# Personal Manifesto

By: [Simi Talkar]

## Table of Contents

# Week 1: Problem Formulation Stage

## Informational Interview - Planning

**Interviewee prospect 1**

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 metrics 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.

**Interviewee prospect 2**

I also approached Sandeep Pawar who works for Cree lighting. He is a Data Analyst and uses advanced  Power BI in his work. He uses machine learning techniques within Power BI and his blogs have taught me both techniques and Power BI at the same time. Here is a link to one of his blogs that I referred to when in the visualization course. I would like to learn about his switch from Mechanical Engineering to Data Analysis. I also want to know about his daily activities.

This was my letter to both my choices on LinkedIn:

As a part of one of the courses (Being a Data Scientist) we have been assigned to interview a Data Scientist in the field. Are you open to an interview 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 your 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?

7) Tell about the drive that led you to working at Streetlight Data.

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

# Reading Responses

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

  **Insight 1:**
  *"To facilitate such qualitative assessment, the data scientist must think about the comprehensibility of the model to stakeholders (not just to the data scientists). And if the model itself is not comprehensible (e.g., maybe the model is a very complex mathematical formula), how can the data scientists work to make the behavior of the model be comprehensible."*
  How can you overcome the challenge in presenting your analysis and findings so others find them actionable and at the very least relevant?
  **Stage** :Data Analysis and modeling
  **What is it:** Question

  **Insight 2:**
  "*Your model is not what the data scientists design, it's what the engineers build. "*

  Transferring the model design to the engineers is the other end of  of this communication channel. These are the developers and deployers who are working on constructing the model for deployment.
  **Stage** : Deployment (Since the ultimate model is the one built and deployed)
  **What is i**t:  Maxim

  **Insight 3:**
  "*The CRISP cycle is based around exploration; it iterates on approaches and strategy rather than on software designs."*

  There is a different cadence to Data Science projects that typical software development Projects,  the exploratory aspect being paramount.
  **Stage** : Deployment (Since the ultimate model is the one built and deployed)
  **What is it**: Maxim

- **Chris Wiggins interview**
  **Insight 1:**
  The key is usually to just keep asking, “So what?” You’ve predicted something to this accuracy? So what? Okay, well, these features turned out to be important. So what? Well, this feature may be related to something that you could make a change to in your product decisions or your marketing decisions. So what?
  The questions when you analyze your data typically lead to insights but often to more questions.
  **Stage** : Modeling and Analysis/Deployment (cycle)
  **What is it**: Maxim

  **Insight 2:**

  *“The great thing about predictions is that you can be wrong, which I think is hugely important. I can’t sleep at night if I’m involved in a scientific field where you can’t be wrong.”*
  **And**
  “*It takes a long time to become an expert in something. It takes years of mistakes.”*

  This is a field where iterative cycles will offer corrections. And so don’t consider the first iteration to be perfection. This is also what I will find most challenging with my “fix-it” engineering mentality.
  **Stage** : Modeling and Analysis
  **What is it**: Expertise

- **Erin Shellman interview**
  **Insight 1:**

  Erin presents an approach to  find insights in data:

  *"The most interesting types are data are those collected for one purpose and used for another. For example, one of our developers, Jason Wilson, had a really cool idea to look at what was purchased when people asked for gift receipts. Then you could make recommendations for the most gifted products for an upcoming holiday."*

  **Stage** : Problem formulation
  **What is it** : Goal

  **Insight 2:**

  Erin talks about the close relationships the data science team keeps with the web team.

  *"We also work hard to have good relationships with the people who are ultimately responsible for getting our work in front of customers—primarily the web team. So we keep in touch with them regularly and let them know what is going on from our side. . We also make sure that when they find bugs that we respond right away."*
  **Stage** : Presentation and Deployment
  **What is it** : Goal

  **Insight 3:**

  *"However, through the course of our chat we learned that she was not interested in our tool because if a customer she chatted with replenished their product online, our stylist wouldn't be credited with the sale. "*

  Erin's team decided not to build a tool that alerted customers when a product they owned was to be replenished, when they learnt that the stylist would lose her commission if this was built. Feeling  the pulse of the stakeholders and responding to it shows humanity

  **Stage** : Problem formulation
  **What is it** : Ethical commitment

- **Jake Porway interview**
  **Insight 1:**

  "*As a data scientist, you may find yourself running a version of what are essentially psychological experiments on users. That's something that people really need to think deeply about."*

  Jake cautions us against finding ourselves doing the above. I heard on NPR a podcast where some of the social media companies were doing just this!

