Personal Manifesto

By: Adam Bakopolus

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

To further engage with the data science community and gain practice-based insights, I will be reading **Caitlin Smallwood**'s interview in the "Data Scientists at Work" book by Sebastian Gutierrez. At the time of the interview, which is already readily collected and available through this course textbook, Smallwood was the Vice President of Science and Algorithms at Netflix. I was particularly interested in this interview as Smallwood, as part of her work at Netflix, would be heavily involved in the implementation and enhancement of their show/movie recommender algorithms. I'm very interested in learning more about the statistical models and approach that went into generating this process not only just as a user of Netflix but also as a current data analyst that would be able to leverage these procedures in my own work. As I work within the healthcare field, I see a clear and immediate need for solutions that allow decision makers and practitioners to clearly identify various populations in need of intervention or additional care to improve outcomes. Understanding the statistical models used and how the solution was rolled out would be invaluable to better understand how this can be leveraged across different fields.

Reading Responses

Readings:

- **Chapter 2 - Business Problems and Data Science Solutions**

Chapter 2's reading made the important note that "iteration is the rule rather than the exception" when it comes to the data mining process. While there may be an initial plan that is reviewed and signed off on for a particular data project for either school or work, the ability to be flexible and change the approach when roadblocks or new findings are discovered is immensely valuable. There will very rarely be a linear path from start to finish for a project, so the willingness to re-tool approaches and rethink strategies will be a great skill to obtain to allow for the end work to be both accurate and meaningful.

This reading also did a really great job of highlighting the key difference between "mining the data to find patterns and build models" and "using the results of data mining". A very effective data mining process and an extremely meaningful finding can all be rendered useless due to a data scientist's poor presentation or inability to relay insights in a digestible way to the audience. A data scientist must be able to effectively bridge the gap between both technical and non-technical audiences, and discuss the process and findings in a way that is both understandable and actionable for the audience to allow for the appropriate next step of utilizing the finding to begin.

- **Chris Wiggins interview**

I really liked Wiggins' note that "the key [when beginning a data problem] is usually to just keep asking, 'So what?' You've predicted something to this accuracy? So what?". This is a very important thought to keep in mind as a data scientist as there is typically a myriad of different paths that can be taken when grappling with a certain problem or question. Before delving in too deeply into the weeds programming-wise or research-wise, it's important to also remain cognisant of the high-level goal of a project. Busy work is not always productive if the discovered insights are meaningless or not actionable.

In Wiggins' interview as well I really liked his discussion around a thought attributed to Einstein that "not everything that can be counted counts, and not everything that counts can be counted." In my own day to day work in the healthcare space, I'm dealing with big data from medical and pharmacy claims and enrollment data that can at times be stored in tables with billions of rows. The ability to reduce data is critical as a data scientist to gather meaningful insights from the often overwhelming datasets that are available for certain projects, and it is very important to know either upfront or early on in the process what the key fields are to quickly aggregate and reduce the data source to a more reasonable level.

- **Erin Shellman interview**

Shellman had an interesting remark that “the most interesting types of data are those collected for one purpose and used for another.” This ties in really well with one of our maxims discussed in the lectures that the original formulation is rarely the right formulation. A framework that was first implemented for one question but was better suited for another is definitely a common case in my own day to day work, but this is a valuable note that I agree with. The ability to be flexible and not abandon work even if it doesn’t quite fit the original purpose is valuable, and the creativity to find alternative uses for these types of solutions is a desirable skill for a data scientist.

Similarly to Wiggins, Shellman also discussed the importance of considering the analysis’ end goal itself prior to really getting into the weeds development-wise as she “typically start[s] from the finish line. Assum[ing] that you’ve built the thing you’re considering, then ask ‘so what?’” Do customers want your product and can they use it? Not only is this consistent with the Wiggins interview, but it also ties in nicely with our own lecture discussions. A successful data analysis when grappling with a problem will continuously ask “why is that”. This approach ensures that the time spent developing a model will not be wasted as there is a true need or role for the finished product.

