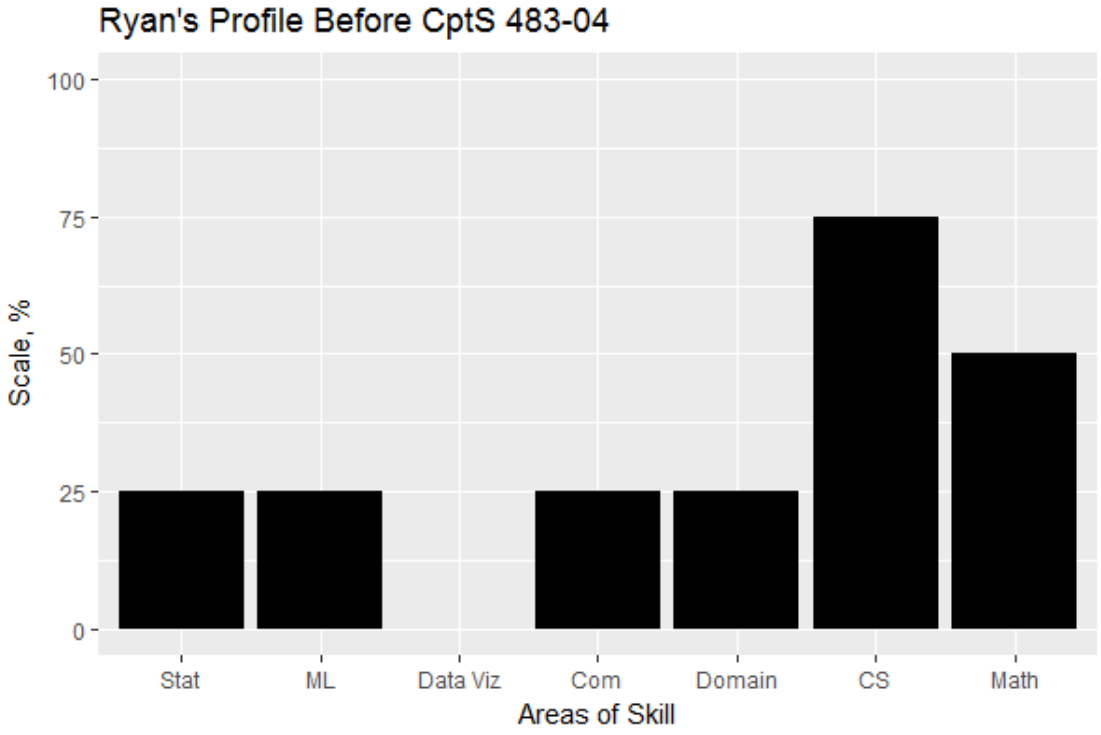
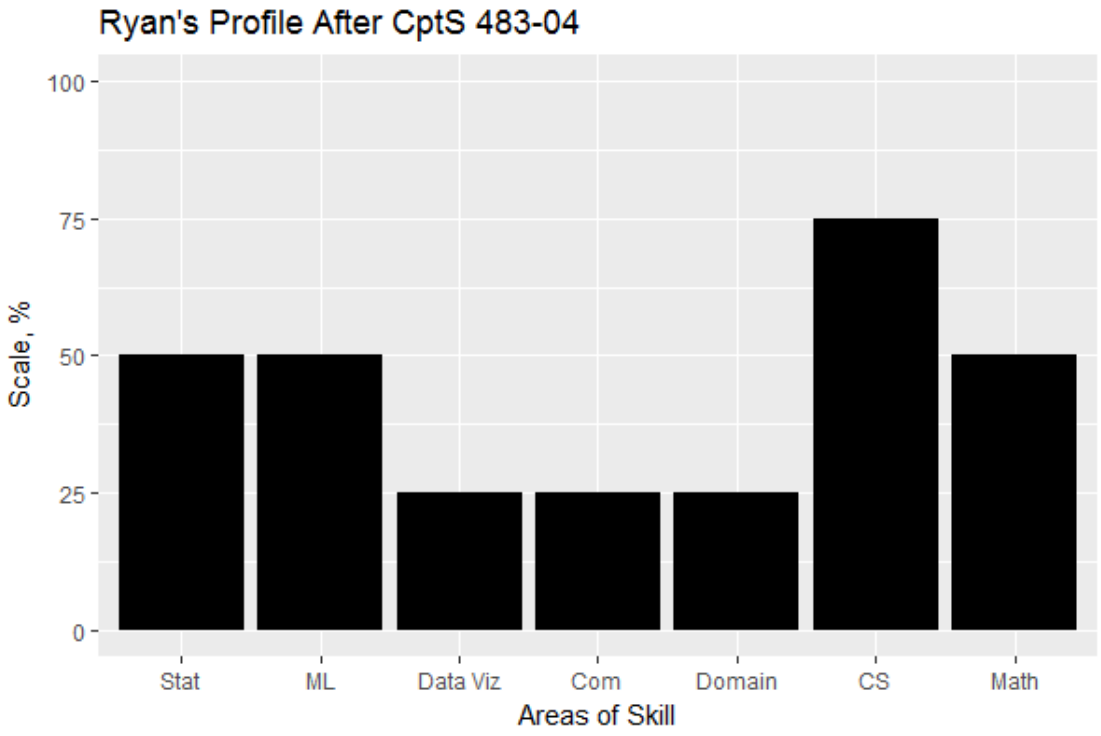
Ryan Torelli
CptS 483-04
Assignment 1
August 30, 2017





Task 1

1.a. The most effective ordering of skills is the choice that renders the profile most readable and natural. One choice, for example, may order skills on a continuum from abstract to concrete, where math rates most extreme in abstractness and communication rates most extreme in concreteness. Another choice, and the one I have made, ranks skills left-to-right from most utilized in data science to least. I rank statistics highest and math lowest. An ordering by any choice, however, leads the reader to interpret skills on a bar graph as ascending or descending proportionally. Clearly, this is false. For my own ordering scheme, statistics and machine learning are most highly utilized and the remaining five skills bunch together as least utilized. Yet, when displayed as equal-width bars on a bar graph, the order of skills appears incremental. A possible correction to imparting this appearance is to change bar width according to utilization or position bars along a number line. The result is more unnatural and less readable; so no correction has not been undertaken.

1.b. There is a skill that should be added to the data science profile. After introducing the concept of a data science profile, in *Doing Data Science*, Schutt and O'Neil describe data science as a process comparable to the scientific method (pp. 41-44). In the opening lecture of this course, Dr. Gebremedhin echoed Dhar in “Data Science and Prediction” that the data science skill set includes "problem formulation to engineer effective solutions" (p. 64). Indeed, the scientific method or an engineering perspective is a distinct skill and one that is separable from the seven skills in the profile. It is crucial to have skill in the process of science or engineering before embarking on tasks that follow a similar arch. Data science is a process that explores data, tests a hypothesis, or engineers a data product - any of which require skill in the research method.

No, no skill should be removed. All seven skills previously stated and the one proposed contribute highly to the skill set of a data scientist.

Task 2

2.a. Dhar opines that data science is different from statistics with respect to assumptions that underlie some methods and capacity to handle unstructured data. Dhar contrasts machine learning with multivariate analysis, writing that machine learning methods "make no or few assumptions about the functional form of relationships among variables" whereas multivariate analysis assumes a model that constrains the "relationship between the dependent and independent variables" (p. 69). Dhar writes that the "raw material" to these methods is "increasingly heterogeneous and unstructured" (p. 64). The tools of computer science adopted by data science enable automated organization of semi-structured or unstructured data. Statistics neither has the tools to organize text, images, or video with relative ease, nor can it operate on such raw data.

2.b. Professor at Stern Business School Explains Data Science to Computer Scientists

- Data science is new and different. It incorporates elements of traditional disciplines to learn patterns or make predictions with high confidence from large volumes of data.
- The data science skill set is interdisciplinary, including machine learning, statistics, computer science, and problem formulation.