Jack Sandberg

INFO 3401

Prof. Szafir

9/23/2018

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?
    a. This statement might be problematic because those who originally gathered/work with the data would have more context on the data than those whom it might get passed on to. Those who gathered or assembled the data would know better how the data could be used to reflect and help accurately those from whom the data came. Combining the roles here also takes away some of the abstraction that comes with working with unfamiliar data.
    b. Naur could have chose to frame data science this way because it allows more room for specialization. They might be able to make more inferences and do more important work, and it might also remove some of the emotional connections that those who gathered the data might have. It might help look at the data from more of an analytical standpoint. Also, by using interdisciplinary fields, more results might be drawn from the data.

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.
    a. In the 1970's, data science was thought of as mostly collecting data. By the 2000's, the way we approached data had changed because a much higher drive to analyze the data became prevalent and gave rise to the field that we now experience today. Between the 1970's and the 2000's, people changed from simply wanting to collect data, to wanting to interpret and draw conclusions from it so as to make informed decisions. This was

in part because it was during this period that there was a huge influx in the amount of data gathered and kept.

2. 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.
        i. Due to the inventions of data lakes, storing data became a lot more viable and practical.
        ii. Data scientists are still realizing the potential of the growing field, meaning that new kinds of analysis jobs are being created every day. Companies are realizing that it is an investment, and jobs in this field can lead directly to growth of the company through data analytics.
        iii. Social media has convinced millions of people into posting and recording their data and life events, and due to this, data is amassing at incredible rates from sites like twitter, youtube, instagram, twitch and other sites.
    b. Where is this new data coming from?
        i. To be quite cheesy, this data is coming from everywhere. From businesses it is coming in logs and records. From social media it is coming in locations, words, and posts. From the government it is coming in logs and the census and records. It is coming from anyone and any company that can record data. Also, with the invention of smartphones, recording data has become a lot easier as we are always connected to the world.

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
    a. Three different data collection methods are interviews, polls, and observations. Interviews are best for finding qualitative, detailed

information about a specific subject or topic, and require lots of time and planning to meet people and write down responses. While they can get you detailed information, they aren't as effective at collecting large amounts of information as polls. Polls are used typically with multiple choice responses and can be given out to many people, and don't require meeting in person. They are effective at amassing quantitative data, but worse than interviews for collecting qualitative data, and still require user participation, which is different than observations. Observations entail going out and simply recording what you see about people. One example of this would be if you were trying to figure out how many people use the bike lanes on campus, your most effective technique would be to go out and sample, by observing and counting people that pass by on a daily basis. If you were trying to find out how much people bike on a weekly basis or categorize the uses, a poll would be better. If you wanted to find detailed uses of why people bike to understand individuals and their intentions, interviews would be best because you could ask many open ended questions and receive detailed, qualitative information. The difference between these examples of observations, polls, and interviews matches the differences between numerical data, quantitative data and qualitative data, and the best data collection methods would incorporate a combination of the three.