Stats 337 Annotated Bibliography: Data Science as an Academic Discipline

Stephen Bates

June 6, 2018

Executive Summary

Data science is increasingly viewed as stand-alone field of research, diverging from existing academic disciplines of statistics and computer science. Universities see potential in this new field, and many have taken action to start interdisciplinary data science institutes. Going even farther, Yale University renamed its statistics department to the Department of Statistics and Data Science [1], and Stanford University recently created the Department of Biomedical Data Science[2]. Despite the intense interest in data science, there is little consensus about how this new field should be situated within the academy.

As a Ph.D. student in statistics, the emergence of data science as a field and it's relationship with academic statistics is highly salient to me. This annotated bibliography is an attempt to clarify what data science would look like as an academic discipline. The has two main components. First, we examine the subject matter that academic data science would study. Since computer science, statistics, and biostatistics are established academic fields interested in related phenomena, we are particularly interested in identify data science research activities that are outside the mainstream research of these existing fields. Along with the question of "what?" comes the question of "how?". A new academic discipline will most likely demand new ways of organizing members and evaluating work to accomplish its goals. In particular, the traditional peer-reviews journal article is unlikely to be the best format for disseminating data science contributions, and similarly the academic publication record is unlikely to be a fair evaluation of the contributions of a data scientist.

A common thread in the readings is that data science is more centered around software than traditional statistics. Data science contributions would often take the form of software packages, and even the contribution is not software itself, the object of study would often be data processing systems. For example, a data scientist might study how academic code is released with papers and propose a new incentive scheme to lead to better open science. Many data scientists expressed the opinion that academic statistics does not appropriately value the creation of software or the analysis of human-software systems as an intellectual contribution. Waller does a wonderful exploring the the clash of cultures in [Wal18]. Data science as an academic field would encompasses a wider range of activities related to the analysis of data, with more of an emphasis on writing software and analyzing software systems than in statistics.

[1] https://statistics.yale.edu/

[2] http://med.stanford.edu/dbds.html

The readings from more established academic statisticians, in contrast, often take the perspective that data science is nothing but statistics with a little more emphasis on computing. These readings typically encourage statisticians to be more applied, do more computing, and teach more messy data analysis in their courses. Although it's true that data science practice involves both statistics and software, saying that data science is only statistics with a little extra engineering dramatically understates the difference in what data science *research* would look like compared with academic statistics. Coming from academic statistics, I see that in my field, the study of the mathematics probabilistic models is the most highly valued and rewarded area of research. It is not realistic that the field of statistics with its hundreds of tenured faculty members would be able to change focus sufficiently quickly to "keep" data science as part of statistics. The nature of the contributions and, perhaps just as importantly, *the evaluation of work* in data science is sufficiently different for it to comprise a new discipline, albeit one with very close connections to statistics.

Top 3 Papers

"The future of data analysis" –John Tukey

Tukey's "The Future of Data Analysis" [Tuk] is a classic that has inspired many generations of statisticians to evaluate the foundations of their discipline. In this article, one of the most gifted statisticians of all time argues that the field has drifted too far from data analysis toward mathematical sophistry. Although Tukey was a very gifted mathematician famous for introducing the fast Fourier transform, he warns statisticians not to be too wrapped up in the mathematics. He argues that statistics should instead create and evaluate data analysis *tools*, if necessary by categorizing the methods available and creating guides for the practicing data analyst.He concludes by arguing for more empiricism and computing (!) in the field of statistics. For an article that is over 50 years old, this commentary is remarkably forward-thinking and fits perfectly with the modern data science moment. This article is my top choice because it simultaneously provides historical context for the current discussion around data science while contributing many original ideas that ring true today.

"Documenting and evaluating data science contributions in academic promotion in departments of statistics and biostatistics" –Lance Waller

Although opinions on data science abound, nobody takes as thoughtful of a look at the situation of the data scientist in the present-day academy than Waller does in this article [Wal18]. He smartly describes the junior faculty member going up for tenor in a statistics department, and the challenges that the scientist and the promotion chair face in their interactions with the the promotion committee. Waller does a wonderful job painting a convincing cast of characters, from the upstart data scientist to the established senior faculty members. He expertly describes the problem of fitting data science in a modern statistics department and goes on to give some concrete tips for young data scientists who are trying to make it work. The key source of contention seems to be the valuing and evaluation of different kinds of work. There is a disconnect between what the senior members of the field value and what the young data scientist contributes, and furthermore even the sympathetic senior statistician has a hard time evaluating the quality of data science contributions that take the form of software packages and conference publications. It is important for data scientists to articulate why the existing academic structure does not optimally enable data

science research, and this article perfectly captures this idea.

