation by stakeholders? If there is none, then Kanban may be an option as it lacks a defined meeting structure. What is the experience level of the team? If the team is all relatively new and inexperienced, then Scrum and DDS may not be the best options as they rely on discipline initiative and strong experimental cultures. Is the data science team working within a strong organizational structure of business? If that’s the case, CRISP-DM is probably going to help bridge the understanding between business and data with well known and well understood standardized processes. I believe the conventionality of frameworks like CRISP-DM help to give structure to the data science team who appear to be iterating “exploratory” projects without the buy-in from key organizational leaders. With a shared understanding and lexicon moving forward, it would subsequently provide legitimacy to the data science in the eyes of company’s business leaders. By using business related goals in a language that stakeholders understand and a stepped process that everyone in the organization would likely be familiar with, the they would be more likely to buy into and even participate in the concept. What is your current access to the data? Is there a lot? Without knowing precisely what the data is and having a precise understanding of it, Scrum will be very difficult due to the challenge estimating time for task completion. In this case, Kanban or DDS may be more appropriate. How open is the team to trying something new? If the answer is yes, then maybe one of the alternate lifecycle frameworks such as Uber or Domino are the answer.

There may be a litany of influences that can play a role in framework selection beyond self-assessment, including customer feedback, the range and/or consistency of projects, organizational culture, business objectives and personalities. There is an alternate to the binary “this or that” approach that seeks to find an optimized fit to most precisely meet the needs of a data science team: a franken approach taking the parts of each that work and then put them all together. One thing that stands out about the different frameworks is the potential for personalization of models to feed preferences, team dynamics, organizational structure and probably institutional (within the data science community) influence. In addition to this, and probably because of this amalgamation of processes, adherence to a single or preferred model/life cycle/approach may not really be feasible or achievable for an organization. Instead, customizing and taking an approach to gather the best pieces that work for your team or your organization or even your project at the moment may optimize the functionality of the team and the work they are doing. Through self-assessment and then careful selection of elements of each framework that fits a part of the teams needs, it may be possible to find a solution that fits that precise teams needs at that precise time. It may not have a name but the solution may work. While an approach for a business or engineering or IT project may not be appropriate for a data science project, the benefit of a structured process has already been established: efficiency, result validity, actionable insights.

So, future you, what framework should you use? For this we will presume a hypothetical of similar projects to what currently exists but with an established framework moving forward. There is a high degree of organizational culture and they (your stakeholders and your team) are highly resistant to change. To that end, I believe the phased structure with iterative nature of CRISP-DM would be tolerable to most of the organization. Stakeholder integration early would be challenging but the structure of CRISP-DM would serve as a forcing function for this action which would refine the business objectives. Your stakeholders only want to read one final report though, they are not interested in iterating products to solicit their feedback. For that reason, I believe integrating the Uber model prototyping loop to iterate the data acquisition or reacquisition, prep and model portion of the process in a separate loop would benefit the project execution. This will give the team time and a structured process to iterate on the data to make sure it fits the business objective(s) before moving into the evaluate the results phase of the lifecycle. The data sets vary and the business objectives vary from project to project therefore Scrum will not work. But the stakeholders still expect some sort of output in a reasonable amount of time or at least evidence of progress being made towards their objective, therefore I don’t believe Kanban would work in this instance. I believe DDS with a Kanban visualization will optimize the functionality of a capability-based task execution with the aid of the lean principles that Kanban supports.

1. For the DDS framework, answer the following questions (1/2 page to 1 page):
   * 1. On a scale of 1 to 5 (5 is highest) How likely are you to suggest using DDS: 5
     2. Do you think DDS is better or worse than Scrum (for data science project). Better. Way Better.
     3. Explain your thoughts in how you answered these two questions.

