Jihoon Jang

INFO 3401

Professor Szafir

23 January 2020

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?

There is always going to be a context to data set. It’s important to have a background to the context. We got to have a better sense of what is going on. Naur argues that data is structured and did not realize that data science was going to be used in many different fields. Data is just not focused in itself, instead, it can produce lots of difference analysis, such as data visualization and data analytics. I think that he lacked to connect data with researching and establishing solutions to real-world problems. Most cases will be patterned. You don’t want to overlook because it’s not always focused questions. There is always some messiness.

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.

Tukey & Naur in the 1970s started as an empirical branch of statistics, arguing that data science is the compiling of data filled with numbers. The newer definition of data science, exemplified by the National Science Foundation, Data Science Journal, and Journal of Data Science, emphasizes that there is a variety of fields that data science falls into, such as science, mathematics, statistics, analytics, management, etc. This shows that data science nowadays is not simply dealing with just the numbers and that it’s a consistently-growing field where it’s used in so many different industries.

3.  Data continues to grow at an exponential rate today. List at least three technological factors that contribute to this growth. 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 emerging growth of data I think are: smartphone improvements, hardware (computer) improvements, and relevancy of data. Every aspect of our everyday lives is becoming technological and data. The internet generally has become very powerful through online shopping, tracing behaviors of consumers, web browsing history, etc. Social media has impacted a lot generating new data as well. For example, there’s a public Twitter data set that generates new data every day. Facebook data as well as Amazon data are used to manipulate individual’s interest and for target advertisements. Recently, I found out that Siri or Google Assistant are always listening to its surroundings that way they are able to answer when they hear the commands like “Hey, Siri” or “Hey, Google”. This emerges a question about people’s privacy because who would want their phones to be listening to what you are saying 24/7? The growth of technology can not only use your data but also surveil your privacy.

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 /  
B. Home – cd ~  
C. Parent – cd ..

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 ~ - goes to home directory  
mkdir ./problem\_set\_1 – make a “problem\_set\_1” directory  
cd .. – goes back out to parent directory  
pwd – check which directory you are in

“/users” would be printed to the command line.