"Software as an academic publication" –Roger Peng

In this insightful blog post [Pen18b], Roger Peng disects the important problem of the place of software in academic research. This topic is crucial for academic data science in at least two ways: (1) credit attribution for scholarly work and (2) dissemination of software design lessons. The last two decades saw the emergence of software journals as an outlet for software-based scholarships. Developers would right a paper accompanying their software package, and then submit it to a software journal. They could then list this as a publication on their CV with their other more traditional journal articles. This was a reasonable compromise at the time, but it requires twice as much work from the developer. Since software is now being taken more seriously, new journals such as the Journal of Open Source Software now publish software as the primary artifact, without a tacked on "paper." This article makes my top three not because the blog post is flashy, but because I believe this is a *very* important topic for academic data science. As Waller argued above, it is challenging for data scientists to be fairly rewarded for software contributions in statistics departments, and a new set of scholarly norms needs to be developed.

References

[ABM+14] Paul Anderson, James Bowring, Renée McCauley, George Pothering, and Christopher Starr. An undergraduate degree in data science: Curriculum and a decade of implementation experience. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, SIGCSE '14, pages 145–150, New York, NY, USA, 2014. ACM.

> The authors describe their experiences with an undergraduate data science program, offering their program as a template for an undergraduate data science major at other universities.

[B+01] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[BW17] Jennifer Bryan and Hadley Wickham. Data science: A three ring circus or a big tent? *Journal of Computational and Graphical Statistics*, 26(4):784–785, 2017.

> Bryan and Wickham, both partially academic statisticians with one foot outside the academy, largely agree with Donoho's article, but the offer diverging opinions about three issues. First, they point out that the misalignment between academic statistics and the practice of modern data science is largely self-imposed. They point out that contributions in the form of mathematical theory and methods are rewarded in the academic culture, whereas other contributions are not. In my opinion, this is mostly just a difference in emphasis to that of Donoho's article.
> Secondly, they disagree with Donoho that learning or developing computational frameworks is a superficial aspect of data science. They agree that some computational tools for big data are specialized and not necessary for all data scientists, but they say this is also true of e.g. measure theory. They claim that practical aspects of software development are fundamental to modern

data analysis and are not sufficiently emphasized in Donoho's formulation. Lastly, they argue that the conceptual contributions of software such as the tidyverse are major accomplishments, despite not being mathematical. While the ideas take form as R code and explicitly deal with computation as part of the subject matter, they transcend any one implementation. Furthermore, they stress that the ecosystem of standards collectively represents a contribution that is greater than the sum of the individual pieces. Overall, I think the authors provide a valuable description of the way in which software engineering is fundamental to modern data science and a concrete description of nature of the contribution of recent data science work.

[Cha93] John M Chambers. Greater or lesser statistics: a choice for future research. *Statistics and Computing*, 3(4):182–184, 1993.

[Cle01] William S Cleveland. Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1):21–26, 2001.

[Com] Computing Research Association. Computing research and the emerging field of data science. `https://cra.org/data-science/`.

In this 2016 post the CRA lays out the broad research questions posed by data science. In contrast to the viewpoint of statisticians, this article mostly emphasizes the data engineering aspects of data science, such as representing data and handling large volumes of data. Only one of their 5 key research areas reference statistics.

[CT] David Culler and Two Sigma. A conversation on data science with prof. david e. culler. `https://www.twosigma.com/insights/article/a-conversation-on-data-science-with-prof-david-e-culler/`.

David Culler shares his thoughts about the intersection of data science with the academy. He notes that activities at the intersection of statistics and databases have been happening at Berkeley for at least a decade, and goes on to describe Berkeley's introductory data science class. He emphasizes that the course always introduces statistical ideas in conjunction with computational problems, unlike traditional statistics courses. He points to data ethics and data provenance as a key research question for data science.

[Dav13] Marie Davidian. Aren't we data science? `http://magazine.amstat.org/blog/2013/07/01/datascience/`, 2013.

In this 2013 blog post, Davidian voices some of the surprise from the statistical establishment that data science seems to often ignore statistics. She shares an anecdote of a data science initiative that included dozens of computer scientists and engineers, but no statisticians. She points out that most tools and even software in use in data science came from statistics. She ends with a exhortation to statisticians to embrace computing in their curriculum.

[DL12] Marie Davidian and Thomas A. Louis. Why statistics? *Science*, 335(6077):12, 4 2012.

Marie Davidian and Thomas Louis argue that statistics is becoming more and more central to data-driven science. They argue that many more statisticians are needed, and that statisticians should be more involved in scientific problems from the beginning to enable more well-designed experiments.

[Don17] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

A wonderful modern day companion to Tukey's article. Donoho examines the current state of data science initiatives across universities, and further exames how the "common task framework" has been an effective way of evaluating work in machine learning.

[GFtSF] Gordan, Betty Moore Foundation, and the Sloan Foundation. Moore-sloan data science environments: Themes. `http://msdse.org/themes/`.

The Gordan and Betty Moore Foundation and the Sloan Foundation partnered to create data science environments within three universities. They bring up six topics related to the institutionalization of data science within universities: careers, education and training, tools and software, physical and intellectual space, and data science studies. They do not give comprehensive analysis of any of these areas, but they note that data science will require new positions, institutional policies, and community organization than existing academic fields.

