Harvard Data Science Review • Issue 2.3, Summer 2020

Insightful Data Science

Hongyu Zhao^{1,2,3}

¹Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, Connecticut, United States of America,

²Department of Genetics, Yale School of Public Health, Yale University, New Haven, Connecticut, United States of America,

³Department of Statistics and Data Science, Yale University, New Haven, Connecticut, United States of America

Published on: Sep 30, 2020

DOI: <https://doi.org/10.1162/99608f92.34d1d59e>

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

I would like to first thank Jeannette Wing, Xuming He, and Xihong Lin for sharing their visions on important research areas in data science, and Xiao-Li Meng for giving me the opportunity to comment on these two excellent articles. The areas outlined by these three data science leaders highlight many challenges and opportunities for aspiring data scientists. Coming from different disciplinary backgrounds (statistics for Dr. He, Biostatistics for Dr. Lin, and computer science for Dr. Wing), coupled with their own research interests, it is no surprise that there are different emphases in the two lists, with He and Lin's list including inferential problems (e.g., postselection inference and study design) and Wing's list including computing systems and data life cycles. It is also a pleasant surprise that there is a large overlap between the two lists (*a statistically significant overlap*), including causal inference, scalability, data privacy, heterogeneous data, theoretical analysis of algorithms and methods, and the need to incorporate domain knowledge. This may have resulted from the active dialogues between statisticians, computer scientists, and researchers from other relevant disciplines on how to define data science, such as those during the planning of curriculum for both undergraduate and graduate data science programs at many universities. It is now well recognized that the field of data science is truly interdisciplinary and as such, needs synergistic efforts from all the fields involved, including statistics, computer science, mathematics, informatics, and domain experts. As Wing elaborated well in her article, it may be too early to define data science as a discipline but a list of top research areas is the first step toward defining the scope and directions of data science in the near future.

In more established disciplines related to data science, such as statistics, biostatistics, and computer science, although many research topics are driven by practical problems, most top journals in these fields still focus on methodology and algorithm developments and theoretical investigations of these methods and algorithms. There tends to be more emphasis on disciplinary contributions than real-world impacts for faculty evaluations and promotions, including many biostatistics departments that are embedded in either public health schools or medical schools. However, for data science to grow, it has to deliver insights and solutions to real-world problems. For example, in the booming field of genomics, where next-generation sequencing and single-cell technologies have generated enormous amounts of data, unless a statistical or computer science solution is implemented and readily available to scientists to manage (often through cloud), analyze, visualize, and interpret these data, the great efforts going into developing methodology (and its theoretical proof of its optimality) will have little or no impacts beyond a publication in a top disciplinary journal. Indeed, the collection, processing, analysis, and dissemination of diverse types of -omics data touch across many research areas on the two lists of data science areas. At Yale University, the Center for Biomedical Data Science was established to bring researchers from biostatistics, computer science, informatics, bioengineering, health economists, and more importantly domain experts in different disease areas to mine and model biomedical data with unprecedented ability.

Data science is clearly pushing statisticians and biostatisticians out of their comfort zone with the need to consider computational issues (in addition to statistical optimality) and the management of large and complex

data in methodology development and theoretical investigations. On the other hand, data science also demands computer scientists to inject inference into their algorithms. These are well reflected in the two lists. More importantly, both statisticians and computer scientists have to seriously address interpretability and causal inference problems. We also need to do trustworthy data science (including FAIR data—defined following—and reproducible and replicable algorithms) and consider ethical and privacy issues, which are very new in the data science era.

Summarizing the important research areas proposed in these two articles, it is clear that data science has to be **INSIGHTFUL** to flourish and to be able to make real impacts on science, technology, business, and society. ‘**I**’ is for interpretability in that the models and tools developed by data scientists have to be interpretable (versus a black box) so that domain knowledge can be incorporated in data analysis, and domain knowledge can also be advanced through data science. A common notation in statistics, ‘**N**’ is for sample size. As implicated from the first area proposed by He and Lin, “Quantitative Precision-X,” the N in data science is not fixed but can be rather fluid. N can range from very large to very small depending on the specific problems to be addressed. We should not take for granted that N is always large for data science problems and creativity is needed to ‘borrow’ information across samples to increase the effective size of N . ‘**S**’ is for scalability, a key feature of data science, which as a method will not be useful if it cannot be scaled up to deal with real-world data. The second ‘**I**’ is for integration because diverse sources and technologies are often used to collect data to answer a specific question, and there is a great need to integrate these data. ‘**G**’ is for generalizability, which is related to both practical and theoretical assessment of model performance on future samples. ‘**H**’ is for heterogeneity, which is related to the N issue because we cannot treat the samples as independently and identically distributed for the observations from large data sets. In fact, heterogeneity is often the focus of many problems, such as the identifications of cell types from millions of single-cell data points. ‘**T**’ is for trustworthiness that encompasses both replicability and reproducibility as well as the need to address ethical and privacy issues so that the tools developed can be really trusted by the users. ‘**F**’ is for FAIR data, which refer to data that meet principles of findability, accessibility, interoperability, and reusability. The lack of FAIR data will limit the impact of data science. ‘**U**’ is for usability in that the results obtained from data science tools have to be actionable to impact science, business, and policies. ‘**L**’ is for the life cycle of data in that we should be all engaged in study design, data collection, curation, and maintenance.

Overall, **INSIGHTFUL** data science should allow scientists, policymakers, and business leaders to make sound causal inference from highly complex and heterogeneous data. The field of real-world evidence, which plays an increasingly important role in health care decisions, is one true test case for data science. Its importance is exemplified by the many decisions that have to be made without conducting randomized clinical trials with the current pandemics. With the accumulations of BIG data, we need to develop **INSIGHTFUL** data science that can guide us with reasoning through a transparent and trustworthy process. In addition, as argued well in the National Science Foundation report on *Statistics at Crossroads: Who Is for the Challenge*, we need to really value and promote future data scientists who are committed to and making significant contributions to

important data science questions, such as those provided in these two articles. Some of these topics may not be appreciated in the current discipline-based university systems.

I thank again the thoughtful lists prepared by Drs. He, Lin, and Wing, and I am sure these two articles will inspire and direct young researchers to many exciting and rewarding paths that will define what data science is in the future.

Disclosure Statement

Hongyu Zhao has no financial or non-financial disclosures to share for this article.

©2020 Hongyu Zhao . This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.