BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 56, Number 1, January 2019, Pages 143–149
http://dx.doi.org/10.1090/bull/1596
Article electronically published on September 5, 2017

Algorithms for data science, by Brian Steele, John Chandler, and Swarna Reddy, Springer, Cham, 2016, xxii+430 pp., ISBN 978-3-319-45797-0, Hardcover US $79.99, eBook US $59.99

When trying to define data science, the ancient Buddhist parable of the blind men and the elephant springs to mind. The entrepreneur, the academic researcher, and the university administrator approach the beast in turn and examine it, but report something different. The entrepreneur sees the training of a generation of data workers to deal with large data sets as a business opportunity. The university administrator hears the siren's call of new programs and students, while the academic researcher is still trying to figure out where the science is. An exact definition of data science remains elusive.

As early as 1962 in *The future of data analysis* [15], John W. Tukey, a chemist turned topologist and finally statistician, wrote: "For a long time, I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt... I have come to feel that my central interest is in data analysis." And later, "Data analysis, and the parts of statistics which adhere to it, must... take on the characteristics of science rather than those of mathematics—data analysis is intrinsically an empirical science." Later, in 1977, he published *Exploratory data analysis* [16], arguing that exploratory and confirmatory statistics are the two complimentary poles of statistics research. This sentiment was echoed years later in Leo Breiman's *Statistical modeling: two cultures* [2], where he chastised statisticians for focusing only on confirmatory analysis (which he referred to as the "data modeling culture") to the exclusion of exploratory analysis (which he called the "culture of algorithmic data modeling"). He went on to note that other fields, notably machine learning and computer science, were rapidly filling this gap. In the same year, Bill Cleveland, then of AT&T Bell Labs, exhorted the statistics community similarly in *Data science: an action plan for expanding the technical areas of the field of statistics* [4], perhaps the first time the term "data science" appeared in print.

According to the Data Science Association (DSA), "Data Science means the scientific study of the creation, validation and transformation of data to create meaning. A "Data Scientist" is a professional who uses scientific methods to liberate and create meaning from raw data—somebody who can play with data, spot trends and learn truths few others know." In the paper *50 years of data science*, presented at the Tukey Centennial Workshop in 2015 [8], David Donoho claimed that to a statistician this sounds "an awful lot like what applied statisticians do." Indeed, the American Statistical Association (ASA) and Institute for Mathematical Statistics (IMS) at first reacted defensively with such articles as *Aren't we data science* [5] (column of ASA President Marie Davidian, AmStat News, July 2013), and *Let us own data science* [17] (IMS presidential address of Bin Yu, reprinted in the IMS Bulletin, October 2014). However, the ASA later came to the view that statistics is a necessary part of, but does not encompass what is currently understood as, the

2010 *Mathematics Subject Classification.* Primary 62-07, 68-01, 68P99, 68Q01, 68W01.

field of data science. In their statement *ASA statement on the role of statistics in data science*) [9], they state,

> While there is not yet a consensus on what precisely constitutes data science, three professional communities are emerging as foundational to data science: (i) Database Management to enable the transformation, conglomeration, and organization of data resources; (ii) Statistics and Machine Learning to convert data into knowledge; and (iii) Distributed and Parallel Systems to provide the computational infrastructure to carry out data analysis. Statistics and machine learning certainly play a central role in data science.

Because much of the impetus for data science comes from entrepreneurial rather than academic motivation, it is difficult to document. But it appears that a consensus for the practical skills of data science, and thus the training in the field, has emerged as some combination of the skills of statistics, computer science, machine learning, mathematics, and substantive domain knowledge in a field (often business). As the DSA says, "The data scientist has a solid foundation in machine learning, algorithms, modeling, statistics, analytics, math and strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge."

The current state of the academic field of data science is less clear. Donoho [8] suggests that there is "a solid case for some entity called 'Data Science' to be created, which would be a true science: facing essential questions of a lasting nature and using scientifically rigorous techniques to attack those questions." Michael Jordan in his 2016 keynote address [11] hearkened back to Breiman's two cultures, imagining a science "between" statistics and computer science that requires both "computation" and "inferential" thinking. He pointed out, "core theories in computer science and statistics developed separately, and there is an oil and water problem." By this he meant that in traditional statistics research, considerations of runtime efficiency and computational resources are overlooked, while in core computational theory there is no appreciation for variability and statistical risk. His vision of a true data science would be one that used the scientific method to attack and solve problems that acknowledge and contain both challenges.

