Zach Lyons

9/23/18

Problem Set 4

**Monday**

* Naur is claiming that the analysis side of data is fully dependent on the data scientist, and that the scientist is not responsible for relating his work to other fields. He claims that the workers within these other fields should do that themselves. I think it is problematic, because data science is such a universal practice and therefore scientists should have the communication skills to represent their findings to the proper stakeholders. The idea should be to work with these stakeholders, not alienate them from the process.

**Wednesday**

* I’d argue that the most drastic change in the definition of data science since the 1970s is the ability for scientists to contextualize their findings. All three of these definitions don’t really mention this aspect of data science, they are more concerned with the technical side of things. I’d argue that the talents of modern data scientists are not only limited to this, they are also able to contextualize it so that anyone can read their findings in layman’s terms.
* What factors do you think may have led to this growth?
  + The increases in computing power in the modern age, and the “quantifiable self” movement are two primary causes for the data explosion. The amount of data being generated is so huge because just about any work application we use on a day-to-day basis generates a large amount of data for companies to use.
* Where is this new data coming from?
  + Just about every business has an online presence or platform these days, which is able to generate data about a user’s habits. This new explosion of data is courtesy of the sheer increase of data-generating platforms.

**Friday**

* Interviews
  + Gets first-hand information about your targeted demographic
  + Human bias can come through in this scenario, as interviewers can direct interviewees into certain frames of mind, while interviewees can withhold information form the interviewer.
  + Provides context that other methods don’t.
* Surveys
  + Good for getting a lot of data in a short amount of time
  + Can also lead to bias, from wording of questions
  + Can lack certain context, especially if survey questions are worded poorly
* Log Data
  + Provides data based on observations of people, or records of their behavior.
  + Observational data, twitter data, online metadata are all examples of this
  + Lacks significant context, especially if the person observed is not able to be contacted