Morgan Baccus  
CptS 475  
Assignment 1

**Task 1**

**1.a.** I believe it makes the most sense for the left most skill on the x-axis to be the most generic and increase in speciality as the skills move to the right. Typically when reading charts like these, the reader starts on the left and moves right. Communication and presentation skills seemed like the most common area out of the seven since it is required in almost every profession and discipline. Math is also a very common skill, but many professions only require basic math. Statistics is similar to math but slightly more advanced and specialized. Data visualization is a much more unique skill as only professions that work with data will require it. Computer science is more narrowed than data visualization as you have to be able to work with data in computer science, but also know much more specific to the area. Machine learning is a specialized type of computer science. Domain expertise is very specific to whichever domain chosen.

**1.b.** A skill that I believe should be added to the data science profile is data mining or gathering data. The skills listed center more around what to do with data once you already have it, not how to collect the data that is then used for analysis. Computer science and machine learning skills can help, but I think that gathering data is a large enough part of the job that it should be its on area.

**Task 2**

**2.a.** The differences between data science and statitics outlined by the author are:

1. The data analyzed in data science is “increasingly heterogenous and unstructured text, images, videos often emanating from networks with complex relationships between their entities” (“Data Science and Predictions”, Vasant Dhar). The complex nature of the data requires the application of data science rather than statistics in order to analyze in a meaningful way.

2. The more complex data that is being generated today “requires integration, interpretation, and sense making that is increasingly derived through tools from computer science, linguistics, economics, sociology, and other disciplines” (“Data Science and Predictions”, Vasant Dhar). The data science discipline is essentailly comprised of elements from many other disciplines whereas statistics is its own separate discipline.

**2.b.** Big data has the power to put both hard and social sciences on firmer ground in terms of theory development. This can be accomplished in the hard sciences by analyzing large amounts of data and finding patterns even where no causal reasons are immediatley obvious. If these observed patterns are predictive, then further inquiry can be done and a theory can be developed. In terms of the social sciences, big data can also be used to find predictive patterns from human behavior. Humans are “simplistic”, yet many factors come into play that normally might not be observed but is with big data.

Errors are significantly reduced through the use of big data. Big data reduces misspecification of the model errors by allowing “us to consider models that make fewer assumptions about functional form than linear or logistic regressions simply because there is a lot more data to test such models and compute reliable error boundsC Additionally, using a sample to estimate the model reduces error as “sample estimates become reasonable proxies for the populationa” (“Data Science and Predictions”, Vasant Dhar). By eliminating error, big data predictions are more reliable and can help put both hard and social sciences on firmer grounds in teerms of theory development.

**2.c.** Headline: Big Data Leads to Predicitve Modeling that Can Answer Any Question

Data science can now make use of big data and build predictive models to answer questions that people didn’t even know to ask.

Patterns that weren’t readily apparent that have been made clear through big data are leading to new theory development in both the hard and social sciences.