PERSPECTIVE

Science and data science

David M. Blei[a,b,c,1] and Padhraic Smyth[d,e]

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved June 16, 2017 (received for review March 15, 2017)

Data science has attracted a lot of attention, promising to turn vast amounts of data into useful predictions and insights. In this article, we ask why scientists should care about data science. To answer, we discuss data science from three perspectives: statistical, computational, and human. Although each of the three is a critical component of data science, we argue that the effective combination of all three components is the essence of what data science is about.

data science | statistics | machine learning

The term "data science" has attracted a lot of attention. Much of this attention is in business (1), in government (2), and in the academic areas of statistics (3, 4) and computer science (5, 6). Here, we discuss data science from the perspective of scientific research. What is data science? Why might scientists care about it?

Our perspective is that data science is the child of statistics and computer science. While it has inherited some of their methods and thinking, it also seeks to blend them, refocus them, and develop them to address the context and needs of modern scientific data analysis. This perspective is not new. Over 50 years ago, Tukey (7) defined "data analysis" as a broad endeavor, much broader than traditional mathematical statistics. In a sense, today's data science, although set against a modern backdrop, is cast from Tukey's original mold.

In modern research, scientists from diverse disciplines are confronting abundant datasets and are confident that there is value in the data for advancing their scientific goals. We give three examples at genomic, social, and galactic scales. First, modern sequencing technology has enabled high-resolution genetic sequencing at massive scale, and geneticists have connected the genetic data to large databases of individuals' behaviors and diseases. These data can potentially aid researchers in studying the human genome, helping them understand how it evolves, and how it governs observed traits. Second, social scientists now have the opportunity to study large archives of digitized texts, often with rich information about human behavior and interactions. These data could help them more effectively navigate and understand the contours of society, finding relevant sources to their work and identifying hard to spot patterns of language that suggest new interpretations and theories. Third, modern telescopes create digital sky surveys that have transformed observational astronomy, generating hundreds of terabytes of raw image data about billions of sky objects. A catalog of these objects, if available, would give astronomers an unprecedented window into the structure of the cosmos.

These examples paint a picture of what might be possible in the modern sciences. However, an issue that pervades many, if not all, scientific disciplines is that scientists cannot yet fully take advantage of their new data. Connecting genes and traits at large scale is a problem that is beyond the limits of classical genome analysis, both computationally and statistically. Building tools for navigating large collections of documents, especially ones that reflect the priorities of social scientists, is a problem that is not solved by classical methods of document analysis. Using digital sky surveys to understand the complex nature of the universe requires computational tools and statistical assumptions beyond those used for the manually curated studies of earlier eras.

Broadly speaking, there is a tension emerging—the existing methods from statistics and computing are not set up to solve the types of problems that face modern scientists. Some issues are computational, such as working with massive datasets and complex metadata. Some issues are statistical, such as the rich interactions of many related variables and the

[a]Department of Computer Science, Columbia University, New York, NY 10027; [b]Department of Statistics, Columbia University, New York, NY 10027; [c]Data Science Institute, Columbia University, New York, NY 10027; [d]Department of Computer Science, University of California, Irvine, CA 92697; and [e]Department of Statistics, University of California, Irvine, CA 92697

Author contributions: D.M.B. and P.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: david.blei@columbia.edu.

theoretical and practical difficulties around high-dimensional statistics. Finally, some issues are fuzzier and philosophical, such as necessarily misspecified models of the world, difficulties in identifying causality from empirical data, and challenges to meeting disciplinary goals around data exploration and understanding.

We believe that this tension has been the catalyst for the new moniker data science. Data science focuses on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention. It emphasizes the value and necessity of approximation and simplification. It values effective communication of the results of a data analysis and of the understanding about the world that we glean from it. It prioritizes an understanding of the optimization algorithms and transparently managing the inevitable tradeoff between accuracy and speed. It promotes domain-specific analyses, where data scientists and domain experts work together to balance appropriate assumptions with computationally efficient methods.

