# Personal Manifesto

By: Varshini Rana

## Table of Contents

# Week 1: Problem Formulation Stage

## Informational Interview - Planning

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

- **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

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. In addition to this, I shall also be perusing through newsletters, blog posts, and tutorials from online resources such as Machine Learning Mastery and Towards Data Science to keep myself abreast of the exciting developments 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. My habit of perusing through external resources such as newsletters, blog posts, and tutorials have also helped me understand how to map problems in my application domain to the repertoire of problem types.

## Application in Domain of Interest

**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:** The creation of a regression model to predict the length of stay of a patient in a hospital.

**Project 2 Problem Type (with explanation):** This is a regression problem. The regression model would attempt to predict a numeric value of the number of days a patient would spend in a hospital, termed as “length of stay”, in order to improve hospital care efficiency. The potential data used for training the model would be patient age, gender, history of past illnesses and hospitalizations, previous lengths of stay if any, lab test values and findings, ailment the patient was admitted for, treatments underwent in the past and present, medications taken, etc.

# 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. One reason is because it will help the data scientist empathize better with the people he is trying to help. If he understands the nuances of the problems they face, he would be better prepared to make things easier for them by creating a focused model expressly with the aim of solving their specific issues, rather than trying to solve irrelevant or low-priority problems. Moreover, the problem statement may change depending on who the results benefit. If doctors and radiologists like the model, but think it takes too much time to provide predictions for the model to be of any use in a real world scenario, the problem statement would have to be modified to include a time constraint to benefit them, whereas something like this may not have bothered executive decision makers who would be using and paying the same cost for the services regardless of how much time the model takes to predict something.

**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, which would probably just be published in a research paper and not utilized in the real world.

**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. This would ensure that both the client and the data scientist are on the same page with regards to how the results will be used. This is especially important in the problem formulation stage, since it would kickstart ideation in the right direction. It would guide the data scientist and provide a focus on the specific needs that he has uncovered and aims to satisfy with his work. It would bring about focus and clarity to the problem design space, ensuring that the eventual results generated would be utilized for the purpose it was intended for rather than being written off as a tangent.

**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.

# Week 2: Data Collection and Cleaning Stage

## Potential Personal Project Tweet

“Can I help prevent pneumonia misdiagnoses by guiding doctors with the power of artificial intelligence?” This was the thought that ignited my endeavor to create an image classification model using chest X-ray images for pneumonia detection, which I would obtain from Kaggle.

Character count (including spaces): 275

# Reading Responses

- **Law of Small Numbers**

1. “Sustaining doubt is harder work than sliding into certainty.” Being unsure about something often requires more mental work than definitively coming up with a decision, as our minds have a natural inclination to gravitate towards conclusiveness.
2. “We are far too willing to reject the belief that much of what we see in life is random.” We often try to find patterns in seemingly random events and are unable to accept the fact that most of what we see in life is purely coincidental, leading to preconceived notions.

- **Statistical Biases Types Explained**

1. “Biased statistics are bad statistics.” Statistical analyses done with biases are not going to be generalizable for the cohort the study was intended to gain insights into, since the underlying assumptions the study was based on itself is erroneous.
2. “If you let the subjects of your analyses select themselves, that means that less proactive people will be excluded.” The subjects who volunteer for the analysis exhibit a certain type of behaviour which may not be representative of the entire population under scrutiny for that particular study.

- **Data Cleaning 101**

1. “Unless you have a good level of confidence in your assumption, you probably should not make the correction or else you may be creating a typo instead of correcting one.” While corrections we make in the data before using it may be well-intentioned, we shouldn’t be making undue modifications without certainty just because we can, as doing this may defeat the purpose of it by introducing erroneous assumptions.
2. “Don’t be afraid of picking up a phone or shooting an email to the source of the data.” It is necessary to be absolutely sure about the authenticity or correctness of the data your project will be based off of, which is why the process of reaching out to the sources of the data to confirm this is paramount.

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

1. “All data science is research.” All science requires systematic inquiry and investigation to generate insights; data science encompasses the same endeavor, but with data.
2. “Manual data manipulation is hidden manipulation.” Manual data manipulation steps often aren’t reproducible. They are difficult to document, track, and repeat. Hence, they should be avoided and a script with general data manipulation steps can be used instead.

