Personal Manifesto

By: Varshini Rana

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

Grading rubric
1 point: Identified who will participate in the interview (or a process for finding an individual)
1 point: Described why the interviewee was chosen
1 point: Identified how the interview will be conducted or collected.

For the Informational Interview, I will be collecting an interview transcript from the book Data Scientists at Work by Sebastian Gutierrez, which is a collection of sixteen interviews of renowned data scientists. The interviewee shall be Eric Jonas, a research scientist in computational neuroscience and signal processing, who is currently a Professor in the Department of Computer Science at the University of Chicago. I chose him as the interviewee for this task because the field of computational neuroscience is something that I have always been intrigued by, and I believe that his industrious work in this field makes him an exemplary authority to talk about it. I am very keen on understanding how he leverages neuroscience data to determine how the brain learns, forms memories and assimilates new information.

Reading Responses

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

Readings:

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

1. “Data mining results should influence and inform the data mining process itself, but the two should be kept distinct.” The data mining process to build a model and the way the model (the results of data mining) is used are distinct in that the process of building a model is often with historical data, while the model so built takes existing or new data points as an input to provide a prediction.
2. “Data mining is a craft.” While data mining involves a fair amount of science and technology, the most effective insights from data mining require quite a bit of creativity to bring it all together. Like any craft, a well-structured approach to data mining with some leeway to allow for creative conceptualization would go a long way in solving the problem in question.

- **Chris Wiggins interview**

1. “The key is usually to just keep asking, “So what?”” Thinking through how your actions impact something that you value by asking “so what?” after every action that you undertake enables you to streamline your thought processes to be clear about what you’re working towards and what you want to accomplish.
2. “The way you become an expert in a field is to make every mistake possible in that field.” Making every possible mistake gives you exposure to every possible scenario where failure occurs so that you’re better prepared for the next time you encounter such scenarios and rise up to the challenge by providing an optimum solution.

- **Erin Shellman interview**

1. “Data’s just the world making noises at you.” Raw data is often messy, crude, and unintelligible. The insights that you can glean from it after massaging it in an appropriate way is what makes the data valuable.
2. “If you talk to somebody who has something you want, follow up.” The field of Data Science often demands the ability to ask for what you want and to be persistent in the pursuit of it if you truly want to obtain it. This would impress upon the mind of the person whose help you wanted about your seriousness and commitment to seeing your project through, in which case they would be more inclined to give you what you want.

- **Jake Porway interview**

1. “I still believe that a data scientist is just a statistician who can program well.” A statistician is able to leverage mathematical techniques to analyse and draw key conclusions from data. A data scientist just extends that thinking by collecting, modeling, visualising, and drawing meaning from data through the judicious use of statistical formulae and computer algorithms, among other things.
2. “There’s almost no limit to where data and data science can be applied.” Every single action (or sometimes even inaction) of ours, who we are, what we do, how we do it, and everything else is a possible data point for someone who looks closely enough and wants to achieve a certain goal related to it.

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 1, Problem Formulation

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")

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

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, most notably on a fraud detection project I was working on in the domain of Health Insurance. During the initial stages of project development, I had to pick the brains of the leaders and members of the fraud investigation team — many of whom were former practicing physicians — of the health insurance companies who had partnered with my start-up. Long discussions were had on the numerous pain points they experienced during the course of their investigations, the available data and any other data they could start collecting during the course of claim processing which they thought would help me in developing an intelligent fraud detector, insights gleaned from their domain expertise, and tiny nuances they pointed out which I would've missed out on by just looking at the data alone. All of these discussions were immensely valuable and indispensable in helping me formulate an actionable problem statement.

2. **a repertoire of problem types**

I do not have this capability yet as I'm still at the budding stage of my Data Science career with exposure to mainly Classification (or Class Probability Estimation) problem types, although I am aware of the existence of other problem types. I plan to acquire this capability through a new job I accepted in an MNC starting in February 2022, which shall expose me to various domains such as Finance, Retail, and Manufacturing. I expect to encounter and work on a variety of

problems such as those related to causal modeling, similarity matching, profiling, and co-occurrence grouping among others. I also plan to develop my repertoire of various problem types through MADS coursework, which includes courses such as SIADS 543: Unsupervised Learning, SIADS 644: Reinforcement Learning Algorithms, SIADS 652: Network Analysis, and SIADS 685: Search and Recommender Systems. The Milestone and Capstone project-based MADS coursework shall also go a long way in helping me develop my repertoire of various problem types in the field of Data Science.

