Assignment 1: Data Science Profile

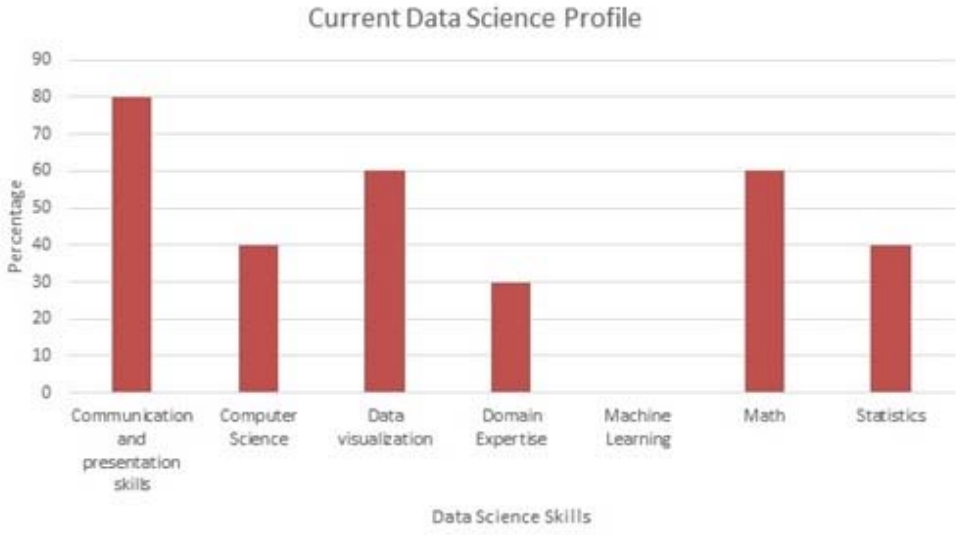
Sheryl Mathew

1 September 2018

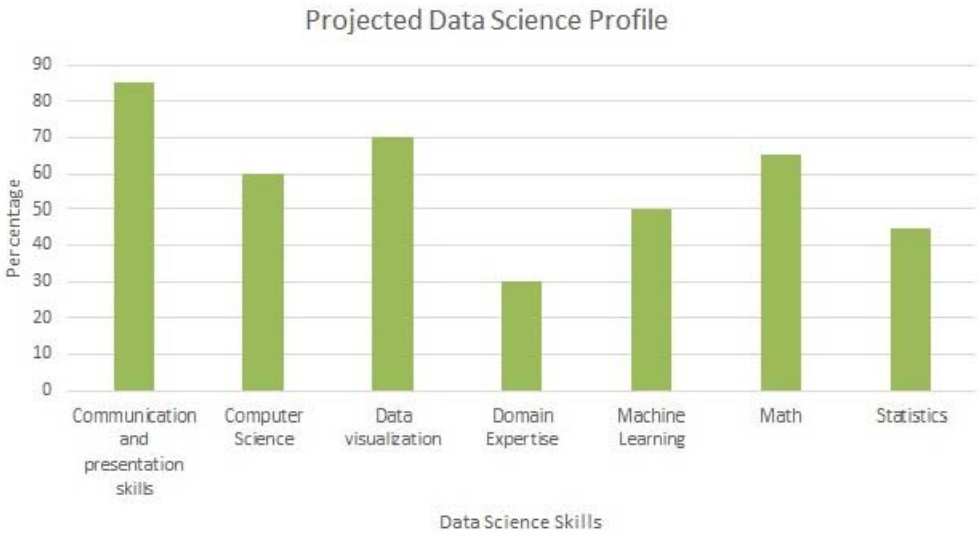
1. **Task 1:** As I explained in class, the purpose of this task is to create a visual data science profile of yourself. Specifically, you will create two instances of profiles. The first will show the way you see yourself now. The second will show how you would like to see yourself by the end of the course. The profile is simple. On the horizontal-axis you will have seven "areas of skills" that could generally be regarded important to Data Science: 1) Computer Science. 2) Math. 3) Statistics. 4) Machine Learning. 5) Domain expertise. 6) Data visualization. 7) Communication and presentation skills. On the vertical-axis you will have a relative scale (think percentage) of your skill level in each of these areas. The area in which you have the strongest skill will be close to 100, and the area in which you think you have very little skill would be close to zero. As an example, see the slide in the lecture slides of Aug 24 (posted on OSBLE under Lectures/082418) that shows the data science profile of the author of the book "Doing Data Science" (one of the reference books listed in the course syllabus). As a context, the author, Rachel Schutt, has a PhD in Stat and has held several senior and executive-level Data Science positions in industry. Your task is to create your own profile – two to be exact, one showing current and the other projected. You are still a student and you may not feel you have a lot of skill in some of these areas. Allow yourself a generous interpretation of skill level and keep in mind that this is on a relative scale. Also, keep in mind that it is perfectly okay to have zero skill level in some of these areas. For example, if you are a computer science major, it is natural that "Domain expertise" would be the area in which you have the lowest skill level among the seven, and it is okay for it to be close to zero. That said, if you have had an internship some place or an interest/hobby that you think has helped you acquire some expertise in an area, you could take that into account in deciding the level of your "Domain expertise". In any event, make sure to mention what the domain is if you indicate your domain expertise to be non-zero. You can use any tool (Excel, Matlab, R, etc) you wish to make the plots. Here are a few associated presentation considerations and discussion points you are asked to address as part of this task.

