Problem Set #4

1. **Peter Naur is famously quoted as saying data science *“deals with the data, while the actual relation of data to what they represent should occur in other fields.”*What might be problematic in this statement? Why do you think he’d choose to frame data science this way?**

I think what Peter Naur is saying is that data science the actual field. For example data scientists should focus on the collection of the data and then specialist in whatever field it is should be separated from who connect the dots and come up with answers to whatever the data seems to represent. He thinks we should not integrate data science and interpretation of the data collected. This should be done by people in other fields, collecting data and interpreting data should be two separate entities. So essentially data science should be a co-producing job and data scientists should have to do all facets of the job. The problem with this is that data scientists may no have the expertise in whatever field it is and vice-versa. I believe that whatever approach is more efficient and accurate should be used in the future.

1. **There was a substantial shift in the ways we define data science between the 1970s and the early 2000s. Describe this shift and why it may have emerged.**

Part of the reason there was a shift was because of the annual conference KDD created in 1989. Essentially data scientists began realizing that collecting data isn’t enough but interpreting it and knowing how to use it is just as important. In the early 2000’s the world became more connected due to the prominence of the Internet and made it easier to collect data and work together with other data scientists and people in other fields. Data is really powerful and during this time span businesses started realizing that they can use it to improve their performance. Companies began hiring people who can actually make sense of the ubiquitous data and data science job listings increased by 15,000%.

1. **The idea of "big data" dominates much of modern data science. However, data is still growing at an exponential rate.**

**A. What factors do you think may have led to this growth? Mention at least three and describe why they have contributed to recent explosions in data volume.**

1. I think that open data and the ability to add onto data that has been done by other people have made a really big difference. Instead of data scientists constantly having to start their own data collection they can now build on other people’s ideas and data.

2. Another big factor in the explosion of data use are big companies realizing the importance of it and hiring people with those skills. A lot of businesses are moving towards statistical changes within their companies and relying on data to predict and create what they need to succeed.

3. Storage and computing power has gotten a lot stronger and cheaper which allows us to collect data and hold for decades. Data in the past could easily be lost.

**B. Where is this new data coming from?**

Data is so ubiquitous that essentially everything can be used for it. Whether it’s data collected from students who bike or from ants in a certain area. Businesses, institutions, governments, and many more all use it to predict things and understand what is going on around us. Companies like facebook and twitter are constantly collecting data about their users and selling it to other businesses. It’s amazing how a company like facebook that is free for use is worth billions of dollars just by collecting data.

**4. Name three different data collection methods. How are they similar? How are they different? Consider using specific scenarios where you may need to collect data to ground your responses.**

1. Quantitative analysis- quantitative analysis is useful for collecting mass amounts of data, for example survey responses.
2. Qualitative analysis- after collecting mass amounts of data, we can use qualitative analysis to dive deep and collect information through interviews for example.
3. Ethnographic research could be used as part of the qualitative research to study people or things in a real-life environment. For example in a previous class we did non-participatory research on ENVD students.