Collaboration: Yizhen Wu & Kathleen Anderson

Dr. Szafir

INFO3401

26 January 2020

Problem Set 1

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?

As a data scientist, the only thing he needs to do is clean up, analyze data, and solve the problem. However, there will be privacy problems that they need to think about, not only just scoping the problem only. If we understand the content we could use those background information to do more work.

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 1970, data is the way we represent facts and ideas. It could communicate through some way. However, in 2002, data science is about everything. And the most important is the applications. Nevertheless, in the 1970s people knew what each data point meant,

but not the methods to put them together. Data at the time was defined as a representation of facts or ideas in a formalized manner capable of being communicated/manipulated by some process. So by 2002, data science meant "almost everything that has something to do with data: Collecting, analyzing, modeling… yet the most important part is its applications--all sorts of applications" (*Journal of Data Science*). Data scientists are "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection" (*The National Science Foundation*).

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

The concept of "data lakes" helped. It was efficient in categorizing the data only when it needed to be put in the database. Hardware is improving constantly, and there's more computer power in our cell phones than ever before; nevertheless, it's faster and cheaper. Relevance and ubiquity are stronger than ever before. Venmo is much easier than everyone pulling out their wallets to split cash evenly. Social media has given data a new category to exponentially grow. You can stream out how much information is given out at any given moment. So all in all, there is cheaper storage, greater bandwidth, faster compute times, better methods, and those ubiquitous technologies bringing in better methods.

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—use '/'

B. Home – Our default directory is the home directory, we don't need to do anything. And it represents as '~'.

C. Parent – use '..'

It's important to have shortcuts for each directory because it could help us to navigate our file structure faster and easier. We will not be lost by using those shortcuts.

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 ~ -- Change directory to home directory. It will go to the home directory.

mkdir ./problem_set_1 –- Make a new directory named 'problem_set_1". It will create a new directory names 'problem_set_1'. If we print ls now, we will see that there is a file names 'problem_set_1" that exists in our directory.

cd .. – Change the current directory to the parent directory. It will print the upper directory we just been through.

pwd – Print working directory. It will print what we are working on right now.