a. **The areas in the horizontal axis could be ordered in a number of different 1 ways. What ordering in your opinion would be most effective (and aesthetically pleasing) and why? Create your profile in the order you chose.**

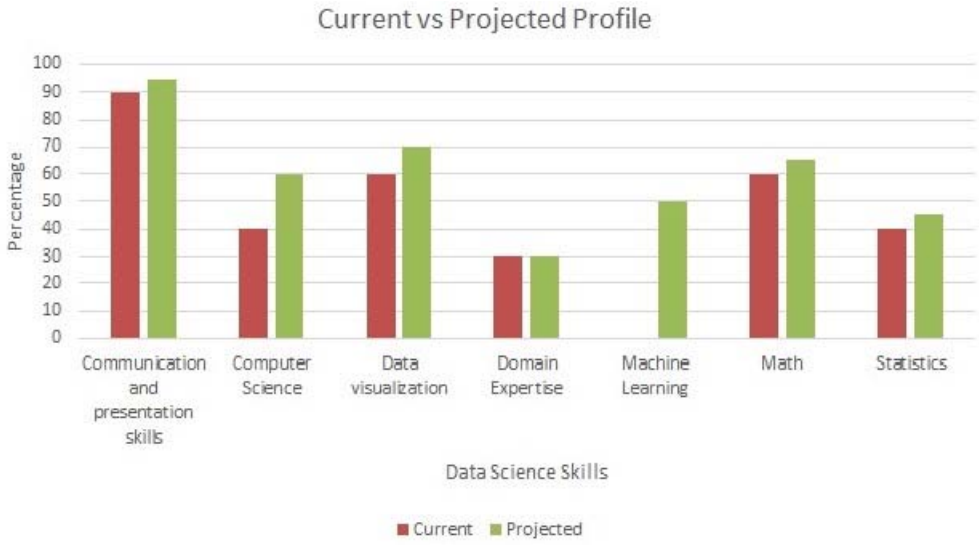
In my opinion, the effective manner to arrange the data skills would be to sort them alphabetically (natural sort). When we want to quickly search for a particular skill set out of the different skills that are present then it would be easier to find it since we know which part of the axis we need to look at instead of traversing through the entire axis one by one (each skill) which is a lot more time consuming.



My current profile shows my communication and presentation skill to be the highest since I have participated in many elocutions and presentations throughout my academic career. My Computer science skill is at 40% is because my knowledge in R is at a beginner level and Python at an intermediate level. Data visualization is at 60% because in the company I worked for a year I was exposed to generating different reports based on the user needs. Domain expertise is at 30% as I am a transfer from Electronics and Communication to Computer Science and I have a basic domain knowledge of the same. Machine Learning is at 0 since I do not have any exposure to it. Math is at 60% as I have a solid foundation from my undergraduate and Statistics is at 40% since though I know the concepts I will need to refresh them.



