Jacob Paul

Problem Set 4

**Monday:**

1. By separating the two responsibilities, it is possible for the data scientist to avoid having any bias or coming with inferences about the data or the results. This could be considered a benefit in some cases. However, more commonly we see that these two responsibilities are both carried out by the data scientist and they usually have to gain some background knowledge in the field that the data comes from. Naur might’ve framed data science in this way because there wasn’t much crossover in professions or it wasn’t normal to have knowledge in more than one profession.

**Wednesday:**

1. In the 1970s, data science mostly focused on statistics that could be derived from data. Through the 80s and 90s more context was added in that data science encompassed more than just the statistics. As datasets grew, more data cleaning and context was involved. Data scientists were required to gather more context for the data and be able to understand it better. Into the 2000s and now, data science became much more focused on the technology that can help us achieve insights such as machine learning, AI, and how to implement them through tools like Python. Being able to visualize and tell stories with data also become a significant aspect of data science around this time.
2. 1. The advent of the Internet enabled data to be gathered on a much larger scale and for companies and individuals alike to be able to gather and access data remotely on many people. Storage and computing power has also gotten exponentially cheaper in the past two decades so the ability to store and process this data has encouraged organizations and individuals to gather more data because the means to do something with the data has become more accessible. There are also more tools being released and becoming more accessible, which allows organizations to make use of the data without having substantially higher expertise. Data also makes companies more competitive and the difference between harnessing this data and not could mean the difference between being pushed out of the market and not.
   2. Recently, more and more products connect to the Internet including our phones, TVs, and watches. These IoT devices introduce many more data sources that were previously unheard of or hard to gather. For example, it is now possible to track a person’s location in near real-time because of the internet connection and GPS that a smartphone provides. Sensor technologies have also gotten smaller, cheaper and more efficient. This makes it easier to have sensors in previously unheard of places, like our pockets or up in the sky.

**Friday:**

1. One common and accessible data collection method is the survey. These are often used to gather the opinions of a group of people. Participation is related to how many people the survey is distributed to and how many of those people choose to participate. Creating a survey is an involved process as there are many principles to follow when designing questions and disseminating the survey. Overall, there is a lot of effort required to get an acceptable response rate for surveys. Web scraping, on the other hand, is a way to capture data that already exists on the internet and put it in an easy to analyze format. This data collection method can be tricky as some websites have rules against web scraping. Similar to surveys, data collected using this method can be messy and can take a lot of work to clean up. Finally, interviewing is the most analog method out of those mentioned and is the best for collecting the thoughts and emotions of other people. It is also the most exploratory in the sense that it allows a researcher to deviate from the guidelines that may be present with something like a survey. Interview probably takes the longest to analyze and sort through out of these methods. Interviews must be transcribed, coded and then analyzed alongside with the other interview in the project.

If I were helping out the state or the city to better understand traffic problems in Boulder, I would collect different kinds of data depending on what the stakeholders request. If they wanted to know what people thought of a certain intersection or street, I might go out to that place and interview people there to gather their candid responses about the issue. This would give me a real look into the emotions of those citizens that wouldn’t come through in traffic data. However, if I wanted to better understand the flow of traffic, I might look at some traffic data to derive some patterns. If I wanted to get an idea of the feelings and sentiment about traffic in general in Boulder, I might use a survey to cast a wide net a get a better variety of responses than pavement-pounding interviews.