# Plan for Knowledge Acquisition

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

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

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, where one of my responsibilities was to develop a fraud detection product in the domain of Health Insurance using structured claim data. The product consisted of a classification algorithm which can predict whether an insurance claim is “fraudulent” or “legitimate”. The initial results I obtained were extremely misleading. On closer inspection of the data, I realised that one of the features used for model training was a field for the “balance sum insured”, which I would typically not be aware of before a claim has been processed. By including this field, I was essentially “cheating” since “fraudulent” claims in the training data had a higher “balance sum insured” value than the “legitimate” claims due to not being used up since these claims were not successfully processed. Another issue I encountered with the data on initial inspection was that all the columns in the tabular data had been shifted by one column to the right, and so all the fields had incorrect headers. Extraneous commas in the CSV file were the reason for this issue and were promptly fixed. Other issues, such as missing values, duplicates, and inconsistencies within fields are all extremely common based on my experience.

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

I already have this capability. During my tenure of two years as an Associate Data Scientist in a start-up, I was encouraged to explore and obtain structured data from the in-house OMOP Common Data Model, hosted on AWS Redshift, which contained anonymised healthcare data (received from clients) devoid of identifiable information. I learnt how to query the databases with SQL via tools such as Aginity Workbench and SQL Workbench, and export the structured data which I required for my analyses in the form of CSV files. I was also able to programmatically access and connect to these databases with Python libraries such as psycopg2. Moreover, I was able to use proprietary APIs to download unstructured data received from clients in the form of bills, lab reports, discharge summaries, and radiology reports as PDF files for my analyses. Externally, I have used and am familiar with several data sources in my application domain, such as Kaggle, data.gov, and the UCI Machine Learning Repository.

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

I already have this capability, owing to my tenure of two years as an Associate Data Scientist in a start-up, wherein I spent a huge chunk of my time scrutinizing datasets, playing around with them, and trying to properly understand them before going ahead with my analyses. I have found that having a high-level overview of the data types and number of values of all of the fields in my datasets go a long way in determining whether everything is as expected. With pandas, I would just use a simple function such as pandas.DataFrame.info(), which would give me a quick summary of the data types and also the number of values (and missing values, if any) of each column in the dataset. This would help me figure out to some extent if each column has the intended values, and what intervention I would need to bring about, if any. Using a function such as pandas.DataFrame.head() would also show me the first 5 rows (along with column names) of the dataset, so that I can quickly see whether everything is as it should be. Similarly, other pandas capabilities would help me summarise, aggregate, and visualise the data so that I could get a glimpse of the data from other perspectives. During the course of my work, I have also had to document some datasets which were given to me by the client, so that others in my organization could use it for their own purposes. I have had to include comprehensive descriptions of the fields in the datasets, along with short accounts of how they were collected, and the possible values (or range of values) that would go into each field. I have also had to include a Data License which would authorise my associates to use this data with the client's consent.

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

I already have this capability. I acquired this capability during my tenure of two years as an Associate Data Scientist in my start-up. I obtained structured data from the in-house OMOP Common Data Model, hosted on AWS Redshift, which contained anonymised healthcare data (received from clients) devoid of identifiable information. I learnt how to query the databases with SQL via tools such as Aginity Workbench and SQL Workbench, and export the structured data which I required for my analyses in the form of CSV files. I was also able to programmatically access and connect to these databases with Python libraries such as psycopg2. The Python requests module helps me to connect to proprietary APIs to download unstructured data received from clients in the form of bills, lab reports, discharge summaries, and radiology reports as PDF files for my analyses. I am also familiar with a pandas functionality pandas.read_html() which can extract relevant tables from the webpage in the URL that would be provided as an argument to the function.

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

I already have this capability, owing to my tenure of two years as an Associate Data Scientist in a start-up, wherein I spent a huge chunk of my time scrutinizing datasets, playing around with them, and trying to properly understand them before going ahead with my analyses. I am quite used to finding missing values, duplicates, and inconsistencies in datasets, and employing pandas functionalities such as pandas.DataFrame.dropna(), fillna(), drop_duplicates(), etc. to clean them efficiently. I am also well-versed in feature engineering, which involves creating new features from existing ones, and dropping redundant features. An example of this during the development of a fraud detection product was that I created a feature called "length of stay in the hospital” by computing the difference of two features — “admission date” and “discharge date”. I have also employed Principal Component Analysis (PCA) and feature selection techniques such as forward selection and backward elimination to arrive at the most optimum set of features sans redundancy for model training.

