CptS 475/575: Data Science Fall 2020

Assignment 1

Ayush Kapoor

9/1/2020

1. **Create Data Science Profile of Yourself and Reflect on an Article on Data Science.**

   a. **The areas in the horizontal axis could be ordered in a number of different ways. What ordering in your opinion would be most effective (and aesthetically pleasing) and why? Create your profile in the order you chose.**

   The areas in the horizontal axis could be ordered in a number of different ways. The order, in my opinion would be:

   Data Visualization → Statistics → Machine Learning → Computer Science → Communication and Presentation Skills → Math→ Domain Expertise.

   Data Visualization will be the major set of skills for the data science process, which basically means that after collecting and modelling the data within various data frames, it will then be visualized for conclusions to be made. Statistics will help find meaningful trends amongst the data. The main goal to be identifying patterns and trends within amongst data received. Machine learning will be using the statistical models to perform tasks with instructions. Understanding the instructions and algorithms within Machine learning models will come from skills obtained in the Computer Science and Math areas.
   These computational methods may become too complex to explain to non-technical audiences. The authority to articulate and present such models of predictive power is constituted in the Communication and Presentation Skills.
   Domain Expertise is given the least priority as it will only broaden its horizons once the rest of the skillsets have been obtained and mastered.

   The following two figures are the depictions of my Data Profiles. Figure 1 depicts my skillset currently whereas Figure 2 is where I would aspire to see myself obtain by the end of the semester.

Figure 1. Before the Semester (BOS)

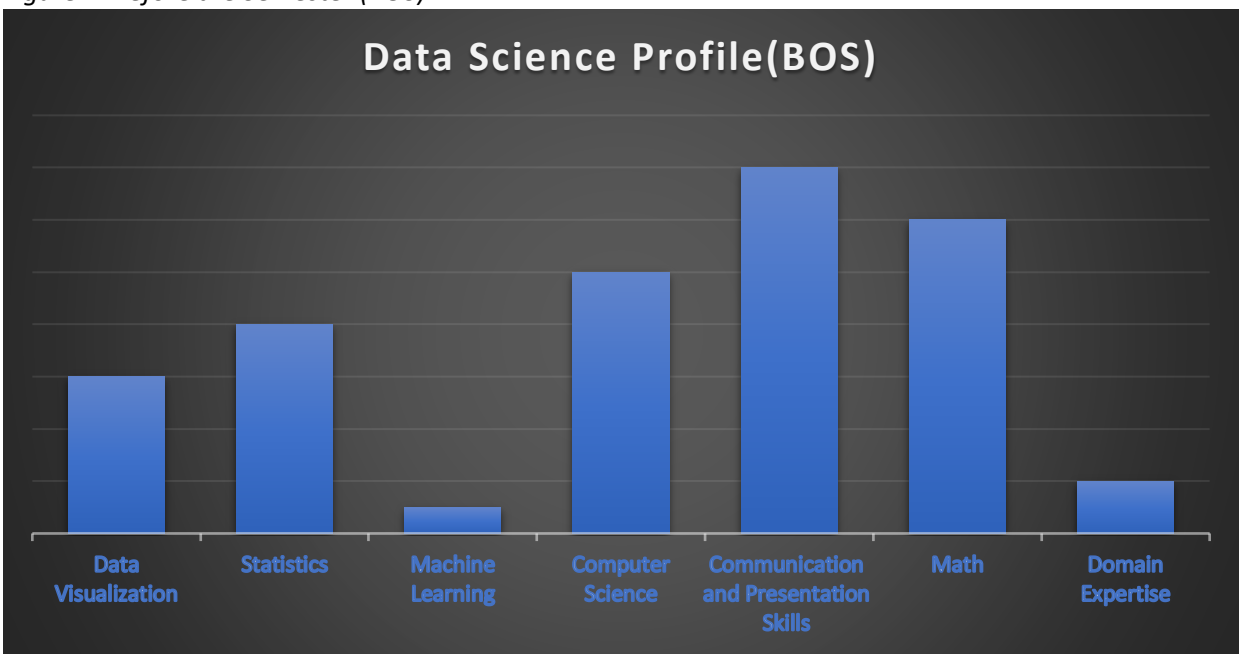
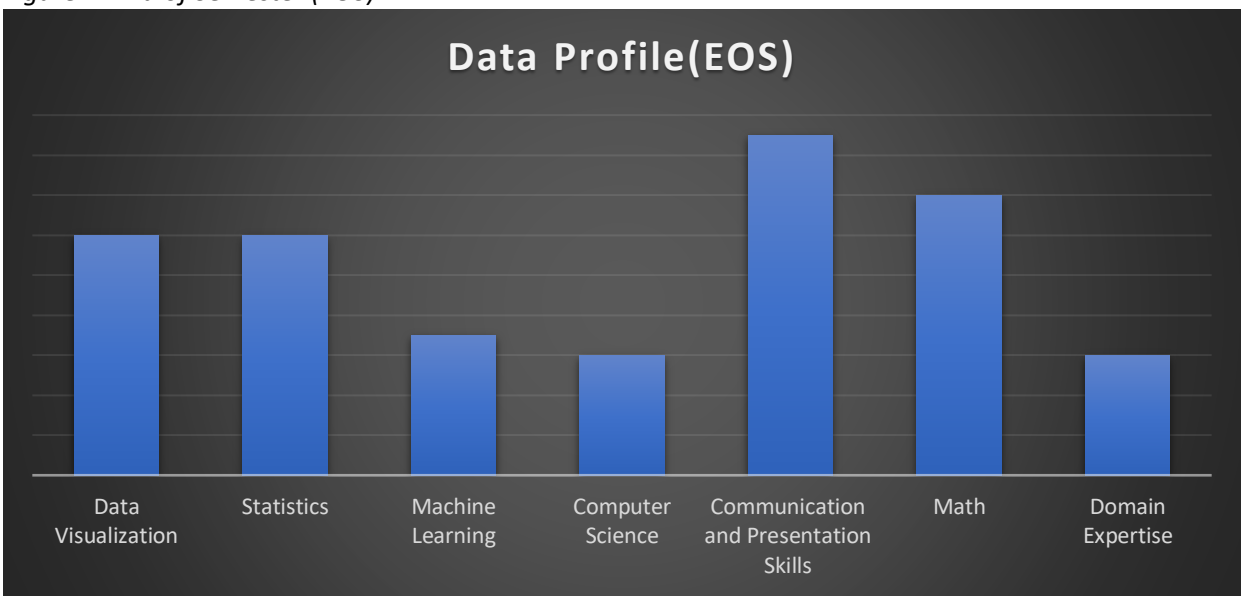