Data driven scrum offers a scrum-like framework, so it’s easy to understand and accessible, but with some major, and very significant, differences. Instead of fixed time-based sprints, DDS has variable length capacity focused iterations that can overlap. Each of these iterations focuses on a task that can create, observe and analyze something in support of the overreaching project. The progress of the tasks and items in the backlog follow a more Kanban like continuous approach than the finite nature of scrum. These capacity-based iterations play to the flexibility that data really needs, the inflexibility and deadlines of scrum belies the challenges of data. It is also important to remember that not all tasks within data take the same amount of time so a time-based sprint may result in tasks that are completed before the end of the sprint and inefficiencies of time or, more likely, ones that can’t be completed in time. DDS also has the flexibility to fit into a number of lifecycle frameworks beyond just CRISP-DM so it has flexibility within and throughout organizations. Adopting DDS can still present challenges within an organization, especially one that employs scrum. As the DDS iterations don’t follow the traditional time basis with deliverables at the end of each, these iterations with not fit into an organization where scrum is executing another part of a project. The below figure shows the conceptual flow of a DDS project, highlighting some of the similarities in roles and meetings with scrum but also highlighting the uniqueness of the iteration cycle with its create, observe and analyze emphasis and team/capabilities based focus.



***Figure 3: Data Driven Scrum Conceptual Flow***

1. Repeat this analysis (i.e., answer the three bullets above) for:
   1. Kanban
      1. On a scale of 1 to 5 (5 is highest) How likely are you to suggest using Kanban: 5
      2. Do you think Kanban is better or worse than Scrum and DDS (for data science project): Better than Scrum worse than DDS

* + 1. Explain your thoughts in how you answered these two questions.

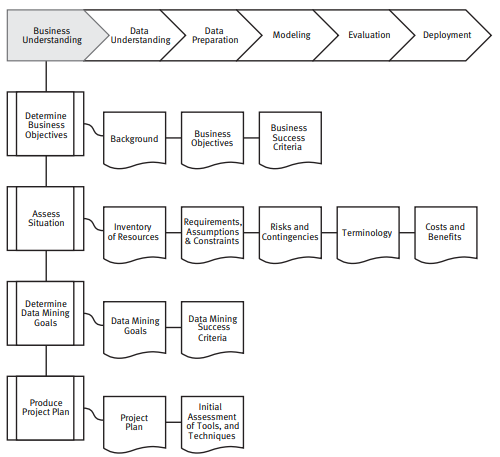
I believe that since Kanban can and frequently is integrated into Scrum and DDS it is slightly difficult to evaluate it as better or worse. I believe that, when DDS incorporates a Kanban style visualization, it is superior to Kanban. A huge limitation of Kanban is that it fails to manage any sort of timeline, have any pre-specified touchpoints or have any specified roles. This makes it difficult to enforce consistency within a project. When incorporated into a larger framework like DDS, however, I believe Kanban provides a powerful visualization and task tracking mechanism that will help ensure the team stays on track. Studies have found almost all (95%) cognitive information is perceived through sight and that most humans (between 60% and 65%) are considered visual learners (<https://www.pmi.org/learning/library/visual-project-management-visual-elements-9862>). This makes the need for a task visualization extremely important. I think Kanban is better than Scrum because I think the scrum format is just incompatible with most data science projects. Even when scrum incorporates the visual benefit of Kanban, its time constraints are too restrictive to optimize the benefits of the WIP limits and ordered columns of Kanban.

* 1. Scrum
     1. On a scale of 1 to 5 (5 is highest) How likely are you to suggest using Scrum: 1
     2. Do you think Scrum is better or worse than DDS and Kanban (for data science project). Scrum is worse than DDS and Kanban
     3. Explain your thoughts in how you answered these two questions.

While Scrum is probably the most well known and widely employed of the agile approaches, it is unpopular (and some would say incompatible with) data science project teams. This aversion is specifically attributable to the time-based sprints in Scrum tied to a precise deliverable which are exceptionally difficult to execute a data science circumstance. Data can be vague and sometimes messy and doesn’t like fitting into a box, time box or sometimes any other box. Therefore, applying estimates, time-constraints, or incremental deliverables may not work as planned. Also, as mentioned, not all data science work can or does take the same increment of time. This time constraint further hampers work by either not empowering the data team to move on with additional work in accordance with the prioritization, or not finishing the work and potentially giving up on that sprint or task. DDS on the other hand works with a capabilities based iteration which identifies a task to complete and has each iteration set to complete a task with a product/deliverable or component therein complete at the end.