## Maxims, Questions, and Commitments

**Question (I will always ask…)**
How much is too much data cleaning?

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

**Meaning in Context**
The entire premise of the image classification model rests on the model being able to detect microscopic subtleties in chest X-ray images, which would otherwise be invisible to the naked eye. While we have a variety of computer vision techniques at our disposal to augment the images, sharpen them, modify the contrast, brightness, rotation, etc. while cleaning or correcting the images, we should be wary of knowing when it becomes too much, as these actions could have unintended repercussions and significantly affect the results, i.e. the microscopic subtleties in the X-ray images could get drastically altered and give the wrong impression.

**Importance for this stage of the project**
Data cleaning should be an extremely conscientious process. It should be done only when and only to the extent that there is some certainty that the resulting dataset would be accurate, free of bias, and without inconsistencies. This is done not only to ensure that the data is easy to work with and train, but also to have some level of confidence in the results of the model trained on it. Just because we have all the capabilities to correct or clean data, it doesn’t mean that we absolutely must use all of them, at all times, and in all situations.

**Maxim (I will always say…)**

Manual data manipulations are not always reproducible.

**Which Project**

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

**Meaning in Context**

Suppose I browse the dataset of chest X-ray images and find a bright red spot on one of the images. I would be tempted to remove this anomaly right then and there using an image editing tool, such as Photoshop, since it's quick and convenient. However, I may have missed fifty other images with similar anomalies since I didn't notice them while idly browsing through the images. This would significantly mess up my classification model as the wrong conclusions would be drawn. Moreover, any person picking up this project after me would be unaware of this anomaly since the correction was undocumented and done on the fly. Hence, it is necessary that any changes I make be done using a scripting language for better coverage, ease of documentation, and reproducibility.

**Importance for this stage of the project**

It is important to ensure that any modifications I make to the raw data at the data collection and cleaning stage of the project be reproducible so that there is some semblance of consistency in the process should it be applied on a repeat basis, and also to enable ease of documentation for the next person to follow, something that would next to impossible were the modifications done manually on the fly. Using a scripting language would help me with this and also make the data cleaning step generalizable for that particular type of data and the issues commonly observed in that data.

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

I will never settle for working with data whose only merit is that it just happens to be easily available, even though it isn't representative of the population I'm considering for my study.

**Which Project**

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

**Meaning in Context**

Pneumonia could be of three different types — bacterial, viral, or mycoplasma pneumonia. Suppose I obtain data just for bacterial pneumonia and healthy control, and fail to find data for the other two types. While it was convenient to have this sample of pneumonia, it isn't representative of the entire gamut of pneumonia conditions which exist in reality. This could bias my model into detecting only bacterial pneumonia and failing to detect the other two types, which could lead to erroneous interpretations of whether a patient has pneumonia or not.

**Importance for this stage of the project**

Knowingly collecting only a specific subset of the population as a sample just because it was easy to access points to an underrepresented population, which could drastically affect the results of the model the data would be trained on when applied in a real-world scenario, and defeat the entire purpose of its development and deployment. Hence, it is important to avoid selection bias by either working to collect the other known representatives of the population, or to modify the problem statement to include and predict for only the types of data you have (for example, in this case, the labels for classification can be "bacterial pneumonia" and "healthy control", as opposed to just "pneumonia" and "healthy control"), in which case the results may or may not be useful depending on the situation.

# Week 3: Data Analysis and Modeling Stage

## Informational Interview - Reflection

The interview transcript of Eric Jonas was collected from the book Data Scientists at Work by Sebastian Gutierrez. Eric Jonas is a research scientist in computational neuroscience and signal processing. He is currently a Professor in the Department of Computer Science at the University of Chicago.

What intrigued me most about Jonas is that he is a forward-thinker. He is always on the lookout for the next problem that would keep him on his toes, and the problems that he chooses to contend with are often those which would have far-reaching consequences into the future. As a budding data scientist, this attitude is something that really inspired me into adopting this way of thinking.

As an industrious person with extensive experience not only in academia but also in the industry, Jonas provides a unique perspective of what success looks like from the lens of a PhD student, an entrepreneur, and an academic researcher. While the nuances differ, he maintains that it all really boils down to "what value are we providing?" **[question - W1 problem formulation]**. This added value could be monetary, related to ROI, something technical such as improving the predictive accuracy of a model for purposes downstream, or even an idea of what necessitates this venture in the first place, which is important to consider before choosing to work on a problem.