- **Jake Porway interview**

Porway made a great note that the next great discovery that is in line with the microscope (viewing of the small) and telescope (viewing of the large and far away) is the “macroscope that lets us look at the complicated patterns of society and nature and the ways that they interact.” With the many fields, like healthcare, for example, constantly changing and becoming more and more complex, the goal of a good data scientist is to be the tool for a project or organization that is finally able to evaluate the wide range of data and knowledge available in a meaningful way that will unlock solutions to the increasingly complex problems facing domains today.

Additionally, Porway also makes an important note that “we shouldn’t underestimate how much little things like that can transform an organization”. This again ties back into the importance of flexibility as a data scientist. While there may be an overarching end goal that a project is moving towards, little findings and insights along the way may also have an impact and time should be allocated to thoroughness and ensuring that even non-anticipated findings can have a meaningful impact.

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 1, Problem Formulation

1. **How to conduct an inquiry in my application domain that leads to a good problem formulation**

Yes, I already have this capability in the healthcare domain. In my day to day work, medical, pharmacy, and dental claims and insurance enrollment data are readily available and serve as a huge database to pull and glean insights from. With all of this data available, inquiries that investigate the quality and cost of care for certain portions of the population can easily be completed through a lot of the problem formulation types discussed in this lecture.

2. **A repertoire of problem types**

Yes, I already have this capability, as well, as my work heavily involves data reduction, which, in turn, allows for many of the other problem types and approaches discussed in this week's lectures to be more easily completed as insights can more easily be gathered from simplified datasets. However, I am very excited to continue building my skills through the MADS program, as Data Manipulation and Math Methods for Data Science have already greatly improved my statistical knowledge and python skills. I'm excited for future courses that continue to build this knowledge set to allow for improved skills and machine learning techniques to enhance a lot of the problems I encounter on a day-to-day basis, like regression analysis and similarity matching, for example.

3. **How to map problems in my application domain to the repertoire of problem types**

Yes, I have mapped problems in my application domain to these various problem types. Following the data reduction of the various healthcare claims and eligibility data, regression analysis to predict certain scores or values for a population or quality measure, clustering certain events or portions of the population together, classifying medical events or demographic details into certain categories or buckets, similarity matching enrolled members together based on certain characteristics like name and date of birth, profiling to identify various outliers for quality measures or cost, causal modeling to determine if a certain event or medical event can be used as a predictor for other medical events, etc. are all various problem types that I have daily familiarity with.

Application in Domain of Interest

Domain: Health care

Project 1 Description: Identify individuals within certain provider organizations or physician groups with a high total cost of care for a measurement period as a means of intervention and providing more preventative care to drive costs down moving forward.

Project 1 Problem Type: This is a profiling problem type that evaluates the total cost of care associated with all patients tied to various provider groups and flags high outlier costs. The high outlier cost patients would be those targeted with additional preventative care.

Problem 2 Description: Create an Adverse Childhood Experiences (ACEs) flag through use of available demographic and medical claims data to identify a portion of the population in need of additional community and social resources to ensure long term opportunities and health are not impacted.

Project 2 Problem Type: This is a classification problem. Through demographic data like first name, last name, date of birth, address, subscriber relationship, etc. and medical claims data around alcohol or drug abuse, for example, an at-risk ACEs child population can be identified and intervened with moving forward. This essentially would be a “Yes”, “No” flag that would make a determination on whether the child is in an at-risk environment, which would allow for early preventative medical and social care.

Questions, Maxims, and Commitments

Question (I will always ask…)
Will the final product or model be actionable for the downstream care providers?

1-Sentence Project Description
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
There will likely be multiple different types of downstream users of this ACEs flag, such as medical care providers, foster care care organizations, social workers, etc. so an overcomplicated model or approach does not need to be shared with these stakeholders as the details/approach to arrive at this flag are not necessary. To be a useful and actionable end product, the flag just needs to be a simple yes or no classification that will allow for tailored care for those who most need it.

Importance

This ties in nicely with a lot of the discussion in both lecture and in this week's selected interviews. An incredibly desired skill of a data scientist is not just to be able to code effectively and create a very advanced model but to have the ability to understand an end user's need for the end product and be able to bridge the gap between the logic and approach to generate the model and how to use it in a day to day application. A data scientist, in not just a classification problem type but in all problem types, needs to be an effective presenter across all knowledge levels to ensure that the work is not "wasted" and can have a meaningful impact. As part of this, this again reiterates the importance of continuously asking "so what" as work progresses towards an end goal.

