Problem Set #4

Marissa Kelley

Due: 9.23.18

**Monday:**

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

Peter Naur’s quote, “deals with the data, while the actual relation of data to what they represent should occur in other fields” has both pros and cons. If you don’t know the problem, you may not know what to do with the data once you get that. Naur might have framed it in this way due to the social structures of the time. If he’s living in a time where people are very specialized, and there’s not a lot of cross-over in fields, then it’s quite logical to say, “if I have information on this one subject, maybe I should pass it to a person who knows that field really well.” At the time, it was likely that there wasn’t a common language or dialogue yet as most people were only specialized in one specific area.

**Wednesday:**

1. There was a substantial shift in the ways we define data science between the 1970s and the early 2000s. Describe this shift and why it may have emerged.

The amount of data was growing rapidly between the 1970s and the early 2000s. Because of this, the shift of how we defined data science could have been caused by the a) influx (of different sorts of) information (from the emergence of the internet), b) the social ways we approached data and c) the amount of people who began to understand the power of (big) data. If more of the population understood the importance of (their) data, then more brain power was put behind the science, hence the definition was altered to accompany the vast variety of data.

Some differences within the definitions include that it’s being defined as more applicable- beyond the hard sciences (Journal of Data Science). It’s more than just a statistical endeavor.

The Data Science Journal brings up new questions about the ramifications of data, such as the applications, publications and legal issues- something we’re still working on.

3. The idea of "big data" dominates much of modern data science. However, data is still growing at an exponential rate.

A. What factors do you think may have led to this growth? Mention at least three and describe why they have contributed to recent explosions in data volume.

Factors that may have led to this growth include the transformation of technology that have grown in the past decade, allowing people to become more digital. This could have been influenced further by businesses that were in the process of learning that they could use their customer’s data to improve their sales and profit and competitive edge (even if they didn’t know how to use the tools quite yet) aka the digital economy. The invention of data lakes seems to have arrived when people realized how much data they had (still not knowing what to do with it) and creating these “lakes” to store it all so they could keep obtaining more. There wasn’t a reason to stop collecting the data because people understood that it would be helpful in the future. It became easier to collect once storage and computing power became cheaper and/or easier to access.

B. Where is this new data coming from?

Today, it seems data is still growing at an exponential rate because it’s coming from everywhere!! Whether people are aware of it or not, it’s being collected right beneath our fingertips. Through our phones, cars, laptops, purchases, credit card transactions, travels and anything else humans touch, see and do, we continue to create data, because we’re (almost) always connected. We’re at a point where we now have so much data that we can (sometimes easily) recognize patterns in our environment, whether modern or in the backcountry. All sorts of documentation are pieces of data, and everything that we post on the web can be considered data- meaning we’re producing it at a ridiculous rate!!!

**Friday:**

1. Name three different data collection methods. How are they similar? How are they different? Consider using specific scenarios where you may need to collect data to ground your responses.
   1. Method 1: One method of data collection are photos and documents. Facial recognition, used to use other data
   2. Method 2: Scanning of things (passports, cards, tickets, etc.)
   3. Method 3: We use surveys, tallying, interviews, scraping off the web, sensor data,