On being asked whether Jonas approaches a data problem model-first or data-first, Jonas insists on not going ahead with the project until he familiarises himself with the data first. He believes that knowing the right questions to ask helps in understanding the data **[maxim - W2 data collection and cleaning]** well enough to start thinking about further analysis and figuring out which modeling approach to employ. He believes that with modeling it is easy to change the approach and try different things, but if the data hasn't even been looked at and played around with first, it is easy to be led astray with fallacious assumptions.

Jonas also is of the opinion that a data scientist cannot just be someone who comes in, does the math, and leaves. He professes that the data scientist actually needs to care about the domain of the project **[professional commitment - W1 problem formulation]**. He states that a data scientist who just wants to apply the tools he has to solve the problem in a domain, but doesn't really have the knowledge or care to learn about the domain of its application would not only have a poor understanding of the problem, but would also not be willing to make the compromises necessary to understand how to guide his own work.

Three follow-up questions I would have liked to ask Jonas are:

1. Hypothetically, if you weren’t a man of science, what would you be doing professionally instead?
2. What is the best challenge that you’ve faced so far, and how did you overcome it?
3. What is one advice that you would give to your past self of ten years ago?

Word count: 496

# Reading Responses

- **Overfitting in Machine Learning: What is it and how to prevent it**

1. “Noise interferes with signal.” The machine learning model would start thinking of the irrelevant information (i.e. noise) in the dataset as a significant pattern and would make erroneous generalizations from it by incorporating it into its calculations, masking the actual insights which could be gleaned from it.
2. “Simple learners tend to have less variance in their predictions but more bias towards wrong outcomes.” Algorithms with a simple, inflexible underlying structure tend to train models which are consistent but are more likely to be inaccurate on average.

- **Common pitfalls in statistical analysis: The perils of multiple testing**

1. “In any study, when two or more groups are compared, there is always a chance of finding a difference between them just by chance.” When it is hypothesized that no difference exists between the groups being compared (i.e. the null hypothesis), there is some probability of falsely rejecting it (i.e. the False-Positive Error rate or Type 1 Error) by finding some difference between them by chance.
2. “Results from single studies should not be used to make treatment decisions.” The researcher should try to gain insights from supporting data and other studies published by the scientific community to validate the results of his study for his insights to be actionable.

- **P-Hacking and the problem with Multiple Comparisons**

1. “So, to me the answer is (to) replicate yourself.” The initial results of the study, deemed to be exploratory and reported as such, can be attempted to be replicated with a new sample for validation and to exhibit the robustness of the results.
2. “Just remember–fishing is fine for tuna, bad for data analysis.” Seeking to obtain a particular result from data analysis which is unfounded or statistically unsupported in practice is a disingenuous way of “hacking” or “gaming” the system, which leads to unsubstantiated assumptions.

- **Correlation vs. Causation: An Example**

1. “Without randomized controlled trials, we cannot say one activity caused another.” A randomized controlled trial would include a randomized population being subjected to the activity under scrutiny and a control population to determine significant differences between the two groups. By not conducting this analysis, we cannot say indubitably that one trend causes another. At the most, we can claim that the trends are correlated.
2. “Humans naturally see patterns where they don’t exist, and we like to tell a cohesive story about what we think is going on.” We often try to find patterns in seemingly random events. We tend to crave structure and reject chaos in favour of something explainable, regardless of whether it is an incorrect assumption to make.

- **Simpson's Paradox in Real Life** or **Ignoring a Covariate: An Example of Simpson's Paradox**

1. "If the populations are separated in parallel into a set of descriptive categories, the population with higher overall incidence may yet exhibit a lower incidence within each such category." Although the overall performance metrics of a population considering all categories under scrutiny may be high, the performance of the individual categories may be lower than those of the other population being considered in the study.
2. "Simpson's paradox is not a contrived pedagogical example." The Simpson's paradox is something that occurs in real-life situations and is not just something that materialized as a teaching contrivance.

- **Conditioning on a collider**

1. "Causal inference from observational data boils down to assumptions you have to make and third variables you have to take into account." Observational studies cannot prove cause and effect by themselves, unlike randomized controlled trials. Hence, certain assumptions and other variables have to be considered while attempting to answer causal questions with only observational data.
2. "If you really care about a cause, don't give mediocre studies an easy time just because they please you." Don't give a lot of weight to studies that seem plausible on the surface but are actually just glorified studies which condition on colliders, which instigate a misrepresented association between two entities where none exists.