Figure 2: End of Semester (EOS)



b. Is there a skill (bucket) you think should be added to this data science profile? A skill you think should be removed. Specify and justify briefly.

A skill I would like to add into this Data Science Profile would be: Cloud Computing. When dealing with such immense datasets, Cloud Computing would enable users to rent and incorporate data servers, storage, and computing power from providers. In my perspective, although this entity would rather come under Computer Science and Data Visualization principles, knowledge, and implementation regarding Cloud Computing too, should be

classified as a skill. It is evident to keep up with the necessary features that circulate the field of Data Science.
There is no skill I would like to remove from my Data Profile. All these skills have their advantages to Data Science and lead the rebellious to make substantial decisions in the development of the logic and reasoning.

2.

a. The author identifies a few ways in which data science differs from statistics. What are those ways?

Statistics and Data Science may complement one another whereas they do have major differences. The author explains, in brief, the different ways in which statistics is an extended practice carried out on specific 'datasets'. This 'data' is considered structured with boundaries, relationships and functions used to generalize and predict results. The information is initially extracted/created from unstructured data and given such measures.

In contrast to the former, data used for Data Science consist of all the 'heterogeneous' and 'unstructured' images, text, videos etc. from various different traces with relationships not too simple to make the data **classified.** This unstructured form of data requires analysis, integration, interpretation, and sense making in order to make computers interpret this information automatically. Various tools from Computer Science, linguistics etc. are used to make machines perform decision-making implementations.

b. The author discusses ways in which "big data" could potentially put domains on both ends of this spectrum on firmer grounds in terms of theory development. Give a brief summary of the ways the author identifies. Do you see any additional ways than what the author sees? (If the discussion in this section of the article resonated in some ways with your own research or work you do, feel free to incorporate that in your answer.)

In the section, author differs in between physical sciences and social sciences in regard to the predictive power of its theories. In Physical sciences, theories are explained and considered 'complete' when certain introduced variables help to explicate the phenomenon. This field heavily follows designed models and describing the factors that will affect its input changes which yields persistence and precision.
In contrary to the former, social sciences lay their emphasis on proposing theories that would concur casualties with not the same predictive power as Physical science phenomenon's. Theories are considered substantial and strong when the accuracy of success is 95%. This may not be the same case when working with Chemistry or Physics formulae.

However, Big Data makes it accessible for computational machines to intervene and exploit data in a manner human would not consider. Therefore, in the latter fields, large amounts of data computation and mining could yield 'accurate predictive models', which can later lead

experts in the correct depth for theory development. These models make less assumptions about functional form due to computation of a large set of data.

c. Imagine you were asked to write a "head-line" (as you see in newspapers) for this article, followed by two or three very telling summary sentences. What would your headline and the summary sentences be?

Although some might consider this topic to be "too technical", it is evident to realize how significant Data Science operates without confusing it with any other entity. Misconceptions arise due to vague knowledge about this phenomenon. Furthermore, it is important for the masses (not just domain experts) to know about Data Science. Progression should normally be in the following order: data → information → knowledge
Therefore, my headline would be:
Data-driven Science for the Confused.
Data Science may be a phenomenon that has cofounded many yet is too consequential to ignore. Data, structured or unstructured, has proven to be of substance in determining the cause of problems. Here arises the explanation and implementation of how this data in abundance, will be modelled with massive computational power and make valuable predictions. This article demonstrate will demonstrate how scientists/corporations use this phenomenon to their advantage.