Challenges and Opportunities in Statistics and Data Science: Ten Research Areas

Xuming He^{1,2} Xihong Lin^{3,4}

¹Department of Statistics, College of Literature Science and the Arts, University of Michigan, Ann Arbor, Michigan, United States of America,

²H.C. Carver Collegiate of Statistics, University of Michigan, Ann Arbor, Michigan, United States of America,

³Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

⁴Department of Biostatistics, Harvard T. H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States of America

Published on: Sep 30, 2020

DOI: <https://doi.org/10.1162/99608f92.95388fcb>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

As a discipline that deals with many aspects of data, statistics is a critical pillar in the rapidly evolving landscape of data science. The increasingly vital role of data, especially big data, in many applications, presents the field of statistics with unparalleled challenges and exciting opportunities. Statistics plays a pivotal role in data science by assisting with the use of data and decision making in the face of uncertainty. In this article, we present 10 research areas that could make statistics and data science more impactful on science and society. Focusing on these areas will help better transform data into knowledge, actionable insights and deliverables, and promote more collaboration with computer and other quantitative scientists and domain scientists.

Keywords: biostatistics, causal inference, cloud computing, data visualization, distributed inference and learning, differential privacy, federated learning, integrative analysis, interpretable machine learning, replicability and reproducibility, scalable statistical inference, study design

Introduction

As we enter into the digital age, data science has been growing rapidly into a vital field that has revolutionized many disciplines along with our everyday lives. Data science integrates a plethora of disciplines to produce a holistic, thorough, and insightful look into complex data, and helps to effectively sift through muddled masses of information to extract knowledge and deliver actionable insights and innovations. The National Science Foundation (NSF) report *Statistics at Crossroads: Who Is for the Challenge* (He et al., 2019) calls for statistics to play a central leadership role in data science, in partnership with other equally important quantitative disciplines such as computer science and informatics. Indeed, as a data-driven discipline, the field of statistics plays a pivotal role in advancing data science, especially by assisting with the keystone of data science, *analysis*, for use in decision making in the face of uncertainty.

In computer science, the appearance of the term data science has been traced back to Naur (1974) who provided a more specific definition than those used today, as “the science of dealing with data once they have been established.” In statistics, the term can be attributed to C. F. Jeff Wu, whose inaugural lecture titled “Statistics = Data Science?” for his appointment to the H. C. Carver Professorship at the University of Michigan in 1997 called for shifting the focus of statistics to center on “large/complex data, empirical-physical approach, representation and exploitation of knowledge.” Cleveland (2001), Breiman (2001), and Donoho (2017) were also instrumental to the modern conception of the field. Today, data science, in the spirit of data + science, has become an interdisciplinary enterprise about data-enabled discoveries and inference with scientific theory and methods, algorithms, and systems. A number of departments of statistics and data science, as well

as schools and institutes of data science, have been established in the United States and around the world. Data science undergraduate majors and master's degree programs are now available in many leading institutions.

The emergence of data science has begun to reshape the research enterprise in the field of statistics and biostatistics. In this article, we present a list of research challenge areas that have piqued the interest of data scientists from a statistical perspective, and vice versa. The selection of these areas takes advantage of the collective wisdom in the NSF report quoted, but also reflects the personal research experience and views of the authors. Our list, organized in no particular order, is by no means exhaustive, but aims to stimulate discussion. We certainly expect other important research areas to emerge and flourish. We also excuse ourselves for not citing references related to the research discussions; the body of work from which we have drawn inspiration is simply too large for this article.

Ten Research Areas

1. Quantitative Precision-X

From personalized health to personalized learning, a common research goal is to identify and develop prevention, intervention, and treatment strategies tailored toward individuals or subgroups of a population. Identification and validation of such subgroups using high-throughput genetic and genomic data, demographic variables, lifestyles, and other idiosyncratic factors is a challenging task. It calls for statistical and machine learning methods that explore data heterogeneity, borrow information from individuals with similar characteristics, and integrate domain sciences. Subgroup analysis calls for the development of integrated approaches to subgroup identification, confirmation, and quantification of differential treatment effects by using different types of data that may come from the same or different sources. Dynamic treatment regimes are increasingly appreciated as adaptive and personalized intervention strategies, but quantification of uncertainty requires more studies, as well as building treatment regimes in the presence of high-dimensional data.

2. Fair and Interpretable Learning and Decision Making

Machine learning has established its value in the data-centric world. From business analytics to genomics, machine learning algorithms are increasingly prevalent. Machine learning methods take a variety of forms; some are based on traditional statistical tools as simple as principle component analysis, while others can be ad hoc and are sometimes referred to as black boxes, which raises issues such as implicit bias and interpretability. Algorithmic fairness is now widely recognized as an important concern, as many decisions rely on automatic learning from existing data. One may argue that interpretability is of secondary importance if prediction is the primary interest. However, in many high-stake cases (e.g., major policy recommendations or treatment choices involving an invasive operation), a good domain understanding is clearly needed to ensure the results are interpretable and insights and recommendations are actionable. By promoting fair and interpretable machine learning methods and taking ethics and replicability as an important metric for evaluation, statisticians have much to contribute to data science.