# Plan for Knowledge Acquisition

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

1. **common mistakes in data analysis that lead to misleading results**

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, where one of my responsibilities was to develop a fraud detection product in the domain of Health Insurance using structured claim data. The product consisted of a classification algorithm which can predict whether an insurance claim is “fraudulent” or “legitimate”. The initial results I obtained were extremely misleading. On closer inspection of the data, I realised that one of the features used for model training was a field for the “balance sum insured”, which I would typically not be aware of before a claim has been processed. By including this field, I was essentially “cheating” since “fraudulent” claims in the training data had a higher “balance sum insured” value than the “legitimate” claims due to not being used up since these claims were not successfully processed. I also made the rookie mistake of including all the available features in the training of my model, not realising that they were introducing a considerable amount of noise into the system, which resulted in overfitting. Some well-planned feature engineering and dimensionality reduction techniques fixed this issue. In a different project involving 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, I initially made the mistake of excluding outlier values based on misguided assumptions, which for this problem, had dire consequences. I realized that outliers need to be considered in the broader context of the analysis depending on the problem at hand.

2. **a repertoire of models and how to estimate, validate, and interpret each of them**

I look forward to strengthening this capability. I’m still at the budding stage of my Data Science career with exposure to mainly Classification (or Class Probability Estimation) problem types, which in turn exposed me to a variety of classification algorithms such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, K-Nearest Neighbors Classifier, Support Vector Machines, etc. Naturally, I learnt how to effectively interpret, estimate, and validate them in various ways. While I am aware of the existence of other model types, I do not have hands-on experience with training, evaluating, and interpreting them. I plan to acquire this hands-on experience 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, which shall naturally expose me to a variety of models. I also plan to develop my repertoire of various models through MADS coursework, which includes courses such as SIADS 542: Supervised Learning, SIADS 543: Unsupervised Learning, SIADS 642: Deep 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 models and algorithms in the field of Data Science. In addition to this, I shall also be perusing through newsletters, blog posts, and tutorials from online resources such as Machine Learning Mastery and Towards Data Science to keep myself abreast of the exciting developments in the field of Data Science.

## Maxims, Questions, and Commitments

**Question (I will always ask…)**
What is the best algorithm to use for this problem?

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

**Meaning in Context**
There are many things to consider while choosing an appropriate algorithm in the context of this project. The underlying structure of the algorithm used in training the model should be such that it is able to support and work well with images (radiological images of chest X-rays in this case), and effortlessly detect subtleties in them which are invisible to the naked eye. Neural networks, such as Convolutional Neural Networks (CNNs) are the widely-accepted standard for image classification. Then comes the question of the architecture of the CNN, which we can build ourselves. If we are using one of the many pretrained models out there, such as VGG19, ResNet50, etc., we should think about fine-tuning them with our own data for more accurate predictions.

**Importance for this stage of the project**
After familiarizing ourselves with the data, the next step is to think about what the best approach would be in the context of analysis and modeling. When thinking about modeling, knowing to employ the right algorithm to train our data on (keeping in mind the capabilities of the model and the bias-variance trade-off) is paramount, as it would not only streamline our efforts in the right direction (as evidenced by similar efforts by other members of the scientific community trying to solve similar problems), but would also help us anticipate the level of effort we can expect to put in to solve the problem at hand. It would also give us an idea of the effort taken to evaluate and explain the results in a manner that actually answers the question that the project was intending to address.

### Maxim (I will always say…)

Fishing is fine for tuna, bad for data analysis.

### Which Project

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

### Meaning in Context

Suppose I split up my training dataset of chest X-ray images into multiple smaller datasets, each with something different being done to them under the guise of “exploratory analysis”. For example, I could enlarge the images of one set by some factor, I could rotate the images of another set by some degree, I could completely change the brightness of the images of a set such that it is completely different from that of the other sets, and so on and so forth. I could then train a model with all of these different datasets in turn hoping for a signal or some significant finding to pop up from at least one of them. These findings are not something I planned for, and are most likely spurious since I was “fishing” for them.

### Importance for this stage of the project

By splitting up the dataset into smaller different datasets with distinct characteristics, I am likely to find something statistically significant from at least one of the smaller datasets. However, we would have no way of knowing if these results were due to the change I made on the dataset to separate it from the other ones, or merely due to an artifact of the sample, or a legitimate judgment call I made after talking to subject matter experts. Since I was “fishing” for some finding, any finding, without knowing what to expect beforehand, I could not say with any certainty that the results I obtained were completely within expected parameters. Such an approach would render my analysis immaterial as it would lead to fallacious assumptions.

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

I will always use hold-out test datasets and cross-validation to mitigate the problem of overfitting.

**Which Project**

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

**Meaning in Context**

The image classification model I would create should be generalizable to make accurate predictions on real-world situations, i.e. with chest X-ray images never seen by the model before. Creating a hold-out test set and using cross-validation would help me validate the model and would give me some confidence that the model has not memorized the noise in the training dataset so much so that it provides wrong predictions for images that it hasn't been exposed to while training.

**Importance for this stage of the project**

The results of the model created would likely lead to an intervention of some sort, as it identifies potential patients of pneumonia who would then be admitted, subjected to further tests and treatment, etc., which results in considerable expenditure of time and money for all parties involved. Hence, it is necessary to ensure that the predictions are as accurate as possible. Overfitting would result in the model only making accurate predictions for the images in the training dataset and giving incorrect predictions for most of the images it has never encountered before. Therefore, it is essential to prevent overfitting by using techniques such as the usage of hold-out test sets and cross-validation to mitigate this problem.

# Week 4: Presenting and Integrating into Action

## Sources for Data Science News

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

- Towards Data Science: I have found that it is a wonderful place not only to learn about new tech in data science but also to gain valuable insights from leaders in the industry who show us a glimpse into their day-to-day lives as well as their overall careers as a data scientist through their blogs.
- Machine Learning Mastery: This goldmine of a website containing easy-to-follow-along tutorials is my go-to resource whenever I want to know why or how to apply a data science concept to solve the data problems I encounter at my workplace, and also just to pick up some nifty tricks when I'm bored.
- KDNuggets: This is a one-stop shop for all of my data science news, dataset, and data science tutorial needs. This resource also offers various free courses one can participate in to augment their knowledge of data science concepts.
- Data Science Central: This is probably one of the most diverse data science blog websites I've come across, including tutorials and tips for applying data science concepts in various programming languages apart from Python such as R, SQL, Julia, etc. and blogs talking about everything from data science to big data to business analytics. It would be a great place to broaden my horizons in terms of thinking beyond regular Python for data science.
- r/datascience and related subreddits: I am quite active on Reddit and have found subreddits (communities within Reddit) such as r/datascience, r/MachineLearning, r/learnmachinelearning, etc. to be extremely helpful whenever I've had a question related to data science concepts. The users of the community are also quite prolific in showcasing their up to date knowledge of the industry and I get regular notifications of these posts, which helps me to keep myself up to date about the exciting developments in the field of data science.

# Reading Responses

- **A History Lesson On the Dangers Of Letting Data Speak For Itself**

1. “All great truths begin as blasphemies.” It is human nature to be resistant to new ideas and changes. We are often reluctant to go against what we’ve always incorrectly accepted as the truth in order to embrace the actual, evident truth.
2. “Know your audience and strive to understand their existing attitudes and beliefs.” In order to connect with the person you’re telling your story to, you have to understand their beliefs and convey your story in a manner that they can relate to.

- **Storytelling for Data Scientists**

1. “We talk, think and dream in stories.” Human beings have a tendency to create stories out of their experiences and perceptions; the decisions they make have a strong emotional basis rather than a purely rational one.
2. “Keep it simple.” Dwelling too much on the technical details might ruin the point you’re trying to make with your story. Focusing on the message you’re trying to bring across in simple terms would go a long way in convincing your audience about your solution to the problem.

- **Interpretability is crucial for trusting AI and machine learning**

1. “The algorithms inside the black box models do not expose their secrets.” Some models are very hard to interpret since they often involve many predictors, parameters and complex transformations, which makes it difficult to figure out why they made a certain prediction.
2. “If we consider that the data in the production environment is not stationary, it can become outdated very quickly.” The data that was used to train a model in its static state may actually be dynamic in nature in real-world scenarios. Hence, the model trained on it may not generalize well in the production environment and will require retraining.