  **Stage** : Presentation and Deployment
  **What is it** : Ethical commitment

  **Insight 2:**
  *"Another lesson from this project was that we shouldn't underestimate how much little things like that can transform an organization. And so finding out about the data was just a simple analysis that found a problem in data quality."*

  Sometime diligence and good observation is all the analysis you need to perform on the data. When statisticians took a good hard look at the data collected in an experiment they realized there were some who were gaming the system.

  **Stage** : Analysis and modeling
  **What is it** : Maxim

# 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**
    - *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*

    We have performed exploratory data analysis in SIADS 521 and data cleaning in SIADS 505 . We have also been introduced to the various problem types in Data Mining (SIADS 532) course. These have prepped me for initial analysis of data and to ask questions - at least some to get started. I attend twitch sessions with Prof Brooks and then I realize that I had not thought of T-tests between the "differences" in the means of projected and actual cases - and this is the kind of thinking I need to strengthen - applying the statistics to glean more from the data.

2. **a repertoire of problem types**
    - *I already have this capability. If so, describe how you acquired it.*

    The class in Data Mining with Prof Mei and this course SIADS 501 having given me quite a bit of exposure to types of problems. I also watch Power BI webinars and subscribe to webinars from Streetlight data and Microsoft research and they give me some information too. Now it is up to me to absorb and apply.

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

    This is where I need to get loads and loads of different kinds of scenarios to apply the knowledge gained so far. I pick up data sets (like the one I used for writing articles on visualization and analysis on Medium) or in UMSI's NLP MyVoice competition. Then I discuss with a classmate how the data can be viewed and labeled and correlated. Much more practice on this front is required though.

    I am a part of Power BI user groups (PUGs) and these groups consist of people who analyze and visualize data in various domains (typically business applications). I watch what insight they want to gain from a particular dataset. Typically "Ranking" and "Regression" features high on the agenda. TopN ( product/region/department) and Year over Year growth trends as well as Outlier detection (Anomaly detection) is what people are after in these groups. I watch Microsoft Research Webinars as well, along with Streetlight Data webinars.

# Application in Domain of Interest

**Domain:**

Real Estate (ownership) /AirBnB (Rental)

I have been picking up data sets that I am interested in from a variety of open sources after each of the four courses taken in the Masters program so far and have been writing articles in the publication Medium as I apply Pandas and Python visualization techniques to the datasets, combined with my database background.The data set that I picked up recently is from Inside AirBnB that web scrapes AirBnB data. I picked this dataset since my husband and I went through an interesting episode over spring/summer 2020 when trying to sell a condo in a town north of Seattle during the Covid pandemic and when doing so were comparing houses that continued to sell while ours did not. We eventually took it off the market and decided to rent it.

**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 specific popular listings.

**Problem 2 Description:**
How to set a rental rate for selling a property in the suburbs of Seattle. As a real estate agent, what is the fair price range for the property that will attract buyers with the market being on the listings for as short a time as possible.
**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

**Question (I will always ask…)**
Who will be using the results? For what decisions?

**Which Project**
**AirBnB data for rentals**

**Meaning in Context**
The data for the rental is rich with possibilities. There are reviews and ratings, amenities list and host ratings as well as locations. These features can be of interest to a number of parties such as renters, landlords(hosts) who are interested in placing their property out to rent, city planners and regulators - who may face suits for not monitoring the AirBnB rentals sufficiently if issues arise, as well as for members of the community in the neighborhoods.

**Importance**
The data that a host or landlord is interested in is how to equip the property, upgrade it and maintain it and advertise it to attract a renter. The features in the dataset that the landlord will focus on will enable them to answer the questions they are after and have to be extracted from all the data available.

**Maxim (I will always say…)**

The original formulation is rarely the right formulation.

**Which Project**

AitBnB rental project

**Meaning in Context**

I have chosen to ask the question “What makes for the popularity of a rental”. I  would like to know if the question is too broad or too narrow. Determining the scope of the project will determine the time it will take, the resources we will seek to tackle it, the data will will look to gather.

**Importance**

Problem formulation sets the tone for the entire project and view of the data. Once the problem formulation is pinned down it is easier to pick relevant data, model and analyze it. This can, in my opinion, easily become the stumbling block in getting the project started.

**Ethical commitment (I will always/never...)**
I will never allow a bias towards  maximizing rent, enter into the analysis

**Which Project**
Selling a condominium in suburb of Seattle

**Meaning in Context**
In this project, I am trying to determine the optimal and fair market value of the property. it is natural to push the envelope and expect as high a price as is a possibility. But when analyzing the data, a clear unbiased eye is required.

**Importance**
If I let my bias to maximize the profit creep in, then I might ignore factors that set the true and fair value. This will lead to the property remaining unsold on the market far too long. If I were the real estate agent helping a client get the property off their hands in a reasonable amount of time, this would hurt their chances and end up in their incurring losses.

# 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?"*

## 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**