INFO 3401
Jason Kibozi-Yocka

Problem Set One

Part One: Histories

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." Why do you think he'd choose to frame data science this way? What might be problematic in this statement?

Peter Naur probably chose to frame data science in this way because he believed that data science should only deal with the data itself, and that it was the responsibility of another field (i.e. the medical field) to interpret and give meaning to that data in the context of their respective field. Naur probably didn't thought of data science as its own thing rather than an interdisciplinary endeavor like we do nowadays. The main issue with Naur framing data science in this way is that data needs context in order to be useful. The data scientist needs to know what the data is going to be used for, what question or issue they are trying to address, and what the data means in the context of the project it is being employed in. If we leave interpretation to another field, they probably won't know how to interpret the data such our work is for naught.

2. In 2002, data science began to gain momentum as its own dedicated subfield. Compare and contrast the definitions of data science at that time, exemplified by the National Science Foundation, Data Science Journal, and Journal of Data Science, to those from Tukey & Naur in the 1970s.

In contrast to how Tukey and Naur thought about data, the definitions of data science presented by the National Science Foundation, Data Science Journal, and Journal of Data Science is that the definitions they formed in 2002 are a lot more broad (less constrained/specific) than those presented by Tukey and Naur. Also, these definitions care more about the effect/impact of the data rather the process of data science itself. I particularly like the Journal of Data Science's definition in that data science is "almost everything that has something to do with data," which I'm prone to agree with.

3. Data continues to grow at an exponential rate today. List at least three technological factors that contribute to this growth and what tole they play. List three major sources of data that contribute to this growth and at least one way they're being used.

Three technological factors that contribute to the exponential growth of data are (1) ubiquitous technology, meaning that technology is everywhere and ever present (most everyone has a cellphone); (2) cheaper storage, in that companies can store more data for a fraction of the cost it was in the past and for less of the spatial cost (look at floppy disks vs the 1TB drives we have now, there are the same size but one has vastly superior storage for less of the cost, if we take into account inflation); and (3) better methods, meaning that since our data analysis process and tools become more advanced and usable, more data is processed by more people more easily. Three major sources of data that contribute to this growth are (1) our phones, which continually collect data about us through sensors and digital trackers; (2) social media, which has formed an economy around selling user data; and (3) the internet, which allows us to share data more easily and widely.

Part Two: Terminal Crash Course

4. How would you move into each of the following directories from the shell? Why do you think it is important to have shortcuts for each of these directories for navigating the file structure?
   a. Root

cd
It's important to be able to quickly jump back to the top of your file structure, especially if you find yourself deep in the file structure.

   b. Home

cd (or) *cd ~*

This is important as the previous one because it saves you time, especially when you’re doing more complex things.

c. Parent

cd -

This allows you to go back to the previous directory, which is super helpful because you don’t want to retype in the path (the just takes unnecessary time).

5. Briefly describe what the following set of commands would achieve. What process would happen and what would be printed to the command line?

```
cd ~
mkdir ./problem_set_1
cd ..
pwd (dir on windows)
```

This code goes to the home directory, then it creates the problem_set_1 (if it doesn’t already exist), then it jumps out of that file back to the home directory, and finally it prints the path of the directory (from root) and its contents.