- **The Signal and the Noise, Chapter 2**

1. “Instead of spitting out just one number and claiming to know exactly what will happen, I instead articulate a range of possible outcomes.” A range of outcomes, a probabilistic response, is an honest representation of uncertainty in real-world scenarios.
2. “Our brains, wired to detect patterns, are always looking for a signal, when instead we should appreciate how noisy the data is.” We often try to make mountains out of molehills with the data we have at hand, refusing to recognise noisy data which would lead us to jumping to misleading conclusions.

- **The Signal and the Noise, Chapter 6**

1. “Political polls are dutifully reported with a margin of error, which gives us a clue that they contain some uncertainty.” A margin of error represents a measure of sampling error in

the polls, which expresses some measure of the uncertainty of the results of the polls, rather than reporting just a single number which would create the perception that the results are extremely accurate when statistically, they won't be.

2. "A probabilistic consideration of outcomes is an essential part of a scientific forecast." There could be multiple possible outcomes in a poll depending on various situations and extenuating circumstances. Hence, providing a range of probabilistic outcomes rather than just one is imperative when conducting a scientific forecast, as this would provide an idea of the uncertainty involved in the event in question, based on which appropriate decisions can be made.

- **How Not to Be Misled by the Jobs Report**

1. "But what if all the worries were based on nothing more than random statistical noise?" Putting too much weight on misleading data (such as data inundated with sampling errors, which is often the case in surveys where individuals volunteer to participate) could lead to groundless generalizations, which could lead to unwanted consequences down the line.
2. "Human beings, unfortunately, are bad at perceiving randomness." Human beings are notorious for trying to find patterns in random events, just to be able to explain the occurrence of the event in a way that makes sense to them since complete randomness is often unfathomable to them.

- **But what is this "machine learning engineer" actually doing?**

1. "When the technology is new, all you are left with is poor documentation and a bunch of blog posts." When the technology is new, not a lot of people have had the opportunity, means, or motivation to use it, resulting in fewer resources and documentation that we can refer to pertaining to it.
2. A person called a Machine Learning Engineer "abuses machine learning libraries to their extremes, often adding new functionalities." A machine learning engineer often is actively involved in customizing standard machine learning frameworks by adding their own tweaks in order to fit the use cases being looked at by the data science team.

- **How we scaled data science to all sides of Airbnb over 5 years of hypergrowth**

1. "Data isn't numbers, it's people." Data often reflects the decisions that people make. A data scientist who can recognise this fact is able to define how to think about the data as well as the cultures and perceptions of the people who constitute the data, which helps ingrain data science in business functions.
2. "Beyond individual teams, where our work is more tactical, we think about the culture of data in the company as a whole (by) educating people on how we think about Airbnb's ecosystem…" A data scientist needs to have a good overall understanding of the domain of the data he is expected to work with, which helps him to better leverage the data to solve problems in that domain.

# Plan for Knowledge Acquisition

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

- **how to present results to domain experts who are not data scientists**

I already have this capability. Having worked in a start-up environment for two years as an Associate Data Scientist, I had to take input from and present the results of my data science experiments to domain experts such as former practicing physicians and health insurance fraud investigators. During the early days of my career, I shadowed my supervisors to see how they tackled the concerns and questions of these domain experts, and eventually was able to tackle them on my own. I learnt the valuable skill of presenting my ideas and findings in simple terms without going into the nitty-gritty technicalities. Once the domain experts understood what I was trying to accomplish, the limitations of the tech we had at our disposal, and the limitations of what we can accomplish with the resources we have bearing in mind the uncertainties in real-world scenarios, their suggestions became far more focused and refined, which is how I knew I had succeeded in what I wanted to convey. Essentially, I was helping them help me.

- **how to work with software engineers to put models into production**

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up. During the course of my career, I trained various models, created pickle files out of them, and pushed my code into a GitLab repository by following a Sprint plan. The software engineers would then take that code and write a wrapper on top of it to productionize it and create an API out of it for clients to use. I would work closely with them during this stage to make sure that we're all on the same page on what exactly should be expected as the output from the model, and troubleshooting any issues that arise during the course of and after the model has been productionized. Some of their inputs often involved the setting of particular coding standards I had to adhere to in order to make the process efficient and also specific library versions I would have to use so that there is no mismatch between those of the development environment and the production environment. I would also modify my code based on their suggestions so that it is easy to scale-up the code based on client requirements.