Below, we explore these ideas from statistical, computational, and human perspectives, identifying which views and attitudes we can draw from to develop data science for science. Statistical thinking is an essential component. Statistics provides the foundational techniques for analyzing and reasoning about data. Computational methods are also key, particularly when scientists face large and complex data and have constraints on computational resources, such as time and memory. Finally, there is the human angle, the reality that data science cannot be fully automated. Applying modern statistical and computational tools to modern scientific questions requires significant human judgment and deep disciplinary knowledge.

Statistical Perspective

Discussions about data science often focus on the large-scale aspects of data and computation. These issues are important, but this focus misses that the foundational goals of data science rely on statistical thinking. Since its inception, statistics has served science to guide data collection and analysis. While many aspects of the relationship between science and data have changed—the domains where we use data analysis, the scale of the data, and the nature of the scientific questions—the basic principles are the same.

Broadly, statistics is about developing methods for making sense of data. As the field has evolved, these methods have largely been cast in the languages of mathematics and probability. Statistics uses a variety of functional and distributional assumptions to model relationships between variables and entities in the real world, and it uses observed data to draw inferences and make predictions about such relationships.

All datasets involve uncertainty. There may be uncertainty about how they were collected, how they were measured, or the process that created them. Statistical modeling helps quantify and reason about uncertainties in a systematic way. It provides tools and theory that guide the inferences and predictions for specific problems and real data. Statistics relates to data science through multiple statistical subfields. Here, we discuss three: complex and structured data, high dimensionality, and causality.

Modern datasets are complex. For example, consider a research problem involving climate data. There may be different types of dependencies in the data: dependencies over time, dependencies across multiple spatial scales, and dependencies among different variables, such as rainfall, pressure, and temperature. Statistics provides a rich language for parsimoniously modeling such dependencies. This language helps encode knowledge of the world into formal probability distributions, share statistical strength across related components of a problem, and capture sequential and spatial regularities among the variables. Many of these benefits are found in Bayesian statistics (8), a framework that helps articulate assumptions about data in a formal model and then prescribes the corresponding methods for analyzing data to make inferences about the world. Bayesian methods and related techniques for expressive probability modeling (9, 10) have the potential to provide the necessary tools for blending scientific domain knowledge with statistical inference from data.

A related subfield of statistics concerns high-dimensional data, where we measure thousands or even millions of variables per data point. As scientific measurement has become increasingly sophisticated, statistical inference from high-dimensional data has become more important to many scientific disciplines. To handle such data, statisticians and computer scientists have developed powerful methods involving robustness, regularization, and stability (11). Furthermore, high-dimensional data often arise in pattern recognition problems, where we make a prediction about an unknown variable based on a large set of related variables or parameters. Machine learning techniques (12), such as deep learning (13), have been particularly effective in this context. They provide flexible ways that the target variable can depend on the predictors, and they can now scale up to very large datasets.

The implicit promise of rich datasets is that they can help deepen our understanding of how the world works, and using data to attain such understanding is the lofty goal of causal inference. Statistical thinking about causality stretches back to the late 1800s, with the development of influential ideas around the difference between correlation and causation and how to design meaningful experiments. Today, causality has grown into a rich field (14–16), with significant contributions from computer science, the social sciences, and statistics. Developing new methods in causal inference—how to scale up to large datasets, how to develop inferences from observational data, how to develop inferences from interacting data (as in a social network), and how to design experiments in the computer age—is a ripe avenue for statistical contributions to data science.

Computational Perspective

Statistical thinking provides methods to answer scientific questions with data. Computational thinking focuses on the algorithmic implementation of those methods, and it provides a way to understand and compare their computational footprints. Computational thinking is particularly important in modern data analysis, where we frequently face a tradeoff between statistical accuracy and computational resources, such as time and memory.

One well-known example of computational thinking revolves around optimization (17). Many data science methods involve maximizing a function of the data. (A primary example of this is when we try to maximize the likelihood of the data with respect to parameters of a probability model.) The most common way to maximize a function is to climb it, iteratively computing the direction to travel and moving its free parameters along that direction. In the context of optimization, computational thinking involves understanding how to best compute the direction, when approximate directions suffice, how far to walk at each iteration, and how much accuracy we sacrifice when we stop climbing early to save computation.

Another example of computational thinking is sampling methods. Sampling methods help compute approximate solutions of data analysis problems where the exact solutions are too complex for direct mathematical analysis. For example, the bootstrap (18) is a way to calculate confidence intervals in very complex situations. It repeatedly samples from the data to approximate the types of quantities that would be impossible (or nearly impossible) to analytically derive. The bootstrap, in its simplicity, has had a major impact on the practice of statistics in modern science. Another widely used application of sampling is in Bayesian data analysis, where one of the most prevalent computational methods is Markov chain Monte Carlo (MCMC) (19, 20). MCMC algorithms sample the parameters of a statistical model to produce approximate posterior distributions, distributions of hidden quantities conditioned on the data. Like the bootstrap, MCMC transforms difficult mathematical calculations into sampling-based procedures. Since the 1990s, this transformation has opened the door to otherwise unimaginable models, methods, and applications for Bayesian statistics.

A final example of computational thinking is in scaling data analysis with distributed computing (21, 22). We can now distribute large datasets across multiple processors (for speed) and multiple storage devices (for memory), and there is a variety of software to support distributed computation. Advances in distributed computing build on 1970s research in large-scale scientific computing as well as more recent innovations developed in the technology industry. The same ideas that allowed technology companies to scale their methods to the growing Internet can allow scientists to scale to their growing datasets.

These examples are just a few of the ways that computational thinking plays a role in data science. More broadly, computational thinking helps guide how we account for resources when analyzing data. While statistical thinking offers a suite of methods for understanding data, computational thinking provides the crucial considerations of how to balance statistical accuracy with limited computational resources (23).

Human Perspective

We described statistical thinking and computational thinking, two essential components of data science that provide general tools for analyzing data. The art of data science is to understand how to apply these tools in the context of a specific dataset and for answering specific scientific questions.

Data science blends statistical and computational thinking, but it shifts their focus and reprioritizes the traditional goals of each. It connects statistical models and computational methods to solve discipline-specific problems (24, 25). In particular, it puts a human face on the data analysis process: understanding a problem domain, deciding which data to acquire and how to process it, exploring and visualizing the data, selecting appropriate statistical models and computational methods, and communicating the results of the analyses. These skills are not usually taught in the traditional statistics or computer science classroom but instead, are gained through experience and collaboration with others.

This perspective of data science is holistic and concrete. For each scientific problem, the data scientist develops an understanding of its context: how the data were collected, existing theories and domain knowledge, and the overarching goals of the discipline. Crucially, the data scientist solves the problem iteratively and collaboratively with the domain expert. (We note they do not need to be two different people; the data scientist and domain expert could simply be two "hats" for the same person.) Together, they develop computational and statistical tools to explore data, questions, and methods in the service of the goals of the discipline.

As an example, consider a computational neuroscientist. New imaging technology lets her image mice neurons while they act in a maze with other mice. Ample funding and equipment let her run hundreds of mice, resulting in terabytes of video data and brain imaging data. With a data scientist, she might develop methods that test existing theories of mouse behavior, produce hypotheses about how behavior is controlled by the brain, and algorithmically handle the high resolution and complexity of the video and brain data. Furthermore, the data scientist helps develop methods that address limitations of the new technology, especially how different runs of the experiment might exhibit different (irrelevant) conditions that confound the results of the analysis. The successful project results in both new neuroscience results and in the development of new data science methods.

The human perspective reveals how aspects of the data analysis process, such as metadata, data provenance, data analysis workflows, and scientific reproducibility, are critical to modern scientific research. We need good software tools and infrastructure that can record, replicate, and facilitate how researchers interact with their data (26, 27). More broadly, the practice of data science is not just a single step of analyzing a dataset. Rather, it cycles between data preprocessing, exploration, selection, transformation, analysis, interpretation, and communication. One of the main priorities for data science is to develop the tools and methods that facilitate this cycle.