3. Postselection Inference

Statistical inference is best justified when carefully collected data (and an appropriately chosen model) are used to infer and learn about an intrinsic quantity of interest. Such a quantity (e.g., a well-defined treatment effect) is not data- or model-dependent. In the big data era, however, statistical inference is often made in practice when the model and sometimes even the quantity of interest is chosen after the data are explored, leading to postselection inference. Interpretability of such quantities and validity of postselection inference have to be carefully examined. We must ensure that postselection inference avoids the bias from data snooping, and maintains statistical validity without unnecessary efficiency losses, and moreover that the conclusions from such inference have a high level of replicability.

4. Balance of Statistical and Computational Efficiencies

When we have limited data, the emphasis on statistical efficiency to make the best use of the available data has naturally become an important focus of statistics research. We do not think statistical efficiency will become irrelevant in the big data era; often inference is made locally and the relevant data that are available to infer around a specific subpopulation remain limited. On the other hand, useful statistical modeling and data analysis must take into account constraints on data storage, communication across sites, and the quality of numerical approximations in the computation. An ‘optimally efficient’ statistical approach is far from optimal in practice if it relies on optimization of a highly nonconvex and nonsmooth objective function, for instance. The need to work with streaming data for real-time actions also calls for a balanced approach. This is where statisticians and computer scientists, as well as experts from related domains (e.g., operation research, mathematics, and subject-matter science) can work together to address efficiency in a holistic way.

5. Cloud-Based Scalable and Distributed Statistical Inference

It is of high importance to develop practical scalable statistical inference for the analysis of real-world massive data. This requires multifaceted strategies. Examples include sparse matrix construction and manipulation, distributed computing and distributed statistical inference and learning, and cloud-based analytic methods. A range of statistical methods have been developed for analysis of high-dimensional data with attractive theoretical properties. However, many of these methods are not readily scalable in real-world settings for analyzing massive data and making statistical inference at scale. Examples include atmospheric data, astronomical data, large-scale biobanks with whole genome sequencing data, electronic health records, and radiomics. Statistical and computational methods, software, and at-scale modules that are suitable for cloud-based open-source distributed computing frameworks, such as Hadoop and Spark, need to be developed and deployed for analyzing massive data. In addition, there is a rapidly increasing trend of moving toward cloud-based data sharing and analysis using the federated data ecosystem (Global Alliance for Genomics and Health, 2016), where data may be distributed across many databases and computer systems around the world. Distributed statistical inference will help researchers to virtually connect, integrate, and analyze data

through software interfaces and efficient communications that allow seamless and authorized data access from different places.

6. Study Design and Statistical Methods for Reproducibility and Replicability

Reproducibility and replicability in science is pivotal for improving rigor and transparency in scientific research, especially when dealing with big data (National Academies of Sciences, Engineering, and Medicine, 2019). This includes data reproducibility, analysis reproducibility/stability, and result replicability (Lin, 2019). A rigorous study design and carefully thought-out sampling plans facilitate reproducible and replicable science by considering key factors at the design stage, including incorporation of both a discovery and a replication phase. Common data models, such as that used in the large scale All of Us Research Program (2019), have become increasingly popular for building federated data ecosystems, especially using the cloud, to assist with data standardization, quality control, harmonization, and data sharing, as well as the development of community standards. Although issues with replicability using statistical significance based on a classical p -value cutoff of, say, 0.05, have been identified and widely debated, there has not been much consensus on what the new norm should be and how to make statistical significance replicable between studies. Limited work has been done on developing formal statistical procedures to investigate whether findings are replicable, especially in the presence of a large number of hypotheses. More such efforts are needed, in collaboration with informaticians and computer scientists.

7. Causal Inference for Big Data

Causal inference for conventional observational studies has been well developed within the potential outcome framework using parametric and semiparametric methods, such as maximum likelihood estimation (MLE), propensity score matching, and G-estimation. As big data are often observational in the real world, they have brought many emerging challenges and opportunities for causal inference, such as causal inference for network data, construction of real-time individualized sequences of treatments using mobile technologies, and adjustment for high-dimensional confounders. For example, for infectious disease network data, and social network data, such as Facebook data, subjects are connected with each other. As a result, the classical causal inference assumption, the stable unit treatment value assumption (SUTVA), which assumes independent subjects, will not hold. Machine learning-based causal inference procedures have emerged in response to such issues, and integration of these procedures into the causal inference framework will be important.

