Review of *50 Years of Data Science*

Grace Yi Chen

This paper talks about the history, current situation and, future vision for data science. Data science is not new and is closely related to statistics. Statistics professors like Cleveland, Chambers, and Tukey raised the concept of data science more than 50 years ago. They argued that the scope of statistics needs to be enlarged and to include broader data analysis activities. There has not been much change in the statistical academic community until 2000. The author proposed the framework of "greater data science" (GDS) with six divisions including data gathering, representation, computation, modeling, visualization, and science. Currently, data modeling is the academic research focus and major shifts are needed in teaching and researching in GDS. In the future, the author visions that reproducible research will become more popular. Numbers in publication will be more accessible and retrievable for meta-analysis. Ample data will be available to measure the performance of algorithms across different settings. Several discussants share their opinions about the article and the future development of data science in teaching and research.

I think the author gives a great overview of the history of data science and calls for a broader scope of statistical research. In my understanding, there is a large overlap between applied statistics and data science. I agree with the proposed GDS framework and six divisions which widen the tent of statistics. However, I share the same concern with the discussants like Peng, Hicks, Bryan and Wickham, especially in the area of teaching and learning data science. With a variety of data generated from different areas in scientific and commercial settings, there is no universal data analysis workflow apply to all scenarios. Domain knowledge is also important which differentiate the required data science techniques. What Zheng proposed sounds like a good way, which is to teach fundamental principles with a capstone course applying these knowledges. Hiring and promoting faculty members in data science is also a challenge for statistics departments. Separating out data science as an independent academic department might be a good idea. Both statisticians and computer scientists should contribute to the development of this new area of research.

Question:
As the definition of science, on Wikipedia, "it is a systematic endeavor that builds and organize knowledge in the form of testable explanation and prediction about the universe." I understand that data science is trying to help explain the universe using data analysis. How important is inference and testable explanations needed in data science?