Summary

We presented a holistic view of data science, a view that has implications for practice, research, and education. It suggests the potential in integrating research that crosses the statistical, computational, and human boundaries. Furthermore, it puts into focus that, to solve real world problems, a data scientist will need to undertake tasks that are beyond their traditional training. Data science is more than the combination of statistics and computer science—it requires training in how to weave statistical and computational techniques into a larger framework, problem by problem, and to address discipline-specific questions. Holistic data science requires that we understand the context of data, appreciate the responsibilities involved in using private and public data, and clearly communicate what a dataset can and cannot tell us about the world.

1 Press G (May 28, 2013) A very short history of data science. *Forbes*. Available at https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#4d91ab9455cf. Accessed June 1, 2017.

2 Persons T, et al. (2016) *Data and Analytics Innovation: Emerging Opportunities and Challenges* (US Government Accountability Office, Washington, DC).

3 Donoho D (2015) *50 Years of Data Science*, Proceedings of the Tukey Centennial Workshop (Princeton). Available at https://pdfs.semanticscholar.org/f564/25ec56586dcfd2694ab83643e9e76f314e91.pdf. Accessed June 30, 2017.

4 Ridgway J (2015) Implications of the data revolution for statistics education. *Int Stat Rev* 84:528–549.

5 Dhar V (2013) Data science and prediction. *Commun ACM* 56:64–73.

6 Getoor L, et al. (2016) Computing research and the emerging field of data science. *CRA Bulletin*. Available at cra.org/data-science/. Accessed December 30, 2017.

7 Tukey JW (1962) The future of data analysis. *Ann Math Stat* 33:1–67.
8 Gelman A, et al. (2014) *Bayesian Data Analysis* (CRC, Boca Raton, FL), 2nd Ed.
9 Murphy K (2013) *Machine Learning: A Probabilistic Approach* (MIT Press, Cambridge, MA).
10 Barber D (2012) *Bayesian Reasoning and Machine Learning* (Cambridge Univ Press, Cambridge, UK).
11 Hastie T, Tibshirani R, Wainwright M (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations* (CRC, Boca Raton, FL).
12 Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349:255–260.
13 LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
14 Pearl J (2009) *Causality* (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
15 Imbens G, Rubin D (2015) *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction* (Cambridge Univ Press, Cambridge, UK).
16 Morgan S, Winship C (2015) *Counterfactuals and Causal Inference* (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
17 Sra S, Nowozin S, Wright S (2012) *Optimization for Machine Learning* (MIT Press, Cambridge, MA).
18 Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap* (Chapman & Hall/CRC, Boca Raton, FL).
19 Robert C, Casella G (2004) *Monte Carlo Statistical Methods*, Springer Texts in Statistics (Springer, New York), 2nd Ed.
20 Green PJ, Łatuszyński K, Pereyra M, Robert CP (2015) Bayesian computation: A summary of the current state, and samples backwards and forwards. *Stat Comput* 25:835–862.
21 Dean J, Ghemawat S (2008) MapReduce: Simplified data processing on large clusters. *Commun ACM* 51:107–113.
22 Bekkerman R, Bilenko M, Langford J, eds (2011) *Scaling Up Machine Learning: Parallel and Distributed Approaches* (Cambridge Univ Press, Cambridge, UK).
23 Jordan MI (2013) On statistics, computation and scalability. *Bern* 19:1378–1390.
24 Cleveland WS (2001) Data science: An action plan for expanding the technical areas of the field of statistics. *Int Stat Rev* 69:21–26.
25 Hardin J, et al. (2015) Data science in statistics curricula: Preparing students to "think with data." *Am Stat* 69:343–353.
26 Goodman A, et al. (2014) Ten simple rules for the care and feeding of scientific data. *PLOS Comput Biol* 10:e1003542.
27 Borgman CL, et al. (2015) Knowledge infrastructures in science: Data, diversity, and digital libraries. *Int J Digit Libr* 16:207–227.