## Maxims, Questions, and Commitments

**Question (I will always ask…)**
Is this a situation where we need an explainable model?

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

**Meaning in Context**
Based on the predictions of this model, an individual could either be hospitalized for pneumonia treatment (if the model predicts that the individual has pneumonia) or could be deemed as not requiring pneumonia treatment (if the model predicts that the individual does not have pneumonia). Since high-risk decisions are likely to be made based on the predictions of the model — the exact weightage of which would be decided by domain experts depending on the situation for decision-making — it is necessary to have some means by which to explain how the model arrived at that prediction so that domain experts are able to attribute some confidence to the predictions. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) would help in this case.

**Importance for this stage of the project**
I believe that it is necessary to consider the explainability of a model especially if the use case for it involves high-risk decisions based on its predictions. This is specifically important to consider at this stage since these predictions need to be communicated to the domain experts and decision makers in order to take appropriate action from it, for which they would require some concrete explanation as to why or how the model arrived at the prediction. In this case, considering worst-case scenarios, if the model predicts that a person had pneumonia when in fact he doesn't (i.e. a false positive), he could end up being hospitalized for no reason, resulting in a waste of time and money for all parties involved. On the other hand, and worse still, if the model predicts that a person does not have pneumonia when in fact he does (i.e. a false negative), he would end up not receiving the treatment he needs, resulting in severe health complications and other bad outcomes. Blindly trusting the predictions of a black-box model could have dire consequences.

**Maxim (I will always say…)**
The data doesn't speak for itself.

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

**Meaning in Context**
We need to leverage data to tell a story with it. While presenting a bunch of facts, figures, and fancy diagrams might appeal to quant-savvy folks, this may not pass muster with the decision makers and domain experts who actually intend to make use of the insights gleaned from the data. For example, in this case, I could just tell them that the precision of my model is 96% and the recall of my model is 94% based on evaluations made on a hold-out test dataset. While this could be considered impressive from a technical standpoint, this is really not helpful when I put myself in their shoes. They would not know the meanings of these terms, or the context of why these terms are used. They would not even know whether this was a good result and something worth investing in and making the important decision of hospitalization over. Hence, it is necessary to present my findings to them in a way that establishes the context of the problem, to what extent I supposedly solved it, and what could be expected in a real-world scenario with well-defined margins of error in simple terms so that they are aware of all of the pros and cons of using my model for decision-making.

**Importance for this stage of the project**
When high-risk decisions are involved based on predictions of the model, it is important to control and direct the narrative of the data in the right direction while presenting the findings to the decision makers and domain experts so that the right actions can be taken based on it. This would not be possible if the results are presented solely as facts and figures. A data scientist's job isn't done after just training the model and deploying it. He needs to ensure that these results are presented in a way so as to appeal to the rational and also the emotional side of decision makers who are the end-users of the model, so that they are able to leverage the functionalities of the model to its fullest based on their own experiences and judgement.

**Professional/Ethical commitment (I will always/never...)**
I will always support and help software developers do their job effectively.

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

**Meaning in Context**
The model that I trained and tested would need to be deployed in a production environment by software developers (or machine learning engineers), who would write the API code for it. If a domain expert, such as a doctor or radiologist wants to consume the predictions of the model I created, he would have to request the API for a response (which would be something like: “person has pneumonia” or “person does not have pneumonia”, or a probability estimation for each class) after feeding it the input data. By supporting the software developers during productionisation, taking an active part in diagnostics in case something goes wrong, and cooperating with them by writing clean, efficient, and scalable code, I would be able to ensure that the deployment of the model and its maintenance thereafter happens smoothly and without hitches.

**Importance for this stage of the project**
My job as a data scientist does not end when I train and test a model, and send the software developers my code. Complete and explicit cooperation among different members of the tech team is vital for the smooth release of a product (which is the productionized model in this case). To that effect, after I have created a working model, it is my responsibility to help the software developers with the smooth integration of the model within the overall tech ecosystem. This is because I would know the details of the model intimately since I created it and can fill any gaps in understanding among the software team. I would also be able to assist in any unprecedented eventualities that may arise in the production environment. I would also support them by modifying my code based on their suggestions such that it is scalable, maintainable, and debuggable. This would ensure a hassle-free deployment of the model for the client’s use.

# Document update information

Updates to this document after the start of class: