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?

The reason why this quote is problematic is because of the fact that there is not just a focus question behind data, there is patterns, messy data, and sometimes answers (in the best case scenarios). Next there is scoping that goes on when analyzing data. This means that we must understand the problem that we want to solve, and this quote leaves out important factors that allow us to properly do our data analysis. There will always be context behind any problem, so having the background information on said context is helpful.

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.

These definitions have evolved in so many ways, because data science is not just data anymore, it has legal and ethical implications versus just some numbers. Data science tends to consider everyone and every aspect to data science. These definitions consider the application of the data that is collected. Lastly, it takes a team in order to succeed at data science projects, not just one person.

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.

Data has grown at an exponential rate and it is still growing today. There have been many factors that have led to this growth. Some of the technical factors include, but are not limited to, the fact that computers are faster and cheaper with constant software improvements. There have also been better methods of finding data that have allowed for an increase in the data and faster computing times. The three sources that contribute to the data growth include the increase of storage. Next people have realized that data is needed, so new data is created. Lastly, smartphone data has been created and it has created a ton of data throughout time.

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 → For root, you would do a C:// for a Windows, which I have
B. Home → For home, you would use a cd~ at the end
C. Parent → For parent, you would do cd . . to change to the parent dictionary

It is important to have shortcuts for this stuff is because we use this so much and we go deeper and deeper into the directories, that this just makes it way easier for us to navigate.

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 ~ → This is the equivalent to saying “Home”
mkdir ./problem_set_1 → This would take you from the current directory to a specific directory
cd .. → This changes the current directory to the parent directory
pwd → This would print the working directory