The Future of Statistics and Data Science

Sofia C. Olhede

Department of Statistical Science, University College London

Patrick J. Wolfe

Department of Statistics, Purdue University

Abstract

The ubiquity of sensing devices, the low cost of data storage, and the commoditization of computing have together led to a big data revolution. We discuss the implication of this revolution for statistics, focusing on how our discipline can best contribute to the emerging field of data science.

Keywords: Algorithmic transparency, Data analysis, Data governance, Predictive analytics, Statistical inference, Structured and unstructured data

2010 MSC: 00-01, 99-00

1. Introduction

The Danish physicist Niels Bohr is said to have remarked: "Prediction is very difficult, especially about the future." Predicting the future of statistics in the era of big data is not so very different from prediction about anything else. Ever since we started to collect data to predict cycles of the moon, seasons, and hence future agriculture yields, humankind has worked to infer information from indirect observations for the purpose of making predictions.

Even while acknowledging the momentous difficulty in making predictions about the future, a few topics stand out clearly as lying at the current and future intersection of statistics and data science. Not all of these topics are of a strictly technical nature, but all have technical repercussions for our field. How might these repercussions shape the still relatively young field of statistics? And what

can sound statistical theory and methods bring to our understanding of the foundations of data science? In this article we discuss these issues and explore how new open questions motivated by data science may in turn necessitate new statistical theory and methods now and in the future.

Together, the ubiquity of sensing devices, the low cost of data storage, and the commoditization of computing have led to a volume and variety of modern data sets that would have been unthinkable even a decade ago. We see four important implications for statistics.

First, many modern data sets are related in some way to human behavior. Data might have been collected by interacting with human beings, or personal or private information traceable back to a given set of individuals might have been handled at some stage. Mathematical or theoretical statistics traditionally does not concern itself with the finer points of human behavior, and indeed many of us have only had limited training in the rules and regulations that pertain to data derived from human subjects. Yet inevitably in a data-rich world, our technical developments cannot be divorced from the types of data sets we can collect and analyze, and how we can handle and store them.

Second, the importance of data to our economies and civil societies means that the future of regulation will look not only to protect our privacy, and how we store information about ourselves, but also to include what we are allowed to do with that data. For example, as we collect high-dimensional vectors about many family units across time and space in a given region or country, privacy will be limited by that high-dimensional space, but our wish to control what we do with data will go beyond that. At the same time, a key problem we face in the wish to regulate or control what we do with data is the increasing complexity of algorithms. On one hand, the concepts behind algorithmic thinking can be challenging to explain in layman's terms so that anyone–for example, a jury of one's peers–can understand the principles by which an algorithm functions, and on the other hand algorithms themselves are becoming so complex that when coupled with the vast amounts of data we can now employ, it can at times be very hard to understand the rationale for their output. Clearly we require a way

to balance algorithmic interpretability with predictive performance—especially if we are making important decisions on the basis of a given algorithm.

Third, the growing complexity of algorithms is matched by an increasing variety and complexity of data. Data sets now come in a variety of forms that can be highly unstructured, including images, text, sound, and various other new forms. These different types of observations have to be understood together, resulting in multimodal data, in which a single phenomenon or event is observed through different types of measurement devices. Rather than having one phenomenon corresponding to single scalar values, a much more complex object is typically recorded. This could be a three-dimensional shape, for example in medical imaging, or multiple types of recordings such as functional magnetic resonance imaging and simultaneous electroencephalography in neuroscience. Data science therefore challenges us to describe these more complex structures, modeling them in terms of their intrinsic patterns.

