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” states that there should be more than one person involved. One person who “deals with the data” or makes sense of it, while there’s another person who is working in that particular field who is supposed to discover/further analyze what the data represents. This can be good if you need more background knowledge in another field or need a sort of specialization. You can create a bridge between the two fields and collaborate. However, this can be problematic on both ends if you don’t understand the problem or what you’re trying to answer, because you may not know what to do with the data once you receive it. It could also lead to a misunderstanding of what’s been interpreted.

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. For example, if you’re dealing with medical data, it would make sense that you may have had to pass off your findings to a person who went to med school since you probably specialized in something else. There was a lot of unknowingness.

**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. The emergence of the internet allowed more people to gain access to the knowledge. 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. Once businesses began understanding its importance, they were collecting the data rapidly- regardless of if they understood it or not.

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, which is 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 (1) the transformation of technology that have grown in the past decade, which allowed people to become more digital. It also allowed people to become more connected. This could have been influenced further by (2) 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 (3) invention of data lakes (likely) 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 essentially 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!

One example that is very useful for me is the data that’s been collected through years and years of observations and calculations about ocean tides and currents. Through this, websites and apps, like deepzoom.com, have been developed to provide accurate timings of what the water will be doing at certain times. It saves a large amount of time instead of doing calculations by hand and continues to become more accurate as researchers/scientists/etc. collect more data.

**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.

One method of data collection are photos. This can be used for facial recognition, travelling or for an emergency crisis. We also use photos to display news and trigger emotion. For example, photos have been used as a way to gain support from other countries when there is a crisis happening, such as the refugee crisis. A second method of data collection includes scanning things, like passports, cards, tickets, food, etc. This collection method allows us to keep track and inventory of what we have or what we’ve done in the past. A final method of data collection are interviews. These are used if we want to gain a more individual perspective of something. All three of these in a way tell stories, with and without words. Photos and scanning are more qualitative while the interviews are more quantitative. The photos and scans are useful for understanding processes while interviews are useful for gaining more answers, clarification and clearing up any ambiguous data. Other forms of data collection include (but is not limited to) surveys, scraping data off the web, focus groups, etc.