* 1. CRISP-DM
     1. On a scale of 1 to 5 (5 is highest) How likely are you to suggest using CRISP-DM: 4
     2. Do you think CRISP-DM is better or worse than Uber (for data science project). That is a tough one but I think generally, CRISP-DM offers better overall execution benefits than Uber.
     3. Explain your thoughts in how you answered these two questions.

CRISP-DM encourages the data science team of a project to understand the project in terms of the business goals – getting after the critical issue of the shared lexicon, and then data understanding in support of those business goals. That is a huge benefit of the CRISP-DM over the UBER process as it keeps the project in line with what the predominance of stakeholders with understand, outcomes driven. While the Uber model has its “define” phase where it seeks to understand the customer needs and define the MVP, its emphasis is not on the concerted alignment of business needs and data work that CRISP-DM is. The figure 4 below shows the level of detail that goes into the business understanding phase, ensuring a consensus understanding of expectations moving forward. In work, it is sometimes necessary to serve as a translator between the business side and the data science side, a role that the project manager or product owner may need to fill. The data science team doesn’t always have the context to understand the problem and business goals from the business perspective and the stakeholders/customer doesn’t always understand what the data can and can’t do.  That can lead to data just for data sake or models that are beautiful and not useful to anyone. Cracking the nut of actionable insights is difficult and I believe that applying the construct of a CRISP-DM cycle with focused business goals plus the iterative prototyping cycle to account for the somewhat discovery nature of our work, might be the optimal fit.



***Figure 4: CRISP-DM Business Understanding Phase Breakout***

* 1. Uber
     1. On a scale of 1 to 5 (5 is highest) How likely are you to suggest using Uber: 3
     2. Do you think Uber is better or worse than CRISP-DM (for data science project). I think Uber is slightly worse than CRISP-DM
     3. Explain your thoughts in how you answered these two questions.:

One significant positive on the Uber model stook out for me. There is something about that cycle within the cycle, the iterative prototype cycle, that speaks to the necessity of some industries to really iterate with the data prior to producing results. Too often, an organization or a data science team may not aware of what the data is going to produce. Too often, our initial question or business problem doesn’t necessarily produce anything or produces something unexpected that may not be immediately meaningful to the stakeholders, or be beyond the scope of the project.  By adding an iterative prototype cycle, a data science team could learn or fail earlier and thereby streamline the overall process. But ultimately it is supplanted by CRISP-DMs emphasis on business understanding as an alignment with data understanding, consistent structure phases and shared lexicon across industries and within an organization.

Discuss potential ethical situations that might arise in a data science project, and how a team's process could reduce the risk of the project doing something that is not ethical (~1 to 2 pages)

Ethics can be a big deal within the data science field since there is so much human influence in the process. Ethics has two sides of it within a data science project. The first is access to information and the second is bias.

Information access comes along with data science. Data science teams have access to a myriad of information that spans a wide degree of sensitivities. This information may include sensitive or even classified information. The first control measure for this type of situation is to ensure that access to such information is only granted to those who have a specific need to access the data. This access should come with training on how to properly handle this type of data and the ethical implications with it. And what are those ethical implications - future you asks? Just because an individual has access to specific information does not mean that they should view it or view certain parts of it. The clearest example of this is medical information. Most medical information is tied back to a specific person but records include some very sensitive details about a person. Just because the records are linked to a person does not mean the analyst has any reason to look up records by person. Data projects and collection should be focused to address the business objective not the personal curiosities of the team members. As sensitive and identifiable information is not pertinent to most data science projects, it should be sanitized at the highest level possible to preclude the option of unethical behaviors.

In many ways, this element of ethics is clearer and more straight forward than bias. Bias is a deviation from expectation in the data or outcome. Humans touch the data at all points and therefore, have an opportunity to inject their bias, either consciously or unconsciously, into the data throughout the process. Ethics is the attempt to remove that bias from the process wherever it exists. Data can only be as precise and free of bias as the individuals who collect, process, observe and analyze it. First and very important to remember is that as humans, we all have bias. We are shaped by our experiences and those experiences become the lens through which we view virtually all things. That lens ends up influencing how we view everything. It is impossible to divest ourselves from our experiences and therefore impossible to divest ourselves entirely from the lens through which we view the world. What we can do is to become aware and cognizant of that lens and look for the implications of it. We can also have multiple people view the same data science project through their own lens. Multiple viewpoints will present varied perspectives and increase the likelihood of identifying bias.