Finally, the types of data sets we now face are far from satisfying the classical statistical assumptions of identically distributed and independent observations. Observations are often "found" or repurposed from other sampling mechanisms, rather than necessarily resulting from designed experiments. They may correspond to a mixture of many heterogeneous populations, with the differences within populations proving challenging to analysis. To remove unwanted artifacts, extensive preprocessing (sometimes aptly described as "data wrangling" [1]) must often take place—leading to an 80/20 rule of thumb amongst practitioners suggesting that four times as much time should be set aside for wrangling than for actual analysis and inference. The complexities of heterogenous, unstructured data requiring substantial preprocessing is challenging to statistical modelers, and calls for new approaches to theoretical concepts and methodological developments, as well as the pipeline that turns these into rigorous applications of modern statistics in practice. Our field will either meet these challenges and become increasingly ubiquitous, or risk rapidly becoming irrelevant to the future of data science and artificial intelligence.

2. Missing the Data Science Boat?

As part of the thirty-eighth Conference on Stochastic Processes and their Applications in 2015, we took part in a debate at the Oxford Union, where we were asked to take opposing side on the following motion: *This house believes that the mathematical scientists will miss the data science boat.* The argument of the "for" side was not that statistics would be prevented from climbing aboard, but rather that statisticians might willingly choose to maroon themselves on shore, forsaking messy data science challenges for the purity of fundamental theoretical challenges in stylized circumstances that have been abstracted away from the reality of modern-day data sets. The argument of the "against" side was that statistics was so integral to data science, the boat would sink without it! Rather dramatically, neither side won, as the house vote of over 100 people left us at a perfect draw. Perhaps neither side won the debate simply because both arguments have merit.

This makes for a rather amusing anecdote, but underneath the superficial there is an element of seriousness. A significant portion of what we have come to call data science is not statistics *per se*, and does not emphasize modeling and inference skills that statisticians have long been trained to value [2]. Part of data science is architecting, understanding how to store and access data; part of it is algorithms, understanding how to implement a chosen analysis method; and part of it is simple common sense. None of these aspects is necessarily well suited to developing statistical theory, though some thought to implementation and analysis trade-offs has started to appear in the literature [3].

The statistical theory and methods that our field will develop in an era of big data must be adapted to the types of data that we encounter in the world around us—or we risk putting ourselves in grave danger of becoming irrelevant. What is more, statistics also needs to adjust to other societal constraints and implications that are becoming apparent. A large part of this is developing an awareness and general broad understanding of the extent to which data will affect everyone's daily lives—not only through technology, but also through

policy, commerce, privacy, and trust.

3. Data Governance

The availability of "big data" poses great opportunities for societal gain, but also threats. Such data sets often comprise observations collected from, or about, human subjects. The potential privacy implications of such volumes of personally identifiable information are enormous, and consequently it has motivated (for example) the development of statistical methods that can calculate meaningful summaries from anonymized data. Ensuring that our field helps contribute in this and other ways to an informed public discourse is crucial for the future potential of data science to be realized—otherwise abuse and misuse of data can generate strong public mistrust. Existing principles for working with human subjects data at smaller scales and levels of pervasiveness–such as the principle of informed consent—are under ever greater pressure thanks to the continuous technological developments of algorithms and analytics.

Part of the solution to this conundrum will inevitably be technological, or at least technical. Recent years have seen considerable innovation in combining statistics with encryption as a means to ensure privacy—for example, understanding how to do inference when encryption has already taken place [4, 5]. Additional technical challenges arise when we consider how to design fail-safe anonymization schemes and methods to analyze anonymized data.

Various international efforts have been launched to determine overarching principles in respect to the usage of personally identifiable information, and also to understand the consequences for analysis when individuals are granted the right to have their data removed from databases (for instance, as a consequence of the forthcoming European General Data Protection Regulation). For example, a recent professional society report [6] spells out recommendations concerning privacy and data access as well as control; similarly a recent British Academy and UK Royal Society report sets out principles for data governance [7, 8]. These documents recognize the rapid advances both of data-

enabled technologies and of data collection activities. As technology advances in this manner, with rapid and widespread adoption, the risk of a major public backlash is considerable.