My projected profile shows my communication and presentation skill to increase to 85% as I would have to present my project work in Machine Learning, Data Science and Proseminar. My Computer science skill is raised to 60% as my familiarity with R and Python will significantly increase. Data visualization is at 70% as I will be learning more about how to do it using R. Domain expertise remains the same. Machine Learning increases to 50% after completing my course in Machine Learning simultaneously. My Math and Statistics will increase after those operations both in Machine Learning and Data Science classes.



This is the combination of both my Current and Projected profile

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.**

   Logical and Problem Solving Skill would be a skill that can be added to the data science profile.

   When we are faced with a business problem we need to foremost understand what the problem is that we are trying to solve, look at the problem from different perspectives, the structure of the data before deciding the models that we need to use to handle and process the data instead of rushing to build a solution when we first come across the problem. This would be possible only if we look at the problem in a logical and critical manner. Therefore this skill would be a good addition to the data skill set.

2. **Task 2:** As you recall, we briefly discussed the article "Data Science and Prediction" by Vasant Dhar in class in connection with the topic "what is data science?" A copy of the article is posted under Lectures/082418. Read the article and briefly answer the following questions.

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

   - **Type of Data:** In Statistics we primarily deal with structured data but in Data Science we work with increasing amounts of heterogeneous and unstructured data (text, images, video)
   - **Required Disciplines:** To perform Statistical analysis we only require knowledge of Mathematics but when dealing with unstructured data we need to have a combination of different disciplines like Computer Science, Mathematics, Linguistics and Sociology amongst others to better analyze the data.
   - **Decision making:** Decision making using Data Science is more efficient when compared to using Statistics since we have a large amount of data that is being generated by humans and computers and the computers are able to understand the data that is being created and take better decisions automatically.
   - **Knowledge Discovery:** Statistics is not optimal for Knowledge Discovery since they provide data based on the given query while in Data Science we are able to find patterns within the same data.

- **Prediction Capability:** One of the key factors that is given the most consideration in Data Science is the predictive accuracy of future observations based on a given set of data while Statistics mainly deals with just the analysis of the given data.

b. In the section of the article headed "Knowledge Discovery" (pages 70 to 72 of the article), the author makes a distinction between domains in terms of the predictive power of their theories (models). Secifically, the author points out that models in the physical sciences are generally expected to be "complete", whereas in the social sciences they are generally "incomplete". 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?

In theory development, whether we start with a pre-conceived idea of the theory or try to develop a theory from the results of the data, we use data science. In physical sciences, we proceed with data analysis after understanding the phenomena (relation) of what we are looking for therefore the explanatory and predictive models are the same. In social sciences we make assumptions about the data and then proceed with the analysis and based on the results of the analysis we extract causal models.

In both physical and social sciences, we encounter three kinds of errors: misspecification of a model, the samples used for estimating parameters and randomness. These errors have been solved using data science. In misspecification of a model which occurs because of choosing an incorrect model for a particular problem is rectified by using large amounts of data to train and test the model such that it makes fewer assumptions and has reliable error bounds. The samples used for estimating parameters causes the model to have a greater bias in its estimate when the sample is small. This issue has been resolved by using high volumes of data such that the sample estimates to be a reasonable proxy data. The error due to randomness occurs because the data that we process is passive and performs predictions for the given information. It does not predict the output when one of the variables are

changed during run time. In short randomness occurs because the data that we receive is not from a controlled environment.

Using data science, physical and social sciences are now able to analyze data at a granular level, draw conclusions from the results, find patterns and connections they would have never otherwise figured out as well predict the future. This leads to formulating theories from the data with less assumptions and bias.

Data science also aides theory development through its powerful data visualization and communication tools. It helps produce a graphical representation of the analysis and the output of theory in a manner which can easily be understood by people from all levels of expertise. Its ability to convert the report on the theory that is created once into multiple formats without additional effort helps researchers to reduce the unnecessary time taken to generate the same report in different formats.

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?**

"Data Science: The key to thrive in the New Era of Data".

In a world where data is increasing exponentially by the millisecond we have found an efficient way to manipulate this data and predict future outcomes using Data Science. The skills needed to perform data science and it's applications to real world problems have been discussed.