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

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, involved in building applications mainly in the Healthcare and Health Insurance domains. While my hands-on exposure is currently limited to a small subset of the repertoire of problem types, I am aware of how to map them with the problems in my domain thanks to various conversations with mentors involved in this line of work. For example, the prediction of the probability that a person with Chronic Kidney Disease would succumb to complications based on lab test values and frequency of dialysis treatments is a class probability estimation problem. In this case, preventive interventions can be put forth to forestall the deterioration of health. Another example would be link predictions in computerized drug discovery, wherein large volumes of biomedical data represented as a network become useful for predicting results from drug-drug and drug-disease interactions, which leads to faster and more efficient drug discovery.

Application in Domain of Interest

Grading rubric (for each project):

- 1 point: Project is clear as described
- 1 point: Involves investigation of a data science problem, meaning that it can be solved through collection and analysis of data, and presentation of results.
    - .5 point deduction: Involves **multiple** data science problems, rather than just one.
- 1 point: It's clear what kinds of data would be used.
- 1 point: It's clear what the results might be used for.
- 1 point: Classification of problem type is correct.
- 1 point: Explanation for classification is correct and specific
- 1 point deduction: Classification is for the wrong problem, not the primary problem in the problem description
- 1 point deduction: Both projects have the same problem type

Domain: Healthcare

Project 1 Description: The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Project 1 Problem Type (with explanation): This is a classification (or class probability estimation) problem. The image classification model would predict either the presence or absence of pneumonia based on lung opacity (an indicator of the presence or absence of lung inflammation) as seen on chest x-rays of pneumonia patients and healthy control respectively.

Problem 2 Description: Information extraction (named entity recognition) of clinical words, phrases and standard medical codes (such as those for diseases and procedures) from unstructured clinical text such as discharge summaries to make health insurance claim processing faster and more efficient.

Project 2 Problem Type (with explanation): I would classify this as an optimization problem. Currently, due to the lack of availability of Electronic Health Record (EHR) resources in my home country, the process of perusing through discharge summaries to determine the cause of admission of the patient to the hospital is being done manually by health insurance claim processors. This is a very cumbersome and time-consuming process. This results in an increase in health insurance claim processing time, money spent on resources, and also the probability of human error. Automatic information retrieval would do away with all of these pitfalls.

Questions, Maxims, and Commitments

Question (I will always ask…)
Whom will these results benefit?

1-Sentence Project Description
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
In this case, it is possible that these results will benefit radiologists or doctors in making accurate diagnoses, which they would've otherwise missed out on due to microscopic subtleties in the X-rays which are invisible to the naked eye and the possibility of human error. These results also have direct implications and benefits for patients undergoing this radiological procedure as these ensure that they receive timely and faultless treatment for the ailment so detected.

Importance for this stage of the project
Knowing who is going to consume these results and how it will benefit them is probably the most important consideration in the problem formulation stage as it will give the data scientist an idea of the return on investment he will obtain if he undertakes this project. By knowing the target market for the results of his project, the data scientist can ensure that he directs his attention towards the goal of making things easier for them (radiologists and doctors in this case) and can anticipate the potential compensation he might receive if he chooses to undertake this project.

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 (problem formulation).

Maxim (I will always say…)
A good data science problem will aim to provide actionable insights, not just make predictions.

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
If it is determined that the results of the predictive model point to a patient having pneumonia, he should be provided the appropriate medical treatment for it. The results of this particular project are such that they demand immediate action and empower the stakeholders to make informed decisions, as opposed to a prediction with just theoretical implications.

Importance for this stage of the project
Formulating a problem statement to work on by considering the possible actions that would be taken based on the potential results would give the data scientist an assurance that the results of his work would have the impact that he intended (i.e. in this case, his prediction of pneumonia or lack thereof are useful and are being taken seriously), since appropriate actions are being taken based on the insights that he generated.