[HJ17] Susan Holmes and Julie Josse. Discussion of '50 years of data science'. *Journal of Computational and Graphical Statistics*, 26(4):768–769, 2017.

[HMV13] Harlan Harris, Sean Murphy, and Marck Vaisman. *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. " O'Reilly Media, Inc.", 2013.

[HV17] Heike Hofmann and Susan VanderPlas. All of this has happened before. all of this will happen again: Data science. *Journal of Computational and Graphical Statistics*, 26(4):775–778, 2017.

In this response to Donoho's "50 years of data science," Hofmann and VanderPlas attempt to put the current tensions between data science and statistics in context of other historical field divisions. They compare the current moment with the formation of statistics departments from mathematics departments, and the formation of computer science departments *not* from mathematics but rather from engineering. They other a very slight reformulation of Donoho's characterization of data science. More interestingly, they claim that the "prediction versus inference" debate in Donoho's article and elsewhere is really a "parametric versus algorithmic" debate. The authors claim that parametric models can equally well be used for prediction, and that the bias in statistics departments is toward parametric models, not necessarily away from prediction. They argue in favor of parametric methods for interpretability reasons. Although there is certainly a preference for parametric methods in statistics departments, it still seems to me that there is a preference for inference problems in statistics departments, and I am not fully convinced by their arguments. The authors close with a exhortation to make data analysis more

central to the the statistics curriculum, including often overlooked practical aspects such as data cleaning.

[Jor] Michael Jordan. Artificial intelligence— the revolution hasn't happened yet. https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7.

[Kel] Adam Kelleher. Causal data science. https://medium.com/causal-data-science/causal-data-science-721ed63a4027.

This medium post gives a few examples of how to do causal inference in industry data science. The article quotes a lot of tools from the statistical literatures, but avoids ever mentioning the field of statistics. The article includes both statistical concepts *and* touches on implementations in python.

[Mar14] M. Martone. Data citation synthesis group: Joint declaration of data citation principles, 2014.

The authors propose a set of principles governing data citations.

[MHB17] Amelia McNamara, Nicholas J. Horton, and Benjamin S. Baumer. Greater data science at baccalaureate institutions. *Journal of Computational and Graphical Statistics*, 26(4):781–783, 2017.

[MNC+18] David Moher, Florian Naudet, Ioana A. Cristea, Frank Miedema, John P. A. Ioannidis, and Steven N. Goodman. Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, 16(3):1–20, 03 2018.

[Pen17] Roger D. Peng. Comment on 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):767–767, 2017.

Peng's response to the famous article by Donoho broadly voices his support for the original article, but adds depth regarding at least two points. Peng comments that it is difficult to teach data cleaning, for example, because it is hard to find general lessons that will apply to future data analysis. Statistical models are useful across many data analysis, whereas data cleaning techniques are specialized, one-off procedures. Peng hypothesizes that because teaching data cleaning is inefficient insofar as few lessons generalize well, statisticians have shied away from teaching these courses.
Although drawing general lessons is not always possible, Peng points to "tidy data" as introduced by Wickham as useful formalism that has contributed a lot to the practice of data analysis because the lessons generalize to many different problems. Despite the lack of mathematization, Peng considers this a seminal contribution to statistics, and is enthusiastic about other similar contributions. Peng closes with a remark that many people cited as revolutionizing data science do not spend their time exclusively in the academy, and he warns statisticians not to miss out on important contributions simply because they do not take a familiar form.

[Pen18a] Roger Peng. Rethinking academic data sharing, 2018.

[Pen18b] Roger Peng. Software as an academic publication, 2018.

[Stu] Lilah Sturges. Plugging the data science talent gap in academia and industry. https://www.elsevier.com/connect/plugging-the-data-science-talent-gap-in-academia-and-industry.

A panel at the HPCC systems summit discuses the perceived talent gap in data science both within universities and industry. They argue that funding from traditional funding sources is insufficient to give data science the fuel attention it deserves. They encourage an interdisciplinary approach to data science where the data scientists work closely with other academic disciplines.

[Tuk] John W. Tukey. The future of data analysis. *Ann. Math. Statist.*, (1):1–67, 03.

[Wal18] Lance A. Waller. Documenting and evaluating data science contributions in academic promotion in departments of statistics and biostatistics. *The American Statistician*, 72(1):11–19, 2018.

[Yu14] Bin Yu. Let us own data science. http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/, 2014.

In her 2014 address to the Institute of Mathematical Statistics, Bin Yu lays out her vision of data science. She describes data science as the intersection of statistics, domain expertise, computing, and collaboration. Her definition is quite similar to definition used by non-statistician data scientists. Yu suggests re-branding statistics as data science because the new name has garnered attention and is just as appropriate as statistics. She concludes by exhorting her statistical colleagues to place more emphasis on computing in their work and teaching.