CptS 475/575: Data Science Fall 2020

Assignment 1: Create Data Science Profile of Yourself and Reflect on an Article on Data Science（By WenLi 11660347）

Task 1

1.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.

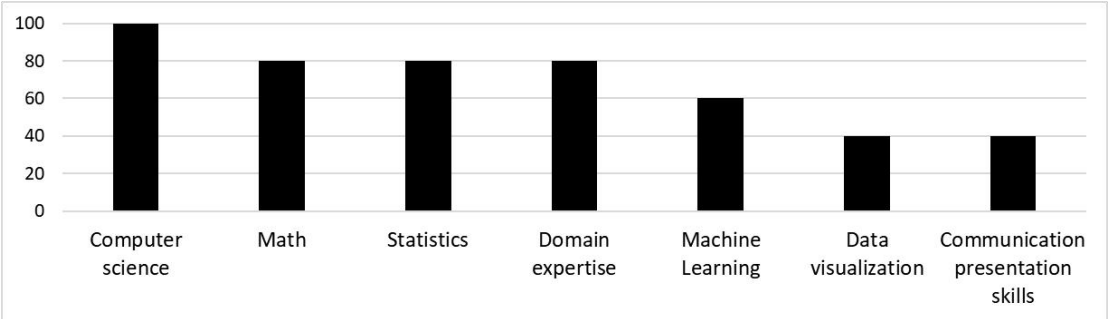

Figure 1 My current profile

In my perspective, the most effective and aesthetically pleasing order is to the descending order of the scale of the seven areas. Hence i can easily divide the seven skills into two groups, i.e., (1) the left four areas and (2) the right three areas, which indicate what i have mastered and what ares i need strengthen respectively.
(For the area of "Domain expertise", i worked as a developer or designer of home gateway and router for more than ten years, so i think i can get 80 in this domain).

Through the study of this course, i expect i can master common approach of data analysis and visualization well and use some machine learning tools skillfully, which can help a lot in my research. The expected profile is shown as Figure 2.

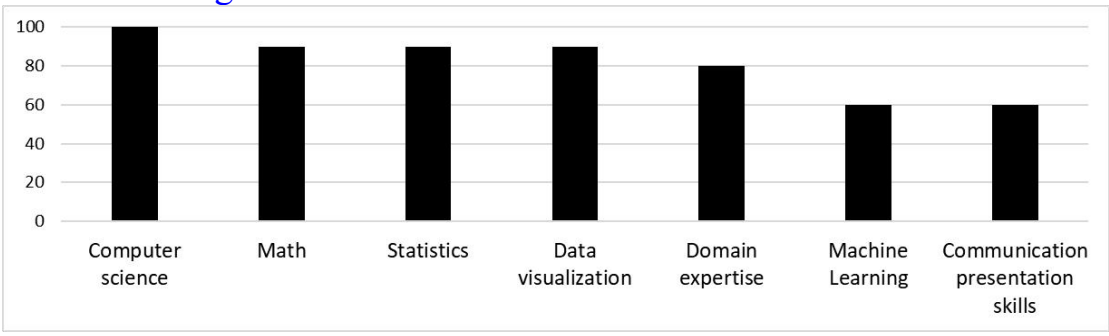

Figure 2 My expected profile at the end of the class

1.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.

I think there are two skills that should be added:Team collaboration and Lifelong learning. And none of the seven skills should be removed.
In the era of big data, it is difficult or even impossible to complete a project by one's ability, so team collaboration is an essential skill in all areas. The reason we need skill of lifelong learning is that data science is interdisciplinary and multidisciplinary. This feature need us to keep learning new knowledge and techniques.

Task 2

2.a. The author identifies a few ways in which data science differs from statistics. What are those ways?
(1) the data that data science target is increasingly heterogeneous and unstructured.
(2) analysis in data science requires requires integration, interpretation, and sense of the combination of unstructured and structured data, which depends on many other disciplines such as computer science, linguistics and so on.
(3) in data science, computer is doing more and more background work and make decision automatically.

2.b. Give a brief summary of the ways the author identifies. Do you see any additional ways than what the author sees?
Summary of the ways the author identifies:
(1) In "hard" science, big data makes it possible to extract causal model. In this field, the model could be assumed complete in practice.
(2) In other fields, if a model can predict results with high precision, it is useful to guide the right direction even without causal insights mined.

In my perspective, lots of causal insights can not be observed by humans, but through some useful theories and tools, we could mine the causality upon the big data. Take a example, in my research on correlation analysis between project properties (age, languages, developers, etc.) and project quality, i did not know whether there exist causality between the two subject, so i constructed a NBR model on a data set about 1.5million counts, the results indicate language number had strong association with project quality, then i did some case studies to verify this prediction. So computer techniques (deep leaning, AI, etc.) could help find hidden causality upon big data.

2.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?

Get close and learn about the power of data science in big data era.

Integrated knowledge across multiple disciplines construct the cornerstone of big data analysis in data science. Based on these knowledge, its real power is to predict the future and mine new knowledge.