50 years of Data Science - Reading Guide

Brief

Data science emerged as a call for academic statistics to evolve. At the forefront of this push 'for a science' were Chambers, Cleveland, and Breiman who advocated for this field just 10-20 years ago. Now we find ourselves in an identity crisis and the current Data Science Moment may be adding to the confusion with many institutions, practitioners and organizations each offering their variation of the definition.

What is clear is that it is a disservice to call Data Science a superset of the field of statistics and machine learning. This only continues to add fuel to the Data Science v Statistics dilemma. To a statistician the definition of data science may encompass applied statistics; if the definition of data science is encompassed by the definition of statistics then data science can be seen under the umbrella of statistics. Some may go as far to say data science is just a rebranding of something that already exists (statistics). This paper by Donoho addresses this misconception, but moreover the main aims of this paper are

(i.) to set out a historical context for the gap data science is filling (as a discipline) and

(ii.) build a conceptual framework, a scaffolding if you will, for this new discipline.

Donoho does this by introducing six divisions of concern and uses these to tease apart the complexities of the overlap between these two distinct fields and highlights the profound differences as well. The six divisions for data science include: (these are key, understand these for the quiz/exam)

- Data exploration
- Data transformation
- Computing
- Modeling (Machine learning, statistical learning)
- Data visualization
- The science of data science

He assesses these different divisions by providing insight from some of the most influential statisticians of the previous and modern era: Tukey, Chambers, Cleveland, Wickham, and Xie (note these individuals have substantial industry experience). This paper serves not only to communicate how data science has come to be, but also as a commentary on the field of statistics. These six divisions are not wholly unique to data science, but you can find significant overlap in the day-to-day operations of statisticians.

Perhaps the most important takeaways is the last division: the science of data science. Why is this important you might ask? Often people will use the skills, or the techniques to define data science to make

data science feel tangible. This is a disservice, e.g. does it seem right that I explain computer science as a way to use programs, code, and databases to make computers do interesting tasks? No, this type of definition misses the meat on the bone in favor of the fat. When Donoho goes through the science of data science, he is firmly planting the discipline in solid ground, removed from technology or skill, instead focused on an academic foundation.

So why then do we have this field of data science? Largely this can be answered by education (i.e. what is taught in stats undergrad/grad programs) and also the demands of industry. Engineering and computation have really been at the forefront of the big data era. These skills tend to be more aligned with software engineering. And the fast-paced movement in industry has in many ways left academic training in the dust. Data science has such a large amount of use cases and so much of it hinges on domain expertise. Statistics as a discipline needs to be concerned with teaching generalizable fundamentals.

The idea of a statistics education is deeply rooted in mathematical analysis, Donoho argues for the need of a balance to applications and analysis. This is a hard problem for stats to answer, because applied analysis can be very nebulous. And largely dependent upon the industry, especially when it comes down to tools and techniques. This is why data science can be used to fill that gap. The gap between Statistical research, and applied analysis (along with all the baggage that comes with it e.g. visualization, distributed computing, and so on).

I don't think Donoho falls short of driving home the two main points I outlined above. However, I feel he does coast through some of the concerns that scalability and engineering bring to producing production ready data science projects (easy to criticize in 2020 after the fact...) and end user products. These products are where the rubber really meets the road and there is a great departure from traditional statistics when you consider the infrastructure required to support productionized data products in today's fast paced world. Engineering is perhaps as much a part of data science as statistics may claim.

Important concepts to review for the quiz/exam

Page 10 The future of data analysis: Tukey's introductory paragraphs

Page 11 Tukey's identified four driving forces in the new science

Four major influences act on data analysis today

1. The formal theories of statistics
2. Accelerating developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The emphasis on quantification in an ever wider variety of disciplines

Page 13 Cleveland's proposed 6 foci of activity

Page 15-16 Breiman's 'Two Cultures', Generative v Predictive Modeling

Page 16 The Common Task Framework (CTF)

Page 16-17 "The secret sauce": (CTF) and Predictive Modeling Culture

Page 22-25 The Greater Data Science, Classified into 6 divisions:

- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science

Page 29-32 The Science of Data Science

Page 34 The future of data science: Science as Data