Maxim (I will always say…)
Data beats emotions.

Which Project
A profiling problem type that evaluates the total cost of care associated with patients, flagging high outlier patient costs for more targeted preventative care.

Meaning in Context
In the case of a profiling problem type that is focused around outliers in medical care costs, the data should override the emotions of perhaps a care provider or a stakeholder involved in the care delivery process. Oftentimes if not always, data is incredibly convincing and clear in what it is portraying and should be weighed more heavily than emotions. If the data, in this profiling case, is showing a patient with extreme ($50k a year, for example) costs, the profiling would quickly flag this as an outlier. The outlier analysis would make clear that a change would be needed, which perhaps would not have been the case if the data was not available and the care decisions were influenced by emotions of a family member or a care provider continuing to stick to the same process of care despite it not being perhaps the optimal approach.

Importance
Oftentimes, it is possible for emotions (whether you have a vested interest in a particular outcome, company, person, etc.) to influence a decision. As a result, this decision may not lead to the optimal outcome, as this optimal outcome may have required actions that were uncomfortable or difficult to make due to this vested interest. This maxim is valuable to keep in mind as it makes clear that data is so valuable and can often be so convincing and clear in what it's presenting that it should be weighed heavily against the raw emotion tied to a preexisting decision or approach. This is often difficult to carry in day to day work as many emotions and decisions from many different coworkers and stakeholders need to be considered, but I believe a reliance on the data is a great way to steer a conversation to the optimal outcome.

Professional/Ethical commitment (I will always/never...)
I will not sacrifice the integrity of the data itself to ensure hitting a quota or goal set by a client.

Which Project
A profiling problem type that evaluates the total cost of care associated with patients, flagging high outlier patient costs for more targeted preventative care.

Meaning in Context
In identifying a high cost outlier, it could be possible to define an outlier in a wide number of ways, either a certain cost threshold ($50k+, for example) or a certain cost category being higher than expected ($5k+ cost in an emergency department, for example). If there was an incentive to identify as many outliers as possible by a client, I would ensure that the definition of an outlier is very clear early on in the formulation process to ensure that all parties involved are on the same page and that there is reliability and consistency in regards to how an outlier patient was flagged.

Importance
This is a constant in my own current day to day work, as a lot of my time is spent not necessarily coding but also meeting with clients or internally to develop specifications for how the reporting will be generated. To make sure that there are a limited number of questions when the final product is generated and that the client or manager is getting the final output that they expected, these types of specification meetings are vital, especially if there are incentives tied to the number of individuals that may be bucketed into various categories. There are a wide number of ways to profile the data, so the problem formulation stage always should include transparency between all involved parties to ensure that the outcome is reliable and in line with expectations.

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

Using enrollment and claims data from health insurance companies, we created a regression model identifying members with a high probability of emergency department usage! This model will allow providers to more effectively target their patients with preventative care.

Reading Responses

- **Law of Small Numbers**

The author offered a recommendation to readers and researchers alike to “replace impression formation by computation whenever possible.” This ties in well with last week’s maxim, as well, in that data beats emotions. It is important as a data scientist to not just settle on a finding that seems reasonable but instead validate and confirm the finding through well formulated analyses. ***Data Collection and Cleaning - Maxim***

The author also asserts that “people are not adequately sensitive to sample size.” The article specifically used a sample of 150 versus 3,000 to make the point that few would be able to determine what the “appropriate” size would need to be for the study to be considered statistically significant and thus meaningful. This is an important note for my own work moving forward, as well, to better understand what is needed for work to be considered significant and ensure that the data collection and cleaning processes are in place to allow for this to happen. ***Data Collection and Cleaning - Maxim***

- ***Statistical Biases Types Explained***

“[Observer bias] can come in many forms, such as (unintentionally) influencing participants...or doing some serious cherry picking.” This is a very important note that is especially relevant currently in my own day to day work, as, in the healthcare space and with COVID, healthcare utilization and costs decreased between 2019 and 2020 almost across the board. However, one should definitely not suggest that there has been improvement in the system due to this trend and instead evaluate a multi-year trend. This portion of the reading was an important reminder to be aware of this bias and always provide appropriate context where it is needed. ***Data Collection and Cleaning - Expertise***

A relatively new form of bias that was introduced to me in this reading was Survivor Bias, or when “the researcher focuses only on that part of the data set that already went through some kind of pre-selection process.” This was an important reminder and acknowledgment that data from outside sources or from a relatively new process should be evaluated and quality checked thoroughly before incorporating into an analysis. This would be an effective step in ensuring completeness and that there were no omissions, intentional or not, that could influence what the data would eventually show in reporting. ***Data Collection and Cleaning - Expertise***

- ***Data Cleaning 101***

A perhaps simple but valuable question posed in this article when evaluating a data set for cleanliness is “does the data match the column label?” Being in the big data healthcare space, this can often be a challenge, albeit a very important one, to ensure. Prior to sending over any tables or datasets to a client, there are thorough quality checks to validate that Personal Identifiable Information (PII) did not sneak into a delivered field (an email address or phone number in a city or zip code field, for example). This type of data cleaning is a significant portion of the day to day role to allow for confidence in a delivered product. ***Data Collection and Cleaning - Question***

Another important note that the author makes in this article is to “communicate with the source.” As I’ve seen again in my own day to day work. It’s very common for there to be data errors or omissions in submitted files or data sets. As opposed to pushing through with perhaps an inaccurate or incomplete dataset, it is a valuable skill as a data scientist to be proactive and follow up with a client or data partner to ensure receipt of the cleanest data possible. Further, it can often, as well, to not just follow up when questions arise but push for documentation at the start to again ensure that both parties are on the same page. ***Data Collection and Cleaning - Maxim***

- ***10 Rules for Creating Reproducible Results in Data Science***

A valuable first rule by the author was that “for every result, keep track of how it was produced.” This is a very important part of the data collection and cleaning as this is a step in the process that is often repeated. Implementing an automated process and providing documentation that would allow for anyone to run the program and reproduce the results are common practices in my own day to day work, as well. ***Data Collection and Cleaning - Expertise***

In a similar way, the author notes the value in “record[ing] all intermediate results, when possible, in standardized formats”. Often when collecting and cleaning data, there are many temporary or base tables that I create as part of the overall build to the final “clean” product. Saving these tables allows for both myself and others to quality check and validate that each step is working as expected and allows for much easier backtracking when trying to identify a particular bug. ***Data Collection and Cleaning - Expertise***

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 2, Data Collection & Cleaning

1. **Common problems with data sets that can lead to misleading results of analyses**

Yes, this is a capability I am familiar with in my own line of work. Oftentimes data being submitted to my organization on a day to day basis has issues within certain fields (either inappropriately left blank or null, having an incorrect value, etc.) so I have become adept at establishing workarounds for this incorrect data. However, I also have become comfortable communicating with the source of the data to confirm that the issue is present and whether the workaround is appropriate or a re-submission would be required.

2. **Potential data sources in my application domain**

Yes, I am also familiar with the data sources in the healthcare domain. Predominantly in my own day to day work we handle millions to billions of rows of medical, pharmacy, and dental claims and enrollment data. Additionally, other data sources like death and birth certificate data, Medicaid capitation files, and electronic medical records are additional files that are available for certain projects depending on the particular client.

3. **How to understand and document data sets**

In addition to my day to day work with large data sets, we also are responsible for documenting our processes and how the data sets are pulled in, cleaned, etc. For this portion of the role, we rely heavily on JIRA and Confluence to track the steps taken and ensure that the process from start to finish is well documented to ensure that any analyst or colleague would be able to pick up the project or dataset and be able to complete the work.

4. **How to write queries and scripts that acquire and assemble data**

This is something that I have become very comfortable with throughout my time at my current company. There are many processes in place that we leverage that utilize Spark SQL as a way of loading claim and enrollment data not only into our system but also transforming and cleaning said data in a predefined way depending on a particular client. More downstream, I also heavily rely on SAS to import and export data from disparate sources and again clean and transform various fields as needed depending on how dirty the data may be (inconsistent field population, errors within certain fields, etc.).

5. **How to clean data sets and extract features**

A large part of my day to day work is to also use SQL to query large datasets and glean meaningful insights and create perhaps more condensed datasets that are more practical and manageable to work with. SQL is a great tool to not only query against tables with perhaps millions and billions of rows, but is also very useful when there are known data issues. SQL logic is very effective at cleaning data essentially “on the fly” and is something that I have become very comfortable with.

Maxims, Questions, and Commitments

Question (I will always ask…)
Does the available data for a particular table or column make sense?

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
For a project around Adverse Childhood Experiences, there would need to be very clean demographic data around first and last name, subscriber relationship data on medical enrollment files, address data, etc. in order for a child to be appropriately linked to an adverse experience. If there are phone numbers in an address field or inconsistent or incomplete population of some of these fields, the generation of the flag would not be as successful if there were a more complete file to draw from.

Importance
This ties in really well with the reading around data cleaning and Professor Resnick's note that most of a data scientist's day to day work will revolve around cleaning data and ensuring that it is sufficient for use in any downstream reporting. Before beginning any project, it is key to evaluate the data sets that will be utilized to ensure that you have what is needed to complete a project satisfactorily for a client or colleague. Taking these data cleaning steps early in the process ensures that sufficient workarounds can be put in place and, if needed, a data provider can be reached out to for any clarification or perhaps even a resubmission of a file in question.

Maxim (I will always say…)
Don't treat the symptom, figure out what the root cause is.

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
As noted earlier, when dealing with large amounts of enrollment and medical claims data, there is a high possibility that data may perhaps be incomplete or submitted to our organization incorrectly. While it would be straightforward to perhaps just establish a workaround (if zip code, for example, is not submitted with its leading 0s, it would be very easy to just add a cleaning step that adds the zero in), it would be best to instead reach out to the data provider or source and bring this issue to their attention. Cleaning up the pipeline itself is a better approach than creating workarounds as, if the latter approach is taken, it will not be long until there are perhaps an overwhelming number of steps that are needed to ensure that the data is as clean as possible.

Importance
As noted above and in the readings and lectures for the week, it is critical to not only just develop the skills needed for data cleaning and manipulation, but a data scientist must become comfortable with the communication portion of their role. With so much data available, especially within the healthcare domain, there will always be situations where there is dirty data. With a lot of these issues often being data entry or population driven, reaching out to the organization that provided the data is often the best way to ensure that the data is not just corrected but is correct for all future iterations. This ties in well as well with the importance of carefully documenting steps within the data cleaning process. Noting the history that a field was previously wrong but since corrected is valuable as you may not always be the analyst working a particular project, and it's important to have the data as accurate as possible to avoid a case where a role is transferred and the cleaning/workaround process is not well fleshed-out.

Professional/Ethical commitment (I will always/never...)
I will always keep track of how a process is implemented to allow for any analyst or colleague to pick up a particular project

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
As noted earlier, while data collection and cleaning was the major objective of this week's lectures, it is also very important to be able to carefully document the steps that were taken to arrive at a "clean dataset". If there were files that needed to be resubmitted, due to perhaps a data entry error, or there were certain fields that required a workaround to allow for appropriate flagging of children with adverse experiences (backfilling the zeros for a zip code, for example) this would all be important to note. Ideally, anyone is able to pick up a project and complete it successfully if the appropriate level of documentation is available.

Importance
This week's lectures and readings generally showed valuable tips and skills to acquire to clean and collect data. However, I believe below the surface that it also highlighted very good practices that should be implemented by all data scientists. Although not specifically related to data collection and cleaning, documentation and relaying findings are part of an invaluable process that ensures a project will not just be dependent on one person and that the minute details are accounted for during any iteration of the collection process.

Week 3: Data Analysis and Modeling Stage

Informational Interview - Reflection

Instructions (delete when submitting):

Synthesizing the information gleaned from the interview that you conducted, read, or listened to, write a 250-500 word reflection on what you have learned about being a data scientist. In your reflection, you must:

1. *Identify and describe at least three insights relevant to course content. These should take the form of one question, one maxim, and one professional (or ethical) commitment.*
2. *Map these three insights to the data science project stages framework, as you have in the weekly maxims, questions, and commitment assignments.*
3. *Brainstorm three additional follow-up questions that you would have liked to ask the interviewee.*

Grading Rubric

- 1 point: Insight in the form of a **Question** is relevant to the course content, and described in adequate detail.
- 1 point: **Question** is correctly mapped to a single data science project stage.
- 1 point: Insight in the form of a **Maxim** is relevant to the course content, and described in adequate detail
- 1 point: **Maxim** is correctly mapped to a single data science project stage.
- 1 point: Insight in the form of a **Commitment** is relevant to the course content, and described in adequate detail
- 1 point: **Commitment** is correctly mapped to a single data science project stage.
- 1 point: Reflection is within recommended word count range, or contains a sufficient amount of detail to demonstrate what has been learned from the interview.
- 1 point: Reflection contains follow up questions.
    - .5 point deduction: Follow up questions have been included, but may be less than required (3), insufficiently specific, or seem unimportant to ask.

Reading Responses

Instructions (Delete these in your submission)

For each required reading, identify and explain two insights that you extracted from it, in the form of a question, maxim, or professional (or ethical) commitment. For each insight, first describe it in 1-3 sentences and then, in bold, label it according to the following framework: ***{Stage the insight is relevant for: problem formulation; data collection & cleaning; modeling & analysis; presentation & deployment} - {which of the following it is: expertise; goal; maxim; question; commitment}***

Here are some examples:

1. *Tait observes that it is important to "avoid manual data manipulation steps." When you clean data by hand, it is not a reproducible step that others can use in the future to validate/repeat your work.* ***Data Collection and Cleaning - Maxim***
2. *"Outcome proxies will be gamed." When you define proxies for the outcomes you really care about, people may start behaving in ways that obscure the natural correlations between the proxy and the real outcome of interest.* ***Problem Formulation - Maxim***
3. *"Who will be using the results and for what decisions?" Knowing who's going to use the results and how they're expecting to use it may shape data collection, analysis, and implementation.* ***Problem Formulation - Question***

Grading rubric (for each of two insights, for each reading):

- .5 point: articulates a meaningful insight that makes reference to something in the reading
- .5 point: correctly interprets the insight in the reader's own words

- ***Overfitting in Machine Learning: What is it and how to prevent it***
- ***Common pitfalls in statistical analysis: The perils of multiple testing***
- ***P-Hacking and the problem with Multiple Comparisons***
- ***Correlation vs. Causation: An Example***
- ***Simpson's Paradox in Real Life*** *or* ***Ignoring a Covariate: An Example of Simpson's Paradox***
- ***Conditioning on a collider***

Plan for Knowledge Acquisition

Instructions (Delete these in your submission):

For each item below, select one of the following:

- *I already have this capability. If so, describe how you acquired it.*
- *I look forward to strengthening this capability. If so, explain how. Mention specific courses where you think it will be covered or outside activities you intend to engage in.*

Note: you only need 1-3 sentences for each, though you are welcome to write more if you want.

Grading rubric (for each of the capabilities for the week):

- 2 points: Describes how capability was already acquired OR
- 2 points: Explains plan for how capability will be acquired
- (Note: maximum of 2 points total; it's OK to describe both how you already learned something about this capability and your plans to learn more, but you can only earn points from one or the other of the two rubric elements).
- Possible deductions
    - -1 Seems to misunderstand the capability
    - -1 Plan is vague; doesn't look ahead to the curriculum or other outside resources to make a guess about where it might be covered.
    - -1 Description of how capability was acquired is vague (e.g., "I have it from my job")

Skills and Knowledge Inventory: Stage 3, Data Analysis & Modeling

- common mistakes in data analysis that lead to misleading results
- a repertoire of models and how to estimate, validate, and interpret each of them

Maxims, Questions, and Commitments

Instructions (Delete these when submitting)

As with any professional, every data scientist has certain beliefs about their work that define how they conduct themselves on a daily basis. Based on what you learn each week about the profession, we will ask you to identify and share beliefs that resonate with you in the form of questions, maxims, and professional (or ethical) commitments. You will have to provide one question, one maxim, and one commitment each week.

For each, you will provide:

- *A **one-sentence statement** of the question, maxim, or commitment.*
    - *Please be sure that it is relevant to the project stage that was covered that week (e.g., problem formulation in week 1).*
- *Which of your two projects from your Application in Domain of Interest you will apply it to. Please just include a one-sentence summary of the project; the reader can refer back to the full description.*
- *One paragraph explaining **what it means.***
    - *Please be sure to explain with respect to the particular context of the hypothetical project**.***
- *One paragraph explaining **why it is valuable** to ask that question, make that statement, or state that commitment. How would it make the particular project go better, or help you avoid some pitfall?*

Question (I will always ask…)

Grading rubric:

- 1 point: Provides a one-sentence question
    - .5 point deduction: Multiple questions rather than a single one.
    - 1 point deduction: The question is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the question means in the context of the project specified
- 1 point: Explains why it is valuable to ask the question by suggesting how it would make the particular project go better.
- 1 point: Question, as applied, applies primarily to the current stage of the project (data analysis and modeling).

Which Project

Meaning in Context

Importance

Maxim (I will always say…)

Grading rubric:

- 1 point: Provides a one-sentence maxim
    - .5 point deduction: Multiple maxims rather than a single one.
    - 1 point deduction: The maxim is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the maxim means in the context of the project specified
- 1 point: Explains why it is valuable to apply the maxim by suggesting how it would make the particular project go better.
- 1 point: Maxim, as applied, applies primarily to the current stage of the project (data analysis and modeling).

Which Project

Meaning in Context

Importance

Professional/Ethical commitment (I will always/never...)
Grading rubric:

- 1 point: Provides a one-sentence commitment
    - .5 point deduction: Multiple commitments rather than a single one.
    - 1 point deduction: The commitment is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the commitment means in the context of the project specified
- 1 point: Explains why it is valuable to articulate the commitment by suggesting what, in the context of the particular project, might create an incentive not to take the action you've committed to.
- 1 point: Commitment, as applied, applies primarily to the current stage of the project (data analysis and modeling).

Which Project

Meaning in Context

Importance

Week 4: Presenting and Integrating into Action

Sources for Data Science News

> ***Instructions (delete before submitting)***
>
> *You will write a brief plan describing what sources of information about data science you plan to follow outside of assigned readings from this program. This could include blogs, podcasts, newsletters, conferences, or other sources. Present it as a short bulleted list, with a sentence describing why you plan to follow that source.*
>
> *When listing which resources you will use, be mindful of how many you are including. Too many resources will be unreasonable to keep up with. Too few resources will not keep you up to date with the industry.*

I plan to follow the following sources of information about data science to keep myself up to date with the industry:

-

Grading rubric:

- 1 point: Provides list of data science news sources.
- 1 point: Each source is accompanied by a short description of why it was chosen, will be useful, what it is, etc.
- 1 point: List is of a reasonable size (e.g. too many resources will be unreasonable to keep up with; too few resources will not keep you up to date with the industry.)

Reading Responses

Instructions (Delete these in your submission)

For each required reading, identify and explain two insights that you extracted from it, in the form of a question, maxim, or professional (or ethical) commitment. For each insight, first describe it in 1-3 sentences and then, in bold, label it according to the following framework: ***{Stage the insight is relevant for: problem formulation; data collection & cleaning; modeling & analysis; presentation & deployment} - {which of the following it is: expertise; goal; maxim; question; commitment}***

Here are some examples:

1. *Tait observes that it is important to "avoid manual data manipulation steps." When you clean data by hand, it is not a reproducible step that others can use in the future to validate/repeat your work.* ***Data Collection and Cleaning - Maxim***
2. *"Outcome proxies will be gamed." When you define proxies for the outcomes you really care about, people may start behaving in ways that obscure the natural correlations between the proxy and the real outcome of interest.* ***Problem Formulation - Maxim***
3. *"Who will be using the results and for what decisions?" Knowing who's going to use the results and how they're expecting to use it may shape data collection, analysis, and implementation.* ***Problem Formulation - Question***

Grading rubric (for each of two insights, for each reading):

- .5 point: articulates a meaningful insight that makes reference to something in the reading
- .5 point: correctly interprets the insight in the reader's own words

- ***A History Lesson On the Dangers Of Letting Data Speak For Itself***
- ***Storytelling for Data Scientists***
- ***Interpretability is crucial for trusting AI and machine learning***
- ***The Signal and the Noise, Chapter 2***
- ***The Signal and the Noise, Chapter 6***
- ***How Not to Be Misled by the Jobs Report***
- ***But what is this "machine learning engineer" actually doing?***
- ***How we scaled data science to all sides of Airbnb over 5 years of hypergrowth***

Plan for Knowledge Acquisition

Instructions (Delete these in your submission):

For each item below, select one of the following:

- *I already have this capability. If so, describe how you acquired it.*
- *I look forward to strengthening this capability. If so, explain how. Mention specific courses where you think it will be covered or outside activities you intend to engage in.*

Note: you only need 1-3 sentences for each, though you are welcome to write more if you want.

Grading rubric (for each of the capabilities for the week):

- 2 points: Describes how capability was already acquired OR
- 2 points: Explains plan for how capability will be acquired
- (Note: maximum of 2 points total; it's OK to describe both how you already learned something about this capability and your plans to learn more, but you can only earn points from one or the other of the two rubric elements).
- Possible deductions
    - -1 Seems to misunderstand the capability
    - -1 Plan is vague; doesn't look ahead to the curriculum or other outside resources to make a guess about where it might be covered.
    - -1 Description of how capability was acquired is vague (e.g., "I have it from my job")

Skills and Knowledge Inventory: Stage 4, Presenting & Integrating into Action

- **how to present results to domain experts who are not data scientists**
- **how to work with software engineers to put models into production**

Maxims, Questions, and Commitments

Instructions (Delete these when submitting)

As with any professional, every data scientist has certain beliefs about their work that define how they conduct themselves on a daily basis. Based on what you learn each week about the profession, we will ask you to identify and share beliefs that resonate with you in the form of questions, maxims, and professional (or ethical) commitments. You will have to provide one question, one maxim, and one commitment each week.

For each, you will provide:

- *A **one-sentence statement** of the question, maxim, or commitment.*
    - *Please be sure that it is relevant to the project stage that was covered that week (e.g., problem formulation in week 1).*
- *Which of your two projects from your Application in Domain of Interest you will apply it to. Please just include a one-sentence summary of the project; the reader can refer back to the full description.*
- *One paragraph explaining **what it means.***
    - *Please be sure to explain with respect to the particular context of the hypothetical project**.***
- *One paragraph explaining **why it is valuable** to ask that question, make that statement, or state that commitment. How would it make the particular project go better, or help you avoid some pitfall?*

Question (I will always ask…)
Grading rubric:

- 1 point: Provides a one-sentence question
    - .5 point deduction: Multiple questions rather than a single one.
    - 1 point deduction: The question is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the question means in the context of the project specified
- 1 point: Explains why it is valuable to ask the question by suggesting how it would make the particular project go better.
- 1 point: Question, as applied, applies primarily to the current stage of the project (presentation and action).

Which Project

Meaning in Context

Importance

Maxim (I will always say…)

Grading rubric:

- 1 point: Provides a one-sentence maxim
    - .5 point deduction: Multiple maxims rather than a single one.
    - 1 point deduction: The maxim is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the maxim means in the context of the project specified
- 1 point: Explains why it is valuable to apply the maxim by suggesting how it would make the particular project go better.
- 1 point: Maxim, as applied, applies primarily to the current stage of the project (presentation and action).

Which Project

Meaning in Context

Importance

Professional/Ethical commitment (I will always/never...)
Grading rubric:

- 1 point: Provides a one-sentence commitment
    - .5 point deduction: Multiple commitments rather than a single one.
    - 1 point deduction: The commitment is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the commitment means in the context of the project specified
- 1 point: Explains why it is valuable to articulate the commitment by suggesting what, in the context of the particular project, might create an incentive not to take the action you've committed to.
- 1 point: Commitment, as applied, applies primarily to the current stage of the project (presentation and action).

Which Project

Meaning in Context

Importance