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?

I think he meant that the science behind data should be separated because then the quality of research would be higher and the quality of data would be more pristine and then the analytics of the data should happen after this is done. This can be problematic because the data should be gathered and analyzed together so that the results of the analysis can be the most accurate. If one person or group is doing both steps, then they can extrapolate the most valuable information.

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.

The 2002 writings state that data should be collected from everything all the time. With the internet, it is really easy to gather data on everything a user does. Tukey and Naur believe that data science is closer to a science than a mathematics. By 2002, it had become both a mathematician and a science and they then had the computing power to be able to gather data about everyone and everything.

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.

Moore's law that states every two years the number of transistors doubles in a computer. This leads to an exponential growth in the amount of processing power that computers have which allows people to search more, faster and allows more data to be collected.

Machine learning and AI also contributes to this growth because the software is learning on incoming information over time and trains itself how to get stronger and better to be able to be more streamlined.

The average person is now able to generate more data because of the technology they use. The technologies of phones and computers are so readily available that almost everyone has one so they are able to produce much more data than before.

Social media is being used constantly and data is consistently being collected on the users. Behind the scenes of social media is all data mining and the data is linked to a specific profile to help link personality and data.

API's that are publicly available allow anyone to access the data from that company and use it to run reports.

The world wide web allows for the collection of mass amounts of data and anyone who wants to find data on anything uses the internet to find it. It is exponentially growing source of information that is publicly accessible to anyone with modern technology.

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 ..

It is important to have these because you are constantly switching between different directories so the shortcuts exist to help streamline using the shell. The parent is relative to your current position so this helps traverse back easily. All of the file system stems from the root so it is always a good starting point.

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

This would create a new directory called problem set 1 in your home directory and you would go back to your root directory.