4. Regulation and Algorithmic Transparency

Data collection and the governance thereof is therefore a subject of significant immediate concern. Rather more futuristically, the regulation of algorithms is also generating considerable debate. For example, the Association for Computing Machinery has weighed in with a statement on algorithmic transparency and accountability [9]. The new European General Data Protection Regulation details the rights of citizens, if affected by a particular algorithmic decision, to an explanation of why that decision was arrived at. In this way regulation is beginning to interact with the latest technological developments in data science.

The notion of algorithmic transparency may seem intuitively obvious, but there are clear problems with this notion when it is subject to the legal definition of the "explanation" to which citizens may be entitled. To arrive at, say, a prediction, we could posit an interpretable model (for example, logistic regression), fit it using explanatory variables, and then predict binary outcomes. If we have too many variables, then we could appeal to modern methods of model choice or sparsification, or we could even pre-process the set of explanatory variables–for example, using principal components analysis. One way or another, we could conceivably arrive at a model with clearly identified explanatory variables, and we would then know why we arrive at a particular predicted value when appealing to the model. To "explain" this model out of its mathematical context, we could give a quantitative description of which variables impacted the prediction or decision, and we could also explain how we arrived at the model itself (e.g., what choice of error metric we used when developing our inferential procedures).

Now suppose that things become gradually more complicated. Assume our mechanism for generating a prediction involves a complex set of iterated non-linear operations. We can still assess predictive performance by keeping sets of

data apart and held out, for example by having a separate training and testing data set. However the explanation as to why we make a given prediction is no longer very clear. We input data, which may be very high dimensional, into some algorithmic procedure, and that procedure outputs a prediction or a decision. The complex interactions which generated the prediction are based on (what may be) interpretable predictors, but the interpretation of their combination is unclear, and if we have very many predictors, their combined usage may correspond to an approximation of variables whose use could reasonably be perceived as discriminatory (for example, proxy variables for race or gender in determining an individual's employment prospects or creditworthiness for a loan). Currently, the more complex a predictive algorithm tends to be, the more difficulty arises when seeking a clear understanding of its mechanisms. Even if, say, our metric for good algorithmic performance in general is well understood– mean squared prediction error as measured on a held-out test data set relative to a training data set, for example–this does not guarantee that our understanding of the mechanism discovered in any given modeling problem yields any insight into "why" a particular prediction is made. A set of key tools remains to be discovered in this field, perhaps most importantly those which attempt to derive interpretability (such as [10], for example).

There are also interesting decision-theoretic problems that relate to our understanding of transparency. It may well be the case that in a given scenario, predictive error will increase as we make models more transparent. What is a reasonable trade-off between prediction error and transparency, and how can we formally study and determine this trade-off? By using proper mathematical methods, with explicit and quantitative optimization criteria, we can envision making this trade-off well defined, and hence eventually equally well understood.

5. Structured and Unstructured Data

Another area where modern statistics stands to contribute to data science is the analysis of large data sets that correspond to more than collections of

just single scalar observations. Such data sets can take the form of shapes such as curves or volumes, or even strongly non-Euclidean objects such as networks [11]. A network represents relationships (edges) between objects (nodes), enabling the modeling and analysis of everything from social networks to physical infrastructure. Recent years have shown how statistics can contribute to the broader field of network analysis, by formulating models that are sufficiently flexible to explain observed variation, but also tractable to analysis and forward simulation (e.g. [12, 13]). Owing in large part to the lack of any unique underlying vector space structure, many basic statistical questions remain open in the study of networks: how do assess and compare model fit, how do we generate more heterogeneous yet structured observations that better match network data sets, and how do we incorporate covariates and other non-network information into our inference mechanisms in a manner that is provably consistent?

Data science faces a number of questions when multiple observations of very large networks are considered: how to store them, how to access them efficiently, and how to compute meaningful summaries of the structures they contain. Statistics can contribute in this setting through the development of theory to quantify precisely what it means to have "more" observations: more replicates of a single network structure, denser observations in time, asympototics in network size and temporal evolution, and a greater degree of heterogeneity in node-to-node connectivity. A number of mathematical challenges stand in the way of defining more complex limiting objects that encompass these ideas, and statistics is needed to relate the network data sets we observe in practice to the corresponding mathematical objects and models.

These questions require theoretical insights and algorithmic advancements to be developed in tandem, especially when considering the computational feasibility of analyzing large networks. Some work is focused on proving results about methods that are known to be efficiently implementable; other work seeks to establish fundamental limits and optimality properties of idealized methods that might not be sufficiently scalable to implement on all but the smallest of networks. As the field matures, we would ideally seek a continuum of meth-

ods, from those that are linear in the number of nodes or edges and hence inherently scalable—with suboptimal but hopefully provable properties—to algorithms that scale with some power of the number of observations but may yield optimal properties. Unfortunately we are often in a scenario where we can prove something about methods that are not scalable, but not about those methods that are. Closing this gap is an important example of how statistics can contribute to a rigorous notion of data science with firm theoretical foundations.

6. Bias, Incompleteness, and Heterogeneity

We have seen how the statistical challenges presented by the present and future of data science look to be growing increasingly more difficult, demanding, and urgent. We now discuss a particularly critical area—that of missing and biased observations—which will require us to develop a new theory for data science. In traditional statistics, we being by specifying a sampling mechanism and a population from which a sample will be drawn. Despite enthusiastic claims that the availability of "all" data precludes the need for statistical models or sampling methodologies, the opposite is in fact true. The more we strive to make sense from repurposed or "found" data, where we may have limited access to information about the sampling design or population composition, the more crucial it is that we take considerations of bias and missingness into account [14]. But as we abandon well-designed experiments, and start to address the properties of unbalanced random designs, much of statistical theory lies underdeveloped. A classic example is the abandonment of "missing at random" assumptions [15], for example when the act of observation is correlated with the quantity that is the subject of inference [16]. For example, studying road conditions using self-reported smartphone data may tell one more about the distribution of smartphones in a city than about the conditions of its roads [17]. Given the promise of the vast quantities and types of data were are now able to collect, new statistical thinking must arise to help us make sense of non-ideal sampling paradigms and develop mechanisms to enable repeatable, defensible

inferential conclusions to be drawn (even if on a very limited basis).

Another important challenge is the analysis of populations with a high degree of heterogeneity [18]. Recognizing heterogeneity and trying to profit from it—by understanding when to aggregate, smooth, and average versus when to disaggregate and stratify—has become increasingly important in areas such as speech recognition technology and precision medicine [19, 20]. Modeling heterogeneity effectively, as well as correctly understanding the sampling replication properties of complex random objects, remains an outstanding problem [21]. Lacking a sufficient wealth of heterogeneity in our models means that we risk failing to reflect accurately all the potentially important structure in our data. Modeling and inference procedures specifically designed for these types of scenarios are desperately needed if data science is ever to be put on a firm inferential footing.

7. Discussion

Much of data science to date has focused on purely predictive "black box" tools rather than classical modeling, inference, and analysis. It is natural with richer sources of data to start by looking for patterns, rather than trying to fit specific models. Yet this remains problematic: how can we determine if patterns are significant if natural variation cannot be quantified by way of models?

The role of statistics is to make our understanding of observed phenomena quantitative and precise. A number of new problems challenge that task: data sets are taking ever more complex forms than previously has been the case. In addition observations are often made without proper experimental design, resulting in biased and incomplete data.

Yet these technical challenges are not sufficient to describe the full set of contributions which statistics must make to the field of data science. Much of "big data" concerns human subjects, raising ethical and governance challenges that we have a responsibility to help resolve. Our teaching of statistics will need to be adapted to encompass a broader degree of training in such areas.

