Anastasiya Zhukova

Assignment 4

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?

What’s problematic about this statement is that it implies that data scientists should simply crunch the numbers but not bother to put their findings into any kind of context and instead hand off their results to someone else who gets to analyze them. Besides making data science a painfully boring job, this would also slow down the process of data analysis and might make it difficult for an analyst to understand where the data person’s findings came from.

I think he chose to frame data science this way because having two separate people that each do part of the data science process might be a way to eliminate bias from the analysis portion of the process. It could have also have come from a desire to create specialization for each part of the data analysis process.

2. There was a substantial shift in the ways we define data science between the 1970s and the early 2000s. What has changed in these definition compared to the 1970s?

In the 70’s, data was defined as the collection of individual data points and was chiefly treated as a domain of statistics. The collection of this data was also treated separately from the analysis part of the data (determining what the data means). By the early 2000s with the emergence of the internet, more data was being generated and collected and data scientists had to ask themselves what the ramifications of data are as well as what to do with it, so they created new definitions that included both the actual collection of the data as well as the methods that they would use to make sense of it.

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 led to the data explosion: the digital economy, storage and computing power has increased, greater bandwidth. The digital economy obviously contributed because businesses discovered that they can analyze the data trails people leave from using new technology and use the insights gained from that data to better target their consumers across various digital platforms. Storage and computing power has contributed to the data explosion because it has allowed us to store more data on our machines with minimal cost as well as analyze large chunks of data much faster. The greater bandwidth has contributed to the data explosion because it has allowed us to send more data much faster over large distances. This allows us to the share lots of information much faster, which in turn generates larger amounts of data.

B. Where is this new data coming from?

This new data is coming from people engaging with new and emerging technologies (smartphones, smartwatches, smart-insertrandomobjecthere-). A large chunk of it comes from people using social media, doing web searches, and engaging in media consumption on the web. Other contributors include the internet of things (toasters, refrigerators, and thermostats hooked up to the internet) as well as public infrastructure (traffic lights that operate through the internet).