Whatever data science is, or turns out to be, it is clearly "hot". The now-famous McKinsey report of 2011 [12] predicted a shortage of 140,000 to 190,000 people with "deep analytical skills" by 2018. More recently, IBM predicted that "by 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000." The lack of a firm vision for academic data science has not stopped business and academic leaders from plunging headlong into the world of data science training. Since the first academic programs (then called analytics) first came online in about 2007,[1] over 500 programs in data science have sprung up, according to the list curated by the Data Science Community.[2] The vast majority of these are master's degree programs, underscoring the practical training aspect on the training and volumes about the state of the art. In fact, of the 563 listed programs, 396 are

[1] Institute for Advanced Analytics, North Carolina State University, `http://analytics.ncsu.edu/?page_id=2`

[2] Data Science Community, `http://datascience.community/colleges`

master's degree programs and another 95 are "certificate" programs, with only 21 doctoral programs.

The variation in the offerings of the 396 master's degree programs is impressive. We certainly didn't have the time to peruse them all, but they range from courses traditionally taught in departments of statistics, computer science, mathematics, and engineering with a mixture of statistics, database systems, visualization, and models courses (as at Johns Hopkins' Whiting School of Engineering) to MBA programs with courses in such areas as profit-focused pricing, game theory, financial engineering, mobile marketing, and data warehousing (as at the University of Santa Clara's Leavey School of Business). Identifying and agreeing on a core set of courses or subjects among these programs seems highly unlikely. The entrepreneurial nature of most programs all but ensures a plethora of paths, courses, and outcomes as a result. Is there a core of such courses? An NSF workshop report on data science education [3] entitled *Strengthening data science education through collaboration* concluded "... there is a widespread agreement that a core would include elements of statistics and elements of computer science. Clearly, data science is inherently interdisciplinary." They went on to recommend that the NSF fund a "broadly interdisciplinary task force to develop curriculum guidelines for degree programs in data science."

University administrators have shown increased interest in establishing undergraduate programs in data science. As of this writing, there are 50 programs listed among the 563 mentioned above, and many are coming online every day. In an attempt to provide some structure for these programs, a Park City Mathematics Institute (PCMI) group met in 2016 to formulate curricular guidelines for undergraduate data science. Their report, *Curriculum guidelines for undergraduate programs in data science* [6], identifies a set of existing core courses typically found in departments of mathematics, computer science, and statistics as a "transition" to a future data science major. They outline a possible set of new and reconfigured courses as a blueprint for future programs (see the appendix to the report [7]). The future curriculum would involve a new alignment of the core topics and a new set of textbooks to cover this material in an integrated set of ten courses. Two of the major challenges of devising such a curriculum involve the decision of what to teach (and more importantly, what not to teach), and in what order. The temptation to teach the superset of topics in a mathematics, statistics, and computer science triple major is strong, but it is certainly not feasible, nor would it be desirable. The data science "unicorn"—a person with a deep understanding of all three areas, plus domain knowledge of another field—is just that, a myth, or, at least, very rare, and not a realistic goal for an undergraduate degree.

The book under review, *Algorithms for data science*, is a recent attempt to start filling the gap in textbooks for the undergraduate data science major. At this time, there are only a few other books that cover similar material, notably the books by Baumer, Kaplan, and Horton [1] and Nolan and Temple Lang [14]. We suspect that there will be flood of such books in the coming decades. New curricula are being proposed but are difficult to implement until there are textbooks designed for the new courses. And new textbooks cannot be written until the curriculum has stabilized. The field is growing fast, but this Catch 22 (or circular dependency) of courses and textbooks is a serious issue. The book under review is one of a group of pioneers into this landscape that tackles the problem head on.

What course is this a book for? As the book states,

> . . . the knowledge base incorporates demanding topics from statistics, computer science and mathematics. On top of that, domain-specific knowledge. . . Preparing students in these three or four areas is necessary. But at some point, the subject areas need to be brought together as a coherent package in what we consider to be a course in data science. This book serves as a backbone for [such a] course.

The audience for the book is either a "practitioner of data science and the allied areas of statistics, mathematics, and computer science" or an " upper-division undergraduate or graduate student in data science, business analytics, mathematics, statistics, and computer science".

So, it seems that this is an introduction to data science as a whole, but it is not a first-year course. Rather, it appears to be an introduction to data science for students who have already committed to one of the three main fields of mathematics, statistics, or computer science. In fact, the prerequisites, in spite of the authors calling them "low", are quite substantial: "The reader with one or two courses in probability or statistics, an exposure to vectors and matrices, and a programming course should succeed at mastering most of the content."

The book is structured into three parts: I. Data Reduction, II. Extracting Information from Data, and III. Predictive Analytics. In part I the authors cover an introduction to computing, the basics of algorithms and then some recent data architectures, namely Hadoop and MapReduce. Part II is a mini course in statistics topics, covering data visualization, linear regression, and cluster analysis. Part III is a mixture of machine learning topics like k-nearest neighbors, naïve Bayes methods, forecasting, and an introduction to streaming data. More sophisticated algorithms of data science, including neural networks, support vector machines, trees, and ensemble methods of bagging and boosting, are ignored. The authors explain that they wanted to restrict attention to algorithms that could be easily coded. While we understand the logic, that seemed a shame. These algorithms are the core of the data analytic part of data science.