It is still unclear how questions of ethics and algorithmic transparency will be

resolved through data governance. Many learned societies, professional organizations, and national academies of science are taking a strong interest. Solutions will inevitably involve technical developments and raise new methodological challenges; these will in turn require strong involvement from statistics.

In the end it is clear that algorithms and the decisions they derive from data will increasingly have an impact across nearly every aspect of society [22]. Many of these decisions will be automated, and harnessing the power of statistics and the efficiency gains of automated decision making can potentially be incredibly beneficial to our world at large. If, however, the inner workings of such procedures remains shrouded in mystery in the public eye, and determining the fairness of algorithmic decisions becomes hard, then data science will lose public trust. Opaque and purely predictive algorithms have served to demonstrate to all of us the power of large-scale analysis, but as automation stands to impact increasingly significant decisions, the need for greater scrutiny and transparency is becoming apparent. Our field has an clear and present window of opportunity to build new theory and methods to meet the current and future challenges of data science, which we must do to not risk missing the data science boat.

References

[1] N. Reid, Statistical science in the world of big data, Special Issue of Statistics and Probability Letters on 'The role of Statistics in the era of big data', to appear.

[2] D. Dunson, Statistics in the big data era: Failures of the machine, Special Issue of Statistics and Probability Letters on 'The role of Statistics in the era of big data', to appear.

[3] A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan, A scalable bootstrap for massive data, Journal of the Royal Statistical Society: Series B 76 (2014) 795–816.

[4] T. Graepel, K. Lauter, M. Naehrig, MIConfidential: machine learning on

encrypted data, in: International Conference on Information Security and Cryptology, Springer, Berlin, Germany, 2012, pp. 1–21.

[5] L. Aslett, P. Esperanca, C. C. Holmes, A review of homomorphic encryption and software tools for encrypted statistical machine learning (2015).

[6] IEEE, Ethically aligned design, v1 (2016).

[7] British Academy & Royal Society, Data management and use: Governance in the 21st century (2017).

[8] S. Olhede, R. Rodrigues, Fairness and transparency in the age of the algorithm, Significance 14 (2) (2017) 8–9.

[9] A. for Computing Machinery US Public Policy Council, Statement on algorithmic transparency and accountability (2017).

[10] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, Pattern Recognition 65 (2017) 211–222.

[11] I. L. Dryden, D. J. Hodge, Journeys in big data statistics, Special Issue of Statistics and Probability Letters on 'The role of Statistics in the era of big data', to appear.

[12] P. J. Bickel, A. Chen, A nonparametric view of network models and newman–girvan and other modularities, Proceedings of the National Academy of Sciences 106 (50) (2009) 21068–21073.

[13] S. C. Olhede, P. J. Wolfe, Network histograms and universality of blockmodel approximation, Proceedings of the National Academy of Sciences 111 (41) (2014) 14722–14727.

[14] W. Davies, How statistics lost their power–and why we should fear what comes next.

[15] R. J. A. Little, D. B. Rubin, Statistical analysis with missing data, 2014.

[16] D. R. Cox, Big data: Some statistical issues, Special Issue of Statistics and Probability Letters on 'The role of Statistics in the era of big data', to appear.

[17] K. Crawford, The hidden biases in big data.
URL hbr.org/2013/04/the-hidden-biases-in-big-data

[18] P. Bühlmann, N. Meinshausen, Magging: maximin aggregation for inhomogeneous large-scale data, Proceedings of the IEEE 104 (1) (2016) 126–135.

[19] E. A. Ashley, The precision medicine initiative: a new national effort, ournal of the American Medical Association 313 (21) (2015) 2119–2120.

[20] E. A. Ashley, Towards precision medicine, Nature Reviews Genetics 17 (2016) 507–522.

[21] P. Buhlmann, S. van der Geer, Statistics for big data: a perspective, Special Issue of Statistics and Probability Letters on 'The role of Statistics in the era of big data', to appear.

[22] S. Olhede, R. Rodrigues, The computer ate my personality, Significance 14 (3) (2017) 6–7.