To understand the potential impact of ethics and bias on data science, it is necessary to identify how it could be introduced into each phase of a data science project. Using the CRISP-DM lifecycle as the methodology for this evaluation gives structure to this evaluation.

Bias in business understanding – When determining business objectives, it may be necessary determine if the business objectives are biased or if the data mining goal is biased by failing to incorporate key points of the data or key elements of analysis. Are there laws or regulations that might be applicable to the project? Using a purely hypothetical example of identifying customer churn in an airline, is the airline evaluating churn by geographic location but failing to identify some of the geographic locations that are being hit with a specific economic impediment? Are there other elements that will bias the ability to thoroughly answer the business understanding?

Bias in data understanding – This phase is a critical area of possible bias. Data collection can be influenced by bias either by limiting (accidentally or constructively) the scope of the data or through the means of collection. Going back to the hypothetical case, if data is only collected from individuals who use the company app, the information produced will be biased as it excludes all other forms of company interaction. Is it possible that the legal rights of the people whose data is collected is or can be infringed? Is their anonymity and privacy assured? If we collected the information, can we be assured that the collection methods were ethical? If we used someone else’s data, can we be assured that it was ethically attained and is authorized for available use? All of this should feed into the data understanding task of verifying data quality.

Bias in data preparation – For data preparation, bias can make its way into such innocuous tasks as data selection and data cleansing, among others. Is the manner of data sampling bias towards a specific variable? Is there a mechanism to make sure the sampling is truly reflective of the total data population? Reference cleaning the data, if some of the participants in a survey leave age blank and we opt to exclude those records, are we biasing against individuals who just chose not to share that one piece of information?

Bias in modeling – Time to build the models, we’re in the clear right? Wrong, but it’s important to remember that models don’t have bias, data and humans have bias. Since humans develop the models there is a chance that individual bias can make its way into model development and test execution. Has that potential modeler bias been identified, and if appropriate, mitigated? How transparent does a model need to be and are the steps to achieve that transparency recorded?

Bias in data and model evaluation – We have ensured that the business objective is unbiased, the data was collected and maintained appropriately and is representative of the population as whole, we have ensured our sample is truly representative and cleaned the data in an unbiased manner and made transparent models devoid of bias. Time to turn our hard work into actionable insights – yet another place for bias to creep into because again, humans are the ones doing the analysis or interpreting the results. It is necessary to ask if there are possible mis-interpretations of the results that could skew interpretations in the final reports. If we go back to our hypothetical and we see results indicating poorer customer responses and greater churn coming out of Minnesota and we attribute it to the customers living in Minnesota (maybe we know people from there and that customer profile fits). That may preclude us from assessing other variables that impact customer satisfaction. Perhaps the flight crew out of Minnesota is less enthusiastic and that is causing greater customer churn, not the personality of our fine people from Minnesota.

Bias in data deployment – We’ve made it to the finish line but need to ensure that the final report and other deliverables are still free of bias, that our maintenance plan is not skewed towards a specific group or area or other means of bias and monitor the project to ensure that bias does not creep into it.

1. Create a FAQ to address what you think would be common question from new DS project managers (~1 to 2 pages)

A top ten (10) to address new DS manager common questions with “a”, but not the, answer. It is important to note however that each organization, leader, project and data science team is unique and perfect-out-of-the-box solutions will likely not work for everyone.

1. How do you encourage stakeholder participation?

Stakeholders are key to any project therefore encouraging their participation necessitates a multipronged approach. First, highlight the benefit of playing an active role in the process. Integrate them early and often: start with an all-hands discussion of the business understanding and then get them integrated into the events/meetings of DDS or similar framework. Actively request and seek their feedback in meetings and after increments. Involve them in the prioritization review process. Get feedback, involve them in the process. Make sure that the data science project is in line with their business goals and objective and they will begin to feel a greater sense of interest and stake in the process. And finally, the best way to ensure participation is to meet their expectation and answer the business objectives. That will result in even greater participation in future projects.