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 (problem formulation).

Professional/Ethical commitment (I will always/never...)
I will always be aware of and respect the licensing status of the data that I want to use in my project.

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
In this case, the project would make use of chest X-rays from patients for training an image classification model. Protecting the privacy of these patients and adhering to the terms of fair use of this data is the ethical responsibility of the data scientist who intends to use this data. Misuse of this data or non-compliance to the specifications of the data license would cause a breach of confidentiality, which could have serious legal, social and financial consequences.

Importance for this stage of the project
The stage of problem formulation is the stage where the data scientist starts thinking about the kind of data he would require for his project and where to obtain it from. It is important that he ensures that the data he chooses to use is legally authorised for his use. In this case, since the data in question is PHI or Protected Health Information, the sensitive nature of the data makes it imperative that he obtain the necessary permissions to use the data as intended beforehand. If he isn't able to obtain the necessary permissions, he may have to source permitted data from elsewhere or rethink the entire project. Hence, being aware of the licensing status of the data and how he hopes to adhere to it also enables him to plan for the project.

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 (problem formulation).

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

Instructions (Delete these when submitting)

Make a plan for a personal project in your application domain of interest. You are not required to complete this personal project as a part of the degree program, but it is a good idea to complete it for your own personal learning and to demonstrate your learning to potential future employers.

The project plan (inspired by step 2 of Monica Rogati's article "How do I become a data scientist?") should be described ***in the form of a tweet (280 character limit).*** *In it, you will explicitly mention the sources of data that would be used and the expected outcome of your project. Including a "hook" is recommended but not required. For examples of project tweets, read "How do I become a Data Scientist?"*

Please include a character count (including spaces) with your tweet submission.

Grading rubric:

- 1 point: The project is clearly described in the tweet.
- 1 point: The expected outcome of the project is mentioned in the tweet.
- 1 point: The source of data to be used is identified in the tweet.
- 1 point: The tweet is not too long (less than ~280 characters).

Reading Responses

Instructions (Delete these in your submission)

For each required reading, identify and explain two insights that you extracted from it. For each, make sure that you reference something from the reading and that you explain it in your words (it's OK to quote, but then explain).

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.*
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.*
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.*

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

- **Law of Small Numbers**
- ***Statistical Biases Types Explained***
- ***Data Cleaning 101***
- ***10 Rules for Creating Reproducible Results in Data Science***

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 2, Data Collection & Cleaning

1. **common problems with data sets that can lead to misleading results of analyses**
2. **potential data sources in my application domain**
3. **how to understand and document data sets**
4. **how to write queries and scripts that acquire and assemble data**
5. **how to clean data sets and extract features**

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 collection and cleaning).

Which Project

Meaning in Context

Importance for this stage of the project

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 collection and cleaning).

Which Project

Meaning in Context

Importance for this stage of the project

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 collection and cleaning).

Which Project

Meaning in Context

Importance for this stage of the project

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. *Include a short description of the person (and professional role) and interview type (e.g. interview you conducted, or that you located and watched online) in your reflection.*
2. *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****. Insights should be clearly labeled.*
3. *Map these three insights to a single stage in the data science project stages framework covered in each week of this class (e.g. W1 problem formulation, W2 data collection and cleaning, etc.).*
    i. *E.g. As the only data scientist on a team with widely diverging areas of expertise (e.g. a product manager, engineer, marketing exec, etc.), Mason described the early stage of their data science projects as an iterative and non-linear process involving negotiation among many different stakeholder viewpoints. In other words, the original formulation is rarely the final formulation.* ***[maxim - W1 problem formulation]***
4. *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 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 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 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. For each, make sure that you reference something from the reading and that you explain it in your words (it's OK to quote, but then explain).

Here are some examples:

4. *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.*
5. *"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.*
6. *"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.*

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 for this stage of the project

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 for this stage of the project

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 for this stage of the project

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. For each, make sure that you reference something from the reading and that you explain it in your words (it's OK to quote, but then explain).

Here are some examples:

7. *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.*
8. *"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.*
9. *"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.*

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 for this stage of the project

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 for this stage of the project

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 for this stage of the project

Document update information

Updates to this document after the start of class: