Personal Manifesto

By: [Simi Talkar]

Table of Contents

[**Week 1: Problem Formulation Stage 2**](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.65bnj6mnuldl)

[Informational Interview - Planning 2](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.s1r8mdz6eptf)

[Reading Responses 3](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.4n1ts0uqnq6o)

[Plan for Knowledge Acquisition 4](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.biz3lrcn7jxd)

[Skills and Knowledge Inventory 4](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.5nqnacbkifrs)

[Application in Domain of Interest 5](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.2o2zyrpfrthv)

[Maxims, Questions, and Commitments 7](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.v5kn5xg5flet)

[**Week 2: Data Collection and Cleaning Stage 10**](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.syotjdos8i1i)

[Potential Personal Project Tweet 10](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.q3ach852s7b3)

[Reading Responses 11](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.8mgq4ado0jb3)

[Plan for Knowledge Acquisition 12](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.14db5del4lq)

[Skills and Knowledge Inventory 12](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.huctn0wah4m1)

[Maxims, Questions, and Commitments 13](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.62pupisxcybu)

[**Week 3: Data Analysis and Modeling Stage 16**](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.o829xx6n9sub)

[Informational Interview - Reflection 16](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.ciozwfe0mbrd)

[Reading Responses 17](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.pius1s786pj9)

[Plan for Knowledge Acquisition 18](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.ozq4ma1woef)

[Skills and Knowledge Inventory 18](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.hy0v7q34imun)

[Maxims, Questions, and Commitments 19](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.xyai7r84ikm6)

[**Week 4: Presenting and Integrating into Action 22**](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.83o9nyr2k2c1)

[Sources for Data Science News 22](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.kfhqf38jrhzf)

[Reading Responses 23](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.az04gna23ruf)

[Plan for Knowledge Acquisition 24](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.ll89azhytjx7)

[Skills and Knowledge Inventory 24](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.ii21kbootlse)

[Maxims, Questions, and Commitments 25](https://docs.google.com/document/d/1yHazEzwdAx3b8hHAeMI4bo2_UihrwCiUe49076oQDyc/edit#heading=h.vslkuzsw81bn)

Stages : ***1) problem formulation; 2) data collection & cleaning; 3) modeling & analysis; 4) presentation & deployment***

***What is it: expertise; goal; maxim; question; ethical commitmen***

Week 1: Problem Formulation Stage

Informational Interview - Planning

***Instructions*** *(****Delete these in your submission)***

*In this assignment you will plan to engage with a practicing data scientist to gain practice-based insights about being a data scientist. We encourage you to find a practicing data scientist and conduct an informational interview with them to understand their career trajectory and what their work as a data scientist entails. Alternatively, you can find a recorded interview or read a transcript. (Note that this should not be one of the required readings from Week 1, or the optional videos that we have included).*

*Submit one paragraph describing your plan for how you will conduct or collect the interview,  who will be the subject of your interview, and why you have chosen the interviewee. You will have until week 3 of the course to complete the interview.*

Since February 2020, I have been fascinatedly following Streetlight Data, a San Francisco based Data Analytics company that can tell city planners interesting things like where to locate charging stations for electric cars and what transit routes to keep and which can be cancelled during the Covid epidemic. I sign up for their webinars and attend them in person or watch the recordings they send me. I have learnt about one of their key metric VMT - vehicle miles travelled from their graphical presentations which show the shift in peak hours(it has moved to the middle of the day for a number of cities during the pandemic) and the possible reasons they provide for it.

And so I chose a relatively new Data Scientist from their team, Claire Douglass. I chose Claire as she has just started out at the company and will be able to give me a good perspective of what skills she had going in. She must have had expectations before joining the team and I would like to understand if they were borne out or whether she was surprised by any aspect of her job.

I also find myself trying to learn everything at once and would like to ask her how she found her focus and planned her track to her goal. I have reached out to Claire on LinkedIn and if she consents to give the interview, I will present it during the final week.

I plan to ask her :

Sandeep,

As a part of one of the course (Being a Data Scientist) we have been assigned to interview a Data Scientist in the field. Are you open to my interviewing you for this assignment.

The insights I will gain from this interview will fall broadly in the categories of "Maxims", "Questions" and "Ethical Commitment".

If you are open to it, then please let me know a convenient time that I can schedule it for - it will take about half an hour. (My deadline is Jan 19, 2021)

I would like the interview to be free flow so as not to stem your thoughts, but these are the broad categories of questions I will have for you.

Questions:

1) Can you describe the most recent project that you willingly stretched yourself on? What consumed most of your energy on it?

2) Do you create or refer to a checklist as you are handed data. If not, how do you create the first stroke on a blank canvas?

3) How do you record the requirements of the stakeholders? How do you clarify your understanding?

4) How do you understand what data is required to get insights.

5) How do you supplement you current learning and do you have a mentor?

6) What tasks are you involved in on a daily basis and how much of it is manual and how much has been automated by you or others?

Thank you are let me know if you have some time for me to conduct this interview with you.

Simi Talkar

*Note: in week 3 you will be asked to write a reflection on what you learned about being a data scientist from the interview, and you will map these insights to course content (e.g. data science project stages, and maxims, questions, and ethical commitments).*

Questions to ask Claire will involve:

Maxims

Questions

Committments

Note: in week 3 you will be asked to write a reflection on what you learned about being a data scientist from the interview, and you will map these insights to course content (e.g. data science project stages, and maxims, questions, and ethical commitments).

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 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; ethical 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***

* **Chapter 2 - Business Problems and Data Science Solutions**
* **Chris Wiggins interview**
* **Erin Shellman interview**
* **Jake Porway interview**

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 1, Problem Formulation

***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.*

1. **how to conduct an inquiry in my application domain that leads to a good problem formulation**
2. **a repertoire of problem types**
3. **how to map problems in my application domain to the repertoire of problem types**

Application in Domain of Interest

***Instructions for Application in Domain of Interest (Delete these in your submission):*** *You will describe two (2) hypothetical projects that you imagine someone could work on as a data scientist, within your chosen domain of interest. A domain could be something like health care, education, finance, sports management, retail, local government, or national security. Use only one domain for both hypothetical projects.*

*Clearly state the goal of the project and then classify the type of problem, according to the taxonomy detailed in Chapter 2 of Business Problems and Data Science Solutions. Use projects for which the problem falls into two distinct taxonomy classifications. You may use one of the same domains as provided in the sample submission, but your projects/problems must be different from those provided in the sample.*

## Application in Domain of Interest

**Project 1 Description:**

Analyze AirBnB data from the perspective of a landlord. This is a project that I am currently working on using the learnings I have gained from Data Manipulation, Data Visualization and Data Mining courses I have taken so far.

The data was gathered from Inside AirBnB. This site webscrapes AirBnB website for information and compiles it for public usage. The dataset contains listings along with locations, pictures, review ratings that are numerical, information about the host as well as amenities. A reviews file has commentary on the listings. A rent pricing file over the past year is also available.

I would like to be able to be able to identify what makes for the popularity of a rental. In other words, for rentals that have high occupancy rates, what contributes to their being constantly rented.

**Project 1 Problem Type:**

I see this as a classification problem. I could label the properties as highly desirable,  average desirability and not popular based on thresholds of occupancy rates.

I also see this as a reduction problem with several features such as rent, location, review ratings, host ratings and a long series of amenities that contribute to the popularity of a rental and I would like to narrow down the features that lead people to renting the popular listings.

**Problem 2 Description:**

How to set a rental rate for selling a property in the suburbs of Seattle

**Project 2 Problem Type:**

To set the rental rate I would ideally like to do some profiling of the people who tend to rent condominiums  in the town to get a sense of what their typical budget is.

I would then like to conduct a similarity assessment for condominiums in the area as in square footage,  number of bedrooms and amenities like exercise room and so forth.

I would also look into the trends and perform a regression analysis to help predict the rent in the next six months.

I think a combination of the trend data and the budget profile data along with the similarity assessment data will give me a price range the property can be listed at.

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 ethical commitments. You will have to provide one question, one maxim, and one ethical commitment each week.*

*For each, you will provide:*

* *A* ***one-sentence statement*** *of the question, maxim, or ethical 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 ethical commitment. How would it make the particular project go better, or help you avoid some pitfall?*

**Question (I will always ask…)**

**Which Project**

**Meaning in Context**

**Importance**

**Maxim (I will always say…)**

**Which Project**

**Meaning in Context**

**Importance**

**Ethical commitment (I will always/never...)**

**Which Project**

**Meaning in Context**

**Importance**

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?*](https://blog.goodaudience.com/how-do-i-become-a-data-scientist-f8074232608e)*”) 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?”*](https://blog.goodaudience.com/how-do-i-become-a-data-scientist-f8074232608e)

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 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; ethical 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***

* **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.*

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 ethical commitments. You will have to provide one question, one maxim, and one ethical commitment each week.*

*For each, you will provide:*

* *A* ***one-sentence statement*** *of the question, maxim, or ethical 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 ethical commitment. How would it make the particular project go better, or help you avoid some pitfall?*

**Question (I will always ask…)**

**Which Project**

**Meaning in Context**

**Importance**

**Maxim (I will always say…)**

**Which Project**

**Meaning in Context**

**Importance**

**Ethical commitment (I will always/never...)**

**Which Project**

**Meaning in Context**

**Importance**

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 ethical statement*
2. *Map these three insights to the data science project stages framework, as you have in the weekly maxims, questions, and ethical commitment assignments.*
3. *Brainstorm three additional follow-up questions that you would have liked to ask the interviewee.*

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 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; ethical 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***

* ***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.*

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 ethical commitments. You will have to provide one question, one maxim, and one ethical commitment each week.*

*For each, you will provide:*

* *A* ***one-sentence statement*** *of the question, maxim, or ethical 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 ethical commitment. How would it make the particular project go better, or help you avoid some pitfall?*

**Question (I will always ask…)**

**Which Project**

**Meaning in Context**

**Importance**

**Maxim (I will always say…)**

**Which Project**

**Meaning in Context**

**Importance**

**Ethical commitment (I will always/never...)**

**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:

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 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; ethical 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***

* ***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.*

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 ethical commitments. You will have to provide one question, one maxim, and one ethical commitment each week.*

*For each, you will provide:*

* *A* ***one-sentence statement*** *of the question, maxim, or ethical 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 ethical commitment. How would it make the particular project go better, or help you avoid some pitfall?*

**Question (I will always ask…)**

**Which Project**

**Meaning in Context**

**Importance**

**Maxim (I will always say…)**

**Which Project**

**Meaning in Context**

**Importance**

**Ethical commitment (I will always/never...)**

**Which Project**

**Meaning in Context**

**Importance**

**Week 2**

Law of small numbers

**Insight 1**

“ if you follow your intuition, you will more often than not err by misclassifying a random event as systematic.”

Even seasoned experts, well aware of the pitfalls of small sample sets, may be led to deriving conclusions.

Stage: Data Collection and cleaning. Presentation and Deployment?

What it is: Maxim

**Insight 2**

However, sustaining doubt is harder work than sliding into certainty. The law of small numbers is a manifestation of a general bias that favors

certainty over doubt.

Stage: Presentation/Deployment (When we come to a conclusion)

What it is : Maxim

**Insight 3**

“The message about the poll contains information of two kinds: the story and the source of the story. Naturally, you focus on the story rather than on the reliability of the results.”

Check the source. Or every article, visual, experiment check where the data came from.

Stage: Data Collection and cleaning

What it is: Maxim

Selection Bias article

Insight 1

“If you let the subjects of your analyses select themselves, that means that less proactive people will be excluded.”

Data collection for experiments should be truly random and of a sufficient size.

Stage : Data Collection and cleaning

What it is: Maxim

Insight 2:

“ Contemporary Predictive Analytics models work pretty much on the principle of “what happened in the past will happen in the future.” This makes these models very vulnerable. If something new is happening on the market, it’s often not calculated in the predictions and it causes major inaccuracy. ”

The bottom line is: don’t expect a predictive model to be accurate for more than 1 or 2 year.

Stage: Data Analysis and modeling

What it is: Maxim

Data Cleaning 101

Insight 1

* Does the data match the column label?

I have gone through personal experience with this. A form was to be submitted by all client with their cloud service usage. Two column Dual Id and Subscription GUID identified every row of consumption. Both these Ids were a mix of numbers and letter as in AG124 and so on . There was a pattern to distinguish the two , but one company always submitted the form with the deal and subscription id interchanged

Status : Data collection and cleaning

What it is : Question

Insight 2

“Don’t be afraid of picking up a phone or shooting an email to the source of the data.”

For the above scenario and others such as splitting up of consumption allocations across the GUIDS, we had to reach out to the customer since in some instances, they submitting consumption far less than or far higher that reported by an internal system.

Status : Data collection and cleaning

What it is : Maxim

# 10 Rules in Data Science Results

Insights

“Cloud-based source code control systems, such as Bitbucket and GitHub, allow the creation of private repositories that can be accessed by any authorized colleagues.”  
And  
“A version control system, such as [Git](https://git-scm.com/), should be used to track versions of your scripts. “

Version control will help you share you work, allow for collaboration as well keep track of changes that will be important to understand. If you get a different result from your model a deployment, it will be useful to know if it was a data or code issue.

Stage: Data collection and cleaning + Data Analysis and Modeling

What is it: Goal (a goal to set when working on a project)

Insight

“Adding footnotes to the text that reference files or URLs containing the specific data that led to the observation in the report”

And  
“if you are using a tool like Excel to draw your charts, make sure you save the underlying data.“

And

“At the very least you need to document the edition and version of all the software used—including the operating system. Minor changes to software can impact results.”

Document everything.

Stage: Data collection and cleaning + Data Analysis and Modeling

What is it: Maxim