One of the choices for a data science curriculum, and thus for the accompanying text, is the choice of a computer language. For the computer scientist, that is most likely *Python.* For the statistician, it is **R**. In many ways these languages are complementary with different strengths, weaknesses, and purposes. In a data science curriculum, the student will most likely encounter both and will (hopefully) learn what each does well and thus when to use each. The authors have decided to use both in the book, starting with *Python* in Chapter 2 (in contrast to the Baumer et al. book [1] which focuses exclusively on **R**). For the *Python* novice, there is far too little introduction, so the warning in section 1.8 should certainly be heeded: "If you are not already familiar with a computer language, you should immerse yourself in a self-study *Python* program for a week or two. If you have not used *Python* before, then you will also benefit from some time spent in self-study." To their credit, the authors provide a nice set of current URLs for becoming proficient in *Python.*

Without an integrated curriculum in place, one of the challenges of trying to teach the principles of data science is the diversity of the audience itself. The authors attempt to fill in gaps in linear algebra with a section 1.10.1 on matrices and vectors. Perhaps similar sections on *Python*, **R**, and some elementary statistics would have saved them explanation later on, but it's not clear that a course on coding of algorithms, statistical analysis, and predictive analytics can be taught effectively to such a diverse set of backgrounds in one book or semester. For a one-semester course in data science, the choice of topics is daunting. We were confused by the section on "associative statistics" (section 3.3). This is not a term used in the field of statistics, and it is easily confused with measures of association (e.g., correlation). Rather, by associative, the authors mean that the computation "scales up" on large data sets. A statistic that is "associative" is one that can be computed separately on each partition of a data set and then collected together to produce the value that would be obtained on the whole set. So the mean is associative, but the median is not. This has implications for efficient coding of statistical algorithms, but we doubt that a student will have a sense of its worth as a concept so early in the course. Then the rest of the section, including a 14-page tutorial, is devoted to constructing a histogram. This seems to be, in a nutshell, the problem with trying to be something to everyone. For the statistics student familiar with **R**, the response will be, "Why spend 14 pages to learn to program an algorithm when I can just type >hist(data) in **R**?" And in fact, in chapter 5, (before introducing **R**), the authors write "The best tool for creating data visualizations in the context of an analysis is Hadley Wickham's ggplot2 available in **R**." So why not start there? Creating a histogram from scratch is painful and is unlikely to give the statistician the inspiration for learning *Python*, nor the computer science major the inspiration for spending 14 pages to learn to make bars.

Chapter 4 is an introduction to modern distributive database architectures for big data with an emphasis on Hadoop and MapReduce running on Amazon Map Services. In an attempt to be current, the authors may have gone too far with a specific choice. The field is changing fast, and while it is important to point out to the beginning student (who has just learned how to code a histogram) that there are scalability issues, the chapter seemed somewhat out of place.

Chapter 5 is an introduction to data visualization. Here, the authors decide not to provide much code but to rely on ggplot. There are many books on this subject, and while the examples are quite good, our experience in teaching both ggplot and the principles of good data visualization tells us that it will require far more time and space than the book allots. The disconnect between the 14-page tutorial on constructing a histogram and the fast pace of the ggplot code here was striking. Strangely, the introduction to **R** is found in the next chapter.

There are similar problems of level throughout the book. Chapter 6 is an introduction to linear models. Having taught regression theory to junior statistics majors at Williams for two decades and to engineers at Princeton for a decade before that, one of the reviewers (Richard) knows that a 16-page summary of conditional expectation, linear models, parameter estimation, standard error of the estimates, and inference is insufficient: confidence intervals and hypothesis tests for parameters will leave all but the best and brightest in the dust.

The rest of the book covers many of the topics of *An introduction to statistical learning* by James et al. [10] with the addition of a chapter on time series and one on streaming data. The difference between this book and both the James and Baumer books is the computer science perspective and the emphasis on coding the algorithms. This perspective benefits those well versed in the practicalities of programming who are looking for a guide to implementing a specific algorithm. The examples, with plenty of documentation, on data retrieval and data wrangling are interesting, plentiful, and well thought out. One of the strengths of the book is how thoroughly it covers the topic of converting real-world data into a workable form, a task known as "data munging". Readers are provided with detailed commands for reading data in from .txt, .csv, .zip, and pickle data files, are taught how to leverage metadata files to interpret categorical values, and are given advice on when to exclude outlier samples from analysis.

With perhaps less introduction to statistics and a little more help starting in *Python*, this might be a nice book for introducing statistics majors to the field. The orientation on teaching how the algorithms are coded rather than a statistical explanation makes the mechanics of implementation very clear, but we wonder what the student with a limited statistics background will take away. This book's de-emphasis on theory along with its conversational tone blurs the line between arbitrary decisions made by the authors (choice of algorithm, choice of dictionary mapping, choice of representation) and decisions about the data informed both by domain knowledge and experience in statistical modeling. The casualness with which the text moves from model, to code, to results on an example data-set, and back to model, could lead to confusion for those without a sufficient background in uncertainty and statistical risk.

One of the book's strengths lies in the choice of examples and tutorials (with the caveat that specific URLs may break as time moves on). The authors have selected a variety of interesting, real-world data sets to illustrate the algorithms and for students to try their skills upon, such as interpreting the Federalist Papers using Naïve Bayes, or attacking the CDC's Behavioral Risk Factor Surveillance System Survey using hierarchical clustering. The authors do not shy away from using large and complex data sets that, at several gigabytes, will expose the students to the trials of working with Big Data. An interactive notebook of code (perhaps in Jupyter, an interpreter for both **R** and *Python*), which was not available at the time of the review, would be a wonderful addition.

The scope of the topics seems a little ambitious, and it is not clear that the stated goal to provide a foundational course in data science has been met. But with the state of data science curriculum at present, we will need multiple attempts until the cycle of course selection and appropriate textbooks is more stable. What the book does deliver is a broad set of hands-on tutorials on topical data sets that will give students needed practice in the process of data science. The breadth of the algorithms and examples, such as the real-time analysis of twitter feeds, the large-scale American health survey, and the look at PAC political spending, will be of interest to students of data science. We applaud the authors for their effort and look forward to seeing the evolution of offerings at this level in the future.

References

[1] B. S. Baumer, D. T. Kaplan, and N. J. Horton, *Modern data science with R*, Chapman and Hall/CRC Press, 2017. `http://mdsr-book.github.io`

[2] L. Breiman, *Statistical modeling: The two cultures*, Statistical Science **16** (2001), no. 3, 199–231.

[3] B. Cassel and H. Topi, *Strengthening data science education through collaboration.* Report on Workshop on Data Science Education, 2015, Funded by the Natl. Sci. Found., Oct. 3–5, Arlington, VA.

[4] W. S. Cleveland, *Data science: An action plan for expanding the technical areas of the field of statistics*, International Statistics Review **60** (2001), no. 1, 21–26.

[5] M. Davidian, *Aren't* we *data science?*, AmStat News, "President's Corner", July 1, 2013.

[6] R. De Veaux, et al., "Curriculum guidelines for undergraduate programs in data science", *Annual Review of Statistics and its Applications*, Vol. 4, pp. 15–30, 2017. `http://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-060116-053930`

[7] R. De Veaux, et al., "Curriculum guidelines for undergraduate programs in data science: Appendix—Detailed courses for a proposed data science major", *Annual Review of Statistics and its Applications*, Vol. 4, appendix, 2017. `http://www.annualreviews.org/doi/suppl/10.1146/annurev-statistics-060116-053930/suppl_file/st04_de_veaux_supmat.pdf`

[8] D. Donoho, *50 years of data science*, presentation at the Tukey Centennial Workshop, Princeton, NJ, September 18, 2015. `http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf`

[9] D. van Dyck, M. Fuentes, M. Jordan, M. Newton, B. K. Ray, D, Temple Lang, H. Wickham, *ASA Statement on the role of statistics in data science*, AmStat News, October 1, 2015.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning. With applications in R*, Springer Texts in Statistics, vol. 103, Springer, New York, 2013. MR3100153

[11] M. I. Jordan, "On computational thinking, inferential thinking and data science", *Proceedings of the 28th ACM Symposium on Parallelism in Algorithms and Architectures*, keynote address, 2016. `http://dl.acm.org/citation.cfm?id=2935826`

[12] McKinsey & Company, *Big data: The next frontier for innovation, competition, and productivity*, 2011. `http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation`

[13] McKinsey & Company, *The age of analytics: Competing in a data-driven world*, December 2016. `http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world`

[14] D. Nolan and D. Temple Lang, *Data science in R: A case studies approach to computational reasoning and problem solving.* Chapman and Hall/CRC Press, 2015.

[15] J. W. Tukey, *The future of data analysis*, The Annals of Mathematical Statistics **33** (1962), no. 1, 1–67.

[16] J. W. Tukey, *Exploratory data analysis*, Addison-Wesley, 1977.

[17] B. Yu, *Let us own data science*, IMS Bulletin Online, October 1, 2014.

Richard D. De Veaux
Williams College
E-mail address: `rdeveaux@williams.edu`

Nicholas R. De Veaux
Center for Computational Biology, Flatiron Institute, Simons Foundation
E-mail address: `nrdeveaux@gmail.com`