1. How do you ensure that everyone has a shared understanding of the business problem?

Each project should start up front with a realistic discussion of the business problem and business understanding. Following up this frank discussion with a discussion on the data understanding (and describing and exploring data options) and then iterating back to business understanding to make sure that the available resources can in fact accomplish the business goal(s) would further mitigate a lack of understanding on both sides.

1. What will be the biggest challenge in synergizing data to answer a business problem?

Is there data to support the business problem and can we get to it? This should start with the business understanding phase of the lifecycle with a thorough determination of business objective and data mining goals. Developing these two items together will seek to identify how they will work together in a nested manner towards the business goal. Then iterating to data understanding to explore and describe the data will give the project manager or product owner a chance to see if they will be able to meet the specified data mining goals – specifically, will the data support the business problem and can we get to that data.

1. How do you assess effort and value?

Assessing effort for data science is best when thought of broadly. How much effort does the data science team think will be required for each potential item? A small amount? Maybe it’s a medium amount. Whatever the measurement, assessing that effort broadly in terms of time and resource output will enable a data science team and specifically the project manager/product owner to scope the larger effort for the project. For value, ask how each potential item will impact the business goal. Combining the value and the effort of each task or item can help a data science manager inform stakeholders of prioritization choices.

1. How do you prioritize?

Data science is a growth industry and there doesn’t appear to be a lack of interest in data science provided insights or data science projects. What there are, however, is a finite number of resources. Managers will need to take in a number of data point such as the business objectives, metrics to define success, expected deliverables, value and effort estimations, uncertainty, dependencies and synergies. The output of these data points will be prioritization: the highest value tasks or deliverables relative to the effort needed to complete them and the uncertainty surrounding their execution. This will produce a prioritization and help project managers refine what projects to focus their resources on.

1. When do you iterate?

Until you have something good enough for a minimally viable product. This will allow you to proceed further through the lifecycle and present that MVP to the stakeholder and see if it meets of business objective. This employs the concept of failing early and failing often but iterating. Once you’ve confirmed that the MVP does in fact fit what the stakeholder/customer is looking for then you iterate again and make it better.

1. What does “done” look like?

Unfortunately, “done” is probably going to look different for different tasks and different projects but it must be codified up front. This may include higher-level metrics to assess if strategic goals that are tied to the organizational metrics have been met as well as lower-level metrics associated with each deliverable. These should be specified before the first iteration and should include stakeholder input so that all individuals have a clear and consistent understanding of the projects planned outputs/goals and when they’ve been met.

1. Is there a clear way to measure if stakeholders/customers are satisfied?

For data scientists, there is always a way to measure anything. Whether it is a subjective or an objective metric is another story. Stakeholder satisfaction falls into that subjective category but can be measured nonetheless. This could be extrapolated from the perception by the project manager of the stakeholder’s satisfaction with the final product or could be asked explicitly. Another way to measure customer satisfaction is through deliverables. The number and value of deliverables provided to the stakeholders/customers can be used to assess the degree to with the data science team is supporting the business objectives and satisfying the customer needs.

1. What are ways to build on and improve processes?

In a word, communication. Fortunately, the collaboration frameworks set the conditions to enable that communication, optimally (I believe) by the DDS framework. The iteration review presents an opportunity to review the state of progress and specifically “done” tasks. At the completion of each iteration, there should be a meeting with the stakeholders to ensure that iteration outputs have met expectations and support larger project objectives. Another opportunity to build and improve the process is iteration planning and prioritization meetings to help identify prioritize and plan subsequent iterations. Focusing more internally, retrospectives discuss process improvement and focus on members of the data science team.

1. When is it time to cut your losses and stop work on a project?

If the data is not available to support the objective and cannot become available, then it may be time. If the stakeholder has lost interest and moved to a different project, then it may be time. If organizational leadership has reprioritized in such a way that work on this project is no longer feasible, it may be time. If additional analysis is determined to add little or no value to the project, yep, it may be time. Resources (time, budget or personnel) are no longer available to support. Or possibly, the customer doesn’t know what they really want or how they would use the results, it is most likely time then.

Best of luck future you!