8. Integrative Analysis of Different Types and Sources of Data

Massive data often consist of different types of data from the same subjects or from different subjects and sources. For example, the U.K. biobank collected whole genome array and sequencing (soon to be available) data, electronic health records (EHRs), epidemiological, biomarker, activity monitor, and imaging data from about 500,000 study participants. Furthermore, data from other sources and different subjects (not from the U.K. biobank) are also available, including data from genome-wide association studies, genomic data such as

ENCODE and GTEx data, and drug target data from Open Targets (Koscielny et al., 2017). There is a strong need to develop statistical methods and tools for integrating different types of data from one or multiple sources. Examples of such methods include causal mediation analysis, Mendelian randomization, and transportable statistical methods for data linkage. It is important to emphasize that statistical methods for data integration need to be driven by scientific questions. Blanket-style data integration methods are likely to be less useful. Close and deep collaborations between statisticians and domain researchers cannot be overemphasized.

9. Statistical Analysis of Privatized Data

With growing emphasis on privacy, data sanitation methods, such as differential privacy, will remain a challenge for statistical analysis. Census data in particular, which are used frequently in social science, public health, internet, and many other disciplines, have raised serious questions regarding the adequacy of available theory and methods for ensuring a desired level of privacy and precision. Current differential privacy frameworks are designed to protect data privacy from any kind of inquiries, which is necessary to guard against the most sophisticated hacking, while still allowing for valid analysis. Researchers advocating differential privacy need to understand the practical concerns of data utility in determining how to balance privacy and precision as well as adoption of federated systems. Indeed, another area of data privacy analytics is federated statistical and machine learning, which allows for analyzing data that cannot leave individual warehouses, such as banks and health care systems. Through building a common data model and analysis protocol, statisticians and data scientists can bring the analysis protocol to the data instead of the traditional way of bringing centralized data to the analysis protocol. To protect privacy, data in different sites are analyzed individually using the common analysis protocol, and the updates are then combined through single or multiple communications. The field of privacy-preserved data science is rapidly evolving. Theoretical and empirical investigations that reflect real-world privacy concerns, as well as approaches to address ethics, social goods, and public policy in computer science and statistics, will be a hallmark of future research in this area.

10. Emerging Data Challenges

Statistical research needs to keep pace with the growing needs for the analysis of new and complex data types, including deliberately generated false data and misinformation. Statisticians have in recent years taken up the challenge in developing new tools and methods for emerging data types, including network data analysis, natural language processing, video, image, and object-oriented data analysis, music, and flow detection. Statisticians need to embrace data engineering to address data challenges. Emerging challenges arising from adversarial machine learning argue for engagement of statisticians, too, and this is becoming more important in the age of information and misinformation. In addition, data visualization and statistical inference for data visualization (e.g., addressing the question ‘Is what we see really there?’) will play increasingly greater roles in data science, especially with massive data in the digital age.

Closing Remarks

Statistics as an ever-growing discipline has always been rooted in and advanced by real-world problems. Statisticians have played vital roles in the agricultural revolution, the industrial revolution, the big data era, and now in the broad digital age. Statistics cannot live successfully outside data science, and data science is incomplete without statistics. We believe that research in statistics and biostatistics should respond to the major challenges of our time by keeping a disciplinary identity, promoting valuable statistical principles, working with other quantitative scientists and domain scientists, and pushing boundaries of data-enabled learning and discovery. To do this well, we need substantially more young talents to join the statistical profession as well as other disciplines that contribute to data science, especially computer science, informatics, and ethical studies. This in turn calls for earlier and broader statistical and scientific data education on a global scale. We therefore encourage members of our field to lead, collaborate, and communicate in data science research and education with open minds, not only as statisticians, but also as scientists.

Disclosure Statement

Xuming He and Xihong Lin have no financial or non-financial disclosures to share for this article.

References

- All of Us Research Program Investigators. (2019). The “All of Us” Research Program. *New England Journal of Medicine*, 381(7), 668–676. <https://doi.org/10.1056/NEJMSr1809937>
- He, X., Madigan, C., Wellner, J., & Yu, B. (2019). *Statistics at a crossroads: Who is for the challenge?* NSF Workshop report. National Science Foundation.
https://www.nsf.gov/mps/dms/documents/Statistics_at_a_Crossroads_Workshop_Report_2019.pdf
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231.
<https://doi.org/10.1214/ss/1009213726>
- Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1), 21–26. <https://doi.org/10.1111/j.1751-5823.2001.tb00477.x>
- Donoho, D. (2017). 50 years of data science. *Journal of Computational & Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Global Alliance for Genomics and Health. (2016). A federated ecosystem for sharing genomic, clinical data. *Science*, 352(6291), 1278–1280. <https://doi.org/10.1126/science.aaf6162>

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., & Pierleoni, A. (2017). Open targets: A platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45(Database Issue), D985–D994. <https://doi.org/10.1093/nar/gkw1055>

Lin, X. (2019). *Reproducibility and replicability in large scale genetic studies*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press.

Naur, P. (1974). *Concise survey of computer methods*. Studentlitteratur.

Wu, J. (1997). *Statistics = Data science?* <https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf>

©2020 Xuming He and Xihong Lin. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.